# Population genomics I: SNPs from RNA-seq

Eyal Privman
http://goo.gl/6tFdI

In this practical we will use Illumina transcriptome data ("RNA-seq") to discover and genotype SNPs (Single Nucleotide Polymorphisms). We will use the same RNA-seq data that we mapped to the reference genome using TopHat. Remember that these samples were chosen for a differential expression study, so this is not a "proper" population sample. We will use the GATK (installed on the Vital-IT system) to identify sequence positions where several different reads show two alternative bases. (Also can detect insertion/deletion polymorphisms)

Start, as usual, by logging on to the Vital-IT system and sourcing the relevant "bashrc" file:

```
$ ssh -t username@prd.vital-it.ch ssh dee-serv0X
$ cd /scratch/cluster/weekly/username
$ source /scratch/cluster/monthly/eprivman/summer2012/GATK.bashrc
```

The GATK bashrc defines paths to the required Java programs.

## A Pre-process bam files

We will use the SAMtools and Picard packages to pre-process the TopHat output files before running GATK. Use the SAMtools manual page on:
http://samtools.sourceforge.net/samtools.shtml

1. We will first process TopHat alignments of M000B RNA-seq lane stored in directory "TopHat". Use `AddOrReplaceReadGroups` from the Picard package to add required read group (RG) information to the BAM file. The Java command line is:

```
$ cd TopHat
$ myDir=M000B
$ java -Xmx1G -jar $PICARD_HOME/AddOrReplaceReadGroups.jar I="$myDir"/
accepted_hits.bam O="$myDir"/accepted_hits.rg.bam SORT_ORDER=coordinate
RGID="$myDir" RGLB="$myDir" RGPL=illumina RGSM="$myDir" RGPU="$myDir"
```

2. Use `samtools index` to index this BAM file (TopHat outputs a sorted BAM). To see a short help type "`samtools index`".

3. Download the BAM file, index, genome fasta ("GenomeSubset/Assembly/SINV_subset_1.fa") and cuffmerge gtf file ("TopHat_full/merged_asm/merged.gtf") to your computer using scp. Launch IGV and import a new genome (download genome fasta as fasta file and gtf as gene file), then load bam file.

> *Q: Take a look at position 760191 on SIgn00002. Would you say it's a SNP?*

4. Repeat the process for M350B.

> *Q: Compare positions 760187 and 760191. Would you say that these are SNPs? Would you say so given just the M000B data?*

5. Use `samtools merge` to merge the two sorted BAM files. You will need to combine their

headers:

```
$ samtools view -H M000B/accepted_hits.bam > header.sam
```

Open this file in an editor and add these two lines:

```
@RG    ID:M000B    SM:M000B
@RG    ID:M350B    SM:M350B
```

(Note: fields are separated by tab characters. If you can't insert them as tabs you can copy a tab character from a previous line)

```
$ samtools cat -h header.sam -o merge.accepted_hits.rg.bam M000B/
accepted_hits.rg.bam M350B/accepted_hits.rg.bam
```

6. Sort and index this merged BAM file.

7. (Optional) Write a script to process also one queen and one worker sample (you can base it on the runTopHat.sh script). Merge them too.

(For extended description of best practices recommended by the GATK people: http://www.broadinstitute.org/gsa/wiki/index.php/Best_Practice_Variant_Detection_with_the_GATK_v2)

## B SNP detection and genotyping

1. Use GATK to discover and genotype SNPs. Use the help information:

```
java -Xmx1G -jar $GATK_HOME/GenomeAnalysisTK.jar --help
```

You will need to use the genomic sequence "GenomeSubset/Assembly/SINV_subset_1.fa". Run `UnifiedGenotyper` on each sample separately and on the merged data. Examine the output VCF files.
Hint: download these files and analyze them locally.

*Q: In which analyses are the two SNP found? (hint: grep 7601)*

2. Count the number of SNPs in the separate analysis of each sample.

*Q: Are there different numbers of SNPs found in different samples? Why?*

3. Count the number of SNPs in each scaffold.

*Q: Which regions of the genome have more SNPs?*
*Q: You have more power for SNP detection in highly-expressed genes than in other parts of the genome. How can you control for this?*