OPINION

Building stem-cell genomics in California and beyond

Natalie D DeWitt, Michael P Yaffe & Alan Trounson

By devoting funding to whole-genome studies, such as epigenetic and copy-number variation in stem cells, research on new genomic technology, and standards for methodologies and data collection/sharing, CIRM can spur both basic and translational research.

The rapid pace of progress in next-generation DNA sequencing and genomics technologies is increasing the feasibility of a wide spectrum of systematic biological studies and bringing us ever closer to the '$1,000 genome' (**Fig. 1**). This trend will soon make genome-scale characterization a practical tool for the analysis of stem cells, not only permitting DNA sequencing, definition of the transcriptome, and elucidation of epigenetic modifications of all stem-cell-derived products and cells used in 'disease in a dish' studies, but also informing fundamental studies of stem-cell biology and potential therapeutic applications. And yet, compared with more established related genomics fields, such as cancer genomics, stem-cell genomics remains in its infancy. There is thus an urgent need in the next few years to ramp up efforts to establish stem cells as a leading model system for understanding human biology and disease states and ultimately to accelerate progress toward clinical translation.

Why should the stem-cell field engage in genomics now? One reason is that both the basic-research and medical communities are moving toward developing integrated platforms where molecular and genomics analyses of patients play a central role in informing diagnostics and therapeutics. Support is also building at a national level; for example, a US National Academy of Sciences (NAS) task force charged with creating a new molecular taxonomy for disease recently issued a report recommending that "the nation needs a live network of information on a person's molecular tests and health care"[1]. The NAS committee proposed that scientists and clinicians engaged in biomedical research, clinical medicine and basic research create a central data base of information, including molecular, environmental, family history and geographical data on individuals in the medical system. Genomic and epigenomic data would be important platforms in such a model. It is our opinion that, to realize its full potential, human stem-cell science should contribute to these activities and intercalate with this proposed knowledge infrastructure by contributing, at a minimum, molecular characterization of induced pluripotent stem cell (iPSC) disease models that can be integrated with longitudinal medical information about the individual who donated the cells. Playing a proactive role in building this knowledge infrastructure will ensure that the needs of all stakeholders in cell therapies are met in the coming decade as medicine transforms to an information-intensive model.

For California to take a firm and lasting grip on leadership in stem-cell research—and, as stated in Proposition 71, "advance the biotech industry in California to world leadership as an economic engine for California's future"—its scientists must have access to these technologies and moreover create a coordinated international enterprise to maximize the reach and impact of stem cell genomics. Genomics is creating a sea change in biomedical research and medicine, and accordingly, the California Institute for Regenerative Medicine (CIRM; San Francisco) can create a process through which stem-cell research can participate and even provide leadership in a new era of medicine. This research will play a key role in providing cellular assays that can drive the development of new compounds and highly refined biomarkers for drug discovery, diagnostics and ultimately cellular therapeutics in the context of personalized medicine. California is well placed to lead this effort, with its superb and well-funded stem-cell science, powerful biocomputing capacities, numerous leading genomics companies and experience in facilitating multinational projects with international funding partners. With judicious expenditure of CIRM funds, it should be possible to use existing resources to rapidly and efficiently build an effective stem-cell genomics infrastructure that will be unique in the world, thus positioning California as a leader in this critical area of basic and translational research while genomic technologies build steam in the next five years.

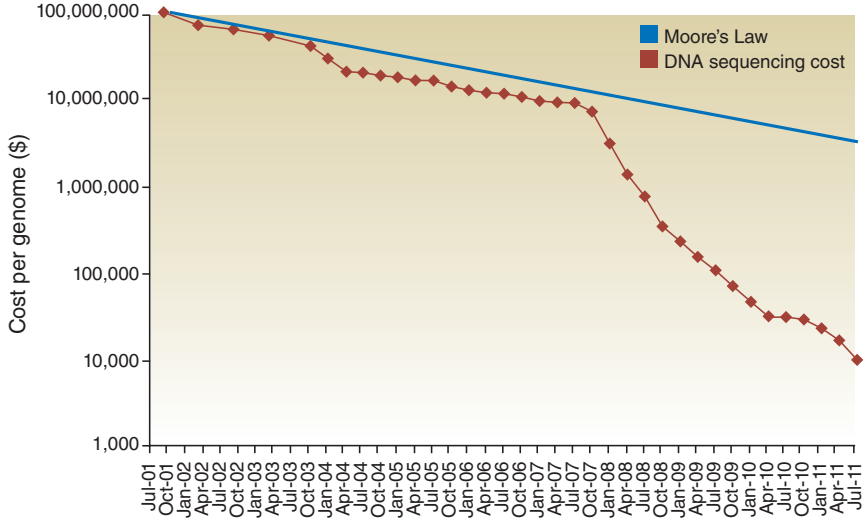

Figure 1 Technology improvements in DNA-sequencing technology are outpacing Moore's Law (the computing industry's trend of doubling computer power every two years; adapted from http://www.genome.gov/sequencingcosts/; accessed 12/06/11).

Natalie D. DeWitt, Michael P. Yaffe & Alan Trounson are at the California Institute for Regenerative Medicine, San Francisco, California, USA.
e-mail: ndewitt@cirm.ca.gov

Why stem-cell science needs genomics

Building a capacity for stem-cell genomics will advance a number of crucial areas of biology and translational research, both in the near and in the long term (**Box 1**). A key question in biology and medicine is how genotype relates to phenotype at the levels of both cells and the individual. Analyses of human iPSC (hiPSC)-based disease models promise to contribute to phenotype characterization. However, substantial structural and sequence variation has been detected in the genomes of wild-type hiPSC and human ESC (hESC) cells under some culture conditions[2–4]. It is not yet clear whether this variation was present in the original cell (for iPSCs), whether it results from the process of deriving the cells or the culture conditions, or even how much of it is experimental noise. It is thus critical that we gain an understanding of what causes this variation, discover to what degree it is present in all cells versus specifically in stem cells under culture and reprogramming conditions, and determine whether the variation leads to changes in cellular behavior before we can understand disease-in-a-dish phenotypes and gain confidence in therapeutic cells. Furthermore, it is important to note that hESCs and iPSCs offer a controlled cellular system for the mapping of molecular changes during development and differentiation, thus providing a superb opportunity for stem cells to contribute to a fundamental understanding of basic biology, as long as intermediate and mature differentiation states can be verified by molecular analyses of endogenous human tissues. Moreover, cellular signaling and gene expression pathways implicated in cancer are increasingly found to be active in ESCs. Next-generation sequencing technology will be important for moving these critical studies forward.

Because stem cells exist in heterogeneous populations, the development of single-cell genomics technology will be key to understanding their regulation and characteristics. To date, most genomics, epigenomics and transcriptome analyses are performed on pooled cells. Therefore, the molecular information is averaged over the entire population, and this step obfuscates critical information about individual cells. An accurate understanding of how the molecular state of a cell drives its behavior is impossible to gain with data averaged from pooled cells[5,6]. Looking forward to applying genomics to the single-cell level, new technologies will need to be created or existing ones adapted to stem-cell applications. Stem cells can be separated into different populations with distinct characteristics. Individual cells can differ in their DNA sequences, mRNA and protein composition, and metabolic and signaling activity. In the body, stem cells exist in niches or circulate accompanied by a variety of other non-stem-cell types. Cultures of cells also comprise a mixture of cells in various degrees of differentiation, or pluripotency states. Isolating and analyzing single cells and comparing their variability at the genomics level will be essential for drilling down to answer questions about how genotype relates to protein function and cellular phenotype[7]. Understanding single-cell behavior and molecular variation observed from cell to cell could build upon pioneering studies of biological noise in prokaryotes[8].

Such cancers as leukemia and breast cancer, which include certain forms that are rooted in aberrant stem cells[9], show the importance of characterizing an individual, disease-causing cancer stem cell. These cancer stem cells can drive tumor progression but are resistant to radiation and chemotherapy[10,11]. To direct therapies specifically to these cells, it will be important to understand how they differ in terms of their DNA or RNA composition from the bulk cells of the tumor. These differences could reveal aberrant molecular pathways to target therapeutically and eventually could be used as biomarkers for drug development and for stratifying patients into groups for individualized therapies. Current methods to target therapies to tumors based on the response of pooled cells, even if enriched for cancer stem cells by surface-marker-based cell sorting, are likely averaging and masking important lesions responsible for tumor progression and drug resistance.

In addition, technologies are needed to detect rare genomic or transcriptional events occurring in only a single or a few cells within a population of cells. The latter application could help ensure that the genomes of therapeutic cells are devoid of rare cells in the population that carry oncogenic alterations, immunogenic human leukocyte antigen (HLA) sequences or known disease-causing genomic variants. For detecting rare oncogenic events in cells destined for patients in clinics, it will be impossible to test batches of cells at the point of delivery, as this would require destruction of the cells themselves. Even so, determining the frequency of these events (if they occur at all) under optimal culture conditions and identifying their causes will not only make it easier to predict whether they will be a concern in clinical applications of cell therapy but also point to standards for cell passage, handling and purification to avoid such events completely.

The stem-cell genomics pipeline

Stem-cell genomics can be viewed as a pipeline comprising a stepwise series of

Box 1 Projected impact of genomics on stem cell biology and therapeutics

CIRM's cross-disciplinary genomics initiative would bring experts in computational biology and genomics together with stem-cell scientists engaged in basic research, cell banking, cell manufacturing and clinical applications. It would have major impacts on the following three areas:

Basic science

- Understand the presence and importance of variation in genomes, epigenomes and transcriptomes of individual cells within an organ or tissue
- Build an atlas of molecular changes (epigenetic and transcriptional) as cells undergo controlled differentiation
- Establish minimally mutated hiPSC and hESC lines to introduce suspected disease-causing mutations for further analysis
- Understand culture-induced genomic instability and its effects on cell behavior in hiPSC-based models

Biomedical research and biomarker development

- Detect sequence variants in patients and correlate with disease-in-a-dish phenotypes and disease onset or severity in patients who provided donor cells
- Develop drugs in iPSC disease models, correlating genomic data to drug responses in cellular models
- Use basic biological studies of epigenomic control of cell state to seed pharmaceutical development and permit the identification of compounds that modulate cell state in cell manufacturing or in clinical applications

Clinical science and cell manufacture

- Test therapeutic cell lines for genomic integrity, immunogenic and oncogenic potential, as well as the presence of disease-associated gene variants
- Stratify patients for personalizing treatments of cancers as stem-cell diseases

technologies, ranging from wet-bench sample preparation to high-throughput massively parallel next-generation sequencing technologies to sophisticated biocomputing capabilities, followed by confirmation and functional testing of hypotheses (**Fig. 2**). In the stem-cell field, projects could originate from a variety of sources, including academic and clinical researchers, biotech companies, and cell manufacturing facilities, with the last likely to be early adopters of genomic analysis tools for characterizing therapeutic cell lines at early stages of expansion (although likely not at point of delivery).

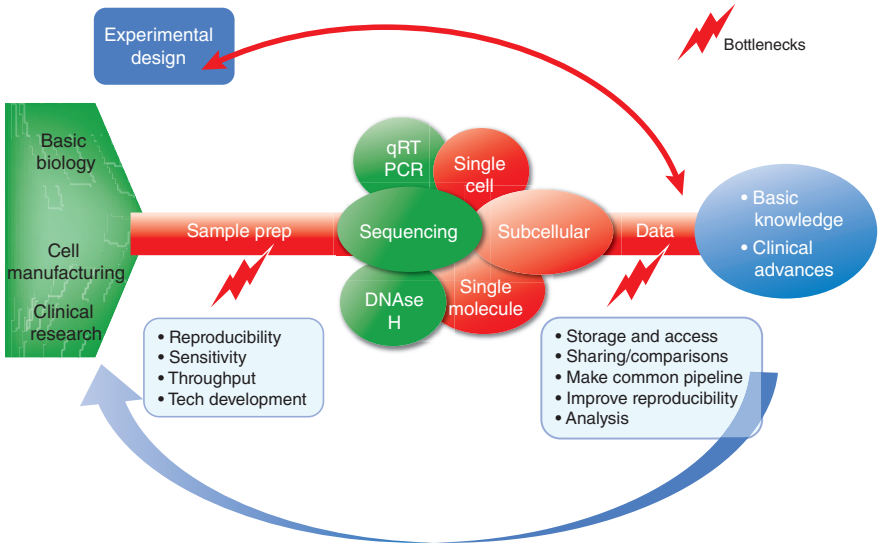

Figure 2 A flow chart depicting major stages of the proposed stem-cell genomics pipeline. From left to right: conception of genomics project from clinical, manufacturing or basic research scientists; sample preparation; next-generation sequencing or deep-sequencing technologies; single-cell analytics (where applicable); data handling and analysis, and transfer of information back to scientists and clinicians. Bottlenecks that are potentially opened by CIRM funding of innovation or provision of resources are indicated by the lightning bolts.

Experimental design. At the stage of experimental design, stem-cell scientists embarking on a costly and long-term sequencing project would benefit from assistance by genomic and computational scientists with specialized genomics expertise. This latter group would be able to help design appropriately controlled and statistically powerful experiments and provide standard operating procedures for sample preparations to minimize experimental noise and allow comparison of data sets from different laboratories and clinics. Companies and academic centers providing sequencing and analytics often perform such a function. For instance, to reduce the risk of experimental noise or failure, facilities carrying out such analyses often do in-house wet-bench manipulations such as library preparation; many companies provide experimental design consultation as a service. Even so, for large-scale projects such as studying genomic variation under culture conditions or epigenetic modification during differentiation, it will be essential to adopt one set of protocols so that data can be compared across labs participating in such projects.

Sample preparation. The time, cost and reproducibility of sample preparation are considered limiting in existing genomics pipelines. Genomic technologies have in the past decade become rapid, high-throughput and massively parallel. With an output of, for example, 3 billion reads in two days, sequencers such as the latest Illumina (San Diego) HiSeq 2000 technology are rapidly exceeding the capacity of scientists to prepare enough samples to maximize each sequencing run.

Sequencing and analytic technologies. At present, obtaining data can be simply a matter of shipping cells or a genomic or cDNA library to a commercial or academic sequence provider, receiving the data by File Transfer Protocol (FTP) or express mail on a data-storage device, and paying the provider for services rendered. To date, such samples as genomic DNA, transcripts or non-coding RNAs are being analyzed on a wide variety of platform technologies, each with its own caveats and benefits, in a range of core facilities, companies and dedicated sequencing centers. These technologies are still quite expensive for the average academic laboratory. Today, the price for a whole genome sequence is ~$5,000. Moreover, to maximize the value of genomic analysis, the comparison of many genomes from different cell lines and those grown in different conditions will be necessary, as will multiple sequencing runs of each line to eliminate noise and error. Fortunately, given the availability of sequence providers, the analytic phase of the pipeline is not considered the most limiting, and the need for new facilities dedicated to DNA sequencing does not appear to be a funding priority.

We anticipate that CIRM funding for a limited number of outstanding projects upon which stem-cell genomics can build as a community, such as epigenomic changes during differentiation, together with support to enhance access of stem-cell scientists to existing technology providers, would be the most flexible way to advance the sequencing and analytics steps in the pipeline.

Data collection, storage, analysis and accession. Integrating stem-cell biologists with data collection and analysis centers will be necessary to permit systematic, data-intensive studies, such as whole-genome epigenetic analysis during differentiation or correlating genomic structural variation with phenotypic data.

This brings us to a bottleneck where the massive amounts of data generated by genomics studies meet the very limited human and infrastructural resources for collection, transfer, storage, analysis and dissemination of these data. Whereas DNA-sequencing facilities are routinely producing sequence runs on demand, few labs in the stem-cell community can access the most advanced informatics technologies necessary to benefit from these rapidly evolving technologies. The problem of assembling, storing and transferring data on tens or hundreds of whole genomes in a secure environment that protects the privacy of experimental subjects while also permitting community access to 'de-identified' medical records and, ideally, longitudinal follow-up studies on medical conditions is considerable. Superimposed is the challenge of standardizing sample preparation and bioinformatics analysis to ensure that where variation is detected, it is biologically meaningful and not experimental noise.

In response to this obstacle, the National Science Foundation recently granted $1.4 million to the University of California at San Diego's Supercomputer Center and the California Institute for Telecommunications and Information Technology (Irvine) to address access to large-scale sequencing data by the scientific community, configuration of

those analyses, and design of efficient workflows. Other funding agencies in various fields of study are similarly concerned. Thus for CIRM to empower the stem-cell community to access genomics, we should also provide a capable data platform tailored to the particular needs of the community.

Well-developed and integrated approaches to handling and analysis of large data sets will be needed for large-scale studies correlating genomic variation with phenotype and disease, as well as tracking the epigenetic changes associated with differentiation. These studies are likely to require on the order of hundreds of whole-genome sequences, exome analyses or equivalent amounts of transcriptome data. Cancer genomics studies are leading the way in setting up such an infrastructure, and this could readily be exploited by the stem-cell community in various ways, both by learning from these models and, where possible, by taking advantage of these existing tools and infrastructure.

Beyond the genomics data

The above data should serve many purposes for the stem-cell community, ranging from generating testable hypotheses for basic biology and disease-in-a-dish studies to providing information on genomic integrity for therapeutic cell lines to informing clinical studies and cell therapies (**Fig. 3**).

For basic research and iPSC-based disease models, using genomics data to construct functionally testable hypotheses is important to ensure that stem-cell genomics remains a robust experimental science as opposed to a theoretical one. In the near future, this would involve mutating genes suspected of playing a role in disease and studying their consequences in disease-in-a-dish models. Looking ahead, synthetic biology promises to provide a means of testing some hypotheses generated by genomics. For instance, synthetic gene networks now can be embedded into cells to observe their impact on cell regulation. Stem-cell genomics will undoubtedly produce hypotheses about how genes and gene networks produce cellular phenotypes and how gene sequences lead to molecular functions of the encoded proteins. With the tools and approaches of synthetic and systems biologists, these hypotheses can be tested by genetically manipulating cellular regulatory circuitry.

A stem-cell genomics platform focusing on variation in iPSCs would dovetail with efforts to bank iPSCs from both healthy individuals and those known to have diseases, to detect and understand disease-causing mutations and epigenomic modifications. With proper consent and data management with privacy controls in place, this information can link back to de-identified patient records to correlate the data with onset and severity of disease and response to therapies. Integrating this information into a 'knowledge network,' such as that proposed by the NAS[1], would provide a powerful platform for incorporating stem-cell research into a comprehensive biomedical information network. For instance, molecular characterization of the genome, epigenome and transcriptome of individuals who contribute to iPSC banks could be used in *in vitro* disease models for drug discovery and research on causes and treatments of human disease.

Finally, as we approach the era of personalized medicine, genomic stem-cell data could provide information to target therapies for a particular individual. The medical field is rapidly moving in this direction, with several studies under way to characterize and compare the genomes and transcriptomes of hundreds or thousands of individuals and correlate this molecular information with medical information. Many drugs and therapies work only in a subset of patient groups, so finding genetic signatures or biomarkers to guide targeted treatments would vastly improve the efficiency and efficacy of medical care.

Developing a California stem-cell genomics infrastructure

Many California stem-cell researchers would benefit from greater access to cutting-edge genomics approaches. Although a number of the major California universities and research institutes have genomics centers that generally offer a variety of fee-based services, including massively parallel (second-generation or deep-) DNA sequencing, the state currently lacks comprehensive genomics centers with a concentration of expertise and capacity like that found at top centers, such as the Broad Institute (Cambridge, MA, USA), Washington University Genome Sequencing Center (St. Louis) and the Human Genome Sequencing Center at Baylor College of Medicine (Houston). This is somewhat surprising because California is home to several companies that produce state-of-the-art DNA-sequencing equipment (including Illumina and Sequenom of San Diego, Life Technologies of Carlsbad, Combimatrix of Irvine, Pacific Biosciences of Menlo Park and Complete Genomics of Mountain View). The private sector is in some instances offering services analogous to what the genomic centers offer in terms of next-generation sequencing and bioinformatics; however, it is beyond their scope to offer data management and infrastructure services to analyze, test and draw conclusions from such data at a community level.

In our view, it would be neither practical nor appropriate for CIRM to attempt to establish one or two genomics institutes on a par with the three national centers mentioned above. Each of those facilities was built with enormous infrastructure investment, and they operate with budgets in the tens of millions of dollars per year. Furthermore, they all support a wide range of research activities, the vast majority of which are not directly related to stem cells. Thus, efforts by CIRM in this field should be

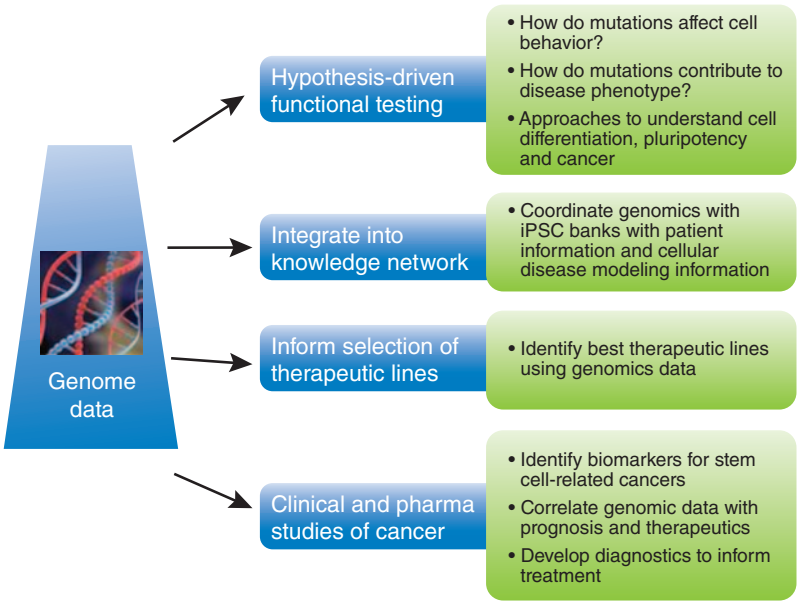

Figure 3 Benefits of genomics data for a broad range of activities, from basic research to clinical applications.

Box 2 Pillars of a California Stem-Cell Consortium

Several capabilities would need to be developed to facilitate integration and analysis of large volumes of stem-cell genome data. We break out some of these in detail below:

- Central data coordination with the capacity for secure data storage, petaflop-scaled data transfer and wide access by the scientific community
- Technology development, such as:
  - Single-cell transcriptomics (quantitative)
  - Detection of rare sequences in cell mixtures ('Amplicon' sequencing)
  - Single-cell methylomes
  - Subcellular transcriptomes
- Coordination to perform outstanding projects, such as:
  - Epigenetics of cell differentiation
  - Genomic variation in stem cells during culture
- Analysis of therapeutic and banked stem-cell lines coordinating with CIRM-funded disease teams and cell banks
- A mechanism whereby California stem-cell principal investigators could submit proposals to the consortium to engage with the consortium's genomics and biocomputing expertise
- Negotiated deals through volume to obtain best prices for CIRM grantees for provision of next-generation sequencing services and equipment
- Through consensus building, the centers could set and implement community standards (for example, cell lines, standard operating procedures and analysis methods)
- The centers' reach could be extended through Collaborative Funding Partnerships
- Development of business models to achieve long-term sustainability for at least a subset of its activities

directed toward enabling California's stem-cell scientists to gain access to technologies and expertise that can specifically benefit their research programs by funding several high-priority projects that will achieve some or all of the research goals listed in **Box 1** (including technology development for key bottlenecks) and developing advanced facilities for integration and analysis of large volumes of stem-cell genome data (**Box 2**).

Such an effort could be modeled along the lines of the ENCODE (Encyclopedia of DNA Elements) project, which is funded by the National Human Genome Research Institute (NHGRI) through a request for applications (RFA) process. Its major objective was to bring together investigators with a diverse set of experimental and computational expertise to identify all functional elements in human DNA, and it does so through an open consortium available to any investigator willing to use its established criteria to maintain standards in data acquisition and analysis. An important function is to develop standards to which consortium members adhere in order to reduce experimental noise and make data comparable across laboratories. In 2003, ENCODE's pilot phase was funded for $15 million to analyze all epigenetic modifications in a small segment of the human genome. Now that ENCODE is in its production phase, NHGRI is providing $80 million over four years to extend this effort genome-wide by funding a data coordination center (based in the University of California at Santa Cruz in conjunction with the University of California at San Diego's Supercomputer Center), a data analysis center (based in the European Bioinformatics Institute in Hinxton, UK) and a Technology Development Effort (comprising six principal investigators, five in the United States and one in Singapore).

In an example of another large-scale genomics project, the US National Institutes of Health Roadmap Epigenomics Mapping Consortium was launched in 2008 to use next-generation sequencing to map normal human epigenomes from stem cells and tissues commonly involved in disease, including human embryonic stem cells. The goal is to provide reference epigenomes as the starting point for a wide range of future studies[12]. The Roadmap project is providing five-year funding for four epigenome mapping centers (the Massachusetts Institute of Technology in Cambridge, MA, USA, the University of California at San Francisco, the Ludwig Center for Cancer Research-San Diego, and the University of Washington in Seattle), as well as a data coordination center at Baylor College of Medicine, technology development projects comprising eight principal investigators, and an initiative to discover new epigenetic marks in mammalian cells (also involving eight principal investigators). The project is also providing a public portal with protocols, downloadable tools and quality metrics for data release.

The centers of excellence created by a CIRM stem-cell genomics initiative could be hosted within established California universities or research institutes and would likely augment current genomics or bioinformatics core facilities to capitalize on existing expertise and infrastructure and interface with private companies providing services, where appropriate. Where possible, the centers would also coordinate with earlier or ongoing initiatives, such as ENCODE and the NIH Roadmap, to make the most of expertise, standards and data sets. In competing for awards, applicant institutions would propose facility structure, administration, policies and activities to promote collaboration and advance the program mission. In addition, institutions would be expected to commit matching funds and other appropriate resources to help establish and maintain the centers, including assurance that the appropriate expertise would be recruited and/or maintained, so that CIRM's grantees would have access to the counsel and collaboration they need. A model for sustainability beyond the lifetime of CIRM would be essential for funding. Because the centers would be expected to serve CIRM-supported scientists from a variety of other California institutions, multi-institutional consortia should be encouraged using CIRM's 'Collaborative Funding Partner' mechanism.

Conclusions

For the stem-cell field to fully benefit from genomics technology, an unprecedented degree of rigor and detail (in terms of defining and adhering to a well-defined set of standards and procedures) and data coordination must be applied in ways that are unfamiliar to most academic stem-cell biologists, outside of those engaged in cell banking and manufacturing. As computational and systems approaches become central to defining the concept of a gene[13] and regulation of cell state[14], the stem-cell community must look to forming working partnerships with computational biologists who think in these terms. Progressively, proteomic and metabolomics data sets will be integrated, but these efforts must build on the standardized sets of cells and protocols initially established by a stem-cell consortium focusing on genomics. Only by adopting standards for sample and data handling that minimize experimental noise will stem-cell 'omics' efforts engage top-rank computational scientists to analyze data sets generated from stem cells under highly controlled states of differentiation and culture conditions. It is hoped that through this level of analysis, stem-cell scientists can join the ranks of cancer cell biologists and

microbiologists to achieve the highest level of understanding of their biological systems currently possible.

By initiating and funding such a cross-disciplinary effort for the California stem-cell community and beyond—one that will impact virtually every aspect of stem-cell research—it is to be hoped that CIRM and its funding partners would leave a formidable and pioneering footprint, marking an indelible contribution to biological science, technological innovation and cell therapies.

COMPETING FINANCIAL INTERESTS
The authors declare no competing financial interests.

ACKNOWLEDGMENTS
The authors thank the many scientists with whom we have had illuminating discussions essential to formulating this article, as well as colleagues at CIRM.

1. National Research Council of the National Academies. Committee on a Framework for Development a New Taxonomy of Disease; National Research Council. *Toward Precision Medicine: Building a Knowledge Network for Biomedical Research and a New Taxonomy of Disease*. (The National Academies Press, Washington, DC, 2011).
2. Mayshar, Y. *et al. Cell Stem Cell* **7**, 521–531 (2010).
3. Gore *et al. Nature* **47**, 63–67 (2011).
4. Hussein, S.M. *et al. Nature* **471**, 58–62 (2011).
5. Subkhankulova, T. *et al. BMC Genomics* **9**, 268 (2008).
6. Stahlberg, A. *et al. Nucleic Acids Res.* **39**, e24 (2011).
7. Kalisky, T. & Quake, S. *Nat. Methods* **8**, 311–314 (2011).
8. Eldar, A. & Elowitz, M.B. *Nature* **467**, 167–173 (2010).
9. Reya, T. *et al. Nature* **414**, 105–111 (2001).
10. Bao, S. *et al. Nature* **444**, 756–760 (2006).
11. Dean, M. *et al. Nat. Rev. Cancer* **5**, 275–284 (2005).
12. Bernstein, B.E. *et al. Nat. Biotechnol.* **28**, 1045–1048 (2010).
13. Gerstein, M.B. *et al. Genome Res.* **17**, 669–681 (2007).
14. Lu, R. *et al. Nature* **462**, 358–362 (2009).