How to generate a SNP gene map using some simple PERL scripts.

1. Generate .txt files containing information about the chromosomal location of:
    a. SNPs
    b. Exons (and their splice junctions)
    c. ECRs (Evolutionary Conserved Regions)
    d. CpG islands
    e. TFBS (transcription factor binding sites)
    f. Regions of clusters of TFBS
    g. Any other region of interest e.g. siRNA/ miRNA binding sites etc.

This information can be downloaded from the UCSC genome browser (a, b, d, e, g), ECR browser (c), cluster buster browser (f) etc. Give each element a unique name e.g. ECR mouse 1, ECR rat 5, TFBS 1 etc. and keep a copy of the raw data downloaded that corresponds to the element name elsewhere.

Note: a visual overview of regions overlapping a gene of interest can be generated using GeneViewer.pl.

The first column of the .txt files contains the name of the element e.g. rs12345 (for a snp), ECR mouse 1…etc. The second and third columns contain the chromosomal start and stop location of the element. In the case of a snp, there will only be two columns; snp name in the first and location in the second.

Example .txt file for exons/ splice junctions:

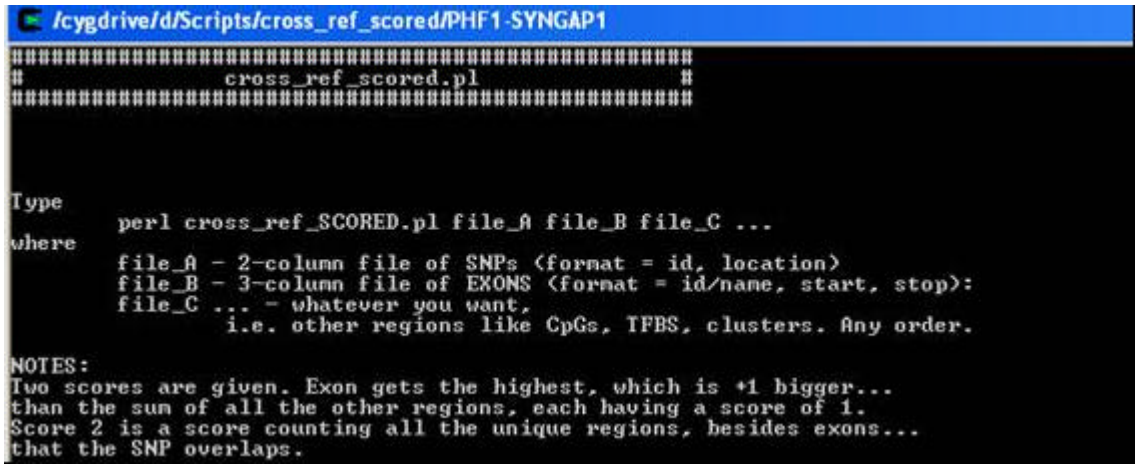
KIFC1 gene, chromosome 5:

```
5' Splice  Junction KIFC1_Exon 1    33467567    33467587
KIFC1_Exon 1                        33467582    33467752
3' Splice  Junction KIFC1_Exon 1    33467747    33467767
5' Splice  Junction KIFC1_Exon 2    33473768    33473788
```

When all files have been generated, save in one folder with the PERL program, cross_ref_SCORED.v3.pl. Open up cygwin, and change directory to the one with all of the saved files and follow these instructions [1,2]:

---

[1] These instructions are given as default output from the program, i.e. are printed to screen when you type “perl cross_ref_SCORES.v3.pl” and press return.
[2] Note: cygwin is not essential. You may have PERL installed on your system and just use the DOS environment.



Result: A .txt file will be generated (that can be viewed more easily in excel) listing all the snps in the snp input file and information about them:

###SNPs.txt,Exons_splice-junctions.txt,Clusters.txt,CpG.txt,TFBS.txt,ECRs.txt,TRF.txt					
SNPID	SNP_START	PRIMARY	SECONDARY	REDUNDANT	OVERLAPPING_REGIONS
rs9394145	33507756	0	1	4	rat ECR 58,mouse ECR 51,opossum scaff 11816 ECR 5,dog ECR 32
rs4231	33491952	6	2	5	rat ECR 19,dog ECR 13,opossum scaff 13488 ECR 18,mouse ECR 16,Cluster Score: 7.35,PHF1_tv2_
rs761583	33489437	0	1	3	dog ECR 10,mouse ECR 9,frog ECR 3
rs34885339	33508463	6	1	7	fugu ECR 3,frog ECR 7,dog ECR 34,opossum scaff 11816 ECR 7,rat ECR 60,mouse ECR 53,zebrafish
rs35180573	33524657	0	1	4	rat ECR 80,opossum scaff 11816 ECR 24,dog ECR 59,mouse ECR 71
rs5875449	33497656	0	1	2	dog ECR 19,rat ECR 36
rs34581314	33488275	6	1	4	rat ECR 9,opossum scaff 13488 ECR 7,dog ECR 9,mouse ECR 7,PHF1_tv1_Exon 3,PHF1_tv2_Exon
rs2247385	33529555	0	1	1	dog ECR 63
rs28361086	33502053	0	1	3	dog ECR 24,rat ECR 44,mouse ECR 38
rs11543058	33492520	6	1	4	rat ECR 21,opossum scaff 13488 ECR 20,mouse ECR 18,dog ECR 14,CUTA_tv1_Exon 6,CUTA_tv5_E
rs211456	33497359	0	1	3	dog ECR 19,rat ECR 35,mouse ECR 30

The first row lists all of the files included for the analysis. A ***primary score*** greater than zero is assigned to snps that are in exonic regions. The ***secondary score*** relates to the number of other regions the snp is in e.g. a TFBS/ECR etc. The ***redundant score*** gives the total number of regions that a snp is in e.g. if a snp is in an ECR of mouse, rat and dog as well as being in a TFBS it will have a secondary score of 2 (ECR and TFBS) but a redundant score of 4 (3 ECRs and 1 TFBS). All regions that the snp overlaps with are listed in the column with the heading “overlapping regions”. If a snp is not in a region of interest it will have “NA” in this column.

The snps can now be sorted based on their scores: in excel, go to Tools➔data➔sort and sort the Primary, Secondary and Redundant scores in descending order.

The positions of exons can also be added to the file and by sorting the file based on chromosomal position; a snp map can be generated.