Characterization of *Escherichia coli* K-12 regulatory networks by bioinformatics integration of high-throughput data

Claire Rioualen[1,2], Julio Collado Vides[2], Jacques van Helden[1]

1. Theories and Approaches of Genomic Complexity (TAGC), Inserm U1090 , Aix-Marseille Université, Marseille, France
2. Computational Genomics Research program, Center for Genomics Sciences, Universidad Nacional Autónoma de México, Cuernavaca, Mexico

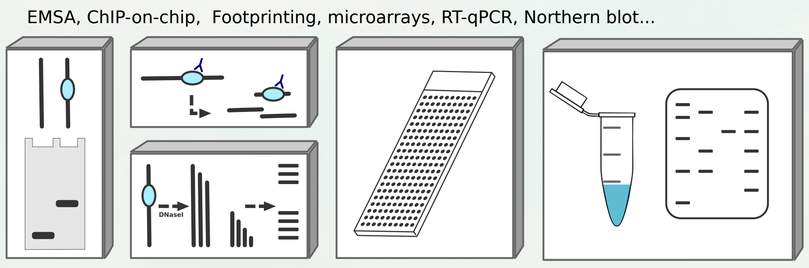
Transcriptional regulation in Bacteria

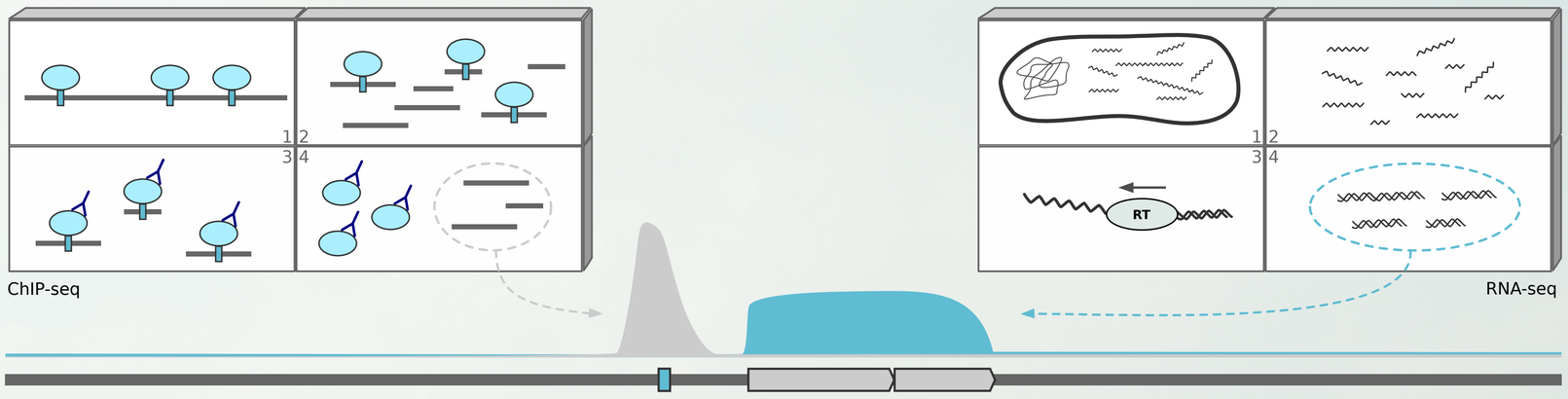
Bacterial genomes are organized into operons, which are defined as a set of two or more genes, that are co-transcribed as a single poly-cistronic unit (Jacob and Monod, 1961). The mRNA can then be translated into several distinct proteins. An operon can contain one or several transcription start sites (TSS) and one or several terminators, giving rise to differents transcription units.

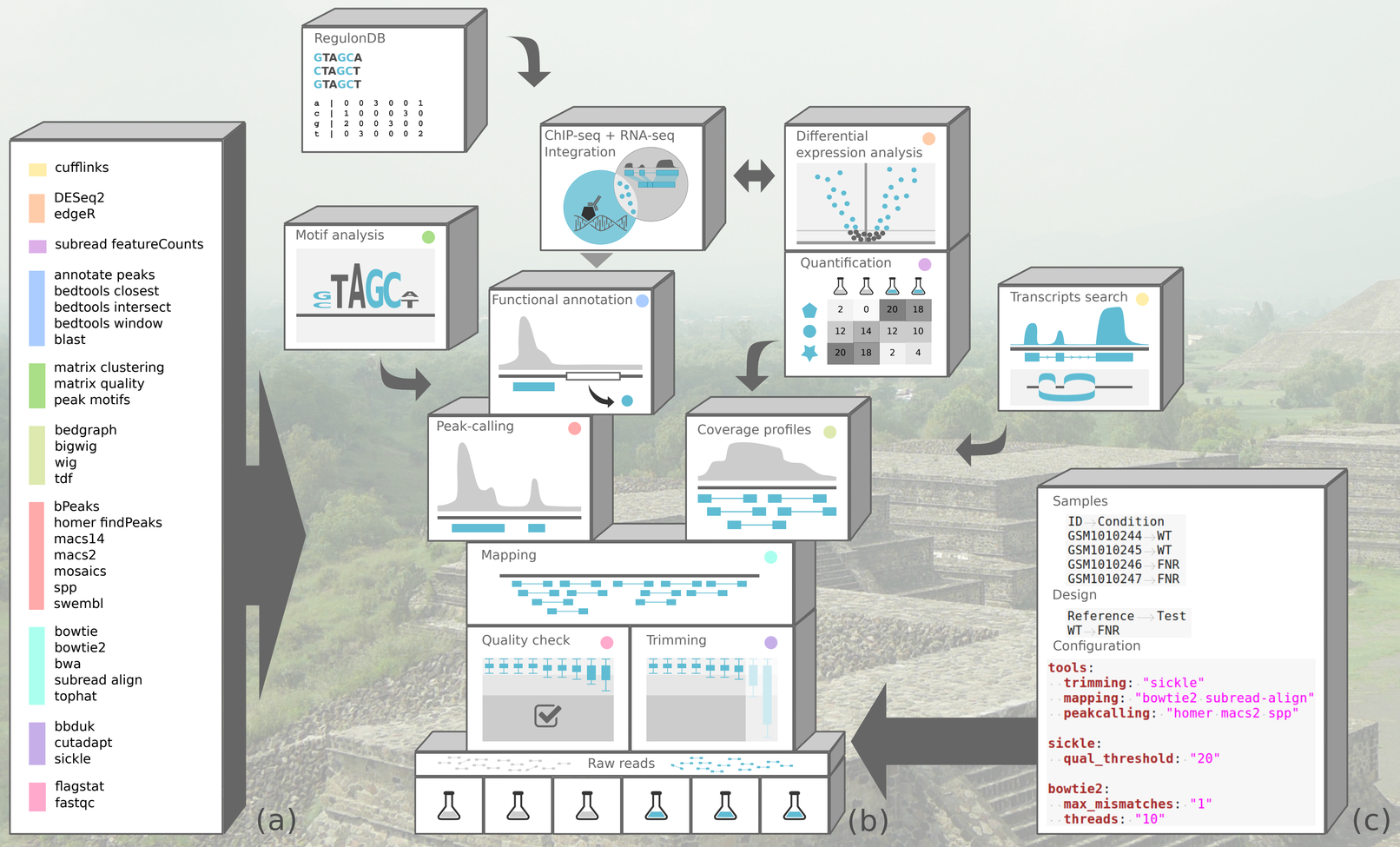
Transcriptional regulation can be achieved through a variety of mechanisms. One of the most important ones is the binding of transcription factors on specific sites of the DNA, called transcription factor binding sites (TFBS). By interfering with the recruitment or action of the RNA polymerase complex, it can activate or repress the transcription of surrounding genes.

Gene regulation has already been well characterized in *Escherichia coli* K-12. RegulonDB (Gama-Castro et al., 2016) is a database on *E.coli* transcriptional regulation, that has been accumulating this knowledge for more than 20 years, by manual curation of published literature on low throughput experiments.

TF binding and gene transcription can be characterized at a genome scale by using NGS technologies like ChIP-seq and RNA-seq. However, so far these technologies have been only very sparsely applied to bacteria, including *E.coli*.







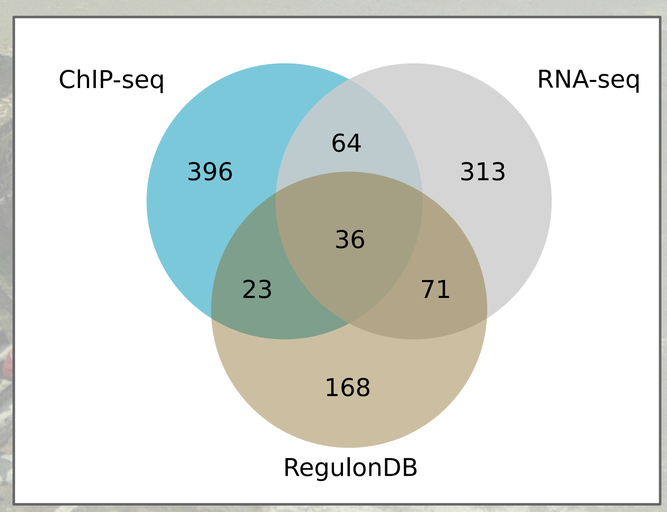
NGS data analysis relies on complex successions of computing tasks and can be affected by the tools used, their parameterization, the full availability of the metadata and the proper organisation of the data files. Ensuring the reproducibility of the analyses can prove to be tedious, but nevertheless indispensable.

We have developed a library of rules for NGS analysis (a) using the Snakemake workflow engine (Köster and Rahmann, 2012). They can be used interchangeably, and be assembled in a modular way into a variety of workflows (b). Finally, the choice of tools and paramaters can be customized easily by using configuration files (c).

Following this logic, we have also developed ready-to-use workflows for the analysis of RNA-seq data, ChIP-seq data, and the integration of transcriptional regulatory data from the RegulonDB database: TFBS, motifs, position-specific scoring matrices... Although our primary focus is the analysis of bacterial genomic data, this framework is totally adapted to process other types of data, for it uses standard tools and file formats.

Selected results

We re-analyzed data from a genome-scale analysis of the FNR transcription factor (Myers et al., 2013), a DNA-binding protein that regulates a large family of genes involved in cellular respiration and carbon metabolism during conditions of anaerobic cell growth.



After integration of ChIP-seq, RNA-seq and data from RegulonDB on the FNR transcription factor, 36 genes fell in the intersection between the three gene lists. The 64 genes reported by both ChIP-seq (FNR binding) and RNA-seq (FNR transcriptional response) but not annotated in RegulonDB are likely to be new direct target genes.

The 313 genes detected as differentially expressed without any associated peaks are likely to include indirect FNR targets, or they could be regulated through distant binding of the TF and specific DNA conformation. The 396 genes associated with ChIP-seq peaks without transcriptional response can result from different effects: false positives of the peak calling, non-functional binding of the FNR factor (missing co-activator, co-binding of a repressor)...

Motif discovery, performed by using RSAT peak-motifs (Thomas-Chollier et al, 2012; Medina-Rivera et al., 2015) revealed the presence of two over-represented motifs, corresponding to FNR and CRP transcription factors. Those TFs are known to co-regulate a number of genes (Myers et al., 2013; Gama-Castro et al., 2016).

Perspectives

Although NGS datasets targeting *Escherichia coli*'s TFs are still sparse, the number is increasing (Aquino et al., 2017). Having such tools at hand will allow to be ready for massive and accurate analysis of the data. Perspectives of this work include optimizing the peak-calling process in order to cater to bacterial genome specificities, by taking advantage of the manually-curated knowledge of well-characterized TFs (Gama-Castro et al., 2016), and the development of tools for automated biocuration of future NGS data (Santos-Zavaleta et al., 2018).

By integrating genome-binding data and transcription data targetting many TFs in *E.coli* K-12, we will be able to characterize new TF targets, unknown genes' function, and identify regulatory relationships involving cooperation, competition, or distant binding of TFs. It will also alow to identify new operons and new transcription units, and reach a very exhaustive knowledge of *E.coli*'s 4,700 genes and 300 TFs. Not only could it shed light on new gene regulatory mechanisms, but it would also be a great improvement in the field of comparative genomics.

References

Aquino et al. (2017) Coordinated regulation of acid resistance in Escherichia coli. BMC Systems Biology.
Gama-Castro et al. (2016) RegulonDB version 9.0: high-level integration of gene regulation, coexpression, motif clustering and beyond. Nucleic Acids Research.
Köster & Rahmann (2012) Snakemake-a scalable bioinformatics workflow engine. Bioinformatics.
Medina-Rivera et al. (2015) RSAT 2015: Regulatory Sequence Analysis Tools. Nucleic Acids Research.
Myers et al. (2013) Genome-scale Analysis of Escherichia coli FNR Reveals Complex Features of Transcription Factor Binding. PLoS Genetics.
Santos-Zavaleta et al. (2018) Towards a unified resource for transcriptional regulation in Escherichia coli K-12: Incorporating high-throughput-generated binding data within the classic framework of regulation of initiation of transcription in RegulonDB.
Thomas-Chollier et al. (2012) RSAT peak-motifs: Motif analysis in full-size ChIP-seq datasets. Nucleic

Acknowledgements

France Génomique, NIH grant GM0110597 & FOINS-CONACYT - Fronteras de la Ciencia 2015 - ID 15