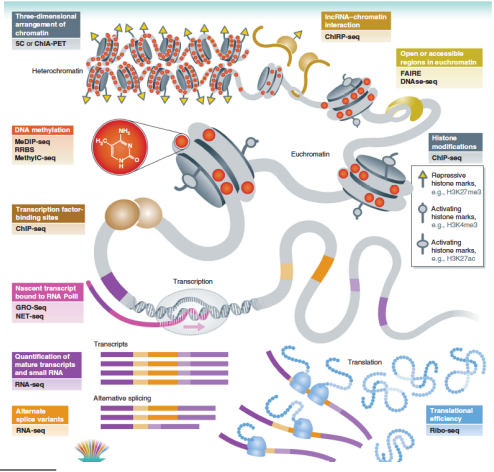
Merging regions

Dan Jiang and Leo Collado

December 17th, 2013

1. Problem setting
   - Background
   - Question
2. Proposed method
3. Example

High-Throughput Genomics Panorama[1]



[1] Wendy Weijia Soon, Manoj Hariharan, and Michael P. Snyder. "High-throughput sequencing for biology and medicine". In: *Molecular Systems Biology* 9.1 (). URL: http://www.nature.com/msb/journal/v9/n1/fig_tab/msb201261_F2.html (visited on 03/05/2013).

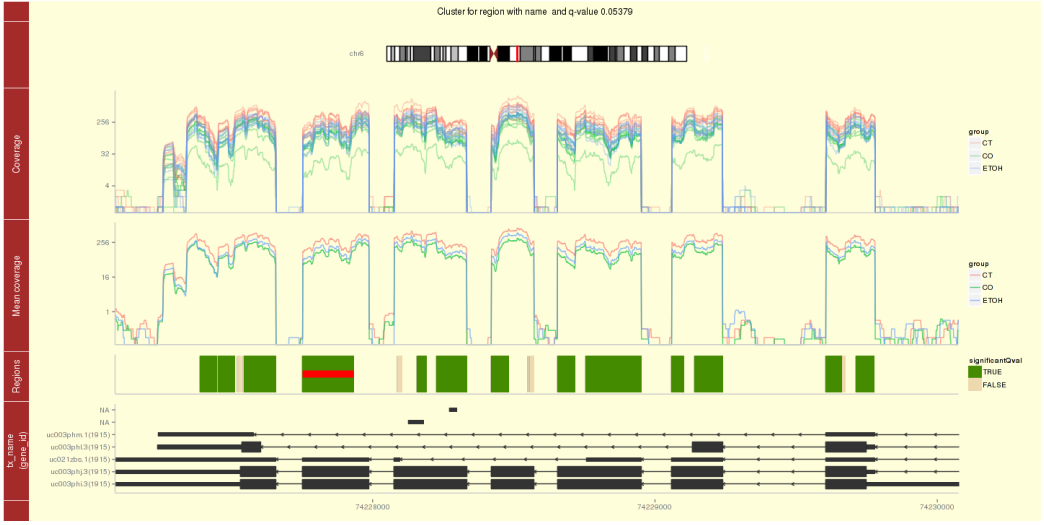
What is common?

- Measurements along the genome (sometimes summarized)
- Two or more groups of samples
- Typical question: are there differences between the groups?
  - $\rightarrow$ Find the candidate regions.

Issue: regions might be highly fragmented.
Why?

1. Biological reasons: regions correspond to two exons (intron is the cause).
2. Measurement not ideal: coverage dips.

Example region cluster (by distance)



Question

Are two adjacent regions *similar*?

1. Can we link them?
2. a

Translating framework

- What is measured?
  - Coverage =: Y
  - Transformed: $\log_2(Y + 32)$
- Individual (cluster of measurements) $\rightarrow$ sample
- Repeated visits $\rightarrow$ individual base pairs (from a given chromosome)
  - Note that the data is correlated!

Consider a region pair:

1. region1: first region
2. regionM: middle part
3. region2: second region

Proposed method

Model for sample i:

$$\log_2(Y_{ijk} + 32) = \alpha + \beta_1 \text{sampleDepth}_i + \boldsymbol{\beta}_2 \text{group}_j + \boldsymbol{\beta}_3 \text{region}_k + \epsilon$$

Using *region1* as the reference, we want to test $\boldsymbol{\beta}_3\,(\text{region}_2) = 0$.

derHippo chr 6

- 25 samples
- 3 groups: CO, ETOH and CT
- 890 pairs; most short:
  - 32.24719% with all regions (1, M, 2) having a width greater than 1
  - 17.97753% greater than 2
- Chose the largest cluster, then the pair starting with the largest region from the cluster.

Data

```
pairs[i, ]

##       start1     end1   startM     endM   start2     end2 cluster
## 553 74227753 74227934 74227935 74228089 74228090 74228091     168
##     width1 widthM width2 widthNoM
## 553    182    155      2      184

dim(covdata[[i]])

## [1] 8475    6

head(covdata[[i]], n = 1)

##   base  region sample coverage sampleDepth group
## 1    1 region1    CO1    6.066       28.25    CO
```

GEE - exchangeable

```
gfit.ex$call

## gee(formula = coverage ~ sampleDepth + group + region, id = sample,
##     data = covdata[[i]], family = gaussian, corstr = "exchangeable",
##     silent = TRUE)

c(gfit.ex$coefficients, dim(gfit.ex$working.correlation)[1])

##   (Intercept)   sampleDepth       groupCO     groupETOH regionregionM
##       -9.2920        0.5920       -0.7108       -0.3104       -0.9189
## regionregion2
##        0.5878      339.0000
```

References

- **Project** code and results: `https://github.com/lcolladotor/756final_code`
- A. Frazee, S. Sabunciyan, K. D. Hansen, R. A. Irizarry, and J. T. Leek (2013). Differential expression analysis of rna-seq data at single base resolution, Biostatistics, *recently accepted*.
- L. Collado-Torres, A. Frazee, M. Love, R. A. Irizarry, A. E. Jaffe, J. T. Leek (2013). derfinder: Software for annotation-agnostic RNA-seq differential expression analysis. Manuscript in preparation.
- `derfinder` package `https://github.com/lcolladotor/derfinder`

Thank you!