Another claim..

- “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.”

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

In silico analysis and reproducibility

- Bioinformaticians gets surprised every time they need to redo/modify previous analyses
- But bench biologists already know the importance of reproducibility!
- You also know that even with a detailed lab journal, reproduction is a challenge
- The question is then how this manifests itself when doing analysis on a computer

What is in silico reproducibility?

- Basically the same issues as at the bench:
  - Materials -> Data sources
  - Experiment conditions -> Analysis parameters
  - Equipment (and models) -> Programs (and versions)
- And the same challenges:
  - Are all relevant conditions described accurately?
  - Will the same materials and equipment be available?

What is the current status of reproducibility?

- Less than half of selected microarray experiments published in Nature Genetics could be reproduced
(Ioannidis et al., Nat Genet 2009)

- More than half [of surveyed papers] do not provide primary data and list neither the version nor the parameters used [for read mapping]
(Nekrutenko and Taylor., Nat Rev Genet 2012)

Why should you care?

(about making your analyses reproducible)

- Because it's the right thing to do!
- .. and the one that's struggling with its reproduction is often the future you
- Journals are becoming aware of the issues
- Reviewers may value it
- Anyway, it's the same as at the bench..

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

Let's introduce some reproducibility

- Your analysis of MS vs "strong enhancer" will now consist of a few steps
- Maybe you want to reproduce the exact same result in a few years from now
- Time for an "in silico lab" journal
- Since you're already at a computer, let's write it there..

You try!

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

- Create a lab journal for the analysis
  - Send to me by email: sveinung.gundersen@medisin.uio.no
  - (just to get an impression, will not be used for any purpose)
- Analyze overlap, MS versus combined state 4 and 5 "strong enhancer"
  - "Extract" tracks, "Remove beginning" on one of them (why?), "Concatenate datasets"

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 hepg2-specific AP as control regions

You try!

- Find cell-type specific regions
  - Extract ..gm12878:AP and ..hepg2:AP
  - “Subtract the intervals” of hepg2 from gm12878
- Perform analysis of constructed track vs MS

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 hepg2-specific AP as control regions

You try!

- Create a target - control track
  - Use the already extracted ..gm12878:AP and ..hepg2:AP tracks
  - Use the shared workflow ("Shared Data"):

    Create case-control track
- Perform analysis "Preferential overlap?" of constructed track vs MS

Now the lab journal becomes more involved..

- You would as before have to document data sources and parameters used in analysis
  - But now also several steps (and implicit methods) of pre-processing data
- But what you need is anyway there already, in the form of the Galaxy history
  - Shouldn't be necessary to duplicate this information, also with risk of introducing errors..
  - A Galaxy Page, linking to history, could be the solution!

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” via Link..

Ten simple rules for reproducibility

- Whenever making a claim, note a reference to supportive data
  - ".. MS occur preferentially inside AP in B-cells [hist:HbLecture-8] .."
- For every result of interest, keep track of how it was produced
  - Solved automatically by redo-functionality if using Galaxy
- Record all intermediate results, when possible in readable formats
  - Intermediate steps of creating case-control are stored as history elements
- Provide public access to scripts, runs and results
  - Provide link to Galaxy Page that embed histories with all runs and results

Ten simple rules for reproducibility (cont.)

- Use executable documentation and verification
  - Galaxy histories document analysis and are executable
- Generate hierarchical analysis output, allowing layers of increasing detail to be inspected
  - HyperBrowser provides conclusion, full table and local results
- Always store raw data behind plots
  - Result plots of HyperBrowser analyses come with underlying numbers

Ten simple rules for reproducibility (cont.)

- Archive all external programs and custom scripts, in the versions that were used
  - Galaxy provides this publicly and explicitly. HyperBrowser is version controlled and can be contacted.
- Avoid manual, non-trackable procedures
  - We have performed all analysis steps in the Galaxy system
- For analyses including randomness, note underlying random seeds
  - HyperBrowser allows a particular random seed to be set (results are then deterministic, like a frozen snapshot of randomness)

Summary

- Reproducibility is important for the field, but also of practical importance to yourself
- The current status of reproducibility is grave
- 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