Systematic analysis of genomic copy number variations in inflammatory bowel diseases

Dissertation zur Erlangung des Doktorgrades
der Mathematisch-Naturwissenschaftlichen Fakultät
der Christian-Albrechts-Universität zu Kiel

vorgelegt von

Hamidreza Saadati

Kiel, Dezember 2016

Referent/in: **Prof. Dr. Andre Franke**

Koreferent/in: **Prof. Dr. Manuela Dittmar**

Tag der mündlichen Prüfung: **15.12.2016**

gez. **Prof. Dr. Natascha Oppelt, Dekanin**

Table of contents

List of Figures

List of Tables

1 Introduction

1.1 Inflammatory bowel diseases

Since the beginning of the 20th century and especially over its second half, there has been a significant increase in the incidences of chronic inflammatory disease, which seems to follow a geographical pattern of industrialization and urban living (Bach *et al.*, 2002, Eder *et al.*, 2006). These inflammatory phenotypes include allergic conditions such as asthma, food allergies and eczema as well as autoimmune disorders such as type 1 diabetes, chronic inflammatory bowel disease and neurodegenerative disease (figure1.1). Inflammatory bowel diseases (IBD; OMIM 601458), as one of such phenotypes, are chronic, relapsing inflammatory disorders of the gastrointestinal tract (see figure 1.2), with a peak age of onset in the second to fourth decades of life (Podolsky *et al.*, 2002). Among the chronic inflammatory diseases, however, the unique feature of IBD is the close apposition of the intestinal immune system to high concentrations of bacteria in the lumen of the gut. The gut contributes essentially to absorb nutrients from digested food and process waste for elimination.

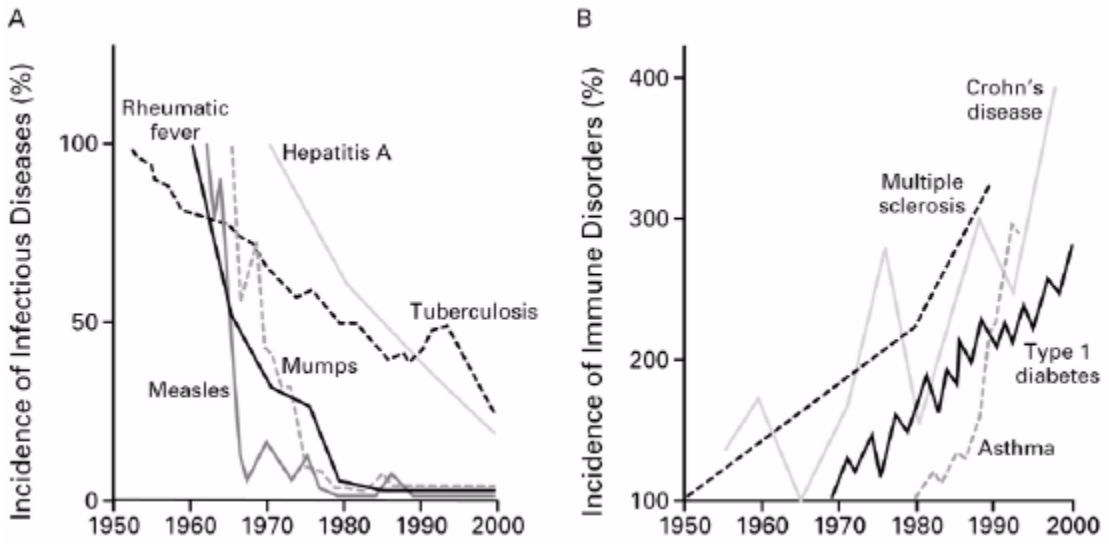


Figure 1.1 Inverse relation between the Incidence of Prototypical Infectious Diseases (A) and the Incidence of Immune Disorders (B) from 1950 to 2000. Reprinted from Bach *et al.* (2002).

These activities are performed at the inner mucosal surface, which consists of a thin, permeable epithelium (see figure 1.2). In the small intestine, this surface is greatly expanded by the presence of fingerlike villi, making the mucosal lining of the gut the largest surface in the body. The lumen is a nutrient-rich microenvironment with a complex microbial population that has coevolved with the host. These commensal microbes perform essential functions, such as digestion of complex carbohydrates and production of vitamins and other small molecules (Bäckhed *et al.*, 2005). Microbial density in the colon has been estimated to reach 10^{11}–10^{12} cells per gram of luminal contents (Whitman *et al.*, 1998).

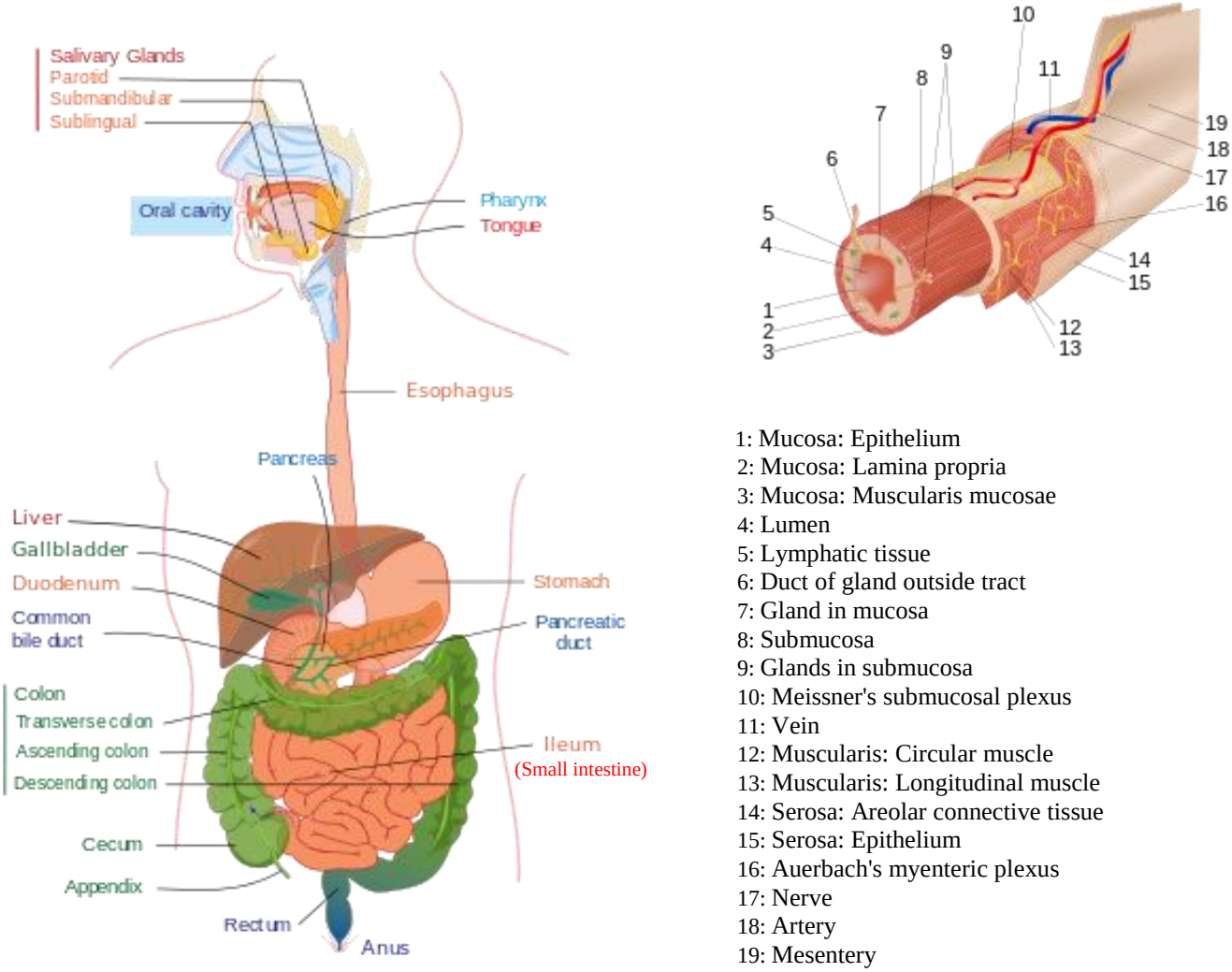


Figure 1.2 Upper and Lower human gastrointestinal tract (right) and General structure of the gut wall (Left). Illustration from https://en.wikipedia.org/wiki/Human_gastrointestinal_tract

More recently modern high throuput metagenomic analysis have shown that the normal gut comprises 100 trillion diverse microbes, mostly bacteria, encompassing over 1100 prevalent species, with at least 160 species in each individual (Qin *et al*, 2010). This creates a key challenge for the intestinal immune system to prevent commensal and pathogenic microbes from crossing the gut epithelial barrier. There is strong evidence to support that dysregulated interaction of the host immune system with the commensal microflora and other luminal antigens leads to IBD (Xavier *et al*, 2007; Kamada *et al*, 2013; Belkaid *et al*, 2014).

IBD patients typically suffer from frequent and chronically relapsing flares, resulting in diarrhea, abdominal pain, rectal bleeding and malnutrition. (Podolsky *et al*, 2002). IBD is divided in two main sub-phenotypes of Crohn's disease (OMIM 266600) and ulcerative colitis (OMIM 191390). Crohn's disase (CD) was first recognized by German surgeon Wilhelm Fabry (aka Guilhelmus Fabricius Hildanus) in 1623 and was later described by and named after the US physician Burril B Crohn (Crohn, *et al*, 1984). Ulcerative colitis (UC) was first described by the British physician Sir Samuel Wilks in 1859. CD can be distinguished from UC, in that it is characterized by inflammation that can extend into all layers of the bowel wall (transmural). Areas of deep ulceration can form localized regions of nodular inflammation (granulomas) (see also figure 1.3) or tube-like connections between loops of the intestines or nearby organs (fistulas).

Table 1.1. **Comparison of CD and UC**.

Feature	Crohn's disease	Ulcerative colitis
Location	Any portion of the gastrointestinal tract, most commonly the ileum and colon	Inflammation confined to the colon
Pathology	Typically discontinuous, often transmural, and with granulomas	Continuous, confined to the mucosa and submucosa
Risk with tobacco	Increased in smokers	Increased in ex-smokers

However, the most commonly affected segment is the terminal ileum. Another distinguishing feature of CD is its segmental distribution, so that regions of inflammation can be separated by tissue with normal appearance. By contrast, the inflammation seen in UC is restricted to the mucosa and, as the disease progresses, the submucosa. Intestinal fistulas, granulomas and deep fissures are not found in UC, and the inflammation usually involves the rectum and extends proximally to include part of or the entire colon. The region of inflammation in UC is continuous without any skipped segments (Podolsky *et al*, 2002). In ~5% of IBD cases, it is not possible to assign a definitive diagnosis of CD or UC, because of the overlap in their pathologies (Friedman, *et al*, 2008). Patients with IBD often have various extra-intestinal symptoms such as arthalgias (inflammatory paints in the joints) and are more likely to have other chronic inflammatory diseases, particularly primary sclerosing cholangitis (Saich *et al*, 2008), ankylosing spondylitis and psoriasis (Lees *et al*, 2011, parkes *et al*, 2013).

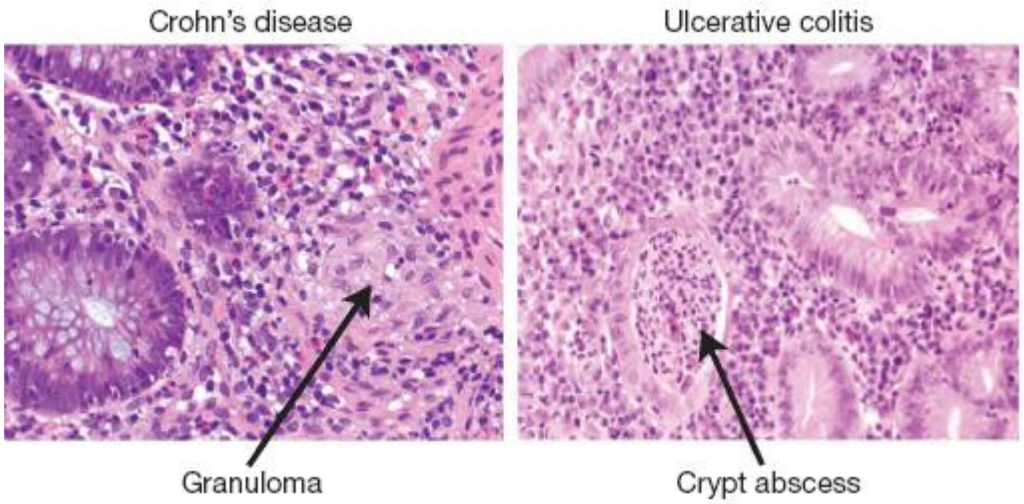


Figure 1.3 Histologic hallmarks of IBD.

Left panel, CD-biopsy from a terminal ileum with active disease. The figure illustrates a discrete granuloma composed of compact macrophages, giant cells and epithelioid cells. Surrounding the nodule there is marked infiltration of lymphoid cells, plasma cells and other inflammatory cells, but there is no necrosis. Right panel, UC-colonic mucosal biopsy taken from a patient with active disease. The crypt abscess is composed of transmigrated neutrophils and the surrounding epithelium exhibits features of acute mucosal injury. Illustration from Xavier *et al.* (2007).

1.1.1 Epidemiology of IBD

The highest prevalence of UC and CD has been reported from northern Europe, the UK, and North America, with the average number of cases ranging from 100 to 200 cases per 100,000 persons (Loftus *et al.*, 2000; Bernstein *et al.*, 1999). Most reports show CD to be more common in European Americans (~ 43 per 100,000) compared with African Americans (~30 per 100,000), with the lowest rates reported in Hispanics and Asians (~ 5 per 100,000) (Binder *et al.*, 1998). Despite this fact, however rates continue to rise in low-incidence areas such as southern Europe, Asia and most developing countries. Within population of European ancestry, CD is more prevalent in persons of Jewish descent than in any other ethnic group (Sandler *et al.*, 1994). The phenomenon of genetic anticipation, which means earlier onset in offspring of people with the disease, has been reported for North American Ashkenazi Jews (Lowe *et al.*, 2009). Different incidence rates, especially the lower incidence of IBD in Asia and Africa compared with North America and Europe could have resulted from different genetic backgrounds. Despite this, epidemiological studies have shown that prevalence of IBD in North America and northern Europe increased rapidly during the early- and mid-20th century and then stabilized at higher levels (Molodecky *et al.*, 2012). Additionally incidence rates of IBD continue to rise in low-incidence areas such as southern Europe, Asia and the majority of the developing world (Loftus *et al.*, 2004). These evidences have indicated that there is a substantial environmental or lifestyle component to disease risk. Factors such as diet, breastfeeding, oral contraceptives and childhood infections have been proposed to contribute to the etiology of IBD (Podolsky *et al.*, 2002; Freidman *et al.*, 2008). However, it has been argued that those with the strongest association are cigarette smoking and appendectomy (Loftus *et al.*, 2004). Beside these factors, etiology of autoimmune and inflammatory diseases like IBD has been also attributed to the 'hygiene hypothesis', which proposes that lack of exposure to certain infectious agents in childhood results in an overactive immune response in later life (Gent *et al.*, 1994). On the other hand, there exist evidences for the hereditary contribution to

IBD. A positive family history is the largest independent risk factor for the disease, as population-based studies have found that 5–10% of patients, have a first-degree family member with IBD (Orholm *et al*, 1991). Further, the relative risk to siblings of affected individuals has been estimated to increase 15–35 fold for CD and 6–9-fold for UC (Binder *et al*, 1998, Bengtson *et al*, 2009). Twin studies provide additional evidence for a genetic contribution in IBD. The concordance rate is significantly greater in MZ twins than DZ ones (Brant, 2011) for both CD (30.3% versus 3.6%) and UC (15.4% versus 3.9%). Although the genetic contribution to risk is stronger in CD, multiple studies show that relatives with either CD or UC are at increased risk of developing either form of IBD, indicating the existence of both phenotype-specific and shared susceptibility mechanisms for UC and CD (Budarf *et al*, 2009). These observations have implicated that IBD, like most other common diseases, have a complex etiology involving multiple genetic and environmental factors.

1.1.2 Genetic architecture of IBD

1.1.2.1 Gene discovery approaches

As discussed in previous section, familial aggregation, significant increased sibling relative risk and twin studies have demonstrated the genetic contribution to IBD. Despite this, the underlying genetic loci were mostly unknown until 2000. In 1996 the first attempts at identifying the CD genetic risk factors used linkage mapping, which identified the peri-centromeric region on chromosome 16 called IBD1 (Hugot *et al*, 1996) and chromosome 12q called IBD2 (Satsangi *et al*. 1996). Family-based **linkage studies** identify large segments of human genomes (mostly megabases in size and therefore containing multiple genes), which are shared among affected relatives more frequently than expected by chance. In 2001, the first IBD gene, *NOD2 (*nucleotide-binding oligomerization domain containing 2; also known as caspase recruitment domain protein 15, *CARD15*) was identified through association mapping of the linkage region on chromosome 16 (Hugot *et al,* 2001). Genome-wide linkage scans

analyzed a relatively small number (300–5,000) of genetic markers and have been highly successful at finding genetic loci with high penetrances in rare single-gene (Mendelian) disease phenotypes (Risch, *et al*, 2000; Rioux *et al*, 2005). After few early findings of genetic linkage studies for IBD, two main developments i.e. the International **HapMap** Project (http://www.hapmap.org/) and the development of microarray genotyping platforms allowed the design of powerful population-based **genome-wide association (GWAs)** studies for common complex diseases like IBD. The sequencing of the human genome (International **Human Genome Sequencing Consortium**, 2004) and the generation of public resources of single nucleotide polymorphisms (SNPs), especially **dbSNP** (Smigielski *et al*, 2000) have contributed essentially to these developments. SNPs are inter-individual differences in a nucleotide base at a given site in the genomic DNA sequence. It has been estimated that in the world's human population, about 10 million sites (that is, one variant per 300 bases on average) vary such that both alleles are observed at a frequency of ≥1% (Kruglyak *et al*, 2001). The specific set of the alleles, who are located on the same chromosome and segregate together, is called a **haplotype**. Sequence variants in a haplotype are said to be in **Linkage disequilibrium (LD),** which refers to the phenomenon that two or more alleles in a chromosomal region occur together more often than accounted for by chance. This mostly indicates that the alleles are in close proximity on the DNA strand and are most likely to be passed on together within a population (The international HapMap project, 2003). Based on the haplotype pattern map in human genome, a genome wide subset (several thousands up to a million) of SNPs, known as **tagSNPs,** are selected and genotyped simultaneously for unrelated disease cases and healthy controls on microarray platforms (Xavier *et al*, 2008). GWAS are largely hypothesis-free approaches that identify common risk alleles, which are significantly more frequent in patients compared to healthy individuals. Since 2006, the year of the first published GWAS study on IBD using > 100,000 SNPs (Duerr *et al*, 2006), there has been an exponential growth in the set of validated genetic risk factors for IBD. This has been

mainly achieved through Meta-analyses of SNP-GWAS data, which provided increased power towards indentifying associated variants with small effect sizes for both CD (Franke at al., 2010) and UC (Anderson, C. A. *et al.* 2011). In 2012, the largest genetic association study for IBD employed GWAs data set of over 75,000 patients and controls and established the association of 163 susceptibility loci (Jostins *et al.*, 2012). Despite this large number of identified associated loci, two main issues still need to be addressed; First, SNPs identified through GWA studies do not mostly constitute the actual causal variants, instead point to genomic regions that vary in size, depending on the extent of the local LD pattern (Kohr *et al.*, 2011). Some of these identified regions contain only one gene, some contain several genes and others do not contain any known coding sequence (gene deserts). An example in

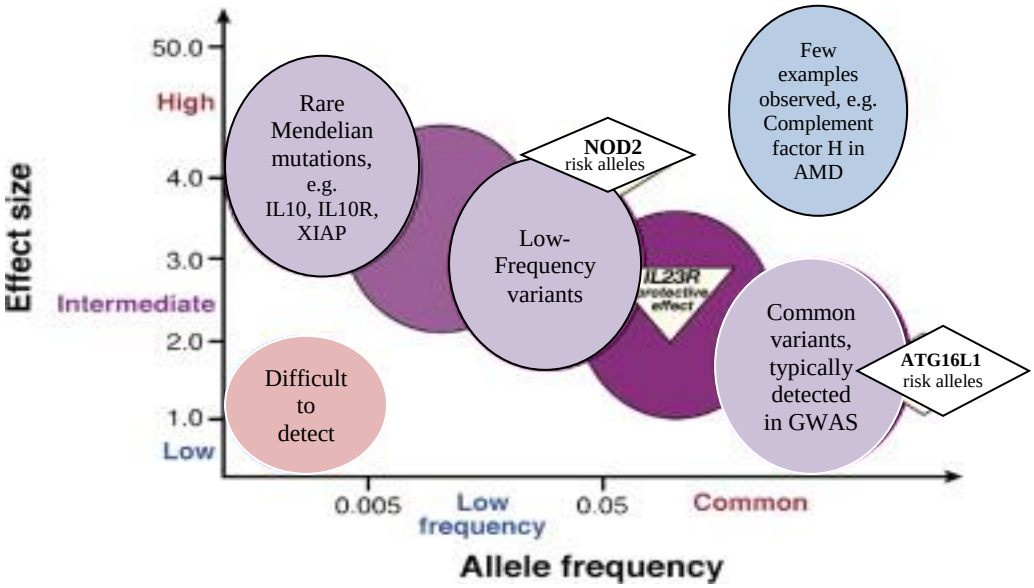


Figure 1.4 Features of disease-associated genetic variation. Most cases of IBD are multifactorial in etiology, reflecting the effects of multiple genetic risk alleles and developmental and environmental factors. Rare cases of early-onset IBD, with extreme phenotypes, might be single-gene, Mendelian disorders (e.g., autosomal recessive mutations in IL10RA or ILRB or X-linked inhibitor of apoptosis protein (XIAP)) - extremely rare mutations with strong effects. In contrast, most of the alleles identified by GWAS are relatively common (allele frequencies > 5%), with modest or low effects (odds ratios < 1.5). Greater effects (intermediate to high) include the relatively low-frequency risk alleles in *NOD2* and the variant that encodes Arg381Gln in *IL23R*. Among complex diseases, it is unusual to have common alleles that have strong effects; an exception is that some alleles of the gene that encodes complement factor H confer high-risk for age-related macular degeneration (AMD). Adopted from Cho *et al.*, 2011.

this context is the association of CD with a region of chromosome 5p13 (Libioulle *et al.*, 2007), which contains no known genes. Further investigation implicated that the identified variants in the region correlated with differential expression of the nearby prostaglandin E receptor 4 (PTGER4) gene (Libioulle *et al.*, 2007), indicating that functional studies are needed to characterize the actual effect of the associated variants and the underlying biological mechanism. The second challenge is that nearly all the identified associated variants have low to modest effect sizes (odds ratio (OR) <1.3) and therefore explain only 13.6% and 7.5% of disease variance for CD and UC respectively (Jostins *et al.*, 2012). Odds ratio is a measure of effect size, defined as the ratio of the odds (the probability of disease divided by 1 minus the probability) of a disease being observed in one group of genotypes and the odds of a disease being observed in another group (Nurminen *et al.*, 1995).

It has been assumed that all risk loci with a minor allele frequency > 5% in the general population and an OR > 1.2 have already been identified in IBD patients with European ancestry by means of GWAS (Liu, J.Z. *et al.*, 2014). Genetic factors such as rare variants, private mutations, structural variants and interactions between genes have, however, not been captured by GWAS studies (Eichler, *et al.*, 2010).

1.2.1.1 Identified genes and implicated pathways

Of the 163 genetic risk loci identified for IBD till now, 110 loci are associated with both disease phenotypes of CD and UC (Jostins *et al.*, 2012), indicating that despite distinct clinical characteristics, both phenotypes have common pathways involved in their pathogenesis. Analyses of the genes and genetic loci identified in IBD show several pathways that are significant for intestinal homeostasis. These include barrier function, epithelial restitution, microbial defense, innate immune regulation, reactive oxygen species (ROS) generation, autophagy, regulation of adaptive immunity, endoplasmic reticulum (ER) stress and metabolic pathways associated with cellular homeostasis (see also Figure 1.6). (Khor *et al.*, 2011; Cho *et al.*,

2011). Intestinal homeostasis involves the coordinated actions of epithelial, innate and adaptive immune cells. Barrier permeability might permit microbial incursion, which is then often detected by the innate immune system and leads to appropriate tolerogenic, inflammatory and restitutive reactions. These responses are mainly activated by releasing extracellular mediators known as cytokines, which instead could recruit other cellular components related to adaptive immune system (Kaser *et al*, 2009). Here some of the most relevant pathways and essential genes contributing to IBD pathogenesis are described. More comprehensive and detailed review of the pathobiological mechanisms and cellular pathways involved in IBD can be seen in Khor *et al*, 2011 and Parkes *et al*, 2013.

Epithelial barrier integrity

In addition to nutrient absorption, intestinal epithelial cells perform both barrier and signal transduction functions, so that they recognize luminal contents through surface receptors and in return secret regulatory products that can mediate appropriate responses in the underlying lamina propria (see figure 1.2). Abnormal intestinal permeability has been observed in IBD patients and in some of their first degree relatives (Kaser *et al*, 2009). Several IBD loci contain genes such as *CDH1*, *GNA12* and *PTPN2,* which contribute to epithelial integrity. *CDH1* encodes adherent junction protein *E-cadherin* and variations leading to truncated form of the protein have been associated with CD (Muise *et al*, 2009). *GNA12* encodes the G protein *Gα12* and its activation leads to phosphorylation of the tight junction proteins *ZO-1* and *ZO-2*, resulting in destabilization of cell junctions in epithelial cell lines (Sabath *et al*, 2008). *In vitro* studies show that the protein tyrosine phosphatase family member *PTPN2* protects against interferon-γ (IFN-γ)-induced epithelial permeability (Scharl *et al*, 2009). Concordantly, *Ptpn2*-deficient mice show increased susceptibility to experimental colitis (Hassan, *et al*, 2010).

Innate immune responses

Mucosal innate immune system defends against pathogenic factors and simultaneously regulates inflammatory responses to maintain a state of controlled responsiveness to

commensal bacteria. Dendritic cells, macrophages, innate lymphoid cells (ILCs) and neutrophils are essential cellular components of the innate immune system during infection or inflammation (Nochi *et al*, 2006). Patients with innate immunodeficiencies such as chronic granulomatous disease and Hermansky–Pudlak syndrome (a syndrome associated with defective responses to bacterial DNA motifs (CpG oligonucleotides) specifically in plasmacytoid dendritic cells) tend to develop IBD (Blasius *et al*, 2010). Some major components of the innate immunity implicated in IBD are introduced here.

Microbe Recognition

Cells of the innate immune system have pattern-recognition receptors that recognize microbe-specific macromolecules, enabling them to target the pathogens. *NOD2* was the first gene to be associated with CD (Hugot *et al*, 2001) and after that several genes interacting with *NOD2* signaling were also identified as associated with IBD. *NOD2* is expressed by many leukocytes, including antigen presenting cells, macrophages and lymphocytes as well as ileal Paneth cells and is an essential component for microbial sensing (Shaw *et al*, 2011). Activation of *NOD2* by microbial ligands activates the transcription factor nuclear factor−κB (*NF-κB*) and mitogen-activated protein kinas signaling and thereby functions as a positive regulator of immune defense (Abraham *et al*, 2006). The most common mutations in *NOD2* that are associated with CD (Arg702Trp [rs2066844], Gly908Arg [rs2066845], and Leu1007fsinsC [rs41450053]) lie either within or near the C-terminal, leucine-rich repeat domain, which is required for microbial sensing (Ogura *et al*, 2001). *NOD2* mutations are consistently associated with ileal and stricturing (transmural) CD (Economou *et al*, 2004).

CARD9 is an adaptor protein that integrates signals from many innate immune receptors that recognize viral, bacterial and fungal motifs. Depending on the stimulus, *CARD9* interacts with distinct signaling complexes and activates different pathways to modulate cytokine environments appropriately (Hsu *et al*, 2007). Defective CARD9 function leads to the immune

deficiency, at least in part owing to failure to promote an adequate T_H17 immune response (Poeck, H. *et al*, 2010).

Autophagy

Autophagy degrades damaged organelles and proteins and is important for the clearance of pathogens (xenophagy), which is required for immunity to multiple different types of bacteria. Genetic analyses have implicated an essential role for autophagy in innate immunity and IBD, indicating two component genes of *ATG16L1* and *IRGM* in IBD pathogenesis (Rioux, *et al.* 2007; McCarroll *et al*, 2008). *ATG16L1 (Autophagy 16-like 1*) is widely expressed, including in small intestinal Paneth cells, where it mediates exocytosis (secretion) of secretory granules that contain antimicrobial peptides (Cadwell *et al*, 2008). A single Thr300Ala (*Threonine* to *Alanine*) substitution in *ATG16L1* results in decreased capability to capture bacteria and has been shown to be associated with CD risk (Kuballa *et al*, 2008).

Additionally leucine-rich repeat kinase 2 (*LRRK2)* also regulate autophagy and is located in the CD-associated region on chromosome 12q12, (Barrett *et al*, 2008) along with *MUC19*. Alterations in autophagy have important roles in pathogenesis of CD, possibly because of the close apposition of microbial components with high cellular turnover of the intestinal environment.

Oxidative stress

The equilibrium between oxidant factors, such as free radicals, reactive oxygen species (ROS) or reactive nitrogen species, and antioxidant components, such as glutathione peroxidase (GPX) and glutathione *S*-transferase enzymes is important in gut homeostasis, as the reduction-oxidation-state affects many signal transduction pathways (Schroeder *et al*, 2011). Genes within several IBD loci may either regulate ROS production or protect against oxidative stress. In particular, *NOD2, CARD9* and *IFN-γ*-regulated leucine-rich repeat kinase 2 (*LRRK2*) contribute to ROS production (Wu *et al*, 2009). In addition to pro-inflammatory pathways, ROS are also involved in T_{reg}-cell polarization and function (Efimova *et al*, 2011).

Adaptive immune responses

If a microbial invasion cannot be controlled by the innate immune system, inflammatory signals that activate the adaptive immune response are released. This response is mediated by lymphocytes (B and T cells) and has functional specificity, but might require several days to reach an effective immune response. Here some of the essential cellular processes and mechanisms of adaptive immunity implicated in IBD are described.

Lymphocyte Activation

In the healthy gut, naïve T cells receive cytokine signals, such as transforming growth factor β (TGF-β) and interleukin-10 (IL-10), which stimulate them to differentiate along a tolerogenic pathway to become regulatory T cells. During infection or chronic inflammation, naïve T cells receive a different set of signals from antigen-presenting cells, which lead them to be differentiated into T_H1, T_H2 or T_H17 helper cells and thereby inflammatory pathways are activated.

Human leukocyte antigens **HLA** class II genes are a major player in T cell activation and have been significantly associated with UC and several other autoimmune diseases (Gregersen *et al*, 2009). These genes are a member of the major histocompatibility complex **(MHC),** which comprises a contiguous 4 Mb region on the short arm of chromosome 6. The extended MHC (xMHC), as its name suggests, spans an even larger 7.6 Mb region and comprises more than 400 annotated genes and pseudogenes (Horton *et al*, 2004). HLA genes are frequently associated with chronic inflammatory genetic disorders, probably because of the enormous genetic and functional diversity contained within this region as well as its essential role in regulating interactions between host cells and pathogens.

IL-23 signaling pathway

One of the strongest associations observed in GWAs of CD is in the gene region encoding interleukin-23 receptor (*IL-23R*). Disease-associated *IL23R* polymorphisms have also been reported in UC patients (Duerr *et al*, 2007). It has been recognized that *IL-23* drives a

pathogenic T cell population with a distinct inflammatory transcription profile, which includes putative cytokines specially *IL17* and contributes essentially to autoimmune inflammation (Langrish *et al*, 2005). Based on this distinct gene transcription profile, harboring unique Janus-kinase (*JAK*)–signal transducer and activator of transcription (*STAT*) pathway, a novel subset of T helper (T_H) cells i.e. Th_{IL-17} or Th17 was discovered (Harrington *et al*, 2005). It has been proposed that an important component in association of *IL23* signalling with IBD is mediated by its roll in inducing *IL-17* expression in T_H17 cells (Gaffen *et al*, 2014). In addition to the receptor, other components of the *IL-23R* signaling pathway i.e. *IL12B, STAT3* and *JAK2* have

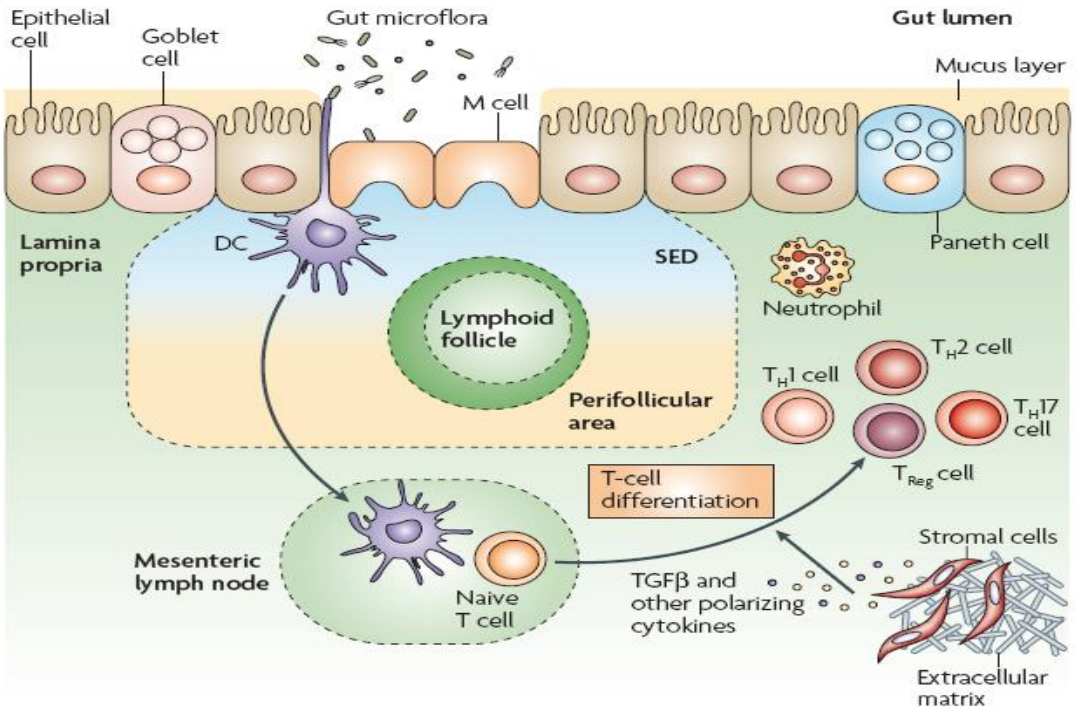


Figure 1.5 key features of the intestinal immune system

The epithelial-cell layer is comprised of absorptive and secretory cells, goblet cells and Paneth cells. Goblet cells contribute to the formation of the protective mucus layer. Microfold cells (M cells) and dendritic cells (DCs) sample intestinal luminal contents. The presence of either pathogenic bacteria or disruption of the epithelial-cell barrier results in activation and migration of DCs to the mesenteric lymph nodes, where they activate naive T cells, which then undergo differentiation under the influence of factors released by DCs and other stromal elements. SED, subepithelial dome; TGFβ transforming growth factor-β TH, T helper; TReg, T regulatory. Illustration from Cho *et al.*, 2008.

also shown strong association with both CD and UC (Barrett *et al*, 2008; Franke *et al*, 2008). The leading hypothesis is that the *IL-23R* signaling pathway contributes to immunopathogenesis of IBD by promoting the pro-inflammatory state. An analysis of the IBD loci list (with 300 prioritized genes) for enrichment in Gene Ontology terms, showed that after excluding high level categories like "immune system processes", the most significant enriched category comprised regulation of cytokine production, specifically interferon-γ, *IL-12*, tumor-necrosis factor-α and *IL-10* signalling (Jostins *et al*, 2012). Lymphocyte activation was the next most significant, with activation of T cells, B cells and natural killer 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 *JAK-STAT* signalling pathway (Jostins *et al*, 2012).

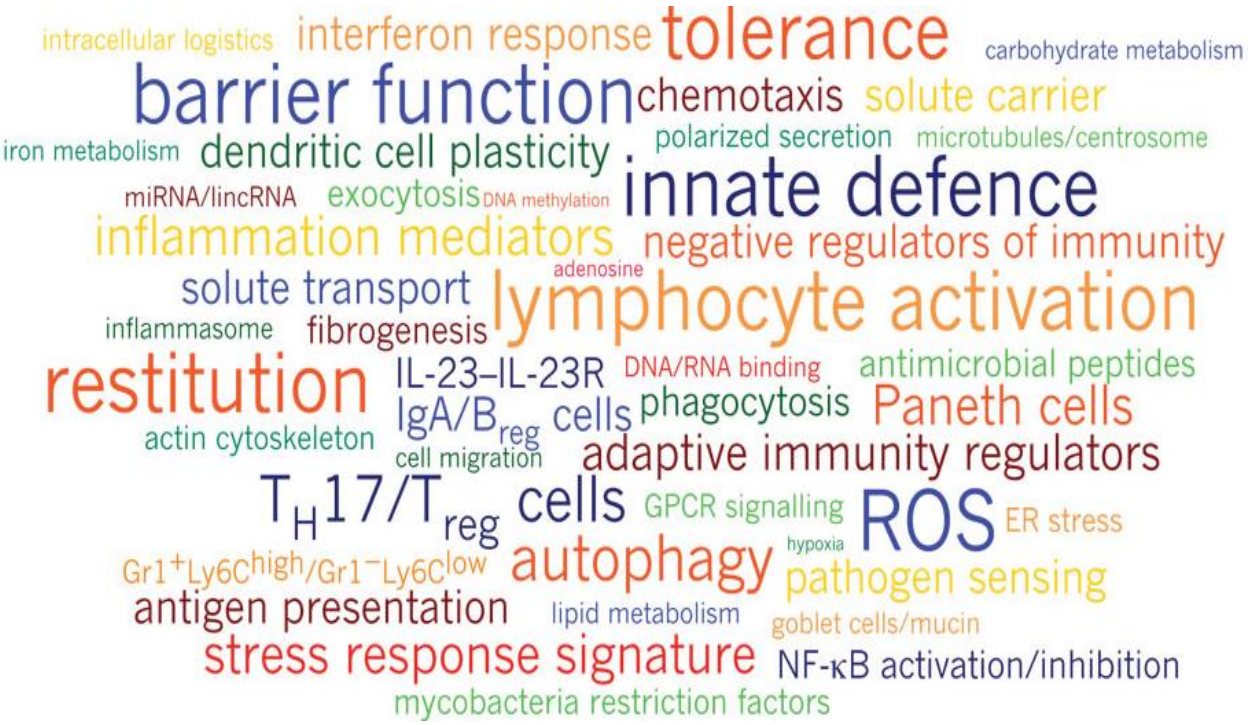


Figure 1.6 Biological processes implicated by IBD loci.

Font sizes are proportional to the number of genes associated with each respective process. Breg cells, B regulatory cells; ER, endoplasmic reticulum; GPCR, G-protein-coupled receptor; IL, interleukin; lincRNA, large intervening non-coding RNA; miRNA, microRNA; ncRNA, non-coding RNA; NF-κB, nuclear factor-κB; ROS, reactive oxygen species; TH17 cells, T helper 17 cells; Treg cells, T regulatory cells. Illustration from Khor *et al*, 2011.

1.2 Copy number variations (CNVs)

The landscape of genetic variations in humans includes single nucleotide substitutions, structural variations, ranging in size from 50 base pairs (bp) to more than one mega-base-pair (Mb), and large chromosomal aberrations (> 1-3 Mb) (see figure 1.7). Variations at either extreme of this spectrum, i.e. single-nucleotide polymorphisms (SNPs) and cytogenetically recognizable chromosomal changes (Jacobs *et al*, 1992) have long been known. About 10 years ago, scientists began to recognize abundant variation of the intermediate-size class known as structural variations (Iafrate *et al*, 2004. Sebat *et al*, 2004.Redon *et al*, 2006. Tuzun *et al*. 2005). Within this class, CNVs include insertions, deletions and duplications of genome sequences, typically defined to be 50 bp or larger. Smaller genomic segments (< 50 bp) with variable copy number are referred to as small insertion/deletions (indels) (Conrad *et al*, 2010; MacDonald *et al*, 2014).

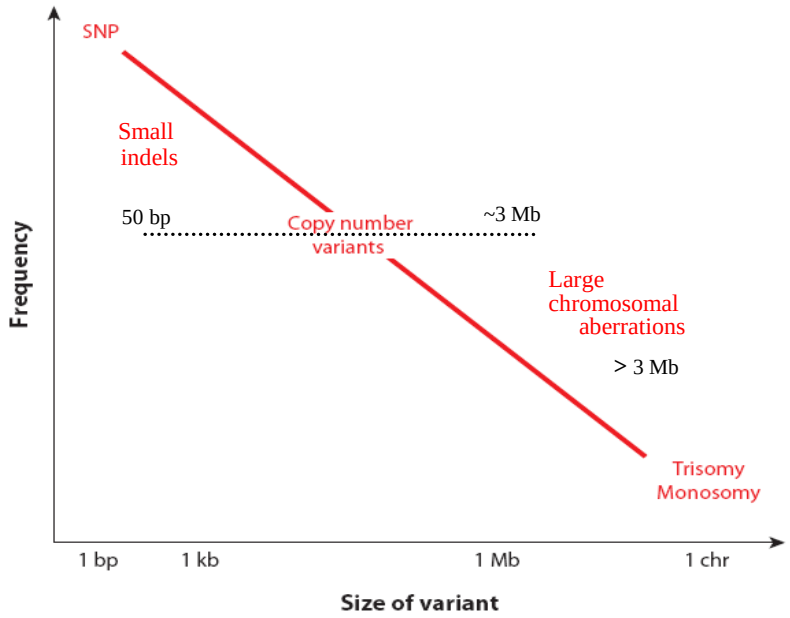


Figure 1.7 Major categories of genetic variants. Copy number variants are imbalanced structural variants within the size range of ~50 bp to 1-3 Mb. Insertions and deletions smaller that 50 bp are usually called small indels. Large chromosomal aberrations (>3 Mb, microscopically visible after G-banding) are rare and often associated with major congenital disorders. Chromosomal monosomies and trisomies represent extreme instances of chromosomal abnormalities (e.g., trisomy 21 associated with Down syndrome). Mb (mega base pair), kb (kilo base pair), bp (base pair).

Other types of structural variations include copy-number-balanced rearrangements such as inversions, which change the orientation of a DNA segment, and translocations, in which a DNA segment is reciprocally exchanged between two chromosomes (see figure 1.8). These rearrangements change the spatial organization of DNA, but unlike CNVs do not result in any net gain or loss of sequence.

The widespread occurrence and high prevalence of CNVs in human genome was first recognized through large-scale population studies, carried out by Sebat *et al.* and Iafrate *et al.* in 2004. Afterwards a survey of 270 individuals from the human HapMap samples estimated that up to 12% of the human genome is subjected to CNVs (Redon *et al.*, 2006).

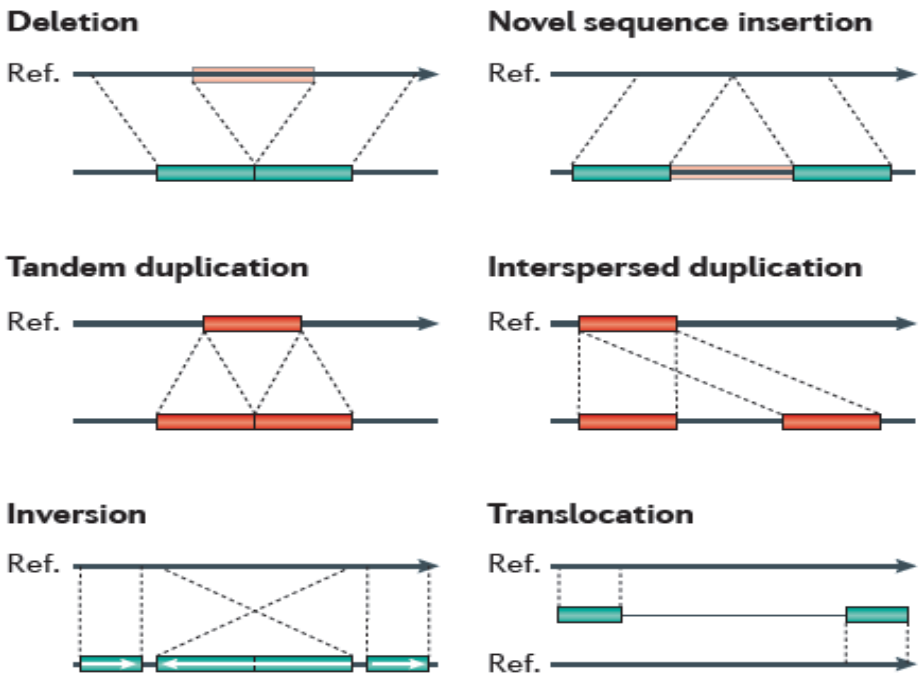


Figure 1.8 Different classes of structural variations. The schematic depicts deletions, novel sequence insertions, tandem and interspersed segmental duplications, inversions and translocations in a test genome (lower line) when compared with the reference genome. Adopted from Alkan *et al.* 2011.

These studies implicated that CNVs are at least as important as SNPs in contributing to inter-individual genome differences and they are actually a major driving force in evolution. Indeed between any two individuals, the number of base pair differences derived by CNVs is >100-fold higher, compared with that by SNPs (Lupski, 2007).

Routinely for nearly all genes in the human genome, each individual inherits one copy from each parent, so that two copies of each gene exist in the nucleus of every diploid cell. However, various studies have indicated that the copy number per genome varies for some genes. The set of the genes that were found to vary in copy number is enriched for those involved in olfaction, immunity and secretion, i.e. genes relevant to the immediate environmental responses (Cooper *et al*, 2007; Conrad *et al*, 2010). For example, a cluster of several β-defensin genes shows common CNVs of between two and seven copies per diploid genome, and occasionally copy numbers are as high as 10 or 11(Hollox *et al*, 2003). Similarly, the salivary amylase gene, *AMY1*, varies in copy number from two to 15 with a mean of seven in the European–American population (Perry *et al*, 2007). The amount of salivary amylase is directly correlated with the copy number of AMY1. The average number of copies of AMY1 is higher in cultures that consume a high level of starch than in cultures that consume little starch (Perry *et al*, 2007), suggesting that a high copy number of AMY1 is advantageous in cultures with a high starch intake and neutral in cultures with low starch intake.

Those CNVs that are present at a greater than 5% frequency in the population and mostly occur in multiple copy number states (0 to 30 copies per diploid genome) are called copy number polymorphisms (CNPs). In contrast, large CNVs (>100 kb) are mostly individually rare (<1% frequency) and exist in fewer copy number states (single copy gain / loss). These rare CNVs are under strong selection pressure, so that their frequency in the population is largely contributed to by *de novo* events and they persist only for a few generations (Turner *et al*, 2008).

1.2.1 Discovery and mapping of CNVs

The systematic discovery and genotyping of CNVs at the genome-wide level has become possible due to advances in whole genome technologies, which have enabled characterization of CNVs that are intermediate between large chromosomal aberrations (>1 Mb) and smaller indels (1–50 bp). The two main approaches to systematically characterize CNVs are microarray-based and sequencing-based methodologies.

Microarray-based approaches

These include two main platforms, namely array comparative genomic hybridization (array CGH) and single nucleotide polymorphism (SNP) microarrays.

Array CGH platforms are based on the principle of comparative hybridization of two labeled samples (test and reference) of genomic DNA to a set of hybridization targets, typically long oligonucleotides or primarily large-insert clones known as bacterial artificial chromosomes (BACs) (Pinkel *et al.*, 1998). The signal ratio between test and sample is normalized and converted to a $\log_2$ ratio, which acts as a proxy for copy number. An increased $\log_2$ ratio represents a gain in copy number in the test compared with the reference; conversely, a decrease indicates a loss in copy number. Early large scale surveys of CNVs have mainly used array CGH. Initially in 2004, two studies reported that CNVs of many large DNA genomic segments exist between normal human individuals. Sebat and colleagues employed an array CGH analysis through 85,000 interrogating probes with an average spacing of 35 kb and identified large-scale (>100-kb) copy number differences between 20 normal individuals (Sebat *et al.* 2004). In total, 221 copy number changes at 76 CNV loci were detected by them. Similarly Using a BAC-CGH array with resolution of ~1 Mb, Iafrate and colleagues investigated large-scale CNVs in 55 unrelated individuals and identified 255 clones with copy number gain or loss (Iafrate *et al.* 2004)

SNP microarray platforms are also based on hybridization, in which signal intensities for the sample being analyzed are compared to a collection of reference hybridizations or the rest of the population being analyzed. However two key differences distinguish SNP microarrays from CGH technologies. First, hybridization in SNP platforms is performed on a single sample per microarray, and log-transformed ratios are generated by clustering the intensities measured at each probe across many samples (McCarroll *et al.*, 2008 Perry *et al.*, 2008). Second, SNP platforms take advantage of probe designs that are specific to single-nucleotide differences between DNA sequences, either by single-base-extension methods (Illumina) or differential hybridization (Affymetrix) (Cooper *et al.*, 2008 Peiffer *et al.*2006). Early SNP arrays demonstrated poor coverage of CNV regions but recent arrays (such as the Affymetrix 6.0 SNP and Illumina 1M platforms) incorporate better SNP selection criteria for complex regions of the genome and non-polymorphic copy-number probes (Winchester *et al.*, 2009).

In 2006 Redon *et al.* constructed the first-generation CNV map of the human genome through the study of 270 HapMap individuals from four populations with ancestries in Europe, Africa or Asia. They employed both SNP genotyping arrays (Affymetrix 500K, 474642 SNPs) and Whole Genome BAC arrays (26574 large insert BAC clones) and found a total of 1,447 CNV regions, covering 360 megabases and consisting 12% of the human genome (Redon *et al.*, 2006).

The Database of Genomic Variants (DGV) (www.projects.tcag.ca/variation), primarily developed by Lafrate and colleagues, is the most up-to-date catalogue for the results of comprehensive genome-wide CNV screenings in peer-reviewed studies (Iafrate *et al.*, 2004, MacDonald *et al.*, 2014). In this database duplications, insertions and deletions of more than 1 kb of genomic DNA sequences are integrated as CNVs and those of 100–1000 bp as small indels. However, further high resolution CNV surveys showed that CNV sizes were often overestimated in initial studies, mainly due to the low resolution of the primary platforms (e.g., BAC arrays) used in CNV screening. With the aid of higher resolution (1 kb) array

CGH, Perry *et al* studied 2191 known CNV regions in 30 individuals from 4 HapMap populations. They detected copy number changes in 1153 loci and narrowed the boundaries of 1020 (88%) CNV regions (Perry *et al*, 2008). Reduced CNV sizes were also reported in another survey of HapMap samples via Affymetrix 6.0 array, as a hybrid SNP-CNV genotyping platform (McCarroll *et al*, 2008). Upon these studies, it was argued that the large-scale CNVs may affect the genome less extensive than initially proposed, encompassing about 5% of the total human genome (McCarroll 2008).

More recent technologies of whole-genome array CGH platforms routinely produce arrays with up to 2.1 million (2.1M) (Roche NimbleGen) and 1M oligonucleotides (Agilent Technologies) per microarray. Detection of a CNV typically requires a signal from at least 3 to 10 consecutive probes.

Sequencing-based approaches

In addition to the array-based platforms, CNVs can also be investigated by DNA sequencing at whole genome levels, either by **Sanger-based sequencing** approaches (Sanger *et al*, 1977) or through **next generation sequencing (NGS)** platforms (McKernan *et al*, 2009; Metzker, 2010). One of the first efforts of using Sanger-based technology for CNV characterization was done in 2006, in which split capillary reads of DNA resequencing traces, generated by shotgun sequencing, were used for identification of insertion and deletion polymorphisms from the genomic DNA of 36 individuals (Mills *et al*, 2006). This survey discovered 415,436 indels and CNVs ranging from 1 to 9989 bp in size.

Balanced SVs such as inversions cannot be identified by array platforms but are detectable through NGS- based **paired-end sequencing** technique, in which both ends of the test DNA segment is sequenced and is compared to the reference genome (Korbel *et al*, 2007). Eichler and colleagues compared fosmid (a phage cloning vector with DNA packaging limited to ~40 kb) DNA sequences from a library constructed from the genomic DNA of the HapMap

individual NA15510 and identified 297 potential SVs, varying in size from 8 kb to 1.9 Mb (Tuzun *et al*, 2005). In a related study, they constructed new fosmid libraries from eight HapMap samples (four Yoruba Africans and four non-African individuals) and sequenced both ends of approximately one million clones per genome (Kid *et al*, 2008). Combined with the previous analysis of the NA15510 fosmid library, they validated 1695 SVs across 9 diploid human genomes, including 747 deletions, 724 insertions, and 224 inversions. 50% of these were found in multiple libraries.

NGS-based platforms typically generate shorter reads than Sanger sequencing-based methods and they can sequence billions of bases in parallel. The first study of structural variation discovery using NGS technology was done by Korbel and colleagues that sequenced paired ends of 3-kb DNA fragments and mapped DNA reads onto the reference genome (Korbel *et al*, 2007). By this strategy they identified deletions, inversions and insertions of ~3 kb or larger. In total, 1297 SVs, including 853 deletions, 322 insertions and 122 inversions, were identified in two female individuals, one African (NA18505) and one European (NA15510).

So far the most high throughput and large scale use of NGS technologies towards a comprehensive map of human genomic variation has been done in the 1000 Genome Project (www.1000genomes.org), which has provided an integrated catalogue of variations from 1092 individuals (1000 Genomes Project Consortium *et al*, 2010 and 2012). Different NGS-based discovery approaches and algorithms for CNV mapping, including read-pair, read-depth and split-read methods, used in the 1000 genome project are explained elsewhere (Medvedev *et al*, 2009 Mills *et al*, 2011) and are beyond the scope of this thesis to be described here. The most important point to be mentioned is that different platforms and technologies have distinct power profiles in CNV genotyping, with regard to size, distribution and type of CNVs and therefore no single discovery strategy can capture the entire spectrum of structural variations in the genome. Table 1.2 shows the numbers of events detected for various categories of genomic structural variations including deletions, novel insertions, inversions and

duplications, as reported in the database of dbVar (Church *et al.*, 2010) for the associated publications. Compared with array-based platforms, NGS-based methods for CNV discovery are biased towards the detection of deletions (Pang *et al.*, 2014). The 1000 Genomes Project studies analyzed a large number of samples (1092) from a wider range of populations but used low-coverage sequencing, which instead limited CNV detection efficiency. Therefore the majority of the CNVs discovered by them were smaller than 400 bp, and duplications as well as larger variations were under-represented (1000 Genomes Project Consortium *et al.*, 2010 and 2012). However, whereas array-based approaches have a limited resolution capacity, sequencing-based approaches provide more accurate sequence-level breakpoint resolutions. The highest-resolution genome-wide array platform was used by Conrad and colleagues, in which the minimum size threshold for CNV detection was 450 bp (Conrad *et al.*, 2010).

Method	Samples	Deletions		Novel Insertions		Inversions		Duplication		References
		Calls	Median length	Calls	Median length	Calls	Median length	Calls	Median length	
SNP microarray*	270	1,122	6,216	_	_	_	_	442	14,122	McCarroll *et al.* ,2008
SNP microarray‡	2,493	9,963	50,265	_	_	_	_	3,880	108,336	Itsara *et al.*,2009
Fosmid ESP	8	1,843	8,657	560	7,594	1,146	77,119	1,768	8,429	Kid *et al.*, 2008
Array CGH§	40	7,909	2,284	_	_	_	_	4,740	5,265	Conrad *et al.*, 2010
Array CGH‖	30	14,597	2,439	_	_	_	_	5,502	3,835	Park *et al.*, 2010
NGS	185	22,025	742	128	98	_	_	501¶	138	1000 Genome proj Con.,2010

Table 1.2 Comparison of different methodologies in identifying CNVs of different types and sizes. Ascertainment depends largely on the platforms and algorithms used. The significant bias of array platforms to deletion events, as well as the use of fosmid and sequencing-based platforms in detecting inversions and novel insertions which are missed by array technologies is highlighted. *Affymetrix 6.0 SNP (CNP calls only). ‡Illumina 300K, 550K and 650K. §Custom 42M probe, NimbleGen (unique CNV loci). ‖Custom 24M probe, Agilent. ¶ Tandem duplications only. ESP: end-sequence pair (Adopted form Alkan *et al.*, 2011).

At present DGV has collected and curated 2,791,408 CNVs (comprising 202,431 CNVRs) that were discovered from 75 peer reviewed studies (www.projects.tcag.ca/variation). On the basis of the approach used for CNV discovery and characterization, these studies could be split into three main categories (i) studies which have used sequencing (NGS and Sanger approaches), (ii) surveys based on oligonucleotide array CGH or SNP arrays and (iii) studies using other methods, for example, fluorescence in situ hybridization (FISH), polymerase chain reaction (PCR), multiplex ligation-dependent probe amplification (MLPA) and optical mapping.

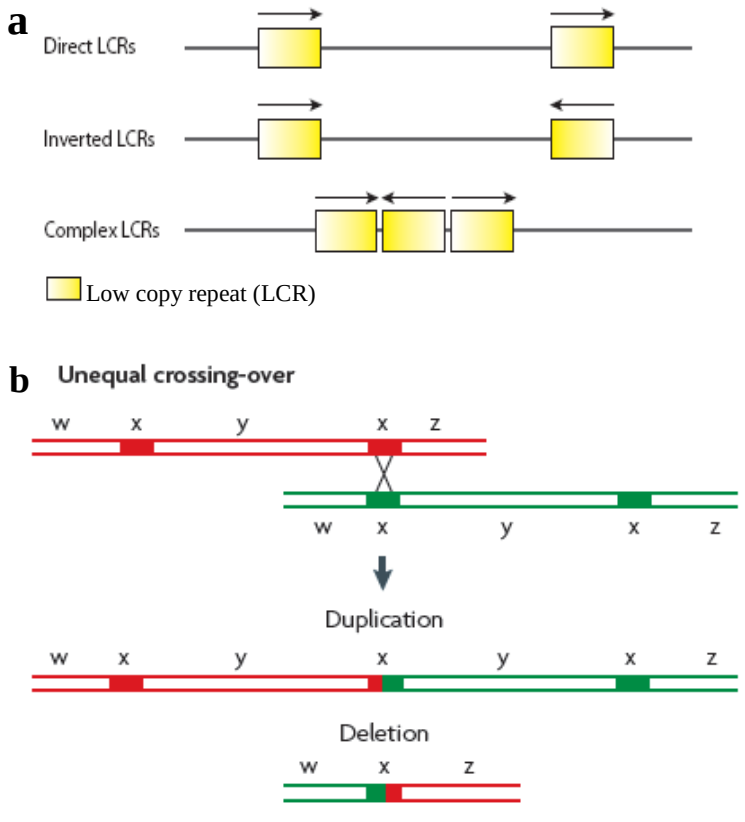
1.2.2 Mechanisms of evolving CNVs

Any change in copy number requires a change in chromosome structure, resulting in joining two formerly separated DNA sequences. Structural variants result from different mutational mechanisms, comprising DNA recombination-, replication- and repair-associated processes (Carvalho *et al.*, 2016). Four major mechanisms, namely, nonallelic homologous recombination; non-homologous end-joining; replication fork stalling and template switching; and L1-mediated retrotransposition generate rearrangements in the human genome and account for the majority of CNVs (reviewed in Gu *et al.*, 2008 and Hasting *et al.*, 2009, Carvalho *et al.*, 2016). These mechanisms are introduced and described here briefly.

Nonallelic Homologous Recombination (NAHR)

NAHR occurs by the alignment and subsequent crossover between two nonallelic (i.e., paralogous) DNA sequence repeats with high similarity to each other (Figure 1.9) (Stankiewicz *et al.*, 2002). *NAHR* has been argued to be the underlying mechanism for the majority of CNVs (Lee *et al.*, 2007; Hasting *et al.*, 2009). Repeats on the same chromosome and in direct orientation mediate duplication and/or deletion, whereas inverted repeats mediate inversion of the genomic interval flanked by the repeats. These recombinations require extensive DNA sequence homology (approximately 50 bp in E. coli (Lovett *et al.*, 2002) and at

least 300 bp in mammalian cells and humans (Reiter *et al*., 1998). Junctions of the recurrent CNVs have been often found to be located in low copy repeats (LCRs), which provide this extensive homology (Stankiewicz 2002). LCRs, also called segmental duplications (SD), are large blocks (>10 kb) of interspersed duplicated sequences with >95% sequence identity and constitute five to six percent of the human genome (Bailey *et al*. 2002; Bailey and Eichler 2006) (See figure 1.9. a). CNPs are not distributed uniformly in the genome, but are enriched four to ten folds in SD regions (Redon *et al*., 2006 Cooper *et al*., 2008. Conrad *et al*., 2009). Due to their high degree of sequence identity, non-allelic copies of LCRs can be aligned in meiosis or mitosis. This leads to subsequent crossover between them, which instead results in genomic



.

Figure 1.9 a) Low Copy Repeats (LCR). b) Non-allelic homologous recombination (*NAHR*) will occur by unequal crossing over if a recombination event uses a direct repeat (x) as homology. In this situation, a crossover outcome leads to products that are reciprocally duplicated and deleted for the sequence between the repeats (y). Adopted form Hastings *et al*. 2009

rearrangements in the form of CNVs in progeny cells (see figure 1.9. b). In addition to LCR, repetitive sequences such as the retrotransposable L1 elements (Han *et al.*, 2008) and Alu repeats (Sen *et al.*, 2006) can act as *NAHR* substrates if they are derived from similar families or have enough sequence identity to facilitate homologous recombination.

Non-Homologous End-Joining (NHEJ)

There are evidences that not all CNVs are associated with large block of repeats such as SDs. At least two main studies proposed that a significant fraction of CNVs are formed by *NHEJ*, which is associated with microhomology rather than with long stretches of sequence identity at CNV breakpoints (Korbel *et al.*, 2007; Perry *et al.*,2008). *NHEJ* is utilized by human cells to repair DNA double-strand breaks (DSBs). DSBs are usually caused by ionizing radiation or reactive oxygen species as pathologic lesions, but are also induced as mediators for some cellular physiological processes such as V(D)J recombination (Lieber *et al.*, 2003; Schwarz *et al.*, 2003). In *NHEJ* mechanism, however, double strand breakage of DNA is followed by end joining in the absence of extensive sequence homology and is associated with small insertions or loss at the junction sites (Lieber *et al.*, 2003) (see also figure 1.10). *NHEJ* was first proposed, when sequencing the breakpoints of non-recurrent deletions in introns 47 and 48 of the human dystrophin gene *(DMD)* in patients with muscular dystrophy (Nobile *et al.*, 2002; Toffolatti *et al.* 2002). These deletions were not flanked by LCRs and the junctions showed only microhomology. Paradoxically, it was later found that the breakpoints of *NHEJ*-mediated rearrangements sometimes overlap with repetitive elements such as Non-long terminal repeats (*LTRs*), long interspersed nuclear elements (*LINEs*), *Alu* repeats and mammalian interspersed repetitive (MIR) elements. These observations suggested that although CNVs, arising due to *NHEJ* mechanism, do not obligatory require LCRs but are still stimulated by certain genomic homology architectures (Stankiewicz *et al.*, 2003).

Replication Fork Stalling and Template Switching (FoSTeS)

As it was difficult to explain the complexity of some CNV rearrangements by either *NAHR* or *NHEJ* recombination mechanisms, Lee and Lupski proposed the *FoSTeS* Model (Lee *et al*, 2007) (see figure 1.10). According to this model, during DNA replication, the DNA replication fork stalls at one position, so that the lagging strand (i) disengages from the original template; (ii) transfers and (iii) then anneals, by virtue of microhomology at the 3' end, to another replication fork in physical proximity (not necessarily adjacent in primary sequence); (iv) 'primes' and restarts the DNA synthesis. Switching to another fork located downstream (forward invasion) would result in a deletion, whereas switching to a fork located upstream (backward invasion) results in duplication. The most distinguished hallmark of this model for CNV formation is that *FoSTeS* is a replication-based mechanism and rearrangement is induced by errors in the replication procedure. *FoSTeS* model has been argued to be the underlying mechanism for gene duplication/triplication and even rearrangements of single exons (Zhang *et al*, 2009). These notions propose *FoSTeS* to be a significant mediator in gene duplication and exon shuffling; two predominant processes in driving new genes and genome evolution.

L1 Retrotransposition

Long interspersed elements-1 (L1) comprise 16.89% of genomic DNA sequence and currently are the only active autonomous transposons in the human genome (Kazazian *et al*, 1998; Goodier *et al*, 2008). Of the 516,000 copies of L1 in our genome, only about 80–100 copies are full length (about 6 kb in size) and have two intact open reading frames (ORF): ORF1 coding for a RNA-binding protein and ORF2 encoding a protein with both endonuclease and reverse transcriptase activity (Babushok *et al*, 2007). L1 transposition occurs via an RNA intermediate that is probably transcribed by RNA polymerase II (Babushok *et al*, 2007). The reverse transcription and integration are thought to occur in a coupled process called target primed

reverse transcription (TPRT) (Ostertag *et al*, 2001). The resultant insertion is flanked by duplicated target sites (TSD) that is a characteristic of TPRT (Ostertag *et al*, 2001).

The predominant mechanisms mediating the majority of CNVs have been reported controversially in different studies. 30% of the SVs, identified through paired-end mapping in HapMap individual NA15510 were attributed to retrotransposition, of which 90% due to L1 elements (Korbel *et al*, 2007). Kidd *et al* however reported 15% of their detected SV events in 9 HapMap individuals to be due to retrotransposition (Kid *et al* 2008). Some initial whole genome CNV detection studies have estimated that only 9% (Korbel *et al*, 2007) or 14% (Perry

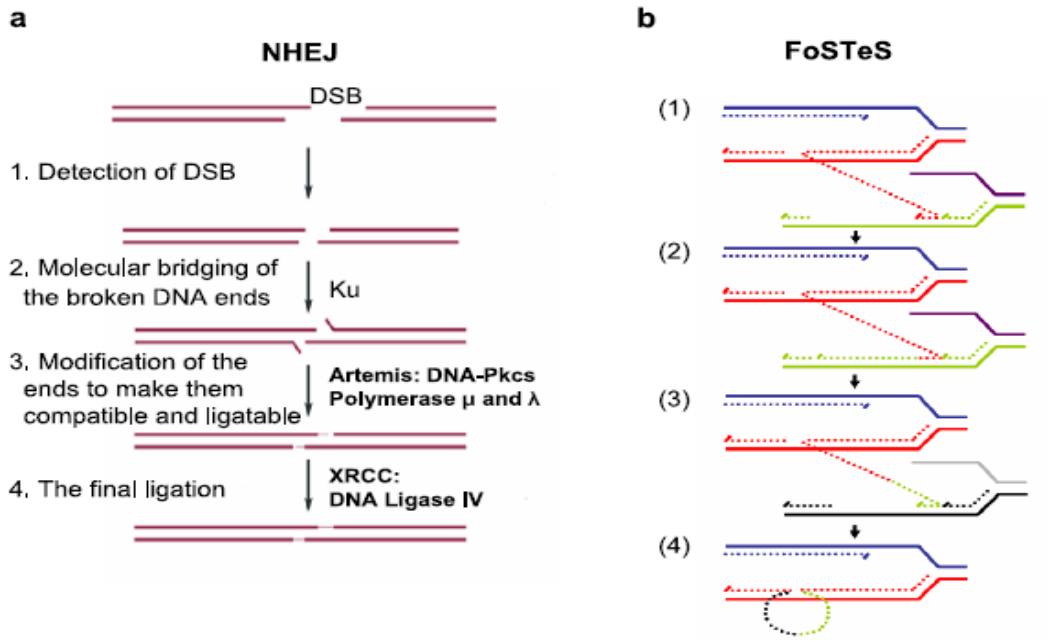


Figure 1.10 Genomic rearrangement mechanisms underlying CNV formation.

a) Non-homologous end-joining (*NHEJ*). Double-stranded DNA break (DSB) occurs and is repaired via *NHEJ* mechanism. The two thick lines depict two DNA strands with DSB, the thin segments in the middle represent the modifications which the ends have gone through before the final ligation. The enzyme machineries catalyzing each step are briefly mentioned. Note at step 3 that in order to repair ends, some addition or deletion of bases may be required, leaving behind a 'signature' of *NHEJ*. **b) *Replication Fork Stalling and Template Switching (FoSTeS)*** After the original stalling of the replication fork (dark blue and red, solid lines), the lagging strand (red, dotted line) disengages and anneals to a second fork (purple and green, solid lines) via microhomology (1), followed by (2) extension of the now 'primed' second fork and DNA synthesis (green, dotted line). After the fork disengages (3), the tethered original fork (dark blue and red, solid lines) with its lagging strand (red and green, dotted lines) could invade a third fork (gray and black, solid lines). Dotted lines represent newly synthesized DNA. Serial replication fork disengaging and lagging strand invasion could occur several times (e.g. *FoSTeS* x 2, *FoSTeS* x 3, ... etc.) before (4) resumption of replication on the original template. Illustration from Gu *et al.*, 2008

et al., 2008) of CNV breakpoints fall within repetitive sequences, suggesting that nonrecurrent mechanisms predominate, whereas others have argued that 47% of breakpoints follow *NAHR* rules (Kidd *et al.*, 2010). Rate of formation of new CNVs is mostly dependent on the underlying mechanism, which instead correlate with local and regional genome architecture. Estimations of locus-specific CNV mutation rates rang from 1.6×10^{-6} to 1.2×10^{-4} per locus per generation, which are 100 to 10,000 times greater than that for SNPs (van Ommen, 2005; Lupski, 2007). Of note, mutation rate of single base substitutions is estimated to be $1.8 – 2.5 \times 10^{-8}$ per base pair per generation (Nachman *et al.*, 2000; Kondrashov, 2003).

1.2.3 Functional consequences of CNVs

Predicting the phenotypic consequences of structural variations has been shown to be complex. In this context the location of CNVs in relation to genes is particularly important. CNVs may occur anywhere, but are more common in regions devoid of genes known as gene deserts (Buchanan *et al.*, 2008). Some CNVs encompass entire genes, the expression of which is assumed to vary according to gene dosage effects. CNVs may also have their breakpoints intersected with coding regions, resulting in gene fusions or truncations. Yet others are in nongenic regions but can have measurable effects on expression of the genes that are hundreds of kilobases away from the rearranged site (see Figure 1.11). When genes are involved, impact of the variant will be dependent on the function(s) of these genes. Essential or housekeeping genes are less likely to be tolerant of any disruption, and *de novo* variants that affect them encounter strong selection (Goh *et al.*, 2007). Studies on human transformed cell lines have shown that globally there is a significant correlation between mRNA levels and gene copy number (Ait Yahya-Graison *et al.*, 2007; Schlattl *et al.*, 2011). Notably, however, for individual genes mRNA levels often deviated from the expected levels; that is, they were not halved when one gene copy was deleted, nor increased by a 3:2 ratio in a trisomic (one copy gain) state. Further to this, not all genes with altered copy number displayed altered

expression, and a small proportion even showed expression changes that were inverse to the copy-number alteration (Schlattl *et al.*, 2011; Vazquez-Mena *et al.*, 2012). It has been argued that part of these aberrant observations might be due to dosage compensation mechanisms, acting on larger regions of the genome (Straub *et al.*, 2007). In addition, structural variants that overlap

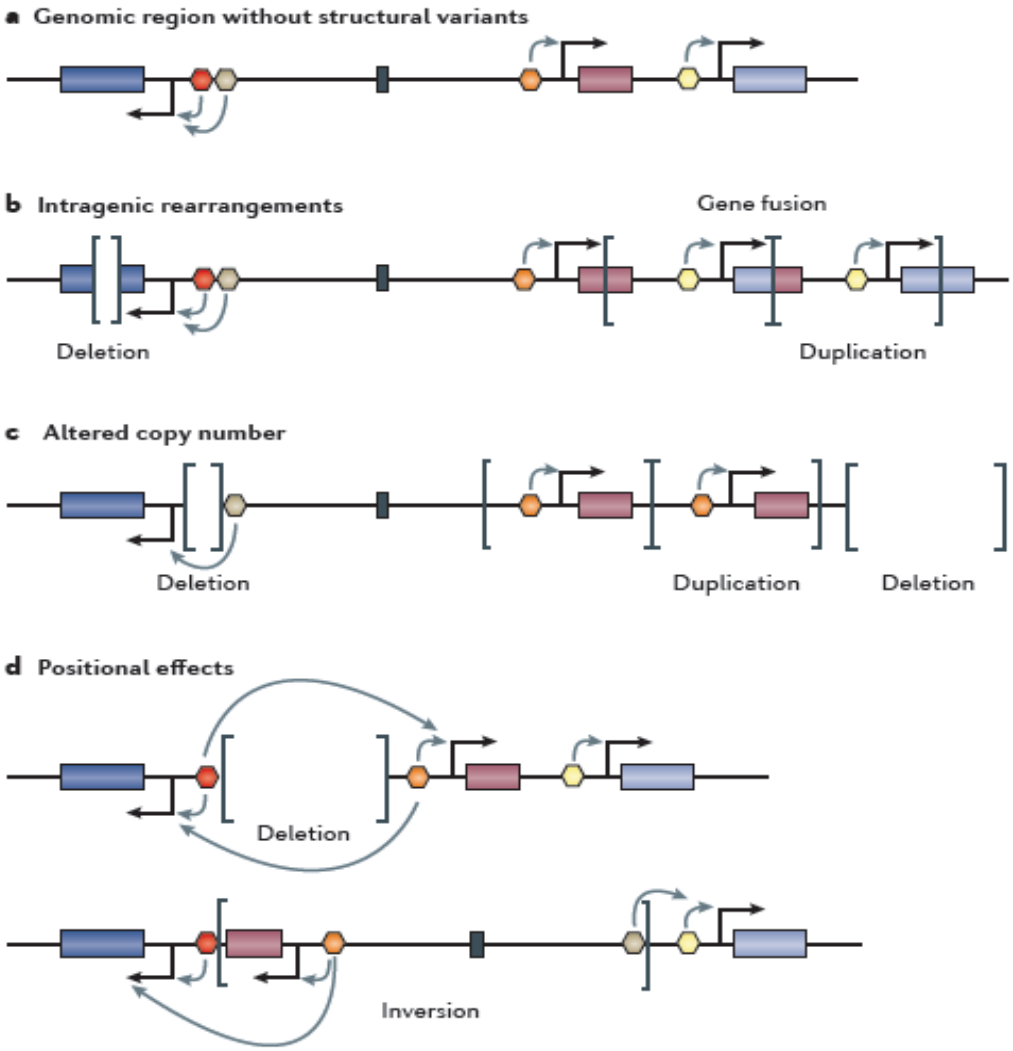


Figure 1.11 Functional consequences of structural variants. a) Genes (boxes) are regulated by the collective and combinatorial input of regulatory elements, including tissue-specific enhancers (hexagons, with different colours indicating tissue-specificity, and arrows pointing to the target gene) and insulators (black rectangles), which block the activity of regulatory elements. **b–d)** Structural variants (shown by square brackets) can have phenotypic consequences by altering coding regions. For example, they can remove part of a coding region or fuse different coding regions after a duplication, resulting in aberrant transcripts **(b)**. Alternatively, deletions or duplications can lead to altered doses of otherwise functionally intact elements **(c)**, resulting in altered regulatory input (left) or altered gene copy number (right). Structural variants can also affect the expression of genes outside of the variants, which is a positional effect **(d)**, thus resulting in a gain or loss of regulatory inputs. (Illustration from Weischenfeldt *et al.*, 2013)

cis-regulatory elements may be inversely correlated with the expected directional change of mRNA abundance. Cis-regulatory elements are genomic regions (e.g. enhancers) that regulate the expression of genes on the same chromosome. It has been shown that deletions can lead to increased gene expression, where the deletion affects a silencer or insulator element (Merla *et al*, 2006).

An **expression quantitative trait locus** (eQTL; a genomic locus that regulates the mRNA expression level of a gene) survey, carried out in inbred mouse strains found that structural variants contribute on average to larger effect sizes on gene expression than single nucleotide variants (SNVs) (Keane *et al*, 2011). The transcriptional effect of CNVs might be distinct in different cell types and tissues. For example a 20 kb deletion, locating immediately upstream the *IRGM* gene, is associated with a reduced expression of the gene in Hela (immortalized cervical cancer) and hepatocellular carcinoma cells, but leads to increased expression in colon carcinoma cells as well as bronchus smooth muscle cells. (McCarroll *et al*, 2008). This differentiated effect could be due to the fact that eQTLs are cell-type-, tissue- and developmental-stage-specific, so that gene expression programs could operate in a highly context-dependent manner (Chaignat, *et al*, 2011)

1.2.4 CNVs and diseases

So far based on the reports of involvement of CNVs in various disease phenotypes, two general types or models of CNV-disease associations have emerged. The first type involves CNVs that are individually rare (<1% frequency), typically involve larger chromosomal segments (>100 kb) and exist in fewer copy number states (single copy gain or loss). These CNVs are under strong selection pressure; their frequency in the population is largely contributed by *de novo* events and they can only persist for a few generations (Girirajan *et al*, 2010). The second model involves CNPs that occur at a population frequency of >1% and

often exist in multicopy number states, ranging from 0 to 30 copies per diploid genome (Sudmant *et al.*, 2010).

Large rare CNVs in neuro-developmental and neuropsychiatric diseases

Extreme representations of deleterious pathogenic CNVs comprise whole or partial chromosomal duplications and deletions as well as sub-chromosomal structural changes of megabases of DNA, detectable by conventional microscopy such as molecular karyotyping or fluorescent in situ hybridization (FISH). A well-known instance in this context is Down syndrome (MIM 190685) caused by duplication of the whole human chromosome 21 (Lejeune *et al.*,1959; Antonarakis *et al.*, 2004). Duplication (trisomy) or deletion (monosomy) of a whole chromosome is called chromosomal aneuploidy and is mostly lethal (Hassold *et al.*, 2001). Trisomies of chromosomes 21, 18, 13 and X are viable (although harboring major congenital abnormalities), whereas except for monosomy X, all chromosomal monosomies are embryonically lethal (O'Connor, 2008). Aside from these extreme chromosomal imbalances, a lot of microdeletions and microduplicatios (megabase-sized variations yet submicroscopic) have been recognized, which lead to sporadic or Mendelian neurodevelopmental syndromes, generally known as genomic disorders (Lupski, 1998; Stankiewicz *et al.*, 2002; Sharp *et al.*, 2006). These large scale CNVs are frequently flanked by large segmental duplications that make these genomic regions prone to recurrent DNA rearrangements (Itsara *et al.*, 2009). This instead leads these CNV events to occur recurrently in multiple unrelated individuals in a relatively short period of time, although they are under strong negative selection in the population (Lupski *et al.*, 2005; Itsara *et al.*, 2010). Some prototypic examples of genomic CNV disorders are Smith–Magenis syndrome (SMS), Williams–Beuren syndrome (WBS) and Potocki–Lupski syndrome (PLS). Clinical phenotypes associated with submicroscopic chromosomal imbalances (including deletions, duplications, insertions, translocations, and inversions) have been archived in DECIPHER (DatabasE of Chromosomal Imbalance and Phenotype in Humans using Ensembl Resources;

http://www.sanger.ac.uk/PostGenomics/decipher/) (Firth *et al*, 2009). Only 20% of the disorders annotated in this database are caused by duplications, whereas 80% are caused by deletions, and duplications in the database are on average associated with less-severe phenotypes. These observations reconfirm the fact that loss of genomic regions is less tolerated than the gain (Firth *et al*, 2009). Large CNV rearrangements are not only associated with sever in-born or early-onset intellectual disabilities and neurodevelopmental malformations with patterns of sporadic or Mendelian inheritance; they have also been implicated in more common complex neuropsychiatric diseases such as autism (MIM 209850) (Sebat *et al*, 2007; Weiss *et al*, 2008), schizophrenia (MIM 181500) (Walsh *et al*, 2008; Xu *et al*, 2008), bipolar disorder and attention deficit hyperactivity disorder (ADHD) (Girirajan *et al*, 2011).

Some studies have reported recurrent large pathogenic CNVs with variable penetrance and expressivity. For example, the 1.5 Mb deletion at 15q13.3 was initially identified in individuals with developmental delay (Sharp *et al*, 2008). Further studies showed that the same deletion was enriched in cases with autism (Pagnamenta *et al*, 2009) and schizophrenia (Stefansson *et al*, 2008) and, in fact accounts for approximately one percent of cases with idiopathic generalized epilepsy (Helbig *et al*, 2009). Similarly, the ~1.6 Mbp microdeletion on 1q21.1 was found to be enriched in individuals with developmental delay (Mefford *et al*, 2008), autism (Szatmari *et al*, 2007) Schizophrenia (Stefansson *et al*, 2008), and cardiac defects (Greenway *et al*, 2009).

CNPs in common complex diseases

CNPs are CNVs that have high frequencies (>=5%) in human population (see also 1.2 page 18). Two types of CNPs may be distinguished; those that simply represent a gain or loss of a particular segment of DNA, referred to as bi-allelic CNPs, and those where the underlying sequence can exist as a series of whole integers, termed multicopy CNPs (Bailey *et al* 2002;

Conrad *et al*, 2006). CNPs have been associated with some complex human diseases (see table 1.3), nevertheless findings of their contribution to the genetic variance of most common diseases have been far less than that for SNPs (Wellcome Trust Case Control Consortium., 2010). In 2010, Conrad *et al* generated a map of over 10,000 common copy-number variants in the human population through high resolution array CGH. Most of common CNPs, characterized in their study were in strong LD with common SNPs (Conrad *et al*, 2010). They therefore argued that the associations of the Common CNVs with complex diseases have already been analyzed in GWAs studies for common SNPs. In contrast to this argument, it has been emphasized that most of multi-allelic CNPs occur in segmental duplication rich regions and therefore are difficult to be assayed either by array-based or NGS-based methodologies. These CNVs even tend to show less LD to SNPs (Alkan *et al*, 2011) and their disease associations have therefore not been dissected comprehensively.

CNPs in the human leukocyte antigen (HLA) region have been associated with multiple diseases, including Crohn's disease, rheumatoid arthritis, and type 1 diabetes (Wellcome Trust Case Control Consortium., 2010). Indeed SNPs in the HLA region have also been strongly associated with many immune-mediated diseases (Shiina *et al*, 2004) and considering the strong extended LD across this locus, it is not surprising that CNPs in this region contribute to risk of these diseases. It has been shown that large number of common CNVs associated with diseases lie in segmental duplications highlighting the role of these regions in genomic pathogenesis (Girirajan *et al*, 2011). Examples of these include association of CNVs of the β-defensin locus on chromosome 8 with psoriasis (Hollox *et al*, 2008) and Crohn's disease (Fellermann *et al*, 2006) as well as association of low copy number of *FCGR3B* with lupus (Fanciulli *et al*, 2007). However follow-up analysis has failed to replicate the association of β-defensin copy number with Crohn's disease using different copy number genotyping methods (Aldhous *et al*, 2010). Therefore, although CNPs in segmental duplications are important for

human genetic diversity, accurate genotyping methods are required to test these variants for association to disease.

SNPs around *IRGM* (immunity-related GTPase family, M) have been associated with CD (Parkes *et al.*, 2007). McCarroll and colleagues showed that a previously known 20-kb deletion polymorphism upstream of *IRGM* is in perfect LD with a CD-associated SNP in this region (McCarroll *et al.*, 2008). The deletion haplotype of *IRGM* was shown to have a distinct expression pattern compared to the reference haplotype. It has been proposed that *IRGM*-upstream deletion may cause CD through the altered level of *IRGM* expression, which instead affects the efficacy of autophagy (McCarroll *et al.*, 2008).

Gene	Disease/trait	Variant type	Associated allele
DEFB4, DEFB103, DEFB104	Psoriasis	Amplification	High copy number
DEFB4, DEFB103, DEFB104	Crohn's disease	Amplification	Low copy number
CCL3L1	HIV/AIDS	Amplification	Low copy number
C4	Lupus	Amplification	Low copy number
FCGR3B	Glomerulonephritis in Lupus patients	Amplification	Low copy number
FCGR3B	Lupus	Amplification	Low copy number
IRGM	Crohn's disease	Upstream deletion	Deletion
CFHR1, CFHR3	Age-related macular degeneration	Deletion	No deletion
CYP2D6	Reduced drug metabolism	Deletion	Deletion
RHD	Rh-negative blood group	Deletion	Deletion
OPN1LW, OPN1MW	Color blindness	Deletion	Deletion
LPA	Coronary heart disease	Amplification	Low copy number
SMN2	Severity of spinal muscular atrophy	Amplification	Low copy number
AZFc region	Spermatogenetic failure	Deletion	Deletion
CYP21A2	Congenital adrenal hyperplasia	Amplification	Two copies per chromosome
UGT2B17	Osteoporosis	Deletion	No deletion
UGT2B17	Graft-versus-host disease	Deletion	Deletion
LCE3B, LCE3C	Psoriasis	Deletion	Deletion
NEGR1	Obesity	Upstream deletion	Deletion
NBPF23	Neuroblastoma	Deletion	Deletion
TSPAN8	Type 2 diabetes	Amplification	Low copy number
HLA	Crohn's disease, rheumatoid arthritis, type 1 diabetes	Multiple CNVs	Various
GPRC5B	Obesity	Upstream deletion	Deletion

Table 1.3 Selected copy number polymorphisms associated with complex diseases. (According to Girirajan *et al.*, 2011)

1.3 Monozygotic (MZ) twins

In the 19th century, the Scottish obstetrician J Matthews Duncan distinguished two types of twins as identical versus non-identical. However, the Britain Francis Galton is usually cited as the first scientist that proposed the possibility of using twins to compare the contributions of nature (heredity) versus nurture (environment) (Galton, 1875; Boomsma *et al*, 2002). MZ twins - also called identical twins - arise from a single ovum (egg), fertilized by one sperm (and therefore from a single zygote), whereas dizygotic (DZ) twins result from two different eggs, fertilized by two different sperms. Generally MZ twins have been assumed to be genetically identical and phenotypic differences between individuals of a MZ pair (discordance) have been classically attributed to non-genetic factors, mainly differential environmental exposures (Schinzel *et al*, 1979). On the other hand, DZ twins share on average only half of their genetic variations (like any non-twin siblings) and therefore phenotypic discordances between them are the result of both genetic and non-genetic factors (see also table 1.4).

The spontaneous rate of monozygotic twinning is about 4 in 1000 human live births around the world (Hall, 2003), although there exists evidence for a familial susceptibility with autosomal dominant inheritance in some cases (Machin, 2009). The rates of twinning have increased in the past decades, in part due to assisted reproductive technologies as well as increases in maternal age (Chang *et al*, 2009). The etiology of monozygotic twinning in humans is mostly unknown. Earlier assumptions considered this as a random process and proposed that stochastic damage to the inner cell mass of the blastomer (see figure 1.12) and subsequent generation of two points of regrowth leads to twinning in humans (Hall *et al*, 1996). More recently it has been proposed that cells within the blastocyst may develop in such a way as to become discordant, recognizing each other as foreign cells and use cell-recognition mechanisms to set up two separate cell masses (Zwijnenburg, 2010). Further it has been argued that post-zygotic epigenetic changes such as differentiated DNA methylation very early in embryonic development (within two weeks after fertilization) could establish two

separate cell masses, which eventually develop to MZ twins (Shur 2009; Machin, 2009). Such a model could be an explanation for increased prevalence of certain (epi-) genetic syndromes among MZ twins compared to singletons, such as Beckwith–Wiedemann syndrome (BWS) (Shur, 2009; Bliek *et al*, 2009). In this model, it is not however clear whether unequal splitting of the inner cell mass results in differential methylation between the two cell masses or alternatively a lack of methylation maintenance in the early zygote results in splitting of the zygote. In both cases, however, a difference in genomic methylation between MZ twins may also account for their phenotypic discordance (Zwijnenburg, 2010).

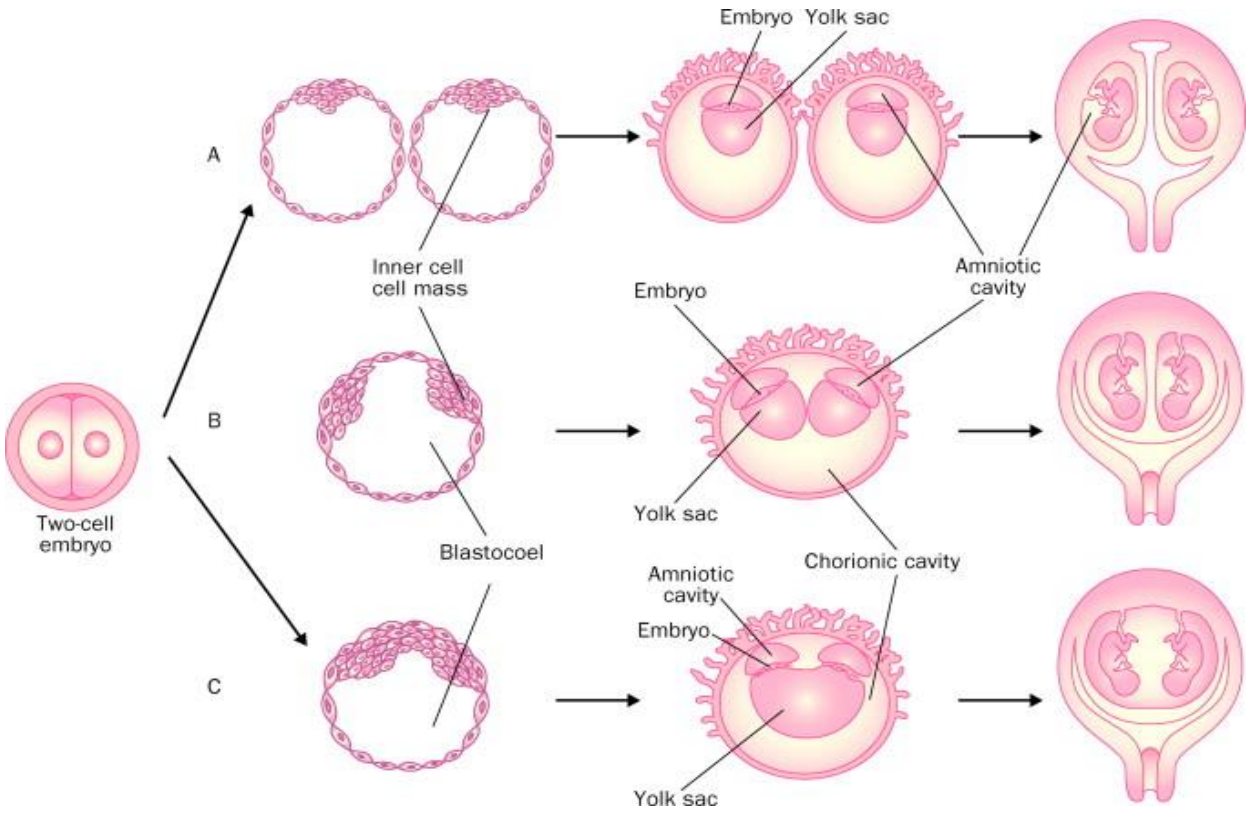


Figure 1.12 Three types of monozygotic placenta and membranes

A) If the zygote splits within 3 days, the twins are dichorionic and diamniotic (18 to 36% of all MZ births). **B)** If splitting occurs after the third but before the seventh day, the twins are monochorionic but diamniotic (60 to 80% of cases). **C)** If division occurs between days 7 and 14, the twins are monochorionic and monoamniotic (2 to 4% of all MZ twins). Conjoined twins arise when splitting happens after the days 13 or 14. A=dichorionic diamniotic pregnancy. B=monochorionic pregnancy. C=monochorionic monoamniotic pregnancy. (Adopted from Hall, 2003)

In addition to post-zygotic genetic and epigenetic events, prenatal intra-uterine exposures could also contribute to phenotypic divergence of MZ twins, despite deriving from a single zygote. The majority of MZ twins (about two-thirds) are monochorionic (MC), that is, twins who are both connected to a truly single (not fused) placenta (Machin, 2009). The remaining one-third of MZ twins are dichorionic (DC), developing in separate sacs (see figure 1.12). Parents of same-sexed DC MZ twins are sometimes informed mistakenly that their twins are dizygotic. In the past, zygosity in same-sex twin pairs was mainly determined by blood group and HLA typing and comparison of physical characteristics (Keith, *et al*, 1997). However, DNA polymorphisms such as microsatellite markers and SNPs are more reliable and used more routinely in zygosity testing (Hall 2003). The most common approach has been analyzing eight variable microsatellite loci by PCR of the genomic DNA from buccal samples. In large-scale epidemiological projects, zygosity is still often determined on a series of questionnaire items (Rietveld *et al*, 2000).

1.3.1 Discordant MZ twins in disease studies

In classical twin studies, comparing the resemblance (concordance) rate of a complex trait or disorder in MZ twins with that in DZ ones is used to estimate the heritability (extent to which genetic variations determines phenotypic variation) of that trait (Boomsma *et al*, 2002). It is expected that any heritable disease will show more concordant rates in MZ twins that in non-identical (DZ) ones. MZ twins are in general expected to be similar; nonetheless an increasing number of reports have been published on MZ twins discordant for congenital malformations, chromosomal abnormalities and Mendelian and complex disorders (Machin 2009; Zwijnenburg *et al*, 2010; Czyz *et al*, 2012). Primary assumptions proposed that discordances in Z twins are generally due to differentiated environmental exposures; although responsible factors have not been well delineated. To date there exist evidences that beside environmental and stochastic effects, genetic and epigenetic differences could also account for discordance in

MZ twins (Czyz *et al*, 2012, Castillo-Fernandez., 2014). It is however appreciated that epigenetic modifications, especially those that occur postnatal in twins are driven by environmental exposures.

1.3.2 Genetic differences in MZ twins: Somatic mosaicism

Genetic differences might arise during prenatal cell divisions and differentiations (during embryonic development) or post-natally during the life time of MZ twins and could potentially contribute to their phenotypic discordance. The presence of cells within an organism that have differences in genetic composition, despite deriving from a single zygote is described as somatic mosaicism, which plays a key role in carcinogenesis, ageing and possibly autoimmunity (Youssoufian *et al*, 2002). Any variations in genomic sequences between MZ twins can also be viewed as extreme examples of somatic mosaicism. Mosaicism for *de novo* genome alterations including point mutations, indels, CNVs and chromosomal

Environmental effects

Phenotype	**Concordance**	Discordance
MZ twin pair	identical genome; similar pre- and post-natal "environment"	different post-natal "environment"
DZ twin pair	similar pre- and post-natal "environment"	different genomes; different post-natal "environment"

Genetic/epigenetic effects

Phenotype	**Concordance**	**Discordance**
MZ twin pair	common (initial) genome	post-zygotic divergence of genome/epigenome; prenatal environment of MC placenta
DZ twin pair	concordant for approximately 50% of alleles	discordant for approximately 50% of alleles

Table 1.4 Original twin model proposed that discordance especially in MZ is due to post-natal environmental factors (upper). More recent data proposes that genetic and epigenetic factors largely account for discordance in both MZ and DZ twin pairs (lower). MC: monochorionic (see also figure 1.12). According to Zwijnenburg *et al.*, 2010

rearrangements have been observed in MZ twins (Bruder *et al*, 2008; Machin 2009; Zwijnenburg *et al*, 2010; Czyz *et al*, 2012). MZ twins discordant for chromosomal abnormality (aneuploidy) have been recognized for a longer time; discordant MZ twins have been reported for monosomy X, trisomy 1, trisomy 13, and trisomy 21 (Zwijnenburg *et al*, 2010). The rate of *de novo* single nucleotide variation (SNV) has been estimated at about 1.18 x 10^{-8} per base pair per generation, which account for ~74 novel SNVs per genome per generation (Veltman *et al*, 2012). Although this estimation primarily explains germline mutations, it has been assumed that most cells in the body carry at least one *de novo* point mutation (Veltman *et al*, 2012), an assumption that can rationally be applied to MZ twins as well. Accordingly, post-zygotic point mutations in MZ twins have been described to be the source of discordance for some disease phenotype. For example, in a pair of MZ twins discordant for Darier's disease

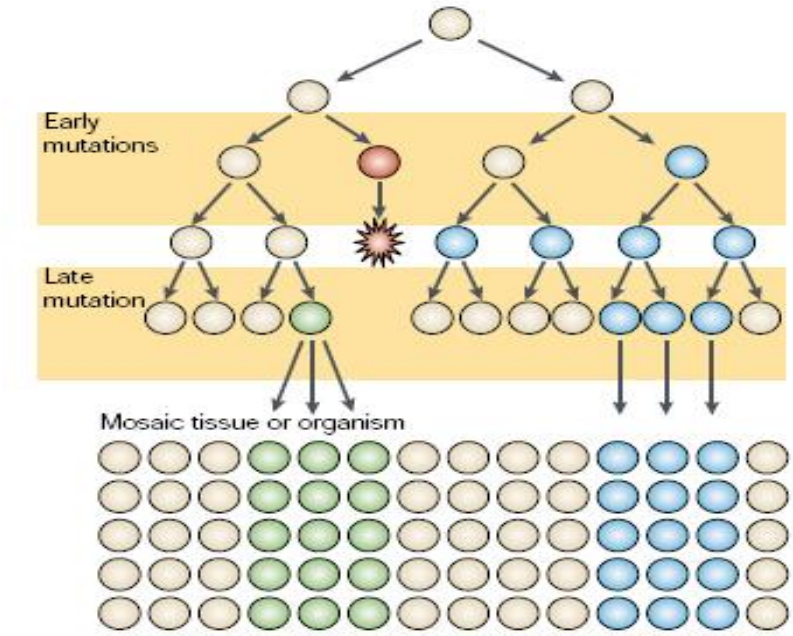


Figure 1.13 Somatic mosaicism. Mosaic populations arise if new mutations that occur early in development (blue circles) do not inhibit cell growth and division or, alternatively, new mutations that occur relatively late in development (green circles) confer a proliferative advantage. Mutations that compromise cell growth early on most likely will not contribute to the mosaic phenotype (red circle). Illustration from Youssoufian *et al.*, 2002.

(an autosomal-dominant skin disorder characterized by loss of adhesion between epidermal cells and abnormal keratinization), discrepancy was attributed to a point mutation in the ATP2A2 gene, which encodes the sarco/endoplasmic reticulum Ca2(+)-ATPase type 2 isoform and is highly expressed in keratinocytes (Sakuntabhai *et al.*, 1999) Furthermore, a mutation in IRF6 was found in the affected twin from a MZ twin pair discordant for Van der Woude syndrome, an autosomal dominant form of cleft lip and palate with lip pits (Kondo *et al.*, 2002). Likewise discordance for a point mutation in FLNA has explained the development of otopalatodigital syndrome in one sibling of a MZ twin pair (Robertson *et al.*, 2006) and mosaicism within a MZ twin pair for a mutation in the COL4A5 gene was suggested to cause discordance of Alport syndrome (Matsukura *et al.*, 2004). It has been hypothesized that the inheritance of a single recessive mutated allele of a gene coupled with a somatic mutation in the normal allele during critical periods of development in the affected twin may result in discordance for Schizophrenia in MZ twins (Singh *et al.*, 2009). These reports highlight the potential value of discordant MZ twins in identifying new genes (Mansilla *et al.*, 2005).

The presence of CNVs has been demonstrated both within concordant and discordant MZ twin pairs. Bruder *et al.* studied a total of 19 pairs of MZ twins, of which 9 were discordant for Parkinson disease or Lewy body dementia, whereas 10 were phenotypically unselected. Comparisons within the twin pairs revealed a considerable number of loci suggestive of putative CNVs in both groups (Bruder *et al.*,2008). It has been even estimated that *de novo* CNVs may occur at a rate of 10% per twinning event (Czyz *et al.*, 2012). *NAHR* as a mechanism underlying CNV formation (see 1.2.3) may also occur in mitosis and thereby result in mosaic populations of somatic cells carrying genomic rearrangements (Lupski, 2007).

1.3.3 MZ twins in IBD

Until 2012, six population-based studies in cohorts of both MZ and DZ twin pairs, in which at least one individual was affected by IBD have been reported (table 1.5). In 1988, the Swedish twin registry identified 34 MZ and 43 same-sex DZ twins with at least one diagnosis of CD or UC (Tysk C 1988). Among 18 MZ twins, each with a CD proband, eight were concordant for CD, whereas in 26 DZ twin pairs only one was CD-concordant. In contrast, the observed MZ and DZ concordance for UC was 6.3% and 0%, respectively. This significantly higher concordance rate for CD (44.4%) versus UC (6.3%) in MZ twins proposed that genetic risk factors play a greater role in CD, whereas UC is influenced more by environmental triggers. Later cohort studies, however, showed that the rate of concordance within MZ twins is more modest for CD and is somewhat greater for UC than that primarily observed in Swedish registry. For example results of a nationwide study of a total of 189 IBD twin pairs (68 MZ and 121 DZ pairs) in Germany revealed MZ-twin-concordance rate to be 35% for CD and 16% for UC (Spehlmann *et al*, 2008). Indeed German cohort reconfirmed the stronger genetic influence in CD than UC, but simultaneously estimated a greater genetic component for UC in comparison to Swedish study. The question, whether Swedish twin study actually overestimated heritability of CD and underestimated that of UC, or CD is really more genetic in the Scandinavian population, was addressed by a long term follow-up of the Swedish IBD cohort by Halfvarson and colleagues. They recognized that primary study had not been age-standardized and only included a relatively small window of time to ascertain IBD cases and therefore they included IBD cases missed during the original study period and extended the time of observation for an additional 15 years (Halfvarson *et al*, 2011). The MZ concordance rate for UC revealed to be higher (15%) in their updated study, which is also very consistent with MZ concordance rate for UC (16%) in German twin cohort (Spehlmann *et al*, 2008). These studies clearly implicated that twins provide a valuable resource to estimate the contribution of genetic versus environmental factors in susceptibility to common complex

disorders like IBD. There are however challenges such as correct zygosity determination, standardization for age and onset of disease as well as variability in disease manifestations that could potentially lead to biased estimations of heritability in twin cohort studies (Brant, 2011).

Year	Population	MZ Twins (No. Concordant: No. Discordant)	Pair Concordance	DZ Twins (No. Concordant: No. Discordant)	Pair Concordance
1988	Sweden Tysk *et al.*	CD 8 : 10	44.4%	CD 1 : 25	3.8%
		UC 1 : 15	6.3%	UC 0 : 17	0.0%
1996	British Thompson *et al.*	CD 5 : 20	20.0%	CD 3 : 43	6.5%
		UC 6 : 32	15.8%	UC 1 : 33	2.9%
2000	Denmark Orholm *et al.*	CD 5 : 4	55.5%	CD 1: 27	3.6%
		UC 3 : 18	14.3%	UC 2: 42	4.5%
2008	Germany Spehlmann *et al.*	CD 11 : 20	35.5%	CD 2 : 56	3.4%
		UC 6 : 31	16.2%	UC 1 : 62	1.6%
2010	Sweden [a] Halfvarson *et al.*	CD 9 : 24	27.3%	CD 1 : 49	2.0%
		UC 6 : 35	14.6%	UC 3: 46	6.1%
2010	Sweden [b] Halfvarson *et al.*	CD 4: 10	28.6%	CD 0 : 14	0.0%
		UC 1 : 5	16.7%	UC 1 : 15	6.3%
1988-2010	All combined non-overlapping cohorts	CD 34 : 78	30.3%	CD 7 : 189	3.6%
		UC 22 : 121	15.4%	UC 8 : 198	3.9%

Table 1.5 IBD Reported in Published Studies of Unselected Twin Cohorts (Adopted from Brant *et al.*, 2011)

Sweden[a] twins born 1886–1958; Sweden[b] twins born 1959–1980.

1.4 Aims of this study

As mentioned earlier, IBD risk loci identified so far confer only a modest effect on disease susceptibility and therefore the majority of the genetic contribution to disease risk remains to be explained. CNVs, as one major form of genomic variations have been shown to be involved in the pathogenesis of some common disease phenotypes. However the limited findings in case-control association studies for CNVs might suggest the need to pursue alternative approaches for dissecting genetic copy number variability and their probable involvement in susceptibility to common diseases. In this thesis we conducted two related studies to examine whether CNVs contribute to IBD risk;

In one, 6 IBD-discordant monozygotic twin pairs were compared genome-wide to explore somatic CNVs. This was conducted to ascertain if twin individuals harbor any genomic copy number differences in comparison to their co-twins and whether the probable CNV regions potentially implicate further susceptibility loci in IBD.

In parallel, we recruited an existing SNP-GWAS data set of UC as well as four other independent UC cohorts and performed a multi-step genome-wide case-control analysis to interrogate the presence of disease-relevant rare CNVs.

We hypothesized that the probable findings of these two studies could provide insights on pathogenic genomic CNVs, which are shared or complementary in germline and somatic tissues.

2 Methods

2.1 Twin sample recruitment

6 pairs of German MZ twins, where in each pair only one sibling was diagnosed as IBD patient, were recruited for this study. Of these 3 pairs were discordant for UC and 3 pairs discordant for CD. These twin individuals were part of a cohort consisting of a total of 189 twin pairs previously gathered by several calls for twin pairs using advertisements and the nationwide newsletter of the German Cohn's and Colitis Association (patient association with more than 20,000 members; DCCV e.V.) and the Competence Network Inflammatory

Bowel Disease (medical expert network; Kompetenznetz Darmerkrankungen e.V.) (Spehlmann *et al*, 2008). In these calls individuals had been asked to participate and complete a questionnaire if they had IBD and were born as one of a twin pair. Among others, questions were related to zygosity, medical history, social status, lifestyle (e.g., former and present smoking status), if they grew up together, and birth order. Diagnosis of CD and UC was confirmed in the previous epidemiological survey conducted on this twin cohort (Spehlmann *et al*, 2008). In that study diagnosis of IBD has been confirmed by review of the patients' original medical records including ileocolonoscopies. Diagnosis of CD had been confirmed when at least 2 of the criteria published by Landers and colleagues (Landers *et al*, 2002) were fulfilled. Diagnosis of UC was based on endoscopic appearance and continuity of inflammation, histology, and proven exclusive involvement of the colon. Patients had been also profiled with regard to age of onset of symptoms, age at time of diagnosis, location of disease (ileal or colonic by CD) or behavior of disease (e.g., mucosal-confined inflammatory, structuring or penetrating by CD), disease extent (in UC), current treatment, and treatment since diagnosis. The non-diseased member of the twins had been evaluated for signs and symptoms of IBD and records of previous ileocolonoscopies (Spehlmann *et al*, 2008).

Table 2.1 describes the characteristics of the 12 twin individuals used in this study.

In addition to twin individuals, the genomic DNA of the anonymous HapMap sample NA15510 and 3 healthy Lithuanian MZ twins were recruited as controls in the way that will be described later.

ID	Clinical status	Gender	Age (y)	Smoking Status	CAI	Duration of the disease	Location	Treatment	Surgery
Hu1	H	M	57	NS	na				
UC1	UC	M	57	NS	4	15	Pancolitis	5-ASA/Steroids	0
Hc2	H	F	36	NS	na				
CD2	CD	F	36	NS	na	5	Ileum	5-ASA	0
Hu3	H	M	42	S	na				
UC3	UC	M	42	NS	3	14	Left-sided	5-ASA/Steroids	1
Hc4	H	F	28	NS	na				
CD4	CD	F	28	NS	na	3	Ileum	na	0
Hu5	H	F	41	NS	na				
UC5	UC	F	41	NS	7	4	Pancolitis	5-ASA/Azathioprin	0
Hc6	H	M	33	S	na				
CD6	CD	M	33	S	na	4	Colon	na	0

Table 2.1. Clinical Data for 6 IBD-Discordant MZ Twin Pairs

UC, ulcerative colitis patient; CD, Crohn's disease patient; Hu Healthy co-twin of UC patient: Hc Healthy co-twin of CD patient; CAI, Colitis Activity Index; na, non applicable; NS, Never smoked; S, Smoker; ASA, mesalamine.

2.2 Twin sample preparation

After recruitment of twin individuals, genomic DNA was isolated from peripheral blood cells as well as biopsy specimens sampled from the sigmoid colon or rather ileum and were checked for quality on agarose gel.

2.2.1 DNA extraction from blood and biopsy samples

Genomic DNA (gDNA) was extracted from EDTA whole blood samples, using the Invisorb® Blood Giga Kit for DNA isolation. Lysis of erythrocytes, while leukocytes stayed intact, was performed by incubating 9 ml of blood for 10 min with 30 ml of cold buffer 1 at room temperature. Afterwards, the suspension was centrifuged for 3 min at 3,000 rpm and the supernatant was carefully discarded. This step was repeated with 20 ml buffer 1 until the leucocyte containing pellet was free of haem. The pellet was resuspended in 3 ml of buffer 2 and50 µl of Proteinase K and incubated for 2 hours in a 60°C water bath under continuous

shaking(95 turns/min) to increase the lysis efficiency. This step leads to the lysis of the leukocytesand their nuclei and therefore to a release of DNA into the suspension. To seperate the DNA from cell and protein fragments, 1.8 ml of buffer 3 were added. Vigorous mixing and a 5 min incubation on ice were subsequently carried out. Then, the mix was centrifuged for 15 min at 5,000 rpm. Afterwards, the cleared supernatant was transferred into a 15 ml centrifuge tube. For the precipitation of the DNA, 10 ml of 96% ethanol were added and the tube carefully inverted several times. If precipitation did not take place, the tube was incubated for 2 h at -20°C. The precipitated DNA was obtained by centrifugation for 3 min at 5,000 rpm and was then collected with a pipette tip and transferred into 2 ml reaction tubes containing 1 ml of 70% ethanol. The DNA pellet was rinsed by vortexing and subsequently centrifuged for 2 min at13,000 rpm. Finally, the ethanol was removed with a pipette and samples were dried for 10min at room temperature. All samples were quantified by measuring the concentration with PicoGreen®. The purified gDNA was resuspended in 500 µl of 1x TE buffer and stored at+4°C for short periods or at -20°C for long periods. Average yields were 200 ng/µl, which corresponds to an amount of 100 µg DNA.

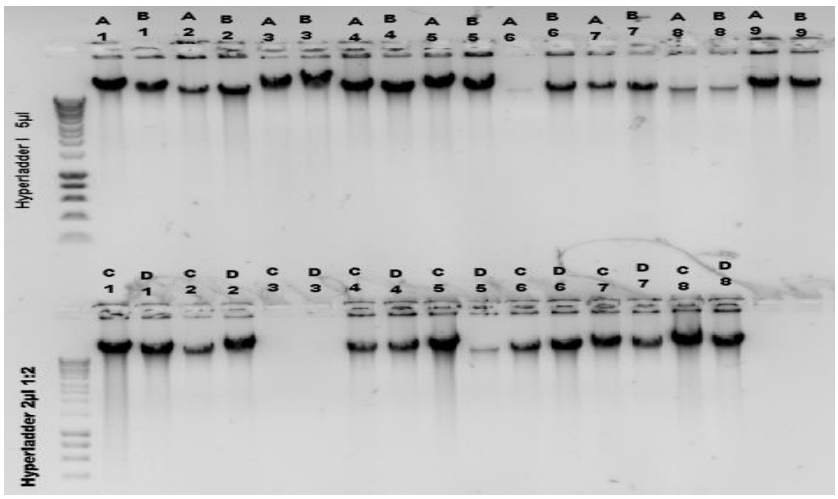


Figure 2.1 Preparation of Agarose Gel Electrophoresis for gDNA from twin individuals It was ensured that gDNA appeared as a single prominent band greater than 12 kb. If the sample appears as more than one band or as a smear, the DNA may be degraded or have a contaminant that could affect the labeling procedure. RNA contamination results in a smear that is less than 200 bp.

To determine the quality of the DNA samples, 250 ng of genomic DNA (gDNA) were run on a 1% agarose gel to ensure that they show no signs of RNA contamination or degradation (see figure 2.1).

All biopsies used in this study were primary tissues from the intestinal mucosa. Biopsies were taken endoscopically from a defined area of the colon, and immediately snap-frozen in liquid nitrogen. DNA was extracted from biopsies using the QIAamp Tissue DNA preparation kit (Qiagen, Hilden, Germany).

2.2.2 Twin Sample combinations

In each twin pair, 4 different combinations were designed and used for the following array experiments. The rationale behind these combinations will be discussed in results and discussion later. Here only the design of four combinations is mentioned as following;

1. Blood DNA from diseased twin individual as test versus blood DNA from healthy co-twin as reference.
2. Biopsy DNA from diseased twin individual as test versus blood DNA form healthy co-twin as reference.
3. Biopsy DNA from diseased twin individual as test versus Blood DNA form the same individual
4. Blood DNA from the HapMap sample individual as test versus blood DNA from healthy twin individual as reference.

Concentrations of the genomic DNA were measured using the PicoGreen® method (Rengarajan *et al*, 2002). The PicoGreen® reagent is a proprietary, unsymmetrical cyanine dye. Free dye is essentially nonfluorescent and exhibits >1000-fold fluorescence enhancement upon binding to double stranded DNA (dsDNA) with excitation and emission maxima of ~500 nm and ~520 nm, respectively. The assay displays a linear correlation between dsDNA concentration and fluorescence and has a detection range extending from 25 pg/mL to 1 μg/mL dsDNA using a

single dye concentration. The assay is highly selective for dsDNA over RNA, single-stranded DNA (ssDNA) and oligonucleotides.

	1	2	3	4	5	6	7	8	9	10	11	12
A	ZS 015 HapMap	ZS 001 HapMap	ZS 007 HapMap	LT008 HapMap	ZS 013 HapMap	LT013 HapMap	LT024 HapMap	ZS 020 HapMap	ZS 024 HapMap			
B												
C	ZS 015	ZS 001	ZS 007	LT008	ZS 013	LT013	LT024	ZS 020	ZS 024			
D	R68	R374	R74	MIIN	R96	FEAU	AZVL	R421	R81			
E	ZS 015	ZS 001	ZS 007	LT008	ZS 013	LT013	LT024	ZS 020	ZS 024			
F	ZS 019	ZS 009	ZS 025	LT007	ZS 004	LT014	LT023	ZS 012	ZS 017			
G	ZS 015	ZS 001	ZS 007	LT008	ZS 013	LT013	LT024	ZS 020	ZS 024			
H	R70	R72	R300	MIZI	R307	KYRA	AZSE	R426	R79			

	1	2	3	4	5	6	7	8	9	10	11	12
A	211.4	250	250	250	250	250	94.1	200.5	221.6			
B	250	250	250	250	250	250	250	250	250			
C	211.4	250	250	250	250	250	94.1	200.5	221.6			
D	74.2	122.4	52.4	68.2	74.7	90.6	102	59.5	67.9			
E	211.4	250	250	250	250	250	94.1	200.5	221.6			
F	185.4	250	60.8	154.5	250	197.7	53.6	250	250			
G	211.4	250	250	250	250	250	94.1	200.5	221.6			
H	57.2	47.1	73.1	63.7	51.4	117.9	126	87.5	222			

Figure 2.2 **Sample combinations with ids on 96-well plate design (upper) and corresponding concentrations (ng/ul) of the samples used as test and reference samples in the array-CGH experiments (lower)**

2.3 Array-CGH experiments

Human CGH 2.1 M Whole-Genome Tiling v2.0 array from Roche NimbleGen (www.nimblegen.com) was used for genome-wide discovery of CNVs in this study. This array platform spans the entire human genome with 2.1 million 60 nucleotide-long probes at a median distance of 1,169 bp, enabling detection of CNVs down to ~5 kb.

Catalog design name / number	Human CGH 2.1M WG Tiling v2.0D / B7074-00-01
Probe length	60mer
Median probe spacing	1,169bp
Total features	2.1 million
Feature size	13µm x 13µm
Array size	62mm x 14mm
Slide size	1" x 3" (25mm x 76mm) glass
Sequence source	UCSC Genome browser, NCBI

Table 2.2 **NimbleGen Human CGH 2.1M Whole-Genome Tiling v2.0D Design Specifications**

2.3.1 Array-CGH workflow

DNA labeling and hybridization

Differential labeling of gDNA samples from twin individuals followed by co-hybridization to arrays slides and scanning with a 5um scanner was performed at the Nimblegen center. Pairs of samples intended for hybridization to the same array should be labeled in parallel using Cy3-random and Cy5-random Nonamers. Roche NimbleGen applied CY3 fluorescent dye for test samples and CY5 fluorescent dye for reference samples and used a high concentration of exo$^-$ Klenow fragment, lacking both proofreading (3' -> 5') and nick translation (5' -> 3') nuclease activities, to incorporate dNTPs while replicating the DNA. As these steps have been provided us by Nimbelgen, I would pass on theme here. A detailed description of these protocols can be found in http://chromatin.bio.fsu.edu/docs/CGH_userguide.

Segmentation and primary CNV prediction

NimbleScan v2.5 was used for the primary analysis of the signal intensity raw data taken from Nimbelgen in the form of ".pair files" and ".tif files" from the arrays. Segmentation algorithm CGH-segMNT implemented in NimbleScan was used for primary segmentation. For calling segments we used the default value of 0.1 for "Min segment difference". This value represents the minimum difference in the $\log_2$ ratio that two segments must exhibit, to be identified as separate segments. Nevertheless we increased the "Min segment length" from Nimbelscan default value of 2 to 5. This value represents the minimum number of consecutive probes that must exhibit a change in $\log_2$ ratio in order to call a segment. Prior to segmentation analysis, spatial correction and qspline fit normalization (Workman, *et al.* 2002) were applied to the raw data. This normalization compensates for inherent difference of signals between the 2 dyes. Finally we chose to generate GFF (General Feature Format, .gff) files which contain the $\log_2$ ratio of Cy3 and Cy5 for each probe plotted versus genomic position. Additionally

segmentation PDF plots as well as Data summary files which contains a summary of predicted segments were used to review the array data.

Visualizing through SignalMap software

SignalMap v1.9 software was used to review and visualize GFF files, generated by the CGH data analysis. Additionally human genome (hg)18 annotation files from www.nimblegen.com/human-annotation were imported into SignalMap. This also gave us the opportunity to see peak correlations between our data and annotation tracks or link to a web site with detailed information for a selected gene (see table 2.3)

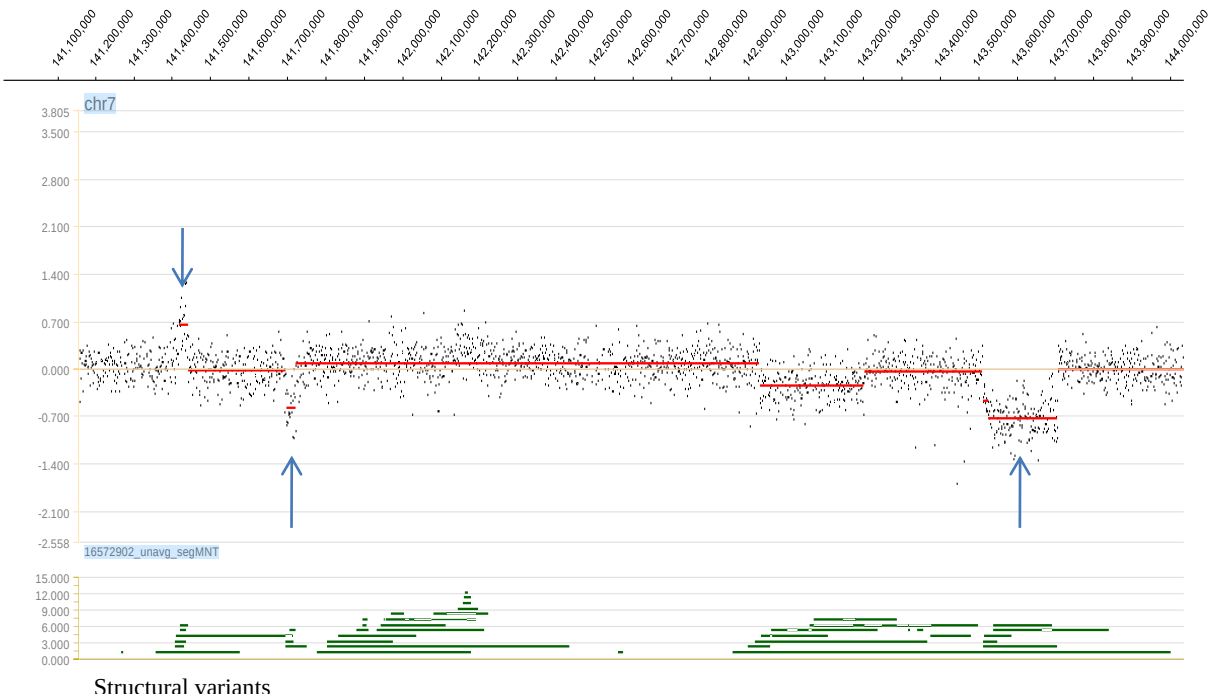


Figure 2.3 GFF file tracks from array-CGH Data visualized through SignalMap. Upper track shows the $\log_2$ ratio of signal intensities plotted versus genomic coordinates in base pairs. Predicted CNV segments (red horizontal lines) are pointed by vertical blue arrow lines. Lower track shows known structural variants (in green) from imported hg18 annotation files.

Annotation files

In order to prioritize predicted CNVs for further validation, a suite of human hg18 annotation files was used. Table 2.3 describes the annotation tracks that was recruited. We imported and viewed these files alongside our microarray data using Roche NimbleGen SignalMap software.

Annotation Files	Description
hg18: Genes.	Indicates all genes for build hg18 as reported in the UCSC Genome browser (http://genome.ucsc.edu). Genes annotated above the baseline in each track represent features identified on the sense strand, while entries below the baseline represent features identified on the antisense strand.
hg18: Genes_Exon-Intron	Indicates the exon-intron boundaries of all genes in build hg18 as reported in the UCSC Genome browser. Exons are denoted as dark blue bars, and introns are denoted as light blue bars
hg18: Transcription_Start_Sites	Indicates all transcription initiation sites for build hg18 as reported in the UCSC Genome browser
hg18: Structural_Variants	Displays all copy number variants as reported in the Database of Genomic Variants (http://projects.tcag.ca/variation)
hg18: 42M_CNV_Regions	Displays validated CNVs identified by the Genome Structural Variation Consortium in a high-resolution CNV discovery project (http://www.sanger.ac.uk/humgen/cnv/42mio). In this study, common CNVs > 500 bp were identified from 20 CEU and 20 YRI HapMap research samples using a set of NimbleGen CGH arrays that contains approximately 42 million probes tiled across the genome
hg18: Segmental_Duplications	Displays regions of genomic duplication > 1 kb in size and with > 90% sequence identity after masking high-copy repeat regions (Bailey, *et al.* 2001; 11:1005-17) and reported in the UCSC Genome browser. The level of similarity is indicated as follows: light to dark gray bars = 90 - 98% similarity, light to dark yellow bars = 98 - 99% similarity, light to dark orange bars ≥ 99% similarity; red = duplications of > 98% that lack sufficient evidence in the Segmental Duplication database
hg18: Cytogenetic_Ideogram	Displays the cytogenetic bands, in grayscale format, for each chromosome as reported in the UCSC Genome browser
hg18: miRNA	Indicates all miRNAs as reported in the miRBase database (http://microrna.sanger.ac.uk/).

Table 2.3 Description of annotation files imported and visualized alongside the array-CGH data.

2.4 UC case-control-study cohorts

We recruited 5 case control sample sets, one as screening (discovery) panel for CNVs and 4 others for follow-up. Here we describe them upon the platform used for CNV genotyping and origin of the samples;

Array-sample sets

Initial screening cohort consisted of 1121 German UC patients and 1770 healthy controls, previously used in a SNP-GWAS experiment using the Affymetrix® Genome-wide Human SNP array 6.0 (Affy6.0) and has been described previously (Franke *et al*, 2010). Additionally the Affy6.0 data sets of two independent cohorts, one Norwegian and one from UK were recruited. The Norwegian study population consisted of 274 clinically well-characterized UC patients and an ethnically and sex-matched group of Norwegian healthy controls (n=282), also studied and described previously (Franke *et al*, 2010). The UK study population was part of the "Welcome Trust case-control consortium 2" (WTCCC2) used for UC GWAS (UK IBD Genetics Consortium, 2009) and contained data sets of 2396 UC cases and 4886 controls after processing and filtrations described later.

TaqMan-sample sets

These samples included two disease cohorts originated from Germany and Lithuania which were genotyped for initially selected CNVs through real-time PCR automated by TaqMan CNV assays (see section 2.5.x). The German UC patients consisted of 245 males and 315 females. The German controls consisted of 779 females of age 18 to 81 (average age: 51) and 637 males of age 27 to 75 (average age: 50). The Lithuanian study population consisted of 443 UC patients and a control group of 1147 ethnically, age and sex-matched healthy blood donors (see also table 2.4).

case/control ->	Discovery Germany (1121/1770	Replication Germany (553/1420)	Replication Norway (274/282)	Replication Lithuania (443/1147)
Sex distribution; % Female				
controls	45.9	59.3	40.6	51.0
cases	56.9	59.7	47.8	50.6
Age distribution; median age (years)				
at sampling **controls**	56 ±12.9	38 ±12.4	31.8 ±6.9	40.2 ±12.8
at sampling **cases**	41 ±13.8	42 ±13.8	39.0 ±15.0	44.3 ±16.6
at onset	27 ±11.7	27 ±11.4	37.7 ±14.9	38.4 ±15.9
Disease extent				
Left sided colitis, %	51.2	56.7	69.4	75.8
Extensive colitis, %	46.8	39.1	30.6	24.2
Colectomy	5.7	11.7	6.0	1.1
Smoking habit at diagnosis, % cases				
current	23.8	36.4	13.1	12.2
previous	24.2	18.1	31.3	28.0
never	51.9	45.5	55.6	59.8

Table 2.4 Characteristics of UC case/control sets used for discovery and replication of CNVs
UC cohort of WTCCC2(UK) is not included in the table

In silico control sample sets

These samples comprised altogether 6724 individuals recruited in previous genotyping studies as; 60 unrelated HapMap CEU samples genotyped with the Illumina 1M Dua SNP array (Costello *et al*, 2005), 445 CEU controls genotyped with the Illumina 500k v3 (Cooper *et al*, 2008), 283 Caucasian controls genotyped with the Illumina Human Hap 300 and 231 Caucasian controls genotyped with the Illumina Human 610-Quad BeadChip (Simon-Sanchez, *et al*, 2007), 653 Caucasian controls genotyped with the Illumina Human Hap 300 and 551 Caucasian controls genotyped with the Illumina Human 610-Quad BeadChip Simon, *et al*, 2006; Albert *et al*, 2001), 3181 European controls genotyped with the Affymetrix® Human SNP array 6.0 (Stone *et al*, 2008).

2.4.1 Patient Recruitment and Ethics

Diagnosis of UC has been based on the review of the patients' original medical records including colonoscopies at the recruiting university hospitals. The currently accepted pathophysiological characteristics of UC include exclusive inflammation of the colon, continuity of inflammation, histological evidence for an inflammation limited to the mucosa, absence of granuloma, intestinal tract architectural changes including crypt abscesses, leukocyte aggregates, distortion of crypt architecture and cryptitis, mucosal edema, and infiltration of neutrophils (Podolsky, 2002). German patients of the discovery and replication panels were recruited either at the Department of General Internal Medicine of the Christian-Albrechts-University Kiel, the Charité University Hospital Berlin, through local outpatient services, or nationwide with the support of the German Crohn and Colitis Foundation. Clinical, radiological, histological, and endoscopic (i.e. type and distribution of lesions) examinations has been required to unequivocally confirm the diagnosis of ulcerative colitis (UC) (Lennard-Jones, 1989). 1214 German healthy control individuals of discovery panel (1703 total) were obtained from the biobankPopGen (http://www.popgen.de).

Popgen targets the population of northern Schleswig-Holstein (1.1 million people) which is surrounded by the Danish border (North), the North Sea and Elbe river (West), the Baltic Sea (East), and the Kiel Canal (South). The remaining 489 German healthy controls in discovery panel were selected from the KORA F4 survey, an independent population-based sample from the general population living in the region of Augsburg, Southern Germany (Wichmann *et al*, 2005). Written, informed consent had been obtained from all study participants and all protocols were approved either by the ethical committee of the University-Hospital Schleswig-Holstein, Center Kiel or through the institutional committee of the "Kompetenznetz Darmerkrankungen" (http://www.kompetenznetz-ced.de/) in Germany. The 274 clinically well-characterized Norwegian UC patients of replication panel were recruited through a population-based incidence study, the Inflammatory Bowel disease in South-Eastern Norway (IBSEN) study (Moum *et al*, 1996). An ethnically and sex-matched group of Norwegian healthy controls (n=282) was

randomly selected from the Norwegian Bone Marrow Donor Registry (NBMDR). The strict criteria (including absence of any autoimmune disease) on inclusion in the NBMDR ensured correct classification of these controls as healthy. Norwegian sample recruitment was approved by the ethics committee of Oslo University Hospital, Rikshospitalet, Norway.

The Lithuanian study population consisted of 443 UC patients recruited at 6 hospitals in Lithuania: Kaunas Medical University Hospital (Kaunas), Vilnius University Hospital Santariskiu Clinic(Vilnius), M. Marcinkevicius Hospital (Vilnius), Klaipeda Seamen's Hospital (Klaipeda),Panevezys Regional Hospital (Panevezys), Siauliai Regional Hospital (Siauliai) and 2hospitals in Latvia: P. Stradin Clinical University Hospital (Riga) andRiga Eastern ClinicalUniversity Hospital,Clinic Linezers (Riga). The diagnosis of UC was based on standard clinical, endoscopic, radiological and histological criteria. The control group consisted of 1157ethnically, age and sex-matched healthy blood donors. Written, informed consent was obtained from all study participants and all protocols were approved by the institutional ethical review committee of the Lithuanian University of Health Sciences, Kaunas, Lithuania.

2.5 CNV calling of Affy6.0 datasets

Affy6.0 platform consists of two main types of probe sets; SNP probes which include 936,000 SNPs (chosen from HapMap and dbSNP) and copy number probes (CN probes) consisting of 940,000 oligo-nucleotide probes which provides a uniform coverage across the genome to directly interrogate CNVs. In this platform a metric called $\log_2$ Ratio (LRR) compares the observed normalized intensity ($R_{observed}$) of the subject sample to the expected intensity (*R* expected) calculated from the collection of reference hybridizations, or the rest of the population being analysed (Peiffer, *et al*, 2006). Additionally a second metric, termed B allele frequency (BAF) is calculated as the proportion of the total allele signal (A + B) explained by a single allele (A). The BAF has a significantly higher per-probe signal-to-noise ratio than the log ratio data and can be interpreted as follows: a BAF of 0 represents the genotype (A/A or A/–), whereas 0.5 represents

(A/B) and 1 represents (B/B or B/–). Different BAF values occur for AAB and ABB genotypes or more complex genotypes (for example, AAAB, AABB and BBBA). Homozygous deletions result in a failure of the BAF to cluster (Cooper *et al*, 2008). These two parameters, LRR and BAF are then plotted along the entire genome for all probes on the array (see figures 2.4 and 2.5). For CNV calling from array data, raw image files were converted into CEL-files by Affymetrix® genotyping console. CEL files were processed with the Affymetrix Power Tools (APT) apt-copynumber-workflow version1.67. The values for contrastQC (based on Affymetrix®GTC 3.0.1 User Manual) and MAPD were extracted and samples that failed default QC values were discarded (MAPD > 0.4 and/or contrastQC < 0.4).

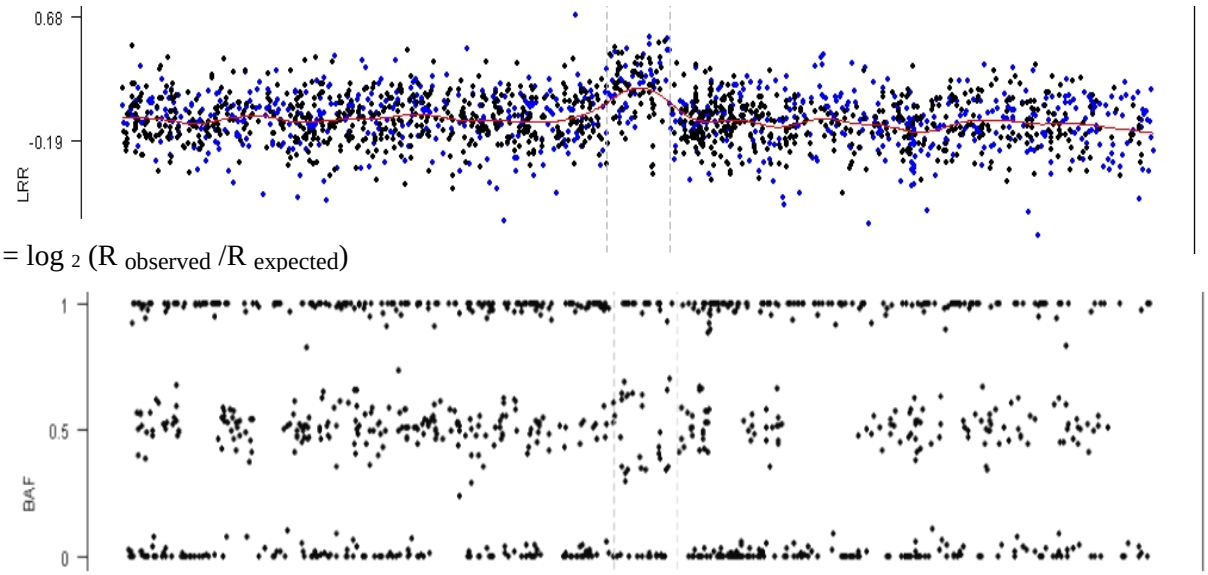


Figure 2.4 Raw data visualization of intensities as LRR and BAF for a predicted duplication on Affy6.0. In the LRR panel (upper) SNP probe sets are in black and copy number probe sets in blue whereas only SNP probes are used for BAF metric (lower panel). Duplication (blue horizontal bar) has spread the SNP probe distribution at frequencies about 0.5 (in BAF panel).

For the remaining samples an identity by state (IBS) and principal component analysis (PCA) was performed as described previously (Franke *et al*, 2010). The output of apt-copynumber-

workflow was used as the input file for CNV data mining tool “CNVineta” (Wittig *et al.*, 2010). CNVineta is an R package for rapid data mining and visualization of CNVs in large case-control datasets genotyped with SNP arrays. A preliminary batch-wise filtering was performed based on the number of called CNVs per samples. This was performed batch-wise, as one batch consists of a sample collection which was prepared in the same process. Outliers were defined as samples which had more CNVs than the 75% quantile plus 1.5 fold of the interquantile range.

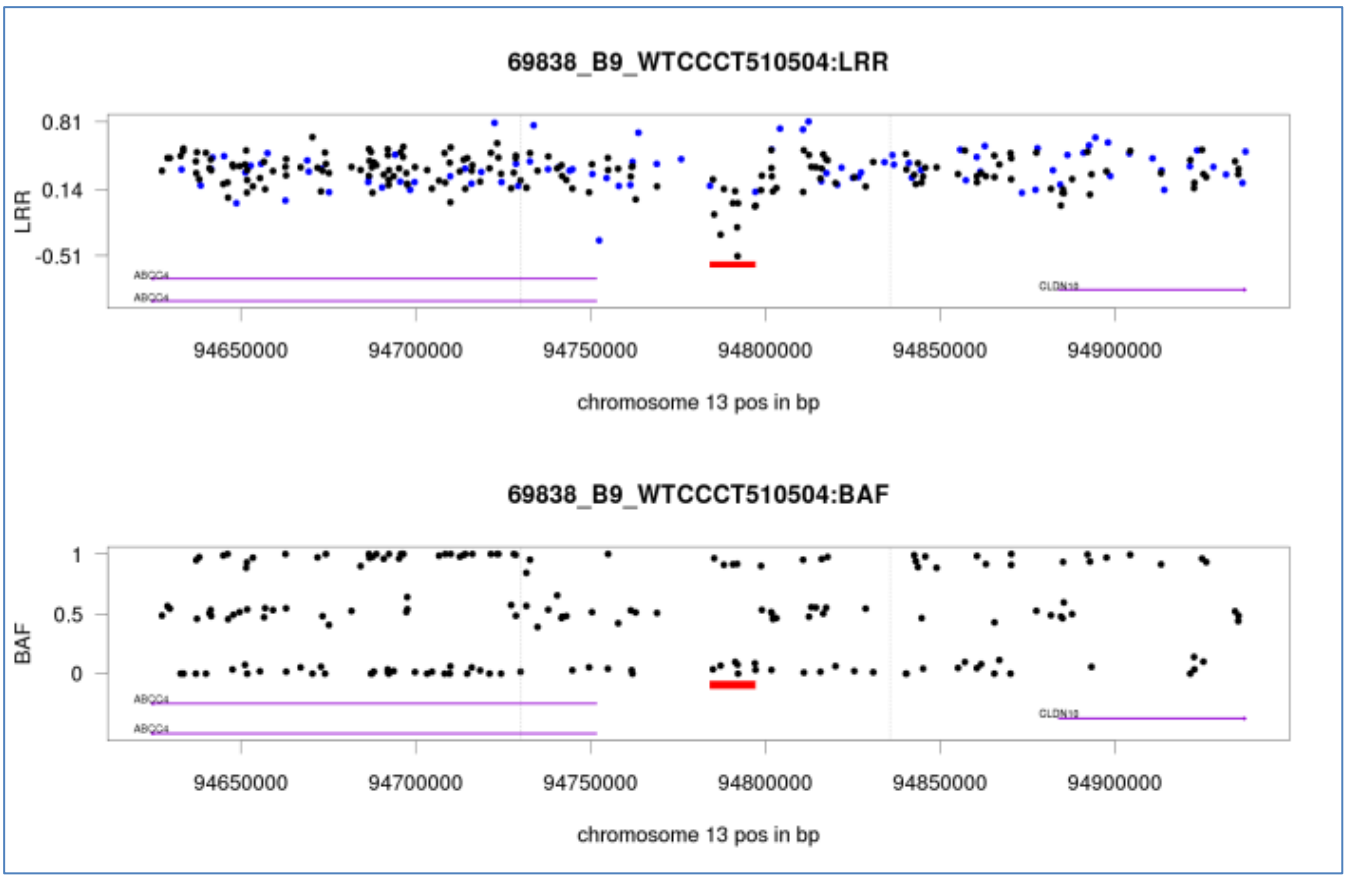


Figure 2.5 Raw data visualization of intensities as LRR and BAF for a predicted deletion in one individual on Affy6.0. In the LRR (upper) panel SNP probe sets are in black and copy number probe sets in blue. Only SNP probes are used for BAF metric (lower panel). Deletion (red horizontal bar) has resulted to loss of SNP probe distribution at frequency 0.5. (lower panel). RefSeq genes are annotated with purple lines.

The screening was followed by a rigorous manual raw data inspection for identifying false-negative and -positive CNVs. This approach especially reduces false-positive CNVs and can also uncover false-negatives at candidate positions. For the whole data mining process, the predicted

CNVs with less than 5 supporting probes per CNV and mean probe set distance less than one kilobase, were ignored.

2.6 Screening for rare CNVs in Affy6.0 datasets

To screen for rare CNVs, we used CNVineta (http://www.ikmb.uni-kiel.de/resources/download-tools/software/cnvineta). In this package a function called "dox0" is implemented which screens for rare deletions, duplications and CNVs without respect to the copy number state (deletion or duplication). In order to identify rare CNVs by this method, three parameters namely "min.diff", "max.one.side" and "cases.more.affected" should be specified. Trigger "min.diff" (default = 5) sets the parameter for the required absolute difference in CNVs within particular regions between cases and controls; "max.one.side" (default = 1) defines the maximum number of individuals with a CNV in a particular region for the group with the lower CNV load (usually controls); "cases.more.affected" (default = TRUE) defines the direction of the scan, i.e. whether cases (TRUE) or controls (FALSE) are screened for rare CNVs.

2.7 Association analysis for common CNVs

Screening of common CNVs and Association testing with UC was done through CNVineta using a logistic regression model (Dobson, 1990) with disease status as the outcome variable and total copy number as the dependent variable. This was undertaken by assigning integer copy number to each individual based on SNP allele hybridization (BAF) and intensity data (LRR) (see figure 2.5), followed by comparing obtained copy number states between cases and controls. To screen for common CNVs, CNVineta provides the *do.log.regression* method, which instead implements the so called "*glm*" function.

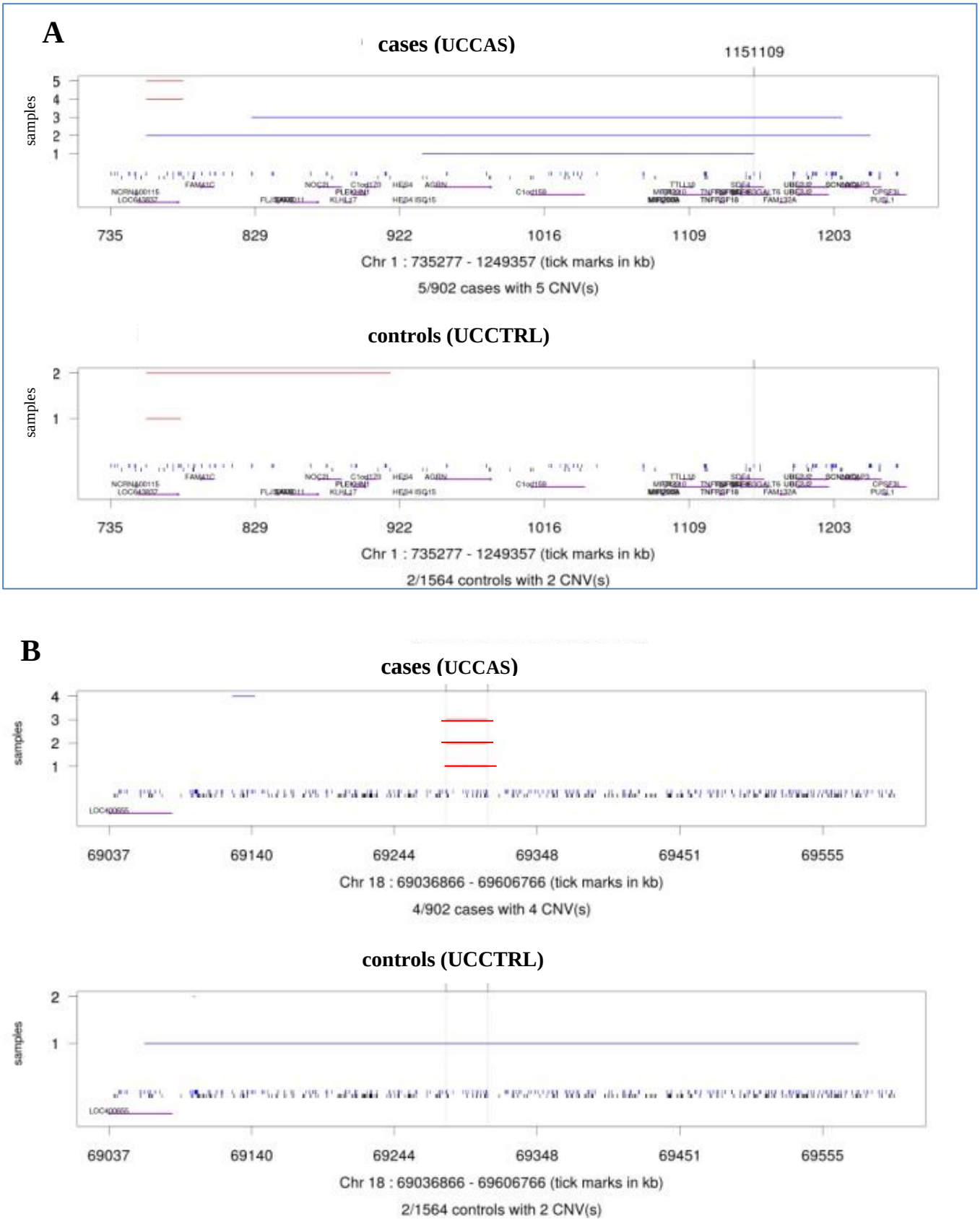


Figure 2.6 Example of rare CNVs, found in our UC German discovery panel screened by CNVineta. Three duplications predicted in cases and no in controls (A). Three deletions predicted in cases and no in controls (B). In each part the predicted CNVs are shown for cases (upper track) and controls (lower track). Duplications are seen in blue and deletions in red. Involved RefSeq genes are annotated. SNP probe sets are seen in black and copy number probe sets in blue.

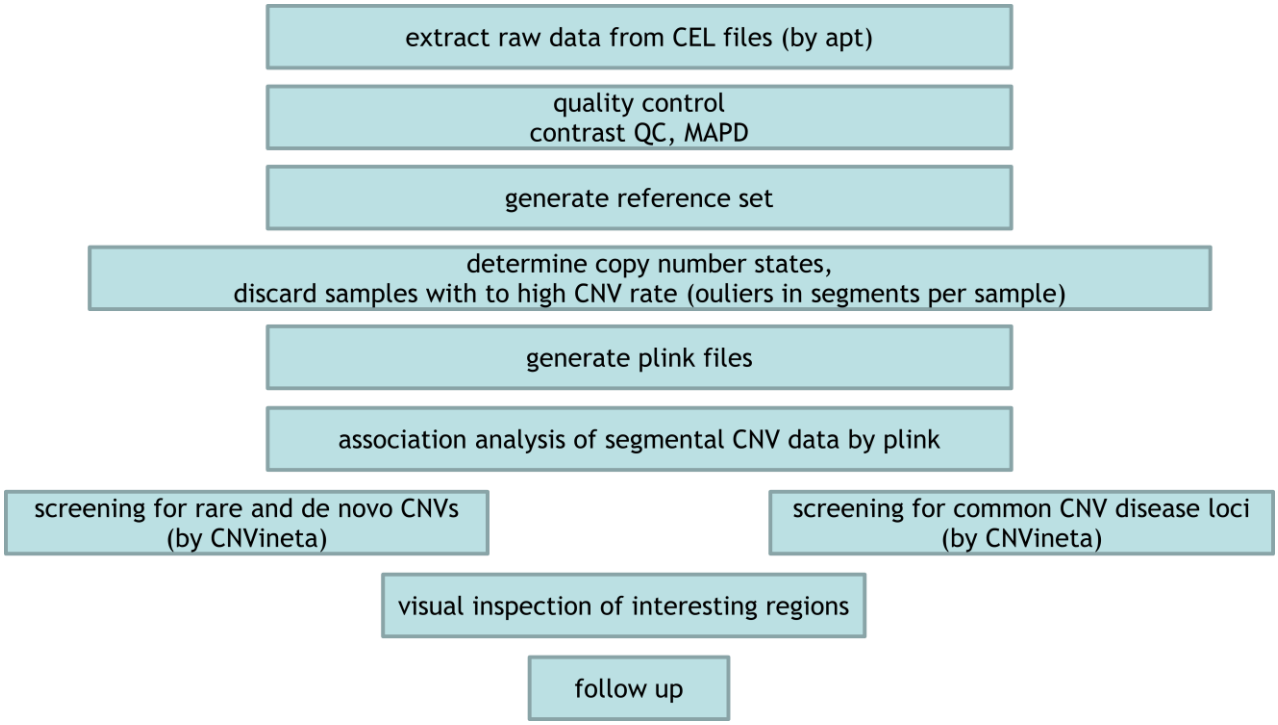


Figure 2.7 SNP-chip CNV analysis pipeline.

2.8 Technical validation and replication of predicted CNVs

For technical verification of the CNVs predicted from array-CGH data in twin analysis, we used TaqMan® CNV genotyping (Mayo *et al.*, 2010). This was also the main independent platform that we used for verification and replication of rare CNVs identified in the discovery panel from Affy6.0 dataset.

2.8.1 TaqMan® copy number analysis

Principles and Chemistry (Adopted from TaqMan® Copy Number Assay Protocol (PN 4397425D))

TaqMan® Copy Number analysis is based on quantitative real-time PCR (qPCR), comparing the signal from the test region against a reference locus and obtaining the ratio. Therefore TaqMan® Copy Number assays run simultaneously with a TaqMan® Copy Number Reference Assay in a duplex PCR. The Copy Number Assay detects the target gene or genomic sequence of interest, and the Reference Assay detects a sequence that is known to exist in two copies in the diploid genome (for example, the human RNase P H1 RNA gene). The number of copies of the target

sequence in each test sample is determined by relative quantization using the comparative CT ($\Delta\Delta C_T$) method.

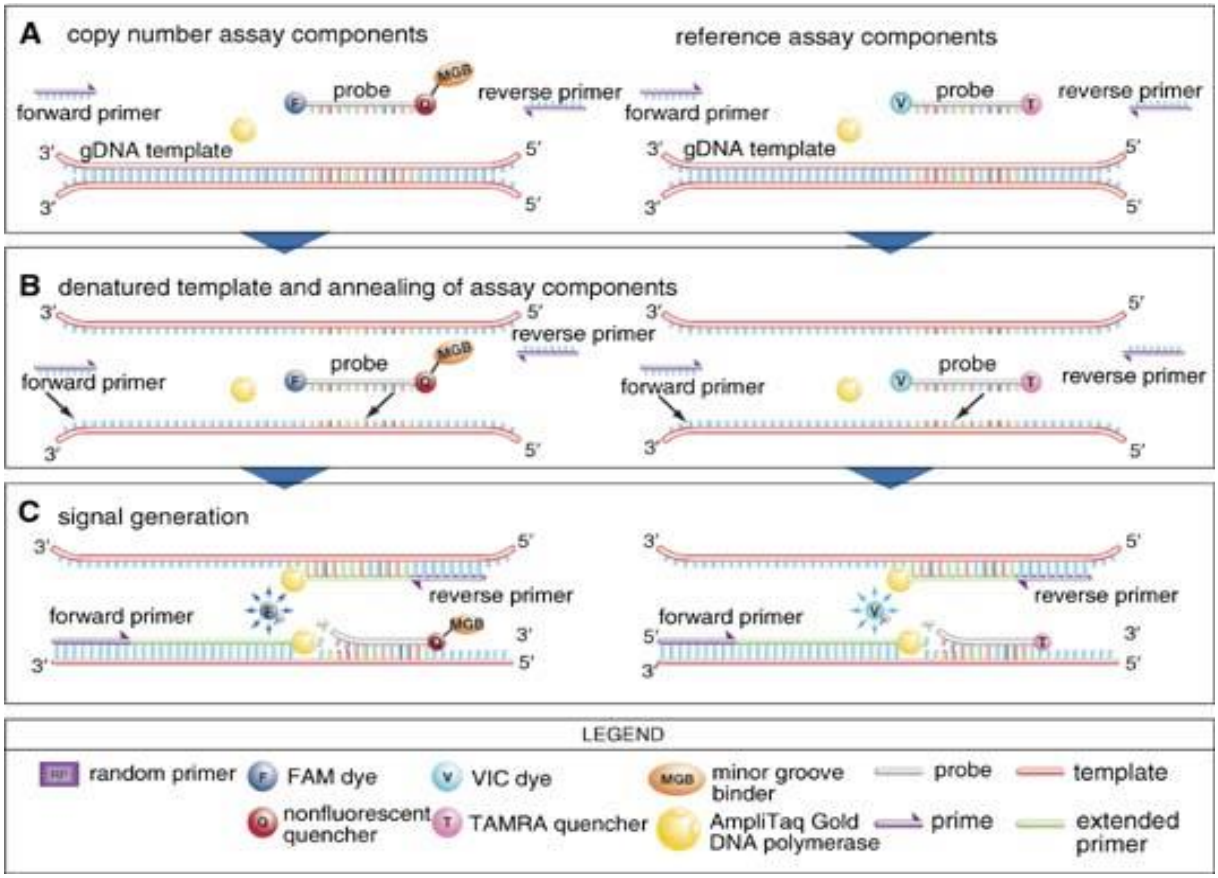


Figure 2.8 PCR and detection of target and reference gDNA sequences in a duplex reaction. A TaqMan® Copy Number Assay, a TaqMan® Copy Number Reference Assay, TaqMan® Genotyping Master Mix, and a gDNA sample are mixed together in a single well or tube (A). The gDNA template is denatured and each set of assay primers anneals to its specific target sequences. Each TaqMan® probe anneals specifically to its complementary sequence between forward and reverse primer binding sites (B). When each oligonucleotide probe is intact, the proximity of the quencher dye to the reporter dye causes the reporter dye signal to be quenched. During each round of PCR, the target and reference sequences are simultaneously amplified by AmpliTaq® Gold DNA Polymerase (C). This enzyme has a 5′ nuclease activity that cleaves probes that are hybridized to each amplicon sequence. When an oligonucleotide probe is cleaved by the AmpliTaq Gold DNA Polymerase 5′ nuclease activity, the quencher is separated from the reporter dye increasing the fluorescence of the reporter. Accumulation of PCRS products can be detected in real time by monitoring the increase in fluorescence of each reporter dye at each PCR cycle. (Printed from TaqMan Copy Number Assay Protocol (PN 4397425D))

This method measures the C_T difference (ΔC_T) between target and reference sequences, then compares the ΔC_T values of test samples to a calibrator sample known to have two copies of the target sequence. In a copy number quantification reaction, genomic DNA is combined with:

1. The TaqMan® Copy Number Assay, containing two primers and a FAM™ dye-labeled MGB probe to detect the genomic DNA target sequence.

2. The TaqMan® Copy Number Reference Assay, containing two primers and a VIC® dye-labeled TAMRA™ probe to detect the genomic DNA reference sequence.

3. The TaqMan® Genotyping Master Mix, containing AmpliTaq Gold® DNA Polymerase, UP (Ultra Pure) and dNTPs required for the PCR reactions.

Figure 2.8 shows the steps in a duplex PCR reaction containing copy number target and reference assays, both of which are 5′ nuclease assays. After amplification, data files containing the sample replicate CT values for each reporter dye can be exported from the real-time PCR instrument software and imported into a software analysis tool for post-PCR data analysis of copy number quantitation experiments.

Preparing the reactions for dried gDNA on 384 well-plates

For assessing the copy number of regions of interest, we used Pre-designed TaqMan® Copy Number Assays from Applied Biosystems. We used four technical replicates for each gDNA sample and applied the same sample name to the wells of each technical replicate group so that later CopyCaller® Software could combine data of replicate wells for calculating copy number.

Amplification reactions (10 μL), which were performed in quadruplicate, consisted of:

• 10 ng gDNA • 1X TaqMan® Copy Number Assay • 1X TaqMan® Copy Number Reference Assay, RNase P • 1X TaqMan® Genotyping Master Mix

Reaction mixture component	Volume per well (μL)	
	384-well plate	96-well plate
2× TaqMan® Genotyping Master Mix‡	5.0	10.0
TaqMan® Copy Number Assay, 20× working stock§	0.5	1.0
TaqMan® Copy Number Reference Assay, 20×	0.5	1.0
Nuclease-free water	4.0	8.0
Total Volume	**10.0**	**20.0**

Running the plates

PCR was performed with an Applied Biosystems 7900HT Fast Real-Time PCR System using the default, universal cycling conditions according to below settings:

Stage	Temperature	Time
Hold	95 °C	10 min
Cycle (40 Cycles)	95 °C	15 sec
	60 °C	60 sec

Cycle threshold (C_T) values were generated by the SDS v2.0 software with manual C_T threshold set to 0.2 according to Applied Biosystem recommendations.

2.8.2 Copy number data analysis of TaqMan® Assays

CopyCaller® Software v1.0 was used for TaqMan® Copy Number Assay data analysis. After amplification, the experiment results table, containing C_T values for the copy number and reference assay for each well, was exported from the Applied Biosystems real-time PCR system software and then imported into CopyCaller® Software for post-PCR data analysis.

CopyCaller® Software performs a comparative C_T ($\Delta\Delta C_T$) relative quantification analysis of the real-time data. The analysis determines the number of copies of the target sequence in each test genomic DNA sample. $\Delta\Delta C_T$ method first calculates the difference (ΔC_T) between the threshold cycles of the target and reference assay sequences. Then, the method compares the ΔC_T values of the test samples to the calibrator sample that contains a known number of copies of the target sequence:

$(\Delta\Delta CT)_{s,t} = \mu(\Delta CT)_{s,t} - \mu(\Delta CT)_{calibrator}$, where: s = sample; t = target copy number assay.

Then the relative quantity (RQ) for the associated sample is calculated as:

$RQ_{(s,t)} = 2^{-(\Delta\Delta CT)s,t}$, where: s = sample; t = target copy number assay.

Finally the copy number for the associated sample is calculated by the relative quantitation to the calibrator sampel as:

$cn_{sample} = RQ_{sample} \times cn_{calibrator}$, where cn = copy number.

In CopyCaller® Software, for each selected copy number assay, a plot displays the calculated or predicted copy number of each sample, and bars that indicate the copy number range for the associated replicate group (see Figure 2.9). Within each bar, a red line (bar) represents the minimum and maximum copy number calculated for the sample replicate group.

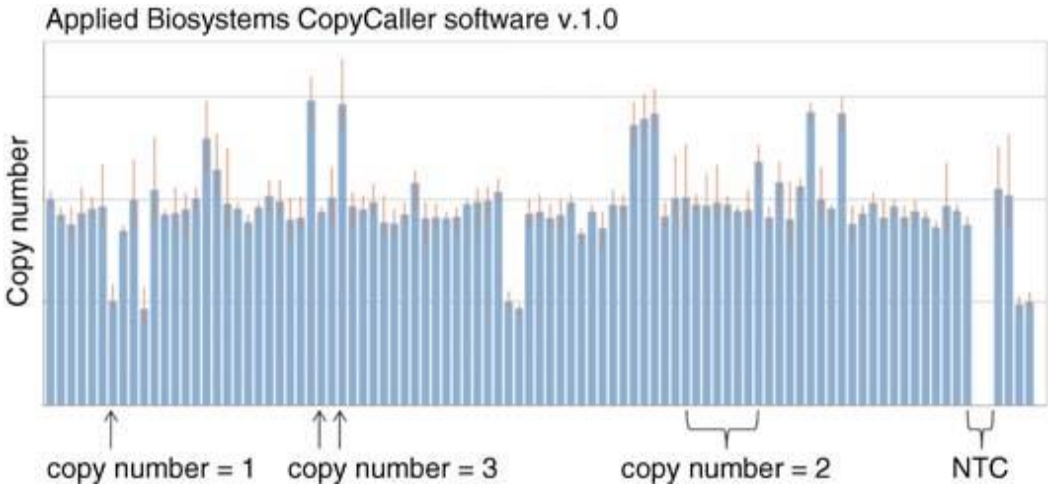


Figure 2.9 Copy Number Plot displayed in CopyCaller® Software. NTC (No Template Controls)

Breakpoint mapping

Based on the genomic resolution provided by custom aCGH for the deletion, five flanking primers at each end of the predicted deletion were designed. The subsequent PCR then only yielded amplicons, if a deletion was present (without deletion the fragment is longer than 15 kb, no long range PCR was performed). All possible primer combinations were tested and an amplified fragment of about 610 nucleotides was used for Sanger sequencing. (Forward Primer: 5'-TCCTTCCAGCATATCCCATC; Reverse Primer: 3'-GAATACTGATAACCACAAACAGACAGA). The resulted sequence was then used for BLAT query with the human genome sequence hg18 reference. For the duplications, breakpoints were derived from the aCGH mapping experiment.

3 Results

3.1 CNV identification in MZ twins

We conducted a comprehensive genome-wide screen for CNVs in 6 monozygotic twin pairs. For each twin pair, two array-CGH experiments were performed. In one, genomic DNA from peripheral blood lymphocytes of the IBD individual was used as the test sample, while in the other the test sample was genomic DNA obtained from the biopsy of the IBD patient. In both arrays for each twin, the peripheral blood DNA of the non-IBD co-twin (referred to as healthy individual) was used as the reference sample. Therefore this study included 12 microarray experiments for genomic comparison of six IBD-discordant twin pairs.

3.1.1 Primary array-CGH results

Primary implementation of segmentation algorithm on normalized intensity ratios of arrays produces a set of segments but does not classify them into gains and losses. In 12 IBD twin arrays, the highest and lowest mean $\log_2$ratio for the segments were 0.57 and -0.6 respectively. We initially selected predicted segments which spanned at least 10 probes and set the $\log_2$ratio thresholds as -+0.25, below/above which a segment represent a potential loss/gain CNV candidate upon previous descriptions. Table 3.1 presents the predicted segments which were suggestive of potential copy number variants after applying these criteria for CNV prediction.

3.1.2 Technical validation of the predicted CNVs

The predefined segments were visualized through SignalMap software, and TaqMan assays with regard to array-predicted breakpoints for 57 promising CNV candidates were designed. None of these regions showed consistent copy number differences in the corresponding twin-DNA samples, analyzed by quantitative PCR. Only two predicted regions in DNA from the biopsy of 2 IBD patients compared to the blood DNA of their healthy co-twins, showed slight sign of copy

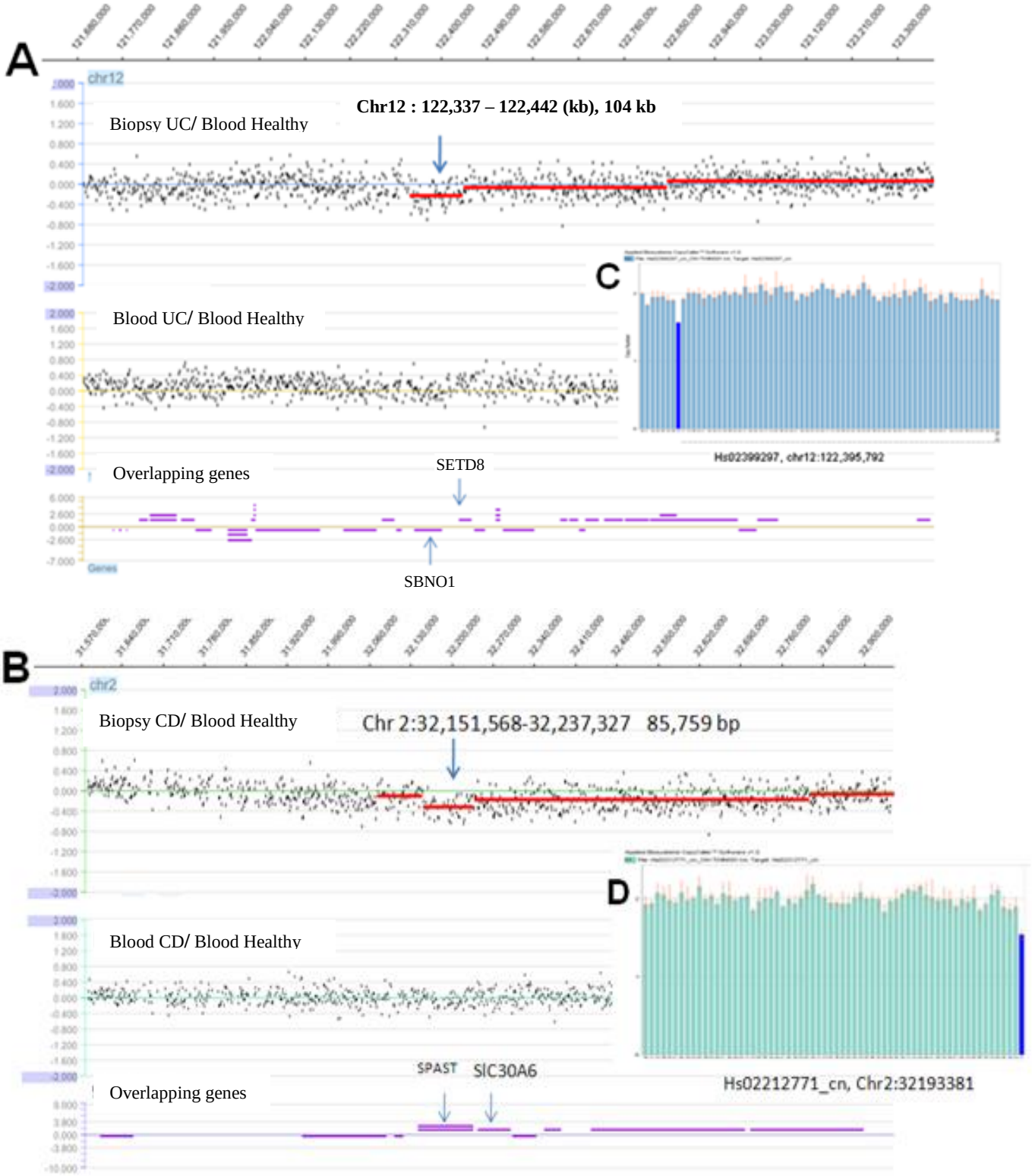


Figure 3.1 Screenshot of two genomic regions suggestive of CNVs between IBD-discordant MZ twins. 104 kb predicted deletion (loss) in genomic DNA of biopsy from a male UC patient relative to blood DNA of the non-IBD co-twin (upper track of A). 85 kb predicted deletion in biopsy DNA from a female CD patient relative to blood of the healthy co-twin (upper track of B).The signal was not predicted in blood DNA from these patients compared to blood DNA of their co-twins (middle track in each A and B). Quantitative PCR assays of the predicted CNVs (C & D)

number difference in quantitative PCR (figure 3.1). However these signals were not reproduced after repeating the real time PCR experiments with newer DNA from biopsies of the corresponding patients.

In order to increase sensitivity for detection of probable CNVs between MZ twins, we set a second less stringent log_2ratio threshold of -+0.15 (Figure 3.2). This was assuming that the probable post-zygotic CNVs might occur in only a proportion of cells and therefore are not reflected by Sharp aberrations in signal intensity ratios on CGH arrays. But applying a more permissive log_2ratio threshold, to increase power for hunting probable hidden CNVs can lead to higher rates of false positive predictions. To somewhat compensate for this, we doubled the minimum number of the probes (to n=20) required to be spanned by a predicted segment, in order for the segment to be selected for further validations. 17 segments fulfilled these criteria and were next assayed by quantitative PCR. Copy number variations could not be verified in these regions. In total we assessed relative copy number status of six twin pair members in 74 genomic regions suggested from array-CGH analysis and we did not find any reproducible variations between individuals of MZ twins.

3.1.3 Efficiency testing of the platform for twin CNV analysis

To have an empirical assessment of the ability of our experiment platform to efficiently detect CNVs, we tested the genomic DNA of the HapMap sample NA15510 against the blood DNA of a non-IBD member of a MZ twin pair on the same array platform used for twin pair comparisons (here referred to as NA15510/twin array). Copy number genomic variants of NA15510 have been characterized previously with high-throughput CNV detection methodologies of paired-end mapping (PEM) and massively parallel sequencing (Korbel *et al*, 2007; Kid *et al*, 2008) and are cataloged in the database of genomic variants (see also 1.2.1).

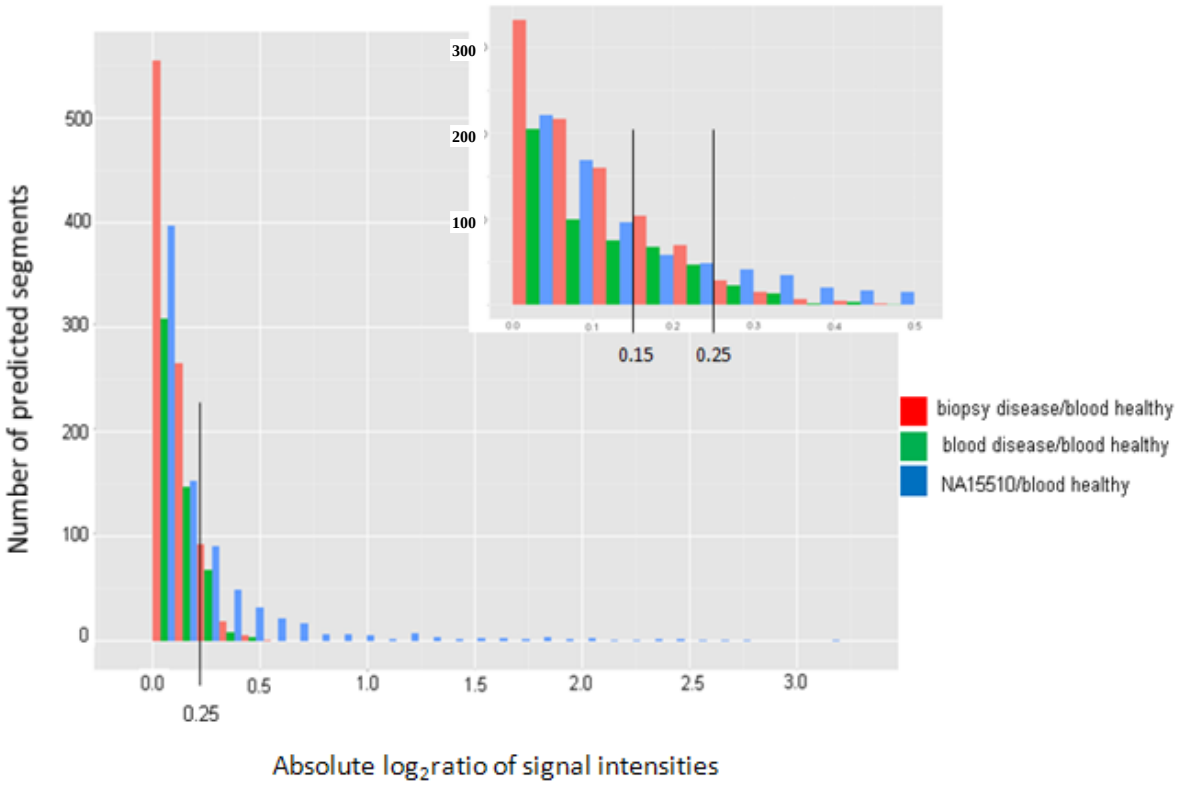


Figure 3.2 Distribution comparison of mean log_2ratio of signal intensities of predicted segments in 3 different array combinations in one twin pair after processing with segmentation algorithm. The predicted segments from 3 different array combinations in one twin pairs are showed with different colors. The absolute log_2ratio 0.25 is the threshold above which a predicted segment potentially represent a CNV. The less conservative threshold 0.15 (upper right plot) was set to include segments that probably represent mosaic CNVs occurring in only a minority of cells. Segments with highly aberrant signal intensities (absolute log_2ratio >=0.5) and covered by at least 10 probes were only seen in the NA15510/ blood healthy (germline inter individual CNVs).

This gave us the opportunity to have an estimation of the reliability of our array-CGH CNV analysis approach. Figure 3.2 shows distribution comparisons of log_2ratios of signal intensities for predicted segments in 3 different array combinations in this twin pair.

In NA15510/twin array, from 191 autosomal predicted segments with the above described criteria (absolute log_2ratio >= 0.25 and probe coverage >=10), 61 segments were overlapping more than 60% with a CNV region reported for NA15510 in DGV (see table 3.1) and hence could be confirmed as true positive predictions. In order to declare that two variants detected on different platforms correspond to the same event, at least 40% reciprocal overlap between them is required (Conrad *et al*, 2010). According to this threshold, the remaining predicted segments were not fined in DGV records for NA15510. The intensity ratios on array-CGH, however shows the relative copy number status of the test versus reference genome. Based on this, even assuming an ideally unbiased array experiment with high specificity for CNV detection, half of the predicted segments on NA15510/blood healthy array (i.e. 191/2) are derived from copy number variants in individual NA15510. Therefore beside 61 truly predicted CNV segments in the test experiment, at least 34 other predicted segments for this individual have not been confirmed and represented the minimum amount of false positive predictions for NA15510 in our test platform. Yet this comparative estimation of performance might be affected after correcting for regions, in which both the test and reference sample harbor copy number variations (Ju *et al*, 2010) and also for complications of comparison between different (PEM versus array-CGH) CNV detection methodologies (Alkan *et al*, 2011)

Chr	Start	End	Size (bp)	No of Probes	Log_2 Ratio	DGV Entry
1	110,025,939	110,056,866	30,927	20	-0.5	dgv276n100
1	150,819,257	150,853,362	34,105	26	0.45	dgv417n100
2	88,943,435	89,161,015	217,580	171	-0.32	dgv3930n100
2	88,943,435	89,161,015	217,580	171	-0.32	dgv3930n100
3	20,262,559	20,604,344	341,785	233	0.3	esv2422249
3	163,996,948	164,009,807	12,859	12	-0.99	nsv460966
3	164,028,400	164,109,371	80,971	58	-0.62	dgv4969n100
4	9,827,621	9,842,922	15,301	15	-1.72	nsv461259
5	150,184,939	150,203,537	18,598	15	-0.45	esv2421997
6	31,464,011	31,558,278	94,267	64	-0.53	nsv470809
6	32,560,897	32,686,692	125,795	92	-0.74	nsv462881
7	38,354,952	38,374,705	19,753	18	0.35	nsv1016904
7	151,556,420	151,621,542	65,122	48	-0.25	nsv1161668
8	6,822,624	6,869,541	46,917	41	0.27	dgv11908n54
8	51,115,286	51,200,865	85,579	61	-0.34	dgv1396e214
10	124,321,737	124,347,792	26,055	23	0.3	nsv1159772
11	4,920,866	4,933,642	12,776	12	-0.69	nsv521088
11	5,741,594	5,763,341	21,747	18	-1.46	nsv1075273
11	55,122,740	55,198,065	75,325	53	-1.16	nsv1053424
11	60,724,304	60,776,689	52,385	40	0.76	dgv1212n100
13	68,143,472	68,163,932	20,460	19	-0.88	esv3632562
14	105,265,168	105,315,345	50,177	14	-0.42	nsv566093
14	105,317,329	105,397,727	80,398	58	-1.44	nsv566139
14	105,406,066	105,428,623	22,557	11	-1.98	nsv821194
14	105,430,230	105,481,651	51,421	37	-0.7	nsv952276
15	19,093,219	19,475,422	382,203	266	0.35	dgv418n67
15	19,763,549	19,874,394	110,845	86	0.47	esv3415650
15	19,885,079	20,199,292	314,213	142	0.42	dgv2170n100
15	32,470,093	32,532,596	62,503	50	-0.4	nsv568863
15	32,534,148	32,584,558	50,410	39	-0.59	dgv2539n100
15	32,585,521	32,659,076	73,555	45	-0.29	esv3892660
16	21,648,297	21,713,878	65,581	50	0.41	esv3638187

16	22,467,708	22,529,247	61,539	47	0.46	nsv433440
16	83,208,964	83,228,402	19,438	16	-0.57	dgv3053n100
17	9,194,999	9,210,169	15,170	13	0.38	nsv833356
17	21,961,590	22,068,041	106,451	83	-0.27	esv3892986
17	31,474,225	31,505,847	31,622	27	1.31	nsv457722
17	31,538,950	31,568,700	29,750	27	0.61	nsv833418
17	31,626,182	31,669,315	43,133	37	0.72	nsv2027
19	46,052,476	46,082,217	29,741	23	-0.47	nsv520061
19	56,827,143	56,839,988	12,845	13	-0.69	esv3644863
19	59,422,687	59,440,139	17,452	15	-1.03	esv28703
20	1,516,595	1,532,009	15,414	13	1.85	esv3644980
20	14,721,773	14,882,268	160,495	130	-0.78	nsv433323
20	14,884,055	14,896,209	12,154	11	0.32	esv29969
20	14,896,722	14,951,317	54,595	42	0.62	esv29969
21	9,997,428	10,202,297	204,869	84	-0.38	esv2723012
22	21,485,112	21,568,623	83,511	69	-0.62	nsv588309
22	22,613,623	22,682,174	68,551	54	-0.38	esv2507410
22	37,689,119	37,718,726	29,607	22	-0.39	nsv471196

Table 3.1 61 CNV segments predicted for HapMap individual NA15510 from our NA15510 / twin array.

Only segments overlapping more than 60% (in size) with a corresponding CNV segment reported for NA15510 in DGV (Database of Genomic Variants) are listed. In NA15510/twin array the genomic DNA of the HapMap individual NA15510 were tested against the blood DNA of a non-IBD member of a MZ twin pair on the same array platform used for twin pair comparisons. Therefore segments with negative $\log_2$ ratios show deletions and positive $\log_2$ ratio segments indicate duplications in NA15510 in our test array. NA15510 was used since its CNVs have been mapped in previous studies (Korbel *et al*., 2007; Kid *et al*., 2008) and are catalogued in DGV (Build GRCh37: Feb. 2009, hg19). Variant regions are assigned an nsv (NCBI; dbVar) or esv (EBI; DGVa) or dgv (database of genomic variants).

3.2 CNV analysis in UC case-control samples

3.2.1 Screening in German discovery panel

Initially a total of 2891 SNP array 6.0 CEL files (1121 German UC patients / 1770 matched controls) were subjected to CNV calling, which left 2466 samples (902 cases / 1564 controls) after discarding outliers with respect to raw data quality, per sample call rate, ethnic origin and relatedness. The primary aim was to identify rare variants in UC patients, which were absent or were underrepresented in controls. Therefore, regions of interest were defined in the primary sample as genomic segments containing CNVs in at least three cases and in no controls. This selection criterion was based on the inspection of the raw data plots where too many false-positives were among the singleton and doubleton predicted events. Furthermore deletions or duplications with one carrier among the controls were also included, when at least five cases contained the corresponding event. Those CNVs occurred in more than two controls were excluded. Upon this setting, the discovery sample yielded 151 CNV regions through screening with our data-mining tool CNVineta.

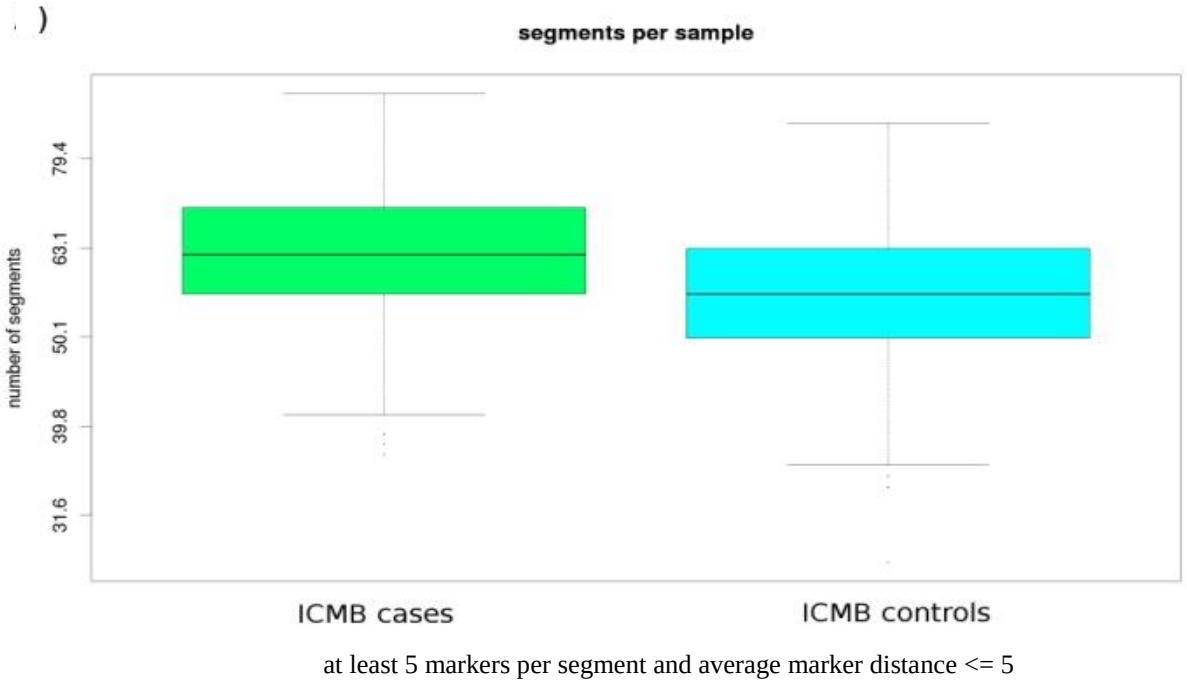


Figure 3.3 Number of copy number segments per sample identified from affy6.0 dataset in UC discovery panel.

3.2.2 Visual inspection of initial CNV regions

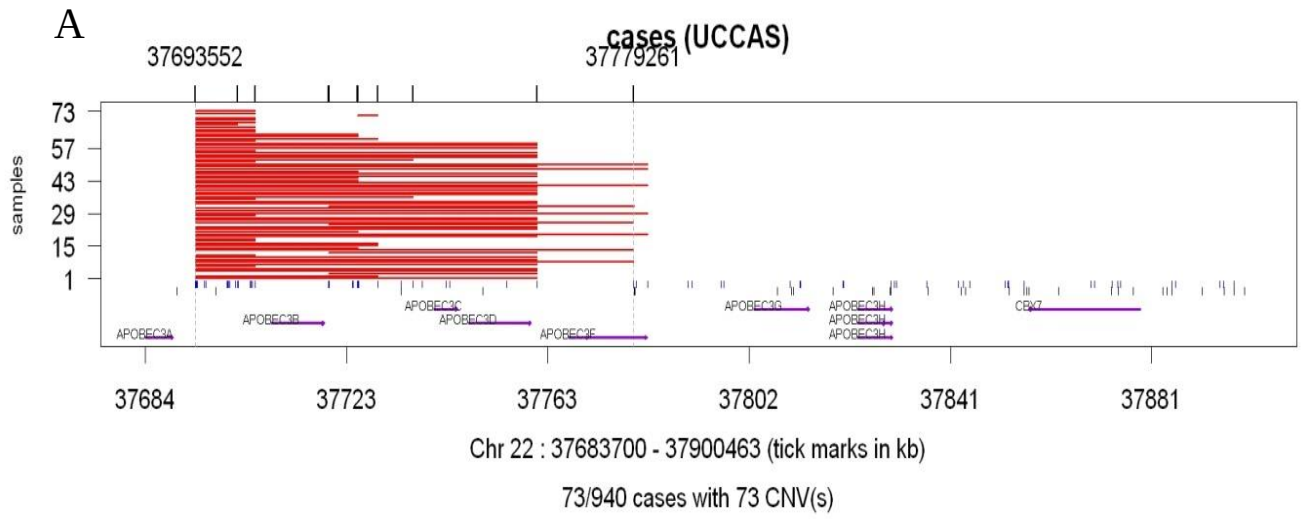
151 rare CNVs as well as 218 common CNV regions were then inspected manually and CNV regions were discarded when; predicted events had a complex breakpoint pattern; covered by less than 10 probe sets or spanned large genomic gaps (e.g. if a predicted CNV contained a gap which was larger than the part(s) covered by array probes). 24 rare candidates (14 deletions and 10 duplications) remained after manual inspection of their *Z*-scores, Log *R* ratio and B-allele frequency traces (See table 3.2).

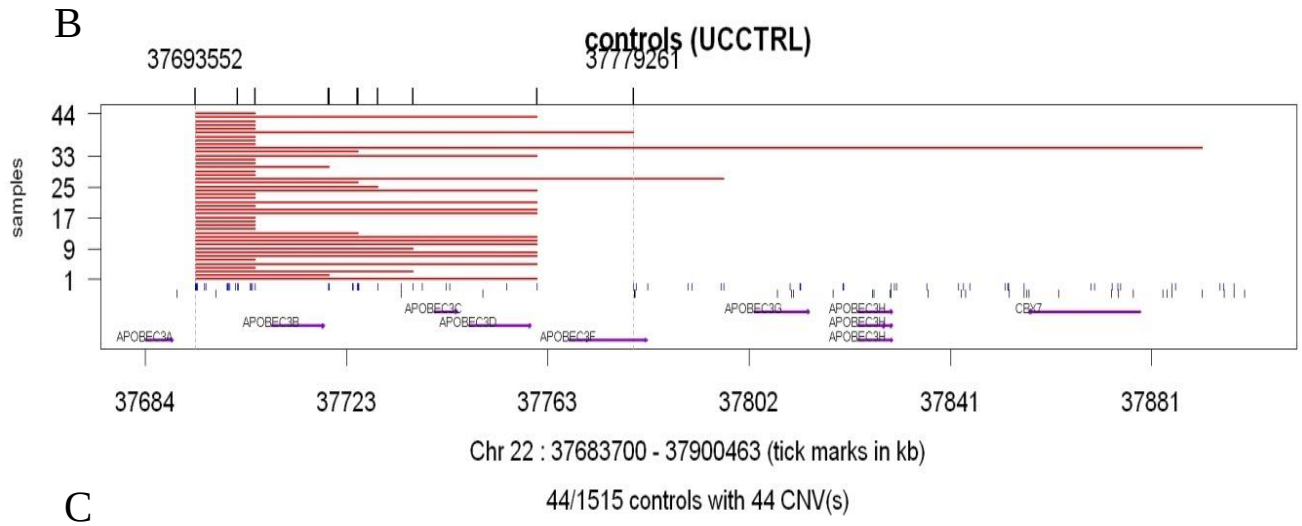
		German discovery CNV counts		WTCCC2 sample CNV counts		German discovery CNV freq in %		WTCCC2 sample CNV freq in %	
genomic position	cytoband	902 cases	1564 controls	2396 cases	4886 controls	cases	controls	cases	controls
chr1:1151109-1151109	1p36.33	3	0	2	0	0.33	0.00	0.08	0.00
chr2:221794031-221797760	2q36.1	3	0	7	11	0.33	0.00	0.29	0.23
chr2:228621916-228622575	2q36.3	23	0	0	0	2.55	0.00	0.00	0.00
chr3:195056555-195066854	3q29	4	0	0	1	0.44	0.00	0.00	0.02
chr3:36657809-36657809	3p22.2	3	0	0	2	0.33	0.00	0.00	0.04
chr4:134431169-134509830	4q28.3	4	0	1	4	0.44	0.00	0.04	0.08
chr4:35052424-35087926	4p15.1	6	1	11	16	0.67	0.06	0.46	0.33
chr4:95967326-95990607	4q22.3	6	0	0	1	0.67	0.00	0.00	0.02
chr5:101083525-101123064	5q21.1	3	0	1	5	0.33	0.00	0.04	0.10
chr6:58882675-58882675	6p11.1	3	0	2	5	0.33	0.00	0.08	0.10
chr7:107596648-107618181	7q31.1	3	0	0	2	0.33	0.00	0.00	0.04
chr7:4584859-4587453	7p22.1	4	0	0	0	0.44	0.00	0.00	0.00
chr7:5937448-5937448	7p22.1	3	0	3	0	0.33	0.00	0.13	0.00
chr7:71459253-71473235	7q11.22	3	0	0	0	0.33	0.00	0.00	0.00
chr8:140523022-40523022	8q24.3	4	0	5	2	0.44	0.00	0.21	0.04
chr10:20830090-20855699	10p12.31	4	0	8	18	0.44	0.00	0.33	0.37
chr11:104259645-104265284	11q22.3	6	0	5	4	0.67	0.00	0.21	0.08
chr11:108369095-108369095	11q22.3	3	0	0	0	0.33	0.00	0.00	0.00
chr13:23991650-23994304	13q12.12	3	0	2	1	0.33	0.00	0.08	0.02
chr13:94784155-94797183	13q32.1	5	1	5	6	0.55	0.06	0.21	0.12
chr14:21768464-22020331	14q11.2	17	1	42	28	1.88	0.06	1.75	0.57
chr18:69281471-69311825	18q22.3	3	0	0	7	0.33	0.00	0.00	0.14
chr19:19801585-19823239	19p12	11	3	6	4	1.22	0.19	0.25	0.08
chr19:58031861-58045469	19q13.41	3	0	2	10	0.33	0.00	0.08	0.20

Table 3.2 Frequencies of the 24 identified rare CNVs in German discovery panel in comparison with the WTCCC2 replication panel. Duplications are marked in blue, deletions in red. The 2 duplications, overrepresented in cases (compared to controls) in both panels (German discovery & WTCCC2) as well as the deletion (significant only in German discovery panel) are marked in green. These 3 regions were followed in additional independent cohorts.

3.2.3 Relevant common CNV regions in German discovery panel

Region : Chr 22_37,693,552 - 37,779,261 (APOBEC3)





C

Chr	Start (kb)	End (kb)	size	Type	Frequency	Gene
22	39,358	39,388	30 kb	Deletion	9 -11 % (European)	APOBEC3B

Figure 3.4 Regional Plots for the common deletion overlapping APOBEC3B

A) Predicted deletions in cases of German discovery panel (hg 18)
B) Predicted deletions in controls of German discovery panel (hg 18)
C) Genomic coordinates of APOBEC3 deletion region in DGV (hg 19)

Region: Chr 7_142,148,564 - 142,171,261 (TRBC1, T Cell Receptor Beta Constanst 1)

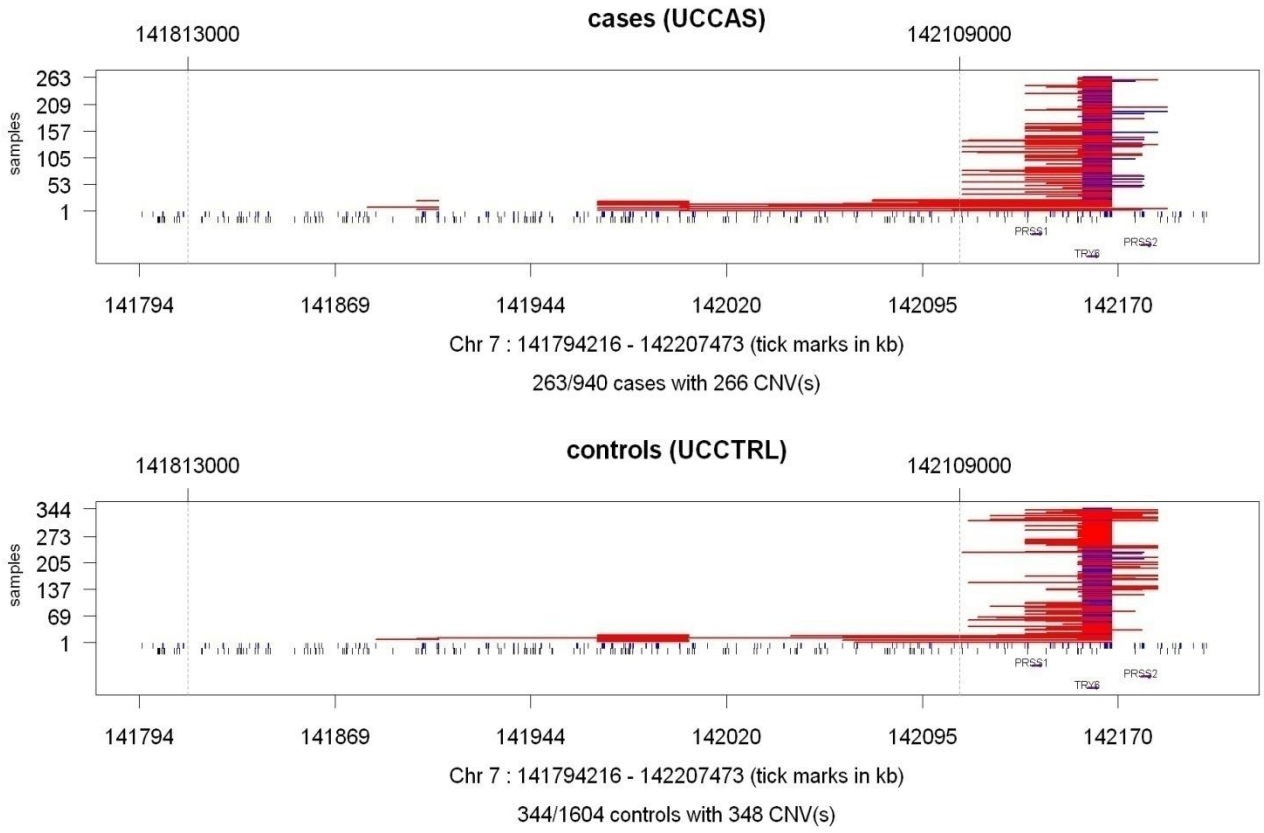


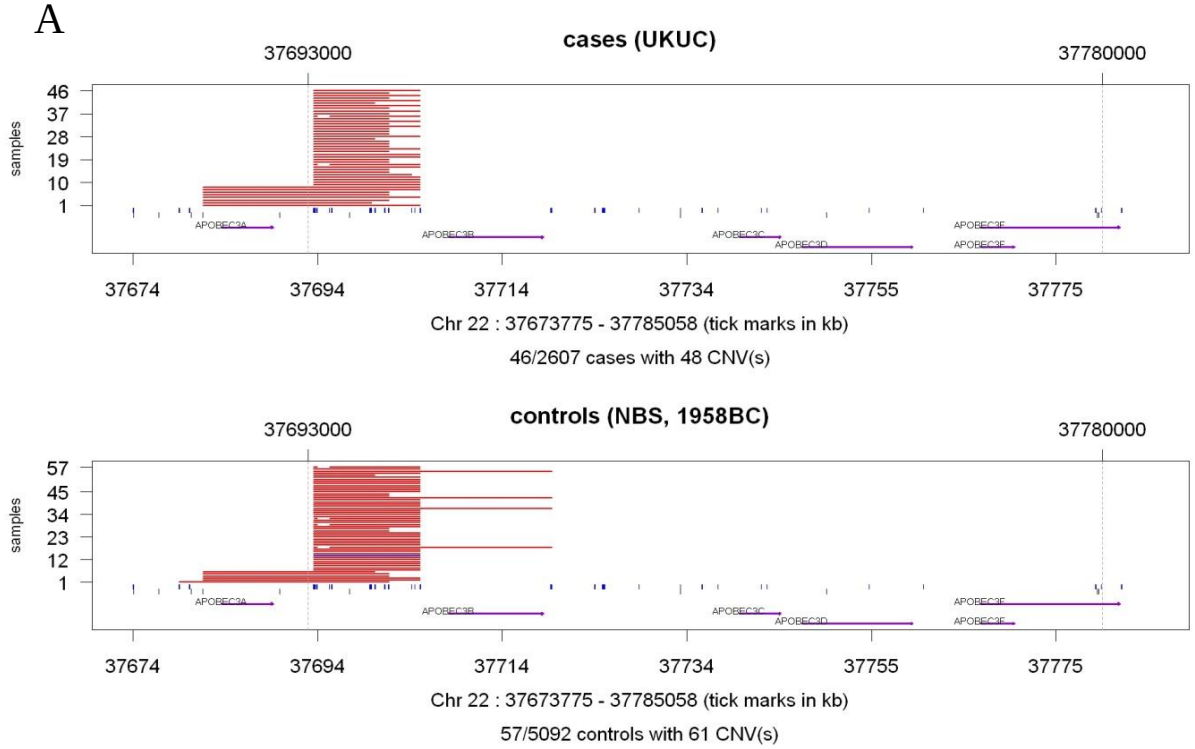
Figure 3. 5 Regional Plot for the common deletion encompassing T Cell Receptor gene in German discovery panel.

3.2.4 Following in independent replication panels

The status of the 24 rare CNVs (table 3.2) and the APOBEC3 common deletion polymorphism were then evaluated in two independent cohorts; the WTCCC2 sample (form UK), with CNV genotypes called from Affy6.0 intensity data, and a German cohort (453 UC patients, 1377 controls) genotyped for the selected CNVs by quantitative PCR through TaqMan CNV assays. In British cohort (2394 cases, 4886 controls), of the 24 rare CNVs evaluated, two duplication events (single copy gains) were the only variants that showed the same distribution trend as the discovery panel (See table 3.2), i.e. more represented in cases compared to controls; A 119 kb large duplicated region at 7p22.1 carried by 3 cases and no control (3/902 cases, 0/1564 controls

in discovery panel) and a 134 kb duplication at 8q24.3 with 0.21% occurrence in cases versus 0.04% in controls, two-sided Fisher's exact $P = 0.058$ ($P = 0.018$ in discovery) (Table 3.3). These two duplications however, were not relevant in the German replication panel, as from the sum of 1830 samples (cases + controls) genotyped for these two variants only one individual (UC case) carried the duplication (Dup8q24.3). Further, of the 24 CNVs followed-up in the German replication panel, a 15.8 kb deletion (single copy loss) at 13q32.1 (chr13:94,781,525-94,797,285), reproduced the trend of association with nominal P-value of 0.005 (P = 0.027 in discovery). (**Table 3.3**). Del13q32.1 was not correlated with UC in the WTCCC2 sample.

A



B

German (cases)	German (ctrls)	Pvalue	Uk (cases)	Uk (ctrls)	Pvalue
0.077	0.03	8.25 E-08	0.017	0.011	0.0209

Figure 3.6 Regional plot of APOBEC3 deleted region in WTCCC2 data (A) and related statistics in German and UK Affy 6.0 data sets (B). ctrls (controls)

3.2.5 Evaluation of the relevant CNVs in *in-silico* controls

The frequency of the three relevant rare CNVs (one deletion and two duplications) were also estimated through SNP array-based genotyping data sets of a total of 6724 individuals recruited

in previous studies (see 2.4 and table 3.3). We observed that 5 of a total of 4501 individuals carried the Del13q32.1, while only 1 of 5788 individuals contained Dup7p22.1 and 3 of 6727 individuals contained Dup8q24.3. Genotypes for these 3 CNVs in *in silico* controls were not considered, when the applied platform in the corresponding study did not covered the CNV region with at least 10 probes. Upon this, genotype calls for Del 13q32.1 from 6 sample sets (highlighted by red background in table 3.3) were not included (as the region was covered by less than 10 probes in the applied arrays of the corresponding studies), yielding 4501 controls for this deletion. For Dup 7p22.1, only two sample sets (red zone in table 3.3) did not fulfil this criterion and were therefore excluded, resulting in 5788 controls for this duplication.

Samples	Ethnicity	Array	PMID	Description	chr13:94781525-94797285	chr7:5786323-5905210	chr8:140390975-140524875
60	CEU	Illumina 1M Duo	18776910	Hapmap	9	63	49
445	Caucasian	Illumina 550K, v3	17116639	NINDS ctrl set from Andy Singleton, II	5	12	29
283	Caucasian	Ill. Human Hap300	16516587	CAP subset of PARC (Caucasian only)	4	5	16
653	Caucasian	Ill. Human Hap300	11434828	PRINCE subset of PARC (Caucasian only)	4	5	16
231	Caucasian	Illumina 610quad	16516587	additional CAP samples	5	34	28
551	Caucasian	Illumina 610quad	11434828	additional PRINCE samples	5	34	28
1320	various European	Affymetrix 6.0	19592680	European unrelated from CHOP DB	13	54	107
3181	various European	Affymetrix 6.0	18668038	Ctrls. from ISC paper on Schizophrenia	13	54	107
6724					4501 controls	5788 controls	6724 controls

Table 3.3 Description of *in silico* controls. The three relevant CNVs were evaluated through the SNP array-based genotyping data sets of a total of 6724 individuals recruited in previous studies (see also 2.4). The number of controls taken from each study, the sample origin, the array type, the PubMed-id (PMID) and a short description of the publication are showed in the table. The three right columns show the number of array probes lying within the genomic region of the 3 relevant CNVs. The red highlighted background indicates the sample sets not used for each of three CNV loci (covered by less than 10 probes in the corresponding study). Last row shows the number of *in silico* controls we used for each CNV locus.

Figure 3. 7 Technical validations of the three relevant CNVs (one deletion and 2 duplications) in the German discovery panel

The left-side plots of each part A-C show the results of TaqMan® CNV assays. While the majority of samples have copy number state two (blue bars), green bars show deletion or duplication carriers. The right-side picture of each A-C illustration shows the Affymetrix Power Tools (APT) prediction of CNVs. For each region the predicted CNVs are shown for cases (upper track) and controls (lower track) with Duplications in blue and deletions in red. Involved RefSeq genes are annotated in purple, SNP probe sets in black and copy number probesets in blue. For all 3 CNV regions TaqMan® results show exact correlation with the predicted CNVs from array dataset, except for one sample (UCGER2_m_117134) with predicted Dup8q24.3, the duplication could not be confirmed with TaqMan® (shown in yellow in part C).

3.2.6 Technical validations of the relevant CNVs

For deletion Del13q32.1 and the two aforementioned duplications, we did a validation step, in which the genotypes of all 13 individuals of the discovery panel, predicted to carry these 3 CNVs (6 with Del13q32.1, 3 with Dup7p22.1 and 4 with Dup8q24.3) were confirmed through TaqMan assays (see Figure 3.4). Furthermore we mapped the physical extent of these 3 CNVs more precisely and beyond the resolution of Affy6.0. Figure 3.6 shows the regional plot of these 3 CNV events, with their breakpoints resolved through custom high density a-CGH. The status of these 3 CNVs was additionally assessed in one small Norwegian sample with affy6.0-based CNV calls as well as a Lithuanian sample (445 cases, 1140 controls) genotyped through corresponding TaqMan CNV assays. The combined study-wide Fisher's exact test P-value for deletion at 13q32.1 was 1.2×10^{-3}(OR = 2.64), the duplication at 7p22.1had a P-value of 2.7×10^{-3}(OR = 8.41) and the duplication at 8q24.3 had a P-value of 8.7×10^{-4} (OR = 4.62). Table 3.4 lists all panel-wise frequencies as well as combined P-values for the three relevant CNVs.

For Del13q32.1, fine mapping through high density a-CGH followed by Sanger sequencing identified the sequence motif 5'-GATCAC-3' at both breakpoints of the deleted segment. As the deletion is flanked by 16 highly identical *Alu* repeats, it is very likely that non-allelic homologous recombination (*NAHR*) is the underlying mechanism of the event (see also 1.2.3).

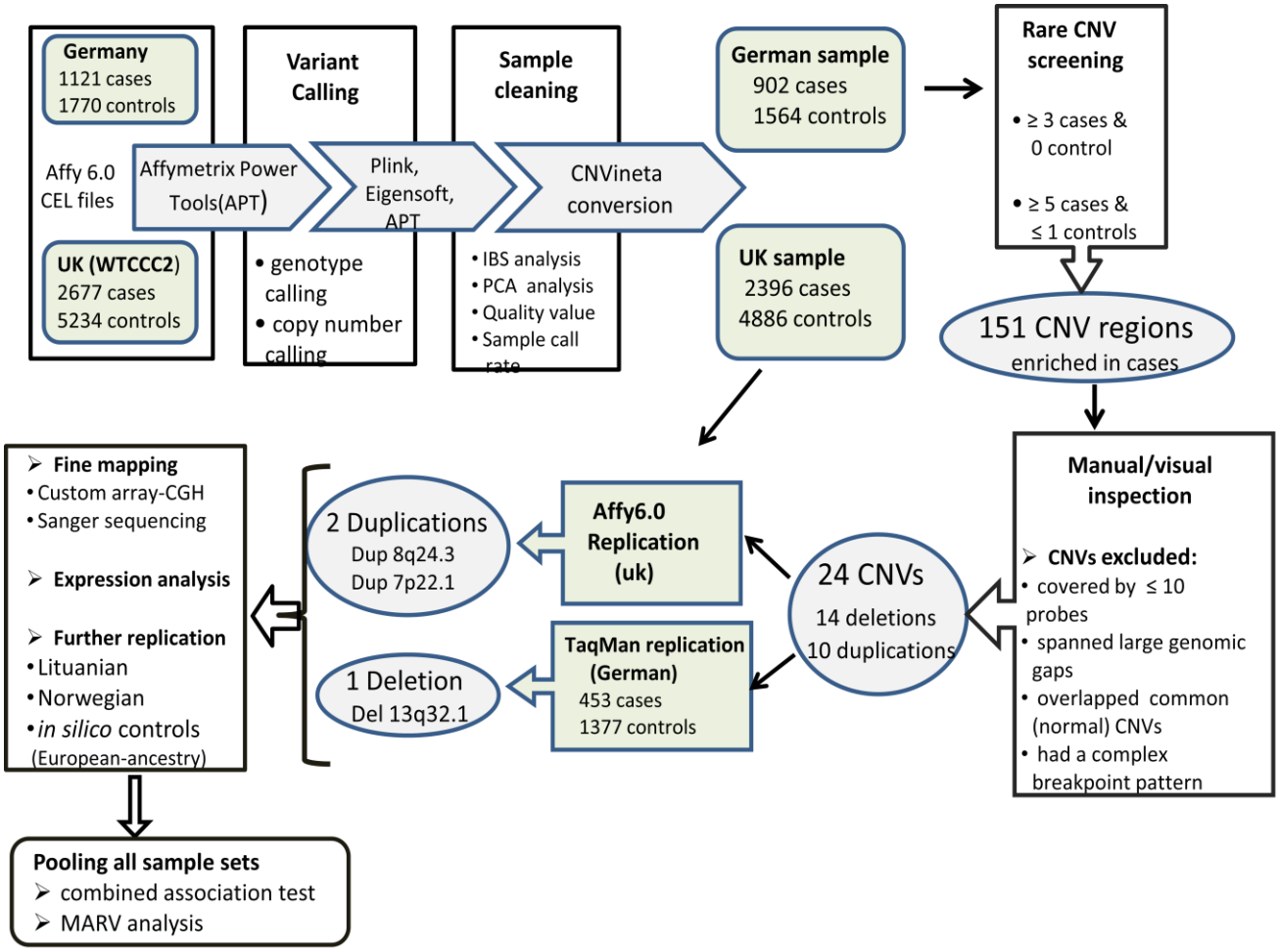


Figure 3.8 Rare CNV Analysis Workflow.

Affymetrix 6.0 data sets for the German (discovery) sample as well as WTCCC2 (UK replication) sample were processed with Affymetrix power tools (APT). Sample cleaning was based on identity by state (IBS) and principal component analysis (PCA) to exclude non-Caucasian samples as well as relatives. The remaining data sets were converted into the CNVineta format. 151 CNVs overrepresented in cases were identified after screening for rare CNVs in the German discovery sample, of which 14 deletion and 10 duplications remained after manual inspection. These 24 CNVs were further evaluated in two independent replication samples, one German and one British (WTCCC2). Dup7p22.1 and Dup8q24.3 were relevant only in UK (Affy6.0) sample, while Del13q32.1was replicated only in the German (TaqMan) sample. Fine mapping for the deletion was done by Sanger sequencing, while custom array-CGH was used for the two duplications. The status of the 3 relevant CNVs was further evaluated in a Norwegian sample (Affy6.0), a Lithuanian sample (TaqMan) and a control sample of various European individuals from previous published studies.

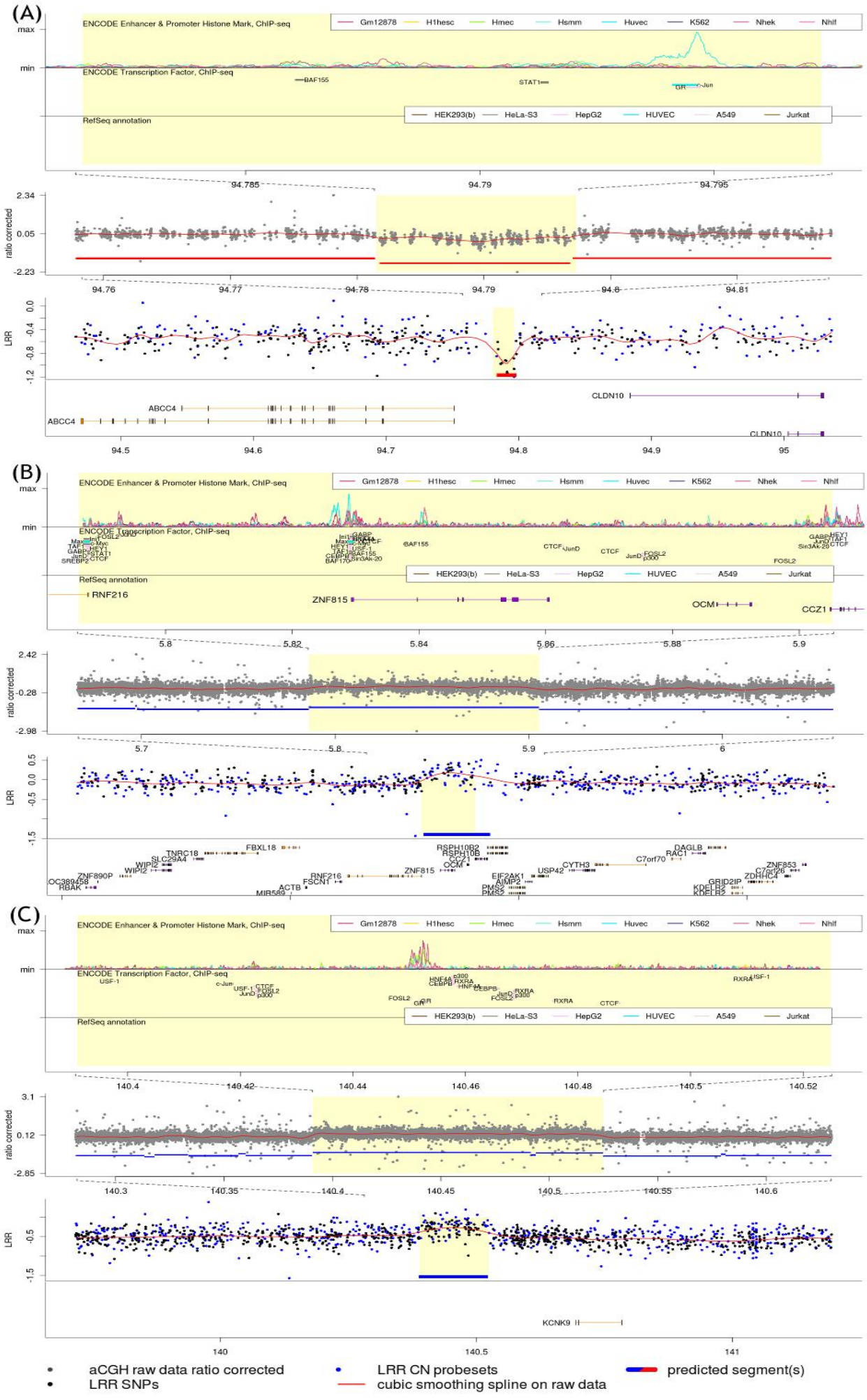
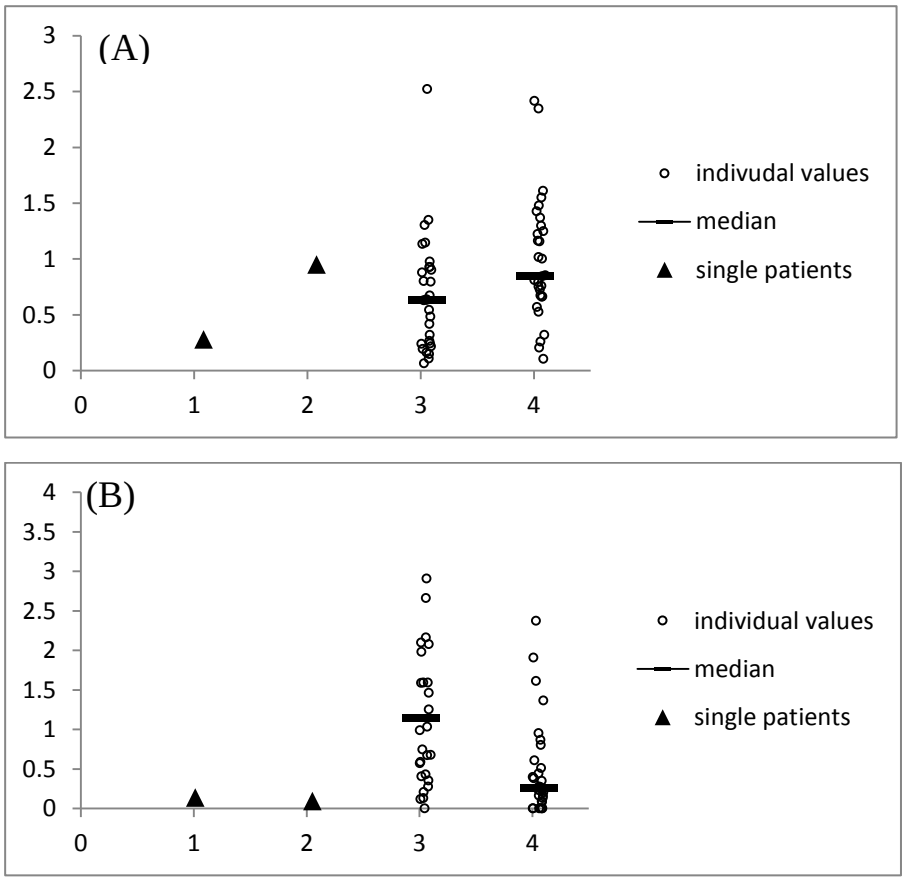

(A)
ENCODE Enhancer & Promoter Histone Mark, ChIP-seq
Gm12878 H1hesc Hmec Hsmm Huvec K562 Nhek Nhlf
ENCODE Transcription Factor, ChIP-seq
RefSeq annotation
HEK293(b) HeLa-S3 HepG2 HUVEC A549 Jurkat
ratio corrected
LRR
CLDN10
ABCC4
(B)
RNF216
ZNF815
OCM
CCZ1
(C)
KCNK9
aCGH raw data ratio corrected
LRR SNPs
LRR CN probesets
cubic smoothing spline on raw data
predicted segment(s)

Figure 3.9 Regional Plots for Del13q32.1 (A), Dup7p22.1 (B) and Dup8q24.3 (C)

For each CNV, SNP array data, array CGH (aCGH) data and breakpoints determined by Sanger sequencing (for the deletion) or array CGH (for the duplications) are illustrated. The lower panel shows the log_2 ratio of raw data from the SNP array together with gene annotation and CNV calling. The middle panel shows array CGH data (aCGH) and is zoomed, which can be seen at the x-axis captions as well as at the slanted dashed lines between the bottom and the middle panel. The top panel shows ENCODE data for transcription factor binding sites and histone marks (enhancer & promotor). There is a zoom between middle and top panel again. The SNP probe sets in the bottom panel are highlighted in black and non-polymorphic probe sets in blue. The RefSeq genes are annotated below with orange genes in reverse orientation and purple in forward orientation. The red horizontal bar represents the prediction of the deleted segment while the blue colored horizontal bar is used for duplications. This prediction was derived from Affymetrix Power Tools. For the middle panel, the predicted copy number state was performed by Nimblegen®'s in-house algorithm and is illustrated by horizontal red bars(s) again. The dots in gray represent the corrected raw data ratio, with a smoothed spline in red. The genomic region with yellow background highlights the deleted region as defined by Sanger sequencing. The upper panel shows encode data for enhancer and promotor histone marks and transcription-factor binding sites. The colors of the illustrated elements and curves correspond with adjacent legend colors for the different cell types used by the ENCODE project. **(A)** The 15.8 kb deletion at chr13:94,781,525-94,797,285 upstream of *ABCC4* and *CLDN10*. **(B)** The 119 kb duplication at chr7:5,786,323-5,905,210encompasses the entire length ofthe genes *ZNF815* and *OCM*, and partially overlaps*CCZ1* and *RNF216.* **(C)** The 134 kb large duplication at 8q24.3 (chr8:140,390,975-140,524,875) located upstream of *KCNK9*. An incidence peak (at 140,450 kb) of cis-acting regulatory elements is annotated in the genomic region affected by Duplication.

3.2.7 Expression analysis of ABCC4 and CLDN10

For interrogating the probable effect of the deletion on the expression of the two nearby genes, we examined two intestinal biopsy samples of unrelated patients carrying the deletion. Compared with the CLDN10 expression level of the deletion-depleted UC patients in inflamed mucosa, the biopsy specimen from the patient with the deletion showed very low levels of expression (Figure 3.10). Yet, this differential expression was not clear for ABCC4. Due to the sparsity of the deletion variation, no more biopsies from distinct patients harboring the deletion were available to further verify the effect of deletion on the expression of these two genes.



1 Del13q32.1 carrier, no inflammation
2 Del13q32.1 carrier, acute inflammation
3 Non-carrier, inflamed
4 Non-carrier, not inflamed

Figure 3.10 Expression analysis of ABCC4 and CLDN10 in intestinal biopsies of a UC patient panel

Real-time PCR through pre-designed TaqMan expression assays was carried out for ABCC4 (A) and CLDN10 (B) genes. Biopsy samples included one deletion carrier with acute inflammation (1), one deletion carrier without acute inflammation (2), thirty UC patients with acute inflammation and wild-type genotype (3) and thirty cases without acute inflammation and wild-type genotype (4). TaqMan® pre-designed expression assays were Hs01075312_m1 for CLDN10 and Hs00988717_m1 for ABCC4.

Locus:	**Deletion 13q32.1**						**Duplication 7p22.1**						**Duplication 8q24.3**					
Breakpoints:	chr13:94,781,525-94,797,285						chr7:5,786,323-5,905,210						chr8:140,390,975-140,524,875					
Previously reported:	DGV(Variation_49277)						DGV(Variation_53516)											
CNV size:	15.8 kb						119 kb						134 kb					
Covered or close genes:	*ABCC4, CLDN10*						*ZNF815, OCM, RNF216, RSPH10B*						*KCNK9*					
	cases		controls		*P-value*	OR	cases		controls		*P-value*	OR	cases		controls		*P-value*	OR
	wt	cc	wt	cc		(95% CI)	wt	cc	wt	cc		(95% CI)	wt	cc	wt	cc		(95% CI)
Germany (discovery) Affy6.0	897	5	1563	1	*0.027*	8.71 (0.97 - 411)	899	3	1564	0	*0.043*	Inf (0.72 - Inf)	898	4	1564	0	0.018	Inf (1.15-Inf)
Germany (replication) TaqMan	445	6	1272	2	*0.005*	8.56 (1.52-87.2)	453	0	1279	0	*> 0.05*		452	1	1377	0	*> 0.05*	
WTCCC2 (UK) Affy6.0	2391	5	4880	6	*> 0.05*		2393	3	4886	0	*0.061*		2391	5	4884	2	*0.058*	3.40 (0.66 - 21.9)
Norwegian Affy6.0	251	1	272	0	*0.36*		252	0	272	0	*> 0.05*		252	0	272	0	*> 0.05*	
Lithuanian TaqMan	442	2	1131	2	*> 0.05*		438	0	1139	1	*> 0.05*		445	0	1134	0	*> 0.05*	
in silico controls European-ancestry	-	-	4496	5	----		-	-	5787	1	-----		-	-	6724	3	-----	
Combined results	**4426** ***0.43%***	**19**	**13619** ***0.11%***	**16**	***1.2×10^{-3}***	**2.64 (1.3-5.2)**	**4435** ***0.13%***	**6**	**14928** ***0.01%***	**2**	***2.7×10^{-3}***	**8.41 (1.4-88)**	**4438** ***0.22%***	**10**	**15955** ***0.03%***	**5**	**8.7×10^{-4}**	**4.62 (1.5-15)**
Bonferroni corrected					***3.6×10^{-3}***						***8.1×10^{-3}***						***2.6×10^{-3}***	
MARV					***4.3×10^{-3}***						***6.2×10^{-3}***						***2.8×10^{-3}***	

Table 3.4 Summary of association statistics for 3 relevant rare CNVs. Frequencies are presented panel-wise and combined for the 3 relevant CNVs. **P-values** were calculated by two-sided Fisher's exact tests for CNV carriership. Odds ratios (**OR**) with 95% confidence intervals (**95% CI; inf**= infinite) are listed when P-values are smaller than 0.05. **cc** refers to **C**NV **c**arrier individuals and **wt** (wild type) to non-carriers. To account for population structure and low frequencies, the M.A.R.V. analysis method (Rivas *et al.*, 2011) was applied to the overall study sample.

4 Discussion

As it was described in the introduction, population-based evidences such as increased disease risk among relatives of IBD patients as well as higher concordance rates for IBD among MZ twins, compared with DZ ones, have implicated the genetic contribution to IBD (see 1.1.1). It was also mentioned that the genetic contribution to IBD has been mainly interrogated through GWAS studies of mostly common single nucleotide polymorphisms represented on oligo-nucleotide microarrays (SNP-array) in the last decade (see also 1.1.2). Subsequent meta analyses of these SNP-GWAS data have substantially increased the number of susceptibility loci (n > 160) for IBD, nevertheless the identified risk alleles are mostly of low to modest effects (odds ratio < 1.5) and explain less than 15% of the overall disease variance for IBD (see 1.1.2.2). It has been assumed that genomic structural variations, including CNVs are among the factors that could potentially account for the bulky missing genetic variance of complex disease phenotypes (Eichler *et al.*, 2010). To examine whether genetic alterations in the form of CNVs contribute to IBD risk, we undertook two parallel studies in this thesis; first recruiting an existing SNP-GWAS data set of UC and 4 other independent UC cohorts, we performed a multi-step genome-wide case-control analysis to interrogate the presence of disease-relevant rare copy number variants. Second, 6 IBD-discordant monozygotic twins were compared genome-wide to explore whether somatic CNVs might have contributed to IBD discordance. Here, I first discuss the results already presented for each of these two studies and their relations to each other and then describe the methodological pitfalls encountered and finally I conclude with remarks and suggestions for future studies.

4.1 Rare CNVs overrepresented in UC cases

Employing UC-GWAS data set of 1121 German UC patients and 1770 healthy controls for CNV calling, coupled with CNV data-mining tool we performed a genome-wide scan for rare CNVs associated with UC. Two main follow-up panels consisted of an independent German cohort of 451 cases and 1274 controls, in which CNVs were assayed through quantitative PCR, and a British cohort of 2396 cases versus 4886 controls with CNV genotypes based on array data. Twenty-four rare copy number variants (14 deletions and 10 duplications), overrepresented in UC patients were identified in the initial screening. Follow-up of these CNV regions in 4 independent case-control series as well as an additional public *in silico* control group (totaling 4,439 UC patients and 15,961 healthy controls) revealed 3 copy number variants, one deletion and two duplications as overrepresented (at a nominal significance) in UC patients compared to controls. Del13q32.1, as a 15.8 kb single copy loss at chr13:94,781,525-94,797,285 showed the trend of association in the discovery panel, which was reproduced in the so called TaqMan replication panel originated from Germany. However, correlation of Del13q32.1 with the disease phenotype was not observed in the WTCCC2-UC panel. The two duplications i.e. Dup7p22.1(chr7:5,786,323-5,905,210) and Dup8q24.3 (chr8:140,390,975-140,524,875) were overrepresented in UC patients of the discovery panel (compared to controls) and this trend was replicated in the WTCCC2 cohort, although no association was seen for these duplications in two independent replication panels with German and Lithuanian origins. The status of these three CNVs were further evaluated in an *in silico* data set comprising a total of 6727 unrelated control individuals of European ancestry. The scarce occurrence of the 3 variants (5 of 4505 for Del13q32.1, 1 of 5788 for Dup7p22.1 and 3 of 6727 for Dup8q24.3), observed in these samples was consistent with their low frequencies in our discovery and replication panels. It should be mentioned that CNV discovery platform used here (Affy6.0), although having a high probe density, has the limited resolution of detecting CNVs that are larger than ~15 kb. Therefore, possible smaller CNV

events (< 15 kb), most probably of disease relevance have not been examined in our study. On the other hand, in comparison to the small median size of common CNVs (~3 kb) in the human genome (McCarroll *et al.*, 2010), the three rare CNVs identified here are intermediate to large genomic alterations, involving regions with multiple genes and can potentially result in functional (deleterious) consequences leading to disease pathogenesis.

Dup7p21.1 indeed lies in a very gene rich region, encompassing either their entire lengths *(ZNF815, OCM)* or overlapping *(RNF216, RSPH10B)* partially. Dup8q24.3 is a 134 kb large duplication upstream of the gene *KCNK9* (*TASK3)*, which encodes a member of the subfamily K of the potassium channel proteins. The ever-increasing knowledge about the involvement of *TASK3 (TWIK-related acid-sensitive potassium)* channels in the pathogenesis of autoimmune inflammation (Bittner *et al.*, 2010; Meuth *et al.*, 2008) have converted them from "mere background" channels to key modulators in pathophysiological conditions.

Del13q32.1 is located 33.7 kb upstream of the gene *ABCC4* (ATP-binding cassette, sub-family C, member 4) also known as *MRP4* (multidrug resistance-associated protein4), and 78.5 kb upstream of *CLDN10* (*claudin 10*). *ABCC4/MRP4* belongs to a large family of trans-membrane proteins that play an important role in regulating cAMP-dependent signaling pathways (Sassi *et al.*, 2008) as well as human dendritic cell migration and thereby modulating immune response (van de Ven *et al.*, 2008). Mapping of *MRP* protein expression among different regions of the human intestinal tract has showed that *MRPs* are higher expressed in the colon compared to the ileum (Zimmermann *et al.*, 2005). Interestingly, one member of this protein family i.e. *ABCC1/MRP1* has been previously associated with severe UC but not with CD (Onnie *et al.*, 2006). The other neighboring gene *CLDN10*, coding for a tight junction adhesion protein is also an intriguing candidate regarding the molecular pathogenesis of UC. Tight junctions contribute essentially to the intestinal epithelial integrity. Barrier disruptions are known to be one of the main hallmarks of both phenotypes (CD and UC) of IBD and various genes

involved in epithelial barrier maintenance have been associated with IBD (Khor *et al*, 2011). Moreover, changes in expression and distribution of *Claudin* 2, 5 and 8 have been shown to result in discontinuous tight junctions and barrier dysfunction in active CD (Zeissig *et al*, 2007). For interrogating the probable effect of the deletion on the expression of the two nearby genes, we examined two intestinal biopsy samples of unrelated patients carrying the deletion. Compared with the *CLDN10* expression level of the deletion-depleted UC patients in inflamed mucosa, the biopsy specimen from the patient with the deletion showed very low levels of expression. Yet, this differential expression was not clear for *ABCC4.* Due to the sparsity of the deletion variation, no more biopsies from distinct patients harboring the deletion were available to further verify the effect of deletion on the expression of these two genes. However, presence of cis-acting regulatory elements such as transcription factor binding sites, showed in ENCODE annotations of the deleted region might be an explanation of the distinct *CLDN10* expression that we observed here.

Overall we find that the rare CNV candidates of this study, verified by visual inspection of the underlying raw data, are true positive CNVs. All three relevant CNVs could technically be validated by independent methods and were followed-up in independent sample sets. In contrast to common variants, disease correlation of rare variants is difficult to be assessed through classical association statistics. Low frequency of these variants impedes to detect associations at the genome-wide level significance by modest or intermediate sample sizes. Power limitations may even increase when these variants do not have high penetrance. In this study, the trend of association with UC was present for each of the three mentioned copy number variants in the discovery sample as well as in at least one replication panel, nevertheless higher statistical power, provided by larger case control samples are needed to confidently evaluate the disease risk of these variants.

4.2 No confirmed CNVs in MZ twins discordant for IBD

As described earlier, discordance with regard to IBD in MZ twins has been well documented (see 1.3.3). It was also mentioned that MZ twins, despite deriving from a single zygote might not actually be genetically identical and could vary in their genomes due to post-zygotic mutational events. CNVs, as one of the main forms of genetic alterations can potentially contribute to these genomic differences (see also 1.3.2). To explore probable IBD-relevant somatic CNVs, genomic DNA from peripheral blood as well as bowl biopsy specimens from 6 IBD patients (3 UC and 3 CD) and their healthy MZ co-twins were screened for copy number differences by means of array-comparative genomic hybridization (array-CGH) followed by quantitative PCR. This survey however did not reveal any validated CNVs within the blood-derived genomic DNA of IBD-discordant MZ twins. Yet two probable confounding issues should be considered here; First, as blood tissue contains high levels of hematopoietic cell lineage, twin blood exchange could lead to lack of ability to detect probable genomic differences in blood-derived DNA of MZ twins (Erlich, 2011). This blood exchange is the result of shared blood circulation in monochorionic twins during embryogenesis (see also Figure 1.12). This instead results in mixed hematopoietic context, where stem cells from one twin are engrafted in the co-twin and vice versa (Greaves *et al*, 2003). About 70% of all MZ twin embryos are monochorionic and experience this blood exchange whereas the dichorionic ones do not (Machin *et al*, 2009). Unfortunately we didn´t have any records of mono- or di-chorionic state of the twin pairs recruited in this study. Second issue is that any postzygotic genomic alteration is likely to be confined to specific tissues and certain cell lineages depending on the timing of the mutational event, resulting in tissue- and (or) organ-specific pathogenesis when mutations are deleterious. In order to somewhat address these two issues, we also analyzed genomic DNA derived from bowel biopsy of the IBD-affected twin individuals against blood DNA of their healthy co-twins through array-CGH, but no copy

number differences could be confirmed as well. It should also be noticed that genomic DNA obtained from bowel biopsies consists of a population of DNA molecules from different cell types and might not high proportionally represent the epithelial cells mainly involved in the pathophysiology of IBD. Therefore, further CNV profiling experiments, assisted with fine histo-pathological cell selection through laser-capture microdissection (Funke *et al.*, 2011) of intestinal epithelial cells from inflamed mucosa specimens might provide more accurate insight in to the landscape of somatic genomics with relevance to IBD. Generally these aspects make identification of probable somatic (and likely disease-relevant) copy number alterations between siblings of MZ twins much more challenging than scoring inter-individual germline CNVs. The latter was performed in our twin study by genomic comparison of the sample NA15510 and one non-IBD individual of a twin pair, mainly for the assessment of our platform performance (see 3.1.3).

The rate of *de novo* somatic CNV formations is not known well, essentially due to the complications of assaying this type of genomic alterations. For *de novo* germline CNVs, however, there are estimations of locus-specific mutation rates ranging from 1.6×10^{-6} to 1.2×10^{-4} per locus per generation, which are 100 to 10,000 times greater than that for SNPs (Lupski, 2007; Itsara *et al.*, 2010). These high mutation rates have been determined for genomic hot spot regions, which contain large blocks of repetitive sequences and are highly susceptible to structural rearrangements mediated by non-allelic homologous recombination (*NAHR*) (see also 1.1.2). There exist evidences that *NAHR* can also occur in mitosis, which results in mosaic populations of somatic cells carrying genomic rearrangements (Gu *et al.*, 2008; Dumanski *et al.*, 2008). Indeed it has been estimated that *de novo* CNVs may occur at a rate of 10% per twinning event (Bruder *et al.*, 2008). Despite this, few studies examining copy number discordance in twins have been reported in disease phenotypes. Our study was an effort to see whether discordance for IBD in MZ twins could be explained, even partially, by copy number differences in their genomic DNA. In 2008 Dumanski and colleagues analyzed blood DNA of

nine pairs of selected MZ twins that were discordant for neurodegenerative phenotypes related to Parkinsonism (Bruder *et al.*, 2008). They reported many loci that were suggestive of putative and mostly large (> 100 kb) CNVs. Their findings proposed that mosaic state with regard to copy number genomic rearrangements is a possible explanation for twin discordance. Noteworthy among their cases, one of the twin individuals had been diagnosed with chronic lymphocytic leukemia. Their finding of genomic CNVs in blood DNA at this case compared to its co-twin has been lacking novelty, since it is rather well recognized that many tumors especially in advanced stages are associated with high genomic instability, resulting in extensive somatic genomic rearrangements (Darai-Ramqvist *et al.*, 2008; Stratton *et al.*, 2009). In another study Baranzini and colleagues examined three pair of monozygotic twins discordant for multiple sclerosis (MS) as a common autoimmune and chronic inflammatory disorder, which has a concordance rate of approximately 30% among MZ twins (Willer *et al.*, 2003). They reported no consistent difference in genomic DNA sequence and genomic DNA methylation in $CD4^+$ lymphocytes between affected and unaffected twins (Baranzini *et al.*, 2010). Further it has failed to identify somatic CNVs as of etiological significance in twin pairs discordant for schizophrenia (Shinji *et al.*, 2010; Lyu *et al.*, 2016).

It should also be emphasized that the applied array-CGH platform in our twin CNV analysis, despite having high probe density, yet provide the capability of detecting only imbalanced structural variations in the form of gain and loss of genomic DNA segments larger than 5 kb. Hence quantitatively balanced genomic alterations like inversions and translocations (see figure 1.8) as well as small CNVs beyond the resolution of our experimental approach have not been explored here and cannot be ruled out.

4.3 Methodological pitfalls

4.3.1 The problem of wave artifacts in twin array-CGH analysis

The results of array-based analysis are graphically represented as scatter plots, in which each spot represents a signal with regard to a genomic position. In the case of array-CGH, results are plotted as the $\log_2$ ratios of the test and reference sample hybridization signals along the genomic coordinates (see 1.2.1). In the absence of CNVs, ratio or signal values form a flat baseline. Deviations from the flat baseline indicate CNVs as genomic insertions (duplications) or deletions. However technical drawbacks, especially high rates of artificial segmentations in signal intensities interfere with the analysis and interpretation of array-based data and have been described by other groups (Marioni *et al.*, 2007; Diskin SJ 2008; Conrad *et al.*, 2010). In our twin CNV analysis, we also experienced an over-segmentation in the array-CGH signal intensity data especially in biopsy samples. This resulted in high rates of false positive CNVs that were not confirmed by the independent quantitative PCR used for validations of array-predicted CNVs. These over-segmentation patterns in $\log_2$ ratio profiles of genomic DNA samples are called wave artifacts, as they appear as oscillating increases and decreases of the hybridization signals regardless of the presence or absence of CNVs. It has been reported that these wave artifacts are correlated with GC content of the targeted genome, so that regions with a low GC content corresponded roughly to peaks of the waves, while regions with high GC probe content corresponded to troughs (Marioni *et al.*, 2007). Some CNV detection algorithms have been developed to correct for the effects of these waves based on the GC content of the probes (Komura *et al.*, 2006; Song *et al.*, 2007) or the GC contents and sizes of genomic DNA fragments (Komura D 2006, Conrad *et al.*, 2010) but none of them yield optimum results. Ylstra and colleagues presented a wave-smoothing method, which is independent from GC content and employs log ratios from samples known to have no CNVs, using multi-variable regression (van de Wiel *et al.*, 2009). However, this method does not

account for the confounding effects from large segmental shifts. It has also been warned that available methods for removing wave artifacts can themselves introduce new artifacts due to skewed regression or even remove signals created by true CNVs due to blind smoothing (Lepretre *et al*, 2010). We considered this warning seriously and did not use any smoothing algorithm to ride of the wave artifacts, which we experienced in our twin CNV analysis. Instead we tried to detect false positives CNVs by verification through quantitative PCR. As false positive CNVs in our experiment were mainly predicted from biopsy samples and not the blood samples, we think that not the CG content but the poor quality (mainly high rate of fragmentation) of genomic DNA form biopsies has been the main source of artifacts in our twin CNV analysis. These difficulties in array-CGH data analysis make even more challenges when looking for somatic CNVs, as the post-zygotic genomic CNV events, if any, might occur in only a limited number of cells and are therefore not reflected by sharp aberrations in signal intensity ratios in array-CGH. We therefore recommend application of at least two replicates for each array design of this kind as suggested elsewhere (Conrad *et al*, 2010), followed by necessarily experimental validation of shared potential CNV signals.

4.3.2 Deficient probe coverage and biases in Multiallelic CNV genotyping

We used high resolution Affymetrix 6.0 SNP microarray in our UC case-control CNV analysis. Although this platform incorporate better SNP and non-polymorphic copy number probe selection criteria for complex regions of the genome, but still tends to have a paucity of probes in duplication-rich regions of the genome and therefore is unable to capture a large number of known CNPs (Kid *et al*, 2008; Alkan *et al*, 2011). It has been shown that around half of simple deletion variants are not well captured by even the highest-density SNP arrays and this deficiency increases when more complex multicopy variants within segmental duplication-rich regions are considered (Cooper 2008).

For CNPs with higher copy numbers often located in repeat-rich and duplicated regions, it is difficult or impossible to assign integer copy numbers from microarray hybridization data. The reason is that array-CGH and SNP platforms assume each location to be diploid in the reference genome but this is not valid for duplicated sequences. For example the signal for a 5 to 4 copy ratio or other complex patterns will not fit the expected results for a diploid reference sequence and may drop below the assay's sensitivity to discriminate signals (Alkan *et al*, 2011). Given the high frequency of CNPs in segmental duplications as mentioned previously (see 1.2.2), it will be important to test these variants for association to disease. One famous example of these complex regions with implications in inflammatory diseases is *ß-defensin* locus. *Beta-defensins* are small secreted antimicrobial peptides, which are encoded by *DEFB* genes in three main gene clusters, two on chromosome 20 and one on 8p23.1(Ganz, 2003). The *beta-defensin* genes on 8p23.1 are on a large repeat unit that is variable in copy number, for which individuals have between 2 and 12 copies per diploid genome (Hollox *et al*,2003). However it is impossible or difficult to assign integer copy numbers from microarray hybridization data and this poses particular challenges for case–control association studies (McCarroll 2008). The influence of genotyping error for *ß-defensin* copy number on the outcome of different studies is evident in reports of discordant findings. In 2006 there was a report of an association between low *beta-defensin* copy number and CD of the colon (Fellermann *et al*, 2006). In contrast, a later study reported an association between CD and higher beta-defensin copy numbers, using real-time PCR measurement in 466 cases and 329 controls (Bentley *et al*, 2010). Schalkwijk and colleagues genotyped *beta-defensin* copy number for two cohorts of psoriasis patients and controls from the Netherlands and Germany (Hollox *et al*, 2008). They used two different methods, namely multiplex amplifiable probe hybridization assay (MAPH) and Paralogue Ratio Test (PRT) on the same set of samples. Unlike CD, increasing copy number was shown to be associated with susceptibility to

psoriasis but the association signal was much stronger in MAPH ($P = 1.65 \times 10^{-6}$) than in PRT ($P = 0.01$). These inconsistencies reflect differential genotyping biases in assaying copy number of multi-allelic complex CNV regions.

It should additionally be mentioned that SNP microarray and array-CGH platforms are designed relative to the reference human genome sequence (International Human Genome Sequencing Consortium, 2004). However, other *de novo* whole genome sequencing efforts have led to the identification of sequence insertions, which are not present in the reference genome assembly (Levy *et al* 2007; Wang J 2008, The 1000 Genomes Project Consortium, 2012). In fact, many of these novel insertions are polymorphic in human populations and thus represent genetic variants that will be missed by using microarrays designed to the reference genome assembly.

4.3.3 Low consistency of CNV detection algorithms

Different computer programs are available to detect CNVs, using the intensity of the hybridization of sample DNA to the array probes. The underlying detection algorithms are generally based on either Hidden Markov model (Korn *et al*, 2008) or circular binary segmentation (Olshen *et al*, 2004). PennCNV and QuantiSNP were first developed based on an HMM-based algorithm for an Illumina platform (Colella *et al*, 2007; Wang *et al*, 2007) and then later were modified to be compatible with Affymetrix platforms as well. Birdseye, another HMM-based approach was developed to detect CNVs in SNP genotyping arrays specifically Affymetrix platforms (Korn *et al*, 2008). Many studies have found considerable variation among the outputs and false call rates for CNVs, comparing different CNV detection programs (Baross *et al* 2007; Winchester *et al*, 2009). Table 4.1 shows the different number of CNVs identified with distinct algorithms and platforms for a single HapMap sample (NA10861). The average number of CNVs, called per individual by different programs varies also with regard to the size of CNVs. Affy6.0 data in 1001 bipolar cases and 1033 controls of

European ancestry have been used to assess CNV detection accuracy of currently used CNV detection softwares (Zhang *et al.*, 2009). For CNVs larger than 100 kb, PennCNV called an average of 8, while HelixTree called an average of 27. The differences declined when the size of CNVs increased in the affy6.0 platform (Zhang *et al.*, 2009).

These evidences led us to the conclusion that without independent experimental genotyping, software-called CNVs based on array data are not reliable. Therefore we used quantitative PCR to independently verify all the CNVs predicted from array-CGH data in our twin CNV analysis as well as Affy6.0 data in the rare CNV analysis for UC. To further compensate for the poor consistency among different CNV programs, we used two distinct programs, namely quantiSNP and Affymetrix power tool for CNV detection. We then followed only those variants that were predicted by both programs (with at least 40% overlapping), as this approach has been suggested by others as well (Pinto *et al.*, 2011).

Algorithm	Platform and array	Number of CN events detected
Birdsuite 1.5.5 (Canary & Birdseye)	Affymetrix 6.0	137
CNAT (Genome Console 3.0.2)	Affymetrix 6.0	10
CNVPartition 1.2.1	Illumina 1M Duo	16
GADA (R 0.7-5)	Affymetrix 6.0	613
GADA (R 0.7-5)	Illumina 1M Duo	87
Nexus Biodiscovery 4.0.1	Affymetrix 6.0	111
Nexus Biodiscovery 4.0.1	Illumina 1M Duo	8
PennCNV (2009Jan06)	Affymetrix 6.0	67
PennCNV (2009Jan06)	Illumina 1M Duo	43
QuantiSNP v2.0	Affymetrix 6.0	193
QuantiSNP v1.1	Illumina 1M Duo	60

Table 4.1 Comparison of CNV numbers detected for a single HapMap sample (NA10861) with different SNP-arrays and algorithms. According to Winchester *et al.*,2009

4.4 Probable sources of discordance in IBD MZ twins

No finding of CNV differences in our examined twins might be in favor of the view that mainly non-genetic factors contribute to phenotypic discordance in MZ twins. Not high concordance rates in MZ twins of IBD (~34% for CD and only ~16% in UC) as well as more frequent IBD records in the first-born sibling of twin pairs have been described (Spehlmann *et al*, 2008) and attributed to pre- and post-natal environmental trigger factors. Here I describe the relevant non-genetic factors that could potentially contribute to MZ twin discordance and also point to evidences and proposed mechanisms of their involvement in IBD pathogenesis, where applicable.

4.4.1 Epigenetic factors

Epigenetics is generally referred to mitotically and/or meiotically heritable changes in gene function that cannot be explained by changes in DNA sequence (Russo *et al*, 1996). The most typical examples of such changes are DNA methylation and histone modifications, both of which serve to regulate gene expression. More recently other components such as ATP-based chromatin remodeling complexes (Vignali *et al*, 2000), non-coding RNAs (Esteller., 2011) and prions (Halfmann *et al*, 2010) have been recognized and increased the complexity of epigenetic regulation mechanisms. Among these factors, genomic DNA methylation of cytosine bases (5-methylcytosine), especially in the context of *CpG* islands has been studied more extensively. *CpG* islands are cytosine-phosphate-guanine dinucleotide rich sequences (mostly around 1 kb) and are found in the transcription start sites and promoter regions of more than half of the genes in the human genome (Jones, 2012). Microarray- and sequencing-based approaches for genome-wide mapping of 5-methylcytosine (methylome) through bisulphate-treated DNA have emphasized the correlation between methylation position and its effect on transcription (Lister *et al*, 2009; Heyn *et al*, 2012). Methylation at the 5' transcription start sites mostly blocks initiation of transcription, whereas methylation in the gene body does not block

transcription and might stimulate transcription elongation and even affect splicing (Jones, 2012). It has been proposed that transcription down-regulation due to methylation is the result of either the inability of specific transcription factors to bind methylated CpGs or the recruitment of methyl-CpG-binding proteins (such as MeCP) with transcription repression activity (Nan *et al.*, 1997; Castillo-Fernandez *et al.*, 2014)

The first comprehensive survey of epigenetic differences in MZ twins was carried out through comparing total content of methyl-cytosines and histone acetylation levels of peripheral lymphocytes (Fraga *et al.*, 2005). In that study, discordances within MZ twin pairs were found for X-inactivation, DNA methylation and histone acetylation, which were correlated with the age of twins and the amount of time they lived together. This observation led to the expected conclusion that a bulk of epigenetic differences accumulate during the life time of MZ twins (Fraga *et al.*, 2005). In 2009, Petronis and colleagues confirmed the presence of DNA methylation differences between co-twins across three different tissues, namely white blood cells, buccal epithelial cells and gut biopsies (Kaminsky *et al.* 2009). They observed greater similarity between DNA methylation profiles within MZ twins than within DZ ones. Furthermore, they suggested a functional stratification of the epigenome, based on the finding of fewer differences between co-twins in CpG islands and promoter regions than in all non-CpG island loci. Extensive region-specific variability in DNA methylation between MZ twins was recognized from profiling of the major histocompatibility complex (MHC) in $CD4^+$ lymphocytes of 49 Norwegian MZ twins (Gervin *et al.*, 2011). In that study, CpG islands, 5' untranslated regions and conserved non-coding regions showed less variability than CpG-poor regions within twins. In contrast to the study of Fraga *et al.* which suggested that epigenetic divergence of MZ twins occur mainly within their life time, differences in neonatal epigenome were observed in a longitudinal study that profiled genome-wide DNA methylation of 10 MZ and 5 DZ twin pairs at birth and at 18 months (Martino *et al.*, 2013). In

their study, epigenetic differences between MZ co-twins appeared to arise early in life (during embryonic development) and surprisingly tended to either diverge or converge in a twin-pair-specific manner.

4.4.2 Gut microbiota

Microbiota of the gut describes the collection of more than 100 trillion microorganisms, mostly bacteria, which colonize the oral–gastrointestinal tract, with highest concentration (10^{11} or 10^{12} cells/g of luminal contents) in the colon (Dave *et al.*, 2010). It is believed that millions of years of co-evolution between the host and these microorganisms have led to a symbiotic relationship, in which the microbiota contributes to many host physiological processes and the host, in turn, provides niches and nutrients for microbial survival (Hooper *et al.*, 2010). Gut homeostasis is provided by the complex interplay between the host immune system and the microbiota as these resident microbes digest substrates inaccessible to host enzymes, educate the immune system and repress the growth of harmful microorganisms (O'Hara *et al.*, 2006). Recent evidences suggest that human intestinal microbiota is seeded even before birth (Rodríguez *et al.*, 2015). Maternal microbiota forms the first microbial profile and prenatal factors such as mode of delivery, diet, genetics and intestinal mucin glycosylation influence microbial colonization (Palmer *et al.*, 2007; Funkhouser *et al.*, 2013). After birth, the microbial diversity increases and shapes an adult-like microbial profile by the end of the first 3–5 years of life (see figure 4.1). After that, the composition of the gut microbiota is relatively stable throughout the life, but may alter as a result of bacterial infections, antibiotic treatment, lifestyle, surgical interventions and long-term changes in diet (Marques *et al.*, 2010; Claesson *et al.*, 2011). A shift in the composition of the gut microbial community from a symbiotic commensal flora to a potential deleterious and pathogenic profile is called dysbiosis and has been correlated with the development of IBD (Kaur *et al.*, 2011). The effect of host genetics versus environment in shaping the gut microbiota has been investigated by analysis of

individuals with varying degree of relatedness. A study with children under the age of 10 indicated that the degree of similarity in bacterial community was higher in MZ twins compared with DZ ones and was lowest in the unrelated control group (Stewart *et al*, 2005). Recently more comprehensive profiling and functional dissection of the gut microbiota between individuals and between healthy and diseased states have become possible, thank to the high throughput metagenomic analysis using next-generation sequencing technologies (Qin *et al*, 2010; Human Microbiome Project Consortium, 2012). A metagenomic study using deep sequencing of samples from 31 monozygotic and 23 dizygotic twin pairs and their parents, however, did not identify significant differences in bacterial diversity between the

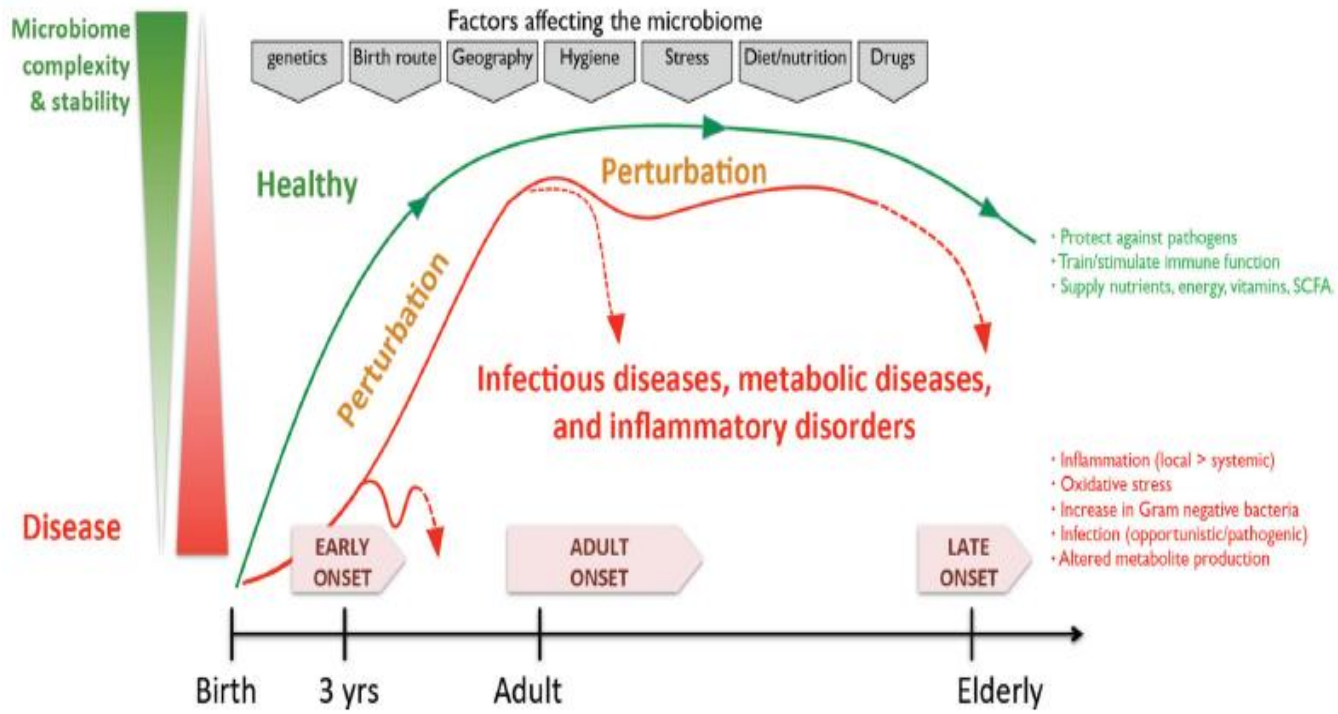


Figure 4.1 Factors affecting the stability and complexity of the gut microbiome in health and disease

Many factors are indicated to impact the microbiome including genetics, diet, medication, among others (marked in the grey boxes at the top of the figure). Some of these factors can introduce perturbations affecting the complexity and stability of the microbiome, leading to microbial dysbiosis. Features of an imbalanced microbiome include, for example, an increase in Gram-negative bacteria linked to an environment of oxidative stress and inflammation, and metabolite production. The first 3 years of life represents the most critical period for dietary interventions aimed at microbiota modulation to improve child growth and development and positively affect health. Illustration from Kostic *et al.*, 2014

two types of twins, although the members of the same family were found to share a higher numbers of bacterial phylotypes (Turnbaugh *et al*, 2010). Several studies have observed a decreased diversity in gut microbiota of UC patients, compared with healthy controls at both mucosal (Nishikawa *et al*, 2009; Ott *et al*, 2008) and fecal (Martinez *et al*, 2008) samples. This reduced diversity of the gut microbiome was also demonstrated in IBD patients within MZ twin pairs discordant for CD (Dicksved *et al*, 2008). Decreased microbial diversity has even been observed within the same patient in inflamed versus non-inflamed tissues, so that CD patients had lower overall bacterial loads at inflamed regions (Sepehri *et al*, 2007). In 2011 Lepage and colleagues used 16S ribosomal DNA libraries to determine Microbiom profiles of sigmoid colon biopsies from 8 MZ twins discordant for UC, 11 healthy DZ twins as well as 7 healthy MZ twins (Lepage *et al*, 2011). They found that Patients with UC had dysbiotic microbiota, characterized by less bacterial diversity and more Actinobacteria and Proteobacteria than that of their healthy siblings. Further they observed that healthy members from discordant twins had more bacteria from the Lachnospiraceae and Ruminococcaceae families, when compared to their diseased sibling but this was not the case within healthy twins (Lepage *et al*, 2011).

4.4.3 Other environmental factors

As described earlier, there exist epidemiological evidences of a putative environmental component in the risk of common idiopathic IBD and especially UC (see 1.1.2). As a result, it has been assumed that differentiated environmental exposures of twin individuals could also contribute to their discordance with regard to IBD. Although these environmental triggers are not fully characterized, however potential relevant influences span the spectrum of life, from mode of childbirth and early-life exposures (such as breastfeeding and antibiotic exposure in infancy) to exposures later on in adulthood including smoking, major life stressors, diet and lifestyle (Ananthakrishnan, 2015). In 2012 Schreiber and colleagues profiled environmental

risk factors in a randomly selected population of 512 German MZ and DZ twins, in which at least one sibling had IBD (Spehlmann *et al.*, 2012). In their study the most significant factors associated with CD or UC included high frequency of antibiotic use, high consumption of processed meat, recall of bacterial gastrointestinal infections and living abroad before the time of diagnosis.

Here I briefly mention environmental risk factors, which have been implicated so far, as they could also contribute to IBD discordance in MZ twins.

Diet

An inverse association with dietary fiber, particularly fruits and vegetables has been observed in IBD (Amre *et al.*, 2007). Pediatric patients with CD had markedly lower intake of fruits and vegetables than controls without IBD (Amre *et al.*, 2007). Further, it has been shown that long-term fiber intake reduces the risk of CD (Ananthakrishnan *et al.*, 2013). A potential mechanism is that soluble fibre (from fruits and vegetables) is metabolized by the intestinal bacteria to short-chain fatty acids that inhibit transcription of pro-inflammatory cytokines (Galvez *et al.*, 2005). Dietary fat particularly saturated fats, Vitamin D, Zinc and iron might also have a role in pathogenesis and course of IBD (Ananthakrishnan, 2015).

Smoking

Increase in risk associated with smoking has been seen in CD, while in UC increased risk has been correlated with former smoking (and not current smoking), demonstrating an inverse association between UC and smoking (Harries *et al.*, 1982; Mahid *et al.*, 2006). Consistent with the direction of effect, smoking is associated with a more aggressive disease course in CD. Smokers are more likely to need immunosuppression and surgery and have greater probability of recurrence after ileocecal (section between small and large intestine) resection (Lakatos *et al.*, 2007). By contrast, in UC smoking cessation is frequently a trigger for disease flares within a year of cessation and current smokers have a milder disease state with less need for strong

immunosuppression or surgery interventions (Cosnes, 2004). It has been proposed that Smoking could alter smooth muscle tone and influence endothelial function through nitric oxide production (Hatoum *et al*, 2006), or affect the integrity of the gut mucous barrier (McGilligan *et al*, 2007) or result in less production against oxidative stress (Bergeron, *et al*, 2012). Smoking cessation is associated with an early change in the microbiome, an effect that may underlie the influence of smoking cessation on UC (Biedermann *et al*, 2014).

Hygiene

The hygiene hypothesis was first proposed to explain the increase in the incidence of autoimmune diseases in the industrialized world (Strachan *et al* 1989). Many studies supporting this hypothesis have shown that factors such as number of siblings, larger family size, drinking unpasteurized milk, living on a farm and exposures to pets particularly early in childhood are inversely associated with risk of IBD (Bernstein *et al*, 2006; Radon *et al*, 2007). These correlations are generally explained in the context of influence of such factors on the parallel development of gut microbiota and maturation of the immune system early on in life (see also 4.4.2)

Lifestyle

IBD has long been associated with psychosocial stressors as well as certain personality types including neuroticism, obsessive–compulsive behavior, dependency and perfectionism (Bernstein *et al*, 2010). Neuroticism is a personality trait characterized by anxiety, fear, moodiness, worry, envy, frustration, jealousy, and loneliness (Thompson, 2008). Obsessive–compulsive behaviour is a mental characteristic, in which people feel the need to check things repeatedly, perform certain routines or have certain thoughts repeatedly (Fenske *et al*, 2009). Stress can influence intestinal inflammation via the hypothalamus–pituitary–adrenal axis and (or) the autonomic nervous system, resulting in increased production of proinflammatory cytokines, activation of macrophages and alteration of intestinal permeability and gut

microbiota composition (Bonaz *et al*., 2013). The association between physical activity and IBD was established, as sitting occupations like administration and office work were shown to increase risk of IBD, whereas heavy manual labour (including building and construction, cleaning and maintenance) associated with a low IBD risk (Sonnenberg *et al*., 1990). Supporting this hypothesis, a prospective cohort study demonstrated that rigorous physical activity reduced the risk of CD by 44% (Khalili *et al*. 2013). Finally disturbed sleep quality is common in society, but is more frequent in patients with IBD and is associated with active disease (Swanson *et al*., 2011)

Infections and Antibiotics

The influence of antibiotics has been shown to be more relevant during early childhood, when the gut microbiota is rather unstable and its perturbation could influence the gut immune response and thereby alter the susceptibility to IBD (Penders *et al*., 2006). In a nested case-control analysis, 58% of pediatric patients with IBD had received an antibiotic in their first year of life, compared with only 39% of the controls (Shaw *et al*., 2010). The association of pediatric IBD with antibiotic use is greater for CD than UC and is stronger for exposure in the first year of life as well as multiple courses of antibiotic use (Kronman *et al*., 2012).

4.5 Missing genetic variance (heritability) and CNVs

The main goal of the CNV analysis in MZ twins as well as case-control samples in this thesis was to explore whether CNVs, either somatic or germline (inherited), contribute to the risk of IBD. We hypothesized that potential findings of pathogenic CNVs could account, although partially, for the missing genetic variance assumed for IBD. The issue of missing variants for IBD and other common disease phenotypes has mainly arisen by the end of the last decade. During the past 15 years, genome-wide association studies (GWAs) have identified more than 2,200 robust associations with more than 300 complex diseases and traits, which are catalogued in https://www.genome.gov/gwastudies. It has been however observed that most of the identified common associated variants have small effect sizes and confer relatively small increases (1.1-1.5 fold) in disease risk and therefore explain only a minority of the estimated heritability (Zuk *et al*, 2012). Heritability refers to the total phenotypic variation of a trait that can be attributed to genetic effects (Falconer *et al*, 1995). There are only few disease phenotypes such as age-related macular degeneration, in which identified associated loci exert large effects in disease risk and have explained more than 50% of the estimated heritability (Maller *et al*, 2006). This is however not the case in IBD and most other common diseases. In IBD the median odds ratio (OR) of more than 160 associated risk loci, identified so far, is ~ 1.1 and these loci account for only 13.6% and 7.5% of the total disease variance for CD and UC respectively (see also 1.1.2.2).

4.5.1 Inflated (Phantom) heritability for IBD?

The proportion of heritability explained by a set of variants is the ratio of the heritability due to these variants (numerator), estimated directly from their observed effects, to the total heritability (denominator) inferred from population data (Visscher *et al*, 2008). As already mentioned, small odds ratio of the GWAs-indentified variants for IBD (and many other common diseases) has resulted to obtain small numbers for numerator (compared to

denominator). Generally the predominant view is that the compensation for missing heritability lies in the numerator, i.e., there are missing variants underlying the susceptibility of the disease, which remain to be discovered. Accordingly many debates and efforts have been raised on how and where this hidden variability is to be hunted (Manolio *et al.*, 2009; Eichler *et al.*, 2010). In contrast, it has been argued that although many genetic risk variants remain to be found, but a significant fraction of the missing heritability is probably due to overestimation of disease heritability itself (denominator), an effect referred to as phantom (inflated) heritability (Zuk *et al.*, 2012).

There are different approaches to estimating the heritability of diseases, with each having their own potential source of biases in sampling strategies and analysis methods, leading to heterogeneous and even confounding results (Tenesa *et al.*, 2013). This ambiguity is classically highlighted in the case of type 2 diabetes, in which heritability estimations using first-, second- and third-degree relatives have resulted in measures ranging from 0.19 to 1.02 (Smith *et al.*, 1972). Two approaches have been mostly used for quantifying heritability in the case of IBD and many other diseases. First approach measures the correlation in liability between relatives, through scoring the disease incidence among the relatives of affected individuals against incidence in the general population (Falconer *et al.*, 1995). Second method, compares the resemblance among MZ twins versus DZ ones (Boomsma *et al.*, 2002). Familial aggregation studies have indicated that 2% to 14% of patients have a family history of CD (Halme *et al.*, 2006), while estimates of the sibling relative risk ratio (the ratio of disease risk among siblings of patients to the risk in the general population) ranged from 15 to 42 (Halme *et al.*, 2006). The large variation in these estimations points to the challenges encountered in obtaining accurate heritability measures for IBD. Confounders also include inconsistent study design (e.g. only counting first degree relatives rather than all relatives), sample selection bias (e.g. using hospital cases that are likely to have a more severe form of the disease than the general IBD

population) and variation in disease prevalence rates, both between different populations and over time (Liu *et al.*, 2014; Ananthakrishnan, 2015). Moreover, it has been almost impossible to distinguish the influence of shared environment from common inheritance in familial clustering of IBD, an ambiguity that could also lead to inflated heritability measures from population-based familial data.

It should be further mentioned that estimations of total heritability classically assume that the disease involves no genetic interactions among loci (Visscher *et al.*, 2008). Indeed, under the model that the disease arise from a strictly additive genetic architecture (no interaction assumed), the identified loci for CD explain only 15% of the estimated heritability (Jostins *et al.*, 2012). However it is now appreciated that additive models for disease risk in IBD are not justified, as many interacting molecular and cellular pathways have been shown to contribute to its pathogenesis (see also 1.1.2.2). Under non-additive models (models counting for various interacting pathways), the total heritability may be much smaller (non-inflated) and thus the proportion of heritability explained much larger (Tenesa *et al.*, 2013). It has been argued that phantom heritability might be around 60% for CD, so that of the current missing heritability, 80% is estimated to be due to unaccounted genetic interactions among different pathways involved in IBD (Zuk *et al.*, 2012).

4.5.2 Where (how) to find the probable missing variants for IBD

It has been emphasized that nearly all IBD-associated variants with frequency greater than 5% and OR greater than 1.2 in individuals of European ancestry have been already identified and the remaining genetic contribution will arise from a combination of common variants with smaller effect sizes and not common variants with large effects (Liu *et al.*, 2014). However, all variants with large effects (OR > 3) and frequency greater than 1% (low frequency variants) have also been uncovered by GWAs and linkage studies (see also 1.1.2). Yet, genetic factors such as rare variants (frequency smaller than 1%), private mutations, CNVs especially small

insertions and deletions (indels) as well as interactions between genes have not been dissected well.

Association studies of rare variants in common diseases mainly suffer from the lack of enough power to detect true associations, especially for those variants which are not highly penetrant. For instance, for an allele that doubles disease risk (OR = 2) and has a frequency of 0.1%, nearly 60,000 cases and a similar number of controls will be required for the variant to reach genome-wide significance (Liu *et al*, 2014). To increase power to detect association, rare variants are often aggregated based on characteristics such as their position within genes, functional features and allele frequencies (Bansal *et al*, 2010). Dozens of these burden tests have been proposed along with methods for meta-analysis and replication (Hu *et al*, 2013).

Eventually it should be mentioned that the missing variability certainly comprise genetic and epigenetic components, which interact with environmental factors mainly microbiota of the gut (see 4.4.2) and likewise the full spectrum of genetic variance in IBD would not resolve completely, unless somatic and tissue-specific variations will be studied adequately.

4.6 Concluding remarks

This thesis was an effort to investigate the contribution of genomic CNVs, either somatic or germline to the risk of IBD.

Somatic CNV events in the blood or bowel biopsies of IBD patients in comparison to their healthy MZ co-twins were not confirmed in 6 IBD twins, examined through comparative genomic hybridization. However elucidating probable genetic variations underlying discordance of MZ twins needs more comprehensive high resolution studies that analyze all types of genomic structural variations as well as single nucleotide variants. Ideally whole genome sequencing of a larger number of IBD-discordant twins, coupled with fine histo-pathological cell targeting through laser-capture microdissection of intestinal epithelial cells

from inflamed mucosa might provide more accurate insights into the landscape of somatic and tissue-specific variations with relevance in IBD pathogenesis.

On the other hand, genome-wide analysis of germline CNVs for UC case-control cohorts in this thesis was an effort of further mining of accumulated SNP-GWAS datasets, towards characterization of additional genetic loci underlying susceptibility to IBD. In contrast to common variants, disease correlation of rare variants is however difficult to be assessed through classical association statistics. Low frequency of these variants impedes to detect associations at the genome-wide level significance by modest or intermediate sample sizes. Power limitations may further increase when these variants do not have high penetrance and confer only small effect sizes to the disease risk. The association signals of the rare CNV loci implicated in this thesis were not strong enough with genome-wide level significance and therefore their contribution to UC needs to be confirmed by other independent case-control as well as functional studies.

5 References

Abraham, C., Cho, J.H. Functional consequences of NOD2 (CARD15) mutations. Inflamm Bowel Dis, 12; 641–650 (2006)

Ananthakrishnan, A. N. *et al.* A prospective study of long-term intake of dietary fiber and risk of Crohn's disease and ulcerative colitis. Gastroenterology 145, 970–977 (2013).

Ananthakrishnan, A.N. Epidemiology and risk factors for IBD. Nature Reviews Gastroenterology & Hepatology. 12: 205–17 (2015).

Albert, M.A, *et al.* Effect of statin therapy on C-reactive protein levels: the pravastatin inflammation/CRP evaluation (PRINCE): a randomized trial and cohort study. JAMA. 286(1):64-70. (2001)

Alkan C, *et al.* Genome structural variation discovery and genotyping. Nat Rev Genet. 12(5):363-76 (2011).

Amre, D. K. *et al.* Imbalances in dietary consumption of fatty acids, vegetables, and fruits are associated with risk for Crohn's disease in children. Am. J. Gastroenterol. 102, 2016–2025 (2007).

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(3): 246–52 (2011).

Antonarakis, S. E., *et al.* Chromosome 21 and Down syndrome: From genomics to pathophysiology. Nature Reviews Genetics 5, 725–738 (2004)

Ait Yahya-Graison, E. *et al.* Classification of human chromosome 21 gene-expression variations in Down syndrome: impact on disease phenotypes. Am. J. Hum. Genet. 81, 475–491 (2007).

Babushok, D. V., Kazazian, H. H. Progress in understanding the biology of the human mutagen LINE-1.Hum Mutat. 28(6):527-39. (2007)

Bach, J. F. The effect of infections on susceptibility to autoimmune and allergic diseases. N Engl J Med 347, 911-20 (2002)

Bäckhed, F. *et al.* Host-bacterial mutualism in the human intestine. Science, 307; 1915–1920 (2005)

Bailey, J.A. *et al.* Recent segmental duplications in the human genome. Science 297: 1003–1007 (2002).

Bailey JA,Eichler EE.Primate segmental duplications: crucibles of evolution, diversity and disease.Nat Rev Genet.7(7):552-64. (2006)

Bansal, V. *et al.* Statistical analysis strategies for association studies involving rare variants. Nat Rev Genet. 11:773–785 (2010)

Baranzini, S.E. *et al.* Genome, epigenome and RNA sequences of monozygotic twins discordant for multiple sclerosis. Nature. 29;464(7293):1351-6. (2010)

Baross A, D. A., Li HI, Nayar T, Flibotte S, *et al.* (2007). "Assessment of algorithms for high throughput detection of genomic copy number variation in oligonucleotide microarray data." BMC Bioinformatics 8: 368.

Bentley, R. *et al.* Association of higher DEFB4 genomic copy number with Crohn's disease. Am. J. Gastroenterol 105: 354-359 (2010).

Bengtson, M.B. *et al.* Clustering in time of familial IBD separates ulcerative colitis from Crohn's disease. Inflamm Bowel Dis. 15(12):1867-74. (2009)

Berkes, J. *et al.* Intestinal epithelial responses to enteric pathogens: effects on the tight junction barrier, ion transport, and inflammation. Gut. 52(3):439–51 (2003).

Bergeron, V. *et al.* Current smoking differentially affects blood mononuclear cells from patients with Crohn's disease and ulcerative colitis: relevance to its adverse role in the disease. Inflamm. Bowel Dis. 18, 1101–1111 (2012).

Bernstein, C. N., *et al.* Epidemiology of Crohn's disease and ulcerative colitis in a central Canadian province: a population-based study. Am. J. Epidemiol. 149, 916–924 (1999).

Bernstein, C. N. *et al.* A population-based case control study of potential risk factors for IBD. Am. J. Gastroenterol. 101, 993–1002 (2006).

Bernstein, C. N. *et al.* A prospective population-based study of triggers of symptomatic flares in IBD. Am. J. Gastroenterol. 105, 1994–2002 (2010).

Biedermann, L. *et al.* Smoking cessation alters intestinal microbiota: insights from quantitative investigations on human fecal samples using FISH. Inflamm. Bowel Dis. 20, 1496–1501 (2014)

Binder, V. Genetic epidemiology in inflammatory bowel disease. Digest. Dis. 16, 351–355 (1998).

Bittner, S. *et al.* From the background to the spotlight: TASK channels in pathological conditions. Brain Pathol. 20(6):999-1009 (2010).

Blasius, A. L. *et al.* Slc15a4, AP-3, and Hermansky–Pudlak syndrome proteins are required for Toll-like receptor signaling in plasmacytoid dendritic cells. Proc. Natl Acad. Sci. USA 107, 19973–19978 (2010).

Bliek J, *et al.* Lessons from BWS twins: Complex maternal and paternal hypomethylation and a common source of haematopoietic stem cells. Eur J Hum Genet 17: 1625–1634. (2009)

Bonaz, B. L. & Bernstein, C. N. Brain–gut interactions in inflammatory bowel disease. Gastroenterology 144, 36–49 (2013).

Boomsma, D. Classical twin studies and beyond. Nat Rev Genet, 3 872–882 (2002)

Brant, S.R. Update on the heritability of inflammatory bowel disease: the importance of twin studies. Inflamm Bowel Dis, 17; 1–5 (2011)

Bruder, C.E. *et al.* Phenotypically concordant and discordant monozygotic twins display different DNA copy-number-variation profiles. Am. J. Hum. Genet. 82:763–771 (2008)

Buchanan, J. A. Scherer, S. W. Contemplating effects of genomic structural variation. Genet. Med 10: 639–647. (2008).

Budarf, M.L. *et al.* GWA studies: rewriting the story of IBD. Trends Genet. 25(3):137-46. (2009)

Cadwell, K., *et al.* A key role for autophagy and the autophagy gene Atg16l1 in mouse and human intestinal Paneth cells. Nature, 456. 259–263 (2008)

Cahan, P., Li, Y., Izumi, M. & Graubert, T. A. The impact of copy number variation on local gene expression in mouse hematopoietic stem and progenitor cells. Nature Genet. 41, 430–437 (2009).

Cardoso, J. *et al.* Genomic profiling by DNA amplification of laser capture microdissected tissues and array CGH. Nucleic Acids Res. 32(19): e146 (2004).

Carvalho, C.M., Lupski, J.R. Mechanisms underlying structural variant formation in genomic disorders Nat Rev Genet. 17(4):224-38. (2016)

Castillo-Fernandez, J. E. Epigenetics of discordant monozygotic twins: implications for disease. Genome Med. 6(7): 60 (2014)

Chaignat, E. *et al.* Copy number variation modifies expression time courses. Genome Res. 21, 106–113 (2011).

Chang, H.J, *et al.* Impact of blastocyst transfer on offspring sex ratio and the monozygotic twinning rate: A systematic review and meta-analysis." FertilSteril 91: 2381–2390 (2009)

Cho, J.H. The genetics and immunopathogenesis of inflammatory bowel disease. Nat Rev Immunol. 8(6):458-66. (2008)

Crohn, B.B, Ginzburg, L. Oppenheimer, G.D. Landmark article Oct 15, 1932. Regional ileitis. A pathological and clinical entity. By Burril B. Crohn, Leon Ginzburg, and Gordon D. Oppenheimer. JAMA. 6; 251(1):73-9 (1984)

Church, D. M. *et al.* Public data archives for genomic structural variation. Nature Genet.42, 813–814 (2010).

Claesson, M.J., *et al.* Composition, variability, and temporal stability of the intestinal microbiota of the elderly. PNAS. 108: 4586–91 (2011)

Colella, S. *et al.* QuantiSNP: an Objective Bayes Hidden-Markov Model to detect and accurately map copy number variation using SNP genotyping data. Nucleic Acids Res 35: 2013–2025 (2007)

Conrad, D.F. *et al.* A high-resolution survey of deletion polymorphism in the human genome. Nat. Genet. 38:75–81 (2006)

Conrad, D. F. *et al.* Origins and functional impact of copy number variation in the human genome. Nature464, 704–712 (2010).

Cooper, G.M. *et al.* Mutational and selective effects on copy-number variants in the human genome.*Nat. Genet.* 39:S22–29 (2007)

Cooper, G. M.*et al.* Systematic assessment of copy number variant detection via genome-wide SNP genotyping. Nature Genet.40, 1199–1203 (2008).

Costello, C.M. *et al.* Dissection of the inflammatory bowel disease transcriptome using genome-wide cDNA microarrays. PLoS Med. 2(8):e199 (2005)

Cosnes, J. Tobacco and IBD: relevance in the understanding of disease mechanisms and clinical practice. Best Pract. Res. Clin. Gastroenterol. 18, 481–496 (2004).

Czyz, W. *et al.* Genetic, environmental and stochastic factors in monozygotic twin discordance with a focus on epigenetic differences. BMC Med. 17;10:93 (2012)

Darai-Ramqvist E, *et al.* Duplications and evolutionary plasticity at tumor chromosome break-prone regions. Genome Res 18: 370–379. (2008)

Dave, M. *et al.* The human gut microbiome: current knowledge, challenges, and future directions Transl Res. 160(4):246-57 (2012)

Dicksved, J. *et al.* Molecular analysis of the gut microbiota of identical twins with Crohn's disease. ISME J. 2(7):716-27 (2008)

Diskin, S.J. *et al.* Adjustment of genomic waves in signal intensities from whole-genome SNP genotyping platforms. Nucleic Acids Res. 36(19):e126. (2008).

Dobson, A. J. An Introduction to Generalized Linear Models. London: Chapman and Hall. (1990)

Duerr, R.H. *et al.* A genome-wide association study identifies IL23R as an inflammatory bowel disease gene. Science, 314 pp. 1461–1463 (2006)

Dumanski, J.P, *et al.* Somatic mosaicism for copy number variation in differentiated human tissues. Human Mutation. 29 (9): 1118-1124. (2008)

Economou, M.*et al.* Differential effects of NOD2 variants on Crohn's disease risk and phenotype in diverse populations: a metaanalysis. Am J Gastroenterol, 99; 2393–2404 (2004)

Eder, W. *et al.* The Asthma Epidemic. N Engl J Med 2006; 355:2226-2235 (2006)

Eichler, E.E, *et al.* Missing heritability and strategies for finding the underlying causes of complex disease.Nat Rev Genet. 11(6):446–50 (2010).

Efimova, O., *et al.* Ncf1 (p47phox) is essential for direct regulatory T cell mediated suppression of CD4+ effector T cells. PLoS ONE 6, e16013 (2011).

Erlich Y. Blood ties: chimerism can mask twin discordance in high-throughput sequencing. Twin Res Hum Genet. 14(2):137-43 (2011)

Esteller, M. Non-coding RNAs in human disease. Nature Reviews Genetics 12, 861-874 (2011)

Falconer, D. S. Inheritance of liability to certain diseases estimated from incidence among relatives. Ann. Hum. Genet. 29, 51–76 (1965).

Falconer, Douglas S.; Mackay, Trudy F. C. Introduction to Quantitative Genetics (4th ed.) (1995).

Fraga, M.F., *et al.* Epigenetic differences arise during the lifetime of monozygotic twins. Proc Natl Acad Sci USA. 102: 10604-10609 (2005)

Fujino, S. *et al.* Increased expression of interleukin 17 in inflammatory bowel disease. Gut 52, 65–70 (2003).

Funke, B. Laser microdissection of intestinal epithelial cells and downstream analysis. Methods Mol Biol. 755:189-96. (2011)

Hollox, E.J *et al.* Extensive normal copy number variation of a beta-defensin antimicrobial-gene cluster. Am J Hum Genet, 73: 591–600 (2003).

Hollox, E.J *et al.* Psoriasis is associated with increased beta-defensin genomic copy number. Nature Genetics 40: 23 – 25 (2008).

Fanciulli, M,. *et al.* FCGR3B copy number variation is associated with susceptibility to systemic, but not organ-specific, autoimmunity. Nat. Genet. 39:721–23. (2007)

Fellermann, K,.*et al.* A chromosome 8 gene-cluster polymorphism with low human beta-defensin 2 gene copy number predisposes to Crohn disease of the colon. Am J Hum Genet 79: 439–448. (2006)

Fenske, J.N., Schwenk, T.L. Obsessive compulsive disorder: diagnosis and management". Am Fam Physician. 80 (3): 239–45 (2009).

Fernando, M.M. *et al.* Defining the role of the MHC in autoimmunity: a review and pooled analysis. PLoS Genet, 4 e1000024 (2008)

Fraga, M.F, *et al.* Epigenetic differences arise during the lifetime of monozygotic twins." Proc Natl AcadSci USA 102: 10604–10609 (2005)

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

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)

Friedman, S., Blumberg, R.S. Inflammatory Bowel Disease. A. Fauci (Ed.), Harrison's Principles of Internal Medicine (17th edn), McGraw-Hill; 1886–1898. (2008)

Funkhouser, L.J. Bordenstein, S.R. Mom knows best: the universality of maternal microbial transmission. PLoS Biol 11: e1001631 (2013)

Galvez, J., Rodriguez-Cabezas, M. E. & Zarzuelo, A. Effects of dietary fiber on inflammatory bowel disease. Mol. Nutr. Food Res. 49, 601–608 (2005).

Galton, F. The history of twins, as a criterion of the relative powers of nature and nurture." J Anthropol Inst 12: 566-576. (1875)

Ganz, T. Defensins: antimicrobial peptides of innate immunity. Nat Rev Immunol. 3(9): 710-720 (2003).

Gent, A. E., *et al.* Inflammatory bowel disease and domestic hygiene in infancy. Lancet. 343, 766–767 (1994).

Gervin, K., *et al.* Extensive variation and low heritability of DNA methylation identified in a twin study. Genome Res. 21: 1813-1821 (2011)

Girirajan, S. *et al.* Human copy number variation and complex genetic disease. Annu Rev Genet. 45:203-26 (2011).

Glessner, J.T, *et al.* Autism genome-wide copy number variation reveals ubiquitin and neuronal genes. Nature. 459(7246):569-73 (2009).

Goodier, J.L., Kazazian,H.H. Retrotransposons revisited: the restraint and rehabilitation of parasites. Cell.3; 135(1):23-35 (2008).

Greaves, M.F. *et al.* Leukemia in twins: lessons in natural history. Blood. 1;102(7):2321-33 (2003)

Gregersen PK, Olsson LM. Recent advances in the genetics of autoimmune disease. Annu. Rev. Immunol. 27:363–91 (2009)

Gu W, Zhang F, Lupski JR. Mechanisms for human genomic rearrangements. Pathogenetics. 3; 1(1):4. (2008)

Halfmann, R., Lindquist, S. Epigenetics in the extreme: prions and the inheritance of environmentally acquired traits. Science. 29;330(6004):629-32. (2010)

Halfvarson, J. Genetics in twins with Crohn's disease: less pronounced than previously believed?" Inflamm Bowel Dis. 17(1): 6-12 (2011).

Hall, J.G. Twins and twinning. Am J Med Genet 61: 202–204. (1996)

Hall, J.G. Twinning. Lancet 362: 735–743. (2003)

Halme L., *et al.* Family and twin studies in inflammatory bowel disease. World J Gastroenterol. 12:3668–3672. (2006)

Han K, *et al.* L1 recombination-associated deletions generate human genomic variation.Proc Natl AcadSci U S A. 105(49):19366-71. (2008)

Hassold, T. *et al.* To err (meiotically) is human: The genesis of human aneuploidy. Nature Reviews Genetics 2, 283. (2001)

Hastings, P. J. *et al.* Mechanisms of change in gene copy number. Nature Reviews Genetics 10, 551-564 (2009)

Hassan, S. W. *et al.* Increased susceptibility to dextran sulfate sodium induced colitis in the T cell protein tyrosine phosphatase heterozygous mouse. PLoS ONE 5, e8868 (2010).

Harries, A. D., Baird, A., Rhodes, J. Non-smoking: a feature of ulcerative colitis. Br. Med. J. (Clin. Res. Ed.) 284, 706 (1982).

Hatoum, O. A., Heidemann, J., Binion, D. G. The intestinal microvasculature as a therapeutic target in inflammatory bowel disease. Ann. NY Acad. Sci. 1072, 78–97 (2006).

Helbig, I. *et al.* 15q13.3 microdeletions increase risk of idiopathic generalized epilepsy. Nat. Genet. 41:160–62 (2009)

Henrichsen, C. N. *et al.* Segmental copy number variation shapes tissue transcriptomes. Nature Genet. 41, 424–429 (2009).

Heyn, H., Esteller, M. DNA methylation profiling in the clinic: applications and challenges. Nature Reviews Genetics 13, 679-692 (2012)

Hollox, E.J., Armour, J.A., Barber, J.C. Extensive normal copy number variation of a beta-defensin antimicrobial-gene cluster.*Am J Hum Genet* ; **73:** 591–600. (2003).

Hooper, L. V. & Macpherson, A. J. Immune adaptations that maintain homeostasis with the intestinal microbiota. Nature Rev. Immunol. 10, 159–169 (2010).

Horton, R. *Et al.* Gene map of the extended human MHC. Nat Rev Genet ;5(12):889-99. (2004)

Hsu, Y.-M. S. *et al.* The adaptor protein CARD9 is required for innate immune responses to intracellular pathogens. Nature Immunol. 8, 198–205 (2007).

Hu Y.-J., *et al.* Meta-analysis of Gene-Level Associations for Rare Variants Based on Single-Variant Statistics. Am J Hum Genet. 93:236–248 (2013)

Hugot, J.P, *et al.* Mapping of a susceptibility locus for Crohn's disease on chromosome 16. Nature. 29; 379(6568):821-3 (1996)

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

Human Microbiome Project Consortium. A framework for human microbiome research. Nature. 13;486(7402):215-21 (2012)

Iafrate, A. J. *et al.* Detection of large-scale variation in the human genome. Nature Genet.36, 949–951 (2004).

International Human Genome Sequencing Consortium. Finishing the euchromatic sequence of the human genome.Nature. 431(7011):931-45 (2004)

International HapMap Consortium. A haplotype map of the human genome. Nature 437, 1299–1320. (2005)

Itsara A, *et al.* Population analysis of large copy number variants and hotspots of human genetic disease. Am. J. Hum. Genet. 84:148–61 (2009)

Itsara, A. *et al.* *De novo* rates and selection of large copy number variation. Genome Res. 20:1469–81(2010).

Jacobs, P. A., Browne, C., Gregson, N., Joyce, C. & White, H. Estimates of the frequency of chromosome abnormalities detectable in unselected newborns using moderate levels of banding. J. Med. Genet. 29, 103–108 (1992).

Jones, P. A. Functions of DNA methylation: islands, start sites, gene bodies and beyond. Nature Reviews Genetics 13, 484-492 (2012)

Jostins, L. *et al.* Host-microbe interactions have shaped the genetic architecture of inflammatory bowel disease. Nature. 491(7422):119–24 (2012).

Ju, Y.S. *et al.* Reference-unbiased copy number variant analysis using CGH microarrays. Nucleic Acids Res. 38(20) e190 (2010)

Levine, B., Deretic, V. Unveiling the roles of autophagy in innate and adaptive immunity. Nat Rev Immunol, 7;767–777 (2007)

Lister, R. *et al.* Human DNA methylomes at base resolution show widespread epigenomic differences. Nature 462, 315–322 (2009).

Liu, J.Z., Anderson, C.A. Genetic studies of Crohn's disease: Past, present and future. Best Pract Res Clin Gastroenterol. 28(3): 373–386 (2014)

Lupsky, J.R. Genomic rearrangements and sporadic disease. Nat. Genet 39: 43–47 (2007)

Kaminsky, Z.A., *et al.* DNA methylation profiles in monozygotic and dizygotic twins. Nat Genet. 41: 240-245 (2009)

Kaser, A. *et al.* Inflammatory Bowel Disease. Annual Review of Immunology. 28: 573-621 (2009)

Kaur, N. *et al.* Intestinal dysbiosis in inflammatory bowel disease. Gut Microbes. 2(4):211-6 (2011)

Keith, L. Machin, G. Zygosity testing: Current status and evolving issues. J Reprod Med 42(11): 699–707. (1997)

Khalili, H. *et al.* Physical activity and risk of inflammatory bowel disease: prospective study from the Nurses' Health Study cohorts. BMJ 347, f 6633 (2013).

Khor, B. *et al.* Genetics and pathogenesis of inflammatory bowel disease. Nature. 15;474(7351):307-17 (2011)

Kidd, J.M, *et al.* Characterization of missing human genome sequences and copy-number polymorphic insertions. Nat. Methods 7: 365–371. (2010)

Kidd, J. M. *et al.* Mapping and sequencing of structural variation from eight human genomes. Nature 453, 56–64 (2008).

Kidd, J. M. *et al.* A human genome structural variation sequencing resource reveals insights into mutational mechanisms. Cell 143: 837–847 (2010)

Komura, D, *et al.* Noise reduction from genotyping microarrays using probe level information. In silico Biol 6(1-2): 79-92 (2006)

Kondo S, Mutations in IRF6 cause Van der Woude and popliteal pterygium syndromes. Nat Genet. 32(2):285-9 (2002)

Korbel, J. O. *et al.* Paired-end mapping reveals extensive structural variation in the human genome. Science 318, 420–426 (2007).

Korn, J. M. *et al.* Integrated genotype calling and association analysis of SNPs, common copy number polymorphisms and rare CNVs. Nature Genet.40, 1253–1260 (2008).

Kondrashov, A.S. Direct estimates of human per nucleotide mutation rates at 20 loci causing Mendelian diseases.Hum Mutat. 21(1):12-27 (2003)

Kostic, A. D., Xavier, R. J. Gevers, D. The Microbiome in Inflammatory Bowel Diseases: Current Status and the Future Ahead. Gastroenterology 146(6): 1489–1499 (2014)

Kronman, M. P. *et al.* Antibiotic exposure and IBD development among children: a population-based cohort study. Pediatrics 130, e794–e803 (2012).

Kruglyak, L. & Nickerson, D. A. Variation is the spice of life. Nature Genet. 27, 234–236 (2001)

Kuballa, P., *et al.* Impaired autophagy of an intracellular pathogen induced by a Crohn's disease associated ATG16L1 variant. PLoS One, 3 e3391 (2008)

Lakatos, P. L., Szamosi, T., Lakatos, L. Smoking in inflammatory bowel diseases: good, bad or ugly? World J. Gastroenterol. 13, 6134–6139 (2007).

Langrish, C.L. *et al.* IL-23 drives a pathogenic T cell population that induces autoimmune inflammation. The Journal of experimental medicine. 201:233–240 (2005).

Landers, C.J. *et al.* Selected loss of tolerance evidenced by Crohn's disease-associated immune responses to auto- and microbial antigens. Gastroenterology. 123:689–699 (2002)

Lees, C.W, *et al.* New IBD genetics: common pathways with other diseases. Gut. 60(12):1739-53 (2011)

Lee, J.A, *et al.* A DNA replication mechanism for generating nonrecurrent rearrangements associated with genomic disorders. Cell 131: 1235–1247.(2007).

Lejeune, J. *et al.* Study of somatic chromosomes from 9 mongoloid children. C R Hebd Seances Acad Sci. 248(11):1721-2. (1959)

Lennard-Jones, J.E. Classification of inflammatory bowel disease. Scand J Gastroenterol Suppl. 170, 2-6; discussion 16-9(1989)

Lepage, P. *et al.* Twin study indicates loss of interaction between microbiota and mucosa of patients with ulcerative colitis. Gastroenterology. 141(1):227-36. (2011)

Levy, S. *et al.* The diploid genome sequence of an individual human. PLoSBiol 4;5(10):e254 (2007)

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).

Lieber, M.R. *et al.* Mechanism and regulation of human non-homologous DNA end-joining.Nat Rev Mol Cell Biol. 4(9):712-20 (2003)

Liu, J.Z. *et al,*. Genetic studies of Crohn's disease: past, present and future. Best Pract Res Clin Gastroenterol;28:373–86 (2014)

Loftus, E. V. Jr *et al.* Ulcerative colitis in Olmsted County, Minnesota, 1940–1993: incidence, prevalence, and survival. Gut 46, 336–343 (2000).

Loftus, E. V. Jr. Clinical epidemiology of inflammatory bowel disease: incidence, prevalence, and environmental influences. Gastroenterology. 126**:** 1504–17 (2004).

Lovett, S. T., *et al.* Crossing over between regions of limited homology in Escherichia coli. RecA-dependent and RecA-independent pathways. Genetics 160: 851–859. (2002).

Lowe, A.M. *et al.* Epidemiology of Crohn's disease in Quebec, Canada.Inflamm Bowel Dis, 15: 429–435. (2009)

Lupski, J.R. Genomic rearrangements and sporadic disease.*Nat. Genet.* 39:S43–47 (2007).

Lyu., Nan. *et al.* Failure to Identify Somatic Mutations in Monozygotic Twins Discordant for Schizophrenia by Whole Exome Sequencing. Chin Med J (Engl). 20; 129(6): 690–695. (2016)

Martino, D. *et al.* Longitudinal, genome-scale analysis of DNA methylation in twins from birth to 18 months of age reveals rapid epigenetic change in early life and pair-specific effects of discordance. Genome Biol. 14: R42. (2013)

Martinez, C., *et al.* Unstable composition of the fecal microbiota in ulcerative colitis during clinical remission. Am J Gastroenterol, 103. 643–648 (2008)

Mayo, P. *et al.* CNV Analysis Using TaqMan Copy Number Assays. Current Protocols in Human Genetics. 67:Chapter 2:Unit2.13. (2010)

Muise, A. M. *et al.* Polymorphisms in E-cadherin (CDH1) result in a mis-localised cytoplasmic protein that is associated with Crohn's disease. Gut 58, 1121–1127 (2009).

Murphy, M., Hey, E. Twinning rates. The Lancet 349: 1398–1399 (1997).

MacDonald, J. R. *et al.*The Database of Genomic Variants: a curated collection of structural variation in the human genome. Nucleic Acids Res. 42, D986–D992 (2014).

Machin, G. Non-identical monozygotic twins, intermediate twin types, zygosity testing, and the non-random nature of monozygotic twinning: A review. Am J Med Genet Part C Semin Med Genet 151C:110–127 (2009)

Mahid, S. S. *et al.* Smoking and inflammatory bowel disease: a meta-analysis. Mayo Clin. Proc. 81, 1462–1471 (2006).

Maller, J. *et al.* Common variation in three genes, including a noncoding variant in CFH, strongly influences risk of age-related macular degeneration. Nature Genet. 38, 1055–1059 (2006)

Manichanh, C., *et al.* Reduced diversity of faecal microbiota in Crohn's disease revealed by a metagenomic approach. Gut. 55:205–21 (2006)

Manolio, T.A, *et al.* Finding the missing heritability of complex diseases. Nature. 8;461(7265):747-53. (2009)

Manolio, T. A. Bringing genome-wide association findings into clinical use. Nature Reviews Genetics 14, 549–558 (2013)

Mansilla, M.A. *et al.* Discordant MZ twins with cleft lip and palate: A model for identifying genes in complex traits. Twin Res Hum Genet 8: 39–46 (2005)

Marioni, J.C. *et al.* Breaking the waves: improved detection of copy number variation from microarray-based comparative genomic hybridization. Genome Biol: 8(10)(R228). (2007).

Matsukura H, *et al.* Discordant phenotypic expression of Alport syndrome in monozygotic twins. ClinNephrol 62: 313–318 (2004)

McCarroll, S.A. *et al.* Deletion polymorphism upstream of *IRGM* associated with altered *IRGM* expression and Crohn's disease. Nat Genet. 40(9):1107-12 (2008)

McCarroll, S. A. *et al.* Integrated detection and population-genetic analysis of SNPs and copy number variation. Nature Genet.40, 1166–1174 (2008).

McGilligan, V. E. *et al.* Hypothesis about mechanisms through which nicotine might exert its effect on the interdependence of inflammation and gut barrier function in ulcerative colitis. Inflamm. Bowel Dis. 13, 108–115 (2007).

McKernan, K. J. *et al.* Sequence and structural variation in a human genome uncovered by short-read, massively parallel ligation sequencing using two-base encoding. Genome Res. 19, 1527–1541 (2009).

Medvedev, P., Stanciu, M. &Brudno, M. Computational methods for discovering structural variation with next-generation sequencing.Nature Methods 6, S13–S20 (2009).

Mefford, H.C. *et al.* Recurrent rearrangements of chromosome 1q21.1 and variable pediatric phenotypes. N. Engl. J. Med. 359:1685–99 (2008)

Merla, G. *et al.* Submicroscopic deletion in patients with Williams-Beuren syndrome influences expression levels of the nonhemizygous flanking genes. Am. J. Hum. Genet. 79, 332–341 (2006)

Metzker, M. L. Sequencing technologies — the next generation. Nature Reviews Genetics11, 31-46 (2010)

Meuth, S.G, *et al.* TWIK-related acid-sensitive K+ channel 1 (TASK1) and TASK3 critically influence T lymphocyte effector functions. J Biol Chem. 283(21):14559–70 (2008).

Mills, R. E. *et al.* Mapping copy number variation at fine scale by population scale genome sequencing. Nature 470, 59–65 (2011).

Molodecky, N.A, *et al.* Increasing incidence and prevalence of the inflammatory bowel diseases with time, based on systematic review. Gastroenterology. 142(1):46-54. (2012)

Moum, B. *et al.* Incidence of ulcerative colitis and indeterminate colitis in four counties of southeastern Norway, 1990-93. A prospective population-based study. The Inflammatory Bowel South-Eastern Norway (IBSEN) Study Group of Gastroenterologists. Scand J Gastroenterol. 1996;31(4):362-6.

Nachman, M.W, Crowell, S.L. Estimate of the mutation rate per nucleotide in humans. Genetics. 156(1):297-304. (2000).

Nannya Y, *et al.* A robust algorithm for copy number detection using high-density oligonucleotide single nucleotide polymorphism genotyping arrays. Cancer Res. 65(14): 6071-6079. (2005)

Nishikawa, J.*et al.* Diversity of mucosa-associated microbiota in active and inactive ulcerative colitis. Scand J Gastroenterol. 44. 180–186 (2009)

Nobile C, T. L.,*et al.* Analysis of 22 deletion breakpoints in dystrophin intron 49. Hum Genet 110: 418–421(2002).

Nurminen, Markku. To Use or Not to Use the Odds Ratio in Epidemiologic Analyses?". European Journal of Epidemiology. 11 (4): 365–371 (1995)

Oates, N.A., *et al.* Increased DNA methylation at the AXIN1 gene in a monozygotic twin from a pair discordant for a caudal duplication anomaly. Am J Hum Genet 79: 155–162 ((2006)

O'Connor, C. Chromosomal abnormalities: Aneuploidies. Nature Education 1(1):172 (2008)

Ogunbi, S. O., *et al.* Inflammatory bowel disease in African-American children living in Georgia. J. Pediatr. 133, 103–107 (1998).

Ogura, Y. *et al.* A frameshift mutation in NOD2 associated with susceptibility to Crohn's disease. Nature, 411: 603–606 (2001)

O'Hara, A.M., Shanahan. F. The gut flora as a forgotten organ. EMBO Rep. 7(7):688-93 (2006)

Olshen, A.B. *et al.* Circular binary segmentation for the analysis of array-based DNA copy number data. Biostatistics. 5: 557–572 (2004)

Onnie, C.M, *et al.* Associations of allelic variants of the multidrug resistance gene (ABCB1 or MDR1) and inflammatory bowel disease and their effects on disease behavior: a case-control and meta-analysis study.Inflamm Bowel Dis. 12(4):263-71 (2006)

Orholm, M. *et al.* Familial occurrence of inflammatory bowel disease. N Engl J Med. 324**:** 84–8 (1991).

Orholm, M. *et al.* Concordance of inflammatory bowel disease among Danish twins. Results of a nationwidestudy. Scand J Gastroenterol. 35:1075–1081(2000)

Ostertag, E.M., Kazazian, H.H. Biology of mammalian L1 retrotransposons. Annu Rev Genet. 35():501-38. (2001)

Ott, S.J., *et al.* Dynamics of the mucosa-associated flora in ulcerative colitis patients during remission and clinical relapse J Clin Microbiol, 46. 3510–3513 (2008)

Pang, A. W. *et al.* Performance of high-throughput sequencing for the discovery of genetic variation across the complete size spectrum.G3 (Bethesda)4, 63–65 (2014).

Park, H. *et al.* Discovery of common Asian copy number variants using integrated high-resolution array CGH and massively parallel DNA sequencing. Nature Genet. 42, 400–405 (2010).

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(7):830-2 (2007)

Parkes, M. *et al.* Genetic insights into common pathways and complex relationships among immune-mediated diseases. Nature Reviews Genetics 14, 661–673 (2013)

Hastings, P. J. *et al.* Mechanisms of change in gene copy number." Nature Reviews Genetics 10: 551-564 (2009).

Peiffer, D. A. *et al.* High-resolution genomic profiling of chromosomal aberrations using Infinium whole-genome genotyping.Genome Res. 16, 1136–1148 (2006).

Perry, G. H. *et al.* The fine-scale and complex architecture of human copy-number variation. Am. J. Hum. Genet.82, 685–695 (2008).

Palmer C, Bik EM, DiGiulio DB, Relman DA, Brown PO. Development of the human infant intestinal microbiota. PLoS Biol 5: e177 (2007)

Penders, J. *et al.* Factors influencing the composition of the intestinal microbiota in early infancy. Pediatrics 118, 511–521 (2006).

Pinkel, D. *et al.* High resolution analysis of DNA copy number variation using comparative genomic hybridization to microarrays. Nature Genet.20, 207–211 (1998).

Pinto, D. *et al.* Comprehensive assessment of array-based platforms and calling algorithms for detection of copy number variants.Nature Biotech.29, 512–520 (2011).

Poeck, H. *et al.* Recognition of RNA virus by RIG-I results in activation of CARD9 and inflammasome signaling for interleukin 1β production. Nature Immunol. 11, 63–69 (2010).

Podolsky, D.K. Inflammatory bowel disease. N Engl J Med. 347(6):417–29.(2002)

Qin, J., *et al.* Human gut microbial gene catalogue established by metagenomic sequencing Nature. 4; 464(7285):59-65 (2010)

Radon, K. *et al.* Contact with farm animals in early life and juvenile inflammatory bowel disease: a case – control study. Pediatrics 120, 354–361 (2007)

Redon, R. *et al.* Global variation in copy number in the human genome.Nature 444, 444–454 (2006).

Reiter, L. T. *et al.* Human meiotic recombination products revealed by sequencing a hotspot for homologous strand exchange in multiple HNPP deletion patients. Am. J. Hum. Genet 62: 1023–1033. (1998)

Rengarajan, K.*et al.* Quantifying DNA concentrations using fluorometry: a comparison of fluorophores. Mol Vis. 6;8:416-21 (2002).

Renz, H., Brandtzaeg, P. & Hornef, M. The impact of perinatal immune development on mucosal homeostasis and chronic inflammation. Nature Rev. Immunol. 12, 9–23 (2011).

Rietveld, M. J. *et al.* Zygosity diagnosis in young twins by parental report. Twin Res 3: 134–14. (2000)

Rioux, J. D. & Abbas, A. K. Paths to understanding the genetic basis of autoimmune disease. Nature 435, 584–589 (2005).

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

Risch, N. J. Searching for genetic determinants in the new millennium. Nature 405, 847–856 (2000)

Rivas, M.A, *et al.* Deep resequencing of GWAS loci identifies independent rare variants associated with inflammatory bowel disease. Nat Genet. 43(11):1066-73. (2011)

Robertson, S.P, *et al.* Postzygotic mutation and germline mosaicism in the otopalatodigital syndrome spectrum disorders. Eur J Hum Genet 14: 549–554. (2006)

Rodríguez, J.M. The composition of the gut microbiota throughout life, with an emphasis on early life. Microb Ecol Health Dis. 2;26:26050. (2015)

Russo, V. E. A., Martienssen, R. A. & Riggs, A. D. Epigenetic Mechanisms of Gene Regulation. Cold Spring Harbor Laboratory Press, Woodbury. (1996)

Sabath, E. *et al.* Ga12 regulates protein interactions within the MDCK cell tight junction and inhibits tight-junction assembly. J. Cell Sci. 121, 814–824 (2008).

Saich, R., Chapman, R. Primary sclerosing cholangitis, autoimmune hepatitis and overlap syndromes in inflammatory bowel disease. World J Gastroenterol. 21; 14(3):331-7. (2008)

Sakuntabhai, A. *et al.* Mutations in ATP2A2, encoding a Ca2+ pump, cause Darier disease. Nat Genet 21: 271–277 (1999)

Sandler, R. S. in Inflammatory bowel disease: from bench to bedside (eds Targan, S. R. & Shanahan, F.) 5–30 Williams and Wilkins, Baltimore, (1994).

Sanger F; Nicklen S; Coulson AR. DNA sequencing with chain-terminating inhibitors". Proc. Natl. Acad. Sci. 74 (12): 5463–7 (1977).

Sassi, Y. *et al.* Multidrug resistance-associated protein 4 regulates cAMP-dependent signaling pathways and controls human and rat SMC proliferation. J Clin Invest. 118(8): 2747-57 (2008)

Satsangi, J., *et al.* Two stage genome-wide search in inflammatory bowel disease provides evidence for susceptibility loci on chromosomes 3, 7 and 12. Nat Genet. 14(2):199-202 (1996)

Scharl, M. *et al.* Protection of epithelial barrier function by the Crohn's disease associated gene protein tyrosine phosphatase N2. Gastroenterology 137, 2030–2040.e5 (2009).

Shinji, Ono. *et al.* Failure to confirm CNVs as of etiological significance in twins pairs discordant for schizophrenia. Twin research and human genetics. 13(5):455-460 (2010)

Schinzel, A.A.G.L, *et al.* Monozygotic Twinning and Structural Defects. Journal of Pediatrics. 95 (6): 921-930. (1979)

Schlattl, A., *et al.* Relating CNVs to transcriptome data at fine resolution: assessment of the effect of variant size, type, and overlap with functional regions. Genome Res. 21, 2004–2013 (2011).

Schmitz, H. *Et al.* Altered tight junction structure contributes to the impaired epithelial barrier function in ulcerative colitis. Gastroenterology. 116(2):301-9 (1999).

Schwarz, K. *et al.* Human severe combined immune deficiency and DNA repair. Bioessays. 25(11):1061-70. (2003)

Sebat, J. *et al.* Large-scale copy number polymorphism in the human genome.Science305, 525–528 (2004).

Sebat, J. *et al.* Strong association of *de novo* copy number mutations with autism. Science316, 445–449 (2007).

Sen, S.K.*et al.* Human genomic deletions mediated by recombination between Alu elements.Am J Hum Genet. 79(1):41-53.(2006)

Sepehri, S. *et al.* Microbial diversity of inflamed and noninflamed gut biopsy tissues in inflammatory bowel disease. Inflamm Bowel Dis. 13(6):675-83 (2007)

Sharp, A.J. *et al.* A recurrent 15q13.3 microdeletion syndrome associated with mental retardation and seizures. Nat. Genet. 40:322–28. (2008).

Shaw, M. H., *et al.* The ever-expanding function of NOD2: autophagy, viral recognition, and T cell activation. Trends Immunol. 32, 73–79 (2011).

Shaw, S. Y., Blanchard, J. F. & Bernstein, C. N. Association between the use of antibiotics in the first year of life and pediatric inflammatory bowel disease. Am. J. Gastroenterol. 105, 2687–2692 (2010).

Singh, S.M. *et al.* Copy number variation showers in schizophrenia: An emerging hypothesis. Mol Psychiatry 14: 356–358 (2009)

Shur, N. The genetics of twinning: From splitting eggs to breaking paradigms. Am J Med Genet Part C 151C: 105–109. (2009).

Simon, J.A., *et al.* Phenotypic predictors of response to simvastatin therapy among African-Americans and Caucasians: the Cholesterol and Pharmacogenetics (CAP) Study. Am J Cardiol. 97(6):843-50. (2006)

Simon-Sanchez, J. *et al.* Genome-wide SNP assay reveals structural genomic variation, extended homozygosity and cell-line induced alterations in normal individuals. Hum Mol Genet. 16(1):1-14 (2007).

Smigielski, E.M. *et al.* dbSNP: a database of single nucleotide polymorphisms. Nucleic Acids Res. 28(1):352-5. (2000)

Smith, C., Falconer, D. S. & Duncan, L. J. P. Statistical and genetic study of diabetes: II. Heritability of liability. Ann. Hum. Genet. 35, 281–299 (1972).

Stankiewicz, P. *et al.*Genome architecture catalyzes nonrecurrent chromosomal rearrangements. Am J Hum Genet. 72(5):1101-16 (2003)

Stankiewicz P, Lupski JR. Genome architecture, rearrangements and genomic disorders.Trends Genet. 18(2):74-82.(2002)

Stefansson, H,.*et al.* Large recurrent microdeletions associated with schizophrenia. Nature 455:232–36 (2008).

Sonnenberg, A. Occupational distribution of inflammatory bowel disease among German employees. Gut 31, 1037–1040 (1990).

Song, J.S., *et al.* Model-based analysis of two-color arrays (MA2C). Genome Biol 8(8):R178. (2007).

Spehlmann, M.E. Epidemiology of inflammatory bowel disease in a German twin cohort: results of a nationwide study. Inflamm Bowel Dis 14: 968–976. (2008)

Spehlmann, M.E. *et al.* Risk factors in German twins with inflammatory bowel disease: results of a questionnaire-based survey. J Crohns Colitis. 6(1):29-42. (2012)

Stankiewicz P, Lupski JR. Genome architecture, rearrangements and genomic disorders.Trends Genet. 18(2):74-82 (2002)

Stecher, B. & Hardt, W. D. Mechanisms controlling pathogen colonization of the gut. Curr. Opin. Microbiol. 14, 82–91 (2011).

Stockard, C.R. Developmental rate and structural expression: an experimental study of twins, double monsters and single deformities and the interaction among embryonic organs during their origin and development. Am J Anat 28: 115–127. (1921)

Stone, J.L, *et al.* Rare chromosomal deletions and duplications increase risk of schizophrenia. Nature. 455(7210):237-41. (2008)

Stratton, M.R, *et al.* The cancer genome. Nature 458:719–24 (2009)

Sudmant, P. H. *et al.* Diversity of human copy number variation and multicopy genes. Science330, 641–646 (2010).

Swanson, G. R., Burgess, H. J. & Keshavarzian, A. Sleep disturbances and inflammatory bowel disease: a potential trigger for disease flare? Expert Rev. Clin. Immunol. 7, 29–36 (2011).

Szatmari, P. *et al.* Mapping autism risk loci using genetic linkage and chromosomal rearrangements. Nat. Genet. 39:319–28 (2007)

Tenesa, A., Haley, C.S. The heritability of human disease: estimation, uses and abuses. Nat Rev Genet. 14(2):139-49. (2013)

The 1000 Genomes Project Consortium. A map of human genome variation from population-scale sequencing.Nature467, 1061–1073 (2010).

The 1000 Genomes Project Consortium *et al.* An integrated map of genetic variation from 1,092 human genomes.Nature 491, 56–65 (2012).

The Int. HapMap Consort.The International HapMap Project.Nature. 2426:789–96 (2003)

The Int. HapMap Consort.A haplotype map of the human genome.Nature. 437:1299–320 (2005)

The International HapMap Consortium. A second generation human haplotype map of over 3.1 million SNPs. Nature 449, 851–861 (2007)

The Int. HapMap Consort. Integrating common and rare genetic variation in diverse human populations. Nature 467, 52–58 (2010)

Thompson, N.P. *et al.* Genetics versus environment in inflammatory bowel disease: results of a British twinstudy. BMJ. 312:95–96 (1996)

Thompson, E.R. Development and Validation of an International English Big-Five Mini-Markers. Personality and Individual Differences. 45 (6): 542–548. (2008)

Tims, S., *et al.* Host genotype and the effect on microbial communities. In: Nelson KE, ed. Metagenomics of the human body. Springer, New York: Springer Science+Business Media, LLC; pp. 15–41.(2011)

Toffolatti, L. *et al.* Investigating the mechanism of chromosomal deletion: characterization of 39 deletion breakpoints in introns 47 and 48 of the human dystrophin gene.Genomics. 80(5):523-30. (2002)

Turner D. J. *et al.* "Germline rates of *de novo* meiotic deletions and duplications causing several genomic disorders." Nat Genet 40: 90–95 (2008)

Tuzun E, *et al.* Fine-scale structural variation of the human genome. Nat. Genet. 37:727–32 (2005).

Tysk, C, *et al.* Ulcerative colitis and Crohn's disease in an unselected population of monozygotic and dizygotic twins.A study of heritability and the influence of smoking. Gut 29: 990–996 (1988)

UK IBD Genetics Consortium. Genome-wide association study of ulcerative colitis identifies three new susceptibility loci, including the HNF4A region. Nature Genet. 41(12):1330-4 (2009)

van de Ven, R. *et al.* A role for multidrug resistance protein 4 (MRP4;ABCC4) in human dendritic cell migration. Blood.112(6):2353-9 (2008)

van de Wiel, M.A, *et al.* Smoothing waves in array CGH tumor profiles. Bioinformatics. 25(9): 1099-1104 (2009)

Van Ommen, G. B. Frequency of new copy number variation in humans. Nat Genet 37: 333–334.(2005).

Vazquez-Mena, O. *et al.* Amplified genes may be overexpressed, unchanged, or downregulated in cervical cancer cell lines. PLoS ONE 7, e32667 (2012).

Veltman, J. A. & Brunner, H. G. *De novo* mutations in human genetic disease. Nature Reviews Genetics 13, 565-575 (2012)

Vignali, M. ATP-Dependent Chromatin-Remodeling Complexes. Mol Cell Biol. 20(6): 1899–1910. (2000)

Visscher, P.M., Hill, W.G. , Wray, N.R. Heritability in the genomics era--concepts and misconceptions. Nat Rev Genet. 9(4):255-66. (2008)

Wang, J. *et al.* The diploid genome sequence of an Asian individual. Nature 456: 60–65 (2008).

Wang, K. *et al.*PennCNV: an integrated hidden Markov model designed for high-resolution copy number variation detection in whole-genome SNP genotyping data. Genome Res 17: 1665–1674 (2007)

Weischenfeldt, J. *et al.* Phenotypic impact of genomic structural variation: insights from and for human disease. Nat Rev Genet. 14(2):125-38. (2013)

Wellcome Trust Case Control Consortium. Genome-wide association study of CNVs in 16,000 cases of eight common diseases and 3,000 shared controls. Nature 464, 713–720 (2010).

Wichmann, H.E., *et al.* KORA-gen--resource for population genetics, controls and a broad spectrum of disease phenotypes. Gesundheitswesen 67 Suppl 1, S 26-30 (2005)

Wilks, S. Morbid appearances in the intestine of Miss Bankes. London Medical Times & Gazette. 2:264 (1859)

Willing, B.P. *et al.* A pyrosequencing study in twins shows that gastrointestinal microbial profiles vary with inflammatory bowel disease phenotypes. Gastroenterology. 139(6):1844-1854.e1. (2010)

Wheeler, D. A. *et al.* The complete genome of an individual by massively parallel DNA sequencing.Nature 452, 872–876 (2008).

Willer, C. J. *et al.* Twin discordance and sibling recurrence rates in multiple sclerosis. Proc. Natl Acad.Sci.100:12877-12882 (2003)

Winchester, L., Yau, C., Ragoussis, J. Comparing CNV detection methods for SNP arrays. Brief. Funct.Genomic.Proteomic.8, 353–366 (2009).

Whitman, W.B. *et al.* Prokaryotes: the unseen majority. Proc. Natl. Acad. Sci. 95 ;6578–6583 (1998)

Wu, W., *et al.* CARD9 facilitates microbe-elicited production of reactive oxygen species by regulating the LyGDI-Rac1 complex. Nature Immunol. 10, 1208–1214 (2009).

Xavier, R. J. & Podolsky, D. K. Unravelling the pathogenesis of inflammatory bowel disease. Nature 448, 427-434 (2007)

Xavier, R. J. & Rioux, J. D Genome-wide association studies: a new window into immune-mediated diseases. Nature Reviews Immunology 8, 631-643 (2008)

Youssoufian H, P. R. Mechanisms and consequences of somatic mosaicism in humans. Nat. Rev. Genet. 3: 748–758 (2002).

Zeissig, S. *et al.* Changes in expression and distribution of claudin 2, 5 and 8 lead to discontinuous tight junctions and barrier dysfunction in active Crohn's disease. Gut. 56(1):61–72 (2007).

Zhang, F. *et al.*The DNA replication *FoSTeS*/MMBIR mechanism can generate genomic, genic and exonic complex rearrangements in humans.Nat Genet. 41(7):849-53. (2009)

Zhang, D. *et al.* Singleton deletions throughout the genome increase risk of bipolar disorder. Mol Psychiatry. 14(4):376-80. (2009)

Zimmermann, C. *et al.* Mapping of multidrug resistance gene 1 and multidrugresistance-associated protein isoform 1 to 5 mRNA expression along the human intestinal tract. Drug Metab Dispos. 33(2):219-24 (2005)

Zuk, O. *et al.* The mystery of missing heritability: Genetic interactions create phantom heritability. Proc Natl Acad Sci U S A. 24;109(4):1193-8. (2012)

Zwijnenburg, P.J. *et al.* Identical but not the same: the value of discordant monozygotic twins in genetic research. Am J Med Genet B Neuropsychiatr Genet Sep;153B(6):1134-1149 (2010)

6 Supplementary Material

Sample Recruitment and Ethics

German patients of the discovery and replication panels were recruited either at the Department of General Internal Medicine of the Christian-Albrechts-University Kiel, the Charité University Hospital Berlin, through local outpatient services, or nationwide with the support of the German Crohn's and Colitis Foundation. Clinical, radiological, histological, and endoscopic (i.e. type and distribution of lesions) examinations were required to unequivocally confirm the diagnosis of ulcerative colitis (UC).1214 German healthy control individuals of discovery panel (1703 total) were obtained from the biobank PopGen (http://www.popgen.de).The remaining 489 German healthy controls in discovery panel were selected from the KORA F4survey, an independent population-based sample from the general population living in the region of Augsburg, Southern Germany[2].Written, informed consent was obtained from all study participants and all protocols were approved either by the ethical committee of the University-Hospital Schleswig-Holstein, Center Kiel or through the institutional committee of the "KompetenznetzDarmerkrankungen" (http://www.kompetenznetz-ced.de/) in Germany.

The 274 clinically well-characterized Norwegian UC patients of replication panel were recruited through a population-based incidence study, the Inflammatory Bowel disease in South-Eastern Norway (IBSEN) study. An ethnically and sex-matched group of Norwegian healthy controls (n=282) was randomly selected from the Norwegian Bone Marrow Donor Registry (NBMDR). The strict criteria (including absence of any autoimmune disease) on inclusion in the NBMDR ensured correct classification of these controls as healthy. Norwegian sample recruitment was approved by the ethics committee of Oslo University Hospital, Rikshospitalet, Norway.

The Lithuanian study population consisted of 443 UC patients recruited at 6 hospitals in Lithuania: Kaunas Medical University Hospital (Kaunas), Vilnius University Hospital Santariskiu Clinic(Vilnius), M. Marcinkevicius Hospital (Vilnius), Klaipeda Seamen's Hospital (Klaipeda), Panevezys Regional Hospital (Panevezys), Siauliai Regional Hospital (Siauliai) and 2 hospitals in Latvia: P. Stradin Clinical University Hospital (Riga) and Riga Eastern ClinicalUniversity Hospital,Clinic Linezers (Riga). The diagnosis of UC was based on standard clinical, endoscopic, radiological and histological criteria. The control group consisted of 1157ethnically, age and sex-matched healthy blood donors. Written, informed consent was obtained from all study participants and all protocols were approved by the institutional ethical review committee of the Lithuanian University of Health Sciences, Kaunas, Lithuania.

SNP array genotyping

The genotyping for the German discovery panel and Norwegian panel - which were both part of the German NGFN GWAS initiative (see press release 04-26-07 on http://www.ngfn.de/englisch/index_368.htm) funded by the NGFN – was performed by an Affymetrix® service facility (South San Francisco, CA, USA) using the Affymetrix® Genome-Wide Human SNP Array 6.0 (1000k) (Santa Clara, CA, USA). The array is based on an assay termed whole-genome sampling analysis (WGSA) developed for highly multiplexed SNP genotyping of complex DNA. This method reproducibly amplifies a subset of the human genome through a single primer amplification reaction using restriction enzyme digested, adapter-ligated human genomic DNA. In brief, 5 μl of genomic DNA samples at 50 ng/ul were aliquoted to the corresponding wells of two 96-well plates. The first run of samples was processed as an entire plate. In the lab, transfers were made with a 12-channel pipette, reducing the risk of sample tracking errors. One plate was digested with NspI and the other plate was digested with StyI. The reaction was incubated at 37 °C for 2 hours and at 65 °C for 20 minutes to deactivate the enzyme. The digested DNA was then ligated to their respective NspI adaptor and StyI adaptor. The ligated product was then PCR-amplified using a common primer. Both NspI PCR product and StyI PCR product were combined, and then purified by ethanol precipitation in combination with membrane filter plate. Purified PCR product was further fragmented with DNaseI then labeled with biotin. Labeled NA was combined with hybridization mix and then injected into array. Arrays were hybridized for 18 to 22 hours at 50°C. DNA samples were recovered from arrays and washed and stained by using Affymetrix® FS450 fluidic stations. Stained arrays were scanned using Affymetrix® GeneChip Scanner 3000 7G.

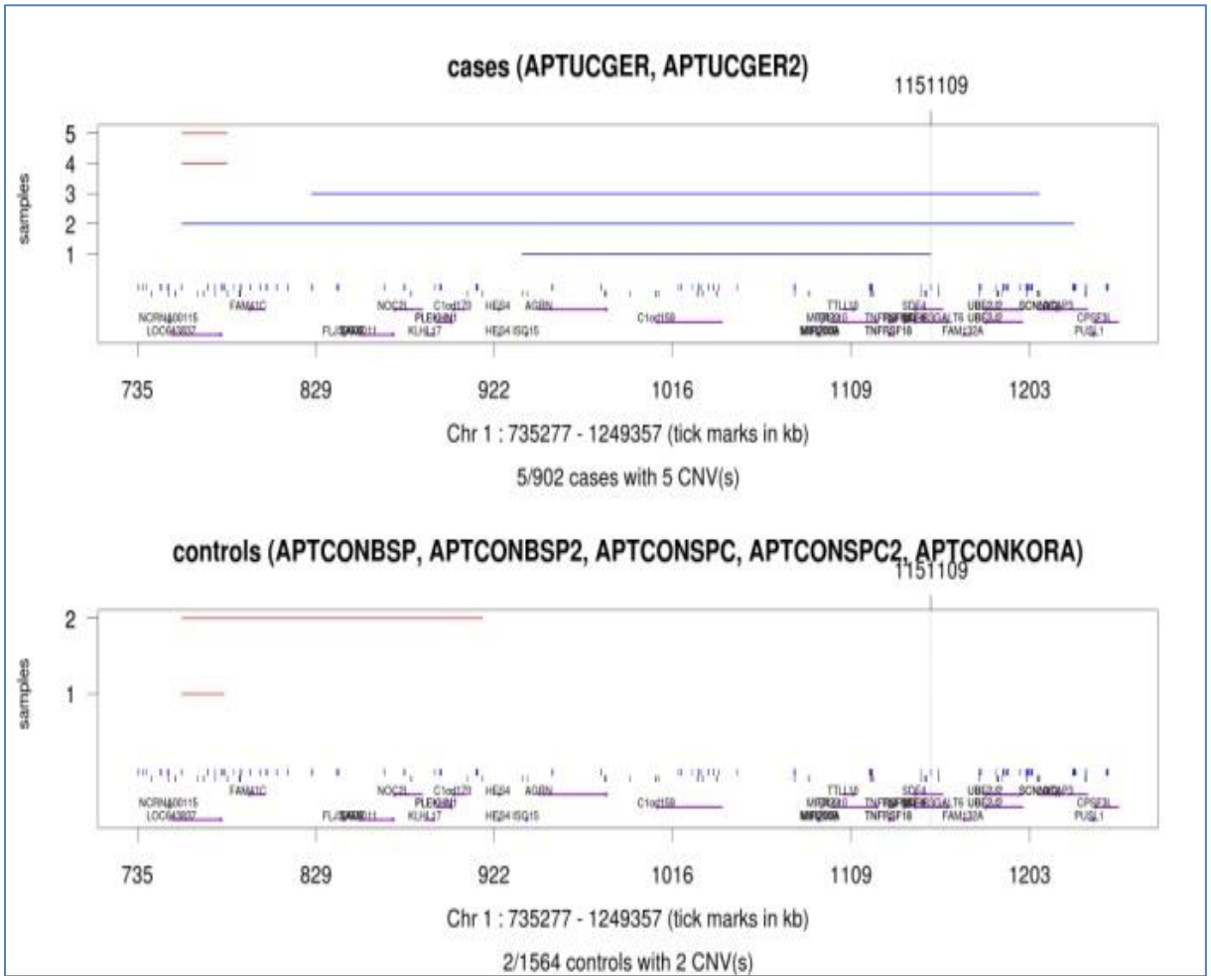
WTCCC2 data processing

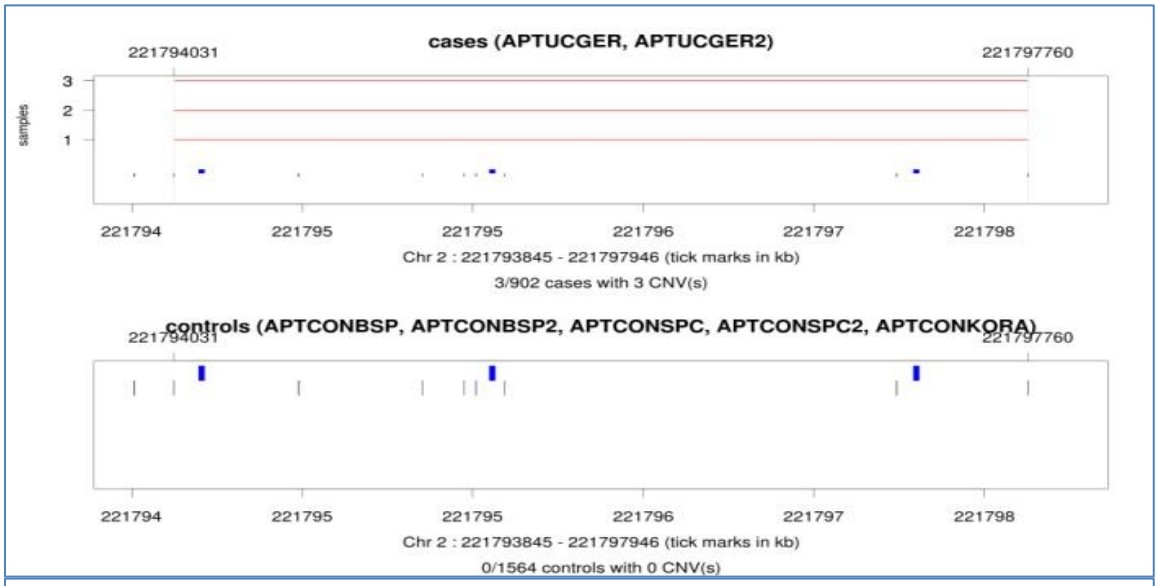
First the dataset was cleaned based on genotype calling. Genotype calling was performed using the Birdseed v2 algorithm implemented in Affymetrix Power Tools version 1.12.0 [4]. Subsequently, genotypes with their corresponding intensity values were converted into the Beagle format by means of python scripts. Because we observed an excess of genotyping artifacts that would result in false positive associations, the software Beagle call v1.0.1 (Browning and Yu 2009) was used to generate accurate genotype calls by using both allele signal intensities and inter-marker correlation. Beagle call's built-in data quality filters excluded any marker with large deviation from the Hardy-Weinberg equilibrium (PHWE < 10-6) or with > 5% of samples having maximal genotype probability < 0.95. We also excluded individuals from each pair of unexpected duplicates or relatives, as well as individuals with outlier heterozygosities of

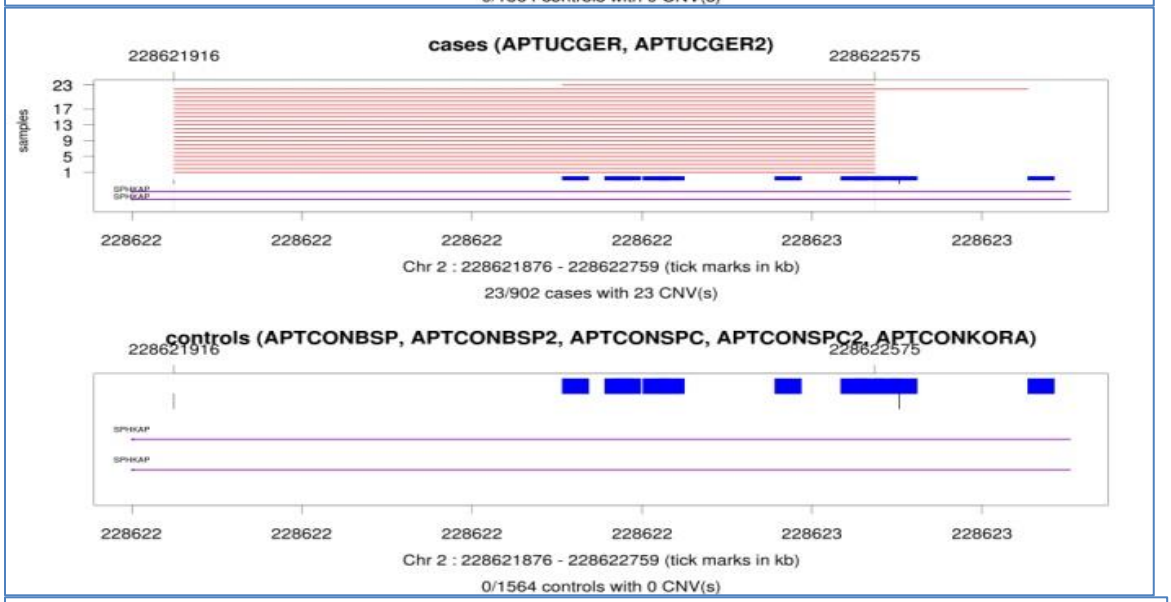
± 5 s.d. away from the mean. The remaining GWAS samples were tested for population stratification using the principal components stratification method, as implemented in EIGENSTRAT [5], and population outliers were subsequently excluded. SNPs that had a minor allele frequency < 1% and exact Hardy-Weinberg equilibrium P controls< 10-4 were finally excluded. For the remaining samples a CNV calling was performed. CEL files were processed with the Affymetrix Power Tools (APT)apt-copynumber-workflow v 1.67. The values for contrastQC and MAPD were extracted and samples that failed default QC values were discarded (MAPD > 0.4 and/or contrastQC< 0.4). The apt-copynumber-workflow output was converted to CNVineta format [6]. A preliminary filtering was performed based on the number of called CNVs per samples. This was performed batch wise, as one batch consists of a sample collection which was prepared in the same process. Outlier were defined as samples which had more CNVs than the 75% quantile plus 1.5 fold of the interquantile range.

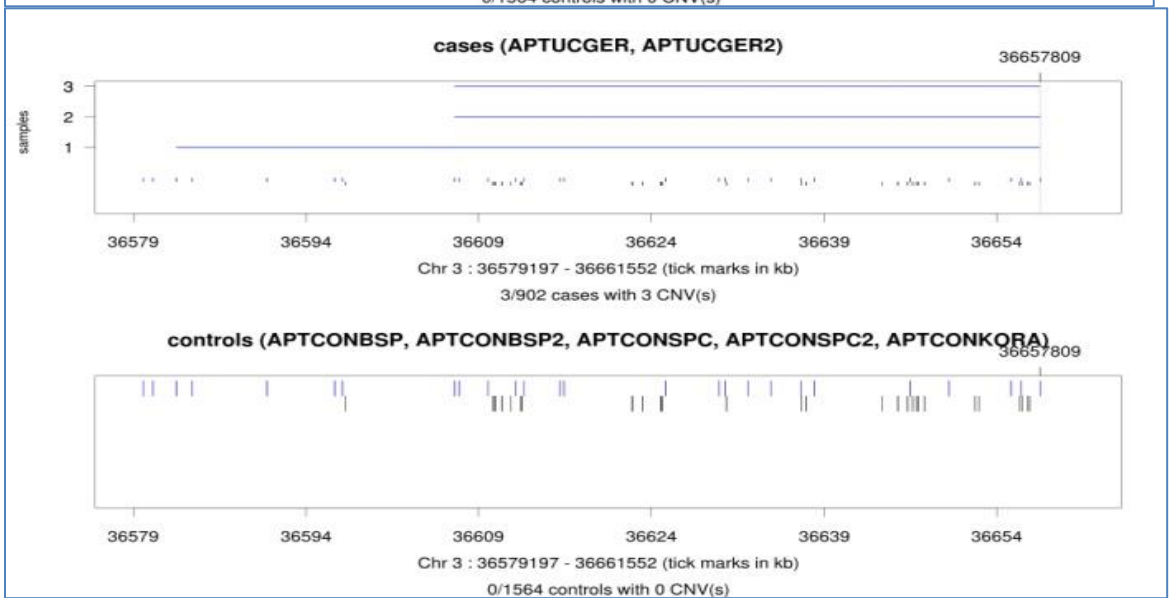
Table 6.1 Twenty-four regions with rare CNVs overrepresented in UC cases of the German discovery panel.

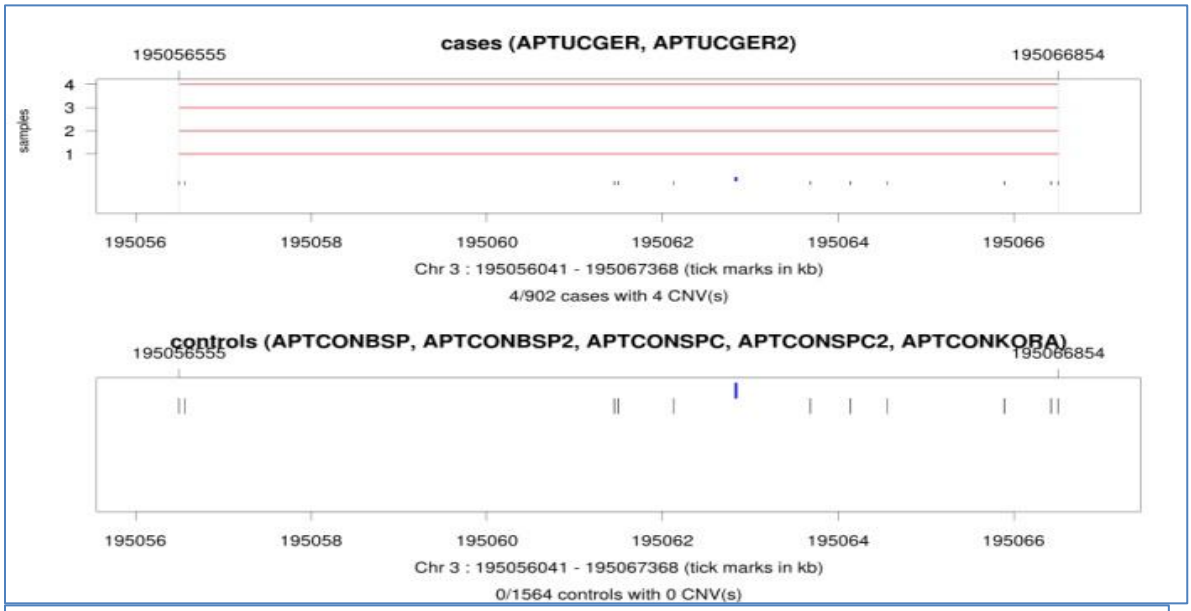
For each region the predicted CNVs are shown for cases (upper panel) and controls (lower panel) with Duplications in blue and deletions in red. Involved RefSeq genes are annotated. SNP probe sets in black and copy number probe sets in blue.

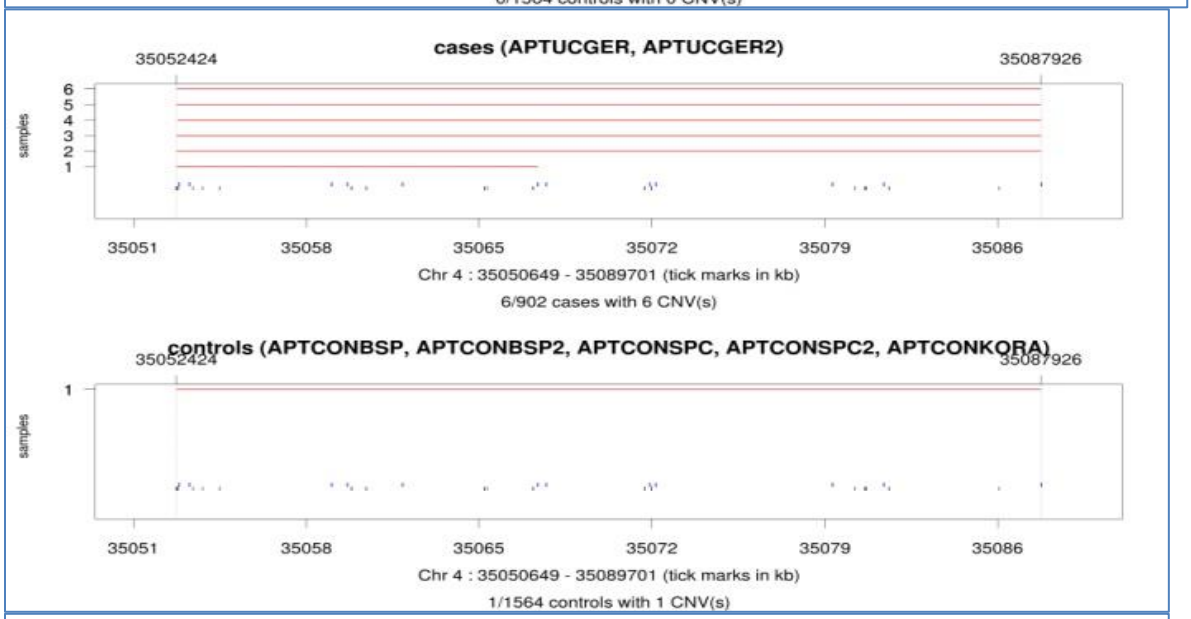


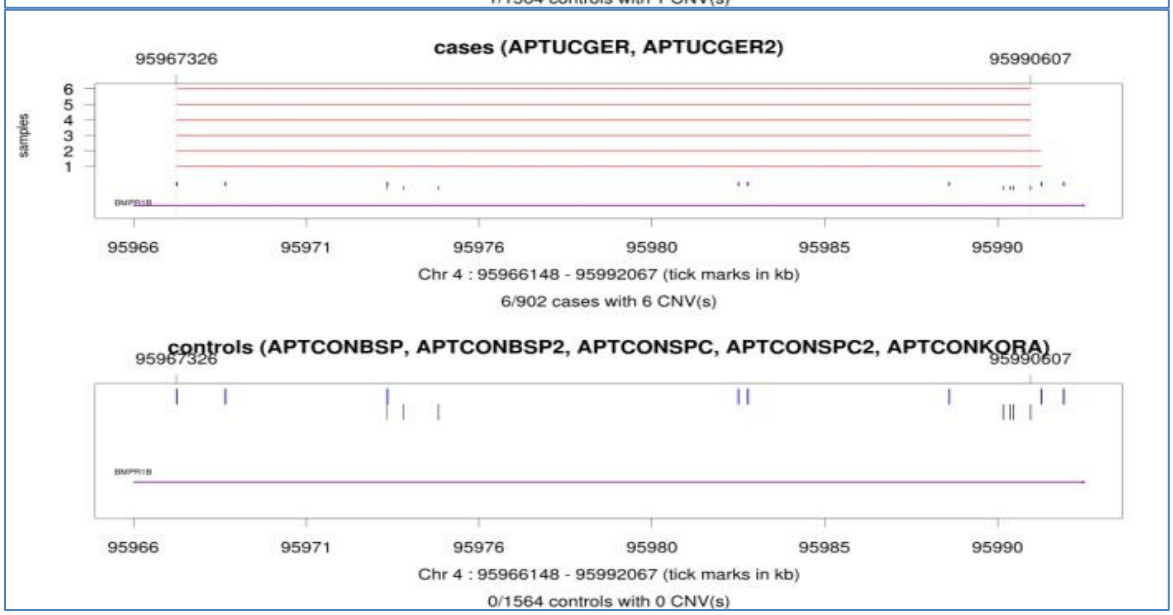

cases (APTUCGER, APTUCGER2)
221794031
221797760
samples
3
2
1
221794
221795
221795
221796
221797
221798
Chr 2 : 221793845 - 221797946 (tick marks in kb)
3/902 cases with 3 CNV(s)
controls (APTCONBSP, APTCONBSP2, APTCONSPC, APTCONSPC2, APTCONKORA)
221794031
221797760
221794
221795
221795
221796
221797
221798
Chr 2 : 221793845 - 221797946 (tick marks in kb)
0/1564 controls with 0 CNV(s)

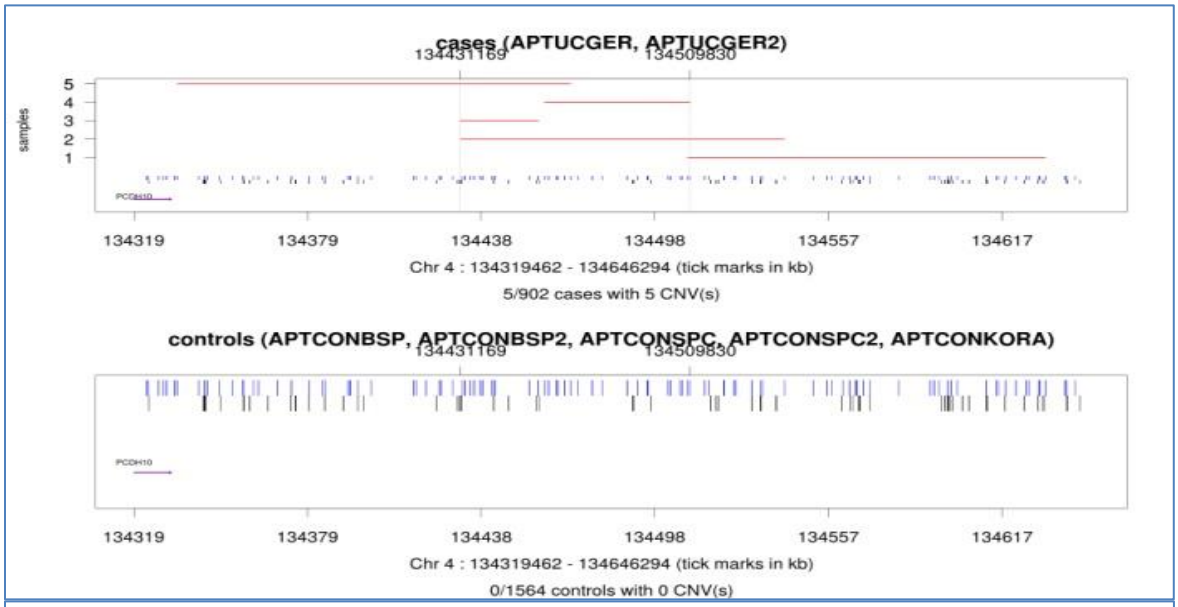

cases (APTUCGER, APTUCGER2)
228621916
228622575
samples
23
17
13
9
5
1
SPHKAP
SPHKAP
228622
228622
228622
228622
228623
228623
Chr 2 : 228621876 - 228622759 (tick marks in kb)
23/902 cases with 23 CNV(s)
controls (APTCONBSP, APTCONBSP2, APTCONSPC, APTCONSPC2, APTCONKORA)
228621916
228622575
SPHKAP
SPHKAP
228622
228622
228622
228622
228623
228623
Chr 2 : 228621876 - 228622759 (tick marks in kb)
0/1564 controls with 0 CNV(s)

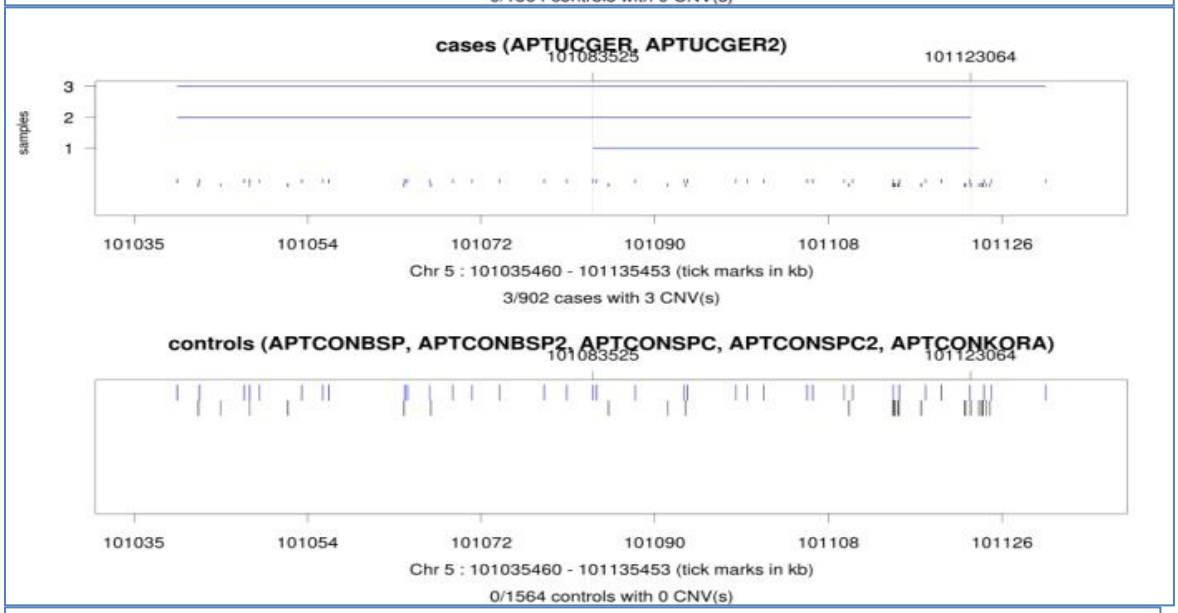

cases (APTUCGER, APTUCGER2)
36657809
samples
3
2
1
36579
36594
36609
36624
36639
36654
Chr 3 : 36579197 - 36661552 (tick marks in kb)
3/902 cases with 3 CNV(s)
controls (APTCONBSP, APTCONBSP2, APTCONSPC, APTCONSPC2, APTCONKORA)
36657809
36579
36594
36609
36624
36639
36654
Chr 3 : 36579197 - 36661552 (tick marks in kb)
0/1564 controls with 0 CNV(s)

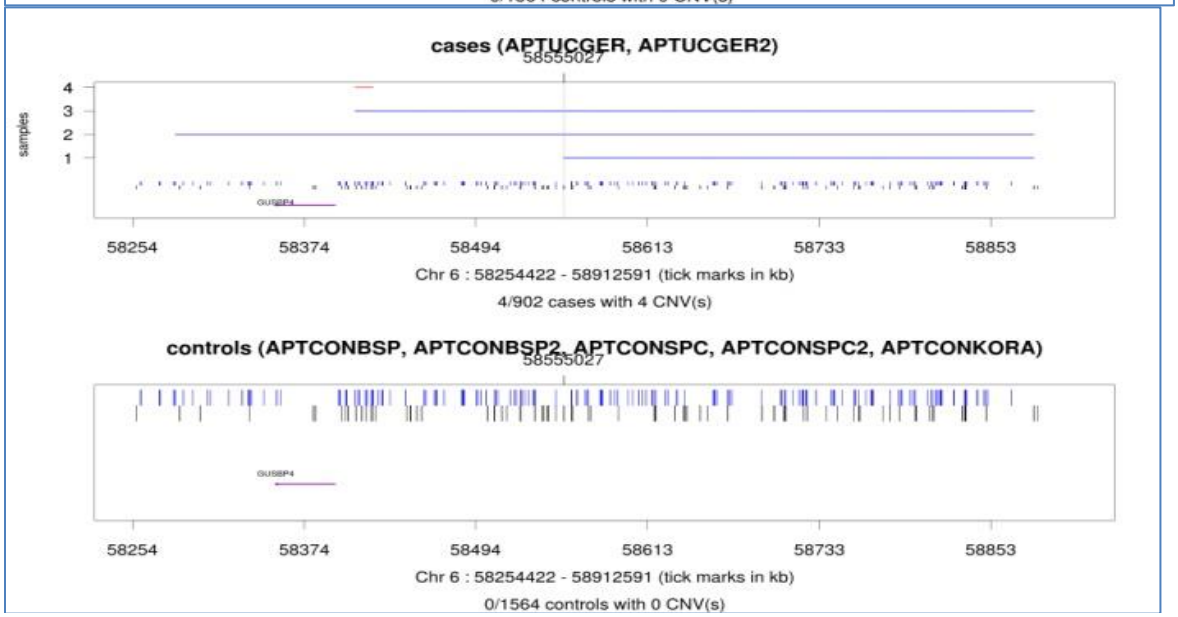

cases (APTUCGER, APTUCGER2)
195056555
195066854
samples
4
3
2
1
195056
195058
195060
195062
195064
195066
Chr 3 : 195056041 - 195067368 (tick marks in kb)
4/902 cases with 4 CNV(s)
controls (APTCONBSP, APTCONBSP2, APTCONSPC, APTCONSPC2, APTCONKORA)
195056555
195066854
195056
195058
195060
195062
195064
195066
Chr 3 : 195056041 - 195067368 (tick marks in kb)
0/1564 controls with 0 CNV(s)

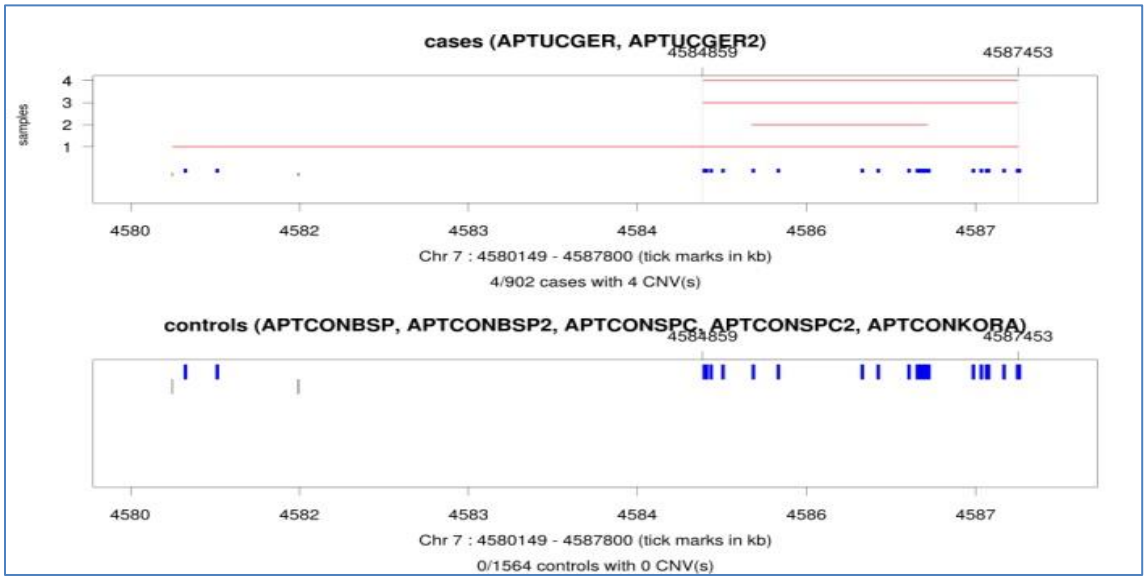

cases (APTUCGER, APTUCGER2)
35052424
35087926
samples
6
5
4
3
2
1
35051
35058
35065
35072
35079
35086
Chr 4 : 35050649 - 35089701 (tick marks in kb)
6/902 cases with 6 CNV(s)
controls (APTCONBSP, APTCONBSP2, APTCONSPC, APTCONSPC2, APTCONKORA)
35052424
35087926
samples
1
35051
35058
35065
35072
35079
35086
Chr 4 : 35050649 - 35089701 (tick marks in kb)
1/1564 controls with 1 CNV(s)

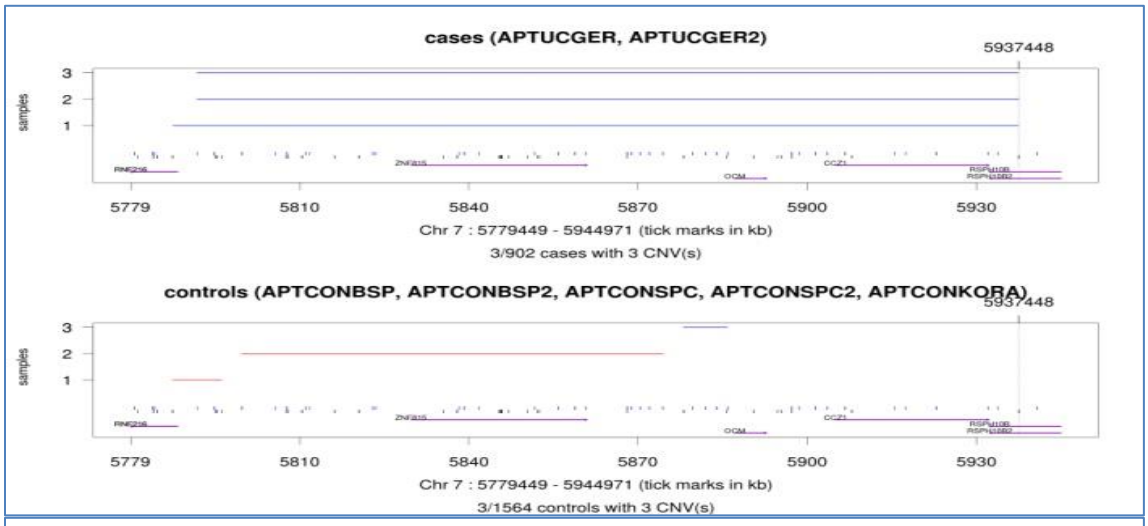

cases (APTUCGER, APTUCGER2)
95967326
95990607
samples
6
5
4
3
2
1
95966
95971
95976
95980
95985
95990
Chr 4 : 95966148 - 95992067 (tick marks in kb)
6/902 cases with 6 CNV(s)
controls (APTCONBSP, APTCONBSP2, APTCONSPC, APTCONSPC2, APTCONKORA)
95967326
95990607
95966
95971
95976
95980
95985
95990
Chr 4 : 95966148 - 95992067 (tick marks in kb)
0/1564 controls with 0 CNV(s)

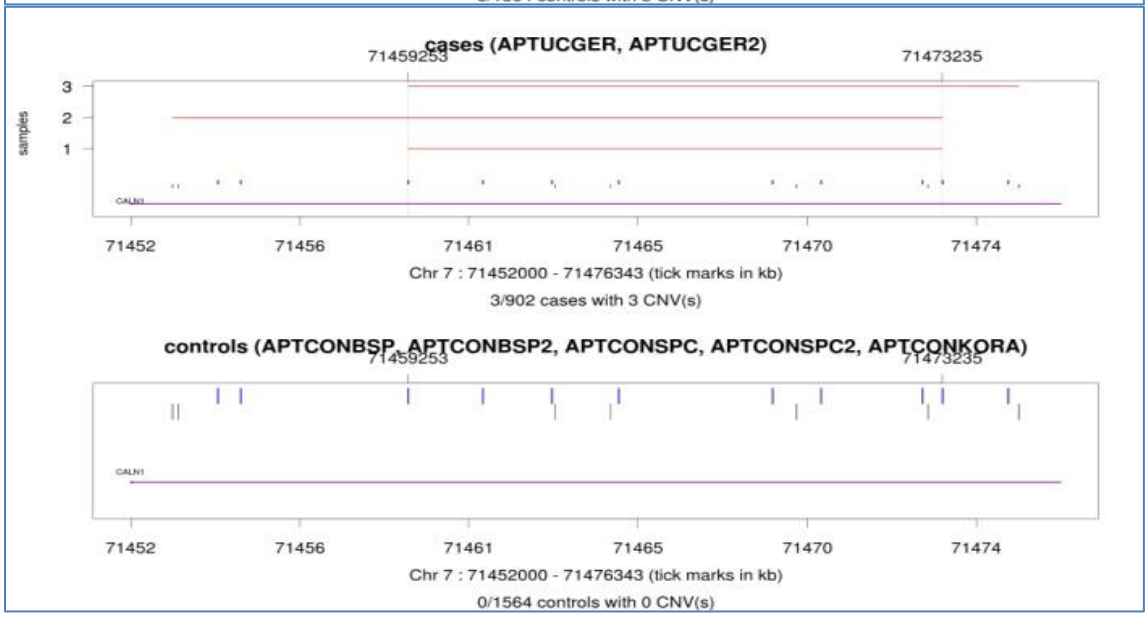

cases (APTUCGER, APTUCGER2)
134431169
134509830
samples
5
4
3
2
1
134319
134379
134438
134498
134557
134617
Chr 4 : 134319462 - 134646294 (tick marks in kb)
5/902 cases with 5 CNV(s)
controls (APTCONBSP, APTCONBSP2, APTCONSPC, APTCONSPC2, APTCONKORA)
134431169
134509830
PCDH10
134319
134379
134438
134498
134557
134617
Chr 4 : 134319462 - 134646294 (tick marks in kb)
0/1564 controls with 0 CNV(s)

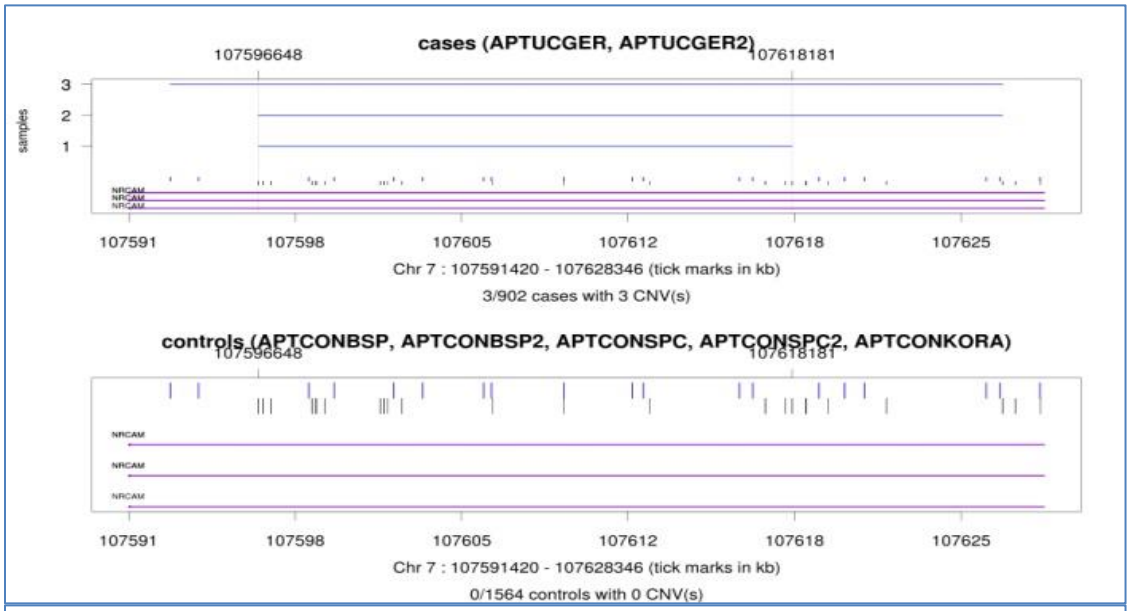

cases (APTUCGER, APTUCGER2)
101083525
101123064
samples
3
2
1
101035
101054
101072
101090
101108
101126
Chr 5 : 101035460 - 101135453 (tick marks in kb)
3/902 cases with 3 CNV(s)
controls (APTCONBSP, APTCONBSP2, APTCONSPC, APTCONSPC2, APTCONKORA)
101083525
101123064
101035
101054
101072
101090
101108
101126
Chr 5 : 101035460 - 101135453 (tick marks in kb)
0/1564 controls with 0 CNV(s)

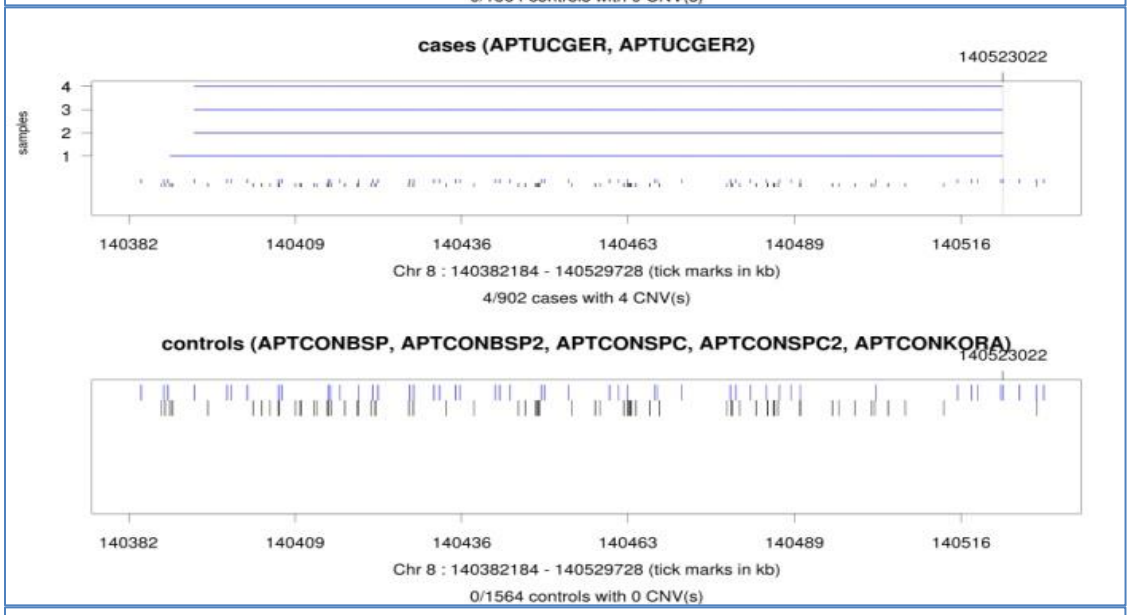

cases (APTUCGER, APTUCGER2)
58555027
samples
4
3
2
1
58254
58374
58494
58613
58733
58853
Chr 6 : 58254422 - 58912591 (tick marks in kb)
4/902 cases with 4 CNV(s)
controls (APTCONBSP, APTCONBSP2, APTCONSPC, APTCONSPC2, APTCONKORA)
58555027
GUSBP4
58254
58374
58494
58613
58733
58853
Chr 6 : 58254422 - 58912591 (tick marks in kb)
0/1564 controls with 0 CNV(s)

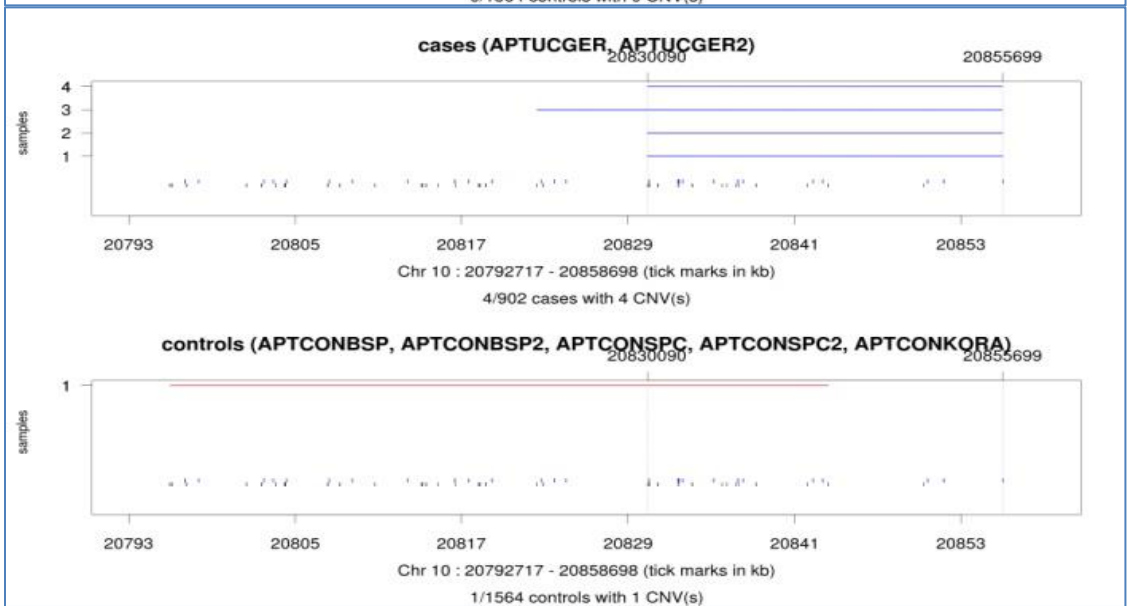

cases (APTUCGER, APTUCGER2)
4584859
4587453
samples
4
3
2
1
4580
4582
4583
4584
4586
4587
Chr 7 : 4580149 - 4587800 (tick marks in kb)
4/902 cases with 4 CNV(s)
controls (APTCONBSP, APTCONBSP2, APTCONSPC, APTCONSPC2, APTCONKORA)
4584859
4587453
4580
4582
4583
4584
4586
4587
Chr 7 : 4580149 - 4587800 (tick marks in kb)
0/1564 controls with 0 CNV(s)

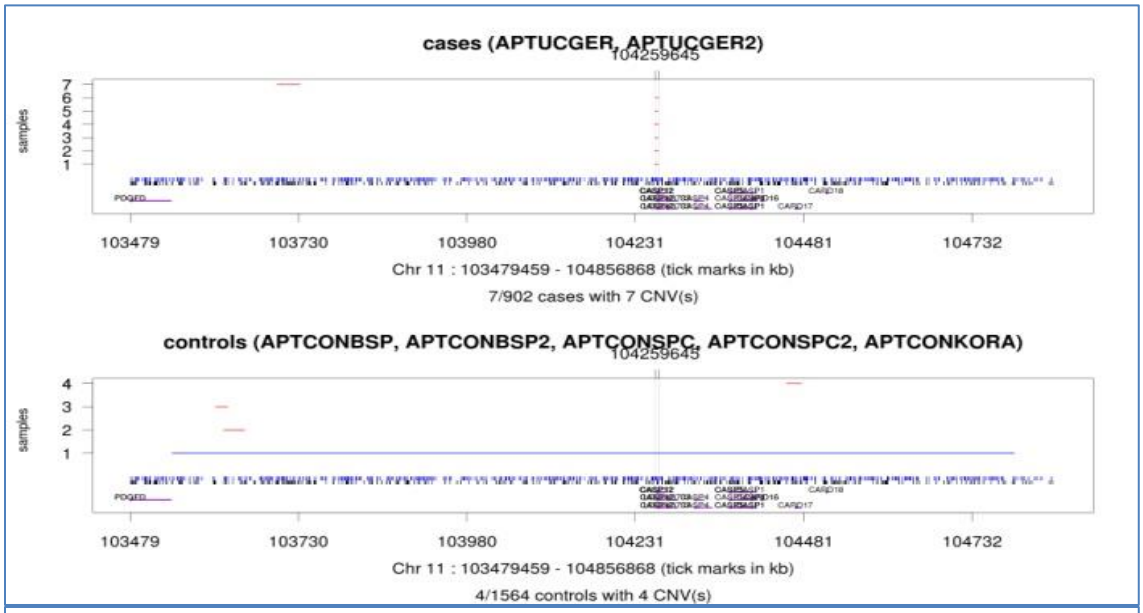

cases (APTUCGER, APTUCGER2)
5937448
samples
3
2
1
5779
5810
5840
5870
5900
5930
Chr 7 : 5779449 - 5944971 (tick marks in kb)
3/902 cases with 3 CNV(s)
controls (APTCONBSP, APTCONBSP2, APTCONSPC, APTCONSPC2, APTCONKORA)
5937448
5779
5810
5840
5870
5900
5930
Chr 7 : 5779449 - 5944971 (tick marks in kb)
3/1564 controls with 3 CNV(s)

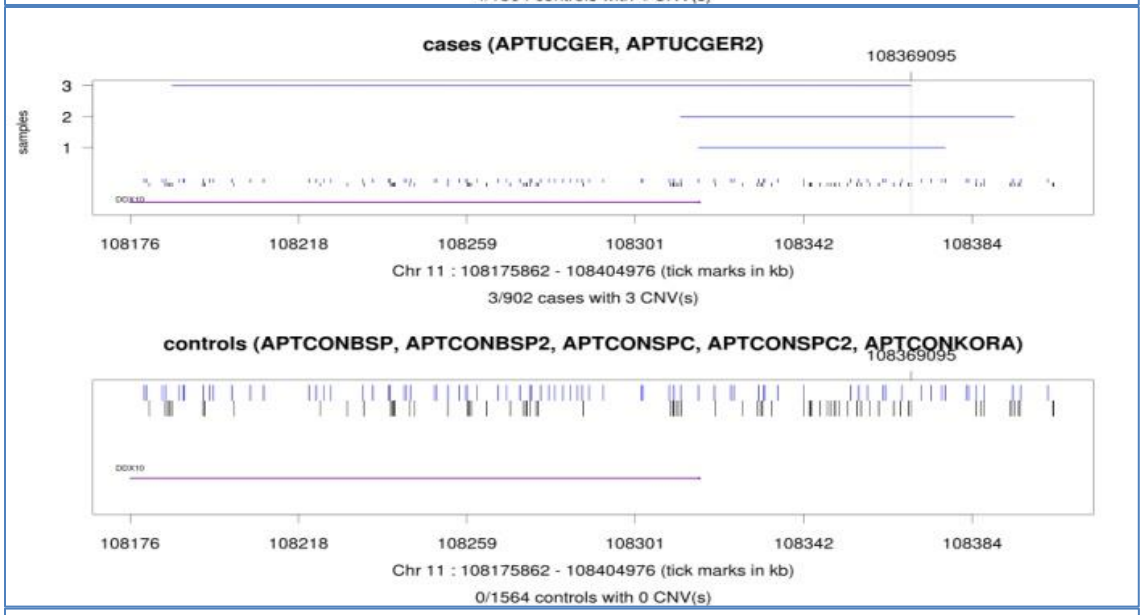

cases (APTUCGER, APTUCGER2)
71459253
71473235
samples
3
2
1
71452
71456
71461
71465
71470
71474
Chr 7 : 71452000 - 71476343 (tick marks in kb)
3/902 cases with 3 CNV(s)
controls (APTCONBSP, APTCONBSP2, APTCONSPC, APTCONSPC2, APTCONKORA)
71459253
71473235
71452
71456
71461
71465
71470
71474
Chr 7 : 71452000 - 71476343 (tick marks in kb)
0/1564 controls with 0 CNV(s)

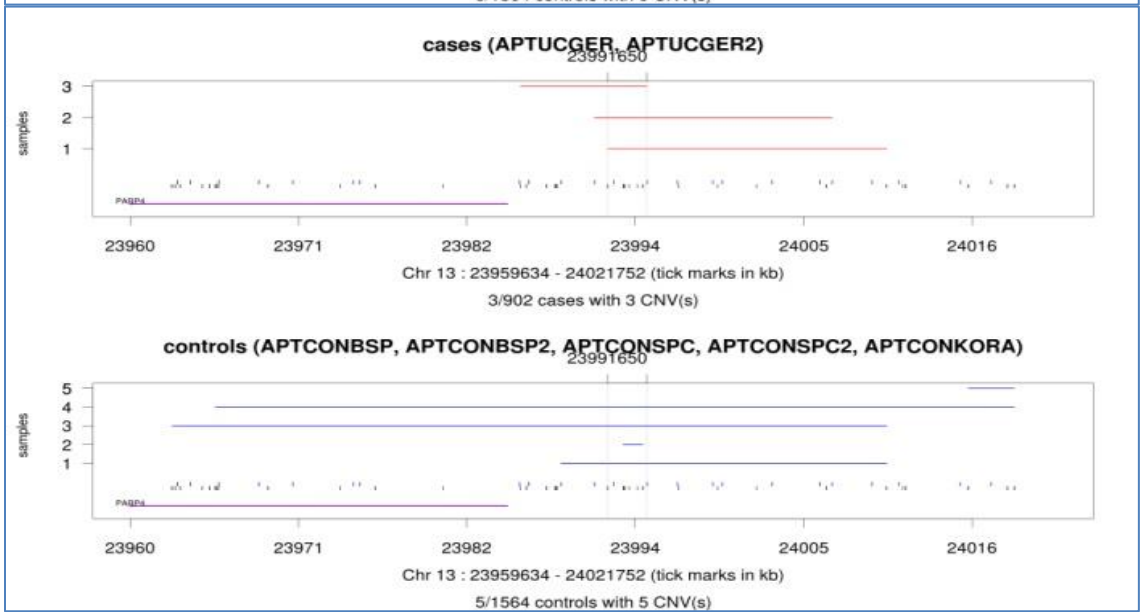

cases (APTUCGER, APTUCGER2)
107596648
107618181
samples
3
2
1
NRCAM
NRCAM
NRCAM
107591
107598
107605
107612
107618
107625
Chr 7 : 107591420 - 107628346 (tick marks in kb)
3/902 cases with 3 CNV(s)
controls (APTCONBSP, APTCONBSP2, APTCONSPC, APTCONSPC2, APTCONKORA)
107596648
107618181
NRCAM
NRCAM
NRCAM
107591
107598
107605
107612
107618
107625
Chr 7 : 107591420 - 107628346 (tick marks in kb)
0/1564 controls with 0 CNV(s)

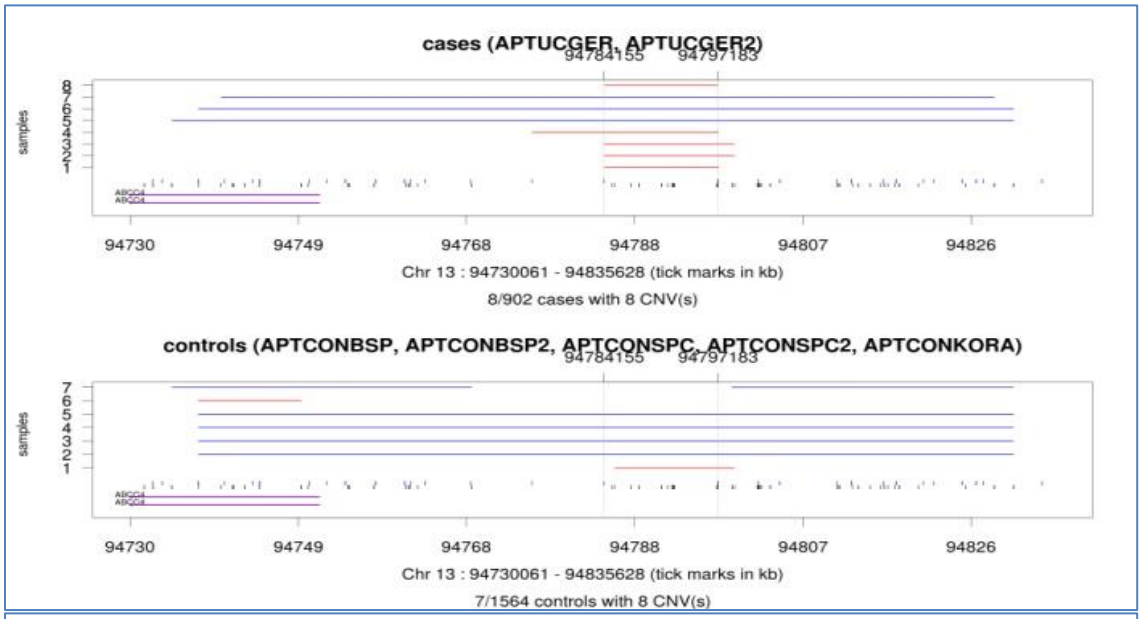

cases (APTUCGER, APTUCGER2)
140523022
samples
4
3
2
1
140382
140409
140436
140463
140489
140516
Chr 8 : 140382184 - 140529728 (tick marks in kb)
4/902 cases with 4 CNV(s)
controls (APTCONBSP, APTCONBSP2, APTCONSPC, APTCONSPC2, APTCONKORA)
140523022
140382
140409
140436
140463
140489
140516
Chr 8 : 140382184 - 140529728 (tick marks in kb)
0/1564 controls with 0 CNV(s)

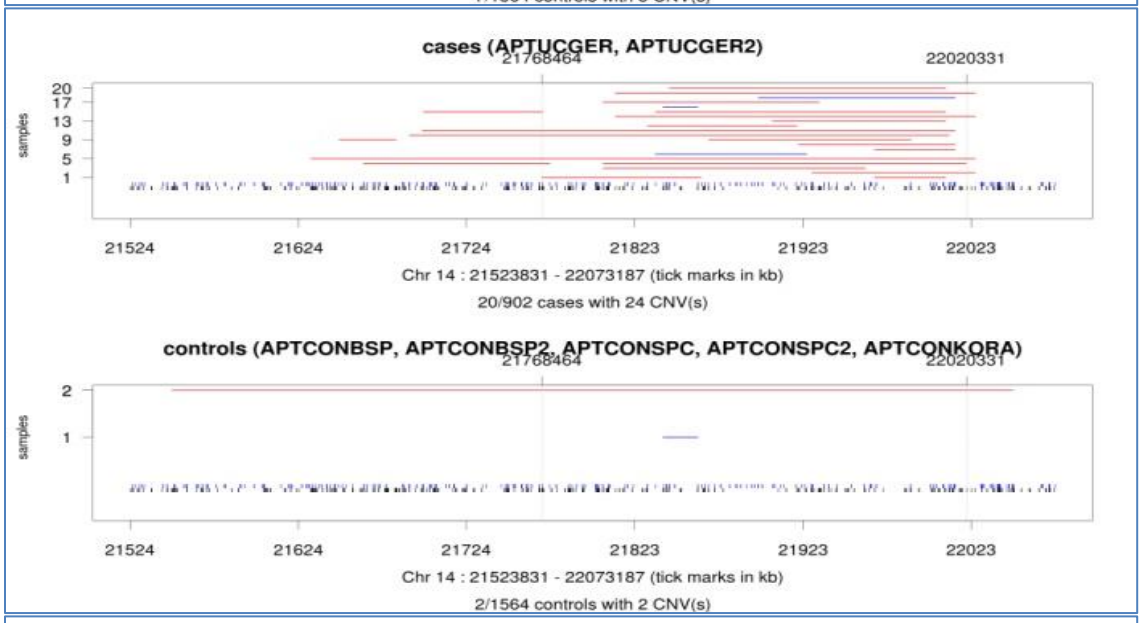

cases (APTUCGER, APTUCGER2)
20830090
20855699
samples
4
3
2
1
20793
20805
20817
20829
20841
20853
Chr 10 : 20792717 - 20858698 (tick marks in kb)
4/902 cases with 4 CNV(s)
controls (APTCONBSP, APTCONBSP2, APTCONSPC, APTCONSPC2, APTCONKORA)
20830090
20855699
samples
1
20793
20805
20817
20829
20841
20853
Chr 10 : 20792717 - 20858698 (tick marks in kb)
1/1564 controls with 1 CNV(s)


cases (APTUCGER, APTUCGER2)
104259645
samples
7
6
5
4
3
2
1
PDGFD
CASP12
CASP9AS1
CARD18
CASP9AS1-SEM1
CASP9AS1-SEM2
CARD16
CASP9AS1-SEM3
CASP9AS1
CARD17
103479
103730
103980
104231
104481
104732
Chr 11 : 103479459 - 104856868 (tick marks in kb)
7/902 cases with 7 CNV(s)
controls (APTCONBSP, APTCONBSP2, APTCONSPC, APTCONSPC2, APTCONKORA)
104259645
samples
4
3
2
1
PDGFD
CASP12
CASP9AS1
CARD18
CASP9AS1-SEM1
CASP9AS1-SEM2
CARD16
CASP9AS1-SEM3
CASP9AS1
CARD17
103479
103730
103980
104231
104481
104732
Chr 11 : 103479459 - 104856868 (tick marks in kb)
4/1564 controls with 4 CNV(s)

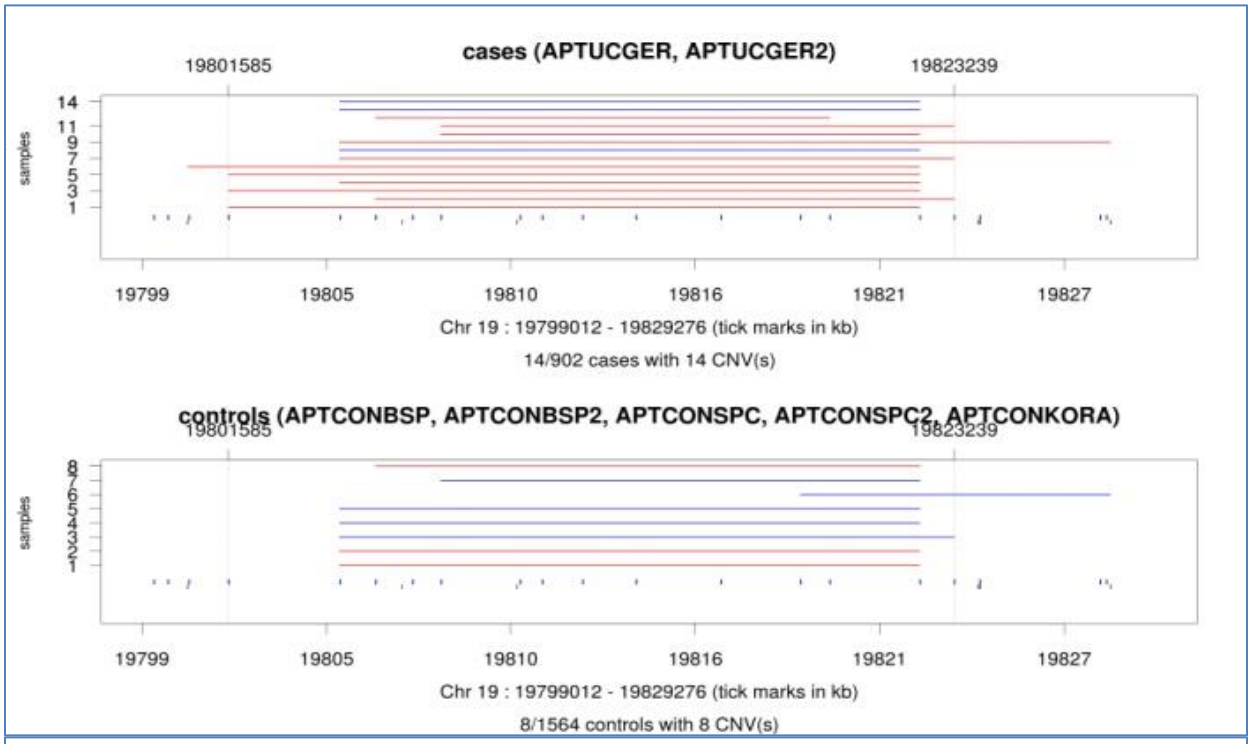

cases (APTUCGER, APTUCGER2)
108369095
samples
3
2
1
DDX10
108176
108218
108259
108301
108342
108384
Chr 11 : 108175862 - 108404976 (tick marks in kb)
3/902 cases with 3 CNV(s)
controls (APTCONBSP, APTCONBSP2, APTCONSPC, APTCONSPC2, APTCONKORA)
108369095
DDX10
108176
108218
108259
108301
108342
108384
Chr 11 : 108175862 - 108404976 (tick marks in kb)
0/1564 controls with 0 CNV(s)


cases (APTUCGER, APTUCGER2)
23991650
samples
3
2
1
PAQR4
23960
23971
23982
23994
24005
24016
Chr 13 : 23959634 - 24021752 (tick marks in kb)
3/902 cases with 3 CNV(s)
controls (APTCONBSP, APTCONBSP2, APTCONSPC, APTCONSPC2, APTCONKORA)
23991650
samples
5
4
3
2
1
PAQR4
23960
23971
23982
23994
24005
24016
Chr 13 : 23959634 - 24021752 (tick marks in kb)
5/1564 controls with 5 CNV(s)


cases (APTUCGER, APTUCGER2)
94784155
94797183
samples
94730
94749
94768
94788
94807
94826
Chr 13 : 94730061 - 94835628 (tick marks in kb)
8/902 cases with 8 CNV(s)
controls (APTCONBSP, APTCONBSP2, APTCONSPC, APTCONSPC2, APTCONKORA)
94784155
94797183
94730
94749
94768
94788
94807
94826
Chr 13 : 94730061 - 94835628 (tick marks in kb)
7/1564 controls with 8 CNV(s)


cases (APTUCGER, APTUCGER2)
21768464
22020331
samples
21524
21624
21724
21823
21923
22023
Chr 14 : 21523831 - 22073187 (tick marks in kb)
20/902 cases with 24 CNV(s)
controls (APTCONBSP, APTCONBSP2, APTCONSPC, APTCONSPC2, APTCONKORA)
21768464
22020331
21524
21624
21724
21823
21923
22023
Chr 14 : 21523831 - 22073187 (tick marks in kb)
2/1564 controls with 2 CNV(s)


cases (APTUCGER, APTUCGER2)
69281471
samples
69037
69140
69244
69348
69451
69555
Chr 18 : 69036866 - 69606766 (tick marks in kb)
4/902 cases with 4 CNV(s)
controls (APTCONBSP, APTCONBSP2, APTCONSPC, APTCONSPC2, APTCONKORA)
69281471
69037
69140
69244
69348
69451
69555
Chr 18 : 69036866 - 69606766 (tick marks in kb)
2/1564 controls with 2 CNV(s)


cases (APTUCGER, APTUCGER2)
19801585
19823239
samples
14
11
9
7
5
3
1
19799
19805
19810
19816
19821
19827
Chr 19 : 19799012 - 19829276 (tick marks in kb)
14/902 cases with 14 CNV(s)
controls (APTCONBSP, APTCONBSP2, APTCONSPC, APTCONSPC2, APTCONKORA)
19801585
19823239
samples
8
19799
19805
19810
19816
19821
19827
Chr 19 : 19799012 - 19829276 (tick marks in kb)
8/1564 controls with 8 CNV(s)

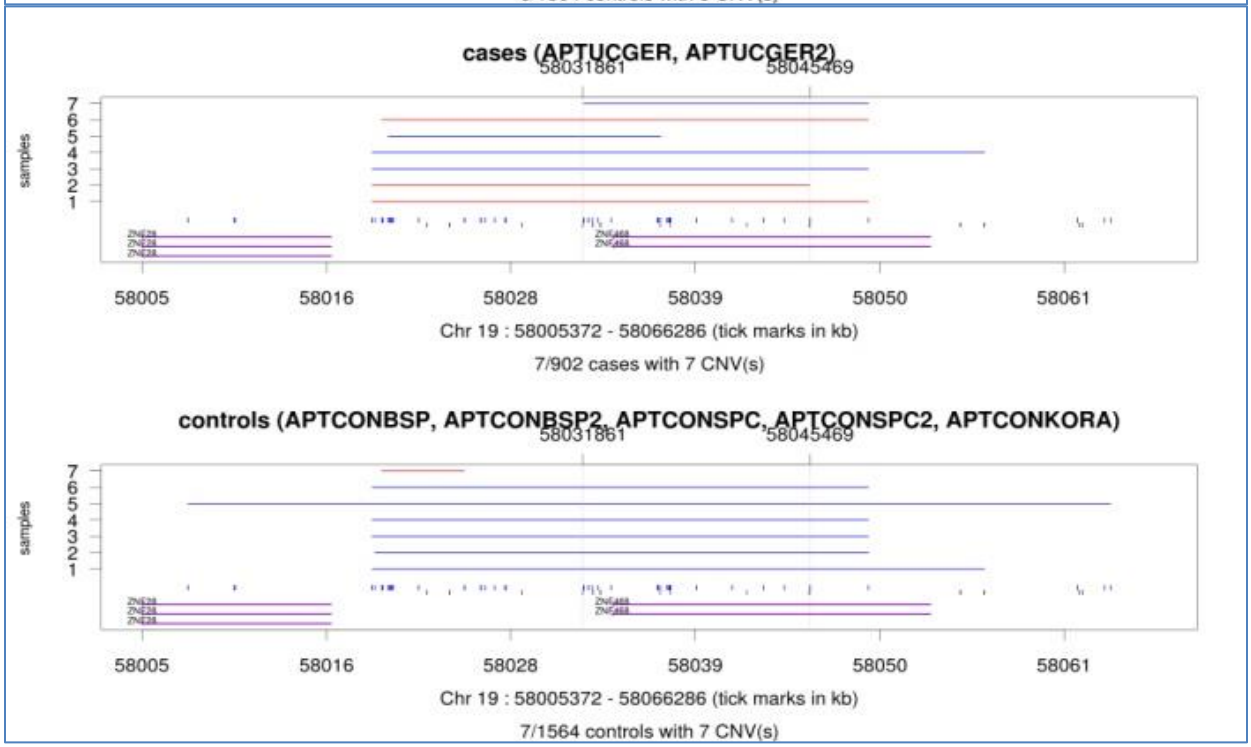
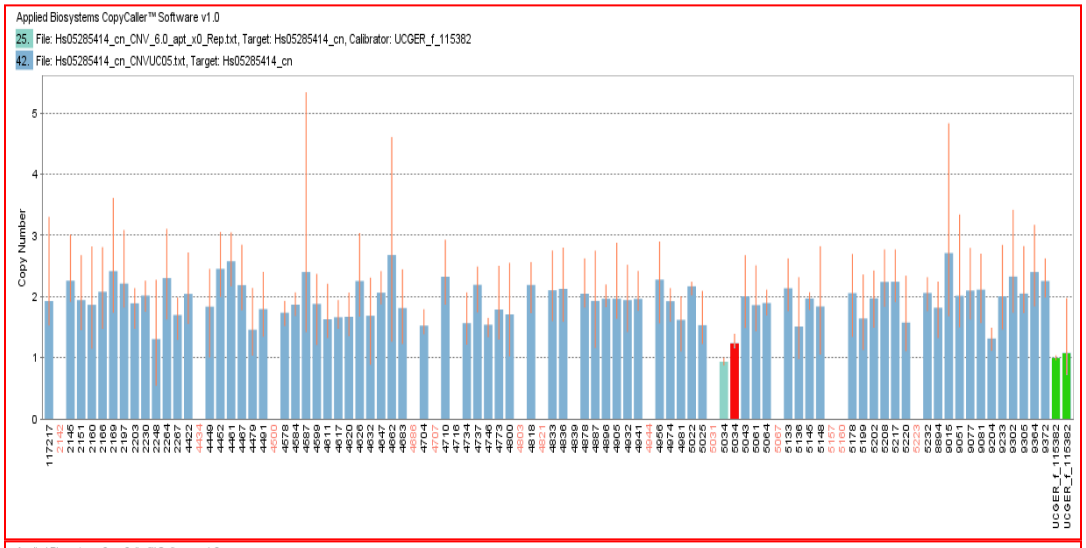
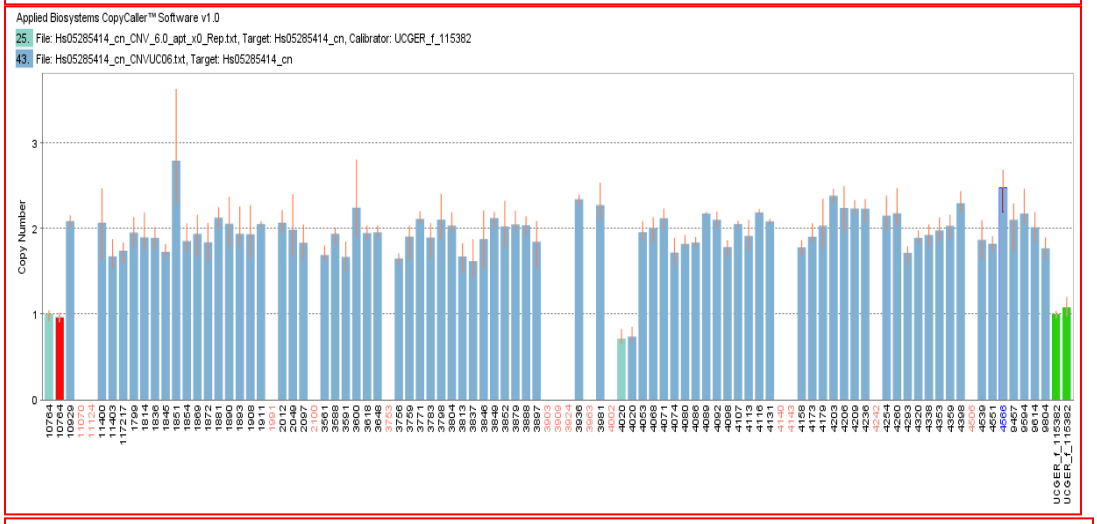
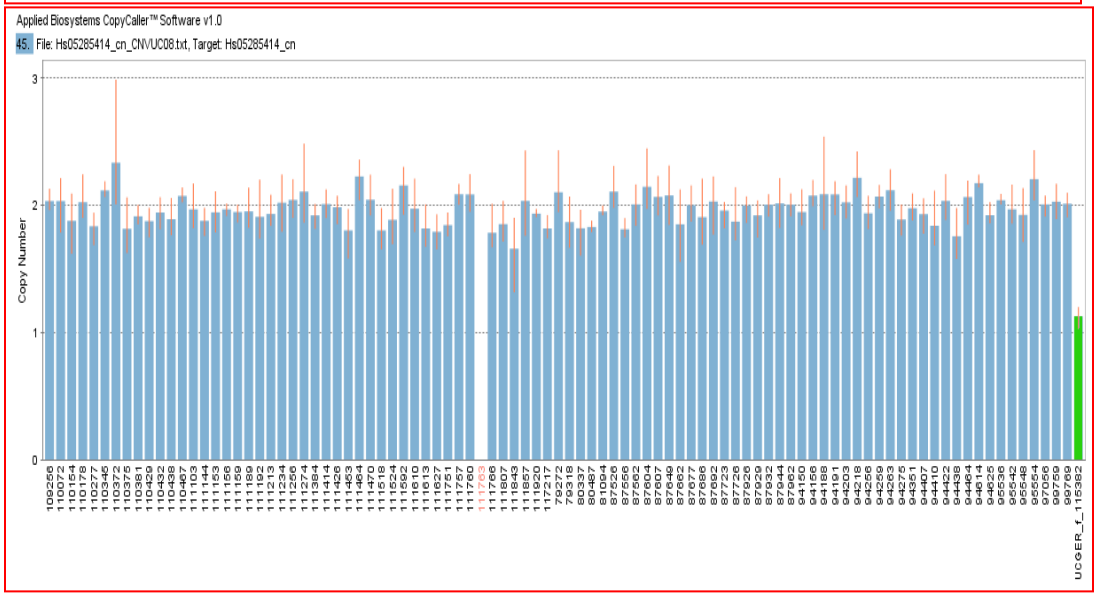
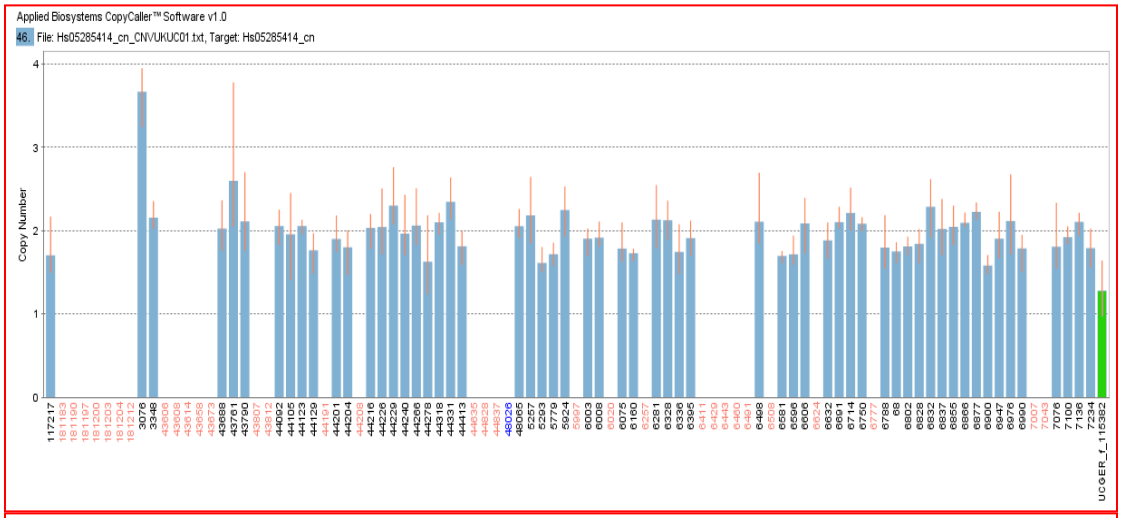

cases (APTUCGER, APTUCGER2)
58031861
58045469
samples
7
6
5
4
3
2
1
58005
58016
58028
58039
58050
58061
Chr 19 : 58005372 - 58066286 (tick marks in kb)
7/902 cases with 7 CNV(s)
controls (APTCONBSP, APTCONBSP2, APTCONSPC, APTCONSPC2, APTCONKORA)
58031861
58045469
samples
7
6
5
4
3
2
1
58005
58016
58028
58039
58050
58061
Chr 19 : 58005372 - 58066286 (tick marks in kb)
7/1564 controls with 7 CNV(s)

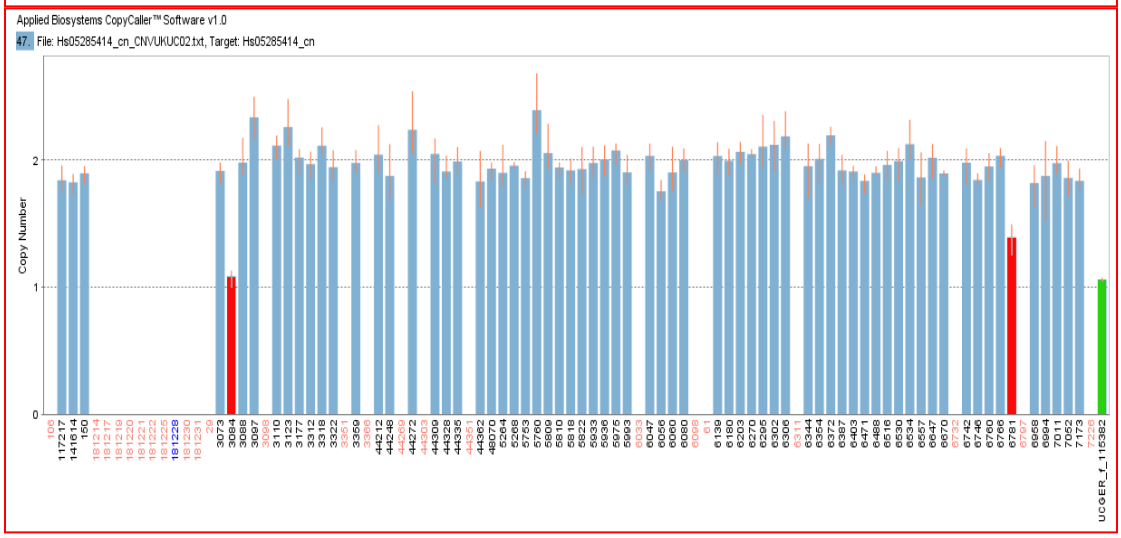
Table 6.2 Independent Replication of Deletion 13q32.1with TaqMan®CNV assays. Each graph of the table shows the result of real time PCR for oneTaqMan® plate. Graphs with red frame are for case plates and blue frames from control plates. Deletion carriers are highlighted by red bars. Samples with green bars are calibrator samples, which are samples with confirmed deletion. These calibrator samples are used to identify the CNV carriers within the replication sample set. If there is a second bar next to a red bar with the same identifier, then it is just a technical replicate. This was sometimes done due to the noisy nature of the data. The two technical replicates of the fourth graph (id 4020) show both copy number state one. Nevertheless this sample was not counted as deletion carrier, because it shows copy number state one for every TaqMan® assay that ran on this plate. TaqMan®quality control was based on confidence greater or equal to 94%.

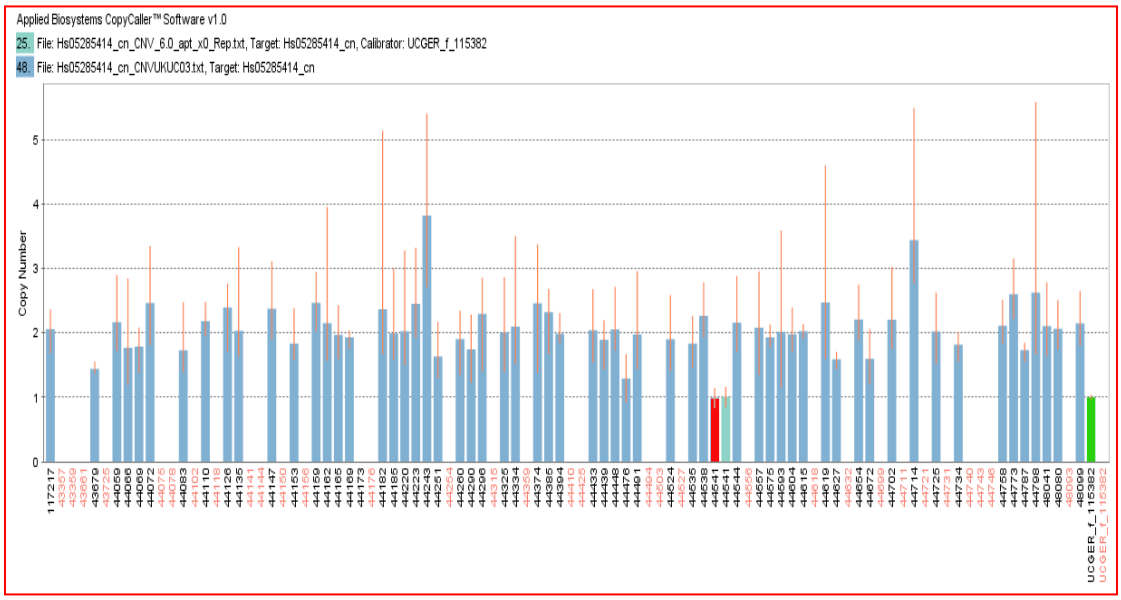

Applied Biosystems CopyCaller™ Software v1.0
25. File: Hs05285414_cn_CNV_6.0_apt_x0_Rep.txt, Target: Hs05285414_cn, Calibrator: UCGER_f_115382
42. File: Hs05285414_cn_CNVUC05.txt, Target: Hs05285414_cn
Copy Number
UCGER_f_115382
UCGER_f_115382

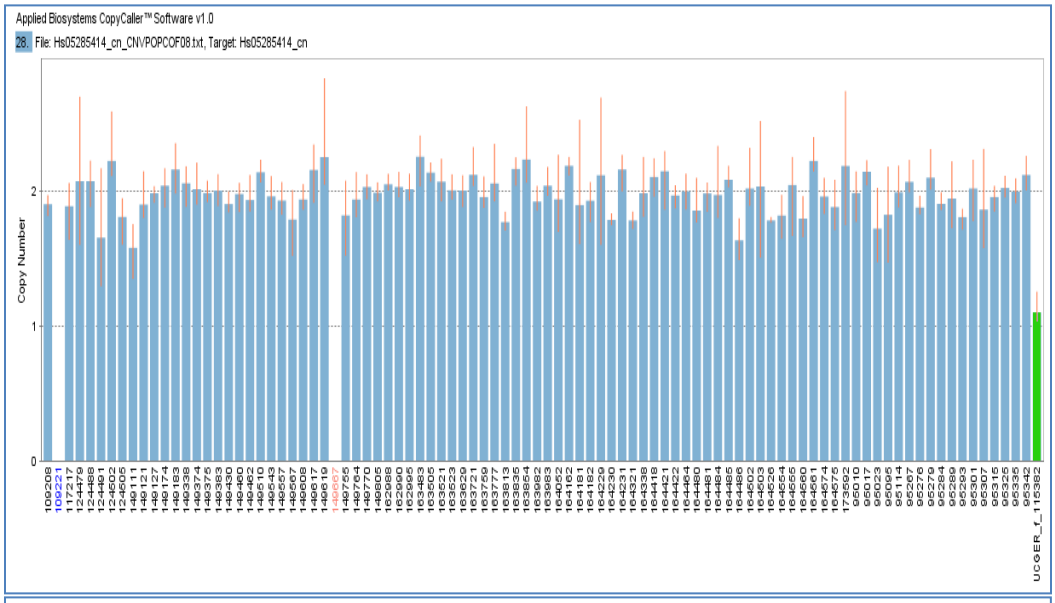

Applied Biosystems CopyCaller™ Software v1.0
25. File: Hs05285414_cn_CNV_6.0_apt_x0_Rep.txt, Target: Hs05285414_cn, Calibrator: UCGER_f_115382
43. File: Hs05285414_cn_CNVUC06.txt, Target: Hs05285414_cn
Copy Number
UCGER_f_115382
UCGER_f_115382

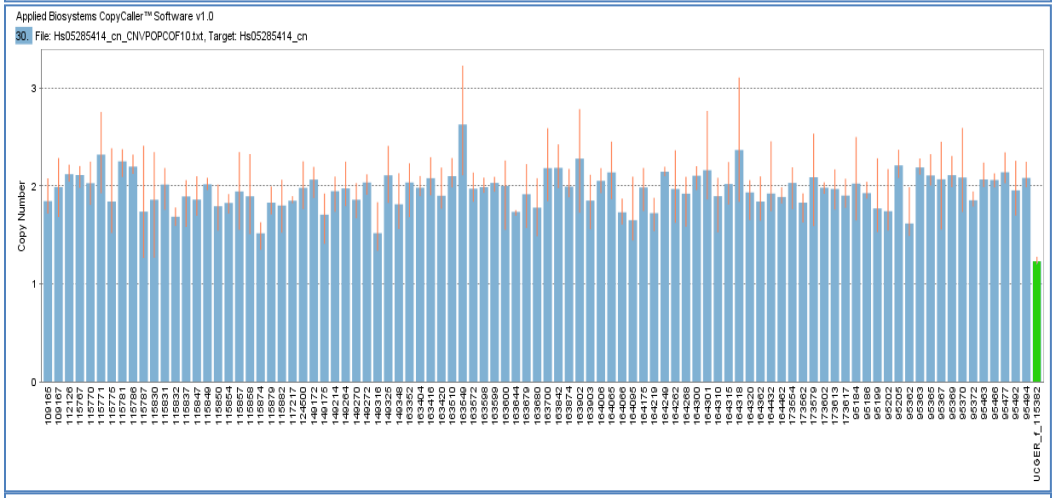

Applied Biosystems CopyCaller™ Software v1.0
45. File: Hs05285414_cn_CNVUC08.txt, Target: Hs05285414_cn
Copy Number
UCGER_f_115382

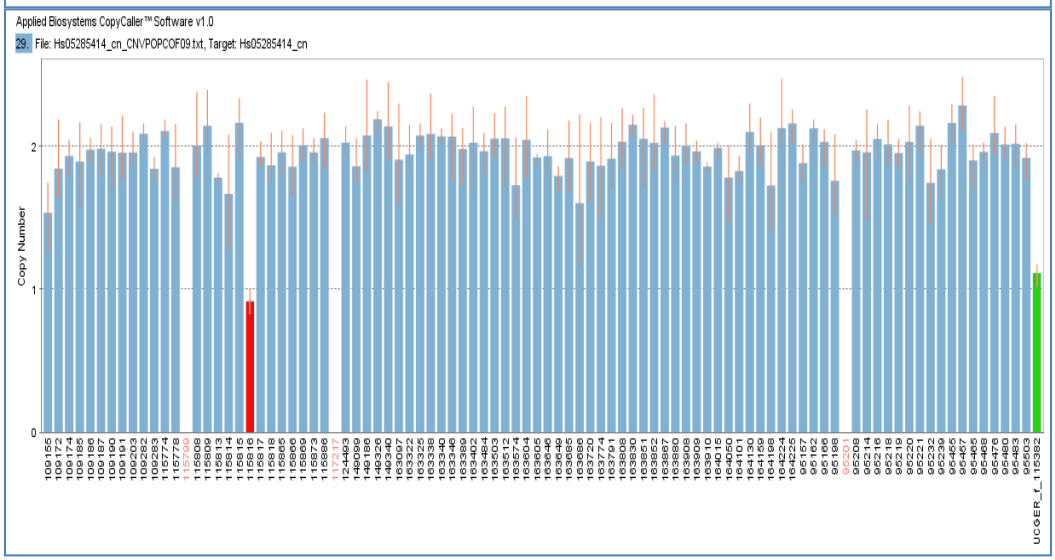

Applied Biosystems CopyCaller™ Software v1.0
46. File: Hs05285414_cn_CNVUKUC01.txt, Target: Hs05285414_cn
Copy Number
UCGER_f_115382

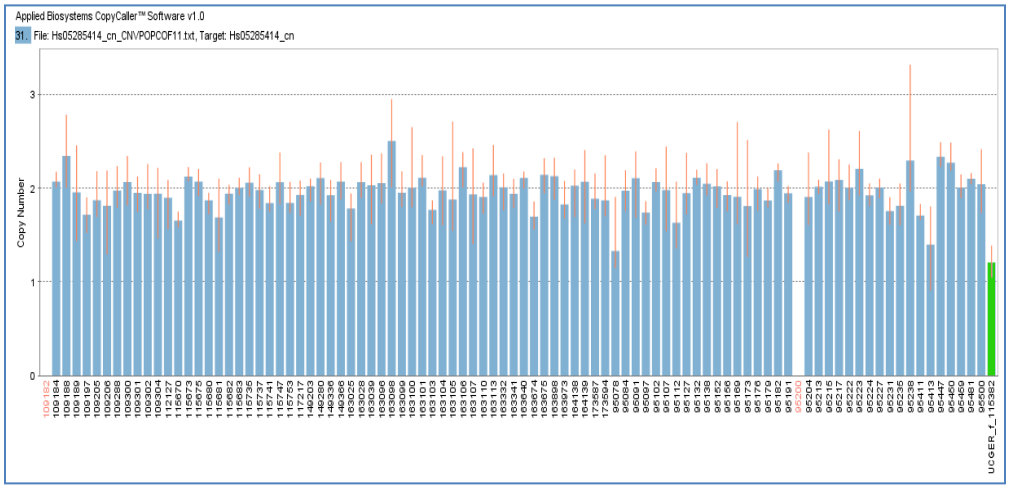

Applied Biosystems CopyCaller™ Software v1.0
47. File: Hs05285414_cn_CNVUKUC02.txt, Target: Hs05285414_cn
Copy Number
UCGER_f_115382

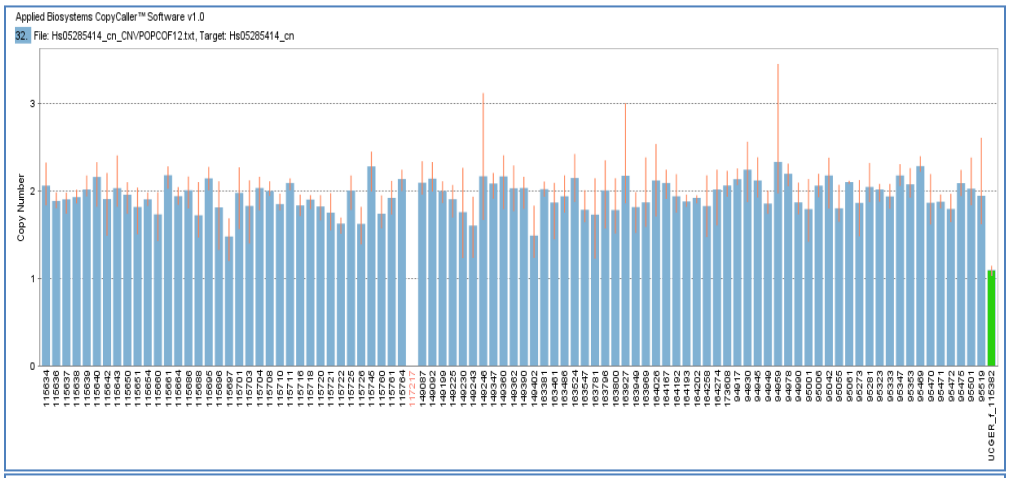

Applied Biosystems CopyCaller™ Software v1.0
25. File: Hs05285414_cn_CNV_6.0_apt_x0_Rep.txt, Target: Hs05285414_cn, Calibrator: UCGER_f_115382
48. File: Hs05285414_cn_CNVUKUC03.txt, Target: Hs05285414_cn
Copy Number
UCGER_f_115382
UCGER_f_115382

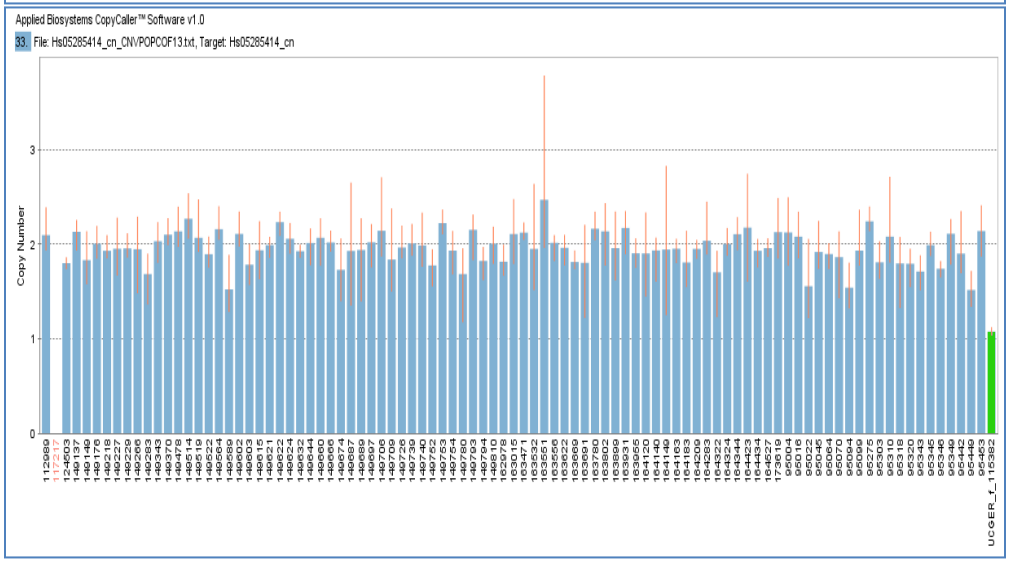

Applied Biosystems CopyCaller™ Software v1.0
28. File: Hs05285414_cn_CNVPOPCOF08.txt, Target: Hs05285414_cn
Copy Number
UCSER_f_115382

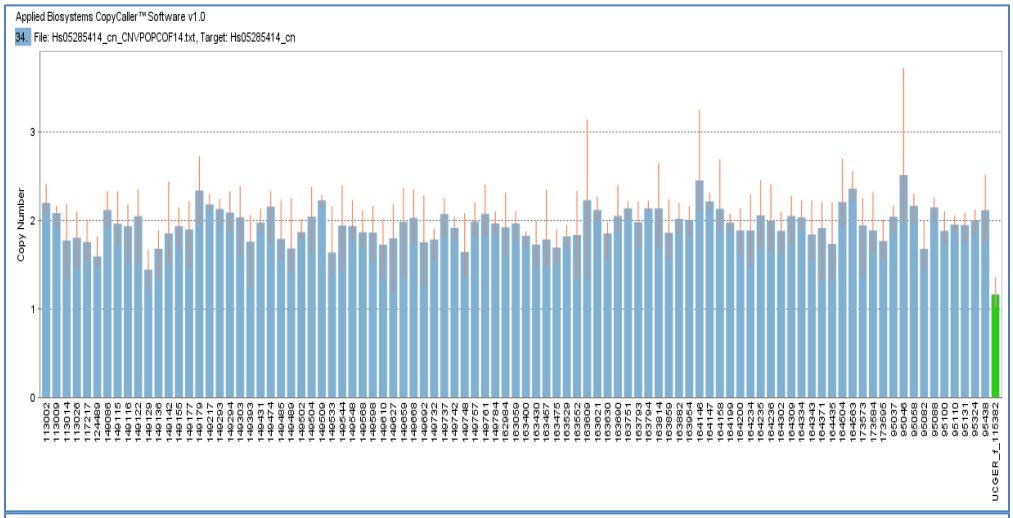

Applied Biosystems CopyCaller™ Software v1.0
30. File: Hs05285414_cn_CNVPOPCOF10.txt, Target: Hs05285414_cn
Copy Number
UCSER_f_115382

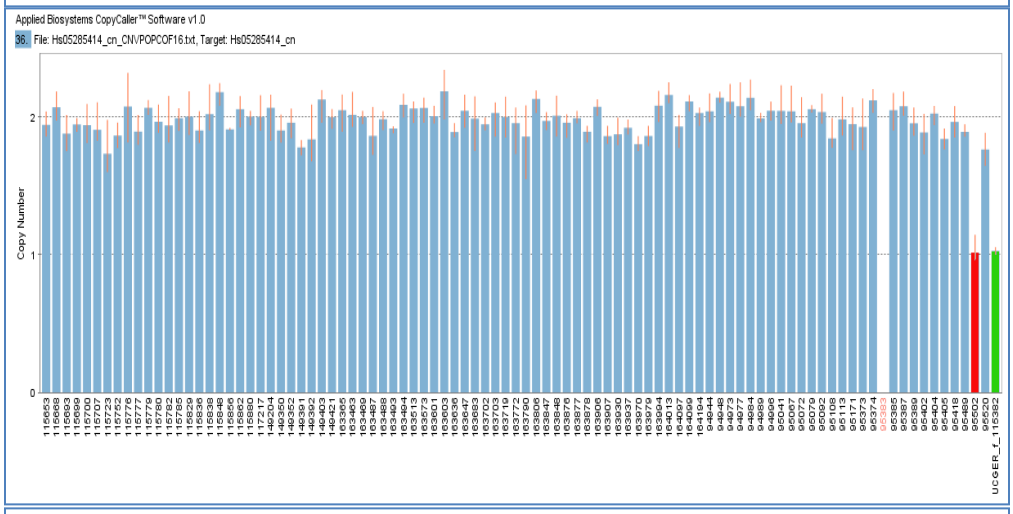

Applied Biosystems CopyCaller™ Software v1.0
29. File: Hs05285414_cn_CNVPOPCOF09.txt, Target: Hs05285414_cn
Copy Number
UCSER_f_115382

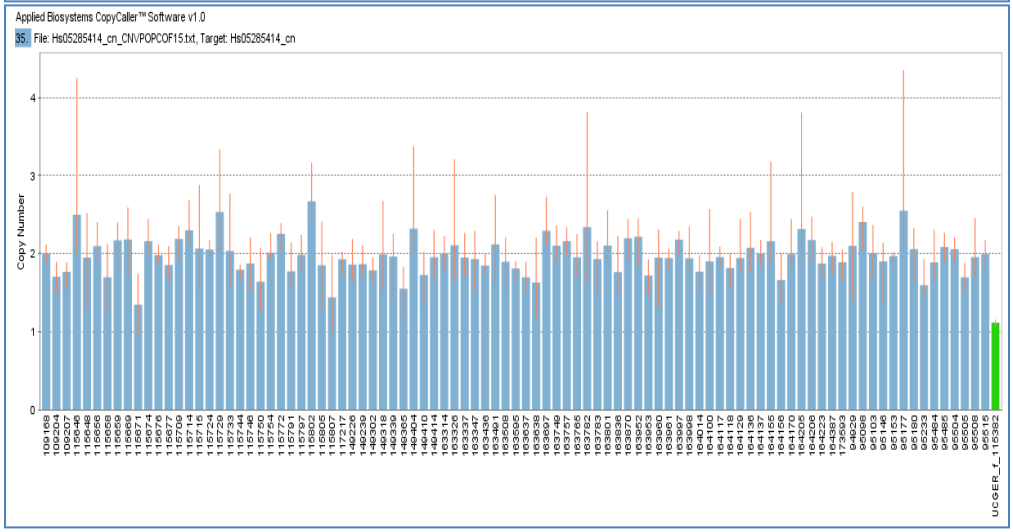

Applied Biosystems CopyCaller™ Software v1.0
31. File: Hs05285414_cn_CNVPOPCOF11.txt, Target: Hs05285414_cn
Copy Number
UCGER_f_115382


Applied Biosystems CopyCaller™ Software v1.0
32. File: Hs05285414_cn_CNVPOPCOF12.txt, Target: Hs05285414_cn
Copy Number
UCGER_f_115382

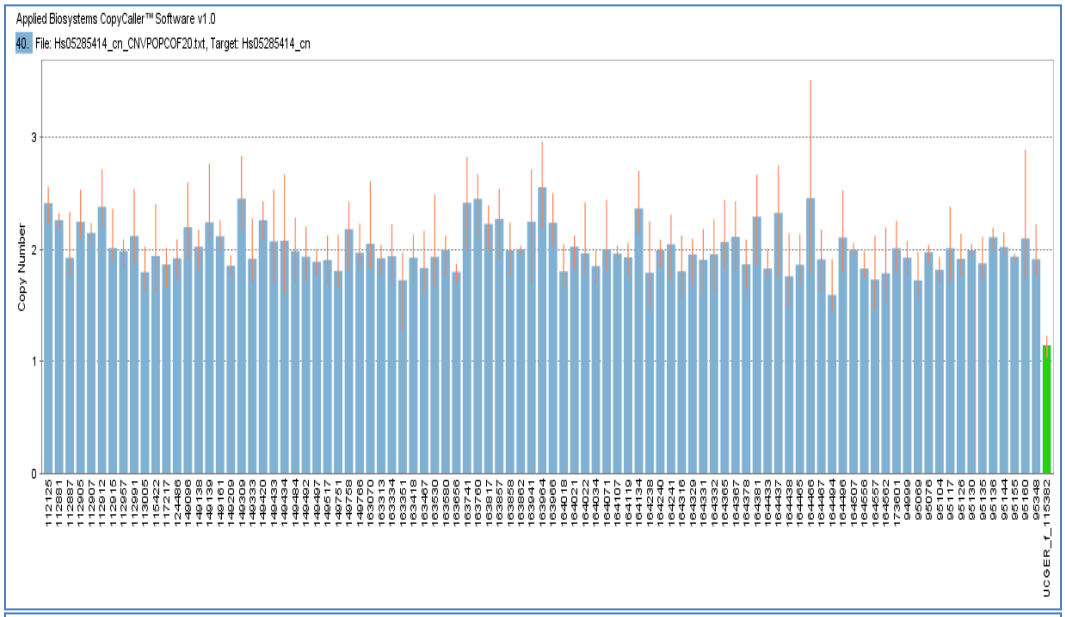

Applied Biosystems CopyCaller™ Software v1.0
33. File: Hs05285414_cn_CNVPOPCOF13.txt, Target: Hs05285414_cn
Copy Number
UCGER_f_115382


Applied Biosystems CopyCaller™ Software v1.0
34. File: Hs05285414_cn_CNVPOPCOF14.txt, Target: Hs05285414_cn
Copy Number
UCGER_f_115382


Applied Biosystems CopyCaller™ Software v1.0
36. File: Hs05285414_cn_CNVPOPCOF16.txt, Target: Hs05285414_cn
Copy Number
UCGER_f_115382


Applied Biosystems CopyCaller™ Software v1.0
35. File: Hs05285414_cn_CNVPOPCOF15.txt, Target: Hs05285414_cn
Copy Number
UCGER_f_115382


Applied Biosystems CopyCaller™ Software v1.0
38. File: Hs05285414_cn_CNVPOPCOF18.txt, Target: Hs05285414_cn
Copy Number
Applied Biosystems CopyCaller™ Software v1.0
39. File: Hs05285414_cn_CNVPOPCOF19.txt, Target: Hs05285414_cn
Copy Number
Applied Biosystems CopyCaller™ Software v1.0
37. File: Hs05285414_cn_CNVPOPCOF17.txt, Target: Hs05285414_cn
Copy Number


Applied Biosystems CopyCaller™ Software v1.0
40. File: Hs05285414_cn_CNVPOPCOF20.txt, Target: Hs05285414_cn
Copy Number
0
1
2
3
UCGER_7_115382

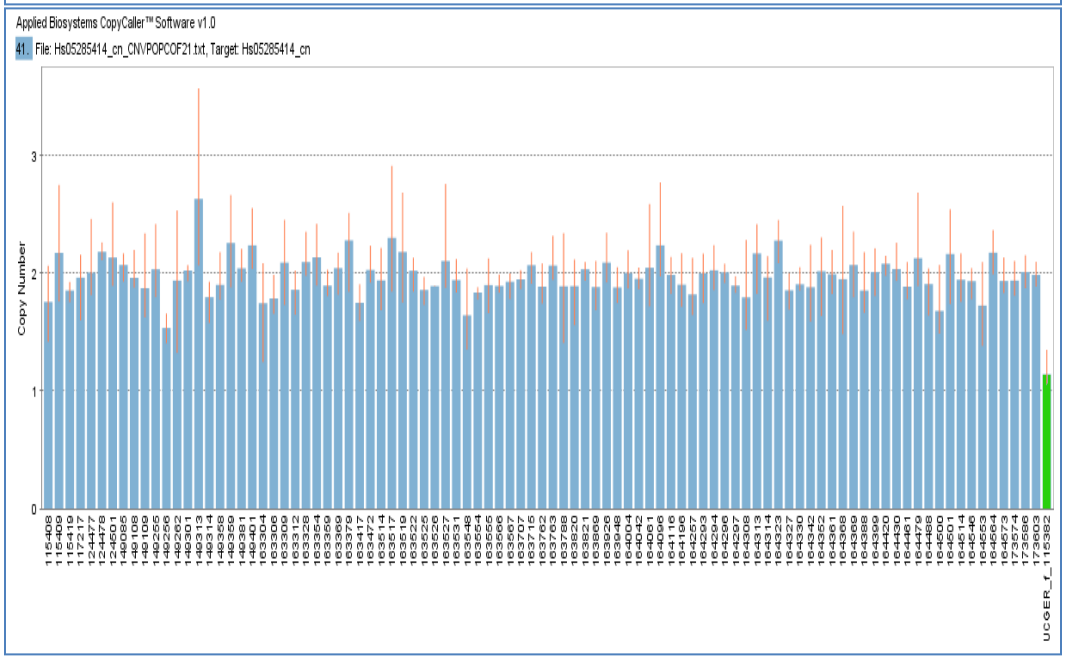
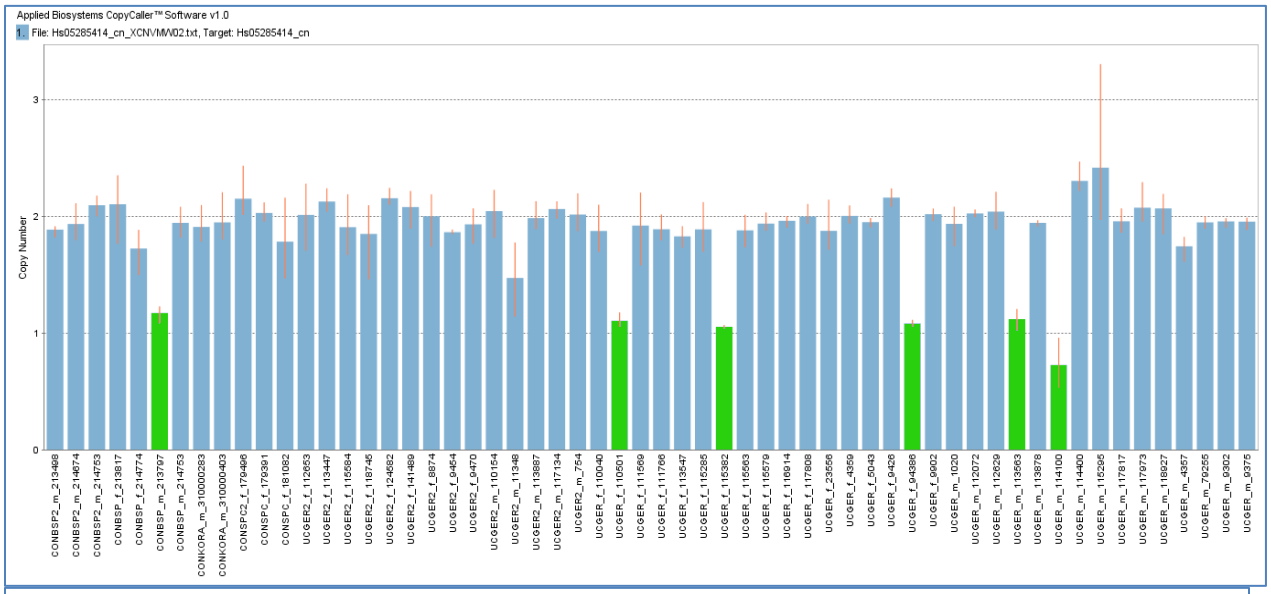
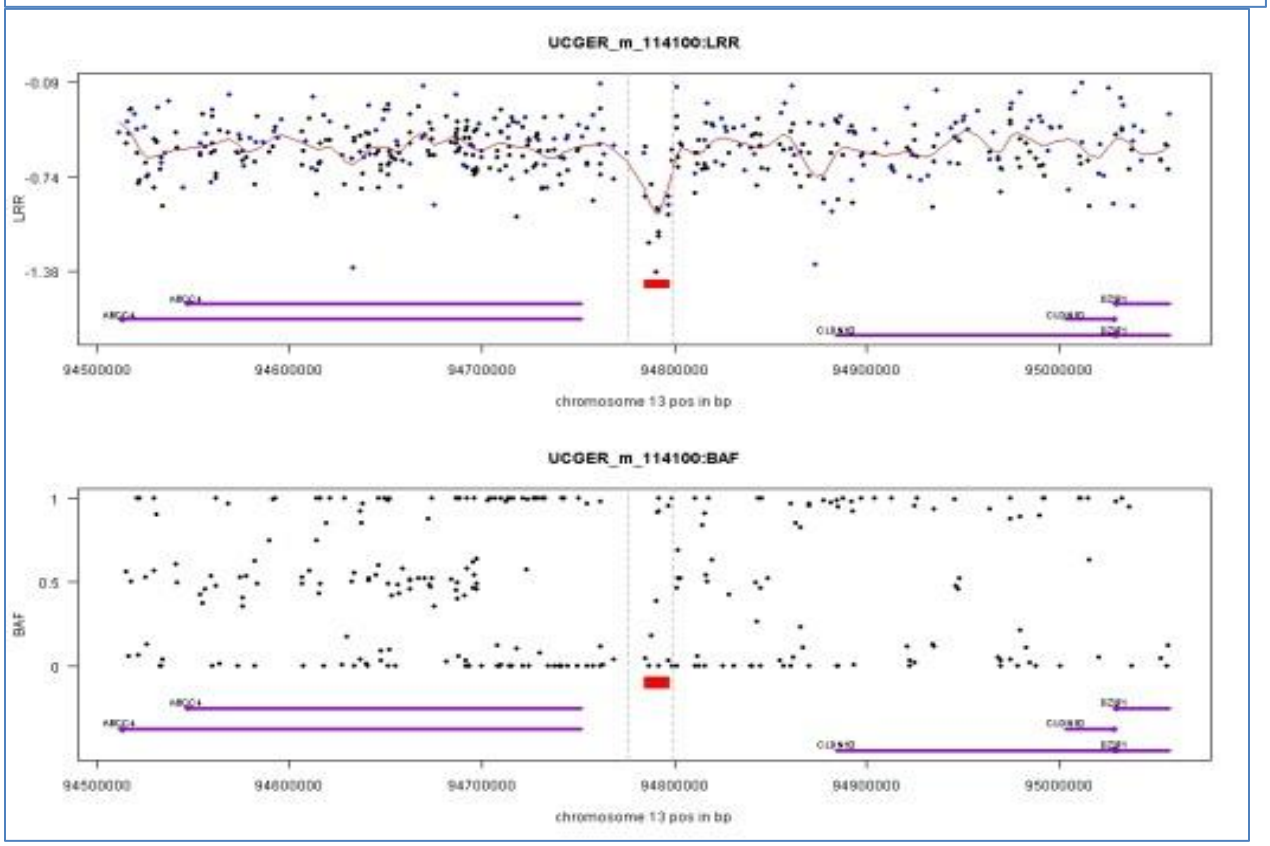
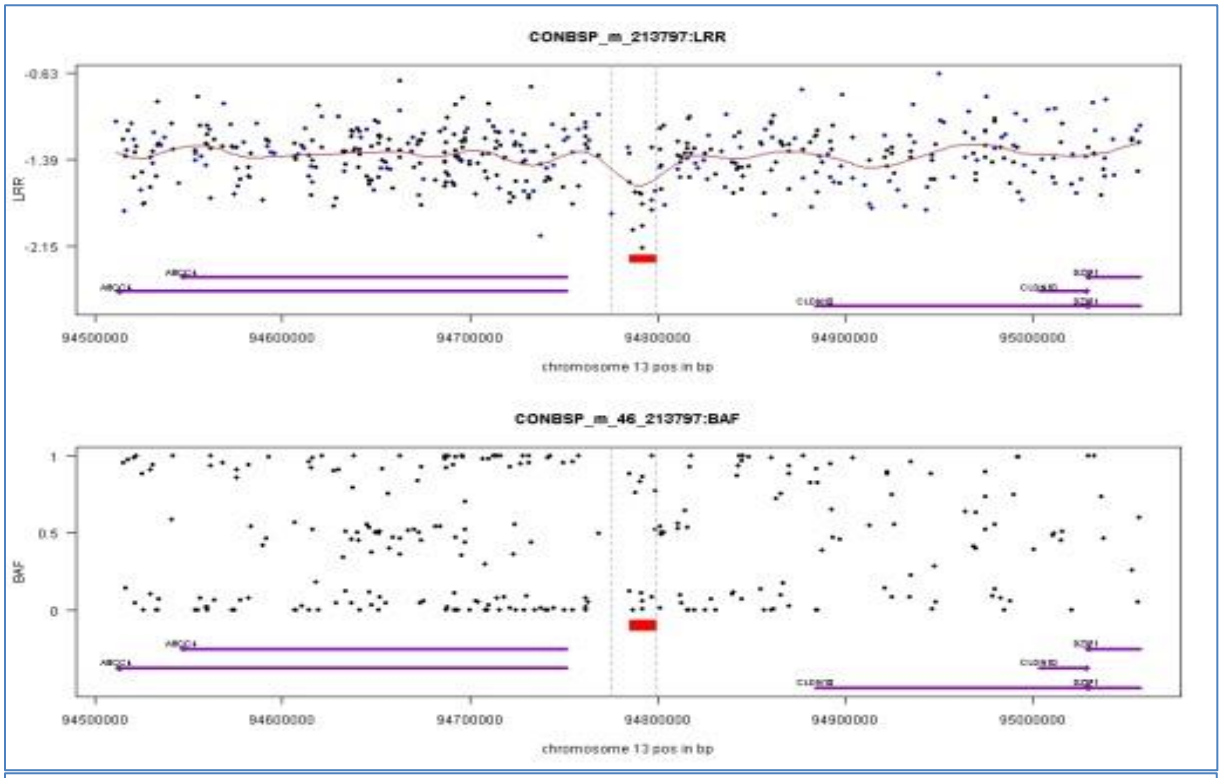
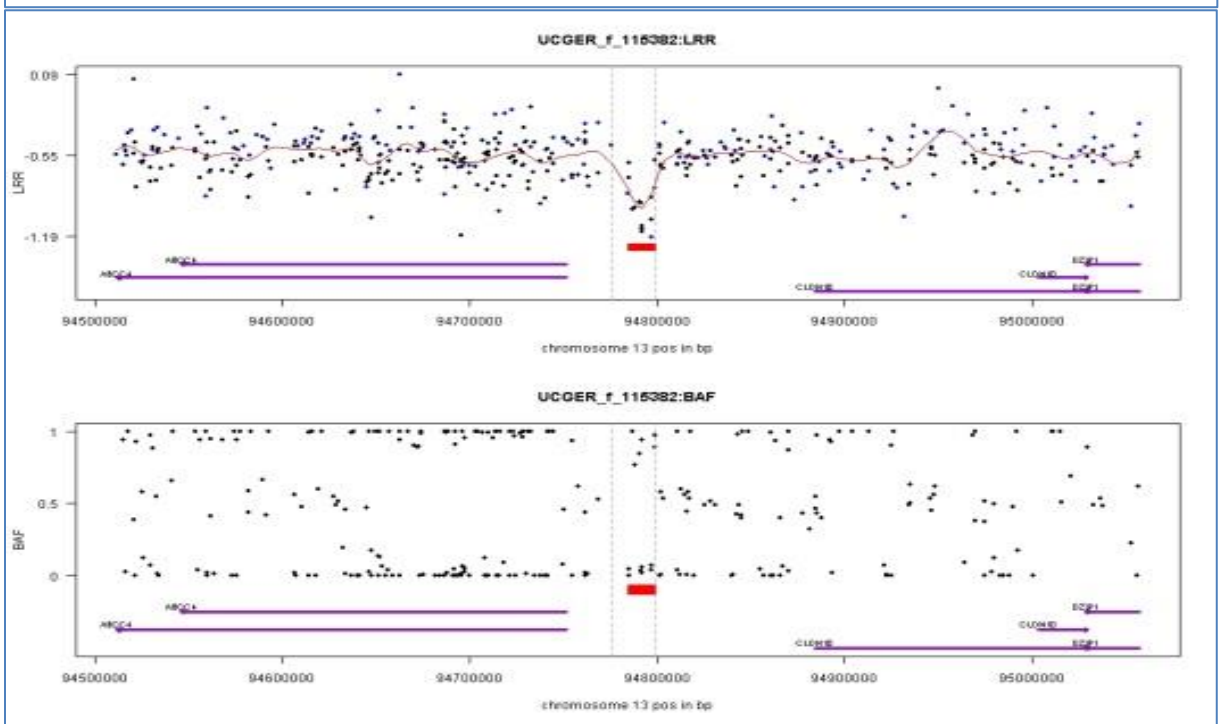

Applied Biosystems CopyCaller™ Software v1.0
41. File: Hs05285414_cn_CNVPOPCOF21.txt, Target: Hs05285414_cn
Copy Number
0
1
2
3
UCGER_7_115382

Table 6.3 Technical validations of Del13q32.1 for the German discovery panel. Results of TaqMan®CNV assay is showed as the top illustration. While the majority of samples have copy number state two (blue bars). Green bars show deletion or duplication carriers The TaqMan®results show exact correlation with the predicted CNVs and raw data quality of the SNP array data is pretty good. For all deletions a loss of signal intensity can be seen corresponding with a loss of heterozygosity. **Raw data visualization of intensities**. In each picture LRR in the top pannel and B allele frequency (BAF) in the lower one. Non polymorphic probe sets are blue and SNP probesets black, a smoothed spline was added for all LRR plots. RefGene annotation is added with purple arrows. The red (deletion) or blue (duplication) horizontal bars visualize the redicted CNV.






CONBSP_m_213797:LRR
LRR
-0.83
-1.39
-2.15
94500000
94600000
94700000
94800000
94900000
95000000
chromosome 13 pos in bp
CONBSP_m_46_213797:BAF
BAF
1
0.5
0
94500000
94600000
94700000
94800000
94900000
95000000
chromosome 13 pos in bp


UCGER_f_115382:LRR
LRR
0.09
-0.55
-1.19
94500000
94600000
94700000
94800000
94900000
95000000
chromosome 13 pos in bp
UCGER_f_115382:BAF
BAF
1
0.5
0
94500000
94600000
94700000
94800000
94900000
95000000
chromosome 13 pos in bp

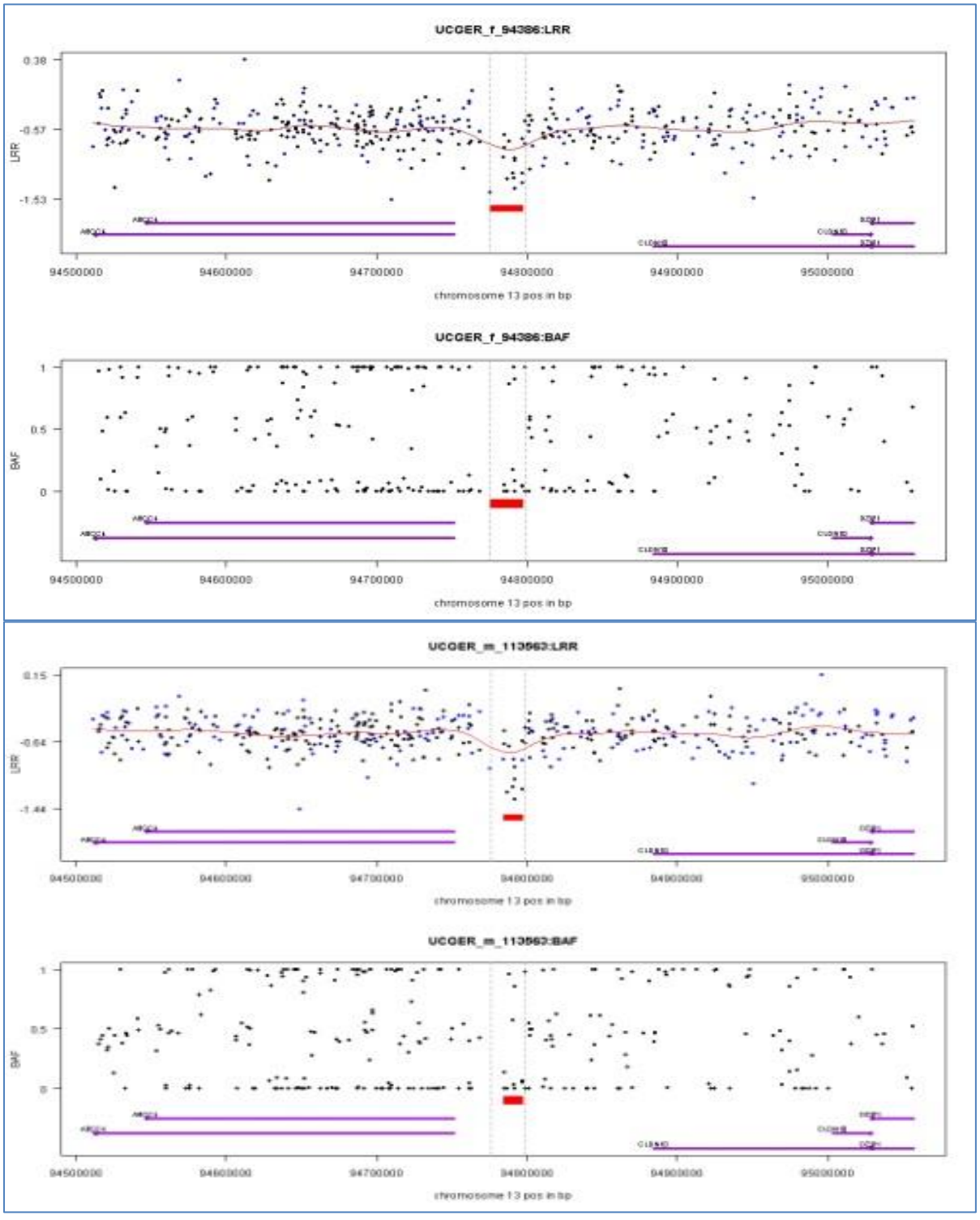
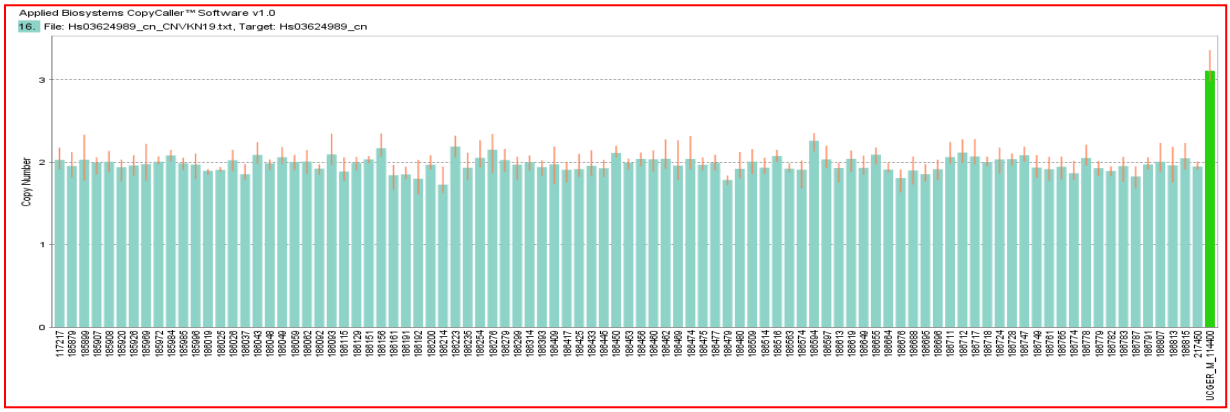
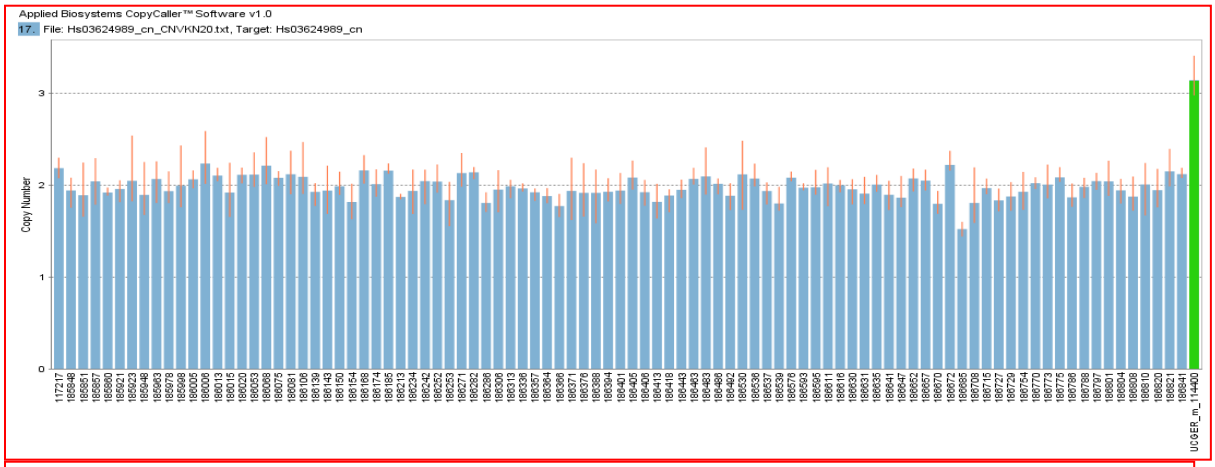
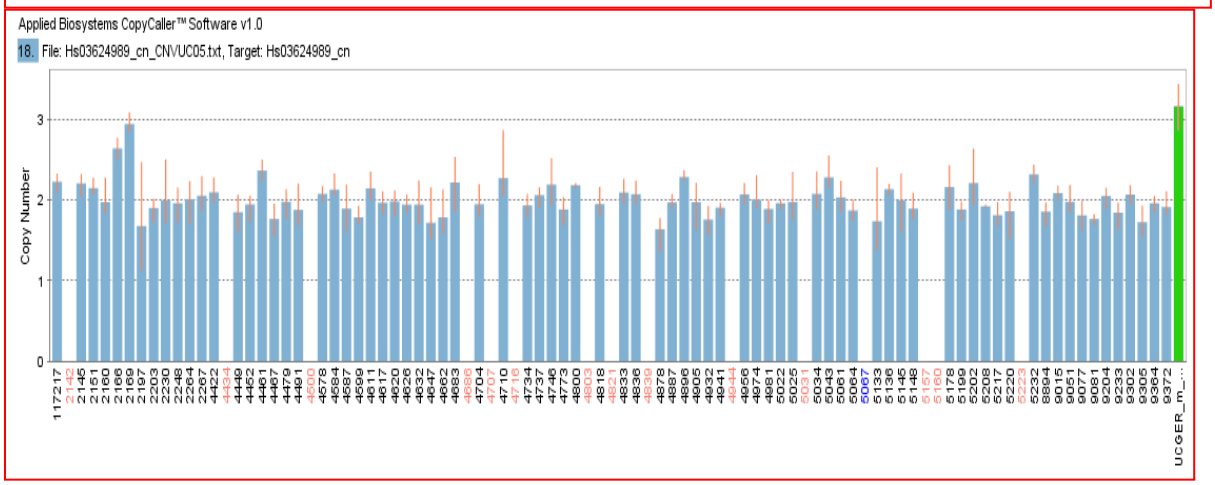
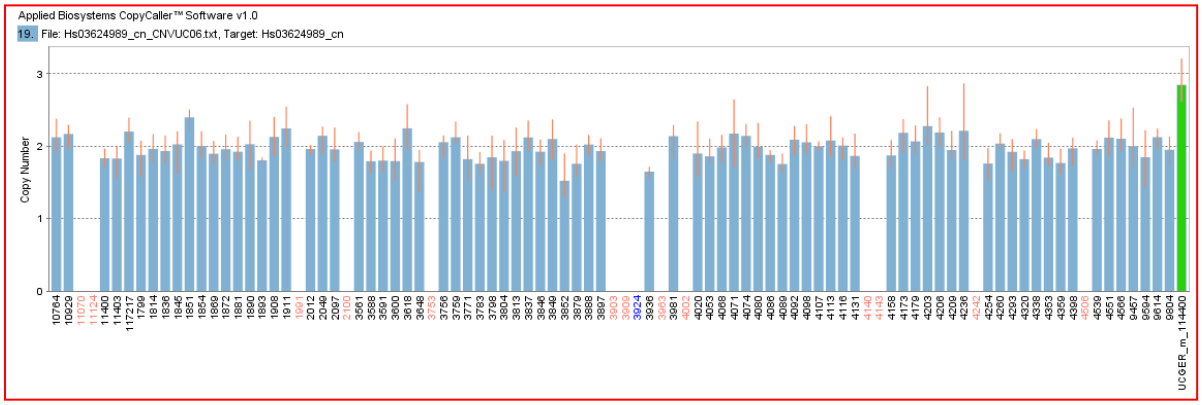

UCGER_f_94386:LRR
LRR
0.38
-0.57
-1.53
94500000
94600000
94700000
94800000
94900000
95000000
chromosome 13 pos in bp
UCGER_f_94386:BAF
BAF
1
0.5
0
94500000
94600000
94700000
94800000
94900000
95000000
chromosome 13 pos in bp
UCGER_m_113562:LRR
LRR
0.15
-0.64
-1.44
94500000
94600000
94700000
94800000
94900000
95000000
chromosome 13 pos in bp
UCGER_m_113562:BAF
BAF
1
0.5
0
94500000
94600000
94700000
94800000
94900000
95000000
chromosome 13 pos in bp

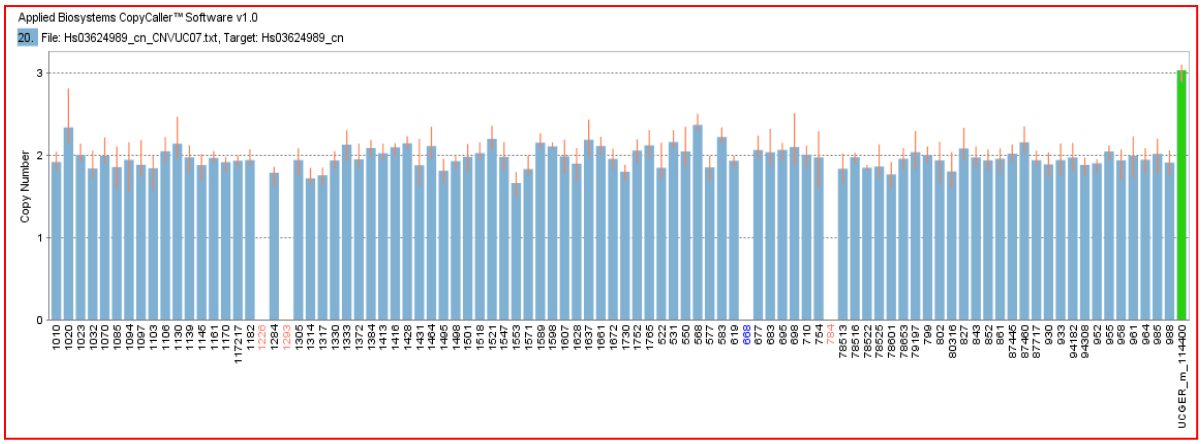
Table 6.4 Independent Replication of Duplication 7p22.1 with TaqMan®CNV assays. Each graph shows the result of one real time PCR TaqMan® plate, with the samples at the x-axis and the intensities at the y-axis. Graphs with red frame are from case plates, blue frames from control plates and green frames from technical replicates. Duplication carriers are highlighted by dark blue bars. Samples with green bars are calibrator samples, which are samples with confirmed duplication. These calibrator samples are used to identify the CNV carrier within the replication sample set. If there is a second bar next to a dark blue bar with the same identifier, then it is just a technical replicate. This was sometimes done due to the noisy nature of the data. TaqMan®quality control was based on confidence greater or equal to 94%.

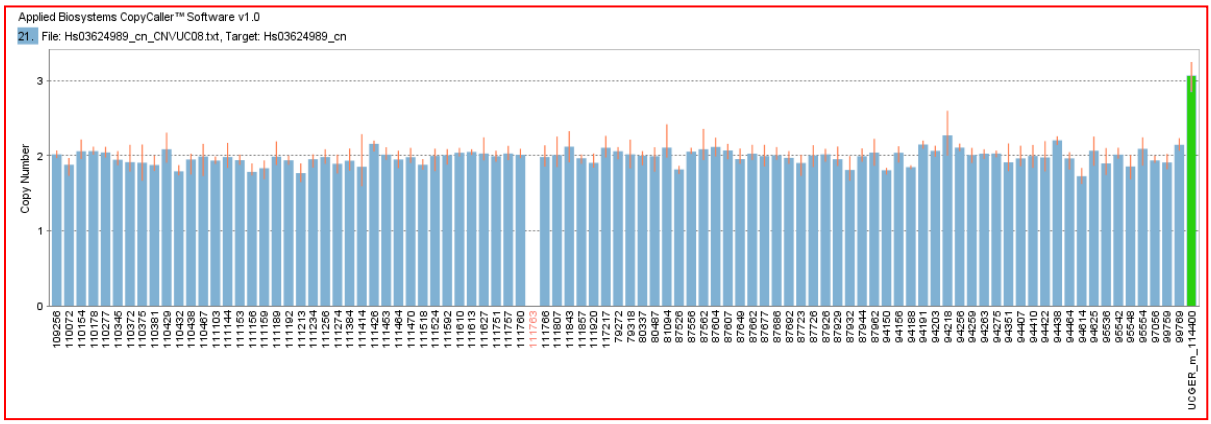


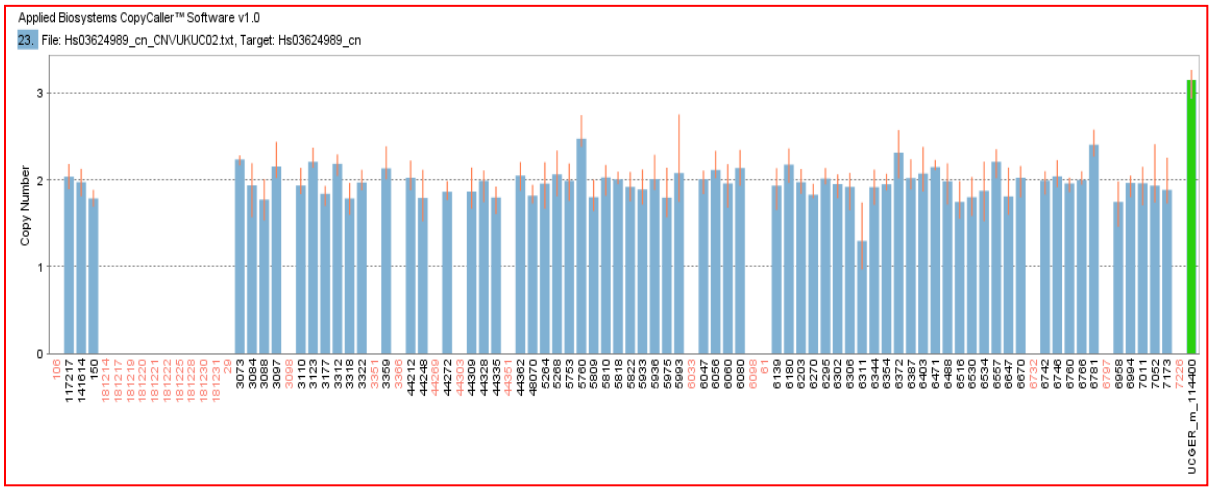


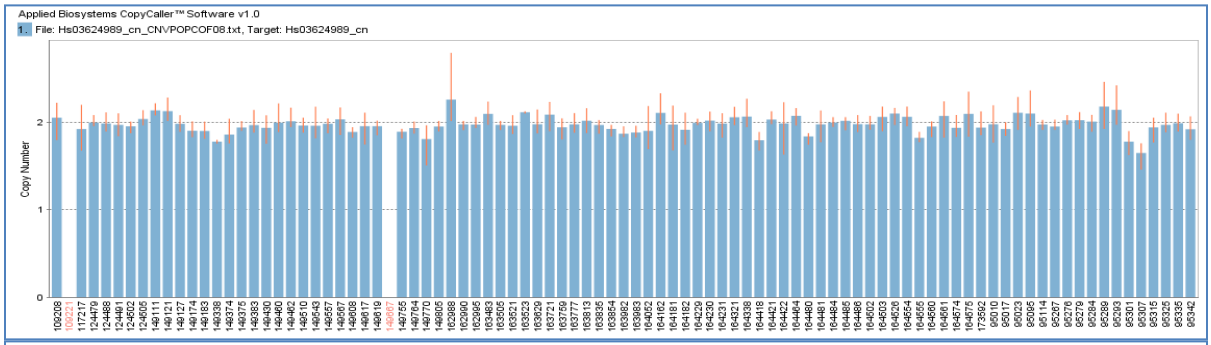


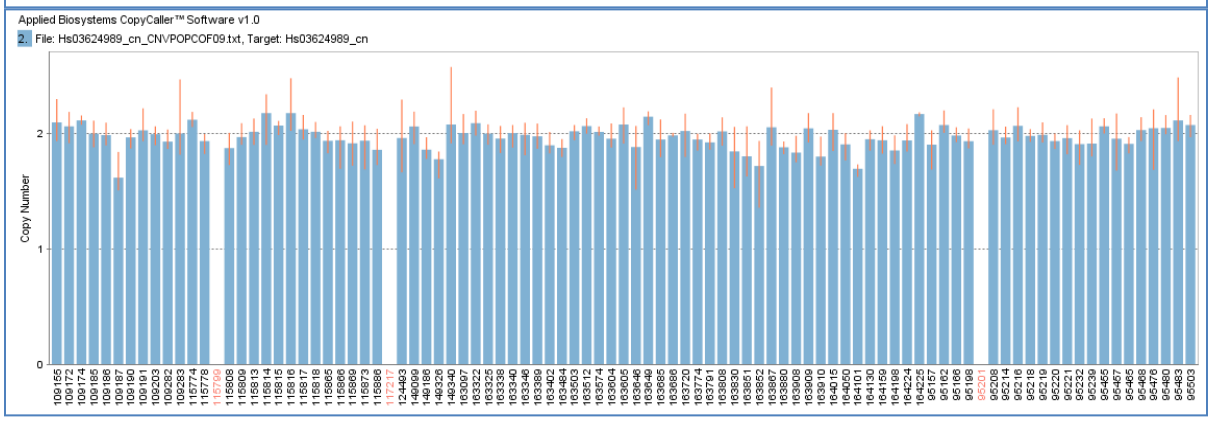

Applied Biosystems CopyCaller™ Software v1.0
19. File: Hs03624989_cn_CNVUC06.txt, Target: Hs03624989_cn
Copy Number
UCGER_m_114400

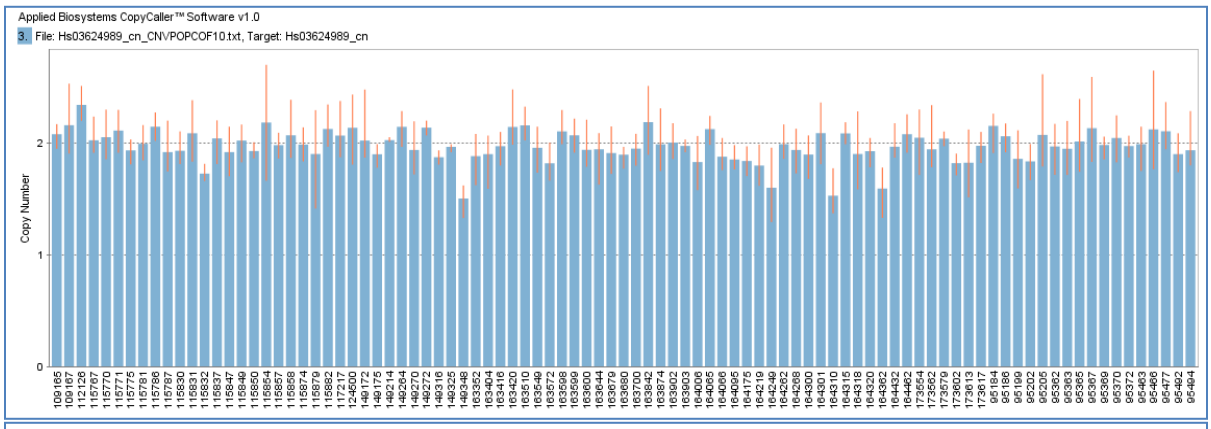

Applied Biosystems CopyCaller™ Software v1.0
20. File: Hs03624989_cn_CNVUC07.txt, Target: Hs03624989_cn
Copy Number
UCGER_m_114400

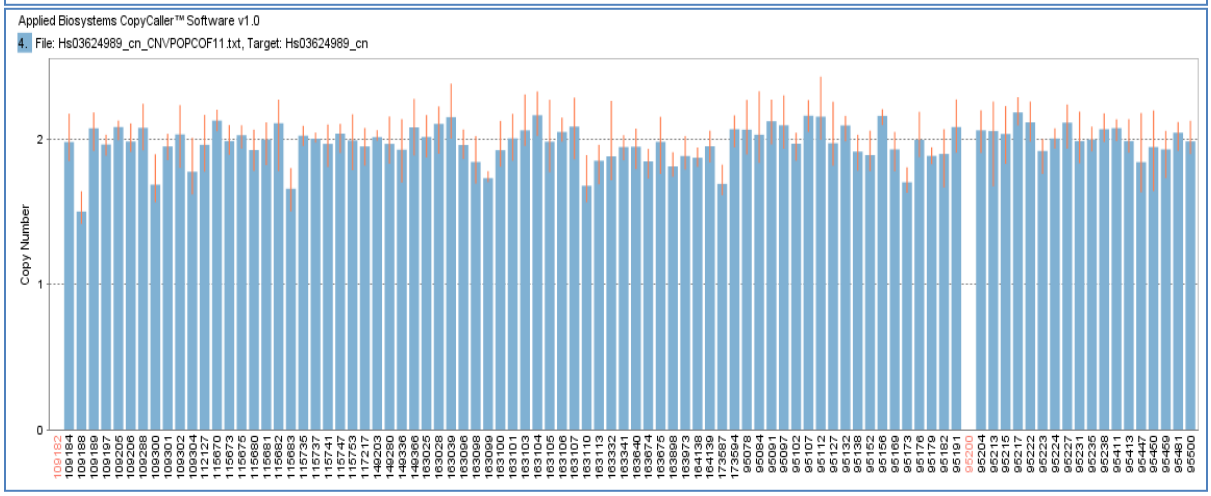

Applied Biosystems CopyCaller™ Software v1.0
21. File: Hs03624989_cn_CNVUC08.txt, Target: Hs03624989_cn
Copy Number
UCGER_m_114400


Applied Biosystems CopyCaller™ Software v1.0
23. File: Hs03624989_cn_CNVUKUC02.txt, Target: Hs03624989_cn
Copy Number
UCGER_m_114400


Applied Biosystems CopyCaller™ Software v1.0
1. File: Hs03624989_cn_CNVPOPCOF08.txt, Target: Hs03624989_cn
Copy Number


Applied Biosystems CopyCaller™ Software v1.0
2. File: Hs03624989_cn_CNVPOPCOF09.txt, Target: Hs03624989_cn
Copy Number


Applied Biosystems CopyCaller™ Software v1.0
3. File: Hs03624989_cn_CNVPOPCOF10.txt, Target: Hs03624989_cn
Copy Number


Applied Biosystems CopyCaller™ Software v1.0
4. File: Hs03624989_cn_CNVPOPCOF11.txt, Target: Hs03624989_cn
Copy Number


Applied Biosystems CopyCaller™ Software v1.0
5. File: Hs03624989_cn_CNVPOPCOF12.txt, Target: Hs03624989_cn
Copy Number
Applied Biosystems CopyCaller™ Software v1.0
6. File: Hs03624989_cn_CNVPOPCOF13.txt, Target: Hs03624989_cn
Copy Number
UCGER_m_114400
Applied Biosystems CopyCaller™ Software v1.0
7. File: Hs03624989_cn_CNVPOPCOF14.txt, Target: Hs03624989_cn
Copy Number
UCGER_m_114400
Applied Biosystems CopyCaller™ Software v1.0
7. File: Hs03624989_cn_CNVPOPCOF14.txt, Target: Hs03624989_cn
Copy Number
UCGER_m_114400

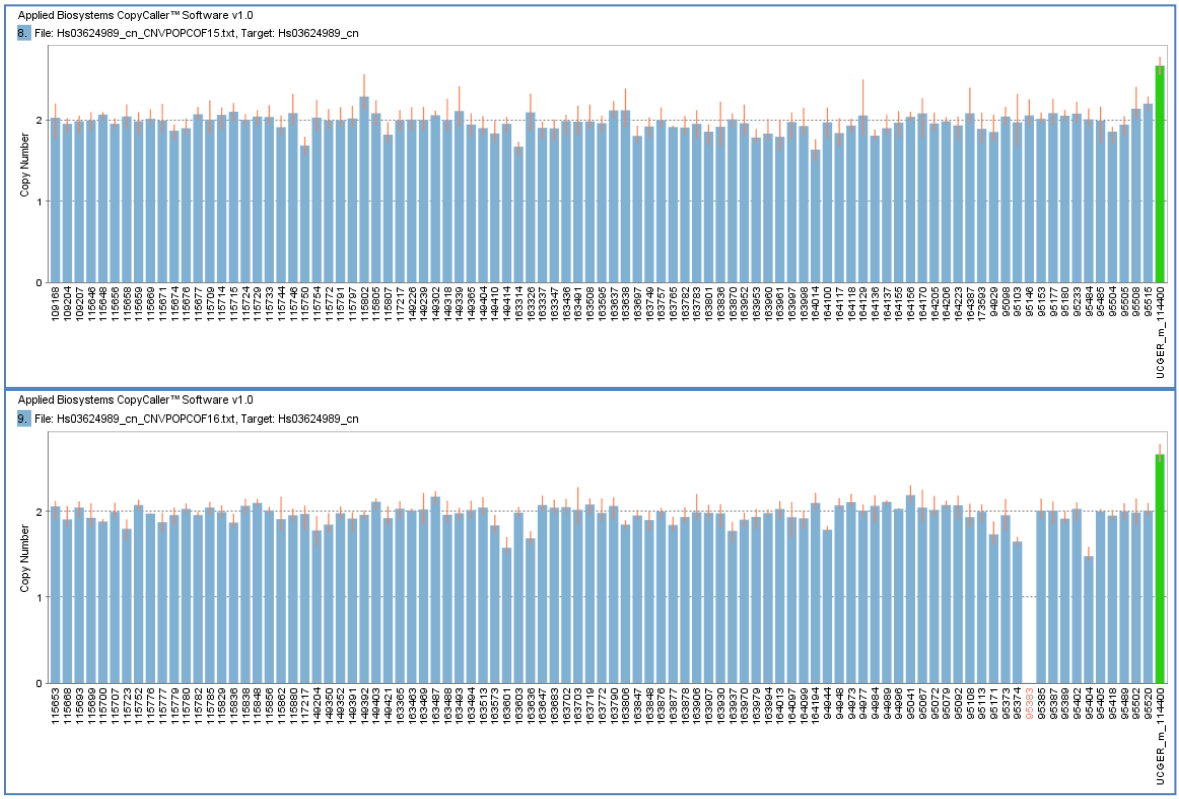
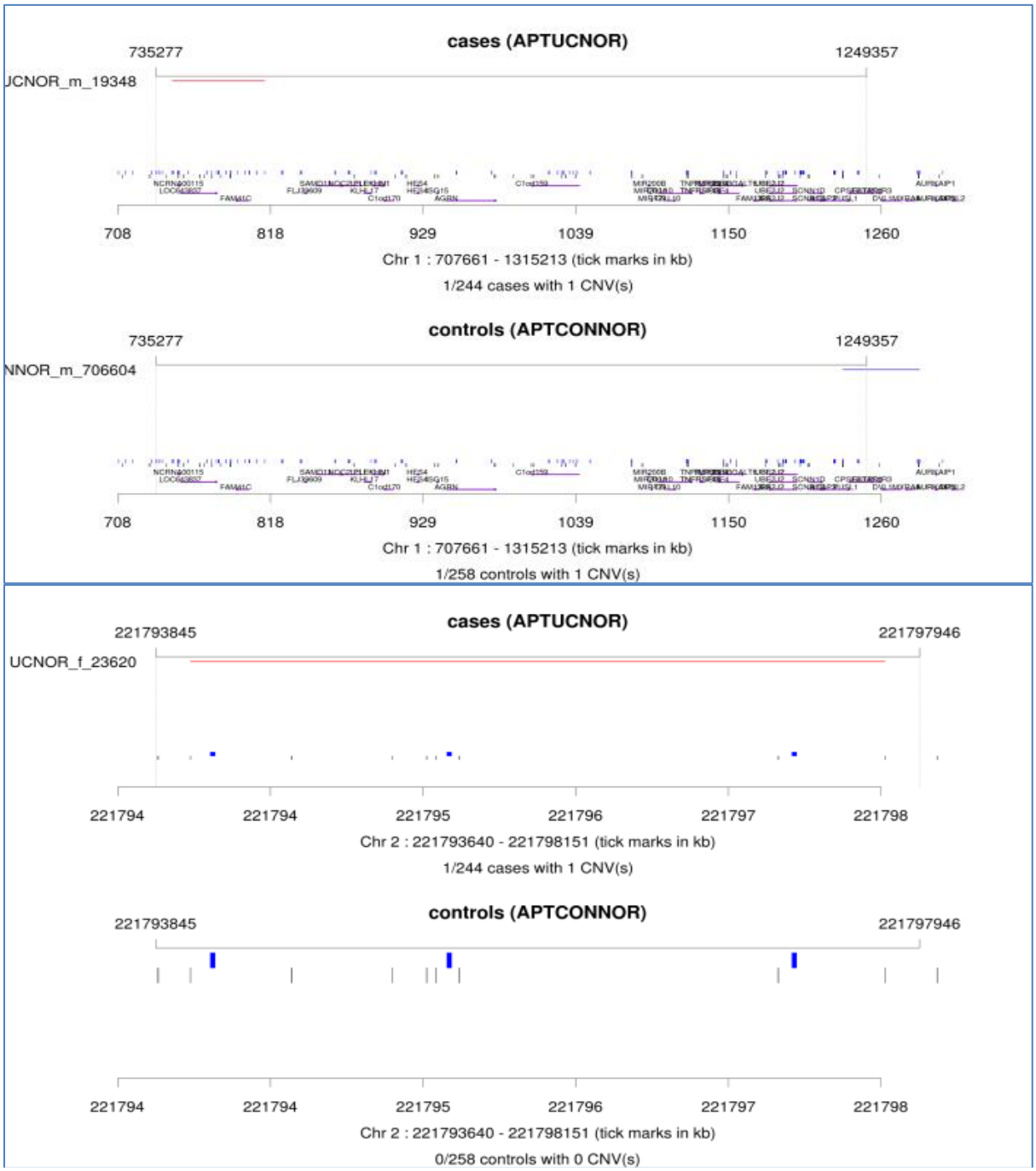
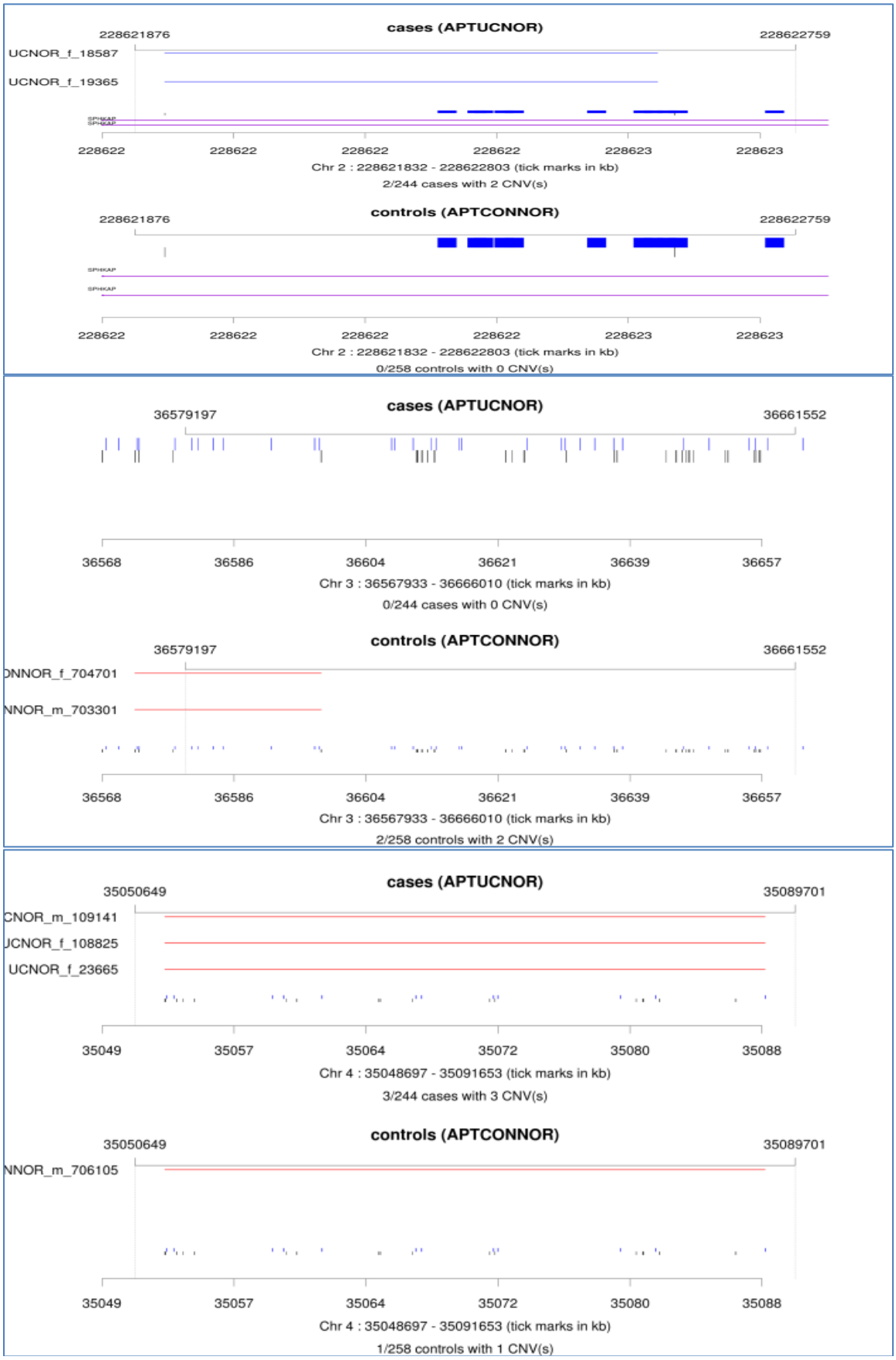
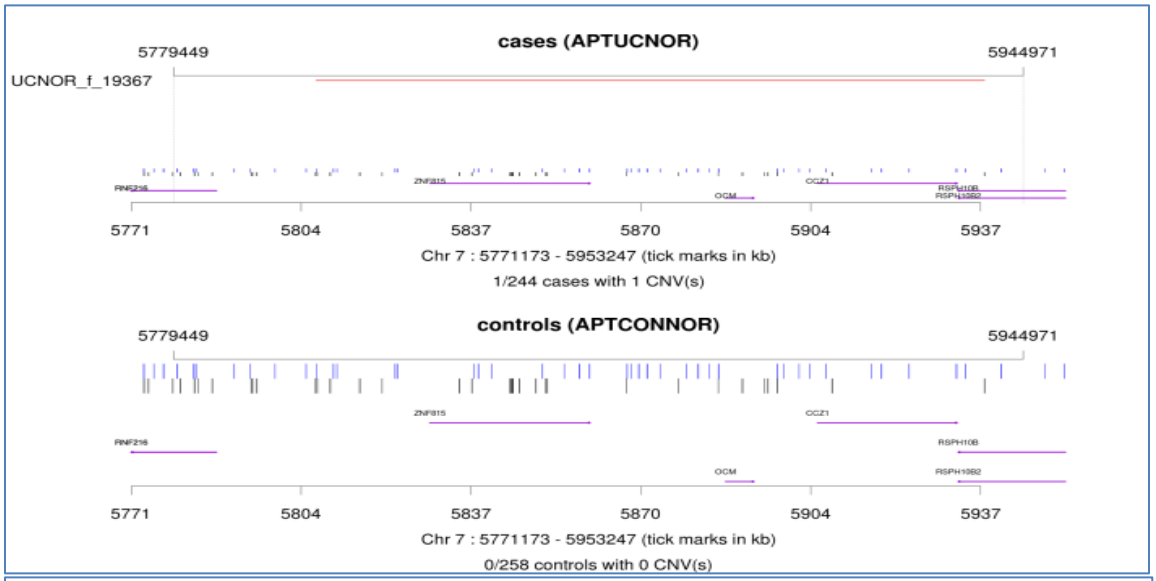

Applied Biosystems CopyCaller™ Software v1.0
8. File: Hs03624989_cn_CNVPOPCOF15.txt, Target: Hs03624989_cn
Copy Number
0
1
2
UCGER_m_114400
Applied Biosystems CopyCaller™ Software v1.0
9. File: Hs03624989_cn_CNVPOPCOF16.txt, Target: Hs03624989_cn
Copy Number
0
1
2
UCGER_m_114400

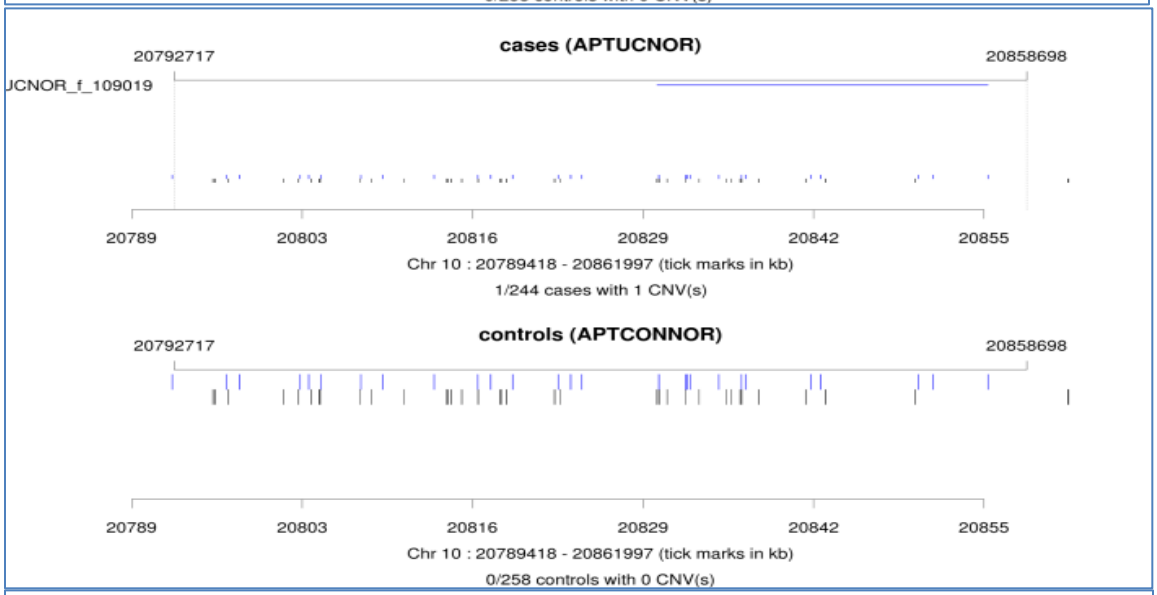
Table 6.5 *In-silico* replication of the 24 (13) rare CNV regions within the Norwegian data set.

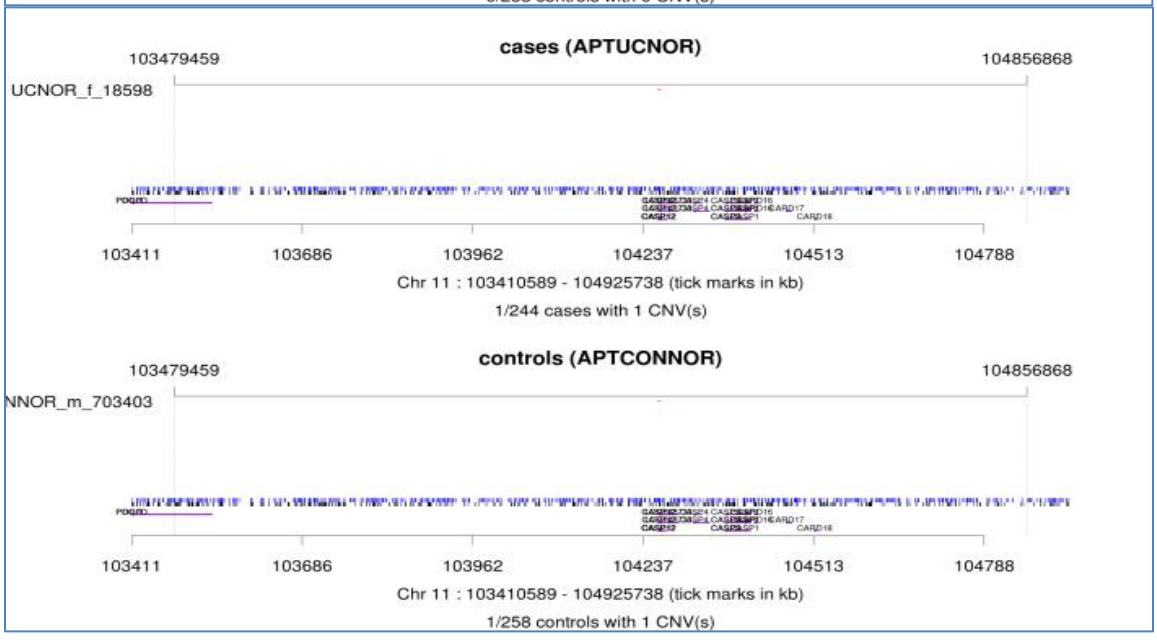
24 rare CNVs identified by initial CNV screening in German discovery panel were evaluated in the Norwegian *in-silico* replication data set. Eleven of the 24 regions have no predicted CNVs within the Norwegian sample set (neither cases nor controls) and are not visualized here. Each picture consists of two main panels. The upper panel shows the case CNVs and the lower panel the control CNVs

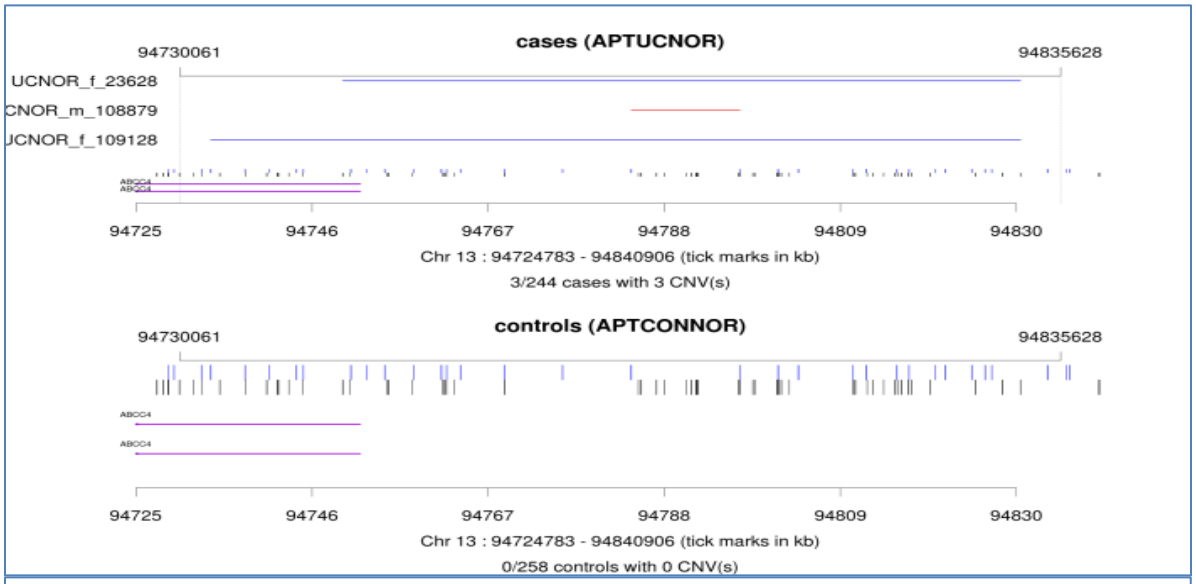


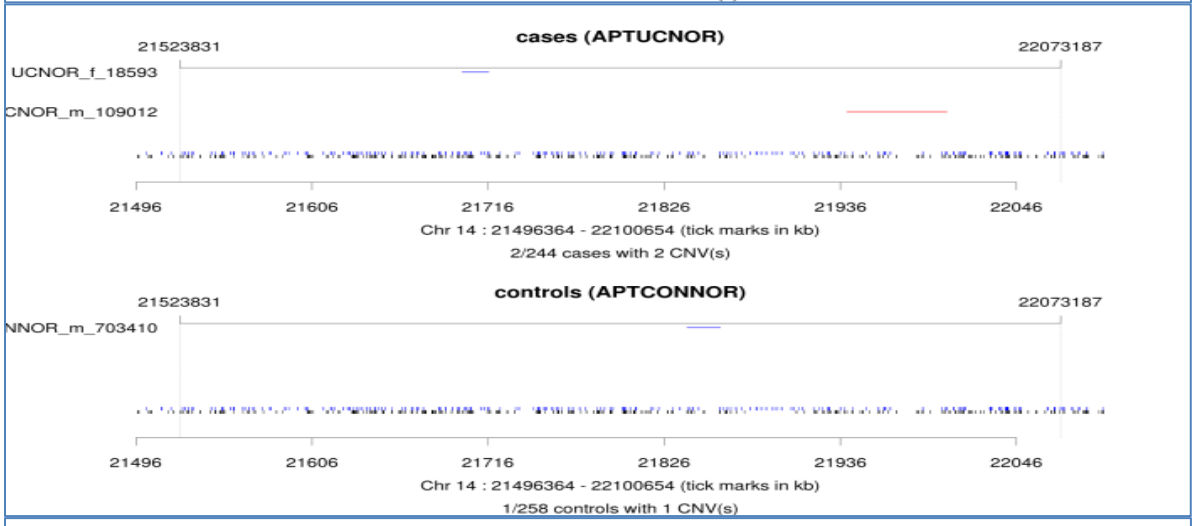

cases (APTUCNOR)
228621876
228622759
UCNOR_f_18587
UCNOR_f_19365
228622
228622
228622
228622
228623
228623
Chr 2 : 228621832 - 228622803 (tick marks in kb)
2/244 cases with 2 CNV(s)
controls (APTCONNOR)
228621876
228622759
228622
228622
228622
228622
228623
228623
Chr 2 : 228621832 - 228622803 (tick marks in kb)
0/258 controls with 0 CNV(s)
cases (APTUCNOR)
36579197
36661552
36568
36586
36604
36621
36639
36657
Chr 3 : 36567933 - 36666010 (tick marks in kb)
0/244 cases with 0 CNV(s)
controls (APTCONNOR)
36579197
36661552
ONNOR_f_704701
NNOR_m_703301
36568
36586
36604
36621
36639
36657
Chr 3 : 36567933 - 36666010 (tick marks in kb)
2/258 controls with 2 CNV(s)
cases (APTUCNOR)
35050649
35089701
CNOR_m_109141
UCNOR_f_108825
UCNOR_f_23665
35049
35057
35064
35072
35080
35088
Chr 4 : 35048697 - 35091653 (tick marks in kb)
3/244 cases with 3 CNV(s)
controls (APTCONNOR)
35050649
35089701
NNOR_m_706105
35049
35057
35064
35072
35080
35088
Chr 4 : 35048697 - 35091653 (tick marks in kb)
1/258 controls with 1 CNV(s)

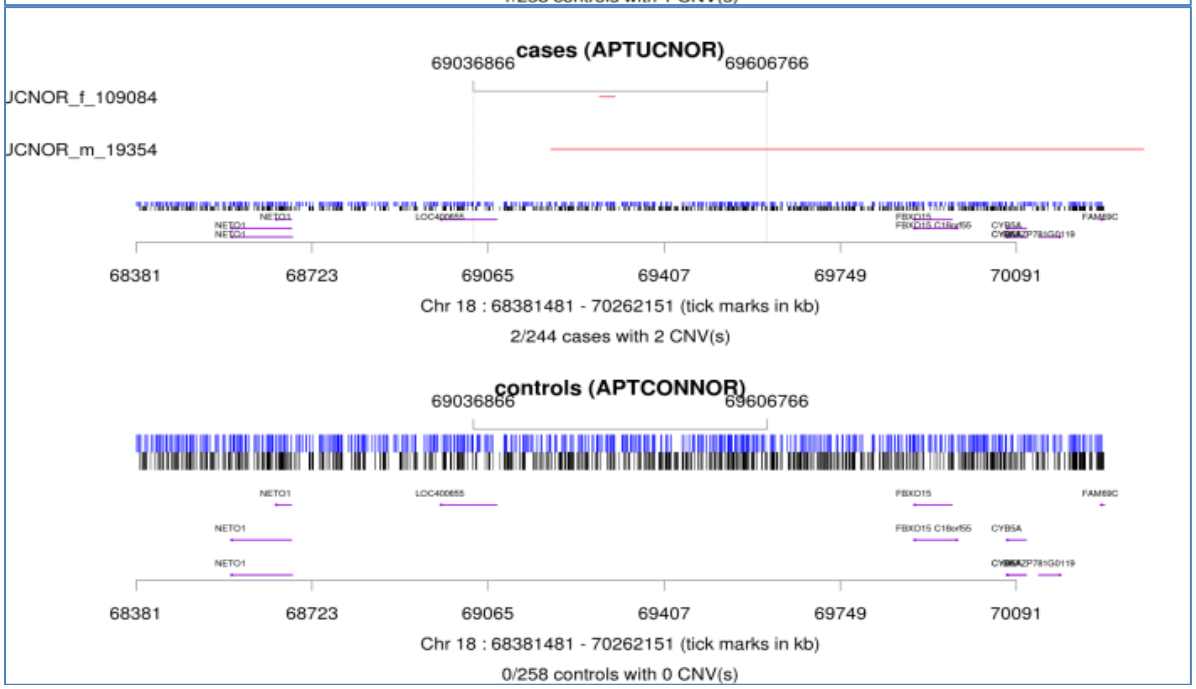

cases (APTUCNOR)
5779449
5944971
UCNOR_f_19367
5771
5804
5837
5870
5904
5937
Chr 7 : 5771173 - 5953247 (tick marks in kb)
1/244 cases with 1 CNV(s)
controls (APTCONNOR)
5779449
5944971
5771
5804
5837
5870
5904
5937
Chr 7 : 5771173 - 5953247 (tick marks in kb)
0/258 controls with 0 CNV(s)

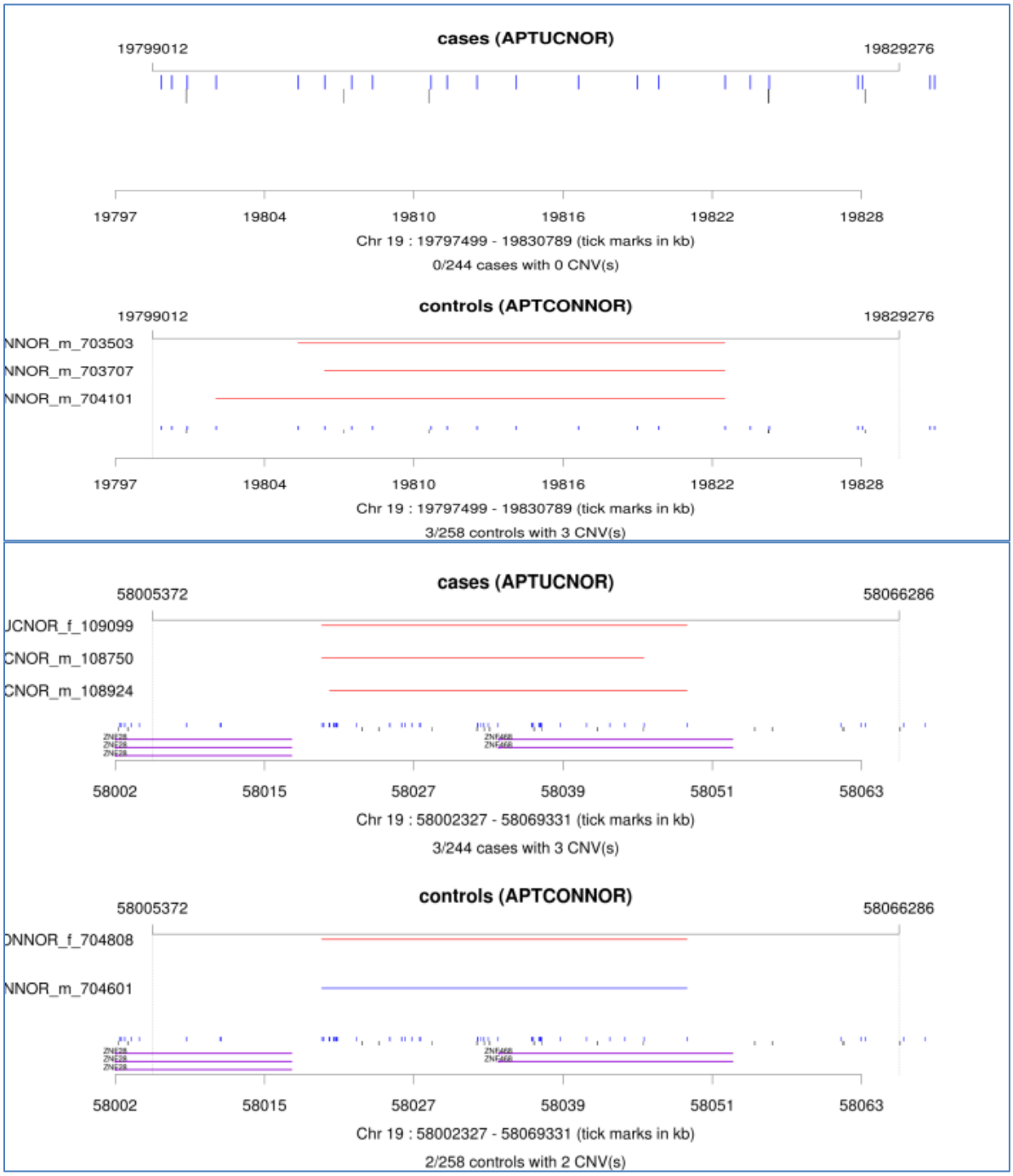

cases (APTUCNOR)
20792717
20858698
UCNOR_f_109019
20789
20803
20816
20829
20842
20855
Chr 10 : 20789418 - 20861997 (tick marks in kb)
1/244 cases with 1 CNV(s)
controls (APTCONNOR)
20792717
20858698
20789
20803
20816
20829
20842
20855
Chr 10 : 20789418 - 20861997 (tick marks in kb)
0/258 controls with 0 CNV(s)


cases (APTUCNOR)
103479459
104856868
UCNOR_f_18598
103411
103686
103962
104237
104513
104788
Chr 11 : 103410589 - 104925738 (tick marks in kb)
1/244 cases with 1 CNV(s)
controls (APTCONNOR)
103479459
104856868
103411
103686
103962
104237
104513
104788
Chr 11 : 103410589 - 104925738 (tick marks in kb)
1/258 controls with 1 CNV(s)

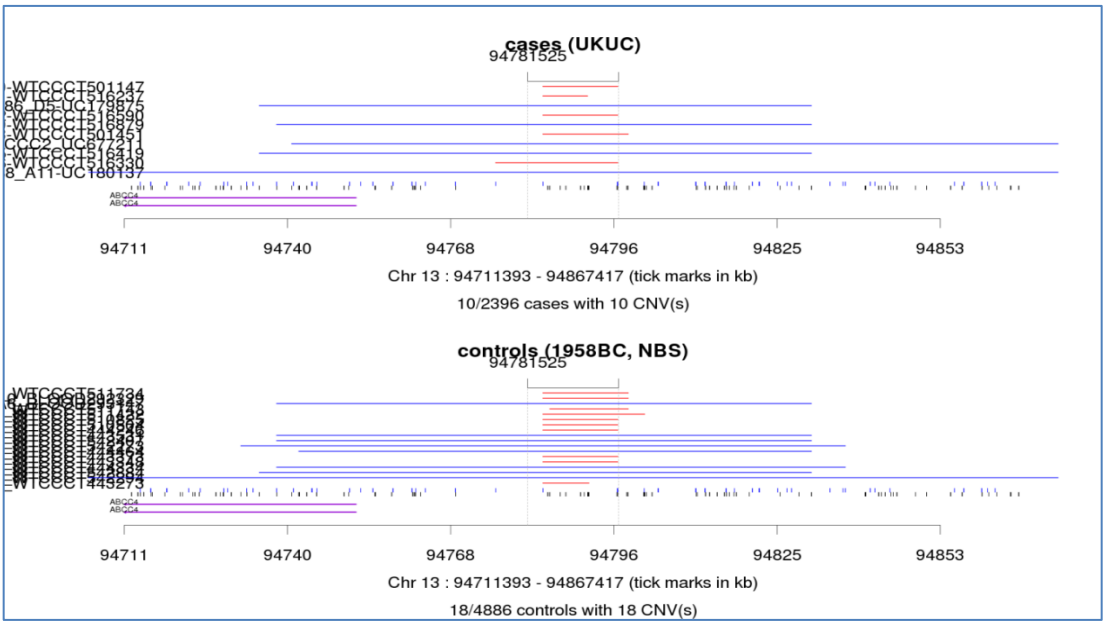

cases (APTUCNOR)
94730061
94835628
UCNOR_f_23628
CNOR_m_108879
JCNOR_f_109128
94725
94746
94767
94788
94809
94830
Chr 13 : 94724783 - 94840906 (tick marks in kb)
3/244 cases with 3 CNV(s)
controls (APTCONNOR)
94730061
94835628
94725
94746
94767
94788
94809
94830
Chr 13 : 94724783 - 94840906 (tick marks in kb)
0/258 controls with 0 CNV(s)

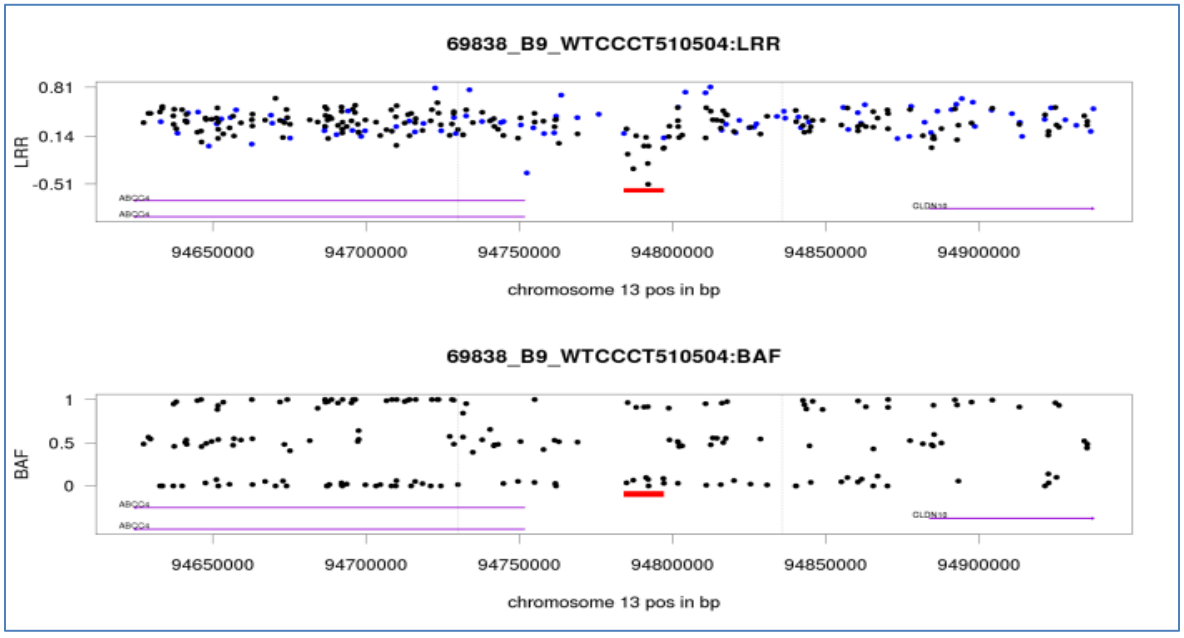

cases (APTUCNOR)
21523831
22073187
UCNOR_f_18593
CNOR_m_109012
21496
21606
21716
21826
21936
22046
Chr 14 : 21496364 - 22100654 (tick marks in kb)
2/244 cases with 2 CNV(s)
controls (APTCONNOR)
21523831
22073187
NNOR_m_703410
21496
21606
21716
21826
21936
22046
Chr 14 : 21496364 - 22100654 (tick marks in kb)
1/258 controls with 1 CNV(s)

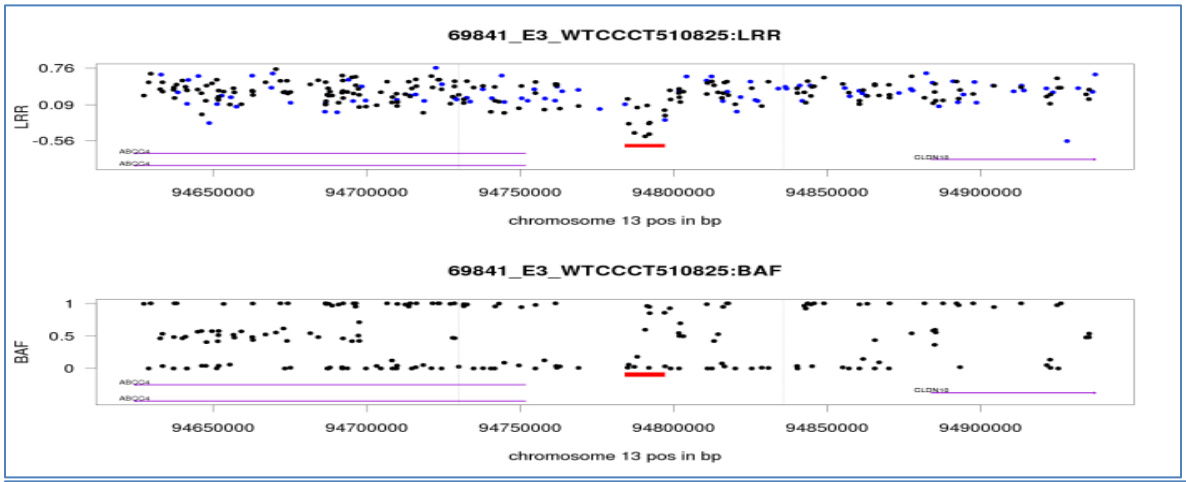

cases (APTUCNOR)
69036866
69606766
JCNOR_f_109084
JCNOR_m_19354
68381
68723
69065
69407
69749
70091
Chr 18 : 68381481 - 70262151 (tick marks in kb)
2/244 cases with 2 CNV(s)
controls (APTCONNOR)
69036866
69606766
68381
68723
69065
69407
69749
70091
Chr 18 : 68381481 - 70262151 (tick marks in kb)
0/258 controls with 0 CNV(s)

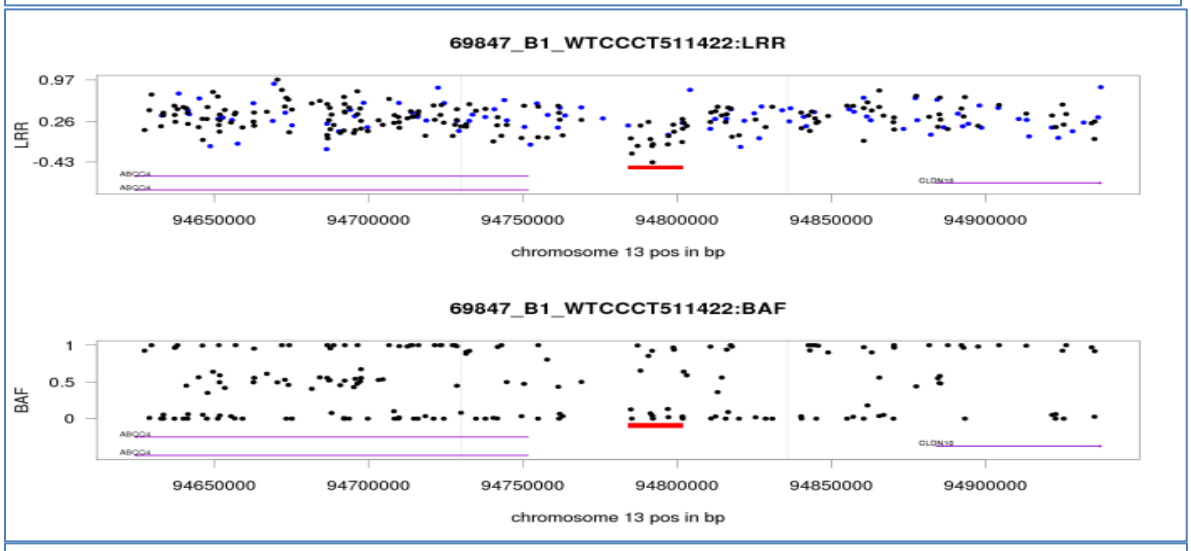

cases (APTUCNOR)
19799012
19829276
19797
19804
19810
19816
19822
19828
Chr 19 : 19797499 - 19830789 (tick marks in kb)
0/244 cases with 0 CNV(s)
controls (APTCONNOR)
19799012
19829276
NNOR_m_703503
NNOR_m_703707
NNOR_m_704101
19797
19804
19810
19816
19822
19828
Chr 19 : 19797499 - 19830789 (tick marks in kb)
3/258 controls with 3 CNV(s)
cases (APTUCNOR)
58005372
58066286
JCNOR_f_109099
CNOR_m_108750
CNOR_m_108924
58002
58015
58027
58039
58051
58063
Chr 19 : 58002327 - 58069331 (tick marks in kb)
3/244 cases with 3 CNV(s)
controls (APTCONNOR)
58005372
58066286
ONNOR_f_704808
NNOR_m_704601
58002
58015
58027
58039
58051
58063
Chr 19 : 58002327 - 58069331 (tick marks in kb)
2/258 controls with 2 CNV(s)

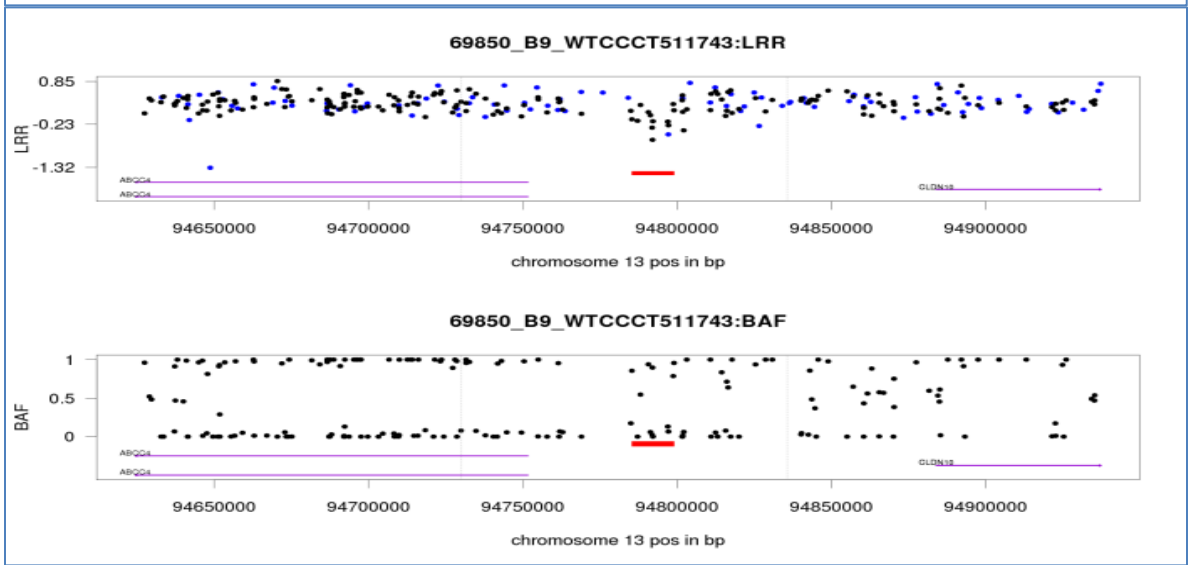
Tble 6.6 Deletion 13q32.1 in WTCCC2 data set. Evaluation of Del13q32.1 within the UK-WTCCC2 data set is visualized here. The first graph of this figure shows an overview of the CNV prediction for that region (Affymetrix Power Tools copy-number-workflow). The pictures show the raw data visualization of LRR in the top part and B allele frequency (BAF) in the lower part. Non polymorphic probe sets are blue and SNP probesets black. RefGene annotation is added with purple arrows. The red bar between RefSeq annotation and raw data highlights the predicted deletion.

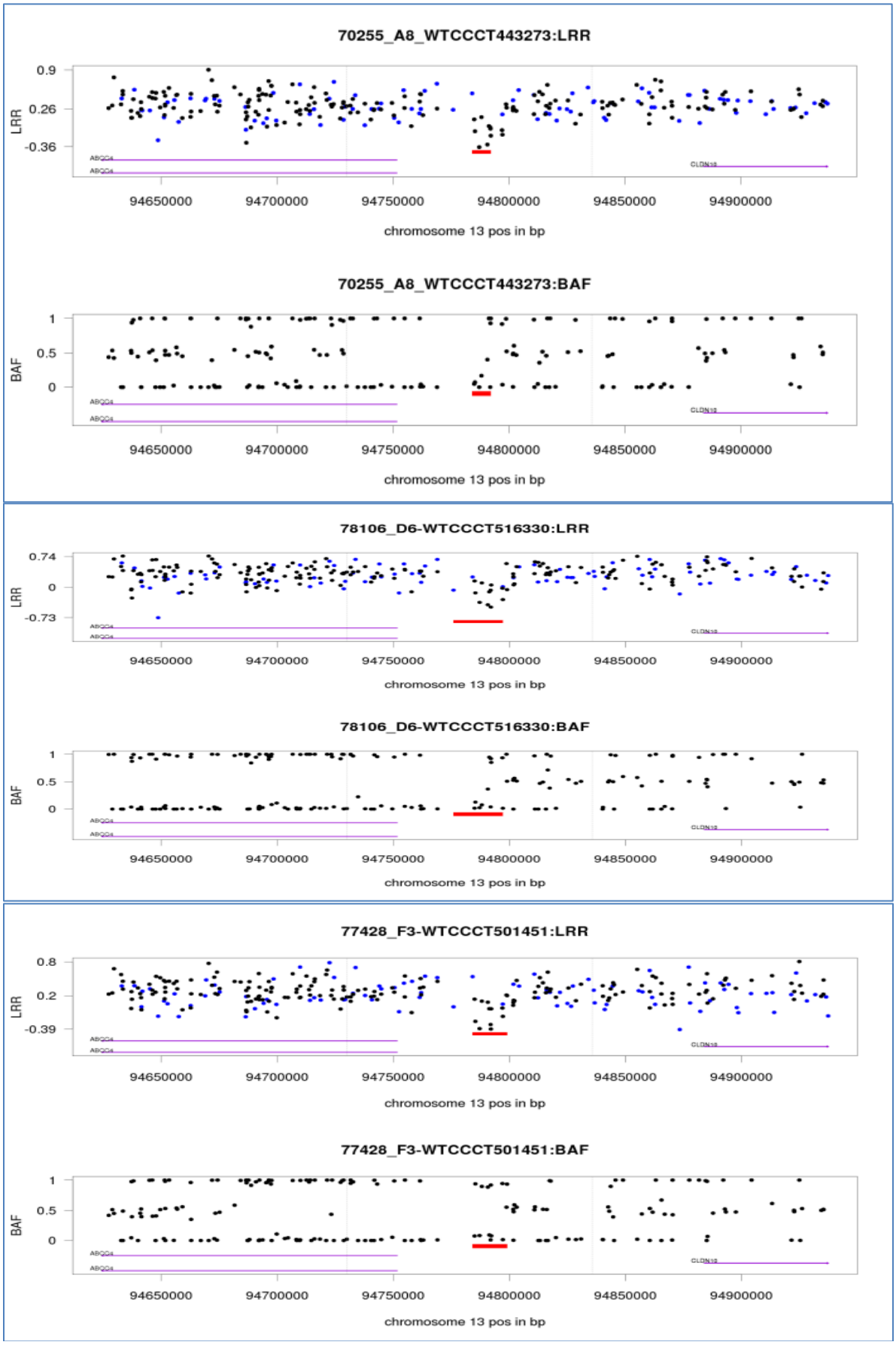


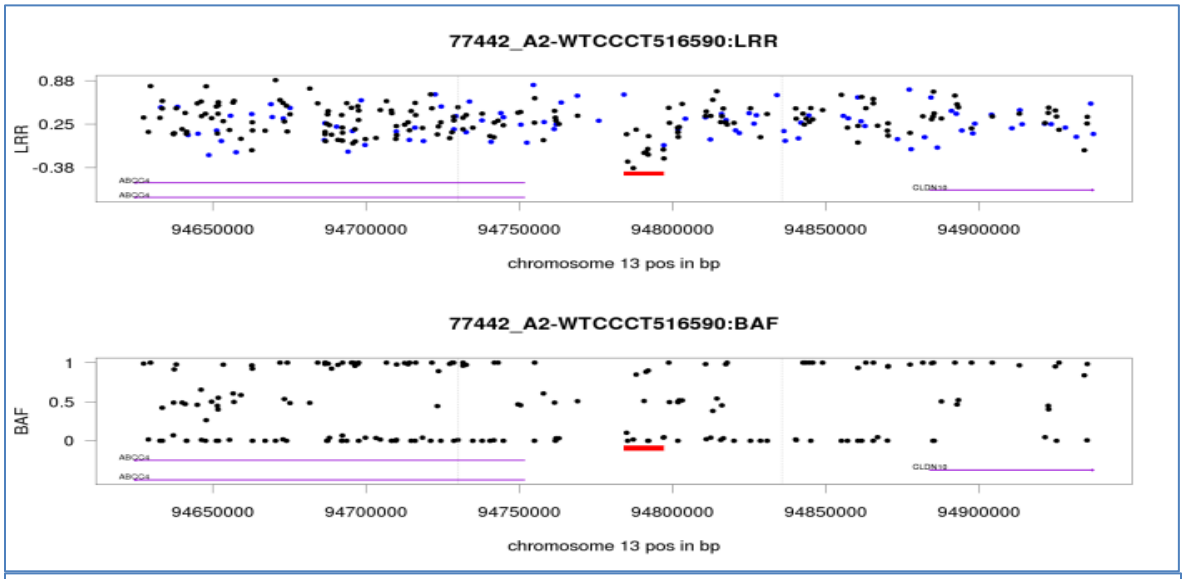


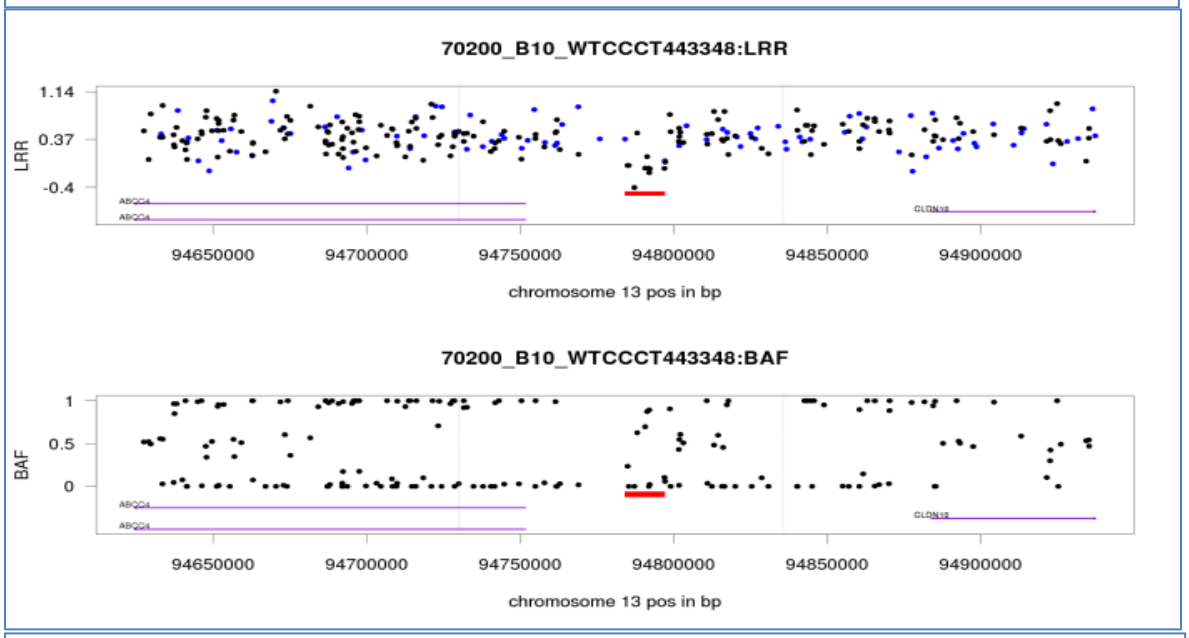

69841_E3_WTCCCT510825:LRR
LRR
0.76
0.09
-0.56
94650000
94700000
94750000
94800000
94850000
94900000
chromosome 13 pos in bp
69841_E3_WTCCCT510825:BAF
BAF
1
0.5
0
94650000
94700000
94750000
94800000
94850000
94900000
chromosome 13 pos in bp

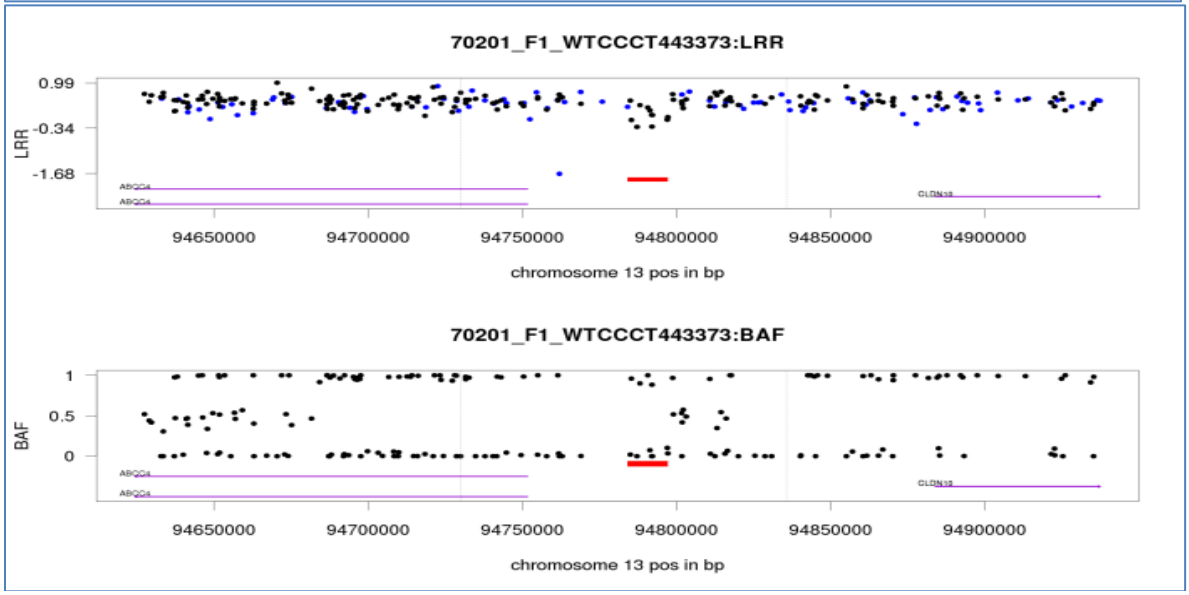

69847_B1_WTCCCT511422:LRR
LRR
0.97
0.26
-0.43
94650000
94700000
94750000
94800000
94850000
94900000
chromosome 13 pos in bp
69847_B1_WTCCCT511422:BAF
BAF
1
0.5
0
94650000
94700000
94750000
94800000
94850000
94900000
chromosome 13 pos in bp

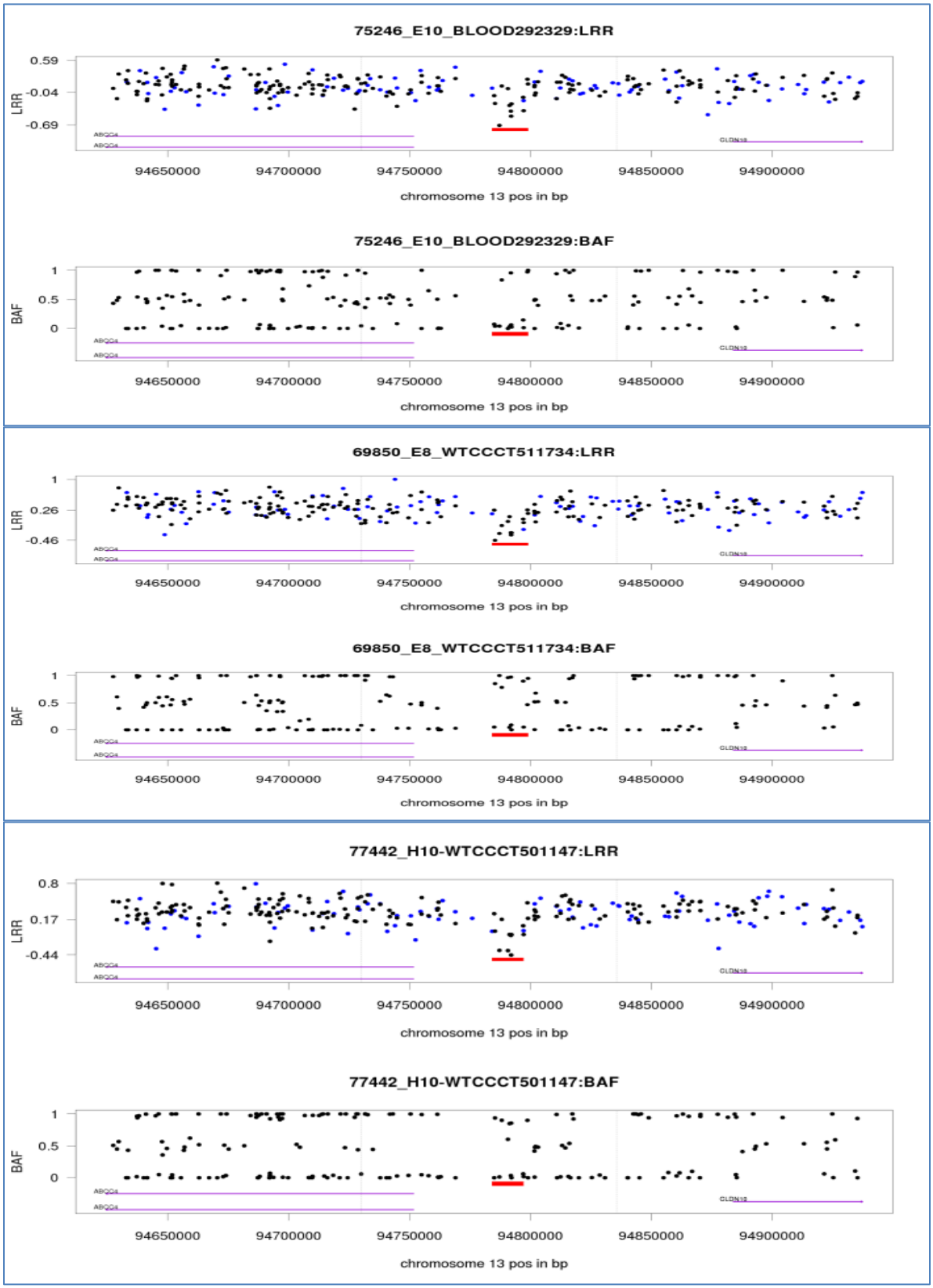

69850_B9_WTCCCT511743:LRR
LRR
0.85
-0.23
-1.32
94650000
94700000
94750000
94800000
94850000
94900000
chromosome 13 pos in bp
69850_B9_WTCCCT511743:BAF
BAF
1
0.5
0
94650000
94700000
94750000
94800000
94850000
94900000
chromosome 13 pos in bp


70255_A8_WTCCCT443273:LRR
LRR
0.9
0.26
-0.36
94650000
94700000
94750000
94800000
94850000
94900000
chromosome 13 pos in bp
70255_A8_WTCCCT443273:BAF
BAF
1
0.5
0
94650000
94700000
94750000
94800000
94850000
94900000
chromosome 13 pos in bp
78106_D6-WTCCCT516330:LRR
LRR
0.74
0
-0.73
94650000
94700000
94750000
94800000
94850000
94900000
chromosome 13 pos in bp
78106_D6-WTCCCT516330:BAF
BAF
1
0.5
0
94650000
94700000
94750000
94800000
94850000
94900000
chromosome 13 pos in bp
77428_F3-WTCCCT501451:LRR
LRR
0.8
0.2
-0.39
94650000
94700000
94750000
94800000
94850000
94900000
chromosome 13 pos in bp
77428_F3-WTCCCT501451:BAF
BAF
1
0.5
0
94650000
94700000
94750000
94800000
94850000
94900000
chromosome 13 pos in bp


77442_A2-WTCCCT516590:LRR
LRR
0.88
0.25
-0.38
94650000
94700000
94750000
94800000
94850000
94900000
chromosome 13 pos in bp
77442_A2-WTCCCT516590:BAF
BAF
1
0.5
0
94650000
94700000
94750000
94800000
94850000
94900000
chromosome 13 pos in bp


70200_B10_WTCCCT443348:LRR
LRR
1.14
0.37
-0.4
94650000
94700000
94750000
94800000
94850000
94900000
chromosome 13 pos in bp
70200_B10_WTCCCT443348:BAF
BAF
1
0.5
0
94650000
94700000
94750000
94800000
94850000
94900000
chromosome 13 pos in bp


70201_F1_WTCCCT443373:LRR
LRR
0.99
-0.34
-1.68
94650000
94700000
94750000
94800000
94850000
94900000
chromosome 13 pos in bp
70201_F1_WTCCCT443373:BAF
BAF
1
0.5
0
94650000
94700000
94750000
94800000
94850000
94900000
chromosome 13 pos in bp


75246_E10_BLOOD292329:LRR
LRR
0.59
-0.04
-0.69
94650000
94700000
94750000
94800000
94850000
94900000
chromosome 13 pos in bp
75246_E10_BLOOD292329:BAF
BAF
1
0.5
0
chromosome 13 pos in bp
69850_E8_WTCCCT511734:LRR
LRR
1
0.26
-0.46
chromosome 13 pos in bp
69850_E8_WTCCCT511734:BAF
BAF
chromosome 13 pos in bp
77442_H10-WTCCCT501147:LRR
LRR
0.8
0.17
-0.44
chromosome 13 pos in bp
77442_H10-WTCCCT501147:BAF
BAF
chromosome 13 pos in bp

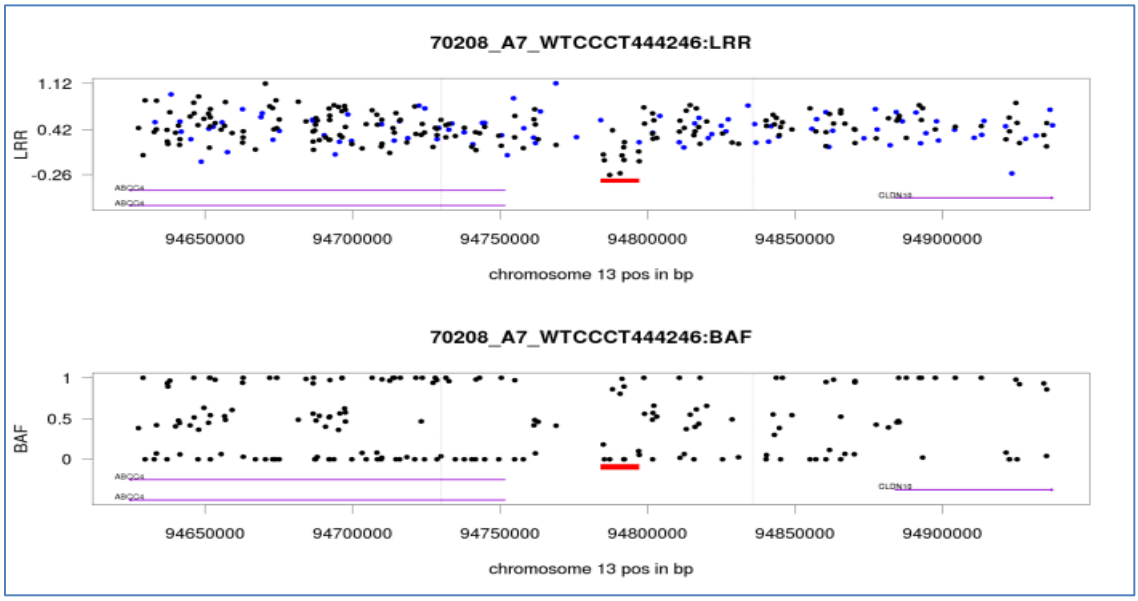
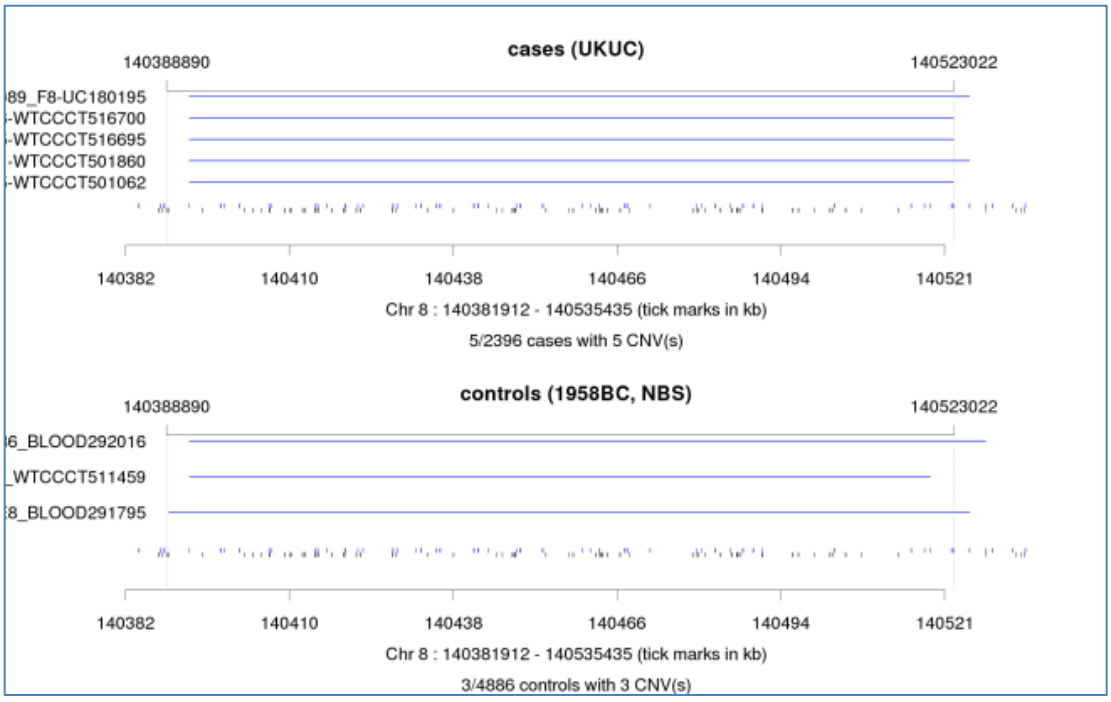
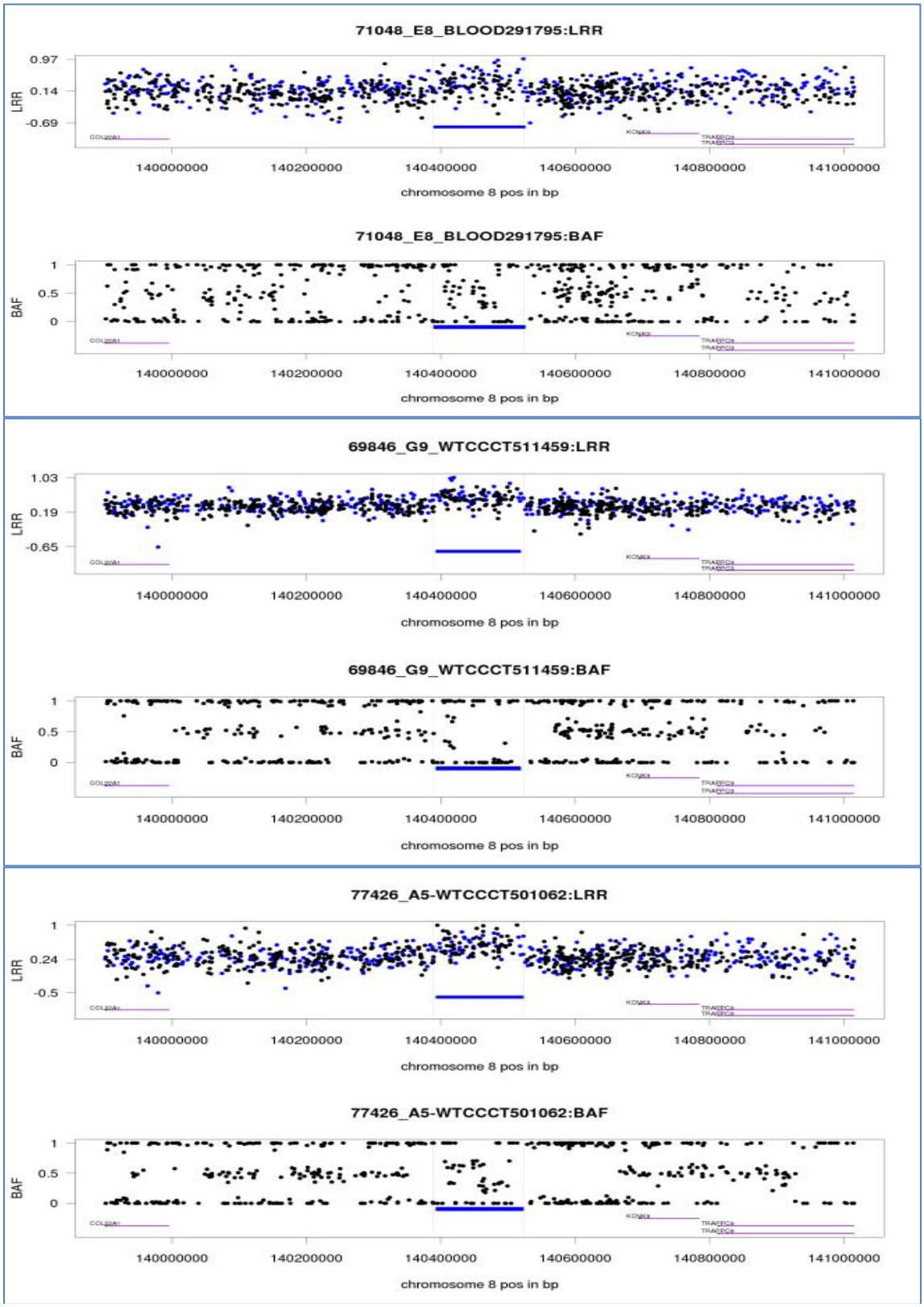
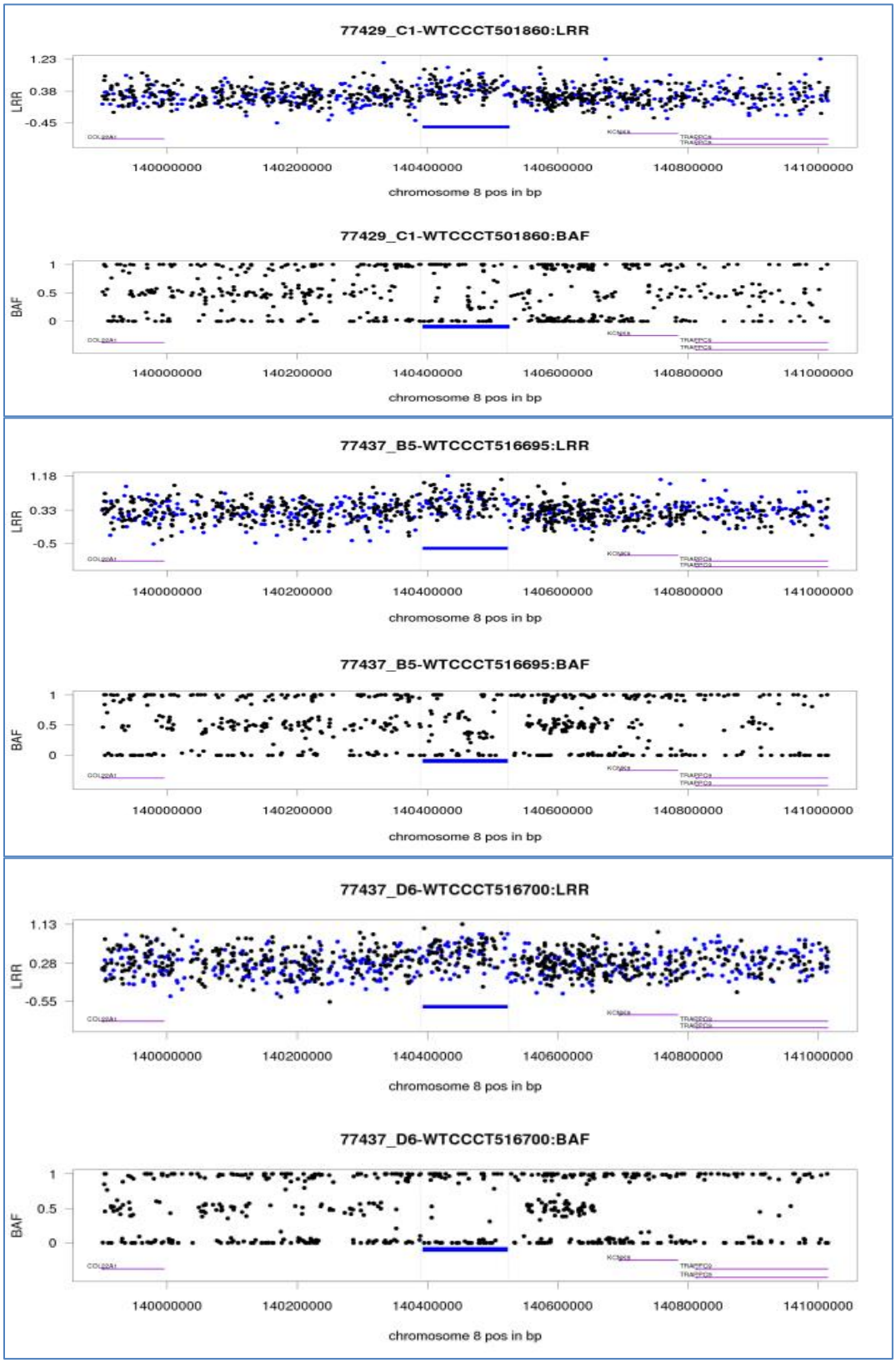
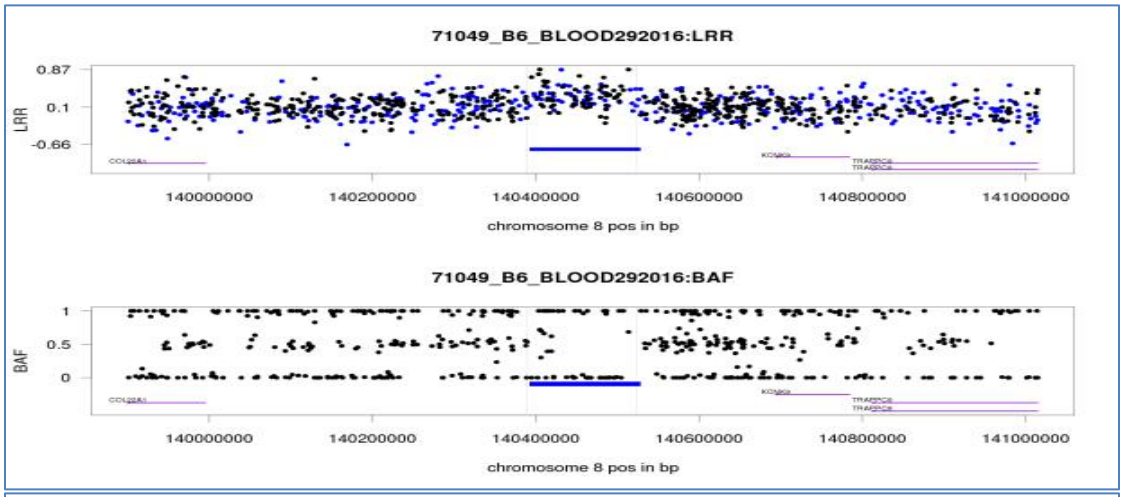


Table 6.7 Duplication 8q24.3 in WTCCC2 data set. In each figure, raw data visualization of LRR in the upper track and B allele frequency (BAF) in the lower track are shown. Non polymorphic probe sets are blue and SNP probe sets black. RefSeq gene annotations are added with purple arrows. The blue bar between RefSeq annotation and raw data highlights the predicted duplication.




71048_E8_BLOOD291795:LRR
LRR
0.97
0.14
-0.69
140000000
140200000
140400000
140600000
140800000
141000000
chromosome 8 pos in bp
71048_E8_BLOOD291795:BAF
BAF
1
0.5
0
chromosome 8 pos in bp
69846_G9_WTCCCT511459:LRR
LRR
1.03
0.19
-0.65
chromosome 8 pos in bp
69846_G9_WTCCCT511459:BAF
BAF
chromosome 8 pos in bp
77426_A5-WTCCCT501062:LRR
LRR
1
0.24
-0.5
chromosome 8 pos in bp
77426_A5-WTCCCT501062:BAF
BAF
chromosome 8 pos in bp


77429_C1-WTCCCT501860:LRR
LRR
1.23
0.38
-0.45
140000000
140200000
140400000
140600000
140800000
141000000
chromosome 8 pos in bp
77429_C1-WTCCCT501860:BAF
BAF
1
0.5
0
140000000
140200000
140400000
140600000
140800000
141000000
chromosome 8 pos in bp
77437_B5-WTCCCT516695:LRR
LRR
1.18
0.33
-0.5
140000000
140200000
140400000
140600000
140800000
141000000
chromosome 8 pos in bp
77437_B5-WTCCCT516695:BAF
BAF
1
0.5
0
140000000
140200000
140400000
140600000
140800000
141000000
chromosome 8 pos in bp
77437_D6-WTCCCT516700:LRR
LRR
1.13
0.28
-0.55
140000000
140200000
140400000
140600000
140800000
141000000
chromosome 8 pos in bp
77437_D6-WTCCCT516700:BAF
BAF
1
0.5
0
140000000
140200000
140400000
140600000
140800000
141000000
chromosome 8 pos in bp


71049_B6_BLOOD292016:LRR
LRR
0.87
0.1
-0.66
140000000
140200000
140400000
140600000
140800000
141000000
chromosome 8 pos in bp
71049_B6_BLOOD292016:BAF
BAF
1
0.5
0
140000000
140200000
140400000
140600000
140800000
141000000
chromosome 8 pos in bp

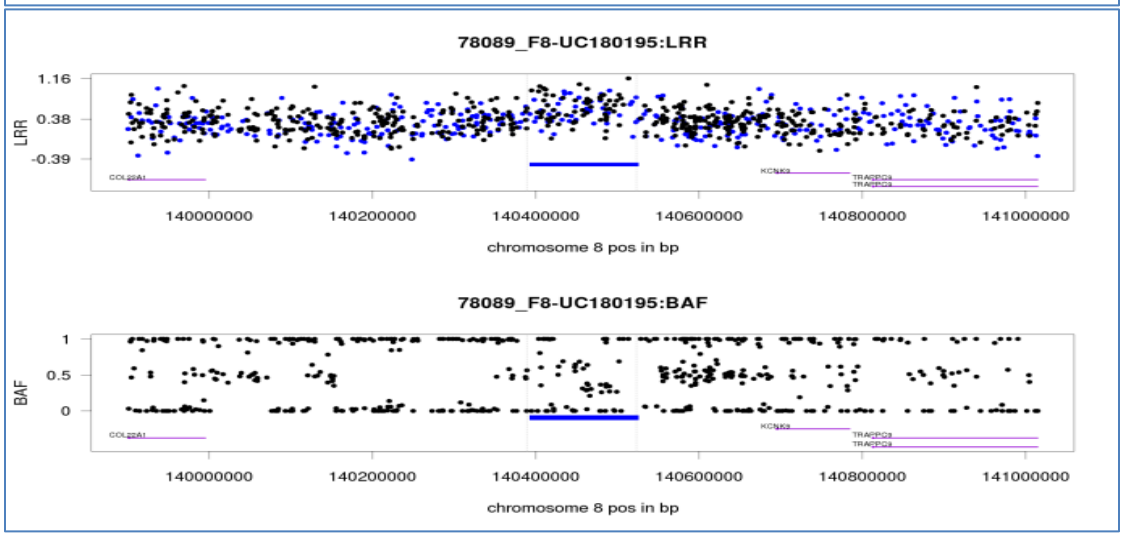
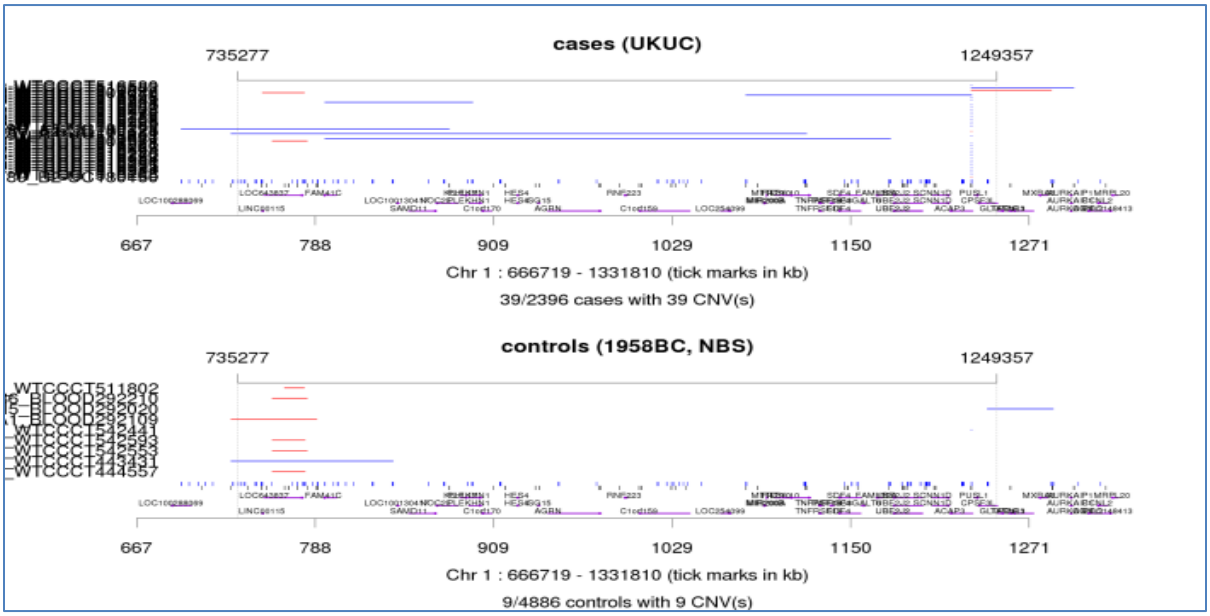
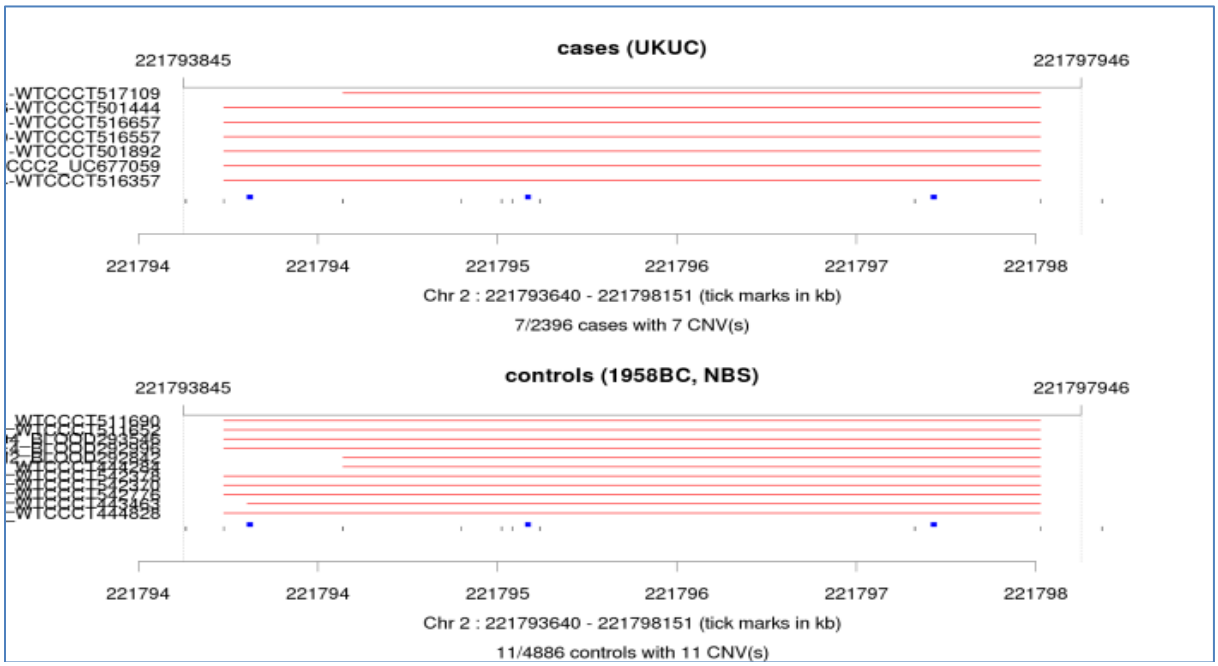
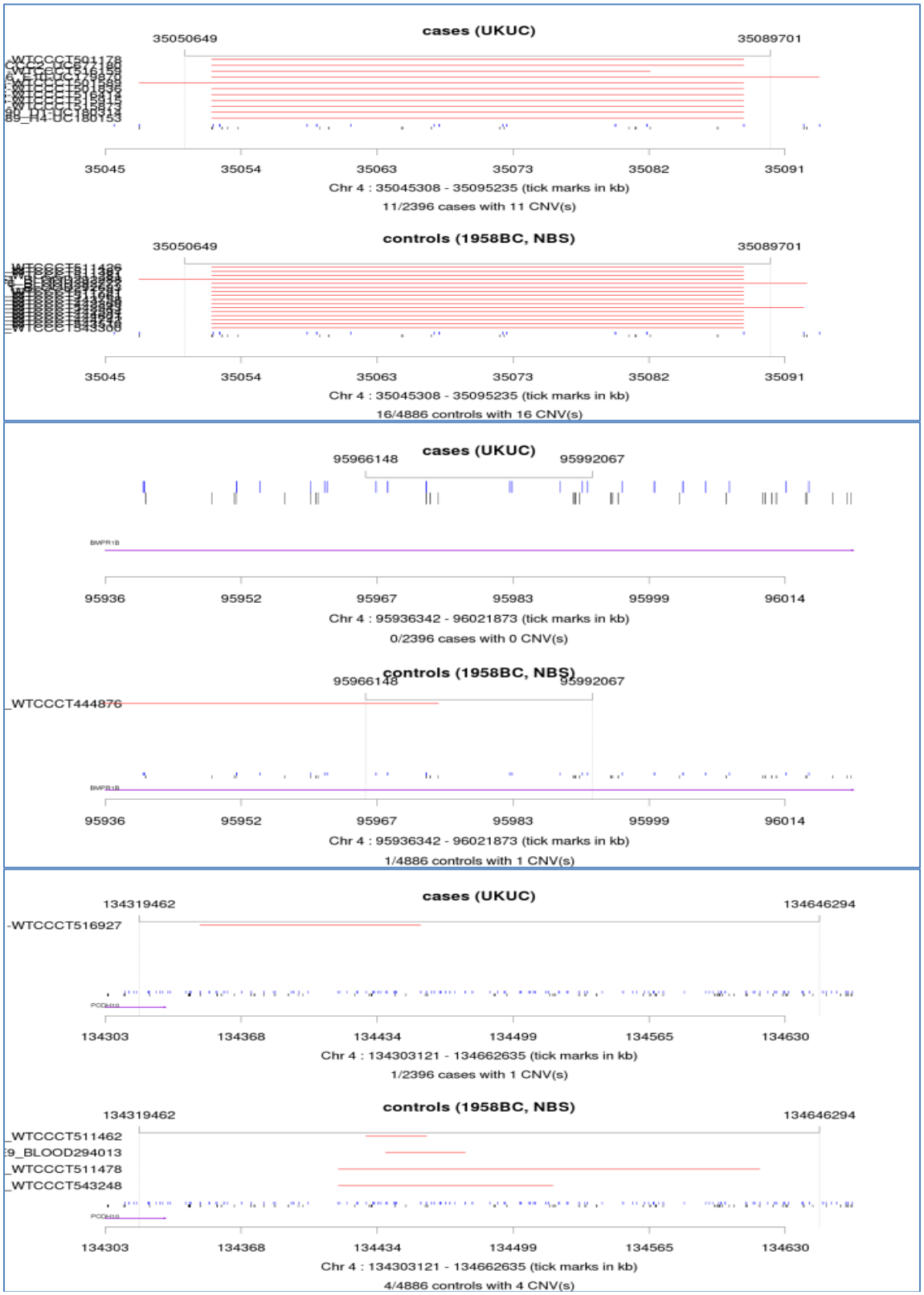

78089_F8-UC180195:LRR
LRR
1.16
0.38
-0.39
140000000
140200000
140400000
140600000
140800000
141000000
chromosome 8 pos in bp
78089_F8-UC180195:BAF
BAF
1
0.5
0
140000000
140200000
140400000
140600000
140800000
141000000
chromosome 8 pos in bp

Table 6.8 Evaluation of the 24 rare CNVs within the UK (WTCCC2) data set. Each figure of the table consists of two main panels. The upper panel shows the predicted CNVs in cases and the lower panel the control CNVs. The RefSeq genes are annotated in purple. SNP probe sets in black and copy number probe sets in blue.

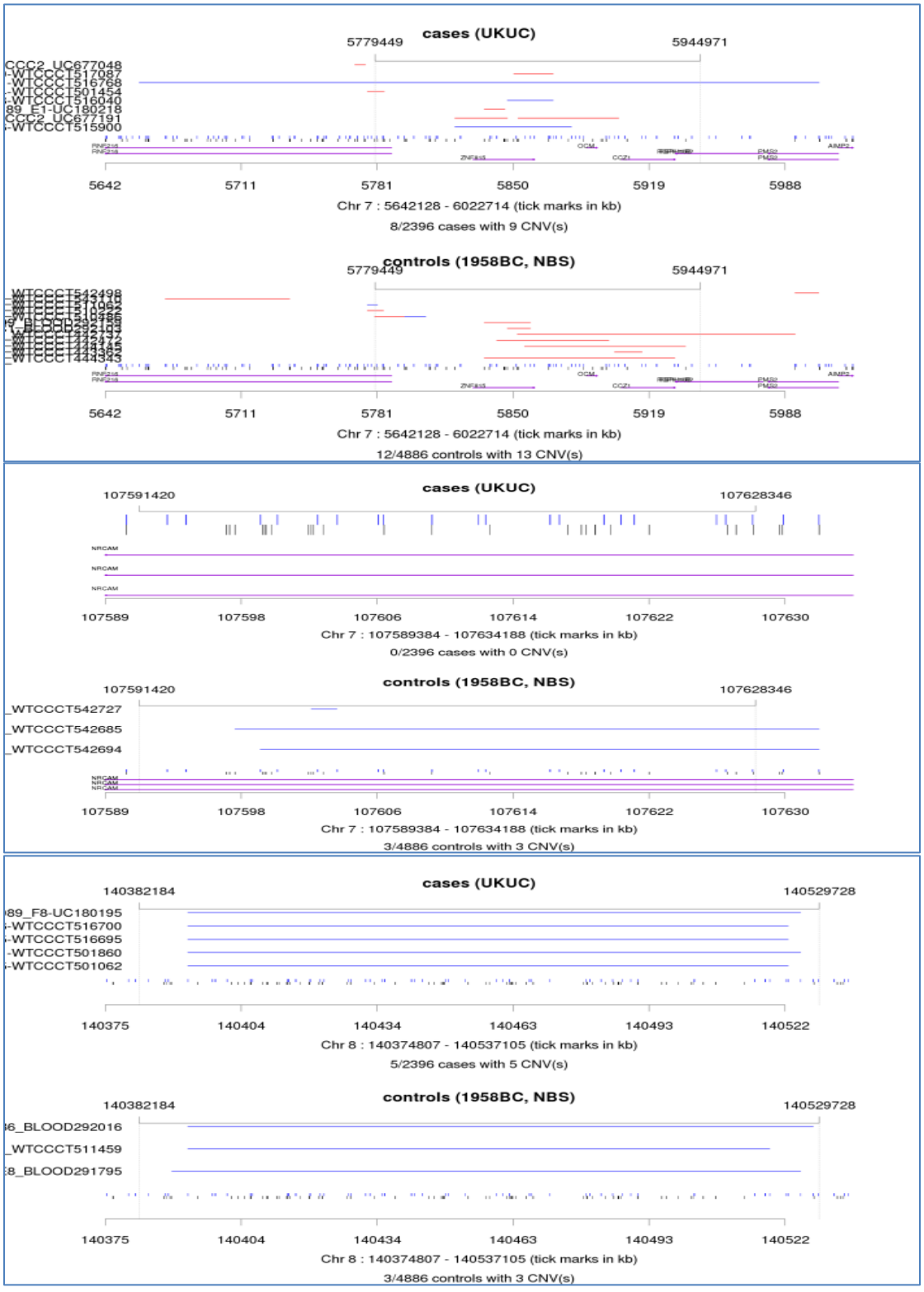





cases (UKUC)
228621876
228622759
228621
228621
228622
228622
228623
228624
Chr 2 : 228620861 - 228623774 (tick marks in kb)
0/2396 cases with 0 CNV(s)
controls (1958BC, NBS)
228621876
228622759
1_BLOOD294544
228621
228621
228622
228622
228623
228624
Chr 2 : 228620861 - 228623774 (tick marks in kb)
1/4886 controls with 1 CNV(s)
cases (UKUC)
36579197
36661552
36484
36534
36583
36633
36682
36732
Chr 3 : 36484489 - 36756260 (tick marks in kb)
0/2396 cases with 0 CNV(s)
controls (1958BC, NBS)
36579197
36661552
_WTCCCT443300
4_BLOOD293366
_WTCCCT444406
36484
36534
36583
36633
36682
36732
Chr 3 : 36484489 - 36756260 (tick marks in kb)
3/4886 controls with 3 CNV(s)
cases (UKUC)
195056041
195067368
195055
195058
195060
195062
195065
195067
Chr 3 : 195055475 - 195067934 (tick marks in kb)
0/2396 cases with 0 CNV(s)
controls (1958BC, NBS)
195056041
195067368
_WTCCCT542700
195055
195058
195060
195062
195065
195067
Chr 3 : 195055475 - 195067934 (tick marks in kb)
1/4886 controls with 1 CNV(s)


cases (UKUC)
35050649
35089701
35045
35054
35063
35073
35082
35091
Chr 4 : 35045308 - 35095235 (tick marks in kb)
11/2396 cases with 11 CNV(s)
controls (1958BC, NBS)
35050649
35089701
35045
35054
35063
35073
35082
35091
Chr 4 : 35045308 - 35095235 (tick marks in kb)
16/4886 controls with 16 CNV(s)
cases (UKUC)
95966148
95992067
BMPR1B
95936
95952
95967
95983
95999
96014
Chr 4 : 95936342 - 96021873 (tick marks in kb)
0/2396 cases with 0 CNV(s)
controls (1958BC, NBS)
95966148
95992067
_WTCCCT444876
BMPR1B
95936
95952
95967
95983
95999
96014
Chr 4 : 95936342 - 96021873 (tick marks in kb)
1/4886 controls with 1 CNV(s)
cases (UKUC)
134319462
134646294
-WTCCCT516927
PCDH10
134303
134368
134434
134499
134565
134630
Chr 4 : 134303121 - 134662635 (tick marks in kb)
1/2396 cases with 1 CNV(s)
controls (1958BC, NBS)
134319462
134646294
_WTCCCT511462
:9_BLOOD294013
_WTCCCT511478
_WTCCCT543248
PCDH10
134303
134368
134434
134499
134565
134630
Chr 4 : 134303121 - 134662635 (tick marks in kb)
4/4886 controls with 4 CNV(s)


cases (UKUC)
101035460
101135453
-WTCCCT516648
CCC2_UC676906
-WTCCCT501663
-WTCCCT501128
100920
100980
101040
101100
101160
101220
Chr 5 : 100920469 - 101250444 (tick marks in kb)
4/2396 cases with 4 CNV(s)
controls (1958BC, NBS)
101035460
101135453
100920
100980
101040
101100
101160
101220
Chr 5 : 100920469 - 101250444 (tick marks in kb)
14/4886 controls with 14 CNV(s)
cases (UKUC)
58254422
58912591
-WTCCCT516661
-WTCCCT516627
-WTCCCT462954
-WTCCCT515921
87_E7-UC179989
-WTCCCT501780
58143
58289
58436
58583
58729
58876
Chr 6 : 58142661 - 58949254 (tick marks in kb)
6/2396 cases with 6 CNV(s)
controls (1958BC, NBS)
58254422
58912591
58143
58289
58436
58583
58729
58876
Chr 6 : 58142661 - 58949254 (tick marks in kb)
30/4886 controls with 30 CNV(s)
cases (UKUC)
4580149
4587800
-WTCCCT501465
4571
4576
4581
4585
4590
4594
Chr 7 : 4571351 - 4596598 (tick marks in kb)
1/2396 cases with 1 CNV(s)
controls (1958BC, NBS)
4580149
4587800
_WTCCCT511260
_WTCCCT542836
4571
4576
4581
4585
4590
4594
Chr 7 : 4571351 - 4596598 (tick marks in kb)
2/4886 controls with 2 CNV(s)


cases (UKUC)
5779449
5944971
5642
5711
5781
5850
5919
5988
Chr 7 : 5642128 - 6022714 (tick marks in kb)
8/2396 cases with 9 CNV(s)
controls (1958BC, NBS)
5779449
5944971
5642
5711
5781
5850
5919
5988
Chr 7 : 5642128 - 6022714 (tick marks in kb)
12/4886 controls with 13 CNV(s)
cases (UKUC)
107591420
107628346
107589
107598
107606
107614
107622
107630
Chr 7 : 107589384 - 107634188 (tick marks in kb)
0/2396 cases with 0 CNV(s)
controls (1958BC, NBS)
107591420
107628346
_WTCCCT542727
_WTCCCT542685
_WTCCCT542694
107589
107598
107606
107614
107622
107630
Chr 7 : 107589384 - 107634188 (tick marks in kb)
3/4886 controls with 3 CNV(s)
cases (UKUC)
140382184
140529728
-WTCCCT516700
-WTCCCT516695
-WTCCCT501860
-WTCCCT501062
140375
140404
140434
140463
140493
140522
Chr 8 : 140374807 - 140537105 (tick marks in kb)
5/2396 cases with 5 CNV(s)
controls (1958BC, NBS)
140382184
140529728
_WTCCCT511459
140375
140404
140434
140463
140493
140522
Chr 8 : 140374807 - 140537105 (tick marks in kb)
3/4886 controls with 3 CNV(s)

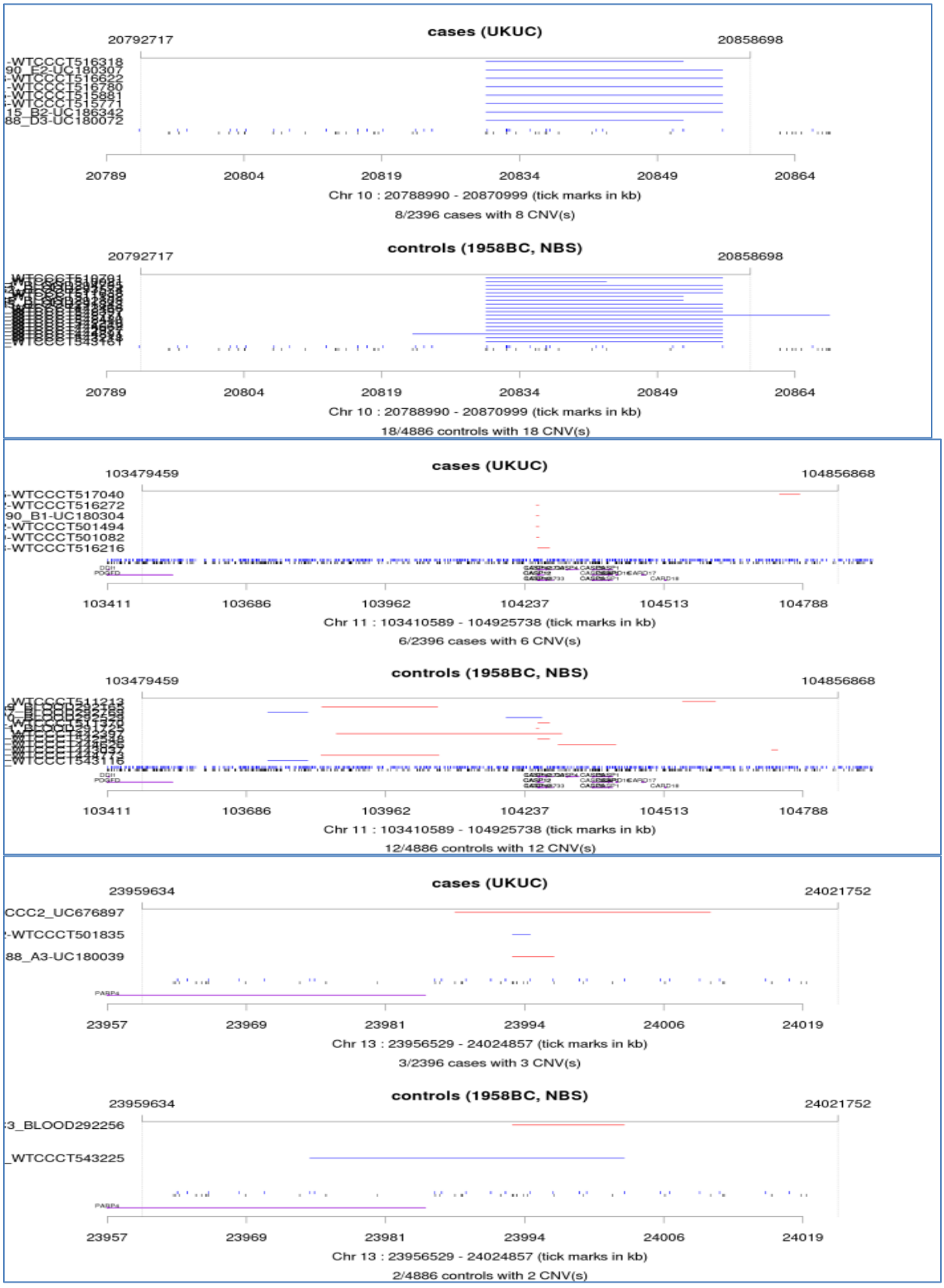
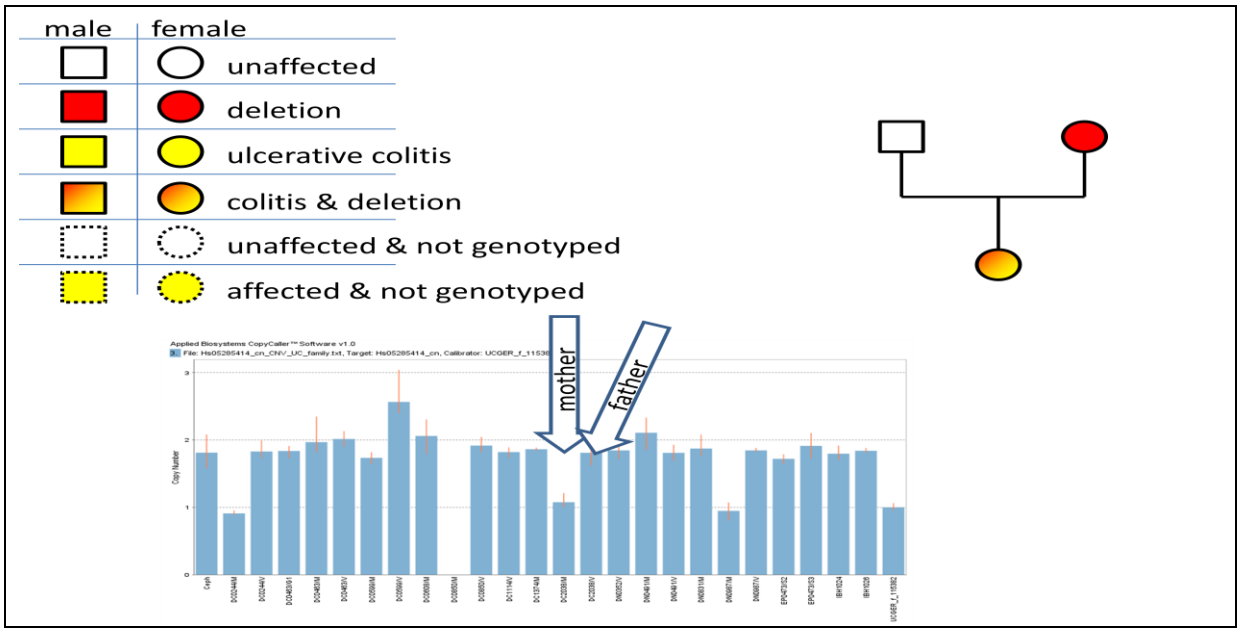
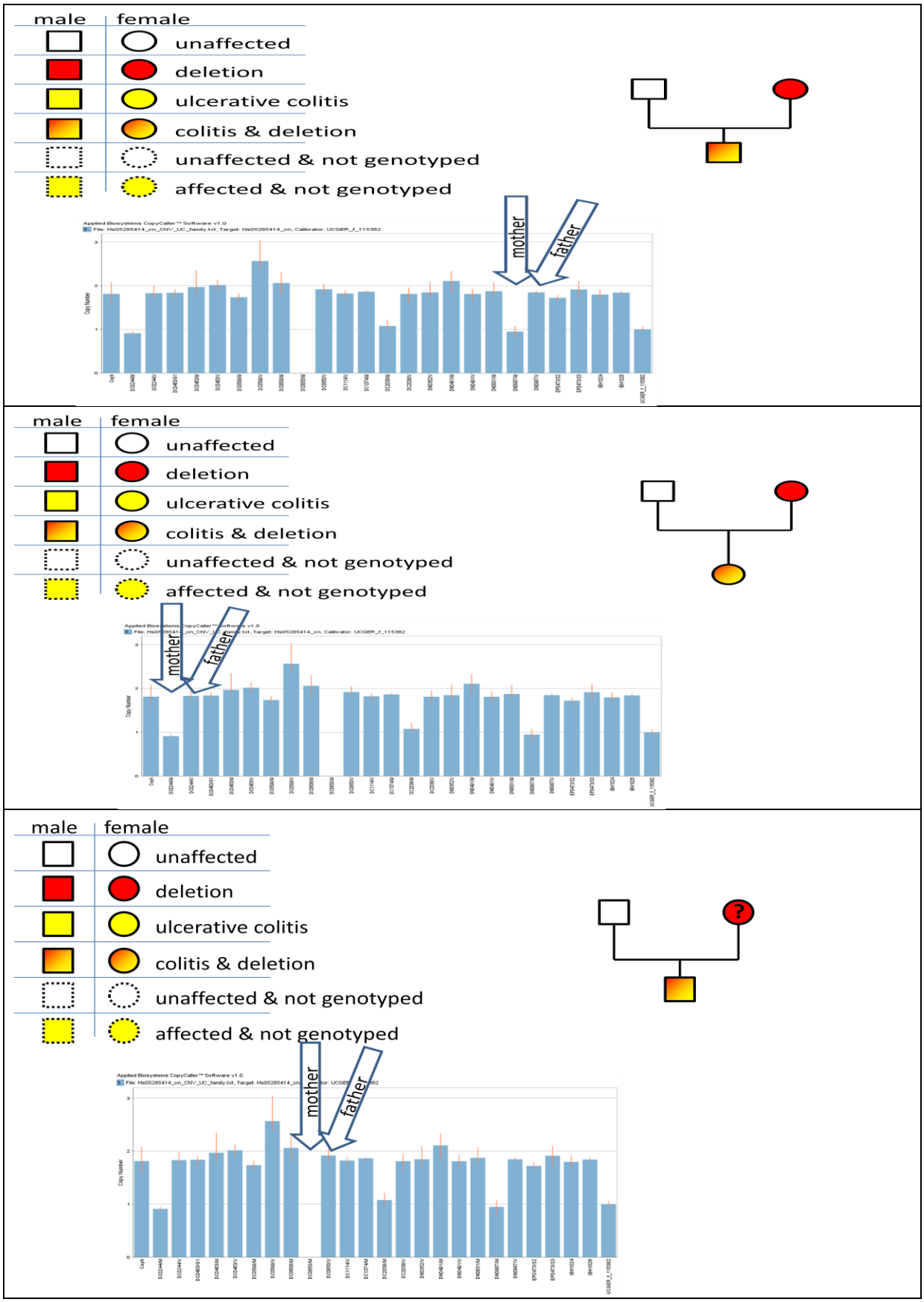

cases (UKUC)
20792717
20858698
20789
20804
20819
20834
20849
20864
Chr 10 : 20788990 - 20870999 (tick marks in kb)
8/2396 cases with 8 CNV(s)
controls (1958BC, NBS)
20792717
20858698
20789
20804
20819
20834
20849
20864
Chr 10 : 20788990 - 20870999 (tick marks in kb)
18/4886 controls with 18 CNV(s)
cases (UKUC)
103479459
104856868
103411
103686
103962
104237
104513
104788
Chr 11 : 103410589 - 104925738 (tick marks in kb)
6/2396 cases with 6 CNV(s)
controls (1958BC, NBS)
103479459
104856868
103411
103686
103962
104237
104513
104788
Chr 11 : 103410589 - 104925738 (tick marks in kb)
12/4886 controls with 12 CNV(s)
cases (UKUC)
23959634
24021752
23957
23969
23981
23994
24006
24019
Chr 13 : 23956529 - 24024857 (tick marks in kb)
3/2396 cases with 3 CNV(s)
controls (1958BC, NBS)
23959634
24021752
23957
23969
23981
23994
24006
24019
Chr 13 : 23956529 - 24024857 (tick marks in kb)
2/4886 controls with 2 CNV(s)

Table 6.9 CNV load in German screening panel and UK replication panel. Table shows the results of the CNV load analysis performed with CNVineta (Wittig et al. 2010). The columns from left to right are the sample set, the size of the sample set, median deletions per sample, median duplications per sample and median CNVs per sample. The next three columns are the calculated p-values based on the CNV counts per sample. P-values were obtained by Wilcoxon rank sum test.

	samples	deletions per sample	duplications per sample	CNVs per sample	pValue-Del	pValue-Dup	pValue-CNV
German cases	902	44	18	62	$< 2.2\times10^{-16}$	$< 2.2\times10^{-16}$	$< 2.2\times10^{-16}$
German Controls	1564	39	17	56			
UK cases	2396	43	17	61	9.5×10^{-3}	0.24	0.17
UK controls	4886	43	18	61			

Table 6.10 Inheritance of Deletion 13q32.1 verified by TaqMan® CNV assays. For all confirmed deletion carriers of the initial screening, for which samples of family members were available, the samples were analyzed with TaqMan® copy number assays. Each graph below represents a family with a legend at the top left, the TaqMan® signal intensities at the bottom and the family tree at the right. For some families samples are missing. The results show that all complete family trees show Mendelian inheritance and that the incomplete family trees at least do not show any inconsistencies.




male
female
unaffected
deletion
ulcerative colitis
colitis & deletion
unaffected & not genotyped
affected & not genotyped
mother
father
?

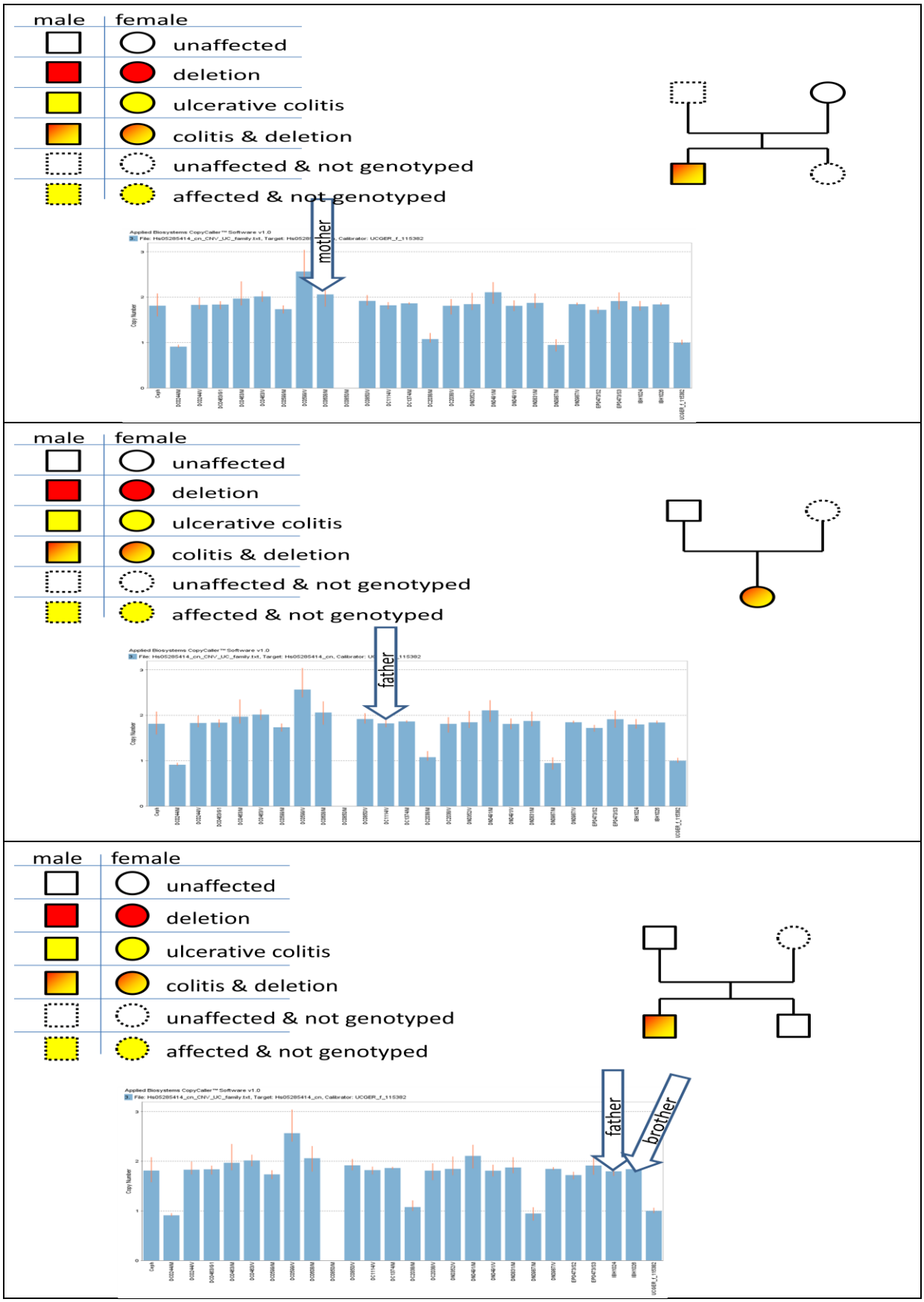
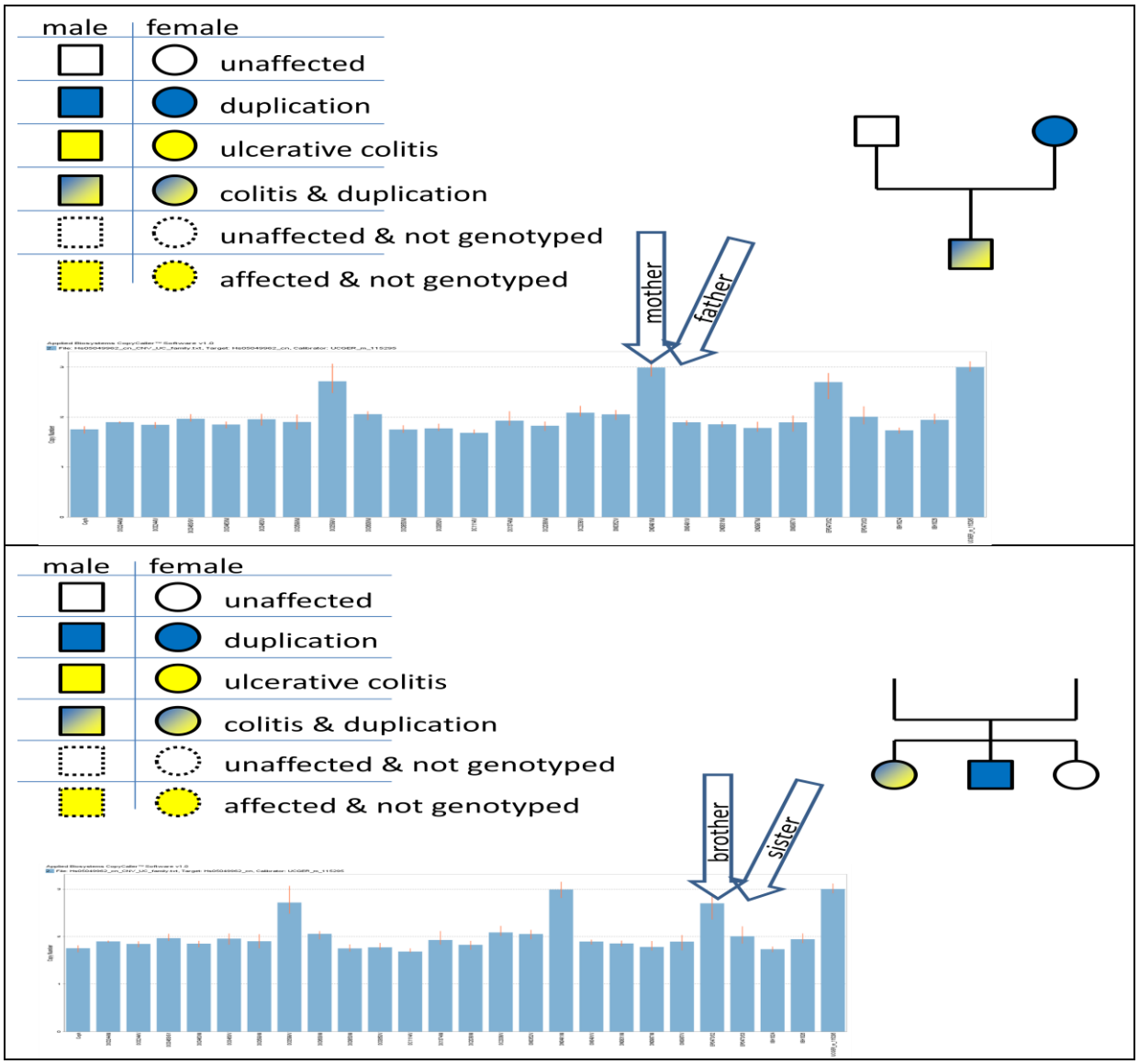

male
female
unaffected
deletion
ulcerative colitis
colitis & deletion
unaffected & not genotyped
affected & not genotyped
mother
male
female
unaffected
deletion
ulcerative colitis
colitis & deletion
unaffected & not genotyped
affected & not genotyped
father
male
female
unaffected
deletion
ulcerative colitis
colitis & deletion
unaffected & not genotyped
affected & not genotyped
father
brother

Table 6.11 Inheritance of Duplication 8q24.3 verified by TaqMan® CNV assays. For all confirmed duplication carriers of the initial screening, for which samples of family members were available, the samples were analyzed with TaqMan® copy number assays. Each graph below represents a family with a legend at the top left, the TaqMan® signal intensities at the bottom and the family tree at the right. For some families samples are missing. The results show that all complete family trees show Mendelian inheritance and that the incomplete family trees at least do not show any inconsistencies.



7 Summary

Background and aims

Crohn's disease (CD) and ulcerative colitis (UC), the two main subtypes of inflammatory bowel disease (IBD) are chronic relapsing inflammatory disorders of the gastrointestinal tract, which are characterized by excessive immune responses to commensal gut microbial flora in genetically susceptible individuals. Genetic contribution to IBD has been mainly interrogated through genome-wide association studies (GWAs) of mostly common single nucleotide polymorphisms (SNPs). However, the identified risk loci confer only modest effects on disease susceptibility and it has been assumed that the majority of the genetic risk remains to be explained. Copy number variations (CNVs), as one major form of genetic variations, comprise insertions, deletions and duplications of genomic sequences and have been shown to be involved in the pathogenesis of some common disease phenotypes. To examine whether genomic CNVs contribute to IBD risk, two parallel studies were conducted in this thesis; first, by recruiting an existing SNP-GWAs data set of UC and 4 other independent UC cohorts, a multi-step genome-wide association analysis was performed to interrogate the presence of disease-relevant germline CNVs. Second, nine monozygotic (MZ) twin pairs, discordant for IBD, were compared genome-wide to explore whether somatic CNVs might have contributed to their IBD discordance.

Results

1121 German UC patients and 1770 healthy controls were initially screened for rare and common deletions and duplications employing SNP-array data. Quantitative PCR, high density custom array-CGH and sequencing were used for validation of identified CNVs as well as fine (break-point) mapping. Twenty-four copy number variants (14 deletions and 10 duplications) overrepresented in UC patients were identified in the initial screening panel. Follow-up of these CNV regions in 4 independent case-control series as well as an additional public in silico control group (totaling 4,439 UC patients and 15,961 healthy controls) revealed three copy number variants enriched in UC patients; a 15.8 kb deletion (0.43% cases, 0.11% controls) upstream of ABCC4 and CLDN10 at 13q32.1, a 119 kb duplication (0.13% cases, 0.01% controls) at 7p22.1, overlapping RNF216, ZNF815, OCM and CCZ1 and a 134 kb large duplication (0.22% carriers among cases, 0.03% carriers among controls) upstream of the KCNK9 gene at 8q24.3. Break-point mapping of the deleted region suggested non-allelic homologous recombination as the mechanism underlying its formation. Expression analysis of the two nearby genes of Del 13q32.1 in intestinal biopsies of a UC patient panel showed differential expression of CLDN10 correlated with the deletion.

Additionally, genomic DNA from peripheral blood as well as intestinal biopsies from nine monozygotic twin pairs discordant for IBD manifestations (4 CD, 5 UC) were recruited and compared for genome-wide CNVs by means of array-comparative genomic hybridization (array-CGH), followed by quantitative PCR and (or) sequencing. Initial CNV calls were also contrasted with expression data of the affected genes in the corresponding biopsy samples. No consistent copy number differences were however revealed in the genomic DNA of discordant twins. Post-zygotic genomic CNVs appear not to be the common cause of IBD-discordance in MZ twins.

Conclusion

The pragmatic approach for effective CNV analysis in this thesis implicated the potential contribution of germline structural variants in the risk of UC. Nevertheless follow-up studies in larger disease cohorts as well as further functional experimental assays are needed to verify the disease relevance of the genetic loci presented here.

No findings of somatic variations within twins examined in this thesis would lend a further support for the notion that environmental and (or) epigenetic factors are the key drivers of IBD discordance in MZ twins.

8 Zusammanfassung

Morbus Crohn und Colitis Ulcerosa sind die beiden häufigsten chronisch-entzündlichen Darmerkrankungen (CED). Diese Erkrankungen treten in genetisch prädisponierten Individuen auf und werden vermutlich durch Überreaktionen des Immunsystems auf ein krankhaft verändertes Darmmikrobiom ausgelöst. Die genetische Veranlagung für CED wurde vor meinen Promotionsarbeiten überwiegend durch genomweite Assoziationsstudien (GWAS) anhand von Einzelnukleotid-Polymorphismen im menschlichen Genom („single nucleotide Polymorphisms“, SNPs) untersucht. Die bisherigen, durch SNP-GWAS identifizierten Risikovarianten besitzen meistens niedrige Risikoeffekte (mediane Effektstärke von 1,1) und erklären deshalb nur einen geringen Anteil (<20%) der kumulativen genetischen Varianz der CED bei Patienten europäischer Abstammung. Man nimm daher an, dass die Mehrheit der genetischen Risikofaktoren für CED noch nicht identifiziert sind. Kopienzahländerungen („Copy number variations“, CNVs) sind eine weitere wichtige Klasse von genetischen Varianten. Hierbei handelt es sich um Insertionen, Deletionen und Duplikationen von DNA-Sequenz-Abschnitten im Genom, die nach allgemeinem Konsensus eine Länge von mindestens 1000 Basenpaaren aufweisen. Für einige CNVs sind bereits vor meinen Arbeiten zum Thema signifikante Assoziationen mit komplexen Erkrankungen gefunden worden. Um die Frage zu beantworten, ob CNVs auch zum CED-Risiko beitragen, habe ich im Rahmen dieser Promotionsarbeit zwei Studien durchgeführt. In meiner ersten Studie habe ich den Datensatz einer SNP-basierten Colitis Ulcerosa GWAS-Studie untersucht, um krankheitsassoziierte CNVs mittels eines mehrstufigen, genomweiten Fall-Kontroll-Analyseverfahrens zu identifizieren. In meiner zweiten Studie habe ich die genomische DNA von sechs CED-diskordanten eineiigen Zwillingen verglichen, um potentielle krankheitsrelevante somatische Mutationen (hier wiederum CNVs) zu detektieren.

Ergebnisse

Im ersten Schritt habe ich 1121 Colitis ulcerosa Patienten und 1770 Kontrollen mittels SNP-Array-Daten auf Insertionen, Deletionen und Duplikationen untersucht. In diesem ersten Screening wurden 24 seltene CNVs (14 Deletionen und 10 Duplikationen) identifiziert, die in Colitis Ulcerosa Patienten überrepräsentiert sind.

Zur Validierung und Feinkartierung der auf diese Weise identifizierten CNVs habe ich Replikationsanalysen an weiteren Kohorten mittels quantitativer PCR bzw. mittels Array-CGH mit hoher Sondendichte durchgeführt. Charakterisierung der initial gefundenen 24 CNV-Regionen mittels Replikationskohorten und zusätzlichen *in-silico*-Kontrollgruppe (insgesamt 4.439 Patienten und 15.961 Gesundheits-Kontrollen) ergab schließlich drei CNVs, die in Patienten häufiger auftraten, nämlich eine 15,8-kb große Deletion vorgelagert vor den Genen

ABCC4 und *CLDN10* auf Chromosom 13q32.1 (0,43% der Fällen vs. 0,11% der Kontrollen), eine 119 kb große Duplikation auf Chromosom 7p22.1, die die Gene *RNF216*, *ZNF815*, *OCM* und *CCZ1* im Ganzen bzw. zum Teil enthält (0,13% der Fälle vs. 0,01% der Kontrollen), und letztlich eine 134 kb große Duplikation vorgelagert vor dem Gen *KCNK9* auf 8q24,3 (0,22% der Fälle vs. 0,03% der Kontrollen). Eine Bruchpunktkartierungsanalyse deutete darauf hin, dass die Deletion durch den Mechanismus der nicht-allelischen homologen Rekombination entstanden sein könnte.

In der Zwillingsstudie habe ich CNVs in genomischer DNA aus den Blutproben sowie Darm-Biopsie-Proben von sechs CED-Patienten (3 Morbus Crohn, 3 Colitis Ulcerosa) und ihren gesunden eineiigen Zwillingen untersucht und verglichen. Hierzu wurden vergleichende Genom-Hybridisierung (Array-CGH) und anschließend quantitative PCR verwendet. Es ließen sich allerdings keine Unterschiede bezüglich der Kopienzahlvarianten in den genomischen Kandidatenregionen der Zwillinge feststellen. Eine hohe Anzahl von CNVs wurde durch die Array-CGH-Analysesoftware ermittelt, die sich aber als falsch-positiv erwiesen.

Schlussfolgerung

In dieser Arbeit wurde ein pragmatischer Ansatz für ein effektives CNV-Screening entwickelt, der den potentiellen Beitrag von drei seltenen genomischen Strukturvarianten zum Risiko von Colitis Ulcerosa impliziert. Es sind weitere Nachuntersuchungen in größeren Kohorten sowie funktionelle Studien erforderlich, um die Assoziation der in dieser Arbeit vorgestellten CED-assoziierten CNV-Loci ultimativ zu belegen. Der fehlende Nachweis von somatischen CNV-Varianten zwischen diskordanten eineiigen Zwillingen in dieser Arbeit deutet darauf hin, dass eher andere genomische Varianten, epigenetischen Modifikationen, oder Umweltfaktoren die Hauptfaktoren der CED-Diskordanz bei eineiigen Zwillingen erklären könnten.

Curriculum vitae

Personal Data

Name.....Hamidreza Saadati

Place of birth.....Esfahan, Iran

Citizenship .....Iranian

Address .....Kopperpahler Allee. 84
D-24119 Kronshagen
Germany

University and scientific background

10/1999 - 8/2003..... **Bachelor** study in Cellular and Molecular Biology,
University of Tehran, Iran

10/2003 - 6/2006.....**Master** study in Biochemistry, University of Tehran
Master Thesis: Analysis of the mutations of the tumor suppressor gene "PTEN" in prostate cancer patients of Iran

08/2006 – 10/2008..... Researcher at Institute of Genetics and Biotechnology, University of Tehran, Iran

03/2009 – 10/2013.....PhD work at Institute of Clinical Molecular Biology, CAU Kiel

10/2013 – 03/2017.....Study Informatics, CAU Kiel

Publications

Saadati HR, Wittig M, Helbig I, Häsler R, Anderson CA, Mathew CG, Kupcinskas L, Parkes M, Karlsen TH, Rosenstiel P, Schreiber S, Franke A. Genome-wide Rare Copy Number Variation Screening in Ulcerative Colitis Identifies potential Susceptibility Loci. BMC Med Genet. 2016 Apr 1;17:26

Skieceviciene J, Kiudelis G, Ellinghaus E, Balschun T, Jonaitis LV, Zvirbliene A, Denapiene G, Leja M, Pranculiene G, Kalibatas V, **Saadati HR**, Ellinghaus D, Andersen V, Valantinas J, Irnius A, Derovs A, Tamelis A, Schreiber S, Kupcinskas L, Franke A. Replication study of ulcerative colitis risk loci in a Lithuanian-Latvian case-control sample**.** Inflamm Bowel Dis. 2013 Oct;19(11):2349-55

Fischer A, Nothnagel M, Franke A, Jacobs G, **Saadati HR**, Gaede KI, Rosenstiel P, Schürmann M, Müller-Quernheim J, Schreiber S, Hofmann S. Association of inflammatory bowel disease risk loci with sarcoidosis, and its acute and chronic subphenotypes. Eur Respir J. 2011 Mar;37(3):610-6

Pourmand G, Ziaee AA, Abedi AR, Mehrsai A, Alavi HA, Ahmadi A, **Saadati HR.** Role of PTEN gene in progression of prostate cancer. Urol J. 2007 4(2):95-100

Declaration

Erklärung

Hiermit erkläre ich, daß diese Dissertation, abgesehen von der Beratung durch meine akademischen Lehrer, nach Inhalt und Form meine eigene Arbeit ist. Sie hat weder im Ganzen noch zum Teil an anderer Stelle im Rahmen eines Promotionsverfahrens vorgelegen.

Ein Teil von Ergebnissen in dieser Arbeit wurden in "BMC Med Genet. 2016 Apr 1;17:26" veröffentlicht (siehe "Publications" auf Seite 178). Daher ist diese Erklärung lediglich zusammen mit der Sektion „Author Contributions" in der oben genannten Referenz vollständig. Die Vorarbeiten der „rare CNV" Studie wurde von Herrn Michael Wittig, dem Zweitautor der Studie durchgeführt.

Declaration

Apart from the advice of my supervisors, this thesis is the result of my own work. No part of it has been submitted to any other board for another qualification.

Some of the results in this thesis have been already published in "BMC Med Genet. 2016 Apr 1;17:26" (see "List of Publications" on page 178). Therefore this declaration is only completed with considerartion of section "Autor Contribution" in the given refrence.

Kiel, 12.12.2016

Hamidreza Saadati

Acknowledgment

This study has been only possible with the kind support of many competent and great people. Therefore;

I am very grateful to ***Prof. Dr. Stefan Schreiber***, for providing me with the chance of doing my PhD work at institute of Clinical Molecular Biology with excellent working conditions, large resources of patient materials and biomolecular technological facilities.

I would like to thank ***Prof. Dr. Andre Franke***, for his supervision during my PhD thesis and for his continuous support as well as valuable guidance and comments on this thesis.

I am very thankful to ***Prof. Dr. Manuela Dittmar*** and ***Prof. Dr. Philip Rosenstiel*** for their valuable considerations and comments on this thesis and their kind scientific supports.

I would like to thank ***Michael Wittig*** for developing tools for CNV analysis as well as his continuous helps and supports during my PhD work. I am especially grateful to him, as this work has not been applicable without his general helps and supports and his outstanding scientific inputs.

I am very thankful to ***Michael Forster*** for revision of the German summery of this thesis.

I would like to thank the ***entire staff*** of DNA lab, Genotyping lab, sequencing lab and IT group of Institute of Clinical Molecular Biology, University of Kiel.

I owe my deepest thanks to ***my family***, for their everlasting support and encouragement throughout all these years.