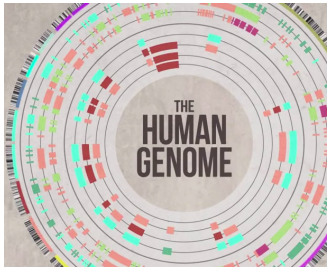
# Introduction to sequencing

Hilary Martin

Wellcome Trust Sanger Institute

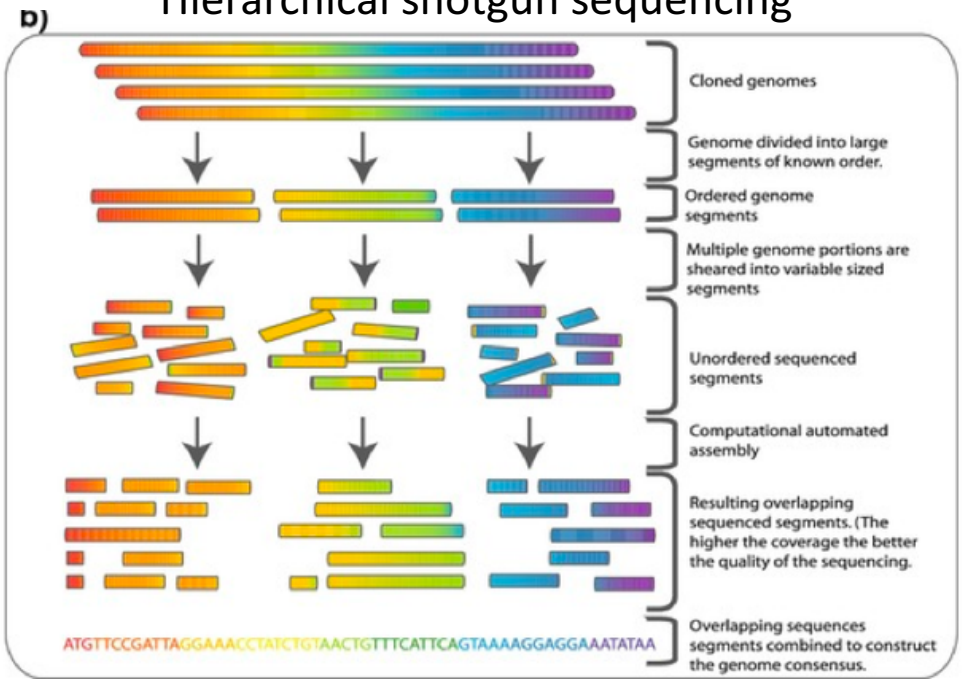
Hinxton (near Cambridge), UK

# Human genome project

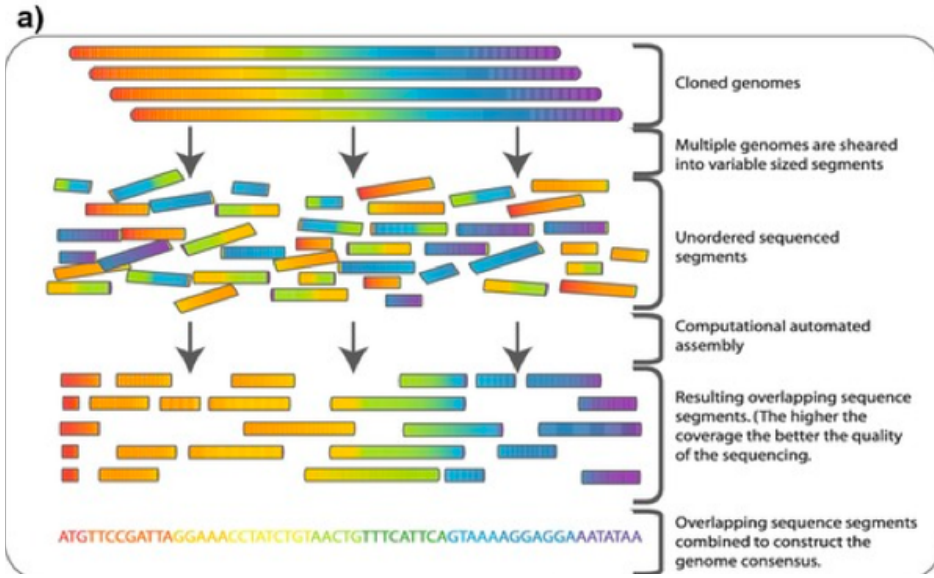



- Public effort - 1990-2003; $3 billion; hierarchical shotgun ("clone by clone")
- Private effort (Celera) – 1998-2001; $300 million; whole-genome shotgun
- Both produced chimeric assemblies of multiple people

Hierarchical shotgun sequencing



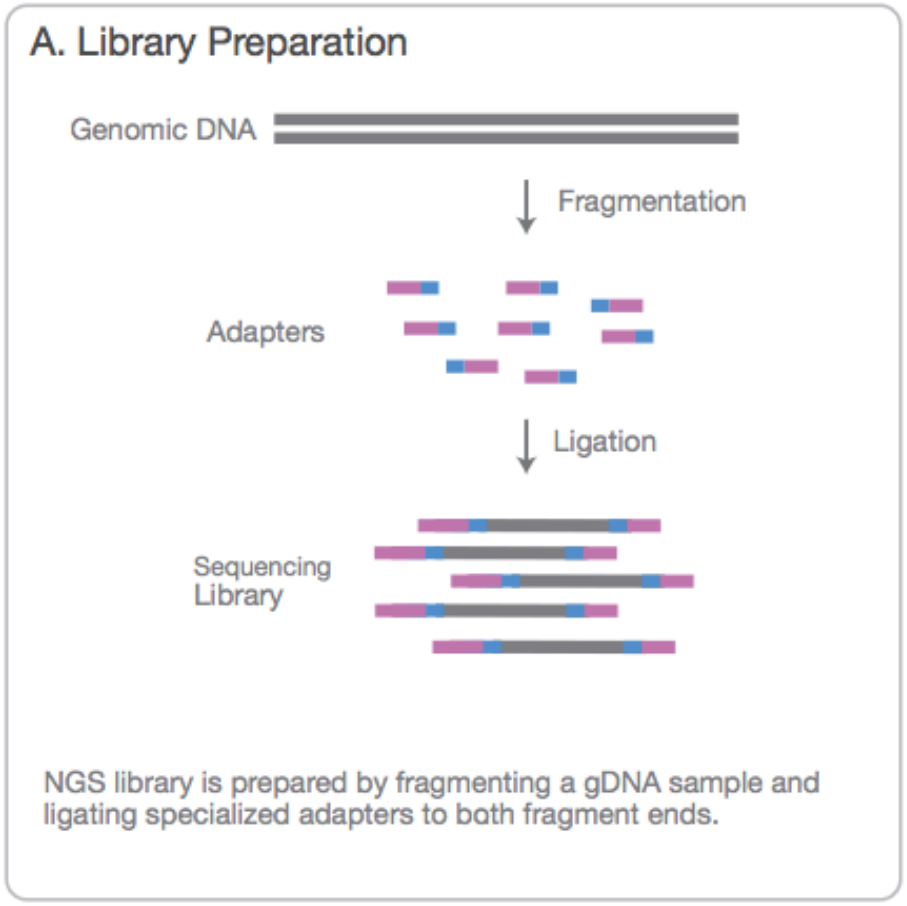

Whole-genome shotgun sequencing

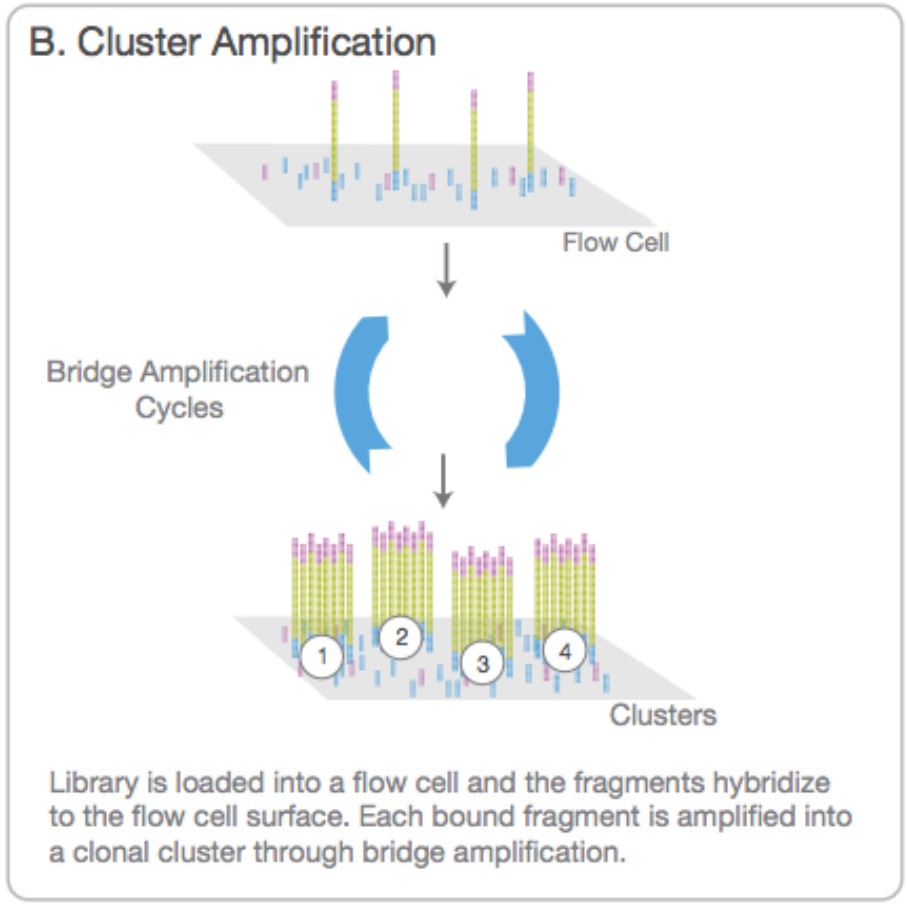


# “Next-generation” sequencing

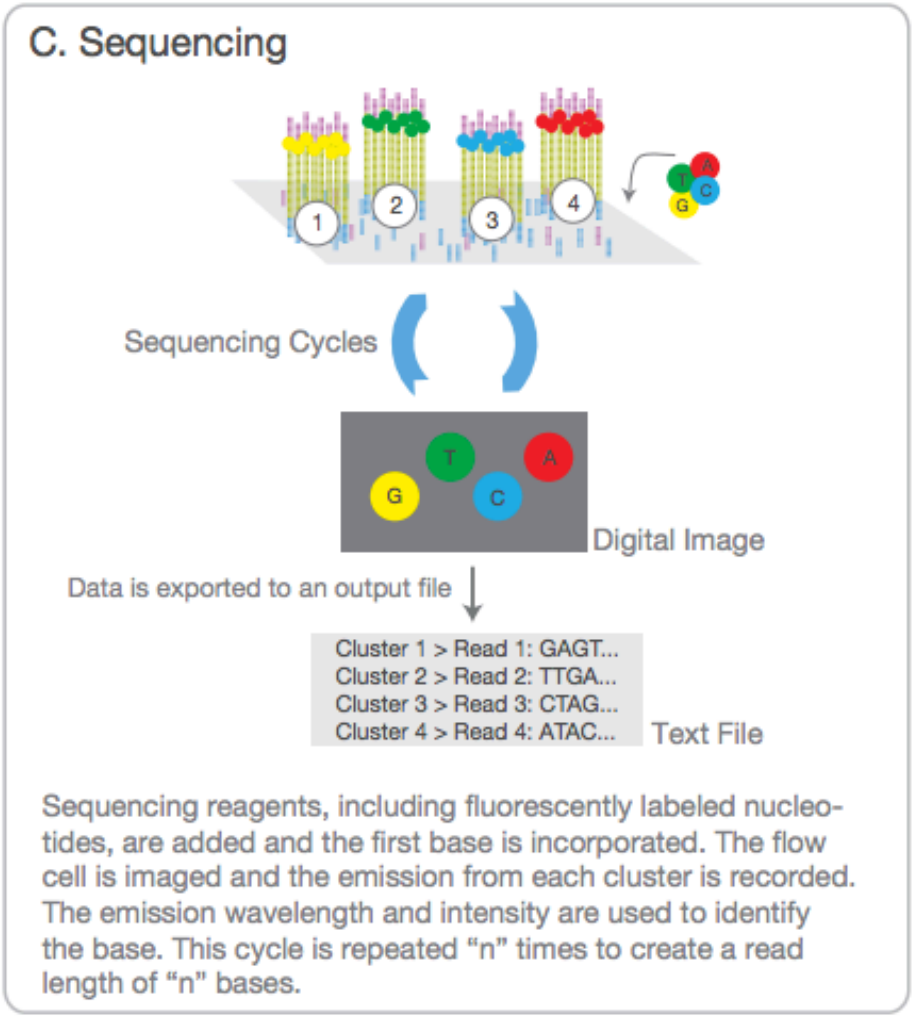
- 2008 – first whole human genome sequenced using “next-generation” technology (James Watson)
  - Used 454 sequencing (pyrosequencing – sequencing by synthesis relying on detection of pyrophosphate release upon nucleotide incorporation)
  - Could sequence 400-600Mb of DNA per 10-hour run
- Several “NGS” technologies emerged:
  - Roche 454 sequencing
  - Ion torrent: Proton / PGM sequencing
  - SOLiD sequencing
  - Illumina (Solexa) sequencing
- Illumina now the most widely used

# Illumina sequencing

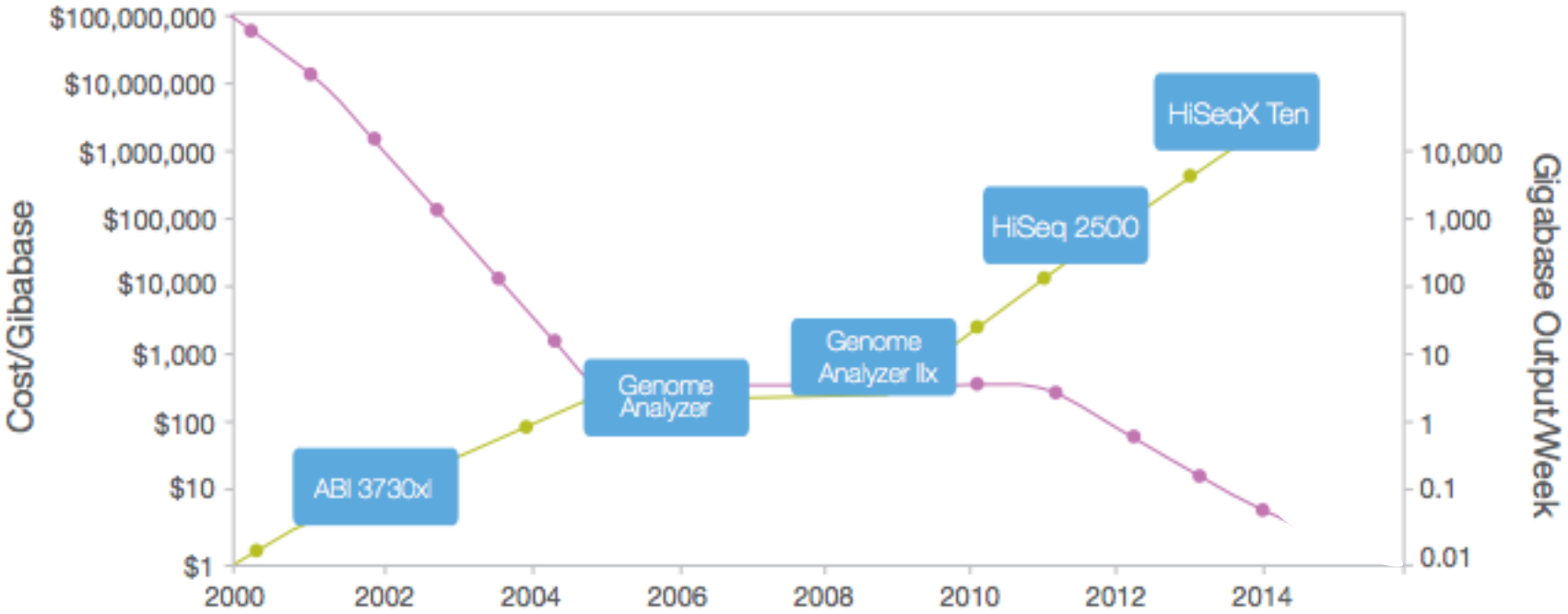


# Illumina sequencing




# Cost of sequencing




- Reminder: human genome 3 Gigabases
- Due to errors, we tend to sequence 20-30X to obtain high quality sequence i.e. 60-90Gb → currently ~$1000/genome

https://www.illumina.com/content/dam/illumina-marketing/documents/products/illumina_sequencing_introduction.pdf

# Direct sequencing has enormous potential

ARTICLES

nature genetics

BRIEF REPORT

Exome
disord

Sarah B Ng[1,1]
Chad D Huff
Michael J Bar

**Making a definitive diagnosis: Successful clinical application of whole exome sequencing in a child with**

REPORT

HUMAN GENETICS

**Whole-Genome Sequencing for Optimized Patient Management**

Matthew N. Ba
Claudia Gonza
Margaret B. M
Shahed Yousaf

ARTICLE

doi:10.1038/nature21062

**Prevalence and architecture of *de novo* mutations in developmental disorders**

Deciphering Developmental Disorders Study

# …and tremendous challenges

- Managing and processing vast quantities of data into variation
- Interpreting millions of variants per individual
  - An individual's genome harbors:
    - ~100,000 exonic variants
    - ~80 point nonsense (loss-of-function) mutations
    - ~100-200 frameshift mutations
    - Tens of splice site mutations, CNV-induced gene disruptions

***For very few of these do we have any conclusive understanding of their medical impact in the population***

# Technical aspects of sequencing studies

# Coverage

Coverage (or depth) is the average number of reads that include a given nucleotide in the reconstructed sequence.

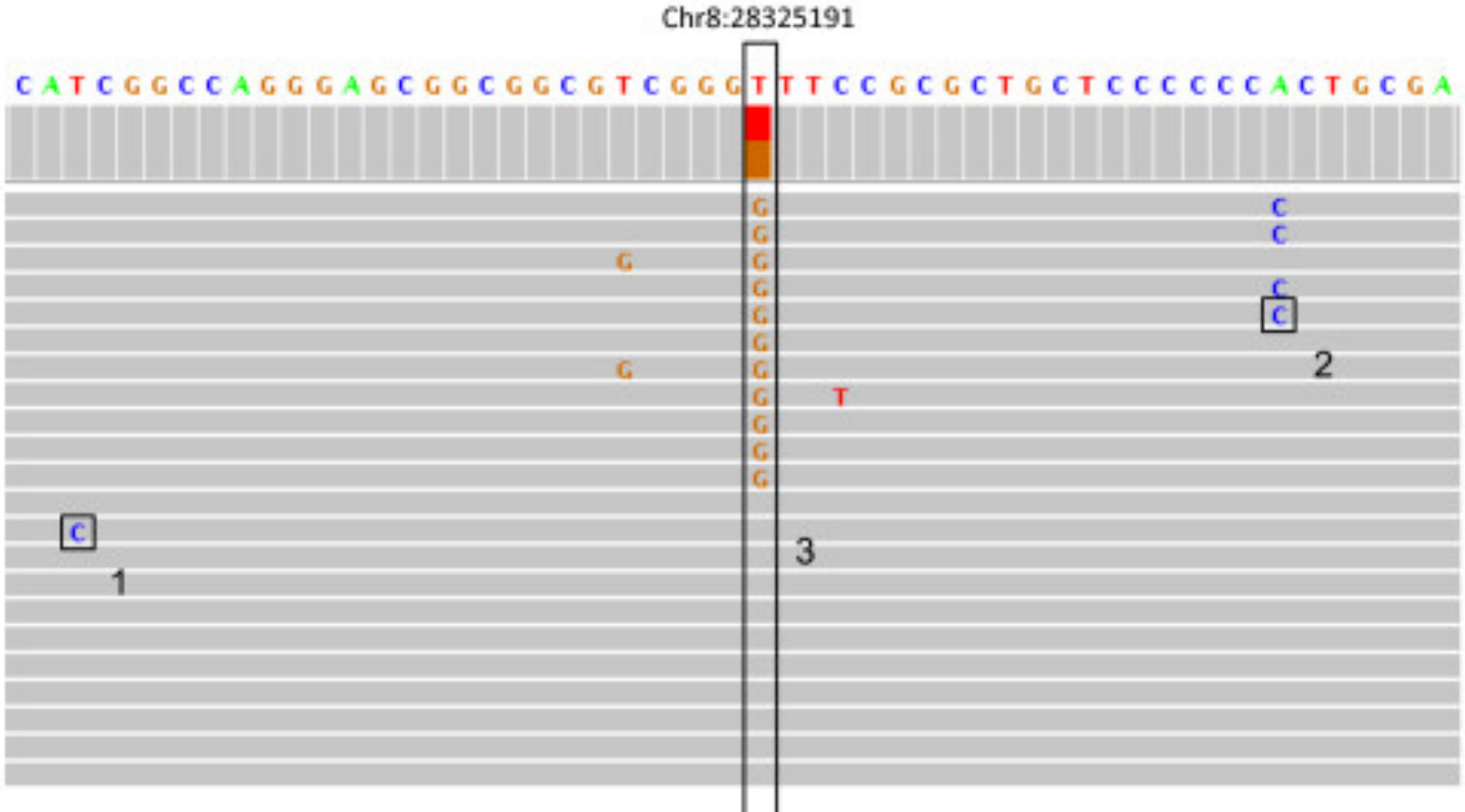

Length of genomic segment: **L**
Number of reads: **n**
Length of each read: **l**

**Definition:** Coverage **C = n l / L**

- Typically use 20-30X coverage to obtain high-quality sequence for human genomes.
- For some purposes, even very low-coverage sequencing (4X, 1X, 0.2X!) is useful.

# Why do we need >1X (or >2X) coverage?

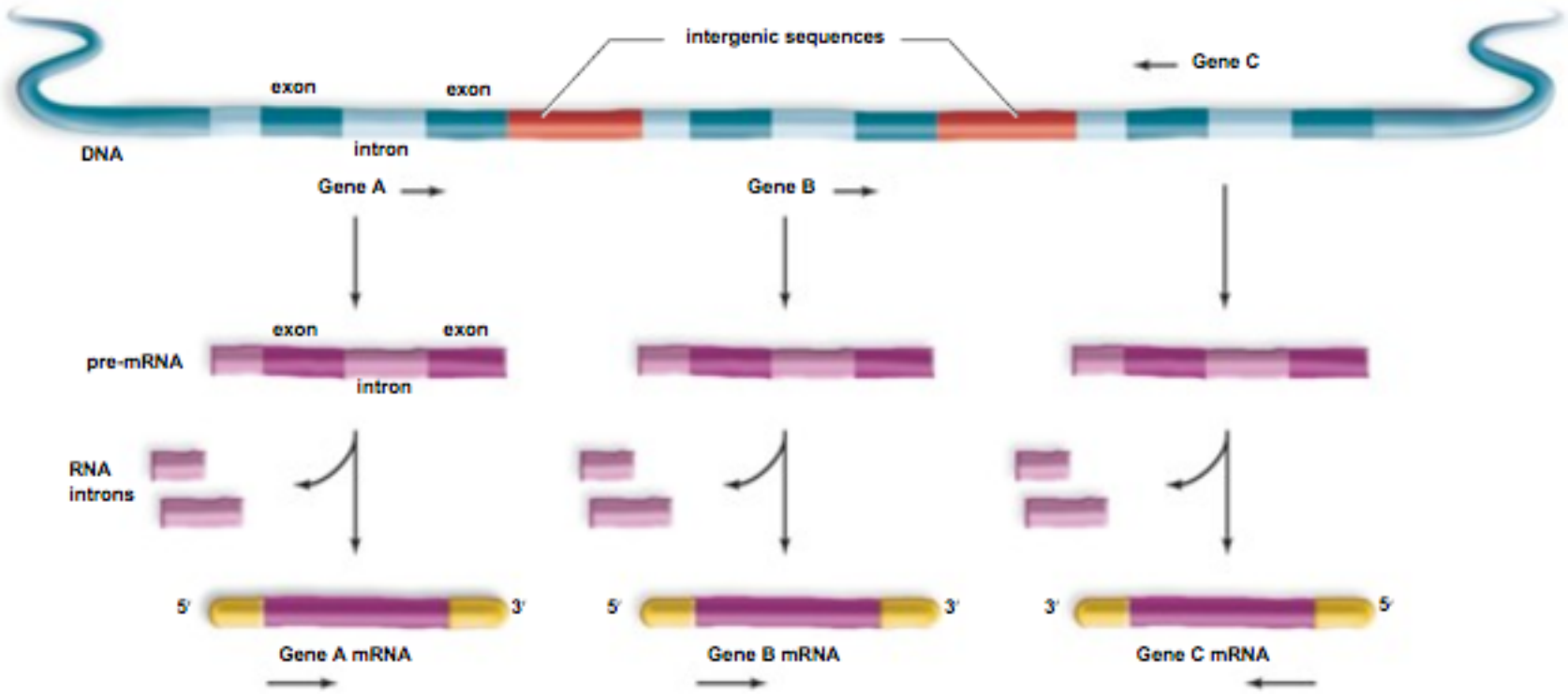
- Humans are diploid – number of reads covering each allele follows a binomial distribution
- Need to distinguish real variants from sequencing errors, especially since some errors are systematic.



# Technologies for sequencing humans

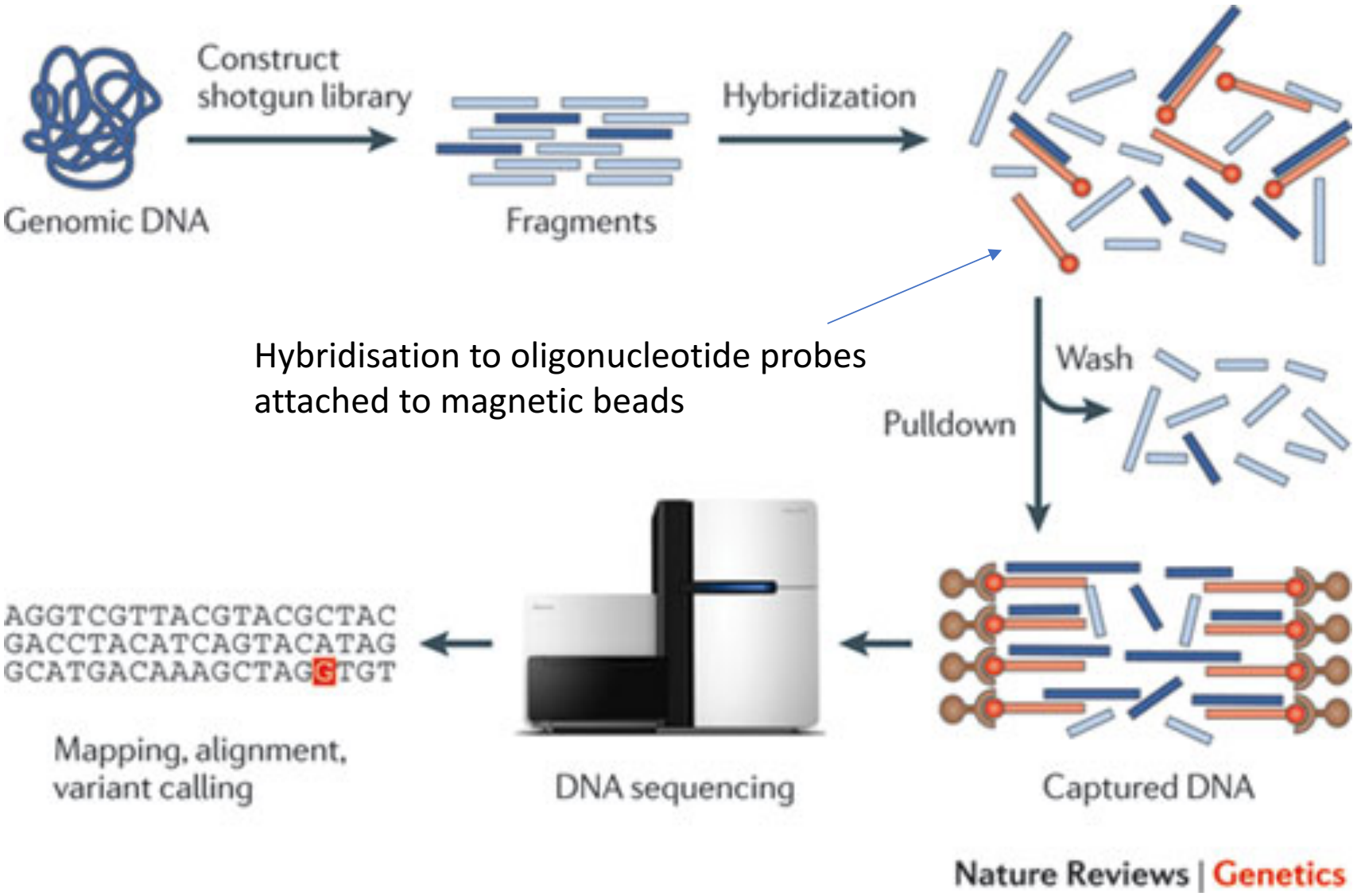
| | **Whole-genome sequencing (WGS)** | **Whole-exome sequencing (WES)** |
|---|---|---|
| Amount of sequence | 3Gb | 30Mb |
| Typical coverage | 30X (for high quality) | Average 60-180X |
| Library preparation | Randomly shear, then do hybridisation-based capture of exonic DNA fragments | Shotgun sequence - randomly shear and capture |
| Advantages | • Covers (most of) the whole sequence<br>• (fairly) unbiased ascertainment | • Cheaper ($200-300)<br>• Focuses on coding regions |
| Disadvantages | • expensive (~$1000 for 30X)<br>• too expensive to do at very high coverage | • Uneven coverage, biases<br>• Harder to call large copy number variants |
| Common applications | • Reference panels for imputation<br>• Complex traits | • Mendelian diseases<br>• Interrogate rare coding variants in complex traits |

# The exome



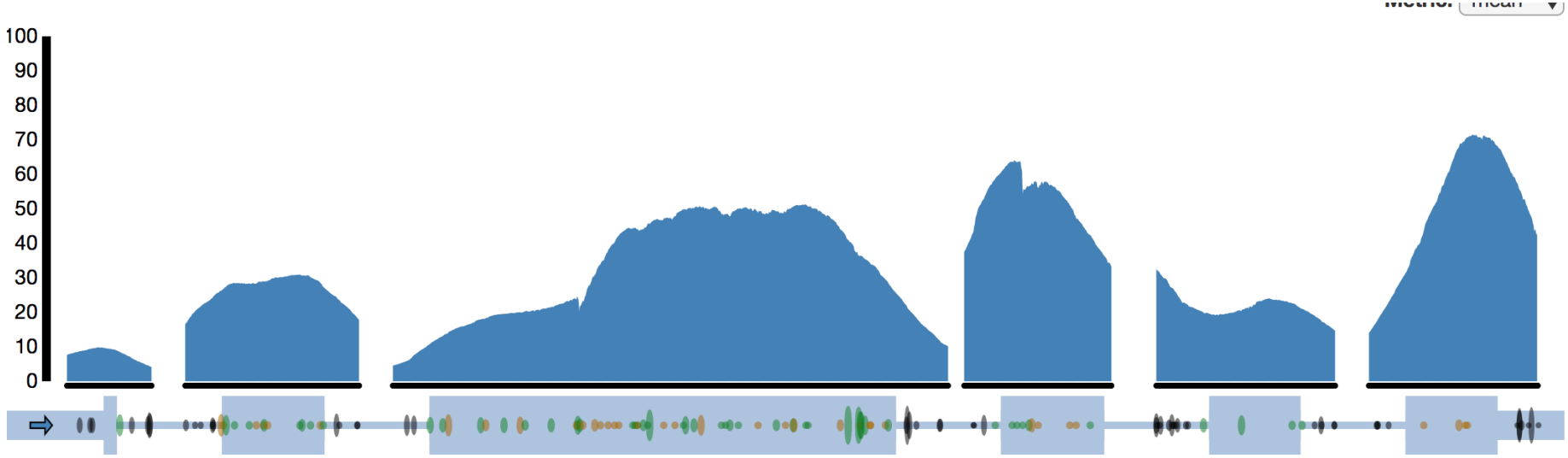

- Exome = all the exons (bits of the genome that encode proteins)

# Targeted exome capture




Bamshad et al., Nature Review Genetics, 2011

# Variable coverage in exome sequencing




Also note that WES shows a greater reference bias than WGS (53% versus 50.3%) – due to both capture probes and mapping bias

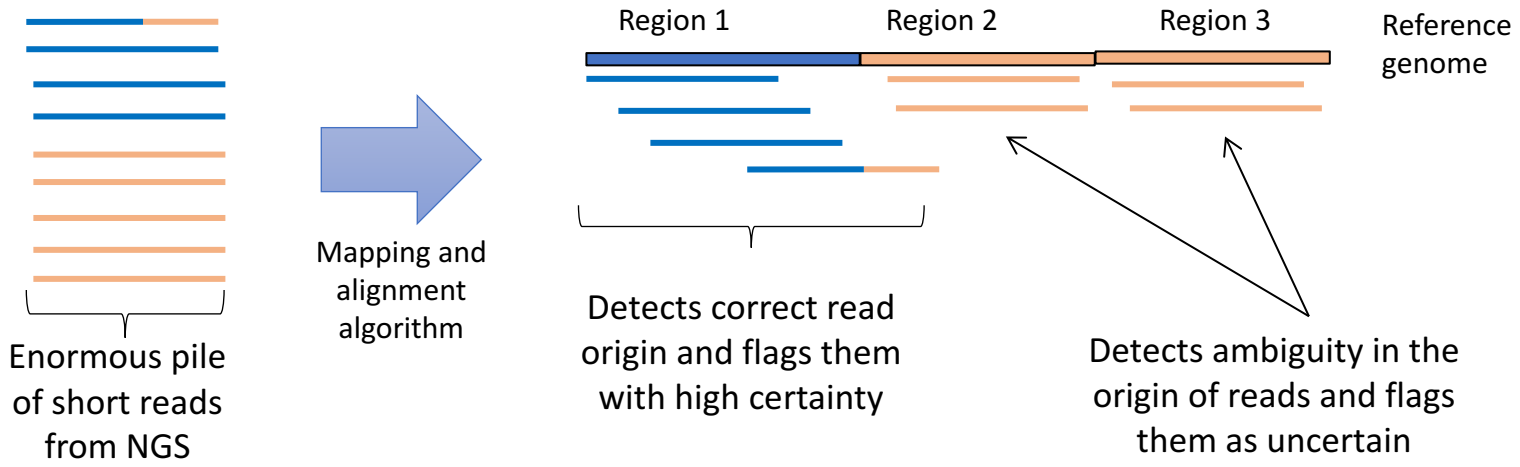
# Depth considerations

- Mendelian disease - need high coverage to be sure rare/*de novo* variants are real (20-30X WGS, or >60X WES)
- Somatic mutations – variants in <<50% of reads, so need high coverage
- Complex disease
  - High coverage needed to interrogate rare variants
  - Low coverage may still be useful to study common variants (genotypes can be improve by imputation)
- Imputation reference panel – want large number of haplotypes, low coverage sufficient for common variants

# Step 1: Aligning to a reference


Torsten Seemann

# Finding the true origin of each read is a computationally demanding and important first step




- Many different alignment programs
- Commonly used aligner: BWA-MEM (Li and Durbin) - robust, accurate ‘gold standard’ – see paper in directory

SAM/BAM files

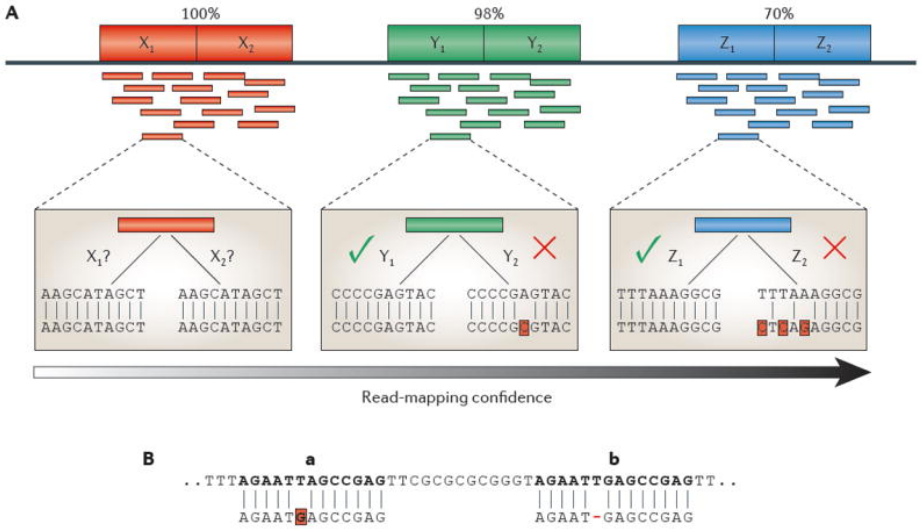
Ben Neale

# Repeats cause problems with sequence data

- Simple repeats
- Paralogs resulting from genome duplication
- Repeated domains found in many different proteins

Reference: TAGTAGTAGTAGTAGTAGTAGTAGT

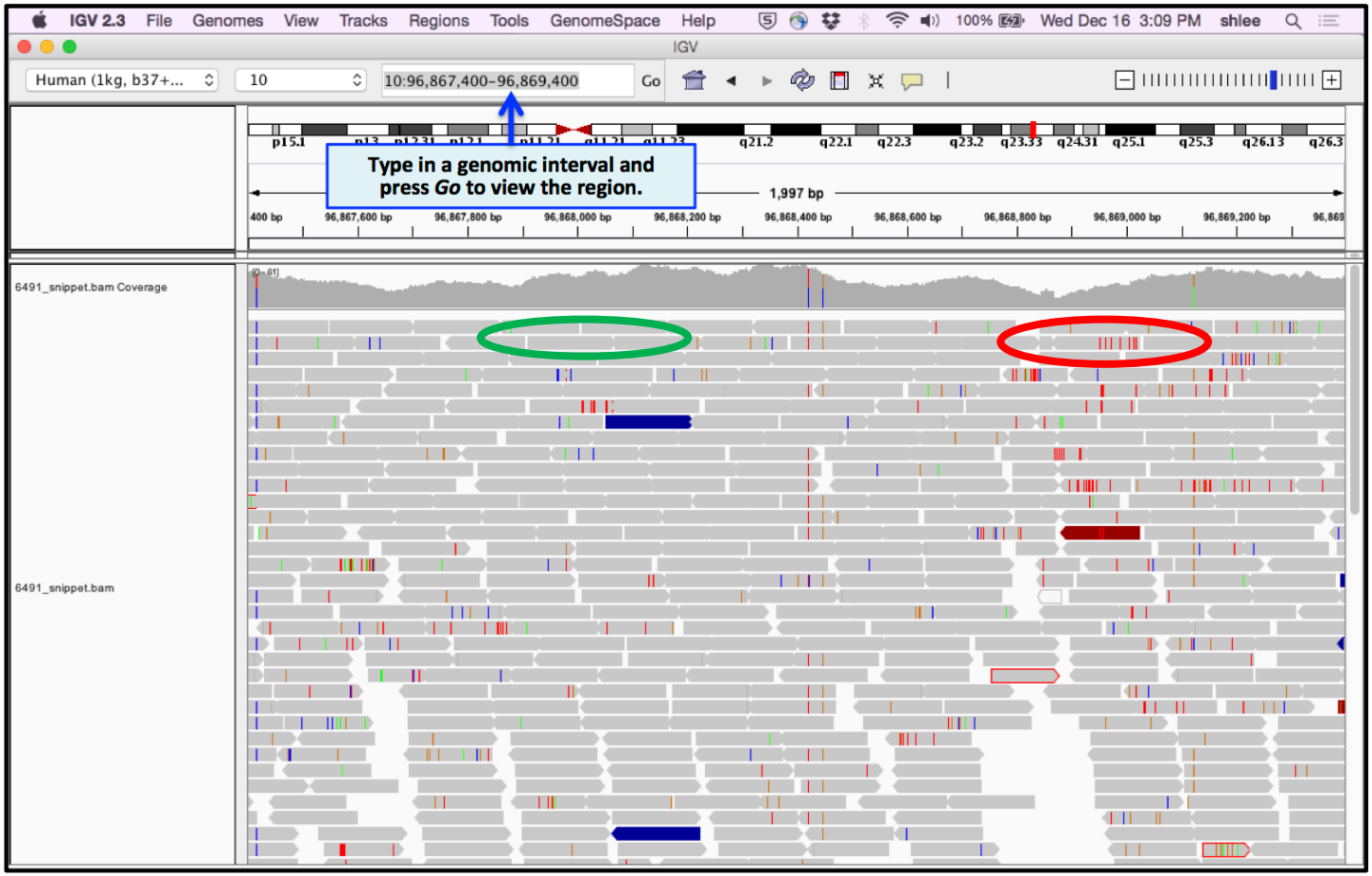
Where to put the read TAGTAGTAGT ?




Treangen and Salzberg, Nat. Rev, Genet., 2011

# Mapping quality

- quantifies the probability that a read is misplaced
- Depends on base quality scores at mismatched bases, and also how many other possible mappings there are throughout the genome



# The SAM/BAM file format

- The Sequence Alignment and Mapping (SAM) file format was designed to capture all of the critical information about NGS data in a single indexed and compressed file
- Contains read sequence, base quality scores, location of alignments, differences relative to reference sequence, MAPQ
- Has enabled sharing of data across centers and the development of tools that work across platforms
- More info at http://samtools.sourceforge.net/

---

**The Sequence Alignment/Map (SAM) Format and SAMtools**

Heng Li [1,*], Bob Handsaker [2,*], Alec Wysoker [2], Tim Fennell [2], Jue Ruan [3], Nils Homer [4], Gabor Marth [5], Goncalo Abecasis [6], Richard Durbin [1,†] and 1000 Genome Project Data Processing Subgroup

[1]Wellcome Trust Sanger Institute, Wellcome Trust Genome Campus, Cambridge, CB10 1SA, UK, [2]Broad Institute of MIT and Harvard, Cambridge, MA 02141, USA, [3]Beijing Institute of Genomics, Chinese Academy of Science, Beijing, 100029, China, [4]Department of Computer Science, University of California Los Angeles, Los Angeles, CA 90095, USA, [5]Department of Biology, Boston College, Chestnut Hill, MA 02467, USA, [6]Center for Statistical Genetics, Department of Biostatistics, University of Michigan, Ann Arbor, MI 48109, USA

Associate Editor: Prof. Alfonso Valencia

Ben Neale

# The Genome Analysis Toolkit (GATK)

- toolkit for processing sequence data (post-alignment), calling and filtering variants

- supports any BAM-compatible aligner

- many tools developed in GATK: base quality score recalibration, HaplotypeCaller, multi-sample genotyping, variant filtering, variant quality score recalibration

- memory and CPU efficient, cluster friendly and are easily parallelized

- being used at many sites around the world

More info: http://www.broadinstitute.org/gsa/wiki/

Ben Neale

# Variant Call Format (VCF)

N.B. differs from A1/A2 on genotyping chips, or minor/major allele

| #CHROM | POS | ID | REF | ALT | QUAL | FILTER | INFO | FORMAT | NA00001 |
|---|---|---|---|---|---|---|---|---|---|
| 20 | 14370 | rs6054257 | G | A | 29 | PASS | NS=3;DP=14;AF=0.5;DB;H2 | GT:GQ:DP:HQ | 0\|0:48:1:51,51 |
| 20 | 17330 | . | T | A | 3 | q10 | NS=3;DP=11;AF=0.017 | GT:GQ:DP:HQ | 0\|0:49:3:58,50 |

Chromosome

Position

SNP ID

Reference Allele

Alternate Allele

Variant quality Score

Filter

INFO field contains meta-data

NS = # samples with data

DP = total depth

AF = ALT allele frequency

DB = in dbSNP

H2 = in HapMap2

FORMAT specifies the genotype format

GT = genotype

GQ = genotype quality

DP = sample depth

HQ = haplotype quality

Individual genotype follows FORMAT structure

# Discovery versus genotyping

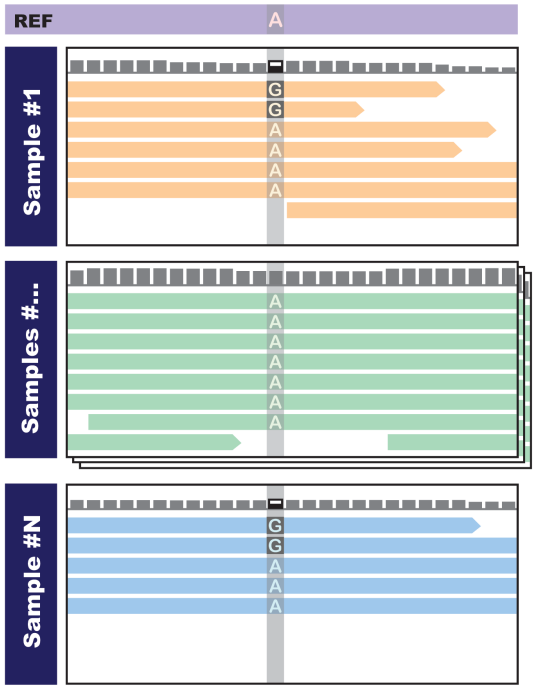
- In genotype data, we know the variants are real – we just need to work out what individuals' genotypes are
- In sequence data, we also have a discovery problem – which variants are real? – as well as a genotyping problem

# What filters do we use?

- Problem: correlated sequencing errors and mapping artefacts drive false positives (cause loss of power, spurious conclusions) → VQSR etc
- The following should be random if the sequencing technology is working as expected:
  - Variant position in read
  - Strand bias – 5'-to-3' and 3'-to-5' reads should give equal representation of alternate allele
  - Allele balance – at heterozygous sites, the number of ALT reads should follow a binomial distribution with p=0.5

# Value of simultaneous variant calling in multiple individuals

- Sensitivity
  - Greater statistical evidence compiled for true variants seen in >1 individual
- Specificity
  - Deviations in metrics that flag false positive sites become much more statistically significant e.g. allele balance, strand bias, proportion of reads with low MAPQ
- Distinguishing missing genotype from homozygous reference





Ben Neale

# Variant filtration strategies are still evolving
## VQSR is a common approach

- Variant quality score recalibration aims to enable variant filtering in order to balance sensitivity and specificity
- VQSR uses machine learning to learn the annotation profile of good versus bad variants across a dataset, by integrating information from multiple QC metrics
- Requires a set of “true sites” as input e.g. HapMap3 sites
- Calculates log odds ratio of being true variant versus being false under trained Gaussian mixture model - VQSLOD added to INFO field

http://gatkforums.broadinstitute.org/gatk/discussion/39/variant-quality-score-recalibration-vqsr

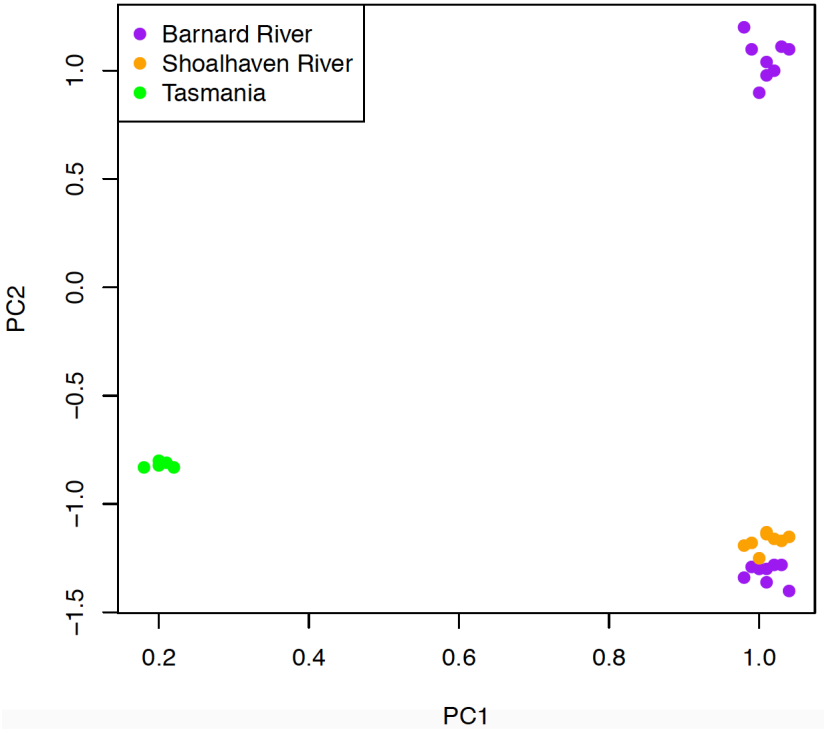
# An important QC metric
## Transition:transversion ratio across the dataset

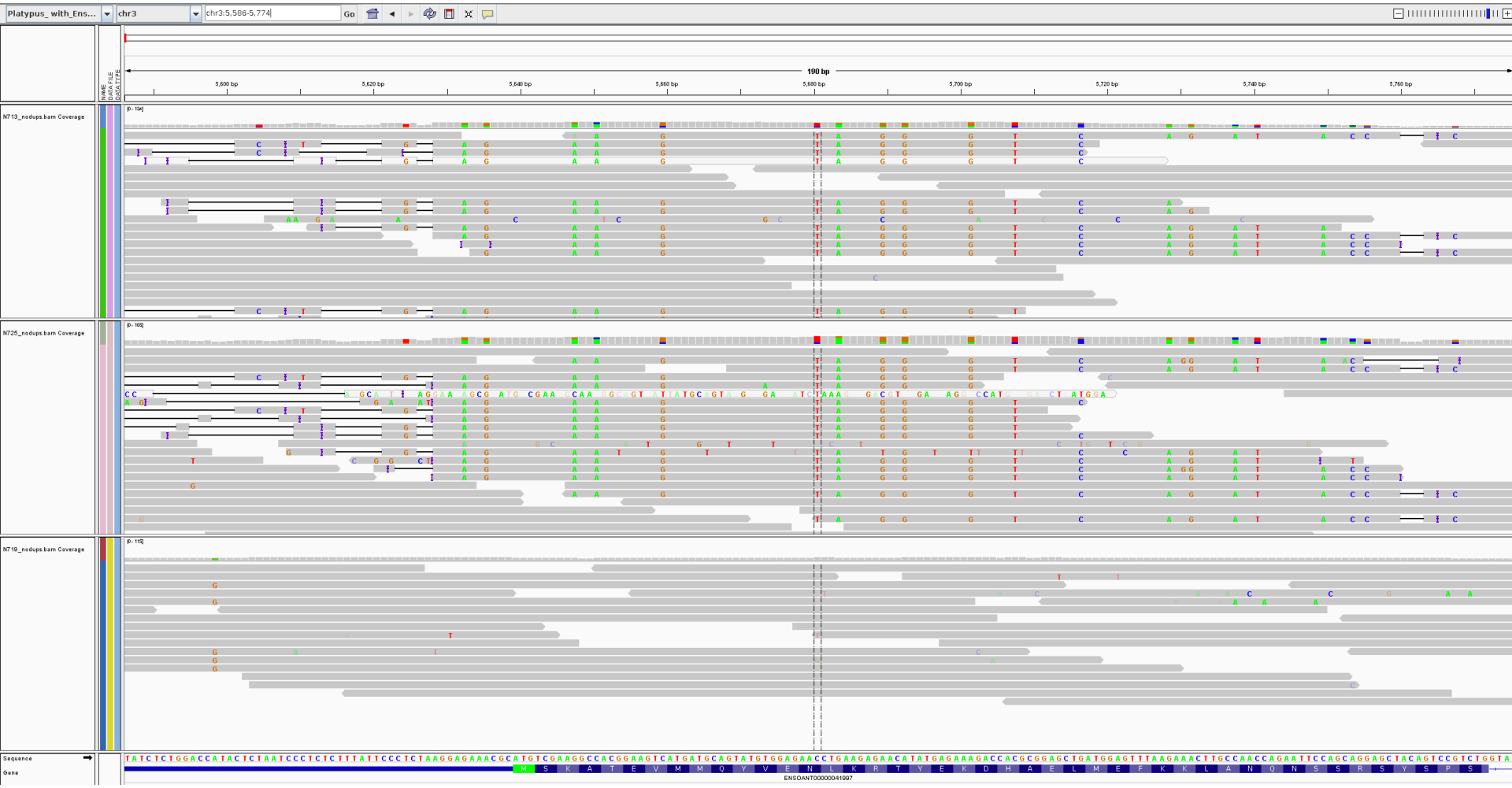
- within vs between type: purine (A & G)  or pyrimidine (C & T)
- transitions are expected to occur twice as frequently as transversions
- across the entire genome Ti:Tv is typically ~2
- in protein coding regions, Ti:Tv is ~3 (higher because transversions are much more likely to change the encoded amino acid, especially in the third base of a codon)
- not relevant for genotype data since we know the variants are real

| | A | C | G | T |
|---|---|---|---|---|
| A | - | Tv | Ti | Tv |
| C | Tv | - | Tv | Ti |
| G | Ti | Tv | - | Tv |
| T | Tv | Ti | Tv | - |

# A cautionary tale: another peril of sequence data

- Sequenced ~60 platypus samples
- Two groups of samples from the same river fell far apart on the PCA
- Noticed that this was driven by dense heterozygous SNPs falling in exons, present only in some lanes in those samples




Platypus_ with_Ens...
chr3
chr3:5,586-5,774
Go
190 bp
5,600 bp
5,620 bp
5,640 bp
5,660 bp
5,680 bp
5,700 bp
5,720 bp
5,740 bp
5,760 bp
N713_nodups.bam Coverage
N725_nodups.bam Coverage
N719_nodups.bam Coverage
Sequence
Gene
ENSOANT00000041997

# A cautionary tale: ~~a new platypus sub-species?~~ contamination

- Sequenced ~60 platypus samples
- Two groups of samples from the same river fell far apart on the PCA
- Noticed that this was driven by dense heterozygous SNPs falling in exons, present only in those samples
- Turns out some sequencing lanes had been contaminated with human exome sequencing libraries
- Human exonic reads still close enough to platypus exons to align
- Would never see something like this with genotype chip data

# More common contamination problems

- Contamination between samples in the same sequencing lane
- Bacterial/viral contamination
- Females who have had multiple sons (fetal DNA remaining in mother's blood)
- People who have had bone marrow transplants

# QC for sequencing versus genotype data

- Error modes greatly differ between sequencing and genotyping chips
- In sequence data, there is a discovery problem as well as a genotyping problem (i.e. the variants may not be real variants at all) – **need to filter sites as well as genotypes**
- Contamination is more of a problem for sequencing than genotyping data
- Spontaneous DNA damage (e.g. at chemically modified nucleotides) leads to false variants in reads – need to avoid calling as variant sites

# Solved and unsolved technical problems in sequencing data processing

- We're now pretty good at SNP calling
- Indel calling still challenging, particularly in low-complexity regions (machine learning approach based on image recognition shows promise - DeepVariant)
- (Structural variants also hard to call)


Figure from Li, Bioinformatics, 2014

# Sequencing studies in practice

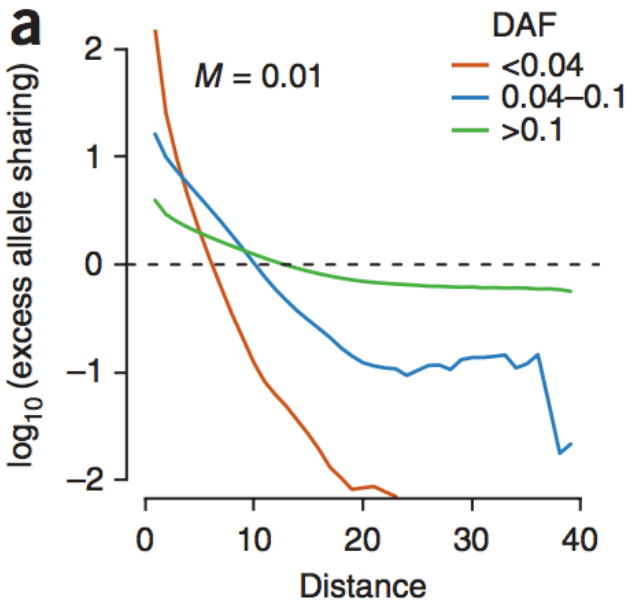
# Importance of controls

- Can’t always afford to sequence both cases and controls, so use publicly available controls (lots of potential artefacts)
- Initially, researchers relied on dbSNP
- Usually interested in rare variants  (otherwise would just genotype)
- Having ancestry-matched controls is very important, especially since rare variants tend to be geographically localised

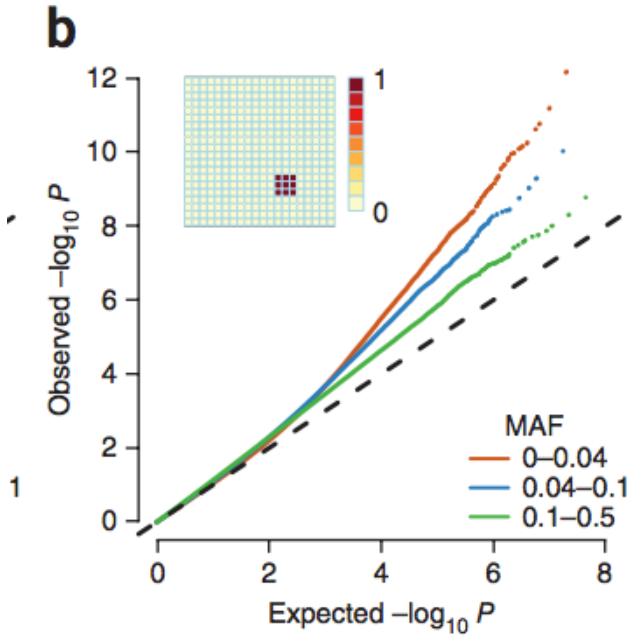
# Population stratification of rare variants

Differential confounding of rare and common variants in spatially structured populations

Iain Mathieson[1] & Gil McVean[1,2]




Plot of excess allele sharing: ratio of how much more likely two individuals at a given spatial distance are to share a derived allele compared to what would be expected in a homogenous population

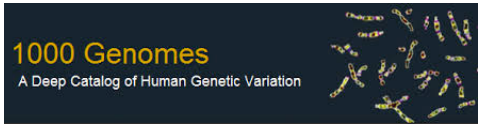



Quantile-quantile plot of association test P values broken down by allele frequency for a small, sharply defined region of constant non-genetic risk

**N.B. the scenarios simulated in this paper are probably more extreme than reality**

# Publicly available controls

- Since 2010, several projects have made large databases of sequence variation in healthy individuals available
- These are very valuable, but if you can afford to sequence in-house controls alongside your cases too, this is even better




2,500 low-coverage whole genomes

NHLBI Exome Sequencing Project (ESP)
Exome Variant Server

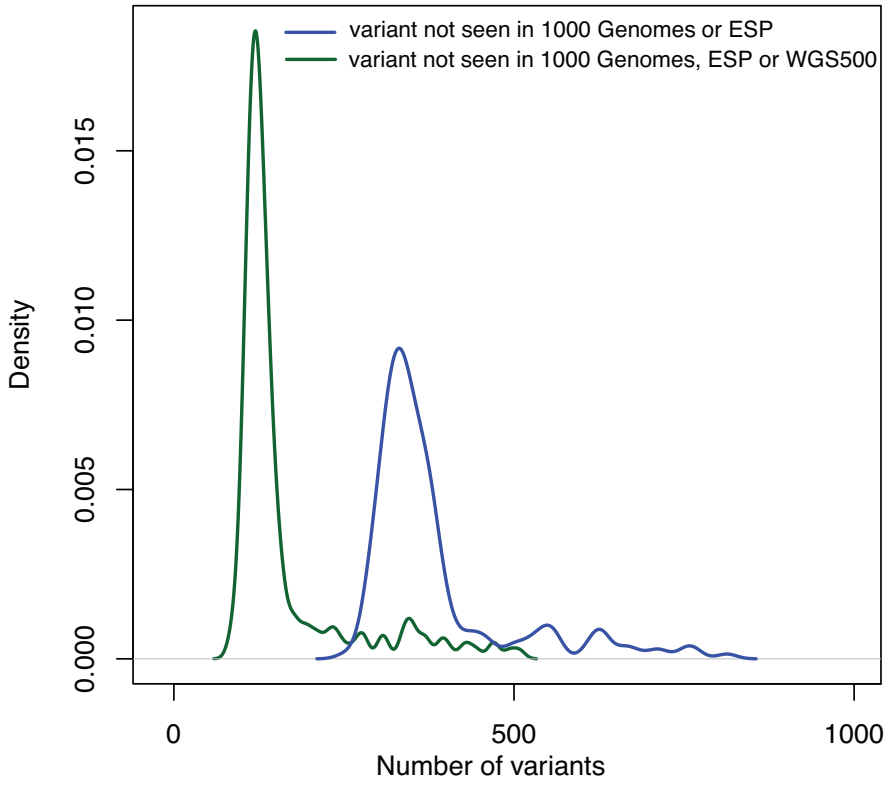
6,500 European and African American exomes
(caveat: focused on heart, lung and blood disorders)

UK 10K

4,000 low-coverage whole genomes (TwinsUK and ALSPAC)
6,000 exomes of people with extreme phenotypes of specific conditions

# Value of in-house controls

- Plot shows distribution of number of “novel” heterozygous protein-altering variants per person, across 500 people in the WGS500 project
- “novel” is defined based on absence from different control datasets (2500 individuals from 1000 Genomes, 6500 from ESP, 499 from WGS500)
- Filtering against in-house control datasets sequenced and processed in same way as patient samples helps to eliminate artefacts (erroneous variant calls)



# The Exome Aggregation consortium (ExAC)

- Largest exome sequencing dataset to date (now gnomAD)
- Samples with severe paediatric disease removed
- All samples called jointly to minimise artefactual differences between studies
- Value of large sample size to estimate allele frequency of rare variants accurately
- N.B. no individual-specific information, just total genotype counts

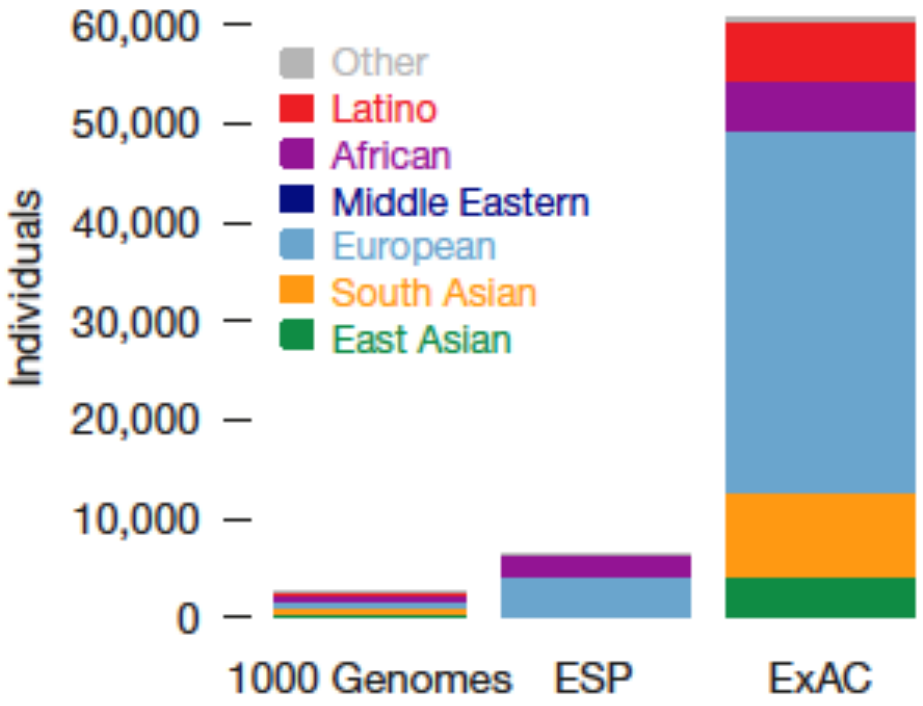
ARTICLE

doi:1(

**Analysis of protein-coding genetic variation in 60,706 humans**

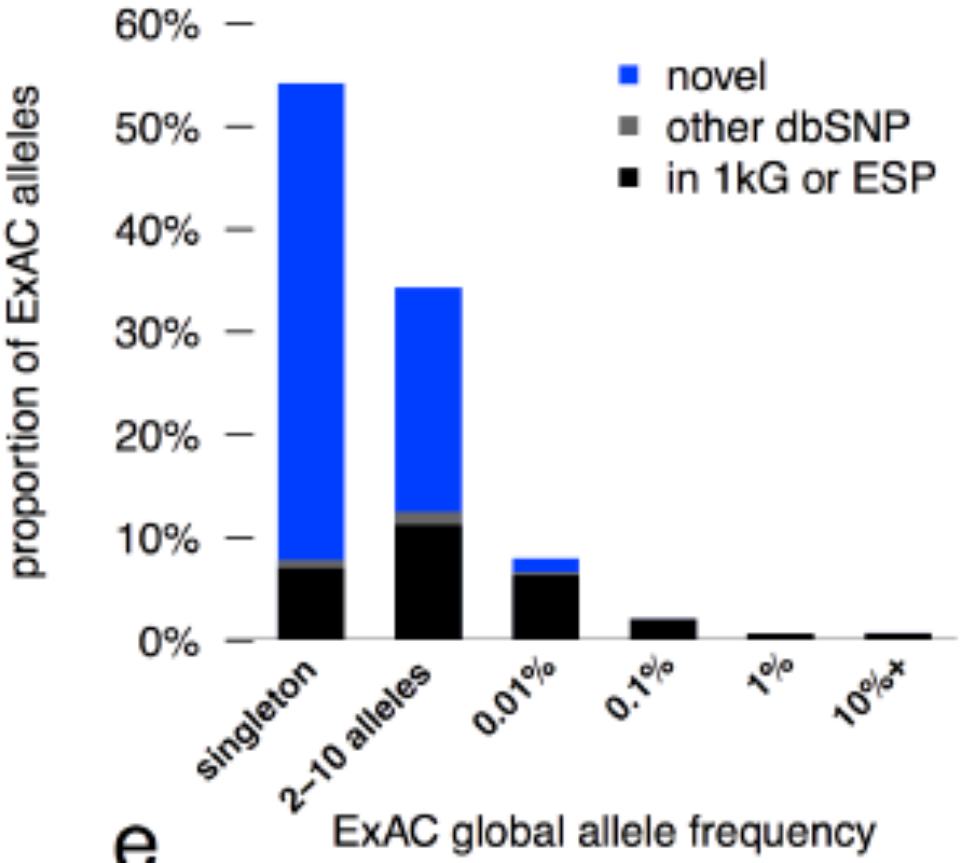
Monkol Lek[1,2,3,4], Konrad J. Karczewski[1,2]*, Eric V. Minikel[1,2,5]*, Kaitlin E. Samocha[1,2,5,6]*, Eric Banks[2], Timothy

Nature 2016



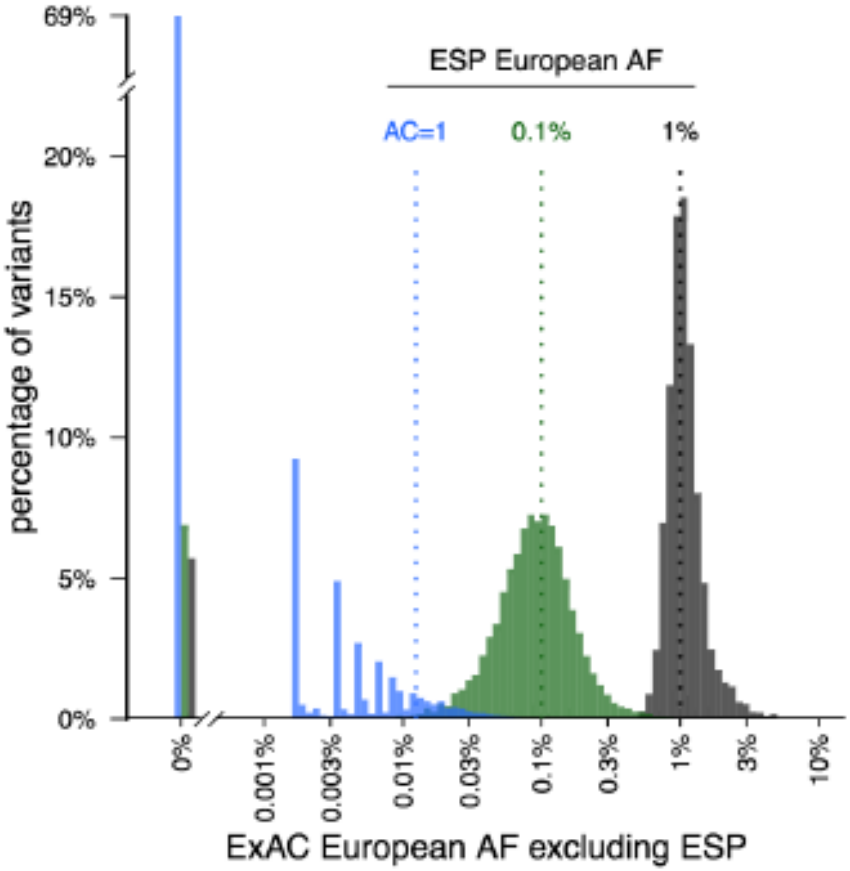
# Basic variant statistics from ExAC

- After filtering, 7.4M variants, of which 317K indels → one variant every 8bp within exons
- 99% have frequency < 1%, 54% are singletons, 72% absent from 1000G+ESP
- 7.9% are have multiple ALT alleles (multiallelic) (cf. <0.5% in 1000G and ESP)

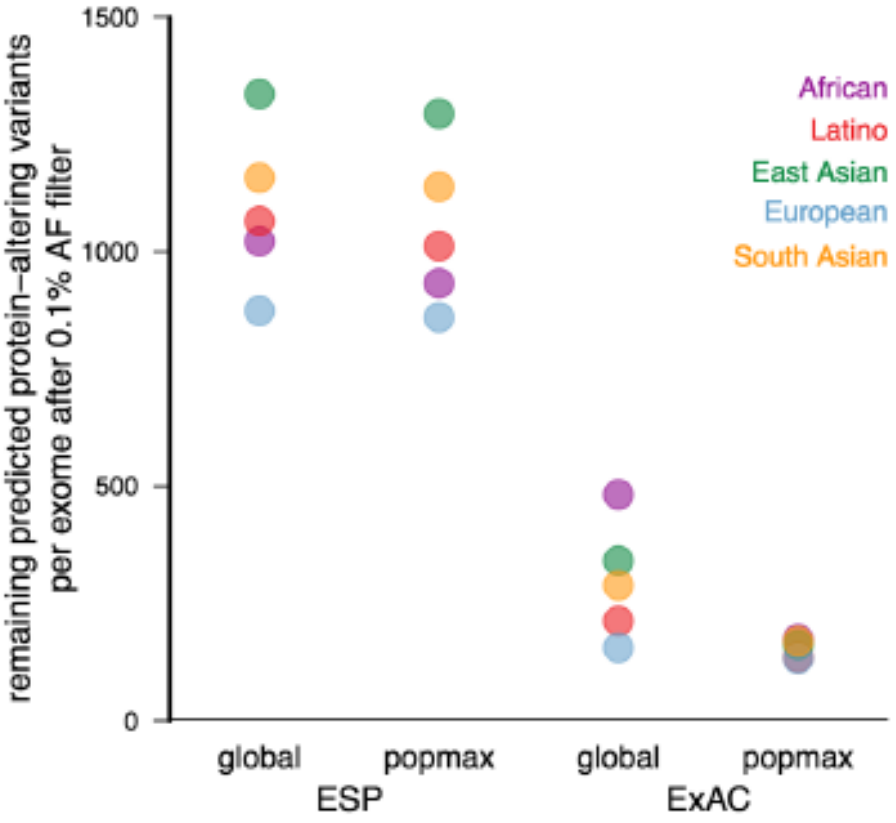


# Use of ExAC for variant interpretation in Mendelian disease

Allele frequency estimates in ESP are unreliable, particularly for very low allele counts (upwardly biased)



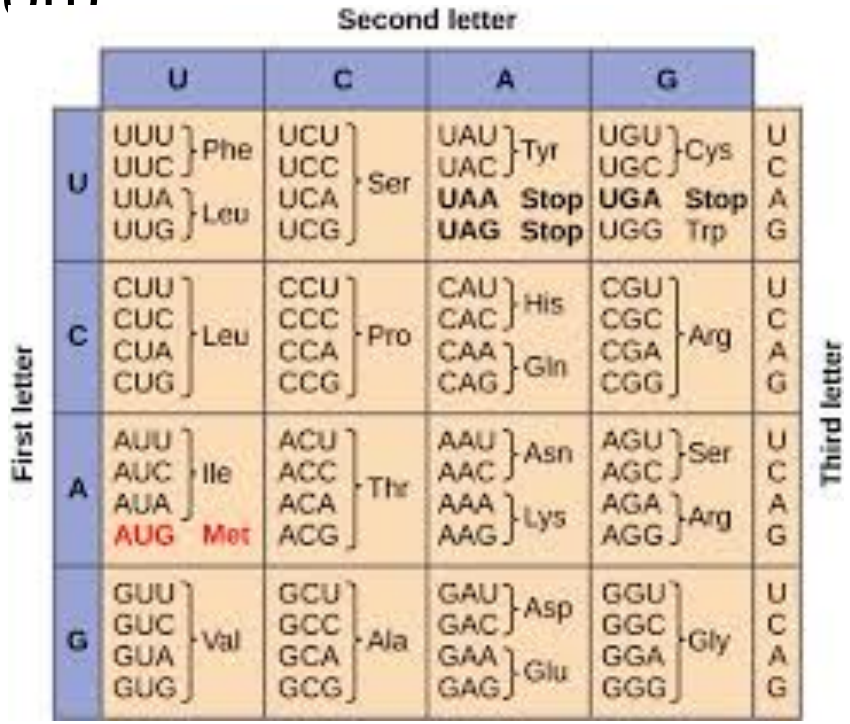

ExAC improves filtering of rare variation compared to ESP



What are the consequences of these variants?
What can we learn about genes?

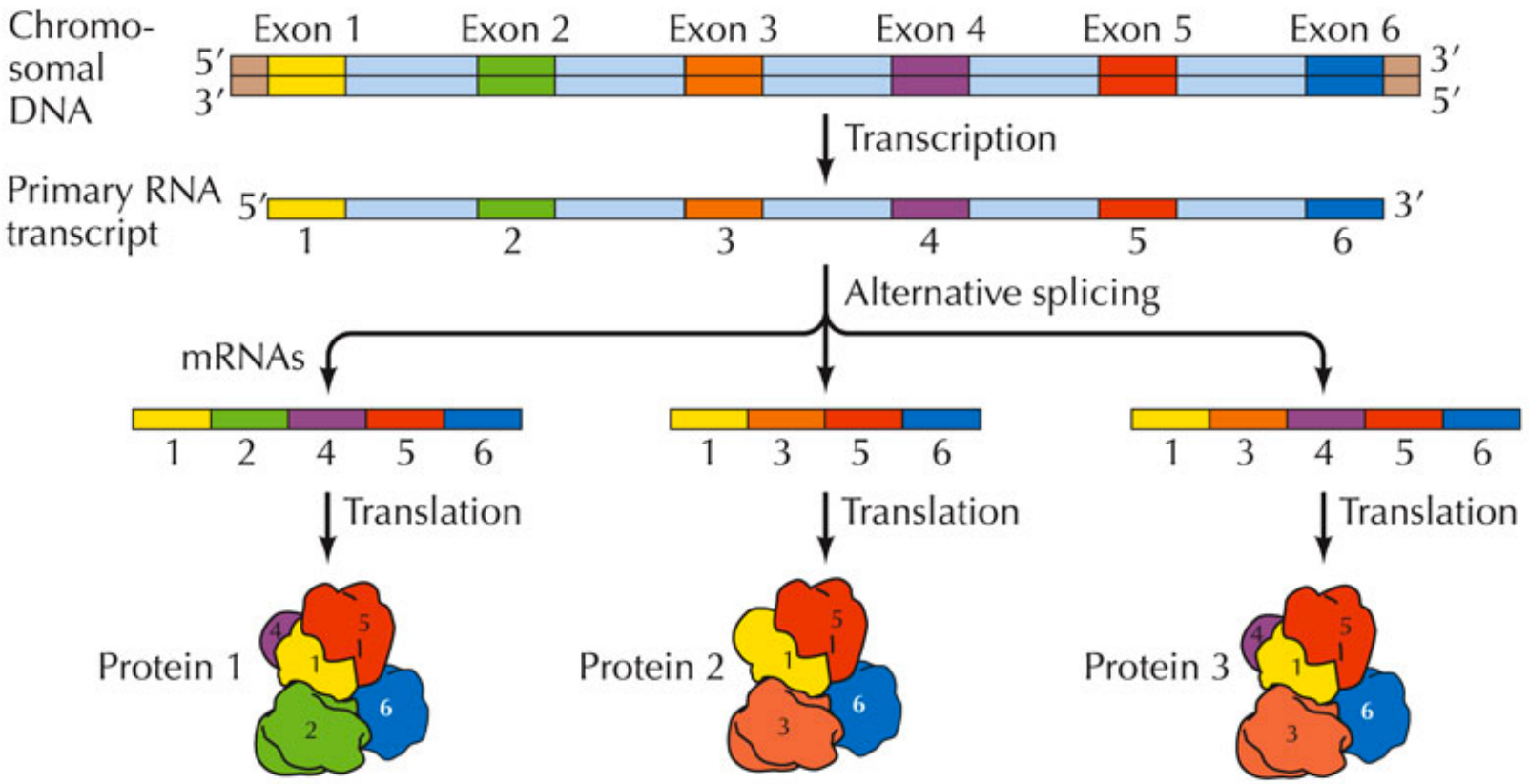
# Exonic variant consequences - revision

- Synonymous (silent) – same amino acid
- Missense (nonsynonymous) – different amino acid
- Nonsense (loss-of-function) – premature stop codon
- Splicing mutation - disrupts splicing (often leading to loss-of-function)




Ben Neale

# Alternative splicing



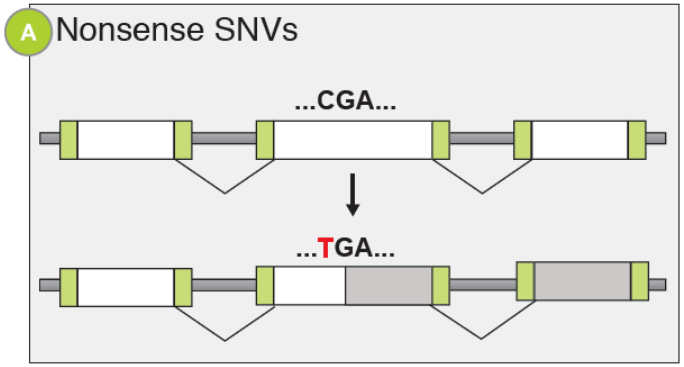
# Annotation

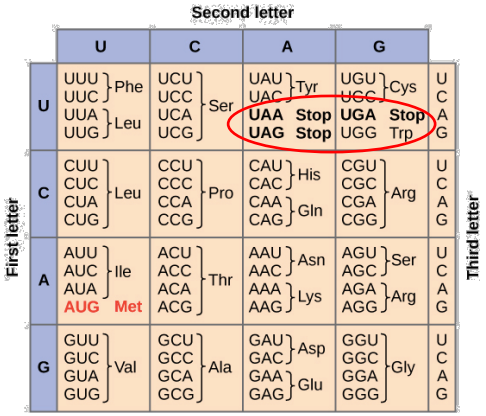
- Process of adding information about frequency, expected functional consequence etc. of variants
- e.g. is the variant found in dbSNP? What is the rs ID? Is it found in 1000 Genomes? At what frequency in each population?
- Functional consequence – synonymous, missense, nonsense, splicing etc.
- Functional consequence often differs depending on transcript (e.g. exon may be present in some both not all transcripts)
- Commonly used tool: Variant Effect Predictor (Ensembl)

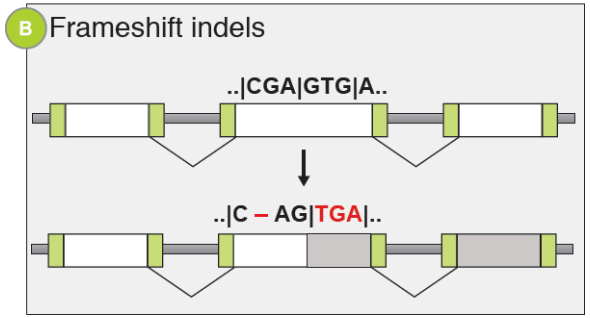
# More on loss-of-function variants (LoFs)

- LoFs are variants that severely affect the function of a protein-coding gene
- Typically do so by deleting it or prompting nonsense-mediated decay (NMD)
- LoFs also called protein truncating variants (PTVs)

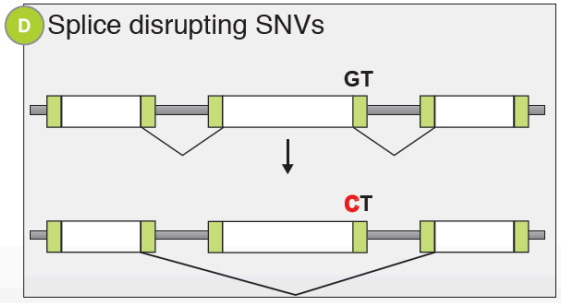
# Different types of LoFs



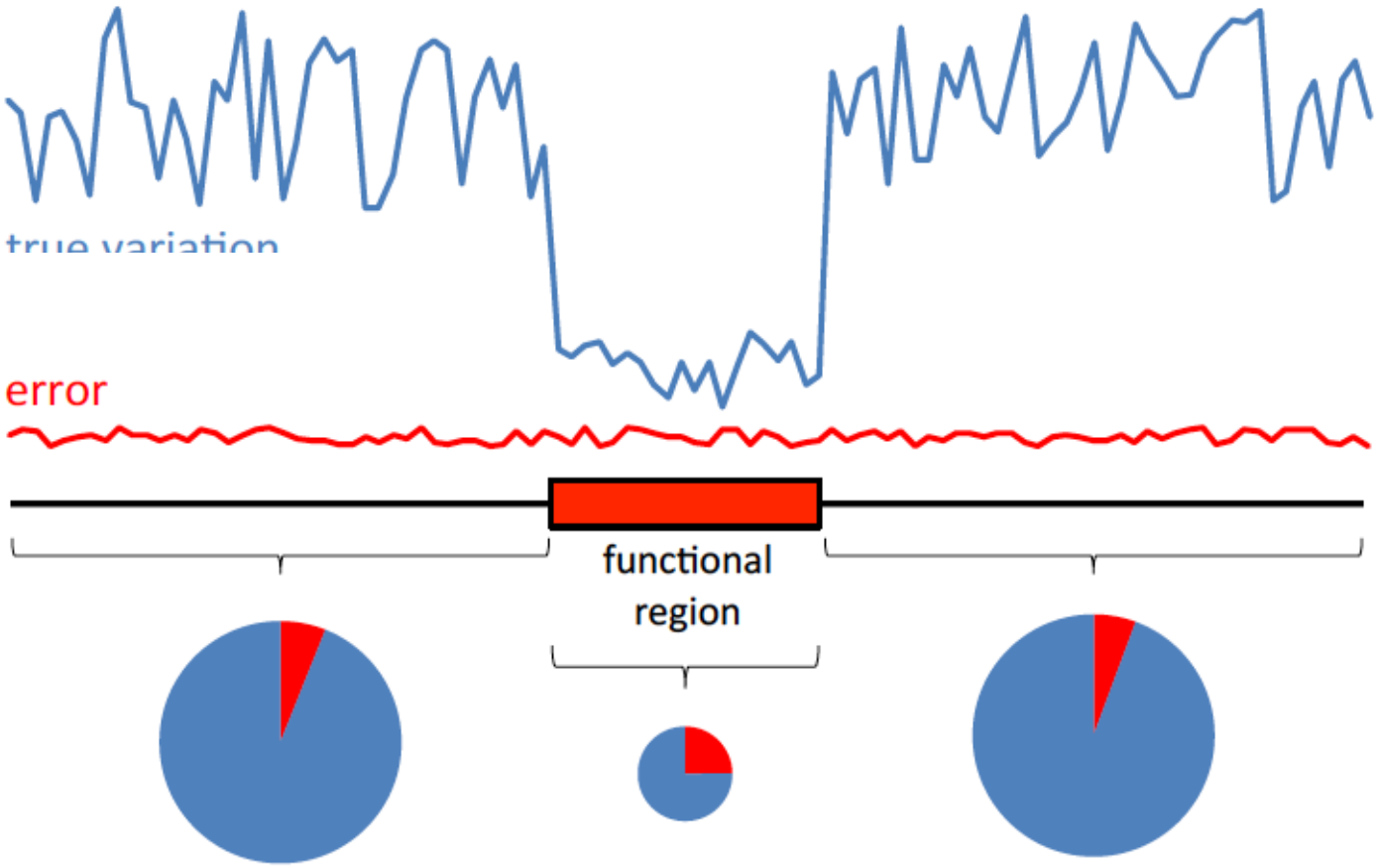




Breaks the GT-AG rule →




- Note all premature stop codons lead to NMD
- LOFTEE – VEP plugin to annotate LoFs as high confidence or low confidence (HC, LC) based on known rules about which variants actually lead to NMD

# Challenges in identifying true LoFs

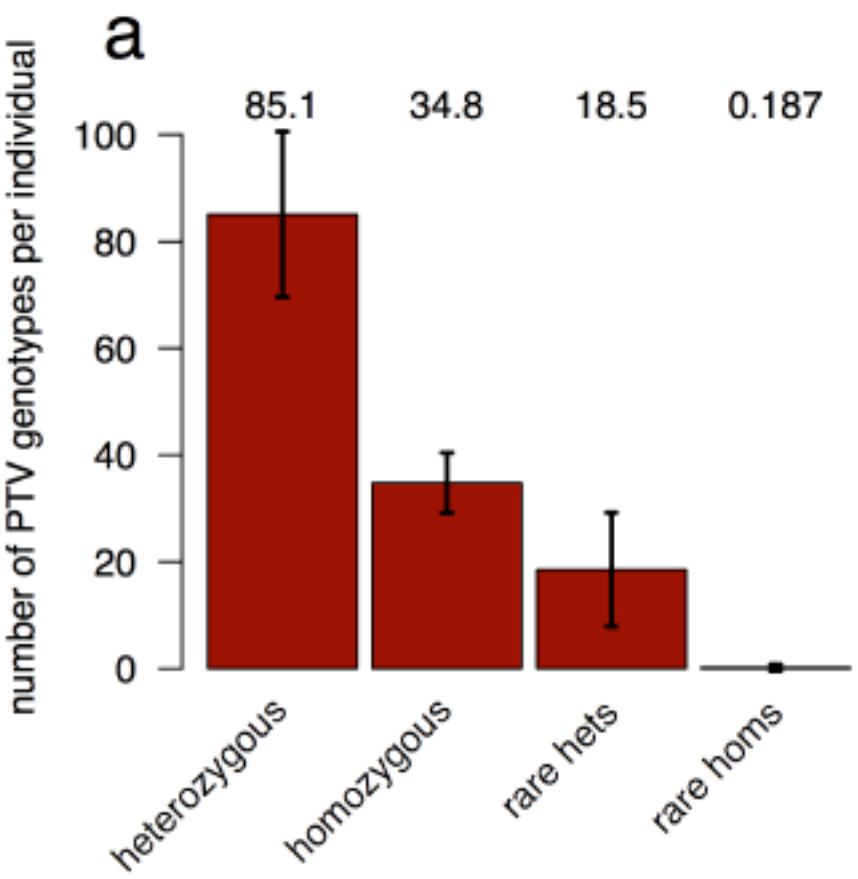


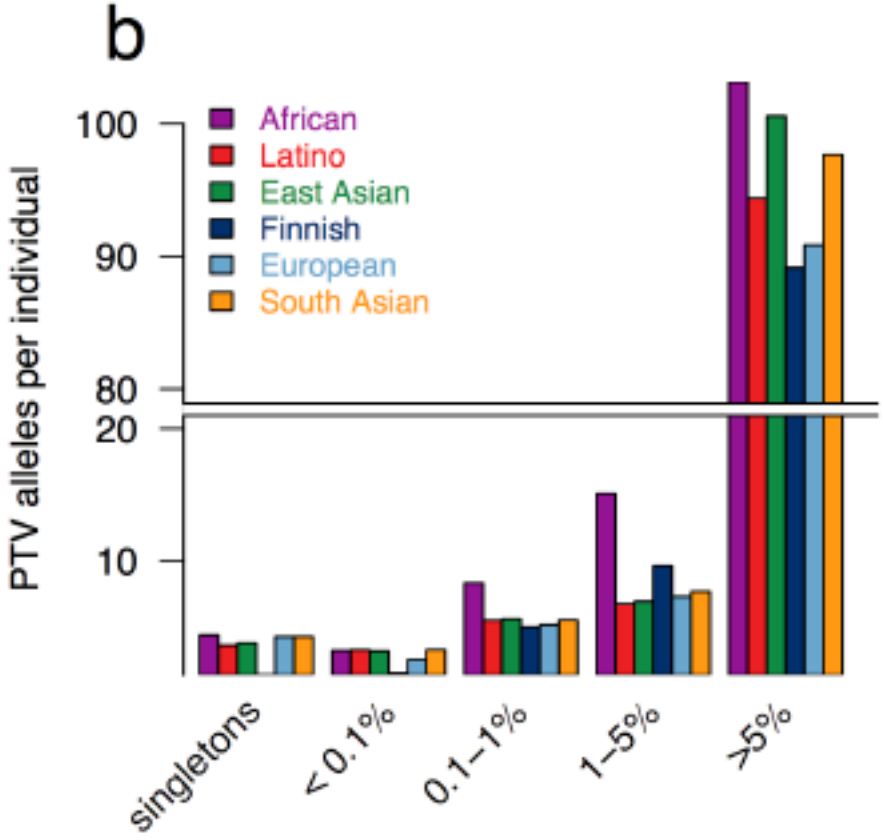

- the fraction of variants that are sequencing/calling errors is higher for LoFs than other types of variants

Konrad Karczewski

# Loss-of-function variants in ExAC

- 180K LoFs, of which 121K are singletons
- Most LoFs are common; each individual has ~2 singleton LoFs

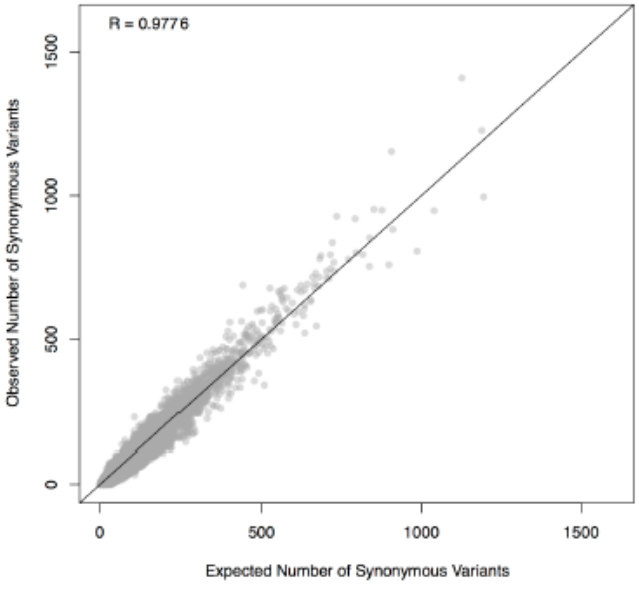


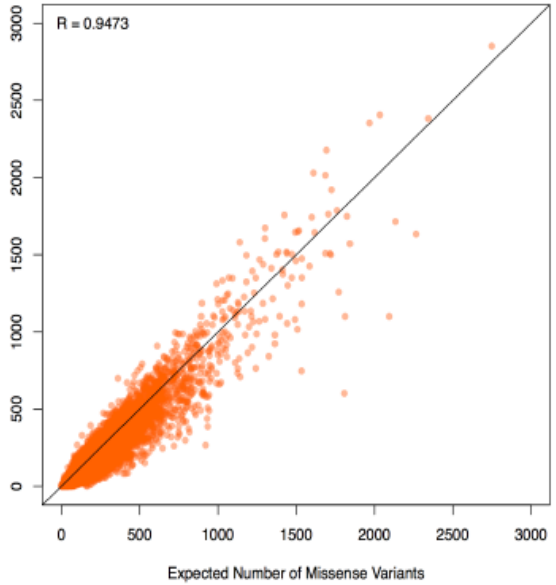
# Inferring gene constraint using ExAC data

- Relies on ratio of # observed to # expected variants in a gene
- Determining # expected variants relies on model for mutation rate in different sequence contexts - see Samocha *et al*. (Nat Gen, 2014) for details
- Model does well at predicting # rare synonymous variants, bue less well for missense and LoFs due to selective constraint

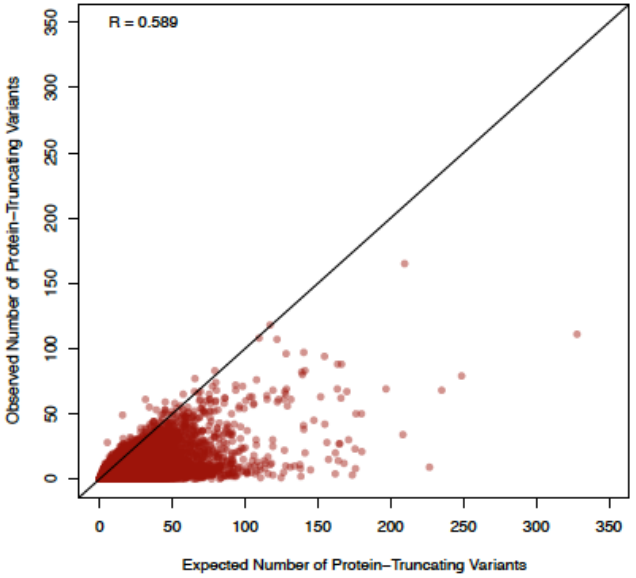
Synonymous



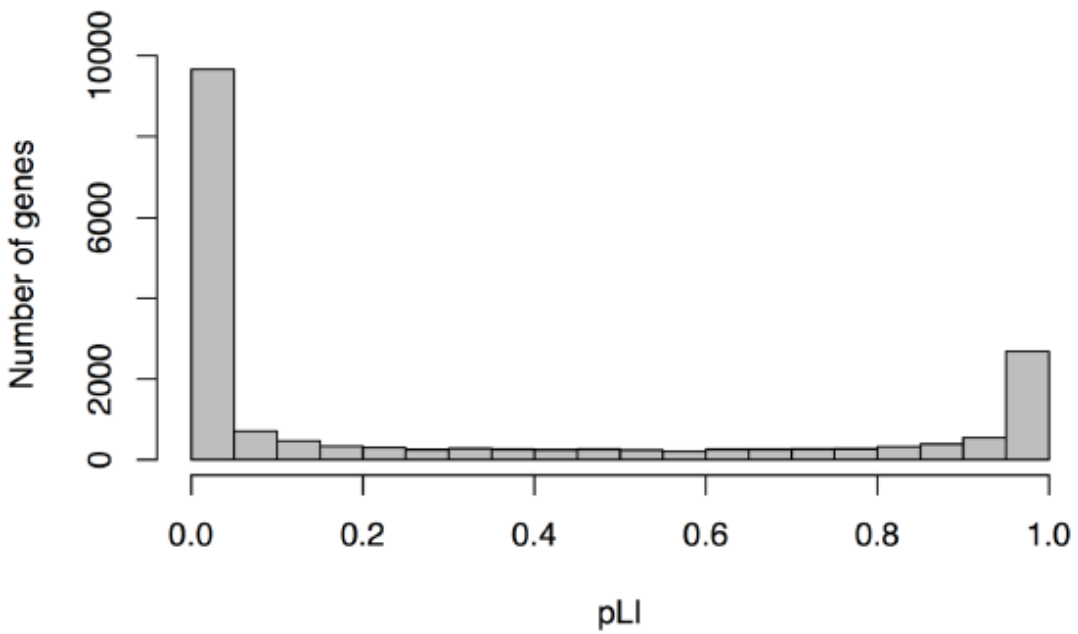

Missense

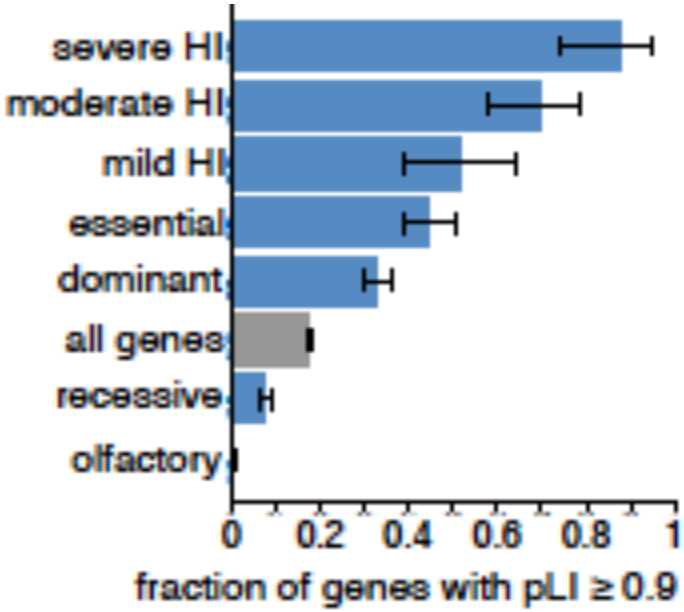



LoF



# pLI: probability of loss-of-function intolerance

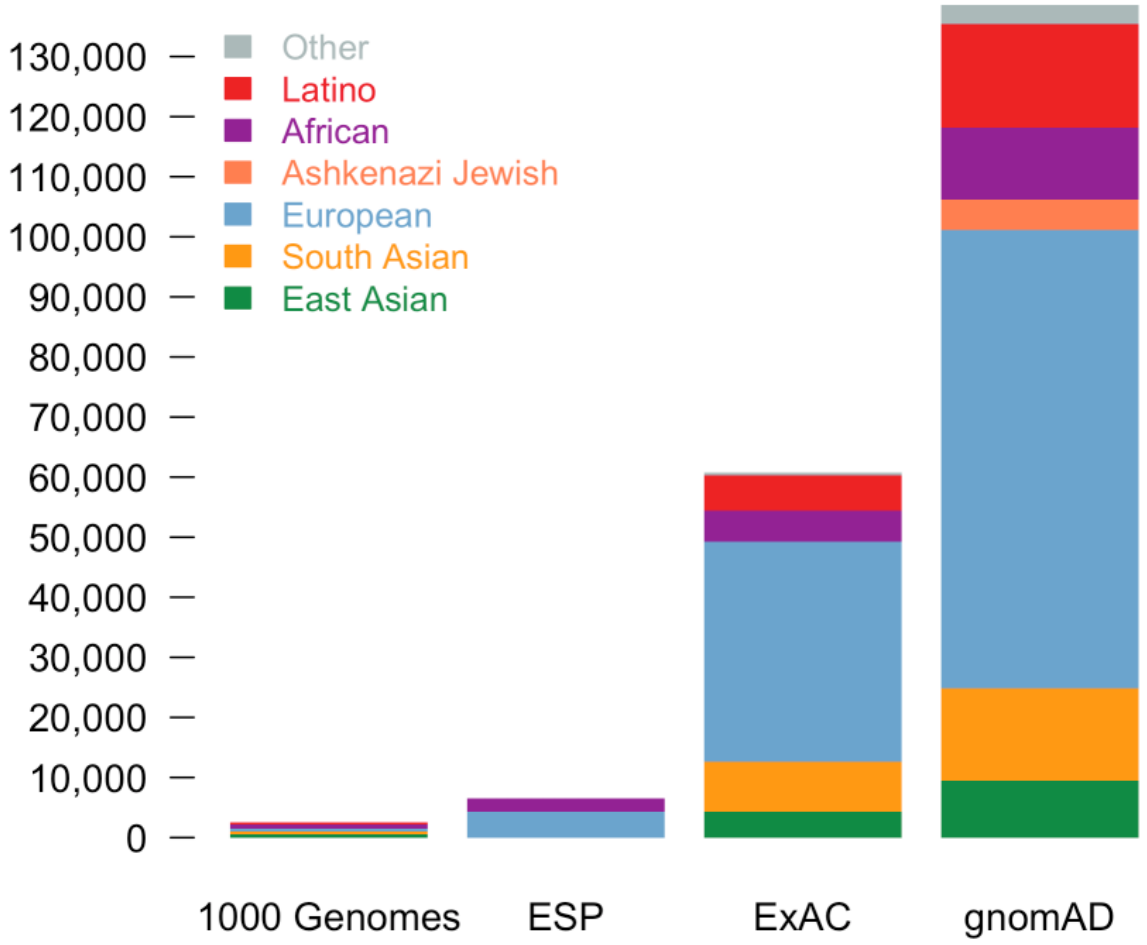



- pLI less correlated with coding sequence length than LoF Z-score (r=0.17 vs. 0.57)
- 10,374 LoF-tolerant genes (pLI ≤ 0.1)
- 3,230 LoF-intolerant genes (pLI ≥ 0.9) → includes almost all known severe haploinsufficient (HI) disease genes; 79% have not yet been assigned a human disease phenotype (could be embryonic lethal, or patients not found yet)

# gnomAD: the new, bigger version of ExAC

**Also ~15,000 jointly-called whole genomes**



# Limitations in using ExAC and gnomAD

- differences in coverage, mapping, variant calling or QC between your dataset and theirs may lead to misestimation of allele frequency for variants in some regions
- these differences become very apparent when doing exome-wide analyses
- beware poorly matched ancestry e.g. a singleton in ExAC may be more common in a tiny Swiss village
- not necessarily useful as controls for complex disease studies because have not been screened for those phenotypes

# Practical

- Variant Effect Predictor (VEP)
- ExAC
- Ensembl for viewing variant frequencies and consequences, and LD structure