Topics in Statistical Genetics

INSIGHT Bioinformatics Webinar 2

August 22nd 2018

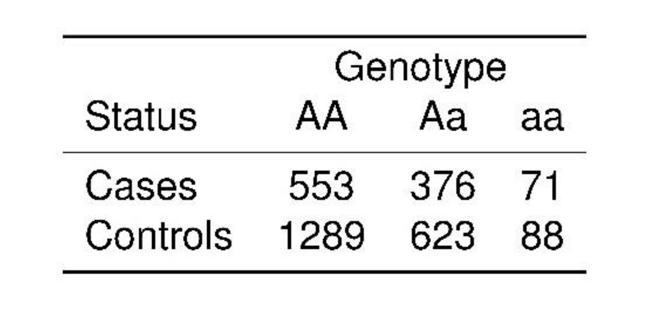
Presented by
Cavan Reilly,
Ph.D. & Brad
Sherman, M.S.

Recap of webinar 1 concepts

- DNA is used to make proteins and proteins are necessary for the creation and maintenance of life
- There is variation in DNA sequences among humans: we frequently look at a particular type of variant called a SNP
- We can use statistical models to relate data on SNPs to traits of interest
- It is possible to get data on up to a million SNPs for thousands of people-we now have this for some INSIGHT trial participants
- There are standard quality control procedures used to set up analysis data sets

Genetic Associations (codominant test)

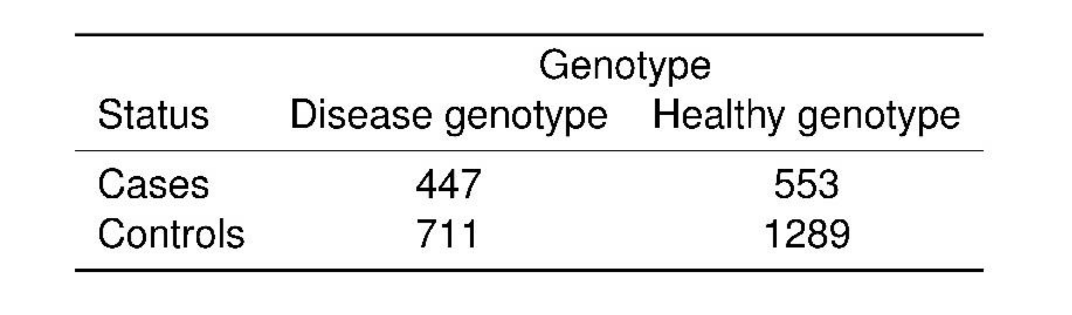
- Frequencies of alleles at a SNP can be compared between cases and controls to determine if there is a statistically significant association between the SNP and case/control status
- Consider the following table:
  - We could conduct Pearson’s chi-square test



	Genotype		
Status	AA	Aa	aa
Cases	553	376	71
Controls	1289	623	88

Genetic Associations (dominant test)

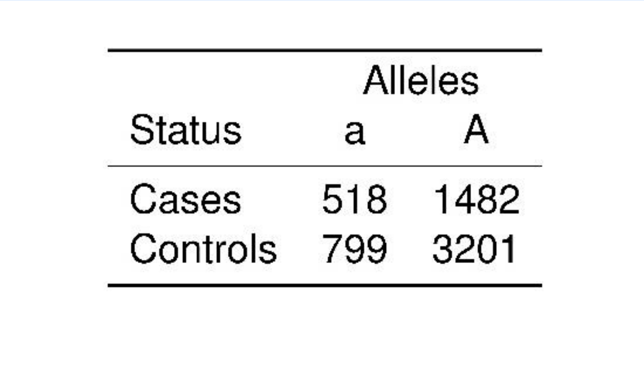
- Alternatively, we could assume that the disease follows a dominant mode of transmission, so that the table becomes as follows if a is the disease allele (we can still use Pearson’s chi-square test)



	Genotype	
Status	Disease genotype	Healthy genotype
Cases	447	553
Controls	711	1289

Genetic Associations (alleles test)

- As another alternative, we can examine the frequency of each allele among cases and controls (and again, use Pearson’s chi-square test if Hardy-Weinberg equilibrium holds)



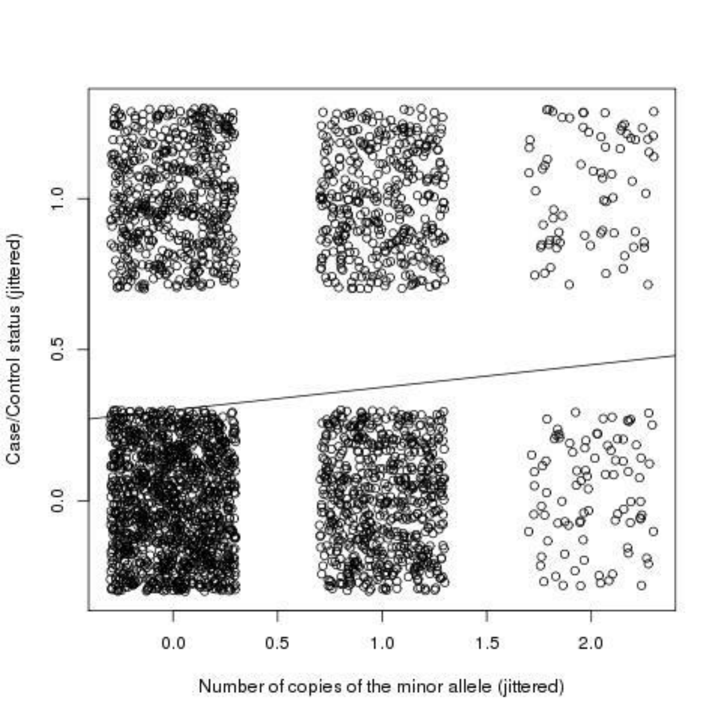
	Alleles	
Status	a	A
Cases	518	1482
Controls	799	3201

Genetic Associations (trend test)

- As yet another alternative, we might assume an *additive* genetic model in which the probability that someone is a case depends linearly on the number of copies of one of the alleles
  - Which allele we select with bi-allelic variants doesn’t matter for computing a p-value: typically use the number of copies of the minor allele
  - We can use simple linear regression to conduct this test (case/control status is the outcome and the number of alleles is the predictor)
- All analyses of the INSIGHT genotypic data that have been conducted thus far have used this additive model

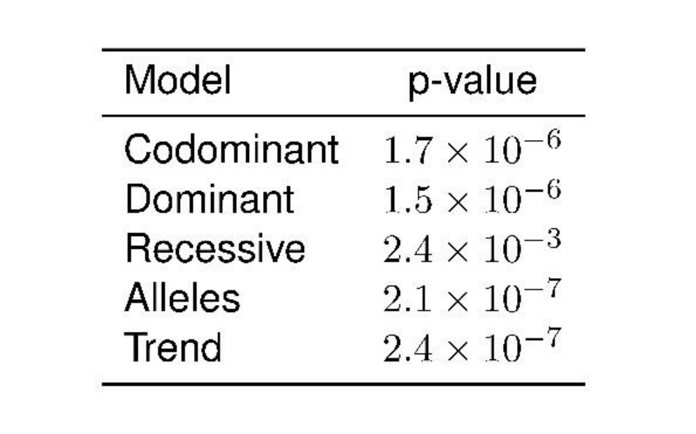
Genetic Associations (trend test)

- Jittered scatterplot and regression line



Comparison of tests

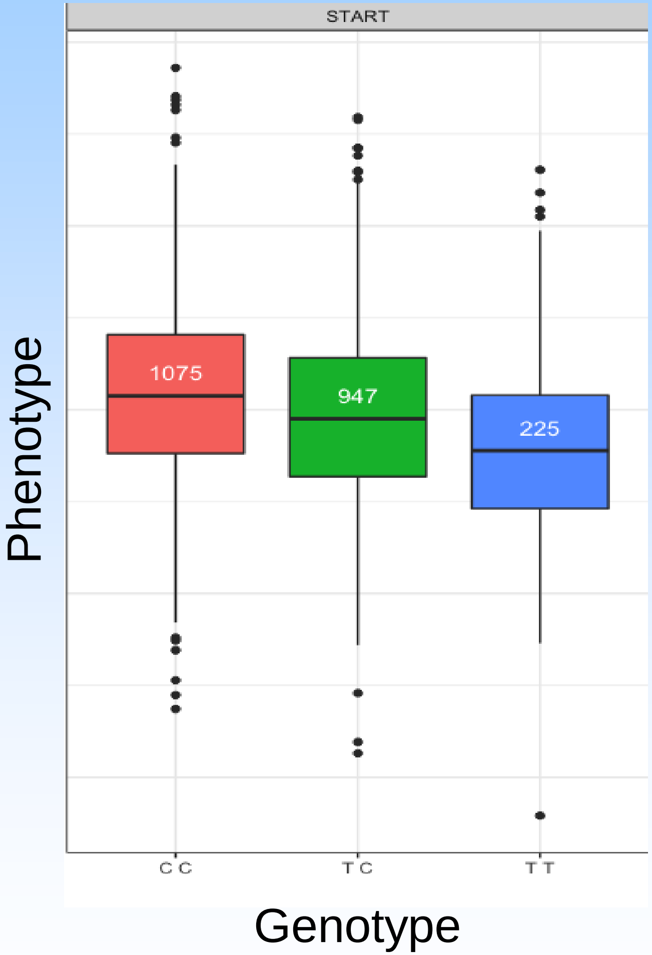
- The p-values one obtains from these various tests can be quite variable:



Model	p-value
Codominant	1.7×10^{-6}
Dominant	1.5×10^{-6}
Recessive	2.4×10^{-3}
Alleles	2.1×10^{-7}
Trend	2.4×10^{-7}

Genetic Associations

- Quantitative variables (e.g. viral load) can be accommodated using the trend test with the quantitative variable as the explanatory variable
    - Quantitative trait locus (QTL)
- With either type of response variable, we can make adjustments for confounders in this regression context by including them as further explanatory variables
- However, when the outcome is dichotomous (e.g. case/control status) one would typically use logistic regression



Multiple hypothesis testing

- Some study designs in contemporary statistical genetics (e.g., GWAS) entail large numbers of hypothesis tests
- Such approaches are subject to false positives unless action is taken to avoid this
- The *family-wise* error rate is the probability of making one or more false positives when conducting more than 1 test
- The Bonferroni correction is widely used despite there being uniformly more powerful approaches to the control of the family-wise error rate: Holm’s method

Multiple hypothesis testing

- In the statistical genetics literature, the use of 5.0×10^{-8} as a cut-off for p-values to be deemed significant is taken for granted
  - If a test statistic has a p-value this extreme, we speak of *genome-wide significance*
- One can interpret this as a Bonferroni correction for a million tests using the usual cut-off for statistical significance
- The Bonferroni correction assumes that the tests are independent: if there is dependence among the tests it is too strict

Multiple hypothesis testing

- Due to linkage disequilibrium we expect tests of association between markers which are close on the genome and a phenotype to be dependent
  - It is not uncommon to find markers which are in complete linkage equilibrium, i.e. there is a perfect association between them
- Some argue that due to linkage disequilibrium there are only about 1 million independent tests possible in a genome, so the use of genome-wide significance is justifiable regardless of the number of markers

Multiple hypothesis testing: new paradigm

- The use of large-scale hypothesis testing became more common in the 1990's-this led to new approaches to corrections for multiple hypothesis testing
- The new paradigm that emerged focused on determining the *false discovery rate* (FDR)
- FDR control is motivated as follows: I conduct tests of many null hypotheses and use some criterion to say a certain number, say X, of the nulls are rejected
  - So I have made X discoveries!

False discovery rates

- Using one of many methods I can estimate the proportion of my discoveries that are likely false (hence the name)
- By choosing a value for the FDR one can control the FDR at that value
- The seminal paper on this topic was authored by Benjamini and Hochberg, and the most commonly used technique bears their name
  - Their technique assumes that the tests are independent, however there are extensions to dependent tests
- The Benjamini Hochberg technique is conservative in that the actual FDR is typically lower than the value at which it is controlled

False discovery rates

- Controlling the FDR is not equivalent to controlling the family-wise error rate
- In fact one can mathematically demonstrate that if one controls the FDR at level α then the family-wise error rate is at least α
  - So if you control the FDR at 5% then the family-wise error rate is 5% or larger
  - So controlling the FDR is less conservative than standard statistical practice
- In practice it is common to see authors control the FDR at 10%, and there are publications in reputable journals where it is controlled at even higher levels (like 30%)

False discovery rates

- The usual justification for this is that one is conducting an exploratory analysis
  - So there is greater concern for type II errors than type I errors
- As such analyses are exploratory one typically follows up with other analyses to investigate the set of SNPs that have been detected to be associated with the outcome

Genetic Associations: INSIGHT

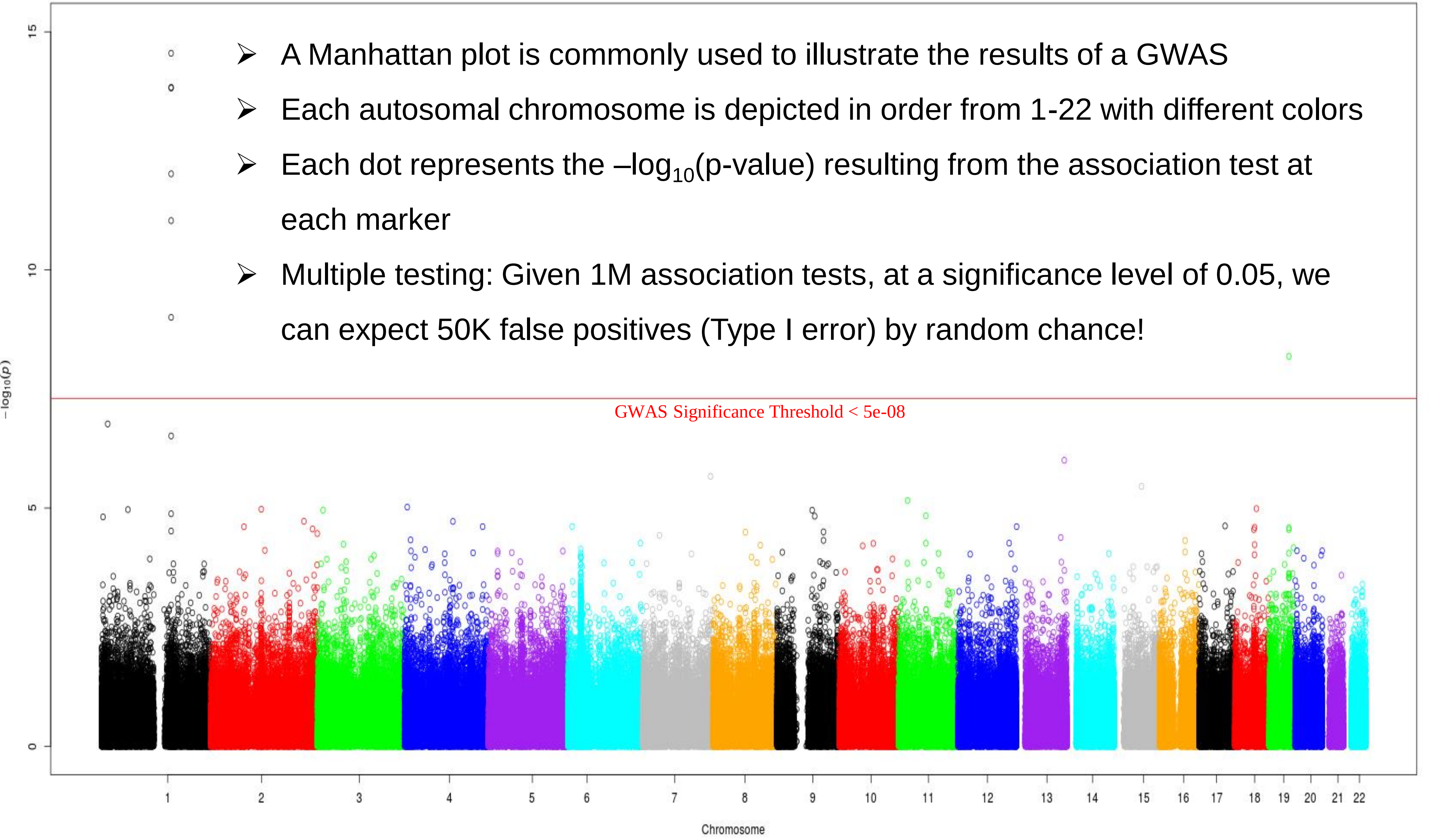
- We have looked at several quantitative traits using a GWAS in this manner thus far

  1. Viral load
  2. D-dimer
  3. hsCRP
  4. IL-6

- These analyses included principle components, gender and sometimes age as additional covariates
- These analyses were also pursued at the individual study level and by combining data across studies

Manhattan plot

- A Manhattan plot is commonly used to illustrate the results of a GWAS
- Each autosomal chromosome is depicted in order from 1-22 with different colors
- Each dot represents the $-\log_{10}$(p-value) resulting from the association test at each marker
- Multiple testing: Given 1M association tests, at a significance level of 0.05, we can expect 50K false positives (Type I error) by random chance!



*GWAS results using START data

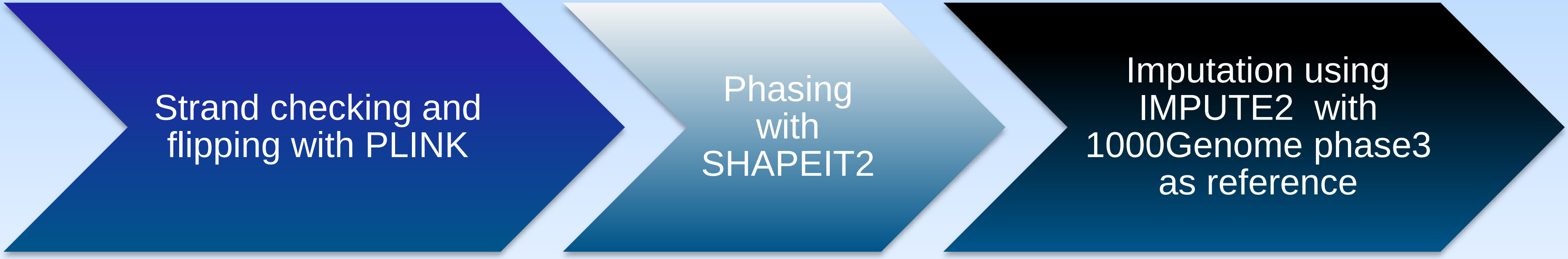
Imputation

- A common approach to the analysis of GWAS data is to impute data for SNPs not originally genotyped
- To conduct this imputation, one needs data from individuals that overlap the set of SNPs that one has genotyped and has additional genetic variants
  - Such datasets are common and publicly available
- The idea: in a small window of the genome compare data that needs imputation to other genomes, and if some are similar to the input genome, use that for reference
  - Then apply to windows that cover the genome

Imputation

- These techniques use hidden Markov models, which makes this fast
- Imputed data will not necessarily be an integer giving the number of minor alleles
- Rather, it will be an estimate of the number of copies of the minor allele for that sample
- Note: we could filter out a SNP for QC then impute data for that SNP

Imputation method



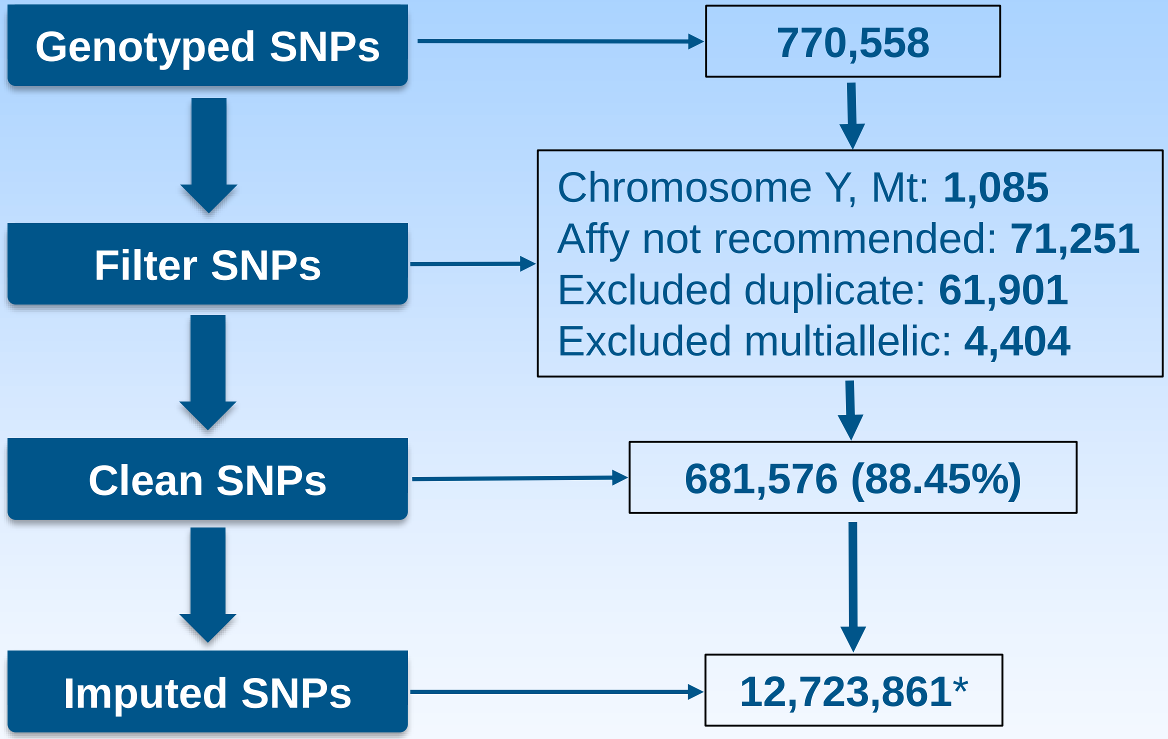
Strand checking and flipping with PLINK

Phasing with SHAPEIT2

Imputation using IMPUTE2 with 1000Genome phase3 as reference

Note: This analysis pipeline is implemented in the genipe tool. (Lemieux Perreault et al. 2016. Bioinformatics)

SNP QC and Imputation Summary



Genotyped SNPs → 770,558

Filter SNPs →

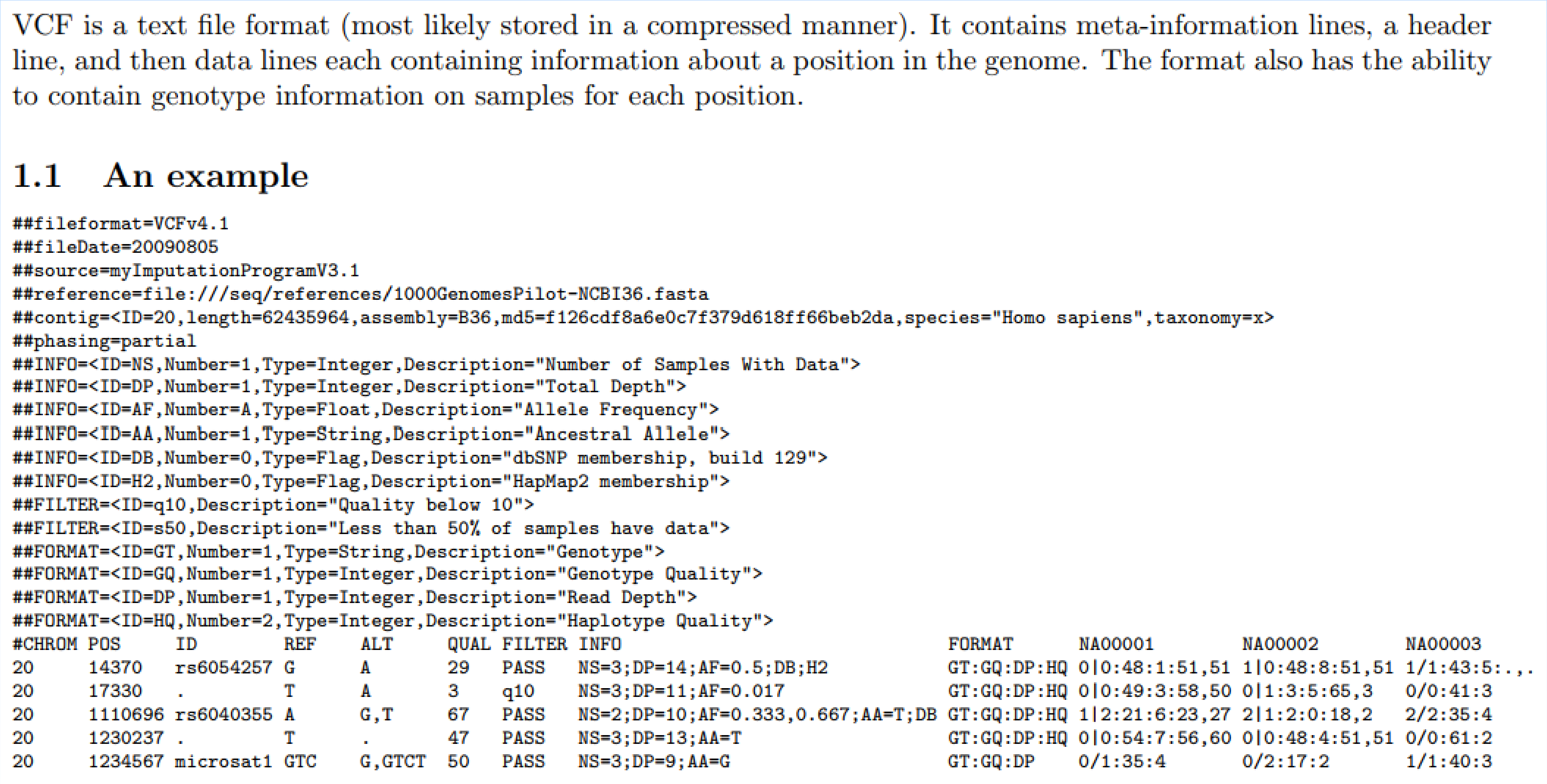
Chromosome Y, Mt: **1,085**
Affy not recommended: **71,251**
Excluded duplicate: **61,901**
Excluded multiallelic: **4,404**

Clean SNPs → 681,576 (88.45%)

Imputed SNPs → 12,723,861*

* Following imputation, SNPs are filtered to include those with an IMPUTE2 INFO score > 0.8 (confidently imputed) and to remove duplicates.

Variant Call Format (VCF)



VCF is a text file format (most likely stored in a compressed manner). It contains meta-information lines, a header line, and then data lines each containing information about a position in the genome. The format also has the ability to contain genotype information on samples for each position.

1.1 An example

```
##fileformat=VCFv4.1
##fileDate=20090805
##source=myImputationProgramV3.1
##reference=file:///seq/references/1000GenomesPilot-NCBI36.fasta
##contig=<ID=20,length=62435964,assembly=B36,md5=f126cdf8a6e0c7f379d618ff66beb2da,species="Homo sapiens",taxonomy=x>
##phasing=partial
##INFO=<ID=NS,Number=1,Type=Integer,Description="Number of Samples With Data">
##INFO=<ID=DP,Number=1,Type=Integer,Description="Total Depth">
##INFO=<ID=AF,Number=A,Type=Float,Description="Allele Frequency">
##INFO=<ID=AA,Number=1,Type=String,Description="Ancestral Allele">
##INFO=<ID=DB,Number=0,Type=Flag,Description="dbSNP membership, build 129">
##INFO=<ID=H2,Number=0,Type=Flag,Description="HapMap2 membership">
##FILTER=<ID=q10,Description="Quality below 10">
##FILTER=<ID=s50,Description="Less than 50% of samples have data">
##FORMAT=<ID=GT,Number=1,Type=String,Description="Genotype">
##FORMAT=<ID=GQ,Number=1,Type=Integer,Description="Genotype Quality">
##FORMAT=<ID=DP,Number=1,Type=Integer,Description="Read Depth">
##FORMAT=<ID=HQ,Number=2,Type=Integer,Description="Haplotype Quality">
#CHROM POS     ID        REF    ALT     QUAL FILTER INFO                              FORMAT      NA00001        NA00002        NA00003
20     14370   rs6054257 G      A       29   PASS   NS=3;DP=14;AF=0.5;DB;H2           GT:GQ:DP:HQ 0|0:48:1:51,51 1|0:48:8:51,51 1/1:43:5:.,.
20     17330   .         T      A       3    q10    NS=3;DP=11;AF=0.017               GT:GQ:DP:HQ 0|0:49:3:58,50 0|1:3:5:65,3   0/0:41:3
20     1110696 rs6040355 A      G,T     67   PASS   NS=2;DP=10;AF=0.333,0.667;AA=T;DB GT:GQ:DP:HQ 1|2:21:6:23,27 2|1:2:0:18,2   2/2:35:4
20     1230237 .         T      .       47   PASS   NS=3;DP=13;AA=T                   GT:GQ:DP:HQ 0|0:54:7:56,60 0|0:48:4:51,51 0/0:61:2
20     1234567 microsat1 GTC    G,GTCT  50   PASS   NS=3;DP=9;AA=G                    GT:GQ:DP    0/1:35:4       0/2:17:2       1/1:40:3
```

START Genotype Data Files

- All samples and controls passing sample QC. No SNP filters.
  - Binary variant call format file (bcf): START.AffyAAS.bcf
  - PLINK format files: START.AffyAAS.bed, START.AffyAAS.bim, START.AffyAAS.fam

- Control samples removed and annotations added for dbSNP RS ID ,gene symbols (+/-5kb) and SNP filters.
  - Binary variant call format file (bcf): START.ann.bcf

- Control samples removed and SNPs failing QC removed.
  - PLINK format files: START.auto.clean.bed, START.auto.clean.bim, START.auto.clean.fam

- Imputed genotype files
  - PLINK format files: START.chr#.imputed.bed, START.chr#.imputed.bim, START.chr#.imputed.fam

Mendelian randomization

- One of Mendel’s laws is *independent assortment:* the alleles at distinct genomic locations are transmitted to offspring independently
  - So if I know both copies of the genome you have and I know which allele one of your sex cell’s has at some position on chromosome 1, I have a 50% chance of predicting the allele you have at a position on another chromosome where you are heterozygous
  - One can exploit the independence of a genotype from other genetic factors under a number of assumptions
  - This is equivalent to using a technique called *instrumental variables* that is common in econometrics with a genetic variant playing the role of the instrument

Mendelian randomization

- A famous example: what is the role of alcohol consumption (a modifiable risk factor) in the development of cardiovascular disease (CVD)?
- There are many potential confounders: age, gender, on and on
- However there is a genotype that is known to have a large impact on alcohol consumption: the ADH1B gene.
- If this
  - Only impacts cardiovascular disease risk through alcohol consumption
  - Is unrelated to the confounders
- We can use the instrumental variable estimator to obtain a method for testing for an association between alcohol consumption and risk for CVD

Instrumental variables

- The instrumental variables estimator is given by the ratio of
  - The regression coefficient from regressing the outcome (CVD) on the instrument (ADH1B genotype)
  - The regression coefficient from regressing the explanatory variable (alcohol consumption) on the instrument (ADH1B genotype)
  - Can get a standard error via a number of methods
- The trick in applying this technique is finding the instrument!
- We need to find a gene that is only related to the outcome via its impact on the explanatory variable

Mendelian randomization, HIV and CVD

- Suppose we knew of a genotype that was associated with contracting HIV
- Suppose we were interested in determining the impact of being HIV positive on the development of CVD
- If the genotype only impacts the development of CVD through its impact on contracting HIV, then such a genotype could be used as an instrument
  - We would need data on the genotype and CVD for HIV negative subjects to compute the instrumental variable estimator
  - But we could assess the relationship between HIV and CVD without worrying about confounders

Haplotypes

- A haplotype is a set of alleles that are on the same chromosome
- Haplotypes can not be observed with data from conventional hybridization based genotyping platforms
- However they can sometimes be deduced
- If we can detect an association between a haplotype and a phenotype then we may have a more specific measure of risk than that based on a single SNP

Haplotypes: an example

- Consider 2 SNPs that each take 2 values
  - SNP 1 has the 2 alleles A and a
  - SNP 2 has the 2 alleles B and b
  - The possible haplotypes are:
    - (A, B), (A, b), (a, B) and (a, b)
  - Everyone has 2 haplotypes: 1 from one’s mother and the other from one’s father: for example (A, B) and (a, b)
    - This person’s genotypes at these 2 markers would be (A, a) and (B, b)-this is the genotypic data one would have
  - If someone’s genotypic data was (A, A) and (b, b) then we know someone has 2 copies of the haplotype (A, b)

Haplotypes: an example

- However there are pairs of SNP genotypes that don’t allow unambiguous determination of haplotypes
  - For example if someone’s genotypes are (A, a) and (B, b) then the possible haplotypes are (A, B) and (a, b) or (A, b) and (a, B)
- Model-based approaches to haplotype estimation have been developed-these typically assume the markers are in Hardy-Weinberg equilibrium
- There also model-based techniques for estimating the risk associated with having haplotypes when haplotypes are not known for everyone with certainty (as is usually the case)

Summary

- There are a variety of statistical tests one can use, however the regression based trend test is common in GWAS
- These regression based tests are easy to extend to accommodate confounders
  - Frequently include principle components to account for ethnicity
- Type I error is a serious problem with large numbers of tests: typically a rather drastic Bonferroni correction is used
  - FDR control can be more powerful and may be more appropriate for exploratory research

Summary

- Some investigations with INSIGHT data have been conducted
  - There is more to do and the data is available in multiple forms
- Contemporary approaches frequently use imputation to obtain data for more SNPs
  - One should probably use a stricter cut-off for statistical significance, but this is not what is commonly done
- One can do more than just test for associations between SNPs and traits
  - Using Mendelian randomization one can investigate causality in the presence of confounders
  - By testing for associations between haplotypes and traits one may be able to develop more specific genetic risk factors

Acknowledgements

- **INSIGHT Patients and Site Coordinators**

- **University of Minnesota**
  - **James Neaton**
  - **Shweta Sharma Mistry**
  - **Jacqueline Nordwall**
  - **Deborah Wentworth**
  - **Greg Thompson**

- **Advanced BioMedical Laboratories**
  - **Marie Hoover**
  - **Norman Gerry**

- **CHIP, Centre of Excellence for Health, Immunity and Infections**
  - **Jens Lundgren**
  - **Man-Hung Eric Tang**
  - **Christina Ekenberg**

- **NIAID, NIH**
  - **Clifford Lane**
  - **Julia Metcalf**

- **Leidos Biomedical Research, Inc.**
  - **Michael Baseler**
  - **Tomozumi Imamichi**
  - **Xiaojun Hu**