Bioinformatic approaches to regulatory genomics and epigenomics

376-1347-00L | week 04

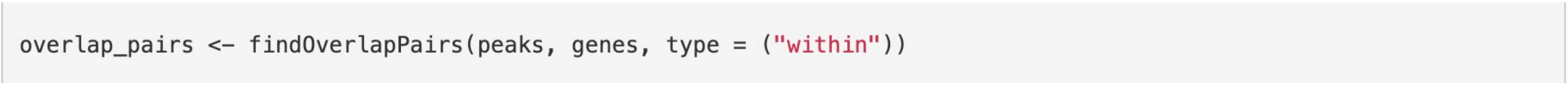
Pierre-Luc Germain

Plan for today

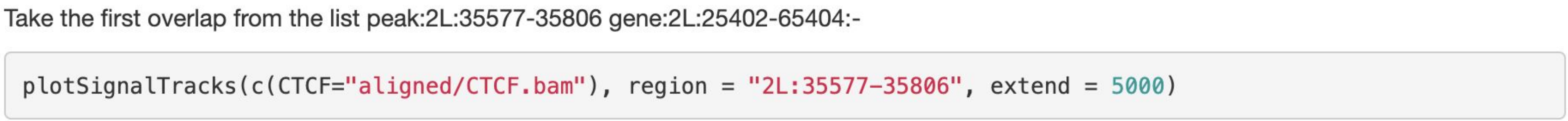
- Debriefing on the assignment
- Coverage track generation
- Manipulating and visualizing peaks
- ENCODE & functional elements
- Finding data from the literature

Debriefing on the assignments: Format

- Please show outputs of code in the rmarkdown document:
    - use `head()` if the object has many entries (e.g. `GRanges`)
    - and do not turn eval=FALSE, unless something runs for very long, please write it then



```
overlap_pairs <- findOverlapPairs(peaks, genes, type = ("within"))
```

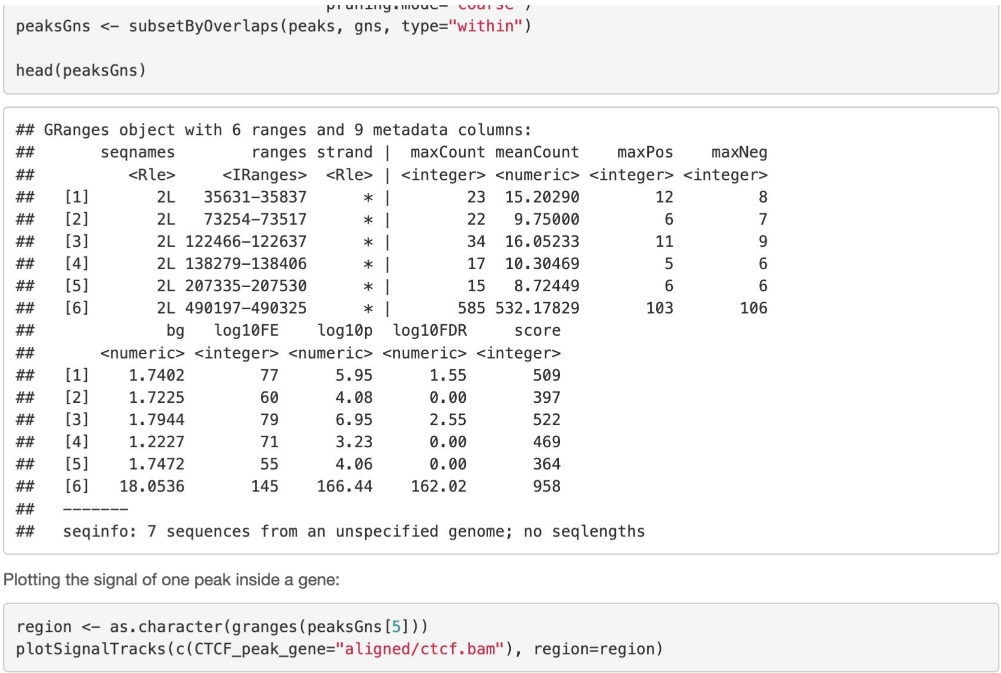


Take the first overlap from the list peak:2L:35577-35806 gene:2L:25402-65404:-

```
plotSignalTracks(c(CTCF="aligned/CTCF.bam"), region = "2L:35577-35806", extend = 5000)
```

Debriefing on the assignments: Format

- Please show outputs of code in the rmarkdown document:
  - use `head()` if the object has many entries (e.g. `GRanges`)



```
                     pruning.mode="coarse")
peaksGns <- subsetByOverlaps(peaks, gns, type="within")

head(peaksGns)
```

```
## GRanges object with 6 ranges and 9 metadata columns:
##       seqnames        ranges strand |  maxCount meanCount    maxPos    maxNeg
##          <Rle>     <IRanges>  <Rle> | <integer> <numeric> <integer> <integer>
##   [1]       2L   35631-35837      * |        23  15.20290        12         8
##   [2]       2L   73254-73517      * |        22   9.75000         6         7
##   [3]       2L 122466-122637      * |        34  16.05233        11         9
##   [4]       2L 138279-138406      * |        17  10.30469         5         6
##   [5]       2L 207335-207530      * |        15   8.72449         6         6
##   [6]       2L 490197-490325      * |       585 532.17829       103       106
##               bg   log10FE    log10p  log10FDR     score
##        <numeric> <integer> <numeric> <numeric> <integer>
##   [1]     1.7402        77      5.95      1.55       509
##   [2]     1.7225        60      4.08      0.00       397
##   [3]     1.7944        79      6.95      2.55       522
##   [4]     1.2227        71      3.23      0.00       469
##   [5]     1.7472        55      4.06      0.00       364
##   [6]    18.0536       145    166.44    162.02       958
##   -------
##   seqinfo: 7 sequences from an unspecified genome; no seqlengths
```

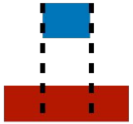
Plotting the signal of one peak inside a gene:

```
region <- as.character(granges(peaksGns[5]))
plotSignalTracks(c(CTCF_peak_gene="aligned/ctcf.bam"), region=region)
```

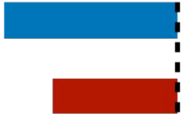
Debriefing on the assignments: findOverlaps

- `findOverlaps, subsetByOverlaps` have a `type` argument

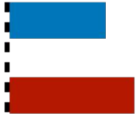
```
findOverlaps(peaks, genes, type=”<…>”)
```



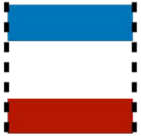
type=within



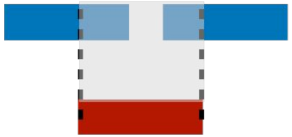
type=end



type=start



type=equal

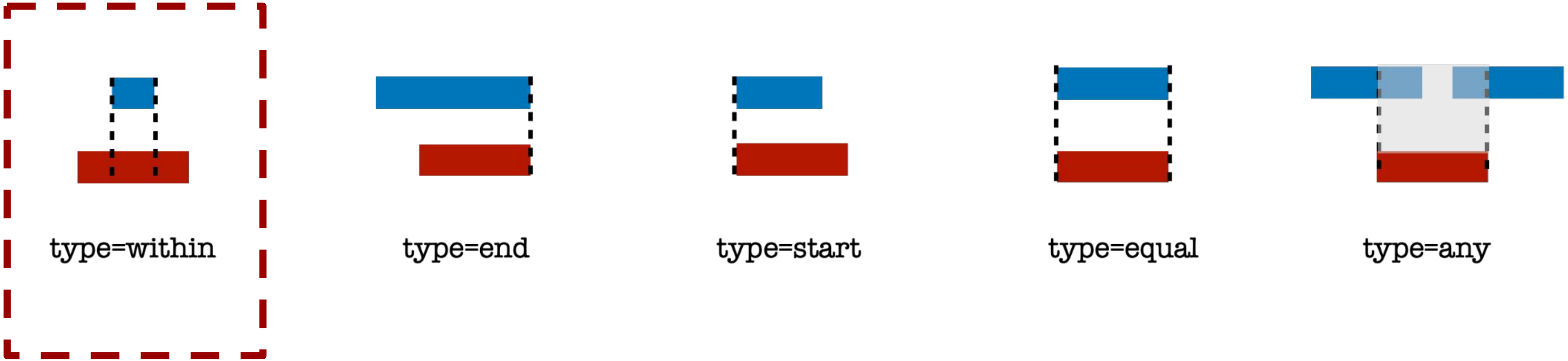


type=any

Debriefing on the assignments: findOverlaps

- Plot the signal around one of the peaks that is located **inside** a gene.

```
findOverlaps(peaks, genes, type=”<…>”)
```



Debriefing on the assignments

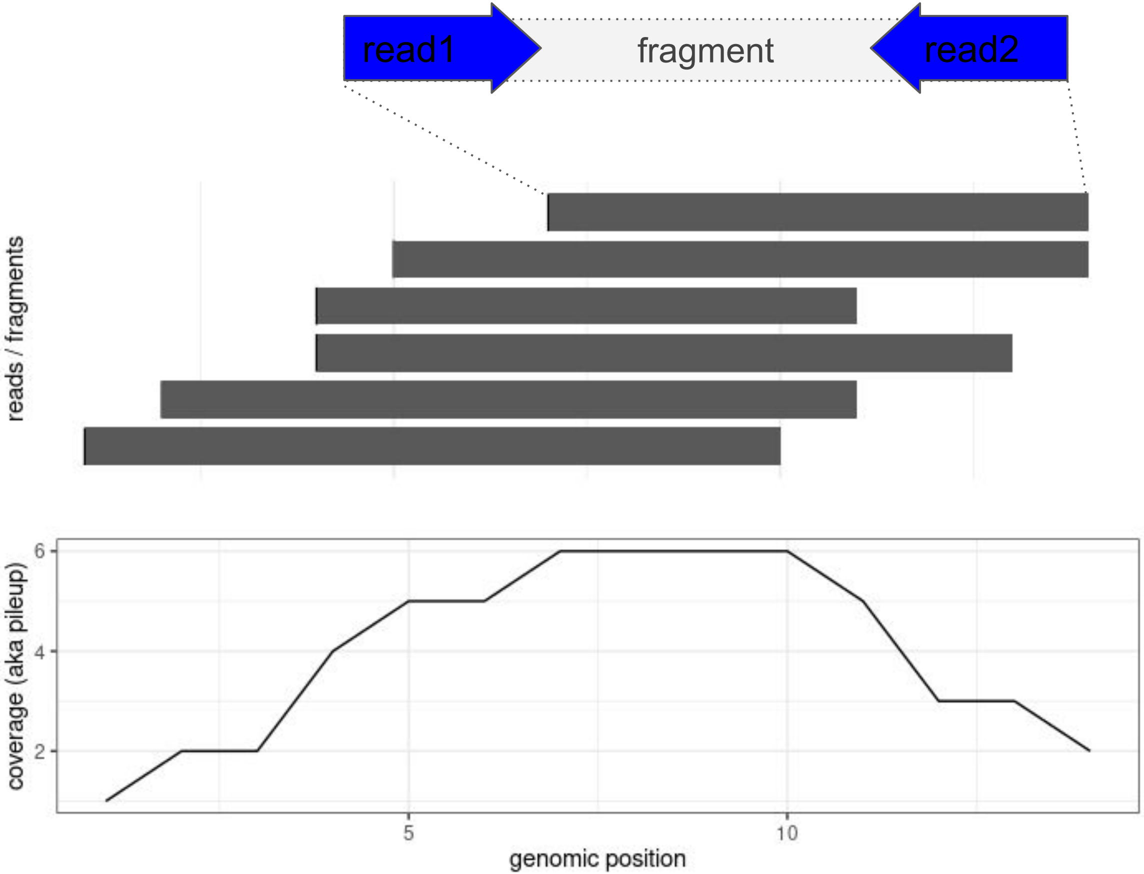
```
Warning: Each of the 2 combined objects has sequence levels not in the other:
  - in 'x': Unmapped_Scaffold_4_D1555_D1692, Unmapped_Scaffold_60_D1601,
                    …
```

This means that the two objects don’t have exactly the same chromosomes (i.e. “seqLevels”). This can be because:

- You are using objects (e.g. an EnsDb and a genome) that don’t match, or
- Your genome contains unlocalised / unplaced scaffolds which are absent from the other object (e.g. gene annotation)

See: http://www.ensembl.org/info/genome/genebuild/chromosomes_scaffolds_contigs.html

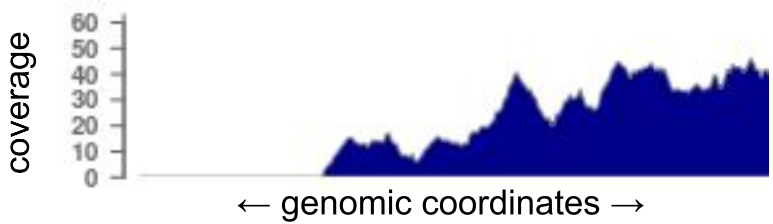
Recap of fragment summarization



Visualizations available in *epiwraps*

Documentation

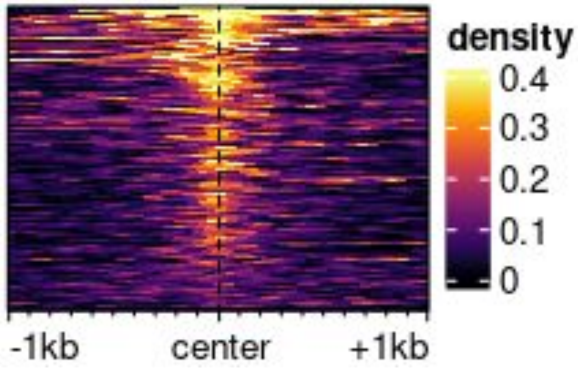
- Signal across one genomic region: `plotSignalTracks`



- `Input: bam/bigwig/bed/GRanges`

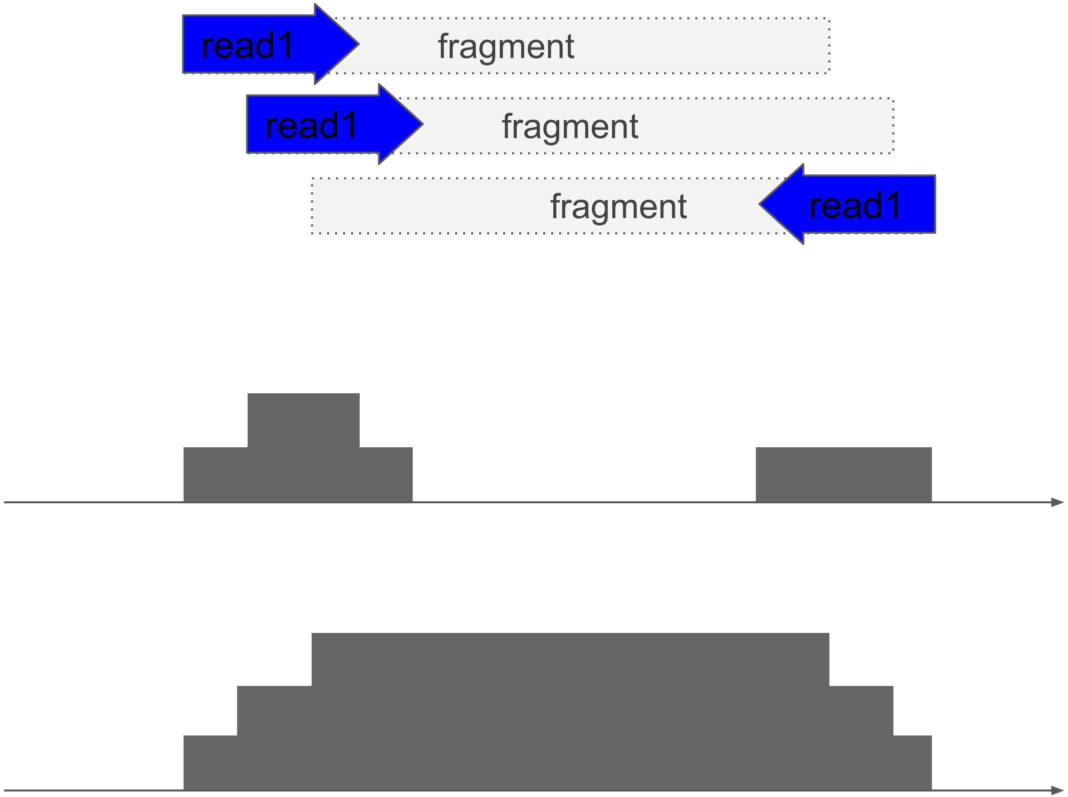
(Based on the *Gviz* R package)

- Signal across several genomic regions: `signal2Matrix` → `plotEnrichedHeatmaps`



(Mainly based on the EnrichedHeatmap R package, itself based on ComplexHeatmap)

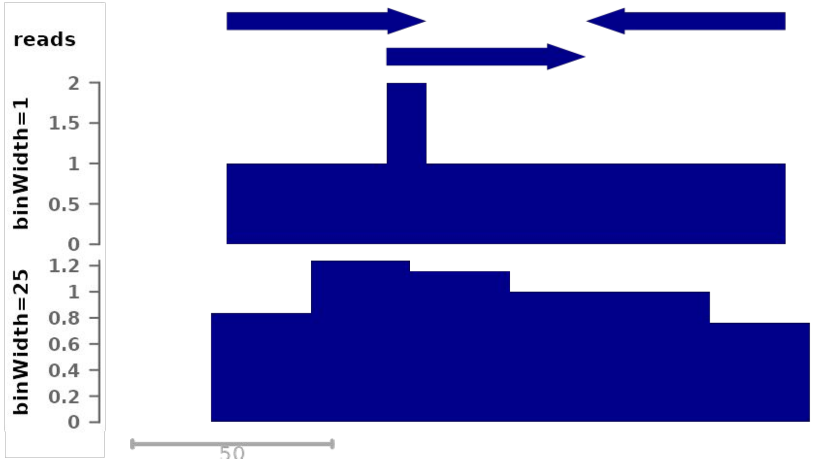
Extension of single-end reads in coverage track generation



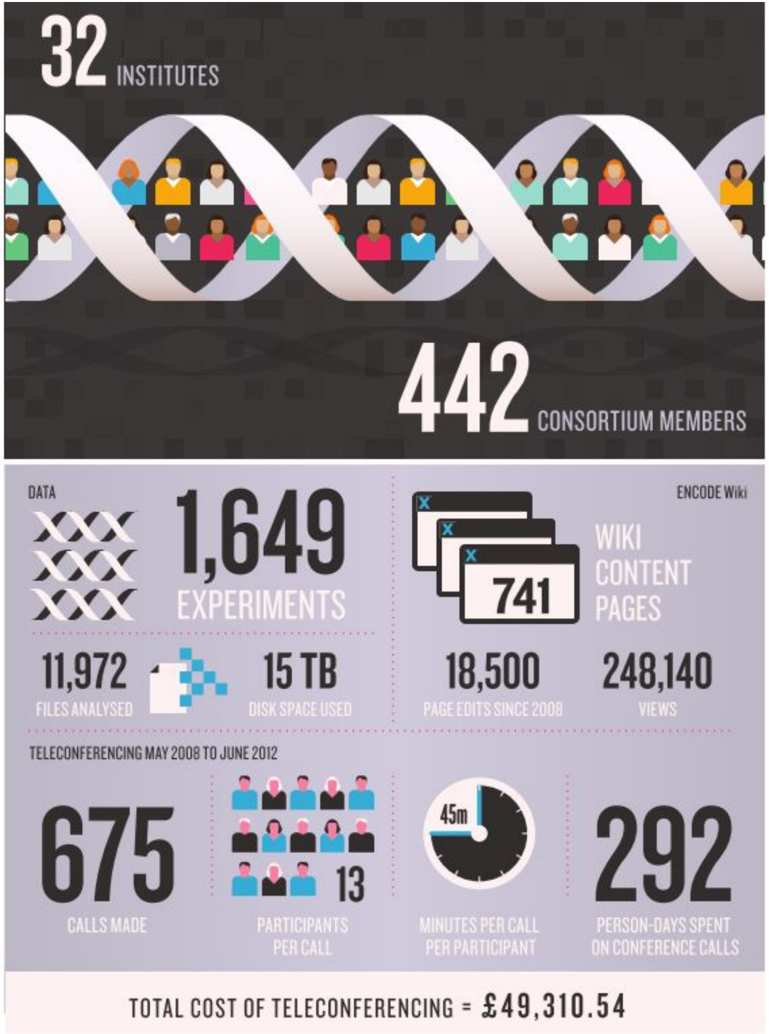
Coverage without read extension

Coverage with read extension

Resolution of coverage tracks



Smaller bin-widths offer higher resolution, but create larger files that are a bit heavier to work with.
(common bin widths are 1, 10, 25 or 50nt)

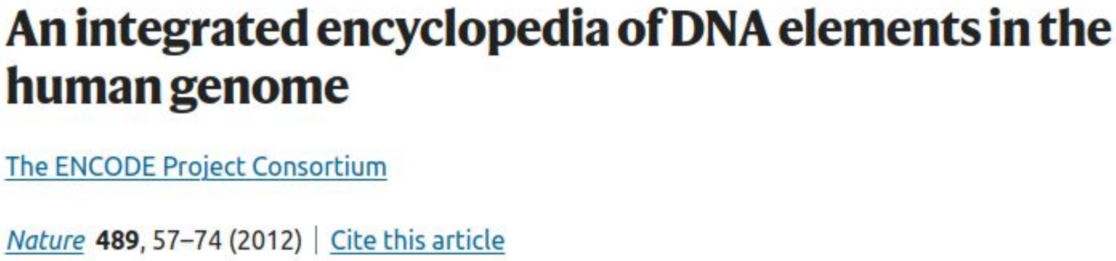

32 INSTITUTES
442 CONSORTIUM MEMBERS
DATA
ENCODE Wiki
1,649 EXPERIMENTS
741
WIKI CONTENT PAGES
11,972 FILES ANALYSED
15 TB DISK SPACE USED
18,500 PAGE EDITS SINCE 2008
248,140 VIEWS
TELECONFERENCING MAY 2008 TO JUNE 2012
675 CALLS MADE
13 PARTICIPANTS PER CALL
45m MINUTES PER CALL PER PARTICIPANT
292 PERSON-DAYS SPENT ON CONFERENCE CALLS
TOTAL COST OF TELECONFERENCING = £49,310.54

The ENCyclopedia Of DNA Elements

~30 publications in September 2012

$288 million USD

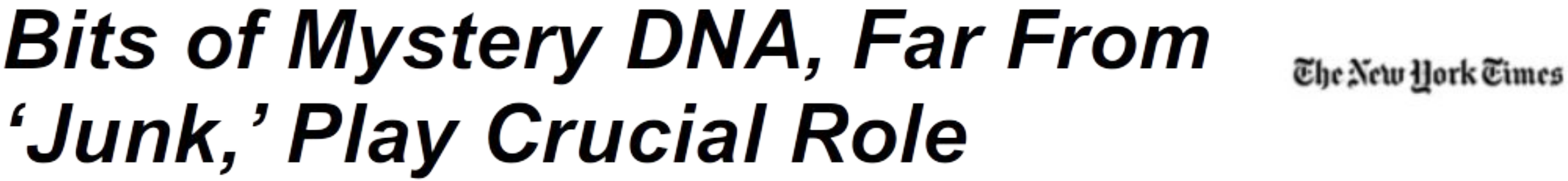
… then an ENCODE2, 3, now working towards the 5…



An integrated encyclopedia of DNA elements in the human genome

The ENCODE Project Consortium

Nature **489**, 57–74 (2012) | Cite this article



Bits of Mystery DNA, Far From 'Junk,' Play Crucial Role

The New York Times

by Gina Kolata

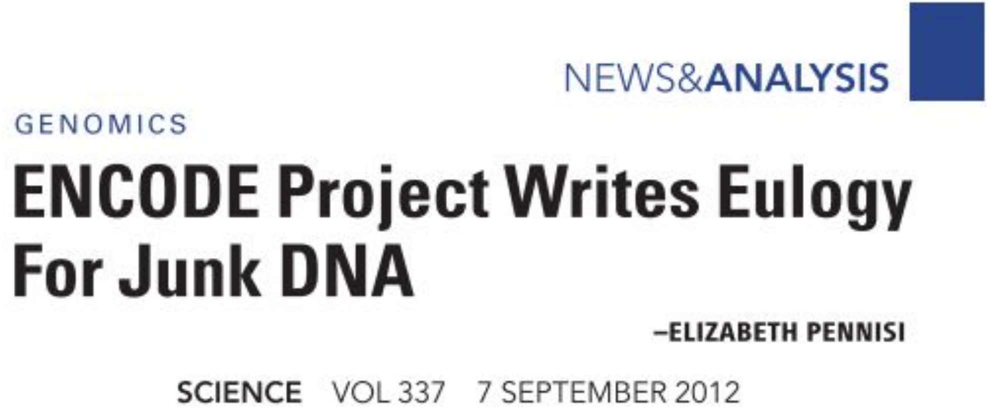
“At least 80 percent of this DNA is *active* and *needed*.”

The evolutionary arguments for junk:

- 1% protein-coding
- ~4 to 10% evolutionarily conserved
- >50% transposable elements
- Onions have a 5 times bigger genome

The very angry response:

- Graur et al., GBE 2013



NEWS&**ANALYSIS**

GENOMICS

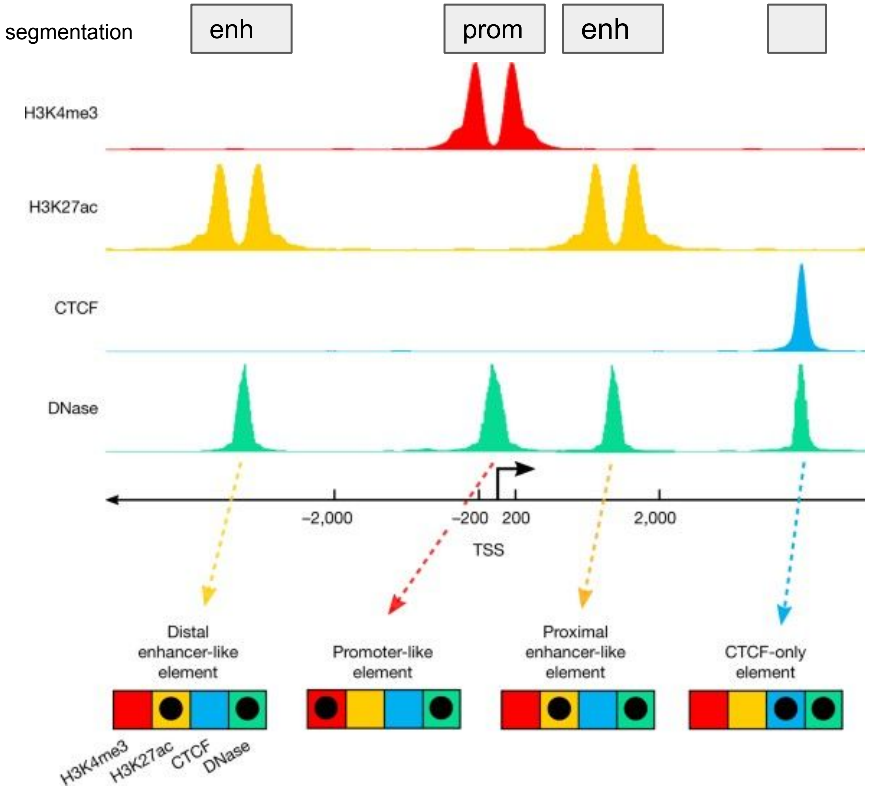
ENCODE Project Writes Eulogy For Junk DNA

–ELIZABETH PENNISI

SCIENCE VOL 337 7 SEPTEMBER 2012

(For more on the topic, see Germain, Ratti and Boem 2014; Ratti and Germain 2022)

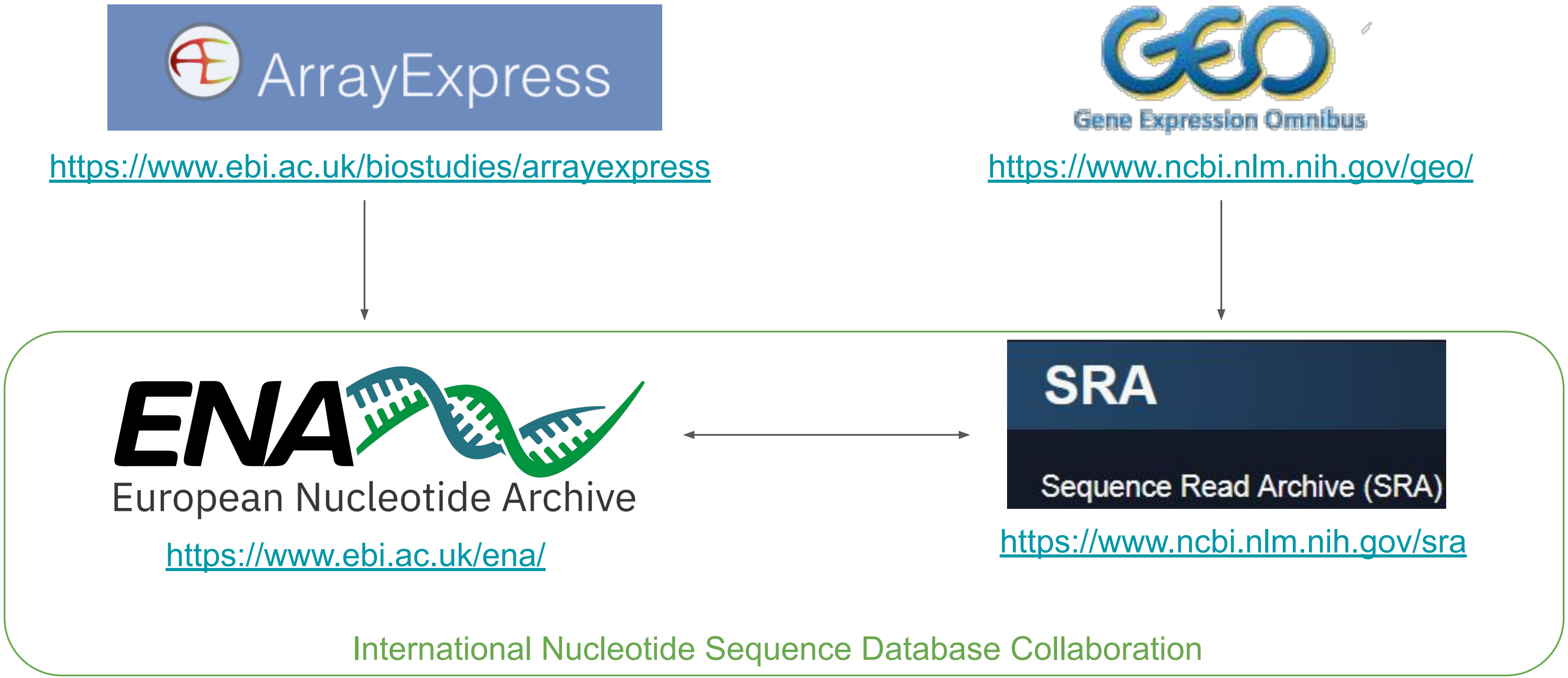
A signature-based encyclopedia of DNA elements



ENCODE's "signature strategy":

- Different types of functional genetic elements are associated with different chemical signatures
- We can identify functional elements by identifying these signatures genome-wide

Generic repositories for NGS data



https://www.ebi.ac.uk/biostudies/arrayexpress

https://www.ncbi.nlm.nih.gov/geo/

https://www.ebi.ac.uk/ena/

https://www.ncbi.nlm.nih.gov/sra

International Nucleotide Sequence Database Collaboration

Quality-controlled and uniformly processed human and mouse NGS datasets

ROADMAP epigenomics PROJECT ↔ ENCODE

www.roadmapepigenomics.org

www.encodeproject.org

(hematopoietic system)

BLUEPRINT epigenome

Assignment

- Find and download from ENCODE the **peaks** (i.e. bed-like format) for the following histone modifications in mouse embryonic stem cells (mESC) from ENCODE:
  - p300, H3K4me3, H3K4me1, H3K27ac, and H3K27me3
  - (when there are replicates, we recommend using the bed file denoted as “`conservative IDR thresholded peaks`”)
- Of the p300 peaks, what proportion overlap each of the marks?
- Don’t forget to upload your assignment as “`assignment.html`” !