Spring 2023 – Epigenetics and Systems Biology
Discussion Session (Systems Biology)
Michael K. Skinner – Biol 476/576
Week 4 (February 2, 2023)

Systems Biology (Omics Technology)

Primary Papers

1. Dahal, et al. (2020) Proteomics. 20917-18):e1900282. (PMID: 32579720)
2. Wagner & Klein (2020) Nature Rev Genet. 21(7):410-427. (PMID: 32235876)
3. ENCODE Consortium (2012) Nature 489:57-74. (PMID: 22955616)

Discussion

Student 7 – Ref #1 above

- What omics protocols and technology are used?
- What integration was required for the technology?
- How can the information be used in genome scale design?

Student 8 – Ref #2 above

- What omics technology was integrated?
- How did single cell omics help the cell lineage analysis?
- Did the single cell molecular insights help understand the physiology?

Student 9 – Ref #3 above

- What is ENCODE?
- What types of technology and data was obtained?
- What novel observations were made?

Proteomics
Proteomics and Systems Biology
www.proteomics-journal.com

Synthesizing Systems Biology Knowledge from Omics Using Genome-Scale Models

*Sanjeev Dahal, James T. Yurkovich, Hao Xu, Bernhard O. Palsson, and Laurence Yang**

Omic technologies have enabled the complete readout of the molecular state of a cell at different biological scales. In principle, the combination of multiple omic data types can provide an integrated view of the entire biological system. This integration requires appropriate models in a systems biology approach. Here, genome-scale models (GEMs) are focused upon as one computational systems biology approach for interpreting and integrating multi-omic data. GEMs convert the reactions (related to metabolism, transcription, and translation) that occur in an organism to a mathematical formulation that can be modeled using optimization principles. A variety of genome-scale modeling methods used to interpret multiple omic data types, including genomics, transcriptomics, proteomics, metabolomics, and meta-omics are reviewed. The ability to interpret omics in the context of biological systems has yielded important findings for human health, environmental biotechnology, bioenergy, and metabolic engineering. The authors find that concurrent with advancements in omic technologies, genome-scale modeling methods are also expanding to enable better interpretation of omic data. Therefore, continued synthesis of valuable knowledge, through the integration of omic data with GEMs, are expected.

1. Introduction

Omic technologies aim to measure the molecular composition of a cell in its entirety. These measurements profile the functional potential (genomics) and activity (transcriptomics, proteomics, metabolomics) of an organism at the systems scale. These entities (genome, transcriptome, proteome, metabolome) are interrelated through expression, metabolism, signaling, and regulation. Understanding and interpreting each of these omic data types individually and combined could help unravel the mechanistic intricacies of biological systems. However, the interconnectedness among these different levels of function within a biological system poses significant challenges for studying the underlying mechanisms and relationships.

Each individual omic data type only describes part of the larger system. Therefore, integrative omic platforms are being developed. For instance, proteogenomics (proteomics with genomics/transcriptomics) can address genetic polymorphisms,[1] improve the detection of novel genes or identify misannotated open reading frames (ORFs),[2] and address the "missing protein problem," which refers to predicted proteins that are not detected in proteomic data.[3] Likewise, metabolomics has been combined with other omic platforms to demonstrate the environmental effects on post-translational modification (PTM) rates,[4] to understand the regulation of metabolite levels[5] and to elucidate complex interactions between the host, commensal bacteria, and pathogens.[6] These diverse datasets can yield a comprehensive understanding of biological mechanisms when they are contextualized and unified into a systems view of biology.

Systems biology is an interdisciplinary field that aims to predict the behavior of biological systems (i.e., phenotype) by considering interactions among biological parts in the context of the whole system. One approach to predicting system behavior is computational modeling such as genome-scale modeling. Genome-scale models (GEMs) have been used to analyze individual and multi-omic data sets.[7] GEMs can be analyzed using various methods including COnstraint-Based Reconstruction and Analysis (COBRA) methods (**Figure 1**).[8]

In general, for COBRA analysis, first the molecular composition of an organism can be represented as a network of interactions in which nodes represent specific entity (e.g., metabolites) and edges represent the interaction between these entities (such as substrate-product conversion). To implement modeling using COBRA framework, these networks are converted to stoichiomet-

Dr. S. Dahal, H. Xu, Prof. L. Yang
Department of Chemical Engineering
Queen's University
19 Division Street, Kingston, ON K7L 3N6, Canada
E-mail: laurence.yang@queensu.ca
Dr. J. T. Yurkovich
Institute for Systems Biology
401 Terry Ave. N. Seattle, WA 98109, USA
Prof. B. O. Palsson
Department of Bioengineering
University of California San Diego
9500 Gilman Drive La Jolla, CA 92093, USA
Prof. B. O. Palsson
Department of Pediatrics
University of California San Diego
9500 Gilman Drive La Jolla, CA 92093, USA

The ORCID identification number(s) for the author(s) of this article can be found under https://doi.org/10.1002/pmic.201900282

DOI: 10.1002/pmic.201900282

 © 2020 WILEY-VCH Verlag GmbH & Co. KGaA, Weinheim

ric matrix (S-matrix) in which rows represent molecular entities and columns represent their interactions. Then, the S-matrix can be analyzed using mathematical optimization as formulated in the COBRA framework.[8] In this approach, the steady state of an organism can be solved by optimizing an objective function (Z). Without adding constraints to this optimization problem, we can get infinite number of solutions (fluxes) that can satisfy the steady-state assumption. Therefore, the optimization problem is subjected to certain constraints which are: $Sv = 0$ (mass balance constraints), and $l \leq v_i \leq u$ (flux bounds). Here, v_i is the flux vector and l, u are the lower and upper bounds of the flux of the ith reaction. Hence, by optimizing Z, one can approximate the flux state of an organism, and identify molecular interactions that lead to such state.

GEMs were traditionally used to model the metabolic state of an organism (metabolic or M-model). In recent years, however, GEMs have also been utilized to compute the metabolic and proteomic state of an organism (metabolism and macromolecular expression or ME-model). Since ME-models deal with metabolism and proteome allocation, additional constraints including coupling constraints and biomass dilution constraints are added (check ref. [9] for more information). In this article, we review how modeling in systems biology has yielded new insights from omic data, and how GEMs can be used to interpret large-scale data. Modeling platforms in systems biology can integrate multiple omics and synthesize knowledge. Such modeling platforms include kinetic modeling (stochastic or deterministic), Boolean formalisms, Bayesian approaches, and COBRA.[8b] We focus on the use of COBRA methods in which steady state of a biological system are modeled by optimizing an objective function subjected to constraints including thermodynamic, stoichiometric, and enzymatic ones. We highlight recent advances in COBRA that were made to integrate multi-omic data types. We organize our review by the omic data types analyzed and COBRA methods used (**Figure 2**).

2. From Annotated Genome Sequences to Genome-Scale Models of Cell Metabolism

The genome encodes the functional capabilities of an organism. Genomics is the study of the whole genome of an organism. Since the first genome sequence of human mitochondria in 1981, there has been a steady increase in the publications that contribute to this field[10] (**Figure 3**). With the explosion of sequenced genomes, tools in comparative genomics have been developed to annotate sequences of previously uncharacterized genomes to unveil their functional potential.[11] With the advent of next generation sequencing technologies, sequencing genomes has become relatively quick, easy, and cheap.[12] However, even though we can sequence an organism, we still do not understand the full functional potential of organisms.[13]

A GEM is a modeling approach for mathematically describing all possible functions that are encoded by the genome, and their interactions, within the context of the full interaction network. For reconstruction of the network of an organism, genomic data and proper annotation are crucial in order to represent the correct interaction between various molecular entities. Following the reconstruction, COBRA methods can be utilized to analyze

Sanjeev Dahal is currently a postdoctoral fellow in Yang Lab in the Department of Chemical Engineering at Queen's University (Canada). He graduated with a B.Sc. in Biology from the University of New Orleans (USA) in 2012. He received his Ph.D. in genome science from the University of Tennessee, Knoxville (USA) in 2019. He is interested in the systems understanding of genotype-environment-phenotype relationship in microbes. Currently, his main project focuses on the genome-scale modeling of metabolism and macromolecular expression of pathogenic bacteria.

Bernhard Palsson is the Distinguished Galletti Professor of Bioengineering, Principal Investigator of the Systems Biology Research Group in the Department of Bioengineering, and Professor of Pediatrics at the University of California, San Diego. He is CEO of the Novo Nordisk Center for Biosustainability in Denmark, working in this capacity since 2011. His research includes developing methods to analyze metabolic dynamics and formulating complete models of cells. He is a member of the National Academy of Engineering and is a Fellow of the AIMBE, AAAS, and AAM.

Laurence Yang is a professor in the Department of Chemical Engineering and Queen's National Scholar in Systems Biology at Queen's University. He received his Ph.D. in Chemical Engineering from the University of Toronto, after which he worked as a scientist at a synthetic biology company. Prior to joining Queen's, he worked as a project scientist in the Department of Bioengineering at the University of California, San Diego, where he also received his postdoctoral training. His research includes developing models of microbial metabolism and macromolecule expression and applying them to human health and biotechnology.

© 2020 WILEY-VCH Verlag GmbH & Co. KGaA, Weinheim

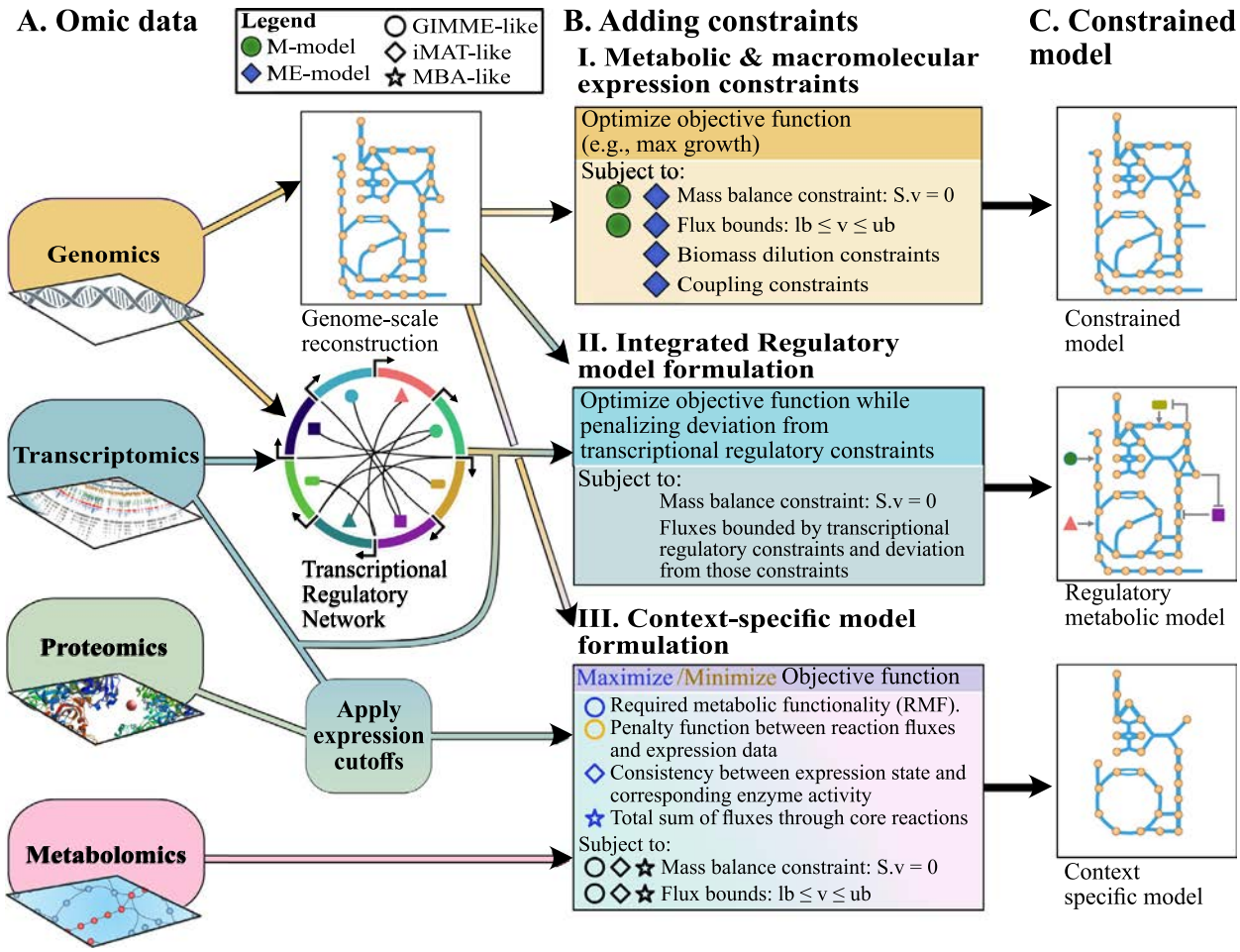


Figure 1. A) Building genome-scale models (GEMs) and integrating them with various omic data types as constraints. Genome-scale models are systems representations of interactions occurring between different molecular components (e.g., metabolites, proteins). These models are built using the annotated genomes of respective organisms. Other omic data can also be used to refine GEMs. B) GEMs need to be constrained to obtain biologically relevant information. General/environmental constraints such as mass balance constraints and flux bounds can be added to GEMs. B-I) For ME-models, additional constraints such as coupling constraints and biomass dilution constraints need to be applied.[9] B-II) Transcriptional regulatory networks (TRNs) combined with transcriptome data can also be used to constrain a GEM to create integrated regulatory genome-scale model. It should be noted that gene expression thresholds are applied in this case as well. B-III) Likewise, multiple omic data (transcriptomic, proteomic, and metabolomic) can be utilized to constrain the model using various approaches. The integration of data leads to new optimization problems (e.g., minimization of inconsistency between fluxes and expression states, maximization of total sum of fluxes through core reactions, etc.) subjected to their own sets of constraints including mass balance and flux constraints. C) The resulting models can be simulated to investigate the genotype-phenotype-environment relationship in the biological system being studied.

the state of the network to identify and predict important features of the organism such as genotype-phenotype-environment relationships, including growth rate, metabolite exchange rates, and gene essentiality.[14] GEMs have also been useful in predicting and analyzing the end result of adaptive evolution.[15] At present, GEMs have been manually reconstructed for at least 183 organisms,[16] and methods are being developed to model microbial communities.[17]

2.1. Using Omic Data to Refine the Genome-Scale Models

Three broad approaches exist to improve GEMs by utilizing omic data. First and foremost, omic data can be directly compared with the flux distribution derived by simulating GEMs to identify any discrepancies between the predicted and experimental data. For instance, one can compare exometabolome with modeling result to determine how accurately the model can predict the secretion profile of an organism under given media condition.

Next, omic data can be used as additional flux constraints on the GEMs to create context- and tissue-specific models[7a,7b,18] (Figure 1). For such purpose, numerous methods have been developed which can be divided into three subcategories—1) use omic data to either indicate presence or absence of enzymes or put relative constraints on enzyme activities (GIMME,[19] GIM3E,[20] REMI[21]), 2) use expression datasets to create context-specific models without prior knowledge of objective function (e.g., iMAT,[22] INIT[23]), and 3) prune non-functional reactions (as extracted from the expression data) to create tissue-specific models (e.g., MBA,[24] mCADRE,[25] CORDA[26]). One inherent

© 2020 WILEY-VCH Verlag GmbH & Co. KGaA, Weinheim

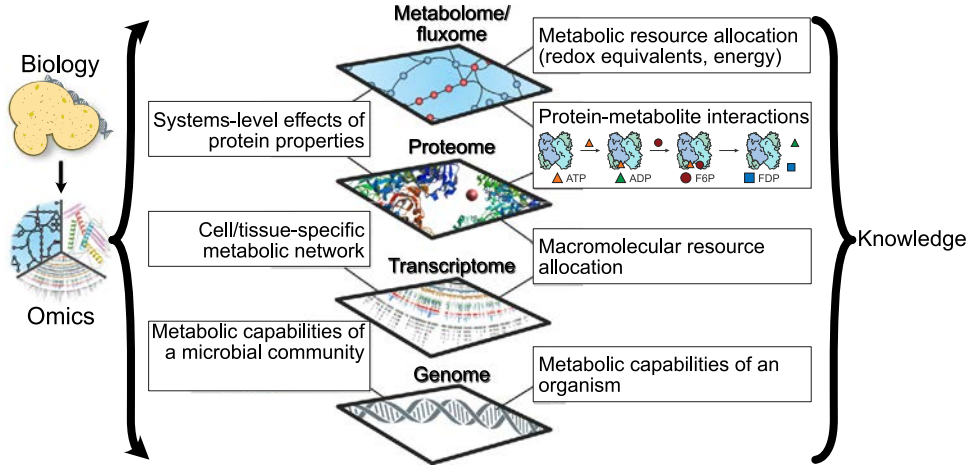


Figure 2. Graphical overview of synthesizing knowledge using omic data and genome-scale models.

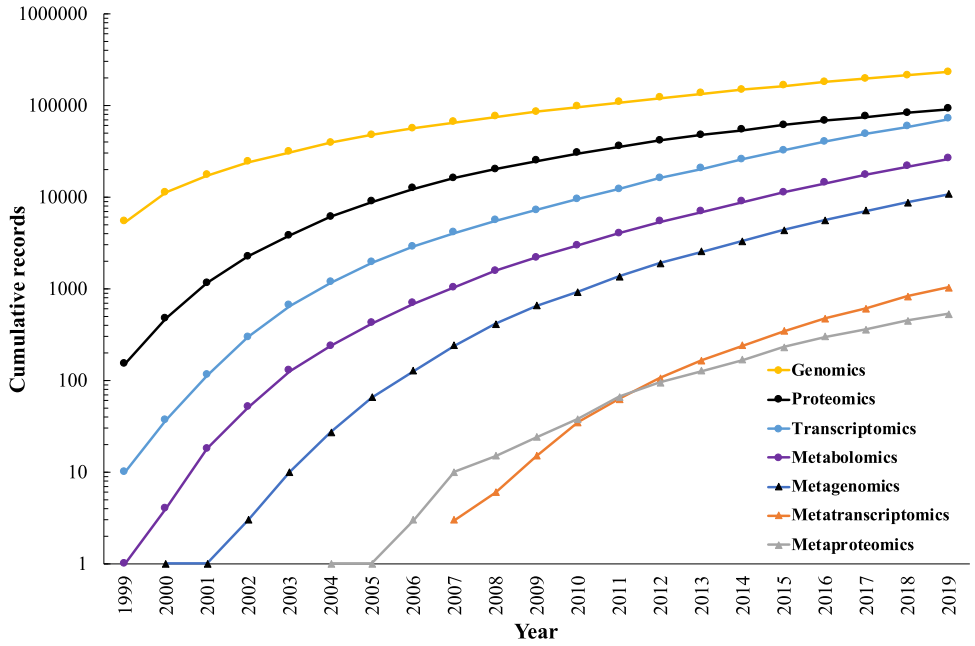


Figure 3. Web of science publications for various omic technologies.

issue with these approaches is that such models can only describe the regulation of metabolism for which the data is integrated into the model.

For predictive models, integrated regulatory metabolic models need to be built. Methods such as probabilistic regulation of metabolism (PROM)[27] exist that can integrate expression datasets (normally transcriptomes) with a transcriptional regulatory network (TRN) and superimpose the information onto a GEM to create an integrated model. In such method, the maximum allowable fluxes of reactions catalyzed by particular enzymes are constrained by the probability of the expression of respective genes given the expression state of a controlling transcription factor (TF). Such TF expression is calculated from the expression datasets across multiple conditions. The rules of TF-target enzyme interaction are governed by the structure of TRN. Since PROM requires a pre-constructed TRN, in recent times, integrated deduced regulation and metabolism (IDREAM) has been developed. IDREAM can create a TRN followed by integration of TRN with GEM and high-throughput expression data (using PROM framework) to create integrated models.[28]

In the next few sections, we will discuss individual omic data types and novel methods that are being introduced to integrate and improve the predictive capabilities of GEMs.

3. Using Transcriptomics to Build Context-Specific Models

Transcriptomics is the quantitative study of all expressed RNA in an organism. Transcriptomic technologies have advanced

© 2020 WILEY-VCH Verlag GmbH & Co. KGaA, Weinheim

from the use of microarrays (i.e., methods involving use of probes to detect specific markers) to exploiting next generation sequencing in the form of RNA-sequencing (RNA-seq). RNA-seq has become more popular because of the ease and availability of advanced sequencing machines.[29] Multiple RNA-Seq algorithms are available with varying degree of accuracy and precision.[30]

Transcriptomics can be used to refine GEMs by integrating the data into model constraints. In this approach, transcriptomics is analyzed by specialized algorithms to create context-specific models by determining the subset of genes that are expressed in a specific cell type, cell line, or tissues.[31] These context-specific models accurately capture tissue specific genotype-phenotype relationships, including gene essentiality.[31] Multiple algorithms are available for building context-specific models, each making different assumptions. Opdam et al.[31] compared six algorithms and showed that the choice of algorithm (and assumptions) had the greatest impact on the model's accuracy of gene essentiality predictions.[31] Once constructed appropriately, context-specific models have yielded important insights into the metabolic mechanisms underlying human diseases.

Gatto et al.[32] constructed context-specific models for 917 primary tumor samples across 13 cancer types, using RNA-Seq and the tINIT algorithm.[33] The models indicated that although cancers can differ in gene expression, their metabolic capabilities are largely similar. Furthermore, cancer metabolic networks overlapped largely with matched normal tissues, suggesting that the metabolic reprogramming—a hallmark of cancer—may reflect cancer cell plasticity to varying conditions. The study also identified a smaller set of 18 metabolic reactions that are present in all the cancers included in the study but without housekeeping functions (such as growth, energy generation, and metabolism) present in normal tissues.[32]

4. Integrating Metabolomics with Genome-Scale Models

Metabolomic technologies quantify the small molecules (molecular mass < 1500 Da) involved in energy metabolism ("metabolites"), representing the most direct way to profile a cell's biochemical activity.[34] Metabolites are involved in the regulation of expression, metabolism, and function of DNA, RNA, and proteins.[35] Research using metabolomic approaches has increased over the past decade (Figure 3) as studies involving identification of disease biomarkers[36] and other important applications[34a,35] have risen. To identify and quantify metabolites, methods such as nuclear magnetic resonance and mass spectrometry (MS) are used. In MS, targeted (hypothesis-driven), untargeted (discovery-based), and recently introduced pseudo-targeted approaches are available.[34b,37] Identification of a metabolite through its spectral signature is crucial to understanding its biological role. However, this approach is limited by the number of available spectra in the available databases.[38] Therefore, multiple methods have been developed for predicting metabolites of which machine learning methods appear promising.[39] Furthermore, a metabolite's function depends on its context-specific interactions with other biological entities. Various computational methods including pathway mapping and network modeling,[40] and GEMs are addressing this need.

Multiple recent studies have used GEMs to integrate metabolomic data. In a recent study, the authors analyzed metabolomic data using GEMs of hepatocytes and identified dopa decarboxylase (DDC) as one of the major cancer-causing enzymes. Following this discovery, authors used the library of integrated network-based cellular signatures program to identify possible drugs that could inhibit expression of DDC.[41] In another study, time-course metabolomic data from human red blood cells (RBCs) stored at different temperatures was analyzed using an RBC GEM. The analysis revealed temperature-dependent metabolic states of RBCs in storage conditions.[42] Recently, a new COBRA method called unsteady-state flux balance analysis (uFBA) has been developed to integrate time-course metabolomic data with GEMs to study the metabolism of RBCs stored in blood bags. The uFBA method predicted that stored RBCs metabolize citric acid cycle intermediates to regenerate key cofactors. These predictions were experimentally confirmed using ^{13}C-metabolic flux analysis.[43]

5. Measuring and Predicting Proteome Allocation Using ME-Models

The proteome represents the functional state of a cell. Proteomics is the quantitative study of all expressed proteins in an organism. Out of several methods used in proteomics, mass spectrometry (MS) is one of the most common platforms. MS can be used in tandem (MS/MS) to provide additional information about a given peptide,[44] and it can be coupled with chromatographic methods to reduce sample complexity and to improve quantification accuracy.[45] Furthermore, it allows multiple properties of the proteome—such as expression, interactions, and modification—to be studied.[46]

Depending on the goal of a study, either targeted or untargeted proteomic approaches can be applied, and sometimes combined for improved analysis.[47] In untargeted proteomics, all possible proteins expressed from a sample are detected and quantified without a priori knowledge. On the other hand, a targeted platform is used to detect specific proteins especially when the desired proteins are known to be present in low abundance a priori.[48] Therefore, targeted approaches are more precise but have lower coverage than untargeted methods.[47] For data acquisition using tandem mass spectrometry (MS/MS), two modes exist—data-dependent acquisition (DDA) mode and data-independent acquisition (DIA) mode. In DDA, a subset of the most abundant precursor ions that exceed a predefined intensity threshold are selected from the first MS scan to the next MS scan. For targeted proteomics, alternative approaches called multiple reaction monitoring and parallel reaction monitoring which select precursor ions for a small set of predetermined peptides for subsequent MS scan are used.[49] DIA, on the other hand, relies on successive isolation and subsequent fragmentation of peptides within a defined mass-to-charge (m/z) window throughout the entire m/z range.[50] SWATH-MS is a DIA method combined with targeted proteomic analysis, and provides good coverage with comparable accuracy and reproducibility.[51] Both DDA and DIA approaches have their advantages and disadvantages

© 2020 WILEY-VCH Verlag GmbH & Co. KGaA, Weinheim

related to sensitivity, dynamic range, accuracy, flexibility, and ease of use.[49,51,52] Finally, for quantification of proteins in a given sample, either relative or absolute quantification methods can be used (please refer to Calderón-Celis et al.[53] for review).

GEMs of metabolism and macromolecular expression (ME-models)[7c,9b,54] directly predict protein expression and proteome allocation (i.e., the relative mass or mole fractions of expressed proteins in a cell). These predictions are validated directly using proteomics, or indirectly using transcriptomics. ME-models predict fluxes for reactions spanning metabolism, transcription, translation, protein modifications, translocation,[55] and protein folding.[56] ME models compute up to 85% protein mass in *Escherichia coli*.[57] ME models are now available for three organisms: *Thermotoga maritima*,[58] *E. coli*,[54a,54b] and *Clostridium ljungdahlii*.[59] Proteomic data has been used to calibrate a ME-model of *E. coli*, decreasing prediction errors of growth rate and metabolic fluxes by 69% and 14%,[54c] and to validate proteomes predicted by a ME-model updated with machine learning-based enzyme turnover rates.

Recently, ME-models were extended to predict cellular response to three stresses: thermal (FoldME),[56] oxidative (OxidizeME),[57] and acid (AcidifyME).[60] By mechanistically reconstructing key molecular responses to each stress, the models successfully predicted phenotypic response (change in growth rate) and differential expression in various growth conditions (i.e., media, supplements, etc.) and stress intensities. These models have been used to explain biological mechanisms by interpreting omics.

A ME-model accounting for the proteostasis network, FoldME,[56] was used to study the global effects caused by the protein stability of dihydrofolate reductase. The experimental (transcriptomic data) and predicted data were quantitatively correlated for the major clusters of orthologous groups. Further analysis using the ME-model suggested that protein destabilizing mutations can lead to chaperone-mediated strategy of systems-level proteome reallocation including downregulation of coenzyme biosynthetic pathways.[56]

In another study, a ME-model accounting for the effect of reactive oxygen species (ROS) on metalloproteins, OxidizeME,[57] was used to explain why the growth rate of *E. coli* was limited when using naphthoquinone (NQ) instead of ubiquinone (UQ) in the electron transport system (ETS).[15b] NQ autoxidizes more readily than ubiquinone (UQ), generating superoxide in the periplasm. OxidizeME showed that the metabolic and protein expression cost of detoxifying periplasmic superoxide strongly decreased growth rate. The reduced ETS efficiency due to electron leakage from NQ toward superoxide generation decreased growth rate further; however, the cost of detoxification was demonstrated to be the primary reason for reduced growth rate.

A modeling approach called metabolism and macromolecular mechanisms (MM) was developed recently for human RBCs.[61] Unlike ME-models, the reactions related to transcription and translation are not present in RBC-MM. In RBC-MM, proteomic data were used to constrain enzyme abundances, which constrained the reaction fluxes. This model simulates metabolism, hemoglobin binding, and the formation and detoxification of ROS.[61]

6. Integrating Multi-Omic Data with Genome-Scale Models

Studies are now combining multi-omic platforms with GEMs to study complex interactions that occur at the molecular level within organisms. In a recent study, metabolomics combined with proteomics was integrated in GEM of *E. coli* to identify pathway engineering strategies to improve biofuel production.[62] A ME-model was recently used to analyze multi-omic (genomic, transcriptomic, ribosomal profiling, proteomic, and fluxomic) data to discover two biological regularities associated with enzyme turnover rates and translation in *E. coli*.[7c] Likewise, a laboratory rat GEM was integrated with transcriptomic, metabolomic, and fluxomic data to identify plasma metabolites that are associated with acetaminophen-induced liver injury.[63]

7. Using Meta-Omic Data to Build and Refine Microbial Community Models

Meta-omic technologies (metagenomics, metatranscriptomics, and metaproteomics) measure the molecular makeup of an entire sample, which can include unculturable organisms. This area has grown steadily since the mid-2000s (Figure 3). Metagenomics provides tools to analyze genomic DNA to determine the abundance of all detectable organisms present in a sample.[64] In metatranscriptomics, RNA is sequenced and analyzed to reveal the functionally active members in a microbial community.[65] Metaproteomics provides platform for the analysis of proteins expressed by the organisms in a given sample.[66] Meta-omic approaches have been applied to environmental (including marine and soil communities),[67] waste management,[68] and clinical samples.[69] These platforms are crucial for generating and analyzing data to understand the dynamics within a community and to study biological systems in nature.

Multiple recent studies have integrated meta-omic data with GEMs to study microbial communities in finer detail. Computational tools have been developed to automatically reconstruct microbial community models using meta-omics. For example, human gut microbiome models can be efficiently reconstructed using metagenomics through the microbiome modeling toolbox.[17b] Another method, MICOM (MIcrobial COMmunity), was developed to build personalized metabolic models for the human gut microbiomes of 186 people using their individual metagenomic samples. The models revealed that changes in microbiome composition and diet have highly personalized effects.[70] Meta-omics in combination with GEMs have also been applied to environmental samples. For example, meta-genomics and meta-proteomics were used to build GEMs of two microbial communities in polyaromatic hydrocarbon contaminated soil.[71]

8. Models Provide a Systems Context for Protein Structures

Structural genomics aims to determine all 3D structures of proteins expressed from an organism's genome, and this field has yielded over 150 000 structures in the Protein Data Bank (PDB).[72] Recent studies have shown that this increasingly

© 2020 WILEY-VCH Verlag GmbH & Co. KGaA, Weinheim

abundant data types can be integrated into GEMs. This integration has expanded the scope of mechanisms and biological questions addressable by computational systems biology. In particular, all three of the recent ME-models that account for stress functions in *E. coli* use 3D structures to perform key computations.[56,57,60]

In the FoldME[56] model that predicts *E. coli*'s thermal stress response, a key feature is to predict protein thermostability. This task required fitting thermodynamic contributions from each type of amino acid using 3D structures of *E. coli* proteins available from PDB. The OxidizeME[57] model that predicts *E. coli*'s response to oxidative stress required a method to predict metal cofactor damage for approximately 43 metalloproteins. Since experimental measurements for every metalloprotein were not available, the probability of metal cofactor damage was computed using protein 3D structural properties. A key feature of the AcidifyME[60] model is to compute (periplasmic) protein stability as a function of pH. This task required applying the multi-conformation continuum electrostatics method to 3D protein structures. In all of the studies above, the availability of high-quality 3D structures was necessary to predict systems-level response to macromolecule properties that change in response to physical and chemical stimuli.

9. Using Machine Learning to Improve Structure-Function Predictions and to Enhance the Predictive Accuracy of GEMs

One gap between structural proteome and cell phenotype is that functional alterations due to the variations in protein structure are still expensive or difficult to predict. More efficient structure-function prediction models, which predict functions of a protein based on its structure, would enable routine computation of mutation effects on function and properties (e.g., solubility, stability, activity, etc.) of proteins in the whole-cell context.

Machine learning (ML) has been successfully used in the computer vision and the natural language processing field. Recently, there has been a significant interest in applying machine ML in the research of protein structure-function prediction.[73] Motivated by the expensive and time-consuming experimental protein functions annotations and aiming to improve the traditional computational approaches, a variety of machine learning methods have been developed to predict protein functions.[73e] The traditional approaches (relying on sequence similarity) might not produce accurate predictions because some proteins might have similar function even with low sequence similarity.[73e] ML methods have improved the prediction performance of such in silico methods that make prediction solely based on the amino acid sequence similarity between proteins by focusing on protein structure itself.[74] Other ML methods focus on predicting the properties of proteins based on more comprehensive features, like protein 3D structure and biological process information.[73a] Current state-of-the-art ML methods for protein structure-function prediction formulate the problem as a supervised classification task. The use of additional information such as the hierarchical structure of gene ontology and protein-protein interactions have been proved helpful to improve prediction capability.[73a,73c–e] However, developing these ML methods is challenging because real biological data tend to be incomplete, noisy, biased and multi-modal.[73e] Nonetheless, continued development of ML techniques to address these data limitations and, more directly, increase the availability of data for ML analysis will make ML a promising approach for predicting protein function from structures.[73a,73c]

ML has already been used to improve GEM predictions, by predicting catalytic turnover rates in *E. coli* from a diverse set of features.[75] These features included network context, protein structure, biochemistry, and assay conditions. The study identified important features for turnover rate prediction: structural (active site depth, active site solvent accessibility, active site exposure), network context (predicted reaction fluxes, reflecting evolutionary selection pressure on turnover rate), and the number of reactions an enzyme promiscuously catalyzes. Using these ML-predicted turnover rates improved the accuracy of GEM predictions: by 20–34%.[75]

10. Using GEMs to Delineate the Network-Level Effects of Post-Translational Modifications

Proteoforms are proteins expressed from one gene but altered through PTMs and that may possess different functions from each other.[76] There are more than 200 types of PTMs recorded in various databases.[77] For proteoform detection, top-down MS-based proteomic approaches which require intact protein separation through methods including serial size exclusion chromatography[78] and capillary zone electrophoresis[79] have been considered. To have a comprehensive understanding of a biological system, knowledge of global effects of PTMs is essential.

There have been some modeling efforts that have examined the network-level effect of PTMs. Brunk et al.[80] identified important branch point enzymes in the metabolic network using a GEM. The authors then integrated the GEM predictions, multiplex automated genome editing, and molecular dynamic simulations to elucidate the mechanisms by which PTMs can affect the protein activity and overall cellular fitness. The authors demonstrated that PTMs can modulate protein interactions (in serine hydroxymethyltransferase), impact substrate binding (transaldolase) and regulate catalytic residues (enolase). These mechanistic insights elucidated how specific PTMs regulate cellular function at multiple biological scales, from individual enzymes to pathway usage and, ultimately, cellular phenotypes.

11. Conclusions and Future Perspectives

Advancements in omic technologies continue to extend our ability to read out the complete molecular makeup of a cell under various conditions of relevance to health, engineering, and knowledge expansion. Each omic technology measures a specific molecular category (RNA, protein, metabolite, etc.) as the cell is "taken apart" and analyzed. Computational systems biology provides a platform to "put together" these disparate data sets and to synthesize knowledge. Literature indicates that this pipeline of measure–model–synthesize is yielding knowledge with consistency and improving accuracy. However, we are also gaining

© 2020 WILEY-VCH Verlag GmbH & Co. KGaA, Weinheim

more appreciation of the complexities associated with integrating multi-omics. Specifically, as the types of omic data types increase, so do the number of interactions we must consider across the different biological layers. Systems biology models, including the GEMs that we focused on here, help to navigate complexity by consolidating existing knowledge to provide context for data. Not all omic types can be interpreted with equal fidelity and resolution, however. Hence, mechanism-elucidating models are used routinely to study metabolic processes using multi-omics, while gene regulation, epigenetics, and signaling require more data-driven or statistical modeling approaches to study system-level phenomena. Furthermore, while structural proteomics has become invaluable for genome-scale modeling in recent years, we require more efficient algorithms to compute the functional effects of genetic and structural perturbations. Recent advances in machine learning in this area show promise. With better predictions and availability of more data, the predictive power of GEMs will continue to rise making GEMs incredibly powerful tools in decoding biological systems.

Acknowledgements

This research was supported by Queen's University (S.D., H.X., and L.Y.), the Institute for Systems Biology's Translational Research Fellows Program (J.T.Y.), and the National Institute of General Medical Sciences of the National Institutes of Health Grant R01GM057089 (B.O.P.).

Conflict of Interest

The authors declare no conflict of interest.

Keywords

computational model, genome-scale model, genomics, machine learning, systems biology

Received: March 8, 2020
Revised: June 13, 2020
Published online: July 12, 2020

[1] T. Y. Low, M. A. Mohtar, M. Y. Ang, R. Jamal, *Proteomics* **2019**, *19*, 1800235.
[2] a) U. Omasits, A. R. Varadarajan, M. Schmid, S. Goetze, D. Melidis, M. Bourqui, O. Nikolayeva, M. Québatte, A. Patrignani, C. Dehio, *Genome Res.* **2017**, *27*, 2083; b) Y. Mao, X. Yang, Y. Liu, Y. Yan, Z. Du, Y. Han, Y. Song, L. Zhou, Y. Cui, R. Yang, *Am. J. Trop. Med. Hyg.* **2016**, *95*, 562; c) A. McAfee, B. A. Harpur, S. Michaud, R. C. Beavis, C. F. Kent, A. Zayed, L. J. Foster, *J. Proteome Res.* **2016**, *15*, 411; d) B. Chapman, M. Bellgard, *Proteomics* **2017**, *17*, 1700197.
[3] a) S. S. Manda, R. S. Nirujogi, S. M. Pinto, M.-S. Kim, K. K. Datta, R. Sirdeshmukh, T. K. Prasad, V. Thongboonkerd, A. Pandey, H. Gowda, *J. Proteome Res.* **2014**, *13*, 3166; b) S. M. Pinto, S. S. Manda, M.-S. Kim, K. Taylor, L. D. N. Selvan, L. Balakrishnan, T. Subbannayya, F. Yan, T. K. Prasad, H. Gowda, *J. Proteome Res.* **2014**, *13*, 2749.
[4] Y. Kori, S. Sidoli, Z.-F. Yuan, P. J. Lund, X. Zhao, B. A. Garcia, *Sci. Rep.* **2017**, *7*, 10296.
[5] B. J. Bennett, T. Q. de A. Vallim, Z. Wang, D. M. Shih, Y. Meng, J. Gregory, H. Allayee, R. Lee, M. Graham, R. Crooke, *Cell Metab.* **2013**, *17*, 49.
[6] B. L. D. Kaiser, J. Li, J. A. Sanford, Y.-M. Kim, S. R. Kronewitter, M. B. Jones, C. T. Peterson, S. N. Peterson, B. C. Frank, S. O. Purvine, *PLoS One* **2013**, *8*, e67155.
[7] a) R. P. Vivek-Ananth, A. Samal, *Biosystems* **2016**, *147*, 1; b) J. Cho, C. Gu, T. Han, J. Ryu, S. Lee, *Curr. Opin. Syst. Biol.* **2019**, *15*, 1; c) A. Ebrahim, E. Brunk, J. Tan, E. J. O'Brien, D. Kim, R. Szubin, J. A. Lerman, A. Lechner, A. Sastry, A. Bordbar, A. M. Feist, B. O. Palsson, *Nat. Commun.* **2016**, *7*, 13091.
[8] a) A. Ebrahim, J. A. Lerman, B. O. Palsson, D. R. Hyduke, *BMC Syst. Biol.* **2013**, *7*, 74; b) A. Bordbar, J. M. Monk, Z. A. King, B. O. Palsson, *Nat. Rev. Genet.* **2014**, *15*, 107; . c) B. Palsson, *Systems Biology*, Cambridge University Press, New York **2015**; d) L. Heirendt, S. Arreckx, T. Pfau, S. N. Mendoza, A. Richelle, A. Heinken, H. S. Haraldsdottir, J. Wachowiak, S. M. Keating, V. Vlasov, S. Magnusdottir, C. Y. Ng, G. Preciat, A. Zagare, S. H. J. Chan, M. K. Aurich, C. M. Clancy, J. Modamio, J. T. Sauls, A. Noronha, A. Bordbar, B. Cousins, D. C. El Assal, L. V. Valcarcel, I. Apaolaza, S. Ghaderi, M. Ahookhosh, M. Ben Guebila, A. Kostromins, N. Sompairac, et al.,*Nat. Protoc.* **2019**, *14*, 639.
[9] a) C. J. Lloyd, A. Ebrahim, L. Yang, Z. A. King, E. Catoiu, E. J. O'Brien, J. K. Liu, B. O. Palsson, *PLoS Comput. Biol.* **2018**, *14*, e1006302; b) L. Yang, J. T. Yurkovich, Z. A. King, B. O. Palsson, *Curr. Opin. Microbiol.* **2018**, *45*, 8.
[10] S. Anderson, A. T. Bankier, B. G. Barrell, M. H. de Bruijn, A. R. Coulson, J. Drouin, I. C. Eperon, D. P. Nierlich, B. A. Roe, F. Sanger, *Nature* **1981**, *290*, 457.
[11] a) J. Parkhill, B. Wren, K. Mungall, J. Ketley, C. Churcher, D. Basham, T. Chillingworth, R. Davies, T. Feltwell, S. Holroyd, *Nature* **2000**, *403*, 665; b) S. Patrick, J. Parkhill, L. J. McCoy, N. Lennard, M. J. Larkin, M. Collins, M. Sczaniecka, G. Blakely, *Microbiology* **2003**, *149*, 915.
[12] a) M. L. Metzker, *Nat. Rev. Genet.* **2010**, *11*, 31; b) N. J. Loman, M. J. Pallen, *Nat. Rev. Microbiol.* **2015**, *13*, 787; c) J. Besser, H. A. Carleton, P. Gerner-Smidt, R. L. Lindsey, E. Trees, *Clin. Microbiol. Infect.* **2018**, *24*, 335.
[13] S. Ghatak, Z. A. King, A. Sastry, B. O. Palsson, *Nucleic Acids Res.* **2019**, *47*, 2446.
[14] N. D. Price, J. L. Reed, B. O. Palsson, *Nat. Rev. Microbiol.* **2004**, *2*, 886.
[15] a) W. R. Harcombe, N. F. Delaney, N. Leiby, N. Klitgord, C. J. Marx, *PLoS Comput. Biol.* **2013**, *9*, e1003091; b) A. Anand, K. Chen, L. Yang, A. V. Sastry, C. A. Olson, S. Poudel, Y. Seif, Y. Hefner, P. V. Phaneuf, S. Xu, R. Szubin, A. M. Feist, B. O. Palsson, *Proc. Natl. Acad. Sci. USA* **2019**, *116*, 25287.
[16] C. Gu, G. B. Kim, W. J. Kim, H. U. Kim, S. Y. Lee, *Genome Biol.* **2019**, *20*, 121.
[17] a) D. Machado, S. Andrejev, M. Tramontano, K. R. Patil, *Nucleic Acids Res.* **2018**, *46*, 7542; b) F. Baldini, A. Heinken, L. Heirendt, S. Magnusdottir, R. M. T. Fleming, I. Thiele, *Bioinformatics* **2019**, *35*, 2332.
[18] S. Opdam, A. Richelle, B. Kellman, S. Li, D. C. Zielinski, N. E. Lewis, *Cell Syst.* **2017**, *4*, 318.
[19] S. A. Becker, B. O. Palsson, *PLoS Comput. Biol.* **2008**, *4*, e1000082.
[20] B. J. Schmidt, A. Ebrahim, T. O. Metz, J. N. Adkins, B. O. Palsson, D. R. Hyduke, *Bioinformatics* **2013**, *29*, 2900.
[21] V. Pandey, N. Hadadi, V. Hatzimanikatis, *PLoS Comput. Biol.* **2019**, *15*, e1007036.
[22] H. Zur, E. Ruppin, T. Shlomi, *Bioinformatics* **2010**, *26*, 3140.
[23] R. Agren, S. Bordel, A. Mardinoglu, N. Pornputtapong, I. Nookaew, J. Nielsen, *PLoS Comput. Biol.* **2012**, *8*, e1002518.
[24] L. Jerby, T. Shlomi, E. Ruppin, *Mol. Syst. Biol.* **2010**, *6*, 401.
[25] Y. Wang, J. A. Eddy, N. D. Price, *BMC Syst. Biol.* **2012**, *6*, 153.
[26] A. Schultz, A. A. Qutub, *PLoS Comput. Biol.* **2016**, *12*, e1004808.

© 2020 WILEY-VCH Verlag GmbH & Co. KGaA, Weinheim

[27] S. Chandrasekaran, N. D. Price, *Proc. Natl. Acad. Sci. USA* **2010**, *107*, 17845.
[28] Z. Wang, S. A. Danziger, B. D. Heavner, S. Ma, J. J. Smith, S. Li, T. Herricks, E. Simeonidis, N. S. Baliga, J. D. Aitchison, N. D. Price, *PLoS Comput. Biol.* **2017**, *13*, e1005489.
[29] A. Byrne, C. Cole, R. Volden, C. Vollmers, *Philos. Trans. R. Soc. B* **2019**, *374*, 20190097.
[30] a) G. Baruzzo, K. E. Hayer, E. J. Kim, B. Di Camillo, G. A. FitzGerald, G. R. Grant, *Nat. Methods* **2017**, *14*, 135; b) K. V. den Berge, K. M. Hembach, C. Soneson, S. Tiberi, L. Clement, M. I. Love, R. Patro, M. D. Robinson, *Annu. Rev. Biomed. Data Sci.* **2019**, *2*, 139.
[31] S. R. Opdam Anne, B. Kellman, S. Li, D. C. Zielinski, N. E. Lewis, *Cell Syst.* **2017**, *4*, 318.
[32] F. Gatto, R. Ferreira, J. Nielsen, *Metab. Eng.* **2020**, *57*, 51.
[33] R. Agren, A. Mardinoglu, A. Asplund, C. Kampf, M. Uhlen, J. Nielsen, *Mol.lar Syst. Biol.* **2014**, *10*, 721.
[34] a) D. S. Wishart, *Physiol. Rev.* **2019**, *99*, 1819; b) M. Yan, G. Xu, *Anal. Chim. Acta* **2018**, *1037*, 41.
[35] a) M. M. Rinschen, J. Ivanisevic, M. Giera, G. Siuzdak, *Nat. Rev. Mol. Cell Biol.* **2019**, *20*, 353; b) J. Simithy, S. Sidoli, B. A. Garcia, *Proteomics* **2018**, *18*, 1700309.
[36] a) Á. López-López, Á. López-Gonzálvez, T. C. Barker-Tejeda, C. Barbas, *Expert Rev. Mol. Diagn.* **2018**, *18*, 557; b) O. D. Rangel-Huerta, B. Pastor-Villaescusa, A. Gil, *Metabolomics* **2019**, *15*, 93.
[37] F. Fenaille, P. B. Saint-Hilaire, K. Rousseau, C. Junot, *J. Chromatogr. A* **2017**, *1526*, 1.
[38] a) F. Allen, R. Greiner, D. Wishart, *Metabolomics* **2015**, *11*, 98; b) R. R. da Silva, P. C. Dorrestein, R. A. Quinn, *Proc. Natl. Acad. Sci. USA* **2015**, *112*, 12549.
[39] a) K. Dührkop, H. Shen, M. Meusel, J. Rousu, S. Böcker, *Proc. Natl. Acad. Sci. USA* **2015**, *112*, 12580; b) H. Ji, Y. Xu, H. Lu, Z. Zhang, *Anal. Chem.* **2019**, *91*, 5629; c) D. H. Nguyen, C. H. Nguyen, H. Mamitsuka, *Briefings Bioinf.* **2018**, *20*, 2028.
[40] C. H. Johnson, J. Ivanisevic, G. Siuzdak, *Nat. Rev. Mol. Cell Biol.* **2016**, *17*, 451.
[41] H.-Q. Wu, M.-L. Cheng, J.-M. Lai, H.-H. Wu, M.-C. Chen, W.-H. Liu, W.-H. Wu, P. M.-H. Chang, C.-Y. F. Huang, A.-P. Tsou, *PLoS Comput. Biol.* **2017**, *13*, e1005618.
[42] J. T. Yurkovich, D. C. Zielinski, L. Yang, G. Paglia, O. Rolfsson, Ó. E. Sigurjónsson, J. T. Broddrick, A. Bordbar, K. Wichuk, S. Brynjólfsson, S. Palsson, S. Gudmundsson, B. O. Palsson, *J. Biol. Chem.* **2017**, *292*, 19556.
[43] A. Bordbar, J. T. Yurkovich, G. Paglia, O. Rolfsson, Ó. E. Sigurjónsson, B. O. Palsson, *Sci. Rep.* **2017**, *7*, 46249.
[44] A. El-Aneed, A. Cohen, J. Banoub, *Appl. Spectrosc. Rev.* **2009**, *44*, 210.
[45] M. Bantscheff, S. Lemeer, M. M. Savitski, B. Kuster, *Anal. Bioanal. Chem.* **2012**, *404*, 939.
[46] X. Han, A. Aslanian, J. R. Yates III, *Curr. Opin. Chem. Biol.* **2008**, *12*, 483.
[47] C. A. Sobsey, S. Ibrahim, V. R. Richard, V. Gaspar, G. Mitsa, V. Lacasse, R. P. Zahedi, G. Batist, C. H. Borchers, *Proteomics* **2019**, *20*, e1900029.
[48] a) E. Borras, E. Sabido, *Proteomics* **2017**, *17*, 1700180; b) S. Saleh, A. Staes, S. Deborggraeve, K. Gevaert, *Proteomics* **2019**, *19*, 1800435.
[49] A. Hu, W. S. Noble, A. Wolf-Yadlin, *F1000Res* **2016**, *5*, 419.
[50] J. D. Venable, M.-Q. Dong, J. Wohlschlegel, A. Dillin, J. R. Yates III, *Nat. Methods* **2004**, *1*, 39.
[51] L. C. Gillet, P. Navarro, S. Tate, H. Röst, N. Selevsek, L. Reiter, R. Bonner, R. Aebersold, *Mol. Cell. Proteomics* **2012**, *11*, 016717.
[52] Y. Kawashima, E. Watanabe, T. Umeyama, D. Nakajima, M. Hattori, K. Honda, O. Ohara, *Int. J. Mol. Sci.* **2019**, *20*, 5932.
[53] a) F. Calderón-Celis, J. R. Encinar, A. Sanz-Medel, *Mass Spectrom. Rev.* **2018**, *37*, 715; b) J. A. Ankney, A. Muneer, X. Chen, *Annu. Rev. Anal. Chem.* **2018**, *11*, 49.
[54] a) I. Thiele, R. M. T. Fleming, R. Que, A. Bordbar, D. Diep, B. O. Palsson, *PLoS One* **2012**, *7*, e45635; b) E. J. O'Brien, J. A. Lerman, R. L. Chang, D. R. Hyduke, B. Ø. Palsson, *Mol. Syst. Biol.* **2013**, *9*, 693; c) L. Yang, J. T. Yurkovich, C. J. Lloyd, A. Ebrahim, M. A. Saunders, B. O. Palsson, *Sci. Rep.* **2016**, *6*, 36734.
[55] J. K. Liu, E. J. O'Brien, J. A. Lerman, K. Zengler, B. O. Palsson, A. M. Feist, *BMC Syst. Biol.* **2014**, *8*, 110.
[56] K. Chen, Y. Gao, N. Mih, E. J. O'Brien, L. Yang, B. O. Palsson, *Proc. Natl. Acad. Sci. USA* **2017**, *114*, 11548.
[57] L. Yang, N. Mih, A. Anand, J. H. Park, J. Tan, J. T. Yurkovich, J. M. Monk, C. J. Lloyd, T. E. Sandberg, S. W. Seo, D. Kim, A. V. Sastry, P. Phaneuf, Y. Gao, J. T. Broddrick, K. Chen, D. Heckmann, R. Szubin, Y. Hefner, A. M. Feist, B. O. Palsson, *Proc. Natl. Acad. Sci. USA* **2019**, *116*, 14368.
[58] J. A. Lerman, D. R. Hyduke, H. Latif, V. A. Portnoy, N. E. Lewis, J. D. Orth, A. C. Schrimpe-Rutledge, R. D. Smith, J. N. Adkins, K. Zengler, B. O. Palsson, *Nat. Commun.* **2012**, *3*, 929.
[59] J. K. Liu, C. Lloyd, M. M. Al-Bassam, A. Ebrahim, J. N. Kim, C. Olson, A. Aksenov, P. Dorrestein, K. Zengler, *PLoS Comput. Biol.* **2019**, *15*, e1006848.
[60] B. Du, L. Yang, C. J. Lloyd, X. Fang, B. O. Palsson, *PLoS Comput. Biol.* **2019**, *15*, e1007525.
[61] J. T. Yurkovich, L. Yang, B. O. Palsson, *bioRxiv* **2019**, https://doi.org/10.1101/797258.
[62] E. Brunk, K. W. George, J. Alonso-Gutierrez, M. Thompson, E. Baidoo, G. Wang, C. J. Petzold, D. McCloskey, J. Monk, L. Yang, *Cell Syst.* **2016**, *2*, 335.
[63] V. R. Pannala, M. L. Wall, S. K. Estes, I. Trenary, T. P. O'Brien, R. L. Printz, K. C. Vinnakota, J. Reifman, M. Shiota, J. D. Young, A. Wallqvist, *Sci. Rep.* **2018**, *8*, 11678.
[64] H. Y. Simon, K. J. Siddle, D. J. Park, P. C. Sabeti, *Cell* **2019**, *178*, 779.
[65] M. Shakya, C.-C. Lo, P. S. Chain, *Front. Genet.* **2019**, *10*, 904.
[66] a) P. Wilmes, A. Heintz-Buschart, P. L. Bond, *Proteomics* **2015**, *15*, 3409; b) N. I. Isaac, D. Philippe, A. Nicholas, D. Raoult, C. Eric, *Clin. Mass Spectrom.* **2019**, *14*, 18.
[67] a) E. A. Eloe-Fadrosh, N. N. Ivanova, T. Woyke, N. C. Kyrpides, *Nat. Microbiol.* **2016**, *1*, 15032; b) I. J. Miller, T. R. Weyna, S. S. Fong, G. E. Lim-Fong, J. C. Kwan, *Sci. Rep.* **2016**, *6*, 34362; c) D. E. Holmes, P. M. Shrestha, D. J. Walker, Y. Dang, K. P. Nevin, T. L. Woodard, D. R. Lovley, *Appl. Environ. Microbiol.* **2017**, *83*, e00223; d) D. S. Jones, B. E. Flood, J. V. Bailey, *ISME J.* **2016**, *10*, 1015; e) S. Kleindienst, F.-A. Herbst, M. Stagars, F. V. Netzer, M. V. Bergen, J. Seifert, J. Peplies, R. Amann, F. Musat, T. Lueders, *ISME J.* **2014**, *8*, 2029; f) H. Teeling, B. M. Fuchs, D. Becher, C. Klockow, A. Gardebrecht, C. M. Bennke, M. Kassabgy, S. Huang, A. J. Mann, J. Waldmann, *Science* **2012**, *336*, 608.
[68] a) E. Bouhajja, S. N. Agathos, I. F. George, *Biotechnol. Adv.* **2016**, *34*, 1413; b) T. P. Delforno, T. Z. Macedo, C. Midoux, G. V. Lacerda Jr., O. Rué, M. Mariadassou, V. Loux, M. B. Varesche, T. Bouchez, A. Bize, *Sci. Total Environ.* **2019**, *649*, 482; c) P. Wilmes, M. Wexler, P. L. Bond, *PLoS One* **2008**, *3*, e1778; d) L. H. Hagen, J. A. Frank, M. Zamanzadeh, V. G. Eijsink, P. B. Pope, S. J. Horn, M. Ø. Arntzen, *Appl. Environ. Microbiol.* **2017**, *83*, e01955.
[69] a) V. Pascal, M. Pozuelo, N. Borruel, F. Casellas, D. Campos, A. Santiago, X. Martinez, E. Varela, G. Sarrabayrouse, K. Machiels, *Gut* **2017**, *66*, 813; b) M. Schirmer, E. A. Franzosa, J. Lloyd-Price, L. J. McIver, R. Schwager, T. W. Poon, A. N. Ananthakrishnan, E. Andrews, G. Barron, K. Lake, *Nat. Microbiol.* **2018**, *3*, 337; c) L. A. Lai, Z. Tong, R. Chen, S. Pan, in *Functional Proteomics* (Eds: X. Wang, M. Kuruc), Springer, Berlin **2019**, p. 123; d) E. Pinto, M. Anselmo, M. Calha, A. Bottrill, I. Duarte, P. W. Andrew, M. L. Faleiro, *Microbiology* **2017**, *163*, 161.

© 2020 WILEY-VCH Verlag GmbH & Co. KGaA, Weinheim

[70] C. Diener, S. M. Gibbons, O. Resendis-Antonio, *mSystems* **2020**, *5*, e00606.
[71] L. Tobalina, R. Bargiela, J. Pey, F.-A. Herbst, I. Lores, D. Rojo, C. Barbas, A. I. Peláez, J. Sánchez, M. von Bergen, *Bioinformatics* **2015**, *31*, 1771.
[72] N. Mih, B. O. Palsson, *Mol. Syst. Biol.* **2019**, *15*, e8601.
[73] a) R. Townshend, R. Bedi, P. Suriana, R. Dror, *NeurIPS* **2019**, *32*, 15642; b) A. Fout, J. Byrd, B. Shariat, A. Ben-Hur, *NeurIPS* **2017**, *30*, 6530; c) V. Gligorijević, M. Barot, R. Bonneau, *Bioinformatics* **2018**, *34*, 3873; d) I. A. Kovács, K. Luck, K. Spirohn, Y. Wang, C. Pollis, S. Schlabach, W. Bian, D.-K. Kim, N. Kishore, T. Hao, M. A. Calderwood, M. Vidal, A.-L. Barabási, *Nat. Commun.* **2019**, *10*, 1240; e) F. Zhang, H. Song, M. Zeng, Y. Li, L. Kurgan, M. Li, *Proteomics* **2019**, *19*, 1900019; f) A. W. Senior, R. Evans, J. Jumper, J. Kirkpatrick, L. Sifre, T. Green, C. Qin, A. Žídek, A. W. Nelson, A. Bridgland, H. Penedones, S. Petersen, K. Simonyan, S. Crossan, P. Kohli, D. T. Jones, D. Silver, K. Kavukcuoglu, D. Hassabis, *Nature* **2020**, *577*, 706.
[74] D. Cozzetto, F. Minneci, H. Currant, D. T. Jones, *Sci. Rep.* **2016**, *6*, 31865.
[75] D. Heckmann, C. J. Lloyd, N. Mih, Y. Ha, D. C. Zielinski, Z. B. Haiman, A. A. Desouki, M. J. Lercher, B. O. Palsson, *Nat. Commun.* **2018**, *9*, 5252.
[76] D. Kumar, G. Bansal, A. Narang, T. Basak, T. Abbas, D. Dash, *Proteomics* **2016**, *16*, 2533.
[77] M. S. Kim, J. Zhong, A. Pandey, *Proteomics* **2016**, *16*, 700.
[78] X. Chen, Y. Ge, *Proteomics* **2013**, *13*, 2563.
[79] a) X. Shen, Z. Yang, E. N. McCool, R. A. Lubeckyj, D. Chen, L. Sun, *TrAC Trends Anal. Chem.* **2019**, *120*, 115644; b) E. N. McCool, R. A. Lubeckyj, X. Shen, D. Chen, Q. Kou, X. Liu, L. Sun, *Anal. Chem.* **2018**, *90*, 5529.
[80] E. Brunk, R. L. Chang, J. Xia, H. Hefzi, J. T. Yurkovich, D. Kim, E. Buckmiller, H. H. Wang, B.-K. Cho, C. Yang, B. O. Palsson, G. M. Church, N. E. Lewis, *Proc. Natl. Acad. Sci. USA* **2018**, *115*, 11096.

© 2020 WILEY-VCH Verlag GmbH & Co. KGaA, Weinheim

Check for updates

Lineage tracing meets single-cell omics: opportunities and challenges

Daniel E. Wagner[1,2]✉ *and Allon M. Klein*[1]✉

Abstract | A fundamental goal of developmental and stem cell biology is to map the developmental history (ontogeny) of differentiated cell types. Recent advances in high-throughput single-cell sequencing technologies have enabled the construction of comprehensive transcriptional atlases of adult tissues and of developing embryos from measurements of up to millions of individual cells. Parallel advances in sequencing-based lineage-tracing methods now facilitate the mapping of clonal relationships onto these landscapes and enable detailed comparisons between molecular and mitotic histories. Here we review recent progress and challenges, as well as the opportunities that emerge when these two complementary representations of cellular history are synthesized into integrated models of cell differentiation.

Cell differentiation
The process by which uncommitted progenitor cells are specified and transform into functional (and typically postmitotic) cells that carry out the specialized tasks of a particular tissue or organ.

Landscape
An informal term for a state manifold, typically used in developmental biology to represent the ensemble of cell states during their differentiation.

[1]*Department of Systems Biology, Harvard Medical School, Boston, MA, USA.*

[2]*Present Address: Department of Obstetrics, Gynecology and Reproductive Science, Center for Reproductive Sciences, Eli and Edythe Broad Center for Regeneration Medicine and Stem Cell Research, University of California San Francisco, San Francisco, CA, USA.*

✉*e-mail: daniel.wagner@ucsf.edu; allon_klein@hms.harvard.edu*

https://doi.org/10.1038/s41576-020-0223-2

Cellular differentiation in composition, organization and function represents one of the major innovations of multicellular life. Determining the molecular mechanisms that govern how cells differentiate in their state is thus a long-standing focus in stem cell and developmental biology[1]. A comprehensive record of changes in cell states as tissues and organs develop can give insights into the molecular mechanisms and order of events by which cells choose their terminal identities during embryogenesis or regeneration. It can provide clues as to how to manipulate cell fates *in vivo*, to predict the origins of developmental pathologies and cancer, and to re-create cell differentiation processes in vitro.

Recent advances in single-cell transcriptomics provide a powerful approach to mapping differentiation dynamics by densely sampling cells at different stages. These sampled cells together can be used to construct a continuum of cell states, or a 'landscape', a term historically inspired by Waddington's metaphorical epigenetic landscape[2]. In this Review, we refer to such depictions as state manifolds, to reflect both their underlying high-dimensional nature and their routine representation as low-dimensional Euclidean surfaces or graphs. State manifolds can provide high-resolution descriptions of cell trajectories as they transition between states during cell differentiation.

While they are powerful, state manifolds and state trajectories offer population-level views of differentiation, without directly revealing the long-term dynamic relationships between individual cells or between cells and their progeny. The gold standard for linking cell states across periods of time is instead through prospective lineage tracing: the practice of labelling an individual cell at an early time point in order to track the state of its clonal progeny at a later time point. Traditionally reliant on microscopy, lineage-tracing approaches have recently evolved to allow the tracking of cell clones via sequencing of inherited DNA sequences, or 'barcodes'. The migration to sequencing platforms has brought several advantages to lineage-tracing efforts: massive throughput, multiplexing and compatibility with other sequencing-based measurements (for example, RNA sequencing (RNA-seq)).

Recently, we and others have developed approaches to carry out single-cell omic-scale profiling while simultaneously reporting lineage information. These methods offer an opportunity to integrate complementary information about both cell lineage and cell state into synthesized views of differentiation dynamics. In this Review, we survey the currently available strategies for single-cell state manifold reconstruction and lineage barcoding, as well as omics methods for combining lineage and state measurements in the same cells. Both the range of single-cell trajectory construction methods and their assumptions have been reviewed extensively elsewhere[3,4], as have foundational molecular strategies for lineage barcoding[5,6]. Here we aim to draw general lessons from reoccurring conflicts that have emerged between state and fate analyses, and we discuss biological results obtained from first applications of combining the two methods. As this is an emerging field, we also discuss current limitations and potential technical pitfalls in their application. Finally, we speculate on the emerging concepts that might arise.

Inferring cell histories from state manifolds

In measuring the instantaneous state of a cell, one might imagine collecting information on the copy number of every molecular species within a cell, their interactions and spatial organization, the position of the cell in its

parent tissue, and its physical and regulatory interactions with other cells. Such a level of information is, of course, impractical. Working definitions of cell state capture only a subset of these attributes and vary dramatically between studies. In the following sections, we describe how cell state designations have evolved from relatively simple observations to quantitative high-dimensional and high-throughput omics measurements. We describe the introduction of cell state manifolds as a relatively recent analytic strategy with important advantages and limitations when inferring cell state relationships.

Defining cell states

A century ago, cells could only be reproducibly defined by simple characteristics: spatial position, morphology, histochemical staining, or basic biochemical or biophysical properties, such as cell density or dye uptake. Accordingly, much of the classical nomenclature associated with cell states (for example, basophilic) reflects these assays. With the advent of molecular biology, cells could be identified more quantitatively by the expression of selected marker genes, through immunocytometry, RNA analysis or the expression of transgenes. The nomenclature of cell state expanded accordingly into marker-based phenotypes (for example, CD34$^+$). The types of measurable determinants of cell identity continue to expand, including epigenetic state (for example, DNA accessibility and conformation, protein–DNA binding, DNA methylation or histone modifications), post-translational protein modifications, protein localization and the metabolic profile of cells.

At present, the most mature technology for genome-scale mapping of cell states is through measurements of the whole transcriptome (single-cell RNA-seq (scRNA-seq)), which can now be carried out rapidly and at low cost, in nanolitre-scale droplets[7,8], in microfluidic wells[9], or using combinatorial split-pool approaches[10]. Transcriptomes contain information about multiple aspects of cell identity (for example, cell cycle phase, metabolic state, cell-specific and tissue-specific molecular signatures, and spatially restricted marker genes). These diverse features may or may not be interrelated, but they reinforce a modern view of cell states as multidimensional vectors[11,12]. Beyond scRNA-seq, recent breakthroughs in single-cell methods capture chromatin accessibility[13,14], methylomes[15], proteomes[16] and metabolic signatures[17], as well as multimodal measurements from the same single cells (for example, mRNA and protein[18–20] or mRNA and DNA[21,22]). These measurements incorporate even further dimensions into routine measurements of cell state. Additionally, some highly multiplexed profiling of cell states is now possible in situ, thus complementing cell-intrinsic state information with detailed information on a cell's local environment and position in tissues[23–27]. Overall, these innovations set up the coming decade to be an exciting time for stem cell and developmental biology, as well as for tissue physiology in general. These new methods are clarifying the changes that occur in cells during development and, ultimately, the mechanisms governing cell behaviour.

Mapping state manifolds

Large single-cell datasets are now being routinely collected to catalogue the distribution and differentiation of cell states in both embryonic and adult tissues, as well as in disease. Recent examples encompassing entire organ systems include the haematopoietic system[28,29], lung[30,31], kidney[32,33], heart[34], gut endoderm[35], somitic mesoderm[36], nervous system[37] and neural crest[38]. Additionally, whole-organism datasets have been generated for *Caenorhabditis elegans*[39,40], *Nematostella vectensis*[41], *Hydra*[42], annelids[43] and planarians[44–46]. Furthermore, time series data for whole embryos have been mapped for zebrafish[47,48], *Xenopus laevis*[49], mouse[50,51], *Drosophila melanogaster*[52] and ascidians[53]. These datasets have revealed novel cell states, and they associate all states with detailed molecular signatures that extend well beyond the previous classifications based on marker genes alone. They also have revealed cells in developmental transitions involving thousands of genes, which change expression at progressive times and between tissues.

Analyses of these and other single-cell data involve several stereotypical steps to predict differentiation dynamics (FIG. 1). First, single-cell datasets noisily sample cells in different states (FIG. 1A). The challenge of data analysis is then to infer the continuum manifold of states from these measurements (FIG. 1B). These manifolds must be constructed, visualized and then used either to predict dynamics directly from cell states or else to represent the measured dynamic information (FIG. 1C). In this section we briefly introduce these steps.

To infer continuum state manifolds, most methods applied to single-cell data to date have been graph-based: they begin by representing individual cells as nodes, which are then connected by edges that reflect pairwise gene expression similarities (FIG. 1B). Graph-based analyses are useful because they convert a set of isolated measurements (single-cell transcriptomes) into a connected structure (the graph), which can then be analysed using a rich set of pre-existing mathematical methods.

To then visualize state manifolds, several algorithms are used that attempt to preserve the structure of the original cell graph when it is plotted in just two or three dimensions (such as uniform manifold approximation and projection (UMAP)[54], SPRING[55] and ForceAtlas2 (REF.[56])). Two-dimensional representations are popular and do capture meaningful biological trends. However, they can be misleading, as they distort high-dimensional structures upon 'flattening' them, and in some cases algorithms force tree-like visual layouts that may further distort the original structure[48,57,58]. Any 2D and 3D visualizations should serve only as aids for representing the results of more powerful forms of data analysis.

Independently of visualization, a multitude of algorithms propose to predict cell state dynamics and/or differentiation hierarchies directly from a manifold (FIG. 1C). These tools for dynamic inference have been reviewed extensively elsewhere[3] and include methods for extracting from the manifold its bare-bones structure, or topology[59]; organizing cells into trajectories[57,58,60–63] along an axis (often called pseudotime); and predicting the future fate of cells on the basis of their state[28,64–68].

State manifolds
Approximate representations of high-dimensional cell states (for example, the whole-animal embryonic cell state atlas Tabula Muris) as lower-dimensional shapes.

State trajectories
The paths taken by individual cells or clones of cells through a state manifold.

Prospective lineage tracing
A lineage-tracing experiment that introduces a label for marking cells in a specified state.

Barcodes
Units of DNA with a large number of sequence possibilities, such as those used to uniquely label cells and their progeny.

Cell lineage
A representation of a series of mitotic events that trace back to a single founder cell.

Cell state
A designation of cell identity (defined with respect to a particular measurement) that can be used to classify or quantify physical or molecular differences between cells (for example, 'basophilic', 'KRT4$^+$', 'columnar', 'RNA-Seq cluster 4').

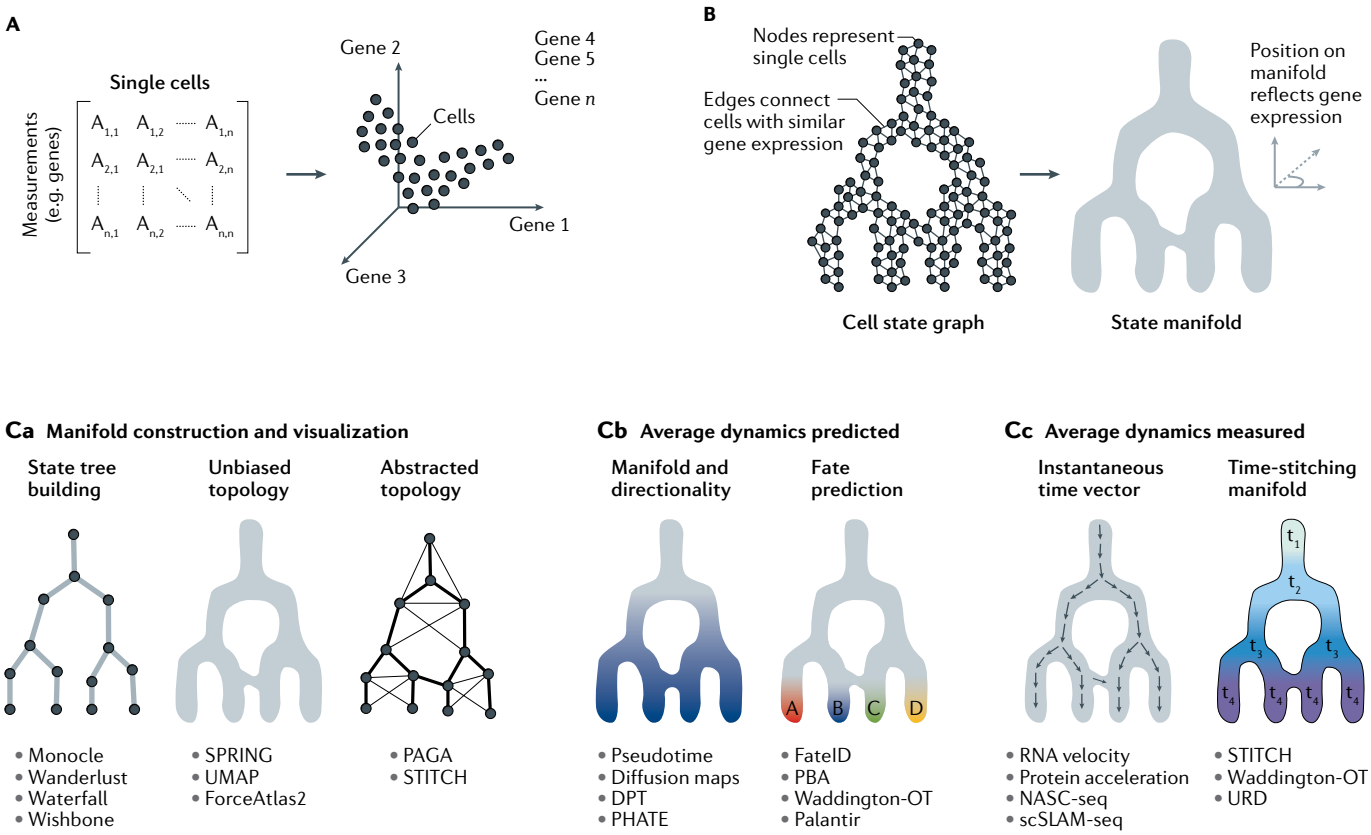


Fig. 1 | **Inferring cell histories from state manifolds. A** | Modern omics-based single-cell datasets, conceptualized as a measurement × cell count matrix or, alternatively, as cells plotted in a high-dimensional Euclidean space. **B** | Single-cell graphs, which link cells according to similarity (for example, Euclidean distance) in gene expression space, can be visualized to reveal underlying state manifolds that reflect gene expression dynamics. **C** | Graph-based tools for constructing and visualizing state manifolds (part **Ca**), computational algorithms for predicting dynamics directly from a state manifold (part **Cb**) and tools for incorporating independently measured state dynamics into a manifold (part **Cc**). DPT, diffusion pseudotime; NASC-seq, new transcriptome alkylation-dependent single-cell RNA sequencing; PAGA, partition-based graph abstraction; PBA, population balance analysis; PHATE, potential of heat diffusion for affinity-based trajectory embedding; scSLAM-seq, single-cell thiol-(SH)-linked alkylation of RNA for metabolic labelling sequencing; SPRING, a force-layout embedding of single-cell data; STITCH, a method for combining time series of single-cell data; UMAP, uniform manifold approximation and projection; URD, a simulated diffusion-based computational approach named after the Norse mythological figure; Waddington-OT, Waddington optimal transport.

To improve these efforts at dynamic inference, some recent studies have succeeded in inferring the instantaneous dynamics of states on the basis of measurements of nascent mRNA abundance, the ratio of spliced to unspliced mRNA (for example, RNA velocity), protein translation or mRNA turnover by metabolite labelling[69–73]. Temporal information can also be integrated into state manifolds when cells are sampled at time intervals[47,48,67] (FIG. 1C). In total, the result of these methods is to order cells along a continuum[74,75], which in turn allows for studying changes in the average, variance and correlation of gene expression across the graph, and for inferring tree-like structures from graphs[57,58,60,76] that organize cells or cell clusters[77,78] into a putative hierarchy.

RNA velocity
The rate of change in mRNA transcript abundance — more specifically, a set of computational techniques for calculating these rates across all genes from measurements of spliced and unspliced transcript abundances.

Limitations of state manifolds for dynamic inference

The representation of cell states as continuous manifolds offers a compelling approach to reconstructing dynamic processes. However, state manifolds average over many individual cells and so lose information on individual dynamics. The missing information includes cell division or death rates, the reversibility of states, and persistent differences between clones, all of which can quantitatively or qualitatively alter the dynamics predicted from snapshot measurements[64]. The dynamics predicted from cell state snapshots should thus be considered hypotheses. In this respect, the tree-like hierarchies of cell states sharply contrast with those obtained by bona fide lineage analysis (FIG. 2a), in which tree edges link cells with an empirical developmental relationship. On a state manifold, branch-points may be hypothetical: cell division may or may not occur at a branch-point, and sister cells from each division may both progress along one branch of a manifold, rather than exploring all branches. By contrast, in lineage trees, each branch-point strictly corresponds to a division event. State trajectories need not even be strictly tree-like, whereas lineage hierarchies are always strictly branching trees. Therefore, although the population-level structure could trace the dynamic sequence of molecular states experienced by single cells (FIG. 2a,b), several specific reasons could obscure or mislead researchers' understanding of the underlying dynamics and/or fate relationships (BOX 1; FIG. 2b–h).

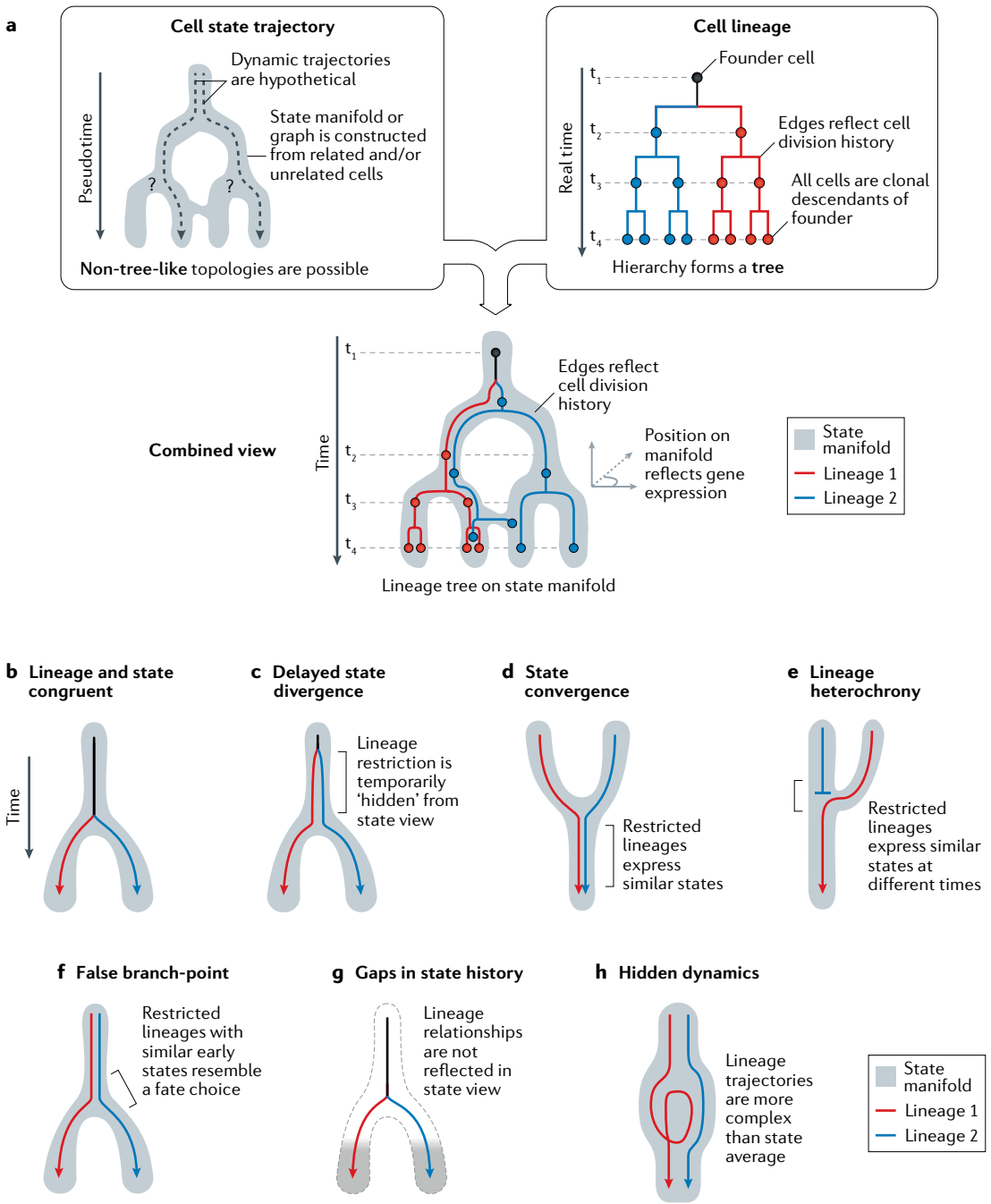


Fig. 2 | **Limitations of cell state manifolds. a** | Clarification of depictions of cell state manifolds versus cell lineage trees. Trajectory relationships are indirectly inferred from gene expression similarities, whereas lineage relationships reflect measured mitotic histories. Below the boxes, a combined representation highlights a clonal hierarchy of related cells, directly revealing its trajectory along a state manifold. **b–h** | Hypothetical scenarios of restricted lineage trajectories unfolding on a state manifold. The behaviours of distinct clonal units are presented in simplified form by coloured arrows. Lineage and state congruent (part **b**): initially all clones share the same fate potential (black); restriction of the clones into distinct trajectories (blue/red) occurs only where the manifold bifurcates. Delayed state divergence (part **c**): cells become committed to distinct trajectories (blue/red) but continue to occupy similar states for some time. This causes the early state to appear seemingly multipotent despite the cells within each clone being fate-restricted. State convergence (part **d**): cells with distinct molecular histories converge into similar states, such that the molecular origin of later cells can no longer be inferred. Lineage heterochrony (part **e**): cells with different origins occupy a sequence of states that implies a false developmental trajectory (blue to red). False branch-point (part **f**): an extreme case of the situation in part **c**, in which an apparent branch-point does not represent a decision made by any cell. Instead, it appears artificially when fate-restricted clones overlap in their early state. Gaps in state manifold (part **g**): disconnected cell states appear when the states of transitional and early progenitors are not represented in the dataset. This occurs when transitional states are very rare or when sampling a developing tissue at a late stage. Hidden dynamics (part **h**): the extent of stochastic or structured fluctuations in clonal dynamics is not visible from snapshots of cell states.

Clonal analysis
A lineage-tracing experiment that involves marking an individual cell, followed by state analysis of that founder cell's clonal descendants.

Retrospective lineage tracing
A lineage-tracing experiment based on phylogenetic reconstruction of endogenous genetic polymorphisms (that is, no experimental intervention).

Inferring cell histories in lineage tracing

Unlike the state of a cell, the lineage history of a cell can be defined without the operational simplification that comes from reducing the dimensionality of thousands of measurements. By 'lineage', we refer to the collective history of cell divisions, as well as the birth, division and death times of a cell's ancestors and clonal relatives. Lineages can be depicted as detailed trees of mitotic events (FIG. 2a) or, alternatively, as clonal units derived from a common progenitor cell. Lineage measurements, however, do not inherently contain information about the states of the cells they comprise, and as such they are typically combined with other measurements (for example, cell position, morphology or gene expression). In the following sections, we describe classic temporal and clonal analysis paradigms for defining cell lineage, referring the reader elsewhere for more complete reviews[5,6]. These paradigms have evolved from their roots in imaging-based studies to the recent use of DNA-barcoding-based systems in the post-genomics era.

Lineage-tracing paradigms

Currently there are two major paradigms for defining cell lineages. One major category of approaches, prospective lineage tracing, attempts to establish lineage relationships forwards in time from cells of a defined starting state. Fate mapping, the practice of associating the position of a cell in the early embryo with the ultimate positions and fates of that cell's descendants, is a form of prospective lineage tracing[79]. Methods based on CRE or FLIP recombinases, which facilitate permanent genetic labelling of progenitor cells based on the activity of a transgenic promoter, can also be used to learn prospective state relationships[80]. Prospective lineage tracing requires that some level of state information be known about a starting cell population, and generally the goal is to correlate this state information with future cell states. By contrast, phylogenetic lineage reconstruction methods seek to map the history of lineage relationships with respect to the cell states queried at a single end point in time. With these methods, state and lineage features are generally only measured at the end of the experiment, and lineage relationships are mapped backward in time in order to infer fate decisions that occurred either early or late. An inherent advantage of phylogenetic approaches is that analyses can be performed retrospectively (that is, without the need for experimental labelling) by analysing endogenous, naturally occurring genetic polymorphisms[81] — these label-free implementations of phylogenetic lineage analyses are thus also known as retrospective lineage tracing. Such approaches, therefore, can be applied to human patient samples and in other cases in which experimental intervention is not possible. Many additional methods perform similar phylogenetic reconstruction of end states but do so by tracking experimental labels rather than endogenous labels. In practice, when experimental labels are specifically introduced into cells of a particular state, the lineage-tracing experiments combine both prospective and phylogenetic paradigms — for example, by reconstructing lineage phylogenies within a specific tissue.

Clonal versus population tracing

Labelling of cells for prospective lineage tracing can be performed at clonal resolution (such as by delivering complex barcode libraries) or, alternatively, by documenting the collective fate of a population of cells (such as by delivering a common label to the cell population). Population lineage-tracing experiments are generally easier to perform, but they leave open the possibility of internal heterogeneity of the labelled population and/or inclusion of off-target cells. The collective activity of a bulk cell population (for example, unfractionated bone marrow) can easily be misinterpreted as representing the

Box 1 | **How do cells traverse single-cell landscapes?**

Single-cell data can be organized into a continuum 'landscape' (or 'manifold') of cell states, representing cells in progressive states of differentiation. However, these landscapes do not directly clarify how cells or clonal lineages explore these states, or how they choose their trajectory at branch-points. Cells in similar states could show the same dynamic progression and remain uncommitted until they reach a branch-point (FIG. 2b). Dynamic behaviours on a landscape can also be unpredictable (FIG. 2c–h) and can deviate from the global averages inferred from many single-cell measurements. These complications may arise from hidden variables, gaps in manifolds and stochastic dynamics, as detailed here.

Hidden variables

Although single-cell technologies aspire to measure cell state comprehensively, they may still miss important cellular properties that are informative as to fate. Such hidden variables may be regulatory molecules that are altogether missing from the state measurement (for example, epigenetic, spatial or post-translational state features that are not captured by single-cell RNA sequencing). They could also be obscured by measurement noise or by ad hoc operational decisions in data processing, such as choices of normalization or dimensionality reduction strategies and of which genes to include in manifold construction. Because cells often participate in multiple dynamic processes, different data-processing choices can emphasize certain biological processes (for example, cell cycle, cell migration or stress) over others and can allow constructing manifolds with qualitatively different structures from the same data. Transcriptional signatures of the cell cycle, for example, can overshadow other state differences between unrelated cell types.

Failure to resolve hidden variables can lead to the appearance of 'delayed state divergence' on a state manifold (FIG. 2c), obscuring the true point at which fate specification occurs. Independent clonal trajectories can also appear to 'converge' temporarily on identical or nearly identical states during the differentiation process (FIG. 2d). In these cases, the distinct ontogeny of the cells cannot be deduced from state information alone. Clones or tissue domains in different stages of differentiation may also form a continuum of states that implies a false trajectory (FIG. 2e). Additionally, state manifolds may imply 'false multipotency' and/or 'false branch-points' by superimposing cells with different fate potentials (FIG. 2f). Some of these problems can be identified by visualizing the state manifold, but this is only possible if the visualization methods used do not force cells to occupy a tree-like hierarchy. Lineage tracing is essential to identifying and resolving lineage restrictions downstream from a point of state convergence (FIG. 2d,e); not even short-term dynamic information (for example, RNA velocity; see the main text) is informative in these situations.

Gaps

Learning differentiation trajectories works best for systems with a strong flux of cells and coverage of multiple time points. When very few cells differentiate at any moment in time (for example, adult neural stem cells) or transitional time points are missing, analyses become more difficult and are prone to creating artefacts. Thus, gaps in the manifold, which may arise from uneven or under-sampling of cell states, can result in apparent discontinuities between clonally related cells (FIG. 2g).

Stochastic dynamics

Even when manifolds faithfully depict the average clonal dynamics, they cannot provide information about distinct dynamic behaviours of cells that appear similar in state, such as stochastic fluctuations about the average or participation in local cycles (FIG. 2h). Collectively resolving the scenarios above would require the ability to track cell state dynamics over both short and long timescales.

output of a single, multifunctional cell type, even when labelled cells could be restricted in their fate potential[82]. Such errors can be resolved by increasing the precision of the labelling process to limit any underlying cellular heterogeneity, or by utilizing single-cell lineage methods to track clonal relationships[83]. Given that even genomics-era state measurements (for example, scRNA-seq) can occasionally fail to fully resolve lineage-restricted groups of cells[47,84–86] (FIG. 2c–h), clonal analysis is still the most robust method for establishing the distribution of lineage outputs of a cell population. Once identified, stereotyped clone behaviours can be used to screen for prospective cell state markers that might correlate with and/or predict different lineage outcomes.

Imaging-based methods for lineage tracing

Prospective lineage-tracing experiments date back to the 19th century and initially relied on direct observations via live microscopy to track blastomere divisions in transparent invertebrate embryos, in particular in annelids[1] and ascidians[87]. Ascidian lineage trees were annotated according to the spatial position of each cell, and owing to determinate cleavage patterns and early fate restriction, this relatively simple level of state information was found to be sufficient to predict future cell fates. A similar direct-observation strategy was applied nearly a century later to the nematode *C. elegans*, again taking advantage of the small size, transparency and determinate embryonic cleavage patterns of this species[88].

Embryos of more complex species (for example, vertebrates) often contain many more cells, and cell divisions are generally indeterminate and more difficult to observe directly. Lineage tracing thus expanded to include a wide range of additional approaches, including the injection of tracer dyes, cell transplantation and *in vivo* genetic recombination methods. The history and applications of these pre-genomic methods have been reviewed extensively elsewhere[80]. More recent advances in *in toto* confocal and light-sheet microscopy have reinvigorated modern versions of the direct-observation approach, enabling the tracking of individual cell division patterns in complex vertebrates such as zebrafish and mouse, together with transgenic reporters[89,90]. One feature common to *in toto* imaging and nearly all pre-genomics methods for live lineage tracing is a reliance on transgenic fluorescent reporters to measure cell state. Thus, these approaches are spectrally limited to relatively few measurements of cell state. Partially countering this limitation, the spatial position of cells and their morphology provide information that may be correlated to molecular state[91]. Furthermore, recent spatial transcriptomics methods overcome the spectral limit by allowing genome-scale measurements in fixed samples in situ. Using such methods subsequent to live imaging or in combination with lineage tracing allows for combining state information with lineage and position information in one experiment[92]. However, such experiments remain extremely challenging, and highly multiplexed spatial transcriptomics methods are still generally restricted to the analysis of tissue sections, which may fail to capture all cells in each clone.

Lineage tracing by barcode-sequencing

Recently, high-throughput sequencing has opened up a new generation of lineage-tracing approaches. These new methods use DNA sequence barcodes to encode clonal information (FIG. 3). Although the number of distinct clones that can be simultaneously queried using fluorescent reporters is intrinsically limited, DNA sequence complexity scales exponentially with the length and multiplicity of the engineered barcodes, which is theoretically sufficient to allow a record of every single division event in an organism. The recorded information is read out retrospectively using high-throughput sequencing and can be readily combined with other sequencing-based omics measurements.

The use of DNA barcodes to reconstruct lineage relationships initially relied on the identification of unique retroviral integration sites and utilized Southern blot or PCR assays to reveal barcode identity[93,94]. In the post-genomics sequencing era there has been a burst of innovation in the creation and deployment of far more complex DNA barcodes for lineage tracing (TABLE 1). A foundational concept for these methods is to use changes in targeted, whole-genome or mitochondrial-genome sequencing data to construct lineage phylogenies[95–97]. Targeted barcoding-based methods generally fall into three thematic categories: first, transgenic integration of exogenous DNA sequences (FIG. 3Aa); second, *in vivo* recombination of transgenic DNA cassettes (FIG. 3Ab); and third, *in vivo* editing of transgenic DNA targets by CRISPR–Cas9 (FIG. 3Ac). In all of these approaches, a DNA-barcoding event permanently alters the genome of an individual cell, the descendants of which inherit the barcode and can be distinguished as a clonal unit (FIG. 3Ba). Importantly, DNA barcodes can be recorded and measured at high throughput, enabling that interrogation of hundreds or thousands of distinct clonal units in parallel. In addition, these modalities can be adapted for cumulative barcoding, which marks successive/nested clonal units and facilitates phylogenetic reconstruction of cell lineage trees (FIG. 3Bb).

The first generation of methods and the logic for sequencing-based lineage tracing have been reviewed extensively elsewhere[5,6]. It is instructive to review the most recent developments, particularly in CRISPR-editing-based barcoding schemes. This family of methods utilizes a cumulative barcoding strategy to reveal lineage hierarchies that terminate at a single end point in time, typically by introducing three transgenic components: CRISPR–Cas9 DNA endonuclease, an array of DNA target sites, and a panel of single guide RNAs (sgRNAs) or homing guide RNAs (hgRNAs). These components generate high-diversity, 'evolving' DNA barcodes within cells by taking advantage of cumulative variability in target sites that results from CRISPR–Cas9 activity. The first methods to demonstrate this principle were genome editing of synthetic target arrays for lineage tracing (GESTALT)[98] and homing CRISPR[99,100]. More recent innovations include the engineering of lineage barcodes into transcribed regions of constitutively expressed or inducible reporter genes, enabling their sequences to be read from mRNA in whole-transcriptome scRNA-seq experiments

Hidden variables
Molecular or environmental properties of a cell that correlate with — or could be used to predict — a cell fate decision, which are obscured from a state manifold.

Direct observations
Lineage-tracing experiments that rely on *in vivo* live imaging of cells as they divide.

Determinate
In the context of developmental processes, when the relationship between lineage and molecular state is tightly controlled at each cell division event and is invariant between individuals.

Indeterminate
In the context of developmental processes, when the relationship between lineage and molecular state can vary greatly between individuals and between cell clones.

Lineage phylogenies
Trees of lineage relationships constructed from end point measurements.

(FIG. 4A). This innovation was first demonstrated by the single-cell GESTALT (scGESTALT)[101], lineage tracing by nuclease-activated editing of ubiquitous sequences (LINNAEUS)[102] and ScarTrace[86] techniques and has become a standard feature in subsequent methods. Other common innovations include the use of barcode arrays, which increase the number of barcoding possibilities, as well as the use of inducible promoters and integrated fluorescent reporters to both control and monitor the barcoding process in real time (FIG. 4B).

Performance, trade-offs and further innovations

DNA-barcoding technologies show considerable potential as future tools for lineage tracing. As this is a rapidly evolving field, the published methods are likely to be revised substantially in the coming years. For this reason, we do not recommend any single published method at present over others. It is helpful instead to appreciate the limitations that are likely to be resolved, as well as some methodological improvements that are already emerging.

DNA-damage-induced toxicity. Most CRISPR–Cas9 barcoding methods rely on random insertions and deletions introduced during the process of double-strand break repair by non-homologous end joining (NHEJ). Recently, CRISPR–Cas9 activity has been shown to cause cell death in human induced pluripotent stem cells (iPSCs)[103] and cell lines[104], and also it can result in developmental delay in mouse embryos[85], raising potential concerns about maintaining continuous endonuclease activity. The extent and effect of potential off-target double-strand breaks also remains generally unaddressed. Going forward, it will therefore be important to validate that these systems do not perturb the developmental dynamics that they are being used to interrogate.

The alternatives to CRISPR–Cas9-based methods may not face the same concern of excessive DNA damage. One alternative is TracerSeq[47], a method for clonal barcoding demonstrated in zebrafish. TracerSeq makes use of ongoing transposase activity to successively integrate a pool of predefined barcodes, delivered as an injected plasmid library into embryos. The progressive integration of plasmids into the genome provides a heritable label of clones and sub-clones without inducing unrepaired double-strand breaks, yet it does require injection or electroporation. Other alternatives that similarly avoid double-strand breaks use genetic recombination[105,106] (for example, PolyLox), CRISPR-associated transposase systems (CAST and *Vibrio cholerae* Tn6677)[107,108] and base-editing enzymes[109,110]. Base-editing enzymes, however, can have substantial off-target effects that could perturb biological function[111–113].

Barcode detection. Failure to detect edited barcode sequences (for example, due to measurement drop-outs) can skew inferred lineage relationships (FIG. 3Cb). Such errors arise, for example, from low or noisy levels of barcode reporter expression or from endogenous silencing of integrated transgenes or lentiviral constructs[114]. We do not at present know the precise barcode detection rates of existing methods, but the extent of such errors for any barcoding method can be estimated in principle through control experiments in which lineage relationships can be independently verified[115]. At a minimum, studies using DNA barcodes should assess the per-cell barcode detection rate, and may need to consider taking steps to improve experimental detection (for example, introducing strong RNA polymerase II promoters that drive the transcription of mRNA-based barcodes).

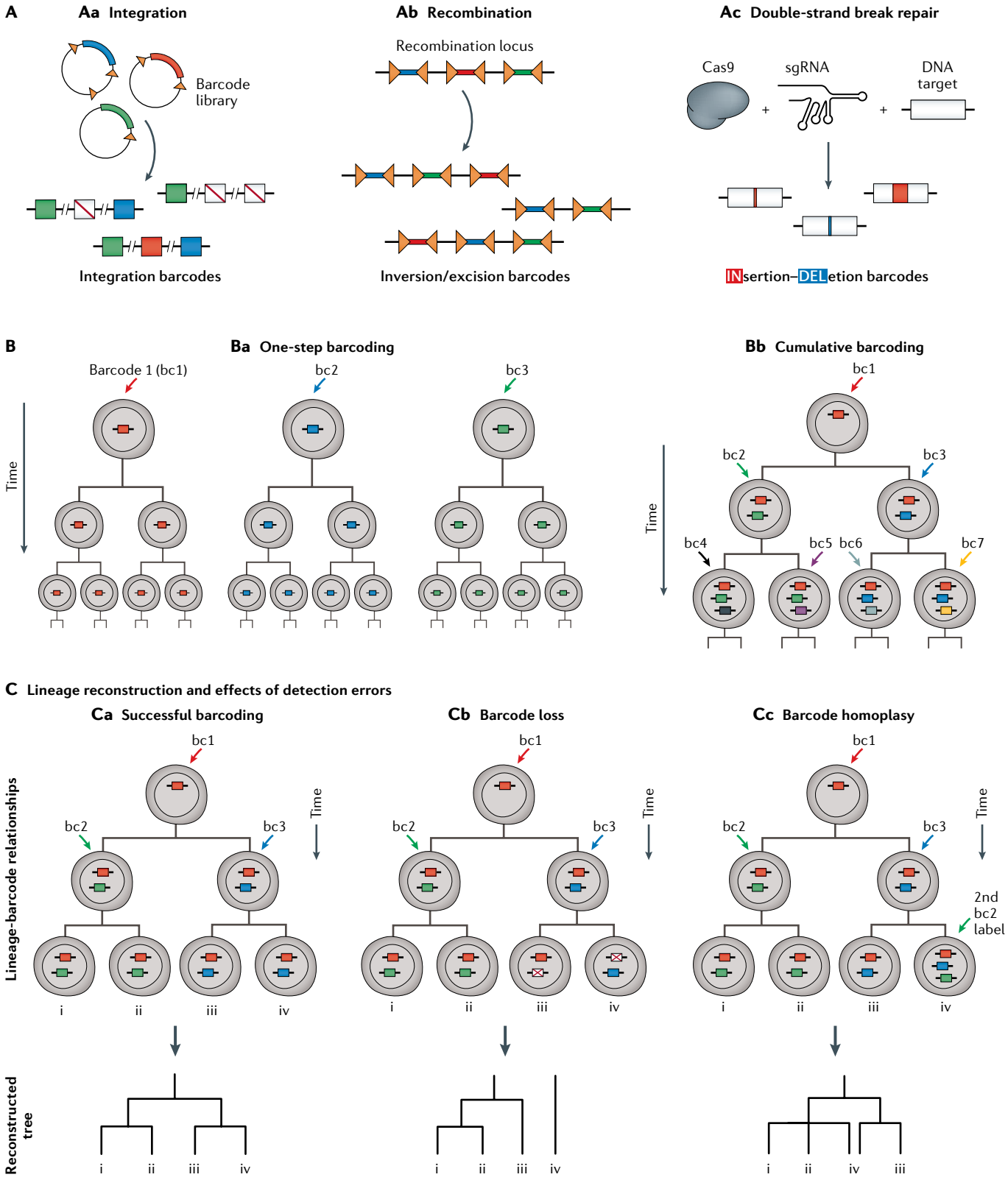
Assay calibration. Because lineage tracing and single-cell omics assays can take weeks to analyse and are expensive, it is desirable to be able to assess the efficiency of barcoding before detection. In integration-based systems, the expression of barcode-linked fluorescent proteins can report on the level and specificity of barcoding activity in live specimens. Some CRISPR–Cas9 systems (for example, LINNAEUS and ScarTrace) target DNA editing to the coding region of a fluorescent transgene, such that loss of fluorescence can be used to monitor the barcoding process. Such live-reporting schemes for barcode generation provide a simple means for sample validation before sequencing.

Barcode diversity. Failures to resolve unique clones (that is, barcode homoplasy or type I errors) occur when cells inherit identical barcode sequences despite having no true lineage relationship (FIG. 3Cc). To avoid such errors, lineage-tracing methods should generate far more barcodes than the number of clones to be analysed. In CRISPR–Cas9 systems, the barcode diversities generated by Cas9 that are quoted in different studies have varied considerably. The true barcode diversity obtained in such systems, however, is likely to be overestimated, in part, because certain errors in double-strand break repair re-occur frequently[102]. In addition, the generation of multiple DNA double-strand breaks in close proximity leads to the excision of intervening sequences, resulting in the loss of previously generated edits[101]. Finally, the

Fig. 3 | **Methods and logic for lineage barcoding experiments. A** | Three major paradigms for introducing unique DNA barcodes into cells: by integration of a high-diversity library of DNA barcodes using a transposase (part **Aa**), by random recombination of an array of recombinase target sites (part **Ab**) and by the accumulation of random errors, insertions and deletions during CRISPR–Cas9 editing of genomic target sites (part **Ac**). **B** | DNA barcoding can be applied in a single, instantaneous pulse, enabling the parallel tracking of many distinct cell clones (part **Ba**). When applied continuously, DNA barcodes can repeatedly label a dividing cell clone at sequential levels of its lineage hierarchy (part **Bb**). **C** | Challenges in lineage reconstruction from cumulative barcoding. The upper diagrams depict hypothetical barcode integration events in a cell lineage. Arrows denote the accumulation of novel barcodes, with each colour indicating a unique DNA barcode sequence. Hypothetical lineage correlation heat maps and trees depict the anticipated results of lineage reconstruction. Lineage phylogenies can be accurately reconstructed from single-cell correlations of the detected barcode labels (part **Ca**), whereby early versus late clones are distinguished on the basis of the number of cells that contain the associated barcode. Errors in barcoding or barcode detection can skew the accuracy of phylogenetic inferences (parts **Cb** and **Cc**). sgRNA, single-guide RNA.

Drop-outs
Type II errors that are common in single-cell omics experiments in which transcripts, lineage barcodes or other features present in cells fail to be detected.

Barcode homoplasy
A type I error in which identical DNA sequence barcodes are randomly recovered from cells with no close lineage relationship.



activity of DNA repair machinery may differ between organisms, tissues and/or species.

To minimize the negative effects of barcode homoplasy, it is possible to utilize biological replicates to empirically identify high-frequency barcodes and exclude them from downstream analyses[85,102]. However, the presence of species-specific and tissue-specific differences in barcode diversity argues that diversity should be evaluated in each experimental system to which the methods are applied. An additional innovation for increasing Cas9-edit barcode diversity includes the use of terminal deoxynucleotidyl transferase (TdT) as an

Table 1 | **Sequencing-based technologies for lineage tracing**

Technology	DNA-editing system	Barcode type	Barcode length (bp)	Uniform barcode frequency?	Frequent barcode homoplasy?	Barcode as mRNA	Barcode generation	Species	In vivo?	Refs
TracerSeq	Tol2	Integration	20	Yes	No	Yes	Continuous	Zebrafish	Yes	47
LARRY	Retrovirus	Integration	28	Yes	No	Yes	Single-step	Mouse	Yes	84
CellTag	Retrovirus	Integration	8	Yes	No	Yes	Multi-step	Human	No	114,116
PolyLox	Cre–*loxP*	Recombination	2,152	No	Yes	No	Continuous	Mouse	Yes	105,106
GESTALT	Cas9	INDEL	266	No	Yes	No	Continuous	Zebrafish	Yes	98
scGESTALT	Cas9	INDEL	363	No	Yes	Yes	Continuous	Zebrafish	Yes	101,119
ScarTrace	Cas9	INDEL	249	No	Yes	Yes	Continuous	Zebrafish	Yes	86
LINNAEUS	Cas9	INDEL	75	No	Yes	Yes	Continuous	Zebrafish	Yes	102
MARC1	Cas9	INDEL + integration	240	No	Yes	No	Continuous, evolvable	Mouse	Yes	99,100
Chan et al.	Cas9	INDEL + integration	350	No	Yes	Yes	Continuous	Mouse	Yes	85
CHYRON	Cas9 + TdT	INDEL (with insertion favoured over deletion)	100	No	Minimal	No	Continuous	Human	No	115

A summary of lineage-tracing methods that make use of sequencing DNA barcodes. CHYRON, cell history recording by ordered insertion; GESTALT, genome editing of synthetic target arrays for lineage tracing; INDEL, insertion or deletion; LARRY, lineage and RNA recovery; LINNAEUS, lineage tracing by nuclease-activated editing of ubiquitous sequences; MARC1, mouse for actively recording cells 1; scGESTALT, single-cell GESTALT; TdT, terminal deoxynucleotidyl transferase.

additional transgenic component expressed at the time of barcoding[115]. In the presence of double-strand breaks, TdT was demonstrated to catalyse the random incorporation of nucleotides at the DNA cut site, resulting in an increased frequency of insertion-based edits over deletion-based edits.

It is also possible to expand barcode diversity by increasing the number of barcoding events per cell, although this strategy can carry experimental trade-offs. In CRISPR–Cas9 systems, barcode diversity can be increased through the parallel editing of several transgenic DNA target sites arranged in tandem or distributed throughout the genome (FIG. 4B). Tandem barcode arrays (FIG. 4Ba) face a practical limit, as they form repeat-rich sequences that are problematic substrates for both molecular cloning and most modern single-cell sequencing pipelines. Most of the recent methods therefore use distributed barcode arrays (FIG. 4Bb), which greatly reduce the number of nucleotides that must be sequenced in order to recover the barcode identity and also provide the advantage of being far less susceptible to internal deletions and information loss[47,84,85,100,116]. Distributed arrays can resolve otherwise identical DNA target sites through the use of an additional layer of integration barcodes that are specific to each transgenic insertion site. Distributed arrays are thus inherently scalable and can increase the barcode-space complexity while avoiding the need for long sequencing reads. However, they too face a limitation, in that a failure to detect some barcodes (type II errors) may lead to partial barcode recovery for many cells. Additionally, distributed arrays may be lost during outbreeding of transgenic animals and/or through endogenous silencing of transgenic or lentiviral constructs.

Barcode diversity may be less of a challenge for integration-based or recombination systems. Integration-based systems use high-diversity, uniform barcode libraries that are both simple to recover by sequencing and straightforward to interpret. TracerSeq barcodes, for example, are sampled evenly from a large sequence space (20-nucleotide sequences, yielding ~10^{12} possible variants), greatly simplifying computational analyses and the assignment of cells to clones[47]. Furthermore, increasing the integration rate expands combinatorial diversity by allowing more than one barcode to label each cell[114,116]. A drawback to the use of defined barcode libraries is that they require the introduction of exogenous transgenic DNA libraries into cells through injection, viral transduction or electroporation/lipofection, which limits their experimental possibilities[47,114]. In recombination systems, the number of barcode possibilities increases with the number of recombination sites. In the PolyLox system, 9 *loxP* sites yields >1.8 million Cre recombination possibilities[105]; this diversity could be further increased by adding more sites.

Barcoding precision. A critical requirement for any lineage-barcoding experiment is the need to capture a minimum of two cells (ideally, many more) per clone. This requirement argues strongly for the need to label small numbers of cells in a defined tissue of interest, in order to ensure adequate sampling of their resulting progeny. In addition, the interpretation of clonal-tracing experiments depends strongly on precisely controlling the time interval in which cells are labelled. To date, published methods have not yet been optimized to achieve both tissue and temporal specificity in barcoding. Targeting clonal labelling to specific tissues can be facilitated by expressing components of the barcoding machinery under the control of tissue-specific promoters. Achieving temporal specificity is a more complex challenge. For CRISPR–Cas9-based methods, an open problem is that of target site 'exhaustion', in which all editing is completed early in the developmental period of interest. We expect the practical challenges of targeting clonal labelling to be resolved in the coming years.

Applications of lineage tracing on state manifolds

Lineage-tracing methods can now integrate high-dimensional state information with clonal and phylogenetic barcoding. In doing so, they greatly increase the number of clones that can be tracked, and they establish clonal composition without requiring prior knowledge of the marker genes. Both of these advantages should greatly reduce transgene-centric observation biases. However, omics lineage-tracing experiments demand novel experimental designs and controls, as compared with traditional methods. These methods also demand far more computational support than do traditional

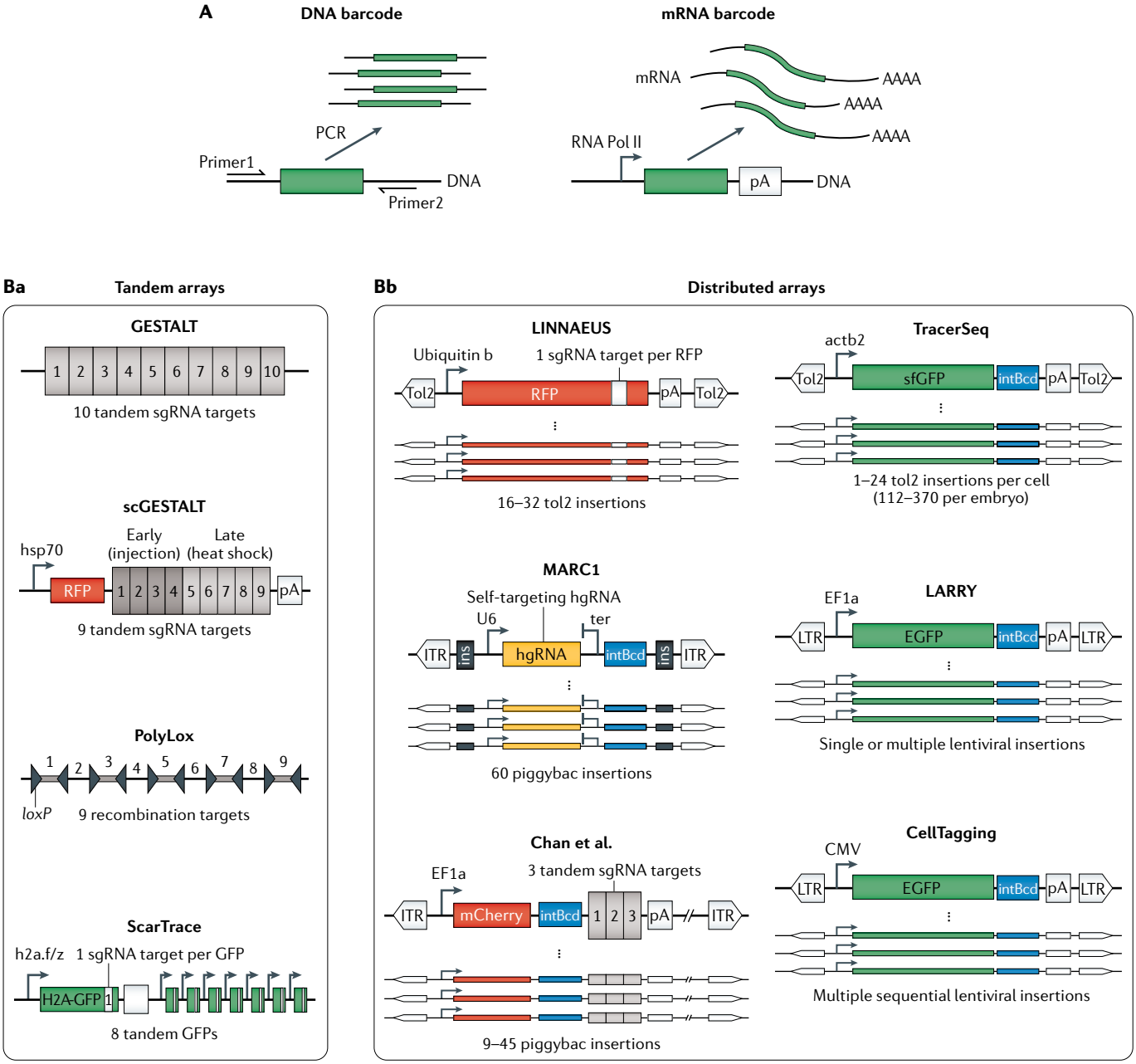
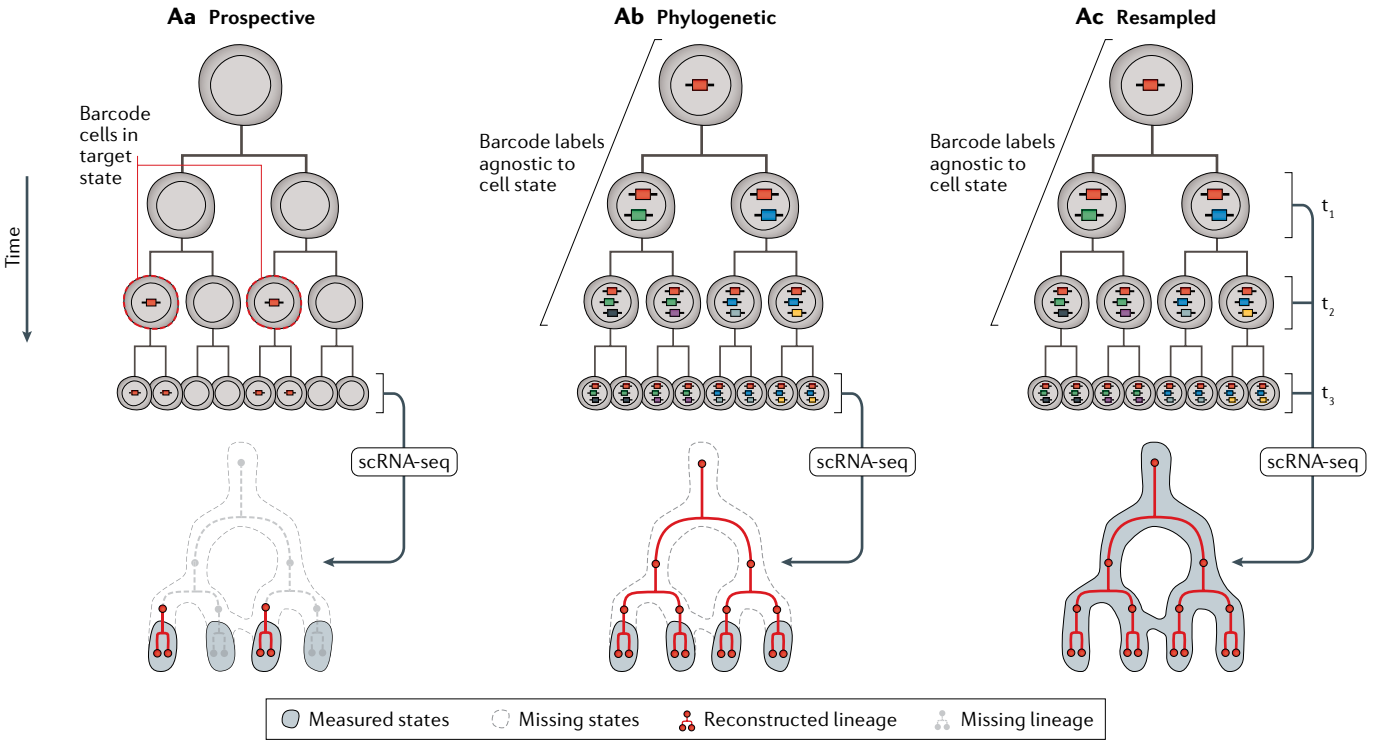
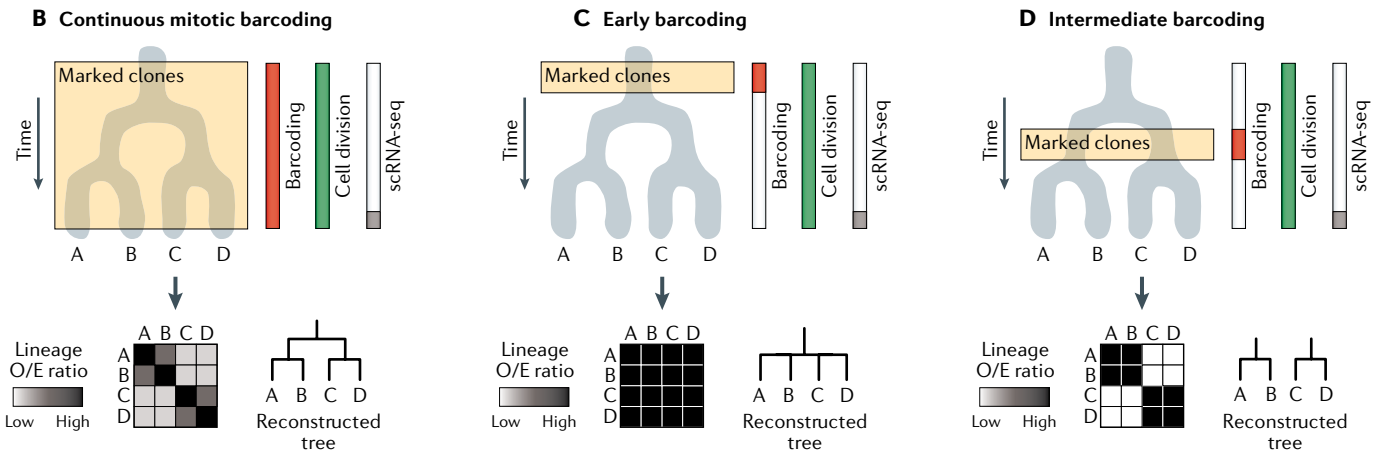
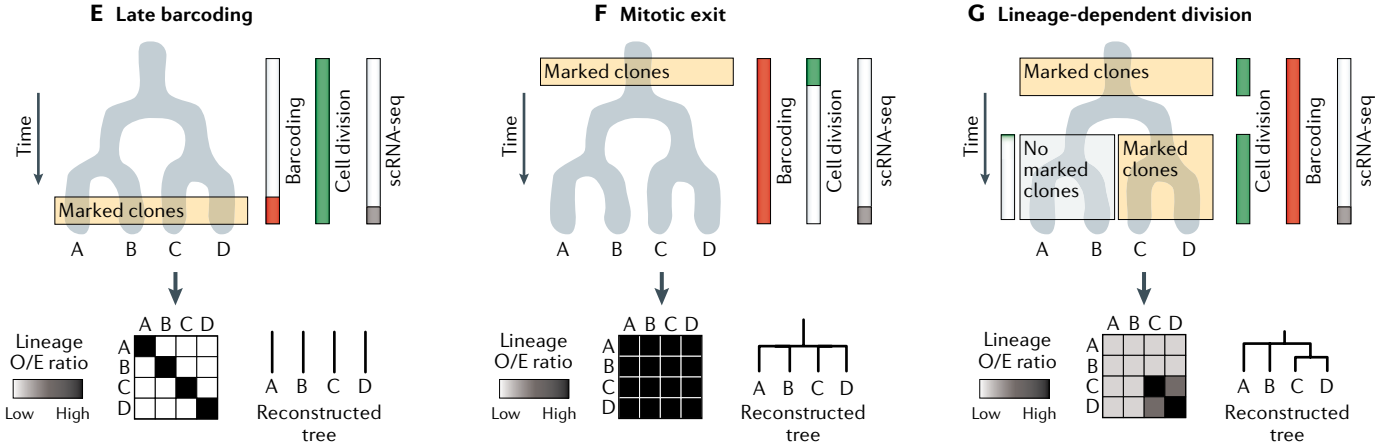


Fig. 4 | **Reading and writing transgenic DNA barcodes. A** | DNA barcodes can be encoded exclusively in genomic DNA (left) or expressed as mRNA, to allow detection concurrent with single-cell RNA sequencing. Reliable detection of barcode sequences requires amplification. For DNA barcodes this is achieved by PCR or in vitro transcription, whereas mRNA-based barcodes are endogenously amplified via RNA polymerase II (Pol II) transcription and can be detected as part of each single-cell transcriptome. **B** | Transgenic strategies for storing and transcribing DNA barcodes. The schematics show the diversity of DNA arrays used to store lineage information for each method. The arrays can be grouped according to whether they store lineage information at a single genomic locus using a tandem array (part **Ba**) or whether they store lineage information at multiple genomic loci using distributed arrays (part **Bb**). Right-angled black arrows indicate promoters used to drive barcode expression for detection by RNA sequencing in a subset of methods. The methods differ in whether they utilize recombination (PolyLox), barcode library integration using a lentivirus or transposase (TracerSeq, LARRY, CellTagging) or CRISPR–Cas9 targeting of single guide RNA (sgRNA) arrays (all remaining methods). GESTALT, genome editing of synthetic target arrays for lineage tracing; hgRNA, homing guide RNA; LARRY, lineage and RNA recovery; LINNAEUS, lineage tracing by nuclease-activated editing of ubiquitous sequences; MARC1, mouse for actively recording cells 1; scGESTALT, single-cell GESTALT.


A Published experimental designs
Aa Prospective
Barcode cells in target state
Time
scRNA-seq
Ab Phylogenetic
Barcode labels agnostic to cell state
scRNA-seq
Ac Resampled
Barcode labels agnostic to cell state
t1
t2
t3
scRNA-seq
Measured states
Missing states
Reconstructed lineage
Missing lineage


B Continuous mitotic barcoding
C Early barcoding
D Intermediate barcoding
Time
Marked clones
Barcoding
Cell division
scRNA-seq
A B C D
Lineage O/E ratio
Low High
Reconstructed tree


E Late barcoding
F Mitotic exit
G Lineage-dependent division
Time
Marked clones
No marked clones
Barcoding
Cell division
scRNA-seq
A B C D
Lineage O/E ratio
Low High
Reconstructed tree

◀ Fig. 5 | **Applications and pitfalls of lineage tracing on state manifolds. A** | Recent studies have highlighted three experimental designs for combining lineage and state measurements. For simplicity, the panels depict largely congruent state–lineage hierarchies. Prospective (part **Aa**): a bulk genetic label is applied to cells of a particular state; labelled cells are subsequently captured and sequenced to reveal the gene expression states and lineage barcodes for each cell. Phylogenetic (part **Ab**): gene expression states and lineage barcodes are measured at a defined end point with respect to a biological process. Prior lineage relationships can be reconstructed retrospectively from the lineage barcodes, whereas state information is limited to the final time point. Resampled (part **Ac**): gene expression states and lineage barcodes are repeatedly sub-sampled over time, enabling the mapping of lineage trends directly on the state manifold. **B**–**G** | Phylogenetic reconstruction of fate hierarchies from end-point state and lineage measurements. The results of hypothetical lineage–state reconstruction analyses are displayed for each scenario; they vary dramatically, depending on the timing of both cell division and lineage barcoding. Heat maps depict the number of shared barcodes observed between each pair of states, normalized by the expected number of barcodes under a null hypothesis in which barcodes are distributed at random ('Lineage O/E ratio'). For a thorough definition of this statistic, see Weinreb et al. (2020)[84]. Lineage relationships can only be inferred at the time points when marked clones are generated and expanded. Given constant cell division rates and identical state manifolds, different time windows of barcode induction will lead to different inferences about lineage relationships. **B** | Continuous lineage barcoding in an actively dividing cell population enables all major lineage restriction events to be well-represented in a lineage–state reconstruction analysis. **C**–**E** | Lineage relationships can only be inferred at time points when marked clones are generated and expanded . Given constant cell division rates and identical state manifolds, different time windows of barcode induction will lead to distinct inferences about lineage–state relationships. **F** | In postmitotic differentiation hierarchies, despite continuous DNA barcoding, an absence of cell division precludes the formation of marked clones containing >1 cell. Barcodes are no longer enriched across the state manifold and cannot be used to reconstruct fate restriction hierarchies. **G** | Lineage inferences require well-sampled barcode data from marked clones. Variable rates of cell division on a state manifold skew clone sizes and, hence, the statistical power to detect lineage–barcode correlations. scRNA-seq, single-cell RNA sequencing.

methods, owing to the high-dimensional nature of omics measurements, the difficulty of studying thousands of lineage trees that may be heterogeneous, and the unique nature of noise in these methods compared with previous approaches. We next survey three general experimental strategies that recently have been utilized to map how lineages unfold on state manifolds. We review outstanding challenges encountered in the computational analysis of state–lineage relationships, as well as potential pitfalls in experimental design.

Prospective lineage tracing on state manifolds

Prospective lineage tracing is still most commonly deployed by marking cells of a defined state at an early time point and establishing their collective cellular products at a later time. This approach has its roots in classical fate-mapping and genetic-labelling methods, which are implemented at either bulk or single-cell resolution. A modern version of this approach combines a sequencing-readable DNA recombination event as the genetic label, cell sorting and downstream analysis by scRNA-seq or some other high-resolution genome-scale measurement (FIG. 5Aa). Such approaches can provide detailed state resolution for the resulting cell populations, but prospective lineage labels are still comparatively low-resolution when they rely on the promoter activity of a single gene.

In an instructive example, Rajagopal and colleagues[30] made use of the classical genetic recombination-based lineage-tracing method to label cells, followed by scRNA-seq to analyse the fates of the labelled cells. This 'pulse-seq' method was applied to study the fate of basal cells in airway epithelium labelled at a defined time using a conditional CreER recombinase expressed from the *Krt5* locus, which activates permanent and heritable expression of a fluorescent reporter gene. At later time points, scRNA-seq established that *Krt5*–CreER-marked basal cells regenerated all epithelial cell types of the airway. Crucially, this approach did not require knowledge of the markers for any of the derivative cell types and thus could be used to establish the basal origin of a novel cell population. In this study, scRNA-seq also revealed that *Krt5*-expressing basal cells are not homogeneous in their transcriptomes. Thus, while pulse-seq could collectively mark all basal cells, it could not distinguish which individual clonal behaviours are stem cell-like, nor could it correlate such clonal behaviours to a particular subset of the *Krt5*-expressing population.

A more refined approach to prospective lineage tracing would make use of DNA barcodes to uniquely label each cell in the initial cell population. Although resolution into the labelled cell state would still be limited to the expression of a single marker gene, the end point measurements would provide information on heterogeneity in the clonal output of the labelled cells. From such data, one could assess whether the entire labelled cell population was equal in its fate potential or whether further fractionation of the labelled cells might be needed to resolve distinct cellular subsets. Such experiments could use scRNA-seq to analyse the clonal progeny. Evolving barcode approaches could also be adapted as a variant of such prospective lineage tracing, by requiring that barcode evolution be conditional on the expression of a state-prognostic transgenic promoter. To our knowledge, however, such applications have not yet been demonstrated.

Lineage phylogenies on state manifolds

Although there are now multiple methods for phylogenetic lineage barcoding (FIG. 4), all share a common goal: to determine the shared division history of the cells collected at a single end point in time. A common innovation in recent lineage-barcoding studies has been the engineering of lineage barcode cassettes into expressed sequences (FIG. 4A, right), enabling the simultaneous measurement of lineage information and whole-transcriptome state measurements for each cell. While these approaches succeed in revealing detailed states for the end-point-sequenced cells (FIG. 5Ab), they fail to capture the transcriptional states of progenitor cells that existed at time points before sequencing. Thus far, early applications of phylogenetic state–lineage approaches have largely recapitulated known developmental hierarchies in proof-of-concept studies. They have, however, revealed a recurring insight: namely, that similar cell states can arise (or 'converge') from qualitatively different developmental origins. Lineage construction using the tools reviewed above (FIGS 3,4) can therefore be useful to identify converging developmental trajectories (FIG. 2d) and to distinguish other trajectories (FIG. 2b–h; BOX 1) that are not immediately highlighted by state manifold approaches alone. They also can be used to identify measured features of cells

Cell ontogeny
The developmental history of a cell.

State convergence
A differentiation scenario in which cells with distinct origins converge onto the same end point on a state manifold.

(for example, novel marker transcripts) that reflect their cell ontogeny.

Three recent studies spanning different embryonic tissues illustrate the recurring observation of state convergence, although these examples are far from exhaustive. In a first example, integration-based barcodes were detected in scRNA-seq data (by TracerSeq; FIG. 3) in zebrafish embryos. From these data, collected at a single time point 24 hours after embryo fertilization, it was possible to determine the shared lineage history of tens of transcriptionally defined cell states. Notably, one set of structures in the embryo, known as the pharyngeal arches, could be seen to arise from different clonal origins, despite appearing transcriptionally similar. These structures arise from either neural crest or lateral plate mesoderm[47]. Once the origin of the cells was established, it became possible to identify genes whose expression in the pharyngeal arches was specific to the crest-derived cells[47]. In a second example, CRISPR–Cas9 barcoding using the ScarTrace system revealed that the zebrafish fin harbours resident immune cells (RICs) with an ontogeny distinct from that of other immune cells[86]. These experiments could reveal precisely which cells were RICs amongst all immune cells in the fin, and they defined *Epcam* as a putative marker for this population. In yet a third example, Chan et al.[85] used cumulative Cas9 editing to study the ontogeny of endodermal tissues in the mouse embryo. These tissues are known to comprise a mixture of visceral and epiblast-derived cells[117]. Chan et al. could resolve between the visceral and epiblast lineages, despite their converging onto similar endodermal gene expression programs. The researchers could then identify differences between the two endodermal lineages in the expression of two genes: *Rhox5* and *Trap1a*. The ubiquity of converging trajectories has been further supported by complementary observations in the mouse extra-embryonic endoderm[35,51], in *C. elegans* embryogenesis[40] and in the parallel progression of excitatory and inhibitory neuronal states in the mouse central nervous system[50]. Collectively, these findings highlight a recurring phenomenon that these methods are particularly suited to address: they resolve different clonal origins among identical or nearly identical cell states (FIG. 2c–e), and they can reveal features of a cellular transcriptome (however subtle) that correlate with lineage behaviour and could be used to label or isolate cell subsets for further study.

Clonal resampling on a state manifold

Recently, several groups have utilized an alternative approach for linking detailed cell states across time. The approach relies on 'clonal resampling': experimentally isolating part of a clone for single-cell transcriptomic analysis recurrently, as the clone differentiates. When scaled to large numbers of cells, this method facilitates the construction of state manifolds on which the trajectories of individual clones may be revealed (FIG. 5Ac). This method requires both that cells be sampled over time without excessively disrupting the behaviour of the surviving cells and that cells divide symmetrically, such that all cells within a clone initially possess similar states. Due to these requirements, this method is best applied to either in vitro systems or regenerative systems in which cells or tissues may be serially removed or transplanted. Early realizations of this approach have been applied in culture, in which individual clones of related cells can be physically split, grown and sampled independently. For example, Tian et al.[118] recently applied this approach to analysing dendritic cell clones derived from single haematopoietic stem cells cultured and assayed in vitro. By physically splitting small clones of cells into separate culture wells, they were able to perform two distinct types of measurement on the clonal 'sister' cells: scRNA-seq at an early time point, to establish the transcriptional features of each clone before differentiation, and in vitro assays, to establish the ability of the same clone to generate three distinct types of differentiated dendritic cell population. This approach, which the researchers termed 'SIS-seq', was able to reveal rich transcriptional features of early progenitor cells that were predictive of the later fate outcomes.

More recent applications of this approach have relied on DNA barcoding, rather than physical isolation, to simultaneously track large numbers of cell clones. In an instructive example, Biddy et al.[114] developed a method, 'CellTagging', to trace the state of cells undergoing direct reprogramming from fibroblast to endoderm progenitors in vitro during serial rounds of passaging. The researchers made use of a lentiviral library to genetically barcode cells by integration of a constitutively expressed GFP-encoding gene with random barcodes engineered into its 3′ untranslated region sequence. During serial passaging, they applied additional rounds of lentiviral barcoding to mark successive lineage restriction events and simultaneously sampled subsets of the growing culture for scRNA-seq analysis. From this analysis, they identified that successful lineage conversion observed late in the reprogramming process correlated with a distinct expression profile of clonally related cells at an earlier time point. Such correlative analyses raise hypotheses for genes whose early expression influences future cell behaviours. In the study, Biddy et al. found that incorporating one such predictor gene, *Mettl7a1*, into the reprogramming procedure increased the efficiency of generating endodermal progenitors. Crucially, in this study neither the initial, transient nor end state of the dynamics had to be resolved in advance, and no marker genes were required to label cells for lineage tracing. A similar logic was applied by Weinreb et al.[84], who also used a lentiviral DNA-barcoding approach to demarcate fate boundaries in haematopoietic progenitor cell differentiation in order to link early biases in gene expression to later fate potential.

Clonal resampling thus offers a powerful approach to fully integrate state manifolds with lineage tracing and can be used to identify prospective fate markers. This approach has been most thoroughly applied in vitro, but it can also be used to interrogate *in vivo* systems that permit physical resampling, such as the haematopoietic system[84] and the regenerative zebrafish fin[86]. A persisting challenge in studying *in vivo* systems in this way is the need to obtain sufficient statistical sampling of each clone of interest, which can be difficult when isolating and sequencing cells from large endogenous populations.

Computational tools for state–lineage mapping

Computational approaches to analyse combined lineage and state datasets are still in their infancy. They are likely to evolve considerably and to require steps that are sensitive to the choice of experimental platform. As choices in data analysis could affect the conclusions drawn from such methods, we briefly review the key steps here.

The first step for DNA-barcoding pipelines is to assign a unique DNA barcode sequence to each cell clone. In doing so, pipelines must eliminate putative sequencing errors, remove cell doublets that could lead to two clonal barcodes appearing in one cell, and correct platform-specific artefacts. For the CRISPR-based and PolyLox methods, some barcodes can be formed with high probability, leading to frequent barcode homoplasy. Computational pipelines must therefore decide which barcodes are informative and which must be discarded from the analysis. Current methods are naive to recombination or error preferences of DNA-modifying enzymes; future methods could learn and incorporate editing biases to correct for observed barcode frequencies.

Computational pipelines then face decisions about how to reconstruct lineage phylogenies from large sets of clonal barcodes. In some cases, tree-building methods established for evolutionary phylogenetics have been applied directly to lineage reconstruction efforts; for example, GESTALT studies have utilized maximum parsimony[98,101,119], whereas homing CRISPR and TracerSeq have utilized neighbour-joining methods[47,100]. However, previously established tree-building methods are not necessarily robust to the frequent detection errors encountered in single-cell measurements[102]. LINNAEUS[102] and Chan et al.[85] have therefore developed custom tree-building algorithms to minimize the influence of drop-outs and have also incorporated empirical likelihood estimates for each barcode in order to minimize the influence of barcode homoplasy on the final inferred tree topology. Inference of lineage relationships from DNA-barcoding data is an active area of research, with several additional groups now favouring maximum-likelihood approaches and ground-truth benchmarking of algorithm performance against empirical[120–122] or simulated[123] datasets.

At present, no universal computational tools exist for end-to-end lineage tree inference, starting from raw single-cell DNA barcode sequences. Given the wide diversity of DNA modification strategies, barcode lengths and barcode probability distributions, the development of a single universal tool might be unlikely. However, for CRISPR–Cas9 editing systems, in particular, community benchmarking efforts such as the DREAM challenge[124] are now providing opportunities to directly compare the performance of dozens to hundreds of independent algorithms. Standardization of metrics and input data types could further enable meta-approaches that draw results from the consensus of multiple different tools. Because all lineage-barcoding methods — including non-CRISPR–Cas9 methods — face similar downstream analysis challenges (for example, tree building, the analysis of large tree ensembles and increasing dataset sizes), the field as a whole will undoubtedly benefit from these and other computational innovations.

In addition to tree construction, there are also other — perhaps simpler — data representations that can reveal intuitive lineage–state relationships. Rather than focusing on the structure of individual lineage trees, one can instead integrate information from multiple trees so as to infer the average lineage relationships between cell states. For many organisms and tissues, such approaches may be crucial, because individual lineage trees can be highly variable. Various metrics can be used for establishing lineage coupling between states, including the covariance of barcode abundances between states or the ratio of the barcodes observed to be shared between two transcriptional states to that expected after data randomization[47,84] (FIG. 5B). Maximum-likelihood frameworks can similarly be leveraged so as to combine individual lineage trees into 'consensus' lineage trees by integrating gene expression and lineage data[122]. This approach permits the integration of information across biological specimens, to separate core systematic trends from chance relationships that occur in just a single lineage tree.

Pitfalls in lineage barcoding on a state manifold

New biological assays can generate unforeseen artefacts that often become appreciated only after technologies mature. In the case of sequencing-based lineage tracing, the details of an experimental design can profoundly affect the relationships encoded in sequencing data. In FIG. 5 we have detailed two parameters that can strongly influence the observed clonal overlaps between states. These include the effects of the timing of barcode induction (FIG. 5B–E) and of changes in endogenous cell division rates (FIG. 5F,G). Altering these parameters can lead to strong differences in apparent lineage structure by affecting both the presence and size — that is, the detectability — of the marked clones. Other barcode detection errors, including both type I and type II errors (FIG. 3C), can similarly interfere with lineage reconstruction efforts. Although the negative effects of detection errors can be minimized by means of certain tree reconstruction algorithms (for example, maximum parsimony), the frequency of such errors for a particular method should be quantified, and minimized wherever possible. Good experimental practices should further ensure that biological conclusions are robust to such errors, including through performing adequate biological/technical replicates and the use of multiple data analysis strategies.

Emerging concepts

State trajectories and lineage codify two distinctive yet complementary aspects of a cell's developmental history, and each type of analysis can provide insights into ontogeny and gene regulation. In this Review we have outlined some important limitations of state manifolds, and we have described the motivation and tools for integrating bona fide lineage measurements with single-cell omics. From the early application of these methods, we propose to highlight three emerging concepts: first, state manifolds as models; second, the modes of coupling of cell state bifurcation with cell division; and third, the validity of trees as descriptions of cell differentiation hierarchies.

State manifolds as models

In this Review we have raised the contradictions that can appear between lineage and state representations (FIG. 2) and discussed how clonal information could be used to clarify such developmental relationships. These contradictions demonstrate that representations of state manifolds are not infallible — rather, they are data-driven models that follow from particular sets of assumptions and data-processing criteria. Currently, most state manifolds are constructed in an unbiased fashion from the most prominent sources of covariation in the original state measurement. Under this practice, the defining features of an scRNA-seq manifold will reflect robust, variable transcriptional signatures and thus are not guaranteed to emphasize cell fate decisions, which might correlate with small sets of regulatory genes expressed at low levels at the time that fate restrictions occur. State manifolds have until now been constructed without incorporating information from clonal data. However, state and lineage relationships need not remain in conflict: once information on lineage is established, it can be used to improve our methods for representing state manifolds. An immediate and simple use of lineage information, for example, is in identifying molecular markers of lineage-biased progenitor cell states. Indeed, novel fate markers have been inferred both from combined lineage and state phylogenetic experiments[85] and from clonal-resampling studies[84,114]. Lineage information could also be used to train algorithms in the construction of state manifolds in a way that avoids errors such as those in FIG. 2. Such actions demand a conceptual shift towards treating state manifolds as models of a particular set of high-dimensional gene expression features, rather than as absolute or universal references on which to overlay cell differentiation trajectories.

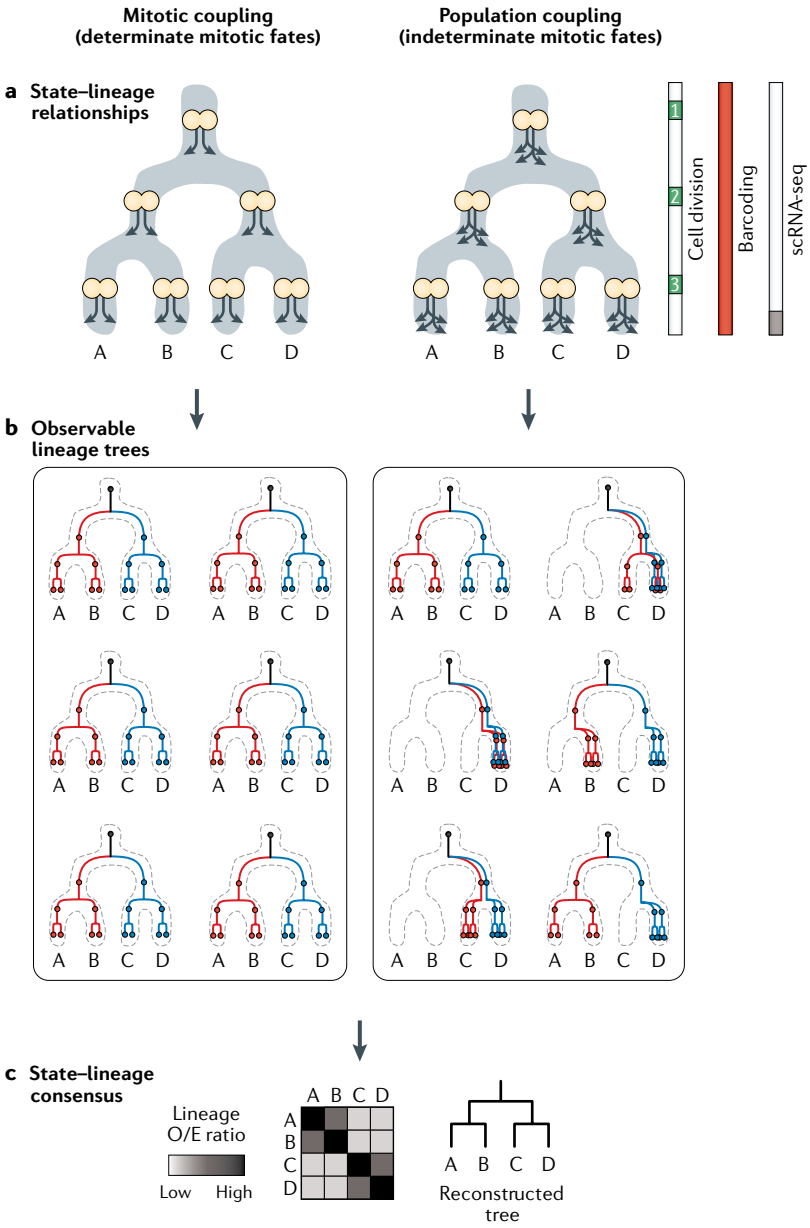


Fig. 6 | **Developmental paradigms that shape state–lineage relationships. a** | State manifold diagrams depicting the timing and fates of mitotic daughter cells. In cases of mitotic coupling (left), cells divide asymmetrically and give rise to distinct daughter states. In cases of population coupling (right), the average flux of cells down branches of the state manifold is maintained, but the fates of individual daughter cells are largely unpredictable. **b** | Examples of observable lineage trees that result from mitotic or population coupling. Mitotic coupling (left) leads to invariant, determinant lineage trees. Population coupling (right) permits a large number of observable lineage tree possibilities (six shown). **c** | Consensus relationships derived from a large number of individual tree observations. Despite the varied possibilities for the individual lineage trees in part **b**, the lineage relationships between states will be similar for both mitotic- and population-coupling scenarios. The heat map plots lineage observed/expected (O/E) ratios (see the FIG. 5 legend and Weinreb et al. (2020)[84] for the definition). scRNA-seq, single-cell RNA sequencing.

Variability of individual lineage trees

Both state manifolds and mitotic lineage trees can define hierarchies. What is the nature of the relationship between these two hierarchies? Drawing on lessons from imaging-based clonal analysis[125], we propose two potential relationships: one of mitotic coupling, and another of population coupling. Mitotic coupling will occur in cases in which a branch-point identified on the cell state manifold closely corresponds to a cell division event (FIG. 6a, left). Determinate lineage trees of ascidians[87] and *C. elegans*[88] stand as instructive examples. Population coupling, by contrast, will occur in cases in which the clonal and division histories do not influence the progression of any individual cell along the manifold or its fate choice. Instead, cell behaviours are indeterminate and can be described by a set of transition probabilities for moving down a particular trajectory (FIG. 6a, right). Accordingly, population coupling can lead to highly variable lineage trees that resemble those from a stochastic branching process and that will not be precisely reproducible within or between organisms (FIG. 6b, right). In such cases, efforts towards high-resolution reconstruction of fate hierarchies may fail to produce a single representative lineage tree of development, but the distribution of state–lineage couplings across multiple observed lineage trees should nonetheless prove highly informative (FIG. 6c).

Is development a tree?

What is the structure of a differentiation hierarchy? Answers to this question depend first on whether one is considering a state manifold or a mitotic lineage.

Mitotic coupling
A class of developmental fate regulation mechanisms that specify states to the daughter cells of a mitotic division, either symmetrically or asymmetrically.

Population coupling
A class of developmental fate regulation mechanisms in which the cell state specification is uncoupled from cell division but the proportion of cells specified to each state is controlled.

State divergence
A scenario in which the asymmetric partitioning of cellular components between two daughters of a single cell division differentiates them rapidly or instantaneously into distinct states.

In the absence of cell fusion, lineages can generally be treated as bifurcating trees, with each branch-point representing a mitotic event. State manifolds can be tree-like but, depending on the biology of the system, they need not be. State manifolds therefore represent an opportunity to discover the structure (that is, the topology) of a cell differentiation process. When state manifolds are integrated with lineage measurements, one has an opportunity to independently reject or confirm specific hypotheses regarding these structures. As we described above, several recent studies have shown evidence for state convergence, in which two or more distinct fate trajectories converge onto the same final position on a state manifold. This end point state thus comprises cells of mixed origins, which may or may not retain distinct functions or potentials. We reviewed examples of state convergence among immune cells[86], neural crest lineages[47] and endodermal populations[35,85]. The reverse scenario (state divergence) has also been observed, in which mitotic sister cells (highly related in lineage) rapidly adopt discontinuous states[40]. State divergence can occur as a result of asymmetric cell division, particularly in cases in which partitioned cytoplasmic components are delivered to only one of the two mitotic daughter cells. Such cases may produce state transitions that lack intermediate states and that thus would not appear as a bifurcation event on a state manifold, at any sampling depth. Both of these scenarios — convergence and divergence — will cause a state manifold to depart from a strict tree structure and can result from well-described biological scenarios. Mapping novel examples of such scenarios from single-cell datasets will therefore require integrated state and lineage measurements.

Conclusions

With the emergence of genome-scale single-cell analyses, representations of differentiation dynamics have shifted in the span of a few years from cartoons of discrete state transitions to data-driven views of dynamic state manifolds. Such representations provide not just predictions for the differentiation dynamics of thousands of genes but also hypotheses for the structure of differentiation hierarchies, including novel transitional and terminal cell states, interactions with cell cycle and the appearance of convergent differentiation that takes the form of 'loops' between cell states. In this Review we described the errors and ambiguities that can arise in inferring dynamics directly from single-cell state measurements, and we argued that the integration of lineage-barcoding data can improve state manifold representations by facilitating a faithful reconstruction of dynamics. Such integrative measurements can identify prospective fate markers, localize fate boundaries on state manifolds, allow the inference of tree-like and non-tree-like differentiation hierarchies, and should allow for resolving consensus fate relationships even when the individual lineage trees are highly variable. We thus anticipate that integrated measurements of cell state and lineage will greatly clarify the key events in cellular differentiation and become an important tool in the arsenal of stem cell, tissue and developmental biologists.

Published online 31 March 2020

1. Whitman, C. O. Memoirs: the embryology of clepsine. *J. Cell Sci.* **s2-18**, 215–315 (1878).
2. Waddington, C. H. *The strategy of the genes. A discussion of some aspects of theoretical biology.* (George Allen & Unwin, Ltd., London, 1957).
3. Saelens, W., Cannoodt, R., Todorov, H. & Saeys, Y. A comparison of single-cell trajectory inference methods. *Nat. Biotechnol.* **37**, 547–554 (2019).
4. Tritschler, S. et al. Concepts and limitations for learning developmental trajectories from single cell genomics. *Development* **146**, dev170506 (2019).
5. McKenna, A. & Gagnon, J. A. Recording development with single cell dynamic lineage tracing. *Development* **146**, dev169730 (2019).
6. Kester, L. & van Oudenaarden, A. Single-cell transcriptomics meets lineage tracing. *Cell Stem Cell* **23**, 166–179 (2018).
7. Klein, A. M. et al. Droplet barcoding for single-cell transcriptomics applied to embryonic stem cells. *Cell* **161**, 1187–1201 (2015).
8. Macosko, E. Z. et al. Highly parallel genome-wide expression profiling of individual cells using nanoliter droplets. *Cell* **161**, 1202–1214 (2015).
9. Gierahn, T. M. et al. Seq-Well: portable, low-cost RNA sequencing of single cells at high throughput. *Nat. Methods* **14**, 395–398 (2017).
10. Cusanovich, D. A. et al. Multiplex single cell profiling of chromatin accessibility by combinatorial cellular indexing. *Science* **348**, 910–914 (2015).
11. Wagner, A., Regev, A. & Yosef, N. Revealing the vectors of cellular identity with single-cell genomics. *Nat. Biotechnol.* **34**, 1145–1160 (2016).
12. Kotliar, D. et al. Identifying gene expression programs of cell-type identity and cellular activity with single-cell RNA-Seq. *eLife* **8**, e43803 (2019).
13. Lareau, C. A. et al. Droplet-based combinatorial indexing for massive-scale single-cell chromatin accessibility. *Nat. Biotechnol.* **37**, 916–924 (2019).
14. Mezger, A. et al. High-throughput chromatin accessibility profiling at single-cell resolution. *Nat. Commun.* **9**, 3647 (2018).
15. Karemaker, I. D. & Vermeulen, M. Single-cell DNA methylation profiling: technologies and biological applications. *Trends Biotechnol.* **36**, 952–965 (2018).
16. Budnik, B., Levy, E., Harmange, G. & Slavov, N. SCoPE-MS: mass spectrometry of single mammalian cells quantifies proteome heterogeneity during cell differentiation. *Genome Biol.* **19**, 161 (2018).
17. Duncan, K. D., Fyrestam, J. & Lanekoff, I. Advances in mass spectrometry based single-cell metabolomics. *Analyst* **144**, 782–793 (2019).
18. Stoeckius, M. et al. Simultaneous epitope and transcriptome measurement in single cells. *Nat. Methods* **14**, 865–868 (2017).
19. Mimitou, E. P. et al. Multiplexed detection of proteins, transcriptomes, clonotypes and CRISPR perturbations in single cells. *Nat. Methods* **16**, 409–412 (2019).
20. Peterson, V. M. et al. Multiplexed quantification of proteins and transcripts in single cells. *Nat. Biotechnol.* **35**, 936–939 (2017).
21. Dey, S. S., Kester, L., Spanjaard, B., Bienko, M. & van Oudenaarden, A. Integrated genome and transcriptome sequencing of the same cell. *Nat. Biotechnol.* **33**, 285–289 (2015).
22. Han, K. Y. et al. SIDR: simultaneous isolation and parallel sequencing of genomic DNA and total RNA from single cells. *Genome Res.* **28**, 75–87 (2018).
23. Lubeck, E., Coskun, A. F., Zhiyentayev, T., Ahmad, M. & Cai, L. Single-cell in situ RNA profiling by sequential hybridization. *Nat. Methods* **11**, 360–361 (2014).
24. Eng, C. L. et al. Transcriptome-scale super-resolved imaging in tissues by RNA seqFISH. *Nature* **568**, 235–239 (2019).
25. Rodriques, S. G. et al. Slide-seq: a scalable technology for measuring genome-wide expression at high spatial resolution. *Science* **363**, 1463–1467 (2019).
26. Chen, K. H., Boettiger, A. N., Moffitt, J. R., Wang, S. & Zhuang, X. Spatially resolved, highly multiplexed RNA profiling in single cells. *Science* **348**, aaa6090 (2015).
27. Lee, J. H. et al. Highly multiplexed subcellular RNA sequencing in situ. *Science* **343**, 1360–1363 (2014).
28. Tusi, B. K. et al. Population snapshots predict early haematopoietic and erythroid hierarchies. *Nature* **555**, 54–60 (2018).
29. Tikhonova, A. N. et al. The bone marrow microenvironment at single-cell resolution. *Nature* **569**, 222–228 (2019).
30. Montoro, D. T. et al. A revised airway epithelial hierarchy includes CFTR-expressing ionocytes. *Nature* **560**, 319–324 (2018).
31. Plasschaert, L. W. et al. A single-cell atlas of the airway epithelium reveals the CFTR-rich pulmonary ionocyte. *Nature* **560**, 377–381 (2018).
32. Park, J. et al. Single-cell transcriptomics of the mouse kidney reveals potential cellular targets of kidney disease. *Science* **360**, 758–763 (2018).
33. Young, M. D. et al. Single-cell transcriptomes from human kidneys reveal the cellular identity of renal tumors. *Science* **361**, 594–599 (2018).
34. de Soysa, T. Y. et al. Single-cell analysis of cardiogenesis reveals basis for organ-level developmental defects. *Nature* **572**, 120–124 (2019).
35. Nowotschin, S. et al. The emergent landscape of the mouse gut endoderm at single-cell resolution. *Nature* **569**, 361–367 (2019).
The authors analyse >100,000 single-cell transcriptomes from developing mouse endoderm and describe the convergence of visceral and definitive lineages into spatially defined transcriptional states.
36. Diaz-Cuadros, M. et al. In vitro characterization of the human segmentation clock. *Nature* https://doi.org/10.1038/s41586-019-1885-9 (2020).
37. Zeisel, A. et al. Molecular architecture of the mouse nervous system. *Cell* **174**, 999–1014.e1022 (2018).
38. Soldatov, R. et al. Spatiotemporal structure of cell fate decisions in murine neural crest. *Science* **364**, eaas9536 (2019).
39. Cao, J. et al. Comprehensive single-cell transcriptional profiling of a multicellular organism. *Science* **357**, 661–667 (2017).

40. Packer, J. S. et al. A lineage-resolved molecular atlas of *C. elegans* embryogenesis at single-cell resolution. *Science* **365**, eaax1971 (2019).
The authors interrogate the temporal dynamics of lineage–state relationships, as well as transcriptional convergence and divergence, in the invariant *C. elegans* embryonic lineage.
41. Sebé-Pedrós, A. et al. Cnidarian cell type diversity and regulation revealed by whole-organism single-cell RNA-seq. *Cell* **173**, 1520–1534.e1520 (2018).
42. Siebert, S. et al. Stem cell differentiation trajectories in *Hydra* resolved at single-cell resolution. *Science* **365**, eaav9314 (2019).
43. Achim, K. et al. Whole-body single-cell sequencing reveals transcriptional domains in the annelid larval body. *Mol. Biol. Evol.* **35**, 1047–1062 (2018).
44. Zeng, A. et al. Prospectively isolated tetraspanin(+) neoblasts are adult pluripotent stem cells underlying planaria regeneration. *Cell* **173**, 1593–1608.e1520 (2018).
45. Fincher, C. T., Wurtzel, O., de Hoog, T., Kravarik, K. M. & Reddien, P. W. Cell type transcriptome atlas for the planarian *Schmidtea mediterranea*. *Science* **360**, eaaq1736 (2018).
46. Plass, M. et al. Cell type atlas and lineage tree of a whole complex animal by single-cell transcriptomics. *Science* **360**, eaaq1723 (2018).
47. Wagner, D. E. et al. Single-cell mapping of gene expression landscapes and lineage in the zebrafish embryo. *Science* **360**, 981–987 (2018).
The authors describe non-tree-like cell state trajectories using combined lineage barcoding and single-cell transcriptomics in zebrafish embryos.
48. Farrell, J. A. et al. Single-cell reconstruction of developmental trajectories during zebrafish embryogenesis. *Science* **360**, eaar3131 (2018).
49. Briggs, J. A. et al. The dynamics of gene expression in vertebrate embryogenesis at single-cell resolution. *Science* **360**, eaar5780 (2018).
50. Cao, J. et al. The single-cell transcriptional landscape of mammalian organogenesis. *Nature* **566**, 496–502 (2019).
This study presents the largest single-cell transcriptome atlas for mouse embryogenesis to date, spanning >2 million cells and 56 cell state trajectories.
51. Pijuan-Sala, B. et al. A single-cell molecular map of mouse gastrulation and early organogenesis. *Nature* **566**, 490–495 (2019).
52. Karaiskos, N. et al. The *Drosophila* embryo at single-cell transcriptome resolution. *Science* **358**, 194–199 (2017).
53. Cao, C. et al. Comprehensive single-cell transcriptome lineages of a proto-vertebrate. *Nature* **571**, 349–354 (2019).
This study performs comprehensive single-cell profiling of ascidian embryos from the early gastrula to larval stages and maps the transcriptomic signatures onto a virtual map of the determinate embryonic lineage tree.
54. Becht, E. et al. Dimensionality reduction for visualizing single-cell data using UMAP. *Nat. Biotechnol.* https://doi.org/10.1038/nbt.4314 (2018).
55. Weinreb, C., Wolock, S. & Klein, A. M. SPRING: a kinetic interface for visualizing high dimensional single-cell expression data. *Bioinformatics* **34**, 1246–1248 (2018).
56. Jacomy, M., Venturini, T., Heymann, S. & Bastian, M. ForceAtlas2, a continuous graph layout algorithm for handy network visualization designed for the Gephi software. *PLoS One* **9**, e98679 (2014).
57. Qiu, X. et al. Reversed graph embedding resolves complex single-cell trajectories. *Nat. Methods* **14**, 979–982 (2017).
58. Trapnell, C. et al. The dynamics and regulators of cell fate decisions are revealed by pseudotemporal ordering of single cells. *Nat. Biotechnol.* **32**, 381–386 (2014).
59. Wolf, F. A. et al. PAGA: graph abstraction reconciles clustering with trajectory inference through a topology preserving map of single cells. *Genome Biol.* **20**, 59 (2019).
This study presents 'PAGA', a graph-based computational approach for mapping non-tree-like topologies in single-cell state landscapes.
60. Setty, M. et al. Wishbone identifies bifurcating developmental trajectories from single-cell data. *Nat. Biotechnol.* **34**, 637–645 (2016).
61. Shin, J. et al. Single-Cell RNA-Seq with Waterfall reveals molecular cascades underlying adult neurogenesis. *Cell Stem Cell* **17**, 360–372 (2015).
62. Haghverdi, L., Buttner, M., Wolf, F. A., Buettner, F. & Theis, F. J. Diffusion pseudotime robustly reconstructs lineage branching. *Nat. Methods* **13**, 845–848 (2016).
63. Bendall, S. C. et al. Single-cell trajectory detection uncovers progression and regulatory coordination in human B cell development. *Cell* **157**, 714–725 (2014).
64. Weinreb, C., Wolock, S., Tusi, B. K., Socolovsky, M. & Klein, A. M. Fundamental limits on dynamic inference from single-cell snapshots. *Proc. Natl Acad. Sci. USA* **115**, E2467–E2476 (2018).
This is one of several studies to provide a framework for predicting fate trajectories from single-cell state manifolds.
65. Herman, J. S., Sagar & Grün, D. FateID infers cell fate bias in multipotent progenitors from single-cell RNA-seq data. *Nat. Methods* **15**, 379–386 (2018).
66. Setty, M. et al. Characterization of cell fate probabilities in single-cell data with Palantir. *Nat. Biotechnol.* **37**, 451–460 (2019).
67. Schiebinger, G. et al. Optimal-transport analysis of single-cell gene expression identifies developmental trajectories in reprogramming. *Cell* **176**, 1517 (2019).
68. Furchtgott, L. A., Melton, S., Menon, V. & Ramanathan, S. Discovering sparse transcription factor codes for cell states and state transitions during development. *eLife* **6**, e20488 (2017).
69. La Manno, G. et al. RNA velocity of single cells. *Nature* **560**, 494–498 (2018).
70. Hendriks, G.-J. et al. NASC-seq monitors RNA synthesis in single cells. *Nat. Commun.* **10**, 3138 (2019).
71. Erhard, F. et al. scSLAM-seq reveals core features of transcription dynamics in single cells. *Nature* **571**, 419–423 (2019).
72. Gorin, G., Svensson, V. & Pachter, L. Protein velocity and acceleration from single-cell multiomics experiments. *Genome Biol.* **21**, 39 (2019).
73. Qiu, X. et al. Mapping vector field of single cells. Preprint at https://doi.org/10.1101/696724 (2019).
74. Haghverdi, L., Buettner, F. & Theis, F. J. Diffusion maps for high-dimensional single-cell analysis of differentiation data. *Bioinformatics* **31**, 2989–2998 (2015).
75. Coifman, R. R. et al. Geometric diffusions as a tool for harmonic analysis and structure definition of data: diffusion maps. *Proc. Natl Acad. Sci. USA* **102**, 7426–7431 (2005).
76. Chen, H. et al. Single-cell trajectories reconstruction, exploration and mapping of omics data with STREAM. *Nat. Commun.* **10**, 1903 (2019).
77. Traag, V. A., Waltman, L. & van Eck, N. J. From Louvain to Leiden: guaranteeing well-connected communities. *Sci. Rep.* **9**, 5233 (2019).
78. Blondel, V. D., Guillaume, J.-L., Lambiotte, R. & Lefebvre, E. Fast unfolding of communities in large networks. *J. Stat. Mech.* https://doi.org/10.1088/1742-5468/2008/10/P10008 (2008).
79. Kimmel, C. B., Warga, R. M. & Schilling, T. F. Origin and organization of the zebrafish fate map. *Development* **108**, 581–594 (1990).
80. Kretzschmar, K. & Watt, F. M. Lineage tracing. *Cell* **148**, 33–45 (2012).
81. Lodato, M. A. et al. Somatic mutation in single human neurons tracks developmental and transcriptional history. *Science* **350**, 94–98 (2015).
82. Wagers, A. J. & Weissman, I. L. Plasticity of adult stem cells. *Cell* **116**, 639–648 (2004).
83. Wagers, A. J., Sherwood, R. I., Christensen, J. L. & Weissman, I. L. Little evidence for developmental plasticity of adult hematopoietic stem cells. *Science* **297**, 2256–2259 (2002).
84. Weinreb, C., Rodriguez-Fraticelli, A., Camargo, F. D. & Klein, A. M. Lineage tracing on transcriptional landscapes links state to fate during differentiation. *Science* **367**, eaaw3381 (2020).
The authors implement the 'LARRY' clonal resampling approach to map single-cell transcriptomes and lineage relationships in differentiating cells in the mouse haematopoietic system.
85. Chan, M. M. et al. Molecular recording of mammalian embryogenesis. *Nature* **570**, 77–82 (2019).
86. Alemany, A., Florescu, M., Baron, C. S., Peterson-Maduro, J. & van Oudenaarden, A. Whole-organism clone tracing using single-cell sequencing. *Nature* **556**, 108–112 (2018).
This study describes the Cas9-editing-based 'ScarTrace' method for simultaneous measurement of single-cell transcriptomes and lineage relationships in the zebrafish embryo and the regenerating fin of its adult form.
87. Conklin, E. G. The organization and cell lineage of the ascidian egg. *J. Acad. Nat. Sci. Phila.* **13**, 1–119 (1905).
88. Sulston, J. E., Schierenberg, E., White, J. G. & Thomson, J. N. The embryonic cell lineage of the nematode Caenorhabditis elegans. *Dev. Biol.* **100**, 64–119 (1983).
89. Keller, P. J., Schmidt, A. D., Wittbrodt, J. & Stelzer, E. H. Reconstruction of zebrafish early embryonic development by scanned light sheet microscopy. *Science* **322**, 1065–1069 (2008).
90. McDole, K. et al. In toto imaging and reconstruction of post-implantation mouse development at the single-cell level. *Cell* **175**, 859–876.e833 (2018).
91. Satija, R., Farrell, J. A., Gennert, D., Schier, A. F. & Regev, A. Spatial reconstruction of single-cell gene expression data. *Nat. Biotechnol.* **33**, 495–502 (2015).
92. Frieda, K. L. et al. Synthetic recording and in situ readout of lineage information in single cells. *Nature* **541**, 107–111 (2017).
93. Keller, G., Paige, C., Gilboa, E. & Wagner, E. F. Expression of a foreign gene in myeloid and lymphoid cells derived from multipotent haematopoietic precursors. *Nature* **318**, 149–154 (1985).
94. Lemischka, I. R., Raulet, D. H. & Mulligan, R. C. Developmental potential and dynamic behavior of hematopoietic stem cells. *Cell* **45**, 917–927 (1986).
95. Ludwig, L. S. et al. Lineage tracing in humans enabled by mitochondrial mutations and single-cell genomics. *Cell* **176**, 1325–1339.e1322 (2019).
96. Woodworth, M. B., Girskis, K. M. & Walsh, C. A. Building a lineage from single cells: genetic techniques for cell lineage tracking. *Nat. Rev. Genet.* **18**, 230–244 (2017).
97. Xu, J. et al. Single-cell lineage tracing by endogenous mutations enriched in transposase accessible mitochondrial DNA. *eLife* **8**, e45105 (2019).
98. McKenna, A. et al. Whole-organism lineage tracing by combinatorial and cumulative genome editing. *Science* **353**, aaf7907 (2016).
99. Kalhor, R., Mali, P. & Church, G. M. Rapidly evolving homing CRISPR barcodes. *Nat. Methods* **14**, 195–200 (2017).
100. Kalhor, R. et al. Developmental barcoding of whole mouse via homing CRISPR. *Science* **361**, eaat9804 (2018).
101. Raj, B. et al. Simultaneous single-cell profiling of lineages and cell types in the vertebrate brain. *Nat. Biotechnol.* **36**, 442–450 (2018).
This study combines the previously established Cas9-editing GESTALT approach for lineage barcoding with inDrops-based single-cell transcriptome analysis to reconstruct developmental trajectories in the zebrafish brain.
102. Spanjaard, B. et al. Simultaneous lineage tracing and cell-type identification using CRISPR–Cas9-induced genetic scars. *Nat. Biotechnol.* **36**, 469–473 (2018).
This study introduces 'LINNAEUS' and a network algorithm for reconstructing Cas9-editing-based lineage phylogenies between cell states of the 5-day-old zebrafish embryo.
103. Ihry, R. J. et al. p53 inhibits CRISPR-Cas9 engineering in human pluripotent stem cells. *Nat. Med.* **24**, 939–946 (2018).
104. Haapaniemi, E., Botla, S., Persson, J., Schmierer, B. & Taipale, J. CRISPR–Cas9 genome editing induces a p53-mediated DNA damage response. *Nat. Med.* **24**, 927–930 (2018).
105. Pei, W. et al. Polylox barcoding reveals haematopoietic stem cell fates realized *in vivo*. *Nature* **548**, 456–460 (2017).
106. Pei, W. et al. Using Cre-recombinase-driven Polylox barcoding for *in vivo* fate mapping in mice. *Nat. Protoc.* **14**, 1820–1840 (2019).
107. Klompe, S. E., Vo, P. L. H., Halpin-Healy, T. S. & Sternberg, S. H. Transposon-encoded CRISPR-Cas systems direct RNA-guided DNA integration. *Nature* **571**, 219–225 (2019).
108. Strecker, J. et al. RNA-guided DNA insertion with CRISPR-associated transposases. *Science* **365**, 48–53 (2019).
109. Hwang, B. et al. Lineage tracing using a Cas9-deaminase barcoding system targeting endogenous L1 elements. *Nat. Commun.* **10**, 1234 (2019).
110. Hess, G. T. et al. Directed evolution using dCas9-targeted somatic hypermutation in mammalian cells. *Nat. Methods* **13**, 1036–1042 (2016).
111. Grunewald, J. et al. Transcriptome-wide off-target RNA editing induced by CRISPR-guided DNA base editors. *Nature* **569**, 433–437 (2019).
112. Jin, S. et al. Cytosine, but not adenine, base editors induce genome-wide off-target mutations in rice. *Science* **364**, 292–295 (2019).
113. Zuo, E. et al. Cytosine base editor generates substantial off-target single-nucleotide variants in mouse embryos. *Science* **364**, 289–292 (2019).

114. Biddy, B. A. et al. Single-cell mapping of lineage and identity in direct reprogramming. *Nature* **564**, 219–224 (2018).
This study introduces the 'CellTag' clonal resampling method for retroviral barcoding of cell lineages with a combined single-cell transcriptomic readout.
115. Loveless, T. B. et al. Ordered insertional mutagenesis at a single genomic site enables lineage tracing and analog recording in mammalian cells. Preprint at https://doi.org/10.1101/639120 (2019).
116. Guo, C. et al. CellTag Indexing: genetic barcode-based sample multiplexing for single-cell genomics. *Genome Biol.* **20**, 90 (2019).
117. Kwon, G. S., Viotti, M. & Hadjantonakis, A.-K. The endoderm of the mouse embryo arises by dynamic widespread intercalation of embryonic and extraembryonic lineages. *Dev. Cell* **15**, 509–520 (2008).
118. Tian, L. et al. SIS-seq, a molecular 'time machine', connects single cell fate with gene programs. Preprint at https://doi.org/10.1101/403113 (2018).
119. Raj, B., Gagnon, J. A. & Schier, A. F. Large-scale reconstruction of cell lineages using single-cell readout of transcriptomes and CRISPR–Cas9 barcodes by scGESTALT. *Nat. Protoc.* **13**, 2685–2713 (2018).
120. Jones, M. G. et al. Inference of single-cell phylogenies from lineage tracing data. Preprint at https://doi.org/10.1101/800078 (2019).
121. Feng, J. et al. Estimation of cell lineage trees by maximum-likelihood phylogenetics. Preprint at https://doi.org/10.1101/595215 (2019).
122. Zafar, H., Lin, C. & Bar-Joseph, Z. Single-cell lineage tracing by integrating CRISPR–Cas9 mutations with transcriptomic data. Preprint at https://doi.org/10.1101/630814 (2019).
123. Salvador-Martínez, I., Grillo, M., Averof, M. & Telford, M. J. Is it possible to reconstruct an accurate cell lineage using CRISPR recorders? *eLife* **8**, e40292 (2019).
124. Synapse. Allen Institute Cell Lineage Reconstruction DREAM Challenge. *Sage Bionetworks* https://www.synapse.org/#!Synapse:syn20692755/wiki/595096 (2019).
125. Klein, A. M. & Simons, B. D. Universal patterns of stem cell fate in cycling adult tissues. *Development* **138**, 3103–3111 (2011).

Acknowledgements
The authors thank R. Ward and S. Mekhoubad for critical reading of the manuscript. D.E.W. is supported by grant R00GM121852.

Author contributions
The authors contributed equally to all aspects of the article.

Competing interests
A.M.K. is a founder of 1CellBio, Inc. D.E.W. declares no competing interests.

Peer review information
Nature Reviews Genetics thanks J. P. Junker and the other, anonymous, reviewer(s) for their contribution to the peer review of this work.

Publisher's note
Springer Nature remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.

© Springer Nature Limited 2020

ARTICLE

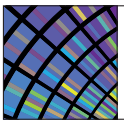
doi:10.1038/nature11247

An integrated encyclopedia of DNA elements in the human genome

The ENCODE Project Consortium*

The human genome encodes the blueprint of life, but the function of the vast majority of its nearly three billion bases is unknown. The Encyclopedia of DNA Elements (ENCODE) project has systematically mapped regions of transcription, transcription factor association, chromatin structure and histone modification. These data enabled us to assign biochemical functions for 80% of the genome, in particular outside of the well-studied protein-coding regions. Many discovered candidate regulatory elements are physically associated with one another and with expressed genes, providing new insights into the mechanisms of gene regulation. The newly identified elements also show a statistical correspondence to sequence variants linked to human disease, and can thereby guide interpretation of this variation. Overall, the project provides new insights into the organization and regulation of our genes and genome, and is an expansive resource of functional annotations for biomedical research.

The human genome sequence provides the underlying code for human biology. Despite intensive study, especially in identifying protein-coding genes, our understanding of the genome is far from complete, particularly with regard to non-coding RNAs, alternatively spliced transcripts and regulatory sequences. Systematic analyses of transcripts and regulatory information are essential for the identification of genes and regulatory regions, and are an important resource for the study of human biology and disease. Such analyses can also provide comprehensive views of the organization and variability of genes and regulatory information across cellular contexts, species and individuals.



The Encyclopedia of DNA Elements (ENCODE) project aims to delineate all functional elements encoded in the human genome[1–3]. Operationally, we define a functional element as a discrete genome segment that encodes a defined product (for example, protein or non-coding RNA) or displays a reproducible biochemical signature (for example, protein binding, or a specific chromatin structure). Comparative genomic studies suggest that 3–8% of bases are under purifying (negative) selection[4–8] and therefore may be functional, although other analyses have suggested much higher estimates[9–11]. In a pilot phase covering 1% of the genome, the ENCODE project annotated 60% of mammalian evolutionarily constrained bases, but also identified many additional putative functional elements without evidence of constraint[2]. The advent of more powerful DNA sequencing technologies now enables whole-genome and more precise analyses with a broad repertoire of functional assays.

Here we describe the production and initial analysis of 1,640 data sets designed to annotate functional elements in the entire human genome. We integrate results from diverse experiments within cell types, related experiments involving 147 different cell types, and all ENCODE data with other resources, such as candidate regions from genome-wide association studies (GWAS) and evolutionarily constrained regions. Together, these efforts reveal important features about the organization and function of the human genome, summarized below.

- The vast majority (80.4%) of the human genome participates in at least one biochemical RNA- and/or chromatin-associated event in at least one cell type. Much of the genome lies close to a regulatory event: 95% of the genome lies within 8 kilobases (kb) of a DNA–protein interaction (as assayed by bound ChIP-seq motifs or DNase I footprints), and 99% is within 1.7 kb of at least one of the biochemical events measured by ENCODE.
- Primate-specific elements as well as elements without detectable mammalian constraint show, in aggregate, evidence of negative selection; thus, some of them are expected to be functional.
- Classifying the genome into seven chromatin states indicates an initial set of 399,124 regions with enhancer-like features and 70,292 regions with promoter-like features, as well as hundreds of thousands of quiescent regions. High-resolution analyses further subdivide the genome into thousands of narrow states with distinct functional properties.
- It is possible to correlate quantitatively RNA sequence production and processing with both chromatin marks and transcription factor binding at promoters, indicating that promoter functionality can explain most of the variation in RNA expression.
- Many non-coding variants in individual genome sequences lie in ENCODE-annotated functional regions; this number is at least as large as those that lie in protein-coding genes.
- Single nucleotide polymorphisms (SNPs) associated with disease by GWAS are enriched within non-coding functional elements, with a majority residing in or near ENCODE-defined regions that are outside of protein-coding genes. In many cases, the disease phenotypes can be associated with a specific cell type or transcription factor.

ENCODE data production and initial analyses

Since 2007, ENCODE has developed methods and performed a large number of sequence-based studies to map functional elements across the human genome[3]. The elements mapped (and approaches used) include RNA transcribed regions (RNA-seq, CAGE, RNA-PET and manual annotation), protein-coding regions (mass spectrometry), transcription-factor-binding sites (ChIP-seq and DNase-seq), chromatin structure (DNase-seq, FAIRE-seq, histone ChIP-seq and MNase-seq), and DNA methylation sites (RRBS assay) (Box 1 lists methods and abbreviations; Supplementary Table 1, section P, details production statistics)[3]. To compare and integrate results across the different laboratories, data production efforts focused on two selected

*Lists of participants and their affiliations appear at the end of the paper.

©2012 Macmillan Publishers Limited. All rights reserved

BOX 1

ENCODE abbreviations

RNA-seq. Isolation of RNA sequences, often with different purification techniques to isolate different fractions of RNA followed by high-throughput sequencing.

CAGE. Capture of the methylated cap at the 5′ end of RNA, followed by high-throughput sequencing of a small tag adjacent to the 5′ methylated caps. 5′ methylated caps are formed at the initiation of transcription, although other mechanisms also methylate 5′ ends of RNA.

RNA-PET. Simultaneous capture of RNAs with both a 5′ methyl cap and a poly(A) tail, which is indicative of a full-length RNA. This is then followed by sequencing a short tag from each end by high-throughput sequencing.

ChIP-seq. Chromatin immunoprecipitation followed by sequencing. Specific regions of crosslinked chromatin, which is genomic DNA in complex with its bound proteins, are selected by using an antibody to a specific epitope. The enriched sample is then subjected to high-throughput sequencing to determine the regions in the genome most often bound by the protein to which the antibody was directed. Most often used are antibodies to any chromatin-associated epitope, including transcription factors, chromatin binding proteins and specific chemical modifications on histone proteins.

DNase-seq. Adaption of established regulatory sequence assay to modern techniques. The DNase I enzyme will preferentially cut live chromatin preparations at sites where nearby there are specific (non-histone) proteins. The resulting cut points are then sequenced using high-throughput sequencing to determine those sites 'hypersensitive' to DNase I, corresponding to open chromatin.

FAIRE-seq. Formaldehyde assisted isolation of regulatory elements. FAIRE isolates nucleosome-depleted genomic regions by exploiting the difference in crosslinking efficiency between nucleosomes (high) and sequence-specific regulatory factors (low). FAIRE consists of crosslinking, phenol extraction, and sequencing the DNA fragments in the aqueous phase.

RRBS. Reduced representation bisulphite sequencing. Bisulphite treatment of DNA sequence converts unmethylated cytosines to uracil. To focus the assay and save costs, specific restriction enzymes that cut around CpG dinucleotides can reduce the genome to a portion specifically enriched in CpGs. This enriched sample is then sequenced to determine the methylation status of individual cytosines quantitatively.

Tier 1. Tier 1 cell types were the highest-priority set and comprised three widely studied cell lines: K562 erythroleukaemia cells; GM12878, a B-lymphoblastoid cell line that is also part of the 1000 Genomes project (http://1000genomes.org)[55]; and the H1 embryonic stem cell (H1 hESC) line.

Tier 2. The second-priority set of cell types in the ENCODE project which included HeLa-S3 cervical carcinoma cells, HepG2 hepatoblastoma cells and primary (non-transformed) human umbilical vein endothelial cells (HUVECs).

Tier 3. Any other ENCODE cell types not in tier 1 or tier 2.

sets of cell lines, designated 'tier 1' and 'tier 2' (Box 1). To capture a broader spectrum of biological diversity, selected assays were also executed on a third tier comprising more than 100 cell types including primary cells. All data and protocol descriptions are available at http://www.encodeproject.org/, and a User's Guide including details of cell-type choice and limitations was published recently[3].

Integration methodology

For consistency, data were generated and processed using standardized guidelines, and for some assays, new quality-control measures were designed (see refs 3, 12 and http://encodeproject.org/ENCODE/dataStandards.html; A. Kundaje, personal communication). Uniform data-processing methods were developed for each assay (see Supplementary Information; A. Kundaje, personal communication), and most assay results can be represented both as signal information (a per-base estimate across the genome) and as discrete elements (regions computationally identified as enriched for signal). Extensive processing pipelines were developed to generate each representation (M. M. Hoffman *et al.*, manuscript in preparation and A. Kundaje, personal communication). In addition, we developed the irreproducible discovery rate (IDR)[13] measure to provide a robust and conservative estimate of the threshold where two ranked lists of results from biological replicates no longer agree (that is, are irreproducible), and we applied this to defining sets of discrete elements. We identified, and excluded from most analyses, regions yielding untrustworthy signals likely to be artefactual (for example, multicopy regions). Together, these regions comprise 0.39% of the genome (see Supplementary Information). The poster accompanying this issue represents different ENCODE-identified elements and their genome coverage.

Transcribed and protein-coding regions

We used manual and automated annotation to produce a comprehensive catalogue of human protein-coding and non-coding RNAs as well as pseudogenes, referred to as the GENCODE reference gene set[14,15] (Supplementary Table 1, section U). This includes 20,687 protein-coding genes (GENCODE annotation, v7) with, on average, 6.3 alternatively spliced transcripts (3.9 different protein-coding transcripts) per locus. In total, GENCODE-annotated exons of protein-coding genes cover 2.94% of the genome or 1.22% for protein-coding exons. Protein-coding genes span 33.45% from the outermost start to stop codons, or 39.54% from promoter to poly(A) site. Analysis of mass spectrometry data from K562 and GM12878 cell lines yielded 57 confidently identified unique peptide sequences in intergenic regions relative to GENCODE annotation. Taken together with evidence of pervasive genome transcription[16], these data indicate that additional protein-coding genes remain to be found.

In addition, we annotated 8,801 automatically derived small RNAs and 9,640 manually curated long non-coding RNA (lncRNA) loci[17]. Comparing lncRNAs to other ENCODE data indicates that lncRNAs are generated through a pathway similar to that for protein-coding genes[17]. The GENCODE project also annotated 11,224 pseudogenes, of which 863 were transcribed and associated with active chromatin[18].

RNA

We sequenced RNA[16] from different cell lines and multiple subcellular fractions to develop an extensive RNA expression catalogue. Using a conservative threshold to identify regions of RNA activity, 62% of genomic bases are reproducibly represented in sequenced long (>200 nucleotides) RNA molecules or GENCODE exons. Of these bases, only 5.5% are explained by GENCODE exons. Most transcribed bases are within or overlapping annotated gene boundaries (that is, intronic), and only 31% of bases in sequenced transcripts were intergenic[16].

We used CAGE-seq (5′ cap-targeted RNA isolation and sequencing) to identify 62,403 transcription start sites (TSSs) at high confidence (IDR of 0.01) in tier 1 and 2 cell types. Of these, 27,362 (44%) are within 100 base pairs (bp) of the 5′ end of a GENCODE-annotated transcript or previously reported full-length messenger RNA. The remaining regions predominantly lie across exons and 3′ untranslated regions (UTRs), and some exhibit cell-type-restricted expression; these may represent the start sites of novel, cell-type-specific transcripts.

Finally, we saw a significant proportion of coding and non-coding transcripts processed into steady-state stable RNAs shorter than 200 nucleotides. These precursors include transfer RNA, microRNA, small nuclear RNA and small nucleolar RNA (tRNA, miRNA, snRNA and snoRNA, respectively) and the 5′ termini of these processed products align with the capped 5′ end tags[16].

©2012 Macmillan Publishers Limited. All rights reserved

Table 1 | Summary of transcription factor classes analysed in ENCODE

Acronym	Description	Factors analysed
ChromRem	ATP-dependent chromatin complexes	5
DNARep	DNA repair	3
HISase	Histone acetylation, deacetylation or methylation complexes	8
Other	Cyclin kinase associated with transcription	1
Pol2	Pol II subunit	1 (2 forms)
Pol3	Pol III-associated	6
TFNS	General Pol II-associated factor, not site-specific	8
TFSS	Pol II transcription factor with sequence-specific DNA binding	87

Protein bound regions

To identify regulatory regions directly, we mapped the binding locations of 119 different DNA-binding proteins and a number of RNA polymerase components in 72 cell types using ChIP-seq (Table 1, Supplementary Table 1, section N, and ref. 19); 87 (73%) were sequence-specific transcription factors. Overall, 636,336 binding regions covering 231 megabases (Mb; 8.1%) of the genome are enriched for regions bound by DNA-binding proteins across all cell types. We assessed each protein-binding site for enrichment of known DNA-binding motifs and the presence of novel motifs. Overall, 86% of the DNA segments occupied by sequence-specific transcription factors contained a strong DNA-binding motif, and in most (55%) cases the known motif was most enriched (P. Kheradpour and M. Kellis, manuscript in preparation).

Protein-binding regions lacking high or moderate affinity cognate recognition sites have 21% lower median scores by rank than regions with recognition sequences (Wilcoxon rank sum P value $<10^{-16}$). Eighty-two per cent of the low-signal regions have high-affinity recognition sequences for other factors. In addition, when ChIP-seq peaks are ranked by their concordance with their known recognition sequence, the median DNase I accessibility is twofold higher in the bottom 20% of peaks than in the upper 80% (genome structure correction (GSC)[20] P value $<10^{-16}$), consistent with previous observations[21–24]. We speculate that low signal regions are either lower-affinity sites[21] or indirect transcription-factor target regions associated through interactions with other factors (see also refs 25, 26).

We organized all the information associated with each transcription factor—including the ChIP-seq peaks, discovered motifs and associated histone modification patterns—in FactorBook (http://www.factorbook.org; ref. 26), a public resource that will be updated as the project proceeds.

DNase I hypersensitive sites and footprints

Chromatin accessibility characterized by DNase I hypersensitivity is the hallmark of regulatory DNA regions[27,28]. We mapped 2.89 million unique, non-overlapping DNase I hypersensitive sites (DHSs) by DNase-seq in 125 cell types, the overwhelming majority of which lie distal to TSSs[29]. We also mapped 4.8 million sites across 25 cell types that displayed reduced nucleosomal crosslinking by FAIRE, many of which coincide with DHSs. In addition, we used micrococcal nuclease to map nucleosome occupancy in GM12878 and K562 cells[30].

In tier 1 and tier 2 cell types, we identified a mean of 205,109 DHSs per cell type (at false discovery rate (FDR) 1%), encompassing an average of 1.0% of the genomic sequence in each cell type, and 3.9% in aggregate. On average, 98.5% of the occupancy sites of transcription factors mapped by ENCODE ChIP-seq (and, collectively, 94.4% of all 1.1 million transcription factor ChIP-seq peaks in K562 cells) lie within accessible chromatin defined by DNase I hotspots[29]. However, a small number of factors, most prominently heterochromatin-bound repressive complexes (for example, the TRIM28–SETDB1–ZNF274 complex[31,32] encoded by the *TRIM28*, *SETDB1* and *ZNF274* genes), seem to occupy a significant fraction of nucleosomal sites.

Using genomic DNase I footprinting[33,34] on 41 cell types we identified 8.4 million distinct DNase I footprints (FDR 1%)[25]. Our *de novo* motif discovery on DNase I footprints recovered ~90% of known transcription factor motifs, together with hundreds of novel evolutionarily conserved motifs, many displaying highly cell-selective occupancy patterns similar to major developmental and tissue-specific regulators.

Regions of histone modification

We assayed chromosomal locations for up to 12 histone modifications and variants in 46 cell types, including a complete matrix of eight modifications across tier 1 and tier 2. Because modification states may span multiple nucleosomes, which themselves can vary in position across cell populations, we used a continuous signal measure of histone modifications in downstream analysis, rather than calling regions (M. M. Hoffman *et al.*, manuscript in preparation; see http://code.google.com/p/align2rawsignal/). For the strongest, 'peak-like' histone modifications, we used MACS[35] to characterize enriched sites. Table 2 describes the different histone modifications, their peak characteristics, and a summary of their known roles (reviewed in refs 36–39).

Our data show that global patterns of modification are highly variable across cell types, in accordance with changes in transcriptional activity. Consistent with previous studies[40,41], we find that integration of the different histone modification information can be used systematically to assign functional attributes to genomic regions (see below).

DNA methylation

Methylation of cytosine, usually at CpG dinucleotides, is involved in epigenetic regulation of gene expression. Promoter methylation is typically associated with repression, whereas genic methylation correlates with transcriptional activity[42]. We used reduced representation bisulphite sequencing (RRBS) to profile DNA methylation quantitatively for an average of 1.2 million CpGs in each of 82 cell lines and tissues (8.6% of non-repetitive genomic CpGs), including CpGs in intergenic regions, proximal promoters and intragenic regions (gene bodies)[43], although it should be noted that the RRBS method preferentially targets CpG-rich islands. We found that 96% of CpGs exhibited differential methylation in at least one cell type or tissue

Table 2 | Summary of ENCODE histone modifications and variants

Histone modification or variant	Signal characteristics	Putative functions
H2A.Z	Peak	Histone protein variant (H2A.Z) associated with regulatory elements with dynamic chromatin
H3K4me1	Peak/region	Mark of regulatory elements associated with enhancers and other distal elements, but also enriched downstream of transcription starts
H3K4me2	Peak	Mark of regulatory elements associated with promoters and enhancers
H3K4me3	Peak	Mark of regulatory elements primarily associated with promoters/transcription starts
H3K9ac	Peak	Mark of active regulatory elements with preference for promoters
H3K9me1	Region	Preference for the 5′ end of genes
H3K9me3	Peak/region	Repressive mark associated with constitutive heterochromatin and repetitive elements
H3K27ac	Peak	Mark of active regulatory elements; may distinguish active enhancers and promoters from their inactive counterparts
H3K27me3	Region	Repressive mark established by polycomb complex activity associated with repressive domains and silent developmental genes
H3K36me3	Region	Elongation mark associated with transcribed portions of genes, with preference for 3′ regions after intron 1
H3K79me2	Region	Transcription-associated mark, with preference for 5′ end of genes
H4K20me1	Region	Preference for 5′ end of genes

©2012 Macmillan Publishers Limited. All rights reserved

assayed (K. Varley *et al.*, personal communication), and levels of DNA methylation correlated with chromatin accessibility. The most variably methylated CpGs are found more often in gene bodies and intergenic regions, rather than in promoters and upstream regulatory regions. In addition, we identified an unexpected correspondence between unmethylated genic CpG islands and binding by P300, a histone acetyltransferase linked to enhancer activity[44].

Because RRBS is a sequence-based assay with single-base resolution, we were able to identify CpGs with allele-specific methylation consistent with genomic imprinting, and determined that these loci exhibit aberrant methylation in cancer cell lines (K. Varley *et al.*, personal communication). Furthermore, we detected reproducible cytosine methylation outside CpG dinucleotides in adult tissues[45], providing further support that this non-canonical methylation event may have important roles in human biology (K. Varley *et al.*, personal communication).

Chromosome-interacting regions

Physical interaction between distinct chromosome regions that can be separated by hundreds of kilobases is thought to be important in the regulation of gene expression[46]. We used two complementary chromosome conformation capture (3C)-based technologies to probe these long-range physical interactions.

A 3C-carbon copy (5C) approach[47,48] provided unbiased detection of long-range interactions with TSSs in a targeted 1% of the genome (the 44 ENCODE pilot regions) in four cell types (GM12878, K562, HeLa-S3 and H1 hESC)[49]. We discovered hundreds of statistically significant long-range interactions in each cell type after accounting for chromatin polymer behaviour and experimental variation. Pairs of interacting loci showed strong correlation between the gene expression level of the TSS and the presence of specific functional element classes such as enhancers. The average number of distal elements interacting with a TSS was 3.9, and the average number of TSSs interacting with a distal element was 2.5, indicating a complex network of interconnected chromatin. Such interwoven long-range architecture was also uncovered genome-wide using chromatin interaction analysis with paired-end tag sequencing (ChIA-PET)[50] applied to identify interactions in chromatin enriched by RNA polymerase II (Pol II) ChIP from five cell types[51]. In K562 cells, we identified 127,417 promoter-centred chromatin interactions using ChIA-PET, 98% of which were intra-chromosomal. Whereas promoter regions of 2,324 genes were involved in 'single-gene' enhancer–promoter interactions, those of 19,813 genes were involved in 'multi-gene' interaction complexes spanning up to several megabases, including promoter–promoter and enhancer–promoter interactions[51].

These analyses portray a complex landscape of long-range gene–element connectivity across ranges of hundreds of kilobases to several megabases, including interactions among unrelated genes (Supplementary Fig. 1, section Y). Furthermore, in the 5C results, 50–60% of long-range interactions occurred in only one of the four cell lines, indicative of a high degree of tissue specificity for gene–element connectivity[49].

Summary of ENCODE-identified elements

Accounting for all these elements, a surprisingly large amount of the human genome, 80.4%, is covered by at least one ENCODE-identified element (detailed in Supplementary Table 1, section Q). The broadest element class represents the different RNA types, covering 62% of the genome (although the majority is inside of introns or near genes). Regions highly enriched for histone modifications form the next largest class (56.1%). Excluding RNA elements and broad histone elements, 44.2% of the genome is covered. Smaller proportions of the genome are occupied by regions of open chromatin (15.2%) or sites of transcription factor binding (8.1%), with 19.4% covered by at least one DHS or transcription factor ChIP-seq peak across all cell lines. Using our most conservative assessment, 8.5% of bases are covered by either a transcription-factor-binding-site motif (4.6%) or a DHS footprint (5.7%). This, however, is still about 4.5-fold higher than the amount of protein-coding exons, and about twofold higher than the estimated amount of pan-mammalian constraint.

Given that the ENCODE project did not assay all cell types, or all transcription factors, and in particular has sampled few specialized or developmentally restricted cell lineages, these proportions must be underestimates of the total amount of functional bases. However, many assays were performed on more than one cell type, allowing assessment of the rate of discovery of new elements. For both DHSs and CTCF-bound sites, the number of new elements initially increases rapidly with a steep gradient for the saturation curve and then slows with increasing number of cell types (Supplementary Figs 1 and 2, section R). With the current data, at the flattest part of the saturation curve each new cell type adds, on average, 9,500 DHS elements (across 106 cell types) and 500 CTCF-binding elements (across 49 cell types), representing 0.45% of the total element number. We modelled saturation for the DHSs and CTCF-binding sites using a Weibull distribution ($r^2 > 0.999$) and predict saturation at approximately 4.1 million (standard error (s.e.) = 108,000) and 185,100 (s.e. = 18,020) sites, respectively, indicating that we have discovered around half of the estimated total DHSs. These estimates represent a lower bound, but reinforce the observation that there is more non-coding functional DNA than either coding sequence or mammalian evolutionarily constrained bases.

The impact of selection on functional elements

From comparative genomic studies, at least 3–8% of bases are under purifying (negative) selection[4–11], indicating that these bases may potentially be functional. We previously found that 60% of mammalian evolutionarily constrained bases were annotated in the ENCODE pilot project, but also observed that many functional elements lacked evidence of constraint[2], a conclusion substantiated by others[52–54]. The diversity and genome-wide occurrence of functional elements now identified provides an unprecedented opportunity to examine further the forces of negative selection on human functional sequences.

We examined negative selection using two measures that highlight different periods of selection in the human genome. The first measure, inter-species, pan-mammalian constraint (GERP-based scores; 24 mammals[8]), addresses selection during mammalian evolution. The second measure is intra-species constraint estimated from the numbers of variants discovered in human populations using data from the 1000 Genomes project[55], and covers selection over human evolution. In Fig. 1, we plot both these measures of constraint for different classes of identified functional elements, excluding features overlapping exons and promoters that are known to be constrained. Each graph also shows genomic background levels and measures of coding-gene constraint for comparison. Because we plot human population diversity on an inverted scale, elements that are more constrained by negative selection will tend to lie in the upper and right-hand regions of the plot.

For DNase I elements (Fig. 1b) and bound motifs (Fig. 1c), most sets of elements show enrichment in pan-mammalian constraint and decreased human population diversity, although for some cell types the DNase I sites do not seem overall to be subject to pan-mammalian constraint. Bound transcription factor motifs have a natural control from the set of transcription factor motifs with equal sequence potential for binding but without binding evidence from ChIP-seq experiments—in all cases, the bound motifs show both more mammalian constraint and higher suppression of human diversity.

Consistent with previous findings, we do not observe genome-wide evidence for pan-mammalian selection of novel RNA sequences (Fig. 1d). There are also a large number of elements without mammalian constraint, between 17% and 90% for transcription-factor-binding regions as well as DHSs and FAIRE regions. Previous studies could not determine whether these sequences are either biochemically active, but with little overall impact on the organism, or under lineage-specific selection. By isolating sequences preferentially inserted into

©2012 Macmillan Publishers Limited. All rights reserved

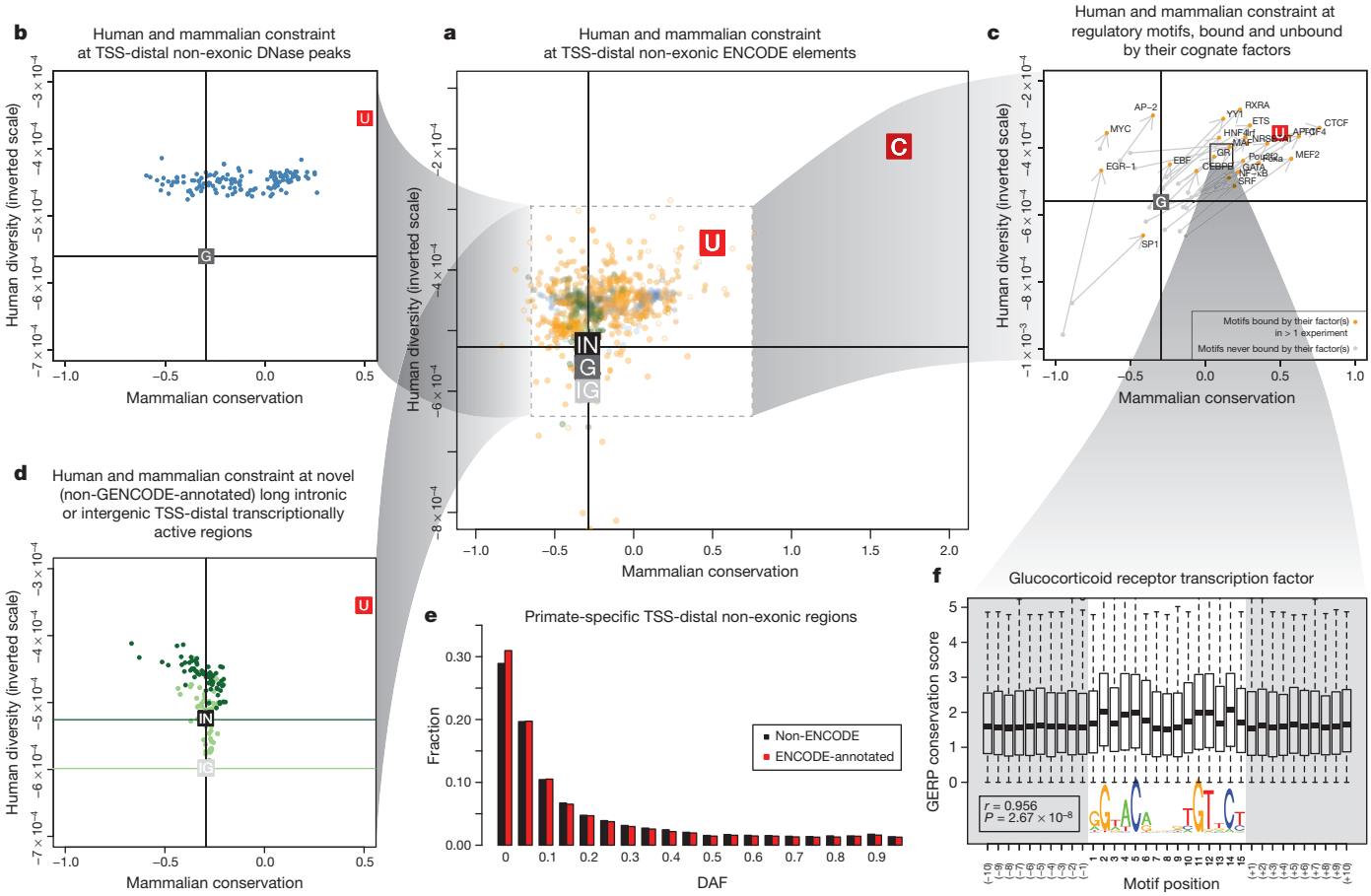


Figure 1 | Impact of selection on ENCODE functional elements in mammals and human populations. **a**, Levels of pan-mammalian constraint (mean GERP score; 24 mammals[8], *x* axis) compared to diversity, a measure of negative selection in the human population (mean expected heterozygosity, inverted scale, *y* axis) for ENCODE data sets. Each point is an average for a single data set. The top-right corners have the strongest evolutionary constraint and lowest diversity. Coding (C), UTR (U), genomic (G), intergenic (IG) and intronic (IN) averages are shown as filled squares. In each case the vertical and horizontal cross hairs show representative levels for the neutral expectation for mammalian conservation and human population diversity, respectively. The spread over all non-exonic ENCODE elements greater than 2.5 kb from TSSs is shown. The inner dashed box indicates that parts of the plot have been magnified for the surrounding outer panels, although the scales in the outer plots provide the exact regions and dimensions magnified. The spread for DHS sites (**b**) and RNA elements (**d**) is shown in the plots on the left. RNA elements are either long novel intronic (dark green) or long intergenic (light green) RNAs. The horizontal cross hairs are colour-coded to the relevant data set in **d**. **c**, Spread of transcription factor motif instances either in regions bound by the transcription factor (orange points) or in the corresponding unbound motif matches in grey, with bound and unbound points connected with an arrow in each case showing that bound sites are generally more constrained and less diverse. **e**, Derived allele frequency spectrum for primate-specific elements, with variations outside ENCODE elements in black and variations covered by ENCODE elements in red. The increase in low-frequency alleles compared to background is indicative of negative selection occurring in the set of variants annotated by the ENCODE data. **f**, Aggregation of mammalian constraint scores over the glucocorticoid receptor (GR) transcription factor motif in bound sites, showing the expected correlation with the information content of bases in the motif. An interactive version of this figure is available in the online version of the paper.

the primate lineage, which is only feasible given the genome-wide scale of this data, we are able to examine this issue specifically. Most primate-specific sequence is due to retrotransposon activity, but an appreciable proportion is non-repetitive primate-specific sequence. Of 104,343,413 primate-specific bases (excluding repetitive elements), 67,769,372 (65%) are found within ENCODE-identified elements. Examination of 227,688 variants segregating in these primate-specific regions revealed that all classes of elements (RNA and regulatory) show depressed derived allele frequencies, consistent with recent negative selection occurring in at least some of these regions (Fig. 1e). An alternative approach examining sequences that are not clearly under pan-mammalian constraint showed a similar result (L. Ward and M. Kellis, manuscript submitted). This indicates that an appreciable proportion of the unconstrained elements are lineage-specific elements required for organismal function, consistent with long-standing views of recent evolution[56], and the remainder are probably 'neutral' elements[2] that are not currently under selection but may still affect cellular or larger scale phenotypes without an effect on fitness.

The binding patterns of transcription factors are not uniform, and we can correlate both inter- and intra-species measures of negative selection with the overall information content of motif positions. The selection on some motif positions is as high as protein-coding exons (Fig. 1f; L. Ward and M. Kellis, manuscript submitted). These aggregate measures across motifs show that the binding preferences found in the population of sites are also relevant to the per-site behaviour. By developing a per-site metric of population effect on bound motifs, we found that highly constrained bound instances across mammals are able to buffer the impact of individual variation[57].

ENCODE data integration with known genomic features

Promoter-anchored integration

Many of the ENCODE assays directly or indirectly provide information about the action of promoters. Focusing on the TSSs of protein-coding transcripts, we investigated the relationships between different ENCODE assays, in particular testing the hypothesis that RNA expression (output) can be effectively predicted from patterns of

©2012 Macmillan Publishers Limited. All rights reserved

chromatin modification or transcription factor binding (input). Consistent with previous reports[58], we observe two relatively distinct types of promoter: (1) broad, mainly (C+G)-rich, TATA-less promoters; and (2) narrow, TATA-box-containing promoters. These promoters have distinct patterns of histone modifications, and transcription-factor-binding sites are selectively enriched in each class (Supplementary Fig. 1, section Z).

We developed predictive models to explore the interaction between histone modifications and measures of transcription at promoters, distinguishing between modifications known to be added as a consequence of transcription (such as H3K36me3 and H3K79me2) and other categories of histone marks[59]. In our analyses, the best models had two components: an initial classification component (on/off) and a second quantitative model component. Our models showed that activating acetylation marks (H3K27ac and H3K9ac) are roughly as informative as activating methylation marks (H3K4me3 and H3K4me2) (Fig. 2a). Although repressive marks, such as H3K27me3 or H3K9me3, show negative correlation both individually and in the model, removing these marks produces only a small reduction in model performance. However, for a subset of promoters in each cell line, repressive histone marks (H3K27me3 or H3K9me3) must be used to predict their expression accurately. We also examined the interplay between the H3K79me2 and H3K36me3 marks, both of which mark gene bodies, probably reflecting recruitment of modification enzymes by polymerase isoforms. As described previously, H3K79me2 occurs preferentially at the 5′ ends of gene bodies and H3K36me3 occurs more 3′, and our analyses support the previous model in which the H3K79me2 to H3K36me3 transition occurs at the first 3′ splice site[60].

Few previous studies have attempted to build qualitative or quantitative models of transcription genome-wide from transcription factor levels because of the paucity of documented transcription-factor-binding regions and the lack of coordination around a single cell line. We thus examined the predictive capacity of transcription-factor-binding signals for the expression levels of promoters (Fig. 2b).

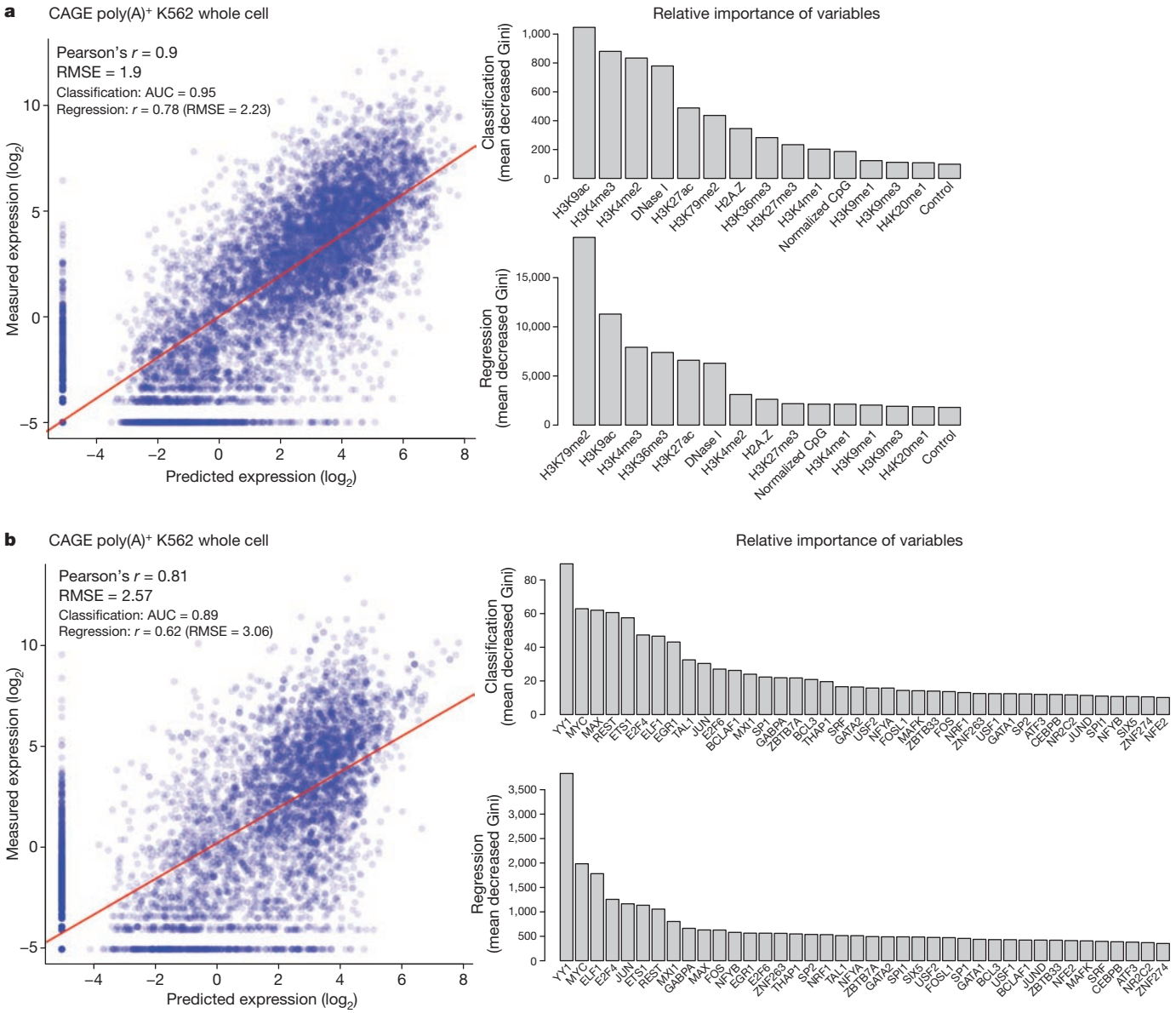


Figure 2 | Modelling transcription levels from histone modification and transcription-factor-binding patterns. **a**, **b**, Correlative models between either histone modifications or transcription factors, respectively, and RNA production as measured by CAGE tag density at TSSs in K562 cells. In each case the scatter plot shows the output of the correlation models (*x* axis) compared to observed values (*y* axis). The bar graphs show the most important histone modifications (**a**) or transcription factors (**b**) in both the initial classification phase (top bar graph) or the quantitative regression phase (bottom bar graph), with larger values indicating increasing importance of the variable in the model. Further analysis of other cell lines and RNA measurement types is reported elsewhere[59,79]. AUC, area under curve; Gini, Gini coefficient; RMSE, root mean square error.

©2012 Macmillan Publishers Limited. All rights reserved

In contrast to the profiles of histone modifications, most transcription factors show enriched binding signals in a narrow DNA region near the TSS, with relatively higher binding signals in promoters with higher CpG content. Most of this correlation could be recapitulated by looking at the aggregate binding of transcription factors without specific transcription factor terms. Together, these correlation models indicate both that a limited set of chromatin marks are sufficient to 'explain' transcription and that a variety of transcription factors might have broad roles in general transcription levels across many genes. It is important to note that this is an inherently observational study of correlation patterns, and is consistent with a variety of mechanistic models with different causal links between the chromatin, transcription factor and RNA assays. However, it does indicate that there is enough information present at the promoter regions of genes to explain most of the variation in RNA expression.

We developed predictive models similar to those used to model transcriptional activity to explore the relationship between levels of histone modification and inclusion of exons in alternately spliced transcripts. Even accounting for expression level, H3K36me3 has a positive contribution to exon inclusion, whereas H3K79me2 has a negative contribution (H. Tilgner *et al.*, manuscript in preparation). By monitoring the RNA populations in the subcellular fractions of K562 cells, we found that essentially all splicing is co-transcriptional[61], further supporting a link between chromatin structure and splicing.

Transcription-factor-binding site-anchored integration

Transcription-factor-binding sites provide a natural focus around which to explore chromatin properties. Transcription factors are often multifunctional and can bind a variety of genomic loci with different combinations and patterns of chromatin marks and nucleosome organization. Hence, rather than averaging chromatin mark profiles across all binding sites of a transcription factor, we developed a clustering procedure, termed the Clustered Aggregation Tool (CAGT), to identify subsets of binding sites sharing similar but distinct patterns of chromatin mark signal magnitude, shape and hidden directionality[30]. For example, the average profile of the repressive histone mark H3K27me3 over all 55,782 CTCF-binding sites in H1 hESCs shows poor signal enrichment (Fig. 3a). However, after grouping profiles by signal magnitude we found a subset of 9,840 (17.6%) CTCF-binding sites that exhibit significant flanking H3K27me3 signal. Shape and orientation analysis further revealed that the predominant signal profile for H3K27me3 around CTCF peak summits is asymmetric, consistent with a boundary role for some CTCF sites between active and polycomb-silenced domains. Further examples are provided in Supplementary Figs 5 and 6 of section E. For TAF1, predominantly found near TSSs, the asymmetric sites are orientated with the direction of transcription. However, for distal sites, such as those bound by GATA1 and CTCF, we also observed a high proportion of asymmetric histone patterns, although independent of motif directionality. In fact, all transcription-factor-binding data sets in all cell lines show predominantly asymmetric patterns (asymmetry ratio >0.6) for all chromatin marks but not for DNase I signal (Fig. 3b). This indicates that most transcription-factor-bound chromatin events correlate with structured, directional patterns of histone modifications, and that promoter directionality is not the only source of orientation at these sites.

We also examined nucleosome occupancy relative to the symmetry properties of chromatin marks around transcription-factor-binding sites. Around TSSs, there is usually strong asymmetric nucleosome occupancy, often accounting for most of the histone modification signal (for instance, see Supplementary Fig. 4, section E). However, away from TSSs, there is far less concordance. For example, CTCF-binding sites typically show arrays of well-positioned nucleosomes on either side of the peak summit (Supplementary Fig. 1, section E)[62]. Where the flanking chromatin mark signal is high, the signals are often asymmetric, indicating differential marking with histone modifications (Supplementary Figs 2 and 3, section E). Thus, we

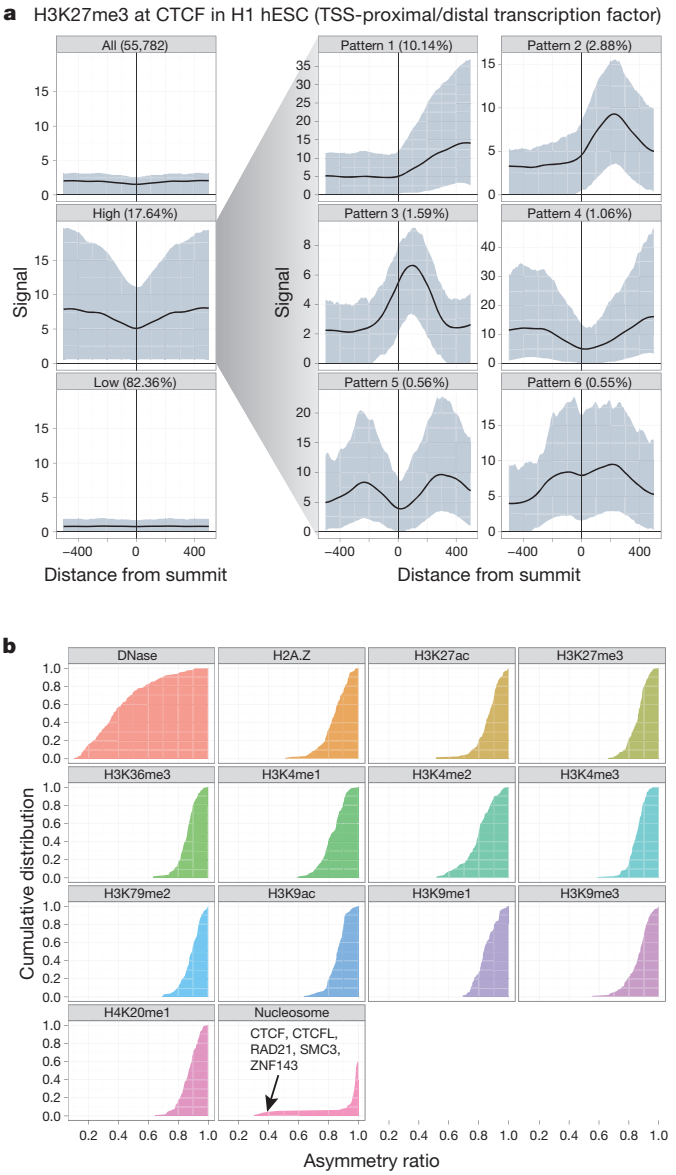


Figure 3 | Patterns and asymmetry of chromatin modification at transcription-factor-binding sites. **a**, Results of clustered aggregation of H3K27me3 modification signal around CTCF-binding sites (a multifunctional protein involved with chromatin structure). The first three plots (left column) show the signal behaviour of the histone modification over all sites (top) and then split into the high and low signal components. The solid lines show the mean signal distribution by relative position with the blue shaded area delimiting the tenth and ninetieth percentile range. The high signal component is then decomposed further into six different shape classes on the right (see ref. 30 for details). The shape decomposition process is strand aware. **b**, Summary of shape asymmetry for DNase I, nucleosome and histone modification signals by plotting an asymmetry ratio for each signal over all transcription-factor-binding sites. All histone modifications measured in this study show predominantly asymmetric patterns at transcription-factor-binding sites. An interactive version of this figure is available in the online version of the paper.

confirm on a genome-wide scale that transcription factors can form barriers around which nucleosomes and histone modifications are arranged in a variety of configurations[62–65]. This is explored in further detail in refs 25, 26 and 30.

Transcription factor co-associations

Transcription-factor-binding regions are nonrandomly distributed across the genome, with respect to both other features (for example, promoters) and other transcription-factor-binding regions. Within the

©2012 Macmillan Publishers Limited. All rights reserved

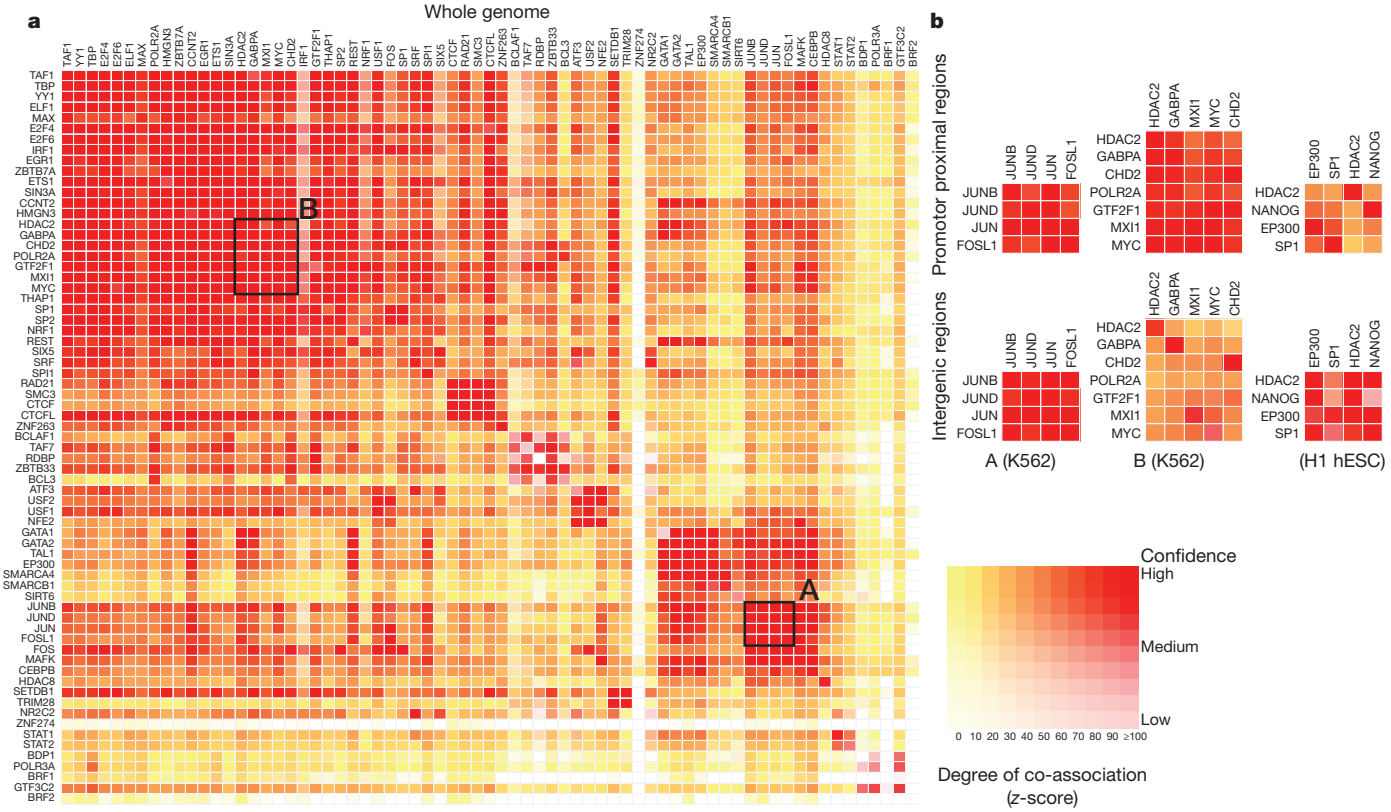


Figure 4 | Co-association between transcription factors. **a**, Significant co-associations of transcription factor pairs using the GSC statistic across the entire genome in K562 cells. The colour strength represents the extent of association (from red (strongest), orange, to yellow (weakest)), whereas the depth of colour represents the fit to the GSC[20] model (where white indicates that the statistical model is not appropriate) as indicated by the key. Most transcription factors have a nonrandom association to other transcription factors, and these associations are dependent on the genomic context, meaning that once the genome is separated into promoter proximal and distal regions, the overall levels of co-association decrease, but more specific relationships are uncovered. **b**, Three classes of behaviour are shown. The first column shows a set of associations for which strength is independent of location in promoter and distal regions, whereas the second column shows a set of transcription factors that have stronger associations in promoter-proximal regions. Both of these examples are from data in K562 cells and are highlighted on the genome-wide co-association matrix (**a**) by the labelled boxes A and B, respectively. The third column shows a set of transcription factors that show stronger association in distal regions (in the H1 hESC line). An interactive version of this figure is available in the online version of the paper.

tier 1 and 2 cell lines, we found 3,307 pairs of statistically co-associated factors ($P < 1 \times 10^{-16}$, GSC) involving 114 out of a possible 117 factors (97%) (Fig. 4a). These include expected associations, such as Jun and Fos, and some less expected novel associations, such as TCF7L2 with HNF4-α and FOXA2 (ref. 66; a full listing is given in Supplementary Table 1, section F). When one considers promoter and intergenic

Table 3 | Summary of the combined state types

Label	Description	Details*	Colour
CTCF	CTCF-enriched element	Sites of CTCF signal lacking histone modifications, often associated with open chromatin. Many probably have a function in insulator assays, but because of the multifunctional nature of CTCF, we are conservative in our description. Also enriched for the cohesin components RAD21 and SMC3; CTCF is known to recruit the cohesin complex.	Turquoise
E	Predicted enhancer	Regions of open chromatin associated with H3K4me1 signal. Enriched for other enhancer-associated marks, including transcription factors known to act at enhancers. In enhancer assays, many of these (>50%) function as enhancers. A more conservative alternative would be *cis*-regulatory regions. Enriched for sites for the proteins encoded by *EP300*, *FOS*, *FOSL1*, *GATA2*, *HDAC8*, *JUNB*, *JUND*, *NFE2*, *SMARCA4*, *SMARCB1*, *SIRT6* and *TAL1* genes in K562 cells. Have nuclear and whole-cell RNA signal, particularly poly(A)− fraction.	Orange
PF	Predicted promoter flanking region	Regions that generally surround TSS segments (see below).	Light red
R	Predicted repressed or low-activity region	This is a merged state that includes H3K27me3 polycomb-enriched regions, along with regions that are silent in terms of observed signal for the input assays to the segmentations (low or no signal). They may have other signals (for example, RNA, not in the segmentation input data). Enriched for sites for the proteins encoded by *REST* and some other factors (for example, proteins encoded by *BRF2*, *CEBPB*, *MAFK*, *TRIM28*, *ZNF274* and *SETDB1* genes in K562 cells).	Grey
TSS	Predicted promoter region including TSS	Found close to or overlapping GENCODE TSS sites. High precision/recall for TSSs. Enriched for H3K4me3. Sites of open chromatin. Enriched for transcription factors known to act close to promoters and polymerases Pol II and Pol III. Short RNAs are most enriched in these segments.	Bright red
T	Predicted transcribed region	Overlap gene bodies with H3K36me3 transcriptional elongation signal. Enriched for phosphorylated form of Pol II signal (elongating polymerase) and poly(A)$^+$ RNA, especially cytoplasmic.	Dark green
WE	Predicted weak enhancer or open chromatin *cis*-regulatory element	Similar to the E state, but weaker signals and weaker enrichments.	Yellow

* Where specific enrichments or overlaps are identified, these are derived from analysis in GM12878 and/or K562 cells where the data for comparison is richest. The colours indicated are used in Figs 5 and 7 and in display of these tracks from the ENCODE data hub.

©2012 Macmillan Publishers Limited. All rights reserved

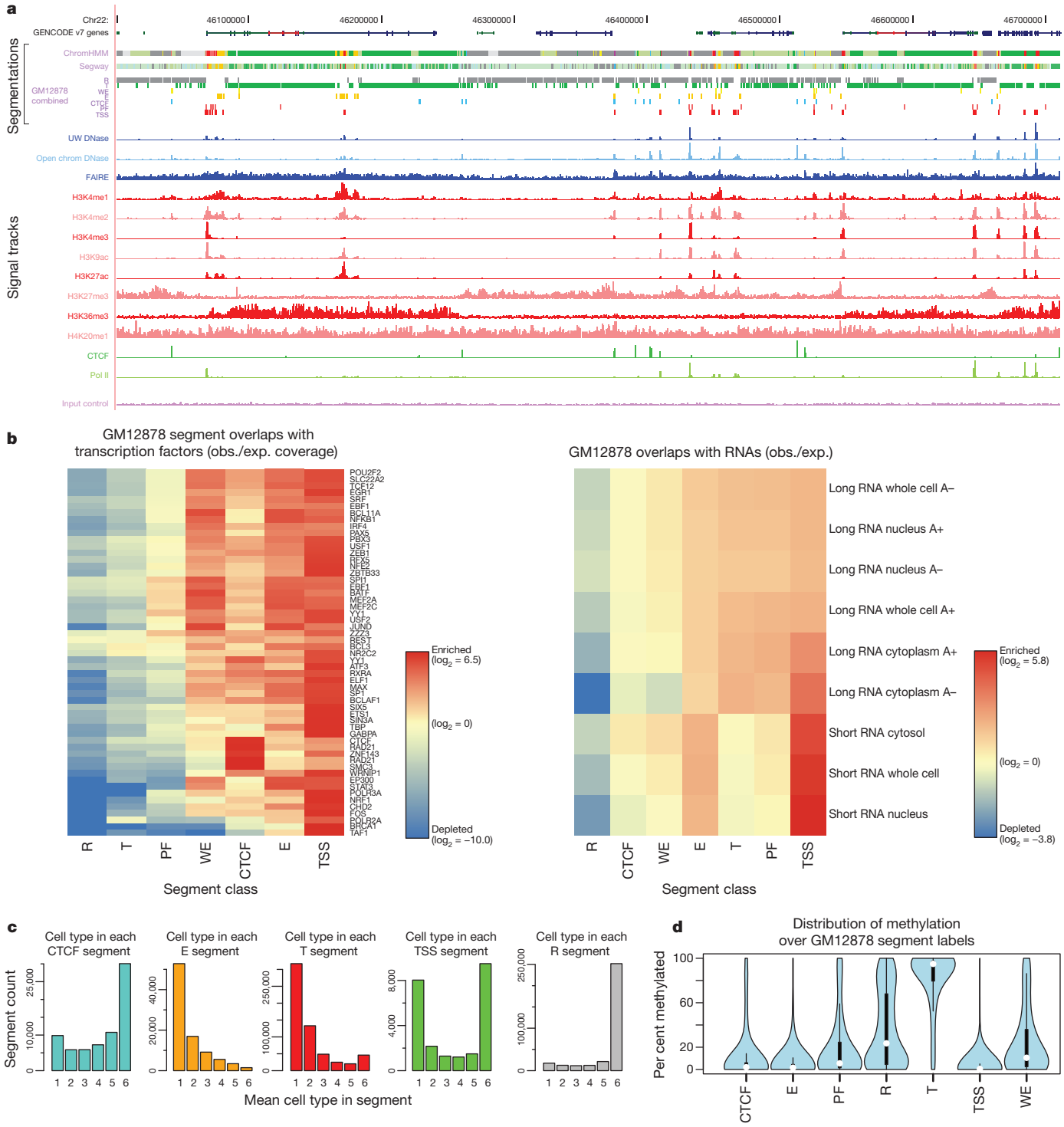


Figure 5 | Integration of ENCODE data by genome-wide segmentation. **a**, Illustrative region with the two segmentation methods (ChromHMM and Segway) in a dense view and the combined segmentation expanded to show each state in GM12878 cells, beneath a compressed view of the GENCODE gene annotations. Note that at this level of zoom and genome browser resolution, some segments appear to overlap although they do not. Segmentation classes are named and coloured according to the scheme in Table 3. Beneath the segmentations are shown each of the normalized signals that were used as the input data for the segmentations. Open chromatin signals from DNase-seq from the University of Washington group (UW DNase) or the ENCODE open chromatin group (Openchrom DNase) and FAIRE assays are shown in blue; signal from histone modification ChIP-seq in red; and transcription factor ChIP-seq signal for Pol II and CTCF in green. The mauve ChIP-seq control signal (input control) at the bottom was also included as an input to the segmentation. **b**, Association of selected transcription factor (left) and RNA (right) elements in the combined segmentation states (*x* axis) expressed as an observed/expected ratio (obs./exp.) for each combination of transcription factor or RNA element and segmentation class using the heat-map scale shown in the key besides each heat map. **c**, Variability of states between cell lines, showing the distribution of occurrences of the state in the six cell lines at specific genome locations: from unique to one cell line to ubiquitous in all six cell lines for five states (CTCF, E, T, TSS and R). **d**, Distribution of methylation level at individual sites from RRBS analysis in GM12878 cells across the different states, showing the expected hypomethylation at TSSs and hypermethylation of genes bodies (T state) and repressed (R) regions.

©2012 Macmillan Publishers Limited. All rights reserved

regions separately, this changes to 3,201 pairs (116 factors, 99%) for promoters and 1,564 pairs (108 factors, 92%) for intergenic regions, with some associations more specific to these genomic contexts (for example, the cluster of HDAC2, GABPA, CHD2, GTF2F1, MXI1 and MYC in promoter regions and SP1, EP300, HDAC2 and NANOG in intergenic regions (Fig. 4b)). These general and context-dependent associations lead to a network representation of the co-binding with many interesting properties, explored in refs 19, 25 and 26. In addition, we also identified a set of regions bound by multiple factors representing high occupancy of transcription factor (HOT) regions[67].

Genome-wide integration

To identify functional regions genome-wide, we next integrated elements independent of genomic landmarks using either discriminative training methods, where a subset of known elements of a particular class were used to train a model that was then used to discover more instances of this class, or using methods in which only data from ENCODE assays were used without explicit knowledge of any annotation.

For discriminative training, we used a three-step process to predict potential enhancers, described in Supplementary Information and ref. 67. Two alternative discriminative models converged on a set of ~13,000 putative enhancers in K562 cells[67]. In the second approach, two methodologically distinct unbiased approaches (see refs 40, 68 and M. M. Hoffman *et al.*, manuscript in preparation) converged on a concordant set of histone modification and chromatin-accessibility patterns that can be used to segment the genome in each of the tier 1 and tier 2 cell lines, although the individual loci in each state in each cell line are different. With the exception of RNA polymerase II and CTCF, the addition of transcription factor data did not substantially alter these patterns. At this stage, we deliberately excluded RNA and methylation assays, reserving these data as a means to validate the segmentations.

Our integration of the two segmentation methods (M. M. Hoffman *et al.*, manuscript in preparation) established a consensus set of seven major classes of genome states, described in Table 3. The standard view of active promoters, with a distinct core promoter region (TSS and PF states), leading to active gene bodies (T, transcribed state), is rediscovered in this model (Fig. 5a, b). There are three 'active' distal states. We tentatively labelled two as enhancers (predicted enhancers, E, and predicted weak enhancers, WE) due to their occurrence in regions of open chromatin with high H3K4me1, although they differ in the levels of marks such as H3K27ac, currently thought to distinguish active from inactive enhancers. The other active state (CTCF) has high CTCF binding and includes sequences that function as insulators in a transfection assay. The remaining repressed state (R) summarizes sequences split between different classes of actively repressed or inactive, quiescent chromatin. We found that the CTCF-binding-associated state is relatively invariant across cell types, with individual regions frequently occupying the CTCF state across all six cell types (Fig. 5c). Conversely, the E and T states have substantial cell-specific behaviour, whereas the TSS state has a bimodal behaviour with similar numbers of cell-invariant and cell-specific occurrences. It is important to note that the consensus summary classes do not capture all the detail discovered in the individual segmentations containing more states.

The distribution of RNA species across segments is quite distinct, indicating that underlying biological activities are captured in the segmentation. Polyadenylated RNA is heavily enriched in gene bodies. Around promoters, there are short RNA species previously identified as promoter-associated short RNAs (Fig. 5b)[16,69]. Similarly, DNA methylation shows marked distinctions between segments, recapitulating the known biology of predominantly unmethylated active promoters (TSS states) followed by methylated gene bodies[42] (T state, Fig. 5d). The two enhancer-enriched states show distinct patterns of DNA methylation, with the less active enhancer state (by H3K27ac/H3K4me1 levels) showing higher methylation. These states also have an excess of RNA elements without poly(A) tails and methyl-cap RNA, as assayed by CAGE sequences, compared to matched intergenic controls, indicating a specific transcriptional mode associated with active enhancers[70]. Transcription factors also showed distinct distributions across the segments (Fig. 5b). A striking pattern is the concentration of transcription factors in the TSS-associated state. The enhancers contain a different set of transcription factors. For example, in K562 cells, the E state is enriched for binding by the proteins encoded by the *EP300*, *FOS*, *FOSL1*, *GATA2*, *HDAC8*, *JUNB*, *JUND*, *NFE2*, *SMARCA4*, *SMARCB1*, *SIRT6* and *TAL1* genes. We tested a subset of these predicted enhancers in both mouse and fish transgenic models (examples in Fig. 6), with over half of the elements showing activity, often in the corresponding tissue type.

The segmentation provides a linear determination of functional state across the genome, but not an association of particular distal regions with genes. By using the variation of DNase I signal across cell lines, 39% of E (enhancer associated) states could be linked to a proposed regulated gene[29] concordant with physical proximity patterns determined by 5C[49] or ChIA-PET.

To provide a fine-grained regional classification, we turned to a self organizing map (SOM) to cluster genome segmentation regions based on their assay signal characteristics (Fig. 7). The segmentation regions were initially randomly assigned to a 1,350-state map in a two-dimensional toroidal space (Fig. 7a). This map can be visualized as a two-dimensional rectangular plane onto which the various signal distributions can be plotted. For instance, the rectangle at the bottom left of Fig. 7a shows the distribution of the genome in the initial randomized map. The SOM was then trained using the twelve different ChIP-seq and DNase-seq assays in the six cell types previously analysed in the large-scale segmentations (that is, over 72-dimensional space). After training, the SOM clustering was again visualized in two dimensions, now showing the organized distribution of genome segments (lower right of panel, Fig. 7a). Individual data sets associated with the genome segments in each SOM map unit (hexagonal cells) can then be visualized in the same framework to learn how each additional kind of data is distributed on the chromatin state map. Figure 7b shows CAGE/TSS expression data overlaid on the randomly initialized (left) and trained map (right) panels. In this way the trained SOM highlighted cell-type-specific TSS clusters (bottom panels of Fig. 7b), indicating that there are sets of tissue-specific TSSs that are distinguished from each other by subtle combinations of ENCODE

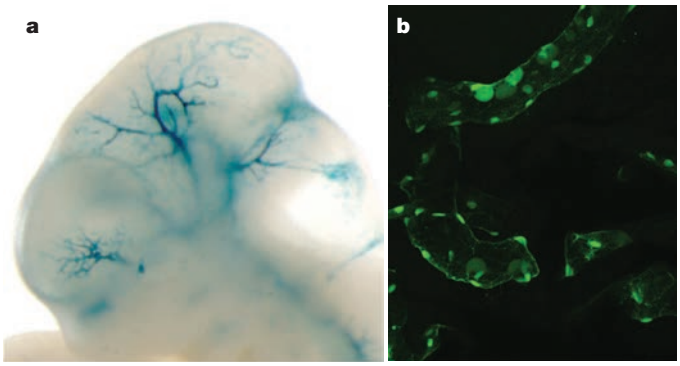


Figure 6 | Experimental characterization of segmentations. Randomly sampled E state segments (see Table 3) from the K562 segmentation were cloned for mouse- and fish-based transgenic enhancer assays. **a**, Representative LacZ-stained transgenic embryonic day (E)11.5 mouse embryo obtained with construct hs2065 (EN167, chr10: 46052882–46055670, GRCh37). Highly reproducible staining in the blood vessels was observed in 9 out of 9 embryos resulting from independent transgenic integration events. **b**, Representative green fluorescent protein reporter transgenic medaka fish obtained from a construct with a basal *hsp70* promoter on meganuclease-based transfection. Reproducible transgenic expression in the circulating nucleated blood cells and the endothelial cell walls was seen in 81 out of 100 transgenic tests of this construct.

©2012 Macmillan Publishers Limited. All rights reserved

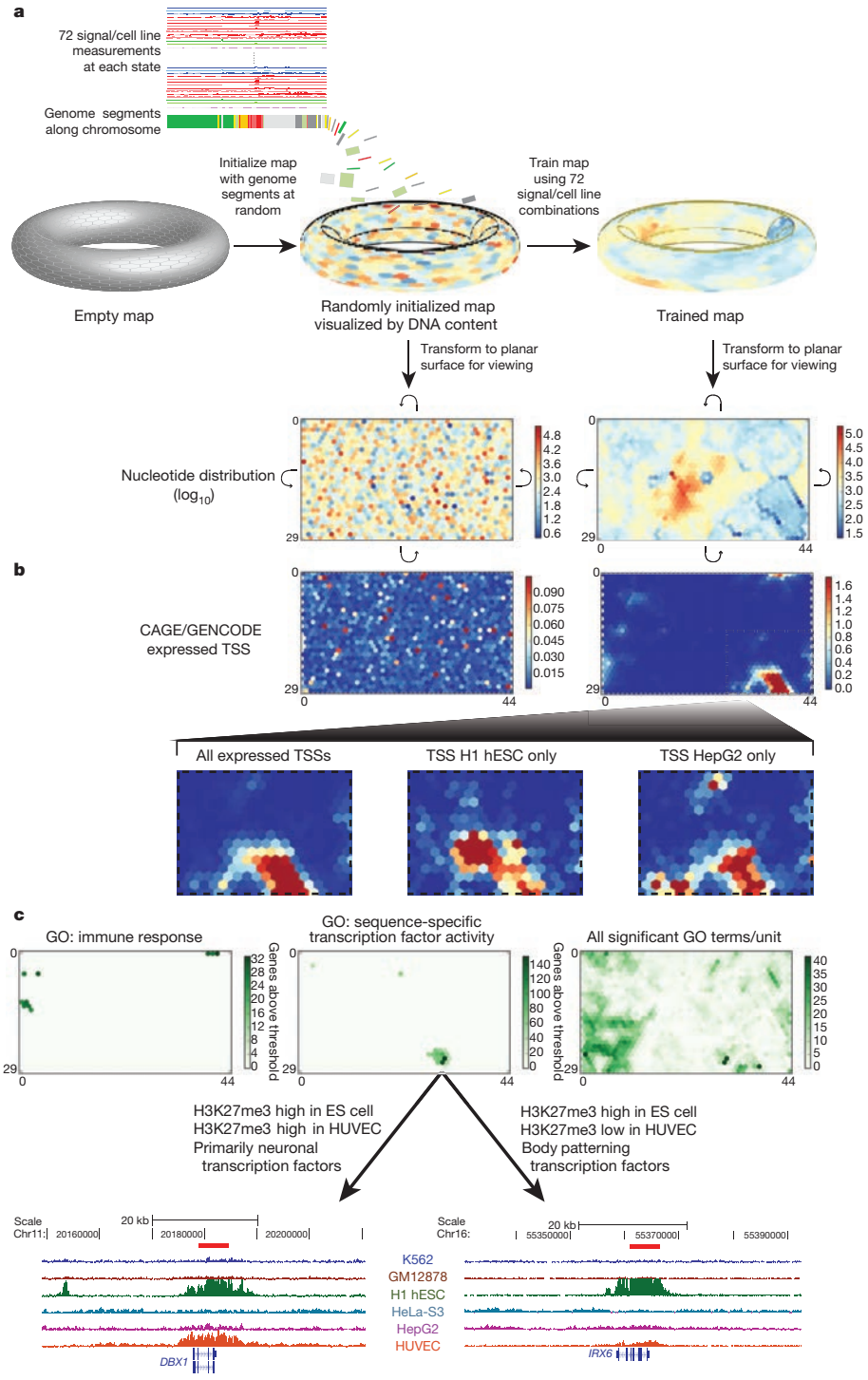


Figure 7 | High-resolution segmentation of ENCODE data by self-organizing maps (SOM). **a**–**c**, The training of the SOM (**a**) and analysis of the results (**b**, **c**) are shown. Initially we arbitrarily placed genomic segments from the ChromHMM segmentation on to the toroidal map surface, although the SOM does not use the ChromHMM state assignments (**a**). We then trained the map using the signal of the 12 different ChIP-seq and DNase-seq assays in the six cell types analysed. Each unit of the SOM is represented here by a hexagonal cell in a planar two-dimensional view of the toroidal map. Curved arrows indicate that traversing the edges of two dimensional view leads back to the opposite edge. The resulting map can be overlaid with any class of ENCODE or other data to view the distribution of that data within this high-resolution segmentation. In panel **a** the distributions of genome bases across the untrained and trained map (left and right, respectively) are shown using heat-map colours for $\log_{10}$ values. **b**, The distribution of TSSs from CAGE experiments of GENCODE annotation on the planar representations of either the initial random organization (left) or the final trained SOM (right) using heat maps coloured according to the accompanying scales. The bottom half of **b** expands the different distributions in the SOM for all expressed TSSs (left) or TSSs specifically expressed in two example cell lines, H1 hESC (centre) and HepG2 (right). **c**, The association of Gene Ontology (GO) terms on the same representation of the same trained SOM. We assigned genes that are within 20 kb of a genomic segment in a SOM unit to that unit, and then associated this set of genes with GO terms using a hypergeometric distribution after correcting for multiple testing. Map units that are significantly associated to GO terms are coloured green, with increasing strength of colour reflecting increasing numbers of genes significantly associated with the GO terms for either immune response (left) or sequence-specific transcription factor activity (centre). In each case, specific SOM units show association with these terms. The right-hand panel shows the distribution on the same SOM of all significantly associated GO terms, now colouring by GO term count per SOM unit. For sequence-specific transcription factor activity, two example genomic regions are extracted at the bottom of panel **c** from neighbouring SOM units. These are regions around the *DBX1* (from SOM unit 26,31, left panel) and *IRX6* (SOM unit 27,30, right panel) genes, respectively, along with their H3K27me3 ChIP-seq signal for each of the tier 1 and 2 cell types. For *DBX1*, representative of a set of primarily neuronal transcription factors associated with unit 26,31, there is a repressive H3K27me3 signal in both H1 hESCs and HUVECs; for *IRX6*, representative of a set of body patterning transcription factors associated with SOM unit 27,30, the repressive mark is restricted largely to the embryonic stem (ES) cell. An interactive version of this figure is available in the online version of the paper.

chromatin data. Many of the ultra-fine-grained state classifications revealed in the SOM are associated with specific gene ontology (GO) terms (right panel of Fig. 7c). For instance, the left panel of Fig. 7c identifies ten SOM map units enriched with genomic regions associated with genes associated with the GO term 'immune response'. The central panel identifies a different set of map units enriched for the GO term 'sequence-specific transcription factor activity'. The two map units most enriched for this GO term, indicated by the darkest green colouring, contain genes with segments that are high in H3K27me3 in H1 hESCs, but that differ in H3K27me3 levels in HUVECs. Gene function analysis with the GO ontology tool (GREAT[71]) reveals that the map unit with high H3K27me3 levels in both cell types is enriched in transcription factor genes with known neuronal functions, whereas the neighbouring map unit is enriched in genes involved in body patterning. The genome browser shots at the bottom of Fig. 7c pick out an example region for each of the two SOM map units illustrating the difference in H3K27me3 signal. Overall, we have 228 distinct GO terms associated with specific segments across

©2012 Macmillan Publishers Limited. All rights reserved

one or more states (A. Mortazavi, personal communication), and can assign over one-third of genes to a GO annotation solely on the basis of its multicellular histone patterns. Thus, the SOM analysis provides a fine-grained map of chromatin data across multiple cell types, which can then be used to relate chromatin structure to other data types at differing levels of resolution (for instance, the large cluster of units containing any active TSS, its subclusters composed of units enriched in TSSs active in only one cell type, or individual map units significantly enriched for specific GO terms).

The classifications presented here are necessarily limited by the assays and cell lines studied, and probably contain a number of heterogeneous classes of elements. Nonetheless, robust classifications can be made, allowing a systematic view of the human genome.

Insights into human genomic variation

We next explored the potential impact of sequence variation on ENCODE functional elements. We examined allele-specific variation using results from the GM12878 cells that are derived from an individual (NA12878) sequenced in the 1000 Genomes project, along with her parents. Because ENCODE assays are predominantly sequence-based, the trio design allows each GM12878 data set to be divided by the specific parental contributions at heterozygous sites, producing aggregate haplotypic signals from multiple genomic sites. We examined 193 ENCODE assays for allele-specific biases using 1,409,992 phased, heterozygous SNPs and 167,096 insertions/deletions (indels) (Fig. 8). Alignment biases towards alleles present in the reference genome sequence were avoided using a sequence specifically tailored to the variants and haplotypes present in NA12878 (a 'personalized genome')[72]. We found instances of preferential binding towards each parental allele. For example, comparison of the results from the POLR2A, H3K79me2 and H3K27me3 assays in the region of *NACC2* (Fig. 8a) shows a strong paternal bias for H3K79me2 and POL2RA and a strong maternal bias for H3K27me3, indicating differential activity for the maternal and paternal alleles.

Figure 8b shows the correlation of selected allele-specific signals across the whole genome. For instance, we found a strong allelic correlation between POL2RA and BCLAF1 binding, as well as negative correlation between H3K79me2 and H3K27me3, both at genes (Fig. 8b, below the diagonal, bottom left) and chromosomal segments (top right). Overall, we found that positive allelic correlations among the 193 ENCODE assays are stronger and more frequent than negative correlations. This may be due to preferential capture of accessible alleles and/or the specific histone modification and transcription factor, assays used in the project.

Rare variants, individual genomes and somatic variants

We further investigated the potential functional effects of individual variation in the context of ENCODE annotations. We divided NA12878 variants into common and rare classes, and partitioned these into those overlapping ENCODE annotation (Fig. 9a and Supplementary Tables 1 and 2, section K). We also predicted potential functional effects: for protein-coding genes, these are either non-synonymous SNPs or variants likely to induce loss of function by frame-shift, premature stop, or splice-site disruption; for other regions, these are variants that overlap a transcription-factor-binding site. We found similar numbers of potentially functional variants affecting protein-coding genes or affecting other ENCODE annotations, indicating that many functional variants within individual genomes lie outside exons of protein-coding genes. A more detailed analysis of regulatory variant annotation is described in ref. 73.

To study further the potential effects of NA12878 genome variants on transcription-factor-binding regions, we performed peak calling using a constructed personal diploid genome sequence for NA12878 (ref. 72). We aligned ChIP-seq sequences from GM12878 separately against the maternal and paternal haplotypes. As expected, a greater

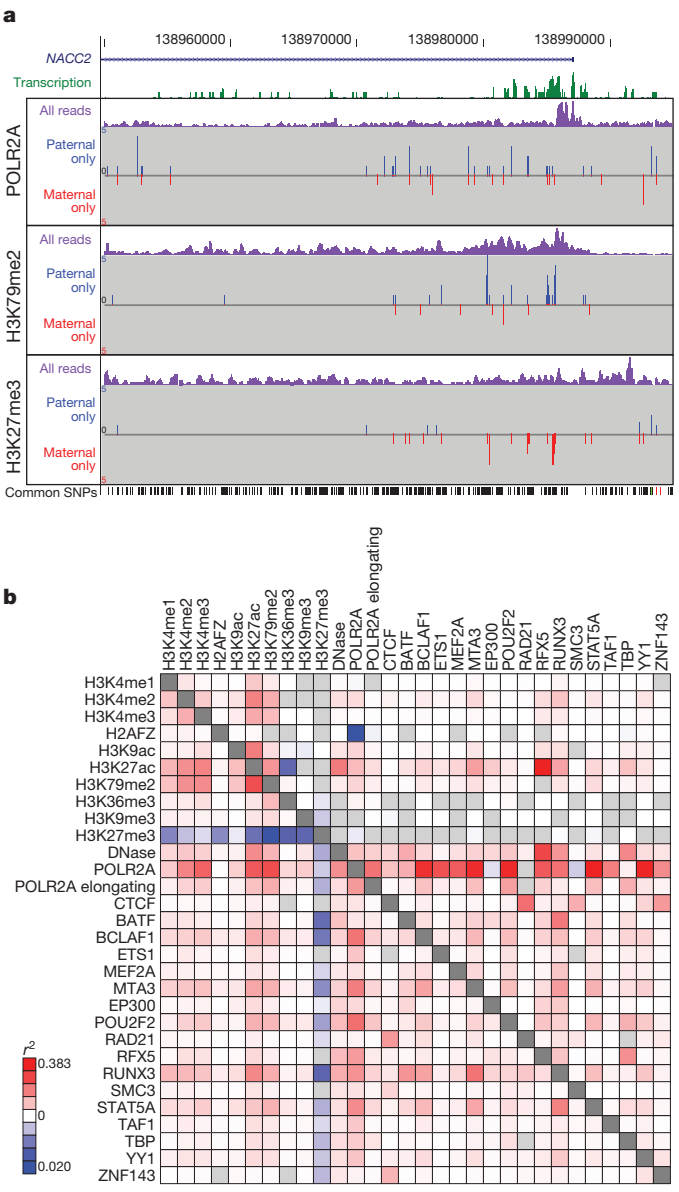


Figure 8 | Allele-specific ENCODE elements. a, Representative allele-specific information from GM12878 cells for selected assays around the first exon of the *NACC2* gene (genomic region Chr9: 138950000–138995000, GRCh37). Transcription signal is shown in green, and the three sections show allele-specific data for three data sets (POLR2A, H3K79me2 and H3K27me3 ChIP-seq). In each case the purple signal is the processed signal for all sequence reads for the assay, whereas the blue and red signals show sequence reads specifically assigned to either the paternal or maternal copies of the genome, respectively. The set of common SNPs from dbSNP, including the phased, heterozygous SNPs used to provide the assignment, are shown at the bottom of the panel. *NACC2* has a statistically significant paternal bias for POLR2A and the transcription-associated mark H3K79me2, and has a significant maternal bias for the repressive mark H3K27me3. **b**, Pair-wise correlations of allele-specific signal within single genes (below the diagonal) or within individual ChromHMM segments across the whole genome for selected DNase-seq and histone modification and transcription factor ChIP-seq assays. The extent of correlation is coloured according to the heat-map scale indicated from positive correlation (red) through to anti-correlation (blue). An interactive version of this figure is available in the online version of the paper.

fraction of reads were aligned than to the reference genome (see Supplementary Information, Supplementary Fig. 1, section K). On average, approximately 1% of transcription-factor-binding sites in GM12878 cells are detected in a haplotype-specific fashion. For instance, Fig. 9b shows a CTCF-binding site not detected using the

©2012 Macmillan Publishers Limited. All rights reserved

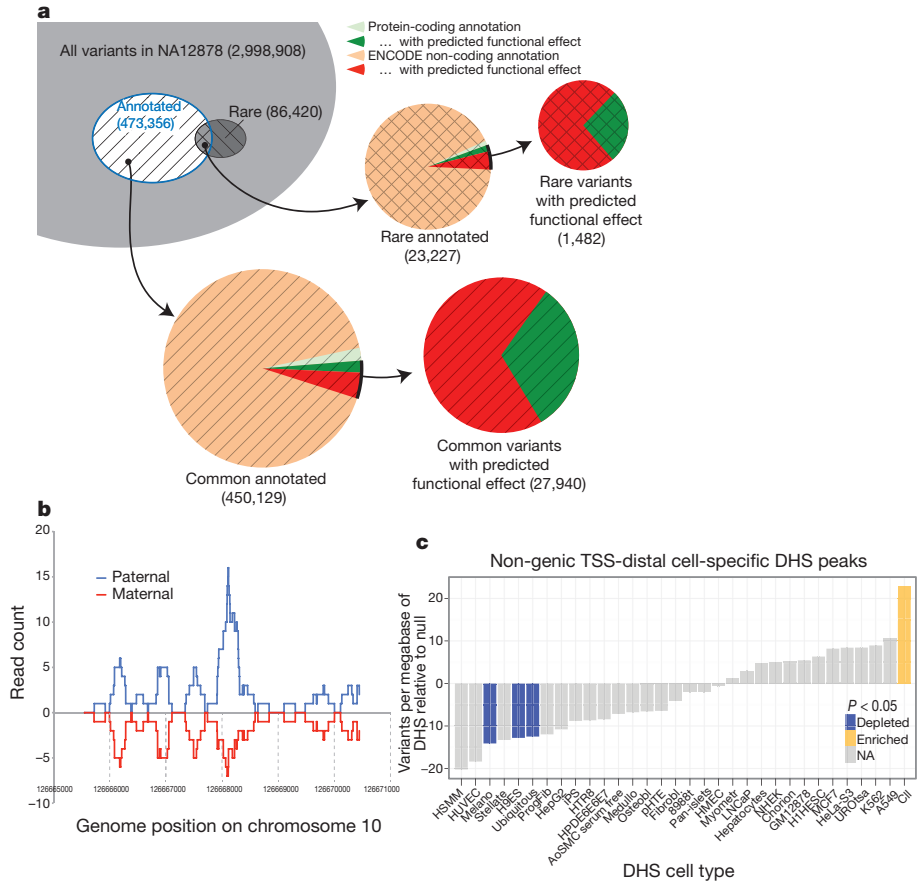


Figure 9 | Examining ENCODE elements on a per individual basis in the normal and cancer genome. **a**, Breakdown of variants in a single genome (NA12878) by both frequency (common or rare (that is, variants not present in the low-coverage sequencing of 179 individuals in the pilot 1 European panel of the 1000 Genomes project[55])) and by ENCODE annotation, including protein-coding gene and non-coding elements (GENCODE annotations for protein-coding genes, pseudogenes and other ncRNAs, as well as transcription-factor-binding sites from ChIP-seq data sets, excluding broad annotations such as histone modifications, segmentations and RNA-seq). Annotation status is further subdivided by predicted functional effect, being non-synonymous and missense mutations for protein-coding regions and variants overlapping bound transcription factor motifs for non-coding element annotations. A substantial proportion of variants are annotated as having predicted functional effects in the non-coding category. **b**, One of several relatively rare occurrences, where alignment to an individual genome sequence (paternal and maternal panels) shows a different readout from the reference genome. In this case, a paternal-haplotype-specific CTCF peak is identified. **c**, Relative level of somatic variants from a whole-genome melanoma sample that occur in DHSs unique to different cell lines. The coloured bars show cases that are significantly enriched or suppressed in somatic mutations. Details of ENCODE cell types can be found at http://encodeproject.org/ENCODE/cellTypes.html. An interactive version of this figure is available in the online version of the paper.

reference sequence that is only present on the paternal haplotype due to a 1-bp deletion (see also Supplementary Fig. 2, section K). As costs of DNA sequencing decrease further, optimized analysis of ENCODE-type data should use the genome sequence of the individual or cell being analysed when possible.

Most analyses of cancer genomes so far have focused on characterizing somatic variants in protein-coding regions. We intersected four available whole-genome cancer data sets with ENCODE annotations (Fig. 9c and Supplementary Fig. 2, section L). Overall, somatic variation is relatively depleted from ENCODE annotated regions, particularly for elements specific to a cell type matching the putative tumour source (for example, skin melanocytes for melanoma). Examining the mutational spectrum of elements in introns for cases where a strand-specific mutation assignment could be made reveals that there are mutational spectrum differences between DHSs and unannotated regions (0.06 Fisher's exact test, Supplementary Fig. 3, section L). The suppression of somatic mutation is consistent with important functional roles of these elements within tumour cells, highlighting a potential alternative set of targets for examination in cancer.

Common variants associated with disease

In recent years, GWAS have greatly extended our knowledge of genetic loci associated with human disease risk and other phenotypes. The output of these studies is a series of SNPs (GWAS SNPs) correlated with a phenotype, although not necessarily the functional variants. Notably, 88% of associated SNPs are either intronic or intergenic[74]. We examined 4,860 SNP–phenotype associations for 4,492 SNPs curated in the National Human Genome Research Institute (NHGRI) GWAS catalogue[74]. We found that 12% of these SNPs overlap transcription-factor-occupied regions whereas 34% overlap DHSs (Fig. 10a). Both figures reflect significant enrichments relative to the overall proportions of 1000 Genomes project SNPs (about 6% and 23%, respectively). Even after accounting for biases introduced by selection of SNPs for the standard genotyping arrays, GWAS SNPs show consistently higher overlap with ENCODE annotations (Fig. 10a, see Supplementary Information). Furthermore, after partitioning the genome by density of different classes of functional elements, GWAS SNPs were consistently enriched beyond all the genotyping SNPs in function-rich partitions, and depleted in function-poor partitions (see Supplementary Fig. 1, section M). GWAS SNPs are particularly enriched in the segmentation classes associated with enhancers and TSSs across several cell types (see Supplementary Fig. 2, section M).

Examining the SOM of integrated ENCODE annotations (see above), we found 19 SOM map units showing significant enrichment for GWAS SNPs, including many SOM units previously associated with specific gene functions, such as the immune response regions.

©2012 Macmillan Publishers Limited. All rights reserved

Thus, an appreciable proportion of SNPs identified in initial GWAS scans are either functional or lie within the length of an ENCODE annotation (~500 bp on average) and represent plausible candidates for the functional variant. Expanding the set of feasible functional SNPs to those in reasonable linkage disequilibrium, up to 71% of GWAS SNPs have a potential causative SNP overlapping a DNase I site, and 31% of loci have a candidate SNP that overlaps a binding site occupied by a transcription factor (see also refs 73, 75).

The GWAS catalogue provides a rich functional categorization from the precise phenotypes being studied. These phenotypic categorizations are nonrandomly associated with ENCODE annotations and there is marked correspondence between the phenotype and the identity of the cell type or transcription factor used in the ENCODE assay (Fig. 10b). For example, five SNPs associated with Crohn's disease overlap GATA2-binding sites (*P* value 0.003 by random permutation or 0.001 by an empirical approach comparing to the GWAS-matched SNPs; see Supplementary Information), and fourteen are located in DHSs found in immunologically relevant cell types. A notable example is a gene desert on chromosome 5p13.1 containing eight SNPs associated with inflammatory diseases. Several are close to or within DHSs in T-helper type 1 (T_H1) and T_H2 cells as well as peaks of binding by transcription factors in HUVECs (Fig. 10c). The latter cell line is not immunological, but factor occupancy detected there could be a proxy for binding of a more relevant factor, such as GATA3, in T cells. Genetic variants in this region also affect expression levels of *PTGER4* (ref. 76), encoding the prostaglandin receptor EP4. Thus, the ENCODE data reinforce the hypothesis that genetic variants in 5p13.1 modulate the expression of flanking genes, and furthermore provide the specific hypothesis that the variants affect occupancy of a GATA factor in an allele-specific manner, thereby influencing susceptibility to Crohn's disease.

Nonrandom association of phenotypes with ENCODE cell types strengthens the argument that at least some of the GWAS lead SNPs are functional or extremely close to functional variants. Each of the associations between a lead SNP and an ENCODE annotation remains a credible hypothesis of a particular functional element

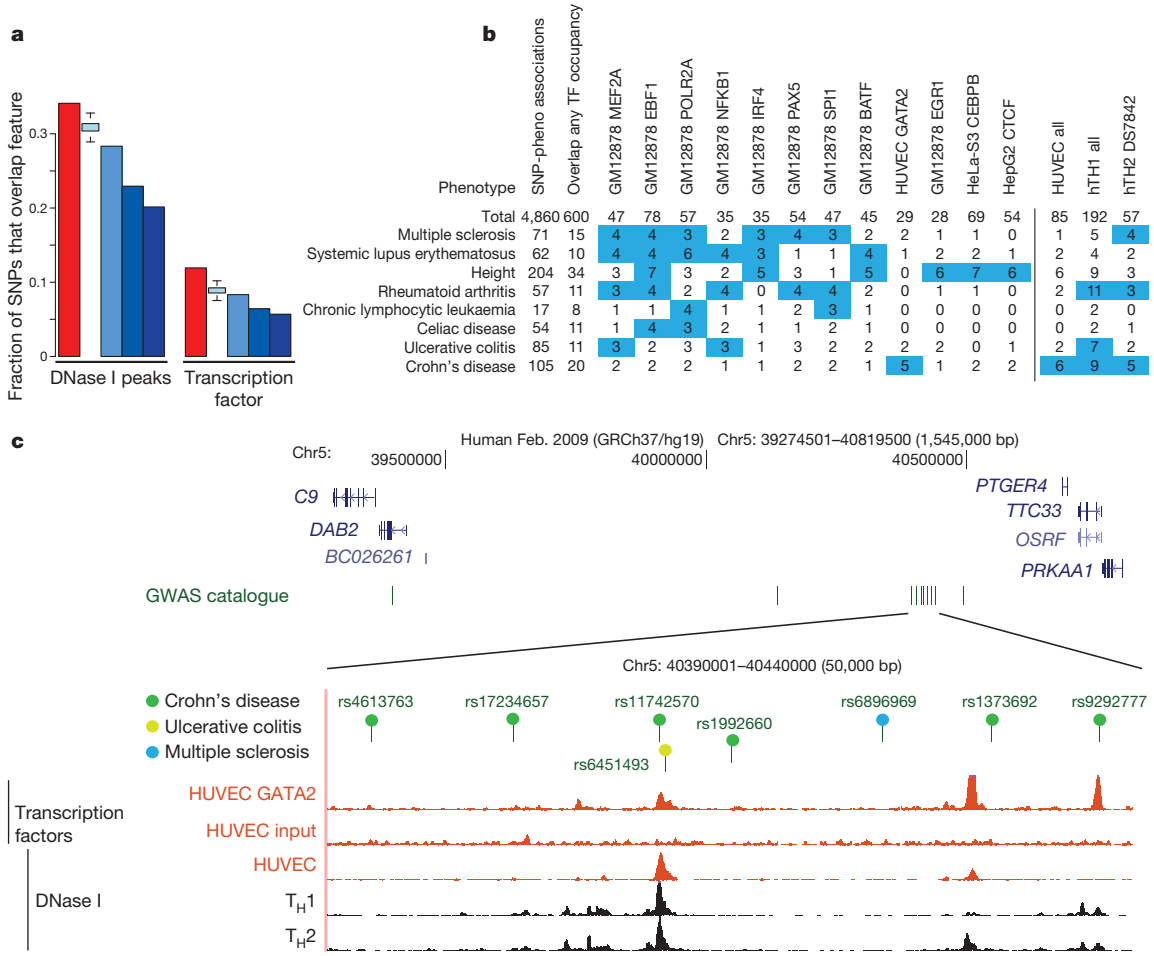


Figure 10 | Comparison of genome-wide-association-study-identified loci with ENCODE data. **a**, Overlap of lead SNPs in the NHGRI GWAS SNP catalogue (June 2011) with DHSs (left) or transcription-factor-binding sites (right) as red bars compared with various control SNP sets in blue. The control SNP sets are (from left to right): SNPs on the Illumina 2.5M chip as an example of a widely used GWAS SNP typing panel; SNPs from the 1000 Genomes project; SNPs extracted from 24 personal genomes (see personal genome variants track at http://main.genome-browser.bx.psu.edu (ref. 80)), all shown as blue bars. In addition, a further control used 1,000 randomizations from the genotyping SNP panel, matching the SNPs with each NHGRI catalogue SNP for allele frequency and distance to the nearest TSS (light blue bars with bounds at 1.5 times the interquartile range). For both DHSs and transcription-factor-binding regions, a larger proportion of overlaps with GWAS-implicated SNPs is found compared to any of the controls sets. **b**, Aggregate overlap of phenotypes to selected transcription-factor-binding sites (left matrix) or DHSs in selected cell lines (right matrix), with a count of overlaps between the phenotype and the cell line/factor. Values in blue squares pass an empirical *P*-value threshold ≤0.01 (based on the same analysis of overlaps between randomly chosen, GWAS-matched SNPs and these epigenetic features) and have at least a count of three overlaps. The *P* value for the total number of phenotype–transcription factor associations is <0.001. **c**, Several SNPs associated with Crohn's disease and other inflammatory diseases that reside in a large gene desert on chromosome 5, along with some epigenetic features indicative of function. The SNP (rs11742570) strongly associated to Crohn's disease overlaps a GATA2 transcription-factor-binding signal determined in HUVECs. This region is also DNase I hypersensitive in HUVECs and T-helper T_H1 and T_H2 cells. An interactive version of this figure is available in the online version of the paper.

©2012 Macmillan Publishers Limited. All rights reserved

class or cell type to explore with future experiments. Supplementary Tables 1–3, section M, list all 14,885 pairwise associations across the ENCODE annotations. The accompanying papers have a more detailed examination of common variants with other regulatory information[19,25,29,73,75,77].

Concluding remarks

The unprecedented number of functional elements identified in this study provides a valuable resource to the scientific community as well as significantly enhances our understanding of the human genome. Our analyses have revealed many novel aspects of gene expression and regulation as well as the organization of such information, as illustrated by the accompanying papers (see http://www.encodeproject.org/ENCODE/pubs.html for collected ENCODE publications). However, there are still many specific details, particularly about the mechanistic processes that generate these elements and how and where they function, that require additional experiments to elucidate.

The large spread of coverage—from our highest resolution, most conservative set of bases implicated in GENCODE protein-coding gene exons (2.9%) or specific protein DNA binding (8.5%) to the broadest, most general set of marks covering the genome (approximately 80%), with many gradations in between—presents a spectrum of elements with different functional properties discovered by ENCODE. A total of 99% of the known bases in the genome are within 1.7 kb of any ENCODE element, whereas 95% of bases are within 8 kb of a bound transcription factor motif or DNase I footprint. Interestingly, even using the most conservative estimates, the fraction of bases likely to be involved in direct gene regulation, even though incomplete, is significantly higher than that ascribed to protein-coding exons (1.2%), raising the possibility that more information in the human genome may be important for gene regulation than for biochemical function. Many of the regulatory elements are not constrained across mammalian evolution, which so far has been one of the most reliable indications of an important biochemical event for the organism. Thus, our data provide orthologous indicators for suggesting possible functional elements.

Importantly, for the first time we have sufficient statistical power to assess the impact of negative selection on primate-specific elements, and all ENCODE classes display evidence of negative selection in these unique-to-primate elements. Furthermore, even with our most conservative estimate of functional elements (8.5% of putative DNA/protein binding regions) and assuming that we have already sampled half of the elements from our transcription factor and cell-type diversity, one would estimate that at a minimum 20% (17% from protein binding and 2.9% protein coding gene exons) of the genome participates in these specific functions, with the likely figure significantly higher.

The broad coverage of ENCODE annotations enhances our understanding of common diseases with a genetic component, rare genetic diseases, and cancer, as shown by our ability to link otherwise anonymous associations to a functional element. ENCODE and similar studies provide a first step towards interpreting the rest of the genome—beyond protein-coding genes—thereby augmenting common disease genetic studies with testable hypotheses. Such information justifies performing whole-genome sequencing (rather than exome only, 1.2% of the genome) on rare diseases and investigating somatic variants in non-coding functional elements, for instance, in cancer. Furthermore, as GWAS analyses typically associate disease to SNPs in large regions, comparison to ENCODE non-coding functional elements can help pinpoint putative causal variants in addition to refinement of location by fine-mapping techniques[78]. Combining ENCODE data with allele-specific information derived from individual genome sequences provides specific insight on the impact of a genetic variant. Indeed, we believe that a significant goal would be to use functional data such as that derived from this project to assign every genomic variant to its possible impact on human phenotypes.

So far, ENCODE has sampled 119 of 1,800 known transcription factors and general components of the transcriptional machinery on a limited number of cell types, and 13 of more than 60 currently known histone or DNA modifications across 147 cell types. DNase I, FAIRE and extensive RNA assays across subcellular fractionations have been undertaken on many cell types, but overall these data reflect a minor fraction of the potential functional information encoded in the human genome. An important future goal will be to enlarge this data set to additional factors, modifications and cell types, complementing the other related projects in this area (for example, Roadmap Epigenomics Project, http://www.roadmapepigenomics.org/, and International Human Epigenome Consortium, http://www.ihec-epigenomes.org/). These projects will constitute foundational resources for human genomics, allowing a deeper interpretation of the organization of gene and regulatory information and the mechanisms of regulation, and thereby provide important insights into human health and disease. Co-published ENCODE-related papers can be explored online via the *Nature* ENCODE explorer (http://www.nature.com/ENCODE), a specially designed visualization tool that allows users to access the linked papers and investigate topics that are discussed in multiple papers via thematically organized threads.

METHODS SUMMARY

For full details of Methods, see Supplementary Information.

Received 24 November 2011; accepted 29 May 2012.

1. ENCODE Project Consortium. The ENCODE (ENCyclopedia Of DNA Elements) Project. *Science* **306,** 636–640 (2004).
2. Birney, E. *et al.* Identification and analysis of functional elements in 1% of the human genome by the ENCODE pilot project. *Nature* **447,** 799–816 (2007).
3. The ENCODE Project Consortium. A user's guide to the encyclopedia of DNA elements (ENCODE). *PLoS Biol.* **9,** e1001046 (2011).
4. Mouse Genome Sequencing Consortium. Initial sequencing and comparative analysis of the mouse genome. *Nature* **420,** 520–562 (2002).
5. Chiaromonte, F. *et al.* The share of human genomic DNA under selection estimated from human-mouse genomic alignments. *Cold Spring Harb. Symp. Quant. Biol.* **68,** 245–254 (2003).
6. Cooper, G. M. *et al.* Distribution and intensity of constraint in mammalian genomic sequence. *Genome Res.* **15,** 901–913 (2005).
7. Parker, S. C., Hansen, L., Abaan, H. O., Tullius, T. D. & Margulies, E. H. Local DNA topography correlates with functional noncoding regions of the human genome. *Science* **324,** 389–392 (2009).
8. Lindblad-Toh, K. *et al.* A high-resolution map of human evolutionary constraint using 29 mammals. *Nature* **478,** 476–482 (2011).
9. Pheasant, M. & Mattick, J. S. Raising the estimate of functional human sequences. *Genome Res.* **17,** 1245–1253 (2007).
10. Ponting, C. P. & Hardison, R. C. What fraction of the human genome is functional? *Genome Res.* **21,** 1769–1776 (2011).
11. Asthana, S. *et al.* Widely distributed noncoding purifying selection in the human genome. *Proc. Natl Acad. Sci. USA* **104,** 12410–12415 (2007).
12. Landt, S. G. *et al.* ChIP-seq guidelines and practices used by the ENCODE and modENCODE consortia. *Genome Res.* http://dx.doi.org/10.1101/gr.136184.111 (2012).
13. Li, Q., Brown, J. B., Huang, H. & Bickel, P. J. Measuring reproducibility of high-throughput experiments. *Ann. Appl. Stat.* **5,** 1752–1779 (2011).
14. Harrow, J. *et al.* GENCODE: The reference human genome annotation for the ENCODE project. *Genome Res.* http://dx.doi.org/10.1101/gr.135350.111 (2012).
15. Howald, C. *et al.* Combining RT-PCR-seq and RNA-seq to catalog all genic elements encoded in the human genome. *Genome Res.* http://dx.doi.org/10.1101/gr.134478.111 (2012).
16. Djebali, S. *et al.* Landscape of transcription in human cells. *Nature* http://dx.doi.org/10.1038/nature11233 (this issue).
17. Derrien, T. *et al.* The GENCODE v7 catalog of human long noncoding RNAs: Analysis of their gene structure, evolution, and expression. *Genome Res.* http://dx.doi.org/10.1101/gr.132159.111 (2012).
18. Pei, B. *et al.* The GENCODE pseudogene resource. *Genome Biol.* **13,** R51 (2012).
19. Gerstein, M. B. *et al.* Architecture of the human regulatory network derived from ENCODE data. *Nature* http://dx.doi.org/10.1038/nature11245 (this issue).
20. Bickel, P. J., Boley, N., Brown, J. B., Huang, H. Y. & Zhang, N. R. Subsampling methods for genomic inference. *Ann. Appl. Stat.* **4,** 1660–1697 (2010).
21. Kaplan, T. *et al.* Quantitative models of the mechanisms that control genome-wide patterns of transcription factor binding during early *Drosophila* development. *PLoS Genet.* **7,** e1001290 (2011).
22. Li, X. Y. *et al.* The role of chromatin accessibility in directing the widespread, overlapping patterns of *Drosophila* transcription factor binding. *Genome Biol.* **12,** R34 (2011).

©2012 Macmillan Publishers Limited. All rights reserved

23. Pique-Regi, R. *et al.* Accurate inference of transcription factor binding from DNA sequence and chromatin accessibility data. *Genome Res.* **21,** 447–455 (2011).
24. Zhang, Y. *et al.* Primary sequence and epigenetic determinants of *in vivo* occupancy of genomic DNA by GATA1. *Nucleic Acids Res.* **37,** 7024–7038 (2009).
25. Neph, S. *et al.* An expansive human regulatory lexicon encoded in transcription factor footprints. *Nature* http://dx.doi.org/10.1038/nature11212 (this issue).
26. Whitfield, T. W. *et al.* Functional analysis of transcription factor binding sites in human promoters. *Genome Biol.* **13,** R50 (2012).
27. Gross, D. S. & Garrard, W. T. Nuclease hypersensitive sites in chromatin. *Annu. Rev. Biochem.* **57,** 159–197 (1988).
28. Urnov, F. D. Chromatin remodeling as a guide to transcriptional regulatory networks in mammals. *J. Cell. Biochem.* **88,** 684–694 (2003).
29. Thurman, R. E. *et al.* The accessible chromatin landscape of the human genome. *Nature* http://dx.doi.org/10.1038/nature11232 (this issue).
30. Kundaje, A. *et al.* Ubiquitous heterogeneity and asymmetry of the chromatin environment at regulatory elements. *Genome Res.* http://dx.doi.org/10.1101/gr.136366.111 (2012).
31. Schultz, D. C., Ayyanathan, K., Negorev, D., Maul, G. G. & Rauscher, F. J. III. SETDB1: a novel KAP-1-associated histone H3, lysine 9-specific methyltransferase that contributes to HP1-mediated silencing of euchromatic genes by KRAB zinc-finger proteins. *Genes Dev.* **16,** 919–932 (2002).
32. Frietze, S., O'Geen, H., Blahnik, K. R., Jin, V. X. & Farnham, P. J. ZNF274 recruits the histone methyltransferase SETDB1 to the 3′ ends of ZNF genes. *PLoS ONE* **5,** e15082 (2010).
33. Boyle, A. P. *et al.* High-resolution genome-wide *in vivo* footprinting of diverse transcription factors in human cells. *Genome Res.* **21,** 456–464 (2011).
34. Hesselberth, J. R. *et al.* Global mapping of protein-DNA interactions *in vivo* by digital genomic footprinting. *Nature Methods* **6,** 283–289 (2009).
35. Zhang, Y. *et al.* Model-based analysis of ChIP-Seq (MACS). *Genome Biol.* **9,** R137 (2008).
36. Kouzarides, T. Chromatin modifications and their function. *Cell* **128,** 693–705 (2007).
37. Li, B., Carey, M. & Workman, J. L. The role of chromatin during transcription. *Cell* **128,** 707–719 (2007).
38. Hon, G. C., Hawkins, R. D. & Ren, B. Predictive chromatin signatures in the mammalian genome. *Hum. Mol. Genet.* **18,** R195–R201 (2009).
39. Zhou, V. W., Goren, A. & Bernstein, B. E. Charting histone modifications and the functional organization of mammalian genomes. *Nature Rev. Genet.* **12,** 7–18 (2011).
40. Ernst, J. *et al.* Mapping and analysis of chromatin state dynamics in nine human cell types. *Nature* **473,** 43–49 (2011).
41. Hon, G., Wang, W. & Ren, B. Discovery and annotation of functional chromatin signatures in the human genome. *PLoS Comput. Biol.* **5,** e1000566 (2009).
42. Ball, M. P. *et al.* Targeted and genome-scale strategies reveal gene-body methylation signatures in human cells. *Nature Biotechnol.* **27,** 361–368 (2009).
43. Meissner, A. *et al.* Genome-scale DNA methylation maps of pluripotent and differentiated cells. *Nature* **454,** 766–770 (2008).
44. Ogryzko, V. V., Schiltz, R. L., Russanova, V., Howard, B. H. & Nakatani, Y. The transcriptional coactivators p300 and CBP are histone acetyltransferases. *Cell* **87,** 953–959 (1996).
45. Lister, R. *et al.* Human DNA methylomes at base resolution show widespread epigenomic differences. *Nature* **462,** 315–322 (2009).
46. Dekker, J. Gene regulation in the third dimension. *Science* **319,** 1793–1794 (2008).
47. Dostie, J. *et al.* Chromosome Conformation Capture Carbon Copy (5C): a massively parallel solution for mapping interactions between genomic elements. *Genome Res.* **16,** 1299–1309 (2006).
48. Lajoie, B. R., van Berkum, N. L., Sanyal, A. & Dekker, J. My5C: web tools for chromosome conformation capture studies. *Nature Methods* **6,** 690–691 (2009).
49. Sanyal, A., Lajoie, B., Jain, G. & Dekker, J. The long-range interaction landscape of gene promoters. *Nature* http://dx.doi.org/10.1038/nature11279 (this issue).
50. Fullwood, M. J. *et al.* An oestrogen-receptor-alpha-bound human chromatin interactome. *Nature* **462,** 58–64 (2009).
51. Li, G. *et al.* Extensive promoter-centered chromatin interactions provide a topological basis for transcription regulation. *Cell* **148,** 84–98 (2012).
52. Borneman, A. R. *et al.* Divergence of transcription factor binding sites across related yeast species. *Science* **317,** 815–819 (2007).
53. Odom, D. T. *et al.* Tissue-specific transcriptional regulation has diverged significantly between human and mouse. *Nature Genet.* **39,** 730–732 (2007).
54. Schmidt, D. *et al.* Five-vertebrate ChIP-seq reveals the evolutionary dynamics of transcription factor binding. *Science* **328,** 1036–1040 (2010).
55. A map of human genome variation from population-scale sequencing. *Nature* **467,** 1061–1073 (2010).
56. King, M. C. & Wilson, A. C. Evolution at two levels in humans and chimpanzees. *Science* **188,** 107–116 (1975).
57. Spivakov, M. *et al.* Analysis of variation at transcription factor binding sites in *Drosophila* and humans. *Genome Biol.* **13,** R49 (2012).
58. Sandelin, A. *et al.* Mammalian RNA polymerase II core promoters: insights from genome-wide studies. *Nature Rev. Genet.* **8,** 424–436 (2007).
59. Dong, X. *et al.* Modeling gene expression using chromatin features in various cellular contexts. *Genome Biol.* **13,** R53 (2012).
60. Huff, J. T., Plocik, A. M., Guthrie, C. & Yamamoto, K. R. Reciprocal intronic and exonic histone modification regions in humans. *Nature Struct. Mol. Biol.* **17,** 1495–1499 (2010).
61. Tilgner, H. *et al.* Deep sequencing of subcellular RNA fractions shows splicing to be predominantly co-transcriptional in the human genome but inefficient for lncRNAs. *Genome Res.* http://dx.doi.org/10.1101/gr.134445.111 (2012).
62. Fu, Y., Sinha, M., Peterson, C. L. & Weng, Z. The insulator binding protein CTCF positions 20 nucleosomes around its binding sites across the human genome. *PLoS Genet.* **4,** e1000138 (2008).
63. Kornberg, R. D. & Stryer, L. Statistical distributions of nucleosomes: nonrandom locations by a stochastic mechanism. *Nucleic Acids Res.* **16,** 6677–6690 (1988).
64. Schones, D. E. *et al.* Dynamic regulation of nucleosome positioning in the human genome. *Cell* **132,** 887–898 (2008).
65. Valouev, A. *et al.* Determinants of nucleosome organization in primary human cells. *Nature* **474,** 516–520 (2011).
66. Frietze, S. *et al.* Cell type-specific binding patterns reveal that TCF7L2 can be tethered to the genome by association with GATA3. *Genome Biol.* **13,** R52 (2012).
67. Yip, K. Y. *et al.* Classification of human genomic regions based on experimentally-determined binding sites of more than 100 transcription-related factors. *Genome Biol.* **13,** R48 (2012).
68. Hoffman, M. M. *et al.* Unsupervised pattern discovery in human chromatin structure through genomic segmentation. *Nature Methods* **9,** 473–476 (2012).
69. Kapranov, P. *et al.* RNA maps reveal new RNA classes and a possible function for pervasive transcription. *Science* **316,** 1484–1488 (2007).
70. Koch, F. *et al.* Transcription initiation platforms and GTF recruitment at tissue-specific enhancers and promoters. *Nature Struct. Mol. Biol.* **18,** 956–963 (2011).
71. McLean, C. Y. *et al.* GREAT improves functional interpretation of cis-regulatory regions. *Nature Biotechnol.* **28,** 495–501 (2010).
72. Rozowsky, J. *et al.* AlleleSeq: analysis of allele-specific expression and binding in a network framework. *Mol. Syst. Biol.* **7,** 522 (2011).
73. Boyle, A. P. *et al.* Annotation of functional variation in personal genomes using RegulomeDB. *Genome Res.* http://dx.doi.org/10.1101/gr.137323.112 (2012).
74. Hindorff, L. A. *et al.* Potential etiologic and functional implications of genome-wide association loci for human diseases and traits. *Proc. Natl Acad. Sci. USA* **106,** 9362–9367 (2009).
75. Schaub, M. A., Boyle, A. P., Kundaje, A., Batzoglou, S. & Snyder, M. Linking disease associations with regulatory information in the human genome. *Genome Res.* http://dx.doi.org/10.1101/gr.136127.111 (2012).
76. Libioulle, C. *et al.* Novel Crohn disease locus identified by genome-wide association maps to a gene desert on 5p13.1 and modulates expression of PTGER4. *PLoS Genet.* **3,** e58 (2007).
77. Vernot, B. *et al.* Personal and population genomics of human regulatory variation. *Genome Res.* http://dx.doi.org/10.1101/gr.134890.111 (2012).
78. Harismendy, O. *et al.* 9p21 DNA variants associated with coronary artery disease impair interferon-γ signalling response. *Nature* **470,** 264–268 (2011).
79. Cheng, C. *et al.* Understanding transcriptional regulation by integrative analysis of transcription factor binding data. *Genome Res.* http://dx.doi.org/10.1101/gr.136838.111 (2012).
80. Schuster, S. C. *et al.* Complete Khoisan and Bantu genomes from southern Africa. *Nature* **463,** 943–947 (2010).

Supplementary Information is available in the online version of the paper.

Acknowledgements We thank additional members of our laboratories and institutions who have contributed to the experimental and analytical components of this project. We thank D. Leja for assistance with production of the figures. The Consortium is funded by grants from the NHGRI as follows: production grants: U54HG004570 (B. E. Bernstein); U01HG004695 (E. Birney); U54HG004563 (G. E. Crawford); U54HG004557 (T. R. Gingeras); U54HG004555 (T. J. Hubbard); U41HG004568 (W. J. Kent); U54HG004576 (R. M. Myers); U54HG004558 (M. Snyder); U54HG004592 (J. A. Stamatoyannopoulos). Pilot grants: R01HG003143 (J. Dekker); RC2HG005591 and R01HG003700 (M. C. Giddings); R01HG004456-03 (Y. Ruan); U01HG004571 (S. A. Tenenbaum); U01HG004561 (Z. Weng); RC2HG005679 (K. P. White). This project was supported in part by American Recovery and Reinvestment Act (ARRA) funds from the NHGRI through grants U54HG004570, U54HG004563, U41HG004568, U54HG004592, R01HG003143, RC2HG005591, R01HG003541, U01HG004561, RC2HG005679 and R01HG003988 (L. Pennacchio). In addition, work from NHGRI Groups was supported by the Intramural Research Program of the NHGRI (L. Elnitski, ZIAHG200323; E. H. Margulies, ZIAHG200341). Research in the Pennachio laboratory was performed at Lawrence Berkeley National Laboratory and at the United States Department of Energy Joint Genome Institute, Department of Energy Contract DE-AC02-05CH11231, University of California.

Author Contributions See the consortium author list for details of author contributions.

Author Information The Supplementary Information is accompanied by a Virtual Machine (VM) containing the functioning analysis data and code. Further details of the VM are available from http://encodeproject.org/ENCODE/integrativeAnalysis/VM. Reprints and permissions information is available at www.nature.com/reprints. This paper is distributed under the terms of the Creative Commons Attribution-Non-Commercial-Share Alike licence, and the online version of the paper is freely available to all readers. The authors declare no competing financial interests. Readers are welcome to comment on the online version of the paper. Correspondence and requests for materials should be addressed to E.B. (birney@ebi.ac.uk).

The ENCODE Project Consortium

Overall coordination (data analysis coordination) Ian Dunham[1], Anshul Kundaje[2]†; **Data production leads (data production)** Shelley F. Aldred[3], Patrick J. Collins[3], Carrie A. Davis[4], Francis Doyle[5], Charles B. Epstein[6], Seth Frietze[7], Jennifer Harrow[8], Rajinder Kaul[9], Jainab Khatun[10], Bryan R. Lajoie[11], Stephen G. Landt[12], Bum-Kyu Lee[13],

©2012 Macmillan Publishers Limited. All rights reserved

Florencia Pauli[14], Kate R. Rosenbloom[15], Peter Sabo[16], Alexias Safi[17], Amartya Sanyal[11], Noam Shoresh[6], Jeremy M. Simon[18], Lingyun Song[17], Nathan D. Trinklein[3]; **Lead analysts (data analysis)** Robert C. Altshuler[19], Ewan Birney[1], James B. Brown[20], Chao Cheng[21], Sarah Djebali[22], Xianjun Dong[23], Ian Dunham[1], Jason Ernst[19]†, Terrence S. Furey[24], Mark Gerstein[21], Belinda Giardine[25], Melissa Greven[23], Ross C. Hardison[25,26], Robert S. Harris[25], Javier Herrero[1], Michael M. Hoffman[16], Sowmya Iyer[27], Manolis Kellis[19], Jainab Khatun[10], Pouya Kheradpour[19], Anshul Kundaje[2]†, Timo Lassmann[28], Qunhua Li[20]†, Xinying Lin[23], Georgi K. Marinov[29], Angelika Merkel[22], Ali Mortazavi[30], Stephen C. J. Parker[31], Timothy E. Reddy[14]†, Joel Rozowsky[21], Felix Schlesinger[4], Robert E. Thurman[16], Jie Wang[23], Lucas D. Ward[19], Troy W. Whitfield[23], Steven P. Wilder[1], Weisheng Wu[25], Hualin S. Xi[32], Kevin Y. Yip[21]†, Jiali Zhuang[23]; **Writing group** Bradley E. Bernstein[6,33], Ewan Birney[1], Ian Dunham[1], Eric D. Green[34], Chris Gunter[14], Michael Snyder[12]; **NHGRI project management (scientific management)** Michael J. Pazin[35], Rebecca F. Lowdon[35]†, Laura A. L. Dillon[35]†, Leslie B. Adams[35], Caroline J. Kelly[35], Julia Zhang[35]†, Judith R. Wexler[35]†, Eric D. Green[34], Peter J. Good[35], Elise A. Feingold[35]; **Principal investigators (steering committee)** Bradley E. Bernstein[6,33], Ewan Birney[1], Gregory E. Crawford[17,36], Job Dekker[11], Laura Elnitski[37], Peggy J. Farnham[7], Mark Gerstein[21], Morgan C. Giddings[10], Thomas R. Gingeras[4,38], Eric D. Green[34], Roderic Guigó[22,39], Ross C. Hardison[25,26], Timothy J. Hubbard[8], Manolis Kellis[19], W. James Kent[15], Jason D. Lieb[18], Elliott H. Margulies[31]†, Richard M. Myers[14], Michael Snyder[12], John A. Stamatoyannopoulos[40], Scott A. Tenenbaum[5], Zhiping Weng[23], Kevin P. White[41], Barbara Wold[29,42]; **Boise State University and University of North Carolina at Chapel Hill Proteomics groups (data production and analysis)** Jainab Khatun[10], Yanbao Yu[43], John Wrobel[10], Brian A. Risk[10], Harsha P. Gunawardena[43], Heather C. Kuiper[43], Christopher W. Maier[43], Ling Xie[43], Xian Chen[43], Morgan C. Giddings[10]; **Broad Institute Group (data production and analysis)** Bradley E. Bernstein[6,33], Charles B. Epstein[6], Noam Shoresh[6], Jason Ernst[19]†, Pouya Kheradpour[19], Tarjei S. Mikkelsen[6], Shawn Gillespie[33], Alon Goren[6,33], Oren Ram[6,33], Xiaolan Zhang[6], Li Wang[6], Robbyn Issner[6], Michael J. Coyne[6], Timothy Durham[6], Manching Ku[6,33], Thanh Truong[6], Lucas D. Ward[19], Robert C. Altshuler[19], Matthew L. Eaton[19], Manolis Kellis[19]; **Cold Spring Harbor, University of Geneva, Center for Genomic Regulation, Barcelona, RIKEN, Sanger Institute, University of Lausanne, Genome Institute of Singapore group (data production and analysis)** Sarah Djebali[22], Carrie A. Davis[4], Angelika Merkel[22], Alex Dobin[4], Timo Lassmann[28], Ali Mortazavi[30], Andrea Tanzer[22], Julien Lagarde[22], Wei Lin[4], Felix Schlesinger[4], Chenghai Xue[4], Georgi K. Marinov[29], Jainab Khatun[10], Brian A. Williams[29], Chris Zaleski[4], Joel Rozowsky[21], Maik Röder[22], Felix Kokocinski[8]†, Rehab F. Abdelhamid[28], Tyler Alioto[22,44], Igor Antoshechkin[29], Michael T. Baer[4], Philippe Batut[4], Ian Bell[45], Kimberly Bell[4], Sudipto Chakrabortty[4], Xian Chen[43], Jacqueline Chrast[46], Joao Curado[22], Thomas Derrien[22]†, Jorg Drenkow[4], Erica Dumais[45], Jackie Dumais[45], Radha Duttagupta[45], Megan Fastuca[4], Kata Fejes-Toth[4]†, Pedro Ferreira[22], Sylvain Foissac[45], Melissa J. Fullwood[47]†, Hui Gao[45], David Gonzalez[22], Assaf Gordon[4], Harsha P. Gunawardena[43], Cédric Howald[46], Sonali Jha[4], Rory Johnson[22], Philipp Kapranov[45]†, Brandon King[29], Colin Kingswood[22,44], Guoliang Li[48], Oscar J. Luo[47], Eddie Park[30], Jonathan B. Preall[4], Kimberly Presaud[4], Paolo Ribeca[22,44], Brian A. Risk[10], Daniel Robyr[49], Xiaoan Ruan[47], Michael Sammeth[22,44], Kuljeet Singh Sandhu[47], Lorain Schaeffer[29], Lei-Hoon See[4], Atif Shahab[47], Jorgen Skancke[22], Ana Maria Suzuki[28], Hazuki Takahashi[28], Hagen Tilgner[22]†, Diane Trout[29], Nathalie Walters[46], Huaien Wang[4], John Wrobel[10], Yanbao Yu[43], Yoshihide Hayashizaki[28], Jennifer Harrow[8], Mark Gerstein[21], Timothy J. Hubbard[8], Alexandre Reymond[46], Stylianos E. Antonarakis[49], Gregory J. Hannon[4], Morgan C. Giddings[10], Yijun Ruan[47], Barbara Wold[29,42], Piero Carninci[28], Roderic Guigó[22,39], Thomas R. Gingeras[4,38]; **Data coordination center at UC Santa Cruz (production data coordination)** Kate R. Rosenbloom[15], Cricket A. Sloan[15], Katrina Learned[15], Venkat S. Malladi[15], Matthew C. Wong[15], Galt P. Barber[15], Melissa S. Cline[15], Timothy R. Dreszer[15], Steven G. Heitner[15], Donna Karolchik[15], W. James Kent[15], Vanessa M. Kirkup[15], Laurence R. Meyer[15], Jeffrey C. Long[15], Morgan Maddren[15], Brian J. Raney[15]; **Duke University, EBI, University of Texas, Austin, University of North Carolina-Chapel Hill group (data production and analysis)** Terrence S. Furey[24], Lingyun Song[17], Linda L. Grasfeder[18], Paul G. Giresi[18], Bum-Kyu Lee[13], Anna Battenhouse[13], Nathan C. Sheffield[17], Jeremy M. Simon[18], Kimberly A. Showers[18], Alexias Safi[17], Darin London[17], Akshay A. Bhinge[13], Christopher Shestak[18], Matthew R. Schaner[18], Seul Ki Kim[18], Zhuzhu Z. Zhang[18], Piotr A. Mieczkowski[50], Joanna O. Mieczkowska[18], Zheng Liu[13], Ryan M. McDaniell[13], Yunyun Ni[13], Naim U. Rashid[51], Min Jae Kim[18], Sheera Adar[18], Zhancheng Zhang[24], Tianyuan Wang[17], Deborah Winter[17], Damian Keefe[1], Ewan Birney[1], Vishwanath R. Iyer[13], Jason D. Lieb[18], Gregory E. Crawford[17,36]; **Genome Institute of Singapore group (data production and analysis)** Guoliang Li[48], Kuljeet Singh Sandhu[47], Meizhen Zheng[47], Ping Wang[47], Oscar J. Luo[47], Atif Shahab[47], Melissa J. Fullwood[47]†, Xiaoan Ruan[47], Yijun Ruan[47]; **HudsonAlpha Institute, Caltech, UC Irvine, Stanford group (data production and analysis)** Richard M. Myers[14], Florencia Pauli[14], Brian A. Williams[29], Jason Gertz[14], Georgi K. Marinov[29], Timothy E. Reddy[14]†, Jost Vielmetter[29,42], E. Christopher Partridge[14], Diane Trout[29], Katherine E. Varley[14], Clarke Gasper[29,42], Anita Bansal[14], Shirley Pepke[29,52], Preti Jain[14], Henry Amrhein[29], Kevin M. Bowling[14], Michael Anaya[29,42], Marie K. Cross[14], Brandon King[29], Michael A. Muratet[14], Igor Antoshechkin[29], Kimberly M. Newberry[14], Kenneth McCue[29], Amy S. Nesmith[14], Katherine I. Fisher-Aylor[29,42], Barbara Pusey[14], Gilberto DeSalvo[29,42], Stephanie L. Parker[14]†, Sreeram Balasubramanian[29,42], Nicholas S. Davis[14], Sarah K. Meadows[14], Tracy Eggleston[14], Chris Gunter[14], J. Scott Newberry[14], Shawn E. Levy[14], Devin M. Absher[14], Ali Mortazavi[30], Wing H. Wong[53], Barbara Wold[29,42]; **Lawrence Berkeley National Laboratory group (targeted experimental validation)** Matthew J. Blow[54], Axel Visel[54,55], Len A. Pennachio[54,55]; **NHGRI groups (data production and analysis)** Laura Elnitski[37], Elliott H. Margulies[31]†, Stephen C. J. Parker[31], Hanna M. Petrykowska[37]; **Sanger Institute, Washington University, Yale University, Center for Genomic Regulation, Barcelona, UCSC, MIT, University of Lausanne, CNIO group (data production and analysis)** Alexej Abyzov[21], Bronwen Aken[8], Daniel Barrell[8], Gemma Barson[8], Andrew Berry[8], Alexandra Bignell[8], Veronika Boychenko[8], Giovanni Bussotti[22], Jacqueline Chrast[46], Claire Davidson[8], Thomas Derrien[22]†, Gloria Despacio-Reyes[8], Mark Diekhans[15], Iakes Ezkurdia[56], Adam Frankish[8], James Gilbert[8], Jose Manuel Gonzalez[8], Ed Griffiths[8], Rachel Harte[15], David A. Hendrix[19], Cédric Howald[46], Toby Hunt[8], Irwin Jungreis[19], Mike Kay[8], Ekta Khurana[21], Felix Kokocinski[8]†, Jing Leng[21], Michael F. Lin[19], Jane Loveland[8], Zhi Lu[57], Deepa Manthravadi[8], Marco Mariotti[22], Jonathan Mudge[8], Gaurab Mukherjee[8], Cedric Notredame[22], Baikang Pei[21], Jose Manuel Rodriguez[56], Gary Saunders[8], Andrea Sboner[58], Stephen Searle[8], Cristina Sisu[21], Catherine Snow[8], Charlie Steward[8], Andrea Tanzer[22], Electra Tapanari[8], Michael L. Tress[56], Marijke J. van Baren[59]†, Nathalie Walters[46], Stefan Washietl[19], Laurens Wilming[8], Amonida Zadissa[8], Zhengdong Zhang[60], Michael Brent[59], David Haussler[61], Manolis Kellis[19], Alfonso Valencia[56], Mark Gerstein[21], Alexandre Reymond[46], Roderic Guigó[22,39], Jennifer Harrow[8], Timothy J. Hubbard[8]; **Stanford-Yale, Harvard, University of Massachusetts Medical School, University of Southern California/UC Davis group (data production and analysis)** Stephen G. Landt[12], Seth Frietze[7], Alexej Abyzov[21], Nick Addleman[12], Roger P. Alexander[21], Raymond K. Auerbach[21], Suganthi Balasubramanian[21], Keith Bettinger[12], Nitin Bhardwaj[21], Alan P. Boyle[12], Alina R. Cao[62], Philip Cayting[12], Alexandra Charos[63], Yong Cheng[12], Chao Cheng[21], Catharine Eastman[12], Ghia Euskirchen[12], Joseph D. Fleming[64], Fabian Grubert[12], Lukas Habegger[21], Manoj Hariharan[12], Arif Harmanci[21], Sushma Iyengar[65], Victor X. Jin[66], Konrad J. Karczewski[12], Maya Kasowski[12], Phil Lacroute[12], Hugo Lam[12], Nathan Lamarre-Vincent[64], Jing Leng[21], Jin Lian[67], Marianne Lindahl-Allen[64], Renqiang Min[21]†, Benoit Miotto[64], Hannah Monahan[63], Zarmik Moqtaderi[64], Xinmeng J. Mu[21], Henriette O'Geen[62], Zhengqing Ouyang[12], Dorrelyn Patacsil[12], Baikang Pei[21], Debasish Raha[63], Lucia Ramirez[12], Brian Reed[63], Joel Rozowsky[21], Andrea Sboner[58], Minyi Shi[12], Cristina Sisu[21], Teri Slifer[12], Heather Witt[7], Linfeng Wu[12], Xiaoqin Xu[62], Koon-Kiu Yan[21], Xinqiong Yang[12], Kevin Y. Yip[21]†, Zhengdong Zhang[60], Kevin Struhl[64], Sherman M. Weissman[67], Mark Gerstein[21], Peggy J. Farnham[7], Michael Snyder[12]; **University of Albany SUNY group (data production and analysis)** Scott A. Tenenbaum[5], Luiz O. Penalva[68], Francis Doyle[5]; **University of Chicago, Stanford group (data production and analysis)** Subhradip Karmakar[41], Stephen G. Landt[12], Raj R. Bhanvadia[41], Alina Choudhury[41], Marc Domanus[41], Lijia Ma[41], Jennifer Moran[41], Dorrelyn Patacsil[12], Teri Slifer[12], Alec Victorsen[41], Xinqiong Yang[12], Michael Snyder[12], Kevin P. White[41]; **University of Heidelberg group (targeted experimental validation)** Thomas Auer[69]†, Lazaro Centanin[69], Michael Eichenlaub[69], Franziska Gruhl[69], Stephan Heermann[69], Burkhard Hoeckendorf[69], Daigo Inoue[69], Tanja Kellner[69], Stephan Kirchmaier[69], Claudia Mueller[69], Robert Reinhardt[69], Lea Schertel[69], Stephanie Schneider[69], Rebecca Sinn[69], Beate Wittbrodt[69], Jochen Wittbrodt[69]; **University of Massachusetts Medical School Bioinformatics group (data production and analysis)** Zhiping Weng[23], Troy W. Whitfield[23], Jie Wang[23], Patrick J. Collins[3], Shelley F. Aldred[3], Nathan D. Trinklein[3], E. Christopher Partridge[14], Richard M. Myers[14]; **University of Massachusetts Medical School Genome Folding group (data production and analysis)** Job Dekker[11], Gaurav Jain[11], Bryan R. Lajoie[11], Amartya Sanyal[11]; **University of Washington, University of Massachusetts Medical Center group (data production and analysis)** Gayathri Balasundaram[70], Daniel L. Bates[16], Rachel Byron[70], Theresa K. Canfield[16], Morgan J. Diegel[16], Douglas Dunn[16], Abigail K. Ebersol[71], Tristan Frum[71], Kavita Garg[72], Erica Gist[16], R. Scott Hansen[71], Lisa Boatman[71], Eric Haugen[16], Richard Humbert[16], Gaurav Jain[11], Audra K. Johnson[16], Ericka M. Johnson[71], Tattyana V. Kutyavin[16], Bryan R. Lajoie[11], Kristen Lee[16], Dimitra Lotakis[71], Matthew T. Maurano[16], Shane J. Neph[16], Fiedencio V. Neri[16], Eric D. Nguyen[71], Hongzhu Qu[16], Alex P. Reynolds[16], Vaughn Roach[16], Eric Rynes[16], Peter Sabo[16], Minerva E. Sanchez[71], Richard S. Sandstrom[16], Amartya Sanyal[11], Anthony O. Shafer[16], Andrew B. Stergachis[16], Sean Thomas[16], Robert E. Thurman[16], Benjamin Vernot[16], Jeff Vierstra[16], Shinny Vong[16], Hao Wang[16], Molly A. Weaver[16], Yongqi Yan[71], Miaohua Zhang[70], Joshua M. Akey[16], Michael Bender[70], Michael O. Dorschner[73], Mark Groudine[70], Michael J. MacCoss[16], Patrick Navas[71], George Stamatoyannopoulos[71], Rajinder Kaul[9], Job Dekker[11], John A. Stamatoyannopoulos[40]; **Data Analysis Center (data analysis)**; Ian Dunham[1], Kathryn Beal[1], Alvis Brazma[74], Paul Flicek[1], Javier Herrero[1], Nathan Johnson[1], Damian Keefe[1], Margus Lukk[74]†, Nicholas M. Luscombe[75], Daniel Sobral[1]†, Juan M. Vaquerizas[75], Steven P. Wilder[1], Serafim Batzoglou[2], Arend Sidow[76], Nadine Hussami[2], Sofia Kyriazopoulou-Panagiotopoulou[2], Max W. Libbrecht[2]†, Marc A. Schaub[2], Anshul Kundaje[2]†, Ross C. Hardison[25,26], Webb Miller[25], Belinda Giardine[25], Robert S. Harris[25], Weisheng Wu[25], Peter J. Bickel[20], Balazs Banfai[20], Nathan P. Boley[20], James B. Brown[20], Haiyan Huang[20], Qunhua Li[20]†, Jingyi Jessica Li[20], William Stafford Noble[16,77], Jeffrey A. Bilmes[78], Orion J. Buske[16], Michael M. Hoffman[16], Avinash D. Sahu[16]†, Peter V. Kharchenko[79], Peter J. Park[79], Dannon Baker[80], James Taylor[80], Zhiping Weng[23], Sowmya Iyer[27], Xianjun Dong[23], Melissa Greven[23], Xinying Lin[23], Jie Wang[23], Hualin S. Xi[32], Jiali Zhuang[23], Mark Gerstein[21], Roger P. Alexander[21], Suganthi Balasubramanian[21], Chao Cheng[21], Arif Harmanci[21], Lucas Lochovsky[21], Renqiang Min[21]†, Xinmeng J. Mu[21], Joel Rozowsky[21], Koon-Kiu Yan[21], Kevin Y. Yip[21]† & Ewan Birney[1]

[1]Vertebrate Genomics Group, European Bioinformatics Institute (EMBL-EBI), Wellcome Trust Genome Campus, Hinxton, Cambridgeshire CB10 1SD, UK. [2]Department of Computer Science, Stanford University, 318 Campus Drive, Stanford, California 94305-5428, USA. [3]SwitchGear Genomics, 1455 Adams Drive Suite 1317, Menlo Park, California 94025, USA. [4]Functional Genomics, Cold Spring Harbor Laboratory, 1 Bungtown Road, Cold Spring Harbor, New York 11724, USA. [5]College of Nanoscale Sciences and Engineering, University ay Albany-SUNY, 257 Fuller Road, NFE 4405, Albany, New York 12203, USA. [6]Broad Institute of MIT and Harvard, 7 Cambridge Center, Cambridge, Massachusetts 02142, USA. [7]Biochemistry and Molecular Biology, USC/Norris Comprehensive Cancer Center, 1450 Biggy Street, NRT 6503, Los Angeles, California 90089, USA. [8]Informatics, Wellcome Trust Sanger Institute, Wellcome Trust Genome Campus, Hinxton, Cambridgeshire CB10 1SA, UK. [9]Department of Medicine, Division of Medical Genetics, University of Washington, 3720 15th Avenue NE, Seattle, Washington 98195, USA. [10]College of Arts and Sciences, Boise State University, 1910 University Drive, Boise, Idaho 83725, USA. [11]Program in Systems Biology, Program in Gene Function and Expression, Department of Biochemistry and Molecular

©2012 Macmillan Publishers Limited. All rights reserved

Pharmacology, University of Massachusetts Medical School, 364 Plantation Street, Worcester, Massachusetts 01605, USA. [12]Department of Genetics, Stanford University, 300 Pasteur Drive, M-344, Stanford, California 94305-5120, USA. [13]Center for Systems and Synthetic Biology, Institute for Cellular and Molecular Biology, Section of Molecular Genetics and Microbiology, The University of Texas at Austin, 1 University Station A4800, Austin, Texas 78712, USA. [14]HudsonAlpha Institute for Biotechnology, 601 Genome Way, Huntsville, Alabama 35806, USA. [15]Center for Biomolecular Science and Engineering, University of California, Santa Cruz, 1156 High Street, Santa Cruz, California 95064, USA. [16]Department of Genome Sciences, University of Washington, 3720 15th Ave NE, Seattle, Washington 98195-5065, USA. [17]Institute for Genome Sciences and Policy, Duke University, 101 Science Drive, Durham, North Carolina 27708, USA. [18]Department of Biology, Carolina Center for Genome Sciences, and Lineberger Comprehensive Cancer Center, The University of North Carolina at Chapel Hill, 408 Fordham Hall, Chapel Hill, North Carolina 27599-3280, USA. [19]Computer Science and Artificial Intelligence Laboratory, Broad Institute of MIT and Harvard, Massachusetts Institute of Technology, 32 Vassar Street, Cambridge, Massachusetts 02139, USA. [20]Department of Statistics, University of California, Berkeley, 367 Evans Hall, University of California, Berkeley, Berkeley, California 94720, USA. [21]Computational Biology and Bioinformatics Program, Yale University, 266 Whitney Avenue, New Haven, Connecticut 06520, USA. [22]Bioinformatics and Genomics, Centre for Genomic Regulation (CRG) and UPF, Doctor Aiguader, 88, Barcelona 08003, Catalonia, Spain. [23]Program in Bioinformatics and Integrative Biology, University of Massachusetts Medical School, 364 Plantation Street, Worcester, Massachusetts 01605, USA. [24]Department of Genetics, The University of North Carolina at Chapel Hill, 120 Mason Farm Road, CB 7240, Chapel Hill, North Carolina 27599, USA. [25]Center for Comparative Genomics and Bioinformatics, The Pennsylvania State University, Wartik Laboratory, University Park, Pennsylvania 16802, USA. [26]Department of Biochemistry and Molecular Biology, The Pennsylvania State University, 304 Wartik Laboratory, University Park, Pennsylvania 16802, USA. [27]Program in Bioinformatics, Boston University, 24 Cummington Street, Boston, Massachusetts 02215, USA. [28]RIKEN Omics Science Center, RIKEN Yokohama Institute, 1-7-22 Suehiro-cho, Tsurumi-ku, Yokohama, Kanagawa 230-0045, Japan. [29]Division of Biology, California Institute of Technology, 156-291200 East California Boulevard, Pasadena, California 91125, USA. [30]Developmental and Cell Biology and Center for Complex Biological Systems, University of California Irvine, 2218 Biological Sciences III, Irvine, California 92697-2300, USA. [31]Genome Technology Branch, National Human Genome Research Institute, 5625 Fishers Lane, Bethesda, Maryland 20892, USA. [32]Department of Biochemistry and Molecular Pharmacology, Bioinformatics Core, University of Massachusetts Medical School, 364 Plantation Street, Worcester, Massachusetts 01605, USA. [33]Howard Hughes Medical Institute and Department of Pathology, Massachusetts General Hospital and Harvard Medical School, 185 Cambridge St CPZN 8400, Boston, Massachusetts 02114, USA. [34]National Human Genome Research Institute, National Institutes of Health, 31 Center Drive, Building 31, Room 4B09, Bethesda, Maryland 20892-2152, USA. [35]National Human Genome Research Institute, National Institutes of Health, 5635 Fishers Lane, Bethesda, Maryland 20892-9307, USA. [36]Department of Pediatrics, Division of Medical Genetics, Duke University School of Medicine, Durham, North Carolina 27710, USA. [37]National Human Genome Research Institute, National Institutes of Health, 5625 Fishers Lane, Rockville, Maryland 20892, USA. [38]Affymetrix, Inc., 3380 Central Expressway, Santa Clara, California 95051, USA. [39]Departament de Ciències Experimentals i de la Salut, Universitat Pompeu Fabra, Barcelona, Catalonia 08002, Spain. [40]Department of Genome Sciences, Box 355065, and Department of Medicine, Division of Oncology, Box 358081, University of Washington, Seattle, Washington 98195-5065, USA. [41]Institute for Genomics and Systems Biology, The University of Chicago, 900 East 57th Street, 10100 KCBD, Chicago, Illinois 60637, USA. [42]Beckman Institute, California Institute of Technology, 156-29 1200 E. California Boulevard, Pasadena, California 91125, USA. [43]Department of Biochemistry and Biophysics, University of North Carolina School of Medicine, Campus Box 7260, 120 Mason Farm Road, 3010 Genetic Medicine Building, Chapel Hill, North Carolina 27599, USA. [44]Centro Nacional de Análisis Genómico (CNAG), C/Baldiri Reixac 4, Torre I, Barcelona, Catalonia 08028, Spain. [45]Genomics, Affymetrix, Inc., 3380 Central Expressway, Santa Clara, California 95051, USA. [46]Center for Integrative Genomics, University of Lausanne, Genopode Building, 1015 Lausanne, Switzerland. [47]Genome Technology and Biology, Genome Institute of Singapore, 60 Biopolis Street, 02-01, Genome, Singapore 138672, Singapore. [48]Computational and Systems Biology, Genome Institute of Singapore, 60 Biopolis Street, 02-01, Genome, Singapore 138672, Singapore. [49]Department of Genetic Medicine and Development, University of Geneva Medical School, and University Hospitals of Geneva, 1 rue Michel-Servet, 1211 Geneva 4, Switzerland. [50]Department of Genetics, The University of North Carolina at Chapel Hill, 5078 GMB, Chapel Hill, North Carolina 27599-7264, USA. [51]Department of Biostatistics, Gillings School of Global Public Health, The University of North Carolina at Chapel Hill, 408 Fordham Hall, Chapel Hill, North Carolina 27599-7445, USA. [52]Center for Advanced Computing Research, California Institute of Technology, MC 158-79, 1200 East California Boulavard, Pasadena, California 91125, USA. [53]Department of Statistics, Stanford University, Sequoia Hall. 390 Serra Mall, Stanford, California 94305-4065, USA. [54]DOE Joint Genome Institute, Walnut Creek, California, USA. [55]Genomics Division, Lawrence Berkeley National Laboratory, One Cyclotron Road, MS 84-171, Berkeley, California 94720, USA. [56]Structural Computational Biology, Spanish National Cancer Research Centre (CNIO), Melchor Fernandez Almagro, 3, 28029 Madrid, Spain. [57]School of Life Sciences, Tsinghua University, School of Life Sciences, Tsinghua University, 100084 Beijing, China. [58]Department of Pathology and Laboratory Medicine, Institute for Computational Biomedicine, Weill Cornell Medical College, 1305 York Avenue, Box 140, New York, New York 10065, USA. [59]Computer Science and Engineering, Washington University in St Louis, St Louis, Missouri 63130, USA. [60]Department of Genetics, Albert Einstein College of Medicine, 1301 Morris Park Avenue, Room 353A, Bronx, New York 10461, USA. [61]Center for Biomolecular Science and Engineering, Howard Hughes Medical Institute, University of California, Santa Cruz, 1156 High Street, Santa Cruz, California 95064, USA. [62]Genome Center, University of California-Davis, 451 Health Sciences Drive, Davis, California 95616, USA. [63]Department of Molecular, Cellular, and Developmental Biology, Yale University, 266 Whitney Avenue, New Haven, Connecticut 06511, USA. [64]Biological Chemistry and Molecular Pharmacology, Harvard Medical School, 240 Longwood Avenue, Boston, Massachusetts 02115, USA. [65]Biochemistry and Molecular Biology, University of Southern California, 1501 San Pablo Street, Los Angeles, California 90089, USA. [66]Department of Biomedical Informatics, Ohio State University, 3172C Graves Hall, 333 W Tenth Avenue, Columbus, Ohio 43210, USA. [67]Department of Genetics, Yale University, Yale University School of Medicine, 333 Cedar Street, New Haven, Connecticut 06510, USA. [68]Department of Cellular and Structural Biology, Children's Cancer Research Institute–UTHSCSA, Mail code 7784- 7703 Floyd Curl Dr, San Antonio, Texas 78229, USA. [69]Centre for Organismal Studies (COS) Heidelberg, University of Heidelberg, Im Neuenheimer Feld 230, 69120 Heidelberg, Germany. [70]Basic Sciences Division, Fred Hutchinson Cancer Research Center, 825 Eastlake Avenue East, Seattle, Washington 98109, USA. [71]Department of Medicine, Division of Medical Genetics, Box 357720, University of Washington, Seattle, Washington 98195-7720, USA. [72]Division of Human Biology, Fred Hutchinson Cancer Research Center, 825 Eastlake Avenue East, Seattle, Washington 98109, USA. [73]Department of Psychiatry and Behavioral Sciences, Box 356560, University of Washington, Seattle, Washington 98195-6560, USA. [74]Microarray Informatics Group, European Bioinformatics Institute (EMBL-EBI), Wellcome Trust Genome Campus, Hinxton, Cambridgeshire CB10 1SD, UK. [75]Genomics and Regulatory Systems Group, European Bioinformatics Institute (EMBL-EBI), Wellcome Trust Genome Campus, Hinxton, Cambridgeshire CB10 1SD, UK. [76]Department of Pathology, Department of Genetics, Stanford University, 300 Pasteur Drive, Stanford, California 94305, USA. [77]Department of Computer Science and Engineering, 185 Stevens Way, Seattle, Washington 98195, USA. [78]Department of Electrical Engineering, University of Washington, 185 Stevens Way, Seattle, Washington 98195, USA. [79]Center for Biomedical Informatics, Harvard Medical School, 10 Shattuck Street, Boston, Massachusetts 02115, USA. [80]Departments of Biology and Mathematics and Computer Science, Emory University, Atlanta, Georgia 30322, USA. †Present addresses: Computer Science and Artificial Intelligence Laboratory, Massachusetts Institute of Technology, 32 Vassar Street, Cambridge, Massachusetts 02139, USA (A.K.); UCLA Biological Chemistry Department, Eli and Edythe Broad Center of Regenerative Medicine and Stem Cell Research at UCLA, Jonsson Comprehensive Cancer Center, 615 Charles E Young Dr South, Los Angeles, California 90095, USA (J.E.); Department of Statistics, 514D Wartik Lab, Penn State University, State College, Pennsylvania 16802, USA (Q.L.); Department of Biostatistics and Bioinformatics and the Institute for Genome Sciences and Policy, Duke University School of Medicine, 101 Science Drive, Durham, North Carolina 27708, USA (T.E.R.); Department of Computer Science and Engineering, The Chinese University of Hong Kong, Shatin, New Territories, Hong Kong (K.Y.Y.); Department of Genetics, Washington University in St Louis, St Louis, Missouri 63110, USA (R.F.L.); Department of Cell Biology and Molecular Genetics, University of Maryland, College Park, Maryland 20742, USA (L.A.L.D.); National Cancer Institute, National Institutes of Health, Bethesda, Maryland 20892, USA (J.Z.); University of California, Davis Population Biology Graduate Group, Davis, California 95616, USA (J.R.W.); Illumina Cambridge Ltd., Chesterford Research Park, Little Chesterford, Saffron Walden, Essex CB10 1XL, UK (E.H.M.); BlueGnome Ltd., CPC4, Capital Park, Fulbourn, Cambridge CB21 5XE, UK (F.K.); Institut de Génétique et Développement de Rennes, CNRS-UMR6061, Université de Rennes 1, F-35000 Rennes, Brittany, France (T.D.); Caltech, 1200 East California Boulevard, Pasadena, California 91125, USA (K.F.-T.); A*STAR-Duke-NUS Neuroscience Research Partnership, 8 College Road, Singapore 169857, Singapore (M.J.F.); St Laurent Institute, One Kendall Square, Cambridge, Massachusetts 02139, USA (P.K.); Department of Genetics, Stanford University, Stanford, California 94305, USA (H.T.); Biomedical Sciences (BMS) Graduate Program, University of California, San Francisco, 513 Parnassus Avenue, HSE-1285, San Francisco, California 94143-0505, USA (S.L.P.); Monterey Bay Aquarium Research Institute, Moss Landing, California 95039, USA (M.J.v.B.); Department of Machine Learning, NEC Laboratories America, 4 Independence Way, Princeton, New Jersey 08540, USA (R.M.); Neuronal Circuit Development Group, Unité de Génétique et Biologie du Développement, U934/UMR3215, Institut Curie–Centre de Recherche, Pole de Biologie du Développement et Cancer, 26, rue d'Ulm, 75248 Paris Cedex 05, France (T.A.); Cancer Research UK, Cambridge Research Institute, Li Ka Shing Centre, Robinson Way, Cambridge CB2 0RE, UK (M.L.); Unidade de Bioinformatica, Rua da Quinta Grande, 6, P-2780-156 Oeiras, Portugal (D.S.); Department of Genome Sciences, University of Washington, 3720 15th Avenue NE, Seattle, Washington 98195-5065, USA (M.W.L.); Center for Bioinformatics and Computational Biology, 3115 Ag/Life Surge Building 296, University of Maryland, College Park, Maryland 20742, USA (A.D.S.).

©2012 Macmillan Publishers Limited. All rights reserved