# MPG NGS workshop: Read-backed haplotype phasing

Menachem Fromer

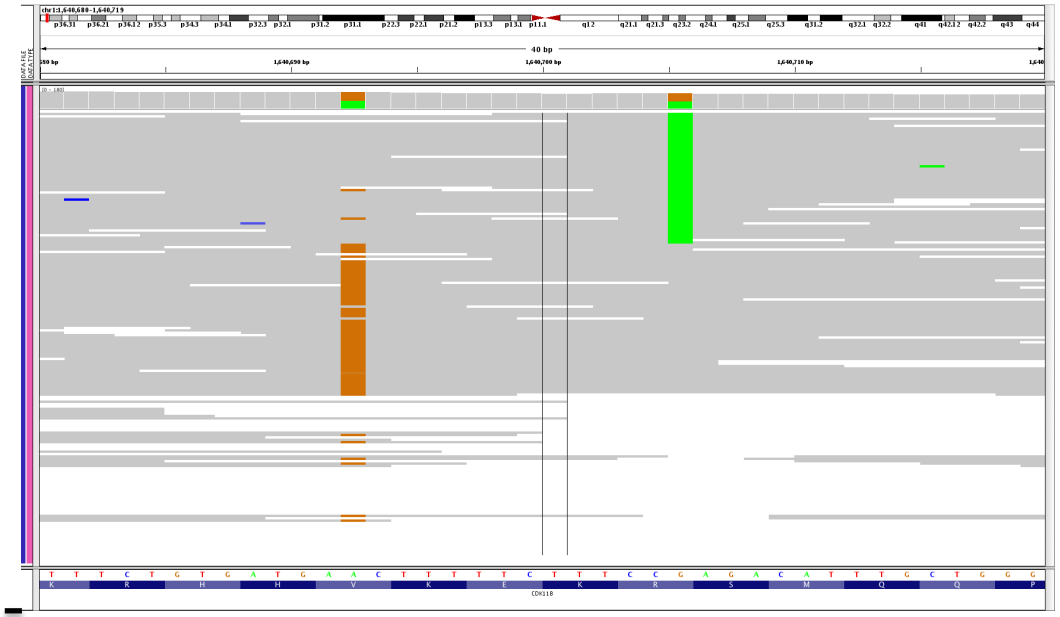
Stanley Center for Psychiatric Research
Genome Sequencing and Analysis, Medical and Population Genetics
Broad Institute of Harvard and MIT

2/17/11

# Local haplotype phasing

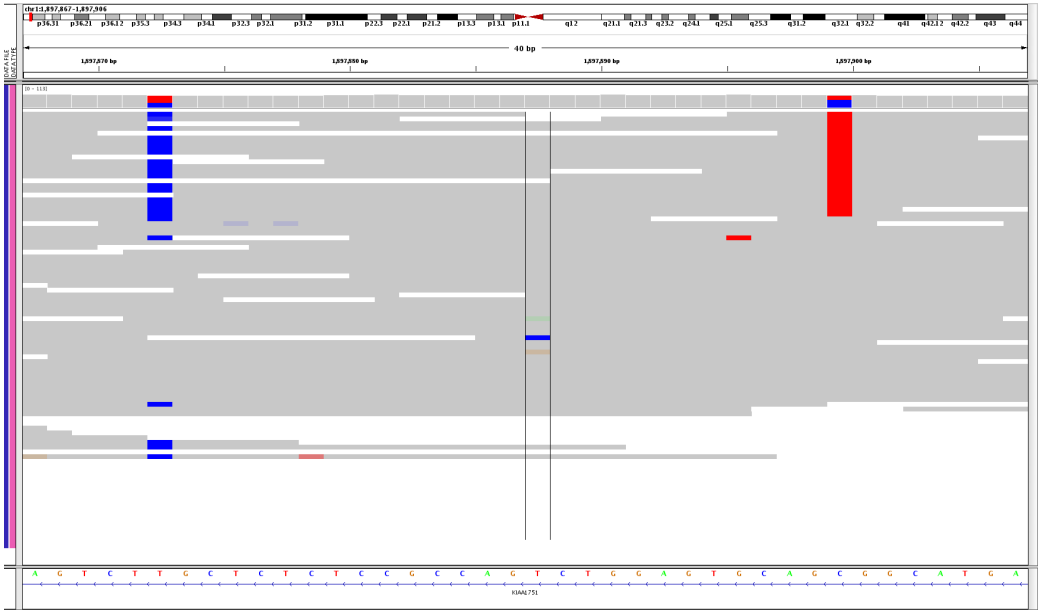
Compound heterozygote





Multi-nucleotide polymorphism





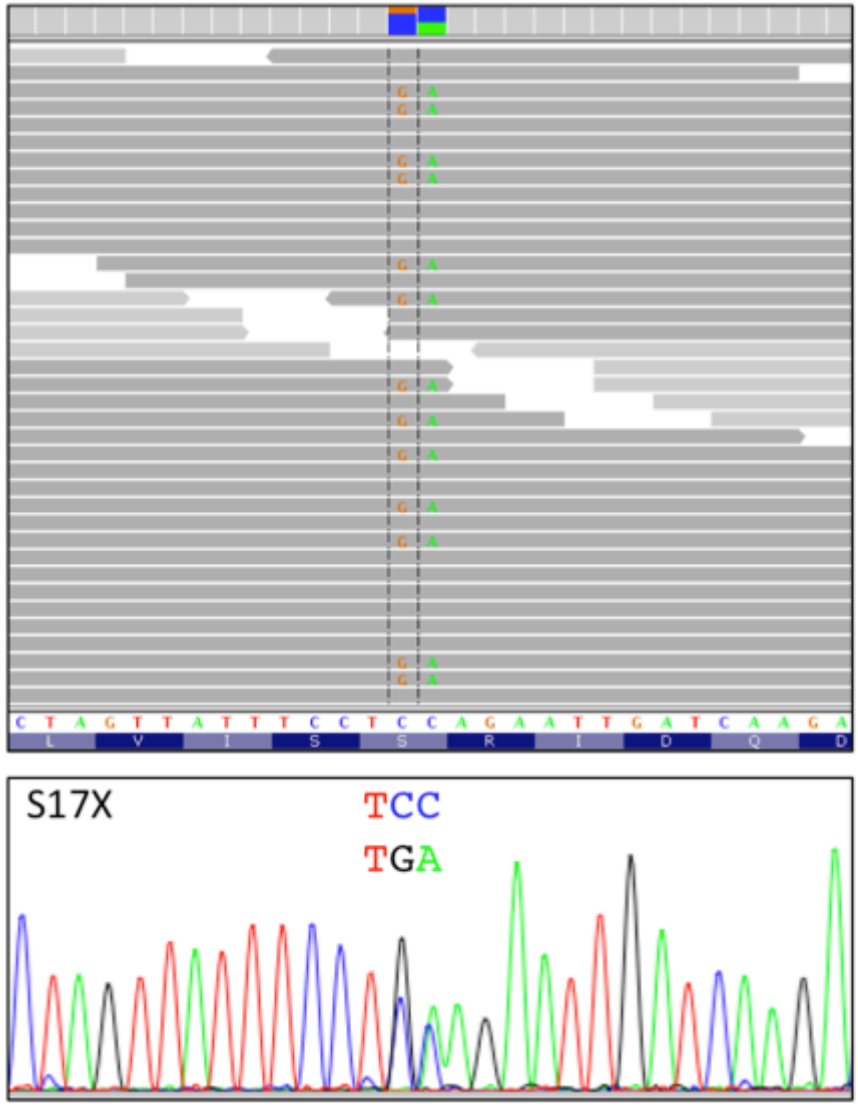
# Disease-associated dinucleotide polymorphism underscores importance of resolving phase

•Exome sequencing of Hypolipidemia patients uncovered nonsense mutation in ANGPTL3 (angiopoietin-like 3 protein)

**If *improperly* taken as two separate SNPs:**
TCC (Serine) -> TCA (Serine), (synonymous)
TCC (Serine) -> TGC (Cysteine), (missense)

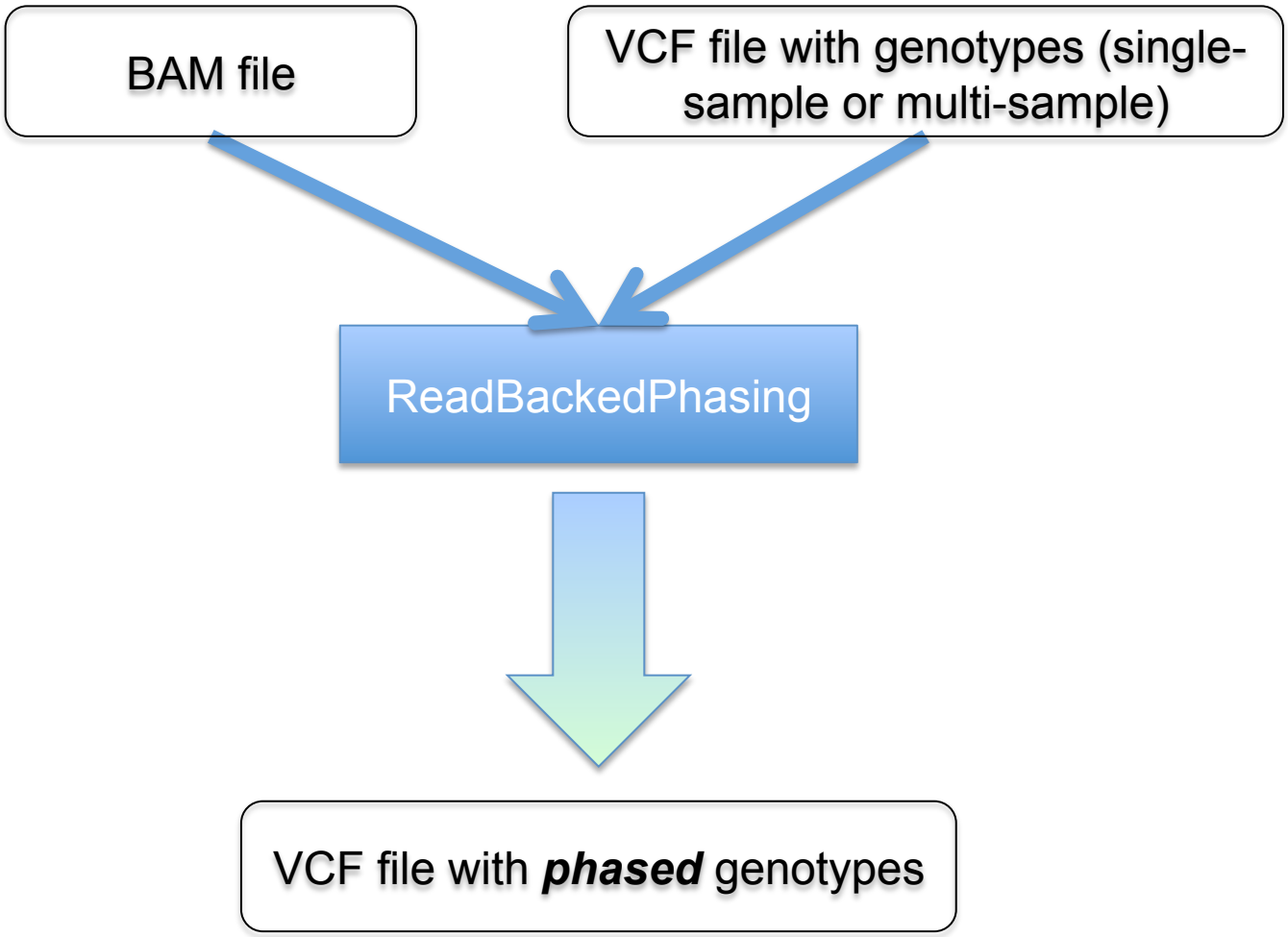
**If *properly* taken as a single DNP:**
TCC (Serine) -> TGA (Opal), (nonsense)

**Proper handling requires examining haplotype information contained in reads.**




K. Musunuru, J. Pirruccello, R. Do et al., NEJM 2010

# Read-backed phasing process



# Running ReadBackedPhasing

```
java -Xmx4g -jar GenomeAnalysisTK.jar
  -T ReadBackedPhasing
  -R Homo_sapiens_assembly19.fasta
  -I reads.bam
  -B:variant,VCF genotype_calls.vcf
  --phaseQualityThresh 20
  -o phased_genotype_calls.vcf
```

Reads to be used in phasing calculation

Genotype calls to be phased

Phasing quality threshold

Phased VCF file (phased_genotype_calls.vcf)

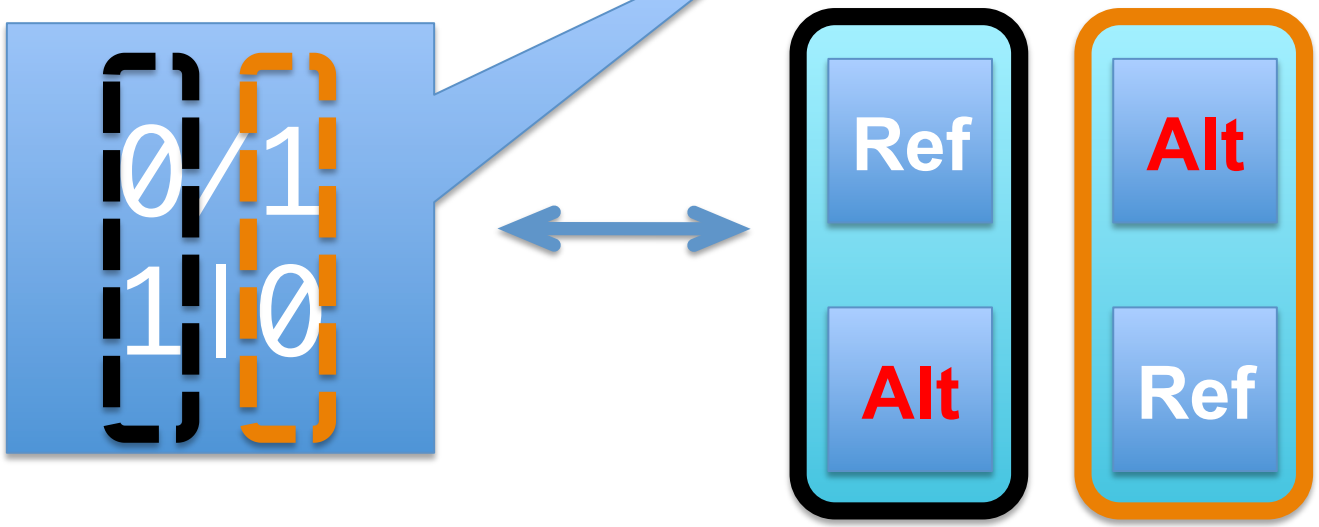
```
#CHROM POS  ID    REF   ALT   QUAL  FILTER INFO FORMAT       SAMP
chr1   1    .     A     G     99    PASS   .    GT:GL:GQ     0/1:-100,0,-100:99
chr1   2    .     A     C     99    PASS   .    GT:GL:GQ:PQ  1|0:-100,0,-100:99:60
chr1   3    .     G     A     99    PASS   .    GT:GL:GQ:PQ  0|1:-100,0,-100:99:70
```

See http://www.broadinstitute.org/gsa/wiki/index.php/Read-backed_phasing_algorithm for more info

# Interpreting phased VCF

- 0/1 = unphased het genotype
- 0|1 = phased het genotype
  - Relative to previous genotype

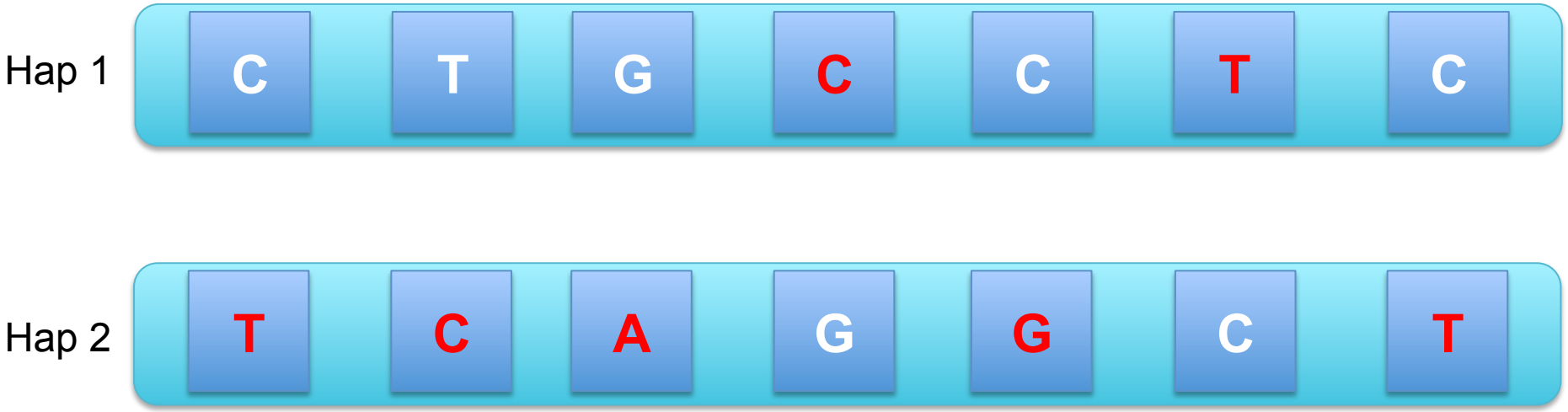
```
#CHROM POS  ID    REF   ALT   QUAL  FILTER INFO FORMAT       SAMP
chr1   1    .     A     G     99    PASS   .    GT:GL:GQ     0/1:-100,0,-100:99
chr1   2    .     A     C     99    PASS   .    GT:GL:GQ:PQ  1|0:-100,0,-100:99:60
chr1   3    .     G     A     99    PASS   .    GT:GL:GQ:PQ  0|1:-100,0,-100:99:70
```



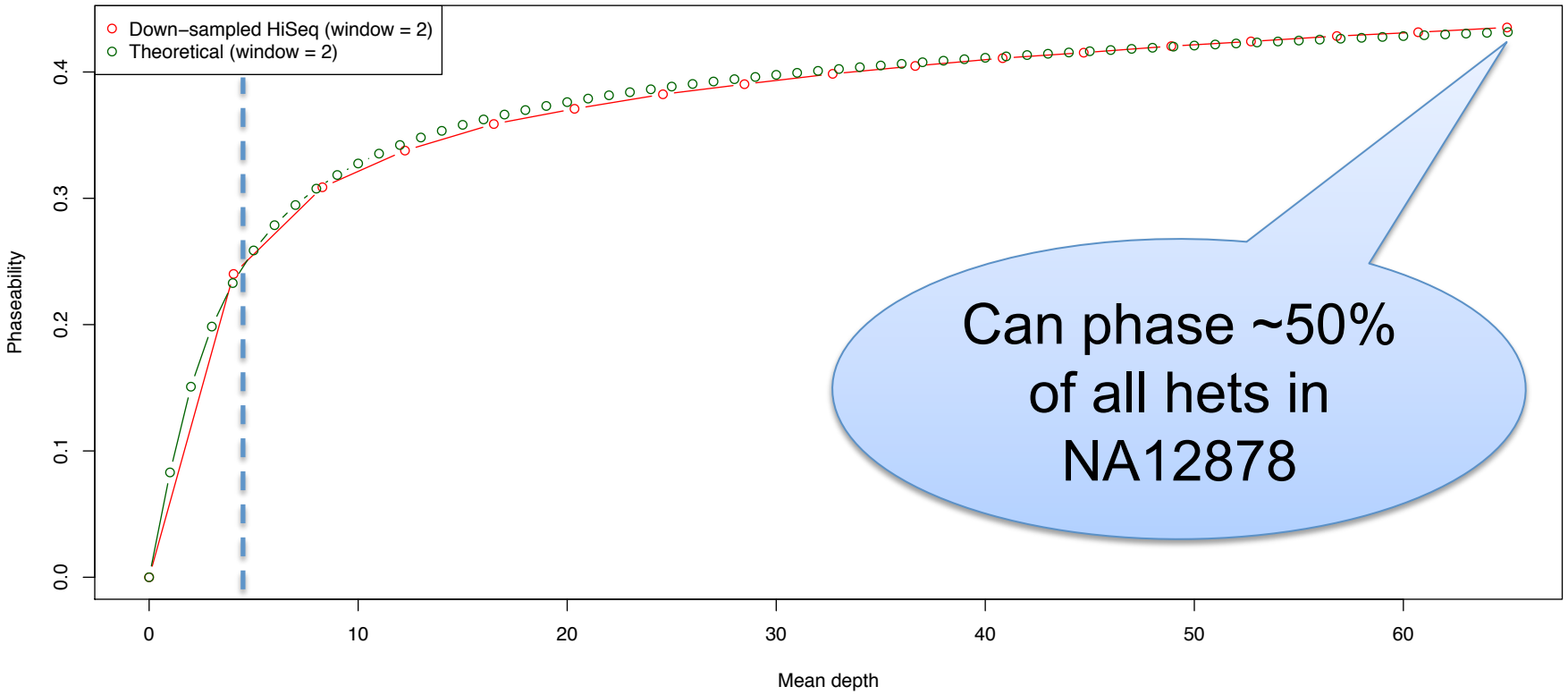
# Chaining phased alleles to yield local haplotype structures

| #CHROM | POS | ID | REF | ALT | QUAL | FILTER | INFO | FORMAT | NA12878 |
|---|---|---|---|---|---|---|---|---|---|
| chr1 | 46501 | rs2691308 | C | T | 220.46 | PASS | . | GT:AD:DP:GQ:PL | 0/1:55,13:65:99:252,0,255 |
| chr1 | 46896 | rs2691311 | T | C | 494.56 | PASS | . | GT:AD:DP:GQ:PL:PQ | 0\|1:55,21:75:99:255,0,255:195.32 |
| chr1 | 46927 | rs2548884 | G | A | 117.24 | PASS | . | GT:AD:DP:GQ:PL:PQ | 0\|1:53,8:60:99:148,0,255:1320.50 |
| chr1 | 47021 | rs6658003 | G | C | 81.72 | PASS | . | GT:AD:DP:GQ:PL:PQ | 1\|0:36,19:6:99:113,0,112:24.75 |
| chr1 | 47109 | rs2691313 | C | G | 348.47 | PASS | . | GT:AD:DP:GQ:PL:PQ | 0\|1:38,13:51:99:255,0,255:24.75 |
| chr1 | 47155 | . | C | T | 1283.62 | PASS | . | GT:AD:DP:GQ:PL:PQ | 1\|0:34,39:60:99:255,0,255:1035.35 |
| chr1 | 47239 | rs2854673 | C | T | 246.16 | PASS | . | GT:AD:DP:GQ:PL:PQ | 0\|1:39,14:49:99:255,0,255:607.08 |

chr1:46501-47718
(1218 bp)



# At 4x coverage, can still phase more than half of het sites phased at 60x



# Phasing information

- Many additional analysis plots available
- Details described on wiki
  - http://www.broadinstitute.org/gsa/wiki/index.php/Read-backed_phasing_algorithm
- Read-backed phasing will soon be activated in the GATK pipeline…