FUNCTIONAL ANNOTATION

Warning

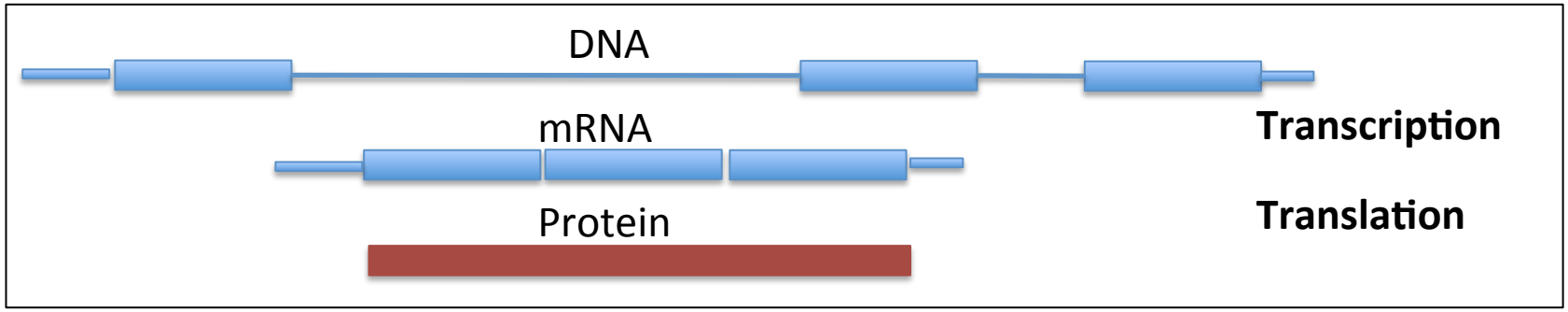
- You will be **rapidly** introduced for many functional annotations in this section

- **Unless you have a good prior background in molecular biology, you will struggle to understand all of these annotations**

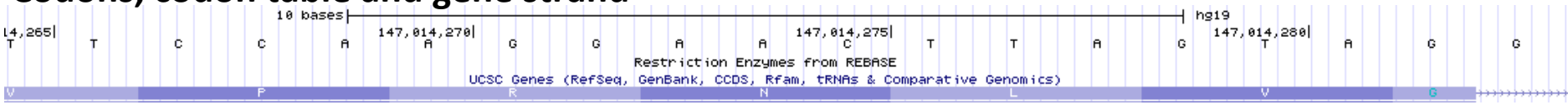
- Keep in mind that the main goal is to use these annotations to distinguish variants with a big phenotypic impact from those with no effect or a very mild effect
  - in the case of monogenic disease, a single variant can be responsible for the entire clinical phenotype.

What is a gene?

Central Dogma

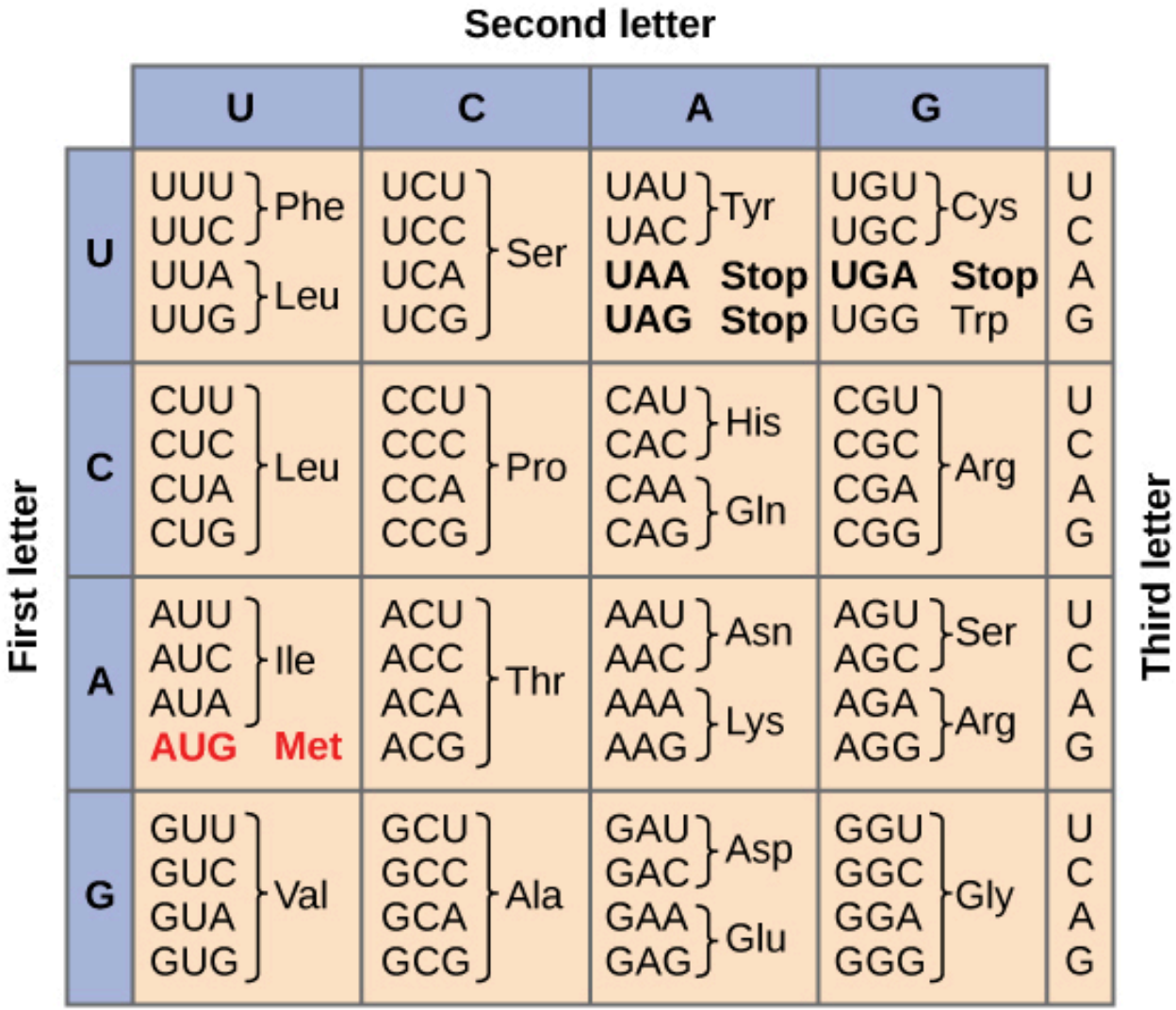


Codons, codon table and gene strand

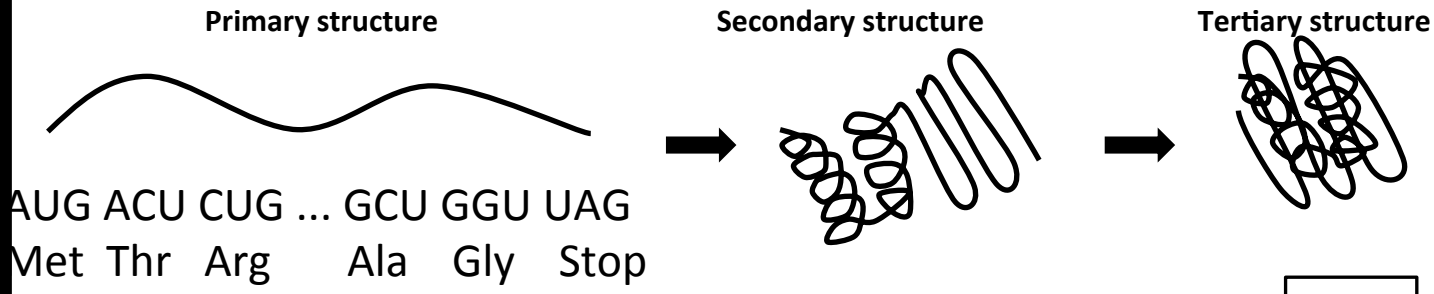


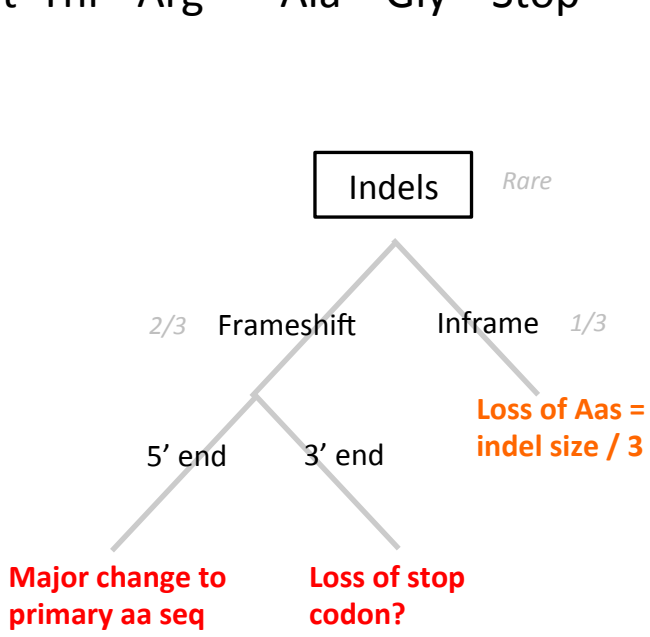
- There are large numbers of genes in multicellular organisms
- Not all organism have multiple exons per gene and thus do not need splicing
- Outside genes are regulatory elements that control when and where genes are expressed

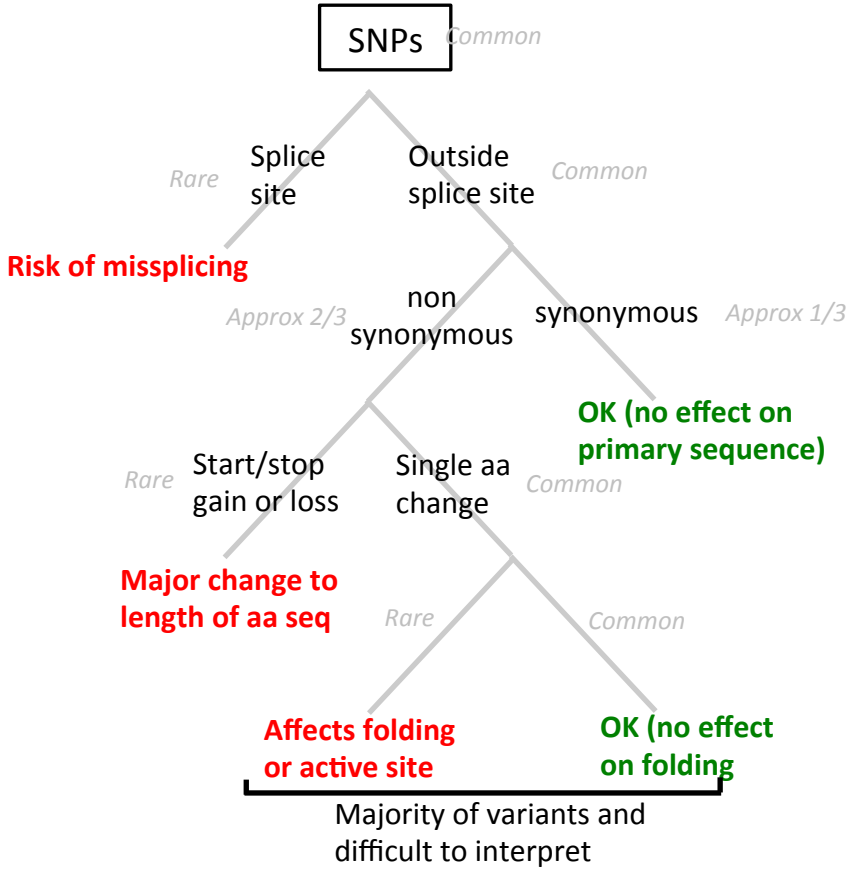
Synonymous and non-synonymous mutations



Effects of variants







http://genetics.bwh.harvard.edu/pph2/dokuwiki/_media/nmeth0410-248.pdf

Polyphen, Sift, etc

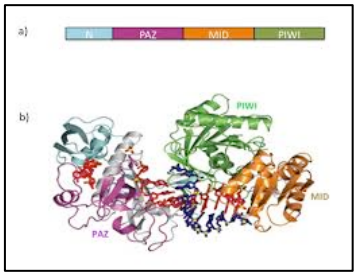
Why are there different isoforms?

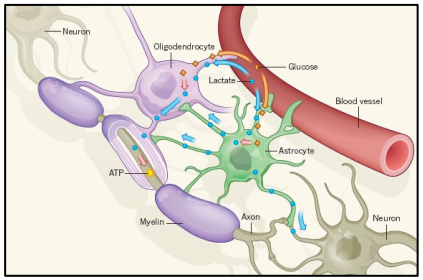
Approx. 25,000 genes in the human genome

DNA >transcription> **RNA** >translation> **PROTEIN**

BUT 1 gene ≠ 1 protein

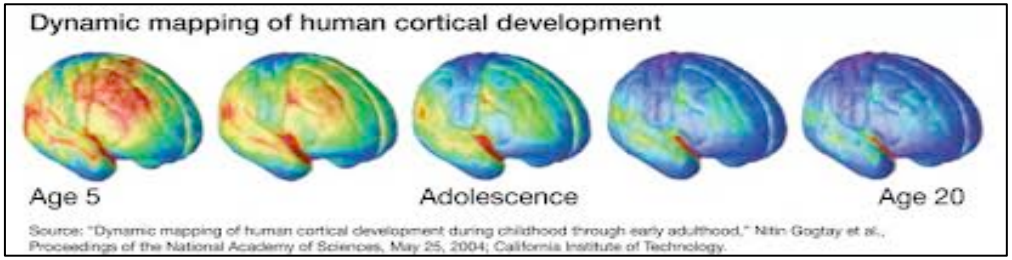
Approx. 25,000 genes in the human genome





Gene A

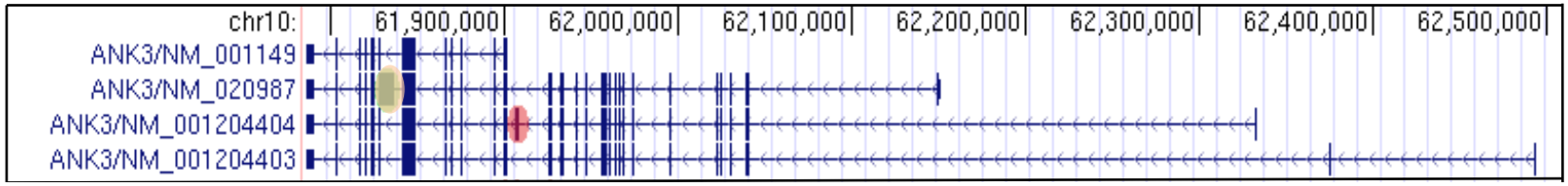
- Transcript A1 → Protein A1
- Transcript A2 → Protein A2
- Transcript A3 → Protein A3
- Transcript A4 → Protein A4



Some transcripts expressed ubiquitously, some highly specific (cell type AND timepoint)

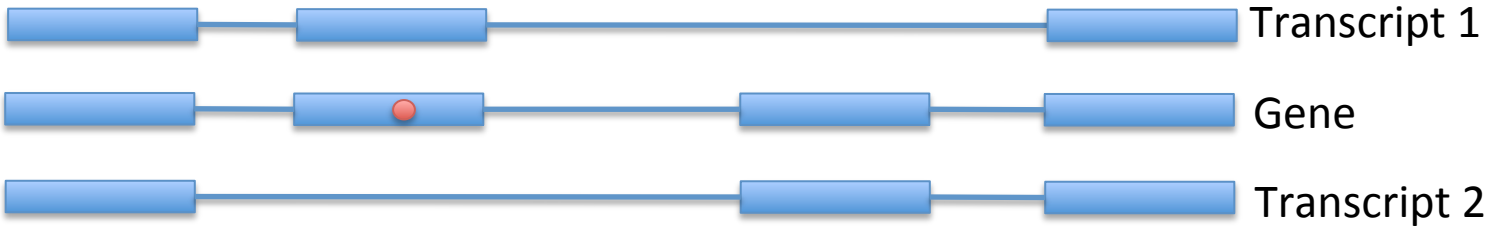
Other domains (spectrin binding)

Ankyrin repeats (membrane binding)



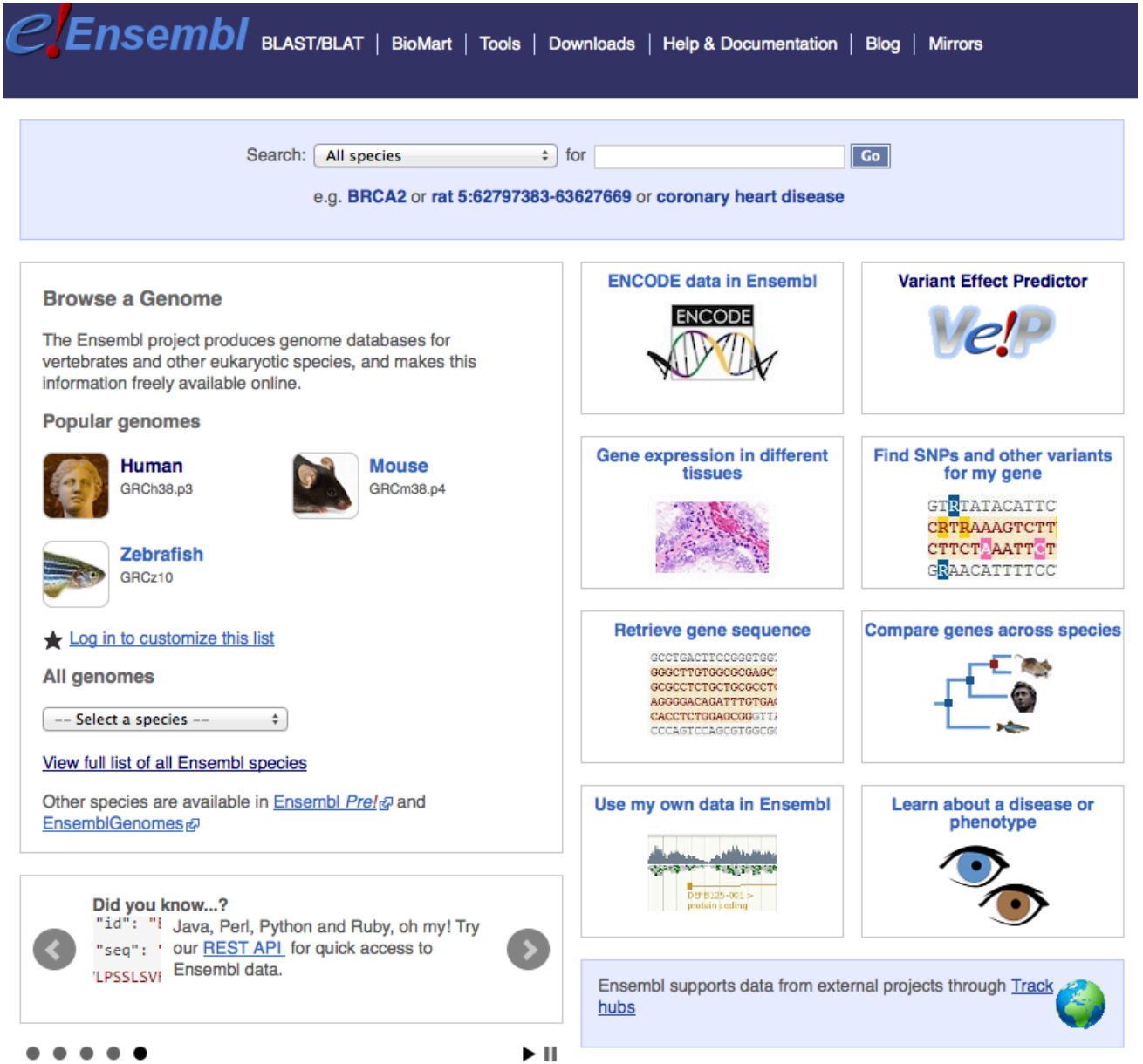
What is annotation?

- Adding information about the variants
- Two broad categories of annotations
  - annotation that **depend on gene models**
    - coding/non-coding
    - if coding: synonymous / non-synonymous
    - if non-synonymous >> what is the impact on protein structure (Polyphen, SIFT, etc)
  - annotations that **do not depend on gene models**
    - variant frequency in different database / different populations
    - degree of conservation across species
- Considerable complications caused by different gene models



- Two approaches to problem
  - decide **ex-ante** what which transcript to use for each gene
  - annotate with all transcript for a given gene and pick the **highest impact effect**

Ensembl – Basis for variant effect prediction



Annotation software

- Two sets of software
  - Annovar
    - provides a wide range of annotations that can be applied with one tool
    - we have experienced some inconsistencies in the results e.g. non-synonymous SNPs without polyphen score
  - SNPEff and dbNSFP (Non-Synoymous Functional Prediction)

- Both tested by GATK team
  - recommended snpEff, but with strict requirements
  - **snpEff version 2.0.5** (not 2.0.5d)
  - db should be **GRCh37.64** (which is the **Ensembl database version 64**)
  - should use the option **-onlyCoding true** (using false can cause erroneous annotation)

- GATKs VariantAnnotator to pick the highest impact.

- Finally, also annotate with **dbNSFP, which contains:**
  - variant frequencies
  - conservation scores
  - protein function effect (Polyphen, SIFT)

snpEff annotation get placed into INFO field

31942920 . G T 683.93 PASS

AC=1;AF=0.50;AN=2;BaseQRankSum=4.358;DP=73;DS;Dels=0.00;FS=0.000;HRun=0;
HaplotypeScore=1.7876;MQ=69.76;MQ0=0;MQRankSum=0.977;QD=9.37;ReadPosR
ankSum=0.508; VQSLOD=1.6292;culprit=QD

SNPEFF_AMINO_ACID_CHANGE=E114*;
SNPEFF_CODON_CHANGE=Gag/Tag;
SNPEFF_EFFECT=STOP_GAINED;
SNPEFF_EXON_ID=exon_22_31942847_31942957;
SNPEFF_FUNCTIONAL_CLASS=NONSENSE;
SNPEFF_GENE_BIOTYPE=processed_transcript;
SNPEFF_GENE_NAME=SFI1;
SNPEFF_IMPACT=HIGH;
SNPEFF_TRANSCRIPT_ID=ENST00000421060;

GT:AD:DP:GQ:PL 0/1:42,31:73:99:714,0,981

Explanation of snpEff fields (bold are important)

Field name	Example	Description
SNPEFF_EFFECT	NON_SYNONYMOUS_CODING	The highest-impact effect resulting from the current variant (or one of the highest-impact effects, if there is a tie)
SNPEFF_IMPACT	MODERATE	Impact of the highest-impact effect resulting from the current variant (**HIGH, MODERATE, LOW, or MODIFIER**)
SNPEFF_FUNCTIONAL_CLASS	MISSENSE	Functional class of the highest-impact effect resulting from the current variant (NONE, SILENT, MISSENSE, or NONSENSE)
SNPEFF_CODON_CHANGE	Tgc/Agc	Old/New codon for the highest-impact effect resulting from the current variant
SNPEFF_AMINO_ACID_CHANGE	C12S	Old/New amino acid for the highest-impact effect resulting from the current variant
SNPEFF_GENE_NAME	SLC6A18	Gene name for the highest-impact effect resulting from the current variant
SNPEFF_GENE_BIOTYPE	protein_coding	Gene biotype for the highest-impact effect resulting from the current variant
SNPEFF_TRANSCRIPT_ID	ENST00000296821	Transcript ID for the highest-impact effect resulting from the current variant
SNPEFF_EXON_ID	exon_5_1225470_1225752	Exon ID for the highest-impact effect resulting from the current variant

Impact classification of all effects

High-Impact effects	Moderate-Impact effe	Low-Impact effects
SPLICE_SITE_ACCEPTOR	NON_SYNONYMOUS_CODING	SYNONYMOUS_START
SPLICE_SITE_DONOR	CODON_CHANGE (note: this ef	NON_SYNONYMOUS_START
START_LOST	CODON_INSERTION	START_GAINED
EXON_DELETED	CODON_CHANGE_PLUS_CODO	SYNONYMOUS_CODING
FRAME_SHIFT	CODON_DELETION	SYNONYMOUS_STOP
STOP_GAINED	CODON_CHANGE_PLUS_CODO	NON_SYNONYMOUS_STOP
STOP_LOST	UTR_5_DELETED	
	UTR_3_DELETED	

All vs. top impact

- SnpEff uses Ensembl gene models and annotates initially with the effect in **all** different transcripts.

- Multiple functional annotations for the same variant make interpretation difficult

- One can then use GATK to pull out the TOP impact ie the most damaging effect

A second source of functional annotation: dbNSFP

- NSFP = Non-synonymous functional prediction
- Limited to non-synonymous variants
- Has many data fields. We use only:
  - dbnsfpSIFT_score
  - dbnsfpPolyphen2_HVAR_pred
  - dbnsfp29way_logOdds
  - dbnsfp1000Gp1_AF

Example of annotation with dbNSFP

766910 rs1809933 C T 556.42 PASS

AC=1;AF=0.50;AN=2;BaseQRankSum=1.366;DB;DP=30;Dels=0.00;FS=0.000;HRun=0;HaplotypeScore=1.8675;MQ=47.46;MQ0=0;MQRankSum=-0.651;QD=18.55;ReadPosRankSum=-1.757;SB=-109.24;

SNPEFF_AMINO_ACID_CHANGE=R42Q;SNPEFF_CODON_CHANGE=cGg/cAg;SNPEFF_EFFECT=NON_SYNONYMOUS_CODING;SNPEFF_EXON_ID=exon_5_766813_767034;SNPEFF_FUNCTIONAL_CLASS=MISSENSE;SNPEFF_GENE_BIOTYPE=processed_transcript;SNPEFF_GENE_NAME=ZDHHC11B;SNPEFF_IMPACT=MODERATE;SNPEFF_TRANSCRIPT_ID=ENST00000382776;

dbnsfp29way_logOdds=3.0289; SiPhy score based on 29 mammals genomes. The **larger** the score, the **more** conserved the site.

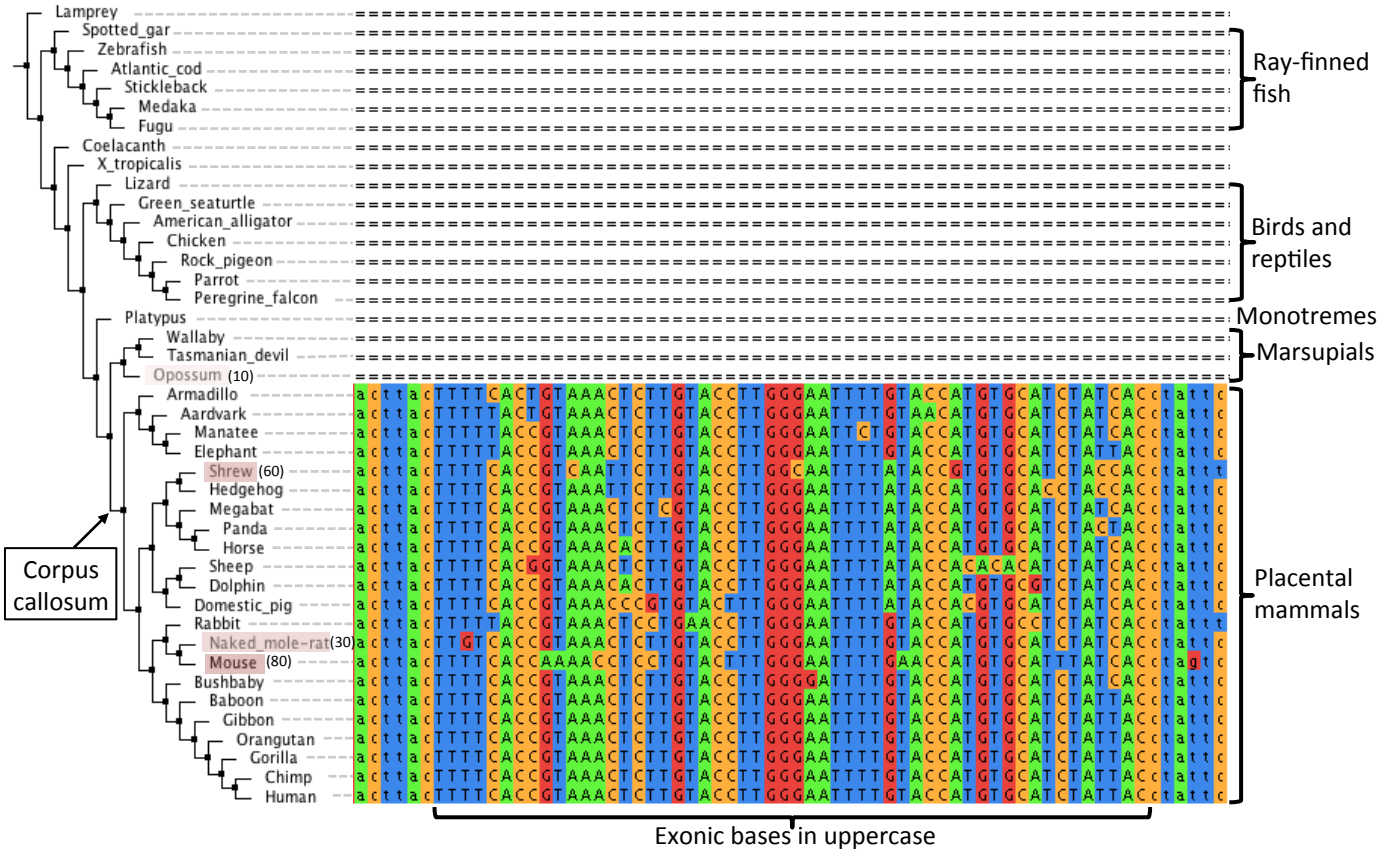
dbnsfp1000Gp1_AF=0.76; Alt. allele frequency in the whole 1000Gp1 data.

dbNSFP_Polyphen2_HVAR_pred=B; Polyphen2 prediction based on HumVar, "D" (”probably damaging"), "P" ("possibly damaging") and "B" ("benign"). Multiple entries separated by ";".

dbNSFP_SIFT_score=0.560000; SIFT score, If a score is smaller than 0.05 the corresponding NS is predicted as "D(amaging)"; otherwise it is predicted as "T(olerated)". SIFT predicts whether an amino acid substitution affects protein function.

GT:AD:DP:GQ:PL 0/1:5,25:30:98:586,0,98

Variation between species



Multiz DNA sequence alignment for the little exon from the UCSC genome browser (note: DNA from the forward strand, but ANK3 is encoded on the reverse strand). Species selected from 100 vertebrate alignments in order to have representative species throughout the phylogeny. Pink shading of four mammalian species with similarly sized cortices approximately proportional to cortical neuron densities (in millions of cells per gram in parenthesis) (Seelke 2014)

Some human variant databases

ExAC Browser (Beta) | Exome Aggregation Consortium

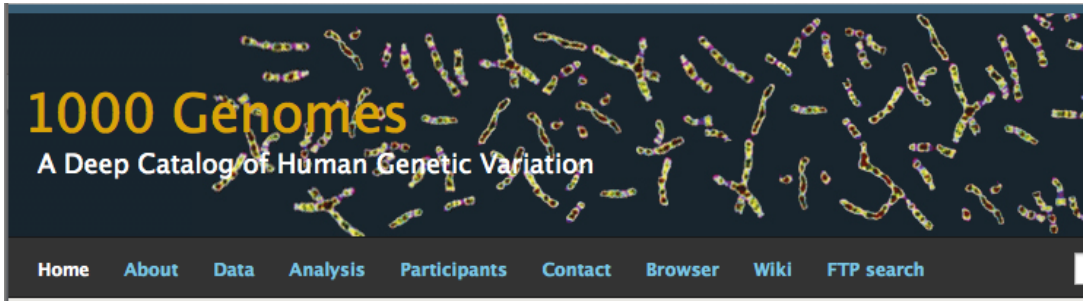
Search for a gene or variant or region

Examples - Gene: PCSK9, Transcript: ENST00000407236, Variant: 22-46615880-T-C, Multi-allelic variant: rs1800234, Region: 22:46615715-46615880

About ExAC

The Exome Aggregation Consortium (ExAC) is a coalition of investigators seeking to aggregate and harmonize exome sequencing data from a wide variety of large-scale sequencing projects, and to make summary data available for the wider scientific community.

The data set provided on this website spans 60,706 unrelated individuals sequenced as part of various disease-specific and population genetic studies. The ExAC Principal Investigators and groups that have contributed data to the current release are listed here.



What are the most useful fields

- The tradeoff
  - Use very strong filters and get very short lists but risk that the causal variant has been excluded
  - Use weaker filters to keep the causal variant “in” but risk getting very long lists

- A typical strategy is to start with very strong filters
  - See whether you can identify a good candidate
  - If not loosen some of the filters

- Strong filters are for example:
  - SNPEFF IMPACT: HIGH or HIGH and MEDIUM
  - Low frequency in 1000G or in Exac
  - Polyphen damaging

HAVING A GO ON YOUR OWN WITH THE REAL DATASET

A full pipeline from fastq to annotated VCF

- 040_functionalAnnotation.bash
- This is a recap of everything we have done in the course
- Again try to be precise.
- **Be quick** on the sections you fully understand
- Ask questions on the sections where you are still unsure.
- Do not copy and paste commands that you do not understand

Hunt for the causal mutant

- Maybe the most fun part of the course

- In **running out of time**: use filtering at the command line in **040_functionalAnnotation.bash**

- **If sufficient time**: work in spreadsheet following the instructions in: **041_findCausalVariantExercise.txt**

MULTIPLE SAMPLES

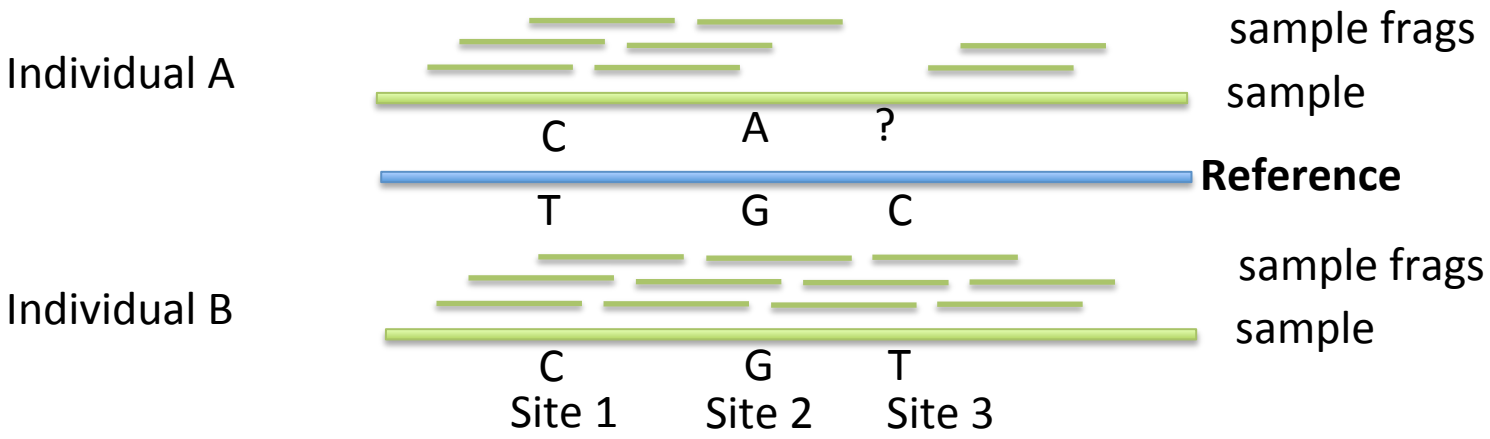
FROM HERE ON IT IS JUST USEFUL INFO

Batch calling / Joint calling

- Better specificity through filtering
  - more variant sites to train on in soft filtering
  - increased power of tests like FS
- Better sensitivity for regions of low coverage (Need to graphically illustrate)
- **Squaring up**
  - **genotyping all samples at all sites where at least one of the samples is variant (see next slide)**

The squaring off problem

Variant files only contain variant sites!



Merging singles

	Indiv. A	Indiv. B
Site 1	Variant	Variant
Site 2	Variant	
Site 3		Variant

Indiv. A	Indiv. B
Variant	Variant
Variant	-
-	Variant

Both sites lack information but we have information on individual B at site 2 (REF G)

Joint calling

Indiv. A	Indiv. B
Variant	Variant
Variant	Ref
No call	Variant

ALL individuals characterised at ALL sites where at least ONE individual is variant

Additional important applications and methods

- Somatic variant calling (e.g. cancer cells)
- Pooled sequencing and allele count estimation
- Trios and denovo variants
- Pipelines like nextgen bcbio
- Structural variation programs like XHMM and BreakDancer
- There used to be a very good website comparing the effect of changing types of sequencing, sequence depth, alignment tool, with or without alignment, and the variant caller but it is now unfortunately offline.

HaplotypeCaller and FreeBayes

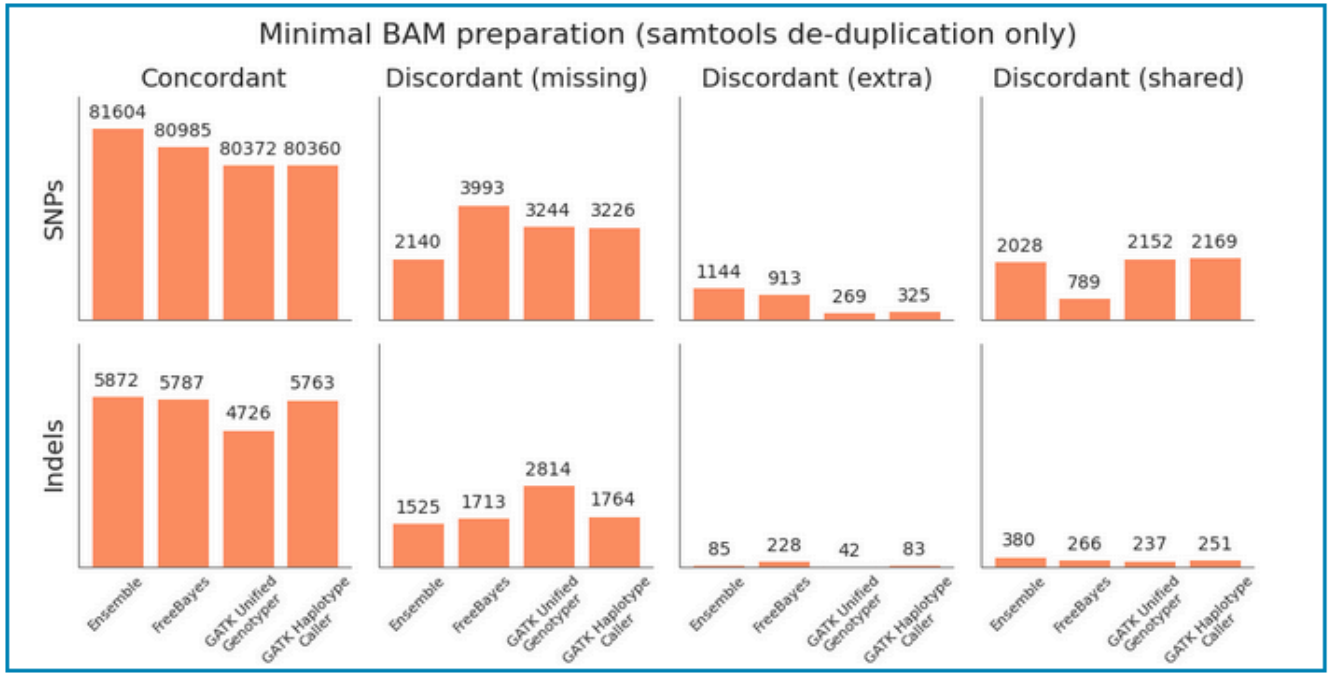
- HaplotypeCaller (in GATK) is an alternative to UnifiedGenotyper which is better at detecting insertions and deletions
  - performs local assembly which gives the same advantage as having longer reads.
- Possibility of dropping the realignment
- FreeBayes is another similar variant caller that does well without realignment.

Note:

1. **Ensembl** is a consensus between many variant callers
2. **Discordant missing:** called by method but not by others
3. **Discordant extra:** missed by method but not by others
4. **Discordant shared:** called by method but discordant on variant **genotype**

Variant caller comparison

from bcbio.wordpress.com (Brad Chapman)



Note: broadly very similar results

Biggest percentage differences:

- Missed indels by UnifiedGenotyper
- Extra indesl and SNPs by FreeBayes

CONCLUDING REMARKS

Reading list

Nielsen et al., Genotype and SNP calling from next-generation sequencing data. Nature Reviews Genetics 2011

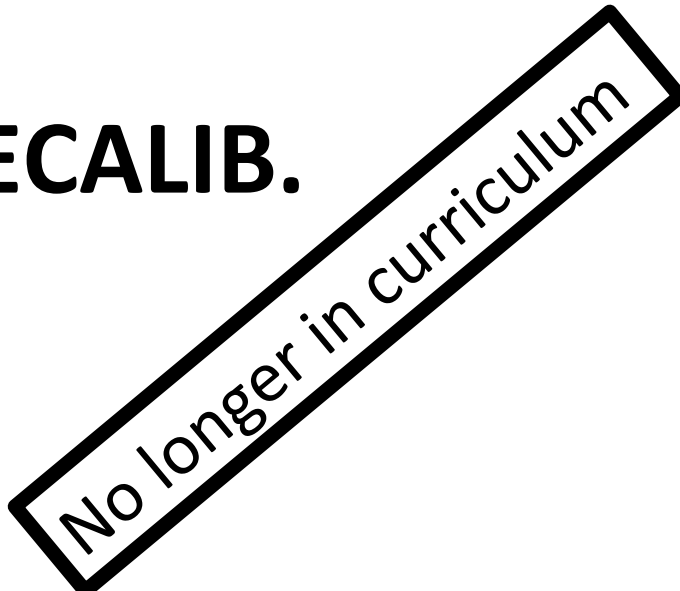
Ng et al., Exome sequencing identifies the cause of a mendelian disorder. Nature Genetics 2010

Summary in plenum of the whole process and all the key concepts

Good luck with the other mountains!!

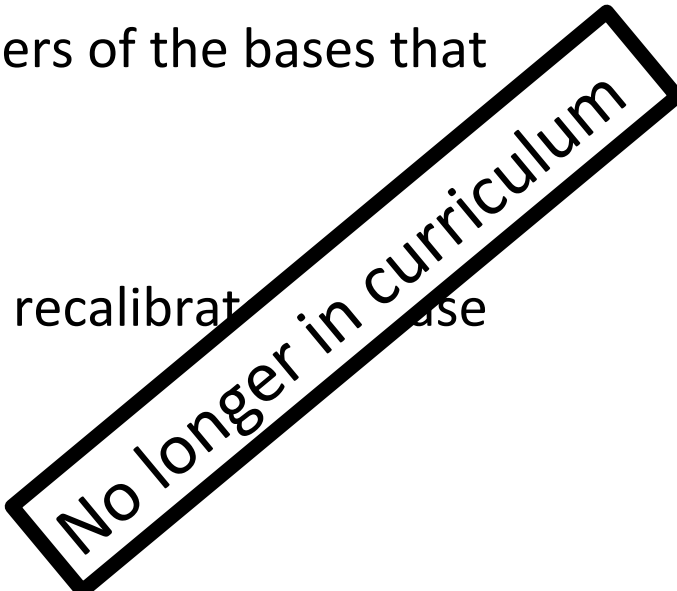
APPENDIX

BASE QUALITY SCORE RECALIB.

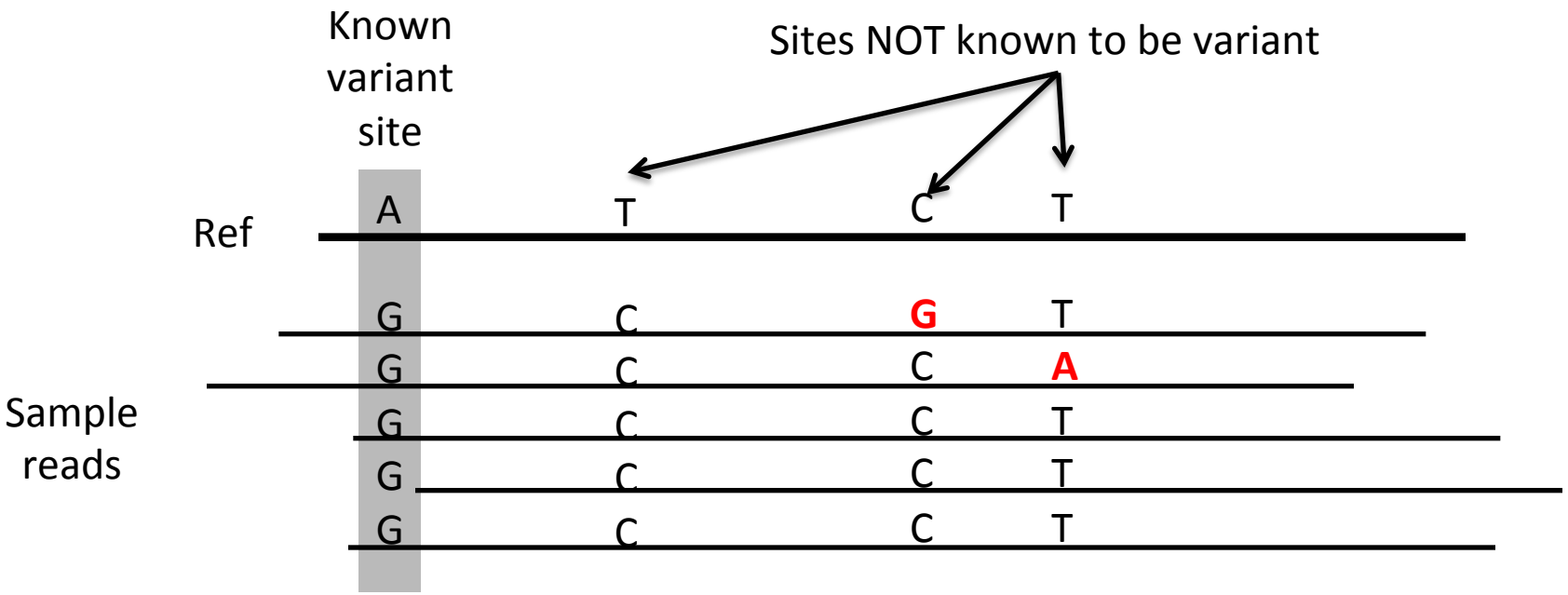


Theoretical vs Empirical error rates / qualities

- The qualities in the fastq file are computed using a model
- This model is not perfect >> there are discrepancies between the model and the empirical error rate
- We can compute a good approximation of the empirical error rate by identifying all sites where there are mismatches between the read and the reference (**being careful to ignore sites with known SNPs**)
- We can analyse whether there are parameters of the bases that covary with the discrepancy
  - e.g. cycle
- We can use these quantified covariances to recalibrate base qualities >> more accurate qualities



Overview of Base Quality recalibration



Sample variant at site **(common)**

Sample variant at site **(rare)**

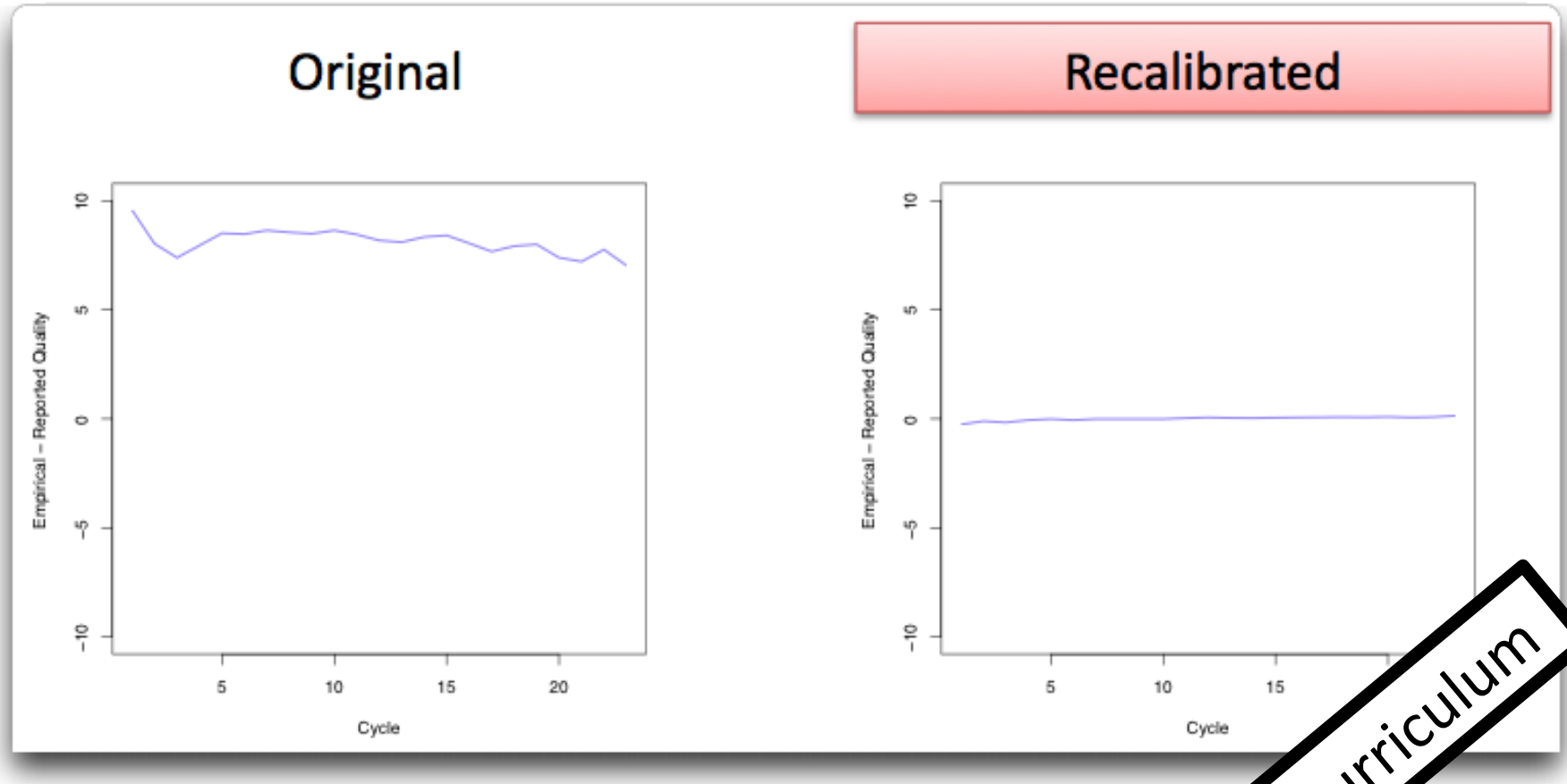
Base call error **(common)**

		Theoretical base Q	Empirical base Q
Cycle	**1**	**28**	**29**
	2	**29**	**30**
	n	**…**	**…**
Dinucleotide	**AG**	**24**	**24**
	AT	**29**	**22**
	AC	**…**	**…**
	…		

Empirical ba… be
comput… e
obse… match rate
b… e read bases
… reference bases
… align **to (at sites that**
are not obviously variant)

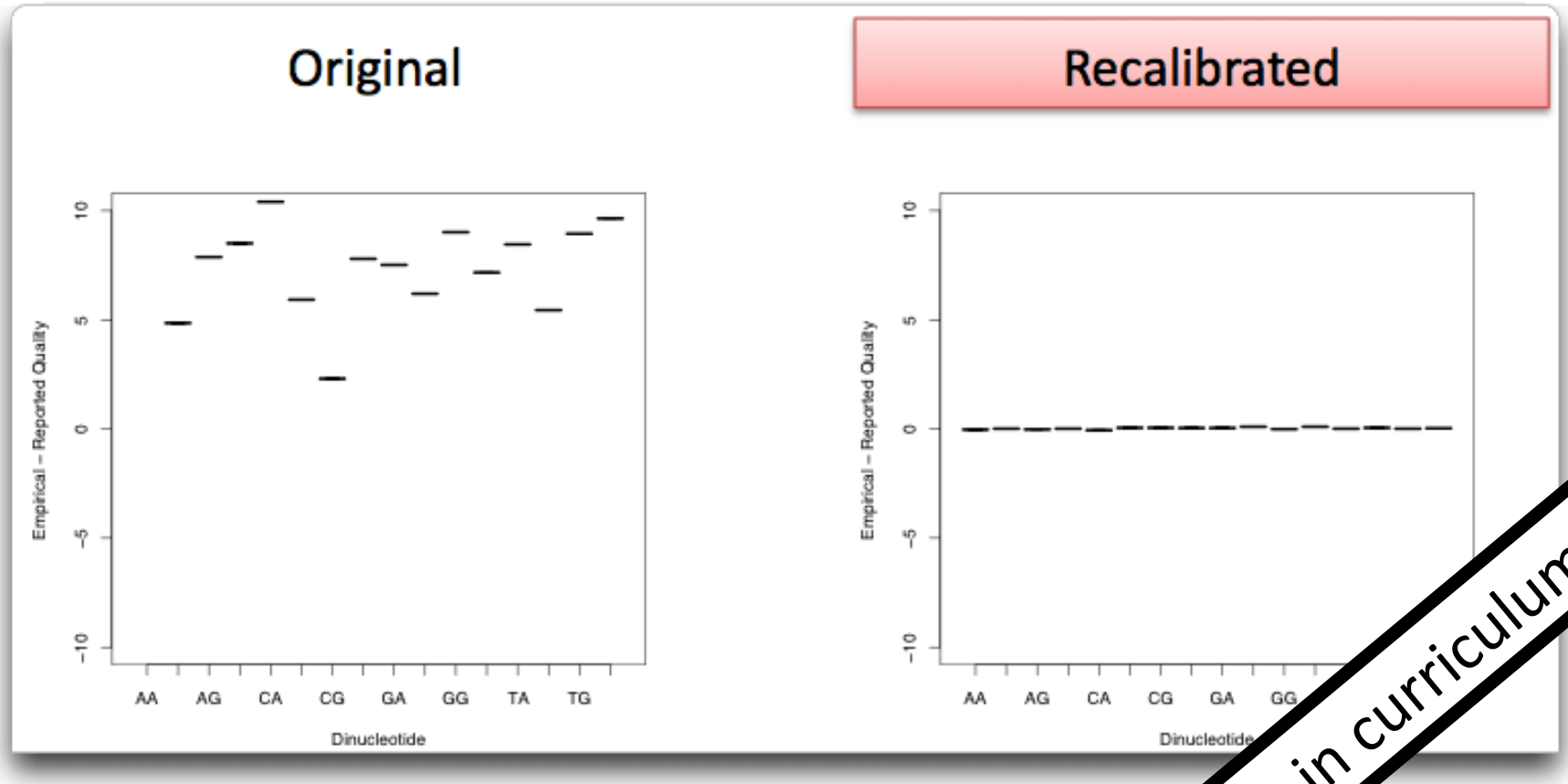
No longer in curriculum

Discrepancy and cycle

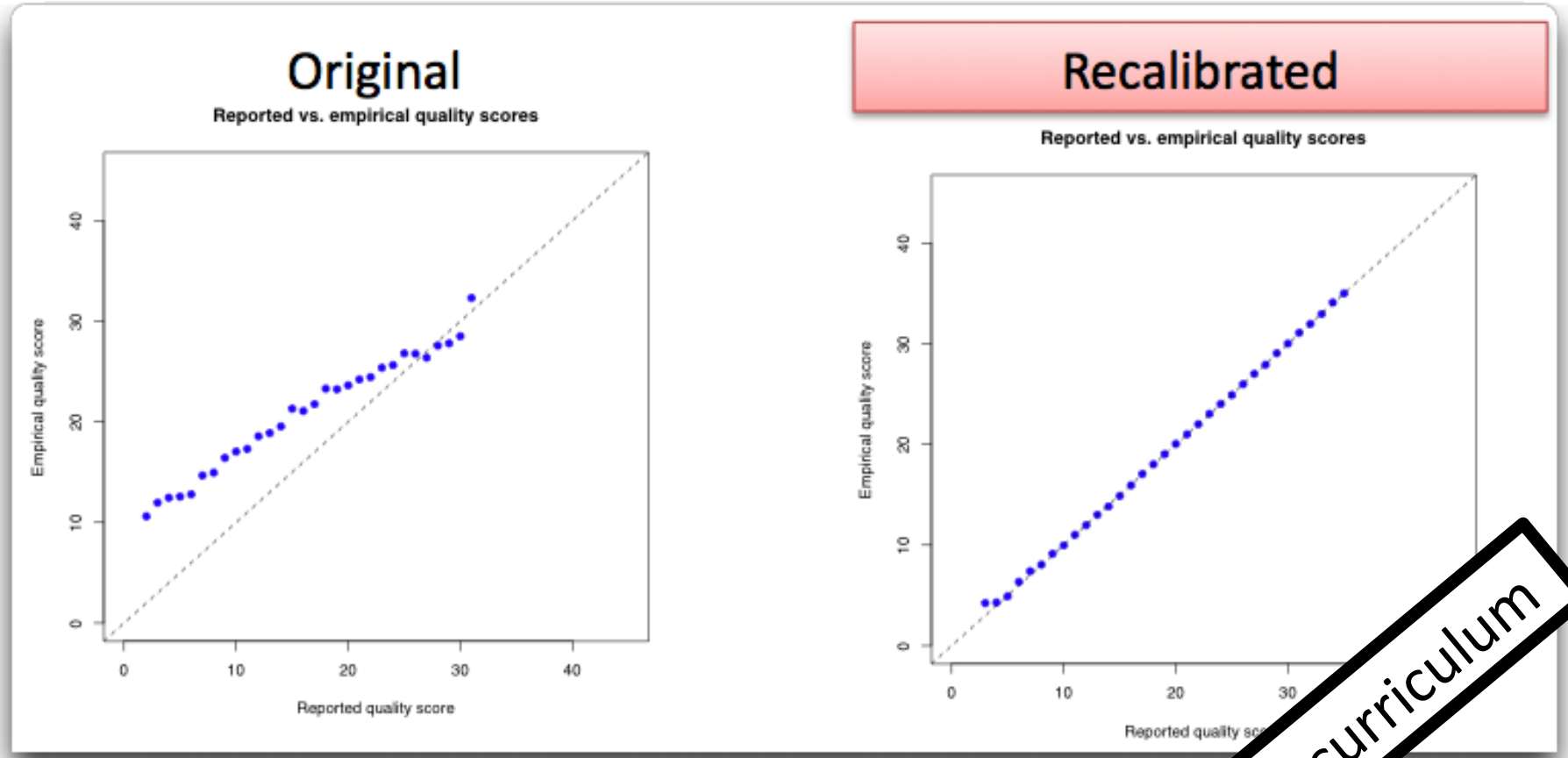


No longer in curriculum

Discrepancy and Dinuc context



Result of recalibration



NB: The theoretical base qualities have become very go ere seems to be little to be gained in recalibrating "modern" fa

No longer in curriculum