Measuring the quality and reproducibility of Hi-C data, update on progress

G. Gürkan Yardımcı

Noble Lab

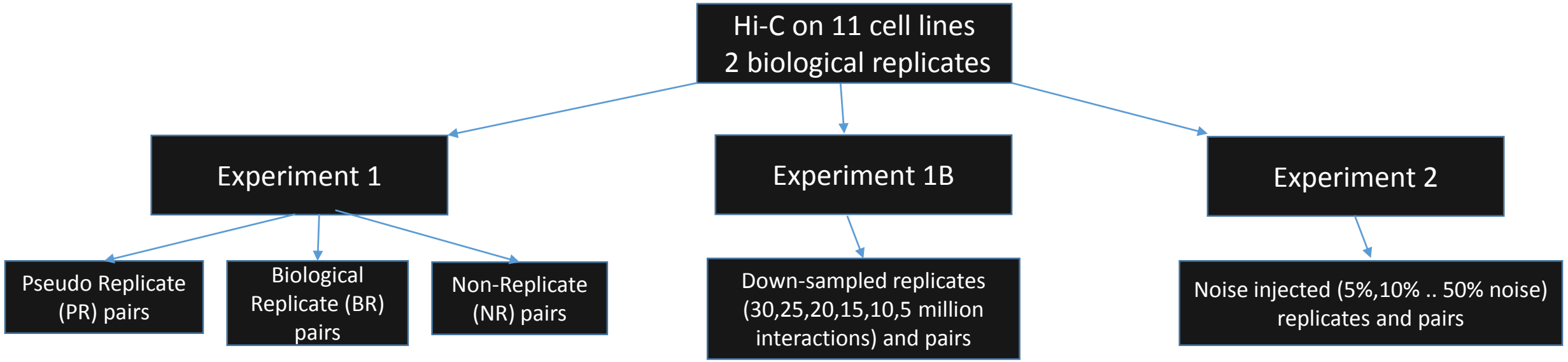
Genome Sciences

University of Washington

Outline

- **Data release**
- Reproducibility analyses
- Quality control analyses

Complete data is now available



- Exp1, Exp1B and Exp2 are to be used for reproducibility analyses. Anonymized replicate pairs are now available
- Exp1B and Exp2 are to be used quality control analyses. Anonymized replicate names will be released next week.

Data available at : http://noble.gs.washington.edu/~gurkan/data/Encode_CompleteRelease/

Some reproducibility analyses are optional

- Sample lines from table containing anonymized replicate pairs:

Matrix144 Matrix156 required

Matrix166 Matrix156 required

Matrix170 Matrix256 optional

Matrix170 Matrix118 optional

- For foreseeable analyses we might conduct, extra pseudo-replicates pairs and large number of non-replicate pairs have been designated. Those pairs are optional.
- Previously, we generated a single pseudo-replicate pair and used only first biological replicates for non-replicates.

Outline

- Data release
- **Reproducibility analyses**
- Quality control analyses

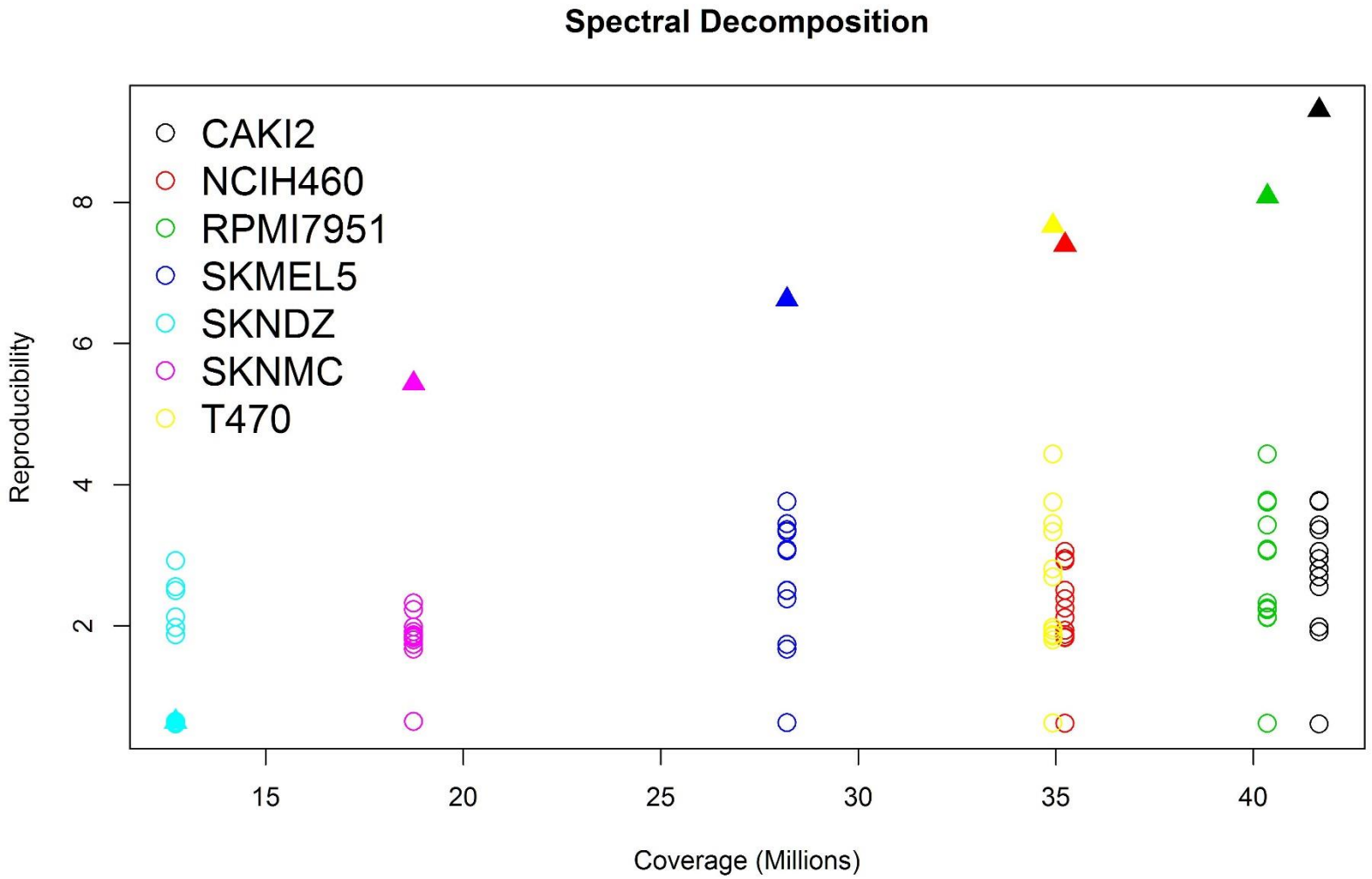
Future reproducibility analyses

- To be done:
  - Calculate reproducibility for all anonymized replicate pairs
  - Run every chromosome (previously we used subsets of chromosomes)
- Write a description of your method for the manuscript
- Link to timeline

Reproducibility metrics require interpretable thresholds or significance values.

- Initial analyses showed that reproducibility metrics do a good job recapturing expected ranks for both experiments.
- Open question: for a given replicate pair we have not seen before, what numerical measure of reproducibility metric indicates sufficient reproducibility?

Biological replicate pairs are more reproducible compared to non-replicate pairs



Rectangles: Biological replicate **pair**
Circles: Non-replicate **pairs**

Comparing reproducibility of PRs, BRs, and NR pairs

- Two potential ideas for comparing the reproducibility of a given biological replicate pair (BRp):

1. Compare against reproducibility of pseudo-replicate pairs
   - Desired outcome: BRp reproducibility close to PRp reproducibility distribution
   - Compare against a (null) distribution of control Hi-C experiments
   - Easy to generate PRs from BRs, control not always available
2. Compare against reproducibility of non-replicates
   - Desired outcome: BR reproducibility is higher than NR reproducibility distribution
   - NRs at a given coverage are not always available, some NR pairs might not be ‘null’
   - We can potentially investigate simulating NRs

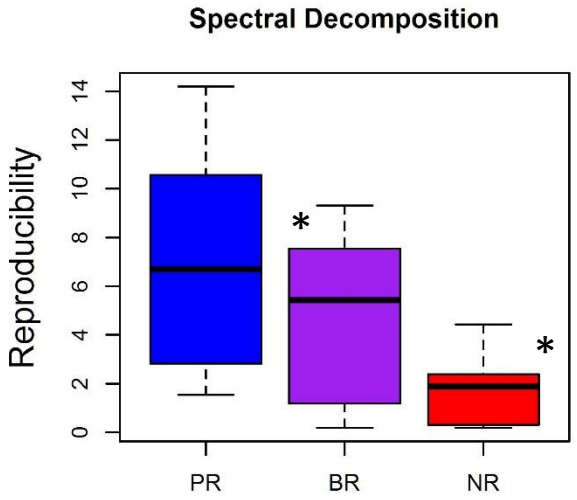
Outline

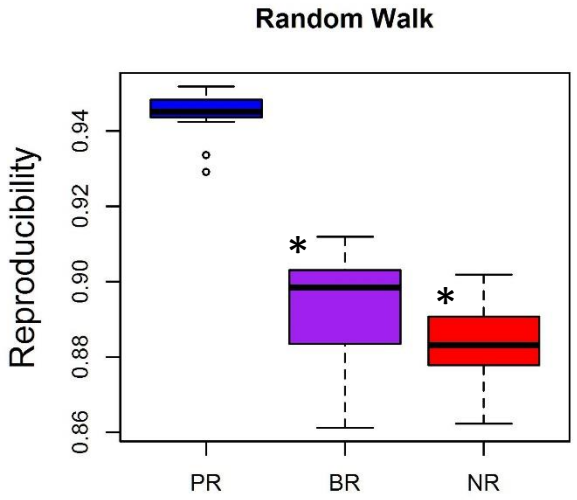
- Data release
- Reproducibility analyses
- **Quality control analyses**

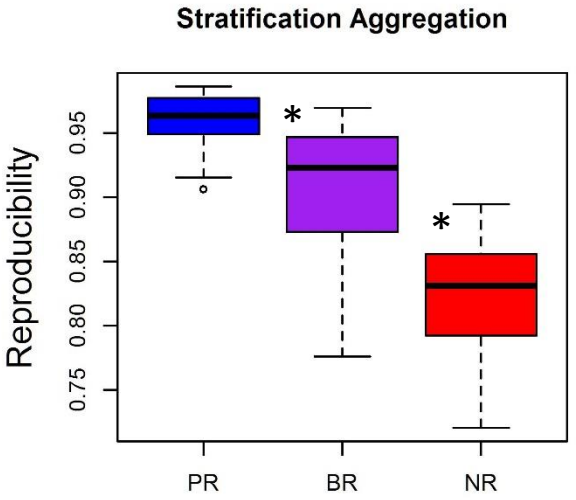
We have different tasks for different QC metrics

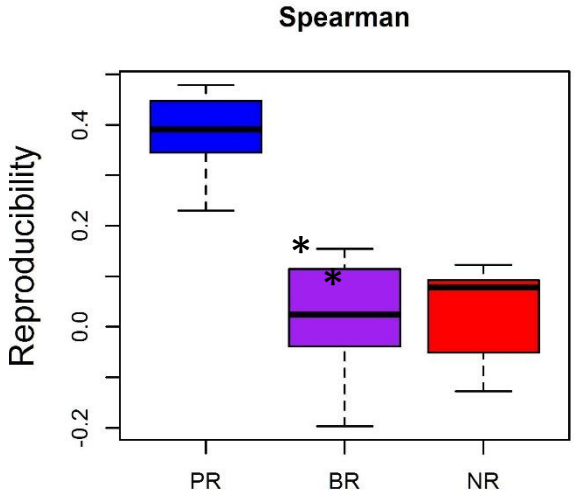
- Spatial consistency
  - Run the analyses on newly released data
- Significant contacts
  - Run the analyses on old + newly released data
  - Investigate shared contacts
- TAD calling
  - Run the analyses on old + newly released data
  - Compare two or more TAD callers
- Different groups ran different subsets of earlier datasets, I will generate a list of replicates to be analyzed for each task
- Link to timeline

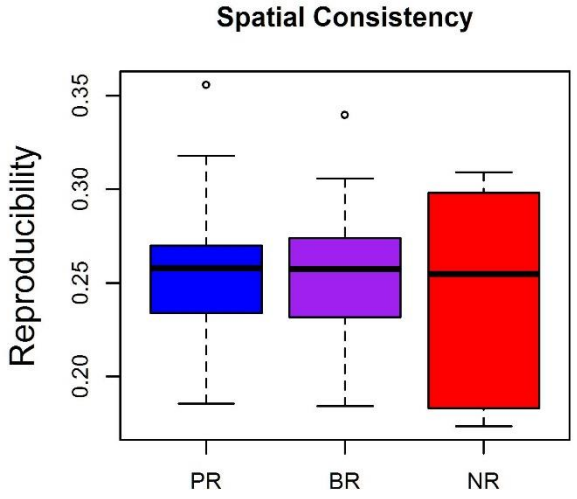
Exp1: Most metrics can separate replicate pair types











- PR: Pseudo-replicate pair
- BR: Biological replicate pairs
- NR: Non-replicate pairs
- Asterisks indicate significant separation between PR vs. BR and BR vs. NR