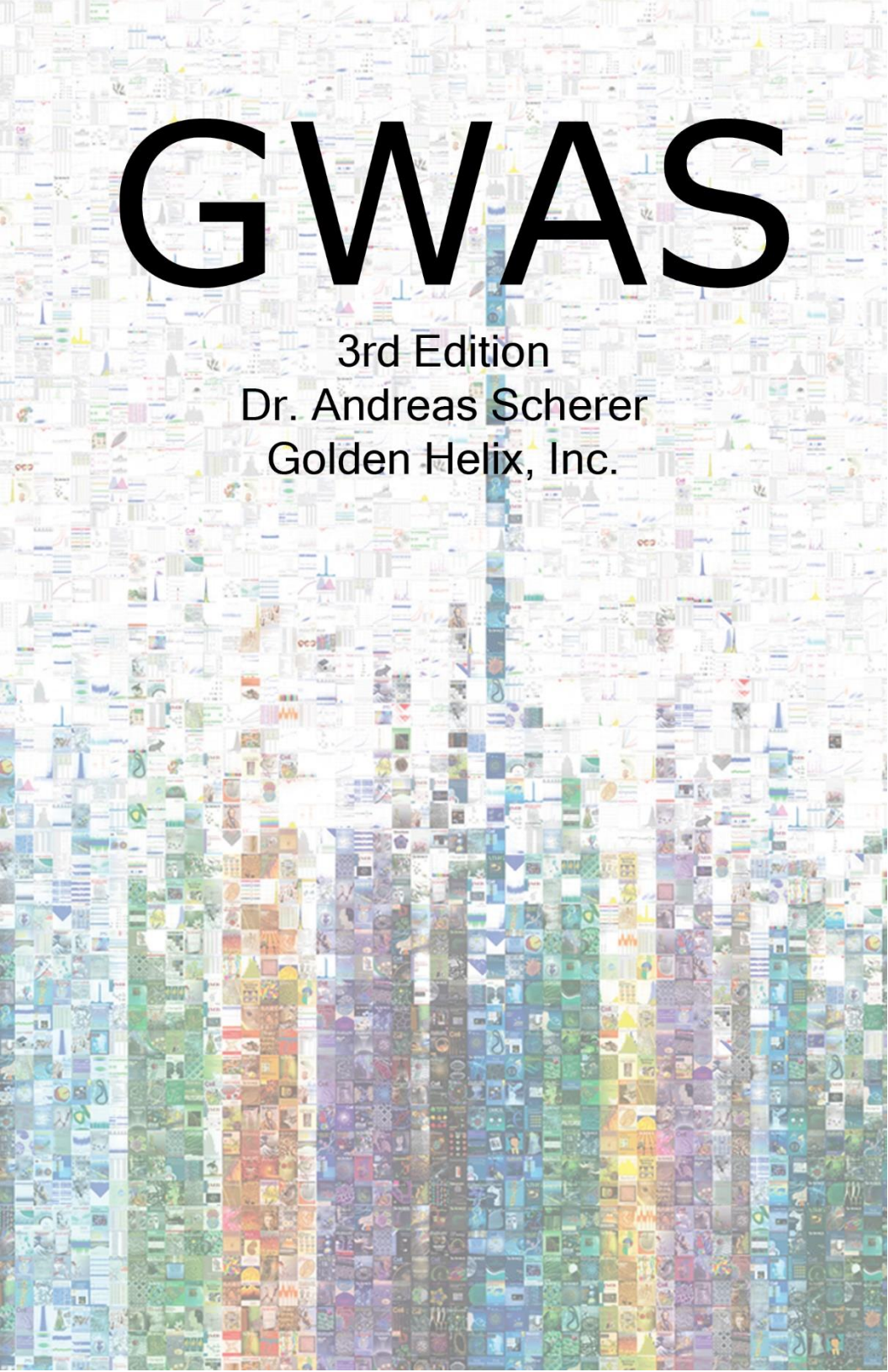
GWAS

3rd Edition

Dr. Andreas Scherer
Golden Helix, Inc.



Copyright © 2017 Andreas Scherer
All rights reserved.
ISBN: 978-0-9986882-0-6

Table of Contents

Preface

Genome-wide association study (GWAS) technology has been a primary method for identifying the genes responsible for diseases and other traits for the past ten years. GWAS continues to be highly relevant as a scientific method. Over 2,000 human GWAS reports now appear in scientific journals.

In fact, we see its adoption increasing beyond the human-centric research into the world of plants and animals. GWAS studies have been beneficial in agrigenomics for identifying genes associated with milk production in the dairy industry, coat color in sheep, along with identifying disease resistance in plants. Identifying the genes of interest for these traits allows farmers to selectively breed for the more desirable trait.

This ebook aims to explain the basic steps and concepts to complete a GWAS experiment and address how these steps are implemented in SVS. In Chapter 1 we start with an introduction to GWAS exploring its biology and origins as well as the practical use of GWAS. Next, we will look at performing a GWAS in the context of the SVS software, discussing quality control, including sample statistics, heterozygosity, LD pruning, population stratification and identity by descent. We also take a look at how to impute data within SVS. From there we move on to genotype association testing and we close with a walk through conducting a Meta-Analysis.

A lot of people in Golden Helix have contributed to this book. It would have been impossible to write without the ingenious work of our product developers who spent many years refining SVS to its current state. Specifically, I'd like to thank Gabe Rudy and Cheryl Rogers for their invaluable contributions. In addition, I am very grateful of the support I received from Dr. Jeffrey Moore, University of Illinois, and Dr. Marcella Devoto, Children's Hospital of Philadelphia.

Andreas Scherer
February 2017
Bozeman, Montana

Chapter 1
Introducing Genome-wide Association Studies

Genome Wide Association Studies (GWAS) were initially developed to study the human genome. The human genome is a sequence of more than three billion DNA bases consisting of four letters: A, C, G or T. Much of the genome sequence is identical or highly conserved across the human population, but every person's genome is unique. On average, a given person's genome sequence is likely to differ from the standard human reference genome at over three million positions. New mutations are introduced to the genome with every passing generation, and there are many old mutations that are now widely observed among all populations. These common mutations are generally called variants or polymorphisms.

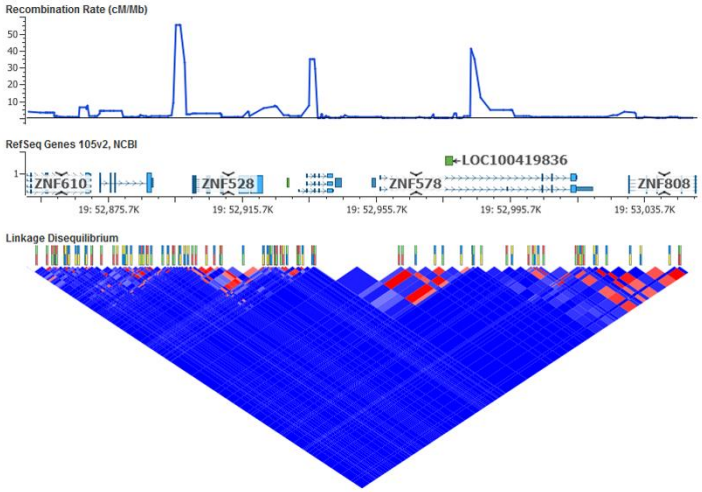


Figure 1: Haplotypes and recombination

The most common type of variants is the single-nucleotide polymorphisms (SNPs). These are changes to an individual DNA base. The different forms of the same gene containing variable SNPs within the same site(s) are typically called alleles. GWAS methods are chiefly concerned with determining alleles associated with various SNPs in each study subject, and making statistical comparisons to identify SNPs or genes associated with a particular trait. If a certain allele is more common among individuals with disease than other healthy ones, this is

interpreted as an evidence that this allele or perhaps another nearby variant may cause the disease or at least increase risk for the disease.

Most SNPs result from one historical mutation event. Because of this ancestry, each new allele is initially associated with the other alleles present on the particular chromosomal background where it arose. The specific set of alleles observed together on a single chromosome, or part of a chromosome, is called a haplotype. New haplotypes are formed by additional mutations or by chromosome recombination (also called crossing-over) during meiotic cell division. Haplotypes tend to be conserved, especially among individuals with recent shared ancestry as can be seen in Figure 1. This figure shows a small region of human chromosome 19. Genes and chromosome physical map coordinates are shown in the middle. The line tracing at the top shows the recombination rate determined from HapMap data—the peaks represent "hotspot" locations for meiotic recombination. The triangular plot in the lower section illustrates linkage disequilibrium (LD) patterns among SNPs in the region, with strong LD shown in red. LD measurements are based on genotypes from 649 individuals of European ancestry. Note that high LD is confined to regions of minimal historic recombination, and does not extend across the recombination hotspots.

Haplotype conservation is a very important factor for GWAS. The genetic variant that causes a particular trait may not be directly tested in the GWAS, but its signature may still be evident through the association of SNPs occurring within the same haplotype (see Figure 2).

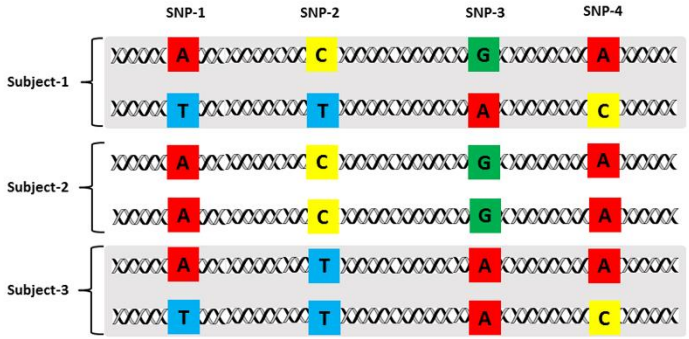


Figure 2: Genotypes and haplotypes

This illustration depicts four SNP loci in the genomes of three subjects. Each subject has two haplotypes, corresponding to the two copies of each chromosome typically present in human cells. Suppose that the C allele at SNP-2 causes a certain trait, but that SNP is not genotyped. The G allele at SNP-3 always occurs on the same haplotype with the causal allele, and if genotyped may serve as a proxy for the causal allele in GWAS tests. Further inspection shows that the causal allele always occurs on the A-C-G-A haplotype, and may also be detected via haplotype association testing. The nonrandom co-occurrence of alleles within a chromosome or haplotype is called linkage disequilibrium, or LD. The degree of LD in a population is shaped by selection, recombination rate, mutation rate, consanguinity, and other factors.

The Origins of GWAS

GWAS became possible as the result of several scientific advances early in the 21st century. The completion of the Human Genome Project greatly improved our knowledge of the human genome and provided a much better context for the study of genetic variants[i]. The International HapMap project, which completed its first phase in 2005, conducted an unprecedented SNP discovery initiative and provided the first detailed human haplotype and LD maps[ii]. These scientific efforts made it possible to identify relatively small numbers of SNPs capable of representing most of the common variation in the human genome. The GWAS era was born as biotechnology companies including Affymetrix, Illumina

and Perlegen launched competing platforms to simultaneously genotype hundreds of thousands of SNPs.

	Primary GWAS Cohort Size			**Replication Cohort Size, if used**		
Year	**Number of Studies**	**Mean** of genotyped subjects	**Median** of genotyped subjects	**Number of Studies**	**Mean** of genotyped subjects	**Median** of genotyped subjects
2005	2	738	738	1	664	664
2006	8	862	821	5	3816	1584
2007	89	2454	1094	63	5957	2519
2008	147	5100	1983	114	9619	4981
2009	235	5748	1984	182	8060	3311
2010	330	7360	2383	223	10733	3835
2011	390	6881	2643	279	9390	3491
2012	382	7575	2662	256	9811	4000
2013	376	8708	2243	252	11276	3609

Table 1: Growth of GWAS

The National Human Genome Research Institute and the European Bioinformatics Institute (NHGRI-EBI) GWAS Catalog[iii] recognizes a 2005's analysis of age-related macular degeneration (AMD) as the first GWAS study. This study analyzed about 100,000 SNPs in just 146 subjects, and identified the *cfh* gene as a major AMD risk factor[iv]. Since then, GWAS has grown to produce hundreds of published reports each year. The volume of published human GWAS studies has plateaued in recent years, but the average size of the study cohorts continues to grow as shown in Table 1. This table shows the number of unique human GWAS papers published per year from 2005 to 2013 according to the NHGRI-EBI GWAS catalog, together with the mean and median number of genotyped subjects analyzed. The number of those reports that included an independent replication cohort is also shown, together with the mean and median number of genotyped

samples analyzed in the replication stage. The largest GWAS studies today may include over 100,000 subjects.

The Practice of GWAS

GWAS studies can be designed to assess the genetic determinants of almost any qualitative or quantitative trait. Several issues must be considered in GWAS study design, including the selection of a genotyping platform, sample size and collection, statistical analysis plans, statistical power, correction for multiple testing and population structure.

Genotype data for GWAS are usually produced with microarray technology allowing the detection of polymorphisms within a population. Microarrays involves three basic principles:

1. The array contains immobilized allele-specific oligonucleotide probes, which are short pieces of synthesized DNA complementary to the sequence of the target DNA.
2. Fragmented nucleic acid sequences of the target, labeled with fluorescent dyes.
3. A detection system that records and interprets hybridization signals measuring essentially genetic similarity.

There are many different microarrays or "chips" available for both human and non-human applications. Some chips are designed to test as many SNPs as practically possible – currently up to about five million. Some chips are specifically designed to test SNPs in coding regions of genes, which make up about 2% of the genome. Other chips may test relatively small numbers of SNPs that have been carefully selected to efficiently represent worldwide haplotype diversity. Some chips are designed for specific ethnic groups or may be enriched with SNPs from genes implicated in particular diseases. In selecting a genotyping chip, it is important to consider the goals of the current project, compatibility with data from past or planned future studies, and the budget available.

The next endeavor required for an effective GWAS study is the collection and recording of the desired phenotype, which can be quantitative (integer or real-valued) or dichotomous (case/control). Quantitative traits can provide more statistical power to show a genetic effect, but the case/control study design can also be effective in identifying multiple genes associated with the phenotype. We see examples of each in the literature [v, vi, vii].

The statistical analysis of genome-wide association can begin once samples have been collected and genotyped. The process begins with a thorough quality control analysis to confirm accuracy of the genotype data[viii]. A statistical hypothesis test is performed for each SNP, with the null hypothesis of no association with the phenotype. There are a number of association tests available depending on which type of trait is being tested. Quantitative traits are generally analyzed using linear regression approaches with the assumptions that the trait is normally distributed, variance within each group is the same, and the groups are independent. Popular analyses include ANOVA and GLM. Binary traits are commonly analyzed using logistic regression, or tests such as a χ^2 or Fisher's Exact Test; logistic regression is popular because it allows adjustment for other covariates[ix]. Specialized tests are available for study designs with family-based collection[x].

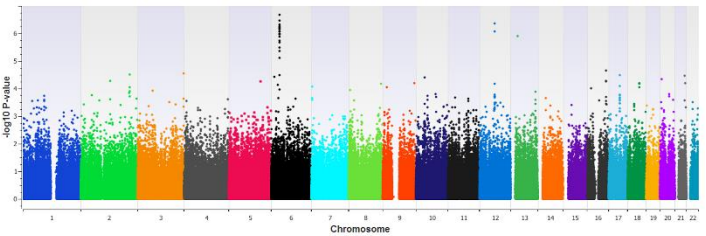


Figure 3: Example Manhattan Plot. GWAS results are often visualized by plotting p-values on a logarithmic scale. The values are plotted in linear order based on the chromosomal locations of the SNPs. This type of figure is commonly called a "Manhattan Plot," alluding to its similarity with a city skyline. In the example above, the most significant SNP in the GWAS is on chromosome 6, with the highest -log10 p-value of 6.68 in the plot.

Statistical power and multiple test correction are important and inseparable issues for GWAS. False positive associations are a great risk when testing large numbers of SNPs, so statistical evidence for association must be held to a high standard. The typical significance threshold used in human GWAS studies is p-

value less than 5e-8, equivalent to a standard Bonferonni correction for one million independent tests[xi]. Populations with greater genetic diversity, such as African populations, may require even greater stringency to determine that a test result is statistically significant. Very large sample sizes may be required to achieve such significance levels, especially for rare disease alleles and alleles with small effect sizes as seen in Figure 4. Power was estimated using the PBAT[xii] Power Calculator implemented in Golden Helix SNP and Variation Suite (SVS)[xiii]. Both figures show the statistical power to detect a true association for a dichotomous trait with significance level $p < 5e\text{-}8$, using an additive genetic model when the true mode of inheritance is also additive. Power is estimated using a simulation procedure for disease allele frequencies between 0.01 and 0.49. Power generally increases when the disease allele has higher frequency. The figure on the left shows the effect of increasing sample size when the effect size is held constant; OR1 (the odds ratio associated with having one copy of the disease allele versus no copies) is fixed at 1.5. The figure on the right shows the power difference to detect causal alleles with various effect sizes. The sample size in this figure is fixed at 1000 cases and 1000 controls.

Statistical power of GWAS is affected by many factors, some of which are beyond the investigator's control. These factors include: complexity of the genetic architecture of the phenotype, frequency and effect size of the disease allele, accuracy of phenotypic measurements and homogeneity of the phenotype, and LD relationships between causal variants and genotyped SNPs[xiv].

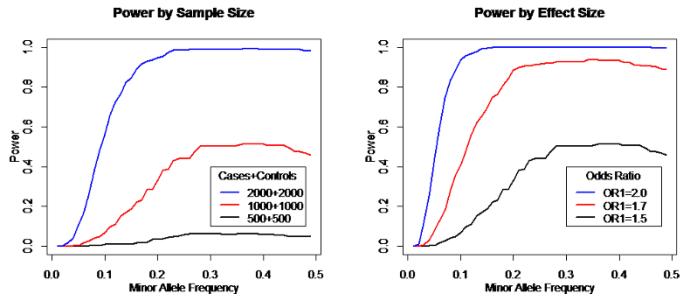


Figure 4: Statistical power in GWAS

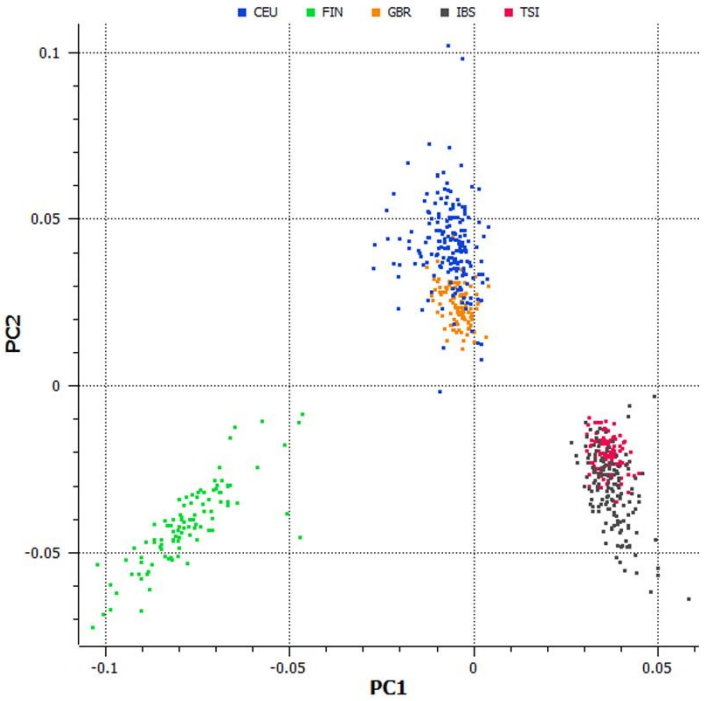


Figure 5: Principal Components Analysis

Standard GWAS test statistics assume that all samples in the analysis are unrelated and selected from a uniform, random-mating population. Any departure from this assumption can cause unexpected results, especially in large study cohorts (groups of subjects encountering a certain event during a particular time period). For example, if individuals of a certain ethnicity are overrepresented in the control group of an experiment, the significance of test results throughout the genome may be consistently inflated due to the unique genetic background of that ethnic group. Principal components analysis (PCA) can be used to stratify subjects based on genomic similarity, and is often used to assess population stratification in GWAS cohorts as shown in Figure 5. This figure shows the first (PC1) and second (PC2) principal components of the GWAS data for a group of samples with European ancestry. The samples are clearly stratified by ancestry and nationality. Samples are colored according to ancestry and geography: CEU = Utah residents (CEPH) with northern and western European ancestry; FIN = Finnish in Finland; GBR = British in England and Scotland; IBS = Iberian population in

Spain; TSI = Toscani in Italy. It is a common practice to adjust GWAS tests for principal components in order to account for the structure of the population. An alternative to PCA-based correction is to account for pairwise allele sharing among all study subjects using mixed linear model (MLM) regression[xv]. MLM methods such as EMMAX[xvi] and GEMMA[xvii] effectively account for population structure in both human and agricultural populations.

Beyond GWAS

GWAS is sometimes called a "hypothesis-generating" process[xviii], as it is often the first step toward understanding the genetic architecture of traits. A successful GWAS will result in one or many SNPs found to be associated with the trait of interest. Researchers may then evaluate the functional consequences of each associated SNP, examine other variants in LD with that SNP, study the function of the gene where the SNP resides, and study the biological pathways in which the gene participates. Indeed, a great number of experiments may be required to fully understand the results of a GWAS. As the biology of the trait is elucidated, it may be possible to develop assays to test for disease risk or to improve disease treatment and prevention programs.

The first decade of GWAS provided many success stories, but debates continue about how to improve GWAS[xix]. Many approaches have been proposed to increase statistical power, reduce false-negative rates, and incorporate biological context in GWAS results[xx]. The coming years are likely to see continued innovations in both technology and analytic methods to make GWAS an even more effective and efficient method to study the underlying biology of diseases and other traits.

Chapter: 2

Conducting a GWAS in SNP and Variation Suite (SVS)

SNP & Variation Suite (SVS) is a project oriented program for the management and analysis of genomic datasets. Both statistically and visually, researchers using SVS can explore the relationships among vast amounts of clinical patient data, environmental factors, and genetic variants to understand the causes of disease and other inherited traits. Applications include candidate gene analysis, genome-wide association studies, copy number analysis, cytogenetic research, and next-generation sequencing studies.

An example GWAS project is available for download and the results of each analysis mentioned below can be explored and visualized with the free SVS viewer.

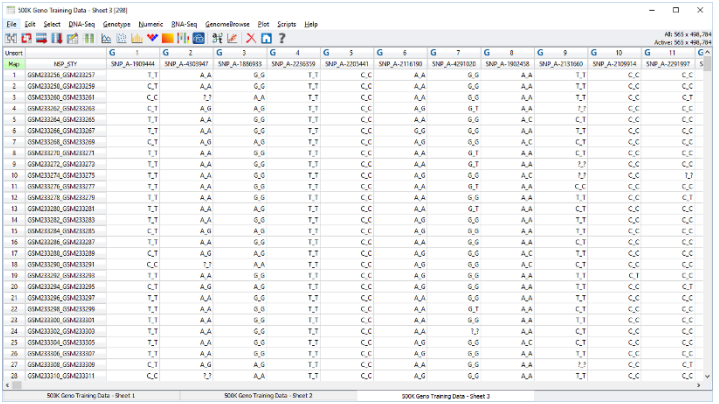


Figure 6: This screenshot (and subsequent) show a simulated dataset from several studies, available through the NCBI Geo database, which used either the Illumina 550K or 610 assay. SVS uses a spreadsheet infrastructure to handle and manage data.

SVS has a variety of features to edit, manipulate, and enrich the data. Immediately upon opening a dataset, the top right corner displays the number of rows (usually the samples) and columns (the phenotypes, environmental factors and/or SNP data). SVS provides a project management interface (Project Navigator) that displays the spreadsheets or plots within a single project. To enable tracking and workflow replication, SVS takes detailed notes on each step that has been performed in a running log that can be accessed from the Project Navigator interface.

One of the first thoughts researchers face when working with genome-wide association studies is the many different data formats available. SVS can handle all standard formats from platforms including Illumina, Affymetrix, and most open source or freeware tools. The software supports spreadsheet-style views of all data types.

Chapter: 3
Quality Control

Sample Statistics

When working with a dataset for the first time, it is helpful to view sample statistics in order to determine information about the samples, where they come from, and if there are any distinguishing features that would set them apart (outliers). The following sections lead researchers through quality assurance procedures that help identify samples of poor quality (low call rates, abnormal heterozygosity, etc.) and those whose identity is of question (mismatched gender, data inconsistent with reported ethnicity, cryptically related, etc.).

Running sample statistics in SVS will produce sample call rates and heterozygosity rates over the entire genome and over autosomes only. Two output spreadsheets are created containing the statistics; the first spreadsheet, Statistics by Sample, contains call rates and heterozygosity rates for all data, for autosomes only, and for each non-autosomal chromosome (including X and Y). The optional second spreadsheet contains the statistics calculated separately from each autosome. It is also possible to summarize these statistics by phenotypic groups.

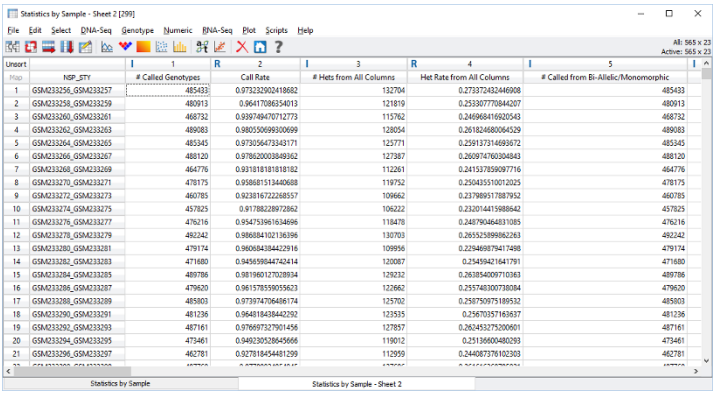


Figure 7: Sample statistics were calculated on the example data.

Sample statistics can tell researchers a great deal about the data, for example the sample call rate, which can be visualized as a histogram to better display the sample distribution. To do this right click on the column and select **Plot Histogram**. In most cases,

samples with low call rates indicate a discrepancy in the quality of DNA. It may also indicate inconsistencies in quantification or other errors that may have occurred during lab handling.

SNPs can also be filtered based on call rate and minor allele frequency within SVS. This allows researchers to simultaneously choose thresholds to filter SNPs failing to meet respective quality control measures and remove them from the set of markers to be used for further analysis.

The call rate filter is used in regards to SNP level quality; for example, if there is a particular SNP that is poorly designed on the chip or has other problems, it will have a low call rate, which indicates that it should be removed from the data. Allele frequency deals with two dimensions, statistical power and rare alleles. Extremely rare alleles do not have good statistical power to compare against a phenotype. Also, older genotyping chips have been known to have quality problems when calling rare alleles because they use clustering algorithms for signal data to determine the different genotype calls.

After performing these filters in SVS, a report spreadsheet will be composed for all of the SNPs containing the criteria used for filtering such as SNP call rate. The report will also contain a column indicating if the SNP should be removed from the analysis based on the specified criteria. It can be helpful to visualize the SNP Call Rates in a histogram (Figure 8). The results of this SNP QC is available in the spreadsheet labeled **Filtering Results** (node 82) in the example project.

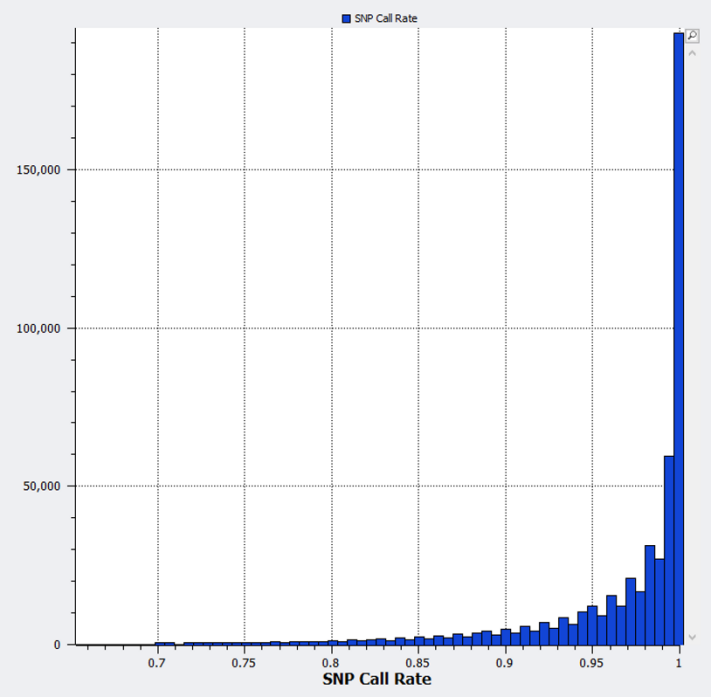


Figure 8: The majority of the SNPs and samples have very high call rates, however, there are a few SNPs with low call rates (below 85%).

Autosomal Heterozygosity Rate

Another useful statistic to focus on is the autosomal heterozygosity rate, which identifies samples with an over- or under-abundance of heterozygous SNPs in the autosomes. Bimodality in the distribution can indicate population stratification of some type. Researchers need to be aware that certain SNPs can be highly polymorphic in one population and less in another. High and low outliers serve as an indication of sample quality. It is most likely that the high outlier samples have been contaminated with DNA from another source, creating a potential quality concern.

Once researchers are satisfied that the call rates are generally good and that there is some stratification amongst the data, it is appropriate to delve deeper into the cause or causes of the data's stratification (Fig. 9).

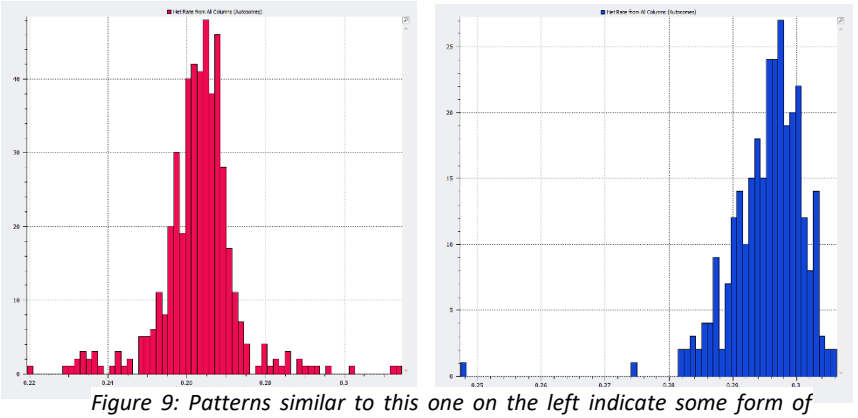


Figure 9: Patterns similar to this one on the left indicate some form of population structure or ethnic stratification. This histogram was generated from the example study data. The histogram on the right is provided for contrast, it is from a single population with two outlier samples.

LD Pruning

Before creating a subset of the filtered SNPs, it is sometimes beneficial to apply a filter based on LD pruning. LD pruning restricts the data to SNPs that have a maximum pairwise LD or SNP-to-SNP correlation of a user-defined threshold. SVS is programed to keep the first SNP and remove the second correlated SNP of the pair.

This filter is applied because a smaller set of SNPs will facilitate certain functions in SVS to run faster as well as exclude unnecessary and redundant information (SNPs) when determining the driving differences in the data. Using universal SNP sets prompts a slightly different output, verses data that has been pruned based on linkage disequilibrium. The difference might be subtle; however, it is important especially when working with large blocks of SNPs that have strong linkage disequilibrium with each other. If ignored, it has the potential to confound calculations in Principal Components Analysis, IBD estimation, and other functions. See **Pruned SNP Subset** (node 46) spreadsheet in the example project.

Population Stratification

The next step is to identify samples that depart from the expected ethnicities. This can be done by preforming a Principal Components Analysis on the data and comparing the first two principal components against reference samples of known ethnicities. Homogeneity of ethnicity is not required for samples as long as the control samples were chosen well to be matched on the ethnicity of case samples. For quantitative traits or cases where a control sample cannot be matched to a case sample by ethnicity the analysis may need to adjust for population structure. There are several ways to perform a Principal Components Analysis. Some recommend using the pruned set of SNPs (done with linkage disequilibrium pruning), some recommend using a filtered set of SNPs (on minor allele frequency and Hardy-Weinberg Equilibrium, for example), and some recommend using the entire SNP set. There are advantages to each and the preferred method depends on the research question.

Two spreadsheets can be created after running PCA and can be found in the example project, the **Principal Components (Additive Model)** spreadsheet (node 62) and the **PC Eigenvalues (Additive Model)** spreadsheet (node 65). To find out how many principal components are required to explain the majority of the population stratification, a Principal Components Analysis plot will be created and the Eigenvectors will be visually inspected. By viewing the specific relationships between samples, as well as within the context of the population, researchers are better able to distinguish sample patterns.

Population stratification can be visualized by plotting the first few components against one another using the **XY Scatter Plot Function**, which with this dataset immediately shows the classic triangular pattern often found in Principal Components Analysis. The clusters become more obvious when each data point is colored according to respective ethnicity (Figure 10).

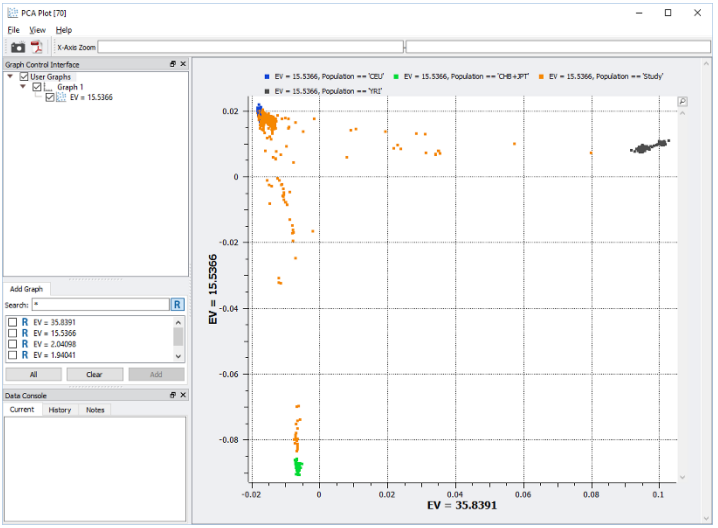


Figure 10: Concentration occurs throughout the left side, primarily with a cluster of green at the bottom indicating East Asia and a clustering of Europeans at the top of the graph in blue. The dark gray reference group to the right of the graph indicates African samples, which are surrounded by scattered orange study samples, indicating that they are likely made up of both European and African-American and Asian-American admixtures.

Clicking on a data point in the SVS plot viewer will provide the information contained in the spreadsheet about a particular subject in the data console. Links in the data console returns researchers to the spreadsheet where individual traits are highlighted.

At this point in a typical GWAS, some decisions need to be made. Should the data be reduced to those who are ethnically cohesive and homogenous? Or should the study continue and some kind of correction factor can be applied? Or is the case/control study well-balanced and proper matching was performed based on criteria such as ethnicity and gender?

Identity by Descent (IBD)

Relatedness is often defined as family-relatedness, but Identity By Descent (IBD) estimations can also detect duplicate samples, duplicate samples from one of a pair of genotyping chips but not the other, or sample contamination. Before performing IBD, standard practice is to first prune the SNPs in LD with one another,

reducing the amount of redundant information used to calculate relatedness. SVS looks at the allele frequency of each SNP to determine the expected rate of sharing, as well as comparing its output to what is actually observed from sample to sample.
The IBD analysis produces a few different output spreadsheets in SVS. Using the Estimated PI spreadsheet to create an N x N table (where N is the number of samples in the dataset) compares individual samples against every other sample. This method provides the proportion of alleles that appear to be from common ancestry. To visually inspect this data, researchers can arrange it into a heat map; unrelated will be white and highly related will be red (Figure 11).

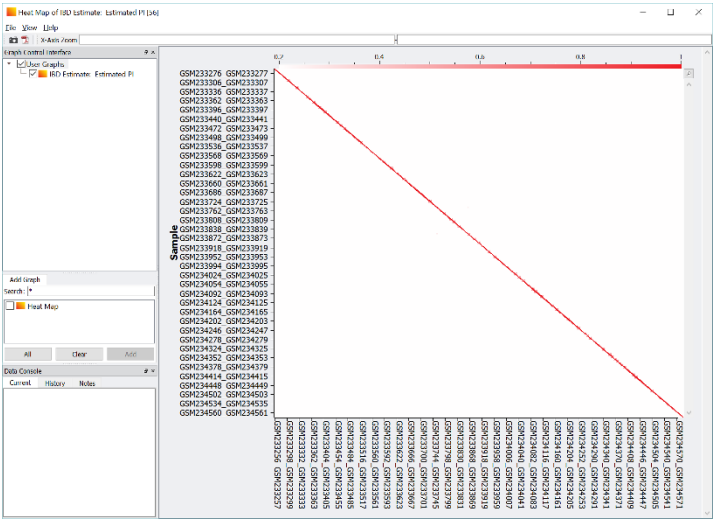


Figure 11: IBD matrix for the example project, some samples show relatedness between one or two other samples but there are no major artifacts or clusters.

To compare the IBD matrix from the example project to another study with more relatedness between samples, populations or sample artifacts examine the IBD matrix in Figure 12. This is another study from the GEO database with more population clusters.

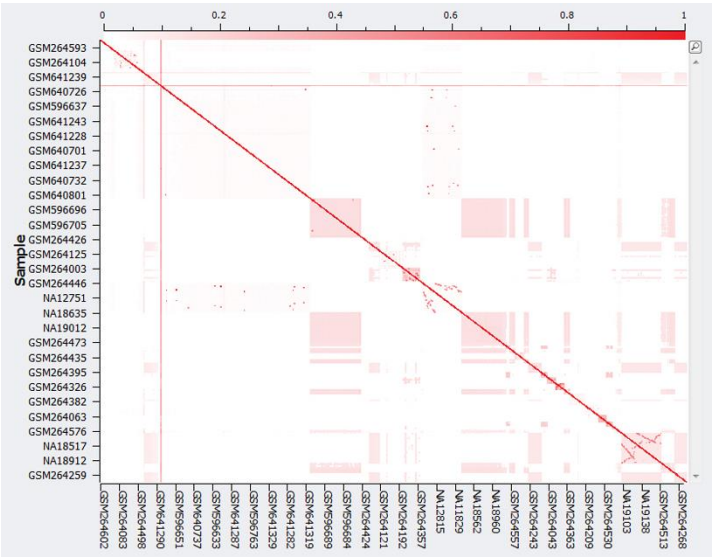


Figure 12: Two particular features really stand out: two heavy streaks of color at the top and a block of color at the bottom indicating samples with background relatedness between the second and third degree.

Whenever streaking appears, showing one sample with excess background relatedness to every other sample, it is most likely a sign of DNA contamination. A large color block of samples appears to indicate a background relatedness within a group of samples. However, IBD estimates are based on observed verses expected rates of allele sharing so those who share alleles based on ancestry or ethnicity tend to appear more related than others, especially if they are a minority within the study data. One of the assumptions of the IBD algorithm is that researchers are working with an ethnically homogenous population, and when they are not, color blocking like this may occur.

Chapter: 4
Imputation

Imputation is an essential capability when conducting genome-wide association studies. It allows researchers to look for evidence of association between genetic markers that have not actually been genotyped. Imputation can increase the power of genome-wide association studies by enabling researchers to combine data from different studies that were conducted using multiple genotyping platforms and methods.

The basic concept of using a relatively small set of genetic variants in individuals to look for potentially useful information in parts of the genome is the fundamental underpinning of GWAS. Early studies used less than 10,000 markers. More recently, technological advances have made it possible to genotype 100,000 – 1,000,000 markers. Genotype imputation can be used in related or unrelated individuals to fill in the data of markers that have not been directly genotyped. In an article by Li et. al (2009), various use cases for genotype imputation of related individuals and unrelated individuals are discussed.

Imputation has enabled researchers to conduct meta-analysis on multiple GWAS studies by combining samples assayed on various genotyping platforms. For example, genotype imputation was used to combine GWAS samples for blood lipid levels (see Kathiresan S et. al 2008 and Willer et. al. 2008) and height (see Sanna S et al 2008). Subsequently, it was used to combine data to study type 2 diabetes (see Zeggini E et al 2008), body-mass index (see Loos RJ 2008) and Crohn's disease (see Barrett JC et al 2008). The benefits of imputation have been discussed for quite some time. The usage of imputation within GWAS is generally well understood.

Researchers have a number of different imputation methods to choose from when conducting this type of analysis. The "usual suspects" are Beagle 4.1, Impute2 and FImpute, among others. Screening the literature, there are several comparative studies and articles that review the output quality of these methods and respectively, lesser known algorithms. The number of articles on this subject is substantial. For the purpose of this e-book, I would like to point out He S (2015) as an example for this type of study.

For a number of reasons, we decided to provide BEAGLE 4.1 as the methods of choice within SVS.

It restricts Hidden Markov model calculations to clusters of markers that are genotyped in the target data. This allows more efficient usage of memory, making computation faster.

Additionally, BEAGLE uses linear interpolation to impute ungenotyped markers. Again, this reduces computational complexity and speeds up the calculation without forgoing too much accuracy.

The method also supports multi-threading. This enables the usage of powerful multi-core servers allowing the parallel computation of some of the more resource intensive calculations.

The method is very memory efficient, which allows it to handle large datasets.

The BEAGLE method leverages Hidden Markov Models to infer haplotypes of individual markers. In the first pass, there is an initialization step imputing missing values based upon allele frequencies. This results in localized haplotype-cluster models, which is essentially a special class of Hidden Markov Models. A forward-backward algorithm is used to estimate the probabilities of each potential haplotype based on the genotype information. Haplotypes are determined based on the conditional probabilities within the model. This is a highly iterative process. The algorithm converges against an optimum selection of the most-likely haplotypes for all individuals. This is the best possible non-mathematical description of the inner workings of BEAGLE. Please find a more detailed description of the method written by the original authors (see Browning and Browning 2016).

Here are a few characteristics of BEAGLE that help to further build an initiation of its wide range of utility in GWAS.

- **Computational Complexity**: Computation time scales linearly with the number of target samples. Moreover, computation increases sublinearly in the number of genotyped markers. This means that researchers who chose a higher resolution in their genotype platform are not unreasonably taxed. The computational complexity is linear in the number of reference samples and linear in the number of reference markers.

- **Parallelization**: This is implemented on the sample level. The input genotype data for the reference panel and target samples are shared across all threads. This has reduced memory consumption as well as input/output overhead.
- **Memory-Efficiency**: The algorithm uses an effective compression of the reference haplotypes and imputed allele probabilities.

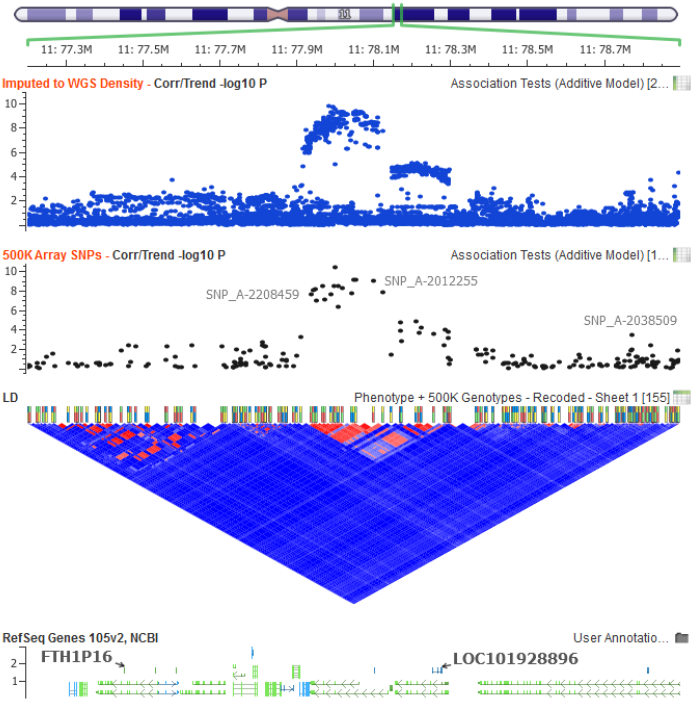


Figure 13: Imputation was used on a 500K SNP array to impute up to the 1000 genome variant density. Around this association site, the localized linkage between the highly-associated SNP and the region can be seen with more clarity with the imputed data. This will help inform any fine-mapping follow up to this GWAS.

Chapter: 5
Genotype Association Testing

In order to run Association Testing, it is important to merge the phenotype data with the genotypes. With just a few clicks in SVS, everything is in one place and ready for the next step. The function that will accomplish this is under **File > Join or Merge Spreadsheets** and or by using the toolbar icon in the upper left corner in the spreadsheet view. In the example project, open **Edited Phenotype + 500k Geno Training Data – Sheet 1** (node 79) to view the joined phenotype and genotype spreadsheet.

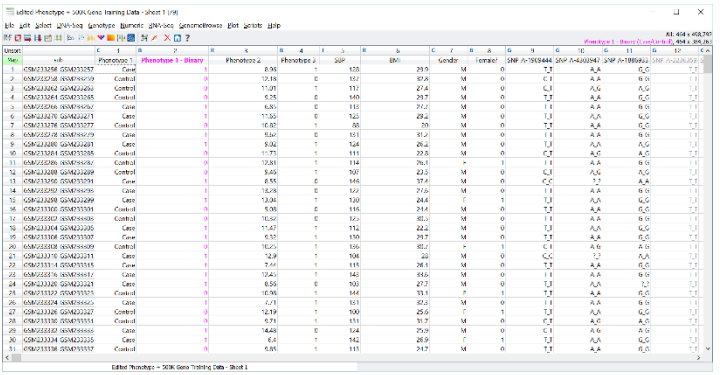


Figure 14: Phenotype 1 - Binary will function as the dependent variable; when the Phenotype 1 - Binary cells are selected, they will turn pink indicating that SVS acknowledges them as its dependent variable.

Before performing genotype association, it is beneficial to re-run genotype filtering with looser thresholds for both SNP call rate and minor allele frequency and including a Hardy-Weinberg Equilibrium filter on controls only. When running this filter previously to prepare the data for population stratification, it is common to have stringent requirements whereas association testing calls for less severe thresholds. The filtered results are found in **Filtered Data for Association Testing** (node 84; also flagged with a green marker).

The Hardy-Weinberg Equilibrium (HWE) was developed as a way to identify biases found on chips. When early GWAS studies began, researchers found thousands of spurious associations due to chip quality, which lead to certain alleles being called more frequently than they should be and ultimately skewing the results. It is applied to controls only at this point in the workflow. When

filtering is complete, a new spreadsheet is produced. For the example project, both a Correlation/Trend test and an Inhibitive Model with an Exact Cochran-Armitage Test were performed, which can be found in under the **Genotypes** Menu in the **Genotype Association Testing** function. The Lambda (inflation factor) should be calculated if there is suspicion that the ethnic stratification is inflating the statistics. Take note that there is a great deal of flexibility in terms of the output from association testing. The results from the Exact Cochran-Armitage Test should be examined in cases where a SNP has a significant p-value but the contingency table of Case Status by Number of Minor Alleles displays at least one count less than five.

The results can be sorted based on p-value (most likely the –log10 p-value), bringing the best results to the top. The map information is carried forward, indicating the location of each p-value. Researchers will now want to graph the p-values in a Manhattan plot in order to see the data visually relative to its genomic context. This is easily accomplished by right clicking on the column containing the p-values and selecting **Plot Variable in Genome Browse**.

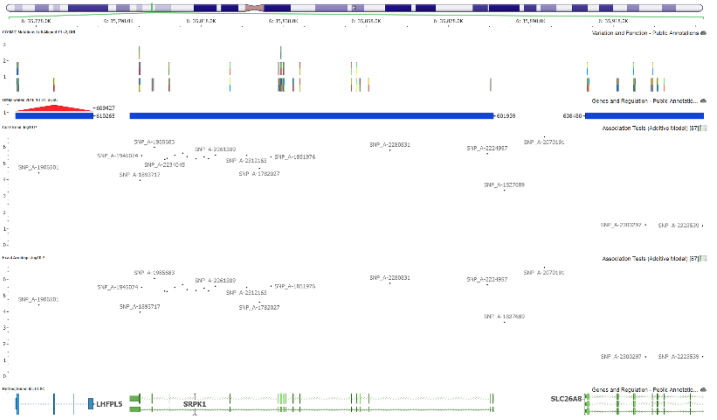


Figure 15: A peak exists in chromosome 6; the gene is SRPK1.

By clicking on an individual SNP in the Genome Browse view, researchers are provided with either a link back to the spreadsheet or links to external databases, which will give more information on the specific SNP. Researchers also have the ability to add annotations within the software. It now needs to be determined whether population structure affects the results. Researchers can get an idea of the severity of the inflation test statistic by plotting the observed versus expected negative log-10 p-values and adding a diagonal line (y = x). This is often called a Quantile-Quantile (Q-Q) or P-P (for p-value) plot. If there is no relationship between the phenotype and the genetic model for all SNPs, the observed p-values would follow a uniform [0, 1] distribution and lie completely on the y = x line. If there is deviation from the expected, the inflation factor is expected be no greater than 1.1. If the observed values fall below the y = x line, this would indicate the data has been over-corrected and any interesting results may be suppressed.

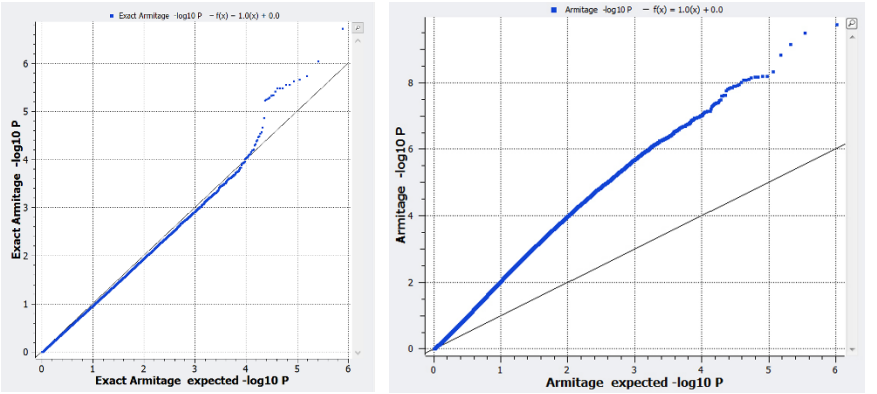


Figure 16: On the left the Q-Q plot from the example project. On the right a Q-Q plot showing severe inflation of the observed versus expected p-values, possibly due to population structure or other batch effects not controlled for in the analysis.

If there is some population structure present, researchers can return to the main spreadsheet and apply a few different Correction Methods. One such method is a Mixed Linear Model Analysis, which can be found in the **Genotype** Menu. This model uses pairwise relatedness between samples as a random effect in a Mixed Model Regression in order to account for pairwise patterns of allele sharing. One mixed model option is the Efficient Mixed-Model Association eXpedited method (EMMAX method)

based on a kinship matrix. Additional phenotypes can be assigned as a fixed effects or covariates.

A Q-Q plot displaying any inflation changes that can also be created from the results of the Mixed Model Analysis. Q-Q plots can be generated by plotting the observed Chi-squared values against the expected Chi-squared values or by plotting the observed versus expected –log10 p-values as above.

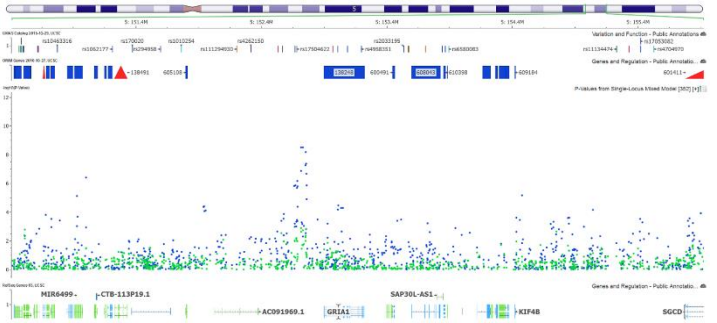


Figure 17: Once the values of the naïve model (green data) and the mixed model analysis (blue data) are added, there is an increased signal from those SNPs that were already significant as well as new significant results.

Chapter: 6
Conducting a Meta-Analysis

Meta-analysis takes the results from existing studies and performs analysis on those results, not directly on the original data. Such an approach is valuable when you want to compare between different published results or compare results between different populations for an alternative to PCA correction or mixed-models analysis.

As an example, take the results from a Correlation-Trend test from three different studies with a simulated case/control phenotype. Two of the studies were from similar population groups and one was from a different population group.

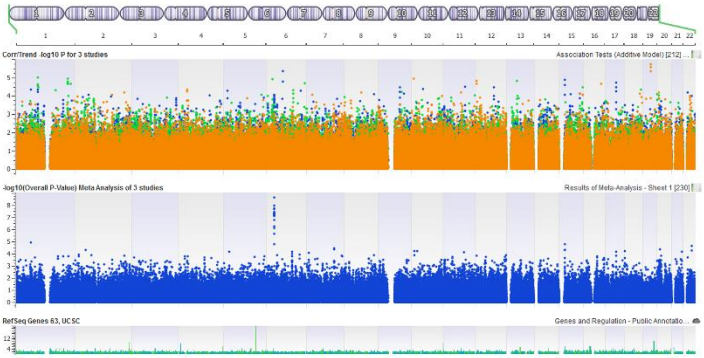


Figure 18: The results of the genome-wide correlation trend test for the three studies are in the top Manhattan plot. The three studies do not show a signal reaching genome-wide significance. A meta-analysis of the three studies, however, demonstrates a significance. A meta-analysis of the three studies, however, demonstrates a significant peak in Chromosome 6.

Examining the significant result further, we can see that this peak is in the SRPK1 gene which seems to be responsible for phosphorylation of SR proteins during the cell cycle in vivo according to OMIM (http://www.omim.org/entry/601939?search=srpk1&highlight=srpk1).

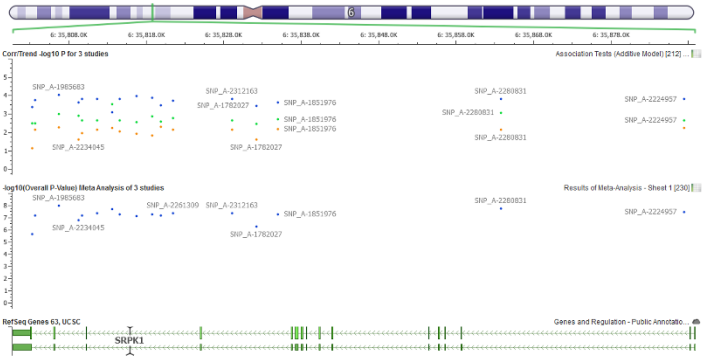


Figure 19: The results of the three studies in the SRPK1 gene are in the top Manhattan plot are not significant but the meta-analysis results in the bottom Manhattan plot show significant results.

Meta-analysis results can also be visualized in a forest-plot to visualize the effect sizes and confidence intervals for the individual studies as well as the meta-analysis summary effect sizes and confidence interval. In a meta-analysis forest plot for effect sizes, if the horizontal lines (confidence intervals) overlap the solid vertical line at 0.0 then the effect sizes cannot be assumed to be different from “no effect” for the individual study. In this example, the confidence intervals from all studies do not overlap with the line at 0.0.

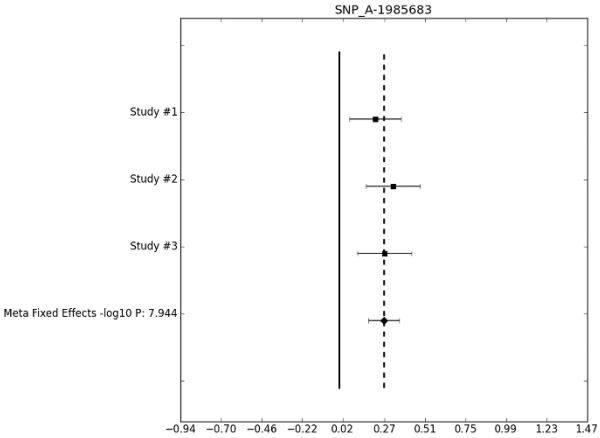


Figure 20: The forest plot for the most significant result in the SRPK1 gene. None of the individual studies overlap the "no effect" line of 0.0. The dotted vertical line represents the overall effect size for the meta-analysis.

Chapter: 7
The Future

GWAS studies had a large and meaningful impact on genetics research, there are many diseases where we have been able to identify risk factors and we have been able to breed cattle that produce more nutritious milk, along with many other success stores. However, technologies have changed to where sequencing an entire genome is becoming less and less expensive and will slowly replace the chip-based technologies. This creates added challenges of data storage, manipulation, quality control and the general bioinformatics infrastructure. Golden Helix with SVS looks forward to meeting these challenges to help our customers discover new and exciting research opportunities and significant results!

End Notes

[i]. International Human Genome Sequencing Consortium. (2004) Finishing the euchromatic sequence of the human genome. *Nature*. **431**, 931-45.

[ii]. The International HapMap Consortium. (2005) A haplotype map of the human genome. *Nature*, **437**, 1299-1320.

[iii]. Welter D, MacArthur J, Morales J, Burdett T, Hall P, Junkins H, Klemm A, Flicek P, Manolio T, Hindorff L, and Parkinson H. (2014) The NHGRI GWAS Catalog, a curated resource of SNP-trait associations. *Nucleic Acids Res*., **42**, D1001-6.

[iv]. Klein RJ, Zeiss C, Chew EY, Tsai JY, Sackler RS, Haynes C, Henning AK, SanGiovanni JP, Mane SM, Mayne ST, Bracken MB, Ferris FL, Ott J, Barnstable C, Hoh J. (2005) Complement factor H polymorphism in age-related macular degeneration. *Science*, **308**, 385-9.

[v]. Teslovich T, Musunuru K, Smith A, Edmondson A, Stylianou I, et al. (2010) Biological, clinical and population relevance of 95 loci for blood lipids. *Nature*, **466**, 707–13.

[vi]. Habek M, Brinar V, Borovecki F. (2010) Genes associated with multiple sclerosis: 15 and counting. *Expert. Rev. Mol. Diagn*., **10**, 857–61.

[vii]. Bush W, Moore J. (2012) Chapter 11: Genome-Wide association studies. *PLOS Comput. Biol.,* **8** (12), e1002822.

[viii]. Laurie CC, Doheny KF, Mirel DB, Pugh EW, Bierut LJ, Bhangale T, Boehm F, Caporaso NE, Cornelis MC, Edenberg HJ, Gabriel SB, Harris EL, Hu FB, Jacobs KB, Kraft P, Landi MT, Lumley T, Manolio TA, McHugh C, Painter I, Paschall J, Rice JP, Rice KM, Zheng X, Weir BS; GENEVA Investigators. (2010) Quality control and quality assurance in genotypic data for genome-wise association studies. *Genet. Epidemiol.*, **34**, 591-602.

[ix]. Clarke GM, Anderson CA, Petterson FH, Cardon LR, Morris AP, Zondervan KT. (2011) Basic statistical analysis in genetic case-control studies. *Nat. Protoc.,* **6**, 121-133.

[x]. Eu-ahsunthonwattana J, Miller EN, Fakiola M, Wellcome Trust Case Control Consortium 2, Jeronimo SMB, Blackwell JM, Cordell HJ. (2014) Comparison of methods to account for relatedness in genome-wide association studies with family-based data. *PLOS Genet,*. **10**, e1004445.
[xi]. Pe'er I, Yelensky R, Altshuler D, Daly MJ. (2008) Estimation of the multiple testing burden for genomewide association studies of nearly all common variants. *Genet. Epidemiol.,* **32**, 381-5.

[xii]. Lange C, DeMeo D, Silverman EK, Weiss ST, Laird NM. (2004) PBAT: Tools for family-based association studies. *Am. J. Hum. Genet.,* **74**:367-9.

[xiii]. Golden Helix Inc., Bozeman, Montana, USA.

[xiv]. Sham PC, Purcell SM. (2014) Statistical power and significance testing in large-scale genetic studies. *Nat. Rev. Genet.,* **15**, 335-46.
[xv]. Vilhjálmsson BJ, Nordborg M. (2013) The nature of confounding in genome-wide association studies. *Nat. Rev. Genet.,* **14**, 1-2.

[xvi]. Kang HM, Sul JH, Service SK, Zaitlen NA, Kong S, Freimer NB, Sabatti C, Eskin E. (2010) Variance component model to account for sample structure in genome-wide association studies. *Nat. Genet.,* **42**, 348-354.

[xvii]. Zhou X, Stephens M. (2012) Genome-wide efficient mixed-model analysis for association studies. *Nat. Genet.,* **44**, 821-24.
[xviii]. Stranger BE, Stahl EA, Raj T. (2011) Progress and promise of genome-wide association studies for human complex trait genetics. *Genetics*, **187**, 367-83.

[xix]. Manolio T, Collins FS, Cox NJ, Goldstein DB, Hindorff LA, Hunter DJ, McCarthy MI, Ramos EM, Cardon LR, Chakravarti A, Cho JH, Guttmacher AE, Kong A, Kruglyak L, Mardis E, Rotimi CN, Slatkin M, Valle D, Whittemore AS, Boehnke M, Clark AG, Eichler EE, Gibson G, Haines JL, Mackay TF, McCarroll SA, Visscher PM. (2009) Finding the missing heritability of complex diseases. *Nature*, **461**, 747-53.

[xx]. Marjoram P, Zubair A, Nuzhdin SV. (2014) Post-GWAS: where next? More Samples, more SNPs or more biology? *Heredity*, **112**, 79-88.

Bibliography

Alizadeh A.A. et al. (2000). "Distinct types of diffuse large B-cell lymphoma identified by gene expression profiling". Nature 403 (6769): 503–511. doi: 10.1038/35000501. PMID 10676951.

American Congress (formerly College) of Obstetricians and Gynecologists Committee on Genetics. “ACOG Committee Opinion No. 486: Update on carrier screening for cystic fibrosis”. Obstet Gynecol. 2011 Apr; 117(4):1028-31.

American College of Medical Genetics. Technical standards and guidelines for CFTR mutation testing. 2006 ed. http://www.acmg.net/Pages/ACMG_Activities/stds-2002/cf.htm. Updated April 26, 2006. Accessed November 21, 2013.

Arrowsmith, J. (2012). “A decade of change.” Nature Reviews Drug Discovery, 11, 17-18.

Bainbridge MN, Wiszniewski W, Murdock DR, Friedman J, Gonzaga-Jauregui C, Newsham I, Reid JG, Fink JK, Morgan MB, Gingras MC, Muzny DM, Hoang LD, Yousaf S, Lupski JR, Gibbs RA. (2011) “Whole-genome sequencing for optimized patient management.” Sci. Transl. Med 3(87).

Balasubramanian et al. (2013) “Maraviroc CCR5 antagonosit for HIV,” Drug discovery, 3(9):39-40 [www.discovery.org.in http://www.discovery.org.in/dd.htm].

Barrett JC, Hansoul S, Nicolae DL, Cho JH, Duerr RH, et al. (2008) Genome-wide association defines more than 30 distinct susceptibility loci for Crohn's disease. Nat Genet. 2008;40:955–62.

Browning BL, Browning SR (2016) “Genotype Imputation with Millions of Reference Samples, Am J Hum Genet, Jan 7 2017, 98(1): 116-126.

Bush W, Moore J. (2012) Chapter 11: Genome-Wide association studies. *PLOS Comput. Biol.,* **8** (12), e1002822.

Clarke GM, Anderson CA, Petterson FH, Cardon LR, Morris AP, Zondervan KT. (2011) Basic statistical analysis in genetic case-control studies. *Nat. Protoc.,* **6**, 121-133.

Cystic Fibrosis Mutation Database: Statistics". Cystic Fibrosis Centre at the Hospital for Sick Children in Toronto. Retrieved 28 July 2014.

Department for Professional Employees (2014) http://dpeaflcio.org/professionals/professionals-in-the-workplace/healthcare-professionals-and-technicians/
Easton DF.

How many more breast cancer predisposition genes are there? Breast Cancer Research 1999; 1(1):14–17.

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

Farwell KD, Shahmirzadi L, El-Khechen,D, Powis Z, Chao EC, Davis BT, Baxter RM, Zeng W, Mroske C, Parra MC, Gandomi SK, Lu I, Li X, Lu H, Lu H-M, Salvador D, Ruble D, Lao M, Fischbach S, Wen J, Lee S, Elliott A, Dunlop CLM, Tang S. (2014) "Enhanced utility of family-centered diagnostic exome sequencing with inheritance model-based analysis: results from 500 unselected families with undiagnosed genetic conditions." Genetics in Medicine 17(7):578-86.

FDA. (2013) "Paving the Way for Personalized Medicine: FDA's Role in a New Era of Medical Product Development."
FDA, (2011)."FDA approves Xalkori with companion diagnostic for a type of late-stage lung cancer". U.S. Food and Drug Administration.

Habek M, Brinar V, Borovecki F. (2010) Genes associated with multiple sclerosis: 15 and counting. *Expert. Rev. Mol. Diagn*., **10**, 857–61.

He S, Zhao Y., Mette MF, Bothe R. Ebmeyer E., Sharbel TF, Reif, JC and Jiang Y. (2015) “Prospects and limits of marker imputation in quantitative genetic studies in European elite wheat (Triticum aestivum L.)” BMC Genomics 2015, 16: 168

Ingelman-Sundberg M. (2005) “Genetic polymorphisms of cytochrome P450 2D6 (CYP2D6): clinical consequences, evolutionary aspects, and functional diversity”. Pharmacogenomics J. 5(1):6-13.

The International HapMap Consortium. (2005) A haplotype map of the human genome. *Nature*, **437**, 1299-1320.

International Human Genome Sequencing Consortium. (2004) Finishing the euchromatic sequence of the human genome. *Nature*. **431**, 931-45.

Ionita-Laza, I, Seunggeun Lee S, Makarov V, Buxbaum JD, Lin X. (2013)“ Kernel Association Tests for the Combined Effect of Rare and Common Variants.” Am J Hum Genet 92(6): 841–853.

Kang HM, Sul JH, Service SK, Zaitlen NA, Kong S, Freimer NB, Sabatti C, Eskin E. (2010) Variance component model to account for sample structure in genome-wide association studies. *Nat. Genet.,* **42**, 348-354.

Kelly, Hillner and Smith, (2014) “Cost Effectiveness of Crizotinib for Anaplastic Lymphoma Kinase-Postive, Non-Small-Cell Lung Cancer: Who is Going to Blink at the Costs?” Journal of Clinical Oncology 32(10):983-985.

Klein RJ, Zeiss C, Chew EY, Tsai JY et al. (2005) "Complement Factor H Polymorphism in Age-Related Macular Degeneration." Science 308 (5720): 385–9. doi:10.1126/science.1109557. PMC 1512523. PMID 15761122.

Lange C, DeMeo D, Silverman EK, Weiss ST, Laird NM. (2004) PBAT: Tools for family-based association studies. *Am. J. Hum. Genet.,* **74**:367-9.

Langfelder-Schwind E, et al. Molecular testing for cystic fibrosis carrier status practice guidelines: recommendations of the

National Society of Genetic Counselors. J Genet Couns. 2014 Feb;23(1):5-15.

Laurie CC, Doheny KF, Mirel DB, Pugh EW, Bierut LJ, Bhangale T, Boehm F, Caporaso NE, Cornelis MC, Edenberg HJ, Gabriel SB, Harris EL, Hu FB, Jacobs KB, Kraft P, Landi MT, Lumley T, Manolio TA, McHugh C, Painter I, Paschall J, Rice JP, Rice KM, Zheng X, Weir BS; GENEVA Investigators. (2010) Quality control and quality assurance in genotypic data for genome-wise association studies. *Genet. Epidemiol.*, **34**, 591-602.

Lee S, Emond MJ, Bamshad MJ, Barnes KC, Rieder MJ, Nickerson DA, NHLBI GO Exome Sequencing Project—ESP Lung Project Team, Christiani DC, Wurfel MM, LinX. (2012) "Optimal Unified Approach for Rare-Variant Association Testing with Application to Small-Sample Case-Control Whole-Exome Sequencing Studies." Am J Hum Genet 91(2): 224–237.

Li Y, Willer C, Sanna S and Abecasis G. (2009) "Genotype Imputation, Annu Rev Genomics Hum Genetl 10: 387-406.

Liu D and Leal SM (2010) "Novel adaptive method for the analysis of next-generation sequencing data to detect complex trait associations with rare variants due to gene main effects and interactions." PLoS Genetics 6(10): e1001156. doi:10.1371/journal.pgen.1001156.

Loos RJ, Lindgren CM, Li S, Wheeler E, Zhao JH, et al. (2008) "Common variants near MC4R are associated with fat mass, weight and risk of obesity" Nat Genet. 2008;40:768–75.

National Human Genome Research Institute. Human Genome Sequence Quality Standards. Oct 2002 http://www.genome.gov/10000923.

Marjoram P, Zubair A, Nuzhdin SV. (2014) Post-GWAS: where next? More Samples, more SNPs or more biology? *Heredity*, **112**, 79-88.

Manolio T, Collins FS, Cox NJ, Goldstein DB, Hindorff LA, Hunter DJ, McCarthy MI, Ramos EM, Cardon LR, Chakravarti A, Cho JH, Guttmacher AE, Kong A, Kruglyak L, Mardis E, Rotimi CN, Slatkin

M, Valle D, Whittemore AS, Boehnke M, Clark AG, Eichler EE, Gibson G, Haines JL, Mackay TF, McCarroll SA, Visscher PM. (2009) Finding the missing heritability of complex diseases. *Nature*, **461**, 747-53.

Manolio TA; Guttmacher, Alan E.; Manolio, Teri A. (2010). "Genomewide association studies and assessment of the risk of disease" N. Engl. J. Med. 363 (2): 166–76. doi:10.1056/NEJMra0905980. PMID 20647212.

Mayer AN, Dimmock DP, Arca MJ, Bick DP, Verbsky JW, Worthey EA, Jacob HJ, Margolis DA (2010) "A timely arrival for genomic medicine." Genetics in Medicine DOI: 10.1097/GIM.0b013e3182095089.

Mcclain, MR, Palomaki, GE, Piper M, Haddow JE. (2008) "A Rapid-ACCE review of CYP2C9 and VKORC1alleles testing to inform warfarin dosing in adults at elevated risk for thrombotic events to avoid serious bleeding." Genetics in Medicine 10(2):89-98.

Kathiresan S, Melander O, Guiducci C, Surti A, Burtt NP, et al (2008) "Six new loci associated with blood low-density lipoprotein cholesterol, high-density lipoprotein cholesterol or triglycerides in humans". Nat Genet. 2008;40:189–97.

Middleton, A, Morley KI, Bragin E, Firth HV, Hurles, ME, Wright CF, Parker M. (2015) "Attitudes of nearly 7000 health professionals, genomic researchers and publics toward the return of incidental results from sequencing research." European Journal of Human Genetics. doi:10.1038/ejhg.2015.58.

Need AC, Shashi V, Hitomi Y, Schoch K, Shianna K, McDonald MT, Meisler M, Goldstein D. (2012) "Clinical application of exome sequencing in undiagnosed genetic conditions." J Med Genet doi: 10.1136/jmedgenet-2012-10081.

Pe'er I, Yelensky R, Altshuler D, Daly MJ. (2008) Estimation of the multiple testing burden for genomewide association studies of nearly all common variants. *Genet. Epidemiol.,* **32**, 381-5.

Rommens JM, Iannuzzi MC, Kerem B, Drumm ML, Melmer G, Dean M, Rozmahel R, Cole JL, Kennedy D, Hidaka N (September 1989).

"Identification of the cystic fibrosis gene: chromosome walking and jumping". Science 245 (4922): 1059–65.

Rope AF, Wang K, Evjenth R, Xing J, Johnston JJ, Swenson J, Johnson WE, Moore B, Huff CD, Bird LM, Carey JC, Opitz JM, Stevens CA, Jiang T, Schank C, Fain HD, Robison R, Dalley B, Chin S, South S, Pysher TJ, Jorde LB, Hakonarson H, Lillehaug JR, Biesecker LG, Yandell M, Arnesen T, Lyon GJ. (2011) "Using VAAST to identify an x-linked disorder resulting in lethality in male infants due to N-terminal acetyltransferase deficiency." AJHG 89(1):28-43.

Ross D.T. et al. (2000). "Systematic variation in gene expression patterns in human cancer cell lines". Nature Genetics 24 (3): 227–235. doi: 10.1038/73432. PMID 10700174.

Sanna S, Jackson AU, Nagaraja R, Willer CJ, Chen WM, et al. (2008) Common variants in the GDF5 region are associated with variation in human height. Nature Genetics. 2008;40:198–203.

Saunders CJ, Miller NA, Soden SE, Dinwiddie DL, Noll A, et al. (2012) Rapid whole-genome sequencing for genetic disease diagnosis in neonatal intensive care units. Sci Transl Med. 4: 154ra135.

Servant N, Romacjon J, Gestraud P, La Rosa P, Lucotte G, Lair S, Bernard V, Zeitouni B, Coffin F, Jules-CIAcment G, Yvon F, LErmine A, Poullet P, Liva S, Pook S, Popova T, Barette C, Prud'homme F, Dick J-G, Kamal M, Le Tourneau C, Barillot E, HupAc P. (2014) "Bioinformatics for precision medicine in oncology: principles and application to the SHIVA clinical trial." Frontiers in Genetics, 5(00152).

Sham PC, Purcell SM. (2014) Statistical power and significance testing in large-scale genetic studies. *Nat. Rev. Genet.,* **15**, 335-46.

Spear, Heath-Chiozzi, Huff (2001) "Clinical application of pharmacogenetics." Trends in Molecular Medicine, 7(5):201-204.
Teh LK, Bertilsson L.(2012) "Pharmacogenomics of CYP2D6: molecular genetics, interethnic differences and clinical importance". Drug Metab. Pharmacokinet. 27(1):55–67. doi:10.2133/dmpk.DMPK-11-RV-121. PMID 22185816.

Stranger BE, Stahl EA, Raj T. (2011) Progress and promise of genome-wide association studies for human complex trait genetics. *Genetics*, **187**, 367-83.

Teslovich T, Musunuru K, Smith A, Edmondson A, Stylianou I, et al. (2010) Biological, clinical and population relevance of 95 loci for blood lipids. *Nature*, **466**, 707–13.

Vilhjálmsson BJ, Nordborg M. (2013) The nature of confounding in genome-wide association studies. *Nat. Rev. Genet.,* **14**, 1-2.

Wall JD, Tang LF, Zerbe B, Kvale MN, Kwok P-Y, Schaefer C, Risch N. (2014) "Estimating genotype error rates from high-coverage next-generation sequence data." Genome Research DOI: 10.1101/GR.168393.113.

Wang B, Yang LP, Zhang XZ, Huang SQ, Bartlam M, Zhou SF.(2009) "New insights into the structural characteristics and functional relevance of the human cytochrome P450 2D6 enzyme". Drug Metab. Rev. 41(4):573–643. doi: 10.1080/03602530903118729. PMID 19645588.

Welter D, MacArthur J, Morales J, Burdett T, Hall P, Junkins H, Klemm A, Flicek P, Manolio T, Hindorff L, and Parkinson H. (2014) The NHGRI GWAS Catalog, a curated resource of SNP-trait associations. *Nucleic Acids Res*., **42**, D1001-6.

Wilson SK. (2008) "Somatic mutations affect key pathways in lung adenocarcinoma." Nature 455(7216):1069-1075.

Willer CJ, Sanna S, Jackson AU, Scuteri A, Bonnycastle LL, et al. (2008), "Genome-Wide Association Scans Identify Novel Loci That Influence Lipid Levels and Risk of Coronary Artery Disease". Nature Genetics. 2008;40:161–9.

Zeggini E, Scott LJ, Saxena R, Voight BF, Marchini JL, et al. (2008) "Meta-analysis of genome-wide association data and large-scale replication identifies additional susceptibility loci for type 2 diabetes". Nat Genet. 2008;40:638–45.

Zhou X, Stephens M. (2012) Genome-wide efficient mixed-model analysis for association studies. *Nat. Genet.,* **44**, 821-24.

About the Author

Dr. Andreas Scherer is CEO of Golden Helix. The company has been delivering industry leading bioinformatics solutions for the advancement of life science research and translational medicine for over a decade. Its innovative technologies and analytic services empower scientists and healthcare professionals at all levels to derive meaning from the rapidly increasing volumes of genomic data produced from microarrays and next-generation sequencing. With its solutions, hundreds of the world's top pharmaceutical, biotech, and academic research organizations are able to harness the full potential of genomics to identify the cause of disease, improve the efficacy and safety of drugs, develop genomic diagnostics, and advance the quest for personalized medicine. Golden Helix products and services have been cited in over 1100 peer-reviewed publications.

He is also Managing Partner of Salto Partners, a management consulting firm headquartered in the DC metro area. He has extensive experience successfully managing growth as well as orchestrating complex turnaround situations. His company, Salto Partners, advises on business strategy, financing, sales and operations. Clients are operating in the high tech and life sciences space.

Dr. Scherer holds a PhD in computer science from the University of Hagen, Germany, and a Master of Computer Science from the University of Dortmund, Germany. He is author and co-author of over 20 international publications and has written books on project management, the Internet and artificial intelligence. His latest book, "Be Fast Or Be Gone", is a prizewinner in the 2012 Eric Hoffer Book Awards competition, and has been named a finalist in the 2012 Next Generation Indie Book Awards!

Connect with Dr. Scherer: