Somatic and germline variant calling

Outline

- Tumor purity and clonality
- Somatic mutation vs. germline mutation vs. germline polymorphism
- Germline variant calling
  - Methodology
  - Tools
  - Quiz
- Somatic variant calling
  - Methodology
  - Tools
  - Quiz
- Variant annotation (VEP)
- File format
- Manual curation
- Mutational signatures
  - General about signatures
  - Microsatellite instability/hypermutation
  - Quiz

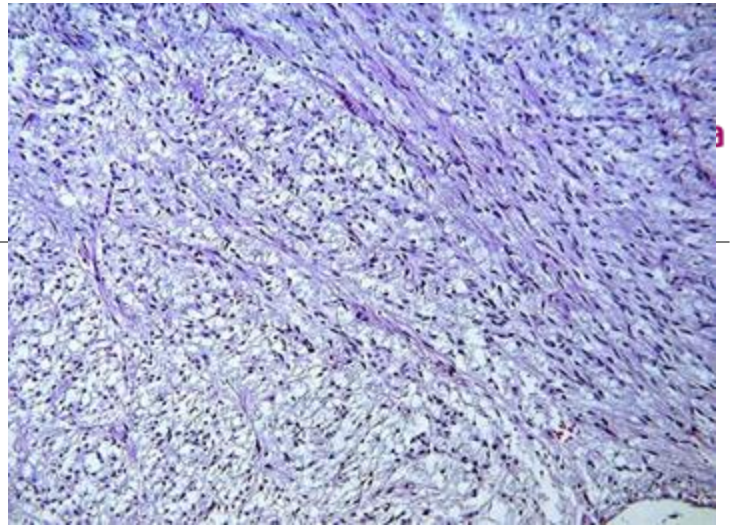
Learning outcomes and course content

- Learning outcomes:
  - Understand how to apply technology to obtain relevant information from the cancer genome.
  - Understand how to apply technology to obtain relevant information from the cancer genome.
  - Call somatic- and germline variation.
  - Annotate somatic- and germline variation.

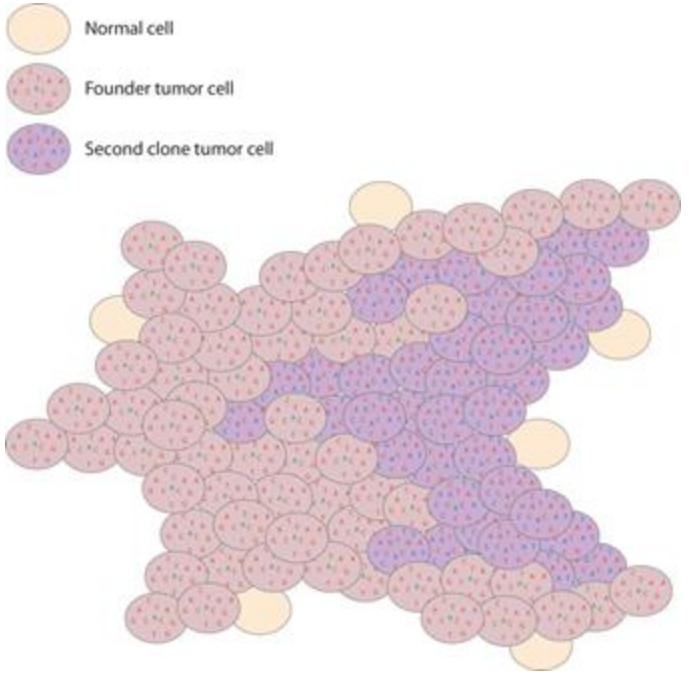
- Course content:
  - Calling somatic- and germline variation:
    - Point mutations and indels.
  - File formats for variant calling.
  - Annotating somatic- and germline variation.

Tumor purity and clonality

- Tumors are often impure
  - Mix of tumor cells and normal (germline) cells
  - Tumor purity expressed as fraction or percentage
  - Cancer DNA fraction is a very closely related term
- Tumors often contain multiple clones
  - Diverse collections of cells harboring different mutations
  - Often one original clone with initial mutations
  - Subclones containing additional mutations may form
  - Treatments may favor one subclone which has resistance mutation, causing it to take over
    - → clonal evolution

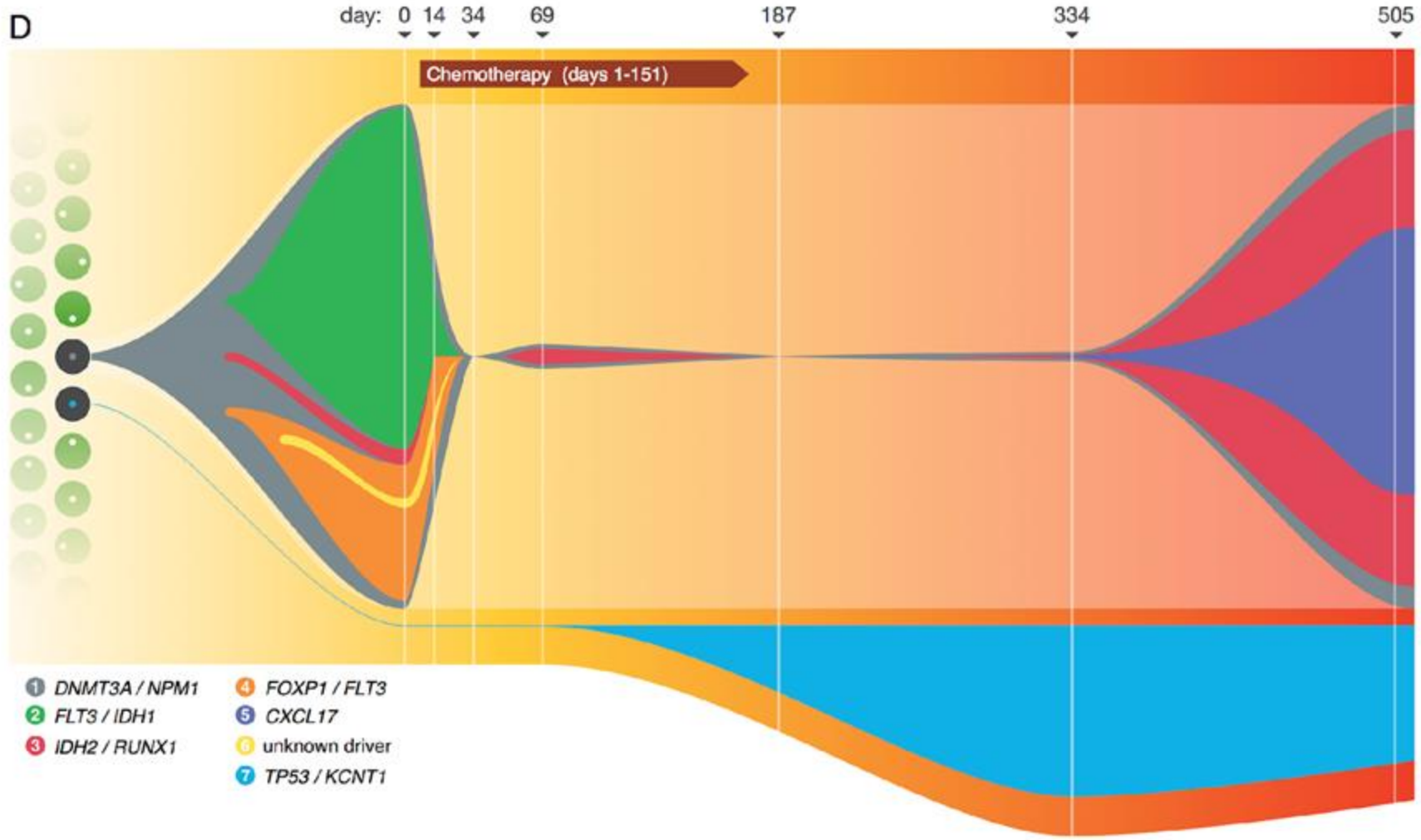


Credit Griffith brothers' lecture #5



Credit Griffith brothers' lecture #5

Clonal evolution



Credit Griffith brothers' lecture #5

Somatic mutation vs. germline mutation vs. germline polymorphism

- Germline mutations
  - Present in egg or sperm
    - All cells of affected offspring
  - Heritable
  - Cause of familial cancers

- Germline polymorphisms
  - Present in egg or sperm
    - All cells of affected offspring
  - Heritable
  - Common in the population, > 1%
  - Generally not impacting disease
  - SNP – Single Nucleotide Polymorphism

- Somatic mutations
  - Occur in non-germline tissues
    - Only tumor cells (breast, lung, blood, etc.)
  - Non-heritable
  - Various reasons
    - Smoking, UV light, oxidation in cells, etc.
  - Cause of sporadic cancers
    - And of familial cancers – in combination with germline mutations

Small variants

- SNVs – single nucleotide variants
  - Change from one base to another at one single position of the genome
- Indels – Small insertions and deletions
  - 1 – ~30 bases added or removed
  - Many more bases can be deleted, but then it’s no longer a ”small variant”, will be covered on Friday

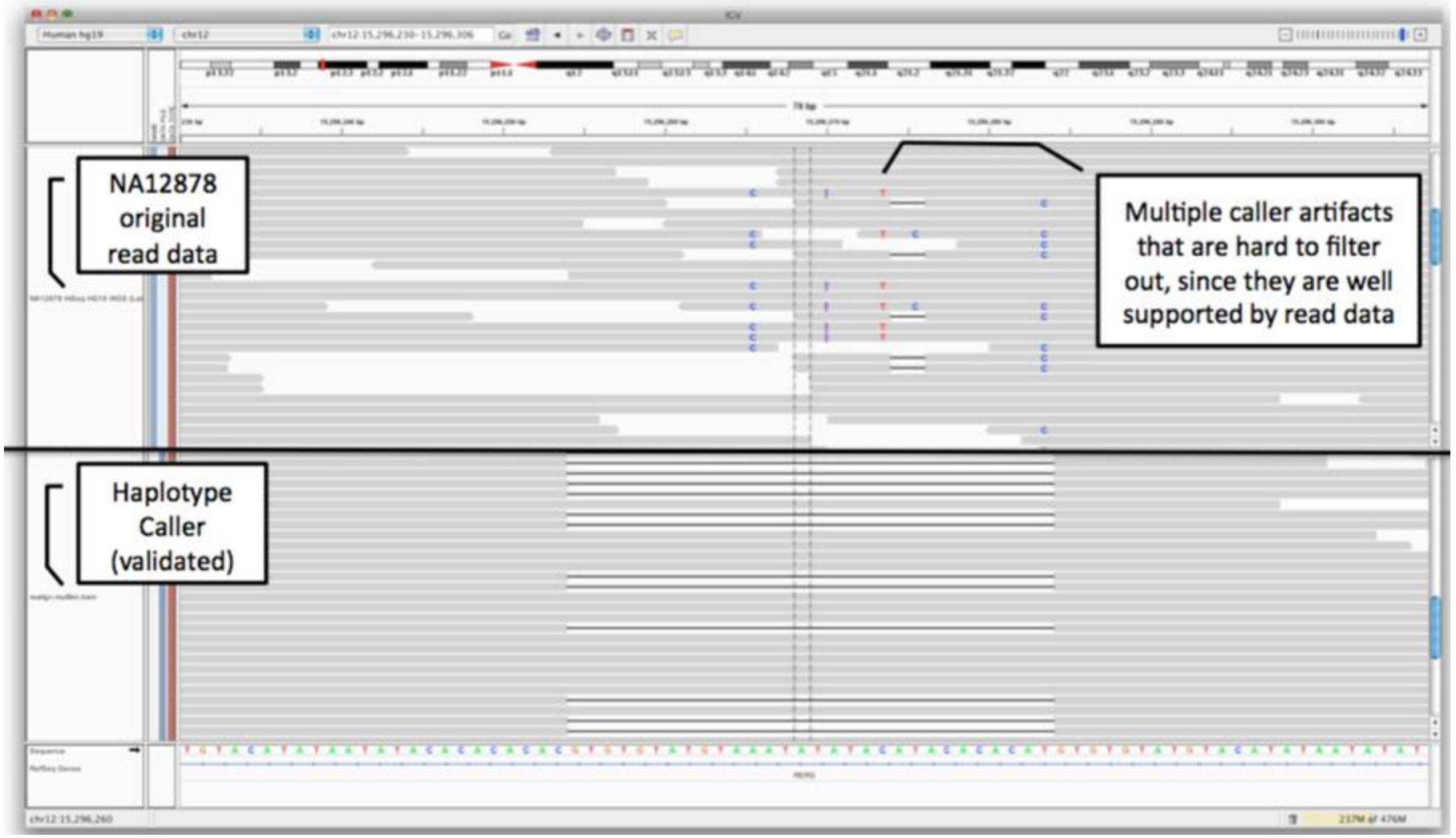
Germline mutation calling – methodology

- Mutation/variant calling – to identify mutations from e.g. sequencing data
  - Focus: DNA data
- Germline sample: often white blood cells, WBCs
- SNVs and small indels
- VAFs (variant allele frequencies): ~50% or 100%
  - One or two mutated alleles out of totally two alleles
- General method:
  - Find positions where a fraction of mapped reads have base deviating from the reference genome – alternate allele
  - If significant difference from the reference – call a germline variant

Germline mutation calling – methodology and tools

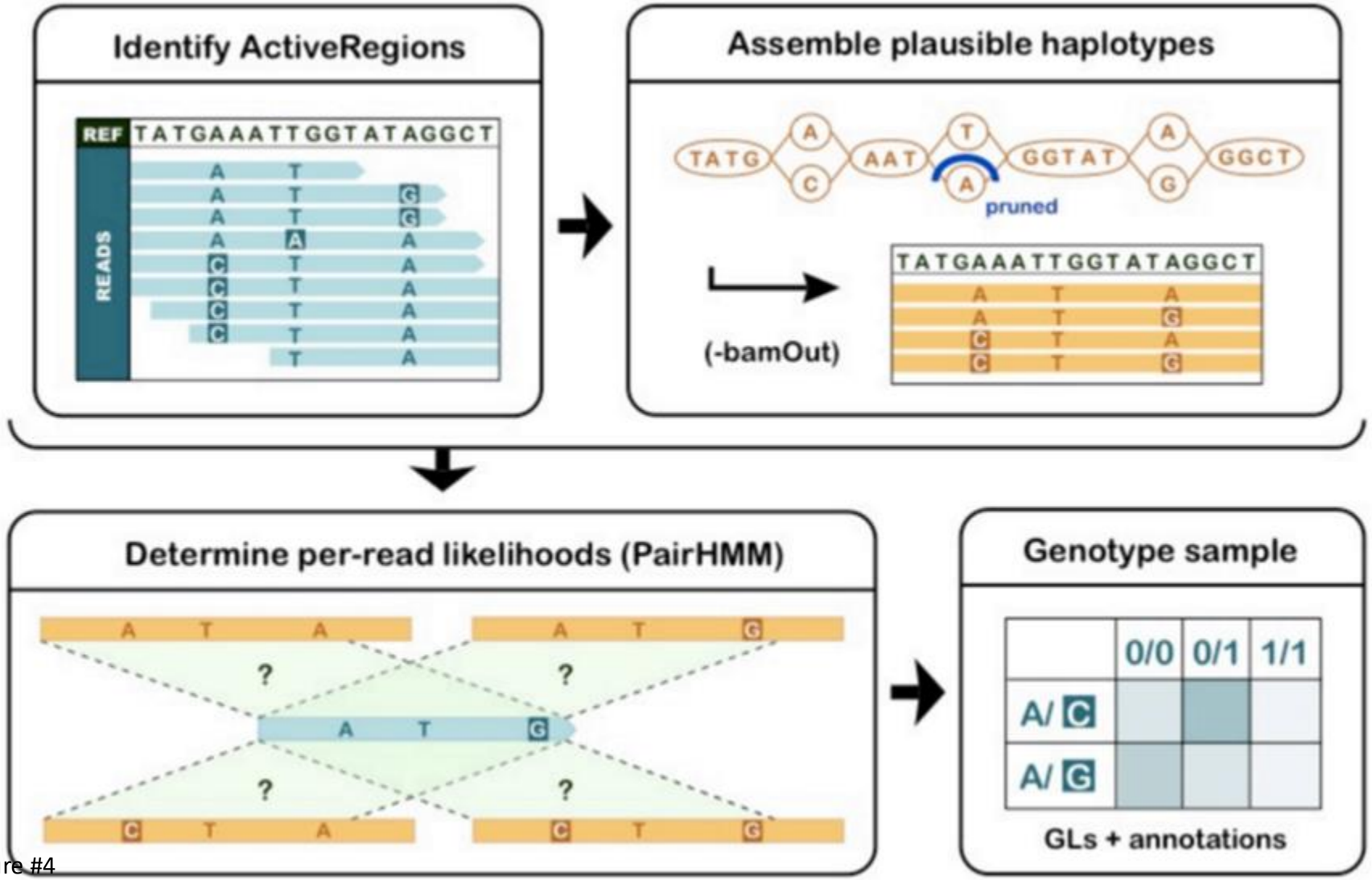
- Call SNVs and Indels separately by considering each variant locus
  - Very fast
  - Assumes bases are independent
- Call SNVs and indels simultaneously via Bayesian genotype likelihood model
  - More computationally intensive
- Call SNVs, indels and SVs simultaneously by performing a local de novo assembly
  - More computationally intensive
  - More accurate – gets rid of many false positives especially indels
  - GATK HaplotypeCaller

Germline mutation calling – methodology



Credit Griffith brothers' lecture #4

Germline mutation calling – HaplotypeCaller

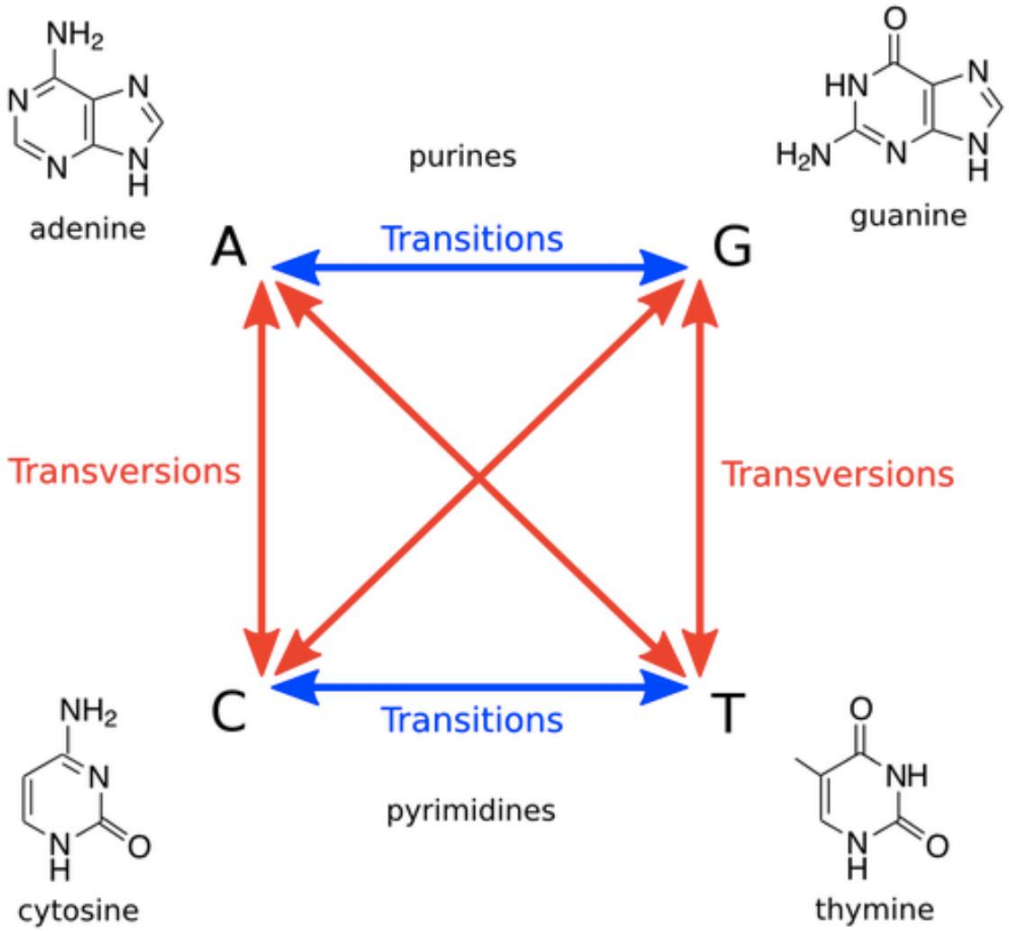


Credit Griffith brothers' lecture #4

GATK recommended filters

- SNPs
  - QD < 2.0 (variant quality/depth of non-ref samples)
  - MQ < 40.0 (Mapping quality)
  - FS > 60.0 (Phred score Fisher's test pvalue for strand bias)
  - SOR > 3.0 (Strand odds ratio, aims to evaluate whether there is strand bias in the data)
  - MQRankSum < -12.5 (mapping quality of reference reads vs alt reads)
  - ReadPosRankSum < -8.0 (distance of alt reads from end of the read)
- Indels
  - QD < 2.0
  - FS > 200.0
  - SOR > 10.0
  - ReadPosRankSum < -20.0
  - InbreedingCoeff < -0.8

Transition/Transversion ratio (Ti/Tv)



Ratios:

Random = 0.5

WGS = 2.0-2.1

Exome = 3-3.5

Watch for major deviation from typical ratio

Credit Griffith brothers' lecture #4

A good paper

- The “gnomAD” paper
- Published in Nature
- Supplementary Information:
  Many details on how to properly call and filter germline variants

Article

The mutational constraint spectrum quantified from variation in 141,456 humans

https://doi.org/10.1038/s41586-020-2308-7

Received: 27 January 2019

Accepted: 26 March 2020

Published online: 27 May 2020

Open access

Check for updates

Konrad J. Karczewski[1,2]✉, Laurent C. Francioli[1,2], Grace Tiao[1,2], Beryl B. Cummings[1,2,3], Jessica Alföldi[1,2], Qingbo Wang[1,2,4], Ryan L. Collins[1,4,5], Kristen M. Laricchia[1,2], Andrea Ganna[1,2,6], Daniel P. Birnbaum[1,2], Laura D. Gauthier[7], Harrison Brand[1,5], Matthew Solomonson[1,2], Nicholas A. Watts[1,2], Daniel Rhodes[8], Moriel Singer-Berk[1,2], Eleina M. England[1,2], Eleanor G. Seaby[1,2], Jack A. Kosmicki[1,2,4], Raymond K. Walters[1,2,9], Katherine Tashman[1,2,9], Yossi Farjoun[7], Eric Banks[7], Timothy Poterba[1,2,9], Arcturus Wang[1,2,9], Cotton Seed[1,2,9], Nicola Whiffin[1,2,10,11], Jessica X. Chong[12], Kaitlin E. Samocha[13], Emma Pierce-Hoffman[1,2], Zachary Zappala[1,2,14], Anne H. O’Donnell-Luria[1,2,15,16], Eric Vallabh Minikel[1], Ben Weisburd[7], Monkol Lek[17], James S. Ware[1,10,11], Christopher Vittal[2,9], Irina M. Armean[1,2], Louis Bergelson[7], Kristian Cibulskis[7], Kristen M. Connolly[18], Miguel Covarrubias[7], Stacey Donnelly[1], Steven Ferriera[18], Stacey Gabriel[18], Jeff Gentry[7], Namrata Gupta[1,18], Thibault Jeandet[7], Diane Kaplan[7], Christopher Llanwarne[7], Ruchi Munshi[7], Sam Novod[7], Nikelle Petrillo[7], David Roazen[7], Valentin Ruano-Rubio[7], Andrea Saltzman[1], Molly Schleicher[1], Jose Soto[7], Kathleen Tibbetts[7], Charlotte Tolonen[7], Gordon Wade[7], Michael E. Talkowski[1,5,19], Genome Aggregation Database Consortium*, Benjamin M. Neale[1,2,9], Mark J. Daly[1,2,6,9] & Daniel G. MacArthur[1,2,150,151]✉

Germline variant calling – Questions?

Germline mutation calling quiz

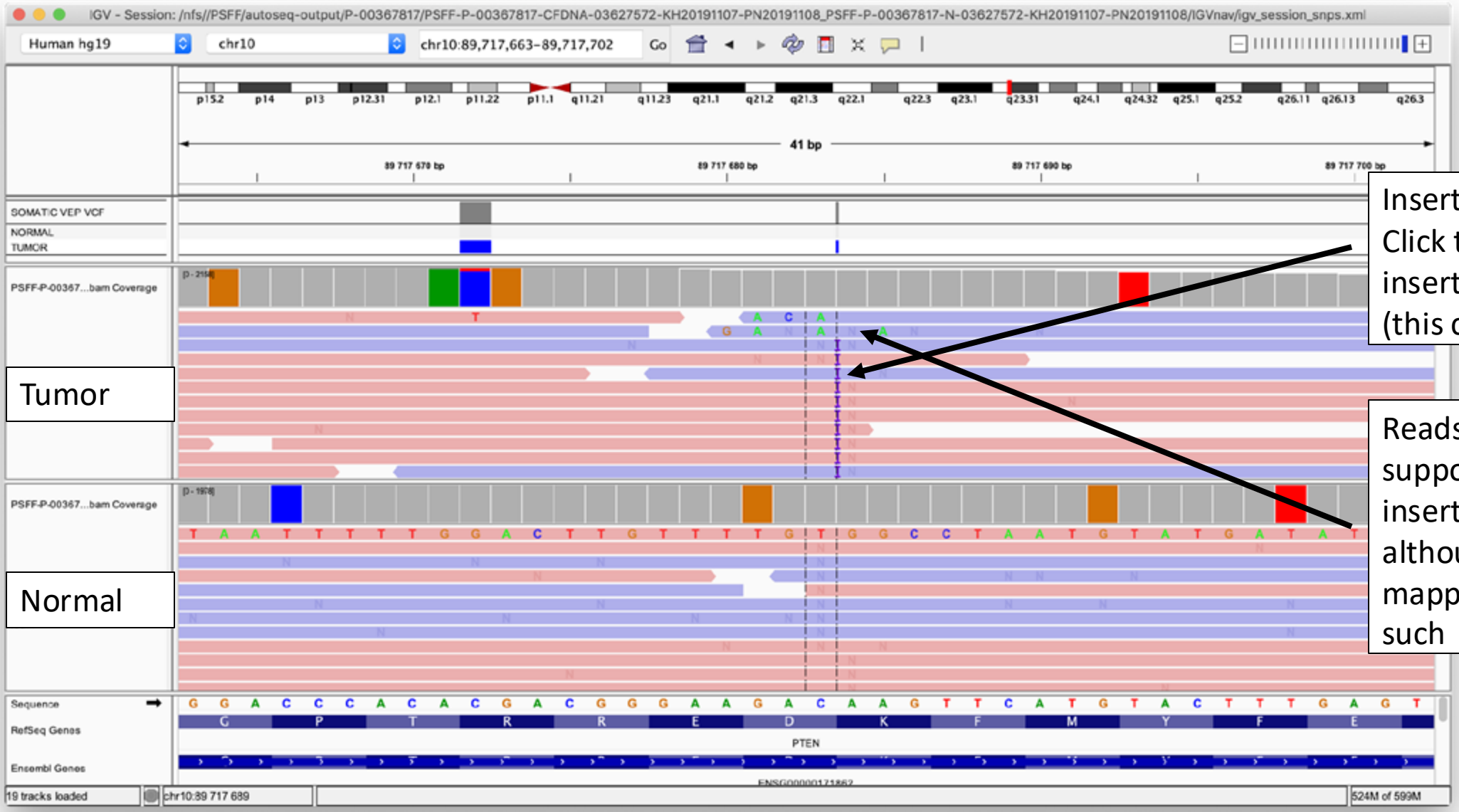
Somatic mutation calling – methodology

- Mutation/variant calling – to identify mutations from e.g. sequencing data
  - Focus: DNA data
- Somatic mutations are best distinguished by comparison of tumor to a matched normal
- Normal sample: Germline sample from the same individual
  - If not available: use healthy donor sample → requires additional filtering
- SNVs and small indels
- VAFs: ~1% - 100%, depending on purity, ploidy, clonality
  - Limited detection capacity for VAF < 1%
- General method:
  - Find positions where a fraction of mapped reads have base deviating from the reference genome – alternate allele
  - Compare with germline sample
  - If significant difference to germline – call a somatic variant

Somatic mutation calling – SNV

Somatic mutation calling – deletion

Somatic mutation calling – insertion



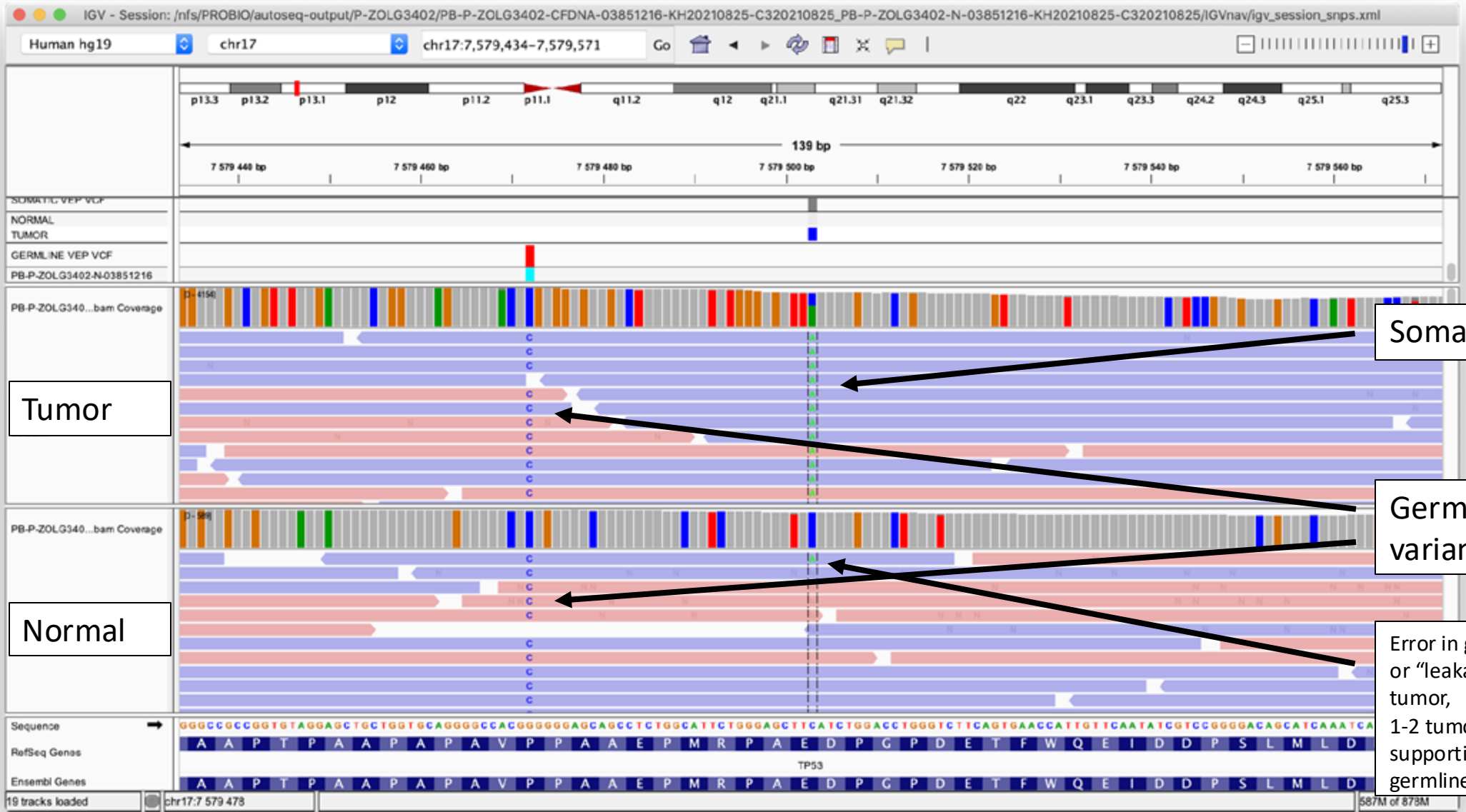
Tumor

Normal

Insertion
Click to see inserted bases (this case A)

Reads also supporting insertion although not mapped as such

Germline vs somatic variant



Somatic mutation calling – tools

- Tools (variant callers):
  - Vardict, Varscan – Counting the reads with ref bases and non-ref bases (alt allele) in tumor and normal, calculating significance test, if significant call the variant – old and outdated
  - GATK Mutect2 – Local de novo assembly of sites with non-ref bases
    - Same basis as HaplotypeCaller but with additional comparison to a normal sample, and not assuming VAF 50% or 100%
  - SAGE (Hartwig) – Identifying sites with non-ref bases and nearby "read context", weighting reads based on the various quality parameters and summing the weights, if high enough weighted support in tumor (and <4% in germline) call the variant
  - Different properties – better at different types of variants

Somatic mutation calling – tools

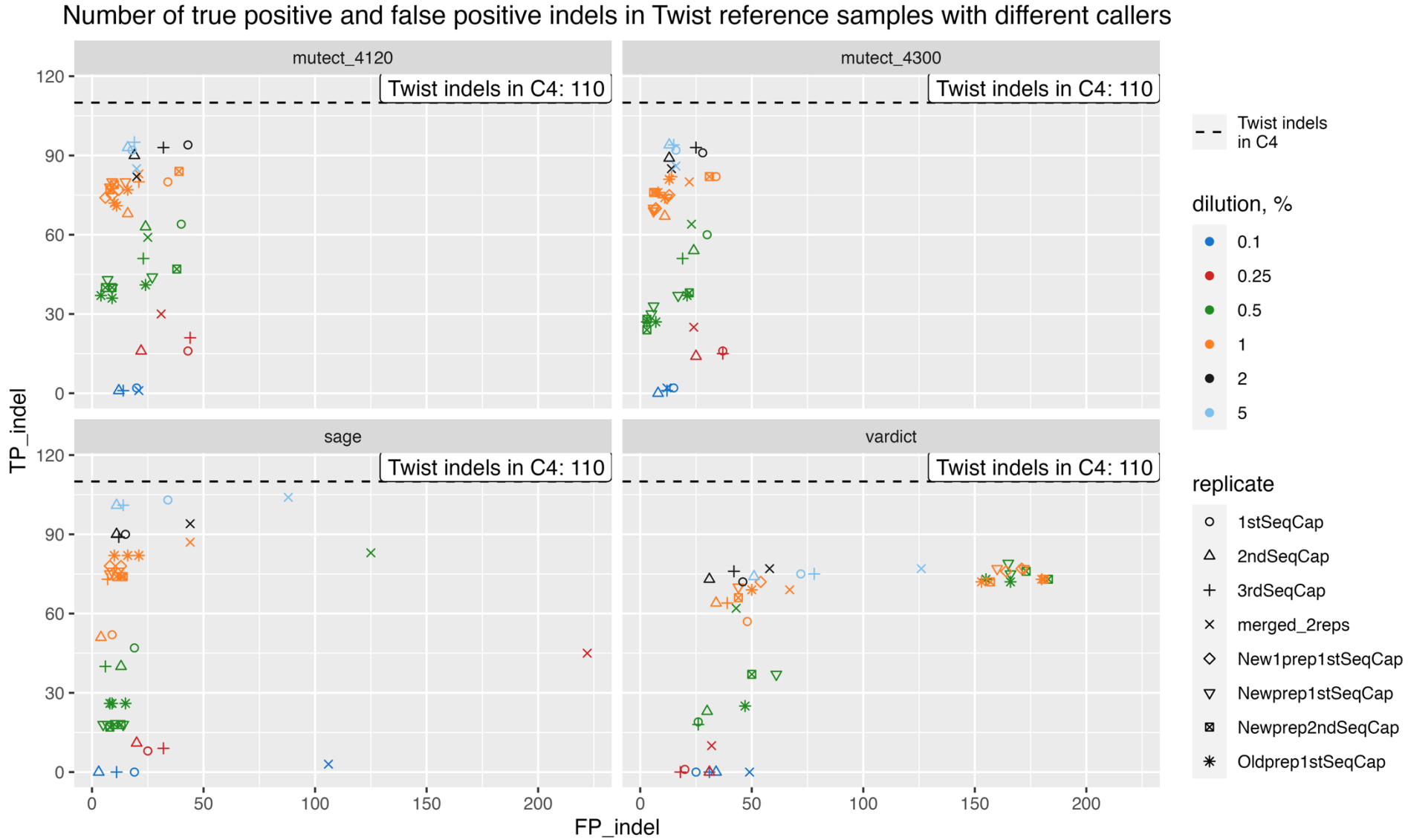
- Union:
  - Not all callers are able to call all variants
  - Take union from multiple callers to get a more complete result
- Merging of results
  - Somaticseq
  - bcftools concat
  - Picard MergeVcfs

Variant caller evaluation

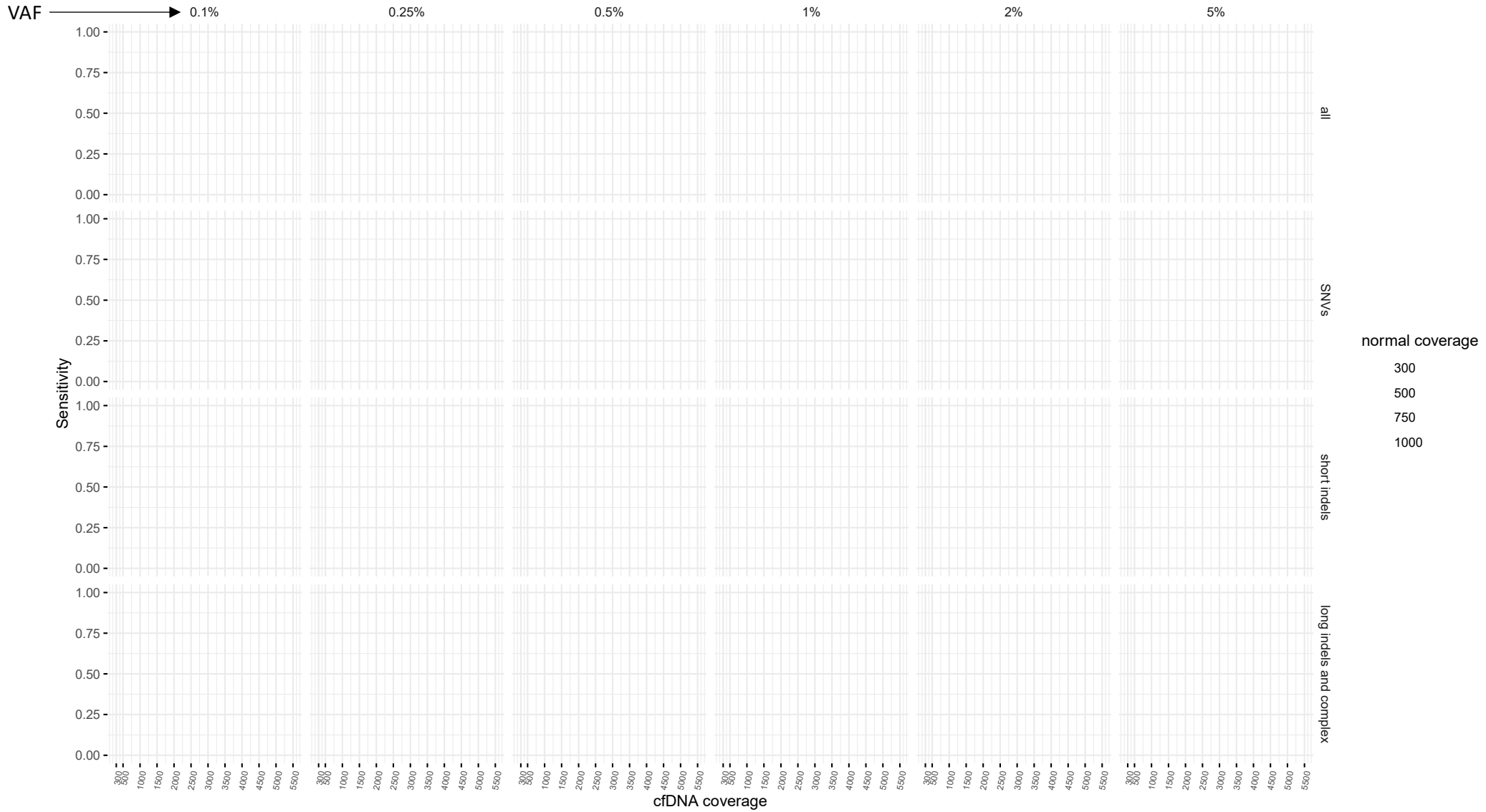
- Literature review in 2023 for relevant callers
  - Compared properties, need of licenses, active maintenance etc.
- VarDict, Mutect2 and SAGE were selected for evaluation
  - Twist reference samples with known variants for sensitivity analysis
  - Healthy anonymous donors from Blodcentralen for specificity/false positive analysis
    - These should not have any somatic variants
- Result: Mutect2 and SAGE selected for use

Comparison of callers - SNVs

Comparison of callers - indels



The need of coverage - sensitivity



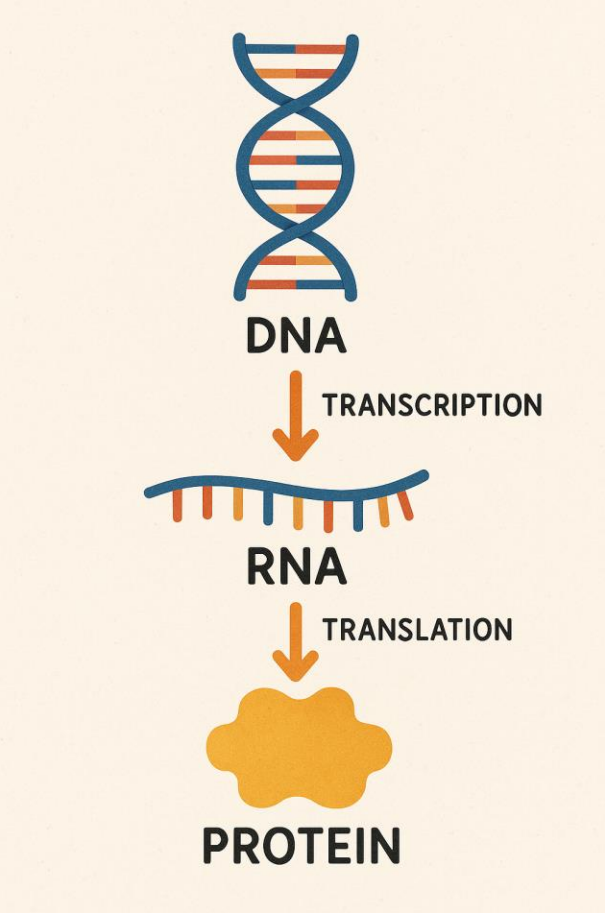
The need of coverage – false positives (specificity)

AD: alternate allele depth, number of reads supporting the variant

False positives dependent on normal coverage

- Big difference in tumor and normal coverage → lots of false positive indels
- Repetitive regions, e.g. TATATATATATATATATATA, are error prone
  - Insertion or deletion of one or multiple repeat units
- Some probability of an error happening
- High coverage → more errors
- Approximately same coverage in tumor and normal → errors will mostly happen with the same amount in both samples
- Much higher tumor coverage → much more errors, sometimes the difference will be significant, and the caller thinks it's a variant because it doesn't see the same thing in the normal

The central dogma

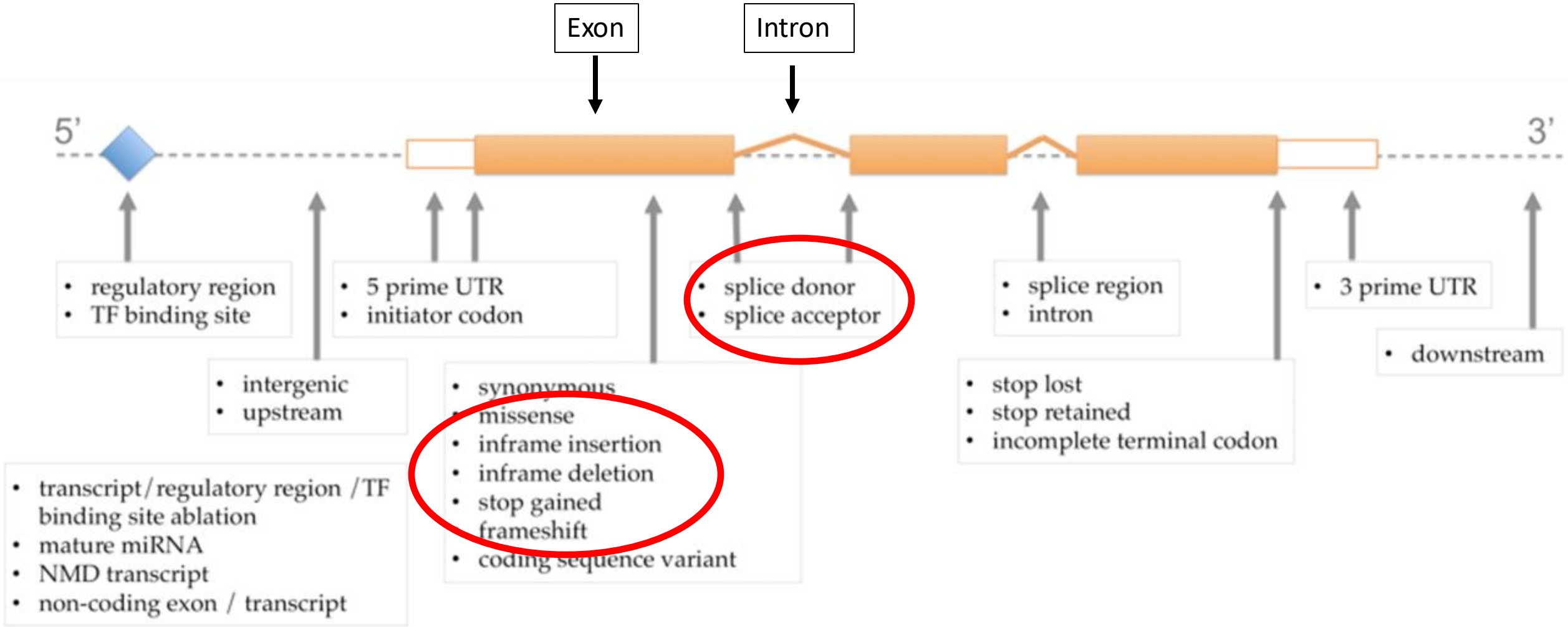


Somatic mutation calling – annotation

- What effects do the mutations have?
  - On proteins
  - On disease
  - On treatment alternatives

- VEP – Variant Effect Predictor
  - From Ensembl
  - “VEP determines the effect of your variants … on genes, transcripts, and **protein sequence**, as well as regulatory regions.” (https://www.ensembl.org/info/docs/tools/vep/index.html)

Somatic mutation calling – VEP annotation categories



Credit Griffith brothers' lecture #4

Mutation calling – file format

- VCF – variant call format
- Header – metadata describing different fields
- Main

```
#CHROM  POS        ID  REF  ALT  QUAL  FILTER  INFO
1       182712438  .   T    G    0     PASS    MVDK=1,1,1,1;NUM_TOOLS=4;SOMATIC;CSQ=G|intergenic_variant|MODIFIER||||||||||||||||1||||SNV|||||||||||||||||||||||||||||||||||||||||
2       178149405  .   CT   C    0     REJECT  MVDK=0,0,1,0;NUM_TOOLS=1;CSQ=-|non_coding_transcript_exon_variant|MODIFIER|AC074286.1|ENSG00000213963|Transcript|ENST00000397057|se
```

```
FORMAT                                                          NORMAL                                                                TUMOR
GT:DP4:CD4:refMQ:altMQ:refBQ:altBQ:refNM:altNM:fetSB:fetCD:zMQ:zBQ:MQ0:VAF  0/0:458,380,0,0:838,0,0,0:60:.:38.9952:.:3.68974:.:1.00:1.00:.:.:0:0         0/1:1580,1177,39
GT:DP4:CD4:refMQ:altMQ:refBQ:altBQ:refNM:altNM:fetSB:fetCD:zMQ:zBQ:MQ0:VAF  0/0:259,345,0,1:604,0,1,1:60:60:38.7517:18:4.35099:14:1.00:1.00:0:-1.52917:0:0.00165  0/1:1188,1475,21
```

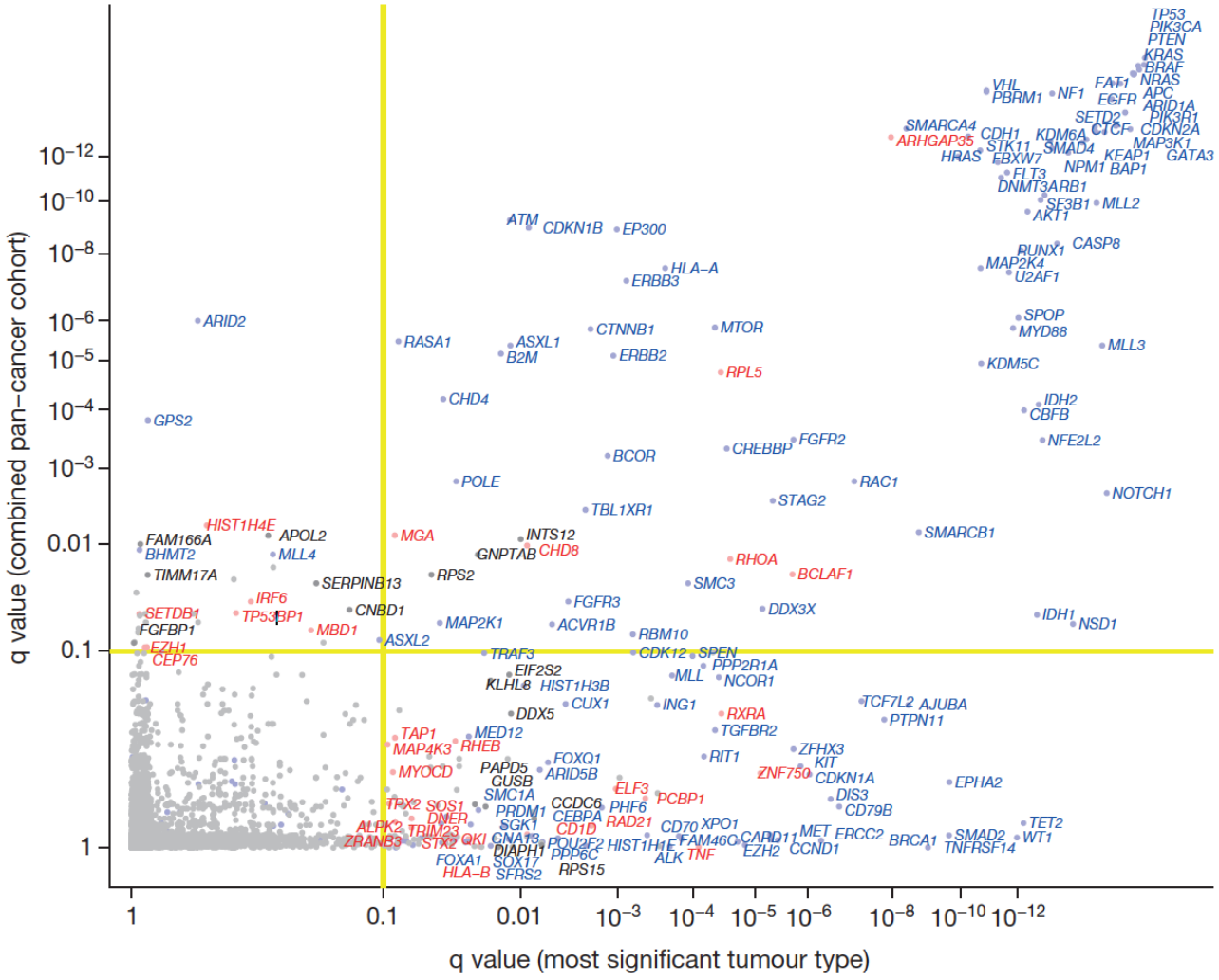
```
TUMOR
0/1:1580,1177,39,34:2756,1,73,0:60:60:41.9666:45.2055:3.19042:4.43836:0.55:1.00:0:1.45553:0:0.0258
0/1:1188,1475,21,27:2663,0,48,0:60:60.9804:41.4889:33.5098:4.05971:5.78431:1.00:1.00:1.20101:-3.31303:0:0
```

Mutation calling – manual curation

- Not all called variants are true – even after consensus
- If filters were sharpened to decrease number of false positives, we would instead miss some variants
- Some properties are hard to account for/discover programmatically
- Manual curation necessary
- Purpose: to identify true variants with impact on the protein structure

Different genes mutated in different cancers

- x-axis: how common in the tumor type where it's most common?
- y-axis: how common in all tumor types together?
- Above yellow line → candidate cancer gene
- Many genes that are candidate cancer genes when looking at a specific tumor type are not significantly mutated when looking at all cancer types together
- Lawrence, M. S., Stojanov, P., Mermel, C. H., Robinson, J. T., Garraway, L. A., Golub, T. R., Meyerson, M., Gabriel, S. B., Lander, E. S., & Getz, G. (2014). Discovery and saturation analysis of cancer genes across 21 tumour types. *Nature*, *505*(7484), 495–501. https://doi.org/10.1038/nature12912



Somatic variant calling – Questions?

Somatic mutation calling quiz

Mutational signatures

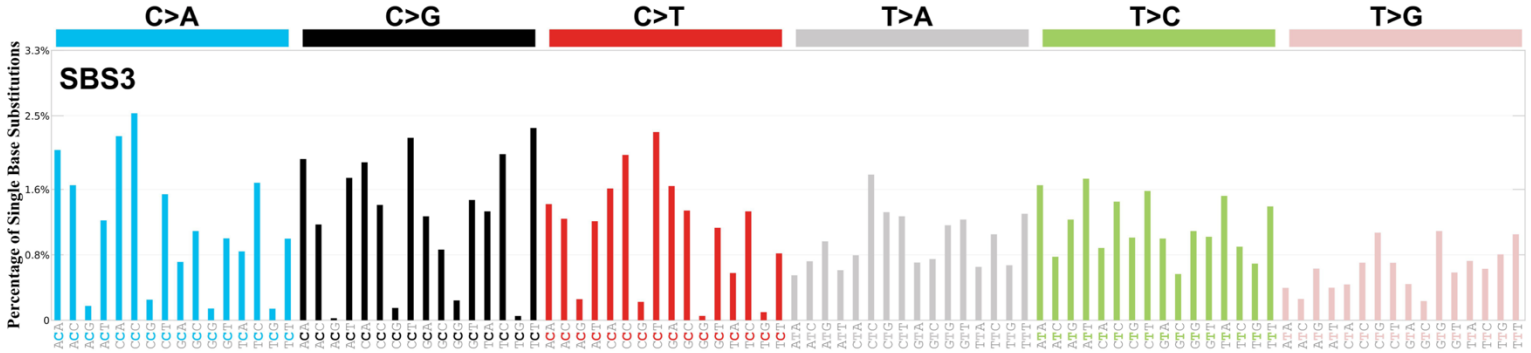
- Somatic mutations are caused by both exogenous and endogenous processes.
  - DNA repair, smoking, sunlight, ageing, chemotherapy etc.
- Mathematical methods have been developed to investigate mutation data to determine the mutational signatures from a cancer genome and its origin.

Statistically significant mutational signature	x	Cancer type association	x	Known explaining mutagenic process? Associated with treatment outcome?

The repertoire of mutational signatures in human cancer, Alexandrov et al., Nature 2020

Mutational signatures

- Single base substitutions (SBS), an example
  - Input data, possible mutations: C>A, C>G, C>T, T>A, T>C and T>G
- Account for 5’ and 3’ base, leads to 96 possibilities. (e.g. T**C**G > T**A**G)
- Account for transcribed or un-transcribed strand
  - 192 possibilities

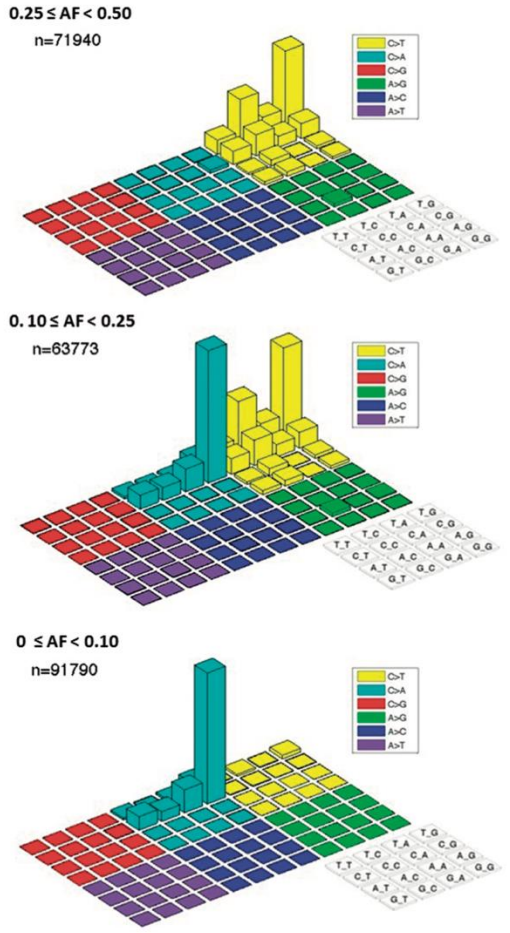


SBS3: signature associated with homologous recombination deficiency and sensitivity to parp inhibitors and carboplatin.

The repertoire of mutational signatures in human cancer, Alexandrov et al., Nature 2020

Mutational signatures and artefacts

- At the Broad Institute noise variants were detected in the TCGA whole exome data
- Another way of displaying the 96 possibilities
- Similarly done for
  - Indels
  - Double base substitutions
  - Vary flanking bases
  - Etc …



Low frequency artefactual mutations increase

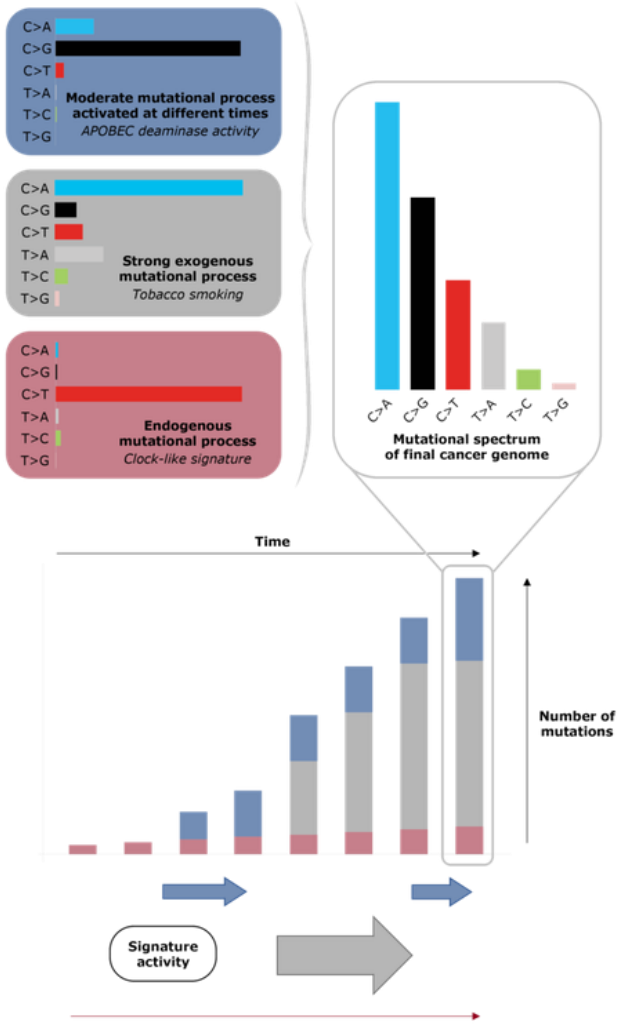
True signature

Artefact signature

Discovery and characterization of artifactual mutations in deep coverage targeted capture sequencing data due to oxidative DNA damage during sample preparation, Costello et al., NAR 2013

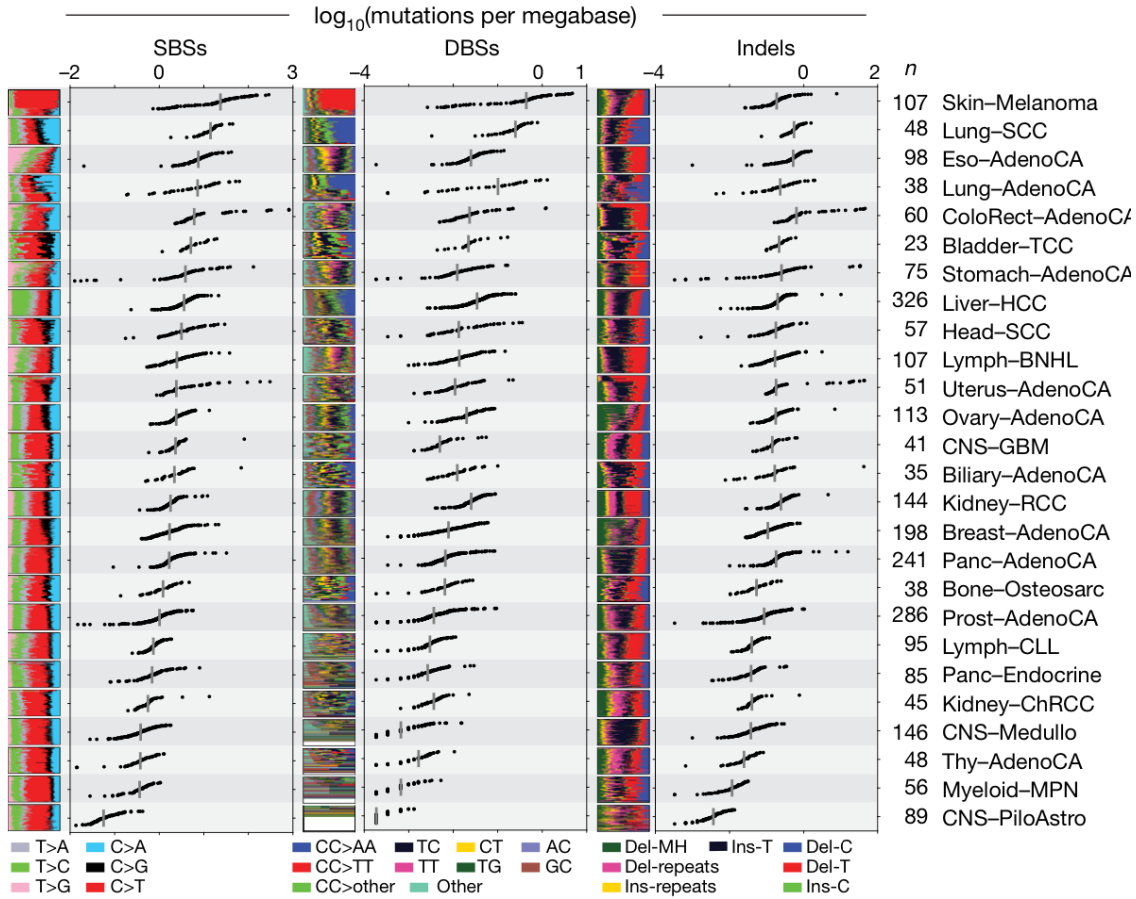
Mutational signatures

Identify the mutational causing processes for an individual cancer



https://cancer.sanger.ac.uk/signatures/

Reflected in the mutation rates of different cancers

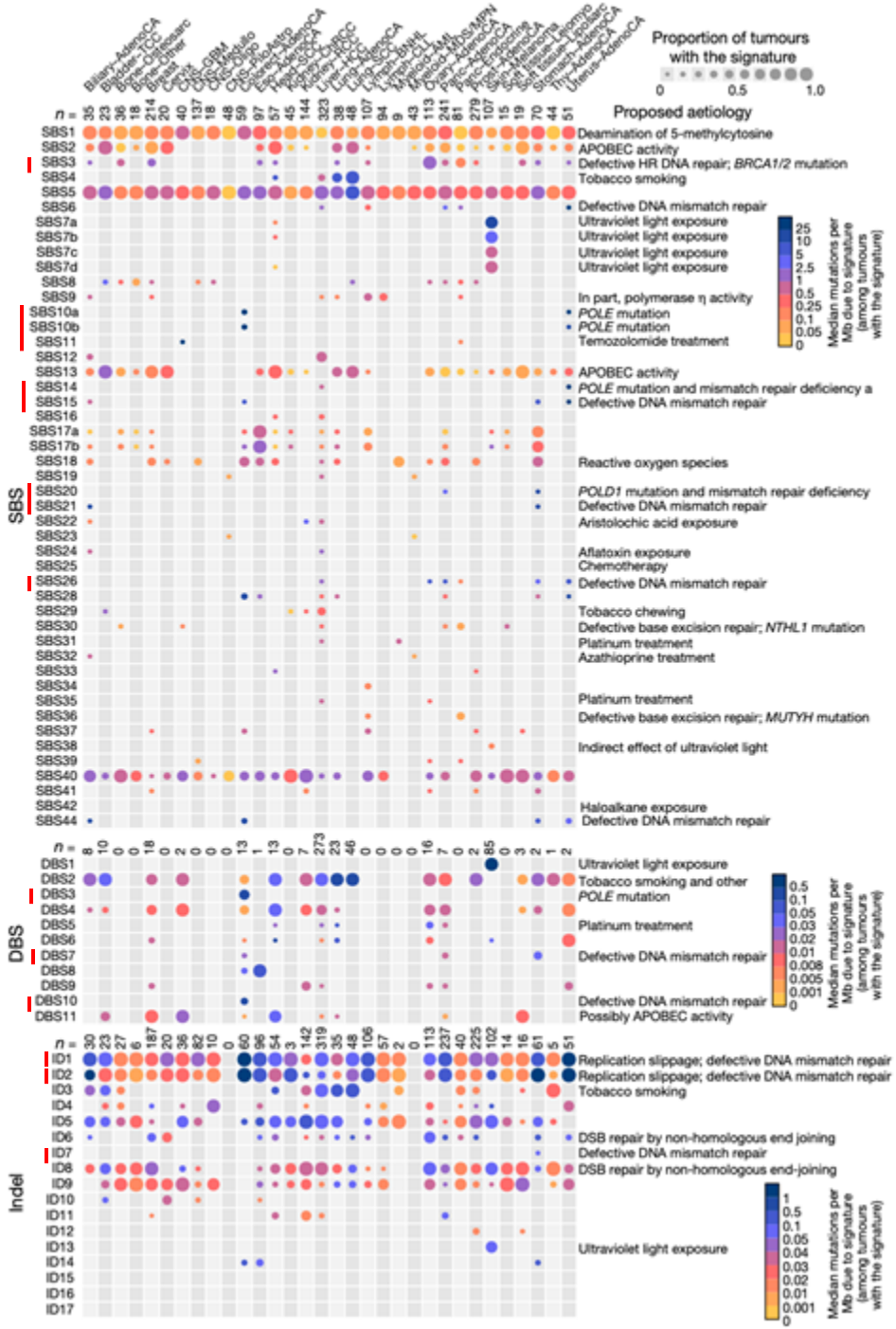


The repertoire of mutational signatures in human cancer, Alexandrov et al., Nature 2020

Mutational signatures

Treatment relevant mutational signatures, e.g. homologous recombination deficiency (SBS3).

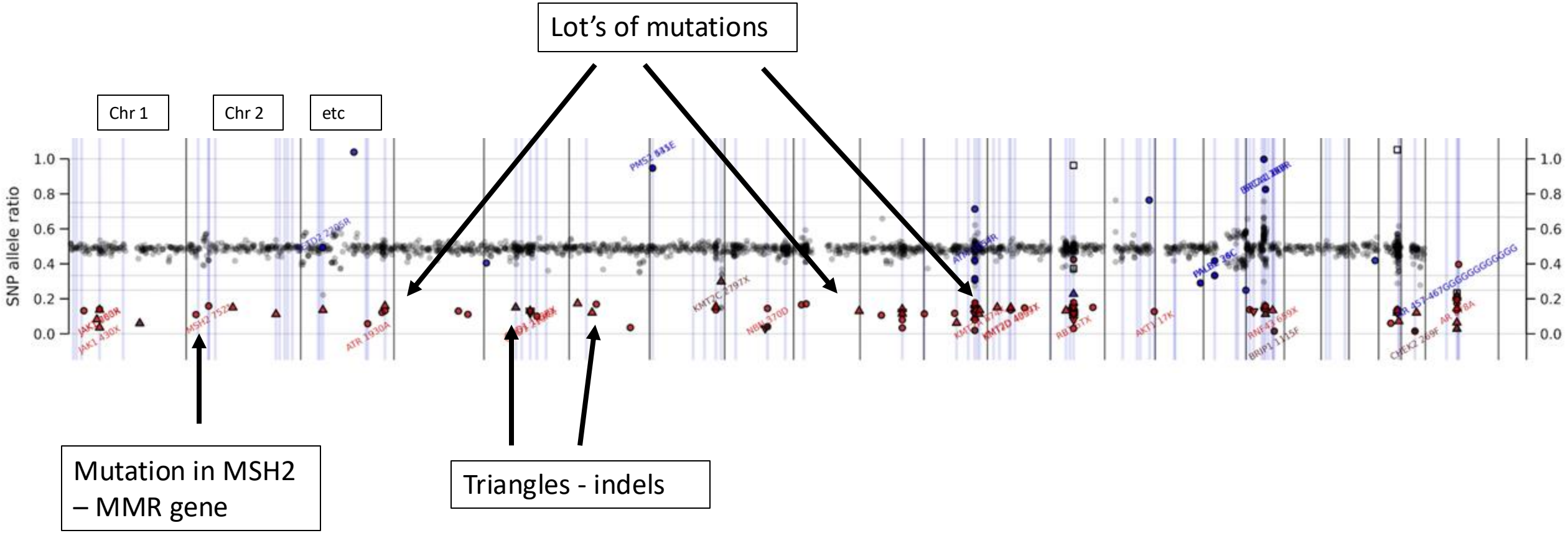
An argument for WGS/WES.
Panel sequencing is often 5x – 50x smaller than WES
Hard to robustly assess mutational signatures with fewer positions sequenced.



MSI – microsatellite instability

- Microsatellite: small repetitive sequence of the genome, e.g. TTTTT or TATATATA
- MSI: increased rate of insertion or deletion of repeated segments in tumor
  - → plenty of insertions and deletions
- Caused by dysfunctional mismatch repair (MMR) mechanism
  - Indels of repeated segments not corrected
- Multiple different mutational signatures recognized as MSI
- Levels:
  - MSI-H: microsatellite instability high
  - (MSI-L: microsatellite instability low)
  - MSS: microsatellite stable
- Many mutations → tumor cells may express lots of weird proteins on their surfaces
  - Immunotherapy may be effective

MSI – an example



MSI - tool

- mSINGS - MicroSatellite Instability detection by Next Generation Sequencing
- Given list of microsatellites, 63 in our designs
- Background control created by ~20 healthy donor samples
- Comparing number of repeats in each microsatellite locus in tumor sample to the same sites in background control
- Locus called as unstable if significantly different from control samples
- Fraction of unstable loci gives mSINGS score
  - Threshold for MSI-H: > ~0.2
  - Confirm by visual inspection of mutation plot

Hypermutation

- MSI is a type of hypermutation – phenotypes with highly increased levels of mutation frequency
- Another type is caused by defective DNA replication repair due to mutations in DNA polymerases
- This gives different mutational signature, with more SNVs than indels
  - SBS14
- Other types of hypermutation can be caused by environmental factors (e.g. UV light, smoking, chemotherapy), and are associated with other specific mutational signatures

CHIP - Clonal Hematopoiesis of Indeterminate Potential

- Sub-population of blood cells carrying the same mutation(s)
- Age-related
- Increased risk of blood cancer and cardiovascular disease
- Shows in germline DNA, often at VAFs ≠ 50% or 100%
- In germline from WBC
  - May show in cfDNA but not tissue tumor samples

Mutational signatures quiz

Credits

- Malachi Griffith, Obi Griffith, Zachary Skidmore, Huiming Xia
  - Lecture notes from the course “Introduction to bioinformatics for DNA and RNA sequence analysis (IBDR01)”, 29 October – 2 November, 2018
  - McDonell Genome Institute, Washington University of St Louis School of Medicine

More questions?