

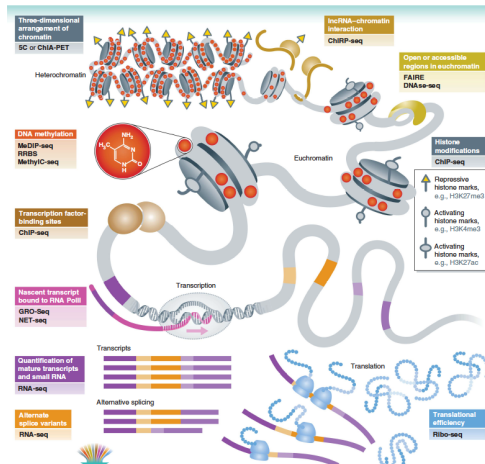
Merging regions

Dan Jiang and Leo Collado

December 17th, 2013

- 1 Problem setting
 - Background
 - Question
- 2 Proposed method
- 3 Example
- 4 Results

High-Throughput Genomics Panorama¹



¹Wendy 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?

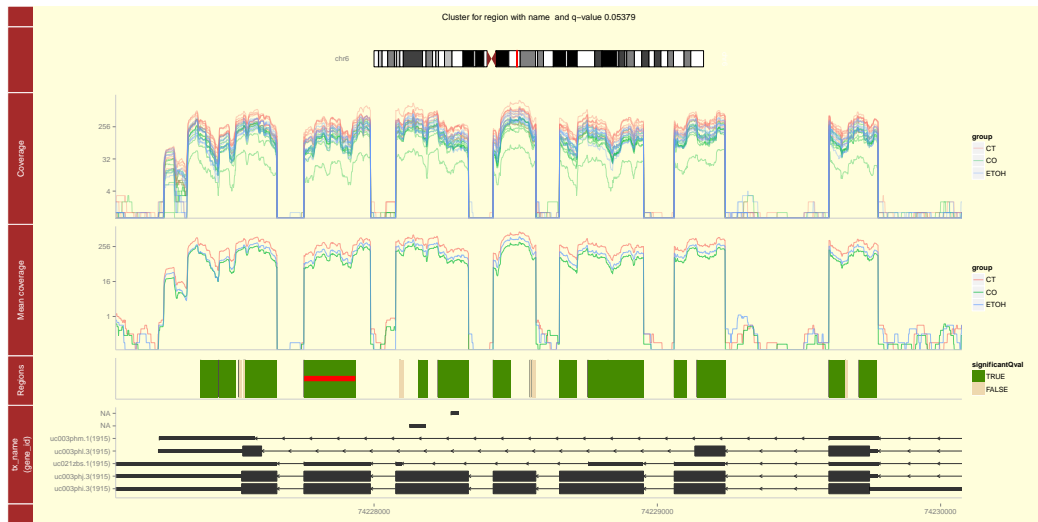
- Measurements along the genome (sometimes summarized)
- Two or more groups of samples
- Typical question: are there differences between the groups?
 - ▶ → Find the candidate regions.

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 Are regions overlapping the same exon more frequently *linked*?

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 :

$$\log_2(Y_{ijk} + 32) = \alpha + \beta_1 \text{sampleDepth}_i + \beta_2 \text{group}_j + \beta_3 \text{region}_k + \epsilon$$

Using *region1* as the reference, we want to test $\beta_3(\text{region}_2) = 0$.

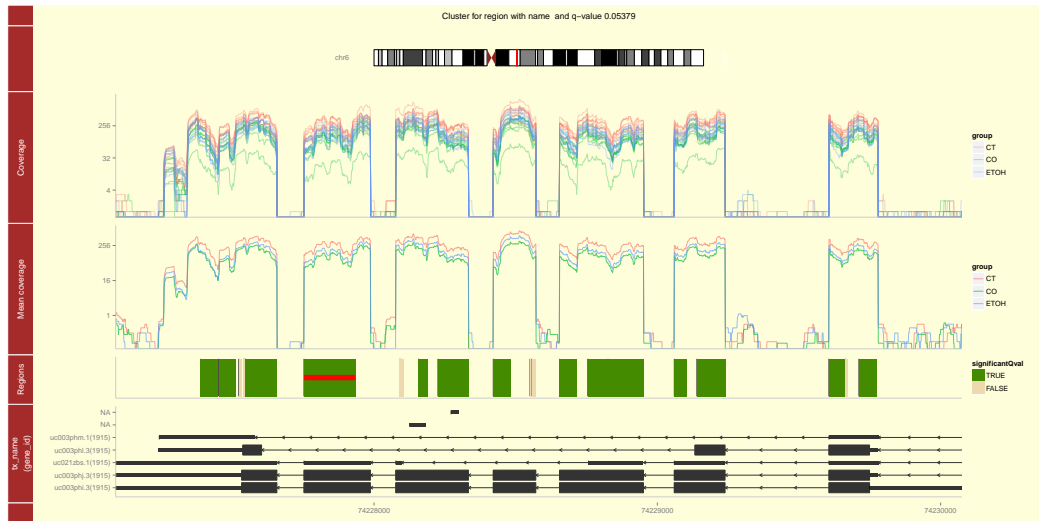
Data sets used

- **derHippo**: RNA-seq brain hippocampus study
- 25 samples
- 3 groups: CO, ETOH and CT
 - chr6 890 pairs
300 ($\sim 33.7\%$) with regions 1 & 2 having a width greater than 1, region $M < 250$
 - chr22 573 pairs
187 ($\sim 32.6\%$) passing the filtering

Example:

chr 6, chose the largest cluster, then the pair starting with the largest region from the cluster.

Example region cluster



Data

```
pairs[i2, ]
```

```
##          start1      end1  startM      endM  start2      end2 cluster
## 552 74227546 74227657 74227658 74227752 74227753 74227934      168
##      width1 widthM width2 widthNoM
## 552      112      95      182      294
```

```
dim(covdata[[i2]])
```

```
## [1] 9725      6
```

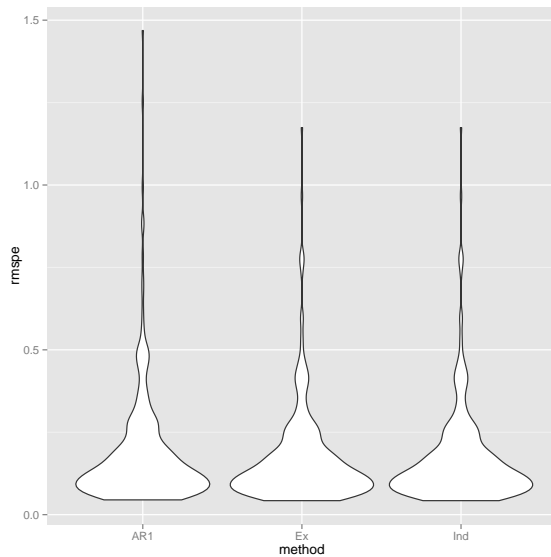
```
head(covdata[[i2]], n = 1)
```

```
##   base  region sample coverage sampleDepth group
## 1    1 region1   C01    7.401      28.25    C0
```

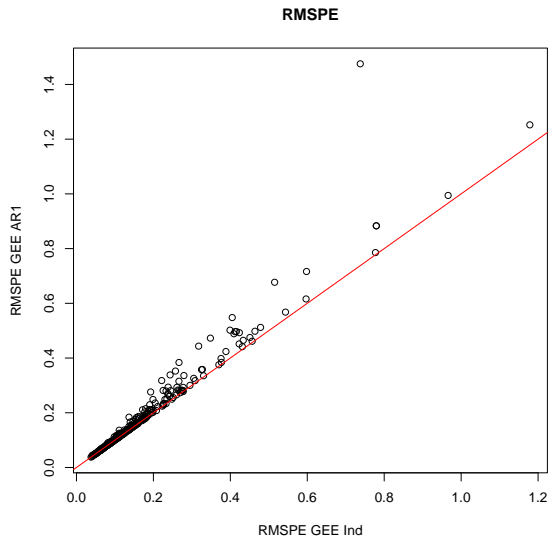
GEE - AR1

```
##
## Call:
## geeglm(formula = coverage ~ sampleDepth + group + region, family = gaussian,
##       data = covdata[[i2]], id = sample, corstr = "ar1")
##
## Coefficients:
##              Estimate      Std.err    Wald Pr(>|W|)
## (Intercept)  -11.02816    3.69521    8.91  0.0028 **
## sampleDepth    0.66137    0.12836   26.55  2.6e-07 ***
## groupCO      -0.76057    0.10828   49.34  2.2e-12 ***
## groupETOH    -0.33737    0.10847    9.67  0.0019 **
## regionregionM -1.79501    0.10960  268.25 < 2e-16 ***
## regionregion2 -0.11747    0.00614  366.09 < 2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Estimated Scale Parameters:
##              Estimate Std.err
## (Intercept)    0.458  0.0224
##
## Correlation: Structure = ar1 Link = identity
##
## Estimated Correlation Parameters:
##              Estimate Std.err
## alpha      0.971 0.00691
## Number of clusters: 25 Maximum cluster size: 389
```

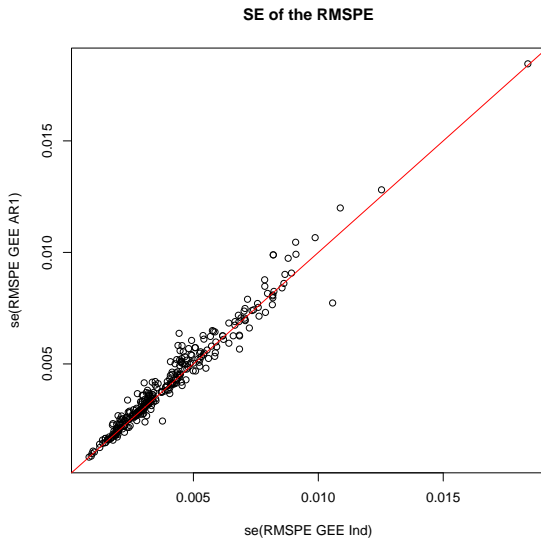
RMSPE chr6



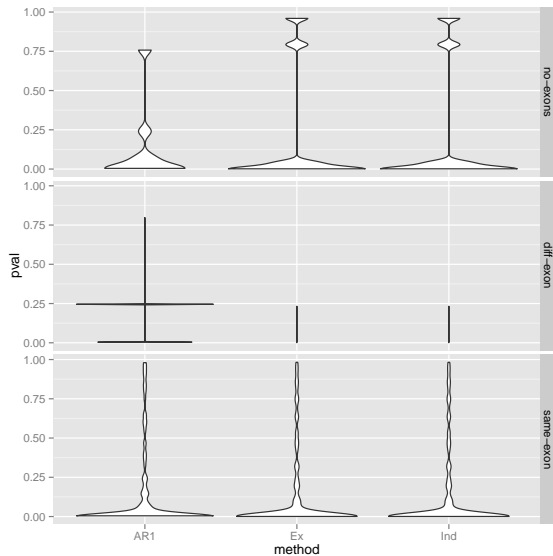
RMSPE chr6: GEE AR1 vs GEE Ind



SE RMSPE chr6: GEE AR1 vs GEE Ind



P-values by exon status: chr6



Test result by exon status

- Adjust for multiple testing by using $q\text{-value} < 0.10$

BothAccept	Ar1Accept	IndAccept	BothReject	chr	ExonStatus
0	1	2	7	chr6	no-exons
0	1	0	13	chr6	diff-exon
19	24	23	210	chr6	same-exon
0	4	1	11	chr22	no-exons
0	0	0	2	chr22	diff-exon
5	15	14	135	chr22	same-exon

Conclusions

- With longer region pairs, fitting GEE takes a significant amount of time.
- GEE with Independence working correlation had lower RMSPE.
- For pairs sharing an exon, 11-20% were linked.

References

- **Project** code and results: https://github.com/lcolladotor/756final_code
- A. Frazee, S. Sabuncuyan, 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!