Toward mapping the biology of the genome

Stephen Chanock[1]

Laboratory of Translational Genomics, Division of Cancer Epidemiology and Genetics, National Cancer Institute, National Institutes of Health, Advanced Technology Center, Bethesda, Maryland 20892-4605, USA

This issue of *Genome Research* presents new results, methods, and tools from The ENCODE Project (ENCyclopedia of DNA Elements), which collectively represents an important step in moving beyond a parts list of the genome and promises to shape the future of genomic research. This collection sheds light on basic biological questions and frames the current debate over the optimization of tools and methodological challenges necessary to compare and interpret large complex data sets focused on how the genome is organized and regulated. In a number of instances, the authors have highlighted the strengths and limitations of current computational and technical approaches, providing the community with useful standards, which should stimulate development of new tools. In many ways, these papers will ripple through the scientific community, as those in pursuit of understanding the "regulatory genome" will heavily traverse the maps and tools. Similarly, the work should have a substantive impact on how genetic variation contributes to specific diseases and traits by providing a compendium of functional elements for follow-up study. The success of these papers should not only be measured by the scope of the scientific insights and tools but also by their ability to attract new talent to mine existing and future data.

As soon as the first draft of a human genome was available in the late 1990s, investigators began to organize a linear sequence of base pairs for each chromosome, constructing a map of the human genome. Completion has not been easy and nearly 10% of the human genome remains resistant to fitting into the current map, mainly because of low complexity and duplicate segments (Bailey et al. 2002; International Human Genome Sequencing Consortium 2004). Before it was possible to envision a whole-genome sequence, genetics had been partly driven by the creation and modification of maps of relative coordinates based on incomplete constructs. Early on, geneticists constructed flimsy topological maps and were forced to generate a toponymy, namely, an understanding of the place names based on empirical evidence of recombination hot spots, to explain the results of mapping studies. The longstanding value of functional elements, here recombination frequencies, served adequately for the mapping of diseases and traits before the draft sequences of genomes began to appear. However, the emergence of a physical map of the human genome has accelerated the mapping of diseases and traits at an unprecedented rate.

As the topology of the human genome has begun to take shape, there has been a natural shift in focus from the linear sequence to a more detailed understanding of the genome space, specifically examining the changes in its structure and interactions with regulatory proteins. Many looked to the possible organization of genes in not only humans but in other species for insights into the biology of the genome. The world of genetics quickly began to catalog regions of nongenic conservation and transcribed elements, some of which possess critical regulatory function. It turns out that it will take more than conservation to develop a comprehensive catalog of functional elements. Early surveys indicated that nearly one half of functional elements are not well conserved (The ENCODE Project Consortium 2007). The field, focusing on cell-specific analyses of transcription networks, began to assign biological meaning to temporal relationships, but lacked precise

definitions of the functional elements of the genome. One of the signature programs to investigate this parallel world has been The ENCODE Project (ENCyclopedia Of DNA Elements), a far-reaching project that is nearly 10 years in gestation (The ENCODE Project Consortium 2004, 2007).

In this issue of *Genome Research*, we find 18 new papers reporting an exciting treasure trove of results, including novel methods and computational tools for navigating The ENCODE Project data (Arvey et al. 2012; Bánfai et al. 2012; Boyle et al. 2012; Charos et al. 2012; Cheng et al. 2012; Derrien et al. 2012; Harrow et al. 2012; Howald et al. 2012; Kundaje et al. 2012; Ladewig et al. 2012; Landt et al. 2012; Natarajan et al. 2012; Park et al. 2012; Schaub et al. 2012; Tilgner et al. 2012; Vernot et al. 2012; H Wang et al. 2012; J Wang et al. 2012). Early on, some questioned the wisdom of creating a reference set for the identification and characterization of "functional elements" in the human genome, but now, the investment in this project has begun to pay substantial dividends, with clearly more to come.

Mapping genome space

The current crop of papers offers new insights into the complexity of transcribed elements in the genome (Cheng et al. 2005; Kapranov et al. 2007). In 2007, following The ENCODE Project Consortium's analyses of 1% of the genome, Greally used the metaphor of the Ishihara test for color deficiencies to point out the gene-centric nature of the early annotation of mammalian genomes (Greally 2007). At that time, Gingeras suggested that the increased and overlapping transcriptional complexity necessitated a reconsideration of the definition of a gene, while Gerstein and colleagues argued that the definition of a gene is predicated on "a coherent set of potentially overlapping functional products" (Gerstein et al. 2007; Gingeras 2007). In this regard, the assignation of a gene is based on functional data superimposed on a physical region of the genome, which can have multiple and complex functional elements operating under distinct conditions or in distinct biological contexts.

The ENCODE surveys expanded to the whole genome, utilizing different approaches to map the RNA space, has stretched our understanding of the scope of transcripts and begun to fulfill

the prophecies of previous assessments of ENCODE data (Gerstein et al. 2007; Gingeras 2007; Greally 2007; Kapranov et al. 2007). In this regard, Howald and colleagues have shown that with RNA-sequencing (RNA-seq) (Wang et al. 2009), a substantial fraction of exons are not well-annotated, and to find these it will require targeted approaches that are mapped locally (Howald et al. 2012). With a validation rate of ~75% for predicted exon–exon junctions, we edge closer to a comprehensive set of transcripts. Still, the endgame for assembling a comprehensive catalog of functional exons could take longer than it took to get to this point. In other words, there is still much to discover and map before we can superimpose a new toponymy, namely, a sophisticated functional interpretation through computational and laboratory analyses.

Based on the recent data stream, annotation of long non-coding RNAs (lncRNAs) not only confirms that lncRNAs are generated by histone modification and splicing, similar to protein-coding genes, but in addition, the majority of lncRNAs are two-exon transcripts that are eventually processed into small RNAs (Derrien et al. 2012). It is not surprising that deep-resequencing studies of subcellular RNA fractions yield a small fraction of lncRNAs, but more often RNA drawn from multiple exons (Tilgner et al. 2012). The results also illustrate that splicing is highly cotranscriptional. Atypical small RNAs, known as mirtrons, roughly the size of a microRNA, can be produced through a non-canonical pathway (Ladewig et al. 2012). They may be useful in the investigation of the role of novel Dicer substrates, particularly as they appear to contribute to regulatory networks.

A survey of RNA editing based on RNA-seq data in 14 cell lines not only provides a new snapshot of RNA editing events, it confirmed that the majority of A-to-G (I) events occur in the 3′ UTR; non-A-to-G (I) map within five bases of exon boundaries, suggesting errors in splice-mapping (Park et al. 2012). Here, it is also evident that many editing events are cell specific, which has important consequences for investigating human diseases and traits.

Tools and consensus

The project should be lauded for its focus on developing novel tools through iterative analyses of an ongoing data stream. The generation of better tools has been central to the task of uncovering the complex relationships of the regulatory genome. Still, the effort has highlighted the limitations of current computational approaches.

One notable example is the consensus report on the guidelines and practices of chromatin immunoprecipitation (ChIP) followed by high-throughput sequencing with next-generation technologies (ChIP-seq) (Johnson et al. 2007; Landt et al. 2012). Since the technology is critical for mapping both transcription-factor binding and histone modification, establishing guidelines for the conduct and interpretation of data represents a seminal task in building a more detailed map of functional elements within the space of the genome. Hidden below the text, which provides the community with important metrics for planning, executing, and analyzing ChIP-seq experiments, are a set of observations that provide boundaries for its utility and the insights. The distillation of the consortium experience offers metrics that can be fruitfully applied to evaluate data quality, which in turn has an impact on the value of the reported insights. This ENCODE story also underscores the dangers of any arbitrary decisions made with respect to antibodies used and conditions employed. By generating and analyzing enough data from standardized pipelines, some of the "dirty laundry" of ChIP-seq has been uncovered, providing a re-

alistic assessment of the technique that clearly will need to develop further to sharpen our view of transcription factor occupancy. Along the way, we still catch a glimpse of the emerging, complex map of transcription-factor binding and histone modification, and the current observations are expected to be refined with new data that will, in turn, enable further technical and analytical modifications of an imperfect yet powerful technique.

Cognizant of the strengths and weaknesses of ChIP-seq, one of the ENCODE groups turned to CTCF occupancy and analyzed 19 human cell types (J Wang et al. 2012). They observed that CTCF binding varied across the genome and, in fact, the regulation of cell-selective occupancy is more complex. When they conducted bisulfite sequencing, it turns out that nearly half of variable CTCF occupancy mapped to differential DNA methylation patterns, refocusing our understanding of the relationship between DNA methylation and CTCF occupancy, highlighting the cell-specific importance as well as differences between immortal and normal cells. This latter point will certainly refuel the intense interest in mutational events that directly or indirectly alter CTCF occupancy in cancer biology (Mulligan et al. 2011). Not only does the map provide a glimpse of the breadth of the global occupancy pattern, but it also provides cancer biologists with genome-wide coordinates for investigation of mutational events uncovered in cancer genome sequencing (Hudson et al. 2010).

Beyond setting standards, other novel tools were developed: for example, a new method for unsupervised pattern discovery, the Clustered Aggregation Tool (CAGT) (Kundaje et al. 2012). When applied to over 5000 data set pairs to explore the relationship between histone modification and nucleosome positioning signal for bound transcription factors, an unexpected degree of heterogeneity in both histone modification and the position of nucleosomes near binding sites was observed, underscoring the difficulty in analyzing a temporal sequence of events.

Understanding the biology of genetic variation and mutation

The new tools and data stream from ENCODE should accelerate the investigation of how genomic variation contributes to disease (Asking for more. [Editorial] 2012). Its value should be felt across the spectrum of genetic diseases, from Mendelian disorders to complex diseases. Mapping diseases or traits rarely provide the final answer; instead it directs investigators to one or more variants/mutations that in turn require corroborative studies—including further fine-mapping studies, in vitro analyses, animal models, and population/family studies to elucidate the underpinnings of the genetic signal (Donnelly 2008; Chung and Chanock 2011).

To date, genome-wide association studies (GWASs) have successfully identified over 1500 loci that are conclusively associated with more than 150 complex traits and diseases in humans (Hindorff et al. 2009; http://www.genome.gov/gwastudies/). The majority of signals identified by GWASs do not map to coding regions and are common markers with allele frequencies of >5% in one or more populations. Tested markers are surrogates for functional variants that explain the underlying association. The statistical challenge of identifying rare variants in unrelated populations becomes more difficult as the minor allele frequency decreases, necessitating larger sample sizes that are often unattainable. Hence, correlative laboratory confirmation will be critical and certainly the ENCODE data will be instrumental in nominating variants for laboratory study. A recent example was published reporting a rare variant in the *MITF* gene that has an

allele frequency of ~1% and increases the risk for melanoma; laboratory investigation revealed that the mutation resulted in impaired sumoylation and differentially regulated *MITF* targets (Yokoyama et al. 2011).

Still, the arduous task of transitioning from a marker for disease to understanding the basic biology of the functional variant is daunting, primarily because each region has to be mapped to find the optimal variants for study. In fact, some groups, as they report new loci using GWAS, cite ENCODE data to highlight plausible candidate genes (Carvajal-Carmona et al. 2011). Notable examples of successful laboratory confirmation of functional single-nucleotide polymorphisms (SNPs) in linkage disequilibrium with the reported SNP marker in the GWAS have used ENCODE data to winnow the candidate variants and focus experimental work on a handful of variants. One group used ENCODE data to focus on allele-specific chromatin modeling in a locus of 17q12 and its contribution to risk for asthma and autoimmune disease (Verlaan et al. 2009). In pursuit of a bladder cancer GWAS signal on 8q24.2, H3K4me1 and H3K4me3 marks zeroed in on variants that regulate the prostate stem cell antigen (*PSCA*) gene that have a demonstrated effect on expression of the *PSCA* gene product (Fu et al. 2012). The ENCODE data was instrumental in determining that variants in 9p21 impair interferon-gamma signaling, which in turn contributes to the risk for coronary artery disease (Harismendy et al. 2011). In each of these cases, the ENCODE data was a useful signpost to direct investigators to variants with a higher prior for functional activity, subsequently confirmed.

Several papers in this set from ENCODE move us a step closer to improved annotation of regulatory variants, but extensive work is needed to provide the proof that the systematic assessment of ENCODE data improves the investigator's chances that promising variants can be validated in laboratory studies (Boyle et al. 2012; Schaub et al. 2012; Vernot et al. 2012). Still, the current set of ENCODE papers are a welcome and useful step forward, especially since the age of next generation sequencing will find many more less-common and rare variants that we will angst over their true significance for disease risk.

Schaub and colleagues present an analysis that systematically looks across multiple types of ENCODE data and crosses GWAS markers with functional elements (Schaub et al. 2012). They suggest that there is a subset of SNPs that fall in regions of high probability for a functional effect on gene regulation and, as expected, the majority of putative functional SNPs are in linkage disequilibrium with the reported markers derived from commercial SNP microarrays. The integrated analyses of ENCODE suggest that a subset of regulatory SNPs is more promising for follow-up work. While the proposed analysis suggests that >75% of GWAS SNP markers can be mapped to variants that reside in functional elements, iterative analyses will be needed to not only uncover the biology (through extensive laboratory work), but also to refine the tool, if it is to be sufficiently robust.

The cross of ENCODE data with GWAS results provides further evidence that alterations in regulatory elements may account for a substantial fraction of genetic susceptibility, particularly for complex diseases. Moreover, it is now possible to look at patterns of regulatory variation both in individuals and across populations, from which we can infer new biological insights (Vernot et al. 2012). The data stream of the next several years should provide the community with a comprehensive annotation of variants that could contribute to human diseases.

With the expectation that whole-genome sequencing will rapidly proliferate, both in research studies and as personal analyses, a challenge arises: how to harness the data from many to infer what is suitable or appropriate for an individual. As the gap widens between what is well established and what is apparent by next-generation sequencing, the ENCODE data set and its tools for navigating the relationship between functional elements and phenotypes will be useful. One of the current ENCODE papers reports a new database, RegulomeDB, similar to HaploReg, which assists investigators in the assessment of variants with either an established or putative regulatory effect (Boyle et al. 2012; Ward and Kellis 2012). However, categorization of variants will continue to be a daunting task because the vast majority will fall into an indeterminate category for the foreseeable future. New laboratory and computational approaches are urgently needed to rigorously and safely interpret genetic variants. Toward this end, J Wang and colleagues integrated sequence features and chromatin structure across 119 different transcription factors and generated a TF-centric web repository, Factorbook, a useful resource in the follow-up of regulatory variants associated with complex diseases (J Wang et al. 2012).

Conclusion

This new set of ENCODE papers has changed our understanding of the landscape of the human genome by providing a finer resolution of its functional elements and uncovering new regulatory patterns. The deeper we probe into the space of the genome, the more complex it becomes. In broader terms, the ENCODE papers take us into new terrains, providing snapshots of transcription-factor occupancy and a widening complex network of RNA transcripts. With more detailed ways of looking into genome space, new connections emerge that promise to have an impact on the future of genomics, particularly as it relates to understanding the "regulatory genome." In this regard, it is likely that in the near future, we will be able to move beyond a recitation of a list of its parts to understand how its functional elements have evolved, and more importantly, how genomic variation contributes to disease.

An important added value of the ENCODE data set is that it is an opportunity to attract new talent to mine existing and future data. In the future, it is likely that its contribution could be measured by more than its discoveries and tools. Perhaps its legacy will partly be that it drew many young creative minds into the field of the regulatory genome.

These papers represent a work in progress, and in this regard, illustrate the limitations of both the current computational approaches and the data sets. Their value will be manifest in not only their widespread use but also in framing the next set of questions that will require the development of novel algorithms and tools. Still, the new insights gained thus far have broadened our understanding of the scope of the inner workings of the genome space, specifically cataloging signposts and markers of transcriptional activity. In this regard, the new tools and compendia of elements across the genome have sharpened our vision of the inner workings of how the genome functions and carries out its appointed business. At the same time, with this added dimension superimposed on the hybrid of physical and genetic map, we can begin to trace the genetic and epigenetic errors that result in traits and, more importantly, human diseases. Indeed, we are developing better lenses to look at the complex map of the genome space and we can expect it will take us to new places as well as refining the familiar. Perhaps the great writer, Marcel Proust, was prescient in his aphorism: "The real voyage of discovery consists not in seeking new landscapes, but in having new eyes."

References

Arvey A, Agius P, Noble WS, Leslie C. 2012. Sequence and chromatin determinants of cell-type–specific transcription factor binding. *Genome Res* (this issue). doi: 10.1101/gr.127712.111.

Asking for More. [Editorial] 2012. *Nat Genet* **44:** 733. doi: 10.1038/ng2345.

Bailey JA, Gu Z, Clark RA, Reinert K, Samonte RV, Schwartz S, Adams MD, Myers EW, Li PW, Eichler EE. 2002. Recent segmental duplications in the human genome. *Science* **5583:** 1003–1007.

Bánfai B, Jia H, Khatun J, Wood E, Risk B, Gundling W, Kundaje A, Gunawardena HP, Yu Y, Xie L, et al. 2012. Long noncoding RNAs are rarely translated in two human cell lines. *Genome Res* (this issue). doi: 10.1101/gr.134767.111.

Boyle AP, Hong EL, Hariharan M, Cheng Y, Schaub MA, Kasowski M, Karczewski KJ, Park J, Hitz BC, Weng S, et al. 2012. Annotation of functional variation in personal genomes using RegulomeDB. *Genome Res* (this issue). doi: 10.1101/gr.137323.112.

Carvajal-Carmona LG, Cazier JB, Jones AM, Howarth K, Broderick P, Pittman A, Dobbins S, Tenesa A, Farrington S, Prendergast J, et al. 2011. Fine-mapping of colorectal cancer susceptibility loci at 8q23.3, 16q22.1 and 19q13.11: Refinement of association signals and use of in silico analysis to suggest functional variation and unexpected candidate target genes. *Hum Mol Genet* **20:** 2879–2888.

Charos AE, Reed BD, Raha D, Szekely AM, Weissman SM, Snyder M. 2012. A highly integrated and complex PPARGC1A transcription factor binding network in HepG2 cells. *Genome Res* (this issue). doi: 10.1101/gr.127761.111.

Cheng J, Kapranov P, Drenkow J, Dike S, Brubaker S, Patel S, Long J, Stern D, Tammana H, Helt G, et al. 2005. Transcriptional maps of 10 human chromosomes at 5-nucleotide resolution. *Science* **308:** 1149–1154.

Cheng C, Alexander R, Min R, Leng J, Yip KY, Rozowsky J, Yan K-K, Dong X, Djebali S, Ruan Y, et al. 2012. Understanding transcriptional regulation by integrative analysis of transcription factor binding data. *Genome Res* (this issue). doi: 10.1101/gr.136838.111.

Chung CC, Chanock SJ. 2011. Current status of genome-wide association studies in cancer. *Hum Genet* **130:** 59–78.

Derrien T, Johnson R, Bussotti G, Tanzer A, Djebali S, Tilgner H, Guernec G, Martin D, Merkel A, Knowles DG, et al. 2012. The GENCODE v7 catalog of human long noncoding RNAs: Analysis of their gene structure, evolution, and expression. *Genome Res* (this issue). doi: 10.1101/gr.132159.111.

Donnelly P. 2008. Progress and challenges in genome-wide association studies in humans. *Nature* **456:** 728–731.

The ENCODE Project Consortium. 2004. The ENCODE (ENCyclopedia Of DNA Elements) Project. *Science* **306:** 636–640.

The ENCODE Project Consortium. 2007. Identification and analysis of functional elements in 1% of the human genome by the ENCODE pilot project. *Nature* **447:** 799–816.

Fu YP, Kohaar I, Rothman N, Earl J, Figueroa JD, Ye Y, Malats N, Tang W, Liu L, Garcia-Closas M, et al. 2012. Common genetic variants in the *PSCA* gene influence gene expression and bladder cancer risk. *Proc Natl Acad Sci* **109:** 4974–4979.

Gerstein MB, Bruce C, Rozowsky JS, Zheng D, Du J, Korbel JO, Emanuelsson O, Zhang ZD, Weissman S, Snyder M. 2007. What is a gene, post-ENCODE? History and updated definition. *Genome Res* **17:** 669–681.

Gingeras TR. 2007. Origin of phenotypes: Genes and transcripts. *Genome Res* **17:** 682–690.

Greally JM. 2007. Encyclopedia of humble DNA. *Nature* **447:** 782–783.

Harismendy O, Notani D, Song X, Rahim NG, Tanasa B, Heintzman N, Ren B, Fu XD, Topol EJ, Rosenfeld MG, et al. 2011. 9p21 DNA variants associated with coronary artery disease impair interferon-γ signalling response. *Nature* **470:** 264–268.

Harrow J, Frankish A, Gonzalez JM, Tapanari E, Diekhans M, Kokocinski F, Aken BL, Barrell D, Zadissa A, Searle S, et al. 2012. GENCODE: The reference human genome annotation for The ENCODE Project. *Genome Res* (this issue). doi: 10.1101/gr.135350.111.

Hindorff LA, Sethupathy P, Junkins HA, Ramos EM, Mehta JP, Collins FS, Manolio TA. 2009. Potential etiologic and functional implications of genome-wide association loci for human diseases and traits. *Proc Natl Acad Sci* **106:** 9362–9367.

Howald C, Tanzer A, Chrast J, Kokocinski F, Derrien T, Walters N, Gonzalez JM, Frankish A, Aken BL, Hourlier T, et al. 2012. Combining RT-PCR-seq and RNA-seq to catalog all genic elements encoded in the human genome. *Genome Res* (this issue). doi: 10.1101/gr.134478.111.

Hudson TJ, Anderson W, Artez A, Barker AD, Bell C, Bernabe RR, Bhan MK, Calvo F, Eerola I, Gerhard DS, et al. 2010. International network of cancer genome projects. *Nature* **464:** 993–998.

International Human Genome Sequencing Consortium. 2004. Finishing the euchromatic sequence of the human genome. *Nature* **431:** 931–945.

Johnson DS, Mortazavi A, Myers RM, Wold B. 2007. Genome-wide mapping of in vivo protein-DNA interactions. *Science* **316:** 1497–1502.

Kapranov P, Cheng J, Dike S, Nix DA, Duttagupta R, Willingham AT, Stadler PF, Hertel J, Hackermüller J, Hofacker IL, et al. 2007. RNA maps reveal new RNA classes and a possible function for pervasive transcription. *Science* **316:** 1484–1488.

Kundaje A, Kyriazopoulou-Panagiotopoulou S, Libbrecht M, Smith CL, Raha D, Winters EE, Johnson SM, Snyder M, Batzoglou S, Sidow A. 2012. Ubiquitous heterogeneity and asymmetry of the chromatin environment at regulatory elements. *Genome Res* (this issue). doi: 10.1101/gr.136366.111.

Ladewig E, Okamura K, Flynt AS, Westholm JO, Lai EC. 2012. Discovery of hundreds of mirtrons in mouse and human small RNA data. *Genome Res* (this issue). doi: 10.1101/gr.133553.111.

Landt SG, Marinov GK, Kundaje A, Kheradpour P, Pauli F, Batzoglou S, Bernstein BE, Bickel P, Brown JB, Cayting P, et al. 2012. ChIP-seq guidelines and practices of the ENCODE and modENCODE consortia. *Genome Res* (this issue). doi: 10.1101/gr.136184.111.

Mulligan CG, Zhang J, Kasper LH, Lerach S, Payne-Turner D, Phillips LA, Heatley SL, Holmfeldt L, Collins-Unerwood JR, Ma J, et al. 2011. CREBBP mutations in relapsed lymphoblastic leukemia. *Nature* **471:** 235–239.

Natarajan A, Yardımcı GG, Sheffield NC, Crawford GE, Ohler U. 2012. Predicting cell-type–specific gene expression from regions of open chromatin. *Genome Res* (this issue). doi: 10.1101/gr.135129.111.

Park E, Williams B, Wold BJ, Mortazavi A. 2012. RNA editing in the human ENCODE RNA-seq data. *Genome Res* (this issue). doi: 10.1101/gr.134957.111.

Schaub MA, Boyle AP, Kundaje A, Batzoglou S, Snyder M. 2012. Linking disease associations with regulatory information in the human genome. *Genome Res* (this issue). doi: 10.1101/gr.136127.111.

Tilgner H, Knowles DG, Johnson R, Davis CA, Chakrabortty S, Djebali S, Curado J, Snyder M, Gingeras TR, Guigó R. 2012. Deep sequencing of subcellular RNA fractions shows splicing to be predominantly co-transcriptional in the human genome but inefficient for lncRNAs. *Genome Res* (this issue). doi: 10.1101/gr.134445.111.

Verlaan DJ, Berlivet S, Hunninghake GM, Madore AM, Larivière M, Moussette S, Grundberg E, Kwan T, Ouimet M, Ge B, et al. 2009. Allele-specific chromatin remodeling in the APBP2/GSDMB/ORMDL3 locus associated with the risk of asthma and autoimmune disease. *Am J Hum Genet* **85:** 377–393.

Vernot B, Stergachis AB, Maurano MT, Vierstra J, Neph S, Thurman RE, Stamatoyannopoulos JA, Akey JM. 2012. Personal and population genomics of human regulatory variation. *Genome Res* (this issue). doi: 10.1101/gr.134890.111.

Wang Z, Gerstein M, Snyder M. 2009. RNA-Seq: A revolutionary tool for transcriptomics. *Nat Rev Genet* **10:** 57–63.

Wang H, Maurano MT, Qu H, Varley KE, Gertz J, Pauli F, Lee K, Canfield T, Weaver M, Sandstrom R, et al. 2012. Widespread plasticity in CTCF occupancy linked to DNA methylation. *Genome Res* (this issue). doi: 10.1101/gr.136101.111.

Wang J, Zhuang J, Iyer S, Lin XY, Whitfield TW, Greven MC, Pierce BG, Dong X, Kundaje A, Cheng Y, et al. 2012. Sequence features and chromatin structure around the genomic regions bound by 119 human transcription factors. *Genome Res* (this issue). doi: 10.1101/gr.139105.112.

Ward LD, Kellis M. 2012. HaploReg: A resource for exploring chromatin states, conservation and regulatory motif alterations within sets of genetically linked variants. *Nucleic Acids Res* **40:** D930–D934.

Yokoyama S, Woods SL, Boyle GM, Aoude LG, MacGregor S, Zismann V, Gartside M, Cust AE, Haq R, Harland M, et al. 2011. A novel recurrent mutation in MITF predisposes to familial and sporadic melanoma. *Nature* **480:** 99–103.