New Statistical Methods for Family-Based Sequencing Studies

Alexandre Bureau (Université Laval),
Kelly Burkett (University of Ottawa),
Jinko Graham (Simon Fraser University),
Ingo Ruczinski (Johns Hopkins University

August 5-10, 2018

1 Overview of the Field

After the era of genome-wide association studies (GWAs) high-throughput DNA sequencing studies became fundamental to isolate the exact causes and contributors of human disease, and to tailor medical treatment to the individual characteristics of each patient ("precision medicine"). While the vast majority of effect sizes for common single nucleotide polymorphisms (SNPs) reported in the GWAS catalogue are small, we now know that much of the heritability for many traits including complex disorders can be explained by rare but highly penetrant variants. These variants have the potential to be used as "genetic biomarkers" in practice, for example to predict disease risk or response to a specific therapy. To detect these rare but highly penetrant variants, family-based study designs are much better suited than population based designs as they provide a way to test co-segregation with disease of variants that are too rare in the population to be tested individually in a conventional case-control study. However, many mathematical and statistical challenges remain in approaches to fully exploit the information in these family based designs. One of the particularly difficult but exciting challenges for bio-mathematicians and statistical geneticists is to account for various types of dependence structures in familial DNA sequencing data, and to develop optimal approaches (in terms of power and scalability) for these time-consuming and expensive studies. Here, dependency can be caused by relatedness among individuals unknown to the investigators, correlation of nearby genetic markers, and/or pleiotropy (i.e. genetic variants affecting multiple traits). This workshop brought together some of the world's leading experts in this field to address these critical and timely issues.

2 Presentation Highlights

2.1 Rare variant sharing

Sequencing DNA in extended multiplex families can help to identify high penetrance disease variants too rare in the population to be detected through tests of association in population based studies, but co-segregating with disease in families. **Alexandre Bureau** and **Ingo Ruczinski** presented a statistical framework based on this paradigm to exploit sequencing data from extended multiplex families [2, 4]. Specifically, when only few affected subjects per family are sequenced, evidence that a rare variant may be causal can be quantified from the probability of sharing alleles by all affected relatives given it was seen in any one family member under the null hypothesis of complete absence of linkage and association. **Ingo Ruczinski** presented the general RVS (rare variant sharing) framework for calculating such sharing probabilities when two or more

affected subjects per family are sequenced, showed how information from multiple families can be combined by calculating a p-value as the sum of the probabilities of sharing events as (or more) extreme, and introduced the concept of "potential p-values" to alleviate the burden due to multiple comparisons. He discussed the importance of the rare variant assumption, and the power of the approach. **Alexandre Bureau** later discussed the effect of cryptic relatedness among family founders on the inference, and presented solutions to address the independence violation (see 2.2). The usefulness of this approach was highlighted in a case study from families with multiple members born with oral clefts, interrogating the sharing patterns of nucleotide and structural variants [3, 6].

Dandi Qiao presented an alternative approach, the gene-based segregation test (GESE) [9]. In contrast to RVS, GESE requires an estimate of variant frequencies to calculate an unconditional probability of segregation patterns (as compared to calculating the probability of sharing conditional on the variant being observed), but otherwise relies on very similar assumptions as RVS. Specifically, GESE (like RVS) assumes only one founder in the family introduced a causal variant in a gene, and limiting the tests to variants with high functional impact is recommended. Further, GESE also calculates the p-value as the sum of the probabilities of all events as or less likely as the observed event. GESE uses the sequence data from affected and unaffected family members, while RVS is based only on sharing among affected subjects, and does not make any assumptions about unaffected subjects.

2.2 Exploiting genealogical databases and isolated populations

Genealogical databases have been developed from a variety of data sources. Databases created from systematic civil registers in founder populations contain nearly the complete history of the population and can serve to answer a number of questions about the distribution and history of genetic variants in the population. The largest such database in Canada is the BALSAC database of the Quebec founder population (balsac.uqac.ca). **Simon Girard** gave an overview of this database hosted at Université du Québec à Chicoutimi and containing over 3 million linked records from the Catholic church.

For rare genetic conditions due to a genetic variant likely introduced by a single population founder, **Simon Gravel** presented an approach to estimate the variant genotype posterior distribution over the population founders and, from that founder distribution, estimate the variant frequency in the various regions of Quebec. Assuming the variant entered the population through a unique founder and the genealogy is correct, the variant genotype founder distribution is inferred by "Monte Carlo climbing simulations", where more promising paths are favored by an importance sampling scheme. The resulting bias in likelihood computation is then corrected. The method implemented in ISGen (github.com/DomNelson/ISGen) was applied to Chronic Atrial et Intestinal Dysrhythmia (CAID), an autosomal recessive condition caused by a variant in gene SGOL1 present in Europe with frequency 0.0002. The distribution of the ancestor most likely to have introduced the variant was inferred based on 11 patients tied to the genealogy, and is concentrated on 4 founders. The allele frequencies inferred in the various regions of Quebec matches estimates from population samples. The genealogical approach predicted high frequencies of the causal variant in regions where the available population samples were too small for reliable estimation (e.g. the Charlevoix region), highlighting its potential for orienting mutation screening.

The problem addressed by **Alexandre Bureau** within the rare variant sharing framework for identification of rare disease-susceptibility variants (see 2.1), was conceptually different but similar in approach. It consisted in estimating the null distribution of rare variant sharing events among present-day affected subjects included in a sequencing study, given a rare variant was seen in any of them, considering the genealogy of these subjects as a single extended pedigree. Solving this problem would have two benefits: controlling for cryptic relatedness (i.e. relatedness that is not captured in typical pedigree structures extending over 3-5 generations) and improving the power of the approach. Like the variant frequency estimation approach of **Gravel**, the present approach has two steps: 1) sampling the distribution of the founder introducing the variant conditional on the genotype of present-day subjects and 2) sampling the transmission of the variant from that founder to the current generation. Step 1) is here much simpler because it conditions on the genotype of a single subject instead of a set of subjects: thus the sampling of the parent transmitting the variant to

his offspring is a simple coin toss performed recursively from a present-day variant carrier to a genealogy founder, while step 2) is the same as for frequency estimation, but applies to a more restricted set of subjects. Estimates of the sharing probabilities given *any* affected subject carries the variant (instead of a specific one) and their associated standard error are obtained by combining results for all present-day affected subjects. This method was attempted on datasets simulated using the 56,800-member genealogy of 217 families from the SaguenayLac-St-Jean (SLSJ) region of Quebec comprising 1018 individuals enrolled in an actual study of asthma, extracted from the BALSAC database. The current implementation is however excessively slow when the number of affected subjects exceeds about 10.

A contrasting example of a smaller and older population isolate without genealogical database was provided by **Arthur Gilly** with the Minoean isolates of the Greek island of Crete. With no genealogical database available, standard rare variant aggregation methods to test association with quantitative and dichotomous traits have been applied to genotyping and sequencing data on 1500 individuals from this population, while controlling for relatedness estimated from the genotype data. This analysis in this special population revealed a burden of rare variants in gene FAM189B driven by isolate-specific rare variants.

2.3 Gene genealogies

For understanding genetic associations with trait data, it can be useful to model the latent ancestries that give rise to the sample's genetic variability. The *gene genealogy* is a graph that describes the ancestry or lines of descent of chromosomal segments sampled from the general population. The neutral coalescent provides a mathematical model for the topology and node times of the graph. Several speakers spoke about their work in incorporating the gene genealogy in mapping methodology, including for the discovery of rare variants and extending the modeling to pedigrees.

Jinko Graham presented her work in developing ancestral tree based statistics to find trait-influencing mutations. The ancestry-based approach is motivated by a similar principle as family-based mapping: the similarity in relationships between haplotypes is reflected in similarity between trait values. This suggests examining statistics that capture the degree to which haplotypes from individuals with similar trait values cluster together in the ancestral tree. She discussed an application in which Pearson correlation between tree-defined clusters and disease status was used as a measure of association. This was applied to a dataset consisting of genotype data for diabetes cases and healthy controls. Trees were sampled from their posterior distribution conditional on imputed haplotype data at 55 focal points in a genomic region. The fuzzy pvalue was used to assess the strength of the association. The results highlighted a small number of adjacent focal points where the median of the fuzzy pvalue was low and the spread of the fuzzy pvalue distribution was small, indicating more certainty about the genealogical tree given the data.

Both **Kelly Burkett** and **Renaud Alie** discussed their progress in extending population-based models for the gene genealogy to family data. Kelly Burkett presented her work in extending a Markov chain Monte Carlo approach to sample an approximation of the gene genealogy conditional on genotype data from trios (mother, father, child). Each parents' two unobserved haplotypes correspond to two tips of the ancestral tree and the child's genotypic data restricts the set of parental haplotype configurations. A proposal distribution is then used to update the parental haplotypes while ensuring the child's genotypes remain the same. She applied the sampler to a previously-analysed Crohn's disease dataset and compared results using the trio-based sampler to sampling conditional on imputed parental haplotypes. Results were similar between the two approaches; however, the trio-based sampler showed signs of poor convergence. Renaud Alie spoke about his work developing the pedigree-based coalescent. He proposed using previously developed models for the inheritance of genetic material within the pedigrees at the tips and using the ancestral recombination graph as the model for the relationships between the pedigree founders.

Tree-based association statistics capture the degree to which sequences from disease-affected individuals cluster in the ancestral tree. However, each individual corresponds to two tips of the ancestral tree - one for each of their sequences. Depending on the disease model, it's possible that only one of these sequences actually carries a risk variant. Therefore, a sequence from a case could be misclassified. **Charith Bhagya**

Karunarathna described his work comparing the performance of multiple tree and non-tree based association methods in localizing a risk variant. In particular, he compared two versions of the tree-based Mantel test: a naive version where both sequences are classified as a case and an informed version where only the sequence carrying the risk variant was classified as a case. He showed that there was a substantially worse performance for the naive Mantel test relative to all other approaches. The informed Mantel test performed well, but it is an idealized test that cannot be implemented in practice without knowledge of the risk variant. This highlights a challenge to be addressed when developing tree-based association statistics.

2.4 Simulation software to test rare variant methods for family data

Simulated genetic data is frequently required for testing new rare variant methods or to compare existing methods. Although many programs for simulating genetic data have been created, they often rely on mathematical models from population genetics that may not be biologically plausible. In addition, programs typically simulate data from unrelated individuals; to generate family data, additional programming or separate software is needed to probabilistically "drop genes" in known family structures. As mentioned in Section 2.2, public databases contain real genealogical (BALSAC) or genomic data (e.g. 1000 Genomes). These databases are valuable resources for simulating realistic datasets or evaluating algorithms based on probabilistic models. Many of the participants presented simulation software developed by themselves or their colleagues that utilicize these public databases to model the reference population for their simulations.

A number of speakers (e.g. **Briollais, Bull, Choi**) used `Sim1000G` as part of the research that they presented. `Sim1000G`, co-developed by **Laurent Briollais**, can be used to simulate genetic data using a reference population stored in a VCF file that is supplied by the user. In particular, the reference population can be phased data from the 1000 Genomes project. Genetic data for new individuals is generated by sampling from this reference population so that the allele frequencies and linkage disequilibrium patterns are maintained. Unrelated individuals and pedigrees of arbitrary size can be simulated. For pedigree data, two recombination models are available.

Christina Nieuwoudt presented two R packages that she and **Jinko Graham** developed for simulating genetic data on pedigrees ascertained to contain a minimum number of affected individuals. `simRVpedigree` [8] is used to simulate the pedigrees and disease status. The disease model for affected pedigree members includes sporadic cases as well as cases caused by a rare variant segregating in the pedigree. `simRVsequences` can then be used to simulate genetic data for the pedigree members using a model inspired by gene dropping and including recombination events. The sequence data in the founders can be obtained using publicly-available datasets.

Simon Gravel described how a coalescent-based simulation algorithm, `msprime`, failed to capture the genomic structure seen with real data. The distributions of both the length and number of IBD segments between individuals with different familial relationships were quite different between data from "23andme" and the simulated data. Although coalescent-based simulators might reasonably approximate real data over smaller genomic regions, the approximation is poor when simulating large regions since it allows more than two parents per offspring. He is working with collaborators on a simulator, `hybride`, that is based on the forward-in-time Wright-Fisher model rather than the coalescent. By incorporating recent improvements to the representation of gene genealogies used in `msprime` [7], computation time for `hybride` is remarkably fast. Though they are still working on improvements, the software will be available for download in the future.

2.5 Variance-components models and identity by descent

Variance-components models are a useful tool for mapping quantitative phenotypes in families. Phenotypes are typically assumed to follow a Gaussian distribution, with a mean that may depend on alleles at a genetic marker of known or hypothesized influence. Overall phenotypic variance within a family is decomposed into *genetic* and *non-genetic* components. In the linear model of the mean phenotype, the genetic component is

characterized by a vector of family-specific random effects. The variance-covariance matrix of these family-specific random effects is the genetic component of variance. To assess the evidence for genetic association between the phenotype and alleles at a given genetic marker, the coefficients for fixed allelic effects are tested in the linear model. To assess the evidence for genetic linkage between the phenotype and a particular genomic location, the variance of the random effects is further decomposed into genome-wide and location-specific components. The genome-wide genetic-variance matrix is written as a scalar "polygenic" variance times a kinship matrix summarizing the pairwise relationships in the pedigree. The location-specific variance matrix is expressed as a scalar, location-specific variance times a matrix of pairwise identity-by-descent (IBD) proportions estimated from genetic markers near the location of interest. Linkage corresponds to the scalar location-specific variance being non-zero. Presentations in the workshop extended and/or applied this variance-components model in novel ways. We heard about extensions to accommodate non-Gaussian (e.g., time-to-event or binary) phenotypes, and heterogeneous variance components across families and across ethnic groups. We also heard about methods for estimating kinship matrices and local IBD sharing. The estimates from these methods allow for linkage analysis without the need to know the pedigree relationships between individuals.

Yun-Hee Choi and **JC Loredo-Osti** presented variance-components models for linkage and association analysis of time-to-event and binary phenotypes, respectively. Their generalized linear models for the mean phenotype specify a linear predictor comprised of fixed effects and a vector of family-specific random effects, as well as a link function that relates the linear predictor to the mean. In classic variance-components style, the covariance between family-specific random effects is decomposed into genome-wide and location-specific components. Fixed effects in the mixed model can be viewed as association parameters. By contrast, the location-specific scalar variance in the variance-component model is a linkage parameter. In particular, large scalar variances indicate that the locus-specific contribution to the genetic random effect varies according to the degree of local relationship (i.e. linkage). Choi's model allows for ascertainment of the families through a proband. The likelihood for randomly-sampled (i.e. unascertained) families is corrected by conditioning or adding a penalty term. Issues such as model identifiability and the appropriate null distribution of test statistics for linkage and/or association were discussed by both speakers. Loredo-Osti commented that unrelated individuals may be incorporated as "families" of size one.

Laura Almasy presented methods that allow for genetic variance components that can vary across families. Her motivation for developing these methods was the family-based Collaborative Study on the Genetics of Alcoholism. So far, linkage-mapping studies of alcoholism and its endophenotypes have identified only broad genomic regions of interest. The rationale for her proposed extension is that more detailed modelling of the genetic variance should improve the linkage resolution and provide insight into families segregating different causal loci. Almasy lets the proportion of total variance attributable to the genome-wide and location-specific genetic components vary by family, while holding constant across all families the total variance. In the linear model, fixed effects are also assumed to be the same across families. Examining the resulting family-specific lod scores for linkage leads to insight into which families are contributory to a given linkage peak. Focusing analysis on these contributory families leads to insight into which genes under a linkage peak are contributory, and therefore improved resolution. The results of her linkage analysis highlight the complex nature of alcoholism. Several promising genes and variants were prioritized for further investigation.

Tim Thornton described a variance-components association analysis for two large consortia comprised of multiple ethnic groups. The data for these consortia were collected from different study designs involving families, founder populations and case-control samples. He used variance-components models to account for correlation between related study subjects. The multi-ethnic nature of the subjects leads to a mix of recent relatedness through family structure and distant relatedness through population structure. To account for population structure, he included fixed effects for genetic principal components in the linear model for the mean response. The new approach improves upon competing methods that under-compensate for population stratification at highly-differentiated markers, and over-compensate for population stratification at weakly-differentiated markers. He found that, in multiethnic samples, allowing for the non-genetic variance to be different for different ethnic groups is crucial for unbiased tests of association.

The Workshop included several case studies involving variance-component models and/or IBD. These case studies highlighted the challenges, rewards and insights offered by analysis of sequencing data. For example, **Mariza de Andrade** described whole-genome sequencing of a family with venous thromboembolism and insights gained from linkage and association analysis. **Heather Cordell** described valuable lessons learned from sequencing under a linkage peak in an ongoing study of families with vesico-ureteric reflux. The process of analysis uncovered some pitfalls of working with sequencing data, including variant calls that are highly dependent on the platforms used for sequencing and on the bioinformatics pipeline. **Janet Sinsheimer** discussed a sequencing study of the microbiome of a beetle that lives in family units. The aim of the study was to estimate the heritability of the microbiome. Her results suggested heritability of important bacterial groups. **Simon Girard** used pairwise IBD sharing in the Quebec founder population to identify genetic variants associated with epilepsy.

Elizabeth Thompson discussed methods for estimating kinship matrices and local IBD sharing. The resulting estimates allow for linkage analysis without knowing the pedigree relationships between individuals. Thompson first presented methods for local IBD estimation. She then described how to combine these local estimates to obtain genome-wide measures of kinship. When these estimates are used in linkage analyses of simulated datasets, the likelihood-ratio (lod) curves correctly identify the linkage regions. **Shelley Bull** described an application of inferred local IBD to mapping risk genes for breast cancer using data on affected sisters. The basic premise is that the siblings share disease because of the genomic regions they share IBD, and in particular the susceptibility variants within those regions. If so, more susceptibility variants are expected on haplotypes shared IBD by the sisters than on haplotypes that are not shared IBD. These idea leads to a statistical test of association in terms of inferred IBD and the number of rare variants in a genomic region for the sibship.

2.6 Accommodating biased sampling of families

Family studies of genetic diseases typically sample families having one or more affected members. The first affected family member to be included in the study is called the proband, and the set of individuals used to determine family eligibility for the study is called the ascertainment set. As ascertained families do not represent a random sample from the population, statistical methods must account for the biased sampling to avoid biased inference. If we view families as independent, we can understand the ascertainment issues in terms of the likelihood for a single family. Let A be the event that the family is ascertained and Y be the phenotypes (e.g., disease status, or age-at-onset of disease). In heritability studies that do not collect genetic data, the likelihood is $P(Y|A)$. In studies that collect genetic data, G, the likelihood is $P(Y, G|A)$.

In studies that collect genetic data, alternatives to the full likelihood, $P(Y, G|A)$, may be obtained by further conditioning. The *prospective* likelihood is based on $P(Y|G, A)$ and the *retrospective* likelihood on $P(G|Y, A)$. As a rule, conditioning ignores information and can lead to a loss of efficiency in statistical inference. However, approaches based on conditional likelihoods may be easier to implement than approaches based on the full likelihood. For example, if ascertainment depends only on Y, then $P(A|Y, G) = P(A|Y)$. We can then argue that the retrospective likelihood is $P(G|Y, A) = P(G|Y)$, so that the ascertainment doesn't matter. Prospective likelihoods can also be easier to implement under certain assumptions. For example, *complete ascertainment* occurs when every eligible family in the population is ascertained into a study. Under complete ascertainment, $P(A|Y, G) = 1$ for every (Y, G) that meets the ascertainment criteria. Choi *et al.* (2008) show that the prospective likelihood is $P(Y|G, A) = P(Y|G)/P(A|G)$, which is a penalized prospective likelihood with penalty term $1/P(A|G)$. This simplification allows ascertainment-adjusted methods to be developed as penalized versions of existing prospective methods.

The workshop featured several talks that included the concept of biased sampling of families and ascertainment adjustment. **Lajmi Lakhal Chaieb** presented work investigating the heritability of psoriatic arthritis and its possible dependence on parent of origin. Families were ascertained through a proband only. The phenotype was age-at-onset of arthritis and no genetic data were collected; hence the likelihood is of the form $P(Y|A)$. The ascertainment event A is that the proband develops disease by the time the family is recruited. **Jooyoung Lee** presented methods to estimate risk parameters from time-to-disease phenotypes

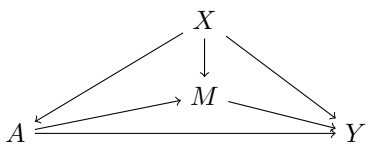
in studies collecting genetic data. The ascertainment event A is that the proband is alive and affected with the disease *and* has multiple relatives alive at the time of recruitment into the study. She described how the complexity of the full likelihood, $P(Y, G|A)$, increases rapidly with the family size. To make the problem tractable, she proposed an approximation to $P(Y, G|A)$ that involves the proband and a *pair* of non-probands from the ascertainment set. Other speakers took retrospective approaches to the analysis of data from their family studies. For example, **Alexandre Bureau** and **Ingo Ruczinski** presented an approach to test for co-segregation of a rare variant with disease. The approach is based on the sharing probabilities, $P(G|Y, A)$, under the null hypothesis of no co-segregation. The ascertainment event A is that there is at least one affected family member and that the rare variant is present in at least one of them. In contrast, **Yun-Hee Choi** took a prospective approach to analysis of her family data. Under complete ascertainment, she adjusts the prospective likelihood by a penalty term to correct for the biased sampling of families. She also assumes that families are recruited through a proband only. The ascertainment event A is that the proband develops disease by the time of recruitment into the study.

Under complex ascertainment, simplifying assumptions for tractable calculations may not be possible. We may then have to resort to Monte-Carlo sampling of ascertained families. Families are randomly sampled from the population and then filtered by the ascertainment criteria. Along these lines, **Christina Nieuwoudt** described her `simRVSequences` R package. This software simulates whole-exome sequence data in families segregating a rare causal variant and allows for different schemes of family ascertainment.

2.7 Causal inference using genetic and epigenetic measurements

Inference of causal effects of exposures on health outcomes from observational data gives rise to a large variety of difficult problems. Two contributors to this workshop considered the setting where an exposure A may have a causal effect on a continuous health outcome Y through a mediator variable M, but some of these causal effects may be confounded by a set of factors X [1] (see Figure 2.7). Both **Karim Oualkacha** and **Xihong Lin** considered studies where DNA methylation at a given genomic site was the mediator variable M, but the problems studied by these two speakers differed. In **Oualkacha**'s talk, the interest lied in establishing the causal effect of M on Y, and A was genotype at a set of genetic markers, to be used as instrumental variables in a Mendelian randomization analysis. Since genotype is fixed at birth, the effect of genotype cannot be confounded by factors acting after birth, so it is assumed that there is no effect of X on A. Mendelian randomization also requires that the genetic markers A selected as instruments have no direct effect on the outcome Y. When these conditions are met, detecting an association between A and Y is evidence of a causal effect of M on Y. With high-dimensional genotype data from genome-wide arrays, selection of the markers to use as instruments A can be performed by penalized least square regression methods such as the Lasso. The difficulty encountered by Oualkacha was that study subjects were grouped in families. A two-step procedure had previously been proposed: 1) use linear mixed models to adjust for relatedness of the subjects and 2) use the residuals from step 1 in variable-selection least-square regression methods. Oualkacha introduced `ggmix`, a two-in-one procedure which controls for relatedness and performs variable selection under linear mixed models. The optimization problem was solved using a block relaxation technique. Initial simulations studies focused on the performance of the approach to select the right set of markers and correctly estimate heritability of M and revealed that `ggmix` is a promising alternative to the two-step approach and the naive Lasso ignoring familial relatedness.

Figure 1: Causal diagram between an exposure A, a mediator M, confounding factors X and an outcome Y



Lin was interested in performing classical mediation analysis in unrelated subjects with an exposure A (e.g. smoking) that may affect DNA methylation M at multiple genomic sites, and thus may have a natural

indirect effect (NIE) on the outcome Y through M, in addition to a possible direct effect. Earlier work had showed that tests of the NIE were conservative under the null case where there is no association between A and M and no association between M and Y, which is likely to be the case for most methylation sites in a genome-wide analysis[1]. She proposed the divide-aggregate or DAT method to correct conservativeness of NIE tests under standard mediation approaches.

3 Outcome of the Meeting

The workshop was an opportunity for exchanges leading to new ideas to solve problems faced by participants. For instance, Alexandre Bureau, Ingo Ruczinski and Simon Gravel determined that the software developed by the Gravel group to sample the transmission of variants from founders of a genealogy to present-day individuals could speed-up the same step in the estimation of the null distribution of rare variant sharing statistics in extended families extracted from a genealogical database attempted by Bureau. Also, the suggestion made by Ken Lange at the workshop to subdivide the genealogy into lineages under each founder was later implemented by Bureau in the `RVS` Bioconductor package and enabled exact computations in larger families than had been possible previously.

The workshop gave speakers an opportunity to demonstrate the use of their software in assessing methods for rare variant discovery with family data. As the software described above is freely available, other participants can now use these programs for their own research. For example, `sim1000G` is currently being used by Kelly Burkett's research group for simulating genetic pathway data on trios (mother, father, child). Because it uses real human genetic data, they are able to easily simulate data for human genes in the pathways of interest.

References

[1] R Barfield, J Shen, AC Just, PS Vokonas, J Schwartz, AA Baccarelli, TJ VanderWeele, X Lin. Testing for the indirect effect under the null for genome-wide mediation analyses *Genetic Epidemiology* **41** (2017), 824–833.

[2] A Bureau, SG Younkin, MM Parker, JE Bailey-Wilson, ML Marazita, JC Murray, E Mangold, H Albacha-Hejazi, TH Beaty, I Ruczinski. Inferring rare disease risk variants based on exact probabilities of sharing by multiple affected relatives. *Bioinformatics* **30** (2014), 2189–2196.

[3] A Bureau, MM Parker, I Ruczinski, MA Taub, ML Marazita, JC Murray, E Mangold, MM Noethen, KU Ludwig, JB Hetmanski, JE Bailey-Wilson, CD Cropp, Q Li, S Szymczak, H Albacha-Hejazi, K Alqosayer, LL Field, YH Wu-Chou, KF Doheny, H Ling, AF Scott, TH Beaty. Whole exome sequencing of distant relatives in multiplex families implicates rare variants in candidate genes for oral clefts. *Genetics* **197**, (2014) 1039–1044.

[4] A Bureau, F Begum, MA Taub, JB Hetmanski, MM Parker, H Albacha-Hejazi, AF Scott, JC Murray, ML Marazita, JE Bailey-Wilson, TH Beaty, I Ruczinski. Inferring disease risk genes from sequencing data in multiplex pedigrees through sharing of rare variants. *Genetic Epidemiology* (2018, to appear).

[5] Y-H Choi, K Kopciuk and L Briollolais. Estimating disease risk associated with mutated genes in family-based designs. *Human Heredity* **66** (2008), 238–251.

[6] J Fu, TH Beaty, AF Scott, J Hetmanski, MM Parker, JE Wilson, ML Marazita, E Mangold, H Albacha-Hejazi, JC Murray, A Bureau, J Carey, S Cristiano, I Ruczinski, RB Scharpf RB. Whole exome association of rare deletions in multiplex oral cleft families. *Genetic Epidemiology* **41** (2017), 61–69.

[7] J. Kelleher, A.M. Etheridge and G. McVean, Efficient Coalescent Simulation and Genealogical Analysis for Large Sample Sizes. *PLoS Computational Biology*, **12** (2016), e1004842.

[8] C. Nieuwoudt, S.J. Jones, A. Brooks-Wilson and J. Graham, Simulating pedigrees ascertained for multiple disease-affected relatives. *Source Code for Biology and Medicine*, **13** (2018), 2.

[9] D Qiao, C Lange, NM Laird, S Won, CP Hersh, J Morrow, BD Hobbs, SM Lutz, I Ruczinski, TH Beaty, EK Silverman, MH Cho MH. Gene-based segregation method for identifying rare variants in family-based sequencing studies. *Genetic Epidemiology* **41** (2017) 309–319.