Automated ChIP-Seq Analysis and Reporting Pipeline

Stephen Kelly[1,2], Igor Dolgalev[1-5], Charalampos Lazaris[3-5] & Aristotelis Tsirigos[1-5]

[1]Applied Bioinformatics Center & [2]Genome Technology Center, NYU School of Medicine, NY 10016, USA, [3]Department of Pathology, [4]NYU Cancer Institute and Helen L. and Martin S. Kimmel Center for Stem Cell Biology, [5]Center for Health Informatics & Bioinformatics,

Objective

Develop a reproducible ChIP-Seq pipeline that can adapt to varying numbers of samples, treatment groups, and parameter sets while allowing for user expansion with custom programs and analysis tasks.

Pipeline usage summary

1. Create new directory for analysis from a clone of the pipeline repository (Figure 8)
2. Set input files (fastq, bam)
3. Generate sample sheet with pipeline-provided scripts
4. (Optional) Modify parameters as needed, add custom tasks
5. Run pipeline
6. Compile automatic report

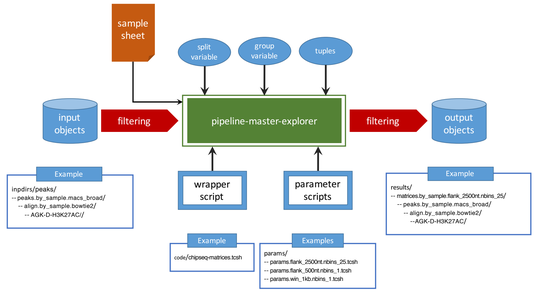

Figure 1: ChIP-Seq pipeline programmatic workflow. All sets of parameters are evaluated for each pipeline task in a combinatorial fashion

ChIP-Seq Pipeline tasks	Programs Used
Alignment	bowtie2
Alignment Stats	bowtie2, **R**
Quality Control	FastQC, deepTools
Peak Calling	MACS
Matrix Correlation	**GenomicTools**
Principal Component Analysis	**R**
Heatmap Clustering	**R**
Differential Peak Binding	DiffBind
Visualization	**R**
Automatic Reporting	**R, LaTeX 2ε**

Table 1: ChIP-Seq pipeline standard components. Internally developed methods listed in bold.

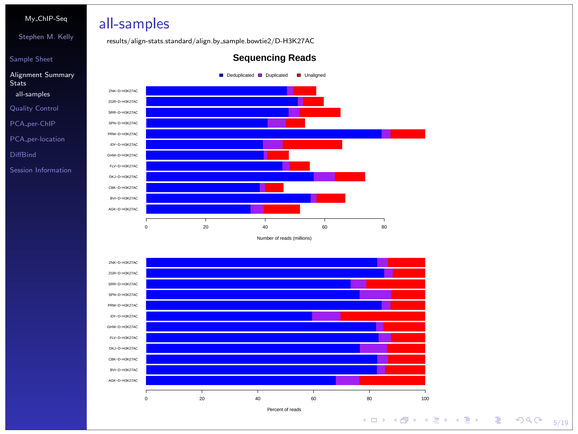

Figure 2: Alignment summary statistics

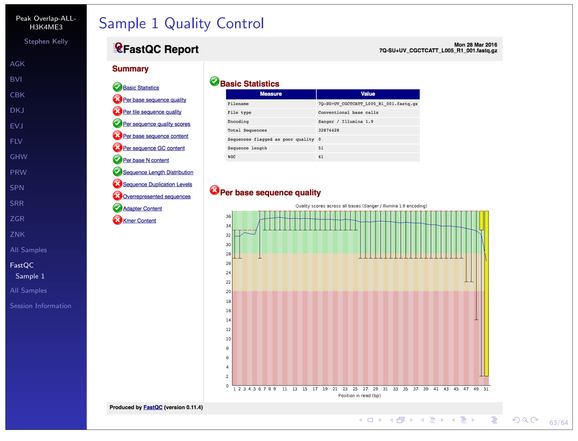

Figure 3: Quality control metrics

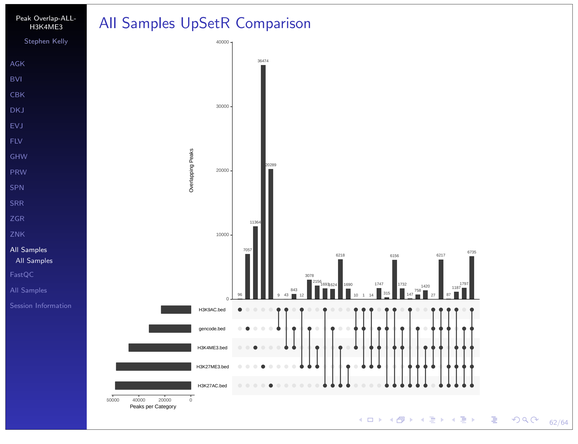

Figure 4: Peak overlap UpSet plot

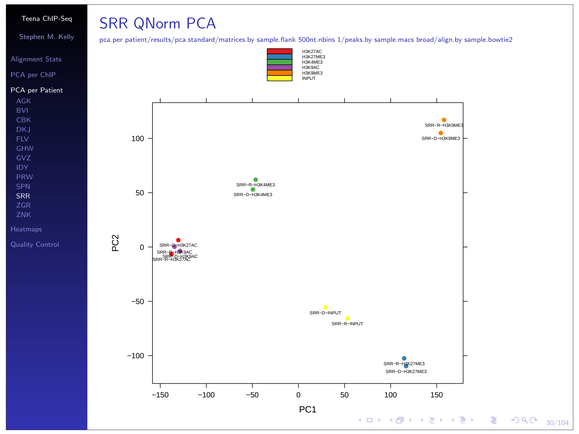

Figure 5: Principal component analysis

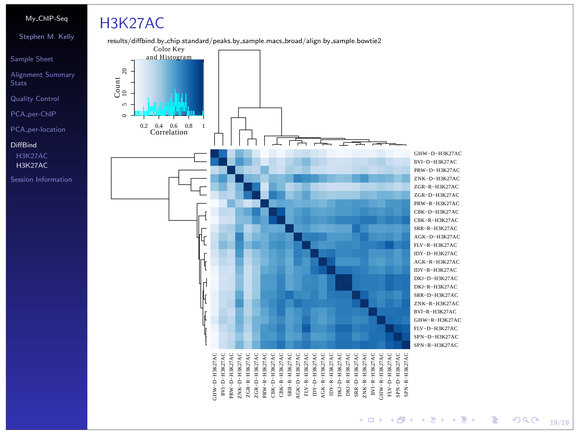

Figure 6: Differential binding heatmaps

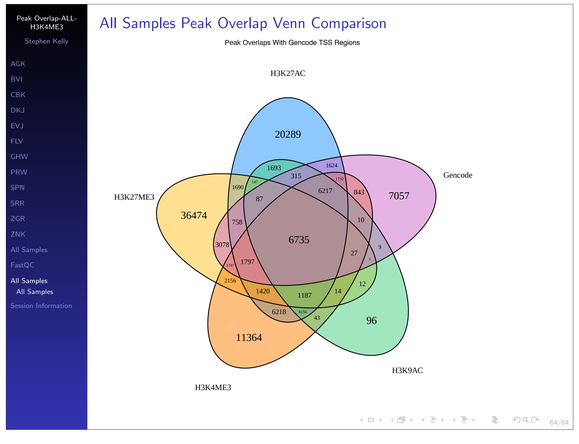

Figure 7: Peak overlap venn diagram

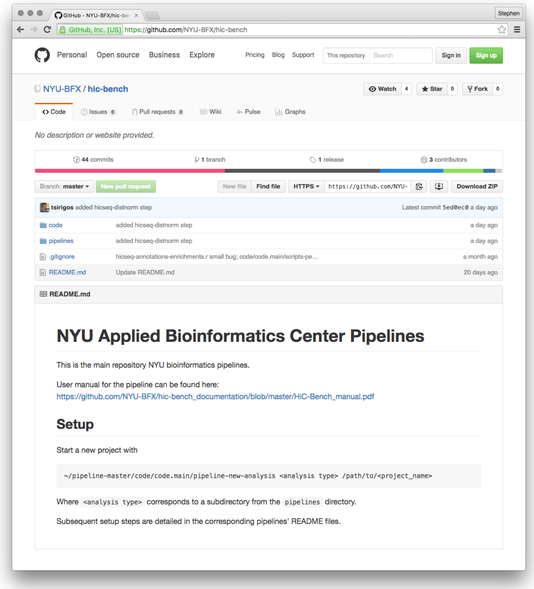

Figure 8: Our ChIP-Seq pipeline is part of the HiC-Bench software package, available on GitHub

Future Developments

- Automated motif analysis with HOMER and MEME-ChIP
- Contaminant analysis
- Peak enrichment analysis

Acknowledgements

This work used computing resources at the Laura and Isaac Perlmutter Cancer Center, which is supported by Cancer Center Support Grant P30CA016087. A. Tsirigos was supported by a Research Scholar Grant, RSG-15-189-01-RMC from the American Cancer Society. This work also used computing resources at the High Performance Computing Facility of the Center for Health Informatics and Bioinformatics at the NYU Langone Medical Center.

Software

- Web: http://www.med.nyu.edu/ocs/applied-bioinformatics-center
- GitHub: https://github.com/NYU-BFX/hic-bench
- Zenodo: https://zenodo.org/record/47676
- Contact: **stephen.kelly@nyumc.org**, **aristotelis.tsirigos@nyumc.org**