Sequence Variants and Phenotype

DAVID L. TABB, PH.D.

Overview

- Technologies for detecting genetic variants: WGS, WES, microarray
- Estimating phenotypic impact of nsSNVs
- Genome Wide Association Studies: admixture and linkage

Whole Genome Shotgun and mapping

UNIVERSITEIT STELLENBOSCH UNIVERSITY

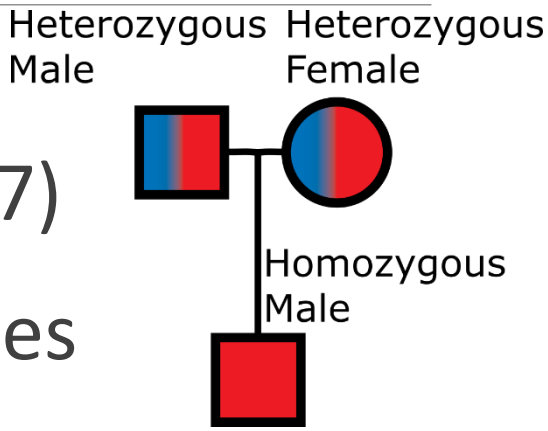
- Sequencing a patient's complete genome *may* inform clinical decisions.
- Short reads are mapped to annotation.
- An individual will differ from reference at more than *two million* different SNVs.
- Desired genes may be low in coverage.
- Different sequencing experiments detect different variants, particularly INDELs.

Single nucleotide variants

Insertions / Deletions

FE Dewey et al. *JAMA* (2014) 311: 1035-1044.

Sequencing triads in rare disease



- Non-transmitted alleles serve as control: Falk, Rubenstein (1987)
- Preferential transmission of alleles among triads: Spielman (1993)
- *Phasing* determines *haplotypes*: which sets of variants come from each parent.
- In all cases, parents are point of comparison for interpreting offspring genotype.

MP Epstein et al. *Am. J. Hum. Genet*. (2005) 76: 592-608.

VarScan 2: detecting variants

How can we discern legitimate sequence variants from sequencing errors?

- Sequence coverage should be high at that position (few reads → error more likely).
- The basecalls declaring the change should have reliable Phred scores.
- Many reads must attest to the variant allele.
- A statistical test establishes significance.

http://dkoboldt.github.io/varscan/

DC Koboldt et al. *Genome Research* (2012) 22: 568-576

Variant Call Format (VCF/BCF)

```
##fileformat=VCFv4.3
##fileDate=20090805
##reference=file:///seq/references/1000GenomesPilot-NCBI36.fasta
##contig=<ID=20,length=62435964,assembly=B36,md5=f126cdf8a6e0c7f379d618ff66beb2da,species="Homo sapiens",taxonomy=x>
#CHROM POS     ID        REF ALT    QUAL FILTER FORMAT      NA00001        NA00002        NA00003
20     14370   rs6054257 G   A      29   PASS   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    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   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   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   GT:GQ:DP    0/1:35:4       0/2:17:2       1/1:40:3
```

omitted INFO sample 1 sample 2 sample 3

"generic format for storing DNA polymorphism data such as SNPs, insertions, deletions and structural variants, together with rich annotations."

http://samtools.github.io/hts-specs/

P. Dancek et al. *Bioinformatics* (2011) 27: 2156-2158

Databases of known variants

ACMG prefers

Not always benign

Single Nucleotide Polymorphism: DNA variant detectable in >1% of population

Not always pathogenic

Mutation (somatic or germline): DNA variant detectable in <1% of population

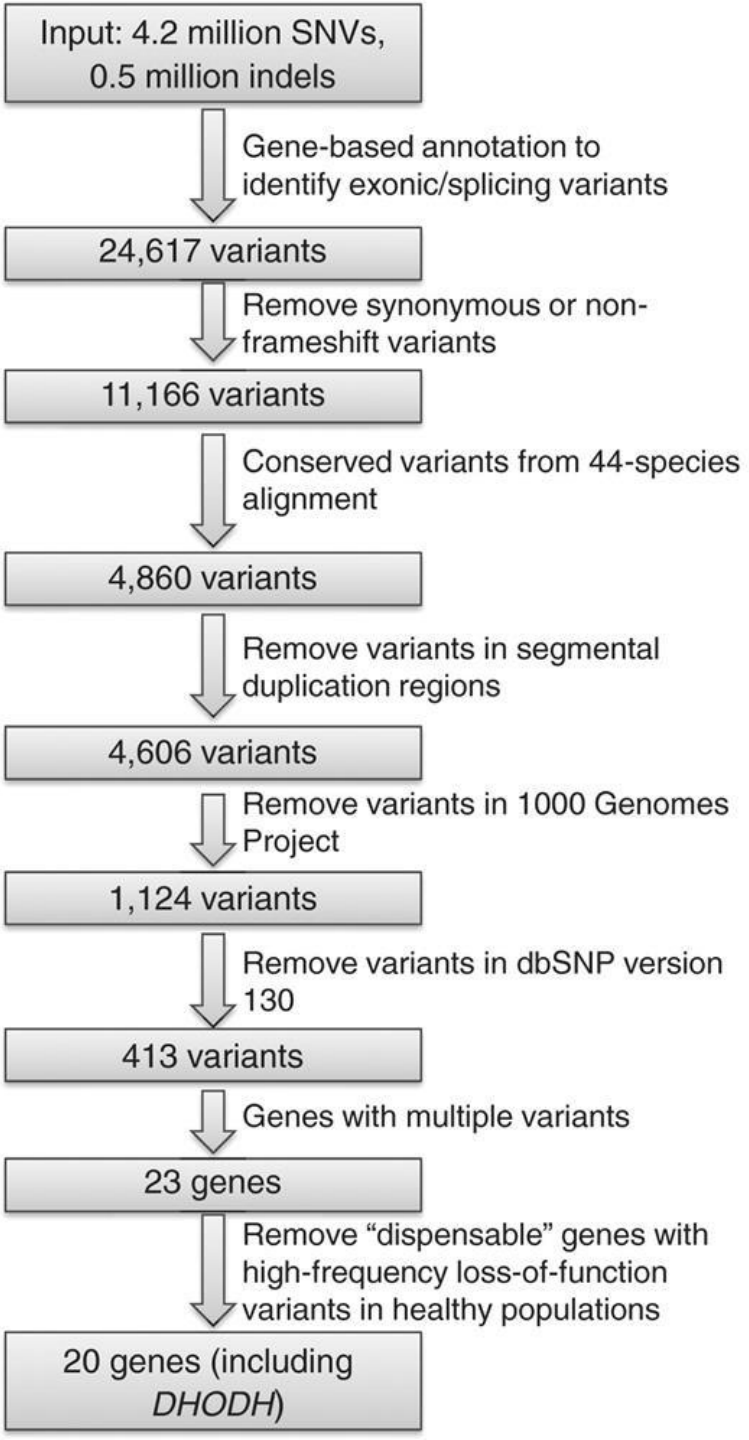
- dbSNP (incl HapMap)
- 1000 Genomes
- Exome Sequencing Project
- Exome Aggregation Consortium (ExAC)
- gnomAd: genome aggregation database

R. Karki et al. *BMC Med Genomics* (2015) 8: 37

Annovar: annotate variants

Given list of variant positions and observed vs. reference,

- Detect protein coding changes
- Recognize variants falling within a genomic region
- Check for presence in databases
- Develop candidate gene lists for Mendelian diseases



K Wang et al. *Nucl. Acids Res*. (2010) 38: e164

UNIVERSITEIT STELLENBOSCH UNIVERSITY

TAPER: from variants of unknown significance to targets

Step 1: Variant annotation using ANNOVAR

↓

Step 2: Removal of all synonymous

↓

Step 3: Removal of all variants in 1000 Genomes Project with a frequency greater than 1%

↓

Step 4: Removal of all variants in the Genome Aggregation Database with a frequency greater than 1%

↓

Step 5: Remove all variants with Genomic Evolutionary Rate Prediction scores less than 0.

↓

Step 6: Remove variants with a FATHMM score of greater than 1.0.

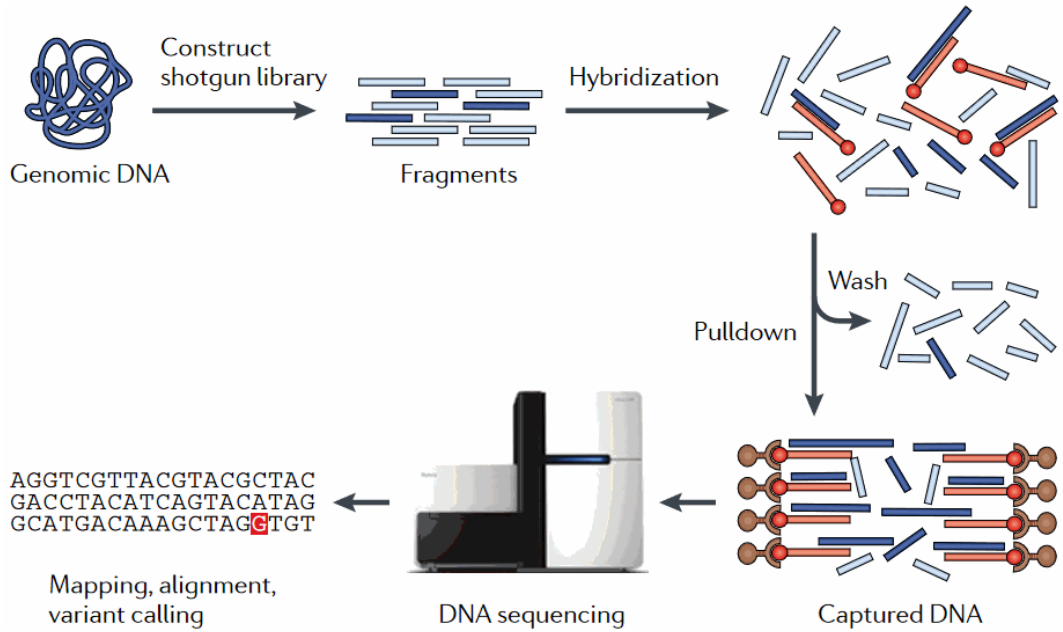
↓

Step 7: Identification of related disorders for prioritized genes.

B Glanzmann et al. *Source Code Biol Med* (2016) 11: 10

Why exome sequencing?

- Focuses sequencer on features of interest.
- More read depth improves sensitivity to variant presence.
- Variants are easier to associate with function.



 MJ Bamshad et al. *Nature Rev. Genet*. (2011) 12: 745-755.

Variability in exome tech

- Enrichment kits target different gene sets or exclude different intronic regions.
- miRNA, promoters, and ultraconserved elements are frequently omitted.
- Probe efficiency varies, and not all sequences map to all annotations (e.g. RefSeq versus Ensembl).

MJ Clark et al. *Nat. Biotech*. (2011) 29: 908-914

AmpliSeq community panels

- SUN Central Analytical Facilities employ Thermo Ion Proton and Ion S5 sequencers.
- Many studies can narrow focus to hundreds rather than thousands of genes; Illumina has a variety of disease-focused panels available.
- Local study of 47 Parkinson's Disease patients found 54 likely deleterious variants.

OG Oluwole et al. *BMC Med. Genet*. (2020) 21: 23.

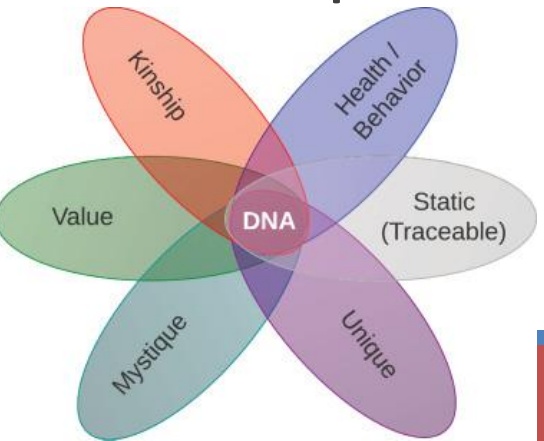
Microarrays are not dead. Integrating with sequencing helps!

- Affymetrix, now part of Thermo Fisher, sells photolithographic Axiom arrays, ranging up to 2.6M markers for AFR, AMR, EAS, EUR, and SAS populations.
- Illumina produced a “bead array” to measure genetic diversity for African populations, using 2.5M markers.
- Sequencing a genome: <$1000 (USD).
  Arraying genomic variants: <$100 (USD).

Delaneau et al. *Nat. Comm*. (2014) 5: 3934.

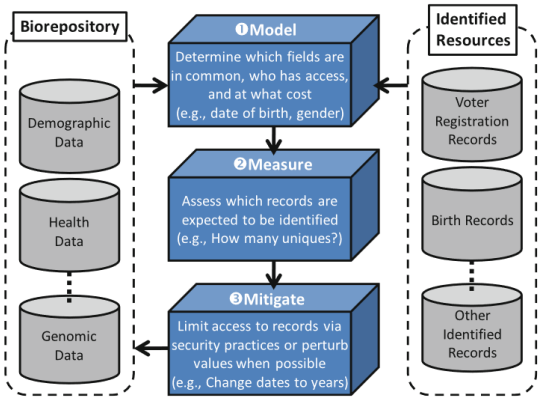
Why we must take care in publishing DNA variants

- DNA predicts aspects of a person's health.
- A person's genome does not change much.
- A genome is usually unique to a person.
- DNA evidence has special standing in the public.
- Sequencing is costly, and data are seen as valuable.
- Biological kinship can be inferred from DNA.



M. Naveed et al. *ACM Comput. Surveys* (2015) 48: 6

Isn't it enough to remove names?



- *Identifiability*: the degree to which materials stored in biobanks can be linked to the name of the individuals from which they were derived.
- "Only about 100 SNPs are required to distinguish an individual's DNA record"
- Reporting only aggregate data (among groups rather than for individuals) offers some protection.
- Nations enact laws to protect patient privacy. HIPAA (USA), GDPR (EU), POPI (ZA)

B Malin et al. *Hum Genet* (2011) 130: 383-392.

L Sweeney. *J Law, Med, and Ethics* (1997): 98-110.

Estimating phenotypic impact

Taxonomy of sequence variants

- Novel, noncoding: some experiments required
- Known, noncoding: possible eQTL association
- In-frame INDELs: gain or loss of AA
- Mis-sense SNVs: non-synonymous AA change
- Frameshifts, splicing, gained stops:
  protein C-terminus abnormal

INDEL: an insertion or deletion in a sequence
non-synonymous: altering the codon to another AA

MJS Dashti and J Gamieldien. *BioTechniques* (2017) 62: 18-30.

An nsSNV changes the resulting polypeptide

- Not all nucleotide sequence changes will alter the amino acid.
- Only changes within ORF can be non-synonymous.
- INDELs and mis-sense variants also change sequence.

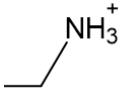
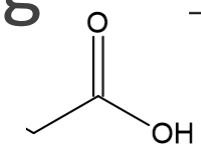
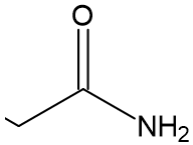
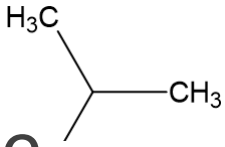
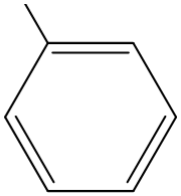
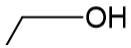
Second letter

First letter	U	C	A	G	Third letter
U	UUU, UUC Phe UUA, UUG Leu	UCU, UCC, UCA, UCG Ser	UAU, UAC Tyr **UAA Stop** **UAG Stop**	UGU, UGC Cys **UGA Stop** UGG Trp	U C A G
C	CUU, CUC, CUA, CUG Leu	CCU, CCC, CCA, CCG Pro	CAU, CAC His CAA, CAG Gln	CGU, CGC, CGA, CGG Arg	U C A G
A	AUU, AUC, AUA Ile **AUG Met**	ACU, ACC, ACA, ACG Thr	AAU, AAC Asn AAA, AAG Lys	AGU, AGC Ser AGA, AGG Arg	U C A G
G	GUU, GUC, GUA, GUG Val	GCU, GCC, GCA, GCG Ala	GAU, GAC Asp GAA, GAG Glu	GGU, GGC, GGA, GGG Gly	U C A G

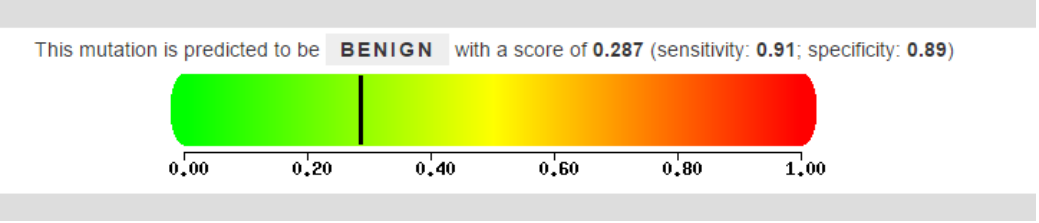
OpenStax College, Biology

Katsonis et al. *Protein Science* (2014) 23: 1650-1666.

Sets of amino acids feature similar biochemistry

- Basic: His Lys Arg 
- Acidic: Asp Glu 
- Amide: Asn Gln 
- Nonpolar hydrophobic: Ala Val Leu Ile 
- Aromatics: Tyr Trp Phe 
- Hydroxyl: Ser Thr 
- Wildcards: Gly Pro

P Katsonis et al. *Protein Sci*. (2014) 23: 1650-1666.



UNIVERSITEIT STELLENBOSCH UNIVERSITY

Is an nsSNV deleterious?

- SIFT (2001) retrieves sequences similar to query, aligns them, and computes probability for substitution.
- Polyphen (2002) adds 3D structure and feature table to phenotype assessment.
- SNAP (2007) adds solvent accessibility and machine learning to calibrate scores.

Ng and Henikoff. *Genome Research* (2001) 11: 863-874.
Ramensky et al. *Nucl. Acids Res*. (2002) 30: 3894-3900.
Bromberg and Rost. *Nucl. Acids Res*. (2007) 35: 3823-3835.

Threading: perturbing a known structure with AA changes

When a structure has already been determined for a closely related sequence, one can estimate the structure for a query sequence by *comparative modelling* or *threading:*

1. Identify structures for related sequences.
2. Align the sequence to the template structure.
3. Build a model reflecting the altered side chains.
4. Assess solvent-accessible surface area for model.

Pieper et al. *Nucl. Acids Res*. (2011) 39: D465-D474 Kroncke et al. *Biochemistry* (2015) 54: 2551

Establishing Linkage via GWAS

Linkage rather than direct impact

- Earlier studies sought genetic markers commonly found in patients, often SNPs.
- If a marker was very close to the risk-conferring DNA, recombination would rarely separate the two.
- How would one go about determining the gene or regulatory element responsible?

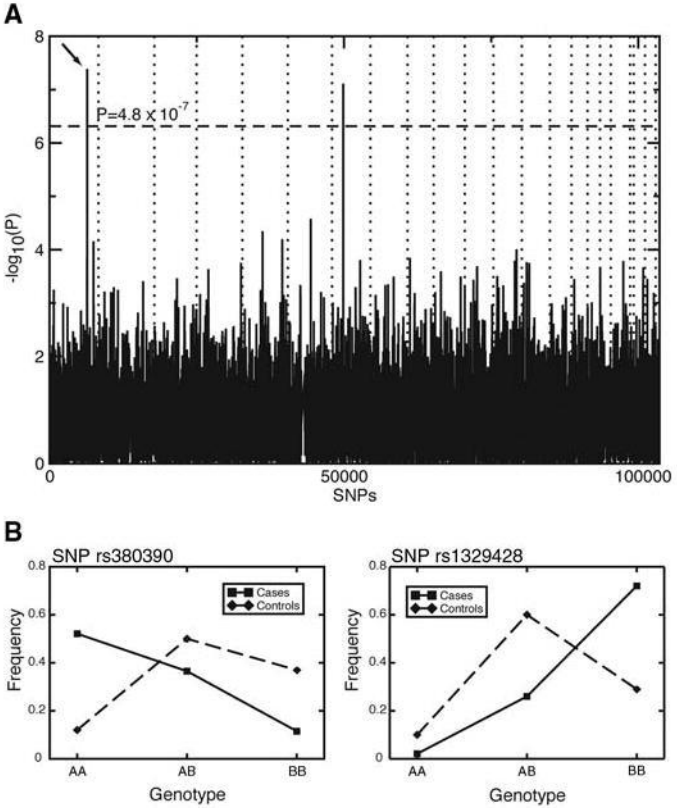
Genome-Wide Association Studies

- If we could amass a hundred patients with a disease and a hundred without, could we find any genetic markers for disease risk?
- Klein *et al* sought markers of age-related macular degeneration with 96 cases and 50 controls, using microarrays measuring 116,204 SNPs. They found two associations.
- The floodgates were opened!

RJ Klein et al. *Science* (2005) 308: 385-389

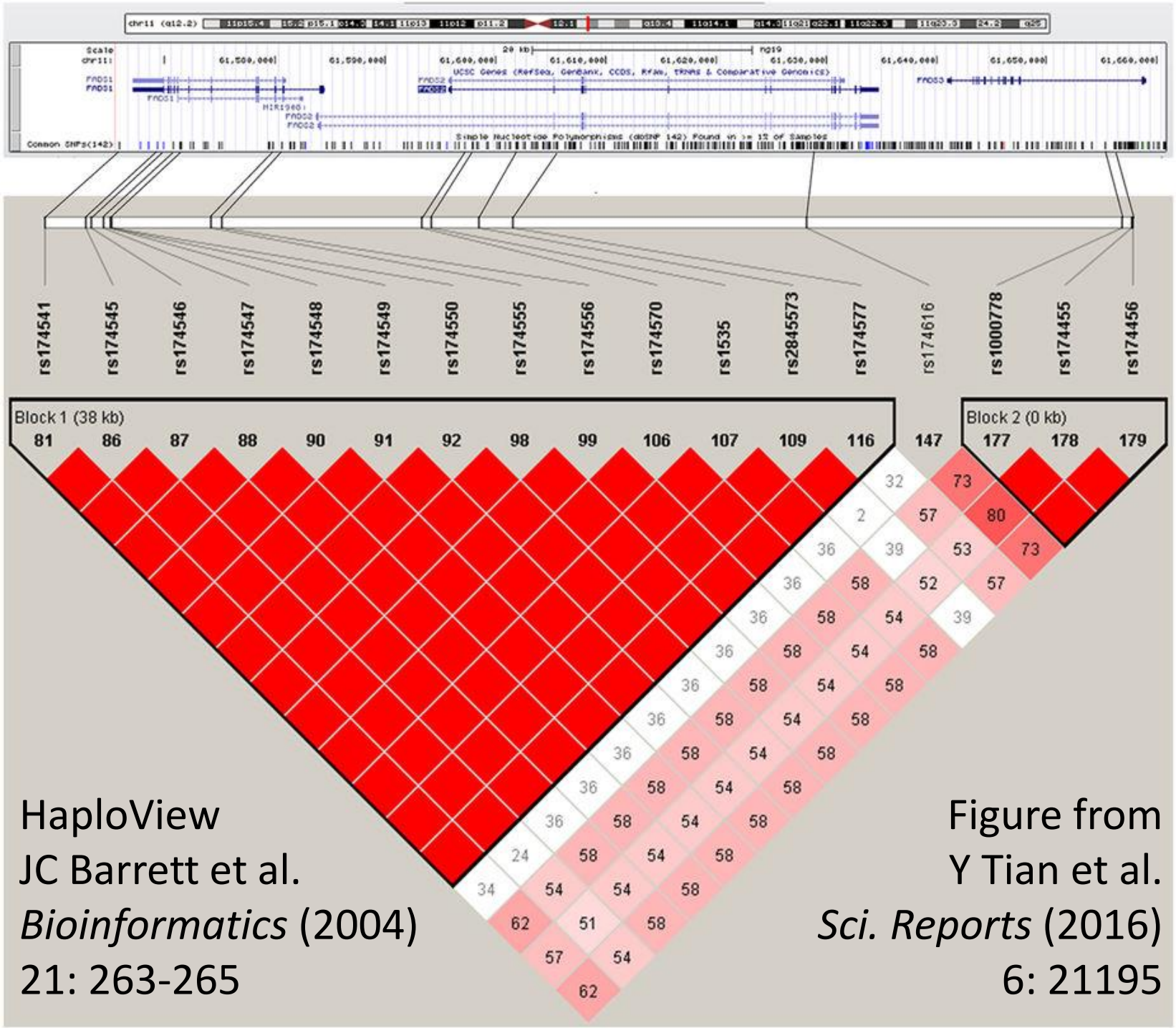
The Manhattan plot

- SNPs arrayed across x-axis.
- Height is –log of p-value.
- Correcting for multiple tests requires $p < \frac{0.05}{103{,}611}$ for a "hit" to protect against any false hits. (Bonferroni)



RJ Klein et al. *Science* (2005) 308: 385-389

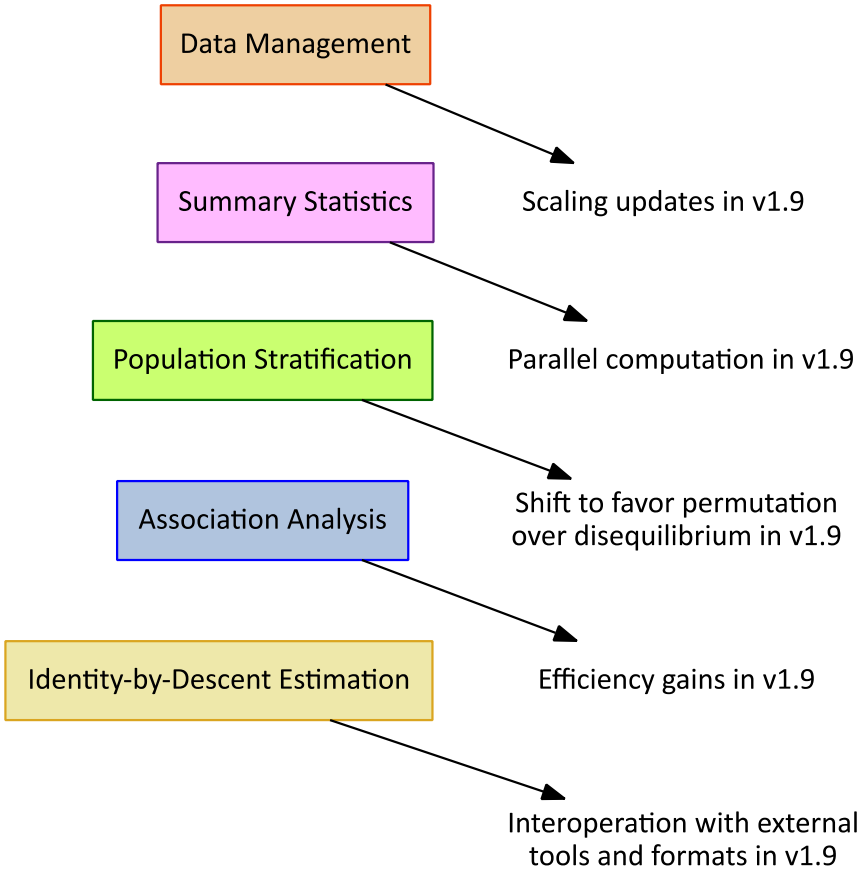
Genetic linkage



HaploView
JC Barrett et al.
Bioinformatics (2004)
21: 263-265

Figure from
Y Tian et al.
Sci. Reports (2016)
6: 21195

PLINK: C++ software for GWAS data analysis



- Open-source tool for Linux, Windows, and Mac OSX. GUI operation included in update.
- Designed for sets where particular markers are measured in many thousands of individuals.

Purcell et al. *Am. J. Hum. Genet*. (2007) 81: 559-575

GWAS Catalog

http://www.ebi.ac.uk/gwas/home

What is *stratification*?

Population stratification (or population structure) is "systematic ancestry differences between cases and controls."

If cases are 60% from Han Chinese ancestry while controls are only 40% Han, GWAS may mislabel Han markers as indicative of disease.

Ancestry Informative Markers measure sites that are associated with particular ancestry.

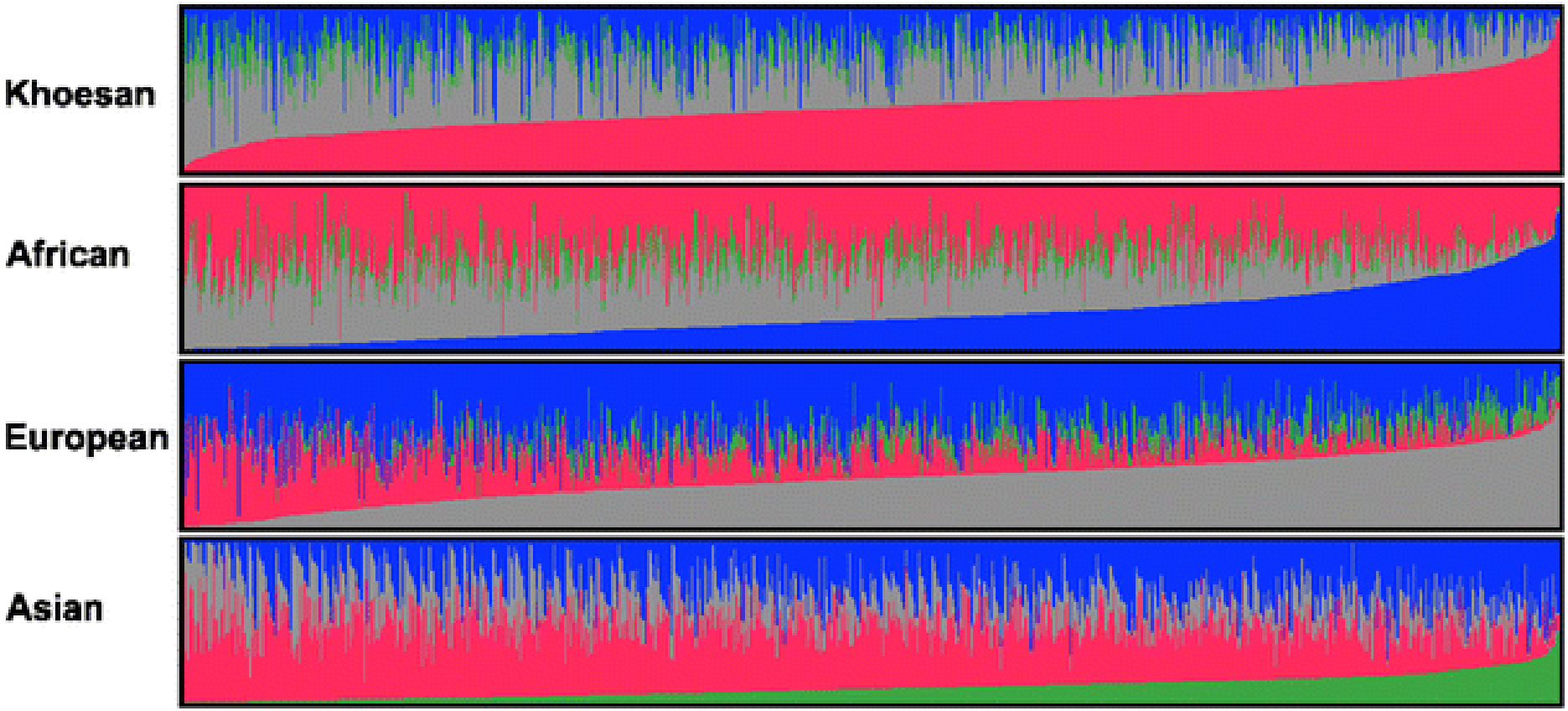
Price et al. *Nat. Reviews Genet*. (2010) 11: 459-463.

Admixture: GWAS in populations with mixed ancestry

- "Populations that came about by the mixing of two or more distant continental populations" in recent history exhibit:
  - Fine scale: correlation among nearby SNPs
  - Segmental scale: syntenic SNPs have not been separated by recombination
- We must estimate each individual's ancestry before we seek disease marker associations.

Pasaniuc et al. *PLOS Genetics*. (2011) pgen.1001371

South African Coloured group reflects worldwide ancestry



http://web.stanford.edu/group/pritchardlab/structure.html

de Wit et al. *Human Genetics* (2010) 128: 145-153

Takeaway messages

- Genetic variation may be measured through whole genome shotgun, whole exome, or microarray / bead array technologies.
- nsSNVs draw particular attention due to recognizable effect on protein sequences.
- Linkage correlates a variant to phenotype, but it can't claim variant *causes* phenotype.