COMPARISON OF GENOME WIDE LINKAGE AND ASSOCIATION ANALYSIS METHODS FOR A QUANTITATIVE TRAIT IN COMPLEX EXTENDED PEDIGREES

-

A CASE STUDY OF MUSICAL APTITUDE

Pro Gradu

Jaana Oikkonen

University of Helsinki

Department of Biosciences

Biotechnology

6/2012

Abbreviations

ANOVA	Analysis of variance
AP	Absolute pitch
DTT	Distorted tunes test
FBAT	Family based association test
g	Genotype
HMM	Hidden Markov model
IBD	Identity by descent
JPSGCS	Java Programs for Statistical Genetics and Computational Statistics package
KMT	Karma musical aptitude test
λ	Variance inflation factor
λ_R	Sibling relative ratio
LD	Linkage disequilibrium
Lod	Logarithm of odds
MCMC	Markov Chain Monte Carlo
P	Likelihood
PCA	Principal components analysis
Ph	Phenotype
PPL	Posterior probability of linkage
Pr	Probability
Q-Q plot	Quantile-quantile plot
QTDT	Quantitative transmission/disequilibrium test
SD	Standard deviation
SNP	Single nucleotide polymorphism
SPT	Seashore pitch test
STT	Seashore time test
TDT	Transmission disequilibrium test
θ, Theta	Recombination fraction
T-Lod	Theta lod
Var	Variance
VC	Variance component
VIF	Variant inflation factor, PLINK

Variables in equations

g	Genetic or genotype
P	Likelihood
Ph	Phenotype
Pr	Probability
Var	Variance

Table of Contents

1 Introduction

Genome wide linkage and association methods are used to map genes affecting traits with genetic predisposition. In this thesis, I compare methods suitable for quantitative trait mapping in complex, extended pedigrees. As a case study, gene-mapping study of musical aptitude is performed with these methods.

Gene mapping methods are introduced in the Section 2. In the methodology, I have concentrated on family-based methods. The most important linkage and association methods are discussed with notes about performing such analyses. In the next section, musical aptitude is introduced in scientific context. Theories of musical aptitude relevant to this study are introduced, as well as tests used to measure musical aptitude. There are only few genetic studies that have been published concerning musical aptitude. Third section covers also them.

Section 4 presents the aim of the study and Section 5 presents the actual outline performing the study. The results are described and discussed in the last sections. In the discussion, the results concerning gene mapping of the musical aptitude are compared between different programs and from other studies. The programs themselves are discussed in respect with usability and reliability.

2 Background

2.1 Complex trait gene mapping

Gene mapping is a process used to locate genes of particular interest. For example, it can be used to search for genes that cause some disease. It has been successfully used to map genes that cause for example Huntington's diseases, Salla disease and lactose intolerance (OMIM database). These diseases are Mendelian, single-gene diseases. But the majority of traits are not Mendelian (Risch 2000) and thus, primary interest in gene mapping has moved from simple Mendelian traits to complex traits. Most common diseases and common features are complex traits with complex inheritance patterns (Plomin et al. 2009).

The difference between simple Mendelian traits and complex traits is the effect of genotypes. While genotypes cause Mendelian diseases, in complex traits genotypes only alter the probability of the disease, usually together with environmental factors.

Here, I call these traits complex, but many names have been used: polygenic, multifactorial and multigenic traits. By calling them complex traits, I am trying to underline the complexity of the traits also outside genetic effects. Complex traits are affected by many genes, but also other effects may occur to make them complex. Environmental factors alone can cause a trait to be complex, even though the genetic effect would be monogenic.

Genetically a trait may be complex due to locus and allelic heterogeneity, pleiotropy or incomplete penetrance (Risch 2000). Locus heterogeneity means that the causative locus genotype of a trait can be different between individuals. In allelic heterogeneity, different alleles in one gene can influence the trait similarly. In pleiotropy, the gene influences multiple phenotypes and different genotypes can thus differently affect separate aspects of the trait. In the case of incomplete penetrance, only some of the carriers of a mutation have an effect on the trait. In that case, it is hard to discover the true effect of the mutation.

Also interactions between separate effects may vary. Epistasis is interaction between genes, where one or several genes modify the effect of other genes. Also complex interactions between genes and the environment may occur.

Evidently, complex traits are different from each other and methodology should be applicable to different situations (Risch 2000). Many methods have been developed for these situations and they are most powerful when applied in a correct setting. For example, the genes for common and rare traits can be found with different methods (see Section 2.4). Also, locus heterogeneity does not cause problems for every method (Risch 2000).

Gene mapping tries to localize susceptibility genes in genome and thus find the cause of a trait. Gene mapping techniques depend on finding areas of genome that are shared by affected individuals more often than is expected by chance. Two most commonly used gene-mapping methods are association and linkage. In this thesis I focus on these two methods.

The linkage method studies the tendency of two loci to be inherited together. It utilizes pedigrees to find an association between a trait and a genetic locus. The association method studies non-random association of alleles: whether an allele or haplotype is overrepresented in affected individuals. Typically, this is done by an unrelated population sample, but also families may be studied. These methods are discussed in detail in Sections 2.5 and 2.6.

Most genetic studies have been performed using qualitative traits (Plomin et al. 2009). Qualitative traits are usually diseases or some other traits, where cases can be clearly separated from controls. In some cases, this division into two parts has not been enough to explain the trait. If genetic effects in the background are complex, a qualitative trait does not include enough information and a quantitative trait should be used instead (Plomin et al. 2009).

However, this question about qualitative and quantitative traits is also a question of study design (Terwilliger and Göring 2009). If extreme cases are collected, a qualitative study is appropriate. If subjects are collected randomly from the population, the number of extreme cases in the sample will usually be too limited and the quantitative approach should be applied.

For genetic mapping, there are two options, a genome wide scan or a candidate gene study. In a candidate gene study, genes of interest are chosen based on prior assumptions. Results from previous studies with similar or related phenotypes or known expression patterns of genes can be taken into account. In contrast, genome wide scans pass through the whole genome. With either method, a genome or some part of it can be sequenced or genotyped using polymorphisms. The genome wide methods are discussed with more detail in the following section.

2.2 Genome wide methods

The technological improvements in genotyping in the last decade have made it possible to perform massive genome-wide analyses. Especially the vast development of marker chip technologies has changed our tools to perform gene-mapping studies. Now, thousands of loci throughout the genome can be genotyped in a limited time. Typically, hundreds of thousands of SNPs can be used. It has also become possible to sequence the whole genome instead of genotyping only some markers, but I will focus here on marker-based approaches.

In the genome wide genetic mapping, the inheritance pattern of a trait is compared against the inheritance pattern of a chromosomal region. The affecting genes are mapped to intervals that are as small as possible. The markers are chosen with prior knowledge of their frequencies in the population to cover most of the genome. The causative polymorphism is not supposed to be included in the marker set, but the aim is, that nearby markers would tag the site. The markers are then tested for association or linkage.

Due to the multiple testing over a large number of markers a high false-positive rate in genome wide methods has become a problem. When almost a million tests are made with any data, there will, inevitably, be a large number of false positives. For example, at a nominal significance level of 0.05, this means there are 50,000 p-values under 0.05 just by chance. Thus, genome wide thresholds for nominal p-values have been developed to keep the genome wide false positive rate tolerable (Ott and Hoh 2000). Also, it has been proposed that results should be replicated to confirm them before announcing them to be definitive (Lohmueller et al. 2003).

2.3 Measures of quantitative genetics

There are several approaches to identify whether a trait is inherited. The methods are usually based on the comparison of relatives and non-relatives. If a trait is inherited, the relatives should be more similar than non-relatives. Rough estimations can be made from the known family history of the trait (King et al. 1984). Methods for estimating the inherited component more precisely include heritability and sibling relative ratio.

Heritability describes the proportion of the trait variability due to genetic effects:

$$H^2 = \frac{Var(g)}{Var(ph)}, \quad (1)$$

where Var(g) is the genetic variance and Var(ph) is the phenotypic variance (Dempster and Lerner 1950). Heritability in the narrow sense includes only the additive genetic effect:

$$h^2 = \frac{Var(a)}{Var(ph)}, \quad (2)$$

where Var(a) is additive genetic variance. The narrow h^2 is used, when epistatic, maternal or paternal effects are to be excluded (Dempster and Lerner 1950).

When studying animals, the heritability can be directly measured by controlling environmental effects. In human studies, heritability is estimated indirectly. The narrow-sense heritability can be estimated from twin or family data (King et al. 1984). Heritability can be estimated from twins either by comparing mono- and

dizygotic twin pairs or by comparing monozygotic twins who have lived in different environments. In family studies, the correlation between trait values is compared to the relationship between the individuals. If the trait is inheritable, close relatives like siblings should resemble each other more than distant relatives.

Heritability is not comparable across populations, because it is conditional on genetic and environmental factors of a population (King et al. 1984). Also differences in measurements of the phenotype affect it greatly and therefore heritability is indicative, not absolute.

Sibling relative ratio (λ_R) tells how much bigger is the disease risk for those who have an affected sibling, than the risk for the general population (Risch 2000). It can be written as:

$$\lambda_R = \frac{risk\ in\ siblings}{risk\ in\ general\ population}. \quad (3)$$

Risk ratios can be defined for any kind of relatives and any value over 1 is considered as evidence for a heritable component. It has to be noted, that the values are relative and cannot be compared between diseases with different prevalence. Obviously, sibling relative risk can only be used for qualitative traits.

2.4 Common and rare variants

Traditional gene-mapping studies have typically identified genetic variants with large effect on the trait. Such variants are rare (<1% in population), because selective pressure is strong against them. As for complex traits, there are supposedly several variants with modest, or even minor, effects (Iles 2008). Unlike variants with large effects, variants with smaller effects can be common (>5% in population) (Lander 1996). The selective pressure is spread over several genotypes for multigenic traits, which makes the pressure weaker and enable the markers to be common. Whether or not the variants influencing a complex trait are actually common or rare has been debated over the last decade.

It has been shown that common variant hypothesis can cause the variability that has been seen in common complex traits (Peng and Kimmel 2007). There are also examples of common variants causing complex disease. One of the best-known examples is APOE gene in late-onset Alzheimer's disease, where the second most common allele is associated with the disease (Saunders et al. 1993). On the other hand, it has also been speculated that many of common allele associations may result from many causative rare alleles (Cirulli and Goldstein 2010). This line of thought was raised by the discovery that even though highly probable associations for disease has been found, causative variants that would alter the gene expression have rarely been identified. The association can result from diluted signal from rare causative alleles or common allele can be modifier for causative alleles, rare or not (Cirulli and Goldstein 2010).

Much of the discussion has been about the methodology. Linkage analysis is best suited for rare variants with large effects, and association analysis will best recover common alleles (Iles 2008). Thus, the debate has also been about proving one method better than the other. But it may be more like Manolio et al. (2009) put it: neither of the methods can explain a complex trait fully, and there is still missing heritability that cannot be found with either of these methods. For example rare alleles with only modest or small effects cannot be found with any method (Manolio et al. 2009). Anyhow, common and rare alleles may both contribute to complex trait variability and hence both methods can reveal true associations between the trait and genetic variants.

2.5 Linkage analysis

Linkage analysis searches for a co-segregation of a marker and the trait of interest, in the context of a sampling unit. A sampling unit can be a sib-pair or a family of arbitrary structure. Here, I will concentrate on methods for families, particularly on extended, multigenerational families.

The linkage method analyse the tendency of two loci to be inherited together. If two loci are near each other, recombination between them during meiosis will only rarely happen and they are said to be linked. Linkage analysis correlates the genotypes of two loci. To perform analysis between locus genotypes and a trait phenotype, the trait needs to be converted into inferred trait locus genotypes (Terwilliger and Göring 2009). In Mendelian traits, where there are only two alleles, the trait phenotypes are virtually the same as the inferred locus genotypes. Continuous traits are more difficult to convert.

A linkage analysis can be performed separately between individual markers and a trait (two-point analysis) or as a multipoint analysis, where adjacent markers and the trait are analysed together. In the following chapters, the linkage methods are introduced in more detail.

2.5.1 Parametric and non-parametric linkage methods

In parametric methods, the mode of trait inheritance needs to be specified. This mode is used to convert the phenotypes into inferred trait-locus genotypes. Inferred locus can then be used to search linkage for real genotypes. Non-parametric methods approach the problem from the other direction: the modes of marker inheritance are evaluated from the data and these modes are used to search correlation for the phenotype. Thus, the non-parametric method uses real inheritance models from the genotypes and the trait inheritance mode need not be specified. This is why non-parametric methods are also called "model-free" methods.

The linkage method is based on Fisher's likelihood inference theory (Fisher 1918). The relationship between phenotype and genotypes can be written as:

$$\begin{aligned} P(Ph|G_M) &= \sum_{g_D} P(Ph|g_D)P(g_D|G_M) \\ &= \sum_{g_D}\sum_{g_M} P(Ph|g_D)P(g_D|g_M)P(g_M|G_M) \end{aligned} \quad (4)$$

(Terwilliger and Göring 2009), where Ph are the observed phenotypes, G_M are the observed genotypes, g_D are the underlying disease-locus genotypes and g_M are the underlying marker genotypes. This equation defines whether the genotypes are

independent of the phenotypes. Information about linkage is contained in the term $P(g_D|G_M)$, in the relationship of the underlying disease-locus genotype and the observed genotypes (Terwilliger and Göring 2009). The null hypothesis for the underlying genotypes is

$$P(g_D, g_M) = P(g_D)P(g_M). \quad (5)$$

This null hypothesis says that underlying genotypes and phenotypes are not linked and thus, the likelihoods are independent. The reason for the deviation from the null hypothesis can be linkage or linkage disequilibrium (Terwilliger and Göring 2009). Linkage disequilibrium means non-random association of alleles of two or more loci and it is better discussed in Section 2.6.

Parametric (or model-based) methods evaluate the likelihood directly from the Equation 4. Thus, the parameters for the equation need to be clarified. Clarification has to include the parameters to calculate the $P(Ph|g_D)$ and $P(g_D|g_M)$. The trait-locus genotypes (g_D) are converted from the known phenotypes (Ph) considering at least allelic influences on phenotype, allele frequencies and penetrance. The computation for $P(g_D|g_M)$ needs trait genotype frequencies and the recombination fraction. These parameters are estimated from the assumed inheritance model, which must be inferred from the known data.

The two main algorithms for parametric analysis are those of Elston-Stewart (Elston and Stewart 1971) and Lander-Green (Lander and Green 1987). Most of the modern methods are based on either of these methods (Ott and Hoh 2000). The Elston-Stewart algorithm goes up the family tree and accounts for every possible genotype of an individual conditioning on parental genotypes and descendants' phenotypes. Elston-Stewart based methods are more suitable for large pedigrees and they have also been extended for general pedigrees (Lange and Elston 1975). The Elston-Stewart based methods for extended families are sometimes called the Lange-Elston methods, but here I will use the name of the original algorithm.

An advance in the Lander-Green method was the ability to analyse large numbers of markers (Lander and Green 1987). The original Elston-Stewart method and

many extensions are capable of analysing only a handful of markers within large pedigrees. Contrary to the Elston-Stewart, the Lander-Green algorithm factorizes the likelihood calculation by markers instead of individuals. Thus, it is capable to analyse large number of markers, but only simple families.

Non-parametric (or model-free) methods estimate the reverse likelihood $P(G_M|Ph)$ directly (Terwilliger and Göring 2009). Most of them are based on the same Elston-Stewart and Lander-Green algorithms as parametric methods. The question is, whether subjects with the same trait status share more marker alleles than expected, otherwise $\frac{P(G_M|Ph)}{P(G_M)}$. Degrees of freedom in equation G_M increase with the pedigree size (Terwilliger and Göring 2009). Thus, the exact non-parametric method is not possible for extended pedigrees, because the computation would be too complex, and inexact methods should be used instead.

Other non-parametric methods, like the variance component (VC) method, are anyhow possible for extended pedigrees. VC is reasoned by an idea that phenotypically similar relatives should share more likely alleles at affective locus (Hopper and Mathews 1982). The VC method calculates within pedigree values for pairs of individuals as a measure of shared alleles at that locus and background genetic similarity (Amos 1994). Measures from different families are summed and genetic parameters are maximized over all families (Amos 1994). With this method, polygenic, environmental and monogenic effects for trait variability can be estimated (Amos 1994). This method is implemented in the software SOLAR (Almasy and Blangero 1998).

Parametric methods may have more power than the non-parametric methods (Kruglyak et al. 1996). However, the parametric method is prone to misspecifications in its genetic model: if the model is misspecified the power may even be lower than with non-parametric methods (Kruglyak et al. 1996). The two-point parametric method is less sensitive to misspecifications of model parameters and thus, the problem is more substantial with multipoint methods. The two-point method has, anyhow, its own flaws. It is sensitive to misspecifications of allele frequencies and to false positives (Kruglyak et al. 1996). However, the power does

depend greatly on the data and different methods are most powerful in different situations (Terwilliger and Göring 2009).

2.5.2 Lod score methods

The results of the linkage methods, parametric or non-parametric, are usually summarized as logarithm of odds (lod) scores. Alternatively also location scores could be used (Lathrop et al. 1984), but they were more useful before the human genome was published and now lod scores are usually employed. A lod score is defined by:

$$LOD(\theta) = log_{10} \frac{P(X|\theta \leq 0.5;\omega)}{P(X|\theta = 0.5;\omega)} \quad (6)$$

(Haldane and Smith 1947; Hauser and Boehnke 1998), where θ is the recombination fraction and ω is the genetic model for trait. This means that a lod score is the ratio of the likelihood of the observed data given linkage divided by the null hypothesis where the recombination fraction (θ) is 0.5 where the loci are unlinked. θ is probability of two loci to recombine during meiosis and varies in usual situations from 0 (totally linked loci) to 0.5 (unlinked loci) (Morton 1955). The distribution of two-point lod scores is a curve, which usually gains its highest score at some point between $0 < \theta \leq 0.5$ and is always 0 at $\theta = 0.5$ (Haldane and Smith 1947). Examples of possible lod score curves can be seen in Figure 1. The value of θ that maximizes lod score is thought to be the "true" value of theta (Haldane and Smith 1947).

In multipoint analysis, the lod score is defined similarly to equation 6 as the likelihood ratio of existing linkage divided by the null hypothesis, where the disease locus is not on the map (Kruglyak et al. 1996). Multipoint lod scores vary from positive values indicating linkage to negative values that oppose linkage.

Families can be employed in linkage analysis either separately or as pooled samples. If families are analysed separately, the lod scores from different families are summed to construct lod scores for the whole set of pedigrees (Morton 1955). Lod scores are maximized over theta values over the whole sample. In pooled

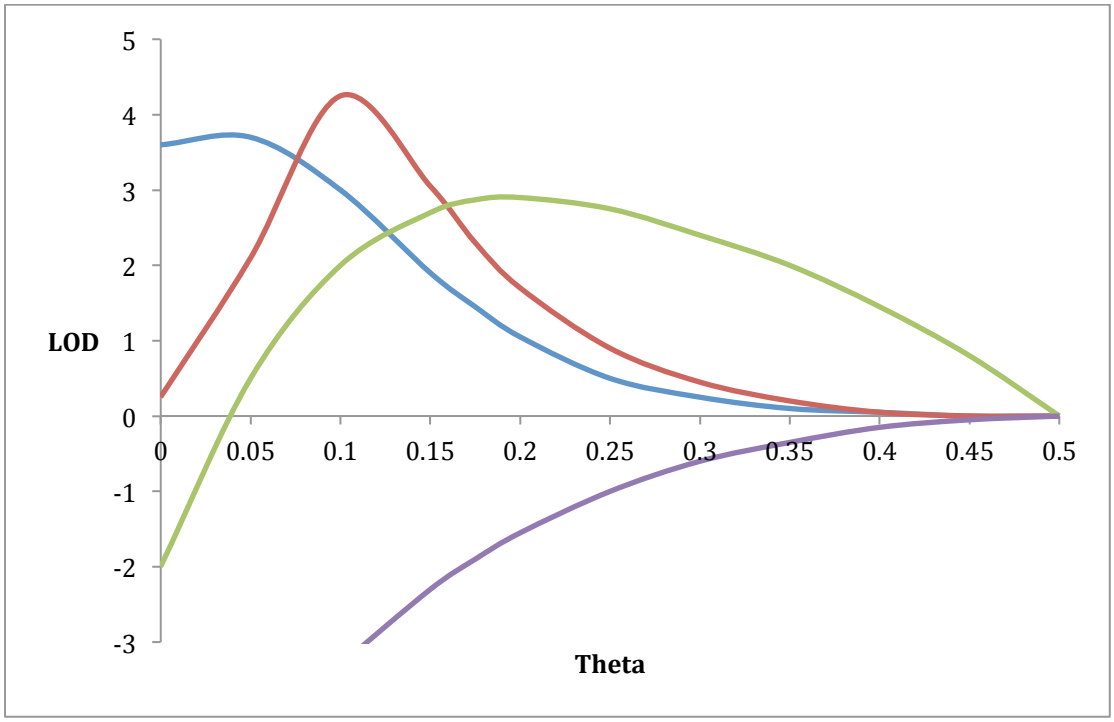


Figure 1 Example of lod score distributions over recombination fraction (θ, theta). A lod score is always 0 when **θ** is 0.5 and gains the highest score at some point between $0 < \boldsymbol{\theta} \leq \mathbf{0.5}$. The value of **θ**, where the lod score have highest value, is used as true value of θ for that locus. For example, the red curve gains highest point at **θ**=0.1 where the lod is 4.25 and the lilac curve at θ=0.5 where the lod is 0.

analysis, all the families are analysed together The pooled analysis is more sensitive for heterogeneity than separate analysis, but considering separate analysis as maximization, the pooled analysis have lower false positive rates

(Weeks et al. 1990). Even though less sensitive to heterogeneity, also separate analysis needs to be combined with heterogeneity analysis to overcome genetic heterogeneity. Multipoint analysis is more sensitive to heterogeneity.

There has been some support for thresholds for significant lod score (for example Morton 1955; Chotai 1984; Lander and Kruglyak 1995; Morton 1998). The thresholds are conditional on the analysis, for example sib-pair analysis and extended family analysis have different values (Lander and Kruglyak 1995). The usual corrections correct for multiple testing in similar ways as Bonferroni correction with some additional factors (Lander and Kruglyak 1995). A widely used threshold in genome wide mapping studies has been a lod of 3.0 (Morton 1955, Morton 1998). This has been presented for two-point studies with the

parametric method. In complex studies, this is however too liberal. Even though, the denser marker maps used today make the tests conditional on adjacent markers and thus the Bonferroni correction is too conservative (Freimer and Sabatti 2004). Lander and Kruglyak (1995) have proposed lod score of 3.3 as a threshold for significant results for complex traits. Morton (1998) says the 3.0 lod score is sufficient, if the power is good and parameters have not been maximized (additionally to θ). Ott and Hoh (2000) claim that the width of the lod score peak might be more important than just the height. Also, lower values might still include true effects.

Lod score methods are widely used for the linkage mapping. Argumentation about significant thresholds has not hindered the use of the lod scores. However, Bayesian statisticians have argued that traditional statistics is maybe not optimal for linkage studies. In traditional statistics, the performed tests raise the possibility of false positives. It is not suited for testing if we should continue to collect larger sample or to try different phenotypes. In contrast, Bayesian methods may be applied to such situations. Thus, Bayesian methods are also proposed for linkage studies. These are covered in the following chapter.

2.5.3 Bayesian linkage methods

Bayesian statistics focus on probability. Probability is understood as degree of belief contrary to the frequency of some event (Savage 1961). Bayesian statistics do not rely on significance level (p-value), and produces more subjective, but direct, probability measure (Smith 1959). A probability scale goes from 0 to 1, where 0 stands for no evidence and 1 stands for proof. Bayesian statistics rely on the Bayes Theorem, which defines the relationship between conditional probabilities. It can be defined as:

$$\Pr(\mathrm{Y}|\mathrm{X}) = \frac{\Pr(X|Y)*\Pr(Y)}{\Pr(X)}, \qquad (7)$$

where X is event that can occur in some way (or reason) Y and we are interested if Y is true (Bayes 1763). This tells how probable it is that X has happened because of Y (considering the probability of Y itself) compared to every reason for X (including Y). This comparison between other options is central in Bayesian

statistics. Even though Y might be unlikely, it can be true if other reasons for the data are even more unlikely (Smith 1959).

In genetics, the posterior probability of linkage (PPL) is the most important paradigm from Bayesian theorem (Vieland 1998). The PPL base for Equation 7 is that X is the genetic data and Y is the model of inheritance. PPL is considered as a model free linkage method and it approximates true likelihood of linkage for every locus.

PPL does utilize prior parameters that may affect the results. However, most significant prior parameters have a distribution of θ, which is also considered in traditional linkage (Smith 1959). Inside one chromosome, a uniform distribution is usually used for θ between 0 and ½, this distribution was originally implemented in the traditional linkage method (Morton 1955).

In PPL, evidence for and against linkage is combined from families sequentially (Vieland 1998). Information about already analysed families is used as prior distribution for the subsequent analyses. This Bayesian framework has been implemented into the package KELVIN (Vieland et al. 2011b). In the program, heterogeneity between subsets of a sample can be considered through sequential update strategy that allows difference between the subsets (Vieland et al. 2011a).

2.5.4 Computational burden in linkage analysis

In genome-wide marker studies the demand for computational resources is not as vast as in sequencing studies. Anyhow, even with these studies the computational burden sets limits for the possible equations that can be used. There are many things that affect the computational time. In the data there are four main things that can increase computational burden. Firstly, the number of loci in the study directly impacts the amount of tests that have to be calculated. Secondly, the number of alleles in each of the loci affects the calculation of that particular locus. The last two aspects are the size and complexity of the pedigrees. Complexity of a pedigree rises especially with untyped individuals (Lange and Sobel 1991). For

example, if parents are genotyped, the inheritance mode of a trait is known, but with missing parents, all possible modes need to be considered, this makes the analysis more complex.

The chosen algorithm greatly affects the computational burden. Lander-Green methods are more suitable for large amount of markers, but computational time grows exponentially with the number of people in a pedigree (Thomas et al. 2000). On the contrary, Elston-Stewart will slow exponentially by the number of markers and also missing data will slow it down substantially (Thomas et al. 2000).

There are also methods that combine features from both of these algorithms. For example, a graphical method combining features from Lander-Green and Elston-Stewart has been introduced. It is included in the package JPSGCS named McLinkage (the older version was McLink) and it allows for large amounts of markers and large pedigrees (Thomas et al. 2000). With this graphical model, the computational time grows linearly with the size of unlooped pedigrees and with the number of markers (Thomas et al. 2000).

In the MORGAN program, the two algorithms are also combined (Tong and Thompson 2008). Here, the pedigree is analysed in segments conditioning on older generations; segments are analysed with Lander-Green and Elston-Stewart is used to generate information over total pedigree. This makes it possible to analyse large families with large number of markers.

Computational burden can be eased with inexact methods (Lange and Sobel 1991). Most of the inexact methods use the Markov chain Monte Carlo (MCMC) algorithm. MCMC algorithms are repeated random sampling methods that approximate target distribution (Lange and Sobel 1991). For example JPSGCS, GeneHunter, MORGAN and SimWalk2 programs utilise MCMC. However, even among these programs there are great differences: SimWalk2 would analyse data for weeks, when the same data could be analysed in only hours with JPSGCS (Service et al. 2006).

Another possibility is to process the data to reduce computational burden. One option is to split the pedigrees. This can though weaken the linkage signals (Gagnon et al. 2003), but is the only way to perform exact multipoint analysis. For example, Lin et al. (2005) showed that there was only a limited effect on the results when the pedigrees were split. However, Terwilliger and Göring (2009) are strongly against the splitting. They say that by breaking the families one "violates every rule of good statistical practise".

Other option for reducing the burden is to reduce the number of markers. In the grid tightening strategy, simple methods are used to prune the markers to be used in more complex methods. For example, a two-point method with relaxed parameters can be used to prune out markers with no evidence of linkage, and then the chosen markers can be re-analysed with stricter, and more complex, methods.

Markers are usually SNPs or microsatellites. Microsatellites are more informative for each locus, but SNPs can be used in higher spatial density to gain the same information content (Ulgen and Li 2005). In linkage analysis LD between SNPs may become a problem. Most linkage algorithms assume that there is no LD between markers, but especially with SNPs it is hard to get rid of all the LD without losing a lot of information with the markers. In the Merlin package, the algorithm allows LD between markers (Abecasis et al. 2002). The allowance is for one solution; another solution would be to use haplotype blocks from SNPs (Service et al. 2006). These blocks could then be used as markers to perform the analysis. The problem with this latter solution is that there is no direct algorithm to get the haplotypes for genome wide data (Service et al. 2006). Thus, the haplotype alternative will work only for restricted areas.

2.6 Association analysis

Association analysis looks for a direct association between the trait and the locus. It searches for differences in allele frequencies between the different values for a

trait. While the linkage disequilibrium between markers limits the amount of markers that can be used in a linkage study, all markers can be used in association analysis. Thus, association is suitable for high-resolution gene mapping. Association is a population-based method and no families are needed.

Association analysis works best in a situation where common variant causes common phenotype as noted in Section 2.4. In contrast to linkage analysis, the genetic cause for the trait needs to be the same between individuals. However, the power to detect association between a trait and a genotype should be higher than in linkage analysis (Risch 2000). In practice, this means that markers with a smaller effect on phenotypes can be detected with the same sample size.

Traditional case-control association analysis relies on the Fisher exact test or the Armitage trend test (also called the Cochran-Armitage or proportion trend test) when testing single SNP against the trait (Balding 2006). These tests search for differences between cases and controls in proportions of different genotypes. For a complex trait, the Armitage test is more powerful, because it detects additive risks more powerfully (Balding 2006). Linear regression, analysis of variance (ANOVA) or score test for association (similar to Armitage test) can be used to test quantitative traits (Balding 2006). In association analysis, every locus is tested against the trait resulting in hundreds of thousands of tests with genome wide analysis.

2.6.1 Family-based association methods

Though traditional population-based association does not and cannot use families, newer approaches do utilize them to ensure that samples are appropriately matched (Spielman et al. 1993). Since family members are not independent, normal association tests are not suitable for families. First family-based tests were designed for trios (parents with affected child), but these methods have been extended to be used with large families as well. Family-based methods are generally called linkage-disequilibrium tests (LD).

Family-based methods can be used to ensure population homogeneity, as association methods are usually sensitive to population stratification (Spielman et al. 1993). Family data might reduce the power compared to population data, but in some cases, for example with rare diseases, the power can even be higher (Laird and Lange 2006). In any case, families are a robust strategy for dealing with population stratification (Bacanu et al. 2002).

For qualitative trait, there are several options. The transmission disequilibrium test (TDT) is widely used and applicable for a qualitative trait (Spielman et al. 1993). Originally it was developed for trios, but it has been widen to use for pedigrees (Martin et al. 2000). The original version explores alleles that are transmitted to affected child from heterozygous parents (Spielman et al. 1993). Heterozygous parents are needed to know the phase of the alleles. The extended versions calculate the associations over the complete sample from associations within individual families (Martin et al. 2000). TDT has little power if association and linkage do not co-exist (Terwilliger and Göring 2009). Other methods for the qualitative trait association with pedigrees include regression and likelihood-based methods (Lange et al. 2002). Population based quantitative trait methods are usually based on general linear models.

For quantitative traits, family-based association tests (FBATs) are developed from population-based methods for qualitative traits: TDT and regression (Lange et al. 2002). The TDT-based methods for a quantitative trait are score tests that test covariance of transmission of alleles and trait values (Lange et al. 2002). The regression-based methods model association between a trait and alleles by regression. These methods account for families by within- and between-families association testing (Abecasis et al. 2000; Lange et al. 2002).

The quantitative transmission disequilibrium test (QTDT) is a regression-based method (Abecasis et al. 2000). QTDT is maximum likelihood based and it can utilize large families by variance components analysis (Abecasis et al. 2000).

Another common program is FBAT, which is a TDT based method (Laird and Lange 2006). It can analyse large pedigrees, but it splits them into nuclear families.

Large families cause computational burden in FBATs similarly to the problem found in linkage studies. Exact likelihood in the family-based method is intractable when using extended pedigrees. Thus, direct family-based methods are preferred in candidate gene studies in contrast to genome wide association studies. However, family structure can also be corrected by population stratification correcting methods, which are computationally less demanding when using large families. Population stratification methods are suitable for genome wide mapping and they are discussed in the following section.

2.6.2 Population stratification methods

Population heterogeneity can affect the results in association study. If a sample consists of subpopulations, and the marker frequencies are different for these groups, the difference can result in spurious associations (Devlin and Roeder 1999). The reason for population stratification can be due to ancestry or separation (Price et al. 2006). Population stratification can be taken into account with several designs in association studies. Here, I will introduce four common methods.

Genomic control corrects for population stratification using the variance inflation factor (λ) (Devlin and Roeder 1999). λ is calculated from a random set of markers. It can be written as

$$\lambda = \frac{Median(T^2)}{Median(\chi^2)}, \quad (8)$$

where T^2 is the empirical distribution of test statistics and χ^2 is assumed chi-squared distribution (Devlin and Roeder 1999). If the frequency of genotypes is the same for cases and controls and there is no population stratification, the test statistics follow the χ^2 distribution (Balding 2006). Markers must be randomly chosen and not assumed to associate with the trait. If population stratification exists, λ is greater than 1, otherwise it is 1. Genomic control adjusts the test

statistics for every marker with this uniform factor, even though stratification would not affect all the markers (Price et al. 2006). This genomic control can also be used in family-based studies (Devlin and Roeder 1999). In family-based studies, λ values can also be less than 1 due to a more similar background than expected.

Structured association splits the sample into a specified number of subpopulations on the basis of genotype differences (Balding 2006). Association is tested on every population cluster (Price et al. 2006). Structured association is computationally more demanding than the genomic control method. The number of subpopulations affects the division greatly and the real subpopulations may be different from the assigned groups (Balding 2006). The structured association method is included in the programs STRUCTURE and ADMIXTURE (Price et al. 2010).

Principal component analysis (PCA) is added to the structured association in the EIGENSTRAT method (Price et al. 2006). In this method, PCA is used to separate a sample into different ancestral groups. These groups are then adjusted to create matched cases and controls (Price et al. 2006). This method should result in more robust subpopulations than the structured association alone. When PCA is used to correct family structure, it may reduce the power or even lead to false grouping, because the samples are correlated but assumed to be uncorrelated (Price et al. 2010).

Population stratification can also be corrected in a more direct way with large number of markers. Kinship methods evaluate the relatedness between every pair of individuals from genotypes and the test statistics are corrected for these relatedness rates (Balding 2006). This kinship method is usually called the mixed models method. The PCA method can be included in mixed models to correct for ancestral groups (Price et al. 2010). These methods are computationally demanding, but recent progress has made the method possible for genome-wide use. The mixed models approach has been implemented in programs EMMAX, TASSEL (Price et al. 2010) and GenAbel (Aulchenko et al. 2007). The mixed models approach, especially with PCA, is the comprehensive approach to correct for family structure (Price et al. 2010). The mixed models approach with PCA is

computationally less demanding for extended families than traditional association methods and genetic relationships are considered as well as more complex population stratification in the sample.

2.6.3 Multiple testing and significance

In genome wide association analysis, as many as millions of tests may be performed. This huge amount of tests causes the number of false positive results to be large. As mentioned in Section 2.2, the nominal significance level of 0.05 means 50,000 false positives by chance when performing one million tests. The traditional p-value thresholds of 0.01 or 0.05 are thus too liberal. One common way to overcome this problem is to use Bonferroni correction. With Bonferroni correction, we can assume that a p-value of $<5*10^{-7}$ would be assigned for a significant result (Freimer and Sabatti 2004). As association results are usually expressed on a logarithmic scale, this threshold corresponds to values over 7 on a logarithmic scale.

However, the markers, and therefore tests, are not usually independent of each other. Thus, Bonferroni correction is usually too conservative with a dense marker map (Freimer and Sabatti 2004). Linkage disequilibrium tells about the dependence between markers and the more there is LD between markers, the more the tests are dependent on each other. Thus, LD between markers does help the problem of multiple testing and lower (logarithmic) values can also be considered to be significant.

One better way to estimate the significance levels would be permutation tests (Clarke et al. 2011). In these tests, the genome-wide significance levels are estimated through permutations. They are conditional on the data and the results cannot be generalized in any other setting (Clarke et al. 2011). They are also computationally demanding, which can make the Bonferroni correction more appealing.

The statistical confidence (and the power of the test) depends greatly on sample size (Freimer and Sabatti 2004). Too small sample sizes increase the possibility of false positive findings even though the possibility of finding any association is decreased. Cardon and Bell (2001) suggest sample sizes from 1,000 to 10,000. After that, even larger sample sizes have been suggested. The reproducibility of the association findings has not been very high and too low sample sizes can be one cause for that.

Also, quality control is important for keeping the false positive and false negative rate low (Anderson et al. 2010). If some markers violate the assumptions of the tests, for example Hardy-Weinberg equilibrium, they raise the error rates. Thus, markers need to be pruned prior to testing and test results should also be checked. Results are usually checked to see if they follow the expected distribution via a Quantile-Quantile-plot (QQ-plot) (McCarthy et al. 2008). In the QQ-plot, values of observed test statistics are compared against values from expected distribution (usually Chi square distribution). The values are compared by pairing observed and expected values that have the same fraction of values below them (otherwise values that are on the same quantile). Deviations from the expected values usually give information about uncorrected differences between cases and controls. Though, the significant results should deviate from the expected distribution as a sharp curve at the high end of the plot if they are true positive values.

The problem of multiple testing can be helped, but not prevented with these methods. Gene mapping results are always probabilistic and even high likely association results may prove to be wrong positive results. This applies also for linkage results. Thus, additional proof is usually needed to confirm gene-mapping results. Duplicate studies, candidate gene studies and functional studies can be used to confirm results.

3 Musical aptitude

3.1 Theories of musical aptitude

Musical ability can be understood in multiple ways. A person may have an ability to understand and perceive differences in pitch, rhythm, timbre, intensity, harmonies or the structures of music. Mainly, the theories of musical aptitude can be divided into two categories: theories of specific, separable capacities and theories of general musical ability (Shuter-Dyson and Gabriel 1981). In the capacity based definition, musical ability is seen as a combination of specific, separable capacities (Shuter-Dyson and Gabriel 1981). Carl Seashore represents this aspect. He includes the recognition of pitch, loudness, time, timbre and musical memory into these special capacities that are needed in musical aptitude (Seashore 1938).

The more general view of musical aptitude concentrates more on the meaning of music and understanding it (Shuter-Dyson and Gabriel 1981). It does not deny that capacities can be separable, but they are thought to be dependent on each other (Shuter-Dyson and Gabriel 1981). In this view, musical aptitude is not constructed from these specific capacities, but they can even obscure the musical aptitude (Wing 1941).

In this thesis I rely on the theory of auditory structuring, a concept described by Kai Karma (2007). This theory belongs to the more general view of musical aptitude. The theory explains music as a combination of melody, harmony and rhythm, as a structured sound. Musical aptitude is understood as the ability to hear patterns in sets of sounds, i.e., auditory structuring (Karma 2002). The theory can also be extended to ability to recognize any kind of patterns, without regard to hearing (Karma 1994).

3.2 Tests of musical aptitude

Musical aptitude is a complex phenotype and measurements include only a part of the concept. Several tests have been demonstrated for testing different aspects of musical ability. One of the earliest standardised tests was Carl Seashore's battery of tests for basic capacities for music (Kirchhubel 2003). The second version of these tests was published in 1939. It includes subtests for different sensory capacities like pitch discrimination and tonal memory (Kirchhubel 2003). Seashore himself said that scores from different subtests should not be totalled, but the reliability of the tests is better if they are summed (Shuter-Dyson and Gabriel 1981).

Herbert Wing's test batteries are also commonly used (Coon and Carey 1989). They include 3 tests for aural capacities and 4 tests for appreciation of music (Kirchhubel 2003). Compared to Seashore's tests, Wing (1941) tries to capture the aesthetic understanding of music with his tests. The reliability and validity of Wing's tests are better than those of Seashore's (Shuter-Dyson and Gabriel 1981). However, Wing's tests are too hard to understand for children and better suited for discrimination of the best from the good ones (Shuter-Dyson and Gabriel 1981).

Wing's and Seashore's tests are relatively short and planned for group testing (Shuter-Dyson and Gabriel 1981), which makes them suitable for research use. The shortness impacts the reliability that is clearly better in Gordon's musical aptitude profile test that lasts for three days (Shuter-Dyson and Gabriel 1981). Even though this longer test is not suitable for genetic studies, it can be used to estimate the differences in results from different tests. It has been shown that all these three tests correlate quite highly and thus, measure somewhat similar capacities (Shuter-Dyson and Gabriel 1981).

Kai Karma has published a test for auditory structuring (Karma 2007). While Karma's tests for musical aptitude are only available in Finnish, no meta-analysis exists on its performance against more common tests. Anyhow, the reliability has been estimated to be from 0.6 to 0.8 (Shuter-Dyson and Gabriel 1981), which is

higher than expected from the amount of items (40) in the test. It measures one's ability to recognise patterns and also musical memory is needed in the test (Shuter-Dyson and Gabriel 1981). The test has been used for example in the brain imaging study of Tervaniemi et al. (1997) and it is also used in music education in Finland.

More specialized tests have also been introduced. Children can be tested with tests formed from Wing's tests (Bentley's test from 1966) or Gordon's test for primary measures (Shuter-Dyson and Gabriel 1981). Pitch recognition can be tested separately with Distorted Tunes test (DTT) (Drayna et al. 2001). In that test, violations of scale structures have to be recognized from popular melodies. This test, however, is culture-bounded, because popular melodies are used.

These tests and theories represented here are developed to measure inborn musical abilities and specifically to measure receptive musical skills. However, musical abilities may also be understood more widely and to include acquired characteristics. For example, Levitin (2012) has proposed testing of any kind of musical ability, acquired or inborn. However, the simpler receptive tests are easier to utilize in genetic studies, especially in family studies where children and adults are both tested. With these simple tests, we might cover only a narrow aspect of the musical ability, which need to be kept in mind when interpreting the results of musical ability studies.

In this study, we have tried to capture a slightly wider perspective of musical abilities by using a combination of different receptive tests. We therefore use Seashore's tests for pitch and time with Karma's auditory structuring test. Supposedly, these three tests measure slightly different aspects of musicality.

3.3 Genetics of musical aptitude

The heritability of musical aptitude has been estimated with many twin studies. Vandenberg (1962) studied musicality tests from Seashore and Wing in a

psychological twin study of 82 twin pairs. He found no significant heritability for pitch tests, but found heritability of 52% for Seashore's rhythm test and 42% for Wing's memory test. More recent studies have showed high heritability for pitch recognition (Pulli et al. 2008). Drayna et al. (2001) estimated heritability to be 0.71 for DTT in a study of 142 twin pairs.

Also extreme phenotypes of pitch recognition have been studied genetically. On the low performing end of variation, there is amusia, otherwise tone deafness (Ayotte et al. 2002). People with amusia cannot separate frequencies, but have no other cognitive impairment or hearing loss. Absolute pitch can be seen as the other end of the variation. Individuals with absolute pitch can name a note of particular pitch without reference (Baharloo et al. 1998).

Amusia was studied by Ayotte et al. 2002. They tested amusia with a reformed DTT. Individuals with amusia had problems recognizing pitches and most of them also on recognizing rhythm. All the subjects had some musical education and a high level of general education. In another study, amusia was studied genetically with an online test (Peretz et al. 2007). Sibling relative risk was estimated in that study as 10.8, which means that there is heritable component also in amusia.

The studies about absolute pitch (AP) are not conclusive. The prevalence of absolute pitch is hard to estimate for the nature of the phenotype. Acquiring AP requires musical training or other tone training in a critical period in childhood (Deutsch et al. 2006). In addition, in early 2000 the studies on perfect pitch based on questionnaires that were later proved to be unreliable (Athos et al. 2007). More recent studies have anyhow tried to conquer these problems. Different tests for AP are demonstrated (Athos et al. 2007, Deutsch et al. 2006) and differences between tone language speakers and non-tone language speakers are known (Deutsch et al. 2006).

There is very limited number of published gene studies about musical ability. Theusch et al. (2009) made a genome-wide linkage study of AP based on a web test. They found linkage (lod score 3.5) in mixed European population sample of

45 families to chromosome 8q. A non-parametric method from the Merlin program was used in that study.

In our group, previous gene mapping study with Karma's musical aptitude test and Seashore's test for pitch and time revealed significant linkage (lod score 3.3) at 4q22 (Pulli et al. 2008). The data consisted of 15 Finnish families with 205 members. It was hypothesised that there is a major locus for musical aptitude. Heritability for the combined score from the tests was estimated to be 0.48 (Pulli et al. 2008). The high lod score with a rather small number of subjects was unexpected.

4 Aims of the study

The aim of this study is to find suitable methods for genome wide gene mapping study with a complex trait and with extended families. The second objective is to perform gene-mapping study of musical aptitude, to replicate and redefine our previous findings (Pulli et al. 2008).

5 Materials and methods

In this study, genome wide linkage and association study is performed with population-based sample. Subjects for this study were collected from the Finnish population between 2003-2011. The study material consists mostly of families.

Previous studies have shown heritability for musical aptitude to be around 50% (Shuter-Dyson and Gabriel 1981). Our pilot study made it clear that it is possible to find linkage for our definition of phenotype (Pulli et al. 2008). Before extending the study materials, power simulations were performed for additional data in linkage analysis (Figure 2). They showed that large effect loci could be found with a sample size of 1000. Simulations also showed that with larger families we gain more power compared to trios when the total sample size remains the same. Thus, we decided to collect more families, focusing on larger families. Also, because the larger families are so powerful, the analysis methods need to be chosen to be appropriate for larger families.

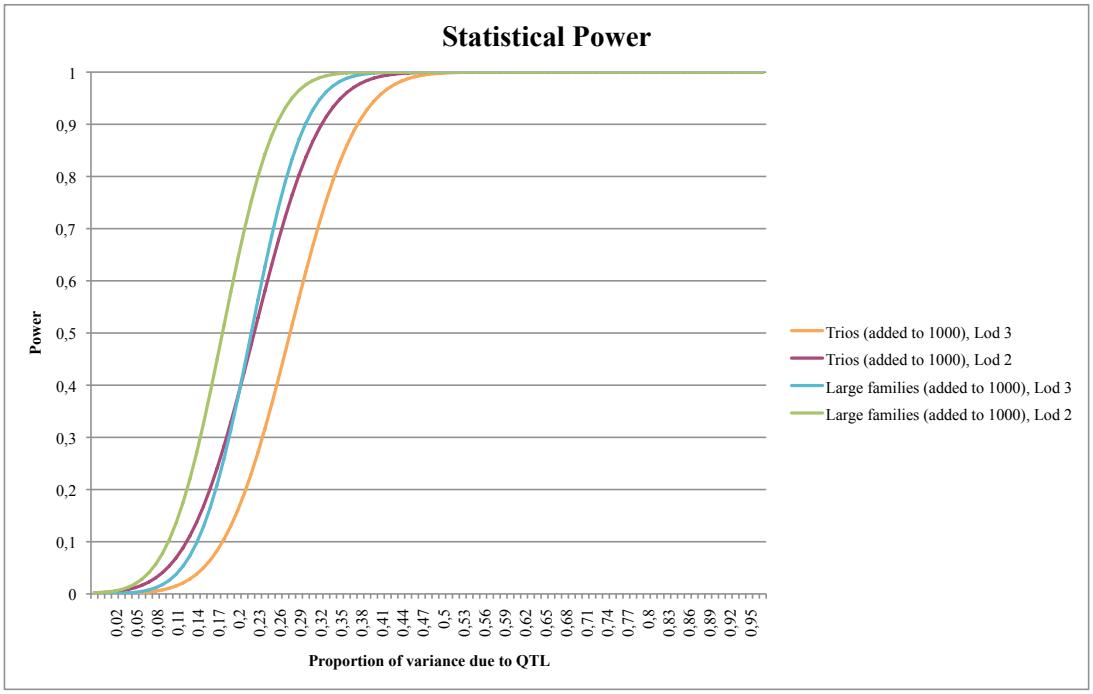


Figure 2 Estimated power for sample size of 1000 with different types of families for finding a QTL that affects musical aptitude. Simulations were made with the SOLAR program by the author. The simulated data includes the original families added to 1000 with trios (yellow and red line) or larger families (green and blue line). The figure shows that with larger families it is possible to find a locus that explains the smaller proportion of the variance of the trait.

Gene mapping through association and linkage will be performed with several programs. Several different programs may give more information about the genetic background of musical aptitude, and different programs will also be compared.

5.1 Materials

5.1.1 Subjects

The data consists of 107 extended families and 93 sporadic subjects, comprising in total of 915 individuals. Each family includes 2 – 50 individuals from 1 to 4 generations. Subjects over 7 years old were tested with musical aptitude tests and DNA was collected from individuals over 12 years old. The participants were 7 – 94 years old and 41% of them were males. The Ethical Committee of Helsinki University Central Hospital approved the study, and informed consent was obtained from all the participants or their parents.

Families have been collected in several batches. The first 15 families were collected for the first gene mapping study (Pulli et al. 2008). They were known to have several professional musicians in the families. In the later batches, the focus shifted towards normal variation of musical aptitude and families were collected with no prior knowledge about their interest in music. The second batch included 18 families (families 16 – 33) and about half of the families still included professional musicians. The rest of the families were collected as a population-based sample via advertisements in magazines, webpages and mailing lists. Open events to participate to the study were arranged in the Finnish Science Center Heureka. In this latest batch, there are musicians only in a few families. Examples of families in different batches are shown in Appendix I.

5.1.2 Phenotype

Musical aptitude was tested with three tests: the Karma auditory structuring ability test (KMT) and Carl Seashore's subtests of pitch (SPT) and time

discrimination (STT). The tests were played through loudspeakers at group sessions. For every test item, participants chose between two options: different/same for KMT, higher/lower for SPT and shorter/longer for STT. KMT includes 40 test items and both SPT and STT include 50 test items (maximum 40, 50 and 50 points, respectively). Three samples of KMT test items can be found from http://www.hi.helsinki.fi/music/naytteet.htm.

Due to the structure of the tests, it is possible, by chance, to guess half of the test items correctly. Thus, half of the test items were decided as the minimum score for every test. In addition, KMT scores were multiplied by 1.25 to allow for direct comparison with the Seashore's tests. SPSS Statistics 20 (IBM) was used to store and analyse the phenotype information.

In order to analyse measured musical aptitude as one variable, a combined test score was created. Musical aptitude scores from the three tests were summed to generate combined music scores. Histograms for original test scores and combined test scores can be seen in Figure 3.

On closer inspection, the ages of the study subjects were found to correlate with the test scores in a non-linear fashion (Figure 4). The combined test scores are lowest among youngest and oldest subjects. Thus, the combined test score was corrected for age through the curve estimation procedure. The best fitting curve was a quadratic curve with following formula:

$$Estimated\ curve\ for\ age = 110.9 + 0.997 * age + 0.014 * age^2 \quad (9)$$

(Figure 4). This curve was chosen to explain the largest amount of musical aptitude scores (R^2=0.185, p-value<0.001).

Additionally, also gender correlates with the combined test score. The combined scores are higher among males (Table 1). Thus, gender was also included in the final regression analysis with the estimated curve for age. Regression analysis with gender and quadratic age curves has an adjusted R squared of 0.189 (p-value<0.001).

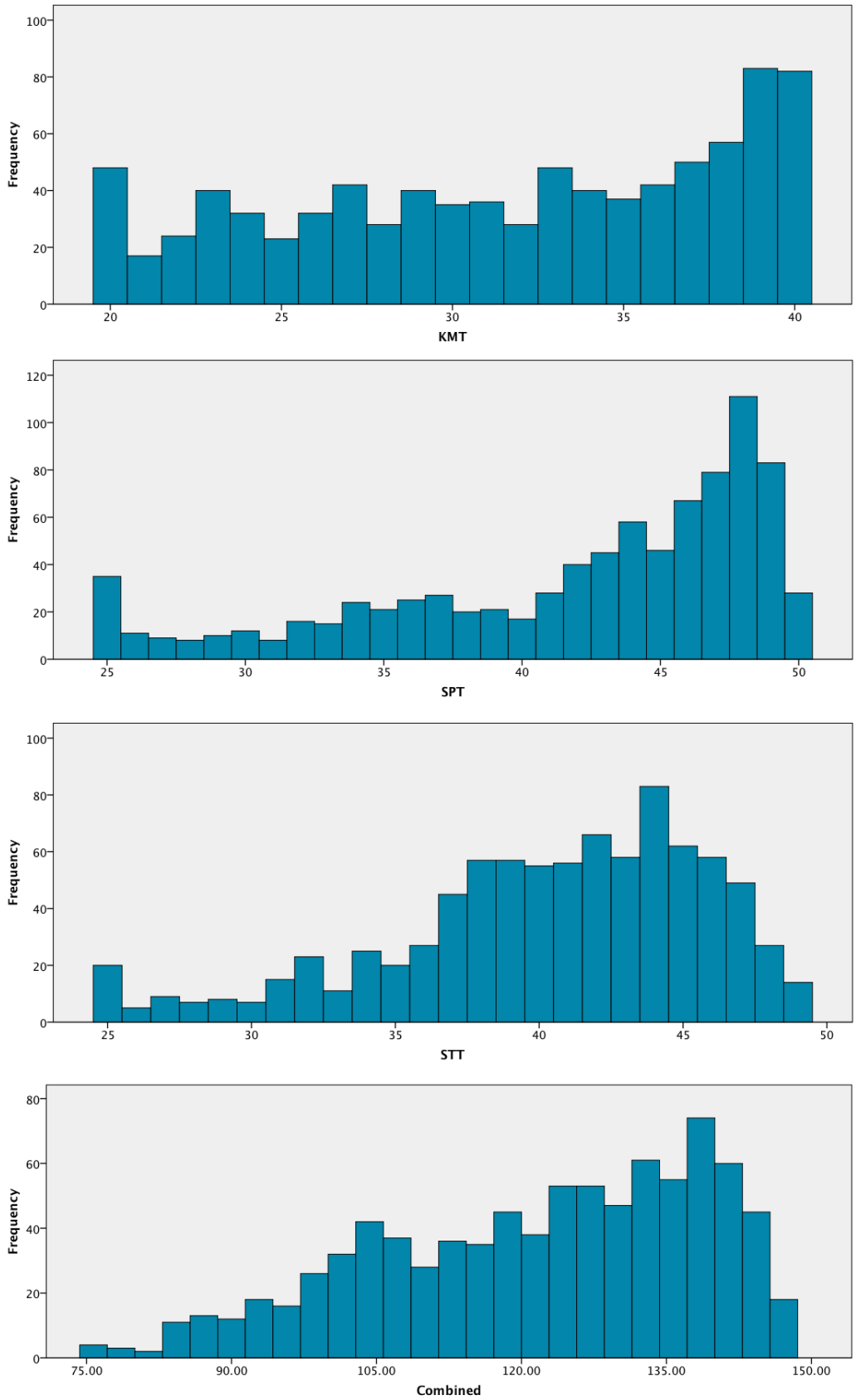


Figure 3 Distribution of musical aptitude test scores. The upper three charts show the original scores of the three musical aptitude tests. The chart on the bottom shows the summed test scores. The total number of participants for these tests is 864, including all members of the families.

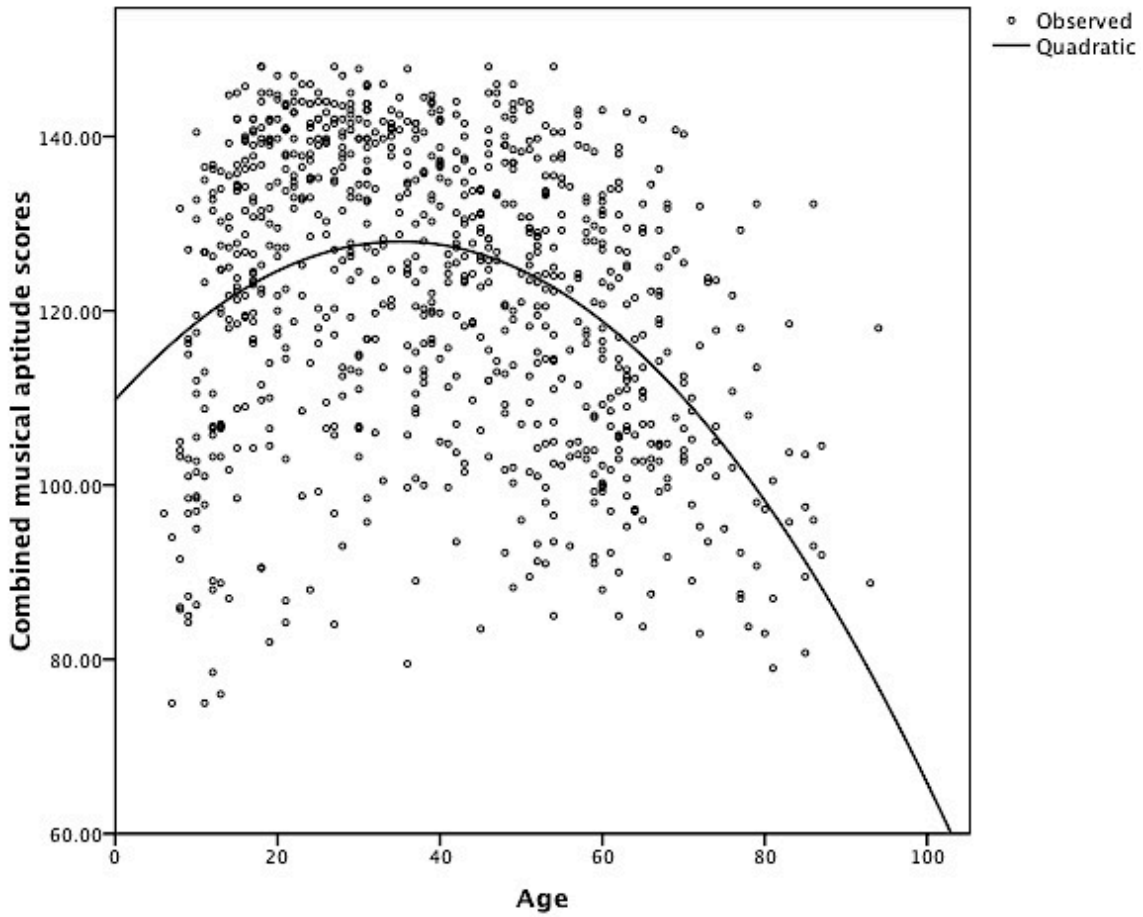


Figure 4 Combined musical aptitude scores differ by age. A quadratic curve was estimated to explain largest amount of musical aptitude by age. The curve was used to correct the effect of age on musical aptitude scores.

The lower scores among youngest can accumulate from concentration problems or problems understanding the questions. For the oldest group, hearing difficulties can explain the issue. Also, the pace of the test is fast and this can explain the lower performance among older subjects. There is no obvious reason for higher musical aptitude scores among males, but the difference is largest in STT.

Participants were also queried about their musical habits and education (see Ukkola et al. 2009 and Ukkola-Vuoti et al. 2011). Musical education was defined as years of active practise of music, and classified into four groups:

0 No music education

1 Some temporary music education (maximum of 2 years)

2 Active amateur

3 Professional

In total, 88 subjects were professional musicians and 278 subjects rehearsed music for more than half an hour a day. 191 subjects did not have any music education. Table 1 shows the characteristics for males and females. In Table 2, the characteristics of the different batches are compared.

Musical aptitude correlates with musical education (Figure 5). However, musical aptitude scores are not corrected by differences in musical education in this study. We assume that musical aptitude partly explains the amount of music education one has obtained. We also assume that only musically talented individuals would become musicians. Therefore, musically educated individuals are assumed to be innately musically talented and by correcting for musical education one would also lose a lot of information about musical aptitude.

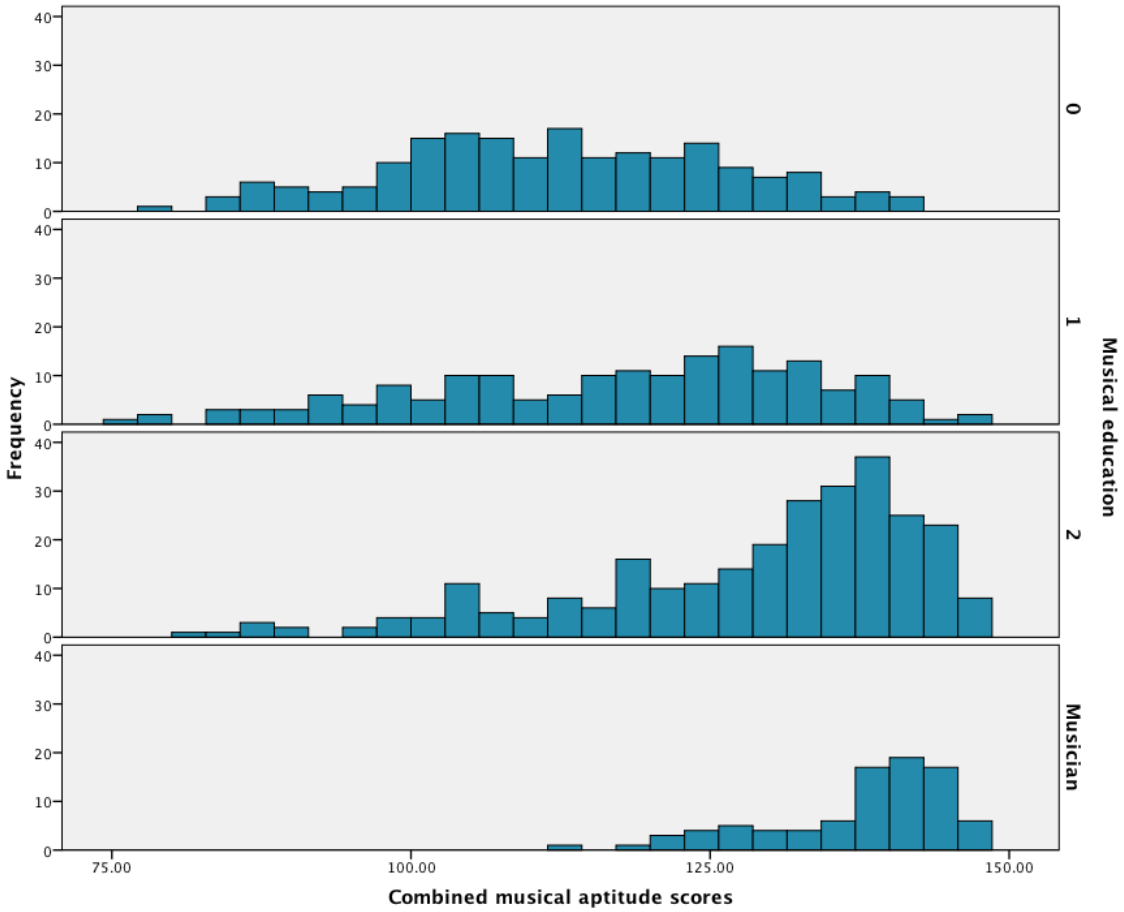


Figure 5 Musical education and original musical aptitude scores. On average, subjects with more musical education gain higher musical aptitude scores. This can be caused by selection on those who get music education, but also the practise of music may improve their performance on the tests. The scores are not normally distributed on the complete data, but scores among the subjects with no musical education (group 0) are normally distributed.

Apart from this examination, it is known that these tests poorly separate well-scoring individuals (Shuter-Dyson and Gabriel 1981). All individuals with high musical aptitude will gain equally high scores and no clear differences can be seen among them. Thus, based on these tests, we cannot separate the best from the good ones. This can also be seen in Figure 5, where the musicians cluster into the high end of the histogram. Other tests should be used for the well-scoring individuals to see any differences between them.

5.1.3 Genotypes

DNA extraction succeeded in 799 samples. The samples were genotyped with Illumina HumanOmniExpress 12 1.0V SNP chip (www.illumina.com). Genotyping was done at the Wellcome Trust Center. The chip includes over 700,000 SNPs. Genotype calls and genotyping quality control was performed with GenomeStudio. GenomeStudio contains Illumina's GenCall algorithm for calling genotypes.

Genotyping failed in 2 samples and they were removed from the following analysis. Call rates for included samples were over 99% (minimum 99.18%, mean 99.74%).

5.1.4 Marker map

Rutgers map v.2 (Matise et al. 2007) was used. Some of the SNPs in the data are situated on the ends of the chromosomes, which the Rutgers map does not cover. For these areas, the map values were generated computationally, using the following equation:

$$1\text{cM} = 1{,}000{,}000\text{bp} \quad (10)$$

(Ulgen and Li 2005).

Table 1 Characteristics by gender. There are differences in musical aptitude scores between males and females in STT and combined test scores; males have higher scores than females. The difference in KMT and SPT scores is not significant. Also, there are more subjects without any music education among males.

Characteristics	**Female**	**Male**	**Total**
Number of individuals	536	379	915
Mean age, years	39.94	41.12	40.42
Mean combined musical aptitude scores	120.8	123.2	121.8
KMT scores	31.5	31.8	31.6
SPT scores	41.8	42.1	41.9
STT scores	39.6	41.2	40.3
Number of musicians	53 (12.2%)	37 (12.2%)	90 (12.2%)
Number of no music education	99 (22.8%)	92 (31.1%)	191 (26.1%)

Table 2 Characteristics of different study subjects by batch. It was known, that the number of musicians differ by batch. However, the latest batch includes the most active amateurs as the number of subjects who practise music reveals.

Characteristics	**Fam. 1-15**	**Fam. 16-33**	**Fam. 34-107**	**Sporadic**	**Total**
Number of individuals	247	263	329	76	915
Males, %	48.7	46.8	45.3	56.1	47.9
Mean of age	43.7	40.3	38.7	37.4	41.2
Mean combined musical aptitude scores	122.3	121.9	121.0	123.1	121.8
Musicians, %	15.5	15.8	7.7	6.7	12.2
No music education, %	19.9	28.1	30.1	23.3	26.1
Does practise music, %	39.4	32.4	49.2	46.0	41.0

5.2 Data handling and quality control

Data handling procedures are shown as a flow chart in Figure 6. The data were first handled with Perl to change the file format. Perl language (version 5.10.0) was used, because it is memory efficient and is suitable for handling large files. With a Perl script, data was compiled and information about genotypes, phenotypes and family relationships were merged into one file set (Appendix IV). After merging, there were totally 1266 individuals in the data. 797 individuals have genotype information and 876 have phenotype information. We needed to create 349 dummy individuals to provide the familial relationships required by several programs.

Quality control was performed to exclude erroneous SNPs and samples. It was performed with PLINK 1.07 (Purcell et al. 2007). Thirty sporadic genotyped subjects were excluded because there was insufficient phenotype data. Relatedness was checked with the identity by descent (IBD) calculation. IBD estimate gives the proportion of alleles shared by a pair of individuals on all their markers. This estimate from genotypes can be compared against expected IBD value from known relationships. For example, the expected proportion alleles shared IBD for mother and son is 0.5. PLINK calculates these expected IBDs for simple familial relationships. For more complex relationships, expected IBDs were calculated by hand.

There were unexpected IBDs in three families. The relationships were confirmed from Finnish Population Information System (Suomen maistraatti). In two families, the genomic relationships were found to be consistent with true relationships and our pedigree information was corrected. In the third family, one subject with ambiguous gender was found to have been interchanged with another sample. Totally two subjects with ambiguous gender and one with substantial rate of Mendelian errors were pruned out. In the remaining data, Mendelian error rate was less than 0.1% per individual. However, it has to be noted that PLINK cannot handle the pedigrees in their entirety. Mendelian errors not evident in nuclear

families are ignored by PLINK. This problem was solved with the linkage data sets, but was not relevant in association analyses.

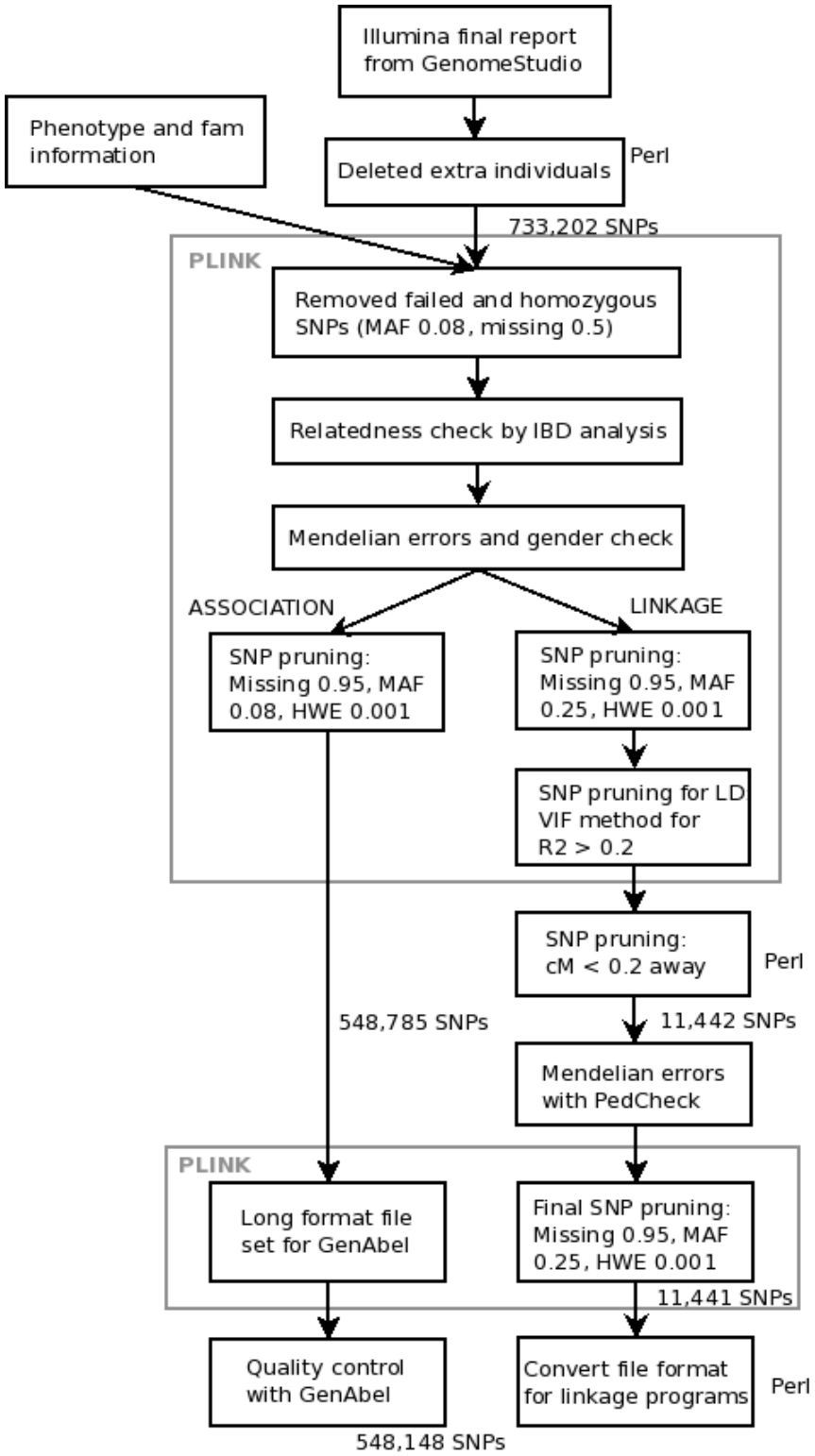


Figure 6 Performed data management and quality control steps. Most of the data quality steps were performed with PLINK.

After these revisions, 764 genotyped individuals remained in the analysis. A total of 686 of these individuals were members of the 107 families and included in linkage data sets. Also, there were identical twins in one family. One of them was excluded for those analyses that could not utilize twin information.

The marker data were pruned for minor allele frequency (among founders, MAF<0.08), genotyping success rate per locus (<95%) and Hardy-Weinberg equilibrium (p-value<0.001 in founders). 548,785 SNPs remained after these revisions.

5.3 Linkage analysis

5.3.1 Data handling

High linkage disequilibrium (LD) between markers can cause false positive signals in linkage analysis. Thus, SNPs in high LD to each other were pruned out. Prior to LD pruning, the minor allele frequency limit was raised to 0.25 in order to keep every family informative for the maximum number of SNPs. Rare alleles can segregate in some families and cause false positives (Horne et al. 2003). Moreover, the power should be higher with more common alleles (Risch 2000). This higher minor allele frequency threshold did not affect the final number of SNPs.

LD pruning was performed with the variant inflation factor method in PLINK, which I call here VIF. The VIF method removes markers with high LD within a sliding window. Within the window, every pair of markers is compared and when a high LD is found, a marker is removed. The VIF value tells us how dependent the markers are from each other. Completely independent SNPs would have a VIF value of 1.0. Dependent markers have higher values. VIF can be transformed into R^2 with this equation:

$$VIF = \frac{1}{1 - R^2} \quad (11)$$

(PLINK manual for Version 1.07).

I used a VIF limit of 1.25 that corresponds to 0.2 as R^2. For the sliding window, I used a size of 50 that shifts 1 marker at each step. I used the smallest step size to prune every marker that exceeds the VIF limit. Additionally, remaining SNPs were pruned for map distance less than 0.2cM. The pruning yielded a subset of 11,442 autosomal SNPs. With PedCheck (O'Connell and Weeks 1998) for Mendelian errors, these SNPs were checked and zeroed out among pedigrees.

Missing rate (>5%) was inspected again after zeroing out genotypes; one SNP was excluded from the final linkage data set.

5.3.2 Linkage analysis programs

Multiple programs were inspected to find the best suited ones for this data. Three programs were chosen, as they were suitable for large pedigrees, a large number of markers and quantitative phenotype. These programs include: the Java Programs for Statistical Genetics and Computational Statistics (JPSGCS) package (Thomas et al. 2000), the Sequential Oligogenic Linkage Analysis Routines (Solar) package (Almasy and Blangero 1998) and the KELVIN package (Vieland et al. 2011b). Also the MORGAN program (Tong and Thompson 2008) would have fulfilled the prerequisites, but faster programs were favoured.

The JPSGCS package includes MCMC methods for linkage analysis and it is based on graphical models combining Elston-Stewart and Lander-Green algorithms. It is uncommon program, but it has been shown to perform similar results as the common program Merlin (Allen-Brady et al. 2007). The Solar package is a variance components method and it is the most common of these three. It was also used successfully in the pilot study. The KELVIN package uses an Elston-Stewart based method with a Bayesian perspective. Thus, it is based on the most common algorithm, even though the Bayesian perspective is new. Altogether, all of these programs seem reliable and they have been proven to work on real data.

Untyped parents and remarriages make large pedigrees very complex to calculate. In our data there is a lot of missing data, mostly parents and grandparents who have not attended the study. These missing parental genotypes make the interpretation of genotype phases computationally demanding and slow. Because of this complexity, the pedigrees can be analysed in their entirety only with a few programs. Even with inexact methods, multipoint analyses are expected to take a long time. Thus, it was not possible to analyse the whole data with all the programs with our computational resources. It has been estimated that even one program of these programs could run for weeks.

The Bayesian program KELVIN could be included in this thesis through collaboration with Veronica Vieland's laboratory. They have the requisite computational resources to analyse the data. Also they have not published the KELVIN version for large pedigrees and our data was used to test the version. Our computational resources were then needed only for JPSGCS and Solar analyses.

Also, the Merlin program was experimented for split families, but the missing data made the pedigrees very complex and splitting would have to be done extensively. For example, one family would have to be split in ten subfamilies. In addition, splitting can end up with noteworthy decrease of statistical power. Thus, Merlin was not used after its trial.

5.3.3 JPSGCS analyses

First, I performed linkage analysis with JPSGCS (version dated October 2011). JPSGCS uses LINKAGE format files (locus and ped file) with some exceptions. Post-MAKEPED pedigree information was produced with the MAKEPED program (author Peter Cartwright, included in the LINKAGE package). The files were then compiled with Perl scripts. The linkage analysis was run both as two- and as multipoint.

The genetic model for the parametric analysis was specified in the following manner: the additive genetic model was chosen, because it is known to be most

robust for model misspecification. Trait mean differences at the unknown QTL were set to -1.5 and +1.5 standard deviation (SD) from the trait mean for the two homozygotes respectively, and the trait mean for heterozygotes (as in Goldgar and Oniki 1992). These values were chosen, because it is known that the linkage method is not powerful enough to detect loci with only modest effect (<1.0 SD) (Risch 2000). In contrast, higher values would be biologically unlikely. For trait allele frequencies, 0.01 and 0.1 were used as in Horne et al. 2003. Thus, two analyses were performed. Marker allele frequencies were estimated with PLINK, from the founders.

The two-point method in JPSGCS produces result for every family separately over theta values {0, 0.01, 0.1, 0.2, 0.3, 0.4, 0.5}. The results were merged with a Perl script. The script sums lod scores from every family for every theta value and the theta value that gains highest total lod score, is selected. Thus, the lod scores were maximized over theta values over all families.

Multipoint analysis was performed with the same parameters as two-point analysis. Multipoint results are produced for a theta value of 0 and for every pedigree separately. Because JPSGCS produces the results for very tight map, the results were not merged directly. The results were first maximized inside each family for the 2cM areas and these maximized lod scores were then summed over the families. Other programs produce results directly for some specified cM area.

Per pedigree lod scores from multipoint analysis were used to recover those pedigrees that are linked to certain area. A lod of 0.55, corresponding approximately to p-value of 0.05, is considered as evidence for linkage in a single family.

5.3.4 Solar analysis

Solar (version 4.2.0) was used for non-parametric linkage analysis. It uses comma-separated files; these were compiled from LINKAGE styled files with Perl scripts.

Solar requires IBD for every SNP and multipoint IBD values for specified cM increments. These were calculated, multipoint IBD values for 2cM increments. Two-point and multipoint analyses for every autosomal chromosome were performed.

5.3.5 KELVIN analysis

Bayesian linkage analysis was performed in Veronica Vieland's laboratory with an unpublished version of the KELVIN program, designed for large families. The algorithm for handling complex pedigrees has been implemented for this version from the JPSGCS program (Veronica Vieland, Battelle Center for Mathematical Medicine, Research Institute at Nationwide Children's Hospital, personal communication, 13 October 2011).

Here, the original uncorrected phenotype was used instead of age- and gender-corrected scores. The Bayesian framework gains from unnormalized data without corrections. The file format for KELVIN is a post-MAKEPED pedigree file with additional files for marker details. These were compiled with Perl scripts from JPSGCS files.

5.4 Association analysis

5.4.1 Program

Association analysis was performed with GenAbel (version 1.6-8) that is an R program for association analysis for both: population-based and family data (Aulchenko et al. 2007). R version 2.15.0 was used (R 2012). GenAbel is a fast association method that is based on approximation of maximum likelihood.

5.4.2 Quality control

All subjects, with both phenotype and genotypes available, were included in the association analysis, which resulted in 764 subjects. 548,785 SNPs were included from PLINK (preliminary quality control explained in Section 5.2).

Because of different set of samples in association analysis, quality control of the SNPs needed to be performed further analysis-wise. GenAbel performs quality control gradually until no individuals or SNPs need to be removed. A total of 637 SNPs were left out for call rate less than 95%. Identical twins were both excluded for association analysis. 548,148 markers and 762 subjects remained in the analysis.

5.4.3 Analysis

GenAbel offers two algorithms for association analysis with family data: Family-based Score Test for Association (FASTA) and Genome-wide Rapid Analysis using Mixed Models And Score Test (GRAMMAS) (Aulchenko 2011). The FASTA method may be used if heritability in the sample is high. If the heritability is low, GRAMMAS should be used. Heritability was here estimated as 0.57, which corresponds to an estimate from Solar. Because this heritability is relatively high, the FASTA method was used.

GenAbel employs the mixed model method (genomic kinship based method) to correct for population stratification and for family structure (see Section 2.6.2). Variance inflation factor statistics (λ) is used to correct genome-wide significance of observed χ^2 statistics. λ represents the background differentiation of the genomic data. In this data lambda was estimated to be below 1, as assumed. This means that it is deflation factor instead of inflation factor and result from the relatedness of the subjects.

In our data, we also found out that some individuals share markers IBD at a in higher level than expected (IBD>0.05), with no prior knowledge about their

relatedness. The amount was the same as if their grandparents would be siblings. Thus, it is assumed that there are distant relatives in this data, even between families. The mixed model method is based on genetic similarity and it can detect and correct also this kind of unknown relatedness.

The data were first analysed with the mixed model method without principal components analysis (PCA). In the second run, PCA was included. PCA divides the sample into genetically different subgroups. The mixed model method with PCA is expected to perform well for family data.

6 Results

6.1 Linkage

Common linkage programs were inspected to see if they were suitable for this project. The programs and their information can be seen in Table 3. From every program, the most suitable analysis options were considered. Linkage analysis was performed with JPSGCS, SOLAR and KELVIN that are all capable to analyse complex pedigrees, a quantitative trait and large amount of markers. These three programs are based on different algorithms and thus, the same calculations are not repeated with different programs. Association analysis was performed with GenAbel with the mixed models approach to correct for familial relationships.

Heritability was estimated for age and gender corrected musical aptitude scores (see Section 5.1.2) to be around 0.57. It was estimated with SOLAR and GenAbel, which both resulted in the same heritability estimate.

Linkage analysis for musical aptitude showed multiple significant results. Linkage analysis was performed as multipoint analysis with JPSGCS, Solar and KELVIN. JPSGCS and SOLAR were used for all autosomes, and KELVIN for all autosomes and also for the X chromosome. The results are plotted in Figures 7, 8 and 9 respectively for JPSGCS, SOLAR and KELVIN.

The best results from the linkage analyses are also shown in Table 4 with examples of genes with neuronal activities in those areas. The same areas with more detailed information about the width of the area and supportive families are shown in Table 5. Supportive families were identified as families with Lod score (JPSGCS) over 0.55. The maximum number of families per area was 7.

The best area was found in chromosome 4 in all three programs. The area spans from 4p13 to 4q13.1, the highest peak settles on 4q12. The result for this area exceeded a lod score of 3.3 (significant result) with JPSGCS. With SOLAR, the multipoint lod score was 2.80 (suggestive result) and KELVIN showed a

probability of 0.63. The multipoint results for chromosome 4 are plotted in Figure 10 from Solar and JPSGCS, and in Figure 11 from KELVIN.

Two-point analyses showed also several significant loci. Solar analysis revealed a lod score of 3.5 at 4q12 and lod score of 3.2 at 4q35.1, which correspond to multipoint results (Table 5, Appendix III). JPSGCS analysis revealed a lod score of 4.1 at 19q13.31 and 3.2 at 16p13.2 (Appendix II). On the chromosome 19q13.31, there is also multipoint lod score of 2.0 from JPSGCS analysis, but neither of these areas were found with any other analysis method.

Additionally, the linkage results were compared to our pilot study (Pulli et al. 2008). Old and new results from the SOLAR program in chromosome 4 were plotted against the map used in this study (Figure 12). Only combined music scores were considered. The best peak from the pilot study map to a smaller peak on the side of the best area found in this study. Chromosomes 8 and 18 showed suggestive linkage in the pilot study, but no linkage in SOLAR analysis in this study. Also, KELVIN results on 8q24 and 18q22 are located outside of old areas in 8q13-q21 and 18q11.2-q21.1.

Table 3 Common linkage programs examined for this project. Programs were studied to see if they were suitable for complex pedigrees, quantitative traits and large amount of markers in a possible computational time. Some programs have several options for analysis, here are shown the most suitable options from every program. Additionally to the chosen programs, also MORGAN could have been used

Program	Parametric	Algorithm	Family size	Loci	QT	MC*	Other
ALLEGRO[1]	Yes / No	Lander-Green	Moderate	Many	Yes	HMM	Uses Fourier transforms
FastLINK[2]	Yes	Elston-Stewart	Any size	Some	No	-	
GENEHUNTER[3]	Yes / No	Lander-Green	Small	Many	Yes	MCMC	Uses Fourier transforms
JPSGCS[4]	Yes	Graphical	Any size	Many	Yes	MCMC	
KELVIN[5]	No	Elston-Stewart, Bayesian PPL	Any size	Many	Yes	MCMC	
LINKAGE[6]	Yes	Elston-Stewart	Any size	Some	Yes	-	
LIPED[7]	Yes	Elston-Stewart	Any size	Two	Yes	-	
Loki[8]	No	Bayesian QT	Large	Many	Yes	MCMC	
MENDEL[9]	Yes	Elston-Stewart	Large	Many	No	**	QT only with moderate sized family
MERLIN[10]	Yes / No	Lander-Green	Moderate	Many	Yes	-	Models LD
MORGAN[11]	Yes / No	Combination of E-S and L-G	Any size	Many	Yes	MCMC	
SAGE[12]	Yes / No	Elston-Stewart	Large	Many	Yes	MCMC	Also multivariate analysis
SimWalk2[13]	Yes / No	Lander-Green	Any size	Many	No	MCMC	
SOLAR[14]	No	VC	Any size	<3000	Yes	-	
VITESSE[15]	Yes	Elston-Stewart	Any size	Some	Yes	-	

*Markov chain methods, including Markov chain Monte Carlo (MCMC) and Hidden Markov model (HMM).
**Includes MCMC methods, but not applicable for extended pedigrees.
Program manuals:
1) www.decode.com/software/; 2) www.ncbi.nlm.nih.gov/CBBresearch/Schaffer/fastlink.html/;
3) linkage.rockefeller.edu/soft/gh/ 4) balance.med.utah.edu/wiki/index.php/JPSGCS;
5) kelvin.mathmed.org/static/doc/; 6) linkage.rockefeller.edu/soft/linkage/;
7) linkage.rockefeller.edu/ott/liped.html/; 8) www.stat.washington.edu/thompson/Genepi/Loki.shtml/;
9) www.genetics.ucla.edu/software/mendel/; 10) www.sph.umich.edu/csg/abecasis/Merlin/;
11) www.stat.washington.edu/thompson/Genepi/MORGAN/morgan303-tut-html/morgan-tut.html/;
12) darwin.cwru.edu/sage/?q=node/9/; 13) www.genetics.ucla.edu/software/simwalk/;
14) bioweb2.pasteur.fr/docs/solar/; 15) watson.hgen.pitt.edu/register/docs/vitesse.html/

Table 4 Comparison of results from Solar, KELVIN and JPSGCS. Here are listed the best results from linkage analyses where lod score exceeds 2.2 or probability score exceeds 0.17. Multipoint lod scores are shown from Solar and JPSGCS if they exceed 0.5. The KELVIN results are shown if the probability score exceeds 0.1. The gene names refer to the RefSeq database (The NCBI Handbook 2002).

	cM (Rutgers)	**JPSGCS** (Lod)	**KELVIN** (probability)	**SOLAR** (Lod)	**Examples of candidate genes in the area**
1q44	242-270	-	0.12	1.77	*KCNK1*, *RYR2*
4q12	64-84	3.26	0.63	2.80	Several genes
4q21.1	88-92	2.62	0.49	2.47	*SHROOM3*
4q35.1	186-198	0.63	0.11	1.15	*ENPP6*
8q24.13	126	-	0.17	-	*HAS2*
13q31.1	75	2.30	-	-	*SLITRK1*
14q11.2	0-6	-	0.11	1.67	*NDRG2*
17q11.2	56-72	-	0.15	0.50	*MYO1D*
18p11.31	24-32	-	0.17	0.50	-
18q22.1	96-100	-	0.42	-	*CDH7*, *CDH19*
18q23	120-122	-	0.18	-	*MBP*
20q13.33	106-109	2.88	0.13	-	*CDH4*
22q11.21	0-10	-	0.18	1.59	*CECR1* and *2*, *BID*

Table 5 A detailed comparison of results from different programs. The width of the area is estimated primarily from KELVIN results, where the area is defined as a continuing curve above probability score of 0.10. Families are considered as linked if the multipoint lod score exceeds 0.55.

AREA				**KELVIN**	**SOLAR**		**JPSGCS**	
Chr	**Genomic area**	**From - to**	**Width** (cM)	**Probability**	**Lod** (multi-point)	**Lod** (two-point)	**Families linked**	**Max. lod per family**
1	1q44	q42.2-q44	2	0.12	1.77	**2.74**	3	0.92 (#12)
4	4q12	p13-q13.1	19	**0.63**	**2.80**	**3.54**	6	1.39 (#10)
4	4q21.1	q13.3-q21.1	3	**0.49**	**2.47**	2.22	4	1.37 (#10)
4	4q35.1	4q35.1	1	0.11	1.15	**3.25**	**7**	1.56 (#15)
8	8q24.13	q24.12-q24.13	2	0.17	-	-	2	0.78 (#17)
13	13q31.1	q31.1	-	-	**-**	1.26	3	1.97 (#6)
14	14q11.2	p13-q11.2	4	0.11	1.67	**2.12**	2	1.04 (#17)
17	17q11.2	q11.2	4	0.15	0.50	**2.64**	2	0.89 (#22)
18	18p11.31	p11.31	1	0.17	0.50	1.74	2	1.06 (#15)
18	18q22.1	q21.33-q22.1	6	**0.42**	-	1.01	4	1.44 (#14)
18	18q23	q23	4	0.18	-	-	5	1.98 (#14)
20	20q13.33	20q13.33	3	0.13	-	1.12	5	1.49 (#17)
22	22q11.21	q11.1-q11.21	6	0.18	1.59	**2.02**	4	1.04 (#21)

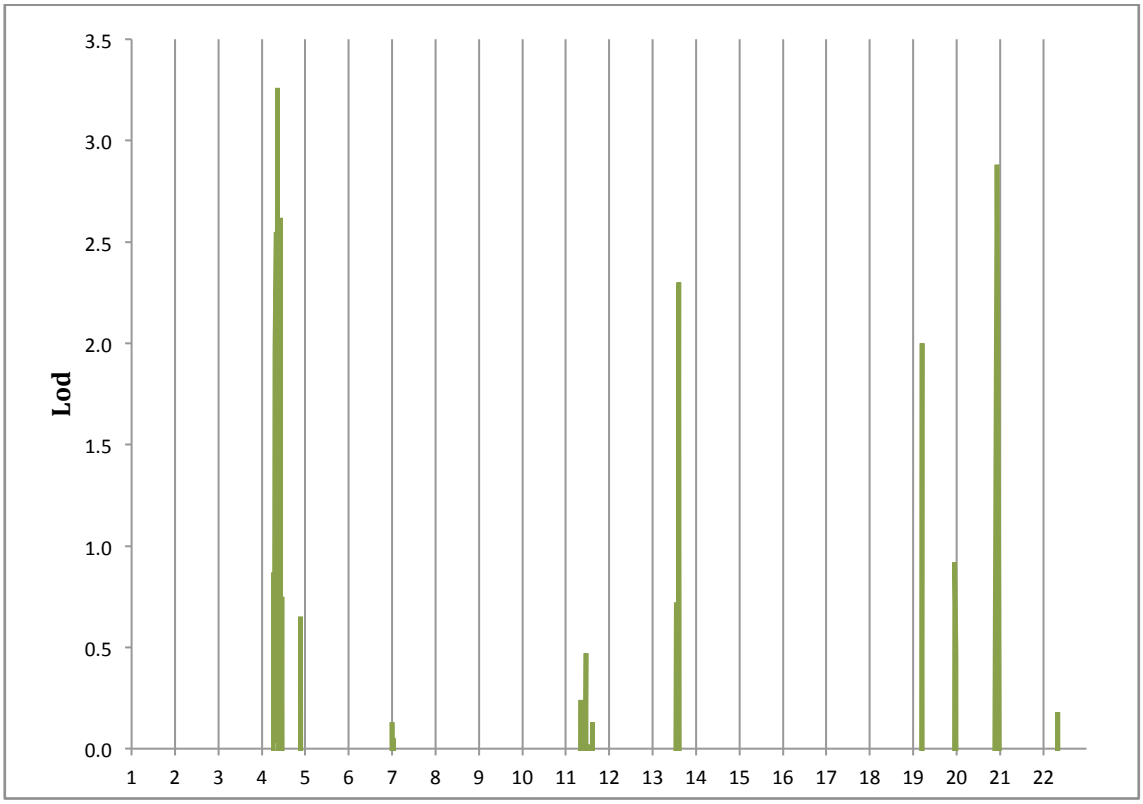


Figure 7 JPSGCS results for multipoint linkage analysis. The x-axis shows the chromosomes and the y-axis shows the result of JPSGCS multipoint analysis as lod scores. Only lod scores above 0 are shown. The best lod score lie on chromosome 4 (lod score of 3.26)

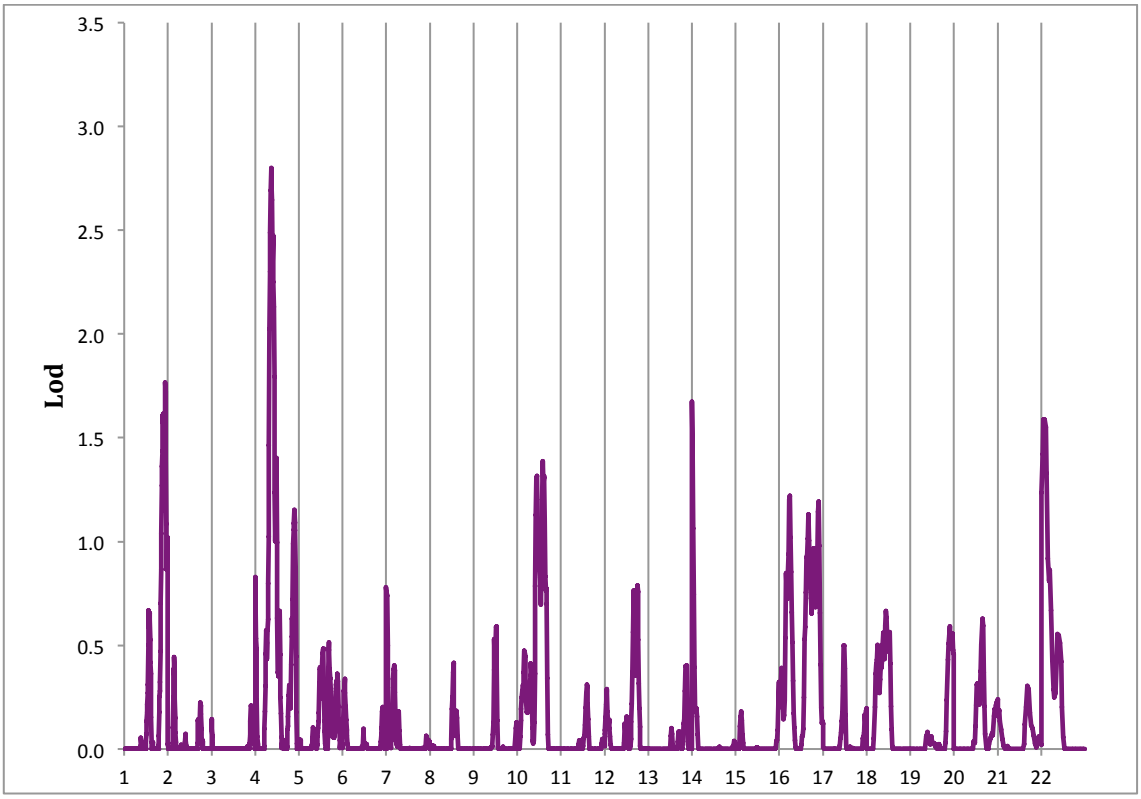


Figure 8 Solar results for multipoint linkage analysis. The x-axis shows the chromosomes and the y-axis shows the result of Solar multipoint analysis as lod scores. The best lod score lie on chromosome 4 (lod score of 2.80)

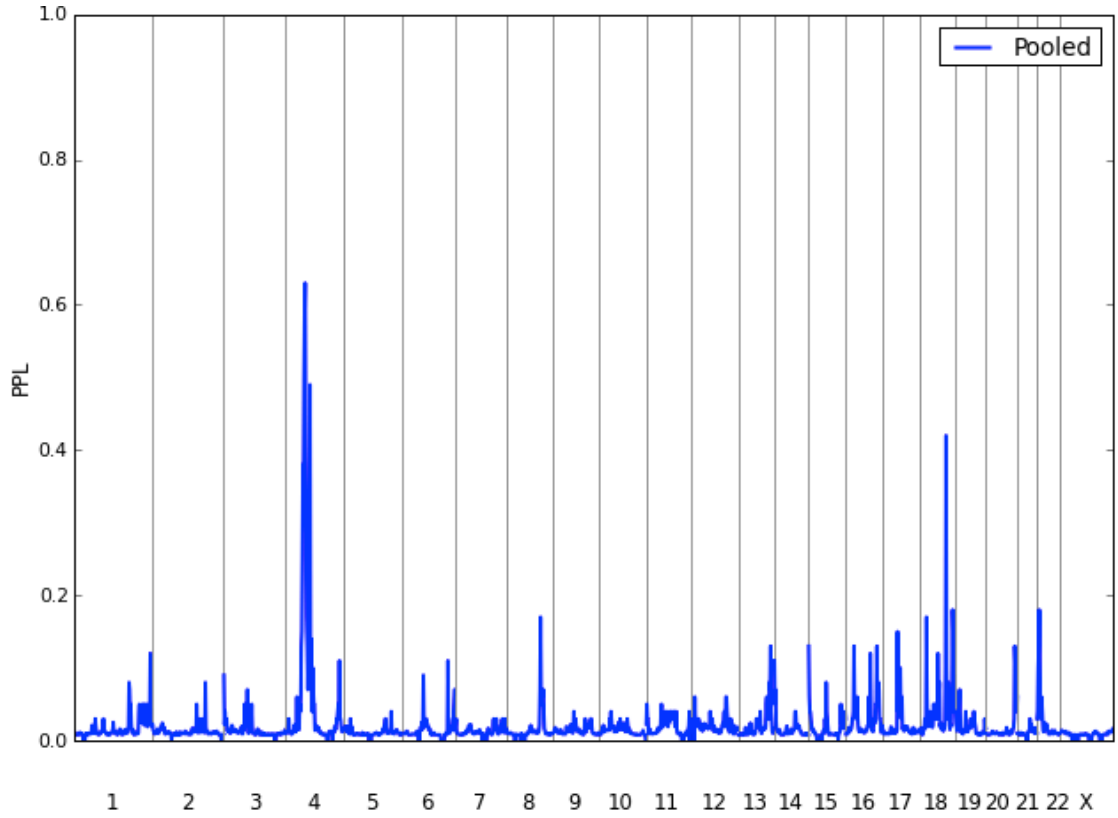


Figure 9 Linkage result for all chromosomes from KELVIN. The x-axis shows the chromosomes and the y-axis shows the result of PPL analysis on probability scale. The highest peak lies on chromosome 4 (probability 0.63).

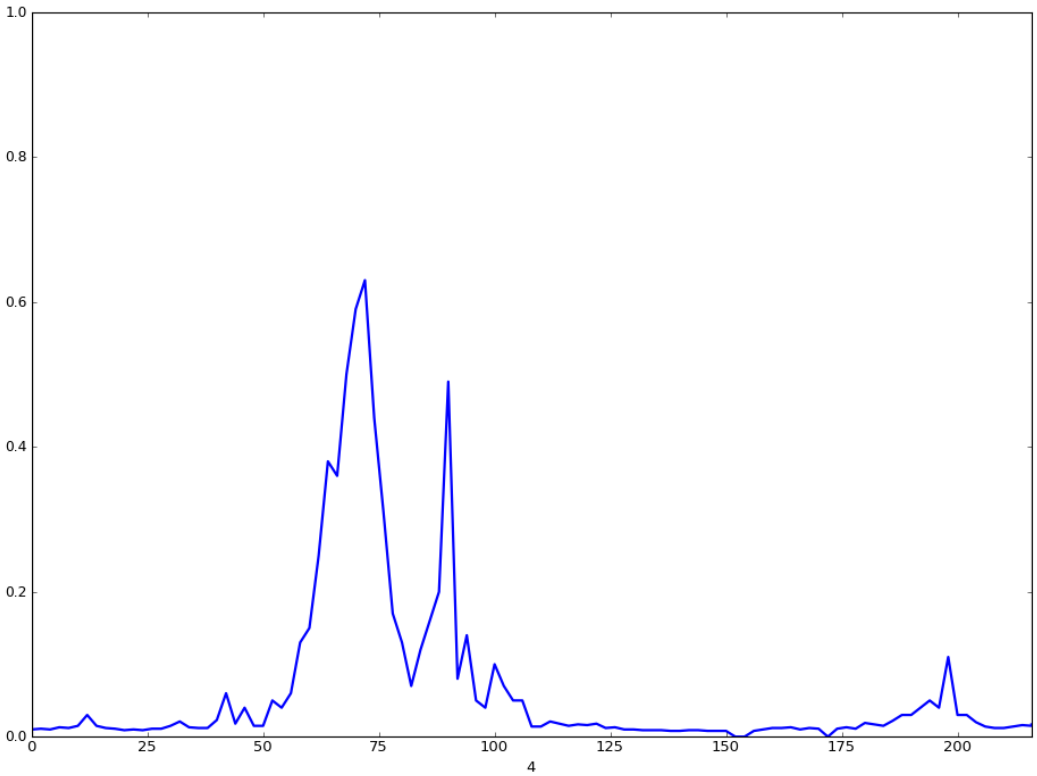


Figure 10 KELVIN result for chromosome 4. The x-axis shows the chromosome on a cM scale and the y-axis shows the result of PPL analysis on a probability scale

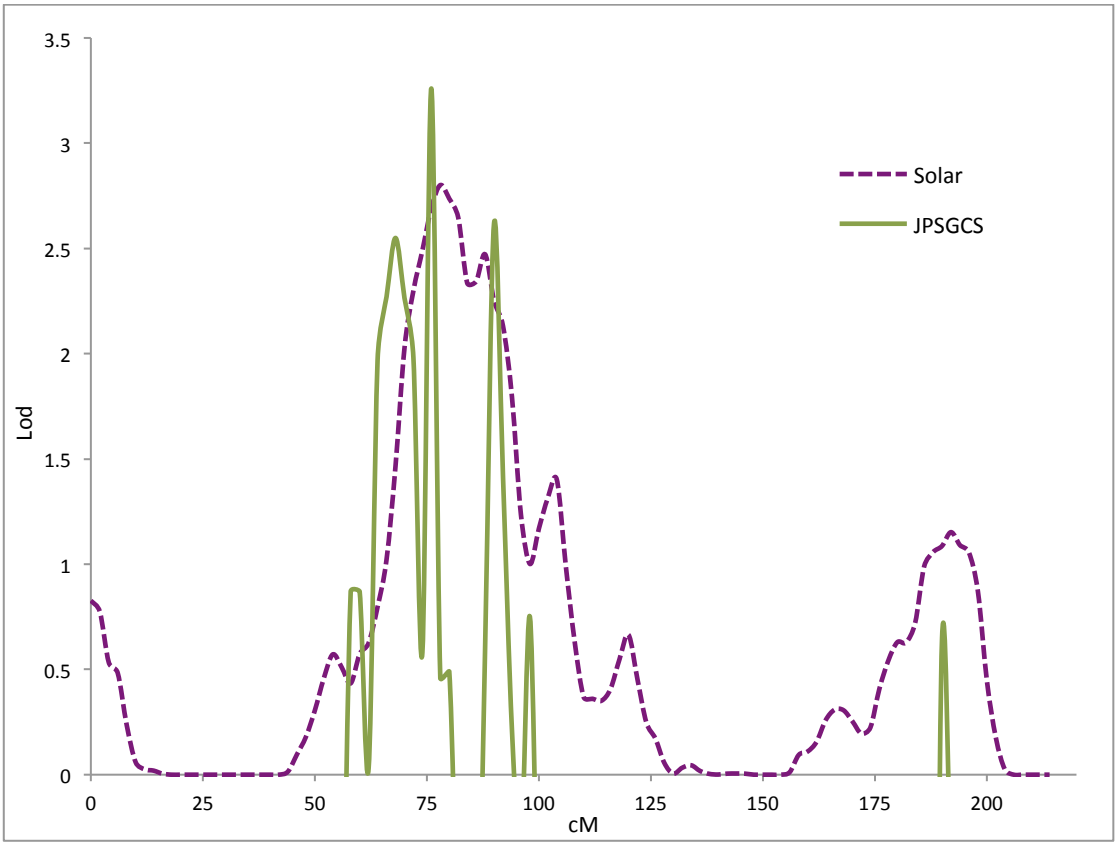


Figure 11 Solar and JPSGCS multipoint results for chromosome 4. The best linkage peaks from the both programs are situated on the same locations.

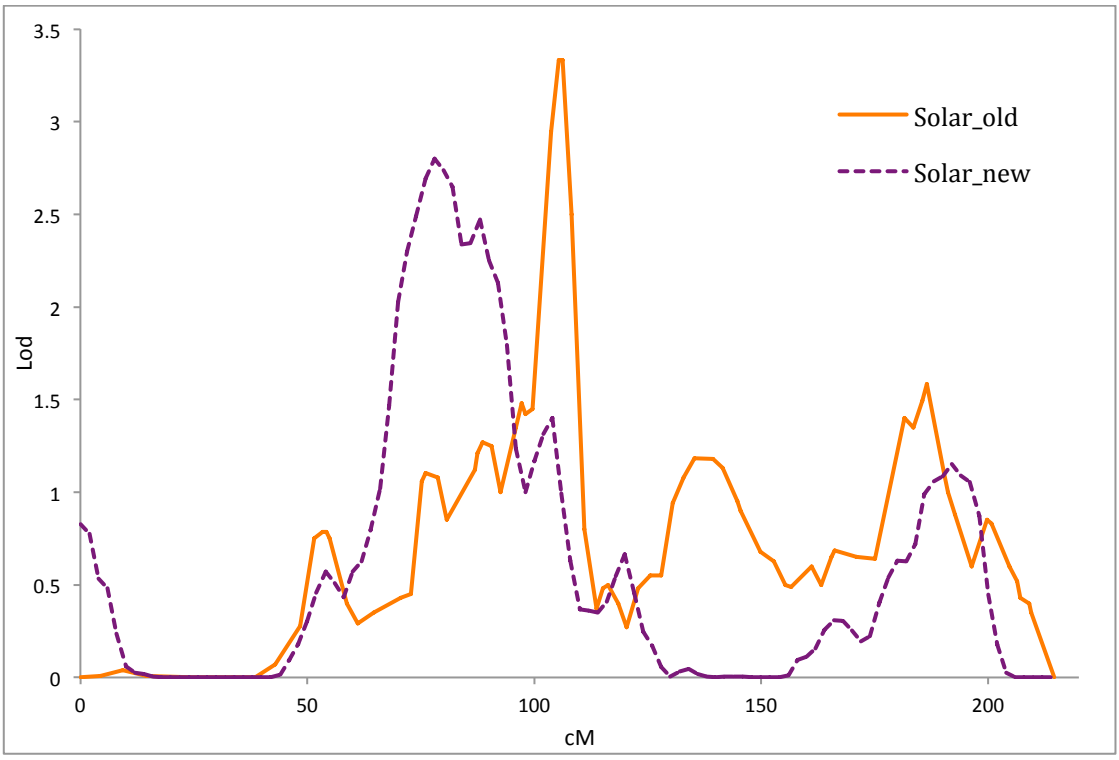


Figure 12 Comparison of results of our pilot study (old) and this study (new) in chromosome 4. Results from old and new Solar analyses are shown here plotted against map used in this study. Thus, the old Solar results were remapped to make the comparison possible.

6.2 Association

Association results were not significant for musical aptitude. Heritability was estimated to be 0.57. Association analysis was first performed with the FASTA algorithm with mixed models correction for population stratification. The results are represented as a Manhattan plot in Figure 13. Interference is quite high in the Manhattan plot, and there are no significantly high peaks. The interference is likely caused by deflation factor that corrects the results to be higher than the original values. A QQ-plot (Figure 14) shows that the population stratification in the sample has been corrected and the results behave as predicted following the expected line. If there would be any true associations, the high end of the curve in QQ-plot should turn to the observed side instead of the expected side. Now, the observed test statistics have less large values than expected.

As mentioned, association analysis was first performed with merely mixed models correction. On the second run, mixed models correction was joined with principal components analysis. In PCA analysis, three subpopulations were identified (Figure 15); these subpopulations were added to the mixed models approach. However, subpopulation structure did not significantly affect the results of association analysis (data not shown). The Manhattan and QQ plots were very similar for both approaches.

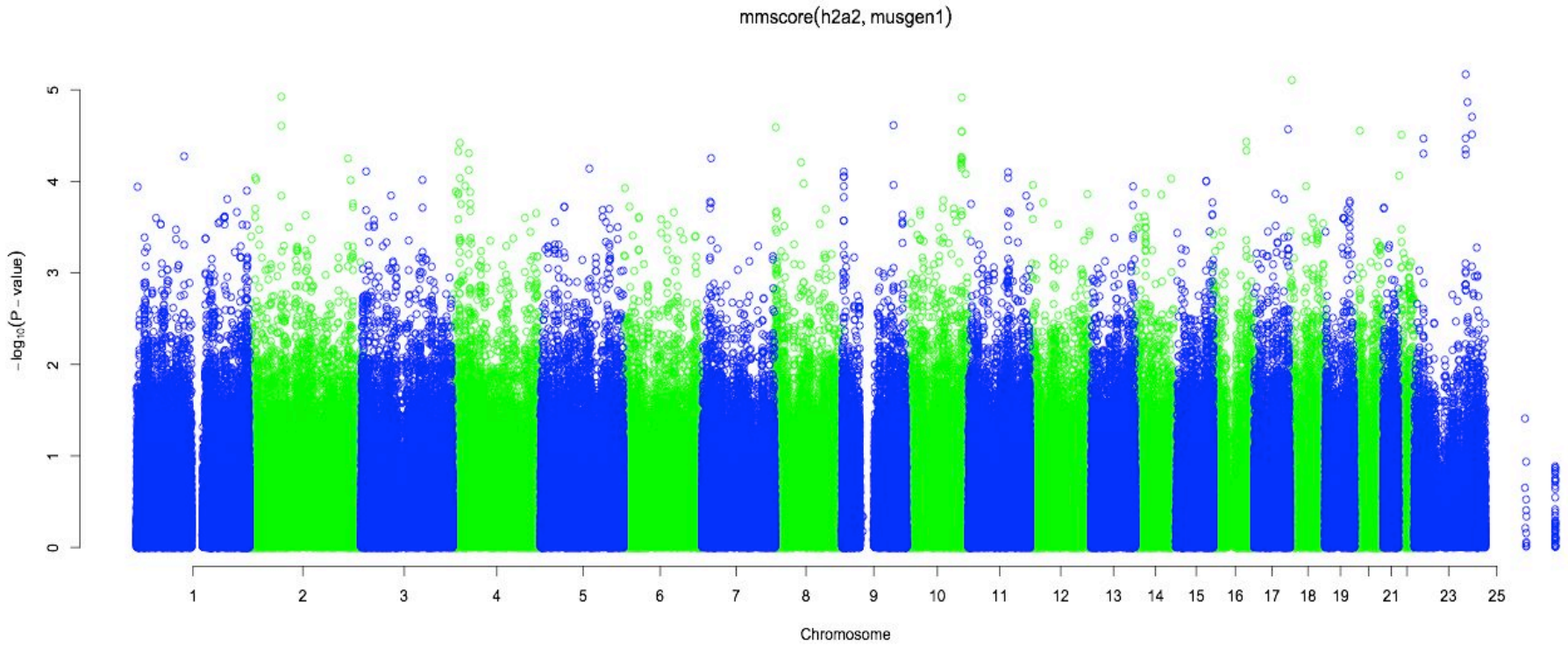


Figure 13 Manhattan plot showing result from association analysis. Chromosomes are separated with colours. Number 23 assigns X chromosome and number 25 assigns for XY SNPs, meaning SNP that are found from areas shared by X and Y chromosomes. Significant results would show here as clear peaks rising from a minimum to a 7 on the logarithmic scale.

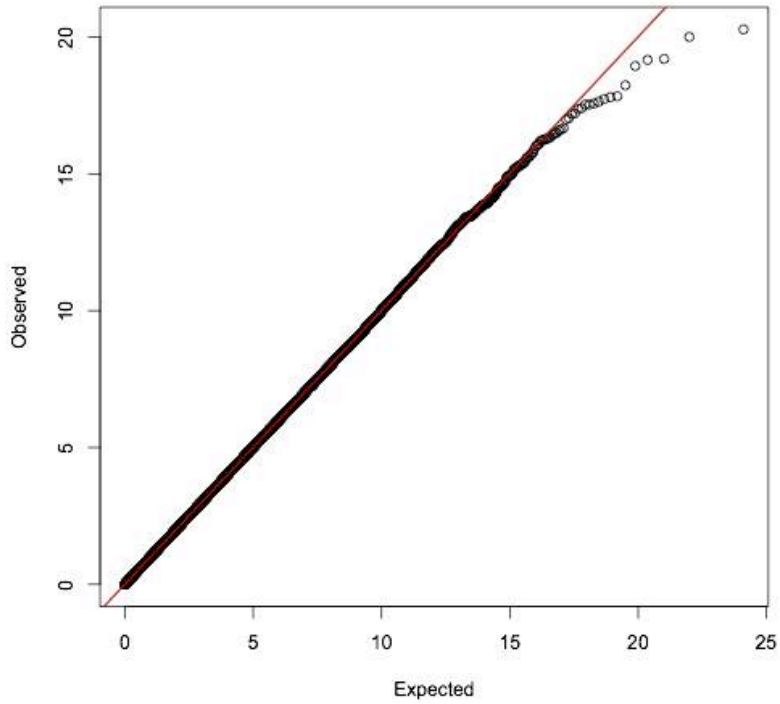


Figure 14 QQ plot of association analysis. Observed test statistics follow the expected distribution. The high end of the peak shows, that there are no significant results, but even less large values than expected.

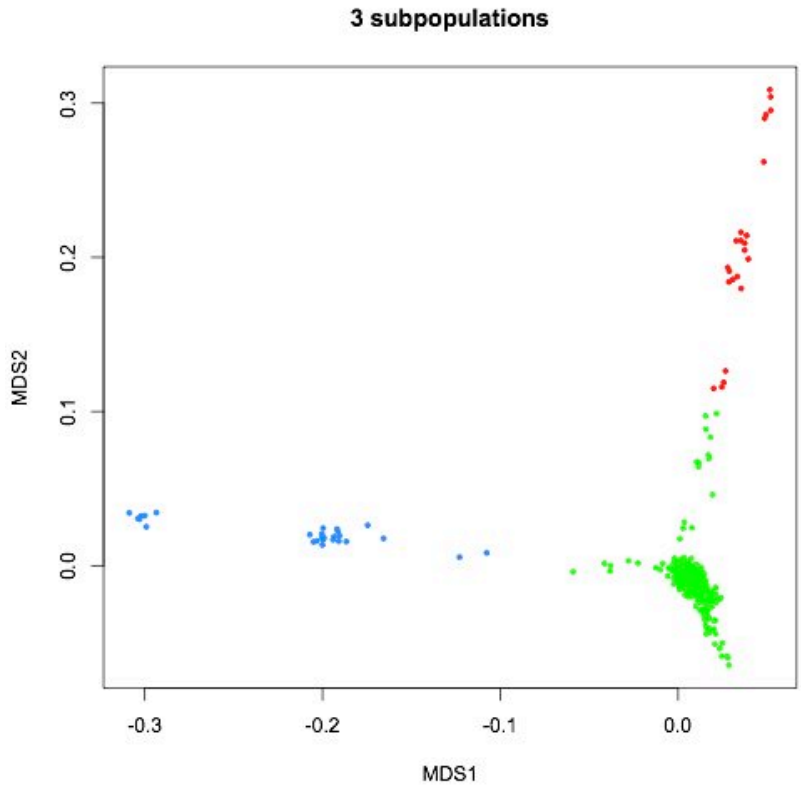


Figure 15 Three subpopulations identified in association analysis showed in different colours. The subjects are drawn on this dot-plot by using multidimensional scaling (MDS) values.

7 Discussion and conclusions

In this study, we were able to find significant or suggestive linkage for musical aptitude in 4q12 using three different linkage approaches. Three linkage programs were successfully used for complex, extended families. Even though the methods were different, most of the results from these programs pointed to the same areas. Association analysis with one program revealed no genome wide significant results. No other association method was used in the scope of this thesis.

7.1 Usability of the programs

7.1.1 Usability of the tested linkage programs

Several linkage programs were inspected to find suitable programs for this project. Programs needed to be suitable for large, complex pedigrees with missing data and for quantitative traits. Commonly used and fast programs were favoured.

Solar, JPSGCS and KELVIN were chosen for the final analyses. Solar was known to be able to handle large pedigrees. Very large pedigrees have been studied with JPSGCS (for example Camp et al. 2003) and it was supposed to be a fast program. KELVIN was utilized in collaboration with Veronica Vieland's group, who are developing the program for extended families. These three programs are also based on different algorithms (see Section 5.3.2) and the results from these different methods can give more information about the trait.

The usability of the linkage programs varied largely. Here, I will discuss the documentation and the usability of the three programs used for linkage analysis. Additionally, during the program selection, the documentation of 12 other programs (listed in Table 3, on page 51) was inspected, and the documentation could be compared between all of these.

Compared to other programs, SOLAR has a good manual with clear instructions and information about the procedures. Options and references to procedures can

be found in the documentation. SOLAR is text based, but the display from the program is relatively clear. Also results are showed in a visually clear format. It is easy to understand why SOLAR has become quite a common program because it is rather uncomplicated to use. However, the results need to be visualized with other tools as SOLAR produces only text files.

KELVIN has a moderately understandable manual where most of the needed information can be found, but it is not very well organized. Most of the information is in an unformatted text file that is not very convenient to use. The order of the information is also defective: for example, the file formats are discussed first, but only later is the user told that the information about different chromosomes need to be in different files. However, KELVIN was the only one of the programs where there is also a graphical configuration interface. This configuration interface compensates for problems with the documentations, as most of the options can be found there. The program reports its progress every two minutes, whereas in other programs it is sometimes hard to discern whether the program has crashed or is still running. The result output options are best in KELVIN: it comes with the KELVIZ program that can be used to visualize the results and no other program is needed to plot the results.

The documentation in JPSGCS is, unfortunately, insufficient. Different analysis options are listed on the webpages, but no real manual or information about different options is available. For example, the file format is said to be LINKAGE format with some exceptions, but the exceptions are not described. Also, the output of the results is impractical. The results are outputted as large tables for every family separately. There are no headers or family identifiers to help to organize the results. Perl scripts had to be written in order to parse the results, as they are not usable without parsing. It has to be noted, however, that JPSGCS is only one of the three used programs that includes error-checking abilities. It can be used to find Mendelian errors, to exclude non-informative subjects and to find non-Mendelian genotyping errors. These abilities can reduce the number of programs needed for quality control.

The largest disadvantage with many of the linkage programs is the need for external quality control. Linkage analysis needs deep quality control and only a few of the linkage programs can themselves perform it. For example, MERLIN, MENDEL and LINKAGE programs include quality control procedures, but they are not suited for large, extended families and hundreds of thousands of markers. Large SNP-based genotyping panels are better suited for association studies and linkage programs have actually not been developed for that kind of data. Some progress in this area has taken place, as, for example, MENDEL can nowadays accept data in PLINK binary data format for some of the quality control procedures.

Within this project, quality control took most of the manual working time. Many of the problems needed to be checked by hand. Some Mendelian errors had to be rechecked in the laboratory. Automatic corrections usually exclude all doubtful data. For example, PedCheck excludes all genotypes of a marker in a family, when it finds a Mendelian problem in even one member of that family. However, we did not want to do that, as we wanted to keep the maximum number of individuals in the analysis. Thus, the errors were corrected, if possible. Naturally, this endeavour slowed down the quality control phase.

PLINK was employed in most of the quality control procedures. The largest advantages in using PLINK are the speed and the file formatting abilities. However, it cannot consider intact large families, but splits them into nuclear families. This causes problems in relatedness error checking as well as with Mendelian problems. Also, allele frequencies are estimated only from the founders, meaning subjects whose both parents are marked as missing. In this data, many parents have not participated in the study, which omitted some of the families from the allele frequency estimation. Splitting of the families may also affect the LD estimation, which is used in marker pruning for linkage analyses. To conclude, external quality control may cause problems even though a linkage program is chosen appropriately.

Linkage analysis took a lot of time to compute. The run-times were compared with desktop computers, but in practise it is advisable to use computer clusters for these time consuming calculations. Two-point analysis was relatively fast with JPSGCS, lasting only for hours. On the contrary, with Solar even two-point analysis took days. Solar is slower because of the IBD analysis that is needed for non-parametric linkage. However, these IBD estimates were also used for multipoint analysis, which speeds up the multipoint analysis. The multipoint analyses took weeks with both JPSGCS and SOLAR.

The KELVIN was run on a computer cluster in Vieland's lab, and thus the run time is not comparable to JPSGCS and SOLAR.

7.1.2 Usability of association programs

Association analysis was performed with GenAbel. It is suitable for large families, as it applies mixed models approach: it does not use the pedigree information *per se*, but corrects for genetic relatedness between every pair of individuals. The mixed models method made GenAbel analysis fast to perform. The program has also good manuals and is easy to use. GenAbel allows for direct graphical visualization of the results and thus no additional visualization tools are needed.

Especially when using genetic data as evidence for kinship, there are only a few errors that would have to be checked by hand. Here, I did use PLINK for preliminary data handling and quality control, but GenAbel could also have been used for most of the phases. PLINK was used because GenAbel cannot correct Mendelian errors, which could have raised the number of possible genotyping errors.

Many features are better designed in PLINK and GenAbel than in any linkage program used or tried out in this project. Starting from flexible file formatting to better manuals and good outputs, it seems that association programs are more focused on usability than linkage programs.

However, the availability of different programs that are able to manage extended families seems as limited as with linkage analysis. More traditional association methods, which use original family information, seem to be either suited for nuclear families or at the most for extended families with a qualitative trait as the response variable. Case-control studies do dominate the field of association analysis. Many studies with extended pedigrees have chosen one trio or nuclear family from every extended pedigree to be able to use common programs, naturally losing a lot of information. This phase would also have to be performed by hand. Thus, the newer methods seem more reliable even though they do not preserve the original family information.

The usability of the programs may contribute to the correct application of the methods. The more common methods are usually included in programs where usability is not a problem. Documentation can help to understand the programs and the methods included in it. Good usability can serve as a guide to perform the analyses in better agreement with the latest development on methods and their application.

7.2 Musical aptitude gene mapping results

This study revealed several promising loci for musical aptitude. The best peak was located in 4q12 and it was found with all of the three linkage programs. However, it was not discovered with association method. The linkage peak in 4q12 spans over centromere, including large area with limited recombination. This makes it difficult to say on which side of the centrosome the possible gene affecting musical aptitude lies on. However, the width of the peak may signify for highly significant result despite of only suggestive lod score (Ott and Hoh 2000). There are several brain-affecting genes in the area and the low resolution in linkage analysis makes it hard to point out any specific gene. However, with no significant results from association analysis, we currently have no successful high-resolution results.

7.2.1 No genome-wide significance with association analysis

The data consists of large extended families that are supposedly more powerful in linkage than in association analysis. Thus, the linkage results were expected to be more revealing. Somewhat surprisingly, there were no results of genome-wide significance from association analysis. One might have assumed that at least the best results from linkage analysis would also show in the association analysis.

The lack of significant results in association analysis may be caused by a too small sample size. There were fewer than 800 samples in association analysis and many of them were close relatives with each other. The use of relatives may lower the power of the analysis compared to the same sample size of randomly chosen individuals (Laird and Lange 2006). Thus, the analysis may have been underpowered to detect any associations. Usually samples over 1000 are accepted, even though those may still be too small. The use of quantitative trait may also decrease the power of association analysis. A simpler phenotype, for example with extreme cases, could have been powerful enough even with this small sample size.

Other reasons for association failing can be found from methodological differences between linkage and association methods. If there is heterogeneity between the families, the association analysis will not work, even though the linkage analysis may still be successful. The musical aptitude is also a complex trait and is probably affected by the environment. In family studies, the environment is kept more constant besides the genetic relationship. For example spouses share common environments even though they are genetically different. In the chosen association method, this information is completely lost, which can affect the results.

With the data used in this study, the linkage method seems to be more powerful than the association method. However, a different sample structure or a different kind of trait will change this assumption. Especially different set of pedigrees, small or large, will affect the power of linkage and association methods differently (Visscher et al. 2008; Almasy and Blangero 2009). The ability of both methods to

find or not to find association between a trait and loci is conditional on the features of the data.

7.2.2 Differences between linkage results

There are also differences between the linkage methods. In this study, the most significant results were found with all of the three programs: KELVIN, SOLAR and JPSGCS. Anyway, there was variation in other peaks between the programs. The results of KELVIN and SOLAR resemble each other for most of the locations of the peaks with only some exceptions. The largest difference between these two programs is the linkage found by KELVIN at 18q22. It is third best result found by KELVIN (probability 0.42) and no other program could identify that location. JPSCGS identified 4 families that are linked to that area (the maximum of families linked to any area was 7), but the overall results show no linkage. The disagreement between the programs may indicate false positive result. Results in 4q12 and 4q21 are found with all of the programs and are thus more reliable.

However, there are also methodological differences that may contribute to the disagreement. For example, KELVIN gains information also from family members that have only phenotype and no genotypes whereas other linkage methods, especially IBD based methods like the one used in SOLAR, usually skip these individuals. Because there is more data for KELVIN to use, there can also be true positive results that will be missed by other methods like SOLAR. Also JPSGCS is capable of using these individuals and on 18q22 it did not agree with KELVIN. There are also other differences between the methods, which may contribute to the differences in the results. Anyhow, the results with disagreement between different methods should be considered with caution.

JPSGCS produced the most diverse result compared to other programs. Two out of five best results were not found in either of the other programs (chromosomes 13 and 19). Also, the consistent results were higher than those of SOLAR, even though SOLAR is thought to be liberal (Kleensang et al. 2010). This superiority of the results can be due to the higher power in parametric methods. However, the

superiority of the significant results and lack of repeatability compared to the other programs increase distrust against JPSGCS, which already existed after problems in usability. Nonetheless, the usability problems raise the possibility of errors that can be caused by the user. For example, there are several ways of parsing the multipoint results and the chosen way to maximize them over specific intervals may not be the best one. Thus, even though the actual method could be good it may be executed erroneously.

JPSGCS was the only parametric method used here and the defined parametric model might also be problematic. Only one additional genetic model with two trait allele frequencies was used. Even though the additive genetic model was supposed to be most powerful, it may miss dominant or recessive signals. Here, it seems that the results of JPSGCS are not reliable, but there might be several reasons that are not necessarily program related.

Also, SOLAR results were somewhat different compared to KELVIN and JPSGCS. For example, the JPSGCS and KELVIN results show two separate peaks on the best area in chromosome 4, whereas SOLAR only points to one large area. Overall, the results from SOLAR seem less precise than the other results: the peaks diffuse over a wider area than the peaks from other programs.

In conclusion, the common results between three different linkage methods seem reliable and thus the areas in chromosome 4 seem probable to contribute to musical aptitude. To refine the results, different phenotypes contributing to the combined test scores could be further analysed. There are interesting genes in every area and no clear candidate genes can be identified even from the best area.

7.2.3 Between-family differences

The families differ in size and there is heterogeneity between families that were collected earlier and later. The families collected earlier were larger and had more musicians in them. Differences are also evident when looking at the significance of individual families for linkage results. The larger families are more powerful and

thus, proportionally, affect the results more. This is consistent with earlier studies, where large families have shown to be notably more powerful than small families (Almasy and Blangero 2009).

There are totally three families that gain maximum lod scores over 2 (families number 13, 14 and 15, which are among the five largest families). Additionally, only 22 of 107 families gain lod scores over 0.55. Therefore, most of the families have only limited additive value on the linkage results. There are a total of 4 areas, where per-family lod scores exceeds 2 (data not shown). None of them lie on the area where there is even suggestive result in whole sample results. Heterogeneity analysis (see Section 2.5.2) might help to clarify if those areas have some true influence on musical aptitude.

Also, a separate analysis for the different batches could have helped with the large differences between the families. The batches were now analysed together, even though the separate analysis would have been more appropriate. At least the families from pilot study should have been analysed separately to actually replicate the previous findings.

7.2.4 Comparison to the results of the pilot study

Linkage results from our pilot study (Pulli et al. 2008) were also compared with these new results. The best area from the pilot study seems to map on the end of the best area found in this study (Figure 12). The additional families here have probably defined the area. In the pilot study, there were 15 families with 205 genotyped individuals, when there are now 107 families with 686 genotyped individuals. Some of the former suggestive results were also now rejected: the results from chromosomes 8 and 18 were not renewed in this study.

7.2.5 Overlap between this study and study of absolute pitch

Theusch et al. (2009) found linkage in 8q24.21 (lod 3.5) for absolute pitch. Our results showed linkage with the KELVIN program for 8q24.13 (probability 0.17).

The highest peaks are located 17cM from each other, but the AP linkage peak seems wide enough to include also the area where our linkage peak is located. Thus, the results may indicate the same location. Here, it would be interesting to compare the results of different musicality tests. SPT could have more in common with the AP test than the combined musicality test. Anyhow, this linkage peak was not found with any other program than KELVIN in this study.

7.3 About gene mapping studies

As the results in this study show there are differences between the methods, especially between the association and linkage, which cannot be captured only by power analysis. Various samples, unrelated population samples or different kinds of pedigrees, are best suited for different approaches. Also, heritability and supposed effect size impact on the choice of a suitable method. Published power analyses are usually made for a certain situation and are not applicable for all cases. Thus, it is important to plan the study to be suitable for that particular case.

The number of family studies has been reduced since association studies have become more popular. However, in complex trait studies especially large families help to control for environmental factors. The family members share a somewhat similar environment for their whole life. Also, the genetic background of the sample is most similar, when they are from the same family. These factors favour the family-studies over the unrelated sample studies, especially with extended families. However, some features found in specific families may have no effect on population level.

The large families can also be problematic computationally. It does not help to collect large families, if they cannot be utilized in the analysis. In many studies families have been split in linkage studies or only some of the samples have been used in association study. Some times there is actual lack of suitable methods, but usual there are suitable methods, but more common methods are used instead. As this study shows, there are suitable methods for large, extended pedigrees with

quantitative trait. However, it may be difficult to identify methods capable to analyse certain data. Specific information on data limitations is usually not publicized.

Gene mapping studies with complex traits have been rarely confirmed. Claims of linkage and association discovery have both been difficult to replicate. Some have blamed poorly planned studies and some others have blamed the complex genetics itself. In association studies, there have been demands for increasingly larger samples. With linkage studies, there have been discussions whether families can reveal anything notable on the population level. Better planning and a better understanding of complex genetics will probably more frequently help to confirm studies. But even if the studies are confirmed, they usually explain only a fraction of the observed variation. Thus, there is still a lot to learn about complex traits before we can explain their nature. Gene mapping may help to solve the mystery, but it does not give a full explanation.

8 Acknowledgements

This study was carried out at the Department of Medical Genetics, University of Helsinki. It was funded by the Academy of Finland.

Firstly, I want to thank my supervisors docent Irma Järvelä and docent Päivi Onkamo. I appreciate the support I have got throughout this project. Thank you also for your valuable comments on the thesis.

I want to thank MSc Liisa Ukkola-Vuoti for guiding me for the project and for all the work in the laboratory. DMus Pirre Raijas I thank for her help in the musical theory. I also thank MSc Minna Ahvenainen and MSc Katri Kantojärvi for helping with this project. Also, other colleagues are thanked for encouragement.

KELVIN analyses were performed in collaboration with Veronica Vieland. I like to thank her, and also Huang Yungui, who performed the analyses.

Finally, to all my friends and family, my warmest thanks for understanding and encouragement throughout this work. In particular, thanks for my beloved Otto, who has supported and fed me during this time.

I dedicate this thesis to my late mother.

Helsinki, June 2012

9 References

9.1 Articles

Abecasis, G. R., Cardon, L. R., and Cookson, W. O. (2000). A general test of association for quantitative traits in nuclear families. *Am J Hum Genet*, **66**:279–292.

Abecasis, G. R., Cherny, S. S., Cookson, W. O., and Cardon, L. R. (2002). Merlin– rapid analysis of dense genetic maps using sparse gene flow trees. *Nat Genet*, **30**:97–101.

Allen-Brady, K., Horne, B. D., Malhotra, A., Teerlink, C., Camp, N. J. and Thomas, A. (2007). Analysis of high-density single-nucleotide polymorphism data: three novel methods that control for linkage disequilibrium between markers in a linkage analysis. *BMC Proceedings*, **1** Suppl 1:S160.

Almasy, L. and Blangero, J. (1998). Multipoint quantitative-trait linkage analysis in general pedigrees. *Am J Hum Genet*, **62**:1198–1211.

Almasy, L. and Blangero, J. (2009). Human QTL linkage mapping. *Genetica*, 136:333–340.

Amos, C. I. (1994). Robust variance-components approach for assessing genetic linkage in pedigrees. *Am J Hum Genet*, **54**:535–543.

Anderson, C. A., Pettersson, F. H., Clarke, G. M., Cardon, L. R., Morris, A. P., and Zondervan, K. T. (2010). Data quality control in genetic case-control association studies. *Nat Protoc*, **5**:1564–1573.

Athos, E. A., Levinson, B., Kistler, A., Zemansky, J., Bostrom, A., Freimer, N., and Gitschier, J. (2007). Dichotomy and perceptual distortions in absolute pitch ability. *PNAS*, **104**:14795–14800.

Aulchenko, Y. S., Ripke, S., Isaacs, A., and van Duijn, C. M. (2007). Genabel: an r library for genome-wide association analysis. *Bioinformatics*, **23**:1294–1296.

Aulchenko, Y. (2011). Abel Tutorial. 2007-2010 edition. Available from: http://www.genabel.org/tutorials/ABEL-tutorial

Ayotte, J., Peretz, I., and Hyde, K. (2002). Congenital amusia: A group study of adults afflicted with a music-specific disorder. *Brain*, **125**:238–251.

Bacanu, S.-A., Devlin, B., and Roeder, K. (2002). Association studies for quantitative traits in structured populations. *Genet Epidemiol*, **22**:78–93.

Baharloo, S., Johnston, P. A., Service, S. K., Gitschier, J., and Freimer, N. B. (1998). Absolute pitch: an approach for identification of genetic and nongenetic components. *Am J Hum Genet*, **62**:224–31.

Balding, D. J. (2006). A tutorial on statistical methods for population association studies. *Nat Rev Genet*, **7**:781–791.

Bayes, T. (1763). An essay towards solving a problem in the doctrine of chances. *Philosophical Transactions, Royal Society of London*, **53**:370–418.

Camp, N., Hopkins, P., Hasstedt, S. J., Coon, H., Malhotra, A., Cawthon, R. M., and Hunt, S. C. (2003). Genome-wide multipoint parametric linkage analysis of pulse pressure in large, extended utah pedigrees. *Hypertension*, **42**:322–328.

Cardon, L. R. and Bell, J. I. (2001). Association study designs for complex diseases. *Nature Reviews Genetics*, **2**:91–99.

Chotai, J. (1984). On the lod score method in linkage analysis. *Ann Hum Genet,* 48:359-378.

Cirulli, E. T. and Goldstein, D. B. (2010). Uncovering the roles of rare variants in common disease through whole-genome sequencing. *Nat Rev Genet*, **11**:415–425.

Clarke, G. M., Anderson, C. A., Pettersson, F. H., Cardon, L. R., Morris, A. P., and Zondervan, K. T. (2011). Basic statistical analysis in genetic case-control studies. *Nat Protoc*, **6**:121–133.

Coon, H. and Carey, G. (1989). Genetic and environmental determinants of musical ability in twins. *Behavior Genetics*, **19**:183–193.

Dempster, E. R. and Lerner, M. (1950). Heritability of threshold characters. *Genetics*, **35**:212–236.

Deutsch, D., Henthorn, T., Marvin, E., and HongShuai, X. (2006). Absolute pitch among american and chinese conservatory students: Prevalence differences, and evidence for a speech-related critical period. *J. Acoust. Soc. Am.*, **119**:719–722.

Devlin, B. and Roeder, K. (1999). Genomic control for association studies. *Biometrics*, **55**:997–1004.

Drayna, D., Manichaikul, A., de Lange, M., Snieder, H., and Spector, T. (2001). Genetic correlates of musical pitch recognition in humans. *Science*, **291**:1969–1972.

Elston, R. C. and Stewart, J. (1971). A general model for the genetic analysis of pedigree data. *Hum Hered*, **21**:523–542.

Fisher, R. A. (1918). The correlation between relatives on the supposition of mendelian inheritance. *Transactions of the Royal Society of Edinburg*, **52**:399–433.

Freimer, N. and Sabatti, C. (2004). The use of pedigree, sib-pair and association studies of common diseases for genetic mapping and epidemiology. *Nat Genet*, **36**:1045–51.

Gagnon, F., Jarvik, G. P., Motulsky, A. G., Deeb, S. S., Brunzell, J. D., and Wijsman, E. M. (2003). Evidence for linkage of hdl level variation to apoc3 in two samples with different ascertainment. *Hum Genet*, **113**:522–533.

Goldgar, D. E. and Oniki, R. S. (1992). Comparison of a multipoint identity-by-descent method with parametric multipoint linkage analysis for mapping quantitative traits. *Am J Hum Genet*, **50**:598–606.

Haldane, J. B. S. and Smith, C. A. B. (1947). A new estimate of the linkage between the genes for colour-blindness and haemophilia in man. *Annals of Eugenics*, **14**:10–31.

Hauser, E. R. and Boehnke, M. (1998). Genetic linkage analysis of complex genetic traits by using affected sibling pairs. *Biometrics*, **54**:1238–1246.

Hopper, J. L. and Mathews, J. D. (1982). Extensions to multivariate normal models for pedigree analysis. *Ann Hum Genet*, **46**:373–383.

Horne, B., Malhotra, A., and Camp, N. (2003). Comparison of linkage analysis methods for genome-wide scanning of extended pedigrees, with application to the tg/hdl-c ratio in the framingham heart study. *BMC Genetics*, **4**:S93.

Iles, M. M. (2008). What can genome-wide association studies tell us about the genetics of common disease? *PLoS Genet*, **4**:e33.

Karma, K. (1994). Auditory and visual temporal structuring: How important is sound to musical thinking? *Psychology of Music*, **22**:20–30.

Karma, K. (2002). Auditory structuring in explaining dyslexia. *Language, Vision, and Music: Selected Papers from the 8th International Workshop on the Cognitive Science of Natural Language Processing, Galway, Ireland.*

Karma, K. (2007). Musical aptitude definition and measure validation: ecological validity can endanger the construct validity of musical aptitude tests. *Psychomusicology*, **19**:79–90.

King, M.-C., Lee, G. M., Spinner, N. B., Thomson, G., and Wrensch, M. R. (1984). Genetic epidemiology. *Annu Rev Public Health*, **5**:1–52.

Kirchhubel, J. (2003). *Adolescent Music Development: The Influence of Pre-Tertiary Specialised Music Training*. PhD thesis, Griffith University, Faculty of Education.

Kleensang, A., Franke, D., Alcais, A., Abel, L., Müller-Myshok, B., and Ziegler, A. (2010). An extensive comparison of quantitative trait loci mapping methods. *Hum Hered*, **69**:202–211.

Kruglyak, L., Daly, M. J., Reeve-Daly, M. P., and Lander, E. S. (1996). Parametric and nonparametric linkage analysis: A unified multipoint approach. *Am J Hum Genet*, **58**:1347–1363.

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

Lander, E. S. and Green, P. (1987). Construction of multilocus genetic linkage maps in humans. *Proc Natl Acad Sci USA*, **84**:2363–2367.

Lander, E. and Kruglyak, L. (1995). Genetic dissection of complex traits: guidelines for interpreting and reporting linkage results. *Nat Genet*, **11**:241–247.

Lander, E. S. (1996). The new genomics: global views of biology. *Science*, **274**:536–539.

Lange, K. and Elston, R. C. (1975). Extensions to pedigree analysis I. likehood calculations for simple and complex pedigrees. *Hum Hered*, **25**:95–105.

Lange, K. and Sobel, E. (1991). A random walk method for computing genetic location scores. *Am J Hum Genet*, **49**:1320–1334.

Lange, C., DeMeo, D. L., and Laird, N. M. (2002). Power and design considerations for a general class of family-based association tests: quantitative traits. *Am J Hum Genet*, **71**:1330–1341.

Lathrop, G. M., Lalouel, J. M., Julier, C., and Ott, J. (1984). Strategies for multilocus linkage analysis in humans. *Proc Natl Acad Sci* USA, **81**:3443–3446.

Levitin, D. J. (2012). What does it mean to be musical? *Neuron*, **73**:633–637.

Lin, S., Ding, J., Dong, C., Liu, Z., Ma, Z. J., Wan, S., and Xu, Y. (2005). Comparisons of methods for linkage analysis and haplotype reconstruction using extended pedigree data. *BMC Genet***, 6** Suppl 1:S76.

Lohmueller, K. E., Pearce, C. L., Pike, M., Lander, E. S., and Hirschhorn, J. N. (2003). Meta-analysis of genetic association studies supports a contribution of common variants to susceptibility to common disease. *Nat Genet*, **33**:177–182.

Manolio, T. A., Collins, F. S., Cox, N. J., Goldstein, D. B., Hindorff, L. A., Hunter, D. J., McCarthy, M. I., Ramos, E. M., Cardon, L. R., Chakravarti, A., Cho, J. H., Guttmacher, A. E., Kong, A., Kruglyak, L., Mardis, E., Rotimi, C. N., Slatkin, M., Valle, D., Whittemore, A. S., Boehnke, M., Clark, A. G., Eichler, E. E., Gibson, G., Haines, J. L., Mackay, T. F. C., McCarroll, S. A., and Visscher, P. M. (2009). Finding the missing heritability of complex diseases. *Nature*, **461**:747–753.

Martin, E. R., Monks, S. A., Warren, L. L., and Kaplan, N. L. (2000). A test for linkage and association in general pedigrees: the pedigree disequilibrium test. *Am J Hum Genet*, **67**:146–54.

Matise, T. C., Chen, F., Chen, W., De La Vega, F. M., Hansen, M., He, C., Hyland, F. C. L., Kennedy, G. C., Kong, X., Murray, S. S., Ziegle, J. S., Stewart, W. C. L., and Buyske, S. (2007). A second-generation combined linkage physical map of the human genome. *Genome Res*, **17**:1783–1786.

McCarthy, M. I., Abecasis, G. R., Cardon, L. R., Goldstein, D. B., Little, J., Ioannidis, J. P. A., and Hirschhorn, J. N. (2008). Genome-wide association studies for complex traits: consensus, uncertainty and challenges. *Nat Rev Genet*, **9**:356–369.

Morton, N. E. (1955). Sequential tests for the detection of linkage. *Am J Hum Genet*, **7**:277–318.

Morton, N. E. (1998). Significance levels in complex inheritance. *Am J Hum Genet*, **62**:690–697.

The NCBI handbook [Internet]. (2002). Bethesda (MD): National Library of Medicine (US), National Center for Biotechnology Information; The Reference Sequence (RefSeq) Project. Chapter 18. Available from http://www.ncbi.nlm.nih.gov/entrez/query.fcgi?db=Books

O'Connell, J. R. and Weeks, D. E. (1998). Pedcheck: a program for identifi- cation of genotype incompatibilities in linkage analysis. *Am J Hum Genet*, **63**:259–266.

Ott, J. and Hoh, J. (2000). Statistical approaches to gene mapping. *Am J Hum Genet*, **67**:289–294.

Peng, B. and Kimmel, M. (2007). Simulations provide support for the common disease-common variant hypothesis. *Genetics*, **175**:763–776.

Peretz, I., Cummings, S., and Dubé, M.-P. (2007). The genetics of congenital amusia (tone deafness): a family-aggregation study. *Am J Hum Genet*, **81**:582–588.

Plomin, R., Haworth, C. M. A., and Davis, O. S. P. (2009). Common disorders are quantitative traits. *Nature Reviews Genetics*, **10**:872–878.

Price, A. L., Patterson, N. J., Plenge, R. M., Weinblatt, M. E., Shadick, N. A., and Reich, D. (2006). Principal components analysis corrects for stratification in genome-wide association studies. *Nat Genet*, **38**:904–909.

Price, A. L., Zaitien, N. A., Reich, D., and Patterson, N. (2010). New approaches to population stratification in genome-wide association studies. *Nature Reviews Genetics*, **11**:459–463.

Pulli, K., Karma, K., Norio, R., Sistonen, P., Göring, H. H. H., and Järvelä, I. (2008). Genome-wide linkage scan for loci of musical aptitude in finnish families: evidence for a major locus at 4q22. J Med Genet, **45**:451–456.

Purcell, S., Neale, B., Todd-Brown, K., Thomas, L., Ferreira, M. A. R., Bender, D., Maller, J., Sklar, P., de Bakker, P. I. W., Daly, M. J., and Sham, P. C. (2007). Plink: a tool set for whole-genome association and population-based linkage analyses. *Am J Hum Genet*, **81**:559–575.

Risch, N. J. (2000). Searching for genetic determinants in the new millenium. *Nature*, **405**:847–856.

Saunders, A. M., Strittmatter, W. J., Schmechel, D., St. George-Hyslop, P. H., Pericak-Vance, M. A., Joo, S. H., Rosi, B. L., Gusella, J. F., Crapper-MacLachlan, D. R., Alberts, M. J., Hulette, C., Crain, B., Goldgaber, D., and Roses, A. D. (1993). Association of

apolipoprotein E allele ε4 with late-onset familial and sporadic Alzheimer's disease. *Neurology*, **43**:1467–1472.

Savage, L. J. (1961). The foundations of statistics reconsidered. In J., N., editor, *Proceedings of the fourth Berkeley symposium on mathematical statistics and probability*, volume **1**, pages 575–585. Berkeley: University of California Press.

Seashore, C. (1938). Psychology of music. Dover books on music, music history. Dover Publications.

Service, S., DeYoung, J., Karayiorgou, M., Roos, J. L., Pretorious, H., Bedoya, G., Ospina, J., Ruiz-Linares, A., Macedo, A., Palha, J. A., Heutink, P., Aulchenko, Y., Oostra, B., van Duijn, C., Jarvelin, M.-R., Varilo, T., Peddle, L., Rahman, P., Piras, G., Monne, M., Murray, S., Galver, L., Peltonen, L., Sabatti, C., Collins, A., and Freimer, N. (2006). Magnitude and distribution of linkage disequilibrium in population isolates and implications for genome-wide association studies. *Nat Genet*, **38**:556–560.

Shuter-Dyson, R. and Gabriel, C. (1981). The psychology of musical ability. Methuen, London, 2nd rev. edition.

Smith, C. A. B. (1959). Some comments on the statistical methods used in linkage investiga- tions. *Am J Hum Genet*, **11**:289–304.

Spielman, R. S., McGinnis, R. E., and Ewens, W. J. (1993). Transmission test for linkage disequilibrium: The insulin gene region and insulin-dependent diabetes mellitus (IDDM). *Am J Hum Genet*, **52**:506–516.

Tervaniemi, M., Ilvonen, T., Karma, K., Alho, K., and Näätänen, R. (1997). The musical brain: brain waves reveal the neurophysiological basis of musicality in human subjects. *Neurosci Lett*, **226**:1–4.

Terwilliger, J. D. and Göring, H. H. H. (2009). Gene mapping in the 20th and 21st centuries: Statistical methods, data analysis, and experimental design. *Human Biology*, **81**:663–728.

Theusch, E., Basu, A., and Gitschier, J. (2009). Genome-wide study of families with absolute pitch reveals linkage to 8q24.21 and locus heterogeneity. *Am J Hum Genet*, **85**:112–119.

Thomas, A., Gutin, A., Abkevich, V., and Bansal, A. (2000). Multilocus linkage analysis by blocked gibbs sampling. *Statistics and computing*, **10**:259–269.

Tong, L. and Thompson, E. (2008). Multilocus lod scores in large pedigrees: Combination of exact and approximate calculations. *Hum Hered*, **65**:142–153.

Ukkola, L. T., Onkamo, P., Raijas, P., Karma, K., and Järvelä, I. (2009). Musical aptitude is associated with avpr1a-haplotypes. *PLoS One*, **4**:e5534.

Ukkola-Vuoti, L., Oikkonen, J., Onkamo, P., Karma, K., Raijas, P., and Järvelä, I. (2011). Association of the arginine vasopressin receptor 1a (avpr1a) haplotypes with listening to music. *Journal of Human Genetics*, **56**:324–329.

Ulgen, A. and Li, W. (2005). Comparing single-nucleotide polymorphism marker-based and microsatellite marker-based linkage analyses. *BMC Genetics*, **6** Suppl 1:S13.

Vandenberg, S. G. (1962). The hereditary abilitites study: Hereditary components in a psychological test battery. *Am J Hum Genet*, **14**:220–237.

Vieland, V. J. (1998). Bayesian linkage analysis, or: Hor i learned to stop worrying and love the posterior probability of linkage. *Am J Hum Genet*, **63**:947–954.

Vieland, V. J., Hallmayer, J., Huang, Y., Pagnamenta, A. T., Pinto, D., Khan, H., Monaco, A. P., Paterson, A. D., Scherer, S. W., Sutcliffe, J. S., Szatmari, P., and The Autism Genome

Project (AGP) (2011a). Novel method for combined linkage and genome-wide association analysis finds evidence of distinct genetic architecture for two subtypes of autism. *J Neurodev Disord*, **3**:113–123.

Vieland, V. J., Huang, Y., Seok, S.-C., Burian, J., Catalyurek, U., O'Connell, J., Segre, A., and Valentine-Cooper, W. (2011b). Kelvin: a software package for rigorous measurement of statistical evidence in human genetics. *Hum Hered*, **72**:276–288.

Visscher, P. M., Andrew, T., and Nyholt, D. R. (2008). Genome-wide association studies of quantitative traits with related individuals: little (power) lost but much to be gained. *Eur J Hum Genet*, **16**:387–390.

Weeks, D. E., Lehner, T., Squires-Wheeler, E., Kaufmann, C., and Ott, J. (1990). Measuring the inflation of the lod score due to its maximization over model parameter values in human linkage analysis. *Genet Epidemiol*, **7**:237–243.

Wing, H. D. (1941). A factorial study of musical tests. *British Journal of Psychology*, General Section, **31**:341-355

9.2 Electronic references

Illumina. (2012). Illumina, Inc. Visited 11.5.2012. URL: www.illumina.com,

OMIM. (2012). Mendelian Inheritance in Man, OMIM®. McKusick-Nathans Institute of Genetic Medicine, Johns Hopkins University (Baltimore, MD). Visited 22.3.2012. URL: omim.org

R. (2012). R project for statistical computing. Visited 20.1.2012. URL: www.r-project.org

RefSeq. (2012). NCBI Reference Sequence. Visited 8.6.2012. URL: www.ncbi.nlm.nih.gov/RefSeq

9.2.1 Program manuals

ALLEGRO. Version 2.0. Available from: www.decode.com/software/

CRANEFOOT. Version 3.2.2. Available from: www.finndiane.fi/software/cranefoot/

FASTLINK. Version 4.1P. Available from: www.ncbi.nlm.nih.gov/CBBresearch/Schaffer/fastlink/

GenAbel. Aulchenko, Y. (2011). Abel Tutorial. 2007-2010 edition. Available from: http://www.genabel.org/tutorials/ABEL-tutorial

GENEHUNTER. Version 2.1. Manual available from: linkage.rockefeller.edu/soft/gh/

JPSGCS. Version dated October 2011. Available from: balance.med.utah.edu/wiki/index.php/JPSGCS

KELVIN. Version 2.3.3. Available from: kelvin.mathmed.org/static/doc/

LINKAGE. Version 5.2. Available from: linkage.rockefeller.edu/soft/linkage/

LIPED. Version dated on June 1995. Available from: linkage.rockefeller.edu/ott/liped.html/

LOKI. Version 2.4.5. Available from: www.stat.washington.edu/thompson/Genepi/Loki.shtml/

MENDEL. Version 12. Available from: www.genetics.ucla.edu/software/mendel/

MERLIN. Version 1.1.2. Available from: www.sph.umich.edu/csg/abecasis/Merlin/

MORGAN. Version 3.0.3. Available from: www.stat.washington.edu/thompson/Genepi/MORGAN/morgan303-tut-html/morgan-tut.html/

PEDCHECK. Version 1.00. Available from: watson.hgen.pitt.edu/register/docs/pedcheck.html/

PLINK. Version 1.07. Available from: http://pngu.mgh.harvard.edu/~purcell/plink/index.shtml

SAGE. Version 6.2. Available from: darwin.cwru.edu/sage/?q=node/9/

SIMWALK2. Version 2.91. Available from: www.genetics.ucla.edu/software/simwalk/

SOLAR. Version 4.2. Available from: bioweb2.pasteur.fr/docs/solar/

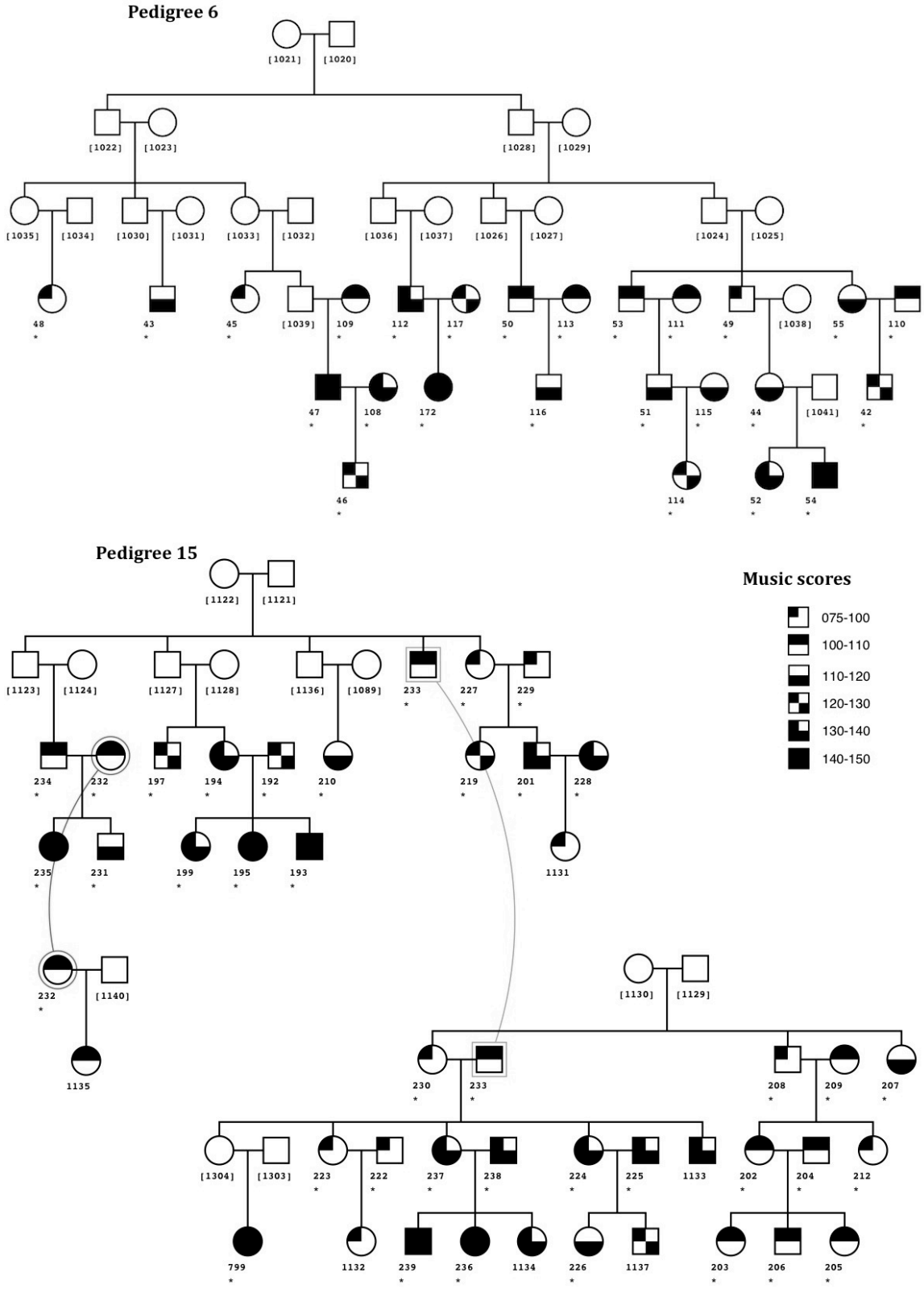
VITESSE. Version 1.0. Available from: watson.hgen.pitt.edu/register/docs/vitesse.html/

Appendices

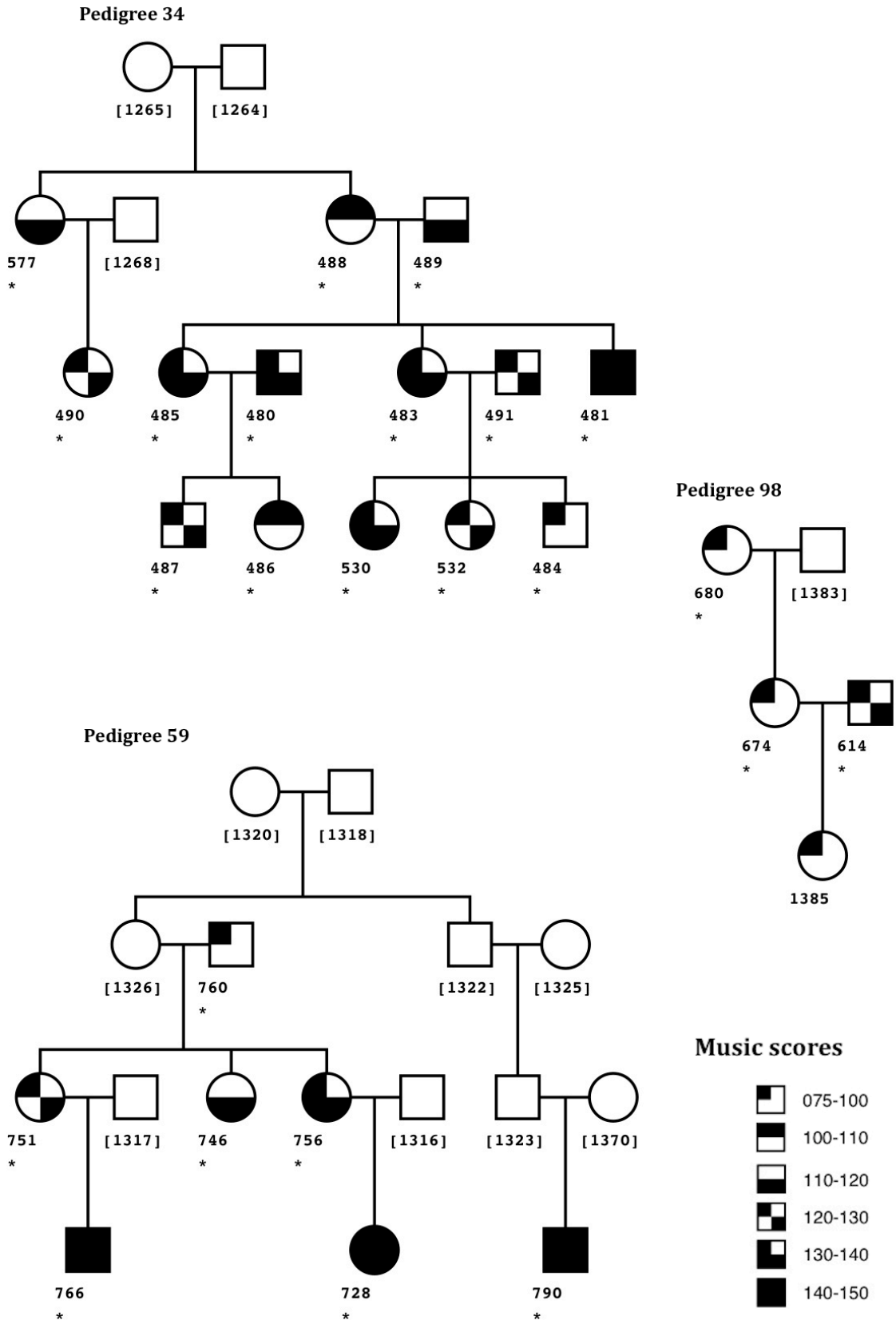
Appendix I: Examples of pedigrees in the different batches

Examples of pedigrees from every collected batch are shown on the following figures. Largest family and some other large families from every batch are included in the figures. The pedigrees are drawn with the Cranefoot program (www.finndiane.fi/software/cranefoot/). Combined music scores are shown as patterns and available dna as * under the individuals. Individuals who have not attended the study (empty individuals) are marked with brackets. Curved links between two nodes assign for same individual who has been drawn twice.

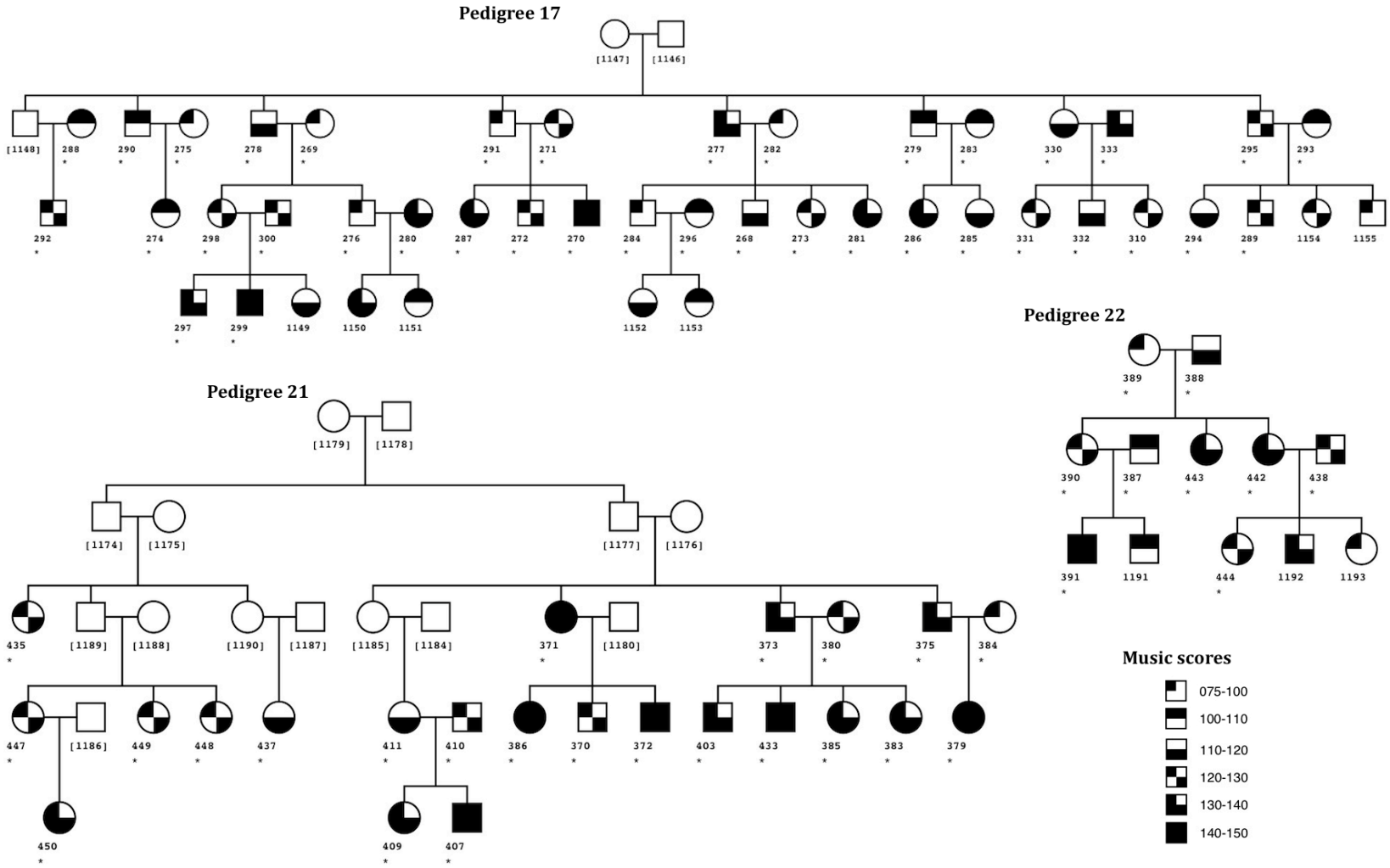
Batch 1


Pedigree 6
[1021]
[1020]
[1022]
[1023]
[1028]
[1029]
[1035]
[1034]
[1030]
[1031]
[1033]
[1032]
[1036]
[1037]
[1026]
[1027]
[1024]
[1025]
48
43
45
[1039]
109
112
117
50
113
53
111
49
[1038]
55
110
47
108
172
116
51
115
44
[1041]
42
46
114
52
54
Pedigree 15
Music scores
075-100
100-110
110-120
120-130
130-140
140-150
[1122]
[1121]
[1123]
[1124]
[1127]
[1128]
[1136]
[1089]
233
227
229
234
232
197
194
192
210
219
201
228
235
231
199
195
193
1131
232
[1140]
1135
[1130]
[1129]
230
233
208
209
207
[1304]
[1303]
223
222
237
238
224
225
1133
202
204
212
799
1132
239
236
1134
226
1137
203
206
205

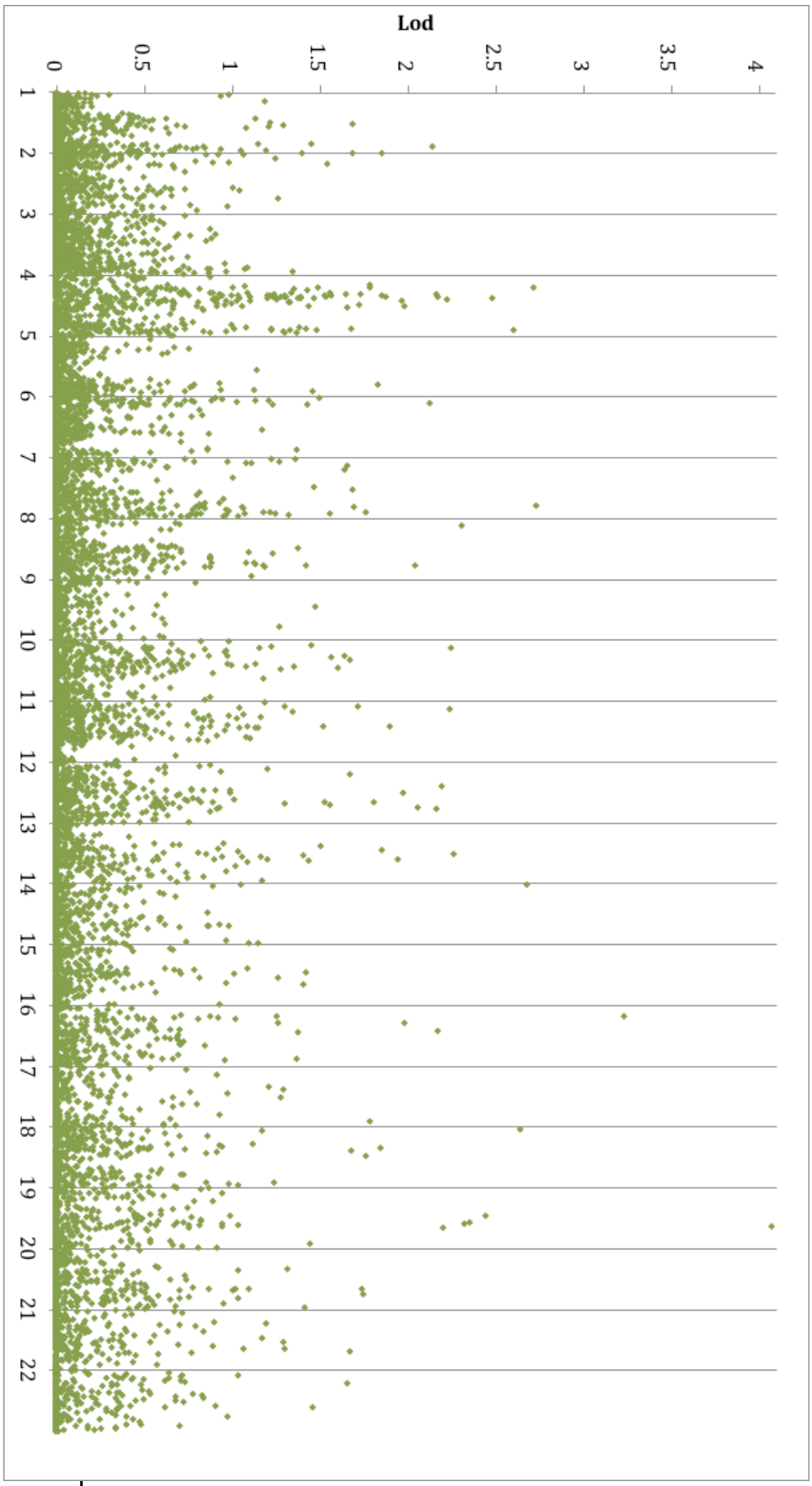
Batch 3


Pedigree 34
[1265]
[1264]
577
*
[1268]
488
*
489
*
490
*
485
*
480
*
483
*
491
*
481
*
487
*
486
*
530
*
532
*
484
*
Pedigree 98
680
*
[1383]
674
*
614
*
1385
Pedigree 59
[1320]
[1318]
[1326]
760
*
[1322]
[1325]
751
*
[1317]
746
*
756
*
[1316]
[1323]
[1370]
766
*
728
*
790
*
Music scores
075-100
100-110
110-120
120-130
130-140
140-150

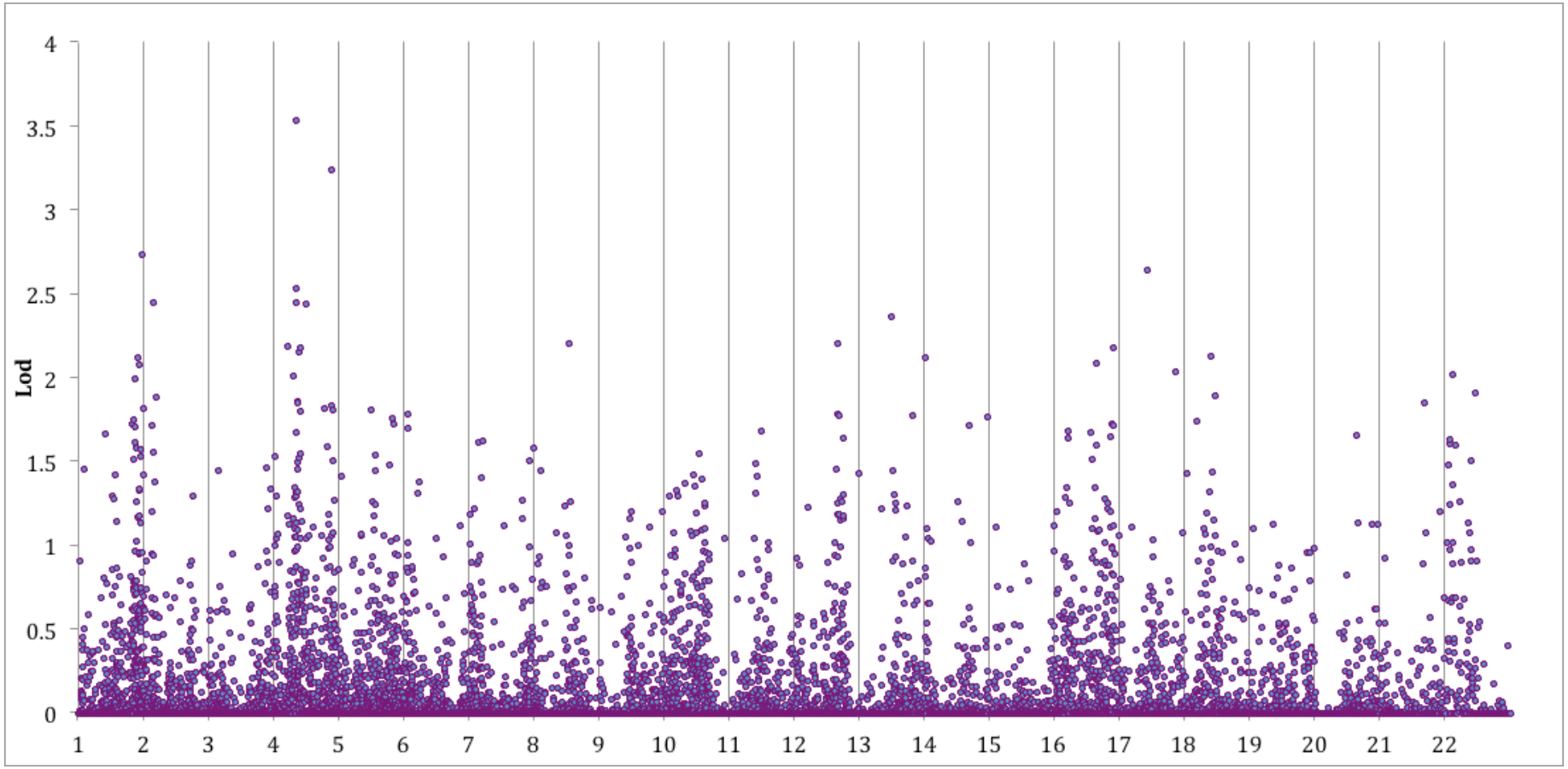
Batch 2


Pedigree 17
[1147]
[1146]
[1148]
288
290
275
278
269
291
271
277
282
279
283
330
333
295
293
292
274
298
300
276
280
287
272
270
284
296
268
273
281
286
285
331
332
310
294
289
1154
1155
297
299
1149
1150
1151
1152
1153
Pedigree 22
389
388
390
387
443
442
438
391
1191
444
1192
1193
Pedigree 21
[1179]
[1178]
[1174]
[1175]
[1177]
[1176]
435
[1189]
[1188]
[1190]
[1187]
[1185]
[1184]
371
[1180]
373
380
375
384
447
[1186]
449
448
437
411
410
386
370
372
403
433
385
383
379
450
409
407
Music scores
075-100
100-110
110-120
120-130
130-140
140-150

Appendix II: JPSGCS two-point linkage result



Appendix III: SOLAR two-point linkage results



Appendix IV: Perl code to convert Illumina data

```
# 16/1/2012
# NB. Uses alleles from Illumina design strand!
#  Printing time of start
$now = localtime;
print "$now\n";
unlink ("musgen.map"); unlink ("musgen.lgen"); unlink ("musgen_draft.fam");
unlink ("musgen_GW_data_without_deleted.txt");

#### READING FINAL RESULT FILES ####
print "
   ************************************
   * Reading Illumina final report files *
   *     Takes time, have coffee         *
   *       Jaana Oikkonen, 2012          *
   ************************************
";
open (FINAL_RESULT, "Irma_Jarvela_MusGen_Project_28032011_FinalReport.txt") ||
die "couldn't open the final result file!";
open (OUTFILE, ">musgen_GW_data_without_deleted.txt");
@delete_persons = qw(429 414 417 431 418 451 425 694 724 858 492 428 104 224
999 849); # Bad result according to genotyping lab
my %bad_persons = map { $_, 1 } @delete_persons;
$rows_deleted = 0;
@delete_persons = ();
## Headings
while($row = <FINAL_RESULT>) { # prints lines (1-10) of the header
       print $row;
       last if $. == 10;
}
open (SNPFILE, ">musgen_illumina_snp_info.txt");
## Data
while($row = <FINAL_RESULT>) { # rest of the file, successfully persons with
design and Plus alleles
       $row =~ s/\R//g; # deleting every kind of newline (also windows CR)
       my($SNP, $person, $allele1_top, $allele2_top, $GCscore, $empty, $famid,
       $sampleID, $SNPid, $SNPAux, $all1_forward, $all2_forward, $all1_design,
       $all2_design, $allele1AB, $allele2AB, $Chr, $position, $GT_Score,
       $Cluster_Sep, $SNP_alleles, $strand_ilmn, $strand_own,
       $Top_Genomic_Sequence, $Theta, $R, $X, $Y, $X_Raw, $Y_Raw,
       $B_Allele_Freq, $Log_R_Ratio, $CNV_Value, $CNV_Confidence,
       $Allele1_Plus, $Allele2_Plus, $Strand) = split(/\t/, $row, 37);
       ## Subjects and their genotypes
       # only successful persons (as genotyping lab reported)
       if ( !$bad_persons{ $person }) {
              # Deleting first character "M" from the family id
              $newfam = int(substr($famid, 1));
              print OUTFILE "$newfam\t$person\t$SNP\t$all1_design\t$all2_design
              \t$Allele1_Plus\t$Allele2_Plus\n";
       }
       else { $rows_deleted += 1; } # Counting the amount of deleted rows
       ## SNP information
       if ($sampleID == 1) { # printing SNP information from first person data,
       assuming all SNPs in this person
              print SNPFILE
              "$SNP\t$SNPid\t$Chr\t$position\t$SNP_alleles\t$Strand\t$Theta\t$R
              \t$X_Raw\t$Y_Raw\t$B_Allele_Freq\t$Log_R_Ratio\t$GT_Score\n";
}}
close FINAL_RESULT; close SNPFILE;
print "\n    Rows deleted: $rows_deleted
    First final result file closed\n\n";
## Replace unsuccessful samples
print "Adding results from 12x file (renewed samples) \n";
open (FINAL_RESULT2, "Irma_Jarvela_12x_repeats_05052011_FinalReport.txt") ||
die "couldn't open the 12x final result file!";
$rows_deleted = 0;
## Heading
```

```
print "    Heading of the second file:\n";
while($row = <FINAL_RESULT2>) {
       print $row;
       last if $. == 10;
       }
## Data
while($row = <FINAL_RESULT2>) {
       $row =~ s/\R//g; # deleting every kind of newline (including windows cr)
       chomp $row; # deleting newline (only lf)
       my($SNP, $person, $allele1_top, $allele2_top, $GCscore, $empty, $famid,
       $sampleID, $SNPid, $SNPAux, $all1_forward, $all2_forward, $all1_design,
       $all2_design, $allele1AB, $allele2AB, $Chr, $position, $GT_Score,
       $Cluster_Sep, $SNP_alleles, $strand_ilmn, $strand_own,
       $Top_Genomic_Sequence, $Theta, $R, $X, $Y, $X_Raw, $Y_Raw,
       $B_Allele_Freq, $Log_R_Ratio, $CNV_Value, $CNV_Confidence,
       $Allele1_Plus, $Allele2_Plus, $Strand) = split(/\t/, $row, 37);
       $newfam = int(substr($famid, 1)); # Deleting first character "M" from
       the family id
       if ($person) {
              print OUTFILE "$newfam\t$person\t$SNP\t$all1_design\t$all2_design
              \t$Allele1_Plus\t$Allele2_Plus\n";
       }
       else { $rows_deleted += 1; }
}
print "    Rows deleted from 12x file: $rows_deleted \n";
$rows_deleted = 0;
close FINAL_RESULT2; close OUTFILE;

print "
   ************************************
   *      Converting the file into       *
   *         PLINK long format           *
   ************************************
";
####### MAKING LGEN FILE #######
## Family IDs:
open (IDFILE, "music_id_2012.txt") || die "couldn't open the family id file!";
open (FAMOUT, ">musgen_draft.fam");
%ids = ();
%missing_family = ();
$realfam = 0;
# Getting new family info into hash (key/value pairs)
while($row = <IDFILE>) {
       chomp $row;
       my($fam, $id, $fid, $mid, $gender, $music) = split(/\t/, $row, 6);
       $ids{ $id } = $fam;
       print FAMOUT "$fam $id $fid $mid $gender $music\n";
}
close IDFILE;
# famid 999 for unrelated individuals
## Genotypes
open (MYINFILE, "musgen_GW_data_without_deleted.txt") || die "couldn't open
the GW data file!";
open (PLINKOUT, ">musgen.lgen");
while ($row = <MYINFILE>) {
       chomp $row; # deleting newline (only LF!)
       # Reads line into variables:
       my($famid, $person, $SNP, $allele1, $allele2, $Allele1_Plus,
       $Allele2_Plus) = split(/\t/, $row, 7);
       # Putting every line into LGEN file with new family IDs
       if(exists $ids{$person}) {$realfam = $ids{ $person };}
       else {
              $realfam = 999;
              if (! exists $missing_family{$person}) {
                     $missing_family{ $person } = $famid;
              }
              print FAMOUT "999 $person 0 0 0 -9";
       }
       print PLINKOUT "$realfam\t$person\t$SNP\t$allele1\t$allele2\n";
}
```

```
close (PLINKOUT); close (MYINFILE); close (FAMOUT);
print "    LGEN file done!\n    Famid 999 for unknown family\n";
# Printing out persons without family information for checking
if(%missing_family) {
       print "\n    Persons without family (check if true):\n ";
       while ( my ($key, $value) = each(%missing_family) ) {
              print "ID $key\t old fam $value\n";
       }
       print "\nThese persons marked with missing gender and phenotype in FAM
       file\n";
}
else { print "\n    All persons have family information"; }
# Deleting trash
@ids = ();
for (keys %missing_family) { delete $missing_family{$_}; }

####### MAKING MAP FILE ######
# Illumina send data with imperfect map information (includes NULL and chr 0)
# Better map data from Rutgers (including cM information)
# Only chromosome heads needs to be calculated separately
print "\n **** Making new and good map file **** \n";
open (RUTGERS, "rutgers.map") || die "couldn't open the file!";
open (MAPOUT, ">musgen.map");
$chromosome = 0; # previous chromosome
$prev_cM = 0; # previous SNP location in cM
$prev_bp = 0; # previous SNP location in bp
$add_cM = 0; # within chromosome cM change from the Rutgers values
$start = 1; # chromosome start with missing data (1) or other situation (0)
$original = 0; # used to calculate $add_cM
# Rutgers map with computational cM information for chromosome heads
#    Heads that are missing from Rutgers are computed from bp info
while($row = <RUTGERS>) {
       chomp $row; # deleting newline (only lf)
       $row =~ s/\R//g; # deleting every kind of newline (including windows cr)
       my($SNP, $chr, $position, $cM, $cM_fem, $cM_male) = split(/,/, $row, 6);
       if ($chr ne $chromosome) { # first SNP in chromosome
              if ($cM eq "NA") { # when no cM data
                     $cM = 0;
                     $start = 1;
              }
              else { $start = 0; } # if no missing cM data at start
              $add_cM = 0; # At first SNP there is nothing to add
       }
       else { # when chromosome continues
              # in the middle, most of SNPs:
              if ($start == 0 && $cM ne "NA") { $cM = $cM + $add_cM; }
              # when no cM, end or start:
              elsif ($cM eq "NA") {
                     $cM = (int($position) - $prev_bp) / 1000000 + $prev_cM;
              }
              # First SNP with nonmissing cM value (end of start)
              elsif ($start == 1 && $cM ne "NA") {
                     $original = $cM;
                     $cM = (int($position) - $prev_bp) / 1000000 + $prev_cM;
                     $add_cM = $cM - $original; # Adding to original values
              }
              else { print "Houston, we have a problem\n"; } # Should never go
              here
       }
       $prev_cM = $cM;
       $prev_bp = $position;
       $chromosome = $chr;
       $cM = sprintf("%.8f", $cM); # only 8 decimals used (program limitations)
       print MAPOUT "$chr\t$SNP\t$cM\t$position\n"; # Always the same out
}
print "\nMap file ready \n";
close RUTGERS; close MAPOUT;
print "
Fam file may include missing data and is named musgen_draft.fam
Other files part of musgen named long format file set (lfile) for PLINK
```

```
Using Design alleles and Rutgers hg18 map \n\n ...Good bye!... \n";
# Printing time
$now = localtime;
print "Script ended $now\n\n";
```