Multi-locus genetic risk score predicts risk for Crohn's disease in Slovenian population

Zupancic, Katarina; Skok, Kristijan; Repnik, Katja; Weersma, Rinse K.; Potocnik, Uros; Skok, Pavel

Published in:
World Journal of Gastroenterology

DOI:
10.3748/wjg.v22.i14.3777

Publication date:
2016

World Journal of Gastroenterology

Submit a Manuscript: http://www.wjgnet.com/esps/
Help Desk: http://www.wjgnet.com/esps/helpdesk.aspx
DOI: 10.3748/wjg.v22.i14.3777

World J Gastroenterol 2016 April 14; 22(14): 3777-3784
ISSN 1007-9327 (print) ISSN 2219-2840 (online)

ORIGINAL ARTICLE

Case Control Study

Multi-locus genetic risk score predicts risk for Crohn's disease in Slovenian population

Katarina Zupančič, Kristijan Skok, Katja Repnik, Rinse K Weersma, Uroš Potočnik, Pavel Skok

Katarina Zupančič, Kristijan Skok, Katja Repnik, Uroš Potočnik, Center for Human Genetics and Pharmacogenomics, Medical Faculty of Maribor, University of Maribor, 2000 Maribor, Slovenia

Rinse K Weersma, Department of Gastroenterology and Hepatology, University of Groningen and University Medical Center Groningen, 9700 Groningen, The Netherlands

Pavel Skok, Department of Gastroenterology, University Medical Center, 2000 Maribor, Slovenia

Pavel Skok, Slovenia and Medical Faculty of Maribor, University of Maribor, 2000 Maribor, Slovenia

Author contributions: Zupančič K and Skok K contributed equally to this work; Zupančič K, Skok K and Potočnik U designed the research; Zupančič K, Skok K and Repnik K performed the research; Zupančič K, Skok K and Skok P analyzed the data; Weersma RK contributed new data and analytic methods; Zupančič K, Skok K, Potočnik U and Skok P wrote the paper; Potočnik U, Repnik K and Skok P revised the article.

Institutional review board statement: The study was approved by the Slovenian National Committee for Medical Ethics (KME 80/10/ 07, 21p / 12/ 07 and 106/05/11).

Informed consent statement: All patients gave informed consent.

Conflict-of-interest statement: No benefits in any form have been received or will be received from a commercial party related directly or indirectly to the subject of this article.

Data sharing statement: Technical appendix, statistical code, and dataset available from the corresponding author at pavel.skok@guest.arnes.si.

Correspondence to: Pavel Skok, MD, PhD, Full Professor, Department of Gastroenterology, University Medical Center, Ljubljanska ulica 5, 2000 Maribor, Slovenia. pavel.skok@guest.arnes.si
Telephone: +386-23212878
Fax: +386-23212845

Received: January 16, 2016
Peer-review started: January 18, 2016
First decision: February 18, 2016
Revised: March 1, 2016
Accepted: March 13, 2016
Article in press: March 14, 2016
Published online: April 14, 2016

Abstract

AIM: To develop a risk model for Crohn's disease (CD) based on homogeneous population.

METHODS: In our study were included 160 CD patients and 209 healthy individuals from Slovenia. The association study was performed for 112 single nucleotide polymorphisms (SNPs). We generated genetic risk scores (GRS) based on the number of risk alleles using weighted additive model. Discriminatory accuracy was measured by area under ROC curve (AUC). For risk evaluation, we divided individuals according to positive and negative likelihood ratios (LR) of a test, with LR > 5 for high risk group and LR < 0.20 for low risk group.

RESULTS: The highest accuracy, AUC of 0.78 was achieved with GRS combining 33 SNPs with optimal sensitivity and specificity of 75.0% and 72.7%, respectively. Individuals with the highest risk (GRS >

5.54) showed significantly increased odds of developing CD (OR = 26.65, 95%CI: 11.25-63.15) compared to the individuals with the lowest risk (GRS < 4.57) which is a considerably greater risk captured than in one SNP with the highest effect size (OR = 3.24). When more than 33 SNPs were included in GRS, discriminatory ability was not improved significantly; AUC of all 74 SNPs was 0.76.

CONCLUSION: The authors proved the possibility of building accurate genetic risk score based on 33 risk variants on Slovenian CD patients which may serve as a screening tool in the targeted population.

Key words: Inflammatory bowel disease; Crohn's disease; Discriminatory accuracy; Genetic risk score; Single nucleotide polymorphisms

Core tip: Genome wide association studies have provided a comprehensive catalogue of susceptibility inflammatory bowel disease (IBD) loci, which now present an important basis for genetic risk prediction. We aimed to develop an accurate Crohn's disease (CD) risk prediction model for the Slovenian cohort. The most optimal 33 SNPs model showed good discriminatory ability, which may be useful for risk stratification in targeted population (gastrointestinal disturbances, positive family history). Individuals in the highest risk group have 27-fold higher odds for CD risk compared to individuals in the lowest risk group. To the best of our knowledge, this is the first population specific genetic prediction based on recently established IBD loci.

Zupančič K, Skok K, Repnik K, Weersma RK, Potočnik U, Skok P. Multi-locus genetic risk score predicts risk for Crohn's disease in Slovenian population. *World J Gastroenterol* 2016; 22(14): 3777-3784 Available from: URL: http://www.wjgnet.com/1007-9327/full/v22/i14/3777.htm DOI: http://dx.doi.org/10.3748/wjg.v22.i14.3777

INTRODUCTION

Crohn's disease (CD) is a multifactorial chronic inflammatory bowel disease (IBD) caused by dysregulated immune response to commensal intestinal microflora in genetically susceptible individuals[1]. CD can affect any segment of the gastrointestinal tract and has a relapsing-remitting course with various patterns of behavior[2,3]. Prevalence and incidence of CD increased significantly over the last decades in developed parts of the world and ranges from 136 to 319 per 100000 and from 6 to 29 per 100000, respectively, indicating its emergence as a global disease[4].

Genome-wide association studies (GWASs) have provided valuable insight into the genetic architecture of CD, thus helping us better understand the mechanisms of innate immunity (*NOD2*), mucosal integrity (*IBD5*), autophagy (*ATG16L1, IRGM*), lymphocyte differentiation and proliferation (*IL23R, STAT3*)[5-7]. Recent meta-analysis of data from 15 existing GWASs and an independent genotype set obtained from DNA microarray (Immunochip) enrolled 75000 IBD cases and controls from International Inflammatory Bowel Diseases Genetic Consortium (IIBDGC) and identified 71 novel variants for a total of 163 IBD loci. Of these loci, 110 conferred risk to both IBD subtypes, whereas 30 were unique to CD[8]. Several studies have investigated risk prediction for CD[9-12]. In CD the strongest known association is with the *NOD2* gene, which was for some time considered to be a possible candidate for genetic screening but ultimately has not been accepted [an area under ROC curve (AUC) of 0.56], because individual markers have small effect on risk and thus have poor predictive ability[13]. In addition, inclusion of only highly significant common variants with strong effect sizes (*NOD2, IL23R, ATG16L1, IRGM*) into a predictive model managed to improve predictive power but it was still insufficient for accurate prediction (AUC of 0.66)[10]. In recent studies, polygenic genetic risk scores (GRS) have been used to summarize risk-associated alleles, weighted by their effect sizes (odds ratios) among an assemble of markers that do not individually achieve significance in association study since it has been suggested that a higher number of included variants with weak to moderate effect sizes can explain larger proportion of heritability[5,14].

Recently established 163 IBD loci now provide an important basis for genetic risk prediction. Attempts have been made to develop accurate genetic risk profiles based on a large number of patients from several population heterogeneous cohorts that were enrolled in IIBDGC[15]. However, limited data exist on developing genetic risk profiles in a single genetically homogeneous population.

We aimed to test the joint contribution of well-established susceptibility loci, obtained from custom designed single nucleotide polimorphism (SNP) genotyping chip "Immunochip" (iCHIP)[8], in order to construct accurate risk prediction for Slovenian CD patients.

MATERIALS AND METHODS

Cases and controls

Study population consisted of 236 healthy individuals in the control group and 202 CD patients from the University Medical Center Ljubljana and Maribor as described previously[16]. All CD patients had reliable clinically and histopathologically confirmed disease. The study was approved by Slovenian National Committee for Medical Ethics (KME 80/10/07, 21p/12/07 and KME 106/05/11). Patients gave informed consent prior to inclusion in the study and the study was performed

in accordance with The Code of Ethics of the World Medical Association (Declaration of Helsinki).

SNP selection

Genotyping data for cases and controls was obtained from the iCHIP, where Slovenian patients and controls were enrolled within the IIBDGC. Genotyping for initial 236 controls and 202 CD patients was performed with hybridization on iCHIP according to protocol (Illumina). Rigorous quality control of the study group (gender matching, call rate, family ties and genetic origin verification) and genotyping information [Hardy-Weinberg equilibrium (HWE)] was applied in PLINK v1.07 and R. Individuals with more than 10% missing genotypes were excluded. SNPs with a call rate below 90% were discarded from further analyses. We also excluded SNPs when the controls showed deviation from HWE with a $P < 0.0001$ in control group. The remaining 160 CD patients, 209 controls, and 112 SNPs from a total of 163 SNPs were included in further statistical analysis.

Statistical analysis

The case-control association analysis of allele frequencies was statistically assessed in SPSS 22.0 (IBM Corp., Armonk, NY) using Fischer exact test and χ^2. The results were presented as *P*-values, OR and 95%CI. We compared risk alleles and their frequencies to those in meta-analysis study by Jostins *et al*[8] and included only those SNPs from a total of 112 SNPs that showed association of risk allele with Slovenian CD patients risk allele. Statistical significant threshold (*P*-value) was at 0.05 and 4.46×10^{-4} after Bonferroni correction for multiple testing.

Genetic risk profiles construction

We constructed GRS in an additive manner using a weighted approach. The number of risk alleles at each locus (2, 1, 0) was multiplied by their corresponding beta-coefficients of effects sizes [log(OR)] and then summed up in GRS that each individual carried. In order to avoid bias we used ORs from recent meta-analysis by Jostins *et al*[8].

For optimal risk construction, we used a ranking *P*-values approach, starting with SNP with the lowest *P*-value and gradually adding one per one in GRS[17]. The most optimal GRS was then compared to GRS of the remaining SNPs, to GRS of marginally significant SNPs with highest OR and to GRS of all included SNPs. ROC with AUC were used to measure discriminatory accuracy at various GRS cut-offs (sensitivity/1-specificity). An area of an AUC < 0.7 represents poor discrimination; 0.7-0.8 acceptable (fair) discrimination, and 0.8-0.9 excellent discrimination[18]. Cut-off values for sensitivities, specificities, positive/negative predictive values and positive/negative likelihood ratios were computed in Medcalc 14.8. Likelihood ratio (LR) describes how likely an affected person is going to have a disease compared to a healthy person with the same result according to a given test (positive or negative). A positive LR above 5 is considered to be a moderate evidence to confirm a disease, LR above 10 strongly confirms a disease, whereas a negative LR below 0.2 is considered moderate, and below 0.1 strong evidence to exclude disease. LR are powerful tools for GRS evaluation since they are not affected by a prevalence and with prior probability (prevalence) they present a basis for post-test probability (predictive value - PV) calculation[19]. Positive PV presents the probability of patient having a disease if the test is positive, while the opposite is true for a negative PV[20]. For post-test probabilities estimations, we used prevalence of disease 1.2/1000, since there are approximately 2000-2500 CD patients in Slovenia.

Statistical analysis of distribution (normal, non-parametric) of selected parameters (GRS, allele frequencies, OR, United States) between cases and controls, and between different GRS models was conducted with Shapiro-Wilk test in SPSS 22.0. *t*-test or rank-sum test (Mann-Whitney) were used according to distribution.

RESULTS

Association analysis

We found that risk alleles frequencies of 74 SNPs (66%) matched with those in the reference study by Jostins *et al*[8]. The risk alleles of remaining 38 SNPs were in the opposite direction and therefore we excluded them from further analysis (Supplement Table 1). This was to be expected since our case-control group was too small to achieve sufficient statistical power comparing to 75.000 cases and controls enrolled in IIBDGC[8]. In present study, we initially confirmed 15 statistically significant associations (for dominant, recessive, allelic model) from a total of 74 SNPs; among these 15 significant associations, only 8 SNPs were statistically significant under allelic model (Supplement Table 1). None of them remained significant ($P < 4.46 \times 10^{-4}$) after Bonferroni correction.

Genetic risk profiles

We used the ranking *P*-value method for GRS construction as described in the methods section. For individual SNPs, AUC ranged from 0.50 to 0.57 (median 0.53) for SNPs with the lowest effects sizes versus SNPs with the highest effect sizes, respectively, but the difference between these two most distant AUC values was not significant ($P = 0.06$). Significantly better ($P < 10^{-4}$) discriminatory ability than in an individual marker was achieved when 8 nominally significant SNPs ($P < 0.05$) were included in a model with AUC of 0.69 (95%CI: 0.64-0.74). Their ORs ranged from 1.42 to 3.24 (median 1.50). The most optimal GRS was achieved with 33 SNPs model with AUC of 0.78 (95%CI: 0.72-0.82) and ORs ranging from 1.15 to

Table 1 Genetic risk profile based on the genetic risk score in 33 single nucleotide polymorphisms model

GRS	Sensitivity	Specificity	+LR	-LR	+PV	-PV	+PV[1]	-PV[1]
> 4.27	98.12%	19.62%	1.22	0.10	0.15%	100.00%	11.94%	98.95%
> 4.57	93.12%	34.45%	1.42	0.20	0.17%	99.98%	13.63%	97.83%
> 5.05	75.00%	72.73%	2.75	0.34	0.33%	99.96%	23.41%	96.32%
> 5.54	35.63%	93.30%	5.32	0.69	0.63%	99.92%	37.14%	92.88%
> 5.81	10.00%	99.04%	10.42	0.91	1.24%	99.89%	53.65%	90.83%
> 5.94	6.25%	99.52%	13.02	0.94	1.54%	99.89%	59.13%	90.53%

[1]PV-positive and negative predictive value for estimated prevalence of CD in targeted population with high risk (10%). LR-positive and negative likelihood ratio, PV-positive and negative predictive value for estimated prevalence of CD in Slovenian population (0.12%). GRS: Genetic risk scores; CD: Crohn's disease; LR: Likelihood ratios; PV: Predictive value.

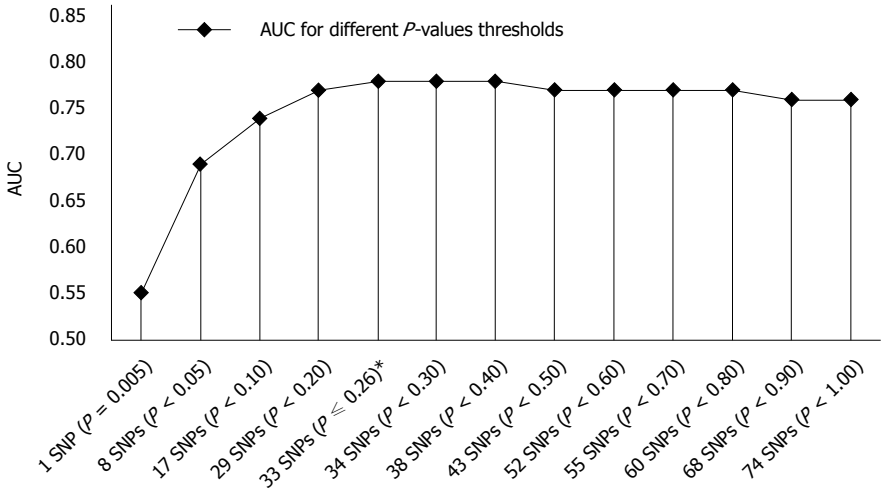

Figure 1 Area under curve for different *P*-values thresholds and corresponding numbers of included single nucleotide polymorphisms. The most optimal AUC was achieved with GRS of 33 risk alleles (AUC of 0.78) at *P*-value threshold of 0.256 (marked with *). Adding 41 SNPs to GRS did not improve GRS significantly, however it presented needless background noise. AUC: Area under curve; GRS: Genetic risk scores; SNPs: Single nucleotide polymorphisms.

3.24 (median 1.32). It performed significantly better ($P = 8 \times 10^{-4}$) than the 8 SNPs model, although the 8 SNPs model had higher ORs median ($P = 0.0012$). Interestingly, mean risk allele frequencies (RAF) of included SNPs did not differ significantly between 8 SNPs and 33 SNPs risk models ($P = 0.110$).

When more than 33 SNPs were included in GRS (Figure 1), discriminatory ability was not improved significantly; AUC of all 74 SNPs was 0.76. These additional SNPs present needless background noise. To confirm this, we also tested accuracy of the remaining 41 SNPs, which achieved poor performance with AUC of 0.58 (95%CI: 0.53-0.63) and this value is similar to an accuracy of an individual SNP. The 41 SNPs model accuracy and median of ORs (ranging from 1.00 to 1.85; median 1.11) of SNPs were significantly lower than in the aforementioned 33 SNPs model ($P < 10^{-4}$). RAF of the included SNPs did not differ significantly between these two models ($P = 0.288$).

Evaluation of 33 SNPs genetic risk model

We presented the applicability of GRS in context of pre-test (prevalence), LR, and corresponding post-test probability of target population for two different prior probabilities (general population, high-risk population) (Table 1). Due to the low CD prevalence (0.12%) only up to 1.5% of population would be correctly confirmed as diseased. On the contrary, high-risk individuals (positive family history, gastrointestinal disturbances) with estimated pre-test probability of 10% have up to 59.1% higher probabilities of having the disease if they test positive.

To evaluate risk between individuals in our study, we divided them into three risk groups according to optimal positive and negative LRs at corresponding GRS cut-offs. LR are powerful tools for model evaluation, since they do not depend on prevalence (for further readings see Methods). The high-risk group included individuals who tested above GRS cut-off 5.54 (positive LR > 5.0). The intermediate group (grey zone) included individuals with GRS between 4.57 and 5.54 and the low risk group included individuals who tested below GRS cut-off 4.57 (negative LR < 0.20). The results showed that the odds of developing CD (OR = 26.65, 95%CI: 11.25-63.15) were significantly increased in individuals in the highest risk group compared to individuals in the lowest risk group (Table 2).

Since the majority of the tested individuals fall into the "grey zone" category (intermediate risk group) (Figure 2) it would be useful to further stratify

Table 2 Comparison of risk in different groups of individuals

Group	OR	*P* value	95%CI
High *vs* low risk	26.65	< 0.0001	11.25-63.15
High *vs* intermediate risk	4.06	< 0.0001	1.32-12.47
Intermediate *vs* low risk	5.13	< 0.0001	2.76-9.56

High-risk group: GRS > 5.54; Intermediate risk group: 4.57 < GRS ⩽ 5.54; low risk group: GRS ⩽ 4.57. GRS: Genetic risk scores.

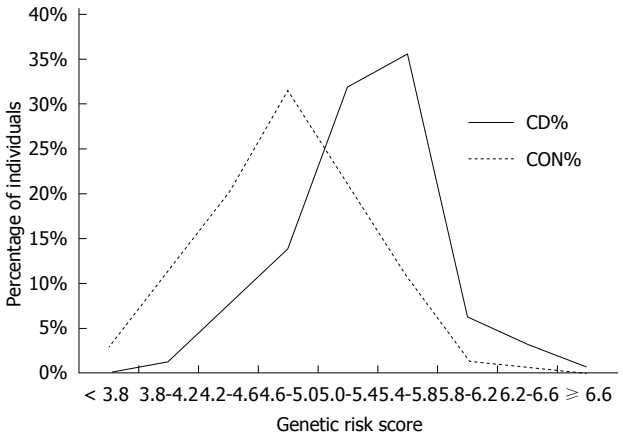

Figure 2 Distribution of Genetic risk scores in Crohn's disease and controls. GRS is normally distributed (Shapiro-Wilk, $P > 0.05$) in CD cases and controls. Cases have significantly higher GRS (5.32 ± 0.47) than controls (4.78 ± 0.53); $P < 10^{-4}$. CD: Crohn's disease; GRS: Genetic risk scores.

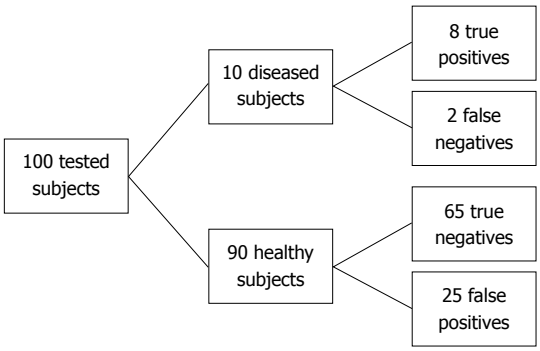

Figure 3 Flow diagram for sensitivity (75%) and specificity (72.7%) in targeted population (Crohn's disease prevalence 10%). If we test 100 subjects and set the cut-off at 5.05, 32 of them will test positive (above cut-off), with post-test probability of disease approximately 24% (8/33).

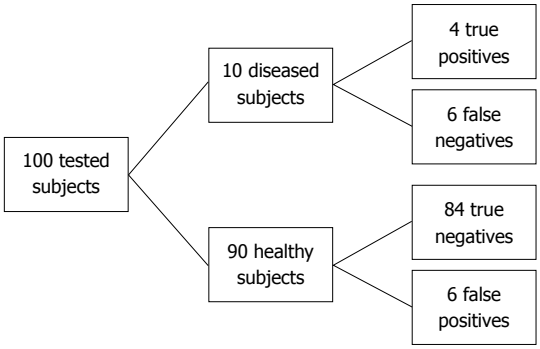

Figure 4 Flow diagram for sensitivity (35.6%) and specificity (93.3%) in targeted population (Crohn's disease prevalence 10%). If we test 100 subjects and set the cut-off at 5.54, 10 of them will test positive (above cut-off), with post-test probability of disease approximately 40% (4/10).

these individuals. According to the best trade-off between sensitivity and specificity we can set GRS cut-off at 5.05 (sensitivity 75.0%, specificity 72.7%, +LR 2.45), where in high risk population 23.4% subjects will be identified as diseased if they are tested positive (Figure 3). Individuals, who test below the cut-off 5.05, have a high probability of 96.5% that they are not affected.

Moderately better risk stratification (+LR > 5, high risk group) can be achieved by setting the threshold to GRS cut-off 5.54 with specificity of 93.3% but at expense of higher number of false negatives (sensitivity of 35.6%) (Figure 4). By setting the threshold even higher (GRS > 5.80) it is possible to yield positive LR greater than 10% and 53.7% individuals will be classified as diseased.

More precise CD exclusion can be achieved with lower GRS cut-off (< 4.57), where we can rule out non-diseased individuals with a sensitivity of 93.12% and a negative PV of 97.8% but at an expense of having more false positives (specificity of 34.65%).

DISCUSSION

In our study, we developed an accurate CD risk prediction model for Slovenian cohort. To the best of our knowledge, this is the first population specific genetic prediction based on recently established IBD loci. The most optimal 33 SNPs model showed good discriminatory ability (AUC of 0.78), which may be useful for risk stratification in targeted population. Individuals in the highest risk group (GRS > 5.54) have 27-fold higher OR for CD risk compared to individuals in the lowest risk group (GRS ⩽ 4.57). This is considerably greater risk captured than in individual SNP with the highest effect size (*IL23R*, OR = 3.24). The highest discriminatory ability of individual SNPs yielded only 0.57 at best, which support findings from study by Jakobsdottir *et al*[10], that "individual markers are poor classifiers even with replicated high effect sizes".

According to literature, genetic risk prediction in CD using common risk loci ranges from AUC of 0.56 to 0.72[10-12]. Recently it has been suggested that robust machine-learning techniques using large sample sizes and wider variant spectrum (including risk allele frequencies less than 0.05) in low linkage disequilibrium importantly improves risk prediction in CD to an AUC greater than 0.85[15,17]. Theoretically, given its high heritability (not accounting the low prevalence of CD), it is possible to achieve AUC up to 0.98 if all CD risk loci have been identified and effect sizes are accurately measured[21,22]. However, these studies are technologically complex and are mostly based on

simulation data and/or require large and heterogeneous patients' cohorts data which is not always available. In our study, in order to avoid methodologic complexity, we used a simplified *P*-value ranking approach to determine optimal number of SNPs, as already described in Methods[17].

Studies have shown that Europe is genetically and ethnically diverse. Genetic differentiation exists not only between the northwest and southeast part of the continent but also within these two distinct groups, creating indispensable inter-population differences[23,24]. In this case SNPs may serve as proxies for shared environments that can differ from population to population. This may be due to interference of different factors modifying genetic architecture (geographical placement, pollutants, adherence to particular diet, *etc.*)[25]. Therefore, population specific risk profiles may offer a more accurate risk estimation for the selected population, which was one of the main reasons why we included the small homogeneous Slovenian cohort[26].

A genetic risk model may be primarily useful for screening purposes in a population with higher prior probabilities of CD; prolonged gastrointestinal disturbances suggestive of CD (abdominal pain, diarrhea, weight loss) and/or positive family history to identify at-risk individuals for further more invasive diagnostic procedures. In addition to sufficiently high prevalence, a risk prediction model also needs to be sensitive and specific with considerably high positive and negative likelihood ratios and cost efficient (as few markers as possible)[19,20,27]. The prevalence of CD in patients with one affected first-degree relative ranges from 2.2% to 16.2%[28]. For Slovenia, there is currently no accurate data for family history in CD patients available, however it has been clinically estimated that around 10% of CD patients have one affected first-degree relative. In this case, positive post-test probability rises up to 59.1% compared to general population, where positive post-test probability due to low CD prevalence yields only 1.5% at best. According to our results, the most optimal cut-off was at GRS > 5.05 with sensitivity of 75.0% and specificity of 72.7%. By raising cut-off to a higher GRS > 5.54 it is possible to stratify at-risk individuals even more precisely with a specificity of 93.3% but at expense of lower sensitivity (35.6%) and therefore a higher number of false negatives. However, due to small number of individuals who test positive (Figure 4) it may be more cost-effective and practical than setting cut-off to a lower GRS. For comparison, prostate-specific antigen test yields similar results with sensitivity and specificity of 20.5% and 93.6%, respectively and is widely used in screening for prostate cancer[29].

To compare risks between high and low risk group, GRS of individuals are usually divided into quartiles or quintiles. However, this approach may be too rigid and therefore may not present all true observations of the series[18]. We thereby propose a more suitable approach, by using test likelihood ratios, which are powerful tools for the GRS evaluation[19]. In our case we divided individuals into three groups according to moderate increase in likelihood of disease (+LR > 5.0; -LR < 0.20) because of a better trade-off between sensitivity and specificity (less false negatives). Individuals in the highest risk group had 27-fold higher OR for CD risk compared to individuals in the lowest risk group. Similar results with high OR between risk groups were presented in rheumatoid arthritis GRS study by Yarwood *et al*[27] (OR = 27.13), and in psoriasis GRS study by Yin *et al*[30] (OR = 28.20) but they used quintiles and quartiles, respectively. Since LR higher than 10 is largely conclusive for disease, in our case it is possible to yield even 55-fold higher OR for CD risk compared to individuals, but at expense of higher number of false negatives (sensitivity 10%).

Our results also highlight the importance of testing published SNPs in population cohort before building models. For example, when all 112 SNP were included in the model we achieved an AUC of only 0.64, since risk alleles were labelled according to reference study by Jostins *et al*[8] and 38 were in the opposite directions than those in the reference study.

It has been suggested that including a higher number of risk variants with weak to moderate effect sizes, that were not all significant in the association study, can improve risk prediction on account of explaining larger proportion of heritability[14,31]. Interestingly, when additional 41 SNPs with weak ORs were included to the model of 33 SNPs the discriminatory ability did not improve significantly. In our case, these SNPs actually present only a background noise and are therefore cost redundant. The discriminatory accuracy of 41 SNPs was in fact similar to an accuracy of an individual SNP, which represents a poor classifier (no better than a chance). Therefore, a cluster of higher numbers of non-significant variants with very weak effect sizes may also explain only a minor proportion of genetic variance in complex diseases (the same as individual SNP). Currently, 163 IBD loci accounts for 13.6% of total CD heritability[8]. Future discovery of rare variants with lower frequencies and high effect sizes through robust gene-mapping studies (next-generation sequencing) could lead to improvement of "missing heritability" and to more accurate genetic risk prediction[22,31].

A limitation of our study is a small number of CD patients and controls. Therefore, results may be biased due to over-fitting and should be validated on a larger case-control sample size. Furthermore, we analyzed only common risk loci from iCHIP, which were strongly confirmed on a large IIBDGC heterogeneous cohort thus balancing out possible inter-population genetic variations. It would be interesting to test, if other variants which have not reached genome-wide levels of significance could improve our population specific risk prediction. And lastly, we used additive model which does not account for possible gene-gene interactions.

In conclusion, to the best of our knowledge, this is

the first population specific study developing predictive models after the 163 IBD SNPs have been discovered[8]. Population specific genetic models may offer more optimal risk stratification for selected population than ethnically and genetically diverse consortiums. In our study we presented "real" issues of small populations, a simpler methodology approach with ranking *P*-values and a comprehensible transition to clinical scenario.

ACKNOWLEDGMENTS

The authors thank Koder S, PhD, for part of the samples, which were used in the study.

COMMENTS

Background

To date genome-wide association studies (GWAs) have identified more than 160 loci in the human genome that contribute to the development of inflammatory bowel disease (IBD). Multi-locus profiles of genetic risk can be used to translate discoveries from GWAs into tools for population health research, such as development of accurate risk profiles for genetic risk prediction.

Research frontiers

Attempts to develop reliable and accurate genetic risk profiles have been made on a large number of patients from several population heterogeneous cohorts. However, limited data exists on developing genetic risk profiles in a single genetically homogeneous population.

Innovations and breakthroughs

To the best of our knowledge, the authors performed one of the first population specific genetic prediction studies based on recently established IBD associated loci. Their results suggest that it is possible to construct a genetic risk model with good discriminatory ability on a homogenous population cohort. The model may be useful for stratifying individuals at higher risk of developing Crohn's disease (CD).

Applications

The genetic risk model may serve as a screening tool in a targeted population with higher risk (gastrointestinal disturbances, positive family history) to develop disease. Similar studies may be performed for other smaller cohorts in order to further improve their population specific genetic risk prediction.

Terminology

Immunochip is an Illumina Infinium single-nucleotide polymorphism (SNP) microarray that includes momentarily approx., 200000 SNPs relevant to multiple different immune-mediated diseases including CD and ulcerative colitis. This chip provides a powerful tool for immunogenetics gene mapping.

Peer-review

The present work has originality. The research study on genetic risk predictor for CD using a reasonable number of SNPs has merits and the only shortcoming is sample size, which was very small. However, the sample population is homogeneous, thus attenuating the effects of the small sample size. The methodology used is adequate and the discussion is consistent. Therefore, the work deserves to be published.

REFERENCES

1 **Khor B**, Gardet A, Xavier RJ. Genetics and pathogenesis of inflammatory bowel disease. *Nature* 2011; **474**: 307-317 [PMID: 21677747 DOI: 10.1038/nature10209]
2 **Ellinghaus D**, Bethune J, Petersen BS, Franke A. The genetics of Crohn's disease and ulcerative colitis--status quo and beyond. *Scand J Gastroenterol* 2015; **50**: 13-23 [PMID: 25523552 DOI: 10.3109/00365521.2014.990507]
3 **Dykes DM**, Towbin AJ, Bonkowski E, Chalk C, Bezold R, Lake K, Kim MO, Heubi JE, Trapnell BC, Podberesky DJ, Denson LA. Increased prevalence of luminal narrowing and stricturing identified by enterography in pediatric Crohn's disease patients with elevated granulocyte-macrophage colony stimulating factor autoantibodies. *Inflamm Bowel Dis* 2013; **19**: 2146-2154 [PMID: 23893081 DOI: 10.1097/MIB.0b013e31829706e0]
4 **Molodecky NA**, Soon IS, Rabi DM, Ghali WA, Ferris M, Chernoff G, Benchimol EI, Panaccione R, Ghosh S, Barkema HW, Kaplan GG. Increasing incidence and prevalence of the inflammatory bowel diseases with time, based on systematic review. *Gastroenterology* 2012; **142**: 46-54.e42; quiz e30 [PMID: 22001864 DOI: 10.1053/j.gastro.2011.10.001]
5 **McGovern DP**, Kugathasan S, Cho JH. Genetics of Inflammatory Bowel Diseases. *Gastroenterology* 2015; **149**: 1163-1176.e2 [PMID: 26255561 DOI: 10.1053/j.gastro.2015.08.001]
6 **Van Limbergen J**, Radford-Smith G, Satsangi J. Advances in IBD genetics. *Nat Rev Gastroenterol Hepatol* 2014; **11**: 372-385 [PMID: 24614343 DOI: 10.1038/nrgastro.2014.27]
7 **Tsianos EV**, Katsanos KH, Tsianos VE. Role of genetics in the diagnosis and prognosis of Crohn's disease. *World J Gastroenterol* 2011; **17**: 5246-5259 [PMID: 22219593 DOI: 10.3748/wjg.v17.i48.5246]
8 **Jostins L**, Ripke S, Weersma RK, Duerr RH, McGovern DP, Hui KY, Lee JC, Schumm LP, Sharma Y, Anderson CA, Essers J, Mitrovic M, Ning K, Cleynen I, Theatre E, Spain SL, Raychaudhuri S, Goyette P, Wei Z, Abraham C, Achkar JP, Ahmad T, Amininejad L, Ananthakrishnan AN, Andersen V, Andrews JM, Baidoo L, Balschun T, Bampton PA, Bitton A, Boucher G, Brand S, Büning C, Cohain A, Cichon S, D'Amato M, De Jong D, Devaney KL, Dubinsky M, Edwards C, Ellinghaus D, Ferguson LR, Franchimont D, Fransen K, Gearry R, Georges M, Gieger C, Glas J, Haritunians T, Hart A, Hawkey C, Hedl M, Hu X, Karlsen TH, Kupcinskas L, Kugathasan S, Latiano A, Laukens D, Lawrance IC, Lees CW, Louis E, Mahy G, Mansfield J, Morgan AR, Mowat C, Newman W, Palmieri O, Ponsioen CY, Potocnik U, Prescott NJ, Regueiro M, Rotter JI, Russell RK, Sanderson JD, Sans M, Satsangi J, Schreiber S, Simms LA, Sventoraityte J, Targan SR, Taylor KD, Tremelling M, Verspaget HW, De Vos M, Wijmenga C, Wilson DC, Winkelmann J, Xavier RJ, Zeissig S, Zhang B, Zhang CK, Zhao H, Silverberg MS, Annese V, Hakonarson H, Brant SR, Radford-Smith G, Mathew CG, Rioux JD, Schadt EE, Daly MJ, Franke A, Parkes M, Vermeire S, Barrett JC, Cho JH. Host-microbe interactions have shaped the genetic architecture of inflammatory bowel disease. *Nature* 2012; **491**: 119-124 [PMID: 23128233 DOI: 10.1038/nature11582]
9 **Weersma RK**, Stokkers PC, Cleynen I, Wolfkamp SC, Henckaerts L, Schreiber S, Dijkstra G, Franke A, Nolte IM, Rutgeerts P, Wijmenga C, Vermeire S. Confirmation of multiple Crohn's disease susceptibility loci in a large Dutch-Belgian cohort. *Am J Gastroenterol* 2009; **104**: 630-638 [PMID: 19174780 DOI: 10.1038/ajg.2008.112]
10 **Jakobsdottir J**, Gorin MB, Conley YP, Ferrell RE, Weeks DE. Interpretation of genetic association studies: markers with replicated highly significant odds ratios may be poor classifiers. *PLoS Genet* 2009; **5**: e1000337 [PMID: 19197355 DOI: 10.1371/journal.pgen.1000337]
11 **Evans DM**, Visscher PM, Wray NR. Harnessing the information contained within genome-wide association studies to improve individual prediction of complex disease risk. *Hum Mol Genet* 2009; **18**: 3525-3531 [PMID: 19553258 DOI: 10.1093/hmg/ddp295]
12 **Kang J**, Kugathasan S, Georges M, Zhao H, Cho JH; NIDDK IBD Genetics Consortium. Improved risk prediction for Crohn's disease with a multi-locus approach. *Hum Mol Genet* 2011; **20**: 2435-2442 [PMID: 21427131 DOI: 10.1093/hmg/ddr116]
13 **Adler J**, Rangwalla SC, Dwamena BA, Higgins PD. The prognostic power of the NOD2 genotype for complicated Crohn's disease: a meta-analysis. *Am J Gastroenterol* 2011; **106**: 699-712

[PMID: 21343918 DOI: 10.1038/ajg.2011.19]
14 **Dudbridge F**. Power and predictive accuracy of polygenic risk scores. *PLoS Genet* 2013; **9**: e1003348 [PMID: 23555274 DOI: 10.1371/journal.pgen.1003348]
15 **Wei Z**, Wang W, Bradfield J, Li J, Cardinale C, Frackelton E, Kim C, Mentch F, Van Steen K, Visscher PM, Baldassano RN, Hakonarson H. Large sample size, wide variant spectrum, and advanced machine-learning technique boost risk prediction for inflammatory bowel disease. *Am J Hum Genet* 2013; **92**: 1008-1012 [PMID: 23731541 DOI: 10.1016/j.ajhg.2013.05.002]
16 **Repnik K**, Potočnik 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* 2011; **46**: 1081-1091 [PMID: 21695374 DOI: 10.1007/s00535-011-0426-6]
17 **Wu J**, Pfeiffer RM, Gail MH. Strategies for developing prediction models from genome-wide association studies. *Genet Epidemiol* 2013; **37**: 768-777 [PMID: 24166696 DOI: 10.1002/gepi.21762]
18 **Hosmer JDW**, Lemeshow S, Sturdivant RX. The Multiple Logistic Regression Model. Applied Logistic Regression: John Wiley & Sons, Inc., 2013: 35-47
19 **Grimes DA**, Schulz KF. Refining clinical diagnosis with likelihood ratios. *Lancet* 2005; **365**: 1500-1505 [PMID: 15850636 DOI: 10.1016/s0140-6736(05)66422-7]
20 **Lalkhen GA**, McCluskey A. Clinical tests: sensitivity and specificity. *Contin Educ Anaest Crit Care Pain* 2008; **8**: 221-223 [DOI: 10.1093/bjaceaccp/mkn041]
21 **Jostins L**, Barrett JC. Genetic risk prediction in complex disease. *Hum Mol Genet* 2011; **20**: R182-R188 [PMID: 21873261 DOI: 10.1093/hmg/ddr378]
22 **Liu JZ**, Anderson CA. Genetic studies of Crohn's disease: past, present and future. *Best Pract Res Clin Gastroenterol* 2014; **28**: 373-386 [PMID: 24913378 DOI: 10.1016/j.bpg.2014.04.009]
23 **Nelis M**, Esko T, Mägi R, Zimprich F, Zimprich A, Toncheva D, Karachanak S, Piskácková T, Balascák I, Peltonen L, Jakkula E, Rehnström K, Lathrop M, Heath S, Galan P, Schreiber S, Meitinger T, Pfeufer A, Wichmann HE, Melegh B, Polgár N, Toniolo D, Gasparini P, D'Adamo P, Klovins J, Nikitina-Zake L, Kucinskas V, Kasnauskiene J, Lubinski J, Debniak T, Limborska S, Khrunin A, Estivill X, Rabionet R, Marsal S, Julià A, Antonarakis SE, Deutsch S, Borel C, Attar H, Gagnebin M, Macek M, Krawczak M, Remm M, Metspalu A. Genetic structure of Europeans: a view from the North-East. *PLoS One* 2009; **4**: e5472 [PMID: 19424496 DOI: 10.1371/journal.pone.0005472]
24 **Alesina AF**, Easterly W, Devleeschauwer A, Kurlat S, Wacziarg RT. Fractionalization (June 2002). Harvard Institute Research Working Paper No. 1959. Available from: URL: http://ssrn.com/abstract=319762
25 **Abraham G**, Inouye M. Genomic risk prediction of complex human disease and its clinical application. *Curr Opin Genet Dev* 2015; **33**: 10-16 [PMID: 26210231 DOI: 10.1016/j.gde.2015.06.005]
26 **Golan D**, Rosset S. Effective genetic-risk prediction using mixed models. *Am J Hum Genet* 2014; **95**: 383-393 [PMID: 25279982 DOI: 10.1016/j.ajhg.2014.09.007]
27 **Yarwood A**, Han B, Raychaudhuri S, Bowes J, Lunt M, Pappas DA, Kremer J, Greenberg JD, Plenge R, Worthington J, Barton A, Eyre S. A weighted genetic risk score using all known susceptibility variants to estimate rheumatoid arthritis risk. *Ann Rheum Dis* 2015; **74**: 170-176 [PMID: 24092415 DOI: 10.1136/annrheumdis-2013-204133]
28 **Ek WE**, D'Amato M, Halfvarson J. The history of genetics in inflammatory bowel disease. *Ann Gastroenterol* 2014; **27**: 294-303 [PMID: 25331623]
29 **Ankerst DP**, Thompson IM. Sensitivity and specificity of prostate-specific antigen for prostate cancer detection with high rates of biopsy verification. *Arch Ital Urol Androl* 2006; **78**: 125-129 [PMID: 17269614]
30 **Yin X**, Cheng H, Lin Y, Wineinger NE, Zhou F, Sheng Y, Yang C, Li P, Li F, Shen C, Yang S, Schork NJ, Zhang X. A weighted polygenic risk score using 14 known susceptibility variants to estimate risk and age onset of psoriasis in Han Chinese. *PLoS One* 2015; **10**: e0125369 [PMID: 25933357 DOI: 10.1371/journal.pone.0125369]
31 **Eichler EE**, Flint J, Gibson G, Kong A, Leal SM, Moore JH, Nadeau JH. Missing heritability and strategies for finding the underlying causes of complex disease. *Nat Rev Genet* 2010; **11**: 446-450 [PMID: 20479774 DOI: 10.1038/nrg2809]

P- Reviewer: Goral V, Sipahi AM **S- Editor**: Gong ZM
L- Editor: A **E- Editor**: Ma S

Published by **Baishideng Publishing Group Inc**
8226 Regency Drive, Pleasanton, CA 94588, USA
Telephone: +1-925-223-8242
Fax: +1-925-223-8243
E-mail: bpgoffice@wjgnet.com
Help Desk: http://www.wjgnet.com/esps/helpdesk.aspx
http://www.wjgnet.com

ISSN 1007-9327