Early Career Investigators Meeting on Quantitative Problems in Human Health and Genetics

Noah Zaitlen (University of California, San Francisco),
Tuuli Lappalainen (New York Genome Center and Columbia University),
Michael Hoffman (Princess Margaret Cancer Centre and University of Toronto),
Julien Ayroles (Princeton University),
Jennifer Listgarten (Microsoft Research)

Jan 10 to Jan 15, 2016

1 Overview of the Field

The results of the last five years of genetic association studies for complex human diseases have revealed that most diseases originate in part from the effects of multiple genetic variants found throughout the human genome. Researchers held that identification, or "mapping" of such risk variants would improve our understanding of the biological mechanisms of complex diseases and assist in accurate diagnosis and risk prediction, eventually leading to better health solutions. However, the common variants identified by GWAS explain only a modest proportion of the genetic variance for most traits, and thus much of the genetic variation contributing to disease remains unknown. Furthermore, merely identifying genetic associations does not lead to better medical interventions without detailed, high-throughput characterization of causal molecular pathways underlying genetic associations to disease.

Large resource efforts like the Human Genome Project [1], the International HapMap Project [16], the 1000 Genomes Project [17] and the ENCODE Project [2] have provided investigators with exceptional resources to facilitate such genetic studies. These resources include sequences of thousands of human genomes with a dense genome-wide catalogue of genetic variation in humans, and an increasingly detailed annotation of functional elements in the genome. These resources are complemented with emerging large data sets from phenotyped cohorts—with variable data access policies—and massive amounts of new data resources arrive every day via high-throughput sequencing, genomic and epigenomic technology. Unfortunately, the pace of technological innovation has outstripped the development of powerful and flexible statistical and computational methods allowing us to fully take advantage of these resources.

As newfound genetic understanding is acquired, strongly held beliefs about the genetic architecture of disease are actively being reshaped. Many factors other than the uncovered genome-wide association study (GWAS) loci must exist to explain the observed variance of human traits, and identifying these is one of the main challenges of modern genetics. Furthermore, while genome-wide assays in functional genomics have enabled simultaneous analyses of thousands of loci, linking these data to both GWAS and phenotype data remains challenging. Thus, it is clear that important information remains buried in existing and emerging data, and that novel, robust statistical analysis is the core foundation upon which to extract it.

2 Recent Developments and Open Problems

Here we present the recent developments and open problems related to the primary fields of interest of discussed during our meeting.

2.1 Genetic regulation of transcription

Integrating genetic and transcriptomic data by associating genetic variation with gene expression levels has become one of the most popular approaches for interpreting how genome variation affects cellular phenotypes. During the recent years, thousands of such genetic associations or expression quantitative trait loci (eQTLs) have been found. A main motivation in eQTL analysis is that most GWAS associations are in regulatory regions, and are enriched in eQTLs. Finding the regulatory variants underlying GWAS associations and understanding their context-specific pattern of activity (for example, tissue-specificity), is a key for understanding biological processes behind each genetic association to disease. This will ultimately make targeted medical interventions possible. Furthermore, eQTLs provide a unique resource for understanding functional genetic variation, due to their high numbers, effect sizes, and relatively simple mechanistic links from genotype to phenotype.

Large-scale eQTL projects such as the Genotype-Tissue-Expression Project (GTEx) [18], DNG [19] and Geuvadis [20] have recently provided unprecedented resources for analyzing genetic effects on gene expression, yielding catalogues of thousands of eQTLs. However, the practical use of these data to interpret GWAS associations has not been entirely straightforward, and the vast majority of GWAS loci lack an eQTL that would provide a plausible cellular mechanism of disease etiology. The effect of a given eQTL variant on gene expression can vary between individuals due to context-dependence of genome function—that is, the effect of a variant in a given cell in a given individual may depend on the individual's genetic background or history of environmental exposures, or on the precise type and state of the cell. These effects are usually not well captured by the existing statistical methods applied on mostly observational data sets. The current eQTL data sets are highly focused on common proximal eQTLs in promoter areas *in cis*, and lack information of rare regulatory variants, eQTL activity in specific cell types, and their interactions with the environment and other genetic variants. Many of the current challenges in functional population genetics are due to limitations both in data and methods. While integrated genome and RNA-sequencing data now exists from several hundreds of individuals, these sample sizes are still too small for many approaches. Furthermore, data from multiple specific cell types or even at the single-cell level would be ideal to capture effects at the true biological unit of the cell. Analysis of RNA-sequencing data is also not entirely trivial, and careful accounting for technical sources of bias and variance is necessary. Finally, statistical methods for efficient mapping of different types of genetic effects on the transcriptome are constantly evolving to push the borders of current knowledge.

2.2 Functional genomics and variant annotation

One of the great challenges in genome science is to understand the functional consequences of sequence variation at the molecular level and determine how and when it contributes to phenotypic differences among individuals. To fully understand the consequences of genetic variation, we need to go beyond a description of the relationships between individual polymorphic sites and phenotypic variation. New experimental and analytical tools now offer an unprecedented view of how dynamic and context dependent our genome really is. Methods ranging from ATACseq (a method for assaying chromatin accessibility genome-wide) to Hi-C (a method to study the three-dimensional architecture of genomes) are finally allowing us to study the regulatory code of the genome [15].

A crucial tool at our disposal to decrypt the regulatory genome is natural genetic variation. The mutation segregating between individuals are an ideal systems level perturbation from which to study consequences. In that context both the external environment and the perturbations in DNA sequence in natural populations can be exploited. Different levels of biological organization (e.g. sequence variation, gene expression, epigenomic variation) can be weaved into networks of interacting elements, allowing us to model correlative and causative relationships [13]. This integrative approach will allow us to draw the causal path from variation in allele frequencies to phenotypic differences.

A key issue remains genome annotation, but recent community efforts, have dramatically improved our ability to interpret functional genomics data and to go behind the relatively simple quantification of RNA level. Project such as ENCODE project (Encyclopedia of DNA Elements [2] are redefining our understanding of what are functional DNA elements, from the role of transcriptional modulators binding sites, to effect of chromatin state.

2.3 Population history and evolutionary genetics

The field of human medical genetics is tasked with identifying and understanding the genetic variants that give rise to disease. To accomplish this goal, a fundamental understanding of the forces that created and shaped genetic variation is required. Furthermore, we have repeatedly observed that the most powerful and robust methods for detecting genotype-phenotype association are informed by population and evolutionary genetics models. Evolutionary genetics, the study of how evolutionary forces shaped our genomes globally or in specific loci, is especially relevant to medical genetics in the sense that fitness, the core phenotype of evolutionary genetics, is effectively the most important measure of human health.

Sequencing large numbers of individuals from around the world and throughout large spans of time has opened the possibility to refine and evaluate the theories of population genetics that were developed over the course of the last hundred years. It has also brought into focus the dramatic effect that human demographic history has played in shaping patterns of genetic variation. The hyper-exponential expansion of human population resulted in a dramatic skew in the minor allele frequency spectrum, with an excess of rare variants segregating in populations. Relating the distribution of effect sizes and allele frequency is central to mapping efforts. As a result, demographic history is also central to understanding complex trait genetics. Medical genomics uses sequencing to study the effect of low-frequency variants ($< 1\%$) on complex traits, but important questions connecting population and quantitative genetics still must be addressed. One of the primary open problems is that of reconstructing demographic of existing or ancient human populations including migrations, admixtures, bottlenecks, and expansions. While there many statistics methods exist, they all have to make simplifying assumptions in order to accommodate mathematical or computational constraints. Other open question include: How do we study the allele frequency spectrum of mutations affecting complex diseases from diverse populations? How has population growth affected the genetic heterogeneity of diseases, and how does this affect the power to detect associations? Does this inflated class of rare mutations play a disproportionate role in disease?

To this day most tests aimed at detecting the contribution of rare variants to disease emergence are ill suited to address these fundamental questions. Most test search for an excess of polymorphic sites within a gene but fails to accommodate a key issue: most individuals who are affected by recessive disorders are heterozygous for two distinct defective alleles. SKAT (SNP-set (Sequence) Kernel Association Test [14]) for example, tests across a population sample, contrasting a control populations to affected cases and asks if there is an excess of rare variants in the cases compared to the control group without regard to the biology leading to disease emergence or the functional consequences of the variants considered. A promising avenues may be found in the analysis of full genome sequences where one can scan a genome for rare variants and determine possible functional mechanisms that may lead a carrier to be afflicted (e.g. disruption of biding site, enhancer).

2.4 Statistical genetics in very large cohorts

Cohorts available for human genetic and epidemiological studies are undergoing a big data revolution driven by exponentially decreasing costs of high-throughput genomic assays, widespread use of electronic medical records, and large-scale enrollment initiatives. The UK Biobank project, and cohorts being constructed at Vanderbilt, the U.S. Department of Veteran Affairs, Kaiser, Brigham and Womens Hospital, and Icahn School of Medicine at Mount Sinai, will each soon exceed 100,000 genotyped individuals tied to electronic medical records with hundreds of clinical measurements and with many individuals containing additional high-throughput transcriptomic, epigenetic, and metabololmic measurements. These very large cohorts (VLCs) are a tremendously useful resource especially for uncovering environmental interactions with genotype effects. However, VLCs also introduce tremendous computational challenges since many standard protocols incorporate quadratic algorithms unfeasible in VLCs.

When relating genotype to phenotype by way of GWASes, population structure, family relatedness, and non-independence of phenotypes can reduce power and cause spurious association. Novel methods capable of integrating such diverse, multi-dimensional information in order to improve association, prediction, and causality analyses include approaches such linear mixed models with various types of genetic similarity to model covariance structures, principal components analysis (PCA), and related approaches such as factor analysis. Although these approaches are well-established in other domains, tailoring and adapting them to problems in genetics is an ongoing open problem with many remaining challenges. Additionally, state-of-the art approaches remain computationally expensive, and much active research in the area focuses on improving the computational efficiency such that modern data sets can be analyzed in full force. Statistical methods development in this area will allow (1) joint examination of disease phenotypes for shared genetic and environmental components, (2) transferring risk prediction protocols between diverse populations as encountered in clinical settings, and (3) methods for integration of genetic, transcriptomic, and epigenomic information in identifying disease loci.

3 Presentation Highlights

This meeting had up to six 50 minute presentations per day from each of the presenters. Participants were encouraged to present unpublished work using a mixture of slides and chalk talks, and at the end presented an open problem to the group for discussion. This informal format fostered deep technical discussions. Staying at a small, remote institute also fostered things like staying up until 3 a.m. discussing population genetics. In this section, we briefly describe each talk.

3.1 Functional genomics and epigenomics.

Several researchers presented work applying quantitative methods to new epigenomics and single-cell technologies. These presentations focused on modeling and understanding the processes of gene regulation.

Michael Hoffman presented a method for joint analysis of dozens of epigenomic datasets [4]. Each one of those datasets describes a particular biochemical property of a cell type's epigenome. This method uses unsupervised machine learning to simultaneously segment and cluster the genome, identifying recurring patterns across multiple datasets. Hoffman also presented a new "expanded alphabet" position weight matrix model for understanding how epigenetic modifications affect transcription factor binding [5].

Cole Trapnell described how to understand single-cell epigenomics and transcriptomics data using manifold learning. Trapnell used his method to understand the complete developmental trajectory of a population of cells ranging from stem cells to terminally differentiated cells. Manifold learning disentangles transcriptional variation from distinct cell types, allowing the elucidation of an embedded cell graph.

Jason de Koning presented models for predicting the functional impact of human genetic variation. First, he presented a theoretical framework for understanding the effects of population-scale genetic processes on variable rates of molecular evolution. Second, he presented computationally tractable methods for Bayesian inference of heterogeneous mutation and selection across the vertebrates.

3.2 Genetic regulation of transcription.

Several participants study how genetic variants affect the transcriptome, and we had excellent presentations on this topic with lively discussion.

Jimmie (Chun) Ye described recent results from the ImmVar consortium [9], with eQTL analysis in dentritic cells collected from a large number of individuals, after in vitro treatments. The large number of eQTLs discovered in the ImmVar study highlighted the power of in vitro analysis to discover eQTLs that respond to environmental conditions such as infectious stimulus. He also described CRISPR validation of some of their QTLs, and results that show a key role of splicing variation in genetic regulatory effects as a response of infectious stimulus. These findings have relevance in understanding cellular response to stimuli, and in interpreting GWAS loci.

Alexis Battle focused on analysis of rare regulatory variants which have been beyond the scope of most eQTL studies, even though the vast majority of genetic variation is rare and multiple lines of evidence suggest that rare regulatory variants have a key role in inter-individual variation in gene expression, as well as

disease risk. She described a novel Bayesian method that integrates multiple layers of evidence from expression levels, allelic expression, variant annotation etc. to maximize the power to capture rare variant effects. Application of her method to GTEx data showed promising signals.

Tuuli Lappalainen gave a chalk talk to raise discussion of the dichotomy between molecular and quantitative genetics, which have distinct historical roots, scientific communities, questions and methods, but these two fields now need each other more than ever [6]. She suggested that functional population genetics studying molecular effects of genetic variants have the opportunity to tie these fields together towards a unified understanding of the human genome and its function. She discussed findings and perceptions that molecular genetics versus genetic association studies have of cell-type specificity, interactions, and context-specificity, and described biological mechanisms and gaps in the data that could explain the discrepancies. Finally, she led an active discussion of the utility of eQTL analysis in understanding GWAS loci—the past expectations, the current reality, and the possible solutions to being able to better use eQTL data to understand GWAS loci.

Julien Ayroles presented recent work on variance QTLs (vQTLs) where genetic variants associate to variance rather than mean of a trait. He described earlier work showing this in *Drosophila* behavioral traits [3], and novel work on variance of gene expression levels. Applying his statistical approach to existing data sets, he found a large number of vQTLs *in cis* and *trans*, which can be due to various biological mechanisms. These loci are particularly interesting: 1) variance-controlling QTLs appear to be relatively common, yet geneticists know little about their mode of action; 2) Because the contribution of v-eQTLs to phenotypic variance is not accessible to traditional GWAS, we have yet to determine their contribution to heritability; 3) v-eQTLs open a window into previously undetected GxG and GxE interactions. His talk stirred a long discussion of the replicability standards that should be applied to variance QTL studies, and the mechanistic models behind variance traits.

Chris Cotsapas presented analysis of integrating existing eQTL data with autoimmune disease loci, discussing the practical challenges, advances and caveats for applying eQTL data to interpret GWAS. These include cell-type specificity and power of eQTL data, GWAS data with shared controls, and complexity of linkage disequilibrium structure. He described findings of some loci with solid co-localization of the two signals, which could be followed up in further functional studies. He also highlighted that in most loci loci the current eQTL data from seemingly relevant tissues does no provide any clues to the functional mechanism, suggesting that the current data and/or methods are insufficient to describe cellular effects of many GWAS loci.

3.3 Population history and evolutionary genetics

Simon Gravel discussed the structure and demographic history of African-Americans and the implications these elements have for medical genetics. He focused specifically on regional variation an African ancestry component throughout the Americas and as well as dating timing and amount of admixture between African, European, and Native American populations. He showed a model for determining demographic history from identity-by-descent (IBD) data in admixed populations, in which IBD track lengths and sharing between local ancestry segments can be used to infer migration times and population sizes. The demographics events inferred from his model recapitulate known historical estimates of population density and the movement rates of African and European individuals around the United States throughout time.

Elinor Karlsson presented artificial selection for behavior in dogs that has occurred due to the history of dog breeding in human populations. Dogs were the first domesticated species with many breeds originating 30–70 generations ago from Europe. The populations sizes dropped from effective population size $N_e = 2000$ to $N_e = 40$–80 for many breeds and she showed that there are many overlaps between dog and human psychiatric disorder including, and obsessive compulsive disorder in dogs is similar to obsessive-compulsive disorder in humans including response to treatment with selective serotonin reuptake inhibitors, a GWAS hit for the gene *CDH2*. She showed that additional genotyping with denser scans gave more hits and confirmation of several with electrophoretic mobility shift assays and replication in other breeds with the genes sharing similar pathways to humans. One of the primary messages of her presentation was the claim that if behavior is selected then the variants should be high frequency and high penetrance, and therefore dogs represent a powerful resource for identifying genes underlying behavioral traits. She also discussed her new project, "Darwin's Dogs", where individuals can send in DNA of their own dogs and answer behavioral question about their dogs to help map new dog behavioral loci (http://darwinsdogs.org/).

3.4 Statistical genetics in very large cohorts

Noah Zaitlen presented two methods that improve the power of detecting associations when a large number of correlated variables have been measured on the same samples. He showed analyses over real and simulated data that provided direct support that large sets of correlated variables can be leveraged to achieve dramatic increases in statistical power equivalent to a two or even three or four fold increase in sample size. The method hinged on an equivalence testing framework to remove covariates from a linear regression model that could either remove signal or induce bias. He then showed an extension to linear mixed models that allowed the multi-phenotype methods to be run in data sets with related individuals.

Bogdan Pasainuc discussed his work to identify the causal mechanism between genetic variation and disease. In the quest to address this gap, post-GWAS studies are experiencing a "big data" revolution driven by the exponentially decreasing costs of high-throughput genomic assays. Multiple layers of data (such as genetic variation, transcriptome levels, epigenetic modifications, localization of tissue-specific regulatory sites, and others) are routinely collected in increasingly large cohorts of individuals. He discussed his PAINTOR software to fine-map and detect enrichment of functional categories in GWAS [7]. He also presented his work on imputation of gene-expression into GWAS cohorts in order to test for association between imputed expression and phenotype [8].

Benjamin Neale broadly examined opportunities and challenges in large scale data sets. He first discussed new genotyping arrays that he and others are developing for use on a large number of samples, potentially reaching into the millions. He described the motivation behind the selection of the SNPs for the array including both medical and population genetic utility. He then discussed the LD-score [11] method for estimating the heritability of phenotypes, the components of heritability of attributed to different functional categories and/or across minor allele frequencies. He showed that the method could be used for meta-analysis or correlated traits to improve power. He discussed the tradeoffs of using inbred vs multiple outbred populations in association studies. He showed the enrichment of rare variants of certain categories in *de novo* mutations in autism and showed that the Exome Aggregation Consortium (ExAC) [12] is useful for finding causal variants. He concluded with a discussion of the Variant Call Format (VCF) file format for sequencing data, which is too big to scale and presented some information on "hail", a distributed compute cloud for genomes.

Joe Pickrell began by raising the questions, "What is an association?" and "What does it mean for a genotype to cause a phenotype?". He suggested that a natural answer in to consider the effect of altering a genotype at the time of conception and the suggested that this model is wrong when parents are considered as in the case of spina bifida and folate levels. He then discussed the utility of parental genotypes in the study of disease and used the UK Biobank data set, which has extensive parental information, to explore their utility. He showed that he was able to replicate phenotype associations after correcting for the 50% sharing between parent and child. He also presented work on a genome-wide scan for genetic variants that influence multiple human phenotypes by comparing large genome-wide association studies (GWAS) of 42 traits or diseases, including anthropometric traits (e.g. nose size and male pattern baldness), immune traits (e.g. susceptibility to childhood ear infections and Crohn's disease), and psychiatric diseases (e.g. schizophrenia and Parkinson's disease). He identified 341 loci (at a false discovery rate of 10%) that influence multiple traits. He used these loci to identify traits that share multiple genetic causes in common. Finally, he developed a method to identify pairs of traits that show evidence of a causal relationship, and used this to identify four such pairs. For example, he showed evidence that increased BMI causally increases triglyceride levels, and that increased liability to hypothyroidism causally decreases adult height.

4 Scientific Progress Made

The discussion of genetic effects on the transcriptome was extensive and of extremely high quality and depth, partially due to the fact that nearly half of the participants work in this field and are familiar with the data, methods, and previous work. Many participants felt that the early expectations of eQTL studies easily explaining much of GWAS results have not currently been met once again, the problem has proven to be harder to solve than first anticipated. Yet, there was optimism that the eQTL approach, together with other cellular QTLs, will be very fruitful in the future, but larger data sets and better statistical methods are needed.

5 Outcome of the Meeting

In addition to the strictly scholarly content at our meeting and following our proposal for this meeting, we had extensive discussions both formally and informally on the challenges of early career scientists and the vision that our generation has for the future of the scientific community in our field.

Enhancing collaboration. Noah Zaitlen described his experience of student exchange with a collaborating lab for a summer, which had been a great way to enhance collaboration between the labs and help the students to build their networks. There was general enthusiasm in encouraging this, between the groups attending this meeting and in the future even broader within our field. We discussed various ways to enhance this and find funding, options including small exchange fellowships within consortia, proposing a fellowship program to NIH/NSF, institutional funding, and using travel funding on grants. Since the necessary funds are not that big, one should watch out for administrative overhead. Several PIs who attended this meeting are considering student exchanges for the following summer.

Training of students. We discussed the current training that students and postdocs have our responsibility, and the responsibility that we have in improving it. Many of us felt that quantitative reasoning and practical skills are too often lacking, even though students are excited and eager to do practical computational work. While PhD programs have some responsibility in providing this training and beginning students up to speed, we have the best insight on what are the skills that students will need, and should be proactive in trying to provide this training. MOOCs are one option; in general a useful format for a course is to take a data set and push it through analysis, QC and interpretation.

Publishing. There was lively discussion on the use of preprints servers. Many of the participants are strong believers in the benefits and value of publishing all (or nearly all) of their work as preprints prior to formal publication in a journal, but some also raised doubts if this would put them—as young investigators—in disadvantage compared to big senior labs that do not release their work as preprints.

Data access. For much of the work that we do, access to relevant data is absolutely essential, and often one of the biggest roadblocks, demonstrated by many anecdotes of ridiculous data access problems that we discussed. When we are producing data ourselves, we should make it as accessible as possible, but the practical constraint in this is often the IRB. This is an acknowledged problem and beyond our influence.

Unique advantages of this meeting. Genomics is a field where young investigator and senior trainees have to learn not only scientific skills but also management of highly collaborative projects, and navigating this environment poses its own challenges. Young investigators have potential to drive the development of the field for the next 40 years, and reinforcing their training, networks and vision has major potential effects.

Many of the participants expressed that this was one of the best or the best meeting they have attended. Participants are committed to continuing to organize this with some overlap of participants over the years. This forum with only early career investigators from a relatively well-defined field was productive in many ways. Being around peers with similar experiences and career stage allowed very honest discussion and confidential atmosphere without the need to try to impress senior scientists or recruit trainees. Small meetings such as this one also allow more direct discussion of scientific problems and caveats with existing approaches. The mixture of 50 min chalk talks, active discussion, and time for deep discussion on diverse topic during the breaks and in the evening led to in-depth on how methods and ideas, in a friendly environment. Also the evening/night gatherings at the lounge, with refreshments provided by our sponsor Illumina, were a valuable forum for informal discussion.

Our group of participants was diverse in terms of location (although mostly US/Canada) and gender, and represented a spectrum of fields with enough of overlap for productive discussion.

Participants left with new scientific knowledge, feedback on their work, and new scientific contacts. But unlike many other meeting formats, we are confident that this meeting will have long-term impacts on its participant. Here, we left with a lasting scientific community, many attendees are now collaborators, are jointly exploring new scientific avenues, are writing grants and publications together. . As a testimony to the

success of this meeting and in spite of busy schedules, most of the attended demanded that such a meeting be organized on a yearly basis.

References

[1] International Human Genome Sequencing Consortium, Initial sequencing and analysis of the human genome, *Nature* **409** (2001), 860–921.

[2] ENCODE Project Consortium, An Integrated Encyclopedia of DNA Elements in the Human Genome, *Nature* **489** (2012), 57–74.

[3] J. F. Ayroles, S. M. Buchanan, C. O'Leary, K. Skutt-Kakaria, J. K. Grenier, A. G. Clark, D. L. Hartl, and B. L. de Bivort, Behavioral idiosyncrasy reveals genetic control of phenotypic variability, *Proc Natl Acad Sci U S A* **112** (2015), 6706–11.

[4] M. M. Hoffman, O. J. Buske, J. Wang, Z. Weng, J. A. Bilmes, and W. S. Noble, Unsupervised pattern discovery in human chromatin structure through genomic segmentation, *Nature Methods* **5** (2012), 473–476.

[5] C. Viner, J. Johnson, N. Walker, H. Shi, M. Sjberg, D. J. Adams, A. C. Ferguson-Smith, T. L. Bailey, and M. M. Hoffman, Modeling methyl-sensitive transcription factor motifs with an expanded epigenetic alphabet, *bioRxiv* (2016), doi:10.1101/043794.

[6] T. Lappalainen, Functional genomics bridges the gap between quantitative genetics and molecular biology, *Genome Research* **25** (2015), 1427–1431.

[7] G. Kichaev, W. Y. Yang, S. Lindstrom, F. Hormozdiari, E. Eskin, A. L. Price, P. Kraft, B. Pasaniuc, Integrating functional data to prioritize causal variants in statistical fine-mapping studies, *PLoS Genet* **10** (2014), e1004722.

[8] Gusev A, Ko A, Shi H, Bhatia G, Chung W, Penninx BW, Jansen R, de Geus EJ,Boomsma DI, Wright FA, Sullivan PF, Nikkola E, Alvarez M, Civelek M, Lusis AJ,Lehtimki T, Raitoharju E, Khnen M, Seppl I, Raitakari OT, Kuusisto J, Laakso M, Price AL, Pajukanta P, Pasaniuc B. Integrative approaches for large-scale transcriptome wide association studies, *Nat Genet* **245** (2016), 26854917.

[9] C. J. Ye, T. Feng, H. K. Kwon, T. Raj, M. T. Wilson, N. Asinovski, C. McCabe, M. H. Lee, I. Frohlich, H. I. Paik, N. Zaitlen, N. Hacohen, B. Stranger, P. De Jager, D. Mathis, A. Regev, C. Benoist, Intersection of population variation and autoimmunity genetics in human T cell activation, *Science* **345** (2014), 1254665.

[10] A. P. J. de Koning, Wanjun Gu, and D. D. Polock, Rapid Likelihood Analysis on Large Phylogenies Using Partial Sampling of Substitution Histories, *Molecular Biology and Evolution* **27** (2010), 249–265.

[11] Bulik-Sullivan BK, Loh PR, Finucane HK, Ripke S, Yang J; Schizophrenia Working Group of the Psychiatric Genomics Consortium, Patterson N, Daly MJ, Price AL, Neale BM. LD Score regression distinguishes confounding from polygenicity in genome-wide association studies. *Nat Genet* **291** 2015, 25642630

[12] Ware, J. S., and Exome Aggregation Consortium. Analysis of protein-coding genetic variation in 60,706 humans. *BioRxiV* **doi:10.1101/030338** 2016.

[13] Civelek M, Lusis AJ. Systems genetics approaches to understand complex traits. *Nature Reviews Genetics* textbf15(1) (2014) 34–48.

[14] 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. *The American Journal of Human Genetics*. **89** (2011), 82–93.

[15] Albert FW, Kruglyak L. The role of regulatory variation in complex traits and disease. *Nat Rev Genet* **16** (2015) 197-212.

[16] International HapMap 3 Consortium, Integrating common and rare genetic variation in diverse human populations, *Nature* **467** (2010), 52–8

[17] 1000 Genomes Project Consortium, A global reference for human genetic variation, *Nature* **7571**) (2015) 68–74

[18] Ardlie KG, Deluca DS, Segr AV, Sullivan TJ, Young TR, Gelfand ET, Trowbridge CA, Maller JB, Tukiainen T, Lek M, Ward LD. The Genotype-Tissue Expression (GTEx) pilot analysis: Multitissue gene regulation in humans. *Science* **348** (2015), 648–60.

[19] Battle A, Mostafavi S, Zhu X, Potash JB, Weissman MM, McCormick C, Haudenschild CD, Beckman KB, Shi J, Mei R, Urban AE, Montgomery SB, Levinson DF, Koller D. Characterizing the genetic basis of transcriptome diversity through RNA-sequencing of 922 individuals. *Genome Research* **1** (2014), 14–24.

[20] Lappalainen T, Sammeth M, Friedlnder MR, 't Hoen PA, Monlong J, Rivas MA, Gonzlez-Porta M, Kurbatova N, Griebel T, Ferreira PG, Barann M, Wieland T, Greger L, van Iterson M, Almlf J, Ribeca P, Pulyakhina I, Esser D, Giger T, Tikhonov A, Sultan M, Bertier G, MacArthur DG, Lek M, Lizano E, Buermans HP, Padioleau I, Schwarzmayr T, Karlberg O, Ongen H, Kilpinen H, Beltran S, Gut M, Kahlem K, Amstislavskiy V, Stegle O, Pirinen M, Montgomery SB, Donnelly P, McCarthy MI, Flicek P, Strom TM; Geuvadis Consortium, Lehrach H, Schreiber S, Sudbrak R, Carracedo A, Antonarakis SE, Hsler R, Syvnen AC, van Ommen GJ, Brazma A, Meitinger T, Rosenstiel P, Guig R, Gut IG, Estivill X, Dermitzakis ET. Transcriptome and genome sequencing uncovers functional variation in humans. *Nature* **501** (2013), 506–511.