Exploring plant epigenomics data

Decoding the regulatory code in plants based on big epigenomics data

***D. Chen, K. Kaufmann**, Humboldt-Universität zu Berlin*

In Short

- Standardized computational pipelines are developed to analyze big epigenomics data.
- More than 25,800 epigenome and >32,000 transcriptome datasets in plants are collated from public databases.
- We establish an integrated platform to the plant community to explore plant epigenomes.

Major research efforts have in the past ten years resulted in generation of around 90,000 experimental epigenomics data series, corresponding to more than 2.5 million biological samples (data in NCBI GEO). While centralized platforms exist that provide standard analysis of epigenomics data in the animal field (e.g. ENCODE, [1]), no such initiative exists in plants. This is a major bottleneck in facilitating efficient use of the existing data by the international plant science community, since multiple tools exist to analyze epigenomics data, and the results are not readily comparable. A systematic analysis of existing data would also allow to assign standardized quality measures and integrative-level annotations. Our lab has a strong focus on epigenomic research, focusing on developmental and evolutionary dynamics of gene-regulatory processes in plants, and studying transcription factors, epigenetic regulators as well as histone modifications ([2], [3], [4], [5], [6], [7], [8], [9]). Therefore, establishing a platform that allows access to standardized data would greatly facilitate our research, and that of many colleagues in the plant science community.

While we have over the last year made significant progress in setting up the pipelines for analysis of RNA-seq, ChIP-seq, DNAse-seq/ATAC-seq and DAP-seq data (see Figure 1; [10]), we are currently in strong need for computing time to apply these pipelines for systematic analysis of available data. Additionally, we have implemented services for the analysis of BS-seq, Hi-C and smRNA-seq data. The goal of this project is therefore to finalize the pipeline, and to systematically apply this methodology for analysis of epigenomes in extended group of model and crop plant species. In this regard, we focus on 64 plant species for which epigenomic datasets based on high-throughput sequencing technology are currently available. We have collected more than 25,800 epigenomic datasets (see Figure 2) for these plant species from public databases and are intensively analyzing these data using our standardized computational pipelines. So far, we have successfully analzyed nearly two thirds of these epigenomic datasets by using the HLRN server. In the continuation of this project, we would also like to extend our analyses by including >32,000 transcriptome datasets.

WWW

https://www.hu-berlin.de/

More Information

[1] The ENCODE project https://www.encodeproject.org/

[2] Kaufmann K, Muino JM, Jauregui R, Airoldi C a, Smaczniak C, Krajewski P, Angenent GC. Target genes of the MADS transcription factor SEPALLATA3: integration of developmental and hormonal pathways in the Arabidopsis flower. *PLoS Biol* **7**, e1000090 (2009). doi:10.1371/journal.pbio.1000090

[3] Kaufmann K, Wellmer F, Muino JM, Ferrier T, Wuest SE, Kumar V, Serrano-Mislata A, Madueno F, Krajewski P, Meyerowitz EM, et al. Orchestration of Floral Initiation by APETALA1. *Science* **328**, 85-89 (2010). doi:10.1126/science.1185244

[4] Immink RGH, Pose D, Ferrario S, Ott F, Kaufmann K, Valentim FL, de Folter S, van der Wal F, van Dijk ADJ, Schmid M, et al. Characterization of SOC1's Central Role in Flowering by the Identification of Its Upstream and Downstream Regulators. *Plant Physiol* **160**, 433-449 (2012). doi:10.1104/pp.112.202614

[5] Pajoro A, Madrigal P, Muino JM, Matus J, Jin J, Mecchia MA, Debernardi JM, Palatnik JF, Balazadeh S, Arif M, et al. Dynamics of chromatin accessibility and gene regulation by MADS-domain transcription factors in flower development. *Genome Biol* **15**, R41 (2014). doi:10.1063/1.3382344

[6] Muino JM, de Bruijn S, Pajoro A, Geuten K, Vingron M, Angenent GC, Kaufmann K. Evolution of DNA-Binding Sites of a Floral Master Regulatory Transcription Factor. *Mol Biol Evol* **33**, 185-200 (2016). doi:10.1186/gb-2014-15-3-r41

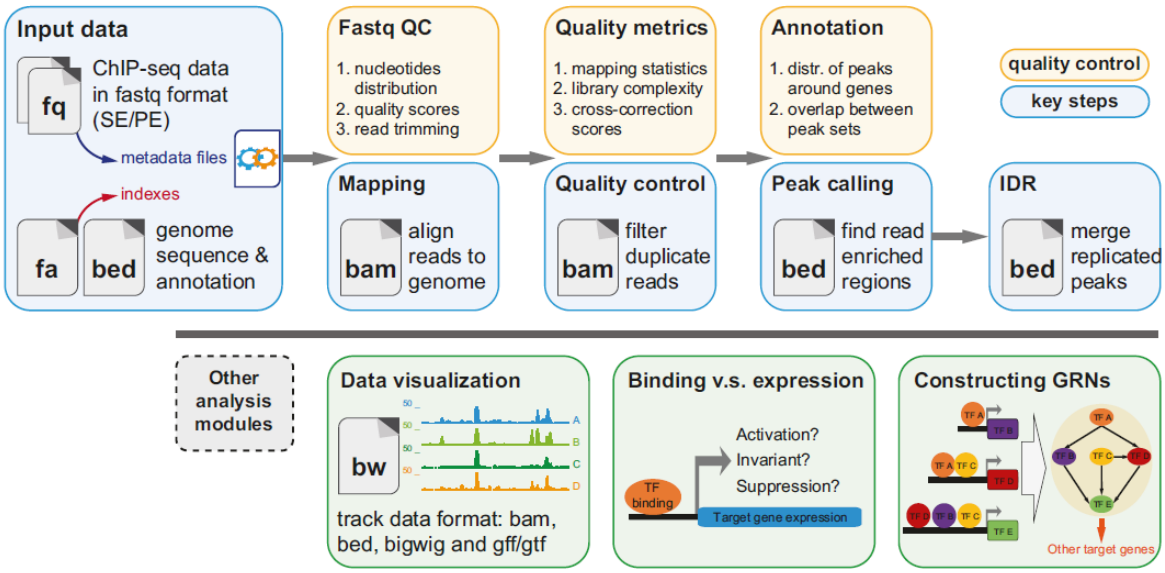

Figure 1: *Conceptual flowchart of data analysis. Top panel: main parts of a typical ChIP-seq data analysis. Blue boxes indicate required steps and yellow boxes optional steps. The goal of these steps is to obtain a list of peaks (read-enriched regions) to represent genome-wide TFBSs for a particular TF. Figure from [10].*

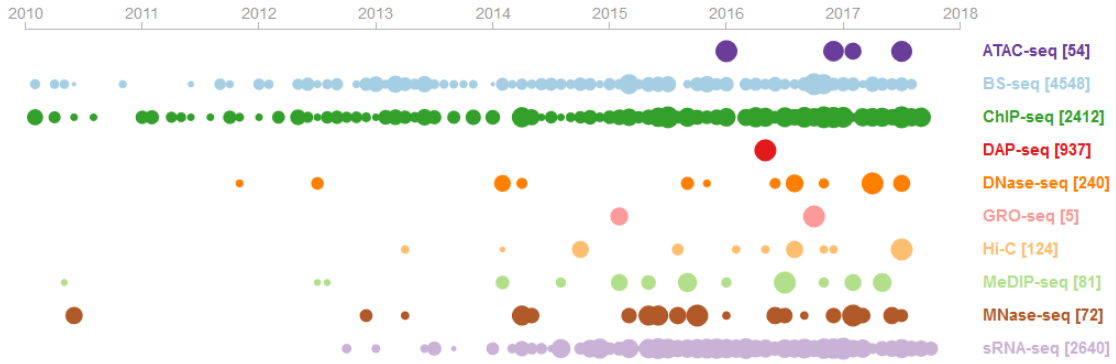

Figure 2: *Summary of epigenomics datasets available in plants.*

[7] Yan W, Chen D, Kaufmann K. Molecular mechanisms of floral organ specification by MADS domain proteins. *Curr Opin Plant Biol* **29**, 154-162 (2016). doi: 10.1016/j.pbi.2015.12.004

[8] Chen D, Yan W, Fu L, Kaufmann K. Architecture of gene regulatory networks controlling flower development in *Arabidopsis thaliana*. *Nat Commun* (2018). doi:10.1038/s41467-018-06772-3.

[9] Yan W, Chen D, Schumacher J, Durantini D, Engelhorn J, Chen M, Carles C, Kaufmann K. Dynamic control of enhancer activity drives stage-specific gene expression during flower morphogenesis. *Nat Commun* (2019). doi: 10.1038/s41467-019-09513-2

[10] Chen D, Kaufmann K. Integration of Genome-Wide TF Binding and Gene Expression Data to Characterize Gene Regulatory Networks in Plant Development. *Methods Mol Biol (Clifton, NJ)* **1629**, 239 (2017). doi:10.1007/978-1-4939-7125-1_16

Project Partners

L.Y. Fu; Humboldt-Universität zu Berlin, Germany; D.H. Hu and M. Chen; Zhejiang University, China.

Funding

the Alexander-von-Humboldt foundation and the BMBF