**SyStem cell biology**

A systems biology approach to pluripotent stem cells

Eran Meshorer

Department of Genetics; The Alexander Silberman Institute of Life Sciences; The Hebrew University of Jerusalem; Edmond J. Safra Campus; Jerusalem, Israel

**Keywords:** stem cells, epigenetics, high throughput, sequencing, epigenomics, chromatin

The past few years have brought us an unprecedented amount of data, which continues to accumulate in an exponentially increasing pace. Advances in high throughput methods, especially of deep sequencing technologies, allowed experimental biologists to move from analyzing responses in single genes or single pathways to the entire genome, be it at the genetic level (i.e., mutations, single nucleotide polymorphisms), the epigenetic level (DNA methylation analyses, DNase I hypersensitivity maps, FAIRE, nucleosome positioning, chromatin modifications, histone variants), the interactome (transcription factor binding, chromatin bound proteins, other trans-acting factors), the RNA level (transcriptome, RNome, ribosome profiling), or the three-dimensional organization level of the genome (Chromosome Conformation Capture related technologies), and its long-range interactions (Fig. 1).

Embryonic stem cells (ESCs) serve as an excellent model system for epigenetic studies as they undergo substantial morphological, structural and functional changes during the early stages of differentiation. Thus, the same genome gives rise to different dynamic outputs in a relatively short timescale, allowing the study of the changes manifested by the epigenome. Indeed, DNA methylation and 5-hydroxy-methylation maps, chromatin state maps, DNase-I hypersensitivity maps, formaldehyde assisted isolation of regulatory elements (FAIRE), nucleosome positioning, transcription factor binding, transcriptome analyses of polyadenylated and non-polyadenylated RNA, and 3-dimensional organization just to name a few (!), have been generated extensively for ESCs, and these maps can serve not only as validation platforms (e.g., to find whether Pol-II is enriched on the promoter of your favorite/newly discovered gene), but also as discovery platforms. An elegant example for such an approach led to the discovery of long intergenic non-coding (linc) RNAs. The authors used the chromatin signature (enrichment for H3K4me3 and H3K36me3) of typical spliced genes in the genome, and looked for similar patterns in un-annotated regions. Remarkably, they identified a large family of spliced and processed RNAs, most of which was previously unnoticed and unexplored, which lack protein coding potential. In follow up studies by the same group and now many others, the functional roles of numerous lincRNAs in ESC pluripotency and differentiation as well as the integrity of chromatin structure was demonstrated.

In a somewhat similar approach, we sought to discover potential novel regulators of histone genes. To this end, we created a database of the currently existing genome-wide maps of transcription factor binding and histone modifications in mouse ESCs, and looked for factors that are enriched at histone gene clusters. We were able to verify that E2f proteins regulate histone genes, although surprisingly, in contrast to the current view, which postulates that some of the histone genes are E2f-independent, we found that essentially all histone genes are E2f-dependent, based on enrichment scores and microarray analysis. In addition, we were able to identify novel positive (e.g., Smad proteins) and negative (e.g., Zfx, Ccf) regulators of histone gene expression, all of which were validated by previously published gene expression studies of relevant knockout cells.

Such approaches demonstrate that we have reached the stage where sufficient genome-wide data has been generated, at least for ESCs, allowing for in silico experiments, which may lead to novel insights. Cross reference of different databases from multiple sources is self-correcting and unbiased, especially when performed by third party laboratories. Such novel insights obtained by computational analyses still require experimental validation, but provide extremely useful “short-cuts”. For example, the ENCODE project included one human ESC line. For stem cell scientists this is not enough, but comparing the different data sets between the human ESC line and other cell lines used in the project may provide some novel insights, and importantly, help researchers plan their next experiments in a more focused manner. Not all predictions may prove correct, but they can no doubt save considerable time and effort. Especially pertinent are approaches that focus on specific categories within the genome, such as the example depicted above for histone gene regulation.

So should we now cease our ‘wet’ experiments and focus on analyzing data? Definitely not. Technologies are advancing at such a rapid pace that each data set that is currently being generated is slightly superior to previous data sets. Moreover, the field is driven forward by users of high throughput technologies, and if experimentalists turn analysts, technologies will soon dry up. But most importantly, we are still tremendously

Correspondence to: Eran Meshorer; Email: meshorer@huji.ac.il
Submitted: 03/14/2013; Accepted: 03/18/2013
http://dx.doi.org/10.4161/auto.9.6.24359

http://www.landesbioscience.com
far from reaching the point where new data generated is mostly, or at least partly, redundant. To advance forward efficiently, we must use the existing resources and data sets, milk them for insights, and plan the following experiments accordingly. We must therefore continue to generate new and better data, and strive to combine these data sets into databases of similar cell types or organisms.

Many examples now demonstrate the power of using existing data sets and combining them with newly generated ones. Platforms that combine all these data sets, allow the addition of data, and provide tools for analyses, are essential for the wider community. Two important ‘veteran’ tools, which are still the preferred choices by many biologists and analysts are Galaxy\(^\text{29}\) and GenePattern,\(^\text{30}\) which were developed by teams from Penn State University and The Broad Institute, respectively. A recent platform with excellent visualization tools and friendly interface, which integrates most of the published data, is ‘GeneProf’\(^\text{31}\), making it essentially accessible to anyone. A more recent interesting attempt was made to create a platform—Spark—for biologists faced with large-scale computational challenges.\(^\text{32}\) Spark was able to identify known epigenetic signatures, but importantly, was also able to predict association between YY1 and CTBP2 in human ESCs, based on their genome-wide binding maps. Such tools make large-scale data sets accessible for the wide scientific community.

**Figure 1.** Genomic and epigenomic layers (left) and the corresponding high-throughput sequencing-based approaches which are used to study them (right). (A) High throughput DNA sequencing (DNA-seq) is used to detect mutations in the primary DNA sequence as well as variations among individuals termed single nucleotide polymorphisms (SNPs). (B) Bisulfite-sequencing and Reduced Representation Bisulfite Sequencing (RRBS) rely on the bisulfite conversion of C to U and comparison of the non-treated to the treated sequence. These methods are used to map DNA methylation patterns. Additional methods (e.g., mDIP) employ antibodies that recognize methylated DNA (or 5-hydroxymethylated DNA), followed by high throughput sequencing. Histone modifications can also be mapped with specific antibodies using Chromatin Immunoprecipitation (ChIP) followed by high throughput sequencing (ChIP-seq). DNase I hypersensitive maps, which are mostly found around regulatory regions are mapped using DNase I digestion of chromatin followed by deep sequencing (DHS-seq). FAIRE (Formaldehyde Assisted Isolation of Regulatory Elements) is a complementary approach enabling the detection of open chromatin regions that are suspected to contain regulatory elements. Finally, to identify the locations of nucleosomes at a genome-wide scale, chromatin can be digested to mono-nucleosomes using micrococcal nuclease (MNase) and sequenced (MNase-seq). (C) ChIP-seq is used to identify the genomic regions bound by transcription factors. In order to identify which factors are bound to a specific genomic region, proteomics of isolated chromatin segments (PICh) can be used, although it has so far been demonstrated successfully only for repetitive regions and requires ample amount of cells. (D) RNA sequencing (RNA-seq) allows studying the transcriptome of cells. While usually RNA-seq employs poly-adenylated RNA, experimental methods are available to enrich for other RNA species such as non-polyadenylated RNA and microRNA. For ribosomes profiling, ribosomes are first purified biochemically, and RNA is then subjected to high throughput sequencing. (E) To study the three-dimensional organization of the genome, distinct genomic regions that are in close proximity can be amplified together following fixation using specific primers, in a method called Chromosome Conformation Capture (3C). When one primer is used as ‘bait’ amplifying adjacent regions, and tiling arrays or high throughput sequencing are applied to detect all amplification products simultaneously, the method is referred to as Circular Chromosome Conformation Capture, or 4C. When two multiplex primers are used to amplify many multiplex PCR reactions the assay is called Carbon Copy Chromosome Conformation Capture, or 5C. When ChIP is performed to first pull-down specific genomic regions bound by a given protein (such as RNA polymerase II), the assay is referred to as ChIP-loop, or ChIA-PET, when using paired end-tag sequencing. Hi-C enables amplifying all genomic regions in close proximity without the requirement of specific baits.\(^\text{1}\)
community, allowing experimental labs with little experimental knowhow to enter the genomic era.

A plethora of new knowledge is concealed beneath many layers of the already existing genome-wide data sets. The years to come will no doubt continue to provide us with increasing amounts of data. In the near future genome-wide data sets will not belong solely to the systems biologist’s realm, but will be an integral part of every experimental lab’s research. Therefore, the development of computational tools and platforms that will allow experimentalists to not only analyze but also integrate such data sets will be essential to fuel and expand the genomic revolution we are currently witnessing.

References


6. Ficz G, Branco MR, Seisenberger S, Santos M. Multilayered chromatin analysis reveals E2f, Smad and Zfx as transcriptional regulators of histones. Nat 2010; 461:46-51; PMID:20452322; http://dx.doi.org/10.1038/nm0601


Disclosure of Potential Conflicts of Interest

No potential conflicts of interest were disclosed.

Acknowledgments

The author would like to thank Yair Aarsonon, Ilana Livyatan and Matan Sorek for helpful comments and suggestions, as well as the Abisch Frenkel Foundation, the Israel Cancer Research Foundation (ICRF), the Human Frontiers Science Foundation (HFSP), the Israel Science Foundation (ISF 657/12; 1252/12), the Israel Ministry of Science, and the European Research Council (ERC-281781) for financial support.