WJG World Journal of Gastroenterology

Submit a Manuscript: http://www.wjgnet.com/esps/
Help Desk: http://www.wjgnet.com/esps/helpdesk.aspx
DOI: 10.3748/wjg.v22.i3.949

World J Gastroenterol 2016 January 21; 22(3): 949-960
ISSN 1007-9327 (print) ISSN 2219-2840 (online)

TOPIC HIGHLIGHT

2016 Inflammatory Bowel Disease: Global view

Application of computational methods in genetic study of inflammatory bowel disease

Jin Li, Zhi Wei, Hakon Hakonarson

Jin Li, Hakon Hakonarson, Center for Applied Genomics, Abramson Research Center, Children's Hospital of Philadelphia, Philadelphia, PA 19104, United States

Zhi Wei, Department of Computer Science, New Jersey Institute of Technology, Newark, NJ 07102, United States

Hakon Hakonarson, Division of Human Genetics, Children's Hospital of Philadelphia, Philadelphia, PA 19104, United States

Hakon Hakonarson, Department of Pediatrics, The Perelman School of Medicine, University of Pennsylvania, Philadelphia, PA 19104, United States

Author contributions: All authors wrote, edited the manuscript and approved the final version.

Conflict-of-interest statement: The authors have no conflict of interest.

Correspondence to: Zhi Wei, PhD, Associate Professor, Department of Computer Science, New Jersey Institute of Technology, Newark, NJ 07102, United States. zhiwei04@gmail.com
Telephone: +1-973-6424497
Fax: +1-973-596577

Received: September 2, 2015
Peer-review started: September 2, 2015
First decision: October 14, 2015
Revised: November 4, 2015
Accepted: November 24, 2015
Article in press: November 24, 2015
Published online: January 21, 2016

Abstract

Genetic factors play an important role in the etiology of inflammatory bowel disease (IBD). The launch of genome-wide association study (GWAS) represents a landmark in the genetic study of human complex disease. Concurrently, computational methods have undergone rapid development during the past a few years, which led to the identification of numerous disease susceptibility loci. IBD is one of the successful examples of GWAS and related analyses. A total of 163 genetic loci and multiple signaling pathways have been identified to be associated with IBD. Pleiotropic effects were found for many of these loci; and risk prediction models were built based on a broad spectrum of genetic variants. Important gene-gene, gene-environment interactions and key contributions of gut microbiome are being discovered. Here we will review the different types of analyses that have been applied to IBD genetic study, discuss the computational methods for each type of analysis, and summarize the discoveries made in IBD research with the application of these methods.

Key words: Inflammatory bowel disease; Computational methods; Genome-wide association study; Pathway analysis; Gene-gene interaction; Gene-environment interaction; Pleiotropy; Risk prediction

Core tip: Computational methods have rapidly progressed during the last a few years, which rendered us the ability to analyze genotype data on a genome-wide level. The application of these methods in inflammatory bowel disease (IBD) genetic study yielded productive results. We discuss the major types of analyses in genome-wide study, and the different computational methods used in each type of analysis.

We also show how these computation methods were used in the IBD genetic study and the major findings achieved.

Li J, Wei Z, Hakonarson H. Application of computational methods in genetic study of inflammatory bowel disease. *World J Gastroenterol* 2016; 22(3): 949-960 Available from: URL: http://www.wjgnet.com/1007-9327/full/v22/i3/949.htm DOI: http://dx.doi.org/10.3748/wjg.v22.i3.949

INTRODUCTION

Inflammatory bowel disease (IBD) consists of two major subtypes: Crohn's disease (CD) and ulcerative colitis (UC) according to the diseases' clinical features. It is a complex disease, the determinants of which involve both genetic factors and environmental influences. The genetic contribution of IBD was first demonstrated by family studies and twin studies. The occurrence of the disease was observed as being aggregated in families, with a relative risk of 13-36 for CD patients' siblings and 7-17 for UC[1]. Several studies found a significantly higher proband concordance rate in monozygotic twins (38.5%-63.6%) than that in dizygotic twins (0%-6.7%)[2]. The difference is less pronounced among UC patients, 6.6%-27.9% in monozygotic twins *vs* 0%-8.0% in dizygotic twins[2]. Similar to other complex diseases, IBD is not a monogenetic disease and does not strictly follow the Mendelian inheritance pattern. Instead, it clearly demonstrates the features of polygenic disease. Linkage based genome-wide scanning studies among IBD families led to the discovery of IBD susceptibility loci IBD1-9[1,3-5], which constitute the early stage of IBD molecular genetics study.

GENOME-WIDE ASSOCIATION STUDIES

The launch of genome-wide association studies (GWAS) focusing on single nucleotide polymorphism (SNP) then resulted in the discovery of numerous susceptibility loci for complex diseases including IBD. In most IBD GWAS designs, the effect allele frequency of each SNP was compared between cases and controls. In this type of case/control study design, the underlying basic statistical test at each SNP is χ^2 test or Fisher's exact test, and logistic regression which allows for the adjustment of various covariates. IBD is one of the successful examples in utilizing GWAS for revealing of its genetic architecture. A total of 163 genetic loci have been found to be genome-wide significantly associated with IBD[6]. Certainly the rapid development of genotyping technology has made it possible to simultaneously examine millions of SNPs. The advance of computation methods also has made essential contributions to the discovery of such a tremendously large number of genetic determinants. Multiple software tools have been developed for the GWAS studies, among which, PLINK is the most popular one[7]. PLINK provides a compact, comprehensive tool-box for GWAS from basic quality control filtering, SNP association testing to advanced features including gene based analysis, annotation and epistasis test. PLINK has been used as the primary GWAS analytical tool in the vast majority of the IBD genetic studies, from the early finding of loci *IL10*, *ARPC2*[8], *HNF4A*, *CDH1*, *CDH3*, *LAMB1*[9] to the more recent identification of 163 IBD loci[6]. In GWAS, sample structure, which includes both population stratification and hidden relatedness[10], is a common issue resulting in inflation of statistics and false positive results. Sample structure is usually estimated using identical-by-descent analysis implemented in PLINK and principal component analysis implemented in EIGENSTRAT[11,12]. During the recent years, linear mixed model (LMM), which captures both fixed and random effects, has been applied to GWAS. Software based on LMM includes EMMAX[10], FaST[13], and GEMMA[14]. They simultaneously correct for population stratification and hidden relatedness. GEMMA can also be used to analyze multiple correlated phenotypes.

Advance in methodology development of imputation and meta-analysis renders GWAS more power for the identification of genome-wide significant genetic loci. Imputation is to infer genotypes of un-genotyped SNPs based on the SNPs that have been assayed for a group of samples and the haplotype information from a reference population set. The aims of using imputation are to fine-map the associated regions, to boost study power by examining more SNPs, and to meta-analyze cohorts which are genotyped on different platforms[15]. The common practice of applying imputation in GWAS mainly involves three steps: haplotype estimation (pre-phasing), genotype imputation and association testing. The first step pre-phasing is to estimate underlying haplotypes from SNP genotype data. Among several existing methods, software packages fastPHASE[16], IMPUTE2[17], MACH[18], BEAGLE[19] and SHAPEIT[20] all employ coalescent-based methods and hidden Markov models[21]. MACH and IMPUTE2 utilize the hidden states as "template haplotypes"[17,18,21]; fastPHASE employs parsimonious haplotype-clustering for a certain number of clusters[16], and BEAGLE adopts a localized haplotype clustering[19]. SHAPEIT further improves by collapsing haplotypes into a graph structure[20]. Several of these software packages also contain the application of imputation, such as IMPUTE V2, MACH, fastPHASE and BEAGLE[21]. They are the most commonly used software for imputation. For example, IMPUTE V2 carries out haploid imputation for the SNPs that are not genotyped but in the reference panel by conditioning on the haplotypes estimated from the genotyped SNPs[17]. BEAGLE iteratively fits the localized haplotype clustering model to the estimated haplotypes and resamples haplotypes conditional on the genotype and diploid HMM model to derive the probability of the un-genotyped SNP[19]. The review by Marchini J. and Howie B. provides a good summary of the pros and cons of each method[15]. For step 3, association testing, one approach is to convert the genotype probability to the bi-allelic format by specifying a probability threshold. Another more commonly used approach is to take imputation uncertainty into account by frequentist tests such as the score test in SNPTEST[22]. A novel approach of Bayesian method to derive Bayes factor has also been incorporated into SNPTEST and BIMBAM[23]. Software has been developed to work seamlessly for the above three steps, for instance, using SHAPEIT for pre-phasing, and then IMPUTE2 for imputation, which was followed by SNPTEST for association testing. Each of the imputation method discussed above has its own advantage. In the two large cohort studies of IBD, BEAGLE was used as the imputation tool, but other methods have also been widely used in the GWAS study of IBD[6,24].

As another popular approach in GWAS to increase statistical power, meta-analysis combines results from individual studies. Identifications of many IBD genome-wide significant loci have benefited from meta-analysis[24-26], with around 30 susceptibility loci being discovered from each of these large scale meta-analyses. In all these three studies, Z-score based inverse variance weighted meta-analysis was carried out, with the assumption of fixed effects. Other popular approaches include traditional *P* value based method with sample size as weight[27], random effect method and Bayesian approach[28]. Popular software packages for GWAS meta-analysis include METAL[27], META[22,29], MetABEL[30], GWAMA[31]; PLINK also has its own meta-analysis option. All these software packages have a fixed effects model implemented, but some also contain random-effect models[32]. When conducting meta-analysis, practitioners should consider the direction of effects and heterogeneity between studies.

PATHWAY ANALYSIS

Though a large number of susceptibility loci have been identified by SNP based GWAS, the disease variance explained by these loci was still limited: 13.6% for CD and 7.5% for ulcerative colitis[6]. At least part of the missing heritability may reside in those genetic loci of moderate effect, which did not reach genome-wide significance due to the limited power in most genetic studies. If genes in these loci function collaboratively, they could make significant contributions to the disease etiology. Pathway studies, which jointly assess the statistical significance of a group of genes, can be applied to identify such loci. From a biological point of view, gene products work cooperatively to carry out certain molecular and cellular functions; and disease etiology is more likely the result of pathway dysregulation. From the statistical point of view, examining the joint effect of a group of genes will allow the identification of those in which each gene member make only a modest contribution. By collapsing SNP statistics to gene level and further pathway level, the number of multiple testing will be reduced. Though challenges and arguments persist in the field of pathway analysis, advance in methodology development has been rapid. The major steps involved in pathway/network analysis are mapping SNPs to genes, aggregating SNPs statistics to gene level, and then assessing the enrichment of pathways or projection of genes into protein-protein interaction (PPI) network. Different methods have been evolved handling the tasks at each step differently. Over 50 pathway analysis software packages have been developed[33]. CD has been used as a common example for the development and application of GWAS-based pathway software. For example, Wang *et al*[34] applied their software GenGen to the study of the Wellcome Trust Case Control Consortium CD dataset and uncovered the significant associations of pathway IL12/IL23 with CD status in several cohorts genotyped on different SNP chips and of different ethnicities[35]. This is one of the earliest GWAS-based pathway analyses of IBD. Another study by Torkamani *et al*[36], revealed another interleukin association with CD - IL3 activation and signaling pathway. The software that they used, MetaCore, is a commercial one which is developed by GeneGo. Significant enrichment of "JAK-STAT signaling pathway", "Cytokine-cytokine receptor interaction" was found by Peng *et al*[37] using statistical method Simes/FDR. Similar findings of several cytokine signaling pathways were made by Carbonetto and Stephens using a model based approach[38]. In addition to IL/cytokine signaling, the involvement of MHC genes was found by Holmans *et al*[39] using software ALIGATOR. Similarly, when searching for enrichment of canonical pathways or Gene Ontology (GO) terms, Jostins *et al*[6] found the highly significant enrichment of GO terms regulation of cytokine production, lymphocyte activation, and pathway JAK-STAT signaling pathway in the largest IBD GWAS so far. Though being common in that these pathways were all interleukin/immune related pathways, the non-overlapping of top significant pathways between different methods could be attributed to different statistical methods and/or the different pathway databases used in their studies. Several non-immune pathways were also found to be significantly associated, such as "calcium signaling", "ChREBP regulation pathway" identified by Torkamani *et al*[36]; "ABC transporters" and "Extracellular matrix-receptor interaction" found by Peng *et al*[37] by Fisher's exact test. It would be interesting to further experimentally test the biological contributions of genes in these pathways to IBD. A lot of convenient online software tools have been developed for pathway analysis based on GWAS data, such as i-GSEA4GWAS[40] and GSEA-SNP[41]. In addition to the methods discussed above, other pathway analysis tools that were originally designed for gene expression data have also been used after concatenating SNP statistics to gene statistics, such as DAVID[42], GeneTrail[43], and WebGestalt[44].

As an alternative to pathway analysis, another way to assess genes' joint effect is PPI network, in which the nodes represent the protein products of the genes, and the edges connecting the nodes represent the biochemical interactions between a pair of proteins. After mapping significant genes to PPI networks in various databases, the resulting nodes, edges and the size of the largest connected component was then computed. Cytoscape[45] is the most commonly used comprehensive toolkits for network analysis and visualization. Other online tools include DAPPLE[46], STRINGS[47], and FUNCOUP[48]. Pathway and network analyses can also be combined, which yield significantly enriched pathways within the network. Such pathway analyses are implemented in software like STRINGS and FUNCOUP; and can be integrated into Cytoscape. Network analysis tools, such as DAPPLE, have been used for prioritization of candidate genes in IBD analysis[6].

GENE-GENE INTERACTION

In addition to pathway analysis in general, gene-gene interaction is also an important topic in IBD research. Gene-gene interaction is often referred to as "epistasis", for which there are its mathematical definition and its biological interpretation. Simplistically, it implicates a departure from independent effect of each gene per se, but we need to be aware that epistasis in mathematical meaning and biological context are not often interchangeable. Mathematically, epistasis testing for complex human diseases including IBDs, was carried out in two ways - hypothesis driven and hypothesis free. In the former, particular groups of genetic variants were selected and tested; and in the later especially in the context of GWASs, pairs of interactions between all SNPs studied were exhaustively searched[49]. Most traditional studies in the IBD research field, which were limited by sample size and resources, focused on epistasis testing between specific candidate genes, using χ^2 test or logistic regression. Because of the huge number of combinations n(n-1)/2 needed to be tested for n SNPs, one can image the tremendous computational burden created to carry out the hypothesis free search. In recent years, the methodology developments have made it possible in human genetics study, which has been summarized in detail in review article by Wei *et al*[49]. They classified the common methods into several classes, such as regression-based methods, LD- and haplotype-based methods, Bayesian methods, machine-learning and data-mining methods, and data filtering methods[49]. Among these methods, the most commonly used one to detect epistasis interactions in either hypothesis free or hypothesis driven setting is the regression-based methods by comparing the statistical model with interaction terms (saturated model) and the one without (reduced model)[49]. Logistic regression has been commonly used in epistasis testing among candidate genes, such as interactions between *CARD8* and *NALP3*[50]. This is relatively easy for testing interactions between a small number of SNPs of interest. For genome-wide epistasis testing, with the rapid development of modern computational technology, this is not an impossible task anymore. Epistasis test implemented in PLINK is a typical example of using regression based method on SNP alleles[7]. Proper application of approximation methods further speeds up the computing time for initial screening purpose. For example, software package "BOOST" employed the Kirkwood superposition approximation[51] at the initial genome-wide screening stage for epistasis interactions involved in disease traits[52], followed by the classical likelihood ratio test only at the testing stage after eliminating the many non-significant interactions first. In addition to the two-step design and utilization of approximation approach, BOOST also transformed genotype data to Boolean representation which allows for fast logic computing[52]. Now together with joint effect tests[53], BOOST algorithm has been incorporated into PLINK 1.9 (http://pngu.mgh.harvard.edu/~purcell/plink2/epistasis.html). Other algorithms have also been introduced into the field of epistasis testing, such as decision tree learning which is a predictive modeling method of machine-learning and data mining[54]; as well as multifactor dimensionality reduction (MDR)[55,56]. These methods also allow for the identification of higher order interactions. Such methods have been used in epistasis analysis for IBD. Most recent publications in this regard include the one by Wang *et al*[57] using regression tree to search for models that contain logical SNP interactions that better explain UC disease compared to single SNP combinations. However, their study was limited to the known genome-wide significant 133 UC loci. MDR is a non-parametric method that select and combine high-dimensional genotypic data into a one-dimensional model with a variable of high and low risk classes[55]. By MDR analysis, Okazaki *et al*[58] found suggestive evidence for high-order interactions between genes *IL23R*, *IBD5* and *ATG16L1* variants in a CD case-control cohort.

Here the epistasis testing that we have discussed refers to its mathematic meaning. The epistasis interactions captured in statistical model often are hard to investigate for its biological mechanism in experimental settings.

GENE-ENVIRONMENT INTERACTION

Gene-environment interaction is another hot topic in IBD research. A broad range of environmental factors have been studied in IBD genetics, including environment factors in traditional definition such as smoking, diet, physical activity, medication, infection and other life style factors[59,60]; and the newly defined "in-vironment" factors, like the microbiome in the gut. In fact, the gut "in-vironment" is affected by environment factors like diet[61]. It has been well recognized that the development of IBD is the result of interplaying between individual's genetic composition and his/her environmental exposure. To study the gene-environment interactions, the χ^2 test and logistic regression methods can be similarly used, just by introducing environmental factors into the regression. Important findings include the interactions between smoking and polymorphisms of important IBD risk genes, *NOD2*, *CYP2A6* and *GSTP1*[62,63]. New methods have been adopted in this regard too, such as the logic regression[64,65], which fits the regression model with Boolean expression of the predictor variables, similarly as we discussed above for gene-gene interactions. In the work by Wang *et al*[57], they not only found gene-gene interactions but also identified interactions between genes and smoking in UC cohorts by logic regression. The same group has also applied this method to a similar study of genetic and environmental factors for CD[66]. By far, smoking is the most studied environmental factors for IBD.

It has long been suggested that the interaction between the intestinal immune system and its microbe environment plays key roles in the pathogenesis of IBD. The highly significant enrichment of pathway "response to molecules of bacterial origin" among the significant IBD GWAS loci reinforced such concept in the largest IBD GWAS study to date[6]. Dysbiosis, which refers to the reduction of gut microbiota diversity or abnormal change of composition, has been known to play a major pathological role in IBD etiology[67]. The Dysbiosis of microbiota in IBD patients was investigated using different experimental methods. Two approaches are commonly used in the microbiome investigation. A traditional approach is the targeted amplicon approach in which the taxonomy and the phylogeny of the microbiota were examined by using several marker genes[68]. One most commonly used marker is ribosomal small subunit --16S ribosomal DNA (rDNA) because of its universal existence across species and its composition of both slowly evolving sequences and fast evolving sequences[68]. A more recent approach is the metagenomics approach which is high-throughput sequencing based microbiota genomic DNA profiling[68]. Many computational tools have been developed for analyzing sequencing data from either approach. The major analytical tools for 16S rDNA data include QIIME (quantitative insights into microbial ecology)[69] established using PyCogent toolkit[70], mothur, which incorporates algorithms in several other tools and new features[71], and VAMPS[72]. For metagenomics data analysis, common software packages, including PhymmBL[73] and MEGAN[74], compare high-throughput sequencing reads to known microbiome genome sequences; while others, such as PhyIOTU[75] and MLtreemap[76], derive phylogenetic origin based on marker genes from the large amounts of sequencing data. Both approaches have been used in microbiome study of IBD pathogenesis. Gut microbiota are shaped by various factors including both host genetic factors, and environment factors, like diet, medication and others. Information from research in this field is emerging. A reduction in diversity of fecal microbiome has been reported for CD patients when compared to healthy controls, which is the result of studying 16S rRNA genes[77], and similar results were found for UC patients[78]. Another microarray study using 16S rDNA from CD patients and healthy controls specifically observed a reduction in bacterial populations within the phylum Firmicutes[79]. Reduced diversity has also been reported when comparing fully inflamed tissue to non-flamed tissue for the same CD or UC patient[80]. On the other hand, increased fungal diversity has been found in CD patients compared with controls[81]. A recent large pediatric treatment-naïve CD cohort examining microbiome indicates increase and reduction of abundance in different bacteria species; and also suggests the influence of antibiotic on such changes[82]. Several studies implicate that host genetics, like *NOD2*, *ATG16L1* variants have an impact on abundance of mucosal microbiota abundance[83,84]. A study focusing on bacterial and metaproteomic analysis of the mucosal-luminal interface has emphasized the host-microbe interactions as one of the important underlying factors for IBD[85].

PLEIOTROPY

IBD is a type of auto-immune disease (AID) and genetic studies have found it shares susceptibility loci, signaling pathways with other immune diseases and non-immune diseases, despite their distinct clinical features. Lees *et al*[86] provided a good summary of susceptibility loci common to IBD and other diseases. Pleiotropy is a common phenomenon among auto-immune diseases. Signaling pathways like cytokine signaling and innate immunity have been linked to several AIDs[87-89], including IBD[90,91]. Results from more recent studies of cohorts genotyped on the immunochip provide further evidence for the genetic sharing among different types of immune-related diseases. Traditionally, when a cohort comprising patients of multiple related diseases were studied, pooled analysis or meta-analysis is employed, including both fixed effects meta-analysis and random effects meta-analysis. However, inherent heterogeneity in these cohorts makes it loose power to find subtypes specific signals or signals that are of opposite directions of effect in related diseases. To compensate for these disadvantages, several methods have been developed. Cotsapas *et al*[92] developed the cross-phenotype meta-analysis method which tests for the

Table 1 Commonly used publically available software for each step in genomic studies

Analysis	Software	URL	Ref.
GWAS	PLINK	http://pngu.mgh.harvard.edu/~purcell/plink/	[7]
	EMMAX	http://genetics.cs.ucla.edu/emmax/	[10]
	FaST	http://research.microsoft.com/en-us/um/redmond/projects/MSCompBio/Fastlmm/	[13]
	GEMMA	http://www.xzlab.org/software.html	[14]
Imputation	SHAPEIT (pre-phasing)	https://mathgen.stats.ox.ac.uk/genetics_software/shapeit/shapeit.html	[20]
	IMPUTE2 (pre-phasing and imputation)	https://mathgen.stats.ox.ac.uk/impute/impute_v2.html	[17]
	MACH (pre-phasing and imputation)	http://csg.sph.umich.edu//abecasis/MACH/tour/imputation.html	[18]
	fastPHASE (pre-phasing and imputation)	https://els.comotion.uw.edu/express_license_technologies/fastphase	[16]
	BEAGLE (pre-phasing and imputation)	http://faculty.washington.edu/browning/beagle/beagle.html	[19]
	SNPTEST (association testing)	https://mathgen.stats.ox.ac.uk/genetics_software/snptest/old/snptest.html	[22]
Meta-analysis	METAL	http://csg.sph.umich.edu//abecasis/Metal/	[27]
	META	https://mathgen.stats.ox.ac.uk/genetics_software/meta/meta.html	[22]
	MetABEL	http://www.genabel.org/packages/MetABEL	[30]
	GWAMA	http://www.well.ox.ac.uk/gwama/download.shtml	[31]
	PLINK	http://pngu.mgh.harvard.edu/~purcell/plink/metaanal.shtml	[7]
Pathway analysis	GenGen	http://gengen.openbioinformatics.org/en/latest/tutorial/pathway/	[34]
	ALIGATOR	http://x004.psycm.uwcm.ac.uk/~peter/	[39]
	i-GSEA4GWAS	http://gsea4gwas.psych.ac.cn/	[40]
	GSEA-SNP	https://www.nr.no/en/projects/software-genomics	[41]
	DAVID	http://david.ncifcrf.gov/summary.jsp	[42]
	GeneTrail	http://genetrail.bioinf.uni-sb.de/	[43]
	WebGestalt	http://bioinfo.vanderbilt.edu/webgestalt/	[44]
Network analysis	Cytoscape	http://www.cytoscape.org/	[45]
	DAPPLE	http://www.broadinstitute.org/mpg/dapple/dappleTMP.php	[46]
	STRINGS	http://string-db.org/	[47]
	FUNCOUP	http://funcoup.sbc.su.se/search/	[48]
Gene-gene interaction	PLINK	http://pngu.mgh.harvard.edu/~purcell/plink/epi.shtml	[7]
	BOOST	http://bioinformatics.ust.hk/BOOST.html	[52]
	MDR	https://ritchielab.psu.edu/mdr-downloads	[55]
Microbiome	QIIME	http://qiime.org/	[69]
	mothur	http://www.mothur.org/wiki/Main_Page	[71]
	VAMPS	https://vamps.mbl.edu/portals/hmp/hmp.php	[72]
	PhymmBL	http://ccb.jhu.edu/software/phymmbl/index.shtml	[73]
	MEGAN5	http://ab.inf.uni-tuebingen.de/software/megan5/	[74]
	PhyIOTU	https://github.com/sharpton/PhyIOTU	[75]
	MLTreeMap	http://mltreemap.org/	[76]
Pleiotropy	CPMA	http://coruscant.itmat.upenn.edu/software.html	[92]
	ASSET	http://www.bioconductor.org/packages/release/bioc/html/ASSET.html	[93]
	GPA	https://github.com/dongjunchung/GPA	[98]
	SMAT	http://www.hsph.harvard.edu/xlin/software.html	[101]

deviation from the null hypothesis that there is no additional association beyond the known ones; and the alternative hypothesis involves associations with two or more phenotypes. They applied this method for the test of pleiotropic effects of 107 SNPs which are known for immune-disease association among 7 diseases including CD; and found evidence of association with more than one disease for 47 SNPs. Bhattacharjee *et al*[93] presented a subset-based meta-analysis, which exhaustively investigates combinations of diseases subsets against the shared control set or the rest of all the samples; logistic regression was iteratively applied for each combination to find the best subset combination that would yield strongest association for each SNP; and the discrete local maxima method[94] was used to get the estimation of *P* value upper boundary. Other methods made extensions to the weighted sum of the univariate test statistics[95], reflecting effect heterogeneity among phenotypes[96,97]. Newly proposed methods like GPA (Genetic analysis incorporating Pleiotropy and Annotation)[98] employed a two-groups model[99] and an Expectation-Maximization (EM) algorithm[100] for estimation of parameters in the GPA model. They also incorporated annotation enrichment into the analysis to prioritize identification of functional signals. Schifano *et al*[101] developed a novel approach SMAT for evaluating shared effect of genetic variants on several secondary continuous phenotypes in a case-control cohort by ways of a scaled marginal model.

RISK PREDICTION

Genetics study is aimed to identify high-risk populations for early prevention and future targeted treatment. Comparing to numerous GWAS on IBD, the number of risk prediction studies using genetic markers is limited. Recently, Wei *et al*[102] used an advanced machine-learning technique and the large IBD Genetic Consortium's cohort of more than 60000

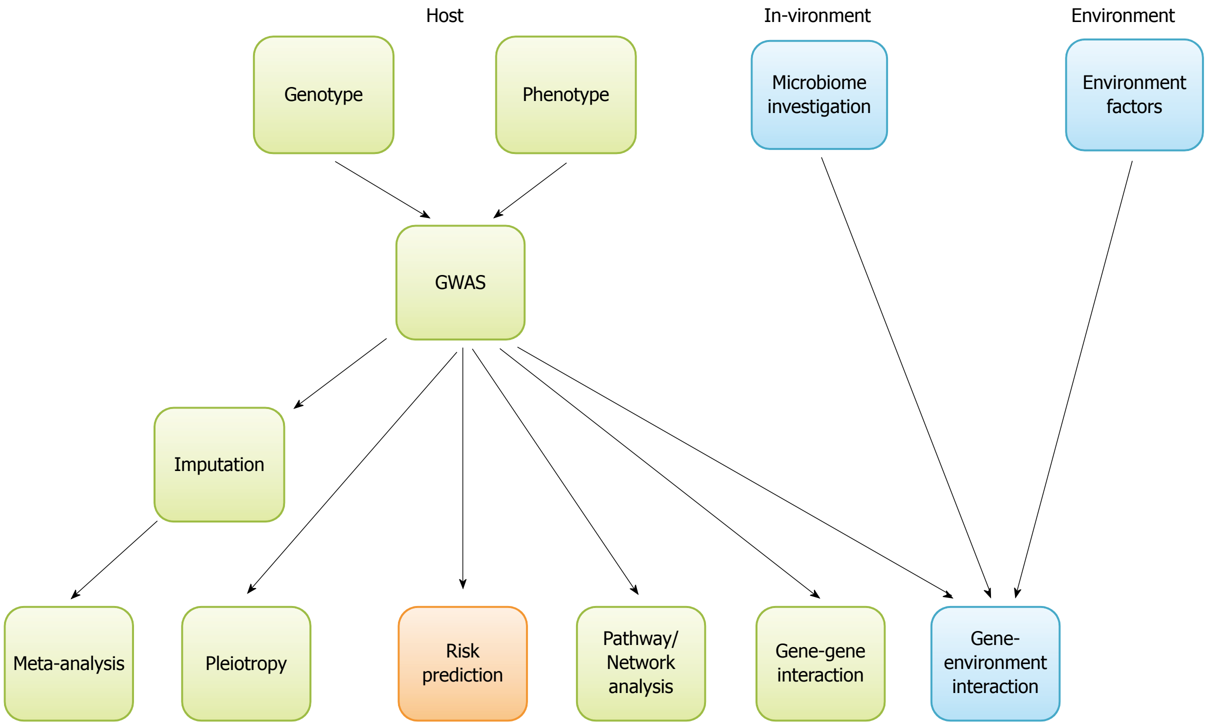

Figure 1 Different analyses for genomic studies of inflammatory bowel disease and other complex human diseases. Multiple different analyses can be carried out based on the disease phenotype, genotype and relationship with other factors. GWAS: Genome-wide association study.

samples. They developed a step-wise risk assessment model. The model achieved the highest areas under the curve (AUCs) so far for CD and UC, which are 0.86 and 0.83 respectively. Compared to previous studies, this is a large improvement. Li *et al*[103] proposed a risk prediction method integrating pleiotropic effects between diseases by using a bivariate ridge regression method. They demonstrated improved prediction accuracy when CD and UC datasets were jointly analyzed.

We listed the commonly used software package for each type of analysis in Table 1 and Figure 1 and summarized the major points of our discussion above into the Table 2 as take-home messages.

CONCLUSION

Genome-wide genotyping results provide a huge reservoir of genetic data to explore. From single variant to gene and to pathway, analyses can be carried out at multiple levels. How host factors interact with environment is another important aspect to consider, especially for IBD. The ultimate goal of genetic studies is to translate discoveries to clinical applications. Thus it is important to develop good methods for risk prediction so as to identify patients at high risk for early prevention. Other applications should also be considered to maximize the value of discoveries from genetic studies.

Through the rapid development of computation methods and other technologies, tremendous advance has been made in the field IBD genetic studies; yet much remains to be explored and investigated.

As a successful example of GWAS, 163 susceptibility loci have been discovered for IBD[6], however, much remains unknown how these loci contribute to IBD etiology, especially the majority of these loci were not in the coding region of any gene. Many questions in this regard need to be addressed, for example, whether these SNPs directly affect gene expression level or gene function, whether they tag for any other SNP. More resources are becoming available for functional annotation of genetic variants, even those in the gene desert regions, such as the ENCODE[104], Epigenome Roadmap[105] for chromatin markers, and GTEx[106] for eQTL. It is expected that integrated analysis of genetic, epigenetic and eQTL data will provide more insight into the understanding of these genetic variants, and will help to prioritize the genetic variants for experimental studies.

In addition to common variants, rare variants can also contribute to IBD susceptibility. The current methods for rare variants identification include exome array, whole exome sequencing and whole genome sequencing. Compared to the well-established pipeline for GWAS and commonly accepted threshold standards to determine GWAS significant loci, much remains to be explored in the field of rare variants study. Rare variants were thought to confer larger effects than common variants, based on experience from Mendelian disease and early findings of complex disease such as variants I1307K and E1317Q in gene *APC* associated with colorectal tumors[107,108], and rare variants in gene *PCSK9* associated with coronary heart disease[109].

Table 2 Take home messages from our discussion on the application of computational methods in genetic study of inflammatory bowel disease

Take home messages
GWAS is an unbiased method to identify common genetic variants that are significantly associated with complex human diseases. Sample structure needs to be carefully handled to avoid false positive results
Imputation is often employed to infer un-genotyped SNPs based on those genotyped ones, followed by meta-analysis to combine results from multiple studies, in order to increase the power in GWAS
Pathway analyses help to identify genetic variants that have modest individual effect but jointly make significant contribution to disease susceptibility
Both gene-gene interactions and gene-environment interactions are important underlying factors for IBD
Pleiotropy studies aim to identify genetic loci shared by IBD and other immune diseases
Risk prediction is one of the ways to translate GWAS discoveries to clinics - to identify patients at high risk

GWAS: Genome-wide association study; IBD: Inflammatory bowel disease.

More studies revealed that the genetic risks attributed to many rare variants are also modest which makes large sample size a necessary factor for such studies, and the effects of rare variants may need to be aggregated for analysis at the gene level to conquer the low power in rare variant analysis[110]. Another feature of rare variants is population specificity, which makes population stratification an outstanding issue in rare variant studies[110]. Because of these features, assumptions in analytical methods for common variants may not hold true for rare variants. Cautions need to be taken when utilizing existing methods that were used for common variants; and new statistical and computational methods development need to meet such demands from the studies of rare variants.

With the combined study for both common variants and rare variants, and integrated functional annotation, more susceptibility and causality loci will be identified for IBD, which will pave the way for the development of suitable treatments for each patient.

REFERENCES

1 **Ahmad T**, Satsangi J, McGovern D, Bunce M, Jewell DP. Review article: the genetics of inflammatory bowel disease. *Aliment Pharmacol Ther* 2001; **15**: 731-748 [PMID: 11380312]

2 **Ek WE**, D'Amato M, Halfvarson J. The history of genetics in inflammatory bowel disease. *Ann Gastroenterol* 2014; **27**: 294-303 [PMID: 25331623]

3 **Hugot JP**, Laurent-Puig P, Gower-Rousseau C, Olson JM, Lee JC, Beaugerie L, Naom I, Dupas JL, Van Gossum A, Orholm M, Bonaiti-Pellie C, Weissenbach J, Mathew CG, Lennard-Jones JE, Cortot A, Colombel JF, Thomas G. Mapping of a susceptibility locus for Crohn's disease on chromosome 16. *Nature* 1996; **379**: 821-823 [PMID: 8587604 DOI: 10.1038/379821a0]

4 **Satsangi J**, Parkes M, Louis E, Hashimoto L, Kato N, Welsh K, Terwilliger JD, Lathrop GM, Bell JI, Jewell DP. Two stage genome-wide search in inflammatory bowel disease provides evidence for susceptibility loci on chromosomes 3, 7 and 12. *Nat Genet* 1996; **14**: 199-202 [PMID: 8841195 DOI: 10.1038/ng1096-199]

5 **Cooney R**, Jewell D. The genetic basis of inflammatory bowel disease. *Dig Dis* 2009; **27**: 428-442 [PMID: 19897957 DOI: 10.1159/000234909]

6 **Jostins L**, Ripke S, Weersma RK, Duerr RH, McGovern DP, Hui KY, Lee JC, Schumm LP, Sharma Y, Anderson CA, Essers J, Mitrovic M, Ning K, Cleynen I, Theatre E, Spain SL, Raychaudhuri S, Goyette P, Wei Z, Abraham C, Achkar JP, Ahmad T, Amininejad L, Ananthakrishnan AN, Andersen V, Andrews JM, Baidoo L, Balschun T, Bampton PA, Bitton A, Boucher G, Brand S, Büning C, Cohain A, Cichon S, D'Amato M, De Jong D, Devaney KL, Dubinsky M, Edwards C, Ellinghaus D, Ferguson LR, Franchimont D, Fransen K, Gearry R, Georges M, Gieger C, Glas J, Haritunians T, Hart A, Hawkey C, Hedl M, Hu X, Karlsen TH, Kupcinskas L, Kugathasan S, Latiano A, Laukens D, Lawrance IC, Lees CW, Louis E, Mahy G, Mansfield J, Morgan AR, Mowat C, Newman W, Palmieri O, Ponsioen CY, Potocnik U, Prescott NJ, Regueiro M, Rotter JI, Russell RK, Sanderson JD, Sans M, Satsangi J, Schreiber S, Simms LA, Sventoraityte J, Targan SR, Taylor KD, Tremelling M, Verspaget HW, De Vos M, Wijmenga C, Wilson DC, Winkelmann J, Xavier RJ, Zeissig S, Zhang B, Zhang CK, Zhao H, Silverberg MS, Annese V, Hakonarson H, Brant SR, Radford-Smith G, Mathew CG, Rioux JD, Schadt EE, Daly MJ, Franke A, Parkes M, Vermeire S, Barrett JC, Cho JH. Host-microbe interactions have shaped the genetic architecture of inflammatory bowel disease. *Nature* 2012; **491**: 119-124 [PMID: 23128233 DOI: 10.1038/nature11582]

7 **Purcell S**, Neale B, Todd-Brown K, Thomas L, Ferreira MA, Bender D, Maller J, Sklar P, de Bakker PI, Daly MJ, Sham PC. PLINK: a tool set for whole-genome association and population-based linkage analyses. *Am J Hum Genet* 2007; **81**: 559-575 [PMID: 17701901 DOI: 10.1086/519795]

8 **Franke A**, Balschun T, Karlsen TH, Sventoraityte J, Nikolaus S, Mayr G, Domingues FS, Albrecht M, Nothnagel M, Ellinghaus D, Sina C, Onnie CM, Weersma RK, Stokkers PC, Wijmenga C, Gazouli M, Strachan D, McArdle WL, Vermeire S, Rutgeerts P, Rosenstiel P, Krawczak M, Vatn MH, Mathew CG, Schreiber S. Sequence variants in IL10, ARPC2 and multiple other loci contribute to ulcerative colitis susceptibility. *Nat Genet* 2008; **40**: 1319-1323 [PMID: 18836448 DOI: 10.1038/ng.221]

9 **Barrett JC**, Lee JC, Lees CW, Prescott NJ, Anderson CA, Phillips A, Wesley E, Parnell K, Zhang H, Drummond H, Nimmo ER, Massey D, Blaszczyk K, Elliott T, Cotterill L, Dallal H, Lobo AJ, Mowat C, Sanderson JD, Jewell DP, Newman WG, Edwards C, Ahmad T, Mansfield JC, Satsangi J, Parkes M, Mathew CG, Donnelly P, Peltonen L, Blackwell JM, Bramon E, Brown MA, Casas JP, Corvin A, Craddock N, Deloukas P, Duncanson A, Jankowski J, Markus HS, Mathew CG, McCarthy MI, Palmer CN, Plomin R, Rautanen A, Sawcer SJ, Samani N, Trembath RC, Viswanathan AC, Wood N, Spencer CC, Barrett JC, Bellenguez C, Davison D, Freeman C, Strange A, Donnelly P, Langford C, Hunt SE, Edkins S, Gwilliam R, Blackburn H, Bumpstead SJ, Dronov S, Gillman M, Gray E, Hammond N, Jayakumar A, McCann OT, Liddle J, Perez ML, Potter SC, Ravindrarajah R, Ricketts M, Waller M, Weston P, Widaa S, Whittaker P, Deloukas P, Peltonen L, Mathew CG, Blackwell JM, Brown MA, Corvin A, McCarthy MI, Spencer CC, Attwood AP, Stephens J, Sambrook J, Ouwehand WH, McArdle WL, Ring SM, Strachan DP. Genome-wide association study of ulcerative colitis identifies three new susceptibility loci, including the HNF4A region. *Nat Genet* 2009; **41**: 1330-1334 [PMID: 19915572 DOI: 10.1038/ng.483]

10 **Kang HM**, Sul JH, Service SK, Zaitlen NA, Kong SY, Freimer NB, Sabatti C, Eskin E. Variance component model to account for

sample structure in genome-wide association studies. *Nat Genet* 2010; **42**: 348-354 [PMID: 20208533 DOI: 10.1038/ng.548]
11 **Price AL**, Patterson NJ, Plenge RM, Weinblatt ME, Shadick NA, Reich D. Principal components analysis corrects for stratification in genome-wide association studies. *Nat Genet* 2006; **38**: 904-909 [PMID: 16862161 DOI: 10.1038/ng1847]
12 **Patterson N**, Price AL, Reich D. Population structure and eigenanalysis. *PLoS Genet* 2006; **2**: e190 [PMID: 17194218 DOI: 10.1371/journal.pgen.0020190]
13 **Lippert C**, Listgarten J, Liu Y, Kadie CM, Davidson RI, Heckerman D. FaST linear mixed models for genome-wide association studies. *Nat Methods* 2011; **8**: 833-835 [PMID: 21892150 DOI: 10.1038/nmeth.1681]
14 **Zhou X**, Stephens M. Efficient multivariate linear mixed model algorithms for genome-wide association studies. *Nat Methods* 2014; **11**: 407-409 [PMID: 24531419 DOI: 10.1038/nmeth.2848]
15 **Marchini J**, Howie B. Genotype imputation for genome-wide association studies. *Nat Rev Genet* 2010; **11**: 499-511 [PMID: 20517342 DOI: 10.1038/nrg2796]
16 **Scheet P**, Stephens M. A fast and flexible statistical model for large-scale population genotype data: applications to inferring missing genotypes and haplotypic phase. *Am J Hum Genet* 2006; **78**: 629-644 [PMID: 16532393 DOI: 10.1086/502802]
17 **Howie BN**, Donnelly P, Marchini J. A flexible and accurate genotype imputation method for the next generation of genome-wide association studies. *PLoS Genet* 2009; **5**: e1000529 [PMID: 19543373 DOI: 10.1371/journal.pgen.1000529]
18 **Li Y**, Willer CJ, Ding J, Scheet P, Abecasis GR. MaCH: using sequence and genotype data to estimate haplotypes and unobserved genotypes. *Genet Epidemiol* 2010; **34**: 816-834 [PMID: 21058334 DOI: 10.1002/gepi.20533]
19 **Browning SR**, Browning BL. Rapid and accurate haplotype phasing and missing-data inference for whole-genome association studies by use of localized haplotype clustering. *Am J Hum Genet* 2007; **81**: 1084-1097 [PMID: 17924348 DOI: 10.1086/521987]
20 **Delaneau O**, Marchini J, Zagury JF. A linear complexity phasing method for thousands of genomes. *Nat Methods* 2012; **9**: 179-181 [PMID: 22138821 DOI: 10.1038/nmeth.1785]
21 **Browning SR**, Browning BL. Haplotype phasing: existing methods and new developments. *Nat Rev Genet* 2011; **12**: 703-714 [PMID: 21921926 DOI: 10.1038/nrg3054]
22 **Marchini J**, Howie B, Myers S, McVean G, Donnelly P. A new multipoint method for genome-wide association studies by imputation of genotypes. *Nat Genet* 2007; **39**: 906-913 [PMID: 17572673 DOI: 10.1038/ng2088]
23 **Guan Y**, Stephens M. Practical issues in imputation-based association mapping. *PLoS Genet* 2008; **4**: e1000279 [PMID: 19057666 DOI: 10.1371/journal.pgen.1000279]
24 **Franke A**, McGovern DP, Barrett JC, Wang K, Radford-Smith GL, Ahmad T, Lees CW, Balschun T, Lee J, Roberts R, Anderson CA, Bis JC, Bumpstead S, Ellinghaus D, Festen EM, Georges M, Green T, Haritunians T, Jostins L, Latiano A, Mathew CG, Montgomery GW, Prescott NJ, Raychaudhuri S, Rotter JI, Schumm P, Sharma Y, Simms LA, Taylor KD, Whiteman D, Wijmenga C, Baldassano RN, Barclay M, Bayless TM, Brand S, Büning C, Cohen A, Colombel JF, Cottone M, Stronati L, Denson T, De Vos M, D'Inca R, Dubinsky M, Edwards C, Florin T, Franchimont D, Gearry R, Glas J, Van Gossum A, Guthery SL, Halfvarson J, Verspaget HW, Hugot JP, Karban A, Laukens D, Lawrance I, Lemann M, Levine A, Libioulle C, Louis E, Mowat C, Newman W, Panés J, Phillips A, Proctor DD, Regueiro M, Russell R, Rutgeerts P, Sanderson J, Sans M, Seibold F, Steinhart AH, Stokkers PC, Torkvist L, Kullak-Ublick G, Wilson D, Walters T, Targan SR, Brant SR, Rioux JD, D'Amato M, Weersma RK, Kugathasan S, Griffiths AM, Mansfield JC, Vermeire S, Duerr RH, Silverberg MS, Satsangi J, Schreiber S, Cho JH, Annese V, Hakonarson H, Daly MJ, Parkes M. Genome-wide meta-analysis increases to 71 the number of confirmed Crohn's disease susceptibility loci. *Nat Genet* 2010; **42**: 1118-1125 [PMID: 21102463 DOI: 10.1038/ng.717]
25 **Anderson CA**, Boucher G, Lees CW, Franke A, D'Amato M, Taylor KD, Lee JC, Goyette P, Imielinski M, Latiano A, Lagacé C, Scott R, Amininejad L, Bumpstead S, Baidoo L, Baldassano RN, Barclay M, Bayless TM, Brand S, Büning C, Colombel JF, Denson LA, De Vos M, Dubinsky M, Edwards C, Ellinghaus D, Fehrmann RS, Floyd JA, Florin T, Franchimont D, Franke L, Georges M, Glas J, Glazer NL, Guthery SL, Haritunians T, Hayward NK, Hugot JP, Jobin G, Laukens D, Lawrance I, Lémann M, Levine A, Libioulle C, Louis E, McGovern DP, Milla M, Montgomery GW, Morley KI, Mowat C, Ng A, Newman W, Ophoff RA, Papi L, Palmieri O, Peyrin-Biroulet L, Panés J, Phillips A, Prescott NJ, Proctor DD, Roberts R, Russell R, Rutgeerts P, Sanderson J, Sans M, Schumm P, Seibold F, Sharma Y, Simms LA, Seielstad M, Steinhart AH, Targan SR, van den Berg LH, Vatn M, Verspaget H, Walters T, Wijmenga C, Wilson DC, Westra HJ, Xavier RJ, Zhao ZZ, Ponsioen CY, Andersen V, Torkvist L, Gazouli M, Anagnou NP, Karlsen TH, Kupcinskas L, Sventoraityte J, Mansfield JC, Kugathasan S, Silverberg MS, Halfvarson J, Rotter JI, Mathew CG, Griffiths AM, Gearry R, Ahmad T, Brant SR, Chamaillard M, Satsangi J, Cho JH, Schreiber S, Daly MJ, Barrett JC, Parkes M, Annese V, Hakonarson H, Radford-Smith G, Duerr RH, Vermeire S, Weersma RK, Rioux JD. Meta-analysis identifies 29 additional ulcerative colitis risk loci, increasing the number of confirmed associations to 47. *Nat Genet* 2011; **43**: 246-252 [PMID: 21297633 DOI: 10.1038/ng.764]
26 **Barrett JC**, Hansoul S, Nicolae DL, Cho JH, Duerr RH, Rioux JD, Brant SR, Silverberg MS, Taylor KD, Barmada MM, Bitton A, Dassopoulos T, Datta LW, Green T, Griffiths AM, Kistner EO, Murtha MT, Regueiro MD, Rotter JI, Schumm LP, Steinhart AH, Targan SR, Xavier RJ, Libioulle C, Sandor C, Lathrop M, Belaiche J, Dewit O, Gut I, Heath S, Laukens D, Mni M, Rutgeerts P, Van Gossum A, Zelenika D, Franchimont D, Hugot JP, de Vos M, Vermeire S, Louis E, Cardon LR, Anderson CA, Drummond H, Nimmo E, Ahmad T, Prescott NJ, Onnie CM, Fisher SA, Marchini J, Ghori J, Bumpstead S, Gwilliam R, Tremelling M, Deloukas P, Mansfield J, Jewell D, Satsangi J, Mathew CG, Parkes M, Georges M, Daly MJ. Genome-wide association defines more than 30 distinct susceptibility loci for Crohn's disease. *Nat Genet* 2008; **40**: 955-962 [PMID: 18587394 DOI: 10.1038/ng.175]
27 **Willer CJ**, Li Y, Abecasis GR. METAL: fast and efficient meta-analysis of genomewide association scans. *Bioinformatics* 2010; **26**: 2190-2191 [PMID: 20616382 DOI: 10.1093/bioinformatics/btq340]
28 **Wellcome Trust Case Control Consortium**. Genome-wide association study of 14,000 cases of seven common diseases and 3,000 shared controls. *Nature* 2007; **447**: 661-678 [PMID: 17554300 DOI: 10.1038/nature05911]
29 **de Bakker PI**, Ferreira MA, Jia X, Neale BM, Raychaudhuri S, Voight BF. Practical aspects of imputation-driven meta-analysis of genome-wide association studies. *Hum Mol Genet* 2008; **17**: R122-R128 [PMID: 18852200 DOI: 10.1093/hmg/ddn288]
30 **Aulchenko YS**, Ripke S, Isaacs A, van Duijn CM. GenABEL: an R library for genome-wide association analysis. *Bioinformatics* 2007; **23**: 1294-1296 [PMID: 17384015 DOI: 10.1093/bioinformatics/btm108]
31 **Mägi R**, Morris AP. GWAMA: software for genome-wide association meta-analysis. *BMC Bioinformatics* 2010; **11**: 288 [PMID: 20509871 DOI: 10.1186/1471-2105-11-288]
32 **Evangelou E**, Ioannidis JP. Meta-analysis methods for genome-wide association studies and beyond. *Nat Rev Genet* 2013; **14**: 379-389 [PMID: 23657481 DOI: 10.1038/nrg3472]
33 **Mooney MA**, Nigg JT, McWeeney SK, Wilmot B. Functional and genomic context in pathway analysis of GWAS data. *Trends Genet* 2014; **30**: 390-400 [PMID: 25154796 DOI: 10.1016/j.tig.2014.07.004]
34 **Wang K**, Li M, Bucan M. Pathway-based approaches for analysis of genomewide association studies. *Am J Hum Genet* 2007; **81**: 1278-1283 [PMID: 17966091 DOI: 10.1086/522374]
35 **Wang K**, Zhang H, Kugathasan S, Annese V, Bradfield JP, Russell RK, Sleiman PM, Imielinski M, Glessner J, Hou C, Wilson DC, Walters T, Kim C, Frackelton EC, Lionetti P, Barabino A, Van

Limbergen J, Guthery S, Denson L, Piccoli D, Li M, Dubinsky M, Silverberg M, Griffiths A, Grant SF, Satsangi J, Baldassano R, Hakonarson H. Diverse genome-wide association studies associate the IL12/IL23 pathway with Crohn Disease. *Am J Hum Genet* 2009; **84**: 399-405 [PMID: 19249008 DOI: 10.1016/j.ajhg.2009.01.026]
36 **Torkamani A**, Topol EJ, Schork NJ. Pathway analysis of seven common diseases assessed by genome-wide association. *Genomics* 2008; **92**: 265-272 [PMID: 18722519 DOI: 10.1016/j.ygeno.2008.07.011]
37 **Peng G**, Luo L, Siu H, Zhu Y, Hu P, Hong S, Zhao J, Zhou X, Reveille JD, Jin L, Amos CI, Xiong M. Gene and pathway-based second-wave analysis of genome-wide association studies. *Eur J Hum Genet* 2010; **18**: 111-117 [PMID: 19584899 DOI: 10.1038/ejhg.2009.115]
38 **Carbonetto P**, Stephens M. Integrated enrichment analysis of variants and pathways in genome-wide association studies indicates central role for IL-2 signaling genes in type 1 diabetes, and cytokine signaling genes in Crohn's disease. *PLoS Genet* 2013; **9**: e1003770 [PMID: 24098138 DOI: 10.1371/journal.pgen.1003770]
39 **Holmans P**, Green EK, Pahwa JS, Ferreira MA, Purcell SM, Sklar P, Owen MJ, O'Donovan MC, Craddock N. Gene ontology analysis of GWA study data sets provides insights into the biology of bipolar disorder. *Am J Hum Genet* 2009; **85**: 13-24 [PMID: 19539887 DOI: 10.1016/j.ajhg.2009.05.011]
40 **Zhang K**, Cui S, Chang S, Zhang L, Wang J. i-GSEA4GWAS: a web server for identification of pathways/gene sets associated with traits by applying an improved gene set enrichment analysis to genome-wide association study. *Nucleic Acids Res* 2010; **38**: W90-W95 [PMID: 20435672 DOI: 10.1093/nar/gkq324]
41 **Holden M**, Deng S, Wojnowski L, Kulle B. GSEA-SNP: applying gene set enrichment analysis to SNP data from genome-wide association studies. *Bioinformatics* 2008; **24**: 2784-2785 [PMID: 18854360 DOI: 10.1093/bioinformatics/btn516]
42 **Huang da W**, Sherman BT, Lempicki RA. Systematic and integrative analysis of large gene lists using DAVID bioinformatics resources. *Nat Protoc* 2009; **4**: 44-57 [PMID: 19131956 DOI: 10.1038/nprot.2008.211]
43 **Backes C**, Keller A, Kuentzer J, Kneissl B, Comtesse N, Elnakady YA, Müller R, Meese E, Lenhof HP. GeneTrail--advanced gene set enrichment analysis. *Nucleic Acids Res* 2007; **35**: W186-W192 [PMID: 17526521 DOI: 10.1093/nar/gkm323]
44 **Zhang B**, Kirov S, Snoddy J. WebGestalt: an integrated system for exploring gene sets in various biological contexts. *Nucleic Acids Res* 2005; **33**: W741-W748 [PMID: 15980575 DOI: 10.1093/nar/gki475]
45 **Shannon P**, Markiel A, Ozier O, Baliga NS, Wang JT, Ramage D, Amin N, Schwikowski B, Ideker T. Cytoscape: a software environment for integrated models of biomolecular interaction networks. *Genome Res* 2003; **13**: 2498-2504 [PMID: 14597658 DOI: 10.1101/gr.1239303]
46 **Rossin EJ**, Lage K, Raychaudhuri S, Xavier RJ, Tatar D, Benita Y, Cotsapas C, Daly MJ. Proteins encoded in genomic regions associated with immune-mediated disease physically interact and suggest underlying biology. *PLoS Genet* 2011; **7**: e1001273 [PMID: 21249183 DOI: 10.1371/journal.pgen.1001273]
47 **Franceschini A**, Szklarczyk D, Frankild S, Kuhn M, Simonovic M, Roth A, Lin J, Minguez P, Bork P, von Mering C, Jensen LJ. STRING v9.1: protein-protein interaction networks, with increased coverage and integration. *Nucleic Acids Res* 2013; **41**: D808-D815 [PMID: 23203871 DOI: 10.1093/nar/gks1094]
48 **Schmitt T**, Ogris C, Sonnhammer EL. FunCoup 3.0: database of genome-wide functional coupling networks. *Nucleic Acids Res* 2014; **42**: D380-D388 [PMID: 24185702 DOI: 10.1093/nar/gkt984]
49 **Wei WH**, Hemani G, Haley CS. Detecting epistasis in human complex traits. *Nat Rev Genet* 2014; **15**: 722-733 [PMID: 25200660 DOI: 10.1038/nrg3747]
50 **Roberts RL**, Topless RK, Phipps-Green AJ, Gearry RB, Barclay ML, Merriman TR. Evidence of interaction of CARD8 rs2043211 with NALP3 rs35829419 in Crohn's disease. *Genes Immun* 2010; **11**: 351-356 [PMID: 20182451 DOI: 10.1038/gene.2010.11]
51 **Matsuda H**. Physical nature of higher-order mutual information: intrinsic correlations and frustration *Phys Rev E Stat Phys Plasmas Fluids Relat Interdiscip Topics* 2000; **62**: 3096-3102 [PMID: 11088803]
52 **Wan X**, Yang C, Yang Q, Xue H, Fan X, Tang NL, Yu W. BOOST: A fast approach to detecting gene-gene interactions in genome-wide case-control studies. *Am J Hum Genet* 2010; **87**: 325-340 [PMID: 20817139 DOI: 10.1016/j.ajhg.2010.07.021]
53 **Ueki M**, Cordell HJ. Improved statistics for genome-wide interaction analysis. *PLoS Genet* 2012; **8**: e1002625 [PMID: 22496670 DOI: 10.1371/journal.pgen.1002625]
54 **Rokach L**, Maimon O. Data mining with decision trees: theory and applications. World Scientific Pub Co Inc., 2008
55 **Ritchie MD**, Hahn LW, Roodi N, Bailey LR, Dupont WD, Parl FF, Moore JH. Multifactor-dimensionality reduction reveals high-order interactions among estrogen-metabolism genes in sporadic breast cancer. *Am J Hum Genet* 2001; **69**: 138-147 [PMID: 11404819 DOI: 10.1086/321276]
56 **Winham SJ**, Motsinger-Reif AA. An R package implementation of multifactor dimensionality reduction. *BioData Min* 2011; **4**: 24 [PMID: 21846375 DOI: 10.1186/1756-0381-4-24]
57 **Wang MH**, Fiocchi C, Zhu X, Ripke S, Kamboh MI, Rebert N, Duerr RH, Achkar JP. Gene-gene and gene-environment interactions in ulcerative colitis. *Hum Genet* 2014; **133**: 547-558 [PMID: 24241240 DOI: 10.1007/s00439-013-1395-z]
58 **Okazaki T**, Wang MH, Rawsthorne P, Sargent M, Datta LW, Shugart YY, Bernstein CN, Brant SR. Contributions of IBD5, IL23R, ATG16L1, and NOD2 to Crohn's disease risk in a population-based case-control study: evidence of gene-gene interactions. *Inflamm Bowel Dis* 2008; **14**: 1528-1541 [PMID: 18521914 DOI: 10.1002/ibd.20512]
59 **Rogler G**, Vavricka S. Exposome in IBD: recent insights in environmental factors that influence the onset and course of IBD. *Inflamm Bowel Dis* 2015; **21**: 400-408 [PMID: 25358064 DOI: 10.1097/MIB.0000000000000229]
60 **Wang MH**, Achkar JP. Gene-environment interactions in inflammatory bowel disease pathogenesis. *Curr Opin Gastroenterol* 2015; **31**: 277-282 [PMID: 26039718 DOI: 10.1097/MOG.0000000000000188]
61 **Fiocchi C**. Genes and 'in-vironment': how will our concepts on the pathophysiology of inflammatory bowel disease develop in the future? *Dig Dis* 2012; **30** Suppl 3: 2-11 [PMID: 23295686 DOI: 10.1159/000342585]
62 **Helbig KL**, Nothnagel M, Hampe J, Balschun T, Nikolaus S, Schreiber S, Franke A, Nöthlings U. A case-only study of gene-environment interaction between genetic susceptibility variants in NOD2 and cigarette smoking in Crohn's disease aetiology. *BMC Med Genet* 2012; **13**: 14 [PMID: 22416979 DOI: 10.1186/1471-2350-13-14]
63 **Ananthakrishnan AN**, Nguyen DD, Sauk J, Yajnik V, Xavier RJ. Genetic polymorphisms in metabolizing enzymes modifying the association between smoking and inflammatory bowel diseases. *Inflamm Bowel Dis* 2014; **20**: 783-789 [PMID: 24651583 DOI: 10.1097/MIB.0000000000000014]
64 **Ruczinski I**, Kooperberg C, LeBlanc ML. Logic regression. *J Comput Graph Statis* 2003; **12**: 475-511
65 **Ruczinski I**, Kooperberg C, LeBlanc ML. Exploring interactions in high-dimensional genomic data: an overview of logic regression, with applications. *J Multivar Anal* 2004; **90**: 178-195
66 **Wang MH**, Fiocchi C, Ripke S. A model integrating genetic and environmental factors interactions can predict Crohn's disease risk. *Am J Gastroenterol* 2014; (**Suppl 109**): S503
67 **Satokari R**. Contentious host-microbiota relationship in inflammatory bowel disease--can foes become friends again? *Scand J Gastroenterol* 2015; **50**: 34-42 [PMID: 25523554 DOI: 10.3109/00365521.2014.966320]
68 **Noval Rivas M**, Burton OT, Wise P, Zhang YQ, Hobson SA, Garcia Lloret M, Chehoud C, Kuczynski J, DeSantis T, Warrington

J, Hyde ER, Petrosino JF, Gerber GK, Bry L, Oettgen HC, Mazmanian SK, Chatila TA. A microbiota signature associated with experimental food allergy promotes allergic sensitization and anaphylaxis. *J Allergy Clin Immunol* 2013; **131**: 201-212 [PMID: 23201093 DOI: 10.1016/j.jaci.2012.10.026]

69 **Caporaso JG**, Kuczynski J, Stombaugh J, Bittinger K, Bushman FD, Costello EK, Fierer N, Peña AG, Goodrich JK, Gordon JI, Huttley GA, Kelley ST, Knights D, Koenig JE, Ley RE, Lozupone CA, McDonald D, Muegge BD, Pirrung M, Reeder J, Sevinsky JR, Turnbaugh PJ, Walters WA, Widmann J, Yatsunenko T, Zaneveld J, Knight R. QIIME allows analysis of high-throughput community sequencing data. *Nat Methods* 2010; **7**: 335-336 [PMID: 20383131 DOI: 10.1038/nmeth.f.303]

70 **Knight R**, Maxwell P, Birmingham A, Carnes J, Caporaso JG, Easton BC, Eaton M, Hamady M, Lindsay H, Liu Z, Lozupone C, McDonald D, Robeson M, Sammut R, Smit S, Wakefield MJ, Widmann J, Wikman S, Wilson S, Ying H, Huttley GA. PyCogent: a toolkit for making sense from sequence. *Genome Biol* 2007; **8**: R171 [PMID: 17708774 DOI: 10.1186/gb-2007-8-8-r171]

71 **Schloss PD**, Westcott SL, Ryabin T, Hall JR, Hartmann M, Hollister EB, Lesniewski RA, Oakley BB, Parks DH, Robinson CJ, Sahl JW, Stres B, Thallinger GG, Van Horn DJ, Weber CF. Introducing mothur: open-source, platform-independent, community-supported software for describing and comparing microbial communities. *Appl Environ Microbiol* 2009; **75**: 7537-7541 [PMID: 19801464 DOI: 10.1128/AEM.01541-09]

72 **Huse SM**, Dethlefsen L, Huber JA, Mark Welch D, Relman DA, Sogin ML. Exploring microbial diversity and taxonomy using SSU rRNA hypervariable tag sequencing. *PLoS Genet* 2008; **4**: e1000255 [PMID: 19023400 DOI: 10.1371/journal.pgen.1000255]

73 **Brady A**, Salzberg S. PhymmBL expanded: confidence scores, custom databases, parallelization and more. *Nat Methods* 2011; **8**: 367 [PMID: 21527926 DOI: 10.1038/nmeth0511-367]

74 **Mitra S**, Rupek P, Richter DC, Urich T, Gilbert JA, Meyer F, Wilke A, Huson DH. Functional analysis of metagenomes and metatranscriptomes using SEED and KEGG. *BMC Bioinformatics* 2011; **12** Suppl 1: S21 [PMID: 21342551 DOI: 10.1186/1471-2105-12-S1-S21]

75 **Sharpton TJ**, Riesenfeld SJ, Kembel SW, Ladau J, O'Dwyer JP, Green JL, Eisen JA, Pollard KS. PhylOTU: a high-throughput procedure quantifies microbial community diversity and resolves novel taxa from metagenomic data. *PLoS Comput Biol* 2011; **7**: e1001061 [PMID: 21283775 DOI: 10.1371/journal.pcbi.1001061]

76 **von Mering C**, Hugenholtz P, Raes J, Tringe SG, Doerks T, Jensen LJ, Ward N, Bork P. Quantitative phylogenetic assessment of microbial communities in diverse environments. *Science* 2007; **315**: 1126-1130 [PMID: 17272687 DOI: 10.1126/science.1133420]

77 **Manichanh C**, Rigottier-Gois L, Bonnaud E, Gloux K, Pelletier E, Frangeul L, Nalin R, Jarrin C, Chardon P, Marteau P, Roca J, Dore J. Reduced diversity of faecal microbiota in Crohn's disease revealed by a metagenomic approach. *Gut* 2006; **55**: 205-211 [PMID: 16188921 DOI: 10.1136/gut.2005.073817]

78 **Martinez C**, Antolin M, Santos J, Torrejon A, Casellas F, Borruel N, Guarner F, Malagelada JR. Unstable composition of the fecal microbiota in ulcerative colitis during clinical remission. *Am J Gastroenterol* 2008; **103**: 643-648 [PMID: 18341488 DOI: 10.1111/j.1572-0241.2007.01592.x]

79 **Kang S**, Denman SE, Morrison M, Yu Z, Dore J, Leclerc M, McSweeney CS. Dysbiosis of fecal microbiota in Crohn's disease patients as revealed by a custom phylogenetic microarray. *Inflamm Bowel Dis* 2010; **16**: 2034-2042 [PMID: 20848492 DOI: 10.1002/ibd.21319]

80 **Sephri S**, Kotlowski R, Bernstein CN, Krause DO. Microbial diversity of inflamed and noninflamed gut biopsy tissues in inflammatory bowel disease. *Inflamm Bowel Dis* 2007; **13**: 675-683 [PMID: 17262808 DOI: 10.1002/ibd.20101]

81 **Ott SJ**, Kühbacher T, Musfeldt M, Rosenstiel P, Hellmig S, Rehman A, Drews O, Weichert W, Timmis KN, Schreiber S. Fungi and inflammatory bowel diseases: Alterations of composition and diversity. *Scand J Gastroenterol* 2008; **43**: 831-841 [PMID: 18584522 DOI: 10.1080/00365520801935434]

82 **Gevers D**, Kugathasan S, Denson LA, Vázquez-Baeza Y, Van Treuren W, Ren B, Schwager E, Knights D, Song SJ, Yassour M, Morgan XC, Kostic AD, Luo C, González A, McDonald D, Haberman Y, Walters T, Baker S, Rosh J, Stephens M, Heyman M, Markowitz J, Baldassano R, Griffiths A, Sylvester F, Mack D, Kim S, Crandall W, Hyams J, Huttenhower C, Knight R, Xavier RJ. The treatment-naive microbiome in new-onset Crohn's disease. *Cell Host Microbe* 2014; **15**: 382-392 [PMID: 24629344 DOI: 10.1016/j.chom.2014.02.005]

83 **Swidsinski A**, Ladhoff A, Pernthaler A, Swidsinski S, Loening-Baucke V, Ortner M, Weber J, Hoffmann U, Schreiber S, Dietel M, Lochs H. Mucosal flora in inflammatory bowel disease. *Gastroenterology* 2002; **122**: 44-54 [PMID: 11781279]

84 **Frank DN**, Robertson CE, Hamm CM, Kpadeh Z, Zhang T, Chen H, Zhu W, Sartor RB, Boedeker EC, Harpaz N, Pace NR, Li E. Disease phenotype and genotype are associated with shifts in intestinal-associated microbiota in inflammatory bowel diseases. *Inflamm Bowel Dis* 2011; **17**: 179-184 [PMID: 20839241 DOI: 10.1002/ibd.21339]

85 **Presley LL**, Ye J, Li X, Leblanc J, Zhang Z, Ruegger PM, Allard J, McGovern D, Ippoliti A, Roth B, Cui X, Jeske DR, Elashoff D, Goodglick L, Braun J, Borneman J. Host-microbe relationships in inflammatory bowel disease detected by bacterial and metaproteomic analysis of the mucosal-luminal interface. *Inflamm Bowel Dis* 2012; **18**: 409-417 [PMID: 21698720 DOI: 10.1002/ibd.21793]

86 **Lees CW**, Barrett JC, Parkes M, Satsangi J. New IBD genetics: common pathways with other diseases. *Gut* 2011; **60**: 1739-1753 [PMID: 21300624 DOI: 10.1136/gut.2009.199679]

87 **Gregersen PK**, Olsson LM. Recent advances in the genetics of autoimmune disease. *Annu Rev Immunol* 2009; **27**: 363-391 [PMID: 19302045 DOI: 10.1146/annurev.immunol.021908.132653]

88 **Rai E**, Wakeland EK. Genetic predisposition to autoimmunity--what have we learned? *Semin Immunol* 2011; **23**: 67-83 [PMID: 21288738 DOI: 10.1016/j.smim.2011.01.015]

89 **Richard-Miceli C**, Criswell LA. Emerging patterns of genetic overlap across autoimmune disorders. *Genome Med* 2012; **4**: 6 [PMID: 22284131 DOI: 10.1186/gm305]

90 **Neurath MF**. Cytokines in inflammatory bowel disease. *Nat Rev Immunol* 2014; **14**: 329-342 [PMID: 24751956 DOI: 10.1038/nri3661]

91 **Gersemann M**, Wehkamp J, Stange EF. Innate immune dysfunction in inflammatory bowel disease. *J Intern Med* 2012; **271**: 421-428 [PMID: 22324936 DOI: 10.1111/j.1365-2796.2012.02515.x]

92 **Cotsapas C**, Voight BF, Rossin E, Lage K, Neale BM, Wallace C, Abecasis GR, Barrett JC, Behrens T, Cho J, De Jager PL, Elder JT, Graham RR, Gregersen P, Klareskog L, Siminovitch KA, van Heel DA, Wijmenga C, Worthington J, Todd JA, Hafler DA, Rich SS, Daly MJ. Pervasive sharing of genetic effects in autoimmune disease. *PLoS Genet* 2011; **7**: e1002254 [PMID: 21852963 DOI: 10.1371/journal.pgen.1002254]

93 **Bhattacharjee S**, Rajaraman P, Jacobs KB, Wheeler WA, Melin BS, Hartge P, Yeager M, Chung CC, Chanock SJ, Chatterjee N. A subset-based approach improves power and interpretation for the combined analysis of genetic association studies of heterogeneous traits. *Am J Hum Genet* 2012; **90**: 821-835 [PMID: 22560090 DOI: 10.1016/j.ajhg.2012.03.015]

94 **Taylor JE**, Worsley KJ, Gosselin F. Maxima of discretely sampled random fields, with an application to 'bubbles'. *Biometrika* 2007; (**94**): 1-18

95 **O'Brien PC**. Procedures for comparing samples with multiple endpoints. *Biometrics* 1984; **40**: 1079-1087 [PMID: 6534410]

96 **Xu X**, Tian L, Wei LJ. Combining dependent tests for linkage or association across multiple phenotypic traits. *Biostatistics* 2003; **4**: 223-229 [PMID: 12925518 DOI: 10.1093/biostatistics/4.2.223]

97 **Yang Q**, Wu H, Guo CY, Fox CS. Analyze multivariate phenotypes in genetic association studies by combining univariate association tests. *Genet Epidemiol* 2010; **34**: 444-454 [PMID: 20583287 DOI: 10.1002/gepi.20497]

98 **Chung D**, Yang C, Li C, Gelernter J, Zhao H. GPA: a statistical approach to prioritizing GWAS results by integrating pleiotropy and annotation. *PLoS Genet* 2014; **10**: e1004787 [PMID: 25393678 DOI: 10.1371/journal.pgen.1004787]

99 **Efron B**. Microarrays, empirical Bayes and the two-groups model. *Statistical Science* 2008; 1-22

100 **Dempster AP**, Laird NM, Rubin DB. Maximum likelihood from incomplete data via the EM algorithm. *J R Stat Soc Series B Stat Methodol* 1977; 1-38

101 **Schifano ED**, Li L, Christiani DC, Lin X. Genome-wide association analysis for multiple continuous secondary phenotypes. *Am J Hum Genet* 2013; **92**: 744-759 [PMID: 23643383 DOI: 10.1016/j.ajhg.2013.04.004]

102 **Wei Z**, Wang W, Bradfield J, Li J, Cardinale C, Frackelton E, Kim C, Mentch F, Van Steen K, Visscher PM, Baldassano RN, Hakonarson H. Large sample size, wide variant spectrum, and advanced machine-learning technique boost risk prediction for inflammatory bowel disease. *Am J Hum Genet* 2013; **92**: 1008-1012 [PMID: 23731541 DOI: 10.1016/j.ajhg.2013.05.002]

103 **Li C**, Yang C, Gelernter J, Zhao H. Improving genetic risk prediction by leveraging pleiotropy. *Hum Genet* 2014; **133**: 639-650 [PMID: 24337655 DOI: 10.1007/s00439-013-1401-5]

104 **ENCODE Project Consortium**. An integrated encyclopedia of DNA elements in the human genome. *Nature* 2012; **489**: 57-74 [PMID: 22955616 DOI: 10.1038/nature11247]

105 **Kundaje A**, Meuleman W, Ernst J, Bilenky M, Yen A, Heravi-Moussavi A, Kheradpour P, Zhang Z, Wang J, Ziller MJ, Amin V, Whitaker JW, Schultz MD, Ward LD, Sarkar A, Quon G, Sandstrom RS, Eaton ML, Wu YC, Pfenning AR, Wang X, Claussnitzer M, Liu Y, Coarfa C, Harris RA, Shoresh N, Epstein CB, Gjoneska E, Leung D, Xie W, Hawkins RD, Lister R, Hong C, Gascard P, Mungall AJ, Moore R, Chuah E, Tam A, Canfield TK, Hansen RS, Kaul R, Sabo PJ, Bansal MS, Carles A, Dixon JR, Farh KH, Feizi S, Karlic R, Kim AR, Kulkarni A, Li D, Lowdon R, Elliott G, Mercer TR, Neph SJ, Onuchic V, Polak P, Rajagopal N, Ray P, Sallari RC, Siebenthall KT, Sinnott-Armstrong NA, Stevens M, Thurman RE, Wu J, Zhang B, Zhou X, Beaudet AE, Boyer LA, De Jager PL, Farnham PJ, Fisher SJ, Haussler D, Jones SJ, Li W, Marra MA, McManus MT, Sunyaev S, Thomson JA, Tlsty TD, Tsai LH, Wang W, Waterland RA, Zhang MQ, Chadwick LH, Bernstein BE, Costello JF, Ecker JR, Hirst M, Meissner A, Milosavljevic A, Ren B, Stamatoyannopoulos JA, Wang T, Kellis M. Integrative analysis of 111 reference human epigenomes. *Nature* 2015; **518**: 317-330 [PMID: 25693563 DOI: 10.1038/nature14248]

106 **GTEx Consortium**. Human genomics. The Genotype-Tissue Expression (GTEx) pilot analysis: multitissue gene regulation in humans. *Science* 2015; **348**: 648-660 [PMID: 25954001 DOI: 10.1126/science.1262110]

107 **Frayling IM**, Beck NE, Ilyas M, Dove-Edwin I, Goodman P, Pack K, Bell JA, Williams CB, Hodgson SV, Thomas HJ, Talbot IC, Bodmer WF, Tomlinson IP. The APC variants I1307K and E1317Q are associated with colorectal tumors, but not always with a family history. *Proc Natl Acad Sci USA* 1998; **95**: 10722-10727 [PMID: 9724771]

108 **Fearnhead NS**, Wilding JL, Winney B, Tonks S, Bartlett S, Bicknell DC, Tomlinson IP, Mortensen NJ, Bodmer WF. Multiple rare variants in different genes account for multifactorial inherited susceptibility to colorectal adenomas. *Proc Natl Acad Sci USA* 2004; **101**: 15992-15997 [PMID: 15520370 DOI: 10.1073/pnas.0407187101]

109 **Cohen JC**, Boerwinkle E, Mosley TH, Hobbs HH. Sequence variations in PCSK9, low LDL, and protection against coronary heart disease. *N Engl J Med* 2006; **354**: 1264-1272 [PMID: 16554528 DOI: 10.1056/NEJMoa054013]

110 **Lettre G**. Rare and low-frequency variants in human common diseases and other complex traits. *J Med Genet* 2014; **51**: 705-714 [PMID: 25185437 DOI: 10.1136/jmedgenet-2014-102437]

P- Reviewer: Cheon JH, Lakatos PL **S- Editor**: Ma YJ
L- Editor: Filipodia **E- Editor**: Liu XM

Baishideng®

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

ISSN 1007-9327