In[1]:=
```
SetDirectory[NotebookDirectory[]]
```

Out[1]=
```
/Users/eterna/runs/sra
```

We investigate whether ALK with an alternative transcript initation site in intron 19 ($ALK^{ATI}$) is potentially expressed on SRA. This notebook requires that the script alk.sh in the “sra” subdirectory of the “runs” repo has been executed. This, in turn, requires the intropolis database of junctions across SRA (intropolis.v1.hg19.tsv.gz). See alk.sh for further information.

$ALK^{ATI}$ should see exons 1-19 largely unexpressed and exons 20-29 expressed. We compare the coverage A of junctions across exons 1-19 with the coverage B of junctions across exons 20-29 by ranking samples in order of decreasing D=(B-A) / (A+B).

alk.sh writes two files: one with junctions from the region spanning exons 1-19 of ALK across SRA, and the other with junctions from the region spannign exons 20-29. These are loaded below.

In[2]:=
```
ALKStartJunctions = Import["!gzip -cd alk_start_junctions.tsv.gz", "TSV"];
ALKEndJunctions = Import["!gzip -cd alk_end_junctions.tsv.gz", "TSV"];
```

startCoverageTotals sums coverage of junctions in exons 1-19, and endCoverageTotals does the same for exons 20-29.

In[3]:=
```
startCoverageTotals =
  Total[SparseArray[Append[(ToExpression /@ StringSplit[ToString[#[[7]]], ","]) + 1,
       21 507] → Append[ToExpression /@ StringSplit[ToString[#[[8]]], ","],
       0]] & /@ ALKStartJunctions];
```

In[4]:=
```
endCoverageTotals =
  Total[SparseArray[Append[(ToExpression /@ StringSplit[ToString[#[[7]]], ","]) + 1,
       21 507] → Append[ToExpression /@ StringSplit[ToString[#[[8]]], ","],
       0]] & /@ ALKEndJunctions];
```

In[5]:=
```
startVsEnd = Transpose[{Range[1, 21 507], startCoverageTotals, endCoverageTotals}];
```

Map sample indexes to SRA accession numbers.

In[6]:=
```
indexes = Import["intropolis.idmap.v1.hg19.tsv"];
```

In[7]:=
```
indexToSrr = Association[
   #[[1]] → #[[2]] & /@ Transpose[{indexes[[All, 1]], indexes[[All, 5]]}]];
```

In[8]:=
```
indexToSrp = Association[
   #[[1]] → #[[2]] & /@ Transpose[{indexes[[All, 1]], indexes[[All, 2]]}]];
```

In[9]:=
```
srrToIndex = Association[
   #[[1]] → #[[2]] & /@ Transpose[{indexes[[All, 5]], indexes[[All, 1]]}]];
```

Find top ten sample hits in order of decreasing D:

In[10]:=
```
ranks = {indexToSrp[#[[1]]], indexToSrr[#[[1]]], #[[2]], #[[3]], #[[4]]} & /@ Reverse[
    SortBy[{#[[1]] - 1, #[[2]], #[[3]], N[(#[[3]] - #[[2]]) / (#[[3]] + #[[2]])]} & /@
      Select[startVsEnd, #[[2]] + #[[3]] ≥ 50 &], Last]][[Range[1, 10]]]
```

Out[10]=
```
{{SRP007461, SRR545713, 0, 139, 1.}, {SRP010166, SRR396804, 0, 172, 1.},
 {SRP017262, SRR620100, 0, 108, 1.}, {SRP042031, SRR1289650, 1, 85, 0.976744},
 {SRP042031, SRR1289651, 1, 77, 0.974359}, {SRP007461, SRR545716, 2, 94, 0.958333},
 {SRP017413, SRR628586, 12, 111, 0.804878}, {DRP001919, DRR016705, 38, 285, 0.764706},
 {SRP007461, SRR545714, 14, 63, 0.636364}, {ERP006077, ERR532612, 16, 53, 0.536232}}
```

Cross-referencing with SRA at http://www.ncbi.nlm.nih.gov/sra, we find that these samples are:

1) NHEM.f_M2: normal human melanocyte cell line sequenced by CSHL for ENCODE
1) non-small cell lung adenocarcinoma
1) leukemia
4) macrophage, part of SRP042031
5) macrophage + fibroblast, part of SRP042031
6) NHEM_M2: normal human melanocyte cell line sequenced by CSHL for ENCODE
7) uveal melanoma
8) H2228, an EML4-ALK-expressing lung adenocarcinoma cell line
9) NHEM.f_M2: normal human melanocyte cell line sequenced by CSHL for ENCODE
10) primary prostate tumor

Find the ALK coverages of other SRA samples associated with the macrophage study SRP042031:

In[11]:=
```
startCoverageTotals[[srrToIndex["SRR1289652"]]]
```

Out[11]= 0

In[12]:=
```
endCoverageTotals[[srrToIndex["SRR1289652"]]]
```

Out[12]= 0

In[13]:=
```
startCoverageTotals[[srrToIndex["SRR1289653"]]]
```

Out[13]= 0

In[14]:=
```
endCoverageTotals[[srrToIndex["SRR1289653"]]]
```

Out[14]= 0

Looks like ALK isn't expressed in these samples!

In[15]:=
```
indexToSrr[19513]
```

Out[15]= SRR1274169