Omics: Fulfilling the Promise

Exploring bacterial epigenomics in the next-generation sequencing era: a new approach for an emerging frontier

Poyin Chen[1,2], Richard Jeannotte[1,2,3], and Bart C. Weimer[1,2]

[1] Department of Population Health and Reproduction, School of Veterinary Medicine, University of California, Davis, CA, USA
[2] Universidad de Tarapacá, Avenida General Velásquez N°1775, Arica, Chile
[3] Facultad de Ciencias, Universidad de Tarapacá, Arica, Chile

Epigenetics has an important role for the success of foodborne pathogen persistence in diverse host niches. Substantial challenges exist in determining DNA methylation to situation-specific phenotypic traits. DNA modification, mediated by restriction-modification systems, functions as an immune response against antagonistic external DNA, and bacteriophage-acquired methyltransferases (MTase) and orphan MTases – those lacking the cognate restriction endonuclease – facilitate evolution of new phenotypes via gene expression modulation via DNA and RNA modifications, including methylation and phosphorothioation. Recent establishment of large-scale genome sequencing projects will result in a significant increase in genome availability that will lead to new demands for data analysis including new predictive bioinformatics approaches that can be verified with traditional scientific rigor. Sequencing technologies that detect modification coupled with mass spectrometry to discover new adducts is a powerful tactic to study bacterial epigenetics, which is poised to make novel and far-reaching discoveries that link biological significance and the bacterial epigenome.

Increasing bacterial genomes and the epigenome

The field of epigenetics is poised to change with the use of multi-omics approaches to examine the epigenome. The routine availability of thousands of bacterial genomes enables new approaches for finding novel genes associated with epigenomics. This review focuses on the integration of omics approaches as applied to DNA modification for high-throughput approaches. The reader will be referred to specific detailed reviews for the nuances of specific genes so that this review can highlight the integration of next-generation sequencing (NGS), metabolomics, and bioinformatics.

The bacterial epigenome is a dynamic feature that changes during growth in response to external stimuli, and thereby facilitating adjustment to varying environmental conditions that controls gene exchange, transcription, and genome stability on a broad scale. The epigenome consists of modifications to the nucleotides with small molecules, such as methylation, or to atoms between nucleotides, such as phosphorothioation (PT) [1]. These modifications are also used to change protein–DNA binding, and thereby altering the biochemical landscape of DNA that directly impacts the phenotype.

DNA modifications were initially identified during bacteriophage transmissibility studies [2]. This discovery heralded the study of the influence of DNA modifications on foreign DNA recognition and ushered in understanding that DNA modification in bacteria is important. Many techniques have been developed to quantify methylation, the number of modified nucleotides, and improve detection resolution (global, site-specific, and genome-wide). Bisulfite sequencing was the first method used for determining DNA methylation via sequencing. Mass spectrometry analysis of digested DNA remains the only approach enabling discovery new modifications [3–5]. With the advent of NGS technologies, bacterial genomes are being sequenced at an exponential pace with epigenome detection not far behind. Among NGS technologies, single molecule real time (SMRT) DNA sequencing technology allows simultaneous acquisition of both genomic and epigenomic information at the nucleotide level [6]. NGS has paved the way for numerous large-scale sequencing efforts, which will probably increase the discovery of methylation density, location, strandedness, and catalytic enzymes. With this in mind, a new approach is possible to examine the genome for genes used in methylation events that lack specific homology, but contain conserved domains to preserve the functional activity.

The 100K Genome Project (Genomics England) aims to sequence the genomes of 100 000 patients with a focus on cancer, rare diseases, and infectious diseases (http://www.genomicsengland.co.uk/). The United Kingdom Food Standards Agency will sequence 1000 *Campylobacter* isolates that will contribute to the characterization of the genomic diversity of *Campylobacter* in the UK (https://fsa-esourcing.eurodyn.com/epps/cft/prepareViewCfTWS.do?resourceId=52167). The 100K Pathogen Genome Project in the USA is a collaborative effort among the FDA, the University of California Davis, and Agilent Technologies.

Corresponding author: Weimer, B.C. (bcweimer@ucdavis.edu).

Keywords: multi-omics; epigenetics; DNA adduct; phosphorothioation; methylation.

0966-842X/

http://dx.doi.org/10.1016/j.tim.2014.03.005

The US-based 100K Pathogen Genome Project will sequence 100 000 foodborne pathogens, 1000 of which will be done using SMRT sequencing, allowing for identification of novel modifications and the extent by which these genomes are modified. These genomes will be released on the 100K Food Pathogen Bioproject web page (http://www.ngbi.nlm.nih.gov/bioproject/186441). The FDA also created the GenomeTrakr bioproject to routinely sequence foodborne pathogens across the USA, releasing the genomes as they are sequenced (http://www.ncbi.nlm.nih.gov/bioproject/183844). Between the US-based projects, >4500 genomes have been deposited at the National Center for Biotechnology Information (NCBI) in 6 months. The current scale of sequencing will produce an additional 10 000 genomes in 2014. Collectively, these projects will radically increase the number of available genomes with the long-term goal of increasing public health.

It is very likely that with this scale of genome availability new restriction modification (RM) systems and methyltransferases (MTases) will be discovered. The sheer scale of these data require new methods for epigenetic studies to mine the information for occurrence and localization, information that can be examined in isolates using traditional approaches to verify bioinformatic predictions. Importantly, these projects will probably uncover new orphan MTases, redefine MTase diversity, and discover new DNA modifications. Use of such large datasets also enables population-based comparisons of domain conservation in addition to gene sequence homology for discovery of new genes used in epigenome modification of single genes and the protein networks used to catalyze DNA modifications.

The explosive production of epigenomes in the coming years foreshadows the need for new computational tools and platforms to conduct large-scale data analysis for genome and epigenetic annotation. Prediction of interaction networks is now possible on a population scale, which will provide strength of estimation based on gene and network distribution. With the likely discovery of new genes for DNA modification and new modifications, it will become increasingly important to use population genetics in bacteria to gain insights into gene distribution, genetic diversity, and discovery of new DNA modification hypotheses. The Genomic Encyclopedia of Bacteria and Archaea (GEBA) project sequenced 3500 genomes at specific phylogenetic branches and increased the Tree of Life [7]. With 50 new genomes, $\sim$1060 new protein families were found. Jacobsen *et al.* [8] increased the number of gene families with 36 *Salmonella* genomes. Accordingly, it is very likely that the nearly 4500 genomes added by the US-based projects will contain many new gene families, which are yet to be examined for genes related to epigenetics.

Creating new high-throughput bioinformatic strategies to predict these features is needed on a wide scale. Current informatics tools are limited in their ability to conduct multi-genome comparisons. Metabolic networks can be created individually using several resources [9,10]. Currently, databases for high-throughput analysis of physical and functional protein–protein interactions include STRING, which predicts protein–protein interaction networks for a single genome (Box 1). REBASE is a tool specific to RM systems that is available and can be used for multi-genome comparisons in a table format (Box 1) [11,12]. Strategic use of these types of databases will allow scientists to assemble information from large-scale genome projects to create new network predictions with many new genomes. Advances in sequencing technology and data processing will help to elucidate DNA modifications and the possible network of proteins that catalyze the modifications. Tools for rapid assessment of biological importance of these genes for current modifications and those that are yet to be discovered remain unclear.

Box 1. Databases and tools

REBASE is an online database (http://rebase.neb.com/rebase/rebase.html) containing all known information regarding known and putative RM system-associated proteins. In addition to recognition sequences, cleavage sites, and source, this database also includes the following information: recognition sequences, cleavage sites, source, commercial availability, sequence data, crystal structure information, isoschizomers, and methylation sensitivity. All genomic sequences uploaded to GenBank are analyzed via data mining techniques followed by manual confirmation for RM systems with those found curated in REBASE. Whereas MTases can be predicted with relative ease and accuracy, REase genes are more highly divergent and as such are usually predicted by proximity to, or co-occurrence with, the MTase [12].

STRING is a public access tool (http://string.embl.de/) that draws together experimental, predicted, and transferred interactions, including interactions predicted through text mining to provide a putative protein–protein interaction map. The premise for STRING is that protein–protein interactions extend beyond physical interactions. Included in protein–protein networks are catalytic interactions in metabolic, transcriptional, and translational pathways as well as proteins that contribute to a larger unit without ever directly interacting. STRING users may input single or multiple queries and preset the target organism in which to build the network. Each node in the resulting network represents an interactor as identified by STRING prediction algorithms, chosen as a result of neighborhood, gene fusion, co-occurrence, co-expression, experimental evidence, databases, text mining, and homology. Supplementary information is provided for each node including protein sequence, structure, and domains, and its homologs. Networks may be expanded or recentered on the desired node and nodes may be categorized by biological relevance such as biological processes, molecular function, and cellular components [11].

Methylation and restriction modification systems

The most comprehensively studied DNA modification is methylation. It is extensively reviewed in numerous reviews [2,13–15]; therefore, only a brief summary of pertinent information from those reviews will be included here to set the stage for the discussion of NGS and high-throughput methods to analyze genomes for these traits. DNA methylation is well established as a part of bacterial RM systems, of which 43 650 RM enzymes are cataloged in over 3600 bacterial isolates (http://REBASE.neb.com/REBASE/REBASE.html). Only a fraction of the newly released genomes from the US-based 100K Pathogen Genome Project and the FDA GenomeTrakr are analyzed, leading to anticipation that a dramatic increase in these enzymes will occur shortly. The challenge will be to experimentally verify the predictions and understand relative biological importance.

RM systems were first described in a series of experiments exploring bacteriophage (phage) infection [2]. Phage