

Meeting report

Large-scale discovery and validation of functional elements in the human genome

Bradley E Bernstein^{*†} and Manolis Kellis^{*‡}

Addresses: ^{*}Broad Institute of Harvard and Massachusetts Institute of Technology, Cambridge, MA 02139, USA. [†]Department of Pathology, Brigham and Women's Hospital and Harvard Medical School, Boston, MA 02115, USA. [‡]MIT Computer Science and Artificial Intelligence Laboratory, The Stata Center, 32 Vassar Street, Cambridge, MA 02139, USA.

Correspondence: Bradley E Bernstein. E-mail: bbernst@fas.harvard.edu. Manolis Kellis. E-mail: manoli@mit.edu

Published: 1 March 2005

Genome Biology 2005, **6**:312

The electronic version of this article is the complete one and can be found online at <http://genomebiology.com/2005/6/3/312>

© 2005 BioMed Central Ltd

A report on the genomics workshop 'Identification of Functional Elements in Mammalian Genomes', Cold Spring Harbor, New York, 11-13 November 2004.

Computational and experimental genomics researchers convened at Cold Spring Harbor Laboratory at the end of 2004 to address the ambitious goal of identifying all the functional elements in the human genome. The functional elements discussed at the meeting included protein-coding genes, regulatory elements, RNA genes and DNA sequences that dictate chromosome structure or replication. The presentations described diverse approaches to the problem, ranging from innovative comparative genomic methods to high-throughput functional assays designed to identify and validate such elements.

The meeting followed a gathering of the ENCODE consortium, which aims to identify a comprehensive 'Encyclopedia of DNA elements' in the human genome [<http://www.genome.gov/10005107>]. The consortium, organized and funded by the National Human Genome Research Institute, is initially focusing on designated regions comprising approximately 1% of the human genome, and it strongly emphasizes technology development. Participating laboratories are developing computational techniques for sequence assembly, gene identification and regulatory motif discovery, as well as experimental methods for identifying transcripts, chromatin structures and regulatory regions. Against this backdrop, most speakers described experimental and computational approaches that generated vast numbers of candidate functional elements, from comprehensive transcript catalogs to lists of highly conserved sequence elements. They were complemented by a smaller number of presentations

dealing with the daunting task of systematically validating these elements.

Genes and transcripts

Mike Snyder (Yale University, New Haven, USA) and Tom Gingeras (Affymetrix, Santa Clara, USA) covered genome-scale technologies for transcript identification and the large numbers of new candidate elements emerging from these studies. Snyder described a complete tiling of the non-repetitive human genome, with 134 arrays containing 52 million oligonucleotide probes. In an effort to identify transcriptionally active regions systematically, his group hybridized RNA extracted from liver against these arrays. Gingeras described similar arrays covering one third of the genome that have been used to screen several human cell lines for transcribed sequences. Both studies identified surprisingly large and diverse collections of transcripts, a high proportion of which do not correspond to existing gene annotations. The presenters pointed out that a major challenge is ahead to define the functional significance of the thousands of novel transcripts identified.

Gustavo Glusman (Institute for Systems Biology, Seattle, USA) presented an orthogonal computational approach to gene identification, which does not observe transcription directly but instead relies on the marks it leaves behind in a genome sequence. He described four computational signatures of transcribed sequences, each relying on a different side-effect of transcription. One signature is the increased frequency of G and T nucleotides observed in the coding strand of genes, which is attributable to a mutational bias introduced during transcription-coupled DNA repair. Another signature is a bias in the orientation of transposable elements within genes, which is attributable to the fact that

polyadenylation signals in the transposon are rejected if they occur early in the coding strand but can be tolerated on the reverse strand. Taken together, the four tests provide a new tool for gene identification, which performs best where other tools fail. For example, the signals observed are strongest for long genes with very small exons, where traditional tools based on hidden Markov models fail. Using these methods on the human genome, Glusman has found evidence for thousands of genes that have not previously been annotated, some of which he has validated experimentally.

Regulatory mechanisms

Tim Hubbard (Wellcome Trust Sanger Institute, Hinxton, UK) presented a new computational approach to the discovery of regulatory motifs in promoter regions. Instead of searching for single motifs, the technique looks for multiple query sequences simultaneously and effectively explores motif space in bulk. The strength of the approach comes from the parallel exploration, which makes it well suited for distinguishing promoters densely populated with interacting regulatory motifs. Although initial results in yeast were promising, Hubbard pointed out a number of computational and implementation challenges in scaling this approach to mammalian genomes. Once these challenges have been overcome, this holistic approach to motif discovery will offer an additional way of obtaining a global understanding of regulatory elements.

John Stamatoyannopoulos (Regulome, Seattle, USA) and Gregory Crawford (National Human Genome Research Institute (NHGRI), Bethesda, USA) presented methodologies for the systematic identification of regions of open chromatin in the human genome. The methods involve mapping DNase I hypersensitive sites, which are known to correlate with many types of functional elements, including promoters, enhancers and insulators (sites at boundaries between open chromatin and inactive heterochromatin). Both investigators have cloned sites cut by DNase I, carried out highly parallel sequencing to map them onto the genome, and integrated the resulting maps with the University of California at Santa Cruz (UCSC) genome browser [<http://genome.ucsc.edu>] for visualization. They found that hypersensitive sites correlate with transcription starts, CpG islands and regions of high sequence conservation at the genome-wide level. They are following up on their findings by screening hypersensitive sites in multiple cell types (such as the ones listed on the ENCODE website [<http://www.genome.gov/10005107>]) to assess their tissue specificities, and to discover additional candidate elements.

Large-scale validation

A smaller number of presentations dealt with the validation problem. Nathan Trinklein (Stanford University, USA) focused on human promoters and described high-throughput

methods for validating computationally predicted regulatory regions. Predicted promoters were cloned upstream of reporter genes and their activity tested in various cell lines using transient transfection. Gabriela Loots (Lawrence Livermore National Laboratory, Livermore, USA) described functional assays for high-throughput validation of genes and regulatory sequences in the tropical frog *Xenopus tropicalis*. Transgenic techniques in frog embryos are used to test the influence of regulatory sequences on gene-expression patterns previously analyzed by *in situ* hybridization.

Greg Elgar (MRC Rosalind Franklin Centre for Genomics Research, Cambridge, UK) described methods for validating noncoding functional elements predicted through comparative genomic analysis. His group has identified more than 1,000 noncoding sequences that are highly conserved between human and pufferfish (*Fugu*) genomes. These candidate elements tended to reside near genes that act as developmental regulators, and were not found in invertebrate genomes. Zebrafish embryos were used to test a subset of the conserved elements for their regulatory potential. Candidate regions were amplified by PCR and co-injected with green fluorescent protein (GFP) reporter constructs. A remarkably high proportion of the highly conserved noncoding sequences tested (23 of 25) were found to enhance GFP expression in a tissue-specific manner.

The many state-of-the-art technologies being applied to the identification of functional elements in genomes are producing huge numbers of candidate regions. High-throughput assays in human cells and model organisms for validating and functionally characterizing these candidates are critical to the overall goal of cataloging functional elements. Given the huge numbers of candidates, however, alternative approaches are also needed. A particularly promising technique for validating, characterizing and prioritizing candidate regions is cross-validation by simultaneous analysis of complementary experimental and computational datasets, and the ENCODE consortium is seeking to maximize its potential by focusing on a well defined subset of the human genome and incorporating computational tools for correlating multiple datasets. Beyond simply cataloging functional elements, this integration should also lead to a description of their complex interactions within the regulatory network of the cell.