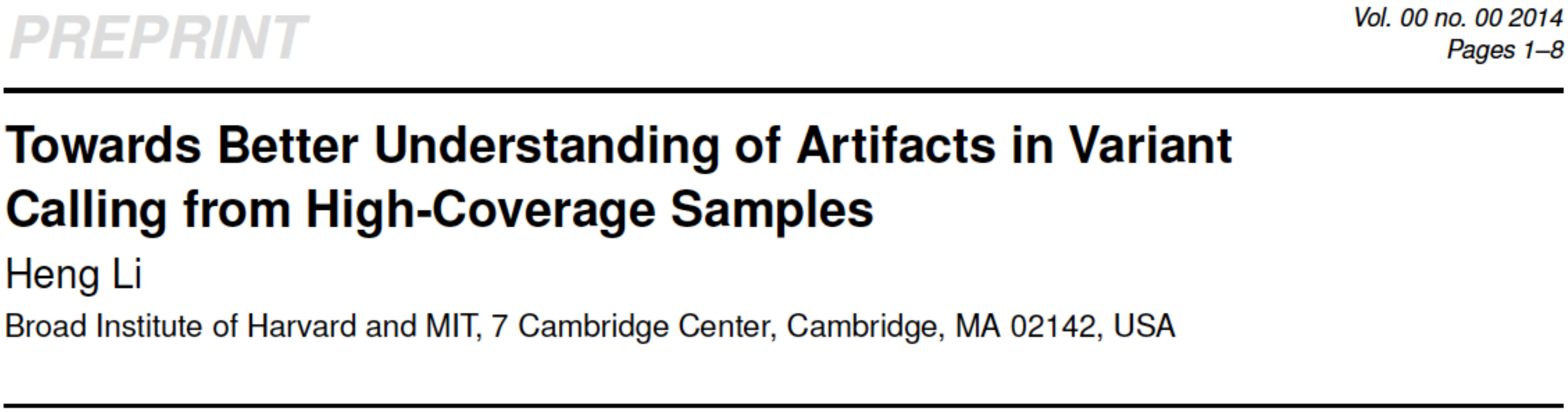
Towards Better Understanding of Artifacts in Variant Calling from High-Coverage Samples

Lyon lab journal club

Han Fang

4/7/2014



PREPRINT

Vol. 00 no. 00 2014
Pages 1–8

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

Heng Li
Broad Institute of Harvard and MIT, 7 Cambridge Center, Cambridge, MA 02142, USA

Previous works

- **Simulate variants** and reads from a reference genome (Li et al.,2008)
- **Sequence a small target region with mature technologies**, and take the resultant sequence as the ground truth (Harismendy et al., 2009)
- **Compare variant calls from different pipelines**, or by comparing calls to variants ascertained with array genotyping or in another study (Clark et al., 2011; Li, 2012; Lam et al., 2012a,b; Boland et al., 2013; Liu et al., 2013; Goode et al., 2013; O'Rawe et al., 2013; Zook et al., 2014; Cheng et al., 2014).
- In the author's view, it is better to evaluate variant calling by **comparing samples from a pedigree** (Zook et al., 2014), or **from the same individual** (Nickles et al., 2012), including cancer samples (L¨ower et al., 2012).

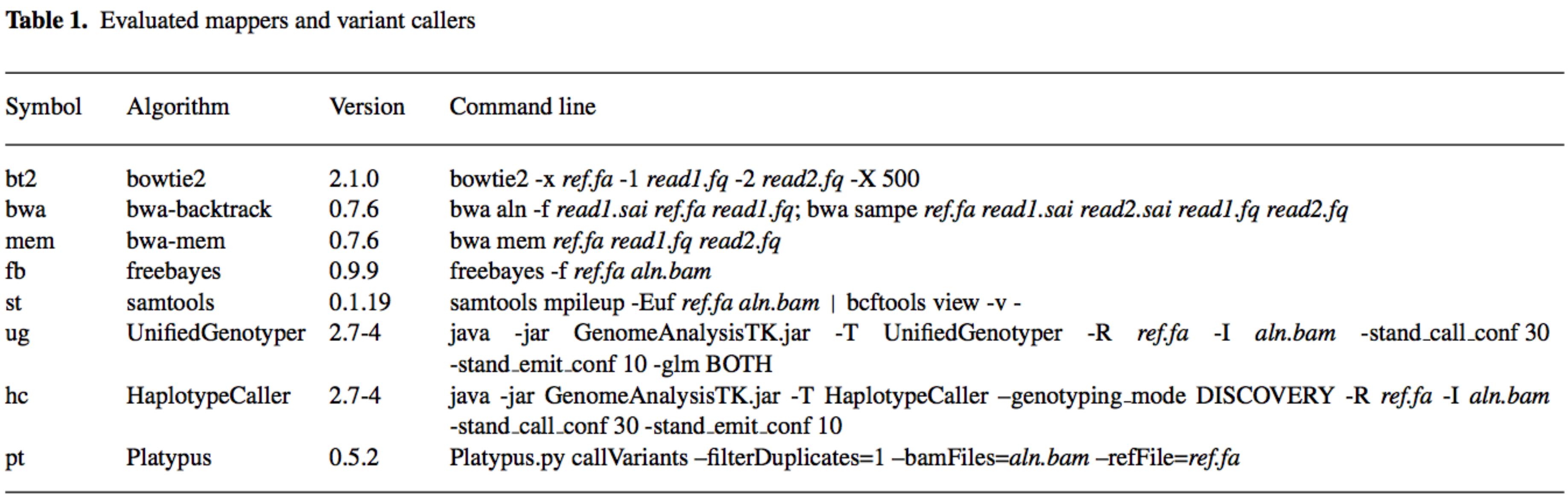


Table 1. Evaluated mappers and variant callers

Symbol	Algorithm	Version	Command line
bt2	bowtie2	2.1.0	bowtie2 -x *ref.fa* -1 *read1.fq* -2 *read2.fq* -X 500
bwa	bwa-backtrack	0.7.6	bwa aln -f *read1.sai ref.fa read1.fq*; bwa sampe *ref.fa read1.sai read2.sai read1.fq read2.fq*
mem	bwa-mem	0.7.6	bwa mem *ref.fa read1.fq read2.fq*
fb	freebayes	0.9.9	freebayes -f *ref.fa aln.bam*
st	samtools	0.1.19	samtools mpileup -Euf *ref.fa aln.bam* | bcftools view -v -
ug	UnifiedGenotyper	2.7-4	java -jar GenomeAnalysisTK.jar -T UnifiedGenotyper -R *ref.fa* -I *aln.bam* -stand_call_conf 30 -stand_emit_conf 10 -glm BOTH
hc	HaplotypeCaller	2.7-4	java -jar GenomeAnalysisTK.jar -T HaplotypeCaller –genotyping_mode DISCOVERY -R *ref.fa* -I *aln.bam* -stand_call_conf 30 -stand_emit_conf 10
pt	Platypus	0.5.2	Platypus.py callVariants –filterDuplicates=1 –bamFiles=*aln.bam* –refFile=*ref.fa*

Variant filtering

- Low-complexity filter (LC)
- Maximum depth filter (DP)
- Allele balance filter (AB)
- Double strand filter (DS)
- Fisher strand filter (FS)
- Quality filter (QU)

Measuring accuracy

- Assumptions:

1) Similar coverage.
2) # called variants per haplotype is very close.
3) # hets calls in CHM1 ≈ # hets errors in NA12878

- N_h - # hets calls in CHM1. (Negative control)
- N_d - # hets calls in NA12878. (Positive control)
- False positive rate: N_h / N_d
- Sensitivity: $N_d - N_h$

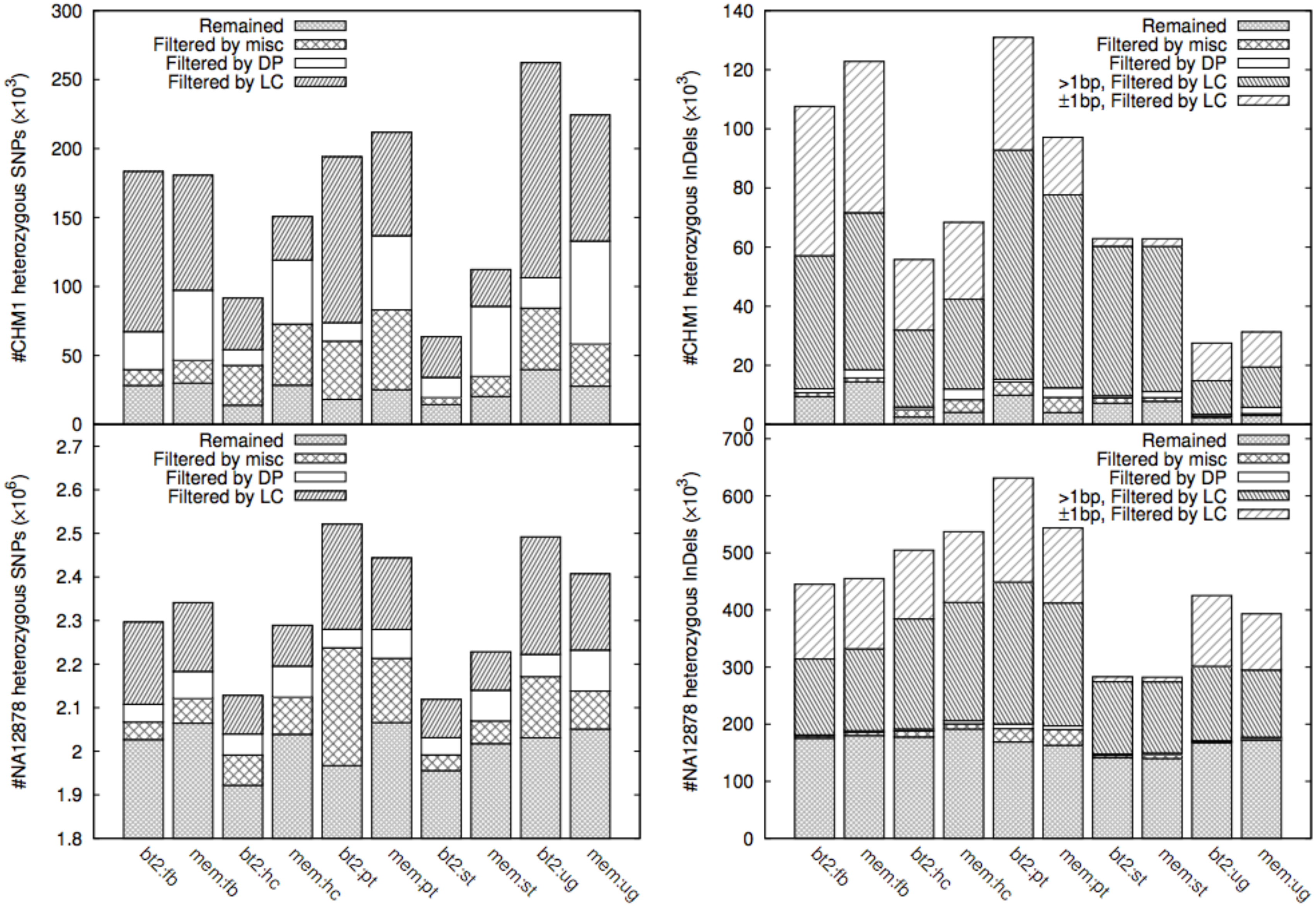
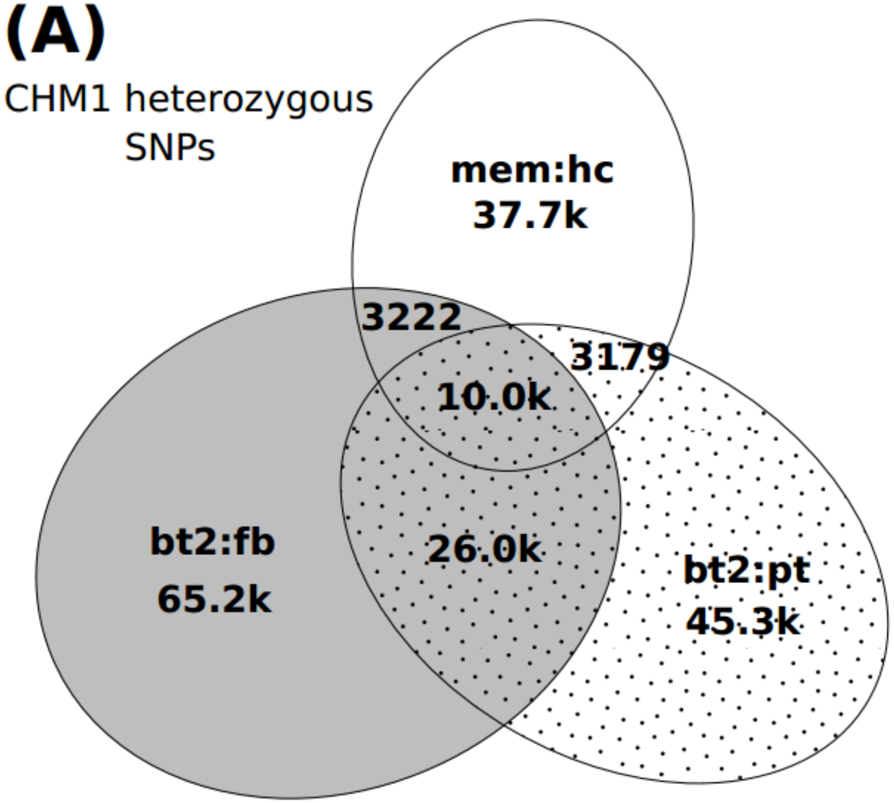
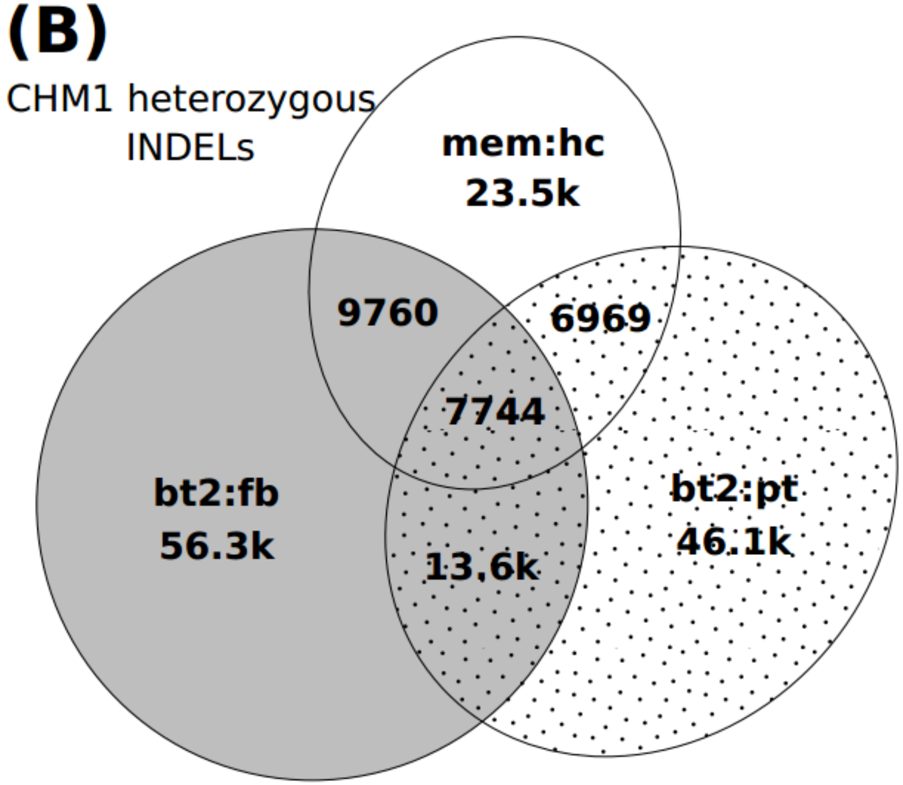


Fig. 1. Effect of filters. Low-complexity filter (LC): not overlapping low-complexity regions identified by the DUST algorithm. Maximum-depth filter (DP): read depth below $d + 3\sqrt{d}$, where d is the average read depth. Miscellaneous filter (misc) includes three filters: allele balance above 30%, variants supported by non-reference reads on both strands and Fisher strand P-value is above 0.01. Filters are applied in the order of LC, DP and misc, with DP applied to variants passing LC, and misc applied to variants passing both LC and DP. For each call set, the total height of the bar gives the number of raw variant calls with the reported quality in VCF no less than 30.


(A)
CHM1 heterozygous SNPs
mem:hc
37.7k
3222
3179
10.0k
bt2:fb
65.2k
26.0k
bt2:pt
45.3k


(B)
CHM1 heterozygous INDELs
mem:hc
23.5k
9760
6969
7744
bt2:fb
56.3k
13.6k
bt2:pt
46.1k

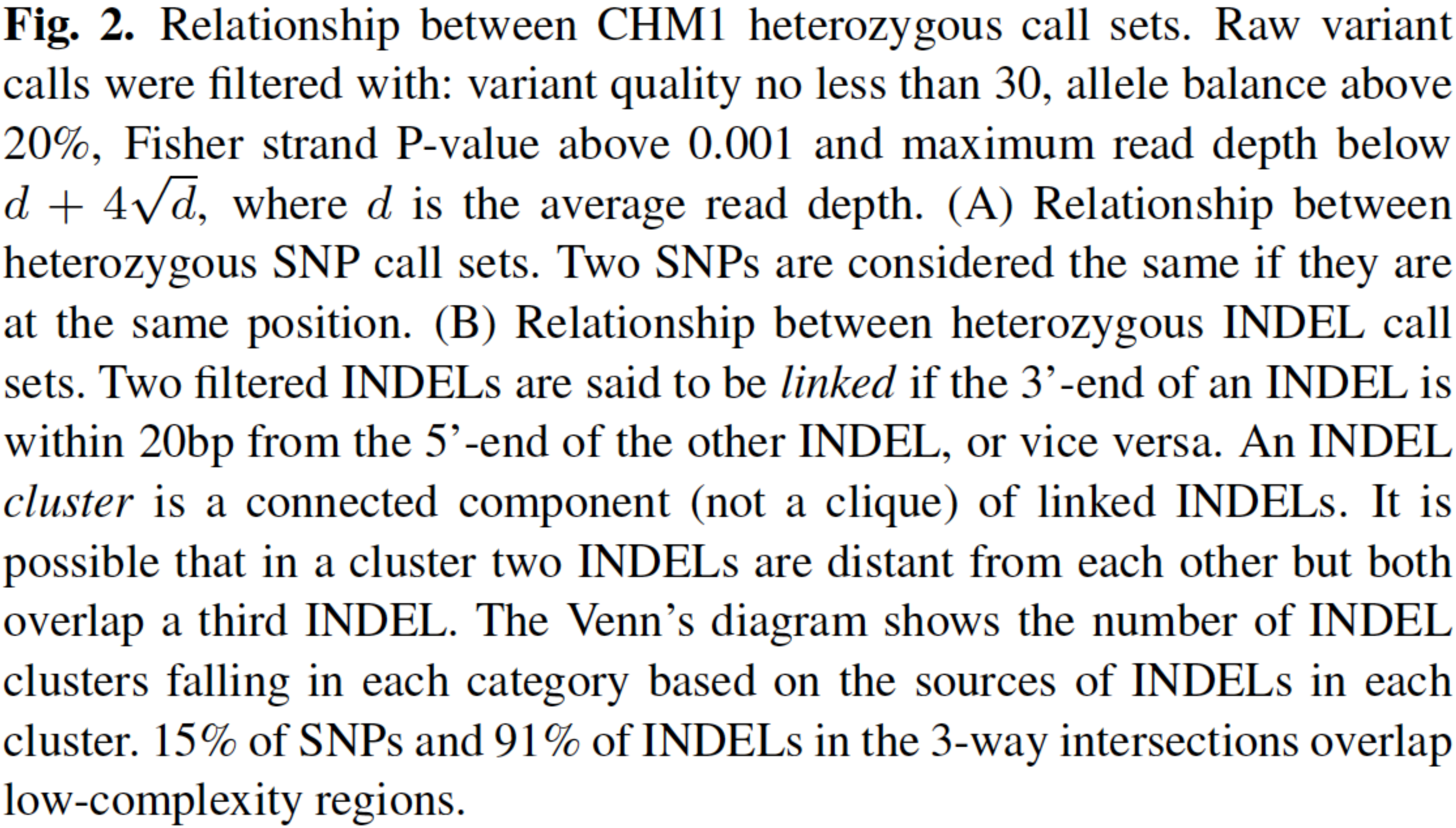


Fig. 2. Relationship between CHM1 heterozygous call sets. Raw variant calls were filtered with: variant quality no less than 30, allele balance above 20%, Fisher strand P-value above 0.001 and maximum read depth below $d + 4\sqrt{d}$, where d is the average read depth. (A) Relationship between heterozygous SNP call sets. Two SNPs are considered the same if they are at the same position. (B) Relationship between heterozygous INDEL call sets. Two filtered INDELs are said to be *linked* if the 3'-end of an INDEL is within 20bp from the 5'-end of the other INDEL, or vice versa. An INDEL *cluster* is a connected component (not a clique) of linked INDELs. It is possible that in a cluster two INDELs are distant from each other but both overlap a third INDEL. The Venn's diagram shows the number of INDEL clusters falling in each category based on the sources of INDELs in each cluster. 15% of SNPs and 91% of INDELs in the 3-way intersections overlap low-complexity regions.

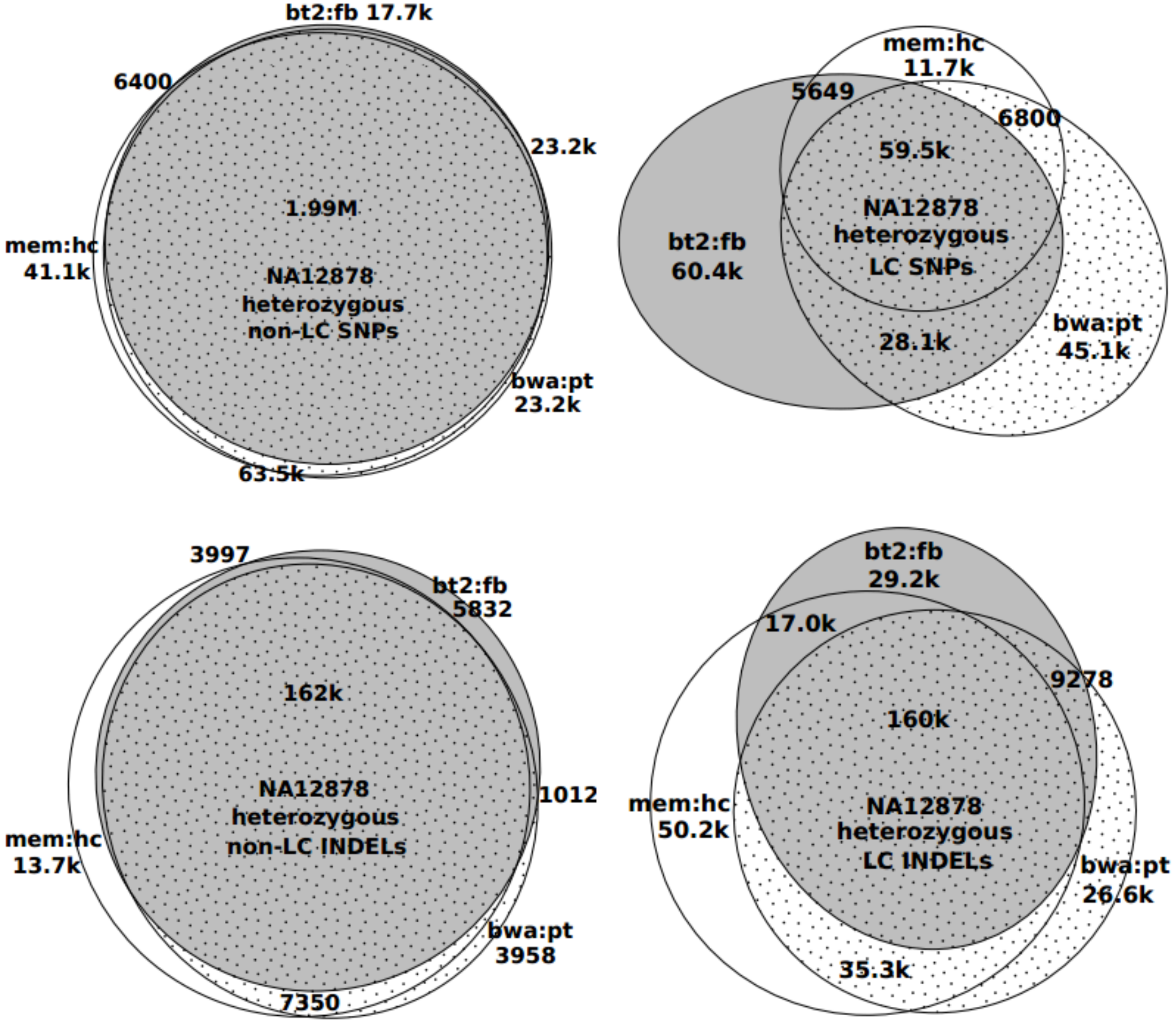
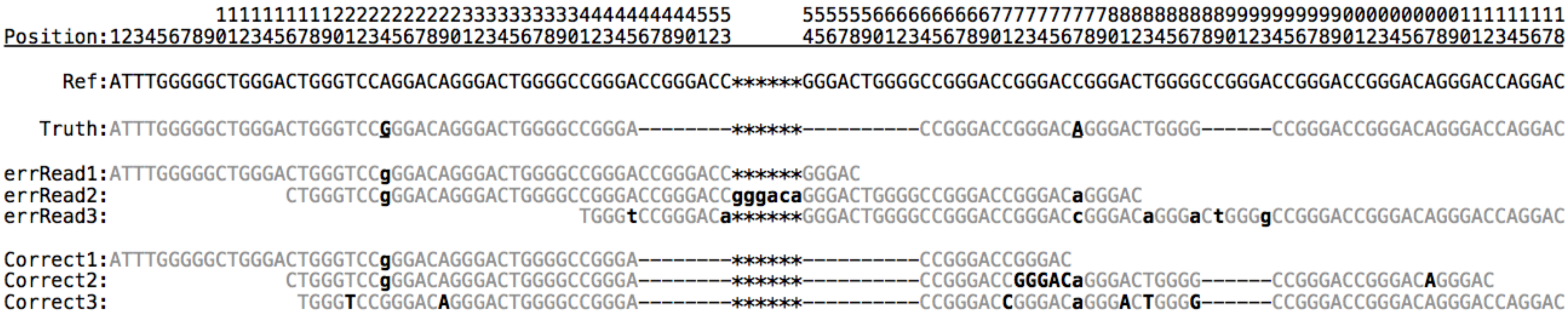
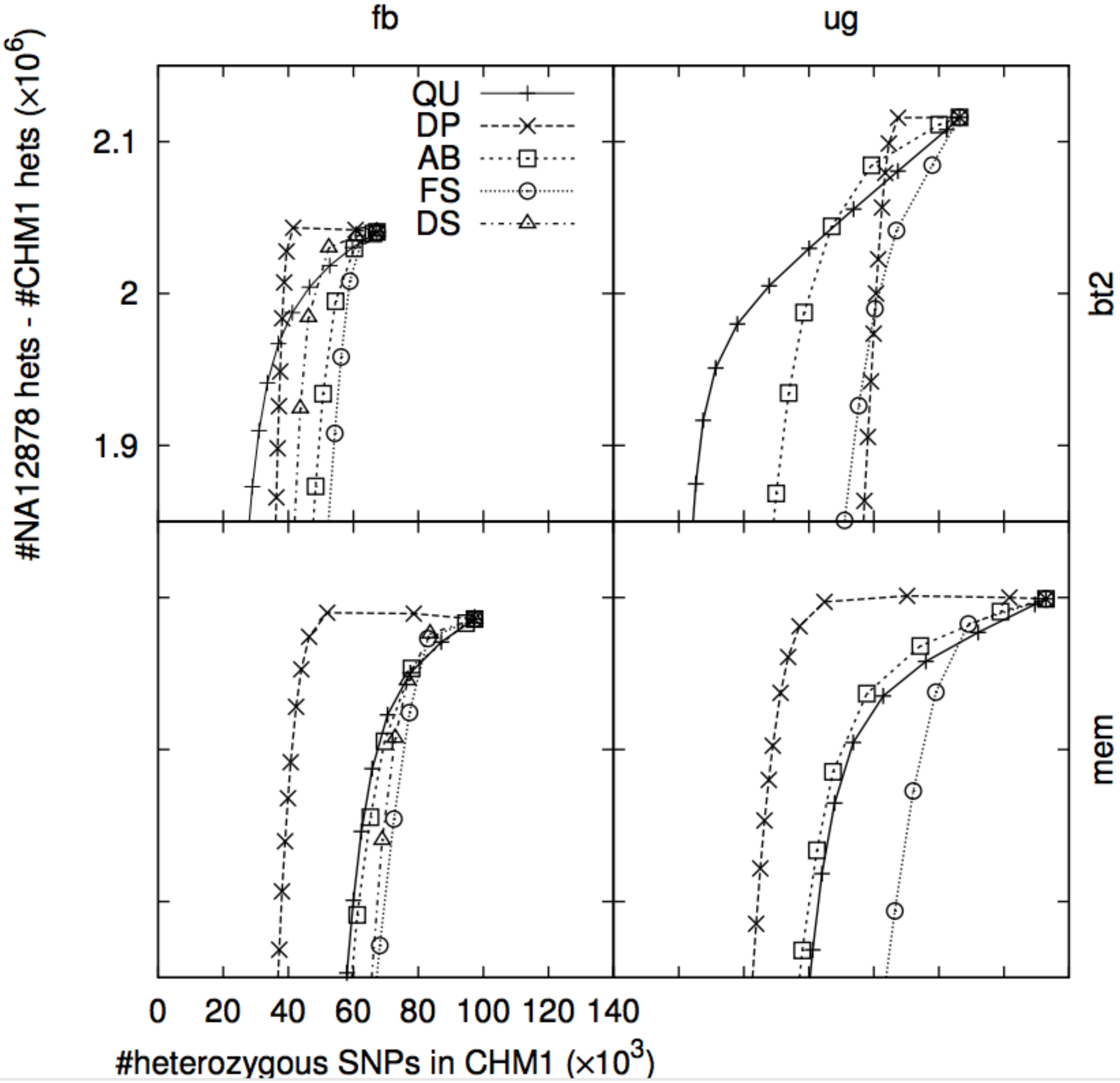


Fig. 3. Relationship between NA12878 heterozygous call sets.



```
                  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
```

Fig. 4. Example of misalignment around chr1:26608841 in CHM1. The truth allele is derived from local assembly. Three erroneous read alignments and their correct alignments are shown below it. Each of the three reads is an exact substring of the truth allele, but their alignments are different. The first read 'errRead1' is aligned without gaps as the 3'-end of the read is a substring of the 18bp deletion. Read 'errRead2' is aligned with a 6bp insertion as this alignment is better than having two long deletions. Read 'errRead3' is also aligned without gaps but with seven mismatches. It is possible for an aligner to find its correct alignment given a small gap extension penalty. On this example, Bowtie2 did not align any reads with gaps. BWA-MEM aligned four reads correctly. Except HaplotypeCaller which locally assembled reads, other callers all called multiple heterozygotes around this region.


fb
ug
bt2
mem
QU
DP
AB
FS
DS
#NA12878 hets - #CHM1 hets (×10⁶)
2.1
2
1.9
0
20
40
60
80
100
120
140
#heterozygous SNPs in CHM1 (×10³)

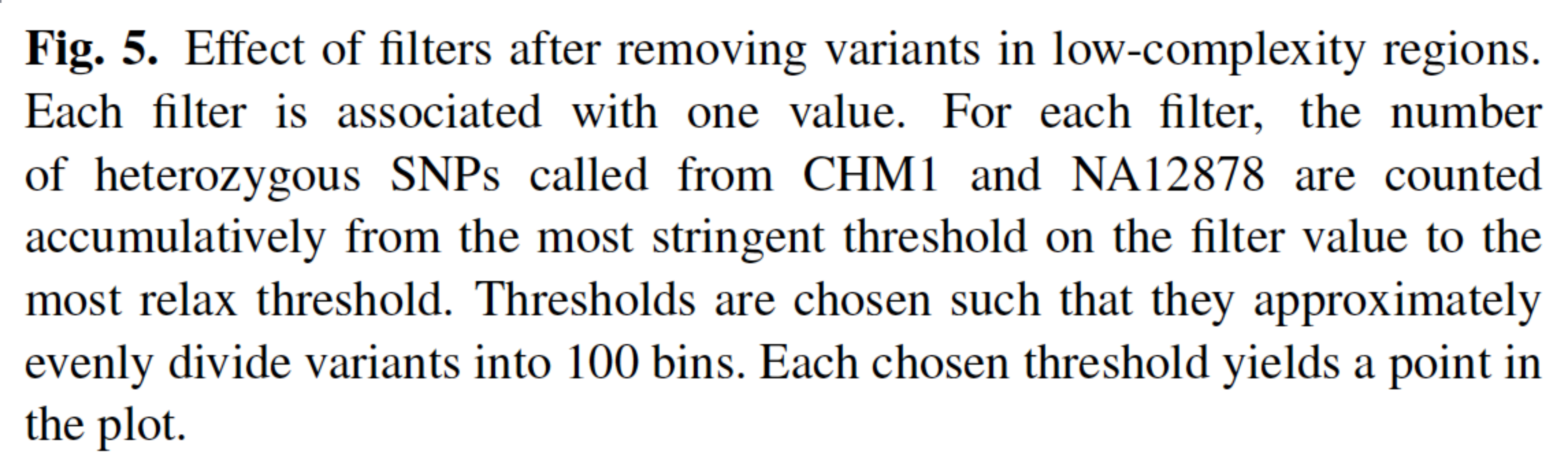


Fig. 5. Effect of filters after removing variants in low-complexity regions. Each filter is associated with one value. For each filter, the number of heterozygous SNPs called from CHM1 and NA12878 are counted accumulatively from the most stringent threshold on the filter value to the most relax threshold. Thresholds are chosen such that they approximately evenly divide variants into 100 bins. Each chosen threshold yields a point in the plot.

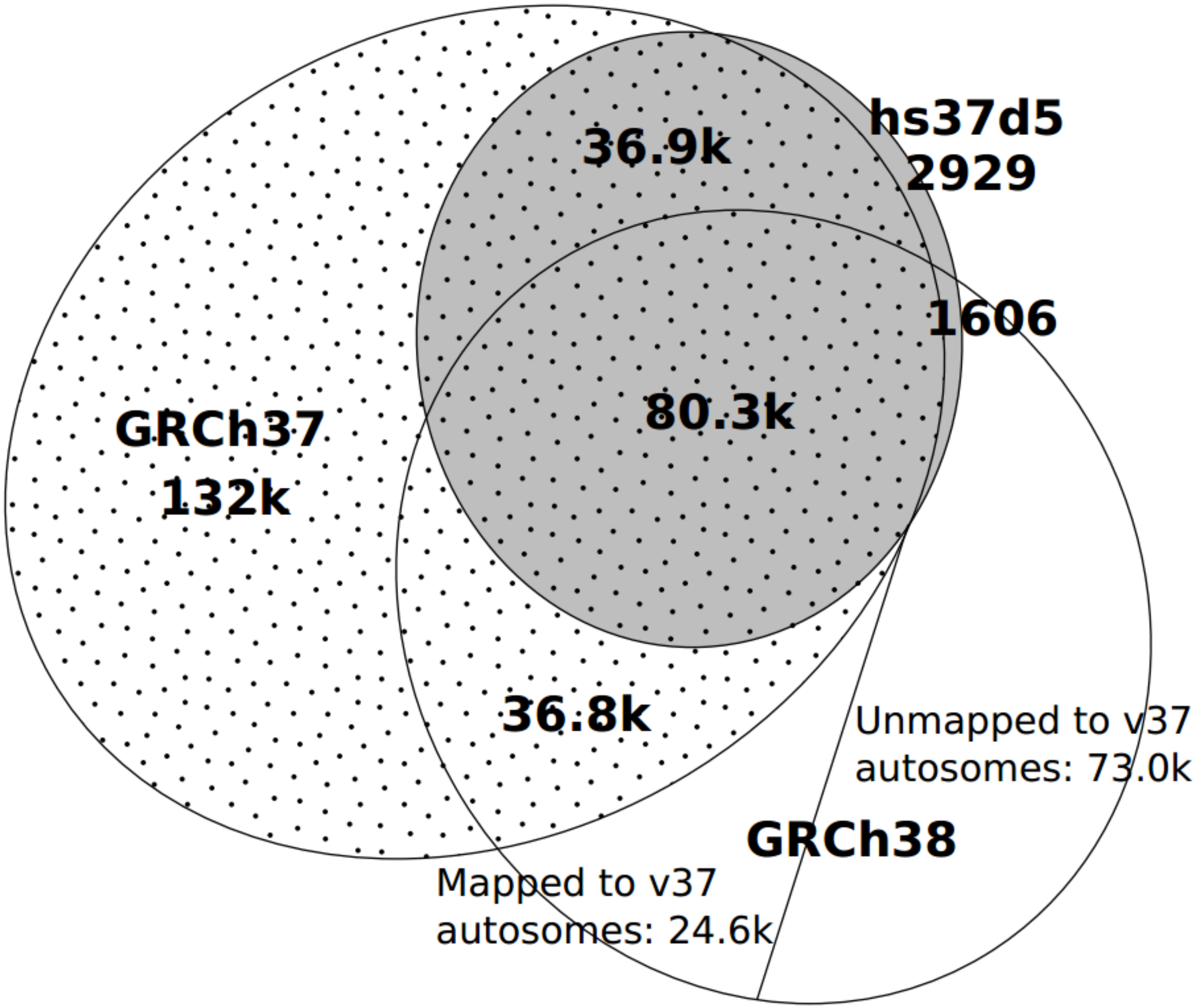


Fig. 6. Relationship of CHM1 heterozygous SNPs called from mappings to different reference genomes. CHM1 reads were mapped with BWA-MEM. Autosomal SNPs were called with GATK HaplotypeCaller and passed the low-complexity filter. Heterozygous calls from GRCh38 were lifted to GRCh37 with the liftOver tool from UCSC under the default setting.

Summary

- Low complexity is the most effective against false hets; filter out them all.
- Realignment in LCRs needs improvements.
- Apply the same filters will make call-set agree better outside LCRs.
- Error rate: 1 per 100-200kb.
- Take the intersection of raw calls and apply caller-oblivious filters.