Research Highlights

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

Discovery and saturation analysis of cancer genes across 21 tumour types
Lawrence, M.S. *et al. Nature* doi:10.1038/nature12912 (5 January 2014)

Rewiring yeast sugar transporter preference through modifying a conserved protein motif
Young, E.M. *et al. PNAS* **111**, 131–136 (2014)

Dynamic imaging of genomic loci in living human cells by an optimized CRISPR/Cas sysem
Chen, B. *et al. Cell* **155**, 1479–1491 (2013)

Genome-scale CRISPR-Cas9 knockout screening in human cells
Shalem, O. *et al. Science* **343**, 84–87 (2014)

Genetic screens in human cells using the CRISPR-Cas9 system
Wang, T. *et al. Science* **343**, 80–84 (2014)

PIQ-ing into chromatin architecture

Sebastian Rieck & Christopher Wright

Decrypting the logic of gene regulation is aided by a breakthrough method for analyzing chromatin-accessibility data.

The gene expression program of any cell depends on complex, dynamic patterns of transcription factor binding. Yet, extracting the biological meaning of such patterns has been limited by a dearth of technologies that can identify exactly where individual transcription factors are bound across an entire genome. In this issue, Sherwood *et al.*[1] describe notable progress toward this goal with a new computational method, called protein interaction quantification (PIQ), which from a single experiment infers the effects on chromatin architecture produced by hundreds of transcription factors. PIQ creates a cumulative catalog of the distinctive imprints caused by each DNA-binding transcription factor, in combination with their associated coregulators, on the patterns of DNA fragments produced by digesting chromatin with the enzyme DNase I. It uses these distinctive, individual DNase-sensitivity profiles to interpret how multiple transcription factors contribute to the genome-wide DNase hypersensitivity (DNase-seq) profile.

The study presents several important advances. First, using PIQ to reveal the bound sites of hundreds of transcription factors could, in some cases, allow a single DNase-seq experiment to replace hundreds of ChIP-seq experiments, each of which detects the binding of a single transcription factor. Second, compared with previous methods, PIQ allows a more sensible comparison of results from multiple experiments—for example, time points along a cell-differentiation process. The authors apply PIQ to a time series of DNase-sensitivity information taken from a cell differentiation model to enable a systematic categorization of transcription factors into the following functional subdivisions: pioneers, which are capable of opening closed chromatin; settlers, which require open chromatin to bind; and migrants, which appear to require more than open chromatin to bind, such as the presence of coregulators (**Fig. 1**). Third, the analysis seems to greatly expand the list of potential pioneers, adding ~120 new ones to the handful identified until now. Finally, the authors identify a new type of pioneer—the 'directional pioneer'—which induces chromatin to open preferentially in one direction.

Ordered, collaborative interactions between transcription factors and their coregulators affect the sequential building of patterns of chromatin and thereby gene activity across the genome. The experimental design and computational output emerging from Sherwood *et al.*[1] provide a template for unraveling the gene regulatory networks that dictate complex phenomena such as the state of cell stemness or differentiation. The authors also speculate that biologically critical gene-regulatory interactions, which depend upon both the distance and relative chromosomal location of motifs that bind directional pioneers and settlers (**Fig. 1b**,**c**), could be a significant component affecting the manner in which an organism's overall genome sequence may evolve over time.

Only a few transcription factors (such as Foxa, GATA, Pu.1 and glucocorticoid receptor) are known to have the ability to 'crack open' condensed chromatin. These pioneers function as gatekeepers: without them, other regulatory proteins cannot access their target sites (**Fig. 1a**). Sometimes even within the same cell, pioneer transcription factors activate or repress different genes; regulated toggling between such dual functions would be a potent influence, for instance, in how cells choose between separate pathways of cell differentiation[2]. Despite huge amounts of work on individual factors, and the invention of methods that attempt to describe chromosome-wide DNA-protein interactions, a systematic decoding of how chromatin architecture is progressively built during differentiation is still far from complete. One hurdle is that genome-wide methods to assay chromatin accessibility are technically challenging and inherently noisy.

The method of Sherwood *et al.*[1] (**Fig. 2**) for discovering bound transcription-factor sites and their particular effects on chromatin structure seems to offer a large improvement over previous approaches for analyzing DNase-seq data sets. DNase-seq involves subjecting carefully prepared chromatin to DNase-I cleavage, and using high-throughput techniques for en masse sequencing of the released DNA fragments. Accurately aligning the DNA fragment ends on the genome produces a DNase-seq profile indicating those regions most sensitive to DNase I and therefore accessible to protein binding. Transcription-factor binding can cause substantial loosening of nearby chromatin structure, increasing the sensitivity of DNA to DNase attack, but directly at the protein-DNA interface there is often a relative protection of DNA against DNase cleavage—a so-called 'footprint'. The end result is a pattern of characteristic local contours in the total DNase-sensitivity profile. The identity of proteins bound to the protected sequences, and thus their local chromatin effects, can be determined *de novo*[3] or conversely by referring to databases of the DNA motifs recognized by transcription factors[4].

Sherwood *et al.*[1] (**Fig. 2**) take the latter approach—using known transcription factor binding motifs—but made three key refinements to improve interpretation of DNase profiles[3,4]. First, PIQ smooths raw sequence read information, decreasing the noise inherent to DNase-seq[5]. In this step, the algorithm can also incorporate data from additional DNase-seq experiments to reduce noise even further, which strengthens the ability to carry

Sebastian Rieck and Christopher Wright are at the Department of Cell and Developmental Biology, Center for Stem Cell Biology and Program in Developmental Biology, Vanderbilt University, Nashville, Tennessee, USA.
e-mail: chris.wright@vanderbilt.edu

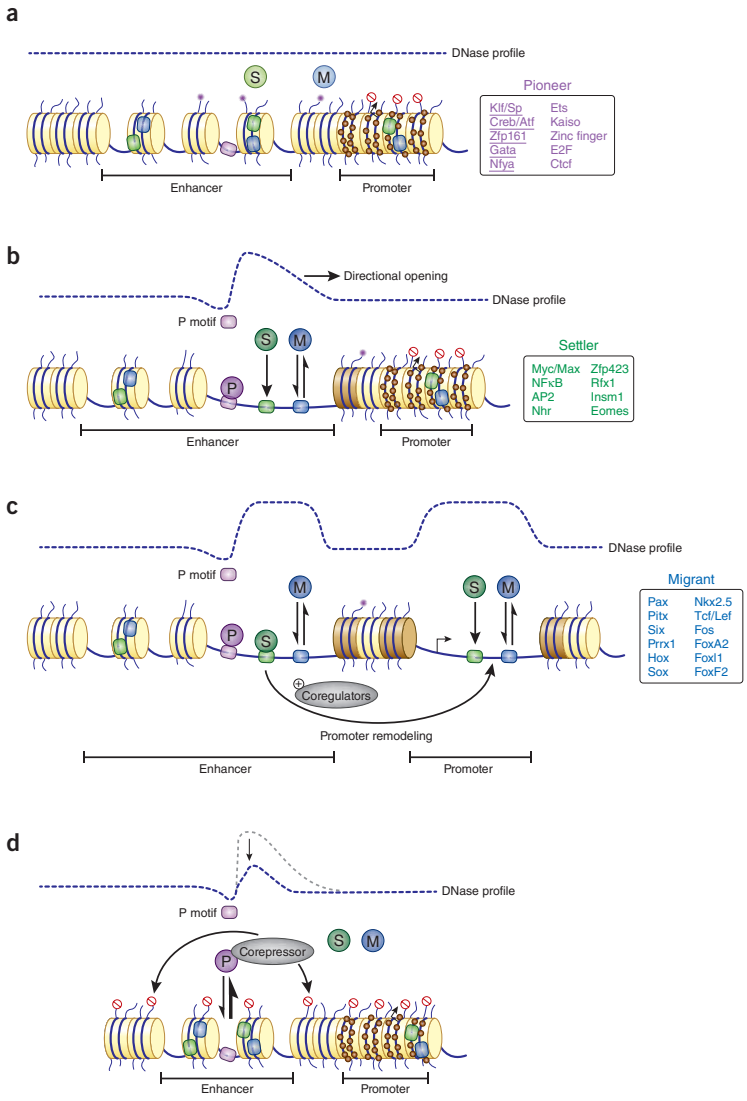

Figure 1 Hierarchical binding of transcription factors regulates gene expression. For the type of gene activation in which an enhancer opens up first[10], three steps are shown (**a**–**c**), with an idealized PIQ-derived DNase profile (blue dashed line) above each step. A non-exhaustive list of transcription factors involved is shown on the right; directional pioneers are underlined (see Supplementary Table 4 of Sherwood *et al.*[1] for a full list). (**a**) The enhancer is initially relatively inaccessible. The promoter is in a closed chromatin configuration, with DNA methylation (brown dots) and repressive histone modifications ('no entry' symbol, red circle with diagonal line). Although present, settler (light green) and migrant (light blue) transcription factors cannot bind their target sites. Note the lack of DNase-hypersensitive regions. (**b**) Certain pioneers (dark purple) can bind to their motif (P motif) within condensed chromatin and open chromatin asymmetrically, exposing a defined set of settler and migrant binding sites. Progressive chromatin opening is facilitated by certain histone variants (brown cylinders). Distinct histone modifications (purple fuzzy circles) may act as pioneer attractors[2]. (**c**) Settler factors (dark green) now access their motifs. Migrants (dark blue) bind sporadically to their binding targets in open chromatin, indicated by the equilibrium symbol. Settler recruitment at the enhancer can initiate self-sustained promoter remodeling, such as changes in nucleosome positioning and sometimes including additional settler and migrant occupancy near the transcriptional start site of the promoter, and gene activation[10]. (**d**) One possible mechanism that restricts the normal sequence of gene activation. Pioneer factors recruit corepressors, resulting in deposition of repressive histone modifications that holds back gene activation by limiting pioneer factor–based chromatin opening, and that limits binding of settlers and migrants.

out between-experiment comparisons. Second, the algorithm learns how the particular shapes of the local contours in the DNase profile are aligned over the target sites for, and discretely connected to, each transcription factor. The associated loosening of chromatin structure may have a specific DNase profile shape and sometimes extends hundreds of nucleotides away from the motif bound by the transcription factor. The contour information can then be used to extricate individual contributions to complicated DNase-seq profiles, such as those arising from the more-condensed motif spacing that often occurs at enhancers and promoters. The analysis allows the identification of transcription factors that have asymmetric effects on chromatin structure, in which chromatin is made much looser on one side of the actual bound motif. The output of PIQ is the probability, for every motif in the genome, that a protein is bound at that location. For a convincing validation of the correctness of these binding calls, Sherwood *et al.*[1] compared their PIQ analysis of the human K562 cell line against more than three hundred ChIP-seq datasets on the same cell type, which came from the ENCODE project[6]. The results were highly concordant, but PIQ was notably more accurate than existing DNase-seq analysis methods.

To further demonstrate the utility of PIQ, Sherwood *et al.*[1] analyze mouse embryonic stem cells being entrained *in vitro* to differentiate into either prepancreatic or intestinal endoderm. The authors generate DNase-seq profiles at six different cell states along a differentiation time course. Overall, binding events are identified for 733 transcription factors, and the observed relationships between protein binding and chromatin architecture are used to build and experimentally validate a rational, step-wise model of transcription factor binding (**Fig. 1b**,**c**).

This study adds to our list of transcription factors with intrinsic chromatin-opening capacity, and takes a large step toward a full categorization of which factors depend on a pre-existing open-chromatin state for binding. Of the 733 transcription factors identified, 120 are estimated to be pioneers, and 131 settlers. Four factors (among other nonverified candidates) are found to be directional pioneers, which open chromatin preferentially to one side or another of the motif. This directional chromatin opening activity is functionally validated using elegant reporter gene assays that compare the effects caused by different relative locations of pioneer and settler sites. The remaining factors seem to be neither pioneers nor settlers and, somewhat ambiguously are classified as 'migrants' to reflect their sporadic binding even to available open motifs. It should be remembered that pioneers and settlers have biologically relevant relationships, and in at least some contexts positively reinforce each other's binding[7].

An exciting potential application of PIQ is the improvement of protocols for controlling cell fate. Learning how to guide interactions between the genome and transcription factors, including how the most-instructive transcription factor genes (especially those encoding pioneers or directional pioneers) are controlled in response to extracellular signals, could speed

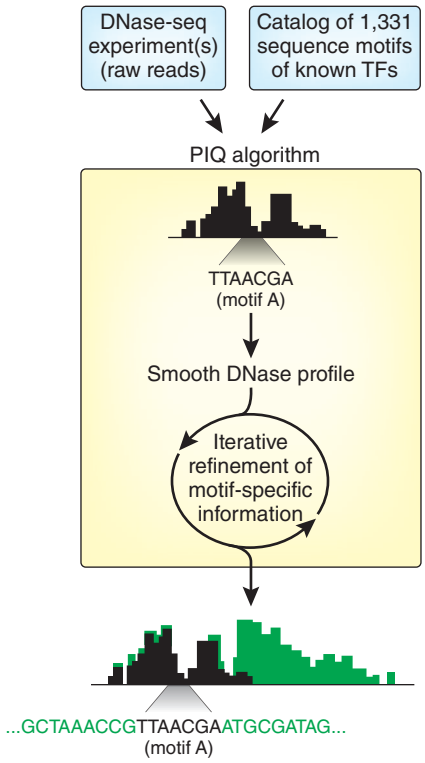

Figure 2 PIQ improves the identification of occupied transcription factor (TF) binding sites *in vivo* from genome-wide DNase I hypersensitivity data. The input for PIQ is one or more DNase-seq experiments and 1,331 transcription factor motifs represented as position-weight matrices. First, PIQ scans a DNase profile to determine whether each individual motif is located at a relative gap in DNase hypersensitivity, indicating that a protein occupies the site. A unique signature profile is shown for a hypothetical motif A (black). Second, raw reads are smoothed, by modeling them as arising from a Gaussian distribution, over each experiment to increase robustness to low coverage and noise. In the last step, estimation of motif-specific information and computer-generated, hypersensitivity-profile building steps are iteratively performed to calculate the probability that each motif is bound. Motif-specific information about the expected DNase profile surrounding a bound site allows for deciphering of complex profiles at enhancers and promoters. As a hypothetical example, motif A (black) is shown to be responsible for the left half of the total DNase profile (green) of a putative enhancer subregion. Integration of time-series data by PIQ can elucidate hierarchical and directional sequences of transcription factor binding.

the development of efficient cell replacement or regeneration therapies for human disease. Current *in vitro* protocols attempting to make large numbers of pancreatic beta cells from human embryonic stem cells mostly induce a slightly incorrect, off-track developmental program[8]. Does mismanagement of pioneer factor–dependent changes in chromatin architecture prevent entry into or failure to maintain a proper differentiation track? As modeled in **Figure 1d**, inefficient entrainment might reflect inappropriate corepressor recruitment by pioneer factors[9]. Deficiencies in the first rounds of settler-migrant binding (either absence or a reduced rate) could obstruct a population-wide change toward the next stage of cell differentiation. Thus, PIQ may identify novel transcriptional manipulations to increase differentiation toward the desired cell lineages.

A linked notion is that scoring by PIQ analysis could provide a deeper assurance that differentiation protocols are moving along correctly. It will be interesting to see to what degree PIQ analyses can be carried out on differentiating cell populations captured from a natural *in vivo* context. Although its robustness to noise implies that PIQ could help analyze DNase-seq data from smaller numbers cells, experiments to define its lower limit of sensitivity should be extremely useful.

PIQ is limited in that it calculates DNase footprints only for single transcription factors rather than explicitly performing a joint modeling of footprints and DNase profiles for multiple transcription factors. Moreover, Sherwood *et al.*[1] did not use PIQ to address if and how transcription factors are capable of mediating chromatin repression. Somewhat conspicuously, they did not identify Foxa1 as a pioneer factor, even though it is one of the few thoroughly validated pioneers[2]. They venture that this 'miss' is either explained by the cell type assayed or because Foxa1 may have both chromatin opening and compacting activity.

The results of Sherwood *et al.*[1] suggest that PIQ is an important new tool for understanding how transcription factor expression and chromatin architecture are integrated at a functional level. It should be useful for assessing how histone modifications and other regulatory mechanisms, such as DNA methylation and passive forms of cooperativity between transcription factors[2,7], enhance or restrict pioneer factor binding *in vivo*. There seems to be no reason why PIQ could not be equally applicable to animal or plant tissues, and to gain insight into the dynamic remodeling of chromatin not only during embryogenesis, organogenesis and normal tissue maintenance but also during the abnormal processes that occur in metaplasia and cancer.

COMPETING FINANCIAL INTERESTS
The authors declare no competing financial interests.

1. Sherwood, R.I. *et al. Nat. Biotechnol.* **32**,171–178 (2014).
2. Zaret, K.S. & Carroll, J.S. *Genes Dev.* **25**, 2227–2241 (2011).
3. Boyle, A.P. *et al. Cell* **132**, 311–322 (2008).
4. Pique-Regi, R. *et al. Genome Res.* **21**, 447–455 2011).
5. Neph, S. *et al. Nature* **489**, 83–90 (2012).
6. ENCODE Project Consortium. *et al. Nature.* **447**, 799–816 (2007).
7. Soufi, A., Donahue, G. & Zaret, K.S. *Cell* **151**, 994–1004 (2012).
8. Yang, Y.P. & Wright, C. *Nat. Chem. Biol.* **5**, 195–196 2009).
9. Rais, Y. *et al. Nature* **502**, 65–70 (2013).
10. Taberlay, P.C. *et al. Cell* **147**, 1283–1294 (2011).

Of snowflakes and natural killer cell subsets

Lewis L Lanier

Mass cytometric analysis reveals a remarkable and unanticipated degree of phenotypic heterogeneity among human peripheral blood natural killer cells.

In the beginning, natural killer (NK) cells were often referred to as 'null' cells because they were thought not to express any defining cell surface markers that could be used to identify and distinguish them from other classes of leukocytes, and because their origins and relationships to other hematopoietic cells were not understood. In a recent issue of *Science Translational Medicine*, Blish and colleagues[1] definitively render this 'null' moniker null and void by revealing the existence of more than 6,000–30,000 phenotypically distinct NK cell subsets in the blood of any single human. Although it will be challenging to elucidate the functional and physiological meaning of each of these NK cell subsets, the appreciation that such extensive heterogeneity exists within the NK cell lineage implies that this diversity may have evolved to deal with specific pathogens and to detect transformed or stressed host cells.

Lewis L. Lanier is in the Department of Microbiology and Immunology and the Cancer Research Institute, University of California San Francisco, San Francisco, California, USA.
e-mail: lewis.lanier@ucsf.edu