Petersen et al. BMC Genetics (2017) 18:14
DOI 10.1186/s12863-017-0479-5

BMC Genetics

REVIEW Open Access

Opportunities and challenges of whole-genome and -exome sequencing

Britt-Sabina Petersen, Broder Fredrich, Marc P. Hoeppner, David Ellinghaus and Andre Franke*

Abstract

Recent advances in the development of sequencing technologies provide researchers with unprecedented possibilities for genetic analyses. In this review, we will discuss the history of genetic studies and the progress driven by next-generation sequencing (NGS), using complex inflammatory bowel diseases as an example. We focus on the opportunities, but also challenges that researchers are facing when working with NGS data to unravel the genetic causes underlying diseases.

Keywords: Whole-genome sequencing, WGS, Whole-exome sequencing, WES, Next-generation sequencing, NGS, Complex diseases, Inflammatory bowel diseases, Variant priorization, Variants of unknown significance (VUS)

Background

Studies of human genetic variation using DNA sequencing have undergone an extraordinary development from their introduction over 40 years ago up to current technologies, which allow for a human genome to be sequenced and analyzed within a matter of days at consumable costs of approximately 1000 USD. The first widely used method was developed by the British chemist Frederick Sanger in the 1970s [1] and he received the Nobel prize in 1980. “Sanger sequencing” relies on nucleotide-specific chain-terminating inhibitors to identify the sequence of a specific fragment of DNA. The method was continuously refined over the years and incorporated in the first generation of automated sequencers. Sanger sequences show very high accuracy but are restricted to a single DNA fragment at a time and a maximum sequence length of 1000 bp. In addition to the low throughput, high costs render this technology unsuitable for routine large scale sequencing projects. The largest effort using the Sanger technique was the Human Genome Project with the goal of identifying the complete sequence of the human genome [2], which is in essence based on different donors from Buffalo (New York, USA) [3, 4]. Completion of the project took over a decade (1990–2003), involved more than 20 institutes all over the world and cost nearly 3 billion dollars. Still, for many years, Sanger sequencing was the prevailing technique to identify causative mutations in monogenic diseases. However, the limitations of the technology meant that finding the one gene responsible for a disease was tedious work. Rather than performing large-scale, indiscriminate sequencing, numerous experiments were often necessary to narrow down candidate regions from microsatellite-based linkage studies and pinpoint to one or a few candidate genes that would then be sequenced. In most cases, these experiments required samples from large pedigrees with several affected individuals to successfully identify candidate regions small enough for further analysis. These issues are further amplified in the study of genetically heterogeneous diseases with causative variants in a number of genes or very large genes, as well as diseases that do not follow a Mendelian inheritance pattern, but instead have a complex genetic background involving tens to hundreds of genes. The common disease-common variant hypothesis assumes that a large part of the heritability of these complex diseases can be attributed to variants with a minor allele frequency above 1% (single nucleotide polymorphisms, SNPs) in the general population, each variant having a small additive or multiplicative effect on the disease phenotype. Addressing questions as complex as these clearly required novel approaches.

However, it was not until the introduction of high-throughput genotyping in the early 2000s, enabling the interrogation of several hundred thousand to millions of genotypes in thousands of cases and controls [5], that genome-wide association studies (GWAS) became a

* Correspondence: a.franke@mucosa.de
Institute of Clinical Molecular Biology, Kiel University, Kiel, Germany

reality. For the first time a quick and unbiased screening of SNPs throughout the whole genome was possible, thereby facilitating the detection of susceptibility regions for complex diseases. What followed can be referred to as the "GWAS era", with genome-wide case-control association studies carried out for numerous complex diseases, identifying more than 25,000 significantly disease-associated genetic loci until today [6]. GWAS studies primarily focused on common SNPs, excluding rare variants. Later approaches like Illumina's human exome genotyping array [7] shifted the focus to include rare, exonic variants. However, it soon became clear that genotyping efforts alone were not sufficient to completely uncover the genetics behind complex diseases [8].

The release of the first "next-generation" sequencing instruments (NGS; see [9] for an overview) in the mid-2000s led to a first revolution in disease study, offering vastly improved speed at significantly lower cost - enabling the generation of a whole human genome sequence in a matter of weeks for 10,000 USD by 2011 [10]. In addition to price and performance, the new sequencing technology also proved to compensate for some of the technical weaknesses of the older sequencing and genotyping technologies, allowing for the genome-wide detection of variants, including novel ones, at a low cost. However, despite the immense drop in sequencing costs for a human genome, large-scale sequencing projects were still costly and therefore not yet carried out for thousands of samples as routinely done in GWAS.

In 2007, Craig Venter published the first diploid genome sequence of a single individual, which was created using the gold-standard Sanger sequencing technology, and which is perhaps still among the most accurate and best-annotated human genomes released to the public domain [11]. However, DNA materials of the donor are, to our knowledge, not available to the public for benchmarking and follow-up studies. This year however, the academic Genome in a Bottle Consortium provided extensive NGS data on seven genomes, including two trios, which serve as open benchmarking data and materials [12].

The next breakthrough for NGS in human genomics arrived with the introduction of targeted enrichment methods, allowing for selective sequencing of regions of interest [13] and thereby dramatically reducing the amount of sequences that needed to be generated. The approach is based on a collection of DNA or RNA probes representing the target sequences in the genome, which can bind and extract the DNA fragments originating from these targeted regions. Whole exome sequencing (WES), which enables sequencing of all protein-coding regions in the human genome (the exome) quickly became the most widely used targeted enrichment method, especially for monogenic ("Mendelian") diseases. This approach enabled the detection of both exonic (coding) as well as splice-site variants, while requiring only approximately 2% of sequencing "load" compared to whole genome sequencing (WGS). The unbiased analysis of all genes eliminates the need for a time-consuming selection of candidate genes prior to sequencing. It has been estimated that the exome harbors about 85% of mutations with large effects on disease-related traits [14]. In addition, exonic mutations were shown to cause the majority of monogenic diseases [15], with missense and nonsense mutations alone accounting for approximately 60% of disease mutations [16]. While these numbers may be in part biased by the difficulty of identifying disease-causing mutations in non-coding regions, the success of exome sequencing studies for monogenic diseases confirms its advantages. In the years following its introduction, exome sequencing led to a vast increase in the identification of Mendelian disease genes [17, 18]. This is reflected for example in almost 2000 new entries in OMIM since 2008 (current total: 4787), describing the molecular basis of a particular phenotype.

Current large-scale genome and exome sequencing projects [19–22] have not only provided crucial information on variant frequencies in different populations, but have also shown that a human genome typically contains an estimated 100 genuine loss-of-function variants, completely inactivating around 20 genes [23]. Therefore, sequencing of healthy individuals or representative population samples can also lead to important insights into disease. Focusing on seemingly "healthy human knockouts" can aid in detecting the true effects of variants previously assumed to be disease-causing [24] and exploring gene function in general, thus elucidating the "resilience" phenomenon further [25].

In recent years, NGS has also been increasingly applied for addressing pharmacogenomic research questions. It is not only possible to detect genetic causes that explain why some patients do not respond to a certain drug, but also try to predict a drug's success based on genetic information [26]. Certain genetic variants can affect the activity of a particular protein and these can be used to estimate the probable efficacy and toxicity of a drug targeting such a protein [27]. NGS therefore has applications far beyond finding disease-causing variants. For inflammatory bowel diseases (IBD) we refer to the exhaustive pharmacogenomics review of Katsanos and colleagues [28].

Progress of genetic research for common complex diseases

Some of the diseases that profited immensely from GWAS are inflammatory bowel diseases. Together with ulcerative colitis (UC), Crohn's disease (CD) is one of the two main sub-phenotypes of IBD. IBD are chronic, relapsing disorders involving inflammation of the

gastrointestinal tract, sometimes accompanied by extra-intestinal manifestations. The disease onset can occur at any age, but the peak for CD as well as UC is in early adulthood (approximately 25 to 35 years of age). In the clinic, symptoms include chronic flare-ups of inflammation, abdominal cramping pain as well as diarrhea and weight loss, thereby greatly affecting the quality of life of patients. In Europe, an estimated 1.4 million people suffer from CD [29] but as of yet, there is no known cure and the current treatment is solely aimed at controlling the symptoms. The current consensus is that IBD is caused by the complex interplay of an overly active immune system and environmental triggers (such as bacterial infections, dietary habits or smoking) in genetically susceptible individuals [30, 31]. The strong genetic component, especially for CD, is reflected by familial clustering of disease occurrence and a concordance of 35% in monozygotic but only 3% in dizygotic twin pairs [32]. The relative risk for developing IBD is estimated to be 15 times higher for first degree relatives of an IBD patient than in the general population [33].

Due to the complex nature of IBD, genetic research focused on the identification of genetic risk factors that increase the susceptibility to the disease, typically common SNP alleles that are significantly more frequent in patients than in healthy controls. The aforementioned methods have all contributed to the discovery of genetic risk factors for IBD in the past. Genome-wide linkage and candidate gene studies during the late 1990s were able to identify the first susceptibility loci for IBD through positional cloning and candidate gene analysis. The first susceptibility gene to be identified for CD was *NOD2* [34, 35], encoding for a member of the cytoplasmic nucleotide-binding oligomerization domain (NOD)-like receptor (NLR) protein family. Over the years, several association studies added significantly to the number of identified loci [36, 37], followed by meta-analyses which combined the data of several individual GWAS-studies from all over the world. The larger sample sizes led to more statistical power and eventually to the discovery of numerous additional susceptibility loci [38–40]. Today, more than 200 loci have been identified for IBD [41] and have highlighted some key pathways involved in the etiology of IBD. Figure 1 illustrates the success of hypothesis-free genome-wide studies. For example, our group first unveiled the link of autophagy to IBD by identifying *ATG16L1* in a genome-wide candidate SNP screen [42]. Before, *NOD2* had been identified as the first and so far best-replicated Crohn's disease susceptibility gene through two independent studies [34, 35]. Gene identification is then ideally followed by numerous validation and in particular functional/mechanistic studies of the respective candidate genes. Bringing disease genes on the radar of the research community leads to further studies, then

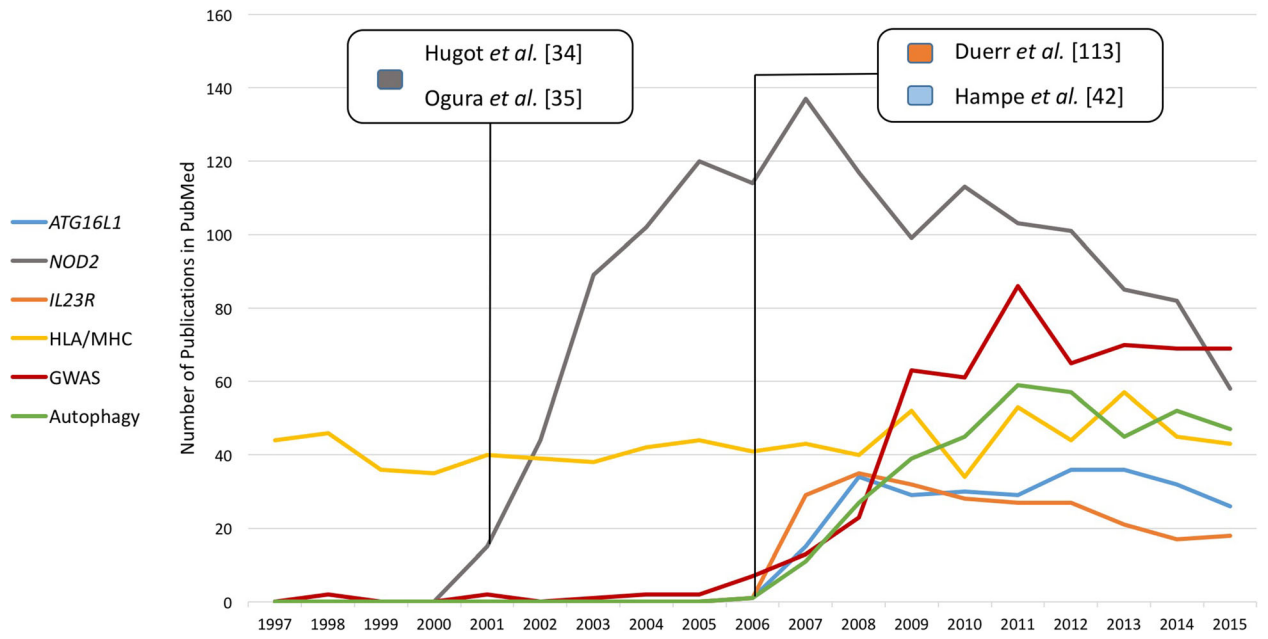

Fig. 1 Number of PubMed citations for *ATG16L1*, *NOD2*, *IL23R*, HLA/MHC, GWAS and autophagy in combination with "inflammatory bowel disease", "Crohn's disease" OR "ulcerative colitis" from the years 1997–2015 depicting a steep increase of follow-up studies for genes and pathways after discovery. Interestingly, the HLA/MHC association signal in IBD has been known for a long time, however, studies for this locus in IBD are rarer and no increase can be observed. We think that this region is understudied, given its importance in disease etiology (in particular in ulcerative colitis), calling for more IBD immunogenetics studies in the future

scientific publications (as shown by the steep increase of publications per year in Fig. 1) and ultimately an improved disease understanding. However, the variants identified in GWAS still explain less than 30% of the estimated genetic variance of IBD [40]. While IBD constitutes perhaps one of the greatest success stories and role models in complex disease genetics research, GWAS have also been quite effective for a number of other complex diseases like psoriasis [43–45], atopic dermatitis [46, 47] and primary sclerosing cholangitis (PSC) [48, 49]. Combined analyses of several of these immune-mediated diseases have even revealed considerable overlap of susceptibility loci, pointing at true pleiotropy and shared disease etiologies beyond CD and UC, while also showing complex disease-specific patterns at shared loci as well as revealing disease-specific loci [50, 51].

Interestingly, the genetic susceptibility factors for IBD are, with a few exceptions (e.g. *NOD2*, *TNFSF15*, *HLA*), the same in European-ancestry and East-Asian IBD patients [41]. Similar results have been obtained for other complex inflammatory diseases, such as systemic lupus erythematosus [52]. Therefore, trans-ethnic studies will clearly aid in identifying consistent genetic risk loci for complex inflammatory diseases, thus increasing also statistical power due to the larger sample sizes. The few differences in the genetic risk maps may also help in pinpointing likely existing different environmental factors in the countries under study.

As previously indicated, GWAS studies focused on SNPs with moderate to high allele frequencies in the general population. A part of the so-called missing heritability may however be found in rare variants with larger effect sizes [53] for some diseases. Results of a recent large-scale sequencing project of more than 2600 genomes and almost 13,000 exomes did not support the idea that lower-frequency variants have a major role in predisposition to type 2 diabetes [54]. For IBD, however, common and rare susceptibility variants have been shown to even coexist in the same genes, as is the case for *NOD2* [34, 55, 56]. Figure 2a illustrates this wide range of IBD-relevant variants concerning their penetrance and the genetic disease complexity and provides an overview of the identified genes from monogenic, fully penetrant genes to those harboring common susceptibility variants. Rare and especially novel variants can best be detected by DNA sequencing and the development of NGS finally made this feasible for complex diseases. Figure 2b depicts the discovery of IBD genes since 2001 employing the different technologies discussed here and shows the great success of GWAS on the one hand, but also the increasing relevance of NGS during the past few years.

Application of NGS for complex diseases

The usage of NGS and especially exome sequencing for Mendelian disorders proved to be extremely successful. Even sequencing of just a single patient could lead to the discovery of the genetic mutation responsible for the disease by filtering the detected variants based on functional consequence (e.g. missense, nonsense, splice-site variants) and allele frequency in the general population, for example in the data of the 1000 genomes project [19, 57] or the Exome Aggregation Consortium (ExAC) [20]. But when dealing with complex diseases, different approaches need to be considered.

One possibility is the application of the GWAS approach to NGS data, aiming for the identification of significant differences between cases and controls. Disease-associated common variants can best be detected by GWAS and sequencing approaches have the potential to complement this by discovering rare variant associations, given that the necessary large sample sizes are considered. However, with a decreasing allele frequency, the power to detect genes or variants of interest also decreases, if effect sizes are small to moderate. Single marker association testing is therefore often "underpowered" for rare variants with frequencies below 0.5% or even 0.1% minor allele frequency (MAF), since the number of observations of such alleles is often not large enough to achieve statistical significance due to small sample sizes [58]. For example, observing an allele once with 0.5 or 0.1% MAF with 99% probability requires sequencing of at least 460 or 2300 individuals, respectively. Assuming a disease-associated variant with 0.1% MAF and an allelic odds ratio (OR) of 1.4, the sample size (cases and controls with equal sized groups) required to achieve 80% power is 540,000, given a disease prevalence of 5% and a significance level of 5×10^{-8} [59]. However, the commonly used significance level of 5×10^{-8} is valid for approximately one million common tag SNPs (MAF ≥ 5%) only if a linkage disequilibrium $r^2 < 0.8$ for pairs of tag SNPs is applied. With 0.1% MAF, we would need a *P*-value threshold level of 1×10^{-8} and $3 \times 10^{-7}$to meet genome-wide and exome-wide significance (at $r^2 < 0.8$), respectively [60]. Several statistical methods have been proposed in the past to perform case-control studies with WES or WGS data, most of them using variant aggregation approaches to address this issue. The two main types of aggregation tests comprise burden and variance component tests [58] or a mixture of both. Burden tests [61, 62] compare the number of variants in a certain region or gene between cases and controls, while variance component tests (e.g. the sequence-based kernel association test, SKAT [63]) can distinguish between protective and risk variants in a single gene, making them more powerful if the gene possesses a mixture of protective and risk variants.

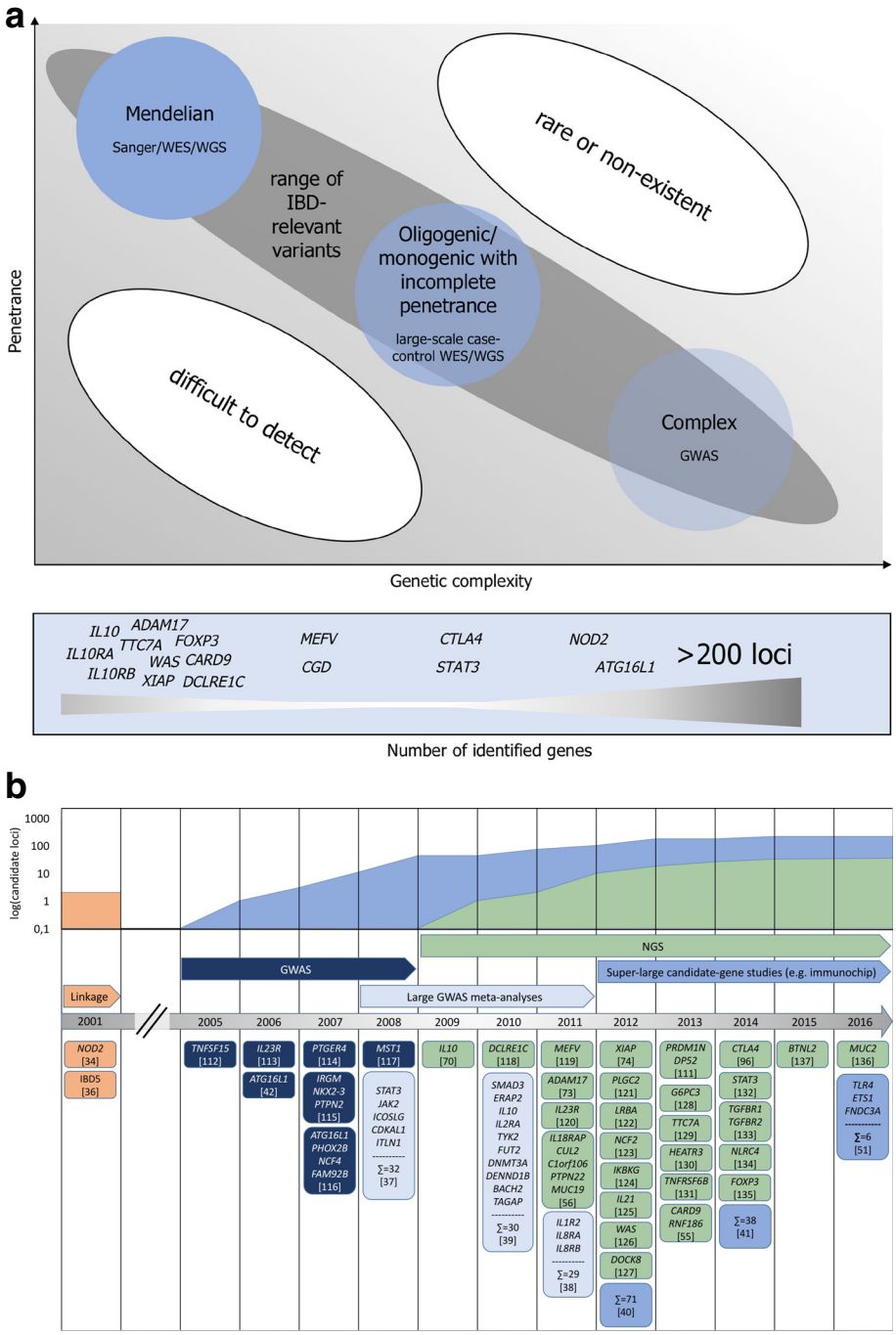

Fig. 2 a Top: Range of IBD-relevant variants based on genetic complexity underlying the disease and variant penetrance. Bottom: Overview of identified IBD genes ranging from monogenic to complex forms based on the highest known penetrance for each gene. For both *NOD2* and *PRDM1*, for example, both common and rare variants have been identified as disease-relevant in patients [110, 111]. **b** Timeline of gene discovery for IBD [112–137]. Top graph shows cumulative number of genes separated by technology (log scale)

Lee et al. [59] provide a comprehensive overview of currently available algorithms.

The successful application of these tests is however limited by sample size, as sequencing studies involving WES or WGS still require a significantly larger sample size than a typical GWAS to identify significant rare variant associations [64]. Despite continuously decreasing prices for sequencing, case-control studies employing thousands of

individuals still remain a costly undertaking compared to GWAS where the latest array generation (e.g. Global Screening Array from Illumina with 700,000 variants) is currently available for less than $40 per sample. The NGS approach is therefore still restricted to large-scale sequencing centers, companies, healthcare providers (e.g. Geisinger Health), consortia involving several institutes and crowd-sourced approaches.

Therefore, the focus currently lies on the analysis of unusual cases to find highly penetrant variants. Here, one possibility is sequencing of large families with several affected individuals to narrow the dataset down to few candidate variants based on those shared by the affected individuals [65]. Clustering of patients within a pedigree may point to variants with larger effects on disease compared to those identified in GWAS and even monogenic forms of IBD. However, it can also simply indicate the accumulation of a large number of common susceptibility alleles in the pedigree [65] and exome sequencing may therefore not necessarily be successful. Apart from multiplex families, the most informative characteristics indicating the presence of a highly penetrant genetic cause are an early age of onset and very severe course of disease. GWAS performed specifically for pediatric IBD (age of onset <18 years) failed to clearly distinguish early onset from adult IBD, identifying known IBD loci or exclusive pediatric loci that were later also identified for adult IBD [66, 67]. There is great overlap between susceptibility genes identified for pediatric and adult-onset IBD (more than 30 loci described [66]). Early-onset cases of IBD, with a disease manifestation during the first 10 years of life, often show a more severe disease course with a higher risk of complications and a higher frequency of indeterminate colitis (IC) diagnoses [68]. Patients classified as very-early-onset even develop the disease during the first 6 years of life. A large spectrum of monogenic diseases, mainly immunodeficiencies, can also present with IBD-like intestinal inflammation [69]. However, several studies have also identified shared genetic factors underlying these monogenic, early-onset and adult-onset IBD cases with rather oligogenic or polygenic causes. Mutations in genes for the interleukin 10 receptor (*IL10R*) subunit proteins [70] and the *IL10* gene itself [71] were shown to be responsible for several cases of severe early-onset IBD (eoIBD). At the same time, *IL10* was also associated with adult-onset UC [72] and CD [39] in GWAS. Other identified causes of eoIBD include a deletion in *ADAM17* (ADAM metallopeptidase domain 17) [73] and mutations or deletions in the *XIAP* (X-linked inhibitor of apoptosis) gene [74–76] in male patients. Although the direct overlap between key genes associated with IBD and IBD-like monogenic disorders is rather low, the affected proteins often interact directly or indirectly with each other and share common signaling cascades that contribute to IBD etiology [69]. Results from monogenic forms therefore have the potential to give important insights into mechanisms contributing to disease. An excellent overview of the genetics of early- and very early-onset forms of IBD is the review by Uhlig et al. [77].

Targeted resequencing of susceptibility regions has also been applied for several immune diseases and has identified additional rare, functional variants in susceptibility genes, which were detected using common variants in GWAS. For instance, gene resequencing for atopic dermatitis identified low-frequency missense variants in the *GARP* gene as significant contributors to disease risk [78]. Perhaps not surprisingly, monogenic disease forms of complex diseases—i.e. patients that carry variants with very high penetrance—have not exclusively been detected for IBD, but also for other diseases. For example, monogenic forms of psoriasis caused by mutations in *CARD14* [79] were revealed through exome sequencing of a family with early-onset psoriasis. Studies of rare variants in Mendelian forms of disorders that are symptomatically similar to systemic lupus erythematosus (SLE) have highlighted pathways also playing a role in the complex disease form. As another example, *TREX1*, encoding for the three prime repair exonuclease 1, has been associated with monogenic Aicardi-Goutières syndrome [80], a disease displaying phenotypic overlap with SLE. More recently, 0.5% of SLE patients were shown to also harbor mutations in this gene [81].

While sequencing of severe early-onset patient exomes greatly facilitated the identification of novel, high penetrance variants, their discovery among the tens of thousands of variants identified in an exome is still a major challenge. Since the first exome studies that relied on allele frequencies from the 1000 genomes pilot [82], several large-scale sequencing studies for genomes and exomes have been undertaken. Databases like EVS [83], ExAC [20] and KAVIAR [84] now provide population-specific allele frequencies from several thousands to more than 60,000 individuals that can be used for filtering of candidate variants. However, some of these databases are "contaminated" with data from patients or yet unknown patients of similar symptoms as the disease of interest, so the data should be used with caution.

The interpretation of non-coding variants has proven to be extremely challenging. The ENCODE project [85] significantly facilitated the understanding of functional elements in the human genome. However, the complex analysis of these sites is not yet routinely carried out in most projects. For exome data, the analysis of non-coding variants is limited from the beginning, due to the nature of the technology with exclusive enrichment of exons and, in some cases, UTRs. Variants from exomes therefore tend to be reduced to those that are most likely to affect protein structure. Nonsense, start-loss,

stop-loss and splice-site variants as well as frameshift insertions and deletions (InDels) have rather clearly defined effects on the protein and are present in comparably low numbers. The interpretation of sometimes hundreds of rare missense variants represents a greater challenge. Several *in silico* prediction tools are available to identify those amino acid changes that most likely affect protein structure. SIFT [86] and Polyphen-2 [87] were the first widely used tools, more recently DANN [88], CADD [89] and FATHMM [90] were introduced. The latter promise improved accuracy and additionally offer predictions for non-coding variants. Other tools specifically focus on identifying splice-altering variants, including those located farther away from the exon-intron boundary [91, 92]. Genes also differ concerning the amount of potentially disruptive genetic variation they can tolerate, expressed for example by the residual variation intolerance score (RVIS) [93]. The prediction of the effect of a variant on the protein structure and thereby its function is however only one of the levels that need to be considered when aiming to detect disease-relevant variants. Variants diminishing the function of a gene do not necessarily manifest as an observable phenotype. This can for example be due to redundancy of the function in several genes, preventing the deficiency of one from having an effect. Filtering and priorization of variants based on these criteria can already significantly reduce the number of candidates. In some cases, this is sufficient to identify a likely causative variant relying on the known function of a well characterized gene or novel variants in a known disease gene. In most cases, however, additional filtering is needed. In general, it is helpful to analyze more than one individual of a family, even when dealing with sporadic cases, since this allows the identification of variants segregating with the disease within the pedigree. For sporadic cases the healthy parents can also be used to detect *de novo* mutations in the patient. These filtering steps can, however, still result in a number of variants remaining, without being able to clearly identify the most likely candidate. Novel genes that haven't previously been implicated in disease or even genes with an unknown function substantially complicate the search. The question then arises, how to proceed with a handful of candidates with a possible but unconfirmed pathogenic effect (variants of unknown significance, VUS) that remain after filtering with all available methods. Functional analyses, especially for genes that are not yet well characterized, can be time-consuming and expensive.

The case of a family with Crohn's disease and autoimmunity in two children [94] nicely illustrates this issue. Exome sequencing was performed and yielded several candidates, among them a rare missense variant in *CTLA4* (Cytotoxic T lymphocyte-associated protein 4). While it represented a likely candidate, it was also present in the asymptomatic mother. This incomplete penetrance, as well as other candidate variants, made interpretation and priorization difficult. Also, heterozygous CTLA-4 deficiency in mice does not induce a phenotype [95], which made the role of the detected variant in disease questionable. Additional evidence pointing to *CTLA4* finally emerged when variants were also identified in other patients with immune phenotypes [96, 97] and functional studies were able to back the role of heterozygous *CTLA4* variants in immune dysregulation. The incomplete penetrance suggests that additional modifying factors yet need to be revealed, requiring the analysis of additional patients with CTLA-4 deficiency.

Developing infrastructure for data sharing

Reliably classifying disease-causing variants often involves finding correlations between different, independent observations, i.e. patients or cohorts with similar clinical phenotypes in which the same (or a functionally related) variant has been observed. For very rare or private variants only a second patient with the same symptoms and the same genetic variant is sufficient for statistical proof of the original finding. Sources of information are usually published studies and public data repositories that need to be searched, manually or with specifically set up local bioinformatics pipelines. However, the complexity of the data at hand (including sometimes dozens of VUS for larger patient cohorts) as well as the vast amount of sequences that is now routinely being generated and deposited, is calling for more efficient and integrated approaches.

Several efforts exist that aim to specifically aggregate relevant clinical data, including databases such as Decipher [98], HGMD [99] or ClinVar [100]. Complementary to these resources, projects are under way to better link national infrastructures and communities. Of note here are, for example, the Belgian "SymBioSys" (http://www.kuleuven.be/symbiosys/) or the German "VarWatch" project (BMBF project ID01EK1506 [101]), both targeting separate issues in the integration of NGS data and clinical variants. The main goal of SymBioSys is to leverage national NGS data and provide efficient access. It does so by building a federated network across sequencing facilities, together with a generic interface that helps in rapidly mining the data for identical variants or study parameters. VarWatch, on the other hand, is focused directly on the clinical context and is designed to function as both a repository and a "monitoring" tool. Clinicians can submit their VUS, together with phenotypic information about the disease, and VarWatch will continuously search for matching cases, both within its own data repository as well as external resources.

While these initiatives are potentially important building blocks towards generating comprehensive clinical resources, they leave the larger issue of how to efficiently access and integrate the globally accumulating information about the genetics of individual patients and their conditions unanswered. A solution that is finding strong support amongst larger databases and bioinformatics institutes is currently being developed by the "Global Alliance for Genomics and Health" (GA4GH), an international consortium of clinicians and bioinformaticians with the goal of providing standards and software for sharing clinical data on a global scale. One product of these activities has been the "Beacon" network, and in extension "MatchMaker Exchange" (MME) [102]. The focus of Beacon and MME is to provide a "connective tissue" between various "information islands", linking databases through a common interface and enabling simple, platform agnostic queries without having to create huge aggregations of data. Databases connected to the beacon network can easily be queried for the presence of specific variants. MME further extends this concept, allowing users not only to find identical variants, but also to include information about the clinical context of the variant (such as observed phenotype). In doing so, it can bring together clinicians and researchers with patients whose variants are not strictly identical, but potentially related on a functional level and thus further help finding diagnoses. Figure 3 depicts the variant filtering of one real-world example from our clinic for trio exome sequencing. While the filtering steps are able to reduce the number of variants from more than 67,000 to only 18 variants potentially of interest, it is still difficult to select the best candidate among these VUS or *in limbo* variants. One possible solution for this problem is the usage of MME which can detect overlaps between the VUS submitted by different scientists or clinicians and establish contact between them, making it possible to pinpoint the causative variant(s) and thus solve the clinical case (statistical significant result through recurrent finding of very rare event).

It is also becoming increasingly clear that in addition to efficient access to distributed variant information, there is also a growing need for metadata standards to describe clinical observations not only genetically, but also phenotypically. While several vocabularies have been proposed over the years, the ones in use – such as the Unified Medical Language System (UMLS) [103] - are focused more on syndromes and less so on the

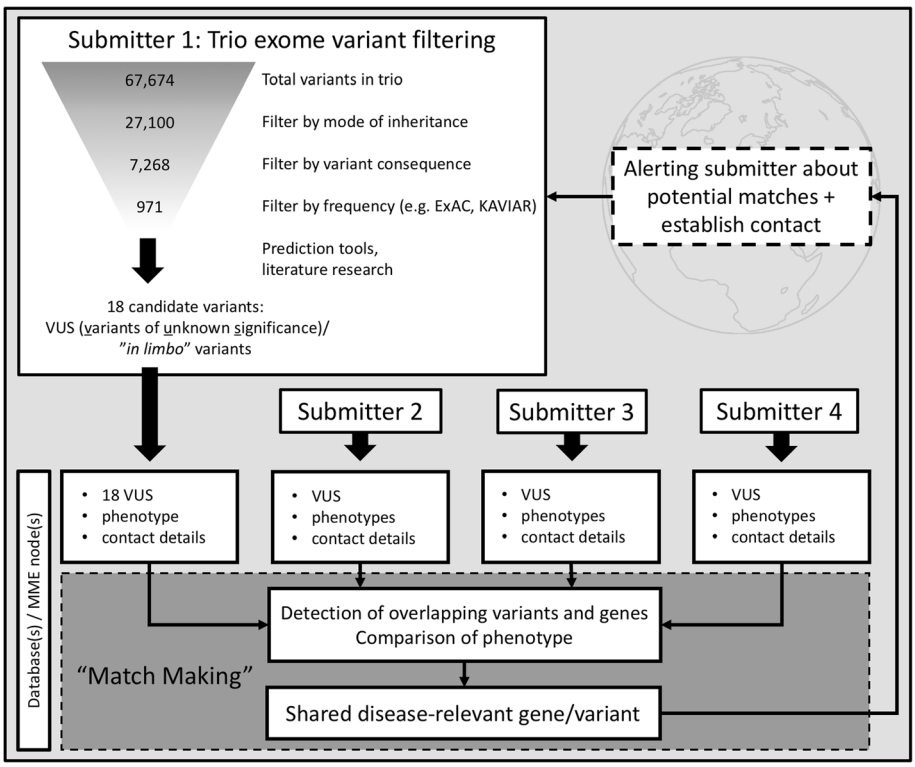

Fig. 3 Course of a typical trio exome project yielding several VUS and benefit of MME for variant selection. Filter by mode of inheritance: recessive or dominant; by variant consequence: missense, nonsense, splice-site, start-loss or stop-loss; by frequency: maximum minor allele frequency of 1% in various databases (ExAC, EVS, in-house controls)

symptoms a patient is presenting with. This information however will likely be vital, especially when trying to match rare variants and rare diseases with poor representation in standard nomenclatures. A promising solution has recently been proposed in the form of the Human Phenotype Ontology (HPO) [104], a collection of hierarchical, phenotypic descriptors organized in an ontological network similar to the already well-established sequence ontology (SO) [105] and gene ontology (GO) [106]. In addition to providing a complementary nomenclature for clinicians to better characterize their findings, the inherent network-like structure of the HPO also allows to measure the distance between any two terms. This enables more complex matching scenarios, for examples when clinicians have used slightly different but related terms or sets of terms to characterize their patients [107].

Future directions

Combination of methods

Apart from genome and exome sequencing, which we focus on here, there are several other NGS applications that we expect to increase in relevance as efforts are concentrating on linking observed mutations to functional consequences beyond putative coding changes. Of great interest here are the detection of modulation in gene activity, for which both the direct sequencing of transcripts through RNAseq as well as the detection of differentially methylated sites (DMS) by means of bisulfite sequencing as proxy for regulation hold great promise. A completely different but equally important line of inquiry is the metagenomic sequencing of the host-associated microbiome to detect possible correlations between the presence or absence of certain genera and disease, as has already been suggested for a decrease in Bacteroides and Firmicutes and a reduced diversity of the microbiota in IBD patients [108]. The combined application of these multi-omics data has the potential to provide an improved overall picture of the characteristics of a certain disease and therefore help to understand its molecular underpinnings. With sequencing costs further decreasing, large case-control studies with sample sizes comparable to GWAS are also slowly becoming a reality and will help detect rare variant associations specifically for complex diseases. The biggest challenge though remains, identifying relevant environmental factors in complex diseases. Genetics and other functional genomics analyses may also help in hinting at the disease-causing environmental factors.

Technological developments

WES and WGS allow for the accurate identification of single-nucleotide variants (SNVs) and small InDels. For the detection of large InDels, copy number variations (CNVs) as well as genomic rearrangements, however, deep sequencing and meticulous analyses are needed, which are mostly not yet part of the common analysis pipelines used in the majority of projects.

NGS is already being applied to the clinic for the diagnosis of certain diseases, mostly through deep sequencing of gene panels. However, relevant variants still require confirmation through Sanger sequencing due to the generally lower quality of NGS data, so it is desirable to further increase the quality of NGS in the near future.

New methods are continuously being developed to use NGS for additional applications or to extract more information from standard applications. 10X Genomics for example offers an additional instrument (Chromium), which is fully compatible with the workflows of available NGS sequencers and enables large-scale phasing of variants and structural variant detection from WGS and even WES as well as single cell applications by generating synthetic long reads. The Chromium instrument uses emulsion to partition DNA. Barcoding and amplification of smaller fragments from the original larger fragments then takes place in droplets called “GEMs” that include all necessary reagents, resulting in the small fragments stemming from one larger molecule carrying the same barcode. These “synthetic long reads” can therefore be linked over larger regions of the genome. The workflow delivers ready-to use libraries for sequencing and software for the analysis and visualization is openly available.

Other companies opt for the development of new sequencing technologies, often called “third-generation sequencers”. Pacific Biosciences performs single molecule, real-time (SMRT) sequencing of DNA fragments using immobilized DNA-polymerases and produces reads of over 10,000 bp average length. Nanopore sequencers, like those developed by Oxford nanopore, detect the DNA sequence of a single-stranded DNA molecule by passing it through a protein pore and measuring a shift in voltage that originates from interactions with the pore. However, these single-molecule technologies are still too expensive and not yet applicable for resequencing larger numbers of complete human genomes.

Looking further into the future, several exciting new technologies are on the horizon. Genia Technologies, which was bought by Roche in 2014, is currently developing a nanopore-based sequencing technique with a focus on diagnostic applications. First results have already been published [109], showing promising proof-of-principle results. However, it will likely still take several years until the method is ready for the market. Illumina is planning the launch of a new semiconductor sequencer in 2017 as part of their Project Firefly, but as of yet, no details have been released to the public. GenapSys by Sigma Aldrich promises a low-cost, portable sequencer with a purely electronic sequencing chip, but more information is currently only available to its testers.

Conclusions

The extraordinary progress in the development of methods for genomic analysis during the past 15 years and especially the breakthrough in NGS in the past decade has led to an enormous increase in the understanding of the human genome and its relation to disease. Improved technologies continuously provide faster, cheaper and more accurate results, allowing us to move from gene panels to exomes to routinely sequencing whole genomes in the clinic in the near future. It has however become increasingly clear, that to make the most of the large, complex datasets being generated, scientists must work together more than ever, to achieve the ultimate goal of translating genomic data into clinically actionable results that patients can directly profit from. With the generation of genomics data continuously becoming easier and cheaper, the interpretation of the large amounts of data and the identification of the relevant disease-causing environmental factors will remain the biggest challenges of the years to come.

Abbreviations

CD: Crohn's disease; CNV: Copy number variation; DMS: Differentially methylated site; ExAC: Exome Aggregation Consortium; GA4GH: Global Alliance for Genomics and Health; GO: Gene ontology; GWAS: Genome-wide association study; HPO: Human Phenotype Ontology; IBD: Inflammatory bowel disease; InDels: Insertions and deletions; MAF: Minor allele frequency; MME: MatchMaker Exchange; NGS: Next-generation sequencing; OR: Odds ratio; PSC: Primary sclerosing cholangitis; RVIS: Residual variation intolerance score; SKAT: Sequence-based kernel association test; SLE: Systemic lupus erythematosus; SMRT: Single molecule, real-time; SNP: Single nucleotide polymorphisms; SNV: Single-nucleotide variant; SO: Sequence ontology; UC: Ulcerative colitis; VUS: Variants of unknown significance; WES: Whole exome sequencing; WGS: Whole genome sequencing

Acknowledgement

The authors would like to thank Thomas F. Wienker (IKMB, Kiel University) for his thoughtful comments on the manuscript.

Funding

The authors are supported by the DFG Cluster of Excellence "Inflammation at Interfaces", the Deutsche Forschungsgemeinschaft (DFG) grant FR 2821/6-1 and the Bundeministerium für Bildung und Forschung E:med/SysInflame grant number 012X1306F.

Availability of data and materials

Not applicable.

Authors' contributions

BP and AF designed the manuscript. All authors wrote the manuscript. BP, BF and AF prepared the figures. All authors read and approved the final manuscript.

Competing interests

The authors declare that they have no competing interests.

Ethics approval and consent to participate

Not applicable.

Received: 27 September 2016 Accepted: 26 January 2017
Published online: 14 February 2017

References

1. Sanger F, Nicklen S, Coulson AR. DNA sequencing with chain-terminating inhibitors. Proc Natl Acad Sci. 1977;74:5463–7.
2. Lander ES, Linton LM, Birren B, Nusbaum C, Zody MC, Baldwin J, et al. Initial sequencing and analysis of the human genome. Nature. 2001;409:860–921. Macmillian Magazines Ltd.
3. Osoegawa K, Mammoser AG, Wu C, Frengen E, Zeng C, Catanese JJ, et al. A bacterial artificial chromosome library for sequencing the complete human genome. Genome Res. 2001;11:483–96.
4. Tuzun E, Sharp AJ, Bailey JA, Kaul R, Morrison VA, Pertz LM, et al. Fine-scale structural variation of the human genome. Nat Genet. 2005;37:727–32.
5. Bumgarner R. Overview of DNA microarrays: types, applications, and their future. Curr Protoc Mol Biol. 2013;Chapter 22:Unit 22.1.
6. Burdett T, Hall PN, Hastings E, Hindorff LA, Junkins HA, Klemm AK, MacArthur J, Manolio TA, Morales J, Parkinson H WD. The NHGRI-EBI Catalog of published genome-wide association studies. [Internet]. [cited 2016 Jul 4]. Available from: www.ebi.ac.uk/gwas.
7. Guo Y, He J, Zhao S, Wu H, Zhong X, Sheng Q, et al. Illumina human exome genotyping array clustering and quality control. Nat Protoc. 2014;9:2643–62. Nature Research.
8. Manolio TA, Collins FS, Cox NJ, Goldstein DB, Hindorff LA, Hunter DJ, et al. Finding the missing heritability of complex diseases. Nature. 2009;461:747–53.
9. Goodwin S, McPherson JD, McCombie WR. Coming of age: ten years of next-generation sequencing technologies. Nat Rev Genet. 2016;17:333–51.
10. Wetterstrand K. DNA Sequencing Costs: Data from the NHGRI Genome Sequencing Program (GSP) Accessed May 2016 [Internet]. [cited 2016 May 30]. Available from: https://www.genome.gov/sequencingcostsdata/.
11. Levy S, Sutton G, Ng PC, Feuk L, Halpern AL, Walenz BP, et al. The diploid genome sequence of an individual human. PLoS Biol. 2007;5:e254.
12. Zook JM, Catoe D, McDaniel J, Vang L, Spies N, Sidow A, et al. Extensive sequencing of seven human genomes to characterize benchmark reference materials. Sci Data. 2016;3:160025.
13. Teer JK, Mullikin JC. Exome sequencing: the sweet spot before whole genomes. Hum Mol Genet. 2010;19:R145–51.
14. Majewski J, Schwartzentruber J, Lalonde E, Montpetit A, Jabado N. What can exome sequencing do for you? J. Med Genet. 2011;48:580–9.
15. Kuhlenbäumer G, Hullmann J, Appenzeller S. Novel genomic techniques open new avenues in the analysis of monogenic disorders. Hum Mutat. 2011;32:144–51.
16. Botstein D, Risch N. Discovering genotypes underlying human phenotypes: past successes for mendelian disease, future approaches for complex disease. Nat Genet. 2003;33(Suppl):228–37.
17. Ng SB, Turner EH, Robertson PD, Flygare SD, Bigham AW, Lee C, et al. Targeted capture and massively parallel sequencing of 12 human exomes. Nature. 2009;461:272–6.
18. Bamshad MJ, Ng SB, Bigham AW, Tabor HK, Emond MJ, Nickerson D a., et al. Exome sequencing as a tool for Mendelian disease gene discovery. Nat Rev Genet. 2011;12:745–55.
19. Abecasis GR, Auton A, Brooks LD, DePristo MA, Durbin RM, Handsaker RE, et al. An integrated map of genetic variation from 1,092 human genomes. Nature. 2012;491:56–65. Nature Publishing Group, a division of Macmillan Publishers Limited. All Rights Reserved.
20. Lek M, Karczewski KJ, Minikel EV, Samocha KE, Banks E, Fennell T, et al. Analysis of protein-coding genetic variation in 60,706 humans. Nature. 2016;536:285–91.
21. Narasimhan V, Hunt KA, Mason D, Baker CL, Karczewski KJ, Barnes MR, et al. Health and population effects of rare gene knockouts in adult humans with related parents. Science (80-). 2016;8624:1–8.
22. Sulem P, Helgason H, Oddson A, Stefansson H, Gudjonsson SA, Zink F, et al. Identification of a large set of rare complete human knockouts. Nat Genet. 2015;47:448–52. Nature Publishing Group, a division of Macmillan Publishers Limited. All Rights Reserved.
23. MacArthur DG, Balasubramanian S, Frankish A, Huang N, Morris J, Walter K, et al. A systematic survey of loss-of-function variants in human protein-coding genes. Science. 2012;335:823–8. Europe PMC Funders.
24. Alkuraya FS. Human knockout research: New horizons and opportunities. Trends Genet. 2015;31:108–15.
25. Narasimhan VM, Xue Y, Tyler-Smith C. Human Knockout Carriers: Dead, Diseased, Healthy, or Improved? Trends Mol Med. 2016;22:341–51.
26. Nelson MR, Tipney H, Painter JL, Shen J, Nicoletti P, Shen Y, et al. The support of human genetic evidence for approved drug indications. Nat Genet. 2015;47:856–60. Nature Research.

27. Plenge RM, Scolnick EM, Altshuler D. Validating therapeutic targets through human genetics. Nat Rev Drug Discov. 2013;12:581–94. Nature Research.
28. Katsanos KH, Papadakis KA. Pharmacogenetics of inflammatory bowel disease. Pharmacogenomics. 2014;15:2049–62.
29. Schirbel A, Fiocchi C. Inflammatory bowel disease: Established and evolving considerations on its etiopathogenesis and therapy. J Dig Dis. 2010;11:266–76.
30. Baumgart DC, Sandborn WJ. Crohn's disease. Lancet. 2012;380:1590–605.
31. Ellinghaus D, Bethune J, Petersen B-S, Franke A. The genetics of Crohn's disease and ulcerative colitis–status quo and beyond. Scand J Gastroenterol. 2015;50:13–23.
32. Spehlmann ME, Begun AZ, Burghardt J, Lepage P, Raedler A, Schreiber S. Epidemiology of inflammatory bowel disease in a German twin cohort: results of a nationwide study. Inflamm Bowel Dis. 2008;14:968–76.
33. Satsangi J, Rosenberg W, Jewell D. The prevalence of inflammatory bowel disease in relatives of patients with Crohn's disease. Eur J Gastroenterol Hepatol. 1994;6:413–6.
34. Hugot JP, Chamaillard M, Zouali H, Lesage S, Cézard JP, Belaiche J, et al. Association of NOD2 leucine-rich repeat variants with susceptibility to Crohn's disease. Nature. 2001;411:599–603.
35. Ogura Y, Bonen DK, Inohara N, Nicolae DL, Chen FF, Ramos R, et al. A frameshift mutation in NOD2 associated with susceptibility to Crohn's disease. Nature. 2001;411:603–6.
36. Rioux JD, Daly MJ, Silverberg MS, Lindblad K, Steinhart H, Cohen Z, et al. Genetic variation in the 5q31 cytokine gene cluster confers susceptibility to Crohn disease. Nat Genet. 2001;29:223–8.
37. Barrett JC, Hansoul S, Nicolae DL, Cho JH, Duerr RH, Rioux JD, et al. Genome-wide association defines more than 30 distinct susceptibility loci for Crohn's disease. Nat Genet. 2008;40:955–62.
38. Anderson CA, Boucher G, Lees CW, Franke A, D'Amato M, Taylor KD, et al. Meta-analysis identifies 29 additional ulcerative colitis risk loci, increasing the number of confirmed associations to 47. Nat Genet. 2011;43:246–52.
39. Franke A, McGovern DPB, Barrett JC, Wang K, Radford-Smith GL, Ahmad T, et al. Genome-wide meta-analysis increases to 71 the number of confirmed Crohn's disease susceptibility loci. Nat Genet. 2010;42:1118–25. Nature Publishing Group.
40. Jostins L, Ripke S, Weersma RK, Duerr RH, McGovern DP, Hui KY, et al. Host-microbe interactions have shaped the genetic architecture of inflammatory bowel disease. Nature. 2012;491:119–24.
41. Liu JZ, van Sommeren S, Huang H, Ng SC, Alberts R, Takahashi A, et al. Association analyses identify 38 susceptibility loci for inflammatory bowel disease and highlight shared genetic risk across populations. Nat Genet. 2015;47:979–86. Nature Publishing Group.
42. Hampe J, Franke A, Rosenstiel P, Till A, Teuber M, Huse K, et al. A genome-wide association scan of nonsynonymous SNPs identifies a susceptibility variant for Crohn disease in ATG16L1. Nat Genet. 2007;39:207–11.
43. Stuart PE, Nair RP, Ellinghaus E, Ding J, Tejasvi T, Gudjonsson JE, et al. Genome-wide association analysis identifies three psoriasis susceptibility loci. Nat Genet. 2010;42:1000–4.
44. Tsoi LC, Spain SL, Knight J, Ellinghaus E, Stuart PE, Capon F, et al. Identification of 15 new psoriasis susceptibility loci highlights the role of innate immunity. Nat Genet. 2012;44:1341–8. Nature Research.
45. Ray-Jones H, Eyre S, Barton A, Warren RB. One SNP at a Time: Moving beyond GWAS in Psoriasis. J Invest Dermatol. 2016;136:567–73.
46. Hoffjan S, Stemmler S. Unravelling the complex genetic background of atopic dermatitis: from genetic association results towards novel therapeutic strategies. Arch Dermatol Res. 2015;307:659–70.
47. Bin L, Leung DYM. Genetic and epigenetic studies of atopic dermatitis. Allergy Asthma Clin Immunol. 2016;12:52.
48. Henriksen EKK, Melum E, Karlsen TH. Update on primary sclerosing cholangitis genetics. Curr Opin Gastroenterol. 2014;30:310–9.
49. Karlsen TH, Chung BK. Genetic risk and the development of autoimmune liver disease. Dig Dis. 2015;33(2):13–24.
50. Fodil N, Langlais D, Gros P. Primary immunodeficiencies and inflammatory disease: a growing genetic intersection. Trends Immunol. 2016;37:126–40.
51. Ellinghaus D, Jostins L, Spain SL, Cortes A, Bethune J, Han B, et al. Analysis of five chronic inflammatory diseases identifies 27 new associations and highlights disease-specific patterns at shared loci. Nat Genet. 2016;48:510–8.
52. Morris DL, Sheng Y, Zhang Y, Wang YF, Zhu Z, Tombleson P, et al. Genome-wide association meta-analysis in Chinese and European individuals identifies ten new loci associated with systemic lupus erythematosus. Nat Genet. 2016;48:940–6.
53. Bodmer W, Bonilla C. Common and rare variants in multifactorial susceptibility to common diseases. Nat Genet. 2008;40:695–701. Nature Publishing Group.
54. Fuchsberger C, Flannick J, Teslovich TM, Mahajan A, Agarwala V, Gaulton KJ, et al. The genetic architecture of type 2 diabetes. Nature. 2016;536:41–7.
55. Beaudoin M, Goyette P, Boucher G, Lo KS, Rivas MA, Stevens C, et al. Deep resequencing of GWAS loci identifies rare variants in CARD9, IL23R and RNF186 that are associated with ulcerative colitis. PLoS Genet. 2013;9:e1003723. Gibson G, editor.
56. Rivas MA, Beaudoin M, Gardet A, Stevens C, Sharma Y, Zhang CK, et al. Deep resequencing of GWAS loci identifies independent rare variants associated with inflammatory bowel disease. Nat Genet. 2011; in press.
57. 1000 Genomes Project Consortium T. A map of human genome variation from population-scale sequencing. Nature. 2010;467:1061–73.
58. Kosmicki JA, Churchhouse CL, Rivas MA, Neale BM. Discovery of rare variants for complex phenotypes. Hum Genet. 2016;135:625–34.
59. Lee S, Abecasis GR, Boehnke M, Lin X. Rare-variant association analysis: study designs and statistical tests. Am J Hum Genet. 2014;95:5–23.
60. Fadista J, Manning AK, Florez JC, Groop L. The (in)famous GWAS P-value threshold revisited and updated for low-frequency variants. Eur J Hum Genet. 2016;24:1–4.
61. Asimit JL, Day-Williams AG, Morris AP, Zeggini E. ARIEL and AMELIA: testing for an accumulation of rare variants using next-generation sequencing data. Hum Hered. 2012;73:84–94.
62. Morgenthaler S, Thilly WG. A strategy to discover genes that carry multi-allelic or mono-allelic risk for common diseases: a cohort allelic sums test (CAST). Mutat Res. 2007;615:28–56.
63. Wu MC, Lee S, Cai T, Li Y, Boehnke M, Lin X. Rare-variant association testing for sequencing data with the sequence kernel association test. Am J Hum Genet. 2011;89:82–93.
64. Bansal V, Libiger O, Torkamani A, Schork NJ. Statistical analysis strategies for association studies involving rare variants. Nat Rev Genet. 2010;11(11):773–85. Nature Publishing Group.
65. Stittrich AB, Ashworth J, Shi M, Robinson M, Mauldin D, Brunkow ME, et al. Genomic architecture of inflammatory bowel disease in five families with multiple affected individuals. Hum Genome Var. 2016;3:15060. Nature Publishing Group.
66. Imielinski M, Baldassano RN, Griffiths A, Russell RK, Annese V, Dubinsky M, et al. Common variants at five new loci associated with early-onset inflammatory bowel disease. Nat Genet. 2009;41:1335–40.
67. Essers JB, Lee JJ, Kugathasan S, Stevens CR, Grand RJ, Daly MJ, et al. Established genetic risk factors do not distinguish early and later onset Crohn's disease. Inflamm Bowel Dis. 2009;15:1508–14.
68. Prenzel F, Uhlig HH. Frequency of indeterminate colitis in children and adults with IBD - a metaanalysis. J Crohns Colitis. 2009;3:277–81.
69. Uhlig HH. Monogenic diseases associated with intestinal inflammation: implications for the understanding of inflammatory bowel disease. Gut. 2013;62:1795–805.
70. Glocker EO, Kotlarz D, Boztug K, Gertz EM, Schäffer A a, Noyan F, et al. Inflammatory bowel disease and mutations affecting the interleukin-10 receptor. N Engl J Med. 2009;361:2033–45.
71. Glocker EO, Frede N, Perro M, Sebire N, Elawad M, Shah N, et al. Infant colitis–it's in the genes. Lancet. 2010;376:1272.
72. Franke A, Balschun T, Karlsen TH, Sventoraityte J, Nikolaus S, Mayr G, et al. Sequence variants in IL10, ARPC2 and multiple other loci contribute to ulcerative colitis susceptibility. Nat Genet. 2008;40:1319–23.
73. Blaydon DC, Biancheri P, Di W-L, Plagnol V, Cabral RM, Brooke MA, et al. Inflammatory skin and bowel disease linked to ADAM17 deletion. N Engl J Med. 2011;365:1502–8.
74. Worthey EA, Mayer AN, Syverson GD, Helbling D, Bonacci BB, Decker B, et al. Making a definitive diagnosis: successful clinical application of whole exome sequencing in a child with intractable inflammatory bowel disease. Genet Med. 2011;13:255–62.
75. Zeissig Y, Petersen B-S, Milutinovic S, Bosse E, Mayr G, Peuker K, et al. XIAP variants in male Crohn's disease. Gut. 2015;64:66–76.
76. Kelsen JR, Dawany N, Martinez A, Grochowski CM, Maurer K, Rappaport E, et al. A de novo whole gene deletion of XIAP detected by exome sequencing analysis in very early onset inflammatory bowel disease:a case report. BMC Gastroenterol. 2015;15:160.

77. Uhlig HH, Schwerd T. From genes to mechanisms: the expanding spectrum of monogenic disorders associated with inflammatory bowel disease. Inflamm Bowel Dis. 2016;22:202–12.
78. Manz J, Rodríguez E, ElSharawy A, Oesau EM, Petersen BS, Baurecht H, et al. Targeted resequencing and functional testing identifies low-frequency missense variants in the gene encoding GARP as significant contributors to atopic dermatitis risk. J Invest Dermatol. 2016;136:2380–86.
79. Signa S, Rusmini M, Campione E, Gueli I, Grossi A, Omenetti A, et al. Severe erytrodermic psoriasis and arthritis as clinical presentation of a CARD14-mediated psoriasis (CAMPS). Pediatr Rheumatol. 2015;13:P57. BioMed Central.
80. Crow YJ, Hayward BE, Parmar R, Robins P, Leitch A, Ali M, et al. Mutations in the gene encoding the 3′-5′ DNA exonuclease TREX1 cause Aicardi-Goutières syndrome at the AGS1 locus. Nat Genet. 2006;38:917–20.
81. Namjou B, Kothari PH, Kelly JA, Glenn SB, Ojwang JO, Adler A, et al. Evaluation of the TREX1 gene in a large multi-ancestral lupus cohort. Genes Immun. 2011;12:270–9.
82. Marth GT, Yu F, Indap AR, Garimella K, Gravel S, Leong WF, et al. The functional spectrum of low-frequency coding variation. Genome Biol. 2011;12:R84.
83. Exome Variant Server. NHLBI Exome Sequencing Project (ESP), Seattle, WA. (http://evs.gs.washington.edu/EVS/). 2012.
84. Glusman G, Caballero J, Mauldin DE, Hood L, Roach JC. Kaviar: an accessible system for testing SNV novelty. Bioinformatics. 2011;27:3216–7.
85. Rosenbloom KR, Dreszer TR, Pheasant M, Barber GP, Meyer LR, Pohl A, et al. ENCODE whole-genome data in the UCSC Genome Browser. Nucleic Acids Res. 2010;38:D620–5.
86. Kumar P, Henikoff S, Ng PC. Predicting the effects of coding non-synonymous variants on protein function using the SIFT algorithm. Nat Protoc. 2009;4:1073–81.
87. Adzhubei IA, Schmidt S, Peshkin L, Ramensky VE, Gerasimova A, Bork P, et al. A method and server for predicting damaging missense mutations. Nat Methods. 2010;7:248–9. Nature Publishing Group.
88. Quang D, Chen Y, Xie X. DANN: a deep learning approach for annotating the pathogenicity of genetic variants. Bioinformatics. 2015;31:761–3.
89. Kircher M, Witten DM, Jain P, O'Roak BJ, Cooper GM, Shendure J. A general framework for estimating the relative pathogenicity of human genetic variants. Nat Genet. 2014;46:310–5.
90. Shihab HA, Rogers MF, Gough J, Mort M, Cooper DN, Day INM, et al. An integrative approach to predicting the functional effects of non-coding and coding sequence variation. Bioinformatics. 2015;31(10):1536–43. btv009-.
91. Jian X, Boerwinkle E, Liu X. In silico prediction of splice-altering single nucleotide variants in the human genome. Nucleic Acids Res. 2014;42:13534–44.
92. Xiong HY, Alipanahi B, Lee LJ, Bretschneider H, Merico D, Yuen RKC, et al. The human splicing code reveals new insights into the genetic determinants of disease. Science (80-). 2014;347:1254806. American Association for the Advancement of Science.
93. Petrovski S, Wang Q, Heinzen EL, Allen AS, Goldstein DB. Genic intolerance to functional variation and the interpretation of personal genomes. PLoS Genet. 2013;9:e1003709.
94. Zeissig S, Petersen B-S, Tomczak M, Melum E, Huc-Claustre E, Dougan SK, et al. Early-onset Crohn's disease and autoimmunity associated with a variant in CTLA-4. Gut. 2015;64:1889–97.
95. Waterhouse P, Penninger JM, Timms E, Wakeham A, Shahinian A, Lee KP, et al. Lymphoproliferative disorders with early lethality in mice deficient in Ctla-4. Science. 1995;270:985–8. American Association for the Advancement of Science.
96. Schubert D, Bode C, Kenefeck R, Hou TZ, Wing JB, Kennedy A, et al. Autosomal dominant immune dysregulation syndrome in humans with CTLA4 mutations. Nat Med. 2014;20:1410–6.
97. Kuehn HS, Ouyang W, Lo B, Deenick EK, Niemela JE, Avery DT, et al. Immune dysregulation in human subjects with heterozygous germline mutations in CTLA4. Science. 2014;345:1623–7. American Association for the Advancement of Science.
98. Firth HV, Richards SM, Bevan AP, Clayton S, Corpas M, Rajan D, et al. DECIPHER: Database of Chromosomal Imbalance and Phenotype in Humans Using Ensembl Resources. Am J Hum Genet. 2009;84:524–33.
99. Stenson PD, Mort M, Ball EV, Shaw K, Phillips A, Cooper DN. The Human Gene Mutation Database: building a comprehensive mutation repository for clinical and molecular genetics, diagnostic testing and personalized genomic medicine. Hum Genet. 2014;133:1–9.
100. Landrum MJ, Lee JM, Benson M, Brown G, Chao C, Chitipiralla S, et al. ClinVar: public archive of interpretations of clinically relevant variants. Nucleic Acids Res. 2016;44:D862–8.
101. Fredrich B, Wienker T, Junge O, Ellinhaus D, Franke A, Hoeppner M, et al. VarWatch – a database for in-limbo variants. Presented at the annual meeting of the Ferman Society for Human Genetics, Lübeck, Germany, 2016.
102. Philippakis AA, Azzariti DR, Beltran S, Brookes AJ, Brownstein CA, Brudno M, et al. The matchmaker exchange: a platform for rare disease gene discovery. Hum Mutat. 2015;36:915–21.
103. Bodenreider O. The Unified Medical Language System (UMLS): integrating biomedical terminology. Nucleic Acids Res. 2004;32:D267–70.
104. Köhler S, Doelken SC, Mungall CJ, Bauer S, Firth HV, Bailleul-Forestier I, et al. The Human Phenotype Ontology project: linking molecular biology and disease through phenotype data. Nucleic Acids Res. 2014;42:D966–74.
105. Eilbeck K, Lewis SE, Mungall CJ, Yandell M, Stein L, Durbin R, et al. The Sequence Ontology: a tool for the unification of genome annotations. Genome Biol. 2005;6:R44. BioMed Central.
106. Ashburner M, Ball CA, Blake JA, Botstein D, Butler H, Cherry JM, et al. Gene ontology: tool for the unification of biology. The Gene Ontology Consortium. Nat Genet. 2000;25:25–9. Nature Publishing Group.
107. Oellrich A, Collier N, Groza T, Rebholz-Schuhmann D, Shah N, Bodenreider O, et al. The digital revolution in phenotyping. Brief Bioinform. 2016;17:819–30.
108. Kostic AD, Xavier RJ, Gevers D. The microbiome in inflammatory bowel disease: current status and the future ahead. Gastroenterology. 2014;146:1489–99.
109. Fuller CW, Kumar S, Porel M, Chien M, Bibillo A, Stranges PB, et al. Real-time single-molecule electronic DNA sequencing by synthesis using polymer-tagged nucleotides on a nanopore array. Proc Natl Acad Sci U S A. 2016;113:5233–8. National Academy of Sciences.
110. Lesage S, Zouali H, Cézard J-P, Colombel J-F, Belaiche J, Almer S, et al. CARD15/NOD2 mutational analysis and genotype-phenotype correlation in 612 patients with inflammatory bowel disease. Am J Hum Genet. 2002;70:845–57.
111. Ellinghaus D, Zhang H, Zeissig S, Lipinski S, Till A, Jiang T, et al. Association Between Variants of PRDM1 and NDP52 and Crohns Disease, Based on Exome Sequencing and Functional Studies. Gastroenterology. 2013;145:339–47.
112. Yamazaki K, McGovern D, Ragoussis J, Paolucci M, Butler H, Jewell D, et al. Single nucleotide polymorphisms in TNFSF15 confer susceptibility to Crohn's disease. Hum Mol Genet. 2005;14:3499–506.
113. Duerr RH, Taylor KD, Brant SR, Rioux JD, Silverberg MS, Daly MJ, et al. A genome-wide association study identifies IL23R as an inflammatory bowel disease gene. Science. 2006;314:1461–3.
114. Libioulle C, Louis E, Hansoul S, Sandor C, Farnir F, Franchimont D, 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. 2007;3:e58.
115. Parkes M, Barrett JC, Prescott NJ, Tremelling M, Anderson CA, Fisher SA, et al. Sequence variants in the autophagy gene IRGM and multiple other replicating loci contribute to Crohn's disease susceptibility. Nat Genet. 2007;39:830–2.
116. Rioux JD, Xavier RJ, Taylor KD, Silverberg MS, Goyette P, Huett A, et al. Genome-wide association study identifies new susceptibility loci for Crohn disease and implicates autophagy in disease pathogenesis. Nat Genet. 2007;39:596–604.
117. Goyette P, Lefebvre C, Ng A, Brant SR, Cho JH, Duerr RH, et al. Gene-centric association mapping of chromosome 3p implicates MST1 in IBD pathogenesis. Mucosal Immunol. 2008;1:131–8. Nature Publishing Group.
118. Rohr J, Pannicke U, Döring M, Schmitt-Graeff A, Wiech E, Busch A, et al. Chronic inflammatory bowel disease as key manifestation of atypical ARTEMIS deficiency. J Clin Immunol. 2010;30:314–20.
119. Egritas O, Dalgic B. Infantile colitis as a novel presentation of familial Mediterranean fever responding to colchicine therapy. J Pediatr Gastroenterol Nutr. 2011;53:102–5.
120. Momozawa Y, Mni M, Nakamura K, Coppieters W, Almer S, Amininejad L, et al. Resequencing of positional candidates identifies low frequency IL23R coding variants protecting against inflammatory bowel disease. Nat Genet. 2011;43:43–7. Nature Publishing Group.
121. Zhou Q, Lee G-S, Brady J, Datta S, Katan M, Sheikh A, et al. A hypermorphic missense mutation in PLCG2, encoding phospholipase Cγ2, causes a

dominantly inherited autoinflammatory disease with immunodeficiency. Am J Hum Genet. 2012;91:713–20.
122. Alangari A, Alsultan A, Adly N, Massaad MJ, Kiani IS, Aljebreen A, et al. LPS-responsive beige-like anchor (LRBA) gene mutation in a family with inflammatory bowel disease and combined immunodeficiency. J Allergy Clin Immunol. 2012;130:481–8.e2.
123. Muise AM, Xu W, Guo C-H, Walters TD, Wolters VM, Fattouh R, et al. NADPH oxidase complex and IBD candidate gene studies: identification of a rare variant in NCF2 that results in reduced binding to RAC2. Gut. 2012;61:1028–35.
124. Mizukami T, Obara M, Nishikomori R, Kawai T, Tahara Y, Sameshima N, et al. Successful treatment with infliximab for inflammatory colitis in a patient with X-linked anhidrotic ectodermal dysplasia with immunodeficiency. J Clin Immunol. 2012;32:39–49.
125. Salzer E, Kansu A, Sic H, Májek P, Ikincioğullari A, Dogu FE, et al. Early-onset inflammatory bowel disease and common variable immunodeficiency-like disease caused by IL-21 deficiency. J Allergy Clin Immunol. 2014;133:1651–9.e12.
126. Catucci M, Castiello MC, Pala F, Bosticardo M, Villa A. Autoimmunity in wiskott-Aldrich syndrome: an unsolved enigma. Front Immunol. 2012;3:209.
127. Sanal O, Jing H, Ozgur T, Ayvaz D, Strauss-Albee DM, Ersoy-Evans S, et al. Additional diverse findings expand the clinical presentation of DOCK8 deficiency. J Clin Immunol. 2012;32:698–708.
128. Bégin P, Patey N, Mueller P, Rasquin A, Sirard A, Klein C, et al. Inflammatory bowel disease and T cell lymphopenia in G6PC3 deficiency. J Clin Immunol. 2013;33:520–5.
129. Avitzur Y, Guo C, Mastropaolo LA, Bahrami E, Chen H, Zhao Z, et al. Mutations in tetratricopeptide repeat domain 7A result in a severe form of very early onset inflammatory bowel disease. Gastroenterology. 2014;146:1028–39. Elsevier.
130. Zhang W, Hui KY, Gusev A, Warner N, Ng SME, Ferguson J, et al. Extended haplotype association study in Crohn's disease identifies a novel, Ashkenazi Jewish-specific missense mutation in the NF-κB pathway gene, HEATR3. Genes Immun. 2013;14:310–6.
131. Cardinale CJ, Wei Z, Panossian S, Wang F, Kim CE, Mentch FD, et al. Targeted resequencing identifies defective variants of decoy receptor 3 in pediatric-onset inflammatory bowel disease. Genes Immun. 2013;14:447–52.
132. Slowik V, Kingsmore S, Dinwiddie D, Septer S, Ciaccio C, Shao L, et al. A novel variant in the STAT3 gene associated with autoimmune enteropathy in a father –Son Duo. J Genomes Exomes. 2014;2014:1. Libertas Academica.
133. Naviglio S, Arrigo S, Martelossi S, Villanacci V, Tommasini A, Loganes C, et al. Severe inflammatory bowel disease associated with congenital alteration of transforming growth factor beta signaling. J Crohns Colitis. 2014;8:770–4.
134. Kitamura A, Sasaki Y, Abe T, Kano H, Yasutomo K. An inherited mutation in NLRC4 causes autoinflammation in human and mice. J Exp Med. 2014;211:2385–96.
135. Okou DT, Mondal K, Faubion WA, Kobrynski LJ, Denson LA, Mulle JG, et al. Exome sequencing identifies a novel FOXP3 mutation in a 2-generation family with inflammatory bowel disease. J Pediatr Gastroenterol Nutr. 2014;58:561–8.
136. Visschedijk MC, Alberts R, Mucha S, Deelen P, de Jong DJ, Pierik M, et al. Pooled resequencing of 122 ulcerative colitis genes in a large dutch cohort suggests population-specific associations of rare variants in MUC2. PLoS One. 2016;11:e0159609.
137. Prescott NJ, Lehne B, Stone K, Lee JC, Taylor K, Knight J, et al. Pooled sequencing of 531 genes in inflammatory bowel disease identifies an associated rare variant in BTNL2 and implicates other immune related genes. PLoS Genet. 2015;11:e1004955. Public Library of Science.