Systems Biomedicine 1:1, 1–3; January/February/March 2013; © 2013 Landes Bioscience

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.[2,3] 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[4,5] and 5-hydroxy-methylation maps,[6-9] chromatin state maps,[10,11] DNase-I hypersensitivity maps,[12,13] formaldehyde assisted isolation of regulatory elements (FAIRE),[14] nucleosome positioning,[15] transcription factor binding,[16,17] transcriptome analyses of polyadenylated[18,19] and non-polyadenylated[20] RNA, and 3-dimentional organization[21,22] 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.[23] 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[24,25] as well as the integrity of chromatin structure[26] 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,[27] 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, Ctcf) regulators of histone gene expression, all of which were validated by previously published gene expression studies of relevant knockout cells.[28]

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 throughout 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

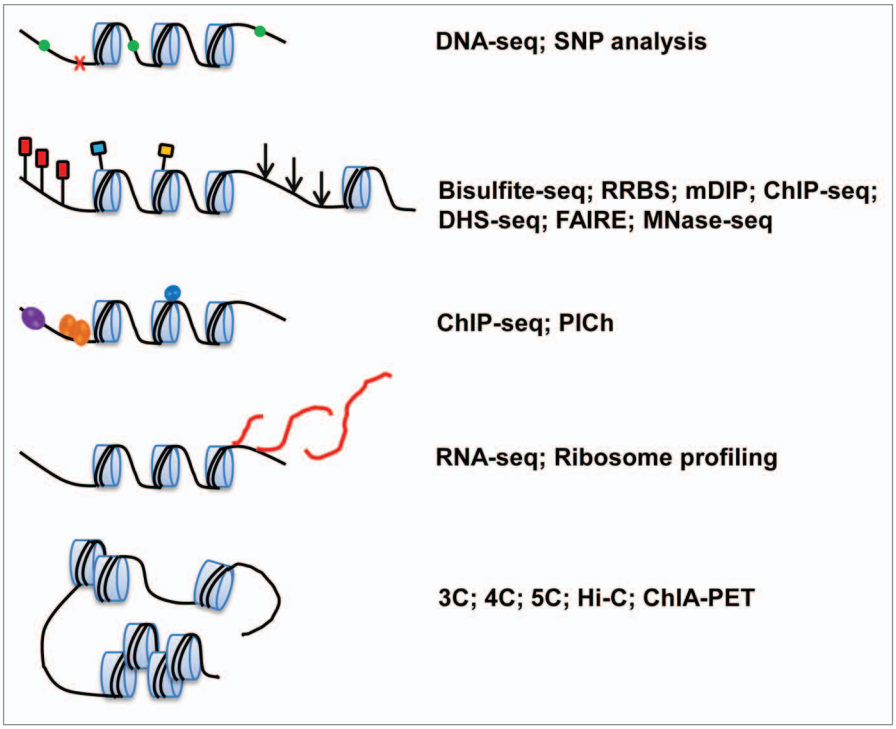

Figure 1. Genomic and epigenomic layers (left) and the corresponding high-throughput sequencing-based approaches which are used to study them (right). (**A**) High throughput DNA sequencing (DNA-seq) is used to detect mutations in the primary DNA sequence as well as variations among individuals termed 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 throughout 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.[1]

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[29] and GenePattern,[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'[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.[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, allowing experimental labs with little computational 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.

Disclosure of Potential Conflicts of Interest

No potential conflicts of interest were disclosed.

Acknowledgments

The author would like to thank Yair Aaronson, 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.

References

1. Hakim O, Misteli T. SnapShot: Chromosome confirmation capture. Cell 2012; 148:1068.e1-2; PMID:22385969; http://dx.doi.org/10.1016/j.cell.2012.02.019
2. Gaspar-Maia A, Alajem A, Meshorer E, Ramalho-Santos M. Open chromatin in pluripotency and reprogramming. Nat Rev Mol Cell Biol 2011; 12:36-47; PMID:21179060; http://dx.doi.org/10.1038/nrm3036
3. Mattout A, Meshorer E. Chromatin plasticity and genome organization in pluripotent embryonic stem cells. Curr Opin Cell Biol 2010; 22:334-41; PMID:20226651; http://dx.doi.org/10.1016/j.ceb.2010.02.001
4. Meissner A, Mikkelsen TS, Gu H, Wernig M, Hanna J, Sivachenko A, et al. Genome-scale DNA methylation maps of pluripotent and differentiated cells. Nature 2008; 454:766-70; PMID:18600261
5. Straussman R, Nejman D, Roberts D, Steinfeld I, Blum B, Benvenisty N, et al. Developmental programming of CpG island methylation profiles in the human genome. Nat Struct Mol Biol 2009; 16:564-71; PMID:19377480; http://dx.doi.org/10.1038/nsmb.1594
6. Ficz G, Branco MR, Seisenberger S, Santos F, Krueger F, Hore TA, et al. Dynamic regulation of 5-hydroxymethylcytosine in mouse ES cells and during differentiation. Nature 2011; 473:398-402; PMID:21460836; http://dx.doi.org/10.1038/nature10008
7. Williams K, Christensen J, Pedersen MT, Johansen JV, Cloos PA, Rappsilber J, et al. TET1 and hydroxymethylcytosine in transcription and DNA methylation fidelity. Nature 2011; 473:343-8; PMID:21490601; http://dx.doi.org/10.1038/nature10066
8. Wu H, D'Alessio AC, Ito S, Wang Z, Cui K, Zhao K, et al. Genome-wide analysis of 5-hydroxymethylcytosine distribution reveals its dual function in transcriptional regulation in mouse embryonic stem cells. Genes Dev 2011; 25:679-84; PMID:21460036; http://dx.doi.org/10.1101/gad.2036011
9. Xu Y, Wu F, Tan L, Kong L, Xiong L, Deng J, et al. Genome-wide regulation of 5hmC, 5mC, and gene expression by Tet1 hydroxylase in mouse embryonic stem cells. Mol Cell 2011; 42:451-64; PMID:21514197; http://dx.doi.org/10.1016/j.molcel.2011.04.005
10. Hawkins RD, Hon GC, Lee LK, Ngo Q, Lister R, Pelizzola M, et al. Distinct epigenomic landscapes of pluripotent and lineage-committed human cells. Cell Stem Cell 2010; 6:479-91; PMID:20452322; http://dx.doi.org/10.1016/j.stem.2010.03.018
11. Mikkelsen TS, Ku M, Jaffe DB, Issac B, Lieberman E, Giannoukos G, et al. Genome-wide maps of chromatin state in pluripotent and lineage-committed cells. Nature 2007; 448:553-60; PMID:17603471; http://dx.doi.org/10.1038/nature06008
12. Schnetz MP, Handoko L, Akhtar-Zaidi B, Bartels CF, Pereira CF, Fisher AG, et al. CHD7 targets active gene enhancer elements to modulate ES cell-specific gene expression. PLoS Genet 2010; 6:e1001023; PMID:20657823; http://dx.doi.org/10.1371/journal.pgen.1001023
13. Thurman RE, Rynes E, Humbert R, Vierstra J, Maurano MT, Haugen E, et al. The accessible chromatin landscape of the human genome. Nature 2012; 489:75-82; PMID:22955617; http://dx.doi.org/10.1038/nature11232
14. Song L, Zhang Z, Grasfeder LL, Boyle AP, Giresi PG, Lee BK, et al. Open chromatin defined by DNaseI and FAIRE identifies regulatory elements that shape cell-type identity. Genome Res 2011; 21:1757-67; PMID:21750106; http://dx.doi.org/10.1101/gr.121541.111
15. Teif VB, Vainshtein Y, Caudron-Herger M, Mallm JP, Marth C, Höfer T, et al. Genome-wide nucleosome positioning during embryonic stem cell development. Nat Struct Mol Biol 2012; 19:1185-92; PMID:23085715; http://dx.doi.org/10.1038/nsmb.2419
16. Chen X, Xu H, Yuan P, Fang F, Huss M, Vega VB, et al. Integration of external signaling pathways with the core transcriptional network in embryonic stem cells. Cell 2008; 133:1106-17; PMID:18555785; http://dx.doi.org/10.1016/j.cell.2008.04.043
17. Marson A, Levine SS, Cole MF, Frampton GM, Brambrink T, Johnstone S, et al. Connecting microRNA genes to the core transcriptional regulatory circuitry of embryonic stem cells. Cell 2008; 134:521-33; PMID:18692474; http://dx.doi.org/10.1016/j.cell.2008.07.020
18. Cloonan N, Forrest AR, Kolle G, Gardiner BB, Faulkner GJ, Brown MK, et al. Stem cell transcriptome profiling via massive-scale mRNA sequencing. Nat Methods 2008; 5:613-9; PMID:18516046; http://dx.doi.org/10.1038/nmeth.1223
19. Efroni S, Duttagupta R, Cheng J, Dehghani H, Hoeppner DJ, Dash C, et al. Global transcription in pluripotent embryonic stem cells. Cell Stem Cell 2008; 2:437-47; PMID:18462694; http://dx.doi.org/10.1016/j.stem.2008.03.021
20. Livyatan I, Harikumar A, Nissim-Rafinia M, Duttagupta R, Gingeras TR, Meshorer E. Non-polyadenylated transcription in embryonic stem cells reveals novel non-coding RNA related to pluripotency and differentiation. Nucleic Acids Res 2013; 41:6300-15; PMID:23630323; http://dx.doi.org/10.1093/nar/gkt316
21. Dixon JR, Selvaraj S, Yue F, Kim A, Li Y, Shen Y, et al. Topological domains in mammalian genomes identified by analysis of chromatin interactions. Nature 2012; 485:376-80; PMID:22495300; http://dx.doi.org/10.1038/nature11082
22. Ryba T, Hiratani I, Lu J, Itoh M, Kulik M, Zhang J, et al. Evolutionarily conserved replication timing profiles predict long-range chromatin interactions and distinguish closely related cell types. Genome Res 2010; 20:761-70; PMID:20430782; http://dx.doi.org/10.1101/gr.099655.109
23. Guttman M, Amit I, Garber M, French C, Lin MF, Feldser D, et al. Chromatin signature reveals over a thousand highly conserved large non-coding RNAs in mammals. Nature 2009; 458:223-7; PMID:19182780; http://dx.doi.org/10.1038/nature07672
24. Guttman M, Donaghey J, Carey BW, Garber M, Grenier JK, Munson G, et al. lincRNAs act in the circuitry controlling pluripotency and differentiation. Nature 2011; 477:295-300; PMID:21874018; http://dx.doi.org/10.1038/nature10398
25. Loewer S, Cabili MN, Guttman M, Loh YH, Thomas K, Park IH, et al. Large intergenic non-coding RNA-RoR modulates reprogramming of human induced pluripotent stem cells. Nat Genet 2010; 42:1113-7; PMID:21057500; http://dx.doi.org/10.1038/ng.710
26. Tsai MC, Manor O, Wan Y, Mosammaparast N, Wang JK, Lan F, et al. Long noncoding RNA as modular scaffold of histone modification complexes. Science 2010; 329:689-93; PMID:20616235; http://dx.doi.org/10.1126/science.1192002
27. van Wijnen AJ, van Gurp MF, de Ridder MC, Tufarelli C, Last TJ, Birnbaum M, et al. CDP/cut is the DNA-binding subunit of histone gene transcription factor HiNF-D: a mechanism for gene regulation at the G1/S phase cell cycle transition point independent of transcription factor E2F. Proc Natl Acad Sci U S A 1996; 93:11516-21; PMID:8876167; http://dx.doi.org/10.1073/pnas.93.21.11516
28. Gokhman D, Livyatan I, Sailaja BS, Melcer S, Meshorer E. Multilayered chromatin analysis reveals E2f, Smad and Zfx as transcriptional regulators of histones. Nat Struct Mol Biol 2013; 20:119-26; PMID:23222641; http://dx.doi.org/10.1038/nsmb.2448
29. Giardine B, Riemer C, Hardison RC, Burhans R, Elnitski L, Shah P, et al. Galaxy: a platform for interactive large-scale genome analysis. Genome Res 2005; 15:1451-5; PMID:16169926; http://dx.doi.org/10.1101/gr.4086505
30. Reich M, Liefeld T, Gould J, Lerner J, Tamayo P, Mesirov JP. GenePattern 2.0. Nat Genet 2006; 38:500-1; PMID:16642009; http://dx.doi.org/10.1038/ng0506-500
31. Halbritter F, Vaidya HJ, Tomlinson SR. GeneProf: analysis of high-throughput sequencing experiments. Nat Methods 2012; 9:7-8; PMID:22205509; http://dx.doi.org/10.1038/nmeth.1809
32. Nielsen CB, Younesy H, O'Geen H, Xu X, Jackson AR, Milosavljevic A, et al. Spark: a navigational paradigm for genomic data exploration. Genome Res 2012; 22:2262-9; PMID:22960372; http://dx.doi.org/10.1101/gr.140665.112