wellcome trust sanger institute

Introduction to sequencing

Hilary Martin

Wellcome Sanger Institute

Hinxton (near Cambridge), UK

Plan for lecture

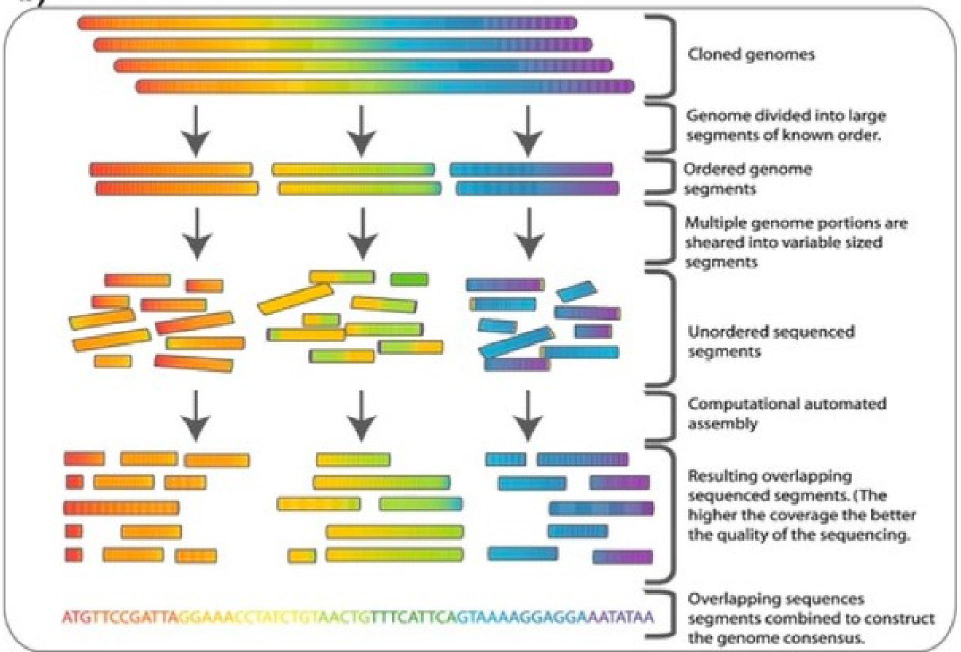
- The sequencing revolution
- Technical aspect of sequencing studies
    - Coverage
    - Exomes versus genomes
    - Alignment
    - Variant calling
    - Quality control
    - Contamination
- Variant consequences and annotation
- Interpretation of *de novo* mutations
- Importance of well-matched controls

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

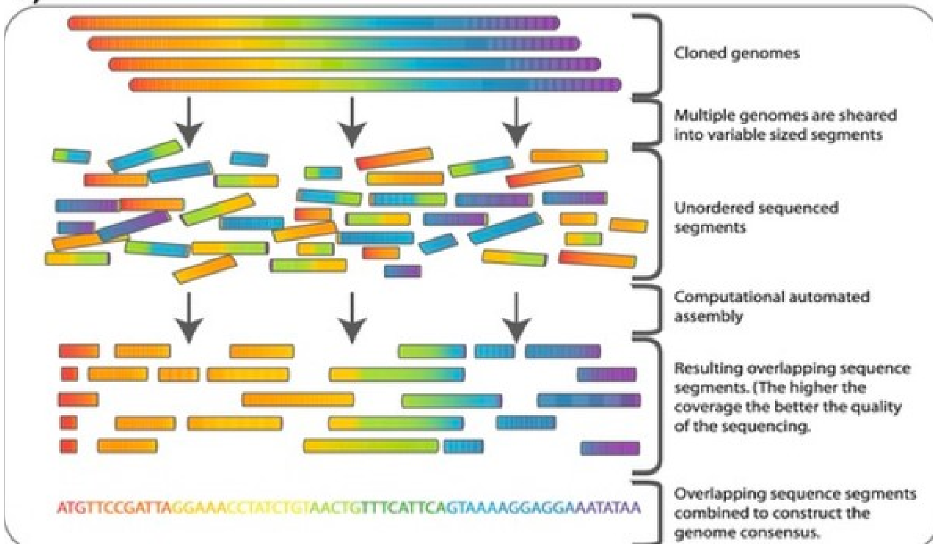
Hierarchical shotgun sequencing

b)

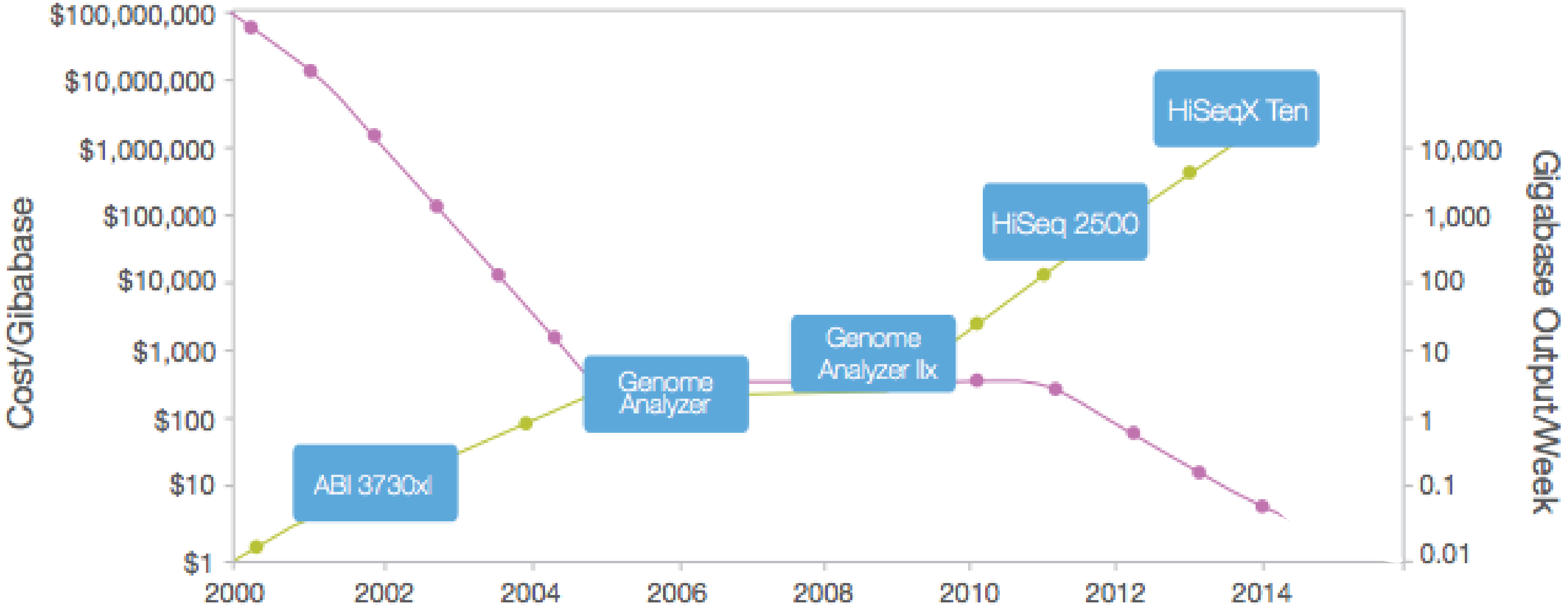

Cloned genomes
Genome divided into large segments of known order.
Ordered genome segments
Multiple genome portions are sheared into variable sized segments
Unordered sequenced segments
Computational automated assembly
Resulting overlapping sequenced segments. (The higher the coverage the better the quality of the sequencing.
ATGTTCCGATTAGGAAACCTATCTGTAACTGTTTCATTCAGTAAAAGGAGGAAATATAA
Overlapping sequences segments combined to construct the genome consensus.

Whole-genome shotgun sequencing

a)


Cloned genomes
Multiple genomes are sheared into variable sized segments
Unordered sequenced segments
Computational automated assembly
Resulting overlapping sequence segments. (The higher the coverage the better the quality of the sequencing.
ATGTTCCGATTAGGAAACCTATCTGTAACTGTTTCATTCAGTAAAAGGAGGAAATATAA
Overlapping sequence segments combined to construct the genome consensus.

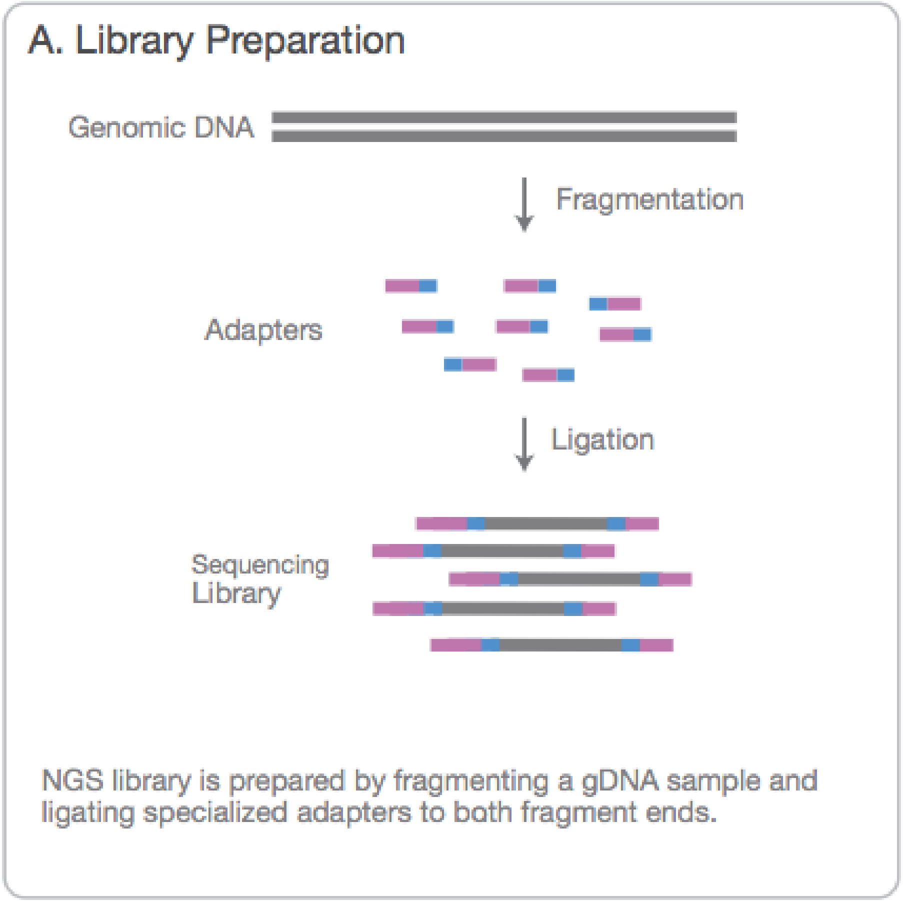
Cost of sequencing



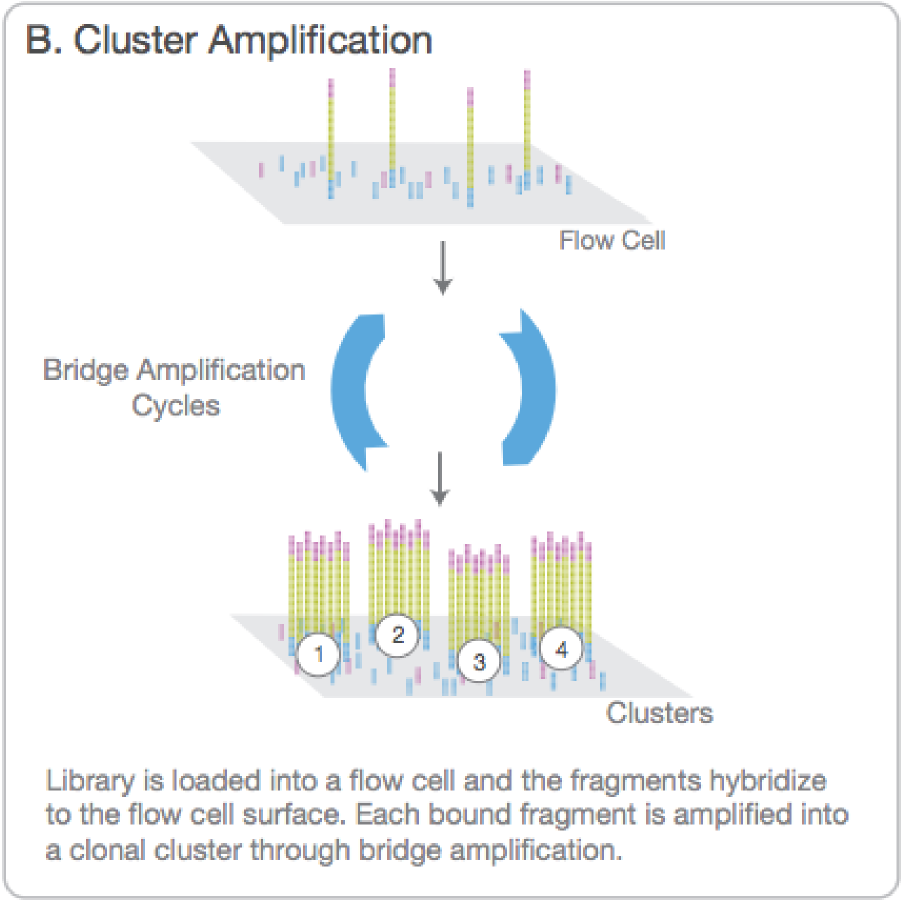
- Reminder: human genome 3 Gigabases
- Due to errors, we tend to sequence 20-30X to obtain high quality sequence i.e. 60-90Gb → currently ~$1000/genome

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

Illumina sequencing



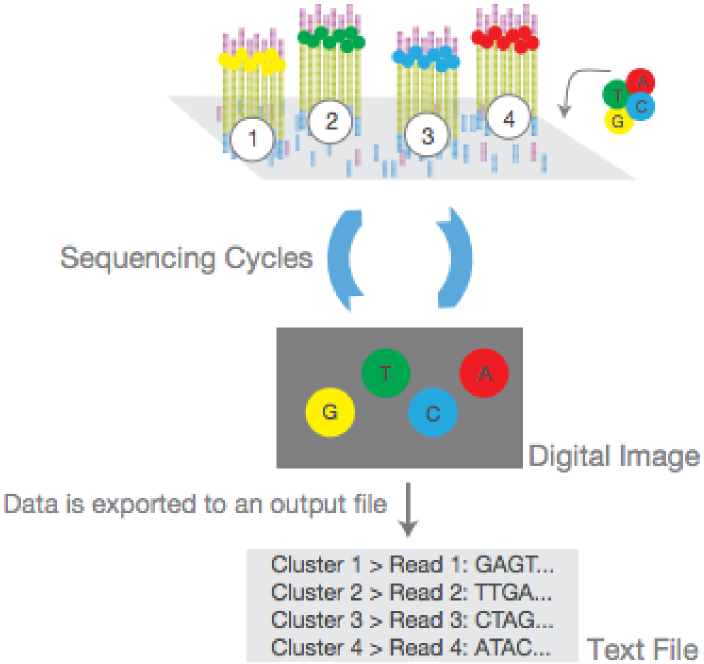
NGS library is prepared by fragmenting a gDNA sample and ligating specialized adapters to both fragment ends.



Library is loaded into a flow cell and the fragments hybridize to the flow cell surface. Each bound fragment is amplified into a clonal cluster through bridge amplification.

Illumina sequencing

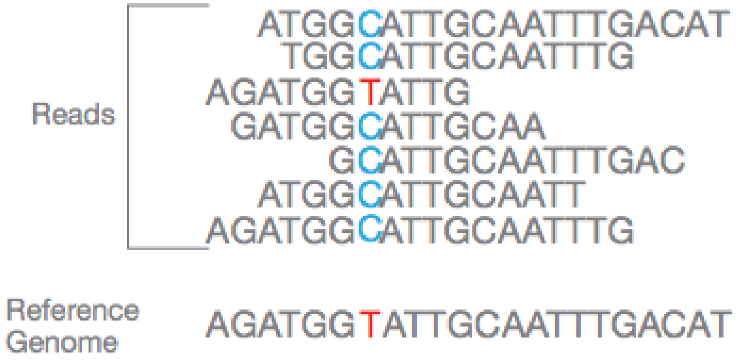
C. Sequencing



```
Cluster 1 > Read 1: GAGT...
Cluster 2 > Read 2: TTGA...
Cluster 3 > Read 3: CTAG...
Cluster 4 > Read 4: ATAC...
```

Sequencing reagents, including fluorescently labeled nucleotides, are added and the first base is incorporated. The flow cell is imaged and the emission from each cluster is recorded. The emission wavelength and intensity are used to identify the base. This cycle is repeated "n" times to create a read length of "n" bases.

D. Alignment & Data Anaylsis

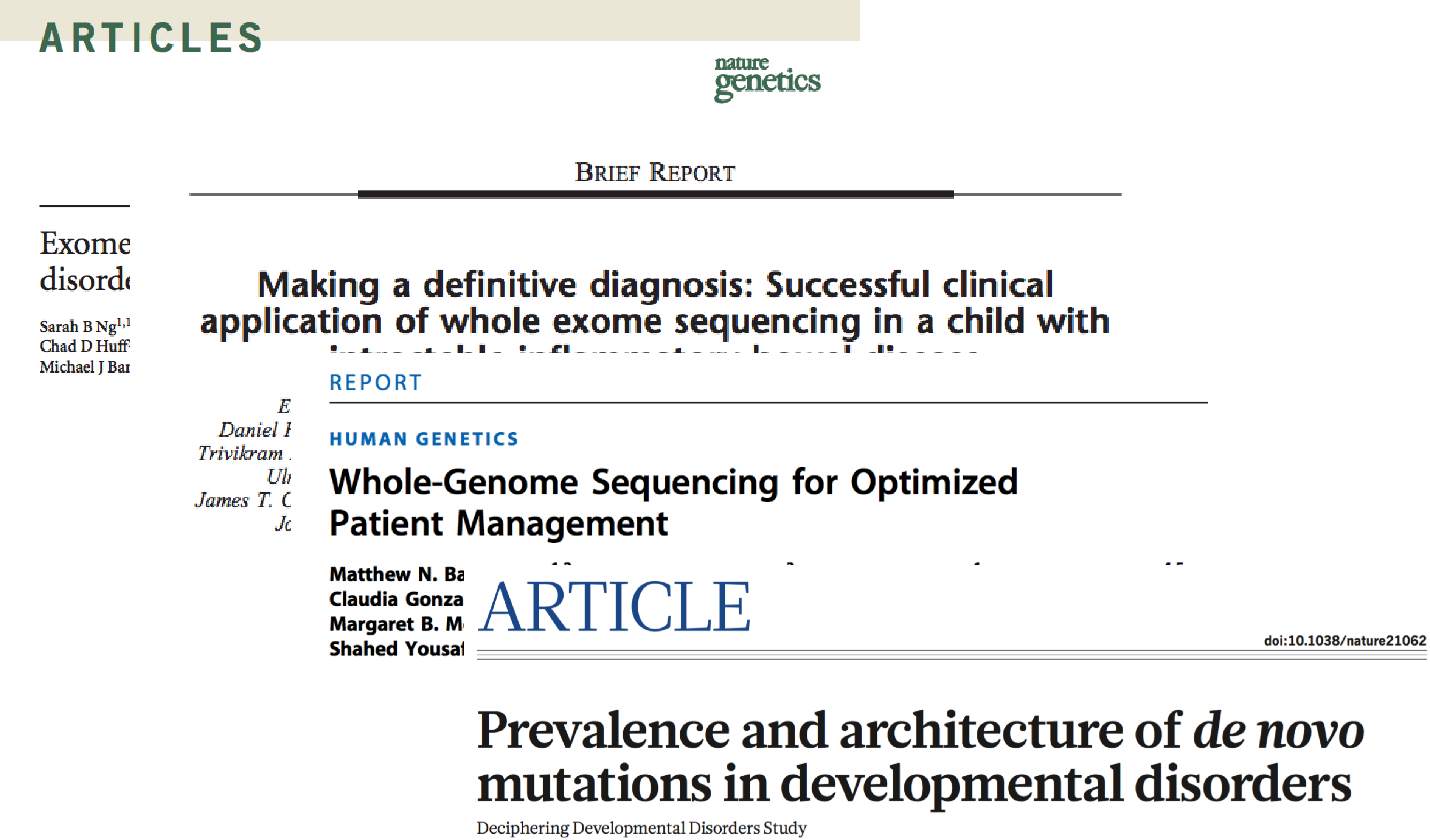


```
  ATGGCATTGCAATTTGACAT
   TGGCATTGCAATTTG
AGATGGTATTG
 GATGGCATTGCAA
     GCATTGCAATTTGAC
  ATGGCATTGCAATT
AGATGGCATTGCAATTTG

AGATGGTATTGCAATTTGACAT
```

Reads are aligned to a reference sequence with bioinformatics software. After alignment, differences between the reference genome and the newly sequenced reads can be identified.

Direct sequencing has enormous potential



ARTICLES

nature genetics

Exome disord

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

BRIEF REPORT

Making a definitive diagnosis: Successful clinical application of whole exome sequencing in a child with intractable inflammatory bowel disease

E
Daniel I
Trivikram .
Uli
James T. C
Jo

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

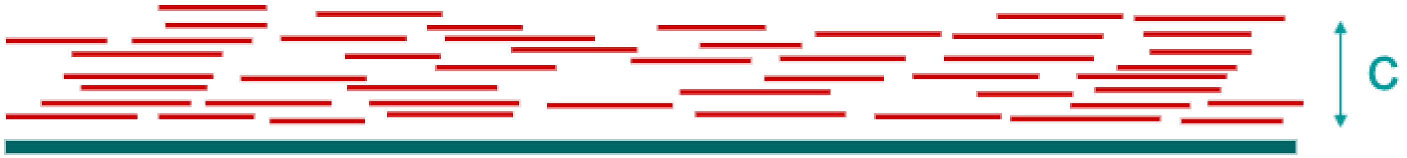
For very few of these do we have any conclusive understanding of their medical impact in the population

Plan for lecture

- The sequencing revolution
- Technical aspect of sequencing studies
  - Coverage
  - Exomes versus genomes
  - Alignment
  - Variant calling
  - Quality control
  - Contamination
- Variant consequences and annotation
- Interpretation of *de novo* mutations
- Importance of well-matched controls

Coverage

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



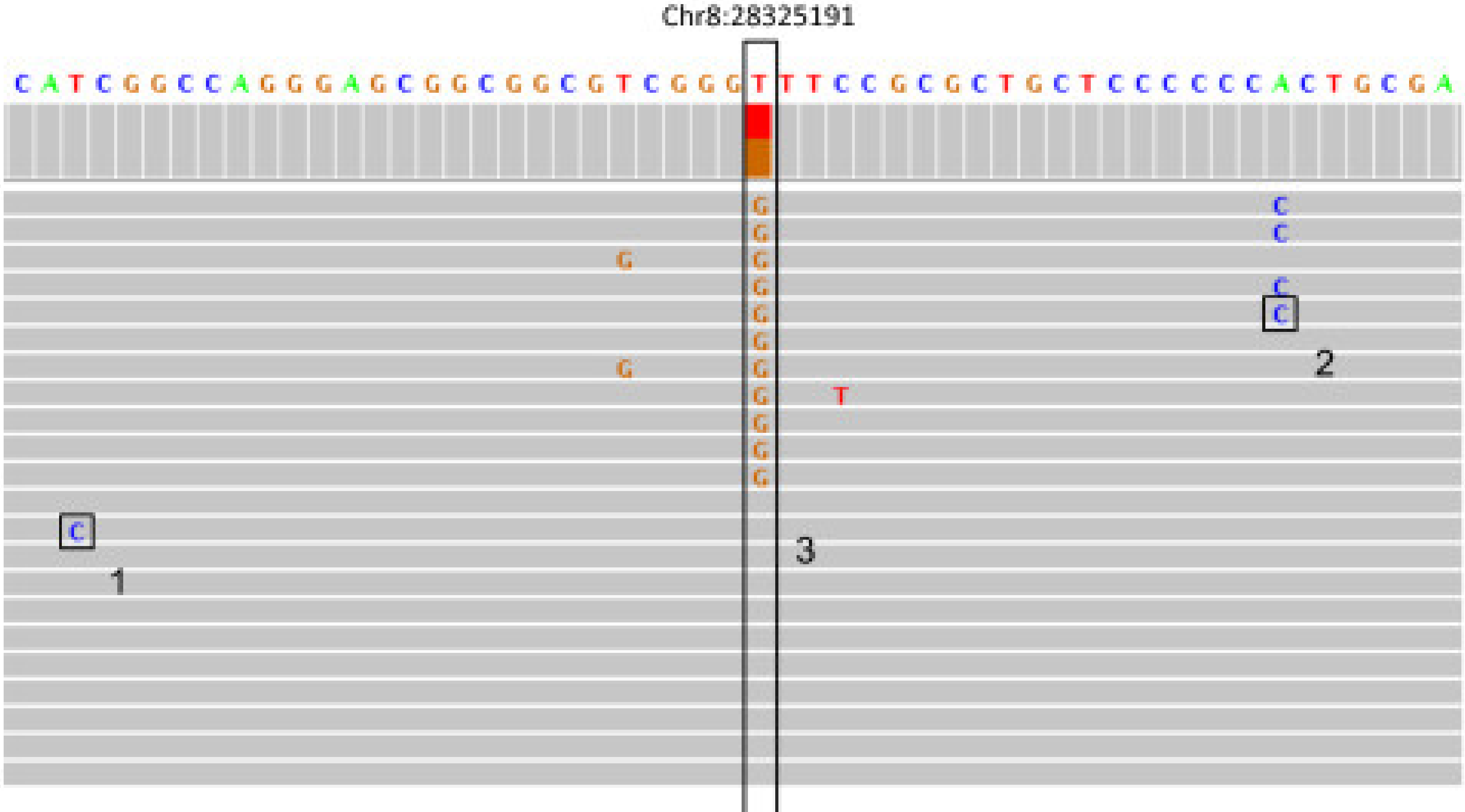
Length of genomic segment:	L
Number of reads:	n
Length of each read:	l

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.



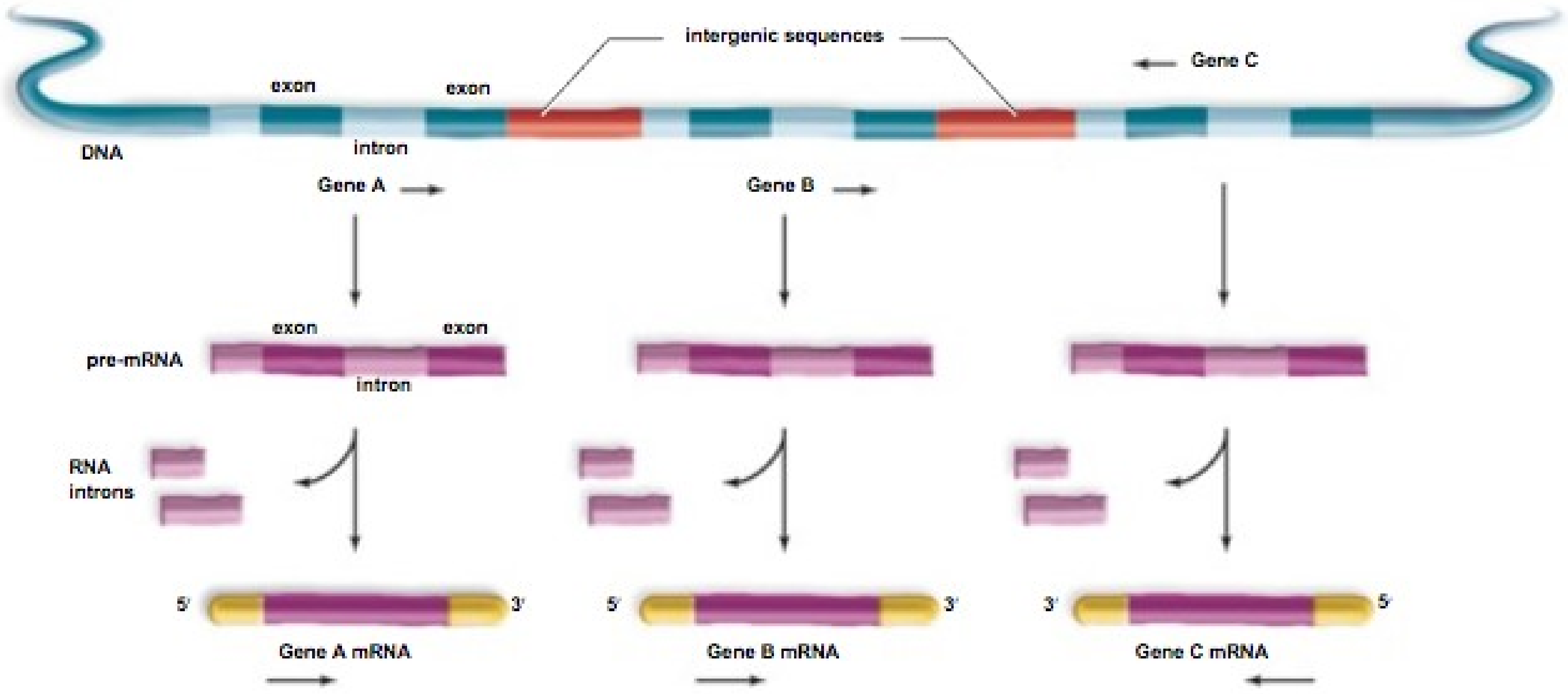
Plan for lecture

- The sequencing revolution
- Technical aspect of sequencing studies
  - Coverage
  - Exomes versus genomes
  - Alignment
  - Variant calling
  - Quality control
  - Contamination
- Variant consequences and annotation
- Interpretation of *de novo* mutations
- Importance of well-matched controls

Technologies for sequencing humans

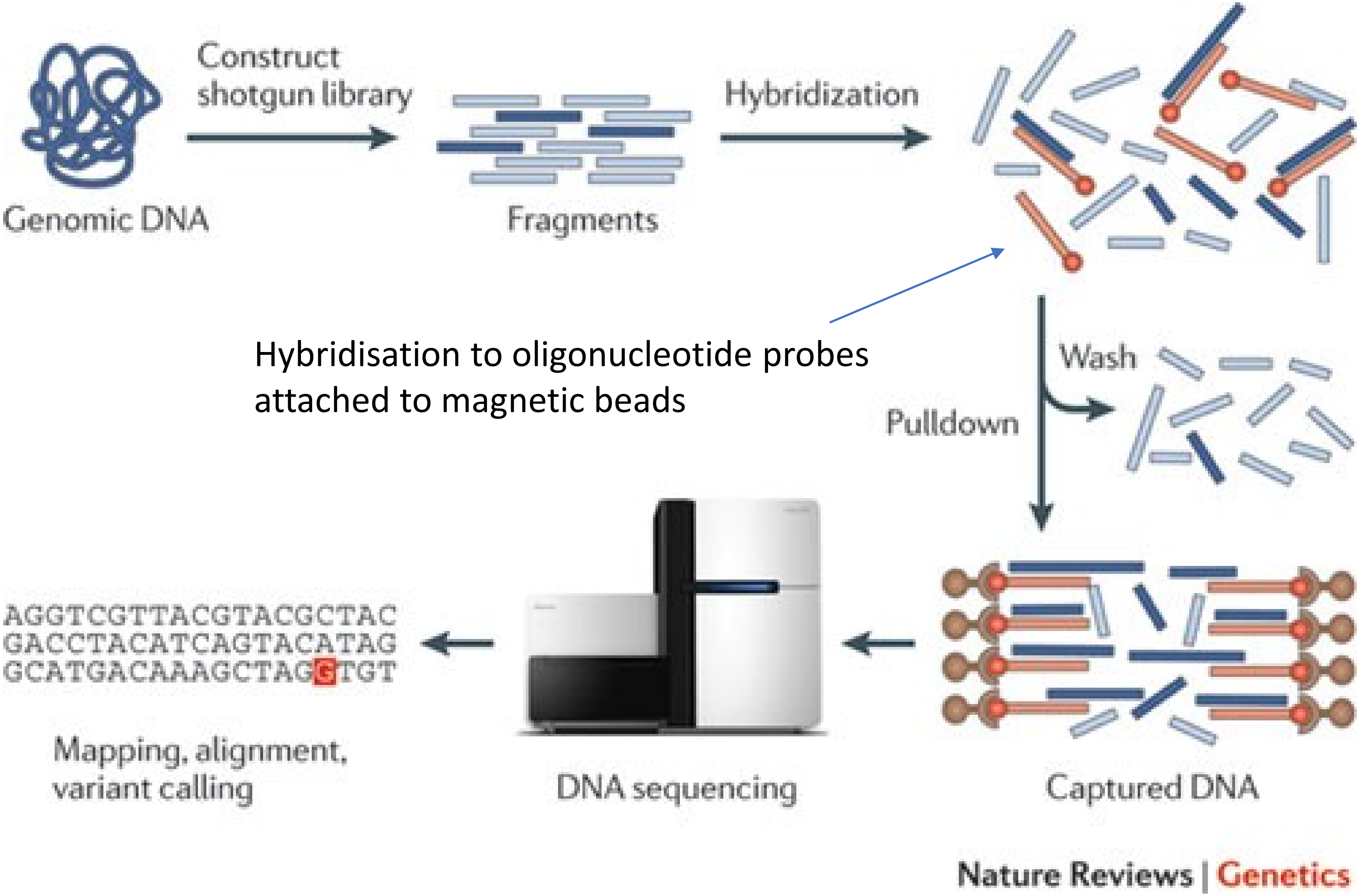
	Whole-genome sequencing (WGS)	**Whole-exome sequencing (WES)**
Amount of sequence	3Gb	30Mb
Typical coverage	30X (for high quality)	Average 60-180X
Library preparation	Randomly shear, then do hybridisation-based capture of exonic DNA fragments	Shotgun sequence - randomly shear and capture
Advantages	•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

The exome



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

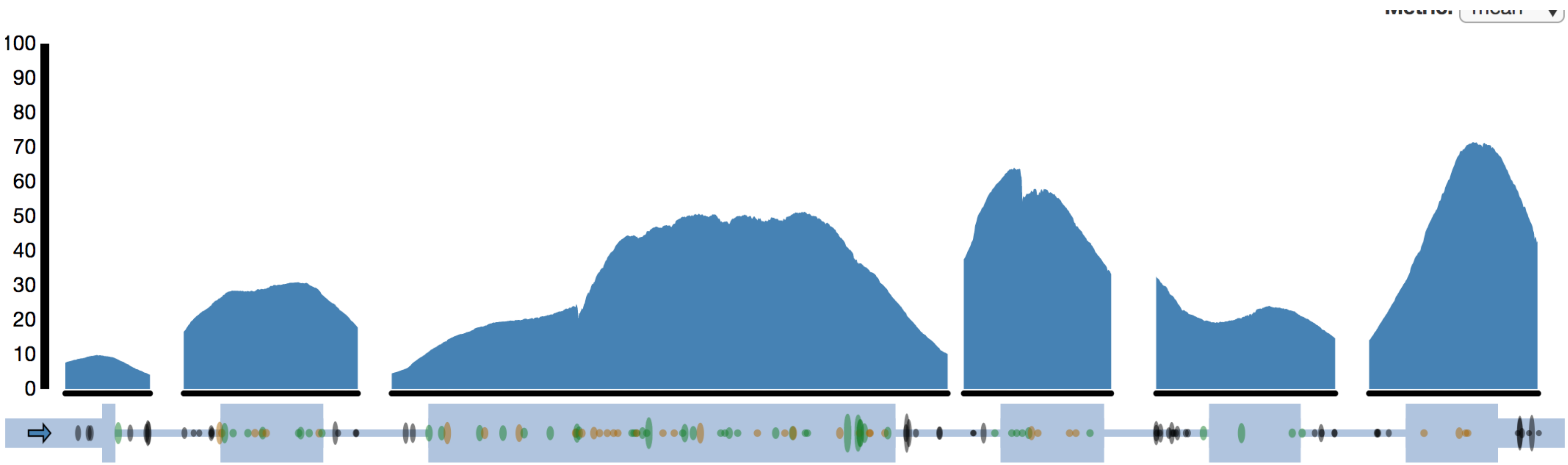
Targeted exome capture



Hybridisation to oligonucleotide probes attached to magnetic beads

Bamshad et al., Nature Review Genetics, 2011

Variable coverage in exome sequencing



- Reference bias: we tend to observe more reads mapping to the reference allele than the alternate allele
- WES shows a greater reference bias than WGS (53% versus 50.3%) – due to capture probes as well as mapping bias

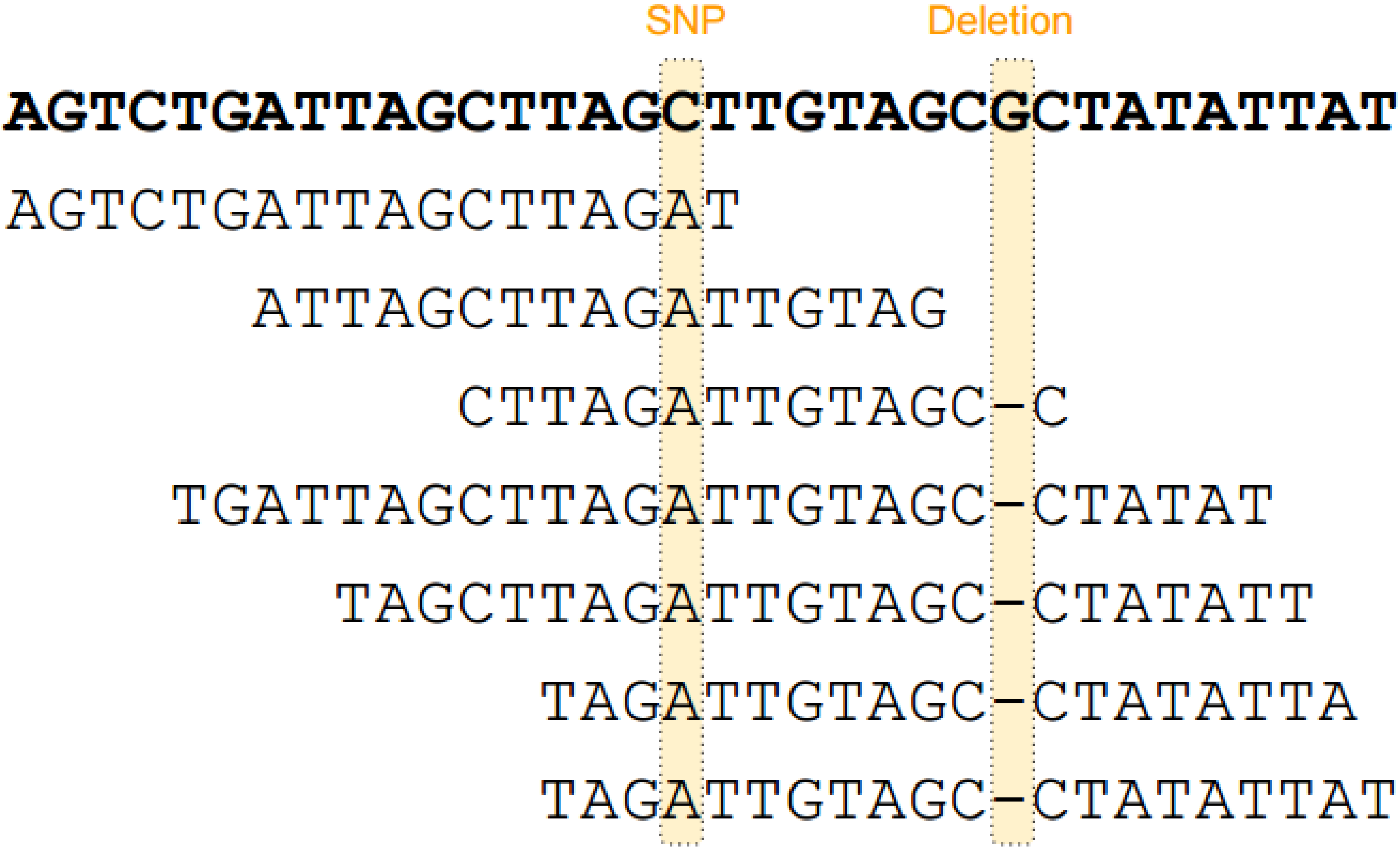
Depth considerations

- Mendelian disease - need high coverage to be sure rare/*de novo* variants are real (20-30X WGS, or >60X WES)
- Complex disease
  - High coverage needed to interrogate rare variants – 15X now considered to get a good balance between sensitivity and specifitiy
  - 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
- Somatic mutations – variants in <<50% of reads, so need high coverage (often >100X for tumours)

Plan for lecture

- The sequencing revolution
- Technical aspect of sequencing studies
  - Coverage
  - Exomes versus genomes
  - Alignment
  - Variant calling
  - Quality control
  - Contamination
- Variant consequences and annotation
- Interpretation of *de novo* mutations
- Importance of well-matched controls

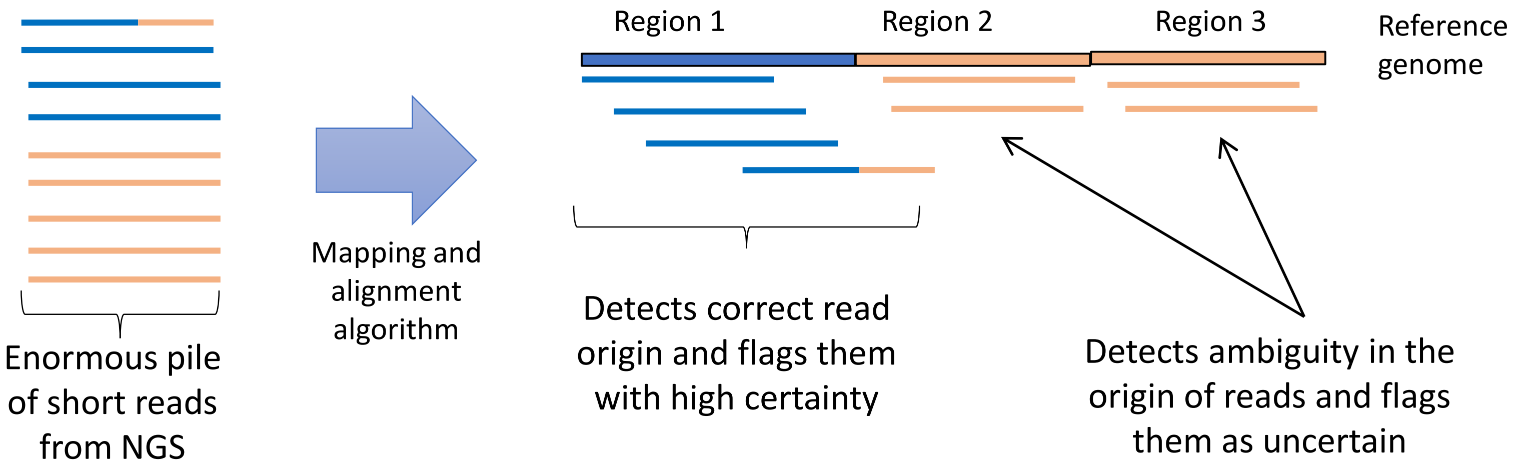
Step 1: Aligning to a reference



```
                SNP     Deletion
**AGTCTGATTAGCTTAGCTTGTAGCGCTATATTAT**
AGTCTGATTAGCTTAGAT
      ATTAGCTTAGATTGTAG
           CTTAGATTGTAGC-C
    TGATTAGCTTAGATTGTAGC-CTATAT
        TAGCTTAGATTGTAGC-CTATATT
             TAGATTGTAGC-CTATATTA
             TAGATTGTAGC-CTATATTAT
```

Torsten Seemann

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



Region 1

Region 2

Region 3

Reference genome

Enormous pile of short reads from NGS

Mapping and alignment algorithm

Detects correct read origin and flags them with high certainty

Detects ambiguity in the origin of reads and flags them as uncertain

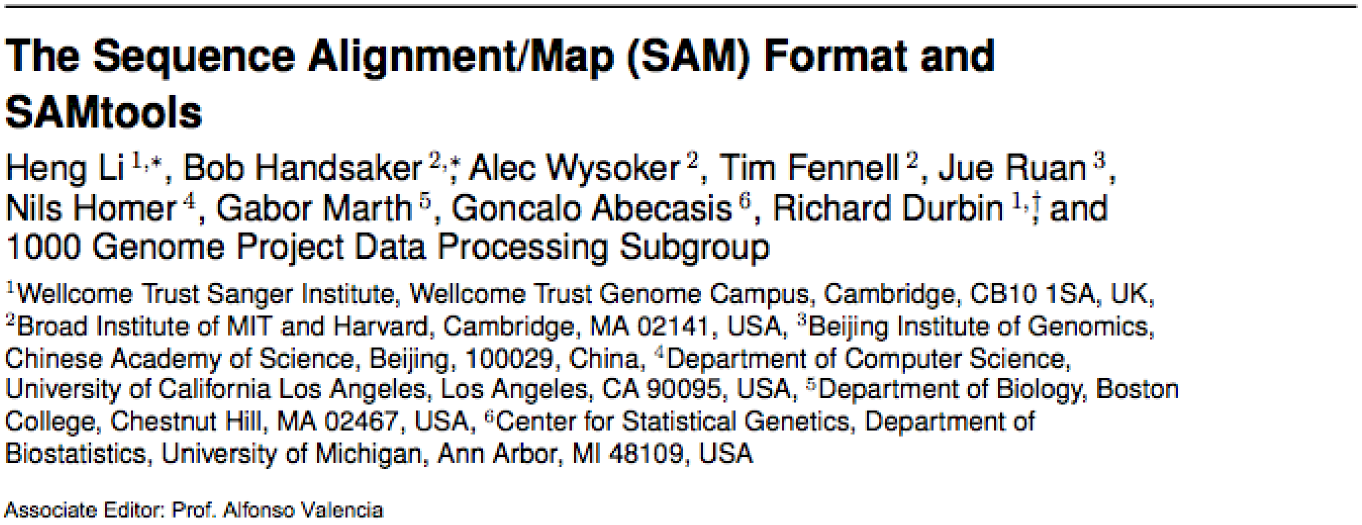
- Many different alignment programs
- Commonly used aligner: BWA-MEM (Li and Durbin) - robust, accurate ‘gold standard’

SAM/BAM/CRAM files

Ben Neale

The SAM/BAM/CRAM file format

- file format was designed to capture all of the critical information about next-generation sequencing 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/
- BAM and CRAM files are compressed versions of SAM



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

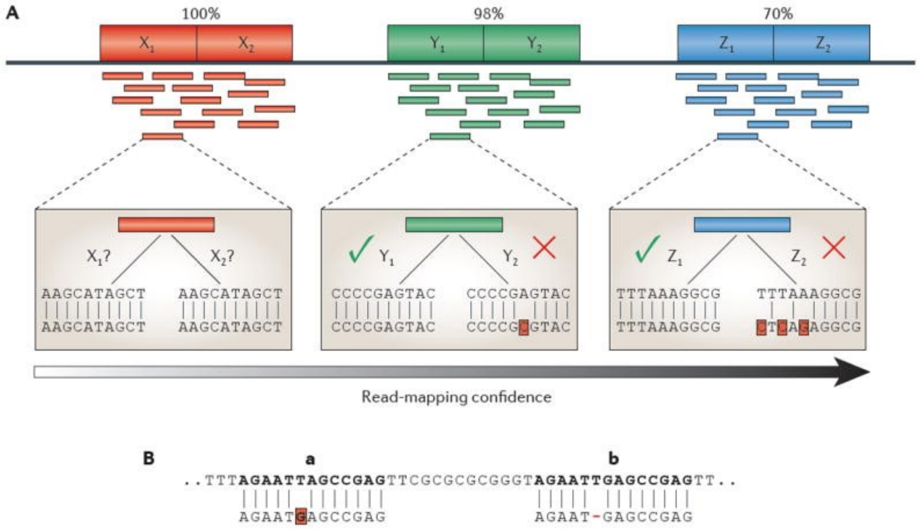
Ben Neale

Repeats cause problems with sequence data

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

Reference: TAGTAGTAGTAGTAGTAGTAGTAGT

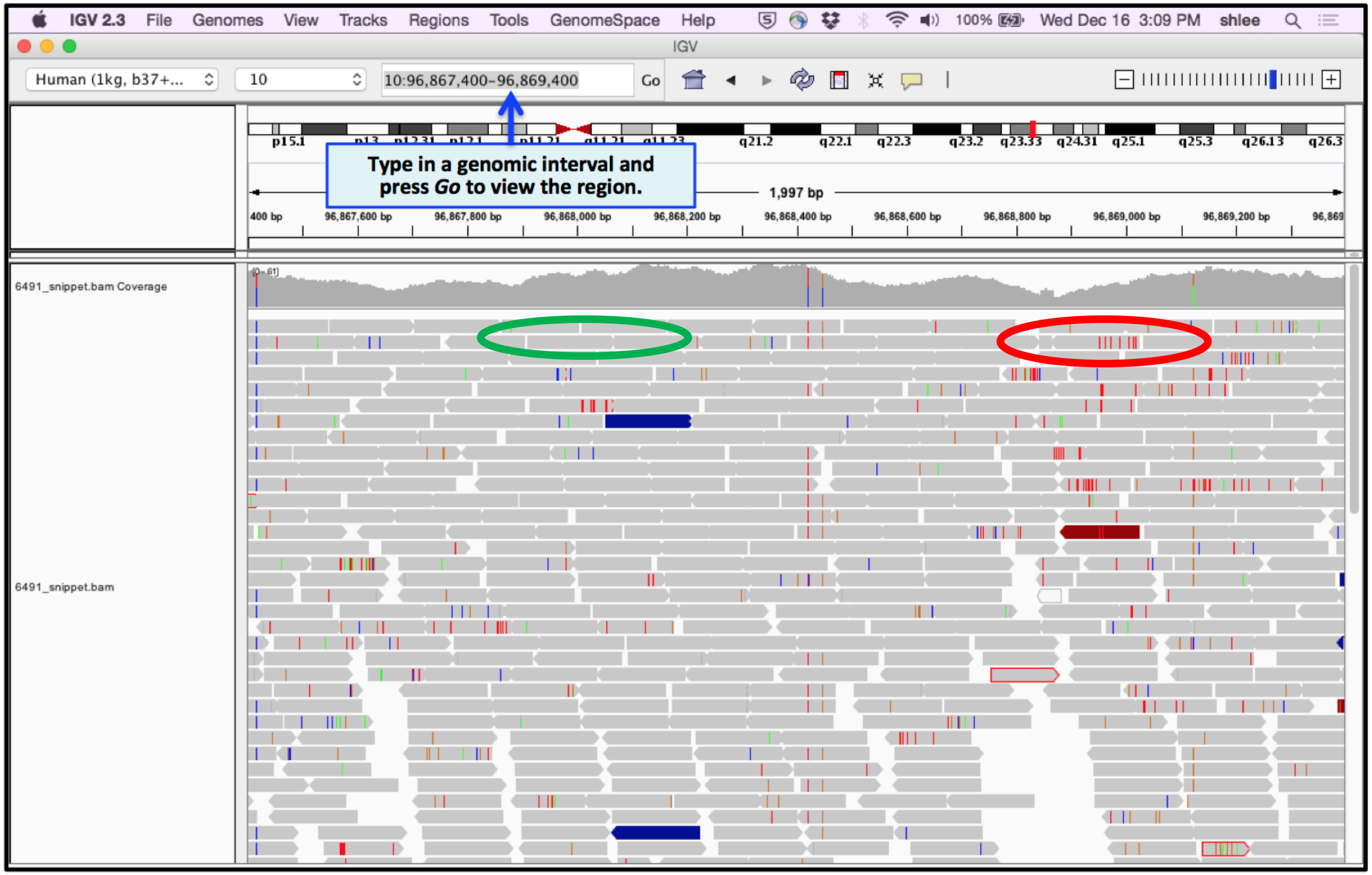
Where to put the read TAGTAGTAGT ?



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



Type in a genomic interval and press *Go* to view the region.

Plan for lecture

- The sequencing revolution
- Technical aspect of sequencing studies
  - Coverage
  - Exomes versus genomes
  - Alignment
  - Variant calling
  - Quality control
  - Contamination
- Variant consequences and annotation
- Interpretation of *de novo* mutations
- Importance of well-matched controls

Variant calling

- The process of ascertaining variants (SNPs, indels, copy number variants, structural variants) in the mapped sequencing reads, and genotyping individuals at those variants

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)

Chromosome	Position	SNP ID	Reference Allele	Alternate Allele	Variant quality Score	Filter
#CHROM	POS	ID	REF	ALT	QUAL	FILTER
chr8	1952745	rs2272608	C	T	771045	PASS
chr8	3219437	rs28455997	T	C	153017	PASS

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

INFO
AC=1;AF=0.125;AN=6;BaseQRankSum=0.124;ClippingRankSum=0;DP=200767;ExcessHet=0.0003; FS=1.214;InbreedingCoeff=0.0426;MLEAC=2036;MLEAF=0.125;MQ=60;MQRankSum=0; QD=16.95;ReadPosRankSum=0.048;SOR=0.837
AC=2;AF=0.078;AN=6;BaseQRankSum=0;ClippingRankSum=0;DP=53124;ExcessHet=0;FS=0; InbreedingCoeff=0.0555;MLEAC=1306;MLEAF=0.081;MQ=59.69;MQRankSum=0;QD=18.37; ReadPosRankSum=0;SOR=0.667

INFO field contains meta-data about the variant

AC, AF, AN = allele count [of the ALT allele], allele frequency, allele number

DP: Approximate read depth across all individuals (N.B. in this case, there were ~8000 individuals in the original VCF)

More on the other variant-level quality metrics in the next few slides

Variant Call Format (VCF)

Chromosome	Position	SNP ID	Reference Allele	Alternate Allele	Variant quality Score	Filter
#CHROM	POS	ID	REF	ALT	QUAL	FILTER
chr8	1952745	rs2272608	C	T	771045	PASS
chr8	3219437	rs28455997	T	C	153017	PASS

INFO
AC=1;AF=0.125;AN=6;BaseQRankSum=0.124;ClippingRankSum=0;DP=200767;ExcessHet=0.0003; FS=1.214;InbreedingCoeff=0.0426;MLEAC=2036;MLEAF=0.125;MQ=60;MQRankSum=0;QD=16.95;ReadPosRankSum=0.048;SOR=0.837
AC=2;AF=0.078;AN=6;BaseQRankSum=0;ClippingRankSum=0;DP=53124;ExcessHet=0;FS=0;InbreedingCoeff=0.0555;MLEAC=1306;MLEAF=0.081;MQ=59.69;MQRankSum=0;QD=18.37;ReadPosRankSum=0;SOR=0.667

FORMAT	person1	person2	person3
GT:AD:DP:GQ:PL	0/0:27,0:27:81:0,81,1070	0/1:17,14:31:99:449,0,613	0/0:31,0:31:87:0,87,1305
GT:AD:DP:GQ:PL	0/0:11,0:11:21:0,21,315	0/1:2,2:4:71:71,0,71	0/1:2,7:9:52:187,0,52

FORMAT field indicates the structure of the GENOTYPE fields

GT: genotype (0/0, 0/1, 1/1); AD: allele depth (ref, alt), DP (depth)

PL: normalized, phred-scaled likelihoods for genotypes; GQ: genotype quality

$$PL = -10 * \log P(Genotype|Data)$$

Multiallelic variants

- Multiple alternate alleles are possible at the same site

#CHROM	POS	ID	REF	ALT	QUAL	FILTER
chr1	236739260	.	C	G,T	4855970	PASS

INFO

AC=1,1;AF=0.084,0.459;AN=6;BaseQRankSum=-0.428;ClippingRankSum=0;DP=272799;ExcessHet=0;FS=0;InbreedingCoeff=0.0499;MLEAC=1368,7505;MLEAF=0.084,0.46;MQ=60.06;MQRankSum=0;QD=23.01;ReadPosRankSum=0.114;SOR=1.078

FORMAT	person1
GT:AD:DP:GQ:PL	0/0:38,0,0:38:99:0,99,1374,99,1374,1374

person2	person3
0/**2**:20,0,11:31:99:345,404,1078,0,674,641	0/**1**:27,22,0:49:99:668,0,804,747,869,1616

Plan for lecture

- The sequencing revolution
- Technical aspect of sequencing studies
  - Coverage
  - Exomes versus genomes
  - Alignment
  - Variant calling
  - Quality control
  - Contamination
- Variant consequences and annotation
- Interpretation of *de novo* mutations
- Importance of well-matched controls

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

Different levels of QC

- Sample-level (e.g. number of heterozygous and non-reference homozygous calls, missingness, contamination, number of singletons)
- Variant-level (e.g. mapping quality, strand bias, overall depth, Hardy-Weinberg)
- Genotype-level (e.g. genotype quality, depth, allele balance)

What filters do we use?

- Problem: correlated sequencing errors and mapping artefacts drive false positives (cause loss of power, spurious conclusions)
- The following should be random if the sequencing technology is working as expected:
  - Strand bias – 5’-to-3’ and 3’-to-5’ reads should give equal representation of alternate allele
  - Base quality – ALT and REF base calls should not differ systematically in quality
  - Variant position in read
  - Allele bias – at heterozygous sites, the number of ALT reads should follow a binomial distribution with p=0.5 (genotype level)

Variant Call Format (VCF)

Chromosome

Position

SNP ID

Reference Allele

Alternate Allele

Variant quality Score

Filter

#CHROM	POS	ID	REF	ALT	QUAL	FILTER
chr8	1952745	rs2272608	C	T	771045	PASS
chr8	3219437	rs28455997	T	C	153017	PASS

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

INFO
AC=1;AF=0.125;AN=6;BaseQRankSum=0.124;ClippingRankSum=0;DP=200767;ExcessHet=0.0003;FS=1.214;InbreedingCoeff=0.0426;MLEAC=2036;MLEAF=0.125;MQ=60;MQRankSum=0;QD=16.95;ReadPosRankSum=0.048;SOR=0.837
AC=2;AF=0.078;AN=6;BaseQRankSum=0;ClippingRankSum=0;DP=53124;ExcessHet=0;FS=0;InbreedingCoeff=0.0555;MLEAC=1306;MLEAF=0.081;MQ=59.69;MQRankSum=0;QD=18.37;ReadPosRankSum=0;SOR=0.667

INFO field contains meta-data about the variant

AC, AF, AN = allele count, allele frequency, allele number

DP: Approximate read depth across all individuals (N.B. in this case, there were ~8000 individuals in the original VCF)

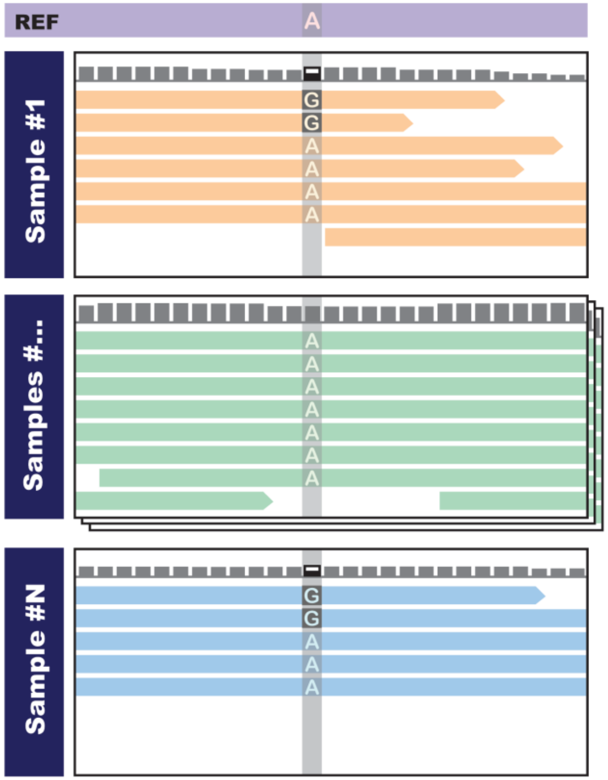
FS: Phred-scaled p-value using Fisher's exact test to detect strand bias

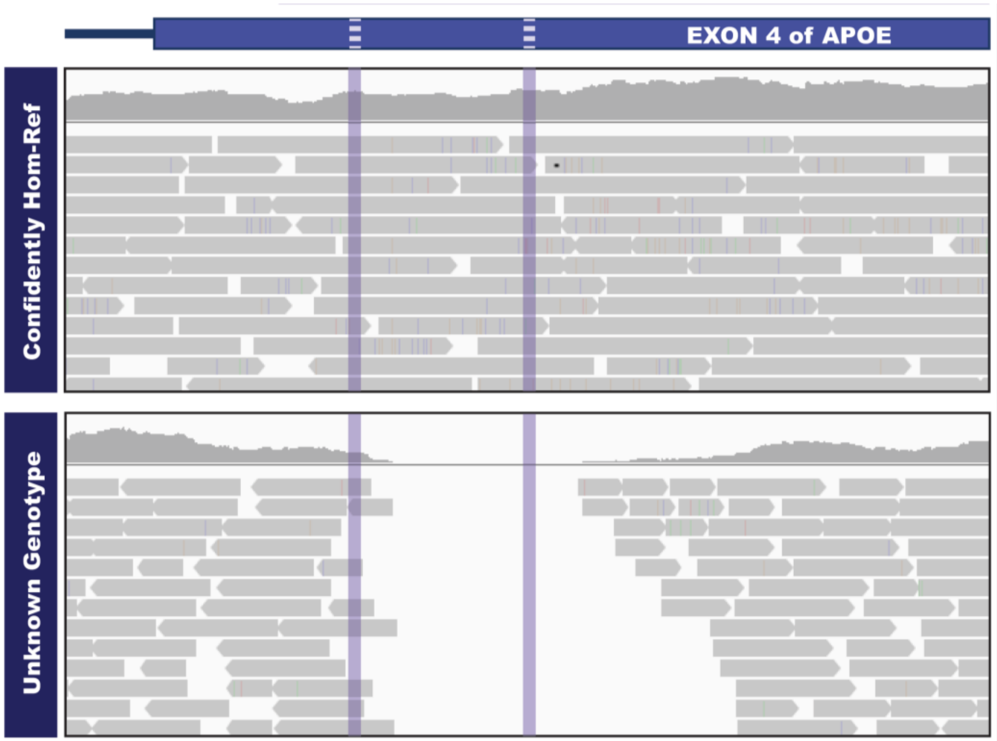
BaseQRankSum: Z-score from Wilcoxon rank sum test of Alt Vs. Ref base qualities

ReadPosRankSum: Z-score from Wilcoxon rank sum test of Alt vs. Ref read position bias

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
- Distinguishing missing genotype from homozygous reference





Ben Neale

Variant filtration strategies are still evolving
VQSR is one approach

- variant quality score recalibration (VQSR) aims to enable variant filtering in order to balance sensitivity and specificity
- 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

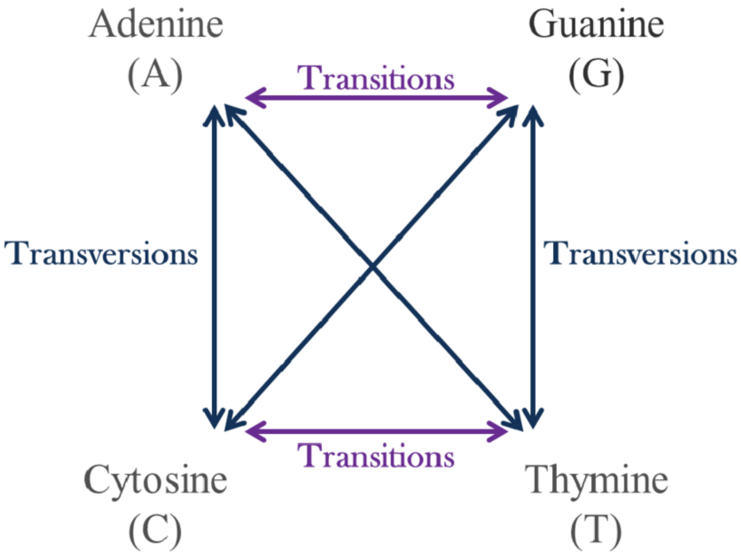
http://gatkforums.broadinstitute.org/gatk/discussion/39/variant-quality-score-recalibration-vqsr

An important variant-level QC metric

Transition:transversion ratio across the dataset

purines

pyrimidines



Transitions (Ti) within purines/pyrimidines

vs transversions (Tv) between them

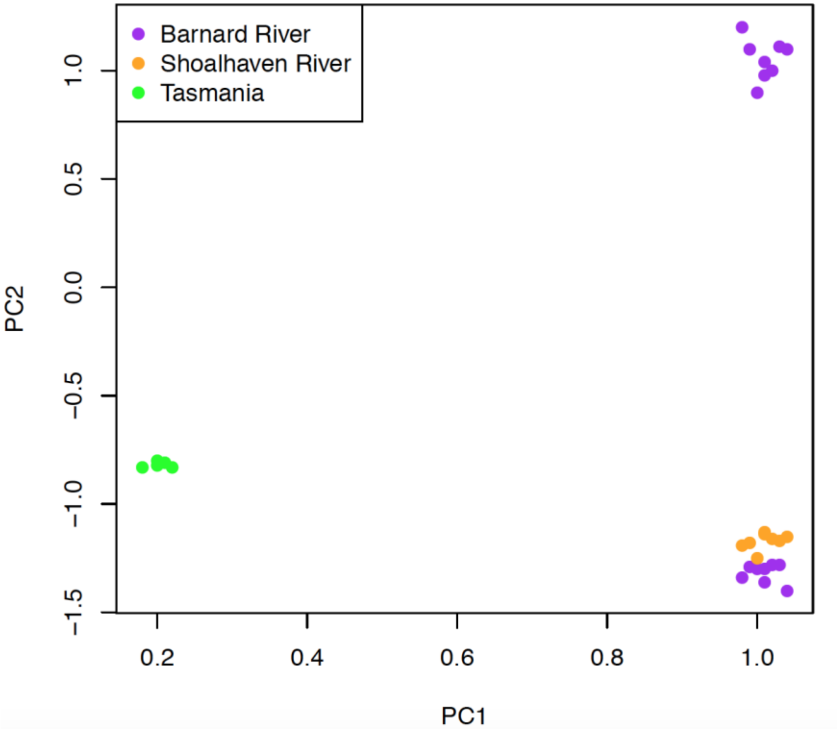
- transitions are expected to occur twice as frequently as transversions
- Ti:Tv is typically ~2 across the whole genome, versus ~3 in protein coding regions
- not relevant for genotype data since we know the variants are real
- most useful at the individual level, as it changes with sample size (larger sample sizes → more recurrent C>T mutations)

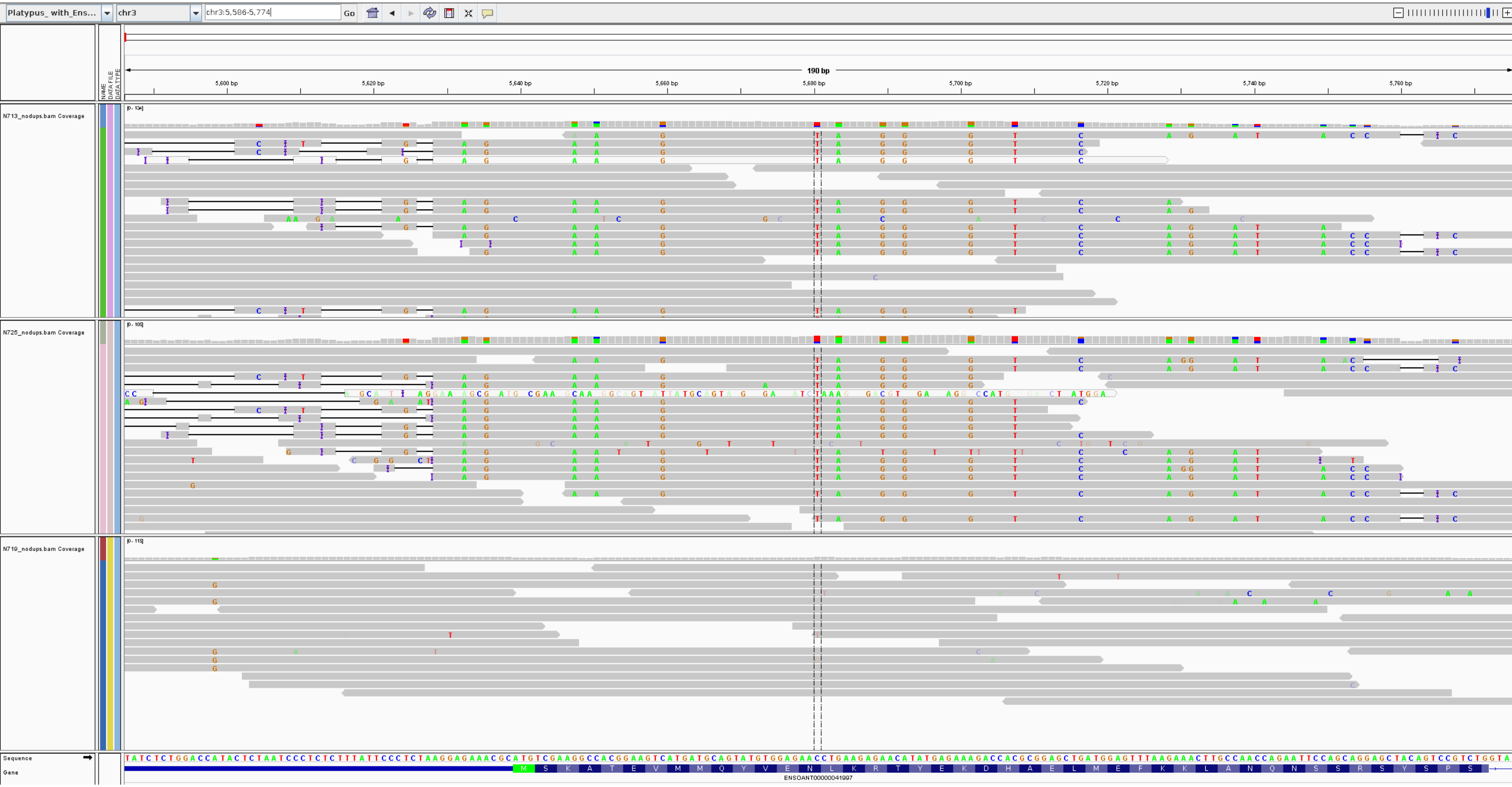
Plan for lecture

- The sequencing revolution
- Technical aspect of sequencing studies
  - Coverage
  - Exomes versus genomes
  - Alignment
  - Variant calling
  - Quality control
  - Contamination
- Variant consequences and annotation
- Interpretation of *de novo* mutations
- Importance of well-matched controls

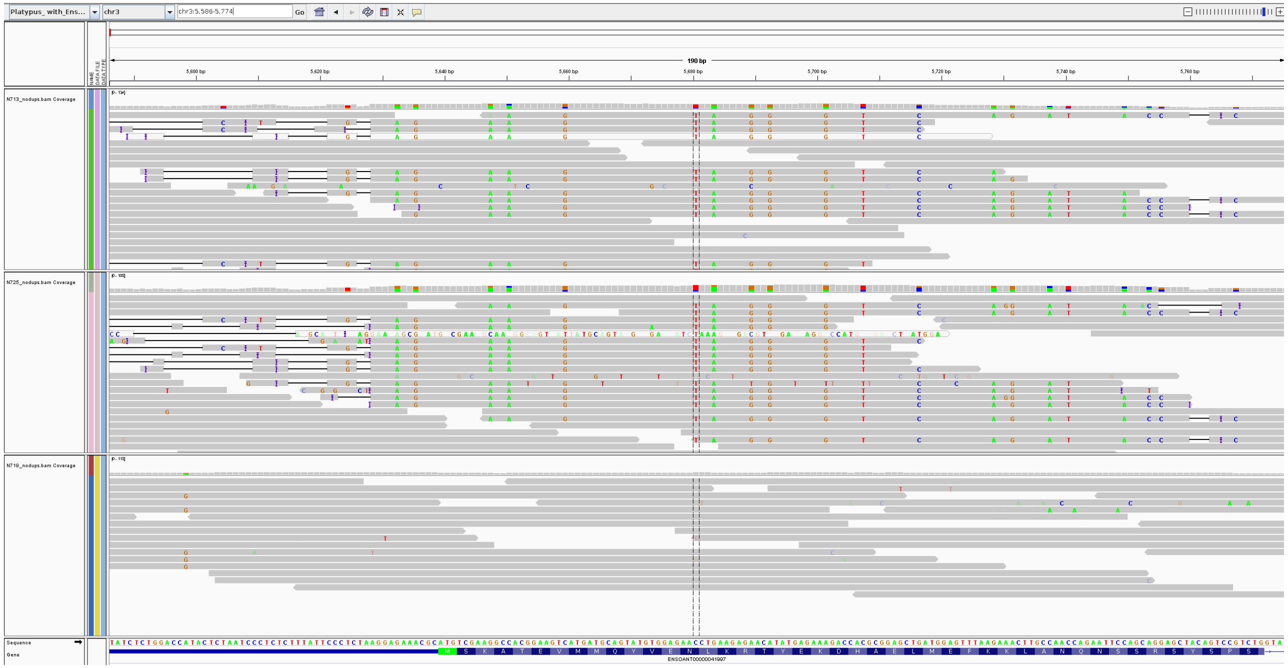
A cautionary tale: another peril of sequence data

- 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
NAME
DATA FILE
DATA TYPE
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
[0 - 134]
N725_nodups.bam Coverage
[0 - 108]
N719_nodups.bam Coverage
[0 - 113]
Sequence
Gene
TATCTCTGGACCATACTCTAATCCCTCTCTTTATTCCCTCTAAGGAGAAACGCATGTCGAAGGCCACGGAAGTCATGATGCAGTATGTGGAGAACCTGAAGAGAACATATGAGAAAGACCACGCGGAGCTGATGGAGTTTAAGAAACTTGCCAACCAGAATTCCAGCAGGAGCTACAGTCCGTCTGGTA
M S K A T E V M M Q Y V E N L K R T Y E K D H A E L M E F K K L A N Q N S S R S Y S P S
ENSOANT00000041997

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



- 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 multiplexed in the same sequencing lane (‘index hopping’)
- people who have just eaten ham for lunch before spitting
- bacterial/viral contamination
- Rarer problems:
    - saliva samples from kids that contains parental saliva
    - people who have had bone marrow transplants

Summary: QC for sequencing versus genotype data

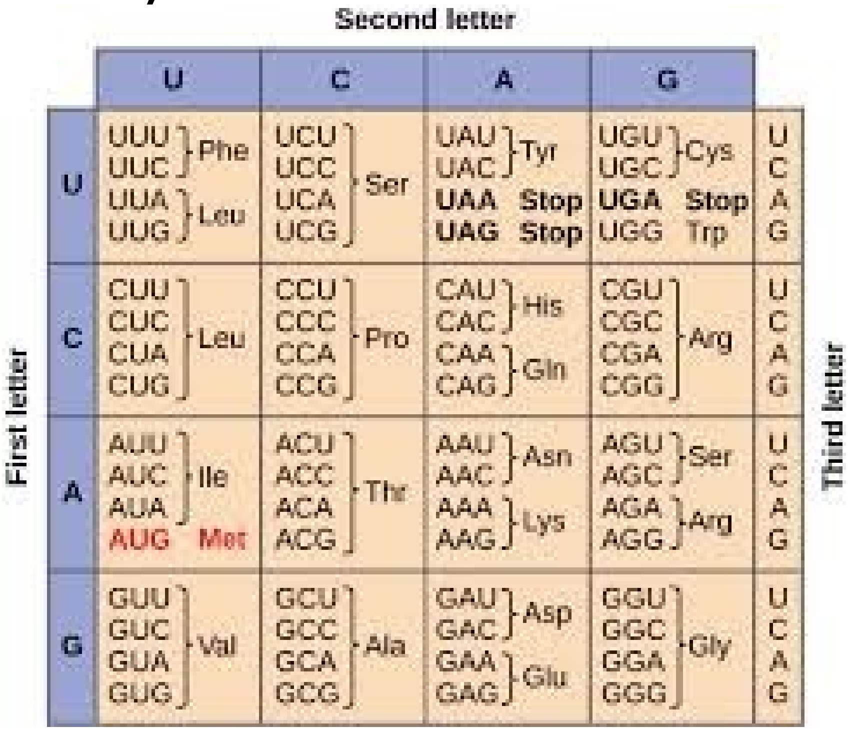
- in sequence data, there is a discovery problem as well as a genotyping problem (i.e. the variants may not be real variants at all) – **need to filter sites as well as genotypes**
- contamination is more of a problem for sequencing than genotyping data
- error modes greatly differ between sequencing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

Plan for lecture

- The sequencing revolution
- Technical aspect of sequencing studies
  - Coverage
  - Exomes versus genomes
  - Alignment
  - Variant calling
  - Quality control
  - Contamination
- Variant consequences and annotation
- Interpretation of *de novo* mutations
- Importance of well-matched controls

Coding variant consequences

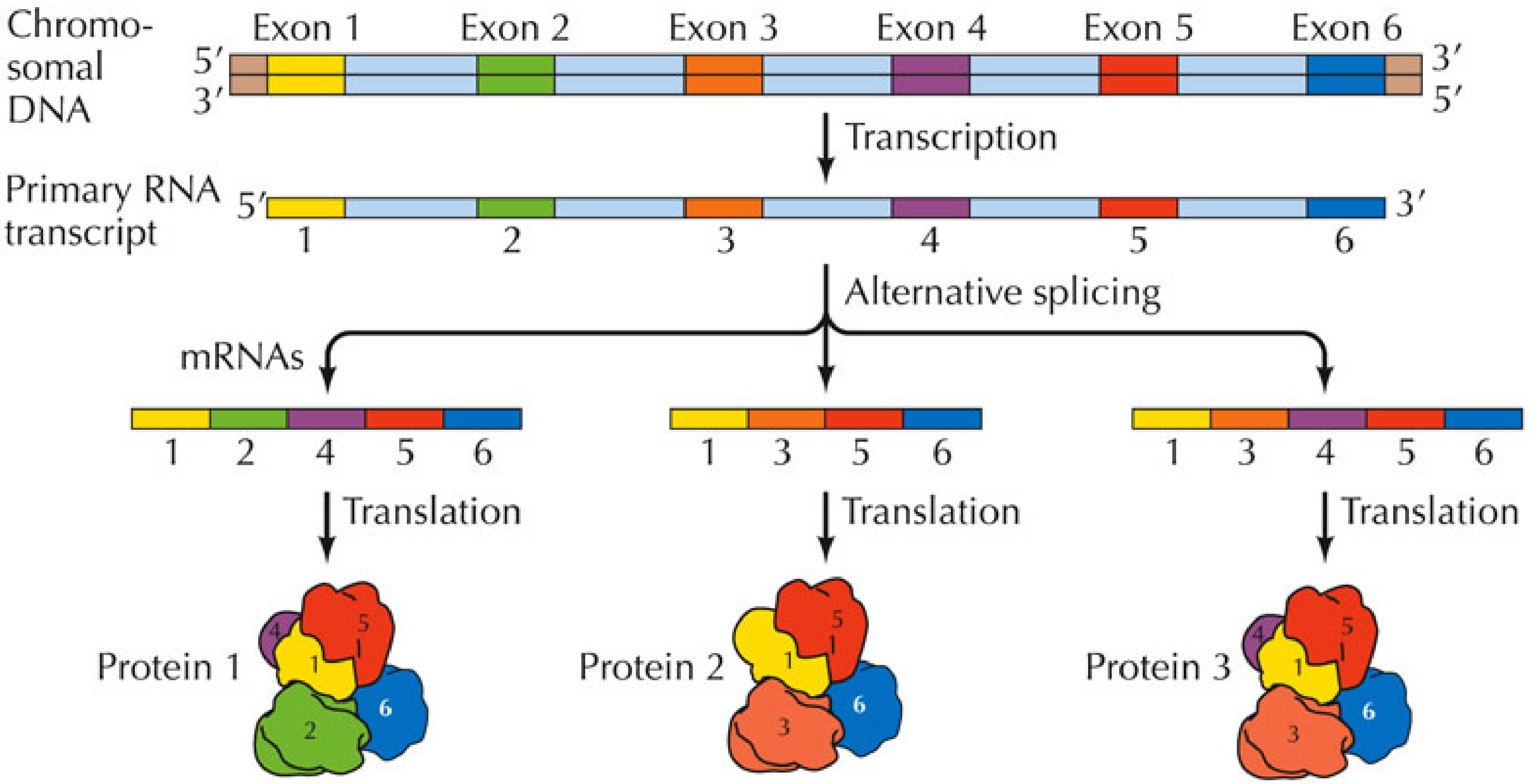
- Synonymous – same amino acid
- Missense – different amino acid
- Nonsense (loss-of-function) – premature stop codon
- Splicing mutation - disrupts splicing (often leading to loss-of-function)



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

Ben Neale

Alternative splicing

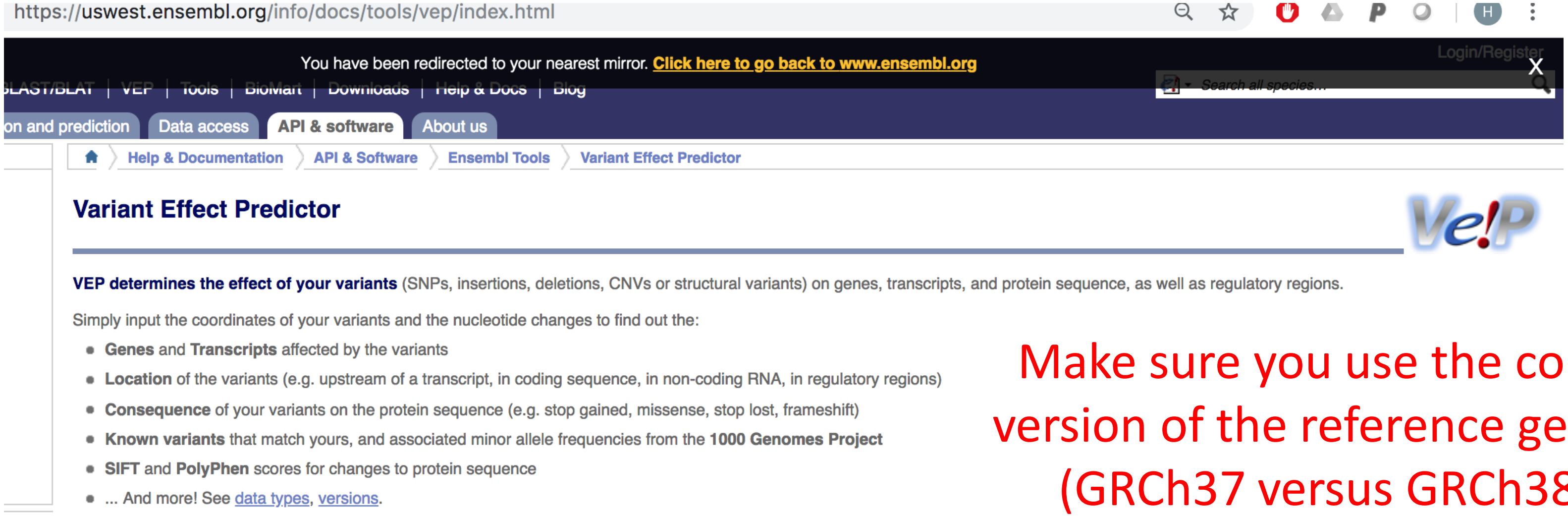


THE CELL, Fourth Edition, Figure 5.5 © 2006 ASM Press and Sinauer Associates, Inc.

Annotation

- process of adding information about frequency, expected functional consequence etc. of variants
  - is the variant found in dbSNP? 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 but not all transcripts)

Variant Effect Predictor



https://uswest.ensembl.org/info/docs/tools/vep/index.html

You have been redirected to your nearest mirror. Click here to go back to www.ensembl.org

Login/Register

LAST/BLAT | VEP | Tools | BioMart | Downloads | Help & Docs | Blog

Search all species...

on and prediction | Data access | API & software | About us

Help & Documentation 〉 API & Software 〉 Ensembl Tools 〉 Variant Effect Predictor

Variant Effect Predictor

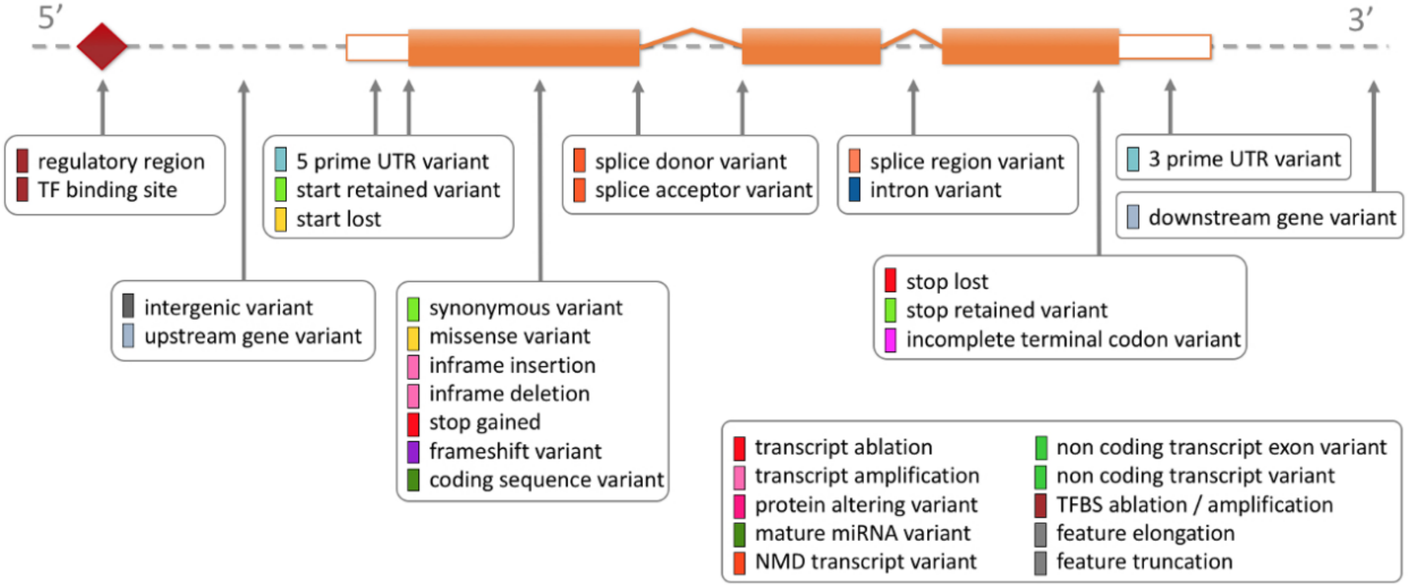
VE!P

VEP determines the effect of your variants (SNPs, insertions, deletions, CNVs or structural variants) on genes, transcripts, and protein sequence, as well as regulatory regions.

Simply input the coordinates of your variants and the nucleotide changes to find out the:

- **Genes** and **Transcripts** affected by the variants
- **Location** of the variants (e.g. upstream of a transcript, in coding sequence, in non-coding RNA, in regulatory regions)
- **Consequence** of your variants on the protein sequence (e.g. stop gained, missense, stop lost, frameshift)
- **Known variants** that match yours, and associated minor allele frequencies from the **1000 Genomes Project**
- **SIFT** and **PolyPhen** scores for changes to protein sequence
- ... And more! See data types, versions.

Make sure you use the correct version of the reference genome (GRCh37 versus GRCh38)!)



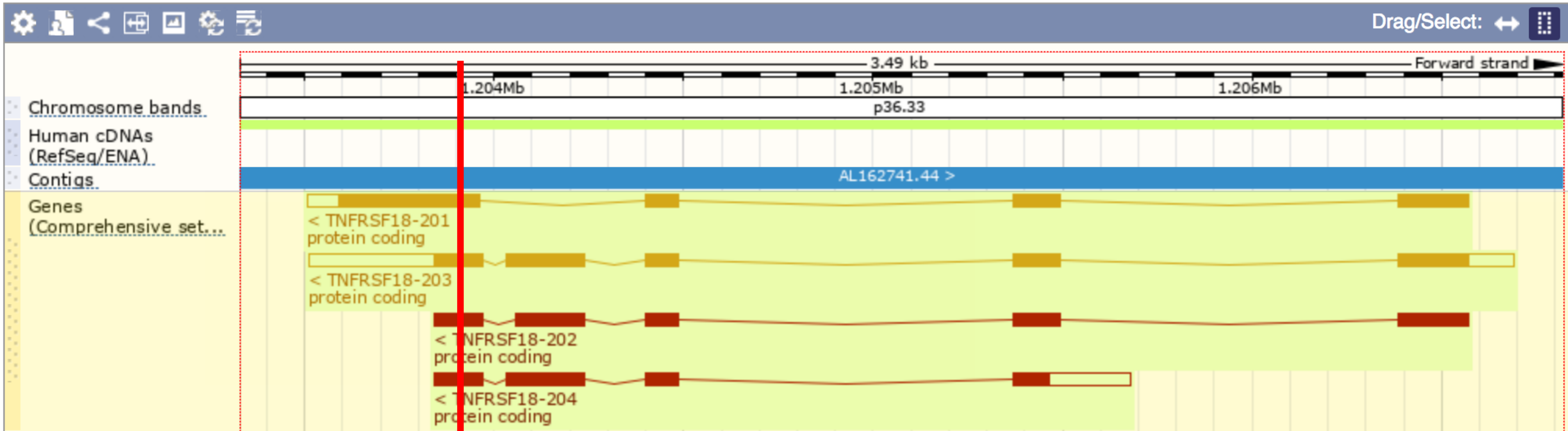
https://uswest.ensembl.org/info/genome/variation/prediction/predicted_data.html

Variant annotation is specific to the alternate allele and the transcript

CHROM	POS	ID	REF	ALT
chr1	1203891	.	C	A,T

SYMBOL	Gene
TNFRSF18	ENSG00000186891

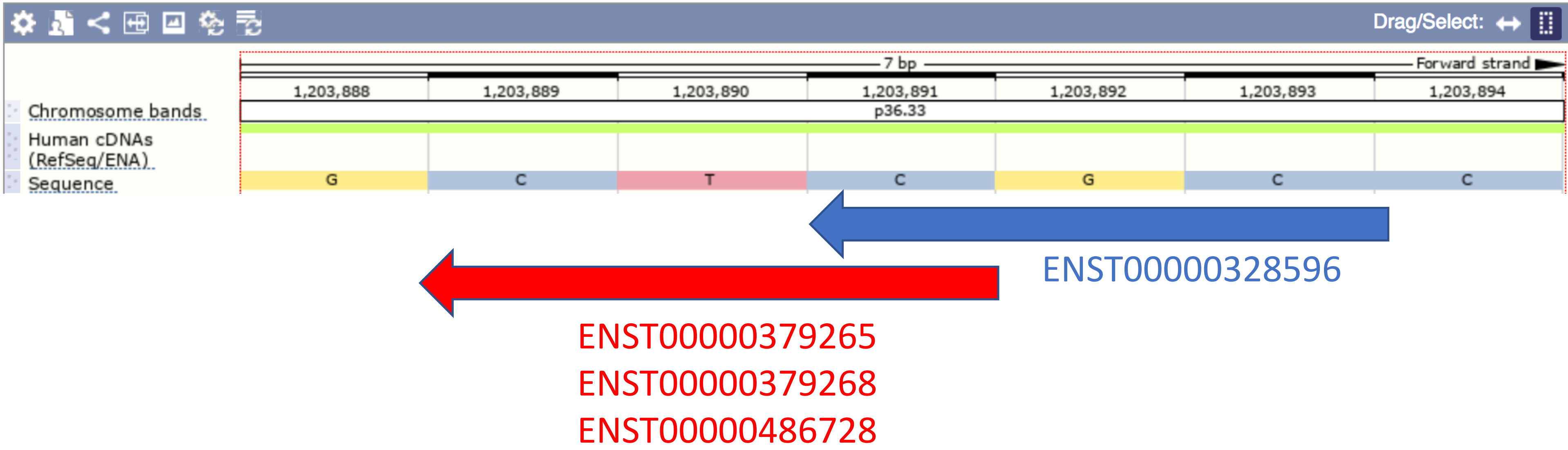
Location	Allele	Consequence	IMPACT	Feature	EXON	Codons
1:1203891-1203891	A	synonymous_variant	LOW	ENST00000328596	4/4	gcG/gcT
1:1203891-1203891	T	synonymous_variant	LOW	ENST00000328596	4/4	gcG/gcA
1:1203891-1203891	A	stop_gained	HIGH	ENST00000379265	5/5	Gag/Tag
1:1203891-1203891	T	missense_variant	MODERATE	ENST00000379265	5/5	Gag/Aag
1:1203891-1203891	A	stop_gained	HIGH	ENST00000379268	5/5	Gag/Tag
1:1203891-1203891	T	missense_variant	MODERATE	ENST00000379268	5/5	Gag/Aag
1:1203891-1203891	A	stop_gained	HIGH	ENST00000486728	4/4	Gag/Tag
1:1203891-1203891	T	missense_variant	MODERATE	ENST00000486728	4/4	Gag/Aag


Drag/Select:
3.49 kb
Forward strand
1.204Mb
1.205Mb
1.206Mb
Chromosome bands
p36.33
Human cDNAs (RefSeq/ENA)
Contigs
AL162741.44 >
Genes (Comprehensive set...
< TNFRSF18-201 protein coding
< TNFRSF18-203 protein coding
< TNFRSF18-202 protein coding
< TNFRSF18-204 protein coding

Variant annotation is specific to the alternate allele and the transcript

CHROM	POS	ID	REF	ALT	SYMBOL	Gene
chr1	1203891	.	C	A,T	TNFRSF18	ENSG00000186891

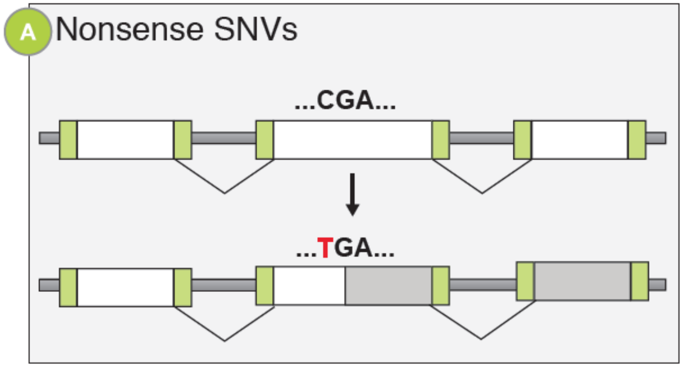
Location	Allele	Consequence	IMPACT	Feature	EXON	Codons
1:1203891-1203891	A	synonymous_variant	LOW	ENST00000328596	4/4	gcG/gcT
1:1203891-1203891	T	synonymous_variant	LOW	ENST00000328596	4/4	gcG/gcA
1:1203891-1203891	A	stop_gained	HIGH	ENST00000379265	5/5	Gag/Tag
1:1203891-1203891	T	missense_variant	MODERATE	ENST00000379265	5/5	Gag/Aag
1:1203891-1203891	A	stop_gained	HIGH	ENST00000379268	5/5	Gag/Tag
1:1203891-1203891	T	missense_variant	MODERATE	ENST00000379268	5/5	Gag/Aag
1:1203891-1203891	A	stop_gained	HIGH	ENST00000486728	4/4	Gag/Tag
1:1203891-1203891	T	missense_variant	MODERATE	ENST00000486728	4/4	Gag/Aag

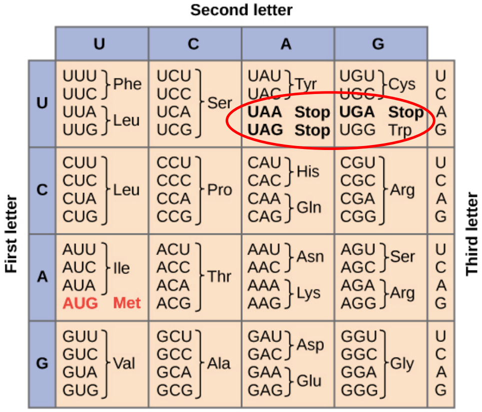


Loss-of-function variants are often of particular interest

- LoFs are variants that severely affect the function of a protein-coding gene
- typically do so by deleting it or prompting nonsense-mediated decay (degradation of mRNA molecules with premature stop codons – protects cells against aberrant proteins that may be deleterious)
- LoFs also called protein truncating variants (PTVs)
- tend to be more deleterious than other types of variants

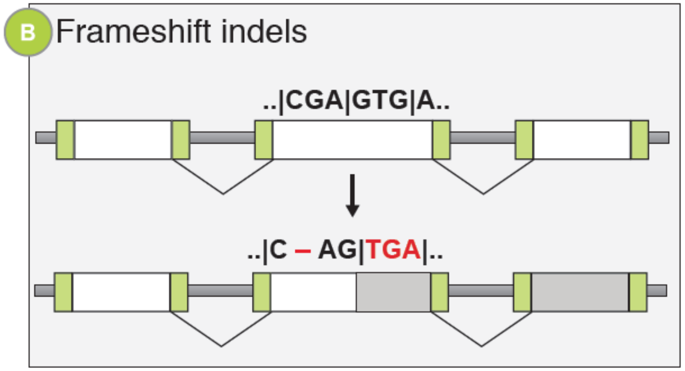
Different types of LoFs



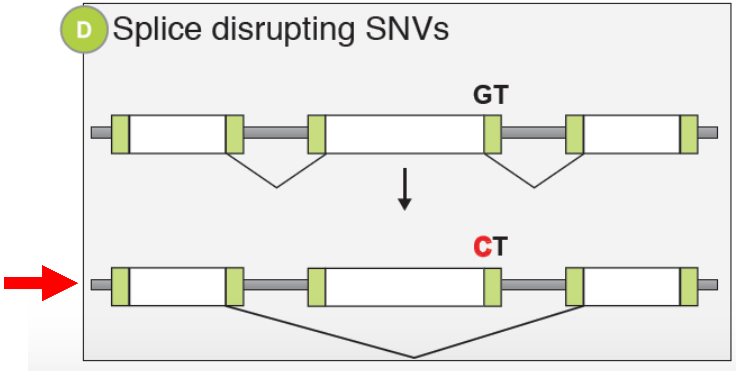


Second letter

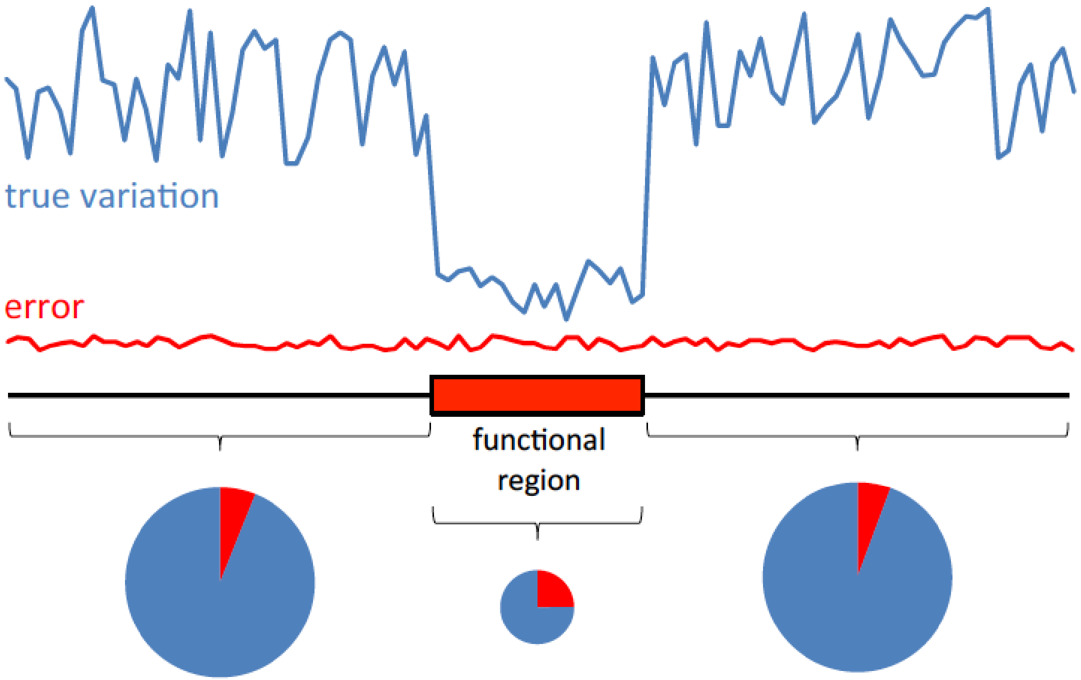
First letter	U	C	A	G	Third letter
U	UUU UUC Phe UUA UUG Leu	UCU UCC UCA UCG Ser	UAU UAC Tyr **UAA Stop** **UAG Stop**	UGU UGC Cys **UGA Stop** UGG Trp	U C A G
C	CUU CUC CUA CUG Leu	CCU CCC CCA CCG Pro	CAU CAC His CAA CAG Gln	CGU CGC CGA CGG Arg	U C A G
A	AUU AUC AUA Ile AUG Met	ACU ACC ACA ACG Thr	AAU 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



Breaks the GT-AG rule



Challenges to identifying true LoFs



- the fraction of variants that are sequencing/calling errors is higher for LoFs than other types of variants
- calling indels and large copy number variants from sequence data is particularly difficult, and they are enriched for LoFs
- validation of variants (usually via Sanger sequencing) is necessary for some applications
- LOFTEE can be used (as a plugin to VEP) to filter out spurious LoFs based on gene/transcript annotation features/errors

Daniel MacArthur

Plan for lecture

- The sequencing revolution
- Technical aspect of sequencing studies
  - Coverage
  - Exomes versus genomes
  - Alignment
  - Variant calling
  - Quality control
  - Contamination
- Variant consequences and annotation
- Interpretation of *de novo* mutations
- Importance of well-matched controls

Why study *de novo* mutations?

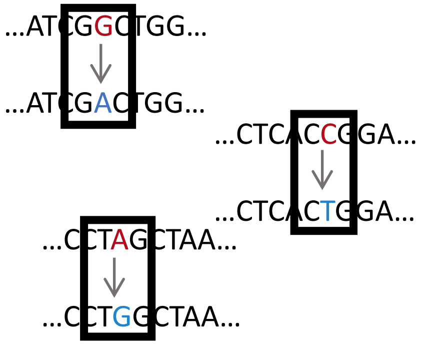
- mutations that occurred in the egg or sperm (or one of their precursor cells) and are hence are not present in all the cells in a parent's body
- the most damaging mutations are likely to be *de novo* – they have not yet been subject to negative selection
- abundant evidence for a large role of *de novo* mutations in severe, early-onset diseases (e.g. developmental disorders)
- some contribution to later onset diseases e.g. schizophrenia, but likely to account for few cases

Interpretation of *de novo* mutations

- multiple *de novo* mutations in a gene in a cohort of disease cases are often used as evidence for that gene's role in disease.
- as we sequence large numbers of individuals, we can easily see recurrent mutations in a particular gene just by chance
- need to understand the expectation for *de novo* variation so we can establish a statistical framework with which to evaluate the results of exome/genome sequencing studies

Slide from Kaitlin Samocha

Creating a model of, and statistical framework for, evaluating *de novo* variation

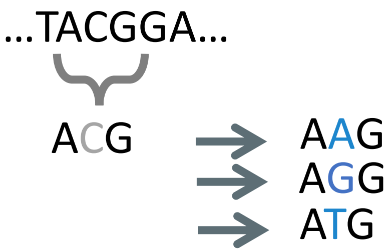


Change	Probability
AAA → ACA	*a*
AAA → AGA	*b*
AAA → ATA	*c*
AAC → ACC	*d*
AAC → AGC	*e*
AAC → ATC	*f*
…	

$$\Pr(AAA \rightarrow ATC) = \lambda \frac{\# AAA > ATC \ variants \ in \ 1000G}{\# AAA \ ancestral \ trucleotides}$$

Created a mutation rate table:
4^3 x 3 = 192 possible mutations

Used the sequence to determine each gene's probability of mutating



Per gene:

Pr(synonymous)
Pr(missense)
Pr(nonsense)
Pr(splice site)

Also corrected for sequencing depth

Slide from Kaitlin Samocha

Samocha *et al.*, **Nat Genet**, 2014

Per-gene probabilities of mutation are small, but consider the number of “candidate” genes and number of samples

Example probabilities of mutation per gene, per trio:

	class	rate
	synonymous	9.88E-6
	missense	2.36E-5
Loss-of-function (LoF)	nonsense	1.14E-6
	Splice site	6.82E-7
	frameshift	1.30E-6

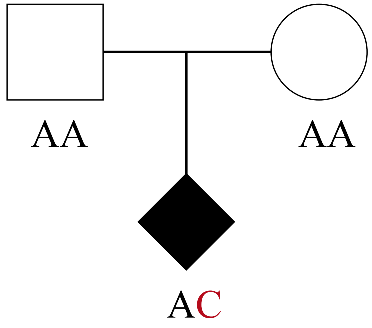
Probability of seeing >1 *de novo* in the same gene is quite high once you have a few hundred samples

sample size	Probability of >1 de novo		
	missense	LoF	either
100	0.053	0.001	0.068
200	0.208	0.004	0.268
300	0.465	0.009	0.597

Probability of *de novo* LoF or missense in a gene expressed in fetal brain = 0.23

Do we see more deleterious *de novo* variants in cases than expected?

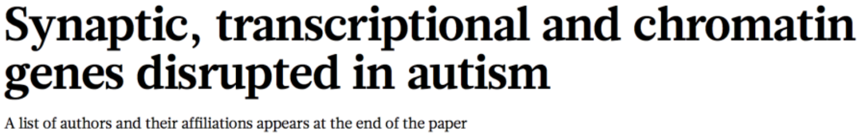
Application to *de novo* variation found in cases with autism spectrum disorders (ASD)



3,982 cases with ASD

2,078 unaffected siblings

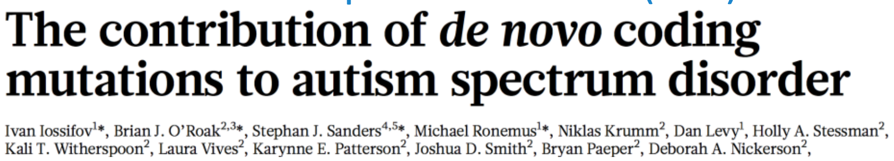
Autism Sequencing Consortium (ASC)



Synaptic, transcriptional and chromatin genes disrupted in autism

A list of authors and their affiliations appears at the end of the paper

Simons Simplex Collection (SSC)



The contribution of *de novo* coding mutations to autism spectrum disorder

Ivan Iossifov[1]*, Brian J. O'Roak[2,3]*, Stephan J. Sanders[4,5]*, Michael Ronemus[1]*, Niklas Krumm[2], Dan Levy[1], Holly A. Stessman[2], Kali T. Witherspoon[2], Laura Vives[2], Karynne E. Patterson[2], Joshua D. Smith[2], Bryan Paeper[2], Deborah A. Nickerson[2],

Slide from Kaitlin Samocha

Genome-wide excess of both missense and loss-of-function (LoF) *de novo* variants in ASD cases

Sample set	N	Consequence	Observed	Expected	one-sided Poisson p-value
affected siblings	3982	synonymous	1048	1092.66	0.91
		missense	2814	2470.03	$7\text{x}10^{-12}$
		LoF	579	341.26	$9\text{x}10^{-32}$
unaffected siblings	2078	synonymous	532	570.20	0.95
		missense	1258	1288.98	0.8
		LoF	190	178.08	0.2

X~Poisson(λ=Expected)

One-sided Poisson test:

$$\Pr(X \geq Observed) = 1 - \Pr(X < Observed) \; = 1 - \sum_{x=0}^{Observed-1} \frac{e^{-\lambda}\lambda^{x}}{x!}$$

Genome-wide burden of synonymous: should have observed≈expected

→ can use this metric to set threshold for calling *de novos* accurately

Samocha et al 2014; De Rubeis et al 2014; Iossifov et al 2014

Is there a significant excess of *de novo* variants in a specific gene?

Six genes cross the significance threshold for harboring multiple *de novo* variants in ASD cases

Gene	# LoFs Observed	# LoFs Expected	p-value
CHD8	7	0.0604	5.51E-13
DYRK1A	5	0.0201	2.71E-11
SYNGAP1	5	0.0313	2.46E-10
ADNP	4	0.0176	3.93E-09
ARID1B	5	0.0674	1.10E-08
DSCAM	4	0.0551	3.69E-07
GRIN2B	3	0.0221	1.77E-06
SCN2A	4	0.0825	1.81E-06
SUV420H1	3	0.0236	2.16E-06
ANK2	4	0.1227	8.57E-06
POGZ	3	0.0583	3.16E-05

27 more genes with at least 2 *de novo* LoF variants not shown

Bonferroni correction for multiple testing

$p< 5 \times 10^{-7}$ (0.01/20,000 genes)

Slide from Kaitlin Samocha

Samocha *et al.* 2014; De Rubeis *et al.* 2014; Iossifov *et al.* 2014

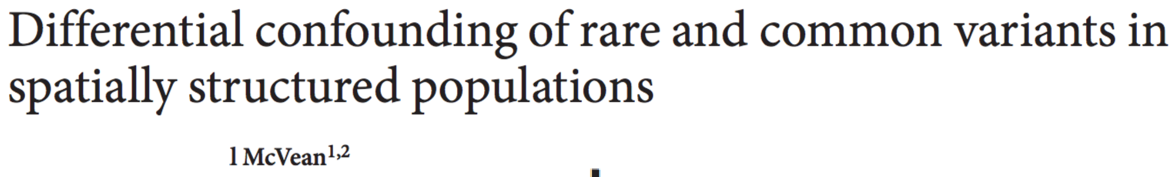
Plan for lecture

- The sequencing revolution
- Technical aspect of sequencing studies
  - Coverage
  - Exomes versus genomes
  - Alignment
  - Variant calling
  - Quality control
  - Contamination
- Variant consequences and annotation
- Interpretation of *de novo* mutations
- Importance of well-matched controls

Case/control studies

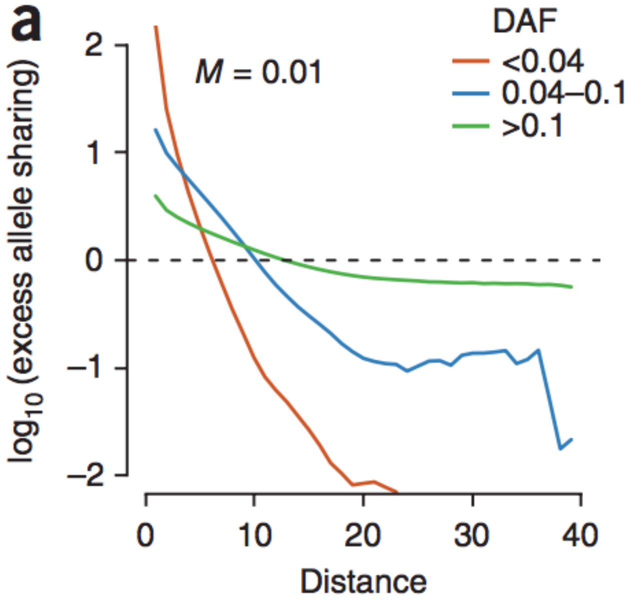
- sequence datasets often used to do per-variant or gene-based burden tests comparing cases and controls
- can’t always afford to sequence both cases and controls, so use publicly available controls → lots of potential artefacts
- as far as possible, we need to harmonise:
    - sequencing (same technology, depth, sequencing centre)
    - read mapping
    - variant calling
- usually interested in rare variants, so having ancestry-matched controls is particularly important, since rare variants tend to be more geographically localized than common variants

Population stratification of rare variants

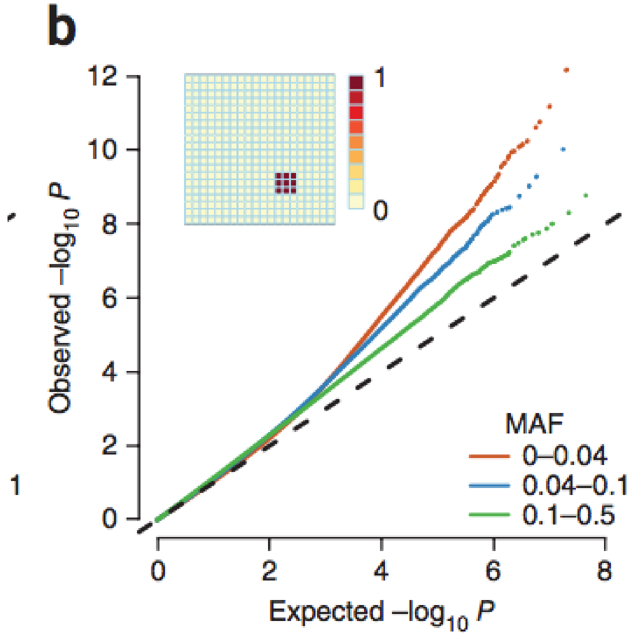


Differential confounding of rare and common variants in spatially structured populations

l 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

1000 Genomes
A Deep Catalog of Human Genetic Variation

2,500 low-coverage whole genomes, various ancestries

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

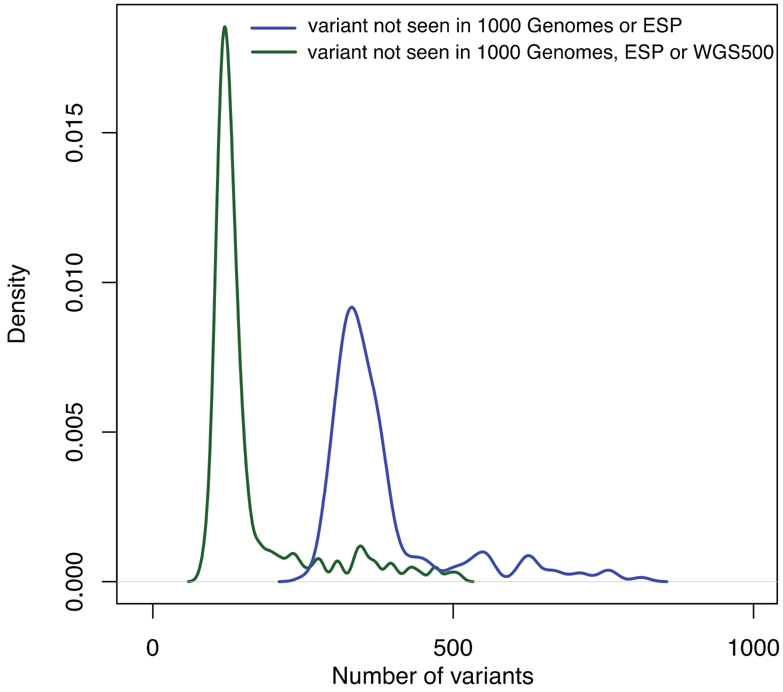
4,000 low-coverage whole genomes (TwinsUK and ALSPAC)
6,000 exomes of people with extreme phenotypes of specific conditions

gnomAD
genome aggregation database

~125k exomes, ~15k genomes, various ancestries, some with complex diseases

Value of in-house controls

- plot shows distribution of number of “novel” heterozygous protein-altering variants per person, across 500 people in a clinical WGS project (WGS500)
- “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)



Limitations in using external sequencing datasets as controls

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

Up next: Konrad Karczewski on gnomAD and constraint