Filtering artefacts in somatic single nucleotide variant calling using a panel of normals

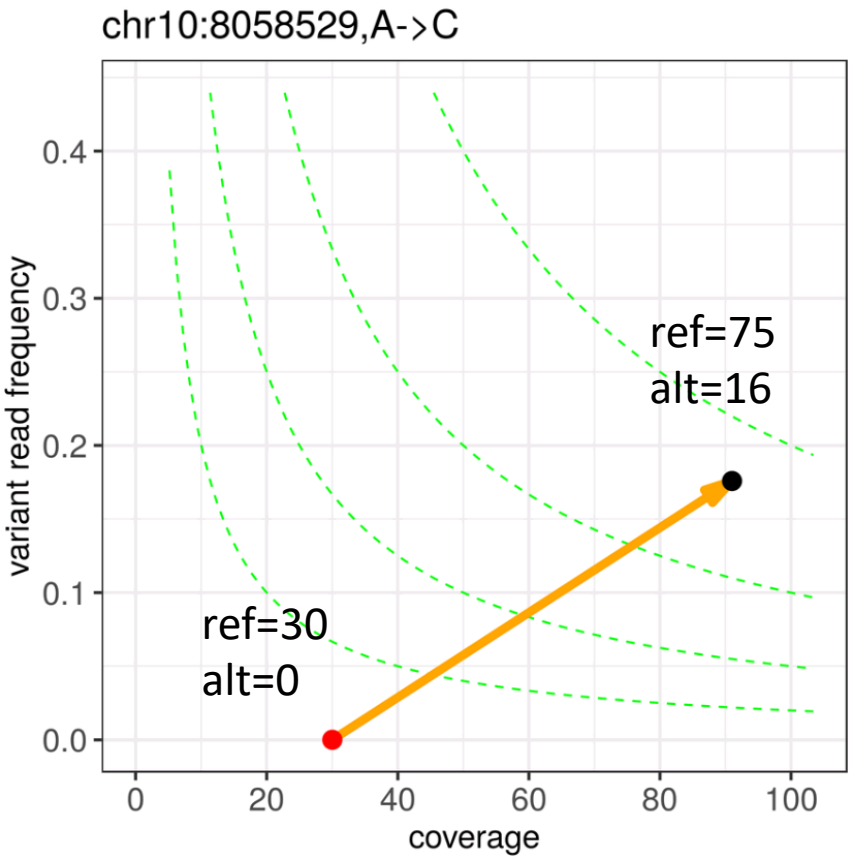
Boris Noyvert
CRUK Birmingham Centre and
Centre for Computational Biology
University of Birmingham

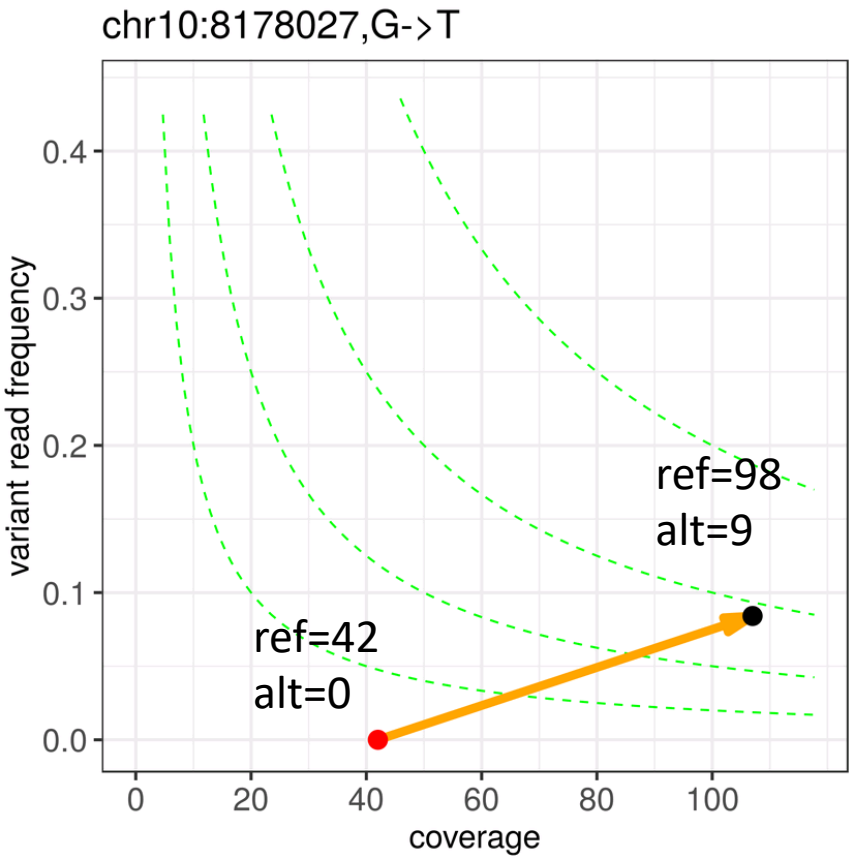
Jonathan Mitchell
Genomics England

Daniel Chubb
Institute of Cancer Research
London

Alona Sosinsky
Genomics England

Somatic mutations detected by Illumina pipeline

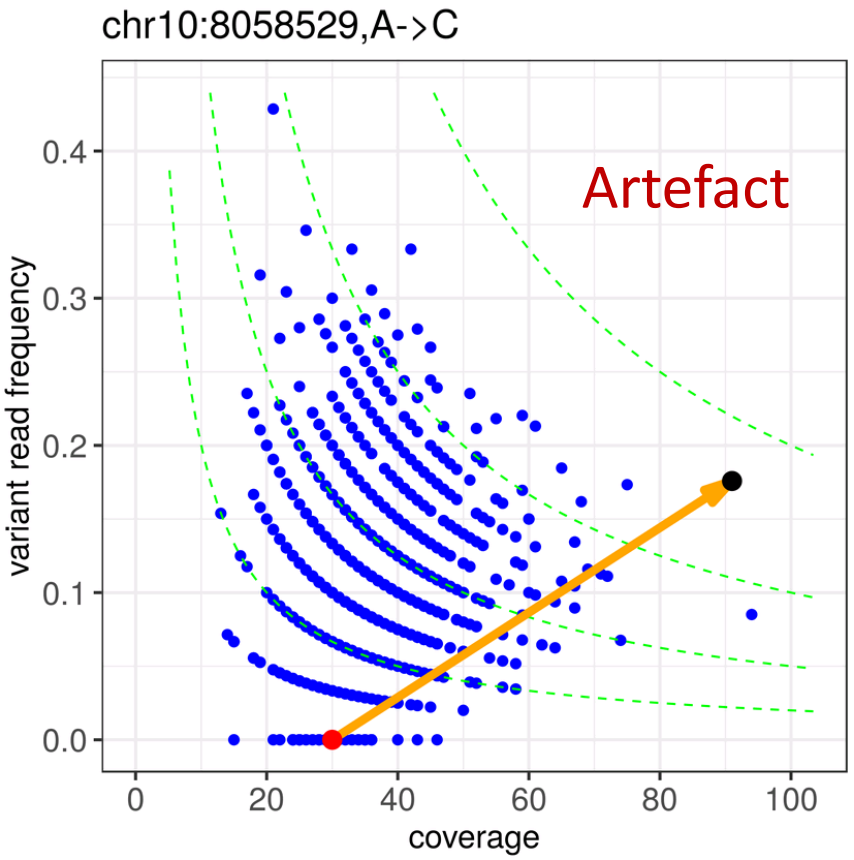


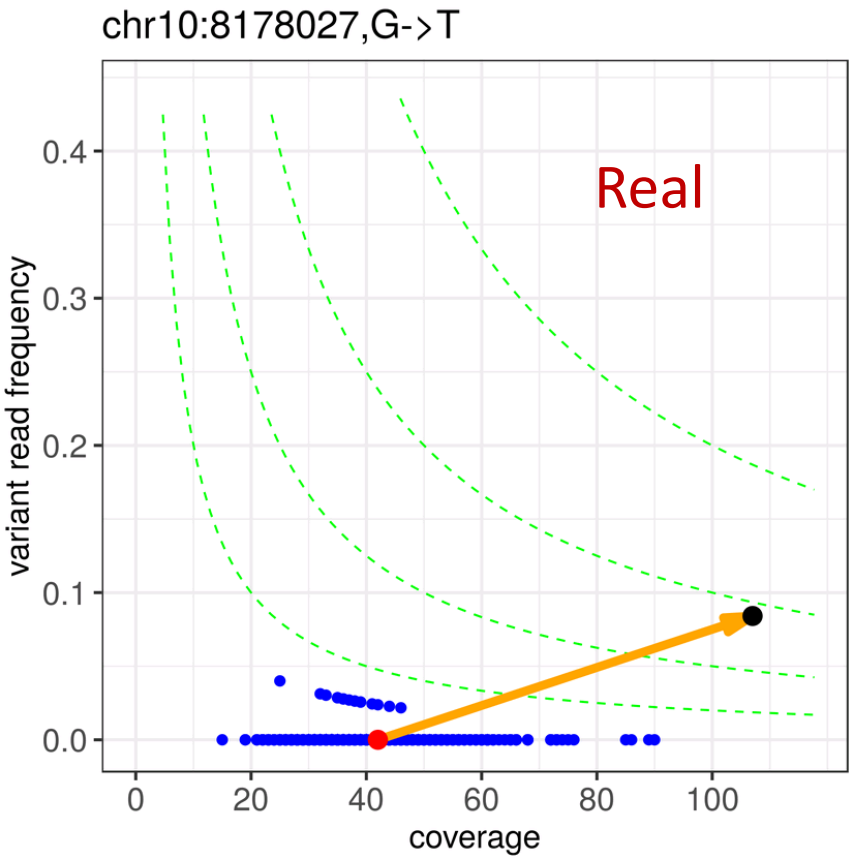


Red circle – germline sample, black circle – tumour sample.
X-axis: total number of reads covering variant site (ref+alt).
Y-axis: variant read frequency = alt/(ref+alt)

Real mutations?
Look real based on read numbers for germline-tumour pairs.

Adding more samples to the picture

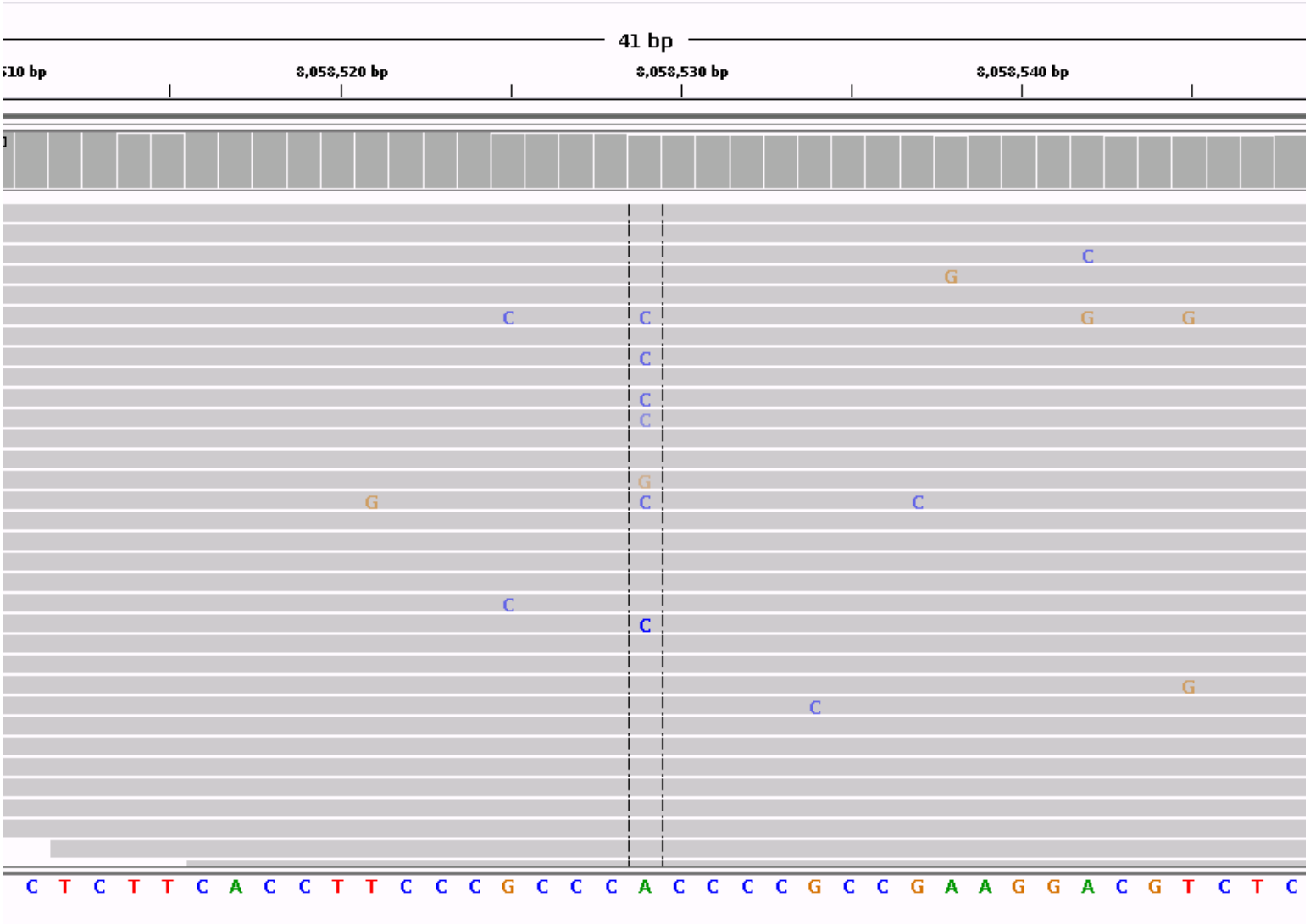




Blue points – 1000 germline samples from the rare disease cohort.

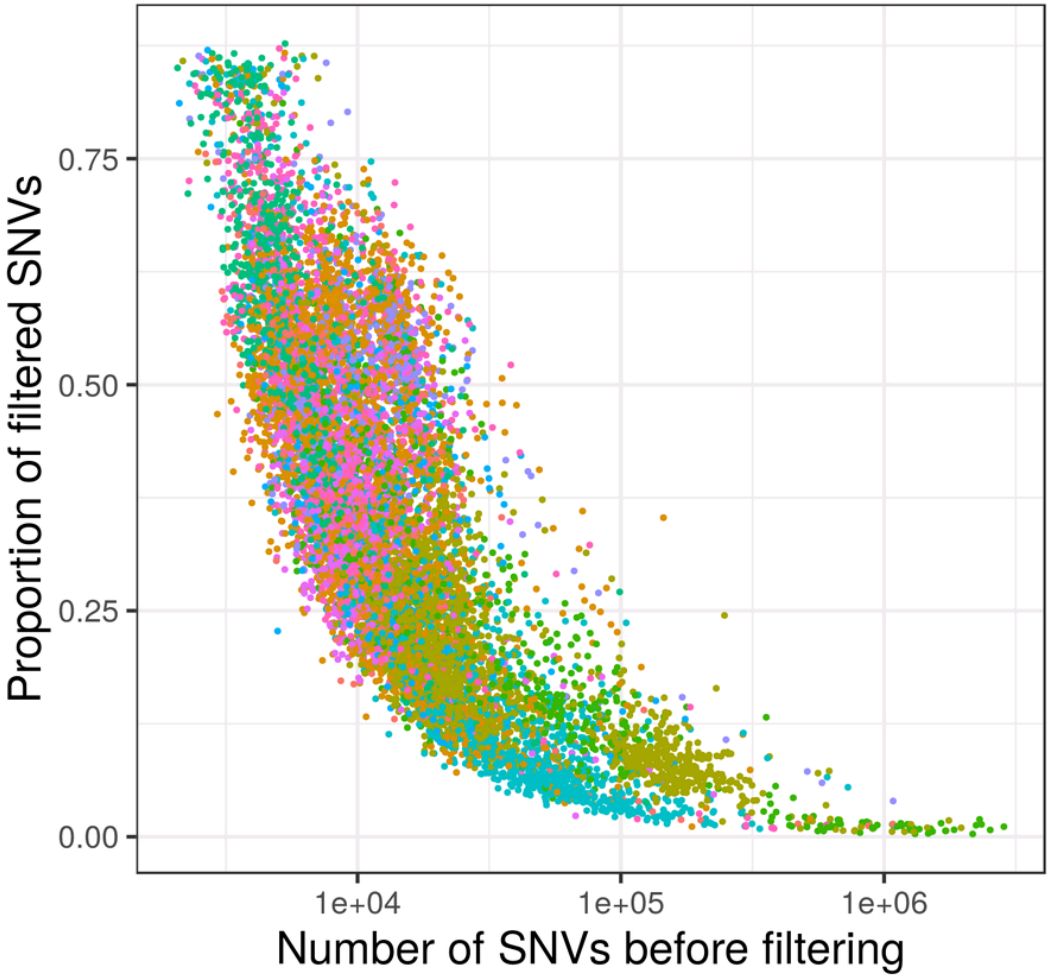
Artefacts are difficult to detect by analysing one sample pair at a time, but are easy to spot on multi-sample diagrams!

Recurring sequencing noise



IGV screenshot for the tumour sample at the artefact site

On average 35% of somatic SNV calls are affected



cancer

ADULT_GLIOMA	0.402
BREAST	0.424
COLORECTAL	0.215
ENDOMETRIAL_	0.299
HAEMONC	0.564
LUNG	0.218
OVARIAN	0.374
PROSTATE	0.506
RENAL	0.387
SARCOMA	0.459

Proportion of SNV calls filtered per sample:
mean=0.356, SD=0.198,
median=0.342.

Summary

- Standard types of somatic variant calling software look at one germline-tumour sample pair at a time.
- To detect sequencing and mapping artefacts one has to look at the suspected mutation site **across a large number of samples (PoN – panel of normals)**.
- The exact nature of the artefact is not important – the same procedure can be used.
- Substantial numbers of single nucleotide variant calls are false positive.

PoN Genomes

- 100K Genomes Project Rare Disease Individuals
- Exclude Proband, and keep one individual per family
- 50% Female, 50% Male
- PCR-free library prep
- Blood samples (non saliva)
- Cross-sample contamination by VerifyBamID < 0.5%
- 7,000 Individuals
- ~ 1 CPU day per Individual

PoN Implementation

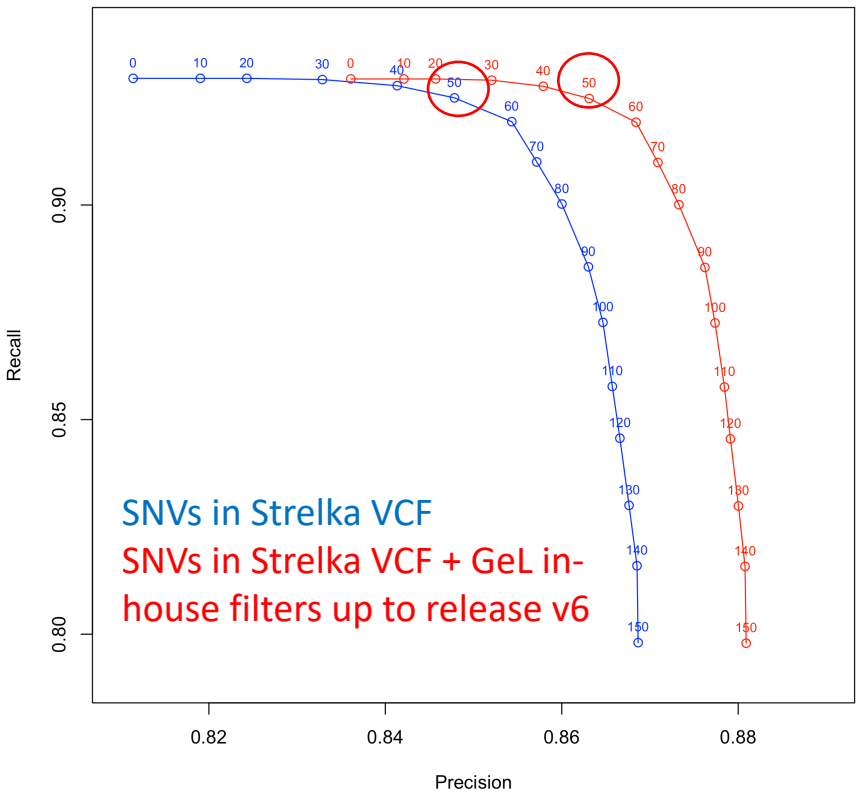
For PoN genomes

- Generate alt/ref counts for each position in genome. Replicate Strelka filters for low quality reads/basecalls:
  - filter reads with mapping quality below 5
  - filter duplicated reads
  - filter basecalls with quality below 5
- Remove counts that support Strelka-called germline variants. We assume that common germline variants and high level noise is already filtered by *CommonGermlineVariant* GeL in-house filter
- Store the ratio of allele depths across PoN genomes for each position in genome

For a patient genome

- Generate alt/ref counts for each somatic SNV (with the same filters as PoN)
- Run Fisher exact test for each somatic SNV.
  - H0: ratio of tumour allele depths is not significantly different from the ratio of allele depths at this site in the PON
- Annotate each somatic SNV with Fisher exact test phred score

Phred score cut-off selection: ROC curve for high-confidence test set

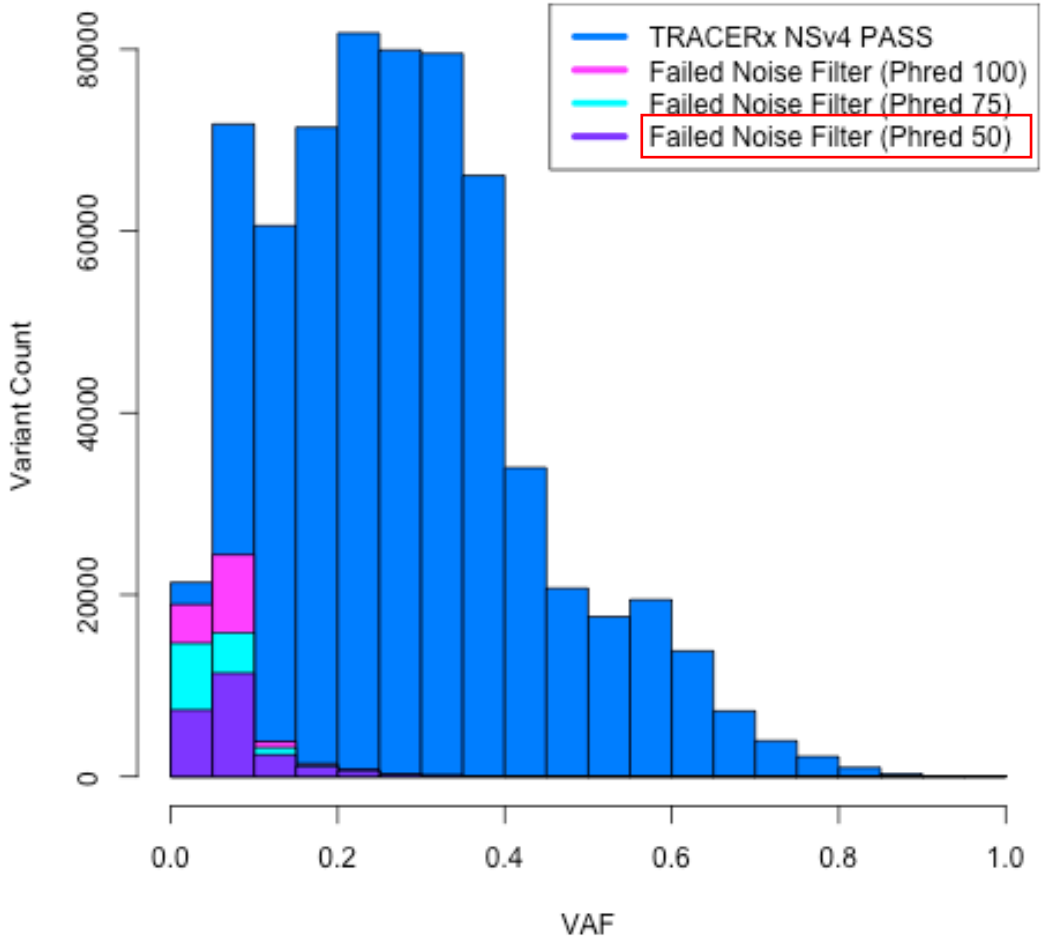


TRACERx small variants test set:

- Ten high-depth exomes (400x)
- Results of exome sequencing had previously been validated for a subset of variants with multiplex PCR and AmpliSeqTM custom panel
- Resulting sensitivity and precision for TRACERx data set was estimated > 99% => high-confidence test set

DNA from the same aliquot underwent WGS and was run through the Genomics England pipeline.

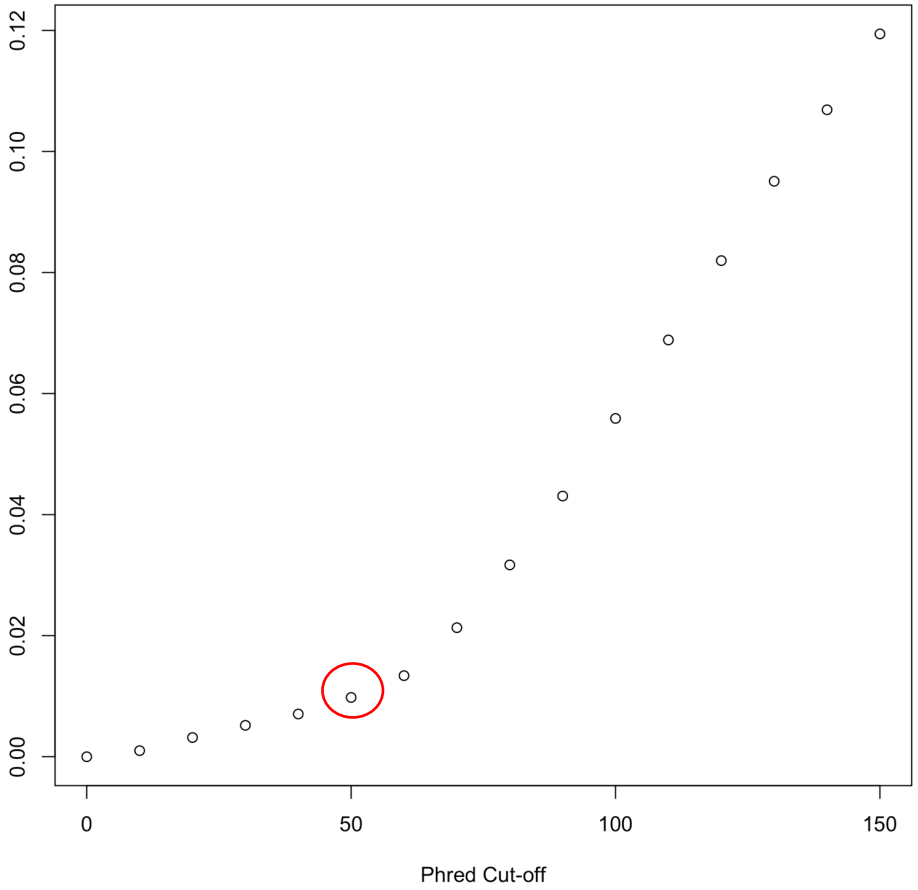
Phred score cut-off selection: Filtering by variant allele frequency



Analysis is performed for 13 TRACERx genomes

Phred score cut-off selection: Filtering potentially actionable variants

Fraction of coding non-synonymous variants in 86 genes associated with approved therapies and clinical trials for colorectal patients that was filtered with PoN filter



Analysis is performed on 1675 genomes from colorectal tumours

Research Environment release V7+ annotated VCF files

FILTERs

- PONnoise50SNV - SomaticFisherPhred below 50, indicating somatic SNV is systematic mapping/sequencing error (applies only to SNVs on primary genome assembly that pass Strelka filters)
- CommonGermlineVariant - Variants with a population germline allele frequency above 1% in a Genomics England cohort
- CommonGnomADVariant - Variants with a population germline allele frequency above 1% in gnomAD dataset
- RecurrentSomaticVariant - Recurrent somatic variants with frequency above 5% in a Genomics England cohort
- BCNoise10Indel - Average fraction of filtered basecalls within 50 bases of the indel exceeds 0.1, FDP50/DP50 > 0.1
- SimpleRepeat - Variants overlapping simple repeats as defined by Tandem Repeats Finder

INFO fields

- HomopolimerIndel - Indels intersecting with reference homopolymers of at least 8 nucleotides
- SomaticFisherPhred,Number=1,Type=Float,Description="Phred score of Fisher's test of somatic allele ratio vs PoN allele ratio (applies only to SNVs that pass Strelka filters)

Conclusions

- By using a large WGS data set systematic false positive somatic mutation calls are filtered.
- The filtering significantly improves precision with little loss in recall.
- Filtered VCF files available in the research environment (V7+).

Acknowledgements

100,000 genomes Colorectal Cancer GeCIP

University of Birmingham
Ian Tomlinson
Boris Noyvert
Archana Sharma-Oates
Albert Menezes

ICR London
Richard Houlston
Daniel Chubb
Alex Cornish

and others!

Andrea Sottoriva
Giulio Caravagna
Luis Zapata Ortiz

Oxford
David Wedge
Andreas Gruber
Anna Frangou

Barts Cancer Institute, QMUL
Trevor Graham
William Cross

Genomics England bioinformatics
Alona Sosinsky
Jonathan Mitchell
Magdalena Zarowiecki
John Ambrose

University of Birmingham
Jean-Baptiste Cazier
Roland Arnold
Anshita Goel
Richard Bryan
Douglas Ward