Imputation of Microsatellite Markers With Tag SNPs

by

Annchen Knodt

Department of Computational Biology and Bioinformatics
Duke University

Date: _____

Approved:

_____
Ahmad Hariri, Supervisor

_____
Sayan Mukherjee

_____
Elizabeth Hauser

_____
Avshalom Caspi

Thesis submitted in partial fulfillment of the requirements for the degree of
Master of Science in the Department of Computational Biology and Bioinformatics
in the Graduate School of Duke University
2012

UMI Number: 1517148

All rights reserved

INFORMATION TO ALL USERS
The quality of this reproduction is dependent on the quality of the copy submitted.

In the unlikely event that the author did not send a complete manuscript
and there are missing pages, these will be noted. Also, if material had to be removed,
a note will indicate the deletion.

UMI 1517148

Copyright 2012 by ProQuest LLC.

All rights reserved. This edition of the work is protected against
unauthorized copying under Title 17, United States Code.

ProQuest LLC.
789 East Eisenhower Parkway
P.O. Box 1346
Ann Arbor, MI 48106 - 1346

Abstract

Imputation of Microsatellite Markers With Tag SNPs

by

Annchen Knodt

Department of Computational Biology and Bioinformatics
Duke University

Date: _____

Approved:

_____
Ahmad Hariri, Supervisor

_____
Sayan Mukherjee

_____
Elizabeth Hauser

_____
Avshalom Caspi

An abstract of a thesis submitted in partial fulfillment of the requirements for the degree of Master of Science in the Department of Computational Biology and Bioinformatics
in the Graduate School of Duke University
2012

Copyright © 2012 by Annchen Knodt
All rights reserved except the rights granted by the
Creative Commons Attribution-Noncommercial Licence

Abstract

Of the two most common forms of genetic variation in the human genome, Single Nucleotide Polymorphisms (SNPs) and Variable Number Tandem Repeat Polymorphisms (VNTRs), SNPs are much more easily and inexpensively assayed in a high-throughput manner. For this reason, we seek to explore methods that can allow us to use the more readily available SNP genotype information to infer VNTR genotypes in nearby genomic regions. We focus in particular on imputing a VNTR polymorphism, 5-HTTLPR, in the promoter region of the serotonin transporter gene in a small sample of individuals from an ongoing neuroimaging genetics study, a portion of whom have both manual 5-HTTLPR genotypes and genome wide SNP data. We investigate four imputation methods: Tagger, Vertex Discriminant Analysis (VDA), IMPUTE2, and BEAGLE. We achieve an accuracy of 93% with VDA in our subsample of Caucasians with manual 5-HTTLPR genotypes. Further, we find that for the entire Caucasian subsample without manual genotypes, a majority of the imputation methods tested make the same 5-HTTLPR genotype call.

Contents

List of Tables

List of Figures

Acknowledgements

I would like to thank Dr. Ahmad Hariri and all members of the Laboratory of Neurogenetics for their continued support, encouragement, and guidance. I would also like to express my gratitude towards all of the faculy, staff, and my fellow students in the Computational Biology and Bioinformatics program. I am grateful for the opportunity to work and study alongside so many talented and accommodating people. This work was made possible in part by funding through a training grant to the Computational Biology and Bioinformatics program at Duke University.

1

Introduction

There exist many types and sources of variation in the human genome. Recent advances in genome tools and technologies have allowed us to study these genetic variations on a large scale. This has led to the discovery of a wide assortment of ways in which individuals differ at the level of DNA. Among these are Single Nucleotide Polymorphisms (SNPs), or single base pair substitutions on the DNA, and Variable Number Tandem Repeat Polymorphisms (VNTRs), or small segments (roughly between 1 and 500 base pairs) of DNA that repeat multiple times in succession, with the number of repeats varying between individuals. While SNPs are easily and inexpensively assayed in a high-throughput manner, genotyping VNTRs involves greater expense and manual effort. For this reason, we seek to explore methods that can allow us to use the more readily available SNP genotype information to infer VNTR genotypes in nearby genomic regions. We focus in particular on using machine learning methods to impute a VNTR polymorphism in the promoter region of the serotonin transporter gene, 5-HTTLPR, in a small sample of individuals from an ongoing neuroimaging genetics study. While most imputation methods are trained and tested on large samples, this study considers their applicability in such situations where

sample size is limited, as is often the case in studies of its kind.

2

Background

2.1 Genetic polymorphisms

Of the 3 billion base pairs present in the human genome, 99.5% are the same from one individual to the next [20]. Our unique individual characteristics, then, arise from sequence variations, or polymorphisms, in the remaining 0.5%. Such polymorphisms can be attributed to a variety of different mechanisms and take on many different forms.

Instances where two individuals differ by a single nucleotide (A, C, G or T) are called Single Nucleotide Polymorphisms (SNPs). Generally, individuals in a population will have one of two possible nucleotides, called alleles, at each such polymorphic site. The occurrence and distribution of SNPs in the genome are influenced by the way in which natural selection causes the fixation of adaptive alleles, as well as other factors such as genetic recombination and mutation rate. SNPs are the most frequently occurring type of genetic polymorphism, with over 10 million having been discovered in the human genome [18]. Many of these variants have been implicated in human disease through genome-wide association studies (GWAS). A major effort,

known as the HapMap project, has been made to develop a map of the common patterns of variation in the human genome - its phase 3 release contains 1.6 million SNPs genotyped in samples from 1184 individuals [1]. In addition, the 1000 Genomes project has been undertaken recently to sequence the genomes of at least one thousand individuals from a variety of populations and uncover even more genetic variation [8].

The second most frequently occurring type of genetic polymorphism is the Variable Number Tandem Repeat (VNTR) polymorphism, with over 600,000 candidate VNTRs in the human genome [2]. This type of polymorphism involves short segments of DNA that are repeated a number of times that differs between individuals. As such, VNTRs can have many more possible alleles than the generally biallelic SNPs. Depending on the length of the repeat unit, VNTRs are also sometimes referred to as microsatellites (1-6 basepair motifs) and minisatellites (6-500 basepair motifs), as well as Simple Sequence Repeats or Short Tandem Repeats. Certain properties of repeated motifs affect the likelihood that their numbers will vary between individuals; those with higher copy numbers and percentage match between repeat units are more likely to be polymorphic, while on the other hand, too many insertions and deletions will stabilize the repeat such that it does not mutate frequently during meiosis [2].

Studies to date have found the distribution of VNTRs in genomes to be random and not associated with genes [9]. Though most VNTRs in higher eukaryotes are thought to be neutral in function, their intrinsic properties and high mutability do make them good candidates for mapping and function. A change in the number of repeat units has a great effect on the chemical-physical properties of the DNA sequence, which can lead to changes in gene transcription, splicing and recombination and ultimately variability in protein sequence and structure in the cases where the repeats code for amino acids [2]. In fact, some such functional VNTRs have been

implicated in monogenic human disease, including a variety of neurological disorders that have been attributed to microsatellites such as CAG triplet repeats occurring within transcribed regions [2].

There exist other types of genetic variation, including those referred to as restriction fragment length polymorphisms and copy number variations, but discussion here will be limited to SNPs and VNTRs.

Polymorphisms can be characterized by their minor allele frequency (MAF), or the rate of occurrence of the less common allele in a population. In an ideal state for a population, both allele and genotype frequencies remain constant, or in Hardy-Weinberg equilibrium. In real-world populations, one or more factors disturbing Hardy-Weinberg equilibrium (HWE), such as non-random mating, random genetic drift, limited population size, mutations, and selection are always in effect. Still, the Hardy-Weinberg assumption is useful for modeling, and tests of allele and genotype frequencies in a population for the extent to which they conform to the Hardy-Weinberg principle provide a good way to detect possible genotyping error and to analyze population change.

The methods that have been developed for identifying polymorphisms vary widely in cost and amount of manual effort required. A common way that SNPs are genotyped is by hybridizing complementary DNA probes to the SNP site. Hundreds of thousands of probes can be arrayed on a small chip, allowing for the simultaneous interrogation of a large number of SNPs [26]. Unfortunately, VNTRs are not similarly conducive to genotyping by such high-throughput array-based techniques. Instead, VNTR genotypes must be determined through a more painstaking procedure involving polymerase chain reaction (PCR) and size separation on gel electrophoresis machines, which is neither highly automatable nor parallelizable. As such, the current cost of genotyping a VNTR is approximately \$0.50 per marker, while that for SNPs is only around \$0.001 [2]. It is desirable, then, to consider ways in which we

might use the more readily available information about SNP genotypes to infer or impute VNTR markers of interest in the surrounding genomic regions.

2.2 5-HTTLPR

This work focuses in particular on a VNTR polymorphism in the human serotonin transporter (5-HTT) gene (*SLC6A4*). This genetic variant, referred to as 5-HTTLPR, was identified in the promoter region of the gene on chromosome 17q11.1-q12 in 1995 [14]. The polymorphism results in two common alleles comprised of different numbers of copies of a 20-23 base pair repeat unit: the short (S) variant with 14 copies and the long (L) variant with 16. The relative frequencies of these two alleles vary across populations, with the S allele having a frequency of approximately 0.40 in European populations [12]. Some studies have found the long allele to be associated with higher levels of 5-HTT mRNA in human cell lines [19], while others have failed to detect significant associations between the 5-HTTLPR and 5-HTT availability [25]. On the level of behavior, it has been demonstrated, among other findings, that individuals carrying the S allele have a slightly higher risk for displaying heightened levels of anxiety compared to L homozygotes [19] and that reduced 5-HTT availability is associated with mood disturbances such as major depression [4]. In 2002, Hariri et al. published work in humans pointing to a neural endophenotype as a possible contributor to the increased levels of fear and anxiety observed in S-carriers - an elevated amygdala reactivity to threatening stimuli as assessed by functional MRI in these individuals [13]. Further, Caspi et al. have demonstrated that 5-HTTLPR serves to moderate the environmental influence of stressful life events on depression, where S-carriers exhibit more depressive symptoms in response to stressful events than L homozygotes [5]. Such findings render the serotonin transporter polymorphism a crucial candidate for study in psychology and neuroscience.

2.3 Existing imputation methods

In instances where genotypes of interest are missing or, as in the case of VNTRs, difficult to obtain, investigators often wish to impute these genotypes based on information from others that are present. Doing so can confer many advantages, including increasing the statistical power of a study, facilitating the identification of susceptibility variants to guide further fine mapping, and allowing for meta-analyses that combine studies genotyped on different sets of variants [15]. Augmenting available data with imputed markers is an especially helpful method for increasing power in rare-variant studies [29].

Many approaches to imputing genetic markers have been taken to date, especially in the case of imputing missing SNPs from other SNPs. Certainly a naive approach to inferring an unknown marker would be assigning it the identity of the major allele (the allele occurring with the highest frequency in the population) for that marker. Fortunately, we can do much better than this. Imputation techniques exploiting properties such as linkage disequilibrium and haplotype phasing will be introduced here and discussed in further detail in Chapter 3.

Linkage disequilibrium (LD) is the non-random association of alleles at different positions in the genome, i.e. the co-occurrence of alleles more or less frequently in a population than would be expected by chance based on their frequencies. A key player in the shaping of patterns of LD in a population is the process of genetic recombination, or the breaking and rejoining of DNA strands that allows homologous chromosomes to crossover during meiosis. Other factors influencing LD include rate of mutation, genetic drift, selection, non-random mating, and population structure. LD is generally reported in units of r^2, which reflects the correlation of the alleles at two given sites. When a locus we wish to impute has a strong correlation with another locus, we say that the second locus "tags" the first, and we can impute

missing genotypes at the first locus with those that are known at the second. The ability to tag a locus of interest is not limited to single loci - generally, certain multi-marker haplotypes can achieve higher r^2 with the locus than just a single marker.

Another property that can be readily capitalized on for the imputation of unknown markers is haplotype phasing. The term haplotype refers to a combination of alleles on adjacent locations of a chromosome that are transmitted together (can be up to an entire chromosome). In diploid organisms, determining haplotype phase amounts to resolving which of the alleles in a set of genotypes from a given chromosomal region of interest collectively originated from one chromosome and which similarly group together on the other. In situations where an individual is homozygous at many or all locations, the haplotype may be unambiguous, but for those ambiguous situations that arise from heterozygosity, algorithms have been developed that can make use of a sampling of individuals to determine haplotype phase.

Here we limit the focus to four existing imputation methods, selected based on their previous application to the imputation of 5-HTTLPR, their relative prominence and demonstrated superiority over similar methods, and their suitability for our dataset.

2.3.1 Tagger

One method that utilizes linkage disequilibrium information to search for single- and multi-marker tags for a locus of interest is the one implemented in a program called Tagger as part of the Haploview package published by the Broad institute [6]. The method searches for up to 3 tags in at most 10,000 allelic predictors and includes steps for the prioritization of certain markers in order to increase tagging efficiency. In 2009, Wray et al. used Tagger to identify a 2-marker SNP haplotype to predict the 5-HTTLPR with an r^2 of 0.72 in a sample of 2823 unrelated individuals from the Australian twin registry. The group limited their search to a 38kb region around

5-HTTLPR bounded by rs140700 in intron 6 and rs7214991 in the promoter. All HapMap (build 35) (www.hapmap.org) [11] SNPs with minor allele frequency > 0.1 were either present in their genotyped data and included in the search or represented by single SNP genotyped proxies with $r^2 > 0.8$, resulting in a search space of 13 SNPs. The tagging SNPs identified were rs4251417 and rs2020934, where the CA haplotype for this pair of SNPs coupled with the S allele of 5-HTTLPR [28]. In 2011, the same group broadened their search region to contain SNPs located ∼25kb downstream and ∼155kb upstream of 5-HTTLPR and reported using Tagger to find that though the highest r^2 between any single SNP and 5-HTTLPR was 0.5, they could again identify a 2-SNP haplotype that tagged 5-HTTLPR: rs2129785 and rs11867581, which tagged the short allele with an r^2 of 0.775. This haplotype is preferential to the original due to the fact that rs2020934 of the first is not usually present in standard GWAS arrays or in the HapMap dataset [27]. The authors concluded based on HapMap3 that no further proxies can be added to the tagging haplotypes that they identified.

2.3.2 Vertex discriminant analysis

A more sophisticated approach capitalizing on linkage disequilibrium between 5-HTTLPR and surrounding SNPs was published by Lu et al. in 2012 with a sample of 1852 European whites genotyped for the both VNTR and surrounding SNPs. In this approach, the authors used stepwise linear regression to select 8 SNPs to be put into a machine learning method for classification of 5-HTTLPR into one of the 3 genotype classes: S/S, S/L, or L/L. The 8 SNPs were selected from those in the 24523266 basepair to 26462684 basepair region on chromosome 17 with a minor allele frequency of greater than 0.05, a Hardy-Weinberg equilibrium p-value of greater than 0.05, and a missing genotypes percentage less than 0.01, with redundant SNPs (pairwise $r^2 > 0.8$) removed [22].

The classification method, called vertex discriminant analysis (VDA) has been

shown by its developers to perform better in multicategory prediction than other methods, having the particular advantage that it allows for a nonlinear relationship between the genotypes and the predictors, which provides greater flexibility. The authors found that their 8-SNP proxy achieved an r^2 of 0.85 with 5-HTTLPR and that their classifier was able to achieve 92.8% accuracy.

2.3.3 IMPUTE2

Another prominent and powerful imputation method, IMPUTE2, makes use of additional large reference panels to impute SNPs with higher accuracy (with especially notable gains at rare SNPs) [15]. The hidden Markov model (HMM) based method is designed to take advantage of the ever-increasing availability of additional control data sets genotyped on multiple SNP chips as well as dense genome-wide haplotypes from the 1000 Genomes project. Like most other methods in its class, it goes beyond relying on linkage disequilibrium values simply calculated from allele frequencies; rather, it seeks to phase the study genotypes at the SNPs that are typed in both the study and reference panels and look for matches between those haplotypes and the corresponding haplotypes in the reference panel, assuming that these match at SNPs that are untyped in the study sample. As such, for these methods, the accuracy of phasing in turn affects the accuracy of imputation, and having previous phase information leads to faster imputation. The authors report being able to obtain higher imputation accuracy with their method than with those that do not use information from additional reference panels. No attempt has been made to date to impute 5-HTTLPR from surrounding SNPs using this method.

2.3.4 BEAGLE

The final method to be considered is a method designed for the inference of haplotype phase as well as the imputation of ungenotyped markers in large sets (hundreds of

thousands of markers genotyped in thousands of samples) of unrelated individuals and parent-offspring pairs and trios [3]. The work is again motivated by the current and foreseen rise in availability of much larger reference panels from initiatives such as HapMap phase 3 and the 1000 Genomes project and the observation that other methods such as IMPUTE do not scale well to larger reference panels. The authors investigated the effect of reference panel size on imputation accuracy and found that while the accuracy of estimated allele frequency does increase as the size of the actual study sample increases, this effect is in fact much smaller than that of the reference panel size on imputation accuracy [3]. The authors found that increasing the reference panel size provided notable gains in imputation accuracy particularly at rare SNPs, and that even unphased reference panels could provide highly accurate genotype imputation [3]. A notable advantage of the algorithm is its versatility: it can accept genotype likelihoods (instead of binary calls) when phasing and imputing, be generalized to impute multiallelic VNTRs, and also perform association analysis.

3

Materials and Methods

3.1 Duke Neurogenetics Study

This work is motivated by the Duke Neurogenetics Study, an ongoing study which assesses a wide range of behavioral and biological traits among nonpatient, young adult, student volunteers. The study is run by the Laboratory of NeuroGenetics at Duke University as a part of the laboratory's wider efforts to understand the neurobiology of individual differences in complex behavioral traits by integrating neuroimaging and genotyping technologies to facilitate the study of the links between genes, brain, and behavior. To date, over 470 undergraduates (ages 18-22) have completed the study, which involves a clinical interview and neuropsychological testing, a large battery of computer-based questionnaires, a functional magnetic resonance imaging (fMRI) session, and a saliva sample for DNA analysis.

The initial cohort of approximately 200 participants were genotyped for VNTRs manually as described below. Beyond that point, samples are sent to the direct-to-consumer personal genomics company, 23andMe, where they are genotyped for over 1,000,000 SNPs on the Illumina OmniExpress Plus chip. Samples are no longer

Table 3.1: Racial distribution in the current sample

Race	Have 5-HTTLPR	Missing 5-HTTLPR	Total
Caucasian	28	131	159
African American	12	26	38
Asian	14	80	94
Multiracial	2	18	20
Other	3	6	9
Total	59	261	320

genotyped for microsatellite markers. Counts of all 320 participants genotyped by 23andMe in the current sample are listed in Table 3.1 by race and whether they have 5-HTTLPR genotypes.

3.1.1 DNA Collection and Genotyping

Saliva samples (2 mL) were collected using Oragene kits in accordance with the manufacturers instructions (DNA Genotek Inc., Toronto, Ontario). Upon collection, saliva samples were stored at room temperature before being shipped to 23andMe for DNA extraction and analysis (23andMe Inc., Mountain View, CA). DNA extraction and genotyping were performed by the National Genetics Institute (NGI), a CLIA-certified clinical laboratory and subsidiary of Laboratory Corporation of America. The Illumina Omni Express chip and a custom array containing an additional 300,000 SNPs were used to provide genome-wide data. The Illumina Omni Express chip included 359 SNPs in the 2MB region surrounding the 5-HTTLPR on chromosome 17 (approximately 1MB upstream and 1MB downstream). These SNPs were extracted from the master database using the freely available software plink (http://pngu.mgh.harvard.edu/purcell/plink/). Genotype distributions within the Caucasian race group, as well as within the entire sample, did not deviate from Hardy-Weinberg equilibrium for 351 (p values > 0.08) and 341 (p values > 0.05) of the 359 SNPs, respectively.

Table 3.2: Distributions of 5-HTTLPR genotypes in the sample. (*) 2 subjects with the L/XL genotype were treated as L/L for our analyses.

Genotype	Caucasians only	Entire sample
S/S	7	16
S/L	9	19
L/L	12	24*
MAF (S)	0.411	0.432

Genotyping for 5-HTTLPR was performed at the Duke Center for Human Genetics. Primer sequences for 5-HTTLPR are described by Gelernter et al. [12], the forward primer having the sequence (5'- ATGCCAGCACCTAACCCCTAATGT-3') and the reverse (5'-GGACCGCAAGGTGGGCGGGA-3'). PCR was conducted using the following cycling conditions: initial 15- min denaturing step at 95°C, followed by 35 cycles of 94°C for 30 sec, 66°C for 30 sec and 72°C for 40 sec, and a final extension phase of 72°C for 15 min. Reactions were performed in 10X reaction Buffer IV (ABgene), 1.5mM MgCl2, 50ng of genomic DNA, 5pmols of each primer, 0.3mM dNTPs and 1 unit of Native Taq (Promega). PCR products were subsequently digested by MspI restriction enzyme for 4 hours at 37°C. The digestion products were separated on a 3% agarose gel (MultiABgarose, ABgene) supplemented with Ethidium bromide (0.03%, BDH) and visualised by ultraviolet transillumination. 5-HTTLPR genotypes in our sample did not differ significantly from Hardy-Weinberg equilibrium for the Caucasian group ($p = 0.1337$), however, there was evidence for a deviation in the entire sample ($p=0.0137$). The distributions are given in Table 3.2.

3.2 Modeling

For these analyses, we considered all SNPs on the Illumina chip in a 2MB region surrounding 5-HTTLPR (approximately 1MB downstream and 1MB upstream), as this region encompassed that explored by each of the methods to be investigated.

This range contained a total of 359 SNPs, 24 of which had a minor allele frequency (MAF) of 0 (90 in the Caucasian sample) and so were uninformative and excluded from all analyses.

3.2.1 Tagger

To select the best tag SNPs for an untyped marker, Tagger utilizes the r^2 metric, calculated as follows:

$$r^2 = \frac{(p_{A1B1} - p_{A1}p_{B1})^2}{p_{A1}p_{B1}p_{A2}p_{B2}}$$

It then searches (in at most 10000 allelic predictors) for optimal tags, which could include haplotypes of up to three loci. It prioritizes certain markers for inclusion by ranking candidate tags by the number of other SNPs that they can serve as a proxy for and keeping the top N (an adjustable parameter). The algorithm picks multi-marker haplotype predictors only if they can capture the alleles originally captured by other potential tags at the user-indicated r^2. Further, it requires tags themselves in multi-marker tests to be in strong LD (logarithm of odds score > 3) with the predicted allele. We ran Tagger in our sample with the pairwise tagging option, as well as the option for aggressive tagging with 2- and 3- marker haplotypes. We force excluded 5-HTTLPR and varied the r^2 and LOD thresholds to investigate whether any suitable tags could be found.

3.2.2 Vertex discriminant analysis

The vertex discriminant analysis (VDA) classification approach involves constructing a learning model with a training dataset and later evaluating it with a test set. In general, to classify a response variable y into one of k categories, VDA finds k equidistant points in the R^{k-1} space, which are then used to assign the coordinates of the variable. In the model learning stage, the algorithm seeks to optimize the

following loss function:

$$Loss(A, b) = \frac{1}{n} \sum_{i=1}^{n} ||y_i - A^T x_i + b||_\epsilon + \lambda \sum_{j=1}^{k-1} ||a_j||^2$$

where (A, b) is the $p \times (k-1)$ matrix of regression coefficients for genotype prediction, n is the number of observations, and x is a vector of p predictors (Lu et al. use eight SNPs) with y as their response values (5-HTTLPR genotypes). The loss function is minimized with the majorization-minimization algorithm [7]. The SNP genotypes used as predictors are coded by allele dosage (number of minor alleles) and standardized for each dosage level by their means and variances in the sample.

All 8 SNPs used in this prediction model [22] are present on the Illumina chip used for genotyping our sample. However, 40 out of 320 subjects (17 out of 159 Caucasians) are missing at least one of the 8 due to difficulty genotyping (these counts drop to 23/320 and 10/159 when excluding rs11651241 which had a particularly low genotyping success rate and is being retyped in our sample). Nonetheless, we were able to use the model to predict 5-HTTLPR for our entire sample.

3.2.3 IMPUTE2

IMPUTE2 is an extension of IMPUTE v1, described in Marchini 2007 [23]. The model at the core of these methods is a Hidden Markov Model (HMM), where the hidden states are a sequence of pairs of the known haplotypes (described in more detail below). Briefly, some aspects of the model include the allowance for recurrent mutation at each SNP and the assumptions of 1) a uniform mutation rate across the genome, 2) the independence of mutations across sites, 3) independence between each individuals genotype vector, and 4) genotype sampling from a population in Hardy-Weinberg equilibrium [23].

In order to estimate the marginal posterior probabilities for missing genotypes,

IMPUTE2 employs a Markov chain Monte Carlo (MCMC) framework consisting of alternate phasing and imputation steps theoretically based on an approximation to a coalescent with recombination process. In short, this coalescent theory, based on Markov chain models originally developed by Kingman in the 1980s [17] and generalized to include recombination by Hudson in 1983 [16], allows for the creation of models relating the distribution of sampled haplotypes to the underlying recombination rate as follows:

$$Pr(h_1, ..., h_n|\rho) = Pr(h_1|\rho)Pr(h_2|h_1;\rho)...Pr(h_n|h_1, ..., h_{n-1};\rho)$$

where $h_1, ..., h_n$ represent the n sampled haplotypes and ρ the recombination parameter. Figure 3.1 gives a further illustration from Li et al. of how this can be used to model haplotypes sampled in the current iteration as imperfect mosaics of observed haplotypes [21].

IMPUTE's algorithm begins by initially phasing the study genotype data at random and performing a small number of burn-in iterations of only phasing. It then proceeds with a large number of main iterations where each study individual i is updated in both phasing and imputation steps.

In the phasing step, individual i's observed genotype is phased by sampling from the conditional distribution given the genotype, all of the current guesses of other study haplotypes, all of the markers that are typed in every subject (call this set of markers T) and each of the N known reference haplotypes at SNPs in T. The HMM for this conditional distribution is represented as follows:

$$Pr(G_i|H) = \sum_{Z_i^{(1)}, Z_i^{(2)}} Pr(G_i|Z_i^{(1)}, Z_i^{(2)}, H)Pr(Z_i^{(1)}, Z_i^{(2)}|H)$$

where the haplotypes to be determined are represented by two sequences of hidden states $Z_i^{(1)} = \{Z_{i1}^{(1)}, ..., Z_{i2}^{(1)}\}$ and $Z_i^{(2)} = \{Z_{i1}^{(1)}, ..., Z_{i2}^{(1)}\}$ for L sites where $Z_{il}^{(j)} \in$

$\{1, ..., N\}$. Here, the SNP to SNP transition probabilities are pulled from a fine-scale, population-scaled recombination map for the region of interest, estimated from HapMap phase II data [24]. The emission probabilities are derived from population genetics theory, which gives more weight to those genotypes that are consistent with the local patterns of linkage disequilibrium, allowing for the inclusion of information from all markers in LD with an untyped marker in such a way that their influence decreases with increasing genetic distance from the untyped marker [23].

In the imputation step, the forward-backward algorithm for HMMs is run independently on each haplotype and the marginal posterior probabilities for the missing alleles are determined. When all iterations have completed, the imputed posterior probabilities are summed across iterations and renormalized to arrive at the final decision on the most probable allele.

An additional strength of the algorithm is an increase in computational efficiency afforded by the use of a constraint on the number of haplotypes used in each of the phasing updates. Since computational burden for the phasing step grows quadratically with the number of haplotypes, greater efficiency can be achieved by using only those k (adjustable parameter) haplotypes closest (as determined by Hamming distance) to those of the individual being updated.

IMPUTE2 was run on the current Caucasian sample using the 28 subjects with both 5-HTTLPR genotype and Illumina SNPs as an unphased reference panel and the 131 subjects without 5-HTTLPR genotypes as the unphased study sample to be imputed. The fine-scale recombination map used was that from the 1000 Genomes project for chromosome 17, as recommended by the authors. Chromosomal locations were updated in the sample to match those on the recombination map.

3.2.4 BEAGLE

BEAGLE is designed to infer haplotype phase and impute missing genotypes using an HMM from a broad general class of models which the authors refer to as haplotype HMMs. The method does not explicitly model recombination and mutation for phasing, rather these are captured implicitly [3]. For this model, observed and current guess haplotypes, with or without missing alleles, are considered as sequences of values emitted from the haplotype HMM. The HMM is fitted to the data by an iterative algorithm that alternates between model building and sampling. For initialization, missing genotypes are imputed at random according to allele frequencies and heterozygous genotypes are phased at random. Next, the model-building step works with current estimates of phased haplotypes, proceeding along a chromosome from marker to marker, with each step involving merging and splitting of haplotype clusters.

At the first level of the HMM (corresponding to the first marker), haplotypes are clustered according to the allele at the first marker. For each successive iteration, then, clusters are first merged based on their similarities at the remaining markers along the chromosome and then split on the basis of the allele at the next successive marker. Emission probabilities are all 0/1, since states represent clusters of haplotypes, which all have the same allele at the current marker due to the splitting step. Transition probabilities are based on the ratio of the number of haplotypes in the destination state to the total number of haplotypes in all of its parent states. The iteration then continues to the sampling step, where new haplotypes are sampled for each individual conditional on the genotype data and the current HMM. Final posterior genotype probabilities are calculated by summing the probabilities of the HMM states corresponding to the different genotypes and averaging over multiple iterations for increased accuracy. The method includes measure of imputation accuracy,

which the authors call allelic R^2, that represents the squared correlation between the allele dosage with the highest posterior probability and the true allele dosage - this measure has the advantage that it can be accurately estimated from posterior genotype probabilities without knowledge of the true genotype.

BEAGLE was run in our sample on the unphased genotypes of all subjects, with 5-HTTLPR marked as missing to be imputed where relevant, using the default settings.

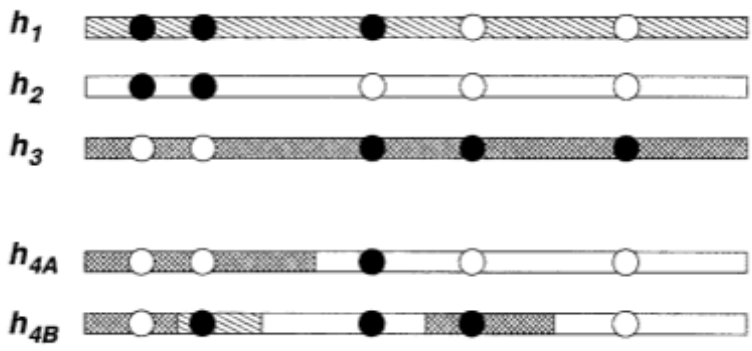


FIGURE 3.1: Illustration from Li et al. of two possible constructions h_{4A} and h_{4B} of imperfect haplotype mosaics by "copying" parts of three observed haplotypes h_1, h_2, and h_3.

4

Results

4.1 Tagger

The highest r^2 with 5-HTTLPR for any marker found by Tagger in the Caucasians only sample was 0.417 for rs2020934 (rs7214014 from Vinkhuyzen et al. came in 6th with an r^2 of 0.344). In the entire sample, the highest r^2 was 0.574 for rs7214014 (rs2020934 was 3rd with an r^2 of 0.543), greater than the 0.50 r^2 with this SNP found by Vinkhuyzen et al. Though the aggressive 2- and 3- marker haplotype tagging option was used, Tagger identified no multi-marker haplotypes in either the Caucasians only or the entire sample. See Figure 4.3 for a linkage disequilibrium matrix comparable to the one in Vinkhuyzen et al.

4.2 Vertex discriminant analysis

In the 59 subjects for whom 5-HTTLPR genotype data was available, the 8-SNP VDA method was able to correctly assign 5-HTTLPR genotype classes for 50 subjects (85%). Considering Caucasians only, the percentage correct rose to 93% (26 out of 28). See Table 4.1 for a breakdown of genotype assignments. Note that out of the 17 subjects that were missing at least one of the 8 SNPs, 13 (76.5%) were still

Table 4.1: Counts for genotypes imputed by VDA verses actual genotypes. Counts for the Caucasians only sample given in parentheses.

Imputed\Actual	S/S	S/L	L/L	L/XL
S/S	13 (6)	1 (0)	1 (0)	0 (0)
S/L	3 (1)	16 (9)	2 (1)	0 (0)
L/L	0 (0)	2 (0)	19 (11)	2 (0)

correctly classified, while 37 out of the 42 (88.1%) with genotype calls for all 8 SNPs were correctly classified. There was no evidence for deviation from HWE for the predicted genotypes in either the Caucasian group (p=0.3009) or the entire sample (p=0.252).

4.3 IMPUTE2 and BEAGLE

Since IMPUTE2 and BEAGLE were both used to impute the untyped 5-HTTLPR using the subjects with typed 5-HTTLPR as a reference panel, we can only assess accuracy by comparing the calls made by these two methods and VDA for those subjects missing 5-HTTLPR genotypes. As both methods output a probability for each of the three possible genotype groups, for our purposes each subject was simply assigned the genotype with the highest probability. Out of the 131 Caucasians in the sample with no 5-HTTLPR genotypes, 106 (81%) were assigned to the same genotype by all three methods, and all 131 were assigned to the same genotype by at least two out of the three methods. Assigning genotypes, then, based on the most common call results in 24 S/S, 64 S/L, and 43 L/L, for a minor allele frequency of 0.416.

Pairwise agreement between the three methods was as follows: VDA and BEAGLE, 108 subjects (82%); VDA and IMPUTE2, 121 subjects (92%); and IMPUTE2 and BEAGLE, 114 subjects (87%). See Figure 4.2 for a Venn diagram illustrating these overlaps and Figure 4.3 for an illustration of the concordance in specific

genotype calls made between the three methods. Figure 4.4 displays the confidence reported by IMPUTE2 and BEAGLE for the most probable genotypes for those subjects where all methods were in agreement and for those where at least one differed.

	Marker	MAF	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15	16	17	18	19	20	21	22	23	24	25
1	rs4583306	0.393	.	1.00	1.00	1.00	1.00	0.53	1.00	0.50	0.85	1.00	1.00	1.00	1.00	0.59	0.88	0.88	0.88	0.75	0.79	0.71	1.00	1.00	0.77	0.54	0.81
2	rs140700	0.089	0.06	.	1.00	1.00	1.00	1.00	1.00	0.35	0.56	0.21	0.21	0.17	1.00	1.00	0.47	0.47	0.47	1.00	1.00	0.34	0.08	0.08	0.59	0.47	1.00
3	rs6354	0.179	0.14	0.45	.	1.00	1.00	0.21	1.00	0.17	0.78	0.37	0.37	0.26	1.00	0.69	0.72	0.72	0.72	1.00	1.00	0.32	0.41	0.41	0.60	0.27	1.00
4	rs2020936	0.179	0.14	0.45	1.00	.	1.00	0.21	1.00	0.17	0.78	0.37	0.37	0.26	1.00	0.69	0.72	0.72	0.72	1.00	1.00	0.32	0.41	0.41	0.60	0.27	1.00
5	rs2066713	0.375	0.39	0.06	0.13	0.13	.	1.00	1.00	0.85	0.37	0.49	0.49	0.63	1.00	1.00	0.60	0.60	0.60	0.25	0.64	0.49	0.47	0.47	0.41	0.18	0.86
6	rs4251417	0.107	0.05	0.01	0.00	0.00	0.07	.	0.21	1.00	1.00	1.00	1.00	0.61	0.23	1.00	1.00	1.00	1.00	1.00	1.00	0.83	0.66	0.66	1.00	1.00	1.00
7	rs8071667	0.179	0.14	0.45	1.00	1.00	0.13	0.00	.	0.17	0.78	0.37	0.37	0.26	1.00	0.69	0.72	0.72	0.72	1.00	1.00	0.32	0.41	0.41	0.60	0.27	1.00
8	5HTTLPR	0.411	0.23	0.01	0.00	0.00	0.31	0.08	0.00	.	0.75	0.67	0.67	1.00	1.00	1.00	0.78	0.78	0.78	0.66	0.71	0.73	0.30	0.30	0.44	0.57	0.64
9	rs1487971	0.393	0.31	0.05	0.20	0.20	0.13	0.08	0.20	0.25	.	1.00	1.00	0.53	1.00	1.00	0.91	0.91	0.91	0.91	0.79	0.23	0.61	0.61	0.67	0.46	0.75
10	rs7214248	0.321	0.31	0.01	0.06	0.06	0.19	0.06	0.06	0.15	0.73	.	1.00	1.00	1.00	1.00	0.89	0.89	0.89	0.88	0.86	0.25	0.56	0.56	0.41	0.36	0.82
11	rs1050565	0.321	0.31	0.01	0.06	0.06	0.19	0.06	0.06	0.15	0.73	1.00	.	1.00	1.00	1.00	0.89	0.89	0.89	0.88	0.86	0.25	0.56	0.56	0.41	0.36	0.82
12	rs6505167	0.107	0.08	0.00	0.04	0.04	0.08	0.01	0.04	0.08	0.05	0.06	0.06	.	0.68	0.95	1.00	1.00	1.00	0.51	1.00	0.47	0.28	0.28	0.61	1.00	1.00
13	rs11651241	0.088	0.05	0.01	0.02	0.02	0.12	0.00	0.02	0.05	0.08	0.06	0.06	0.30	.	0.23	1.00	1.00	1.00	1.00	1.00	1.00	0.80	0.80	1.00	1.00	1.00
14	rs2129785	0.125	0.08	0.01	0.02	0.02	0.09	0.84	0.02	0.10	0.09	0.07	0.07	0.02	0.00	.	1.00	1.00	1.00	1.00	1.00	1.00	0.00	0.00	1.00	1.00	1.00
15	rs7214014	0.446	0.40	0.03	0.14	0.14	0.26	0.10	0.14	0.34	0.67	0.46	0.46	0.15	0.10	0.12	.	1.00	1.00	0.69	0.91	0.60	0.53	0.53	0.61	0.36	0.90
16	rs11867581	0.446	0.40	0.03	0.14	0.14	0.26	0.10	0.14	0.34	0.67	0.46	0.46	0.15	0.10	0.12	1.00	.	1.00	0.69	0.91	0.60	0.53	0.53	0.61	0.36	0.90
17	rs8073378	0.446	0.40	0.03	0.14	0.14	0.26	0.10	0.14	0.34	0.67	0.46	0.46	0.15	0.10	0.12	1.00	1.00	.	0.69	0.91	0.60	0.53	0.53	0.61	0.36	0.90
18	rs3110454	0.446	0.29	0.12	0.27	0.27	0.05	0.10	0.27	0.24	0.67	0.46	0.46	0.04	0.09	0.12	0.47	0.47	0.47	.	0.91	0.56	0.66	0.66	0.74	1.00	0.90
19	rs3110095	0.5	0.40	0.10	0.22	0.22	0.24	0.12	0.22	0.35	0.40	0.35	0.35	0.12	0.08	0.14	0.67	0.67	0.67	0.67	.	1.00	0.56	0.56	0.68	1.00	1.00
20	rs3110093	0.214	0.09	0.04	0.01	0.01	0.11	0.02	0.01	0.10	0.02	0.04	0.04	0.10	0.30	0.04	0.12	0.12	0.12	0.11	0.27	.	1.00	1.00	1.00	1.00	1.00
21	rs6505179	0.286	0.26	0.00	0.09	0.09	0.15	0.02	0.09	0.03	0.23	0.26	0.26	0.02	0.03	0.00	0.14	0.14	0.14	0.22	0.13	0.11	.	1.00	1.00	1.00	1.00
22	rs4533339	0.286	0.26	0.00	0.09	0.09	0.15	0.02	0.09	0.03	0.23	0.26	0.26	0.02	0.03	0.00	0.14	0.14	0.14	0.22	0.13	0.11	1.00	.	1.00	1.00	1.00
23	rs8066222	0.321	0.18	0.07	0.17	0.17	0.13	0.06	0.17	0.06	0.33	0.17	0.17	0.09	0.05	0.07	0.22	0.22	0.22	0.32	0.22	0.13	0.84	0.84	.	1.00	1.00
24	rs887469	0.143	0.03	0.13	0.06	0.06	0.01	0.02	0.06	0.04	0.06	0.05	0.05	0.02	0.02	0.02	0.03	0.03	0.03	0.21	0.17	0.61	0.07	0.07	0.08	.	1.00
25	rs3794794	0.463	0.52	0.09	0.20	0.20	0.38	0.12	0.20	0.35	0.31	0.27	0.27	0.11	0.08	0.15	0.56	0.56	0.56	0.55	0.86	0.25	0.33	0.33	0.40	0.15	.

FIGURE 4.1: Linkage disequilibrium (r^2 below diagonal, D' above) with 5-HTTLPR for 24 SNPs in Vinkhuyzen et al.

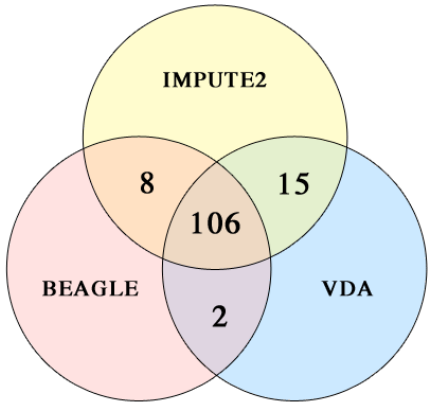


FIGURE 4.2: Venn diagram illustrating counts of overlap between 5-HTTLPR imputed genotype calls by IMPUTE2, BEAGLE, and VDA methods in the Caucasians only sample (N=131).

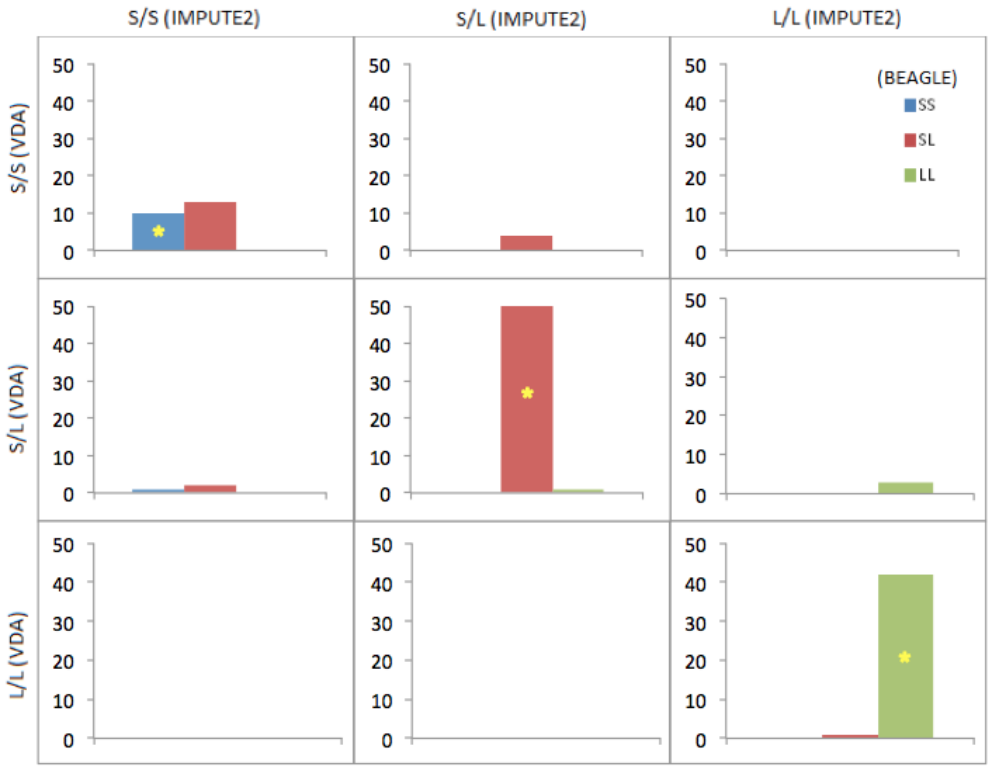
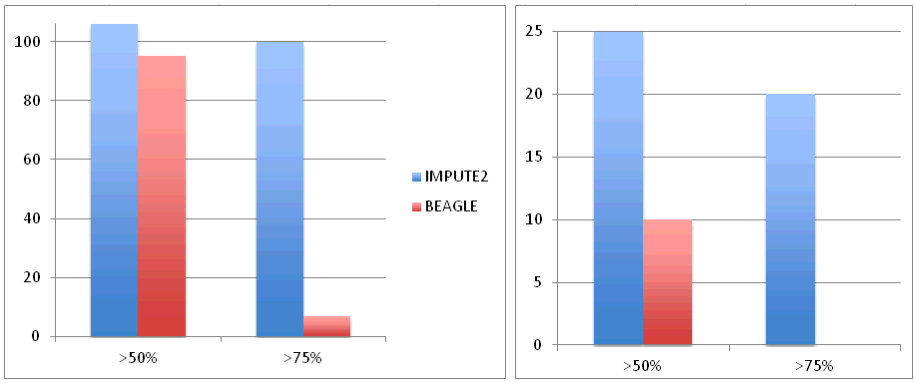


FIGURE 4.3: Concordance between 5-HTTLPR genotypes imputed with IMPUTE2, BEAGLE, and VDA methods by genotype call in the Caucasians only sample. Rows correspond to genotype calls made by VDA, columns to IMPUTE2, and individual bars to BEAGLE. Heights of bars represent the number of subjects classified by the combination of the three methods corresponding to the position in the chart. The three bars corresponding to cases where all three methods made the same call are indicated with a yellow *.



(a) (b)

FIGURE 4.4: Confidence reported by IMPUTE2 and BEAGLE for those 106 subjects where all methods were in agreement (a) and for those 25 where at least two methods made differing genotype calls (b).

5

Discussion

Perhaps the most encouraging result in the investigation of these various methods for the imputation of 5-HTTLPR from SNP data is the correct identification (as determined by manual genotyping) of 5-HTTLPR genotypes in 93% of the Caucasian subjects in our current sample with the vertex discriminant analysis method by Lu et al. This percentage is comparable to the 92.4% accuracy that the authors of the method were able to achieve on their test set, but given that one of the key SNPs in the model is sporadically missing in our sample, we can certainly be optimistic about the possibility for improvement upon solving this problem. Another promising finding is the fact that for all 131 Caucasian subjects missing 5-HTTLPR genotypes, at least two out of the three imputation methods investigated made the same genotype call, resulting in a reasonable minor allele frequency for the imputed genotypes.

Certainly a major issue faced by this investigation and most such imaging genetics studies of its kind is the small sample size. This does not pose a problem for the VDA method, as it was trained by its developers on a larger sample of similar ethnicity. Since we found that it performed nearly as well in our sample as in tests by its

authors, it presents a promising avenue to continue down. Still, if we wish to have the flexibility for imputing other markers in other ethnic groups, it is important to consider whether the methods can perform in smaller samples and how this limitation can be overcome. Naturally we expect to obtain more accurate imputation estimates as we collect and genotype more samples. Another way to increase sample size in our study would be to allow for the inclusion of subjects of other races by accounting for the effects of population stratification.

Population structure and its effects on imputation accuracy is a primary concern for any imputation method. Differences in LD patterns between the study sample and the reference panel (often HapMap) are likely to reduce the accuracy of imputation. Further, when conducting tests of association, whether imputed or genotype data are used, population structure within a study sample can result in false-positives [23]. Lu et al. also express the concern that patterns of LD in non-White populations could render their 8-SNP model less effective in predicting 5-HTTLPR. Indeed, we did find that in accuracy of the VDA method was slightly lower in the non-Caucasians in our sample. In addition, care must be taken when assessing imputation methods across different studies, as factors such as SNP density and the similarity of LD patterns between the study data and reference populations will affect imputation accuracy [23]. However, it should be noted that imputation may still work well if there is population structure within a study sample, since imputation depends on haplotypes similar to those in the study sample occurring in reference sample used. The authors of IMPUTE suggest that accurate imputation with their model can be extended to less homogeneous studies by including and conditioning on different reference panels (i.e. different HapMap populations) and including a model of ancestry [10]. Finally, there exist approaches for dealing with population structure for genotyped SNPs which should work just as well when applied to imputed markers [23]. Statistics such as the CMAT developed by Zawistowski et al. allow for the

incorporation of qualitative covariates and can thus correct for confounders such as population stratification [29]. Future work including items such as these could allow for the use of more of the available data and more accurate and confident imputation. This becomes especially important as we move on to use imputed genotypes in studies of association with phenotypes of interest.

Bibliography

[1] D. M. Altshuler and R. e. a. Gibbs. Integrating common and rare genetic variation in diverse human populations. *Nature*, 467(7311):52–8, Sept. 2010.

[2] G. Breen. Practical Informatics Approaches to Microsatellite and Variable Number Tandem Repeat Analysis. *Methods in Molecular Biology*, 628, 2010.

[3] B. L. Browning and S. R. Browning. A unified approach to genotype imputation and haplotype-phase inference for large data sets of trios and unrelated individuals. *American journal of human genetics*, 84(2):210–23, Feb. 2009.

[4] A. Caspi, K. Sugden, T. E. Moffitt, A. Taylor, I. W. Craig, H. Harrington, J. McClay, J. Mill, J. Martin, A. Braithwaite, and R. Poulton. Influence of life stress on depression: moderation by a polymorphism in the 5-HTT gene. *Science (New York, N.Y.)*, 301(5631):386–9, July 2003.

[5] A. Caspi, K. Sugden, T. E. Moffitt, A. Taylor, I. W. Craig, H. Harrington, J. McClay, J. Mill, J. Martin, A. Braithwaite, and R. Poulton. Influence of life stress on depression: moderation by a polymorphism in the 5-HTT gene. *Science (New York, N.Y.)*, 301(5631):386–9, July 2003.

[6] P. I. W. de Bakker, R. Yelensky, I. Pe'er, S. B. Gabriel, M. J. Daly, and D. Altshuler. Efficiency and power in genetic association studies. *Nature genetics*, 37(11):1217–23, Nov. 2005.

[7] J. De Leeuw and W. Heiser. Convergence of correction matrix algorithms for multidimensional scalings. *Geometric representations of relational data*, pages 735–752, 1977.

[8] R. M. Durbin, D. L. Altshuler, G. A. R. Abecasis, D. R. Bentley, A. Chakravarti, A. G. Clark, F. S. Collins, and E. Al. A map of human genome variation from population scale sequencing. *Nature*, 467(7319):1061–1073, 2011.

[9] J. Epplen, W. Maueler, and E. Santos. On GATAGATA and other junk in the barren stretch of genomic desert. *Cytogenet Cell Genet*, 80:75–82, 1998.

[10] D. Falush, M. Stephens, and J. K. Pritchard. Inference of population structure using multilocus genotype data: linked loci and correlated allele frequencies. *Genetics*, 164(4):1567–87, Aug. 2003.

[11] K. Frazer, D. Ballinger, D. Cox, D. Hinds, L. Stuve, G. RA, and et al. A second generation human haplotype map of over 3.1 million SNPs. *Nature*, 449(7164):851–61, Oct. 2007.

[12] J. Gelernter, H. Kranzler, and J. F. Cubells. Serotonin transporter protein (SLC6A4) allele and haplotype frequencies and linkage disequilibria in African- and European-American and Japanese populations and in alcohol-dependent subjects. *Human genetics*, 101(2):243–6, Dec. 1997.

[13] A. R. Hariri, V. S. Mattay, A. Tessitore, B. Kolachana, F. Fera, D. Goldman, M. F. Egan, and D. R. Weinberger. Serotonin transporter genetic variation and the response of the human amygdala. *Science (New York, N.Y.)*, 297(5580):400–3, July 2002.

[14] A. Heils, A. Teufel, S. Petri, M. Seemann, D. Bengel, U. Balling, P. Riederer, and K. P. Lesch. Functional promoter and polyadenylation site mapping of the human serotonin (5-HT) transporter gene. *Journal of Neural Transmission*, 102(3):247–254, Oct. 1995.

[15] B. N. Howie, P. Donnelly, and J. Marchini. A flexible and accurate genotype imputation method for the next generation of genome-wide association studies. *PLoS genetics*, 5(6):e1000529, June 2009.

[16] R. R. Hudson. Properties of a neutral allele model with intragenic recombination. *Theoretical population biology*, 23(2):183–201, Apr. 1983.

[17] J. Kingman. The coalescent. *Stochastic Processes and their Applications*, 13(3):235–248, Sept. 1982.

[18] R. Laing, P. Hess, Y. Shen, J. Wang, and S. Hu. The role and impact of SNPs in pharmacogenomics and personalized medicine. *Curr Drug Metab*, 12(5):460–86, 2011.

[19] K. Lesch, D. Bengel, A. Heils, S. Sabol, B. Greenberg, S. Petri, and et al. Association of anxiety-related traits with a polymorphism in the serotonin transporter gene regulatory region. *Science*, 274:1527–1531, 1996.

[20] S. Levy, G. Sutton, P. C. Ng, L. Feuk, A. L. Halpern, B. P. Walenz, N. Axelrod, J. Huang, E. F. Kirkness, G. Denisov, Y. Lin, J. R. MacDonald, A. W. C. Pang, M. Shago, T. B. Stockwell, A. Tsiamouri, V. Bafna, V. Bansal, S. a. Kravitz, D. a. Busam, K. Y. Beeson, T. C. McIntosh, K. a. Remington, J. F. Abril, J. Gill, J. Borman, Y.-H. Rogers, M. E. Frazier, S. W. Scherer, R. L.

Strausberg, and J. C. Venter. The diploid genome sequence of an individual human. *PLoS biology*, 5(10):e254, Sept. 2007.

[21] N. Li and M. Stephens. Modeling Linkage Disequilibrium and Identifying Recombination Hotspots Using Single-Nucleotide Polymorphism Data. *Genetics*, 2233(December):2213–2233, 2003.

[22] A. T.-H. Lu, S. Bakker, E. Janson, S. Cichon, R. M. Cantor, and R. a. Ophoff. Prediction of serotonin transporter promoter polymorphism genotypes from single nucleotide polymorphism arrays using machine learning methods. *Psychiatric genetics*, pages 1–7, Apr. 2012.

[23] J. Marchini, B. Howie, S. Myers, G. McVean, and P. Donnelly. A new multipoint method for genome-wide association studies by imputation of genotypes. *Nature genetics*, 39(7):906–13, July 2007.

[24] S. Myers, L. Bottolo, C. Freeman, G. McVean, and P. Donnelly. A fine-scale map of recombination rates and hotspots across the human genome. *Science (New York, N.Y.)*, 310(5746):321–4, Oct. 2005.

[25] A. Patkar, W. Berrettini, P. Mannelli, R. Gopalakrishnan, M. Hoehe, L. Bilal, and et al. Relationship between serotonin transporter gene polymophisms and platelet serotonin transporter sites among African-American cocaine-dependent individuals and healthy volunteers. . *Psychiatric Genetics*, 14:25–32, 2004.

[26] R. Rapley and S. Harbron, editors. *Molecular Analysis and Genome Discovery.* John Wiley & Sons, Ltd, Chichester, UK., 2004.

[27] a. a. E. Vinkhuyzen, T. Dumenil, L. Ryan, S. D. Gordon, a. K. Henders, P. a. F. Madden, a. C. Heath, G. W. Montgomery, N. G. Martin, and N. R. Wray. Identification of tag haplotypes for 5HTTLPR for different genome-wide SNP platforms. *Molecular psychiatry*, pages 1–2, June 2011.

[28] N. R. Wray, M. R. James, S. D. Gordon, T. Dumenil, L. Ryan, W. L. Coventry, D. J. Statham, M. L. Pergadia, P. a. F. Madden, A. C. Heath, G. W. Montgomery, and N. G. Martin. Accurate, Large-Scale Genotyping of 5HTTLPR and Flanking Single Nucleotide Polymorphisms in an Association Study of Depression, Anxiety, and Personality Measures. *Biological psychiatry*, 66(5):468–76, Sept. 2009.

[29] M. Zawistowski, S. Gopalakrishnan, J. Ding, Y. Li, S. Grimm, and S. Zöllner. Extending rare-variant testing strategies: analysis of noncoding sequence and imputed genotypes. *American journal of human genetics*, 87(5):604–17, Nov. 2010.