Using High-Throughput Genotyping and Large Families to Reduce Sequencing Costs

Authors

Stewart, WCL[1,2,3]

Bartlett, CW[1,3]

Affiliations

[1]The Research Institute at Nationwide Children's Hospital, and [2]Departments of Statistics and [3]Pediatrics at Ohio State University

Correspondence

Email: William.Stewart@nationwidechildrens.org

Abstract

For modern linkage studies involving many small families, there exists an efficient estimator of disease gene location (denoted $\tilde{\theta}$) that averages location estimates computed from random subsamples of the data. This estimator has lower mean squared error than competing estimators and yields narrower confidence intervals (CIs) as well. However, when the number of families is small and the pedigree structure is large (possibly extended), the computational feasibility and statistical properties of $\tilde{\theta}$ are not known. Using simulation and real data, this research shows that (for this extremely important but often overlooked study design), CIs based on $\tilde{\theta}$ are narrower than CIs based on a single subsample, and that the corresponding percent reduction in CI length is bounded above by the square root of the percent reduction in variance. As a proof of principle, $\tilde{\theta}$ was applied to the dense SNP data of four large, extended, specific language impairment (SLI) pedigrees, and the single subsample CI was reduced by 18%. In summary, confidence intervals based on $\tilde{\theta}$ should minimize re-sequencing costs beneath linkage peaks, and should reduce the number of genes to investigate in follow-up candidate gene studies.

Introduction

Modern linkage studies are extremely effective at identifying and defining genetic loci (i.e. broad chromosomal regions) that influence highly heritable traits[2-5]. But, identifying the disease gene(s) within a genetic locus is often difficult, expensive, and slow because the number of candidate genes spanned by a genetic locus is typically large. Here, the simple and efficient estimator of disease gene location (denoted $\tilde{\theta}$) that was first introduced by Stewart et al.[1] for studies involving a large number of small families genotyped on a dense panel of SNPs (single nucleotide polymorphisms) is extended. Because of these extensions, CIs (confidence intervals) based on $\tilde{\theta}$ can now be constructed from studies involving a small number of large families, which can reduce the size of the region implicated by a single subsample CI by more than 18 percent (see Results).

Most modern linkage studies contain high-throughput genotype data on millions of roughly equi-spaced SNPs. Although these dense SNP panels are considerably informative for linkage[6], SNPs in *close* proximity to one another are correlated, which can negatively affect the precision and accuracy of estimators of disease gene location[1]. The correlation between SNPs is known as linkage disequilibrium (LD), and in practice, the most effective solution is to consider one or more LE (linkage equilibrium) subsamples of the dense SNP data. Within each LE subsample, the SNPs are approximately uncorrelated; and because the intervening dense SNPs contain very little additional information about inheritance, each LE subsample retains almost all of the original information for linkage. Moreover, the maximum likelihood estimator of θ (denoted $\hat{\theta}$), which is based on a single LE subsample, remains consistent.

For modern linkage studies involving a large number of small families, one can substantially improve the precision of single subsample estimation by averaging location estimates over multiple subsamples. For example, an accurate estimate of the variance of $\tilde{\theta}$ (denoted τ^2) follows from the sequence $(\tilde{\theta}^{(1)},\cdots,\tilde{\theta}^{(k)})$, where each $\tilde{\theta}^{(j)}$ is computed from the j^{th} bootstrap resample. The resampling occurs over

independent families, and ($\tilde{\theta} \pm 1.96\tau$) is an approximate but efficient 95% CI for θ that leverages the information on trait location contained in LE subsamples of the original dense SNP data. By contrast, the nonparametric bootstrap procedure is no longer valid for studies involving a small number of large families; therefore, although the computation of $\tilde{\theta}$ remains straightforward, a new approach is needed to compute τ^2 for a small number of large families.

In Methods, a formal definition of $\tilde{\theta}$ is given, and key mathematical relationships between various statistical quantities of interest are established. These relationships can inform the estimation (or approximation) of τ^2, and shed light on the efficiency gains of CIs based on $\tilde{\theta}$ relative to CIs based on $\hat{\theta}$. In Results, the two estimators are compared using data simulated under a variety of trait models with a realistic pattern of LD. As proof of principle, $\tilde{\theta}$ was applied to three large, extended specific language impairment (SLI) pedigrees. Finally, some of the limitations, and several potentially high-impact implications of the proposed estimator and its corresponding CI are discussed.

Methods

For a dense SNP linkage study, let $\mathbf{G}$ and $\mathbf{T}$ be the multilocus genotype and trait data, respectively. Ideally, inference about θ should be made on the basis of the observed-data likelihood: $L(\theta; \mathbf{G}, \mathbf{T}, \phi)$, where θ represents the location of a hypothesized trait locus, and ϕ represents all other parameters in the analysis (e.g. LD structure, trait model, gene frequencies, genetic map, etc.). In the presence of LD and missing data, the observed-data likelihood is computationally feasible only for small to moderate sized pedigrees, and only when the LD structure is assumed to follow a haplotype-block model[7]. These constraints are quite restrictive, and as such, inference about θ is usually made on the basis of the transformed-data likelihood: $L(\theta; \mathbf{M}, \mathbf{T}, \phi)$ [1], where $\mathbf{M} \equiv h(\mathbf{G}, \mathbf{S})$ is a random LE subsample of $\mathbf{G}$ that retains the genotypes of SNPs in the random subpanel $\mathbf{S}$. The SNPs in

[1] In the case of nonparametric likelihoods, $L(\theta; \mathbf{M}, \mathbf{T}, \phi)$ is usually replaced by the Kong & Cox likelihood: $L(\hat{\delta}; \mathbf{M}, \mathbf{T}, \theta)$[8].

$\mathbf{S}$ are intentionally chosen to be approximately uncorrelated (i.e. in LE), so that $L(\theta;\mathbf{M},\mathbf{T},\phi)$ is computationally feasible, even for large, extended pedigrees.

In practice, the unknown ϕ is replaced by a consistent estimate $\hat{\phi}$ and estimates of θ (and CIs for θ) are based on the $\text{argmax}_\theta L(\theta;\mathbf{M},\mathbf{T},\hat{\phi})$, which depends on a single (randomly chosen) LE subsample and approximates $\hat{\theta}$ for a large number of parent-offspring transmissions among affected families. However for modern linkage studies, the number of approximately uncorrelated subpanels is large, and because each corresponding subsample provides slightly different information about θ, an estimator that combines information from multiple LE subsamples should outperform one that does not (i.e. $\hat{\theta}$). Let's define $\tilde{\theta} \equiv \mathbf{E}[\hat{\theta} \mid \mathbf{G}]$, then the following equation shows that increased precision is guaranteed:

$$Var(\hat{\theta}(\mathbf{M})) = Var(\mathbf{E}[\hat{\theta} \mid \mathbf{G}]) + \mathbf{E}[Var(\hat{\theta} \mid \mathbf{G})]. \quad \text{Eq. 1}$$

In words, Eq. (1) says that the variance of $\hat{\theta}$ is equal to the variance of $\tilde{\theta}$ plus the average Monte Carlo error, where the Monte Carlo error refers to the conditional variance of $\hat{\theta}$ over LE subsamples $\mathbf{M}$ given dense SNP data, $\mathbf{G}$. Note that while an LD model is not needed to generate *iid* realizations of $\mathbf{M}$ (unconditionally, or conditionally for a given $\mathbf{S}$), $Var(\hat{\theta}(\mathbf{M})) \neq Var(\hat{\theta}(\mathbf{M}) \mid \mathbf{S})$, as common practice would suggest. Moreover, Eq. (1) implies that $Var(\tilde{\theta}) = Var(\hat{\theta}) - \mathbf{E}[Var(\hat{\theta} \mid \mathbf{G})]$, which together with the previously mentioned note, implies that the real obstacle to constructing narrower CIs based $\tilde{\theta}$ is evaluation of the outer expectation, as this computation *does* require an LD model. For modern linkage studies with large families, this computation is not feasible. Therefore, an upper bound on $Var(\tilde{\theta})$ is attractive, especially if it is strictly less than $Var(\hat{\theta})$, and does not require an LD model.

Remarkably, one such upper bound

exists: the minimum of $[Var(\hat{\theta}\,|\,\mathbf{S}_1),\cdots,Var(\hat{\theta}\,|\,\mathbf{S}_k)]$ To understand why this is a useful upper bound, it is helpful to recast the problem as a random effects model:

$$\hat{\theta}(\mathbf{G},\mathbf{S}) = \tilde{\theta}(\mathbf{G}) + \varepsilon(\mathbf{G},\mathbf{S}), \qquad \text{Eq. 2}$$

where $\varepsilon(\mathbf{G},\mathbf{S})$ is the random effect of subpanel $\mathbf{S}$ on the dense SNP data $\mathbf{G}$. By fixing $\mathbf{S}$, and considering the variation attributable to $\mathbf{G}$, Eq. 2 implies that

$$Var(\hat{\theta}\,|\,\mathbf{S}) = Var(\tilde{\theta}) + \gamma^2, \qquad \text{Eq. 3}$$

where $\gamma^2 = Var(\varepsilon\,|\,\mathbf{S})$. Note that, the minimum of $[Var(\hat{\theta}\,|\,\mathbf{S}_1),\cdots,Var(\hat{\theta}\,|\,\mathbf{S}_k)]$ is strictly less than the average of $[Var(\hat{\theta}\,|\,\mathbf{S}_1),\cdots,Var(\hat{\theta}\,|\,\mathbf{S}_k)]$ (which by definition is the variance of $\hat{\theta}$). Furthermore, the minimum of $[Var(\hat{\theta}\,|\,\mathbf{S}_1),\cdots,Var(\hat{\theta}\,|\,\mathbf{S}_k)]$ can be computed (or estimated) without an LD model.

Data Description

To quantify the potential gains in precision of $\tilde{\theta}$ over $\hat{\theta}$, simulated data and real dense SNP linkage data on extended families were analyzed. Dominant and recessive trait models with incomplete penetrance were simulated separately, and each trait locus was positioned in the middle of 132 equi-spaced haplotype-blocks (average spacing 0.5 centi-Morgans between each block). Each block was comprised of 3 SNPs, with each SNP separated by 0.25 centi-Morgans (cMs). For the dominant trait model, a disease allele frequency of 1%, a phenocopy rate of 1%, and a penetrance of 20% were assumed; each replicate dense SNP data set contained five 3-generation families (Figure 1). As for the recessive trait, the corresponding parameters were10%, 1%, and 50%, respectively. Each replicate dense SNP linkage data set with the recessive trait model contained seven 3-generation families (Figure 2).

Figure 1

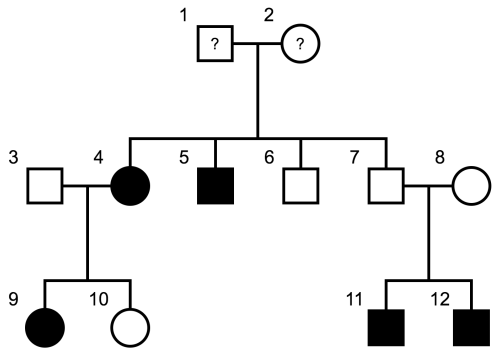

Figure 1: The pedigree structure used for simulating dense SNP data linked to a dominant trait with incomplete penetrance and LD. The question mark symbol "?" denotes individuals who are unobserved for both genotypic and phenotypic data. Filled shapes represent affected individuals, and unfilled shapes represent unaffected individuals.

Figure 2

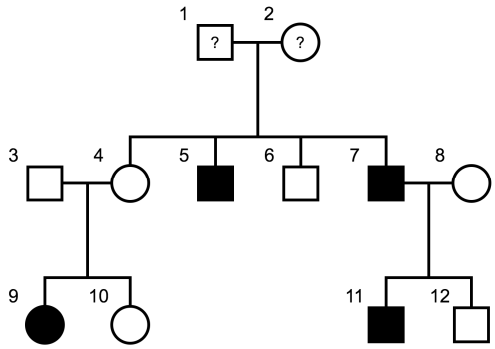

Figure 2: The pedigree structure used for simulating dense SNP data linked to a recessive trait with incomplete penetrance and LD. The question mark symbol "?" denotes individuals who are unobserved for both genotypic and phenotypic data. Filled shapes represent affected individuals, and unfilled shapes represent unaffected individuals.

Note that, in a haplotype-block model, the blocks are uncorrelated, but the SNPs within each block are highly correlated. The LD pattern used in the simulations (Figure 3) mimicked the empirical LD pattern used in Stewart et al. [1]. With this model and these parameters, the outer expectation (and variances) in Eq. (1) were estimated based on 20,000 independent replicates

of dense SNP data $\mathbf{G}$, which allowed us to quantify the gains in precision of $\tilde{\theta}$ over $\hat{\theta}$.

Figure 3

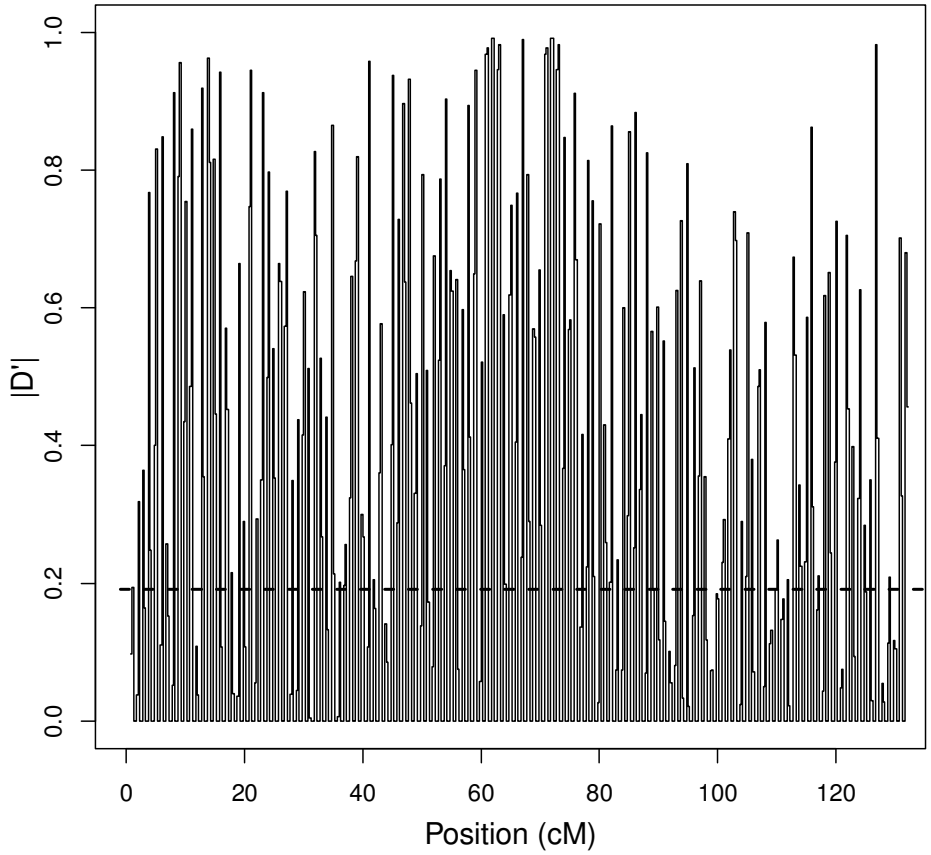

Figure 3: Absolute D-prime as a function of 395 contiguous SNP-SNP intervals comprising 132 haplotype blocks, with 3 SNPs per block.

In the real data analysis, involving 3 extended multiplex SLI pedigrees[9] containing 165 individuals, with 105 subjects genotyped at 8,736 SNPs spanning chromosome 13 (132 cMs), the outer expectation in Eq. (1) is not computationally feasible. Therefore, the proposed upper bound on $Var(\tilde{\theta})$ (as described in Methods) was used to

estimate the gain in precision. Because $L(\theta;\mathbf{M},\mathbf{T},\hat{\phi})$ is computationally intractable for the SLI data, the Markov chain Monte Carlo (MCMC) program LM_MARKERS[10] was used to obtain $\hat{\theta}$ and $\tilde{\theta}$; this program is just one of several in the MORGAN suite. In computing $\tilde{\theta}$, the average was taken over 20 LE subpanels, and each subpanel was sampled conditional on the observed dense SNP data $\mathbf{G}$ using the program EAGLET[1, 4, 11-13].

Results

From the analysis of simulated data, there is excellent agreement with Eq. (1) for dense SNP linkage studies involving large (possibly extended) multiplex families with missing data (Table 1). Note that, in Table 1, the estimator of $Var(\hat{\theta})$ is the average of $(Var[\hat{\theta}\,|\,\mathbf{S}_1],\cdots,Var[\hat{\theta}\,|\,\mathbf{S}_k])$ for $k = 20$.

Table 1: Decomposing the Variance of $\hat{\theta}$

Trait Model	$Var(\hat{\theta})$	$Var(\tilde{\theta})$	$\mathbf{E}[Var(\hat{\theta}\,|\,\mathbf{G})]$
Dominant	46.39	38.43	8.84
Recessive	63.13	53.01	11.24

Furthermore, given that the simulated LD model mimics the LD-pattern of a real data set, it is not unreasonable to expect precision gains (i.e. reductions in variance) in the neighborhood of 15-20 percent (Table 1). Similarly, the 95% CIs based on $\tilde{\theta}$ are 10% and 9% narrower than the 95% CIs based on $\hat{\theta}$, for dominant and recessive models respectively. Interestingly, under the assumption that $\hat{\theta}$ and $\tilde{\theta}$ are each normally distributed with mean θ, it follows from Eq. (1) that the percent reduction in CI length is bounded above by the square root of the percent reduction in variance. Hence, the percent reduction in CI length tends to be large when, relative to the variance of $\tilde{\theta}$, the average Monte Carlo error is large.

From the analysis of the real SLI data, the original 95% CI was narrowed from 10.9 cMs[14] to 8.9 cMs (Figure 4). Because the 105 genes of the original CI

are not evenly distributed, the 18% reduction in CI length corresponds (in this case) to a 9.5% reduction in the number of candidate genes. However, it is noteworthy that the 95% CI based on $\tilde{\theta}$ does not include an interesting biological candidate *PCDH9* (protocadherin 9), suggesting that the re-sequencing of other genes beneath this peak should take priority.

Figure 4

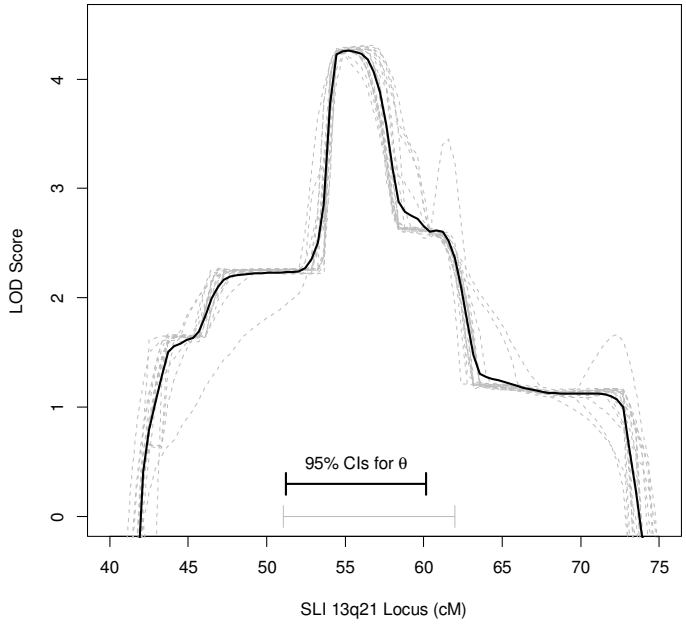

Figure 4: The solid curve is the average lod score, where the average is taken over 20 random subsamples. The dotted gray curves are the lod scores from each of the 20 subsamples. The solid and gray intervals are the 95% CIs based on the proposed estimator and a single subsample, respectively.

Discussion

This research has shown that the simple, but efficient estimator of trait location, first proposed by Stewart et al.[1] for modern linkage studies involving a large number of small families, can also be used to obtain narrower confidence intervals for studies involving a small number of large (possibly extended) pedigrees. This research also showed that gains in precision are possible, despite the fact that the observed-data

likelihood: $L(\theta;\mathbf{G},\mathbf{T},\hat{\phi})$ is intractable for large families in the presence of missing data and LD. That said, an important area of open research involves the development of methods to compute (or simulate) $L(\theta;\mathbf{G},\mathbf{T},\hat{\phi})$ on arbitrary pedigrees across a wide range of LD patterns. Such methods should maximize the gain in precision, which in turn, should lead to the greatest reduction in candidate gene re-sequencing costs.

Although a method for constructing narrow CIs from modern linkage studies containing a mixture of many small and several large families has not yet been devised, one could easily imagine implementing either a pooled data approach, or a meta-analysis approach. Conceptually, the pooled data approach is simpler because all of the data could be analyzed jointly by LM_MARKERS, which uses exact calculation where possible and MCMC otherwise. However, given that $\tilde{\theta}_l$ (which is based on large families) and $\tilde{\theta}_s$ (which is based on small families) are both approximately normally distributed, a meta-analysis approach that averages $\tilde{\theta}_l$ and $\tilde{\theta}_s$ with weights that vary inversely in proportion to their marginal variances is straightforward too. Moreover, it's not immediately clear which approach would be better, or that the difference (if any) would have any practical importance with respect to follow-up re-sequencing or fine-mapping efforts.

Finally, given the number of existing orphan linkage peaks (i.e. the approximately 3000 linkage peaks for highly heritable traits for which no known disease gene has yet been found[15]), and given the affordability of genome-wide genotyping and whole-exome sequencing, these two technologies could be paired with the proposed estimator of trait location to substantially expedite the rate at which disease genes are discovered, while simultaneously reducing the overall costs. In principle, 95% CIs based on $\tilde{\theta}$ should facilitate progress on the much more difficult, and more widely spread, problem of identifying non-exonic pathogenic variants for complex traits.

Acknowledgements

We are grateful to The Research Institute of Nationwide Children's Hospital for it's generous support of this important work.

Web Resources

The URLs for software used in this research are:

EAGLET

http://u.osu.edu/stewart.1212

MERLIN

http://csg.sph.umich.edu/abecasis/Merlin/download

MORGAN

https://www.stat.washington.edu/thompson/Genepi/pangaea.shtml

References

1. Stewart, W.C.L., A.L. Peljto, and D.A. Greenberg, *Multiple subsampling of dense SNP data localizes disease genes with increased precision.* Hum Hered, 2010. **69**(3): p. 152-9.
2. Shugart, Y.Y., et al., *An SNP linkage scan identifies significant Crohn's disease loci on chromosomes 13q13.3 and, in Jewish families, on 1p35.2 and 3q29.* Genes Immun, 2008. **9**(2): p. 161-7.
3. Huyghe, J.R., et al., *Genome-wide SNP-based linkage scan identifies a locus on 8q24 for an age-related hearing impairment trait.* Am J Hum Genet, 2008. **83**(3): p. 401-7.
4. Rodriguez-Murillo, L., et al., *Novel loci interacting epistatically with bone morphogenetic protein receptor 2 cause familial pulmonary arterial hypertension.* J Heart Lung Transplant, 2010. **29**(2): p. 174-80.
5. Costantino, F., et al., *Whole-genome single nucleotide polymorphism-based linkage analysis in spondyloarthritis multiplex families reveals a new susceptibility locus in 13q13.* Ann Rheum Dis, 2016. **75**(7): p. 1380-5.
6. Evans, D.M. and L.R. Cardon, *Guidelines for genotyping in genomewide linkage studies: single-nucleotide-polymorphism maps versus microsatellite maps.* Am J Hum Genet, 2004. **75**(4): p. 687-92.
7. Abecasis, G.R. and J.E. Wigginton, *Handling marker-marker linkage disequilibrium: pedigree analysis with clustered markers.* Am J Hum Genet, 2005. **77**(5): p. 754-67.
8. Kong, A. and N.J. Cox, *Allele-sharing models: LOD scores and accurate linkage tests.* Am J Hum Genet, 1997. **61**(5): p. 1179-88.
9. Bartlett, C.W., et al., *A major susceptibility locus for specific language impairment is located on 13q21.* Am J Hum Genet, 2002. **71**(1): p. 45-55.

10. Wijsman, E.M., J.H. Rothstein, and E.A. Thompson, *Multipoint linkage analysis with many multiallelic or dense diallelic markers: Markov chain-Monte Carlo provides practical approaches for genome scans on general pedigrees.* Am J Hum Genet, 2006. **79**(5): p. 846-58.
11. Stewart, W.C.L., E.N. Drill, and D.A. Greenberg, *Finding disease genes: a fast and flexible approach for analyzing high-throughput data.* Eur J Hum Genet, 2011. **19**(10): p. 1090-4.
12. Kambhampati, S., et al., *Managing Tiny Tasks for Data-Parallel, Subsampling Workloads.* 2014 Ieee International Conference on Cloud Engineering (Ic2e), 2014: p. 225-234.
13. Stewart, W.C., et al., *Next-generation linkage and association methods applied to hypertension: a multifaceted approach to the analysis of sequence data.* BMC Proc, 2014. **8**(Suppl 1 Genetic Analysis Workshop 18Vanessa Olmo): p. S111.
14. Simmons, T.R., et al., *Increasing genotype-phenotype model determinism: application to bivariate reading/language traits and epistatic interactions in language-impaired families.* Hum Hered, 2010. **70**(4): p. 232-44.
15. McKusick, V.A., *Mendelian Inheritance in Man and its online version, OMIM.* Am J Hum Genet, 2007. **80**(4): p. 588-604.