Track Data and Bedtools
Adam m France
afrance3@uncc.edu

Part 3:

What format is the output that you get?

```
$ bedtools intersect -a genome_Gibas.vcf -b common_snp.BED | head -5
chr1    1025301 rs9442400       T       C       .       .       .       GT      1/1
chr1    1113121 rs12092254      G       .       .       .       .       GT      0/0
chr1    1366830 rs873927        A       .       .       .       .       GT      0/0
chr1    1715011 rs742359        C       A       .       .       .       GT      0/1
chr1    1875267 rs2262190       a       G       .       .       .       GT      0/1
```

This looks like VCF format

What happens if you switch things up so that the -a file is the BED file and the -b file is the VCF?

```
$ bedtools intersect -a common_snp.BED -b genome_Gibas.vcf | head -5
chr1    1025300 1025301 rs144619388     0       +
chr1    1113120 1113121 rs7518814       0       +
chr1    1366829 1366830 rs75051470      0       +
chr1    1715010 1715011 rs3820004       0       +
chr1    1875266 1875267 rs141122345     0       +
```

This looks like BED format

Which human autosome has the highest per-base-pair representation of SNPs in the 23andMe assay?

I used the command below, but repeated for chr 1 - 23, with the highest WC match being for chromosome 9 with 7,053.

$ bedtools intersect -b genome_Gibas.vcf -a common_snp_wg.BED | grep chr9 | wc 7053

Which SNPs in the 23andMe assay intersect with SNPs in the GWAS (Genome Wide Association Study) Catalog (in the Phenotype and Literature track category)? Do you think the number that you get from a simple intersection is correct? What happens when you include flanking sequence, either via bedtools window or by selecting different options at track download time?

$ bedtools intersect -a genome_Gibas.vcf -b common_snp_wg.BED > overlaps.output

Using wc -l on the output file, we see how many snps overlap:
$ wc -l overlaps.output
14915 overlaps.output

Problem 1: get the coordinate intervals for 100 bases flanking each feature in the GFF.

$ bedtools flank -i NC_007898.gff -g NC_007898.genome -b 100 | head -5

NC_007898.3 RefSeq gene 71536 71635 . - .
ID=gene0;Dbxref=GeneID:3950426;Name=rps12;exception=trans-splicing;gbkey=Gene;gene=rps12;locus_tag=LyesC2p022;part=1/2
NC_007898.3 RefSeq gene 71750 71849 . - .
ID=gene0;Dbxref=GeneID:3950426;Name=rps12;exception=trans-splicing;gbkey=Gene;gene=rps12;locus_tag=LyesC2p022;part=1/2
NC_007898.3 RefSeq gene 99139 99238 . - .
ID=gene0;Dbxref=GeneID:3950426;Name=rps12;exception=trans-splicing;gbkey=Gene;gene=rps12;locus_tag=LyesC2p022;part=2/2
NC_007898.3 RefSeq gene 100033 100132 . - .
ID=gene0;Dbxref=GeneID:3950426;Name=rps12;exception=trans-splicing;gbkey=Gene;gene=rps12;locus_tag=LyesC2p022;part=2/2
NC_007898.3 RefSeq CDS 71536 71635 . - 0
ID=cds0;Parent=gene0;Dbxref=Genbank:YP_008563067.1;Name=YP_008563067.1;exception=

trans-splicing;gbkey=CDS;gene=rps12;product=ribosomal protein S12;protein_id=YP_008563067.1;transl_table=11

Problem 2: for each feature in a file, find its nearest physical neighbors on the same strand (say, within 200 bp) in the chloroplast

$ cat *.gff | grep CDS > output.cds.gff

$ bedtools window -abam BC30.bam -b output.cds.gff -w 200 > prob22.BED -bed
$ head -5 prob22.BED
NC_007898.3 67 342 SC998:02260:01254 24 + NC_007898.3 RefSeq CDS 537 1598 . -0 ID=cds1;Parent=gene2;Dbxref=Genbank:YP_008563068.1;Name=YP_008563068.1;gbkey=CDS;gene=psbA;product=photosystem II protein D1;protein_id=YP_008563068.1;transl_table=11
NC_007898.3 103 385 SC998:01683:02712 23 - NC_007898.3 RefSeq CDS 537 1598 . -0 ID=cds1;Parent=gene2;Dbxref=Genbank:YP_008563068.1;Name=YP_008563068.1;gbkey=CDS;gene=psbA;product=photosystem II protein D1;protein_id=YP_008563068.1;transl_table=11
NC_007898.3 118 343 SC998:01587:02448 40 + NC_007898.3 RefSeq CDS 537 1598 . -0 ID=cds1;Parent=gene2;Dbxref=Genbank:YP_008563068.1;Name=YP_008563068.1;gbkey=CDS;gene=psbA;product=photosystem II protein D1;protein_id=YP_008563068.1;transl_table=11
NC_007898.3 118 355 SC998:02278:01285 24 - NC_007898.3 RefSeq CDS 537 1598 . -0 ID=cds1;Parent=gene2;Dbxref=Genbank:YP_008563068.1;Name=YP_008563068.1;gbkey=CDS;gene=psbA;product=photosystem II protein D1;protein_id=YP_008563068.1;transl_table=11
NC_007898.3 123 363 SC998:01312:00848 24 + NC_007898.3 RefSeq CDS 537 1598 . -0 ID=cds1;Parent=gene2;Dbxref=Genbank:YP_008563068.1;Name=YP_008563068.1;gbkey=CDS;gene=psbA;product=photosystem II protein D1;protein_id=YP_008563068.1;transl_table=11

Problem 3: for each feature in the GFF, report only the features that have NO reads overlapping them in the BAM file

$ bedtools window -abam BC30.bam -b output.cds.gff -v > prob33.BED -bed

```
$ head prob33.BED
NC_007898.3	0	102	SC998:00281:01611	3	+
NC_007898.3	0	149	SC998:00509:00586	0	+
NC_007898.3	0	112	SC998:00707:01173	0	+
NC_007898.3	0	82	SC998:00815:01895	0	+
NC_007898.3	0	131	SC998:01085:02291	0	+
NC_007898.3	0	147	SC998:01175:00421	0	+
NC_007898.3	0	206	SC998:01279:02241	0	+
NC_007898.3	0	192	SC998:01286:02666	8	+
NC_007898.3	0	155	SC998:01311:01763	3	+
NC_007898.3	0	148	SC998:01408:01722	0	+
$ wc -l prob33.BED
39022 prob33.BED
```