Methods xxx (2014) xxx–xxx

Contents lists available at ScienceDirect

Methods

journal homepage: www.elsevier.com/locate/ymeth

Navigating and mining modENCODE data

Nathan Boley [a], Kenneth H. Wan [b], Peter J. Bickel [c], Susan E. Celniker [b,*]

[a] Department of Biostatistics, University of California Berkeley, Berkeley, CA, United States
[b] Department of Genome Dynamics, Lawrence Berkeley National Laboratory, Berkeley, CA, United States
[c] Department of Statistics, University of California Berkeley, Berkeley, CA, United States

ARTICLE INFO

Article history:
Received 9 January 2014
Revised 4 March 2014
Accepted 6 March 2014
Available online xxxx

Keywords:
modENCODE
Mining
Drosophila
Transcription
Chromatin

ABSTRACT

modENCODE was a 5 year NHGRI funded project (2007–2012) to map the function of every base in the genomes of worms and flies characterizing positions of modified histones and other chromatin marks, origins of DNA replication, RNA transcripts and the transcription factor binding sites that control gene expression. Here we describe the *Drosophila* modENCODE datasets and how best to access and use them for genome wide and individual gene studies.

1. Introduction

In 1998 and 2000, the complete genome sequences for *Caenorhabditis elegans* [1] and *Drosophila melanogaster* [2–4], respectively, were published. Thus emerged the modENCODE project [5], whose goal was to map functional elements to the genomes of these model organisms. The modENCODE consortium was formed as a network of research laboratories, comprised of 11 research projects: four projects for worm, six for fly and one contributing to both organisms. The six fly projects are: domains of gene structure; transcription factor binding sites; histone modifications; chromatin structure; origins of DNA replication and timing; and RNA expression profiling. Here we describe the *Drosophila* modENCODE datasets [6], and how best to access and use them for genome wide and individual gene studies. The data can be accessed through the modENCODE website (http://www.modencode.org/). All data from this project is publicly available and was vetted by a data coordinating center (DCC) to ensure consistency and completeness [7]. All the software of the project is publicly available.

2. DNA sequencing and replication

2.1. Sequencing technologies

Four sequencing technologies – Sanger, 454, Illimunia, and SOLiD – were used in modENCODE. We describe these briefly here; for comprehensive reviews see [8–10].

Standard Sanger sequencing [11] produces long reads (up to about 900 bp), has easy reaction setup, and can assign sequence to individual clones. The major disadvantage is the low throughput compared to newer technologies. The *D. melanogaster* genome was primarily sequenced using the Sanger Sequencing technology.

Roche's 454 instruments use pyrosequencing [12], which can produce read lengths up to 1 kb, but has difficulties sequencing long homopolymer stretches and requires a time consuming sample preparation step. modENCODE used Roche's 454 instruments to produce the bulk of the 5′ RACE data (Transcriptomics – Transcription Start Sites).

The Illumina sequencing platform, the most common "second generation" sequencing technology, was used to produce most of the modENCODE DNA and RNA data. It is high throughput, costs relatively little per base, and has a relatively easy sample preparation step. However, it can only produce short reads lengths (currently up to 300 nt, but only 100 nt during the majority of modENCODE).

Applied Biosystem's (Life Technologies) SOLiD platform [13] is also high throughput and has better accuracy than Illumina [8], but requires a relatively difficult sample preparation step, under-represents AT- and GC-rich regions with substitutions as the most common error type [8]. However, the SOLiD platform was the first to produce stranded data and was used to sequence the total RNA-seq time course through embryogenesis.

2.2. Whole genome sequencing

For comparative studies, eight additional *Drosophila* species were sequenced. The read data are available in the Sequence Read

[*] Corresponding author.

http://dx.doi.org/10.1016/j.ymeth.2014.03.007
1046-2023/© 2014 Elsevier Inc. All rights reserved.

Table 1
Whole genome sequencing experiments.

Species	SRA accession	Genbank ID
D. biarmipes	SRP007984	Dbia_1.0
D. eugracilis	SRP007991	Deug_1.0
D. ficusphila	SRP008002	Dfic_1.0
D. takahashii	SRP008019	Dtak_1.0
D. elegans	SRP008020	Dele_1.0
D. rhopaloa	SRP008021	Drho_1.0
D. bipectinata	SRP008024	Dbip_1.0
D. kikkawai	SRP008029	Dkik_1.0

Archive (SRA), and the assembled genomes are available from Genbank (Table 1).

2.3. Copy number variation

Immortalized cell line genomes typically differ substantially from the reference genome. To investigate these differences, whole genome sequencing experiments were conducted in 21 different *Drosophila* cell lines. For the cell lines with two independent cultures, S2-DRSC and Cl.8, expression, gene ontology, and chromatin analyses were also conducted [14]. The data is available at data.modencode.org under the "Copy Number Variation" project category.

3. Epigenetics and transcription regulation

Chromatin is subject to a diverse array of post-translational modifications, which are known to play important roles in gene regulation. To study the set of histone modifications and elucidate their function, the modENCODE consortium conducted 535 ChIP-chip and 322 ChIP-seq experiments which localized 42 distinct histone modifications, 64 chromatin-binding factors and 61 transcription factors. The major results of this study can be found in [15]; an online browsing tool for the chromatin marks can be found at http://compbio.med.harvard.edu/flychromatin/. The data can be downloaded from http://data.modencode.org/.

3.1. Assays to identify regions bound by DNA binding proteins

Briefly, ChIP-chip and ChIP-seq assays work by first cross-linking the protein to DNA (e.g. by formaldehyde fixation). The cells are lysed, and the DNA is sheared (e.g. by sonication) resulting in DNA fragments, where some are bound to the protein of interest. An antibody is then introduced that binds to the protein of interest, the antibody bound fragments are separated, the cross-linking is reversed, and the DNA fragments are purified. This process results in DNA fragments enriched for regions that were bound to the protein of interest, in this case DNA bound to a specifically modified histone protein. ChIP-chip and ChIP-seq assays vary from this stage forward.

In ChIP-chip assays, the DNA fragments are amplified, denatured, labeled with a flourescent tag, and poured over the surface of a DNA microarray. The DNA microarray is spotted with short single stranded DNA sequences tiling the genome, so the poured fragments can hybridize with complementary sequences on the array. The array is illuminated with a fluorescent light, and the fluorescence signals are captured and associated with their underlying sequence. Identifying bound genomic regions from the fluorescence requires substantial post-experimental analysis, although software packages exist which largely automate this process [16].

In ChIP-seq assays, the DNA fragments are amplified, and then sequenced. The sequenced reads are mapped to a reference genome, and then summarized by a file containing the read coverage at each given base. Generally, regions with high read coverage can be identified as bound; however, biases in the assay make naive analysis unreliable. Rather, a negative control experiment is performed, and bound regions are identified by locations in which the read coverage in significantly higher than the negative control experiment. This process is called peak calling, and sophisticated tools have been developed which perform this analysis http://www.ebi.ac.uk/~anshul/public/softwareRepo/spp_package.tar.gz. As the cost of sequencing has dropped, ChIP-seq has largely replaced ChIP-chip. For a review of ChIP-seq and analysis challenges and tools, see Park [17] and Wilbanks et al. [18]. Aleksic et al. wrote an introduction to ChIP entitled, "modENCODE Data Gentle Introduction Guides", which is available as a public Google document (https://docs.google.com/document/d/1-BQfalYIZ58POE29dnzIw903FMhrIwiTIorGOcTCQWg/).

3.2. DNA accessibility

Genes located in tightly packaged chromatin cannot be transcribed; therefore, identifying regions with low nucleosome density provides important information about the active gene set in a particular biological sample. To identify these regions, the modENCODE consortium performed DNase I hypersensitive site Sequencing (DNase-seq) [19] in three cell lines. DNase-seq works by introducing DNase I into a population of lysed cells, and sequencing the resulting fragments. Since open chromatin regions are dis-proportionally digested by DNase I, the resulting sequenced fragments provide information about DNA accessibility. The raw and mapped reads are available from GEO (http://www.ncbi.nlm.nih.gov/geo/query/acc.cgi?acc=GSE25321) and a browser track with the average read coverage across the genome is available at (http://compbio.med.harvard.edu/flychromatin/).

Mapping biases, sequencing biases, and copy number variation means that some care must be taken in the identification of open chromatin regions. Several statistical methods have been developed [20–22], but they all work by identifying regions in which the DNase-seq read counts are significantly higher than counts from a negative control. The modENCODE consortium used standard DNA sequencing data as the negative control; positions were identified as enriched in a 300 bp scanning window when compared to the DNA sequencing data [15]. The identified regions are available in the supplement of [15]. However, because chromatin region boundaries show substantial variation between different tissues and cell lines [23], it is important to ensure that samples are properly matched when using previously identified chromatin states.

3.3. Histone modifications

Chemical modifications to the ends of histones alter chromatin structure, and this altered structure correlates with specific transcriptional processes. These modifications can occur simultaneously, and the combinatorics permit for billions of potential distinct combinations. However, by analyzing histone modification patterns, and comparing them to transcription data, patterns of combinations have emerged. To identify these patterns and elucidate their corresponding biological function, the modENCODE consortium conducted ChIP-chip and ChIP-seq experiments on 51 histone modifications in four cell lines. A select number of marks were also analyzed at several developmental timepoints. The full set of histone modification data can be found at