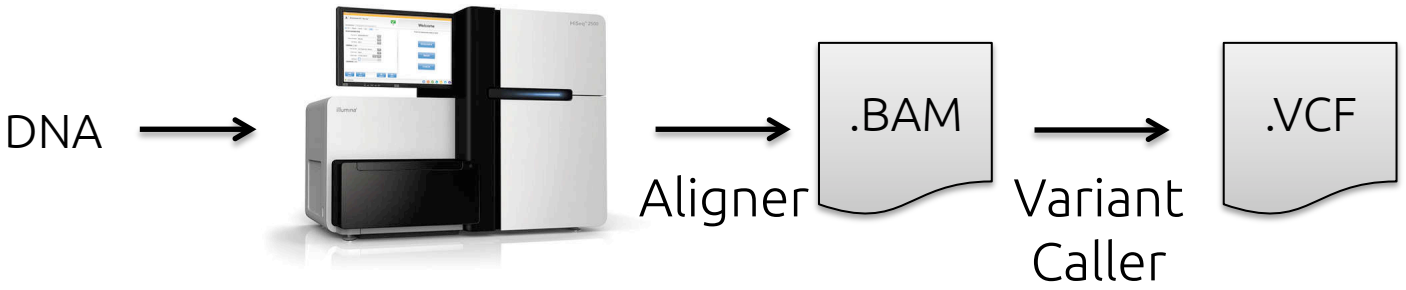
Variant visualisation and quality control

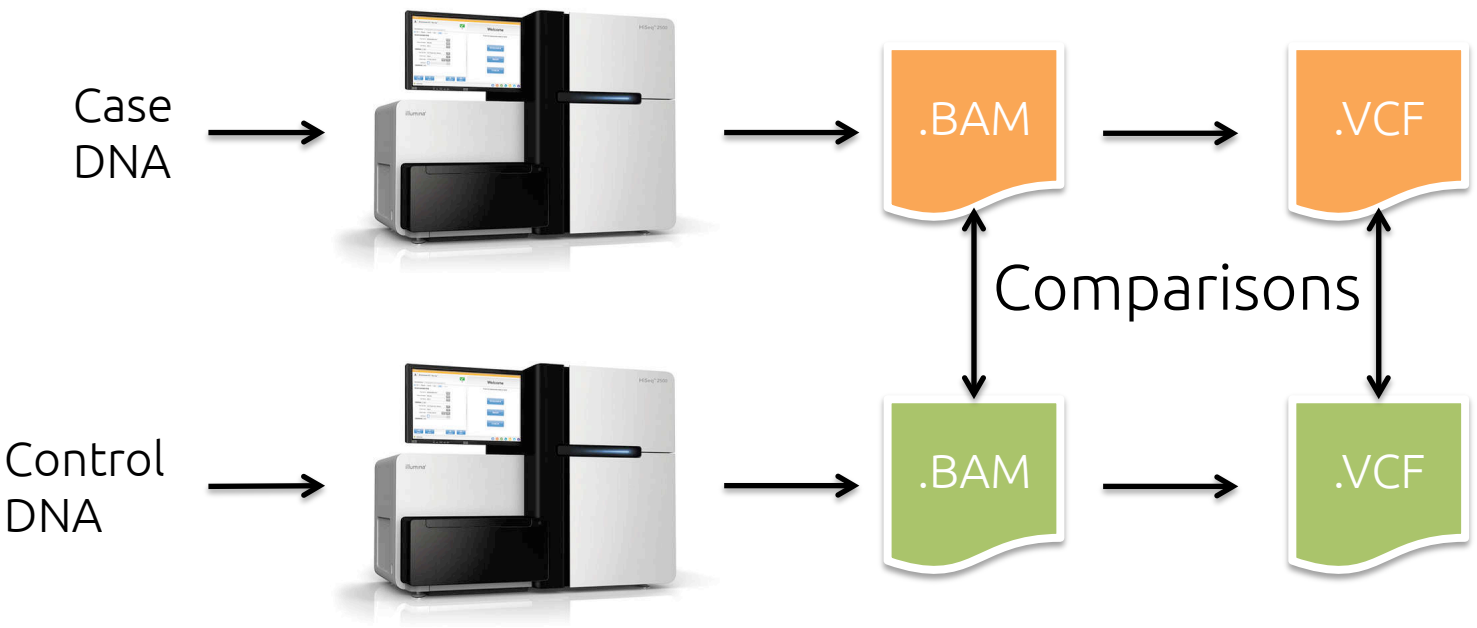
You really should be making plots!

Classical Sequencing Example



A single sample sequencing run

Comparative Sequencing Example



A comparative genomics example

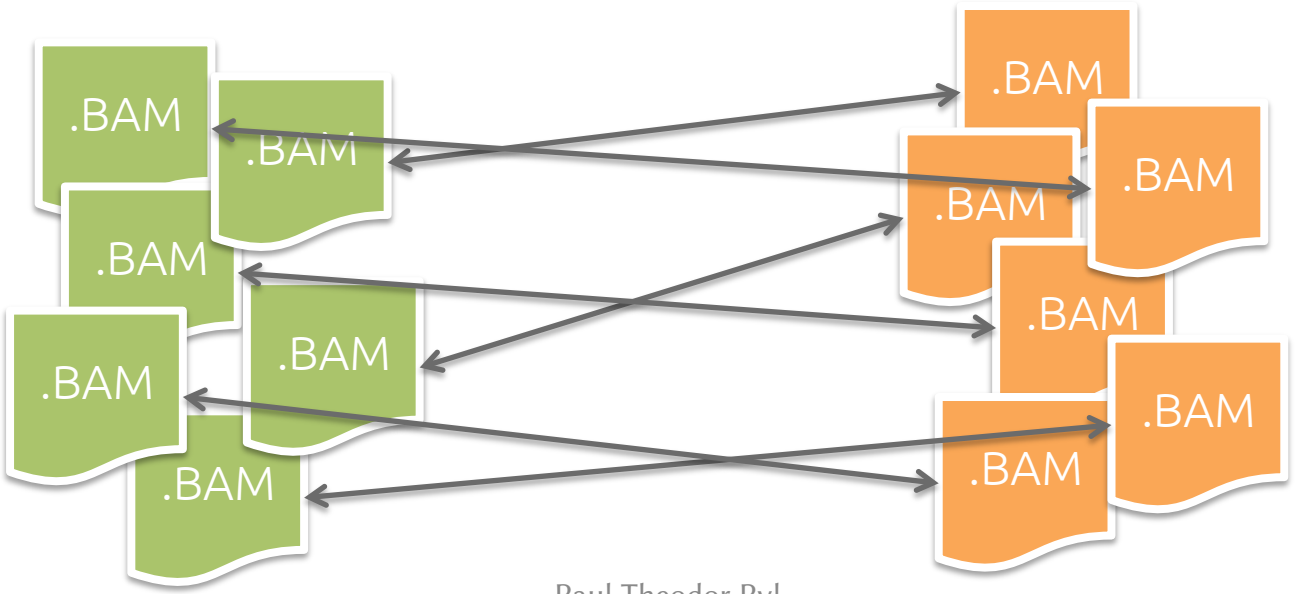
Scary Sequencing Example

DNA
(many Samples)

.BAM

.VCF

What to do with 1000 .BAM files (~200GB each)



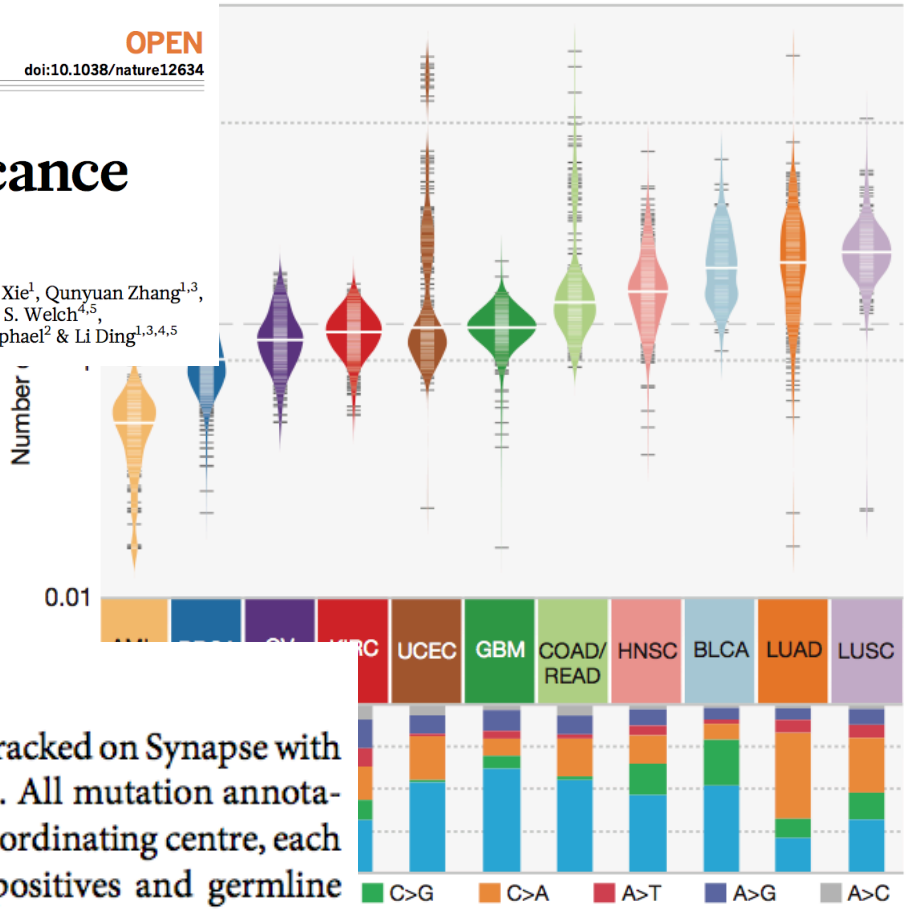
How are we dealing with this?

ARTICLE

OPEN
doi:10.1038/nature12634

Mutational landscape and significance across 12 major cancer types

Cyriac Kandoth[1]*, Michael D. McLellan[1]*, Fabio Vandin[2], Kai Ye[1,3], Beifang Niu[1], Charles Lu[1], Mingchao Xie[1], Qunyuan Zhang[1,3], Joshua F. McMichael[1], Matthew A. Wyczalkowski[1], Mark D. M. Leiserson[2], Christopher A. Miller[1], John S. Welch[4,5], Matthew J. Walter[4,5], Michael C. Wendl[1,3,6], Timothy J. Ley[1,3,4,5], Richard K. Wilson[1,3,5], Benjamin J. Raphael[2] & Li Ding[1,3,4,5]



METHODS SUMMARY

Mutation data were standardized for 12 cancer types and tracked on Synapse with documentation (http://dx.doi.org/10.7303/syn1729383.2). All mutation annotation format files were downloaded from the TCGA data coordinating centre, each being reprocessed to eliminate known, recurrent false positives and germline single nucleotide polymorphisms (SNP) present in the dbSNP database. All vari-

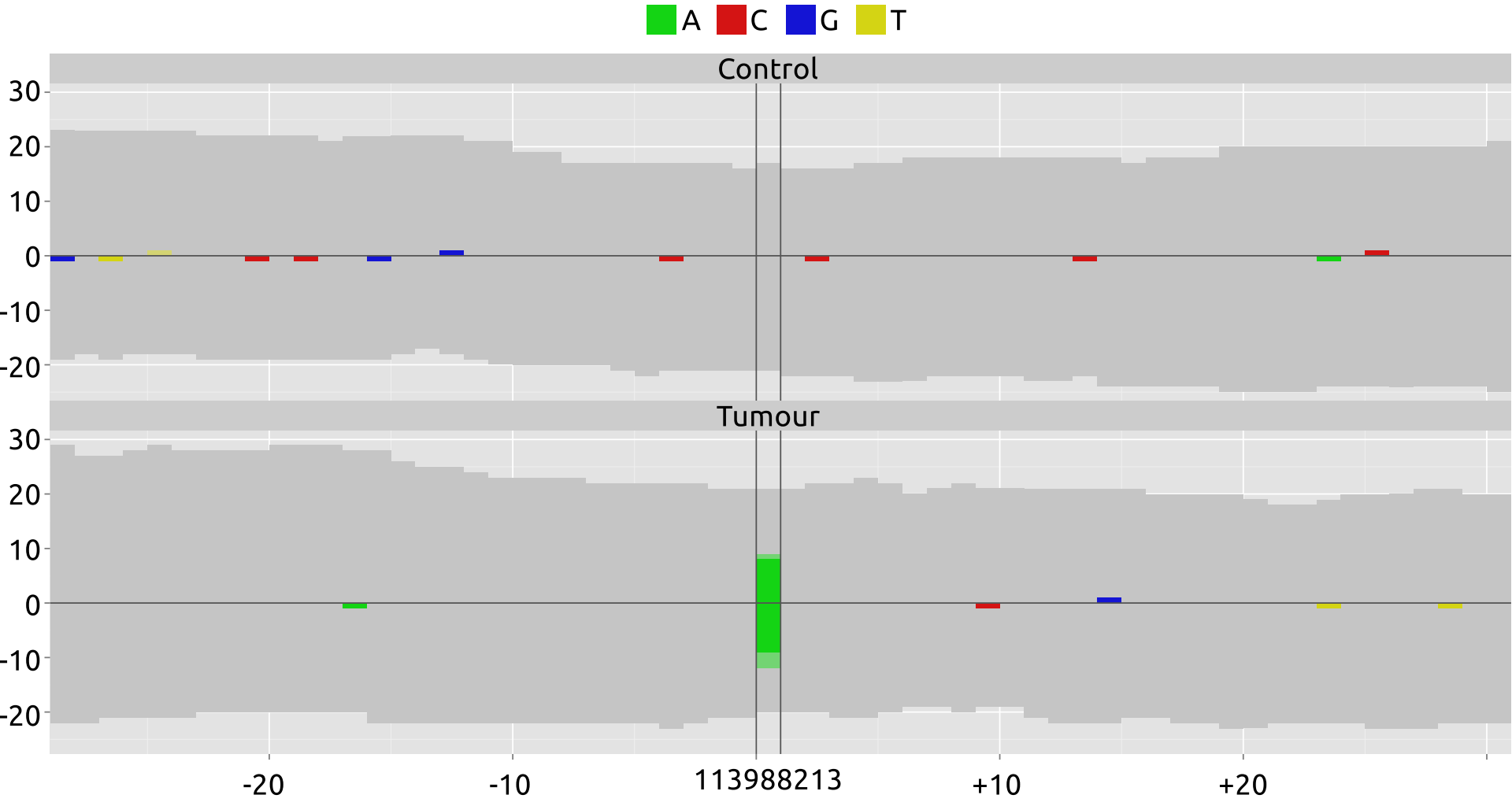
SNV Calling

- Pretty well established for diploid monoclonal populations (i.e. non-cancer human sample); e.g. GATK; samtools
- Can be more problematic in interesting samples, e.g. cancer:
  - Math might make unreasonable assumptions (copy number, clonality, etc. …)
- Specialised tools exists: e.g. MuTect

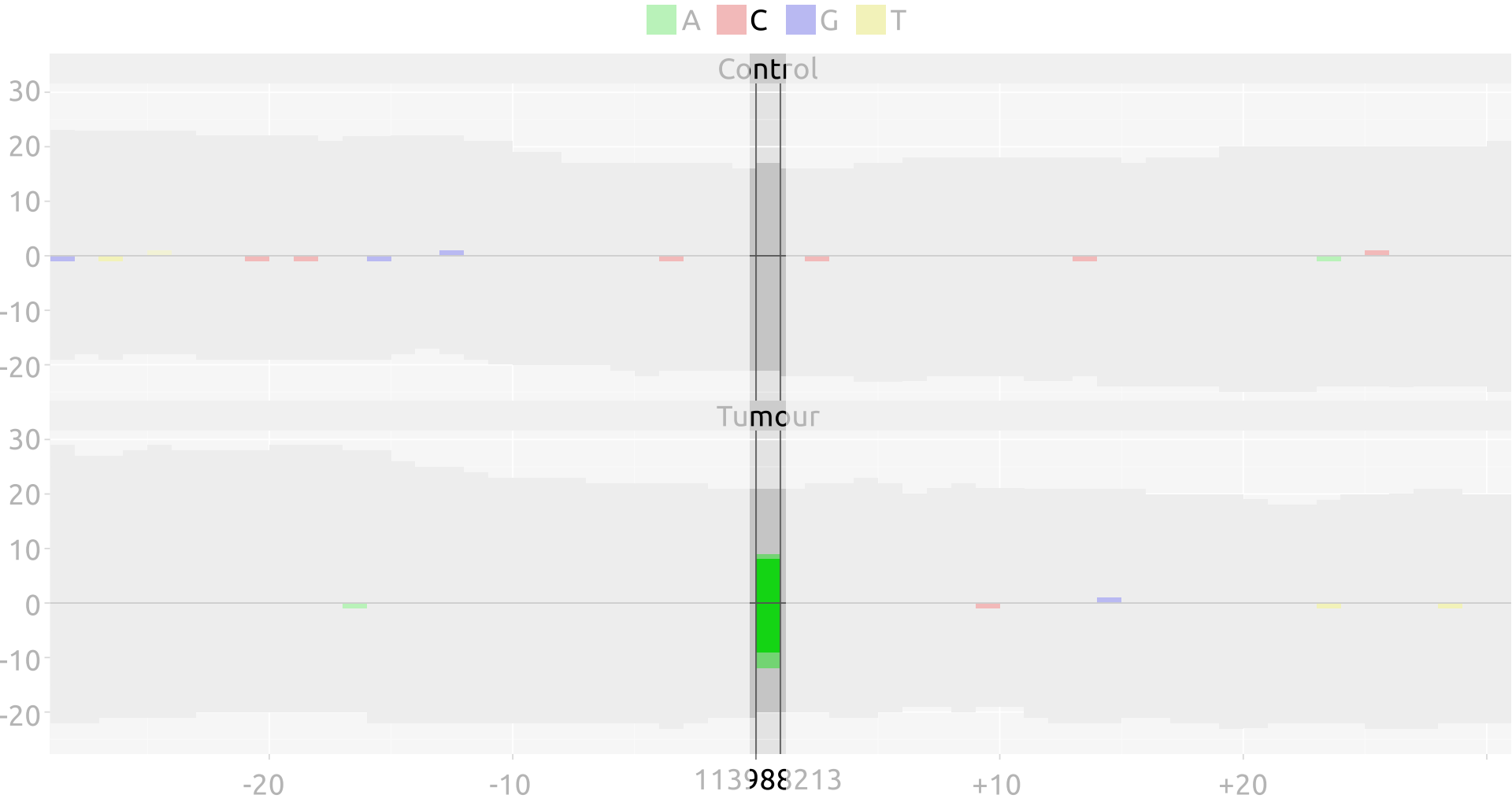
Example VCF

```
#CHROM    POS        ID   REF  ALT  QUAL FILTER    INFO       Format    Control   Tumour
1         113988213  rs…  C    A    65   PASS      GMAF=0.02  AD:DP:GT  0:42:0/0  24:38:0/1
2         101733683  -    G    C    60   PASS      GMAF=0.3   AD:DP:GT  0:18:0/0  5:14:0/1
…
```

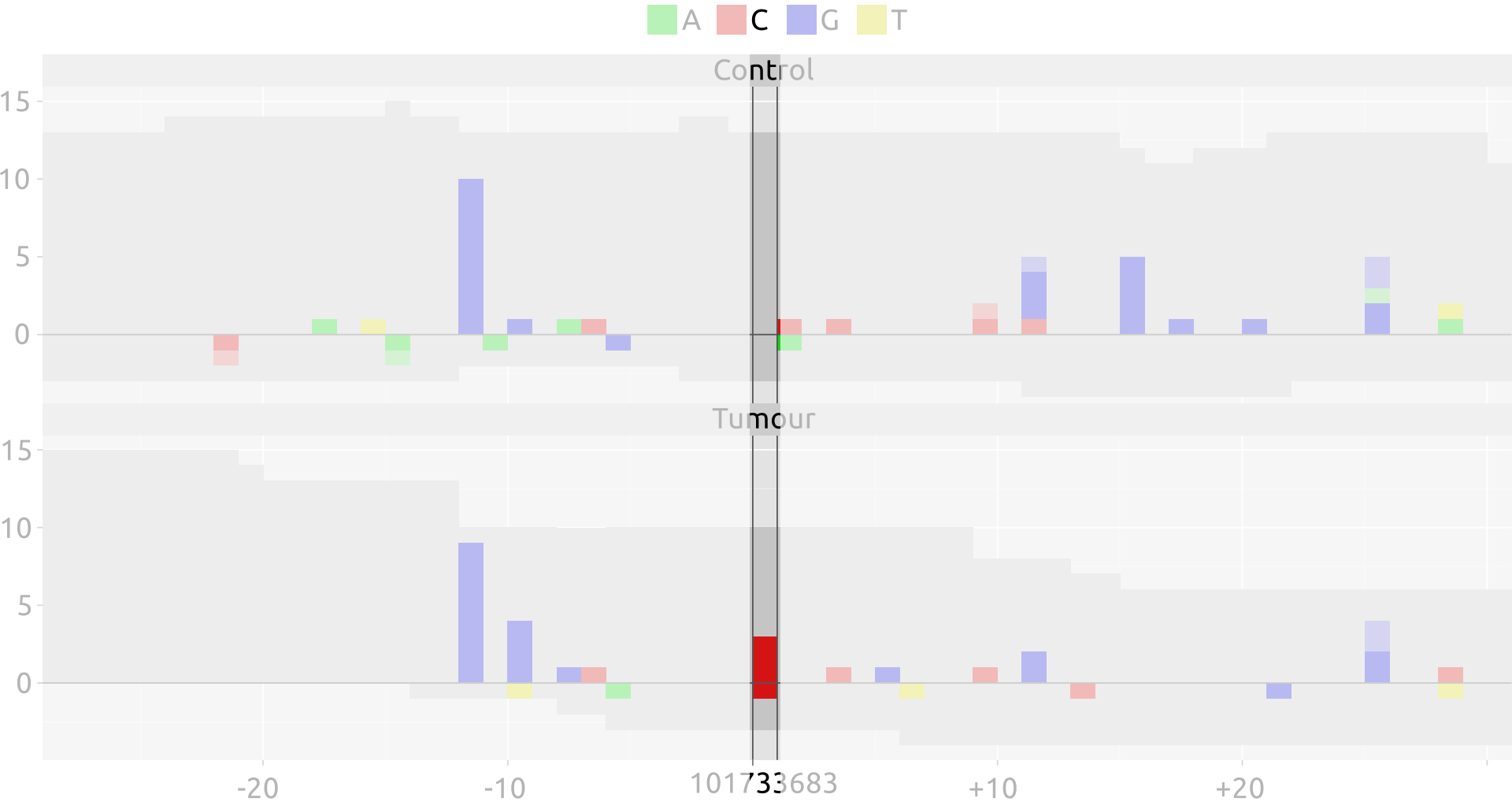
Visualisation is Key



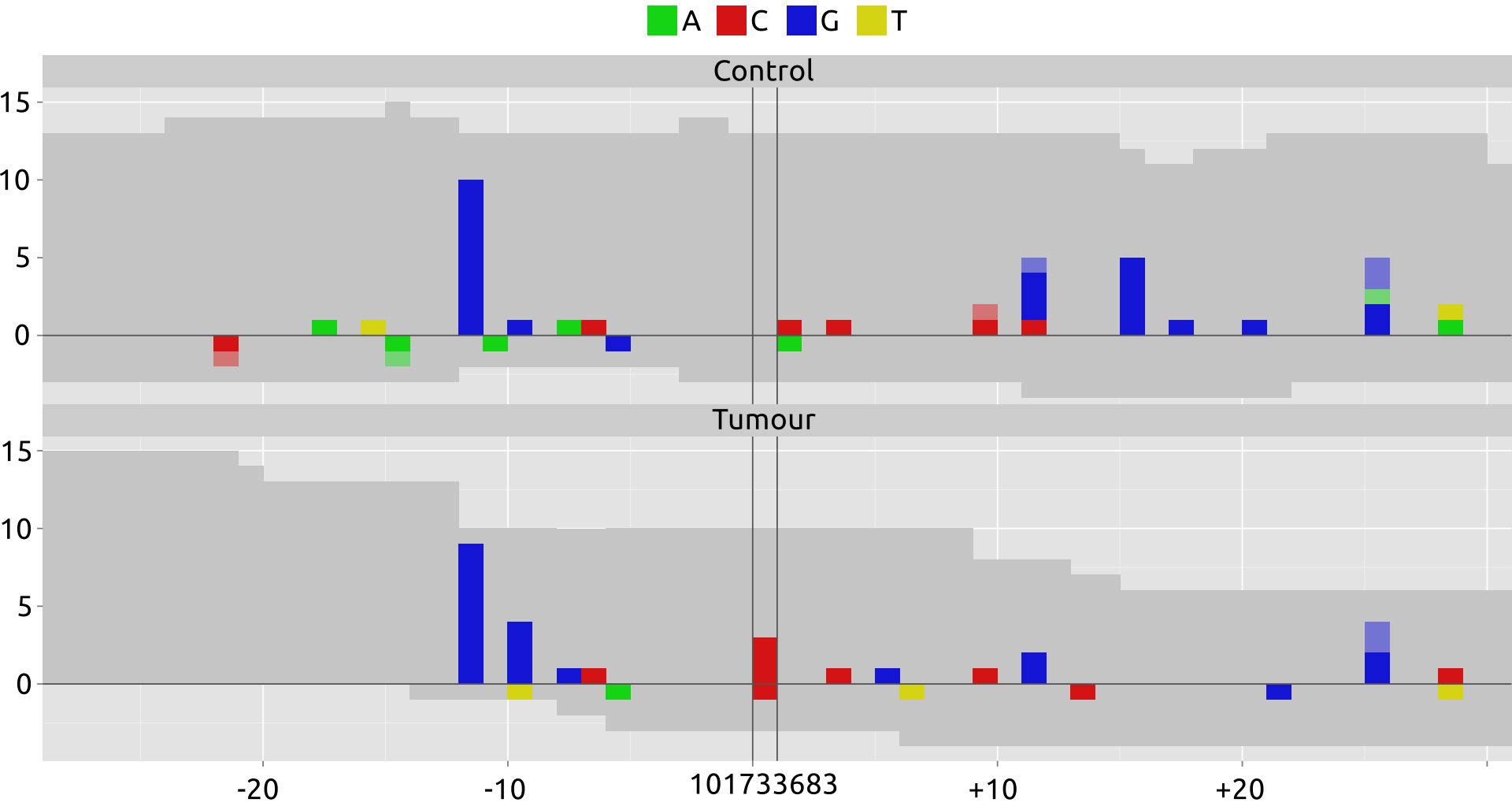
What the VCF file will tell you



What the VCF file won't tell you

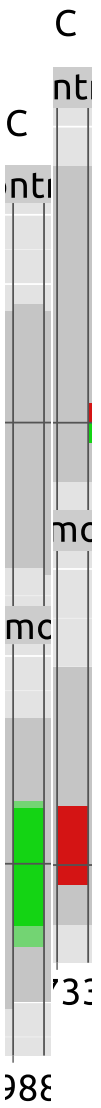


What the VCF file won't tell you

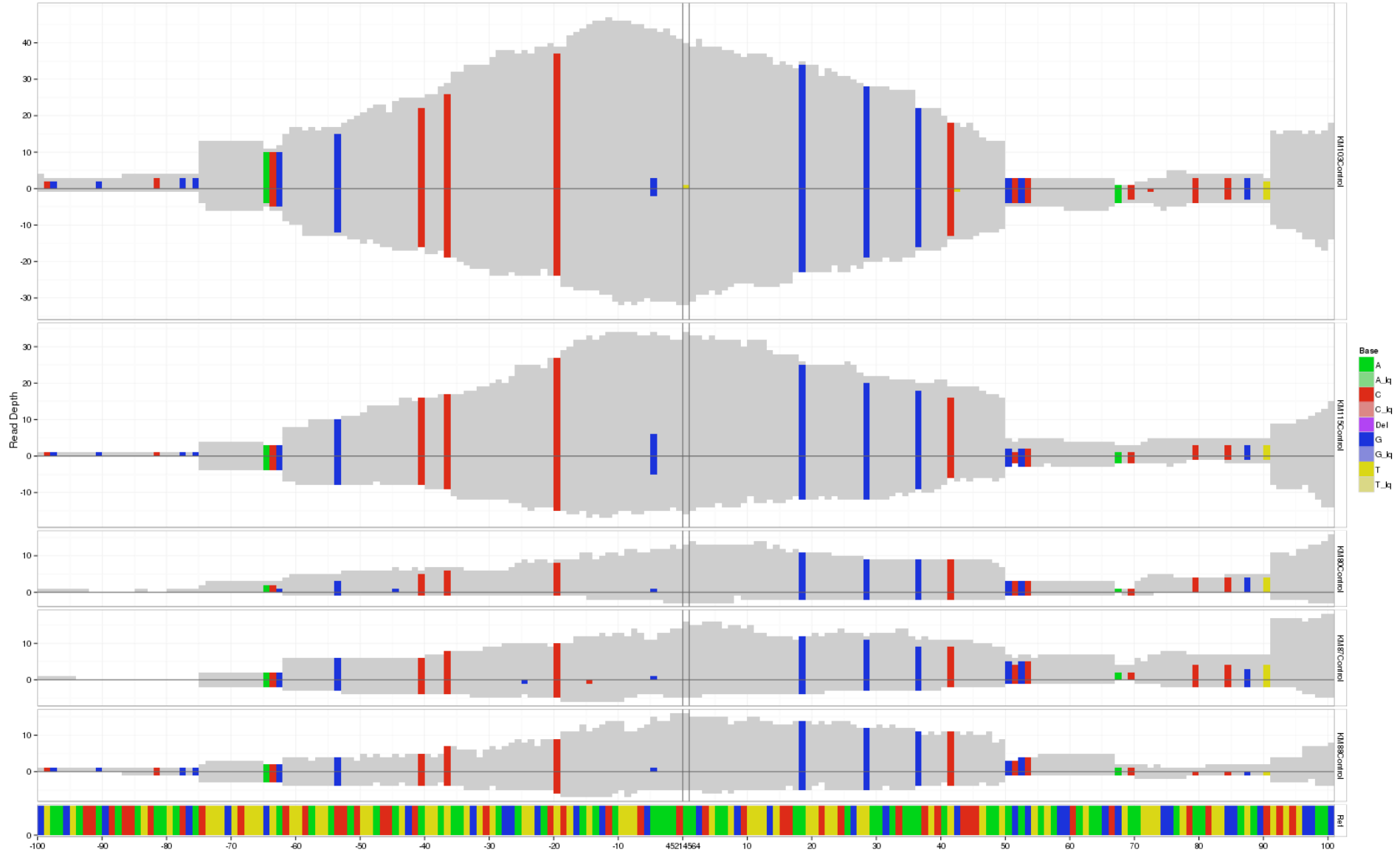


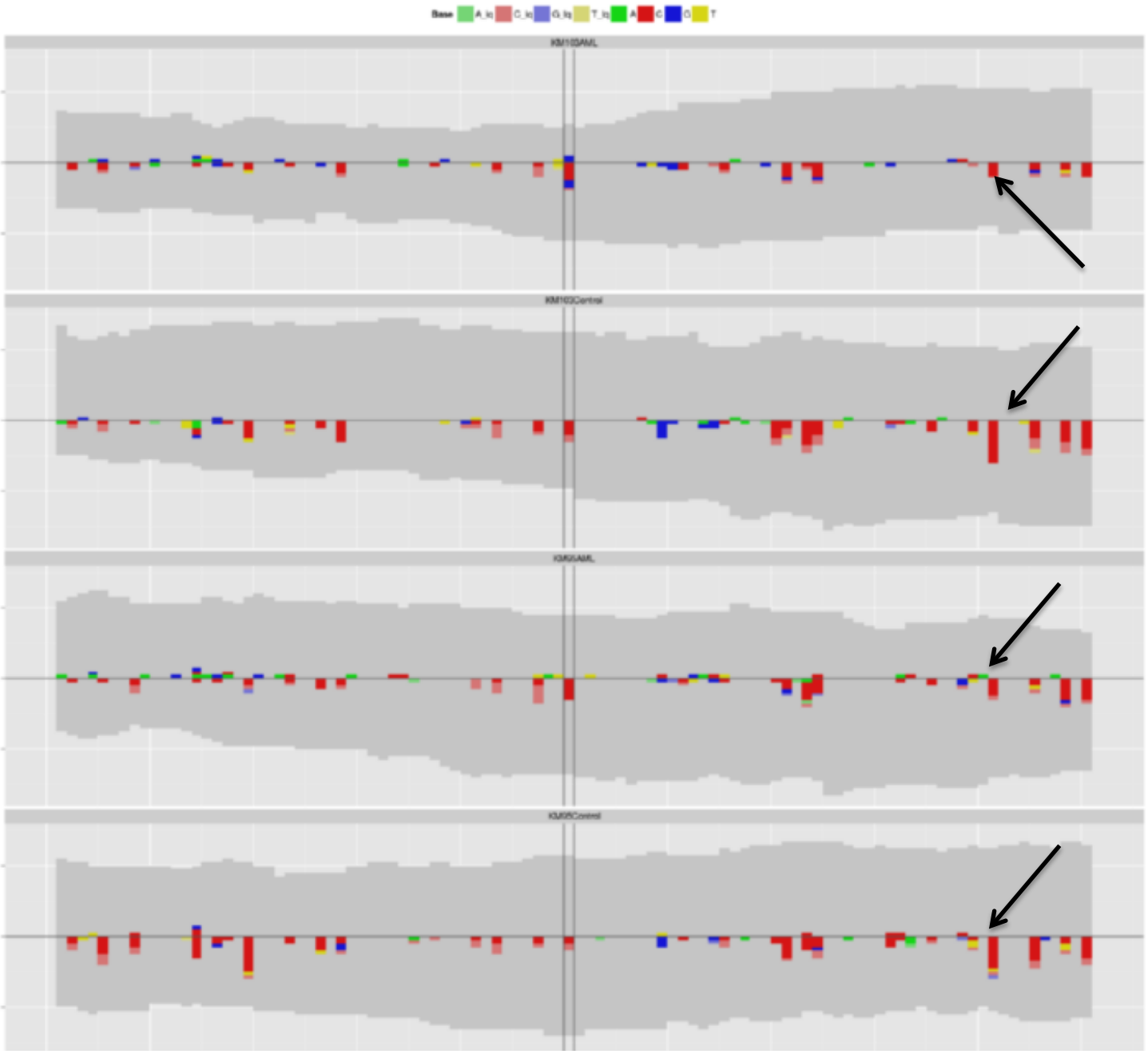
Example VCF revisited

#CHROM	POS	ID	REF	ALT	QUAL	FILTER	INFO	Format	Control	Tumour
1	113988213	rs…	C	A	65	PASS	GMAF=0.02	AD:DP:GT	0:42:0/0	24:38:0/1
2	101733683	-	G	C	60	PASS	GMAF=0.3	AD:DP:GT	0:18:0/0	5:14:0/1
…										



Example: CDC27





Half-time Summary

- Visualisation gives context
- A list of positons (i.e. VCF file) is likely to miss out on some of that
- Good to know:
  - Regions that are always hard (e.g. CDC27)
  - Regions that show specific artifacts from library prep / sequencer (if the samples are all processed the same)

Important post-processing Steps (Alignments)

- After alignment:
  - Remove duplicates
  - InDel realignment (GATK)

```
ATTAC--ACAC
 TTAC--ACAC
GATTACTTACAC
```

```
ATTAC--ACAC
  TTACACAC
GATTACTTACAC
```

```
ATTACACAC
 TTACACAC
GATTACTTACAC
```

Important post-processing Steps (Variants)

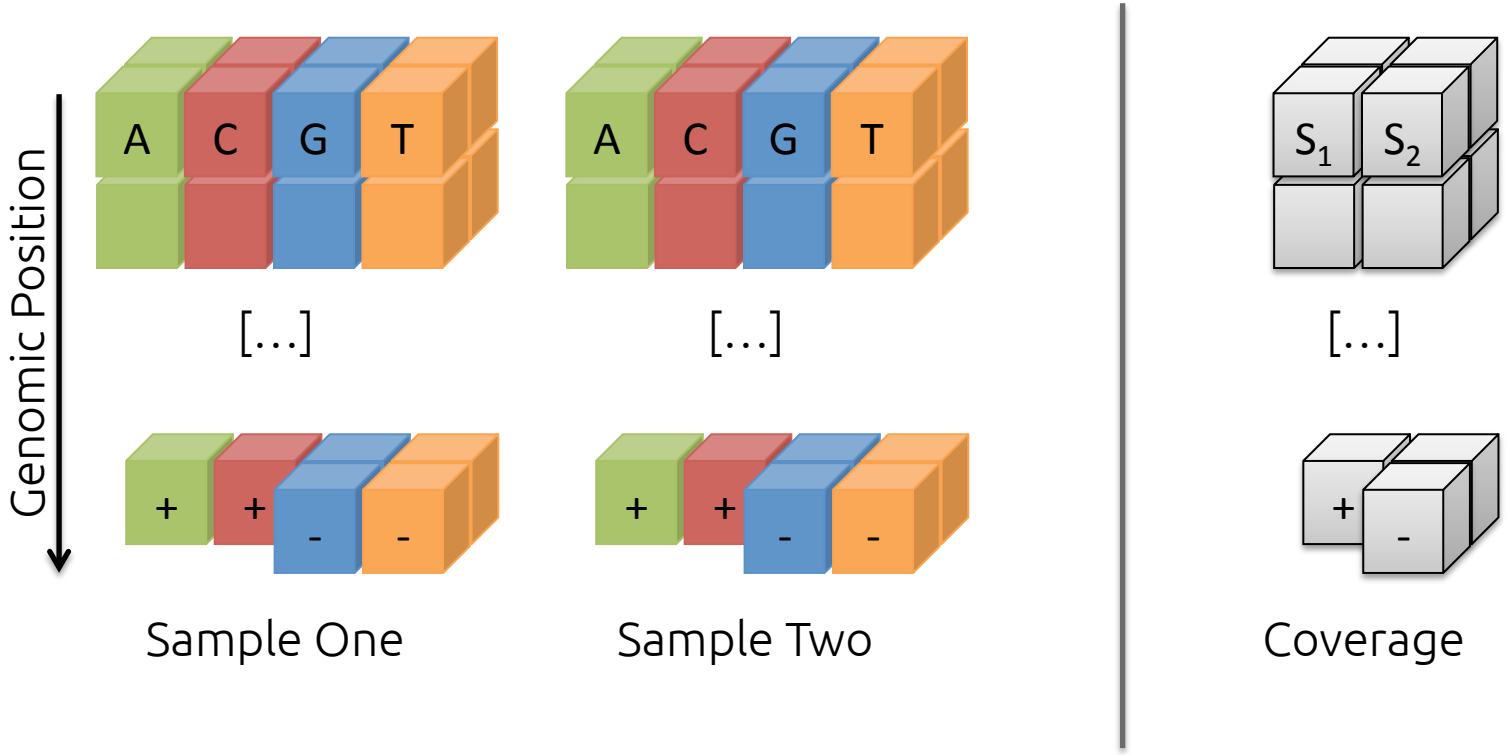
- Ensembl Variant Effect Predictor
  - R package: ensemblVEP (or use command line tool)
    - Location / overlapping genes etc.
    - GMAF (1000genomes or HapMap)
    - SIFT / PolyPhen Scores
- Annotate with available data
  - e.g. mismatch rates in other samples of the same cohort
  - Local mismatch rate within a sample (e.g. genomic distance to the next 10 mismatches)

Visualisation Tools

- Genome Browser, e.g. IGV
  - Programmatic access? (IGV can be scripted)
- h5vc R/Bioconductor package
  - Processing BAM files into nucleotide tallies
  - Analysing and visualising on those
  - Shamelessly advertising my own software ☺

Nucleotide Tallies

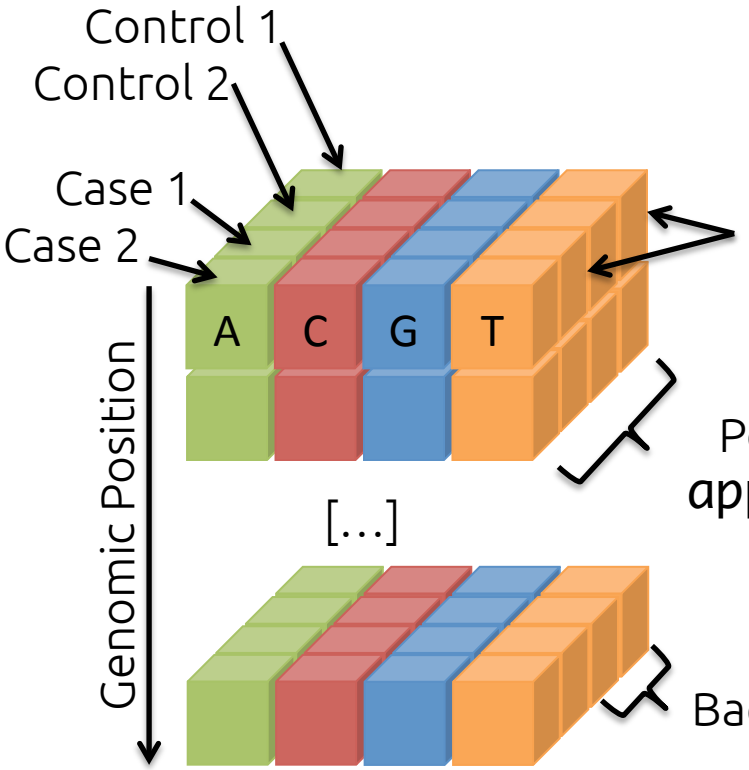
- Table of (mis)matches, coverages, deletions, insertions, softclips, …



Genomics Analyses on Tallies

- Having the data as a matrix:
  - Easy subsetting (e.g. selecting all controls)
  - Easy building of summary statistics
    - applying functions to the matrix
    - E.g. summarise control samples
- Many Analyses, especially variant calling and visualisation, can be performed on tallies (we don't need the BAM's for it)

Genomics Analyses on Tallies



Pairwise Comparisons

```
a[,c(2,4),]
```

Population level statistics

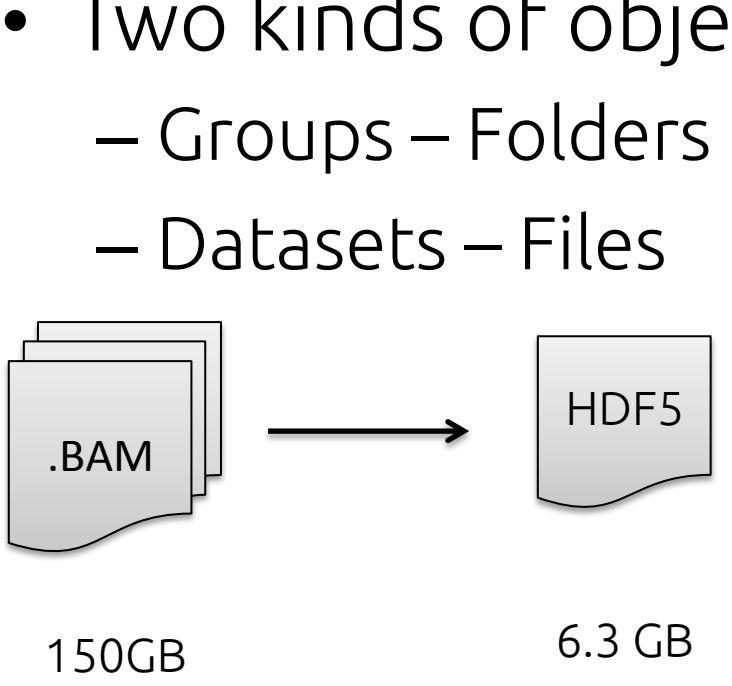
```
apply(a, c(1,3), sum)
```

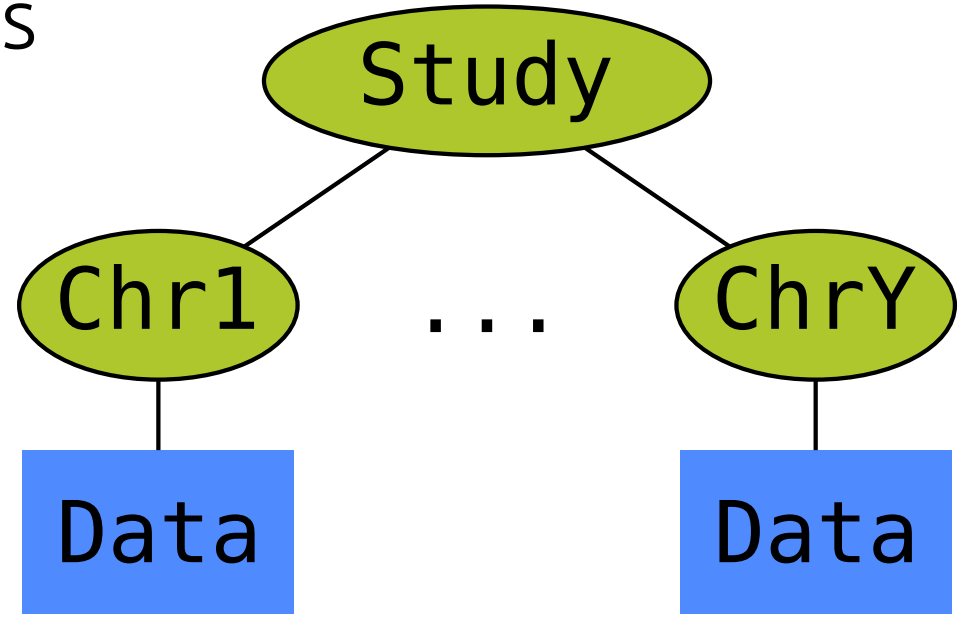
Background of mismatches in control samples

```
apply(a[,3:4,], c(1,3), mean)
```

What is HDF5

- Hierarchical Data Format
  - Efficient storage of numerical data
- Two kinds of objects
  - Groups – Folders
  - Datasets – Files





HDF5 – A brief overview

- Introduced in 1987
  - National Center for Supercomputing
- Maintained by the HDFGroup
- Production Use
  - NASA
  - Imaging
  - The Lord of the Rings

What to store in our HDF5 file

- 4 data-sets per Chromosome
- Counts
  - 4D : [bases x samples x strands x positions]
- Coverages
  - 3D : [samples x strands x positions]
- Deletions
  - 3D : [samples x strands x positions]
- Reference
  - 1D : [positions]

The ‘h5vc’ package

- Available in R/Bioconductor
- Functionality:
  - creating / interacting with HDF5 tally files
  - Variant calling
  - Data exploration
  - Plotting
  - …

Bioinformatics Advance Access published February 5, 2014

BIOINFORMATICS **APPLICATIONS NOTE** *2014, pages 1–3*
doi:10.1093/bioinformatics/btu026

Genome analysis Advance Access publication January 21, 2014

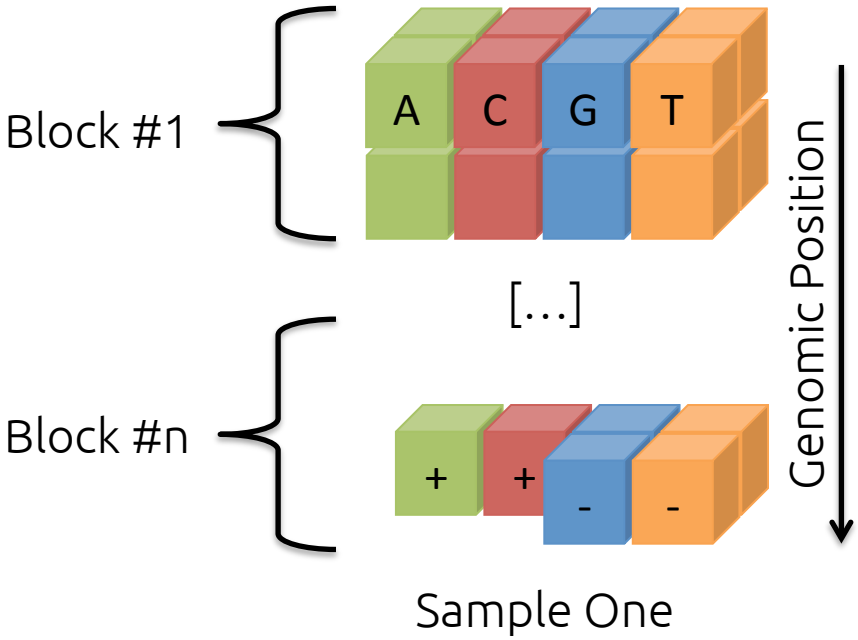
h5vc: scalable nucleotide tallies with HDF5

Paul Theodor Pyl*, Julian Gehring, Bernd Fischer and Wolfgang Huber*
EMBL Heidelberg, Genome Biology Unit, Meyerhofstr. 1, 69117 Heidelberg, Germany

Associate Editor: John Hancock

Applying Functions Block-wise

```
variantCalls <- h5dapply(
    filename = "example.tally.hfs5",
    group = "/ExampleStudy/16",
    blocksize = 100000,
    names = c("Counts", "Coverages"),
    dims = c(4, 3),
    range = c(29000000, 30000000),
    FUN = callVariants,
    sampledata = sampleData
)
```



Tutorial Tomorrow

- Example Workflow
  - Creating Tally Files
  - Variant Calling
  - Visualisation and Quality Control
  - Creating Reports
  - …

My Current Workflow

- Alignment (e.g. gsnap)
- Postprocessing (GATK)
  - Remove PCR duplicates
  - InDel realignment
- Tallying (h5vc)
- Variant Calling (e.g. h5vc)
- Ensembl VEP
- ReportingTools (Interactive HTML tables)

Final Summary

- (comparative) variant calling is not completely solved yet
  - ➢ We need to do some quality control
- Plotting variants can be helpful
  - – Tables of variants risk missing important context
- We should try to formalise the intuitions we use for visual inspection (we're working on it)
- HDF5-based nucleotide tallies allow for analysis and visualisation of SNVs in context

Acknowledgement

- EMBL and the Huber Lab
  - Wolfgang
  - **Bernd Fischer**
  - Simon Anders
  - Julian Gehring

Tutorial this afternoon

- http://192.168.0.9/materials/4_Thursday/labs/
  - ExampleData.zip
  - Tutorial.Rmd
  - Tutorial.R
  - Tutorial.pdf
- Get newest version of h5vc:

```
source("http://192.168.0.9/biocLite.R")
biocLite(”h5vc")
```