Use perl to extract sequences given IDs (get-seq4.pl)

Use bioperl to extract sequences given IDs (get-seq-bioperl.pl)

Use bioperl to change sequence format (format-bioperl.pl)

Use bioperl to extract sub-sequences given tabular file with positions (get-subseq-bioperl.pl)

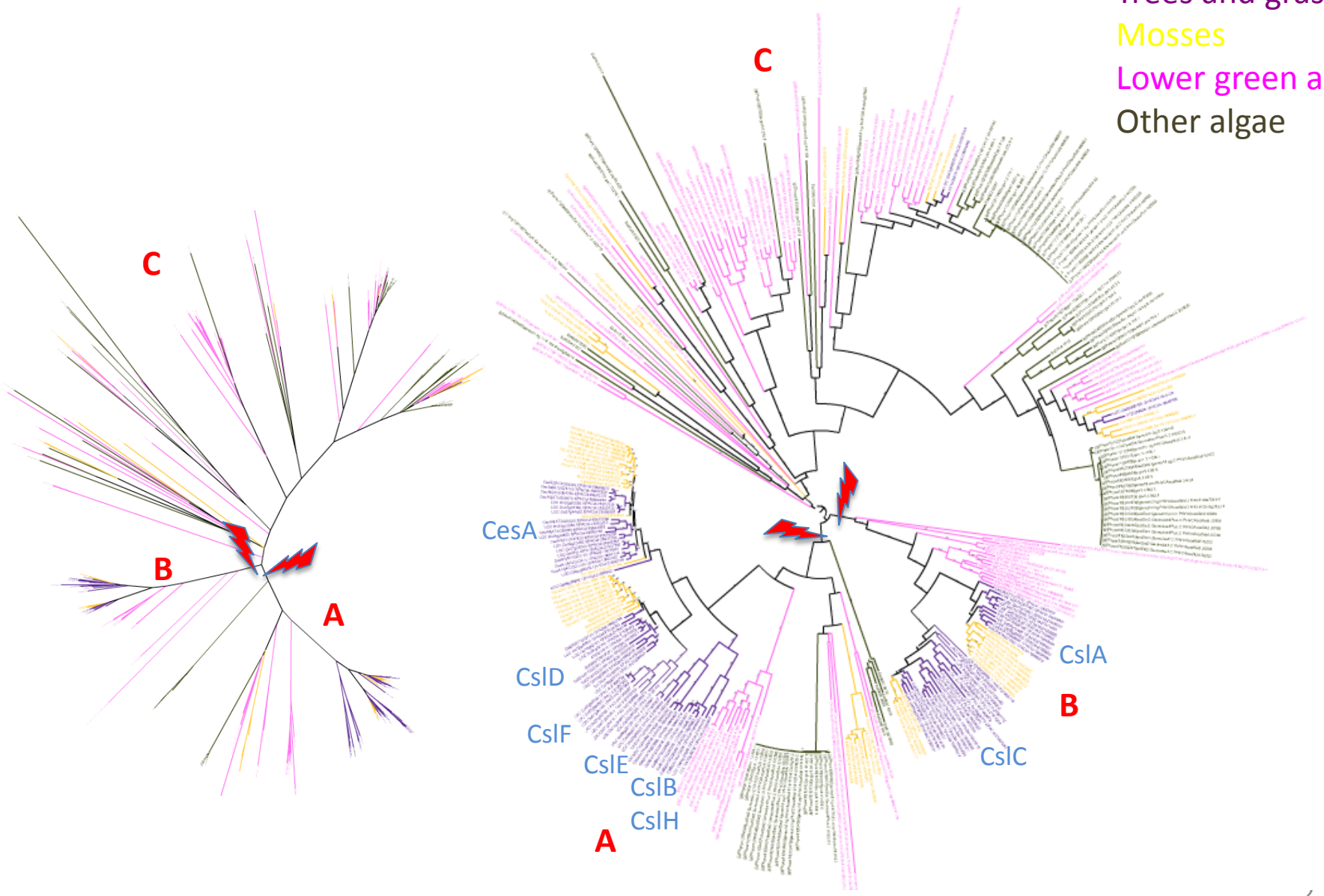
Given GenBank IDs, get fasta or genbank format from NCBI remotely (get-genbank.pl)

Bioperl to parse newick trees (get-treeid.pl)

Connecting two ids file to find common or different things (join.pl)

http://www.biogem.org/downloads/notes/BioPerl.pdf

How to parse newick format tree text file to retrieve IDs?



yyin@glu:~/work/class$ less full-genom.sel.fa.l.fasttree.nwk.2-sub1

Newick format: not human-friendly

```
yyin@glu:~/work/class$ less full-genom.sel.fa.l.fasttree.nwk.2-sub1
(((jgi|Ost9901_3|25890|estExt_fgenesh1_pg.C_Chr_200213:0.15902,jgi|OstRCC809_2|61153|fgenesh1_pm.chr_20_#_228
:0.27533):1.83361[1.00],(((g6025.t1|PACid-26893837:0.0,g6025.t2|PACid-26893838:0.0):0.13115,Vocar20003015m|PA
Cid-23131228:0.11793):0.96427[1.00],(((jgi|ChlNC64A_1|139598|IGS.gm_25_00149:0.40939,(jgi|Coc_C169_1|2833|gw1
.7.154.1:0.14667,(jgi|Astpho1|33644|e_gw1.00149.81.1:0.15812,jgi|Coc_C169_1|66059|estExt_Genemark1.C_70365:0.
15753):0.09256[0.98]):0.13253[0.99]):0.33178[1.00],((jgi|OstRCC809_2|60557|fgenesh1_pm.chr_16_#_137:0.25330,(
jgi|Ost9901_3|18489|fgenesh1_pg.C_Chr_16000054:0.18305,jgi|Ostta4|9161|fgenesh1_pm.C_Chr_17.0001000010:0.1944
7):0.04122[0.41]):0.33634[1.00],(jgi|MicpuC3|23128|MicpuC2.e_gw1.16.71.1:0.19723,jgi|MicpuN3|97997|fgenesh2_p
m.C_Chr_10000079:0.30466):0.04300[0.26]):0.26563[1.00]):0.24381[0.89],((CslC6|AT3G07330.1|PACid-19663909:0.25
950,((CslC12|AT4G07960.1|PACid-19644865:0.18820,((LOC_Os07g03260.1|PACid-21899603:0.09811,LOC_Os03g56060.1|PA
Cid-21911464:0.03241):0.08245[0.99],(LOC_Os01g56130.1|PACid-21908220:0.07294,LOC_Os05g43530.1|PACid-21939856:
0.05842):0.09024[1.00]):0.07664[0.97]):0.10087[1.00],(((140200|PACid-15405032:0.18777,442658|PACid-15412386:0
.13170):0.05478[0.93],(((Pp1s89_286V6.1|PACid-18046098:0.02931,Pp1s224_44V6.1|PACid-18053335:0.04428):0.15389
[1.00],(Pp1s162_130V6.1|PACid-18043543:0.06375,(Pp1s373_28V6.1|PACid-18059424:0.04625,Pp1s19_258V6.1|PACid-18
063508:0.04614):0.04348[0.99]):0.07078[1.00]):0.06561[0.98],(Pp1s164_5V6.1|PACid-18038263:0.03271,Pp1s15_115V
6.1|PACid-18050525:0.02650):0.20137[1.00]):0.07517[0.99]):0.08272[0.99],(CslC4|AT3G28180.1|PACid-19662684:0.2
4547,(LOC_Os08g15420.1|PACid-21888437:0.21486,(LOC_Os09g25900.1|PACid-21926657:0.22883,(CslC8|AT2G24630.1|PAC
id-19638556:0.07641,CslC5|AT4G31590.1|PACid-19646979:0.03274):0.11628[1.00]):0.04293[0.44]):0.03353[0.45]):0.
06688[0.99]):0.05248[0.61]):0.11373[0.65]):0.61724[1.00],(LOC_Os09g39920.1|PACid-21925118:0.71591,((LOC_Os08g
33740.1|PACid-21888939:0.25395,LOC_Os02g51060.1|PACid-21924463:0.47471):0.06092[0.14],((LOC_Os06g12460.1|PACi
d-21930248:0.37581,((LOC_Os07g43710.1|PACid-21901112:0.17583,LOC_Os03g26044.1|PACid-21914244:0.15585):0.30519
[1.00],(LOC_Os10g26630.1|PACid-21882652:0.16652,(LOC_Os09g26770.2|PACid-21927293:0.56424,LOC_Os03g07350.1|PAC
id-21911250:0.00015):0.08626[1.00]):0.16088[1.00]):0.09931[0.96]):0.13144[1.00],(((CslA9|AT5G03760.1|PACid-19
669977:0.16676,(CslA2|AT5G22740.1|PACid-19673036:0.21878,(LOC_Os02g09930.1|PACid-21919891:0.15886,LOC_Os06g42
020.1|PACid-21932621:0.15291):0.07181[0.85]):0.06807[0.75]):0.04907[0.78],(230176|PACid-15411642:0.23068,(Pp1
s36_214V6.1|PACid-18053528:0.04313,(Pp1s36_62V6.1|PACid-18053619:0.03188,Pp1s65_194V6.1|PACid-18057777:0.0375
7):0.01312[0.49]):0.29524[1.00]):0.07358[0.94]):0.03123[0.02],(CslA7|AT2G35650.1|PACid-19639665:0.34578,((CslA1|AT4G16590.1|PACid-19644242:0.23037,(CslA10|AT1G24070.1|PACid-19656371:0.14229,(CslA15|AT4G13410.1|PACid-19
645825:0.13816,CslA11|AT5G16190.1|PACid-19666574:0.10137):0.02807[0.41]):0.05314[0.89]):0.21219[1.00],(CslA3|
AT1G23480.1|PACid-19653790:0.13931,CslA14|AT3G56000.1|PACid-19665013:0.35565):0.09725[0.98]):0.09176[0.98]):0
.08783[0.98]):0.11718[1.00]):0.05389[0.15]):0.08084[0.10]):0.41999[1.00]):0.24745[0.98]):0.31142[0.51]):0.786
97[0.99]):0.20304[0.50],ConsensusfromContig37575-snap_masked-ConsensusfromContig37575-abinit-gene-0.2-mRNA-1-
cds-7119/2753-2927-0-+:3.26499);
```

perldoc Bio::TreeIO

vi get-treeid.pl

```
#!/usr/bin/perl -w

use Bio::TreeIO;

$treeio=Bio::TreeIO->new(-format=>"newick", -
file=>$ARGV[0]);

while($tree=$treeio->next_tree){
  for $node ($tree->get_nodes){
    print $node->id."\n";
  }
}
```

```
perl get-treeid.pl /home/yyin/work/class/full-genom.sel.fa.l.fasttree.nwk.2-sub1
| less

perl get-treeid.pl /home/yyin/work/class/full-genom.sel.fa.l.fasttree.nwk.2-sub1
| sed '/^$/d' | less

perl get-treeid.pl /home/yyin/work/class/full-genom.sel.fa.l.fasttree.nwk.2-sub1
| sed '/^$/d' > sub1.id
```

How do we join two tabular files according to a common column?

Consider this task: find out how many cancer-related genes are transcription regulators?

The cancer Gene Census

http://www.nature.com/nrc/journal/v4/n3/abs/nrc1299.html

http://cancer.sanger.ac.uk/cancergenome/projects/census/
Download the excel sheet and open it
Copy and paste to create a file
/home/yyin/work/class/cancer-census

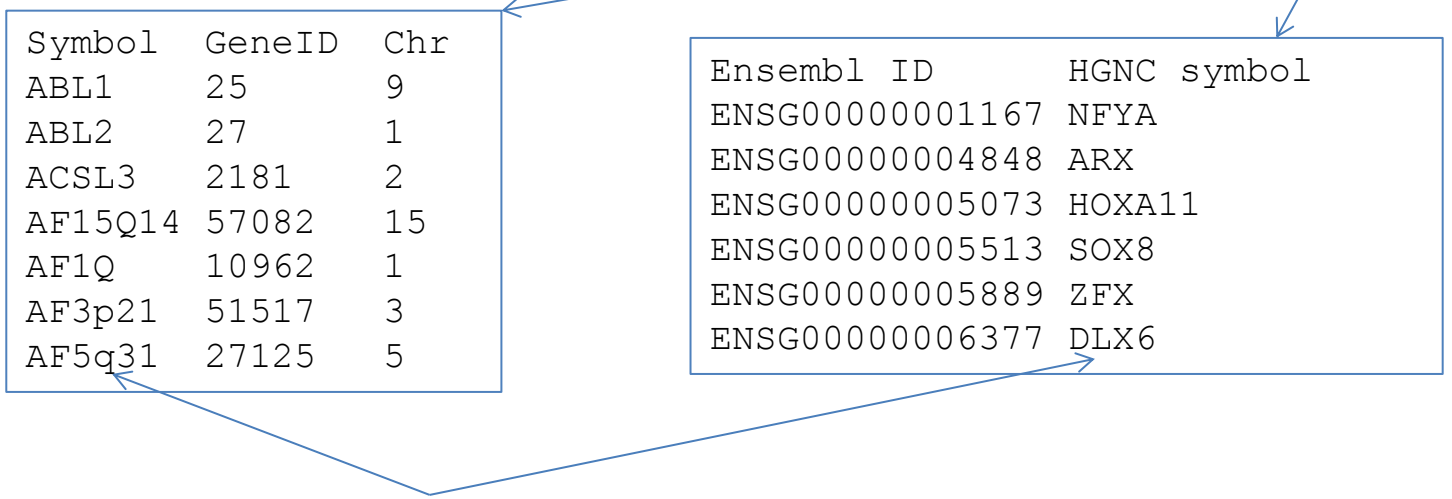
A census of Human Transcription Factors

http://www.nature.com/nrg/journal/v10/n4/suppinfo/nrg2538.html
Wget the Supplementary information S3
/home/yyin/work/class/nrg2538-s3.txt

What we need to do is to join the two tables together based on common col

First let's clean the two tables first:

```
cat nrg2538-s3.txt | cut -f2,6 > nrg2538-s3.txt.cut
less cancer-census | cut -f1,3,4 > cancer-census.cut
```



Symbol	GeneID	Chr
ABL1	25	9
ABL2	27	1
ACSL3	2181	2
AF15Q14	57082	15
AF1Q	10962	1
AF3p21	51517	3
AF5q31	27125	5

Ensembl ID	HGNC symbol
ENSG00000001167	NFYA
ENSG00000004848	ARX
ENSG00000005073	HOXA11
ENSG00000005513	SOX8
ENSG00000005889	ZFX
ENSG00000006377	DLX6

Common col is

Now we need a perl script to join the two tables:
using the gene symbol as the key to create a hash

vi join.pl

```
#!/usr/bin/perl

open(IN,$ARGV[0]);
open(IN2,$ARGV[1]);

while(<IN>){
  chomp $_;
  @col=split(/\t/,$_);
  $hash{$col[1]}=$_;
}

while(<IN2>){
  chomp $_;
  @col2=split(/\t/,$_);
  if(defined $hash{$col2[0]}){
    print $_,"\t",$hash{$col2[0]},"\n";
  }
  else{
    print $_,"\n";
  }
}
```

```
perl join.pl nrg2538-s3.txt.cut cancer-census.cut | less
perl join.pl nrg2538-s3.txt.cut cancer-census.cut | awk '$4!=""' | less
perl join.pl nrg2538-s3.txt.cut cancer-census.cut | awk '$4!=""' | wc -l
```

April 30: Student presentation of project 1:

Group 1 (GH9): Bill, Brenda, Steve, Tom
Group 2 (GH10): Jenny, Matt, Shannon, William

Each group have 30 minutes

Indicate who contributed which analyses of the project

Include the background, the experiment design, the methods and the results