Emerging Topics in Life Sciences (2017) **1** 121–124
https://doi.org/10.1042/ETLS20170129

Commentary

Big knowledge from big data in functional genomics

Chris P. Ponting

MRC Human Genetics Unit, The Institute of Genetics and Molecular Medicine, University of Edinburgh, Western General Hospital, Crewe Road, Edinburgh EH4 2XU, U.K.

Correspondence: Chris P. Ponting (chris.ponting@igmm.ed.ac.uk)

With so much genomics data being produced, it might be wise to pause and consider what purpose this data can or should serve. Some improve annotations, others predict molecular interactions, but few add directly to existing knowledge. This is because sequence annotations do not always implicate function, and molecular interactions are often irrelevant to a cell's or organism's survival or propagation. Merely correlative relationships found in big data fail to provide answers to the *Why* questions of human biology. Instead, those answers are expected from methods that causally link DNA changes to downstream effects without being confounded by reverse causation. These approaches require the controlled measurement of the consequences of DNA variants, for example, either those introduced in single cells using CRISPR/Cas9 genome editing or that are already present across the human population. Inferred causal relationships between genetic variation and cellular phenotypes or disease show promise to rapidly grow and underpin our knowledge base.

Single-gene studies in model or cellular systems have substantially advanced knowledge in the life sciences. Progress has relied on scientific acumen and on technological advances that provide detailed insights into processes at the atomic, molecular, multisubunit complex, cellular and sometimes organismal levels. These many successes, however, should not blind us as to how our knowledge is incomplete and error-prone. Virtually all (99.85%) protein sequences have no associated experimental evidence at the protein level and for 52% their annotations are flagged as containing possible errors (www.ebi.ac.uk/uniprot/TrEMBLstats). Furthermore, scientific knowledge from targeted studies has been gained unevenly: of all human brain-expressed genes for example, science has focused on very few, with the top 5% of such genes being the subject of 70% of the literature [1].

Whole-genome experiments seek to address these deficiencies of uneven coverage and incompleteness. These are aided by technological innovations that inexorably generate ever larger data sets. Critically, however, big data analysis *per se* reveals not mechanistic causes, but rather correlations and patterns, and leaves questions starting *Why* unanswered [2]. Even when subsequent experiments address more narrowly defined hypotheses while exploiting this data, these also often fail to determine causality. Correlations and patterns may describe the data set well, but they need to be supplemented by causal inferences in order to predict phenomena reliably. The transformation of large, unstructured data sets to insights (Figure 1) and predictive biology is challenging and rarely attained.

In human genomics, data and annotations have grown rapidly. The 3.2 billion base reference genome is partitioned currently into 20 338 protein-coding and 22 521 non-protein-coding gene annotations that are transcribed into 200 310 transcripts (www.ensembl.org/Homo_sapiens/Info/Annotation) that start from 308 214 locations [3]. Binding sites, often considered to regulate the activity of these genes, have been assigned to 636 336 regions occupying 8.1% of the genome [4]. Nevertheless, experiments imply that many protein–DNA and protein–RNA binding events are not consequential (i.e. are not functional) [5,6]: molecular events are often 'noise' that have no subsequent bearing on whether a cell or organism thrives and propagates [7].

Received: 6 September 2017
Revised: 12 September 2017
Accepted: 12 September 2017

Version of Record published:
14 November 2017

Emerging Topics in Life Sciences (2017) 1 121–124
https://doi.org/10.1042/ETLS20170129

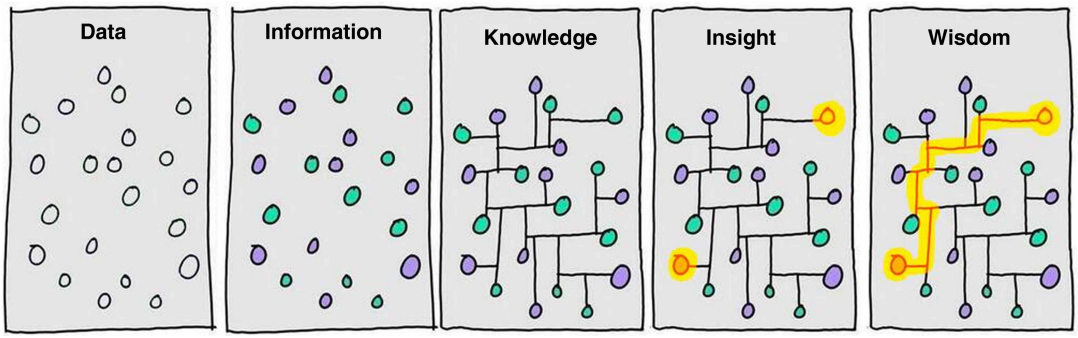

Figure 1. *Information isn't*.

'Data is not information, information is not knowledge, knowledge is not understanding, understanding is not wisdom' (Clifford Stoll and Gary Schubert). The drawing nicely captures some of the distinguishing features of these concepts. Wisdom should also permit reliable prediction. Illustration by David Somerville from original drawings by Hugh McLeod, reproduced with permission (personal communication).

Human genome annotation is an incomplete, undoubtedly biased and error-prone molecular parts list. Long-read [8] or targeted RNA sequencing [9] often reveals new or erroneous transcript models, even in well-annotated loci, and predictions of enhancers produce largely discordant results [10]. The value of annotations generated by large-scale big data 'omics projects', such as ENCODE [4], FANTOM5 [3] and the Human Cell Atlas [11], is not an immediate gain of new biological knowledge. Instead, their value is technical, by providing new standards, analytical approaches and reagents, as well as data that is accessible, standardised, reusable and extensive that can be exploited by anyone in order to frame new mechanistic hypotheses.

Despite these issues, genomics is not destined forever to produce only large annotation data sets of uneven completeness, quality and predictive potential. Rather the introduction of three novel approaches, based on emerging technologies and analytical methods, could transform its ability to make causal inferences and to address hypotheses. Critically, each is founded on DNA changes that causally lead to downstream effects; reverse causation — DNA mutation caused by phenotypic change — is excluded.

The first of these combines two recent technologies, namely single-cell genomics and multiplexed genome sequence editing by CRISPR/Cas9 [12–15], to couple individual genomic perturbations to transcriptomic read-outs in each of many single cells. Applications have investigated the downstream cellular effects of knocking out transcription factors [14] or genes involved in the unfolded protein [12] or the immune response [15]. When applied across the human genome, there is potential to determine how each DNA lesion causally alters a cell's survival, differentiation or proliferation thereby aiding the generation of more targeted functional hypotheses. Beyond the manipulation of single human cells and the editing of their genomes, it is hoped that real-time image-based high-content screening [16], real-time sampling of a living cell's contents [17] and spatial transcriptomics [18] will together provide the infrastructure required to generate and test functional hypotheses at the genome scale.

The second innovation also links DNA variation to phenotype, but at the human population not cellular level. Sequencing exomes of large cohorts, over half-a-million strong, is predicted to identify at least 7.5% of all possible loss-of-function variants, defined as point substitutions that either introduce stop codons or disrupt splice sites in protein-coding genes [19]. These variants are naturally occurring alleles whose deleterious effects result in their preferential loss from the population and cause their population frequencies to be lower than otherwise expected. Population-scale genome sequencing [20] thus will reveal an increasing number of functional sites whose mutation reduces reproductive success.

The final innovation is Mendelian randomisation. This approach applies the framework of randomised controlled trials to DNA variants that have a robust correlation with a modifiable exposure or biological intermediate [21,22]. In a first step, DNA variants are identified that predict the life-long levels of, and thus exposure to, a molecule. In the next step, these variants are tested for the extent by which they explain a complex trait or disease risk. For example, four DNA variants were found that showed genome-wide significance in their prediction of 25-hydroxyvitamin D (25OHD) levels; then, it was calculated that a two-fold increase in multiple

sclerosis disease risk is conferred by a combination of these alleles that reduces 25OHD levels, in a genetically determined manner, by an amount equal to 1 s.d. in log-transformed values [23]. The applicability of Mendelian randomisation has been substantially broadened by exploiting DNA variants that predict RNA [24,25] and protein [26,27] levels to test for a causal effect on traits or disease risk. While challenges need to be overcome, most specifically that of horizontal pleiotropy [28], Mendelian randomisation has potential to reveal causal relationships between DNA variant and trait, and between trait pairs [29].

As sequence data becomes cheaper and easier to generate, its acquisition will be ever more torrential. Nevertheless, in order to generate knowledge, this data needs first to be structured into reliable annotation before being used with approaches that predict causal relationships. Correlation alone will never be sufficient to determine function over effect or causation over statistical association. In time, our currently patchy knowledge will grow and join up. How big will we need big knowledge to be? Measuring a single phenotype caused by the substitution or deletion of each nucleotide in a human genome in, say, 2000 cell types would result in over 24 trillion observations. Yet, even this experiment would fail to account for cellular variation due to state, development, cancer, epistasis or external stimuli. Clearly, this is a path we are just beginning to tread.

Summary

- Life sciences are awash with data, but relatively bereft of knowledge.
- Human genome sequence annotations are extensive yet are incomplete, often inconsistent and error-prone, and fail to represent functional knowledge.
- Answering *Why* questions requires a detailed understanding of cause-and-effect, rather than correlations and statistical associations.
- Coupling single-cell transcriptomics to CRISPR/Cas9 genome editing, population-scale genome sequencing and Mendelian randomisation each has the potential to identify functions without being confounded by reverse causation.
- Genomic data is easy and relatively cheap to generate. The critical question is not whether such data can be generated, but whether it ought to be: specifically, whether it will generate new knowledge and have a high predictive value.

Abbreviations
25OHD, 25-hydroxyvitamin D.

Funding
Ponting group research is funded by the UK Medical Research Council and the Wellcome Trust.

Acknowledgements
I am grateful to Martin Taylor, Colin Semple, Luis Sanchez-Pulido, Oscar Bedoya-Reina and Tamir Chandra for insightful comments.

Competing Interests
The Author declares that there are no competing interests associated with the manuscript.

References
1 Pandey, A.K., Lu, L., Wang, X., Homayouni, R. and Williams, R.W. (2014) Functionally enigmatic genes: a case study of the brain ignorome. *PLoS ONE* **9**, e88889 https://doi.org/10.1371/journal.pone.0088889
2 Mazzocchi, F. (2015) Could big data be the end of theory in science? A few remarks on the epistemology of data-driven science. *EMBO Rep.* **16**, 1250–1255 https://doi.org/10.15252/embr.201541001

3 Forrest, A.R.R., Kawaji, H., Rehli, M., Baillie, J.K., de Hoon, M.J.L., Haberle, V. et al. (2014) A promoter-level mammalian expression atlas. *Nature* **507**, 462–470 https://doi.org/10.1038/nature13182
4 Dunham, I., Kundaje, A., Aldred, S.F., Collins, P.J., Davis, C.A., Doyle, F. et al. (2012) An integrated encyclopedia of DNA elements in the human genome. *Nature* **489**, 57–74 https://doi.org/10.1038/nature11247
5 Cusanovich, D.A., Pavlovic, B., Pritchard, J.K. and Gilad, Y. (2014) The functional consequences of variation in transcription factor binding. *PLoS Genet.* **10**, e1004226 https://doi.org/10.1371/journal.pgen.1004226
6 Davidovich, C., Wang, X., Cifuentes-Rojas, C., Goodrich, K.J., Gooding, A.R., Lee, J.T. et al. (2015) Toward a consensus on the binding specificity and promiscuity of PRC2 for RNA. *Mol. Cell* **57**, 552–558 https://doi.org/10.1016/j.molcel.2014.12.017
7 Doolittle, W.F., Brunet, T.D.P., Linquist, S. and Gregory, T.R. (2014) Distinguishing between 'function' and 'effect' in genome biology. *Genome Biol. Evol.* **6**, 1234–1237 https://doi.org/10.1093/gbe/evu098
8 Weirather, J.L., de Cesare, M., Wang, Y., Piazza, P., Sebastiano, V., Wang, X.-J. et al. (2017) Comprehensive comparison of Pacific Biosciences and Oxford Nanopore Technologies and their applications to transcriptome analysis [version 2; referees: 2 approved]. *F1000Research* **6**, 100 https://doi.org/10.12688/f1000research.10571.2
9 Mercer, T.R., Gerhardt, D.J., Dinger, M.E., Crawford, J., Trapnell, C., Jeddeloh, J.A. et al. (2011) Targeted RNA sequencing reveals the deep complexity of the human transcriptome. *Nat. Biotechnol.* **30**, 99–104 https://doi.org/10.1038/nbt.2024
10 Benton, M.L., Talipineni, S.C., Kostka, D. and Capra, J.A. (2017) Genome-wide enhancer maps differ significantly in genomic distribution, evolution, and function. *bioRxiv* https://doi.org/10.1101/176610
11 Regev, A., Teichmann, S., Lander, E.S., Amit, I., Benoist, C., Birney, E. et al. (2017) The human cell atlas. *bioRxiv* https://doi.org/10.1101/121202
12 Adamson, B., Norman, T.M., Jost, M., Cho, M.Y., Nunez, J.K., Chen, Y. et al. (2016) A multiplexed single-cell CRISPR screening platform enables systematic dissection of the unfolded protein response. *Cell* **167**, 1867–1882.e21 https://doi.org/10.1016/j.cell.2016.11.048
13 Datlinger, P., Rendeiro, A.F., Schmidl, C., Krausgruber, T., Traxler, P., Klughammer, J. et al. (2017) Pooled CRISPR screening with single-cell transcriptome readout. *Nat. Methods* **14**, 297–301 https://doi.org/10.1038/nmeth.4177
14 Dixit, A., Parnas, O., Li, B., Chen, J., Fulco, C.P., Jerby-Arnon, L. et al. (2016) Perturb-seq: dissecting molecular circuits with scalable single-cell RNA profiling of pooled genetic screens. *Cell* **167**, 1853–1866.e17 https://doi.org/10.1016/j.cell.2016.11.038
15 Jaitin, D.A., Weiner, A., Yofe, I., Lara-Astiaso, D., Keren-Shaul, H., David, E. et al. (2016) Dissecting immune circuits by linking CRISPR-pooled screens with single-cell RNA-seq. *Cell* **167**, 1883–1896.e15 https://doi.org/10.1016/j.cell.2016.11.039
16 Conway, J.R.W., Carragher, N.O. and Timpson, P. (2014) Developments in preclinical cancer imaging: innovating the discovery of therapeutics. *Nat. Rev. Cancer* **14**, 314–328 https://doi.org/10.1038/nrc3724
17 Actis, P., Maalouf, M.M., Kim, H.J., Lohith, A., Vilozny, B., Seger, R.A. et al. (2014) Compartmental genomics in living cells revealed by single-cell nanobiopsy. *ACS Nano* **8**, 546–553 https://doi.org/10.1021/nn405097u
18 Moor, A.E. and Itzkovitz, S. (2017) Spatial transcriptomics: paving the way for tissue-level systems biology. *Curr. Opin. Biotechnol.* **46**, 126–133 https://doi.org/10.1016/j.copbio.2017.02.004
19 Zou, J., Valiant, G., Valiant, P., Karczewski, K., Chan, S.O., Samocha, K. et al. (2016) Quantifying unobserved protein-coding variants in human populations provides a roadmap for large-scale sequencing projects. *Nat. Commun.* **7**, 13293 https://doi.org/10.1038/ncomms13293
20 Goldfeder, R.L., Wall, D.P., Khoury, M.J., Ioannidis, J.P.A. and Ashley, E.A. (2017) Human genome sequencing at the population scale: a primer on high-throughput DNA sequencing and analysis. *Am. J. Epidemiol.* 1–10 https://doi.org/10.1093/aje/kww224
21 Evans, D.M. and Davey Smith, G. (2015) Mendelian randomization: new applications in the coming age of hypothesis-free causality. *Annu. Rev. Genomics Hum. Genet.* **16**, 327–350 https://doi.org/10.1146/annurev-genom-090314-050016
22 Swerdlow, D.I., Kuchenbaecker, K.B., Shah, S., Sofat, R., Holmes, M.V., White, J. et al. (2016) Selecting instruments for Mendelian randomization in the wake of genome-wide association studies. *Int. J. Epidemiol.* **45**, 1600–1616 https://doi.org/10.1093/ije/dyw088
23 Mokry, L.E., Ross, S., Ahmad, O.S., Forgetta, V., Smith, G.D., Goltzman, D. et al. (2015) Vitamin D and risk of multiple sclerosis: a Mendelian randomization study. *PLoS Med.* **12**, e1001866 https://doi.org/10.1371/journal.pmed.1001866
24 Mancuso, N., Shi, H., Goddard, P., Kichaev, G., Gusev, A. and Pasaniuc, B. (2017) Integrating gene expression with summary association statistics to identify genes associated with 30 complex traits. *Am. J. Hum. Genet.* **100**, 473–487 https://doi.org/10.1016/j.ajhg.2017.01.031
25 Zhu, Z., Zhang, F., Hu, H., Bakshi, A., Robinson, M.R., Powell, J.E. et al. (2016) Integration of summary data from GWAS and eQTL studies predicts complex trait gene targets. *Nat. Genet.* **48**, 481–487 https://doi.org/10.1038/ng.3538
26 Sun, B.B., Maranville, J.C., Peters, J.E., Stacey, D., Staley, J.R., Blackshaw, J. et al. (2017) Consequences of natural perturbations in the human plasma proteome. *bioRxiv* https://doi.org/10.1101/134551
27 Yao, C., Chen, G., Song, C., Mendelson, M., Huan, T., Laser, A. et al. (2017) Genome-wide association study of plasma proteins identifies putatively causal genes, proteins, and pathways for cardiovascular disease. *bioRxiv* https://doi.org/10.1101/136523
28 Davey Smith, G. and Hemani, G. (2014) Mendelian randomization: genetic anchors for causal inference in epidemiological studies. *Hum. Mol. Genet.* **23**, R89–R98 https://doi.org/10.1093/hmg/ddu328
29 Hemani, G., Bowden, J., Haycock, P.C., Zheng, J., Davis, O., Flach, P. et al. (2017) Automating Mendelian randomization through machine learning to construct a putative causal map of the human phenome. *bioRxiv* https://doi.org/10.1101/173682