Towards Better Understanding of Artifacts in Variant Calling from High-Coverage Samples

by Heng Li (2014)

Benjamin L. Moore

8th June 2015

Aims

- How important is choice of aligner, variant caller and filtering steps?
- What are the sources of errors and disagreements?
- What's a reasonable estimate for the global error rates of variant calls?

Study design

Measure accuracy using real data rather than simulations

CHM1(hTERT)	**NA12878**
"Complete hydatidiform mole" cell line with haploid genome	Illumina platinum genomes (PCR free + deeply sequenced)

Handy in this case because heterozygous calls in CHM1 should (in theory) all be erroneous. . .

Study design

Read mapping:

- bowtie2
- bwa-backtrack
- bwa-mem

Variant callers:

- FreeBayes
- samtools
- UnifiedGenotyper
- HaplotypeCaller
- Platypus

Study design

Read mapping:

- `bowtie2`
- `bwa-backtrack`
- `bwa-mem`

Variant callers:

- `FreeBayes`
- `samtools`
- `UnifiedGenotyper`
- `HaplotypeCaller`
- `Platypus`

Broad comparison of popular tools but doesn't investigate:

- Aligner and variant caller parameters
- Pragmatic conerns: throughput, compute resources

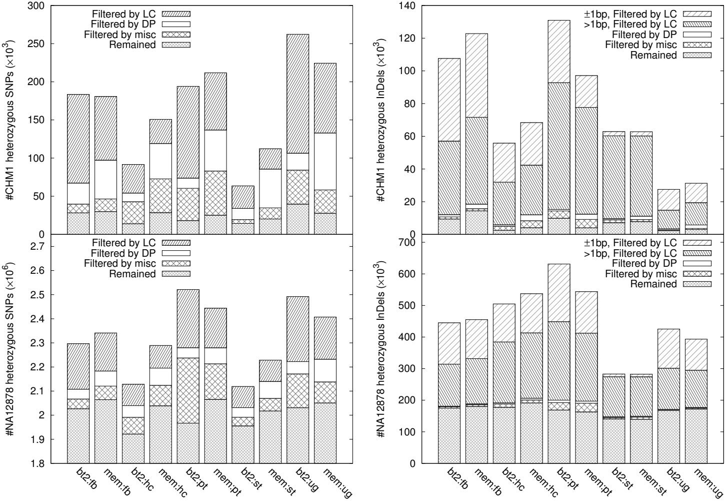
Variant filtering

Compare“universal filters”, i.e. not those embedded in callers:

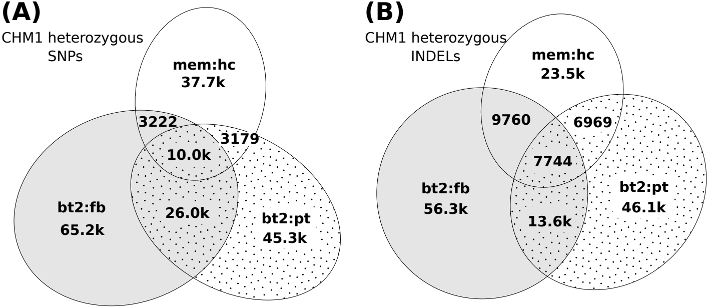
1. **Low complexity**: remove vars in LCRs*
2. **Max-depth**: filter if suspiciously high coverage
3. **Allele balance**: filter if not roughly 1 or .5
4. **Double strand**: var should be represented on both strands
5. **Fisher strand**: reference/non- match forward/reverse
6. **Quality**: threshold by reported variant quality

*alignment and caller independent

Low complexity, max depth filters ++effective

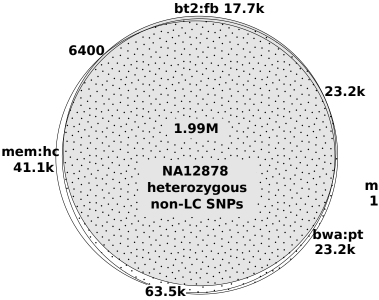


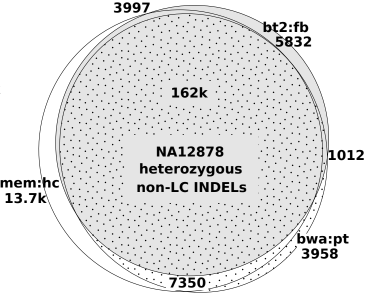
Inconsistencies suggest non-biological errors



If problems were with ploidy or mutations, we'd expect more agreement between aligners + callers.

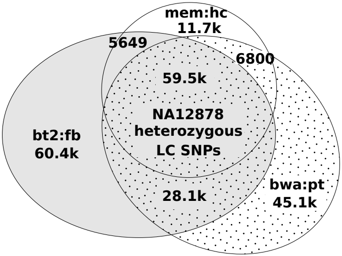
Methods agree in diploid line

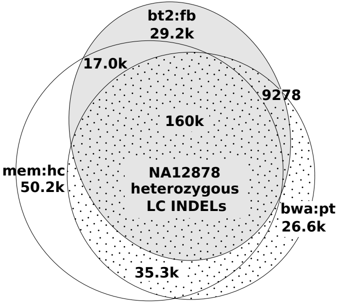




Low-hanging fruit + well-developed algorithms

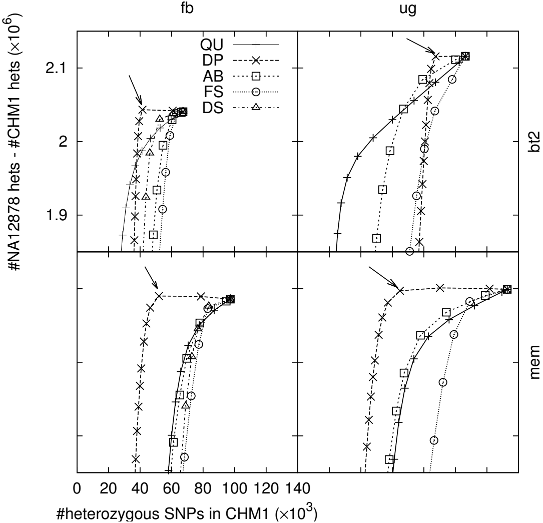
. . . but not in low-complexity regions





Maybe variants in LC regions should be ignored until methods improve, or can be resolved with long-read tech

ROC-ish plot



≈FP on x-axis
≈TP on y-axis

Again max-depth stands out, optimally: DP $< d + [3\sqrt{d}, 4\sqrt{d}]$

Investigating problematic regions

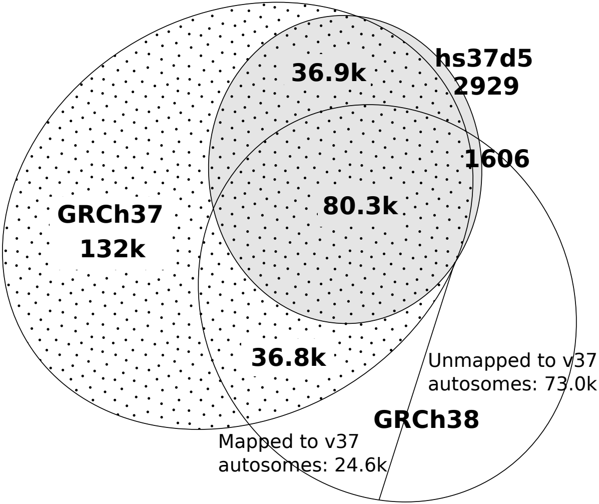
Interesting to look at where things are going wrong and why

```
                  11111111112222222222333333333344444444445555      55555566666666667777777777888888888899999999990000000000111111111
Position:12345678901234567890123456789012345678901234567890123      45678901234567890123456789012345678901234567890123456789012345678

     Ref:ATTTGGGGGCTGGGACTGGGTCCAGGACAGGGACTGGGGCCGGGACCGGGACC******GGGACTGGGGCCGGGACCGGGACCGGGACTGGGGCCGGGACCGGGACCGGGACAGGGACCAGGAC

   Truth:ATTTGGGGGCTGGGACTGGGTCCGGGACAGGGACTGGGGCCGGGA--------******----------CCGGGACCGGGACAGGGACTGGGG------CCGGGACCGGGACAGGGACCAGGAC

errRead1:ATTTGGGGGCTGGGACTGGGTCCgGGACAGGGACTGGGGCCGGGACCGGGACC******GGGAC
errRead2:               CTGGGTCCgGGACAGGGACTGGGGCCGGGACCGGGACCgggacaGGGACTGGGGCCGGGACCGGGACaGGGAC
errRead3:                                        TGGGtCCGGGACa******GGGACTGGGGCCGGGACCGGGACcGGGACaGGGaCtGGGgCCGGGACCGGGACAGGGACCAGGAC

Correct1:ATTTGGGGGCTGGGACTGGGTCCgGGACAGGGACTGGGGCCGGGA--------******----------CCGGGACCGGGAC
Correct2:               CTGGGTCCgGGACAGGGACTGGGGCCGGGA--------******----------CCGGGACCGGGACaGGGACTGGGG------CCGGGACCGGGACAGGGAC
Correct3:                TGGGTCCGGGACAGGGACTGGGGCCGGGA--------******----------CCGGGACCGGGACaGGGACTGGGG------CCGGGACCGGGACAGGGACCAGGAC
```

Here mapping errors lead to variant calls instead of recognising insertion (over-penalising gap extension?)

Example of where assembling reads can help (HaplotypeCaller)

Genome build matters



Headline statistics

1. **Raw variant calls**: 1 error per 10-15 kb
2. **After filtering**: 1 error per 100-200 kb

Headline statistics

1. **Raw variant calls**: 1 error per 10-15 kb
2. **After filtering**: 1 error per 100-200 kb

. . . confirmatory.
Matches estimates by Bentley *et al.* (2008) and Nickles *et al.* (2012).

Sources of errors

Largest sources of error:

1. Low complexity regions, incl. caller realignments
2. Incomplete reference genome

Sources of errors

Largest sources of error:

1. Low complexity regions, incl. caller realignments
2. Incomplete reference genome

Read assembly can help with both: long synthetic reads can bridge low complexity regions and can be assembled *de novo*, independent of reference.

Advised best practices

Now:

Run ≥two pipelines, take intersection of raw calls and apply universal filters

Future:

De novo assembly using long reads (PacBio, ONT or something like Moleculo/TruSeq Synthetics)

Map to multiple possible genotypes instead of a single reference