ONETOOL for the analysis of family-based big data – Supplementary material

Yeunjoo E. Song[1‡], Sungyoung Lee[2‡], Kyungtaek Park[2], Robert C. Elston[1], Hyeon-Jong Yang[3,4*] and Sungho Won[2,5,6*]

[1] Department of Population and Quantitative Health Sciences, Case Western Reserve University, Cleveland, OH, USA.
[2] Interdisciplinary Program of Bioinformatics, Seoul National University, Seoul, South Korea
[3] SCH Biomedical Informatics Research Unit, Soonchunhyang University Seoul Hospital, Seoul, Korea
[4] Pediatric Allergy and Respiratory Center, Department of Pediatrics, Soonchunhyang University Seoul Hosplital, Soonchunhyang University College of Medicine, Seoul, Korea
[5] Department of Public Health Science, Graduate School of Public Health, Seoul National University, Seoul, Korea
[6] Institute of Health and Environment, Seoul National University, Seoul, South Korea

[‡] Contributed equally

[*]Correspondence:

Sungho Won, Ph.D.
Department of Public Health Science
Graduate School of Public Health
Seoul National University
Seoul, Korea
sunghow@gmail.com

Hyeon Jong Yang, MD, Ph.D.
Department of Pediatrics
Soonchunhyang University Seoul Hospital
Soonchunhyang University College of Medicine
Seoul, Korea
pedyang@schmc.ac.kr

Table 1. Available tools for family-based sequence data analysis.

Name	1	2	3	4	5	6	7	8	9	10	11	12	Reference
Olorin		√		√									Morris et al. (2012)
VAR-MD		√											Sincan et al. (2012)
PEDCMC									√				Zhu and Xiong (2012)
Mendel	√	√	√			√		√	√				Lange et al. (2013)
famBT/famSKAT									√				Chen et al. (2013)
FB-SKAT									√				Ionita-Laza et al. (2013)
PEDGENE									√				Schaid et al. (2013)
FARVAT									√				Choi et al. (2014)
MendelScan		√				√	√						Koboldt et al. (2014)
FamAnn		√											Yao et al. (2014)
pVAAST		√						√	√				Hu et al. (2014)
rvTDT									√				Jiang et al. (2014)
RarePedSim	√												Li et al. (2015)
PBAP		√	√										Nato et al. (2015)
F-SKAT									√				Yan et al. (2015)
SEQLINAKGE								√					Wang et al. (2015)
FamPipe		√					√						Chung et al. (2016)
FCVPP		√				√							Forsti et al. (2016)
RVTESTS		√							√	√			Zhan et al. (2016)
RV-GDT/RV-PDT									√				He et al. (2017)
Merlin											√		Burdick et al. (2006)
GIGI											√		Cheung et al. (2013)
PedBLIMP											√		Chen et al. (2014)
PRIMAL											√		Livine et al. (2015)
GIGI-Quick											√		Kunji et al. (2018)

1: design/simulation, **2**: variant QC/filtering/ranking/annotation, **3**: pedigree description/summary, **4**: pedigree plot, **5**: familial aggregation, **6**: segregation, **7**: IBD mapping, **8**: linkage, **9**: association (genotype), **10**: meta-analysis, **11**: family-based imputation, **12**: association (dosage)

Table 2. Analyses available in ONETOOL.

Main	Sub-category	Detail	Reference and software
InfoQC analysis	Variant Information	F_{ST}, Ts/Tv ratio, MAF, HWE, PCA	
	Sample Information	Het, Het/Hom	
	Pedigree Information	Description and summary, plot, relative pairs	S.A.G.E. (2016) – PEDINFO Sinnwell et al. (2014) – kinship2 Song and Elston (2013) – PEDWIZ
	Error Detection	Mendelian error	
	Relatedness matrix	Kinship, IBS, GRM	Balding and Nichols (1995)
Trait Analysis	Familial Aggregation	Correlation	S.A.G.E. (2016) - FCOR
	Heritability	Based on Kinship, IBS, GRM	
	Segregation Analysis	Mode of inheritance	S.A.G.E. (2016) - SEGREG
Linkage Analysis	Model-based	Two-point, utilizing segregation analysis	S.A.G.E. (2016) - LODLINK
	Model-free	Multipoint, modeling LD	Abecasis et al (2002) - MERLIN
Association Analysis	Single variant	Generalized Score test, regression$^+$, Fisher's exact test$^+$	
		Transmission disequilibrium test – TDT, SDT	Spielman et al. (1993) Spielman and Ewens (1998)
		Likelihood ratio test – MQLS, FQLS, E(xtendedF)QLS	Thornton and McPeek (2007) Park et al. (2015) Won et al. (2015)
		Linear mixed model method - GEMMA	Zhou and Stephens (2012)
	Gene-based	CMC - no weight (default) Collapsing	Li and Leal (2008) Morris and Zeggini (2010)
		Burden test – PEDCMC, wSum$^+$, aSum$^+$	Madsen and Browning (2009) Han and Pan (2010) Zhu and Xiong (2012)
		Variable threshold method – famVT, VT$^+$	Price et al. (2010)
		Kernel method – FARVAT, KBAC$^+$, SKAT$^+$, SKATO$^+$	Liu and Leal (2010) Wu et al. (2011) Lee et al. (2012) Choi et al. (2014) Wang et al. (2016) Choi et al. (2016)
		Burden & Kernel - PEDGENE	Schaid etal. (2013)
	Epistasis$^+$	MDR$^+$, GMDR$^+$	

F_{ST}: fixation index, **Ts/Tv**: transition and transversion ratio, **MAF**: minor allele frequency, **HWE**: Hardy–Weinberg Equilibrium, **PCA**: principle component analysis, **Het/Hom**: heterozygote and homozygote ratio, **IBS**: identity by state, **GRM**: genetic relation matrix, **LD**: linkage disequilibrium

Note:

1. Both binary and quantitative variables can be analyzed in trait analysis and linkage analysis, both as main traits and as covariates. For heritability estimation of binary variable, ONETOOL first estimates the heritability by assuming the binary trait is a quantitative trait and then the heritability of its liability is estimated on the logistic scale (Lee et al., 2011). Therefore, the estimated heritability of a dichotomized existing quantitative variable should be understood as being measured on the logistic scale.
2. The additional association analysis methods available for the independent samples are marked with $^+$.

Table 3. The variable type and covariate support in association analyses in ONETOOL.

		Trait type		Covariate	Family data structure	Note
		binary	continuous			
Single variant analysis (suitable for common SNPs)	Score test*	N	Y	Y	general pedigree	usually efficient for randomly selected samples
	TDT	Y	N	N	trio	parental genotype need to be known but not used
	SDT	Y	N	N	nuclear family	need the genotype data of unaffected sibs
	MQLS	Y	N	N	general pedigree	efficient for ascertained families
	FQLS	Y	Y	N	general pedigree	efficient for ascertained families
	GEMMA*	N	Y	Y	general pedigree	usually efficient for randomly selected samples
	EQLS*	Y	Y	Y	general pedigree	efficient for ascertained families
Gene-based analysis (suitable for rare variants)	CMC*	Y	Y	N	general pedigree	efficient when effects of rare variants are homogeneous
	PEDCMC*	Y	N	N	general pedigree	efficient when effects of rare variants are homogeneous
	FAMVT*	N	Y	N	general pedigree	efficient if rarer variants have stronger effect on disease
	FARVAT*	Y	Y	Y	general pedigree	robust to the heterogeneity of effects of rare variants
	PEDGENE*	Y	Y	Y	general pedigree	conditioning on phenotypes, treating the genotype data random, for pedigrees sampled because of multiple affected members
	FBSKAT	Y	N	N	general pedigree	efficient if rare variants with both positive and negative effect on disease are grouped to a single set
	RVTDT	Y	N	N	trio	efficient if rare variants with both positive and negative effect on disease are grouped to a single set

Note:

The association analysis methods marked with * can utilize the dosage data.

Table 4. Recommended association analysis method for different types of data.

Trait Type		Heritability	
		Small or zero (<0.3)	Large(>0.3)
Continuous trait		• Random sample - Logistic regression with/without PC scores depending on the presence of population structure • Family-based sample - GEMMA	• Random sample - GEMMA • Family-based sample - GEMMA
Prevalence of binary trait	Small	• Random sample - Logistic regression with/without PC scores depending on the population substructures • Family-based sample - FARVAT/MQLS/FQLS	• ascertained family / trio - TDT/SDT • ascertained family / general pedigree - FARVAT/MQLS/FQLS
	Large	• Independent sample - survival analysis / age-of-onset analysis	• Ascertained family / trio - Age-of-onset analysis

Text 1. Input and Output

ONETOOL supports two different sets of input files, a PLINK set and a VCF set. The PLINK set consists of three files (i.e., .fam, .bed, and .bim) that are used to run PLINK, and the VCF set consist of a plink format family file (.fam) and a Variant Call Format (.vcf). The additional phenotypes and covariates are supported through an optional input file (.pheno) for both sets of input files. ONETOOL also support two different ways to specify the desired analysis options, through a command line and a script file. Each method in ONETOOL outputs the result file with the appropriate extension, so that the user can recognize it easily. It has the familiar user interface and the same or similar analysis option names as the existing tools, so no, or only a minimal, learning curve is needed.

Text 2. Imputation

ONETOOL provides an option to impute the missing genotypes for typed genotypes. Expected missing genotypes for typed variants are imputed based on the familial relationship, and if phenotypes of any subjects with missing genotypes are available, genotypes imputed with family members' genotypes can improve statistical power. Let m_i and m_i' be the vector of subjects with observed genotypes and missing genotypes for the i^{th} variant, respectively. Then the observed genotypes of the i^{th} variant G_{im} can be estimated using the equation $G_{im} = \Phi_{mm'}\Phi_{m'm'}^{-1}(G_{im'} - 2f_i) + 2f_i$, where $\Phi_{mm'}$ and $\Phi_{m'm'}$ denote the relationship matrices between subjects with observed and missing genotypes, respectively, and f_i denotes minor allele frequency of the i^{th} variant.

The efficiency of the proposed method was evaluated with simulated data. We considered the pedigree that consists of 97 subjects (see Figure 1). Genotypes were selected from 1000 Genome Project and 100,000 variants were randomly selected. Then, MAFs for those 100,000 variants were calculated and founders' genotypes were randomly generated from binomial distribution under Hardy-Weinberg equilibrium. We assume there is no *de novo* mutation and non-founders' genotypes were randomly chosen with Mendelian transmission. Genotypes of all variants for 1%, 2%, 5%, 10% and 20% family members were randomly masked and then their genotypes were imputed using their relatives' genotypes. The dosage of the imputed genotypes shown in Figure 2 shows the accuracy of the imputed genotypes according to the MAF. Results show that the accuracy of imputed genotypes is up to 99% for rare variants, and imputed genotypes are reasonably accurate even with a substantial amount of missing data.

Figure 1. Pedigree used for simulations. This pedigree was randomly chosen from GAW19 data (Blangero et al., 2016).

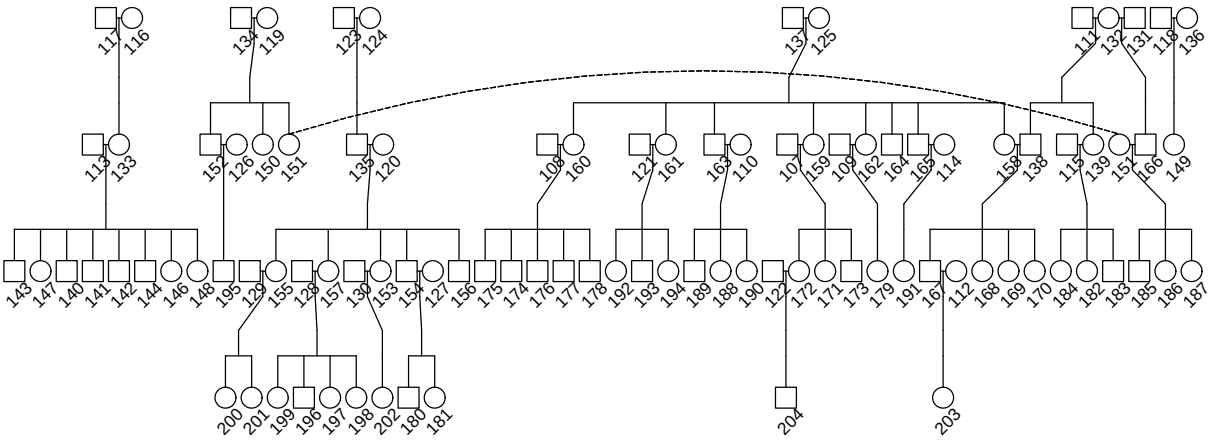
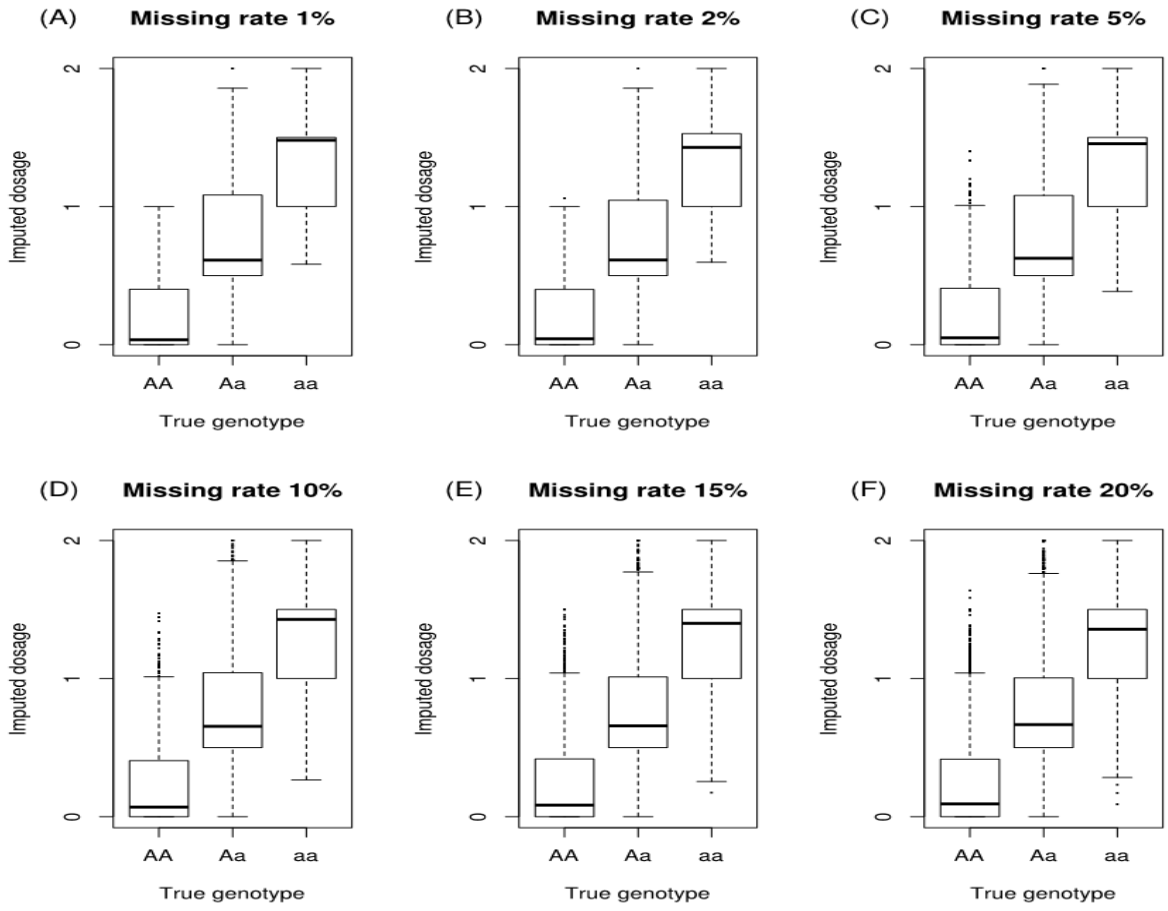


Figure 2. Boxplots of imputed dosages. Results were provided for various missing rates: (A) missing rate: 1%, (B) missing rate: 2%, (C) missing rate: 5%, (D) missing rate: 10%, (E) missing rate: 15%, and (F) missing rate:20%. In each plot, three boxplots of imputed dosage for AA (homozygote major), Aa (heterozygote), and aa (homozygote recessive) are provided.



Text 3. Association analysis with imputed genotypes

ONETOOL can take the dosage and genotype probability files from several popular imputation tools available for population data (Table 5). The list of statistical analyses which can utilize the dosage data is marked by * in Table 1 above.

Table 5. Dosage and genotype probability formats that are supported in ONETOOL.

File type	Impute toolset	Extension
Genotype probability formats	**IMPTUE2**	.impute2
	Beagle	.bgl.gprobs
	minimac3	.vcf (version 4.3)
Dosage formats	**MACH (minimac2)**	.mldose
	Beagle	.bgl.dose
	minimac3	.vcf (version 4.3)

Text 4. Performance Evaluation

ONETOOL is implemented in C++ to provide the best performance. It uses an R plugin for the pedigree plot functionality. A multi-thread option is available for various analyses with the '--thread' option. In Table 5 and 6, we show the performance of ONETOOL.

First, we compared the performance of ONETOOL to RVTESTS (Zhan et al., 2016) to evaluate speed. Though RVTESTS is not specifically designed for family-based data, it provides different association analysis methods that can accommodate both unrelated and related data, thus making a good comparison case for the performance of ONETOOL's association module.

For the dataset, we used chromosome 21 of GAW19 simulation dataset (Blangero et al., 2016). It contains 464 subjects with 191,664 variants. For the comparison of the gene-based analysis methods, we used the refFlat gene table for hg19 reference downloaded from their website. It provided the analysis of 302 genes consisting of 78,791 low-frequency variants. All analyses were conducted in a Linux workstation with four Intel E7-4850 CPUs and 1T RAM. For each method, the computation time was averaged over ten runs.

Table 6 shows the average time each analysis method took in ONETOOL and RVTESTS. ONETOOL consistently took less time to finish the analysis than RVTESTS, with up to around 200x acceleration folds, except for Balding-Nichols empirical kinship calculation.

Table 6. Performance comparison between ONETOOL and RVTESTS.

		RVTESTS	ONETOOL	Acceleration Folds
Univariate test	**Wald test**	75	33	2.27
	Fisher's Exact Test	89	22	4.05
Gene-level test	**CMC**	34	15	2.27
	Collapsing test	34	15	2.27
	SKAT	2,160	640	3.38
	SKAT-o	87,981	404	217.77
	KBAC	254	70	3.63
	VT	356	270	1.32
Relatedness computation	**GRM**	45	95	0.47
	IBS	82	42	1.95

(Unit: seconds, times)

Second, we evaluated the time to run several analyses in ONETOOL in two different family data sets. The first set (Data1) is the example data set available in our website. It consists of 10 simulated nuclear families and 100 variants. Each family contains 2 parents and 6 offsprings, so total 100 individuals. It is a complete data set, so genotyping rate is 100%. We ran ONETOOL analyzing the default binary trait included in .fam file. The second set (Data2) again is GAW19 real dataset. We analyzed a subset of families without any loops, so it contains 800 people from 12 pedigrees of size range 27 to 97. The total number of individuals is 800. We analyzed chromosome 21 again 12,842 SNPs and the genotyping rate is 66.94%. All analyses were conducted in a Linux server with four Intel Intel(R) Xeon(R) CPUs and 16G RAM. For each

method, the computation time was averaged over five runs. The run time each analysis took is shown in Table 7. Note that the last column indicates which analyses are included into the results shown in Table 2 of the main manuscript.

Table 7. Run time of each analysis in ONETOOL (in second).

Type	Method	Data1	Data2	Evaluated
InfoQC analysis	freq	0.245	2.069	+
	hwe	0.192	2.191	+
	pca 5	0.219	3.678	+
	het	0.153	2.175	+
	hethom	0.208	2.136	+
	mendel	0.223	2.25	+
	pedinfo	0.099	0.224	+
	relpair	0.256	1.063	+
	famuniq	0.165	0.304	+
Trait Analysis	fcor	0.301	39.289	+
	heritability	0.151	0.452	+
	makecor	0.275	2.447	+
	segreg	7.714	42.522	
Linkage Analysis	lodlink	27.613	12711.080	
	merlin	0.196	*	
Single variant association analysis	scoretest	NA	7.129	
	tdt	0.367	NA	
	sdt	0.209	NA	
	mqls	0.230	NA	
	fqls	0.501	7.809	
	gemma	NA	11.905	
	multifqls	0.195	2.908	+
Gene-based association analysis	collapsing	3.319	2.873	
	pedcmc	0.259	NA	
	famvt	NA	35.354	
	farvat	0.234	3.365	+
	pedgene	0.368	5.183	
	fbskat	0.421	NA	
	rvtdt	16.163	NA	

NA: Not applicable, *****: were too big to run by Merlin, **+**: included in the run time evaluation

Text 5. Discussion

The advantages of family-based genetic studies have been emphasized by the ample amount of literatures and researchers (Ott et al., 2011, Clerget-Darpoux and Elston, 2007; Stein and Elston, 2009, Bailey-Wilson and Wilson, 2011; Wijsman, 2012). The importance of family-based designs has been repeatedly stressed for analyses with sequence data because of the genetic homogeneity between family members (Laird and Lange, 2006). Family study designs provide not only the enrichment of genetic loci containing rare variants, but also methods to control for genetic heterogeneity and population stratification.

As next generation sequence (NGS) data become more and more readily available for genetic and genomic analyses, the need for tools to integrate the various sources and analyze the vast amount of data is inevitable. This need has led to a plethora of such tools already developed and used, as reported in Pabinger et al. (2014). They surveyed 205 such tools for whole-genome/whole-exome sequencing data analysis and reported 32 selected tools. However, most of them are designed for analyzing population-based NGS data, not for family-based data.

We developed a novel tool, ONETOOL, to fill that gap and pipeline the genetic analysis process for pedigree data. It is designed to be as convenient as PLINK, as versatile as S.A.G.E., and as fast as Merlin. Input files for ONETOOL have the most popularly used format, so users familiar with how to use PLINK can easily use ONETOOL without much of a learning curve. Also, the outputs from the S.A.G.E. and Merlin modules are in the same as the original formats, which makes the comparison and the interpretation of results much easier for the many users who are already familiar with those tools. As pointed out by Eu-Ahsunthornwattana et al. (2014), the choice of analysis tool is often made on the basis of speed and convenience, given the strong concordance between the results from the different approaches and implementations in the different tools. In that respect, ONETOOL stands out among other tools as it is specifically designed to provide speed and convenience.

More time took to calculate GRM by ONETOOL than by RVTESTS. This seems to be due to the different approach each program is taking to process the genotype data while reading in a VCF file. In ONETOOL, all genotypes are pre-loaded before any analysis begins while RVTESTS reads in the VCF file sequentially, line by line for each variant, and processes to calculate the GRM. The sequential approach has less I/O burden, so provides the faster calculation for the GRM itself compare to the pre-load approach. However, the sequential approach has the major disadvantages in the down-road analyses. First, it is very limited for any sample-wise analyses. Second, it is computationally very inefficient for any gene-level analyses because the I/O pattern becomes random. The pre-load approach in ONETOOL provides the computational efficiency and the superiority in the rare-variant analysis.

Currently, ONETOOL can be used to analyze the genetic data with bi-allelic variants for the association analyses. For the loci with more than 2 variants will be automatically filtered and reported in the log file it automatically generates for every run.

Though this initial version of ONETOOL implements many different analysis methods for analyzing family data, there are many more. With a modular design, each analysis module within ONETOOL is independent of the others, so it is very easy to extend and add more tools.

References

Abecasis GR, Cherny SS, Cookson WO, Cardon LR (2002) Merlin-rapid analysis of dense genetic maps using sparse gene flow tree. Nat Genet 30:97–101.

Bailey-Wilson JE, Wilson AF (2011) Linkage analysis in the next generation sequencing era. Hum Hered 72:228–236.

Balding DJ, Nichols RA (1995) A method for quantifying differentiation between populations at multi-allelic loci and its implications for investigating identity and paternity. Genetica 96:3–12.

Blangero J, Teslovich TM, Sim X, Almeida MA, Jun G, Dyer TD, Johnson M, Peralta JM, Manning AK, Wood AR, et al. (2016) Omics squared: Human genomic, transcriptomic, and phenotypic data for Genetic Analysis Workshop 19. BMC Proc 9 Suppl 7:S20.

Burdick JT, Chen WM, Abecasis GR, Cheung VG (2006) In silico method for inferring genotypes in pedigrees. Nat Genet 38(9):1002-1004.

Chen H, Meigs JB, Dupuis J. (2012) Sequence kernel association test for quantitative traits in family samples. Genet Epidemiol 37(2):196–204.

Chen W, Schaid DJ. 2012. PedBLIMP: Extending Linear Predictors to Impute Genotypes in Pedigrees. Genet Epidemiol 37(2):196–204.

Choi S, Lee S, Cichon S, Nöthen MM, Lange C, Park T, Won S (2014) FARVAT: a family-based rare variant association test. Bioinformatics 30:3197–3205.

Choi S, Lee S, Qiao D, Hardin M, Cho MH, Silverman EK, Park T, Won S (2016) FARVATX: Family-Based Rare Variant Association Test for X-Linked Genes. Genet Epidemiol. 40(6):475-85.

Cheung,C.Y. et al. (2013) GIGI: an approach to effective imputation of dense genotypes on large pedigrees. Am. J. Hum. Genet., 92, 504–516.

Chung RH, Tsai WY, Kang CY, Yao PJ, Tsai HJ, Chen CH (2016) FamPipe: An Automatic Analysis Pipeline for Analyzing Sequencing Data in Families for Disease Studies. PLoS Comput Biol 12(6):e1004980.

Clerget-Darpoux F, Elston RC (2007) Are linkage analysis and the collection of family data dead? Prospects for family studies in the age of genome-wide association. Hum Hered 64:91–96.

Eu-Ahsunthornwattana J, Miller EN, Fakiola M, Wellcome Trust Case Control Consortium 2, Jeronimo SM, Blackwell JM, Cordell HJ (2014) Comparison of methods to account for relatedness in genome-wide association studies with family-based data. PLoS Genet.10:e1004445.

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

Försti A, Kumar A, Paramasivam N, Schlesner M, Catalano C, Dymerska D, … Hemminki K (2016). Pedigree based DNA sequencing pipeline for germline genomes of cancer families. Hered Cancer Clin Pract, 14:16. Doi:10.1186/s13053-016-0058-1.

Gazal S, Gosset S, Verdura E, Bergametti F, Guey S, Babron MC, Tournier-Lasserve E (2016). Can whole-exome sequencing data be used for linkage analysis? Eur J Hum Genet 24:581-586.

Han F, Pan W (2010) A Data-Adaptive Sum Test for Disease association with multiple common or rare variants. Hum Hered 70:42–54.

He Z, Zhang D, Renton AE, Li B, Zhao L, Wang GT, Goate AM, Mayeux R, Leal SM (2017). The Rare-Variant Generalized Disequilibrium Test for Association Analysis of Nuclear and Extended Pedigrees with Application to Alzheimer Disease WGS Data. Am J Hum Genet 100:193-204.

Hu H, Roach JC, Coon H, Guthery SL, Voelkerding KV, Margraf RL, Durtschi JD, Tavtigian SV, Shankaracharya, Wu W, Scheet P, Wang S, Xing J, Glusman G, Hubley R, Li H, Garg V, Moore B, Hood L, Galas DJ, Srivastava D, Reese MG, Jorde LB, Yandell M, Huff CD (2014). A unified test of linkage analysis and rare-variant association for analysis of pedigree sequence data. Nat Biotechnol 32:663-669

Jiang D, McPeek MS (2014). Robust Rare Variant Association Testing for Quantitative Traits in Samples with Related Individuals. Genet Epidemiol 38:10-20.

Kim Y, Lee Y, Lee S, Kim NH, Lim J, Kim YJ, Oh JH, Min H, Lee M, Seo HJ, Lee SH, Sung J, Cho NH, Kim BJ, Han BG, Elston RC, Won S, Lee J (2015). On the estimation of heritability with family-based and population-based samples. BioMed Res Int 2015:671349.

Koboldt DC, Larson DE, Sullivan LS, Bowne SJ, Steinberg KM, Churchill JD, Buhr AC, Nutter N, Pierce EA, Blanton SH, Weinstock GM, Wilson RK, Daiger SP (2014). Exome-based mapping and variant prioritization for inherited Mendelian disorders. Am J Hum Genet 94:373-384.

Kong A, Cox NJ (1997) Allele-sharing models: LOD scores and accurate linkage tests. Am J Hum Genet 61:1179-1188.

Kunji K, Ullah E, Nato AQ, Wijsman EM, Sadd M (2018). GIGI-Quick: a fast approach to impute missing genotypes in genome-wide association family data. Bioinformatics doi: 10.1093/bioinformatics/btx782.

Laird NM, Lange C (2006). Family-based designs in the age of large-scale gene-association studies. Nat Rev Genet 7:385–394.

Lange K, Papp JC, Sinsheimer JS, Sripracha R, Zhou H, Sobel EM (2013). Mendel: the Swiss army knife of genetic analysis programs. Bioinformatics 29:1568–1570.

Lee S, Emond MJ, Bamshad MJ, Barnes KC, Rieder MJ, Nickerson DA; NHLBI GO Exome Sequencing Project—ESP Lung Project Team, Christiani DC, Wurfel MM, Lin X. (2012) Optimal Unified Approach for Rare-Variant Association Testing with Application to Small-Sample Case-Control Whole-Exome Sequencing Studies. Am J Hum Genet 91:224-237.

Lee SH, Wray NR, Goddard ME, Visscher PM (2011). Estimating missing heritability for disease from genome-wide association studies. Am J Hum Genet 88:294-305.

Li B, Leal SM (2008) Methods for detecting associations with rare variants for common diseases: application to analysis of sequence data. Am J Hum Genet 83:311-321.

Li B, Wang GT, Leal SM (2015). Generation of sequence-based data for pedigree-segregating Mendelian or Complex trait. Bioinformatics 31:3706–3708.

Liu DJ, Leal SM (2010) A novel adaptive method for the analysis of next-generation sequencing data to detect complex trait association rare variants due to gene main effects and inteactions. PLOD Genet 6, e1001156.

Livne OE, Han L, Alkorta-Aranburu G, Wentworth-Sheilds W, Abney M, Ober C, Nicolae DL (2015) PRIMAL: Fast and Accurate Pedigree-based Imputation from Sequence Data in a Founder Population. PLOS ComBio DOI:10.1371/journal.pcbi.1004139.

Madsen BE, Browning SR (*2009*) A Groupwise Association Test for Rare Mutations Using a Weighted Sum Statistic. PLoS Genet 5:e1000384.

Morris JA, Barrett JC (2012). Olorin: combining gene flow with exome sequencing in large family studies of complex disease. Bioinformatics 28:3320–3321.

Morris AP, Zeggini E (2010). An evaluation of statistical approachs to rare variant analysis in genetic association studies. Genet Epidemiol 34:188–193.

Nato AQ Jr., Chapman NH, Sohi HK, Nguyen HD, Brkanac Z (2015). PBAP: a pipeline for file processing and quality control of pedigree data with dense genetic markers. Bioinformatics 31:3790–3798.

Ott J, Kamatani Y, Lathrop M (2011) Family-based designs for genome-wide association studies. Nat Rev Genet 12:465-474.

Ott J, Wang L, Leal SM (2015) Genetic linkage analysis in the age of whole-genome sequencing. Nat Rev Genet 16:275-284.

Pabinger S, Dander A, Fisher M, Snajder R, Sperk M, Efremova M, Krabichler B, Speicher MR, Trajanoski Z (2013) A survey of tools for variant analysis of next-generation genome sequencing data. Briefings in Bioinformatics 15:256-278.

Park S, Lee S, Lee Y, Herold C, Hooli B, Mullin K, Park T, Park C, Bertram L, Lange C, Tanzi R, Won S (2015) Adjusting heterogeneous ascertainment bias for genetic association analysis with extended families. BMC Med Genet 16:62.

Price AL, Kryukov GV, de Bakker PI, Purcell SM, Staples J, Wei LJ, Sunyaev SR (2010) Pooled association tests for rare variants in exon-resequencing studies. Am J Hum Genet 86:832-838.

Purcell S, Neale B, Todd-Brown K, Thomas L, Ferreira MA, Bender D, Maller J, Sklar P, de Bakker PI, Daly MJ, Sham PC (2007) PLINK: a tool set for whole-genome association and population-based linkage analyses. Am J Hum Genet 81:559-575.

S.A.G.E. 6.4 (2016) Statistical Analysis for Genetic Epidemiology http://darwin.cwru.edu/sage/.

Schaid DJ, McDonnell SK, Sinnwell JP, Thibodeau SN (2013) Multiple genetic variant association testing by collapsing and kernel methods with pedigree or population structured data. Genet Epidemiol 37:409-418.

Schnell AH, Sun X (2012) Model-based linkage analysis of a quantitative trait. Methods Mol Biol 850:263-283.

Sincan M, Simeonov DR, Adams D, Markello TC, Pierson TM, Toro C, Gahl WA, Boerkoel CF (2012) VAR-MD: A Tool to Analyze Whole Exome–Genome Variants in Small Human Pedigrees with Mendelian Inheritance. Hum Mutat Apr;33(4):593-8. doi: 10.1002/humu.22034. Epub 2012 Feb 24.

Sinnwell JP, Therneau TM, Schaid DJ (2014) The kinship2 R package for pedigree data. Hum hered 78:91-93.

Song YE, Elston RC (2013) PedWiz: a web-based tool for pedigree informatics. Front Genet 4:189. 10.3389/fgene.2013.00189.

Speed D, Balding DJ (2015) Relatedness in the post-genomic era: is it still useful? Nat Rev Gen 16:33–44.

Spielman RS, Ewens WJ (1998) A Sibship Test for Linkage in the presence of association: the sib transmission/disequilibrium test. Am J Hum Genet 62:450-458.

Spielman RS, McGinnis RE, Ewens WJ (1993) Transmission test for linkage disequilibrium: the insulin gene region and insulin-dependent diabetes mellitus (IDDM). Am J Hum Genet 52:506-516.

Stein CM, Elston RC (2009) Finding genes underlying human disease. Clin Genet 75:101-106.

Thornton T, McPeek MS (2007) Case-Control Association Testing with Related Individuals: A More Powerful Quasi-Likelihood Score Test. Am J Hum Genet 81:321-337.

Wang GT, Zhang D, Li B, Dai H, Leal SM (2015) Collapsed haplotype pattern method for linkage analysis of next-generation sequence data. Eur J Hum Genet 23:1739-1743.

Wang L, Lee S, Gim J, Qiao D, Cho M, Elston RC, Silverman EK, Won S. Family-Based Rare Variant Association Analysis: A Fast and Efficient Method of Multivariate Phenotype Association Analysis. Genet Epidemiol. 2016 Sep;40(6):502-11.

Wijsman EM (2012) The role of large pedigrees in an era of high-throughput sequencing. Hum Genet 131:1555-1563.

Won S, Kim W, Lee S, Lee Y, Sung J, Park T (2015) Family-based association analysis: a fast and efficient method of multivariate association analysis with multiple variants. BMC Bioinformatics 16:46.

Wu MC, Lee S, Cai T, Li Y, Boehnke M, Lin X. (2011) Rare-variant association testing for sequencing data with the sequence kernel association test. Am J Hum Genet 89:82-93.

Yan Q, Tiwari HK, Yi N, Gao G, Zhang K, Lin WY, Lou XY, Cui X, Liu N (2015) A Sequence Kernel Association Test for Dichotomous Traits in Family Samplesunder a Gene ralized Linear Mixed Model. Hum Hered 79:60-68.

Yao J, Zhang KX, Kramer M, Pellegrini M, McCombie WR (2014) FamAnn: an automated variant annotation pipeline to facilitate target discovery for family-based sequencing studies. Bioinformatics 30:1175-1176.

Zhan X, Hu Y, Abecasis GR, Liu DJ (2016) RVTESTS: an efficient and comprehensive tool for rare variant association analysis using sequencing data. Bioinformatics 32:1423-1426.

Zhou X, Stephens M (2012) Genome-wide efficient mixed-model analysis for association studies. Nat Genet 44: 821-824.

Zhu Y, Xiong M (2012) Family-based association studies for next-generation sequencing. Am J Hum Genet. 90:1028–1045.