Workshop 6

Computational Genomics

Phasing

- What is phasing?

Phasing

- First.., let’s talk about haplotype resolution:
  - haplotype is a group of allele inherited from a single parent.
  - knowing a genotype doesn’t always uniquely define a haplotype.
  - Examples?

An Example

- Assume there being **2 SNPs** located on the same chromosome.
- SNP 1: A or T
- SNP 2: C or G
- **3** possible genotypes, **9** possible haplotypes.

possible genotypes	**GG**	**GC**	**CC**
AA	AG, AG	AG, AC	AC, AC
AT	AG, TG	AG, TC or AC, TG	AC, TC
TT	TG, TG	TG, TC	TC, TC

An Example

- Ambiguous phase

possible genotypes	GG	GC	CC
AA	AG, AG	AG, AC	AC, AC
AT	AG, TG	AG, TC or AC, TG	AC, TC
TT	TG, TG	TG, TC	TC, TC

Phasing

- How to resolve ambiguity?
- Discuss in regards to homozygous and heterozygous genotypes. Are homozygous genotypes useful for phasing?

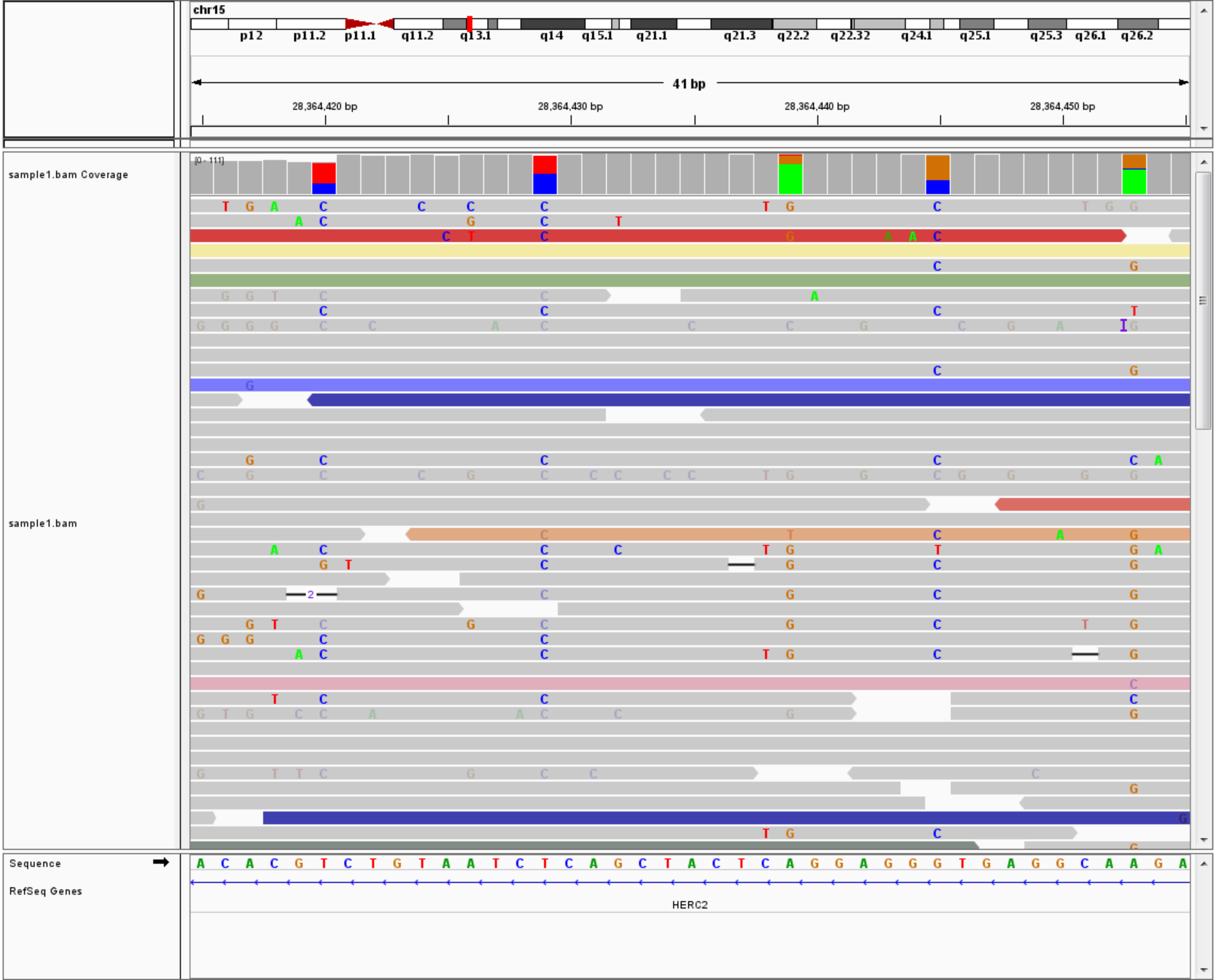
Phasing

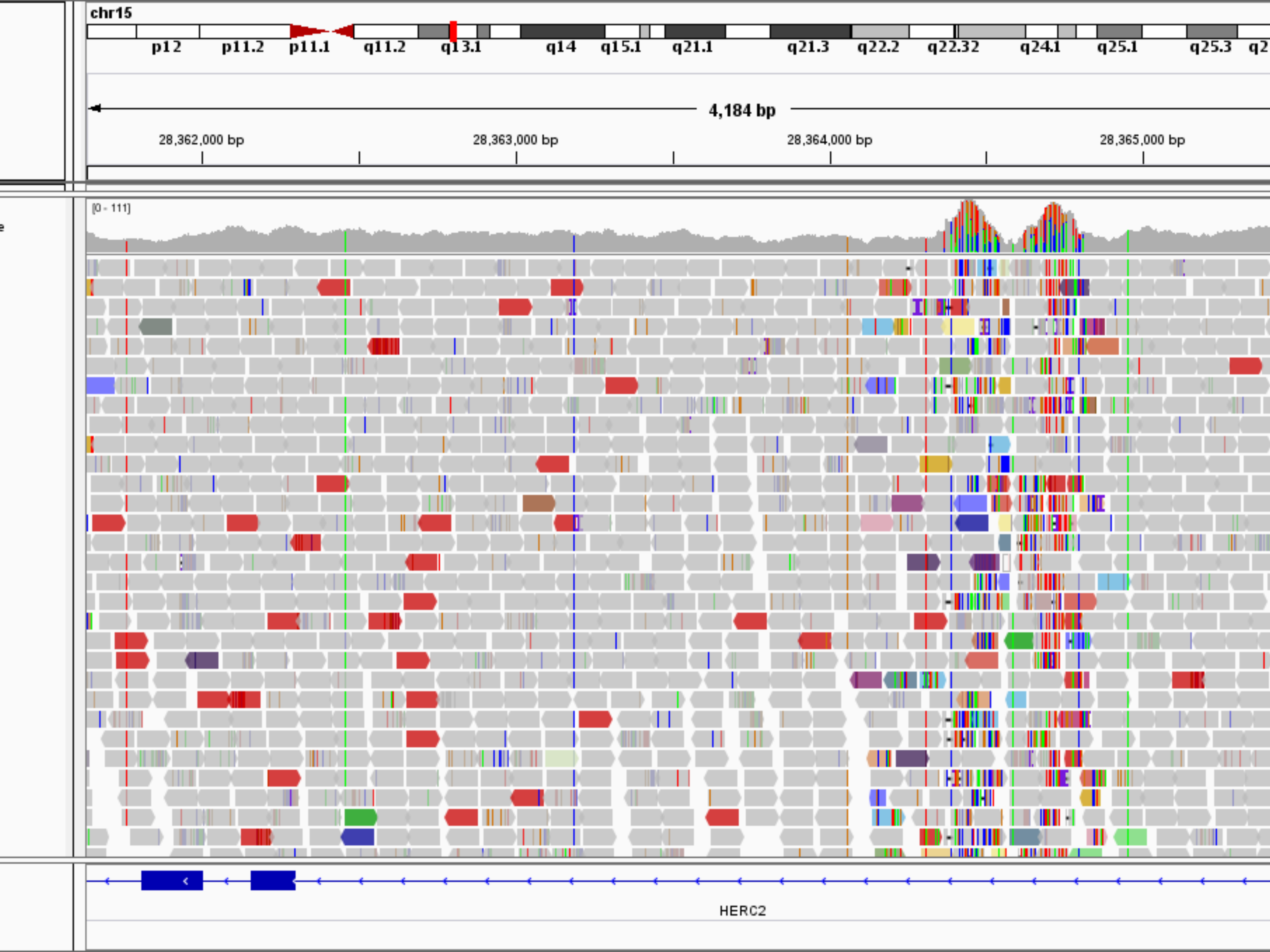
- Only **heterozygous** genotypes are useful for phasing. Why?

Phasing in Sequencing Data

- Look at output from last Workshop. Can any two of the detected SNPs be phased?

chr15	28356858	C	0.421052631579	33.4375
chr15	28356858	T	0.552631578947	29.5238095238
chr15	28359220	A	0.787878787879	25.3076923077
chr15	28359220	C	0.212121212121	4.28571428571
chr15	28360426	G	0.633333333333	27.5263157895
chr15	28360426	T	0.333333333333	4.9
chr15	28360427	C	0.7	25.619047619
chr15	28360427	T	0.266666666667	2.0
chr15	28360438	G	0.258064516129	2.0
chr15	28360438	T	0.677419354839	29.0
chr15	28360638	G	0.367346938776	3.66666666667
chr15	28360638	T	0.612244897959	29.9666666667
chr15	28360660	G	0.230769230769	2.0
chr15	28360660	T	0.769230769231	30.375
chr15	28361477	C	0.761904761905	32.0625
chr15	28361477	G	0.214285714286	31.2222222222
chr15	28361543	C	0.685714285714	33.1666666667
chr15	28361543	G	0.257142857143	30.7777777778
chr15	28361552	A	0.794117647059	34.7777777778
chr15	28361552	G	0.205882352941	27.1428571429
chr15	28362966	A	0.756756756757	28.8571428571
chr15	28362966	C	0.243243243243	2.0
chr15	28363414	G	0.21875	5.42857142857
chr15	28363414	T	0.78125	20.8
chr15	28363852	G	0.217391304348	2.8
chr15	28363852	T	0.739130434783	25.1764705882
chr15	28364366	C	0.222222222222	35.5
chr15	28364366	T	0.777777777778	33.6571428571


chr15
p12
p11.2
p11.1
q11.2
q13.1
q14
q15.1
q21.1
q21.3
q22.2
q22.32
q24.1
q25.1
q25.3
q26.1
q26.2
41 bp
28,364,420 bp
28,364,430 bp
28,364,440 bp
28,364,450 bp
sample1.bam Coverage
[0 - 111]
sample1.bam
Sequence
A C A C G T C T G T A A T C T C A G C T A C T C A G G A G G G T G A G G C A A G A
RefSeq Genes
HERC2


chr15
p12
p11.2
p11.1
q11.2
q13.1
q14
q15.1
q21.1
q21.3
q22.2
q22.32
q24.1
q25.1
q25.3
q2
4,184 bp
28,362,000 bp
28,363,000 bp
28,364,000 bp
28,365,000 bp
[0 - 111]
HERC2

Phasing in Sequencing Data

- Let’s investigate the distance between variants more generally. How many variant pairs could be phased this way? Write your program and find out.

```
import pysam
import sys

bamfile = pysam.AlignmentFile(sys.argv[1],'rb')

r1 = bamfile.fetch("chr15", 28364418, 28364419)

rn1 = set([read.query_name for read in r1])

r2 = bamfile.fetch("chr15", 28364427, 28364428)

rn2 = set([read.query_name for read in r2])

print "reads overlapping SNP1: ",len(rn1)
print "reads overlapping SNP2: ",len(rn2)

print "reads fit for phasing: ",len(rn1.intersection(rn2))
```