Analysis of Genome Wide Association Studies (GWAS) Lecture 20

David K. Gifford

Massachusetts Institute of Technology

Today's Narrative Arc

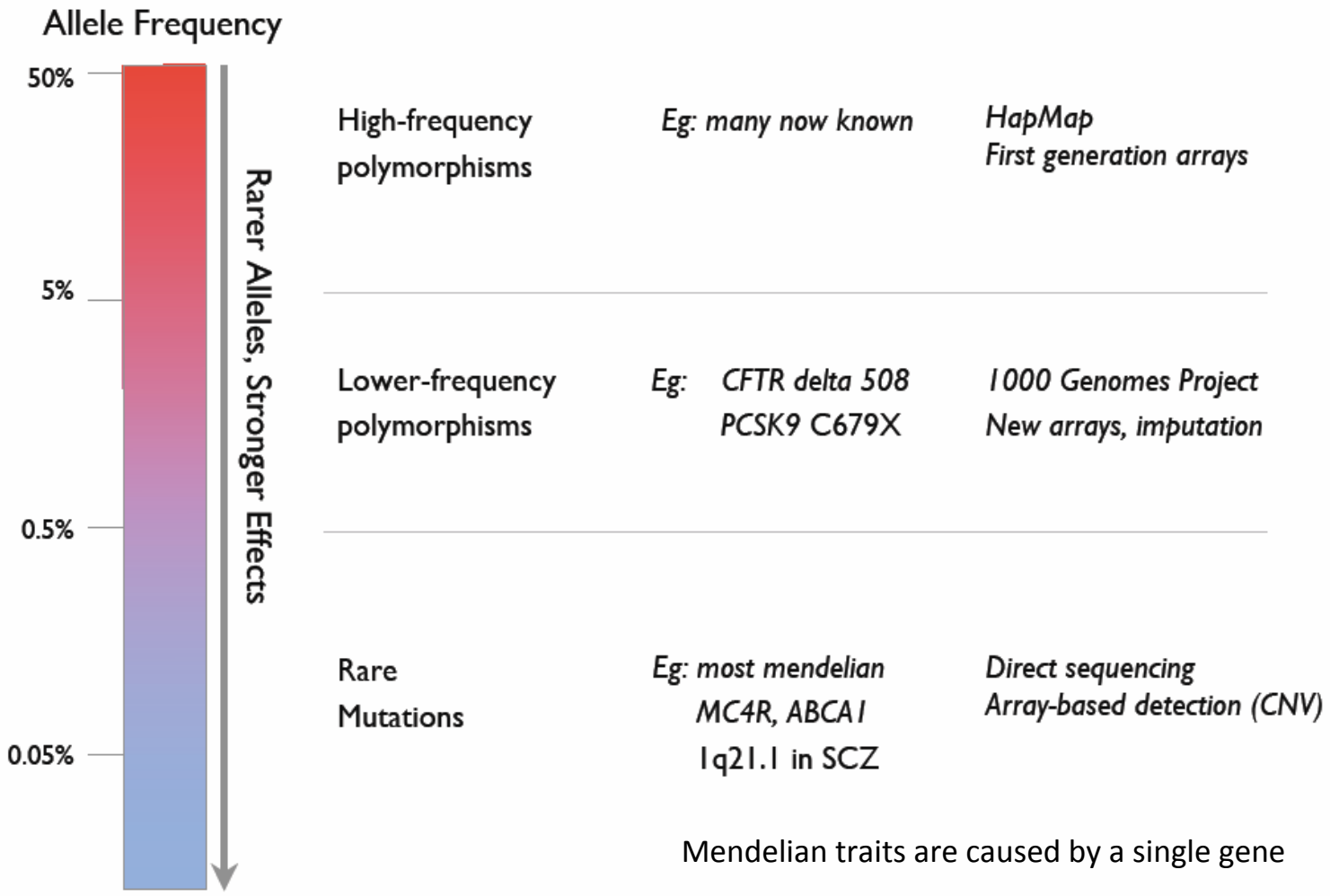
1. We can discover human variants that are associated with a phenotype by studying the genotypes of case and control populations
   - **Approach 1 – Use allelic counts from SNP arrays (SNPs called from microarray data)**
   - Approach 2 – Use read counts from sequencing (multiple reads per variant per individual)
2. We can prioritize variants based upon their estimated importance
3. Follow up confirmation is important because correlation is not equivalent to causality

Today's Computational Approaches

1. Contingency tables for allelic association tests and genotypic association tests.
2. Methods of testing - Chi-Square tests, Fisher's exact test
3. Likelihood based tests of case/control posterior genotypes

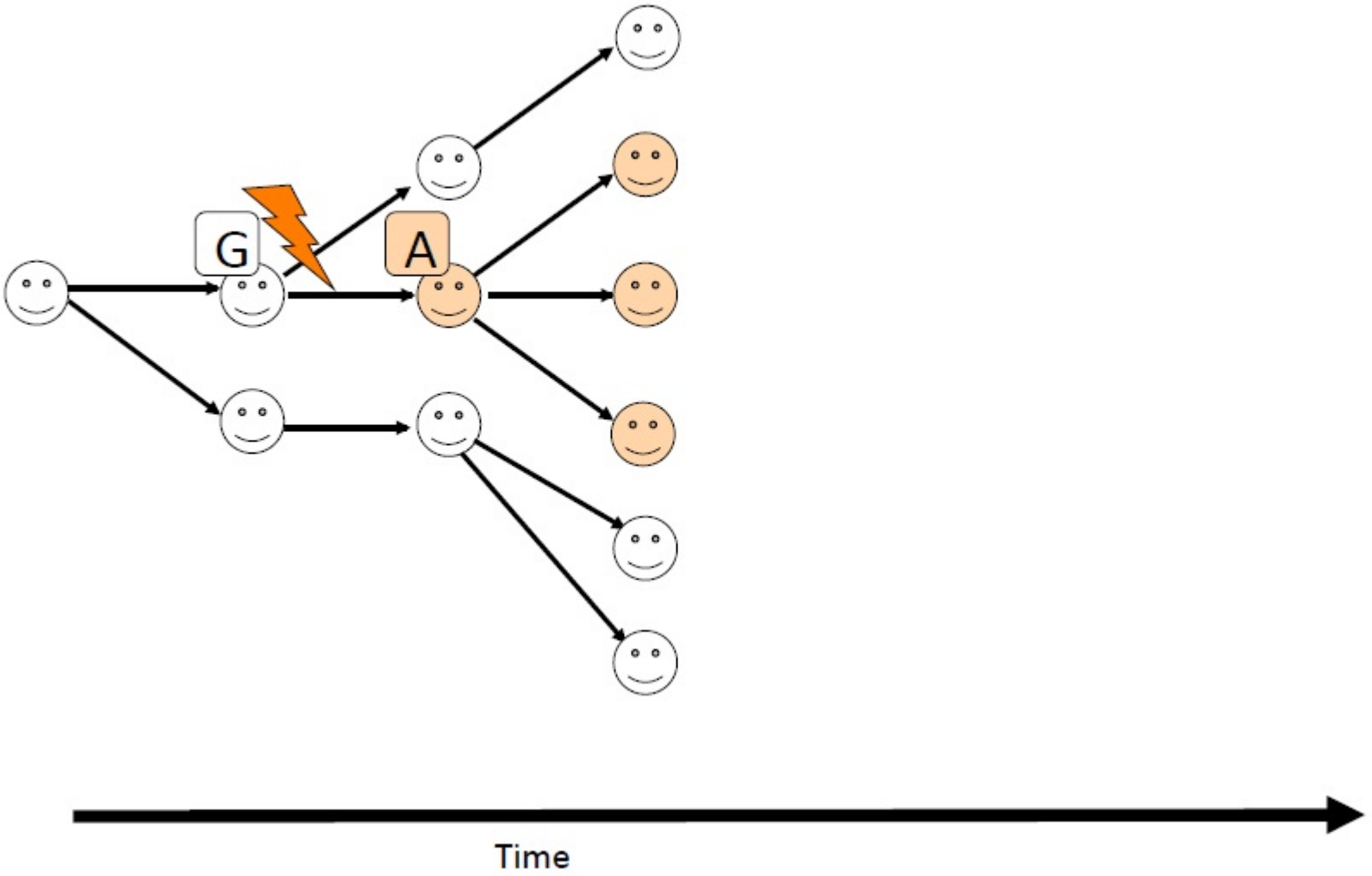
Out of scope for today

1. Non-random genotyping failure
2. Methods to correct for population stratification
3. Structural variants (SVs) and copy number variations (CNVs)

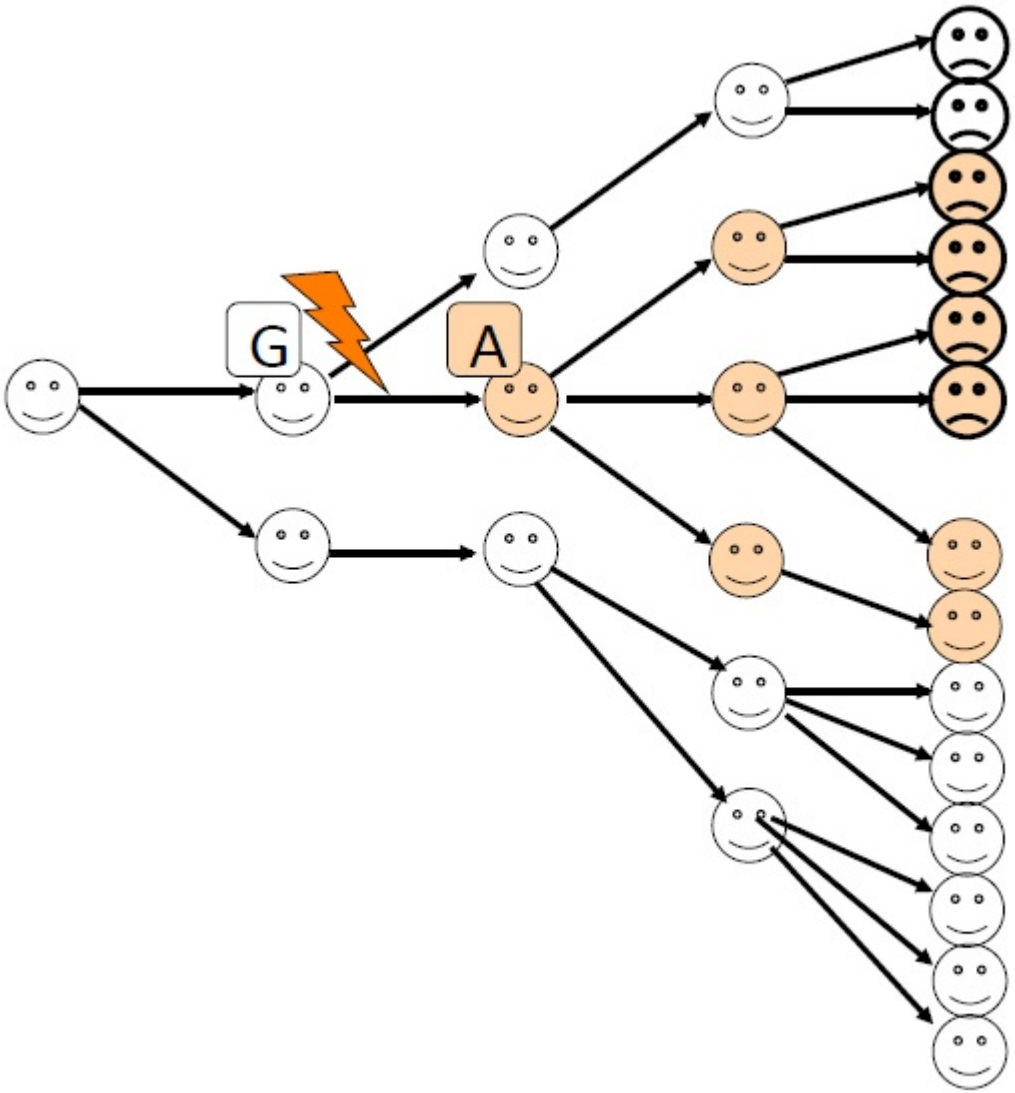


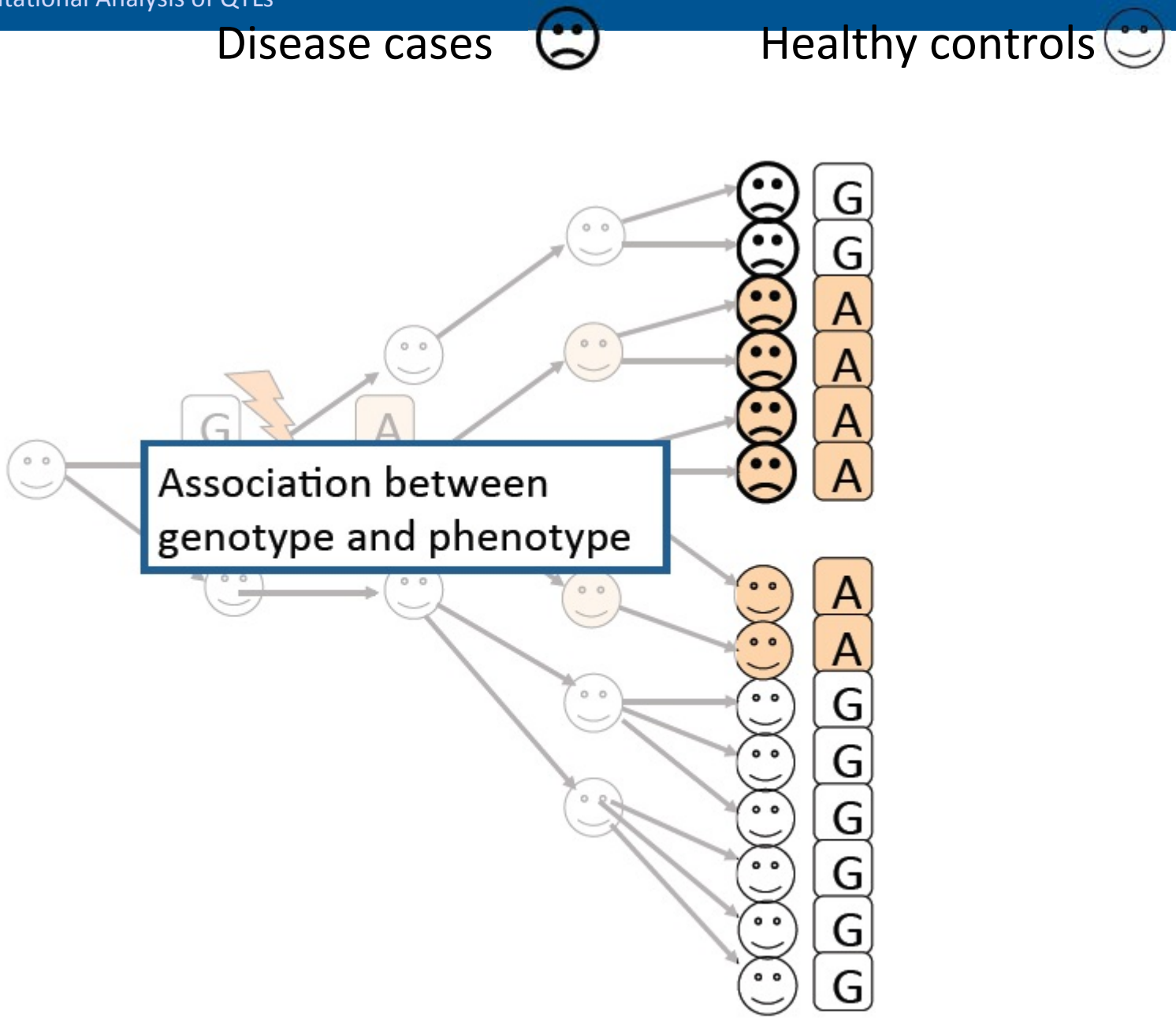
Courtesy of David Altshuler. Used with permission.

Slide courtesy of David Altshuler, HMS/Broad


G
A
Time

Disease cases Healthy control




Disease cases
Healthy controls
G
A
Association between genotype and phenotype
G
G
A
A
A
A
A
A
G
G
G
G
G
G

Age-related macular degeneration

Cohort – 2172 unrelated European descent individuals at least 60 years old

2004: Little known about cause of AMD

934 controls

1238 cases

Photographs are in the public domain.

SNP rs1061170

1238 individuals with AMD and 934 controls

2172 individuals / 4333 alleles

Allele	Cases (with AMD)	Controls (without AMD)	Total Alleles
C	1522 (a)	670 (b)	2192
T	954 (c)	1198 (d)	2152
Total Alleles	2476	1868	4344

$$\chi^2 = \frac{(ad-bc)^2(a+b+c+d)}{(a+b)(c+d)(b+d)(a+c)}$$

$X^2 = 279$ Df = (2 rows-1)x(2 columns-1) = 1

P-value = 1.2×10^{-62}

Contingency Tables – χ^2 test

Allele	Cases (with AMD)	Controls (without AMD)	Total Alleles
C	a	b	a+b
T	c	d	c+d
Total Alleles	a+c	b+d	a+b+c+d

$$E_1 = \frac{(a+b)(a+c)}{(a+b+c+d)} \qquad X^2 = \sum_{i=1}^{n} \frac{(O_i - E_i)^2}{E_i}$$

Df = (2 rows-1)x(2 columns-1) = 1

Contingency Tables – Fisher's Exact Test

Allele	Cases (with AMD)	Controls (without AMD)	Total Alleles
C	a	b	a+b
T	c	d	c+d
Total Alleles	a+c	b+d	a+b+c+d

$$p = \frac{\binom{a+b}{a}\binom{c+d}{c}}{\binom{a+b+c+d}{a+c}}$$

Sum all probabilities for observed and all more extreme values with same marginal totals to compute probability of null hypothesis

Does the affected or control group exhibit Population Stratification?

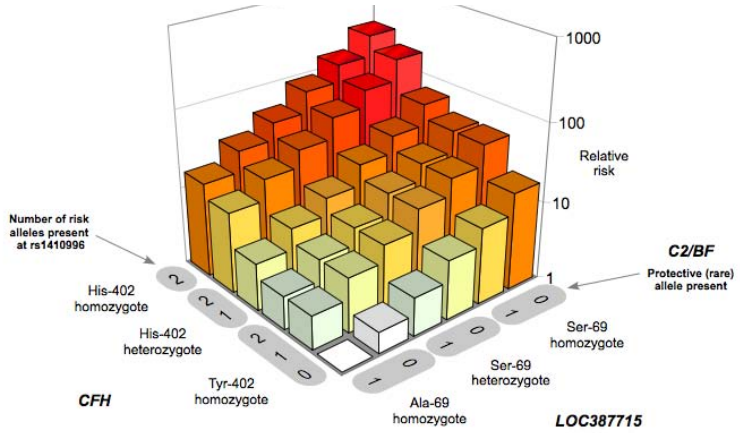
- Population stratification is when subpopulations exhibit allelic variation because of ancestry
- Can cause false positives in an association study if there are SNP differences in the case and control population structures
- Control for this artifact by testing control SNPs for general elevation in χ^2 distribution between cases and controls

Age-related macular degeneratio

2004: Little known about cause of AMD

Photographs are in the public domain.

2006: Three genes (5 common variants)
Together explain >50% of risk



Courtesy of Macmillan Publishers Limited. Used with permission.
Source: Maller, Julian, Sarah George, et al. "Common Variation in Three Genes, Including a Noncoding Variant in CFH, Strongly Influences Risk of Age-related Macular Degeneration." *Nature Genetics* 38, no. 9 (2006): 1055-9.

Relative risk plotted as a function of the genetic load of the five variants that influence risk of AMD. Two variants are in the CFH gene on chromosome 1: Y402H and rs1410996. Another common variant (A69S) is in hypothetical gene LOC387715 on chromosome 10. Two relatively rare variants are observed in the C2 and BF genes on chromosome 6. We find no evidence for interaction between any of these variants, suggesting an independent mode of action.

Edwards et al, Klein et al, Haines et al *Science* (2005); Jakobsdottir et al, *AJHG* (2005); Gold et al *Nature Genetics* (2006), Maller, George, Purcell, Fagerness, Altshuler, Daly, Seddon, Nature Genetics (2006)

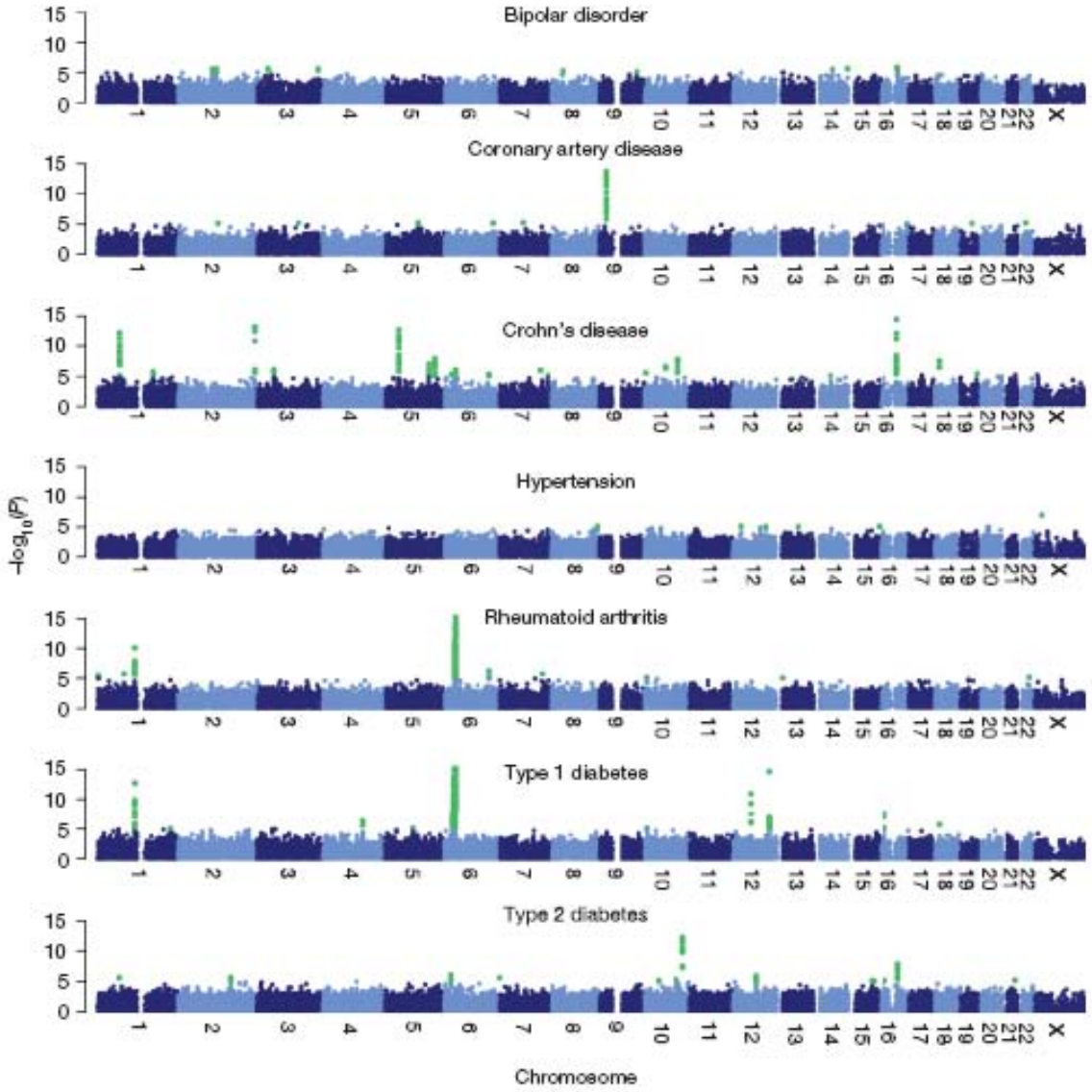


Figure 4 | Genome-wide scan for seven diseases. For each of seven diseases $-\log_{10}$ of the trend test P value for quality-control-positive SNPs, excluding those in each disease that were excluded for having poor clustering after visual inspection, are plotted against position on each chromosome. Chromosomes are shown in alternating colours for clarity, with P values $<1 \times 10^{-5}$ highlighted in green. All panels are truncated at $-\log_{10}(P \text{ value}) = 15$, although some markers (for example, in the MHC in T1D and RA) exceed this significance threshold.

Source: Burton, Paul R., David G. Clayton, et al. "Genome-wide Association Study of 14,000 Cases of Seven Common Diseases and 3,000 Shared Controls." *Nature* 447, no. 7145 (2007): 661-78.

Nature Vol 447|7 June 2007| doi:10.1038/nature05911

Linkage Disequilibrium (LD) between two loci L1 and L2 in gametes

At locus L1

p_A probability L1 is A

q_a probability L1 is a

At locus L2

p_B probability L2 is B

q_b probability L2 is b

	L2 B	L2 b
L1 A	$P_{AB} = p_A p_B + D$	$P_{Ab} = p_A q_b - D$
L1 a	$P_{aB} = q_a p_B - D$	$P_{ab} = q_a q_b + D$

D = Measure of linkage disequilibrium

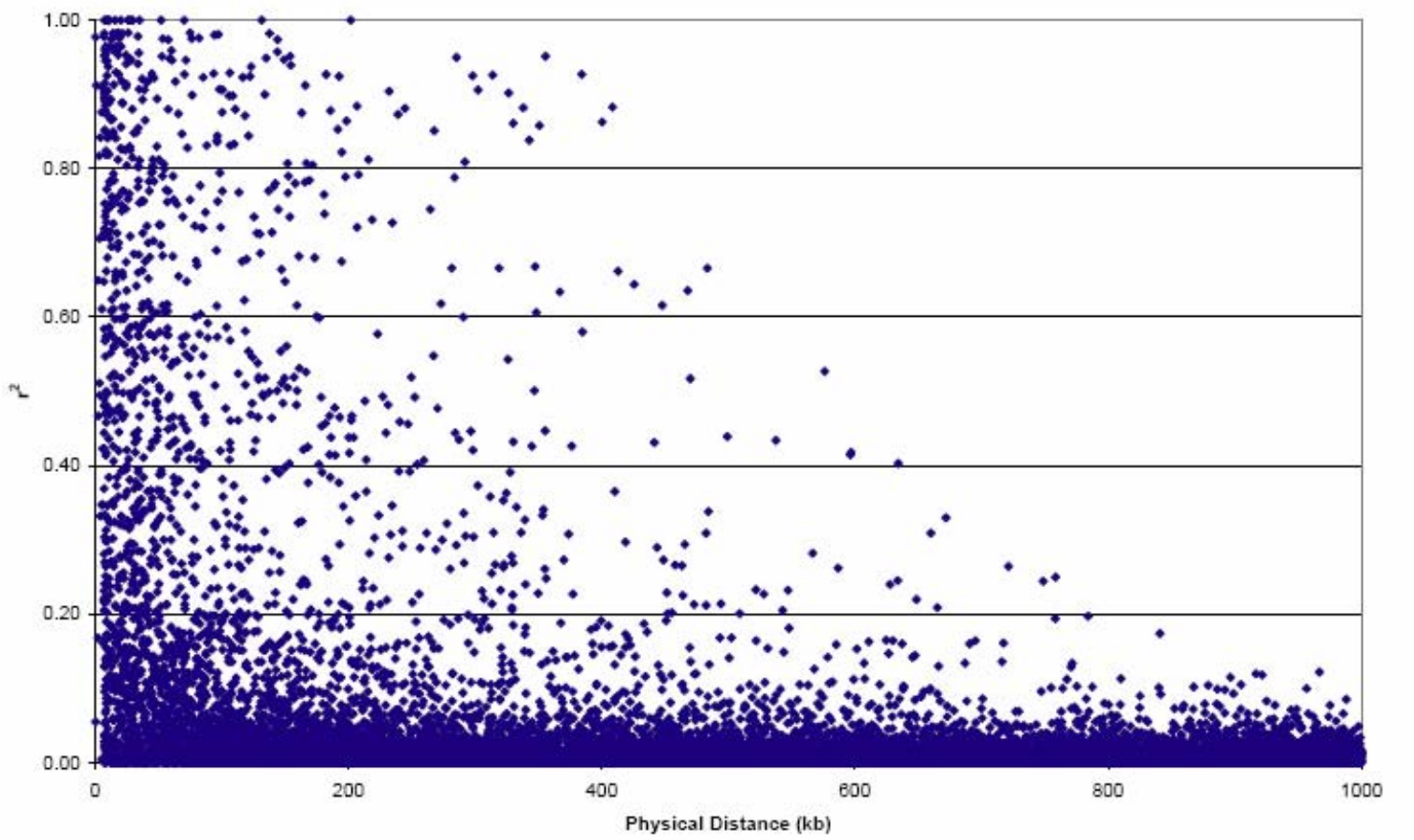
= 0 when L1 and L2 are in equilibrium

$D = P_{AB}P_{ab} - P_{Ab}P_{aB}$

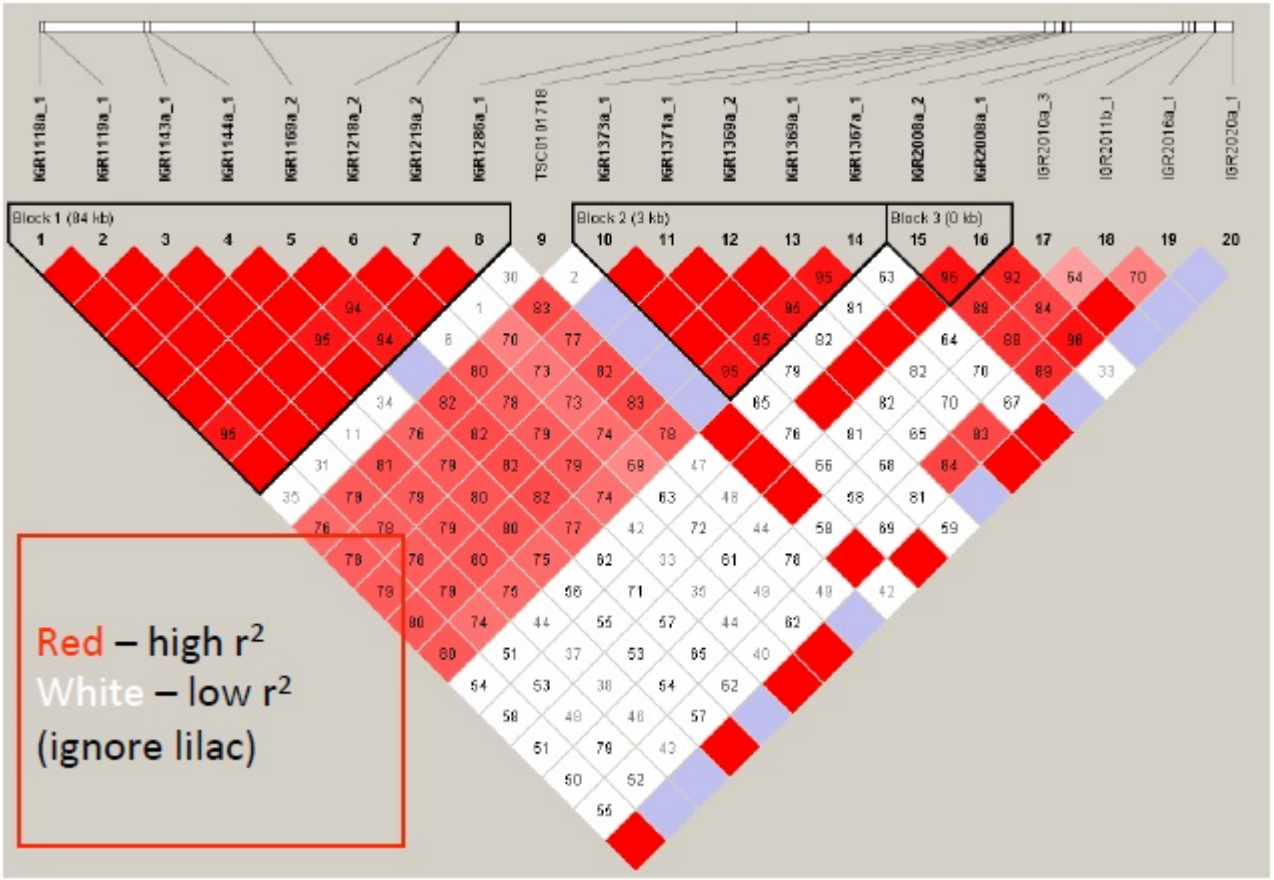
$r^2 = D^2 / (p_A q_a p_B q_b)$

r is [0,1] and is the correlation coefficient between allelic states in L1 and L2

r^2 from human chromosome 22

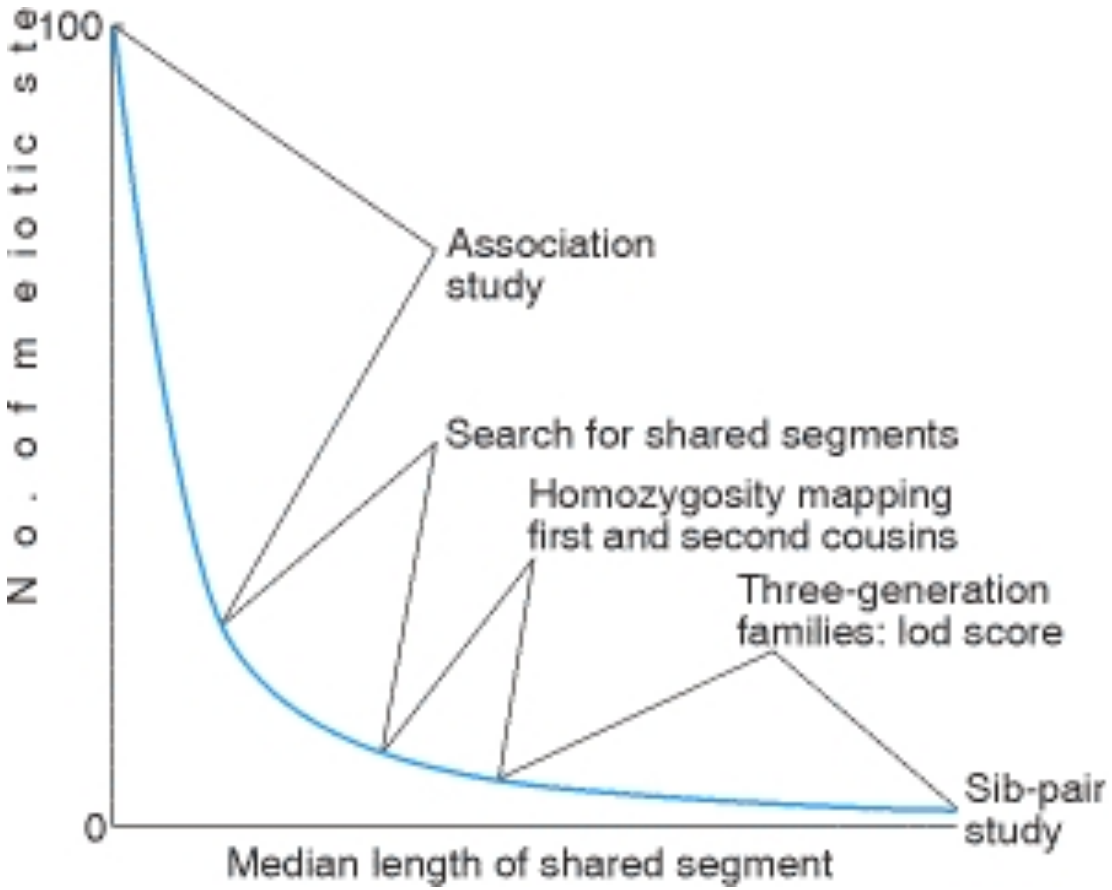


LD organizes the genome into haplotype blocks

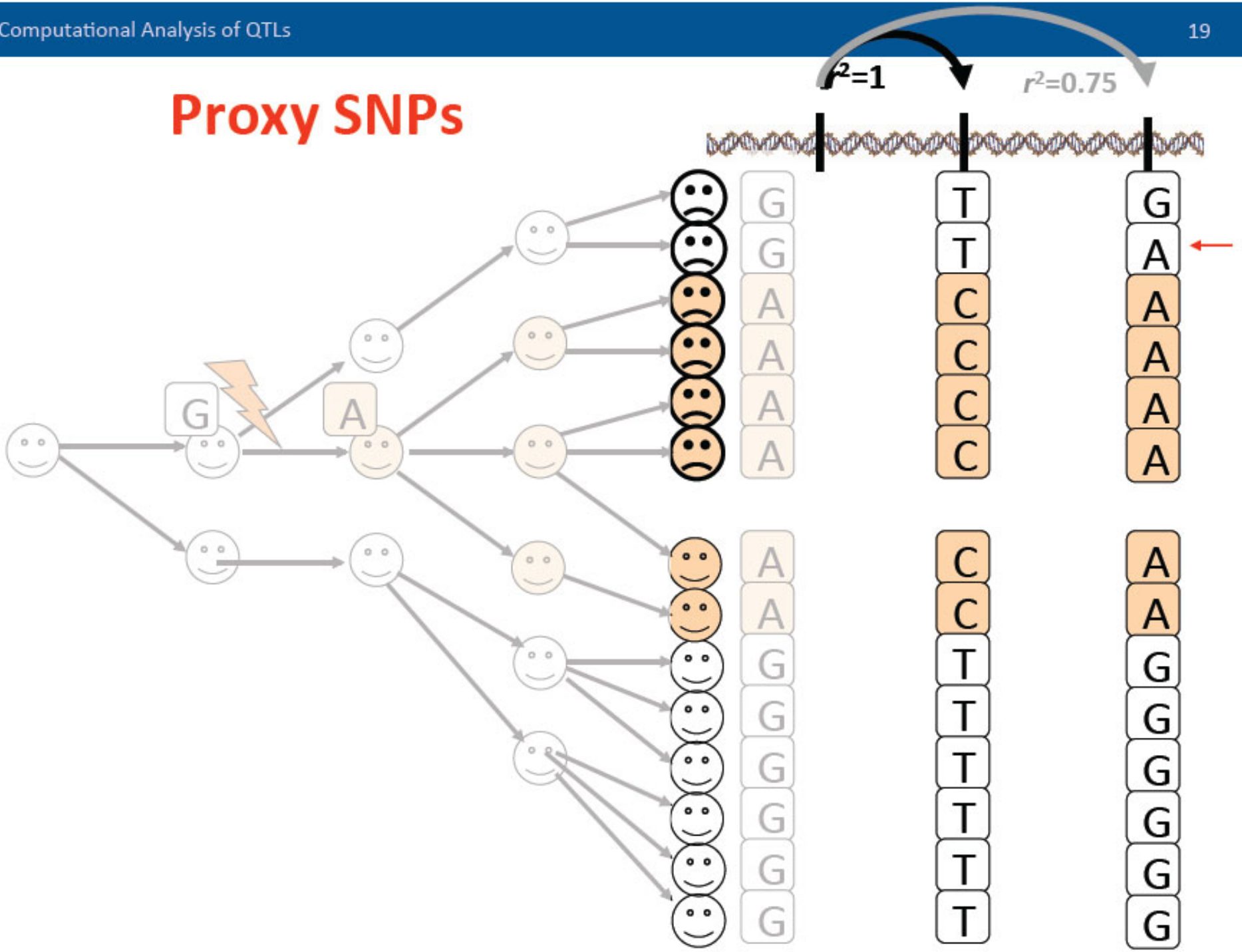


Human genome 5q31 region (associated with Inflammatory Bowel Disease)

The length of haplotype blocks vs time



Proxy SNPs



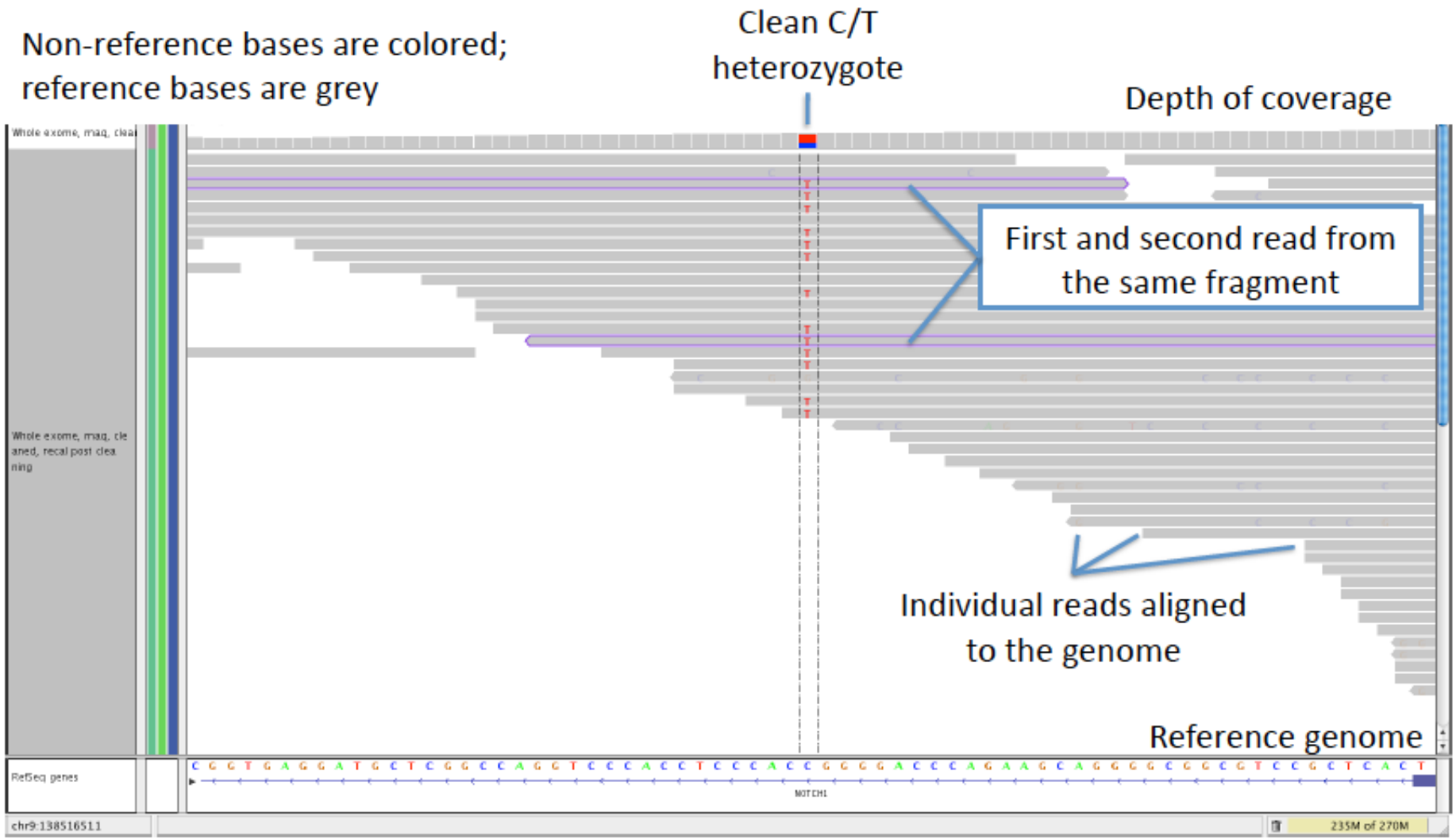
Variant Phasing

1. Phasing assigns alleles to their parental chromosome
2. Set of ordered alleles along a chromosome is a haplotype
3. Known haplotypes can assist with phasing
4. Phasing is critical for understanding the functional status of genes with more than one important SNPs (are the non-reference alleles on different chromosome? If so, the gene may not be functional)
5. New long read sequencing technologies phase variants in observed reads

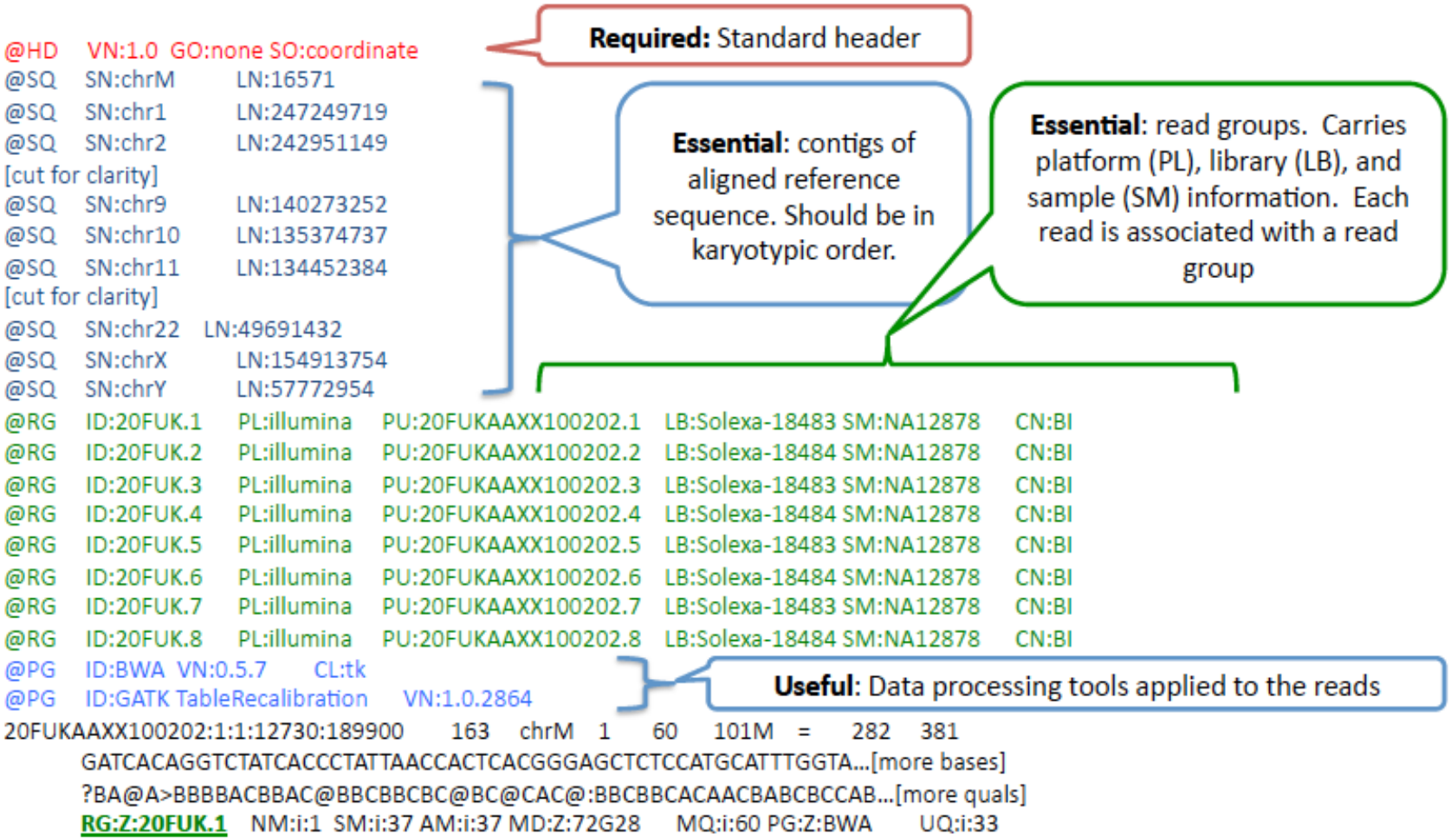
Today's Narrative Arc

1. We can discover human variants that are associated with a phenotype by studying the genotypes of case and control populations
   - Approach 1 – Use allelic counts from SNP arrays (SNPs called from microarray data)
   - **Approach 2 – Use read counts from sequencing (multiple reads per variant per individual)**
2. We can prioritize variants based upon their estimated importance
3. Follow up confirmation is important because correlation is not equivalent to causality

Prototypical IGV screenshot representing aligned NGS reads



BAM headers: an essential part of a BAM file



Official specification in http://samtools.sourceforge.net/SAM1.pdf

VCF Files store variant information

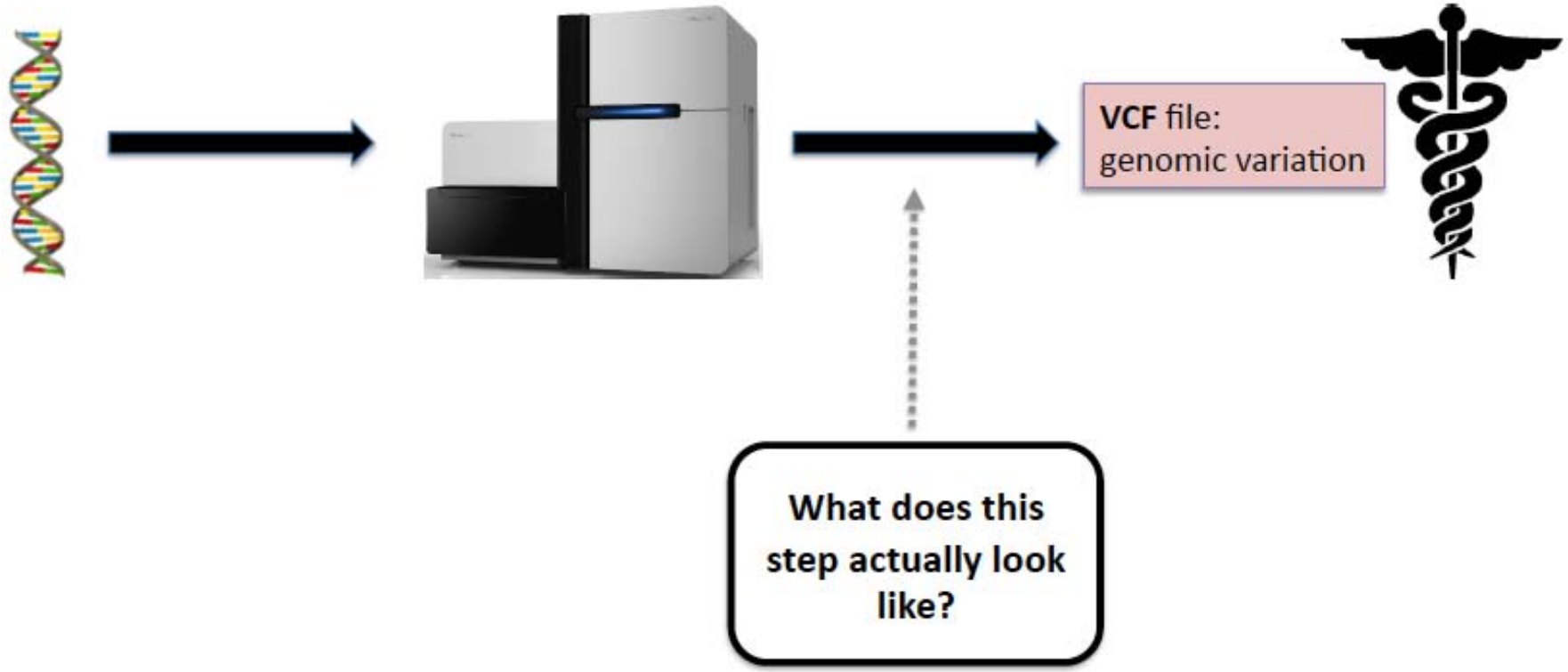
```
##fileformat=VCFv4.1
##reference=1000GenomesPilot-NCBI36
##INFO=<ID=DP,Number=1,Type=Integer,Description="Total Depth">
##INFO=<ID=AF,Number=A,Type=Float,Description="Allele Frequency">
##INFO=<ID=DB,Number=0,Type=Flag,Description="dbSNP membership, build 129">
##FILTER=<ID=s50,Description="Less 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">
#CHROM POS     ID        REF ALT    QUAL FILTER INFO
    FORMAT      NA00001        NA00002        NA00003

20     14370   rs6054257 G      A       29   PASS   DP=14;AF=0.5;DB
    GT:GQ:DP 0|0:48:1 1|0:48:8 1/1:43:5
20     1110696 rs6040355 A      G,T     67   PASS   DP=10;AF=0.333,0.667;DB
    GT:GQ:DP 1|2:21:6 2|1:2:0   2/2:35:4
20     1230237 .         T      .       47   PASS   DP=13
    GT:GQ:DP 0|0:54:7 0|0:48:4 0/0:61:2
20     1234567 microsat1 GTCT   G,GTACT 50   PASS   DP=9
    GT:GQ:DP    0/1:35:4       0/2:17:2       1/1:40:3
```

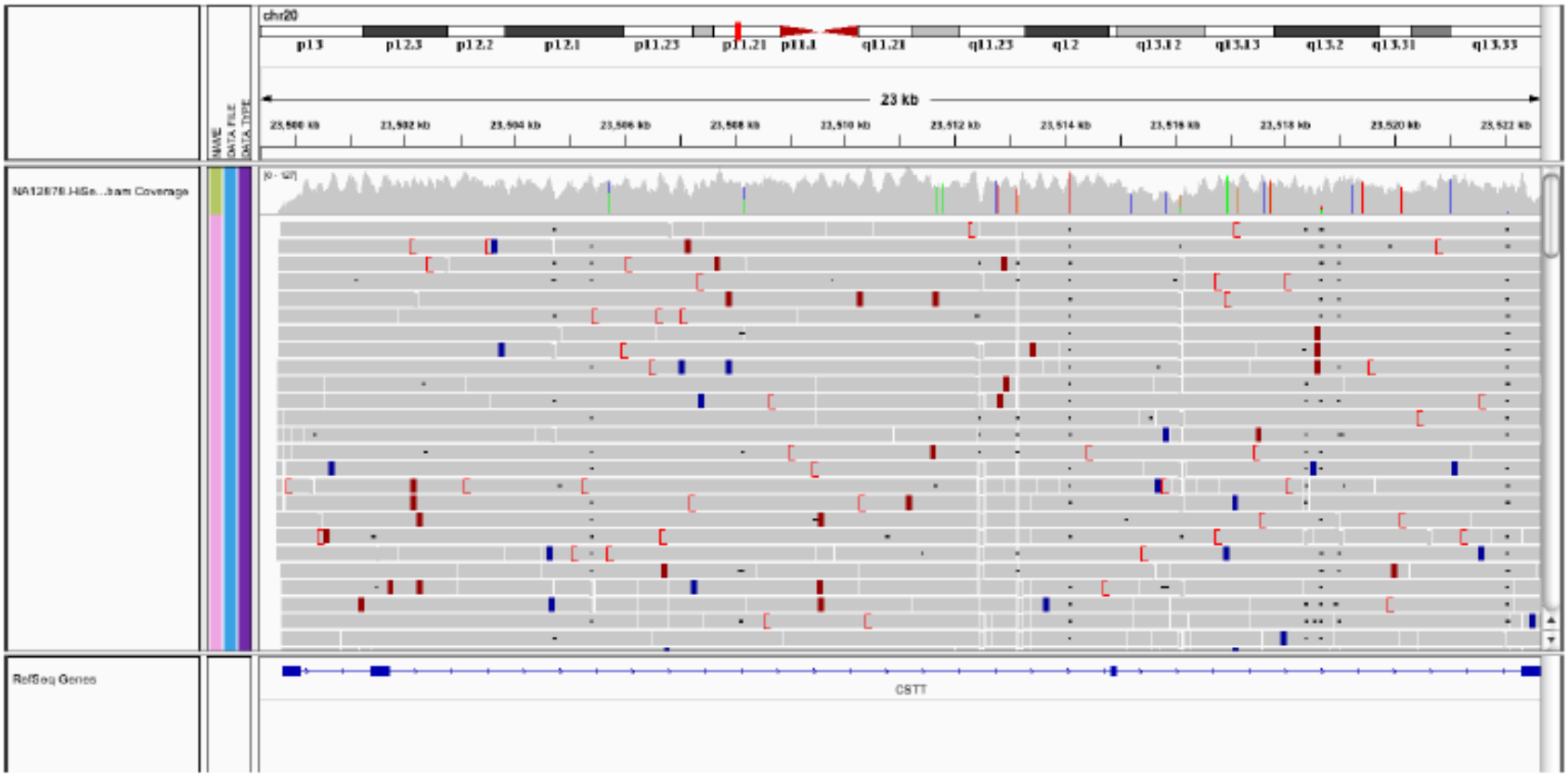
Header

Variant records

Is processing/analysis of NGS data really that easy?



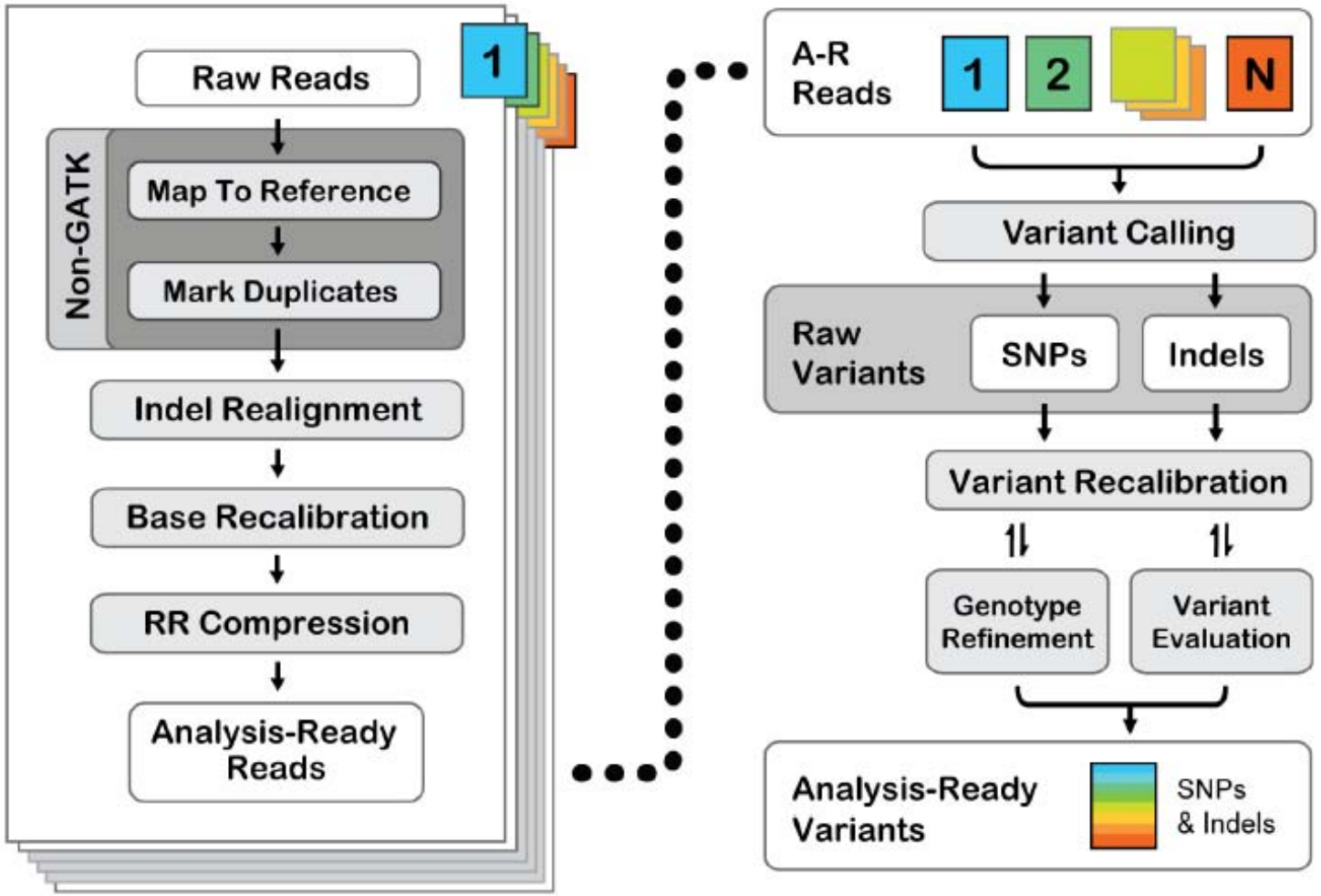
It's going to involve dealing with messy situations like this:



How can we tell which mismatches represent real mutations and which are just noise?

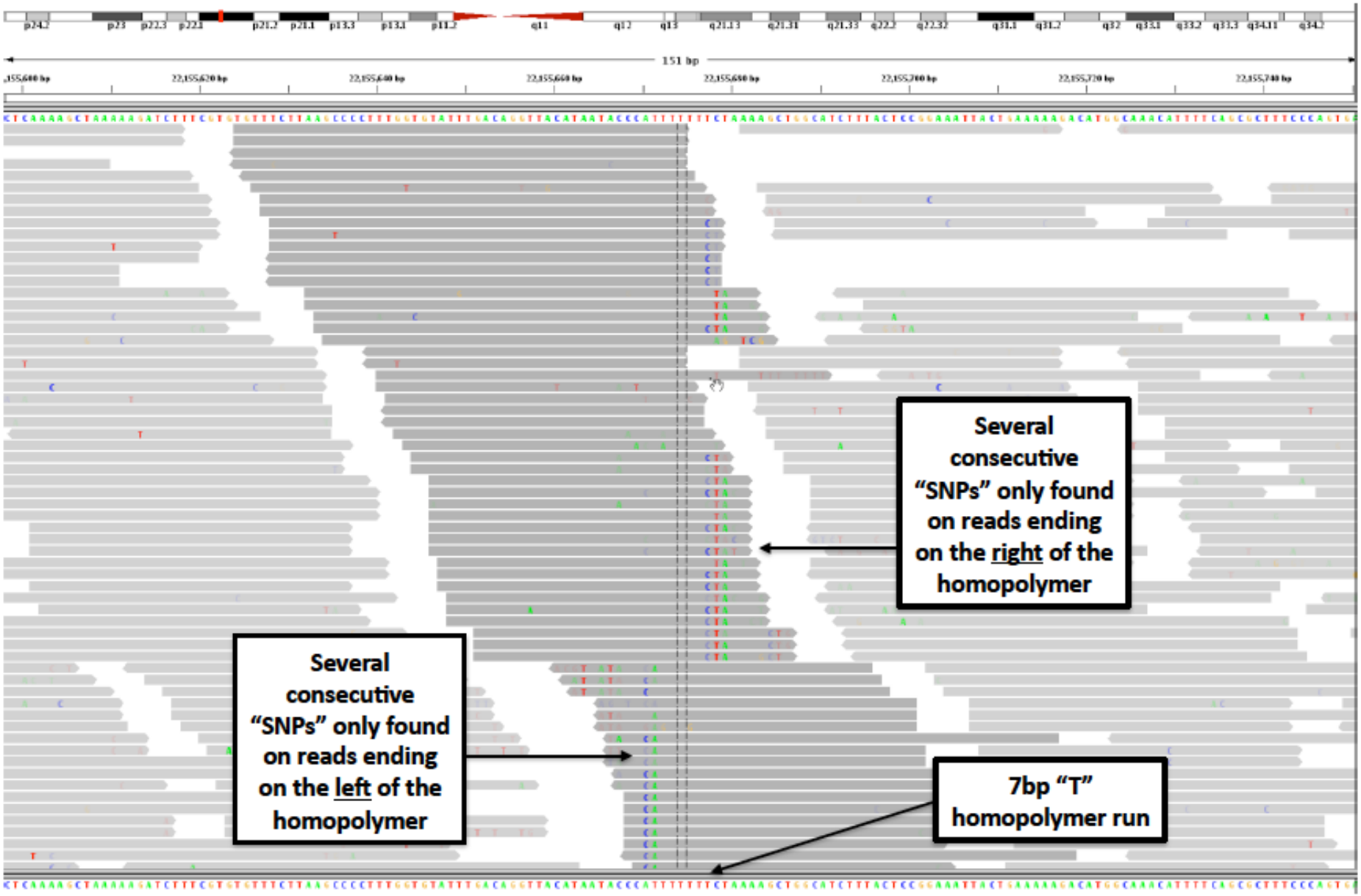
Genome Analysis Tool Kit (GATK)

Scope and schema of the Best Practices



Courtesy of the Broad Institute. Used with permission. The most recent best practices can be found at this website: https://www.broadinstitute.org/gatk/guide/best-practices.

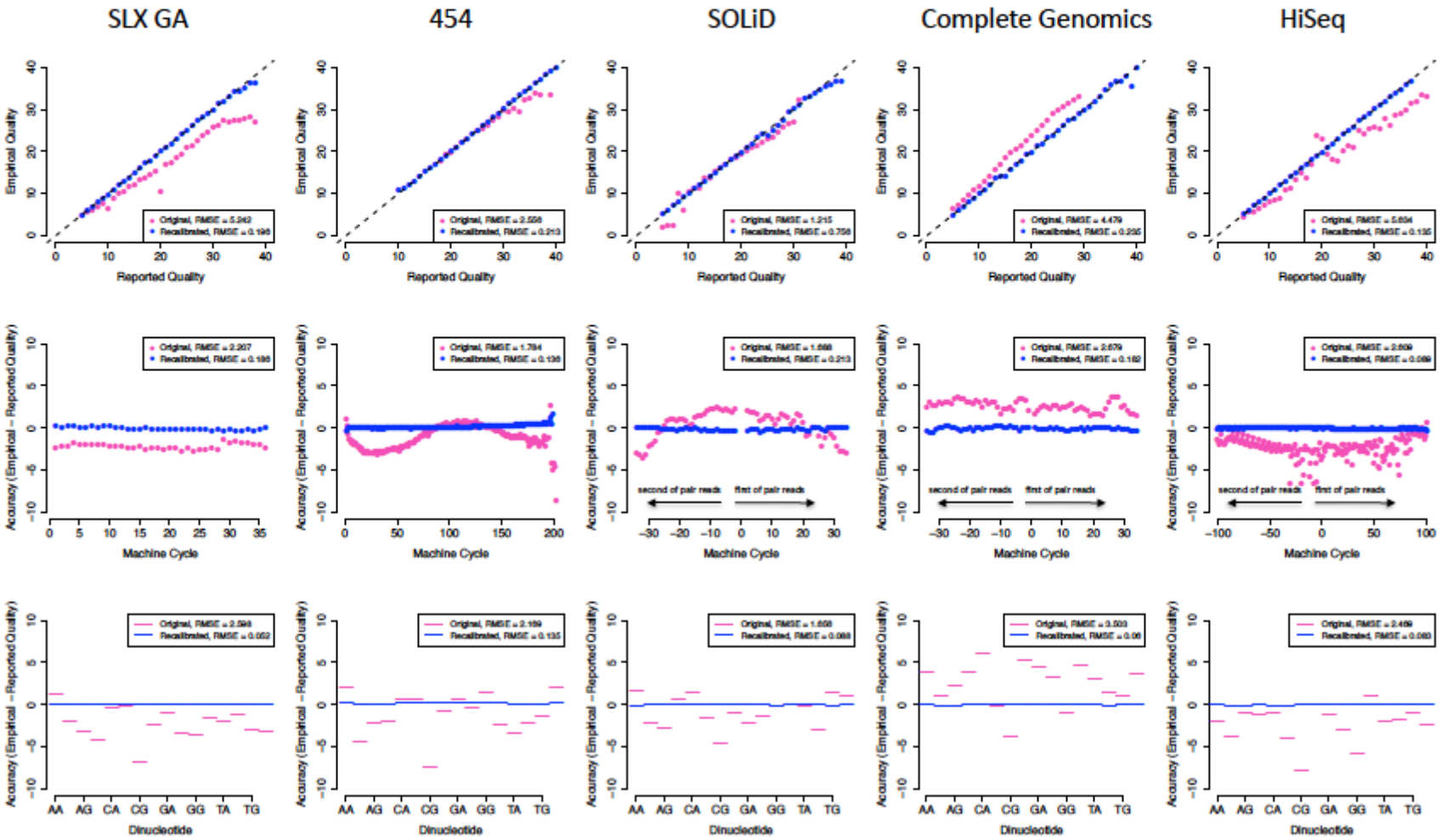
An example of a strand-discordant locus



Local realignment uncovers the hidden indel in these reads and eliminates all the potential FP SNPs

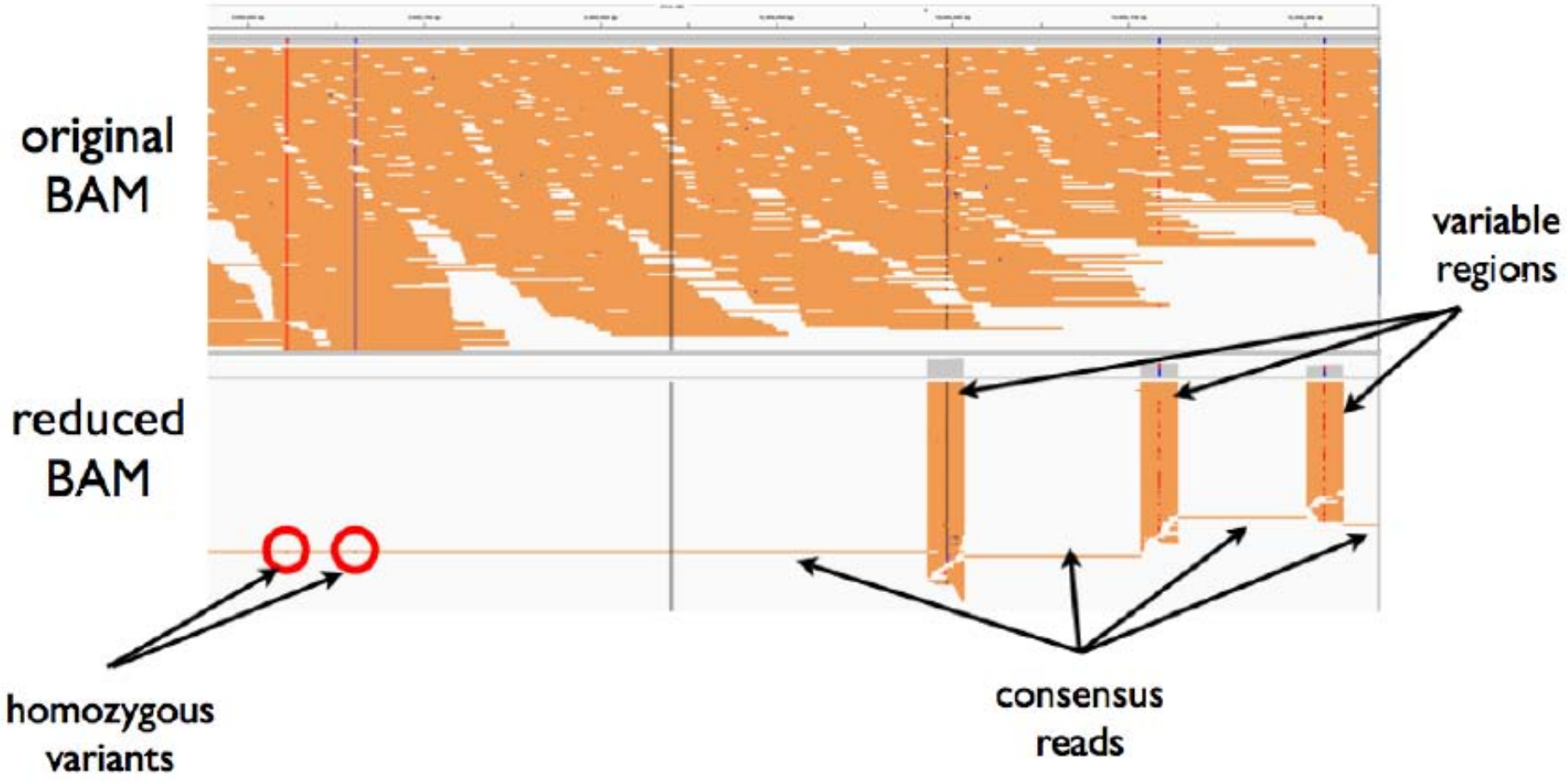
Highlighted as one of the major methodological advances of the 1000 Genomes Pilot Project!

Base Quality Score Recalibration provides a calibrated error model from which to make mutation calls

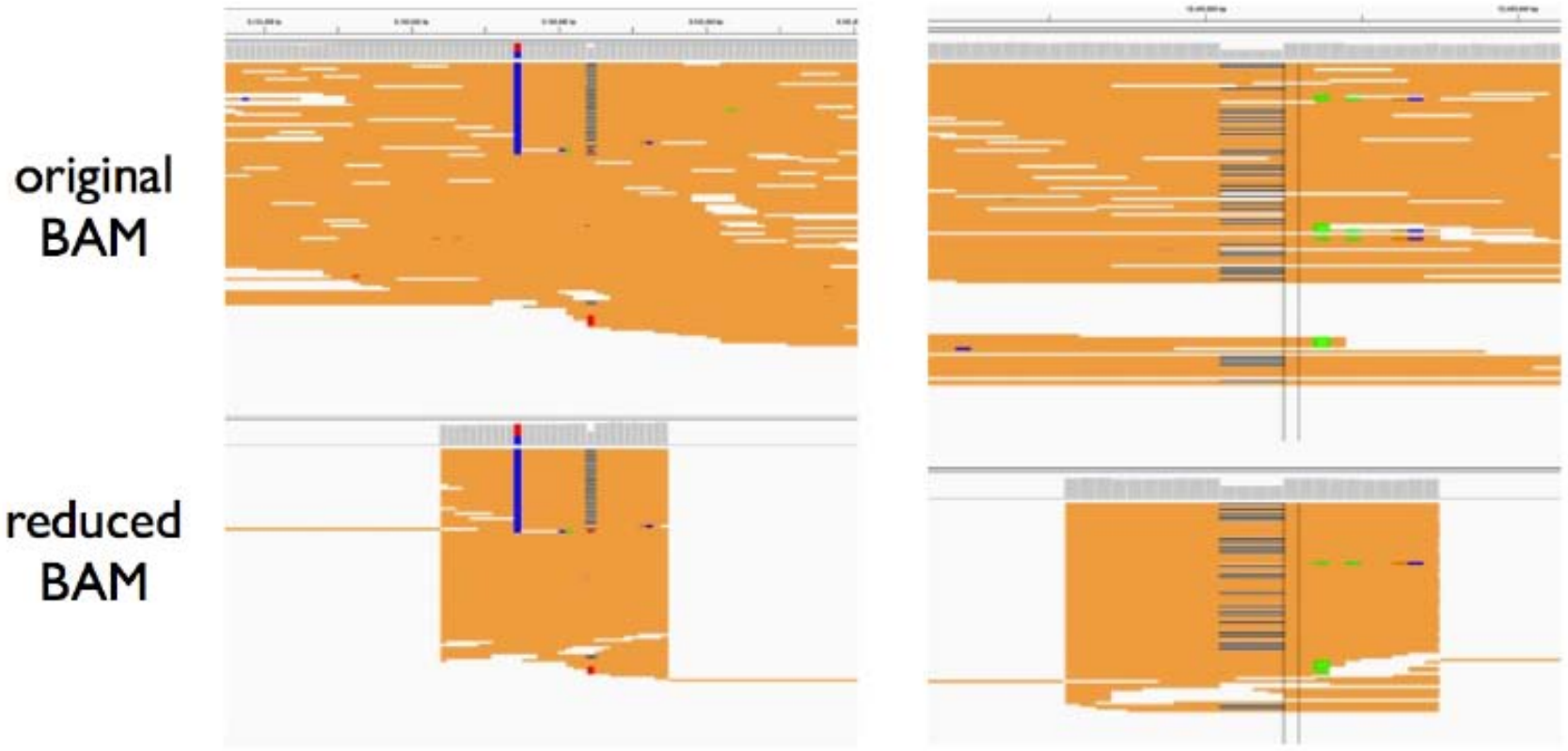


Courtesy of the Broad Institute. Used with permission. The most recent best practices can be found at this website: https://www.broadinstitute.org/gatk/guide/best-practices.

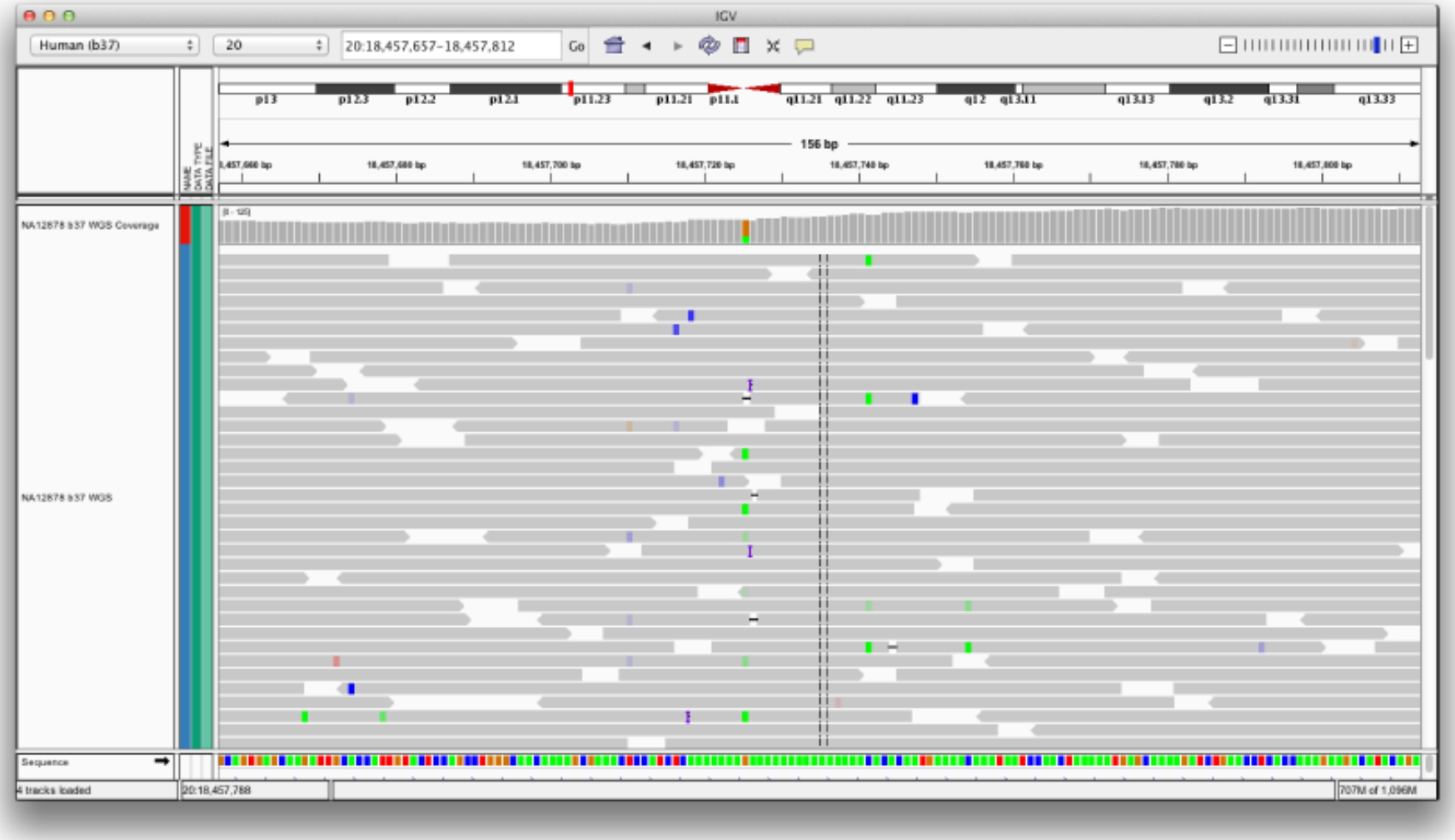
This is what a compressed BAM looks like



Important to handle complex cases properly



Real mutations are hidden in the noise

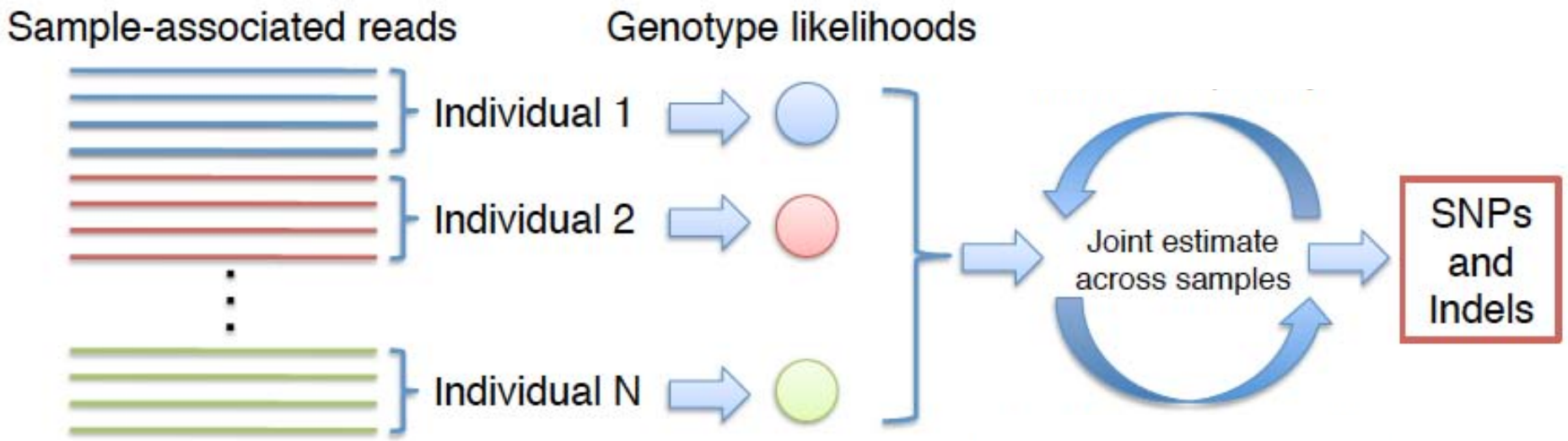


Computing genotypes

$$1 = \sum_{G \in \{AA, AC, \ldots, TT\}} P(G)$$

Given the reads we observe we wish to compute $P(G_p)$ at a SNP for a population p (p could be cases or controls)

Joint estimation of genotype frequencies



- Simultaneous estimation of:
  - Allele frequency (AF) spectrum Pr{AF = i | D}
  - The probability that a variant exists Pr{AF > 0 | D}
  - Assignment of genotypes to each sample

Compute Bayesian posterior genotype frequencies (G) for each individual from their reads (D)

Bayesian model

$$\Pr\{G|D\} = \frac{\Pr\{G\}\Pr\{D|G\}}{\Sigma_i \Pr\{G_i\}\Pr\{D|G_i\}}, \text{ [Bayes' rule]}$$

Prior of the genotype; Likelihood of the genotype

$$\Pr\{D|G\} = \prod_j \left(\frac{\Pr\{D_j|H_1\}}{2} + \frac{\Pr\{D_j|H_2\}}{2} \right) \text{ where } G=H_1,H_2$$

Diploid assumption

$$\Pr\{D|H\} \text{ is the haploid likelihood function}$$

- Inference: what is the genotype G of each sample given read data D for each sample?
- Calculate via Bayes' rule the probability of each possible G
- Product expansion assumes reads are independent
- Relies on a likelihood function to estimate probability of sample data given proposed haplotype

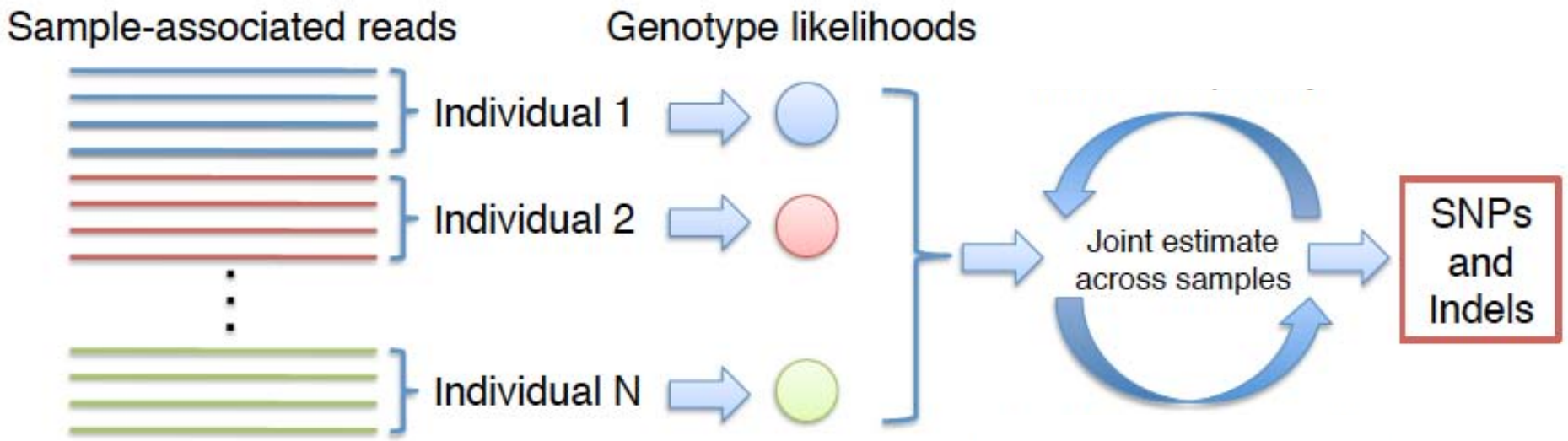
Haploid likelihood considers the probability of errors

$$\Pr\{D_j|H\} = \Pr\{D_j|b\}, \ [D_j \text{ is a single read}]$$

$$\Pr\{D_j|b\} = \begin{cases} 1-\epsilon_j & D_j = b, \\ \epsilon_j & \text{otherwise.} \end{cases}$$

- All diploid genotypes (AA, AC, …, GT, TT) considered at each base
- Likelihood of genotype computed using only pileup of bases and associated quality scores at given locus
- Only "good bases" are included: those satisfying minimum base quality, mapping read quality, pair mapping quality, NQS

Joint estimation of genotype frequencies



- Simultaneous estimation of:
  - Allele frequency (AF) spectrum Pr{AF = i | D}
  - The probability that a variant exists Pr{AF > 0 | D}
  - Assignment of genotypes to each sample

EM can be used to improve the estimate of $P(G_p)$

$$P(G_p)^{(t+1)} = \frac{1}{n}\sum_{i=1}^{n}\frac{P(D_i \mid G_p)P(G_p)^{(t)}}{\sum_{G_p'} P(D_i \mid G_p')P\left(G_p'\right)^{(t)}}$$

Testing for associations

Assume a reference allele (A) and a single non-reference allele (a)

$$\psi = P(A)$$
$$(1-\psi) = P(a)$$
$$\varepsilon_0 = P(AA)$$
$$\varepsilon_1 = P(Aa)$$
$$\varepsilon_2 = P(aa)$$

Testing for Hardy Weinberg Equilibrium (HWE)

When a population is in HWE we can compute genotypic frequencies from allelic frequencies

We can test for HWE as follows –

$$T_3 = 2\log\frac{P(D\,|\,\varepsilon_0,\varepsilon_1,\varepsilon_2)}{P(D\,|\,(1-\psi)^2,2\psi(1-\psi),\psi^2)}$$

Testing for associations

$$T_1 = 2\log\frac{P(D^{[1]} \mid \psi^{[1]})P(D^{[2]} \mid \psi^{[2]})}{P(D \mid \psi)}$$

[1] and [2] are cases and controls. Do not use T_2 when population is in HWE as it will be underpowered (too many DOF)

$$T_2 = 2\log\frac{P(D^{[1]} \mid \varepsilon_0^{[1]}, \varepsilon_1^{[1]}, \varepsilon_2^{[1]})P(D^{[2]} \mid \varepsilon_0^{[2]}, \varepsilon_1^{[2]}, \varepsilon_2^{[2]})}{P(D \mid \varepsilon_0, \varepsilon_1, \varepsilon_2)}$$

HaplotypeCaller method overview

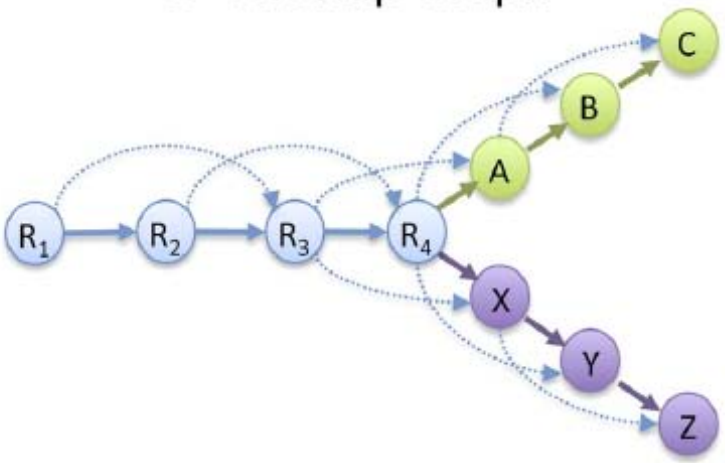
- Call SNPs, indels, and some SVs simultaneously by performing a local *de-novo* assembly
  - Determine if a region has the potential to be variable
  - Construct a deBruijn assembly of the region
  - The paths in the graph are potential haplotypes that need to be evaluated
  - Calculate haplotype likelihoods given the data using the PairHMM model
  - Determine if there are any variants on the most likely haplotypes
  - Compute the allele frequency distribution to determine most likely allele count, and emit a variant call if determined
  - If we are going to emit a variant, assign a genotype to each sample

Propose haplotypes with local de novo assembly via DeBruijn graphs

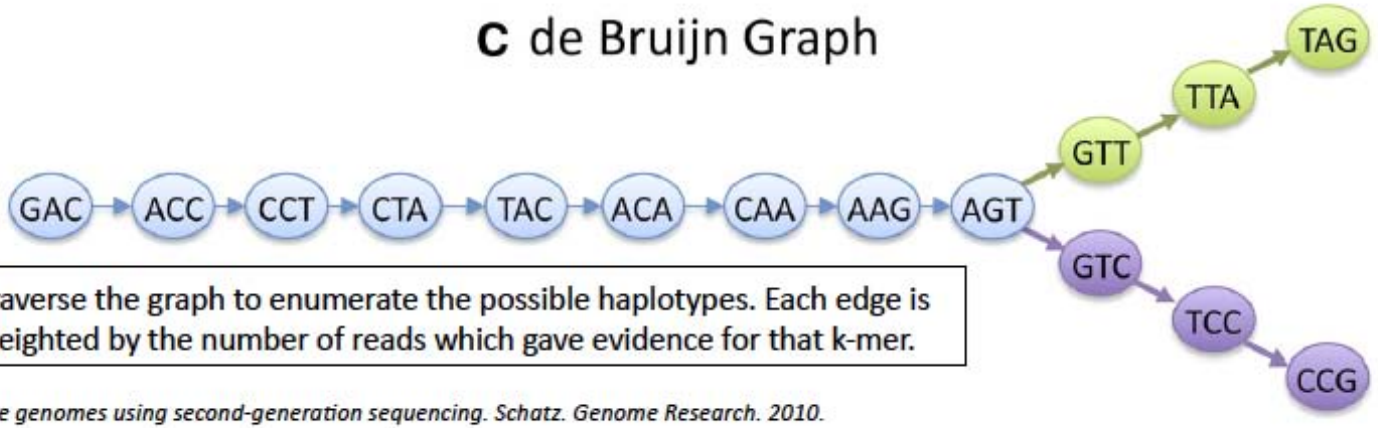
A Read Layout

```
R1: GACCTACA
R2:   ACCTACAA
R3:    CCTACAAG
R4:     CTACAAGT
A:       TACAAGTT
B:        ACAAGTTA
C:         CAAGTTAG
X:       TACAAGTC
Y:        ACAAGTCC
Z:         CAAGTCCG
```

B Overlap Graph



C de Bruijn Graph

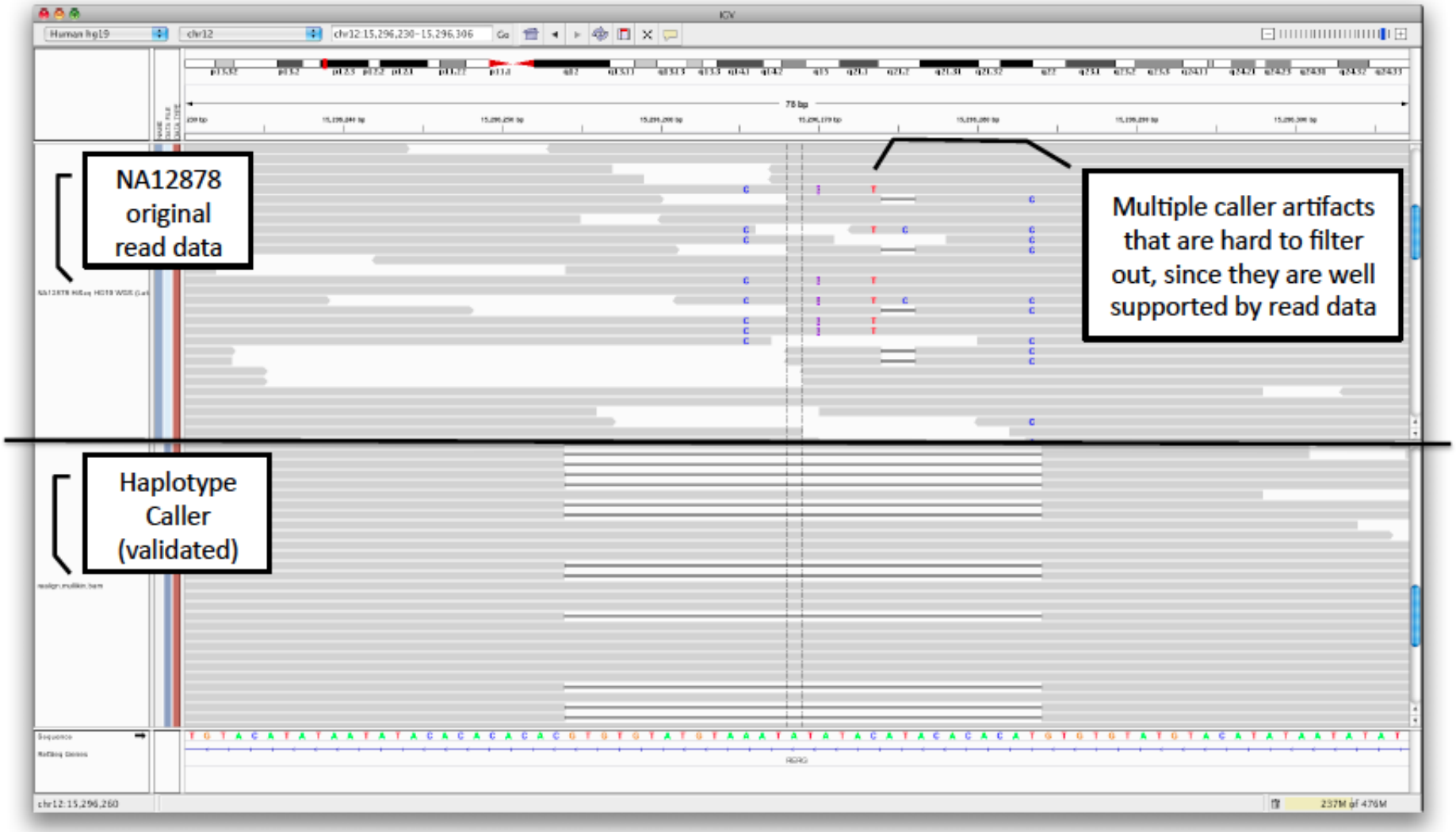


Traverse the graph to enumerate the possible haplotypes. Each edge is weighted by the number of reads which gave evidence for that k-mer.

20

Assembly of large genomes using second-generation sequencing. Schatz. Genome Research. 2010.

Artifactual SNPs and small indels caused by large indel recovered by assembly



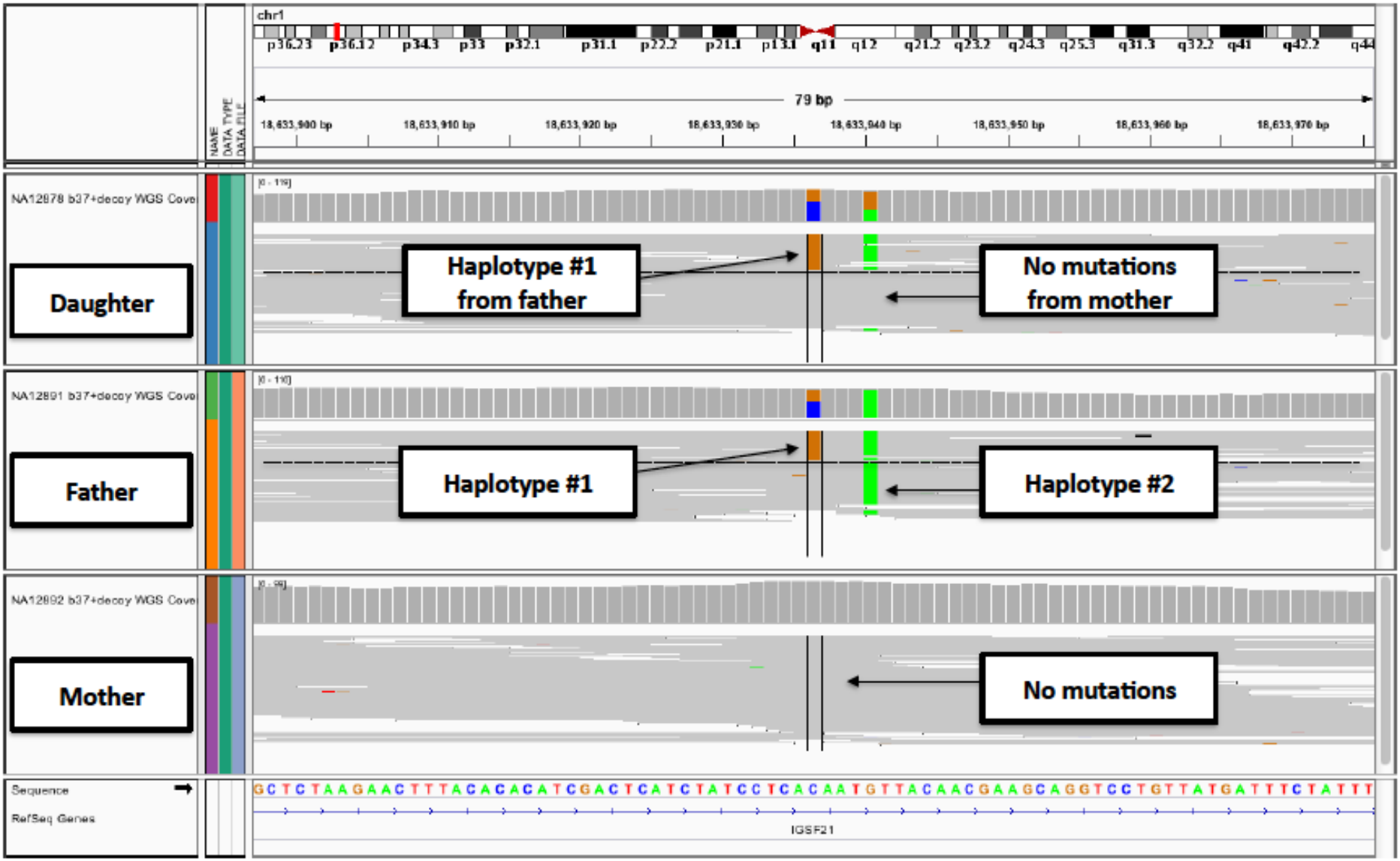
Many downstream genetic analyses need accurate genotypes and/or phasing information

- E.g. Mendelian disease caused by Loss Of Function event
  - Homozygous mutation causing disease (both copies affected)
  - Compound heterozygote (het mutations on different copies)

- Critical in population genetics studies to determine haplotype structure

☑ **Refining and phasing genotypes empowers downstream medical and population genetics analyses that require accurate determination of haplotype structure.**

Example site showing Mendelian inheritance in a trio



Original VCF

#CHROM	POS	ID	REF	ALT	QUAL	FILTER	INFO	FORMAT	MOTHER	FATHER	CHILD
1	10109	.	A	T	99	PASS	.	GT:PL	0/0:0,50,200	0/0:0,40,200	0/1:30,0,200
1	10147	.	C	A	99	PASS	.	GT:PL	**0/0**:0,30,200	**0/0**:0,50,200	**1/1**:200,40,0
1	10150	.	C	T	99	PASS	.	GT:PL	0/0:0,40,200	0/1:30,0,200	1/1:200,50,0

Phased VCF

#CHROM	POS	ID	REF	ALT	QUAL	FILTER	INFO	FORMAT	MOTHER	FATHER	CHILD
1	10109	.	A	T	99	PASS	.	GT:PL:TP	0|0:0,50,200:10	0|0:0,40,200:10	0|0:30,0,200:10
1	10147	.	C	A	99	PASS	.	GT:PL:TP	**1|0**:0,30,200:10	**0|0**:0,50,200:10	**1|0**:200,40,0:10
1	10150	.	C	T	99	PASS	.	GT:PL:TP	1|0:0,40,200:10	1|0:30,0,200:10	1|1:200,50,0:10

The convention is:
Allele From Mother | Allele From Father

- Simplified VariantAnnotator annotation of most egregious effect

```
SNPEFF_EFFECT=SPLICE_SITE_ACCEPTOR;SNPEFF_FUNCTIONAL_CLASS=NONE;SNPEFF_GENE_BIOTYPE=
protein_coding;SNPEFF_GENE_NAME=AURKAIP1;SNPEFF_IMPACT=HIGH;SNPEFF_TRANSCRIPT_ID=ENS
T00000470457
```

Today's Narrative Arc

1. We can discover human variants that are associated with a phenotype by studying the genotypes of case and control populations
   - Approach 1 – Use allelic counts from SNP arrays (SNPs called from microarray data)
   - Approach 2 – Use read counts from sequencing (multiple reads per variant per individual)
2. **We can prioritize variants based upon their estimated importance**
3. Follow up confirmation is important because correlation is not equivalent to causality

LETTERS

nature genetics

Recessive mutations in a distal *PTF1A* enhancer cause isolated pancreatic agenesis

Michael N Weedon[1,12], Inês Cebola[2–4,12], Ann-Marie Patch[1,12], Sarah E Flanagan[1], Elisa De Franco[1], Richard Caswell[1], Santiago A Rodríguez-Seguí[2,3], Charles Shaw-Smith[1], Candy H-H Cho[5], Hana Lango Allen[1], Jayne A L Houghton[1], Christian L Roth[6], Rongrong Chen[7], Khalid Hussain[8,9], Phil Marsh[10], Ludovic Vallier[5], Anna Murray[1], International Pancreatic Agenesis Consortium[11], Sian Ellard[1,13], Jorge Ferrer[2–4,13] & Andrew T Hattersley[1,13]

The contribution of cis-regulatory mutations to human disease remains poorly understood. Whole-genome sequencing can identify all noncoding variants, yet the discrimination of causal regulatory mutations represents a formidable challenge. We used epigenomic annotation in human embryonic stem cell (hESC)-derived pancreatic progenitor cells to guide the interpretation of whole-genome sequences from individuals with isolated pancreatic agenesis. This analysis uncovered six different recessive mutations in a previously uncharacterized ~400-bp sequence located 25 kb downstream of *PTF1A* (encoding pancreas-specific transcription factor 1a) in ten families with pancreatic agenesis. We show that this region acts as a developmental enhancer of *PTF1A* and that the mutations abolish enhancer activity. These mutations are the most common cause of isolated pancreatic agenesis. Integrating genome sequencing and epigenomic annotation in a disease-relevant cell type can thus uncover new noncoding elements underlying human development and disease.

Most individuals with syndromic pancreatic agenesis have heterozygous dominant mutations in *GATA6* (refs. 1,2). Extrapancreatic features in these individuals include cardiac malformations, biliary tract defects, and gut and other endocrine abnormalities. Four families have been reported with syndromic pancreatic agenesis, with severe neurological features and cerebellar agenesis caused by recessive coding mutations in *PTF1A*[3–5]. Most cases of isolated, non-syndromic pancreatic agenesis remain unexplained, with the only cause described being recessive coding mutations in *PDX1* that were reported in two families[6,7]. We previously noted that individuals with unexplained pancreatic agenesis were often born to consanguineous parents and rarely had extrapancreatic features[1]. These observations suggested an autosomal recessive defect underlying isolated pancreatic agenesis.

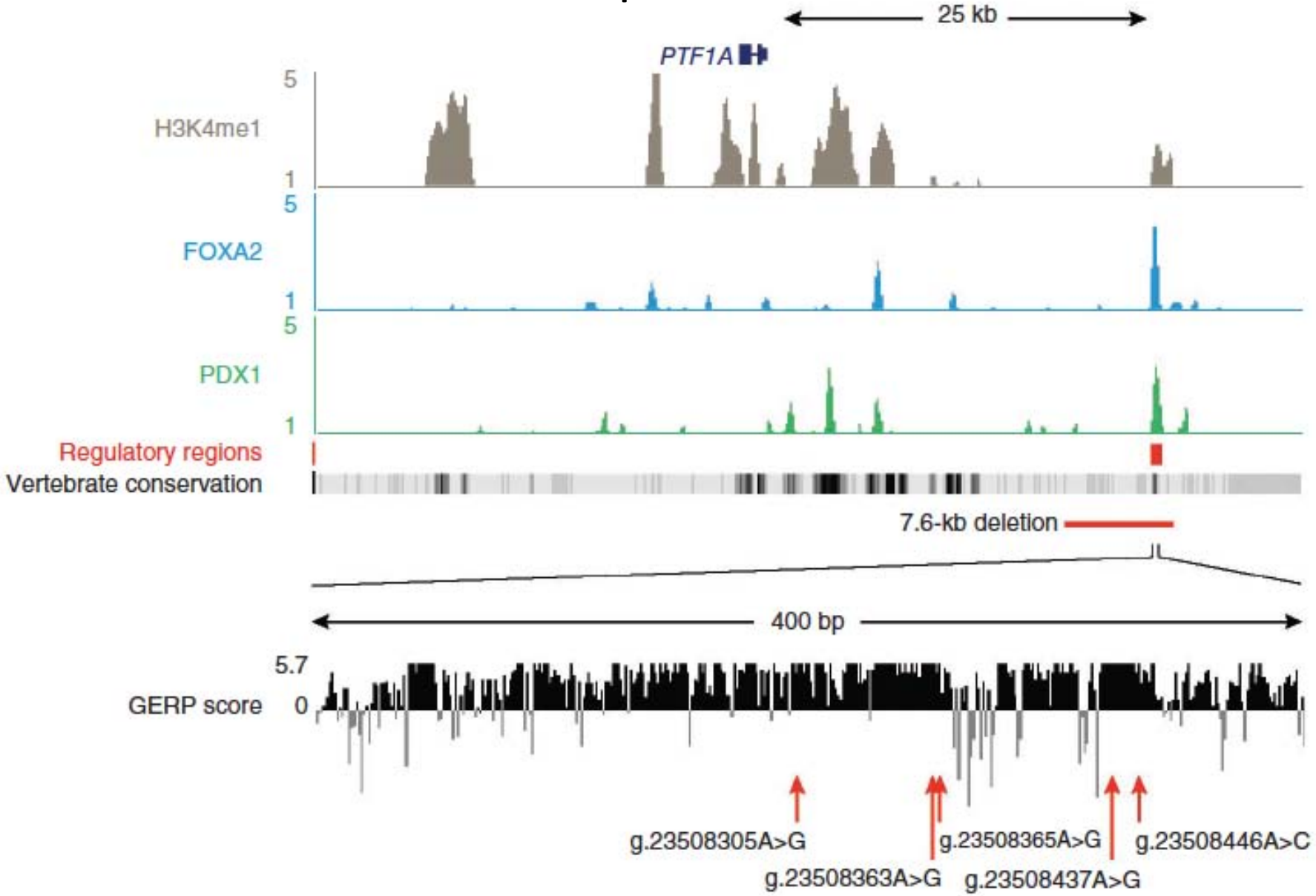
To identify recessive mutations causing isolated pancreatic agenesis, we used linkage and whole-genome sequencing analyses. Initially, we performed homozygosity mapping in six affected subjects and one unaffected subject from three unrelated consanguineous families (**Supplementary Fig. 1**). This analysis highlighted a single shared locus on chromosome 10 that included *PTF1A*, but mutations in the coding and promoter sequences of *PTF1A* and in the coding sequences of 24 other genes in the region were excluded by Sanger sequencing (**Supplementary Fig. 1** and **Supplementary Table 1**). We next performed whole-genome sequencing of probands from the two families with multiple affected individuals. We first looked for homozygous coding mutations in the exomes of the two individuals for whom whole-genome sequencing was performed. Each genome contained ~3.6 million variants, from which we filtered out any that were present in 81 control genomes or that were present at a frequency of >1% in 1000 Genomes Project data[8]. This filtering left 2,868 and 3,188 rare or newly identified homozygous single-nucleotide variants (SNVs) and indels per subject. Of these, 8 and 19 per subject were annotated as missense, nonsense, frameshift or essential splice site (**Supplementary Table 2**). However, these coding variants either did not cosegregate with disease or were not considered plausible candidates for having a role in pancreas development (**Supplementary Table 2**).

We next searched for noncoding disease-causing mutations among the remaining candidate homozygous variants. We reasoned that any causal variant should disrupt a noncoding genomic element that is active in cells that are relevant to this disease. As isolated pancreatic agenesis must be the result of a defect in early pancreas development, we determined whether any of the rare or newly identified

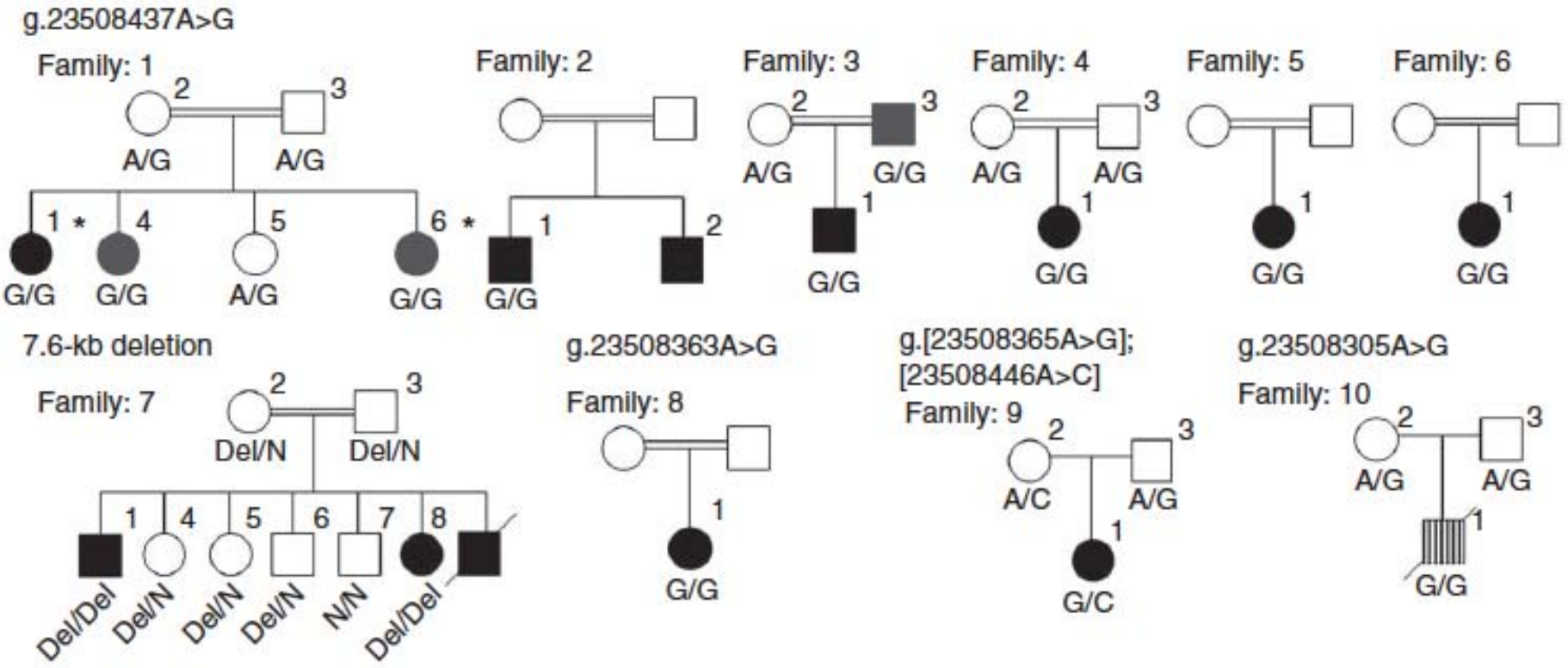
[1]Institute of Biomedical and Clinical Science, University of Exeter Medical School, Exeter, UK. [2]Genomic Regulation of Pancreatic Beta-Cells Laboratory, Institut d'Investigacions Biomèdiques August Pi i Sunyer, Barcelona, Spain. [3]Centro de Investigación Biomédica en Red de Diabetes y Enfermedades Metabólicas, Barcelona, Spain. [4]Department of Medicine, Imperial College, London, UK. [5]Wellcome Trust–Medical Research Council Cambridge Stem Cell Institute, Anne McLaren Laboratory for Regenerative Medicine, Cambridge, UK. [6]Seattle Children's Hospital Research Institute, Seattle, Washington, USA. [7]School of Biomedical Science, Waterloo Campus, King's College London, London, UK. [8]London Centre for Paediatric Endocrinology and Metabolism, in partnership with the Great Ormond Street Hospital for Children National Health Service Trust, London, UK. [9]Institute of Child Health, University College London, London, UK. [10]Diabetes Research Group, Diabetes and Nutritional Sciences Division, School of Medicine, King's College London, London, UK. [11]A full list of members and affiliations appears in the **Supplementary Note**. [12]These authors contributed equally to this work. [13]These authors jointly directed this work. Correspondence should be addressed to A.T.H. (a.t.hattersley@exeter.ac.uk) or J.F. (j.ferrer@imperial.ac.uk).

Source: Weedon, Michael N., Inês Cebola, et al. "Recessive Mutations in a Distal PTF1A Enhancer Cause Isolated Pancreatic Agenesis." *Nature Genetics* (2013).

Identification of possible functional variants



Association of variants with pedigrees

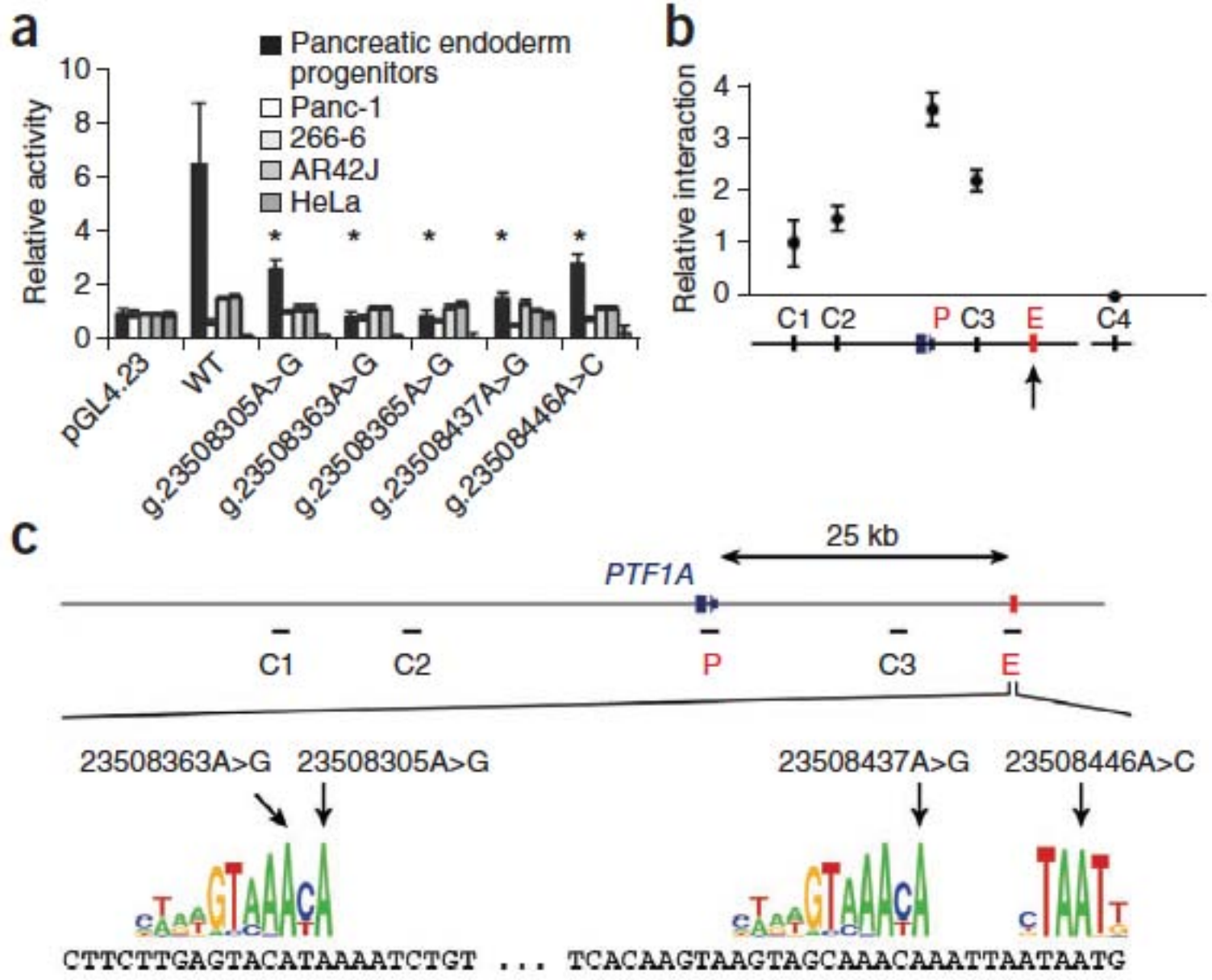


Courtesy of Macmillan Publishers Limited. Used with permission.
Source: Weedon, Michael N., Inês Cebola, et al. "Recessive Mutations in a Distal PTF1A Enhancer Cause Isolated Pancreatic Agenesis." *Nature Genetics* (2013).

Today's Narrative Arc

1. We can discover human variants that are associated with a phenotype by studying the genotypes of case and control populations
   - Approach 1 – Use allelic counts from SNP arrays (SNPs called from microarray data)
   - Approach 2 – Use read counts from sequencing (multiple reads per variant per individual)
2. We can prioritize variants based upon their estimated importance
3. **Follow up confirmation is important because correlation is not equivalent to causality**

Confirmation of variant function



Courtesy of Macmillan Publishers Limited. Used with permission.
Source: Weedon, Michael N., Inês Cebola, et al. "Recessive Mutations in a Distal PTF1A Enhancer Cause Isolated Pancreatic Agenesis." *Nature Genetics* (2013).

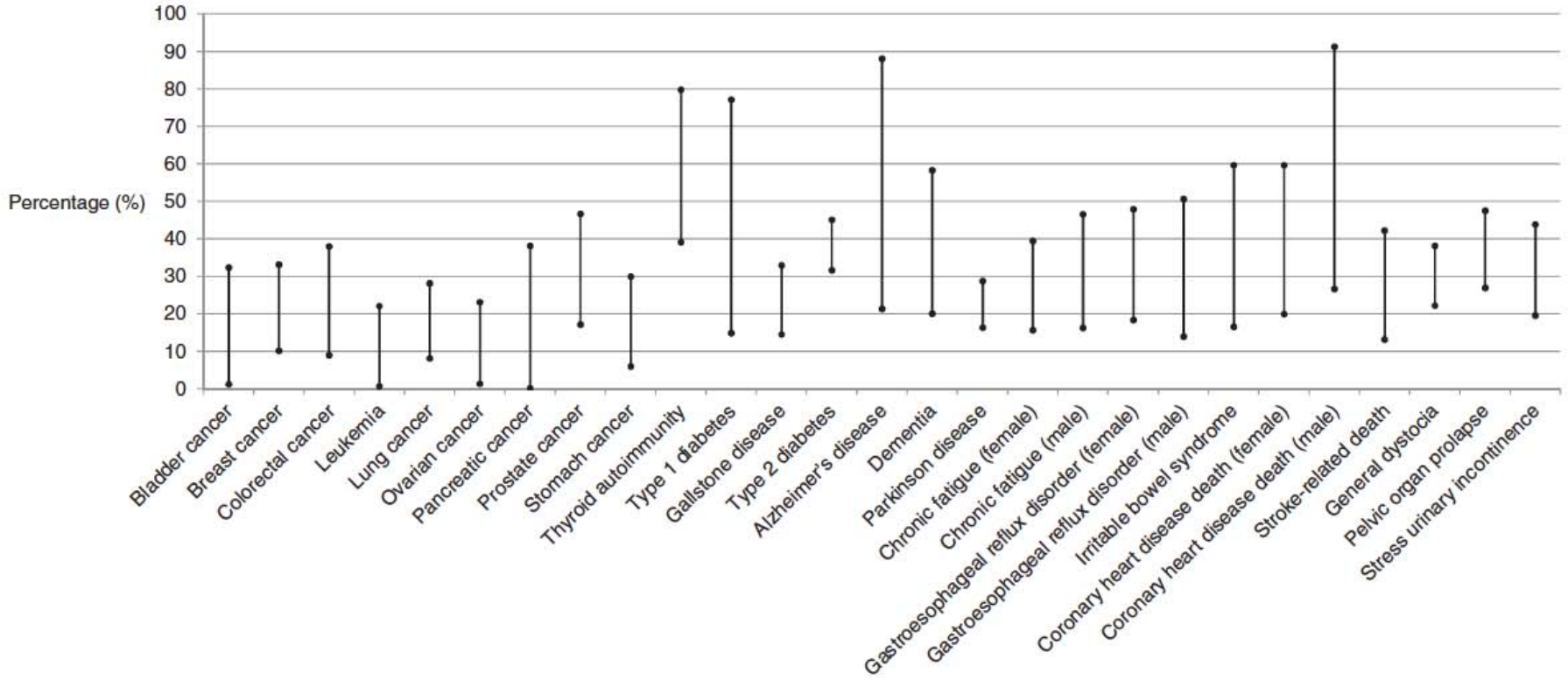


Fig. 1. The fraction of cases (that is, patients with disease) who would test positive by whole-genome sequencing. For each disease, the maximum and minimum fraction of cases that would test positive using the thresholds defined in table S1 are plotted.

Source: Roberts, Nicholas J., Joshua T. Vogelstein, et al. "The Predictive Capacity of Personal Genome Sequencing." *Science Translational Medicine* 4, no. 133 (2012): 133ra58.

Science Translational Medicine AAAS

The Predictive Capacity of Personal Genome Sequencing
Nicholas J. Roberts *et al.*
Sci Transl Med **4**, 133ra58 (2012);
DOI: 10.1126/scitranslmed.3003380

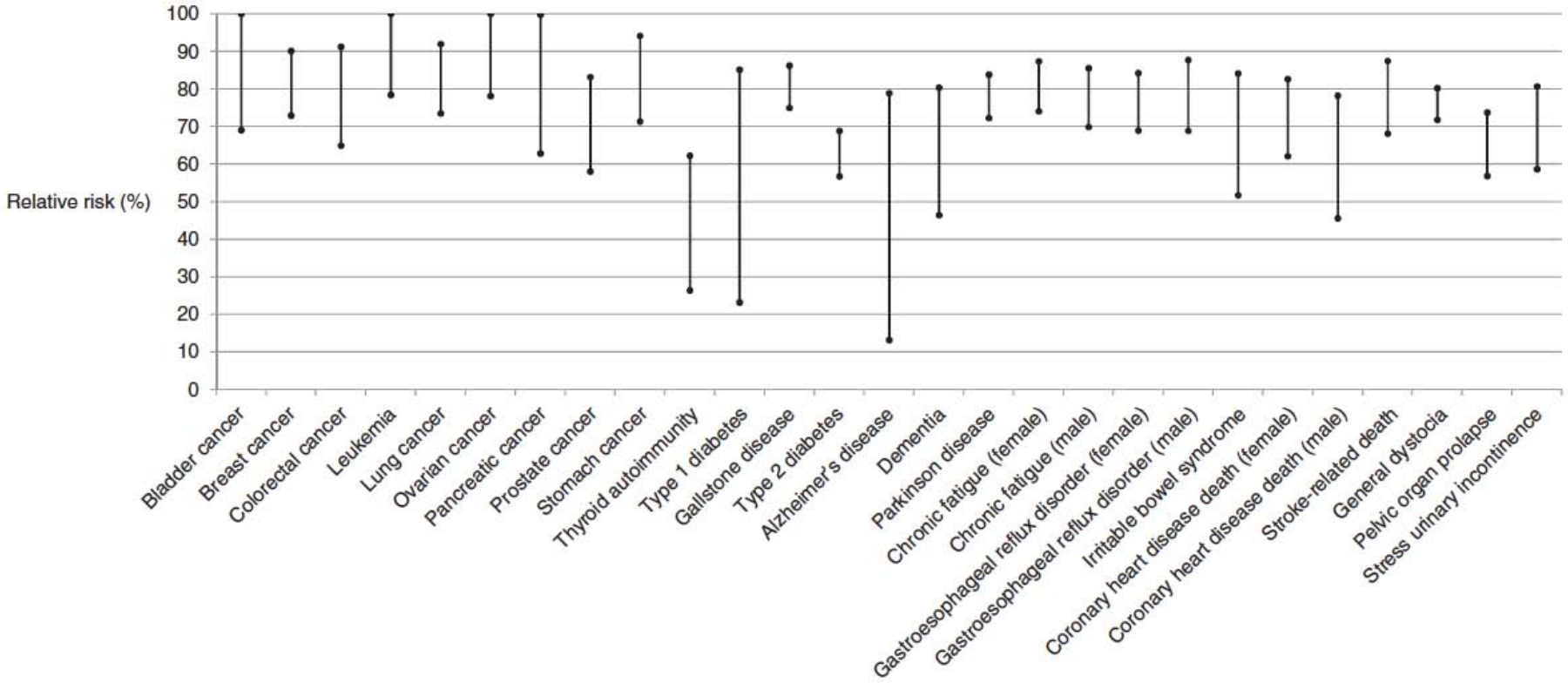


Fig. 3. Relative risk of disease in individuals testing negative by whole-genome sequencing. A relative risk of 100% represents the same risk as the general population, that is, the cohort risk. Relative risks were calculated using the genometype frequencies and genometype genetic risks that maximized or minimized sensitivity for disease detection.

Source: Roberts, Nicholas J., Joshua T. Vogelstein, et al. "The Predictive Capacity of Personal Genome Sequencing." *Science Translational Medicine* 4, no. 133 (2012): 133ra58.

Science Translational Medicine AAAS

The Predictive Capacity of Personal Genome Sequencing
Nicholas J. Roberts *et al.*
Sci Transl Med **4**, 133ra58 (2012);
DOI: 10.1126/scitranslmed.3003380

FIN

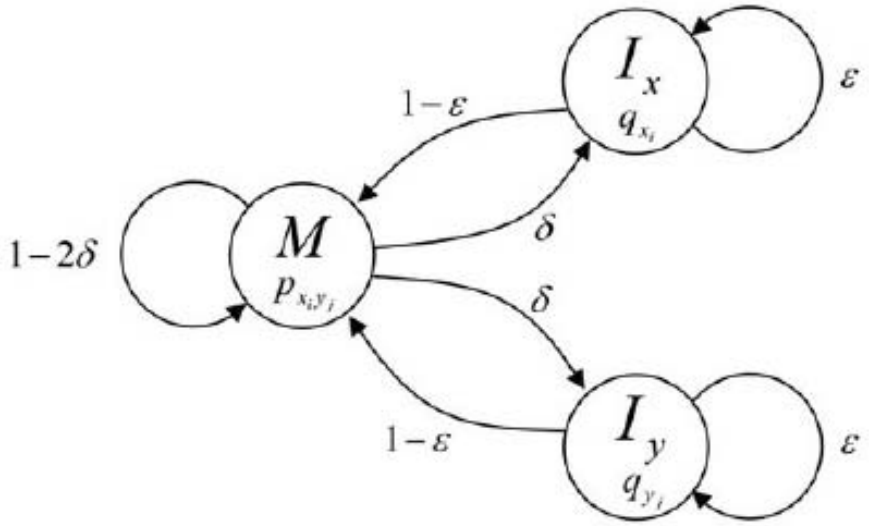
Evaluate haplotypes with Pair HMM

Bayesian model

$$\Pr\{G|D\} = \frac{\overbrace{\Pr\{G\}}^{\text{Prior of the genotype}}\overbrace{\Pr\{D|G\}}^{\text{Likelihood of the genotype}}}{\Sigma_i \Pr\{G_i\}\Pr\{D|G_i\}}, \text{ [Bayes' rule]}$$

$$\Pr\{D|G\} = \prod_j \left(\frac{\Pr\{D_j|H_1\}}{2} + \frac{\Pr\{D_j|H_2\}}{2}\right) \text{ where } \overbrace{G = H_1 H_2}^{\text{Diploid assumption}}$$

$\Pr\{D|H\}$ is the haploid likelihood function



Empirical gap penalties derived from data using new BQSR.

Base mismatch penalties are the base quality scores.

7.91J / 20.490J / 20.390J / 7.36J / 6.802J / 6.874J / HST.506J Foundations of Computational and Systems Biology

Spring 2014