Further Issues and Considerations

Andy Lynch

CRUK CI

July 2016

Outline

This afternoon:

- InDels
- MNVs
- Single sample signatures
- Precision of estimates

Further Issues and Considerations

InDels

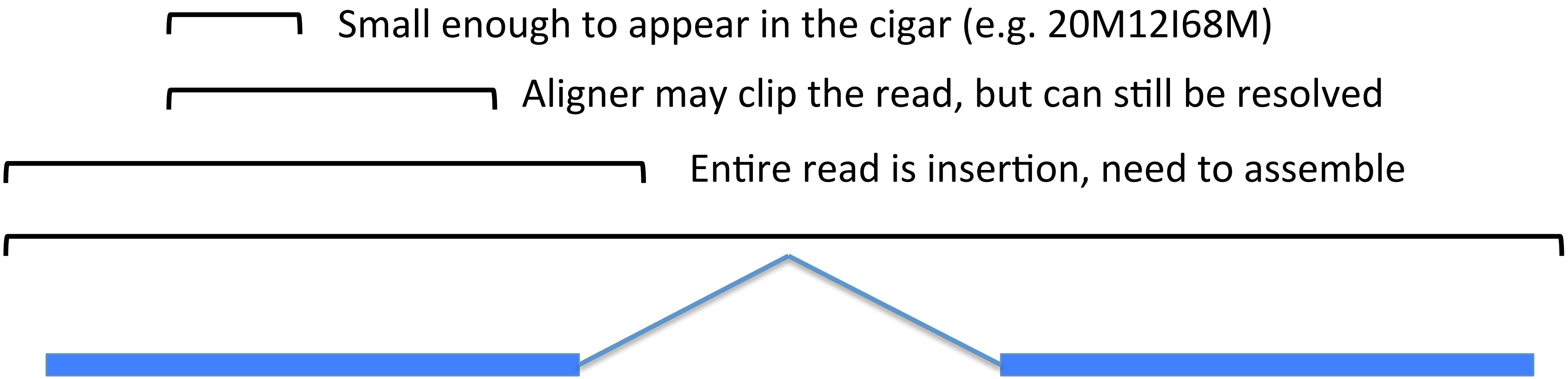
InDels

We are generally much worse at calling somatic InDels than somatic SNVs

The problems we have with InDels include

- **InDels are not one type of mutation**

Obviously there are insertions and deletions, but we also have to worry about lengths (particularly with insertions).



Small enough to appear in the cigar (e.g. 20M12I68M)

Aligner may clip the read, but can still be resolved

Entire read is insertion, need to assemble

InDels

We are generally much worse at calling somatic InDels than somatic SNVs

The problems we have with InDels include

- **InDels are often ambiguous**

We may know what our new sequence is, and what the original was, but microhomologies prevent us from specifying the removed pattern

```
CGTGTTACTTAC
CGTGT    TAC
CGTGTT    AC
CGTG    TTAC
```

Individual reads covering the InDel may align in a different place.

InDels

We are generally much worse at calling somatic InDels than somatic SNVs

The problems we have with InDels include

- **InDels are often in repeat seqeunces**

The indel may simply be a change in the length of a polymer.

For example: CAG repeat lengths in the Androgen Receptor gene are of interest in prostate cancer.

Unfortunately, sequencers and aligners both tend to perform worse in such regions, and there tends to be natural variation in the population.

Establishing the ‘allele fraction’ of an InDel is particularly tricky.

InDel-calling tools

We are going to want a tool that performs assembly

I won’t list all of them here, nor is this an endorsement, but a few worth looking at follow. One key characteristic is what size (k) chunks of reads to use to try assembling what size region (g) the genome.

- **GATK HaplotypeCaller – tries larger k until resolution**
- **Platypus – assembly needs to be specified (g=1500 by default)**
- **Scalpel – self tuning to find k, g~600**
- **EBCall – performs well in recent review (PLoS ONE http://dx.doi.org/10.1371/journal.pone.0151664)**
- **MonoSeq – specifically calls changes in mononucleotide runs**

Here more than ever, it is likely that we will wish to combine several tools

Further Issues and Considerations

Multinucleotide Variants

MNVs

So far, multinucleotide variants are appearing in the vcf as multiple SNVs

- **The impact of MNVs is likely to be greater than the sum of the parts, meaning that annotations will be wrong**

- **Inference of characteristics such as signatures is likely to be biased if substantial numbers of MNVs are treated as independent SNVs**

- **Due to the filtering employed for SNVs, MNVs may be under-called**

- **They are unlikely just to be neighbouring SNVs.** With approximately 30,000 SNVs in a genome we would still only expect 1/3 cancers to show an MNV by chance.

- **MNVs may be interesting in their own right**

Expected MNVs

Some common and COSMIC-reported ones:

- **CA -> CG (transition) -> TG (methylation + deamination)**
- **CC>AA (Cosmic signature 4/29) – likely tobacco consequence?**
- **CC>TT (Cosmic signature 7) – UV light**
- **GA>TT, GC>AA – Pol Zeta errors (Harris et al. 2014)**

Some specific dinucleotide changes in KRAS and TP53

But need to know that neighbouring 'SNV's are on the same allele and not a double hit.

MNV tools

- **GATK –** can record directly in the vcf?
- **MAC (no not that one or that one) – Wei et al. 2015** – takes in BAM files and SNV calls and corrects the annotation of SNVs (using ANNOVAR, SnoEff or VEP)

Further Issues and Considerations

Single sample signatures

Single Sample Signatures

- **De novo discover of signatures is time-consuming**
- **De novo discover of signatures requires large amounts of data**
- **De novo discover of signatures is fairly robust (given those two) but new data could change the signatures assigned to previous sample**

If I have a sample of clinical interest…

- **I don’t want to have to wait until I have many more samples in order to derive signatures**
- **I don’t want to have to repeatedly rerun my analysis as I get more data (it is time-consuming and I don’t want the numbers to change)**
- **I don’t want to end up with signatures that I can’t interpret**

Single Sample Signatures

- I want to be able to decompose a vector of observations into the known signatures.

- The results will be stable – no need to rerun
- It could be done quickly
- Even if we don’t know the mechanism behind a signature, we know other samples in which it has been seen

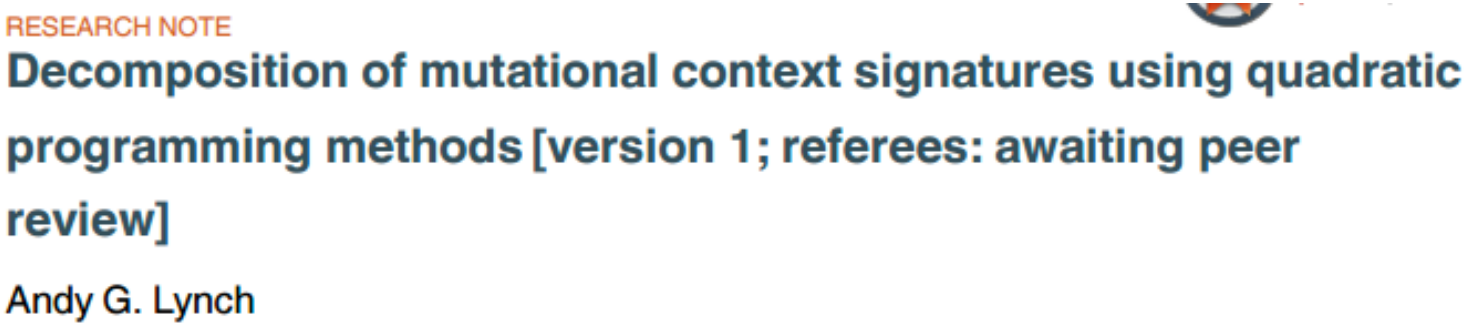
Single Sample Signature Tools

SOFTWARE Open Access

CrossMark

deconstructSigs: delineating mutational processes in single tumors distinguishes DNA repair deficiencies and patterns of carcinoma evolution

Rachel Rosenthal[1,2,4], Nicholas McGranahan[1,2,3], Javier Herrero[4*], Barry S. Taylor[5,6,7*] and Charles Swanton[1,2*]



RESEARCH NOTE

Decomposition of mutational context signatures using quadratic programming methods [version 1; referees: awaiting peer review]

Andy G. Lynch

Single Sample Signatures process

- **Annotate variants with mutational context**
- **Tabulate variant counts**
- **Find linear combination of signatures that best fits the variant counts**

Should we be normalizing for genome-wide context rates?

Estimating allele-frequencies

Validation of SNVs

We have an SNV call, and approximate allele frequency, but with a denominator of ~50 and 60% cellularity the precision of our estimate is tricky.

- **Clonality** – Precision of the AF estimate is required to determine whether the mutation is fully clonal (or even to determine the number of copies in each cell)

- **Phylogeny reconstruction** – similar to clonality

- **Monitoring** – If we want to monitor the presence of a mutation as a marker for the disease, we need precision to detect small changes

More precision = bigger denominator = throw sequencing at it

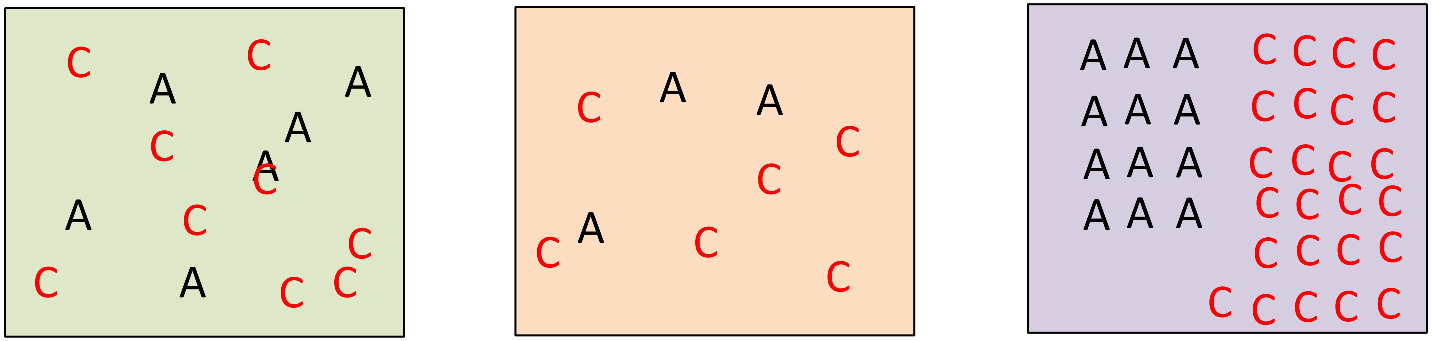
Further Issues and Considerations

Precision of estimates

Estimating allele-frequencies

Diminishing returns

There are sources of variation that extra sequencing can't help.



Tissue: 40% A

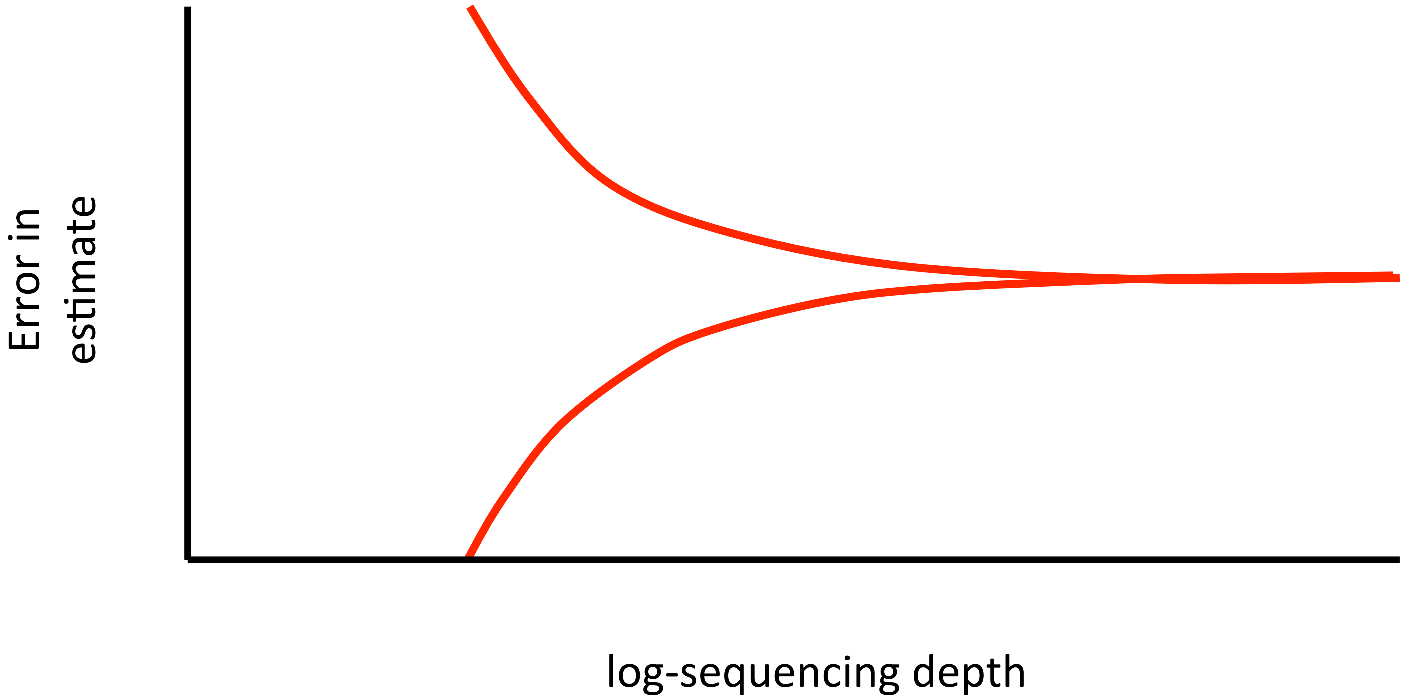
Sample: 33% A

Amplicon sequencing: approaching something close to 33%

Estimating allele-frequencies

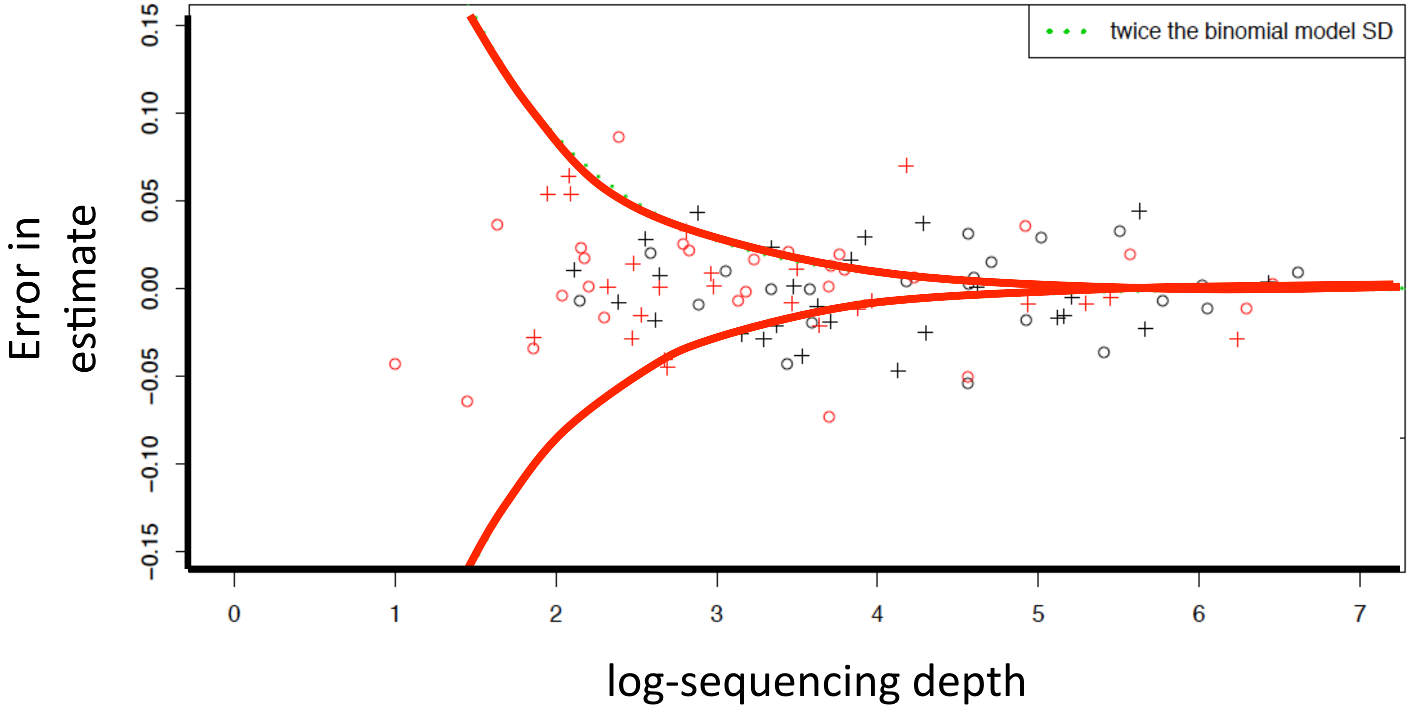
Synthetic experiment

From Daniel Andrew’s thesis (2015)



Estimating allele-frequencies

Precision doesn’t diminish with depth



References

- Krøigård A. B. et al. (2016). Evaluation of Nine Somatic Variant Callers for Detection of Somatic Mutations in Exome and Targeted Deep Sequencing Data. Plos One, 11(3), e0151664.
- Wei L. et al. (2015). MAC: identifying and correcting annotation for multi-nucleotide variations. BMC Genomics, 16(1), 569
- Rosenthal R. et al. (2016). deconstructSigs: delineating mutational processes in single tumors distinguishes DNA repair deficiencies and patterns of carcinoma evolution. Genome Biology, 17(1), 31.
- Lynch, A. G. (2016). Decomposition of mutational context signatures using quadratic programming methods. F1000Research, 5, 1253
- D. Andrews (2015) Statistical models of PCR for quantification of target DNA by sequencing. PhD Thesis, University of Cambridge