# Meta-structure of a Bacterial Genome

Byung-Kwan Cho, Karsten Zengler, Yu Qiu, Young-Seoub Park, Eric M. Knight, Christian L. Barrett,
and Bernhard Ø. Palsson

*Short Abstract* — **Bacterial genomes are organized in terms of structural and functional components. These components include promoters, transcription start (TSSs) and termination sites (TTSs), open reading frames (ORFs), regulatory non-coding regions (RNCRs), untranslated regions (UTRs) and transcription units (TUs), that together comprise the meta-structure of a genome. The meta-structure of the *Escherichia coli* K-12 MG1655 genome is described here and it was obtained by iterative integration of multiple genome-scale measurements. The meta-structure is comprised of 3,139 independently addressable modular units amounting to an experimental annotation of the genome. The meta-structure provides a foundation on which genome-scale transcriptional and translational regulatory networks are based.**

*Keywords* — **Meta-structure, transcription unit, ChIP-chip, gene expression profiling, proteomics, transcriptional regulatory network, high-throughput data integration**

## I. Purpose

TREMEMDOUS progress has been made toward determining whole genome sequences of bacteria as well as in describing their transcriptomes and proteomes during last decade [1]. However, the in-depth organizational structure, or meta-structure, of bacterial genomes has not yet been fully elucidated. Bacterial genomes are highly organized in various structural and functional components. These components include (but not limited to) promoters, transcription start (TSSs) and termination sites (TTSs), open reading frames (ORFs), regulatory non-coding regions, untranslated regions (UTRs), and transcription units (TUs) that all together comprise the genomic meta-structure. Since sequence information by itself is not suitable for a comprehensive elucidation of these components, multiple simultaneous genome-scale measurements are therefore needed to determine all these components, their location, and relationship to the genome sequence. Here we describe a systems approach that iteratively integrates multiple genome-scale measurements on the basis of genetic information flow to identify meta-structural components and map those onto the genome sequence.

## II. Results

We experimentally determined a fully integrated modular model of the *E. coli* genome that is represented by its meta-structural components. The modular model is composed of 3,139 independently addressable modular units, which represent (i) promoter regions obtained from RNA polymerase ChIP-chip, (ii) TSSs from 5'-RACE using high-throughput sequencing with unique RNA adapter, (iii) transcribed regions from gene expression profiling on whole-genome tiling microarrays, (iv) ORFs from proteomics using high-resolution mass spectrometry. The modular genome model represents a comprehensive scaffold describing multifunctional states of the bacterial genome. This model can be readily utilized for determining TU and subsequently the TRN. We determined 4,661 TUs, of which 3,946 (~86%) were fully supported by all meta-structural components. This represents an increase of > 310% compared to the currently validated 1,261 TUs. A total of 3,010 TUs (~65%) are monocistronic, while 1,652 TUs comprise more than one ORF (polycistronic). Most TUs (~91%) were defined by an independent single modular unit. However, 398 TUs (~9%) were comprised of multiple modular units that are nested within each other, defining a convoluted modular genome structure. These nested TUs might therefore increase the flexibility of bacterial genomes without increasing genome size.

## III. Conclusion

Conceptually, the TRN consists of nodes (TUs and regulatory proteins) and links (their interaction). The meta-structure elucidation greatly advances the TRN reconstruction effort by providing a nearly complete set of nodes, that in turn, now requires condition-dependent location-analysis of sigma factors, transcription factors and other participating components for its full elucidation. Taken together, the extensive experimental results presented demonstrate how the meta-structure of the bacterial genome can be experimentally obtained. The meta-structure of *E. coli* K-12 MG1655 genome notably improves our knowledge and understanding of this widely studied genome. The process developed and implemented here can be applied to other prokaryotic organisms. The result is an experimental annotation of a genome and it provides the scaffold on which the transcriptional and translational regulatory network will be built.

## References

[1] MacLean D, Jones JD, Studholme DJ (2009) Application of 'next-generation' sequencing technologies to microbial genetics. *Nat. Rev. Microbiol.* **7**, 287-296.

Acknowledgements: This work was funded by NIH grant GM062791 and DOE DE-FC02-02ER63446.

Department of Bioengineering, University of California, San Diego. E-mail: (bcho@ucsd.edu; bpalsson@ucsd.edu)