present, are often not perfectly predictive in up to 30% of patients. This is likely caused by modifier genes elsewhere in the genome, which can now be more easily identified and used to increase predictive precision.

The results from the ENCODE project will also breathe new life into diagnostics for common, chronic, noncommunicable diseases that are caused by a complex interplay between many genes and the environment. Such tests have often been designed based on the results of genome-wide association studies, which compare DNA samples of patients with and without disease using high-density genome-scanning approaches[2]. In many of these studies, the genomic regions identified as conferring disease risk were very distant from genes[3–5]. Diagnostic products based on this idea—the more risk factors, the higher the absolute risk of disease—were commercialized more than 5 years ago by several companies, but there was initial skepticism about these risk factors given that the identified genomic regions were often located far from genes and had no known function, despite validated and reproducible correlations with disease.

With the ENCODE data in hand, it should now be possible to tease out the biological basis of such risk-enhancing variants (for example, links to specific genes and biological processes as well as quantitative estimates of the impact of transcriptional dysregulation). As a result, the diagnostics field is likely to embrace compilations of low-effect-size functional variants, perhaps in combination with environmental factors, for personalized assessment of presymptomatic risk. Because there is usually an environmental component to risk that can be modulated and because screening and drug regimens are more effective with early detection, clinical intervention is straightforward in diseases such as macular degeneration, diabetes and myocardial infarction.

A related diagnostic approach developed over the past decade by several companies is RNA-based signatures of disease. One can easily imagine new expression signatures being developed and refined by an improved understanding, based on ENCODE data, of the causal relationships between transcriptional circuitry and diseases or drug responses, rather than by correlational studies, which are riddled with noise and chance coexpression. Moreover, the ENCODE data reveal new correlations between specific functional regions and gene-expression levels. These findings suggest that gene expression signatures could be replaced by tests that assay DNA sequences or proteins, both of which are less susceptible to degradation compared with RNA, allowing the community to move away from the use of fast-frozen tissues, which are expensive to collect and difficult to store while preserving their high quality.

We are approaching the day when it will be less expensive to sequence a patient's entire genome than to do targeted sequencing, particularly if more than one genomic locus is to be examined. A year ago, sequencing a human genome cost ~$4,000. Today, in an environment regulated by the Clinical Laboratory Improvement Amendments, it costs <$2,000 and may drop to <$1,000 within a year. These prices are within the range of diagnostic tests reimbursed by insurance companies. Thus, it seems that the low cost of clinical-grade genome sequences together with deeper understanding of disease-associated mutations made possible by large-scale functional genomics efforts will usher in a new era in diagnostics. A 'clinical genome project,' analogous in scale and funding to the Human Genome Project, is warranted to understand the correlations between the multitude of annotations and disease states, disease predisposition, drug response and host-pathogen interactions[6].

COMPETING FINANCIAL INTERESTS
The authors declare competing financial interests: details are available in the online version of the paper.

1. ENCODE Project Consortium. *Nature* **489**, 57–74 (2012).
2. Padhukasahasram, B. *PLoS One* **5**, e14338 (2010).
3. Boyle, A.P. *et al. Genome Res.* **22**, 1790–1797 (2012).
4. Hindorff, L.A. *et al. Proc. Natl. Acad. Sci. USA* **106**, 9362–9367 (2009).
5. Schaub, M.A., Boyle, A.P., Kundaje, A., Batzoglou, S. & Snyder, M. *Genome Res.* **22**, 1748–1759 (2012).
6. Ghadar, F., Sviokla, J. & Stephan, D.A. *Harv. Bus. Rev.* **90** (7-8), 25–27 (2012).

Uniting ENCODE with genome-wide proteomics

Young-Ki Paik & William S Hancock

An important goal of future omics research is to determine how the annotated regions of the genome control the production of protein isoforms.

The recent set of publications from the Encyclopedia of DNA Elements (ENCODE) Consortium reveals rapid progress in deciphering the 'parts list' of the human genome[1]. With this treasure trove of new data, now is the time for a close collaboration between ENCODE and the other omics research communities. Earlier this year, the Human Proteome Organization launched a similar large-scale initiative for proteomics—the chromosome-centric human proteome project (C-HPP)—which has a ten-year goal of characterizing the parts list of the human proteome[2]. A combined effort between the two initiatives would go a long way toward deciphering how the interacting genomic elements documented by ENCODE—including polygenes, transcription factor networks and single-nucleotide polymorphisms—control the families of isoforms generated at the protein level (**Fig. 1**).

What are the major points of synergy between ENCODE and C-HPP, and what can be gained by integrating the two data sets? Examples include the search for 'missing' proteins that have not been identified in proteomic studies, understanding the relationship of transcription-factor expression patterns to sets of protein variants, and defining the biochemical signatures of certain noncoding genomic regions and resulting protein modules. Despite the low abundance of most transcription factors, immunoprecipitation combined with high-sensitivity mass spectrometry platforms[3] can be used to study how transcription factors mediate the formation of chromosomal loops and heterochromatin stability. The role of long noncoding RNAs, sequence-specific factors and histone modifications can also be analyzed by cell-based proteomics with top-down mass spectrometry[4]. Given that only one-third of protein-coding genes annotated in GENCODE have been validated by mass

Young-Ki Paik is in the Department of Integrated Omics for Biomedical Science, Yonsei Proteome Research Center, Yonsei University, Seoul, Korea, and William S. Hancock is in Barnett Institute and Department of Chemistry and Chemical Biology, Northeastern University, Boston, USA, and World Class University Program, Yonsei University, Seoul, Korea.
e-mail: paikyk@yonsei.ac.kr or wi.hancock@neu.edu

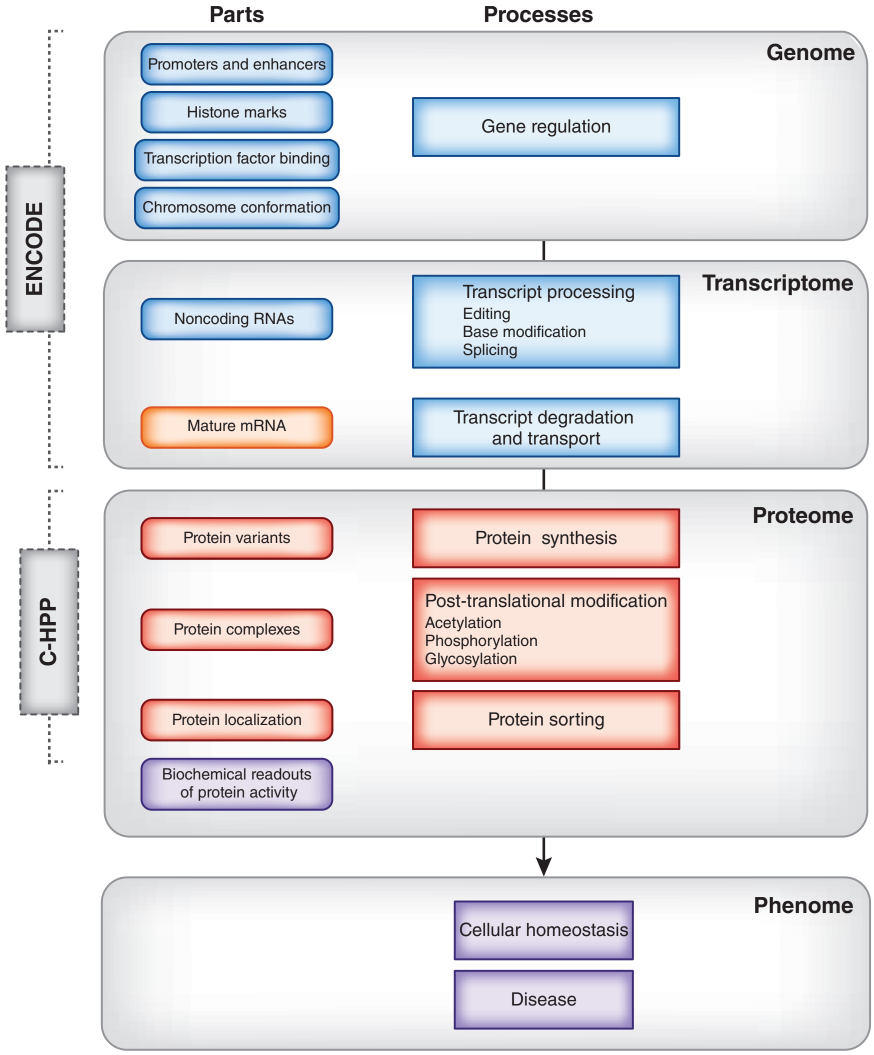

Figure 1 Synergies to be achieved by uniting ENCODE and C-HPP. Defining the flow of genetic information from genome to transcriptome to proteome will enable characterization of the relationship of transcription factor expression patterns to sets of protein variants and define biochemical signatures of certain noncoding genomic regions and the resulting protein modules. Colors delineate the purview and overlap of ENCODE and C-HPP: blue, ENCODE; red, C-HPP; orange, both; purple, neither.

spectrometry analysis[5], C-HPP will provide peptide evidence for newly identified coding loci[3].

The goals of C-HPP include characterizing all protein-coding gene products (including the major gene products and the alternative splicing variants), three major post-translational modifications (phosphorylation, acetylation and glycosylation) and common amino-acid polymorphisms[2] (**Fig. 1**). To date, the global effort in proteomics has resulted in the archiving of some 2×10^8 mass spectrometric measurements[6] and has characterized ~75% of the protein-coding genome, with the remainder of protein-coding loci expressed at low levels, in 'rare' sample sets or perhaps not at all[7].

The initial target of C-HPP is low-abundance proteins; such proteins are likely to exhibit both tissue and subcellular specificity and to be involved in gene regulation in poorly understood regions of the genome[7]. Another key challenge for the proteomics community has been the complexity of protein isoforms present in a given cell. Because of the large dynamic range of proteins in biological samples, low-abundance isoforms, which may be associated with disease[8], can be extremely difficult to measure. Furthermore, proteomic analysis has shown that post-translational modifications can be different in alternative splicing variants[2], and that amino-acid polymorphisms can generate additional variants[6].

Proteomics can also reveal the effects of high- and low-penetrance nucleotide polymorphisms on transcriptional mechanisms (such as alternative splicing), levels of translation and post-translational modifications. An example is the variation in phosphorylation status of a single-nucleotide variant of the important breast cancer oncogene *ERBB2* (ENSP00000269571)[2], with a substitution of tyrosine for serine at residue 1051. Such a substitution would require the activation of a tyrosine protein kinase rather than serine/threonine protein kinases, and although both phosphorylated forms could be active in downstream signaling, one could expect pathway perturbations and changes in regulatory mechanisms such as phosphatase activity. Thus, outputs like this from ENCODE could be followed up with proteomic exploration of the spectrum of ERBB2 protein isoforms present in biological and clinical samples. In addition, functional measurements, such as mapping of (altered) interaction partners and metabolomic effects, are required to establish the significance of this single-nucleotide variant.

Ultimately, ENCODE and C-HHP data should be integrated into modules (metabolic or signaling pathways, gene sets and chromosome regions) and networks to develop system-wide models of biological processes[9]. For example, *ERBB2* is associated with an amplicon that spans a significant region of chromosome 17 (q22 to q24), which correlates with increased expression of adjacent genes implicated in the oncogenic process (*PGAP3*, *GRB7*, *RARA* and *TOP2A*)[10]. Given that each of the genes in the amplicon has a set of isoforms that can be modulated in tumor tissues by ENCODE-defined regulatory elements, measuring the protein isoform profile is a desirable first step before module analysis. Of the 14 potential *ERBB2* alternative splicing transcripts[6] listed in Ensembl, proteomic measurements have potentially identified six.

ENCODE has already achieved an integration of databases containing genomic regulatory elements, transcriptomics and some proteomics data, whereas the Human Proteome Project has developed systems, such as the ProteomeXchange Consortium, to process mass spectrometry spectra from 25 teams in 21 countries and from the PRIDE, Peptide Atlas and GPMDB databases. The output of proteomic data curated at the Peptide Atlas and GPMDB sites is presented in a chromosome-centric format and with antibody-based tissue localization data (Protein Atlas)[6]. Such an output is well aligned with ENCODE data. C-HPP groups are already integrating transcriptomics data with proteomics, and C-HPP relies on RNA sequencing methods to guide genome-wide proteomic analysis. There are, however, a number of technical issues to solve, such as the signal-to-noise cutoff in assessing both types of data, elimination of false positives, temporal aspects of transcript and protein expression and, most importantly, cross-analysis of common sample sets.

At present, only a small fraction of the complexity of the proteome is available to complement the genomic knowledge provided by ENCODE. Why are certain parts of the proteome expressed rarely if at all? C-HPP needs the insights of the regulatory environment of gene expression to contextualize the proteomic signal. Conversely, our view of the significance of changes in gene regulation is hindered by a lack of understanding of the consequences at the protein level. Thus, the integration of C-HPP and ENCODE outputs as well as metabolomic data will set the stage for defining the full complexity of phenomes in different biological states.

COMPETING FINANCIAL INTERESTS

The authors declare no competing financial interests.

1. ENCODE Project Consortium. *Nature* **489**, 57–74 (2012).
2. Paik, Y.K. *et al. Nat. Biotechnol.* **30**, 221–223 (2012).
3. Sidoli, S. *et al. J. Proteomics* **75**, 3419–3433 (2012).
4. Tran, J.C. *et al. Nature* **480**, 254–258 (2011).
5. Harrow, J. *et al. Genome Res.* **22**, 1760–1774 (2012).
6. http://www.ensembl.org/, http://www.genecards.org/, http://www.nextprot.org/, http://gpmdb.thegpm.org/, http://www.peptideatlas.org/, http://www.proteinatlas.org/.
7. Lundberg, E. *et al. Mol. Syst. Biol.* **6**, 450 (2010).
8. Germann, S. *et al. J. Nucleic Acids.* **2012**, 269570 (2012).
9. Califano, A. *et al. Nat. Genet.* **44**, 841–847 (2012).
10. Kauraniemi, P. *et al. Cancer Res.* **61**, 8235–8240 (2001).

Research Highlights

Papers from the literature selected by the Nature Biotechnology editors (Follow us on Twitter, @NatureBiotech #nbtHighlight)

Offspring from oocytes derived from *in vitro* primordial germ cell–like cells in mice
Hayashi, K. *et al. Science* doi:10.1126/science.1226889 (4 October 2012)

A physically transient form of silicon electronics
Hwang, S.-W. *et al. Science* **337**, 1640–1644 (2012)

Loss of 5-hydroxymethylcytosine is an epigenetic hallmark of melanoma
Guo Lian, C. *et al. Cell* **150**, 1135–1146 (2012)

Super-resolution fluorescence imaging of organelles in live cells with photoswitchable membrane probes
Shim, S.-H. *et al. Proc. Nat. Acad. Sci.* **109**, 13978–13983 (2012)

Complete *Plasmodium falciparum* liver-stage development in liver-chimeric mice
Vaughan, A.M *et al. J. Clin. Invest.* **122**, 3618–3628 (2012)