SNP and Sequencing-based genome-wide association studies (GWAS)

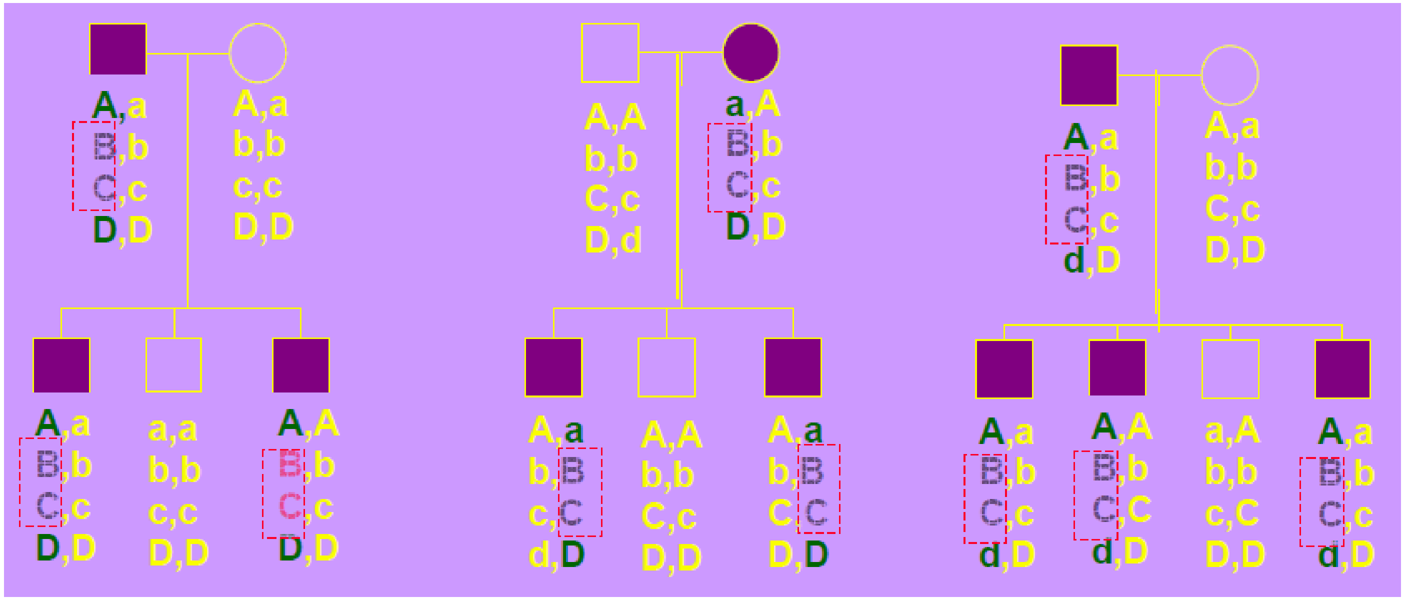
2019 Dragon Star Bioinformatics Course (Day 4)

Important Questions in Human Genetics

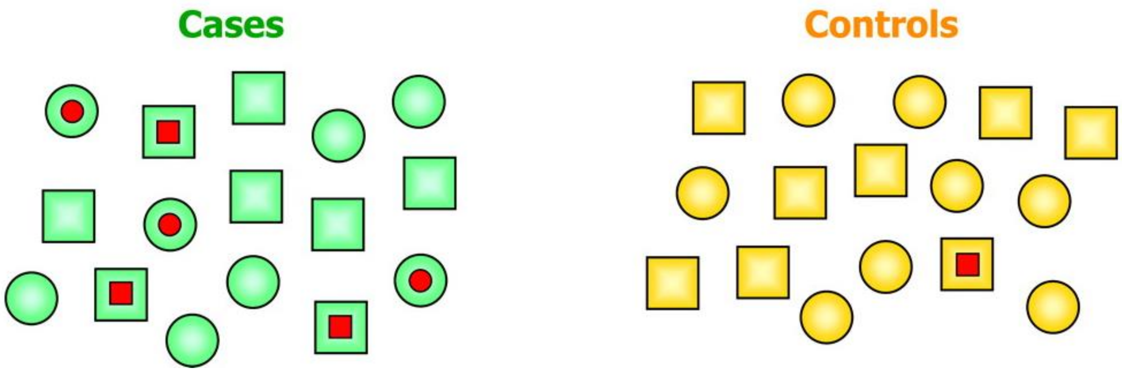
- Central Goal of human genetics:
  - Identify genetic risk factors for common, complex diseases; e.g., schizophrenia, type 2 diabetes, and rare Mendelian diseases such as cystic fibrosis.
- Genome-wide association studies (GWAS)
  - Population-based study to analyze DNA variations across the entire human genome to identify genetic risk factors associated with diseases

Linkage vs. Association

Family-based study



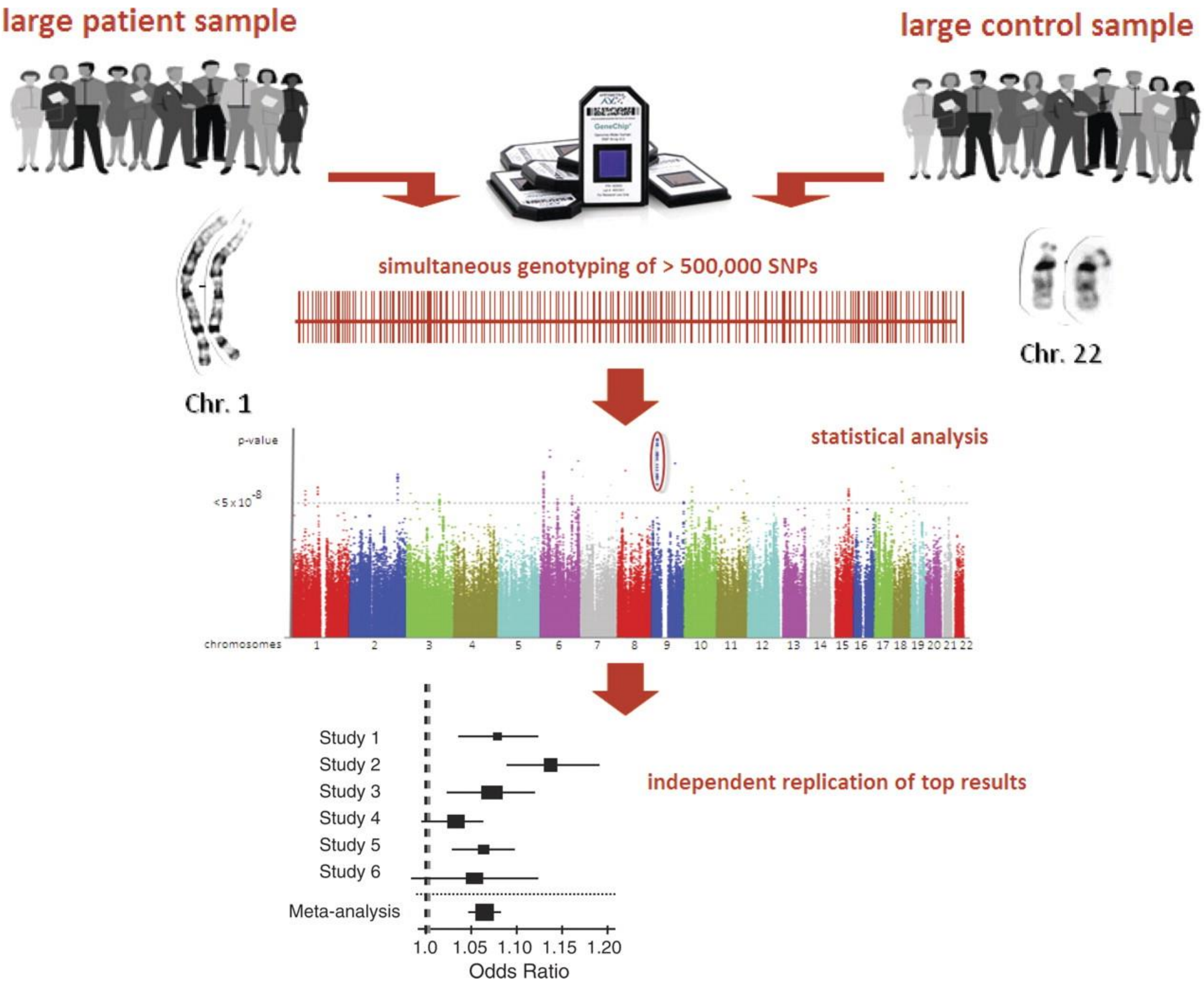
Population-based study



Linkage vs. Association

- Linkage studies
    - Pros: can scan genome with fewer markers
    - Cons: can only detect alleles with large effect; limited resolution (identify broad region, not individual genes); requires data on multiple family members
- Association studies
    - Pros: can detect subtle effects; very fine resolution; doesn’t require families
    - Cons: requires 0.5 to 1 million or even more markers to cover whole genome; requires large sample size

Genome-wide Association Studies (GWAS)



large patient sample

large control sample

simultaneous genotyping of > 500,000 SNPs

Chr. 1

Chr. 22

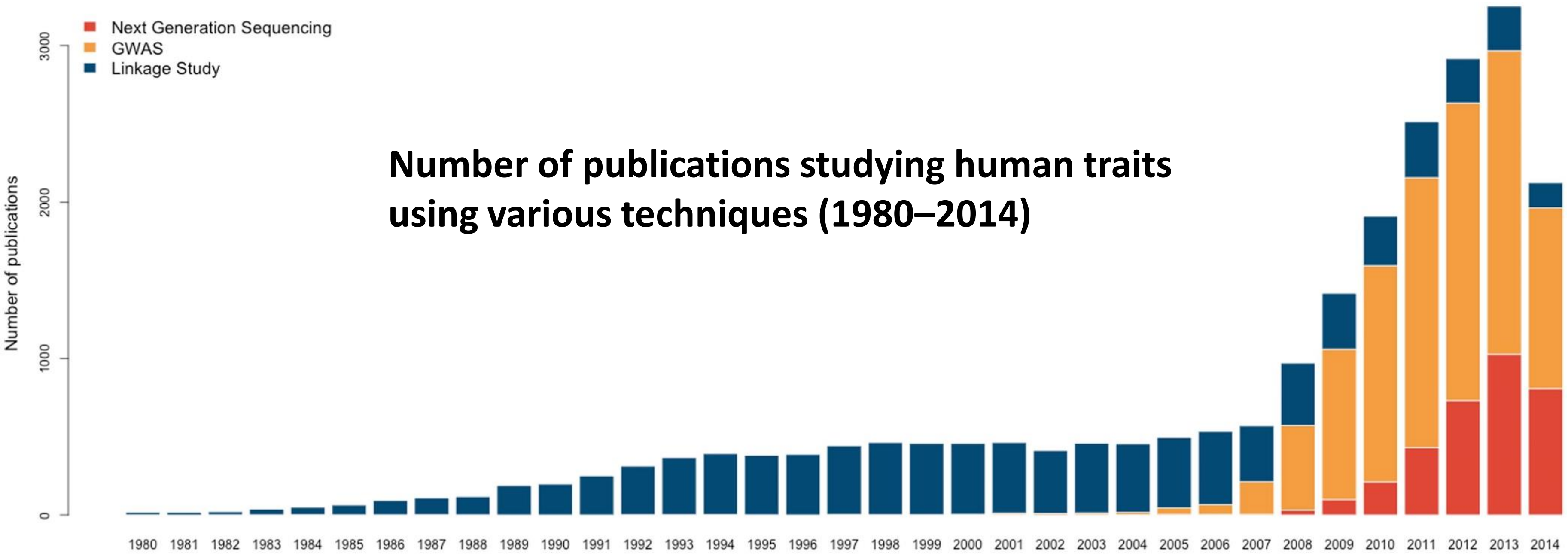
statistical analysis

independent replication of top results

Common Disease Common Variant Hypothesis

- **Hypothesis:** common disorders are likely influenced by genetic variation that is also common in the population.
- **Implications of this hypothesis:**
  - If common genetic variants influence disease, the effect size (or penetrance) for any one variant must be small relative to that found for rare disorders
    - For example, if a SNP with 40% frequency in the population causes a highly deleterious amino acid change that directly leads to a disease phenotype, nearly 40% of the population would have that phenotype
  - So under this common disease common variant hypothesis, common variants cannot have high penetrance
  - If common alleles have small genetic effects, then multiple common alleles must influence disease susceptibility together to explain heritability of common diseases

No. of Publications Studying Human Traits using Different Techniques (1980-2014)



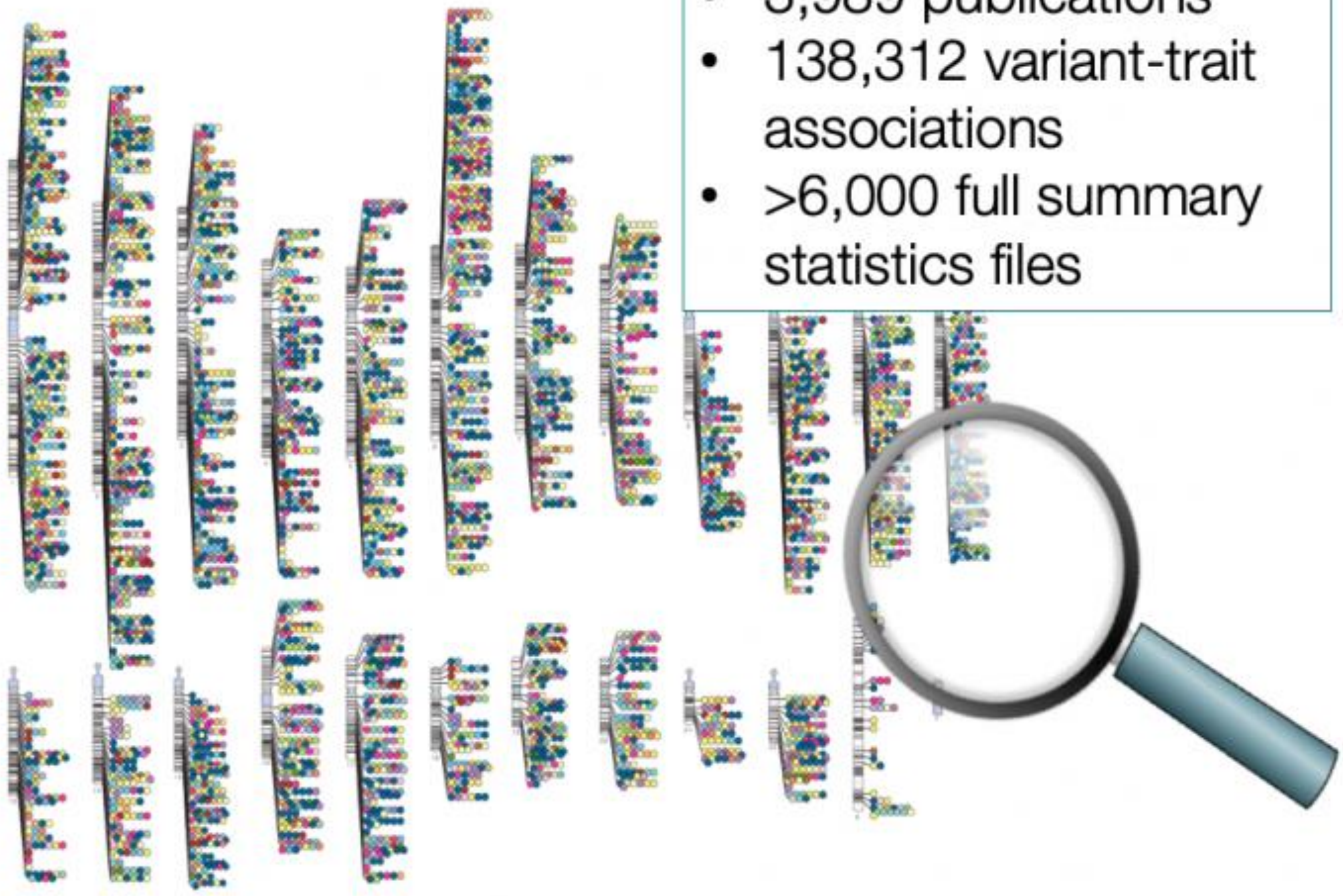
Number of publications studying human traits using various techniques (1980–2014)

Wang et al, Front. Genet., 2015

GWAS Catalog

As of May 2019

- 3,989 publications
- 138,312 variant-trait associations
- >6,000 full summary statistics files



Capturing Common Variation

- To test the common disease/common variant hypothesis, we need a systematic approach to interrogate much of the common variations in the human genome
  - First, need to have the location and density of common SNPs in the human genome
  - Second, need to catalogue population specific differences in genetic variation so that studies of phenotypes in different populations can be conducted with proper design
  - Third, need to determine correlations among common genetic variants so that genetic studies do not collect redundant information

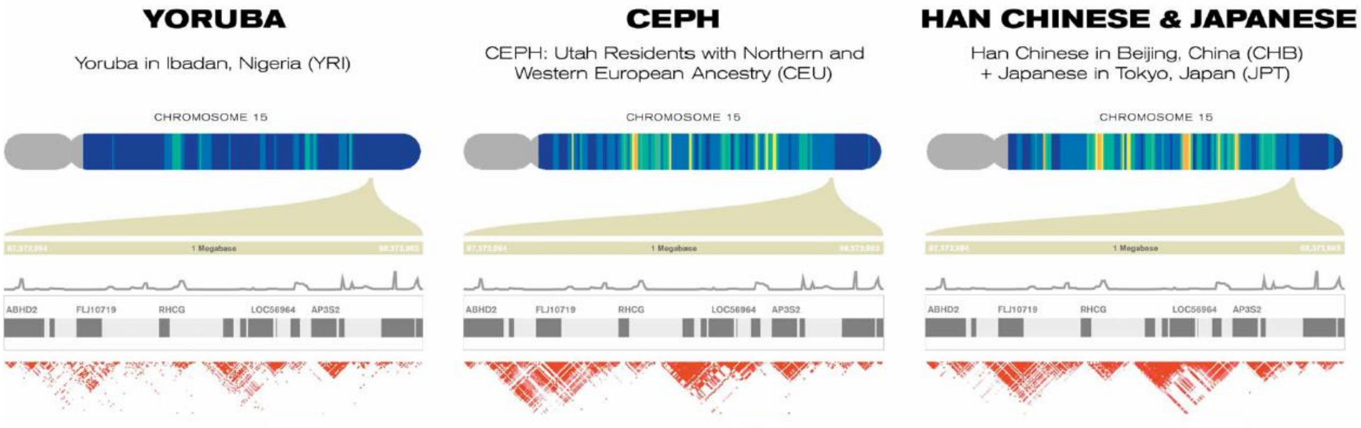
The International HapMap Project

www.hapmap.org

-Phase I and II: common SNPs in CEU, CHB, JPT, YRI

-HapMap3: 11 populations

-Patterns of linkage disequilibrium and haplotypes defined genome-wide


YORUBA

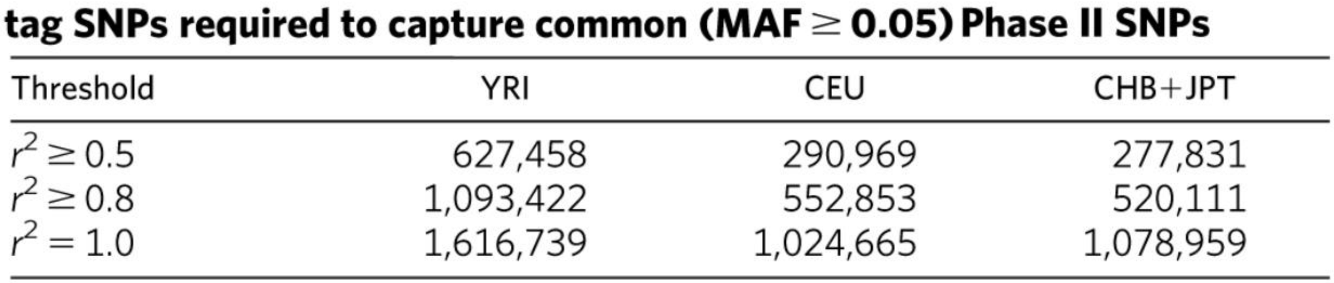
Yoruba in Ibadan, Nigeria (YRI)

CEPH

CEPH: Utah Residents with Northern and Western European Ancestry (CEU)

HAN CHINESE & JAPANESE

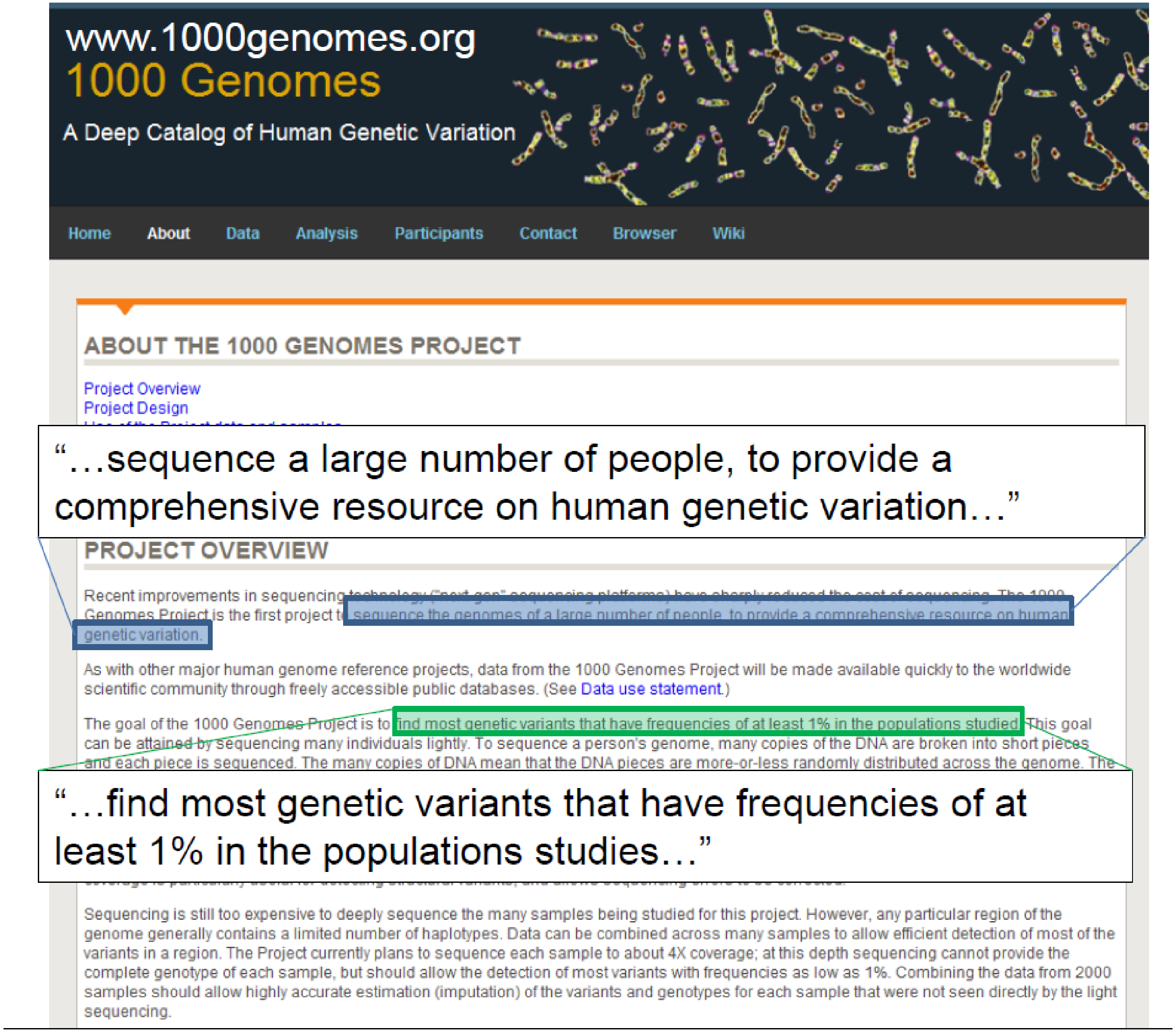
Han Chinese in Beijing, China (CHB) + Japanese in Tokyo, Japan (JPT)



tag SNPs required to capture common (MAF ≥ 0.05) Phase II SNPs

Threshold	YRI	CEU	CHB+JPT
$r^2 \geq 0.5$	627,458	290,969	277,831
$r^2 \geq 0.8$	1,093,422	552,853	520,111
$r^2 = 1.0$	1,616,739	1,024,665	1,078,959

The 1000 Genomes Project



ABOUT THE 1000 GENOMES PROJECT

Project Overview
Project Design
Use of the Project data and samples

“…sequence a large number of people, to provide a comprehensive resource on human genetic variation…”

PROJECT OVERVIEW

Recent improvements in sequencing technology ("next-gen" sequencing platforms) have sharply reduced the cost of sequencing. The 1000 Genomes Project is the first project to sequence the genomes of a large number of people, to provide a comprehensive resource on human genetic variation.

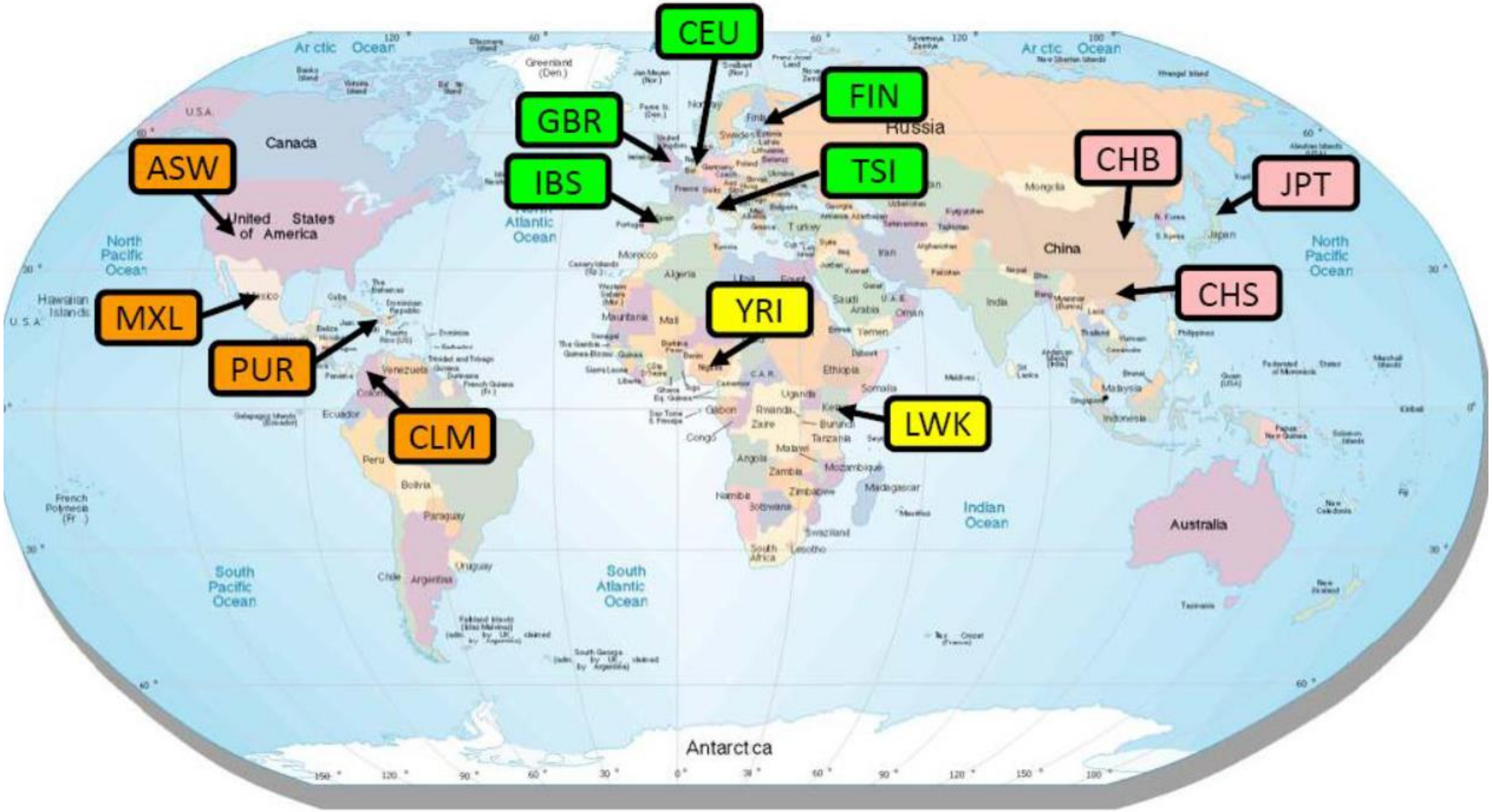
As with other major human genome reference projects, data from the 1000 Genomes Project will be made available quickly to the worldwide scientific community through freely accessible public databases. (See Data use statement.)

The goal of the 1000 Genomes Project is to find most genetic variants that have frequencies of at least 1% in the populations studied. This goal can be attained by sequencing many individuals lightly. To sequence a person's genome, many copies of the DNA are broken into short pieces and each piece is sequenced. The many copies of DNA mean that the DNA pieces are more-or-less randomly distributed across the genome. The

“…find most genetic variants that have frequencies of at least 1% in the populations studies…”

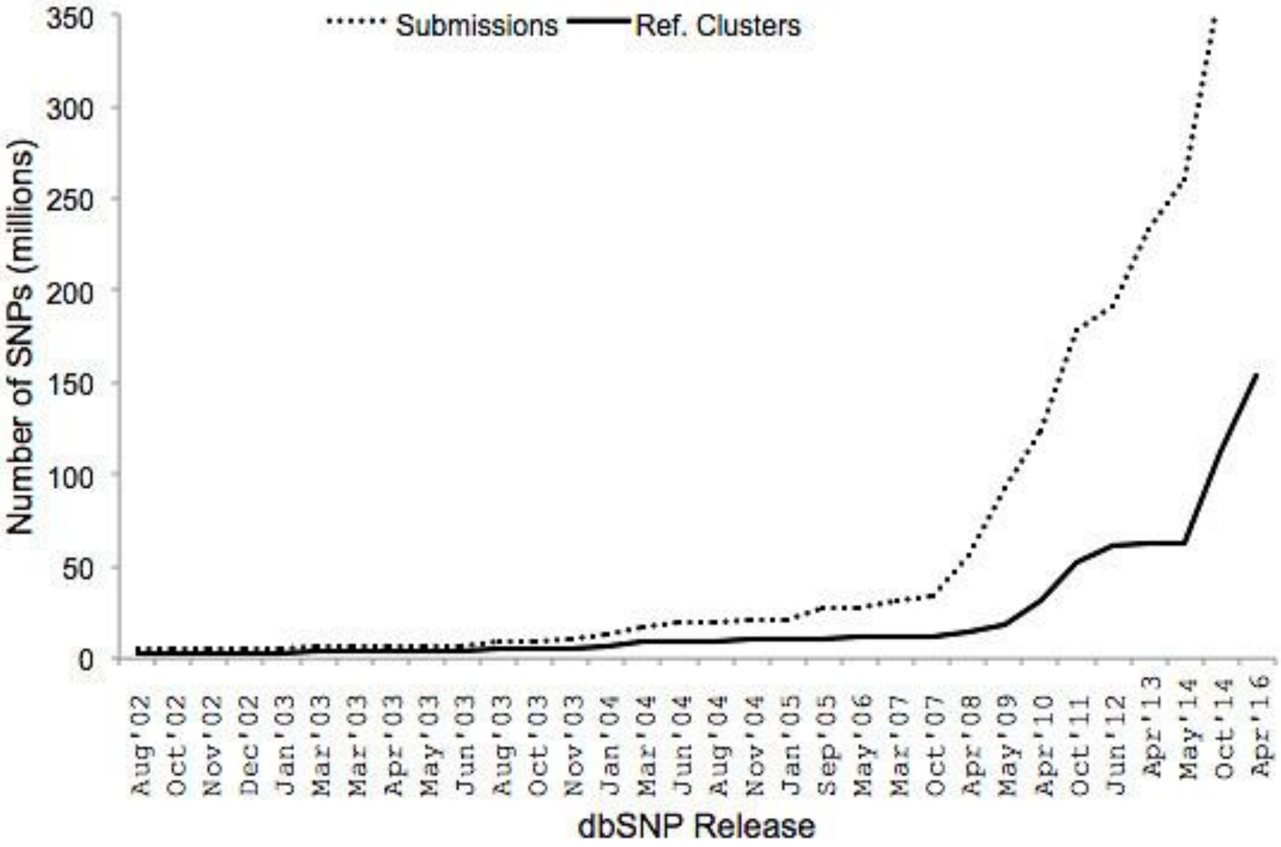
Sequencing is still too expensive to deeply sequence the many samples being studied for this project. However, any particular region of the genome generally contains a limited number of haplotypes. Data can be combined across many samples to allow efficient detection of most of the variants in a region. The Project currently plans to sequence each sample to about 4X coverage; at this depth sequencing cannot provide the complete genotype of each sample, but should allow the detection of most variants with frequencies as low as 1%. Combining the data from 2000 samples should allow highly accurate estimation (imputation) of the variants and genotypes for each sample that were not seen directly by the light sequencing.

1000G Phase I populations



Concepts Underlying the GWAS Design

- The modern unit of genetic variation is SNP
- GWAS become feasible due to success of the HapMap Project and technology improvement.

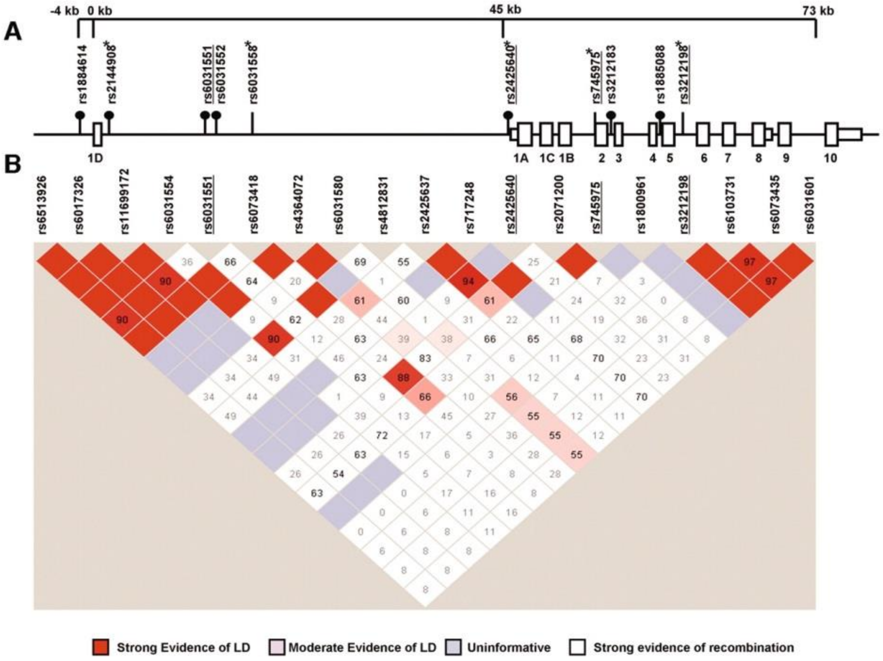


Linkage Disequilibrium (LD)

- What is LD?
  - Describes the degree to which an allele of one SNP is correlated with an allele of another SNP within a population
  - Different from linkage (correlation within a family)
- Different human subpopulations have different degrees and patterns of LD.
  - African-descent populations are the most ancestral and have smaller regions of LD due to accumulation of more recombination events
  - European-descent and Asian-descent populations were created by founder events (a sampling of chromosomes from the African population), so these populations on average have larger regions of LD than Africans

Measures of LD

		Locus B		Totals
		B	b	
Locus A	A	p_{AB}	p_{Ab}	p_A
	a	p_{aB}	p_{ab}	p_a
Totals		p_B	p_b	1.0



$$\text{LD coefficient}: \; D_{AB} = p_{AB} - p_A p_B$$

Two Most Popular Measures of LD

- D’:

$$D'_{AB} = \frac{D_{AB}}{D_{max}} = \begin{cases} \dfrac{D_{AB}}{\min(p_A p_B, p_a p_b)} & D_{AB} < 0 \\ \dfrac{D_{AB}}{\min(p_A p_b, p_a p_B)} & D_{AB} > 0 \end{cases}$$

Ranges between –1 and +1

- More likely to take extreme values when allele frequencies are small

- ±1 implies at least one of the four haplotypes is not observed

Two Most Popular Measures of LD

- **r^2:**

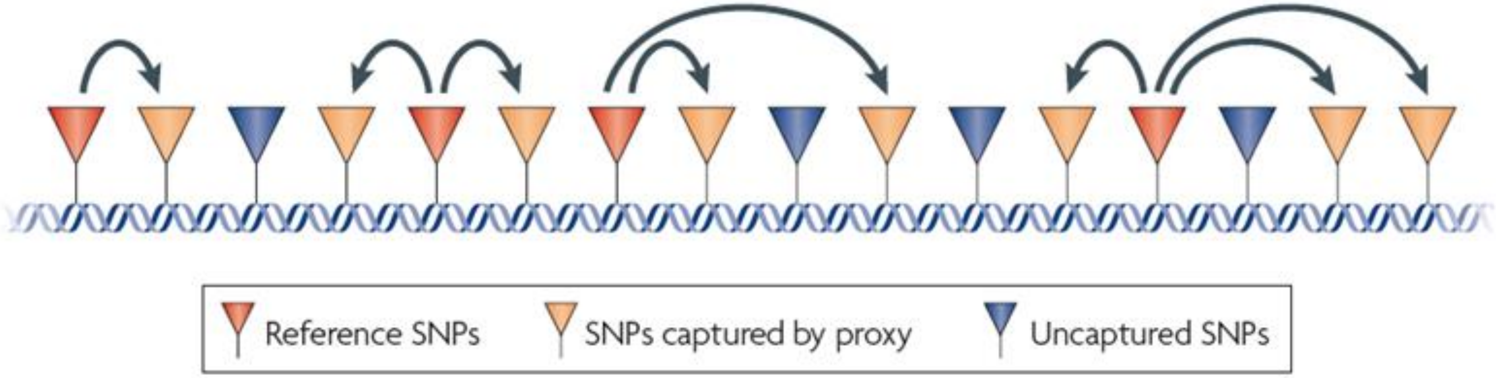
$$r^2 = \frac{D_{AB}^2}{p_A(1-p_A)p_B(1-p_B)} = \frac{(p_{AB}-p_Ap_B)^2}{p_A(1-p_A)p_B(1-p_B)} = \frac{\chi^2}{2n}$$

Ranges between 0 and 1

- $r^2 = 1$ when the two markers provide identical information
- r^2 is the squared statistical correlation.
- For low allele frequencies, r^2 has more reliable sample properties than D’

Direct vs. Indirect Association

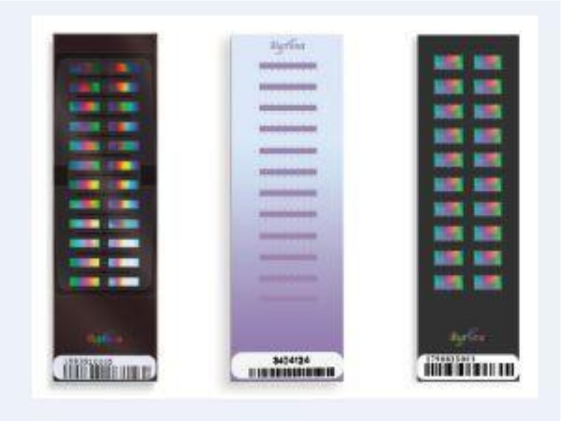
- **Direct association:**
  - The SNP that leads to the phenotype is directly genotyped and found associated with the trait.
- **Indirect association:**
  - The influential SNP is not directly genotyped, but a tag SNP in high LD with the influential SNP is typed and associated with the trait.
  - Due to the possibility of indirect association, a significant SNP association from a GWAS should not be interpreted as “causal”.

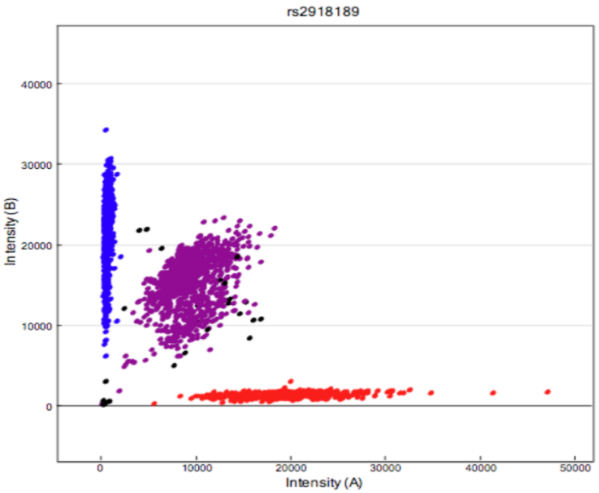


Calling SNPs in GWAS Array

Affymetrix

Illumina

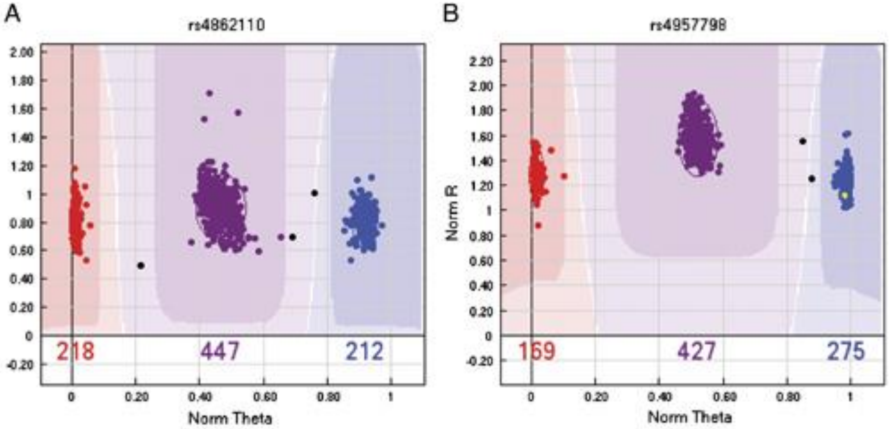




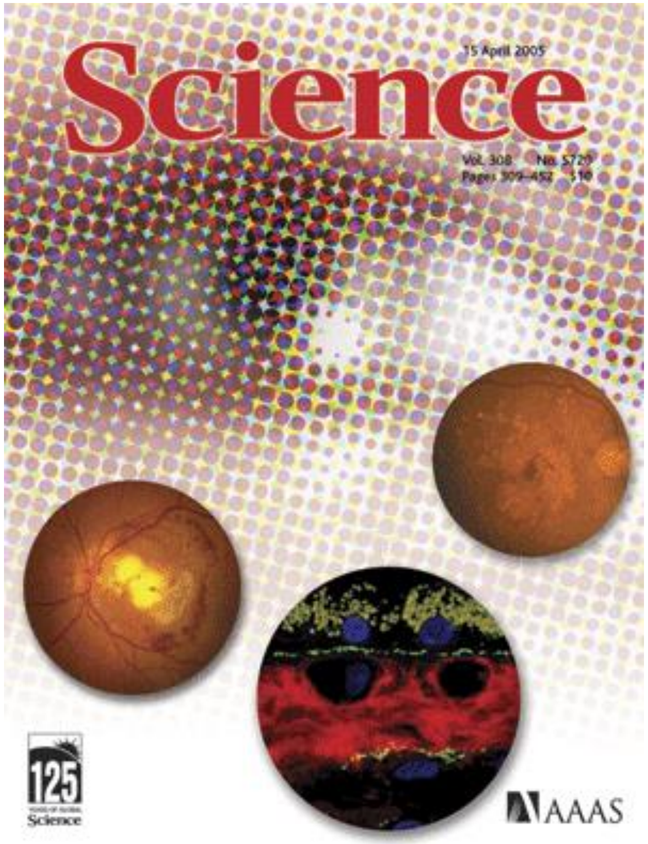
Ratio of intensities from two channels

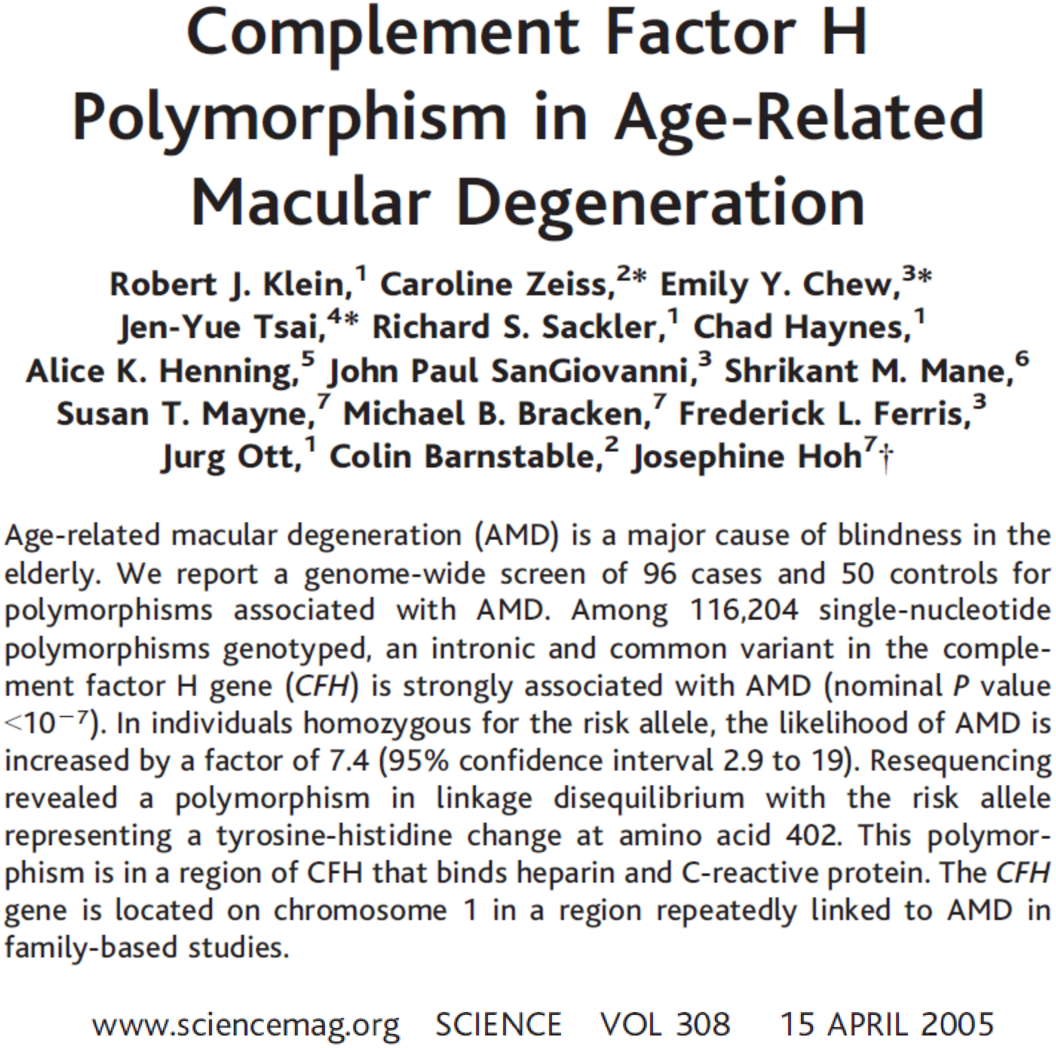
Calls = 2461

No calls = 27



The First GWAS (2005)





Complement Factor H Polymorphism in Age-Related Macular Degeneration

Robert J. Klein,[1] Caroline Zeiss,[2]* Emily Y. Chew,[3]*
Jen-Yue Tsai,[4]* Richard S. Sackler,[1] Chad Haynes,[1]
Alice K. Henning,[5] John Paul SanGiovanni,[3] Shrikant M. Mane,[6]
Susan T. Mayne,[7] Michael B. Bracken,[7] Frederick L. Ferris,[3]
Jurg Ott,[1] Colin Barnstable,[2] Josephine Hoh[7]†

Age-related macular degeneration (AMD) is a major cause of blindness in the elderly. We report a genome-wide screen of 96 cases and 50 controls for polymorphisms associated with AMD. Among 116,204 single-nucleotide polymorphisms genotyped, an intronic and common variant in the complement factor H gene (*CFH*) is strongly associated with AMD (nominal P value $<10^{-7}$). In individuals homozygous for the risk allele, the likelihood of AMD is increased by a factor of 7.4 (95% confidence interval 2.9 to 19). Resequencing revealed a polymorphism in linkage disequilibrium with the risk allele representing a tyrosine-histidine change at amino acid 402. This polymorphism is in a region of CFH that binds heparin and C-reactive protein. The *CFH* gene is located on chromosome 1 in a region repeatedly linked to AMD in family-based studies.

www.sciencemag.org SCIENCE VOL 308 15 APRIL 2005

Steps in GWAS Analysis

- Store large amounts of genotype data
- Quality control analysis
- Generate initial association analysis results
- Visualize results
- Impute missing SNP genotypes
- Store results and plan specialized analysis

GWAS genotype data is large

- 500,000 SNPs * 2000 cases + controls = 1,000,000,000 genotypes!
- Need programs to select and write out genotype data in multiple formats
- FOR GWAS data, the most commonly used program is PLINK; there are also R packages that integrates PLINK

Quality Control in GWAS

Sample level quality control

- Remove samples with low genotype call rate
- Remove samples with excess/deficient heterozygosity
- Check for gender inconsistency
- Check for cryptic relatedness, remove one from each relative pair
- Identify genetically homogeneous group, remove population outliers

Quality Control in GWAS

SNP level quality control

- Remove SNPs with low genotype call rate
- Remove SNPs with low minor allele frequency (MAF)
- Remove SNPs that fail Hardy-Weinberg equilibrium test
- Remove SNPs that show non-random missingness in cases and controls

Check of Call Rate and Heterozygosity

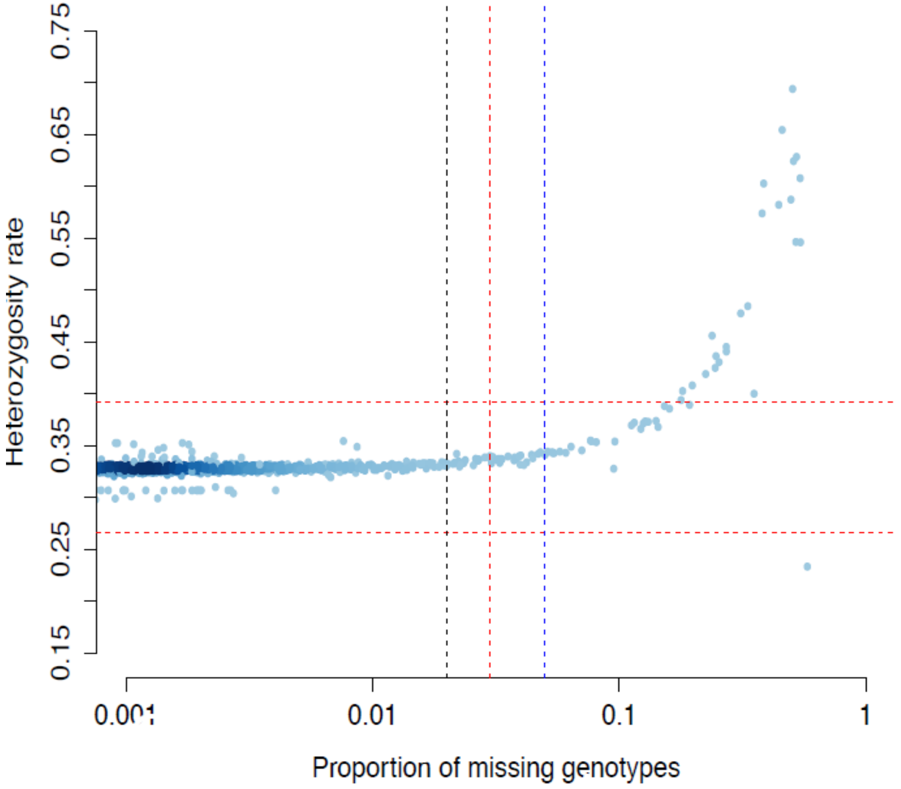
Genotyping call rate

- Per sample (individual) rate
- Number of **non-missing genotypes** divided by the total number of **genotyped** markers.
- Low genotyping call rate indicate problem with **sample DNA** like low concentration.
- Thresholds used generally vary between 3% and 7%

Genotyping call rate and heterozygosity rate are generally plotted together. Cutoffs are selected so as to identify outlier individuals based on both the statistics

Heterozygosity Rate

- Per sample (individual) rate
- Number of (total non-missing genotypes(N) – homozygous(0)) genotypes divided by total non-missing genotypes(N)
- Excess heterozygosity - Possible sample **contamination**
- Less than expected heterozygosity- Possibly **inbreeding**
- Threshold for inclusion is generally Mean ± 3 std.dev. over all samples



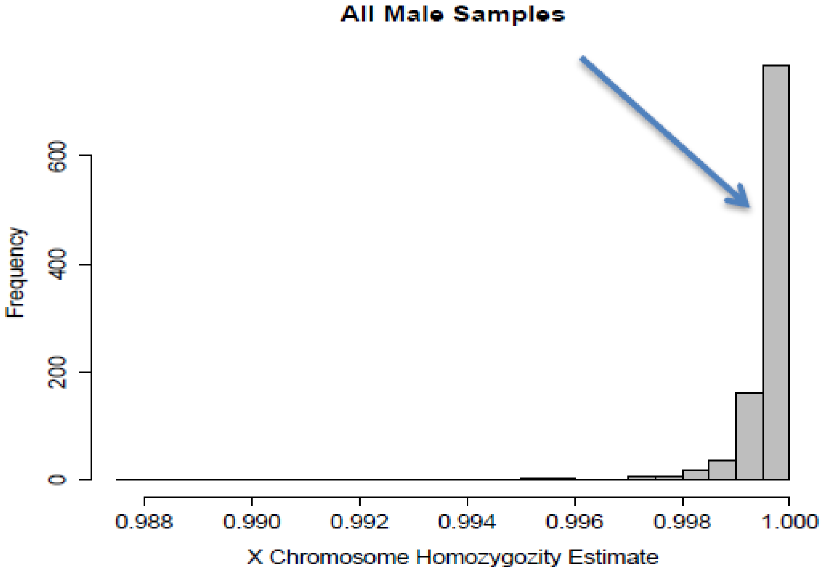
Check of Gender Concordance

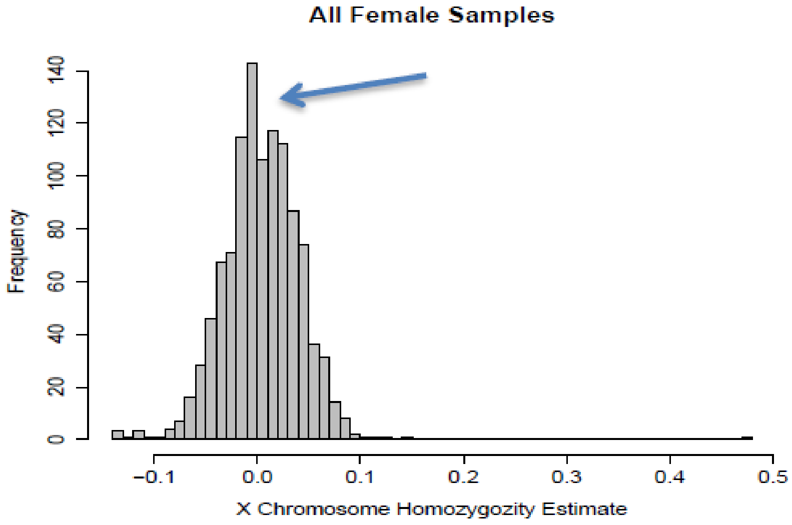
- Males have a single X chromosome and therefore can be estimated to be homozygous for all the X chromosome SNPs (other than those in the pseudo autosomal region(PAR)).
- Therefore, X chromosome homozygosity estimate for males(XHE) is 1

- Plink assigns sex based on XHE estimate (F or inbreeding coefficient) :
- Male (1) : XHE >0.80
- Female (2) : XHE <0.20
- No sex (0) : 0.20 <XHE <0.80

- Comparisons of predicted and observed sex can be used to identify miscoded sex or **sample mix-ups**, etc.

- Samples with discordant sex information are removed

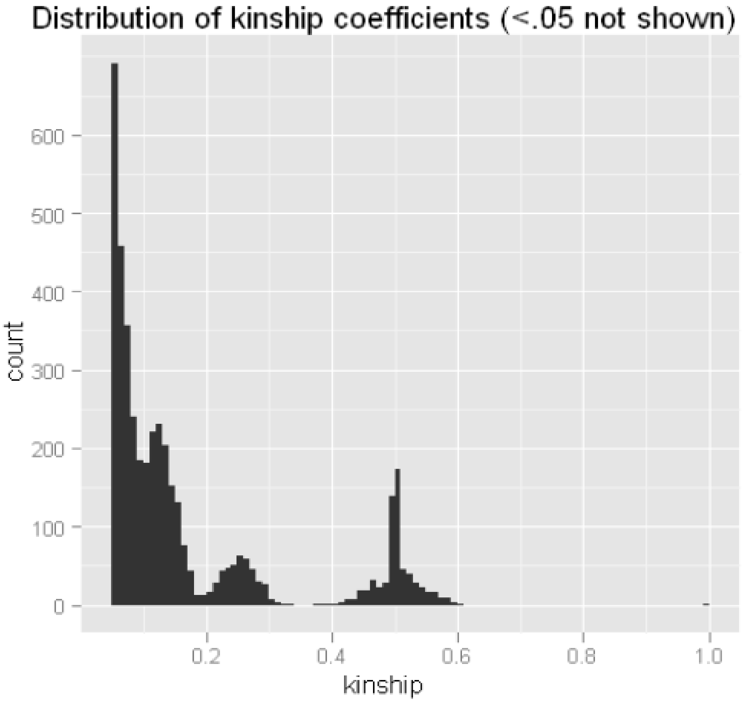

All Male Samples


All Female Samples

Check of Sample Relatedness

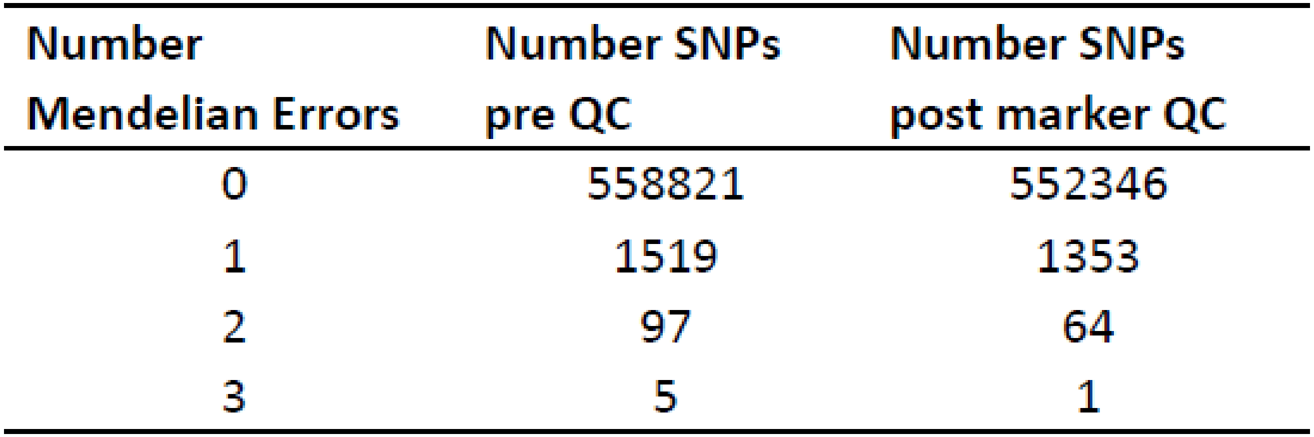
- IBD used to be inferred for family data only
- Now with large-scale genetic data available, it is possible to infer IBD even for population-based data
- This can help identify cryptically related individuals

Z0	Z1	Z2	Kinship	Relationship
0.0	0.0	1.0	1.0	MZ twin or duplicate
0.0	1.0	0.0	0.50	Parent-offspring
0.25	0.50	0.25	0.50	Full siblings
0.50	0.50	0.0	0.25	Half siblings
0.75	0.25	0.0	0.125	Cousins
1.0	0.0	0.0	0.0	Unrelated



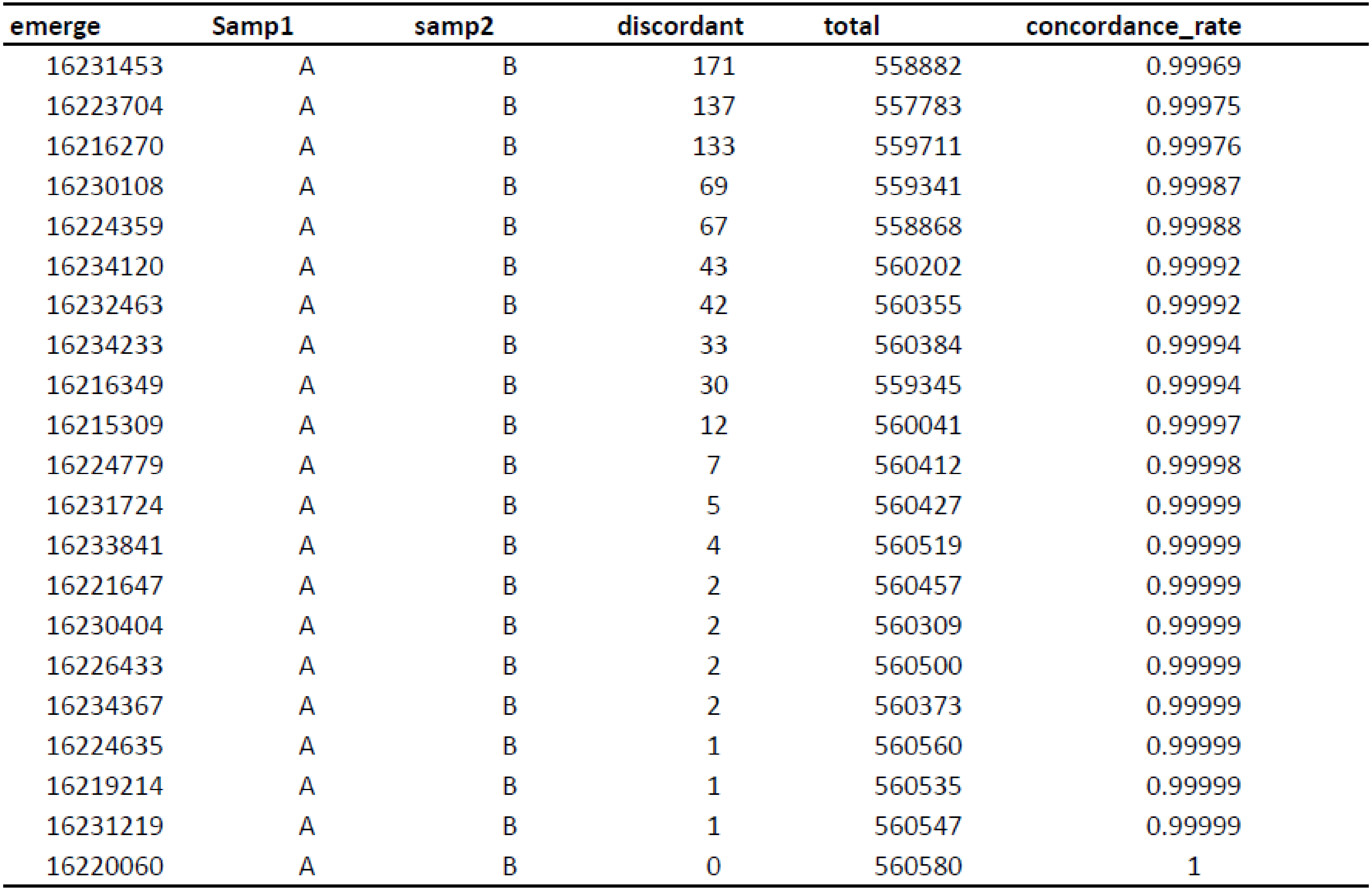
Check of Mendelian Inheritance Errors

- Even with Case/control data, HapMap trios are typically plated with study samples for QC purpose.



Number Mendelian Errors	Number SNPs pre QC	Number SNPs post marker QC
0	558821	552346
1	1519	1353
2	97	64
3	5	1

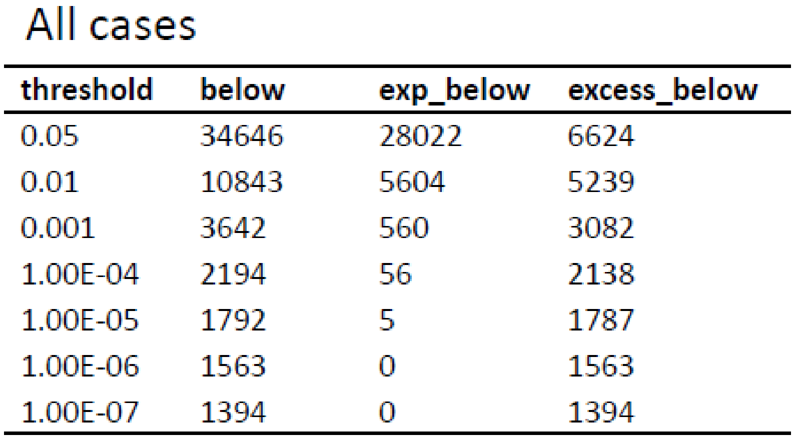
Sample Replicate Concordance



emerge	Samp1	samp2	discordant	total	concordance_rate
16231453	A	B	171	558882	0.99969
16223704	A	B	137	557783	0.99975
16216270	A	B	133	559711	0.99976
16230108	A	B	69	559341	0.99987
16224359	A	B	67	558868	0.99988
16234120	A	B	43	560202	0.99992
16232463	A	B	42	560355	0.99992
16234233	A	B	33	560384	0.99994
16216349	A	B	30	559345	0.99994
16215309	A	B	12	560041	0.99997
16224779	A	B	7	560412	0.99998
16231724	A	B	5	560427	0.99999
16233841	A	B	4	560519	0.99999
16221647	A	B	2	560457	0.99999
16230404	A	B	2	560309	0.99999
16226433	A	B	2	560500	0.99999
16234367	A	B	2	560373	0.99999
16224635	A	B	1	560560	0.99999
16219214	A	B	1	560535	0.99999
16231219	A	B	1	560547	0.99999
16220060	A	B	0	560580	1

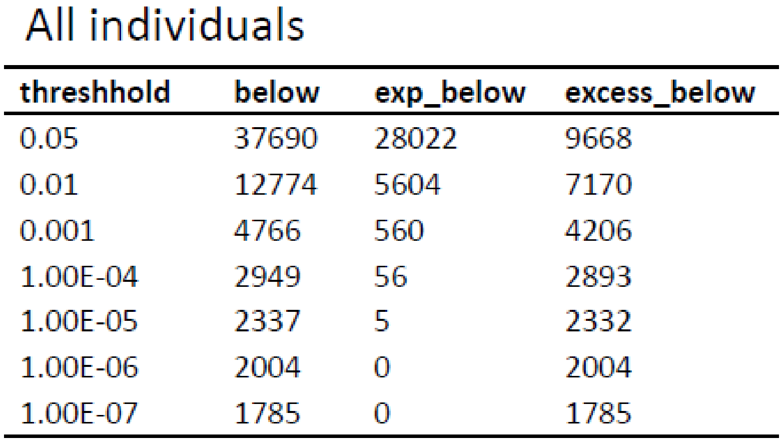
Check Hardy-Weinberg Equilibrium (HWE)

For a single locus with two alleles denoted A and a with frequencies f(A) = p and f(a) = q, the expected genotype frequencies under random mating are f(AA) = p^2 for the AA homozygotes, f(aa) = q^2 for the aa homozygotes, and f(Aa) = 2pq for the heterozygotes



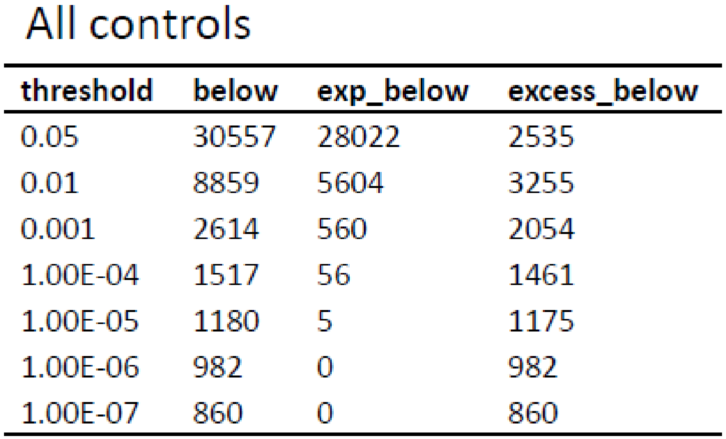
All cases

threshold	below	exp_below	excess_below
0.05	34646	28022	6624
0.01	10843	5604	5239
0.001	3642	560	3082
1.00E-04	2194	56	2138
1.00E-05	1792	5	1787
1.00E-06	1563	0	1563
1.00E-07	1394	0	1394



All individuals

threshhold	below	exp_below	excess_below
0.05	37690	28022	9668
0.01	12774	5604	7170
0.001	4766	560	4206
1.00E-04	2949	56	2893
1.00E-05	2337	5	2332
1.00E-06	2004	0	2004
1.00E-07	1785	0	1785

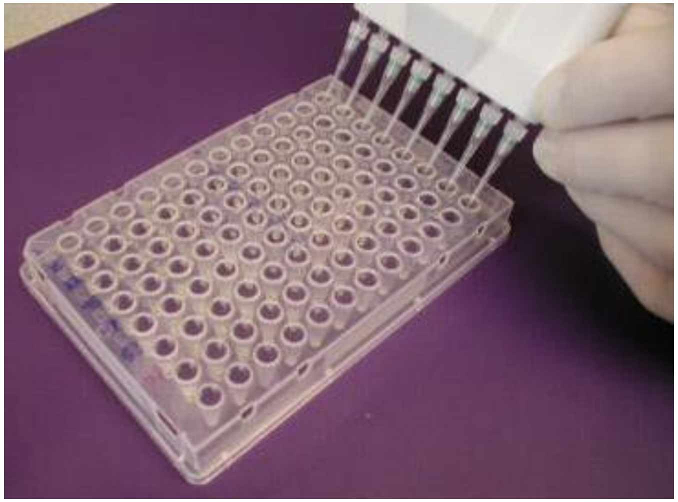


All controls

threshold	below	exp_below	excess_below
0.05	30557	28022	2535
0.01	8859	5604	3255
0.001	2614	560	2054
1.00E-04	1517	56	1461
1.00E-05	1180	5	1175
1.00E-06	982	0	982
1.00E-07	860	0	860

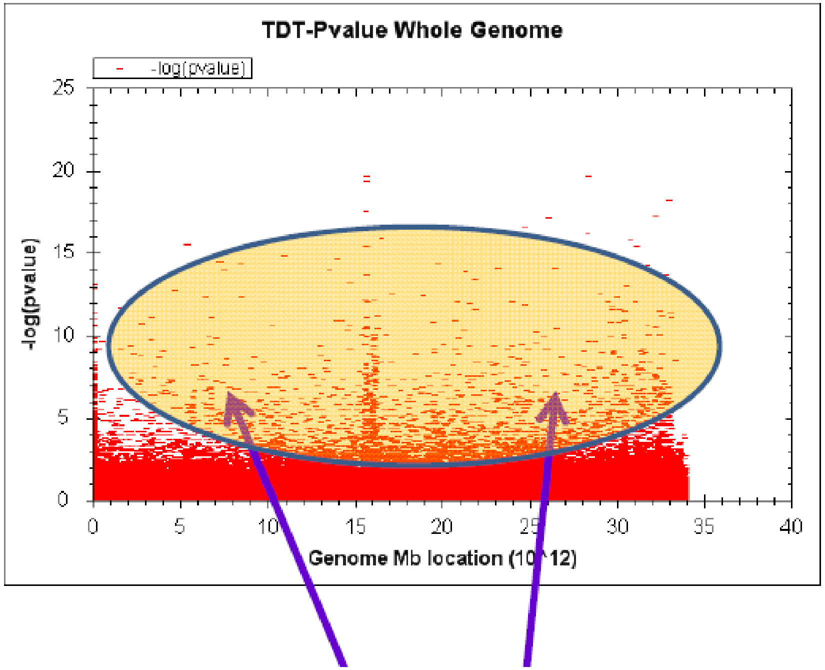
Check of Batch Effects

- Spurious association due to allele frequency difference in different plates
- Careful consideration when creating plate maps
  - Randomize cases and controls
  - Randomize by race, gender, age, BMI, …
- After genotyping, look for plate effects
  - MAF differences by plate
  - Call rate by plate
  - Association tests (one plate vs. all others)

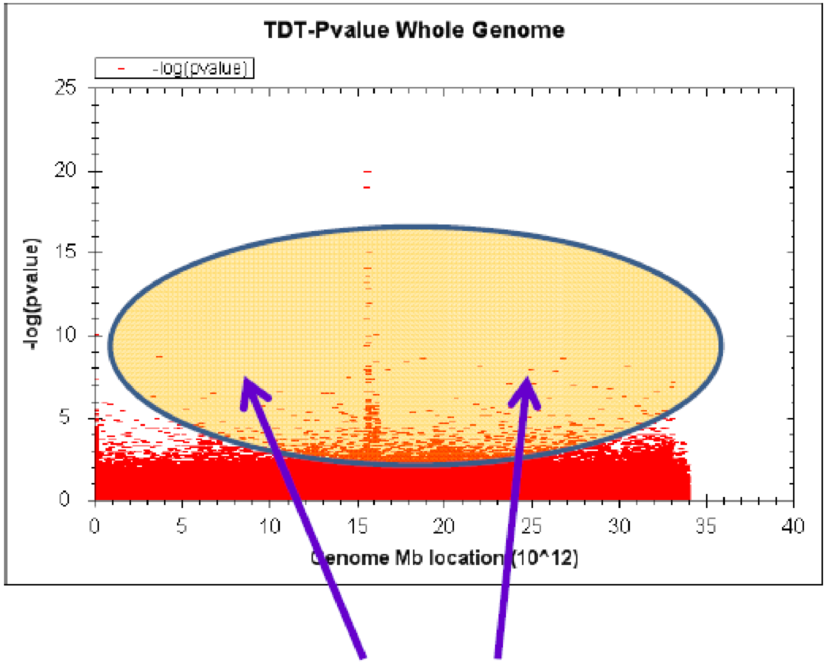


Importance of QC

Pre-QC Thresholds



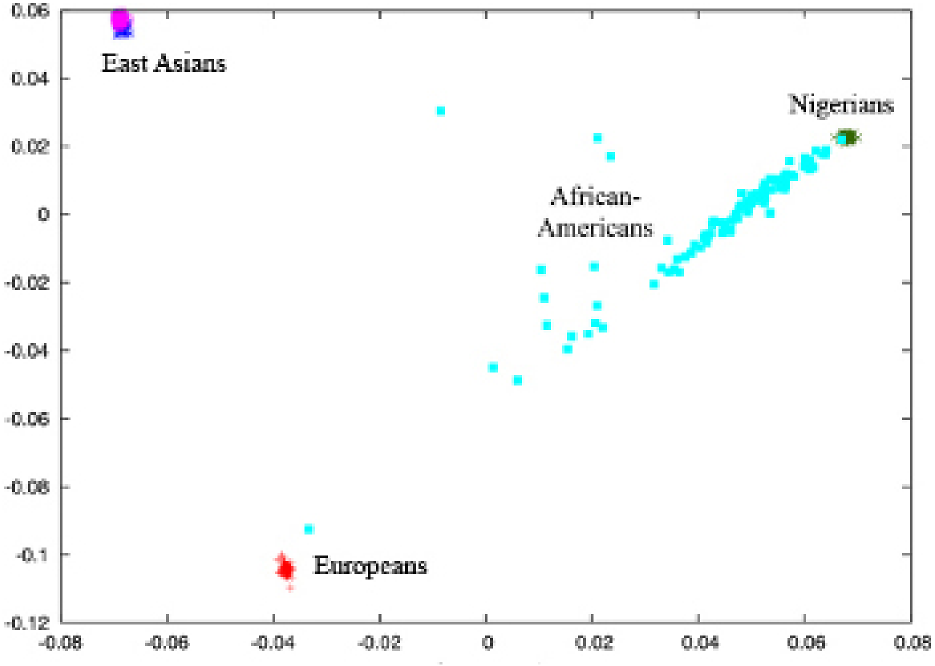
Post-QC Thresholds

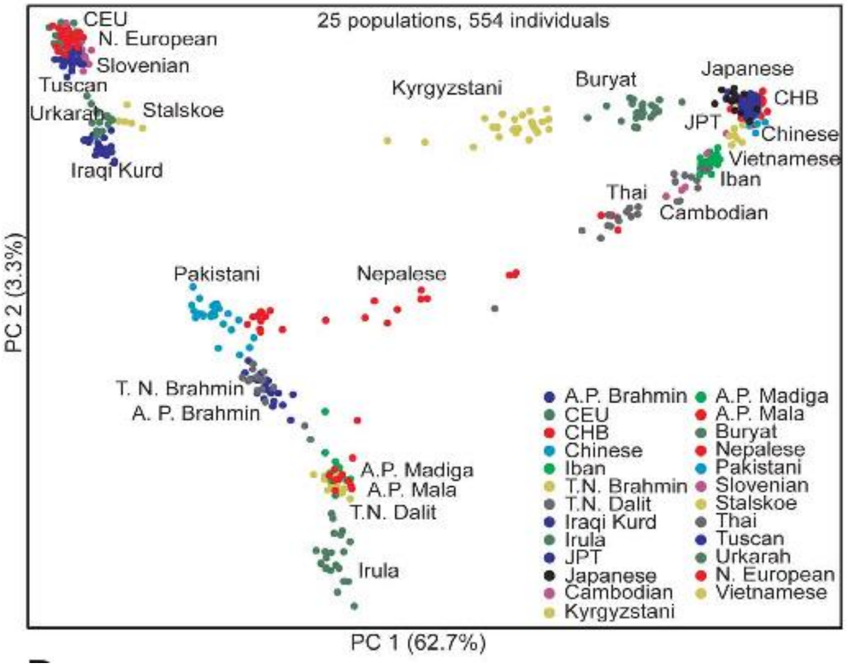


Many false positives disappear after QC

Genetic Race

Two-dimensional visualization of genotype data, with samples from different populations





http://blogs.discovermagazine.com/gnxp/2010/07/one-principal-component-to-rule-them-all

Association Test in GWAS

- Basic statistical methods are usually applied to test for genetic association in GWAS

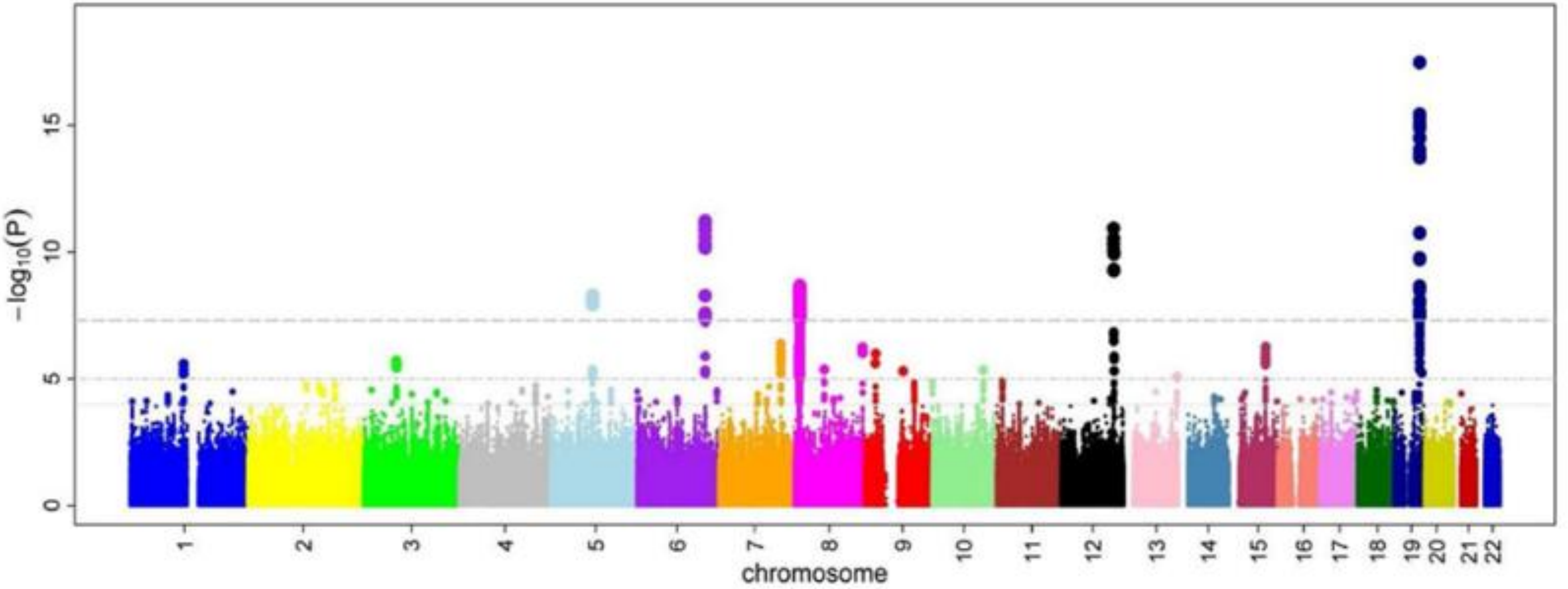
- Quantitative traits
  - Linear regression with appropriate transformations

- Binary traits
  - Logistic regression (case/control design)
  - Family-based association (family-based design)

- Adjustment of covariates is critical to avoid confounding

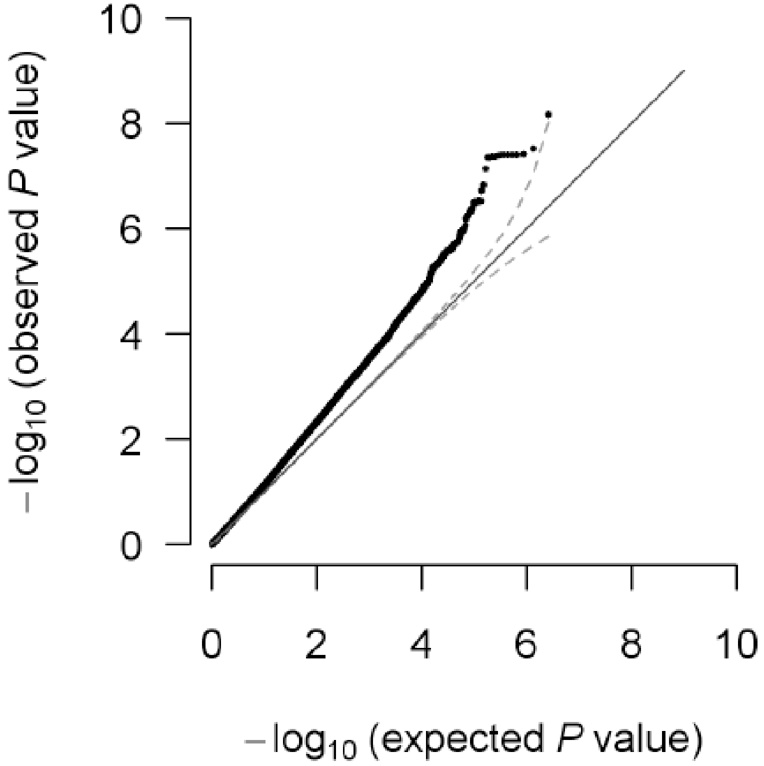
Graphical Display of Results

- An illustration of a **<u>Manhattan plot</u>** depicting several strongly associated risk loci.
  - In GWAS Manhattan plots, genomic coordinates are displayed along the X-axis, with –log10 of the association p-value for each SNP displayed on the Y-axis, meaning that each dot on the Manhattan plot signifies a SNP.

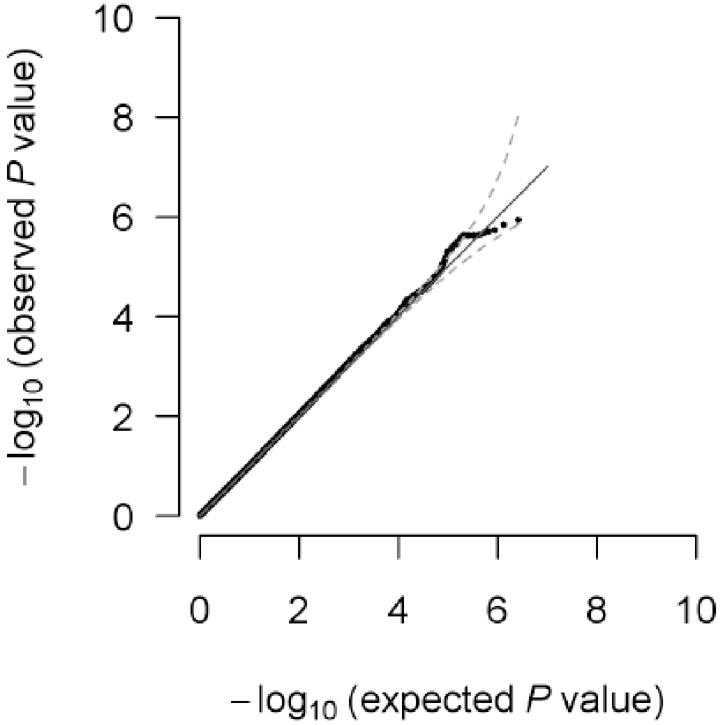


Graphical Display of Results

- QQ plot



Problem: early departure possibly due to population stratification



No strong evidence of population stratification

Remarks on QQ Plot

- Rank observed –log10(p-values) from most significant to least
- Pair these with expected values from order statistics of a Uniform(0,1) distribution (the distribution of p-values under Ho).
- Plot the matched pairs
- Construct confidence bands (bands get wider at end because more variability at ends of distribution and because of –log10 transformation.
- Look for early departures. Late departures are to be expected if there are really truly causal variants.

Multiple Testing Correction

- Perform 1,000,000 tests in a typical GWAS
  - If set type I error rate at 5%, we can expect 50,000 false positive results, too many false positives!
  - Bonferroni correction: too stringent due to LD among SNPs.
  - The exact threshold
    - Varies by study
    - Conventional threshold is $\mathbf{5\times10^{-8}}$ to be significant in the face of hundreds of thousands to millions of tested SNPs.
    - This threshold is obtained by estimating the total number of “independent” SNPs in the genome.

Replication

- Now required for consideration of publication
- Ideally should be interchangeable with the first sample in every way
- Need all the covariates you used in the first dataset

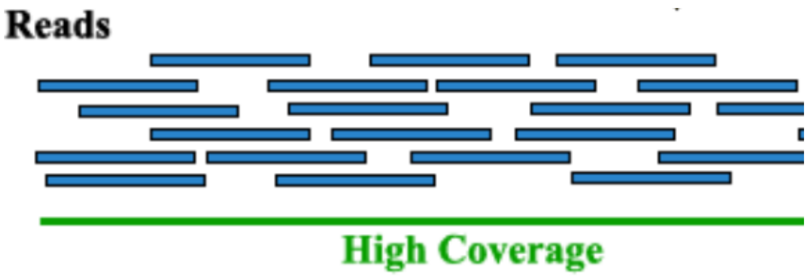
From SNP based to sequencing based GWAS

- Dissecting complex traits will require whole genome sequencing of thousands of individuals

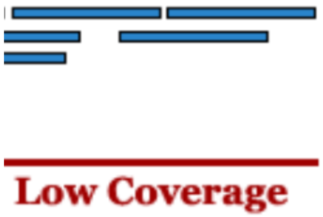
- How to sequence thousands of individuals cost-effectively?

Current Genome Scale Approaches

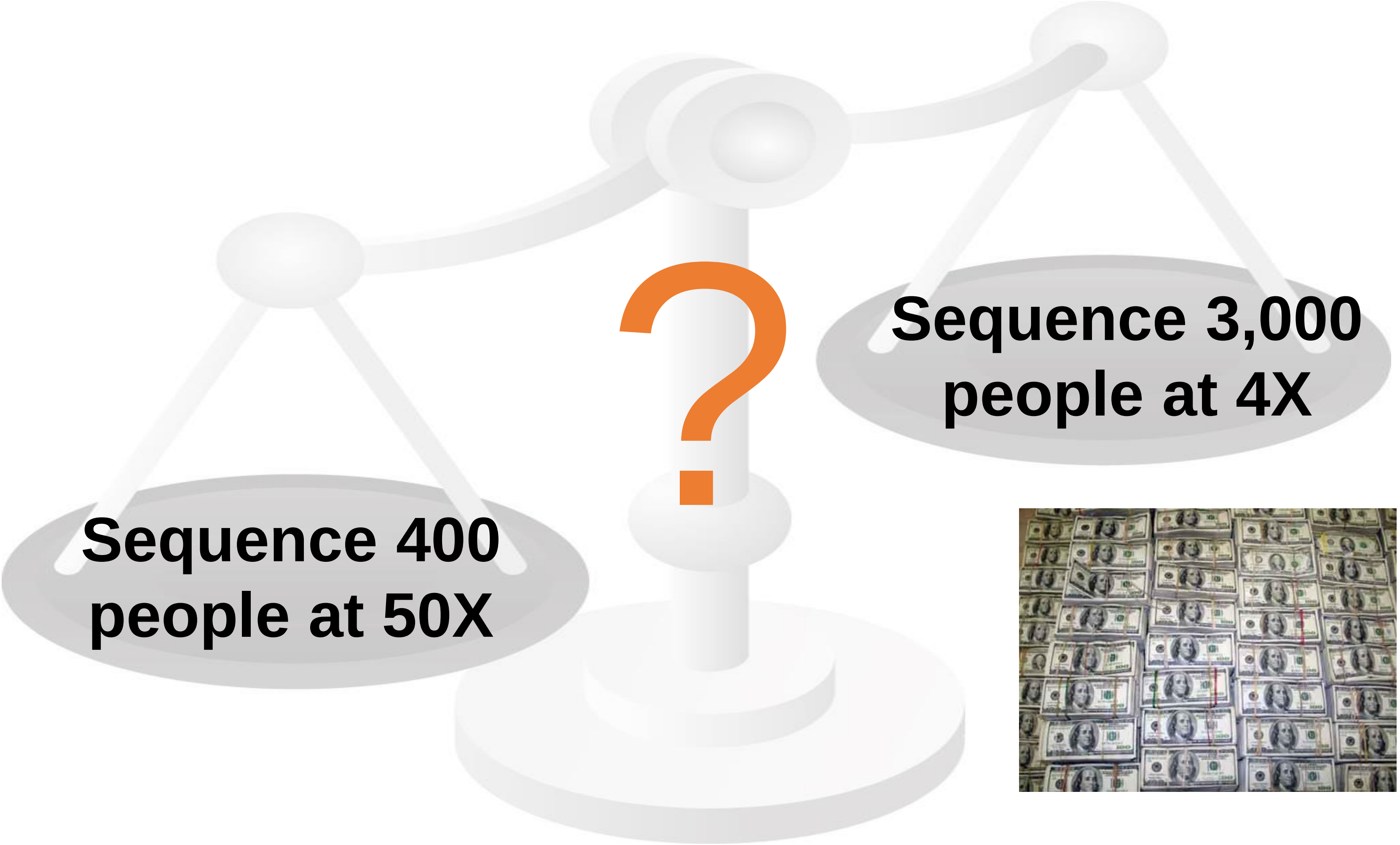
- **High coverage whole genome sequencing**
  - Can only be applied to limited number of samples
  - Most complete ascertainment of genetic variations



- **Low coverage whole genome sequencing**
  - Can be applied to moderate numbers of samples
  - Very complete ascertainment of shared variations across samples
  - Less complete ascertainment of rare variants



High vs Low Coverage Sequencing

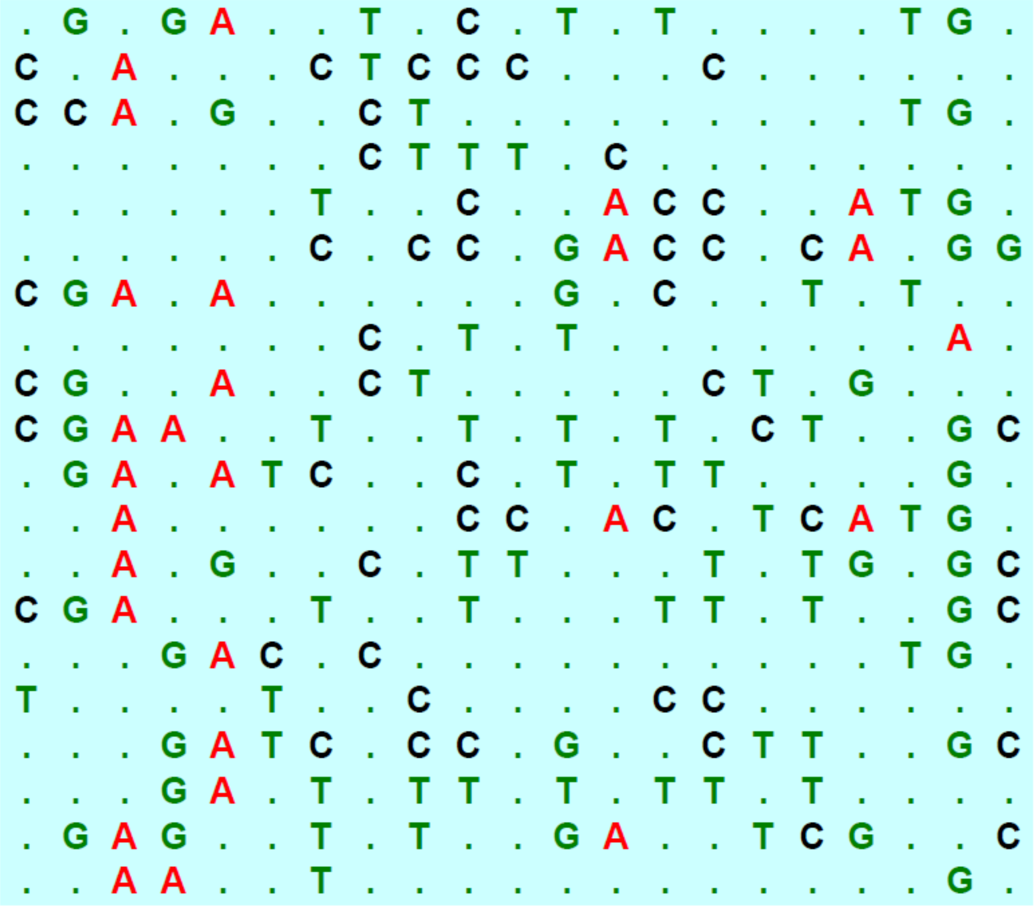


?

Sequence 3,000 people at 4X

Sequence 400 people at 50X

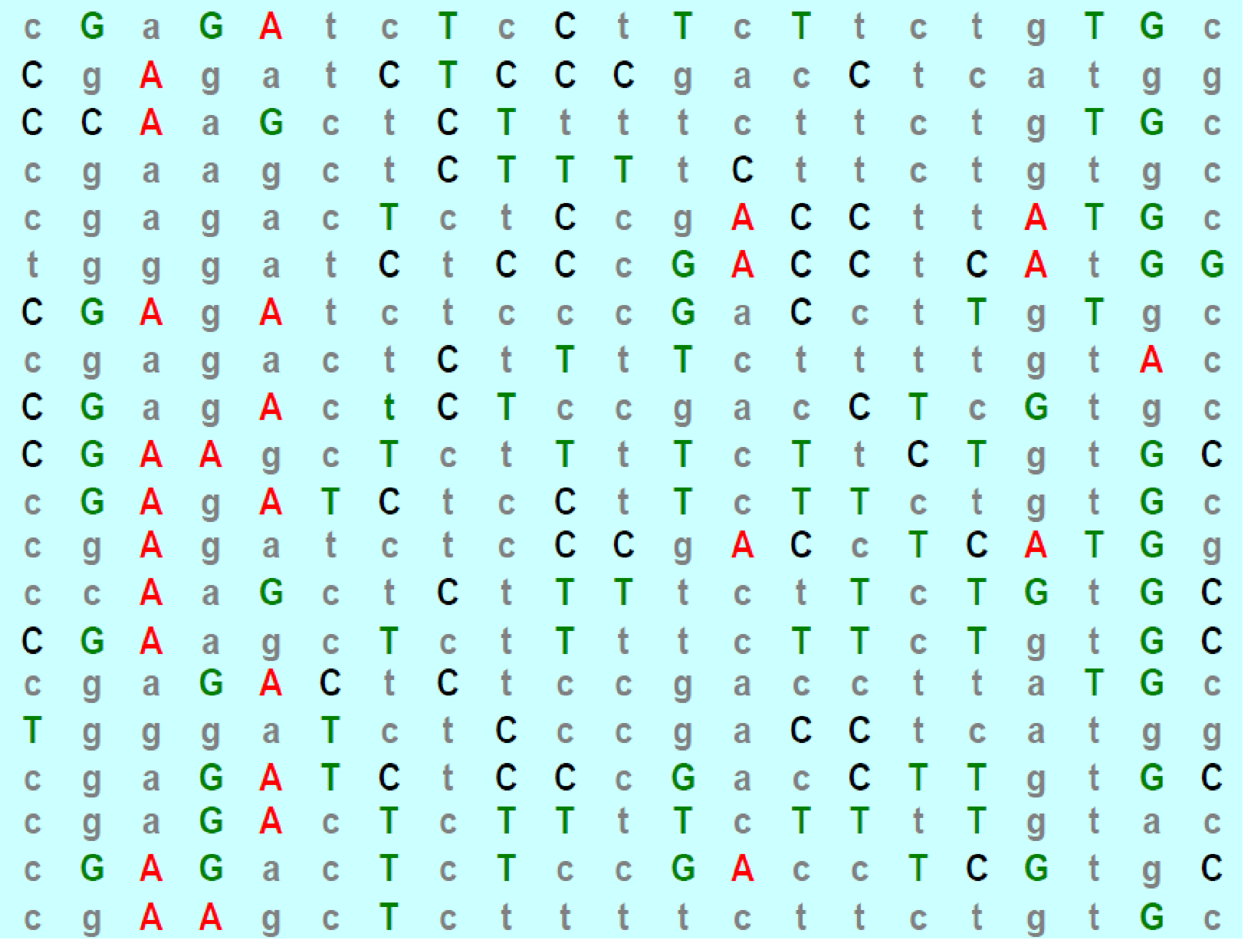
Cartoon View of Low Coverage NGS Data



.	G	.	G	A	.	.	T	.	C	.	T	.	T	.	.	.	.	T	G	.
C	.	A	.	.	.	C	T	C	C	C	.	.	.	C	.	.	.	.	.	.
C	C	A	.	G	.	.	C	T	.	.	.	.	.	.	.	.	.	T	G	.
.	.	.	.	.	.	.	C	T	T	T	.	C	.	.	.	.	.	.	.	.
.	.	.	.	.	.	T	.	.	C	.	.	A	C	C	.	.	A	T	G	.
.	.	.	.	.	.	C	.	C	C	.	G	A	C	C	.	C	A	.	G	G
C	G	A	.	A	.	.	.	.	.	.	G	.	C	.	.	T	.	T	.	.
.	.	.	.	.	.	.	C	.	T	.	T	.	.	.	.	.	.	.	A	.
C	G	.	.	A	.	.	C	T	.	.	.	.	.	C	T	.	G	.	.	.
C	G	A	A	.	.	T	.	.	T	.	T	.	T	.	C	T	.	.	G	C
.	G	A	.	A	T	C	.	.	C	.	T	.	T	T	.	.	.	.	G	.
.	.	A	.	.	.	.	.	.	C	C	.	A	C	.	T	C	A	T	G	.
.	.	A	.	G	.	.	C	.	T	T	.	.	.	T	.	T	G	.	G	C
C	G	A	.	.	.	T	.	.	T	.	.	.	T	T	.	T	.	.	G	C
.	.	.	G	A	C	.	C	.	.	.	.	.	.	.	.	.	.	T	G	.
T	.	.	.	.	T	.	.	C	.	.	.	.	C	C	.	.	.	.	.	.
.	.	.	G	A	T	C	.	C	C	.	G	.	.	C	T	T	.	.	G	C
.	.	.	G	A	.	T	.	T	T	.	T	.	T	T	.	T	.	.	.	.
.	G	A	G	.	.	T	.	T	.	.	G	A	.	.	T	C	G	.	.	C
.	.	A	A	.	.	T	.	.	.	.	.	.	.	.	.	.	.	.	G	.

Because of low coverage, alleles are only reliably called for some polymorphic sites in each individual, resulting in missing data in the remaining sites.

Cartoon View of Low Coverage NGS Data



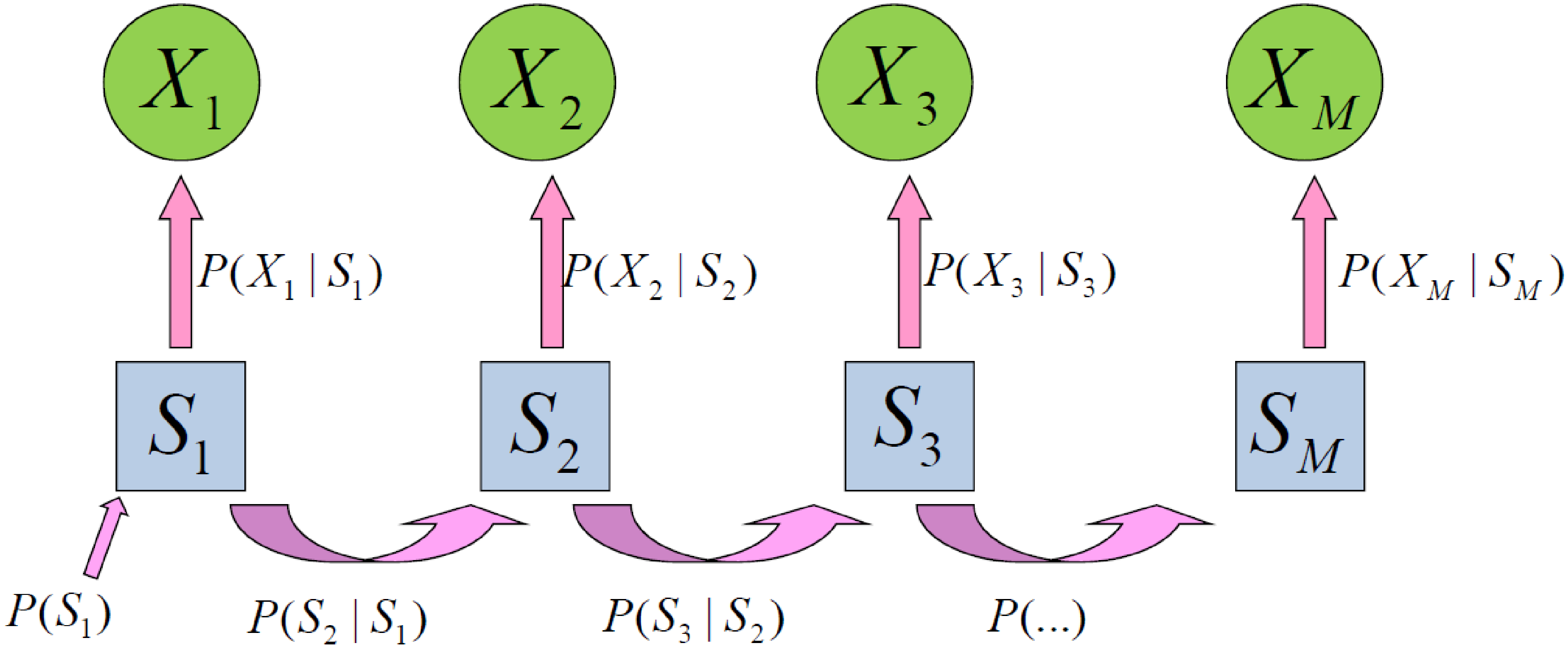
c	G	a	G	A	t	c	T	c	C	t	T	c	T	t	c	t	g	T	G	c
C	g	A	g	a	t	C	T	C	C	C	g	a	c	C	t	c	a	t	g	g
C	C	A	a	G	c	t	C	T	t	t	t	c	t	t	c	t	g	T	G	c
c	g	a	a	g	c	t	C	T	T	T	t	C	t	t	c	t	g	t	g	c
c	g	a	g	a	c	T	c	t	C	c	g	A	C	C	t	t	A	T	G	c
t	g	g	g	a	t	C	t	C	C	c	G	A	C	C	t	C	A	t	G	G
C	G	A	g	A	t	c	t	c	c	c	G	a	C	c	t	T	g	T	g	c
c	g	a	g	a	c	t	C	t	T	t	T	c	t	t	t	t	g	t	A	c
C	G	a	g	A	c	t	C	T	c	c	g	a	c	C	T	c	G	t	g	c
C	G	A	A	g	c	T	c	t	T	t	T	c	T	t	C	T	g	t	G	C
c	G	A	g	A	T	C	t	c	C	t	T	c	T	T	c	t	g	t	G	c
c	g	A	g	a	t	c	t	c	C	C	g	A	C	c	T	C	A	T	G	g
c	c	A	a	G	c	t	C	t	T	T	t	c	t	T	c	T	G	t	G	C
C	G	A	a	g	c	T	c	t	T	t	t	c	T	T	c	T	g	t	G	C
c	g	a	G	A	C	t	C	t	c	c	g	a	c	c	t	t	a	T	G	c
T	g	g	g	a	T	c	t	C	c	c	g	a	C	C	t	c	a	t	g	g
c	g	a	G	A	T	C	t	C	C	c	G	a	c	C	T	T	g	t	G	C
c	g	a	G	A	c	T	c	T	T	t	T	c	T	T	t	T	g	t	a	c
c	G	A	G	a	c	T	c	T	c	c	G	A	c	c	T	C	G	t	g	C
c	g	A	A	g	c	T	c	t	t	t	t	c	t	t	c	t	g	t	G	c

But borrowing information from other individuals, we can infer missing alleles in the remaining sites.

Recipe for Imputation with NGS Data

- Start with some plausible configuration for each individual
- Use Markov model to update one individual conditional on all other individuals
- Repeat previous step many times
- Generate a consensus set of genotypes and haplotypes for each individual

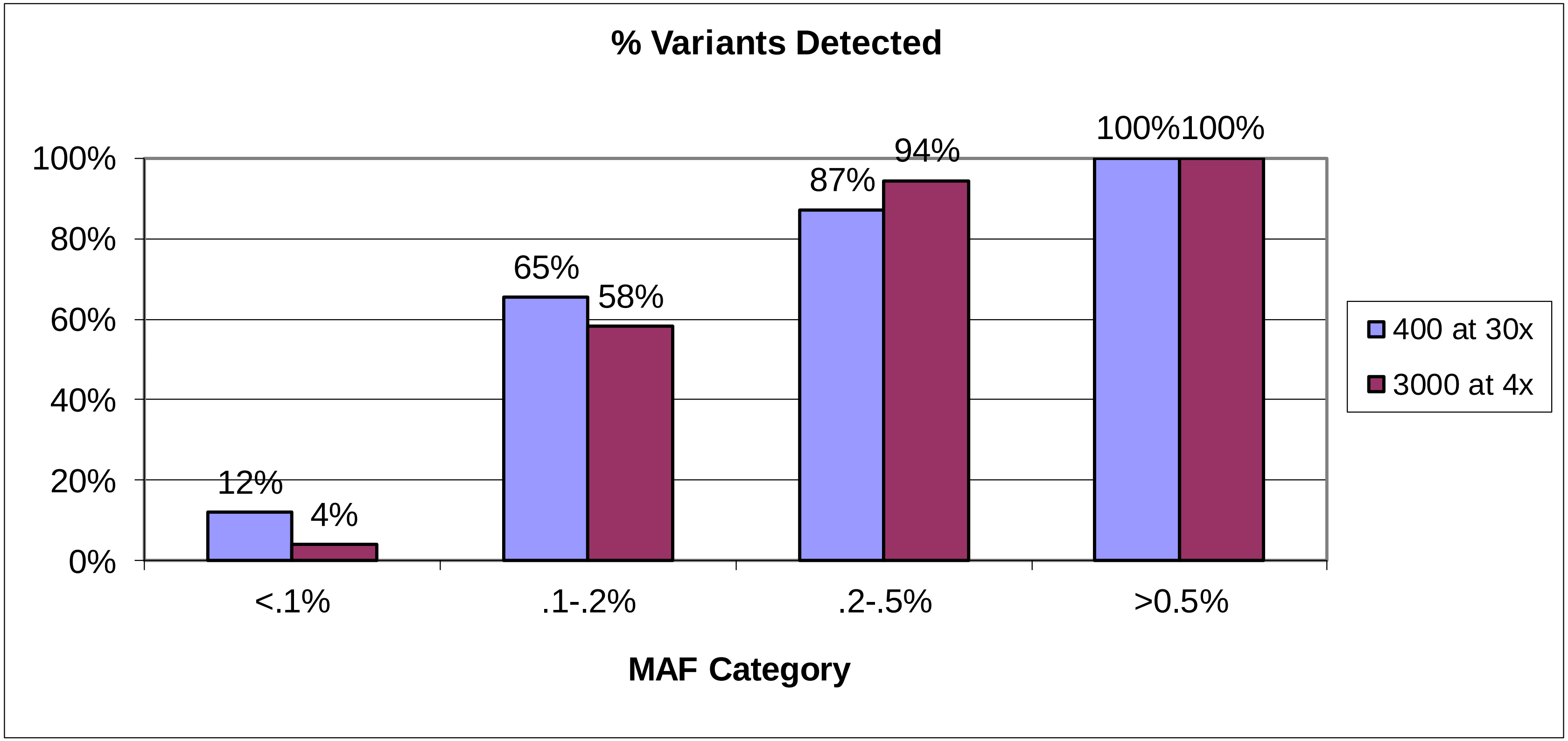
Hidden Markov Model for Genotype Imputation



Comparison Summary

- Multi-sample callers better performance than single-sample callers
  - More variants detected
  - Better genotype calling quality

High vs Low Coverage Sequencing



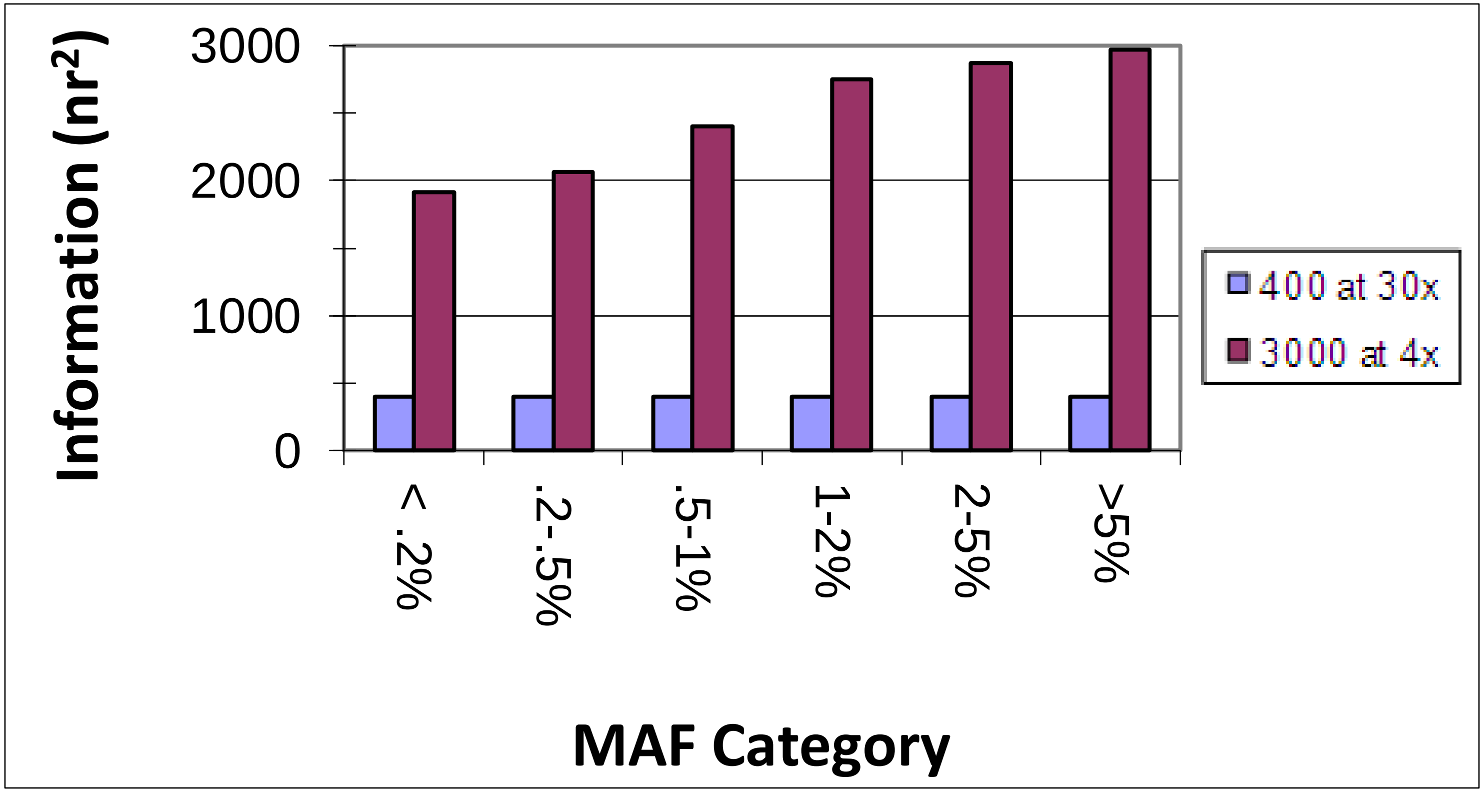
% is out of the total number of polymorphisms in the population

Li et al. (2011) Genome Research

Comparison Summary: Variant Detection

- Both designs had near 100% power to detect variants with MAF>0.5%
- Low-depth design provided greater power to detect less common variants with MAF 0.2-0.5%
- Neither design had much chance to detect the rarest SNPs (MAF<0.1%)
    - For high-depth designs, the minor allele for rare SNPs was often absent in the sequenced sample.
    - For low-depth designs, it was not possible to distinguish true variants from sequencing errors confidently with a small number of reads carrying the alternative allele.

High vs Low Coverage Sequencing



Information (nr^2), measuring the effective sample size of a study, is directly related to power for association analysis.
n is the number of individuals sequenced; r^2 is squared Pearson correlation between called and true genotypes.

Li et al. (2011) Genome Research

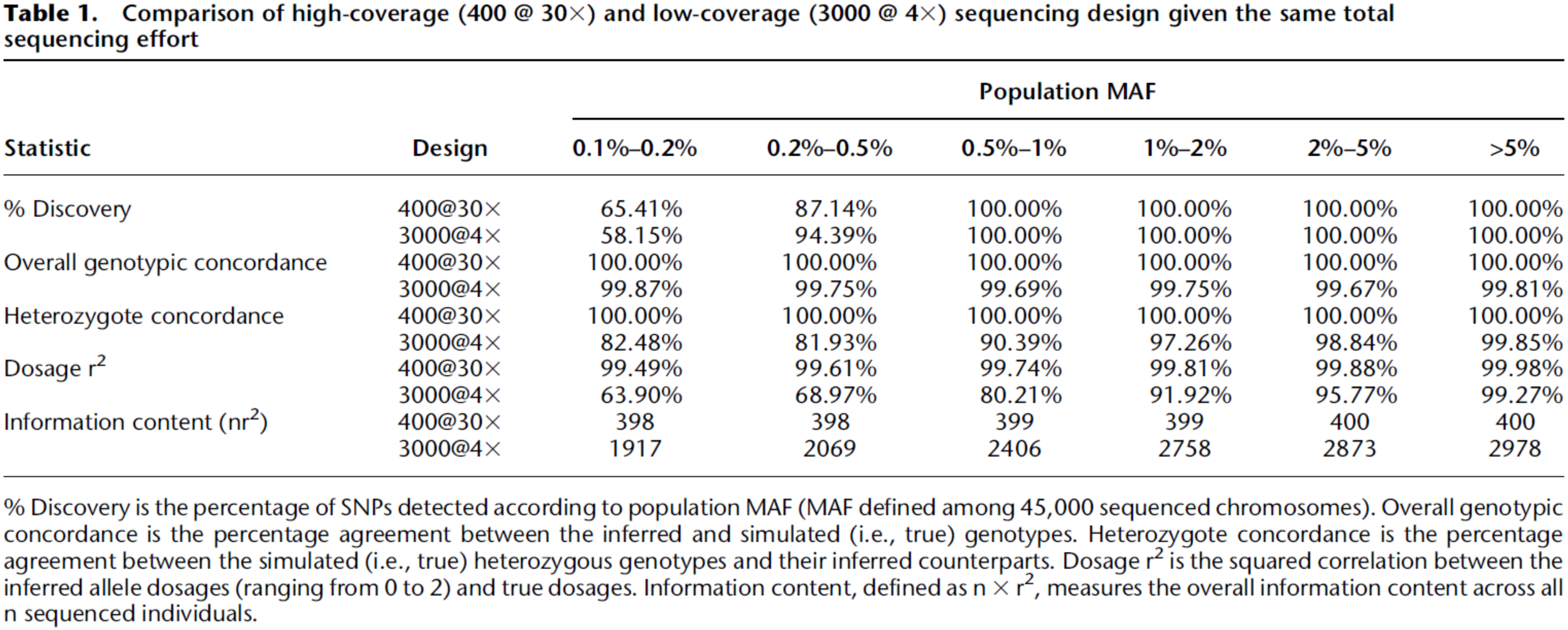


Table 1. Comparison of high-coverage (400 @ 30×) and low-coverage (3000 @ 4×) sequencing design given the same total sequencing effort

		Population MAF					
Statistic	**Design**	**0.1%–0.2%**	**0.2%–0.5%**	**0.5%–1%**	**1%–2%**	**2%–5%**	**>5%**
% Discovery	400@30×	65.41%	87.14%	100.00%	100.00%	100.00%	100.00%
	3000@4×	58.15%	94.39%	100.00%	100.00%	100.00%	100.00%
Overall genotypic concordance	400@30×	100.00%	100.00%	100.00%	100.00%	100.00%	100.00%
	3000@4×	99.87%	99.75%	99.69%	99.75%	99.67%	99.81%
Heterozygote concordance	400@30×	100.00%	100.00%	100.00%	100.00%	100.00%	100.00%
	3000@4×	82.48%	81.93%	90.39%	97.26%	98.84%	99.85%
Dosage r^2	400@30×	99.49%	99.61%	99.74%	99.81%	99.88%	99.98%
	3000@4×	63.90%	68.97%	80.21%	91.92%	95.77%	99.27%
Information content (nr^2)	400@30×	398	398	399	399	400	400
	3000@4×	1917	2069	2406	2758	2873	2978

% Discovery is the percentage of SNPs detected according to population MAF (MAF defined among 45,000 sequenced chromosomes). Overall genotypic concordance is the percentage agreement between the inferred and simulated (i.e., true) genotypes. Heterozygote concordance is the percentage agreement between the simulated (i.e., true) heterozygous genotypes and their inferred counterparts. Dosage r^2 is the squared correlation between the inferred allele dosages (ranging from 0 to 2) and true dosages. Information content, defined as $n \times r^2$, measures the overall information content across all n sequenced individuals.

Li et al. (2011) Genome Research

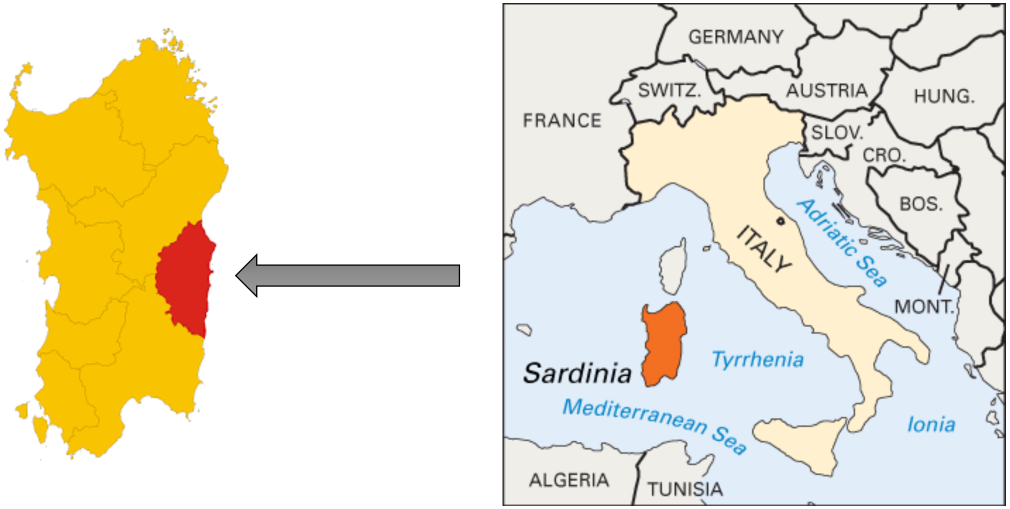
Comparison Summary: Variant Detection

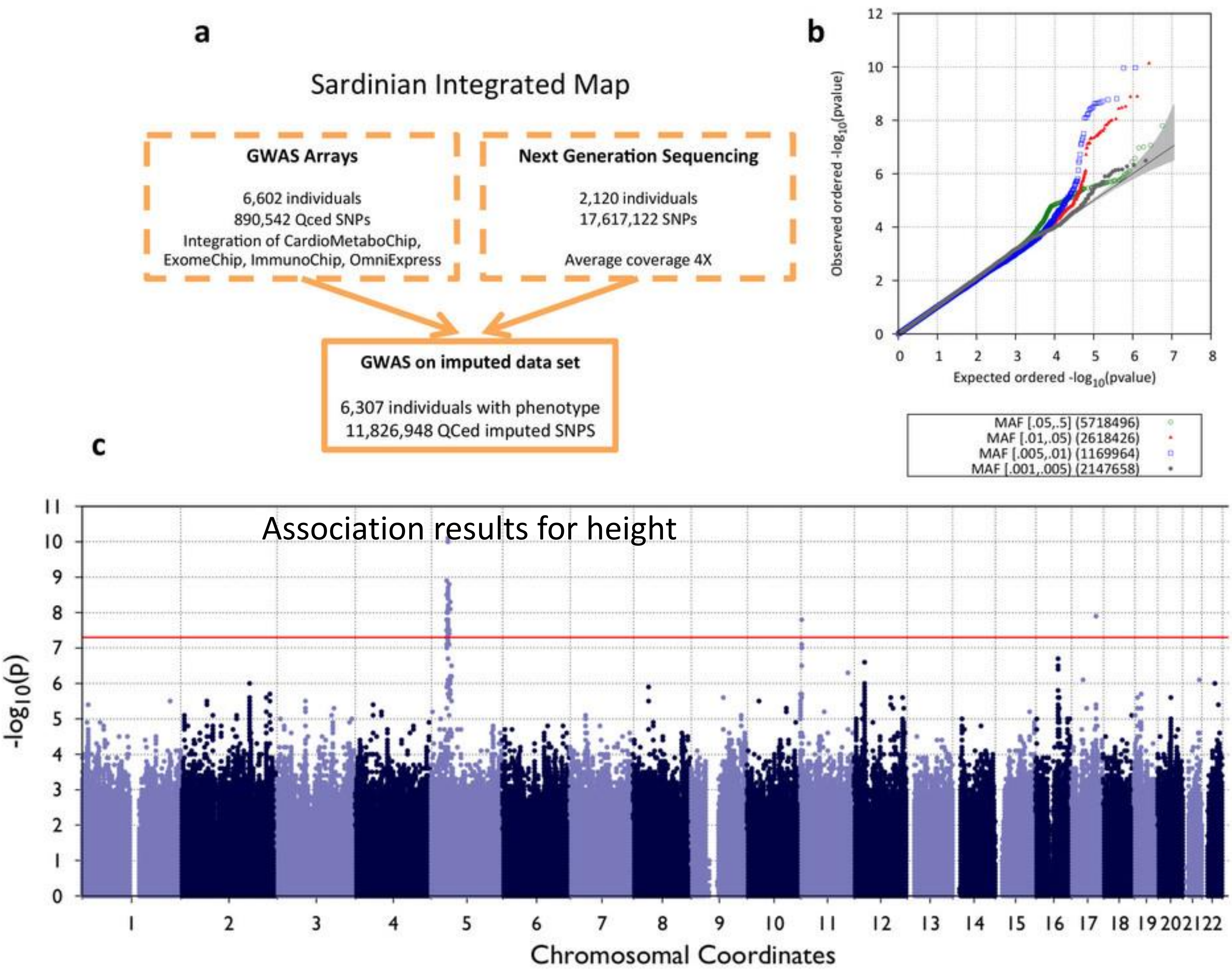
- At detected variants, genotype accuracy was reduced for low-depth compared to high-depth designs but was still impressive
    - For example, for variants with MAF>1%, the genotypic concordance, albeit not 100% as in the high-depth design, is always >99.67% and concordance at heterozygous sites >97%.

- Thus, low-depth designs substantially increase the overall information content (genotypes are individually not as good but, in aggregate, contain more information), holding the overall sequencing investment constant.

The SardiNIA Project

- **SardiNIA Whole Genome Sequencing Study**
  - ~7000 Sardinians from Sardinia, Italy
  - Recruited among population of ~60,000 individuals (founder, homogeneous population)
  - Sample includes >34,000 relative pairs
- Measured ~100 aging related quantitative traits
- Aim to sequence ~2,000 individuals at 2x to obtain genomes, and then genotype all individuals, impute sequences into relatives





Sardinian Integrated Map

GWAS Arrays

6,602 individuals
890,542 Qced SNPs
Integration of CardioMetaboChip,
ExomeChip, ImmunoChip, OmniExpress

Next Generation Sequencing

2,120 individuals
17,617,122 SNPs

Average coverage 4X

GWAS on imputed data set

6,307 individuals with phenotype
11,826,948 QCed imputed SNPS