Statistical analysis of genome-scale data

Geir Kjetil Sandve,
the HyperBrowser team,
University of Oslo
The way we do this

• “Try out yourself” means try out yourself
  • No use in me demonstrating one step at a time
  • I provide the key terms - you sort out the clicking
  • If you’re stuck - ask us or neighbor
  • I will anyway demo each task afterwards

• And science is more than clicking, anyway
  • We will focus just as much on the underlying concepts
Outline of session

• Analyzing genomic tracks
• The gospel: powerful analyses through simple means
• The cautionary tale: challenges with data and assumptions
• Conclusion
Outline of session

• Analyzing genomic tracks

• The gospel: powerful analyses through simple means

• The cautionary tale: challenges with data and assumptions

• Conclusion
What are genes?

This! : 

[Diagram showing the location of genes within a genome]
What are genes?

Reference genome acts like coordinate system for genomic data

| chr21 | 10079666 | 10120808 | NM_001187 |
| chr21 | 13332357 | 13412442 | NR_026916 |
| chr21 | 13700575 | 13700652 | NR_036164 |
| chr21 | 13904368 | 13935777 | NM_174981 |
| chr21 | 14137324 | 14142556 | NR_026755 |
Classic examples of genomic track data

- Gene locations, gene expression
- Evolutionary conserved regions
- Repeating elements
ENCODE, FANTOM, GEO, Roadmap Epigenomics ..

• By now, Big Science provides:
  • Chromatin accessibility (DHSs) for ~350 cell samples
  • Binding of ~100 TFs in several cell types
  • Most histone modifications in several cell types
  • Gene expression for thousands of setups
  • TSS and active promoters in ~950 cell samples
  • DNA methylation, 3D genome structure, ...
Delineating basic types of genomic tracks

- Bins
- Points
- Segments
- Function
And what about analysis?
Example analyses

- A relation between methylation patterns and repeating elements? (Genome Res. 2009 19: 221-233)

- Distinct methylation for tissue-specific genes? (Genome Res. 2010 20: 1493-1502)

- Cooperative histone modifications? (Nat Genet 2008 40:897-903)
Example analyses (cont.)

- Fragile sites, breakpoints and repeats? (Genome Biology 2006 7:R115)

- Copy number variation, repeats, duplications and genes? (Genome Res. 2009 19: 1682-1690)

- Methylation and active genes at T-Cell G0->G1 (Genome Res. 2009 19: 1325-1337)
Example analyses (cont.)

- Virus integration vs genes, CpG, GC-content  
  (Journal of Virology 2007 6731–6741)

- Methylation patterns in embryonic cells  
  (PNAS 2010 107:10783–10790)
Example analyses (cont.)

- 80.4% of the genome participates in at least one biochemical/chromatin-associated event (Nature 2012, 489:57)

- Motifless TF ChIP-seq peaks vs high TRF occupancy (HOT) (Genome Biol 2012, 13:R48)

- [Almost every ENCODE article has many examples, really]
This can’t be it?!
Co-location of genomic features

- Common question:
  *do genomic feature X and Y occur (more than expected) at the same locations in the genome?*

- Used to discover novel relations

- May indicate a direct causal relation, or hint to indirect association.
How does this look at the drawing board?

overlap > expected?
How does this look at the drawing board?

Issues in practice:

- How to represent data (easy)
- How to count overlap (easy)
- How to conclude on relation or not (challenging)
Outline of session

- Analyzing genomic tracks
- The gospel: powerful analyses through simple means
- The cautionary tale: challenges with data and assumptions
- Conclusion
Technical note (1): Wifi access

- Network name (ssid): conferences
- Password (WPA2): uio202aar
Technical note (2): Server load

- We will today be 40 people running on a 32-CPU server
  - Might be some queueing of runs or slow GUI
  - Might be some DB operational errors (just refresh..)
  - [we are in transition to a larger server]
B-cells important for multiple sclerosis?

- “Results suggest an important role of B cells in the pathology of MS”
- “MS associated genomic regions co-localized with regions which are functionally active in B cells”
- “MS SNPs including 0.25cM flanks overlap more than expected with regions of chromatin state AP in gm12878”
Getting to know MS

- Go to HyperBrowser:
  - "hyperbrowser.uio.no"

- Import MS from published Page:
  - "Training material"

- Expand history element, and:
  - "Perform HyperBrowser analysis"

- Keep defaults, and "Start analysis"
Do MS overlap unexpectedly with AP regions in gm12878?

Select tool: “Analyze genomic tracks”

Genome: “hg18” (!)

Track1: “--From history--” -> MS

Track2: “Chromatin / Chromatin state../..Gm12878../1 Active Promoter”

Analysis: “Overlap?”

Keep defaults and “Start analysis”
But, something isn’t right!

- Unexpected overlap between MS and B-cell AP does not confirm a role of B-cells in MS!
  - (why not?)

- Restrict to AP regions specific to B-cell
  - “Subtract” intervals of other cell from B-cell
  - But still not right! (and why not?)

- Must instead use a case-control analysis..
Track customized for analysis:
Create B-cell AP vs hepatocyte AP

Menu “HyperBrowser track repository”:

- “Extract track from HyperBrowser repository”
- “hg18”

1. “Chromatin / Chromatin state../..Gm12878../.. AP
2. “Chromatin / Chromatin state../..Hepg2../.. AP

Menu “Customize tracks”:

- “Combine two BED files into single case-control track”
Do MS overlap preferentially with B-cell AP vs hepatocyte AP?

Tool: “Analyze genomic tracks”

Genome: hg18, Track1: MS
  · shortcut: “perform HyperBrowser analysis”

Track2: Customized case-control track

Question: “Preferential overlap?”

Keep defaults and “Start analysis”
Outline of session

• Analyzing genomic tracks
• The gospel: powerful analyses through simple means
• The cautionary tale: challenges with data and assumptions
• Conclusion
A second case:
Do HPVs integrate preferentially inside genes?
Do HPV and genes overlap?

- Import HPV from (the same) published Page:
  - "Training material"
  - "Analyze genomic tracks", "hg19" (!)
    - (shortcut: "perform HyperBrowser analysis")
  - Track1: "--From history..." -> HPV
  - Track2: Find genes in track collection ..
  - "Located inside?", "Start analysis"
How does this look at the drawing board?

overlap > expected?
How does this look at the drawing board?

Points

Segments

inside $>$ outside?
<table>
<thead>
<tr>
<th>P</th>
<th>P</th>
<th>Different frequencies?</th>
</tr>
</thead>
<tbody>
<tr>
<td>P</td>
<td>P</td>
<td>Located nearby?</td>
</tr>
<tr>
<td>P</td>
<td>S</td>
<td>Located inside?</td>
</tr>
<tr>
<td>P</td>
<td>S</td>
<td>Located nonuniformly inside?</td>
</tr>
<tr>
<td>P</td>
<td>S</td>
<td>Located nearby?</td>
</tr>
<tr>
<td>S</td>
<td>S</td>
<td>Similar segments?</td>
</tr>
<tr>
<td>S</td>
<td>S</td>
<td>Overlap?</td>
</tr>
<tr>
<td>S</td>
<td>S</td>
<td>Located nearby?</td>
</tr>
<tr>
<td>F</td>
<td>F</td>
<td>Correlated?</td>
</tr>
<tr>
<td>P</td>
<td>F</td>
<td>Higher values at locations?</td>
</tr>
<tr>
<td>S</td>
<td>F</td>
<td>Higher values inside?</td>
</tr>
<tr>
<td>P</td>
<td>VS</td>
<td>Located in segments with high values?</td>
</tr>
<tr>
<td>S</td>
<td>VP</td>
<td>Higher values inside segments?</td>
</tr>
<tr>
<td>VP</td>
<td>VP</td>
<td>Nearby values similar?</td>
</tr>
<tr>
<td>P</td>
<td>VS (c/c)</td>
<td>Located in case segments</td>
</tr>
<tr>
<td>VS (c/c)</td>
<td>S</td>
<td>Preferential overlap?</td>
</tr>
<tr>
<td>VP (cat)</td>
<td>VS (cat)</td>
<td>Category pairs differentially co-located?</td>
</tr>
<tr>
<td>LGP</td>
<td>P</td>
<td>Colocalized in 3D?</td>
</tr>
</tbody>
</table>
Making justified choices is indeed hard!

- The choice of data may influence results
  - Both source and exact version of genes might matter
  - Can sometimes justify, e.g. based on sensitivity/specificity trade-off
  - Should ideally show how results vary with choice of data
  - Should at least be very precise in what was done (accessibility, transparency, reproducibility)
Making justified choices is indeed hard (2)

• There is usually more than one possible test for a given biological question
  • The choice has to be made, and can’t be resolved automatically
  • Statistical and biological implications play together to determine what may be reasonable
  • Should at least expose the different possibilities
Hypothesis testing

- Alternative hypothesis (H1)
  - What you really want to show (more HPV in genes)

- Null hypothesis (H0)
  - A neutral baseline (HPV equally inside/outside)

- P-value
  - How likely is observation (or more extreme), given H0
  - Observation unlikely -> reject H0, left with H1
Hypothesis testing: the challenges

• Alternative hypothesis (H1)
  • What you really want to show (more HPV in genes)

• Null hypothesis (H0)
  • A neutral baseline (HPV equally inside/outside)

• P-value
  • How likely is observation (or more extreme), given H0
  • Observation unlikely -> reject H0, left with H1
How to compute p-value?

- Look at normal distribution table? Run a t-test?
- But where to put in HPV and genes?
The quest for a distribution

- Can we find a suited distribution?
  (for number of HPV sites inside genes under H0)
  - Statistician may find that “yes: a binomial distribution”
  - Would you be comfortable assuming a binomial distribution?
    Or better: Would you have any clue on the implications?
The quest for a distribution

- The implication of using a binomial distribution
  - What is binomially distributed - HPV or genes?
  - Neither.. This only applies to the measure.
  - Instead, HPV assumed independently and uniformly distributed
  - Not trivial to see, and if found: is this acceptable?
HPV integration sites
How to compute p-value?

• Look at normal distribution table? Run a t-test?

• But where to put in HPV and genes?

• Turns out that thinking about standard tests and distributions becomes awkward

• Instead, do it the modern way..
Meet Monte Carlo

• Null model:
  • How to randomize data (precise rendition of H0)
  • Where could HPV be located under H0?

• Test-statistic:
  • How to measure aspect of interest
  • Number of HPV sites located inside genes

• P-value:
  • How often is test-statistic from null model more extreme than for observation?
  • How often are 78 or more random HPV inside genes?
Monte Carlo test on “points inside segments”

- Randomize point (HPV) locations (null model)
Monte Carlo test on “points inside segments”

- Randomize point (HPV) locations (null model)
- Count random points (HPV) inside segments (genes) - test statistic

Diagram:

- HPV locations
- HPV randomized track
- Genes
Monte Carlo test on “points inside segments”

- Randomize point (HPV) locations (null model)
- Count random points (HPV) inside segments (genes) - test statistic

Histogram of test_statistic

Frequency

0.0 0.2 0.4 0.6 0.8 1.0

5 10 15 20
test_statistic

HPV randomized track

HPV

Genes
Monte Carlo test on “points inside segments”

- Randomize point (HPV) locations (null model)
- Count random points (HPV) inside segments (genes) - test statistic
- Repeat a number of times

Histogram of test_statistic
Monte Carlo test on “points inside segments”

- Randomize point (HPV) locations (null model)
- Count random points (HPV) inside segments (genes) - test statistic
- Repeat a number of times
- Build histogram

Histogram of test_statistic
Monte Carlo test on “points inside segments”

- Randomize point (HPV) locations (null model)
- Count random points (HPV) inside segments (genes) - test statistic
- Repeat a number of times
- Build histogram
- Compare with number of observed points (HPV) inside segments (genes)
Monte Carlo test on “points inside segments”

- Randomize point (HPV) locations (null model)
- Count random points (HPV) inside segments (genes) - test statistic
- Repeat a number of times
- Build histogram
- Compare with number of observed points (HPV) inside segments (genes)
- p-value = area to the right if H1 is more (if less, area to the left)

p-value = 0.08
Back to HPV and genes

- Didn’t like implications of binomial distribution?
- With Monte Carlo, you can shuffle how you like
  - Throw HPV around uniformly and independently (like binomial)
  - Keep clustering tendency of HPV (shuffle HPV spacings)
  - Keep HPV as is, only shuffle genes (in various ways)
Exploring alternative data and assumptions

- Try different gene data sources and assumptions (null models) on HPV-gene relation
- Use back button or redo functionality
- Who gets the best p-value;)

(Replay button image)
Data and assumptions matter!

- HPV inside Ensembl genes? (default assumptions)
  - Yes! ($p$-value=0.006)

- HPV inside Refseq genes? (default assumptions)
  - No! ($p$-value=0.5)

- Inside Ensembl (v2)? (Preserve inter-HPV distances)
  - Still yes ($p$-value=0.005)

- Inside Ensembl (v3)? (Randomize genes)
  - Maybe.. ($p$-value=0.02)
An example of alternative assumptions

- Duan, [...] and Noble (Nature, 2011):
  - Extensive significant 3D co-localization of functional elements, assessed by hypergeometric distribution

- Witten and Noble (NAR, 2012):
  - Hypergeometric test had unrealistic implications. Telomeres and breakpoints may not be co-located after all. (cancelled 4 of 11 findings)
Other important issues

• Selecting appropriate test statistic

• Handling confounders
Conclusion

• Genomic tracks provide a powerful, generic basis for statistical analysis

• Sophisticated statistical testing can be performed through simple means (web GUI)

• The devil may be in the details, and selection of data and assumptions can’t be outsourced
Further into statistical details: the test-statistic

- Maybe viruses integrate close, rather than inside?
- Let’s instead analyze distance to TSS!
HPV close to genes?

- Same data as for last analysis!
  - Use redo - only slight changes are needed..
- Question: “Located nearby?”
- Options looks okay !?
- “Start analysis”
Back to drawing board: the test-statistic

- For “located inside”:
  - Could simply count the number of HPV sites falling inside genes
Back to drawing board:
Must quantify “close”
But that’s trivial, sure:
Just count bp distance?!

• But which distances - not all vs all?!
But that’s trivial, sure: Just count bp distance!?

- But which distances - not all vs all?!
  - Only shortest!
But that’s trivial, sure:
Just count bp distance!?

- But which distances - not all vs all?!
  - Only shortest! From 1 to 2!
But that’s trivial, sure: Just count bp distance!?

- But which distances - not all vs all?!
  - Only shortest! From 1 to 2! But MC needs a single number..
But that’s trivial, sure: Just count bp distance!?

- But which distances - not all vs all?!
  - Only shortest! From 1 to 2! But MC needs a single number..
  - Just use sum/average of distances!!
Same degree of close?!

- Two scenarios with same (arithmetic) average...
  - Scenario A indicates relation, but not B !?
  - If so, can be captured by instead using geometric average
You try!

• Can you find a significant HPV-gene relation?
• Would you be comfortable reporting (publishing) this relation?
• If so, what would be an acceptable way to report it?
Any rules of thumb?
(for the statistical testing)

• Maybe:
  • Use test-statistic that gives best (lowest) p-value
  • Use null model that gives worst (highest) p-value

• Reasoning:
  • Use measure that best catches relation of interest
  • Use the most realistic model of nature (null model)
Conclusion

• Genomic tracks provide a powerful, generic basis for statistical analysis

• Sophisticated statistical testing can be performed through simple means (web GUI)

• The devil may be in the details, and selection of data and assumptions can’t be outsourced
Handling confounders

- Exons are associated with heightened DNA melting temperature
  - But both exons and DNA melting are also directly associated with GC content
  - Are exon regions really associated mechanistically with DNA melting, beyond the relation through a common association with GC?

- Analyzing exon-melting while controlling for confounders
Controlling for confounders

• **Tutorial 5** of “Analyze genomic tracks”
Conclusion

• Genomic tracks provide a powerful, generic basis for statistical analysis

• Sophisticated statistical testing can be performed through simple means (web GUI)

• The devil may be in the details, and selection of data and assumptions can’t be outsourced