Univerza v Mariboru

MEDICINSKA FAKULTETA

DOKTORSKA DISERTACIJA

ASOCIACIJSKA ANALIZA NA CELOTNEM GENOMU PRI SLOVENSKIH BOLNIKIH S KRONIČNO VNETNO ČREVESNO BOLEZNIJO

APRIL 2013

MITJA MITROVIČ

Univerza v Mariboru

MEDICINSKA FAKULTETA

Doktorska disertacija

Asociacijska analiza na celotnem genomu pri slovenskih bolnikih s kronično vnetno črevesno boleznijo

April 2013

Mitja Mitrovič

Mentor: izr. prof. dr. Uroš Potočnik
Somentorja: doc. dr. Rinse K. Weersma
prof. dr. Pavel Skok

»Science is a way of trying not to fool yourself.
The first principle is that you must not fool yourself
and you are the easiest person to fool.«

- Richard P. Feynman

»The capacity to blunder slightly is the real marvel of DNA.
Without this special attribute,
we would still be anaerobic bacteria
and there would be no music.«

- Lewis Thomas

VSEBINSKO KAZALO

Str.

POVZETEK

V preteklih desetletjih je bil v raziskave in razumevanje patogeneze KVČB vložen precejšen napor, vendar natančni vzroki nastanka bolezni še niso povsem pojasnjeni. Po zdaj najbolj uveljavljeni hipotezi se bolezen pojavi zaradi pretiranega imunskega odziva na normalno črevesno mikrofloro pri gensko dovzetnih posameznikih. Bolniki s KVČB zbolijo za eno od dveh precej podobnih oblik kroničnega vnetja črevesja: ulceroznim kolitisom (UK) ali Crohnovo boleznijo (CB). V 10−15 odstotkih primerov pa bolniki ne kažejo jasnih kliničnih, endoskopskih in histoloških značilnosti CB ali UK, in govorimo o intermediarnem kolitisu (IK). Največjo pogostost bolezni so ugotovili v Kanadi in Severni Evropi, najmanjšo pa v deželah v razvoju in v Aziji, pri čemer je UK pogostejši kot CB. Za slovensko populacijo sicer ni konkretnih epidemioloških podatkov, vendar je bilo ob koncu 90. let v Sloveniji 1150 bolnikov s KVČB, kar nas uvršča na dno evropske lestvice. Na nastanek bolezni vplivajo različni dejavniki okolja, iz epidemioloških raziskav družin in dvojčkov pa nedvomno izhaja, da k tveganju za KVČB prispevajo tudi genski dejavniki. Kljub številnim dokazom o genskem ozadju bolezni se je natančno določanje genskih determinant začelo šele v zadnjih petnajstih letih. Temeljilo je na številnih raziskovalnih pristopih. Asociacijske študije na celotnem genomu (GWAS) so z odkritjem povezav med KVČB in 99 lokusi največ prispevale k razumevanju genskega ozadja bolezni. Med pomembnejše izzive v obdobju po GWAS spadata odkrivanje vzročnih variant in analiza njihovih funkcionalnih posledic, s tem pa posledično tudi rešitev problema manjkajočega dednostnega deleža pri kompleksnih boleznih. Zaradi tega so bili predlagani številni pristopi, kot so npr. fino kartiranje, analiza eQTL-ov in GWAS pri družinah obolelih, vključno s preučevanjem učinkov starševskega izvora (angl. parent of origin − POO) na izražanje genov. Ob tem pa so številne raziskovalne skupine, ki se ukvarjajo z genetiko KVČB, pod okriljem konzorcija International Inflammatory Bowel Disease Consortium (IIBDGC) združile moči v projektu ImmunoChip (iCHIP), ki predstavlja tarčno mikromrežo DNK za analiziranje imunsko posredovanih bolezni. V raziskavi smo izvedli z iCHIP-om prvo tarčno študijo GWAS pri 227 bolnikih s KVČB in 210 zdravih posameznikih. Zaradi razmeroma omejenega števila preiskovancev sicer nismo imeli zadostne statistične moči za odkrivanje statistično značilnih povezav, vendar smo nakazali na nominalne povezave z boleznijo na kromosomih 1, 5, 8, 13 in 16. V sodelovanju z drugimi člani projekta iCHIP, ki je v metaanalizi vključeval približno 75.000 bolnikov s KVČB in zdravih posameznikov, smo prispevali k odkritju 71 novih statistično značilnih povezav z boleznijo, tako da se je skupno število bolezenskih lokusov z 99 povzpelo na 163. Glede odkrivanja skupnih mehanizmov kompleksnih bolezni je pomembna tudi ugotovitev, da je več kot dve tretjini lokusov KVČB impliciranih v patogenezi drugih imunsko posredovanih bolezni, kot so ankilizirajoči spondilitis, psoriaza in mikobakterijske okužbe. V raziskavi smo ločeno izvedli tudi asociacijsko analizo SNP-jev genov *NOD2* in *IL23R* ter prvič povezali omenjena gena s patogenezo KVČB pri slovenskih bolnikih. Genotipske podatke smo pridobili z optimizacijo metode analize DNK s talilnimi krivuljami visoke ločljivosti (HRMA), ki se je izkazala kot hitra, preprosta, natančna in cenovno ugodna metoda za odkrivanje redkih mutacij in za gensko tipizacijo. V študiji smo s kartiranjem *cis*-eQTL-ov v regiji s tremi kandidatnimi geni (*IL18R1*, *IL18RAP* in *IL1RL1*) ugotovili, da polimorfizma rs10178214, rs1041973 in redka varianta ccc-2-102248784-A-G statistično značilno vplivajo na izražanje gena *IL1RL1* v periferni venski krvi. Nismo pa opazili statistično značilnega izražanja tarčnih genov v črevesnih biopsijah bolnikov s CB. V raziskavi smo pri nizozemskih bolnikih s KVČB odkrili omejen učinek POO genov *IL12B*, *PRDM1* in *NOD2*, ki pa nam ga v večji neodvisni kohorti ni uspelo replicirati. Nominalno značilen učinek POO smo opazili tudi za gen *IL10* pri Indijcih. S kombinacijo strojnega učenja s podpornimi vektorji in bioinformatske analize smo odkrili enega od genov (*NCAM1*), ki najverjetneje vpliva na raznolikost odziva bolnikov s CB na zdravljenje s kortikosteroidi.

Rezultati naše študije so delno osvetlili zapleten genski sestav pri KVČB. Za potrditev nekaterih rezultatov so potrebne dodatne asociacijske analize in metaanalize z večjim številom preiskovancev. Pojasnitev mehanizmov in sprožilcev, s katerimi so posamezni kandidatni geni udeleženi v patogenezi KVČB, pa zahteva dodatne funkcijske študije na molekularni ravni. V prihodnjih raziskavah bo najverjetneje več pozornosti namenjene raziskavam interakcij mukusnega imunskega sistema in črevesne mikroflore.

Ključne besede: kronična vnetna črevesna bolezen, asociacijska analiza na celotnem genomu, ImmunoChip

UDK: 616.34-002-056.7:601.4.008.7(043.3)

SUMMARY

Genome-wide association analysis of Slovenian inflammatory bowel disease patients

Crohn’s disease (CD) and ulcerative colitis (UC) are the two main forms of chronic relapsing inflammatory bowel diseases (IBD). The etiology of IBD has been extensively studied in the past few decades however, disease pathogenesis is not fully understood. It is a chronic disease characterized by recurring inflammation of the gut, and is thought to arise in response to the commensal microflora in a genetically susceptible host. IBD is believed to be associated with industrialization of nations, with the highest incidence rates and prevalence of IBD in North America and Northern Europe. There are no precise epidemiological data available for Slovenians, yet in the late 90. there were 1.190 individuals affected with IBD. The abundance of epidemiological data is suggesting influence of different environmental and genetic determinants in the disease pathogenesis. Complex disease genetics has been revolutionised in recent years by the advent of genome-wide association (GWA) studies. The chronic inflammatory bowel diseases (IBDs), Crohn's disease and ulcerative colitis have seen notable successes culminating in the discovery of 99 published susceptibility loci/genes to date. One of the future challenges in this post-GWAS era is to identify potential sources of the remaining heritability. Such sources may include common variants with limited effect size, rare variants with higher effect sizes, or even more complicated mechanisms such as epistatic and epigenetic interactions. Many approaches to meet these challenges have been suggested, including fine-mapping, eQTL mapping and family studies, focused on parent of origin (POO) analysis. To fully unravel the hidden heritability of IBD, collaborations between genome research centers are crucial, since the solutions to identify the hidden heritability are either costly or require a huge number of cases and controls. The IIBDGC is a good example of what can be achieved by performing large meta-analyses, and it is currently joining forces in the ImmunoChip project (iCHIP), a custom made array encompasing loci from GWAS of 12 immune-mediated diseases. We conducted iCHIP based association analysis on 227 IBD patients and 210 healthy controls. Unfortunately, due to small sample size we did not have sufficient power to detect significant associations. However we show a trend towards association on chromosomes 1, 5, 8, 13 and 16. In collaboration with other IIBDGC members we expanded on the knowledge of relevant pathways by undertaking a meta-analysis of Crohn's disease and ulcerative colitis genome-wide association scans, followed by extensive validation of significant findings, with a combined total of more than 75,000 cases and controls. We identify 71 new associations, for a total of 163 IBD loci, that meet genome-wide significance thresholds. Many IBD loci are also implicated in other immune-mediated disorders, most notably with ankylosing spondylitis and psoriasis. We also observe considerable overlap between susceptibility loci for IBD and mycobacterial infection. We also conducted a separate association analysis for SNPs in genes *NOD2* and *IL23R*, which were associated with Slovenian IBD patients for the first time. Genotypes were obtained by high-resolution melting analysis (HRMA). The results of this study suggest that HRMA yields significant savings on analysis time and costs and has proven as a simple high-throughput technique for screening for polymorphisms. In this study, we also conducted eQTL mapping of *IL18RAP* locus, which contains 3 candidate genes *IL18R1*, *IL18RAP* in *IL1RL1.* Polymorphisms rs10178214, rs1041973 and a rare variant ccc-2-102248784-A-G, were influencing expression of *IL1RL1* in whole-blood samples. We did not observe any changes of gene expression of the candidate genes when intestinal biopsies were taken into consideration. We identified POO effects for *NOD2* for Dutch CD trios; these results could not be replicated in an independent cohort. A POO effect in IBD was observed for *IL12B* and *PRDM1*. In the Indian trios the *IL10* locus showed a nominal POO effect. With a combination of Support vector learning machines and bioinformatic analysis we identified a compelling candidate (*NCAM1*) for future follow-up functional studies. This will improve our knowledge of this complex disease and hopefully provide future strategies of disease prevention and treatment. Collectively, our findings begin to shed light on these questions and provide a rich source of clues to the pathogenic mechanisms underlying this archetypal complex disease. Some of the results presented here will need replication in functional follow-up studies. Most of the future efforts will be focused on the interaction between the host mucosal immune system and microbes. In particular, they raise the question, in the context of this burden of IBD-susceptibility genes, of what triggers components of the commensal microbiota to switch from a symbiotic to a pathogenic relationship with the host.

Key words: inflammatory bowel disease, targeted genome-wide association analysis, ImmunoChip

UDK: 616.34-002-056.7:601.4.008.7(043.3)

1 UVOD

Človeške dedne bolezni so vse večji problem sodobne družbe. Zato ne preseneča, da ima preučevanje genskih nepravilnosti osrednjo vlogo v sodobni medicini. Skupino dednih bolezni, ki nastanejo kot posledica nepravilnih sprememb večjega števila genov, imenujemo večgenske (poligenske) bolezni. Večina jih je tudi večfaktorskih, kar pomeni, da so bolezenski procesi posledica spleta genskih in drugih dejavnikov, npr. dejavnikov iz okolja. Med omenjene bolezni prištevamo npr. bolezni osrednjega živčevja, rak in imunsko posredovana stanja, tj. kronične vnetne in avtoimunske bolezni. Za obe skupini je značilen iztirjen imunski odziv na telesu lastne celice in tkiva. Imunski odziv je lahko omejen le na določeno vrsto celic, kot so npr. beta celice Langerhansovih otočkov pri diabetesu tipa 1 ali oligodendrociti pri multipli sklerozi. Lahko pa je usmerjen na širši spekter celic in tkiv, kot so npr. jedrni antigeni pri sistemskem eritematoznem lupusu ali bakterijska mikroflora in črevesne epitelijske celice pri kronični vnetni črevesni bolezni (KVČB).

V preteklih desetletjih je bil v raziskovanje in razumevanje patogeneze KVČB vložen precejšen napor, vendar natančni vzroki nastanka bolezni še niso povsem pojasnjeni. Po za zdaj najbolj uveljavljeni hipotezi se bolezen pojavi zaradi pretiranega imunskega odziva na normalno črevesno mikrofloro pri gensko dovzetnih posameznikih.[1] Pretiran imunski odziv naj bi bil posledica povečane prepustnosti črevesne bariere, ki sicer preprečuje vdor vsebine črevesnega lumna (večinoma bakterij) v globlje plasti črevesne stene. Zaradi povečane prepustnosti črevesne bariere lahko bakterijski antigeni pridejo v intenziven stik s celicami sluzničnega imunskega sistema, kar sproži silovit imunski odziv, ki posledično pripelje do vnetja in poškodb črevesne sluznice.[2]

Odkrivanje in določanje genskih dejavnikov KVČB sta temeljili na številnih raziskovalnih pristopih. V uvodnem poglavju so podane osnove za razumevanje širšega konteksta področja epidemioloških in genskih raziskav te bolezni. V drugem poglavju so opisane metode in postopki, ki smo jih uporabili za odkrivanje genov in variant, povezanih s KVČB. V poglavju z rezultati so strnjene ugotovitve in nove domnevne povezave polimorfizmov in genov s KVČB, do katerih smo prišli z asociacijsko študijo, meritvami genskega izražanja in različnimi bioinformatskimi pristopi. Interpretacije rezultatov in uporabljenih pristopov so predstavljene v četrtem poglavju, njihov vpliv na razumevanje zapletene patogeneze bolezni in prihodnji izzivi v genetiki KVČB pa so orisani v petem poglavju.

1.1 Bolezenski fenotipi in zdravljenje KVČB

Bolniki s KVČB zbolijo za eno od dveh precej podobnih oblik kroničnega vnetja črevesja: za ulceroznim kolitisom (UK) ali Crohnovo boleznijo (CB). V 10–15 odstotkih primerov pa bolniki ne kažejo jasnih kliničnih, endoskopskih in histoloških značilnosti CB ali UK.[3] Takrat je postavitev dokončne diagnoze težavna in se praviloma diagnosticira nedoločena oblika KVČB, t. i. intermediarni kolitis (IK). Pri nekaterih bolnikih z IK je natančnejša diagnoza mogoča šele ob poznejšem napredovanju bolezni, bodisi do CB ali do UK.[3]

1.1.1 Crohnova bolezen

Čeprav se CB lahko pojavi pri kateri koli starosti, sta značilna dva vrha obolevnosti. Najpogosteje se pojavlja med 15. in 30. letom življenja, drugi, manjši, vrh obolevnosti pa se giblje med 50. in 70. letom.[4] Pri približno 25 odstotkih bolnikov bolezen nastopi v otroštvu in adolescenci.[5] Tak zgodnji nastop je običajno povezan s hudimi in hitro napredujočimi oblikami bolezni.[6]

Vnetje lahko zajame katerikoli del prebavnega trakta (od ustne votline do zadnjične odprtine), čeprav najpogosteje prizadene terminalni ileum. Pri tem je značilno menjavanje vnetih in nevnetih, zdravih delov črevesne sluznice.[7] Sprva je vnetni proces omejen na sluznico, kjer se kmalu pojavijo majhne razjede, ki jih običajno spremlja tudi infiltracija nevtrofilcev in makrofagov. Pozneje, ko se vnetje razširi in prodre v vse sloje črevesne stene, pa ga občasno spremljajo abscesi (ognojki), strikture in fistule od črevesja do kože, mehurja ali drugih predelov črevesja. Razpon simptomov je odvisen od oblike bolezni in je temu primerno širok: od bolečin v trebuhu in izgube telesne teže do povišane temperature, krvi v blatu, malabsorbcije in motenj v razvoju.[7]

Odziv bolnikov na zdravljenje je raznolik, tako kot pri večini bolezni, ki zahtevajo dolgotrajno zdravljenje. Najbolj uveljavljen pristop pri zdravljenju KVČB je t. i. okrepitveni (angl. step-up) pristop, pri katerem se glede na (ne)odzivnost in (ali) slabšanje stanja postopoma prehaja na zdravljenje z učinkovitejšimi zdravili (npr.

imunosupresivi, biološkimi zdravili).[8] Prvi korak pri zdravljenju blagih in zmernih oblik CB so praviloma aminosaliciati (npr. sulfasalazin in mesalamin) ter antibiotiki (npr. metronidazol in ciprofloksacin).[9] Kortikosteroidi (npr. prednison in budesonid) se uporabljajo pri zdravljenju hudih oblik CB in se zaradi stranskih učinkov dajejo občasno, zlasti ob nenadnih zagonih bolezni. Kadar kortikosteroidi ne zadostujejo, se preide na zdravljenje z imunosupresivi, kot so azatioprin, 6-merkaptopurin in metotreksat.[9] V zadnjih letih se hude oblike CB, pri katerih ne zadostujejo niti kortikosteroidi niti imunosupresivi, zdravijo s kombinacijo imunosupresivov in zaviralcev dejavnika tumorske nekroze alfa (TNF-α), kot sta infliksimab in adalimumab.[7] Kirurški poseg se praviloma izvede, ko bolezen napreduje od vnetne oblike do obsežnih kompliciranih oblik s fistulami in z abscesi.[10] Novejše klinične raziskave pa kažejo, da je v določenih primerih ustreznejše zdravljenje z učinkovitimi zdravili v zgodnjih fazah poteka bolezni (angl. top-down).[9] Med omenjene primere uvrščamo zgodnji nastop bolezni, agresivne oblike CB z obsežnimi anatomskimi poškodbami črevesja in CB s hudimi zunajčrevesnimi zapleti.[9]

1.1.2 Ulcerozni kolitis

UK se praviloma pojavi med 20. in 30. letom življenja, lahko pa prizadene tudi precej mlajše ali starejše posameznike.[4] Vnetni proces je omejen na kolon, pri bolnikih s hudimi oblikami pankolitisa pa včasih zajame tudi terminalni ileum.[7] Vnetje poteka neprekinjeno, prizadeti sta sluznica in podsluznica; pri blagih in zmernih oblikah bolezni se na sluznici kažejo pordečitve in manjše krvavitve, medtem ko so za hude oblike značilni obsežna ulceracija (razjedenost) in psevdopolipi.[11] Pri večini bolnikov se simptomi kažejo kot bolečine v spodnjem delu trebušne votline in kri v blatu z velikimi količinami izločene sluzi.[7] Dolgotrajna izpostavljenost vnetju lahko izzove maligno displazijo sluznice, pri približno desetih odstotkih bolnikov s pankolitisom pa se razvije karcinom debelega črevesa.[12]

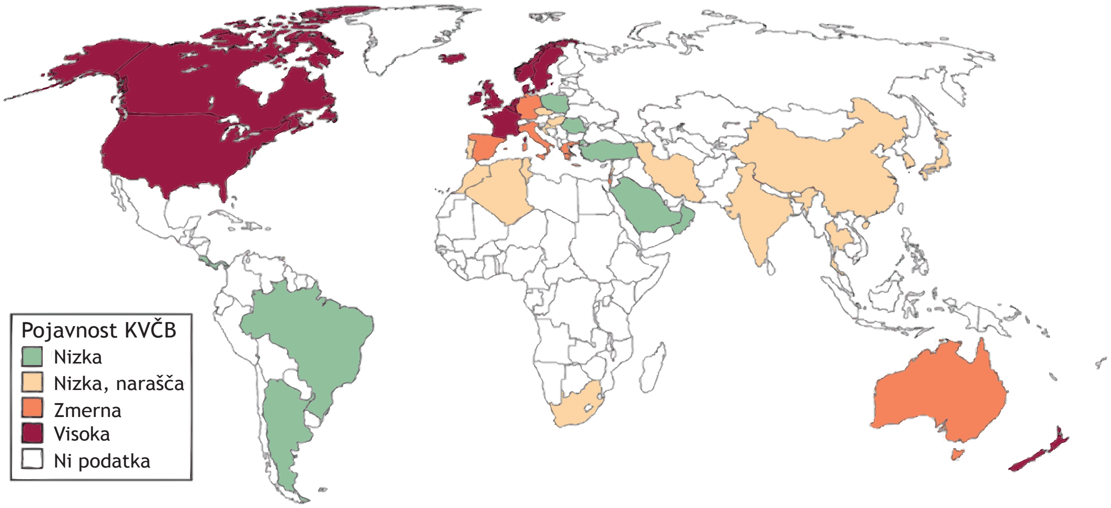
V okviru standardne terapije (angl. step-up) se za sprožanje in vzdrževanje remisije (mirovanja bolezni oz. popuščanja bolezenskih znakov) blagih in zmernih oblik UK najpogosteje uporabljata mesalazin in sulfazalin, če je treba, pa tudi kortikosteroidi.[9] Imunosupresivi se uporabljajo pri zdravljenju akutnih oblik UK in ob ponovnih zagonih bolezni.[7, 9, 12] Kirurško zdravljenje se izvaja pri hitro napredujočih oblikah bolezni, neodzivnosti na standardno zdravljenje in pri pojavu displazije.[13,14] Pri tem gre običajno za poseg v dveh korakih, začenši s proktokolektomijo z ileostomijo. V drugem koraku pa se iz zadnjega dela tankega črevesa oblikuje vrečka, katere odprtina se prišije na anus.[13,14] Ker je UK praviloma omejen le na kolon, lahko popolno ozdravitev dosežemo s proktokolektomijo.[13,14]

1.2 Epidemiološke raziskave in prvi znaki genske pogojenosti KVČB

Pogostost (prevalenca) in pojavnost (incidenca) KVČB se med populacijami na različnih zemljepisnih območjih zelo razlikujeta. Največjo pogostost so ugotovili v Kanadi in Severni Evropi, najmanjšo pa v deželah v razvoju ter v Aziji.[15] UK je s povprečno pogostostjo 235 primerov na 100.000 prebivalcev v zahodnih državah in 20 na 100.000 v državah v razvoju pogostejši kot CB.[15] V Severni Ameriki se pogostost CB giblje okoli 200 na 100.000, v deželah v razvoju pa okoli 10 na 100.000.[15]

Tudi pojavnost KVČB sledi podobnim trendom kot pogostost, saj so najvišje letne vrednosti izmerili v zahodnih državah, najnižje pa v Aziji.[15] Pojavnost narašča v veliki večini držav po svetu, kar nakazuje na zaskrbljujoče dejstvo, da KVČB postaja globalni problem (slika 1-1).[16] V zahodnih državah se je v zadnjih letih stopnja letne pojavnosti ustalila, stalno in strmo naraščanje pa beležijo v Južni Evropi, deželah v razvoju ter v Aziji.[16]

Za slovensko populacijo sicer ni konkretnih epidemioloških podatkov, vendar so Ferkolj in sod. ugotovili, da je bilo ob koncu 90. let v Sloveniji 1150 bolnikov s KVČB.[17] Slednje nas, s približno prevalenco 50 na 100.000, uvršča na dno evropske lestvice.[16]



Slika 1-1: Pregled pojavnosti KVČB po svetu. Države so razvrščene v skupine glede na stopnjo letne pojavnosti: rdeča 10 na 100.000, oranžna 5–10 na 100.000, zelena > 4 na 100.000, rumena – nizka pojavnost, ki stalno narašča. Za države, obarvane belo, ni bilo mogoče pridobiti ustreznih podatkov. (Prirejeno in objavljeno z dovoljenjem Jacquesa Cosnesa.)[18]

Raziskave pogostosti in pojavnosti torej pričajo o naraščajočem številu obolelih za KVČB po vsem svetu, zlasti v industrializiranih, t. i. visokorazvitih družbah. Še presenetljiveje je, da se pogostost bolezni določenih etničnih skupin po migraciji v urbanizirano in razvitejše okolje izenači z lokalno pogostostjo bolezni v eni do dveh generacijah.[15] Pri tem je v prvi generaciji potomcev migrantov tveganje za KVČB povečano.[19] Omenjeni pojavi torej jasno kažejo, da določeni dejavniki iz »zahodnjaškega« načina življenja in okolja (npr. zaradi industrializacije in prehrane) povečujejo tveganje za KVČB.

Med najbolje raziskane dejavnike iz okolja spadata kajenje in apendektomija (kirurška odstranitev slepiča).[20] Vpliv kajenja je paradoksalen: opustitev kajenja je povezana z razvojem UK, medtem ko kajenje vpliva na nastanek in potek CB.[21] Na drugi strani pa apendektomija deluje zaščitno pri UK, pri CB pa zaradi nasprotujočih si rezultatov slika ni povsem jasna.[22,23] Tudi vloga drugih dejavnikov okolja, kot so npr. okužba s *Helicobacter pylori*, velikost družine (v večjih družinah je pogostost bolezni manjša), odraščanje v urbanem okolju, prehrana, dojenje in uporaba antibiotikov, ostaja zaradi nasprotujočih si rezultatov in neskladno zasnovanih raziskav za zdaj nejasna (preglednica 1-1).[20]

Preglednica 1-1: Vpliv dejavnikov iz okolja na tveganje za KVČB

Dejavnik tveganja	Ulcerozni kolitis	Crohnova bolezen
Kajenje		
Kadilec	–	+
Nekdanji kadilec	+	+
Nekadilec	+	–
Pasivni kadilec (prenatalni)	0?	0?
Pasivni kadilec (v otroštvu)	0	0
Apendektomija	–	+?
Prehrana		
Sladkorji	+?	+?
Maščobe	+?	+?
Sadje in zelenjava	–?	–?
Vlaknine	0	–?
Dojenje	–	–
Okužba/cepljenje		
Mycobacterium avium paratuberculosis	0	+?
Helicobacter pylori	–?	–?
Ošpice	0?	0?
Cepljenje proti ošpicam	0?	0?
Adherentna invazivna *Escherichia Coli*	0?	+?
Psihrotrofne bakterije	0?	+?
Perinatalne okužbe	+?	+?
Antibiotiki	+?	+?
Nesteroidna protivnetna zdravila (NSAID)	+?	+?
Velikost družine	–?	–?
Urbano okolje	+?	+?
Kontracepcijska sredstva	+	+

V preglednici 1-1 so povzeti učinki večine doslej preiskovanih dejavnikov iz okolja, ki naj bi vplivali na tveganje za KVČB. S + so označeni dejavniki, ki večajo tveganje za KVČB, zaščitni (ti zmanjšujejo tveganje) so označeni z –, pri tistih, ki so označeni z 0, pa ni znanih vplivov na tveganje. ? označuje, da vpliv določenega dejavnika zaradi nasprotujočih si rezultatov še ni potrjen.

Že dolgo je znano, da med pomembne dejavnike, ki povečujejo tveganje za KVČB, spada tudi pozitivna družinska anamneza. Začetno zanimanje za molekularnogenetske raziskave bolezni je temeljilo ravno na rezultatih epidemioloških raziskav družin in dvojčkov, obolelih za KVČB. Iz teh namreč nedvomno izhaja, da ob dejavnikih iz okolja k tveganju za KVČB prispevajo tudi genski dejavniki.

Raziskave družin so pokazale, da je obolevnost za KVČB večja pri sorodnikih obolelih kot pa pri sorodnikih zdravih posameznikov. Halme in sod. so ugotovili, da je bilo 2–14 odstotkov bolnikov s CB tudi s pozitivno družinsko anamnezo CB, v 5–16 odstotkih primerov pa UK.[24] Delež bolnikov z UK, pri katerih je ta bolezen tudi v družini, je 7–11 odstotkov, v 8–14 odstotkih primerov pa so družinski člani oboleli za CB.[24]

Družinsko kumulacijo bolezni lahko kvantitativno izrazimo kot razmerje med stopnjo tveganja bratov ali sester in pogostostjo bolezni v populaciji, tj. kot relativno tveganje bratov ali sester (λ_s). Vrednosti λ_s so za KVČB (preglednica 1-2), zlasti za CB, podobne ali večje kot za druge kompleksne bolezni, kot so npr. diabetes tipa 1 ($\lambda_s = 15$), diabetes tipa 2 ($\lambda_s < 10$), shizofrenija ($\lambda_s < 10$) in celiakija ($\lambda_s = 7–30$).[24]

Preglednica 1-2: Epidemiološki podatki za KVČB, CB in UK

Epidemiološki parameter	Vrednost	Referenca
Pogostost KVČB	8–294/100.000 prebivalcev*	18
Pojavnost KVČB	4–10/100.000 prebivalcev/leto*	15
Konkordanca enojajčnih dvojčkov – CB	37–58 %	25
Konkordanca dvojajčnih dvojčkov – CB	4–12 %	25
Konkordanca enojajčnih dvojčkov – UK	6–17 %	25
Konkordanca dvojajčnih dvojčkov – UK	0–5 %	25
Relativno tveganje bratov ali sester (λ_s) – CB	25–42	26
Relativno tveganje bratov in sester (λ_s) – UK	8–15	26

* Povprečna vrednost v Evropski uniji

Čeprav maloštevilne, so se raziskave dvojčkov izkazale kot pomembno orodje pri razlikovanju prispevanja okoljskih in genskih dejavnikov k tveganju za razvoj bolezni, kar najlažje ponazorimo z naslednjima skrajnima primeroma:

i) nastanek in razvoj bolezni pogojujejo le genski dejavniki; v tem primeru bi bilo fenotipsko ujemanje oz. konkordanca enojajčnih (identičnih) dvojčkov blizu 100 odstotkov, dvojajčnih dvojčkov pa 50 odstotkov;

ii) nastanek in razvoj bolezni pogojujejo le zunanji dejavniki; v tem primeru pa bi bili konkordanci obeh tipov dvojčkov podobni.

Pri CB se konkordanca enojajčnih dvojčkov giblje med 37 in 58 odstotki, dvojajčnih pa med 4 in 12 odstotki. Pri UK je konkordanca nekoliko manjša in pri enojajčnih dvojčkih znaša 6–17 odstotkov, pri dvojajčnih dvojčkih pa 0–5 odstotkov, kar kaže na nekoliko večji vpliv oz. penetranco genskih dejavnikov pri CB kot pri UK.[27] Konkordanci enojajčnih dvojčkov v nobenem od podtipov KVČB ne presegata 50 odstotkov, kar nakazuje na to, da vloga dejavnikov iz okolja v patogenezi KVČB ni zanemarljiva.

Spoznanja epidemioloških raziskav, ki so jih izvajali v preteklih desetletjih, so nedvomno prispevala k boljšemu poznavanju dejavnikov, ki privedejo do nastanka KVČB. Kljub številnim neovrgljivim dokazom o genskem ozadju bolezni pa se je natančno določanje genskih determinant začelo šele v zadnjih petnajstih letih.

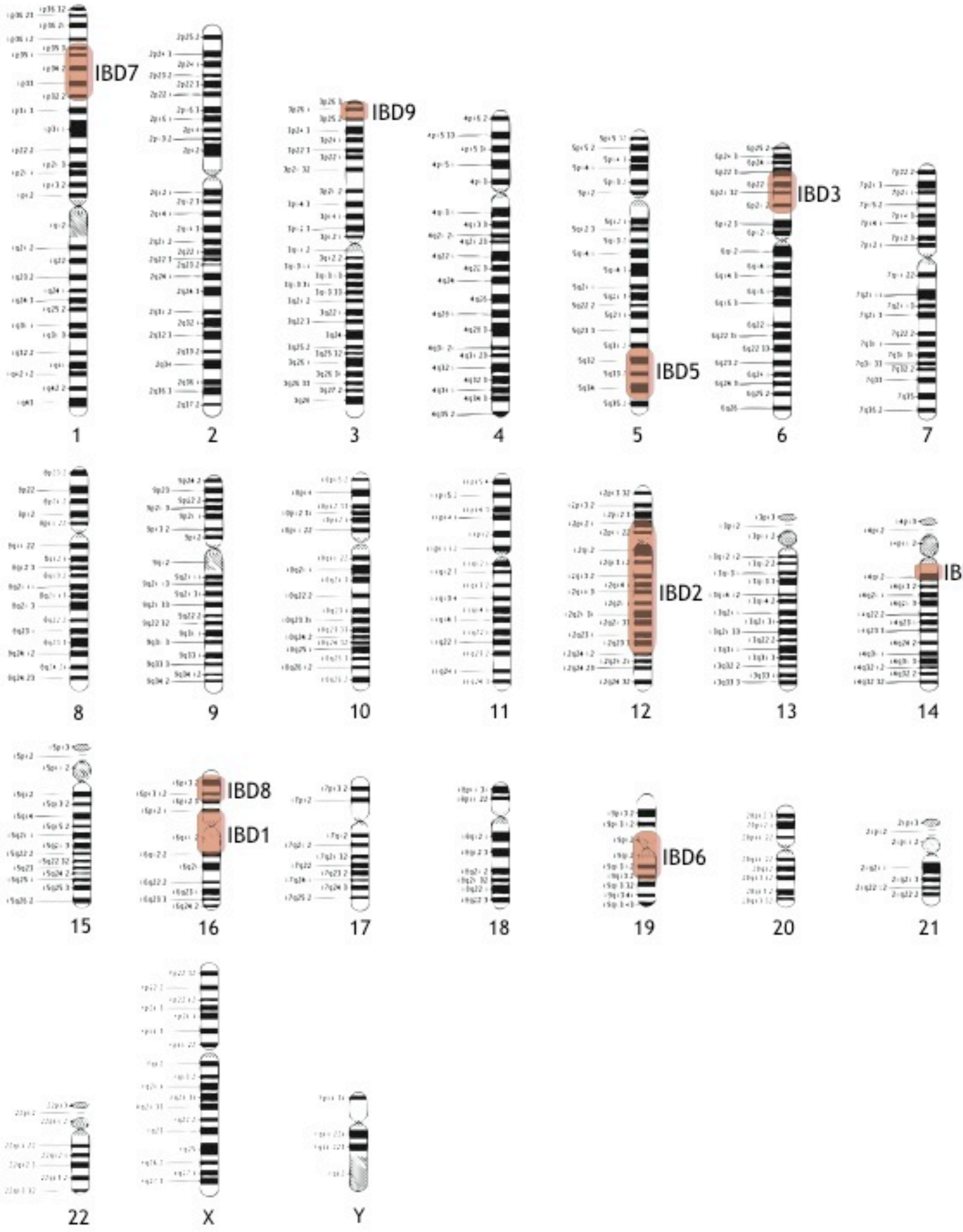
1.3 Študijski pristopi pri odkrivanju genskih vzrokov KVČB

Odkrivanje in določanje genskih dejavnikov KVČB je temeljilo na številih raziskovalnih pristopih. Med temi so k razumevanju genskega ozadja bolezni največ prispevale asociacijske študije na celotnem genomu (angl. genome-wide association studies – GWAS). Zaradi njihovega pomena v raziskavah kompleksnih bolezni in učinka »snežne kepe« ločimo v kronološkem smislu različna raziskovalna obdobja.

1.3.1 Raziskave in izsledki v obdobju pred pojavom GWAS

Večja obolevnost posameznikov iz družin s KVČB je privedla do prepričanja, da bi lahko vzročne gene z večjim vplivom na tveganje za KVČB odkrili z analizo genske povezanosti (angl. linkage analysis).[28] Prav tako se je analiza genske povezanosti izkazala kot zelo uspešen pristop pri odkrivanju bolezenskih oz. vzročnih genov monogenskih bolezni.[29] Zato so v devetdesetih letih 20. stoletja začeli iskati vzročne gene kompleksnih bolezni s študijami genske povezanosti na celotnem genomu (angl. genome–wide linkage studies – GWLs).

GWLs je metoda, s katero preiskujemo genom in lokaliziramo gene glede na hkratno dedovanje genskih označevalcev in bolezni v družinah z več obolelimi člani. Temeljni princip metode je, da si bodo družinski člani s KVČB delili skupne gene pogosteje kot po naključju, kar se pokaže statistično z logaritmom obetov (angl. logarithm of odds – LOD). Za statistično značilno povezana področja je dogovorjen LOD ≥ 2,2.[30] V enajstih neparametričnih GWLs so za KVČB odkrili devet kromosomskih področij (*IBD1–IBD9*) z LOD ≥ 2,2, ki so si jih oboleli družinski člani delili pogosteje kot po naključju (slika 1-2).[31]



Slika 1-2: Kromosomska področja *IBD1–IBD9*. S KVČB so jih povezali v devetdesetih letih 20. stoletja v študijah genske povezanosti na celotnem genomu (GWLs). Na vsakem od področij je od nekaj deset do več sto potencialnih kandidatnih genov. Ta kromosomska področja se raztezajo čez več deset milijonov baznih parov (bp). Posledično je na posameznem področju lahko tudi več sto potencialnih kandidatnih genov. Zato je primarnim odkritjem sledilo t. i. pozicijsko kloniranje. Pri tem gre za iskanje vzročnega gena s postopnim oženjem širših kromosomskih področij s kombinacijo različnih kliničnih, laboratorijskih in statističnih analiz.

S tem postopkom so leta 2001 tri skupine raziskovalcev neodvisno odkrile povezavo med CB in variantami v genu *NOD2* na kromosomu 16q12 (*IBD1*).[32-34] *NOD2* kodira znotrajcelični receptor, ki ga aktivira minimalno biološko aktivna komponenta peptidoglikana – muramildipeptid (MDP). Ta je v celični steni tako po Gramu pozitivnih kot po Gramu negativnih bakterij. NOD2 je v največji meri izražen v monocitih in Panethovih celicah črevesnega epitelija.[35,36]

Omenjene variante gena *NOD2* so polimorfizmi posameznega nukleotida (angl. single nucleotide polymorphism – SNP), in sicer: dve drugačnosmiselni mutaciji R702W (Arg702Trp) in G908R (Gly908Arg) ter premik bralnega okvirja L1007fs (Leu1007fs). Ti SNP-ji so na C–terminalnem koncu levcinsko bogatih ponovitev (angl. leucine-rich repeats – LRR), ki je domena za zaznavanje bakterijskih produktov.[37] Zato se zdi, da ti SNP-ji zmanjšujejo odzivnost proteina pri zaznavanju bakterijskih produktov, kar posledično vpliva na prirojeni imunski odziv.[37]

Čeprav vloga NOD2 v imunopatogenezi CB še ni povsem pojasnjena, je znano, da je pri heterozigotnih nosilcih katerega koli od treh vzročnih alelov omenjenih SNP-jev 2–4-krat večje tveganje za CB, medtem ko je tveganje pri homozigotih in sestavljenih heterozigotih 17-krat večje.[38, 39]

Lesage in sod. so ugotovili, da vzročni aleli omenjenih SNP-jev zastopajo več kot 80 odstotkov celotnega prispevanja gena *NOD2* k tveganju za CB.[40] Zato je toliko presenetljiveje, da so ti vzročni aleli v znatni meri prisotni tudi pri zdravih Kavkazcih.[41] Iz tega izhaja, da omenjene mutacije niso niti zadostne niti nujne za razvoj CB, kar pa nakazuje na poligensko in kompleksno ozadje bolezni.

Na drugih kromosomskih področjih (*IBD2–IBD9*) pa se iskanje vzročnih genov s kombinacijo GWLs in pozicijskega kloniranja ni obneslo tako kot v primeru *NOD2*.[42] Zato je bil glavni izsledek GWLs, da KVČB ne povzroča malo mutacij z visoko penetranco, temveč več vzročnih genov (oz. alelov) s šibkim vplivom na tveganje za KVČB.[30] Vzročne alele gena *NOD2* so namreč uspešno zaznali predvsem zaradi njihove pozicije na splošnem haplotipskem bloku SNP5.[43] V primerjavi s splošnimi mutacijami bi za odkrivanje redkih mutacij (tj. z majhno penetranco in s šibkim vplivom) na drugih kromosomskih področjih potrebovali zelo veliko število obolelih družin.[30, 44]

Slednje je raziskovalce spodbudilo k preizkušanju novih pristopov in paradigem pri iskanju vzročnih genov in posledično privedlo k izvajanju asociacijskih študij kandidatnih genov (angl. candidate gene association study – CGAS).[28]

V CGAS gre za iskanje povezave med določenim genom s primerjavo alelnih, genotipskih ali haplotipskih frekvenc med bolniki in zdravimi posamezniki iz splošne populacije, v t. i. študijah primerov in kontrol.[45] Pri tem gre za statistični prikaz sopojavnosti določenega alela in bolezni. Povedano drugače, nek alel bo povezan z boleznijo (asociacija), če bo statistično značilno bolj (ali manj) zastopan med bolniki kot med zdravimi posamezniki.

Kmalu po prvem valu CGAS se je izkazalo, da so za tovrstni pristop značilne številne pomanjkljivosti.[42] Glavna težava je bila neponovljivost oz. nekonsistentnost rezultatov. Novo odkritih asociacij neke raziskovalne skupine ni uspelo replicirati v drugih raziskovalnih skupinah ali pa so celo prišli do nasprotujočih zaključkov.[46-56] Metaanaliza več kot tristotih objav o 25 najpogosteje preučevanih asociacijah pri enajstih različnih boleznih, ki so jo izvedli Lohmueller in sod., je pokazala, da je bila ustrezno ponovljiva le ena tretjina asociacij. Razlogi za to izvirajo iz zasnove CGAS, saj so rezultati v največji meri odvisni od izbora preiskovanih polimorfizmov, raznolikosti genskih in okoljskih ozadij preiskovancev, števila bolnikov ter zdravih posameznikov in vključitvenih kriterijev bolnikov in zdravih posameznikov.[28,57,58]

Med najuspešneje replicirane objave CGAS spada asociacija med KVČB in vzročnimi aleli poglavitnega histokompatibilnega kompleksa (angl. major histocompatibility complex – MHC).[59] Satsangi in sod. so odkrili povezavo med hudo obliko UK in redkim alelom DRB1*0103, Silverberg in sod. pa so ta alel povezali tudi s CB, omejenim na kolon.[60,61]

Obe objavi sta bili uspešno replicirani v številnih študijah, ki so jih izvedli na kavkaški populaciji.[62-68]

1.3.2 GWAS in ključne ugotovitve

Napredovanje pri odkrivanju vzročnih genov KVČB je pospešila združitev ugodnih lastnosti dotlej uporabljenih pristopov, tj. »genomske širine« GWLs in statistične moči za iskanje vzročnih alelov s šibkim vplivom CGAS (zasnovanih kot študije primerov in kontrol). K prehodu od GWLs in CGAS do prvih asociacijskih študij na celotnem genomu (angl. genome-wide association studies – GWAS) so prispevali tudi nekatera spoznanja iz molekularne biologije in napredek genske tehnologije:

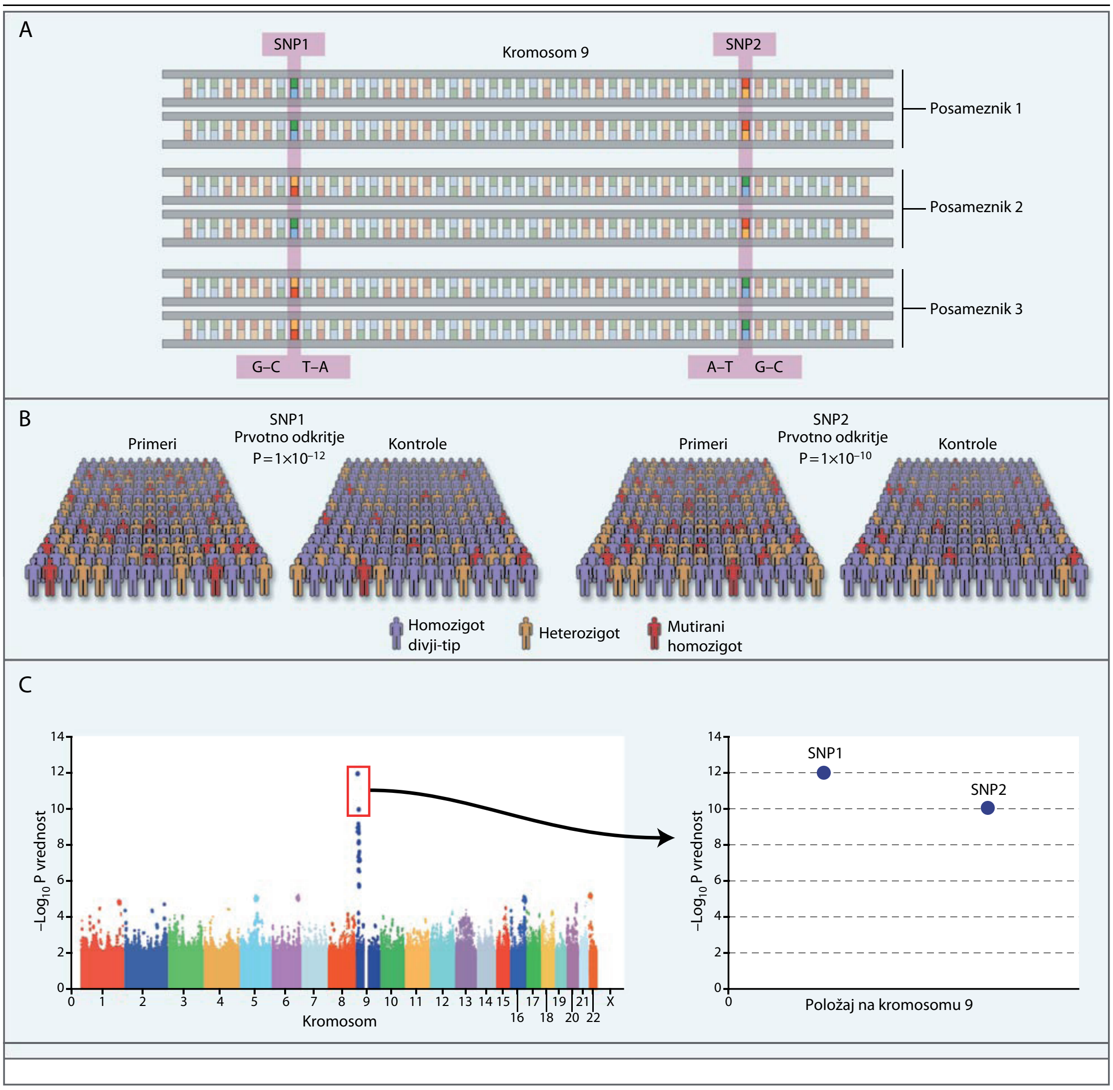
i) Razkritje človeškega genoma s projektoma »Human Genome Project« in »HapMap«.[69-72] Pri tem je bila ključna katalogizacija SNP-jev in razkritje široko razširjene korelacije med njimi, t. i. vezavnega neravnotežja (angl. linkage disequlibrium – LD). SNP-ji, ki so v LD, se dedujejo skupaj (v t. i. haplotipskih blokih) pogosteje kot po naključju. Moč korelacije običajno izrazimo s korelacijskim koeficientom r^2, pri čemer $r^2 = 0$ pomeni, da korelacije ni, $r^2 = 1$ pa popolno ujemanje med SNP-ji. Ko želimo analizirati nek haplotipski blok, zadostuje, da genotipiziramo le enega od SNP-jev, ki je v korelaciji $r^2 > 0,8$ z drugimi SNP-ji.

ii) Hiter razvoj nove tehnologije mikromrež visokih zmogljivosti, s katerimi je mogoče genotipizirati več 100.000 SNP-jev naenkrat.[73]

iii) Razvoj statističnih metod, ki omogočajo posredno pripisovanje genotipov za SNP-je, ki niso na mikromreži (angl. imputation methods).[74]

V običajno zasnovani študiji GWAS testiramo več 100.000 polimorfizmov pri večjem številu bolnikov (primeri) in zdravih posameznikov (kontrole), kot je prikazano na sliki 1-3. Ta zasnova je najpreprostejša in najpogostejša, glede na izbor preiskovancev pa je mogoče GWAS zasnovati tudi kot študijo triov (bolnik in njegovi starši) ali kot kohortno študijo.[75]

Ker v GWAS testiramo več SNP-jev naenkrat, je temu primerno treba prilagoditi prag statistične značilnosti, in sicer tako, da običajno mejo (P = 0,05) delimo s številom testov (tj. Bonferronijeva korekcija). Po dogovoru je prag statistične značilnosti na celotnem genomu postavljen pri P = 5 x 10^{-8}.[76]



Slika 1-3: Asociacijska študija na celotnem genomu (GWAS). GWAS je običajno zasnovan kot študija primerov in kontrol, v kateri genotipiziramo več SNP-jev na celotnem genomu. Na sliki A je kratek segment kromosoma 9, kjer sta SNP-ja SNP1 in SNP2, ki ju želimo genotipizirati za posameznike 1, 2 in 3. Zaradi poenostavitve sta tukaj prikazana dva SNP-ja treh posameznikov, v pravi študiji genotipiziramo več 100.000 polimorfizmov na več tisočih primerov in kontrol. Na sliki B je shematsko prikazana moč asociacije med SNP-jema in boleznijo. Pogostost SNP-jev se med kontrolami in primeri izmeri za vsak

SNP na mikromreži. V tem primeru se statistično značilno razlikuje, kar je izraženo s P-vrednostma $P_{SNP1} = 10^{-12}$ in $P_{SNP2} = 10^{-10}$. Na sliki C je t. i. diagram Manhattan s prikazom P-vrednosti vseh genotipiziranih SNP-jev na celotnem genomu, ki so prestali kontrolo kakovosti. Posamezni kromosomi so zaradi preglednosti obarvani različno. (Prirejeno in objavljeno z dovoljenjem Teri Manolio.[77])

Izvedba študije GWAS običajno poteka v štirih osnovnih korakih:

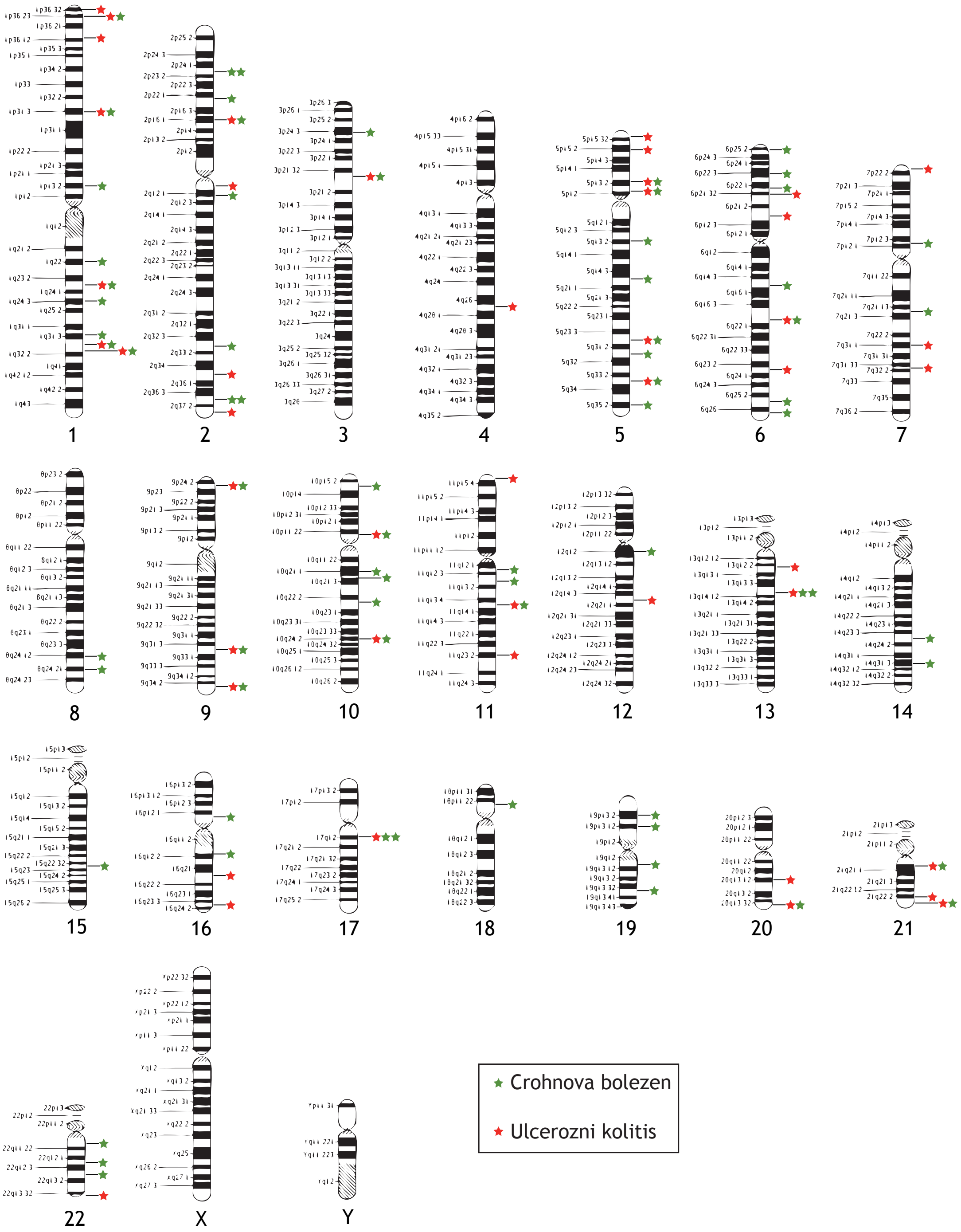
i) izbor večjega števila bolnikov in zdravih posameznikov, pri čemer se morata obe skupini preiskovancev ujemati v spolu in starosti ter drugih kliničnih značilnostih, ki jih želimo testirati;
ii) izolacija DNK, genotipizacija na izbrani mikromreži in kontrola kakovosti;
iii) asociacijska analiza polimorfizmov in preiskovancev, ki so prestali kontrolo kakovosti;
iv) replikacija odkritih asociacij pri drugi neodvisni populaciji in (ali) potrditev s funkcijsko analizo.

Yamazaki in sod. so leta 2005 izvedli prvo študijo GWAS za KVČB. Testirali so 80.000 SNP-jev na japonskih primerih in kontrolah ter odkrili povezavo med CB in variantami v genu *TNFSF15*, ki kodira citokin iz družine dejavnikov tumorske nekroze (TNF).[78] To povezavo so pozneje uspešno potrdili tudi pri evropski populaciji.[79] Kakuta in sod. pa so v funkcijski študiji ugotovili, da je pri nosilcih vzročne variante povečano izražanje TNFSF15 po stimulaciji T-celic.[80]

Iz kataloga objavljenih študij GWAS[81] lahko razberemo, da jih je bilo doslej za KVČB izvedenih dvajset (šest za UK,[82-87] deset za CB[88-97] in štiri za KVČB[98-101]). S povezovanjem posameznih raziskovalnih skupin in centrov v mednarodne konzorcije, kot je npr. International Inflammatory Bowel Disease Genetics Consortium (IIBDGC), in z izvajanjem metaanaliz se je število preiskovancev občutno povečalo, kar je še posebej ugodno pri odkrivanju redkih variant z alelno frekvenco < 1 odstotka.[102]

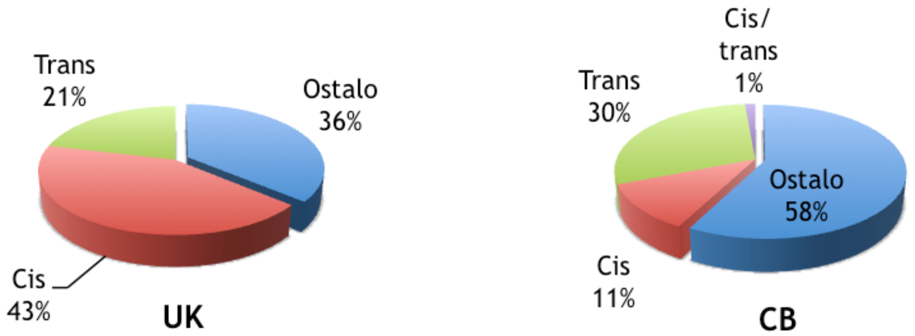
Pred prvima metaanalizama GWAS je bilo znanih 32 lokusov z vplivom na tveganje za CB in 18 za UK.[87,94] Franke in sod. so v nedavni metaanalizi za CB, v katero je bilo v okviru konzorcija IIBDGC vključenih > 20.000 primerov in > 28.000 kontrol, odkrili 39 novih lokusov, tako da za zdaj poznamo skupaj 71 lokusov z vplivom na tveganje za CB.[103]

Anderson in sod. so na začetku leta 2011 v metaanalizi za UK, v katero je bilo v okviru konzorcija IIBDGC vključenih > 15.000 primerov in > 30.000 kontrol, povečali število lokusov z vplivom na tveganje za UK na 47.[104] Obema podtipoma bolezni je skupnih 28 lokusov (slika 1-4).



Slika 1-4: Pregled lokusov KVČB iz zadnjih metaanaliz študij GWAS.[103, 104] Za zdaj je znanih 99 lokusov z vplivom na tveganje za KVČB, od tega jih je 71 je specifičnih za CB, 47 za UK, 28 pa je skupnih obema podtipoma bolezni.

Kljub številčnosti s KVČB povezanih lokusov (in kandidatnih genov) je doslej znanih presenetljivo malo pravih vzročnih variant ali celo vzročnih genov. Večina povezanih SNP-jev je namreč v nekodirajočih, tj. intra- oziroma intergenskih regijah ali pa celo v regijah brez znanih genov (t. i. genskih puščavah): pri CB 60/71, pri UK 39/47.[81] Približno polovica vseh povezanih SNP-jev je cis- ali transekspresijskih kvantitativnih lokusov (angl. expression quantitative trait locus – eQTL), kar pomeni, da alelnospecifično izražanje genov precej prispeva k tveganju za nastanek bolezni[105] (slika 1-5).



Slika 1-5: Porazdelitev SNP-jev iz intergenskih regij, povezanih s KVČB, glede na funkcijo genotipa oz. vpliv na izražanje genov (eQTL)

V lokusih z večjim številom potencialnih kandidatnih genov je bilo s funkcijskimi analizami ugotovljeno, da so ti geni vključeni v ključne biološke poti za vzdrževanje črevesne homeostaze (preglednica 1-3).[38,106]

Preglednica 1-3: Biološke poti, povezane s KVČB in z vzdrževanjem črevesne homeostaze

S KVČB povezani procesi	**CB-geni**	**UK-geni**	**KVČB-geni**
Integriteta črevesne bariere	*MUC19, ITLN1*$^{+\circ}$	*GNA12*$^{+\circ}$*, HNF4A, CDH1, ERRFI1*	/
Obnavljanje epitelija	*STAT3*	*ERRFI1, HNF4A, PLA2G2A/E*	*REL, PTGER4*$^{\circ}$*, NKX2-3*
Ionski prenašalci	*SLC9A4, SLC22A5*$^{\circ}$*, SLC22A4*$^{+}$	*AQP12A/B, SLC9A3, SLC26A3*	/
Panethove celice	*ITLN1*$^{+}$*, NOD2*$^{+}$*, ATG16L1*$^{+}$	/	*XBP1*$^{+}$

Preglednica 1-3: Biološke poti, povezane s KVČB in z vzdrževanjem črevesne homeostaze (nadaljevanje)

S KVČB povezani procesi	**CB-geni**	**UK-geni**	**KVČB-geni**
Prirojeni mukusni imunski odziv	*NOD2*$^{+}$, *ITLN1*$^{+}$	*SLC11A1*, *FCGR2A*$^{+}$/*B*$^{\circ}$	*CARD9*$^{+\circ}$, *REL*
Pridobivanje/priklic novih imunskih celic	*CCL11/CCL2/CCL7/CCL8*, *CCR6*	*IL8RA/IL8RB*$^{\circ}$	*MST1*$^{+\circ}$
Prikaz antigena	*ERAP2*$^{+}$, *LNPEP*, *DENND1B*	/	/
Biološka pot IL23/Th17	*STAT3*	*IL21*	*IL23R*$^{+}$, *JAK2*, *TYK2*$^{+}$, *ICOSLG*, *TNFSF15*$^{+}$
Regulacija T-celic	*NDFIP1*, *TAGAP*, *IL2RA*	*IL2*, *TNFRSF9*, *PIM3*$^{\circ}$, *IL7R*$^{+\circ}$, *TNFSF8*$^{\circ}$, *IFNG*, *IL21*	*TNFSF8*, *IL12B*, *IL23R*$^{+}$, *PRDM1*, *ICOSLG*
Regulacija B-celic	*IL5*, *IKZF1*, *BACH2*	*IL7R*$^{+\circ}$, *IRF5*	/
Imunotoleranca	*IL27*$^{+}$, *SBNO2*, *NOD2*$^{+}$	*IL1R1/IL1R2*$^{\circ}$	*IL10*, *CREM*$^{\circ}$
Celični odzivi			
Avtofagija	*ATG16L1*$^{+}$, *IRGM*, *NOD2*$^{+}$, *LRRK2*	*PARK7*, *DAP*$^{\circ}$	*CUL2*
Stres ER	*CPEB4*$^{\circ}$	*SERINC3*$^{\circ}$	*ORMDL3*$^{\circ}$, *XBP1*$^{+}$
Znotrajcelični transport	*VAMP3*, *FGFR1OP*	*TTLL8*, *CEP72*, *TPPP*	*KIF21B*
Celična migracija	/	*ARPC2*$^{\circ}$, *LSP1*$^{\circ}$, *AAMP*$^{\circ}$	/
Apoptoza	*FASLG*, *THADA*$^{+}$	*DAP*$^{\circ}$	*PUS10*, *MST1*$^{\circ+}$
Metabolizem ogljikovih hidratov	*GCKR*$^{+}$	/	*SLC2A4RG*$^{\circ}$
Oksidativni stres	*PRDX5*, *BACH2*, *ADO*, *GPX4*, *GPX1*$^{+}$, *SLC22A4*, *LRRK2*, *NOD2*$^{+}$	*HSPA6*$^{\circ}$, *DLD*, *PARK7*	*CARD9*$^{\circ+}$, *UTS2*$^{+}$, *PEX13*

Črevesna homeostaza zahteva usklajeno delovanje epitelijskih celic ter celic prirojenega in pridobljenega imunskega odziva. V preglednici 1–3 so geni z visoko korelacijo LD (r^2 > 0,8) s SNP-ji iz GWAS za KVČB. Znak $^{+}$pomeni gene, kjer je z boleznijo povezan SNP v eksonu (kodirajoča varianta), znak $^{\circ}$ pa gene, na katere vplivajo *cis*-eQTL-i.

Ob omenjeni funkcijski porazdelitvi SNP-jev in povezavah KVČB z različnimi biološkimi potmi, ki so del črevesne homeostaze, lahko iz študij GWA in metaanaliz za KVČB strnemo še naslednje ugotovitve:

i) Povezave med komponentami biološke poti IL23/Th17 (IL23R, IL12B, STAT3, JAK2 in TYK2) in KVČB so postavile v ospredje pomen pridobljenega imunskega odziva v patogenezi bolezni in pri vzdrževanju črevesne homeostaze.[107]

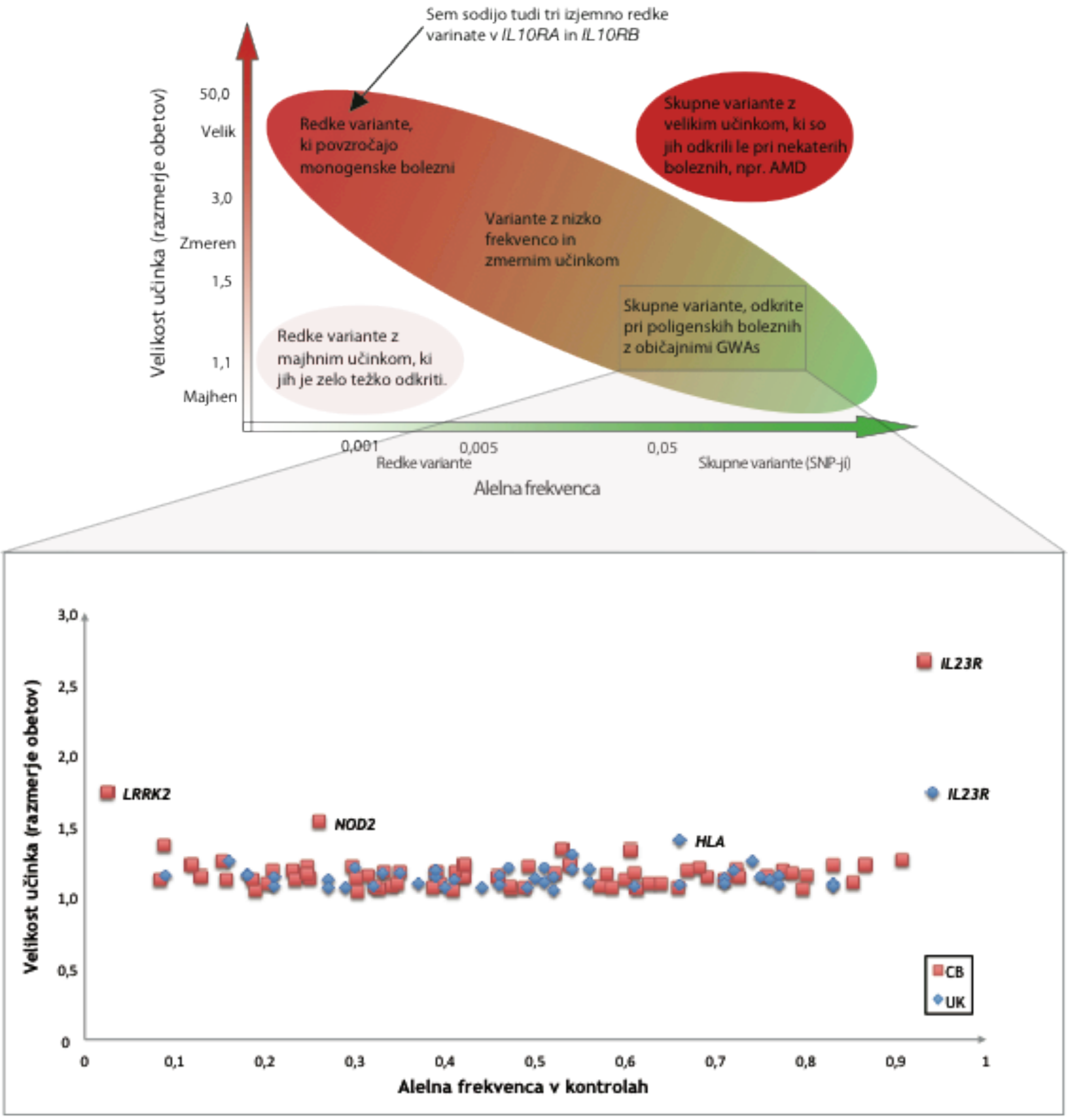
ii) Nedavne raziskave funkcijskih variant v genih *IL10R1* in *IL10R2*, ki se manifestirajo kot primarna imunodeficienca in hude oblike KVČB, kažejo, da ima osrednjo vlogo signalna pot IL10 v patogenezi KVČB.[108]

iii) Povezave med SNP-ji v genih *NOD2*, *ATG16L1* in *IRGM* ter CB so izpostavile prirojeni imunski odziv v patogenezi CB.[90,91,97] Posebej presenetljiva je bila raziskava Riouxa in sod., v kateri so prvič povezali CB in avtofagijo, ki je pred tem sploh niso prištevali med CB-procese.[89] Pri avtofagiji gre za razgradnjo poškodovanih organelov in proteinov ter odstranjevanje patogenih organizmov iz celic. Iniciacija avtofagije je odvisna od NOD2, ki nato prek RIPK2, ATG5 in ATG7 prikliče ATG16L1 do mesta bakterijskega vdora na celični membrani. Zato imajo dendritske celice pri bolnikih s CB, pri katerih sta mutirana *NOD2* in (ali) *ATG16L1i*, okvarjeno avtofagijo, zaznavanje bakterij in prezentacijo antigenov.[37,109] Ob tem je zanimivo, da variante *NOD2* in drugih genov, povezanih z avtofagijo in s prepoznavanjem patogenih bakterij, ne vplivajo na tveganje za UK.

iv) Povezave med SNP-ji v genih *ECM1, CDH1, HNF4A, LAMB1* in *GNA12* ter UK izpostavljajo pomen okvar črevesne epitelijske bariere v patogenezi UK.[84,104,110]

v) Na lokusih KVČB so odkrili več kot deset miRNK (angl. micro RNA − miRNA) in približno 40 velikih interferenčnih nekodirajočih RNK (angl. large intervening non-coding RNA − lincRNA). Nekatere od omenjenih molekul RNK reagirajo s histonskim kompleksom »polycomb« (PRC2), ki ima pomembno vlogo pri utišanju genov, kar najverjetneje pomeni, da imajo tudi miRNK in lincRNA vlogo v patogenezi KVČB.[111]

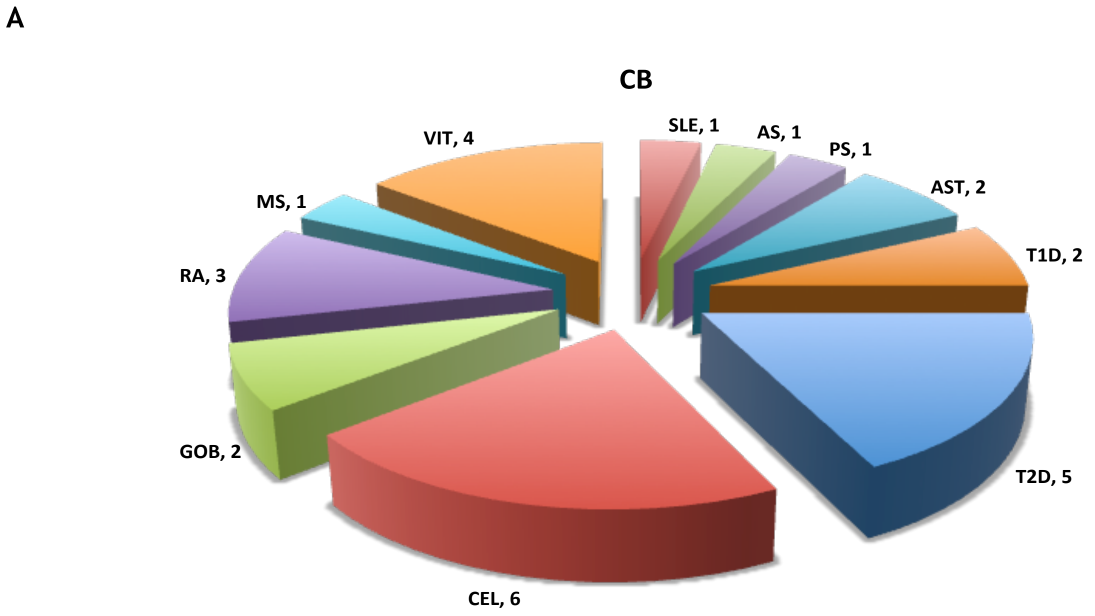
vi) Doslej odkriti SNP-ji so v območju zmernega vpliva na tveganje za KVČB (slika 1-6). Presenetljivo je, da je z njimi mogoče razložiti le približno 23 odstotkov dednostnega deleža (heritabilitete) za CB in 16 odstotkov za UK.[103,104] V raziskavi Parka in sod. so avtorji z ekstrapolacijo doslej zbranih rezultatov napovedali, da bi bilo tudi z metanalizo s 125.000 preiskovanci (tj. z desetkrat večjo, kot je trenutna) mogoče odkriti največ 142 neodvisnih lokusov, ki bi razložili »le« dodatnih 20 odstotkov dednostnega deleža za CB.[112] Danes bi se morda zdelo naivno, vendar GWAS v osnovi temelji na hipotezi splošna bolezen – splošna (skupna) varianta (angl. common disease – common variant).[113] Zadnjo zavrača prej omenjeni problem manjkajočega dednostnega deleža, saj se je pri kompleksnih boleznih brez izjeme izkazalo, da je mogoče s skupnimi oz. splošnimi variantami razložiti le manjši dednostni delež.[114] Problem manjkajočega dednostnega deleža je izzval burne razprave, zato zdaj predvidevajo, da lahko manjkajoči genski prispevek k tveganju za nastanek bolezni pripišemo: a) večjemu številu redkih variant z velikim vplivom na tveganje (model redkih alelov),[115] ki so slabo zastopane na trenutnih platformah GWAS, b) večjemu številu preostalih skupnih variant, ki pokrivajo celoten spekter alelnih frekvenc in je značilen majhen vpliv na tveganje za bolezen (infinitezimalni model)[116] ali kombinaciji genotipskih, epistatskih in epigenetskih interakcij (splošni model dednosti).[117,118] Na drugi strani je lahko problem manjkajočega dednostnega deleža tudi posledica neustrezne metodologije in precenjenih ocen dednostnega deleža pri tveganju za nastanek bolezni.[119]

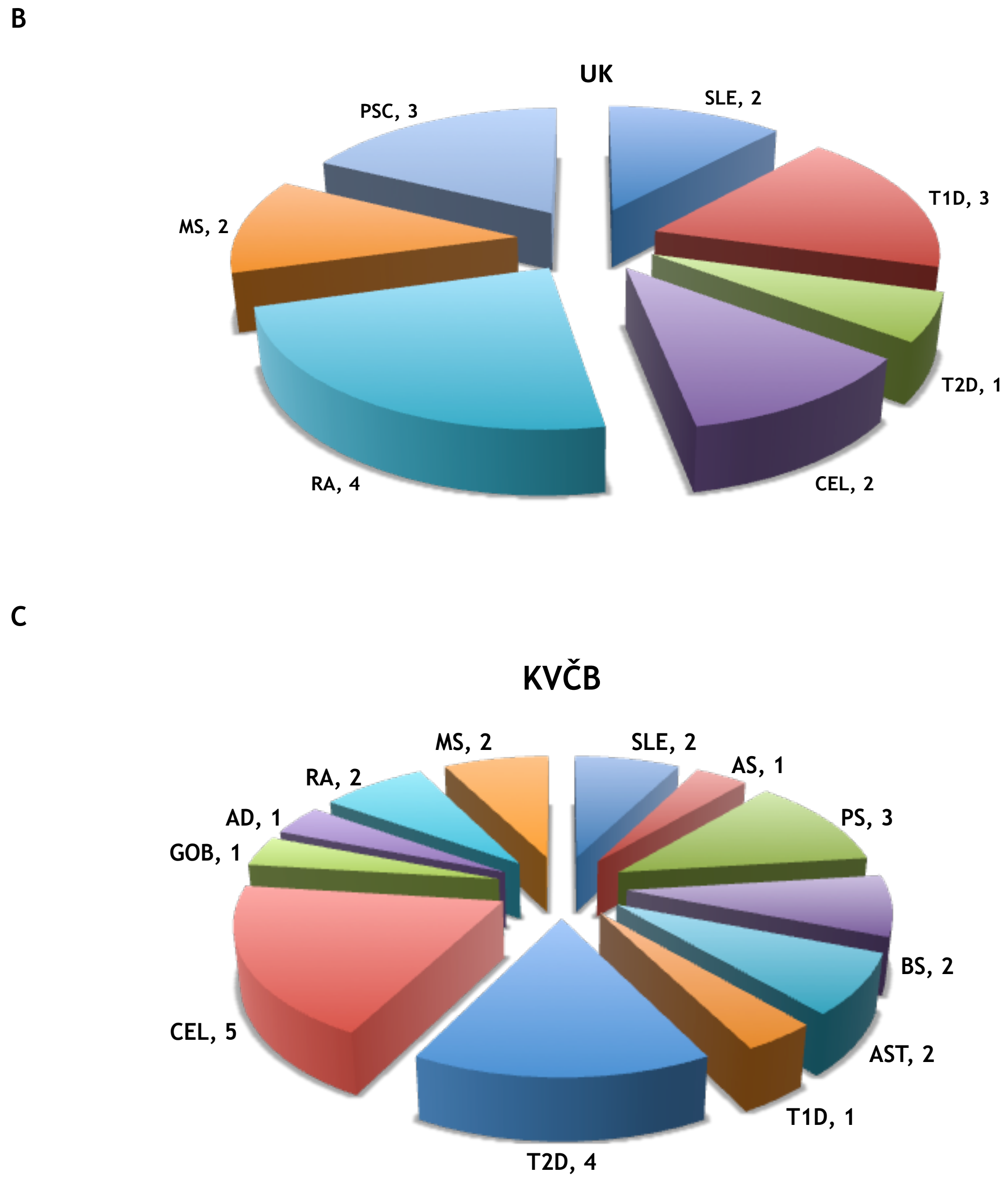


Slika 1-6: Lastnosti variant, povezanih z boleznijo, in njihov vpliv na izvedljivost GWAS

Večina variant, odkritih z GWAS, je skupnih oz. je njihov najmanj pogost alel v populaciji zastopan s frekvenco enega odstotka ali več. Ob tem je za te variante značilen šibek oz. zmeren vpliv na tveganje za KVČB (razmerje obetov < 1,5). Le v redkih primerih take variante zelo vplivajo na tveganje za razvoj bolezni. Slednje so Glocker in sod. odkrili v genih *IL10RA* in *B* ter jih povezali s hudimi oblikami CB pri otrocih iz konsangvinih družin.[108] Za skupne variante, povezane s kompleksnimi boleznimi, je velik vpliv na tveganje za nastanek bolezni prej izjema kot pravilo. Tako izjemo predstavlja npr. povezava med starostno degeneracijo očesne mrežnice (angl. age-related macular degeneration – AMD) in nekaterimi aleli v genu *CFH*, ki kodira dejavnik H komplementa.[120]

vii) Na podlagi za zdaj dostopne literature lahko ugotovimo, da se približno polovica lokusov KVČB prekriva tudi z lokusi 23 različnih bolezenskih stanj, kot so gastrointestinalna stanja (celiakija, kolorektalni rak, primarni sklerozni kolangitis), avtoimunske bolezni (multipla skleroza, revmatoidni artritis, psoriaza, atopični dermatitis, sistemski lupus eritematozus, diabetes T1 in vitiligo), astma, mikobakterijske okužbe (gobavost in verjetno tuberkuloza) ter neimunske bolezni (diabetes T2, dislipidemija, debelost in osteoporoza) (slika 1-7).[87,104,121-123] Zato ne preseneča dejstvo, da so geni, ki so skupni največjemu številu bolezni (*IL2RA, PTPN22* (petim), *FCGR2A* (štirim), *IRF5, IL10, IL23R, IL2*/*IL21* in *ORMDL3* (trem)), ključni dejavniki v imunskem odzivu.[124,125] Vendar je ob tem treba poudariti, da prekrivanje genov ne pomeni tudi enakih učinkov na bolezen. Za nekatere gene se je izkazalo, da na različne bolezni učinkujejo različno. Kodirajoča varianta R620W v genu *PTPN22* na primer močno prispeva k tveganju za diabetes tipa 1, revmatoidni artritis in sistemski eritematozni lupus (SLE), medtem ko pri CB deluje zaščitno.[126] Podobno je tudi v primeru variante H131R v genu FCGR2A, ki povečuje tveganje za UK in deluje zaščitno pri SLE in multipli sklerozi.[85,124,127]





Slika 1-7: Prekrivanje lokusov, povezanih s KVČB in z drugimi bolezenskimi stanji. Na sliki A je prikazano prekrivanje 71 lokusov CB, na sliki B 47 lokusov UK in na sliki C 99 lokusov KVČB. Zaradi preglednosti prikaza smo upoštevali le lokuse pod pragom statistične značilnosti na celotnem genomu (P = 5 x 10^{-8}). (Pomen kratic: AD – atopični dermatitis, AS – ankilozirajoči spondilitis, AST – astma, BS – Bechetov sindrom, CEL – celiakija, GOB – gobavost (lepra), MS – multipla skleroza, PS – psoriaza, PSC – primarni sklerozantni holangitis, RA – revmatoidni artritis, SLE – sistemski eritematozni lupus, T1D – diabetes tipa 1 (juvenilni tip), T2D – diabetes tipa 2, VIT – vitiligo.)

Prekrivanje genskih lokusov dodatno potrjuje predhodne epidemiološke raziskave o pogostosti ene bolezni v drugi, v katerih so pri bolnikih s KVČB zabeležili povečano pogostost astme, ankilozirajočega spondilitisa, atopičnega dermatitisa, psoriaze, multiple skleroze, revmatoidnega artritisa in miastenije gravis.[128,129]

1.3.3 Raziskave in pristopi po prvem valu GWAS

Z odkritjem povezav med KVČB in 99 lokusi, ki vplivajo na tveganje za nastanek in razvoj bolezni, je bil v razumevanju nekaterih temeljnih značilnosti genskega in biološkega ozadja te bolezni dosežen velik napredek. Med pomembnejše izzive v obdobju po GWAS spadata razrešitev problema manjkajočega dednostnega deleža ter odkrivanje in raziskovanje funkcionalnih posledic vzročnih variant in bolezenskih genov.

Za soočenje z omenjenimi izzivi so bile predlagane številne strategije, npr. fino kartiranje, analiza eQTL-ov in GWAS pri družinah obolelih, vključno s preučevanjem učinkov starševskega izvora na izražanje genov.[118,130]

Fino kartiranje bolezenskih lokusov

Pri finem kartiranju gre za določanje vzročnih variant, ki temelji na genski tipizaciji s po meri narejenimi mikromrežami ali na sekvenciranju (eksonskem ali genomskem) bolezenskih lokusov.[131] K temu bodo zagotovo prispevali tudi izsledki projekta 1000 genomov (angl. 1000 Genomes project), s katerim skušajo zajeti skupne in redke variante različnih etničnih skupin s kombinacijo sekvenciranja genoma in tarčnega sekvenciranja eksonov.[131-133]

Sekvenciranje bolezenskih lokusov

V nedavnih študijah so s tarčnim sekvenciranjem genov *CARD9, NOD2* in *IL23R* odkrili neodvisne redke variante z enakim učinkom na tveganje za bolezen (npr. zaščitni v primeru variante R381Q v *IL23R*), kot ga povzročajo splošne variante, odkrite v GWAS.[133,134] Kmalu za tem so zaščitni učinek variante R381Q potrdili tudi v funkcijski študiji na *ex vivo* celicah Th-17, saj so celice z omenjeno varianto izločale manj vnetnih citokinov.[135] To je dober zgled, kako lahko tarčno sekvenciranje in funkcionalne raziskave podkrepijo predhodna odkritja iz GWAS.

Študije na populacijah neevropskega izvora

Vzorci LD se med različnimi etničnimi skupinami zelo razlikujejo.[71,136] Če redke variante precej prispevajo k manjkajočemu dednostnemu deležu KVČB, so razlike med populacijami še posebej informativne, saj so redke variante v večji meri populacijsko specifične.[130] Genotipiziranje oz. sekvenciranje npr. afriških populacij, ki so gensko najbolj raznolike in pri katerih so haplotipski bloki v primerjavi z Evropejci bolj razdrobljeni, bi lahko prispevalo k oženju bolezenskih lokusov in zmanjšanju števila potencialnih vzročnih variant.[71,136] Doslej je bil ta pristop uspešno uporabljen pri diabetesu tipa 2 in raku dojke.[137,138] Študije na populacijah neevropskega izvora so pokazale tudi na razlike pri nekaterih lokusih, ki prispevajo k tveganju za bolezen. Tako so npr. *NOD2, IL23R* in *ATG16L1* povezani s CB pri Afroameričanih, ne pa tudi v vzhodnoazijskih populacijah.[139-141] Zaradi 2–4-krat večje pogostosti KVČB v primerjavi z drugimi populacijami in velike homogenosti alelov, ki je posledica visoke stopnje sokrvja, je bilo veliko zanimanja za raziskave skupnosti aškenaških Judov.[142] Kenny in sod. so v nedavni študiji GWAS potrdili že znane CB-regije, hkrati pa odkrili pet novih, in sicer na kromosomih 2p15, 5q21.1, 8q21.11, 10q26.3 in 11q12.1.[143]

Analiza izražanja genov in kartiranje eQTL-ov

Ker je večina s KVČB povezanih SNP-jev v nekodirajočih intra- oziroma intergenskih regijah, je očitna potreba po kartiranju alelov, ki posredno vplivajo na izražanje genov.[105,144] Spremembe genskega izražanja, ki jih povzročajo SNP-ji, so lahko kvalitativne ali kvantitativne. Primer kvalitativne spremembe je alternativno izrezovanje intronov. Rivas in sod. so z odkritjem alternativne spojitvene variante v *CARD9*, ki povzroči nastanek zaščitne spojitvene oblike gena, pokazali, da igra alternativno izrezovanje intronov pomembno vlogo v patogenezi KVČB.[133] Labbe in sod. so najprej s finim kartiranjem regije 19p (*IBD6*) s KVČB povezali gen *MAST3*, nato pa z genomsko analizo genskega izražanja ugotovili, da je ta gen ključni modulator vnetnega odziva pri bolnikih z UK.[145,146]

Na drugi strani je učinek eQTL-ov na izražanje genov primer kvantitativne spremembe. Fransen in sod. so z analizo *cis*-eQTL-ov v regijah, ki so bile v študiji GWAS nominalno povezane s CB ($5 \times 10^{-8} < P < 0{,}05$), odkrili dva nova potencialna gena, *UBE2L3* in *BCL3*.[147] Nedavno je bilo ugotovljeno tudi, da SNP-ji, ki so povezani s kompleksnimi boleznimi, v večji meri tkivno specifično vplivajo na izražanje genov.[148] Fu in sod. so pokazali, da slednje velja zlasti za kodirajoče istosmiselne SNP-je kot tudi za SNP-je v regulatornih regijah transkripcijskih faktorjev ter v 3'- in 5'-neprevedljivih regijah.[148]

Vpliv eQTL-ov pa ni omejen le na molekule mRNK, temveč tudi na nekodirajoči RNK, mikro RNK (miRNK) in daljše nekodirajoče molekule RNK (lncRNK).[38] Zwiers in sod. so na primer ugotovili, da SNP rs10889677, ki je v 3'-neprevedljivi regiji gena *IL23R,* povzroči spremembo v vezavnem mestu miRNK (Let-7), kar posledično vodi do povečane produkcije IL23R.[149] Bolj kontroverzni primer je istosmiselna varianta rs10065172 v genu *IRGM*. Sprva je prevladovalo mnenje, da je varianta v popolnem LD ($r^2 = 1$) z različico v številu kopij (angl. copy number variant – CNV), ki povzroči delecijo 20 kb pred genom, kjer so tudi vezavna mesta nekaterih transkripcijskih dejavnikov.[150,151] To naj bi zaviralo izražanje samega gena in vodilo do okvar v avtofagiji pri bolnikih s CB.[150,151] Nedavna študija pa je pokazala, da istosmiselna varianta povzroči spremembo vezavnega mesta miR-196, ki uravnava izražanje *IRGM*.[152] To vodi do zmanjšanega izražanja gena in do okvar ksenofagije (tj. posebne oblike avtofagije), kar posledično pripelje do trdovratnih bakterijskih okužb in vnetja črevesne stene.[153] Najnovejše raziskave so pokazale, da se ekspresijska profila molekul miRNK med CB in UK jasno razlikujeta.[154,155] Te razlike in vse pomembnejša vloga miRNK v patogenezi KVČB bodo v prihodnje najverjetneje pomembno prispevale k izdelavi biooznačevalcev za napoved zagonov in posledic bolezni.[156]

Študije družin, obolelih za KVČB

Ena od ugotovitev, ki izhaja iz študije GWAS KVČB je, da družinske kumulacije bolezni ni mogoče v celoti pojasniti le z doslej odkritimi bolezenskimi aleli, saj ti bolj ali manj enako vplivajo na tveganje tako za družinske kot tudi za sporadične oblike KVČB. Zato se zdi, da bi lahko večje relativno tveganje pri družinskih oblikah KVČB v primerjavi s sporadičnimi pojasnili z upoštevanjem epigenetskih in (ali) paragenetskih dejavnikov (starost matere ob rojstvu otroka, število že rojenih otrok, spol otroka ipd.). Mehanizmi delovanja teh dejavnikov za zdaj ostajajo neznani, lahko pa sklepamo, da bi z njihovim razumevanjem pomembno prispevali tudi k reševanju problema t. i. manjkajočega dednostnega deleža.[114,157] Akolkar in sod. so namreč ugotovili, da je bilo pri potomcih staršev s CB povečano tveganje za bolezen le, kadar so za njo bolehale matere.[158] Slednjo ugotovitev so pred nedavnim podkrepili tudi Zelinkova in sod., ki so hkrati predpostavili, da v razvoju vzpostavljen vzorec metilacije DNK precej prispeva k tveganju za CB v družinah obolelih.[159] Fenomen, pri katerem se epigenetske spremembe kažejo v neenakomernem, od spola odvisnem izražanju genov, imenujemo učinek starševskega izvora (angl. parent-of-origin effect – POO).[160,161] Povečano obolevnost žensk za družinskimi oblikami CB bi lahko pojasnili z vsaj dvema mehanizmoma POO: i) z genskim vtisnjenjem in deaktivacijo očetovega alela ii) z materinim vplivom – prehranjevalne navade in genotip matere vplivajo na okolje razvijajočega se zarodka, in sicer tako, da njeni proteini in (ali) cirkulirajoče RNK-molekule prehajajo skozi posteljico in spreminjajo epigenom zarodka, s čimer vplivajo na fenotip.

1.4 Cilji in teze doktorske disertacije

Študije GWAS so prispevale k boljšemu razumevanju patogeneze in genskega sestava pri KVČB. Ključna pomanjkljivost standardnih študij GWAS je v tem, da omogočajo detekcijo variant z MAF > 1 odstotka, ki pa niso nujno tudi resnične vzročne variante, ki prispevajo k nastanku in razvoju kompleksnih bolezni, ali pa je z njimi mogoče razložiti le majhen delež t. i. manjkajočega dednostnega deleža (heritabilitete). Po za zdaj uveljavljeni hipotezi bodo k razrešitvi problema manjkajoče heritabilitete kompleksnih bolezni precej prispevale redke variante (MAF < 1 %). S trenutnim tehnološkim znanjem je mogoče redke variante detektirati in analizirati s sekvenciranjem (eksonov ali celotnega genoma) ali s finim kartiranjem (angl. fine-mapping) s tarčnimi DNK-mikromrežami. Ena takih je tudi »po meri« izdelana DNK-mikromreža Immunochip (iCHIP), ki vključuje redke variante iz projekta 1000 Genomes[26] in drugih projektov resekvenciranja in SNP-je iz GWAS 12 imunsko posredovanih bolezni (tudi KVČB). Za iCHIP je v primerjavi s standardnimi DNK-mikromrežami (npr. Hap550, Illumina) značilna 20–30-krat večja gostota variant v kandidatnih bolezenskih lokusih. Slednje bo bistveno prispevalo k izboljšanju genske ločljivosti za natančno določanje vzročnih variant v doslej s KVČB povezanih lokusih. Ker so v iCHIP-u zbrane variante 12 imunsko posredovanih bolezni, bo mogoče prvič enovito in natančno preiskati skupne in specifične bolezenske regije ter patogenetske mehanizme kompleksnih bolezni sploh.

Glavni namen te študije je izvedba standardnega asociacijskega testa primerov in kontrol (test hi-kvadrat z eno prostostno stopnjo), s katerim bomo primerjali alelne frekvence variant iz iCHIP-a med slovenskimi bolniki s KVČB in zdravimi posamezniki. Potencialne statistično značilne variante iz slovenske populacije bomo primerjali z rezultati iz nizozemske študije iCHIP pri KVČB. S to primerjavo želimo potrditi ali zavreči obstoj za slovensko populacijo specifičnih lokusov KVČB. Ločeno bomo izvedli tudi asociacijsko analizo SNP-jev genov *NOD2* in *IL23R* z najvišjim faktorjem vpliva v patogenezi KVČB.[10, 11] V ta namen bomo razvili in optimizirali metodo analize DNK s talilnimi krivuljami visoke ločljivosti (angl. high resolution melting analysis – HRMA), ki se je izkazala kot hitra, preprosta, natančna in cenovno ugodna metoda za odkrivanje redkih mutacij in za gensko tipizacijo.[27, 28, 29]

Kljub številčnosti s KVČB povezanih lokusov (in kandidatnih genov) je bilo doslej odkritih presenetljivo malo pravih vzročnih variant ali celo vzročnih genov. Večina povezanih SNP-jev je namreč v nekodirajočih, tj. intra- oziroma intergenskih regijah ali pa celo v regijah brez znanih genov. Zato bomo za izbrane, statistično značilne intergenske SNP-je, ki so v lokusih z več kandidatnimi geni, izvedli kartiranje eQTL-ov. S tem bomo skušali ugotoviti, kateri od kandidatnih genov je posredno povezan z boleznijo, in podkrepili vlogo *cis*-eQTL-ov v patogenezi KVČB.

V tej študiji nameravamo dokazati obstoj fenomena POO s predpostavljenima mehanizmoma delovanja pri bolnikih s KVČB in njihovih starših (klasični trio, tj. bolniki z eno od oblik KVČB in zdravi starši). Ker med slovenskimi preiskovanci ni bilo na voljo družin s KVČB ali pa je eden od staršev manjkal, bo ta raziskava potekala v sodelovanju s kolegi iz Univerzitetnega kliničnega centra Groningen, ki so v ta namen zbrali zadostno število klasičnih triov. Učinek starševskega izvora nameravamo dokazati s testom razmerja verjetij (angl. likelihood ratio test), ki so ga razvili Weinberg in sod.[162] Z omenjenim testom je bil učinek starševskega izvora uspešno dokazan pri diabetesu T1.[163]

Z metodo podpornih vektorjev nameravamo dokazati, da lahko na osnovi izbranih SNP-jev, povezanih s KVČB, tj. genskega profila, znotraj skupine bolnikov s CB ločimo podskupino bolnikov z refraktorno obliko CB. Analiza slednjih je poseben izziv, saj se ti bolniki ne odzivajo na standardno terapijo in razvijejo težko obliko bolezni s fistulami, zaradi česar so primerni kandidati za zdravljenje z biološkimi zdravili, ki delujejo kot antagonisti TNF-α. Ugotovitve nameravamo podkrepiti z analizo *in silico* z ustreznimi bioinformatskimi orodji (npr. DAPPLE[164] in GRAIL[165]).

1.4.1 Pričakovani izvirni znanstveni prispevki

Po za zdaj dostopni literaturi in poizvedbah bo naša raziskava prva tarčna GWA-študija, izvedena pri slovenski populaciji. Pričakujemo, da bomo z rezultati dobili podroben katalog variant, povezanih s KVČB pri slovenski populaciji. Glede na to, da v okviru študije sodelujemo tudi v projektu iCHIP mednarodnega konzorcija IIBDGC, prav tako pričakujemo, da bomo prispevali k odkrivanju novih bolezenskih regij in zapletenega genskega sestava pri KVČB.

Na podlagi ugotovitev kartiranja eQTL-ov pričakujemo, da nam bo v izbranih regijah uspelo določiti ali vsaj zožiti nabor resničnih bolezenskih genov.

Pričakujemo, da bomo z razvojem in optimizacijo metode HRMA poenostavili in pocenili prenos genskih testov za gena *NOD2* in *IL23R* v klinično prakso.

Pričakujemo, da nam bo z rezultati testiranja starševskega izvora (POO) uspelo dokazati obstoj fenomena POO z vsaj enim od predpostavljenih mehanizmov delovanja.

Na podlagi analize podpornih vektorjev bolnikov z refraktorno obliko CB pričakujemo, da bomo lahko na osnovi genskega profila predvideli, kateri bolniki se ne bodo odzivali na standardno terapijo. Na podlagi testiranja s takim prediktivnim modelom bi jih lahko vključili v terapijo z biološkimi zdravili že v zgodnjih fazah razvoja bolezni.

1.4.2 Predpostavke in morebitne omejitve

Iz zasnove preteklih GWA-študij je razvidno, da je treba za odkrivanje novih povezav v raziskavo vključiti več tisoč posameznikov. Ferkolj in sod.[166] so na podlagi poizvedbene ankete po zdravstvenih domovih ugotovili, da je bilo leta 1998 v Sloveniji približno 1150 bolnikov s KVČB. V okviru te raziskave, ki poteka v Centru za humano molekularno genetiko in farmakogenomiko na Medicinski fakulteti Univerze v Mariboru, nam je v sodelovanju s specialisti iz Univerzitetnega kliničnega centra Maribor in Univerzitetnega kliničnega centra Ljubljana uspelo zbrati biološke vzorce 588 bolnikov s KVČB in 256 zdravih posameznikov. Iz lastnih in tujih izkušenj upravičeno predpostavljamo, da manjšega deleža vzorcev DNK ne bo mogoče uspešno hibridizirati v iCHIP. Ob tem predpostavljamo, da določen delež preiskovancev ne bo prestal strogih kriterijev protokola kontrole kakovosti genotipskih podatkov, zaradi česar bo okrnjena statistična moč asociacijske analize in drugih analiz, ki izhajajo iz slednje. Posledično bomo za nekatere polimorfizme, ki prispevajo k tveganju za KVČB, lahko dokazali le nominalno povezavo z boleznijo, za polimorfizme z nižjo pojavnostjo (MAF < 1 %) pa to zaradi premajhnega vzorca najverjetneje ne bo mogoče.

2 PREISKOVANCI IN METODE

2.1 Preiskovanci

V sodelovanju s specialisti iz Univerzitetnega kliničnega centra Maribor in Univerzitetnega kliničnega centra Ljubljana smo v študijo vključili 588 bolnikov s KVČB, od tega 354 s CB, 197 z UK in 37 z nedoločeno obliko KVČB. Diagnoza je bila postavljena v skladu s standardnimi kliničnimi kriteriji (endoskopija, radiologija, histopatologija).[167] V študijo smo kot kontrolno skupino vključili 256 zdravih posameznikov, biološke vzorce pa so jim odvzeli specialisti v Univerzitetnem kliničnem centru Maribor.

Med 354 bolniki s CB smo posebej obravnavali skupino 92 bolnikov, ki se ne odzivajo na standardno terapijo in razvijejo težko obliko bolezni s fistulami, zaradi česar so primerni kandidati za zdravljenje z biološkim zdravilom adalimumab (Humira), ki deluje kot antagonist TNF-α.

Vsi preiskovanci so podpisali izjavo o zavestni privolitvi k sodelovanju v raziskavi. Raziskava je bila izvedena v skladu z etičnim kodeksom Svetovne medicinske organizacije[168] in jo je odobrila Komisija Republike Slovenije za medicinsko etiko (KME št. 77/09/11).

Za testiranje učinka starševskega izvora (POO) smo uporabili 249 klasičnih triov, od tega 115 s CB in 134 z UK. Vseh 115 triov s CB je bilo zahodnoevropskega izvora (vzorci so bili odvzeti na Nizozemskem v Univerzitetnem kliničnem centru Groningen). Od 134 triov z UK jih je bilo 72 zahodnoevropskega, 62 pa indijskega izvora (vzorci so bili odvzeti v Dayanand Medical College and Hospital, Ludhiana, Punjab, Indija). Podrobnejše informacije so podane v razdelku 6.6.4 o materialih in metodah v izvirnem znanstvenem članku 3.

S študijo sodelujemo tudi v projektu iCHIP pod okriljem mednarodnega konzorcija IIBDGC. Podrobnejše informacije o preostalih 75.000 preiskovancih metaanalize so v razdelku 6.6.5 o materialih in metodah v izvirnem znanstvenem članku 4.

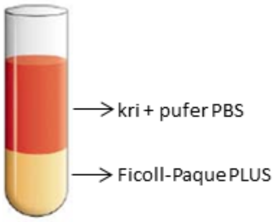
2.2 Metode

2.2.1 Izolacija DNK, RNK in genska tipizacija

Prvi del raziskave je zajemal izolacijo DNK in RNK iz periferne krvi in črevesnih biopsij. Za potrebe te raziskave je zadostovalo 12 mL venske krvi oziroma majhna biopsija črevesnega tkiva (~ 0,25 cm^2).

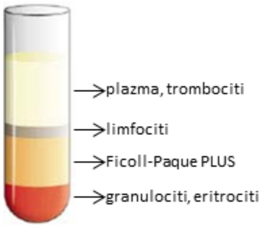
Izolacija DNK in RNK iz krvi

Kri je bila najprej odvzeta v epruveto z antikoagulantom EDTA, nato pa smo jo prenesli v sterilno centrifugirko, ji dodali 12 mL pufra PBS in vsebino rahlo premešali. Zatem smo v novo sterilno centrifugirko nalili 12 mL raztopine Ficoll-Paque PLUS (GE Healthcare) ter dodali suspenzijo krvi in pufra PBS. Pri tem smo poskrbeli, da se kri in Ficoll nista pomešala, kot je prikazano na sliki 2–1.



Slika 2-1: Nanašanje mešanice krvi in pufra PBS na raztopino Ficoll-Paque PLUS

Sledi 30-minutno centrifugiranje pri 18 °C in s hitrostjo 1800 obr./min. Po centrifugiranju se je vzorec ločil v štiri sloje, kot je prikazano na sliki 2–2.



Slika 2-2: Sloji različnih krvnih celic krvi po centrifugiranju

Po centrifugiranju smo najprej odpipetirali plazmo, nato pa limfocite prenesli v novo sterilno centrifugirko. Nato smo k limfocitom ponovno dodali pufer PBS. Sledilo je 15-minutno centrifugiranje pri 18 °C in s hitrostjo 1400 obr./min. Po centrifugiranju smo odpipetirali supernatant, dodali 5 mL pufra PBS, preostalo raztopino močno premešali s stresalnikom in ponovno 15 min centrifugirali pri 18 °C in s hitrostjo 1400 obr./min. Po centrifugiranju smo odpipetirali supernatant in izolirali limfocite.

Limfocitom smo nato dodali 1500 µL TRI reagenta (Sigma-Aldrich). Celično vsebino smo homogenizirali s ponavljajočim hitrim pipetiranjem in z inkubacijo 5 min pri sobni temperaturi. Nato smo homogenat prenesli v mikrocentrifugirke, dodali 300 µL kloroforma in močno premešali s stresalnikom. Sledila sta inkubacija pri sobni temperaturi 2–15 min in centrifugiranje pri 4 °C 15 min in s hitrostjo 12.000 x g. Pri tem se je mešanica ločila v sloj s proteini, sloj z DNK in sloj z RNK. Raztopino z RNK smo prenesli v svežo sterilno mikrocentrifugirko ter dodali 750 µL 2-propanola in premešali z obračanjem. Nato smo mešanico z vzorcem RNK inkubirali pri sobni temperaturi 5–10 min. Sledilo je desetminutno centrifugiranje pri 4 °C in s hitrostjo 12.000 x g. Nato smo dekantirali supernatant in dodali 1500 µL 75-odstotnega etanola, premešali z obračanjem in 5 min centrifugirali pri 4 °C in s hitrostjo 12.000 x g. Ponovno smo odlili supernatant, vzorec sušili na zraku pri sobni temperaturi in ga raztopili v 60 µL prečiščene vode brez ribonukleaz. Vzorec z RNK smo takoj uporabili v nadaljnji analizi. V nasprotnem primeru pa smo ga shranili v zamrzovalniku pri temperaturi −80 °C.

Za izolacijo DNK smo uporabili sloj, iz katerega smo predhodno odstranili RNK. DNK smo precipitirali tako, da smo najprej dodali 450 µL absolutnega etanola. Nato smo mešanico premešali z obračanjem in inkubirali 2–3 min pri sobni temperaturi. Sledilo je petminutno centrifugiranje pri 4 °C in s hitrostjo 2000 x g. Za tem smo dekantirali supernatant, pelet DNK pa dvakrat sprali v 1500 µL 0,1 M natrijevega citrata/10 odstotkov etanola. Med vsakim spiranjem smo pelet DNK z občasnim mešanjem inkubirali 30 min pri sobni temperaturi. Sledilo je petminutno centrifugiranje pri 4 °C in s hitrostjo 2000 x g. Pelet DNK smo resuspendirali v 1500 µL 75-odstotnega etanola in inkubirali 10–20 min pri sobni temperaturi. Sledilo je ponovno petminutno centrifugiranje pri 4 °C in s hitrostjo 2000 x g, supernatant smo dekantirali, DNK pa sušili na zraku in ga končno raztopili v 400 µL vode. Tako izoliran vzorec DNK smo uporabili v nadaljnji analizi ali pa shranili v hladilniku pri temperaturi 4 °C. Koncentracijo in čistost izoliranih RNK- in DNK-vzorcev smo izmerili s spektrofotometrom NanoDrop 2000 UV-Vis Spectrophotometer (Thermo Scientific, ZDA).

Izolacija DNK in RNK iz črevesnih biopsij

RNK in DNK smo izolirali tudi iz črevesnih biopsij bolnikov s KVČB. Košček tkiva je bil odvzet med endoskopskim pregledom črevesa ter je bil takoj po odvzemu shranjen v raztopini RNAlater (Ambion). Nato smo biopsije najprej homogenizirali, nato pa z reagentom TRI reagent, v skladu s prej opisanim protokolom, izolirali DNK in RNK.

Genska tipizacija

Za gensko tipizacijo izbranih SNP-jev smo uporabili DNK bolnikov in zdravih posameznikov. Genotipe smo pridobili na več načinov: a) s kombinacijo verižne reakcije s polimerazo (angl. polymerase chain reaction − PCR) in tehniko polimorfizmov dolžin restrikcijskih fragmetov (angl. restriction length polymorphism − RFLP), ki jima je sledilo odčitavanje fragmentov z agarozno gelsko elektroforezo, b) z analizo talilnih krivulj visoke ločljivosti in c) s hibridizacijo v iCHIP-u v skladu s protokolom proizvajalca (Illumina) v laboratorijih Univerzitetnega kliničnega centra Groningen na Nizozemskem.

Podroben protokol genske tipizacije s PCR-RFLP in gelsko elektroforezo ter za to uporabljeni reagenti in instrumenti so predstavljeni v razdelkih 6.6.1 in 6.6.2 o materialih in metodah v izvirnih znanstvenih člankih 1 in 2.

2.2.2 Optimizacija in genska tipizacija s HRMA

Optimizacija qPCR-HRMA je potekala z instrumentom LightCycler 480 (Roche), in sicer z variiranjem parametrov, kot so npr. koncentracija $MgCl_2$, temperatura prileganja in koncentracija začetnih oligonukleotidov, število ciklov in volumen dodanega dimetilsulfoksida (DMSO), pri čemer so bili drugi parametri pri konstantnih vrednostih. Optimalne parametre smo določili s primerjavo ujemanja rezultatov PCR-RFLP in qPCR-HRMA; optimalen parameter smo izbrali, ko smo dosegli 99-odstotno ujemanje rezultatov med metodama genske tipizacije. Podrobnejši protokol optimizacije, dobljeni optimalni parametri, protokol genske tipizacije ter za to uporabljeni reagenti in instrumenti so predstavljeni v razdelkih 6.6.1 in 6.6.2 o materialih in metodah v izvirnih znanstvenih člankih 1 in 2.

2.2.3 Zasnova DNK-mikromreže ImmunoChip

Immunochip (iCHIP) je po meri izdelana DNK-mikromreža (proizvajalec podjetje Illumina, tehnologija izdelave Infinium BeadChip), ki zajema 196.524 SNP-jev ter kratkih insercij in delecij, izbranih na osnovi predhodnih GWAS 12 avtoimunskih bolezni (avtoimunske bolezni ščitnice, ankilozirajočega spondilitisa, Crohnove bolezni, celiakije, imunodeficience IgA, multiple skleroze, primarne biliarne ciroze, psoriaze, revmatoidnega artritisa, sistemskega lupusa eritematozusa, diabetesa T1 in ulceroznega kolitisa). iCHIP je namenjen finemu kartiranju 187 različnih lokusov in replikaciji domnevno povezanih SNP-jev iz predhodnih GWAS. Posamezna kandidatna regija za fino kartiranje je definirana tako, da se razteza 0,1 cM okoli najbolj statistično značilnega SNP-ja iz GWAS in zajema vse SNP-je, kratke insercije in delecije iz pilotne faze projekta 1000 genomov[26] (1000 Genomes project) in variante iz različnih eksperimentov resekvenciranja. Preostale SNP-je v iCHIP-u predstavljajo replikacijski kandidati 12 avtoimunskih bolezni in 3120 SNP-jev iz GWAS shizofrenije, psihoze in bralno-pisalnih sposobnosti (slednji se uporabljajo za kontrolo kakovosti genske tipizacije).

Gensko tipizacijo smo izvedli s programskim orodjem Genome Studio Data Analysis (Illumina) in podatkovno zbirko s pozicijami genotipskih klastrov, posebej prirejeno za slovensko in nizozemsko populacijo. Genomske pozicije SNP-jev smo odčitali iz manifestne datoteke proizvajalca (Immuno_BeadChip_11419691_B.bpm), temeljijo pa na NCBI-različici 36 (hg18).

2.2.4 Kontrola kakovosti genotipskih podatkov iz iCHIP-a

iCHIP vključuje precej variant, ki niso izpolnjevale kakovostnih standardov proizvajalca. Hkrati pa so bili v študijo vključeni tudi vzorci DNK slabše kakovosti. Zaradi tega je bilo treba najprej izdelati protokol z rigoroznimi kriteriji kontrole kakovosti (KK). Protokol KK smo izvedli z bioinformatskim orodjem PLINK v1.07[169] in s statističnim programskim paketom R.[170] Ker z iCHIP-om testiramo več SNP-jev, lahko tudi majhna napaka v genski tipizaciji odločilno vpliva na rezultate. Vpliv izločitve neustreznih posameznikov na statistično moč (in rezultate) je manjši kot vpliv izločitve neustreznih SNP-jev, saj vsak izločeni SNP predstavlja potencialno spregledano povezavo med boleznijo in določenim lokusom. Zato smo protokol KK razdelili na dva dela in najprej izvedli KK posameznikov, nato pa KK SNP-jev.

Kontrola kakovosti preiskovancev

i) Preverjanje ujemanja spola

Odkrivanje in izločanje posameznikov z nepravilno določenim spolom smo izvedli z analizo SNP-jev na kromosomu X. Ker imajo moški le eno kopijo kromosoma X, ne morejo biti heterozigoti, razen za SNP-je v psevdoavtosomni regiji kromosoma Y. Običajno bodo posamezniki moškega spola za izbrane SNP-je na kromosomu X v celoti homozigoti (delež homozigotnih SNP-jev ~ 1), medtem ko pri ženskih posameznicah pričakujemo manj kot 20 odstotkov homozigotnih SNP-jev. Primer: moški z napačno določenim spolom bo imel večji delež homozigotnih SNP-jev, kot bi pričakovali za vzorec DNK ženske (> 20 %).

ii) Delež napačno določenih genotipov in delež heterozigotnih SNP-jev

Vzorci DNK slabše kakovosti in (ali) nižjih koncentracij bodo z nadpovprečnim deležem napačno določenih in (ali) manjkajočih genotipov (angl. call rate – CR). Mejno vrednost določimo na osnovi porazdelitve manjkajočih genotipov pri vseh preiskovanih posameznikih. Preiskovance z deležem manjkajočih genotipov pod mejno vrednostjo izločimo iz analize. Pri običajnih GWAS se mejna vrednost giblje med tremi in sedmimi odstotki, medtem ko smo jo za iCHIP določili pri dveh odstotkih.

Kontaminacijo DNK ali sorodstvene vezi lahko zaznamo z nenavadno velikim odstopanjem od povprečnega deleža heterozigotnih SNP-jev iz avtosomnih kromosomov. Povprečni delež heterozigotnih SNP-jev je podan z naslednjo enačbo:

$$H_{povp.} = (N - O)/N,$$

pri čemer je N število SNP-jev z nemanjkajočimi genotipi, O pa število homozigotnih SNP-jev pri določenem posamezniku. $H_{povp.}$ se bo razlikovala tako med populacijami kot med različnimi mikromrežami.

iii) Preverjanje sorodstvenih vezi

Ena od temeljnih predpostavk GWAS je, da preiskovanci niso v sorodu (v praksi zadostuje, da je katerikoli par preiskovancev v sorodu do največ tretjega kolena (tretjestopenjska sorodnost)). Izločanje preiskovancev z bližjimi sorodstvenimi vezmi je pomembno zato, ker bodo pri takih posameznikih določeni aleli zastopani pogosteje kot v preostali populaciji. Duplikate in bližnje sorodnike odkrivamo z metodo enakosti po stanju (angl. identity by state – IBS), in sicer tako, da primerjamo povprečni delež enakih alelov avtosomnih SNP-jev med pari preiskovancev. Povprečni delež IBS alelov populacije je funkcija alelne frekvence testiranih SNP-jev. Metoda deluje najbolje, kadar jo uporabimo v zbirki neodvisnih SNP-jev, ki jo dobimo tako, da izključimo regije z visokim LD in redke variante z MAF < 5 odstotkov. V študiji smo izdelali zbirko 8858 SNP-jev, ki so imeli na intervalu 50 kb LD $r^2 < 0{,}2$ ($r^2 = 1$ nakazuje, da sta SNP-ja v popolni korelaciji, 0 pa, da sta neodvisna drug od drugega). S tako zbirko SNP-jev za vsak par preiskovancev izračunamo vrednost IBS, ki narašča premosorazmerno s stopnjo sorodnosti, tako da je za duplikate IBS ~ 1.

Iz podatkov IBS lahko v PLINK-u[169] izračunamo delež nedavno skupnih alelov z metodo enakosti po izvoru (angl. identity by descent – IBD). Pričakovane teoretične vrednosti IBD so naslednje:

- duplikati in enojajčni dvojčki IBD = 1,
- prvostopenjski sorodniki IBD = 0,5 (bratje–sestre; starši-otroci),
- drugostopenjski sorodniki IBD = 0,25 (stari starši-vnuki/-nje), strici, tete–nečaki/-nje); polbratje–polsestre),
- tretjestopenjski sorodniki IBD = 0,125 (bratranci-sestrične; prastarši–pravnuki/-nje).

Zaradi napak v genski tipizaciji, LD in raznolikosti sestave preiskovane populacije prihaja do odstopanj od teoretičnih vrednosti IBD. Po dogovoru iz analize izločimo posameznike z IBD > 0,1875.

iv) Preverjanje izvora

Posameznike z divergentnim izvorom (tj. kadar posameznikovo gensko ozadje precej odstopa od povprečja preiskovane/referenčne populacije) zaznamo z večrazsežnim skaliranjem (angl. multidimensional scaling – MDS). Z MDS smo v PLINK-u[169] izračunali principalne komponente iz podatkovne matrice (N x N), sestavljene iz vrednosti IBD vseh parov preiskovancev za 8858 SNP-jev. Prva komponenta je zajela največ variance dane podatkovne zbirke. Izvor preiskovancev smo ugotavljali z referenčnimi populacijami iz projekta HapMap 3.[72] Posamezne komponente smo grafično prikazali v programskem paketu R.[170]

Kontrola kakovosti SNP-jev

Podobno kot pri KK posameznikov smo tudi v tem primeru arbitrarno določili mejno vrednost napačno določenih in (ali) manjkajočih genotipov (angl. call rate − CR) in izločili SNP-je s CR < 0,98. Prevelika odstopanja od HWE lahko kažejo na napake v genski tipizaciji, zato smo SNP-je, ki so statistično značilno (p < 0,0001) odstopali od HWE, izločili iz nadaljnje analize. Pri tem smo upoštevali le odstopanja v kontrolni skupini, saj so odstopanja od HWE pri bolnikih pričakovana in kažejo na selekcijo (zlasti pri SNP-jih v bolezenskih lokusih). Iz analize smo izločili tudi SNP-je s precejšnjimi razlikami v CR med primeri in kontrolno skupino (p < 0,00001), saj bi se lahko kazali kot lažno pozitivne povezave z boleznijo.

Validacija kontrole kakovosti

Uspešnost KK smo ocenili z diagrami QQ (kvantil-kvantil) in izračunom faktorja genomske inflacije (λ).[169] QQ-diagrame smo uporabili kot metodo za grafično analizo pričakovane in opažene porazdelitve *p*-vrednosti testiranih SNP-jev. Na diagramih so kvantili dotičnih porazdelitev *p*-vrednosti. Če dani SNP ni povezan z boleznijo, bosta pričakovana in opažena p-vrednost približno enaki (na diagramu se to vidi kot premica y = x). Za merodajno oceno KK je slovenska populacija premajhna, zato smo v analizo vključili še nizozemsko populacijo. Diagrame smo načrtovali s statističnim paketom R.[170]

2.2.5 Merjenje genskega izražanja

Vzorce RNK bolnikov s KVČB in zdravih posameznikov smo uporabili pri meritvah genskega izražanja z metodo kvantitativnega PCR (qPCR) v realnem času na instrumentu LightCycler 480 (Roche). Vzorce RNK (γ = 0,1 mg/μL) smo s kitom High Capacity Reverse Transcription (Applied Biosystems) najprej z reverzno transkripcijo prepisali v cDNK. Sestava reakcijske mešanice za prepis je prikazana v preglednici 2–1. Prepisovanje je potekalo po naslednjem temperaturnem protokolu v cikličnem termostatu T1 (Biometra): 10 min pri 25 °C, 120 min pri 37 °C, 5 s pri 85 °C in hlajenje pri 4 °C.

Preglednica 2-1: Sestava reakcijske mešanice za reverzno transkripcijo

Reagent	Volumen/reakcijo [mL]
10 x RT Buffer	2
25 x dNTP Mix (100 mM)	0,8
10 RT Random Primers	2
MultiScribe^{TM} Reverse Transcriptase	1
Nuclease-free H2O	4,2
RNK*	10
Skupaj	**20**

* Začetna koncentracija RNK je bila 0,1 mg/μL.

Preliminarni rezultati asociacijskega testa pri slovenskih bolnikih s CB in zdravih posameznikih so pokazali na statistično značilno povezavo s SNP-jem rs7600901, ki je v neposredni bližini treh kandidatnih genov (*IL18R1*, *IL18RAP* in *IL1RL1*). Prav tako so ta lokus na kromosomu 2q11.2 povezali s CB v nedavni metaanalizi.[103] Zato smo z multipleksno reakcijo qPCR v realnem času izmerili izražanje genov s TaqMan (Applied Biosystems) *IL18R1* (*Hs00977691_m1*), *IL18RAP* (Hs00977695_m1) in *IL1RL1* (Hs00545033_m1) v 369 vzorcih RNK iz krvi in v 52 črevesnih biopsijah. Vsak vzorec smo ponovili v duplikatih in pri izračunih upoštevali povrečno vrednost obeh ponovitev. Sonde kandidatnih genov so bile označene z barvilom FAM. Kot referenčni gen smo uporabili *B2M*, ki je konstantno izražen v vseh celicah (angl. housekeeping gene). Sonda gena *B2M* je bila označena z barvilom VIC. Reakcijo smo izvedli z Maxima Probe qPCR Master Mix (Fermentas). Reakcijski pogoji za qPCR v realnem času so bili: začetna denaturacija pri 95 °C 10 min, nato pa 40 ciklov pri 95°C 15 sekund in pri 60°C 1 min. Nato smo po t. i. Ct-metodi, ki sta jo opisala Livak in Schmittgen,[72] izračunali izražanje *IL18R1*, *IL18RAP* in *IL1RL1* relativno z izražanjem referenčnega gena.

Po tej metodi potrebujemo za izračun relativnega izražanja vrednost Ct (angl. za threshold cycle). Gre za število ciklov, pri katerem količina pomnoženega tarčnega gena doseže prag fluorescence. Nato smo izračunali razliko med Ct(kandidatni gen) in Ct(referenčni gen) oz. vrednost ΔCt. Tako dobljene vrednosti smo kalibrirali s povprečjem vseh meritev, tako da smo v naslednjem koraku izračunali vrednost ΔΔCt kot razliko med ΔCt in ΔCt(povprečni). Relativno količino izraženega gena smo nato v tretjem koraku izrazili kot $2^{-\Delta\Delta Ct}$.

2.2.6 Statistične metode

Del statističnih analiz smo izvedli s programskim paketom SPSS 17.0, del pa z bioinformatskim orodjem PLINKv1.07[169] in s statističnim paketom R.[170] Asociacijsko študijo primerov in kontrol ter razlike v frekvencah genotipov in alelov smo statistično ovrednotili s testoma hi-kvadrat (1 prostostna stopnja) in s Fischer exact. Za izračune in ugotavljanje zadovoljive statistične moči pri asociacijski analizi smo uporabili bioinformatski program Quanto.[171]

Izražanje kandidatnih genov smo izmerili z omenjeno metodo ddCp, ki sta jo opisala Livak in Schmittgen,[72] statistično pa smo ga ovrednotili z neparametričnima testoma Mann-Whitney in Kruskal-Wallis.

Za testiranje učinka starševskega izvora (POO) smo uporabili test razmerja verjetij (angl. likelihood ratio test), ki so ga razvili Weinberg in sod.[162]

Genske profile, s katerimi smo ločili bolnike z refraktorno obliko CB od drugih bolnikov s CB, smo dobili s klasifikacijskimi in z regresijskimi algoritmi po metodi podpornih vektorjev (SVM). Pri tem smo uporabili programski paket CARET 5.15.[172] Podrobnejši opis paketa in delovanje algoritmov sta predstavljena v razdelku 6.6.6 o statističnih metodah v izvirnem znanstvenem članku 5. Ugotovitve smo podkrepili z analizo *in silico* z bioinformatskima orodjema DAPPLE[164] in GRAIL.[165]

Statistične metode, uporabljene v metaanalizi iCHIP konzorcija IIBDGC, so opisane v razdelku 6.6.5 o materialih in metodah ter dodatnih informacijah v izvirnem znanstvenem članku 4.

3 REZULTATI

3.1 Rezultati kontrole kakovosti podatkov iz iCHIP-a

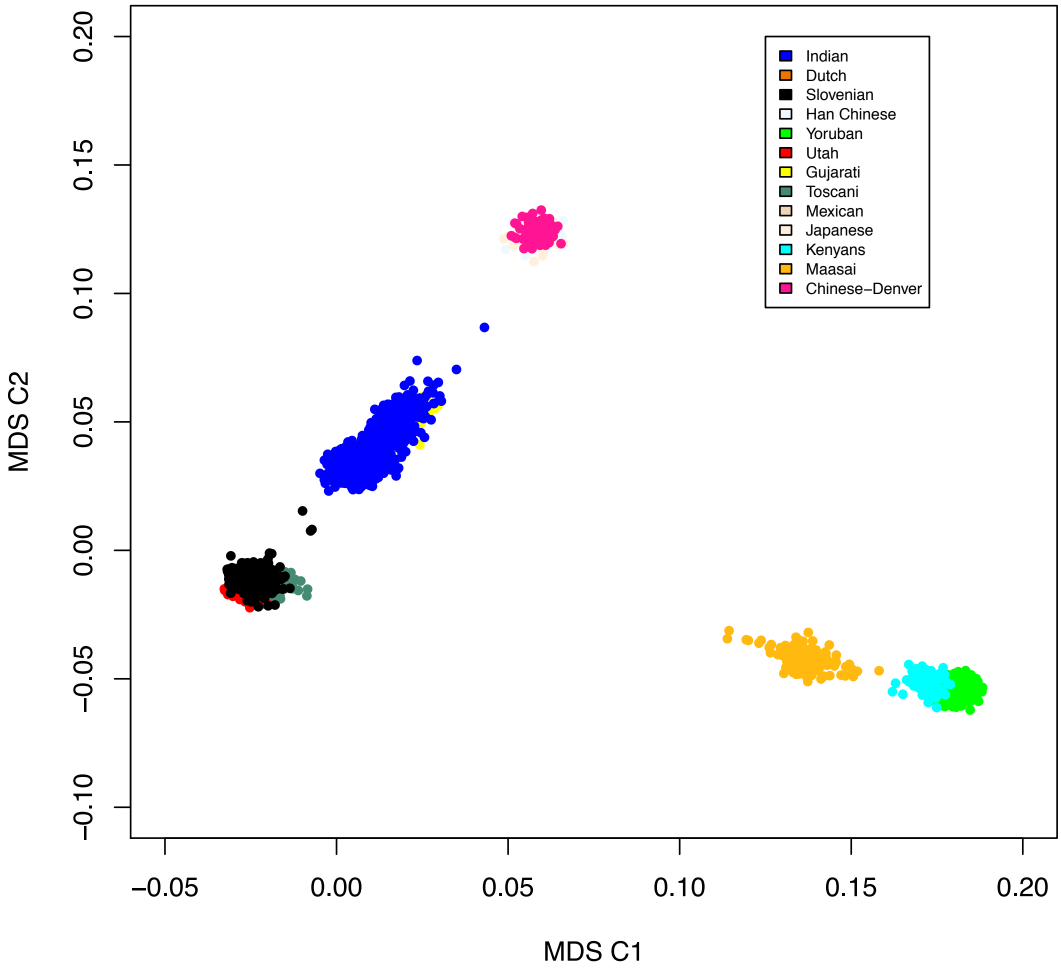
Od 588 slovenskih bolnikov s KVČB (354 s CB, 197 z UK in 37 z IK) ter 256 zdravih posameznikov smo v iCHIP-u uspešno hibridizirali 256 bolnikov in 236 zdravih posameznikov, ki so bili nato vključeni v protokol kontrole kakovosti (KK). Specifičnost in občutljivost nekaterih testov protokola KK sta odvisni od statistične moči, tj. od velikosti preiskovanega vzorca. Zaradi tega smo slovenskim preiskovancem pridružili še 4068 nizozemskih (2262 bolnikov s KVČB in 1806 zdravih posameznikov) ter izvedli KK. Število preiskovancev pred in po KK je povzeto v preglednici 3–1.

Preglednica 3-1: Pregled preiskovancev pred kontrolo kakovosti in po njej

Populacija	Pred KK				Po KK			
	CB	UK	IK	Kontrole	CB	UK	IK	Kontrole
Slovenci	202	39	15	236	179	37	11	210
Nizozemci	1370	892	2	1806	1050	671	2	1374
Skupaj	**1572**	**931**	**17**	**2042**	**1229**	**708**	**13**	**1584**

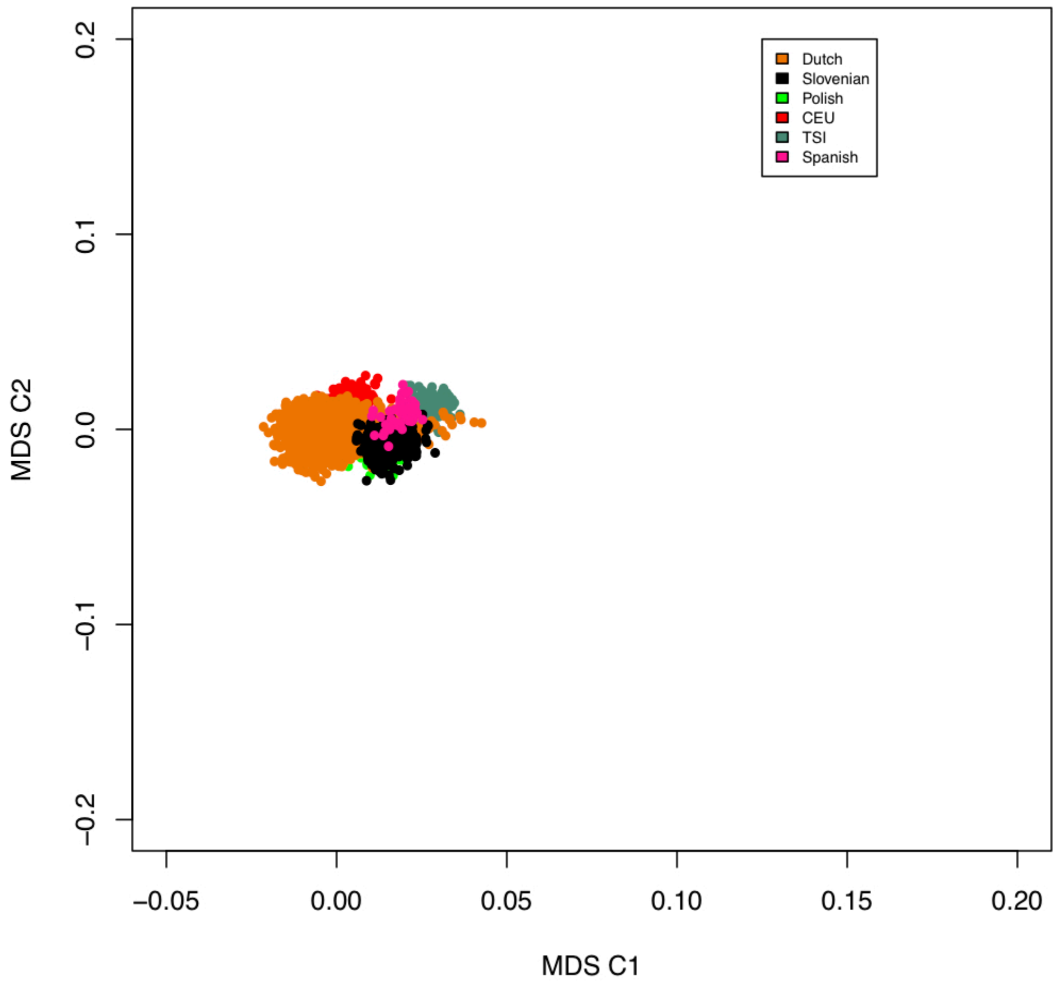
Skupno smo izločili 55 (11 %) slovenskih in 971 (24 %) nizozemskih preiskovancev. Posamezno pa smo iz skupine slovenskih preiskovancev izločili pet z neustrezno določenim spolom, 15 s CR < 0,98, 37 z IBD > 0,1875 in tri z divergentnim izvorom. (Opomba: Nekateri preiskovanci so bili označeni za izločitev v več testih hkrati.)

Izvor preiskovancev smo najprej preverili z globalno različico MDS z referenčnimi populacijami iz projekta HapMap 3,[72] kot je prikazano na sliki 3–1.



Slika 3-1: Globalna različica MDS. Za preverjanje izvora smo uporabili populacije iz projekta HapMap 3.[72] Vsaka točka predstavlja enega preiskovanca, različne barve pa različne populacije (črna Slovence). Komponenti MDS globalnih populacij tvorita »trikotnik«, in sicer tako, da so v levem oglišču zbrane evropske populacije, okoli desnega afriške, ob zgornjem pa azijske. Na sliki so lepo vidni tudi trije slovenski preiskovanci z divergentnim izvorom, ki so načrtani med evropskimi populacijami in dvema indijskima populacijama (modra in rumena).

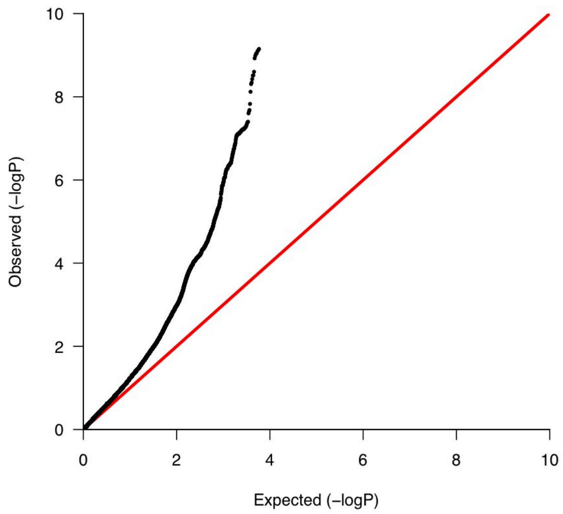
Z lokalno različico MDS, ki smo jo izvedli le na nekaterih evropskih populacijah, nismo zaznali slovenskih preiskovancev z divergentnim izvorom (slika 3-2).



Slika 3-2: Lokalna različica MDS. Na sliki je prikazano sovpadanje slovenske z nekaterimi drugimi evropskimi populacijami. Vsaka točka predstavlja enega preiskovanca, različne barve pa različne populacije (črna Slovence).

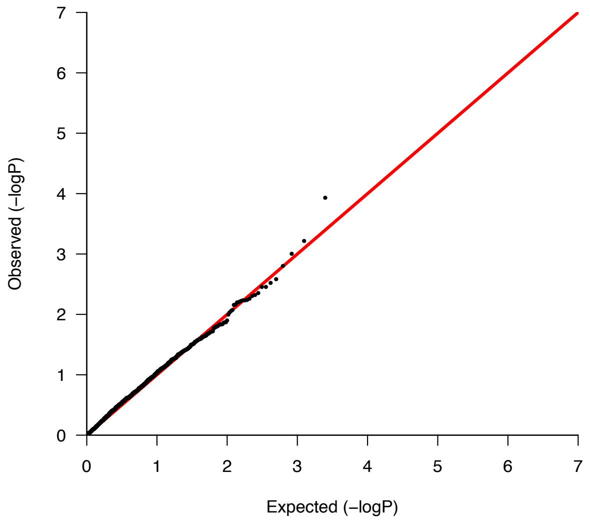
Pri slovenskih in nizozemskih preiskovancih, ki so prestali KK, smo izvedli KK SNP-jev, pri čemer smo od 196.524 variant izločili 26.975 neustreznih. Posamezno smo izločili 17.442 variant z neustreznimi genotipskimi klastri, 8329 variant zaradi različnih CR med primeri in kontrolami, 19 s CR < 0,98 in 1190 zaradi odstopanj od HWE v kontrolah. Tako nam je po KK za asociacijsko analizo ostalo 169.549 variant.

Validacijo KK smo izvedli s QQ-diagrami. Porazdelitev pričakovane in opažene porazdelitve *p*-vrednosti za 169.549 variant pri 1950 bolnikih s KVČB in 1584 zdravih posameznikih (slovenskih in nizozemskih) je prikazana na sliki 3-3.



Slika 3-3: QQ-diagram 169.549 variant pri 1950 bolnikih s KVČB in 1584 zdravih posameznikih (slovenskih in nizozemskih). Faktor genomske inflacije je znašal λ = 1,53, kar je precej nad pričakovano zgornjo mejno vrednostjo (λ = 1,05). Vendar slednje ne preseneča, saj je inflacija posledica zasnove iCHIP-a, ki ga sestavljajo regije, povezane z imunsko posredovanimi boleznimi.

KK smo ustrezno validirali z 2583 SNP-ji iz GWAS shizofrenije, psihoze in bralno-pisalnih sposobnosti, ki niso povezani z imunsko posredovanimi boleznimi, in smo jih uporabili kot negativne kontrole.



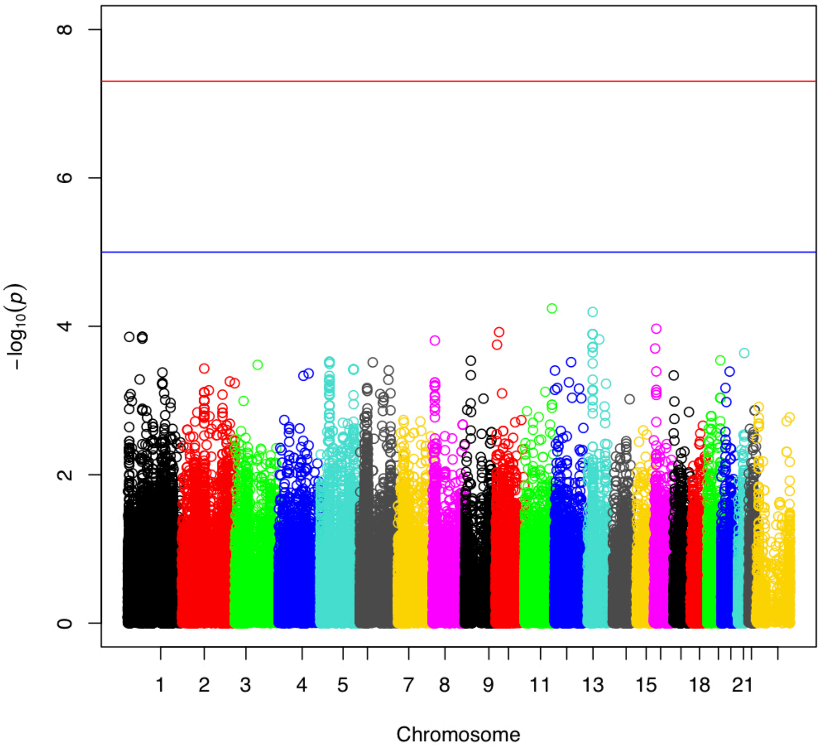
Slika 3-4: QQ-diagram 2583 SNP-jev iz GWAS neimunskih bolezni pri 1950 bolnikih s KVČB in 1584 zdravih posameznikih (slovenskih in nizozemskih). Faktor genomske inflacije je v tem primeru znašal λ = 1,01, kar pomeni ustrezno izvedbo KK.

3.2 Asociacijska analiza

Rezultati metaanalize, v kateri smo sodelovali pod okriljem konzorcija IIBDGC, so predstavljeni v razdelku 6.6.5 o rezultatih v izvirnem znanstvenem članku 4.

3.2.1 Asociacijska analiza slovenskih preiskovancev

Asociacijsko analizo po KK smo izvedli pri 227 bolnikih s KVČB in 210 zdravih posameznikih za 169.549 variant iz iCHIP-a. Rezultati so grafično prikazani na diagramu Manhattan na sliki 3-5.



Slika 3-5: Diagram Manhattan za 227 bolnikov s KVČB in 210 zdravih posameznikov

S slike 3-5 je razvidno, da testirane variante niso dosegle praga statistične značilnosti, ki je po dogovoru določen pri p = 5 x 10^{-8}. K temu je prispevalo premajhno število preiskovancev, kar smo potrdili tudi z izračuni statistične moči, ki jo lahko dosežemo s 436 slovenskimi preiskovanci. Za detekcijo variante z MAF = 0,02 in OR = 1,5 na primer dosežemo le pribl. 30 odstotkov statistične moči. Minimalna zahtevana statistična moč se giblje nad 80 odstotkov, kar bi dosegli pri SNP-jih z MAF > 0,3 in OR = 1,5. Kljub dejstvu, da nismo imeli zadostne statistične moči za detekcijo statistično značilnih povezav med testiranimi variantami in boleznijo, pa na sliki 3-5 izstopajo nekateri klastri variant in nakazujejo na domnevne (nominalne) povezave na kromosomih 1, 5, 8, 13 in 16. Variante, ki so se po statistični značilnosti v asociacijski analizi bolnikov s KVČB in zdravih posameznikov uvrstile najvišje, so prikazane v preglednici 3–2.

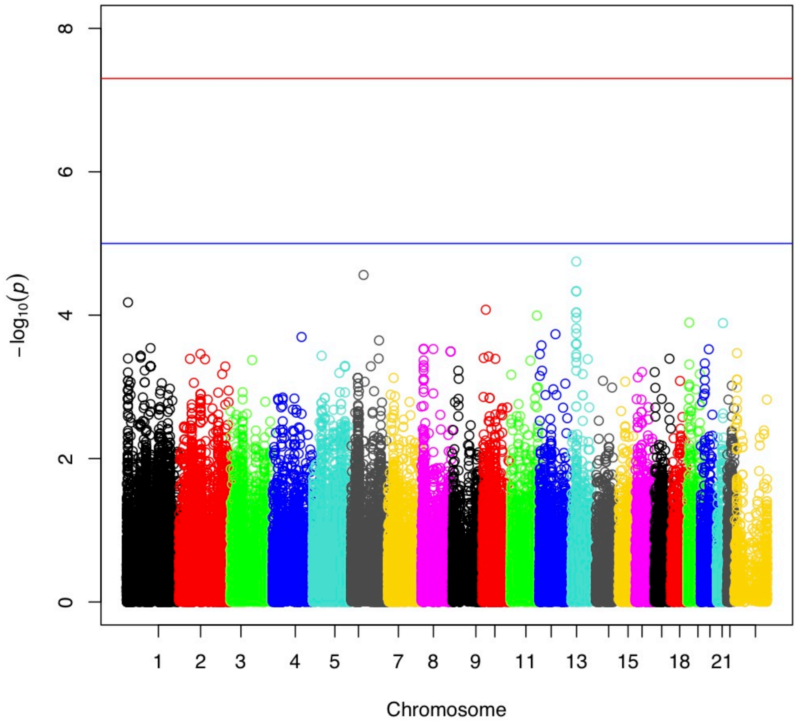
Preglednica 3-2: Deset statistično najbolj povezanih variant iz asociacijske analize 227 bolnikov s KVČB in 210 zdravih posameznikov

SNP	Kr.	Alelna fr.*	p-vrednost	OR	Regija	Lokacija SNP-ja
rs7110733	11	0,2679	5,73E-05	1,807	GRAMD1B	INTRON
1kg_13_41875439	13	0,1851	6,40E-05	1,916	AKAP11 | TNFSF11	INTERGENSKI
imm_16_11371833	16	0,1507	0,0001077	1,961	LOC388210	INTRON
rs11254394	10	0,4282	0,0001197	1,701	CUBN | TRDMT1	INTERGENSKI
imm_1_7932572	1	0,004808	0,0001388	10,1	TNFRSF9 | PARK7	INTERGENSKI
imm_1_67507948	1	0,05981	0,0001389	0,1841	IL23R | IL12RB2	INTERGENSKI
1kg_8_10824467	8	0,07416	0,0001554	2,332	XKR6 | LOC100129441	INTERGENSKI
rs11255182	10	0,2512	0,0001769	1,754	LOC728777 | ITIH5	INTERGENSKI
rs1424121	16	0,1522	0,0001989	0,4307	LOC100129334 | A2BP1	INTERGENSKI
imm_2_102282003	2	0,2033	0,00037	1,762	IL1RL2 | IL1RL1	INTERGENSKI

* Alelna frekvenca v skupini zdravih posameznikov

Najvišji signal prihaja s kromosoma 11, in sicer iz introna kandidatnega gena *GRAMD1B*. Pri tem gre najverjetneje za lažno pozitivni signal, saj gre za osamljen signal, in ne več signalov v klastru, kot je vidno v primeru druge uvrščene regije na kromosomu 13.

V nadaljevanju analize smo skupino 227 bolnikov s KVČB razdelili na klinične podskupine (CB, UK, CB refraktorni) in jih ločeno primerjali z 210 zdravimi posamezniki. Na sliki 3–7 so prikazani rezultati asociacijske analize 179 bolnikov s CB in 210 zdravih posameznikov.



Slika 3-6: Diagram Manhattan za 179 bolnikov s CB in 210 zdravih posameznikov

Variante, ki so se po statistični značilnosti v asociacijski analizi bolnikov s CB in zdravih posameznikov uvrstile najvišje, so prikazane v preglednici 3-3.

Preglednica 3-3: Dvajset statistično najbolj povezanih variant iz asociacijske analize 179 bolnikov s CB in 210 zdravih posameznikov

SNP	Kr.	Alelna fr.*	p-vrednost	OR	Regija	Lokacija SNP-ja
1kg_13_41875439	13	0,1851	1,79E-05	2,057	AKAP11 | TNFSF11	INTERGENSKI
rs1547226	6	0,2297	2,75E-05	0,4338	RAB23	INTRON
imm_1_7932572	1	0,004808	6,63E-05	11,02	TNFRSF9 | PARK7	INTERGENSKI
rs11254394	10	0,4282	8,39E-05	1,769	CUBN | TRDMT1	INTERGENSKI
rs7110733	11	0,2679	0,0001012	1,817	GRAMD1B	INTRON
imm_19_10352706	19	0,09569	0,0001268	0,2716	TYK2 | CDC37	INTERGENSKI
rs2839591	21	0,5168	0,0001288	0,5711	WDR4	INTRON
rs10160994	12	0,08612	0,000184	2,266	E2F7 | NAV3	INTERGENSKI
rs12503254	4	0,3756	0,0002012	1,719	LOC729551 | LOC646316	INTERGENSKI
1kg_8_10824467	8	0,07416	0,0002938	2,315	XKR6 | LOC100129441	INTERGENSKI
rs4737297	8	0,04306	0,0002965	0,06225	LOC100129092 | XKR4	INTERGENSKI
rs6072091	20	0,3254	0,0002991	0,5495	DHX35 | MAFB	INTERGENSKI
rs2610112	8	0,3636	0,0003209	1,692	KHDRBS3 | FLJ45872	INTERGENSKI
rs2613841	8	0,3636	0,0003209	1,692	KHDRBS3 | FLJ45872	INTERGENSKI

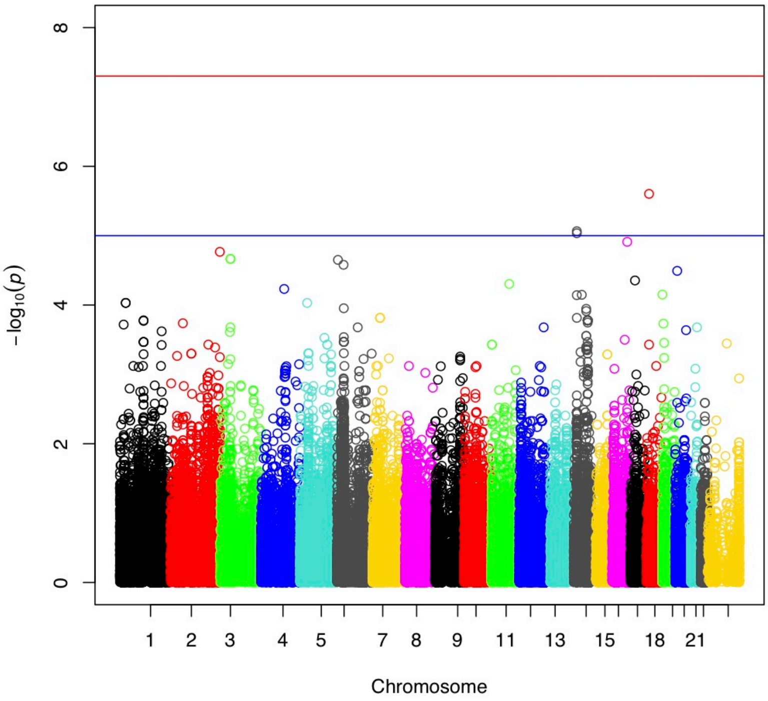
Preglednica 3-3: Dvajset statistično najbolj povezanih variant iz asociacijske analize 179 bolnikov s CB in 210 zdravih posameznikov (nadaljevanje)

SNP	Kr.	Alelna fr.*	p-vrednost	OR	Regija	Lokacija SNP-ja
rs2255407	8	0,3636	0,0003209	1,692	KHDRBS3 | FLJ45872	INTERGENSKI
1kg_X_12837680	23	0,284	0,0003382	0,4849	TLR8	INTRON
imm_2_102282003	2	0,2033	0,0003476	1,807	IL1RL2 | IL1RL1	INTERGENSKI
imm_12_2290787	12	0,122	0,000349	2,005	CACNA1C	INTRON
imm_1_67507948	1	0,05981	0,0003697	0,1776	IL23R | IL12RB2	INTERGENSKI
imm_1_67503956	1	0,05981	0,0003904	0,1786	IL23R | IL12RB2	INTERGENSKI

* Alelna frekvenca v skupini zdravih posameznikov

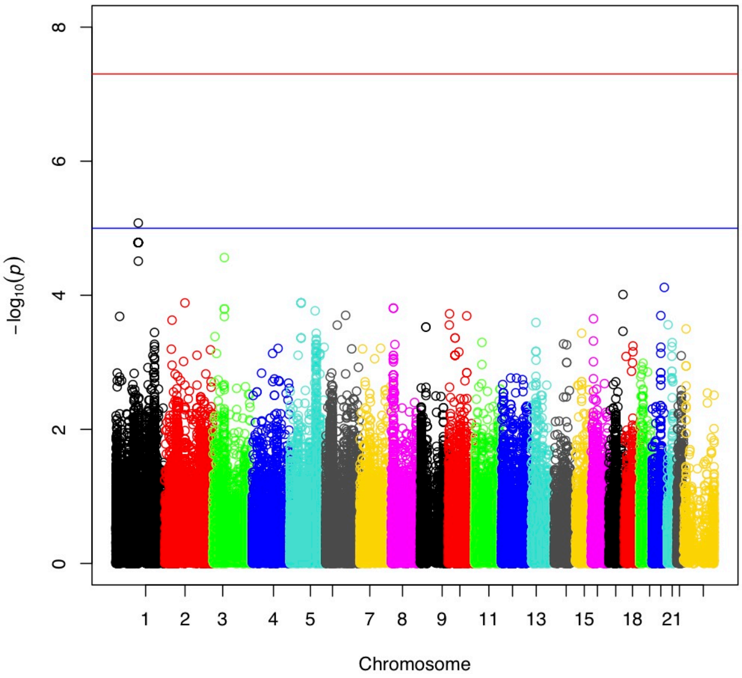
Z ločeno obravnavo bolnikov s CB smo odstranili domnevno lažno pozitivni signal na 11. kromosomu. Najvišji signal je na kromosomu 13, in sicer v intergenski regiji med genoma *AKAP11* in *TNFSF11*. Slednji je bil v nedavni metaanalizi povezan s CB in domnevno pleiotropsko vpliva na tveganje različnih imunsko posredovanih bolezni.[103,173]

Z ločenim obravnavanjem 37 bolnikov z UK smo dodatno okrnili statistično moč. Čeprav so rezultati na sliki 3–7 videti celo statistično značilnejši v primerjavi s predhodnima analizama, pa gre v tem primeru najverjetneje za lažno pozitivne povezave zaradi premajhnega števila bolnikov z UK.



Slika 3-7: Diagram Manhattan za 37 slovenskih bolnikov z UK in 210 zdravih posameznikov

Z ločeno analizo bolnikov z refraktorno obliko CB smo želeli ugotoviti, ali se bodo statistično najbolj povezane regije razlikovale od tistih, ki smo jih odkrili pri vseh bolnikih s CB (slika 3-8).



Slika 3-8: Diagram Manhattan za 92 slovenskih bolnikov z refraktorno obliko CB in 210 zdravih posameznikov

Variante, ki so se po statistični značilnosti v asociacijski analizi bolnikov z refraktorno obliko CB in zdravih posameznikov uvrstile najvišje, so prikazane v preglednici 3–4.

Preglednica 3-4: Dvajset statistično najbolj povezanih variant iz asociacijske analize 92 bolnikov z refraktorno obliko CB in 210 zdravih posameznikov

SNP	Kr.	Alelna fr.*	p-vrednost	OR	Regija	Lokacija SNP-ja
imm_1_113885628	1	0,04306	8,35E-06	3,822	LOC100131737 | MAGI3	INTERGENSKI
imm_3_58509590	3	0,2909	2,74E-05	2,139	ACOX2 | FAM107A	INTERGENSKI
imm_1_114163293	1	0,02153	3,09E-05	4,928	PTPN22	INTRON
imm_20_61798924	20	0,0601	7,63E-05	3,041	TNFRSF6B	EKSON
rs4789241	17	0,4689	9,74E-05	0,4829	SRP68	INTRON
imm_2_102030110	2	0,002392	0,0001298	18,95	IL1R2 | LOC100131131	INTERGENSKI
imm_5_55460169	5	0,002392	0,0001298	18,95	ANKRD55 | LOC727984	INTERGENSKI
1kg_8_10772091	8	0,11	0,0001555	2,392	PINX1 | XKR6	INTERGENSKI
rs10040899	5	0,2967	0,0001706	0,4256	GRAMD3 | ALDH7A1	INTERGENSKI
rs11255182	10	0,2512	0,0001889	2,005	LOC728777 | ITIH5	INTERGENSKI

Preglednica 3-4: Dvajset statistično najbolj povezanih variant iz asociacijske analize 92 bolnikov z refraktorno obliko CB in 210 zdravih posameznikov (nadaljevanje)

SNP	Kr.	Alelna fr.*	p-vrednost	OR	Regija	Lokacija SNP-ja
rs9402255	6	0,1077	0,0001991	2,377	FBXL4	INTRON
imm_20_44204972	20	0,4436	0,0002009	1,951	CD40 | CDH22	INTERGENSKI
rs7078243	10	0,4498	0,0002026	1,947	KIF11	UTR
rs6668463	1	0,1531	0,0002063	2,179	TMCO4	INTRON
imm_16_11371884	16	0,2249	0,0002239	0,3738	LOC388210	INTRON
rs2706807	2	0,01914	0,0002347	4,549	HEATR5B	INTRON
1kg_13_41875439	13	0,1851	0,0002552	2,078	AKAP11 | TNFSF11	INTERGENSKI
rs7284040	21	0,2847	0,0002748	0,4321	C21orf91 | CHODL	INTERGENSKI
imm_10_6472283	10	0,0311	0,0002773	3,587	DKFZp667F0711 | PRKCQ	INTERGENSKI
rs1547226	6	0,2297	0,0002785	0,3862	RAB23	INTRON

* Alelna frekvenca v skupini zdravih posameznikov

Primerjava najvišjih signalov med bolniki z refraktorno obliko CB in drugimi bolniki s CB kaže na tranzicijo signala na kromosom 1. Najvišji signal je dosegla varianta imm_1_113885628 med genoma *LOC100131737* in *MAGI3*. V neposredni bližini je tudi signal, ki ga izpostavlja 3. uvrščena varianta imm_1_114163293 in sovpada z regijo kandidatnega gena *PTPN22*. Slednji je ključni igralec v imunskem odzivu in v patogenezi večine imunsko posredovanih bolezni.[124] Zanimiv je tudi četrtouvrščeni signal (imm_20_61798924), ki sovpada z drugim eksonom gena *TNFRSF6B*. Ta varianta povzroča drugačnosmiselno spremembo, in sicer iz stopkodona v triptofan. Gen *TNFRSF6B* kodira receptor, ki skrbi za nevtralizacijo citotoksičnih ligandov in je bil v predhodnih raziskavah povezan s CB, z lupusom, revmatoidnim artritisom in zdravljenjem diabetesa T2.[174-177]

V nedavni metaanalizi CB so Franke in sod.[103] s CB povezali 71 neodvisnih variant in opisali domnevne kandidatne gene v njihovi bližini. Od 71 variant jih je 59 prestalo protokol KK. Za slednje smo za slovenske bolnike s CB in zdrave posameznike izračunali p-vrednosti in OR, ki so podani v preglednici 3–5.

Preglednica 3-5: Replikacija 59 variant iz uveljavljenih lokusov CB pri slovenskih bolnikih s CB

SNP	Kr.	p-meta	OR	Kandidatni gen(i)	p-slo CB	OR
rs11209026	1p31	1,00 × 10−64	2,66	IL23R	**0,003265**	0,2784
rs2476601	1p13	4,47 × 10−9	1,26	PTPN22	0,2969	1,31
rs4656940	1q23	6,17 × 10−7	1,15	CD244, ITLN1	0,534	1,114
rs7517810	1q24	1,51 × 10−15	1,22	TNFSF18, TNFSF4, FASLG	0,2136	1,239
rs7554511	1q32	1,58 × 10−7	1,14	C1orf106, KIF21B	0,1529	0,776
rs3792109	2q37	6,76 × 10−41	1,34	ATG16L1	0,1755	0,8223
rs3197999	3p21	6,17 × 10−17	1,22	MST1, GPX1, BSN	0,5326	1,105
rs11742570	5p13	7,08 × 10−36	1,33	PTGER4	0,1684	0,8155
rs12521868	5q31	1,41 × 10−20	1,23	SLC22A4, SLC22A5, IRF1, IL3	0,08142	1,288
rs7714584	5q33	7,76 × 10−19	1,37	IRGM	**0,01977**	1,822
rs6556412	5q33	5,37 × 10−14	1,18	IL12B	0,1578	1,235
rs6908425	6p22	1,41 × 10−8	1,17	CDKAL1	0,07487	0,7283
rs1799964	6p21	3,98 × 10−11	1,19	LTA, HLA-DQA2, TNF, LST1, LTB	**0,03225**	0,6863
rs415890	6q27	2,51 × 10−12	1,17	CCR6	0,08184	0,7763
rs1456896	7p12	1,20 × 10−8	1,14	IKZF1, ZPBP, FIGNL1	0,7604	0,9517
rs4871611	8q24	1,51 × 10−12	1,17		**0,02824**	1,386
rs10758669	9p24	1,00 × 10−13	1,18	JAK2	**0,0359**	1,363
rs3810936	9q32	1,00 × 10−15	1,21	TNFSF15, TNFSF8	0,4985	1,113
rs12242110	10p11	1,10 × 10−9	1,15	CREM	0,05575	1,347
rs10761659	10q21	4,37 × 10−22	1,23	ZNF365	0,5845	0,9225
rs4409764	10q24	2,29 × 10−20	1,22	NKX2-3	0,8155	0,9669
rs7927997	11q13	5,62 × 10−13	1,17	C11orf30	0,2287	1,198
rs11564258	12q12	6,17 × 10−21	1,74	MUC19, LRRK2	0,2612	1,888
rs3764147	13q14	1,41 × 10−10	1,17	C13orf31	0,5612	0,9059
rs2076756	16q12	3,98 × 10−69	1,53	NOD2	0,8942	0,9783
rs2872507	17q21	1,51 × 10−9	1,14	GSMDL, ZPBP2, ORMDL3, IKZF3	0,4567	0,8968
rs1893217	18p11	1,29 × 10−14	1,25	PTPN2	0,9427	0,9858
rs1736020	21q21	9,33 × 10−12	1,16		0,8904	0,98
rs2838519	21q22	2,09 × 10−14	1,18	ICOSLG	0,6451	1,07
rs2797685	1p36	2,69 × 10−10	1,05	VAMP3	0,2581	1,25
rs1998598	1q31	4,90 × 10−9	1,04	DENND1B	0,6948	0,9417
rs3024505	1q32	8,32 × 10−9	1,12	IL10, IL19	0,05357	1,453
rs13428812	2p23	1,41 × 10−8	1,06	DNMT3A	0,4056	1,136
rs10495903	2p21	7,70 × 10−8	1,14	THADA	0,8239	0,9545
rs10181042	2p16	6,61 × 10−9	1,14	C2orf74, REL	0,1025	1,266
rs2058660	2q12	1,58 × 10−12	1,19	IL18RAP, IL12RL2, IL18R1, IL1RL1	0,197	1,234
rs6738825	2q33	1,82 × 10−7	1,06	PLCL1	0,5266	0,9127
rs13073817	3p24	8,20 × 10−7	1,08		0,7494	0,951
rs11167764	5q31	1,10 × 10−9	1,06	NDFIP1	0,6124	1,093
rs359457	5q35	5,25 × 10−8	1,08	CPEB4	0,2249	1,192
rs17309827	6p25	6,16 × 10−7	1,1		0,3078	0,8552

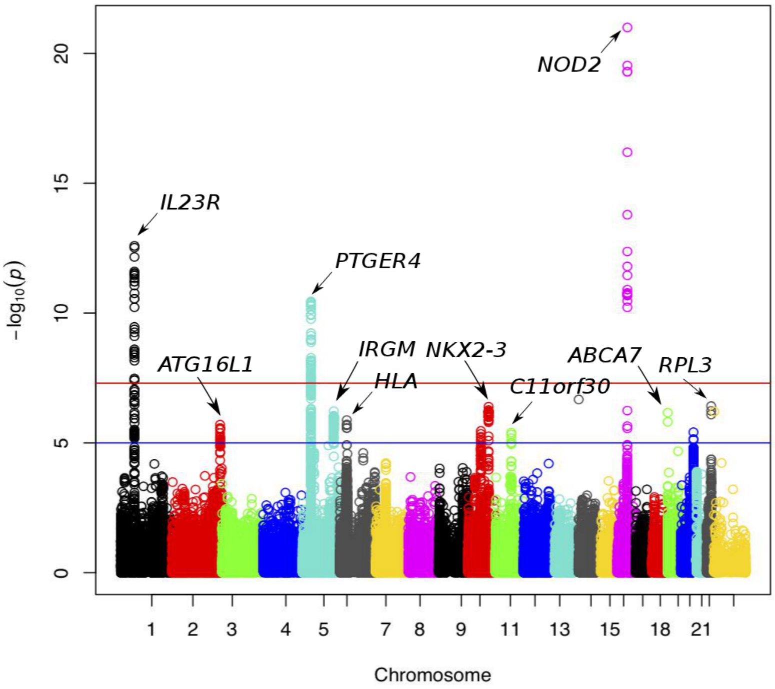
Preglednica 3-5: Replikacija 59 variant iz uveljavljenih lokusov CB pri slovenskih bolnikih s CB (nadaljevanje)

SNP	Kr.	p-meta	OR	Kandidatni gen(i)	p-slo CB	OR
rs1847472	6q15	3,63 × 10−6	1,07	BACH2	0,662	0,9355
rs212388	6q25	1,41 × 10−7	1,1	TAGAP	0,7801	1,044
rs6651252	8q24	2,29 × 10−6	1,23		0,7929	0,9446
rs4077515	9q34	4,37 × 10−19	1,18	CARD9, SNAPC4	0,3234	1,156
rs12722489	10p15	8,51 × 10−6	1,11	IL2RA	0,05769	0,6639
rs1250550	10q22	2,00 × 10−10	1,19	ZMIZ1	0,3478	1,159
rs102275	11q12	7,24 × 10−8	1,08	FADS1	0,7142	1,059
rs694739	11q13	3,38 × 10−7	1,1	PRDX5, ESRRA	0,696	1,061
rs2062305	13q14	2,00 × 10−6	1,1	TNFSF11	0,1081	1,261
rs8005161	14q35	1,29 × 10−8	1,23	GALC, GPR65	0,5085	0,84
rs17293632	15q22	1,41 × 10−13	1,12	SMAD3	0,1108	1,301
rs151181	16p11	1,10 × 10−10	1,07	IL27, SH2B1, EIF3C, LAT, CD19	0,1623	1,223
rs3091315	17q12b	1,70 × 10−13	1,2	CCL2, CCL7	0,168	0,7952
rs12720356	19p13	9,20 × 10−10	1,12	TYK2, ICAM1, ICAM3	0,2473	1,386
rs281379	19q13	8,60 × 10−10	1,07	FUT2, RASIP1	0,1085	1,264
rs4809330	20q13	2,51 × 10−12	1,12	RTEL1, TNFRSF6B, SLC2A4RG	0,4375	0,8818
rs713875	22q12	5,70 × 10−9	1,08	MTMR3	0,3998	1,129
rs2413583	22q13	1,70 × 10−10	1,23	MAP3K7IP1	0,9584	1,011

Iz preglednice 3–5 je razvidno, da smo z asociacijsko analizo slovenskih bolnikov s CB statistično značilno (nominalno) replicirali štiri variante oz. naslednje kandidatne regije: *IL23R*, *IRGM*, *HLA-DQA*, 8q24 in *JAK2*. Pri tem je zanimivo, da je učinek variant ob genih *IL23R* in *IRGM* pri slovenskih bolnikih s CB nasproten (OR) kot pri drugih kohortah iz metaanalize.

3.2.2 Asociacijska analiza slovenskih in nizozemskih preiskovancev

Skupno asociacijsko analizo smo izvedi s testom Cochran-Mantel-Haenszel, ki upošteva morebitne razlike v klastrih med slovenskimi in nizozemskimi preiskovanci. Na sliki 3-9 in v preglednici 3–6 so rezultati asociacijske analize 1229 bolnikov s CB in 1584 zdravih posameznikov.



Slika 3-9: Diagram Manhattan za 1229 bolnikov s CB in 1584 zdravih posameznikov

Na sliki 3-9 so vsi lokusi nad mejo nominalne statistične povezanosti ($p < 10^{-5}$) označeni z najverjetnejšim kandidatnim genom. V preglednici 3-6 pa so podane le variante nad pragom statistične značilnosti GWAS ($p < 10^{-8}$).

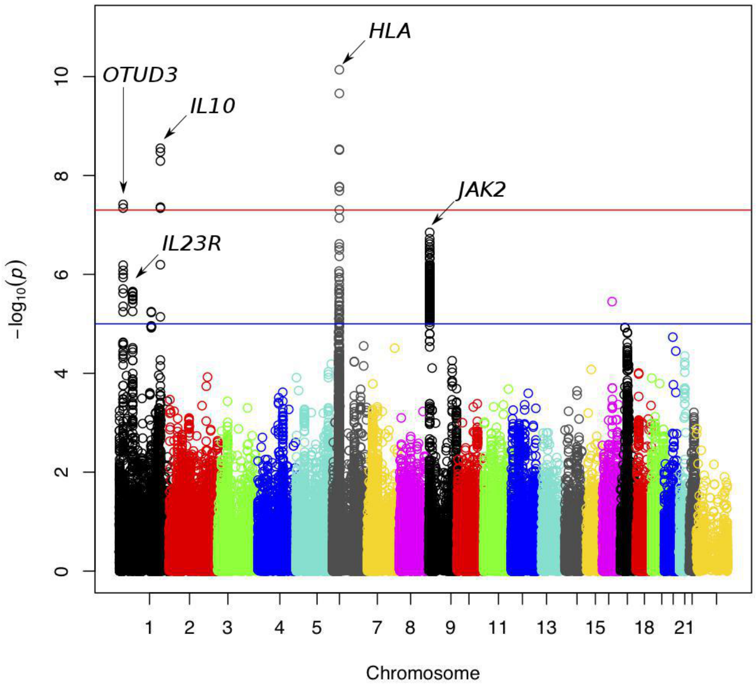
Preglednica 3-6: Statistično značilne variante iz asociacijske analize 1229 bolnikov s CB in 1584 zdravih posameznikov

SNP	Kr.	Alelna fr.*	p-vrednost	OR	Regija	Lokacija SNP-ja
imm_16_49303427	16	0,03311	1,59E-22	3,25	NOD2	EKSON
rs2066844	16	0,03311	1,59E-22	3,25	NOD2	EKSON
rs2066847	16	0,02693	9,01E-22	3,503	NOD2	EKSON
imm_1_67454257	1	0,4534	3,13E-13	0,658	IL23R	INTRON
imm_5_40434853	5	0,3861	9,69E-12	0,6668	PTGER4	INTERGENSKI
ccc-10-101272594-C-A	10	0,46	5,26E-08	1,357	NKX2-3	INTERGENSKI
imm_10_101264048	10	0,4596	5,39E-08	1,357	NKX2-3	INTERGENSKI
rs10136056	14	0,4189	9,50E-08	1,351	HNRNPC | RPGRIP1	INTERGENSKI

* Alelna frekvenca v skupini zdravih posameznikov

S slike 3-9 in iz preglednice 3-6 je razvidno, da smo s skupno asociacijsko analizo replicirali najpomembnejše lokuse, ki so bili v predhodnih raziskavah povezani s CB. Vseeno pa tudi s 3813 preiskovanci nismo imeli dovolj statistične moči za odkrivanje novih bolezenskih lokusov. Pri varianti rs10136056 iz preglednice 3–6 gre najverjetneje za lažno pozitiven signal.

Na sliki 3-10 in v preglednici 3–7 so rezultati asociacijske analize 708 bolnikov z UK in 1584 zdravih posameznikov.



Slika 3-10: Diagram Manhattan za 708 bolnikov z UK in 1584 zdravih posameznikov

Na sliki 3–10 so vsi lokusi nad mejo nominalne statistične povezanosti ($p < 10^{-5}$) označeni z najverjetnejšim kandidatnim genom. V preglednici 3–7 pa so podane le variante nad pragom statistične značilnosti GWAS ($p < 10^{-8}$).

Preglednica 3-7: Statistično značilni varianti iz asociacijske analize 708 bolnikov z UK in 1584 zdravih posameznikov

SNP	Kr.	Alelna fr.*	p-vrednost	OR	Regija	Lokacija SNP-ja
rs2097431	6	0,3372	6,82E-09	1,449	HLA-DRB1 | HLA-DQA1	INTERGENSKI
imm_1_205006527	1	0,1642	5,1E-09	1,494	L10	INTRON
imm_1_20044447	1	0,4778	6,51E-08	0,7112	RNF186 | OTUD3	INTERGENSKI

* Alelna frekvenca v skupini zdravih posameznikov

S slike 3–10 in iz preglednice 3–7 je razvidno, da smo s skupno asociacijsko analizo replicirali nekatere pomembnejše lokuse, ki so bili v predhodnih raziskavah povezani z UK.

3.3 Optimizacija HRMA in asociacijska analiza kandidatnih genov *IL23R* in *NOD2*

Optimizacijo HRMA smo izvedli pri 295 bolnikih s KVČB (159 CB in 136 UK) in 345 zdravih posameznikih. Z izborom optimalnih parametrov smo na podlagi različnih temperatur taljenja amplikonov ločili obe skupini homozigotov (GG in TT) polimorfizma gena *IL23R* (rs7517847). Na podlagi različnih oblik talilnih krivulj pa nam je uspelo ločiti obe skupini homozigotov od heterozigotov. Referenčne genotipe vseh preiskovancev smo dobili z metodo PCR-RFLP. Primerjava referenčnih genotipov z genotipi iz HRMA je pokazala na 98,6-odstotno ujemanje. Z asociacijsko analizo s testom Fischer exact smo potrdili povezavo med CB in polimorfizmom gena *IL23R* ($p < 0{,}001$, $OR = 0{,}588$) v slovenski populaciji.

Optimizacija HRMA za tri polimorfizme gena *NOD2* (rs2066844, rs2066845 in rs2066847) je potekala podobno kot za *IL23R*, vendar smo v primeru polimorfizma rs2066845 morali načrtovati in uporabiti neoznačeno sondo, s čimer smo izboljšali ločevanje homozigotov (GG in CC). Referenčne genotipe vseh preiskovancev smo dobili z metodo PCR-RFLP. Primerjava referenčnih genotipov z genotipi iz HRMA je za vse preiskovane polimorfizme pokazala na 100-odstotno ujemanje. V asociacijski analizi s testom Fischer exact smo potrdili povezavo CB s polimorfizmoma rs2066847 ($p = 0{,}001$, $OR = 3{,}011$) in rs2066845 ($p = 2{,}62 \times 10−4$, $OR = 14{,}117$) v slovenski populaciji. Podrobnejši pregled rezultatov je v razdelkih 6.6.1 in 6.6.2 o rezultatih v izvirnih znanstvenih člankih 1 in 2.

3.4 Rezultati kartiranja eQTL-ov

Med variantami, ki so v asociacijskem testu pri slovenskih bolnikih s CB in zdravih posameznikih pokazali na najmočnejšo nominalno povezavo, je tudi SNP imm_2_102282003 ($p = 3{,}476 \times 10^{-4}$, $OR = 1{,}807$), ki je v neposredni bližini treh kandidatnih genov: *IL18R1*, *IL18RAP* in *IL1RL1*. Prav tako so ta lokus na kromosomu 2q11.2 s CB povezali v nedavni metaanalizi.[103]

V iCHIP-u je v tem tarčnem lokusu 779 variant, od tega 130 LD neodvisnih variant ($r^2 < 0,2$). Za 12 od 130 variant, ki so bile v asociacijskem testu slovenskih bolnikov s CB in zdravih posameznikov nominalno statistično povezane (p < 0,05) s CB, smo izmerili in statistično ovrednotili vpliv na izražanje tarčnih kandidatnih genov v 369 vzorcih RNK iz periferne venske krvi. Rezultati so v preglednici 3–8.

Preglednica 3-8: Rezultati kartiranja eQTL-ov v lokusu *IL18RAP* (periferna venska kri)

SNP	p-vrednost*	OR	*IL18RAP*		*IL18R1*		*IL1RL1*	
			M-W p-vrednost	K-W p-vrednost	M-W p-vrednost	K-W p-vrednost	M-W p-vrednost	K-W p-vrednost
imm_2_102282003	0,0003476	1,807	0,764	0,909	0,245	0,019	0,167	0,102
imm_2_102225353	0,006225	0,6171	0,117	0,263	0,341	0,3	0,026	**0,001**
imm_2_102245323	0,01457	1,424	0,455	0,749	0,03	0,065	0,406	0,646
imm_2_102321900	0,001714	0,5998	0,718	0,689	0,902	0,527	0,004	**0,001**
imm_2_102239801	0,01751	1,41	0,431	0,721	0,031	0,075	0,44	0,571
ccc-2-102248784-A-G	0,02094	0,6345	0,048	0,137	0,339	0,004	0,243	**0,003**
imm_2_102277489	0,045	1,392	0,093	0,073	0,587	0,844	0,983	0,976
imm_2_102286017	0,02735	0,5843	0,254	0,128	0,064	0,006	0,829	0,366
imm_2_102286469	0,001265	1,594	0,972	0,805	0,943	0,986	0,019	0,034
imm_2_102326078	0,03302	1,614	0,284	0,486	0,184	0,335	0,883	0,984
imm_2_102328782	0,02547	0,5973	0,106	0,084	0,021	0,02	0,734	0,597
imm_2_102382648	0,04302	0,7446	0,184	0,34	0,559	0,868	0,052	0,12

M-W – Mann-Whitney, K-W – Kruskal-Wallis

* p-vrednost iz asociacijskega testa slovenskih bolnikov s CB in zdravih posameznikov

Iz preglednice 3–8 so razvidne tri s CB povezane variante (imm_2_102225353, imm_2_102321900, ccc-2-102248784-A-G), za katere smo izmerili statistično značilen vpliv na izražanje gena *IL1RL1*. Mejna vrednost statistične značilnosti je bila pri p < 0,004.

V predhodnih raziskavah je bil ugotovljen večji tkivno specifični vpliv na izražanje genov pri SNP-jih, povezanih s kompleksnimi boleznimi.[148] Zato smo izražanje tarčnih genov izmerili tudi v 52 črevesnih biopsijah bolnikov s CB, vendar nismo odkrili statistično značilnih razlik. Dodatno smo preverili, ali so razlike v izražanju tarčnih genov med RNK-vzorci 22 biopsij z znaki vnetja in 30 normalnimi biopsijami, vendar nismo odkrili statistično značilnih razlik v izražanju tarčnih genov.

3.5 Rezultati analize učinka starševskega izvora (POO)

Z analizo učinka starševskega izvora pri nizozemskih triih smo odkrili povezavo med varianto gena *NOD2* (L1007fs; OR = 21, *p*-vrednost = 0,013) in CB, ki pa nam je ni uspelo replicirati v neodvisni kohorti (OR = 0,97, *p*-vrednost = 0,95). S skupno analizo triov s CB in UK smo zaznali tudi povezavo med genoma *IL12B* (OR = 3,2, *p*-vrednost = 0,019) in *PRDM1* (OR = 5,6, *p*-vrednost = 0,04). Z analizo indijskih triov nam je uspelo dokazati povezavo z lokusom IL10 (OR = 0,2, *p*-vrednost = 0,03). Podrobnejši pregled rezultatov analize je podan v razdelku 6.6.4 o rezultatih v izvirnem znanstvenem članku 3.

3.6 Rezultati analize prediktivnih modelov

Z metodo podpornih vektorjev smo najboljšo ločitev 92 bolnikov z refraktorno obliko CB od drugih bolnikov s CB dosegli z 59 SNP-ji (AUC = 0,64). Z asociacijsko analizo teh 59 SNP-jev smo odkrili povezavo s statistično značilnima SNP-jema (rs9592040 in rs346818). SNP-ja sta v bližini genov *PHACTR2* in *KIR3DL3*, ki sta vpletena v patogenezo drugih imunsko posredovanih bolezni. V analizi z bioinformatskima orodjema DAPPLE in GRAIL smo dokazali statistično značilno funkcijsko povezavo s petimi geni rs395561 (*EFNA5*), rs2690262 (*UNCX*), rs9720889 (*SCXB*), rs10992979 (*BARX1*), rs7127817 (*NCAM1*). Z analizo *in silico* genskega izražanja smo ugotovili, da je gen *NCAM1* precej in specifično izražen v celicah NK. Te ugotovitve smo podkrepili s poizvedbami po literaturi in ugotovili, da je gen *NCAM1* nenormalno izražen v mukusu debelega črevesa pri bolnikih s CB in da igra vlogo pri zdravljenju multiple skleroze. Iz tega izhaja, da smo s kombinacijo strojnega učenja (podporni vektorji) in bioinformatske analize odkrili enega od genov, ki najverjetneje vpliva na raznolikost odziva bolnikov s CB na zdravljenje s kortikosteroidi. Podrobnejši opis je podan v razdelku 6.6.6 o rezultatih v izvirnem znanstvenem članku 5.

4 RAZPRAVA

Raziskava spada v obdobje raziskav po prvem valu GWAS, v katerem so napori raziskovalcev po vsem svetu v glavnem usmerjeni v odkrivanje vzročnih variant in analizo njihovih funkcionalnih posledic, s tem pa posledično k iskanju rešitev problema manjkajočega dednostnega deleža kompleksnih bolezni, kakršna je tudi KVČB. V ta namen so številne raziskovalne skupine, ki se ukvarjajo z genetiko KVČB pod okriljem konzorcija International Inflammatory Bowel Disease Consortium (IIBDGC), združile moči v projektu ImmunoChip (iCHIP). S sodelovanjem v tem projektu, ki je v metaanalizi vključeval približno 75.000 bolnikov s KVČB in zdravih posameznikov, smo prispevali k odkritju 71 novih statistično značilnih povezav z boleznijo, tako da se je skupno število bolezenskih lokusov z 99 povzpelo na 163. Večina novo odkritih lokusov (110) je povezana z obema podtipoma bolezni, manjši del pa je specifičen za CB (30) in UK (23). Za približno polovico (50) od 110 skupnih lokusov je značilen enak faktor vpliva pri obeh podtipih bolezni, 60 pa jih vpliva tako, da pri enem podtipu deluje zaščitno, pri drugem pa veča tveganje in nasprotno. Glede odkrivanja skupnih mehanizmov kompleksnih bolezni je pomembna tudi ugotovitev, da je več kot dve tretjini lokusov KVČB (113 od 163) impliciranih v patogenezi drugih imunsko posredovanih bolezni, kar je kar 8,6-krat več, kot bi pričakovali po naključju.[178] Prekrivanje lokusov KVČB je bilo največje z lokusi, ki so povezani z ankilozirajočim spondilitisom in s psoriazo. Zanimivo je, da precejšnje število lokusov KVČB vsebuje gene, vpletene v patogenezo primarnih imunodeficienc. Slednjim sta skupna deregulacija imunskega sistema in zmanjšano število T-celic, kar pa lahko vodi do resnih infektov.[179] Največje prekrivanje genov je bilo z geni, ki večajo tveganje za razvoj mikobakterijskih okužb, kot je gobavost (leprosija).[180] Na domnevni vpliv okužb na razvoj KVČB lahko sklepamo tudi na podlagi tega, da so bile okvare KVČB genov *STAT3*[181,182] in *CARD9*[183] povezane s kožnimi okužbami s stafilokoki in s kandidozami.

V raziskavi smo z iCHIP-om izvedli prvo tarčno študijo GWAS pri slovenski populaciji, pri čemer smo izdelali podroben katalog variant, povezanih s KVČB. Zaradi relativno omejenega števila preiskovancev sicer nismo imeli zadostne statistične moči za odkritje statistično značilnih povezav, vendar so klastri variant na kromosomih 1, 5, 8, 13 in 16 nakazovali na nominalne povezave z boleznijo. V asociacijski analizi bolnikov s CB je največji signal prihajal iz intergenske regije med genoma *AKAP11* in *TNFSF11*. Slednji je bil v nedavni metaanalizi povezan s CB in domnevno pleiotropsko vpliva na tveganje različnih imunsko posredovanih bolezni.[103,173] Asociacijska analiza bolnikov z refraktorno obliko CB je izpostavila regiji kandidatnih genov *TNFRSF6B* in *PTPN22*. Slednji je ključni dejavnik v imunskem odzivu in v patogenezi večine imunsko posredovanih bolezni.[124] Drugačnosmiselna mutacija v genu *TNFRSF6B* povzroča spremembo iz stopkodona v triptofan. Gen *TNFRSF6B* kodira receptor, ki skrbi za nevtralizacijo citotoksičnih ligandov in je bil v predhodnih raziskavah povezan s CB, z lupusom, revmatoidnim artritisom in zdravljenjem diabetesa T2.[174-177] Še bolj preseneča dejstvo, da je *TNFRSF6B* eden od ključnih prediktivnih genov za razvoj in zdravljenje KVČB in je bil leta 2011 tudi zaščiten s patentom (US 2011/0177502 A1). Z asociacijsko analizo slovenskih bolnikov s CB smo nominalno replicirali povezavo s 4/59 kandidatnimi regijami (*IL23R*, *IRGM*, *HLA-DQA*, 8q24 in *JAK2*). Ko smo slovenskim preiskovancem pridružili več nizozemskih, smo v ločenih asociacijskih analizah CB in UK repilicirali večino uveljavljenih oz. predhodno znanih povezav s KVČB. Slednje ponovno kaže, da smo bili v raziskavah slovenskih preiskovancev omejeni z velikostjo vzorca, zato bodo za potrditev naših rezultatov potrebne nadaljnje asociacijske študije z večjim številom preiskovancev in (ali) metaanalize.

V raziskavi smo ločeno izvedli tudi asociacijsko analizo SNP-jev genov *NOD2* in *IL23R*, za katera je značilen največji faktor vpliva v patogenezi KVČB,[10, 11] in prvič povezali omenjena gena s patogenezo KVČB pri slovenskih bolnikih. Genotipske podatke smo pridobili z optimizacijo metode analize DNK s talilnimi krivuljami visoke ločljivosti (HRMA), ki se je v izkazala kot hitra, preprosta, natančna in cenovno ugodna za odkrivanje redkih mutacij in za gensko tipizacijo.[27, 28, 29] Duerr in sod. so polimorfizem rs7517847 (int6G/T) v genu *IL23R* prvi povezali s CB.[98] To povezavo je potrdila tudi sočasna študija GWAS, ki so jo izvedli Rioux in sod.[89] IL-23 prek svojega receptorja IL23R spodbuja sekrecijo interferona gama in proliferacijo celic Th, provnetno pa deluje tudi na podlagi spodbujanja sinteze IL-17 v aktiviranih limfocitih Th.[142] Kaskadna pot IL-23/IL-17 je vpletena v patogenezo nekaterih drugih imunsko posredovanih bolezni, kot so psoriaza, multipla skleroza in revmatoidni artritis.[124] V tej asociacijski študiji smo pri slovenski populaciji prvič potrdili povezavo med CB in polimorfizmoma gena *NOD2* rs2066847 in rs2066845. Prvo povezavo med CB in *NOD2* so sočasno odkrile tri skupine raziskovalcev.[32-34] *NOD2* kodira znotrajcelični receptor, ki ga aktivira minimalno biološko aktivna komponenta peptidoglikana – muramildipeptid (MDP). Ta je v celični steni tako po Gramu pozitivnih kot po Gramu negativnih bakterij. NOD2 je v največji meri izražen v monocitih in Panethovih celicah črevesnega epitelija.[35,36] Omenjene vzročne variante gena *NOD2* so polimorfizmi posameznega nukleotida (angl. single nucleotide polymorphism – SNP), in sicer dve drugačnosmiselni mutaciji R702W (Arg702Trp) in G908R (Gly908Arg) ter skrajševalna mutacija L1007fs (Leu1007fs). Ti SNP-ji so na C-terminalnem koncu levcinsko bogatih ponovitev (angl. leucine-rich repeats – LRR), ki je domena za zaznavanje bakterijskih produktov.[37] Zato se zdi, da ti SNP-ji zmanjšujejo odzivnost proteina pri zaznavanju bakterijskih produktov, kar posledično vpliva na prirojeni imunski odziv.[37]

V študiji smo s kartiranjem *cis*-eQTL-ov v regiji s tremi kandidatnimi geni (*IL18R1*, *IL18RAP* in *IL1RL1*) ugotovili, da polimorfizma rs10178214, rs1041973 in redka varianta ccc-2-102248784-A-G statistično značilno vplivajo na izražanje gena *IL1RL1*. To je v nasprotju z rezultati raziskave, v kateri so Franke in sod. ugotovili, da polimorfizem rs2058660 vpliva na izražanje gena *IL18RAP*.[103] Z omenjenim polimorfizmom pri slovenskih bolnikih s CB nismo odkrili nikakršne statistično značilne povezave. Nedavno je bilo ugotovljeno tudi, da SNP-ji v večji meri tkivno specifično vplivajo na izražanje genov.[148] Vendar v raziskavi nismo opazili statistično značilnega izražanja tarčnih genov v črevesnih biopsijah bolnikov s CB.

V raziskavi smo pri nizozemskih bolnikih s KVČB prvič ugotavljali učinek POO v bolezenskih regijah in pri tem odkrili omejen učinek genov *IL12B*, *PRDM1* in *NOD2*, ki pa nam ga v večji neodvisni kohorti ni uspelo replicirati. Nominalno značilen učinek POO smo opazili tudi za gen *IL10* pri Indijcih. Kljub temu da dopuščamo možnost lažno pozitivnih rezultatov, pa naše ugotovitve nesporno nakazujejo na utišanje očetovskega alela, s čimer izključujemo njegov vpliv na tveganje za bolezen pri omenjenih genih. To je skladno z epidemiološkimi študijami, ki kažejo, da se KVČB pogosteje prenaša na potomce z matere kot z očeta. Izvor učinkov POO za zdaj ostaja nepojasnjen. Človek lahko kot diploidni organizem preživi tudi kot nosilec, v povprečju, 500 recesivnih mutacij, saj izbris nekega alela kompenzira njegov homolog.[184] Genomsko vtisnjenje pa z inaktivacijo enega od starševskih haplotipov precej vpliva na redukcijo diploidnosti, posledično pa tudi na kompenzacijski reševalni mehanizem. Ta fenomen opisuje tudi za zdaj najbolj uveljavljena hipoteza t. i. hipoteza starševskega konflikta (angl. parental conflict hypothesis), po kateri oba spola tekmujeta v prenašanju genskega materiala v naslednjo generacijo. Vendar s tem na primer ne moremo razložiti genomskega vtisnjenja genov imunskega sistema.[185] Hipotetično je lahko tako zato, ker je s tem preprečen vpliv škodljivih imunskih odzivov matere na zarodek, oz. ker je za preživetje zarodka pomembno, da sta si imunska odziva obeh čim podobnejša, so preferenčno izraženi le geni matere, očetovi pa so utišani. Za neponovljivost naših odkritij v neodvisni in večji kohorti so na voljo vsaj tri razlage: i) omejena velikost našega vzorca; ii) pri človeku za zdaj še ni znano, pri koliko zaporednih generacijah lahko še zaznamo stabilen učinek genomskega vtisnjenja (raziskave pri miših so pokazale, da je bil učinek genomskega vtisnjenja stabilen vsaj tri generacije[185]); iii) na genomsko vtisnjenje vplivajo tudi dejavniki okolja, kar bi lahko pomenilo, da bodo v različnih populacijah vtisnjeni različni geni.[186,187]

V študiji smo z metodo podpornih vektorjev izdelali genski profil, ki ga je sestavljalo 59 SNP-jev. Po za zdaj dostopni literaturi in poizvedbah je to prvi primer uporabe takega pristopa za ločevanje bolnikov z refraktorno obliko CB od drugih bolnikov s CB. V asociacijski analizi smo odkrili statistično značilno povezavo s SNP-jema, ki sta v bližini genov *PHACTR2* in *KIR3DL3*. PHACTR2 deluje kot regulator aktina ter je bil povezan s Parkinsonovo boleznijo in z multiplo sklerozo.[188,189] *KIR3DL3* kodira imunoglobulinski receptor celic NK in je bil vpleten v patogenezo ankilizirajočega spondilitisa in multiple skleroze.[190,191] Zanimivo je, da so Vidal-Castineira in sod.[192] ugotovili, da so se bolniki z virusom hepatitisa C, ki so bili hkrati nosilci določenih genotipov *KIR,* slabše odzivali na zdravljenje z biološkimi zdravili. V bioinformatski analizi smo dokazali statistično značilno funkcijsko povezavo s petimi geni rs395561 (*EFNA5*), rs2690262 (*UNCX*), rs9720889 (*SCXB*), rs10992979 (*BARX1*), rs7127817 (*NCAM1*). Z *in silico* analizo genskega izražanja smo ugotovili, da je gen *NCAM1* precej in specifično izražen v celicah NK. Te ugotovitve smo podkrepili s poizvedbami po literaturi in ugotovili, da je gen *NCAM1* nenormalno izražen v mukusu debelega črevesa pri bolnikih s CB in da igra vlogo pri zdravljenju multiple skleroze z biološkim zdravilom natalizumab.[176,177] Na podlagi tega ugotavljamo, da smo s kombinacijo strojnega učenja s podpornimi vektorji in bioinformatske analize odkrili enega od genov, ki najverjetneje vpliva na raznolikost odziva bolnikov s CB na zdravljenje s kortikosteroidi. Našo ugotovitev bo treba podkrepiti v prihodnjih funkcijskih študijah.

Če povzamemo, so rezultati predstavljene študije delno osvetlili zapleten genski sestav pri KVČB. Za potrditev nekaterih rezultatov so potrebne dodatne asociacijske in metaanalize z večjim številom preiskovancev. Pojasnitev mehanizmov in sprožilcev, s katerimi so posamezni kandidatni geni udeleženi v patogenezi KVČB, pa zahteva tudi dodatne funkcijske študije na molekularni ravni. V prihodnjih raziskavah bo najverjetneje treba več pozornosti posvetiti raziskavam interakcij mukusnega imunskega sistema in črevesne mikroflore, zlasti pa ugotoviti, kako bolezenski geni vplivajo na imunski sistem, da se ta v danem trenutku začne odzivati sovražno do simbiotskih mikroorganizmov.

5 VIRI IN LITERATURA

1. Xavier, R.J. & Podolsky, D.K. Unravelling the pathogenesis of inflammatory bowel disease. *Nature* **448**, 427-34 (2007).

2. Kaser, A., Zeissig, S. & Blumberg, R.S. Inflammatory bowel disease. *Annu Rev Immunol* **28**, 573-621 (2010).

3. Guindi, M. & Riddell, R.H. Indeterminate colitis. *J Clin Pathol* **57**, 1233-44 (2004).

4. Johnston, R.D. & Logan, R.F. What is the peak age for onset of IBD? *Inflamm Bowel Dis* **14 Suppl 2**, S4-5 (2008).

5. Henderson, P., van Limbergen, J.E., Wilson, D.C., Satsangi, J. & Russell, R.K. Genetics of childhood-onset inflammatory bowel disease. *Inflamm Bowel Dis* **17**, 346-61 (2011).

6. Sauer, C.G. & Kugathasan, S. Pediatric inflammatory bowel disease: highlighting pediatric differences in IBD. *Med Clin North Am* **94**, 35-52 (2010).

7. Baumgart, D.C. & Sandborn, W.J. Inflammatory bowel disease: clinical aspects and established and evolving therapies. *Lancet* **369**, 1641-57 (2007).

8. Blonski, W., Buchner, A.M. & Lichtenstein, G.R. Inflammatory bowel disease therapy: current state-of-the-art. *Curr Opin Gastroenterol* **27**, 346-57 (2011).

9. Burger, D. & Travis, S. Conventional medical management of inflammatory bowel disease. *Gastroenterology* **140**, 1827-1837 e2 (2011).

10. Bernstein, C.N. *et al.* World Gastroenterology Organization Practice Guidelines for the diagnosis and management of IBD in 2010. *Inflamm Bowel Dis* **16**, 112-24 (2010).

11. Danese, S. & Fiocchi, C. Ulcerative colitis. *N Engl J Med* **365**, 1713-25 (2011).

12. Bernstein, C.N., Blanchard, J.F., Kliewer, E. & Wajda, A. Cancer risk in patients with inflammatory bowel disease: a population-based study. *Cancer* **91**, 854-62 (2001).

13. Caprilli, R., Viscido, A. & Latella, G. Current management of severe ulcerative colitis. *Nat Clin Pract Gastroenterol Hepatol* **4**, 92-101 (2007).

14. Vella, M., Masood, M.R. & Hendry, W.S. Surgery for ulcerative colitis. *Surgeon* **5**, 356-62 (2007).

15. Logan, I. & Bowlus, C.L. The geoepidemiology of autoimmune intestinal diseases. *Autoimmun Rev* **9**, A372-8 (2010).

16. Molodecky, N.A. *et al*. Increasing incidence and prevalence of the inflammatory bowel diseases with time, based on systematic review. *Gastroenterology* **142**, 46-54 e42; quiz e30 (2012).

17. Ocepek, A., Skok, P. Kronicne vnetne crevesne bolezni in rak debelega crevesa in danke. *Zdrav. Vest.* **75**, 99-103 (2006).

18. Cosnes, J., Gower-Rousseau, C., Seksik, P. & Cortot, A. Epidemiology and natural history of inflammatory bowel diseases. *Gastroenterology* **140**, 1785-94 (2011).

19. Bernstein, C.N. & Shanahan, F. Disorders of a modern lifestyle: reconciling the epidemiology of inflammatory bowel diseases. *Gut* **57**, 1185-91 (2008).

20. Molodecky, N.A. & Kaplan, G.G. Environmental risk factors for inflammatory bowel disease. *Gastroenterol Hepatol (N Y)* **6**, 339-46 (2010).

21. van der Heide, F. *et al*. Differences in genetic background between active smokers, passive smokers, and non-smokers with Crohn's disease. *Am J Gastroenterol* **105**, 1165-72 (2010).

22. Koutroubakis, I.E. *et al*. Appendectomy, tonsillectomy, and risk of inflammatory bowel disease: case-controlled study in Crete. *Dis Colon Rectum* **42**, 225-30 (1999).

23. Radford-Smith, G.L. *et al*. Protective role of appendicectomy on onset and severity of ulcerative colitis and Crohn's disease. *Gut* **51**, 808-13 (2002).

24. Halme, L. *et al*. Family and twin studies in inflammatory bowel disease. *World J Gastroenterol* **12**, 3668-72 (2006).

25. Hanauer, S.B. Inflammatory bowel disease: epidemiology, pathogenesis, and therapeutic opportunities. *Inflamm Bowel Dis* **12 Suppl 1**, S3-9 (2006).

26. Russell, R.K. & Satsangi, J. IBD: a family affair. *Best Pract Res Clin Gastroenterol* **18**, 525-39 (2004).

27. Spehlmann, M.E. *et al*. Epidemiology of inflammatory bowel disease in a German twin cohort: results of a nationwide study. *Inflamm Bowel Dis* **14**, 968-76 (2008).

28. Daly, M.J. & Rioux, J.D. New approaches to gene hunting in IBD. *Inflamm Bowel Dis* **10**, 312-7 (2004).

29. Altmuller, J., Palmer, L.J., Fischer, G., Scherb, H. & Wjst, M. Genomewide scans of complex human diseases: true linkage is hard to find. *Am J Hum Genet* **69**, 936-50 (2001).

30. Brant, S.R. & Shugart, Y.Y. Inflammatory bowel disease gene hunting by linkage analysis: rationale, methodology, and present status of the field. *Inflamm Bowel Dis* **10**, 300-11 (2004).

31. Van Limbergen, J., Wilson, D.C. & Satsangi, J. The genetics of Crohn's disease. *Annu Rev Genomics Hum Genet* **10**, 89-116 (2009).

32. Hugot, J.P. *et al*. Association of NOD2 leucine-rich repeat variants with susceptibility to Crohn's disease. *Nature* **411**, 599-603 (2001).

33. Ogura, Y. *et al*. A frameshift mutation in NOD2 associated with susceptibility to Crohn's disease. *Nature* **411**, 603-6 (2001).

34. Hampe, J. *et al*. Association between insertion mutation in NOD2 gene and Crohn's disease in German and British populations. *Lancet* **357**, 1925-8 (2001).

35. Lala, S. *et al*. Crohn's disease and the NOD2 gene: a role for paneth cells. *Gastroenterology* **125**, 47-57 (2003).

36. Rosenstiel, P. *et al*. TNF-alpha and IFN-gamma regulate the expression of the NOD2 (CARD15) gene in human intestinal epithelial cells. *Gastroenterology* **124**, 1001-9 (2003).

37. Travassos, L.H. *et al*. Nod1 and Nod2 direct autophagy by recruiting ATG16L1 to the plasma membrane at the site of bacterial entry. *Nat Immunol* **11**, 55-62 (2010).

38. Khor, B., Gardet, A. & Xavier, R.J. Genetics and pathogenesis of inflammatory bowel disease. *Nature* **474**, 307-17 (2011).

39. Economou, M., Trikalinos, T.A., Loizou, K.T., Tsianos, E.V. & Ioannidis, J.P. Differential effects of NOD2 variants on Crohn's disease risk and phenotype in diverse populations: a metaanalysis. *Am J Gastroenterol* **99**, 2393-404 (2004).

40. Lesage, S. *et al*. CARD15/NOD2 mutational analysis and genotype-phenotype correlation in 612 patients with inflammatory bowel disease. *Am J Hum Genet* **70**, 845-57 (2002).

41. Hugot, J.P. *et al*. Prevalence of CARD15/NOD2 mutations in Caucasian healthy people. *Am J Gastroenterol* **102**, 1259-67 (2007).

42. Zhang, H., Massey, D., Tremelling, M. & Parkes, M. Genetics of inflammatory bowel disease: clues to pathogenesis. *Br Med Bull* **87**, 17-30 (2008).

43. Lee, J.C. & Parkes, M. Genome-wide association studies and Crohn's disease. *Brief Funct Genomics* **10**, 71-6 (2011).

44. Risch, N. & Merikangas, K. The future of genetic studies of complex human diseases. *Science* **273**, 1516-7 (1996).

45. Cardon, L.R. & Bell, J.I. Association study designs for complex diseases. *Nat Rev Genet* **2**, 91-9 (2001).

46. Arnott, I.D. *et al*. NOD2/CARD15, TLR4 and CD14 mutations in Scottish and Irish Crohn's disease patients: evidence for genetic heterogeneity within Europe? *Genes Immun* **5**, 417-25 (2004).

47. Cario, E. & Podolsky, D.K. Differential alteration in intestinal epithelial cell expression of toll-like receptor 3 (TLR3) and TLR4 in inflammatory bowel disease. *Infect Immun* **68**, 7010-7 (2000).

48. Torok, H.P. *et al.* Crohn's disease is associated with a toll-like receptor-9 polymorphism. *Gastroenterology* **127**, 365-6 (2004).

49. Zouali, H. *et al.* CARD4/NOD1 is not involved in inflammatory bowel disease. *Gut* **52**, 71-4 (2003).

50. Yang, H. *et al.* Intercellular adhesion molecule 1 gene associations with immunologic subsets of inflammatory bowel disease. *Gastroenterology* **109**, 440-8 (1995).

51. Hong, J. *et al.* Polymorphisms in NFKBIA and ICAM-1 genes in New Zealand Caucasian Crohn's disease patients. *J Gastroenterol Hepatol* **22**, 1666-70 (2007).

52. van Bodegraven, A.A. *et al.* Genetic variation in myosin IXB is associated with ulcerative colitis. *Gastroenterology* **131**, 1768-74 (2006).

53. Schwab, M. *et al.* Association between the C3435T MDR1 gene polymorphism and susceptibility for ulcerative colitis. *Gastroenterology* **124**, 26-33 (2003).

54. Croucher, P.J., Mascheretti, S., Foelsch, U.R., Hampe, J. & Schreiber, S. Lack of association between the C3435T MDR1 gene polymorphism and inflammatory bowel disease in two independent Northern European populations. *Gastroenterology* **125**, 1919-20; author reply 1920-1 (2003).

55. Friedrichs, F. *et al.* The Crohn's disease susceptibility gene DLG5 as a member of the CARD interaction network. *J Mol Med (Berl)* **86**, 423-32 (2008).

56. Stoll, M. *et al.* Genetic variation in DLG5 is associated with inflammatory bowel disease. *Nat Genet* **36**, 476-80 (2004).

57. Lohmueller, K.E., Pearce, C.L., Pike, M., Lander, E.S. & Hirschhorn, J.N. Meta-analysis of genetic association studies supports a contribution of common variants to susceptibility to common disease. *Nat Genet* **33**, 177-82 (2003).

58. Gaya, D.R., Russell, R.K., Nimmo, E.R. & Satsangi, J. New genes in inflammatory bowel disease: lessons for complex diseases? *Lancet* **367**, 1271-84 (2006).

59. Stokkers, P.C., Reitsma, P.H., Tytgat, G.N. & van Deventer, S.J. HLA-DR and -DQ phenotypes in inflammatory bowel disease: a meta-analysis. *Gut* **45**, 395-401 (1999).

60. Satsangi, J. *et al.* Contribution of genes of the major histocompatibility complex to susceptibility and disease phenotype in inflammatory bowel disease. *Lancet* **347**, 1212-7 (1996).

61. Silverberg, M.S. *et al.* A population- and family-based study of Canadian families reveals association of HLA DRB1*0103 with colonic involvement in inflammatory bowel disease. *Inflamm Bowel Dis* **9**, 1-9 (2003).

62. Lappalainen, M. *et al*. Association of IL23R, TNFRSF1A, and HLA-DRB1*0103 allele variants with inflammatory bowel disease phenotypes in the Finnish population. *Inflamm Bowel Dis* **14**, 1118-24 (2008).

63. Fernandez, L. *et al*. IBD1 and IBD3 determine location of Crohn's disease in the Spanish population. *Inflamm Bowel Dis* **10**, 715-22 (2004).

64. Yamamoto-Furusho, J.K. *et al*. Polymorphisms in the promoter region of tumor necrosis factor alpha (TNF-alpha) and the HLA-DRB1 locus in Mexican mestizo patients with ulcerative colitis. *Immunol Lett* **95**, 31-5 (2004).

65. Newman, B. *et al*. CARD15 and HLA DRB1 alleles influence susceptibility and disease localization in Crohn's disease. *Am J Gastroenterol* **99**, 306-15 (2004).

66. Puzanowska, B., Prokopowicz, D., Ziarko, S., Radziwon, P. & Lapinski, T.W. The incidence of HLA DRB1*0103 in ulcerative colitis patients in north-eastern Poland. *Hepatogastroenterology* **50**, 1436-8 (2003).

67. Fernandez, L. *et al*. A recombined haplotype in the major histocompatibility region contains a cluster of genes conferring high susceptibility to ulcerative colitis in the Spanish population. *Inflamm Bowel Dis* **11**, 785-91 (2005).

68. Annese, V. *et al*. HLA-DRB1 alleles may influence disease phenotype in patients with inflammatory bowel disease: a critical reappraisal with review of the literature. *Dis Colon Rectum* **48**, 57-64; discussion 64-5 (2005).

69. Gabriel, S.B. *et al*. The structure of haplotype blocks in the human genome. *Science* **296**, 2225-9 (2002).

70. International HapMap, C. A haplotype map of the human genome. *Nature* **437**, 1299-320 (2005).

71. International HapMap, C. *et al*. A second generation human haplotype map of over 3.1 million SNPs. *Nature* **449**, 851-61 (2007).

72. International HapMap, C. *et al*. Integrating common and rare genetic variation in diverse human populations. *Nature* **467**, 52-8 (2010).

73. Ragoussis, J. Genotyping technologies for genetic research. *Annu Rev Genomics Hum Genet* **10**, 117-33 (2009).

74. Marchini, J. & Howie, B. Genotype imputation for genome-wide association studies. *Nat Rev Genet* **11**, 499-511 (2010).

75. Pearson, T.A. & Manolio, T.A. How to interpret a genome-wide association study. *JAMA* **299**, 1335-44 (2008).

76. Yang, Q., Cui, J., Chazaro, I., Cupples, L.A. & Demissie, S. Power and type I error rate of false discovery rate approaches in genome-wide association studies. *BMC Genet* **6 Suppl 1**, S134 (2005).

77. Manolio, T.A. Genomewide association studies and assessment of the risk of disease. *N Engl J Med* **363**, 166-76 (2010).

78. Yamazaki, K. *et al*. Single nucleotide polymorphisms in TNFSF15 confer susceptibility to Crohn's disease. *Hum Mol Genet* **14**, 3499-506 (2005).

79. Tremelling, M. *et al*. Contribution of TNFSF15 gene variants to Crohn's disease susceptibility confirmed in UK population. *Inflamm Bowel Dis* **14**, 733-7 (2008).

80. Kakuta, Y. *et al*. TNFSF15 transcripts from risk haplotype for Crohn's disease are overexpressed in stimulated T cells. *Hum Mol Genet* **18**, 1089-98 (2009).

81. Hindorff, L.A. A catalog of published genome-wide association studies. Vol. 2012 (2012).

82. Franke, A. *et al*. Sequence variants in IL10, ARPC2 and multiple other loci contribute to ulcerative colitis susceptibility. *Nat Genet* **40**, 1319-23 (2008).

83. Silverberg, M.S. *et al*. Ulcerative colitis-risk loci on chromosomes 1p36 and 12q15 found by genome-wide association study. *Nat Genet* **41**, 216-20 (2009).

84. Consortium, U.I.G. *et al*. Genome-wide association study of ulcerative colitis identifies three new susceptibility loci, including the HNF4A region. *Nat Genet* **41**, 1330-4 (2009).

85. Asano, K. *et al*. A genome-wide association study identifies three new susceptibility loci for ulcerative colitis in the Japanese population. *Nat Genet* **41**, 1325-9 (2009).

86. McGovern, D.P. *et al*. Genome-wide association identifies multiple ulcerative colitis susceptibility loci. *Nat Genet* **42**, 332-7 (2010).

87. Franke, A. *et al*. Genome-wide association study for ulcerative colitis identifies risk loci at 7q22 and 22q13 (IL17REL). *Nat Genet* **42**, 292-4 (2010).

88. Libioulle, C. *et al*. Novel Crohn disease locus identified by genome-wide association maps to a gene desert on 5p13.1 and modulates expression of PTGER4. *PLoS Genet* **3**, e58 (2007).

89. Rioux, J.D. *et al*. Genome-wide association study identifies new susceptibility loci for Crohn disease and implicates autophagy in disease pathogenesis. *Nat Genet* **39**, 596-604 (2007).

90. Parkes, M. *et al*. Sequence variants in the autophagy gene IRGM and multiple other replicating loci contribute to Crohn's disease susceptibility. *Nat Genet* **39**, 830-2 (2007).

91. Wellcome Trust Case Control, C. Genome-wide association study of 14,000 cases of seven common diseases and 3,000 shared controls. *Nature* **447**, 661-78 (2007).

92. Franke, A. *et al*. Systematic association mapping identifies NELL1 as a novel IBD disease gene. *PLoS One* **2**, e691 (2007).

93. Raelson, J.V. *et al.* Genome-wide association study for Crohn's disease in the Quebec Founder Population identifies multiple validated disease loci. *Proc Natl Acad Sci U S A* **104**, 14747-52 (2007).

94. Barrett, J.C. *et al.* Genome-wide association defines more than 30 distinct susceptibility loci for Crohn's disease. *Nat Genet* **40**, 955-62 (2008).

95. Franke, A. *et al.* Genome-wide association analysis in sarcoidosis and Crohn's disease unravels a common susceptibility locus on 10p12.2. *Gastroenterology* **135**, 1207-15 (2008).

96. McGovern, D.P. *et al.* Fucosyltransferase 2 (FUT2) non-secretor status is associated with Crohn's disease. *Hum Mol Genet* **19**, 3468-76 (2010).

97. Hampe, J. *et al.* A genome-wide association scan of nonsynonymous SNPs identifies a susceptibility variant for Crohn disease in ATG16L1. *Nat Genet* **39**, 207-11 (2007).

98. Duerr, R.H. *et al.* A genome-wide association study identifies IL23R as an inflammatory bowel disease gene. *Science* **314**, 1461-3 (2006).

99. Kugathasan, S. *et al.* Loci on 20q13 and 21q22 are associated with pediatric-onset inflammatory bowel disease. *Nat Genet* **40**, 1211-5 (2008).

100. Imielinski, M. *et al.* Common variants at five new loci associated with early-onset inflammatory bowel disease. *Nat Genet* **41**, 1335-40 (2009).

101. Okada, Y. *et al.* HLA-Cw*1202-B*5201-DRB1*1502 haplotype increases risk for ulcerative colitis but reduces risk for Crohn's disease. *Gastroenterology* **141**, 864-871 e1-5 (2011).

102. International Inflammatory Bowel Disease Genetics Consortium. (2001).

103. Franke, A. *et al.* Genome-wide meta-analysis increases to 71 the number of confirmed Crohn's disease susceptibility loci. *Nat Genet* **42**, 1118-25 (2010).

104. Anderson, C.A. *et al.* Meta-analysis identifies 29 additional ulcerative colitis risk loci, increasing the number of confirmed associations to 47. *Nat Genet* **43**, 246-52 (2011).

105. Nicolae, D.L. *et al.* Trait-associated SNPs are more likely to be eQTLs: annotation to enhance discovery from GWAS. *PLoS Genet* **6**, e1000888 (2010).

106. Ballard, D., Abraham, C., Cho, J. & Zhao, H. Pathway analysis comparison using Crohn's disease genome wide association studies. *BMC Med Genomics* **3**, 25 (2010).

107. Brand, S. Crohn's disease: Th1, Th17 or both? The change of a paradigm: new immunological and genetic insights implicate Th17 cells in the pathogenesis of Crohn's disease. *Gut* **58**, 1152-67 (2009).

108. Glocker, E.O. *et al*. Inflammatory bowel disease and mutations affecting the interleukin-10 receptor. *N Engl J Med* **361**, 2033-45 (2009).

109. Cooney, R. *et al*. NOD2 stimulation induces autophagy in dendritic cells influencing bacterial handling and antigen presentation. *Nat Med* **16**, 90-7 (2010).

110. Thompson, A.I. & Lees, C.W. Genetics of ulcerative colitis. *Inflamm Bowel Dis* **17**, 831-48 (2011).

111. Khalil, A.M. *et al*. Many human large intergenic noncoding RNAs associate with chromatin-modifying complexes and affect gene expression. *Proc Natl Acad Sci U S A* **106**, 11667-72 (2009).

112. Park, J.H. *et al*. Estimation of effect size distribution from genome-wide association studies and implications for future discoveries. *Nat Genet* **42**, 570-5 (2010).

113. Botstein, D. & Risch, N. Discovering genotypes underlying human phenotypes: past successes for mendelian disease, future approaches for complex disease. *Nat Genet* **33 Suppl**, 228-37 (2003).

114. Manolio, T.A. *et al*. Finding the missing heritability of complex diseases. *Nature* **461**, 747-53 (2009).

115. Cirulli, E.T. & Goldstein, D.B. Uncovering the roles of rare variants in common disease through whole-genome sequencing. *Nat Rev Genet* **11**, 415-25 (2010).

116. Visscher, P.M., Hill, W.G. & Wray, N.R. Heritability in the genomics era--concepts and misconceptions. *Nat Rev Genet* **9**, 255-66 (2008).

117. Eichler, E.E. *et al*. Missing heritability and strategies for finding the underlying causes of complex disease. *Nat Rev Genet* **11**, 446-50 (2010).

118. Fransen, K., Mitrovic, M., van Diemen, C.C. & Weersma, R.K. The quest for genetic risk factors for Crohn's disease in the post-GWAS era. *Genome Med* **3**, 13 (2011).

119. Zuk, O., Hechter, E., Sunyaev, S.R. & Lander, E.S. The mystery of missing heritability: Genetic interactions create phantom heritability. *Proc Natl Acad Sci U S A* **109**, 1193-8 (2012).

120. Edwards, A.O. *et al*. Complement factor H polymorphism and age-related macular degeneration. *Science* **308**, 421-4 (2005).

121. Gregersen, P.K. & Olsson, L.M. Recent advances in the genetics of autoimmune disease. *Annu Rev Immunol* **27**, 363-91 (2009).

122. Zhernakova, A., van Diemen, C.C. & Wijmenga, C. Detecting shared pathogenesis from the shared genetics of immune-related diseases. *Nat Rev Genet* **10**, 43-55 (2009).

123. Janse, M. *et al*. Three ulcerative colitis susceptibility loci are associated with primary sclerosing cholangitis and indicate a role for IL2, REL, and CARD9. *Hepatology* **53**, 1977-85 (2011).

124. Lees, C.W., Barrett, J.C., Parkes, M. & Satsangi, J. New IBD genetics: common pathways with other diseases. *Gut* **60**, 1739-53 (2011).

125. Ramos, P.S. *et al*. A comprehensive analysis of shared loci between systemic lupus erythematosus (SLE) and sixteen autoimmune diseases reveals limited genetic overlap. *PLoS Genet* **7**, e1002406 (2011).

126. Wang, K. *et al*. Comparative genetic analysis of inflammatory bowel disease and type 1 diabetes implicates multiple loci with opposite effects. *Hum Mol Genet* **19**, 2059-67 (2010).

127. International Consortium for Systemic Lupus Erythematosus, G. *et al*. Genome-wide association scan in women with systemic lupus erythematosus identifies susceptibility variants in ITGAM, PXK, KIAA1542 and other loci. *Nat Genet* **40**, 204-10 (2008).

128. Bernstein, C.N., Wajda, A. & Blanchard, J.F. The clustering of other chronic inflammatory diseases in inflammatory bowel disease: a population-based study. *Gastroenterology* **129**, 827-36 (2005).

129. Cohen, R. *et al*. Autoimmune disease concomitance among inflammatory bowel disease patients in the United States, 2001-2002. *Inflamm Bowel Dis* **14**, 738-43 (2008).

130. Freedman, M.L. *et al*. Principles for the post-GWAS functional characterization of cancer risk loci. *Nat Genet* **43**, 513-8 (2011).

131. Via, M., Gignoux, C. & Burchard, E.G. The 1000 Genomes Project: new opportunities for research and social challenges. *Genome Med* **2**, 3 (2010).

132. Genomes Project, C. A map of human genome variation from population-scale sequencing. *Nature* **467**, 1061-73 (2010).

133. Rivas, M.A. *et al*. Deep resequencing of GWAS loci identifies independent rare variants associated with inflammatory bowel disease. *Nat Genet* **43**, 1066-73 (2011).

134. Momozawa, Y. *et al*. Resequencing of positional candidates identifies low frequency IL23R coding variants protecting against inflammatory bowel disease. *Nat Genet* **43**, 43-7 (2011).

135. Di Meglio, P. *et al*. The IL23R R381Q gene variant protects against immune-mediated diseases by impairing IL-23-induced Th17 effector response in humans. *PLoS One* **6**, e17160 (2011).

136. Saccone, N.L. *et al*. In search of causal variants: refining disease association signals using cross-population contrasts. *BMC Genet* **9**, 58 (2008).

137. Zheng, W. *et al*. Genome-wide association study identifies a new breast cancer susceptibility locus at 6q25.1. *Nat Genet* **41**, 324-8 (2009).

138. Yasuda, K. *et al*. Variants in KCNQ1 are associated with susceptibility to type 2 diabetes mellitus. *Nat Genet* **40**, 1092-7 (2008).

139. Yamazaki, K., Takazoe, M., Tanaka, T., Kazumori, T. & Nakamura, Y. Absence of mutation in the NOD2/CARD15 gene among 483 Japanese patients with Crohn's disease. *J Hum Genet* **47**, 469-72 (2002).

140. Yang, S.K. *et al*. Contribution of IL23R but not ATG16L1 to Crohn's disease susceptibility in Koreans. *Inflamm Bowel Dis* **15**, 1385-90 (2009).

141. Wang, M.H. *et al*. Contribution of higher risk genes and European admixture to Crohn's disease in African Americans. *Inflamm Bowel Dis* (2012).

142. Cho, J.H. & Brant, S.R. Recent insights into the genetics of inflammatory bowel disease. *Gastroenterology* **140**, 1704-12 (2011).

143. Kenny, E.E. *et al*. A genome-wide scan of ashkenazi jewish Crohn's disease suggests novel susceptibility Loci. *PLoS Genet* **8**, e1002559 (2012).

144. Saunders, M.A., Liang, H. & Li, W.H. Human polymorphism at microRNAs and microRNA target sites. *Proc Natl Acad Sci U S A* **104**, 3300-5 (2007).

145. Labbe, C. *et al*. MAST3: a novel IBD risk factor that modulates TLR4 signaling. *Genes Immun* **9**, 602-12 (2008).

146. Labbe, C. *et al*. Genome-wide expression profiling implicates a MAST3-regulated gene set in colonic mucosal inflammation of ulcerative colitis patients. *Inflamm Bowel Dis* (2011).

147. Fransen, K. *et al*. Analysis of SNPs with an effect on gene expression identifies UBE2L3 and BCL3 as potential new risk genes for Crohn's disease. *Hum Mol Genet* **19**, 3482-8 (2010).

148. Fu, J. *et al*. Unraveling the regulatory mechanisms underlying tissue-dependent genetic variation of gene expression. *PLoS Genet* **8**, e1002431 (2012).

149. Zwiers, A. *et al*. Cutting edge: a variant of the IL-23R gene associated with inflammatory bowel disease induces loss of microRNA regulation and enhanced protein production. *J Immunol* **188**, 1573-7 (2012).

150. McCarroll, S.A. *et al*. Deletion polymorphism upstream of IRGM associated with altered IRGM expression and Crohn's disease. *Nat Genet* **40**, 1107-12 (2008).

151. Prescott, N.J. *et al*. Independent and population-specific association of risk variants at the IRGM locus with Crohn's disease. *Hum Mol Genet* **19**, 1828-39 (2010).

152. Brest, P. *et al*. A synonymous variant in IRGM alters a binding site for miR-196 and causes deregulation of IRGM-dependent xenophagy in Crohn's disease. *Nat Genet* **43**, 242-5 (2011).

153. Brest, P., Lapaquette, P., Mograbi, B., Darfeuille-Michaud, A. & Hofman, P. Risk predisposition for Crohn disease: a "menage a trois" combining IRGM allele, miRNA and xenophagy. *Autophagy* **7**, 786-7 (2011).

154. Wu, F. *et al*. Peripheral blood microRNAs distinguish active ulcerative colitis and Crohn's disease. *Inflamm Bowel Dis* **17**, 241-50 (2011).

155. Paraskevi, A. *et al*. Circulating MicroRNA in inflammatory bowel disease. *J Crohns Colitis* (2012).

156. Duttagupta, R. *et al*. Genome-Wide Maps of Circulating miRNA Biomarkers for Ulcerative Colitis. *PLoS One* **7**, e31241 (2012).

157. Feng, R., Wu, Y., Jang, G.H., Ordovas, J.M. & Arnett, D. A powerful test of parent-of-origin effects for quantitative traits using haplotypes. *PLoS One* **6**, e28909 (2011).

158. Akolkar, P.N. *et al*. Differences in risk of Crohn's disease in offspring of mothers and fathers with inflammatory bowel disease. *Am J Gastroenterol* **92**, 2241-4 (1997).

159. Zelinkova, Z. *et al*. Maternal imprinting and female predominance in familial Crohn's disease. *J Crohns Colitis* (2012).

160. Reik, W. & Walter, J. Genomic imprinting: parental influence on the genome. *Nat Rev Genet* **2**, 21-32 (2001).

161. Glaser, R.L., Ramsay, J.P. & Morison, I.M. The imprinted gene and parent-of-origin effect database now includes parental origin of de novo mutations. *Nucleic Acids Res* **34**, D29-31 (2006).

162. Rutgeerts, P. *et al*. Comparison of scheduled and episodic treatment strategies of infliximab in Crohn's disease. *Gastroenterology* **126**, 402-13 (2004).

163. Sandborn, W.J. *et al*. Adalimumab induction therapy for Crohn disease previously treated with infliximab: a randomized trial. *Ann Intern Med* **146**, 829-38 (2007).

164. Anderson, C.A. *et al*. Data quality control in genetic case-control association studies. *Nat Protoc* **5**, 1564-73 (2010).

165. Colombel, J.F. *et al*. Adalimumab for maintenance of clinical response and remission in patients with Crohn's disease: the CHARM trial. *Gastroenterology* **132**, 52-65 (2007).

166. Goldstein, B.A., Hubbard, A.E., Cutler, A. & Barcellos, L.F. An application of Random Forests to a genome-wide association dataset: methodological considerations & new findings. *BMC Genet* **11**, 49 (2010).

167. Podolsky, D.K. Inflammatory bowel disease. *N Engl J Med* **347**, 417-29 (2002).

168. Sandborn, W.J. *et al.* Adalimumab for maintenance treatment of Crohn's disease: results of the CLASSIC II trial. *Gut* **56**, 1232-9 (2007).

169. Purcell, S. *et al.* PLINK: a tool set for whole-genome association and population-based linkage analyses. *Am J Hum Genet* **81**, 559-75 (2007).

170. Wider, C. *et al.* Phactr2 and Parkinson's disease. *Neurosci Lett* **453**, 9-11 (2009).

171. Jiao, Y.L. *et al.* Polymorphisms of KIR gene and HLA-C alleles: possible association with susceptibility to HLA-B27-positive patients with ankylosing spondylitis. *J Clin Immunol* **30**, 840-4 (2010).

172. Kuhn, M. Building predictive models in R using the caret package. *Journal of Statistical Software* **28**, 1-26 (2008).

173. Sanseau, P. *et al.* Use of genome-wide association studies for drug repositioning. *Nat Biotechnol* **30**, 317-20 (2012).

174. Amre, D.K. *et al.* Investigation of reported associations between the 20q13 and 21q22 loci and pediatric-onset Crohn's disease in Canadian children. *Am J Gastroenterol* **104**, 2824-8 (2009).

175. Jacob, C.O. *et al.* Identification of novel susceptibility genes in childhood-onset systemic lupus erythematosus using a uniquely designed candidate gene pathway platform. *Arthritis Rheum* **56**, 4164-73 (2007).

176. Perdigones, N. *et al.* Evidence of epistasis between TNFRSF14 and TNFRSF6B polymorphisms in patients with rheumatoid arthritis. *Arthritis Rheum* **62**, 705-10 (2010).

177. Bailey, S.D. *et al.* Variation at the NFATC2 locus increases the risk of thiazolidinedione-induced edema in the Diabetes REduction Assessment with ramipril and rosiglitazone Medication (DREAM) study. *Diabetes Care* **33**, 2250-3 (2010).

178. Lorentzen, A.R. *et al.* Killer immunoglobulin-like receptor ligand HLA-Bw4 protects against multiple sclerosis. *Ann Neurol* **65**, 658-66 (2009).

179. Vidal-Castineira, J.R. *et al.* Effect of killer immunoglobulin-like receptors in the response to combined treatment in patients with chronic hepatitis C virus infection. *J Virol* **84**, 475-81 (2010).

180. Koh, E., Devendra, K. & Tan, L.K. B-Lynch suture for the treatment of uterine atony. *Singapore Med J* **50**, 693-7 (2009).

181. Satyanarayana, C.R. *et al.* Genetic variations and haplotypes of the 5' regulatory region of CYP2C19 in South Indian population. *Drug Metab Pharmacokinet* **24**, 185-93 (2009).

182. Bering, B. & Devendra, D. Latent autoimmune diabetes in the young. *Clin Med* **9**, 93; author reply 93-4 (2009).

183. Skidmore, S. *et al.* A case study of delayed HIV-1 seroconversion highlights the need for Combo assays. *Int J STD AIDS* **20**, 205-6 (2009).

184. Hassanein, M., Bravis, V., Hui, E. & Devendra, D. Ramadan-focused education and awareness in type 2 diabetes. *Diabetologia* **52**, 367-8 (2009).

185. Yazbek, S.N., Spiezio, S.H., Nadeau, J.H. & Buchner, D.A. Ancestral paternal genotype controls body weight and food intake for multiple generations. *Hum Mol Genet* **19**, 4134-44 (2010).

186. Thompson, S.L. *et al.* Environmental effects on genomic imprinting in mammals. *Toxicol Lett* **120**, 143-50 (2001).

187. Wang, S., Yu, Z., Miller, R.L., Tang, D. & Perera, F.P. Methods for detecting interactions between imprinted genes and environmental exposures using birth cohort designs with mother-offspring pairs. *Hum Hered* **71**, 196-208 (2011).

188. Dieker, J.W. *et al.* Apoptosis-induced acetylation of histones is pathogenic in systemic lupus erythematosus. *Arthritis Rheum* **56**, 1921-33 (2007).

189. Jacobi, A.M. *et al.* Differential effects of epratuzumab on peripheral blood B cells of patients with systemic lupus erythematosus versus normal controls. *Ann Rheum Dis* **67**, 450-7 (2008).

190. Finke, D. *et al.* Elevated levels of endogenous apoptotic DNA and IFN-alpha in complement C4-deficient mice: implications for induction of systemic lupus erythematosus. *Eur J Immunol* **37**, 1702-9 (2007).

191. Strombeck, B. & Jacobsson, L.T. The role of exercise in the rehabilitation of patients with systemic lupus erythematosus and patients with primary Sjogren's syndrome. *Curr Opin Rheumatol* **19**, 197-203 (2007).

192. Perdigones, N. *et al.* Study of chromosomal region 5p13.1 in Crohn's disease, ulcerative colitis, and rheumatoid arthritis. *Hum Immunol* **71**, 826-8 (2010).

6 PRILOGE

6.1 Kazalo preglednic

6.2 Kazalo slik

6.3 Seznam krajšav

APC	antigen predstavitvene celice
AMD	starostna degeneracija očesne mrežnice
AUC	površina pod krivuljo
bp	bazni par
CEPH	Center za študijo polimorfizmov v človeškem genomu
CB	Crohnova bolezen
CR	angl. call rate
cDNK	komplementarni DNK
CGAS	asociacijska študija kandidatnih genov
cM	centimorgan, enota za merjenje genske vezanosti
CNV	različica v številu kopij
DNK	deoksiribonukleinska kislina
DC	dendritska celica
dNTP	deoksinukleozid trifosfat
EDTA	etilendiamin tetraocetna kislina
ER	endoplazemski retikulum
eQTL	ekspresijski kvantitativni lokus
GWAS	asociacijska študija na celotnem genomu
GWLs	študija genske povezanosti na celotnem genomu
HCl	klorovodikova kislina
HLA	humani levkocitni antigen
HRMA	analiza talilnih krivulj visoke ločljivosti
iCHIP	ImmunoChip, tarčna DNK-mikromreža
IBS	enakost po stanju
IBD	enakost po izvoru
IFN	interferon
Ig	imunoglobulin
IK	intermediarni kolitis
IL	interlevkin
KCl	kalijev klorid
KK	kontrola kakovosti
KVČB	kronična vnetna črevesna bolezen
LD	vezno neravnotežje

LOD	logaritem obetov
MAF	frekvenca najmanj pogostega alela
MDS	večrazsežno skaliranje
MDP	muramil dipeptid
MHC	poglavitni histokompatibilnostni kompleks
$MgCl_2$	magnezijev klorid
mRNK	informacijski RNK
miRNK	mikro RNK
NaCl	natrijev klorid
ncRNK	nekodirajoči RNK
NFKB	nuklearni dejavnik kappa B
NK	naravne celice ubijalke
OR	razmerje obetov
PCR	verižna reakcija s polimerazo
POO	učinek starševskega izvora
QQ	kvantil-kvantil
RE	restrikcijski encim
RFLP	polimorfizem dolžin restrikcijskih fragmentov
RNK	ribonukleinska kislina
SDS	natrijev dodecilsulfat
SNP	polimorfizem posameznega nukleotida
SVM	metoda podpornih vektorjev
Th	T-celica pomagalka
TNF	dejavnik tumorske nekroze
Treg	regulatorna celica T
TRIS	2-amino-2-hidroksimetil-1,3-propandiol
UK	ulcerozni kolitis
UTR	neprevedljiva regija

6.4 Seznam simbolov

Oznaka	Veličina	Enota
M	molska masa	g/mol
a	dolžina	cm
T	temperatura	°C
m	masa	g; mg
V	volumen	L; mL; µL
γ	masna koncentracija	mg/mL; µg/µL
c	množinska koncentracija	mol/L; mM
f	število obratov	vrt./min; g
w	masni delež	%
t	čas	s; min; h
U	napetost	V
ρ	gostota	g/L
θ	količnik rekombinacije	/
λ_s	valovna dolžina	nm
λ	valovna dolžina	nm

Uporabljeni simboli so v skladu z mednarodnim standardom ISO/IEC 80000.

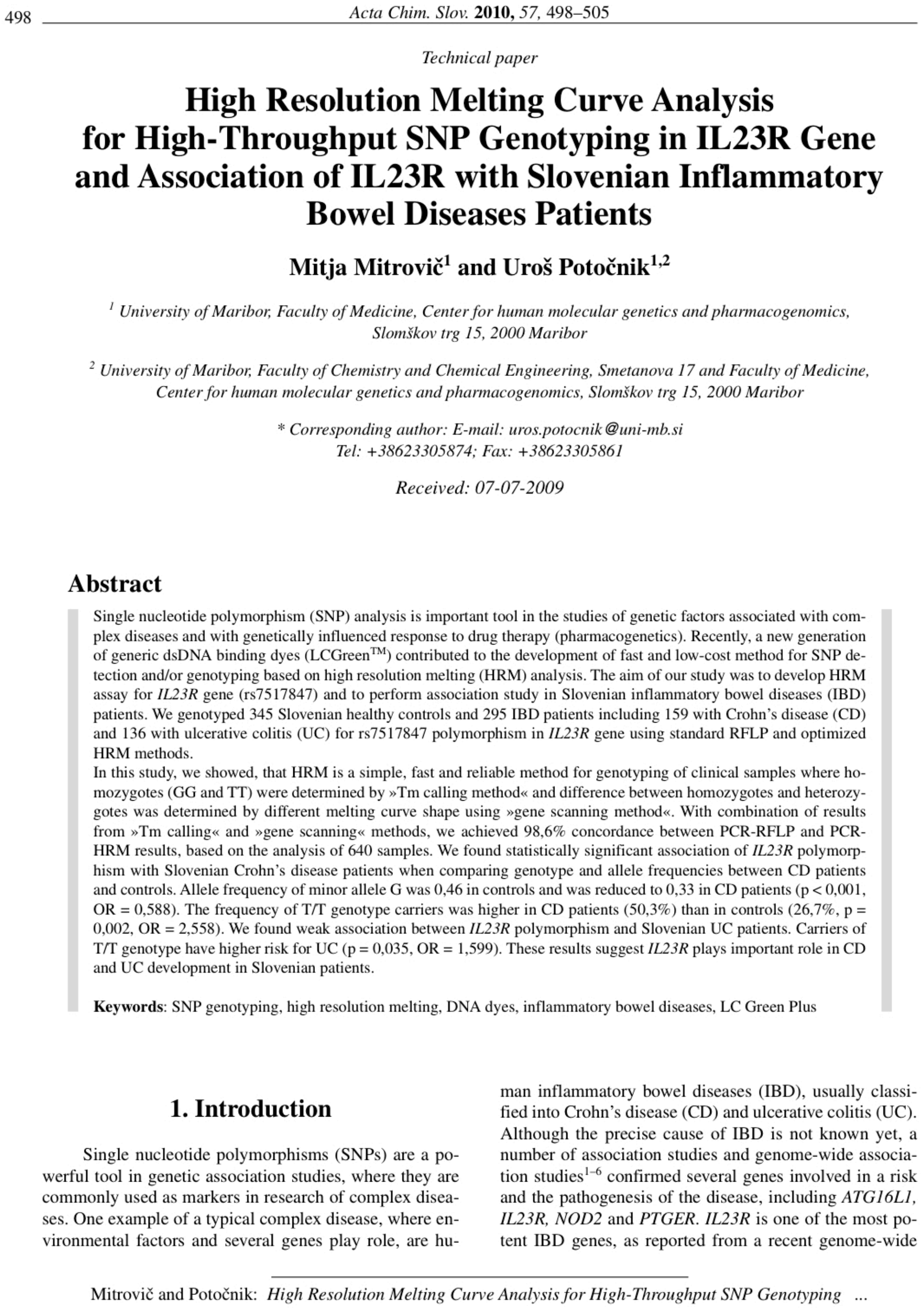
6.5 Seznam in krajšave aminokislin

Ime	Enočrkovna oznaka	Tričrkovna oznaka
Alanin	A	Ala
Cistein	C	Cys
Asparaginska kislina	D	Asp
Glutaminska kislina	E	Glu
Fenilalanin	F	Phe
Glicin	G	Gly
Histidin	H	His
Izolevcin	I	Ile
Lizin	K	Lys
Levcin	L	Leu
Metionin	M	Met
Asparagin	N	Asn
Prolin	P	Pro
Glutamin	Q	Gln
Arginin	R	Arg
Serin	S	Ser
Treonin	T	Thr
Valin	V	Val
Triptofan	W	Trp
Tirozin	Y	Tyr

6.6 Članki kot sestavni del doktorske naloge

1. Mitrovic, M., & Potočnik, U. High resolution melting curve analysis forhigh-throughput SNP genotyping in IL23R gene and association of IL23R with Slovenian inflammatory bowel disease patients. *Acta Chim Slov* **57**, 498−505 (2010).
2. Mitrovic, M., & Potočnik, U. High-resolution melting curve analysis for high throughput genotyping of *NOD2/CARD15* mutations and distribution of these mutations in Slovenian inflammatory bowel diseases patients. *Dis Markers* **30**, 265−74 (2011).
3. Fransen, K., Mitrovic, M., van Diemen, C. C., & Weersma, R. K. The quest for genetic risk factors for Crohn's disease in the post-GWAS era. *Genome Med* **25**,13−23 (2011).
4. Fransen, K., Mitrovic, M, van Diemen, C. C., Thelma, B. K., Sood, A., Franke, A., Schreiber, S., Midha, V., Juyal, G., Potocnik, U., Fu, J., Nolte, I., & Weersma, R. K. Limited Evidence for Parent-of-Origin Effects in Inflammatory Bowel Disease Associated Loci. *PLoS One*. 7(9):e45287 (2012).
5. Jostins, L., et *al*. Host–microbe interactions have shaped the genetic architecture of inflammatory bowel disease. *Nature* **491**, 119–124 (2012).
6. Mitrovic, M., Stiglic, G., Weersma, R. K., & Potocnik, U. Support vector machine model for identification of Crohn’s disease patients requiring biological therapy. Poslano v revizijo k *Genome Medicine*.

6.6.1 Izvirni znanstveni članek 1



498 Acta Chim. Slov. **2010,** *57,* 498–505

Technical paper

High Resolution Melting Curve Analysis for High-Throughput SNP Genotyping in IL23R Gene and Association of IL23R with Slovenian Inflammatory Bowel Diseases Patients

Mitja Mitrovič[1] and Uroš Potočnik[1,2]

[1] *University of Maribor, Faculty of Medicine, Center for human molecular genetics and pharmacogenomics, Slomškov trg 15, 2000 Maribor*

[2] *University of Maribor, Faculty of Chemistry and Chemical Engineering, Smetanova 17 and Faculty of Medicine, Center for human molecular genetics and pharmacogenomics, Slomškov trg 15, 2000 Maribor*

** Corresponding author: E-mail: uros.potocnik@uni-mb.si*
Tel: +38623305874; Fax: +38623305861

Received: 07-07-2009

Abstract

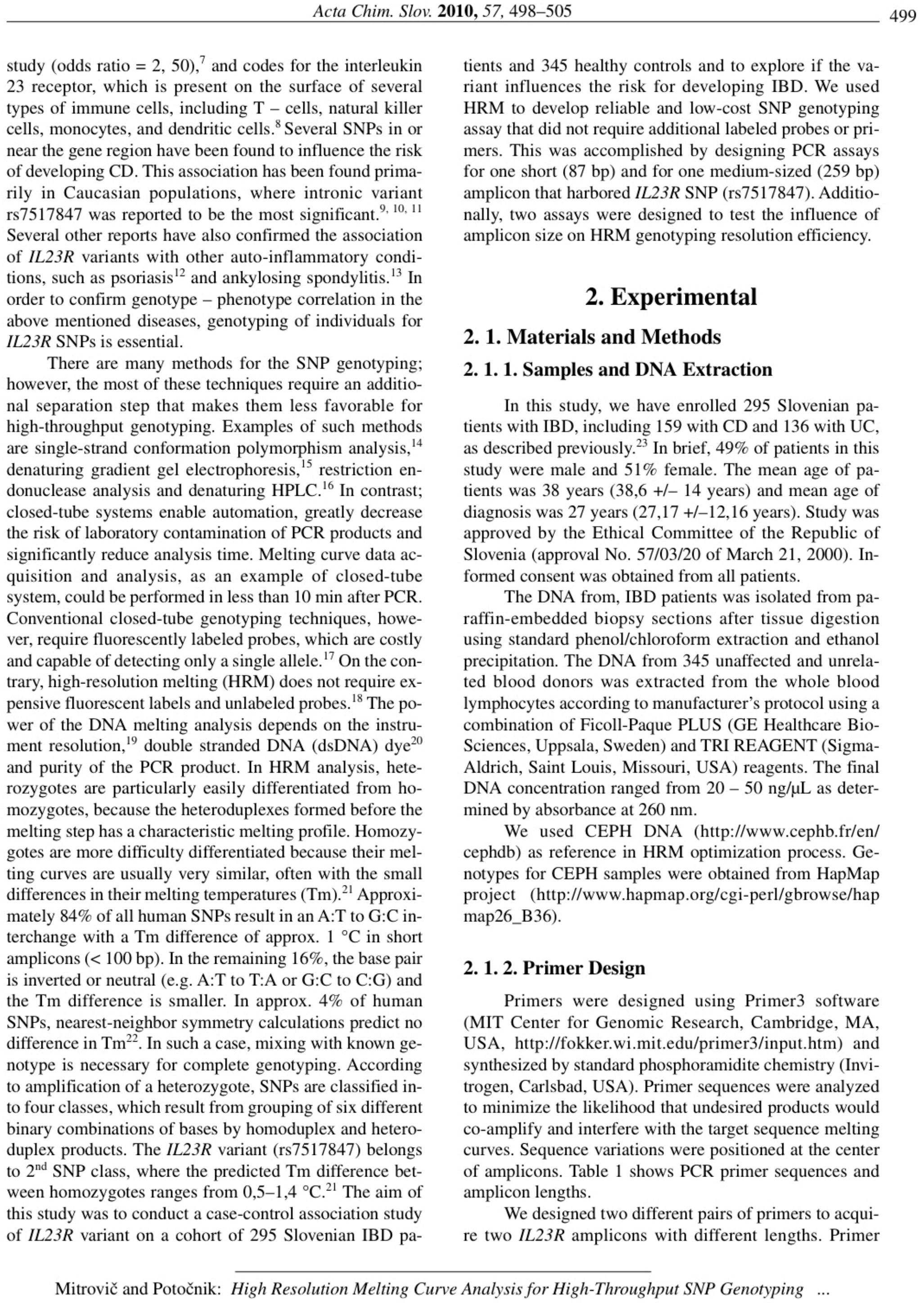
Single nucleotide polymorphism (SNP) analysis is important tool in the studies of genetic factors associated with complex diseases and with genetically influenced response to drug therapy (pharmacogenetics). Recently, a new generation of generic dsDNA binding dyes (LCGreen^{TM}) contributed to the development of fast and low-cost method for SNP detection and/or genotyping based on high resolution melting (HRM) analysis. The aim of our study was to develop HRM assay for *IL23R* gene (rs7517847) and to perform association study in Slovenian inflammatory bowel diseases (IBD) patients. We genotyped 345 Slovenian healthy controls and 295 IBD patients including 159 with Crohn's disease (CD) and 136 with ulcerative colitis (UC) for rs7517847 polymorphism in *IL23R* gene using standard RFLP and optimized HRM methods.

In this study, we showed, that HRM is a simple, fast and reliable method for genotyping of clinical samples where homozygotes (GG and TT) were determined by »Tm calling method« and difference between homozygotes and heterozygotes was determined by different melting curve shape using »gene scanning method«. With combination of results from »Tm calling« and »gene scanning« methods, we achieved 98,6% concordance between PCR-RFLP and PCR-HRM results, based on the analysis of 640 samples. We found statistically significant association of *IL23R* polymorphism with Slovenian Crohn's disease patients when comparing genotype and allele frequencies between CD patients and controls. Allele frequency of minor allele G was 0,46 in controls and was reduced to 0,33 in CD patients ($p < 0,001$, OR = 0,588). The frequency of T/T genotype carriers was higher in CD patients (50,3%) than in controls (26,7%, p = 0,002, OR = 2,558). We found weak association between *IL23R* polymorphism and Slovenian UC patients. Carriers of T/T genotype have higher risk for UC (p = 0,035, OR = 1,599). These results suggest *IL23R* plays important role in CD and UC development in Slovenian patients.

Keywords: SNP genotyping, high resolution melting, DNA dyes, inflammatory bowel diseases, LC Green Plus

1. Introduction

Single nucleotide polymorphisms (SNPs) are a powerful tool in genetic association studies, where they are commonly used as markers in research of complex diseases. One example of a typical complex disease, where environmental factors and several genes play role, are human inflammatory bowel diseases (IBD), usually classified into Crohn's disease (CD) and ulcerative colitis (UC). Although the precise cause of IBD is not known yet, a number of association studies and genome-wide association studies[1–6] confirmed several genes involved in a risk and the pathogenesis of the disease, including *ATG16L1, IL23R, NOD2* and *PTGER. IL23R* is one of the most potent IBD genes, as reported from a recent genome-wide



study (odds ratio = 2, 50),[7] and codes for the interleukin 23 receptor, which is present on the surface of several types of immune cells, including T – cells, natural killer cells, monocytes, and dendritic cells.[8] Several SNPs in or near the gene region have been found to influence the risk of developing CD. This association has been found primarily in Caucasian populations, where intronic variant rs7517847 was reported to be the most significant.[9, 10, 11] Several other reports have also confirmed the association of *IL23R* variants with other auto-inflammatory conditions, such as psoriasis[12] and ankylosing spondylitis.[13] In order to confirm genotype – phenotype correlation in the above mentioned diseases, genotyping of individuals for *IL23R* SNPs is essential.

There are many methods for the SNP genotyping; however, the most of these techniques require an additional separation step that makes them less favorable for high-throughput genotyping. Examples of such methods are single-strand conformation polymorphism analysis,[14] denaturing gradient gel electrophoresis,[15] restriction endonuclease analysis and denaturing HPLC.[16] In contrast; closed-tube systems enable automation, greatly decrease the risk of laboratory contamination of PCR products and significantly reduce analysis time. Melting curve data acquisition and analysis, as an example of closed-tube system, could be performed in less than 10 min after PCR. Conventional closed-tube genotyping techniques, however, require fluorescently labeled probes, which are costly and capable of detecting only a single allele.[17] On the contrary, high-resolution melting (HRM) does not require expensive fluorescent labels and unlabeled probes.[18] The power of the DNA melting analysis depends on the instrument resolution,[19] double stranded DNA (dsDNA) dye[20] and purity of the PCR product. In HRM analysis, heterozygotes are particularly easily differentiated from homozygotes, because the heteroduplexes formed before the melting step has a characteristic melting profile. Homozygotes are more difficulty differentiated because their melting curves are usually very similar, often with the small differences in their melting temperatures (Tm).[21] Approximately 84% of all human SNPs result in an A:T to G:C interchange with a Tm difference of approx. 1 °C in short amplicons (< 100 bp). In the remaining 16%, the base pair is inverted or neutral (e.g. A:T to T:A or G:C to C:G) and the Tm difference is smaller. In approx. 4% of human SNPs, nearest-neighbor symmetry calculations predict no difference in Tm[22]. In such a case, mixing with known genotype is necessary for complete genotyping. According to amplification of a heterozygote, SNPs are classified into four classes, which result from grouping of six different binary combinations of bases by homoduplex and heteroduplex products. The *IL23R* variant (rs7517847) belongs to 2nd SNP class, where the predicted Tm difference between homozygotes ranges from 0,5–1,4 °C.[21] The aim of this study was to conduct a case-control association study of *IL23R* variant on a cohort of 295 Slovenian IBD patients and 345 healthy controls and to explore if the variant influences the risk for developing IBD. We used HRM to develop reliable and low-cost SNP genotyping assay that did not require additional labeled probes or primers. This was accomplished by designing PCR assays for one short (87 bp) and for one medium-sized (259 bp) amplicon that harbored *IL23R* SNP (rs7517847). Additionally, two assays were designed to test the influence of amplicon size on HRM genotyping resolution efficiency.

2. Experimental

2. 1. Materials and Methods

2. 1. 1. Samples and DNA Extraction

In this study, we have enrolled 295 Slovenian patients with IBD, including 159 with CD and 136 with UC, as described previously.[23] In brief, 49% of patients in this study were male and 51% female. The mean age of patients was 38 years (38,6 +/– 14 years) and mean age of diagnosis was 27 years (27,17 +/–12,16 years). Study was approved by the Ethical Committee of the Republic of Slovenia (approval No. 57/03/20 of March 21, 2000). Informed consent was obtained from all patients.

The DNA from, IBD patients was isolated from paraffin-embedded biopsy sections after tissue digestion using standard phenol/chloroform extraction and ethanol precipitation. The DNA from 345 unaffected and unrelated blood donors was extracted from the whole blood lymphocytes according to manufacturer's protocol using a combination of Ficoll-Paque PLUS (GE Healthcare Bio-Sciences, Uppsala, Sweden) and TRI REAGENT (Sigma-Aldrich, Saint Louis, Missouri, USA) reagents. The final DNA concentration ranged from 20 – 50 ng/μL as determined by absorbance at 260 nm.

We used CEPH DNA (http://www.cephb.fr/en/cephdb) as reference in HRM optimization process. Genotypes for CEPH samples were obtained from HapMap project (http://www.hapmap.org/cgi-perl/gbrowse/hapmap26_B36).

2. 1. 2. Primer Design

Primers were designed using Primer3 software (MIT Center for Genomic Research, Cambridge, MA, USA, http://fokker.wi.mit.edu/primer3/input.htm) and synthesized by standard phosphoramidite chemistry (Invitrogen, Carlsbad, USA). Primer sequences were analyzed to minimize the likelihood that undesired products would co-amplify and interfere with the target sequence melting curves. Sequence variations were positioned at the center of amplicons. Table 1 shows PCR primer sequences and amplicon lengths.

We designed two different pairs of primers to acquire two *IL23R* amplicons with different lengths. Primer

Table 1. Primers used for real-time PCR amplification and SNP genotyping.

Gene	Sense primer 5' – > 3'	Antisense primer 5' – > 3'	Amplicon length (bp)
IL23R	CCATCTCACTGTCTCCTCTC	GGCTCCAGTTTCTAGCCTAC	87
IL23R	TCTGCCAATTCCCTAAAC	AAGTAGGTGTGGATTGCC	259

pair that resulted in 87 bp long amplicon was used in HRM. As the amplicon length increases, the difference in Tm between genotypes becomes smaller.[24] Primer pair that resulted in 259 bp long amplicon was used in RFLP. The 259 bp long amplicon was cleaved with restriction enzyme to produce fragments with different lengths needed for discrimination between genotypes. Both sets of primers were used to test the influence of amplicon size on HRM genotyping efficiency resolution.

2. 1. 3 PCR-RFLP

PCR amplification and restriction fragment length polymorphism (RFLP) were used to obtain reference genotypes from CEPH reference DNA samples and from DNA samples of patients and controls. PCR was performed in 0,2 mL PCR strip tubes on T1 Thermocycler (Biometra, Germany), using 10-μL final reaction volumes with 2 μL of template DNA (20–50 ng/μL). The reaction mixture contained 1 x PCR Buffer, 2 mM $MgCl_2$, 0,2 mM of each dNTP, 0,5 μM of each primer, 0,5 U of *Taq* Polymerase (Fermentas, Lithuania) and PCR grade water. Cycling conditions were performed using the following protocol: initial denaturation at 95 °C for 5 min, followed by 35 cycles of denaturation (95 °C, 30 s), annealing (55 °C, 30 s), extension (72 °C, 30 s) and final extension (72 °C, 5 min). After amplification the resulting PCR products were mixed with restriction enzyme *Hpy* F3I (Fermentas, Lithuania) and cleaved at 37 °C for 16 hours. Resulting fragments were separated by length on agarose gel electrophoresis.

2. 1. 4 PCR-HRM

PCR and HRM reactions were performed on a LightCycler 480 2.0 instrument (Roche Diagnostics, Indianapolis, IN, USA), using 10-μL reaction volumes with 2 μL of DNA, 1 x PCR Buffer, 3 mM $MgCl_2$, 0,2 mM of each dNTP, 0,5 μM of each HRM primer, 1 x LC Green Plus (Idaho Technology, USA) and 0,5 U of *Taq* Polymerase (Fermentas, Lithuania). Cycling conditions were performed using the following protocol: initial denaturation at 95 °C for 1 min, followed by 40 cycles of denaturation (95 °C, 1s), annealing (58 °C, 1s) and extension (72 °C, 10s). After amplification, the samples were heated to 95 °C for 1 min and rapidly cooled to 40 °C for 1 min at rate of 1 °C/s, to induce heteroduplex formation before melting. Melting curve data were obtained by continuous fluorescence acquisition from 55 to 90 °C with a thermal transition rate of 0,1 °C/s. Genotyping was based on negative first derivate melting curves and comparison of unknowns to genotyped controls. Data were analyzed using software provided with the LC480 instrument, where two methods were used to determine sample genotypes. »Gene scanning« method is based on normalization, temperature shift of fluorescence data and calculation of difference plot to discriminate between different genotypes. »Tm calling« method, on the other hand, is based on calculation of negative of the first derivative of the fluorescence data, where melting curve peaks are obtained to discriminate between different genotypes.

3. Results and Discussion

3. 1. High Resolution Melting Analysis of *IL23R*

High-resolution dye LCGreen Plus used in our study allowed the identification of both heterozygous and homozygous samples, thus making scanning and genotyping of PCR products possible. Although both, 87 bp and 259 bp amplicons were tested for HRM analysis, only genotype data from the shorter amplicon was in full coherence with the results obtained by RFLP analysis. Altogether 640 samples were genotyped by PCR-RFLP and PCR-HRM methods. Melting curve analysis of 87 bp amplicon by »gene scanning« method is shown in Figures 1A, B, C, D.

The original HRM data (Figure 1A) were normalized by defining linear baselines before and after melting transitions, which were designated values of 100% and 0%, respectively. Within each sample, the fluorescence of each acquisition was calculated as a percentage of fluorescence between the top and bottom baselines at each acquisition temperature, as depicted on Figure 1B.

Different genotypes were easily distinguished after normalization (Figure 1B). Normalized melting curves were adjusted to eliminate slight temperature and salt variation between samples by shifting each curve along the temperature axis for 5% of normalized fluorescence, as shown in Figure 1C.

Figure 1C shows, that heterozygotes (G/T) have different melting curve shapes and were unambiguously distinguished from homozygotes G/G and T/T. Note that homozygotes G/G and T/T have similar curve shape, whereas heterozygotes G/T were in between with broader melting transition. To further discriminate between genotypes, difference plots were obtained by subtracting the temperature-overlaid, normalized curves from one of the wild-type (G/G) curves, as shown in Figure 1D.

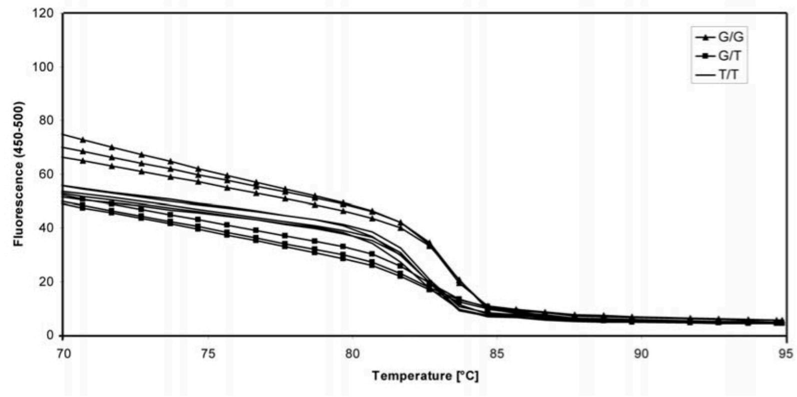


Figure 1A. Original HRM data. Melting curves include three different individuals for each of three genotypes and each individual run in triplicate.

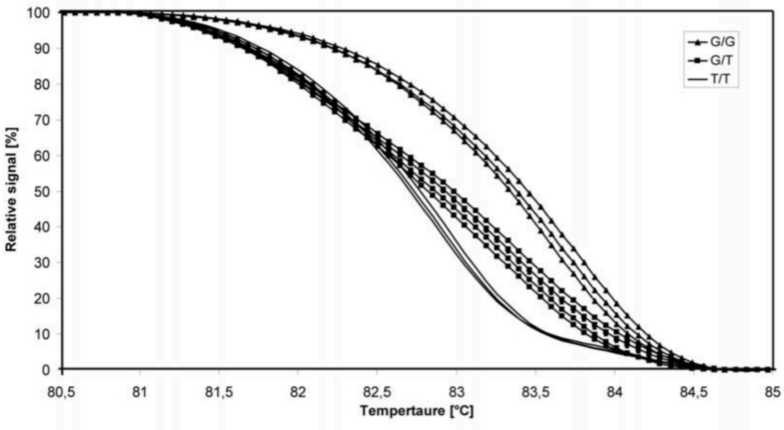


Figure 1B. Normalized HRM data.

The average peak of relative signal difference in Figure 1D for T/T homozygotes was at 83,19 °C, and 82,44 °C for G/T heterozygotes. Note that, because of the shift in the temperature of the curves in Figures 1C and D, the temperature axis no longer reflects absolute temperatures, but rather reflects temperature differences relative to superimposed segments of the curves. A total of 627 of 640 tested samples (98%) were determined correctly by HRM »gene scanning« method when compared to RFLP results. Four homozygotes G/G (0,6%), were incorrectly grouped as T/T homozygotes and were subsequently correctly determined by »Tm calling« method. The other nine samples (1,4 %) were negative and subsequent measurements on spectrophotometer showed poor A260/280 (0,56–1,14) and A260/230 (0,44–0,98) ratios, suggesting low DNA quality and the presence of small molecule contaminants in the DNA preparation.

We also performed the alternative »Tm calling« method to additionally differentiate between genotypes as shown in Figure 1E

Figure 1E shows that homozygotes G/G and T/T melted at different temperatures and could be easily distinguished from each other. The data obtained from »Tm calling« method are summarized in Table 2.

The calculation of theoretical Tm is based on nearest-neighbor thermodynamic model, but practical Tm is the one measured by the instrument and could differ from theoretical Tm due to fluctuations in reaction conditions (e.g. salt concentration), that influence melting and are not taken into account in the model calculation.

The efficiency of Tm calling method for 12 and 92 samples was calculated by comparing genotyping results with results from RFLP. The efficiency of ho-

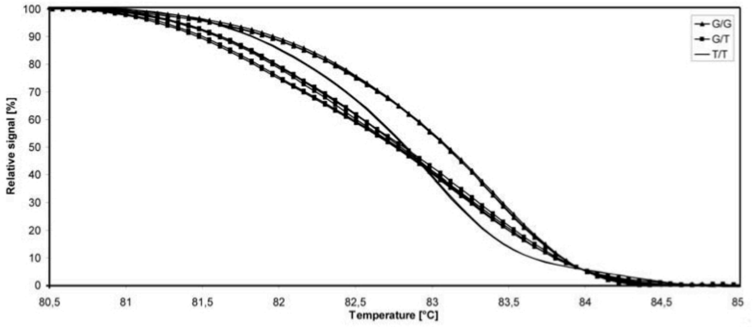


Figure 1C. Normalized and temperature shifted HRM data.

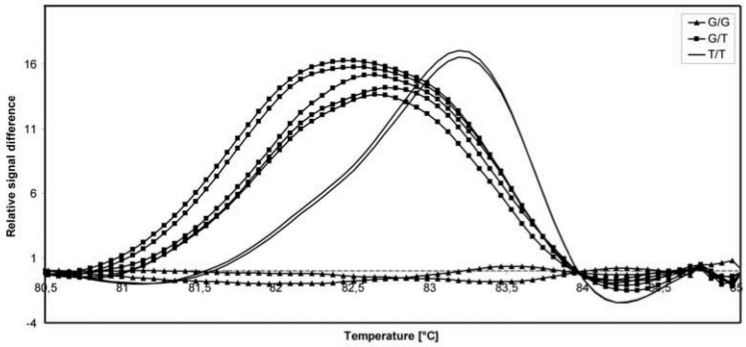


Figure 1D. Normalized and temperature shifted HRM difference plot.

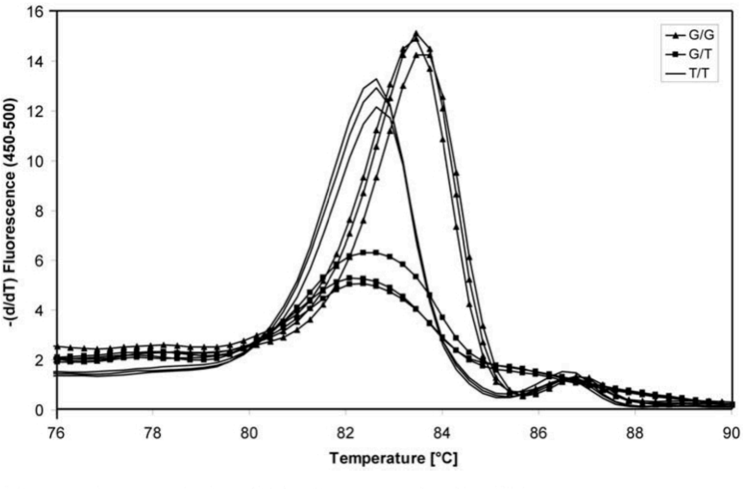


Figure 1E. Derivative melting curves of amplicon (87bp) melting for genotyping of the IL23R gene.

Table 2. Tm calling data.

No. Samples	T_m [°C] (theoretical)	ΔT_m [°C] (theoretical)	T_m [°C] (practical)	DT_m [°C] (practical)	STD	Homozygotes differentiation	Heterozygotes differentiation
12	G: 72,9 T: 72,4	0,5	G:82,90 T: 82,37	0,53	G:0,40 T: 0,18	100%	100%
92	G: 72,9 T: 72,4	0,5	G:83,07 T: 82,44	0,63	G:0,44 T: 0,33	95%	79%

*STD = standard deviation. * T_m = melting temperature.
*ΔT_m = difference in melting temperatures (e.g. between G/G and T/T genotypes).

mozygote determination by Tm calling method was 100% in our initial analysis of 12 samples, however when the sample size increased to 92 samples the efficiency decreased to 95%. The efficiency of heterozygote determination was 100% in 12 samples analyzed, yet it decreased to 79% in 92 samples analyzed. Although »Tm calling« did not prove as a very robust method, it was successfully used to complement »gene scanning« method.

With combination of results from »Tm calling« and »gene scanning« methods, we achieved 98,6% concordance between PCR-RFLP and PCR-HRM results based on the analysis of 640 samples. Using combination of both, »gene scanning« and »Tm calling« methods we were able to unambiguously distinguish different genotypes, therefore no mixing of DNA samples with known genotypes (spike-in principle) was needed to enhance the differences between the melting curves.

3. 2. RFLP Analysis

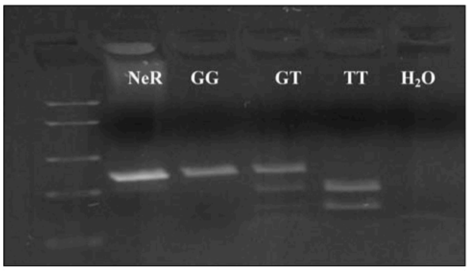
The restriction analysis of *IL23R* gene revealed 259 bases long fragment for G/G genotype, 168 and 91 bases long fragments for T/T genotype and 259, 168 and 91 bases long fragments for G/T genotype as shown on Figure 2. RFLP was used as a reference method to evaluate the accuracy of high resolution amplicon melting. 345 Slovenian healthy controls and a total of 295 IBD patients were genotyped by RFLP method.

3. 3. Association Study Analysis

We genotyped 345 Slovenian healthy controls and a total of 295 IBD patients (159 with CD and 136 UC), for intronic rs7517847 polymorphism in *IL23R* gene. We found statistically significant association of *IL23R* polymorphism with Crohn's disease patients when comparing genotype and allele frequencies between CD patients and controls (Tables 3 and 4).

The allele frequency of minor allele G was 0,46 in controls and was reduced to 0,33 in CD patients ($p < 0,001$, OR = 0,588). The frequency of protective allele G was also decreased in case of UC compared to controls, but only weak association was found (p = 0,13, OR = 1,252), suggesting a significantly less important protective effect of the minor allele for UC compared with CD.

As shown in Table 4, the frequency of T/T genotype carriers was significantly higher in CD patients (50,3%) than in controls (26,7%, p = 0,002, OR = 2,558). The frequency of T/T genotype carriers was also significantly higher in UC patients (p = 0,035, OR = 1,599) compared to healthy controls, suggesting a major effect on susceptibility to CD and a more modest effect on UC. Our results are in concordance with two other association studies, in Caucasians in Great Britain and in Ashkenazi Jews in the USA, which also confirmed the association of *IL23R* variant with IBD.[10,11] Compared to Slovenian healthy controls, the allele frequencies of minor allele G were similar in both healthy populations (SLO: 0,45; GB: 0,45; USA, 0,44;). Interestingly, the allele frequency of the minor allele G was also very similar in both populations of IBD



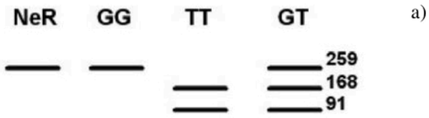


Figure 2. Restriction analysis of IL23R gene (rs7517847). a) Restriction scheme for IL23R gene, where NeR represents unrestricted sample, G/G, G/T and T/T genotype standards and H_2O a blank sample.

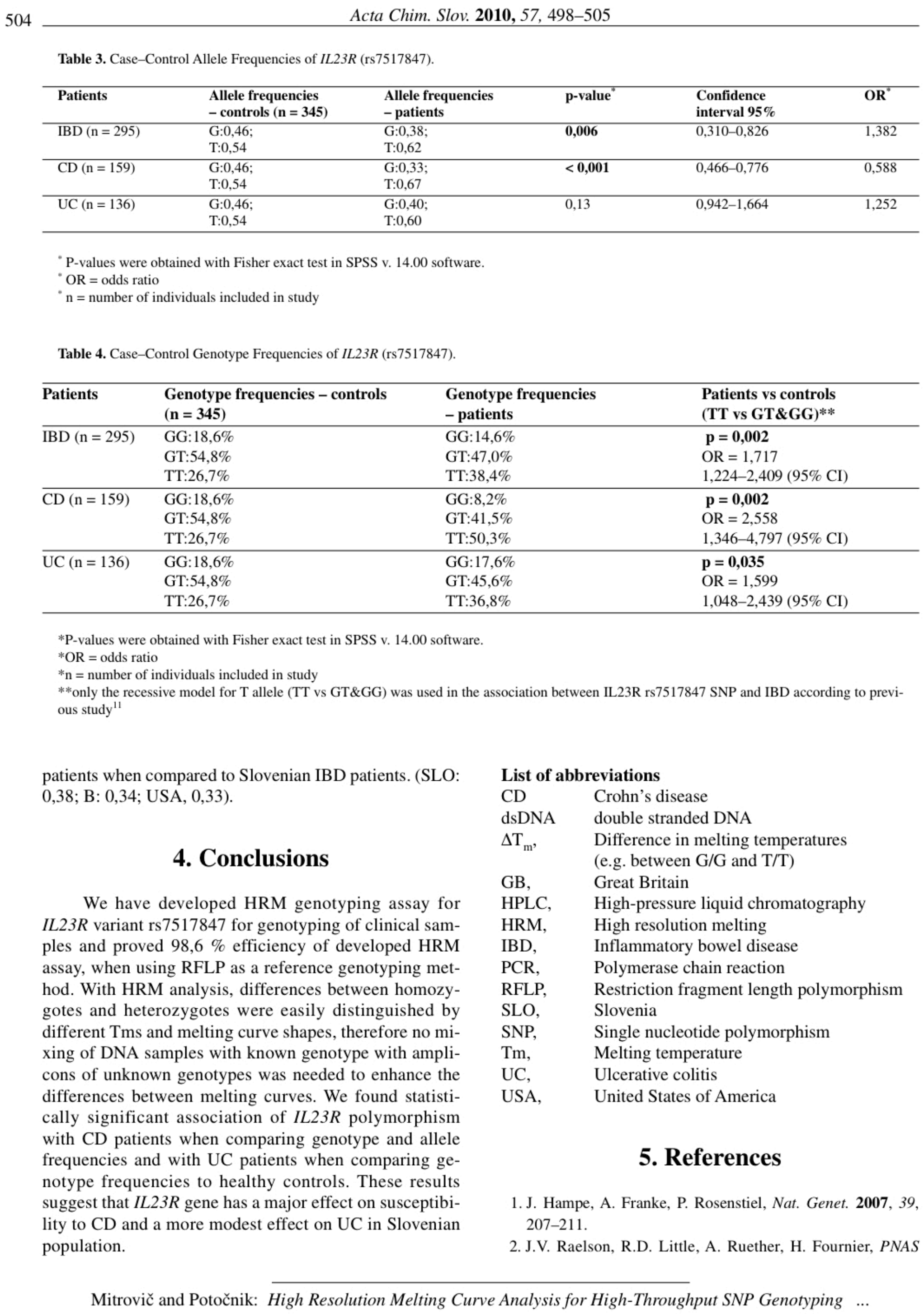


Table 3. Case–Control Allele Frequencies of *IL23R* (rs7517847).

Patients	**Allele frequencies – controls (n = 345)**	**Allele frequencies – patients**	**p-value***	**Confidence interval 95%**	**OR***
IBD (n = 295)	G:0,46; T:0,54	G:0,38; T:0,62	**0,006**	0,310–0,826	1,382
CD (n = 159)	G:0,46; T:0,54	G:0,33; T:0,67	**< 0,001**	0,466–0,776	0,588
UC (n = 136)	G:0,46; T:0,54	G:0,40; T:0,60	0,13	0,942–1,664	1,252

* P-values were obtained with Fisher exact test in SPSS v. 14.00 software.
* OR = odds ratio
* n = number of individuals included in study

Table 4. Case–Control Genotype Frequencies of *IL23R* (rs7517847).

Patients	**Genotype frequencies – controls (n = 345)**	**Genotype frequencies – patients**	**Patients vs controls (TT vs GT&GG)****
IBD (n = 295)	GG:18,6% GT:54,8% TT:26,7%	GG:14,6% GT:47,0% TT:38,4%	**p = 0,002** OR = 1,717 1,224–2,409 (95% CI)
CD (n = 159)	GG:18,6% GT:54,8% TT:26,7%	GG:8,2% GT:41,5% TT:50,3%	**p = 0,002** OR = 2,558 1,346–4,797 (95% CI)
UC (n = 136)	GG:18,6% GT:54,8% TT:26,7%	GG:17,6% GT:45,6% TT:36,8%	**p = 0,035** OR = 1,599 1,048–2,439 (95% CI)

*P-values were obtained with Fisher exact test in SPSS v. 14.00 software.
*OR = odds ratio
*n = number of individuals included in study
**only the recessive model for T allele (TT vs GT&GG) was used in the association between IL23R rs7517847 SNP and IBD according to previous study[11]

patients when compared to Slovenian IBD patients. (SLO: 0,38; B: 0,34; USA, 0,33).

4. Conclusions

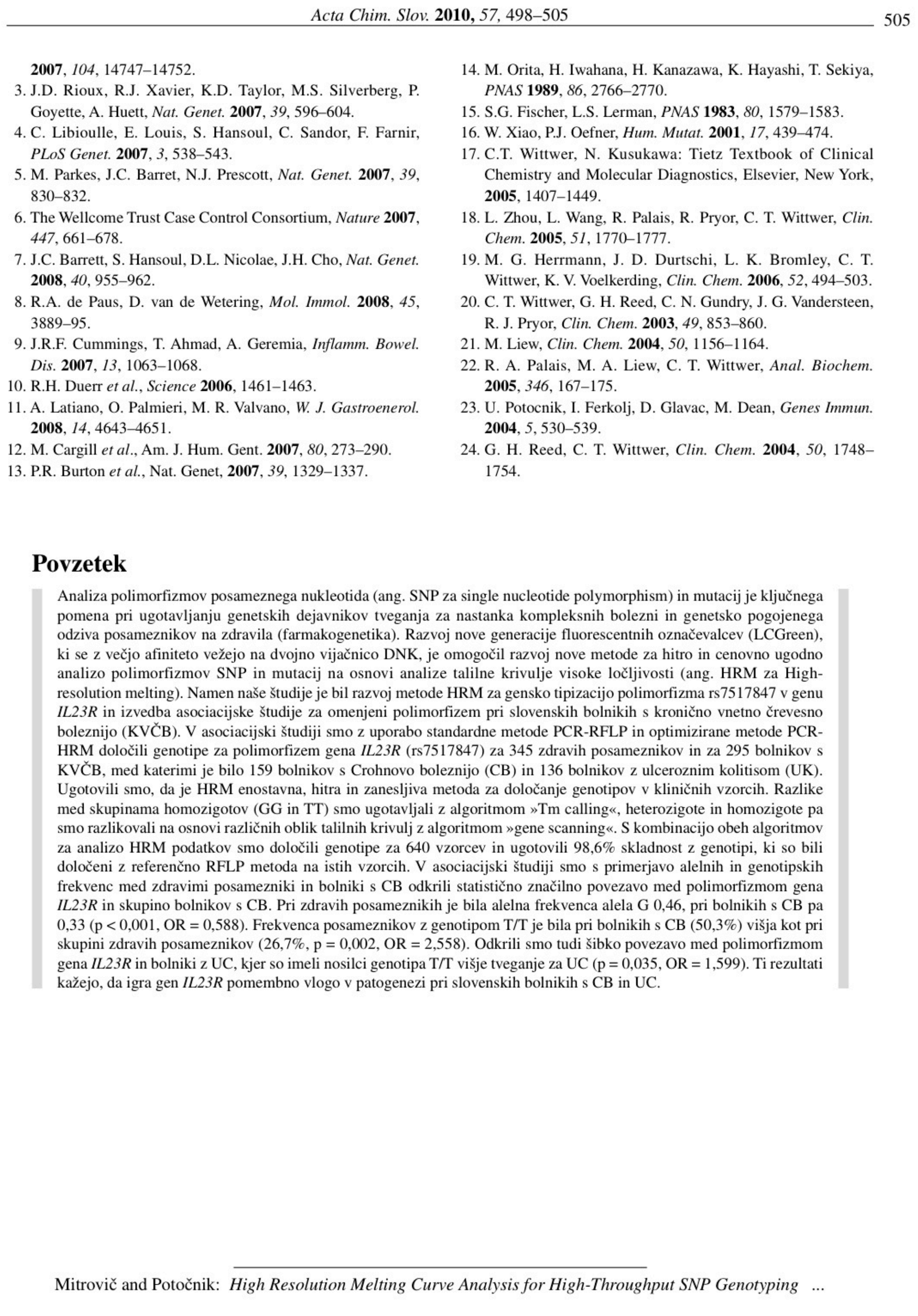
We have developed HRM genotyping assay for *IL23R* variant rs7517847 for genotyping of clinical samples and proved 98,6 % efficiency of developed HRM assay, when using RFLP as a reference genotyping method. With HRM analysis, differences between homozygotes and heterozygotes were easily distinguished by different Tms and melting curve shapes, therefore no mixing of DNA samples with known genotype with amplicons of unknown genotypes was needed to enhance the differences between melting curves. We found statistically significant association of *IL23R* polymorphism with CD patients when comparing genotype and allele frequencies and with UC patients when comparing genotype frequencies to healthy controls. These results suggest that *IL23R* gene has a major effect on susceptibility to CD and a more modest effect on UC in Slovenian population.

List of abbreviations

CD	Crohn's disease
dsDNA	double stranded DNA
ΔT_m,	Difference in melting temperatures (e.g. between G/G and T/T)
GB,	Great Britain
HPLC,	High-pressure liquid chromatography
HRM,	High resolution melting
IBD,	Inflammatory bowel disease
PCR,	Polymerase chain reaction
RFLP,	Restriction fragment length polymorphism
SLO,	Slovenia
SNP,	Single nucleotide polymorphism
Tm,	Melting temperature
UC,	Ulcerative colitis
USA,	United States of America

5. References

1. J. Hampe, A. Franke, P. Rosenstiel, *Nat. Genet.* **2007**, *39*, 207–211.
2. J.V. Raelson, R.D. Little, A. Ruether, H. Fournier, *PNAS*



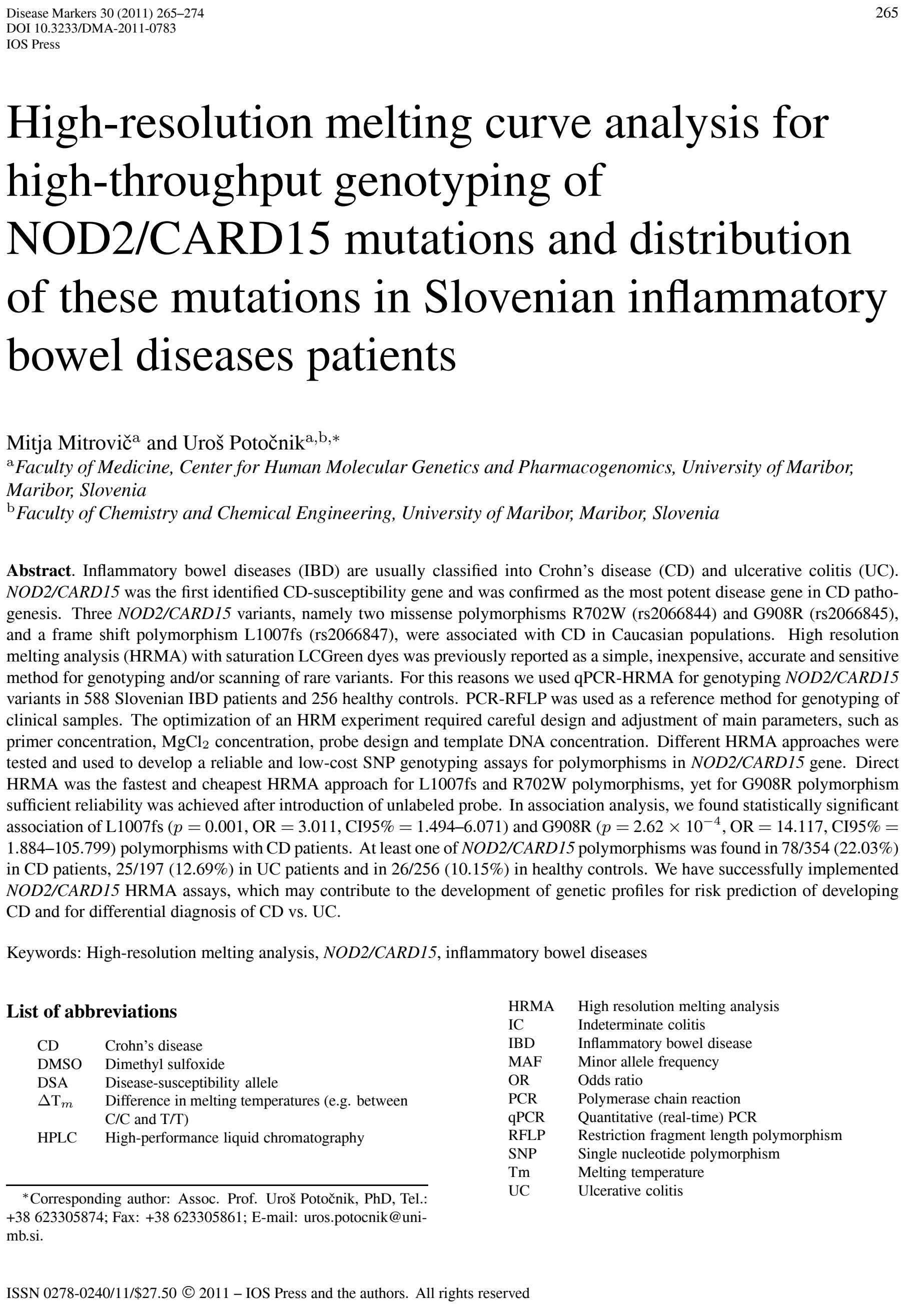
2007, *104*, 14747–14752.

3. J.D. Rioux, R.J. Xavier, K.D. Taylor, M.S. Silverberg, P. Goyette, A. Huett, *Nat. Genet.* **2007**, *39*, 596–604.
4. C. Libioulle, E. Louis, S. Hansoul, C. Sandor, F. Farnir, *PLoS Genet.* **2007**, *3*, 538–543.
5. M. Parkes, J.C. Barret, N.J. Prescott, *Nat. Genet.* **2007**, *39*, 830–832.
6. The Wellcome Trust Case Control Consortium, *Nature* **2007**, *447*, 661–678.
7. J.C. Barrett, S. Hansoul, D.L. Nicolae, J.H. Cho, *Nat. Genet.* **2008**, *40*, 955–962.
8. R.A. de Paus, D. van de Wetering, *Mol. Immol.* **2008**, *45*, 3889–95.
9. J.R.F. Cummings, T. Ahmad, A. Geremia, *Inflamm. Bowel. Dis.* **2007**, *13*, 1063–1068.
10. R.H. Duerr *et al.*, *Science* **2006**, 1461–1463.
11. A. Latiano, O. Palmieri, M. R. Valvano, *W. J. Gastroenerol.* **2008**, *14*, 4643–4651.
12. M. Cargill *et al*., Am. J. Hum. Gent. **2007**, *80*, 273–290.
13. P.R. Burton *et al.*, Nat. Genet, **2007**, *39*, 1329–1337.
14. M. Orita, H. Iwahana, H. Kanazawa, K. Hayashi, T. Sekiya, *PNAS* **1989**, *86*, 2766–2770.
15. S.G. Fischer, L.S. Lerman, *PNAS* **1983**, *80*, 1579–1583.
16. W. Xiao, P.J. Oefner, *Hum. Mutat.* **2001**, *17*, 439–474.
17. C.T. Wittwer, N. Kusukawa: Tietz Textbook of Clinical Chemistry and Molecular Diagnostics, Elsevier, New York, **2005**, 1407–1449.
18. L. Zhou, L. Wang, R. Palais, R. Pryor, C. T. Wittwer, *Clin. Chem.* **2005**, *51*, 1770–1777.
19. M. G. Herrmann, J. D. Durtschi, L. K. Bromley, C. T. Wittwer, K. V. Voelkerding, *Clin. Chem.* **2006**, *52*, 494–503.
20. C. T. Wittwer, G. H. Reed, C. N. Gundry, J. G. Vandersteen, R. J. Pryor, *Clin. Chem.* **2003**, *49*, 853–860.
21. M. Liew, *Clin. Chem.* **2004**, *50*, 1156–1164.
22. R. A. Palais, M. A. Liew, C. T. Wittwer, *Anal. Biochem.* **2005**, *346*, 167–175.
23. U. Potocnik, I. Ferkolj, D. Glavac, M. Dean, *Genes Immun.* **2004**, *5*, 530–539.
24. G. H. Reed, C. T. Wittwer, *Clin. Chem.* **2004**, *50*, 1748–1754.

Povzetek

Analiza polimorfizmov posameznega nukleotida (ang. SNP za single nucleotide polymorphism) in mutacij je ključnega pomena pri ugotavljanju genetskih dejavnikov tveganja za nastanka kompleksnih bolezni in genetsko pogojenega odziva posameznikov na zdravila (farmakogenetika). Razvoj nove generacije fluorescentnih označevalcev (LCGreen), ki se z večjo afiniteto vežejo na dvojno vijačnico DNK, je omogočil razvoj nove metode za hitro in cenovno ugodno analizo polimorfizmov SNP in mutacij na osnovi analize talilne krivulje visoke ločljivosti (ang. HRM za High-resolution melting). Namen naše študije je bil razvoj metode HRM za gensko tipizacijo polimorfizma rs7517847 v genu *IL23R* in izvedba asociacijske študije za omenjeni polimorfizem pri slovenskih bolnikih s kronično vnetno črevesno boleznijo (KVČB). V asociacijski študiji smo z uporabo standardne metode PCR-RFLP in optimizirane metode PCR-HRM določili genotipe za polimorfizem gena *IL23R* (rs7517847) za 345 zdravih posameznikov in za 295 bolnikov s KVČB, med katerimi je bilo 159 bolnikov s Crohnovo boleznijo (CB) in 136 bolnikov z ulceroznim kolitisom (UK). Ugotovili smo, da je HRM enostavna, hitra in zanesljiva metoda za določanje genotipov v kliničnih vzorcih. Razlike med skupinama homozigotov (GG in TT) smo ugotavljali z algoritmom »Tm calling«, heterozigote in homozigote pa smo razlikovali na osnovi različnih oblik talilnih krivulj z algoritmom »gene scanning«. S kombinacijo obeh algoritmov za analizo HRM podatkov smo določili genotipe za 640 vzorcev in ugotovili 98,6% skladnost z genotipi, ki so bili določeni z referenčno RFLP metoda na istih vzorcih. V asociacijski študiji smo s primerjavo alelnih in genotipskih frekvenc med zdravimi posamezniki in bolniki s CB odkrili statistično značilno povezavo med polimorfizmom gena *IL23R* in skupino bolnikov s CB. Pri zdravih posameznikih je bila alelna frekvenca alela G 0,46, pri bolnikih s CB pa 0,33 (p < 0,001, OR = 0,588). Frekvenca posameznikov z genotipom T/T je bila pri bolnikih s CB (50,3%) višja kot pri skupini zdravih posameznikov (26,7%, p = 0,002, OR = 2,558). Odkrili smo tudi šibko povezavo med polimorfizmom gena *IL23R* in bolniki z UC, kjer so imeli nosilci genotipa T/T višje tveganje za UC (p = 0,035, OR = 1,599). Ti rezultati kažejo, da igra gen *IL23R* pomembno vlogo v patogenezi pri slovenskih bolnikih s CB in UC.

6.6.2 Izvirni znanstveni članek 2



Disease Markers 30 (2011) 265–274
DOI 10.3233/DMA-2011-0783
IOS Press

High-resolution melting curve analysis for high-throughput genotyping of NOD2/CARD15 mutations and distribution of these mutations in Slovenian inflammatory bowel diseases patients

Mitja Mitrovič[a] and Uroš Potočnik[a,b,*]
[a] *Faculty of Medicine, Center for Human Molecular Genetics and Pharmacogenomics, University of Maribor, Maribor, Slovenia*
[b] *Faculty of Chemistry and Chemical Engineering, University of Maribor, Maribor, Slovenia*

Abstract. Inflammatory bowel diseases (IBD) are usually classified into Crohn's disease (CD) and ulcerative colitis (UC). *NOD2/CARD15* was the first identified CD-susceptibility gene and was confirmed as the most potent disease gene in CD pathogenesis. Three *NOD2/CARD15* variants, namely two missense polymorphisms R702W (rs2066844) and G908R (rs2066845), and a frame shift polymorphism L1007fs (rs2066847), were associated with CD in Caucasian populations. High resolution melting analysis (HRMA) with saturation LCGreen dyes was previously reported as a simple, inexpensive, accurate and sensitive method for genotyping and/or scanning of rare variants. For this reasons we used qPCR-HRMA for genotyping *NOD2/CARD15* variants in 588 Slovenian IBD patients and 256 healthy controls. PCR-RFLP was used as a reference method for genotyping of clinical samples. The optimization of an HRM experiment required careful design and adjustment of main parameters, such as primer concentration, $MgCl_2$ concentration, probe design and template DNA concentration. Different HRMA approaches were tested and used to develop a reliable and low-cost SNP genotyping assays for polymorphisms in *NOD2/CARD15* gene. Direct HRMA was the fastest and cheapest HRMA approach for L1007fs and R702W polymorphisms, yet for G908R polymorphism sufficient reliability was achieved after introduction of unlabeled probe. In association analysis, we found statistically significant association of L1007fs ($p = 0.001$, OR $= 3.011$, CI95% $= 1.494$–6.071) and G908R ($p = 2.62 \times 10^{-4}$, OR $= 14.117$, CI95% $=$ 1.884–105.799) polymorphisms with CD patients. At least one of *NOD2/CARD15* polymorphisms was found in 78/354 (22.03%) in CD patients, 25/197 (12.69%) in UC patients and in 26/256 (10.15%) in healthy controls. We have successfully implemented *NOD2/CARD15* HRMA assays, which may contribute to the development of genetic profiles for risk prediction of developing CD and for differential diagnosis of CD vs. UC.

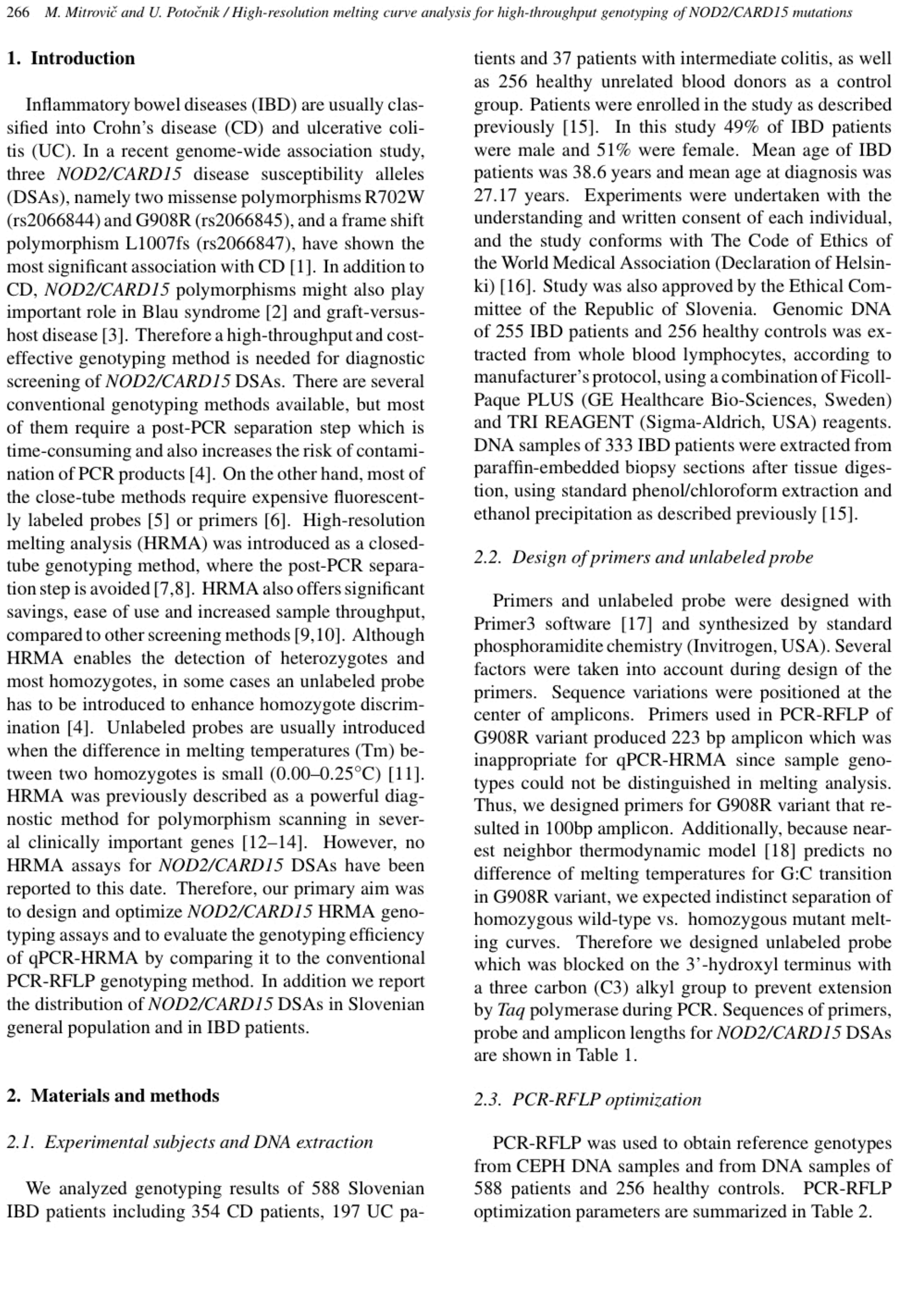
Keywords: High-resolution melting analysis, *NOD2/CARD15*, inflammatory bowel diseases

List of abbreviations

CD	Crohn's disease
DMSO	Dimethyl sulfoxide
DSA	Disease-susceptibility allele
ΔT_m	Difference in melting temperatures (e.g. between C/C and T/T)
HPLC	High-performance liquid chromatography
HRMA	High resolution melting analysis
IC	Indeterminate colitis
IBD	Inflammatory bowel disease
MAF	Minor allele frequency
OR	Odds ratio
PCR	Polymerase chain reaction
qPCR	Quantitative (real-time) PCR
RFLP	Restriction fragment length polymorphism
SNP	Single nucleotide polymorphism
Tm	Melting temperature
UC	Ulcerative colitis

[*] Corresponding author: Assoc. Prof. Uroš Potočnik, PhD, Tel.: +38 623305874; Fax: +38 623305861; E-mail: uros.potocnik@uni-mb.si.

ISSN 0278-0240/11/$27.50 © 2011 – IOS Press and the authors. All rights reserved



1. Introduction

Inflammatory bowel diseases (IBD) are usually classified into Crohn's disease (CD) and ulcerative colitis (UC). In a recent genome-wide association study, three *NOD2/CARD15* disease susceptibility alleles (DSAs), namely two missense polymorphisms R702W (rs2066844) and G908R (rs2066845), and a frame shift polymorphism L1007fs (rs2066847), have shown the most significant association with CD [1]. In addition to CD, *NOD2/CARD15* polymorphisms might also play important role in Blau syndrome [2] and graft-versus-host disease [3]. Therefore a high-throughput and cost-effective genotyping method is needed for diagnostic screening of *NOD2/CARD15* DSAs. There are several conventional genotyping methods available, but most of them require a post-PCR separation step which is time-consuming and also increases the risk of contamination of PCR products [4]. On the other hand, most of the close-tube methods require expensive fluorescently labeled probes [5] or primers [6]. High-resolution melting analysis (HRMA) was introduced as a closed-tube genotyping method, where the post-PCR separation step is avoided [7,8]. HRMA also offers significant savings, ease of use and increased sample throughput, compared to other screening methods [9,10]. Although HRMA enables the detection of heterozygotes and most homozygotes, in some cases an unlabeled probe has to be introduced to enhance homozygote discrimination [4]. Unlabeled probes are usually introduced when the difference in melting temperatures (Tm) between two homozygotes is small (0.00–0.25°C) [11]. HRMA was previously described as a powerful diagnostic method for polymorphism scanning in several clinically important genes [12–14]. However, no HRMA assays for *NOD2/CARD15* DSAs have been reported to this date. Therefore, our primary aim was to design and optimize *NOD2/CARD15* HRMA genotyping assays and to evaluate the genotyping efficiency of qPCR-HRMA by comparing it to the conventional PCR-RFLP genotyping method. In addition we report the distribution of *NOD2/CARD15* DSAs in Slovenian general population and in IBD patients.

2. Materials and methods

2.1. Experimental subjects and DNA extraction

We analyzed genotyping results of 588 Slovenian IBD patients including 354 CD patients, 197 UC patients and 37 patients with intermediate colitis, as well as 256 healthy unrelated blood donors as a control group. Patients were enrolled in the study as described previously [15]. In this study 49% of IBD patients were male and 51% were female. Mean age of IBD patients was 38.6 years and mean age at diagnosis was 27.17 years. Experiments were undertaken with the understanding and written consent of each individual, and the study conforms with The Code of Ethics of the World Medical Association (Declaration of Helsinki) [16]. Study was also approved by the Ethical Committee of the Republic of Slovenia. Genomic DNA of 255 IBD patients and 256 healthy controls was extracted from whole blood lymphocytes, according to manufacturer's protocol, using a combination of Ficoll-Paque PLUS (GE Healthcare Bio-Sciences, Sweden) and TRI REAGENT (Sigma-Aldrich, USA) reagents. DNA samples of 333 IBD patients were extracted from paraffin-embedded biopsy sections after tissue digestion, using standard phenol/chloroform extraction and ethanol precipitation as described previously [15].

2.2. Design of primers and unlabeled probe

Primers and unlabeled probe were designed with Primer3 software [17] and synthesized by standard phosphoramidite chemistry (Invitrogen, USA). Several factors were taken into account during design of the primers. Sequence variations were positioned at the center of amplicons. Primers used in PCR-RFLP of G908R variant produced 223 bp amplicon which was inappropriate for qPCR-HRMA since sample genotypes could not be distinguished in melting analysis. Thus, we designed primers for G908R variant that resulted in 100bp amplicon. Additionally, because nearest neighbor thermodynamic model [18] predicts no difference of melting temperatures for G:C transition in G908R variant, we expected indistinct separation of homozygous wild-type vs. homozygous mutant melting curves. Therefore we designed unlabeled probe which was blocked on the 3'-hydroxyl terminus with a three carbon (C3) alkyl group to prevent extension by *Taq* polymerase during PCR. Sequences of primers, probe and amplicon lengths for *NOD2/CARD15* DSAs are shown in Table 1.

2.3. PCR-RFLP optimization

PCR-RFLP was used to obtain reference genotypes from CEPH DNA samples and from DNA samples of 588 patients and 256 healthy controls. PCR-RFLP optimization parameters are summarized in Table 2.

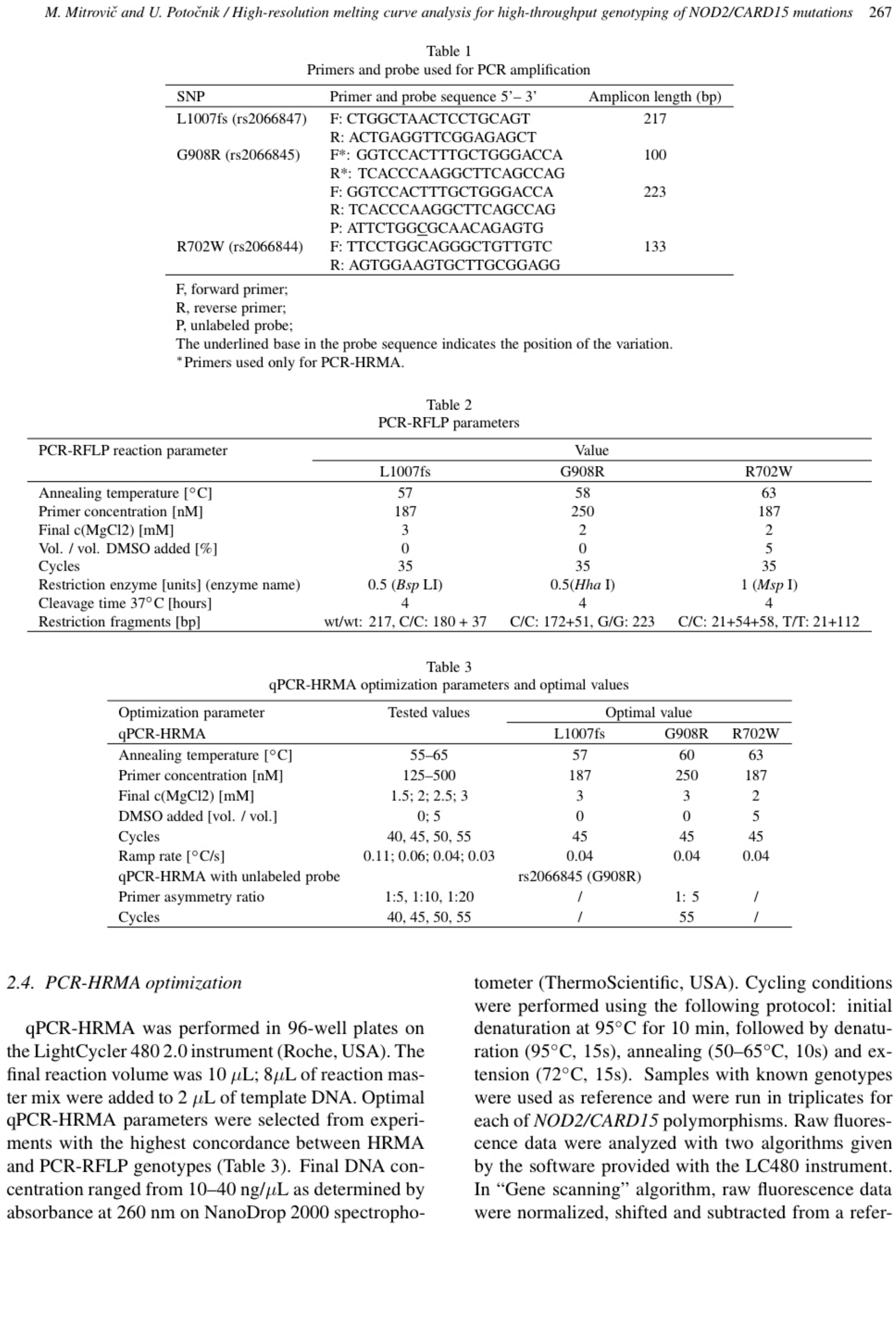


Table 1
Primers and probe used for PCR amplification

SNP	Primer and probe sequence 5’– 3’	Amplicon length (bp)
L1007fs (rs2066847)	F: CTGGCTAACTCCTGCAGT	217
	R: ACTGAGGTTCGGAGAGCT	
G908R (rs2066845)	F*: GGTCCACTTTGCTGGGACCA	100
	R*: TCACCCAAGGCTTCAGCCAG	
	F: GGTCCACTTTGCTGGGACCA	223
	R: TCACCCAAGGCTTCAGCCAG	
	P: ATTCTGGCGCAACAGAGTG	
R702W (rs2066844)	F: TTCCTGGCAGGGCTGTTGTC	133
	R: AGTGGAAGTGCTTGCGGAGG	

F, forward primer;
R, reverse primer;
P, unlabeled probe;
The underlined base in the probe sequence indicates the position of the variation.
*Primers used only for PCR-HRMA.

Table 2
PCR-RFLP parameters

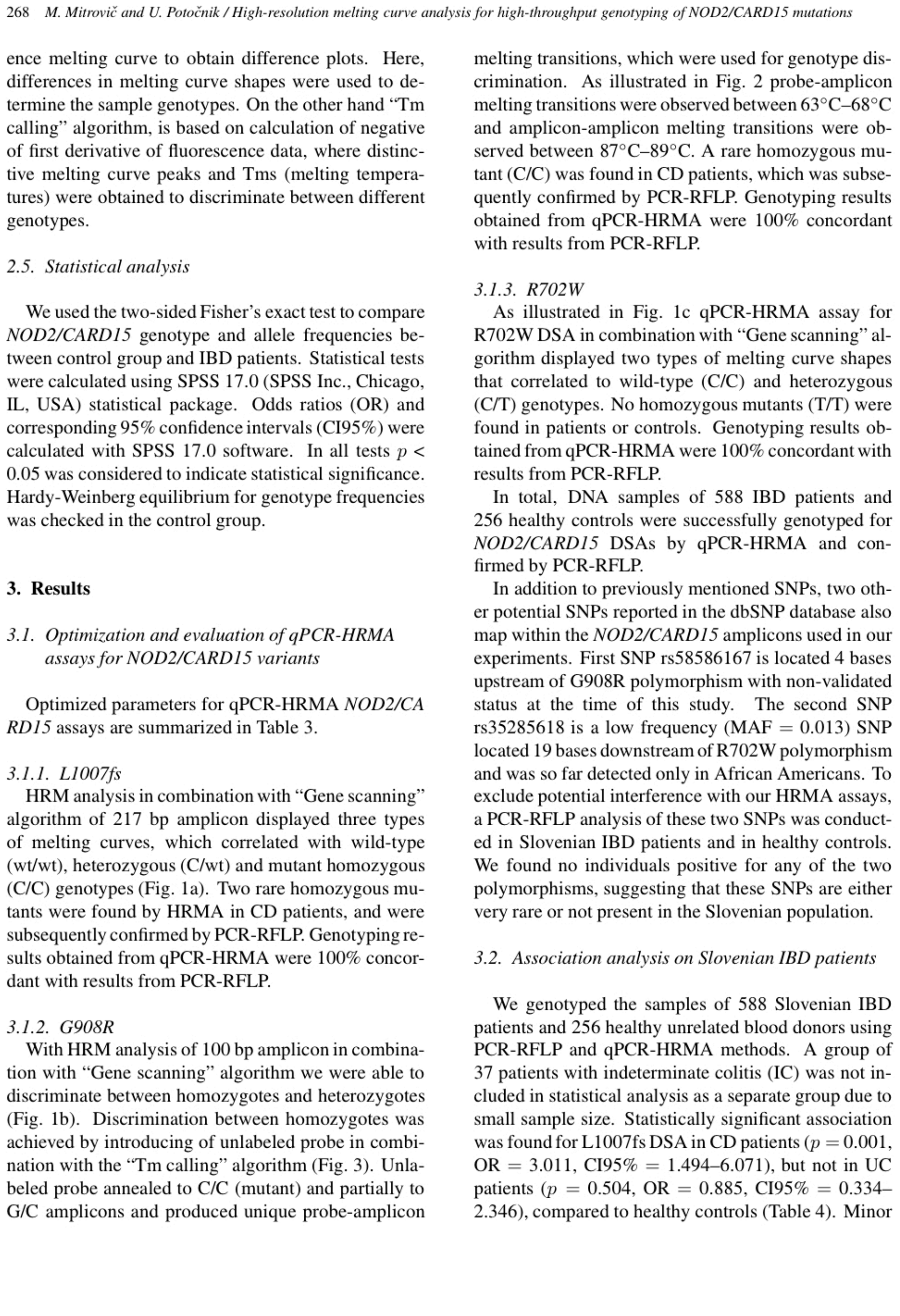
PCR-RFLP reaction parameter	Value		
	L1007fs	G908R	R702W
Annealing temperature [°C]	57	58	63
Primer concentration [nM]	187	250	187
Final c(MgCl2) [mM]	3	2	2
Vol. / vol. DMSO added [%]	0	0	5
Cycles	35	35	35
Restriction enzyme [units] (enzyme name)	0.5 (*Bsp* LI)	0.5(*Hha* I)	1 (*Msp* I)
Cleavage time 37°C [hours]	4	4	4
Restriction fragments [bp]	wt/wt: 217, C/C: 180 + 37	C/C: 172+51, G/G: 223	C/C: 21+54+58, T/T: 21+112

Table 3
qPCR-HRMA optimization parameters and optimal values

Optimization parameter	Tested values	Optimal value		
qPCR-HRMA		L1007fs	G908R	R702W
Annealing temperature [°C]	55–65	57	60	63
Primer concentration [nM]	125–500	187	250	187
Final c(MgCl2) [mM]	1.5; 2; 2.5; 3	3	3	2
DMSO added [vol. / vol.]	0; 5	0	0	5
Cycles	40, 45, 50, 55	45	45	45
Ramp rate [°C/s]	0.11; 0.06; 0.04; 0.03	0.04	0.04	0.04
qPCR-HRMA with unlabeled probe		rs2066845 (G908R)		
Primer asymmetry ratio	1:5, 1:10, 1:20	/	1: 5	/
Cycles	40, 45, 50, 55	/	55	/

2.4. PCR-HRMA optimization

qPCR-HRMA was performed in 96-well plates on the LightCycler 480 2.0 instrument (Roche, USA). The final reaction volume was 10 μL; 8μL of reaction master mix were added to 2 μL of template DNA. Optimal qPCR-HRMA parameters were selected from experiments with the highest concordance between HRMA and PCR-RFLP genotypes (Table 3). Final DNA concentration ranged from 10–40 ng/μL as determined by absorbance at 260 nm on NanoDrop 2000 spectrophotometer (ThermoScientific, USA). Cycling conditions were performed using the following protocol: initial denaturation at 95°C for 10 min, followed by denaturation (95°C, 15s), annealing (50–65°C, 10s) and extension (72°C, 15s). Samples with known genotypes were used as reference and were run in triplicates for each of *NOD2/CARD15* polymorphisms. Raw fluorescence data were analyzed with two algorithms given by the software provided with the LC480 instrument. In “Gene scanning” algorithm, raw fluorescence data were normalized, shifted and subtracted from a refer-



ence melting curve to obtain difference plots. Here, differences in melting curve shapes were used to determine the sample genotypes. On the other hand "Tm calling" algorithm, is based on calculation of negative of first derivative of fluorescence data, where distinctive melting curve peaks and Tms (melting temperatures) were obtained to discriminate between different genotypes.

2.5. Statistical analysis

We used the two-sided Fisher's exact test to compare *NOD2/CARD15* genotype and allele frequencies between control group and IBD patients. Statistical tests were calculated using SPSS 17.0 (SPSS Inc., Chicago, IL, USA) statistical package. Odds ratios (OR) and corresponding 95% confidence intervals (CI95%) were calculated with SPSS 17.0 software. In all tests $p < 0.05$ was considered to indicate statistical significance. Hardy-Weinberg equilibrium for genotype frequencies was checked in the control group.

3. Results

3.1. Optimization and evaluation of qPCR-HRMA assays for NOD2/CARD15 variants

Optimized parameters for qPCR-HRMA *NOD2/CARD15* assays are summarized in Table 3.

3.1.1. L1007fs

HRM analysis in combination with "Gene scanning" algorithm of 217 bp amplicon displayed three types of melting curves, which correlated with wild-type (wt/wt), heterozygous (C/wt) and mutant homozygous (C/C) genotypes (Fig. 1a). Two rare homozygous mutants were found by HRMA in CD patients, and were subsequently confirmed by PCR-RFLP. Genotyping results obtained from qPCR-HRMA were 100% concordant with results from PCR-RFLP.

3.1.2. G908R

With HRM analysis of 100 bp amplicon in combination with "Gene scanning" algorithm we were able to discriminate between homozygotes and heterozygotes (Fig. 1b). Discrimination between homozygotes was achieved by introducing of unlabeled probe in combination with the "Tm calling" algorithm (Fig. 3). Unlabeled probe annealed to C/C (mutant) and partially to G/C amplicons and produced unique probe-amplicon melting transitions, which were used for genotype discrimination. As illustrated in Fig. 2 probe-amplicon melting transitions were observed between 63°C–68°C and amplicon-amplicon melting transitions were observed between 87°C–89°C. A rare homozygous mutant (C/C) was found in CD patients, which was subsequently confirmed by PCR-RFLP. Genotyping results obtained from qPCR-HRMA were 100% concordant with results from PCR-RFLP.

3.1.3. R702W

As illustrated in Fig. 1c qPCR-HRMA assay for R702W DSA in combination with "Gene scanning" algorithm displayed two types of melting curve shapes that correlated to wild-type (C/C) and heterozygous (C/T) genotypes. No homozygous mutants (T/T) were found in patients or controls. Genotyping results obtained from qPCR-HRMA were 100% concordant with results from PCR-RFLP.

In total, DNA samples of 588 IBD patients and 256 healthy controls were successfully genotyped for *NOD2/CARD15* DSAs by qPCR-HRMA and confirmed by PCR-RFLP.

In addition to previously mentioned SNPs, two other potential SNPs reported in the dbSNP database also map within the *NOD2/CARD15* amplicons used in our experiments. First SNP rs58586167 is located 4 bases upstream of G908R polymorphism with non-validated status at the time of this study. The second SNP rs35285618 is a low frequency (MAF = 0.013) SNP located 19 bases downstream of R702W polymorphism and was so far detected only in African Americans. To exclude potential interference with our HRMA assays, a PCR-RFLP analysis of these two SNPs was conducted in Slovenian IBD patients and in healthy controls. We found no individuals positive for any of the two polymorphisms, suggesting that these SNPs are either very rare or not present in the Slovenian population.

3.2. Association analysis on Slovenian IBD patients

We genotyped the samples of 588 Slovenian IBD patients and 256 healthy unrelated blood donors using PCR-RFLP and qPCR-HRMA methods. A group of 37 patients with indeterminate colitis (IC) was not included in statistical analysis as a separate group due to small sample size. Statistically significant association was found for L1007fs DSA in CD patients ($p = 0.001$, OR $= 3.011$, CI95% $= 1.494$–6.071), but not in UC patients ($p = 0.504$, OR $= 0.885$, CI95% $= 0.334$–2.346), compared to healthy controls (Table 4). Minor

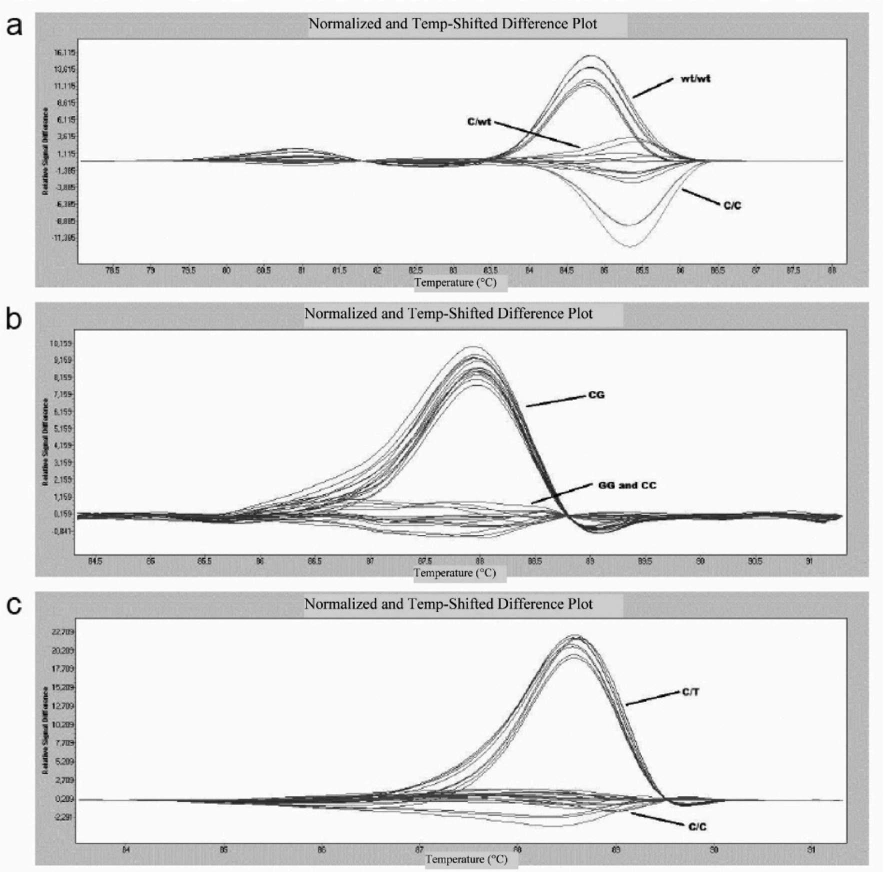


Fig. 1. Difference plots of *NOD2/CARD15* polymorphisms. Difference plots were obtained subsequent to raw fluorescence data normalization, Tm-shift and subtraction from a reference melting curve by "Gene scanning" algorithm. a) L1007fs (rs2066847). Genotypes of L1007fs polymorphism were clearly distinguished and grouped into three groups. wt/wt represents melting curves of wild-type homozygotes, C/wt melting curves of heterozygotes and C/C melting curves of mutant homozygotes. b) G908R (rs2066845). Heterozygotes were unambiguously distinguished from both groups of homozygotes due to altered melting curve shapes. Differentiation of homozygotes was difficult because of similar melting curve shapes of G/G and C/C homozygotes. c) R702W (rs2066844). Wild-type homozygotes and C/T heterozygotes were clearly distinguished. No homozygous mutants (T/T) were found in our study. C/C represents wild-type homozygotes and C/T heterozygotes.

allele frequency (MAF) of L1007fs polymorphism was 2% in control group, 5.9% in CD group and 1.8% in UC group. L1007fs MAF was also significantly higher in CD patients as compared to UC patients ($p = 0.001$, OR = 3.404, CI95% = 1.512–7.663). In case of G908R DSA, we found statistically significant association in CD patients ($p = 2.62 \times 10^{-4}$, OR = 14.117, CI95% = 1.884–105.799) and UC patients ($p = 0.013$, OR = 9.389, CI95% = 1.150–76.641) as compared to healthy controls. G908R MAF was 0.2% in control group,

Table 4
Results of association analysis of *NOD2/CARD15* DSAs in Slovenian IBD patients and healthy controls

	L1007fs			G908R			R702W		
	C	CD	UC	C	CD	UC	C	CD	UC
Homozygotes (wt)	237	311	188	252	330	184	238	323	185
Heterozygotes	10	37	7	1	17	7	15	30	11
Homozygotes (mt)	0	2	0	0	1	0	0	0	0
Total	247	350	195	253	348	191	253	353	196
Homozygotes (wt) [%]	96.0	88.86	96.41	99.6	94.82	96.34	94.1	91.50	94.39
Heterozygotes [%]	4.0	10.57	3.59	0.4	4.89	3.66	5.9	8.50	5.61
Homozygotes (mt) [%]	0	0.57	0	0	0.29	0	0	0	0
MAF	0.02	0.059	0.018	0.002	0.027	0.018	0.03	0.042	0.028
p-value		0.001	0.504		2.62×10^{-4}	0.013		0.156	0.527
OR		3.011	0.885		14.117	9.389		0.450	0.945
CI95%		1.494–6.071	0.334–2.346		1.884–105.799	1.150–76.641		0.773–2.729	0.429–2.081

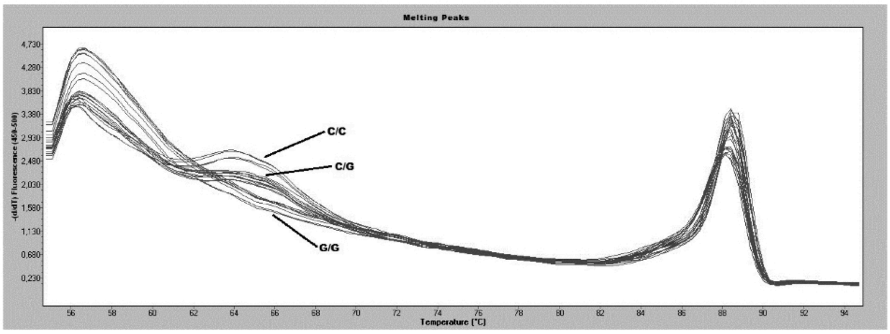


Fig. 2. Melting peaks of G908R (rs2066845) amplicon and unlabeled probe melting. Negative of first derivative of raw fluorescence data (-dF/dT) was calculated by "Tm calling" algorithm and presented on y – axis against temperature (T) on x–axis. Unlabeled probe annealed to C/C (homozygous-mutant) and partially to heterozygous G/C amplicons. This resulted in a unique probe-amplicon melting transitions (62°C–68°C), which were used for differentiation of G908R genotypes. Differentiation of genotypes by amplicon melting was not possible, since the amplicon melting peaks were confined to a narrow temperature interval (88.2°C–88.5°C).

2.7% in CD patients and 1.8% in UC patients. We found no statistically significant association of R702W DSA with CD and UC patients. Detailed results of association analysis are summarized in Table 4. The frequency of carriers of any polymorphism was significantly higher in CD patients compared to healthy controls ($p = 6.53 \times 10^{-5}$, OR $= 2.5$, CI95% $= 1.551$–4.028) and also when compared to UC patients ($p = 0.004$, OR $= 1.944$, CI95% $= 1.192$–3.171) (Table 5). At least one of *NOD2/CARD15* DSAs was found in 78/354 (22.03%) CD patients, 25/197 (12.69%) in UC patients and 26/256 (10.15%) in healthy controls. Additionally, we found 9/354 (2.54%) compound heterozygotes in CD patients and none in UC patients or in control group. In addition, mutant homozygotes for L1007fs and G908R polymorphisms were found only in group of CD patients.

Table 5
Prevalence of at least one *NOD/CARD15* DSA in Slovenian IBD patients and controls

	At least one mutation		
	C	CD	UC
Any variant	0.1015	0.2203	0.1269
p-value		6.53×10^{-5}	0.243
OR		2.500	1.286
CI95%		1.551–4.028	0.717–2.304

OR, odds ratio; CI95%, confidence interval 95%; C, controls; CD, Crohn's disease; UC, ulcerative colitis.

4. Discussion

We have developed three qPCR-HRMA genotyping assays for *NOD2/CARD15* DSAs and compared them to standard PCR-RFLP assays. As reported in previous studies [19,20], we also observed that the success of

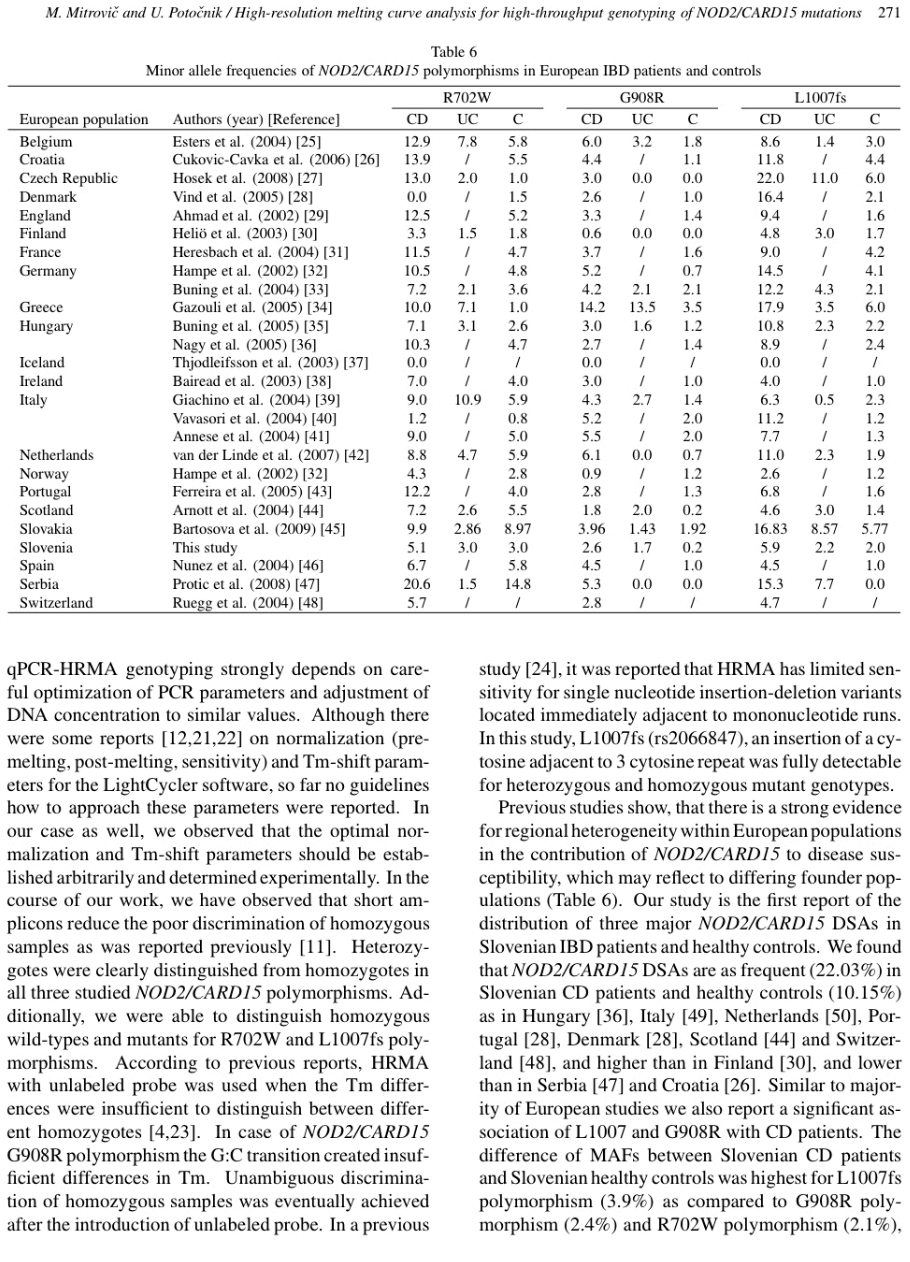
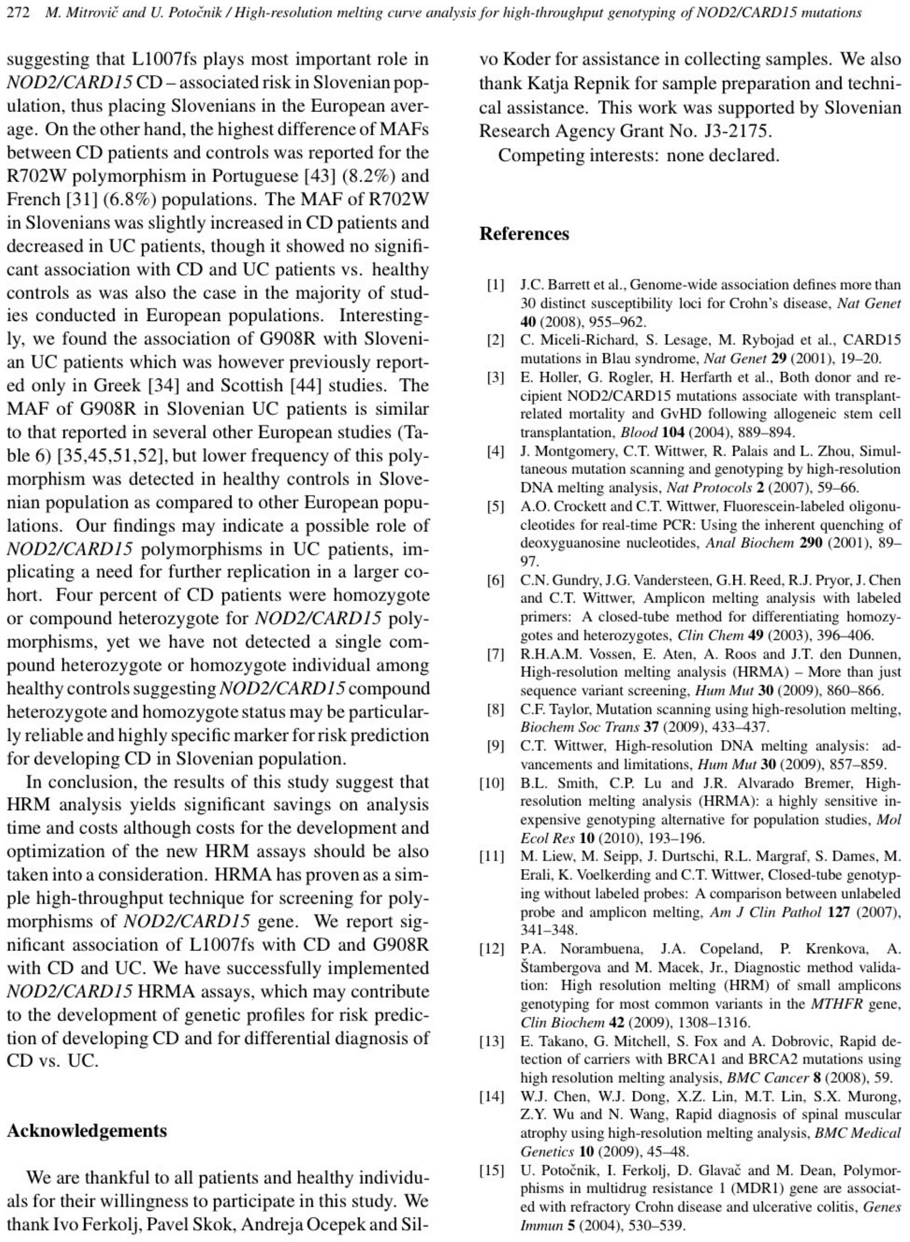


Table 6
Minor allele frequencies of *NOD2/CARD15* polymorphisms in European IBD patients and controls

		R702W			G908R			L1007fs		
European population	Authors (year) [Reference]	CD	UC	C	CD	UC	C	CD	UC	C
Belgium	Esters et al. (2004) [25]	12.9	7.8	5.8	6.0	3.2	1.8	8.6	1.4	3.0
Croatia	Cukovic-Cavka et al. (2006) [26]	13.9	/	5.5	4.4	/	1.1	11.8	/	4.4
Czech Republic	Hosek et al. (2008) [27]	13.0	2.0	1.0	3.0	0.0	0.0	22.0	11.0	6.0
Denmark	Vind et al. (2005) [28]	0.0	/	1.5	2.6	/	1.0	16.4	/	2.1
England	Ahmad et al. (2002) [29]	12.5	/	5.2	3.3	/	1.4	9.4	/	1.6
Finland	Heliö et al. (2003) [30]	3.3	1.5	1.8	0.6	0.0	0.0	4.8	3.0	1.7
France	Heresbach et al. (2004) [31]	11.5	/	4.7	3.7	/	1.6	9.0	/	4.2
Germany	Hampe et al. (2002) [32]	10.5	/	4.8	5.2	/	0.7	14.5	/	4.1
	Buning et al. (2004) [33]	7.2	2.1	3.6	4.2	2.1	2.1	12.2	4.3	2.1
Greece	Gazouli et al. (2005) [34]	10.0	7.1	1.0	14.2	13.5	3.5	17.9	3.5	6.0
Hungary	Buning et al. (2005) [35]	7.1	3.1	2.6	3.0	1.6	1.2	10.8	2.3	2.2
	Nagy et al. (2005) [36]	10.3	/	4.7	2.7	/	1.4	8.9	/	2.4
Iceland	Thjodleifsson et al. (2003) [37]	0.0	/	/	0.0	/	/	0.0	/	/
Ireland	Bairead et al. (2003) [38]	7.0	/	4.0	3.0	/	1.0	4.0	/	1.0
Italy	Giachino et al. (2004) [39]	9.0	10.9	5.9	4.3	2.7	1.4	6.3	0.5	2.3
	Vavasori et al. (2004) [40]	1.2	/	0.8	5.2	/	2.0	11.2	/	1.2
	Annese et al. (2004) [41]	9.0	/	5.0	5.5	/	2.0	7.7	/	1.3
Netherlands	van der Linde et al. (2007) [42]	8.8	4.7	5.9	6.1	0.0	0.7	11.0	2.3	1.9
Norway	Hampe et al. (2002) [32]	4.3	/	2.8	0.9	/	1.2	2.6	/	1.2
Portugal	Ferreira et al. (2005) [43]	12.2	/	4.0	2.8	/	1.3	6.8	/	1.6
Scotland	Arnott et al. (2004) [44]	7.2	2.6	5.5	1.8	2.0	0.2	4.6	3.0	1.4
Slovakia	Bartosova et al. (2009) [45]	9.9	2.86	8.97	3.96	1.43	1.92	16.83	8.57	5.77
Slovenia	This study	5.1	3.0	3.0	2.6	1.7	0.2	5.9	2.2	2.0
Spain	Nunez et al. (2004) [46]	6.7	/	5.8	4.5	/	1.0	4.5	/	1.0
Serbia	Protic et al. (2008) [47]	20.6	1.5	14.8	5.3	0.0	0.0	15.3	7.7	0.0
Switzerland	Ruegg et al. (2004) [48]	5.7	/	/	2.8	/	/	4.7	/	/

qPCR-HRMA genotyping strongly depends on careful optimization of PCR parameters and adjustment of DNA concentration to similar values. Although there were some reports [12,21,22] on normalization (premelting, post-melting, sensitivity) and Tm-shift parameters for the LightCycler software, so far no guidelines how to approach these parameters were reported. In our case as well, we observed that the optimal normalization and Tm-shift parameters should be established arbitrarily and determined experimentally. In the course of our work, we have observed that short amplicons reduce the poor discrimination of homozygous samples as was reported previously [11]. Heterozygotes were clearly distinguished from homozygotes in all three studied *NOD2/CARD15* polymorphisms. Additionally, we were able to distinguish homozygous wild-types and mutants for R702W and L1007fs polymorphisms. According to previous reports, HRMA with unlabeled probe was used when the Tm differences were insufficient to distinguish between different homozygotes [4,23]. In case of *NOD2/CARD15* G908R polymorphism the G:C transition created insufficient differences in Tm. Unambiguous discrimination of homozygous samples was eventually achieved after the introduction of unlabeled probe. In a previous study [24], it was reported that HRMA has limited sensitivity for single nucleotide insertion-deletion variants located immediately adjacent to mononucleotide runs. In this study, L1007fs (rs2066847), an insertion of a cytosine adjacent to 3 cytosine repeat was fully detectable for heterozygous and homozygous mutant genotypes.

Previous studies show, that there is a strong evidence for regional heterogeneity within European populations in the contribution of *NOD2/CARD15* to disease susceptibility, which may reflect to differing founder populations (Table 6). Our study is the first report of the distribution of three major *NOD2/CARD15* DSAs in Slovenian IBD patients and healthy controls. We found that *NOD2/CARD15* DSAs are as frequent (22.03%) in Slovenian CD patients and healthy controls (10.15%) as in Hungary [36], Italy [49], Netherlands [50], Portugal [28], Denmark [28], Scotland [44] and Switzerland [48], and higher than in Finland [30], and lower than in Serbia [47] and Croatia [26]. Similar to majority of European studies we also report a significant association of L1007 and G908R with CD patients. The difference of MAFs between Slovenian CD patients and Slovenian healthy controls was highest for L1007fs polymorphism (3.9%) as compared to G908R polymorphism (2.4%) and R702W polymorphism (2.1%),



suggesting that L1007fs plays most important role in *NOD2/CARD15* CD – associated risk in Slovenian population, thus placing Slovenians in the European average. On the other hand, the highest difference of MAFs between CD patients and controls was reported for the R702W polymorphism in Portuguese [43] (8.2%) and French [31] (6.8%) populations. The MAF of R702W in Slovenians was slightly increased in CD patients and decreased in UC patients, though it showed no significant association with CD and UC patients vs. healthy controls as was also the case in the majority of studies conducted in European populations. Interestingly, we found the association of G908R with Slovenian UC patients which was however previously reported only in Greek [34] and Scottish [44] studies. The MAF of G908R in Slovenian UC patients is similar to that reported in several other European studies (Table 6) [35,45,51,52], but lower frequency of this polymorphism was detected in healthy controls in Slovenian population as compared to other European populations. Our findings may indicate a possible role of *NOD2/CARD15* polymorphisms in UC patients, implicating a need for further replication in a larger cohort. Four percent of CD patients were homozygote or compound heterozygote for *NOD2/CARD15* polymorphisms, yet we have not detected a single compound heterozygote or homozygote individual among healthy controls suggesting *NOD2/CARD15* compound heterozygote and homozygote status may be particularly reliable and highly specific marker for risk prediction for developing CD in Slovenian population.

In conclusion, the results of this study suggest that HRM analysis yields significant savings on analysis time and costs although costs for the development and optimization of the new HRM assays should be also taken into a consideration. HRMA has proven as a simple high-throughput technique for screening for polymorphisms of *NOD2/CARD15* gene. We report significant association of L1007fs with CD and G908R with CD and UC. We have successfully implemented *NOD2/CARD15* HRMA assays, which may contribute to the development of genetic profiles for risk prediction of developing CD and for differential diagnosis of CD vs. UC.

Acknowledgements

We are thankful to all patients and healthy individuals for their willingness to participate in this study. We thank Ivo Ferkolj, Pavel Skok, Andreja Ocepek and Silvo Koder for assistance in collecting samples. We also thank Katja Repnik for sample preparation and technical assistance. This work was supported by Slovenian Research Agency Grant No. J3-2175.

Competing interests: none declared.

References

[1] J.C. Barrett et al., Genome-wide association defines more than 30 distinct susceptibility loci for Crohn's disease, *Nat Genet* **40** (2008), 955–962.

[2] C. Miceli-Richard, S. Lesage, M. Rybojad et al., CARD15 mutations in Blau syndrome, *Nat Genet* **29** (2001), 19–20.

[3] E. Holler, G. Rogler, H. Herfarth et al., Both donor and recipient NOD2/CARD15 mutations associate with transplant-related mortality and GvHD following allogeneic stem cell transplantation, *Blood* **104** (2004), 889–894.

[4] J. Montgomery, C.T. Wittwer, R. Palais and L. Zhou, Simultaneous mutation scanning and genotyping by high-resolution DNA melting analysis, *Nat Protocols* **2** (2007), 59–66.

[5] A.O. Crockett and C.T. Wittwer, Fluorescein-labeled oligonucleotides for real-time PCR: Using the inherent quenching of deoxyguanosine nucleotides, *Anal Biochem* **290** (2001), 89–97.

[6] C.N. Gundry, J.G. Vandersteen, G.H. Reed, R.J. Pryor, J. Chen and C.T. Wittwer, Amplicon melting analysis with labeled primers: A closed-tube method for differentiating homozygotes and heterozygotes, *Clin Chem* **49** (2003), 396–406.

[7] R.H.A.M. Vossen, E. Aten, A. Roos and J.T. den Dunnen, High-resolution melting analysis (HRMA) – More than just sequence variant screening, *Hum Mut* **30** (2009), 860–866.

[8] C.F. Taylor, Mutation scanning using high-resolution melting, *Biochem Soc Trans* **37** (2009), 433–437.

[9] C.T. Wittwer, High-resolution DNA melting analysis: advancements and limitations, *Hum Mut* **30** (2009), 857–859.

[10] B.L. Smith, C.P. Lu and J.R. Alvarado Bremer, High-resolution melting analysis (HRMA): a highly sensitive inexpensive genotyping alternative for population studies, *Mol Ecol Res* **10** (2010), 193–196.

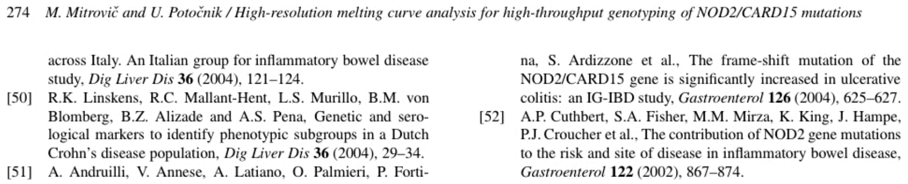
[11] M. Liew, M. Seipp, J. Durtschi, R.L. Margraf, S. Dames, M. Erali, K. Voelkerding and C.T. Wittwer, Closed-tube genotyping without labeled probes: A comparison between unlabeled probe and amplicon melting, *Am J Clin Pathol* **127** (2007), 341–348.

[12] P.A. Norambuena, J.A. Copeland, P. Krenkova, A. Štambergova and M. Macek, Jr., Diagnostic method validation: High resolution melting (HRM) of small amplicons genotyping for most common variants in the *MTHFR* gene, *Clin Biochem* **42** (2009), 1308–1316.

[13] E. Takano, G. Mitchell, S. Fox and A. Dobrovic, Rapid detection of carriers with BRCA1 and BRCA2 mutations using high resolution melting analysis, *BMC Cancer* **8** (2008), 59.

[14] W.J. Chen, W.J. Dong, X.Z. Lin, M.T. Lin, S.X. Murong, Z.Y. Wu and N. Wang, Rapid diagnosis of spinal muscular atrophy using high-resolution melting analysis, *BMC Medical Genetics* **10** (2009), 45–48.

[15] U. Potočnik, I. Ferkolj, D. Glavač and M. Dean, Polymorphisms in multidrug resistance 1 (MDR1) gene are associated with refractory Crohn disease and ulcerative colitis, *Genes Immun* **5** (2004), 530–539.



across Italy. An Italian group for inflammatory bowel disease study, *Dig Liver Dis* **36** (2004), 121–124.

[50] R.K. Linskens, R.C. Mallant-Hent, L.S. Murillo, B.M. von Blomberg, B.Z. Alizade and A.S. Pena, Genetic and serological markers to identify phenotypic subgroups in a Dutch Crohn's disease population, *Dig Liver Dis* **36** (2004), 29–34.

[51] A. Andruilli, V. Annese, A. Latiano, O. Palmieri, P. Fortina, S. Ardizzone et al., The frame-shift mutation of the NOD2/CARD15 gene is significantly increased in ulcerative colitis: an IG-IBD study, *Gastroenterol* **126** (2004), 625–627.

[52] A.P. Cuthbert, S.A. Fisher, M.M. Mirza, K. King, J. Hampe, P.J. Croucher et al., The contribution of NOD2 gene mutations to the risk and site of disease in inflammatory bowel disease, *Gastroenterol* **122** (2002), 867–874.

6.6.3 Pregledni znanstveni članek

Fransen *et al. Genome Medicine* 2011, 3:13
http://genomemedicine.com/content/3/2/13

Genome Medicine

REVIEW

The quest for genetic risk factors for Crohn's disease in the post-GWAS era

Karin Fransen[†1,2], Mitja Mitrovic[†1,3], Cleo C van Diemen[1] and Rinse K Weersma[*2]

Abstract

Multiple genome-wide association studies (GWASs) and two large scale meta-analyses have been performed for Crohn's disease and have identified 71 susceptibility loci. These findings have contributed greatly to our current understanding of the disease pathogenesis. Yet, these loci only explain approximately 23% of the disease heritability. One of the future challenges in this post-GWAS era is to identify potential sources of the remaining heritability. Such sources may include common variants with limited effect size, rare variants with higher effect sizes, structural variations, or even more complicated mechanisms such as epistatic, gene-environment and epigenetic interactions. Here, we outline potential sources of this hidden heritability, focusing on Crohn's disease and the currently available data. We also discuss future strategies to determine more about the heritability; these strategies include expanding current GWAS, fine-mapping, whole genome sequencing or exome sequencing, and using family-based approaches. Despite the current limitations, such strategies may help to transfer research achievements into clinical practice and guide the improvement of preventive and therapeutic measures.

Background

Crohn's disease (CD) is one of the two main forms of inflammatory bowel disease (IBD), the other being ulcerative colitis (UC). It is a chronic disease characterized by recurring inflammation of the gut, and is thought to arise in response to the commensal microflora in a genetically susceptible host [1]. It can affect the entire gastrointestinal tract, although the most common locations are the terminal ileum and the colon. Symptoms can be diffuse, and include (bloody) diarrhea, abdominal discomfort, weight loss and anemia, and there may also be extra-intestinal symptoms such as arthritis, and eye and skin disorders. Complications such as strictures often occur in CD, and since the inflammation is transmural, fistulas and abscesses can develop, and these eventually require surgical treatment [2]. Most of the medications have significant side effects, and they are expensive, and often ineffective. CD is a major burden on healthcare services, with a prevalence of 100 to 150 cases per 100,000 persons per year in the western world and with a peak age of onset between 10 and 30 years of age [3]. CD is partly heritable; this is reflected in the higher concordance rate in monozygotic twins compared with dizygotic twins. The concordance for CD in dizygotic twins is 4%, and for monozygotic twins it is as high as 56% [4].

Prior to the introduction of genome-wide association studies (GWASs), only a few genetic factors (for example, *NOD2*, which encodes nucleotide binding oligomerization domain 2) had unequivocally been associated with CD. However, multiple GWASs have now been performed for CD, and a recent meta-analysis carried out by Franke *et al.* [5] has unveiled 71 genetic variants as associated with CD; Table 1 highlights some noteworthy genes from that study. Many of the genes cluster in several different molecular pathways and gene networks. In particular, results from GWASs have indicated the importance of the immune system in disease pathogenesis by identifying genes involved in innate and adaptive immunity. Hence, the association of *IRGM*, encoding immunity-related GTPase family M, and *ATG16L1*, encoding autophagy-related 16-like 1, with CD has implicated the process of autophagy [6]. The association of *NOD2*, *CARD9*, which encodes caspase recruitment domain family member 9, and *TLR4*, which encodes Toll-like receptor 4, indicates the involvement of pattern recognition mechanisms of the innate immune system [7]. Other genes are involved in pro-inflammatory pathways (T helper 1 cells and T helper 17 cells) and in anti-inflammatory pathways (regulatory T cells and IL-10),

[†] Equal contributors
[*] Correspondence: r.k.weersma@mdl.umcg.nl
[2] Department of Gastroenterology and Hepatology, University Medical Centre Groningen, University of Groningen, Groningen, the Netherlands
Full list of author information is available at the end of the article

BioMed Central © 2011 BioMed Central Ltd

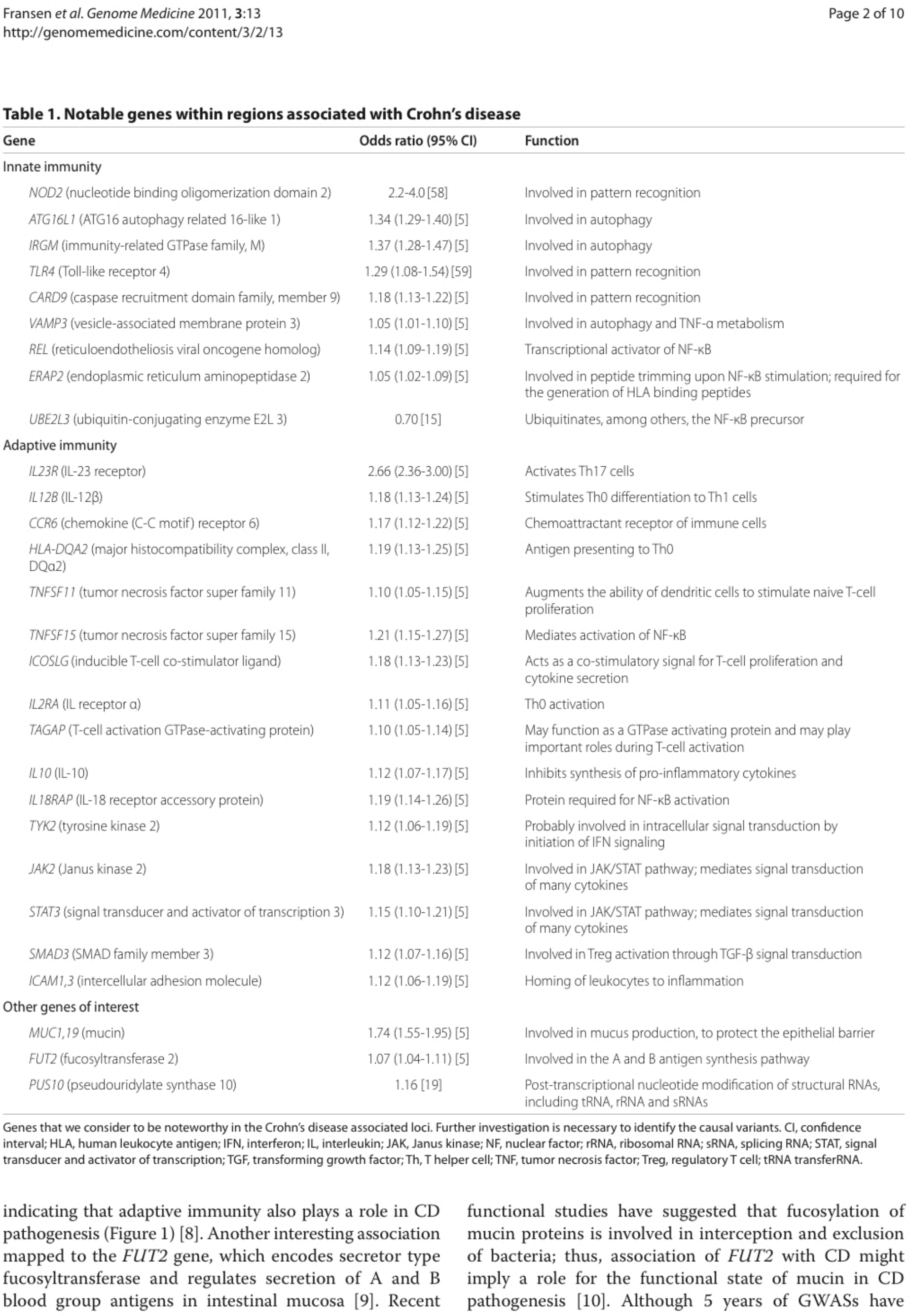


Table 1. Notable genes within regions associated with Crohn's disease

Gene	Odds ratio (95% CI)	Function
Innate immunity		
NOD2 (nucleotide binding oligomerization domain 2)	2.2-4.0 [58]	Involved in pattern recognition
ATG16L1 (ATG16 autophagy related 16-like 1)	1.34 (1.29-1.40) [5]	Involved in autophagy
IRGM (immunity-related GTPase family, M)	1.37 (1.28-1.47) [5]	Involved in autophagy
TLR4 (Toll-like receptor 4)	1.29 (1.08-1.54) [59]	Involved in pattern recognition
CARD9 (caspase recruitment domain family, member 9)	1.18 (1.13-1.22) [5]	Involved in pattern recognition
VAMP3 (vesicle-associated membrane protein 3)	1.05 (1.01-1.10) [5]	Involved in autophagy and TNF-α metabolism
REL (reticuloendotheliosis viral oncogene homolog)	1.14 (1.09-1.19) [5]	Transcriptional activator of NF-κB
ERAP2 (endoplasmic reticulum aminopeptidase 2)	1.05 (1.02-1.09) [5]	Involved in peptide trimming upon NF-κB stimulation; required for the generation of HLA binding peptides
UBE2L3 (ubiquitin-conjugating enzyme E2L 3)	0.70 [15]	Ubiquitinates, among others, the NF-κB precursor
Adaptive immunity		
IL23R (IL-23 receptor)	2.66 (2.36-3.00) [5]	Activates Th17 cells
IL12B (IL-12β)	1.18 (1.13-1.24) [5]	Stimulates Th0 differentiation to Th1 cells
CCR6 (chemokine (C-C motif) receptor 6)	1.17 (1.12-1.22) [5]	Chemoattractant receptor of immune cells
HLA-DQA2 (major histocompatibility complex, class II, DQα2)	1.19 (1.13-1.25) [5]	Antigen presenting to Th0
TNFSF11 (tumor necrosis factor super family 11)	1.10 (1.05-1.15) [5]	Augments the ability of dendritic cells to stimulate naive T-cell proliferation
TNFSF15 (tumor necrosis factor super family 15)	1.21 (1.15-1.27) [5]	Mediates activation of NF-κB
ICOSLG (inducible T-cell co-stimulator ligand)	1.18 (1.13-1.23) [5]	Acts as a co-stimulatory signal for T-cell proliferation and cytokine secretion
IL2RA (IL receptor α)	1.11 (1.05-1.16) [5]	Th0 activation
TAGAP (T-cell activation GTPase-activating protein)	1.10 (1.05-1.14) [5]	May function as a GTPase activating protein and may play important roles during T-cell activation
IL10 (IL-10)	1.12 (1.07-1.17) [5]	Inhibits synthesis of pro-inflammatory cytokines
IL18RAP (IL-18 receptor accessory protein)	1.19 (1.14-1.26) [5]	Protein required for NF-κB activation
TYK2 (tyrosine kinase 2)	1.12 (1.06-1.19) [5]	Probably involved in intracellular signal transduction by initiation of IFN signaling
JAK2 (Janus kinase 2)	1.18 (1.13-1.23) [5]	Involved in JAK/STAT pathway; mediates signal transduction of many cytokines
STAT3 (signal transducer and activator of transcription 3)	1.15 (1.10-1.21) [5]	Involved in JAK/STAT pathway; mediates signal transduction of many cytokines
SMAD3 (SMAD family member 3)	1.12 (1.07-1.16) [5]	Involved in Treg activation through TGF-β signal transduction
ICAM1,3 (intercellular adhesion molecule)	1.12 (1.06-1.19) [5]	Homing of leukocytes to inflammation
Other genes of interest		
MUC1,19 (mucin)	1.74 (1.55-1.95) [5]	Involved in mucus production, to protect the epithelial barrier
FUT2 (fucosyltransferase 2)	1.07 (1.04-1.11) [5]	Involved in the A and B antigen synthesis pathway
PUS10 (pseudouridylate synthase 10)	1.16 [19]	Post-transcriptional nucleotide modification of structural RNAs, including tRNA, rRNA and sRNAs

Genes that we consider to be noteworthy in the Crohn's disease associated loci. Further investigation is necessary to identify the causal variants. CI, confidence interval; HLA, human leukocyte antigen; IFN, interferon; IL, interleukin; JAK, Janus kinase; NF, nuclear factor; rRNA, ribosomal RNA; sRNA, splicing RNA; STAT, signal transducer and activator of transcription; TGF, transforming growth factor; Th, T helper cell; TNF, tumor necrosis factor; Treg, regulatory T cell; tRNA transferRNA.

indicating that adaptive immunity also plays a role in CD pathogenesis (Figure 1) [8]. Another interesting association mapped to the *FUT2* gene, which encodes secretor type fucosyltransferase and regulates secretion of A and B blood group antigens in intestinal mucosa [9]. Recent functional studies have suggested that fucosylation of mucin proteins is involved in interception and exclusion of bacteria; thus, association of *FUT2* with CD might imply a role for the functional state of mucin in CD pathogenesis [10]. Although 5 years of GWASs have

Fransen *et al. Genome Medicine* 2011, **3**:13
http://genomemedicine.com/content/3/2/13

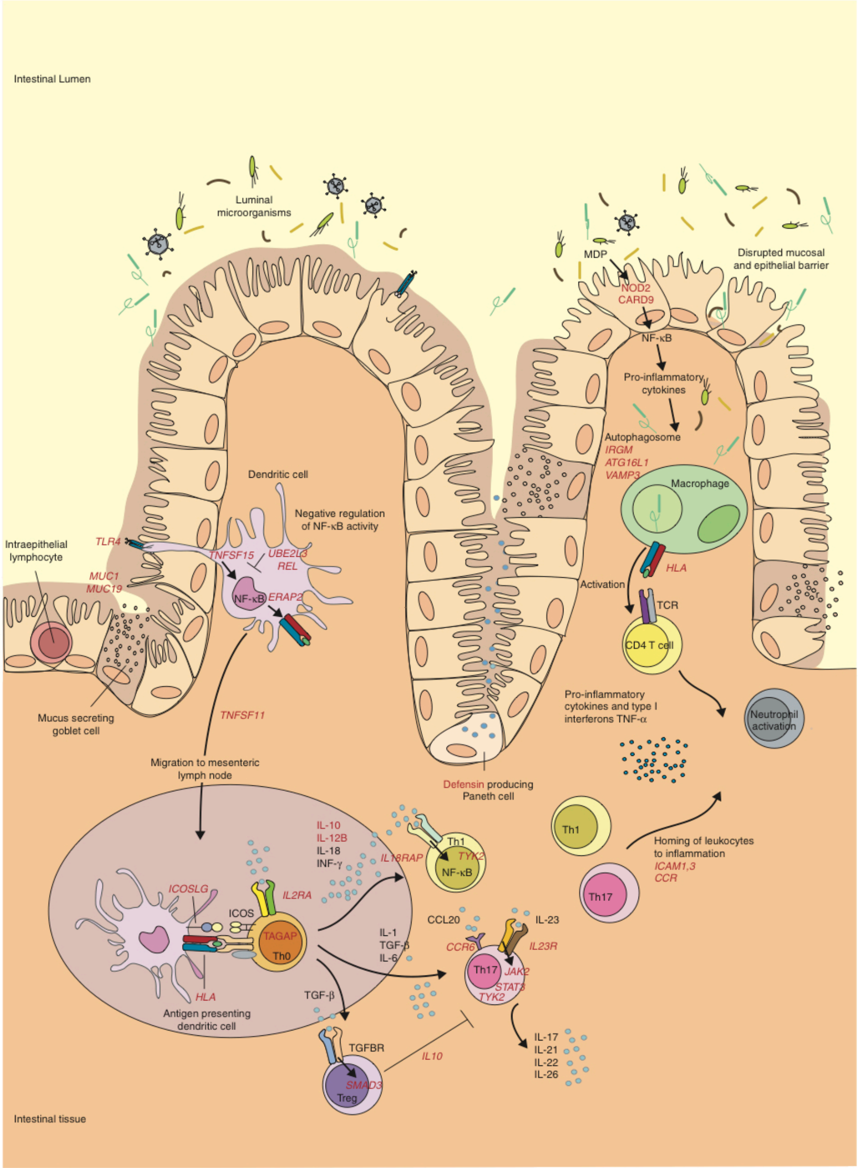


Figure 1. See next page for legend.

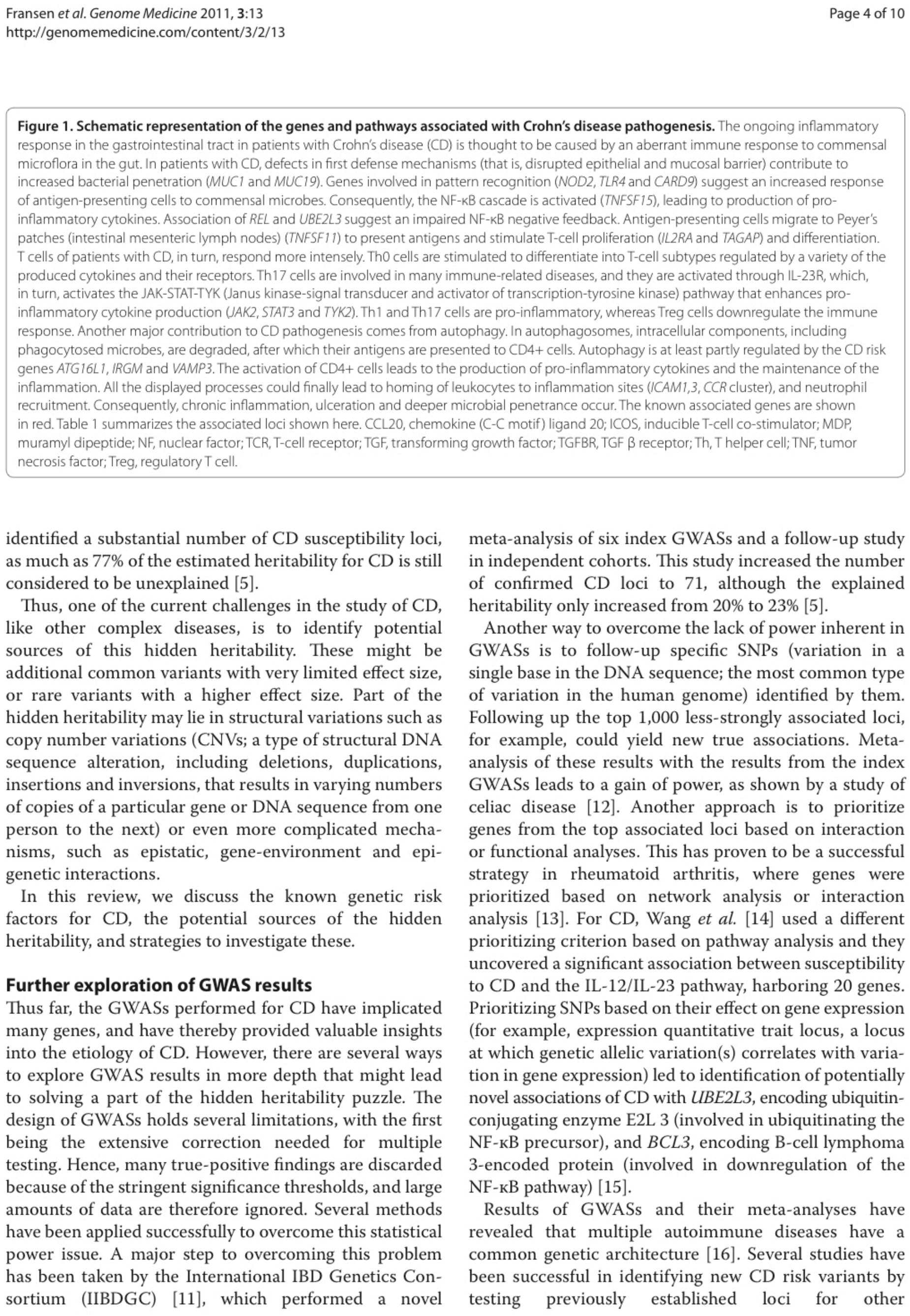


Figure 1. Schematic representation of the genes and pathways associated with Crohn's disease pathogenesis. The ongoing inflammatory response in the gastrointestinal tract in patients with Crohn's disease (CD) is thought to be caused by an aberrant immune response to commensal microflora in the gut. In patients with CD, defects in first defense mechanisms (that is, disrupted epithelial and mucosal barrier) contribute to increased bacterial penetration (*MUC1* and *MUC19*). Genes involved in pattern recognition (*NOD2*, *TLR4* and *CARD9*) suggest an increased response of antigen-presenting cells to commensal microbes. Consequently, the NF-κB cascade is activated (*TNFSF15*), leading to production of pro-inflammatory cytokines. Association of *REL* and *UBE2L3* suggest an impaired NF-κB negative feedback. Antigen-presenting cells migrate to Peyer's patches (intestinal mesenteric lymph nodes) (*TNFSF11*) to present antigens and stimulate T-cell proliferation (*IL2RA* and *TAGAP*) and differentiation. T cells of patients with CD, in turn, respond more intensely. Th0 cells are stimulated to differentiate into T-cell subtypes regulated by a variety of the produced cytokines and their receptors. Th17 cells are involved in many immune-related diseases, and they are activated through IL-23R, which, in turn, activates the JAK-STAT-TYK (Janus kinase-signal transducer and activator of transcription-tyrosine kinase) pathway that enhances pro-inflammatory cytokine production (*JAK2*, *STAT3* and *TYK2*). Th1 and Th17 cells are pro-inflammatory, whereas Treg cells downregulate the immune response. Another major contribution to CD pathogenesis comes from autophagy. In autophagosomes, intracellular components, including phagocytosed microbes, are degraded, after which their antigens are presented to CD4+ cells. Autophagy is at least partly regulated by the CD risk genes *ATG16L1*, *IRGM* and *VAMP3*. The activation of CD4+ cells leads to the production of pro-inflammatory cytokines and the maintenance of the inflammation. All the displayed processes could finally lead to homing of leukocytes to inflammation sites (*ICAM1,3*, *CCR* cluster), and neutrophil recruitment. Consequently, chronic inflammation, ulceration and deeper microbial penetrance occur. The known associated genes are shown in red. Table 1 summarizes the associated loci shown here. CCL20, chemokine (C-C motif) ligand 20; ICOS, inducible T-cell co-stimulator; MDP, muramyl dipeptide; NF, nuclear factor; TCR, T-cell receptor; TGF, transforming growth factor; TGFBR, TGF β receptor; Th, T helper cell; TNF, tumor necrosis factor; Treg, regulatory T cell.

identified a substantial number of CD susceptibility loci, as much as 77% of the estimated heritability for CD is still considered to be unexplained [5].

Thus, one of the current challenges in the study of CD, like other complex diseases, is to identify potential sources of this hidden heritability. These might be additional common variants with very limited effect size, or rare variants with a higher effect size. Part of the hidden heritability may lie in structural variations such as copy number variations (CNVs; a type of structural DNA sequence alteration, including deletions, duplications, insertions and inversions, that results in varying numbers of copies of a particular gene or DNA sequence from one person to the next) or even more complicated mechanisms, such as epistatic, gene-environment and epigenetic interactions.

In this review, we discuss the known genetic risk factors for CD, the potential sources of the hidden heritability, and strategies to investigate these.

Further exploration of GWAS results

Thus far, the GWASs performed for CD have implicated many genes, and have thereby provided valuable insights into the etiology of CD. However, there are several ways to explore GWAS results in more depth that might lead to solving a part of the hidden heritability puzzle. The design of GWASs holds several limitations, with the first being the extensive correction needed for multiple testing. Hence, many true-positive findings are discarded because of the stringent significance thresholds, and large amounts of data are therefore ignored. Several methods have been applied successfully to overcome this statistical power issue. A major step to overcoming this problem has been taken by the International IBD Genetics Consortium (IIBDGC) [11], which performed a novel meta-analysis of six index GWASs and a follow-up study in independent cohorts. This study increased the number of confirmed CD loci to 71, although the explained heritability only increased from 20% to 23% [5].

Another way to overcome the lack of power inherent in GWASs is to follow-up specific SNPs (variation in a single base in the DNA sequence; the most common type of variation in the human genome) identified by them. Following up the top 1,000 less-strongly associated loci, for example, could yield new true associations. Meta-analysis of these results with the results from the index GWASs leads to a gain of power, as shown by a study of celiac disease [12]. Another approach is to prioritize genes from the top associated loci based on interaction or functional analyses. This has proven to be a successful strategy in rheumatoid arthritis, where genes were prioritized based on network analysis or interaction analysis [13]. For CD, Wang *et al.* [14] used a different prioritizing criterion based on pathway analysis and they uncovered a significant association between susceptibility to CD and the IL-12/IL-23 pathway, harboring 20 genes. Prioritizing SNPs based on their effect on gene expression (for example, expression quantitative trait locus, a locus at which genetic allelic variation(s) correlates with variation in gene expression) led to identification of potentially novel associations of CD with *UBE2L3*, encoding ubiquitin-conjugating enzyme E2L 3 (involved in ubiquitinating the NF-κB precursor), and *BCL3*, encoding B-cell lymphoma 3-encoded protein (involved in downregulation of the NF-κB pathway) [15].

Results of GWASs and their meta-analyses have revealed that multiple autoimmune diseases have a common genetic architecture [16]. Several studies have been successful in identifying new CD risk variants by testing previously established loci for other immune-related diseases [17,18]. Festen *et al.* [19] developed a new method to identify shared risk loci of two immune-mediated diseases with a partially shared genetic background, namely celiac disease and CD. To increase the statistical power, they performed a combined analysis of GWAS results from celiac disease and CD, and identified *TAGAP*, which encodes T-cell activation GTPase-activating protein, and *PUS-10*, which encodes tRNA pseudouridylate synthase, as new shared loci [19].

The second limitation of the GWAS design is that it does not lead to the identification of causal variants, since the tested SNPs are merely tagging SNPs in linkage disequilibrium (LD; a non-random association of alleles at two or more loci as a result of a recent mutation, genetic drift, selection, or non-random mating) with the causal variants. Therefore, the effect sizes of known CD loci may be an underestimation of their actual relative risk. To further investigate the known risk loci and identify new SNPs, either as causal or close-to-causal variants, extensive fine-mapping is currently being performed by the IIBDGC using a custom-made GWA chip. In addition, cross-ethnicity fine-mapping has proven successful in exploring conserved haplotype structures (that is, LD blocks) [20]. The most common LD blocks occur in all populations; however, their frequencies vary among different ethnicities [20]. For example, common *NOD2* and *IL23R* variants that are well established in Caucasians could not be replicated in an Indian population, implying that additional variants in these or other candidate genes may play a role in the pathogenesis of CD in Indians [21]. This principle was also successfully applied in analyzing the *IL2*/*IL21* LD block, which is strongly conserved in Caucasians as opposed to Han Chinese, in which the *IL2* and *IL21* genes reside on two distinct LD blocks. Both *IL2* and *IL21* could be identified as separate UC risk loci in Han Chinese [22].

Park *et al.* [23] proposed a method to evaluate statistical power and risk prediction of future GWASs. They estimated that there are, in total, 142 CD susceptibility loci with effect sizes similar to the loci reported in the current GWASs, and that a sample size of approximately 50,000 would be needed to uncover them. However, even if a GWAS with hundreds of thousands of cases were to provide new CD susceptibility loci and explain more of the genetic variance, it seems unlikely that it would capture even half of the estimated heritability since 142 loci only explain 20% of the sibling relative risk for CD. We can speculate that identification of the true causal variants could amplify the effect size for some of the known loci and could consequently increase the discriminatory power of risk models.

Another potential source of hidden heritability could lie in sample mix-ups that occur accidentally during sample collection, genotyping or data management. Some genetic variants influence gene expression phenotypes (expression quantitative trait loci); this allows checking for concordance between phenotypic measurements and genetic variants that affect these phenotypes. Westra *et al.* (personal communication) found that 3% of sample mix-ups decrease the number of loci normally discovered by 23% for a trait with a heritability of 50% and 500 loci explaining the total heritability. Thus, sample mix-ups may explain part of the hidden heritability and it will be possible to detect them as long as databases encompass sufficient numbers of phenotypes that are strongly determined by known genetic variants.

GWASs are most likely to remain an important approach for investigating the hidden heritability, since the potential of their results can be enhanced by: performing meta-analyses (for example, between multiple GWASs or between similar disease phenotypes); following-up prioritized SNPs based on pathway, functional or interaction analyses; studying SNPs that have been associated with other immune-related diseases; and expanding the design of GWASs to include samples from non-Caucasians.

Low frequency and rare variants

Common variants identified by GWASs represent only a small fraction of the phenotypic variation. Thus, much speculation about the hidden heritability has focused on the contribution of variants with low allele frequencies, defined as 0.5% < minor allele frequency (MAF; proportion of the less common of two alleles in a population) < 5%, or from rare variants with MAF <0.5%, that are not sufficiently frequent to be captured by current GWA arrays, nor sufficiently penetrant to be captured by traditional, family-based linkage studies [24]. Detecting such variants will be facilitated by advances in high-throughput sequencing technologies and by the wide-ranging catalog of variants with MAF >1% generated by the 1000 Genomes Project [25]. Current efforts to identify rare variants by sequencing are likely to focus on the regions of most significant GWAS SNPs and around genes already implicated in CD pathogenesis or treatment. Resequencing of selected susceptibility loci has led recently to the discovery of three *IL23R* (the gene encoding IL-23 receptor) coding variants that offer protection against CD [26]. The results of this particular study confirmed an increase in effect size with decreasing variant frequency, although rare variants explained less of the heritability than common variants.

In addition to resequencing efforts, whole-genome/exome sequencing will be needed to detect rare high-risk variants beyond the LD reach of tag SNPs. Although the costs of next-generation sequencing remain high, they are dropping fairly rapidly as the technologies improve and the process time per sample is becoming shorter; so this method is becoming more and more feasible and accessible for researchers. Evaluating such signals and determining the real causal variant will, however, be a difficult task. Feng and Zhu [27] developed an alternative method for searching for rare variants in previously published GWAS datasets. Their method relies on haplotype analysis across the genome and the hypothesis that multiple rare variants can be captured by many haplotypes. Using this method, they confirmed nine previously established loci and also discovered four new CD susceptibility loci [27].

Another approach that may prove to be important is performing resequencing studies of individuals with extreme phenotypes in lipid levels; these studies have shown that such individuals seem more likely to be the carriers of rare, yet non-synonymous, variants [28]. A large number of rare variants may have distinct effects on the phenotype. Therefore, pooling variants of similar effect and locus-specific matching of cases with specific CD subphenotypes and controls throughout the genome may help to reveal some of the hidden heritability [29].

Structural variation

It has been estimated that chromosomal rearrangements (that is, duplications, deletions, insertions and inversions), collectively named CNVs, comprise 12% of the human genome [30]. Currently, more than 15,000 CNV loci are catalogued in the Database of Genomic Variants [31]. Some CNVs have been linked to complex disorders, such as autism, neuroblastoma and systematic lupus erythematosus [32-34]. A recent study suggested that CNVs are enriched in genomic regions containing genes that influence immunity [35]. In particular, low and high copy numbers of the β-defensin gene (*HBD2*), which acts as an antimicrobial peptide and as a cytokine, have been found to predispose to colonic CD [36,37]. Yet, in a recent study, Aldhous *et al.* [38] failed to replicate both of the previously published associations. Moreover, they argued that these two associations could be due to measurement error because of a general deficiency of real-time PCR to distinguish multiple CNV clusters. In addition to the β-defensins, a fine-mapping study of the *IRGM* susceptibility locus revealed a 20-kb deletion polymorphism immediately upstream of *IRGM* that was associated with CD risk and *IRGM* expression [39]. Furthermore, a recent GWAS of CNVs from the Wellcome Trust Case Control Consortium has confirmed these CNVs for CD, and also discovered new CNVs in the *IRGM* and human leukocyte antigen (5.1 kb) regions [40]. The Wellcome Trust Case Control Consortium study also showed that the most common CNVs are well tagged by SNPs in current GWAS chips, and that they are unlikely to make much contribution to the hidden heritability in common diseases. More work is needed to elucidate the functional consequences and impact of high copy-number repeats (for example, long interspersed nuclear elements), and of rare CNVs on clinical phenotypes, such as CD.

Family-based approaches

Since the possibility of chip-based GWASs became available, linkage analysis and family-based approaches have been largely discarded. However, now that the opportunities for gene detection by conventional GWASs have been almost exhausted, researchers are shifting back towards family-based approaches. These approaches can be helpful when GWASs fail to detect signals from rare variants and are biased by population stratification, which is defined as a presence of subpopulations in a supposedly homogeneous population. Subpopulations arise from differences in allele frequencies between individuals as a consequence of distinct ancestral and/or demographic origin. Family-based studies may also be advantageous since the low frequency risk alleles (SNPs with MAF <5%) are likely to be more prevalent in large families with several affected members and should therefore be easier to detect. By assessing GWAS data in such families, large regions of identity-by-descent may be identified and found to include genes associated with CD; this approach has already proved to be a powerful tool in classical linkage analysis. However, the shared environment of family members is an alternative explanation for familial clustering that should be taken into account. Glocker *et al.* [41] identified loss-of-function mutations in two loci by considering early onset colitis as a monogenic trait in two consanguineous families. They performed a genetic linkage analysis followed by candidate gene sequencing and identified the *IL10RA* (the gene encoding IL-10 receptor α) and *IL10RB* (the gene encoding IL-10 receptor β) loci as being associated with early-onset enterocolitis. However, it is most likely that in this particular case a private variant, not present in the general population, is responsible for the disease.

Akolkar *et al.* [42] found that CD is subject to a parent-of-origin effect, indicating that loci affected by genomic imprinting play a role in CD pathogenesis. In genomic imprinting, the expression of an inherited variant is determined by the parent from whom that variant is inherited. If the maternal allele, for instance, is inactivated by genomic imprinting, then expression of the locus is determined by the paternal allele only. If this effect is not taken into account, a significant loss in the statistical power of the study might develop [43].

Family-based approaches may be useful in the search for the hidden heritability since low-frequency variants accumulate in families with multiple affected individuals; moreover, low-frequency variants are not affected by population stratification and they also include parent-of-origin effects. However, the causal variants identified in such families may prove to be private variants or the shared environment may play a major role.

GWAS aftermath: epistatic, gene-environment and epigenetic interactions

Given that a large proportion of the heritability of CD and its complex architecture is as yet unexplained, one might speculate other aspects of inheritance, such as epistasis, gene-environment interactions or epigenetic effects, might be involved. GWASs may be missing higher-order genetic effects that arise from the interaction of two or more SNPs [44]. The underlying idea for such epistatic effects is that a significant proportion of the hidden heritability is not due to single common variants, nor to single rare variants, but rather to rare combinations of common variants. Since typical GWASs examine the association of single SNPs with a phenotype, SNPs that contribute epistatically will not be revealed by such an analysis. A recent pair-wise analysis of variants related to the *IL17-IL23* pathway showed an increasing odds ratio for CD when the 'risk' haplotypes for these genes were combined [45]. Analysis of epistatic interactions in better-powered datasets, and the use of more efficient computational approaches that can account for the complex nature of biomolecular networks, may yield new genetic risk factors for CD [46,47].

An even more complex source for the hidden heritability might lie in gene-environment interactions, which are defined as the joint effect of one or more genes with one or more environmental factors that cannot be readily explained by their separate marginal effects [48]. The strongest and best replicated environmental risk factor for CD is smoking, which increases both the risk and severity of CD. However, a recent, moderately sized study found remarkable differences in associated loci between smoking and non-smoking CD patients, thereby implying that a complex gene-environment interaction must be at work [49]. Another example of the complex interaction between genetic and environmental factors was shown in a study by Cadwell *et al.* [50] where *Atg16L1*-deficient mice infected with a specific strain of norovirus developed CD-like phenotypes in a model of intestinal injury induced by dextran sodium sulfate. In particular, structural Paneth cell abnormalities and decreased production of antimicrobial granules in the mice resembled those found in CD patients who are homozygous carriers of the *ATG16L1* risk alleles. Remarkably, the severity of intestinal injury induced by dextran sodium sulfate was not only dependent on aberrant *Atg16L1* function and norovirus infection, but also on the timing of infection, secretion of the pro-inflammatory cytokines TNF-α and IFN-γ, and the presence of commensal bacteria in the mouse intestine.

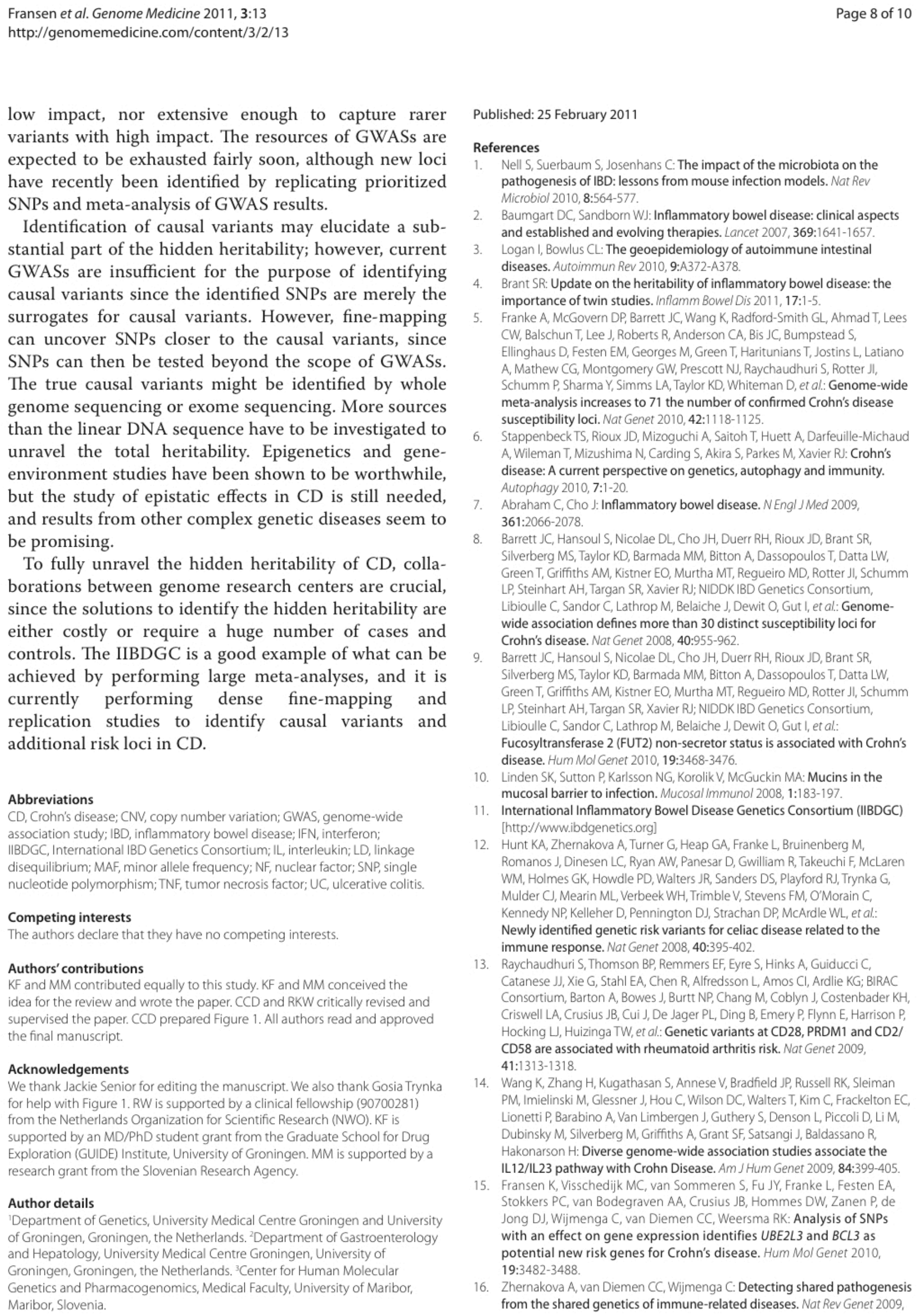
Other environmental factors, such as appendectomy, diet and domestic hygiene habits, may also play a role in CD, but the evidence for each of these factors is much weaker. To study gene-environment interactions will require careful consideration of the epidemiologic study design, exposure assessment, and methods of analysis, paying particular attention to ways of harmonizing these features across consortia.

An additional source of the hidden heritability might not lie in the genome sequence itself, but in subtle mechanisms interfering with genome functions, such as gene expression. These mechanisms include histone modification, methylation and gene inactivation, and are covered by the study of epigenetics. However, there is much controversy on this topic. Its role in CD is unknown, but there are some hints that methylation plays a role in other complex diseases: type 2 diabetes, rheumatoid arthritis and neurodegenerative diseases [51-53]. Epigenetics is also correlated with age, gender and nutrition, and it is likely that there are other environmental factors to be discovered [54,55]. It has been shown that changes in DNA methylation in mice can be provoked by dietary alterations and subsequently transmitted across generations [56]. Thus, sequence-independent epigenetic effects (beyond imprinting) that might be environmentally induced and transmitted across several generations [57] could represent a revolutionary glimpse into the enigmatic world of the heritability of complex diseases.

Conclusions

CD is a complex genetic disorder with an estimated heritability of 50% and it is characterized by a recurring inflammation of the gastrointestinal tract. Two decades of research have led to the discovery of 71 risk loci, which have improved our understanding of the disease pathogenesis. At the moment, approximately 23% of the heritability can be explained. To fully understand the disease pathogenesis and link current insights to clinically relevant knowledge, it is important to continue our quest to identify more genetic risk factors in CD. In this review, we have presented various potential sources for the hidden heritability of complex diseases given the current knowledge on CD.

It is unlikely that conventional GWASs alone can solve the puzzle of the hidden heritability. They are not powerful enough to detect signals from common variants with



low impact, nor extensive enough to capture rarer variants with high impact. The resources of GWASs are expected to be exhausted fairly soon, although new loci have recently been identified by replicating prioritized SNPs and meta-analysis of GWAS results.

Identification of causal variants may elucidate a substantial part of the hidden heritability; however, current GWASs are insufficient for the purpose of identifying causal variants since the identified SNPs are merely the surrogates for causal variants. However, fine-mapping can uncover SNPs closer to the causal variants, since SNPs can then be tested beyond the scope of GWASs. The true causal variants might be identified by whole genome sequencing or exome sequencing. More sources than the linear DNA sequence have to be investigated to unravel the total heritability. Epigenetics and gene-environment studies have been shown to be worthwhile, but the study of epistatic effects in CD is still needed, and results from other complex genetic diseases seem to be promising.

To fully unravel the hidden heritability of CD, collaborations between genome research centers are crucial, since the solutions to identify the hidden heritability are either costly or require a huge number of cases and controls. The IIBDGC is a good example of what can be achieved by performing large meta-analyses, and it is currently performing dense fine-mapping and replication studies to identify causal variants and additional risk loci in CD.

Abbreviations

CD, Crohn's disease; CNV, copy number variation; GWAS, genome-wide association study; IBD, inflammatory bowel disease; IFN, interferon; IIBDGC, International IBD Genetics Consortium; IL, interleukin; LD, linkage disequilibrium; MAF, minor allele frequency; NF, nuclear factor; SNP, single nucleotide polymorphism; TNF, tumor necrosis factor; UC, ulcerative colitis.

Competing interests

The authors declare that they have no competing interests.

Authors' contributions

KF and MM contributed equally to this study. KF and MM conceived the idea for the review and wrote the paper. CCD and RKW critically revised and supervised the paper. CCD prepared Figure 1. All authors read and approved the final manuscript.

Acknowledgements

We thank Jackie Senior for editing the manuscript. We also thank Gosia Trynka for help with Figure 1. RW is supported by a clinical fellowship (90700281) from the Netherlands Organization for Scientific Research (NWO). KF is supported by an MD/PhD student grant from the Graduate School for Drug Exploration (GUIDE) Institute, University of Groningen. MM is supported by a research grant from the Slovenian Research Agency.

Author details

[1]Department of Genetics, University Medical Centre Groningen and University of Groningen, Groningen, the Netherlands. [2]Department of Gastroenterology and Hepatology, University Medical Centre Groningen, University of Groningen, Groningen, the Netherlands. [3]Center for Human Molecular Genetics and Pharmacogenomics, Medical Faculty, University of Maribor, Maribor, Slovenia.

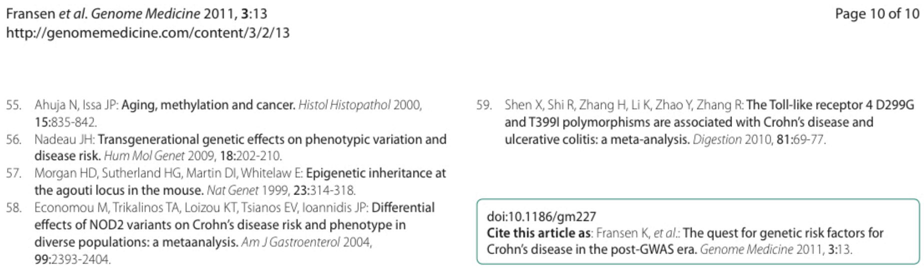
Published: 25 February 2011

References

1. Nell S, Suerbaum S, Josenhans C: **The impact of the microbiota on the pathogenesis of IBD: lessons from mouse infection models.** *Nat Rev Microbiol* 2010, **8**:564-577.
2. Baumgart DC, Sandborn WJ: **Inflammatory bowel disease: clinical aspects and established and evolving therapies.** *Lancet* 2007, **369**:1641-1657.
3. Logan I, Bowlus CL: **The geoepidemiology of autoimmune intestinal diseases.** *Autoimmun Rev* 2010, **9**:A372-A378.
4. Brant SR: **Update on the heritability of inflammatory bowel disease: the importance of twin studies.** *Inflamm Bowel Dis* 2011, **17**:1-5.
5. Franke A, McGovern DP, Barrett JC, Wang K, Radford-Smith GL, Ahmad T, Lees CW, Balschun T, Lee J, Roberts R, Anderson CA, Bis JC, Bumpstead S, Ellinghaus D, Festen EM, Georges M, Green T, Haritunians T, Jostins L, Latiano A, Mathew CG, Montgomery GW, Prescott NJ, Raychaudhuri S, Rotter JI, Schumm P, Sharma Y, Simms LA, Taylor KD, Whiteman D, *et al.*: **Genome-wide meta-analysis increases to 71 the number of confirmed Crohn's disease susceptibility loci.** *Nat Genet* 2010, **42**:1118-1125.
6. Stappenbeck TS, Rioux JD, Mizoguchi A, Saitoh T, Huett A, Darfeuille-Michaud A, Wileman T, Mizushima N, Carding S, Akira S, Parkes M, Xavier RJ: **Crohn's disease: A current perspective on genetics, autophagy and immunity.** *Autophagy* 2010, **7**:1-20.
7. Abraham C, Cho J: **Inflammatory bowel disease.** *N Engl J Med* 2009, **361**:2066-2078.
8. Barrett JC, Hansoul S, Nicolae DL, Cho JH, Duerr RH, Rioux JD, Brant SR, Silverberg MS, Taylor KD, Barmada MM, Bitton A, Dassopoulos T, Datta LW, Green T, Griffiths AM, Kistner EO, Murtha MT, Regueiro MD, Rotter JI, Schumm LP, Steinhart AH, Targan SR, Xavier RJ; NIDDK IBD Genetics Consortium, Libioulle C, Sandor C, Lathrop M, Belaiche J, Dewit O, Gut I, *et al.*: **Genome-wide association defines more than 30 distinct susceptibility loci for Crohn's disease.** *Nat Genet* 2008, **40**:955-962.
9. Barrett JC, Hansoul S, Nicolae DL, Cho JH, Duerr RH, Rioux JD, Brant SR, Silverberg MS, Taylor KD, Barmada MM, Bitton A, Dassopoulos T, Datta LW, Green T, Griffiths AM, Kistner EO, Murtha MT, Regueiro MD, Rotter JI, Schumm LP, Steinhart AH, Targan SR, Xavier RJ; NIDDK IBD Genetics Consortium, Libioulle C, Sandor C, Lathrop M, Belaiche J, Dewit O, Gut I, *et al.*: **Fucosyltransferase 2 (FUT2) non-secretor status is associated with Crohn's disease.** *Hum Mol Genet* 2010, **19**:3468-3476.
10. Linden SK, Sutton P, Karlsson NG, Korolik V, McGuckin MA: **Mucins in the mucosal barrier to infection.** *Mucosal Immunol* 2008, **1**:183-197.
11. **International Inflammatory Bowel Disease Genetics Consortium (IIBDGC)** [http://www.ibdgenetics.org]
12. Hunt KA, Zhernakova A, Turner G, Heap GA, Franke L, Bruinenberg M, Romanos J, Dinesen LC, Ryan AW, Panesar D, Gwilliam R, Takeuchi F, McLaren WM, Holmes GK, Howdle PD, Walters JR, Sanders DS, Playford RJ, Trynka G, Mulder CJ, Mearin ML, Verbeek WH, Trimble V, Stevens FM, O'Morain C, Kennedy NP, Kelleher D, Pennington DJ, Strachan DP, McArdle WL, *et al.*: **Newly identified genetic risk variants for celiac disease related to the immune response.** *Nat Genet* 2008, **40**:395-402.
13. Raychaudhuri S, Thomson BP, Remmers EF, Eyre S, Hinks A, Guiducci C, Catanese JJ, Xie G, Stahl EA, Chen R, Alfredsson L, Amos CI, Ardlie KG; BIRAC Consortium, Barton A, Bowes J, Burtt NP, Chang M, Coblyn J, Costenbader KH, Criswell LA, Crusius JB, Cui J, De Jager PL, Ding B, Emery P, Flynn E, Harrison P, Hocking LJ, Huizinga TW, *et al.*: **Genetic variants at CD28, PRDM1 and CD2/CD58 are associated with rheumatoid arthritis risk.** *Nat Genet* 2009, **41**:1313-1318.
14. Wang K, Zhang H, Kugathasan S, Annese V, Bradfield JP, Russell RK, Sleiman PM, Imielinski M, Glessner J, Hou C, Wilson DC, Walters T, Kim C, Frackelton EC, Lionetti P, Barabino A, Van Limbergen J, Guthery S, Denson L, Piccoli D, Li M, Dubinsky M, Silverberg M, Griffiths A, Grant SF, Satsangi J, Baldassano R, Hakonarson H: **Diverse genome-wide association studies associate the IL12/IL23 pathway with Crohn Disease.** *Am J Hum Genet* 2009, **84**:399-405.
15. Fransen K, Visschedijk MC, van Sommeren S, Fu JY, Franke L, Festen EA, Stokkers PC, van Bodegraven AA, Crusius JB, Hommes DW, Zanen P, de Jong DJ, Wijmenga C, van Diemen CC, Weersma RK: **Analysis of SNPs with an effect on gene expression identifies *UBE2L3* and *BCL3* as potential new risk genes for Crohn's disease.** *Hum Mol Genet* 2010, **19**:3482-3488.
16. Zhernakova A, van Diemen CC, Wijmenga C: **Detecting shared pathogenesis from the shared genetics of immune-related diseases.** *Nat Rev Genet* 2009,

10:43-55.

17. Wang K, Baldassano R, Zhang H, Qu HQ, Imielinski M, Kugathasan S, Annese V, Dubinsky M, Rotter JI, Russell RK, Bradfield JP, Sleiman PM, Glessner JT, Walters T, Hou C, Kim C, Frackelton EC, Garris M, Doran J, Romano C, Catassi C, Van Limbergen J, Guthery SL, Denson L, Piccoli D, Silverberg MS, Stanley CA, Monos D, Wilson DC, Griffiths A, *et al.*: **Comparative genetic analysis of inflammatory bowel disease and type 1 diabetes implicates multiple loci with opposite effect.** *Hum Mol Gen* 2010, **19:**2059-2067.
18. Danoy P, Pryce K, Hadler J, Bradbury LA, Farrar C, Pointon J; Australo-Anglo-American Spondyloarthritis Consortium, Ward M, Weisman M, Reveille JD, Wordsworth BP, Stone MA; Spondyloarthritis Research Consortium of Canada, Maksymowych WP, Rahman P, Gladman D, Inman RD, Brown MA: **Association of variants at 1q32 and *STAT3* with Ankylosing Spondilitis suggests genetic overlap with Crohn's disease.** *PLoS Genet* 2010, **6:**e1001195.
19. Festen EAM, Goyette P, Green T, Beauchamp C, Boucher G, Trynka G: **A meta-analysis of genome wide association scans identifies TAGAP and PUS10 as shared risk loci for Crohn's disease and celiac disease.** *PLoS Genet* 2011, **7:**e1001283.
20. Gabriel SB, Schaffner SF, Nguyen H, Moore JM, Roy J, Blumenstiel B, Higgins J, DeFelice M, Lochner A, Faggart M, Liu-Cordero SN, Rotimi C, Adeyemo A, Cooper R, Ward R, Lander ES, Daly MJ, Altshuler D: **The structure of haplotype blocks in the human genome.** *Science* 2002, **296:**2225-2229.
21. Mahurkar S, Banerjee R, Rani SV, Thakur N, Guduru VR, Duvvuru NR, Chandak GR: **Common variants in *NOD2* and *IL23R* are not associated with inflammatory bowel disease in Indian patients.** *J Gastroenterol Hepatol* 2010, in press. doi: 10.1111/j.1440-1746.2010.06533.x
22. Shi J, Lu Z, Zhernakova A, Qian J, Zhu F, Sun G, Zhu L, Ma X, Dijkstra G Wijmenga C, Faber KN, Lu X, Weersma RK: **Haplotype-based analysis of ulcerative colitis risk loci identifies both *IL2* and *IL21* as susceptibility genes in Han Chinese.** *Inflamm Bowel Dis* 2010, in press.
23. Park JH, Wacholder S, Gail MH, Peters U, Jacobs KB, Chanock SJ, Chatterjee N: **Estimation of effect size distribution from genome-wide association studies and implications for future discoveries.** *Nat Genet* 2010, **42:**570-575.
24. Manolio TA, Collins FS, Cox NJ, Goldstein DB, Hindorff LA, Hunter DJ, McCarthy MI, Ramos EM, Cardon LR, Chakravarti A, Cho JH, Guttmacher AE, Kong A, Kruglyak L, Mardis E, Rotimi CN, Slatkin M, Valle D, Whittemore AS, Boehnke M, Clark AG, Eichler EE, Gibson G, Haines JL, Mackay TF, McCarroll SA, Visscher PM: **Finding the missing heritability of complex disease.** *Nature* 2009, **461:**747-753.
25. **1000 Genomes - A Deep Catalog of Human Genetic Variation** [http://www.1000genomes.org]
26. Momozawa Y, Mni M, Nakamura K, Coppieters W, Almer S, Amininejad L: **Resequencing of positional candidates identifies low frequency *IL23R* coding variants protecting against inflammatory bowel disease.** *Nat Genet* 2011, **43:**43-47.
27. Feng T, Zhu X: **Genome-wide searching of rare genetic variants in WTCCC data.** *Hum Genet* 2010, **128:**269-280.
28. Romeo S, Pennacchio LA, Fu Y, Boerwinkle E, Tybjaerg-Hansen A, Hobbs HH, Cohen JC: **Population-based resequencing of *ANGPTL4* uncovers variations that reduce triglycerides and increase HDL.** *Nat Genet* 2007, **39:**513-516.
29. Li B, Leal SM: **Methods for detecting associations with rare variants for common diseases: application to analysis of sequence data.** *Am J Hum Genet* 2008, **83:**311-321.
30. Redon R, Ishikawa S, Fitch KR, Feuk L, Perry GH, Andrews TD, Fiegler H, Shapero MH, Carson AR, Chen W, Cho EK, Dallaire S, Freeman JL, González JR, Gratacòs M, Huang J, Kalaitzopoulos D, Komura D, MacDonald JR, Marshall CR, Mei R, Montgomery L, Nishimura K, Okamura K, Shen F, Somerville MJ, Tchinda J, Valsesia A, Woodwark C, Yang F, *et al.*: **Global variation in copy number in the human genome.** *Nature* 2006, **444:**444-454.
31. **Database of Genomic Variants - a curated catalogue of structural variation in the human genome** [http://projects.tcag.ca/variation/]
32. Glessner JT, Wang K, Cai G, Korvatska O, Kim CE, Wood S, Zhang H, Estes A, Brune CW, Bradfield JP, Imielinski M, Frackelton EC, Reichert J, Crawford EL, Munson J, Sleiman PM, Chiavacci R, Annaiah K, Thomas K, Hou C, Glaberson W, Flory J, Otieno F, Garris M, Soorya L, Klei L, Piven J, Meyer KJ, Anagnostou E, Sakurai T, *et al.*: **Autism genome-wide copy number variation reveals ubiquitin and neuronal genes.** *Nature* 2009, **459:**569-573.
33. Diskin SJ, Hou C, Glessner JT, Attiyeh EF, Laudenslager M, Bosse K, Cole K, Mossé YP, Wood A, Lynch JE, Pecor K, Diamond M, Winter C, Wang K, Kim C, Geiger EA, McGrady PW, Blakemore AIF, London WB, Shaikh TH, Bradfield J, Grant SFA, Li H, Devoto M, Rappaport ER, Hakonarson H, Maris JM: **Copy number variation at 1q21.1 associated with neuroblastoma.** *Nature* 2009, **459:**987-991.
34. Willcocks LC, Lyons PA, Clatworthy MR, Robinson JI, Yang W, Newland SA, Plagnol V, McGovern NN, Condliffe AM, Chilvers ER, Adu D, Jolly EC, Watts R, Lau YL, Morgan AW, Nash G, Smith KG: **Copy number of *FCGR3B*, which is associated with systematic lupus erythematosus, correlates with protein expression and immune complex uptake.** *J Exp Med* 2008, **205:**1573-1582.
35. Schaschl H, Aitman TJ, Vyse TJ: **Copy number variation in the human genome and its implication in autoimmunity.** *Clin Exp Immunol* 2009, **156:**12-16.
36. Fellermann K, Stange DE, Schaeffeler E, Schmalzl H, Wehkamp J, Bevins CL, Reinisch W, Teml A, Schwab M, Lichter P, Radlwimmer B, Stange EF: **A chromosome 8 gene-cluster polymorphism with low human β-defensin 2 gene copy number predisposes to Crohn's disease of the colon.** *Am J Hum Genet* 2006, **79:**439-448.
37. Bentley R, Pearson J, Gearry R, Barclay M, McKinney C, Merriman T, Roberts R: **Association of higher *DEFB4* genomic copy number with Crohn's disease.** *Am J Gastroenterol* 2010, **105:**354-359.
38. Aldhous MC, Abu Bakar S, Prescott NJ, Palla R, Soo K, Mansfield JC, Mathew CG, Satsangi J, Armour JA: **Measurement methods and accuracy in copy number variation: failure to replicate associations of beta-defensin copy number with Crohn's disease.** *Hum Mol Gen* 2010, **19:**4930-4938.
39. McCarroll SA, Huett A, Kuballa P, Chilewski SD, Landry A, Goyette P, Zody MC, Hall JL, Brant SR, Cho JH, Duerr RH, Silverberg MS, Taylor KD, Rioux JD, Altshuler D, Daly MJ, Xavier RJ: **Deletion polymorphism upstream of *IRGM* associated with altered *IRGM* expression and Crohn's disease.** *Nat Genet* 2008, **40:**1107-1112.
40. The Wellcome Trust Consortium: **Genome-wide association study of CNVs in 16,000 cases of eight common diseases and 3,000 shared controls.** *Nature* 2010, **464:**713-720.
41. Glocker EO, Kotlarz D, Boztug K, Gertz EM, Schäffer AA, Noyan F, Perro M, Diestelhorst J, Allroth A, Murugan D, Hätscher N, Pfeifer D, Sykora KW, Sauer M, Kreipe H, Lacher M, Nustede R, Woellner C, Baumann U, Salzer U, Koletzko S, Shah N, Segal AW, Sauerbrey A, Buderus S, Snapper SB, Grimbacher B, Klein C: **Inflammatory bowel disease and mutations affecting the interleukin-10 receptor.** *N Engl J Med* 2009, **361:**2033-2045.
42. Akolkar PN, Gulwani-Akolkar B, Heresbach D, Lin XY, Fisher S, Katz S, Silver J: **Differences in risk of Crohn's disease in offspring of mothers and fathers with inflammatory bowel disease.** *Am J Gastroenterol* 1997, **92:**2241-2244.
43. Hanson RL, Kobes S, Lindsay RS, Knowler WC: **Assessment of parent-of-origin effects in linkage analysis of quantitative traits.** *Am J Hum Genet* 2001 **68:**951-962.
44. Moore JH, Williams SM: **Epistasis and its implications for personal genetics.** *Am J Hum Genet* 2009, **85:**309-320.
45. McGovern DP, Rotter JI, Mei L, Haritunians T, Landers C, Derkowski C, Dutridge D, Dubinsky M, Ippoliti A, Vasiliauskas E, Mengesha E, King L, Pressman S, Targan SR, Taylor KD: **Genetic epistasis of IL23/IL17 related genes in Crohn's disease.** *Inflamm Bowel Dis* 2009, **15:**883-889.
46. Marchini J, Donnelly P, Cardon LR: **Genome-wide strategies for detecting multiple loci that influence complex diseases.** *Nat Genet* 2005, **37:**413-417.
47. Cordell HJ: **Detecting gene-gene interactions that underlie human diseases.** *Nat Rev Genet* 2009, **10:**392-404.
48. Thomas D: **Gene-environment-wide association studies: emerging approaches.** *Nat Rev Genet* 2010, **11:**259-272.
49. Van der Heide F, Nolte IM, Kleibeuker JH, Wijmenga C, Dijkstra G, Weersma RK: **Differences in genetic background between active smokers, passive smokers, and non-smokers with Crohn's disease.** *Am J Gastroenterol* 2010, **105:**1165-1172.
50. Cadwell K, Patel KK, Maloney NS, Liu TC, Ng AC, Storer CE, Head RD, Xavier R, Stappenbeck TS, Virgin HW: **Virus-plus-susceptibility gene interaction determines Crohn's disease gene *Atg16L1* phenotypes in intestine.** *Cell* 2010, **141:**1135-1145.
51. Maier, S and Olek A: **Diabetes: a candidate disease for efficient DNA methylation profiling.** *J Nutr* 2002, **132:**2440S-2443S.
52. Kim, YI Logan JW, Mason JB, Roubenoff R: **DNA hypomethylation in inflammatory arthritis: reversal with methotrexate.** *J Lab Clin Med* 1996, **128:**165-172.
53. Cara Terribas CJ, Gonzalez Guijarro L: **Hypomethylation and multiple sclerosis, the susceptibility factor?** *Neurologia* 2002, **17:**132-135.
54. Issa JP: **Epigenetic variation and human disease.** *J Nutr* 2002, **132:**2388S-2392S.



55. Ahuja N, Issa JP: **Aging, methylation and cancer.** *Histol Histopathol* 2000, **15:**835-842.
56. Nadeau JH: **Transgenerational genetic effects on phenotypic variation and disease risk.** *Hum Mol Genet* 2009, **18:**202-210.
57. Morgan HD, Sutherland HG, Martin DI, Whitelaw E: **Epigenetic inheritance at the agouti locus in the mouse.** *Nat Genet* 1999, **23:**314-318.
58. Economou M, Trikalinos TA, Loizou KT, Tsianos EV, Ioannidis JP: **Differential effects of NOD2 variants on Crohn's disease risk and phenotype in diverse populations: a metaanalysis.** *Am J Gastroenterol* 2004, **99:**2393-2404.
59. Shen X, Shi R, Zhang H, Li K, Zhao Y, Zhang R: **The Toll-like receptor 4 D299G and T399I polymorphisms are associated with Crohn's disease and ulcerative colitis: a meta-analysis.** *Digestion* 2010, **81:**69-77.

doi:10.1186/gm227
Cite this article as: Fransen K, *et al.*: **The quest for genetic risk factors for Crohn's disease in the post-GWAS era.** *Genome Medicine* 2011, **3:**13.

6.6.4 Izvirni znanstveni članek 3

OPEN ACCESS Freely available online

PLOS | ONE

Limited Evidence for Parent-of-Origin Effects in Inflammatory Bowel Disease Associated Loci

Karin Fransen[1,2¶], Mitja Mitrovic[1,3¶], Cleo C. van Diemen[1], Thelma B. K.[4], Ajit Sood[5], Andre Franke[6], Stefan Schreiber[6,7], Vandana Midha[5], Garima Juyal[4], Uros Potocnik[3,9], Jingyuan Fu[1], Ilja Nolte[8], Rinse K. Weersma[2]*

1 Department of Genetics, University of Groningen, University Medical Centre Groningen, Groningen, The Netherlands, **2** Department of Gastroenterology and Hepatology, University of Groningen, University Medical Centre Groningen, Groningen, The Netherlands, **3** Center for Human Molecular Genetics and Pharmacogenomics, Medical Faculty, University of Maribor, Maribor, Slovenia, **4** Department of Genetics, University of Delhi, South Campus, New Delhi, India, **5** Department of Medicine, Dayanand Medical College and Hospital, Ludhiana, India, **6** Institute for Clinical Molecular Biology, Christian-Albrechts-University, Kiel, Germany, **7** Department of General Internal Medicine, University Clinic Schleswig-Holstein, Kiel, Germany, **8** Department of Epidemiology, University of Groningen, University Medical Centre Groningen, Groningen, The Netherlands, **9** Faculty of Chemistry and Chemical Engineering, University of Maribor, Maribor, Slovenia

Abstract

Background: Genome-wide association studies of two main forms of inflammatory bowel diseases (IBD), Crohn's disease (CD) and ulcerative colitis (UC), have identified 99 susceptibility loci, but these explain only ~23% of the genetic risk. Part of the 'hidden heritability' could be in transmissible genetic effects in which mRNA expression in the offspring depends on the parental origin of the allele (genomic imprinting), since children whose mothers have CD are more often affected than children with affected fathers. We analyzed parent-of-origin (POO) effects in Dutch and Indian cohorts of IBD patients.

Methods: We selected 28 genetic loci associated with both CD and UC, and tested them for POO effects in 181 Dutch IBD case-parent trios. Three susceptibility variants in *NOD2* were tested in 111 CD trios and a significant finding was re-evaluated in 598 German trios. The UC-associated gene, *BTNL2*, reportedly imprinted, was tested in 70 Dutch UC trios. Finally, we used 62 independent Indian UC trios to test POO effects of five established Indian UC risk loci.

Results: We identified POO effects for *NOD2* (L1007fs; OR = 21.0, P-value = 0.013) for CD; these results could not be replicated in an independent cohort (OR = 0.97, P-value = 0.95). A POO effect in IBD was observed for *IL12B* (OR = 3.2, P-value = 0.019) and *PRDM1* (OR = 5.6, P-value = 0.04). In the Indian trios the *IL10* locus showed a POO effect (OR = 0.2, P-value = 0.03).

Conclusions: Little is known about the effect of genomic imprinting in complex diseases such as IBD. We present limited evidence for POO effects for the tested IBD loci. POO effects explain part of the hidden heritability for complex genetic diseases but need to be investigated further.

Citation: Fransen K, Mitrovic M, van Diemen CC, BK T, Sood A, et al. (2012) Limited Evidence for Parent-of-Origin Effects in Inflammatory Bowel Disease Associated Loci. PLoS ONE 7(9): e45287. doi:10.1371/journal.pone.0045287

Editor: Zhaoxia Yu, University of California Irvine, United States of America

Received April 23, 2012; **Accepted** August 14, 2012; **Published** September 27, 2012

Copyright: © 2012 Fransen et al. This is an open-access article distributed under the terms of the Creative Commons Attribution License, which permits unrestricted use, distribution, and reproduction in any medium, provided the original author and source are credited.

Funding: K. Fransen was supported by a MD/PhD grant from GUIDE at University Medical Center Groningen. M. Mitrovic was supported by a PhD grant from Slovenian Research Agency (Grant No. 630-39/2012-1). U. Potocnik was supported by a grant from Slovenian Research Agency (Grant No. J3-2175). R.K. Weersma is supported by a clinical fellowship grant (90.700.281) from the Netherlands Organization for Scientific Research (NWO) and the Broad Medical Research Program of The Broad Foundation. The funders had no role in study design, data collection and analysis, decision to publish, or preparation of the manuscript.

Competing Interests: The authors have declared that no competing interests exist.

* E-mail: r.k.weersma@umcg.nl

¶ These authors contributed equally to this work.

Introduction

Crohn's disease (CD) and ulcerative colitis (UC) are the two main forms of chronic relapsing inflammatory bowel diseases (IBD). With a cumulative prevalence of up to 800 per 100,000 in Europe and 570 in North America [1], it is considered one of the most common immune-related diseases worldwide. Typically, from their second or third decade on, patients suffer from a chronic relapsing inflammation of the gut, which is often accompanied by extra-intestinal manifestations and complications that can be extremely debilitating and severe. Treatments are costly and often insufficient and can be accompanied by severe side-effects [2]. Hence, there is an urgent need for new therapeutic targets and curative medication. The pathogenesis is largely unknown and it is currently thought that an aberrant immune response to commensal microflora in a genetically susceptible host underlies the disease [3].

Prior to the introduction of genome-wide association studies (GWAS), only three loci had been consistently associated with either form of IBD. Over the past six years, multiple GWAS and meta-analyses have yielded a lengthening list of variants associated with CD (71 confirmed independent genetic risk loci) and UC (47 loci) [4,5]. Nevertheless, despite this encouraging progress, as much as 77% of the estimated heritability for CD and 72% for UC is still considered to be unexplained [4,5]. Thus, one of the challenges in the post-GWAS era is to identify potential sources of this 'hidden heritability' [6], which may reside in associated variants with lower odds ratios, gene-gene interactions, gene-environment interactions, and/or structural variation.

In addition, parent-of-origin effects (POO) may comprise a piece of the missing heritability puzzle in IBD, as suggested by Akolkar et al. [7]. They show that offspring of mothers with CD are at higher risk for CD than when fathers are affected. More recently, Zelinkova et al. showed there was maternal imprinting and female predominance in familial Crohn's disease [8]. This could be explained with at least two distinct types of POO mechanisms (Fig. 1). If the paternal allele is inactivated by genomic imprinting, then expression of the locus is determined only by the maternal allele (Fig. 1a). If this effect is not taken into account, there may be a significant loss in the statistical power of genetic association studies [9,10]. Secondly, maternal effects such as diet or genotype affect the environment for the developing fetus (Fig. 1b). It is thought that maternal proteins or circulating RNA passes the placental barrier and may cause changes in the epigenome of fetal DNA, thereby influence its phenotype.

In this study, we tested for these two types of POO effects in IBD by using a likelihood ratio test developed by Weinberg [11]. A previous version of this method (parental asymmetry test) has already successfully identified a POO effect in another complex genetic disease, type 1 diabetes [12].

Materials and Methods

Ethical Considerations

This study was approved by the institutional review boards (Institutional ethical committee, Dayanand Medical College and Hospital, Ludhiana and Institutional ethical committee, University of Delhi South Campus; Ethical Review Board of the Medical Faculty of the Christian-Albrechts-University of Kiel; Institutional review board, University Medical Centre Groningen, The Netherlands) of each of the hospitals and written informed consent was obtained from all subjects personally.

Subjects

All patients were diagnosed according to standard clinical criteria by endoscopy, radiology and histopathology [13]. A total of 249 classical offspring-parent trios (one affected offspring with two unaffected parents) were included in our initial POO analysis. Of these trios, 115 offspring had CD and 134 had UC. All 115 CD trios were of western European descent from the Netherlands and were collected at the IBD Center in the Department of Gastroenterology and Hepatology, University Medical Center Groningen (UMCG), the Netherlands. For UC there were 72 trios of Dutch ancestry (collected at the UMCG) and 62 of Indian ancestry (collected at Dayanand Medical College and Hospital, Ludhiana, Punjab, India). Tables 1 and 2 show the clinical characteristics of the studied cohorts. Genotype data of an independent replication cohort, consisting of 598 CD parent-offspring trios from Germany, were obtained for the *NOD2* L1007fs variant from the Institute of Clinical Molecular Biology, Christian-Albrechts-University of Kiel, Germany. These data were available from a previous study, no phenotypic data is available [14].

Genotyping and Quality Control

All Dutch and Indian subjects were genotyped using the Illumina Immunochip (iCHIP) (Illumina Inc., San Diego, California, United States of America), which is a custom-made genotyping array that contains ~200,000 single nucleotide polymorphisms (SNPs) focusing on immune-mediated diseases [15,16]. Genotyping was performed according to the manufacturer's protocol. Genotyping clusters of the SNPs included in the current analysis were checked manually using GenomeStudio software by Illumina Inc. [15,16]. Individuals with a call rate $<95\%$ and/or discordant gender information, and SNPs with a call rate $<98\%$ were removed from further analysis. Identity-by-descent analysis by Plink software was used to test for incorrect family relations (Mendelian errors) in the trios, but no mismatches were identified.

SNP Selection

For this study we used several strategies to select SNPs. First, to gain power we pooled the Dutch UC and CD trios and tested for POO effects in 28 established IBD loci that are both associated to CD and UC [4]. To avoid losing significance due to multiple testing correction we tested the 28 overlapping loci instead of all 99 associated risk loci. Variant rs736289 is not present on iCHIP and no proxy ($r^2>0.5$) could be found. Two variants (rs12261843, rs181359) were also not captured by iCHIP, but we identified perfect proxies using SNAP software: rs12261843 was represented by rs12254167 ($r^2=1$; D' = 1) and rs181359 was represented by rs2266961 ($r^2=1$; D' = 1) [17]. Second, we aimed to test for the existence of POO effects in the UC risk SNPs established in Indians [18]. In addition, we aimed to include SNPs from known imprinted genes in our analysis. For this, a publicly available database of known imprinted genes was compared with all 99 IBD-associated loci [19]. The associated locus was defined as the region of $r^2>0.5$ flanking the most significantly associated SNP, then extended to the nearest recombination hot-spot, and from there for an extra 100 kb. Comparison of IBD-associated genes with the known imprinted genes resulted in the inclusion of one extra gene, *BTNL2*; since this is a UC-specific locus it was only analyzed in the Dutch and Indian UC cohorts. SNP rs9268853 is the reported UC risk SNP in the Caucasians and was tested in the Dutch trios, rs3763313 was tested for POO effect in the Indian trios since this is the reported risk SNP in the Indian population. Lastly, we included *NOD2* since it is the most strongly associated gene and is most replicated in association studies of CD in populations of western European descent. Three common disease-susceptibility variants (G809R, R702W, and L1007fs) were therefore tested for POO effects in the Dutch CD trios, and subsequently the L1007fs variant was tested in the German replication cohort [20].

Statistical Analysis

A power analysis was performed with Quanto software [21] and showed that in Dutch trios (n = 181) we had more than 80% power to detect POO effects of OR ≥ 3 in variants with MAF ≥ 0.025. In Indian trios (n = 62) we had 80% power to detect POO effects of OR ≥ 3 in variants with MAF ≥ 0.075 (see Fig. S1). POO effects were calculated by a log-likelihood ratio test, which is a statistical test used to compare the fit of the null hypothesis (i.e. no evidence/presence of POO in our case) and the alternative hypothesis. The test is based on the likelihood ratio, which expresses how many times more likely the data are under one model than the other and can be used to decide whether to reject the null model in favor of the alternative model. Weinberg *et al.* [11] have developed a log-linear model when a case-parents triad is genotyped and jointly classified according to the number of copies of a particular allele carried by the mother, father, and child (denoted as "M," "P," and "C," respectively), there are 15 possible outcomes (i.e. mating

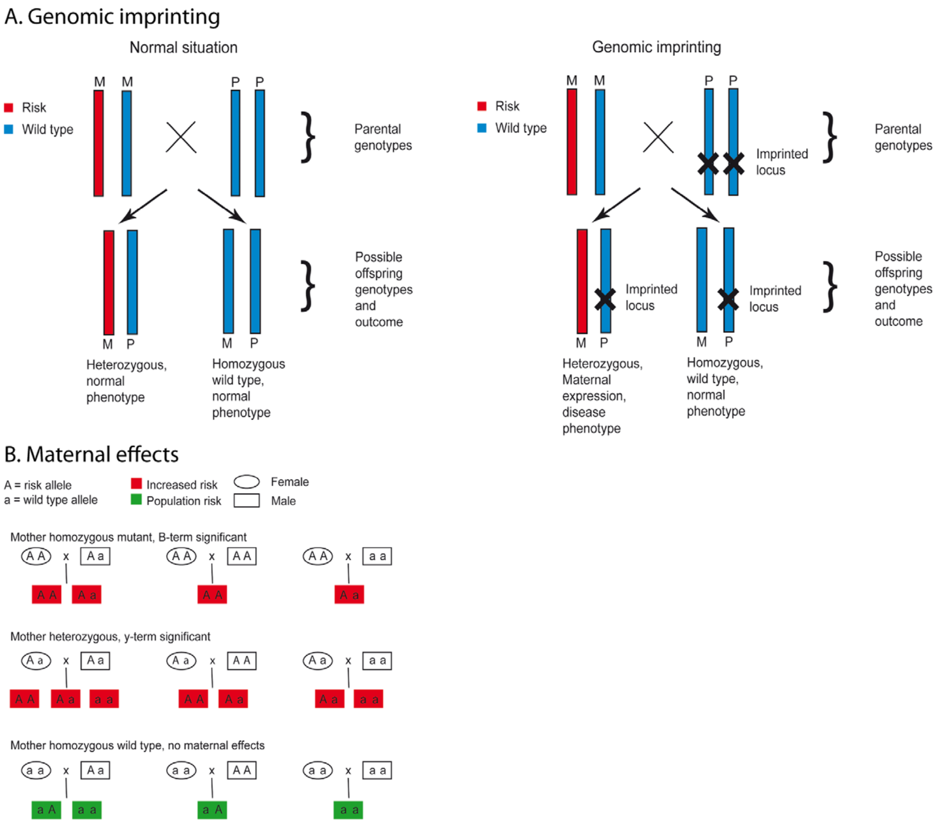

A. Genomic imprinting
Normal situation
Genomic imprinting
Risk
Wild type
M M
P P
Parental genotypes
Imprinted locus
M P
Possible offspring genotypes and outcome
Heterozygous, normal phenotype
Homozygous wild type, normal phenotype
Heterozygous, Maternal expression, disease phenotype
Homozygous, wild type, normal phenotype
B. Maternal effects
A = risk allele
a = wild type allele
Increased risk
Population risk
Female
Male
Mother homozygous mutant, B-term significant
AA x Aa
AA Aa
AA x AA
AA
AA x aa
Aa
Mother heterozygous, γ-term significant
Aa x Aa
AA Aa aa
Aa x AA
AA Aa
Aa x aa
Aa aa
Mother homozygous wild type, no maternal effects
aa x Aa
aA aa
aa x AA
aA
aa x aa
aa

Figure 1. Distinct types of parent-of-origin mechanisms tested in this study. Fig. 1a. Genomic imprinting: Genomic imprinting is characterized by consequent silencing of one allele, depending on the parental origin. In the example shown above a normal situation is displayed on the left and the genomic imprinting is shown on the right; red is the risk allele and bleu is the wild type allele. The maternal genotype is heterozygous, the father's genotype is homozygous wild-type. Offspring in the left scenario have a normal phenotype since the paternal wild-type allele is expressed in the heterozygous offspring and the mutated allele of the mother is thus rescued by the paternal allele. On the right genomic imprinting is shown, reflecting the α-term in the method used to test for parent of origin effects. In this example there is a significant genomic imprinting effect and the OR >1 so the paternal allele is silenced (see materials and methods section statistical analysis). We assume an additive or recessive model of inheritance. Two possible outcomes are listed, if the offspring inherits the risk allele from the mother and the wild-type allele from the father is subjected to genomic imprinting, then only the risk allele is expressed, thus the offspring is affected by the mutated allele from the mother. Fig. 1b. Maternal effects: Maternal effects are effects of the maternal genotype on the fetal phenotype, irrespective of the fetal genotype, these effects are reflected by the β- and γ-terms in the likelihood ratio test that was used to test for parent of origin effects in our study. In the example given above, the β- and γ-terms are significant with an OR >1, meaning that the risk of disease is higher if the mother carries two or one risk allele respectively. A recessive or co-dominant model is assumed, and higher expression of the mutant allele leads to disease. If the genotype of the offspring is red, then maternal effects cause increased disease risk and if it is green than the normal population risk applies. If the mother is homozygous wild-type, no maternal effects occur. If she is homozygous mutant or heterozygous for the risk allele, the offspring is subjected to maternal effects and thus has an increased disease risk. Note that the wild-type homozygous offspring has a higher disease risk if both parents are heterozygous.
doi:10.1371/journal.pone.0045287.g001

types). The family-specific outcomes (i.e., the cell into which a particular triad is classified) are independent, provided that each family contributes only one case. The counts based on classification of the triads studied can therefore be thought of as distributed according to a 15-cell multinomial. The method is based on consideration of mating types in which the mother and father carry unequally many copies of the variant allele, with further stratification on the number of inherited copies of the allele, C. This second level of conditioning (on C) effectively removes any effects related jointly to the inherited number of copies and the

Table 1. Phenotypic characterization of subjects with Crohn's disease.

Cohort	Number of trios	No. of males (%)	AOO	Ileal	Colonic	Ileocolon	Upper GI
Dutch CD	115	39 (34%)	24	23	28	64	11
German CD	598	N/A	N/A	N/A	N/A	N/A	N/A

doi:10.1371/journal.pone.0045287.t001

parental-allele counts M, P. In short, the method is valid if inheritance of allele is Mendelian, if there is parental symmetry within mating types in the population studied, and if the gene under study is not in linkage disequilibrium with another disease-susceptibility gene. First, the relative penetration of the risk allele of the child is established by determining the parental origin of the risk allele. In the latter, the difference in disease risk is compared for the varying amounts of risk alleles carried by the mothers; the genotype of the child is not relevant. The likelihood ratio test calculates α-, ß-, and γ-terms. The α-term indicates the significance level for genomic imprinting effects: if OR >1, the risk allele is transmitted more often from the mother to the patient and if OR <1 then it is transmitted more often from the father. The ʅ- and γ-terms indicate the prenatal effect of the maternal genotype when the mother carries two risk alleles or one risk allele, respectively. When the ß- or γ-term is significant and OR >1 then the child has more chance of getting the disease due to maternal effects. If OR <1 then the child has less chance of developing the disease as a consequence of this prenatal effect. Bonferroni multiple testing corrections were applied to the four different analyses.

Results

DNA of 249 complete IBD trios was available for our study, of which four CD (4/115) and two Dutch UC trios (2/72) did not pass the quality control. Therefore 243 IBD trios (111 CD, 70 Dutch UC & 62 Indian UC) were available for the discovery phase of the study. Our findings were then replicated in an independent replication cohort consisting of 598 German CD trios.

Parent-of-origin Analysis in Dutch IBD Trios

A nominally significant genomic imprinting effect was found in the *IL12B* gene (α term: P = 0.019; OR = 3.2), with OR >1 indicating that the risk allele is more often transmitted from the mother to the child. In addition, the β term was nominally significant (p-value = 0.003; OR = 0.2) with OR <1, indicating that offspring have less chance of getting the disease if their mothers carry two risk alleles. The *PRDM1* gene showed a nominally significant maternal effect if the mother carried two risk alleles (β term: p-value = 0.04; OR = 5.6), with OR >1 indicating that the offspring have more chance of getting the disease. The other tests did not result in significant POO effects. However, none of these associations were significant after the Bonferroni correction (table 3).

NOD2 in Dutch and German CD Trios

Three established CD variants in *NOD2* (G809R, R702W, L1007fs) were tested for POO effect in the 111 CD trios [20]. After correcting for multiple testing, a significant genomic imprinting effect was detected for the L1007fs variant (α term: p-value = 0.013; OR = 21.0). The risk allele was transmitted more often from the mother than the father. Given the high OR we aimed to replicate this finding in an independent German cohort for which *NOD2* genotyping data was available. Unfortunately, our results could not be replicated in this cohort (α term: p-value = 0.95; OR = 0.97) (table 4).

Known Imprinted Gene BTNL2 in Dutch UC Trios

No significant POO effects were detected the established UC SNP in the *BTNL2* locus (rs9268853) in 72 Dutch UC trios (table 5).

Indian UC Analysis

The established Indian UC SNPs were tested for POO effects in 62 Indian trios [18]. We found a nominally significant genomic imprinting effect in the *IL10* locus (p-value = 0.03; OR = 0.16) where the OR <1 indicates that the risk allele is more often transmitted from the father. This association does not, however, pass the multiple testing correction. This SNP was also tested in the population of western European descent and we could not detect any significant imprinting effect. The *NOD2* variant that showed association in the Indian population could not be tested for POO effects since only homozygous wild-type fathers were available, hence all the trios were uninformative (table 6).

Discussion

For the first time parent-of-origin effects have been tested in IBD on a genetic level for the overlapping IBD-associated loci. We found limited evidence that POO effects exist in IBD in the Dutch population for *IL12B*, *PRDM1* and *NOD2* in our discovery cohort, but the large POO effect for *NOD2* could not be replicated in an independent German replication cohort. Moreover, we found a nominally significant POO effect in *IL10* in our Indian population. Although the results from the Dutch trios might be false-positive,

Table 2. Phenotypic characterization of subjects with ulcerative colitis.

Cohort	Number of trios	No. of males (%)	AOO	Proctitis	Left-sided	Extended	Unknown
Dutch UC	72	30 (42%)	25	8	19	38	5
Indian UC	62	45 (73%)	28	22	15	20	5

N/A not available, AOO average age of onset. Cases and disease location are given according to the Montreal classification. for CD L1, L2, L3 and L4; for UC E1, E2, E3. No phenotypic information was available for the German cohort.

doi:10.1371/journal.pone.0045287.t002

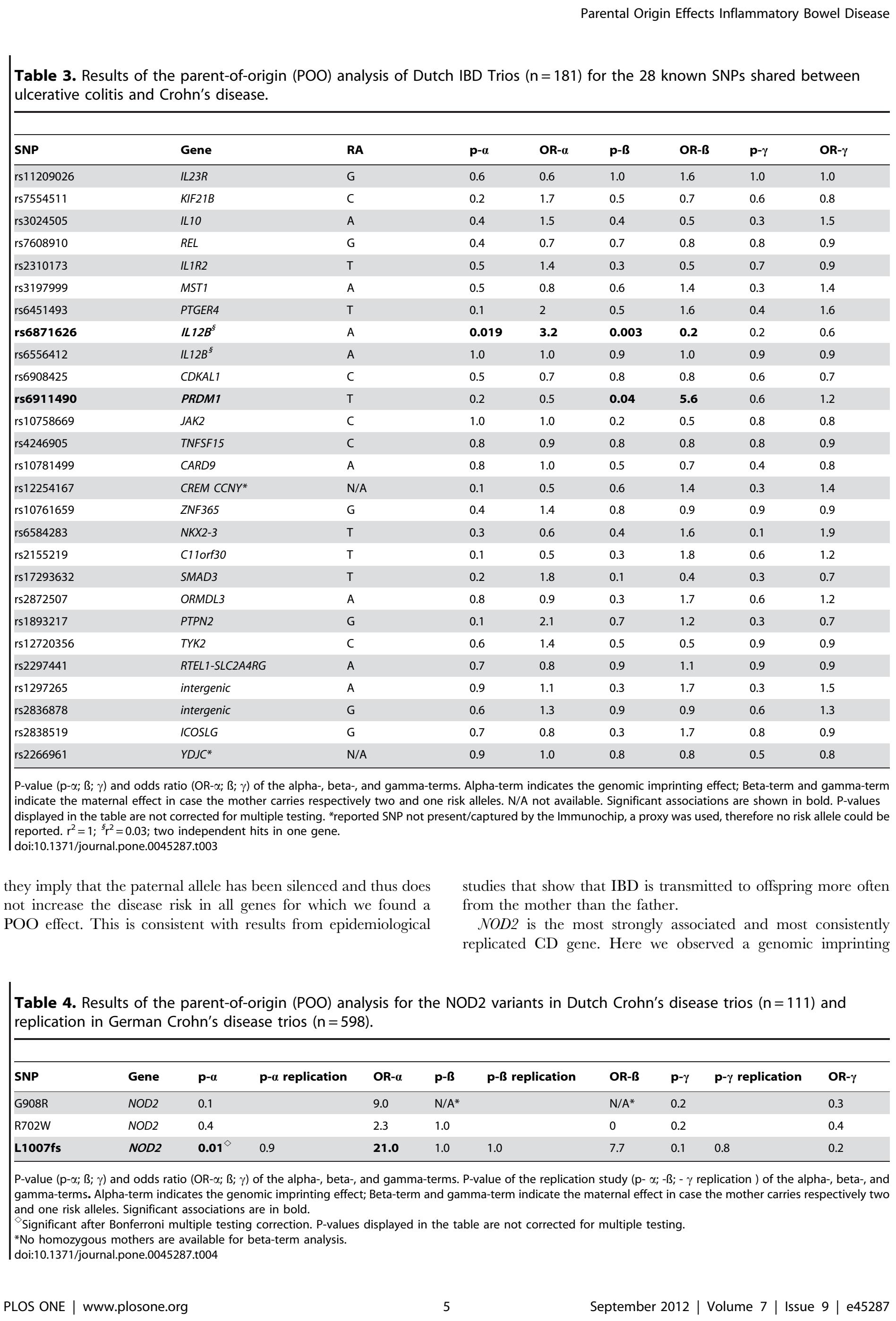


Table 3. Results of the parent-of-origin (POO) analysis of Dutch IBD Trios (n = 181) for the 28 known SNPs shared between ulcerative colitis and Crohn's disease.

SNP	Gene	RA	p-α	OR-α	p-ß	OR-ß	p-γ	OR-γ
rs11209026	*IL23R*	G	0.6	0.6	1.0	1.6	1.0	1.0
rs7554511	*KIF21B*	C	0.2	1.7	0.5	0.7	0.6	0.8
rs3024505	*IL10*	A	0.4	1.5	0.4	0.5	0.3	1.5
rs7608910	*REL*	G	0.4	0.7	0.7	0.8	0.8	0.9
rs2310173	*IL1R2*	T	0.5	1.4	0.3	0.5	0.7	0.9
rs3197999	*MST1*	A	0.5	0.8	0.6	1.4	0.3	1.4
rs6451493	*PTGER4*	T	0.1	2	0.5	1.6	0.4	1.6
rs6871626	***IL12B***§	A	**0.019**	**3.2**	**0.003**	**0.2**	0.2	0.6
rs6556412	*IL12B*§	A	1.0	1.0	0.9	1.0	0.9	0.9
rs6908425	*CDKAL1*	C	0.5	0.7	0.8	0.8	0.6	0.7
rs6911490	***PRDM1***	T	0.2	0.5	**0.04**	**5.6**	0.6	1.2
rs10758669	*JAK2*	C	1.0	1.0	0.2	0.5	0.8	0.8
rs4246905	*TNFSF15*	C	0.8	0.9	0.8	0.8	0.8	0.9
rs10781499	*CARD9*	A	0.8	1.0	0.5	0.7	0.4	0.8
rs12254167	*CREM CCNY**	N/A	0.1	0.5	0.6	1.4	0.3	1.4
rs10761659	*ZNF365*	G	0.4	1.4	0.8	0.9	0.9	0.9
rs6584283	*NKX2-3*	T	0.3	0.6	0.4	1.6	0.1	1.9
rs2155219	*C11orf30*	T	0.1	0.5	0.3	1.8	0.6	1.2
rs17293632	*SMAD3*	T	0.2	1.8	0.1	0.4	0.3	0.7
rs2872507	*ORMDL3*	A	0.8	0.9	0.3	1.7	0.6	1.2
rs1893217	*PTPN2*	G	0.1	2.1	0.7	1.2	0.3	0.7
rs12720356	*TYK2*	C	0.6	1.4	0.5	0.5	0.9	0.9
rs2297441	*RTEL1-SLC2A4RG*	A	0.7	0.8	0.9	1.1	0.9	0.9
rs1297265	*intergenic*	A	0.9	1.1	0.3	1.7	0.3	1.5
rs2836878	*intergenic*	G	0.6	1.3	0.9	0.9	0.6	1.3
rs2838519	*ICOSLG*	G	0.7	0.8	0.3	1.7	0.8	0.9
rs2266961	*YDJC**	N/A	0.9	1.0	0.8	0.8	0.5	0.8

P-value (p-α; ß; γ) and odds ratio (OR-α; ß; γ) of the alpha-, beta-, and gamma-terms. Alpha-term indicates the genomic imprinting effect; Beta-term and gamma-term indicate the maternal effect in case the mother carries respectively two and one risk alleles. N/A not available. Significant associations are shown in bold. P-values displayed in the table are not corrected for multiple testing. *reported SNP not present/captured by the Immunochip, a proxy was used, therefore no risk allele could be reported. $r^2 = 1$; $^{\S}r^2 = 0.03$; two independent hits in one gene.

doi:10.1371/journal.pone.0045287.t003

they imply that the paternal allele has been silenced and thus does not increase the disease risk in all genes for which we found a POO effect. This is consistent with results from epidemiological studies that show that IBD is transmitted to offspring more often from the mother than the father.

NOD2 is the most strongly associated and most consistently replicated CD gene. Here we observed a genomic imprinting

Table 4. Results of the parent-of-origin (POO) analysis for the NOD2 variants in Dutch Crohn's disease trios (n = 111) and replication in German Crohn's disease trios (n = 598).

SNP	Gene	p-α	p-α replication	OR-α	p-ß	p-ß replication	OR-ß	p-γ	p-γ replication	OR-γ
G908R	*NOD2*	0.1		9.0	N/A*		N/A*	0.2		0.3
R702W	*NOD2*	0.4		2.3	1.0		0	0.2		0.4
L1007fs	***NOD2***	**0.01**$^{\diamond}$	0.9	**21.0**	1.0	1.0	7.7	0.1	0.8	0.2

P-value (p-α; ß; γ) and odds ratio (OR-α; ß; γ) of the alpha-, beta-, and gamma-terms. P-value of the replication study (p- α; -ß; - γ replication) of the alpha-, beta-, and gamma-terms**.** Alpha-term indicates the genomic imprinting effect; Beta-term and gamma-term indicate the maternal effect in case the mother carries respectively two and one risk alleles. Significant associations are in bold.

$^{\diamond}$Significant after Bonferroni multiple testing correction. P-values displayed in the table are not corrected for multiple testing.

*No homozygous mothers are available for beta-term analysis.

doi:10.1371/journal.pone.0045287.t004

Table 5. Results of the parent-of-origin (POO) analysis in the BTNL2 locus in Dutch UC ulcerative colitis trios (n = 72).

SNP	Gene	RA	p-α	OR-α	p-ß	OR-ß	p-γ	OR-γ
rs9268853	*BTNL2*	T	1.0	1.0	0.7	0.7	0.6	0.6

P-value (p-α; ß; γ) and odds ratio (OR-α; ß; γ) of the alpha-, beta-, and gamma-term. Alpha-term indicates the genomic imprinting effect; Beta-term and gamma-term indicate the maternal effect in case the mother carries respectively two and one risk alleles. Significant associations are in bold. P-values displayed in the table are not corrected for multiple testing.
doi:10.1371/journal.pone.0045287.t005

effect for the L1007fs mutation in Dutch CD trios, yet we failed to replicate this in an independent German cohort. The results in our initial analysis might be a false-positive finding, although we had sufficient power to detect effects in the Dutch trios. Although both cohorts were of western European descent, we question whether population specific and environmental factors might play a major role in POO effects and explain part of the lack of replication. We will further elaborate on this later in the discussion. TheL1007fs mutation seems to have a predominant role in CD families since recently it has been shown in a case report that all family members were carrying the mutation and had CD [22]. Moreover in cases of homozygosity this variant will lead to ileal stenosis [23,24], implying a strong effect of the L1007fs mutation on the disease phenotype.

Our results suggest a possible contradictory effect of the two types of POO effects we studied for the *IL12B* gene both with a nominal significance. The genomic imprinting analysis (α-term) showed inheritance of the disease risk from the mother, while the analysis of independent maternal effect showed protection for disease if the mother carries two risk alleles (β-term). This suggests that the maternal risk allele is expressed and causes a higher disease risk, but simultaneously and independently, if the mother carries two risk alleles the child has a lower risk for IBD due to *in utero* effects on the fetus. *IL12B* resides on the established and consistently replicated IBD locus on chromosome 5q33 and it encodes a sub-unit of IL23, which is involved in Th17/IL23R signaling. This pathway has been implicated in several chronic, immune-related diseases such as psoriasis, rheumatoid arthritis, and ankylosing spondylitis [25,26,27].

The *PRDM1* gene showed a nominally significant maternal effect in the POO analysis in the population of western European descent when mothers carried two risk alleles; the OR of 5.6 supports results from others that IBD is more often transmitted from the mother than the father. The environment in which the fetus develops causes changes that increase the risk for CD. *PRDM1* has been associated to several immune-related diseases, but also to various types of lymphomas [28,29,30,31]. It encodes a protein that represses the expression of the β-interferon gene. Hypothetically, if the maternal effect causes altered expression of *PRDM1,* an aberrant immune response could increase CD risk.

In the Indian trio study we found a nominally significant genomic imprinting effect for the *IL10* locus. In contrast to the findings in the population of western European descent, the risk allele was more often transmitted from the father to the child. No epidemiological studies of the Indian population are available to validate the paternal transmission. We could not replicate the POO effects from the Dutch population in the Indian trios nor *vice versa*. This might indicate that POO effects are population specific. Later in the discussion we will discuss this in further detail.

Weinberg's method to detect POO effect is a robust one. Moreover it takes genomic imprinting and maternal effects into consideration simultaneously. It therefore has less power than the standard parental asymmetry test (PAT) that only tests for genomic imprinting, but PAT is invalid if maternal effects are present [11]. The importance of these maternal effects has been shown in mice, with a knockout of the serotonin 1A receptor gene leading to an anxiety-like phenotype. Implantation of wild-type embryos into knockout mothers and cross-fostering of the pups with wild-type mothers showed the full anxiety-phenotype, indicating that the maternal genotype influences the phenotype and that this effect persists after birth [32]. This is supported by our evidence for maternal effects in the *PRDM1* and *IL12B* loci.

We do not know why POO effects occur. Humans are diploid organisms and as such, can survive the, on average, 500 recessive mutations that are present in every human being, since most deleterious effects are rescued by the other allele [33]. Genomic imprinting significantly deduces diploidy by consequently inactivating one haplotype depending on its parental origin and thus impairing the rescue mechanism. The most cited and best supported hypothesis for the existence of this counter-intuitive phenomenon is the parental conflict hypothesis, in which both sexes have a need to pass on their genetic information to the next generation. Yet this does not explain the existence of genomic imprinting in immune-related genes, for example [34]. Hypothetically, to prevent adverse reactions passing from mother to fetus, it

Table 6. Results of the parent-of-origin (POO) analysis in Indian UC trios (n = 62).

SNP	Gene	RA	p-α	OR-α	p-ß	OR-ß	p-γ	OR-γ
rs6426833	*RNF186*	A	0.2	0.3	0.2	4.3	0.1	4.8
rs3024505	***IL10***	A	**0.03**	**0.16**	0.8	1.3	0.5	1.5
rs3763313	*BTNL2*	T	0.8	0.8	1.0	2.0	1.0	4.0
rs2395185	*HLA-DRA*	A	0.3	2.3	0.8	1.3	0.4	3.0

P-value (p-α; ß; γ) and odds ratio (OR-α; ß; γ) of the alpha-, beta-, and gamma-term. Alpha-term indicates the genomic imprinting effect; Beta-term and gamma-term indicate the maternal effect in case the mother carries respectively two and one risk alleles. Significant associations are in bold. P-values displayed in the table are not corrected for multiple testing.
doi:10.1371/journal.pone.0045287.t006

is important that the immune responses are alike and thus preferably maternal immune genes are expressed.

In our study we had sufficient power (>80%) to detect POO effects with an OR of three or higher for each SNP in our study. By adding more tests the significance level must be adjusted accordingly and the power to detect differences is lower. Therefore we chose to only test the 28 overlapping IBD risk loci in a pooled cohort of Dutch CD and UC trios instead of all 99 risk loci. Consequently, bigger cohorts are needed to test the remaining IBD loci for POO effects.

None of our findings in one population could be replicated in another population. At least three reasons could explain this fact. First, it might be that the initial findings are false positive findings: the cohorts have a limited size and thus more variation around the mean, resulting in a higher chance of false positive findings. Second, it is unknown for how long genomic imprinting effects are stable in humans. It has been shown from mouse studies that genomic imprinting is stable for at least 3 generations [35]. No data is available in human studies. Moreover, genomic imprinting was shown to be influenced by environmental factors [36,37], which could mean that although the imprinting mechanism is global, distinct genes may be imprinted in different populations because they were exposed to distinct environmental effects. The latter two could indicate that even within populations different imprinting effects occur.

In conclusion, we aimed to identify genomic imprinting effects and maternal effects acting on the risk alleles of IBD and we showed, for the first time, that *IL12B*, *NOD2* and *PRDM1* might be involved in these phenomena in Dutch IBD trios. It has already been shown that POO effects exist in type 1- and type 2 diabetes, which like IBD are complex genetic disorders and show a substantial overlap of disease susceptibility loci with IBD [10,12]. Given the high OR in *NOD2* we sought to replicate our findings, but could not confirm the POO effect for *NOD2* in an independent German replication cohort. In the Indian population we did identify POO effects in the *IL10* gene. We could neither replicate our findings from the Dutch trios in the Indian or in the German population nor our findings from the Indian population in the Dutch cohort. This suggests that POO effects are either false-positive findings or prone to be population specific. We anticipate that future investigations, using larger, multi-ethnical cohorts will help to shed light on these complex and currently little known relationships. Parent-of-origin effects can take various forms and are not restricted to imprinting, but may involve a variety of mechanisms including gender effects, epistasis, epigenetic effects, and environmental influences during pre- or postnatal development. Better understanding of such effects will probably require detailed studies of model organisms in which breeding and environment can be carefully controlled. Given that alleles identified through GWAS account for a relatively small fraction of heritability, parent of origin effects may underlie some of the missing heritability problem. With appropriate family-based study designs data analysis methods and international collaborative efforts it will be possible to screen for parent of origin effects across the entire genome. In addition, epigenetic profiling on a genome scale will likely lead to the identification of novel epigenetic marks in a variety of disorders that may provide a bridge among the parental genome, parental environment, and offspring phenotype. We anticipate that the investigations of alternative models of inheritance, appropriate study design and application of novel technologies will enable a more complete picture of heritability in human traits, leading to new insights in the field of genetics of complex diseases.

Supporting Information

Figure S1 Power analysis of a. the Dutch trio analysis (181 trios) and b. the Indian trios (62 trios). S1a. Power calculation of the Dutch trio analysis: The power is shown on the x-axis, the different odds ratios (OR) are shown on the y-axis. The different lines represent SNPs with different minor allele frequencies. In red, the regular 80% power cut-off is shown. With an OR of 3, we have sufficient power to detect parent-of-origin effects in SNPs with a MAF of 2.5%. S1b. Power calculation of the Indian trio analysis: The power is shown on the x-axis, the different odds ratios (OR) are shown on the y-axis. The different lines represent SNPs with different minor allele frequencies. In red, the 80% power cut-off is shown. With an OR of 4, we have sufficient power to detect parent-of-origin effects in SNPs with a MAF of 4.0%.
(TIF)

Acknowledgments

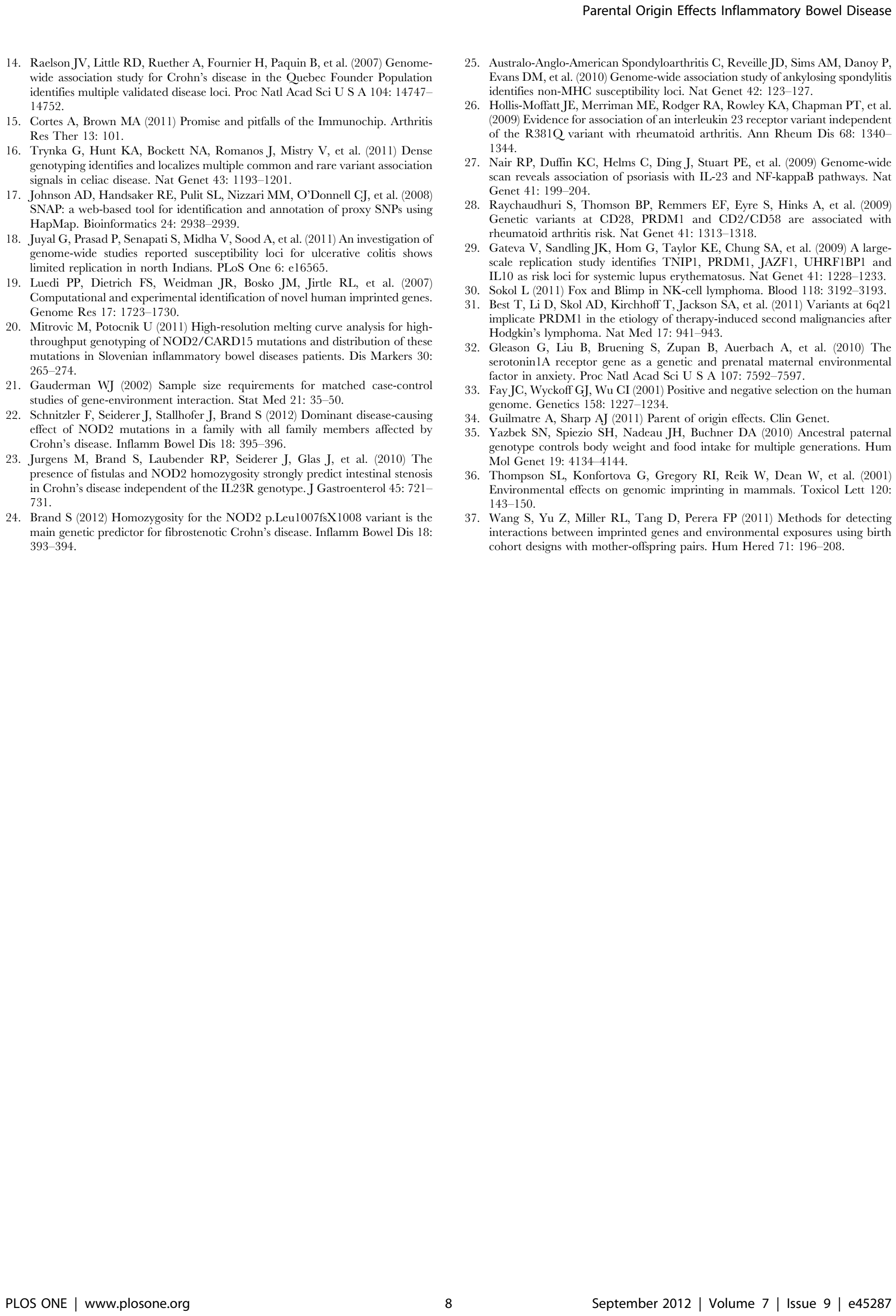
We thank all the patients who participated in this study and Jackie Senior for editing the manuscript.

Author Contributions

Conceived and designed the experiments: KF MM RKW. Performed the experiments: KF MM AF IN. Analyzed the data: KF MM IN CCD. Wrote the paper: KF MM. Recruited and interviewed the participating patients and their families: RKW UP TBK AS AF SS GJ VM. Critically revised and supervised the paper: RKW CCD JYF IN. Read and approved the final manuscript: KF MM CCD TBK AS AF SS VM GJ UP JYF IN RKW.

References

1. Molodecky NA, Soon IS, Rabi DM, Ghali WA, Ferris M, et al. (2012) Increasing incidence and prevalence of the inflammatory bowel diseases with time, based on systematic review. Gastroenterology 142: 46–54 e42.
2. Di Sabatino A, Liberato L, Marchetti M, Biancheri P, Corazza GR (2011) Optimal use and cost-effectiveness of biologic therapies in inflammatory bowel disease. Intern Emerg Med 6 Suppl 1: 17–27.
3. Nell S, Suerbaum S, Josenhans C (2010) The impact of the microbiota on the pathogenesis of IBD: lessons from mouse infection models. Nat Rev Microbiol 8: 564–577.
4. Anderson CA, Boucher G, Lees CW, Franke A, D'Amato M, et al. (2011) Meta-analysis identifies 29 additional ulcerative colitis risk loci, increasing the number of confirmed associations to 47. Nat Genet 43: 246–252.
5. Franke A, McGovern DP, Barrett JC, Wang K, Radford-Smith GL, et al. (2010) Genome-wide meta-analysis increases to 71 the number of confirmed Crohn's disease susceptibility loci. Nat Genet 42: 1118–1125.
6. Fransen K, Mitrovic M, van Diemen CC, Weersma RK (2011) The quest for genetic risk factors for Crohn's disease in the post-GWAS era. Genome Med 3: 13.
7. Akolkar PN, Gulwani-Akolkar B, Heresbach D, Lin XY, Fisher S, et al. (1997) Differences in risk of Crohn's disease in offspring of mothers and fathers with inflammatory bowel disease. Am J Gastroenterol 92: 2241–2244.
8. Zelinkova Z, Stokkers PC, van der Linde K, Kuipers EJ, Peppelenbosch MP, et al. (2012) Maternal imprinting and female predominance in familial Crohn's disease. J Crohns Colitis.
9. Hanson RL, Kobes S, Lindsay RS, Knowler WC (2001) Assessment of parent-of-origin effects in linkage analysis of quantitative traits. Am J Hum Genet 68: 951–962.
10. Kong A, Steinthorsdottir V, Masson G, Thorleifsson G, Sulem P, et al. (2009) Parental origin of sequence variants associated with complex diseases. Nature 462: 868–874.
11. Weinberg CR (1999) Methods for detection of parent-of-origin effects in genetic studies of case-parents triads. Am J Hum Genet 65: 229–235.
12. Wallace C, Smyth DJ, Maisuria-Armer M, Walker NM, Todd JA, et al. (2010) The imprinted DLK1-MEG3 gene region on chromosome 14q32.2 alters susceptibility to type 1 diabetes. Nat Genet 42: 68–71.
13. Podolsky DK (2002) Inflammatory bowel disease. N Engl J Med 347: 417–429.



14. Raelson JV, Little RD, Ruether A, Fournier H, Paquin B, et al. (2007) Genome-wide association study for Crohn's disease in the Quebec Founder Population identifies multiple validated disease loci. Proc Natl Acad Sci U S A 104: 14747–14752.
15. Cortes A, Brown MA (2011) Promise and pitfalls of the Immunochip. Arthritis Res Ther 13: 101.
16. Trynka G, Hunt KA, Bockett NA, Romanos J, Mistry V, et al. (2011) Dense genotyping identifies and localizes multiple common and rare variant association signals in celiac disease. Nat Genet 43: 1193–1201.
17. Johnson AD, Handsaker RE, Pulit SL, Nizzari MM, O'Donnell CJ, et al. (2008) SNAP: a web-based tool for identification and annotation of proxy SNPs using HapMap. Bioinformatics 24: 2938–2939.
18. Juyal G, Prasad P, Senapati S, Midha V, Sood A, et al. (2011) An investigation of genome-wide studies reported susceptibility loci for ulcerative colitis shows limited replication in north Indians. PLoS One 6: e16565.
19. Luedi PP, Dietrich FS, Weidman JR, Bosko JM, Jirtle RL, et al. (2007) Computational and experimental identification of novel human imprinted genes. Genome Res 17: 1723–1730.
20. Mitrovic M, Potocnik U (2011) High-resolution melting curve analysis for high-throughput genotyping of NOD2/CARD15 mutations and distribution of these mutations in Slovenian inflammatory bowel diseases patients. Dis Markers 30: 265–274.
21. Gauderman WJ (2002) Sample size requirements for matched case-control studies of gene-environment interaction. Stat Med 21: 35–50.
22. Schnitzler F, Seiderer J, Stallhofer J, Brand S (2012) Dominant disease-causing effect of NOD2 mutations in a family with all family members affected by Crohn's disease. Inflamm Bowel Dis 18: 395–396.
23. Jurgens M, Brand S, Laubender RP, Seiderer J, Glas J, et al. (2010) The presence of fistulas and NOD2 homozygosity strongly predict intestinal stenosis in Crohn's disease independent of the IL23R genotype. J Gastroenterol 45: 721–731.
24. Brand S (2012) Homozygosity for the NOD2 p.Leu1007fsX1008 variant is the main genetic predictor for fibrostenotic Crohn's disease. Inflamm Bowel Dis 18: 393–394.
25. Australo-Anglo-American Spondyloarthritis C, Reveille JD, Sims AM, Danoy P, Evans DM, et al. (2010) Genome-wide association study of ankylosing spondylitis identifies non-MHC susceptibility loci. Nat Genet 42: 123–127.
26. Hollis-Moffatt JE, Merriman ME, Rodger RA, Rowley KA, Chapman PT, et al. (2009) Evidence for association of an interleukin 23 receptor variant independent of the R381Q variant with rheumatoid arthritis. Ann Rheum Dis 68: 1340–1344.
27. Nair RP, Duffin KC, Helms C, Ding J, Stuart PE, et al. (2009) Genome-wide scan reveals association of psoriasis with IL-23 and NF-kappaB pathways. Nat Genet 41: 199–204.
28. Raychaudhuri S, Thomson BP, Remmers EF, Eyre S, Hinks A, et al. (2009) Genetic variants at CD28, PRDM1 and CD2/CD58 are associated with rheumatoid arthritis risk. Nat Genet 41: 1313–1318.
29. Gateva V, Sandling JK, Hom G, Taylor KE, Chung SA, et al. (2009) A large-scale replication study identifies TNIP1, PRDM1, JAZF1, UHRF1BP1 and IL10 as risk loci for systemic lupus erythematosus. Nat Genet 41: 1228–1233.
30. Sokol L (2011) Fox and Blimp in NK-cell lymphoma. Blood 118: 3192–3193.
31. Best T, Li D, Skol AD, Kirchhoff T, Jackson SA, et al. (2011) Variants at 6q21 implicate PRDM1 in the etiology of therapy-induced second malignancies after Hodgkin's lymphoma. Nat Med 17: 941–943.
32. Gleason G, Liu B, Bruening S, Zupan B, Auerbach A, et al. (2010) The serotonin1A receptor gene as a genetic and prenatal maternal environmental factor in anxiety. Proc Natl Acad Sci U S A 107: 7592–7597.
33. Fay JC, Wyckoff GJ, Wu CI (2001) Positive and negative selection on the human genome. Genetics 158: 1227–1234.
34. Guilmatre A, Sharp AJ (2011) Parent of origin effects. Clin Genet.
35. Yazbek SN, Spiezio SH, Nadeau JH, Buchner DA (2010) Ancestral paternal genotype controls body weight and food intake for multiple generations. Hum Mol Genet 19: 4134–4144.
36. Thompson SL, Konfortova G, Gregory RI, Reik W, Dean W, et al. (2001) Environmental effects on genomic imprinting in mammals. Toxicol Lett 120: 143–150.
37. Wang S, Yu Z, Miller RL, Tang D, Perera FP (2011) Methods for detecting interactions between imprinted genes and environmental exposures using birth cohort designs with mother-offspring pairs. Hum Hered 71: 196–208.

6.6.5 Izvirni znanstveni članek 4

LETTER

doi:10.1038/nature11582

Host–microbe interactions have shaped the genetic architecture of inflammatory bowel disease

A list of authors and their affiliations appears at the end of the paper.

Crohn's disease and ulcerative colitis, the two common forms of inflammatory bowel disease (IBD), affect over 2.5 million people of European ancestry, with rising prevalence in other populations[1]. Genome-wide association studies and subsequent meta-analyses of these two diseases[2,3] as separate phenotypes have implicated previously unsuspected mechanisms, such as autophagy[4], in their pathogenesis and showed that some IBD loci are shared with other inflammatory diseases[5]. Here we expand on the knowledge of relevant pathways by undertaking a meta-analysis of Crohn's disease and ulcerative colitis genome-wide association scans, followed by extensive validation of significant findings, with a combined total of more than 75,000 cases and controls. We identify 71 new associations, for a total of 163 IBD loci, that meet genome-wide significance thresholds. Most loci contribute to both phenotypes, and both directional (consistently favouring one allele over the course of human history) and balancing (favouring the retention of both alleles within populations) selection effects are evident. Many IBD loci are also implicated in other immune-mediated disorders, most notably with ankylosing spondylitis and psoriasis. We also observe considerable overlap between susceptibility loci for IBD and mycobacterial infection. Gene co-expression network analysis emphasizes this relationship, with pathways shared between host responses to mycobacteria and those predisposing to IBD.

We conducted an imputation-based association analysis using autosomal genotype-level data from 15 genome-wide association studies (GWAS) of Crohn's disease and/or ulcerative colitis (Supplementary Fig. 1 and Supplementary Table 1). We imputed 1.23 million single-nucleotide polymorphisms (SNPs) from the HapMap3 reference set (Supplementary Methods 1a), resulting in a high-quality data set with reduced genome-wide inflation (Supplementary Figs 2 and 3) compared with previous meta-analyses of subsets of these data[2,3]. The imputed GWAS data identified 25,075 SNPs that were associated ($P < 0.01$) with at least one of the Crohn's disease, ulcerative colitis, or combined IBD analyses. A meta-analysis of GWAS data with Immunochip[6] validation genotypes from an independent, newly genotyped set of 14,763 Crohn's disease cases, 10,920 ulcerative colitis cases and 15,977 controls was performed (Supplementary Fig. 1 and Supplementary Table 1). Principal-components analysis resolved geographic stratification, as well as Jewish and non-Jewish ancestry (Supplementary Fig. 4), and reduced inflation to a level consistent with residual polygenic risk, rather than other confounding effects (from a median test statistic inflation (λ_{GC}) = 2.00 to $\lambda_{GC} = 1.23$ when analysing all IBD samples; Supplementary Fig. 5 and Supplementary Methods 1b).

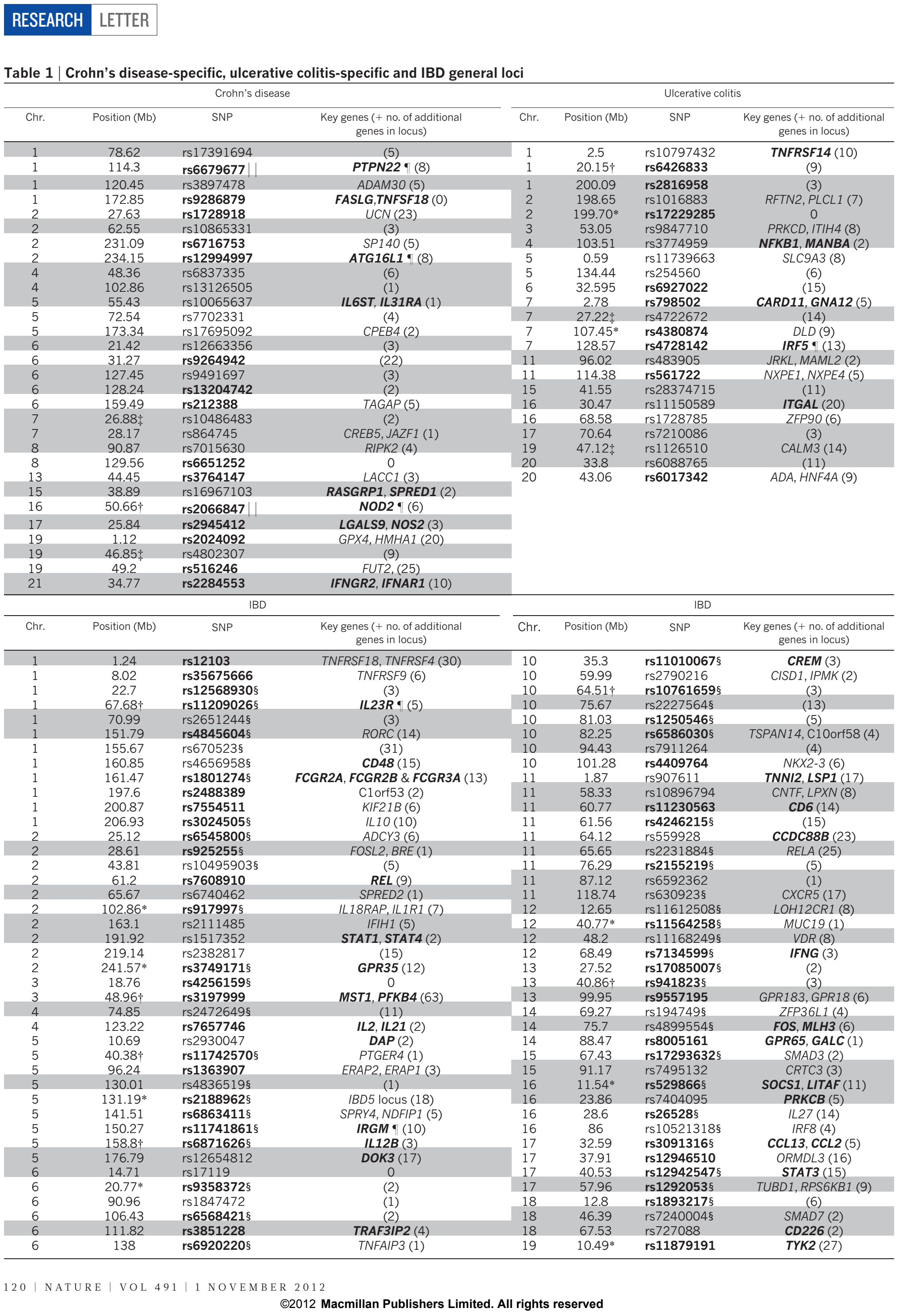
Our meta-analysis of the GWAS and Immunochip data identified 193 statistically independent signals of association at genome-wide significance ($P < 5 \times 10^{-8}$) in at least one of the three analyses (Crohn's disease, ulcerative colitis, IBD). Because some of these signals (Supplementary Fig. 6) probably represent associations to the same underlying functional unit, we merged these signals (Supplementary Methods 1b) into 163 regions, 71 of which are reported here for the first time (Table 1 and Supplementary Table 2). Fig. 1a shows the relative contributions of each locus to the total variance explained in ulcerative colitis and Crohn's disease. We have increased the total disease variance explained (variance being subject to fewer assumptions than heritability[7]) from 8.2% to 13.6% in Crohn's disease and from 4.1% to 7.5% in ulcerative colitis (Supplementary Methods 1c). Consistent with previous studies, our IBD risk loci seem to act independently, with no significant evidence of deviation from an additive combination of log odds ratios.

Our combined genome-wide analysis of Crohn's disease and ulcerative colitis enables a more comprehensive analysis of disease specificity than was previously possible. A model-selection analysis (Supplementary Methods 1c) showed that 110 out of 163 loci are associated with both disease phenotypes; 50 of these have an indistinguishable effect size in ulcerative colitis and Crohn's disease, whereas 60 show evidence of heterogeneous effects (Table 1). Of the remaining loci, 30 are classified as Crohn's-disease-specific and 23 as ulcerative-colitis-specific. However, 43 of these 53 loci show the same direction of effect in the non-associated disease (Fig. 1b; overall $P = 2.8 \times 10^{-6}$). Risk alleles at two Crohn's disease loci, *PTPN22* and *NOD2*, show significant ($P < 0.005$) protective effects in ulcerative colitis, exceptions that may reflect biological differences between the two diseases. This degree of sharing of genetic risk suggests that nearly all of the biological mechanisms involved in one disease have some role in the other.

The large number of IBD associations, far more than reported for any other complex disease, increases the power of network-based analyses to prioritize genes within loci. We investigated the IBD loci using functional annotation and empirical gene network tools (Supplementary Table 2). Compared with previous analyses that identified candidate genes in 35% of loci[2,3] our updated GRAIL[8] -connectivity network identifies candidates in 53% of loci, including increased statistical significance for 58 of the 73 candidates from previous analyses. The new candidates come not only from genes within newly identified loci, but also integrate additional genes from previously established loci (Fig. 1c). Only 29 IBD-associated SNPs are in strong linkage disequilibrium ($r^2 > 0.8$) with a missense variant in the 1000 Genomes Project data, which reinforces previous evidence that a large fraction of risk for complex disease is driven by non-coding variation. By contrast, 64 IBD-associated SNPs are in linkage disequilibrium with variants known to regulate gene expression (Supplementary Table 2). Overall, we highlighted a total of 300 candidate genes in 125 loci, of which 39 contained a single gene supported by two or more methods.

Seventy per cent (113 out of 163) of the IBD loci are shared with other complex diseases or traits, including 66 among the 154 loci previously associated with other immune-mediated diseases[9], which is 8.6-times the number that would be expected by chance ($P < 10^{-16}$; Fig. 2a and Supplementary Fig. 7). Such enrichment cannot be attributed to the immune-mediated focus of the Immunochip (Supplementary Methods 4 and Supplementary Fig. 8), as the analysis is based on our combined GWAS–Immunochip data. Comparing overlaps with specific diseases is confounded by the variable power in studies of different diseases. For instance, although type 1 diabetes shares the largest number of loci (20 out of 39; tenfold enrichment) with IBD, this is partially driven by the large number of known type 1 diabetes associations. Indeed, seven other immune-mediated diseases

©2012 Macmillan Publishers Limited. All rights reserved



RESEARCH LETTER

Table 1 | Crohn's disease-specific, ulcerative colitis-specific and IBD general loci

Crohn's disease				Ulcerative colitis			
Chr.	Position (Mb)	SNP	Key genes (+ no. of additional genes in locus)	Chr.	Position (Mb)	SNP	Key genes (+ no. of additional genes in locus)
1	78.62	rs17391694	(5)	1	2.5	rs10797432	***TNFRSF14*** (10)
1	114.3	**rs6679677**‖	***PTPN22*** ¶ (8)	1	20.15†	**rs6426833**	(9)
1	120.45	rs3897478	*ADAM30* (5)	1	200.09	**rs2816958**	(3)
1	172.85	**rs9286879**	***FASLG***,***TNFSF18*** (0)	2	198.65	rs1016883	*RFTN2*, *PLCL1* (7)
2	27.63	**rs1728918**	*UCN* (23)	2	199.70*	**rs17229285**	0
2	62.55	rs10865331	(3)	3	53.05	rs9847710	*PRKCD*, *ITIH4* (8)
2	231.09	**rs6716753**	*SP140* (5)	4	103.51	rs3774959	***NFKB1***, ***MANBA*** (2)
2	234.15	**rs12994997**	***ATG16L1*** ¶ (8)	5	0.59	rs11739663	*SLC9A3* (8)
4	48.36	rs6837335	(6)	5	134.44	rs254560	(6)
4	102.86	rs13126505	(1)	6	32.595	**rs6927022**	(15)
5	55.43	rs10065637	***IL6ST***, ***IL31RA*** (1)	7	2.78	**rs798502**	***CARD11***, ***GNA12*** (5)
5	72.54	rs7702331	(4)	7	27.22‡	rs4722672	(14)
5	173.34	rs17695092	*CPEB4* (2)	7	107.45*	**rs4380874**	*DLD* (9)
6	21.42	rs12663356	(3)	7	128.57	**rs4728142**	***IRF5*** ¶ (13)
6	31.27	**rs9264942**	(22)	11	96.02	rs483905	*JRKL*, *MAML2* (2)
6	127.45	rs9491697	(3)	11	114.38	**rs561722**	*NXPE1*, *NXPE4* (5)
6	128.24	**rs13204742**	(2)	15	41.55	rs28374715	(11)
6	159.49	**rs212388**	*TAGAP* (5)	16	30.47	rs11150589	***ITGAL*** (20)
7	26.88‡	rs10486483	(2)	16	68.58	rs1728785	*ZFP90* (6)
7	28.17	rs864745	*CREB5*, *JAZF1* (1)	17	70.64	rs7210086	(3)
8	90.87	rs7015630	*RIPK2* (4)	19	47.12‡	rs1126510	*CALM3* (14)
8	129.56	**rs6651252**	0	20	33.8	rs6088765	(11)
13	44.45	**rs3764147**	*LACC1* (3)	20	43.06	**rs6017342**	*ADA*, *HNF4A* (9)
15	38.89	rs16967103	***RASGRP1***, ***SPRED1*** (2)				
16	50.66†	**rs2066847**‖	***NOD2*** ¶ (6)				
17	25.84	**rs2945412**	***LGALS9***, ***NOS2*** (3)				
19	1.12	**rs2024092**	*GPX4*, *HMHA1* (20)				
19	46.85‡	rs4802307	(9)				
19	49.2	**rs516246**	*FUT2*, (25)				
21	34.77	**rs2284553**	***IFNGR2***, ***IFNAR1*** (10)				

IBD				IBD			
Chr.	Position (Mb)	SNP	Key genes (+ no. of additional genes in locus)	Chr.	Position (Mb)	SNP	Key genes (+ no. of additional genes in locus)
1	1.24	**rs12103**	*TNFRSF18*, *TNFRSF4* (30)	10	35.3	**rs11010067**§	***CREM*** (3)
1	8.02	**rs35675666**	*TNFRSF9* (6)	10	59.99	rs2790216	*CISD1*, *IPMK* (2)
1	22.7	**rs12568930**§	(3)	10	64.51†	**rs10761659**§	(3)
1	67.68†	**rs11209026**§	***IL23R*** ¶ (5)	10	75.67	rs2227564§	(13)
1	70.99	rs2651244§	(3)	10	81.03	**rs1250546**§	(5)
1	151.79	**rs4845604**§	*RORC* (14)	10	82.25	**rs6586030**§	*TSPAN14*, C10orf58 (4)
1	155.67	rs670523§	(31)	10	94.43	rs7911264	(4)
1	160.85	rs4656958§	***CD48*** (15)	10	101.28	**rs4409764**	*NKX2-3* (6)
1	161.47	**rs1801274**§	***FCGR2A***, ***FCGR2B*** & ***FCGR3A*** (13)	11	1.87	rs907611	***TNNI2***, ***LSP1*** (17)
1	197.6	**rs2488389**	C1orf53 (2)	11	58.33	rs10896794	*CNTF*, *LPXN* (8)
1	200.87	**rs7554511**	*KIF21B* (6)	11	60.77	**rs11230563**	***CD6*** (14)
1	206.93	**rs3024505**§	*IL10* (10)	11	61.56	**rs4246215**§	(15)
2	25.12	**rs6545800**§	*ADCY3* (6)	11	64.12	rs559928	***CCDC88B*** (23)
2	28.61	**rs925255**§	*FOSL2*, *BRE* (1)	11	65.65	rs2231884§	*RELA* (25)
2	43.81	rs10495903§	(5)	11	76.29	**rs2155219**§	(5)
2	61.2	**rs7608910**	***REL*** (9)	11	87.12	rs6592362	(1)
2	65.67	rs6740462	*SPRED2* (1)	11	118.74	rs630923§	*CXCR5* (17)
2	102.86*	**rs917997**§	*IL18RAP*, *IL1R1* (7)	12	12.65	rs11612508§	*LOH12CR1* (8)
2	163.1	rs2111485	*IFIH1* (5)	12	40.77*	**rs11564258**§	*MUC19* (1)
2	191.92	rs1517352	***STAT1***, ***STAT4*** (2)	12	48.2	rs11168249§	*VDR* (8)
2	219.14	rs2382817	(15)	12	68.49	**rs7134599**§	***IFNG*** (3)
2	241.57*	**rs3749171**§	***GPR35*** (12)	13	27.52	**rs17085007**§	(2)
3	18.76	**rs4256159**§	0	13	40.86†	**rs941823**§	(3)
3	48.96†	**rs3197999**	***MST1***, ***PFKB4*** (63)	13	99.95	**rs9557195**	*GPR183*, *GPR18* (6)
4	74.85	rs2472649§	(11)	14	69.27	rs194749§	*ZFP36L1* (4)
4	123.22	**rs7657746**	***IL2***, ***IL21*** (2)	14	75.7	rs4899554§	***FOS***, ***MLH3*** (6)
5	10.69	rs2930047	***DAP*** (2)	14	88.47	**rs8005161**	***GPR65***, ***GALC*** (1)
5	40.38†	**rs11742570**§	*PTGER4* (1)	15	67.43	**rs17293632**§	*SMAD3* (2)
5	96.24	**rs1363907**	*ERAP2*, *ERAP1* (3)	15	91.17	rs7495132	*CRTC3* (3)
5	130.01	rs4836519§	(1)	16	11.54*	**rs529866**§	***SOCS1***, ***LITAF*** (11)
5	131.19*	**rs2188962**§	*IBD5* locus (18)	16	23.86	rs7404095	***PRKCB*** (5)
5	141.51	**rs6863411**§	*SPRY4*, *NDFIP1* (5)	16	28.6	**rs26528**§	*IL27* (14)
5	150.27	**rs11741861**§	***IRGM*** ¶ (10)	16	86	rs10521318§	*IRF8* (4)
5	158.8†	**rs6871626**§	***IL12B*** (3)	17	32.59	**rs3091316**§	***CCL13***, ***CCL2*** (5)
5	176.79	rs12654812	***DOK3*** (17)	17	37.91	**rs12946510**	*ORMDL3* (16)
6	14.71	rs17119	0	17	40.53	**rs12942547**§	***STAT3*** (15)
6	20.77*	**rs9358372**§	(2)	17	57.96	**rs1292053**§	*TUBD1*, *RPS6KB1* (9)
6	90.96	rs1847472	(1)	18	12.8	**rs1893217**§	(6)
6	106.43	**rs6568421**§	(2)	18	46.39	rs7240004§	*SMAD7* (2)
6	111.82	**rs3851228**	***TRAF3IP2*** (4)	18	67.53	rs727088	***CD226*** (2)
6	138	**rs6920220**§	*TNFAIP3* (1)	19	10.49*	**rs11879191**	***TYK2*** (27)

©2012 Macmillan Publishers Limited. All rights reserved

Table 1 | Continued

IBD				IBD			
Chr.	Position (Mb)	SNP	Key genes (+ no. of additional genes in locus)	Chr.	Position (Mb)	SNP	Key genes (+ no. of additional genes in locus)
6	143.9	rs12199775	*PHACTR2* (5)	19	33.73	**rs17694108**	*CEBPG (8)*
6	167.37	**rs1819333**§	***CCR6***, ***RPS6KA2*** (4)	19	55.38	rs11672983	(19)
7	50.245*	**rs1456896**	*ZPBP*, *IKZF1* (4)	20	30.75	rs6142618§	*HCK* (10)
7	98.75	rs9297145	*SMURF1* (6)	20	31.37	rs4911259	*DNMT3B* (8)
7	100.34	**rs1734907**§	***EPO*** (21)	20	44.74	**rs1569723**§	***CD40*** (13)
7	116.89	rs38904§	(6)	20	48.95	rs913678	*CEBPB* (5)
8	126.53	**rs921720**§	*TRIB1* (1)	20	57.82	rs259964	*ZNF831*, *CTSZ* (5)
8	130.62	rs1991866	(2)	20	62.34	**rs6062504**	*TNFRSF6B* (26)
9	4.98	**rs10758669**	***JAK2*** (4)	21	16.81	**rs2823286**§	0
9	93.92	rs4743820§	*NFIL3* (2)	21	40.46	**rs2836878**§	(3)
9	117.60†	**rs4246905**	*TNFSF15* (4)	21	45.62	**rs7282490**	*ICOSLG* (9)
9	139.32*	**rs10781499**§	***CARD9*** (22)	22	21.92	**rs2266959**	(13)
10	6.08	rs12722515§	***IL2RA***, ***IL15RA*** (6)	22	30.43	**rs2412970**	***LIF***, ***OSM*** (9)
10	30.72	rs1042058§	***MAP3K8*** (3)	22	39.69*	**rs2413583**§	*TAB1* (18)

The position given is the middle of the locus window, with all positions relative to human reference genome GRCh37. Bolded rs numbers indicate SNPs with P values less than 1×10^{-13}. Grey shading indicates newly discovered loci. Listed are genes implicated by one or more candidate gene approaches. Bolded genes have been implicated by two or more candidate gene approaches. For each locus, the top two candidate genes are listed. A complete listing of gene prioritization is provided in Supplementary Table 2. *Additional genome-wide significant associated SNP in the region. †Two or more additional genome-wide significant SNPs in the region. ‡These regions have overlapping but distinct ulcerative colitis and Crohn's disease signals. §Heterogeneity of odds ratios. ‖ Crohn's disease risk allele is significantly protective in ulcerative colitis. ¶Gene for which functional studies of associated alleles have been reported. Chr., chromosome; Mb, megabase.

show stronger enrichment of overlap, with the largest being ankylosing spondylitis (8 out of 11; 13-fold) and psoriasis (14 out of 17; 14-fold).

IBD loci are also markedly enriched (4.9-fold; $P < 10^{-4}$) in genes involved in primary immunodeficiencies (PIDs; Fig. 2a), which are characterized by a dysfunctional immune system resulting in severe infections[10]. Genes implicated in this overlap correlate with reduced levels of circulating T cells (*ADA*, *CD40*, *TAP1*, *TAP2*, *NBN*, *BLM*, *DNMT3B*) or of specific subsets, such as T-helper cells producing IL-17 (T_H17 cells) (*STAT3*), memory (*SP110*) or regulatory T cells (*STAT5B*). The subset of PID genes leading to Mendelian susceptibility to mycobacterial disease (MSMD)[10–12] is enriched still further; six of the eight known autosomal genes linked to MSMD are located within IBD loci (*IL12B*, *IFNGR2*, *STAT1*, *IRF8*, *TYK2*, *STAT3*; 46-fold enrichment; $P = 1.3 \times 10^{-6}$), and a seventh, *IFNGR1*, narrowly missed genome-wide significance ($P = 6 \times 10^{-8}$). Overlap with IBD is also seen in complex mycobacterial disease; we find IBD associations in seven out of eight loci identified by leprosy GWAS[13], including six cases in which the same SNP is implicated. Furthermore, genetic defects in *STAT3* (refs 14, 15) and *CARD9* (ref. 16), also within IBD loci, lead to PIDs involving skin infections with *Staphylococcus* and candidiasis, respectively. The comparative effects of IBD and infectious-disease-susceptibility-risk alleles on gene function and expression are summarized in Supplementary Table 3, and include both opposite (for example, *NOD2* and *STAT3*; Supplementary Fig. 9) and similar (for example, *IFNGR2*) directional effects.

To extend our understanding of the fundamental biology of IBD pathogenesis we conducted searches across the IBD locus list: (1) for enrichment of specific Gene Ontology terms and canonical pathways; (2) for evidence of selective pressure acting on specific variants and pathways; and (3) for enrichment of differentially expressed genes across immune-cell types. We tested the 300 prioritized genes (see above) for enrichment in Gene Ontology terms (Supplementary Methods 4a) and identified 286 Gene Ontology terms and 56 pathways demonstrating significant enrichment in genes contained within IBD loci (Supplementary Figs 10 and 11 and Supplementary Table 4). Excluding high-level Gene Ontology categories such as 'immune system processes' ($P = 3.5 \times 10^{-26}$), the most significantly enriched term is regulation of cytokine production ($P = 2.7 \times 10^{-24}$), specifically

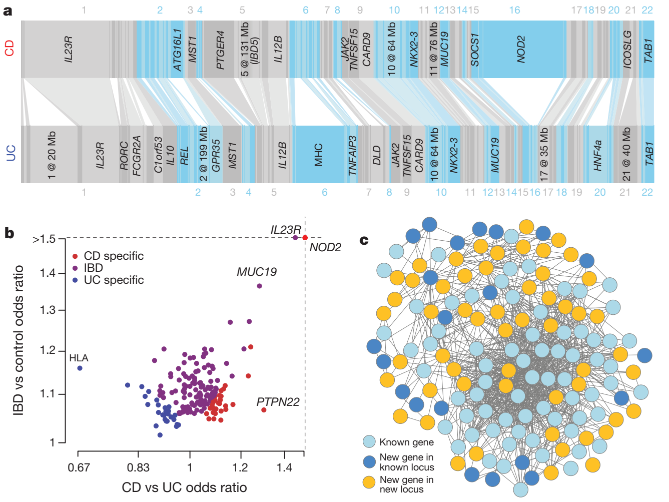


Figure 1 | The IBD genome. **a**, Variance explained by the 163 IBD loci. Each bar, ordered by genomic position, represents an independent locus. The width of the bar is proportional to the variance explained by that locus in Crohn's disease (CD) and ulcerative colitis (UC). Bars are connected together if they are identified as being associated with both phenotypes, and loci are labelled if they explain more than 1% of the total variance explained by all loci for that phenotype. Labels are either the best-supported candidate gene in Table 1, or the chromosome and position of the locus if either no, or multiple, well-supported candidates exist. **b**, The 193 independent signals, plotted by total IBD odds ratio and phenotype specificity (measured by the odds ratio of Crohn's disease relative to ulcerative colitis), and coloured by their IBD phenotype classification from Table 1. Note that many loci (for example, *IL23R*) show very different effects in Crohn's disease and ulcerative colitis despite being strongly associated to both. **c**, GRAIL network for all genes with GRAIL $P < 0.05$. Genes included in our previous GRAIL networks in both phenotypes are shown in light blue, newly connected genes in previously identified loci in dark blue, and genes from newly associated loci in gold. The gold genes reinforce the previous network (light blue) and expand it to include dark blue genes.

©2012 Macmillan Publishers Limited. All rights reserved

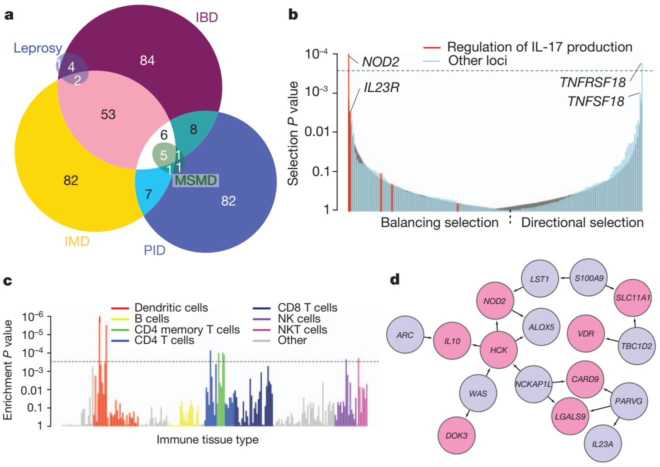


Figure 2 | Dissecting the biology of IBD. **a**, Number of overlapping IBD loci with other immune-mediated diseases (IMD), leprosy and Mendelian PIDs. Within PID, we highlight MSMD. **b**, Signals of selection at IBD SNPs, from strongest balancing on the left to strongest directional on the right. The grey curve shows the 95% confidence interval for randomly chosen frequency-matched SNPs, illustrating our overall enrichment ($P = 5.5 \times 10^{-6}$), and the dashed line represents the Bonferroni significance threshold. SNPs highlighted in red are annotated as being involved in the regulation of IL-17 production, a key IBD functional term related to bacterial defence, and are enriched for balancing selection. **c**, Evidence of enrichment in IBD loci of differentially expressed genes from various immune tissues. Each bar represents the empirical *P* value in a single tissue, and the colours represent different cell type groupings. The dashed line is Bonferroni-corrected significance for the number of tissues tested. **d**, *NOD2*-focused cluster of the IBD causal sub-network. Pink genes are in IBD-associated loci, blue are not. Arrows indicate inferred causal direction of regulation of expression.

interferon-γ, interleukin (IL)-12, tumour-necrosis factor-α and IL-10 signalling. Lymphocyte activation was the next most significant ($P = 1.8 \times 10^{-23}$), with activation of T cells, B cells and natural killer (NK) cells being the strongest contributors to this signal. Strong enrichment was also seen for response to molecules of bacterial origin ($P = 2.4 \times 10^{-20}$), and for the Kyoto Encyclopedia of Genes and Genomes (KEGG) JAK-STAT signalling pathway ($P = 4.8 \times 10^{-15}$). We note that no enriched terms or pathways showed specific evidence of Crohn's disease or ulcerative colitis specificity.

As infectious organisms are known to be among the strongest agents of natural selection, we investigated whether the IBD-associated variants are subject to selective pressures (Supplementary Table 5 and Supplementary Methods 4c). Directional selection would imply that the balance between these forces shifted in one direction over the course of human history, whereas balancing selection would suggest an allele-frequency-dependent scenario typified by host–microbe co-evolution, as can be observed with parasites. Two SNPs show Bonferroni-significant selection: the most significant signal, in *NOD2*, is under balancing selection ($P = 5.2 \times 10^{-5}$), and the second most significant, in the receptor *TNFRSF18*, showed directional selection ($P = 8.9 \times 10^{-5}$). The next most significant variants were in the ligand of that receptor, *TNFSF18* (directional; $P = 5.2 \times 10^{-4}$), and *IL23R* (balancing; $P = 1.5 \times 10^{-3}$). As a group, the IBD variants show significant enrichment in selection (Fig. 2b) of both types ($P = 5.5 \times 10^{-6}$). We discovered an enrichment of balancing selection (Fig. 2b) in genes annotated with the Gene Ontology term 'regulation of interleukin-17 production' ($P = 1.4 \times 10^{-4}$). The important role of IL-17 in both bacterial defence and autoimmunity suggests a key role for balancing selection in maintaining the genetic relationship between inflammation and infection, and this is reinforced by a nominal enrichment of balancing selection in loci annotated with the broader Gene Ontology term 'defense response to bacterium' ($P = 0.007$).

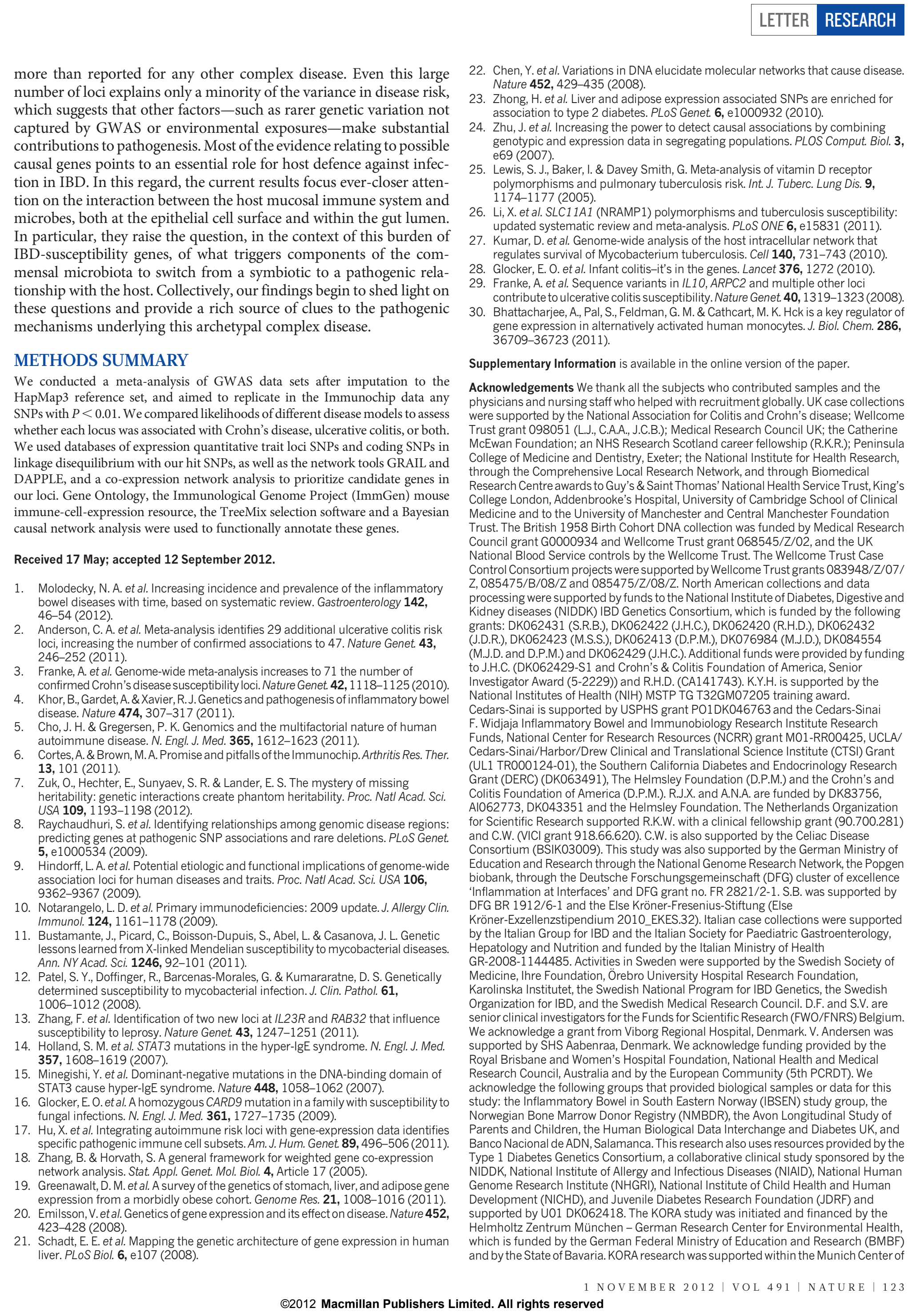
We tested for enrichment of cell-type expression specificity of genes in IBD loci in 223 distinct sets of sorted, mouse-derived immune cells from the Immunological Genome Consortium[17]. Dendritic cells showed the strongest enrichment, followed by weaker signals that support the Gene Ontology analysis, including CD4$^+$ T cells, NK cells and NKT cells (Fig. 2c). Notably, several of these cell types express genes near our IBD associations much more specifically when stimulated; our strongest signal, a lung-derived dendritic cell, had $P_{\text{stimulated}} < 1 \times 10^{-6}$ compared with $P_{\text{unstimulated}} = 0.0015$, consistent with an important role for cell activation.

To further our goal of identifying likely causal genes within our susceptibility loci and to elucidate networks underlying IBD pathogenesis, we screened the associated genes against 211 co-expression modules identified from weighted gene co-expression network analyses[18], conducted with large gene-expression data sets from multiple tissues[19–21]. The most significantly enriched module comprised 523 genes from omental adipose tissue collected from morbidly obese patients[19], which was found to be 2.9-fold enriched for genes in the IBD-associated loci ($P = 1.1 \times 10^{-13}$; Supplementary Fig. 12 and Supplementary Table 6). We constructed a probabilistic causal gene network using an integrative Bayesian network-reconstruction algorithm[22–24], which combines expression and genotype data to infer the direction of causality between genes with correlated expression. The intersection of this network and the genes in the IBD-enriched module defined a sub-network of genes enriched in bone marrow-derived macrophages ($P < 10^{-16}$) and is suggestive of dynamic interactions relevant to IBD pathogenesis. In particular, this sub-network featured close proximity among genes connected to host interaction with bacteria, notably *NOD2*, *IL10* and *CARD9*.

A *NOD2*-focused inspection of the sub-network prioritizes multiple additional candidate genes within IBD-associated regions. For example, a cluster near *NOD2* (Fig. 2d) contains multiple IBD genes implicated in the *Mycobacterium tuberculosis* response, including *SLC11A1*, *VDR* and *LGALS9*. Furthermore, both *SLC11A1* (also known as *NRAMP1*) and *VDR* have been associated with *M. tuberculosis* infection by candidate gene studies[25,26], and *LGALS9* modulates mycobacteriosis[27]. Of interest, *HCK* (located in our new locus on chromosome 20 at 30.75 megabases) is predicted to upregulate expression of both *NOD2* and *IL10*, an anti-inflammatory cytokine associated with Mendelian[28] and non-Mendelian [29] IBD. *HCK* has been linked to alternative, anti-inflammatory activation of monocytes (M2-group macrophages)[30]; although not identified in our aforementioned analyses, these data implicate *HCK* as the causal gene in this new IBD locus.

We report one of the largest genetic experiments involving a complex disease undertaken to date. This has increased the number of confirmed IBD susceptibility loci to 163, most of which are associated with both Crohn's disease and ulcerative colitis, and is substantially

©2012 Macmillan Publishers Limited. All rights reserved



more than reported for any other complex disease. Even this large number of loci explains only a minority of the variance in disease risk, which suggests that other factors—such as rarer genetic variation not captured by GWAS or environmental exposures—make substantial contributions to pathogenesis. Most of the evidence relating to possible causal genes points to an essential role for host defence against infection in IBD. In this regard, the current results focus ever-closer attention on the interaction between the host mucosal immune system and microbes, both at the epithelial cell surface and within the gut lumen. In particular, they raise the question, in the context of this burden of IBD-susceptibility genes, of what triggers components of the commensal microbiota to switch from a symbiotic to a pathogenic relationship with the host. Collectively, our findings begin to shed light on these questions and provide a rich source of clues to the pathogenic mechanisms underlying this archetypal complex disease.

METHODS SUMMARY

We conducted a meta-analysis of GWAS data sets after imputation to the HapMap3 reference set, and aimed to replicate in the Immunochip data any SNPs with $P < 0.01$. We compared likelihoods of different disease models to assess whether each locus was associated with Crohn's disease, ulcerative colitis, or both. We used databases of expression quantitative trait loci SNPs and coding SNPs in linkage disequilibrium with our hit SNPs, as well as the network tools GRAIL and DAPPLE, and a co-expression network analysis to prioritize candidate genes in our loci. Gene Ontology, the Immunological Genome Project (ImmGen) mouse immune-cell-expression resource, the TreeMix selection software and a Bayesian causal network analysis were used to functionally annotate these genes.

Received 17 May; accepted 12 September 2012.

1. Molodecky, N. A. *et al.* Increasing incidence and prevalence of the inflammatory bowel diseases with time, based on systematic review. *Gastroenterology* **142,** 46–54 (2012).
2. Anderson, C. A. *et al.* Meta-analysis identifies 29 additional ulcerative colitis risk loci, increasing the number of confirmed associations to 47. *Nature Genet.* **43,** 246–252 (2011).
3. Franke, A. *et al.* Genome-wide meta-analysis increases to 71 the number of confirmed Crohn's disease susceptibility loci. *Nature Genet.* **42,** 1118–1125 (2010).
4. Khor, B., Gardet, A. & Xavier, R. J. Genetics and pathogenesis of inflammatory bowel disease. *Nature* **474,** 307–317 (2011).
5. Cho, J. H. & Gregersen, P. K. Genomics and the multifactorial nature of human autoimmune disease. *N. Engl. J. Med.* **365,** 1612–1623 (2011).
6. Cortes, A. & Brown, M. A. Promise and pitfalls of the Immunochip. *Arthritis Res. Ther.* **13,** 101 (2011).
7. Zuk, O., Hechter, E., Sunyaev, S. R. & Lander, E. S. The mystery of missing heritability: genetic interactions create phantom heritability. *Proc. Natl Acad. Sci. USA* **109,** 1193–1198 (2012).
8. Raychaudhuri, S. *et al.* Identifying relationships among genomic disease regions: predicting genes at pathogenic SNP associations and rare deletions. *PLoS Genet.* **5,** e1000534 (2009).
9. Hindorff, L. A. *et al.* Potential etiologic and functional implications of genome-wide association loci for human diseases and traits. *Proc. Natl Acad. Sci. USA* **106,** 9362–9367 (2009).
10. Notarangelo, L. D. *et al.* Primary immunodeficiencies: 2009 update. *J. Allergy Clin. Immunol.* **124,** 1161–1178 (2009).
11. Bustamante, J., Picard, C., Boisson-Dupuis, S., Abel, L. & Casanova, J. L. Genetic lessons learned from X-linked Mendelian susceptibility to mycobacterial diseases. *Ann. NY Acad. Sci.* **1246,** 92–101 (2011).
12. Patel, S. Y., Doffinger, R., Barcenas-Morales, G. & Kumararatne, D. S. Genetically determined susceptibility to mycobacterial infection. *J. Clin. Pathol.* **61,** 1006–1012 (2008).
13. Zhang, F. *et al.* Identification of two new loci at *IL23R* and *RAB32* that influence susceptibility to leprosy. *Nature Genet.* **43,** 1247–1251 (2011).
14. Holland, S. M. *et al. STAT3* mutations in the hyper-IgE syndrome. *N. Engl. J. Med.* **357,** 1608–1619 (2007).
15. Minegishi, Y. *et al.* Dominant-negative mutations in the DNA-binding domain of STAT3 cause hyper-IgE syndrome. *Nature* **448,** 1058–1062 (2007).
16. Glocker, E. O. *et al.* A homozygous *CARD9* mutation in a family with susceptibility to fungal infections. *N. Engl. J. Med.* **361,** 1727–1735 (2009).
17. Hu, X. *et al.* Integrating autoimmune risk loci with gene-expression data identifies specific pathogenic immune cell subsets. *Am. J. Hum. Genet.* **89,** 496–506 (2011).
18. Zhang, B. & Horvath, S. A general framework for weighted gene co-expression network analysis. *Stat. Appl. Genet. Mol. Biol.* **4,** Article 17 (2005).
19. Greenawalt, D. M. *et al.* A survey of the genetics of stomach, liver, and adipose gene expression from a morbidly obese cohort. *Genome Res.* **21,** 1008–1016 (2011).
20. Emilsson, V. *et al.* Genetics of gene expression and its effect on disease. *Nature* **452,** 423–428 (2008).
21. Schadt, E. E. *et al.* Mapping the genetic architecture of gene expression in human liver. *PLoS Biol.* **6,** e107 (2008).
22. Chen, Y. *et al.* Variations in DNA elucidate molecular networks that cause disease. *Nature* **452,** 429–435 (2008).
23. Zhong, H. *et al.* Liver and adipose expression associated SNPs are enriched for association to type 2 diabetes. *PLoS Genet.* **6,** e1000932 (2010).
24. Zhu, J. *et al.* Increasing the power to detect causal associations by combining genotypic and expression data in segregating populations. *PLOS Comput. Biol.* **3,** e69 (2007).
25. Lewis, S. J., Baker, I. & Davey Smith, G. Meta-analysis of vitamin D receptor polymorphisms and pulmonary tuberculosis risk. *Int. J. Tuberc. Lung Dis.* **9,** 1174–1177 (2005).
26. Li, X. *et al. SLC11A1* (NRAMP1) polymorphisms and tuberculosis susceptibility: updated systematic review and meta-analysis. *PLoS ONE* **6,** e15831 (2011).
27. Kumar, D. *et al.* Genome-wide analysis of the host intracellular network that regulates survival of Mycobacterium tuberculosis. *Cell* **140,** 731–743 (2010).
28. Glocker, E. O. *et al.* Infant colitis–it's in the genes. *Lancet* **376,** 1272 (2010).
29. Franke, A. *et al.* Sequence variants in *IL10*, *ARPC2* and multiple other loci contribute to ulcerative colitis susceptibility. *Nature Genet.* **40,** 1319–1323 (2008).
30. Bhattacharjee, A., Pal, S., Feldman, G. M. & Cathcart, M. K. Hck is a key regulator of gene expression in alternatively activated human monocytes. *J. Biol. Chem.* **286,** 36709–36723 (2011).

Supplementary Information is available in the online version of the paper.

Acknowledgements We thank all the subjects who contributed samples and the physicians and nursing staff who helped with recruitment globally. UK case collections were supported by the National Association for Colitis and Crohn's disease; Wellcome Trust grant 098051 (L.J., C.A.A., J.C.B.); Medical Research Council UK; the Catherine McEwan Foundation; an NHS Research Scotland career fellowship (R.K.R.); Peninsula College of Medicine and Dentistry, Exeter; the National Institute for Health Research, through the Comprehensive Local Research Network, and through Biomedical Research Centre awards to Guy's & Saint Thomas' National Health Service Trust, King's College London, Addenbrooke's Hospital, University of Cambridge School of Clinical Medicine and to the University of Manchester and Central Manchester Foundation Trust. The British 1958 Birth Cohort DNA collection was funded by Medical Research Council grant G0000934 and Wellcome Trust grant 068545/Z/02, and the UK National Blood Service controls by the Wellcome Trust. The Wellcome Trust Case Control Consortium projects were supported by Wellcome Trust grants 083948/Z/07/Z, 085475/B/08/Z and 085475/Z/08/Z. North American collections and data processing were supported by funds to the National Institute of Diabetes, Digestive and Kidney diseases (NIDDK) IBD Genetics Consortium, which is funded by the following grants: DK062431 (S.R.B.), DK062422 (J.H.C.), DK062420 (R.H.D.), DK062432 (J.D.R.), DK062423 (M.S.S.), DK062413 (D.P.M.), DK076984 (M.J.D.), DK084554 (M.J.D. and D.P.M.) and DK062429 (J.H.C.). Additional funds were provided by funding to J.H.C. (DK062429-S1 and Crohn's & Colitis Foundation of America, Senior Investigator Award (5-2229)) and R.H.D. (CA141743). K.Y.H. is supported by the National Institutes of Health (NIH) MSTP TG T32GM07205 training award. Cedars-Sinai is supported by USPHS grant PO1DK046763 and the Cedars-Sinai F. Widjaja Inflammatory Bowel and Immunobiology Research Institute Research Funds, National Center for Research Resources (NCRR) grant M01-RR00425, UCLA/Cedars-Sinai/Harbor/Drew Clinical and Translational Science Institute (CTSI) Grant (UL1 TR000124-01), the Southern California Diabetes and Endocrinology Research Grant (DERC) (DK063491), The Helmsley Foundation (D.P.M.) and the Crohn's and Colitis Foundation of America (D.P.M.). R.J.X. and A.N.A. are funded by DK83756, AI062773, DK043351 and the Helmsley Foundation. The Netherlands Organization for Scientific Research supported R.K.W. with a clinical fellowship grant (90.700.281) and C.W. (VICI grant 918.66.620). C.W. is also supported by the Celiac Disease Consortium (BSIK03009). This study was also supported by the German Ministry of Education and Research through the National Genome Research Network, the Popgen biobank, through the Deutsche Forschungsgemeinschaft (DFG) cluster of excellence 'Inflammation at Interfaces' and DFG grant no. FR 2821/2-1. S.B. was supported by DFG BR 1912/6-1 and the Else Kröner-Fresenius-Stiftung (Else Kröner-Exzellenzstipendium 2010_EKES.32). Italian case collections were supported by the Italian Group for IBD and the Italian Society for Paediatric Gastroenterology, Hepatology and Nutrition and funded by the Italian Ministry of Health GR-2008-1144485. Activities in Sweden were supported by the Swedish Society of Medicine, Ihre Foundation, Örebro University Hospital Research Foundation, Karolinska Institutet, the Swedish National Program for IBD Genetics, the Swedish Organization for IBD, and the Swedish Medical Research Council. D.F. and S.V. are senior clinical investigators for the Funds for Scientific Research (FWO/FNRS) Belgium. We acknowledge a grant from Viborg Regional Hospital, Denmark. V. Andersen was supported by SHS Aabenraa, Denmark. We acknowledge funding provided by the Royal Brisbane and Women's Hospital Foundation, National Health and Medical Research Council, Australia and by the European Community (5th PCRDT). We acknowledge the following groups that provided biological samples or data for this study: the Inflammatory Bowel in South Eastern Norway (IBSEN) study group, the Norwegian Bone Marrow Donor Registry (NMBDR), the Avon Longitudinal Study of Parents and Children, the Human Biological Data Interchange and Diabetes UK, and Banco Nacional de ADN, Salamanca. This research also uses resources provided by the Type 1 Diabetes Genetics Consortium, a collaborative clinical study sponsored by the NIDDK, National Institute of Allergy and Infectious Diseases (NIAID), National Human Genome Research Institute (NHGRI), National Institute of Child Health and Human Development (NICHD), and Juvenile Diabetes Research Foundation (JDRF) and supported by U01 DK062418. The KORA study was initiated and financed by the Helmholtz Zentrum München – German Research Center for Environmental Health, which is funded by the German Federal Ministry of Education and Research (BMBF) and by the State of Bavaria. KORA research was supported within the Munich Center of

©2012 **Macmillan Publishers Limited. All rights reserved**

Health Sciences (MC Health), Ludwig-Maximilians-Universität, as part of LMUinnovativ.

Author Contributions R.K.W., R.H.D., D.P.M., C.G.M., J.D.R., E.E.S., M.J.D., A.F., M.P. and S.V. contributed equally to the manuscript. J.H.C., J.C.B., R.K.W., R.H.D., D.P.M., A.F., M.P., C.G.M., J.D.R., S.V., M.D.A. and V. Annese conceived, designed and managed the study and managed the funding. J.H.C., J.C.B., L.J., S. Ripke, R.K.W., R.H.D., D.P.M., M.J.D., M.P. and C.G.M. were involved in manuscript preparation. J.H.C., J.C.B., L.J., S. Ripke, R.K.W., K.Y.H., C.A.A., J.E., K.N., S.L.S., S. Raychaudhuri, Z.W., C.A., A.C., G.B., M.H., X.H., B.Z., C.K.Z., H.Z., J.D.R., E.E.S. and M.J.D. performed or supervised statistical and computational analyses. R.K.W., R.H.D., D.P.M., J.C.L., L.P.S., Y.S., P.G., J.-P.A., T.A., L.A., A.N.A., V. Andersen, J.M.A., L.B., P.A.B., A.B., S.B., C.B., S.C., M.D.A., D.D.J., K.L.D., M.D., C.E., L.R.F., D.F., M.G., C.G., R.G., J.G., A.H., C.H., T.H.K., L.K., S.K., A.L., D.L., E.L., I.C.L., C.W.L., A.R.M., C.M., G.M., J.M., W.N., O.P., C.Y.P., U.P., N.J.P., M.R., J.I.R., R.K.R., J.D.S., M.S., J. Satsangi, S.S., L.A.S., J. Sventoraityte, S.R.T., M.T., H.W.V., M.D.V., C.W., D.C.W., J.W., R.J.X., S.Z., M.S.S., V. Annese, H.H., IIBDGC, S.R.B., J.D.R., G.R.S., C.G.M., A.F., M.P., S.V. and J.H.C. were involved in study subject recruitment and assembling phenotypic data. R.K.W., R.H.D., D.P.M., L.P.S., Y.S., M.M., I.C., E.T., T.B., D.E., K.F., T.H., K.D.T., C.G.M., A.F., M.P. and J.H.C. established DNA collections, genotyping and data management. All authors read and approved the final manuscript before submission.

Author Information Data have been deposited in the NCBI database of Genotypes and Phenotypes under accession numbers phs000130.v1.p1 and phs000345.v1.p1. Summary statistics for imputed GWAS are available at http://www.broadinstitute.org/mpg/ricopili/. Summary statistics for the meta-analysis markers are available at http://www.ibdgenetics.org/. The 523 causal gene network cytoscape file is available on request. Reprints and permissions information is available at www.nature.com/reprints. The authors declare no competing financial interests. Readers are welcome to comment on the online version of the paper. Correspondence and requests for materials should be addressed to J.H.C. (judy.cho@yale.edu).

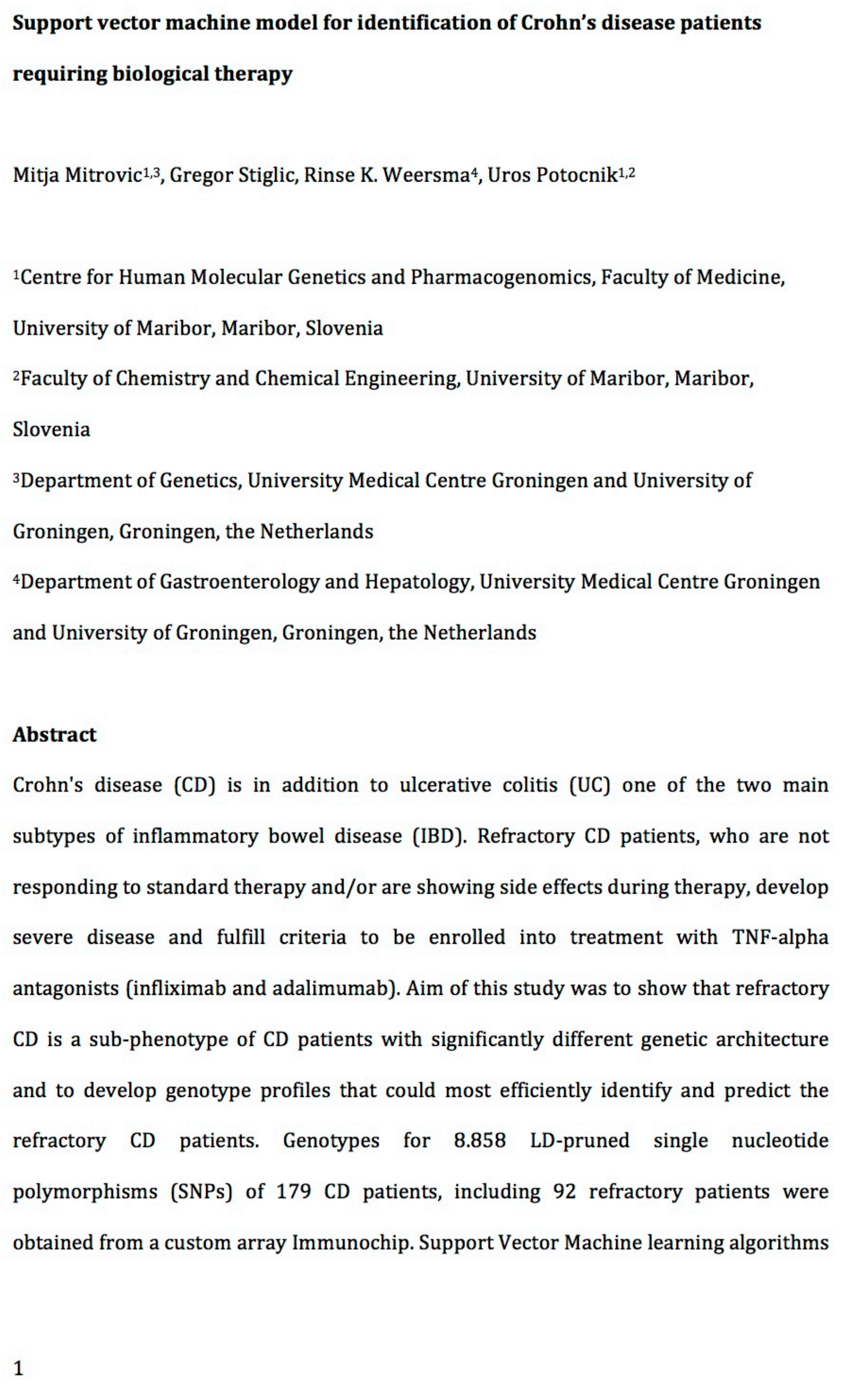
Luke Jostins[1]*, Stephan Ripke[2,3]*, Rinse K. Weersma[4], Richard H. Duerr[5,6], Dermot P. McGovern[7,8], Ken Y. Hui[9], James C. Lee[10], L. Philip Schumm[11], Yashoda Sharma[12], Carl A. Anderson[1], Jonah Essers[13], Mitja Mitrovic[14,15], Kaida Ning[12], Isabelle Cleynen[16], Emilie Theatre[17,18], Sarah L. Spain[19], Soumya Raychaudhuri[20,21,22], Philippe Goyette[23], Zhi Wei[24], Clara Abraham[12], Jean-Paul Achkar[25,26], Tariq Ahmad[27], Leila Amininejad[28], Ashwin N. Ananthakrishnan[29], Vibeke Andersen[30], Jane M. Andrews[31], Leonard Baidoo[5], Tobias Balschun[32], Peter A. Bampton[33], Alain Bitton[34], Gabrielle Boucher[23], Stephan Brand[35], Carsten Büning[36], Ariella Cohain[37], Sven Cichon[38], Mauro D'Amato[39], Dirk De Jong[4], Kathy L. Devaney[29], Marla Dubinsky[40], Cathryn Edwards[41], David Ellinghaus[32], Lynnette R. Ferguson[42], Denis Franchimont[28], Karin Fransen[5,43], Richard Gearry[44,45], Michel Georges[17], Christian Gieger[46], Jürgen Glas[34], Talin Haritunians[8], Ailsa Hart[47], Chris Hawkey[48], Matija Hedl[12], Xinli Hu[20], Tom H. Karlsen[49], Limas Kupcinskas[50], Subra Kugathasan[51], Anna Latiano[52], Debby Laukens[53], Ian C. Lawrance[54], Charlie W. Lees[55], Edouard Louis[18], Gillian Mahy[56], John Mansfield[57], Angharad R. Morgan[42], Craig Mowat[58], William Newman[59], Orazio Palmieri[52], Cyriel Y. Ponsioen[60], Uros Potocnik[14,61], Natalie J. Prescott[19], Miguel Regueiro[5], Jerome I. Rotter[8], Richard K. Russell[62], Jeremy D. Sanderson[63], Miquel Sans[64,65], Jack Satsangi[55], Stefan Schreiber[32,66], Lisa A. Simms[67], Jurgita Sventoraityte[50], Stephan R. Targan[7], Kent D. Taylor[7,8], Mark Tremelling[68], Hein W. Verspaget[69], Martine De Vos[53], Cisca Wijmenga[43], David C. Wilson[62,70], Juliane Winkelmann[71], Ramnik J. Xavier[29,72], Sebastian Zeissig[66], Bin Zhang[37], Clarence K. Zhang[73], Hongyu Zhao[73], The International IBD Genetics Consortium (IIBDGC)†, Mark S. Silverberg[74], Vito Annese[52,75], Hakon Hakonarson[76,77], Steven R. Brant[78], Graham Radford-Smith[67,79], Christopher G. Mathew[19], John D. Rioux[23], Eric E. Schadt[37], Mark J. Daly[2,3], Andre Franke[32], Miles Parkes[10], Severine Vermeire[16,80], Jeffrey C. Barrett[1]* & Judy H Cho[9,12]*

[1]Wellcome Trust Sanger Institute, Wellcome Trust Genome Campus, Hinxton, Cambridge CB10 1HH, UK. [2]Analytic and Translational Genetics Unit, Massachusetts General Hospital, Harvard Medical School, Boston, Massachusetts 02114, USA. [3]Broad Institute of MIT and Harvard, Cambridge, Massachusetts 02142, USA. [4]Department of Gastroenterology and Hepatology, University of Groningen and University Medical Center Groningen, Groningen 9700 RB, The Netherlands. [5]Division of Gastroenterology, Hepatology and Nutrition, Department of Medicine, University of Pittsburgh School of Medicine, Pittsburgh, Pennsylvania 15261, USA. [6]Department of Human Genetics, University of Pittsburgh Graduate School of Public Health, Pittsburgh, Pennsylvania 15261, USA. [7]F. Widjaja Foundation Inflammatory Bowel and Immunobiology Research Institute, Los Angeles, California 90048, USA. [8]Medical Genetics Institute, Cedars-Sinai Medical Center, Los Angeles, California 90048, USA. [9]Department of Genetics, Yale School of Medicine, New Haven, Connecticut 06520, USA. [10]Inflammatory Bowel Disease Research Group, Addenbrooke's Hospital, University of Cambridge, Cambridge CB2 0QQ, UK. [11]Department of Health Studies, University of Chicago, Chicago, Illinois 60637, USA. [12]Department of Internal Medicine, Section of Digestive Diseases, Yale School of Medicine, New Haven, Connecticut 06520, USA. [13]Center for Human Genetic Research, Massachusetts General Hospital, Harvard Medical School, Boston, Massachusetts 02114, USA. [14]University of Maribor, Faculty of Medicine, Center for Human Molecular Genetics and Pharmacogenomics, Maribor 2000, Slovenia. [15]University Medical Center Groningen, Department of Genetics, Groningen 9700 RB, The Netherlands. [16]Department of Clinical and Experimental Medicine, Gastroenterology section, KU Leuven, Leuven 3000, Belgium. [17]Unit of Animal Genomics, Groupe Interdisciplinaire de Génoprotéomique Appliquée (GIGA-R) and Faculty of Veterinary Medicine, University of Liège, Liège 4000, Belgium. [18]Division of Gastroenterology, Centre Hospitalier Universitaire, Université de Liège, Liège 4000, Belgium. [19]Department of Medical and Molecular Genetics, Division of Genetics and Molecular Medicine, King's College London School of Medicine, Guy's Hospital, London SE1 9RT, UK. [20]Division of Rheumatology Immunology and Allergy, Brigham and Women's Hospital, Boston, Massachusetts 02115, USA. [21]Program in Medical and Population Genetics, Broad Institute, Cambridge, Massachusetts 02142, USA. [22]Division of Genetics, Brigham and Women's Hospital, Boston, Massachusetts 02115, USA. [23]Université de Montréal and the Montreal Heart Institute, Research Center, Montréal, Québec H1T 1C8, Canada. [24]Department of Computer Science, New Jersey Institute of Technology, Newark, New Jersey 07102, USA. [25]Department of Gastroenterology & Hepatology, Digestive Disease Institute, Cleveland Clinic, Cleveland, Ohio 44195, USA. [26]Department of Pathobiology, Lerner Research Institute, Cleveland Clinic, Cleveland, Ohio 44195, USA. [27]Peninsula College of Medicine and Dentistry, Exeter EX1 2LU, UK. [28]Erasmus Hospital, Free University of Brussels, Department of Gastroenterology, Brussels, 1070 Belgium. [29]Massachusetts General Hospital, Harvard Medical School, Gastroenterology Unit, Boston, Massachusetts 02114, USA. [30]Viborg Regional Hospital, Medical Department, Viborg 8800, Denmark. [31]Inflammatory Bowel Disease Service, Department of Gastroenterology and Hepatology, Royal Adelaide Hospital, and School of Medicine, University of Adelaide, Adelaide 5000, Australia. [32]Institute of Clinical Chemistry, Christian-Albrechts-University, Kiel 24105, Germany. [33]Department of Gastroenterology and Hepatology, Flinders Medical Centre and School of Medicine, Flinders University, Adelaide 5000, Australia. [34]Division of Gastroenterology, McGill University Health Centre, Royal Victoria Hospital, Montréal, Québec H3A 1A1, Canada. [35]Department of Medicine II, University Hospital Munich-Grosshadern, Ludwig-Maximilians-University, Munich 80336, Germany. [36]Department of Gastroenterology, Charité, Campus Mitte, Universitätsmedizin Berlin, Berlin 10117, Germany. [37]Department of Genetics and Genomic Sciences, Mount Sinai School of Medicine, New York City, New York 10029, USA. [38]Department of Genomics, Life & Brain Center, University Hospital Bonn, Bonn 53012, Germany. [39]Department of Biosciences and Nutrition, Karolinska Institutet, Stockholm 14 183, Sweden. [40]Department of Pediatrics, Cedars Sinai Medical Center, Los Angeles, California 90048, USA. [41]Torbay Hospital, Department of Gastroenterology, Torbay, Devon TQ2 7AA, UK. [42]School of Medical Sciences, Faculty of Medical & Health Sciences, The University of Auckland, Auckland 1142, New Zealand. [43]University of Groningen, University Medical Center Groningen, Department of Genetics, Groningen T9700 RB, The Netherlands. [44]Department of Medicine, University of Otago, Christchurch 8140, New Zealand. [45]Department of Gastroenterology, Christchurch Hospital, Christchurch 8011, New Zealand. [46]Institute of Genetic Epidemiology, Helmholtz Zentrum München - German Research Center for Environmental Health, Neuherberg 85764, Germany. [47]St Mark's Hospital, Watford Road, Harrow, Middlesex HA1 3UJ, UK. [48]Nottingham Digestive Diseases Centre, Queens Medical Centre, Nottingham NG7 1AW, UK. [49]Research Institute of Internal Medicine, Oslo University Hospital Rikshospitalet, Oslo 0424, Norway. [50]Kaunas University of Medicine, Department of Gastroenterology, Kaunas 44307, Lithuania. [51]Department of Pediatrics, Emory University School of Medicine, Atlanta, Georgia 30322, USA. [52]Unit of Gastroenterology, Istituto di Ricovero e Cura a Carattere Scientifico-Casa Sollievo della Sofferenza (IRCCS-CSS) Hospital, San Giovanni Rotondo 71013, Italy. [53]Ghent University Hospital, Department of Gastroenterology and Hepatology, Ghent 9000, Belgium. [54]School of Medicine and Pharmacology, The University of Western Australia, Fremantle, Western Australia 6009, Australia. [55]Gastrointestinal Unit, Molecular Medicine Centre, University of Edinburgh, Western General Hospital, Edinburgh EH4 2XU, UK. [56]Department of Gastroenterology, The Townsville Hospital, Townsville, Queensland 4810, Australia. [57]Institute of Human Genetics, Newcastle University, Newcastle upon Tyne NE1 7RU, UK. [58]Department of Medicine, Ninewells Hospital and Medical School, Dundee DD1 9SY, UK. [59]Genetic Medicine, MAHSC, University of Manchester, Manchester M13 9PL, UK. [60]Academic Medical Center, Department of Gastroenterology, Amsterdam 1105 AZ, The Netherlands. [61]University of Maribor, Faculty for Chemistry and Chemical Engineering, Maribor 2000, Slovenia. [62]Royal Hospital for Sick Children, Paediatric Gastroenterology and Nutrition, Glasgow G3 8SJ, UK. [63]Guy's & St Thomas' NHS Foundation Trust, St Thomas' Hospital, Department of Gastroenterology, London SE1 7EH, UK. [64]Department of Gastroenterology, Hospital Clinic/Institut d'Investigacions Biomédiques August Pi i Sunyer (IDIBAPS), Barcelona, Spain. [65]Centro de Investigación Biomédica en Red de Enfermedades Hepáticas y Digestivas (CIBER-EHD), Barcelona 08036, Spain. [66]Department for General Internal Medicine, Christian-Albrechts-University, Kiel, Kiel 24118, Germany. [67]Inflammatory Bowel Diseases, Genetics and Computational Biology, Queensland Institute of Medical Research, Brisbane 4029, Australia. [68]Norfolk and Norwich University Hospital, Norwich NR4 7UY, UK. [69]Department of Gastroenterology, Leiden University Medical Center, Leiden 2333 ZA, The Netherlands. [70]Child Life and Health, University of Edinburgh, Edinburgh, Scotland EH9 1UW, UK. [71]Institute of Human Genetics and Department of Neurology, Technische Universität München, Munich 80336, Germany. [72]Center for Computational and Integrative Biology, Massachusetts General Hospital, Boston, Massachusetts 02114, USA. [73]Department of Biostatistics, School of Public Health, Yale University, New Haven, Connecticut 06520, USA. [74]Mount Sinai Hospital Inflammatory Bowel Disease Centre, University of Toronto, Toronto, Ontario M5G 1X5, Canada. [75]Azienda Ospedaliero Universitaria (AOU) Careggi, Unit of Gastroenterology SOD2, Florence 50134, Italy. [76]Center for Applied Genomics, The Children's Hospital of Philadelphia, Philadelphia, Pennsylvania 19104, USA. [77]Department of Pediatrics, Center for Pediatric Inflammatory Bowel Disease, The Children's Hospital of Philadelphia, Philadelphia, Pennsylvania 19104, USA. [78]Meyerhoff Inflammatory Bowel Disease Center, Department of Medicine, School of Medicine, and Department of Epidemiology, Bloomberg School of Public Health, Johns Hopkins University, Baltimore, Maryland 21205, USA. [79]Department of Gastroenterology, Royal Brisbane and Women's Hospital, and School of Medicine, University of Queensland, Brisbane 4029, Australia. [80]Department of Gastroenterology, University Hospital Leuven, Leuven 3000, Belgium.

*These authors contributed equally to this work.
†Lists of participants and their affiliations appear in the Supplementary Information.

©2012 Macmillan Publishers Limited. All rights reserved

6.6.6 Izvirni znanstveni članek 5



Support vector machine model for identification of Crohn's disease patients requiring biological therapy

Mitja Mitrovic[1,3], Gregor Stiglic, Rinse K. Weersma[4], Uros Potocnik[1,2]

[1]Centre for Human Molecular Genetics and Pharmacogenomics, Faculty of Medicine, University of Maribor, Maribor, Slovenia

[2]Faculty of Chemistry and Chemical Engineering, University of Maribor, Maribor, Slovenia

[3]Department of Genetics, University Medical Centre Groningen and University of Groningen, Groningen, the Netherlands

[4]Department of Gastroenterology and Hepatology, University Medical Centre Groningen and University of Groningen, Groningen, the Netherlands

Abstract

Crohn's disease (CD) is in addition to ulcerative colitis (UC) one of the two main subtypes of inflammatory bowel disease (IBD). Refractory CD patients, who are not responding to standard therapy and/or are showing side effects during therapy, develop severe disease and fulfill criteria to be enrolled into treatment with TNF-alpha antagonists (infliximab and adalimumab). Aim of this study was to show that refractory CD is a sub-phenotype of CD patients with significantly different genetic architecture and to develop genotype profiles that could most efficiently identify and predict the refractory CD patients. Genotypes for 8.858 LD-pruned single nucleotide polymorphisms (SNPs) of 179 CD patients, including 92 refractory patients were obtained from a custom array Immunochip. Support Vector Machine learning algorithms were used for genotype profile modeling. Measuring AUC assessed the efficiency of the genetic models.

We have found that most efficient genotype profile distinguished refractory patients from other CD patients with the best-achieved AUC 0.64 and consisted of 59 discriminator SNPs. Inclusion of demographic data (smoking, age, sex) did not significantly improve the predictive model. In addition, discriminator SNPs were tested in a standard association test between CD and CD refractory patients and two SNPs (rs9592040 and rs346818) remained significantly associated after Bonferroni correction. Comparison to previously associated CD loci revealed an overlap with gene regions *PHACTR2* and *KIR3DL3* that were previously implicated in pathogenesis of other immune-mediated diseases. Protein-protein interaction (DAPPLE) and connectivity analysis (GRAIL) were conducted for gene products coded in loci flanking the 59 discriminator SNPs. Although DAPPLE results showed no evidence that the PPI between tested genes was greater than chance, GRAIL analysis revealed, a total of 5 genes rs395561 (*EFNA5*), rs2690262 (*UNCX*), rs9720889 (*SCXB*), rs10992979 (*BARX1*), rs7127817 (*NCAM1*), with significant association to functional connectivity.

Functional enrichment and gene expression analysis of these 5 genes showed that only *NCAM1* was largely and specifically expressed in NK cells as compared to other tested tissues. Interestingly, recent evidence shows that this gene is abnormally expressed in inactive colonic mucosa of patients with CD, and was implicated in treatment of multiple sclerosis. Thus, this gene seems a compelling candidate for future follow-up functional studies. Identification of patients that will not respond to standard therapy is important as these patients could be enrolled into biological therapy at early stages.

Key words: Support vector machines, refractory Crohn's disease, TNF-alpha antagonists

Introduction

Crohn's disease (CD) is one of the two major subtypes of inflammatory bowel disease (IBD) and is characterized as a chronic heterogeneous disorder with differences in severity, location, behavior (i.e. inflammatory, fistulizing, stenosing) and age at onset of inflammation[1]. Refractory CD patients, who are not responding to standard therapy, develop severe disease and fulfill criteria to be enrolled into treatment with TNF-alpha antagonists (infliximab and adalimumab)[2-4]. The mechanism of lack of efficacy to the standard treatment is unclear with its potential origin in specific genetic background, eventually reflected in abnormal immunological, biochemical, and clinical parameters.

Genetic background in CD has been extensively investigated[5]. Recent meta-analysis of 15 GWASs and an independent genotype dataset (obtained using Immunochip[6]), identified 71 new associations, for a total of 163 IBD loci, yet as much as 86 % of the total disease variance for CD is still considered to be unexplained[7]. One possible explanation for these weak relative and attributable risks is, that risk may vary across clinically and biologically distinct subgroups of CD patients. Moreover, Cleynen *et al.* identified 6 genetic-based subgroups of CD patients using 43 established SNPs, but were underpowered to allocate them to clinically derived sub-phenotypes[8]. Recently, polymorphisms in *SLC22A5*[9] and *ABCB1/MDR1*[10] were significantly associated with refractory CD patients but not with other CD patients. Therefore, we sought to scrutinize potential differences in genetic background between CD patients responding to the standard therapy and refractory CD patients. To this end, we used state-of-the-art classification and regression algorithm Support Vector Machines (SVM) on a large number of SNPs from immune-related regions. Additionally, the optimal number of SNPs needed for the best classification performance was obtained using Support Vector Machines Recursive Feature Elimination (SVM-RFE) technique.

Materials And Methods

Subjects. All patients were diagnosed according to standard clinical criteria by endoscopy, radiology and histopathology[11]. A total of 179 CD patients were included in our study between years 2000 and 2010 from the Department of Gastroenterology in University Medical Center Ljubljana and University Clinical Center Maribor as described previously[9,12]. In brief, the patients' mean age was 40 years (range 19-60) and mean disease duration was 10 years (range 1-22). Out of 179 CD patients, 92 CD refractory patients, including patients with the active luminal or fistulizing form of CD for at least 6 months, were included in the study (Table S1). Inclusion criteria for active luminal disease comprised: moderately to severely active ileal and/or colonic form of CD (40 patients) not adequately responding to prior conventional treatment with 5-aminosalicylates (5-ASA) 4.5 g/day given for at least 2 months, antibiotics (metronidazole 1.2 g/day and/or ciprofloxacin 1g/day) given for at least 1 month, 6-methyl-prednisolone up to 42 mg/day given for at least 2 months and azathioprine (AZA) 2.5 mg/kg/day given for at least 6 months. The activity of the disease was evaluated by Crohn's disease activity index (CDAI)[13,14]. CDAI scores were calculated from patient diaries within a week before the start of the study. The mean CDAI score for the group was 280 (range 180-320). The following inclusion criteria were used for fistulizing disease: single or multiple draining fistulas occurring as a complication of CD (52 patients), with no prior response to standard treatment with metronidazole 0.8g/day and/or ciprofloxacin 1g/day given for at least 2 months, and AZA 2.5mg/kg/day given for at least 6 months. Exclusion criteria were: other complications of CD (e.g. symptomatic strictures or stenosis, abscesses, previous total colectomy), a history of allergy to murine proteins, a serious infection in the previous 3 months, and prior treatment with any anti TNF-α medication. The protocol allowed concomitant treatment with 5-ASA, antibiotics, corticosteroids or AZA, but the dosage had to be stable for at least 2 months before enrolment. All 19 patients who tolerated AZA, regardless of their previous response, continued to receive the drug at a stable dose of 2.5 mg/kg/day.

Ethical considerations. Experiments were undertaken with the understanding and written consent of each individual. This study conforms with The Code of Ethics of the World Medical Association[15] and was approved by The Ethical Committee of the Republic of Slovenia (approval no. 57/03/20 of March 21, 2000).

DNA extraction, genotyping and genotype calling. Genomic DNA was extracted from whole blood lymphocytes, according to manufacturer's protocol, using a combination of Ficoll- Paque PLUS (GE Healthcare Bio-Sciences, Sweden) and TRI REAGENT (Sigma-Aldrich, USA) reagents. DNA samples were genotyped using the Immunochip as described previously[7]. In short, Immunochip is a custom Illumina Infinium high-density array, consisting of 196,524 variants, compiled largely from variants identified in previous GWAS of 12 different immune-mediated diseases[16]. Genotype calling was conducted using Illumina's Genome Studio Data Analysis software and custom generated cluster file of Trynka *et al.*[16].

Data quality control. Data quality control was conducted as proposed by Anderson *et al.*[17]. In brief, from a set of 196,524 SNPs and 179 individuals, we first excluded individuals and variants with call rate <99 % and substantial deviation from the mean heterozygosity value, as computed in PLINK[18]. To assess potential ethnic outliers in the dataset, multidimensional scaling was used with HapMap 3 individuals as a reference[19].

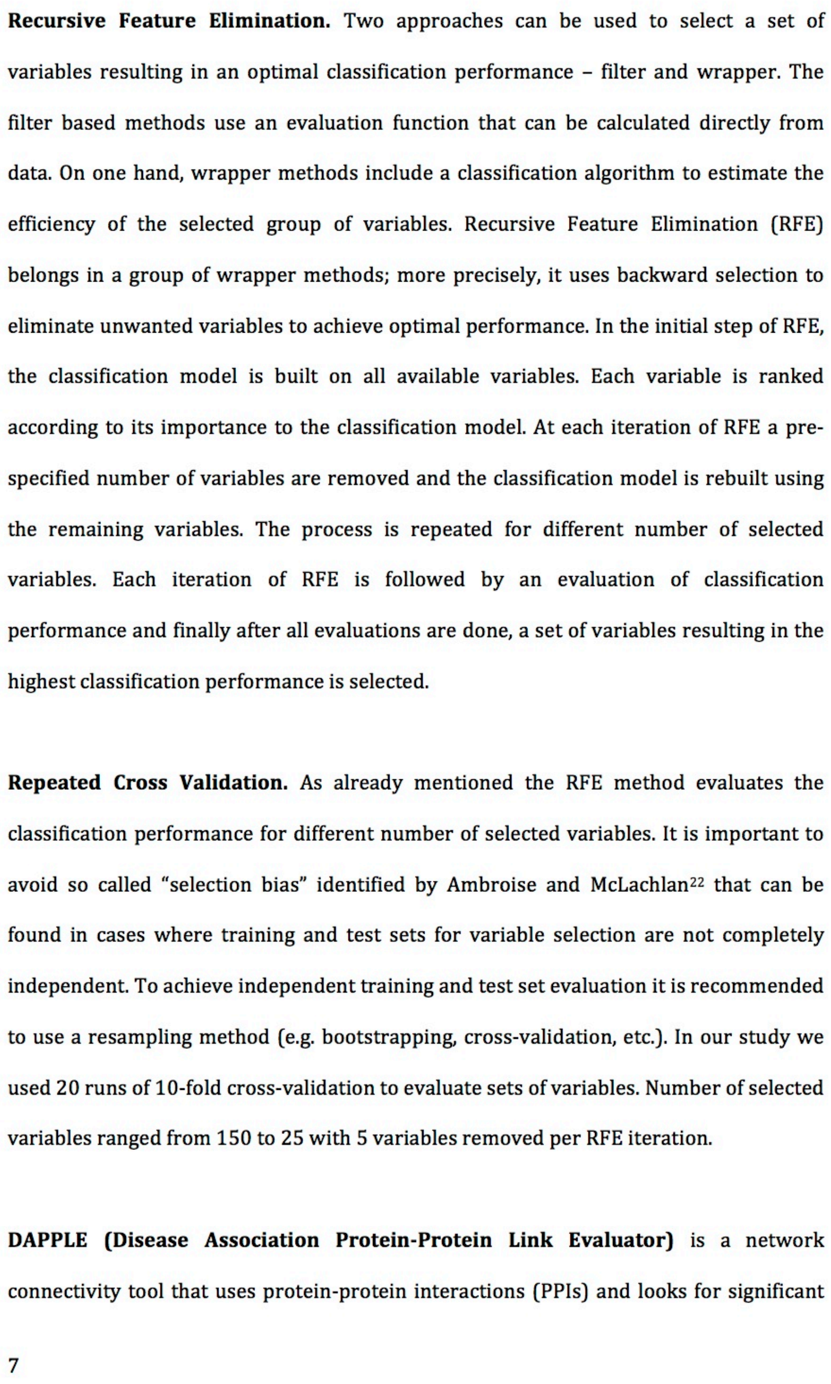
5

Identity-by-descent and identity-by-state analyses were used to identify possible duplicated and highly related individuals (first- or second-degree relatives). In addition we excluded variants with deviation from Hardy-Weinberg equilibrium in controls (P < 0.0001) and/or differential missingness in no-call genotypes between cases and controls (P < 0.001). After the quality control procedure, our dataset consisted of 169,548 autosomal or X-chromosome genotypes for 179 individuals.

SNP subset selection. Given a relatively small sample size in our study we restricted our analysis to a representative subset of SNPs to achieve sufficient statistical power. Thus, we extracted a subset of 8,858 SNPs with minimal linkage disequilibrium (LD) among them from the initial 169,548 set of variants. This was achieved in PLINK[20] by calculating LD between the original set of SNPs with a window size of 50 SNPs, sliding 5 SNPs in each step and excluding one of a pair of SNPs if the LD was greater than r2 < 0.5.

Statistical Methods

Support Vector Machines. This study utilizes state-of-the-art classification and regression algorithm Support Vector Machines (SVM). It aims to learn a classifier from a training set consisting of positive and negative samples. The built model is later used to classify new unlabeled test samples. The SVM builds a classification model by mapping the input training samples into a high-dimensional variable space, where a hyper-plane separating two groups of samples from different classes is constructed. A hyper-plane is constructed in a way that maximizes the margin, i.e. the distance between the hyper-plane and the nearest points of the opposite class. Classification And Regression Training (CARET) package[21] version 5.15 was used in our experiments.



Recursive Feature Elimination. Two approaches can be used to select a set of variables resulting in an optimal classification performance – filter and wrapper. The filter based methods use an evaluation function that can be calculated directly from data. On one hand, wrapper methods include a classification algorithm to estimate the efficiency of the selected group of variables. Recursive Feature Elimination (RFE) belongs in a group of wrapper methods; more precisely, it uses backward selection to eliminate unwanted variables to achieve optimal performance. In the initial step of RFE, the classification model is built on all available variables. Each variable is ranked according to its importance to the classification model. At each iteration of RFE a pre-specified number of variables are removed and the classification model is rebuilt using the remaining variables. The process is repeated for different number of selected variables. Each iteration of RFE is followed by an evaluation of classification performance and finally after all evaluations are done, a set of variables resulting in the highest classification performance is selected.

Repeated Cross Validation. As already mentioned the RFE method evaluates the classification performance for different number of selected variables. It is important to avoid so called "selection bias" identified by Ambroise and McLachlan[22] that can be found in cases where training and test sets for variable selection are not completely independent. To achieve independent training and test set evaluation it is recommended to use a resampling method (e.g. bootstrapping, cross-validation, etc.). In our study we used 20 runs of 10-fold cross-validation to evaluate sets of variables. Number of selected variables ranged from 150 to 25 with 5 variables removed per RFE iteration.

DAPPLE (Disease Association Protein-Protein Link Evaluator) is a network connectivity tool that uses protein-protein interactions (PPIs) and looks for significant physical connectivity among proteins encoded for by genes in loci associated to disease[23]. The hypothesis behind DAPPLE is that causal genetic variation affects a limited set of underlying mechanisms that are detectable by protein-protein interactions. Hence each gene is measured in either direct or indirect interactions with genes in other loci, and an empirical p-value is calculated by permutation.

GRAIL (Gene Relationships Across Implicated Loci) is a bioinformatic annotation tool that, given several genomic regions or SNPs associated with a particular phenotype or disease, searches for similarities in the published scientific text among the associated genes[24]. It scores regions for functional relatedness by defining associated regions based on the interval between recombination hotspots flanking furthest neighboring SNPs with r2 >0.5 to the index SNP, and identifies overlapping genes in that region. Based on textual relationships between genes (as determined from a download of PubMed abstracts on 16 December 2006), GRAIL assigns a P value to each region suggesting its degree of functional connectivity, and picks the best candidate gene after taking into account multiple comparisons.

Functional enrichment and gene expression analysis. To assess the statistical enrichment of functional gene sets from molecular function, biological process categories and pathways for candidate genes, we used a prediction query program PANTHER[13,25]. Expression data were gathered from BioGPS, an online gene annotation database that reports individual gene expression levels for a number of human tissues and cell types[26].

Results

To select an optimal number of SNPs needed for top classification performance between refractory CD patients and CD patients responding to the standard therapy, we measured average AUC over 20 randomized 10-fold cross-validations using SVM for classification and RFE for the subset of 8,858 LD-pruned SNPs. Figure 1 shows an average AUC for different number of selected SNPs along with sensitivity and specificity of the predictive model (linear SVM). Number of tested SNPs was increased in steps of 5 due to the computational complexity of repeated cross-validation process. The top performance in all three measurements was achieved using 59 SNPs. The same procedure was used to test higher number of chosen SNPs (up to 1,000 SNPs in steps of 25), however the performance of the predictive models with higher number of SNPs did not match the overall best classifier using 59 SNPs.

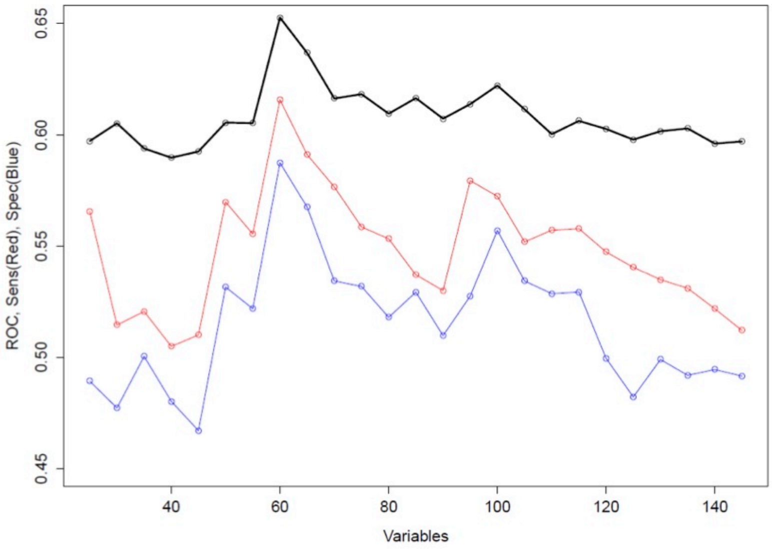


Figure 1. Comparison of classification performance metrics (AUC, sensitivity, specificity) for different numbers of SNPs (variables).

9

Figure 2 shows results of 20 cross-validation runs in terms of optimal feature sets. Optimal set of SNPs is selected for each run, where Figure 2 presents the number of SNPs in those optimal sets. It can be observed that in 10 out of 20 runs, the optimal set consists of 59 classifier SNPs.

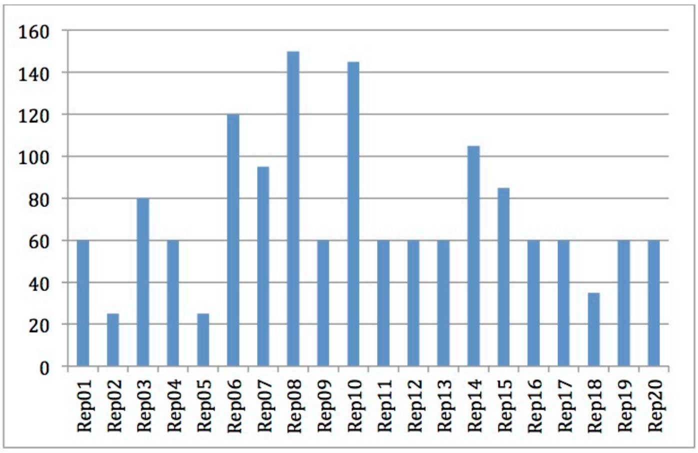


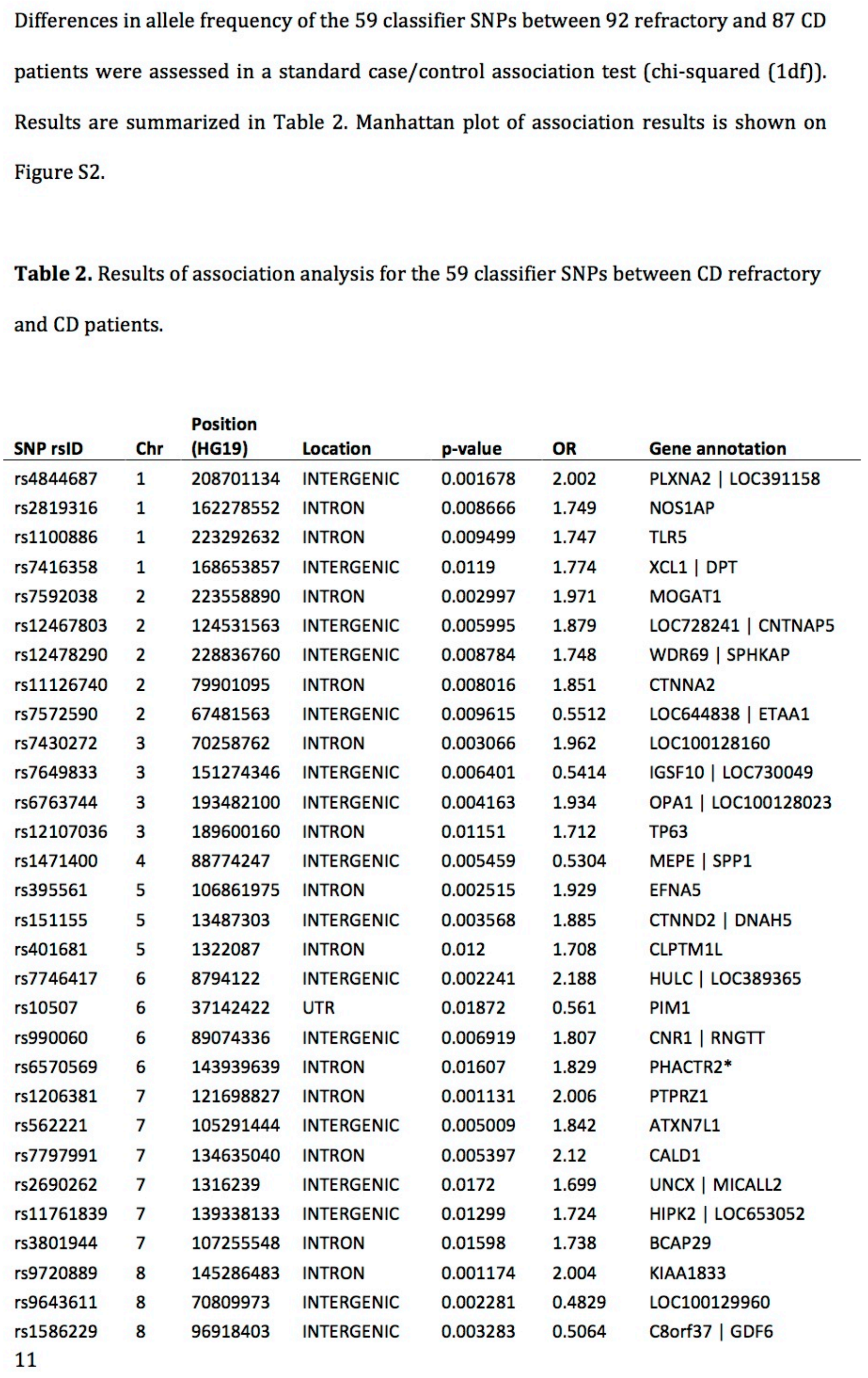
Figure 2. Bar plot of selected optimal SNP set sizes for 20 cross-validation runs

Stability of SNP selection was also assessed with frequency tables (Table 1). In our case two SNPs were selected in more than half of the optimal sets (rs346818 and rs9592040).

Table 1. Most frequently selected SNPs in 200 iterations of repeated cross-validation process.

SNP rsID	No. of selections in 200 iterations
rs346818	169
rs9592040	121
rs1470943	70
rs9720889	62
rs4844687	53

10

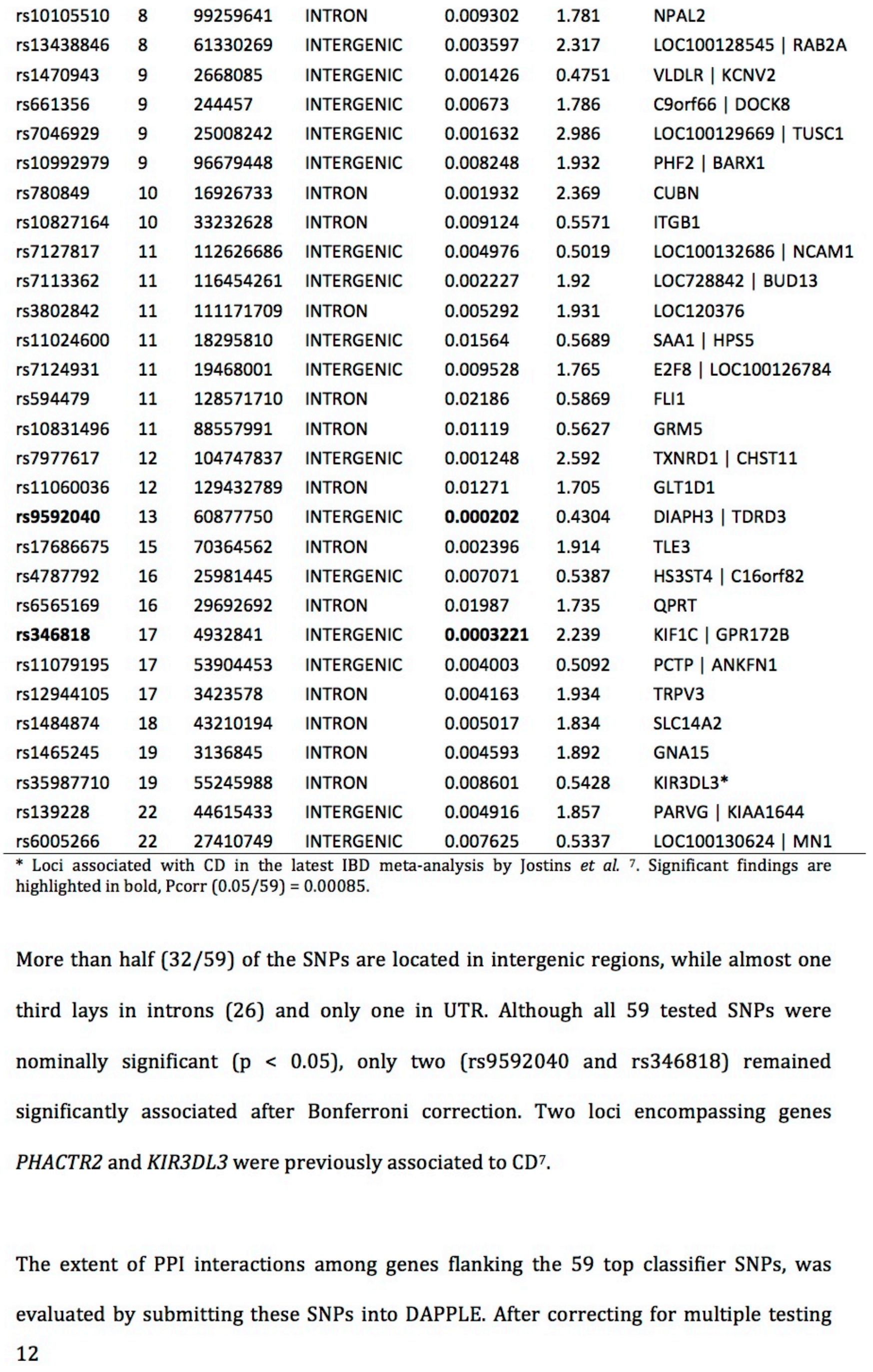


Differences in allele frequency of the 59 classifier SNPs between 92 refractory and 87 CD patients were assessed in a standard case/control association test (chi-squared (1df)). Results are summarized in Table 2. Manhattan plot of association results is shown on Figure S2.

Table 2. Results of association analysis for the 59 classifier SNPs between CD refractory and CD patients.

SNP rsID	Chr	Position (HG19)	Location	p-value	OR	Gene annotation
rs4844687	1	208701134	INTERGENIC	0.001678	2.002	PLXNA2 | LOC391158
rs2819316	1	162278552	INTRON	0.008666	1.749	NOS1AP
rs1100886	1	223292632	INTRON	0.009499	1.747	TLR5
rs7416358	1	168653857	INTERGENIC	0.0119	1.774	XCL1 | DPT
rs7592038	2	223558890	INTRON	0.002997	1.971	MOGAT1
rs12467803	2	124531563	INTERGENIC	0.005995	1.879	LOC728241 | CNTNAP5
rs12478290	2	228836760	INTERGENIC	0.008784	1.748	WDR69 | SPHKAP
rs11126740	2	79901095	INTRON	0.008016	1.851	CTNNA2
rs7572590	2	67481563	INTERGENIC	0.009615	0.5512	LOC644838 | ETAA1
rs7430272	3	70258762	INTRON	0.003066	1.962	LOC100128160
rs7649833	3	151274346	INTERGENIC	0.006401	0.5414	IGSF10 | LOC730049
rs6763744	3	193482100	INTERGENIC	0.004163	1.934	OPA1 | LOC100128023
rs12107036	3	189600160	INTRON	0.01151	1.712	TP63
rs1471400	4	88774247	INTERGENIC	0.005459	0.5304	MEPE | SPP1
rs395561	5	106861975	INTRON	0.002515	1.929	EFNA5
rs151155	5	13487303	INTERGENIC	0.003568	1.885	CTNND2 | DNAH5
rs401681	5	1322087	INTRON	0.012	1.708	CLPTM1L
rs7746417	6	8794122	INTERGENIC	0.002241	2.188	HULC | LOC389365
rs10507	6	37142422	UTR	0.01872	0.561	PIM1
rs990060	6	89074336	INTERGENIC	0.006919	1.807	CNR1 | RNGTT
rs6570569	6	143939639	INTRON	0.01607	1.829	PHACTR2*
rs1206381	7	121698827	INTRON	0.001131	2.006	PTPRZ1
rs562221	7	105291444	INTERGENIC	0.005009	1.842	ATXN7L1
rs7797991	7	134635040	INTRON	0.005397	2.12	CALD1
rs2690262	7	1316239	INTERGENIC	0.0172	1.699	UNCX | MICALL2
rs11761839	7	139338133	INTERGENIC	0.01299	1.724	HIPK2 | LOC653052
rs3801944	7	107255548	INTRON	0.01598	1.738	BCAP29
rs9720889	8	145286483	INTRON	0.001174	2.004	KIAA1833
rs9643611	8	70809973	INTERGENIC	0.002281	0.4829	LOC100129960
rs1586229	8	96918403	INTERGENIC	0.003283	0.5064	C8orf37 | GDF6

11



rs10105510	8	99259641	INTRON	0.009302	1.781	NPAL2
rs13438846	8	61330269	INTERGENIC	0.003597	2.317	LOC100128545 | RAB2A
rs1470943	9	2668085	INTERGENIC	0.001426	0.4751	VLDLR | KCNV2
rs661356	9	244457	INTERGENIC	0.00673	1.786	C9orf66 | DOCK8
rs7046929	9	25008242	INTERGENIC	0.001632	2.986	LOC100129669 | TUSC1
rs10992979	9	96679448	INTERGENIC	0.008248	1.932	PHF2 | BARX1
rs780849	10	16926733	INTRON	0.001932	2.369	CUBN
rs10827164	10	33232628	INTRON	0.009124	0.5571	ITGB1
rs7127817	11	112626686	INTERGENIC	0.004976	0.5019	LOC100132686 | NCAM1
rs7113362	11	116454261	INTERGENIC	0.002227	1.92	LOC728842 | BUD13
rs3802842	11	111171709	INTRON	0.005292	1.931	LOC120376
rs11024600	11	18295810	INTERGENIC	0.01564	0.5689	SAA1 | HPS5
rs7124931	11	19468001	INTERGENIC	0.009528	1.765	E2F8 | LOC100126784
rs594479	11	128571710	INTRON	0.02186	0.5869	FLI1
rs10831496	11	88557991	INTRON	0.01119	0.5627	GRM5
rs7977617	12	104747837	INTERGENIC	0.001248	2.592	TXNRD1 | CHST11
rs11060036	12	129432789	INTRON	0.01271	1.705	GLT1D1
rs9592040	13	60877750	INTERGENIC	**0.000202**	0.4304	DIAPH3 | TDRD3
rs17686675	15	70364562	INTRON	0.002396	1.914	TLE3
rs4787792	16	25981445	INTERGENIC	0.007071	0.5387	HS3ST4 | C16orf82
rs6565169	16	29692692	INTRON	0.01987	1.735	QPRT
rs346818	17	4932841	INTERGENIC	**0.0003221**	2.239	KIF1C | GPR172B
rs11079195	17	53904453	INTERGENIC	0.004003	0.5092	PCTP | ANKFN1
rs12944105	17	3423578	INTRON	0.004163	1.934	TRPV3
rs1484874	18	43210194	INTRON	0.005017	1.834	SLC14A2
rs1465245	19	3136845	INTRON	0.004593	1.892	GNA15
rs35987710	19	55245988	INTRON	0.008601	0.5428	KIR3DL3*
rs139228	22	44615433	INTERGENIC	0.004916	1.857	PARVG | KIAA1644
rs6005266	22	27410749	INTERGENIC	0.007625	0.5337	LOC100130624 | MN1

* Loci associated with CD in the latest IBD meta-analysis by Jostins *et al.* [7]. Significant findings are highlighted in bold, Pcorr (0.05/59) = 0.00085.

More than half (32/59) of the SNPs are located in intergenic regions, while almost one third lays in introns (26) and only one in UTR. Although all 59 tested SNPs were nominally significant ($p < 0.05$), only two (rs9592040 and rs346818) remained significantly associated after Bonferroni correction. Two loci encompassing genes *PHACTR2* and *KIR3DL3* were previously associated to CD[7].

The extent of PPI interactions among genes flanking the 59 top classifier SNPs, was evaluated by submitting these SNPs into DAPPLE. After correcting for multiple testing

12

we found no evidence that the PPI between these genes was greater than by chance alone.

Next, we used GRAIL to perform functional relatedness analysis among the regions flanking 59 SNPs. We could not find appropriate proxies for 7/59 SNPs, hence 52 SNPs were submitted into GRAIL search among PubMed articles till August 2012. We selected all genes with $p < 0.05$ as GRAIL implicated loci. A total of 5 genes (Table 3, Figure 3): rs395561 (*EFNA5*), rs2690262 (*UNCX*), rs9720889 (*SCXB*), rs10992979 (*BARX1*), rs7127817 (*NCAM1*), out of 120 (Table S3) had a significant association with functional connectivity.

Table 3. Significant candidate genes from GRAIL.

SNP rsID	GRAIL p-value	CANDIDATE GENE(S)
rs395561	0.011664265	*EFNA5*
rs2690262	0.027882239	*UNCX*
rs9720889	0.015961279	*SCXB*
rs10992979	0.046930008	*BARX1*
rs7127817	0.008606012	*NCAM1*

In addition, the following keywords describing functional connections were produced: 'ncam', 'integrin', 'sulfate', 'adhesion', 'heparan', 'matrix', 'ephrin', 'neural', 'extracellular', 'receptor', 'signaling', 'patterning', 'cannabinoid', 'notch', 'mice', 'receptors', 'risk', 'development', 'bone', 'somite'.

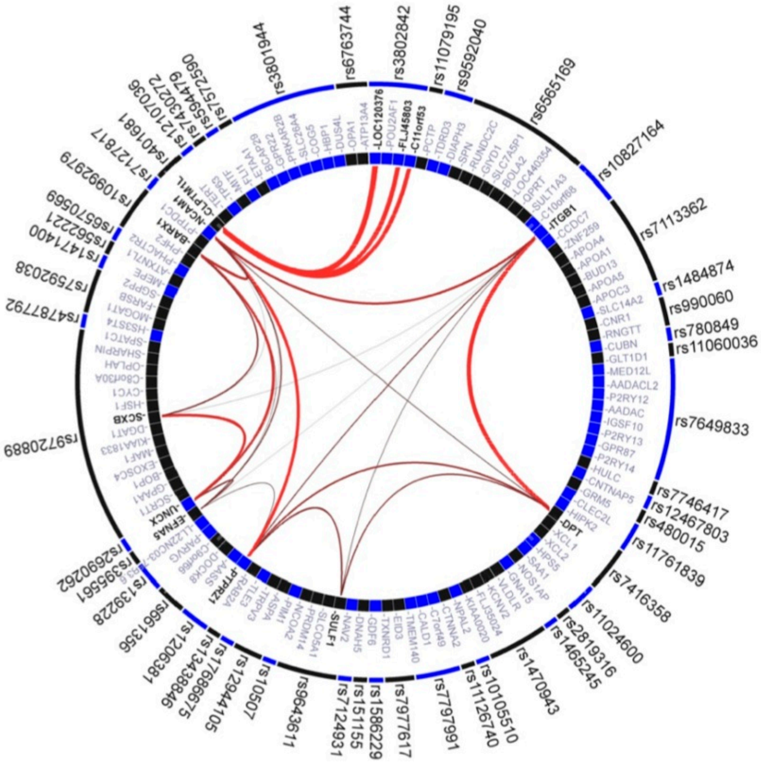


Figure 3. Gene relationships across implicated loci (GRAIL) pathway analysis.

Links between genes flanking 52 classifier SNPs which scored $P < 0.01$ using GRAIL. In total, there were 120 genes implicated by proximity to these 52 SNPs. Each observed association was scored with GRAIL, which takes each gene mapping within associated intervals and evaluates for each whether it is non-randomly linked to the other genes through word usage in PubMed abstracts. The 52 SNPs shown in the outer circle are significant at $P < 0.01$, indicating that the regions which they tag contain genes which are more significantly linked to genes in the other regions than expected by chance at that level. The lines between genes represent individually significant connections that contribute to the positive signal, with the thickness of the lines being inversely

14

proportional to the probability that a literature-based connection would be seen by chance.

In order to functionally annotate our GRAIL results, we estimated the enrichment of functional gene sets from molecular function, biological process categories and pathways for the 5 candidate genes (Table S3, Figure 4).

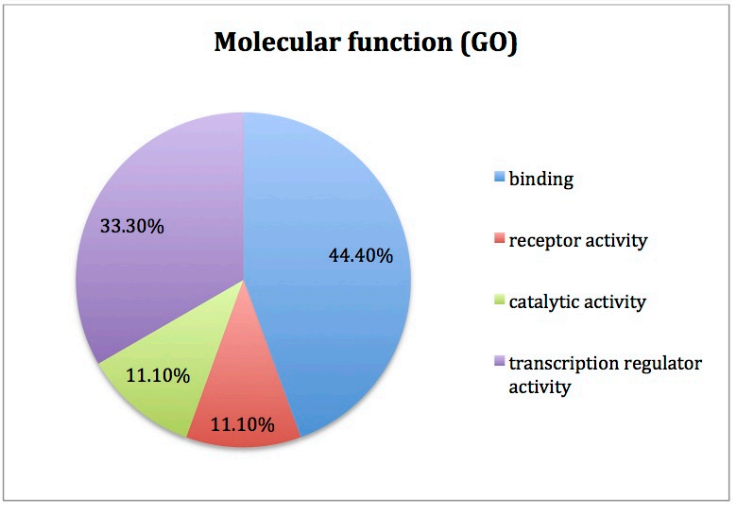


Figure 4. Molecular function for 5 candidate genes obtained from PANTHER analysis.

Analysis of GO results showed that 5 candidate genes are transcription factors involved in development of various tissues including immune system (Table S4).

Finally, we used the gene portal BioGPS, which contains gene expression data on a variety of human tissues and cell types[26]. For our analysis we considered 9 immune cell types and 8 non-immune tissues. We submitted the list of significant candidate genes (n = 5) obtained from GRAIL analysis and for each gene we obtained a different genetic expression value in every tissue or cell type tested, with exception of *SCXB*, because the expression data were not available. Because of different background characteristics between each probe set, a direct comparison of expression across different genes was not possible. Therefore, we decided to standardize the expression values of each single gene across different tissues. Figure S1 shows the standardized expression values in the 17 tissues and cell types tested. We did not observe significant differences across tested genes and tissues, except for *NCAM1,* where expression appeared particularly high in natural killer (NK) cells (CD56+_NKCells).

Discussion

In this study we hypothesized that refractory CD patients are a subgroup of CD patients on a molecular level, with its origin in a specific genetic background, eventually reflected in response to the standard therapy. In particular, we've measured average AUC over 20 randomized 10-fold cross-validations using SVM for classification and RFE for the subset of 8,858 LD-pruned SNPs. The field of machine learning provides a variety of methods to approach these challenges, such as linear or logistic regression techniques, decision trees, random forests and Bayesian approaches, which can provide an improvement over standard regression approaches by including a statistical analysis in the context of Bayesian inference. Recently, the support vector machine (SVM) approach was proposed to perform genome-wide disease risk predictions based on GWAS data[34,35] and was shown to outperform logistic regression on type 2 diabetes (T2D) dataset[36]. An important issue in predictive modeling for SNP studies is also the feature selection problem. A study by Yingjie et al.[37] demonstrates an approach to select the most appropriate SNPs to achieve the best classification performance. The applied approach identified 59 SNPs that distinguished refractory from other CD patients. This was also

16

confirmed by a follow-up association analysis where all identified SNPs had significantly different allele frequencies between the groups of patients and two SNPs remained significant after Bonferroni correction. SNP rs9592040 is in the intergenic region 140kb downstream from *DIAPH3*, which encodes a protein involved in actin remodeling and regulates cell movement and adhesion and 93kb upstream from *TDRD3*, which encodes a scaffolding protein that specifically recognizes and binds dimethylarginine-containing protein and acts as a coactivator in the nucleus. The second associated SNP, rs346818, lays 1kb downstream of *KIF1C*, which encodes a motor protein, required for the retrograde transport of Golgi vesicles to the endoplasmic reticulum and 3kb upstream of *GPR172B*, which encodes riboflavin transporter, highly expressed in the small intestine. PHACTR2 is involved in phosphate and actin regulation and has been implicated in Parkinson's disease and multiple sclerosis[38,39]. *KIR3DL3* encodes a killer cell immunoglobulin-like receptor, previously implicated in ankylosing spondylitis and multiple sclerosis[40,41]. Interestingly, one study showed that, patients with hepatitis C virus (HCV) who carried the certain *KIR* genotypes were less prone to respond to treatment with pegylated alpha interferon and ribavirin[42].

When we tested the PPI interactions among genes flanking the 59 top classifier SNPs, we were unable to reach significance in number of direct and/or indirect interactions. This could be due to the fact that they are indeed more general genes for CD, and might be less useful to differentiate refractory from other CD patients. Alternatively, this could be explained by the possible lack of true association among SNPs from our list and was confirmed by cross comparison with results from latest IBD meta-analysis [7]. However, 5 significant instances were obtained from GRAIL analysis. To further validate the GRAIL results, we looked at tissue specific expression of the candidate genes. Using the BioGPS database we were able to show that candidate genes were expressed in immune cells.

Gene ontology analysis also confirmed the immune-related functions of these genes. Of note is the association with *NCAM1,* largely and specifically expressed in NK cells as compared to other tissues. This gene encodes a cell adhesion protein, which is a member of the immunoglobulin superfamily. The encoded protein is a well-known NK cell marker has been shown to be involved in the expansion of T cells and dendritic cells which play an important role in immune surveillance[27]. Study of Fasseu *et al.* showed evidence of this gene being abnormally expressed in inactive colonic mucosa of patients with CD[43]. Recent Swedish study showed that the protein levels of NCAM1 were significantly decreased in MS patients after therapy with natalizumab[44]. Thus, this gene seems a compelling candidate for future follow-up functional studies. This will improve our knowledge of this complex disease and hopefully provide future strategies of disease prevention and treatment.

In conclusion, we showed that classification and regression algorithm Support Vector Machines (SVM) can be used to as a model produce a prioritized list of SNPs distinguishing the response to therapy in CD patients. The specific SNP combination determining the refractory CD patients could be promising disease therapy predictors, and deserve future study in a larger cohort. Similar approaches, using GWAs data, could be interesting also for pharmacogenetics of other complex diseases.

Acknowledgements

Authors are thankful to Karin Fransen, Mathieu Platteel and Katja Repnik for excellent technical and scientific support, sample selection, DNA extraction and genotyping.

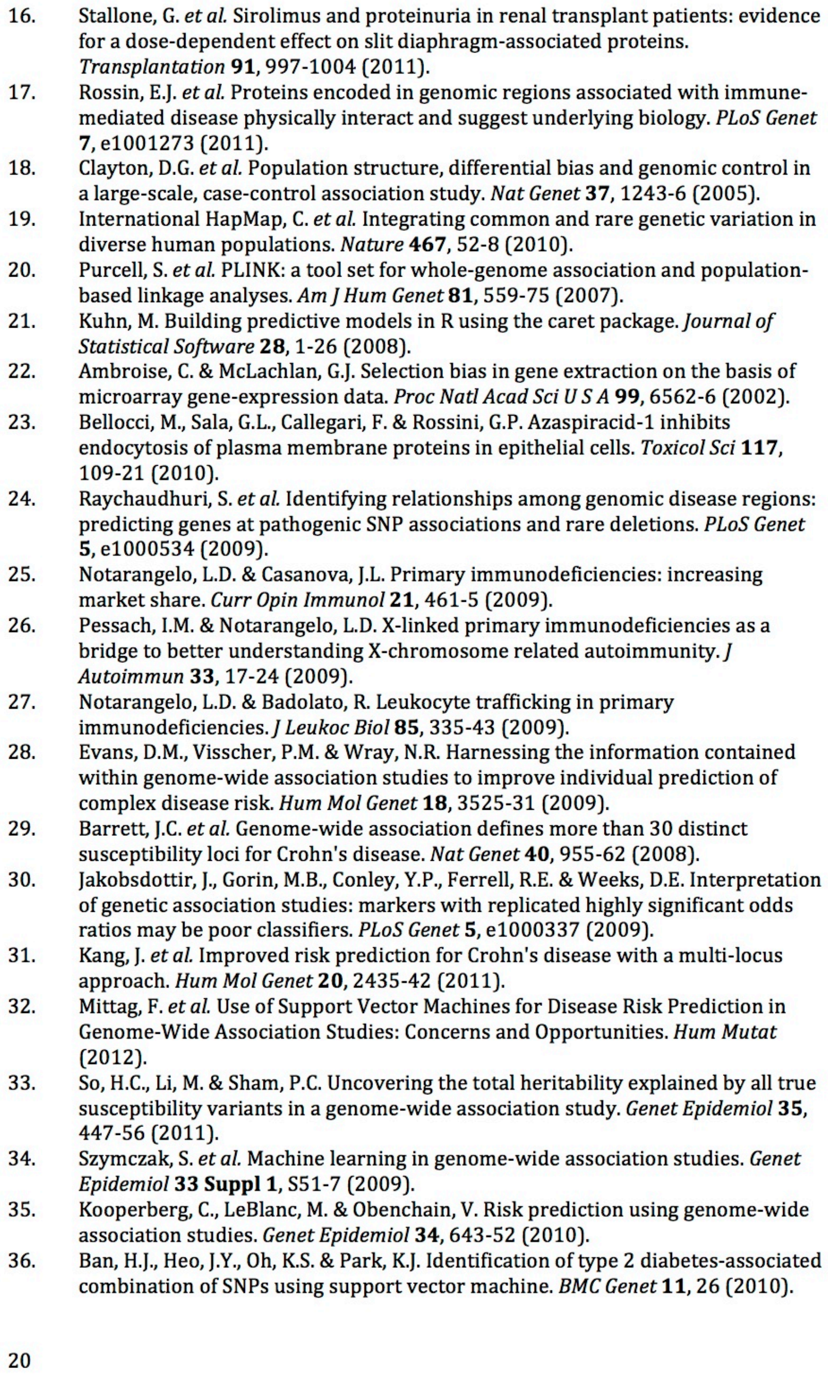
Author contributions

Conceived and designed the experiments: MM, GS, UP, RKW. Performed the experiments: MM, GS. Analyzed the data: MM, GS. Wrote the paper: MM, GS. Contributed reagents/materials/analysis tools: RKW, GS, UP.

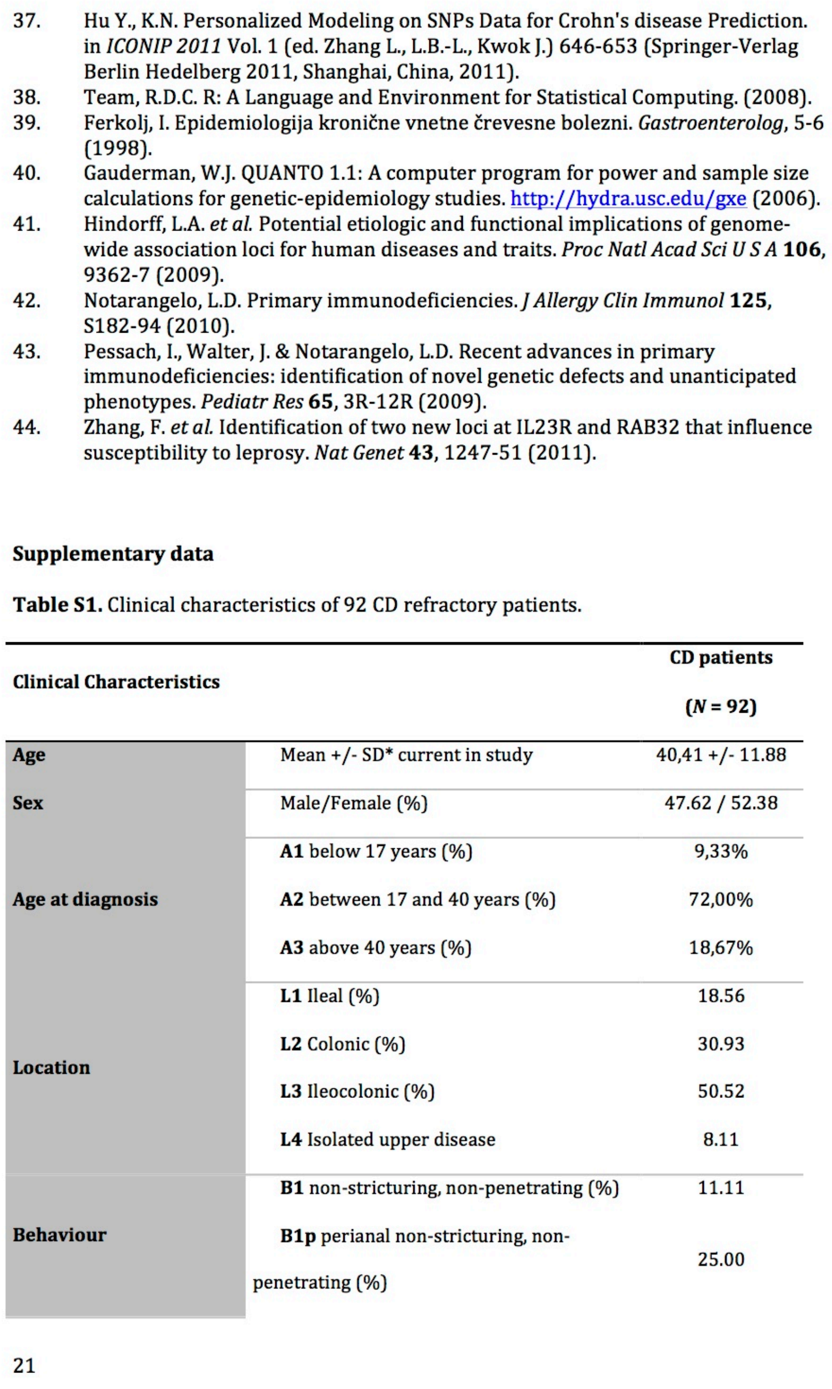
Competing interests: none declared.

References

1. Weersma, R.K. *et al.* Molecular prediction of disease risk and severity in a large Dutch Crohn's disease cohort. *Gut* **58**, 388-95 (2009).
2. Sandborn, W.J. *et al.* Ustekinumab induction and maintenance therapy in refractory Crohn's disease. *N Engl J Med* **367**, 1519-28 (2012).
3. Blackhouse, G. *et al.* Canadian cost-utility analysis of initiation and maintenance treatment with anti-TNF-alpha drugs for refractory Crohn's disease. *J Crohns Colitis* **6**, 77-85 (2012).
4. Roblin, X. *et al.* Prevalence of cytomegalovirus infection in steroid-refractory Crohn's disease. *Inflamm Bowel Dis* **18**, E1396-7 (2012).
5. Fransen, K., Mitrovic, M., van Diemen, C.C. & Weersma, R.K. The quest for genetic risk factors for Crohn's disease in the post-GWAS era. *Genome Med* **3**, 13 (2011).
6. Cortes, A. & Brown, M.A. Promise and pitfalls of the Immunochip. *Arthritis Res Ther* **13**, 101 (2011).
7. Stacchiotti, A. *et al.* Stress proteins in experimental nephrotoxicity: a ten year experience. *Ital J Anat Embryol* **115**, 153-8 (2010).
8. Cleynen, I. *et al.* Molecular reclassification of Crohn's disease by cluster analysis of genetic variants. *PLoS One* **5**, e12952 (2010).
9. Repnik, K. & Potocnik, U. Haplotype in the IBD5 region is associated with refractory Crohn's disease in Slovenian patients and modulates expression of the SLC22A5 gene. *J Gastroenterol* **46**, 1081-91 (2011).
10. Potocnik, U., Ferkolj, I., Glavac, D. & Dean, M. Polymorphisms in multidrug resistance 1 (MDR1) gene are associated with refractory Crohn disease and ulcerative colitis. *Genes and Immunity* **5**, 530-539 (2004).
11. Podolsky, D.K. Inflammatory bowel disease. *N Engl J Med* **347**, 417-29 (2002).
12. Mitrovic, M. & Potocnik, U. High-resolution melting curve analysis for high-throughput genotyping of NOD2/CARD15 mutations and distribution of these mutations in Slovenian inflammatory bowel diseases patients. *Dis Markers* **30**, 265-74 (2011).
13. International Union of Immunological Societies Expert Committee on Primary, I. *et al.* Primary immunodeficiencies: 2009 update. *J Allergy Clin Immunol* **124**, 1161-78 (2009).
14. Team, R.D.C. R: A Language and Environment for Statistical Computing. (R Foundation for Statistical Computing, Vienna, Austria, 2008).
15. Rickham, P.P. Human Experimentation. Code of Ethics of the World Medical Association. Declaration of Helsinki. *Br Med J* **2**, 177 (1964).



16. Stallone, G. *et al.* Sirolimus and proteinuria in renal transplant patients: evidence for a dose-dependent effect on slit diaphragm-associated proteins. *Transplantation* **91**, 997-1004 (2011).
17. Rossin, E.J. *et al.* Proteins encoded in genomic regions associated with immune-mediated disease physically interact and suggest underlying biology. *PLoS Genet* **7**, e1001273 (2011).
18. Clayton, D.G. *et al.* Population structure, differential bias and genomic control in a large-scale, case-control association study. *Nat Genet* **37**, 1243-6 (2005).
19. International HapMap, C. *et al.* Integrating common and rare genetic variation in diverse human populations. *Nature* **467**, 52-8 (2010).
20. Purcell, S. *et al.* PLINK: a tool set for whole-genome association and population-based linkage analyses. *Am J Hum Genet* **81**, 559-75 (2007).
21. Kuhn, M. Building predictive models in R using the caret package. *Journal of Statistical Software* **28**, 1-26 (2008).
22. Ambroise, C. & McLachlan, G.J. Selection bias in gene extraction on the basis of microarray gene-expression data. *Proc Natl Acad Sci U S A* **99**, 6562-6 (2002).
23. Bellocci, M., Sala, G.L., Callegari, F. & Rossini, G.P. Azaspiracid-1 inhibits endocytosis of plasma membrane proteins in epithelial cells. *Toxicol Sci* **117**, 109-21 (2010).
24. Raychaudhuri, S. *et al.* Identifying relationships among genomic disease regions: predicting genes at pathogenic SNP associations and rare deletions. *PLoS Genet* **5**, e1000534 (2009).
25. Notarangelo, L.D. & Casanova, J.L. Primary immunodeficiencies: increasing market share. *Curr Opin Immunol* **21**, 461-5 (2009).
26. Pessach, I.M. & Notarangelo, L.D. X-linked primary immunodeficiencies as a bridge to better understanding X-chromosome related autoimmunity. *J Autoimmun* **33**, 17-24 (2009).
27. Notarangelo, L.D. & Badolato, R. Leukocyte trafficking in primary immunodeficiencies. *J Leukoc Biol* **85**, 335-43 (2009).
28. Evans, D.M., Visscher, P.M. & Wray, N.R. Harnessing the information contained within genome-wide association studies to improve individual prediction of complex disease risk. *Hum Mol Genet* **18**, 3525-31 (2009).
29. Barrett, J.C. *et al.* Genome-wide association defines more than 30 distinct susceptibility loci for Crohn's disease. *Nat Genet* **40**, 955-62 (2008).
30. Jakobsdottir, J., Gorin, M.B., Conley, Y.P., Ferrell, R.E. & Weeks, D.E. Interpretation of genetic association studies: markers with replicated highly significant odds ratios may be poor classifiers. *PLoS Genet* **5**, e1000337 (2009).
31. Kang, J. *et al.* Improved risk prediction for Crohn's disease with a multi-locus approach. *Hum Mol Genet* **20**, 2435-42 (2011).
32. Mittag, F. *et al.* Use of Support Vector Machines for Disease Risk Prediction in Genome-Wide Association Studies: Concerns and Opportunities. *Hum Mutat* (2012).
33. So, H.C., Li, M. & Sham, P.C. Uncovering the total heritability explained by all true susceptibility variants in a genome-wide association study. *Genet Epidemiol* **35**, 447-56 (2011).
34. Szymczak, S. *et al.* Machine learning in genome-wide association studies. *Genet Epidemiol* **33 Suppl 1**, S51-7 (2009).
35. Kooperberg, C., LeBlanc, M. & Obenchain, V. Risk prediction using genome-wide association studies. *Genet Epidemiol* **34**, 643-52 (2010).
36. Ban, H.J., Heo, J.Y., Oh, K.S. & Park, K.J. Identification of type 2 diabetes-associated combination of SNPs using support vector machine. *BMC Genet* **11**, 26 (2010).



37. Hu Y., K.N. Personalized Modeling on SNPs Data for Crohn's disease Prediction. in *ICONIP 2011* Vol. 1 (ed. Zhang L., L.B.-L., Kwok J.) 646-653 (Springer-Verlag Berlin Hedelberg 2011, Shanghai, China, 2011).
38. Team, R.D.C. R: A Language and Environment for Statistical Computing. (2008).
39. Ferkolj, I. Epidemiologija kronične vnetne črevesne bolezni. *Gastroenterolog*, 5-6 (1998).
40. Gauderman, W.J. QUANTO 1.1: A computer program for power and sample size calculations for genetic-epidemiology studies. http://hydra.usc.edu/gxe (2006).
41. Hindorff, L.A. *et al.* Potential etiologic and functional implications of genome-wide association loci for human diseases and traits. *Proc Natl Acad Sci U S A* **106**, 9362-7 (2009).
42. Notarangelo, L.D. Primary immunodeficiencies. *J Allergy Clin Immunol* **125**, S182-94 (2010).
43. Pessach, I., Walter, J. & Notarangelo, L.D. Recent advances in primary immunodeficiencies: identification of novel genetic defects and unanticipated phenotypes. *Pediatr Res* **65**, 3R-12R (2009).
44. Zhang, F. *et al.* Identification of two new loci at IL23R and RAB32 that influence susceptibility to leprosy. *Nat Genet* **43**, 1247-51 (2011).

Supplementary data

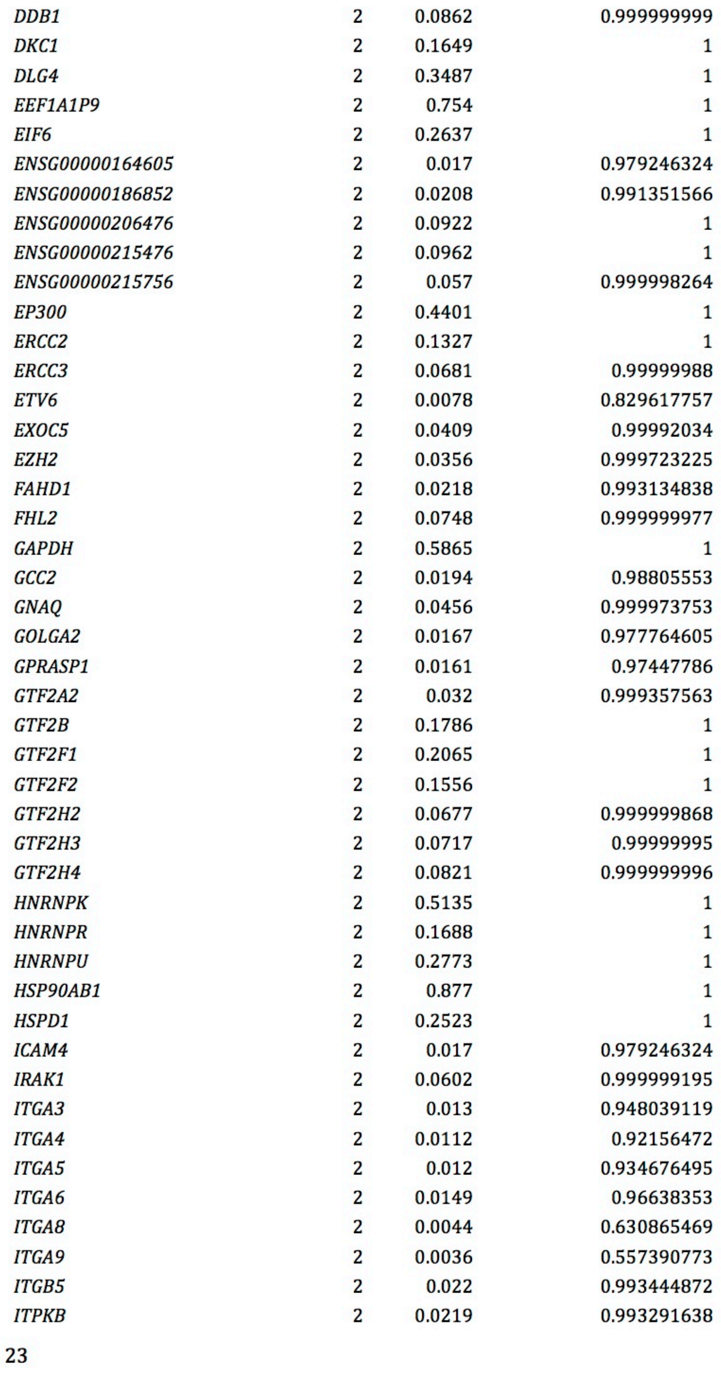
Table S1. Clinical characteristics of 92 CD refractory patients.

Clinical Characteristics		**CD patients** **(*N* = 92)**
Age	Mean +/- SD* current in study	40,41 +/- 11.88
Sex	Male/Female (%)	47.62 / 52.38
Age at diagnosis	**A1** below 17 years (%)	9,33%
	A2 between 17 and 40 years (%)	72,00%
	A3 above 40 years (%)	18,67%
Location	**L1** Ileal (%)	18.56
	L2 Colonic (%)	30.93
	L3 Ileocolonic (%)	50.52
	L4 Isolated upper disease	8.11
Behaviour	**B1** non-stricturing, non-penetrating (%)	11.11
	B1p perianal non-stricturing, non-penetrating (%)	25.00

	B2 stricturing (%)	16.67
	B2p perianal structuring (%)	22.22
	B3 penetrating (%)	13.89
	B3p perianal penetrating (%)	11.11
Fistula	Yes/No (%)	53.03 / 46.97
Smoker	Yes/No (%)	21.15 / 78.85
Previous surgery	Yes/No (%)	62.82 / 37.18

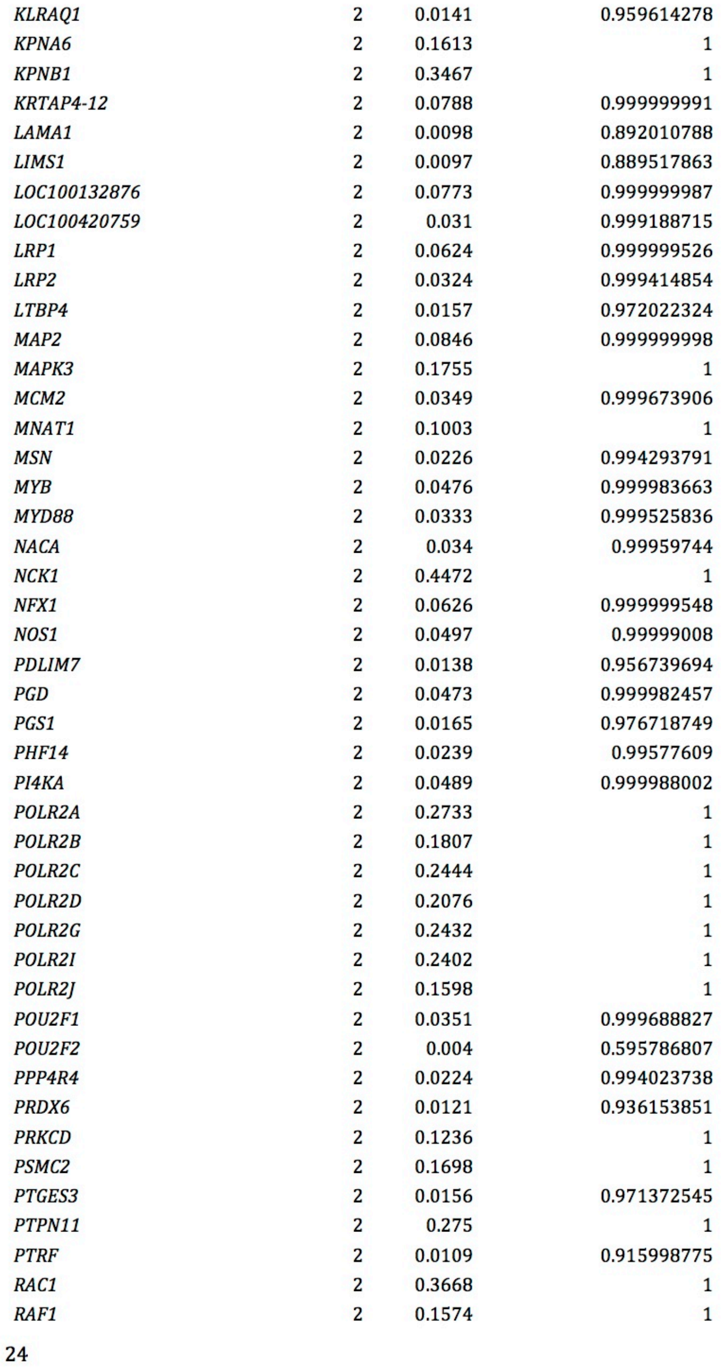
Table S2. List of all genes from DAPPLE analysis.

PROTEIN	NUM_BINDERS	P_VALUE	P_VALUE_CORRECTED
ACTN1	2	0.0671	0.999999848
ACTN2	2	0.1824	1
ALB	2	0.0817	0.999999996
APOA1	2	0.059	0.999998925
ATP13A3	2	0.0208	0.991351566
BAD	2	0.0464	0.999978284
BEST2	2	0.0231	0.994916945
BICC1	2	0.0167	0.977764605
C12orf30	2	0.0206	0.990943038
CAMK2A	2	0.1242	1
CCNE1	2	0.0583	0.999998729
CCNH	2	0.0925	1
CD209	2	0.0229	0.994676258
CD36	2	0.0062	0.754770742
CD3EAP	2	0.0033	0.526229989
CD44	2	0.0275	0.998167467
CDC27	2	0.0331	0.999503142
CDK7	2	0.1912	1
CDS1	2	0.0228	0.994551693
CHRNA4	2	0.0203	0.990293952
CLDN11	2	0.0176	0.981921268
CLU	2	0.0314	0.999260992
CNOT6L	2	0.0235	0.995366299
COX6A1	2	0.0279	0.998330167
CPSF4	2	0.0073	0.80907054
CREBBP	2	0.3731	1
CRK	2	0.3745	1
CSNK2A2	2	0.3047	1



DDB1	2	0.0862	0.999999999
DKC1	2	0.1649	1
DLG4	2	0.3487	1
EEF1A1P9	2	0.754	1
EIF6	2	0.2637	1
ENSG00000164605	2	0.017	0.979246324
ENSG00000186852	2	0.0208	0.991351566
ENSG00000206476	2	0.0922	1
ENSG00000215476	2	0.0962	1
ENSG00000215756	2	0.057	0.999998264
EP300	2	0.4401	1
ERCC2	2	0.1327	1
ERCC3	2	0.0681	0.99999988
ETV6	2	0.0078	0.829617757
EXOC5	2	0.0409	0.99992034
EZH2	2	0.0356	0.999723225
FAHD1	2	0.0218	0.993134838
FHL2	2	0.0748	0.999999977
GAPDH	2	0.5865	1
GCC2	2	0.0194	0.98805553
GNAQ	2	0.0456	0.999973753
GOLGA2	2	0.0167	0.977764605
GPRASP1	2	0.0161	0.97447786
GTF2A2	2	0.032	0.999357563
GTF2B	2	0.1786	1
GTF2F1	2	0.2065	1
GTF2F2	2	0.1556	1
GTF2H2	2	0.0677	0.999999868
GTF2H3	2	0.0717	0.99999995
GTF2H4	2	0.0821	0.999999996
HNRNPK	2	0.5135	1
HNRNPR	2	0.1688	1
HNRNPU	2	0.2773	1
HSP90AB1	2	0.877	1
HSPD1	2	0.2523	1
ICAM4	2	0.017	0.979246324
IRAK1	2	0.0602	0.999999195
ITGA3	2	0.013	0.948039119
ITGA4	2	0.0112	0.92156472
ITGA5	2	0.012	0.934676495
ITGA6	2	0.0149	0.96638353
ITGA8	2	0.0044	0.630865469
ITGA9	2	0.0036	0.557390773
ITGB5	2	0.022	0.993444872
ITPKB	2	0.0219	0.993291638

23



KLRAQ1	2	0.0141	0.959614278
KPNA6	2	0.1613	1
KPNB1	2	0.3467	1
KRTAP4-12	2	0.0788	0.999999991
LAMA1	2	0.0098	0.892010788
LIMS1	2	0.0097	0.889517863
LOC100132876	2	0.0773	0.999999987
LOC100420759	2	0.031	0.999188715
LRP1	2	0.0624	0.999999526
LRP2	2	0.0324	0.999414854
LTBP4	2	0.0157	0.972022324
MAP2	2	0.0846	0.999999998
MAPK3	2	0.1755	1
MCM2	2	0.0349	0.999673906
MNAT1	2	0.1003	1
MSN	2	0.0226	0.994293791
MYB	2	0.0476	0.999983663
MYD88	2	0.0333	0.999525836
NACA	2	0.034	0.99959744
NCK1	2	0.4472	1
NFX1	2	0.0626	0.999999548
NOS1	2	0.0497	0.99999008
PDLIM7	2	0.0138	0.956739694
PGD	2	0.0473	0.999982457
PGS1	2	0.0165	0.976718749
PHF14	2	0.0239	0.99577609
PI4KA	2	0.0489	0.999988002
POLR2A	2	0.2733	1
POLR2B	2	0.1807	1
POLR2C	2	0.2444	1
POLR2D	2	0.2076	1
POLR2G	2	0.2432	1
POLR2I	2	0.2402	1
POLR2J	2	0.1598	1
POU2F1	2	0.0351	0.999688827
POU2F2	2	0.004	0.595786807
PPP4R4	2	0.0224	0.994023738
PRDX6	2	0.0121	0.936153851
PRKCD	2	0.1236	1
PSMC2	2	0.1698	1
PTGES3	2	0.0156	0.971372545
PTPN11	2	0.275	1
PTRF	2	0.0109	0.915998775
RAC1	2	0.3668	1
RAF1	2	0.1574	1

24

RBL2	2	0.3142	1
REPS1	2	0.0118	0.931618913
RGS12	2	0.0094	0.881690085
RGS3	2	0.0091	0.873310325
RRN3	2	0.0199	0.989355924
RUVBL2	2	0.4575	1
SEPSECS	2	0.0162	0.975057446
SFRS8	2	0.0206	0.990943038
SGTA	2	0.015	0.967146012
SHC1	2	0.2364	1
SMAD2	2	0.1779	1
SMAD9	2	0.1234	1
SMARCC2	2	0.2398	1
SMOX	2	0.0206	0.990943038
SNAPIN	2	0.0171	0.979718049
SNRPG	2	0.1694	1
SORCS2	2	0.1345	1
SORT1	2	0.1126	1
SPTBN1	2	0.0873	0.999999999
SRF	2	0.0488	0.999987713
SSR4	2	0.0799	0.999999993
SUB1	2	0.0096	0.88696765
SUMO1P3	2	0.4515	1
SUPT16H	2	0.1918	1
SYN1	2	0.0665	0.999999824
TAF1	2	0.0821	0.999999996
TAF11	2	0.0395	0.999889234
TAF12	2	0.0551	0.999997263
TAF13	2	0.0468	0.999980249
TAF1B	2	0.0324	0.999414854
TAF1C	2	0.0168	0.978269857
TAF5	2	0.1579	1
TAF9	2	0.1283	1
TDRD3	2	0.0228	0.994551693
THBS1	2	0.0274	0.998124384
TJP1	2	0.1084	1
TLE1	2	0.0616	0.999999425
TP73	2	0.0404	0.999910382
TTF1	2	0.0159	0.973278158
TUBA4A	2	0.2749	1
TULP4	2	0.0252	0.996874585
UBTF	2	0.028	0.998368542
UMPS	2	0.0311	0.999207418
WDR82	2	0.0204	0.990515303
WSB1	2	0.0093	0.878960022

25

XPNPEP1	2	0.002	0.363933781
ABL1	3	0.1752	1
ACTN4	3	0.0158	0.972657418
ADRA1B	3	0.001	0.202372096
C14orf21	3	0.0026	0.444765764
CCT2	3	0.0151	0.967891275
COL4A1	3	0.0003	0.065562161
CSNK2A1P	3	0.4498	1
ENSG00000137379	3	0.2597	1
ENSG00000183311	3	0.2952	1
EWSR1	3	0.0793	0.999999992
FLNA	3	0.0685	0.999999892
FN1	3	0.0489	0.999988002
FYN	3	0.4352	1
GRM1	3	0.0041	0.604855887
HSP90AA2	3	0.1185	1
ITGAV	3	0.0027	0.457205919
ITGB3	3	0.0119	0.93316511
LGALS1	3	0.0004	0.086450836
MED31	3	0.0171	0.979718049
PIK3R1	3	0.267	1
PLCB1	3	0.0062	0.754770742
PLCG1	3	0.0974	1
POLR1A	3	0.0187	0.98596519
POLR1B	3	0.0133	0.95148914
POLR1C	3	0.015	0.967146012
POLR1D	3	0.0066	0.776097378
POLR2E	3	0.1085	1
POLR2F	3	0.0785	0.999999991
POLR2K	3	0.0576	0.999998496
POLR2L	3	0.1681	1
PPP1CC	3	0.018	0.983510927
PRKCB	3	0.0505	0.999991799
PRKCG	3	0.0154	0.970027573
PTN	3	0.0092	0.876167247
PXN	3	0.0674	0.999999858
RUVBL1	3	0.1411	1
SMAD1	3	0.0443	0.999964299
TBP	3	0.2921	1
TUBB	3	0.338	1
XRCC5	3	0.0233	0.995146797
XRCC6	3	0.1001	1
ACTG1	4	0.2539	1
CALM1	4	0.1212	1
CALM2	4	0.1251	1

26

CALM3	4	0.1192	1
CDC2	4	0.0465	0.999978793
CTNNB1	4	0.0945	1
DAXX	4	0.0001	0.022347638
ENSG00000204523	4	0.1659	1
ENSG00000206206	4	0.0003	0.065562161
ENSG00000206279	4	0.0002	0.044198019
GNB2L1	4	0.1274	1
ILK	4	0.0012	0.237660169
POLR2H	4	0.0195	0.988327681
PRKCE	4	0.0056	0.718931528
RPA1	4	0.0059	0.737459276
TP53	4	0.0489	0.999988002
ACTB	5	0.3806	1
GRB2	5	0.2203	1
HSPA8	5	0.1086	1
PRKCA	5	0.0238	0.995677156
SRC	5	0.1027	1

Table S3. List of all genes from GRAIL analysis.

GENE	GRAIL p-value
SCXB	0.001148628
NCAM1	0.008606012
EFNA5	0.011664265
FLJ45803	0.014495673
BARX1	0.015894637
ITGB1	0.017767359
C11orf53	0.021781178
PTPRZ1	0.02709751
UNCX	0.027882239
SULF1	0.028879408
CLPTM1L	0.030328932
DPT	0.042140848
LOC120376	0.044831624
NOS1AP	0.054991293
SHARPIN	0.057602401
MITF	0.060887937
MEPE	0.068883874
GDF6	0.089439978
PARVG	0.097736196

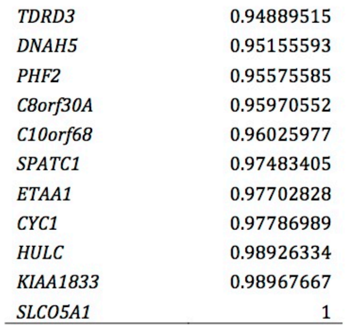
27

SCRT1	0.10862729
P2RY13	0.10887249
CNR1	0.11933196
SPN	0.12218555
CNTNAP5	0.12346167
FLI1	0.12802777
POU2AF1	0.13091422
APOA5	0.142541
VLDLR	0.15914673
TLE3	0.16129687
GRM5	0.16198902
HS3ST4	0.17305263
APOC3	0.21599063
DGAT1	0.21818261
CTNNA2	0.22377091
AASS	0.22606778
P2RY14	0.24246232
XCL2	0.24602333
TRPV3	0.24785219
P2RY12	0.25023948
APOA1	0.25100634
QPRT	0.26524557
TP63	0.26590481
SULT1A3	0.26811529
PHACTR2	0.32207847
GPR87	0.33674991
MOGAT1	0.3661129
HIPK2	0.38463158
ZNF259	0.38796092
NCOA2	0.40606951
HPS5	0.44438222
GPR22	0.44641479
SLC7A5P1	0.45452143
CALD1	0.46258854
SAA1	0.47376876
MED12L	0.49143007
DIAPH3	0.50869909
CUBN	0.51415596
HBP1	0.52923366
PCTP	0.54660408
PRDM14	0.56411462
ATP13A4	0.58650857
APOA4	0.60231621
IGSF10	0.60650531
ATXN7L1	0.6302091

28

SLC26A4	0.63330183
RUNDC2C	0.63366035
LOC440354	0.63366035
PIM1	0.63433247
MAF1	0.63857319
RAB2A	0.66352336
CLEC2L	0.66610432
KCNV2	0.69334347
DOCK8	0.71168294
OPA1	0.72616991
COG5	0.72656324
FLJ35024	0.73324721
DUS4L	0.73717725
XCL1	0.73749243
FARSB	0.7470411
AADACL2	0.75504121
RNGTT	0.75581509
GLT1D1	0.76792504
NPAL2	0.77309904
BUD13	0.77906397
ASPA	0.78337746
PRKAR2B	0.78403952
TXNRD1	0.78597238
LL22NC03-75B3.6	0.79132002
C9orf66	0.79724318
GIYD1	0.80083624
GNA15	0.80646166
SGPP2	0.80744261
SLC14A2	0.84916695
TERT	0.85875734
NAV2	0.86696994
EID3	0.86883834
EXOSC4	0.87767848
OPLAH	0.88361214
TMEM140	0.88437071
GPAA1	0.88804139
PTPDC1	0.89858484
BOP1	0.89909615
KIAA0020	0.9004611
BOLA2	0.90727594
BCAP29	0.90734909
AADAC	0.91708474
HSF1	0.93312755
CCDC7	0.94059481
C7orf49	0.94188183

29



TDRD3	0.94889515
DNAH5	0.95155593
PHF2	0.95575585
C8orf30A	0.95970552
C10orf68	0.96025977
SPATC1	0.97483405
ETAA1	0.97702828
CYC1	0.97786989
HULC	0.98926334
KIAA1833	0.98967667
SLCO5A1	1

30

Table S4. Gene ontology results from PANTHER analysis.

Gene ID	Family/Subfamily	Molecular function (GO)	Biological process (GO)	Cellular component (GO)	Protein class
UNCX	N/A	transcription factor activity	regulation of transcription from RNA polymerase II promoter;segment specification;ectoderm development;gut mesoderm development;embryonic development;skeletal system development;nervous system development;heart development;muscle organ development	N/A	homeobox transcription factor;DNA binding protein
BARX1	N/A	transcription factor activity	regulation of transcription from RNA polymerase II promoter;ectoderm development;mesoderm development;skeletal system development;nervous system development	N/A	homeobox transcription factor;DNA binding protein
EFNA5	EPHRIN-A5	receptor binding	transmembrane receptor protein tyrosine kinase signaling pathway;cell-cell signaling;transmembrane receptor protein tyrosine kinase signaling pathway;cell-cell signaling;ectoderm development;nervous system development	N/A	membrane-bound signaling molecule
SCX	SCLERAXIS	transcription factor activity	regulation of transcription from RNA polymerase II promoter;mesoderm development;skeletal system development;angiogenesis;heart development	N/A	basic helix-loop-helix transcription factor;nucleic acid binding
NCAM1	NEURAL CELL ADHESION MOLECULE 1 (N-CAM 1)	phosphoprotein phosphatase activity;receptor activity	immune system process;muscle contraction;neurological system process;induction of apoptosis;cell cycle;cell surface receptor linked signal transduction;cell-cell signaling;cell-matrix adhesion;cell-cell adhesion;protein modification process;cell motion;cell cycle;cell surface receptor linked signal transduction;cell-cell signaling;cell-matrix adhesion;cell-cell adhesion;ectoderm development;mesoderm development;angiogenesis;nervous system development;muscle organ development	extracellular matrix	immunoglobulin receptor superfamily;protein phosphatase;protein phosphatase;extracellular matrix linker protein;immunoglobulin receptor superfamily;immunoglobulin superfamily cell adhesion molecule

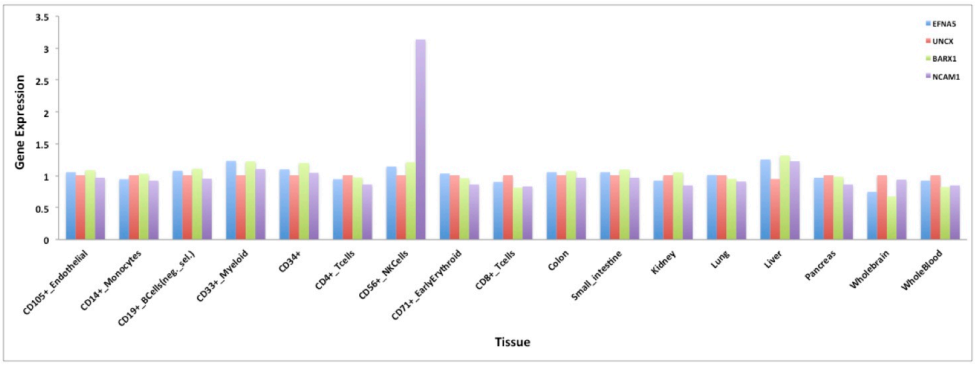


Figure S1. Standardized expression values for 4 candidate genes in 17 tissues and cell types obtained from BioGPS database.

33

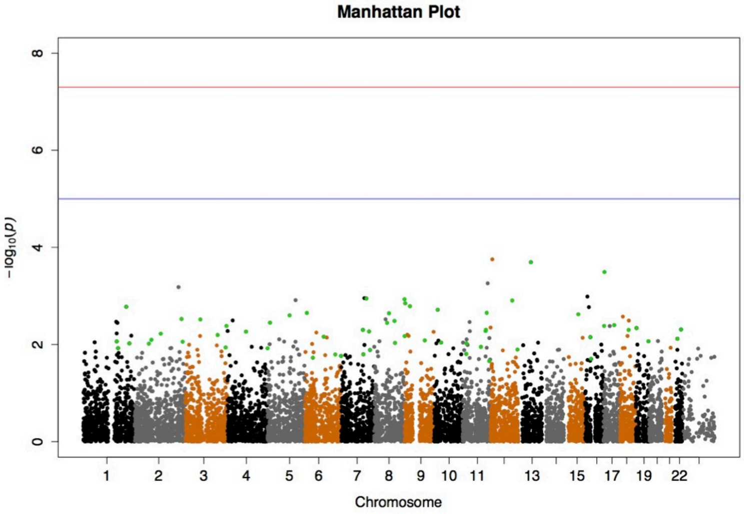


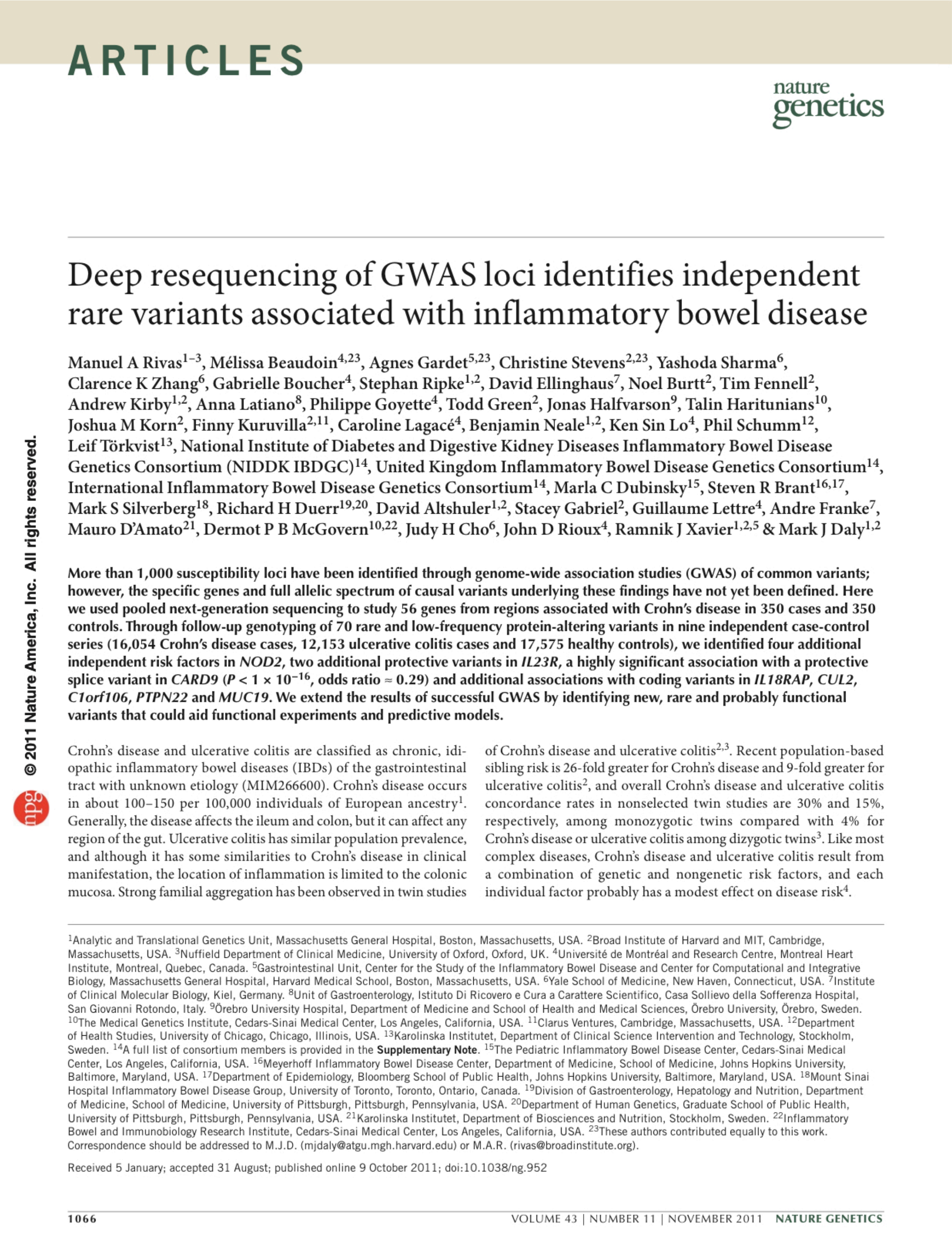
Figure S2. Manhattan plot of 8.858 iCHIP SNPs for 87 CD and 92 non-responding patients. The 59 discriminator SNPs are highlighted with green colour.

34

6.7 Drugi članki v okviru doktorskega študija

1. Rivas, M. A., Beaudoin, M., Gardet, A., et *al*. Deep resequencing of GWAS loci identifies independent rare variants associated with inflammatory bowel disease. *Nat Genet* **43**, 1066–73 (2011).
2. Trynka, G., Hunt, A. K., Bockett, N. A., et *al*. Dense genotyping reveals and localizes multiple common and rare variant association signals in celiac disease. *Nat Genet* **44,** 1193–201 (2011).
3. Liu, J., et *al*. Dense genotyping of immune-related disease regions identifies nine new risk loci for primary sclerosing cholangitis. V tisku pri *Nature Genetics*.

6.7.1 Izvirni znanstveni članek 6



nature genetics

Deep resequencing of GWAS loci identifies independent rare variants associated with inflammatory bowel disease

Manuel A Rivas[1–3], Mélissa Beaudoin[4,23], Agnes Gardet[5,23], Christine Stevens[2,23], Yashoda Sharma[6], Clarence K Zhang[6], Gabrielle Boucher[4], Stephan Ripke[1,2], David Ellinghaus[7], Noel Burtt[2], Tim Fennell[2], Andrew Kirby[1,2], Anna Latiano[8], Philippe Goyette[4], Todd Green[2], Jonas Halfvarson[9], Talin Haritunians[10], Joshua M Korn[2], Finny Kuruvilla[2,11], Caroline Lagacé[4], Benjamin Neale[1,2], Ken Sin Lo[4], Phil Schumm[12], Leif Törkvist[13], National Institute of Diabetes and Digestive Kidney Diseases Inflammatory Bowel Disease Genetics Consortium (NIDDK IBDGC)[14], United Kingdom Inflammatory Bowel Disease Genetics Consortium[14], International Inflammatory Bowel Disease Genetics Consortium[14], Marla C Dubinsky[15], Steven R Brant[16,17], Mark S Silverberg[18], Richard H Duerr[19,20], David Altshuler[1,2], Stacey Gabriel[2], Guillaume Lettre[4], Andre Franke[7], Mauro D'Amato[21], Dermot P B McGovern[10,22], Judy H Cho[6], John D Rioux[4], Ramnik J Xavier[1,2,5] & Mark J Daly[1,2]

More than 1,000 susceptibility loci have been identified through genome-wide association studies (GWAS) of common variants; however, the specific genes and full allelic spectrum of causal variants underlying these findings have not yet been defined. Here we used pooled next-generation sequencing to study 56 genes from regions associated with Crohn's disease in 350 cases and 350 controls. Through follow-up genotyping of 70 rare and low-frequency protein-altering variants in nine independent case-control series (16,054 Crohn's disease cases, 12,153 ulcerative colitis cases and 17,575 healthy controls), we identified four additional independent risk factors in *NOD2*, two additional protective variants in *IL23R*, a highly significant association with a protective splice variant in *CARD9* ($P < 1 \times 10^{-16}$, odds ratio ≈ 0.29) and additional associations with coding variants in *IL18RAP*, *CUL2*, *C1orf106*, *PTPN22* and *MUC19*. We extend the results of successful GWAS by identifying new, rare and probably functional variants that could aid functional experiments and predictive models.

Crohn's disease and ulcerative colitis are classified as chronic, idiopathic inflammatory bowel diseases (IBDs) of the gastrointestinal tract with unknown etiology (MIM266600). Crohn's disease occurs in about 100–150 per 100,000 individuals of European ancestry[1]. Generally, the disease affects the ileum and colon, but it can affect any region of the gut. Ulcerative colitis has similar population prevalence, and although it has some similarities to Crohn's disease in clinical manifestation, the location of inflammation is limited to the colonic mucosa. Strong familial aggregation has been observed in twin studies of Crohn's disease and ulcerative colitis[2,3]. Recent population-based sibling risk is 26-fold greater for Crohn's disease and 9-fold greater for ulcerative colitis[2], and overall Crohn's disease and ulcerative colitis concordance rates in nonselected twin studies are 30% and 15%, respectively, among monozygotic twins compared with 4% for Crohn's disease or ulcerative colitis among dizygotic twins[3]. Like most complex diseases, Crohn's disease and ulcerative colitis result from a combination of genetic and nongenetic risk factors, and each individual factor probably has a modest effect on disease risk[4].

[1]Analytic and Translational Genetics Unit, Massachusetts General Hospital, Boston, Massachusetts, USA. [2]Broad Institute of Harvard and MIT, Cambridge, Massachusetts, USA. [3]Nuffield Department of Clinical Medicine, University of Oxford, Oxford, UK. [4]Université de Montréal and Research Centre, Montreal Heart Institute, Montreal, Quebec, Canada. [5]Gastrointestinal Unit, Center for the Study of the Inflammatory Bowel Disease and Center for Computational and Integrative Biology, Massachusetts General Hospital, Harvard Medical School, Boston, Massachusetts, USA. [6]Yale School of Medicine, New Haven, Connecticut, USA. [7]Institute of Clinical Molecular Biology, Kiel, Germany. [8]Unit of Gastroenterology, Istituto Di Ricovero e Cura a Carattere Scientifico, Casa Sollievo della Sofferenza Hospital, San Giovanni Rotondo, Italy. [9]Örebro University Hospital, Department of Medicine and School of Health and Medical Sciences, Örebro University, Örebro, Sweden. [10]The Medical Genetics Institute, Cedars-Sinai Medical Center, Los Angeles, California, USA. [11]Clarus Ventures, Cambridge, Massachusetts, USA. [12]Department of Health Studies, University of Chicago, Chicago, Illinois, USA. [13]Karolinska Institutet, Department of Clinical Science Intervention and Technology, Stockholm, Sweden. [14]A full list of consortium members is provided in the **Supplementary Note**. [15]The Pediatric Inflammatory Bowel Disease Center, Cedars-Sinai Medical Center, Los Angeles, California, USA. [16]Meyerhoff Inflammatory Bowel Disease Center, Department of Medicine, School of Medicine, Johns Hopkins University, Baltimore, Maryland, USA. [17]Department of Epidemiology, Bloomberg School of Public Health, Johns Hopkins University, Baltimore, Maryland, USA. [18]Mount Sinai Hospital Inflammatory Bowel Disease Group, University of Toronto, Toronto, Ontario, Canada. [19]Division of Gastroenterology, Hepatology and Nutrition, Department of Medicine, School of Medicine, University of Pittsburgh, Pittsburgh, Pennsylvania, USA. [20]Department of Human Genetics, Graduate School of Public Health, University of Pittsburgh, Pittsburgh, Pennsylvania, USA. [21]Karolinska Institutet, Department of Biosciences and Nutrition, Stockholm, Sweden. [22]Inflammatory Bowel and Immunobiology Research Institute, Cedars-Sinai Medical Center, Los Angeles, California, USA. [23]These authors contributed equally to this work. Correspondence should be addressed to M.J.D. (mjdaly@atgu.mgh.harvard.edu) or M.A.R. (rivas@broadinstitute.org).

Received 5 January; accepted 31 August; published online 9 October 2011; doi:10.1038/ng.952

© 2011 Nature America, Inc. All rights reserved.

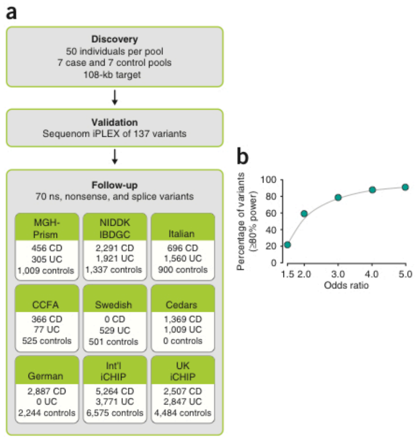


Figure 1 Overview. (**a**) Schematic of Crohn's disease rare-variant phenotype project. (**b**) Power to detect single-marker rare-variant association in follow-up sample sets. We report the results of the Crohn's disease pooled resequencing project with follow-up genotypes in 13,167 Crohn's disease cases, 12,153 ulcerative colitis cases and 15,331 healthy controls. We report that of the 70 markers successfully genotyped, 22%, 60%, 79%, 88% and 91% have at least 80% power to detect association at minor allele frequency ORs of 1.5, 2, 3, 4 and 5, respectively (see also **Supplementary Fig. 3a**,**b**), suggesting that we can address the contribution of rare and low-frequency polymorphisms in GWAS loci to IBD. OR, odds ratio.

© 2011 Nature America, Inc. All rights reserved.

Common immune-mediated diseases such as IBD have a genetic basis. However, until recently, identifying disease susceptibility genes has been challenging for common, polygenic disease[5,6]. With the development of HapMap and GWAS technology, the number of bona fide risk loci that have been identified and replicated for complex trait genetics in general, and for IBD in particular, has markedly increased. In Crohn's disease, individual GWAS scans and follow-up meta-analyses have identified >71 susceptibility loci and have provided insights beyond the two loci established before the GWAS era[7,8]. Similarly, in ulcerative colitis, GWAS efforts have identified 47 susceptibility loci[9,10] and, after accounting for the many alleles associated with both diseases, 99 distinct associations have been documented for IBD. Although these findings have clarified disease pathways, the common SNPs identified are generally of modest effect and explain only ~23% of the overall variance in Crohn's disease risk. Moreover, most of the associated variants do not have known or obvious function, and many implicate regions with multiple genes, limiting biological extrapolation.

SNPs implicated by GWAS are tightly correlated with other SNPs in the region and are probably in linkage disequilibrium (LD) with the causal variant rather than causal themselves. A complete catalog of all variation is required in the search for causal variants[11,12]. However, even in denser reference data from the 1000 Genomes Project, most GWAS hits are not correlated with an obvious functional variant and therefore do not conclusively implicate a unique gene. If independently associated rare coding variation is discovered in a gene within a region implicated by GWAS, the gene harboring such variants is directly implicated. Furthermore, additional heritability can be explained and specific alleles identified for direct functional experimentation. In Crohn's disease, multiple independent associated alleles have been documented at *NOD2* and *IL23R*[13,14]. Exhaustive sequencing of genomic regions has recently become possible for the first time with the advent of next-generation sequencing (NGS) technologies. Growing collections of genome sequences through international efforts like the 1000 Genomes Project are driving the development of laboratory study designs and analytic methods for using large-scale genomic sequencing in human genetic discovery[15].

Targeted sequencing of pooled samples allows researchers to efficiently and cost-effectively capture all variation in a limited target region that has been selectively amplified in multiple DNA samples[16,17]. This approach allows efficient use of NGS technologies, which generate billions of base pairs per experimental unit yet introduce challenges in data processing and analysis, for the discovery of new variants and assess their potential association with disease. We describe here a pooled NGS study of 350 Crohn's disease cases and 350 controls across coding exons of 56 genes contained in regions of confirmed association with Crohn's disease[7], and we introduce new SNP calling methods for pooled targeted sequencing projects implemented in the software Syzygy. We further evaluated new, potentially functional rare variants identified in the survey in nine independent case-control series, confirming a role for functional, rare variants in *CARD9*, *NOD2*, *IL23R* and *IL18RAP* and identifying others in *MUC19*, *CUL2*, *PTPN22* and *C1orf106* that were significantly associated with IBD. The results lend further support to an emerging paradigm in both rare diseases (Hirschsprung's disease and Bardet-Biedl syndrome) and common phenotypes (serum lipids, QT-interval and height and type 1 diabetes) in which both common, low-penetrance and rarer, often higher-penetrance alleles exist in the same gene. They also suggest that deep sequencing of regions implicated by GWAS may be effective in increasing knowledge about the heritability of specific functional alleles in complex disease[16,18–21].

RESULTS

Discovery of new variants using pooled sequencing

We selected 350 Crohn's disease cases and 350 healthy controls of European ancestry from among samples collected by the NIDDK IBDGC with genome-wide SNP data[14,22]. We pooled samples in batches of 50 cases or 50 controls matched for European ancestry using GWAS data. One pool of 50 cases was drawn from self-reported and empirically confirmed (by GWAS data[22]) Jewish ancestry and was matched with one pool of 50 equivalently defined Jewish controls. The remaining pools of cases and controls were selected from the non-Jewish European-American samples. Samples were pooled only after two rounds of quantification and normalization to ensure that the initial DNA pool accurately reflected sample allele frequencies. For each pool, we carried out PCR amplification to capture the 107.5-kb target region, which included 645 nuclear-encoded exons (**Supplementary Tables 1** and **2**). We amplified each sample in 593 PCR reactions. The successful PCR amplicons were combined in equimolar amounts, concatenated and sheared to construct libraries. The 14 libraries were sequenced using Illumina Genome Analyzer flow cells, with one pool per lane (see Online Methods; **Fig. 1a**). High-throughput sequencing yielded large amounts of high-quality data for each pool. We captured 91% of our nuclear target regions at ≥100× coverage and achieved 1,500× median coverage per pool (corresponding to 30× per sample or 15× per individual chromosome; **Supplementary Fig. 1**).

We next aimed to identify rare and low-frequency single-nucleotide variants (SNVs) in the pooled samples. We developed a variant calling method, Syzygy, to accommodate the specific pooled study design

© 2011 Nature America, Inc. All rights reserved.

Table 1 Variant discovery summary

Category	High quality	Moderate quality
Variants identified	429	173
dbSNP (%)	45	24
NS/S	1.4	1.7
Ti/Tv	2.3	1.4

Using Syzygy, we detected 429 high-confidence variants (240 nonsynonymous sites, 169 synonymous sites and 20 variants within 5 bp of the nearest splice site) within our 107.5-kb targeted region with a dbSNP rate of 45%, NS/S of 1.42, and Ti/Tv of 2.3 in the pooled sequencing experiment with 350 Crohn's disease cases and 350 healthy controls.

and identify rare variants (see **Supplementary Methods**). Through empirical modeling of the sequencing error processes and filters to remove sites with strand inconsistency or clusters of variants suggestive of read misalignment, Syzygy detected 429 putatively high-confidence variants (240 nonsynonymous sites, 169 synonymous sites and 20 intronic variants within 5 bp of a splice junction) within our 107.5-kb targeted region, with 45% of the variants already included in dbSNP using dbSNP version 132, nonsynonymous/synonymous ratio (NS/S) of 1.42 and transition/transversion ratio (Ti/Tv) of 2.3 (**Table 1**). Because we designed our experiments to detect variants correctly at the limit of machine quality, we estimated the proposed set of false-positive SNPs that would need to be eliminated in subsequent genotyping. Both the proportion of variants in dbSNP and the Ti/Tv suggest a relatively high true-positive rate in this data set. Specifically, high-depth individual-level sequencing of 1,000 genes carried out by the 1000 Genomes Project (called Pilot 3) in 697 samples identified a high-quality SNP set with the same dbSNP percentage (dbSNP version 129), whereas the Ti/Tv detected here suggests a ~90% true-positive rate[23]. To confirm this, we selected a random subset of 137 high-confidence functional nonsynonymous, nonsense and putative splice-variant SNPs for Sequenom iPLEX genotyping of all samples in the sequenced pools and validated 91.2% of them (**Fig. 1a**). Using a canonical expectation of $\left(\theta \times \sum_{i=1}^{n-1}\frac{1}{i} \times Nbases\right)$, where n denotes the number of chromosomes and Nbases represents the targeted bases, or the rate observed directly in 1000 Genomes Pilot 3, we would expect to see ~470 variants across the successfully queried target. Sensitivity for singletons, however, is incomplete at the lower end of coverage in our experiment (**Supplementary Fig. 1**) and accounts for the modest deficit in our study.

A challenge in pooled genotyping or sequencing experiments is accurate recovery of allele frequencies. We were surprised to observe a strong correlation between genotype frequencies and frequencies estimated for sequence data ($r^2 \approx 0.99$) using the method in Syzygy, suggesting that accurate quantification of DNAs in the pooling steps led to experimental recovery of the pool composition. We also observed a strong correlation between the case-control test statistic estimated with the pooled data and the test statistic in the genotype data ($r^2 \approx 0.925$; **Supplementary Fig. 2**).

To test the role of these rare variants, we identified all nonsynonymous, nonsense or splice-site variants that occurred in two or more copies up to a frequency of 5%, for a total of 115 variants (**Supplementary Table 3**). Excluding known GWAS-associated low-frequency coding variants at *NOD2*, *IL23R* and *LRRK2-MUC19*, we carried out follow-up genotyping for 70 of these markers in nine independent case-control series totaling 16,054 Crohn's disease cases, 12,153 ulcerative colitis cases and 17,575 healthy controls. These included (i) samples from the Prospective Registry in IBD Study at MGH (PRISM); (ii) samples assembled from throughout North America and Australia by the NIDDK IBDGC; (iii) an Italian-Dutch case-control sample; (iv) Crohn's and Colitis Foundation of America (CCFA) Repository Collection; (v) Swedish samples; (vi) Cedars samples; (vii) German samples and Immunochip genotype data provided by (viii) the International IBD Genetics Consortium and (ix) UK IBD Genetics Consortium (rare coding variants discovered in this study contributed to the Immunochip design; **Fig. 1a**). Samples i and iii–vii were genotyped for sets of markers using Sequenom iPLEX. Sample ii genotyping was done as part of a larger NIDDK IBDGC Illumina GoldenGate study. Because of design constraints and assay failures, not all markers were examined in all nine follow-up sample sets (see **Supplementary Methods** for details of follow-up genotyping). We demonstrate that the current study design is well positioned to address the overall contribution of variants in coding regions of GWAS loci to IBD (**Fig. 1b**, **Supplementary Fig. 3** and **Supplementary Methods**).

The few nonreference alleles expected for many of these variants in each substudy precludes the use of asymptotic statistics common to most association studies. Population structure is probably an even more substantial problem at low frequencies, demanding a stratified analysis retaining strict population case-control matching. Therefore, we used a mega-analysis of rare variants (MARV) that provides a permutation-based estimate of significance, constraining all permutations to be within each subgroup and thus accommodating arbitrary numbers of sample subsets of diverse population and case-control origin without power loss for single-marker and group-marker analysis (see Online Methods). Given a target set of 70 variants, we would expect <1 SNP to exceed $P < 0.01$ by chance in the follow-up analyses and define traditional experiment-wise significance to be $P = 0.0007$. As we explored both Crohn's disease and ulcerative colitis in follow-up studies, our primary analysis compared all IBD (Crohn's disease and ulcerative colitis) cases versus controls to maximize power for genes in which the same common variants have been conclusively associated with both diseases with similar effect (such as *CARD9*). For genes specifically associated with Crohn's disease only (such as *NOD2*), the ulcerative colitis group was combined with controls (for details, see ref. 10).

New protective splice variant in *CARD9*

CARD9 is associated with both Crohn's disease and ulcerative colitis risk, with a common coding variant (rs4077515 creating substitution p.S12N with both alleles of roughly equal frequency) that represents a 'typical' GWAS hit (odds ratio (OR) ≈ 1.2 in both diseases)[8,9]. In the pooled sequencing, we identified a splice-site variant in *CARD9* (**Fig. 2** and **Supplementary Fig. 4**) that altered the first base after exon 11 in six controls and zero cases, suggesting a potentially strong protective effect. Follow-up analyses confirmed a highly significant association ($P < 10^{-16}$), with the allele appearing in about 0.20% of cases and 0.64% of controls (OR ≈ 0.3; **Table 2** and **Supplementary Table 4**).

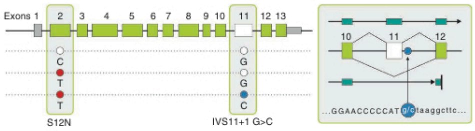


Figure 2 *CARD9* protective splice-site variant and predicted transcript. Splice-site variant IVS11+1C>G (OR = 0.29), conferring protection against Crohn's disease with predicted transcript. This hypothetical transcript has been observed in spleen, lymph-node and PBMC-derived cDNA libraries. We predict exon 11 to be skipped and the alternative transcript to include exon 9 mRNA sequence continuing to exon 12, including 21 amino acids before reaching a premature stop.

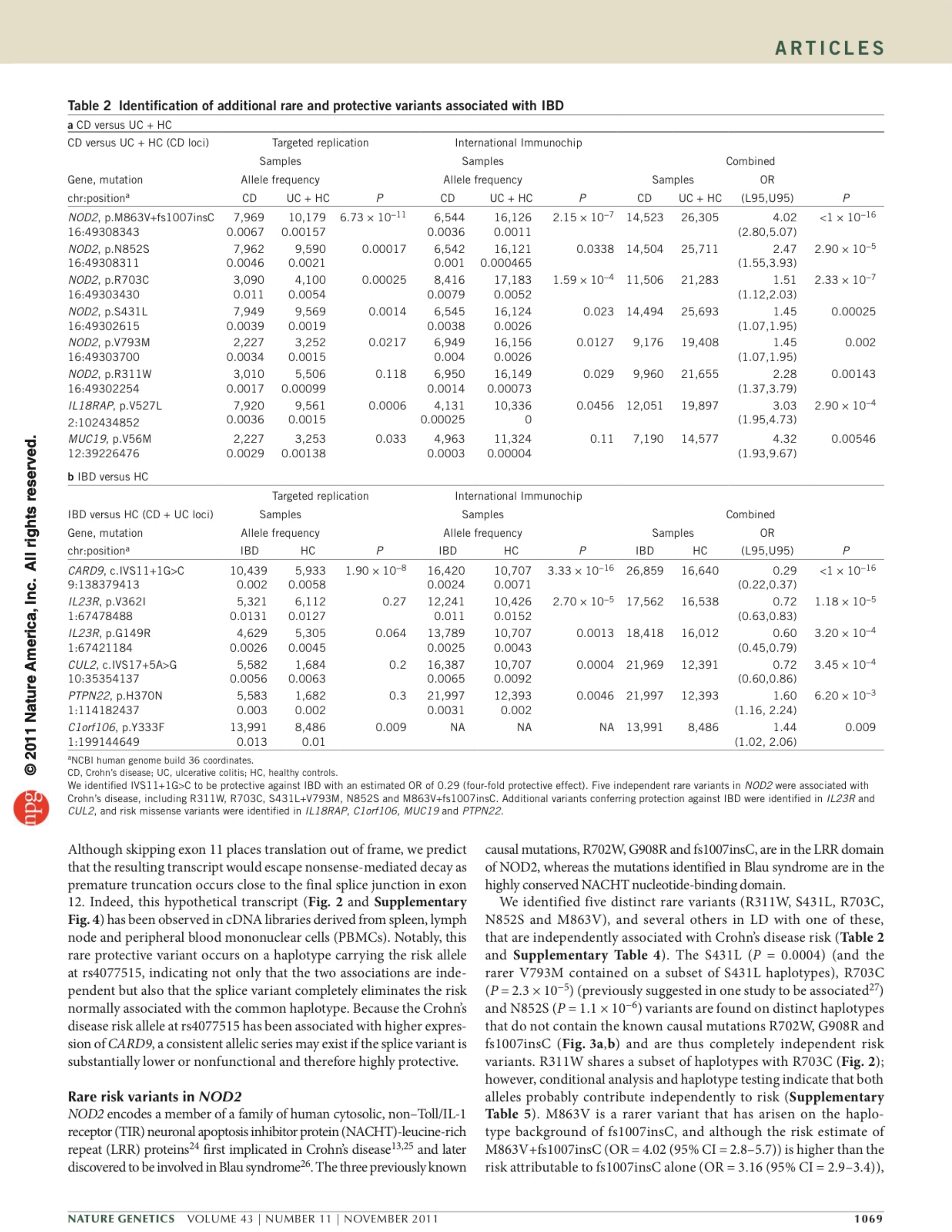


Table 2 Identification of additional rare and protective variants associated with IBD

a CD versus UC + HC

CD versus UC + HC (CD loci)	Targeted replication			International Immunochip						
	Samples			Samples					Combined	
Gene, mutation	Allele frequency			Allele frequency			Samples		OR	
chr:position[a]	CD	UC + HC	*P*	CD	UC + HC	*P*	CD	UC + HC	(L95,U95)	*P*
NOD2, p.M863V+fs1007insC 16:49308343	7,969 0.0067	10,179 0.00157	6.73×10^{-11}	6,544 0.0036	16,126 0.0011	2.15×10^{-7}	14,523	26,305	4.02 (2.80,5.07)	$<1 \times 10^{-16}$
NOD2, p.N852S 16:49308311	7,962 0.0046	9,590 0.0021	0.00017	6,542 0.001	16,121 0.000465	0.0338	14,504	25,711	2.47 (1.55,3.93)	2.90×10^{-5}
NOD2, p.R703C 16:49303430	3,090 0.011	4,100 0.0054	0.00025	8,416 0.0079	17,183 0.0052	1.59×10^{-4}	11,506	21,283	1.51 (1.12,2.03)	2.33×10^{-7}
NOD2, p.S431L 16:49302615	7,949 0.0039	9,569 0.0019	0.0014	6,545 0.0038	16,124 0.0026	0.023	14,494	25,693	1.45 (1.07,1.95)	0.00025
NOD2, p.V793M 16:49303700	2,227 0.0034	3,252 0.0015	0.0217	6,949 0.004	16,156 0.0026	0.0127	9,176	19,408	1.45 (1.07,1.95)	0.002
NOD2, p.R311W 16:49302254	3,010 0.0017	5,506 0.00099	0.118	6,950 0.0014	16,149 0.00073	0.029	9,960	21,655	2.28 (1.37,3.79)	0.00143
IL18RAP, p.V527L 2:102434852	7,920 0.0036	9,561 0.0015	0.0006	4,131 0.00025	10,336 0	0.0456	12,051	19,897	3.03 (1.95,4.73)	2.90×10^{-4}
MUC19, p.V56M 12:39226476	2,227 0.0029	3,253 0.00138	0.033	4,963 0.0003	11,324 0.00004	0.11	7,190	14,577	4.32 (1.93,9.67)	0.00546

b IBD versus HC

	Targeted replication			International Immunochip						
IBD versus HC (CD + UC loci)	Samples			Samples					Combined	
Gene, mutation	Allele frequency			Allele frequency			Samples		OR	
chr:position[a]	IBD	HC	*P*	IBD	HC	*P*	IBD	HC	(L95,U95)	*P*
CARD9, c.IVS11+1G>C 9:138379413	10,439 0.002	5,933 0.0058	1.90×10^{-8}	16,420 0.0024	10,707 0.0071	3.33×10^{-16}	26,859	16,640	0.29 (0.22,0.37)	$<1 \times 10^{-16}$
IL23R, p.V362I 1:67478488	5,321 0.0131	6,112 0.0127	0.27	12,241 0.011	10,426 0.0152	2.70×10^{-5}	17,562	16,538	0.72 (0.63,0.83)	1.18×10^{-5}
IL23R, p.G149R 1:67421184	4,629 0.0026	5,305 0.0045	0.064	13,789 0.0025	10,707 0.0043	0.0013	18,418	16,012	0.60 (0.45,0.79)	3.20×10^{-4}
CUL2, c.IVS17+5A>G 10:35354137	5,582 0.0056	1,684 0.0063	0.2	16,387 0.0065	10,707 0.0092	0.0004	21,969	12,391	0.72 (0.60,0.86)	3.45×10^{-4}
PTPN22, p.H370N 1:114182437	5,583 0.003	1,682 0.002	0.3	21,997 0.0031	12,393 0.002	0.0046	21,997	12,393	1.60 (1.16, 2.24)	6.20×10^{-3}
C1orf106, p.Y333F 1:199144649	13,991 0.013	8,486 0.01	0.009	NA	NA	NA	13,991	8,486	1.44 (1.02, 2.06)	0.009

[a]NCBI human genome build 36 coordinates.

CD, Crohn's disease; UC, ulcerative colitis; HC, healthy controls.

We identified IVS11+1G>C to be protective against IBD with an estimated OR of 0.29 (four-fold protective effect). Five independent rare variants in *NOD2* were associated with Crohn's disease, including R311W, R703C, S431L+V793M, N852S and M863V+fs1007insC. Additional variants conferring protection against IBD were identified in *IL23R* and *CUL2*, and risk missense variants were identified in *IL18RAP*, *C1orf106*, *MUC19* and *PTPN22*.

© 2011 Nature America, Inc. All rights reserved.

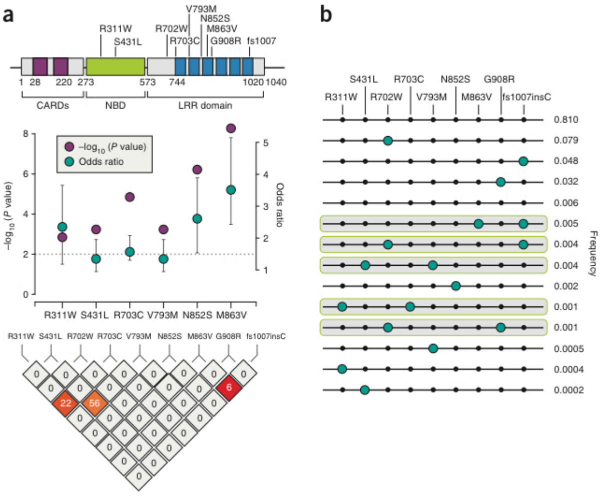
Although skipping exon 11 places translation out of frame, we predict that the resulting transcript would escape nonsense-mediated decay as premature truncation occurs close to the final splice junction in exon 12. Indeed, this hypothetical transcript (**Fig. 2** and **Supplementary Fig. 4**) has been observed in cDNA libraries derived from spleen, lymph node and peripheral blood mononuclear cells (PBMCs). Notably, this rare protective variant occurs on a haplotype carrying the risk allele at rs4077515, indicating not only that the two associations are independent but also that the splice variant completely eliminates the risk normally associated with the common haplotype. Because the Crohn's disease risk allele at rs4077515 has been associated with higher expression of *CARD9*, a consistent allelic series may exist if the splice variant is substantially lower or nonfunctional and therefore highly protective.

Rare risk variants in *NOD2*

NOD2 encodes a member of a family of human cytosolic, non–Toll/IL-1 receptor (TIR) neuronal apoptosis inhibitor protein (NACHT)-leucine-rich repeat (LRR) proteins[24] first implicated in Crohn's disease[13,25] and later discovered to be involved in Blau syndrome[26]. The three previously known causal mutations, R702W, G908R and fs1007insC, are in the LRR domain of NOD2, whereas the mutations identified in Blau syndrome are in the highly conserved NACHT nucleotide-binding domain.

We identified five distinct rare variants (R311W, S431L, R703C, N852S and M863V), and several others in LD with one of these, that are independently associated with Crohn's disease risk (**Table 2** and **Supplementary Table 4**). The S431L ($P = 0.0004$) (and the rarer V793M contained on a subset of S431L haplotypes), R703C ($P = 2.3 \times 10^{-5}$) (previously suggested in one study to be associated[27]) and N852S ($P = 1.1 \times 10^{-6}$) variants are found on distinct haplotypes that do not contain the known causal mutations R702W, G908R and fs1007insC (**Fig. 3a**,**b**) and are thus completely independent risk variants. R311W shares a subset of haplotypes with R703C (**Fig. 2**); however, conditional analysis and haplotype testing indicate that both alleles probably contribute independently to risk (**Supplementary Table 5**). M863V is a rarer variant that has arisen on the haplotype background of fs1007insC, and although the risk estimate of M863V+fs1007insC (OR = 4.02 (95% CI = 2.8–5.7)) is higher than the risk attributable to fs1007insC alone (OR = 3.16 (95% CI = 2.9–3.4)),

Figure 3 Identification of additional rare variants in *NOD2* associated with Crohn's disease. (**a**) Five additional risk variants were discovered in *NOD2*. $-\log_{10}P$ and minor allele ORs with 95% confidence intervals indicated with error bars and haplotype block, where *D*′ taking values from 0–1 (white to red) represents the extent of linkage disequilibrium between markers and numbers represent r^2 between markers. (**b**) *NOD2* haplotypes observed in 700 individuals with overlapping genotype data (R311W, S431L, R702W, R703C, V793M, N852S, M863V, G908R and fs1007insC). S431L and V793M are in tight LD and we regard this as one unit (S431L V793M). R703C is at a higher frequency than R311W although they share haplotypes. Conditional analysis (**Supplementary Table 3**) demonstrates independent contributions. M863V lies on background haplotype of fs1007insC.



© 2011 Nature America, Inc. All rights reserved.

the low frequency of M863V makes its functionality unclear. Thus, in later calculations of variance explained, we did not count this as an additional risk factor.

Functional assessment of additional associated alleles in *NOD2*

Through assays to identify the effect of the mutations on NOD2 intracellular localization, we found that the S431L mutant and the well-studied insertion mutant (fs1007insC) did not localize to the membrane area, in contrast with N852S (**Fig. 4**). We next determined whether the NOD2 mutants S431L and N852S activated NF-κB in response to the NOD2 ligand muramyl dipeptide (MDP). HEK293T cells were transfected with the point mutants, wild-type NOD2 and the well-studied fs1007insC mutant (**Fig. 4**). Western blot analysis showed that the point mutations did not affect expression compared with the wild-type protein (**Fig. 4**). As published earlier[28,29], the fs1007insC mutant did not induce NF-κB activation after MDP stimulation. The MDP-induced NF-κB activation was also impaired in the presence of S431L and N852S (**Fig. 4**).

Together, these results indicate that the N852S alteration in the LRR domain may perturb MDP recognition without affecting NOD2 intracellular localization, similarly to the common R702W and G908R alterations[28]. This contrasts with the fs1007insC mutation, which also affects targeting of NOD2 to the membrane area. The S431L variant is in the nucleotide-binding domain of the protein and impairs both localization and MDP-induced NF-κB activation. These findings are similar to earlier studies demonstrating that critical residues in the nucleotide-binding domain region attenuate MDP-dependent NF-κB activation[29]. Further studies are needed to determine the instructive role of NOD2 mutants in coordinating autophagy, control of cellular stress signals and adaptive immune responses.

N852S and M863V variants are more common in Ashkenazi Jewish individuals

The highest reported prevalence of Crohn's disease is in subjects of Ashkenazi Jewish descent, with two to four times higher prevalence than non-Jewish populations[30]. An earlier study[31] has screened the *NOD2* gene for rare variants and identified five previously unreported changes (D113N, D357A, I363F, L550V and N852S). N852S occurs only in Ashkenazi Jewish individuals and has been proposed to predispose disease, with seven transmissions and only one nontransmission from heterozygous parents to affected offspring in an Ashkenazi Jewish family collection, concordant with the case-control observations in this study. In our study, Ashkenazi Jewish individuals had a much higher frequency of both N852S and M863V (4% and 2%, respectively, in Ashkenazi Jewish Crohn's disease cases, and 0.5% for N852S and M863V in non-Ashkenazi Jewish Crohn's disease cases). This accounts for the greater incidence of these alleles in the first replication column of **Table 2**, as NIDDK studies in particular had a specific and significant Ashkenazi Jewish collection.

We examined the haplotype carrying N852S in Ashkenazi Jewish individuals (determined given the existence of two homozygote cases) and in non-Ashkenazi Jewish individuals in the subset of samples with existing GWAS genotype data[8,9,14,22]. We found that the N852S variant in Ashkenazi individuals lies on a unique extended haplotype of several megabases (≥2 Mb to the left and right). However, the N852S variant in non-Ashkenazi Jewish individuals does not share the extended background haplotype. In Ashkenazi individuals, the average shared distance between a pair of N852S chromosomes is ≥4 Mb, whereas in non-Ashkenazi individuals, the average shared distance between a pair of N852S chromosomes is 0.5 Mb (**Supplementary Fig. 5**), suggesting that the variant is reasonably old but that a single copy was stochastically enriched in the recent Ashkenazi bottleneck ~25 generations ago.

Rare protective variants in *IL23R*

We also identified significant protective effects of substitutions G149R ($P = 3.2 \times 10^{-4}$) and V362I ($P = 1.2 \times 10^{-5}$) in *IL23R*. This confirms recent findings[32] and indicates that each of these variants have a protective effect equivalent to that of the more common R381Q substitution (**Table 2** and **Supplementary Table 4**), although they arose on different haplotype backgrounds and are not in LD with R381Q. Despite the large follow-up sample size (n = 31,747), we did not find evidence for a protective effect of the previously reported R86Q variant (P = 0.94). *IL23R* signaling is attenuated in T helper type 17 (T_H17) cells generated from healthy subjects carrying the R381Q substitution, leading to a decrease of IL-17A secretion in response to IL-23, and indicating that R381Q is associated with reduced T_H17

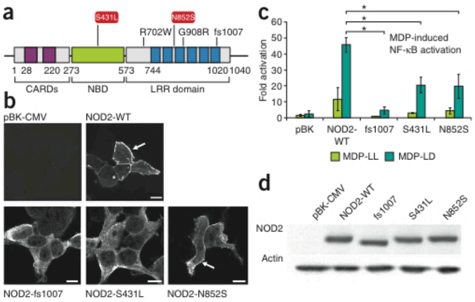


Figure 4 Functional analyses of NOD2 variants. (**a**) Schematic of NOD2 protein domains and localization. (**b**) HEK293T cells were transfected with NOD2 constructs and fixed using 4% paraformaldehyde at 24 h after transfection. Cells were then subjected to immunofluorescent staining to detect NOD2 and fluorescence was collected using a confocal microscope. Image gallery of a single confocal section. (**c**) HEK293T cells were transfected with NOD2 constructs and reporter plasmids encoding firefly luciferase cloned under a promoter containing NF-κB elements and with a plasmid encoding renilla luciferase as a transfection control. After 24 h, cells were stimulated with MDP–L-alanine–L-glutamine (LL) or MDP–L-alanine–D-glutamine (LD) (10 μg/ml) for 6 h. Transcriptional activation was quantified by ratios of firefly luciferase activity to renilla luciferase activity. Data were normalized to unstimulated condition with empty vector transfection. Statistical analyses were carried out using Student's *t*-test (*$P < 0.05$). Error bars represent 95% CI for fold activation. (**d**) Cell lysates were also collected and subjected to western blot analysis to detect NOD2 and actin expression levels. Scale bars, 10 μm.

npg © 2011 Nature America, Inc. All rights reserved.

responses[33]. In addition, recent studies have highlighted a role for IL-23 in T_H17 cell lineage commitment without TGF-β. This alternate mode of T_H17 differentiation, dependent on *IL23R* expression, seems to have a greater pathogenic role than TGF-β–induced T_H17 differentiation, highlighting the value of discovering protective variants in autoimmunity[34]. Future therapies for autoimmune disease should consider the phenotypic characteristics of pathogenic T_H17 cells generated without TGF-β, and their signaling pathways as possible targets.

Additional rare risk and protective variants

Although Crohn's disease and ulcerative colitis do not share an association with the common variant rs2058660 in *IL18RAP* (minor allele frequency (MAF) = 0.23, OR ≈ 1.19 for Crohn's disease, chr2:102.17–102.67 Mb), an association of rs2058660 with celiac disease has recently been documented[35]. We identified a rare risk missense variant, V527L (MAF = 0.003), in *IL18RAP* with an estimated minor allele OR of 2.79 for Crohn's disease. In addition, a low-frequency missense variant, Y333F (MAF = 0.008), in *C1orf106* was associated with risk of both Crohn's disease and ulcerative colitis.

A common *CUL2* variant (rs12261843, MAF = 0.30, OR ≈ 1.15) is associated with both Crohn's disease and ulcerative colitis risk[8,9]. In the pooled sequencing experiment, we identified a splice-site variant in *CUL2* altering a nucleotide five bases downstream of exon 17 with an estimated OR of 0.72 in the follow-up samples (MAF = 0.007). Notably, several members of the ubiquitin proteosome are present in the autophagy interaction network, including *CUL2*, suggesting cross-talk between these processes in intracellular quality control and immunity[36].

A common missense variant (risk allele frequency = 0.90; OR = 1.31, rs2476601) in *PTPN22* is associated with Crohn's disease[7,8], type 1 diabetes[37], rheumatoid arthritis[38] and vitiligo[39]. In this instance, the direction of association differs in different diseases, with the minor allele (Trp) strongly associated with type 1 diabetes, rheumatoid arthritis and vitiligo but highly protective against Crohn's disease. Analysis of rare variants in IBD cases versus healthy controls showed a modest risk effect ($P = 0.00026$, minor allele OR = 1.6) for a rare (MAF = 0.003) *PTPN22* missense variant (H370N). Ongoing studies in other autoimmune diseases may help clarify the relevance of H370N and rs2476601 in different conditions.

Through examination of haplotype structure (**Supplementary Fig. 6**) and formal conditional analysis (**Supplementary Table 6**), we found that the rare variants highlighted in *IL18RAP*, *MUC19*, *C1orf106*, *PTPN22* and *CUL2* are independent of the common associated GWAS variants. Specifically, the rare variants at *IL18RAP* and *MUC19* reside on the common higher-risk background but confer independently significant risk, the rare variants at *PTPN22* and *C1orf106* occur on the common low-risk background and are therefore independent, and the rare variant at *CUL2* is protective and in weak LD with common-risk variants at that locus.

Heritability estimates of rare associated variants

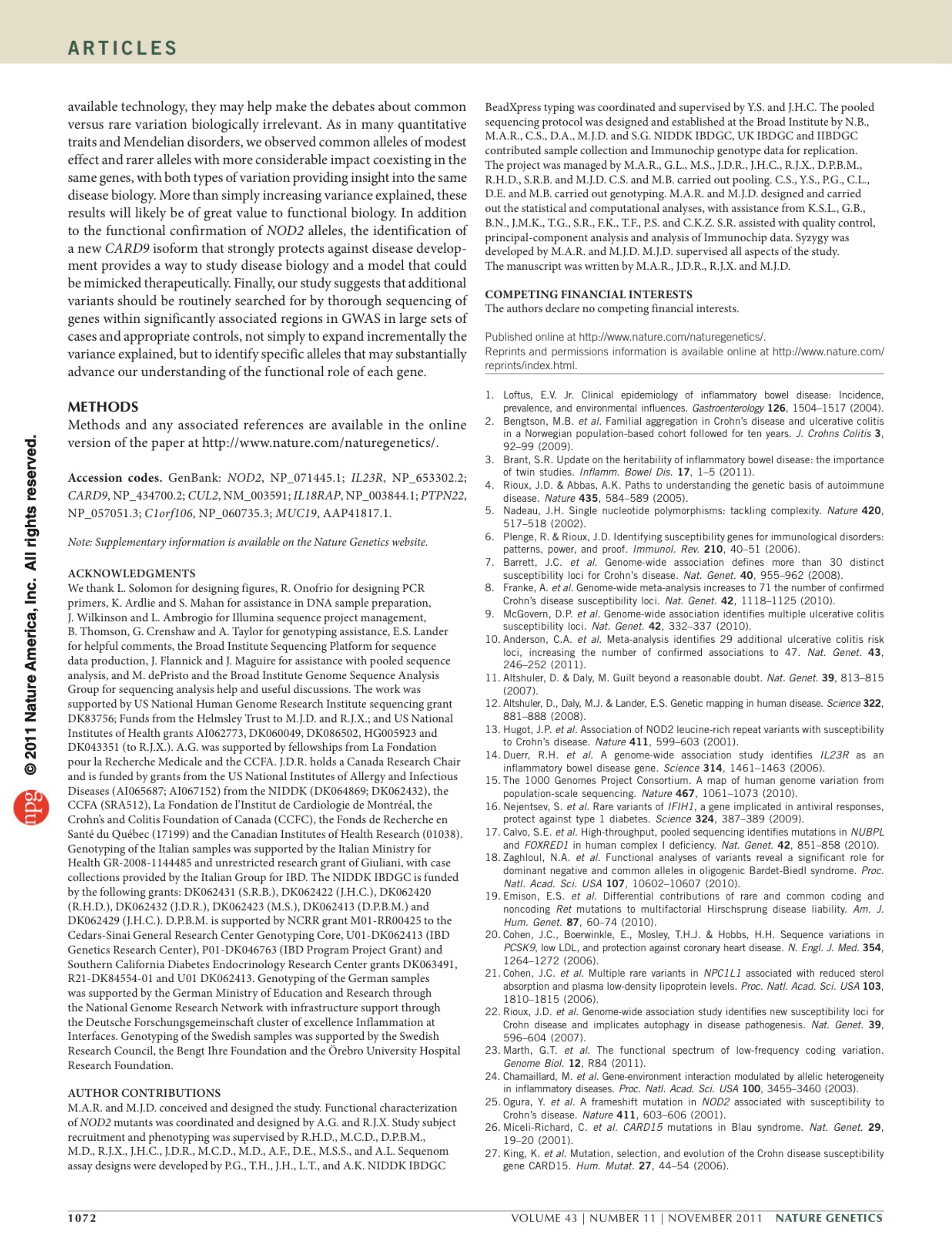
We estimated the fraction of additive genetic variance explained using the liability threshold model[40,41], which assumes an additive effect at each locus and shifts the mean of a normal distribution of disease liability for each genotype class. We assumed a prevalence of Crohn's disease of 4 per 1,000 and a total narrow-sense heritability of 50% (ref. 42). We estimate that the discovered rare and low-frequency variants associated with Crohn's disease in this study contribute another 1–2% genetic variance over all populations and 2–3% genetic variance to the Ashkenazi Jewish population (**Supplementary Table 7**).

DISCUSSION

Genome-wide association has been highly successful in IBD, with 99 confirmed associations providing new insight into disease biology. However, it is the ~75% of heritability yet to be explained that fuels most debate in human genetics. NGS offers potential insights into both the biology and the heritable component explained by GWAS results by completing the allelic spectrum of functional alleles in cases and controls, including rare variation.

Using a targeted pooled approach, we carried out an efficient and cost-effective scan for rare and low-frequency polymorphisms in genes from regions identified as relevant through GWAS. After extensive follow-up genotyping, we identified highly significant variants at *CARD9*, *NOD2*, *CUL2* and *IL18RAP* that contribute to risk independently of previously defined variants at these loci, and we showed the functionality of the newly implicated *NOD2* variants. We report additional protective variants at *IL23R* and identify nominally significant variants in *MUC19*, *PTPN22* and *C1orf106* more frequently than expected by chance.

The results of this experiment are relevant to ongoing debates in human genetics. Although we found little support for the hypothesis that common-variant associations are simply an indirect LD-driven by-product of higher penetrance rare alleles, additional independently acting low-frequency alleles in genes implicated by common-variant association are documented. In the case of the *CARD9* splice variant, this newly discovered allele explains more of the overall population variance in risk than does the common S12N variant (about 0.3% and 0.2%, respectively). If these observations become commonplace through



available technology, they may help make the debates about common versus rare variation biologically irrelevant. As in many quantitative traits and Mendelian disorders, we observed common alleles of modest effect and rarer alleles with more considerable impact coexisting in the same genes, with both types of variation providing insight into the same disease biology. More than simply increasing variance explained, these results will likely be of great value to functional biology. In addition to the functional confirmation of *NOD2* alleles, the identification of a new *CARD9* isoform that strongly protects against disease development provides a way to study disease biology and a model that could be mimicked therapeutically. Finally, our study suggests that additional variants should be routinely searched for by thorough sequencing of genes within significantly associated regions in GWAS in large sets of cases and appropriate controls, not simply to expand incrementally the variance explained, but to identify specific alleles that may substantially advance our understanding of the functional role of each gene.

METHODS

Methods and any associated references are available in the online version of the paper at http://www.nature.com/naturegenetics/.

Accession codes. GenBank: *NOD2*, NP_071445.1; *IL23R*, NP_653302.2; *CARD9*, NP_434700.2; *CUL2*, NM_003591; *IL18RAP*, NP_003844.1; *PTPN22*, NP_057051.3; *C1orf106*, NP_060735.3; *MUC19*, AAP41817.1.

Note: Supplementary information is available on the Nature Genetics website.

© 2011 Nature America, Inc. All rights reserved.

ACKNOWLEDGMENTS

We thank L. Solomon for designing figures, R. Onofrio for designing PCR primers, K. Ardlie and S. Mahan for assistance in DNA sample preparation, J. Wilkinson and L. Ambrogio for Illumina sequence project management, B. Thomson, G. Crenshaw and A. Taylor for genotyping assistance, E.S. Lander for helpful comments, the Broad Institute Sequencing Platform for sequence data production, J. Flannick and J. Maguire for assistance with pooled sequence analysis, and M. dePristo and the Broad Institute Genome Sequence Analysis Group for sequencing analysis help and useful discussions. The work was supported by US National Human Genome Research Institute sequencing grant DK83756; Funds from the Helmsley Trust to M.J.D. and R.J.X.; and US National Institutes of Health grants AI062773, DK060049, DK086502, HG005923 and DK043351 (to R.J.X.). A.G. was supported by fellowships from La Fondation pour la Recherche Medicale and the CCFA. J.D.R. holds a Canada Research Chair and is funded by grants from the US National Institutes of Allergy and Infectious Diseases (AI065687; AI067152) from the NIDDK (DK064869; DK062432), the CCFA (SRA512), La Fondation de l'Institut de Cardiologie de Montréal, the Crohn's and Colitis Foundation of Canada (CCFC), the Fonds de Recherche en Santé du Québec (17199) and the Canadian Institutes of Health Research (01038). Genotyping of the Italian samples was supported by the Italian Ministry for Health GR-2008-1144485 and unrestricted research grant of Giuliani, with case collections provided by the Italian Group for IBD. The NIDDK IBDGC is funded by the following grants: DK062431 (S.R.B.), DK062422 (J.H.C.), DK062420 (R.H.D.), DK062432 (J.D.R.), DK062423 (M.S.), DK062413 (D.P.B.M.) and DK062429 (J.H.C.). D.P.B.M. is supported by NCRR grant M01-RR00425 to the Cedars-Sinai General Research Center Genotyping Core, U01-DK062413 (IBD Genetics Research Center), P01-DK046763 (IBD Program Project Grant) and Southern California Diabetes Endocrinology Research Center grants DK063491, R21-DK84554-01 and U01 DK062413. Genotyping of the German samples was supported by the German Ministry of Education and Research through the National Genome Research Network with infrastructure support through the Deutsche Forschungsgemeinschaft cluster of excellence Inflammation at Interfaces. Genotyping of the Swedish samples was supported by the Swedish Research Council, the Bengt Ihre Foundation and the Örebro University Hospital Research Foundation.

AUTHOR CONTRIBUTIONS

M.A.R. and M.J.D. conceived and designed the study. Functional characterization of *NOD2* mutants was coordinated and designed by A.G. and R.J.X. Study subject recruitment and phenotyping was supervised by R.H.D., M.C.D., D.P.B.M., M.D., R.J.X., J.H.C., J.D.R., M.C.D., M.D., A.F., D.E., M.S.S., and A.L. Sequenom assay designs were developed by P.G., T.H., J.H., L.T., and A.K. NIDDK IBDGC BeadXpress typing was coordinated and supervised by Y.S. and J.H.C. The pooled sequencing protocol was designed and established at the Broad Institute by N.B., M.A.R., C.S., D.A., M.J.D. and S.G. NIDDK IBDGC, UK IBDGC and IIBDGC contributed sample collection and Immunochip genotype data for replication. The project was managed by M.A.R., G.L., M.S., J.D.R., J.H.C., R.J.X., D.P.B.M., R.H.D., S.R.B. and M.J.D. C.S. and M.B. carried out pooling. C.S., Y.S., P.G., C.L., D.E. and M.B. carried out genotyping. M.A.R. and M.J.D. designed and carried out the statistical and computational analyses, with assistance from K.S.L., G.B., B.N., J.M.K., T.G., S.R., F.K., T.F., P.S. and C.K.Z. S.R. assisted with quality control, principal-component analysis and analysis of Immunochip data. Syzygy was developed by M.A.R. and M.J.D. M.J.D. supervised all aspects of the study. The manuscript was written by M.A.R., J.D.R., R.J.X. and M.J.D.

COMPETING FINANCIAL INTERESTS

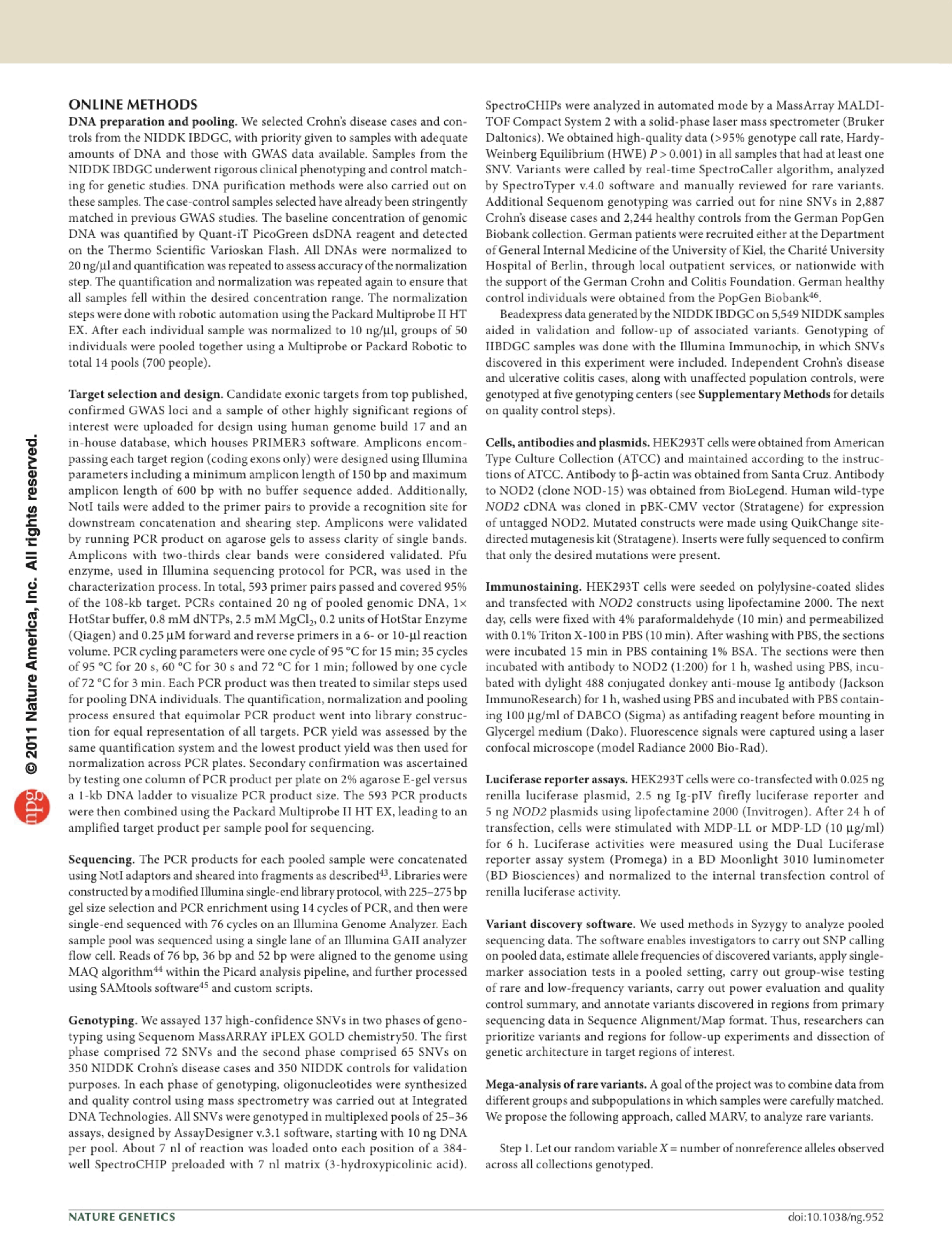
The authors declare no competing financial interests.

Published online at http://www.nature.com/naturegenetics/.
Reprints and permissions information is available online at http://www.nature.com/reprints/index.html.

1. Loftus, E.V. Jr. Clinical epidemiology of inflammatory bowel disease: Incidence, prevalence, and environmental influences. *Gastroenterology* **126**, 1504–1517 (2004).
2. Bengtson, M.B. *et al.* Familial aggregation in Crohn's disease and ulcerative colitis in a Norwegian population-based cohort followed for ten years. *J. Crohns Colitis* **3**, 92–99 (2009).
3. Brant, S.R. Update on the heritability of inflammatory bowel disease: the importance of twin studies. *Inflamm. Bowel Dis.* **17**, 1–5 (2011).
4. Rioux, J.D. & Abbas, A.K. Paths to understanding the genetic basis of autoimmune disease. *Nature* **435**, 584–589 (2005).
5. Nadeau, J.H. Single nucleotide polymorphisms: tackling complexity. *Nature* **420**, 517–518 (2002).
6. Plenge, R. & Rioux, J.D. Identifying susceptibility genes for immunological disorders: patterns, power, and proof. *Immunol. Rev.* **210**, 40–51 (2006).
7. Barrett, J.C. *et al.* Genome-wide association defines more than 30 distinct susceptibility loci for Crohn's disease. *Nat. Genet.* **40**, 955–962 (2008).
8. Franke, A. *et al.* Genome-wide meta-analysis increases to 71 the number of confirmed Crohn's disease susceptibility loci. *Nat. Genet.* **42**, 1118–1125 (2010).
9. McGovern, D.P. *et al.* Genome-wide association identifies multiple ulcerative colitis susceptibility loci. *Nat. Genet.* **42**, 332–337 (2010).
10. Anderson, C.A. *et al.* Meta-analysis identifies 29 additional ulcerative colitis risk loci, increasing the number of confirmed associations to 47. *Nat. Genet.* **43**, 246–252 (2011).
11. Altshuler, D. & Daly, M. Guilt beyond a reasonable doubt. *Nat. Genet.* **39**, 813–815 (2007).
12. Altshuler, D., Daly, M.J. & Lander, E.S. Genetic mapping in human disease. *Science* **322**, 881–888 (2008).
13. Hugot, J.P. *et al.* Association of NOD2 leucine-rich repeat variants with susceptibility to Crohn's disease. *Nature* **411**, 599–603 (2001).
14. Duerr, R.H. *et al.* A genome-wide association study identifies *IL23R* as an inflammatory bowel disease gene. *Science* **314**, 1461–1463 (2006).
15. The 1000 Genomes Project Consortium. A map of human genome variation from population-scale sequencing. *Nature* **467**, 1061–1073 (2010).
16. Nejentsev, S. *et al.* Rare variants of *IFIH1*, a gene implicated in antiviral responses, protect against type 1 diabetes. *Science* **324**, 387–389 (2009).
17. Calvo, S.E. *et al.* High-throughput, pooled sequencing identifies mutations in *NUBPL* and *FOXRED1* in human complex I deficiency. *Nat. Genet.* **42**, 851–858 (2010).
18. Zaghloul, N.A. *et al.* Functional analyses of variants reveal a significant role for dominant negative and common alleles in oligogenic Bardet-Biedl syndrome. *Proc. Natl. Acad. Sci. USA* **107**, 10602–10607 (2010).
19. Emison, E.S. *et al.* Differential contributions of rare and common coding and noncoding *Ret* mutations to multifactorial Hirschsprung disease liability. *Am. J. Hum. Genet.* **87**, 60–74 (2010).
20. Cohen, J.C., Boerwinkle, E., Mosley, T.H.J. & Hobbs, H.H. Sequence variations in *PCSK9*, low LDL, and protection against coronary heart disease. *N. Engl. J. Med.* **354**, 1264–1272 (2006).
21. Cohen, J.C. *et al.* Multiple rare variants in *NPC1L1* associated with reduced sterol absorption and plasma low-density lipoprotein levels. *Proc. Natl. Acad. Sci. USA* **103**, 1810–1815 (2006).
22. Rioux, J.D. *et al.* Genome-wide association study identifies new susceptibility loci for Crohn disease and implicates autophagy in disease pathogenesis. *Nat. Genet.* **39**, 596–604 (2007).
23. Marth, G.T. *et al.* The functional spectrum of low-frequency coding variation. *Genome Biol.* **12**, R84 (2011).
24. Chamaillard, M. *et al.* Gene-environment interaction modulated by allelic heterogeneity in inflammatory diseases. *Proc. Natl. Acad. Sci. USA* **100**, 3455–3460 (2003).
25. Ogura, Y. *et al.* A frameshift mutation in *NOD2* associated with susceptibility to Crohn's disease. *Nature* **411**, 603–606 (2001).
26. Miceli-Richard, C. *et al.* *CARD15* mutations in Blau syndrome. *Nat. Genet.* **29**, 19–20 (2001).
27. King, K. *et al.* Mutation, selection, and evolution of the Crohn disease susceptibility gene CARD15. *Hum. Mutat.* **27**, 44–54 (2006).

28. Barnich, N., Aguirre, J.E., Reinecker, H.C., Xavier, R.J. & Podolsky, D.K. Membrane recruitment of NOD2 in intenstinal epithelial cells is essential for nuclear factor-κB activation in muramyl dipeptide recognition. *J. Cell Biol.* **170**, 21–26 (2005).
29. Tanabe, T. *et al.* Regulatory regions and critical residues of NOD2 involved in muramyl dipeptide recognition. *EMBO J.* **23**, 1587–1597 (2004).
30. Roth, M.P. *et al.* Geographic origins of Jewish patients with inflammatory bowel disease. *Gastroenterology* **97**, 900–904 (1989).
31. Tukel, T. *et al.* Crohn disease: frequency and nature of *CARD15* mutations in Ashkenazi and Sephardi/Oriental Jewish families. *Am. J. Hum. Genet.* **74**, 623–636 (2004).
32. Momozawa, Y. *et al.* Resequencing of positional candidates identifies low frequency *IL23R* coding variants protecting against inflammatory bowel disease. *Nat. Genet.* **43**, 43–47 (2011).
33. Di Meglio, P. *et al.* The IL23R R381Q gene variant protects against immune-mediated diseases by impairing IL-23-induced T_H17 effector response in humans. *PLoS ONE* **6**, e17160 (2011).
34. Ghoreschi, K. *et al.* Generation of pathogenic T(H)17 cells in the absence of TGF-β signalling. *Nature* **467**, 967–971 (2010).
35. Festen, E.A. *et al.* A meta-analysis of genome-wide association scans identifies *IL18RAP, PTPN2, TAGAP*, and *PUS10* as shared risk loci for Crohn's disease and celiac disease. *PLoS Genet.* **7**, e1001283 (2011).
36. Behrends, C., Sowa, M.E., Gygi, S.P. & Harper, J.W. Network organize of the human autophagy system. *Nature* **466**, 68–76 (2010).
37. Barrett, J.C. *et al.* Genome-wide association study and meta-analysis find that over 40 loci affect risk of type 1 diabetes. *Nat. Genet.* **41**, 703–707 (2009).
38. Stahl, E.A. *et al.* Genome-wide association study meta-analysis identifies seven new rheumatoid arthritis risk loci. *Nat. Genet.* **42**, 508–514 (2010).
39. Jin, Y. *et al.* Variant of *TYR* and autoimmunity susceptibility loci in generalized vitiligo. *N. Engl. J. Med.* **362**, 1686–1697 (2010).
40. Pearson, K. Mathematical contributions to the theory of evolution VIII: On the inheritance of characters not capable of exact quantitative measurement. *Phil. Trans. R. Soc. Lond. A* **195**, 79–150 (1900).
41. Fisher, R.A. The correlation between relatives on the supposition of Mendelian inheritance. *Trans. R. Soc. Edinb.* **52**, 399–433 (1918).
42. Ahmad, T., Satsangi, J., McGovern, D., Bunce, M. & Jewell, D.P. The genetics of inflammatory bowel disease. *Aliment. Pharmacol. Ther.* **15**, 731–748 (2001).

© 2011 Nature America, Inc. All rights reserved.



© 2011 Nature America, Inc. All rights reserved.

ONLINE METHODS

DNA preparation and pooling. We selected Crohn's disease cases and controls from the NIDDK IBDGC, with priority given to samples with adequate amounts of DNA and those with GWAS data available. Samples from the NIDDK IBDGC underwent rigorous clinical phenotyping and control matching for genetic studies. DNA purification methods were also carried out on these samples. The case-control samples selected have already been stringently matched in previous GWAS studies. The baseline concentration of genomic DNA was quantified by Quant-iT PicoGreen dsDNA reagent and detected on the Thermo Scientific Varioskan Flash. All DNAs were normalized to 20 ng/μl and quantification was repeated to assess accuracy of the normalization step. The quantification and normalization was repeated again to ensure that all samples fell within the desired concentration range. The normalization steps were done with robotic automation using the Packard Multiprobe II HT EX. After each individual sample was normalized to 10 ng/μl, groups of 50 individuals were pooled together using a Multiprobe or Packard Robotic to total 14 pools (700 people).

Target selection and design. Candidate exonic targets from top published, confirmed GWAS loci and a sample of other highly significant regions of interest were uploaded for design using human genome build 17 and an in-house database, which houses PRIMER3 software. Amplicons encompassing each target region (coding exons only) were designed using Illumina parameters including a minimum amplicon length of 150 bp and maximum amplicon length of 600 bp with no buffer sequence added. Additionally, NotI tails were added to the primer pairs to provide a recognition site for downstream concatenation and shearing step. Amplicons were validated by running PCR product on agarose gels to assess clarity of single bands. Amplicons with two-thirds clear bands were considered validated. Pfu enzyme, used in Illumina sequencing protocol for PCR, was used in the characterization process. In total, 593 primer pairs passed and covered 95% of the 108-kb target. PCRs contained 20 ng of pooled genomic DNA, 1× HotStar buffer, 0.8 mM dNTPs, 2.5 mM $MgCl_2$, 0.2 units of HotStar Enzyme (Qiagen) and 0.25 μM forward and reverse primers in a 6- or 10-μl reaction volume. PCR cycling parameters were one cycle of 95 °C for 15 min; 35 cycles of 95 °C for 20 s, 60 °C for 30 s and 72 °C for 1 min; followed by one cycle of 72 °C for 3 min. Each PCR product was then treated to similar steps used for pooling DNA individuals. The quantification, normalization and pooling process ensured that equimolar PCR product went into library construction for equal representation of all targets. PCR yield was assessed by the same quantification system and the lowest product yield was then used for normalization across PCR plates. Secondary confirmation was ascertained by testing one column of PCR product per plate on 2% agarose E-gel versus a 1-kb DNA ladder to visualize PCR product size. The 593 PCR products were then combined using the Packard Multiprobe II HT EX, leading to an amplified target product per sample pool for sequencing.

Sequencing. The PCR products for each pooled sample were concatenated using NotI adaptors and sheared into fragments as described[43]. Libraries were constructed by a modified Illumina single-end library protocol, with 225–275 bp gel size selection and PCR enrichment using 14 cycles of PCR, and then were single-end sequenced with 76 cycles on an Illumina Genome Analyzer. Each sample pool was sequenced using a single lane of an Illumina GAII analyzer flow cell. Reads of 76 bp, 36 bp and 52 bp were aligned to the genome using MAQ algorithm[44] within the Picard analysis pipeline, and further processed using SAMtools software[45] and custom scripts.

Genotyping. We assayed 137 high-confidence SNVs in two phases of genotyping using Sequenom MassARRAY iPLEX GOLD chemistry50. The first phase comprised 72 SNVs and the second phase comprised 65 SNVs on 350 NIDDK Crohn's disease cases and 350 NIDDK controls for validation purposes. In each phase of genotyping, oligonucleotides were synthesized and quality control using mass spectrometry was carried out at Integrated DNA Technologies. All SNVs were genotyped in multiplexed pools of 25–36 assays, designed by AssayDesigner v.3.1 software, starting with 10 ng DNA per pool. About 7 nl of reaction was loaded onto each position of a 384-well SpectroCHIP preloaded with 7 nl matrix (3-hydroxypicolinic acid). SpectroCHIPs were analyzed in automated mode by a MassArray MALDI-TOF Compact System 2 with a solid-phase laser mass spectrometer (Bruker Daltonics). We obtained high-quality data (>95% genotype call rate, Hardy-Weinberg Equilibrium (HWE) $P > 0.001$) in all samples that had at least one SNV. Variants were called by real-time SpectroCaller algorithm, analyzed by SpectroTyper v.4.0 software and manually reviewed for rare variants. Additional Sequenom genotyping was carried out for nine SNVs in 2,887 Crohn's disease cases and 2,244 healthy controls from the German PopGen Biobank collection. German patients were recruited either at the Department of General Internal Medicine of the University of Kiel, the Charité University Hospital of Berlin, through local outpatient services, or nationwide with the support of the German Crohn and Colitis Foundation. German healthy control individuals were obtained from the PopGen Biobank[46].

Beadexpress data generated by the NIDDK IBDGC on 5,549 NIDDK samples aided in validation and follow-up of associated variants. Genotyping of IIBDGC samples was done with the Illumina Immunochip, in which SNVs discovered in this experiment were included. Independent Crohn's disease and ulcerative colitis cases, along with unaffected population controls, were genotyped at five genotyping centers (see **Supplementary Methods** for details on quality control steps).

Cells, antibodies and plasmids. HEK293T cells were obtained from American Type Culture Collection (ATCC) and maintained according to the instructions of ATCC. Antibody to β-actin was obtained from Santa Cruz. Antibody to NOD2 (clone NOD-15) was obtained from BioLegend. Human wild-type *NOD2* cDNA was cloned in pBK-CMV vector (Stratagene) for expression of untagged NOD2. Mutated constructs were made using QuikChange site-directed mutagenesis kit (Stratagene). Inserts were fully sequenced to confirm that only the desired mutations were present.

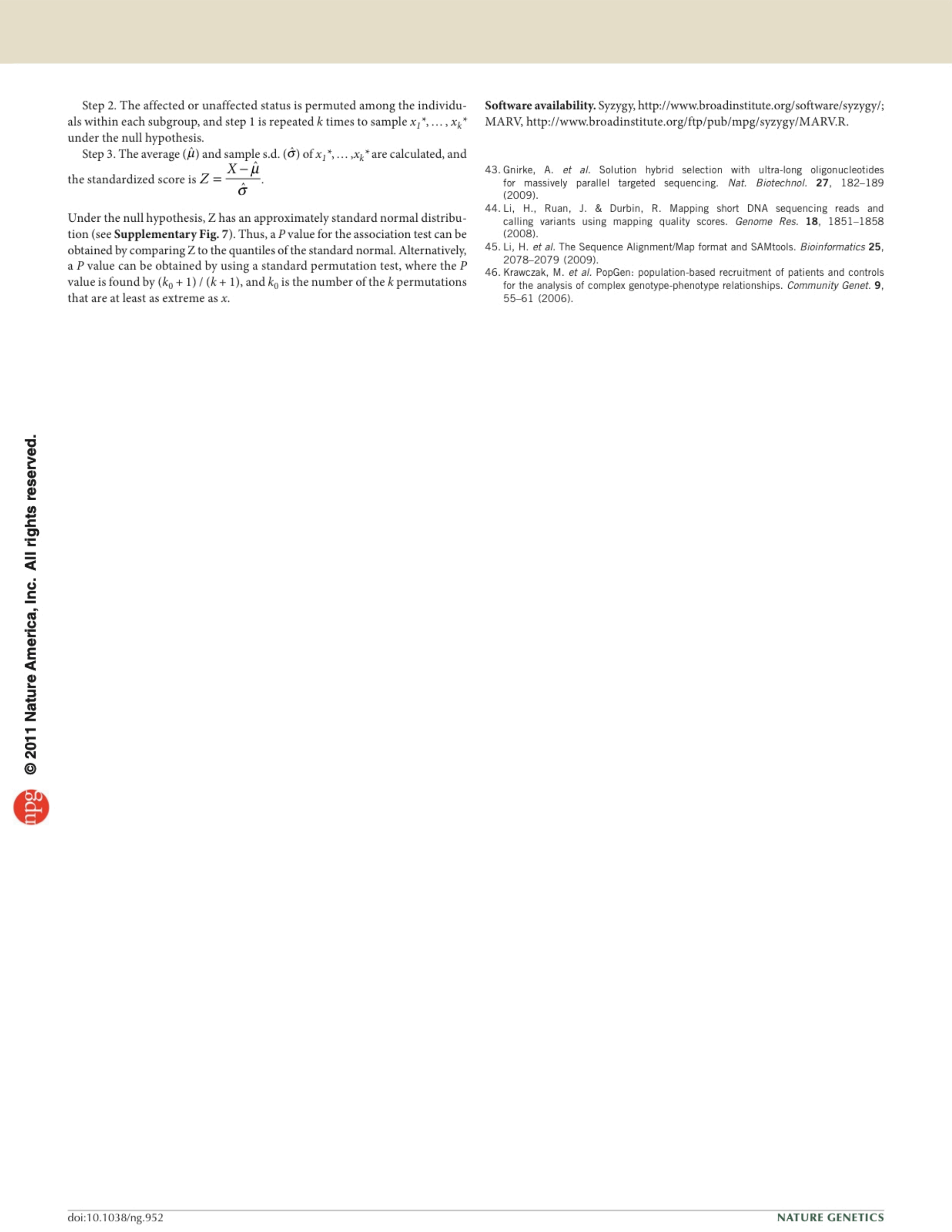
Immunostaining. HEK293T cells were seeded on polylysine-coated slides and transfected with *NOD2* constructs using lipofectamine 2000. The next day, cells were fixed with 4% paraformaldehyde (10 min) and permeabilized with 0.1% Triton X-100 in PBS (10 min). After washing with PBS, the sections were incubated 15 min in PBS containing 1% BSA. The sections were then incubated with antibody to NOD2 (1:200) for 1 h, washed using PBS, incubated with dylight 488 conjugated donkey anti-mouse Ig antibody (Jackson ImmunoResearch) for 1 h, washed using PBS and incubated with PBS containing 100 μg/ml of DABCO (Sigma) as antifading reagent before mounting in Glycergel medium (Dako). Fluorescence signals were captured using a laser confocal microscope (model Radiance 2000 Bio-Rad).

Luciferase reporter assays. HEK293T cells were co-transfected with 0.025 ng renilla luciferase plasmid, 2.5 ng Ig-pIV firefly luciferase reporter and 5 ng *NOD2* plasmids using lipofectamine 2000 (Invitrogen). After 24 h of transfection, cells were stimulated with MDP-LL or MDP-LD (10 μg/ml) for 6 h. Luciferase activities were measured using the Dual Luciferase reporter assay system (Promega) in a BD Moonlight 3010 luminometer (BD Biosciences) and normalized to the internal transfection control of renilla luciferase activity.

Variant discovery software. We used methods in Syzygy to analyze pooled sequencing data. The software enables investigators to carry out SNP calling on pooled data, estimate allele frequencies of discovered variants, apply single-marker association tests in a pooled setting, carry out group-wise testing of rare and low-frequency variants, carry out power evaluation and quality control summary, and annotate variants discovered in regions from primary sequencing data in Sequence Alignment/Map format. Thus, researchers can prioritize variants and regions for follow-up experiments and dissection of genetic architecture in target regions of interest.

Mega-analysis of rare variants. A goal of the project was to combine data from different groups and subpopulations in which samples were carefully matched. We propose the following approach, called MARV, to analyze rare variants.

Step 1. Let our random variable X = number of nonreference alleles observed across all collections genotyped.



Step 2. The affected or unaffected status is permuted among the individuals within each subgroup, and step 1 is repeated k times to sample $x_1^*, \ldots, x_k^*$ under the null hypothesis.

Step 3. The average ($\hat{\mu}$) and sample s.d. ($\hat{\sigma}$) of $x_1^*, \ldots, x_k^*$ are calculated, and the standardized score is $Z = \dfrac{X - \hat{\mu}}{\hat{\sigma}}$.

Under the null hypothesis, Z has an approximately standard normal distribution (see **Supplementary Fig. 7**). Thus, a *P* value for the association test can be obtained by comparing Z to the quantiles of the standard normal. Alternatively, a *P* value can be obtained by using a standard permutation test, where the *P* value is found by $(k_0 + 1) / (k + 1)$, and k_0 is the number of the k permutations that are at least as extreme as x.

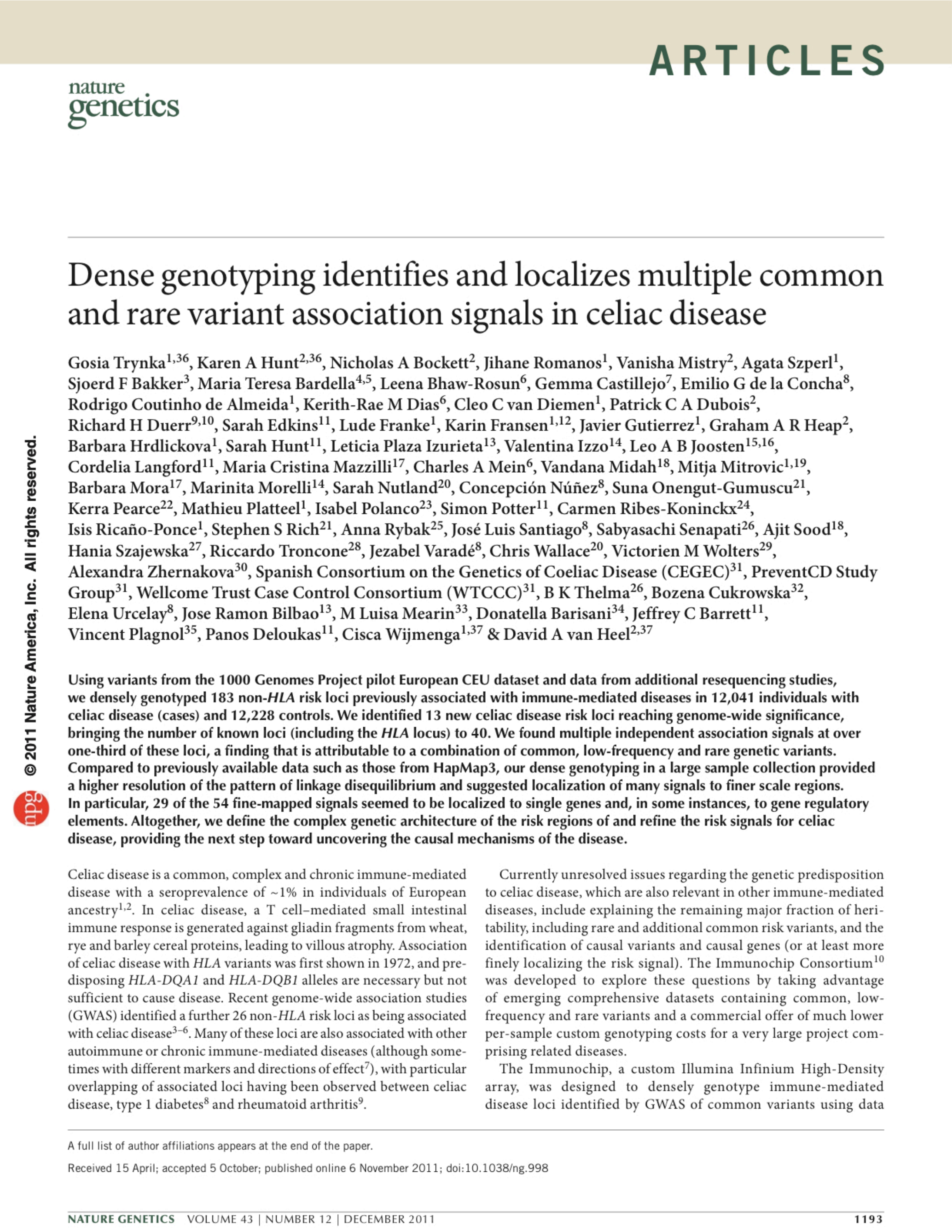
Software availability. Syzygy, http://www.broadinstitute.org/software/syzygy/; MARV, http://www.broadinstitute.org/ftp/pub/mpg/syzygy/MARV.R.

43. Gnirke, A. *et al.* Solution hybrid selection with ultra-long oligonucleotides for massively parallel targeted sequencing. *Nat. Biotechnol.* **27**, 182–189 (2009).
44. Li, H., Ruan, J. & Durbin, R. Mapping short DNA sequencing reads and calling variants using mapping quality scores. *Genome Res.* **18**, 1851–1858 (2008).
45. Li, H. *et al.* The Sequence Alignment/Map format and SAMtools. *Bioinformatics* **25**, 2078–2079 (2009).
46. Krawczak, M. *et al.* PopGen: population-based recruitment of patients and controls for the analysis of complex genotype-phenotype relationships. *Community Genet.* **9**, 55–61 (2006).

© 2011 Nature America, Inc. All rights reserved.

doi:10.1038/ng.952

6.7.2 Izvirni znanstveni članek 7



ARTICLES

nature genetics

© 2011 Nature America, Inc. All rights reserved.

Dense genotyping identifies and localizes multiple common and rare variant association signals in celiac disease

Gosia Trynka[1,36], Karen A Hunt[2,36], Nicholas A Bockett[2], Jihane Romanos[1], Vanisha Mistry[2], Agata Szperl[1], Sjoerd F Bakker[3], Maria Teresa Bardella[4,5], Leena Bhaw-Rosun[6], Gemma Castillejo[7], Emilio G de la Concha[8], Rodrigo Coutinho de Almeida[1], Kerith-Rae M Dias[6], Cleo C van Diemen[1], Patrick C A Dubois[2], Richard H Duerr[9,10], Sarah Edkins[11], Lude Franke[1], Karin Fransen[1,12], Javier Gutierrez[1], Graham A R Heap[2], Barbara Hrdlickova[1], Sarah Hunt[11], Leticia Plaza Izurieta[13], Valentina Izzo[14], Leo A B Joosten[15,16], Cordelia Langford[11], Maria Cristina Mazzilli[17], Charles A Mein[6], Vandana Midah[18], Mitja Mitrovic[1,19], Barbara Mora[17], Marinita Morelli[14], Sarah Nutland[20], Concepción Núñez[8], Suna Onengut-Gumuscu[21], Kerra Pearce[22], Mathieu Platteel[1], Isabel Polanco[23], Simon Potter[11], Carmen Ribes-Koninckx[24], Isis Ricaño-Ponce[1], Stephen S Rich[21], Anna Rybak[25], José Luis Santiago[8], Sabyasachi Senapati[26], Ajit Sood[18], Hania Szajewska[27], Riccardo Troncone[28], Jezabel Varadé[8], Chris Wallace[20], Victorien M Wolters[29], Alexandra Zhernakova[30], Spanish Consortium on the Genetics of Coeliac Disease (CEGEC)[31], PreventCD Study Group[31], Wellcome Trust Case Control Consortium (WTCCC)[31], B K Thelma[26], Bozena Cukrowska[32], Elena Urcelay[8], Jose Ramon Bilbao[13], M Luisa Mearin[33], Donatella Barisani[34], Jeffrey C Barrett[11], Vincent Plagnol[35], Panos Deloukas[11], Cisca Wijmenga[1,37] & David A van Heel[2,37]

Using variants from the 1000 Genomes Project pilot European CEU dataset and data from additional resequencing studies, we densely genotyped 183 non-*HLA* risk loci previously associated with immune-mediated diseases in 12,041 individuals with celiac disease (cases) and 12,228 controls. We identified 13 new celiac disease risk loci reaching genome-wide significance, bringing the number of known loci (including the *HLA* locus) to 40. We found multiple independent association signals at over one-third of these loci, a finding that is attributable to a combination of common, low-frequency and rare genetic variants. Compared to previously available data such as those from HapMap3, our dense genotyping in a large sample collection provided a higher resolution of the pattern of linkage disequilibrium and suggested localization of many signals to finer scale regions. In particular, 29 of the 54 fine-mapped signals seemed to be localized to single genes and, in some instances, to gene regulatory elements. Altogether, we define the complex genetic architecture of the risk regions of and refine the risk signals for celiac disease, providing the next step toward uncovering the causal mechanisms of the disease.

Celiac disease is a common, complex and chronic immune-mediated disease with a seroprevalence of ~1% in individuals of European ancestry[1,2]. In celiac disease, a T cell–mediated small intestinal immune response is generated against gliadin fragments from wheat, rye and barley cereal proteins, leading to villous atrophy. Association of celiac disease with *HLA* variants was first shown in 1972, and predisposing *HLA-DQA1* and *HLA-DQB1* alleles are necessary but not sufficient to cause disease. Recent genome-wide association studies (GWAS) identified a further 26 non-*HLA* risk loci as being associated with celiac disease[3–6]. Many of these loci are also associated with other autoimmune or chronic immune-mediated diseases (although sometimes with different markers and directions of effect[7]), with particular overlapping of associated loci having been observed between celiac disease, type 1 diabetes[8] and rheumatoid arthritis[9].

Currently unresolved issues regarding the genetic predisposition to celiac disease, which are also relevant in other immune-mediated diseases, include explaining the remaining major fraction of heritability, including rare and additional common risk variants, and the identification of causal variants and causal genes (or at least more finely localizing the risk signal). The Immunochip Consortium[10] was developed to explore these questions by taking advantage of emerging comprehensive datasets containing common, low-frequency and rare variants and a commercial offer of much lower per-sample custom genotyping costs for a very large project comprising related diseases.

The Immunochip, a custom Illumina Infinium High-Density array, was designed to densely genotype immune-mediated disease loci identified by GWAS of common variants using data

A full list of author affiliations appears at the end of the paper.

Received 15 April; accepted 5 October; published online 6 November 2011; doi:10.1038/ng.998

© 2011 Nature America, Inc. All rights reserved.

Table 1 Sample collections

Population sample	Cases[a]	Controls
UK	7,728	8,274[b]
The Netherlands	1,123	1,147
Poland	505	533
Spain—CEGEC[c]	545	308
Spain—Madrid[c]	537	320
Italy—Rome, Milan and Naples	1,374	1,255
India—Punjab	229	391
Total	12,041	12,228

The collections from the UK, The Netherlands, Poland, Spain (Madrid) and Italy contained essentially the same sample set as our 2010 GWAS of celiac disease[5] but had substantial additional samples from the UK and The Netherlands and excluded amplified DNA samples from the Spanish collections. The Indian collection was not previously studied. Our 2010 GWAS contained several collections not studied here.

[a]Cases are defined as individuals with Celiac disease. [b]This data includes 5,430 UK 1958 Birth Cohort participants and 2,844 UK Blood Services Common Controls. [c]We considered the two Spanish population samples separately because the samples were genotyped in different laboratories.

from the 1000 Genomes Project and any other available disease-specific resequencing data. The 1000 Genomes Project pilot CEU low-coverage whole-genome–sequencing dataset captures 95% of the variants of minor allele frequency (MAF) = 0.05, and although it is underpowered to comprehensively detect variants of rarer allele frequency, the dataset still identifies 60% of variants of MAF = 0.02 and 30% of variants of MAF = 0.01 (ref. 11). The Immunochip Consortium selected 186 distinct loci containing markers reaching genome-wide significance ($P < 5 \times 10^{-8}$) from 12 diseases (autoimmune thyroid disease, ankylosing spondylitis, Crohn's disease, celiac disease, IgA deficiency, multiple sclerosis, primary biliary cirrhosis, psoriasis, rheumatoid arthritis, systemic lupus erythematosus, type 1 diabetes and ulcerative colitis). We submitted all sample variants from the 1000 Genomes Project low-coverage pilot CEU population[11] (September 2009 release) that were in 0.1-cM (HapMap3 CEU) recombination blocks around each GWAS region lead marker for array design. We did not apply any filtering on correlated variants (linkage disequilibrium (LD)). Further case and control regional resequencing data were submitted by several groups (Online Methods and **Supplementary Note**), as well as a small amount of investigator-specific undisclosed content, including GWAS results of intermediate significance.

Most GWAS were performed using common SNPs (typically with MAF > 5%) further selected for low inter-marker correlation and/or even genomic spacing. In contrast to GWAS, the Immunochip Consortium represents an opportunity to in depth and comprehensively dissect the architecture of both rare and common genetic variation at immunobiologically relevant genomic regions in human diseases. Because of the presence of the majority of the polymorphic genetic variants from the 1000 Genomes Project pilot CEU dataset (as well as additional resequencing at some loci) in our final Immunochip dataset, the true causal variants at many risk loci may have been directly genotyped and analyzed.

RESULTS

Overview of the study design

We submitted a total of 207,728 variants for Immunochip assay design, and 196,524 variants passed manufacturing quality control at Illumina. After extensive and stringent data quality control (Online Methods), we analyzed a near-complete dataset (overall, there were only 0.008% missing genotype calls) comprising 12,041 cases with celiac disease, 12,228 controls (from seven geographic regions; **Table 1**) and 139,553 polymorphic (defined here as at least two observed genotype groups) markers. We assayed 634 biallelic SNPs in duplicate; at these SNPs, we observed 189 of the 15,384,884 (0.0012%) genotype calls to be discordant. Considering the intended 207,728 variants submitted for design and an observed ~9.1% non-polymorphic rate in our data after quality control filtering, we estimated that we had high quality genotype data on ~74% of the complete set of the 1000 Genomes Project pilot CEU true polymorphic variants at the fine-mapped regions.

We observed that 36 of the 183 non-*HLA* immune-mediated disease loci selected for dense 1000-Genomes–based genotyping using the Immunochip reached genome-wide significance ($P < 5 \times 10^{-8}$) for celiac disease in either the current study or in our previous GWAS[5] (the summary association statistics for all markers are available in T1DBase (see URLs)). All variants reaching genome-wide significance were common (MAF > 5%). We also observed marked enrichment for celiac disease association signals of intermediate significance (for example, rs6691768, at the *NFIA* locus, $P = 5.3 \times 10^{-8}$) at a proportion of the remaining 147 densely genotyped non-celiac autoimmune disease regions (**Supplementary Fig. 1**). Variants from three densely genotyped regions selected on Immunochip for a non–immune-mediated trait (bipolar disorder) showed no excess of association signals (**Supplementary Fig. 1**).

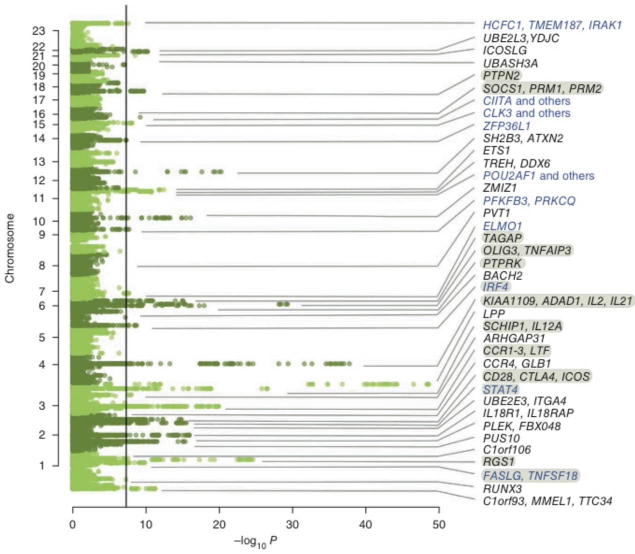


Figure 1 Manhattan plot of association statistics for previously known and newly discovered celiac disease risk loci. Newly discovered loci are indicated in blue; loci with multiple signals are shown in a gray highlighted box. The significance threshold used was $P = 5 \times 10^{-8}$.

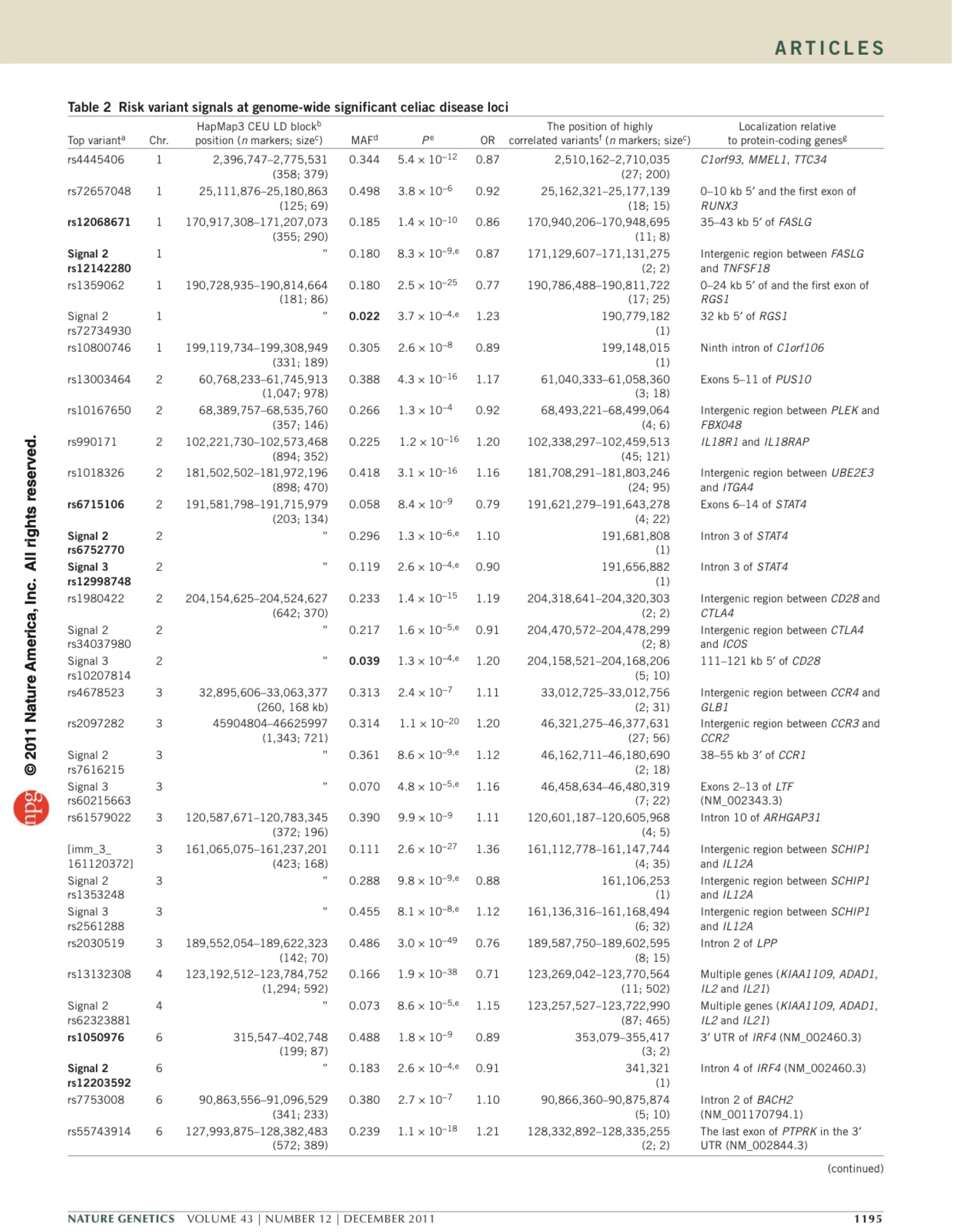
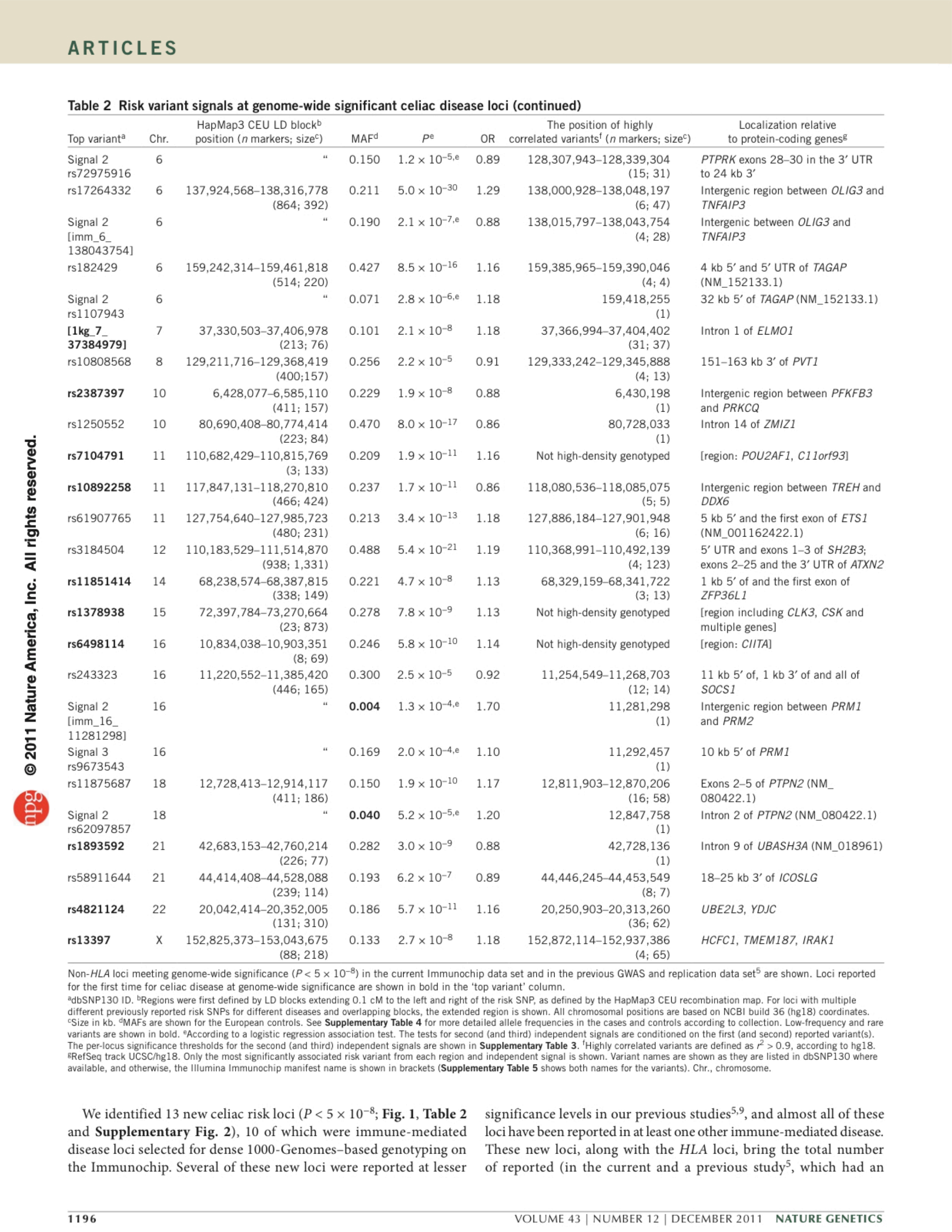


Table 2 Risk variant signals at genome-wide significant celiac disease loci

Top variant[a]	Chr.	HapMap3 CEU LD block[b] position (*n* markers; size[c])	MAF[d]	*P*[e]	OR	The position of highly correlated variants[f] (*n* markers; size[c])	Localization relative to protein-coding genes[g]
rs4445406	1	2,396,747–2,775,531 (358; 379)	0.344	5.4×10^{-12}	0.87	2,510,162–2,710,035 (27; 200)	*C1orf93*, *MMEL1*, *TTC34*
rs72657048	1	25,111,876–25,180,863 (125; 69)	0.498	3.8×10^{-6}	0.92	25,162,321–25,177,139 (18; 15)	0–10 kb 5′ and the first exon of *RUNX3*
rs12068671	1	170,917,308–171,207,073 (355; 290)	0.185	1.4×10^{-10}	0.86	170,940,206–170,948,695 (11; 8)	35–43 kb 5′ of *FASLG*
Signal 2 rs12142280	1	"	0.180	8.3×10^{-9}[e]	0.87	171,129,607–171,131,275 (2; 2)	Intergenic region between *FASLG* and *TNFSF18*
rs1359062	1	190,728,935–190,814,664 (181; 86)	0.180	2.5×10^{-25}	0.77	190,786,488–190,811,722 (17; 25)	0–24 kb 5′ of and the first exon of *RGS1*
Signal 2 rs72734930	1	"	**0.022**	3.7×10^{-4}[e]	1.23	190,779,182 (1)	32 kb 5′ of *RGS1*
rs10800746	1	199,119,734–199,308,949 (331; 189)	0.305	2.6×10^{-8}	0.89	199,148,015 (1)	Ninth intron of *C1orf106*
rs13003464	2	60,768,233–61,745,913 (1,047; 978)	0.388	4.3×10^{-16}	1.17	61,040,333–61,058,360 (3; 18)	Exons 5–11 of *PUS10*
rs10167650	2	68,389,757–68,535,760 (357; 146)	0.266	1.3×10^{-4}	0.92	68,493,221–68,499,064 (4; 6)	Intergenic region between *PLEK* and *FBX048*
rs990171	2	102,221,730–102,573,468 (894; 352)	0.225	1.2×10^{-16}	1.20	102,338,297–102,459,513 (45; 121)	*IL18R1* and *IL18RAP*
rs1018326	2	181,502,502–181,972,196 (898; 470)	0.418	3.1×10^{-16}	1.16	181,708,291–181,803,246 (24; 95)	Intergenic region between *UBE2E3* and *ITGA4*
rs6715106	2	191,581,798–191,715,979 (203; 134)	0.058	8.4×10^{-9}	0.79	191,621,279–191,643,278 (4; 22)	Exons 6–14 of *STAT4*
Signal 2 rs6752770	2	"	0.296	1.3×10^{-6}[e]	1.10	191,681,808 (1)	Intron 3 of *STAT4*
Signal 3 rs12998748	2	"	0.119	2.6×10^{-4}[e]	0.90	191,656,882 (1)	Intron 3 of *STAT4*
rs1980422	2	204,154,625–204,524,627 (642; 370)	0.233	1.4×10^{-15}	1.19	204,318,641–204,320,303 (2; 2)	Intergenic region between *CD28* and *CTLA4*
Signal 2 rs34037980	2	"	0.217	1.6×10^{-5}[e]	0.91	204,470,572–204,478,299 (2; 8)	Intergenic region between *CTLA4* and *ICOS*
Signal 3 rs10207814	2	"	**0.039**	1.3×10^{-4}[e]	1.20	204,158,521–204,168,206 (5; 10)	111–121 kb 5′ of *CD28*
rs4678523	3	32,895,606–33,063,377 (260, 168 kb)	0.313	2.4×10^{-7}	1.11	33,012,725–33,012,756 (2; 31)	Intergenic region between *CCR4* and *GLB1*
rs2097282	3	45904804–46625997 (1,343; 721)	0.314	1.1×10^{-20}	1.20	46,321,275–46,377,631 (27; 56)	Intergenic region between *CCR3* and *CCR2*
Signal 2 rs7616215	3	"	0.361	8.6×10^{-9}[e]	1.12	46,162,711–46,180,690 (2; 18)	38–55 kb 3′ of *CCR1*
Signal 3 rs60215663	3	"	0.070	4.8×10^{-5}[e]	1.16	46,458,634–46,480,319 (7; 22)	Exons 2–13 of *LTF* (NM_002343.3)
rs61579022	3	120,587,671–120,783,345 (372; 196)	0.390	9.9×10^{-9}	1.11	120,601,187–120,605,968 (4; 5)	Intron 10 of *ARHGAP31*
[imm_3_ 161120372]	3	161,065,075–161,237,201 (423; 168)	0.111	2.6×10^{-27}	1.36	161,112,778–161,147,744 (4; 35)	Intergenic region between *SCHIP1* and *IL12A*
Signal 2 rs1353248	3	"	0.288	9.8×10^{-9}[e]	0.88	161,106,253 (1)	Intergenic region between *SCHIP1* and *IL12A*
Signal 3 rs2561288	3	"	0.455	8.1×10^{-8}[e]	1.12	161,136,316–161,168,494 (6; 32)	Intergenic region between *SCHIP1* and *IL12A*
rs2030519	3	189,552,054–189,622,323 (142; 70)	0.486	3.0×10^{-49}	0.76	189,587,750–189,602,595 (8; 15)	Intron 2 of *LPP*
rs13132308	4	123,192,512–123,784,752 (1,294; 592)	0.166	1.9×10^{-38}	0.71	123,269,042–123,770,564 (11; 502)	Multiple genes (*KIAA1109*, *ADAD1*, *IL2* and *IL21*)
Signal 2 rs62323881	4	"	0.073	8.6×10^{-5}[e]	1.15	123,257,527–123,722,990 (87; 465)	Multiple genes (*KIAA1109*, *ADAD1*, *IL2* and *IL21*)
rs1050976	6	315,547–402,748 (199; 87)	0.488	1.8×10^{-9}	0.89	353,079–355,417 (3; 2)	3′ UTR of *IRF4* (NM_002460.3)
Signal 2 rs12203592	6	"	0.183	2.6×10^{-4}[e]	0.91	341,321 (1)	Intron 4 of *IRF4* (NM_002460.3)
rs7753008	6	90,863,556–91,096,529 (341; 233)	0.380	2.7×10^{-7}	1.10	90,866,360–90,875,874 (5; 10)	Intron 2 of *BACH2* (NM_001170794.1)
rs55743914	6	127,993,875–128,382,483 (572; 389)	0.239	1.1×10^{-18}	1.21	128,332,892–128,335,255 (2; 2)	The last exon of *PTPRK* in the 3′ UTR (NM_002844.3)

(continued)

© 2011 Nature America, Inc. All rights reserved.



© 2011 Nature America, Inc. All rights reserved.

Table 2 Risk variant signals at genome-wide significant celiac disease loci (continued)

Top variant[a]	Chr.	HapMap3 CEU LD block[b] position (*n* markers; size[c])	MAF[d]	P^e	OR	The position of highly correlated variants[f] (*n* markers; size[c])	Localization relative to protein-coding genes[g]
Signal 2 rs72975916	6	"	0.150	$1.2 \times 10^{-5,e}$	0.89	128,307,943–128,339,304 (15; 31)	*PTPRK* exons 28–30 in the 3′ UTR to 24 kb 3′
rs17264332	6	137,924,568–138,316,778 (864; 392)	0.211	5.0×10^{-30}	1.29	138,000,928–138,048,197 (6; 47)	Intergenic region between *OLIG3* and *TNFAIP3*
Signal 2 [imm_6_ 138043754]	6	"	0.190	$2.1 \times 10^{-7,e}$	0.88	138,015,797–138,043,754 (4; 28)	Intergenic between *OLIG3* and *TNFAIP3*
rs182429	6	159,242,314–159,461,818 (514; 220)	0.427	8.5×10^{-16}	1.16	159,385,965–159,390,046 (4; 4)	4 kb 5′ and 5′ UTR of *TAGAP* (NM_152133.1)
Signal 2 rs1107943	6	"	0.071	$2.8 \times 10^{-6,e}$	1.18	159,418,255 (1)	32 kb 5′ of *TAGAP* (NM_152133.1)
[1kg_7_ 37384979]	7	37,330,503–37,406,978 (213; 76)	0.101	2.1×10^{-8}	1.18	37,366,994–37,404,402 (31; 37)	Intron 1 of *ELMO1*
rs10808568	8	129,211,716–129,368,419 (400;157)	0.256	2.2×10^{-5}	0.91	129,333,242–129,345,888 (4; 13)	151–163 kb 3′ of *PVT1*
rs2387397	10	6,428,077–6,585,110 (411; 157)	0.229	1.9×10^{-8}	0.88	6,430,198 (1)	Intergenic region between *PFKFB3* and *PRKCQ*
rs1250552	10	80,690,408–80,774,414 (223; 84)	0.470	8.0×10^{-17}	0.86	80,728,033 (1)	Intron 14 of *ZMIZ1*
rs7104791	11	110,682,429–110,815,769 (3; 133)	0.209	1.9×10^{-11}	1.16	Not high-density genotyped	[region: *POU2AF1*, *C11orf93*]
rs10892258	11	117,847,131–118,270,810 (466; 424)	0.237	1.7×10^{-11}	0.86	118,080,536–118,085,075 (5; 5)	Intergenic region between *TREH* and *DDX6*
rs61907765	11	127,754,640–127,985,723 (480; 231)	0.213	3.4×10^{-13}	1.18	127,886,184–127,901,948 (6; 16)	5 kb 5′ and the first exon of *ETS1* (NM_001162422.1)
rs3184504	12	110,183,529–111,514,870 (938; 1,331)	0.488	5.4×10^{-21}	1.19	110,368,991–110,492,139 (4; 123)	5′ UTR and exons 1–3 of *SH2B3*; exons 2–25 and the 3′ UTR of *ATXN2*
rs11851414	14	68,238,574–68,387,815 (338; 149)	0.221	4.7×10^{-8}	1.13	68,329,159–68,341,722 (3; 13)	1 kb 5′ of and the first exon of *ZFP36L1*
rs1378938	15	72,397,784–73,270,664 (23; 873)	0.278	7.8×10^{-9}	1.13	Not high-density genotyped	[region including *CLK3*, *CSK* and multiple genes]
rs6498114	16	10,834,038–10,903,351 (8; 69)	0.246	5.8×10^{-10}	1.14	Not high-density genotyped	[region: *CIITA*]
rs243323	16	11,220,552–11,385,420 (446; 165)	0.300	2.5×10^{-5}	0.92	11,254,549–11,268,703 (12; 14)	11 kb 5′ of, 1 kb 3′ of and all of *SOCS1*
Signal 2 [imm_16_ 11281298]	16	"	**0.004**	$1.3 \times 10^{-4,e}$	1.70	11,281,298 (1)	Intergenic region between *PRM1* and *PRM2*
Signal 3 rs9673543	16	"	0.169	$2.0 \times 10^{-4,e}$	1.10	11,292,457 (1)	10 kb 5′ of *PRM1*
rs11875687	18	12,728,413–12,914,117 (411; 186)	0.150	1.9×10^{-10}	1.17	12,811,903–12,870,206 (16; 58)	Exons 2–5 of *PTPN2* (NM_ 080422.1)
Signal 2 rs62097857	18	"	**0.040**	$5.2 \times 10^{-5,e}$	1.20	12,847,758 (1)	Intron 2 of *PTPN2* (NM_080422.1)
rs1893592	21	42,683,153–42,760,214 (226; 77)	0.282	3.0×10^{-9}	0.88	42,728,136 (1)	Intron 9 of *UBASH3A* (NM_018961)
rs58911644	21	44,414,408–44,528,088 (239; 114)	0.193	6.2×10^{-7}	0.89	44,446,245–44,453,549 (8; 7)	18–25 kb 3′ of *ICOSLG*
rs4821124	22	20,042,414–20,352,005 (131; 310)	0.186	5.7×10^{-11}	1.16	20,250,903–20,313,260 (36; 62)	*UBE2L3*, *YDJC*
rs13397	X	152,825,373–153,043,675 (88; 218)	0.133	2.7×10^{-8}	1.18	152,872,114–152,937,386 (4; 65)	*HCFC1*, *TMEM187*, *IRAK1*

Non-*HLA* loci meeting genome-wide significance ($P < 5 \times 10^{-8}$) in the current Immunochip data set and in the previous GWAS and replication data set[5] are shown. Loci reported for the first time for celiac disease at genome-wide significance are shown in bold in the 'top variant' column.

[a] dbSNP130 ID. [b] Regions were first defined by LD blocks extending 0.1 cM to the left and right of the risk SNP, as defined by the HapMap3 CEU recombination map. For loci with multiple different previously reported risk SNPs for different diseases and overlapping blocks, the extended region is shown. All chromosomal positions are based on NCBI build 36 (hg18) coordinates. [c] Size in kb. [d] MAFs are shown for the European controls. See **Supplementary Table 4** for more detailed allele frequencies in the cases and controls according to collection. Low-frequency and rare variants are shown in bold. [e] According to a logistic regression association test. The tests for second (and third) independent signals are conditioned on the first (and second) reported variant(s). The per-locus significance thresholds for the second (and third) independent signals are shown in **Supplementary Table 3**. [f] Highly correlated variants are defined as $r^2 > 0.9$, according to hg18. [g] RefSeq track UCSC/hg18. Only the most significantly associated risk variant from each region and independent signal is shown. Variant names are shown as they are listed in dbSNP130 where available, and otherwise, the Illumina Immunochip manifest name is shown in brackets (**Supplementary Table 5** shows both names for the variants). Chr., chromosome.

We identified 13 new celiac risk loci ($P < 5 \times 10^{-8}$; **Fig. 1**, **Table 2** and **Supplementary Fig. 2**), 10 of which were immune-mediated disease loci selected for dense 1000-Genomes–based genotyping on the Immunochip. Several of these new loci were reported at lesser significance levels in our previous studies[5,9], and almost all of these loci have been reported in at least one other immune-mediated disease. These new loci, along with the *HLA* loci, bring the total number of reported (in the current and a previous study[5], which had an

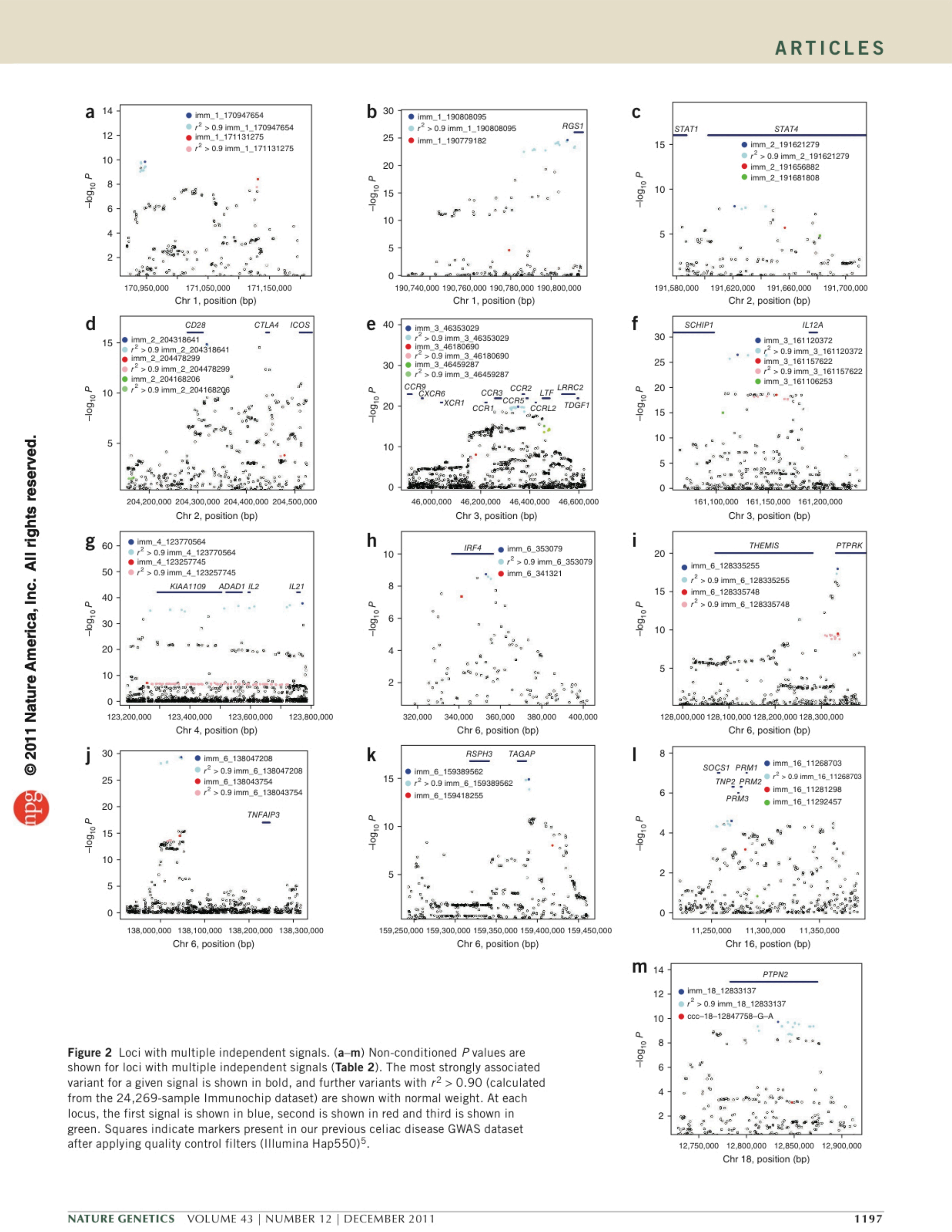
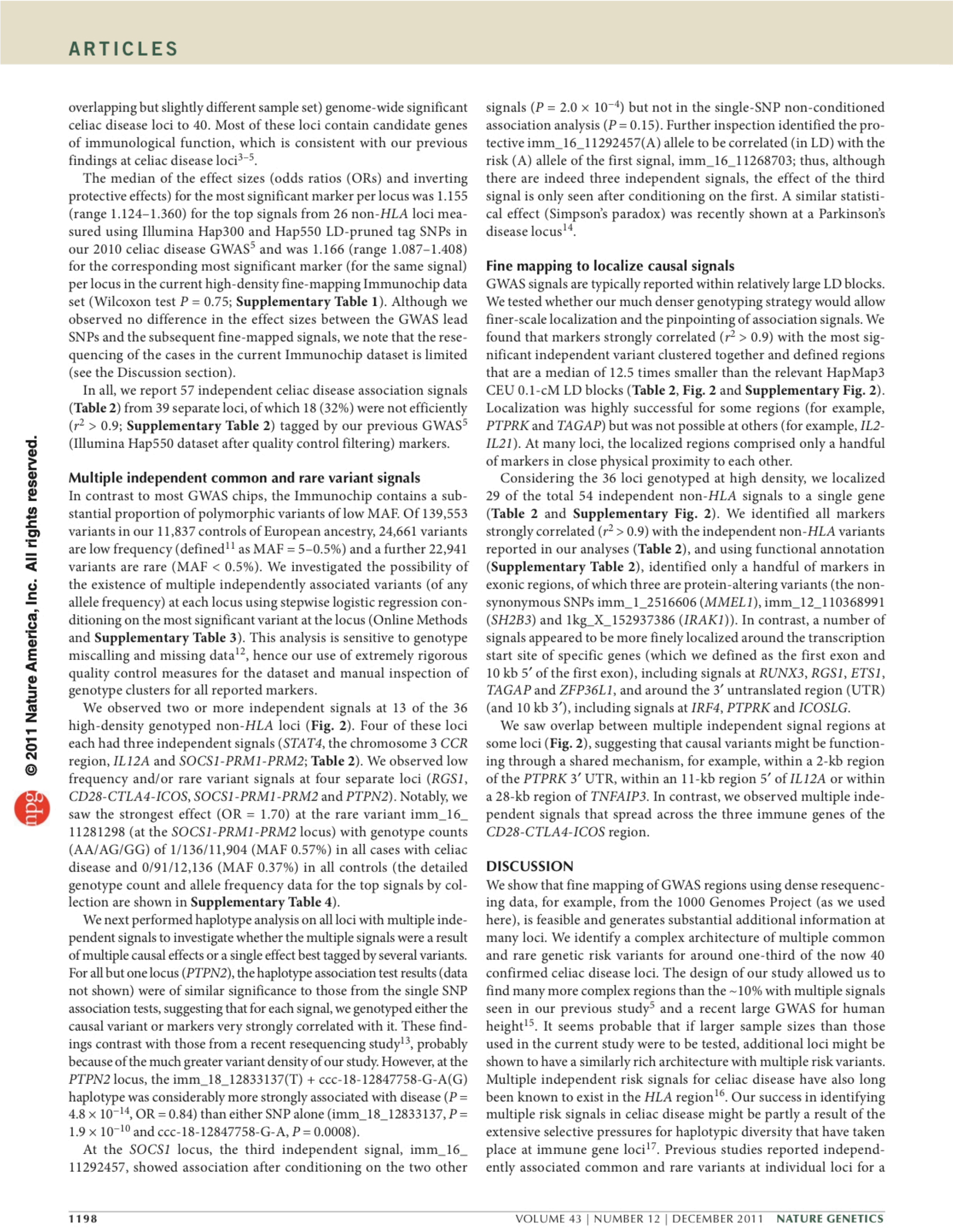


Figure 2 Loci with multiple independent signals. (**a**–**m**) Non-conditioned *P* values are shown for loci with multiple independent signals (**Table 2**). The most strongly associated variant for a given signal is shown in bold, and further variants with $r^2 > 0.90$ (calculated from the 24,269-sample Immunochip dataset) are shown with normal weight. At each locus, the first signal is shown in blue, second is shown in red and third is shown in green. Squares indicate markers present in our previous celiac disease GWAS dataset after applying quality control filters (Illumina Hap550)[5].

© 2011 Nature America, Inc. All rights reserved.



© 2011 Nature America, Inc. All rights reserved.

overlapping but slightly different sample set) genome-wide significant celiac disease loci to 40. Most of these loci contain candidate genes of immunological function, which is consistent with our previous findings at celiac disease loci[3–5].

The median of the effect sizes (odds ratios (ORs) and inverting protective effects) for the most significant marker per locus was 1.155 (range 1.124–1.360) for the top signals from 26 non-*HLA* loci measured using Illumina Hap300 and Hap550 LD-pruned tag SNPs in our 2010 celiac disease GWAS[5] and was 1.166 (range 1.087–1.408) for the corresponding most significant marker (for the same signal) per locus in the current high-density fine-mapping Immunochip data set (Wilcoxon test $P = 0.75$; **Supplementary Table 1**). Although we observed no difference in the effect sizes between the GWAS lead SNPs and the subsequent fine-mapped signals, we note that the resequencing of the cases in the current Immunochip dataset is limited (see the Discussion section).

In all, we report 57 independent celiac disease association signals (**Table 2**) from 39 separate loci, of which 18 (32%) were not efficiently ($r^2 > 0.9$; **Supplementary Table 2**) tagged by our previous GWAS[5] (Illumina Hap550 dataset after quality control filtering) markers.

Multiple independent common and rare variant signals

In contrast to most GWAS chips, the Immunochip contains a substantial proportion of polymorphic variants of low MAF. Of 139,553 variants in our 11,837 controls of European ancestry, 24,661 variants are low frequency (defined[11] as MAF = 5–0.5%) and a further 22,941 variants are rare (MAF < 0.5%). We investigated the possibility of the existence of multiple independently associated variants (of any allele frequency) at each locus using stepwise logistic regression conditioning on the most significant variant at the locus (Online Methods and **Supplementary Table 3**). This analysis is sensitive to genotype miscalling and missing data[12], hence our use of extremely rigorous quality control measures for the dataset and manual inspection of genotype clusters for all reported markers.

We observed two or more independent signals at 13 of the 36 high-density genotyped non-*HLA* loci (**Fig. 2**). Four of these loci each had three independent signals (*STAT4*, the chromosome 3 *CCR* region, *IL12A* and *SOCS1-PRM1-PRM2*; **Table 2**). We observed low frequency and/or rare variant signals at four separate loci (*RGS1*, *CD28-CTLA4-ICOS*, *SOCS1-PRM1-PRM2* and *PTPN2*). Notably, we saw the strongest effect (OR = 1.70) at the rare variant imm_16_11281298 (at the *SOCS1-PRM1-PRM2* locus) with genotype counts (AA/AG/GG) of 1/136/11,904 (MAF 0.57%) in all cases with celiac disease and 0/91/12,136 (MAF 0.37%) in all controls (the detailed genotype count and allele frequency data for the top signals by collection are shown in **Supplementary Table 4**).

We next performed haplotype analysis on all loci with multiple independent signals to investigate whether the multiple signals were a result of multiple causal effects or a single effect best tagged by several variants. For all but one locus (*PTPN2*), the haplotype association test results (data not shown) were of similar significance to those from the single SNP association tests, suggesting that for each signal, we genotyped either the causal variant or markers very strongly correlated with it. These findings contrast with those from a recent resequencing study[13], probably because of the much greater variant density of our study. However, at the *PTPN2* locus, the imm_18_12833137(T) + ccc-18-12847758-G-A(G) haplotype was considerably more strongly associated with disease ($P = 4.8 \times 10^{-14}$, OR = 0.84) than either SNP alone (imm_18_12833137, $P = 1.9 \times 10^{-10}$ and ccc-18-12847758-G-A, $P = 0.0008$).

At the *SOCS1* locus, the third independent signal, imm_16_11292457, showed association after conditioning on the two other signals ($P = 2.0 \times 10^{-4}$) but not in the single-SNP non-conditioned association analysis ($P = 0.15$). Further inspection identified the protective imm_16_11292457(A) allele to be correlated (in LD) with the risk (A) allele of the first signal, imm_16_11268703; thus, although there are indeed three independent signals, the effect of the third signal is only seen after conditioning on the first. A similar statistical effect (Simpson's paradox) was recently shown at a Parkinson's disease locus[14].

Fine mapping to localize causal signals

GWAS signals are typically reported within relatively large LD blocks. We tested whether our much denser genotyping strategy would allow finer-scale localization and the pinpointing of association signals. We found that markers strongly correlated ($r^2 > 0.9$) with the most significant independent variant clustered together and defined regions that are a median of 12.5 times smaller than the relevant HapMap3 CEU 0.1-cM LD blocks (**Table 2**, **Fig. 2** and **Supplementary Fig. 2**). Localization was highly successful for some regions (for example, *PTPRK* and *TAGAP*) but was not possible at others (for example, *IL2-IL21*). At many loci, the localized regions comprised only a handful of markers in close physical proximity to each other.

Considering the 36 loci genotyped at high density, we localized 29 of the total 54 independent non-*HLA* signals to a single gene (**Table 2** and **Supplementary Fig. 2**). We identified all markers strongly correlated ($r^2 > 0.9$) with the independent non-*HLA* variants reported in our analyses (**Table 2**), and using functional annotation (**Supplementary Table 2**), identified only a handful of markers in exonic regions, of which three are protein-altering variants (the non-synonymous SNPs imm_1_2516606 (*MMEL1*), imm_12_110368991 (*SH2B3*) and 1kg_X_152937386 (*IRAK1*)). In contrast, a number of signals appeared to be more finely localized around the transcription start site of specific genes (which we defined as the first exon and 10 kb 5′ of the first exon), including signals at *RUNX3*, *RGS1*, *ETS1*, *TAGAP* and *ZFP36L1*, and around the 3′ untranslated region (UTR) (and 10 kb 3′), including signals at *IRF4*, *PTPRK* and *ICOSLG*.

We saw overlap between multiple independent signal regions at some loci (**Fig. 2**), suggesting that causal variants might be functioning through a shared mechanism, for example, within a 2-kb region of the *PTPRK* 3′ UTR, within an 11-kb region 5′ of *IL12A* or within a 28-kb region of *TNFAIP3*. In contrast, we observed multiple independent signals that spread across the three immune genes of the *CD28-CTLA4-ICOS* region.

DISCUSSION

We show that fine mapping of GWAS regions using dense resequencing data, for example, from the 1000 Genomes Project (as we used here), is feasible and generates substantial additional information at many loci. We identify a complex architecture of multiple common and rare genetic risk variants for around one-third of the now 40 confirmed celiac disease loci. The design of our study allowed us to find many more complex regions than the ~10% with multiple signals seen in our previous study[5] and a recent large GWAS for human height[15]. It seems probable that if larger sample sizes than those used in the current study were to be tested, additional loci might be shown to have a similarly rich architecture with multiple risk variants. Multiple independent risk signals for celiac disease have also long been known to exist in the *HLA* region[16]. Our success in identifying multiple risk signals in celiac disease might be partly a result of the extensive selective pressures for haplotypic diversity that have taken place at immune gene loci[17]. Previous studies reported independently associated common and rare variants at individual loci for a

© 2011 Nature America, Inc. All rights reserved.

handful of phenotypes, for example, fetal hemoglobin[13], sick sinus syndrome[18], Crohn's disease[19] and hypertriglyceridemia[20]. To the best of our knowledge, this is the first study to have comprehensively surveyed the genetic architecture of all known risk loci for a trait.

In part, our identification of rare variants at risk regions relies on the prior discovery of a genome-wide significant common variant association signal at each locus. This then permits a per-locus correction rather than a genome-wide multiple-testing correction when searching for additional independent association signals. Only particularly strong rare variant signals would, on their own, generate significance levels reaching the genome-wide threshold typically used in GWAS ($P < 5 \times 10^{-8}$). Alternative methods, such as collapsing rare variant signals across a gene or functional categories of genes, have therefore been suggested as approaches to this problem[21]. Although a rare variant may occur on a recent haplotypic background and thus show LD at a substantially longer range than common variants, we deliberately restricted our search to around the common-variant LD blocks because to do otherwise would have incurred a considerably greater penalty from multiple testing. Therefore, although our study provides considerable support for exome and whole-genome sequencing efforts aimed at identifying rare risk variants (and those not necessarily restricted to GWAS loci) in common complex diseases, it further highlights the statistical challenges of establishing associations for rare variants.

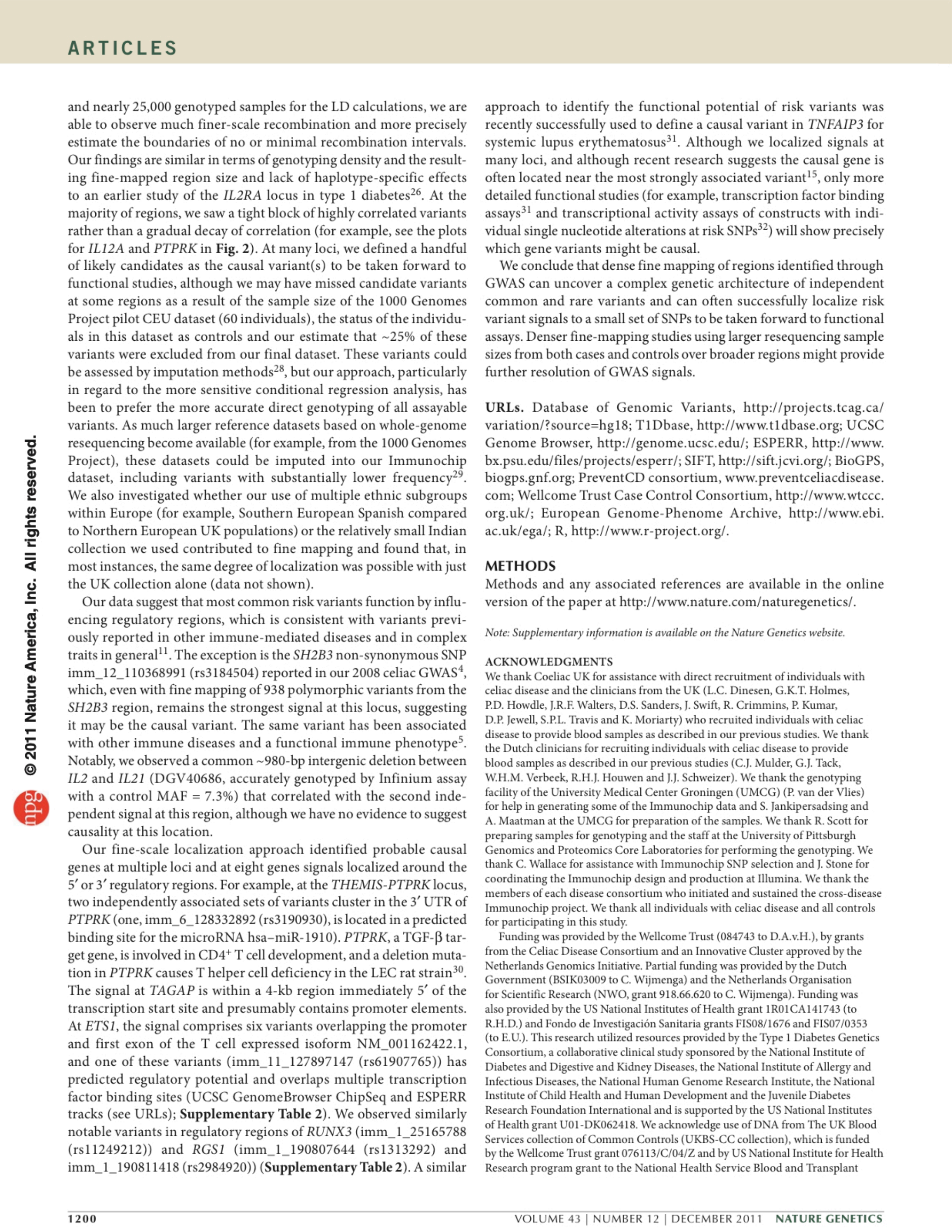
We used a dense genotyping strategy and a stepwise conditional association analysis but did not identify any rare highly penetrant variants that might explain the genome-wide significant common SNP signals at any of the 39 loci. Our study does have limitations in this regard, particularly: (i) the restriction of the analysis to 0.1-cM LD blocks; (ii) the limited control resequencing sample size of the 1000 Genomes Project pilot CEU dataset; (iii) the limited case resequencing sample size; and (iv) case resequencing being limited to three loci for celiac disease and to selected loci for other immune diseases. We observed a weak trend toward a lower MAF ($P = 0.042$, Wilcoxon test; **Supplementary Table 1**) for the best fine-mapping SNP from the Immunochip analysis compared to the lead SNP from our 2010 tag SNP GWAS (determined by measuring the MAF in a subset of samples genotyped in both datasets). One signal showed a substantially higher MAF (>25% change) using fine mapping and four signals showed a substantially lower MAF using fine mapping (**Supplementary Table 1**), however, all fine-mapping variants corresponding to the lead GWAS SNPs remained common (MAF > 0.10). We suggest that these changes in the MAFs of the lead GWAS SNPs using fine mapping simply reflect a more precise measurement of common frequency risk haplotypes. Although we cannot exclude the possibility that a single high-penetrance lower-frequency variant explains most of the association signal at a locus, especially without more comprehensive resequencing of the cases, we found no evidence to support this possibility in the current fine-mapping analysis. Similarly, although our stepwise selection procedure cannot robustly refute the 'synthetic association' hypothesis—in particular, that a combination of multiple rare variants jointly explains the association signal[22]—we have so far not observed any evidence supporting this possibility.

We identified 13 new loci for celiac disease at genome-wide significance, most of which have been reported previously at lesser significance levels or in another immune-mediated disease. The Illumina Hap550 chip (used in our 2010 GWAS) would have detected 10 of the 13 new loci and, in total, 39 of the 57 independent non-*HLA* signals that we report here. A current genotyping platform, the Illumina Omni2.5 chip, would have detected 12 of the 13 new loci and, in total, 50 of the 57 independent non-*HLA* signals that we report here. However, neither of these chips would have provided the finer-scale localization of the Immunochip. The 13 new loci contain many candidate genes with an immunological function ($P = 0.0002$ for enrichment of the Gene Ontology term 'immune system process'[23]), which is in line with expectations based on our previous studies. We also found evidence suggesting that substantial additional signals exist at other immune-mediated disease loci that are below the genome-wide significance threshold applied to the current dataset. It is a point of debate whether such strict ($P < 5 \times 10^{-8}$) criteria should apply; for example, an analyst might apply a higher Bayesian prior at a locus already reported in another immune-mediated disease. Alternatively, an Immunochip-wide P value with a Bonferroni correction for independent SNPs, as was used recently in the Cardiochip custom genotyping project[24], of $P < 1.9 \times 10^{-6}$ (Online Methods) would yield 16 new celiac disease loci in addition to the 13 we identified here. These 16 loci also mostly contain immune system genes. An analysis of these signals of intermediate significance would gain substantial additional power in a meta-analysis across the several hundred thousand samples from multiple immune-mediated disease collections currently being run on the Immunochip.

We found that our previous GWAS using tag SNPs gave very similar estimates of effect size as our current fine-mapping experiment (**Supplementary Table 1**), which is in contrast to a simulation study that suggested that GWAS markers often underestimate risk[14]. However, we found substantial evidence for multiple additional signals at known loci and report many new loci. In individuals of European ancestry, the 39 non-*HLA* loci explain 13.7% of the genetic variance of celiac disease (*HLA* variants account for a further ~40%). We also show a long list of effects of weaker significance, which will explain substantial additional heritability.

Only one of the variants reported here was discovered in a disease-specific resequencing study: ccc-18-12847758-G-A (rs62097857), a marker identified by the WTCCC's resequencing of cases with Crohn's disease and controls (**Supplementary Note**) and that is also present in the Watson genome. We submitted for Immunochip analysis ~4,000 variants from high-throughput resequencing of pools of 80 cases with celiac disease for extended genomic regions at three loci (*RGS1*, *IL12A* and *IL2-IL21*; **Supplementary Note**). These loci did not contribute any signals in addition to those obtained from the 1000 Genomes Project pilot CEU variants, although they did increase the number of variants correlated with each signal (that is, the set of markers that probably contains the causal variant(s)) and more precisely define the boundaries of the signal localization. We note that larger-scale resequencing of cases (for example, using many hundreds of samples) would identify a spectrum containing more rare variants than the current study, and this method has previously been used with success at selected genes and phenotypes.

The possibility of performing fine-scale mapping of GWAS regions using, for example, 1000 Genomes Project data, has been discussed as a natural follow-up strategy for such studies[25,26] and has been used recently to identify risk variants in *APOL1* in African-Americans with renal disease[27]. To our knowledge, our current report is the first to test such a strategy on a large scale in a complex disease. At multiple regions, we were able to refine the signal to a handful of variants over a few kb or tens of kb, although some regions (for example, *IL2-IL21*) were resistant to this approach, presumably because of the presence of particularly strong LD. Most GWAS report signals mapping to an LD block based on HapMap recombination rates (with a sample size of 60 families from the CEU dataset). In our data, where we have both much denser genotyping than GWAS chips (with a mean of 13.6× the genotyping density at celiac loci compared to the Illumina Hap550 chip)



© 2011 Nature America, Inc. All rights reserved.

and nearly 25,000 genotyped samples for the LD calculations, we are able to observe much finer-scale recombination and more precisely estimate the boundaries of no or minimal recombination intervals. Our findings are similar in terms of genotyping density and the resulting fine-mapped region size and lack of haplotype-specific effects to an earlier study of the *IL2RA* locus in type 1 diabetes[26]. At the majority of regions, we saw a tight block of highly correlated variants rather than a gradual decay of correlation (for example, see the plots for *IL12A* and *PTPRK* in **Fig. 2**). At many loci, we defined a handful of likely candidates as the causal variant(s) to be taken forward to functional studies, although we may have missed candidate variants at some regions as a result of the sample size of the 1000 Genomes Project pilot CEU dataset (60 individuals), the status of the individuals in this dataset as controls and our estimate that ~25% of these variants were excluded from our final dataset. These variants could be assessed by imputation methods[28], but our approach, particularly in regard to the more sensitive conditional regression analysis, has been to prefer the more accurate direct genotyping of all assayable variants. As much larger reference datasets based on whole-genome resequencing become available (for example, from the 1000 Genomes Project), these datasets could be imputed into our Immunochip dataset, including variants with substantially lower frequency[29]. We also investigated whether our use of multiple ethnic subgroups within Europe (for example, Southern European Spanish compared to Northern European UK populations) or the relatively small Indian collection we used contributed to fine mapping and found that, in most instances, the same degree of localization was possible with just the UK collection alone (data not shown).

Our data suggest that most common risk variants function by influencing regulatory regions, which is consistent with variants previously reported in other immune-mediated diseases and in complex traits in general[11]. The exception is the *SH2B3* non-synonymous SNP imm_12_110368991 (rs3184504) reported in our 2008 celiac GWAS[4], which, even with fine mapping of 938 polymorphic variants from the *SH2B3* region, remains the strongest signal at this locus, suggesting it may be the causal variant. The same variant has been associated with other immune diseases and a functional immune phenotype[5]. Notably, we observed a common ~980-bp intergenic deletion between *IL2* and *IL21* (DGV40686, accurately genotyped by Infinium assay with a control MAF = 7.3%) that correlated with the second independent signal at this region, although we have no evidence to suggest causality at this location.

Our fine-scale localization approach identified probable causal genes at multiple loci and at eight genes signals localized around the 5′ or 3′ regulatory regions. For example, at the *THEMIS-PTPRK* locus, two independently associated sets of variants cluster in the 3′ UTR of *PTPRK* (one, imm_6_128332892 (rs3190930), is located in a predicted binding site for the microRNA hsa–miR-1910). *PTPRK*, a TGF-β target gene, is involved in CD4$^+$ T cell development, and a deletion mutation in *PTPRK* causes T helper cell deficiency in the LEC rat strain[30]. The signal at *TAGAP* is within a 4-kb region immediately 5′ of the transcription start site and presumably contains promoter elements. At *ETS1*, the signal comprises six variants overlapping the promoter and first exon of the T cell expressed isoform NM_001162422.1, and one of these variants (imm_11_127897147 (rs61907765)) has predicted regulatory potential and overlaps multiple transcription factor binding sites (UCSC GenomeBrowser ChipSeq and ESPERR tracks (see URLs); **Supplementary Table 2**). We observed similarly notable variants in regulatory regions of *RUNX3* (imm_1_25165788 (rs11249212)) and *RGS1* (imm_1_190807644 (rs1313292) and imm_1_190811418 (rs2984920)) (**Supplementary Table 2**). A similar approach to identify the functional potential of risk variants was recently successfully used to define a causal variant in *TNFAIP3* for systemic lupus erythematosus[31]. Although we localized signals at many loci, and although recent research suggests the causal gene is often located near the most strongly associated variant[15], only more detailed functional studies (for example, transcription factor binding assays[31] and transcriptional activity assays of constructs with individual single nucleotide alterations at risk SNPs[32]) will show precisely which gene variants might be causal.

We conclude that dense fine mapping of regions identified through GWAS can uncover a complex genetic architecture of independent common and rare variants and can often successfully localize risk variant signals to a small set of SNPs to be taken forward to functional assays. Denser fine-mapping studies using larger resequencing sample sizes from both cases and controls over broader regions might provide further resolution of GWAS signals.

URLs. Database of Genomic Variants, http://projects.tcag.ca/variation/?source=hg18; T1Dbase, http://www.t1dbase.org; UCSC Genome Browser, http://genome.ucsc.edu/; ESPERR, http://www.bx.psu.edu/files/projects/esperr/; SIFT, http://sift.jcvi.org/; BioGPS, biogps.gnf.org; PreventCD consortium, www.preventceliacdisease.com; Wellcome Trust Case Control Consortium, http://www.wtccc.org.uk/; European Genome-Phenome Archive, http://www.ebi.ac.uk/ega/; R, http://www.r-project.org/.

METHODS

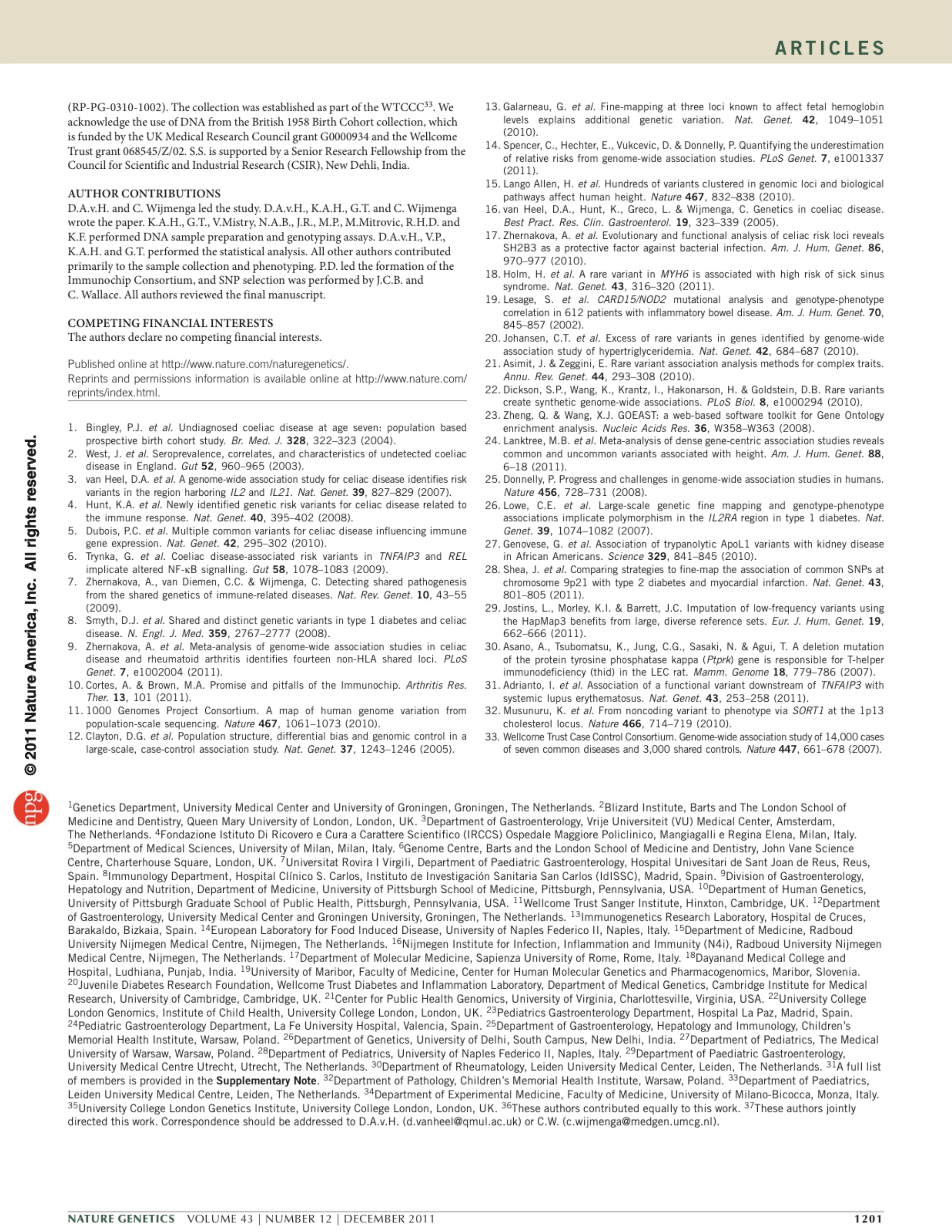
Methods and any associated references are available in the online version of the paper at http://www.nature.com/naturegenetics/.

Note: Supplementary information is available on the Nature Genetics website.

ACKNOWLEDGMENTS

We thank Coeliac UK for assistance with direct recruitment of individuals with celiac disease and the clinicians from the UK (L.C. Dinesen, G.K.T. Holmes, P.D. Howdle, J.R.F. Walters, D.S. Sanders, J. Swift, R. Crimmins, P. Kumar, D.P. Jewell, S.P.L. Travis and K. Moriarty) who recruited individuals with celiac disease to provide blood samples as described in our previous studies. We thank the Dutch clinicians for recruiting individuals with celiac disease to provide blood samples as described in our previous studies (C.J. Mulder, G.J. Tack, W.H.M. Verbeek, R.H.J. Houwen and J.J. Schweizer). We thank the genotyping facility of the University Medical Center Groningen (UMCG) (P. van der Vlies) for help in generating some of the Immunochip data and S. Jankipersadsing and A. Maatman at the UMCG for preparation of the samples. We thank R. Scott for preparing samples for genotyping and the staff at the University of Pittsburgh Genomics and Proteomics Core Laboratories for performing the genotyping. We thank C. Wallace for assistance with Immunochip SNP selection and J. Stone for coordinating the Immunochip design and production at Illumina. We thank the members of each disease consortium who initiated and sustained the cross-disease Immunochip project. We thank all individuals with celiac disease and all controls for participating in this study.

Funding was provided by the Wellcome Trust (084743 to D.A.v.H.), by grants from the Celiac Disease Consortium and an Innovative Cluster approved by the Netherlands Genomics Initiative. Partial funding was provided by the Dutch Government (BSIK03009 to C. Wijmenga) and the Netherlands Organisation for Scientific Research (NWO, grant 918.66.620 to C. Wijmenga). Funding was also provided by the US National Institutes of Health grant 1R01CA141743 (to R.H.D.) and Fondo de Investigación Sanitaria grants FIS08/1676 and FIS07/0353 (to E.U.). This research utilized resources provided by the Type 1 Diabetes Genetics Consortium, a collaborative clinical study sponsored by the National Institute of Diabetes and Digestive and Kidney Diseases, the National Institute of Allergy and Infectious Diseases, the National Human Genome Research Institute, the National Institute of Child Health and Human Development and the Juvenile Diabetes Research Foundation International and is supported by the US National Institutes of Health grant U01-DK062418. We acknowledge use of DNA from The UK Blood Services collection of Common Controls (UKBS-CC collection), which is funded by the Wellcome Trust grant 076113/C/04/Z and by US National Institute for Health Research program grant to the National Health Service Blood and Transplant



(RP-PG-0310-1002). The collection was established as part of the WTCCC[33]. We acknowledge the use of DNA from the British 1958 Birth Cohort collection, which is funded by the UK Medical Research Council grant G0000934 and the Wellcome Trust grant 068545/Z/02. S.S. is supported by a Senior Research Fellowship from the Council for Scientific and Industrial Research (CSIR), New Dehli, India.

AUTHOR CONTRIBUTIONS

D.A.v.H. and C. Wijmenga led the study. D.A.v.H., K.A.H., G.T. and C. Wijmenga wrote the paper. K.A.H., G.T., V.Mistry, N.A.B., J.R., M.P., M.Mitrovic, R.H.D. and K.F. performed DNA sample preparation and genotyping assays. D.A.v.H., V.P., K.A.H. and G.T. performed the statistical analysis. All other authors contributed primarily to the sample collection and phenotyping. P.D. led the formation of the Immunochip Consortium, and SNP selection was performed by J.C.B. and C. Wallace. All authors reviewed the final manuscript.

COMPETING FINANCIAL INTERESTS

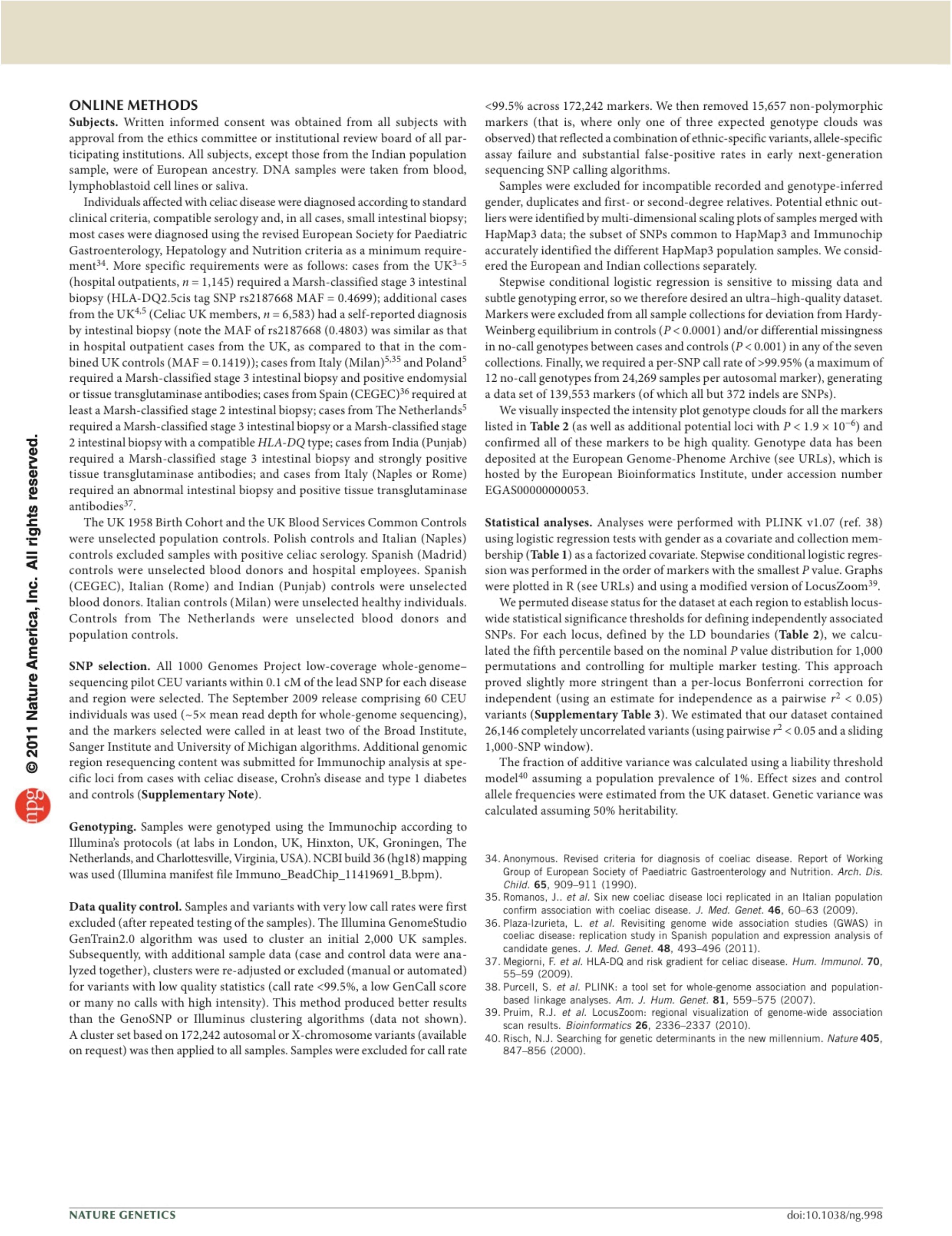
The authors declare no competing financial interests.

Published online at http://www.nature.com/naturegenetics/.
Reprints and permissions information is available online at http://www.nature.com/reprints/index.html.

1. Bingley, P.J. *et al.* Undiagnosed coeliac disease at age seven: population based prospective birth cohort study. *Br. Med. J.* **328**, 322–323 (2004).
2. West, J. *et al.* Seroprevalence, correlates, and characteristics of undetected coeliac disease in England. *Gut* **52**, 960–965 (2003).
3. van Heel, D.A. *et al.* A genome-wide association study for celiac disease identifies risk variants in the region harboring *IL2* and *IL21*. *Nat. Genet.* **39**, 827–829 (2007).
4. Hunt, K.A. *et al.* Newly identified genetic risk variants for celiac disease related to the immune response. *Nat. Genet.* **40**, 395–402 (2008).
5. Dubois, P.C. *et al.* Multiple common variants for celiac disease influencing immune gene expression. *Nat. Genet.* **42**, 295–302 (2010).
6. Trynka, G. *et al.* Coeliac disease-associated risk variants in *TNFAIP3* and *REL* implicate altered NF-κB signalling. *Gut* **58**, 1078–1083 (2009).
7. Zhernakova, A., van Diemen, C.C. & Wijmenga, C. Detecting shared pathogenesis from the shared genetics of immune-related diseases. *Nat. Rev. Genet.* **10**, 43–55 (2009).
8. Smyth, D.J. *et al.* Shared and distinct genetic variants in type 1 diabetes and celiac disease. *N. Engl. J. Med.* **359**, 2767–2777 (2008).
9. Zhernakova, A. *et al.* Meta-analysis of genome-wide association studies in celiac disease and rheumatoid arthritis identifies fourteen non-HLA shared loci. *PLoS Genet.* **7**, e1002004 (2011).
10. Cortes, A. & Brown, M.A. Promise and pitfalls of the Immunochip. *Arthritis Res. Ther.* **13**, 101 (2011).
11. 1000 Genomes Project Consortium. A map of human genome variation from population-scale sequencing. *Nature* **467**, 1061–1073 (2010).
12. Clayton, D.G. *et al.* Population structure, differential bias and genomic control in a large-scale, case-control association study. *Nat. Genet.* **37**, 1243–1246 (2005).
13. Galarneau, G. *et al.* Fine-mapping at three loci known to affect fetal hemoglobin levels explains additional genetic variation. *Nat. Genet.* **42**, 1049–1051 (2010).
14. Spencer, C., Hechter, E., Vukcevic, D. & Donnelly, P. Quantifying the underestimation of relative risks from genome-wide association studies. *PLoS Genet.* **7**, e1001337 (2011).
15. Lango Allen, H. *et al.* Hundreds of variants clustered in genomic loci and biological pathways affect human height. *Nature* **467**, 832–838 (2010).
16. van Heel, D.A., Hunt, K., Greco, L. & Wijmenga, C. Genetics in coeliac disease. *Best Pract. Res. Clin. Gastroenterol.* **19**, 323–339 (2005).
17. Zhernakova, A. *et al.* Evolutionary and functional analysis of celiac risk loci reveals SH2B3 as a protective factor against bacterial infection. *Am. J. Hum. Genet.* **86**, 970–977 (2010).
18. Holm, H. *et al.* A rare variant in *MYH6* is associated with high risk of sick sinus syndrome. *Nat. Genet.* **43**, 316–320 (2011).
19. Lesage, S. *et al.* *CARD15/NOD2* mutational analysis and genotype-phenotype correlation in 612 patients with inflammatory bowel disease. *Am. J. Hum. Genet.* **70**, 845–857 (2002).
20. Johansen, C.T. *et al.* Excess of rare variants in genes identified by genome-wide association study of hypertriglyceridemia. *Nat. Genet.* **42**, 684–687 (2010).
21. Asimit, J. & Zeggini, E. Rare variant association analysis methods for complex traits. *Annu. Rev. Genet.* **44**, 293–308 (2010).
22. Dickson, S.P., Wang, K., Krantz, I., Hakonarson, H. & Goldstein, D.B. Rare variants create synthetic genome-wide associations. *PLoS Biol.* **8**, e1000294 (2010).
23. Zheng, Q. & Wang, X.J. GOEAST: a web-based software toolkit for Gene Ontology enrichment analysis. *Nucleic Acids Res.* **36**, W358–W363 (2008).
24. Lanktree, M.B. *et al.* Meta-analysis of dense gene-centric association studies reveals common and uncommon variants associated with height. *Am. J. Hum. Genet.* **88**, 6–18 (2011).
25. Donnelly, P. Progress and challenges in genome-wide association studies in humans. *Nature* **456**, 728–731 (2008).
26. Lowe, C.E. *et al.* Large-scale genetic fine mapping and genotype-phenotype associations implicate polymorphism in the *IL2RA* region in type 1 diabetes. *Nat. Genet.* **39**, 1074–1082 (2007).
27. Genovese, G. *et al.* Association of trypanolytic ApoL1 variants with kidney disease in African Americans. *Science* **329**, 841–845 (2010).
28. Shea, J. *et al.* Comparing strategies to fine-map the association of common SNPs at chromosome 9p21 with type 2 diabetes and myocardial infarction. *Nat. Genet.* **43**, 801–805 (2011).
29. Jostins, L., Morley, K.I. & Barrett, J.C. Imputation of low-frequency variants using the HapMap3 benefits from large, diverse reference sets. *Eur. J. Hum. Genet.* **19**, 662–666 (2011).
30. Asano, A., Tsubomatsu, K., Jung, C.G., Sasaki, N. & Agui, T. A deletion mutation of the protein tyrosine phosphatase kappa (*Ptprk*) gene is responsible for T-helper immunodeficiency (thid) in the LEC rat. *Mamm. Genome* **18**, 779–786 (2007).
31. Adrianto, I. *et al.* Association of a functional variant downstream of *TNFAIP3* with systemic lupus erythematosus. *Nat. Genet.* **43**, 253–258 (2011).
32. Musunuru, K. *et al.* From noncoding variant to phenotype via *SORT1* at the 1p13 cholesterol locus. *Nature* **466**, 714–719 (2010).
33. Wellcome Trust Case Control Consortium. Genome-wide association study of 14,000 cases of seven common diseases and 3,000 shared controls. *Nature* **447**, 661–678 (2007).

© 2011 Nature America, Inc. All rights reserved.

[1]Genetics Department, University Medical Center and University of Groningen, Groningen, The Netherlands. [2]Blizard Institute, Barts and The London School of Medicine and Dentistry, Queen Mary University of London, London, UK. [3]Department of Gastroenterology, Vrije Universiteit (VU) Medical Center, Amsterdam, The Netherlands. [4]Fondazione Istituto Di Ricovero e Cura a Carattere Scientifico (IRCCS) Ospedale Maggiore Policlinico, Mangiagalli e Regina Elena, Milan, Italy. [5]Department of Medical Sciences, University of Milan, Milan, Italy. [6]Genome Centre, Barts and the London School of Medicine and Dentistry, John Vane Science Centre, Charterhouse Square, London, UK. [7]Universitat Rovira I Virgili, Department of Paediatric Gastroenterology, Hospital Univesitari de Sant Joan de Reus, Reus, Spain. [8]Immunology Department, Hospital Clínico S. Carlos, Instituto de Investigación Sanitaria San Carlos (IdISSC), Madrid, Spain. [9]Division of Gastroenterology, Hepatology and Nutrition, Department of Medicine, University of Pittsburgh School of Medicine, Pittsburgh, Pennsylvania, USA. [10]Department of Human Genetics, University of Pittsburgh Graduate School of Public Health, Pittsburgh, Pennsylvania, USA. [11]Wellcome Trust Sanger Institute, Hinxton, Cambridge, UK. [12]Department of Gastroenterology, University Medical Center and Groningen University, Groningen, The Netherlands. [13]Immunogenetics Research Laboratory, Hospital de Cruces, Barakaldo, Bizkaia, Spain. [14]European Laboratory for Food Induced Disease, University of Naples Federico II, Naples, Italy. [15]Department of Medicine, Radboud University Nijmegen Medical Centre, Nijmegen, The Netherlands. [16]Nijmegen Institute for Infection, Inflammation and Immunity (N4i), Radboud University Nijmegen Medical Centre, Nijmegen, The Netherlands. [17]Department of Molecular Medicine, Sapienza University of Rome, Rome, Italy. [18]Dayanand Medical College and Hospital, Ludhiana, Punjab, India. [19]University of Maribor, Faculty of Medicine, Center for Human Molecular Genetics and Pharmacogenomics, Maribor, Slovenia. [20]Juvenile Diabetes Research Foundation, Wellcome Trust Diabetes and Inflammation Laboratory, Department of Medical Genetics, Cambridge Institute for Medical Research, University of Cambridge, Cambridge, UK. [21]Center for Public Health Genomics, University of Virginia, Charlottesville, Virginia, USA. [22]University College London Genomics, Institute of Child Health, University College London, London, UK. [23]Pediatrics Gastroenterology Department, Hospital La Paz, Madrid, Spain. [24]Pediatric Gastroenterology Department, La Fe University Hospital, Valencia, Spain. [25]Department of Gastroenterology, Hepatology and Immunology, Children's Memorial Health Institute, Warsaw, Poland. [26]Department of Genetics, University of Delhi, South Campus, New Delhi, India. [27]Department of Pediatrics, The Medical University of Warsaw, Warsaw, Poland. [28]Department of Pediatrics, University of Naples Federico II, Naples, Italy. [29]Department of Paediatric Gastroenterology, University Medical Centre Utrecht, Utrecht, The Netherlands. [30]Department of Rheumatology, Leiden University Medical Center, Leiden, The Netherlands. [31]A full list of members is provided in the **Supplementary Note**. [32]Department of Pathology, Children's Memorial Health Institute, Warsaw, Poland. [33]Department of Paediatrics, Leiden University Medical Centre, Leiden, The Netherlands. [34]Department of Experimental Medicine, Faculty of Medicine, University of Milano-Bicocca, Monza, Italy. [35]University College London Genetics Institute, University College London, London, UK. [36]These authors contributed equally to this work. [37]These authors jointly directed this work. Correspondence should be addressed to D.A.v.H. (d.vanheel@qmul.ac.uk) or C.W. (c.wijmenga@medgen.umcg.nl).



© 2011 Nature America, Inc. All rights reserved.

ONLINE METHODS

Subjects. Written informed consent was obtained from all subjects with approval from the ethics committee or institutional review board of all participating institutions. All subjects, except those from the Indian population sample, were of European ancestry. DNA samples were taken from blood, lymphoblastoid cell lines or saliva.

Individuals affected with celiac disease were diagnosed according to standard clinical criteria, compatible serology and, in all cases, small intestinal biopsy; most cases were diagnosed using the revised European Society for Paediatric Gastroenterology, Hepatology and Nutrition criteria as a minimum requirement[34]. More specific requirements were as follows: cases from the UK[3–5] (hospital outpatients, $n = 1{,}145$) required a Marsh-classified stage 3 intestinal biopsy (HLA-DQ2.5cis tag SNP rs2187668 MAF = 0.4699); additional cases from the UK[4,5] (Celiac UK members, $n = 6{,}583$) had a self-reported diagnosis by intestinal biopsy (note the MAF of rs2187668 (0.4803) was similar as that in hospital outpatient cases from the UK, as compared to that in the combined UK controls (MAF = 0.1419)); cases from Italy (Milan)[5,35] and Poland[5] required a Marsh-classified stage 3 intestinal biopsy and positive endomysial or tissue transglutaminase antibodies; cases from Spain (CEGEC)[36] required at least a Marsh-classified stage 2 intestinal biopsy; cases from The Netherlands[5] required a Marsh-classified stage 3 intestinal biopsy or a Marsh-classified stage 2 intestinal biopsy with a compatible *HLA-DQ* type; cases from India (Punjab) required a Marsh-classified stage 3 intestinal biopsy and strongly positive tissue transglutaminase antibodies; and cases from Italy (Naples or Rome) required an abnormal intestinal biopsy and positive tissue transglutaminase antibodies[37].

The UK 1958 Birth Cohort and the UK Blood Services Common Controls were unselected population controls. Polish controls and Italian (Naples) controls excluded samples with positive celiac serology. Spanish (Madrid) controls were unselected blood donors and hospital employees. Spanish (CEGEC), Italian (Rome) and Indian (Punjab) controls were unselected blood donors. Italian controls (Milan) were unselected healthy individuals. Controls from The Netherlands were unselected blood donors and population controls.

SNP selection. All 1000 Genomes Project low-coverage whole-genome–sequencing pilot CEU variants within 0.1 cM of the lead SNP for each disease and region were selected. The September 2009 release comprising 60 CEU individuals was used (~5× mean read depth for whole-genome sequencing), and the markers selected were called in at least two of the Broad Institute, Sanger Institute and University of Michigan algorithms. Additional genomic region resequencing content was submitted for Immunochip analysis at specific loci from cases with celiac disease, Crohn's disease and type 1 diabetes and controls (**Supplementary Note**).

Genotyping. Samples were genotyped using the Immunochip according to Illumina's protocols (at labs in London, UK, Hinxton, UK, Groningen, The Netherlands, and Charlottesville, Virginia, USA). NCBI build 36 (hg18) mapping was used (Illumina manifest file Immuno_BeadChip_11419691_B.bpm).

Data quality control. Samples and variants with very low call rates were first excluded (after repeated testing of the samples). The Illumina GenomeStudio GenTrain2.0 algorithm was used to cluster an initial 2,000 UK samples. Subsequently, with additional sample data (case and control data were analyzed together), clusters were re-adjusted or excluded (manual or automated) for variants with low quality statistics (call rate <99.5%, a low GenCall score or many no calls with high intensity). This method produced better results than the GenoSNP or Illuminus clustering algorithms (data not shown). A cluster set based on 172,242 autosomal or X-chromosome variants (available on request) was then applied to all samples. Samples were excluded for call rate <99.5% across 172,242 markers. We then removed 15,657 non-polymorphic markers (that is, where only one of three expected genotype clouds was observed) that reflected a combination of ethnic-specific variants, allele-specific assay failure and substantial false-positive rates in early next-generation sequencing SNP calling algorithms.

Samples were excluded for incompatible recorded and genotype-inferred gender, duplicates and first- or second-degree relatives. Potential ethnic outliers were identified by multi-dimensional scaling plots of samples merged with HapMap3 data; the subset of SNPs common to HapMap3 and Immunochip accurately identified the different HapMap3 population samples. We considered the European and Indian collections separately.

Stepwise conditional logistic regression is sensitive to missing data and subtle genotyping error, so we therefore desired an ultra–high-quality dataset. Markers were excluded from all sample collections for deviation from Hardy-Weinberg equilibrium in controls ($P < 0.0001$) and/or differential missingness in no-call genotypes between cases and controls ($P < 0.001$) in any of the seven collections. Finally, we required a per-SNP call rate of >99.95% (a maximum of 12 no-call genotypes from 24,269 samples per autosomal marker), generating a data set of 139,553 markers (of which all but 372 indels are SNPs).

We visually inspected the intensity plot genotype clouds for all the markers listed in **Table 2** (as well as additional potential loci with $P < 1.9 \times 10^{-6}$) and confirmed all of these markers to be high quality. Genotype data has been deposited at the European Genome-Phenome Archive (see URLs), which is hosted by the European Bioinformatics Institute, under accession number EGAS00000000053.

Statistical analyses. Analyses were performed with PLINK v1.07 (ref. 38) using logistic regression tests with gender as a covariate and collection membership (**Table 1**) as a factorized covariate. Stepwise conditional logistic regression was performed in the order of markers with the smallest *P* value. Graphs were plotted in R (see URLs) and using a modified version of LocusZoom[39].

We permuted disease status for the dataset at each region to establish locus-wide statistical significance thresholds for defining independently associated SNPs. For each locus, defined by the LD boundaries (**Table 2**), we calculated the fifth percentile based on the nominal *P* value distribution for 1,000 permutations and controlling for multiple marker testing. This approach proved slightly more stringent than a per-locus Bonferroni correction for independent (using an estimate for independence as a pairwise $r^2 < 0.05$) variants (**Supplementary Table 3**). We estimated that our dataset contained 26,146 completely uncorrelated variants (using pairwise $r^2 < 0.05$ and a sliding 1,000-SNP window).

The fraction of additive variance was calculated using a liability threshold model[40] assuming a population prevalence of 1%. Effect sizes and control allele frequencies were estimated from the UK dataset. Genetic variance was calculated assuming 50% heritability.

34. Anonymous. Revised criteria for diagnosis of coeliac disease. Report of Working Group of European Society of Paediatric Gastroenterology and Nutrition. *Arch. Dis. Child.* **65**, 909–911 (1990).
35. Romanos, J.. *et al.* Six new coeliac disease loci replicated in an Italian population confirm association with coeliac disease. *J. Med. Genet.* **46**, 60–63 (2009).
36. Plaza-Izurieta, L. *et al.* Revisiting genome wide association studies (GWAS) in coeliac disease: replication study in Spanish population and expression analysis of candidate genes. *J. Med. Genet.* **48**, 493–496 (2011).
37. Megiorni, F. *et al.* HLA-DQ and risk gradient for celiac disease. *Hum. Immunol.* **70**, 55–59 (2009).
38. Purcell, S. *et al.* PLINK: a tool set for whole-genome association and population-based linkage analyses. *Am. J. Hum. Genet.* **81**, 559–575 (2007).
39. Pruim, R.J. *et al.* LocusZoom: regional visualization of genome-wide association scan results. *Bioinformatics* **26**, 2336–2337 (2010).
40. Risch, N.J. Searching for genetic determinants in the new millennium. *Nature* **405**, 847–856 (2000).

6.7.3 Izvirni znanstveni članek 8

Dense genotyping of immune-related disease regions identifies nine new risk loci for primary sclerosing cholangitis

Running title: Immunochip in PSC

Jimmy Z. Liu[1,*], Johannes Roksund Hov[2,3,4,5,*], Trine Folseraas[2,3,4, *], Eva Ellinghaus[6,*] , Simon M. Rushbrook[7], Nadezhda T. Doncheva[8], Ole A. Andreassen[4,9], Rinse K. Weersma[10], Tobias J. Weismüller[11,12,13], Bertus Eksteen[14], Pietro Invernizzi[15], Gideon M. Hirschfield[16], Daniel Nils Gotthardt[17] , Albert Pares[18], David Ellinghaus[6], Tejas Shah[1], Brian D. Juran[19], Piotr Milkiewicz[20], Christian Rust[21], Christoph Schramm[22], Tobias Müller[23], Brijesh Srivastava[24], Georgios Dalekos[25], Markus M. Nöthen[26,27], Stefan Herms[26,27], Juliane Winkelmann[28,29], Mitja Mitrovic[30], Felix Braun[31], Cyriel Y. Ponsioen[32], Peter J. P. Croucher[33], Martina Sterneck[34], Andreas Teufel[35], Andrew L. Mason[36], Janna Saarela[37], Virpi Leppa[38], Virpi Pelkonen[38], Ruslan Dorfman[39], Domenico Alvaro[40], Annarosa Floreani[41], Suna Onengut-Gumuscu[42], Stephen S. Rich[43], Wesley K. Thompson[44], Andrew J. Schork[45], Sigrid Næss[2,3,4], Ingo Thomsen[6], Gabriele Mayr[8], Inke R. König[46], Kristian Hveem[47], Isabelle Cleynen[48], Javier Gutierrez-Achury[30], Isis Ricaño-Ponce[30], David van Heel[49], Einar Björnsson[50], Richard N. Sandford[24], Peter R. Durie[51], Espen Melum[2,3,4], Morten H Vatn[4,5,52], Mark Silverberg[53], Richard H. Duerr[54,55], Leonid Padyukov[56], Stephan Brand[57], Miquel Sans[58], Vito Annese[59,60], Jean-Paul Achkar[61,62], Kirsten Muri Boberg[2,5], Hanns-Ulrich Marschall[63], Olivier Chazouillères[64], Christopher L. Bowlus[65], Cisca Wijmenga[30], Erik Schrumpf[2,4,5], Severine Vermeire[49,66], Mario Albrecht[8,67], The UK-PSC Consortium[68], The International IBD Genetics Consortium[68], John D. Rioux[69], Graeme Alexander[70], Annika Bergquist[71], Judy Cho[72], Stefan Schreiber[6,73,74], Michael P. Manns[11,12], Martti Färkkilä[75], Anders M. Dale[76,77,] Roger W. Chapman[78], Konstantinos N. Lazaridis[19], The International PSC Study Group[68], Andre Franke[6,§], Carl A. Anderson[1,§], Tom H. Karlsen[2,3,5,79,§]

[1] Wellcome Trust Sanger Institute, Wellcome Trust Genome Campus, Hinxton, Cambridge, UK,

[2] Norwegian PSC Research Center, Division of Cancer Medicine, Surgery and Transplantation, Oslo University Hospital, Rikshospitalet, Oslo, Norway,

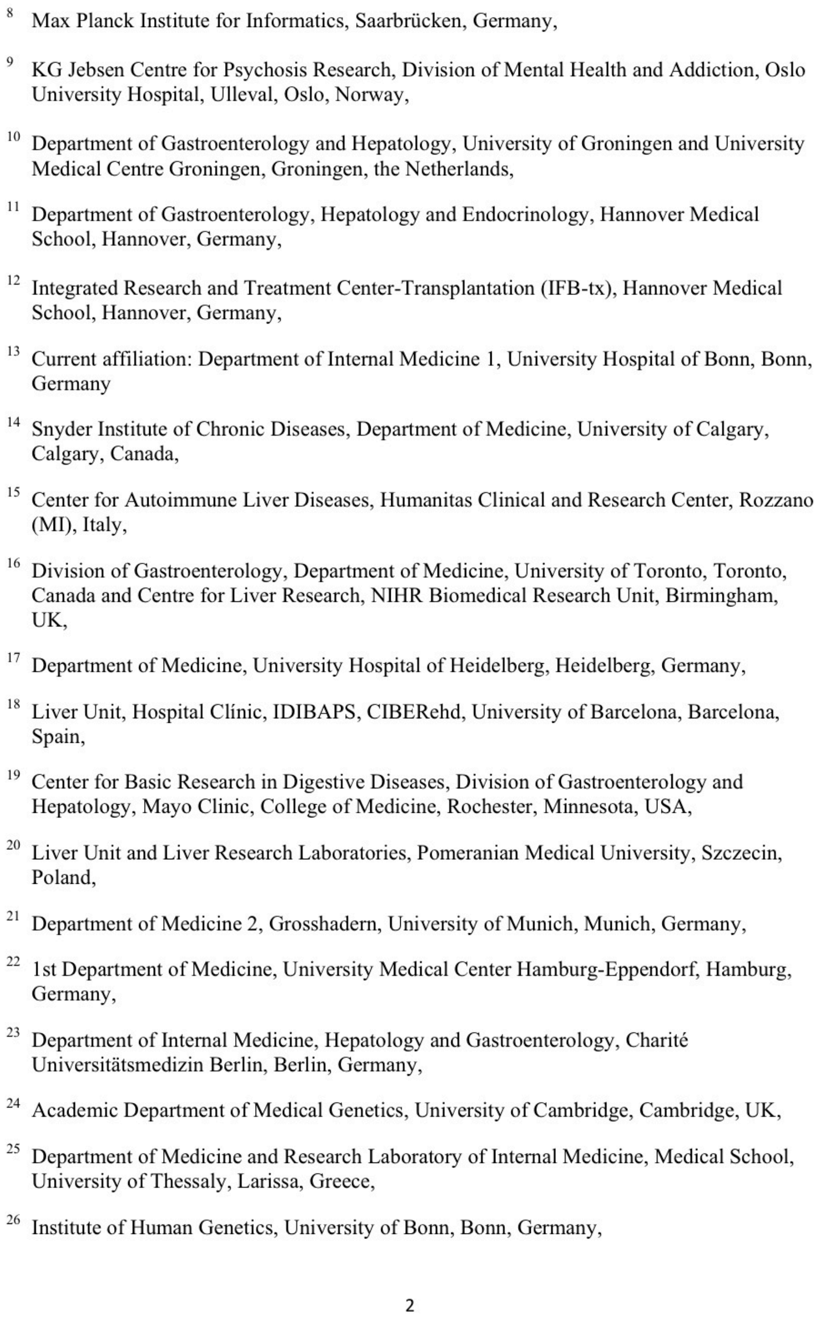
[3] Research Institute of Internal Medicine, Oslo University Hospital, Rikshospitalet, Oslo, Norway,

[4] Institute of Clinical Medicine, University of Oslo, Oslo, Norway,

[5] Section of Gastroenterology, Department of Transplantation Medicine, Division of Cancer, Surgery and Transplantation, Oslo University Hospital, Rikshospitalet, Oslo, Norway,

[6] Institute of Clinical Molecular Biology, Christian-Albrechts-University of Kiel, Kiel, Germany,

[7] Department of Gastroenterology and Hepatology, Norfolk and Norwich, University Hospitals NHS Trust, Norwich, UK,



[8] Max Planck Institute for Informatics, Saarbrücken, Germany,

[9] KG Jebsen Centre for Psychosis Research, Division of Mental Health and Addiction, Oslo University Hospital, Ulleval, Oslo, Norway,

[10] Department of Gastroenterology and Hepatology, University of Groningen and University Medical Centre Groningen, Groningen, the Netherlands,

[11] Department of Gastroenterology, Hepatology and Endocrinology, Hannover Medical School, Hannover, Germany,

[12] Integrated Research and Treatment Center-Transplantation (IFB-tx), Hannover Medical School, Hannover, Germany,

[13] Current affiliation: Department of Internal Medicine 1, University Hospital of Bonn, Bonn, Germany

[14] Snyder Institute of Chronic Diseases, Department of Medicine, University of Calgary, Calgary, Canada,

[15] Center for Autoimmune Liver Diseases, Humanitas Clinical and Research Center, Rozzano (MI), Italy,

[16] Division of Gastroenterology, Department of Medicine, University of Toronto, Toronto, Canada and Centre for Liver Research, NIHR Biomedical Research Unit, Birmingham, UK,

[17] Department of Medicine, University Hospital of Heidelberg, Heidelberg, Germany,

[18] Liver Unit, Hospital Clínic, IDIBAPS, CIBERehd, University of Barcelona, Barcelona, Spain,

[19] Center for Basic Research in Digestive Diseases, Division of Gastroenterology and Hepatology, Mayo Clinic, College of Medicine, Rochester, Minnesota, USA,

[20] Liver Unit and Liver Research Laboratories, Pomeranian Medical University, Szczecin, Poland,

[21] Department of Medicine 2, Grosshadern, University of Munich, Munich, Germany,

[22] 1st Department of Medicine, University Medical Center Hamburg-Eppendorf, Hamburg, Germany,

[23] Department of Internal Medicine, Hepatology and Gastroenterology, Charité Universitätsmedizin Berlin, Berlin, Germany,

[24] Academic Department of Medical Genetics, University of Cambridge, Cambridge, UK,

[25] Department of Medicine and Research Laboratory of Internal Medicine, Medical School, University of Thessaly, Larissa, Greece,

[26] Institute of Human Genetics, University of Bonn, Bonn, Germany,

[27] Department of Genomics, Life & Brain Center, University of Bonn, Bonn, Germany,

[28] Institute of Human Genetics and Department of Neurology, Technische Universität München, Munich, Germany,

[29] Institute of Human Genetics, Helmholtz Zentrum München-German Research Center for Environmental Health, Neuherberg, Germany,

[30] Department of Genetics, University of Groningen and University Medical Centre Groningen, Groningen, The Netherlands,

[31] Department of General, Visceral, Thoracic, Transplantation and Pediatric Surgery, University Medical Centre Schleswig-Holstein, Campus Kiel, Germany,

[32] Department of Gastroenterology and Hepatology, Academic Medical Center, Amsterdam, the Netherlands,

[33] Department of Environmental Science, Policy, and Management, University of California, Berkeley, United States of America,

[34] Department of Hepatobiliary Surgery and Transplantation, University Medical Center Hamburg-Eppendorf, Hamburg, Germany,

[35] 1st Department of Medicine, University of Mainz, Mainz, Germany,

[36] Division of Gastroenterology and Hepatology, University of Alberta, Edmonton, Alberta, Canada,

[37] Institute for Molecular Medicine Finland (FIMM), University of Helsinki, Helsinki, Finland,

[38] Public Health Genomics Unit, Institute for Molecular Medicine Finland FIMM, University of Helsinki and National Institute for Health and Welfare, Helsinki, Finland,

[39] Program in Genetics and Genome Biology, Hospital for Sick Children, Toronto, Canada,

[40] Department of Clinical Medicine, Division of Gastroenterology, Sapienza University of Rome, Rome, Italy

[41] Dept. of Surgical, Oncological and Gastroenterological Sciences, University of Padova, Padova, Italy,

[42] Center for Public Health Genomics and Department of Internal Medicine, Division of Endocrinology & Metabolism, University of Virginia, Charlottesville, USA,

[43] Center for Public Health Genomics and Department of Public Health Sciences, University of Virginia, Charlottesville, USA,

[44] Department of Psychiatry, University of California, San Diego, La Jolla, CA, USA,

[45] Graduate Program in Cognitive Science, University of California, San Diego, La Jolla, CA, USA

[46] Institute of Medical Biometry and Statistics, University of Lübeck, Lübeck, Germany,

[47] Department of Public Health, Faculty of Medicine, Norwegian University of Science and Technology, Trondheim, Norway

[48] Department of Clinical and Experimental Medicine, KU Leuven, Leuven, Belgium,

[49] Blizard Institute, Barts and The London School of Medicine and Dentistry, Queen Mary University of London, London, UK,

[50] Department of Internal Medicine, Division of Gastroenterology and Hepatology, Landspitali University Hospital, Reykjavik, Iceland,

[51] Physiology and Experimental Medicine, Research Institute, Hospital for Sick Children, Toronto, Ontario, Canada,

[52] EpiGen, Campus AHUS, Akershus University Hospital, Nordbyhagen, Norway,

[53] Inflammatory Bowel Disease (IBD) Group, Zane Cohen Centre for Digestive Diseases, Mount Sinai Hospital Toronto, Ontario, Canada,

[54] Division of Gastroenterology, Hepatology, and Nutrition, Department of Medicine, University of Pittsburgh School of Medicine, Pittsburgh, Pennsylvania, USA,

[55] Department of Human Genetics, Graduate School of Public Health, University of Pittsburgh, Pittsburgh, Pennsylvania, USA,

[56] Rheumatology Unit, Department of Medicine, Karolinska Institutet and Karolinska University Hospital Solna, Stockholm, Sweden,

[57] Department of Medicine II, University Hospital Munich-Grosshadern, Ludwig-Maximilians-University Munich, Germany,

[58] Department of Digestive Diseases, Centro Médico Teknon, Barcelona, Spain,

[59] Division of Gastroenterology, Istituto di Ricovero e Cura a Carattere Scientifico-Casa Sollievodella Sofferenza Hospital, San Giovanni Rotondo, Italy,

[60] Unit of Gastroenterology SOD2, Azienda Ospedaliero Universitaria Careggi, Florence, Italy,

[61] Department of Gastroenterology and Hepatology, Digestive Disease Institute, Cleveland Clinic, Cleveland, OH, USA,

[62] Department of Pathobiology, Lerner Research Institute, Cleveland Clinic, Cleveland, OH, USA,

[63] Department of Internal Medicine, Institute of Medicine, Sahlgrenska Academy and University Hospital, Gothenburg, Sweden

[64] AP-HP, Hôpital Saint Antoine, Department of Hepatology, UPMC Univ Paris 06, Paris, France,

[65] Division of Gastroenterology and Hepatology, University of California Davis, Davis, CA, USA,

[66] Department of Gastroenterology, University Hospitals Leuven, Leuven, Belgium,

[67] Department of Bioinformatics, Institute of Biometrics and Medical Informatics, University Medicine Greifswald, Greifswald, Germany,

[68] A full list of consortium members is provided in the Supplementary Material,

[69] Université de Montréal and the Montreal Heart Institute, Research Center, Montreal, Quebec, Canada,

[70] Department of Medicine, Division of Hepatology, University of Cambridge, Cambridge, UK,

[71] Department of Gastroenterology and Hepatology, Karolinska University Hospital Huddinge, Karolinska Institutet, Stockholm, Sweden,

[72] Department of Medicine, Section of Digestive Diseases, Yale University, New Haven, Connecticut, USA,

[73] Department for General Internal Medicine, Christian-Albrechts-University, Kiel, Germany,

[74] Popgen Biobank, University Hospital Schleswig-Holstein, Christian-Albrechts-University, 24105 Kiel, Germany,

[75] Division of Gastroenterology, Department of Medicine, Helsinki University Hospital, Finland,

[76] Department of Radiology, University of California, San Diego, La Jolla, CA, USA,

[77] Department of Neurosciences, University of California, San Diego, La Jolla, CA, USA,

[78] Department of Hepatology, John Radcliffe University Hospitals NHS Trust, Oxford, UK,

[79] Division of Gastroenterology, Institute of Medicine, University of Bergen, Bergen, Norway,

* These authors contributed equally to this work
[§] Shared senior authorship

[†]**Contact Information:**
Prof. Tom H. Karlsen, MD, PhD., Norwegian PSC Research Center,

Division of Cancer Medicine, Surgery and Transplantation, Oslo University Hospital
Rikshospitalet, Postboks 4950 Nydalen, N-0424 Oslo, Norway

Tel.: +47 23 07 2469; *Fax*: +47 2307 3928; *E-mail*: t.h.karlsen@medisin.uio.no

Dr Carl. A. Anderson, PhD.,

Wellcome Trust Sanger Institute, Wellcome Trust Genome Campus, Hinxton,
Cambridgeshire, CB10 1SA, United Kingdom,

Tel.: +44 1223 492371; *Fax*: ++44 1223 496826; *Email*: carl.anderson@sanger.ac.uk

Key words: genetic association study, disease genetics, immunogenetics, liver

Introductory paragraph

Primary sclerosing cholangitis (PSC) is a severe liver disease of unknown etiology leading to fibrotic destruction of the bile ducts and ultimately the need for liver transplantation[1-3]. We compared 3,789 European ancestry PSC cases to 25,079 population controls across 130,422 single-nucleotide polymorphisms (SNPs) genotyped using Immunochip[4]. We identified 12 genome-wide significant associations outside the human leukocyte antigen (HLA) complex, nine of which were novel, thereby increasing the number of known PSC risk loci to 16. Despite comorbidity with inflammatory bowel disease (IBD) in 72% of the patients, six of the 12 loci showed significantly stronger association with PSC than IBD, suggesting an overlapping yet distinct genetic architecture. We incorporated pleiotropy with seven diseases clinically co-occurring with PSC and found suggestive evidence for 33 additional PSC risk loci. Together with network analyses, these findings further complete the genetic risk map of PSC and considerably expand on the relationship between PSC and other immune-mediated diseases.

Text body

The pathogenesis of PSC is poorly understood, and due to lack of effective medical therapy, PSC remains a leading indication for liver transplantation in Northern Europe and the US[5], despite the relatively low prevalence (1/10,000). Affected individuals are diagnosed at a median age of 30-40 years and suffer from an increased frequency of IBD (60-80%)[5,6] and autoimmune diseases (25%)[7]. Conversely, approximately only 5% of patients with IBD develop PSC[5,6]. A 9-39-fold sibling relative risk indicates a strong genetic component to PSC risk[8]. In addition to multiple strong associations within the HLA complex, recent association studies have identified genome-wide significant loci at 1p36 (*MMEL1/TNFRSF14*), 2q13 (*BCL2L11*), 2q37 (*GPR35*), 3p21 (*MST1*), 10p15 *(IL2RA)* and 18q21 (*TCF4*)[9-13].

Several theories have been proposed to explain the development of PSC[5]. The strong HLA associations and the clinical co-occurrence of immune-mediated diseases suggest that autoimmunity plays a role. To further characterize the genetic etiology of PSC, we recruited PSC patients throughout Europe and North America, more than doubling the number of ascertained cases included in previous genetic studies[11]. We genotyped 196,524 SNPs in 4,228 PSC cases and 27,077 population controls (see **Online Methods** and **Supplementary**

Methods) using the Immunochip[4,14], a targeted genotyping array, not aiming for genome-wide coverage, with dense marker coverage across 186 known disease loci from 12 immune-mediated diseases (see **Online Methods**). Outside these 186 loci, Immunochip also assays thousands of SNPs of intermediate significance from multiple meta-analyses of immune-mediated diseases.

Following quality control (QC; see **Online Methods**), 130,422 SNPs from 3,789 PSC cases and 25,079 population controls were available for analysis (**Supplementary Tables 1** and **2, Supplementary Figures 1** and **2**). We imputed a further 80,183 SNPs located in the Immunochip fine mapping regions using the 1000 Genomes reference panel (**Online Methods**). We performed case-control association tests using a linear mixed model as implemented in MMM[15] to minimize the effect of population stratification and sample relatedness ($\lambda_{GC} = 1.02$, estimated using 2,544 "null" SNPs, see **Online Methods**).

We identified twelve non-HLA genome-wide significant ($P<5\times10^{-8}$) susceptibility loci (**Table 1**), nine of which were novel (**Fig. 1**). The most associated SNP within each locus was a common variant (all risk allele frequencies >0.18) of moderate effect (odds ratios (ORs) between 1.15 and 1.4) (**Table 1**). Imputed genotypes and stepwise conditional regressions[16] within each locus did not identify additional independent genome-wide significant signals, nor did genotype-genotype or gender-genotype interaction analyses.

For seven of the nine novel loci, the most significantly associated SNP in the locus was the same SNP or was in strong linkage disequilibrium (LD; $r^2>0.8$) with the original association reports for another disease (**Supplementary Table 3**). The two exceptions were 11q23, where only independent disease associations ($r^2<0.01$) have so far been reported[17], and 6q15, where the most significantly associated PSC variant, rs56258221 (OR=1.23, $P=8.36\times10^{-12}$), is in low-to-moderate LD with the previously reported *BACH2* variants in Crohn's disease ($r^2=0.23$) and type 1 diabetes ($r^2=0.12$). Three out of four known non-HLA PSC risk loci present on the Immunochip passed genotyping QC and were confirmed in our analysis (1p36, 3p21 and 10p15; see **Supplementary Results** and **Supplementary Fig. 3**).

To prioritize candidate genes within the non-HLA genome-wide significant loci, we searched for functional consequences of the most associated SNPs or SNPs in high LD (r^2>0.8), i.e. missense SNPs **(Supplementary Table 4** and **Supplementary Fig. 4**) and expression quantitative trait loci (eQTLs) (**Supplementary Table 5**), and we functionally annotated risk loci using data from the ENCODE project (**Supplementary Table 6** and **Supplementary Methods**)[18]. We also constructed networks based on functional similarity measures (**Supplementary Fig. 5** and **Online Methods**), known protein-protein interactions (DAPPLE[19], **Supplementary Table 7** and **Supplementary Methods**), and the published literature (GRAIL[20], **Supplementary Fig. 6** and **Supplementary Methods**) to identify important disease-relevant genes. For six of the 12 genome-wide significant loci, the same gene (*MMEL1*, *CD28*, *MST1*, *SH2B3*, *CD226* and *SIK2*) was annotated by more than one method (**Supplementary Table 7**), suggesting these as candidates for further investigation at these loci.

Two newly associated loci are located outside of the Immunochip fine mapping regions (**Figures 1d** and **1e**). At 11q23, the most strongly associated SNP, rs7937682 (OR=1.17, P=3.18×10^{-9}), is located in an intron of salt-inducible kinase 2 (*SIK2*), which both influences the expression of interleukin-10 in macrophages and Nur77, an important transcription factor in leukocytes[21]. The association at 12q13 is with an intronic SNP (rs11168249, OR=1.15, P=5.49×10^{-9}) within the histone deacetylase 7 (*HDAC7*) gene, which has also been associated with IBD[22]. HDAC7 has been implicated in negative selection of T cells in the thymus[23], a key factor in the development of immune tolerance. A role for *HDAC7* in PSC etiology is supported by the novel association at 19q13, where the most associated SNP, rs60652743 (OR=1.25, P=6.51×10^{-10}) is located within an intron of serine-threonine protein kinase D2 (*PRKD2*). When T cell receptors of thymocytes are engaged, PRKD2 phosphorylates HDAC7, leading to nuclear exclusion of HDAC7 and loss of its gene regulatory functions, ultimately resulting in apoptosis and negative selection of immature T cells[24,25]. Interestingly, this negative selection takes place due to a loss of HDAC7-mediated repression of Nur77 (regulated by SIK2)[26], linking three novel PSC loci to this pathway.

The associations at the HLA complex at 6p21 were refined by imputing alleles at *HLA-A, HLA-C, HLA-B, HLA-DRB1, HLA-DQB1, HLA-DQA* and *HLA-DPB1* (see **Supplementary Methods**)[27]. The top associated SNP (rs4143332) was in almost perfect LD (r^2=0.996) with

HLA-B*08:01 (**Supplementary Results**). In a stepwise conditional analysis including both SNP and HLA allele genotypes, rs4143332 (tagging HLA-B*08:01) and a complex HLA class II association signal determined by HLA-DQA1*01:03 and SNPs rs532098, rs1794282 and rs9263964 **(Supplementary Fig. 7)** explain most of the HLA association signal in PSC. When performing a stepwise regression of the HLA alleles only, apart from a novel association with HLA-DQA1*01:01, the class II associations are coherent with previous reports (see **Supplementary Results** and **Supplementary Tables 8-10**)[9,28,29]. The HLA-DRB1*15:01 association overlaps with that of ulcerative colitis (UC; risk increasing) and Crohn's disease (CD; risk decreasing)[30,31]. Since imputed genotypes at the class II region were only available for four (*HLA-DRB1, HLA-DQB1, HLA-DQA1* and *HLA-DPB1*) out of 20 loci[32], further studies involving direct sequencing of all HLA class II loci along with assessments of their protein structure and peptide binding are required to causally resolve the link between this HLA subregion and PSC development[33,34].

Although 72% of the PSC patients in this study have a diagnosis of concomitant IBD (**Supplementary Table 11)**, only half of our genome-wide significant loci were associated with IBD in the recent International IBD Genetics Consortium (IIBDGC) Immunochip analysis (**Fig. 2a, Supplementary Table 3** and **Supplementary Fig. 8**)[22], despite the greater sample size of that study (25,683 cases and 15,977 controls). Across the 12 non-HLA PSC loci we observed greater similarity between the OR estimates for PSC and UC than for PSC and CD. We used the CD and UC OR estimates for the 163 IBD-associated loci to predict PSC case/control status in our sample (**Online Methods**)[22], and found a significantly greater area under the receiver operating characteristic curve (AUC) when applying UC ORs compared to CD ORs (UC AUC=0.62, CD AUC=0.56, P=1.2x10^{-57}, **Fig. 2b**). This suggests that PSC is genetically more similar to UC than CD, and is consistent with clinical observations of greater comorbidity between PSC with UC than CD[35]. To further compare the genetic profile of PSC and IBD, we combined our genome-wide significant PSC loci with the 163 confirmed IBD loci[22] in a functional similarity network (**Fig. 2c**). The figure shows that the PSC loci are distributed throughout the IBD loci, suggesting that there is no particular functional subcluster of IBD susceptibility genes associated with PSC and vice versa.

While we consider only those loci reaching a stringent significance threshold (P<5x10^{-8}) to be conclusively associated to PSC, it is likely that additional true associations lie among SNPs with weaker associations. An alternative approach for controlling for multiple hypothesis testing is false discovery rate (FDR) control, which regulates the expected proportion of incorrectly rejected null hypotheses. FDR is well suited to focused genotyping platforms such as Metabochip[36] and Immunochip because it implicitly accounts for the expected enrichment of association. To further increase this enrichment, we exploited the known pleiotropy between related immune-mediated traits[37], and calculated the FDR[38-40] for association with PSC conditional on previously published summary statistics from each of the related phenotypes (yielding a per SNP conditional FDR)[41] (**Online Methods**). We identified 33 non-HLA loci with a conditional FDR<0.001 in this analysis (**Fig. 3**), all of which showed suggestive significance ($5x10^{-5} \leq P < 5x10^{-8}$) in the standard association analysis (**Supplementary Table 12** and **Supplementary Figures 9-11)**. These loci were integrated in the functional similarity network analysis (**Supplementary Fig. 12)**, highlighting potential candidate susceptibility genes.

In conclusion, the present study increases the number of genome-wide significant loci in PSC from seven to 16 (including the HLA complex). The nine novel variants together explain 0.9% of variance in PSC liability, increasing the total amount of variance explained by the 16 known loci to 7.3% (**Online Methods**). The data convincingly show that genetic susceptibility to PSC extends considerably beyond the risk factors involved in the closely related IBD phenotype and into autoimmune pathophysiology. Furthermore, analysis of pleiotropic immune-related genetic variants highlights 33 additional suggestive loci in PSC, overall representing major new avenues for research into disease pathogenesis.

URLs

1000 Genomes Omni2.5M genotype data:
(ftp://ftp.1000genomes.ebi.ac.uk/vol1/ftp/technical/working/20110921_phase2_omni_genotypes/Omni25_genotypes_1856_samples.b36.20110921.vcf.gz)
Mangrove (http://cran.r-project.org/web/packages/Mangrove/)
University of Chicago eQTL Browser (http://eqtl.uchicago.edu/cgi-bin/gbrowse/eqtl/)

Acknowledgements

We thank all PSC patients and healthy controls for their participation, and we are indebted to all physicians and nursing staff who recruited patients for this study. We thank TanjaWesse, Tanja Henke, Sanaz Sedghpour Sabet, Rainer Vogler, Gunnar Jacobs, Ilona Urbach, Kristian Holm, Hege Dahlen Sollid, Bente Woldseth, Jarl Andreas Anmarkrud and Liv Wenche Torbjørnsen for expert technical help. Prof. Ulrich Beuers, Dr. Felix Braun, Dr. Wolfgang Kreisel, Dr. Thomas Berg and Dr. Rainer Günther are acknowledged for contributing German PSC patients. Benedicte A. Lie and The Norwegian Bone Marrow Donor Registry at Oslo University Hospital, Rikshospitalet in Oslo and the Nord-Trøndelag Health Study (HUNT) are acknowledged for sharing the healthy Norwegian controls. HUNT is a collaboration between HUNT Research Centre (Faculty of Medicine, Norwegian University of Science and Technology NTNU), Nord-Trøndelag County Council, Central Norway Health Authority, and the Norwegian Institute of Public Health. Banco Nacional de ADN, Salamanca, Spain is acknowledged for providing Spanish controls samples.This study makes use of genotyping data generated by the DILGOM (http://www.aka.fi/en-GB/A/Programmes-and-Cooperation/Research-programmes/Ongoing/ELVIRA/Projects/DILGOM/), the Cooperative Research in the Region of Augsburg (KORA) study and by the Heinz Nixdorf Recall (Risk Factors, Evaluation of Coronary Calcification, and Lifestyle) study (HNR). We acknowledge the members of the International PSC Study Group, the NIDDK Inflammatory Bowel Disease Genetics Consortium (IBDGC), the UK-PSC Consortium and the Alberta IBD Consortium for their participation. John Todd, Neil Walker, Olly Burren and Jeffrey Barrett are acknowledged for the use of type 1 diabetes summary statistics. J.B. is also acknowledged for contributions to the design of the Immunochip experiment. Eli A. Stahl *et al.* is acknowledged for the use of rheumatoid arthritis summary statistics. We thank Matti Pirinen for sharing his MMM mixed linear model software ahead of publication. The study was supported by The Norwegian PSC Research Center (http://ous-research.no/nopsc/), by the German Ministry of Education and Research (BMBF) through the National Genome Research Network (NGFN;

01GS0809-GP7), by the Deutsche Forschungsgemeinschaft (DFG), grant no. FR 2821/2-1, by the EU Seventh Framework Programme FP7/2007-2013, grant number 262055, ESGI, by the Integrated Research and Treatment Center - Transplantation (reference number: 01EO0802) and the PopGen biobank (http://www.popgen.de).The project received infrastructure support through the DFG cluster of excellence “Inflammation at Interfaces”. The British 1958 Birth Cohort DNA collection was funded by Medical ResearchCouncil grant G0000934 and Wellcome Trust grant 068545/Z/02, and the UK National Blood Service controls by the Wellcome Trust. The Wellcome Trust Case Control Consortium projects were supported by Wellcome Trust grants 083948/Z/07/Z, 085475/B/08/Z and 085475/Z/08/Z. Genotyping of samples at the University of Virginia utilized resources provided by the Type 1 Diabetes Genetics Consortium, a collaborative clinical study sponsored by the National Institute of Diabetes and Digestive and Kidney Diseases (NIDDK), the National Institute of Allergy and Infectious Diseases (NIAID), the National Human Genome Research Institute (NHGRI), the National Institute of Child Health and Human Development (NICHD), and the Juvenile Diabetes Research Foundation International (JDRF) and is supported by U01-DK-062418.
The HNR study is supported by the Heinz Nixdorf Foundation (Germany). Additionally, the HNR study is funded by the German Ministry of Education and Science and the German Research Council (DFG; Project SI 236/8-1, SI236/9-1, ER 155/6-1). J.Z.L., T.S., C.A.A. are supported by a grant from the Wellcome Trust (098051). R.K.W. is supported by a clinical fellowship grant (90.700.281) from the Netherlands Organization for Scientific Research (NWO). B.E. receives support from the Medical Research Council, UK. T.M. is supported by the Deutsche Forschungsgemeinschaft, Grant MU 2864/1-1. D.G. was supported by a grant from the Deutsche Forschungsgemeinschaft. M.S. is funded by Fundació Miarnau, Barcelona, Spain. A.P. is supported by Centro de Investigación Biomédica en Red de Enfermedades Hepáticas y Digestivas (CIBERehd), and grant no. PI071318 Instituto de Salud Carlos III, Ministerio de Ciencia e Innovación, Spain. Grant support to P.D. was from the Canadian Institutes of Health research (CIHR) and Genome Canada. S.B. is supported by grants from the Deutsche Forschungsgemeinschaft (DFG), grant number BR 1912/6-1 and the Else-Kröner-Fresenius-Stiftung (Else Kröner-Exzellenzstipendium 2010-EKES.32). C.W. is supported by grants from the Celiac Disease Consortium (BSIK03009) and The Netherlands Organization for Scientific Research (NWO, VICI grant918.66.620). KNL is supported by a NIDDK grant (DK 84960).

Competing Interests Statement

The authors declare no competing interests.

Author contributions

J.Z.L., J.R.H., T.F., E.E., N.T.D., O.A.A., W.K.T, A.M.D., T.S. and C.A.A. performed data and statistical analysis. A.F., C.A.A. and T.H.K. coordinated the project and supervised the data-analysis. J.Z.L., J.R.H., T.F., E.E., A.F., C.A.A. and T.H.K. drafted the manuscript. All other authors contributed primarily to the patient ascertainment, sample collection, genotyping and/or clinical data. All authors revised the manuscript for critical content and approved the final version.

Figure legends

Figure 1. Association results across the nine newly associated primary sclerosing cholangitis (PSC) loci

Regional association plots of the 9 loci newly associated with PSC at genome-wide significance ($P<5\times10^{-8}$). Filled-in circles are genotyped and hollow triangles imputed (see **Online Methods**) single-nucleotide polymorphisms (SNPs). The color of the marker (see legends in panel **a)** illustrates linkage disequilibrium with the most associated SNP. Since the most associated SNPs in panels **d)** and **e)** are located outside Immunochip fine mapping regions, association results from the discovery panel of the largest PSC genome-wide association study to date[12] are shown as hollow circles and the most associated SNP is a hollow diamond (genotyped and imputed HapMap release 22 SNPs, cases overlap with the current study).

Figure 2. Genetic and functional similarity of primary sclerosing cholangitis (PSC) and inflammatory bowel disease (IBD) associated loci.

a) Comparison of odds ratios (ORs) between the most associated risk allele in PSC and the same allele in Crohn's disease (CD) and ulcerative colitis (UC) across the 12 genome-wide significant PSC-associated loci. UC and CD ORs, and the denomination of IBD loci as CD, UC or IBD, were obtained from Jostins *et al*[22]. [*]The PSC associated alleles at 6q15 (*BACH2*), 10p15 (*IL2RA*) and 19q13 (*PRKD2*) are independent of the reported IBD associations ($r^2<0.3$) but locate to the same broad genetic region; for this reason these loci are defined as PSC-IBD loci in panel **c)**.

b) Predicting PSC using OR estimates across 163 IBD-associated loci. The green and orange lines represent the receiver operating characteristic (ROC) curves for discriminating PSC cases from population controls using 163 UC or CD ORs, respectively[22]. The dashed diagonal line is $y=x$, and specifies the ROC curve of a random predictor.

c) Functional similarity network for PSC and IBD associated loci. The protein-coding genes closest to the most associated SNP in the 12 non-HLA genome-wide significant PSC loci and the 163 confirmed IBD loci[22] were used to construct a functional similarity network (see **Online Methods**). The network contains 90 gene nodes that are connected by 292 similarity edges. Genes associated with only PSC are represented by large red nodes, with only IBD by small green nodes, and with both PSC and IBD by large violet nodes. Grey edge lines indicate strong functional similarity between the connected genes based on their Gene Ontology annotations (see **Online Methods**). Genes and their nodes that are not connected to any other node in the network are omitted from the figure (see **Supplementary Table 13** for full listing).

Figure 3. Pleiotropic primary sclerosing cholangitis (PSC) loci

Manhattan plot of conditional associations in PSC calculated as stratified false discovery rates (FDRs) based on the results of the present PSC analysis and genetic associations previously reported in seven immune-mediated diseases (Crohn's disease, celiac disease, psoriasis, rheumatoid arthritis, sarcoidosis; type 1 diabetes and ulcerative colitis) (see **Online Methods** and **Supplementary Figures 9-11**). Single-nucleotide polymorphisms (SNPs) in red represent genome-wide significant findings from the regular association analysis, while SNPs listed in black are significantly associated with PSC conditional on their pleiotropic effects across the related immune-mediated diseases. The horizontal red line represents a threshold of FDR<0.001 (**Supplementary Table 12),** while the blue line represents a threshold of FDR<0.01 (see **Supplementary Table 14** for full listing).

Table 1: Association results of twelve non-HLA genome-wide significant risk loci for primary sclerosing cholangitis (PSC).

Chr	SNP[a]	RA	RAF cases	RAF controls	*P*-value	OR (95%CI)	LD region[b] (Kb)	RefSeq genes in LD region	Notable nearby gene(s)[c]	Functional annotation[d]
1p36	rs3748816	A	0.698	0.656	7.41×10^{-12}	1.21 (1.14-1.27)	2,398-2,775	9	*MMEL1, TNFRSF14*	eQTL,MS, OC, PB, HM
2q33	**rs7426056**	A	0.277	0.229	1.89×10^{-20}	1.3 (1.23-1.37)	204,155-204,397	1	*CD28*	HM, OC
3p21	rs3197999	A	0.352	0.285	2.45×10^{-26}	1.33 (1.26-1.4)	48,388-51,358	90	*MST1*	eQTL,MS, OC, PB HM
4q27	**rs13140464**	C	0.871	0.836	8.87×10^{-13}	1.3 (1.21-1.4)	123,204-123,784	4	*IL2, IL21*	OC, PB
6q15	**rs56258221**	G	0.213	0.183	8.36×10^{-12}	1.23 (1.16-1.31)	90,967-91,150	1	*BACH2*	OC, PB
10p15	rs4147359	A	0.401	0.349	8.19×10^{-17}	1.24 (1.18-1.3)	6,070-6,206	2	*IL2RA*	PB
11q23	**rs7937682**	G	0.298	0.265	3.17×10^{-09}	1.17 (1.11-1.24)	110,824-111,492	19	*SIK2*	OC, PB, HM
12q13	**rs11168249**	G	0.506	0.466	5.49×10^{-09}	1.15 (1.1-1.21)	46,442-46,534	3	*HDAC7*	OC, PB, HM
12q24	**rs3184504**	A	0.527	0.488	5.91×10^{-11}	1.18 (1.12-1.24)	110,186-111,512	16	*SH2B3, ATXN2*	MS, OC, HM
18q22	**rs1788097**	A	0.518	0.483	3.06×10^{-08}	1.15 (1.1-1.21)	65,633-65,721	2	*CD226*	MS, OC, PB, HM
19q13	**rs60652743**	A	0.864	0.836	6.51×10^{-10}	1.25 (1.16-1.34)	51,850-51,998	6	*PRKD2, STRN4*	OC, PB, HM
21q22	**rs2836883**	G	0.777	0.728	3.19×10^{-17}	1.28 (1.21-1.36)	39,374-39,404	-	*PSMG1*	OC, PB, HM

Chr: chromosome; CI: confidence interval; eQTL: expression quantitative trait locus, HM: overlaps a region of histone modification; Kb: kilobasepairs; LD: linkage disequilibrium; MS: missense mutation; OC: overlaps known region of open chromatin; OR: odds ratio; PB: overlaps a region of protein binding; RA: risk allele; RAF: risk allele frequency

[a]SNPs from novel PSC-associated loci are shown in bold. [b]LD regions around lead SNPs were calculated by extending in both directions a distance of 0.1 centimorgans as defined by the HapMap recombination map. [c]Candidate gene(s) within same LD region as the associated SNPs. [d]Denotes if there are SNPs with r^2>0.8 with the hit SNP that have functional annotations (**Supplementary Tables 4-7**).

Online Methods

Study Subjects. Recruitment of PSC patients was performed in 14 countries in Europe and North America (see **Supplementary Methods, Supplementary Table 15a** and **15b**). The diagnosis of PSC was based on standard clinical, biochemical, cholangiographic and histological criteria with exclusion of secondary causes of sclerosing cholangitis[42]. Controls were recruited from blood donors or population-based studies as part of this study or via the International Immunochip consortium (see **Supplementary Methods**).

Ethical approval. The patient recruitment was approved by the ethics committees or institutional review boards of all participating centers or countries. Written informed consent was obtained from all study participants.

Quality control. SNPs with a call rate <80% were removed prior to commencing sample QC (n=235). Per individual genotype call rate and heterozygosity rate were calculated using PLINK[43] and outlying samples were identified using Aberrant[44], which automatically identifies outliers from otherwise Gaussian distributions (**Supplementary Fig. 1**). A set of 20,837 LD-pruned (r^2<0.1) SNPs with MAF>10% present on both the Immunochip and the Illumina Omni2.5-8 array used in the 1000 Genomes project (see URLs) were used to estimate identity by descent and ancestry. For each pair with estimated identity by descent≥0.9, the lower call rate sample was removed (unless case/control status was discordant between the pair, in which case both samples were removed, n=92). Related individuals (0.1875<identity by descent<0.9) remained in the analysis to maximize power because the mixed model association analysis can correctly account for the relatedness between individuals. Principal components analysis, implemented in SMARTPCA (Eigenstrat)[45], was used to identify samples of non-European ancestry. Principal components were defined using population samples from the 1000 Genomes project[46] genotyped using the Illumina Omni2.5-8 genotyping array (see URLs) and then projected into cases and controls (**Supplementary Fig. 2**)[14,22,47]. Following sample QC, 3,789 PSC cases and 25,079 controls remained. SNPs with a minor allele frequency less than 0.1%, Hardy-Weinberg equilibrium $P<10^{-5}$, call rate lower than 98%, or failing the PLINK v1.07 non-random differential missing data rate test between cases and controls ($P<10^{-5}$) were excluded. After completion of marker QC (**Supplementary Table 2**), 131,220 SNPs were available for analysis, further reduced to 130,422 after cluster plot inspection (see below).

19

Statistical methods.

Genomic inflation factor. The Immunochip contains 3,120 SNPs that were part of a bipolar disease replication effort and other non-immune-related studies. After QC, 2,544 of these were used as null markers to estimate the overall inflation of the distribution of association test statistics.

Imputation. Using 85,747 post-QC SNPs located in the Immunochip fine mapping regions, additional genotypes were imputed using IMPUTE2 with the 1000 Genomes Phase 1 (March, 2012) reference panel of 1,092 individuals[48] and 744,740 SNPs. Imputation was performed separately in ten batches (nine batches of 2,887 individuals each, and one of 2885 individuals), with the case:control and country of origin ratios constant across batches. SNPs with a posterior probability less than 0.9 and those with differential missingness ($P<10^{-5}$) between the 10 batches were removed, as were those SNPs that failed the same exclusion thresholds used for the original SNP QC. After imputation, a total of 163,379 SNPs in the Immunochip fine mapping regions, including 153,857 SNPs from the reference panel, were available for analysis.

Association analysis. Case-control association tests were performed using a linear mixed model as implemented in MMM[15]. A covariance matrix, *R*, of a random effects component was included in the model to explicitly account for confounding due to relatedness between individuals. This method has been shown to better control for population stratification than correction for principal components or meta-analyses of matched subgroups of cases and controls [49-51]. *R* is a symmetric $n \times n$ matrix with each entry representing the relative sharing of alleles between two individuals compared to the average in the sample, and is typically estimated using genome-wide SNP data[15]. To avoid biases in the estimation of *R* due to the design of the Immunochip, SNPs were first pruned for LD ($r^2<0.1$). Of the remaining SNPs, we then removed those that lie in the HLA region or have a minor allele frequency<10%. Finally, we excluded SNPs that showed modest association ($P<0.005$) with PSC in a linear regression model fitting the first 10 principal components as covariates. A total of 17,260 SNPs were used to estimate *R*.

Due to computational limitations, we estimated the *R* matrix and performed all association analyses applying *R* separately for UK (n=9,696) and non-UK (n=19,172) samples, and then combined the results using a fixed-effects (inverse-variance weighting) meta-analysis, as done previously[49]. This reduced the $\lambda_{GC,}$ estimated using the 2,544 "null" SNPS, from 1.24 to 1.02 (**Supplementary Fig. 13**), showing excellent control for population stratification. Stepwise conditional regression was used to identify possible independent associations at genome-wide significant loci. SNP×SNP interactions between all pairs of genome-wide significant SNPs were tested using the PLINK --epistasis command. Signal intensity plots of all non-HLA loci with association P-value$<5\times10^{-6}$ were visually inspected using Evoker[52]. SNPs that clustered poorly were removed (n=798).

Prediction of PSC using IBD SNPs. ORs for CD and UC in 163 IBD-associated SNPs were obtained from Jostins *et al.*[22]. We used the R package Mangrove (see URLs) to estimate each individual's probability of developing PSC among our 3,789 PSC cases and 25,079 controls assuming additive risk (log-additive OR). The performance of our predictor using either CD or UC ORs, was assessed by constructing a ROC curve, which shows the proportion of true and false positives at each probability threshold. The AUC was calculated to compare the predictive power of the UC and CD ORs. The DeLong method was used to test if the AUC using UC ORs was significantly greater than the AUC using CD ORs[53].

Functional similarity networks were generated for different sets of genes. In functional similarity networks, each network edge represents strong functional similarity of two genes based on the similarity of annotated Gene Ontology terms as determined by the functional similarity measure rfunSim, which considers Biological Process and Molecular Function ontologies[54]. The rfunSim similarity values above the recommended cutoff 0.8 were retrieved using the FunSimMat web service[55]. The resulting networks were visualized and analyzed using Cytoscape[56].

To construct PSC-specific networks from the functional similarity networks that contain more than one gene per locus (**Supplementary Figures 5** and **12**), the connectivity of each gene was assessed by computing three different topology measures for the corresponding node: (1) degree (number of direct edges to other nodes), (2) shortest path closeness (inverted average shortest path distance to all other nodes) and (3) shortest path betweenness (the fraction of shortest paths that pass through the node). Similarity edges between genes in the same locus as well as gene nodes that were not contained in the resulting largest connected subnetworks were ignored. The genes were first ranked according to each measure and then assigned the best of the three resulting ranks. The PSC-specific network was generated from the top ranked genes in their respective locus.

Pleiotropy analysis. We included summary statistics from genome-wide association studies in seven PSC-associated diseases (CD, celiac disease, psoriasis, rheumatoid arthritis, sarcoidosis, type 1 diabetes, UC, see **Supplementary Table 16**). For all diseases we constructed conditional stratified Q-Q plots of empirical quantiles of nominal $-\log_{10}(P)$ values for SNP association with PSC for all SNPs (see **Supplementary Fig. 9**), and for different overlapping subsets of SNPs determined by the significance of their association with the PSC-associated autoimmune disorder (SNP subsets defined by $P<1$, $P<0.1$, $P<0.01$ and $P<0.001$ in the pleiotropic phenotype, respectively). For a given PSC associated phenotype, 'enrichment' for pleiotropic signals in PSC can be observed as an increasing leftward deflection from the expected null line with lower *P*-value thresholds in the second phenotype (**Supplementary Methods**). The 'enrichment' in the stratified Q-Q plots is directly interpretable in terms of the true discovery rate (TDR), equivalent to one minus the FDR[57]. Specifically, it can be shown that a conservative estimate of FDR can be calculated from the horizontal shift of the Q-Q curve from the expected line x=y, with a larger shift corresponding to a smaller FDR for a given nominal *P*-value (see **Supplementary Methods**). We calculated the conditional TDR as a function of *P*-value in PSC across a series of *P*-value thresholds in the pleiotropic trait (**Supplementary Fig. 9**).

In order to assess significance of the association with PSC, we assigned a pleiotropic (conditional) FDR value for PSC per SNP. The pleiotropic FDR value for each SNP is based on the *P*-value of the SNP in PSC relative to the *P*-value distribution of other SNPs in the same conditioning subset, where subsets are defined by the pleiotropic association (lowest *P*-value among associated diseases) of the SNP. Importantly, the conditioning procedure is blind to the *P*-value of the SNP with respect to PSC. The pleiotropic FDR is then interpolated from conditional FDR curves using established stratified FDR methods[41,58] (see **Supplementary Methods**). The increase in power from using pleiotropic FDR is demonstrated by dividing the total sample in half and observing that empirical replication rates between the training and test halves increases with decreasing *P*-value in the pleiotropic disease (**Supplementary Fig. 14**). The SNP with lowest FDR within each LD block (as defined by 1000 Genomes) was considered the lead SNP of a new pleiotropic PSC locus, if below a 0.001 threshold (loci defined by FDR<0.001 and FDR<0.01 shown in **Supplementary Tables 12 and 14**). All test

statistics were adjusted for population stratification by genomic control (see **Supplementary Methods**).

Variance explained and heritability. The proportion of variance explained by the genome-wide significant loci and HLA alleles was calculated using a liability threshold model[59] assuming a disease prevalence of 10/100,000 and multiplicative risk.

References

1. Aadland, E. *et al.* Primary sclerosing cholangitis: a long-term follow-up study. *Scand J Gastroenterol* **22**, 655-64 (1987).
2. Broome, U. *et al.* Natural history and prognostic factors in 305 Swedish patients with primary sclerosing cholangitis. *Gut* **38**, 610-5 (1996).
3. Farrant, J.M. *et al.* Natural history and prognostic variables in primary sclerosing cholangitis. *Gastroenterology* **100**, 1710-7 (1991).
4. Cortes, A. & Brown, M.A. Promise and pitfalls of the Immunochip. *Arthritis Res Ther* **13**, 101 (2011).
5. Karlsen, T.H., Schrumpf, E. & Boberg, K.M. Update on primary sclerosing cholangitis. *Dig Liver Dis* **42**, 390-400 (2010).
6. Karlsen, T.H. & Kaser, A. Deciphering the genetic predisposition to primary sclerosing cholangitis. *Semin Liver Dis* **31**, 188-207 (2011).
7. Saarinen, S., Olerup, O. & Broome, U. Increased frequency of autoimmune diseases in patients with primary sclerosing cholangitis. *Am J Gastroenterol* **95**, 3195-9 (2000).
8. Bergquist, A. *et al.* Increased risk of primary sclerosing cholangitis and ulcerative colitis in first-degree relatives of patients with primary sclerosing cholangitis. *Clin Gastroenterol Hepatol* **6**, 939-43 (2008).
9. Karlsen, T.H. *et al.* Genome-wide association analysis in primary sclerosing cholangitis. *Gastroenterology* **138**, 1102-11 (2010).
10. Srivastava, B. *et al.* Fine mapping and replication of genetic risk loci in primary sclerosing cholangitis. *Scand J Gastroenterol* **47**, 820-6 (2012).
11. Folseraas, T. *et al.* Extended analysis of a genome-wide association study in primary sclerosing cholangitis detects multiple novel risk loci. *J Hepatol* **57**, 366-75 (2012).
12. Melum, E. *et al.* Genome-wide association analysis in primary sclerosing cholangitis identifies two non-HLA susceptibility loci. *Nat Genet* **43**, 17-9 (2011).
13. Ellinghaus, D. *et al.* Genome-wide association analysis in sclerosing cholangitis and ulcerative colitis identifies risk loci at GPR35 and TCF4. *Hepatology* (2012).
14. Trynka, G. *et al.* Dense genotyping identifies and localizes multiple common and rare variant association signals in celiac disease. *Nat Genet* **43**, 1193-201 (2011).
15. Pirinen, M., Donnelly, P. & Spencer, C. Efficient computation with a linear mixed model on large-scale data sets with applications to genetic studies. *Ann Appl Stat* **In press**(2012).
16. Cordell, H.J. & Clayton, D.G. A unified stepwise regression procedure for evaluating the relative effects of polymorphisms within a gene using case/control or family data: application to HLA in type 1 diabetes. *Am J Hum Genet* **70**, 124-41 (2002).

17. Peters, U. *et al.* Meta-analysis of new genome-wide association studies of colorectal cancer risk. *Hum Genet* **131**, 217-34 (2012).

18. Gerstein, M.B. *et al.* Architecture of the human regulatory network derived from ENCODE data. *Nature* **489**, 91-100 (2012).

19. Rossin, E.J. *et al.* Proteins encoded in genomic regions associated with immune-mediated disease physically interact and suggest underlying biology. *PLoS Genet* **7**, e1001273 (2011).

20. Raychaudhuri, S. *et al.* Identifying relationships among genomic disease regions: predicting genes at pathogenic SNP associations and rare deletions. *PLoS Genet* **5**, e1000534 (2009).

21. Hanna, R.N. *et al.* The transcription factor NR4A1 (Nur77) controls bone marrow differentiation and the survival of Ly6C- monocytes. *Nat Immunol* **12**, 778-85 (2011).

22. Jostins, L. *et al.* Host-microbe interactions have shaped the genetic architecture of inflammatory bowel disease. *Nature* **491**, 119-24 (2012).

23. Kasler, H.G. *et al.* Histone deacetylase 7 regulates cell survival and TCR signaling in CD4/CD8 double-positive thymocytes. *J Immunol* **186**, 4782-93 (2011).

24. Dequiedt, F. *et al.* HDAC7, a thymus-specific class II histone deacetylase, regulates Nur77 transcription and TCR-mediated apoptosis. *Immunity* **18**, 687-98 (2003).

25. Dequiedt, F. *et al.* Phosphorylation of histone deacetylase 7 by protein kinase D mediates T cell receptor-induced Nur77 expression and apoptosis. *J Exp Med* **201**, 793-804 (2005).

26. Clark, K. *et al.* Phosphorylation of CRTC3 by the salt-inducible kinases controls the interconversion of classically activated and regulatory macrophages. *Proc Natl Acad Sci U S A* (2012).

27. Raychaudhuri, S. *et al.* Five amino acids in three HLA proteins explain most of the association between MHC and seropositive rheumatoid arthritis. *Nat Genet* **44**, 291-6 (2012).

28. Schrumpf, E. *et al.* HLA antigens and immunoregulatory T cells in ulcerative colitis associated with hepatobiliary disease. *Scand J Gastroenterol* **17**, 187-91 (1982).

29. Spurkland, A. *et al.* HLA class II haplotypes in primary sclerosing cholangitis patients from five European populations. *Tissue Antigens* **53**, 459-69 (1999).

30. Stokkers, P.C., Reitsma, P.H., Tytgat, G.N. & van Deventer, S.J. HLA-DR and -DQ phenotypes in inflammatory bowel disease: a meta-analysis. *Gut* **45**, 395-401 (1999).

31. Okada, Y. *et al.* HLA-Cw*1202-B*5201-DRB1*1502 haplotype increases risk for ulcerative colitis but reduces risk for Crohn's disease. *Gastroenterology* **141**, 864-871 e1-5 (2011).

32. Horton, R. *et al.* Gene map of the extended human MHC. *Nature* **5**, 889-99 (2004).

33. Hov, J.R. *et al.* Electrostatic modifications of the human leukocyte antigen-DR P9 peptide-binding pocket and susceptibility to primary sclerosing cholangitis. *Hepatology* **53**, 1967-76 (2011).

34. Hovhannisyan, Z. *et al.* The role of HLA-DQ8 beta57 polymorphism in the anti-gluten T-cell response in coeliac disease. *Nature* **456**, 534-8 (2008).

35. Broome, U. & Bergquist, A. Primary sclerosing cholangitis, inflammatory bowel disease, and colon cancer. *Semin Liver Dis* **26**, 31-41 (2006).

36. The, C.D.C. *et al.* Large-scale association analysis identifies new risk loci for coronary artery disease. *Nat Genet* (2012).

37. Zhernakova, A., van Diemen, C.C. & Wijmenga, C. Detecting shared pathogenesis from the shared genetics of immune-related diseases. *Nat Rev Genet* **10**, 43-55 (2009).

38. Benjamini, Y. & Hochberg, Y. Controlling the False Discovery Rate: A Practical and Powerful Approach to Multiple Testing. *J Roy Statist Soc Ser B* **57**, 289-300 (1995).

39. Storey, J.D. The positive false discovery rate: A Bayesian interpretation and the q-value. *Ann Statist* **31**, 2013-2035 (2003).

40. Efron, B. Simultaneous Inference: When Should Hypothesis Testing Problems Be Combined? *Ann Appl Statist* **2**, 197-223 (2008).

41. Sun, L., Craiu, R.V., Paterson, A.D. & Bull, S.B. Stratified false discovery control for large-scale hypothesis testing with application to genome-wide association studies. *Genet Epidemiol* **30**, 519-30 (2006).

42. Chapman, R.W. *et al.* Primary sclerosing cholangitis: a review of its clinical features, cholangiography, and hepatic histology. *Gut* **21**, 870-7 (1980).

43. Purcell, S. *et al.* PLINK: a tool set for whole-genome association and population-based linkage analyses. *Am J Hum Genet* **81**, 559-75 (2007).

44. Bellenguez, C., Strange, A., Freeman, C., Donnelly, P. & Spencer, C.C. A robust clustering algorithm for identifying problematic samples in genome-wide association studies. *Bioinformatics* **28**, 134-5 (2012).

45. Patterson, N., Price, A.L. & Reich, D. Population structure and eigenanalysis. *PLoS Genet* **2**, e190 (2006).

46. 1000 Genomes Project Consortium. A map of human genome variation from population-scale sequencing. *Nature* **467**, 1061-73 (2010).

47. Liu, J.Z. *et al.* Dense fine-mapping study identifies new susceptibility loci for primary biliary cirrhosis. *Nat Genet* (2012).

48. Howie, B.N., Donnelly, P. & Marchini, J. A flexible and accurate genotype imputation method for the next generation of genome-wide association studies. *PLoS Genet* **5**, e1000529 (2009).

49. Sawcer, S. *et al.* Genetic risk and a primary role for cell-mediated immune mechanisms in multiple sclerosis. *Nature* **476**, 214-9 (2011).

50. Korte, A. *et al.* A mixed-model approach for genome-wide association studies of correlated traits in structured populations. *Nat Genet* **44**, 1066-71 (2012).

51. Tsoi, L.C. *et al.* Identification of 15 new psoriasis susceptibility loci highlights the role of innate immunity. *Nat Genet* **44**, 1341-8 (2012).

52. Morris, J.A., Randall, J.C., Maller, J.B. & Barrett, J.C. Evoker: a visualization tool for genotype intensity data. *Bioinformatics* **26**, 1786-7 (2010).

53. DeLong, E.R., DeLong, D.M. & Clarke-Pearson, D.L. Comparing the areas under two or more correlated receiver operating characteristic curves: a nonparametric approach. *Biometrics* **44**, 837-45 (1988).

54. Schlicker, A., Domingues, F.S., Rahnenfuhrer, J. & Lengauer, T. A new measure for functional similarity of gene products based on Gene Ontology. *BMC Bioinformatics* **7**, 302 (2006).

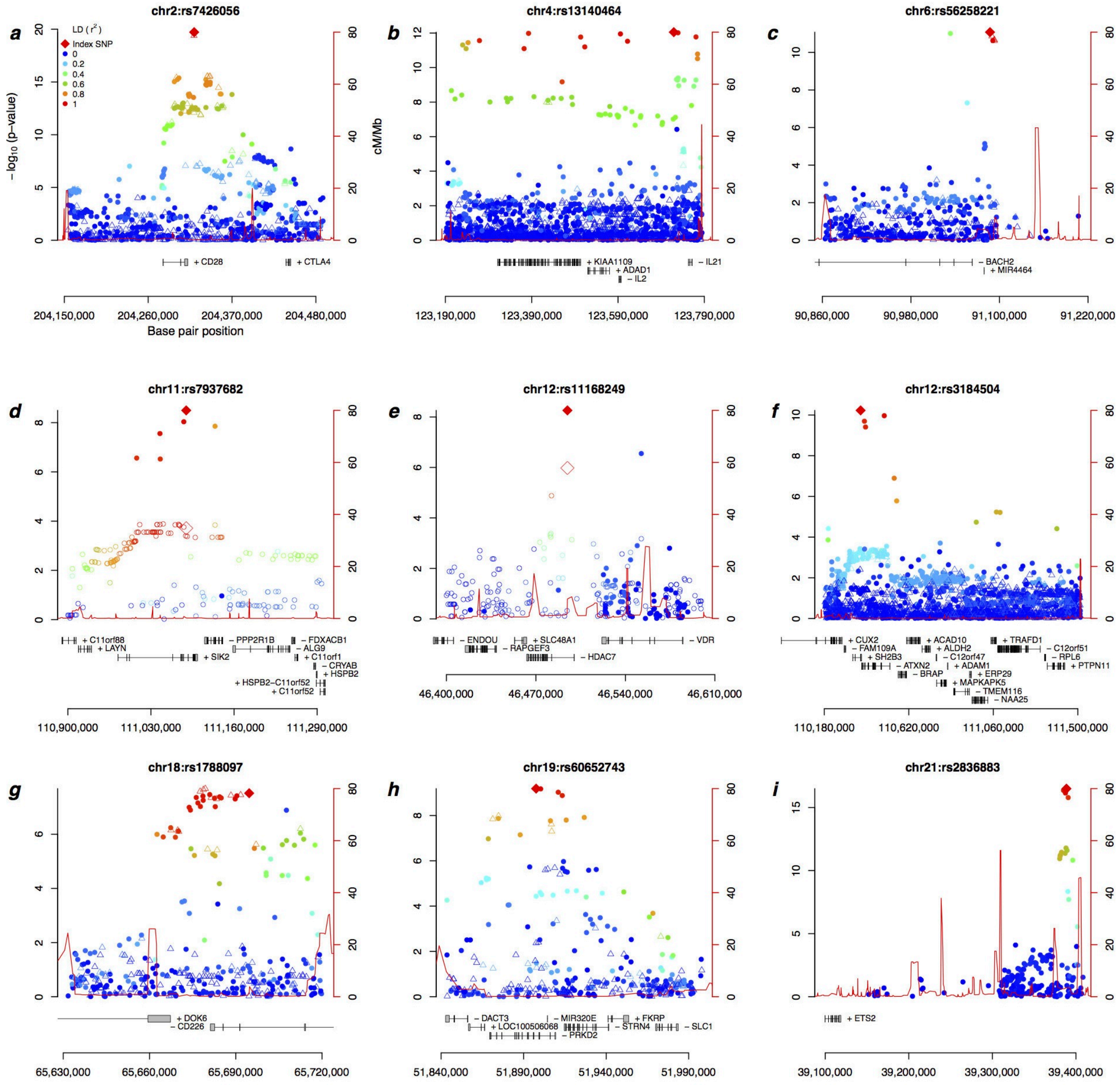
55. Schlicker, A. & Albrecht, M. FunSimMat update: new features for exploring functional similarity. *Nucleic Acids Res* **38**, D244-8 (2010).

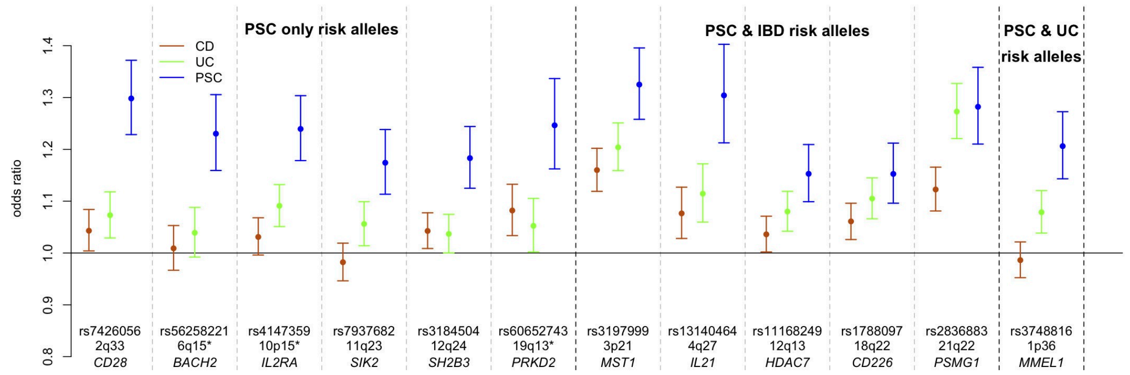
56. Shannon, P. *et al.* Cytoscape: a software environment for integrated models of biomolecular interaction networks. *Genome Res* **13**, 2498-504 (2003).

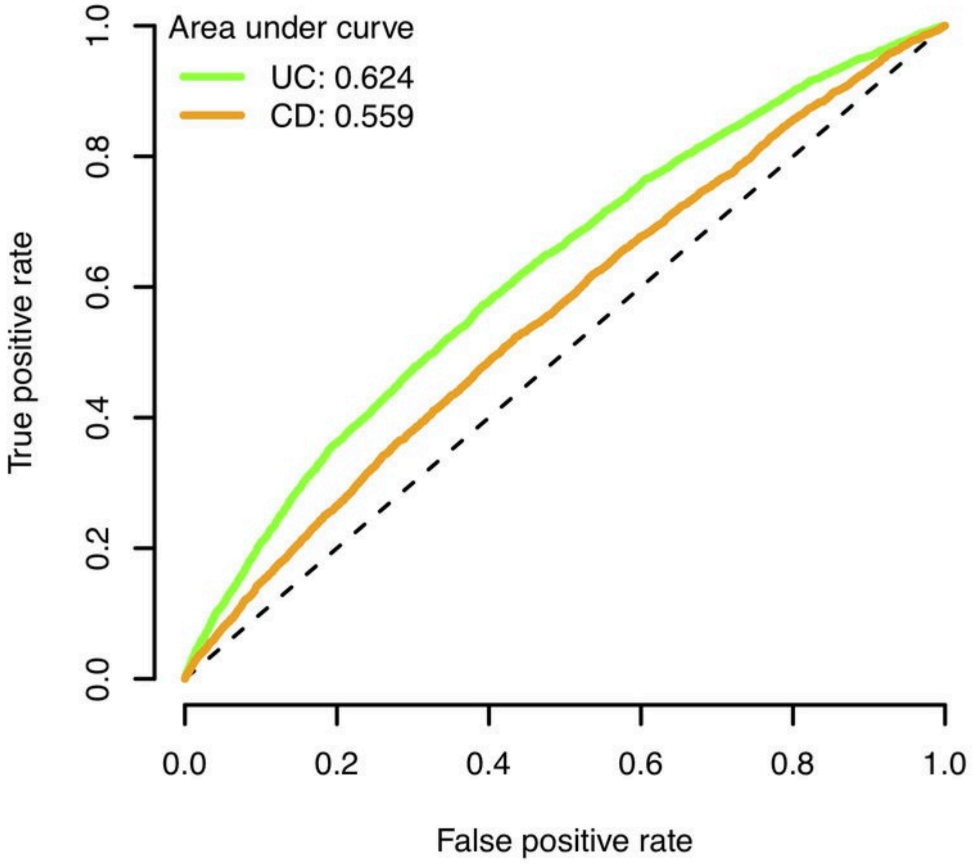
57. Efron, B. Size, power and false discovery rates. *Ann Statist* **35**, 1351-77 (2007).

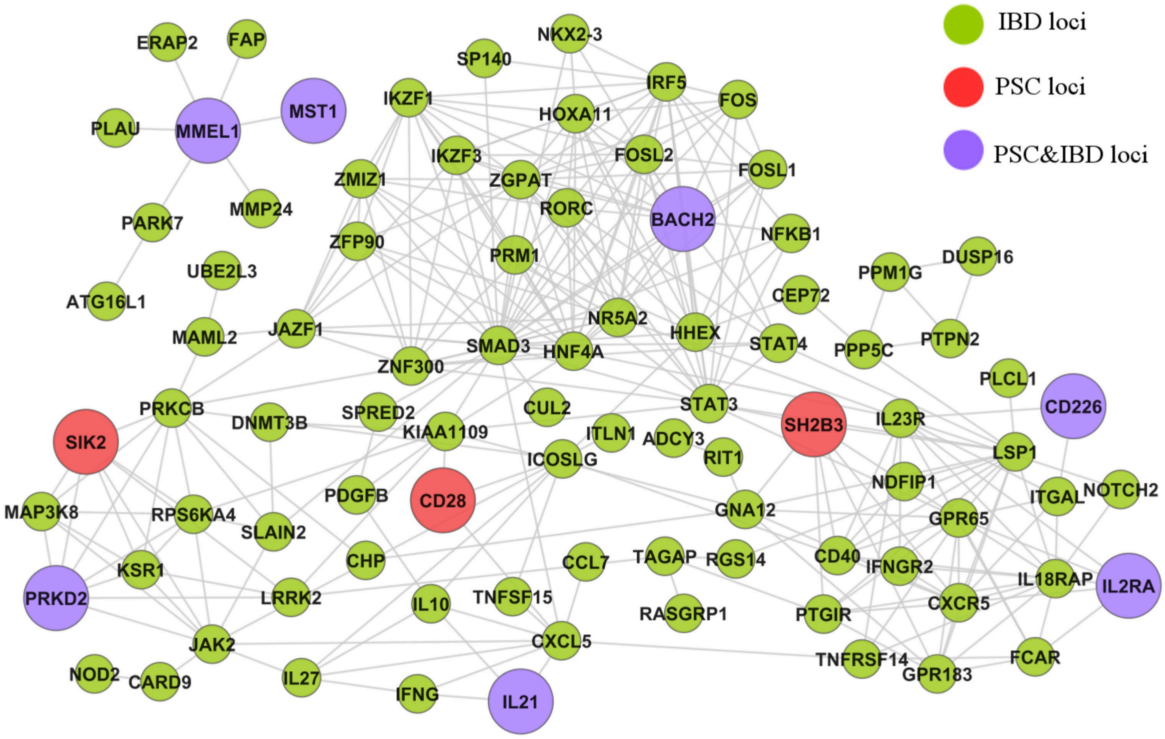
58. Yoo, Y.J., Pinnaduwage, D., Waggott, D., Bull, S.B. & Sun, L. Genome-wide association analyses of North American Rheumatoid Arthritis Consortium and Framingham Heart Study data utilizing genome-wide linkage results. *BMC Proc* **3 Suppl 7**, S103 (2009).

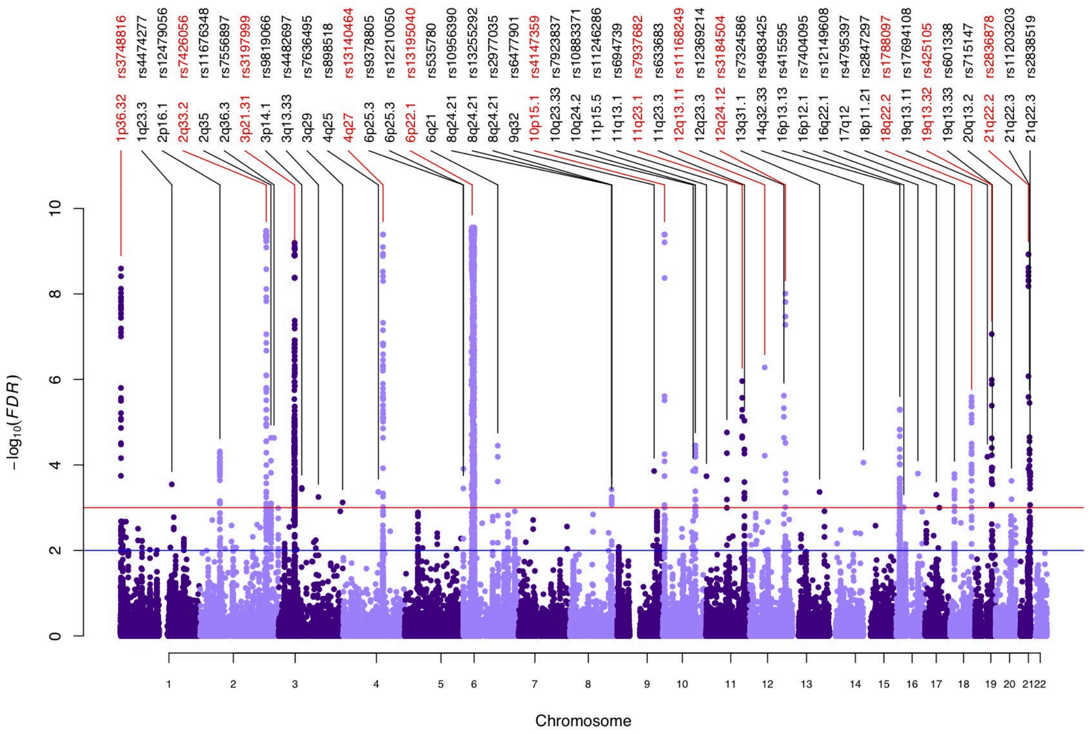
59. So, H.C., Gui, A.H., Cherny, S.S. & Sham, P.C. Evaluating the heritability explained by known susceptibility variants: a survey of ten complex diseases. *Genet Epidemiol* **35**, 310-7 (2011).


a
chr2:rs7426056
LD (r²)
Index SNP
0
0.2
0.4
0.6
0.8
1
− log10 (p−value)
cM/Mb
+ CD28
+ CTLA4
204,150,000
204,260,000
204,370,000
204,480,000
Base pair position
b
chr4:rs13140464
+ KIAA1109
+ ADAD1
− IL2
− IL21
123,190,000
123,390,000
123,590,000
123,790,000
c
chr6:rs56258221
− BACH2
+ MIR4464
90,860,000
90,980,000
91,100,000
91,220,000
d
chr11:rs7937682
+ C11orf88
+ LAYN
+ SIK2
− PPP2R1B
− ALG9
− FDXACB1
+ C11orf1
− CRYAB
+ HSPB2
+ HSPB2−C11orf52
+ C11orf52
110,900,000
111,030,000
111,160,000
111,290,000
e
chr12:rs11168249
− ENDOU
− RAPGEF3
+ SLC48A1
− HDAC7
− VDR
46,400,000
46,470,000
46,540,000
46,610,000
f
chr12:rs3184504
+ CUX2
− FAM109A
+ SH2B3
− ATXN2
− BRAP
+ ACAD10
+ ALDH2
− C12orf47
+ ADAM1
+ MAPKAPK5
+ ERP29
− TMEM116
− NAA25
+ TRAFD1
− C12orf51
− RPL6
+ PTPN11
110,180,000
110,620,000
111,060,000
111,500,000
g
chr18:rs1788097
+ DOK6
− CD226
65,630,000
65,660,000
65,690,000
65,720,000
h
chr19:rs60652743
− DACT3
+ LOC100506068
− MIR320E
− PRKD2
+ FKRP
− STRN4
− SLC1
51,840,000
51,890,000
51,940,000
51,990,000
i
chr21:rs2836883
+ ETS2
39,100,000
39,200,000
39,300,000
39,400,000


PSC only risk alleles
PSC & IBD risk alleles
PSC & UC risk alleles
CD
UC
PSC
odds ratio
0.8
0.9
1.0
1.1
1.2
1.3
1.4
rs7426056 2q33 CD28
rs56258221 6q15* BACH2
rs4147359 10p15* IL2RA
rs7937682 11q23 SIK2
rs3184504 12q24 SH2B3
rs60652743 19q13* PRKD2
rs3197999 3p21 MST1
rs13140464 4q27 IL21
rs11168249 12q13 HDAC7
rs1788097 18q22 CD226
rs2836883 21q22 PSMG1
rs3748816 1p36 MMEL1


Area under curve
UC: 0.624
CD: 0.559
True positive rate
0.0
0.2
0.4
0.6
0.8
1.0
0.0
0.2
0.4
0.6
0.8
1.0
False positive rate


IBD loci
PSC loci
PSC&IBD loci
ERAP2
FAP
MST1
PLAU
MMEL1
PARK7
MMP24
ATG16L1
UBE2L3
MAML2
JAZF1
SP140
NKX2-3
IKZF1
HOXA11
IRF5
FOS
IKZF3
FOSL2
FOSL1
ZMIZ1
ZGPAT
RORC
BACH2
NFKB1
ZFP90
PRM1
DUSP16
PPM1G
CEP72
NR5A2
HHEX
STAT4
PPP5C
PTPN2
SMAD3
HNF4A
ZNF300
PLCL1
CD226
PRKCB
DNMT3B
SPRED2
CUL2
STAT3
SH2B3
IL23R
SIK2
KIAA1109
ITLN1
ADCY3
RIT1
LSP1
ICOSLG
NDFIP1
ITGAL
NOTCH2
PDGFB
CD28
MAP3K8
RPS6KA4
GNA12
GPR65
SLAIN2
CHP
CCL7
TAGAP
RGS14
CD40
IFNGR2
KSR1
IL18RAP
IL2RA
PRKD2
LRRK2
IL10
TNFSF15
RASGRP1
PTGIR
CXCR5
JAK2
CXCL5
NOD2
CARD9
IL27
TNFRSF14
GPR183
FCAR
IFNG
IL21


rs3748816
rs4474277
rs12479056
rs7426056
rs11676348
rs7556897
rs3197999
rs9819066
rs4482697
rs7636495
rs898518
rs13140464
rs9378805
rs12210050
rs13195040
rs535780
rs10956390
rs13255292
rs2977035
rs6477901
rs4147359
rs7923837
rs10883371
rs11246286
rs694739
rs7937682
rs633683
rs11168249
rs12369214
rs3184504
rs7324586
rs4983425
rs415595
rs7404095
rs12149608
rs4795397
rs2847297
rs1788097
rs17694108
rs425105
rs601338
rs715147
rs2836878
rs11203203
rs2838519
1p36.32
1q23.3
2p16.1
2q33.2
2q35
2q36.3
3p21.31
3p14.1
3q13.33
3q29
4q25
4q27
6p25.3
6p25.3
6p22.1
6q21
8q24.21
8q24.21
8q24.21
9q32
10p15.1
10q23.33
10q24.2
11p15.5
11q13.1
11q23.1
11q23.3
12q13.11
12q23.3
12q24.12
13q31.1
14q32.33
16p13.13
16p12.1
16q22.1
17q12
18p11.21
18q22.2
19q13.11
19q13.32
19q13.33
20q13.2
21q22.2
21q22.3
21q22.3
$-\log_{10}(FDR)$
0
2
4
6
8
10
Chromosome
1
2
3
4
5
6
7
8
9
10
11
12
13
14
15
16
17
18
19
20
21
22

6.8 Življenjepis

CURRICULUM VITAE **Mitja Mitrovič**

OSEBNI PODATKI

Ime in priimek, naziv	Mitja Mitrovič, univ. dipl. inž. kem. tehnol.
Službeni naslov	Univerza v Mariboru, Medicinska fakulteta, Center za humano molekularno genetiko in farmakogenomiko (CHMGF), Slomškov trg 15
Kraj	Maribor
Poštna št.	SI – 2000
E-naslov	mitja_mitrovic@yahoo.com mitja.mitrovic@yale.edu
Tel. št.	040 413 838
Datum rojstva	17. 12. 1981
Državljanstvo	slovensko

IZOBRAZBA

Okt. 2002–apr. 2008	diploma iz biokemijskega inženirstva na Fakulteti za kemijo in kemijsko tehnologijo UM
Okt. 2008–mar. 2013	podiplomski študent na Medicinski fakulteti UM; mladi raziskovalec v CHMGF MF UM
Jun. 2012–	podoktorsko izobraževanje na Yale School of Medicine, Department of Neurology, Cotsapas Lab
Zdajšnje področje dela	genetika avtoimunskih bolezni

DELOVNE IN RAZISKOVALNE IZKUŠNJE

Apr. 2010–avg. 2011 usposabljanje v laboratoriju prof. Cisce Wijmenga in dr. Rinse K. Weersmaja na oddelku za genetiko, Univerzitetni klinični center Groningen, Nizozemska; delo z mikromrežami, kontrola kakovosti in različne analize projekta Immunochip pri pribl. 5000 bolnikih s kronično vnetno črevesno boleznijo in zdravih prostovoljcih

OPRAVLJENI PODIPLOMSKI PREDMETI IN TEČAJI

Okt. 2008–feb. 2009 **Molekularna biologija**, podiplomski predmet na Medicinski fakulteti Univerze v Mariboru

Okt. 2008–feb. 2009 **Bioinformatika**, podiplomski predmet na Medicinski fakulteti Univerze v Mariboru

Okt. 2008–feb. 2009 **Biokemija**, podiplomski predmet na Medicinski fakulteti Univerze v Mariboru

Apr.–jun. 2009 **Farmacevtska biotehnologija z gensko tehnologijo**, podiplomski predmet na Medicinski fakulteti Univerze v Mariboru

Okt. 2009–feb. 2010 **Farmakogenomika,** podiplomski predmet na Medicinski fakulteti Univerze v Mariboru

Okt. 2009–feb. 2010 **Uporabna biostatistika v klinični praksi**, podiplomski predmet na Medicinski fakulteti Univerze v Mariboru

Okt. 2009–feb. 2010 **Molekularna imunologija v klinični praksi**, podiplomski predmet na Medicinski fakulteti Univerze v Mariboru

19.–30. apr. 2010 **Epidemiologija in uporabna statistika**, podiplomska šola GUIDE, Univerza v Groningenu, Nizozemska

9.–11. mar 2011 **From DNA to phenotype**, posebna konferenca v organizaciji Kraljeve akademije znanosti in umetnosti (KNAW), Amsterdam, Nizozemska

KONFERENCE IN SIMPOZIJI

22. jun 2007	Farmakogenetika v klinični praksi, satelitsko srečanje ob 15. mednarodni konferenci o citokromih P450, v organizaciji Medicinske fakultete Univerze v Ljubljani. ***Predstavitev posterja***.
26.–29. sep. 2007	7. srečanje biokemijskega društva z mednarodno udeležbo, Maribor. ***Predstavitev posterja***.
23.–26. maj 2009	European Society of Human Genetics, Dunaj, Avstrija. ***Predstavitev posterja***.
20.–23. sep. 2009	Skupni kongres Slovenskega biokemijskega društva in Genetskega društva Slovenije z mednarodno udeležbo. ***Predstavitev posterja***.
24.–25. sep. 2009	Slovenski kemijski dnevi 2009, Maribor. ***Predavanje***.
23. in 24. sep. 2010	Slovenski kemijski dnevi 2010, Maribor. ***Predstavitev posterja***.
24. sep. 2011	Kolokvij iz genetike, Maribor. ***Predstavitev posterja***.
28.–31. maj 2011	European Society of Human Genetics, Amsterdam, Nizozemska. ***Predstavitev posterja***.
14.–16. sep. 2011	Slovenski kemijski dnevi 2011, Portorož. ***Predstavitev posterja***.
16. sep. 2011	Kolokvij iz genetike, Piran. ***Predstavitev posterja***.

NAGRADE IN ŠTIPENDIJE

Apr.–okt. 2008	štipendija Slovenske znanstvene fundacije za raziskave pri slovenskih bolnikih s Crohnovo boleznijo v laboratoriju prof. Potočnika
Apr.–jul. 2010	štipendija Javnega sklada RS za razvoj kadrov in štipendije za sofinanciranje raziskovalnega dela v tujini

DRUGE DEJAVNOSTI

2009−2012	izdelava in urejanje spletnih strani Centra za humano molekularno genetiko in farmakogenomiko Medicinske fakultete Univerze v Mariboru
Konjički	eksperimentiranje v kuhinji, dobri filmi, dobre knjige, dober bluz z vinilk, mediteran-je

ZNANJE JEZIKOV

Angleški j.	aktivno
Nemški j.	pasivno
Francoski j.	pasivno
Srbski in hrvaški j.	aktivno

ZAHVALE

Mentorju prof. dr. Urošu Potočniku se zahvaljujem za podporo in nasvete pri zasnovi in izdelavi doktorske disertacije. Prav tako sem vam izjemno hvaležen za izkazano potrpežljivost med mojo preobrazbo v genetika in da ste mi omogočili, da sem lahko del raziskav opravil na Nizozemskem.

Za številne spodbudne misli, nasvete in pomoč pri zasnovi različnih projektov sem hvaležen somentorju doc. dr. Rinsu Weersmaju. Rinse, prav tako se ti iskreno zahvaljujem za topel sprejem v raziskovalno skupino, prenos raziskovalne vneme pri reševanju genske uganke KVČB in da si mi zagotovil izvrstno izhodišče za prihodnje karierne izzive. Dankje wel chef!

Prof. dr. Cisci Wijmenga se zahvaljujem za pronicljive komentarje in kritične misli pri zasnovi in spremljanju različnih projektov v okviru doktorske naloge in bivanja na Nizozemskem. Cisca, tvoj osebni moto »there are no problems - only challanges« mi je pomagal izpiliti znanstveni profil in vlil zdravo mero samozavesti.

Hvala vsem sodelavcem iz Centra za humano molekularno genetiko in farmakogenomiko. Hvala tudi drugemu osebju iz centra, Medicinske fakultete Univerze v Mariboru in dekanu prof. dr. Ivanu Krajncu ter sodelavcem iz Univerzitetnega kliničnega centra Groningen, ki so poskrbeli, da sem lahko nemoteno opravljal svoje delo.

Posebej se zahvaljujem Katji Repnik za kolegialnost in nesebično pomoč pri administrativnih opravilih in eksperimentalnem delu v laboratoriju. Katja, iskrena hvala, da si mi velikokrat olajšala življenje s tem, da si brez zadržkov in z dobro voljo uredila tekoče posle v laboratoriju.

Karin Fransen se zahvaljujem za pomoč pri pisanju skupnih člankov, podpori in zabavnem krajšanju časa (npr. »the lamest joke contest«) med eksperimentalnim delom v laboratoriju. Hvala ti tudi za nagajive nasvete o nenavadnih šegah in rabi nizozemskih narečij. Dankje wel part!/»Svaka ti čast!«

Prav tako se zahvaljujem Gosii Trynka. Hvala ti, da si me med pisanjem podpirala in z menoj delila svojo neomajno motivacijo in ljubezen do znanosti. Hvala tudi, da si mi predstavila tradicionalne specialitete iz poljske kuhinje: žurek in boršč »na žlico« in vodko »na eks«. Dziękuję bardzo!

Somentorju prof. dr. Pavlu Skoku, prim. Cvetki Pernat, Andreji Ocepek in mag. Silvu Kodru iz UKC Maribor se zahvaljujem za posredovanje kliničnih podatkov in vzorcev.

Bolnikom in njihovim družinskim članom se iskreno zahvaljujem za sodelovanje v raziskavi.

Posebne zasluge imajo tudi trije učitelji, ki so bili med mojim 16-letnim šolanjem ob pravem času na pravem mestu: učitelj angleščine Zlatko Veber, ki mi je s svojim vihravim smislom za angleški humor in narečja angleško govorečega sveta približal, obogatil in dal dobro osnovo za jezik, ki ga zdaj rabim v službi in zasebnem življenju; profesorica kemije mag. Milena Pintarič, ki je skrbno bedela in me spodbujala pri mojih prvih »ekskurzijah« v svet eksperimentalne kemije; prof. dr. Črtomir Stropnik, ki me je z odličnim podajanjem svojega enciklopedičnega znanja organske kemije in številnih pogovorih o znanosti, Mediteranu in sploh vsem spodbudil k vpisu na doktorski študij.

Tina! Tvoj prihod v moje življenje, še posebej pa tvoj prihod v ZDA, me je dodatno vzpodbudil k zaključku disertacije. Končno rešeno! :)

Najbolj pa sem hvaležen vsem domačim, ki so me spodbujali in mi omogočili, da sem, kar sem. Draga starši, Dina in Vojko! Hvala vama, da sta mi privzgojila občutek za korenine in krila. Hvaležen sem vama tudi za vse (pre)skoke na Mediteran, ki je tako postal zavetišče mojih misli in telesa.

ACKNOWLEDGEMENTS

Dear Rinse! I am deeply grateful for all your encouraging thoughts, guidance and support. Thank you for the transmission of the good research spirit, warm welcome into your research team and providing me with an excellent base for my future career. Dankje wel chef!

Dear Cisca! Thank you very much for offering me a front row seat and opportunity to learn, observe and take part in the remarkable research done in your lab. Your proverb »there are no problems - only challanges«, has helped me to carve up my scientific profile and build up a healthy dose of self-confidence.

Dear Karin! Thanks for being such a good and caring friend. I'm grateful for all the help with writing papers, but especially for sharing all that funny stuff while working in the lab (e.g. »lamest joke contest«) and for introducing me to unusual local Dutch dialects and habits. Dankje wel part!/»Svaka ti čast!«

Dear Gosia! Many thanks for the encouragment and supporting me when needed. Thanks for sharing your unshakeable love and motivation for science! I'm also very grateful for introducing me to the traditional Polsih cuisine, especially to żurek & barszcz czerwony »on the spoon« and vodka »on bottoms-up«. Dziękuję bardzo!

Izjava doktorskega kandidata

UNIVERZA V MARIBORU

MEDICINSKA FAKULTETA

IZJAVA DOKTORSKEGA KANDIDATA

Podpisani **Mitja Mitrovič**, vpisna številka **30806257,**

izjavljam,

da je doktorska disertacija z naslovom **Asociacijska analiza na celotnem genomu pri slovenskih bolnikih s kronično vnetno črevesno boleznijo**

- rezultat lastnega raziskovalnega dela,
- da predložena disertacija v celoti ali v delih ni bila predložena za pridobitev kakršnekoli izobrazbe po študijskem programu druge fakultete ali univerze,
- da so rezultati korektno navedeni in
- da nisem kršil avtorskih pravic oz. pravic intelektualne lastnine drugih.

Podpis doktorskega kandidata

Mitja Mitrovič

Asociacijska analiza na celotnem genomu pri slovenskih bolnikih s kronično vnetno črevesno boleznijo
doktorska disertacija, Medicinska fakulteta Univerze v Mariboru

Copyright © 2013 Mitja Mitrovič. Vse pravice pridržane.
Natisnjeno na recikliranem papirju v Sloveniji.

Lektoriranje: mag. Suzana Jakoša, prof. slov. - Založba Sentenca, d. o. o.
Oblikovanje prve strani: Staš Mitrovič, univ. dipl. inž. arh.
Tisk in vezava: Labar biro − Posredništvo Barbara Lazar, s. p.

UDK: 616.34-002-056.7:601.4.008.7(043.3)