CAMBRIDGE Cancer CENTRE

Cambridge science is
changing the way we treat cancer

SNV Introduction and calling germline SNPs

Andy Lynch

CRUK CI

July 2016

UNIVERSITY OF CAMBRIDGE

CANCER RESEARCH UK

Cambridge University Hospitals NHS
NHS Foundation Trust

Outline

This morning's session:

- Introduction
- SNP calling
- SNPs and SNVs
- SNV-calling experiments
- Power
- VCF files

SNV Introduction and calling germline SNPs

Introduction

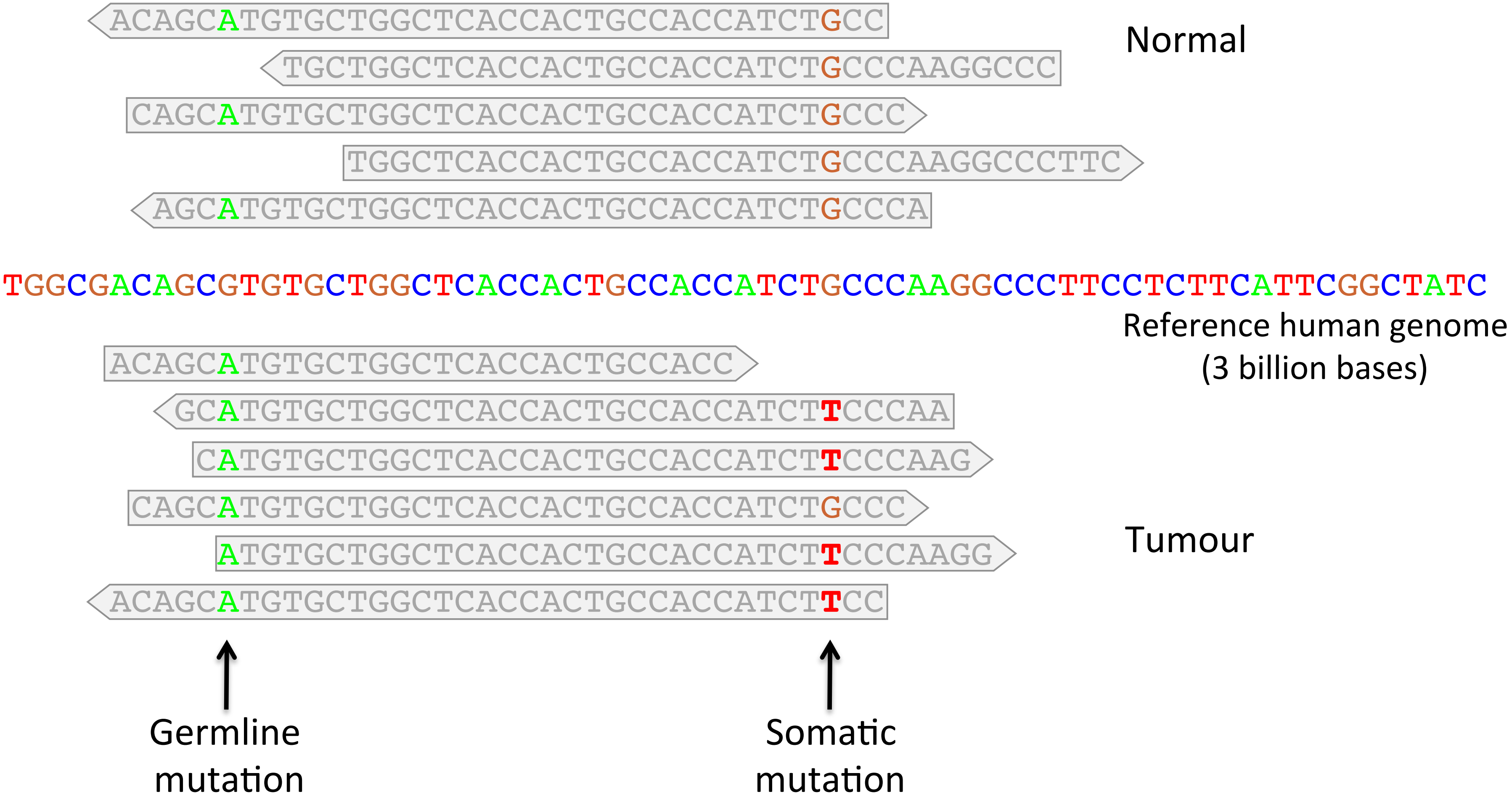
Mutations

Three types of small mutation we will consider

We'll start out by supposing that we are considering whole-genome sequencing data from matched tumour and normal samples.

- **Single-nucleotide polymorphisms (SNPs) [or Germline SNVs].** Single nucleotide 'mutations' that were inherited from the patient's parents. We expect to see >3,000,000 per patient
- **Single-nucleotide variants (SNVs).** Single nucleotide mutations that have been acquired since birth. We expect to see from e.g. 100 (pilocytic astrocytoma) to >50,000 (lung adenocarcinoma) per patient (Khurana et al. 2016)
- **Indels.**

Cartoon idea



ACAGCATGTGCTGGCTCACCACTGCCACCATCTGCC
TGCTGGCTCACCACTGCCACCATCTGCCCAAGGCCC
CAGCATGTGCTGGCTCACCACTGCCACCATCTGCCC
TGGCTCACCACTGCCACCATCTGCCCAAGGCCCTTC
AGCATGTGCTGGCTCACCACTGCCACCATCTGCCCA

Normal

TGGCGACAGCGTGTGCTGGCTCACCACTGCCACCATCTGCCCAAGGCCCTTCCTCTTCATTCGGCTATC

Reference human genome
(3 billion bases)

ACAGCATGTGCTGGCTCACCACTGCCACC
GCATGTGCTGGCTCACCACTGCCACCATCTTCCCAA
CATGTGCTGGCTCACCACTGCCACCATCTTCCCAAG
CAGCATGTGCTGGCTCACCACTGCCACCATCTGCCC
ATGTGCTGGCTCACCACTGCCACCATCTTCCCAAGG
ACAGCATGTGCTGGCTCACCACTGCCACCATCTTCC

Tumour

Germline
mutation

Somatic
mutation

Why doesn’t this always work

If you ‘call’ an SNV, there are four possibilities

- **It really is an SNV:** Yippee etc.

- **It is actually a SNP:** This is what we will be talking about now

- **It is an artefact:** This would usually be the result of consistent alignment or sequencing errors, and Matt will be talking about the filtering methods we apply to avoid these.

- **It is an artefact caused by a different somatic variant:** This could be a structural variant or an indel (or even an SNV) causing localized and consistent alignment errors. I separate these out from the previous category because they are still interesting. We don’t for example wish to exclude a call entirely because we can’t tell whether it is an SNV or an indel if it is disrupting e.g. KRAS in either case.

SNV Introduction and calling germline SNPs

SNP calling

Identifying SNPs

Four reasons for identifying SNPs

Let’s be clear at this point that there are multiple different reasons why we would wish to call SNPs in a sample, and the sensitivity/specificity trade off we want will depend on that purpose (a lesson we might keep in mind when it comes to filtering SNVs later)

- **We want to perform germline analyses (GWAS, fine-mapping, etc.).** In this case, we really want to have reliable calls, but will presumably be using cross-population linkage info to supplement our results and won’t want to sacrifice too much sensitivity for specificity.

- **We want to eliminate false SNV calls*.** Suppose we expect to see 3,000,000 SNPs and 3,000 SNVs – only 0.1% of our SNPs need to get through filters for them to outnumber the called SNVs. Since we probably wouldn’t trust an SNV called in the same location as an artefactual SNP, we can sacrifice specificity to enhance sensitivity

- **We want to call copy number states.** As Oscar will show later, SNPs (in particular heterozygous SNPs) are hugely valuable for estimating copy number states. The statistical methods that are used for calling copy number are not always robust to artefactual SNP calls, and we don’t need many to get a good profile so here we would want to sacrifice sensitivity to enhance specificity

- **We want to filter nearby SNV calls.** An SNV will only arise in one allele, so if there is a heterozygous SNP in the vicinity we may be able to check this. Again specificity is probably more important.

Tools to identify SNPs

I'll just mention three

All of these assume a pre-processing of alignment, duplicate marking etc.

- **GATK.** (Broad) HaplotypeCaller + filtering. Benefits from processing multiple normal samples at once. Also calls indels.

- **Platypus.** (Rimmer 2014, WTHGC) Simple to use, good performance. Can deal with multiple related samples. Also calls MNPs, indels, rearrangements etc. (Also want to give a shout for CRUK CI's multiSNV)

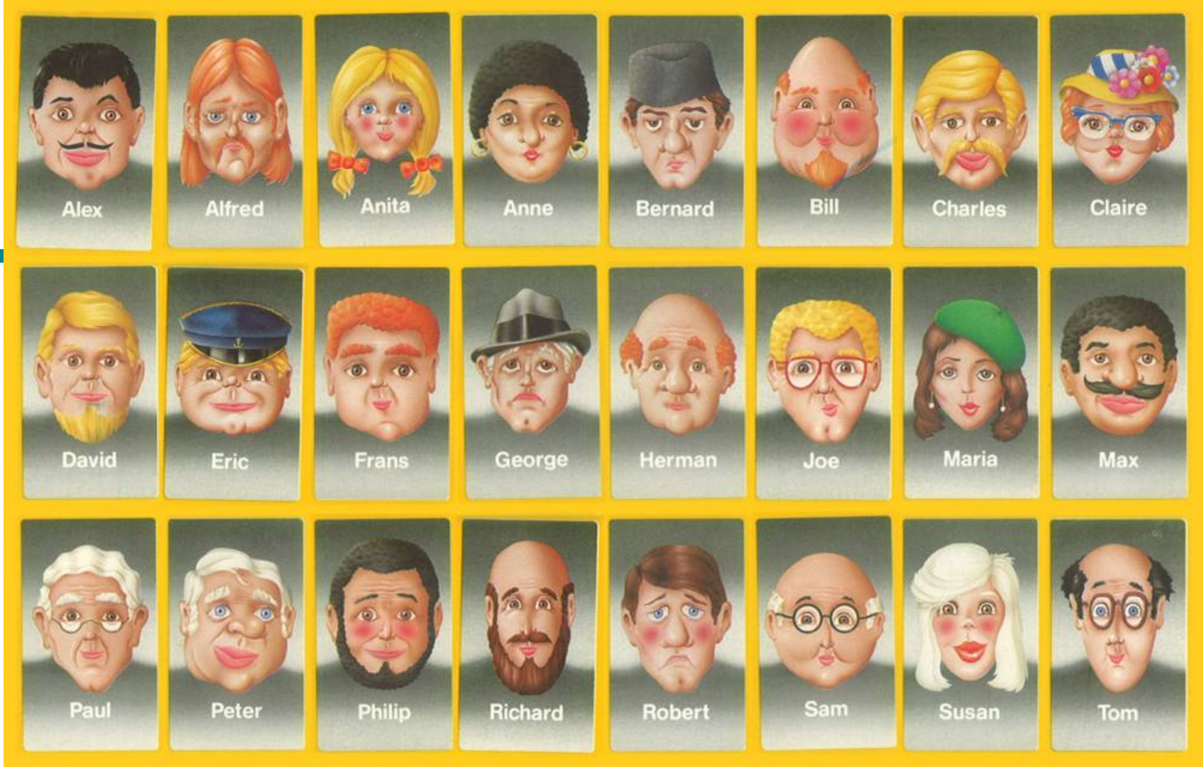
- **FreeBayes.** (Garrisson 2012, Boston College) The one we will be using today. Again would benefit from processing multiple samples at once. Works with haplotype blocks, but requires well-mapping regions at each end. Can return other mutation types also. Possibly the easiest to generalize to other species.

Bayes Theorem

P(Glasses) = 5/24
P(Bald) = 6/24 *

P(Glasses|Bald) = 2/6
P(Bald|Glasses) = 2/5

P(Bald & Glasses) = 2/24



P(Bald & Glasses) = P(Glasses)P(Bald|Glasses) = P(Bald)P(Glasses|Bald)

P(Bald|Glasses)P(Glasses) = P(Glasses|Bald)P(Bald)

$$P(\text{Bald}|\text{Glasses}) = \frac{P(\text{Glasses}|\text{Bald})P(\text{Bald})}{P(\text{Glasses})}$$

$$P(\text{Bald}|\text{Glasses}) \propto P(\text{Glasses}|\text{Bald})P(\text{Bald})$$

$$P(A|B) \propto P(B|A)P(A)$$

*George is fooling nobody

FreeBayes

FreeBayes (and most other tools we discuss) apply this rule to genotyping

$P(A|B) \propto P(B|A)P(A)$

$P(\text{Genotype}|\text{Data}) \propto P(\text{Data}|\text{Genotype})P(\text{Genotype})$

If we have low coverage data then the **P(Genotype)** part will be more influential. If we have high coverage data, then the information derived from a large cohort will be less important.

Remember that these calculations are not being done a locus at a time, but in haplotype blocks. This is important for filtering and downstream interpretation.

Haplotype Blocks

- **Blocks/Clusters exist.** We typically sequence ~100bp reads in pairs from fragments that are ~500bp wide. Germline variants cluster such that 80% lie within 60 bases of another.

- **Blocks/Clusters help.** If we have several variants in a region, they should appear in complementary reads. This helps separate out artefacts and recover variants that might otherwise have been missed.

  Having the phasing information will be advantageous for copy number calling and SNV filtering later.

- **Blocks/Clusters are important.** Recall these are germline variants. If they cluster together, we need to consider them together when we annotate their effects.

  Two variants in a gene? We need to know which combinations we are seeing!

  Eight variants?

Source: An introduction to (small) variant detection - 2013 - Erik Garrison

SNV Introduction and calling germline SNPs

SNVs and SNPs

Discriminating SNVs and SNPs

- **Subtraction.** Early studies called variants in the tumour sample and in the normal sample, and subtracted one from the other.

- **Reverend Bayes.** Most modern approaches extend the types of method we saw FreeBayes use for germline variants.

Why doesn’t this work perfectly?

The germline sample doesn’t always divide neatly into sites with zero variation and sites with 50% variation.

We may want to allow for the possibility of contamination of the normal sample

Should we change our approach depending on the type of normal sample?

Other resources

Can anything else help separate out SNVs and SNPs

- **1000 Genomes project/dbSNP.** We have good catalogues of common variants. Can filter against these

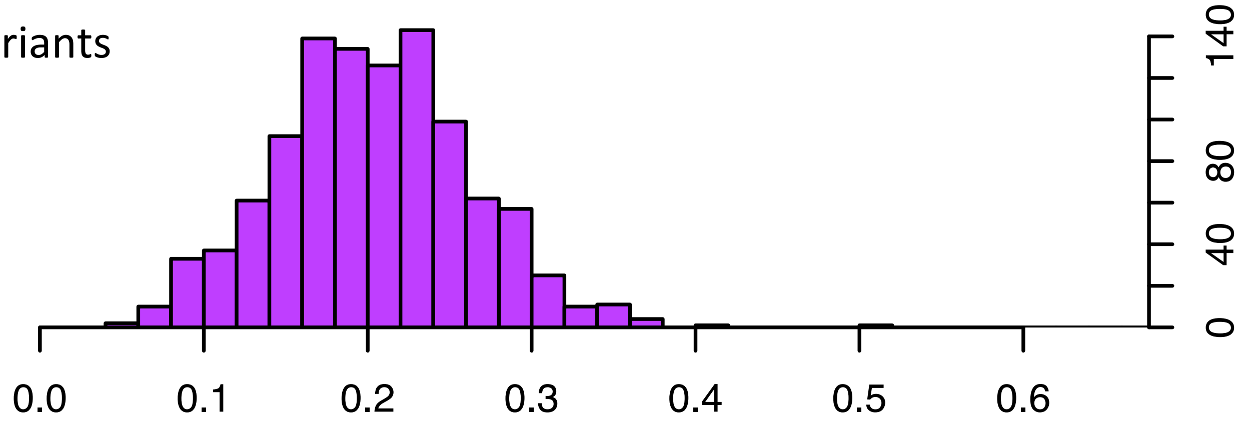
WARNING: dpSNP includes some somatic variants

- **Impure tumour samples.** If you are looking at a cell line, or an astonishingly pure tissue sample, SNVs and SNPs look the same. Otherwise they show different allele frequencies.

WARNING: Could be a field effect/early somatic event or somatic event in the normal sample

Allele frequencies of all variants that remain in a diploid heterozygous state.

Tumour has approx 40% cellularity.



Do we need to call germline variants?

Sacrilege

Suppose we had only sequenced tumour samples. We could filter out all common variants and some based on allele frequency. What would be the effect?

- **False positives.** SNVs that are really SNPs. Not common SNPs though, so if they target an area or feature of the genome that would be noteworthy.

Not clear that they would harm a cohort analysis, and for the individual is a rare germline variant more or less likely to have a role in the disease than a somatic variant?

- **False negatives.** SNVs not called because they mimic a common variant. Are they credible actors in the disease? Does it matter that we missed them?

Would these not have come up in a GWAS?

Of course, there are good reasons for performing sequencing of matched normals, but if paying for it, be sure to know where your value is.

SNV Introduction and calling germline SNPs

SNV-calling experiments

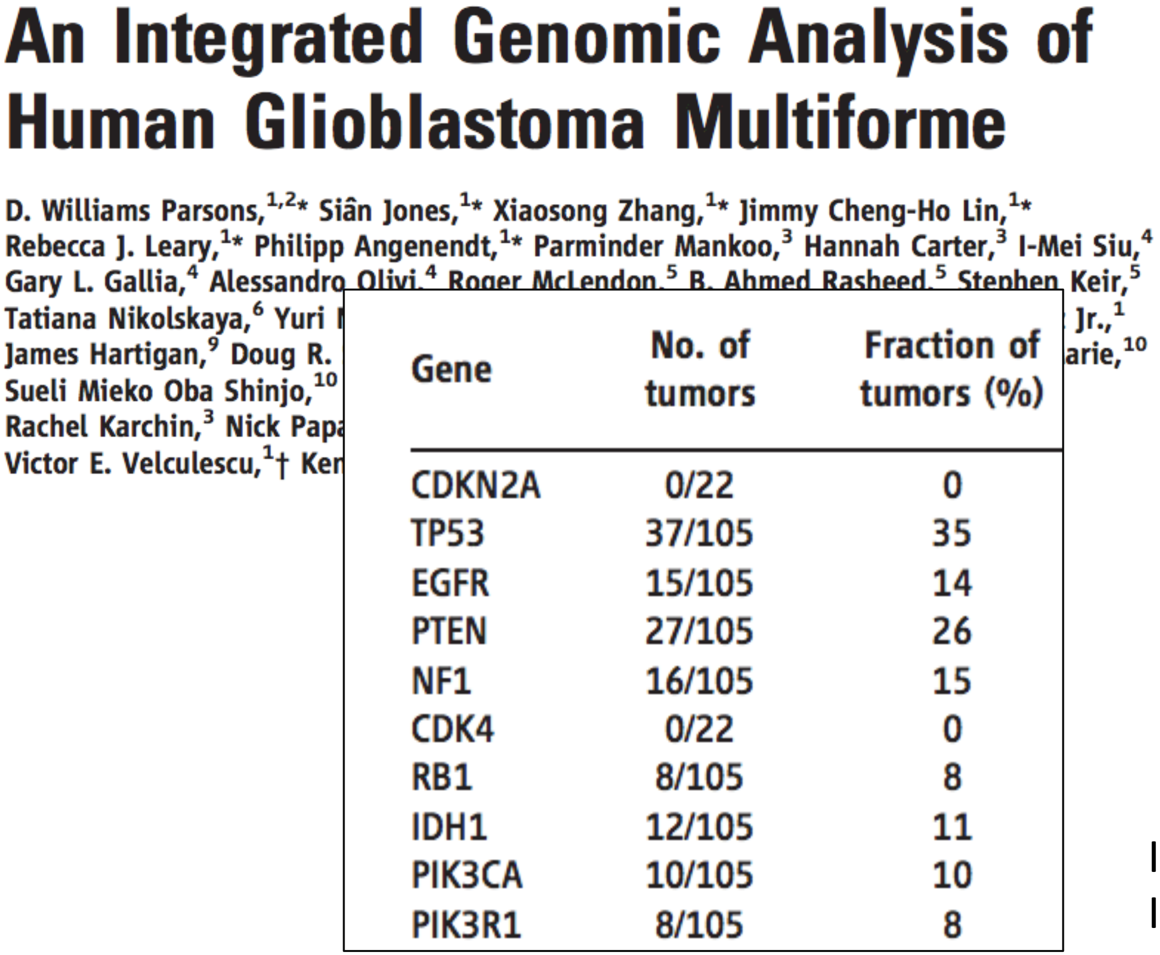
Purposes of experiment

Mwenifumbo et al. 2013 identify four main purposes for cancer sequencing

The detail of

- **Discovering driver mutations**
- **Identifying somatic mutational signatures**
- **Characterizing clonal evolution**
- **Personalizing treatment**

Discovering Driver Mutations examples



An Integrated Genomic Analysis of Human Glioblastoma Multiforme

D. Williams Parsons,[1,2*] Siân Jones,[1*] Xiaosong Zhang,[1*] Jimmy Cheng-Ho Lin,[1*]
Rebecca J. Leary,[1*] Philipp Angenendt,[1*] Parminder Mankoo,[3] Hannah Carter,[3] I-Mei Siu,[4]
Gary L. Gallia,[4] Alessandro Olivi,[4] Roger McLendon,[5] B. Ahmed Rasheed,[5] Stephen Keir,[5]
Tatiana Nikolskaya,[6] Yuri I Jr.,[1]
James Hartigan,[9] Doug R. arie,[10]
Sueli Mieko Oba Shinjo,[10]
Rachel Karchin,[3] Nick Papa
Victor E. Velculescu,[1]† Ken

Gene	No. of tumors	Fraction of tumors (%)
CDKN2A	0/22	0
TP53	37/105	35
EGFR	15/105	14
PTEN	27/105	26
NF1	16/105	15
CDK4	0/22	0
RB1	8/105	8
IDH1	12/105	11
PIK3CA	10/105	10
PIK3R1	8/105	8

In 2008, Parsons et al. identified IDH1 as a novel driver gene in GBM

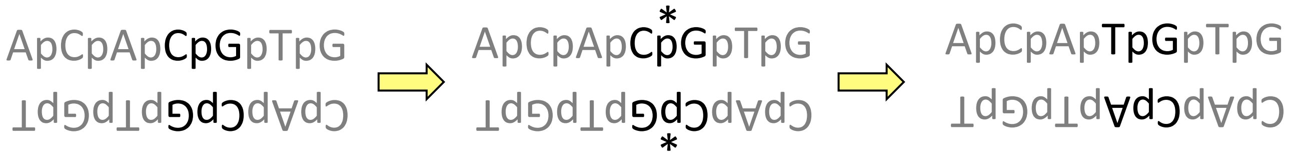
Mutational signatures

There are processes that lead to distinct patterns of SNVs

- **Causes.** Generally classifiable as errors in DNA replication or the failure of DNA repair mechanisms to accurately counter the presence of mutagens

- **Mutational contexts.** Typically the signatures consider 96 contexts = 4 preceding bases x 6 mutations (C>A, C>G, C>T, T>A, T>C, T>G) x 4 following bases, although larger/different contexts can be defined.

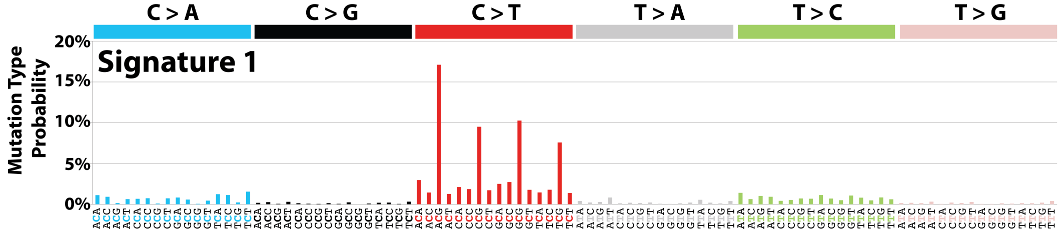
- **Known associations.** Signatures associated with e.g. tobacco have been known for some time (see e.g. Pfeifer *et al*. 2004). One well known ‘aging’ signature sees methylated CpG islands de-aminate to a TG sequence the complementary pair of which may be corrected to “CA”

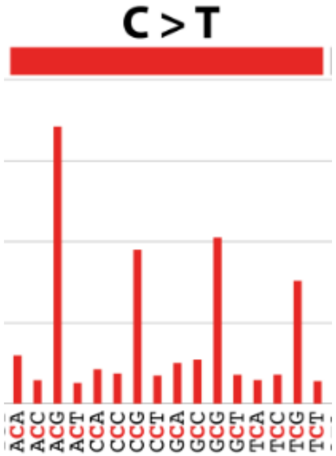


ApCpApCpGpTpG
CpApCpGpTpGpT

*
ApCpApCpGpTpG
CpApCpGpTpGpT
*

ApCpApTpGpTpG
CpApCpApTpGpT





Signature software

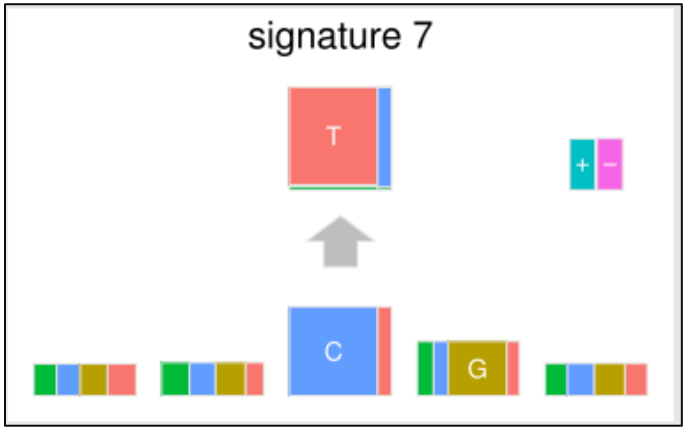
Signatures detected by considering large numbers of contrasting samples

If two signatures always act together, they might be difficult to separate.

- **Alexandrov et al.** The original option. De facto standard? Identifies signatures via non-negative matrix factorization.

- **Gehring et al.** The SomaticSignatures bioconductor package. Enables easy, transparent, application of Alexandrov's methods. Also allows greater flexibility in terms of the matrix factorization

- **Shiraishi et al.** The pmsignature R package. A different model and representation, but possibly more robust. Perhaps harder to interpret (?harder to misinterpret?), easier to extend to a wider context.



Two things to think about re. signatures

- **True counts or normalize to genome frequencies?** Alexandrov et al. work from the counts, but present results normalized for genome frequencies. Shiraishi et al don't seem to normalize (I may have missed it). Gehring et al. offer a normalize option in their plots.

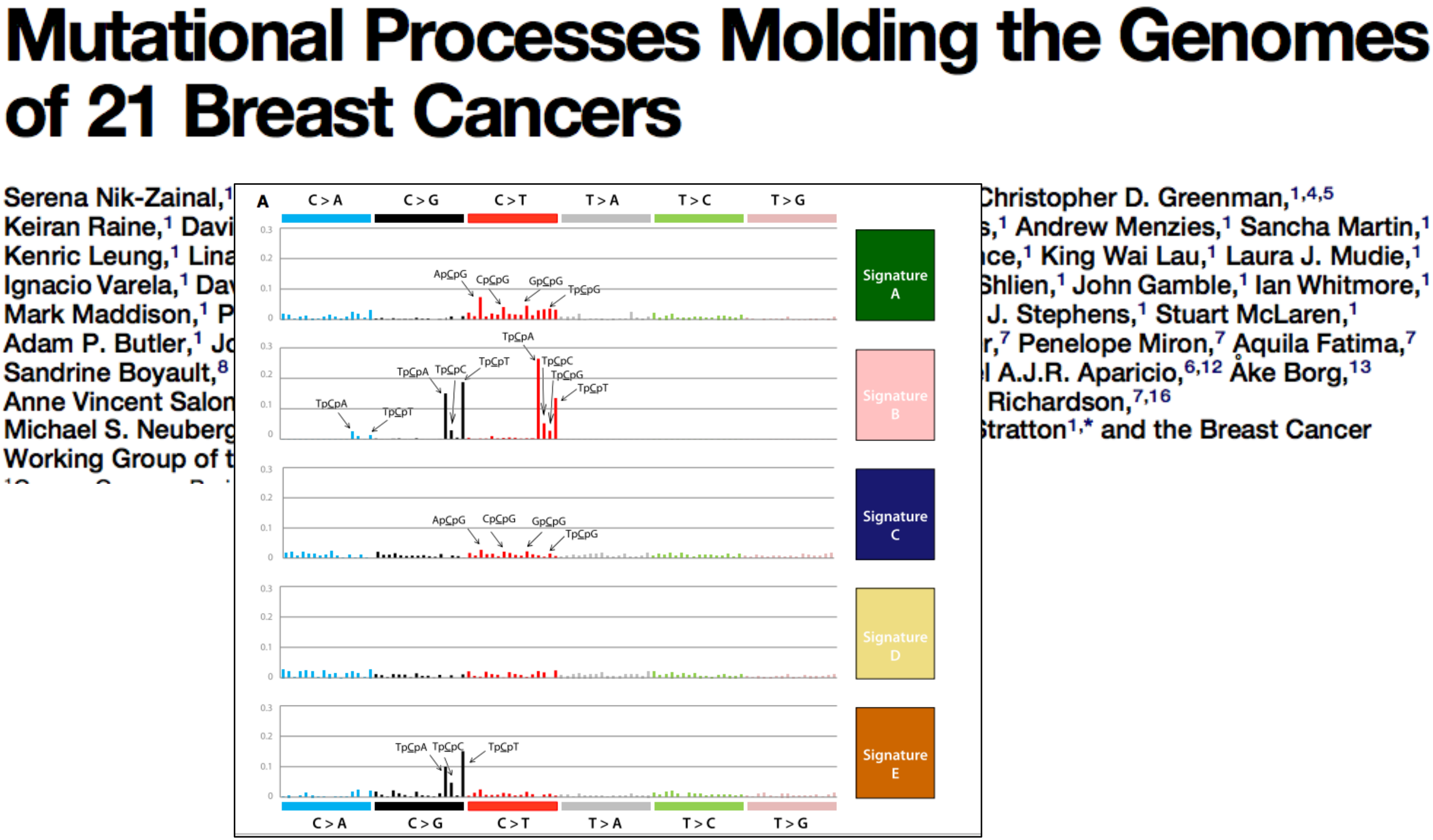
Normalization may be important

a) If combining exomes and genomes (even aneuploid tumours?)

b) If we think that the rate limiting step for mutational signatures is the availability of nucleotides in the correct context.

- **Strand.** We usually have six identifiable mutations as C>T is equivalent to G>A etc. In a gene region, we can distinguish the C and G as one will be transcribed and one won't (in the simple story). Therefore, especially if considering exome data, we may want to have 12 mutations. Shiraishi et al offer this.

Mutational signatures examples

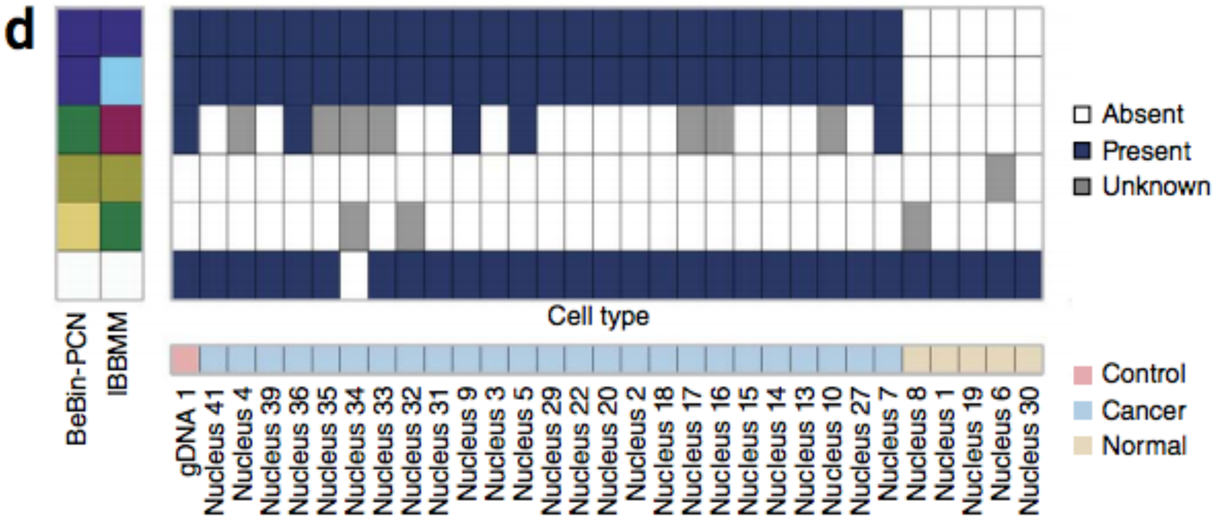


Mutational Processes Molding the Genomes of 21 Breast Cancers

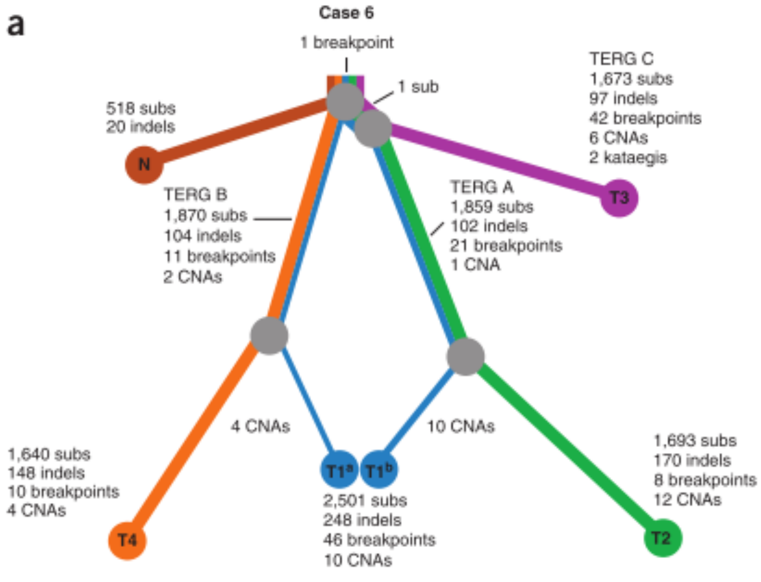
Serena Nik-Zainal,[1]
Keiran Raine,[1] Davi
Kenric Leung,[1] Lina
Ignacio Varela,[1] Day
Mark Maddison,[1] P
Adam P. Butler,[1] Jo
Sandrine Boyault,[8]
Anne Vincent Salon
Michael S. Neuberg
Working Group of t

Christopher D. Greenman,[1,4,5]
s,[1] Andrew Menzies,[1] Sancha Martin,[1]
ce,[1] King Wai Lau,[1] Laura J. Mudie,[1]
Shlien,[1] John Gamble,[1] Ian Whitmore,[1]
J. Stephens,[1] Stuart McLaren,[1]
r,[7] Penelope Miron,[7] Aquila Fatima,[7]
l A.J.R. Aparicio,[6,12] Åke Borg,[13]
Richardson,[7,16]
tratton[1,*] and the Breast Cancer

Evolution/Inferring clone examples

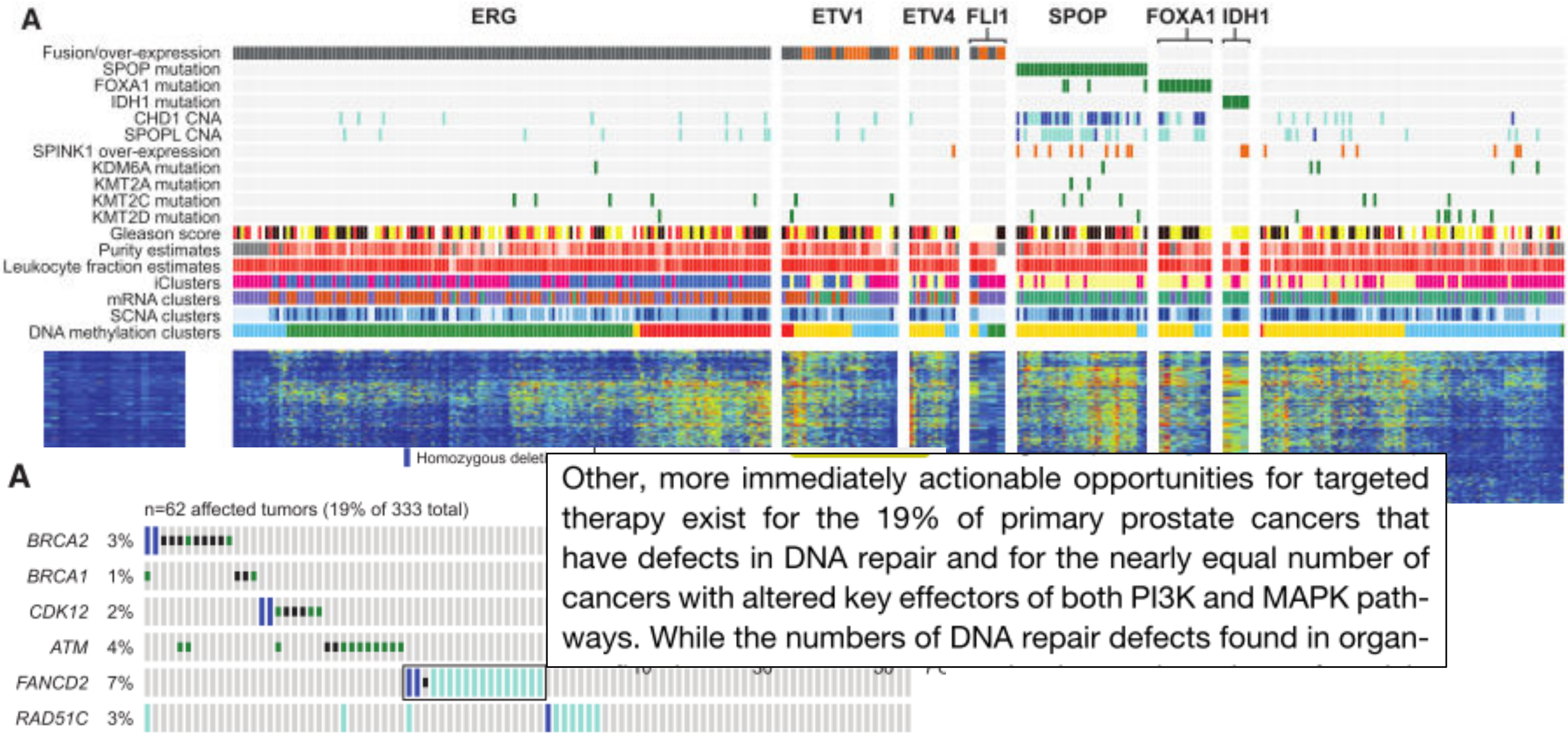


Roth et al. 2014 use PyClone to identify clusters of mutations, and annotate single cells accordingly



Cooper et al. 2015 use unique and shared SNVs to define the relationships between samples taken from the same prostate

Personalized treatment examples



Other, more immediately actionable opportunities for targeted therapy exist for the 19% of primary prostate cancers that have defects in DNA repair and for the nearly equal number of cancers with altered key effectors of both PI3K and MAPK pathways. While the numbers of DNA repair defects found in organ-

Abeshouse 2015

Design of experiment

	More focus on fewer samples	Balance depth and copy number	More samples less sensitivity for each individaul
Discovering Driver mutations			
Identifying Signatures			
Clonal evolution			
Personalized treatment			

SNV Introduction and calling germline SNPs

Power

Power and things that affect power – single SNV

Sequencing factors

- **Tumour Depth.** More sequencing = more reads supporting mutations

- **Normal Depth.** Assuming 2000000 heterozygous sites, Only ~5000 would be missed at 20x coverage. Perhaps as few as 250 wouldn't be filtered with dbSNP. 30x coverage should deal with these, but the extra 10x might offer even more to the tumour.

SNV factors

- **Copies in a cell.** e.g. if there are two copies of a mutant allele in each cell, we will see twice the number of reads as if there were only one.

- **Somatic copy number/Germline copy number.** Doesn't directly affect power to call variant, but allele frequencies are affected, so the ability to discriminate between SNV and SNP may be changed

Sample factors

- **Cellularity.** High cellularity means that more of the sequencing is spent on the tumour, and so the average number of reads supporting a mutation is increased.

- **Tumour ploidy.** The higher the average ploidy, the more genome there is to share DNA amongst. The average number of reads supporting a 'one-copy' mutation is decreased

Power and things that affect power – driver genes

Factors

- **Population frequency.** TP53 is mutated in nearly all ovarian cancer – easy to detect. IDH1 is mutated in 1% of prostate cancers – much less power.

- **Sample size.** The more samples we look at, the more that will be mutated in the gene of interest.

- **Power to detect an SNV.** If we are missing too many SNVs, we'll find it hard to spot a recurrently mutated gene. We have just seen the factors that affect this.

Power isn't linear in sequencing depth, so to a degree we should favour more samples at lower individual power.

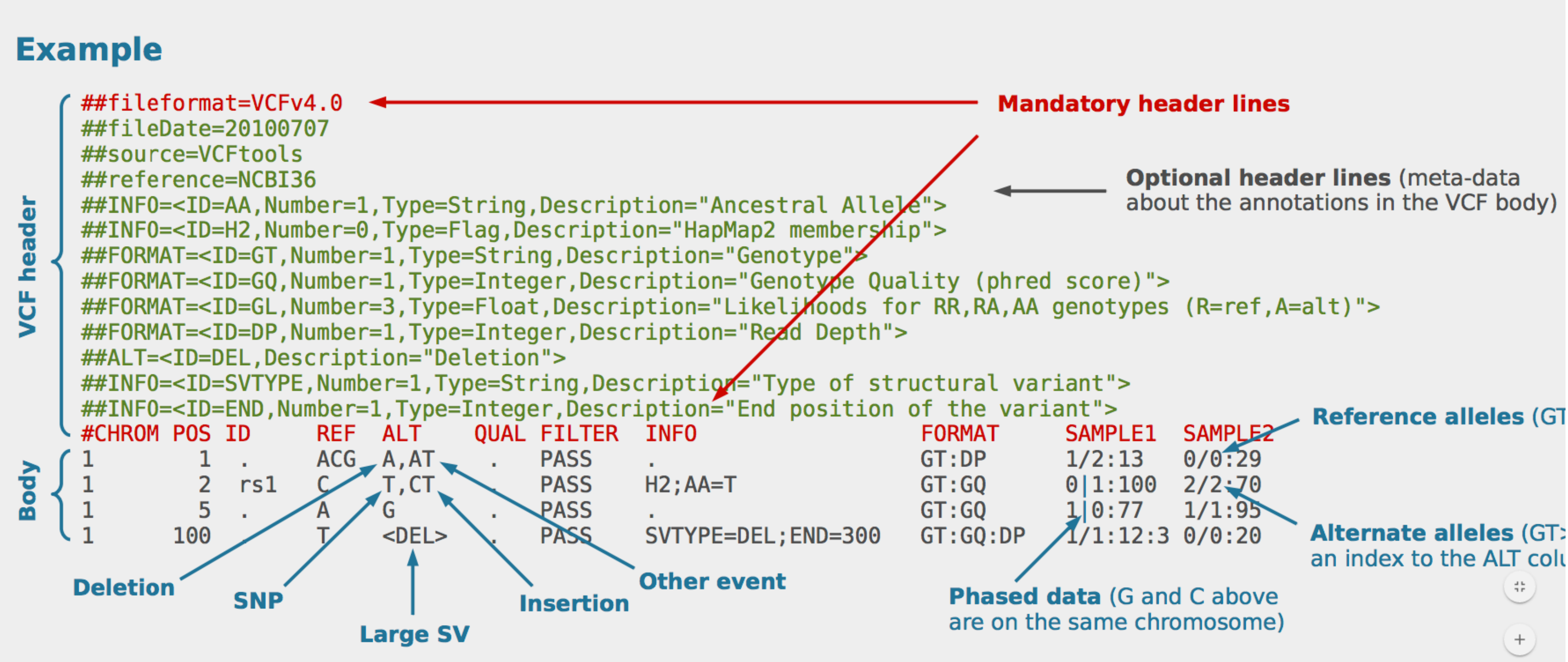
- **Mutation rate.** Cancers with highly mutated genomes will see genes affected by higher numbers of hits by chance – and so we need to see more affected samples for any truly affected gene

SNV Introduction and calling germline SNPs

VCF files

What information do we record?

VCF files



Example

VCF header

```
##fileformat=VCFv4.0
##fileDate=20100707
##source=VCFtools
##reference=NCBI36
##INFO=<ID=AA,Number=1,Type=String,Description="Ancestral Allele">
##INFO=<ID=H2,Number=0,Type=Flag,Description="HapMap2 membership">
##FORMAT=<ID=GT,Number=1,Type=String,Description="Genotype">
##FORMAT=<ID=GQ,Number=1,Type=Integer,Description="Genotype Quality (phred score)">
##FORMAT=<ID=GL,Number=3,Type=Float,Description="Likelihoods for RR,RA,AA genotypes (R=ref,A=alt)">
##FORMAT=<ID=DP,Number=1,Type=Integer,Description="Read Depth">
##ALT=<ID=DEL,Description="Deletion">
##INFO=<ID=SVTYPE,Number=1,Type=String,Description="Type of structural variant">
##INFO=<ID=END,Number=1,Type=Integer,Description="End position of the variant">
#CHROM POS ID     REF  ALT    QUAL FILTER  INFO                 FORMAT     SAMPLE1  SAMPLE2
1        1  .     ACG  A,AT    .   PASS    .                    GT:DP      1/2:13   0/0:29
1        2  rs1   C    T,CT    .   PASS    H2;AA=T              GT:GQ      0|1:100  2/2:70
1        5  .     A    G       .   PASS    .                    GT:GQ      1|0:77   1/1:95
1      100  .     T    <DEL>   .   PASS    SVTYPE=DEL;END=300   GT:GQ:DP   1/1:12:3 0/0:20
```

Body

Mandatory header lines

Optional header lines (meta-data about the annotations in the VCF body)

Reference alleles (GT

Alternate alleles (GT:
an index to the ALT colu

Deletion

SNP

Large SV

Insertion

Other event

Phased data (G and C above are on the same chromosome)

taken from http://vcftools.sourceforge.net/VCF-poster.pdf

Do we want a single set of calls?

Different tasks benefit from different designs, so do we want different lists of SNVs for different tasks?

- **Possibly best to have a master list, but apply different selections for different tasks.**

Later

Coming up after copy numbers

The detail of

- **Calling SNVs**
- **Filtering SNVs**
- **Using SNVs**
- **Misc.**

References

- Khurana E. et al. (2016). Role of non-coding sequence variants in cancer. Nat Rev Genet, 17(2), 93–108.
- Garrison E, Marth G. Haplotype-based variant detection from short-read sequencing. arXiv preprint arXiv:1207.3907 [q-bio.GN] 2012
- Rimmer A. et al. (2014). Integrating mapping-, assembly- and haplotype-based approaches for calling variants in clinical sequencing applications. Nature Genetics, 46(8), 912–918.
- Mwenifumbo J. C. & Marra M. A. (2013). Cancer genome-sequencing study design. Nature Publishing Group, 14(5), 321–332.
- Parsons D. W. (2008). An Integrated Genomic Analysis of Human Glioblastoma Multiforme. Science, 321(5897), 1807–1812.
- Pfeifer G. P. (2002). Tobacco smoke carcinogens, DNA damage and p53 mutations in smoking-associated cancers. Oncogene, 21(48), 7435–7451.
- Alexandrov L. B. et al. (2013). Signatures of mutational processes in human cancer. Nature, 500, 415–21.
- Gehring J. S. et al. (2015). SomaticSignatures: Inferring mutational signatures from single-nucleotide variants. Bioinformatics, 31(22), 3673–3675.
- Shiraishi, Y. et al. (2015). A simple model-based approach to inferring and visualizing cancer mutation signatures, 1–22.

References

- Nik-Zainal S. et al. (2012). Mutational processes molding the genomes of 21 breast cancers. Cell, 149(5), 979–993.
- Roth A. et al. (2014). PyClone: statistical inference of clonal population structure in cancer. Nature Methods, 11(4), 396–398.
- Cooper C. S. et al. (2015). Analysis of the genetic phylogeny of multifocal prostate cancer identifies multiple independent clonal expansions in neoplastic and morphologically normal prostate tissue. Nature Genetics, 6.
- Abeshouse A. et al. (2015). The Molecular Taxonomy of Primary Prostate Cancer. Cell, 163(4), 1011–1025.