Lecture 7: Genome-wide Association Study

Studies of Human Genetics

- Human genetic variants and sample sizes over past 20 years

Year	No. of Samples	No. of Markers	Publication
Ongoing	120,000	600 million	NHLBI Precision Medicine Cohorts / TopMed
2016	32,488	40 million	Haplotype Reference Consortium (Nature Genetics)
2015	2,500	80 million	The 1000 Genomes Project (Nature)
2012	1,092	40 million	The 1000 Genomes Project (Nature)
2010	179	16 million	The 1000 Genomes Project (Nature)
2010	100,184	2.5 million	Lipid GWAS (Nature)
2008	8,816	2.5 million	Lipid GWAS (Nature Genetics)
2007	270	3.1 million	HapMap (Nature)
2005	270	1 million	HapMap (Nature)
2003	80	10,000	Chr. 19 Variation Map (Nature Genetics)
2002	218	1,500	Chr. 22 Variation Map (Nature)
2001	800	127	Three Region Variation Map (Am J Hum Genet)
2000	820	26	T-cell receptor variation (Hum Mol Genet)

International HapMap Consortium, Nature, 2005, 437:1299–1320,
The HapMap project is a concerted effort to centralize this work and to provide the scientific community with a comprehensive LD-catalog for the entire human genome across four different ethnicities.

Samples:

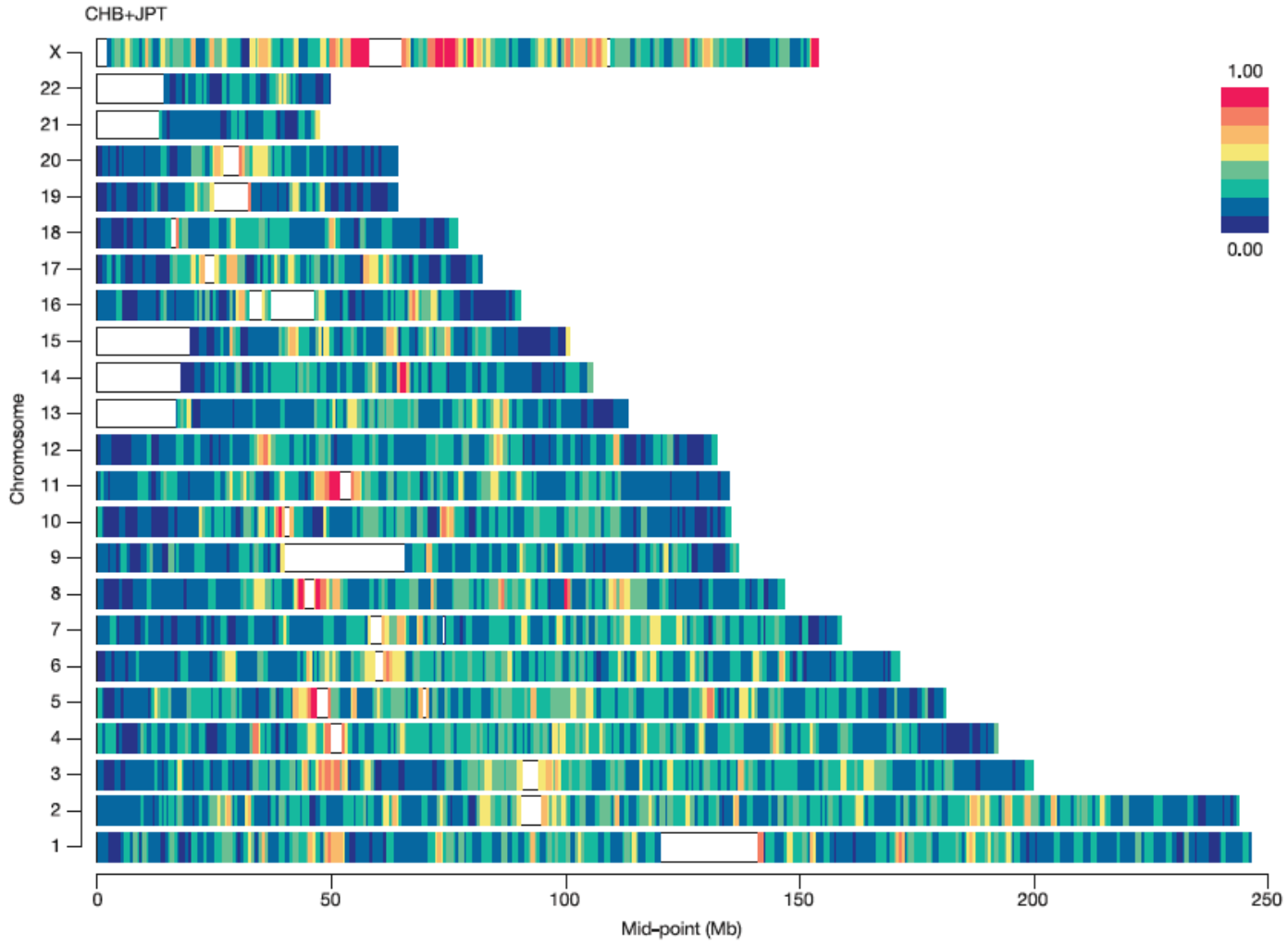
- (YRI) 30 parent-offspring trios from the Yoruba people in Ibadan, Nigeria
- (CEU) 30 CEPH trios (largely Western European background)
- (JPT) 45 unrelated Japanese from Tokyo
- (CHB) 45 unrelated Han Chinese from Beijing
- Additional samples from other populations have been added.

Accomplishments

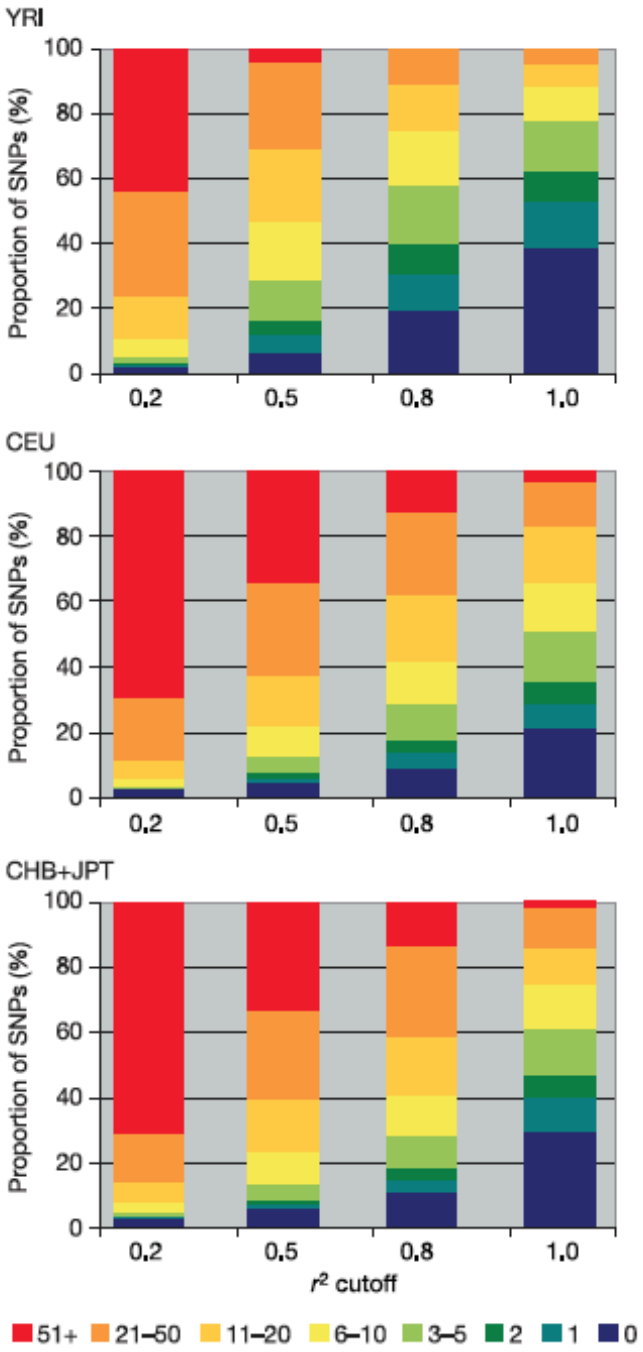
- Capture sites of common variation in the human genome
- Generate genotypes on $>$ 3 million SNPs on 269 individuals for LD estimation
- Store the genotypes in a public database
- Develop many new technologies for probing the genome

HapMap: Genome-Wide LD Patterns

– Average LD r^2 presented here for CHB+JPT in 1Mb grid
– Highest LD on X chromosome and near centromeres
– Lowest LD in telomeres



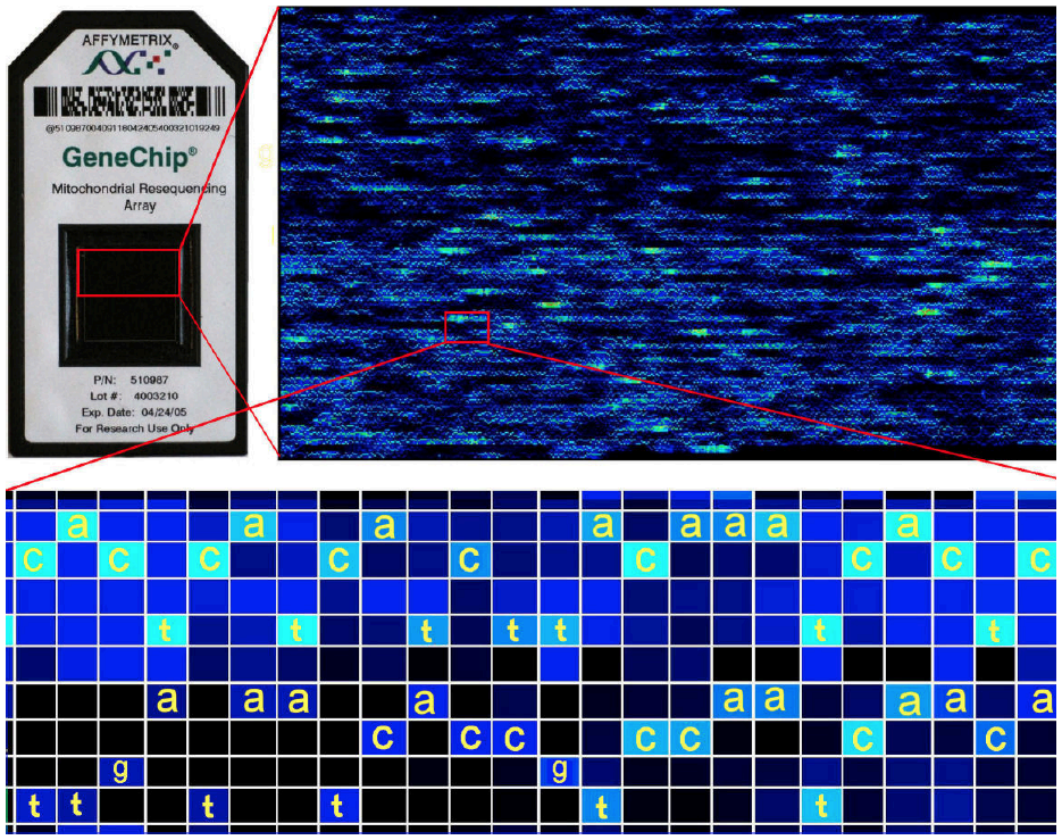
HapMap: Proxy SNPs

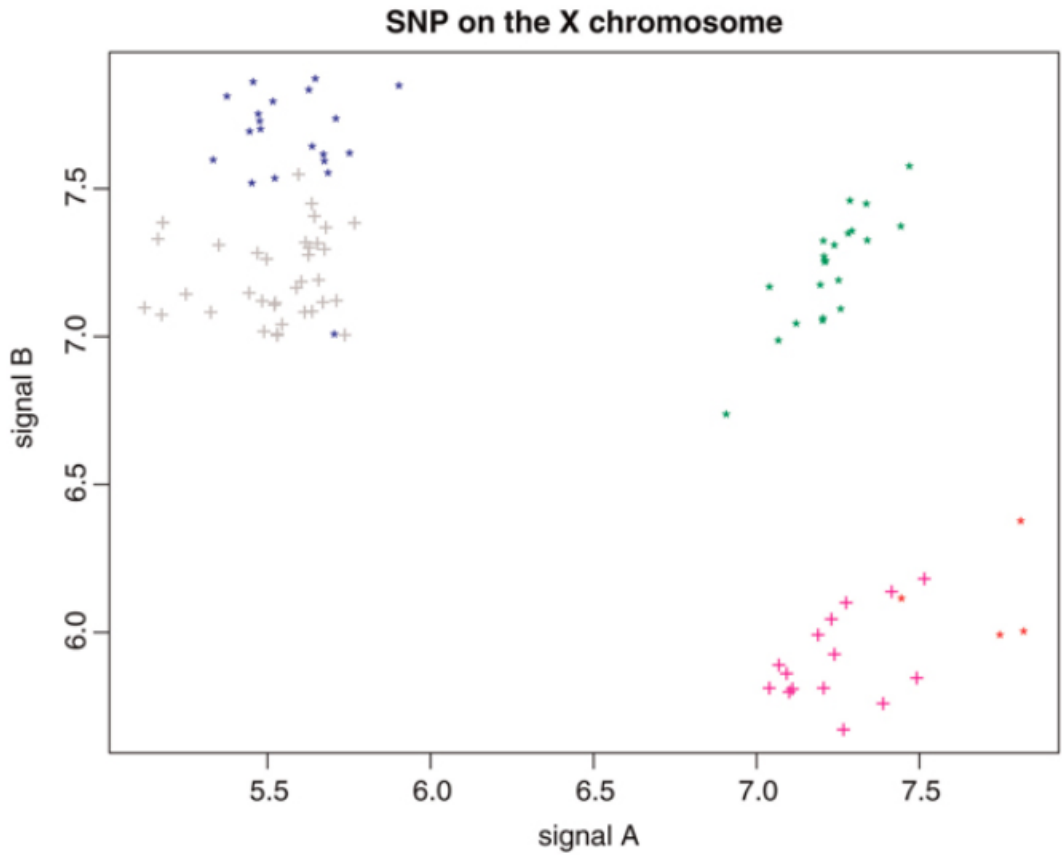


- The number of proxies per SNP as a function of the threshold for squared-correlation (r^2, quantifies LD).
- Proportions of SNPs that have $0, 1, 2, \cdots$ SNPs (redundant SNPs) in high LD
  - Highest redundancy in CEU.
  - Least redundancy in YRI.

DNA Microarrays (CHIP)

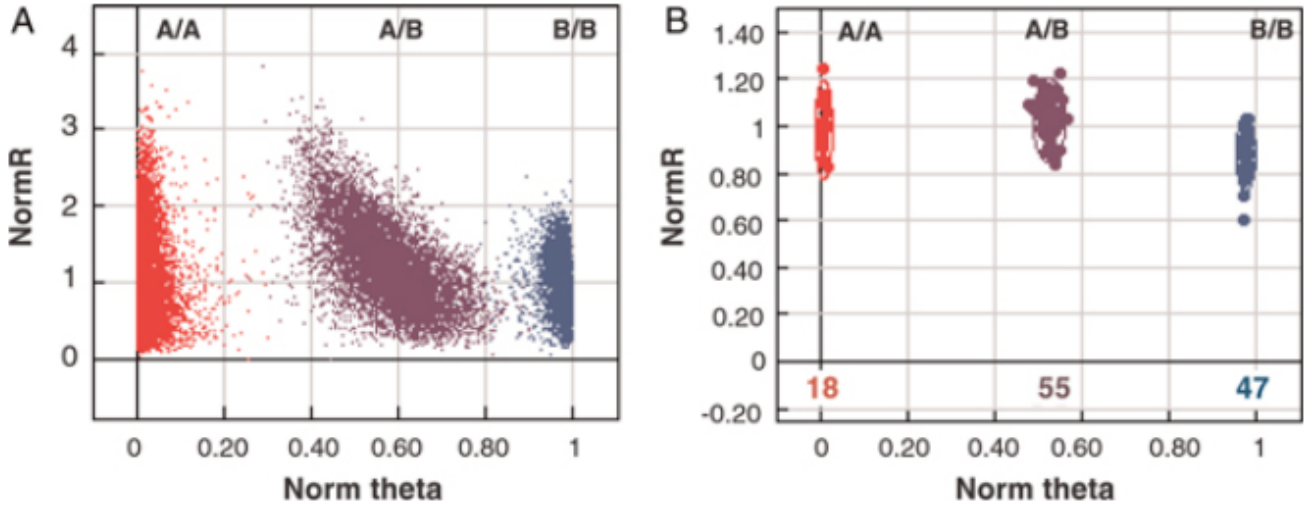
- Microarrays (Illumina and Affymetrix) are used to genotype $0.5M - 1M$ SNPs genome-wide
- LD-information of the HapMap project has been incorporated so that the chips provide adequate coverage of the entire human genome for most ethnicities.
- Customize chips with densely spaced SNPs within known genes regions





Normalised and summarised allele intensities from the Affymetrix GeneChip array. Each SNP is represented by a pair of intensity values (A, B) for the A and B alleles, respectively (here, on a log-scale). An X chromosome SNP is shown, clearly indicating separation into distinct genotype clusters. The plot also shows that different copy numbers can be distinguished. Males are haploid for the particular SNP (ie either AY or BY) and show up as homozygous but with reduced allele intensity. Grey: BY; blue: BB; green: AB; red: AA; and pink: AY.

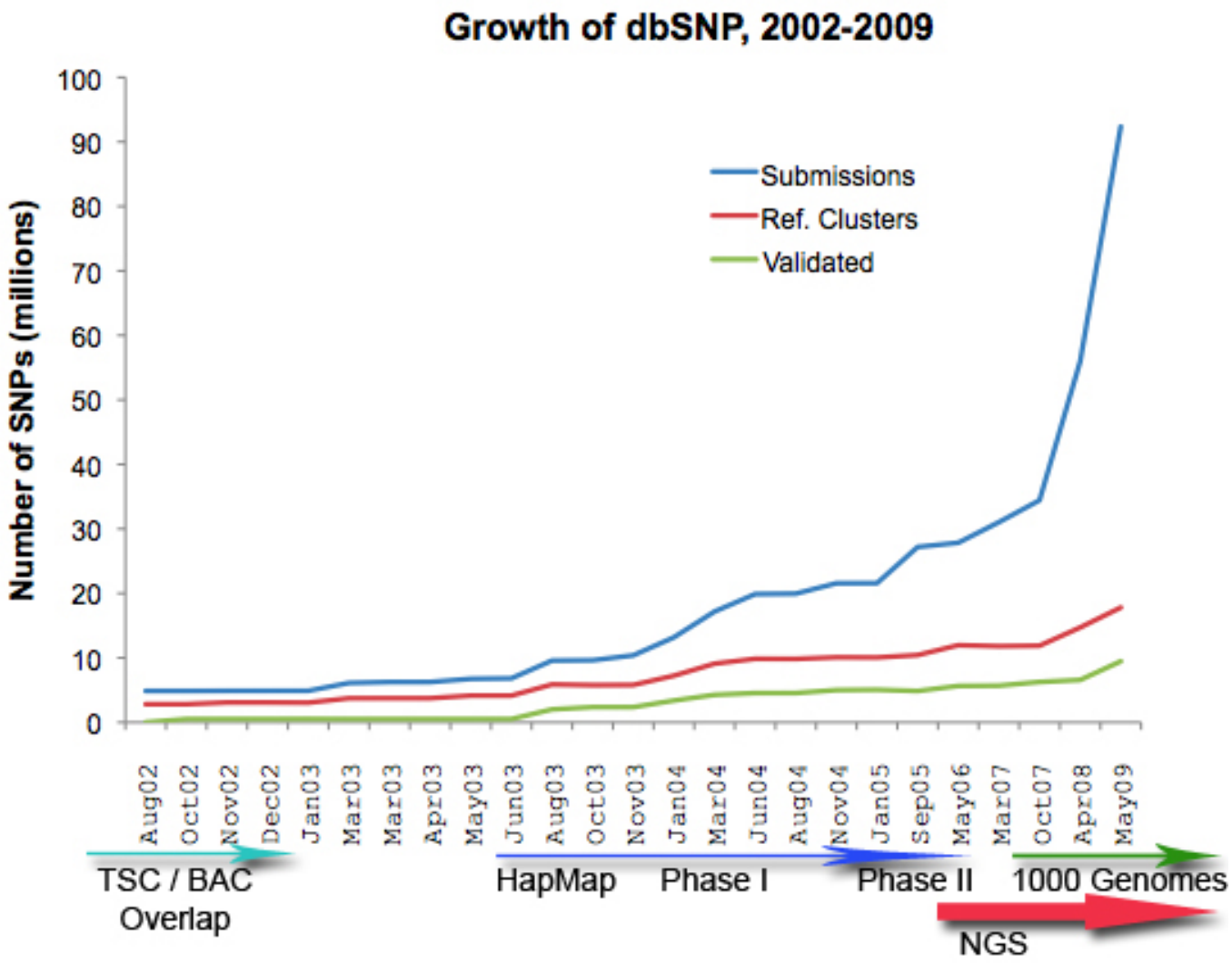
doi:10.1186/1479-7364-5-4-304



Normalised and summarised allele intensities from the Illumina BeadChip array. The intensities are shown in transformed polar coordinates: the theta-coordinate represents the angle from the x-axis (the angle from the x-axis to the vector [A, B] of the two allele intensities), and the R-coordinate represents the copy number (the length of the vector). (A) Intensities for a single nucleotide polymorphism (SNP) from 120 arrays, clearly separating the intensities into three groups (A/A, A/B, B/B). (B) Data from 317,000 SNPs (from the same 120 arrays). This plot clearly indicates that signal strength varies considerably with the SNP, a factor that must be taken into account when genotyping individual SNPs and deriving copy numbers. The figure is reproduced with the permission of Gunderson *et al.* [15]

doi:10.1186/1479-7364-5-4-304

Growth of dbSNP

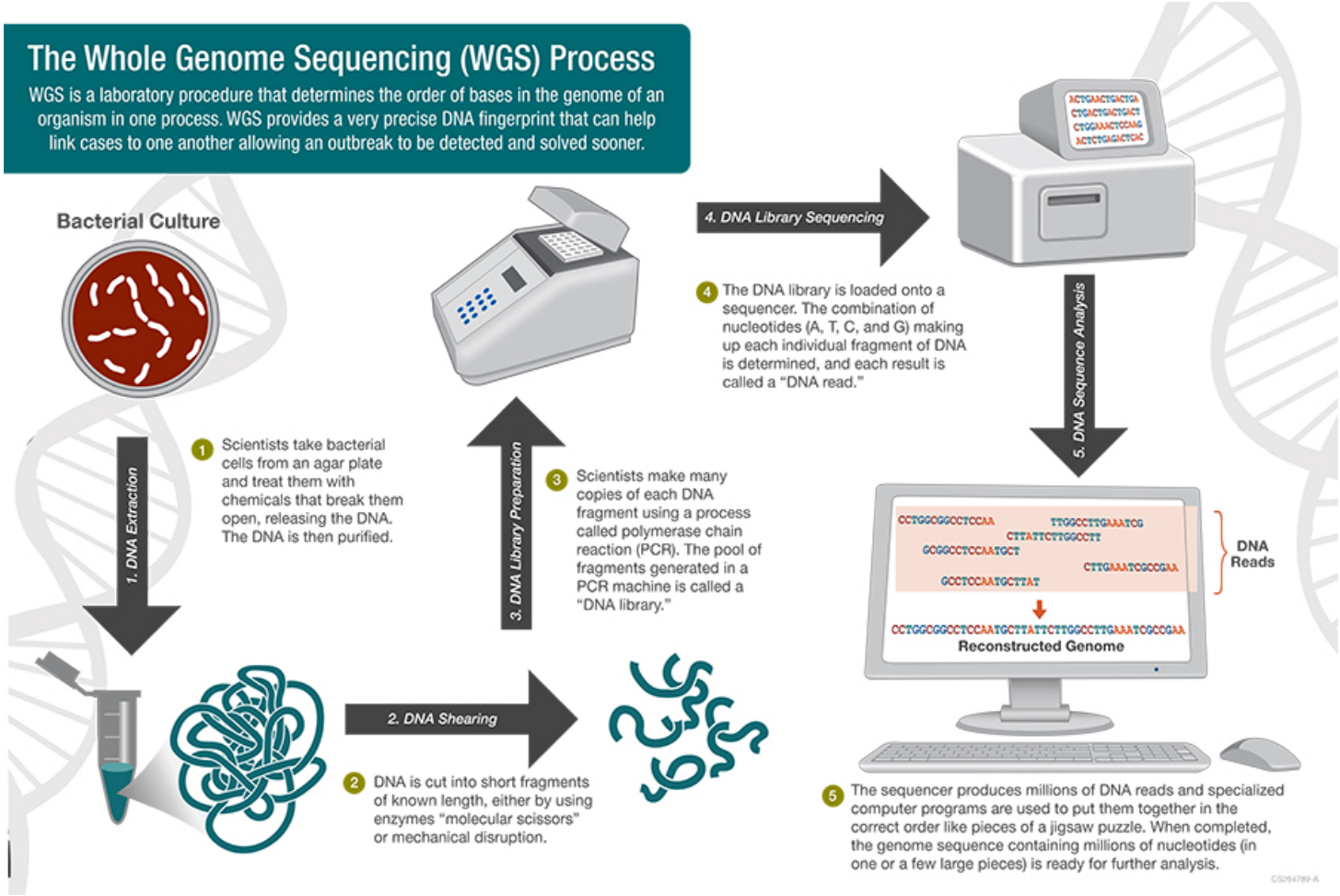


NGS: Next-generation Sequencing

Whole Genome Sequencing

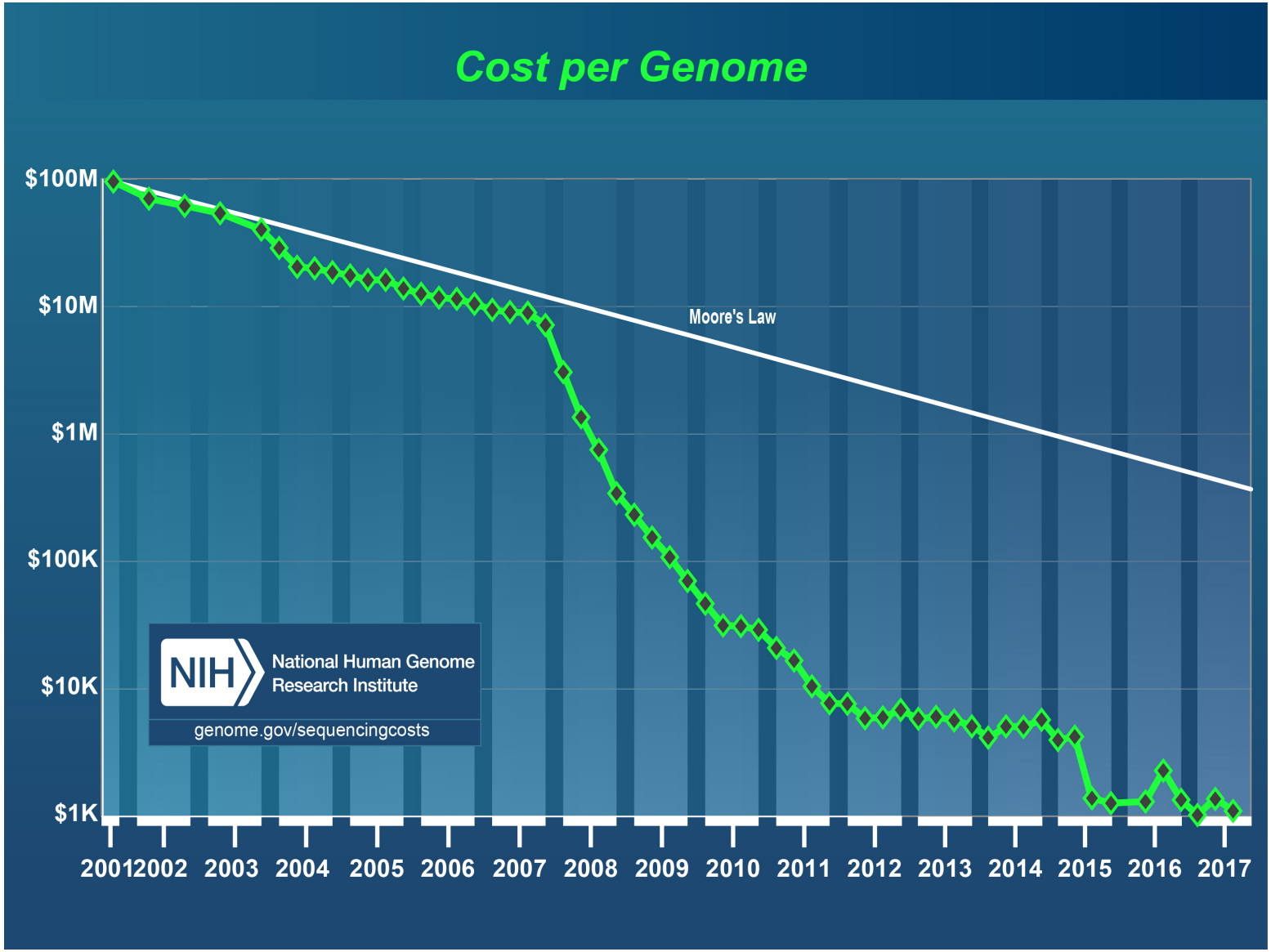
Introduction video of Illumina Sequencing technology

https://www.youtube.com/watch?v=fCd6B5HRaZ8



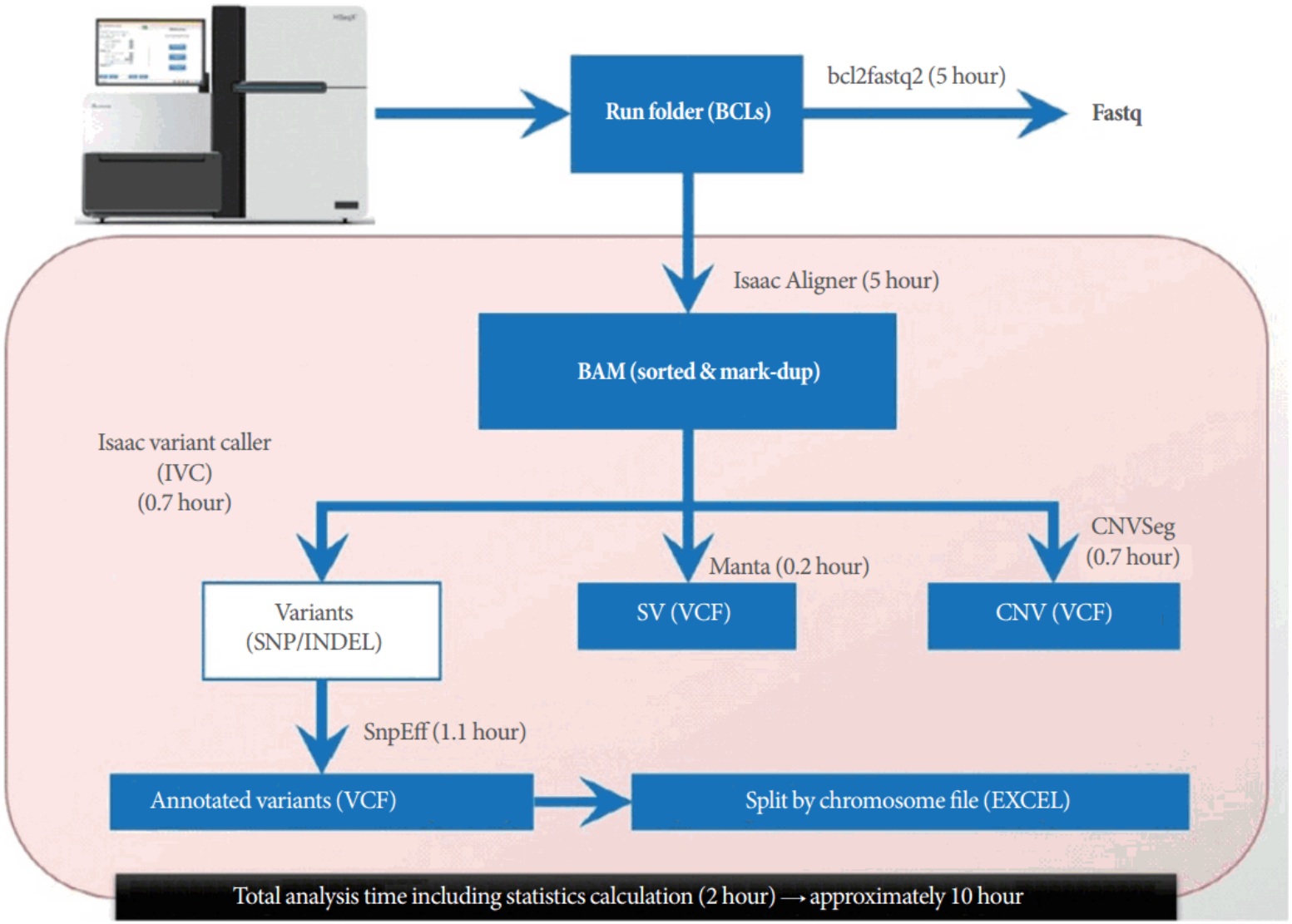
https://www.cdc.gov/pulsenet/pathogens/wgs.html

21st Century Sequencing Costs



http://genome.gov/sequencingcostsdata

WGS Analysis Workflow

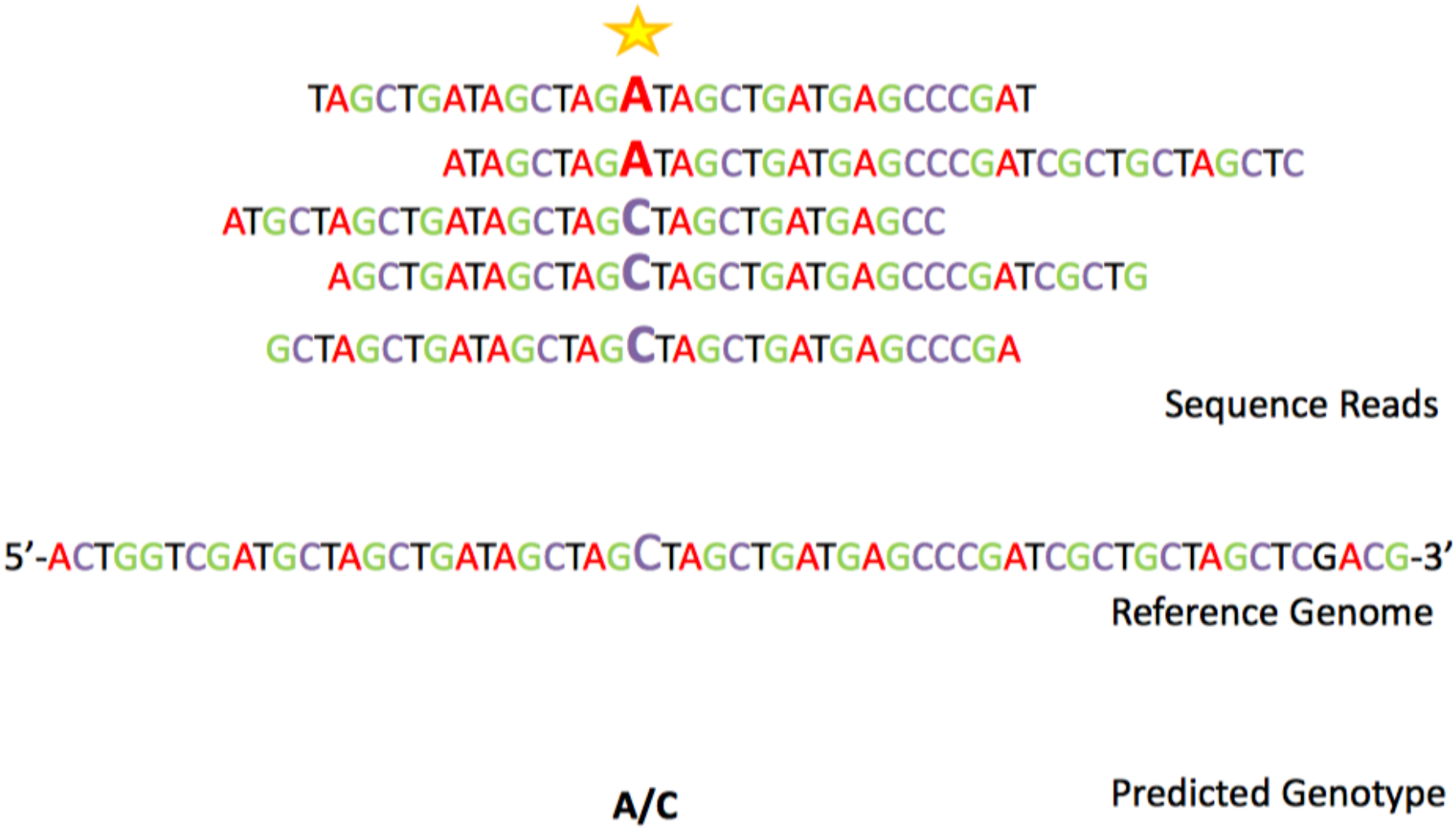


DOI:10.5213/inj.1632742.371

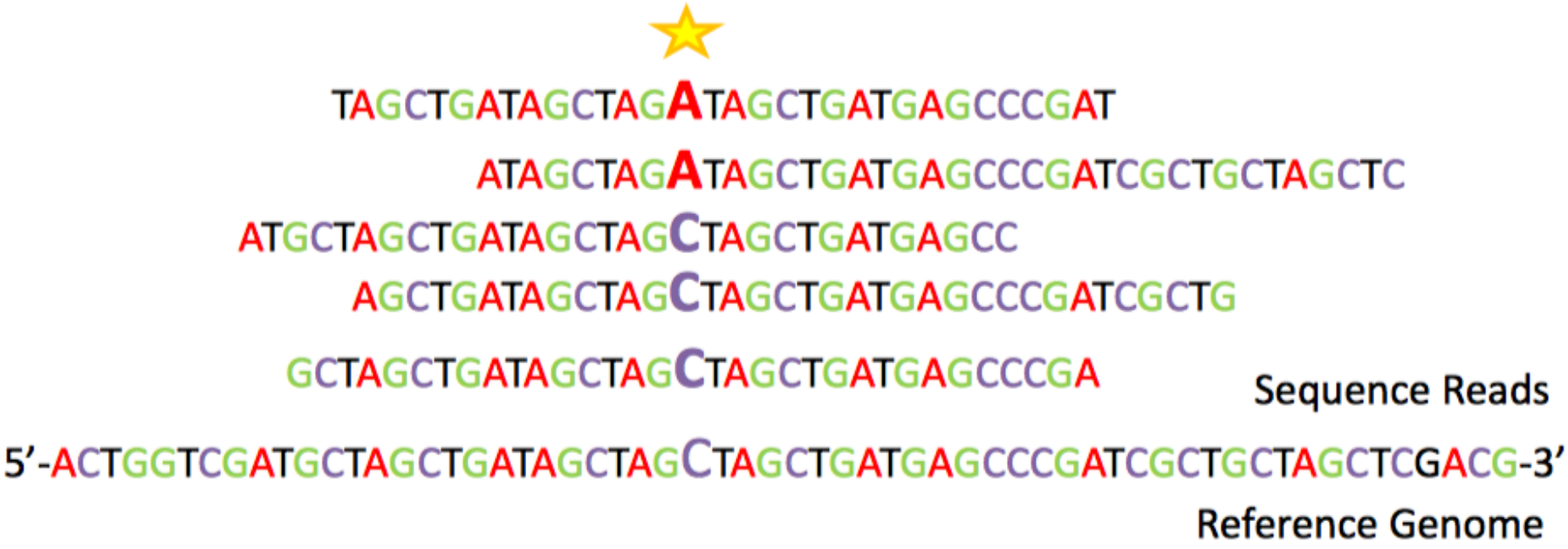
Shotgun Sequence Reads

- Typical short read might be <50-150 bp long and not very informative on its own
- Reads must be arranged (*aligned*) relative to each other to reconstruct longer sequences

From Sequence to Genotype



From Sequence to Genotype



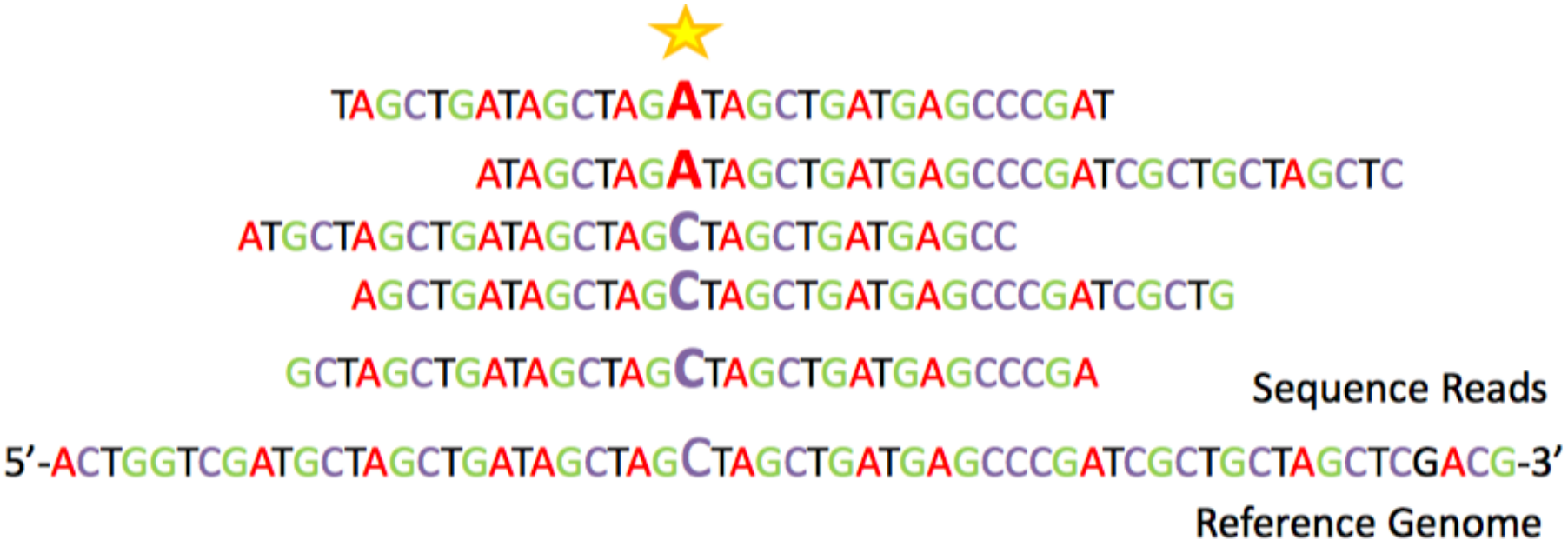
$$P(Genotype|reads) = \frac{P(reads|Genotype)Prior(Genotype)}{\sum_G P(reads|G)Prior(G)}$$

Combine these likelihoods with a prior incorporating information from other individuals and flanking sites to assign a genotype.

Ingredients That Go Into Prior

- Most sites don't vary
  - P(non-reference base) ~ 0.001
- When a site does vary, it is usually heterozygous
  - P(non-reference heterozygote) ~ 0.001 * 2/3
  - P(non-reference homozygote) ~ 0.001 * 1/3
- Mutation model
  - Transitions account for most variants (C↔T or A↔G)
  - Transversions account for minority of variants

From Sequence to Genotype



P(reads|A/A)= 0.00000098 **Prior(A/A)** = 0.00034 **Posterior(A/A)** = <.001

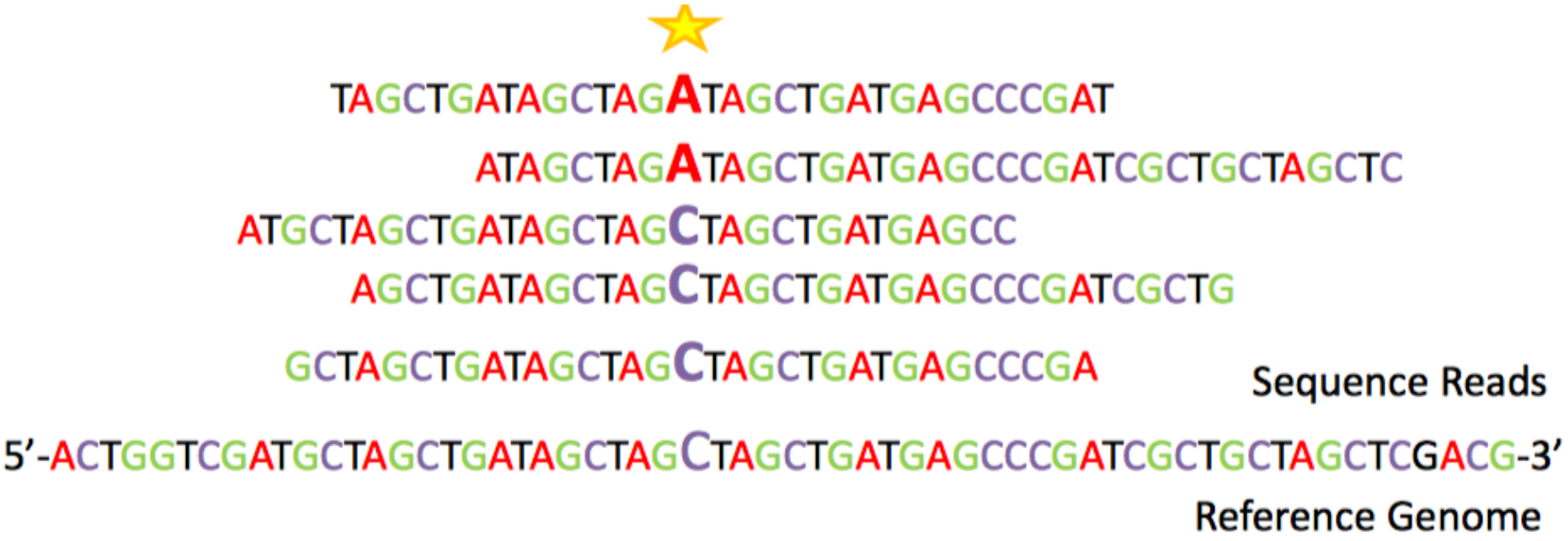
P(reads|A/C)= 0.03125 **Prior(A/C)** = 0.00066 **Posterior(A/C)** = 0.175

P(reads|C/C)= 0.000097 **Prior(C/C)** = 0.99900 **Posterior(C/C)** = 0.825

Individual Based Prior: Every site has 1/1000 probability of varying.

Assume sequence error rate 0.01

From Sequence to Genotype



P(reads|A/A)= 0.00000098 **Prior(A/A)** = 0.04 **Posterior(A/A)** = <.001

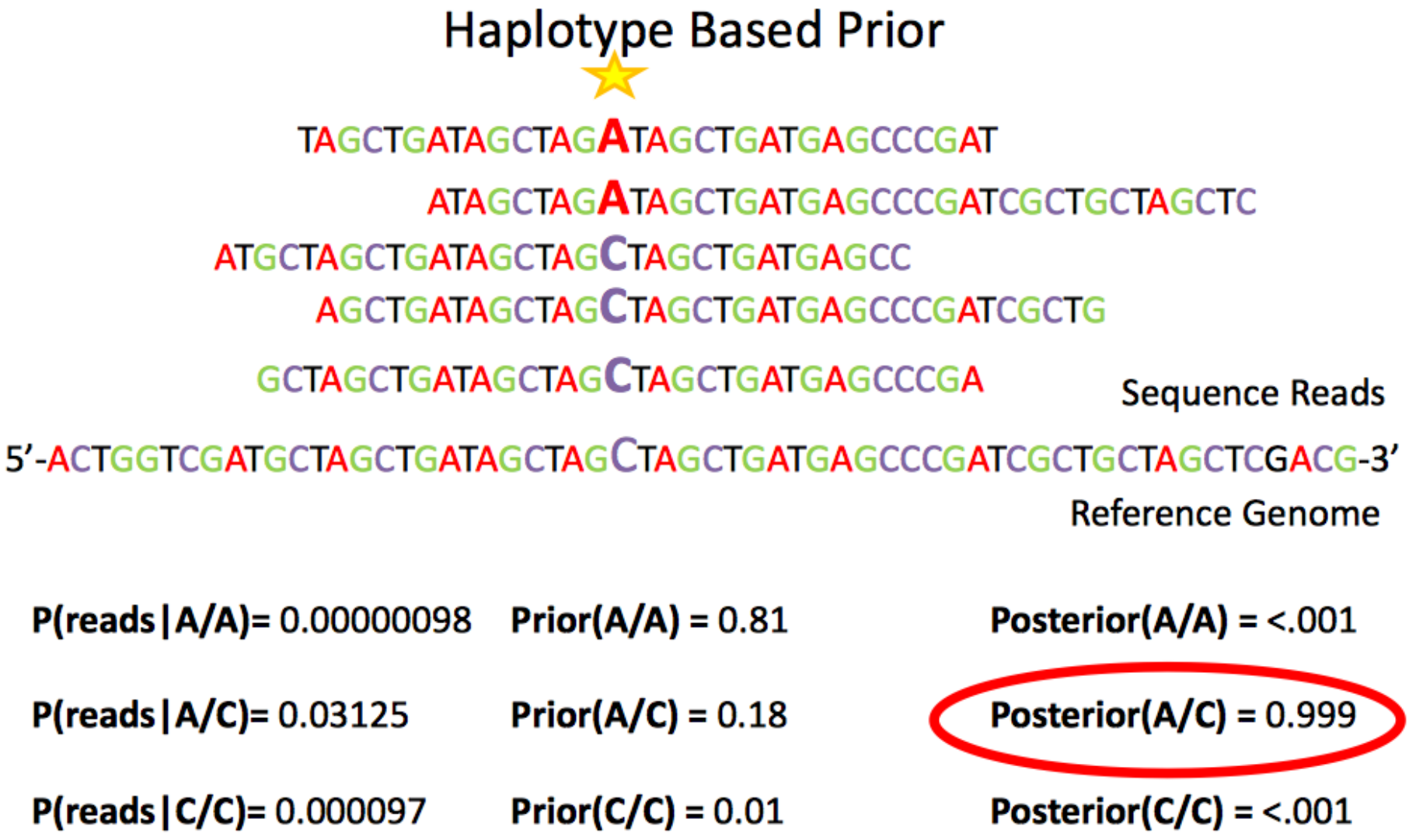
P(reads|A/C)= 0.03125 **Prior(A/C)** = 0.32 **Posterior(A/C)** = 0.999

P(reads|C/C)= 0.000097 **Prior(C/C)** = 0.64 **Posterior(C/C)** = <.001

Population Based Prior: Use frequency information from examining others at the same site.

In the example above, we estimated P(A) = 0.20

From Sequence to Genotype



Haplotype Based Prior: Examine other chromosomes that are similar at locus of interest.

In the example above, we estimated that 90% of similar chromosomes carry allele A.

From Sequence to Genotype

- **Individual Based Prior**
  - Assumes all sites have an equal probability of showing polymorphism
  - Specifically, assumption is that about 1/1000 bases differ from reference
  - If reads where error free and sampling Poisson …
  - … 14x coverage would allow for 99.8% genotype accuracy
  - … 30x coverage of the genome needed to allow for errors and clustering

- **Population Based Prior**
  - Uses frequency information obtained from examining other individuals
  - Calling very rare polymorphisms still requires 20-30x coverage of the genome
  - Calling common polymorphisms requires much less data

- **Haplotype Based Prior or Imputation Based Analysis**
  - Compares individuals with similar flanking haplotypes
  - Calling very rare polymorphisms still requires 20-30x coverage of the genome
  - Can make accurate genotype calls with 2-4x coverage of the genome
  - Accuracy improves as more individuals are sequenced

1000 Genome Project 2008 - 2015

GOAL: Find most genetic variants with MAF $\geq 1\%$ in populations across the world.

- First project to sequence the genomes of a large number of people (2,504 samples)
- Largest public catalogue of human variation and genotype data
- http://www.internationalgenome.org/
- 26 Different populations under 5 super populations

  - **AFR**, African
  - **AMR**, Ad Mixed American
  - **EAS**, East Asian
  - **EUR**, European
  - **SAS**, South Asian

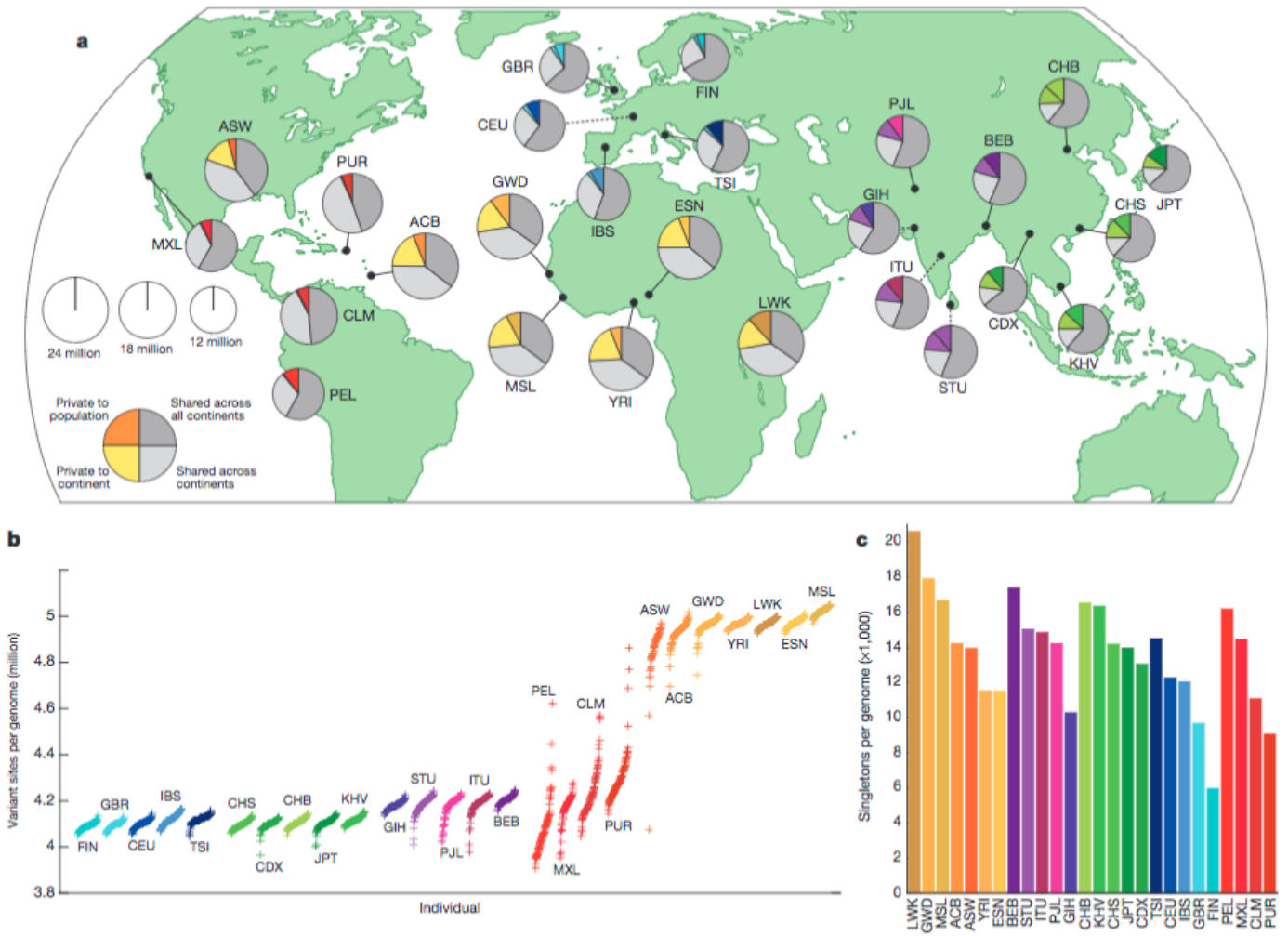


Figure 1 | Population sampling. **a**, Polymorphic variants within sampled populations. The area of each pie is proportional to the number of polymorphisms within a population. Pies are divided into four slices, representing variants private to a population (darker colour unique to population), private to a continental area (lighter colour shared across continental group), shared across continental areas (light grey), and shared across all continents (dark grey). Dashed lines indicate populations sampled outside of their ancestral continental region. **b**, The number of variant sites per genome. **c**, The average number of singletons per genome.

The 1000 Genomes Project Consortium. Nature volume 526, pages 68-74 (2015).

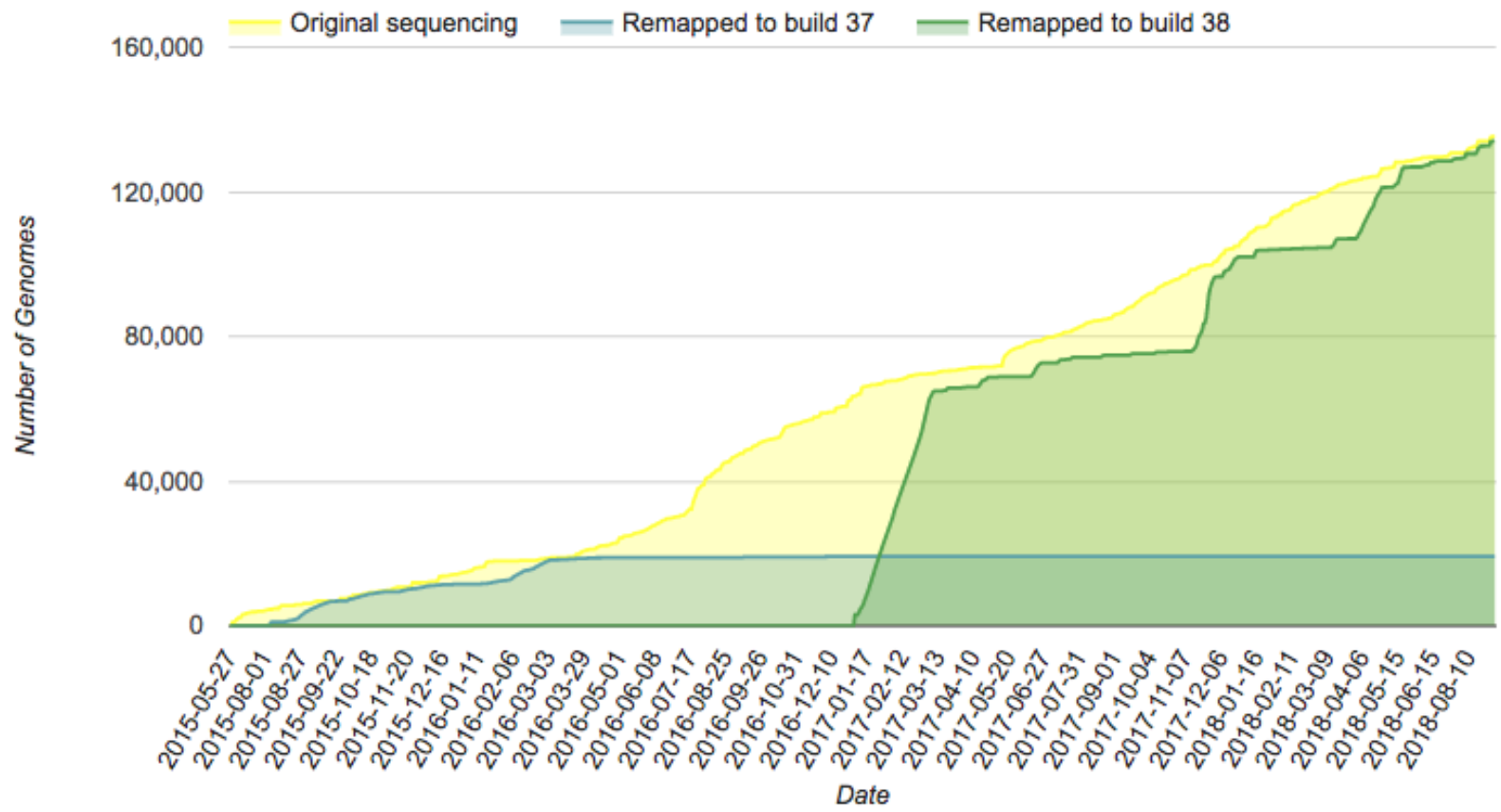
Population Code	Population Description	Super Population Code
CHB	Han Chinese in Beijing, China	EAS
JPT	Japanese in Tokyo, Japan	EAS
CHS	Southern Han Chinese	EAS
CDX	Chinese Dai in Xishuangbanna, China	EAS
KHV	Kinh in Ho Chi Minh City, Vietnam	EAS
CEU	Utah Residents (CEPH) with Northern and Western European Ancestry	EUR
TSI	Toscani in Italia	EUR
FIN	Finnish in Finland	EUR
GBR	British in England and Scotland	EUR
IBS	Iberian Population in Spain	EUR
YRI	Yoruba in Ibadan, Nigeria	AFR
LWK	Luhya in Webuye, Kenya	AFR
GWD	Gambian in Western Divisions in the Gambia	AFR
MSL	Mende in Sierra Leone	AFR
ESN	Esan in Nigeria	AFR
ASW	Americans of African Ancestry in SW USA	AFR
ACB	African Caribbeans in Barbados	AFR
MXL	Mexican Ancestry from Los Angeles USA	AMR
PUR	Puerto Ricans from Puerto Rico	AMR
CLM	Colombians from Medellin, Colombia	AMR
PEL	Peruvians from Lima, Peru	AMR
GIH	Gujarati Indian from Houston, Texas	SAS
PJL	Punjabi from Lahore, Pakistan	SAS
BEB	Bengali from Bangladesh	SAS
STU	Sri Lankan Tamil from the UK	SAS
ITU	Indian Telugu from the UK	SAS

TOPMed WGS Project 2015 - Present

- NIH National Heart, Lung, and Blood Institute (NHLBI) sponsored the Trans-Omics for Precision Medicine (TOPMed) program https://www.nhlbiwgs.org/
- Deep (30x coverage) whole genome sequencing for all of the collected samples from ongoing disease-specific research projects
- WGS data are generated by seven sequencing centers
- University of Washington group is designated as the Data Coordinating Center (DCC) and will coordinate phenotype information
- University of Michigan group is designated as the Informatics Research Center (IRC) with responsibility for creating a unified variant call set
- The sequence and genotype data will be deposited to dbGaP https://www.ncbi.nlm.nih.gov/gap

Summary of total sequencing progress over time

This chart shows a summary of the total genomes received by IRC over time



Samples that have completed QC: 135,235 (as of 2/4/2019).

10^{16} sequenced bases, $100\times$ more data than the 1000 Genome Project.

http://nhlbi.sph.umich.edu

10^{16} sequenced bases

US corn production in 2014: 1.3 x 10^{15} kernels

Image: Patrick Porter @ Smug Mug

How Much Variation is There?

1000 Genome Project / TOPMed Project

Type	Variant sites / genome
SNPs	~3,800,000
Indels	~570,000
Mobile Element Insertions	~1000
Large Deletions	~1000
CNVs	~150
Inversions	~11

Variant Type	Category	# PASS	# FAIL	% dbSNP (PASS)	Known/Novel Ts/Tv (PASS)
SNP	**All**	**438M**	**85M**	**22.9%**	**1.93 / 1.69**
	Singleton	202M	24M	8.5%	1.23 / 1.54
	Doubleton	69M	8.8M	12.6%	1.61 / 1.74
	Tripleton ~ 0.1%	142M	24M	34.9%	2.23 / 1.99
	0.1% ~ 1%	13M	4.5M	98.2%	2.17 / 1.79
	1 ~ 10%	6.5M	2.9M	99.6%	1.82 / 1.75
	>10%	5.3M	2.0M	99.8%	2.11 / 1.88
Indels	**All**	**33.4M**	**26.2M**	**20.1%**	
	Singleton	15.7M	4.7M	10.1%	
	Doubleton	5.3M	1.8M	12.6%	
	Tripleton ~ 0.1%	10.7M	8.0M	26.7%	
	0.1% ~ 1%	2.8M	968K	88.9%	
	1 ~ 10%	432K	2.3M	98.5%	
	>10%	298K	1.4M	99.6%	

Genome-Wide Association Study (GWAS)

GWAS: independent single-variant tests across all genome-wide variants

- Quality control (QC) of the study dataset
- Choose a model/test for the phenotype of interest (e.g., linear regression model for quantitative traits, logistic regression model for dichotomous traits, other association tests from previous lecture)
- Significance level $\alpha = 5 \times 10^{-8}$
- Report nearby genes of significant SNPs

Genotype Quality

Data quality is one of the key factors affecting the validity of findings.

Example factors affecting genotype quality:

- Quality of DNA samples, depending on the sample source (e.g., blood, buccal swab, spit kit)
- Handling and storage of the sample (e.g., sample contamination)
- Genotyping platforms/chips
- Sequence errors
- Variant calling

Quality Control (QC)

- Filter SNPs
  - Marker genotyping missing rate (e.g., $> 2\%$)
  - Mapping quality for sequence data (based on mapping quality scores)
  - Hardy-Weinberg Equilibrium (HWE) Testing (e.g., p-value $< 10^{-6}$)
  - MAF (e.g., $< 5\%$)
  - Control sample reproducibility
  - Mendelian Errors (e.g., $> 1\%$ families, or > 5 errors) for family-based studies
- Filter samples
  - Sex inconsistencies and chromosomal anomalies
  - Relatedness for population-based studies (how to quantify relatedness given genotype data?)
  - Ethnicity
  - Sample genotyping efficiency/call rate (e.g., $< 98\%$)

Check Relatedness by PLINK

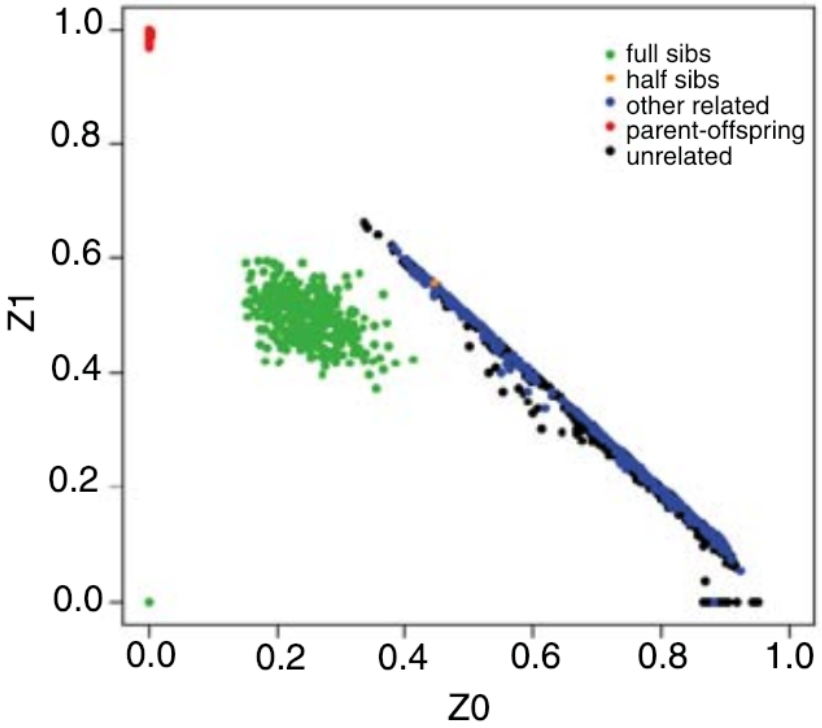
Identity-by-descent

These calculations are not LD-sensitive. It is usually a good idea to perform some form of LD-based pruning before invoking them.

```
--genome <gz> <rel-check> <full> <unbounded> <nudge>
--ppc-gap [distance in kbs]
--min [minimum PI_HAT value]
--max [maximum PI_HAT value]
```

--genome invokes an IBS/IBD computation, and then writes a report with the following fields to plink.genome:

FID1	Family ID for first sample
IID1	Individual ID for first sample
FID2	Family ID for second sample
IID2	Individual ID for second sample
RT	Relationship type inferred from .fam/.ped file
EZ	IBD sharing expected value, based on just .fam/.ped relationship
Z0	P(IBD=0)
Z1	P(IBD=1)
Z2	P(IBD=2)
PI_HAT	Proportion IBD, i.e. P(IBD=2) + 0.5*P(IBD=1)
PHE	Pairwise phenotypic code (1, 0, -1 = AA, AU, and UU pairs, respectively)
DST	IBS distance, i.e. (IBS2 + 0.5*IBS1) / (IBS0 + IBS1 + IBS2)
PPC	IBS binomial test
RATIO	HETHET : IBS0 SNP ratio (expected value 2)



Additional Factors Important for GWAS: batch effects, population stratification

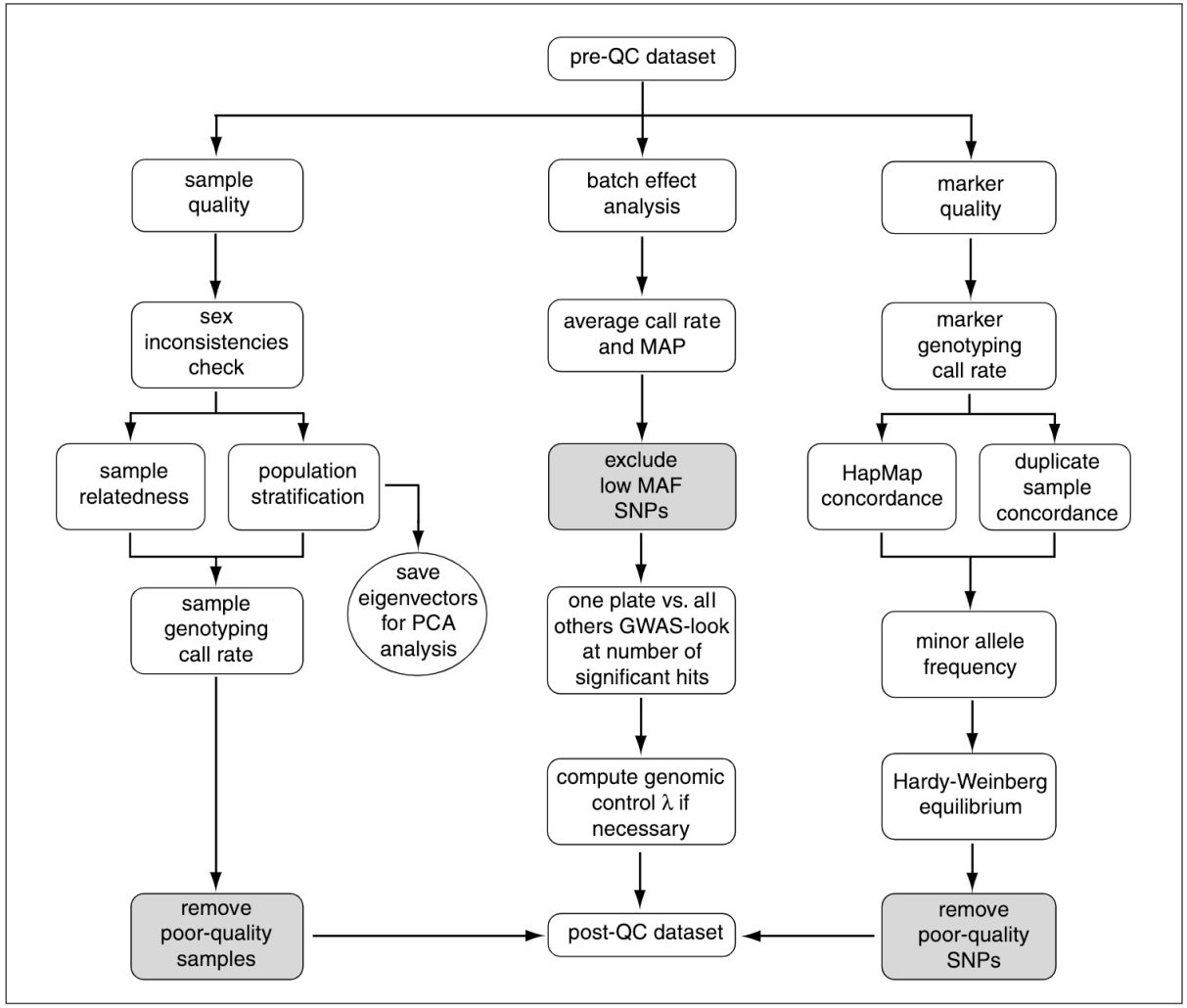


Figure 1.19.1 A flowchart overview of the entire GWAS QC process. Each topic is discussed in detail in the corresponding section in the text. Squares represent steps, ovals represent input or output data, and trapezoids represent filtering of data.

DOI:10.1002/0471142905.hg0119s68

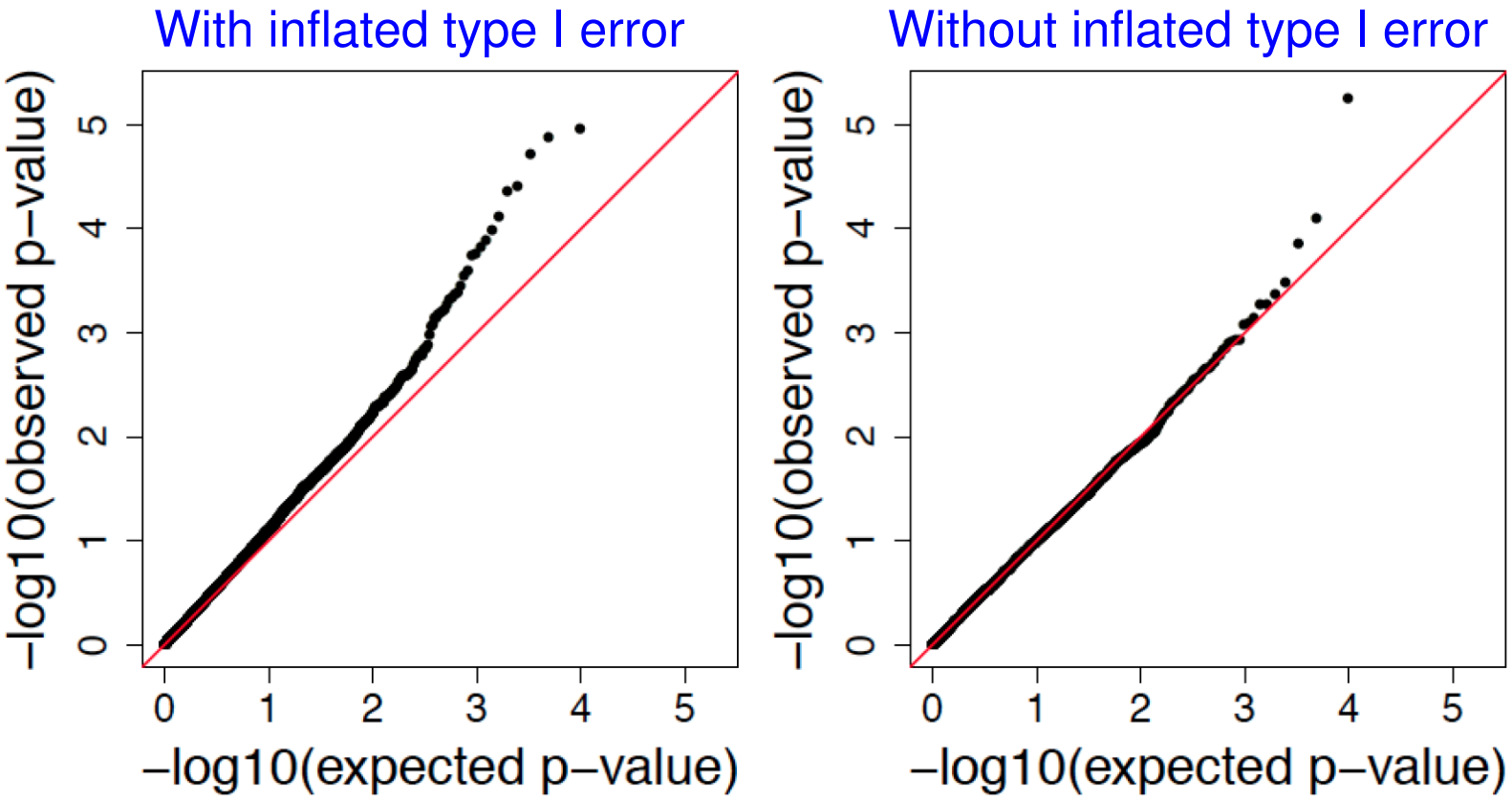
Quantile-Quantile (QQ) Plot

– Obtained $-\log 10$(p-values) from GWAS

– Sort all $-\log 10$(p-values) from most significant to least

– Pair these with the expected values of order statistics of a Uniform$(0, 1)$ distribution

– Under NULL hypothesis (no association), p-values follow a Uniform$(0, 1)$ distribution

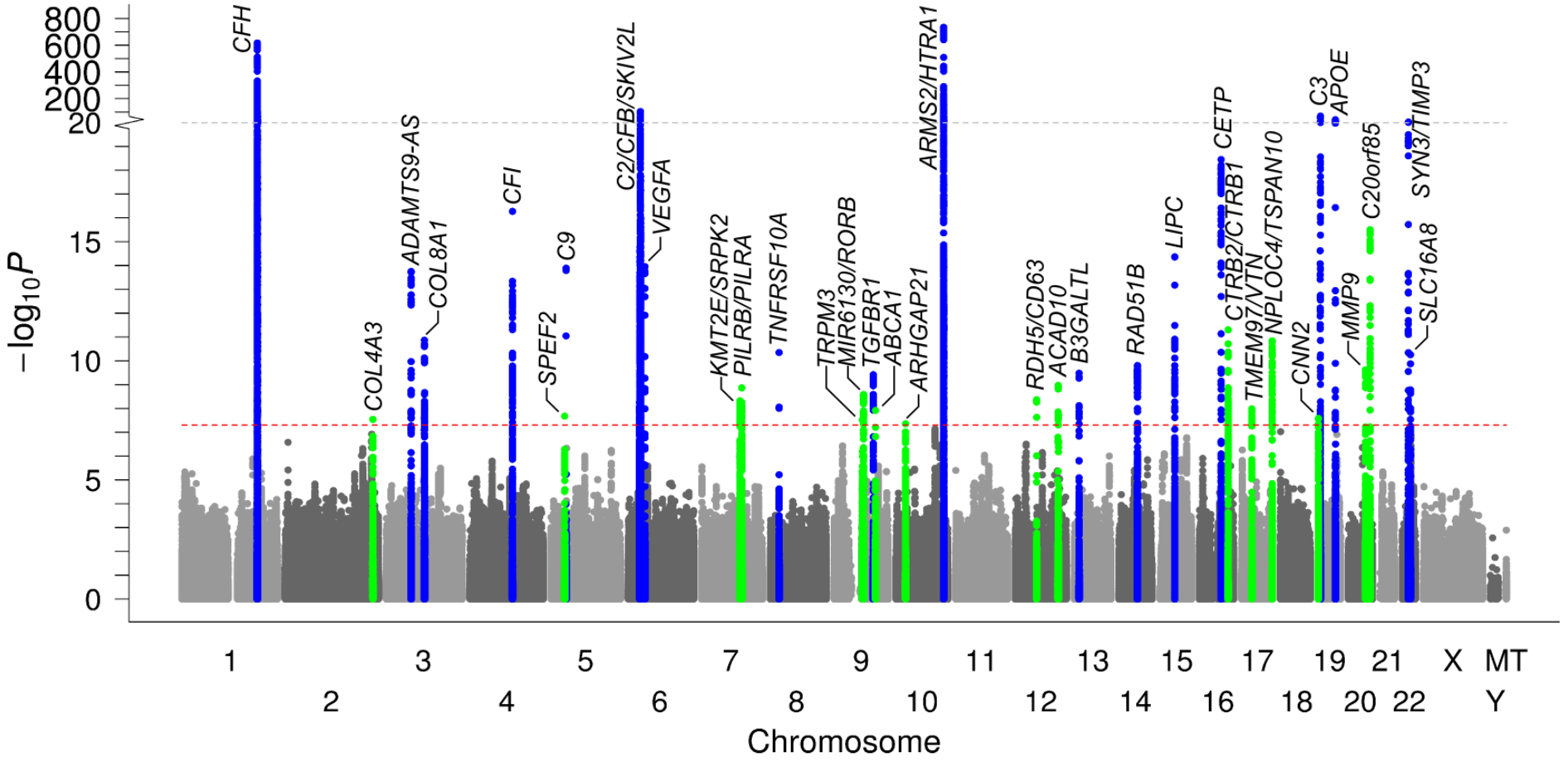


Visualize GWAS Results: Manhattan Plot

– Scatter plot of $-\log 10$(p-values) across all genome-wide variants

– Visualize signal peaks

GWAS Results



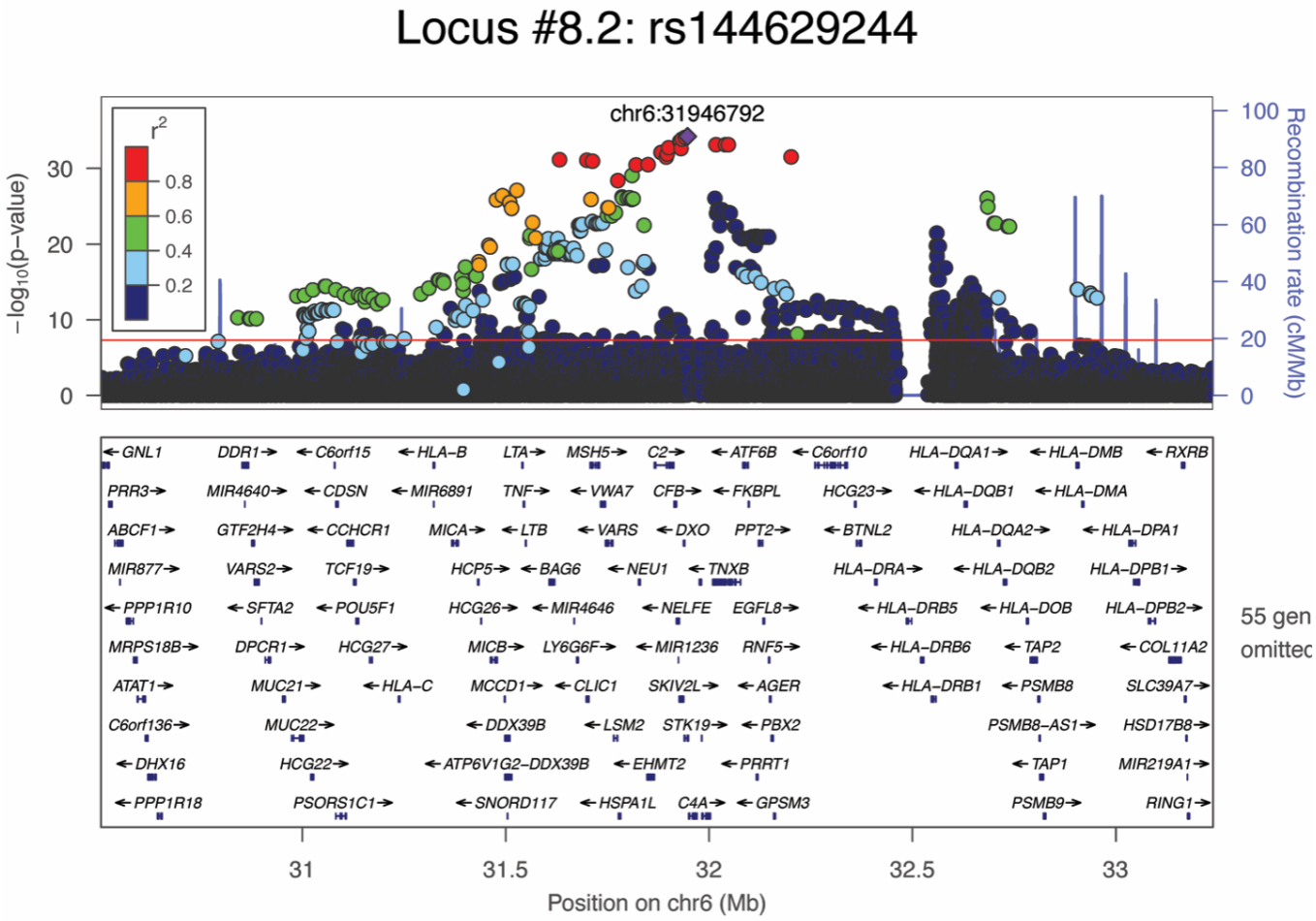
18 known AMD loci and **16 novel AMD loci**

Visualize GWAS Results: Locus Zoom Plot

– Zoom into the peak region with gene annotations: http://locuszoom.org/

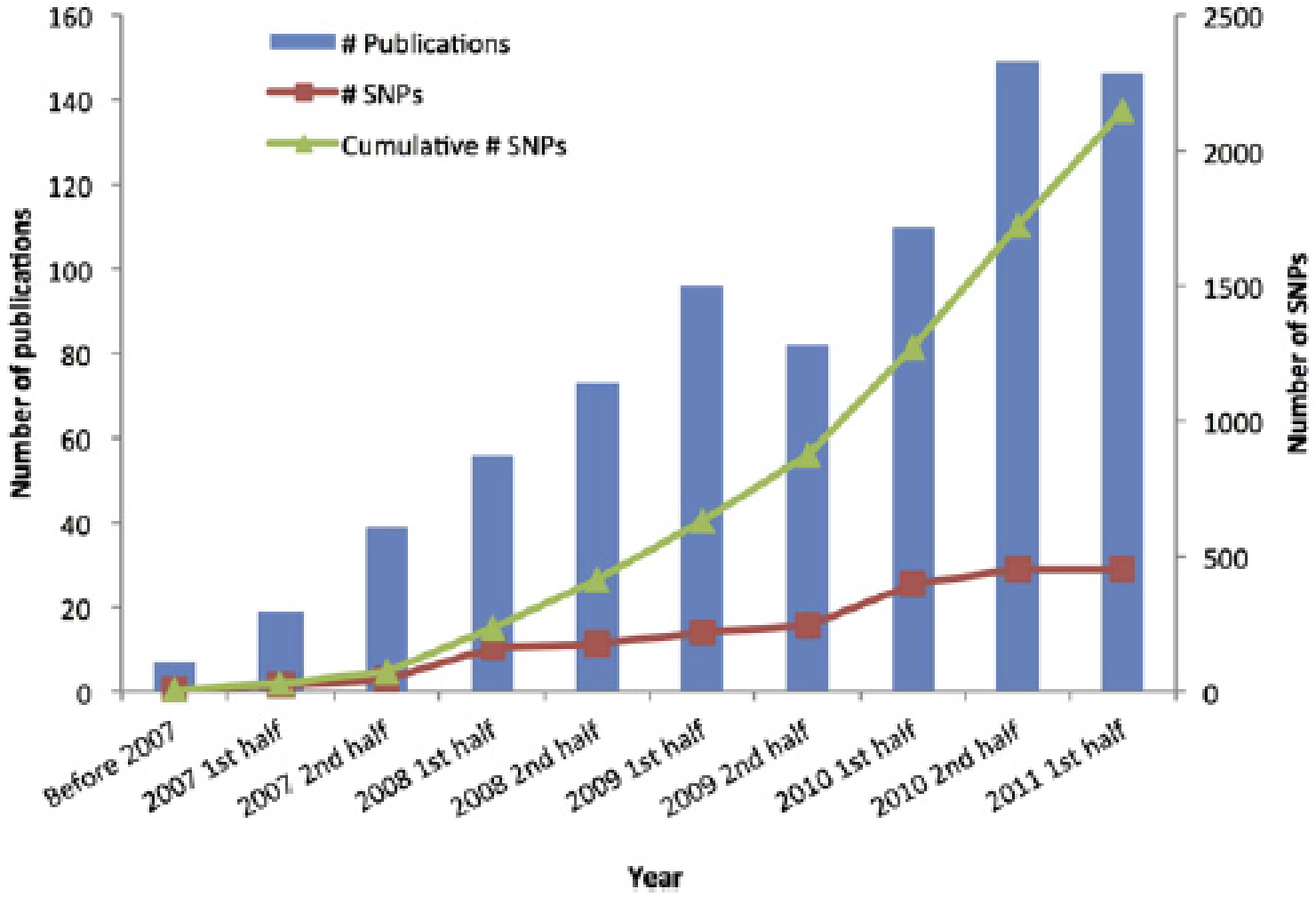
– Visualize r^2 between the specified significant (purple diamond) signal and its neighbor SNPs

– Visualize recombination rate



NHGRI-EBI GWAS Catalog

https://www.ebi.ac.uk/gwas/



2018 Apr

Associations: 69,885

Studies: 5,152

Papers: 3,378

www.ebi.ac.uk/gwas