Another type of data

- A. 57 genomic regions associated with MS from a recent GWAS study:
  - Sawcer S, et al. (2011) Genetic risk and a primary role for cell-mediated immune mechanisms in multiple sclerosis. Nature 476: 214–219.
- B. Genome-wide chromatin profiling from another recent Nature paper:
  - Ernst J, et al. (2011) Mapping and analysis of chromatin state dynamics in nine human cell types. Nature 473: 43–49.
- A + B + the Genomic HyperBrowser =
  - Disanto G, et al. (2012) Genomic Regions Associated with Multiple Sclerosis Are Active in B Cells. PLoS ONE, e32281.

You try!

- **First: Inspect the data!**
- **MS** -> https://wiki.uio.no/projects/clsi/index.php/INF-BIOX121_H14_course_material
  - (Text Manipulation: Convert delimiters to tab)
- **Chromatin states in B-cells** -> Chromatin:Chromatin state segmentation:..Gm12878..:1 Active Promoter
- **Genome**: hg18!!
- Descriptive statistics -> Enrichment:
  - > 1.0 is enriched with MS, < 1.0 is depleted with MS
- Region and scale: chromosomes

How about the other chromatin states?

- Use the "Execute batch commands" functionality:
  - Run one analysis
  - Inspect parameters of the analysis
  - Corresponding batch command line
  - Edit as needed:
    - (Put * instead of last level of track name for all subtypes)
  - Execute

One of the claims..

- “We hypothesized that Multiple Sclerosis (MS) associated genomic regions co-localized with regions which are functionally active in B cells. Results confirm the important role of B cells in MS.”

You try!

- "We hypothesized that MS associated genomic regions co-localized with regions which are functionally active in B cells. Results confirm the important role of B cells in MS."

MS -> Galaxy page "HyperBrowser lecture, MS case"
Active regions in B-cells -> Chromatin:Chromatin state segmentation:..Gm12878..: 1 Active Promoter
Genome: hg18

(see http://genome-mirror.duhs.duke.edu/ENCODE/cellTypes.html for cell type info)

Comparing MS with SE instead of AP

- Just change track from “.. active promoter” to “.. strong enhancer”
  - But: there are both states 4 and 5 “strong enhancer”
- Must get tracks to history and combine

You try!

(minor tip: Tools-Options -> Show tool search)

- Analyze overlap, MS versus combined state 4 and 5 "strong enhancer"
  - "Extract" tracks, "Remove beginning" on one of them (why?), "Concatenate datasets"
  - "Perform analysis" with MS vs combined data set

But, something isn't right!

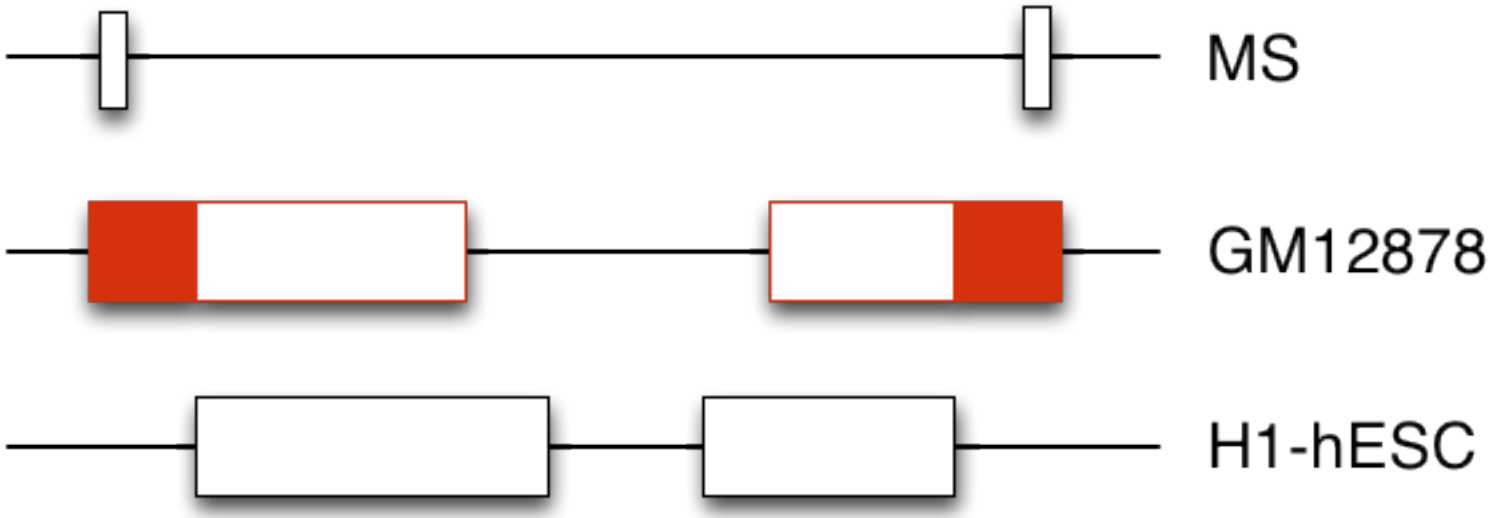
- "We hypothesized that MS associated genomic regions co-localized with regions which are functionally active in B cells. Results confirm the important role of B cells in MS."

  Do the overlap between MS and Active Promoters/Strong Enhancers really confirm a role of B cells in MS? Why not?

Only cell-specific regions!

- Some of the AP/SE regions may be common for several or all cell types
- Overlap between MS regions and such regions does not really say anything about B-cell specificity
- The question should be:
  - Do MS overlap more with AP/SE regions specific to B cells than expected by chance?
  - Let's use gm12878-specific AP as case regions and h1hesc-specific AP as control regions

Only cell-specific regions!

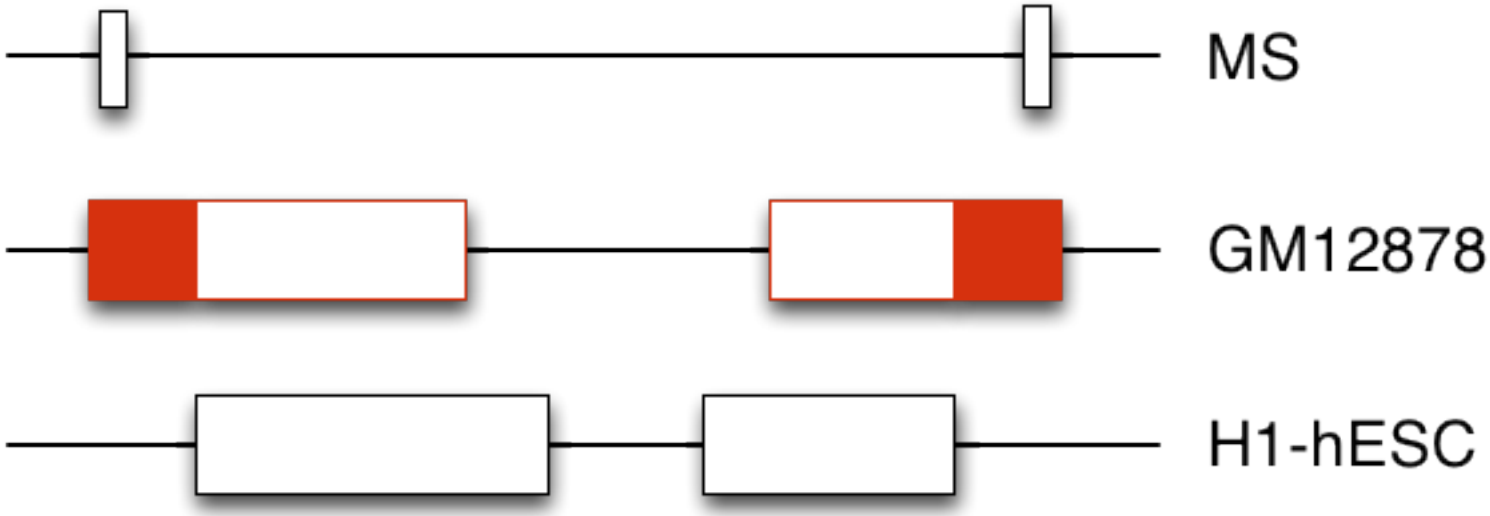


- Only look in filled red regions

You try!

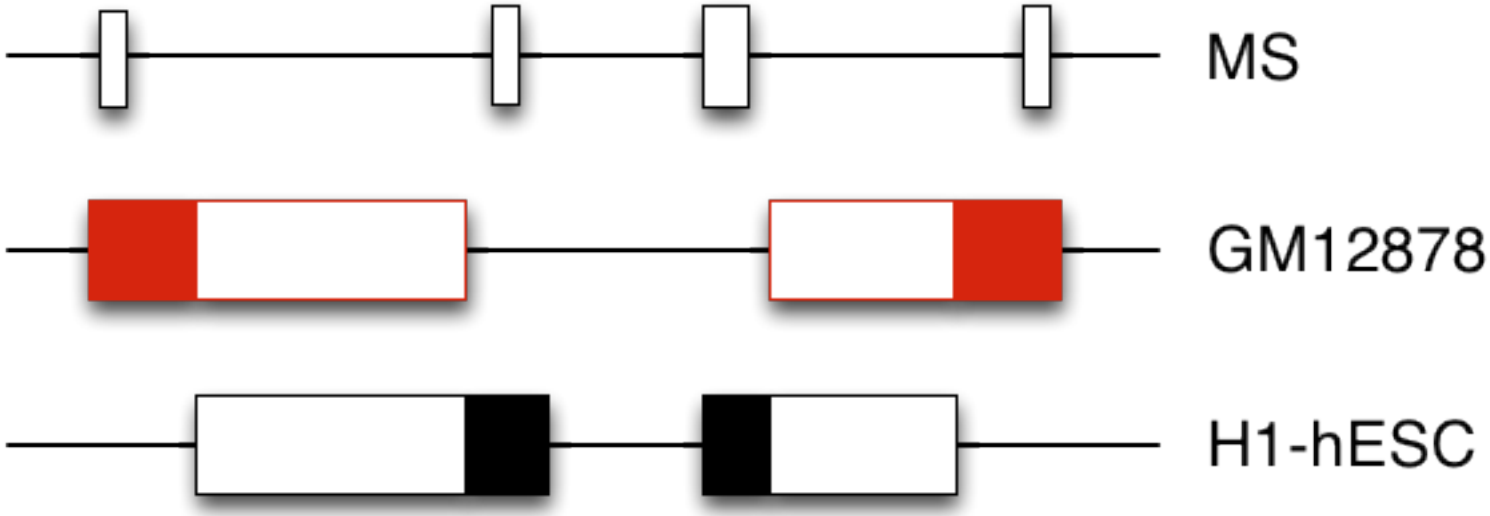
- Find cell-type specific regions
  - Extract
    - Chromatin:Chromatin state segmentation:..gm12878..:AP
    - Chromatin:Chromatin state segmentation:..H1hesc..:AP
  - "Subtract the intervals" of H1-hESC from GM12878
- Perform analysis of constructed track vs MS

Hold your horses!



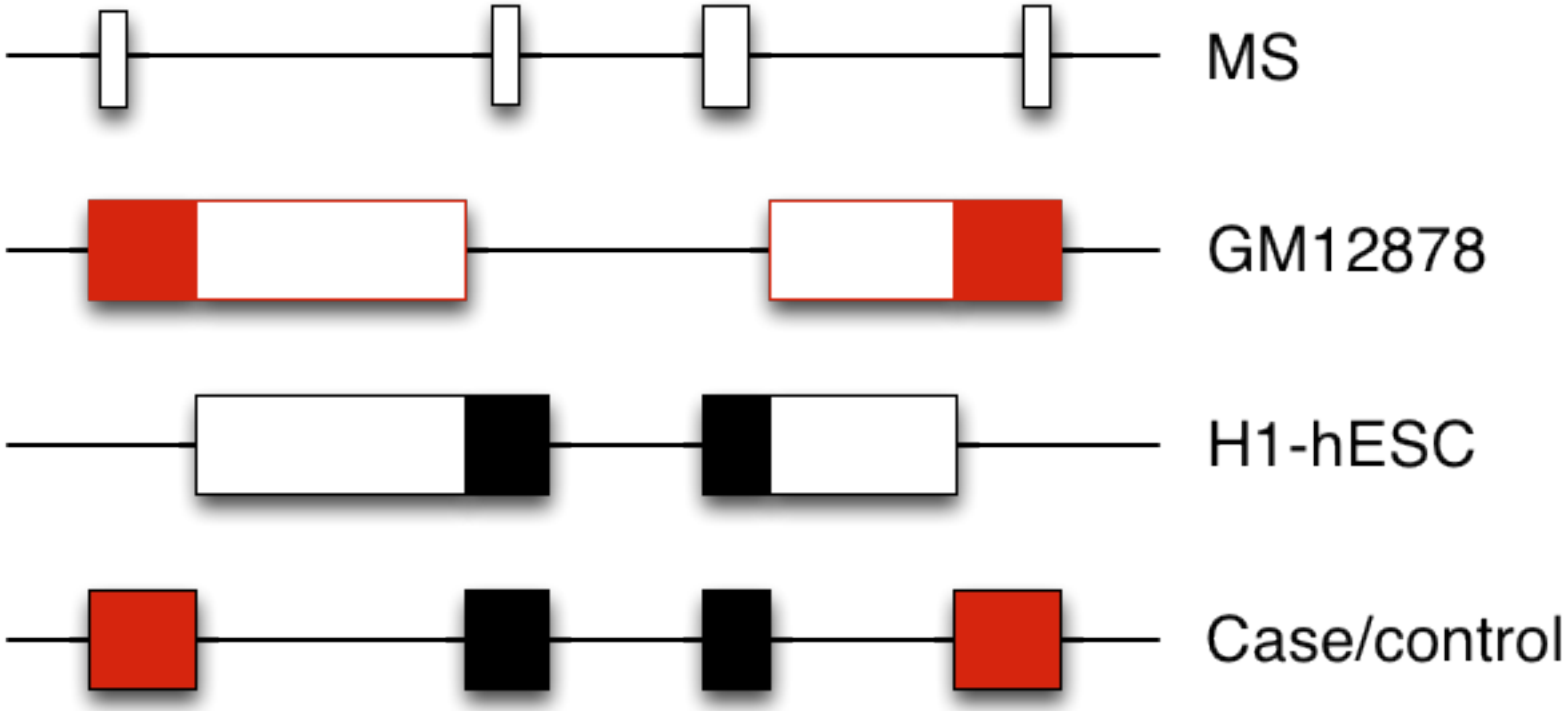
- The way we thought before

Hold your horses!



- But what about this case?
- MS regions in cell-specific regions of both GM12878 and H1-hESC

Hold your horses!



- We need a track of case/control regions!
- Case are red and control is black in this example

Hold your horses!

- What if MS likes active regions specific to all cell types (not only B cells)?
- We need control regions
- The question should be:
  - Do MS overlap more with AP/SE regions specific to B cells than such regions specific to other cell types?
  - Let's use again use gm12878-specific AP as case regions and H1-hESC-specific AP as control regions

You try!

- Create a target - control track
  - Use the already extracted ..gm12878:AP and ..h1hesc:AP tracks
  - "Combine two BED files into single case-control track"
- Perform analysis "Preferential overlap?" of constructed track vs MS

Let's introduce some reproducibility

- You may want to publish a paper, and have other researchers reproduct your results
- Maybe you want to reproduce the exact same result in a few years from now
- We have previously introduced Galaxy Pages, as a way of publishing reproducible research in Galaxy

But: clean up the mess!

- A bunch of history elements with default names
- Name of history: “Untitled history”
- A bunch of dead ends

You try!

- Only the last analysis is correct, ignore the first tries
- Copy the needed steps into a new history (Options > Copy datasets)
- Name the history
- Name the elements

You try!

- Create a Galaxy Page
  - User->Saved Pages (you will have to register a user)
  - “Add new page”
  - Click chosen name under “Title” and “edit content”
  - “Embed Galaxy object”->history
  - For now, just write very brief explanatory text
  - “Save”, “Close” and “Share or Publish”
  - Share the history with “sveinugu@gmail.com”

Summary

- What you compare with (control data/ region) is important!
- Reproducibility is important for the field, but also of practical importance to yourself
- Some simple habits and a Galaxy history is (almost) all you need
- By referring to a Galaxy page with runs and results, the analysis in your publication becomes transparent and reproducible

Conclusion (for both sessions)

- There is usually more than one way to ask and answer a biological (genomic) question
- Statistical testing in genome analysis has pitfalls and ambiguities, but often work out using MC
- Reproducibility in genome analysis is currently grave, but is within reach through habits, histories and Galaxy Pages