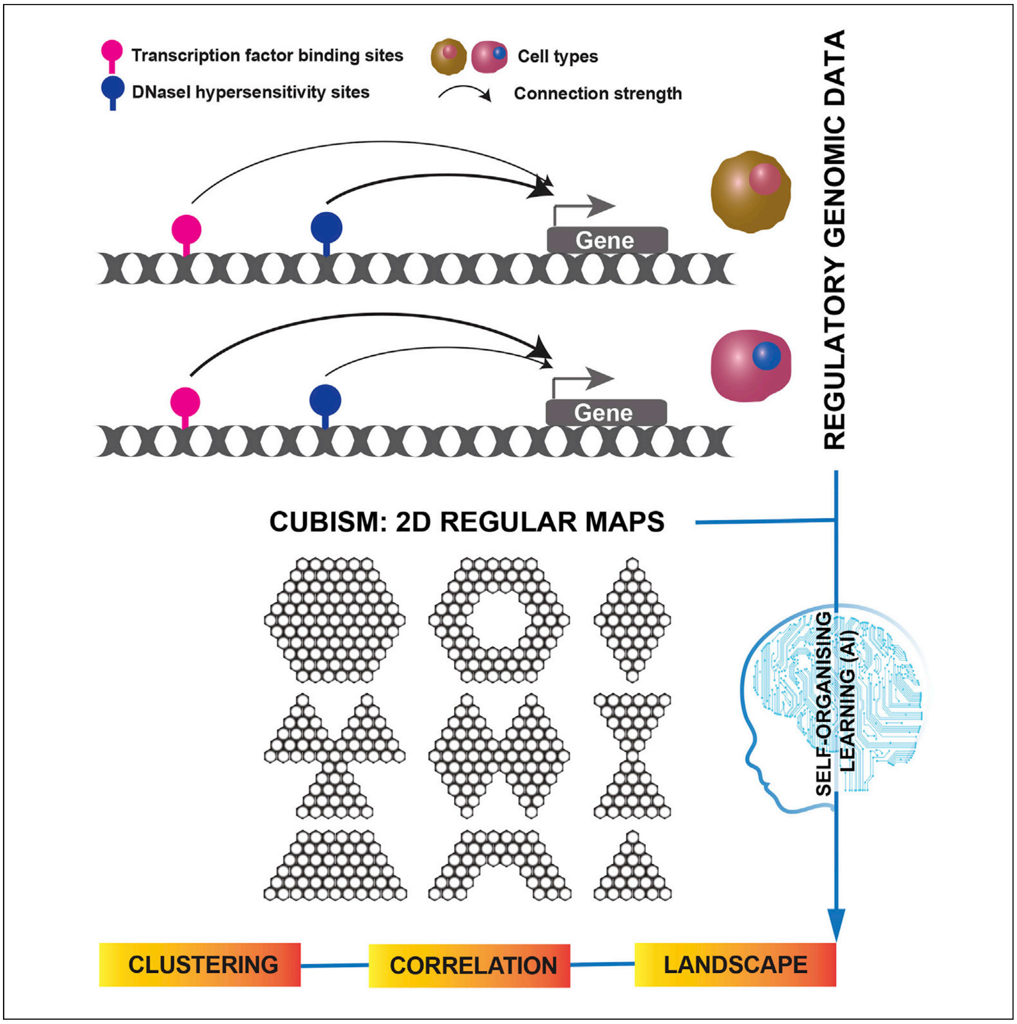
iScience

CellPress

Article

Regulatory Genomic Data Cubism



Hai Fang, Kankan Wang

hfang@well.ox.ac.uk (H.F.)
kankanwang@shsmu.edu.cn (K.W.)

HIGHLIGHTS

Realization of Picasso's cubism in regulatory genomics

Enables representation and comparison of patterns in regulatory genomic data

Able to analyze data at multilayer levels and involving different cell types

A general strategy for regulatory genomic data analysis

Fang & Wang, iScience 3, 217–225
May 25, 2018 © 2018 The Author(s).
https://doi.org/10.1016/j.isci.2018.04.017

iScience

CellPress

Article

Regulatory Genomic Data Cubism

Hai Fang[1,2,3,*] and Kankan Wang[1,*]

SUMMARY

A regularly shaped grid is useful for analyzing data particularly at multilayer levels, where patterns can be visually represented and analytically compared—conceptually similar to Picasso's cubism. Here we introduce ATLAS, featuring a suite of spatially ordered maps designed for representation and comparison of patterns seen in regulatory genomic data. It produces a landscape learned from input data and enables landscape-guided correlation with additional data. We illustrate its use for multilayer data comparison on the same cell type, and for comparisons involving different cell types, revealing information in a scientifically insightful and also visually intuitive way. The data-driven and visual-aided ability of ATLAS presents a general strategy for regulatory genomic data analysis.

INTRODUCTION

Now we are in the era of doing genomic data science; an important element of this concept is how to represent and model genomic data in a scientifically sound and also visually arresting way. Identifying and comparing the occurrence of patterns hidden in genomic data is the very first challenge to data scientists (Marx, 2013). How to intuitively characterize regulatory genomic data? How to analytically compare data at the multilayer levels? Could such characterizations and comparisons be made at the target gene level? We address these challenges by considering routinely generated types of regulatory genomic data. Regulatory genomic data are essentially in the form of non-coding genomic regions associated with signals thereof, such as transcription factor (TF) binding and DNase I hypersensitivity sites (DHSs), that are profiled in the same cell type or the same data type produced across cell types (The ENCODE Project Consortium, 2012).

Data-driven models are needed for achieving effective representation, comparison, and exploration of data. Picasso's cubist painting is a representation of the natural forms reduced into basic geometric regulars on the 2D map, known as "cubism". The use of a regularly shaped grid to "see" (model) data is attractive. We human beings are good at perceiving and comparing regular grids such as hexagons, and rightly using them could provide an information-rich overview of data. Designing such grids, however, is not trivial; *a priori* knowledge of data structure is usually unknown or limited, and the need for rich regular shapes should be met.

In addition to regular grids, models that are able to learn data are desirable. A self-organizing learning algorithm is a special type of artificial intelligence (AI), suitable for this purpose (Zhang and Fang, 2012). We previously described such model based on a supra-hexagonal map (Fang and Gough, 2014a), successfully applied to genomic data analysis (Allman et al., 2016; Jiang et al., 2016). The map grid of this shape, however, is not without limitations; the underlying structure of data is not necessarily distributed as a radial symmetry, and this limitation should be overcome.

We have created a system to realize regulatory genomic data cubism, designed for the self-organized representation and comparison of the occurrence of the patterns seen in regulatory genomic data. We refer to such system as a taught landscape analytic system, or called the "ATLAS": *taught*, because of the data-driven ability in a self-organizing manner; *landscape*, because of a global view in a visual-friendly way; and *analytic*, because of the support for quantitative analyses. This system is also highlighted by the support for linking regulatory elements to target genes and the support for correlation analysis involving data at multilayer levels. We demonstrate the use by showing how to correlate multilayer regulatory genomic data in a leukemia cell type, how to correlate the same layered genomic data involving two cell types, and how to link TFs to datasets generated from clustered regularly interspaced short palindromic repeats (CRISPR). In a wider context, ATLAS adds a new strategy to the repository of AI applied to the big regulatory genomic data analysis.

[1]State Key Laboratory of Medical Genomics and Shanghai Institute of Hematology, Ruijin Hospital, Shanghai Jiao Tong University School of Medicine, Shanghai 200025, China

[2]Wellcome Trust Centre for Human Genetics, University of Oxford, Oxford OX3 7BN, UK

[3]Lead Contact

*Correspondence:
hfang@well.ox.ac.uk (H.F.),
kankanwang@shsmu.edu.cn
(K.W.)

https://doi.org/10.1016/j.isci.2018.04.017

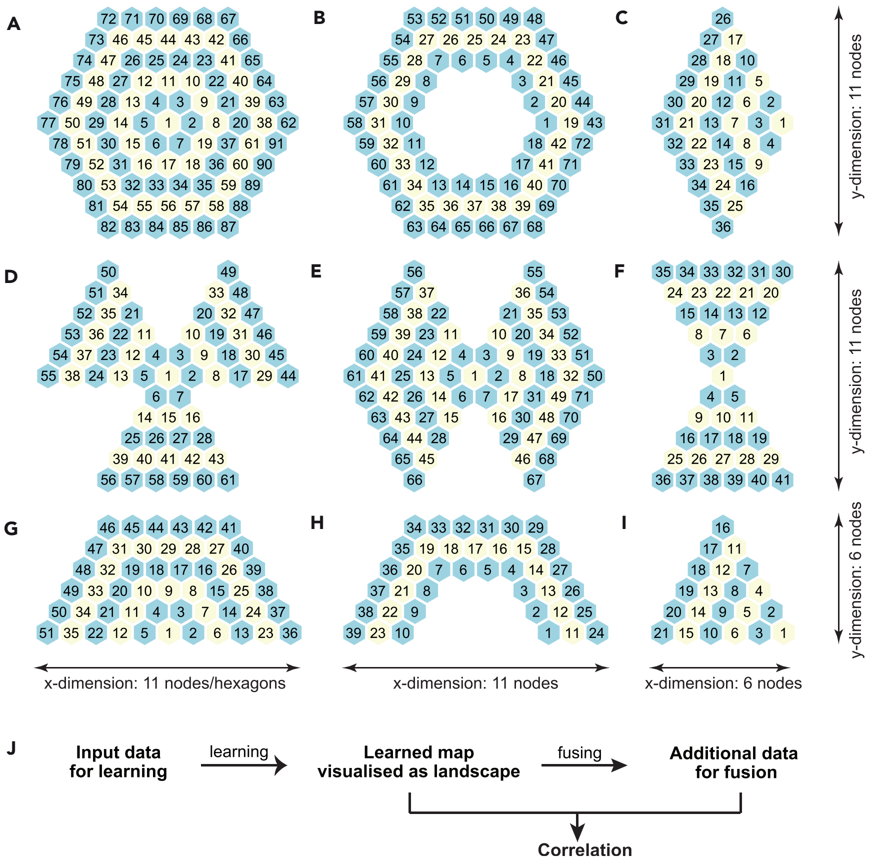


Figure 1. Realization of Cubism by Creating a Suite of Spatially Ordered 2D Maps

(A) A supra-hexagonal map. Uniquely determined by the radius, node indexed and expanding outward (colored alternatively in yellow and blue).

(B–I) Map variants sharing the basis of architectural design. They all have the same radius and nodes indexed in the same way, including a ring-shaped map (B), a diamond map (C), a trefoil map (D), a butterfly map (E), an hourglass map (F), a ladder map (G), a bridge map (H), and a triangle-shaped map (I).

(J) Backbone of analysis: a landscape learned from input data and then the landscape-guided correlation with additional data.

RESULTS

Realizing Cubism by Creating a Suite of Spatially Ordered 2D Maps

We devised a range of spatially ordered 2D maps, all designed on the basis of architecture used in a supra-hexagonal map. Illustrated in Figure 1A is the supra-hexagonal map with nodes indexed in a way that they radiate circularly outward. Its map variants (Figures 1B–1I) are all indexed in the same way. We achieved this by designing, for example, the ladder-shaped map (Figure 1H) derived from the top half part of the supra-hexagonal map (Figure 1A). In addition to the same architectural design, we modeled maps of all shapes following the same principle: learned from input data using the self-organizing learning algorithm (Transparent Methods). In brief, the learning achieves the conversion from the input data matrix to the codebook matrix associated with the learned map, constrained by the map shape. The learning process consists of (1) the choice of the map shape, empirically determined according to the input data (that is, informed by visualizing the input data onto a 2D hyperspace by principle component analysis [PCA]) and (2) the learning of the map, automatically achieved (that is, iteratively identifying the winner node and updating its neighbors). The process is largely data-driven, although the user can explicitly choose map shapes, if *a priori* knowledge strongly suggests doing so. The learned map (the codebook matrix) is visualized as the landscape and is fused with additional data for correlation analysis (Figure 1J). In summary, we have provided wide choices of 2D regular maps (“cubism”), enabling input data to be represented on a case-by-case basis, and subsequently effective comparison with additional data (usually another layered data).

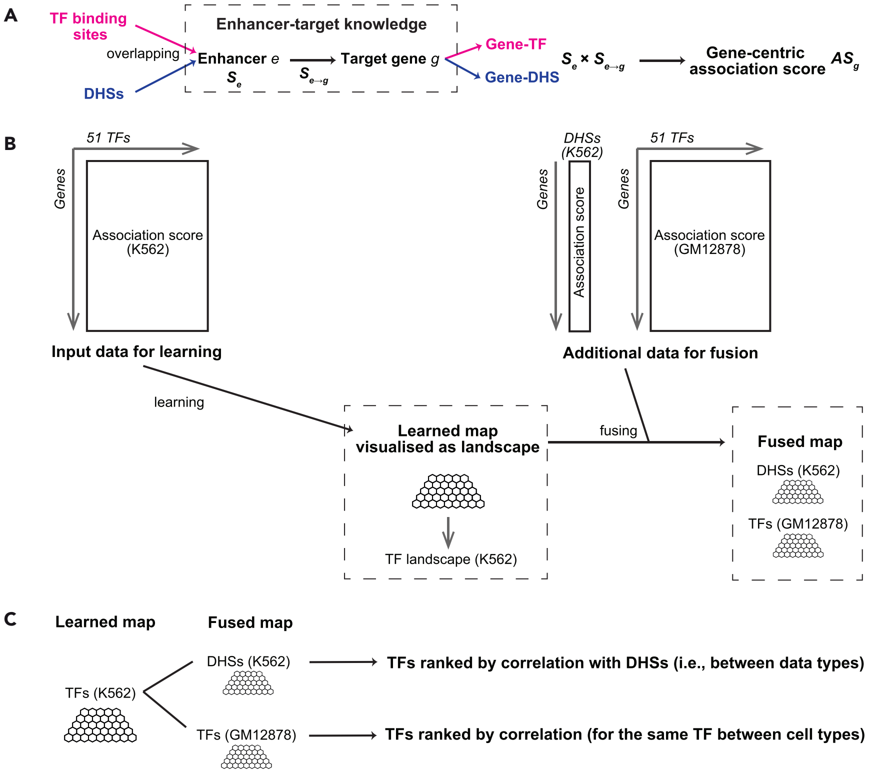


Figure 2. Overview of How to Correlate TFs and/or DHSs

(A) Association score for genes calculated from an input list of TF binding sites (or DHSs) by utilizing unified enhancer-target knowledge.

(B) The learning and fusing of regulatory genomic data.

(C) Correlation involving multilayer data types (TFs and DHSs) in the same cell type and/or involving two cell types (K562 and GM12878).

Comparing Regulatory Genomics Involving Multiple Data Types and/or between Different Cell Types

In this section, we describe a typical procedure used in ATLAS to carry out integrated tasks analyzing regulatory genomic data (Figure 2), including (1) how to select a map *tailored to* and *learned by* the input data and (2) how to use the learned map (visualized as a landscape) as a scaffold/reference for correlating with additional data. We mainly focus on two leukemia cell types: K562 (a chronic myeloid leukemia cell line) and GM12878 (a lymphoblastoid cell line); a total of 51 TFs common to both cell types were assayed (The ENCODE Project Consortium, 2012). Together with TF binding data, we also include DHS data in K562 (Thurman et al., 2012) for comparisons involving multiple data types produced in the same cell type. Taking into account enhancer-target knowledge (Fishilevich et al., 2017) (Figure 2A), we first calculated gene-centric association scores from genome-wide TF binding sites or DHSs, with higher scores implying higher chance of a gene (in rows) being targeted by a TF or a DHS (in columns), and then used these gene-centric scores for the learning and fusion (Figure 2B; Transparent Methods). The correlation was estimated based on the learned and fused map (Figure 2C).

A Ladder-Shaped Map Models TF Targeting in K562 and Defines Inter-TF Taxonomy

Based on the calculated gene-centric association scores for 51 TFs in K562 (Table S1), we observed a ladder-like distribution of target genes (Figure S1) and hence chose a ladder-shaped map (Figure 1H) for the learning. We visualized the learned map as the TF landscape (Figure 3A) and built a TF tree using a neighbor-joining algorithm (Paradis et al., 2004) (Figure S2). The basis of inter-TF taxonomy becomes much clearer when coupled with the TF landscape visualization, indicative of co-occupancy on putative target genes. For example, binding/targeting profiles of seven TFs (YY1, ELF1, EGR1, MAZ, POLR2A, MYC, and MAX) were similar both in the number and strength of target genes, being grouped together in the tree; they are all essential for the basic transcriptional regulation in K562.

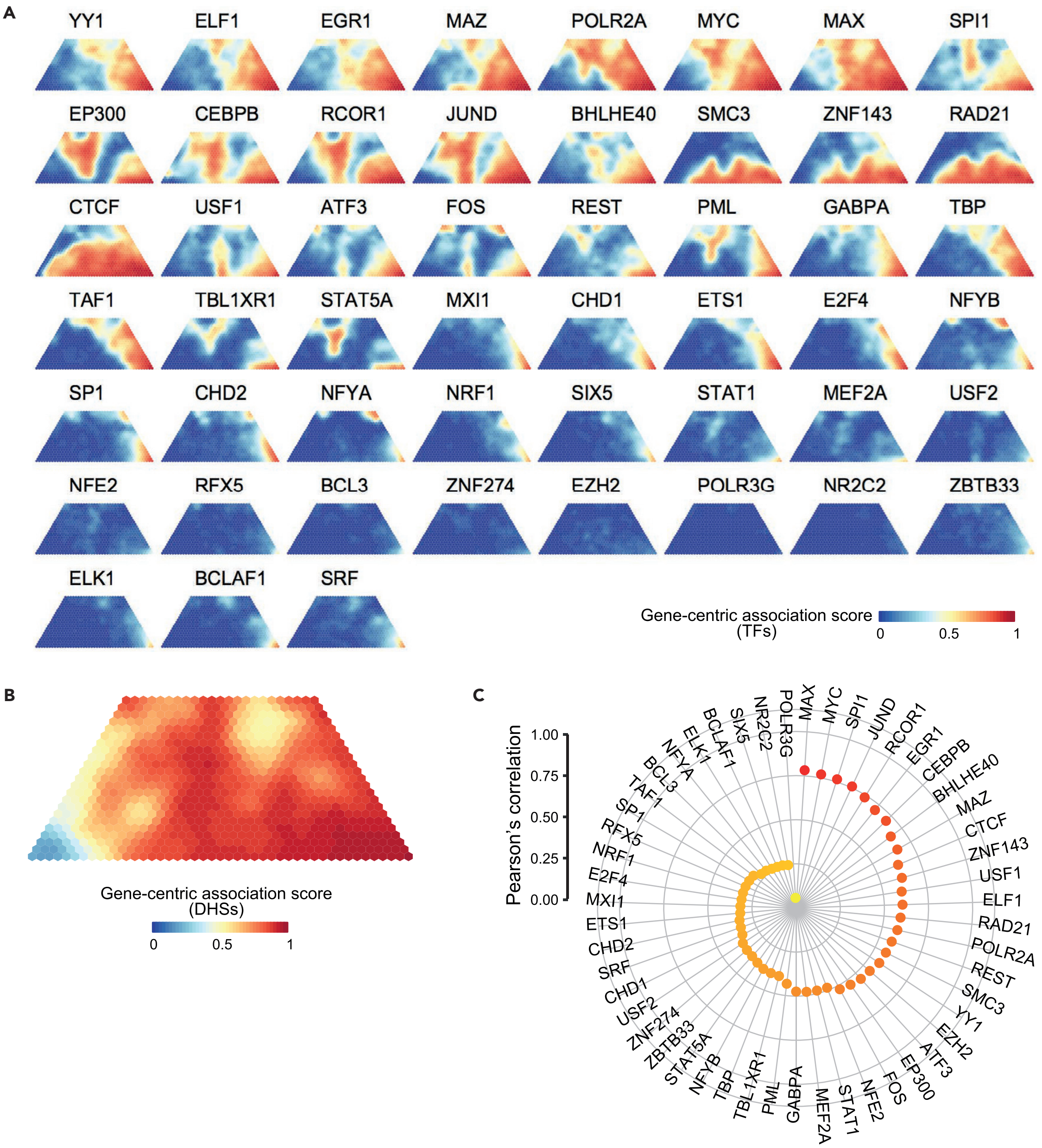


Figure 3. Correlating TFs and/or DHSs in K562

(A) TF landscape in K562. The landscape visualized based on a ladder-shaped map learned from gene-centric TF association scores. The color bar represents gene association scores, with the red for the highest score and the green for the lowest.

(B) The DHS map produced by fusing the DHS data onto the learned TF map.

(C) The polar plot of TFs ranked by correlation to DHS.

See also Figures S1 and S2, and Tables S1 and S2.

K562 TF Map Is Fused with Additional DHS Targeting Data for Multilayer Data Comparison in the Same Cell Type

Next, we calculated the DHS-derived gene-centric association scores (Figure 3A and Table S2) and fused DHS targeting data onto the learned map, producing a fused DHS map (Figure 3B). Based on the degree of similarity in target gene profiles for DHSs, TFs were ranked with the top two, MAX and MYC (Pearson's correlation >0.75 in Figure 3C). Both factors form the dimers as a transcriptional activator binding to the E box; how they select target genes largely depends on the chromatin context (Sabò and Amati,

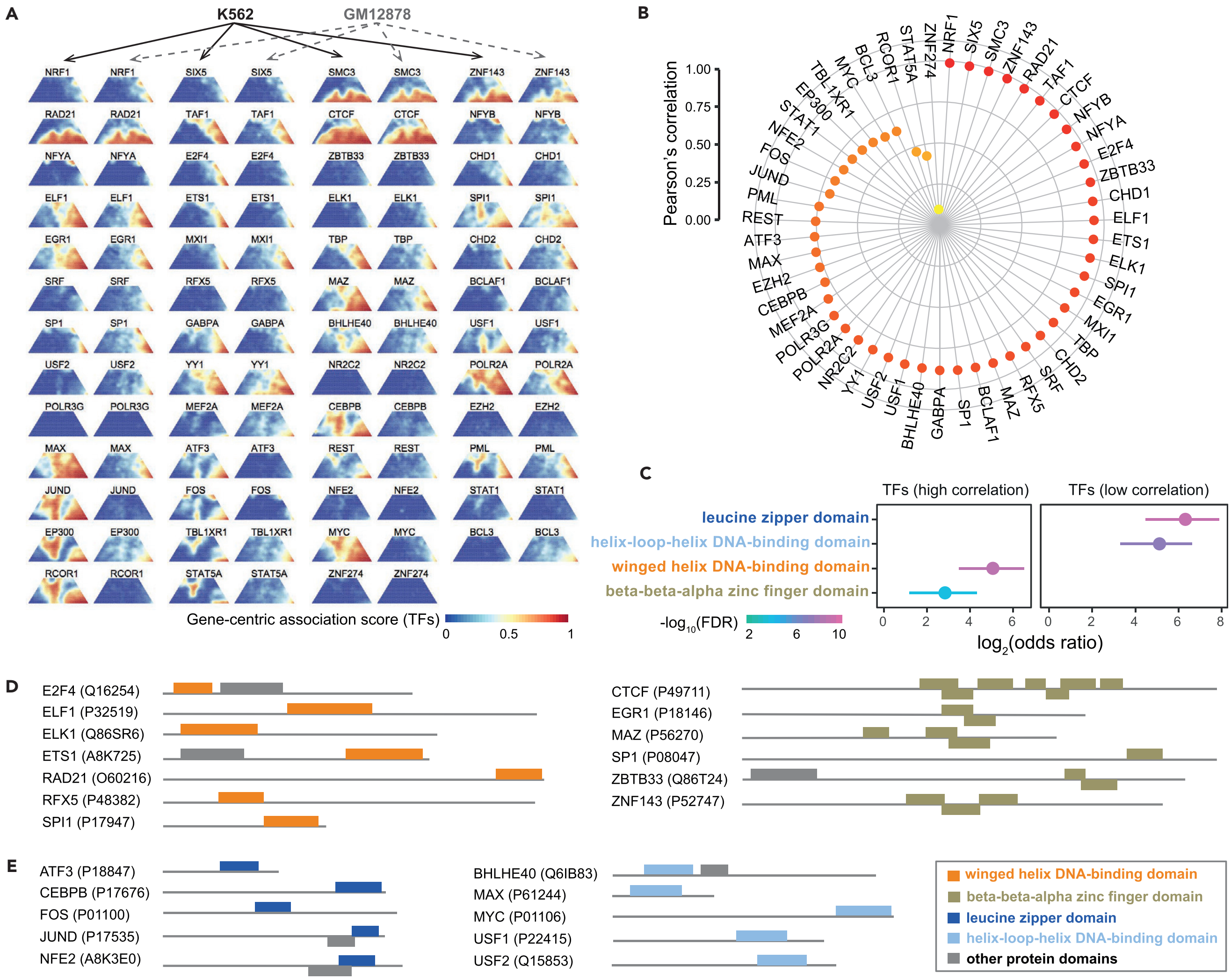


Figure 4. Correlating TFs between K562 and GM12878

(A) Side-by-side comparisons of TFs. Fusion of the learned K562 TF map with the TF data in GM12878 produces the GM12878 TF map.

(B) TFs ranked by correlation between K562 and GM12878; illustrated in the polar correlation plot of TFs.

(C) Protein domains enriched in top-ranked TFs and least-ranked TFs. Odds ratio (and 95% confidence interval) based on Fisher's exact test.

(D and E) Protein domain architectures for TFs. Domain architecture for a TF is based on the longest protein sequence (represented by UniProt ID). Highly correlated TFs (D) and lowly correlated TFs (E).

See also Table S3.

2014). The other top TFs are SPI1 (also known as PU.1), which acts as a chromatin accessibility factor (Marecki et al., 2004), and JUND, a functional component of the AP1 complex that has been shown to potentiate chromatin accessibility (Biddie et al., 2011). Multilayer data comparison revealed relationships between the transcription factor binding events and chromatin accessibility in terms of targeting potential.

K562 TF Map Is Used for Comparison Involving Two Cell Types

A more interesting case of comparison involves two cell types having the same data type. For this, we fused the K562 TF learned map with the TF binding/targeting data in GM12878 (that is, gene-centric association scores for TFs in GM12878; Table S3), producing a fused TF map in GM12878. The per-TF map comparisons between the two are illustrated in Figure 4A, with correlations shown in Figure 4B. Given a wide range of correlations observed across TFs, we hypothesized that the differences in the structural characteristics of TFs might explain the cellular basis of differential binding/targeting events. To test this hypothesis, we used the dnet package (Fang and Gough, 2014b) to perform enrichment analysis for 25 highly correlated

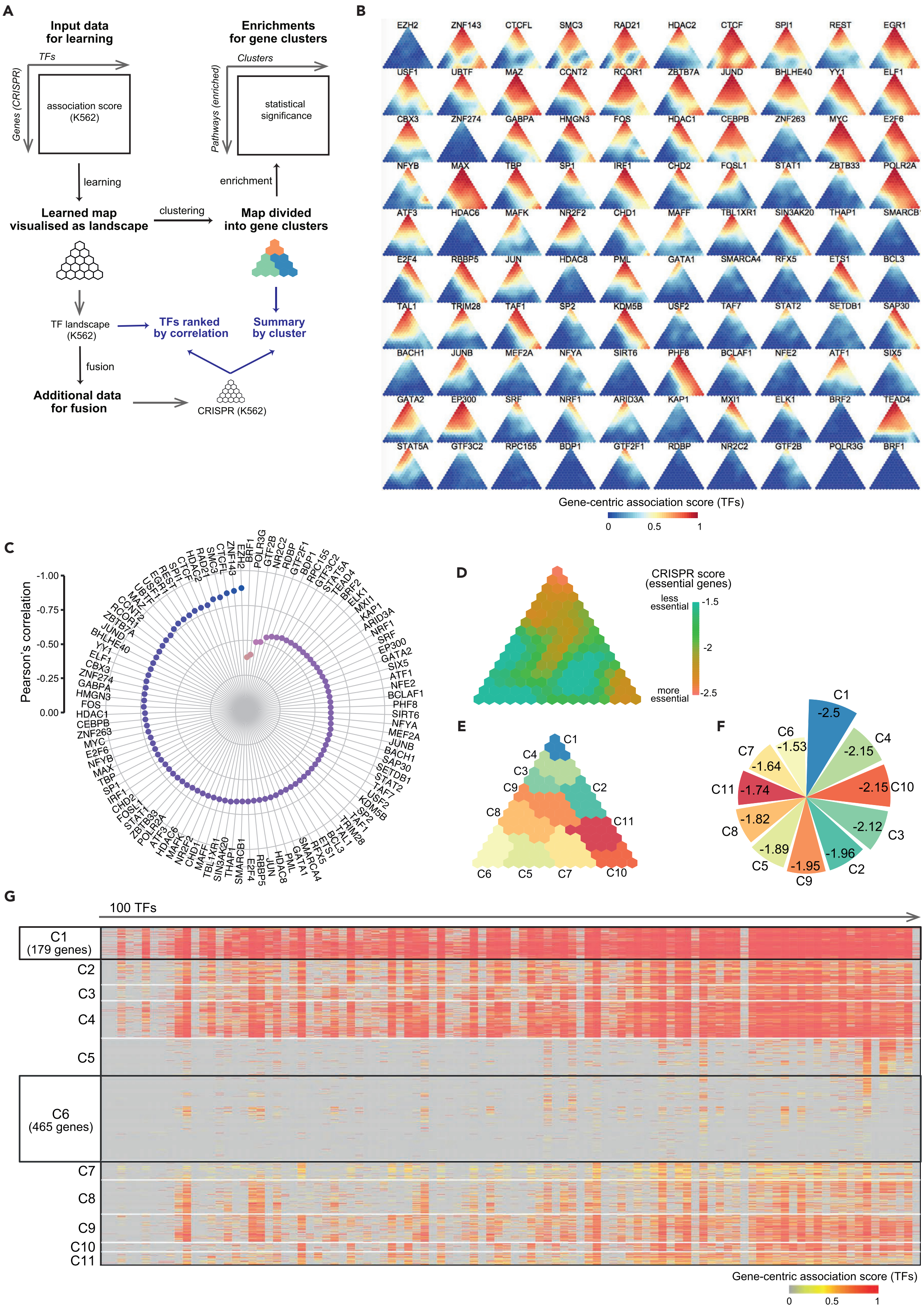

A
Input data
for learning
TFs
Genes (CRISPR)
association score
(K562)
learning
Learned map
visualised as landscape
clustering
Map divided
into gene clusters
Enrichments
for gene clusters
Clusters
Pathways (enriched)
statistical
significance
enrichment
TF landscape
(K562)
TFs ranked
by correlation
Summary
by cluster
fusion
Additional data
for fusion
CRISPR (K562)
B
EZH2 ZNF143 CTCFL SMC3 RAD21 HDAC2 CTCF SPI1 REST EGR1
USF1 UBTF MAZ CCNT2 RCOR1 ZBTB7A JUND BHLHE40 YY1 ELF1
CBX3 ZNF274 GABPA HMGN3 FOS HDAC1 CEBPB ZNF263 MYC E2F6
NFYB MAX TBP SP1 IRF1 CHD2 FOSL1 STAT1 ZBTB33 POLR2A
ATF3 HDAC6 MAFK NR2F2 CHD1 MAFF TBL1XR1 SIN3AK20 THAP1 SMARCB1
E2F4 RBBP5 JUN HDAC8 PML GATA1 SMARCA4 RFX5 ETS1 BCL3
TAL1 TRIM28 TAF1 SP2 KDM5B USF2 TAF7 STAT2 SETDB1 SAP30
BACH1 JUNB MEF2A NFYA SIRT6 PHF8 BCLAF1 NFE2 ATF1 SIX5
GATA2 EP300 SRF NRF1 ARID3A KAP1 MXI1 ELK1 BRF2 TEAD4
STAT5A GTF3C2 RPC155 BDP1 GTF2F1 RDBP NR2C2 GTF2B POLR3G BRF1
Gene-centric association score (TFs)
0 0.5 1
C
Pearson's correlation
-1.00
-0.75
-0.50
-0.25
0.00
D
CRISPR score
(essential genes)
less
essential
more
essential
-1.5
-2
-2.5
E
C1 C4 C3 C9 C2 C8 C11 C6 C5 C7 C10
F
C1 -2.5
C4 -2.15
C10 -2.15
C3 -2.12
C2 -1.96
C9 -1.95
C5 -1.89
C8 -1.82
C11 -1.74
C7 -1.64
C6 -1.53
G
100 TFs
C1
(179 genes)
C2
C3
C4
C5
C6
(465 genes)
C7
C8
C9
C10
C11
Gene-centric association score (TFs)
0 0.5 1

Figure 5. Linking TFs to CRISPR in K562

(A) Schematic overview for correlating TF binding/targeting with CRISPR screens in K562. Calculating gene-centric association scores per TF is illustrated in Figure 2A. Also illustrated are identification and enrichment analysis of gene clusters.

(B) TF landscape visualized based on a triangle-shaped map learned from gene-centric TF association scores in K562.

(C) The polar correlation plot of TFs ranked by correlation to CRISPR.

(D) CRISPR map produced by fusing the CRISPR data onto the learned TF map.

(E) Gene clusters identified from the learned TF map. Clusters are color coded and labeled.

(F) Summary of CRISPR gene essential scores per cluster; valued is the per-cluster average.

(G) Heatmap of gene clusters.

See also Figure S4 and Table S4.

TFs versus 25 lowly correlated ones using SCOP structural domains (de Lima Morais et al., 2011), revealing the distinctive protein domain compositions (Figure 4C). TFs with high correlation tend to contain winged helix DNA binding domains (false discover rate [FDR] = 1.5×10^{-9}; E2F4, ELF1, ELK1, ETS1, RAD21, RFX5, and SPI1) and beta-beta-alpha zinc finger domains (FDR = 1.2×10^{-4}; CTCF, EGR1, MAZ, SP1, ZBTB33, and ZNF143), with their domain architectures (obtained via the dcGO Predictor Batch Query [Fang and Gough, 2013]) illustrated in Figure 4D. By contrast, TFs with low correlation are unique in leucine zipper domains (FDR = 6.7×10^{-10}; ATF3, CEBPB, FOS, JUND, and NFE2) and helix-loop-helix DNA binding domains (FDR = 3.9×10^{-8}; BHLHE40, MAX, MYC, USF1, and USF2) (Figures 4C and 4E).

Linking TFs to CRISPR, Both in K562

In this section, we consider all 100 TFs available in K562 (The ENCODE Project Consortium, 2012) and demonstrate that linking TFs to CRISPR reveals a greater number TF binding/targeting events for essential genes. The CRISPR-based screen identified genes required for survival in K562 (Wang et al., 2015). The CRISPR score measures the essential genes; the lower the score, the higher the fitness cost imposed by gene inactivation (that is, the more essential genes are). Following the process outline in Figure 5A, we first produced the binding/targeting landscape of TFs (Figure 5B) and ranked them according to correlation with the fused CRISPR map (Figures 5C and 5D). Next, we implemented a region-growing algorithm to partition the learned map into gene clusters, the number of gene clusters determined based on the distance matrix of the map nodes and each cluster covering continuous regions (Transparent Methods). In doing so, we obtained 11 gene clusters (C1-C11; Figure 5E and Table S4). For each cluster, we subsequently summarized the CRISPR gene essential scores (Figure 5F). We observed that essential genes with the lowest CRISPR scores (C1) have greater number of TF binding events (Figure 5G) and that the TF binding events are rarely seen for genes with the highest CRISPR scores (C6; Figure 5G). Through enrichment analysis using Reactome pathways (Fabregat et al., 2018), we also found enrichment of essential biological processes/pathways, such as translation and nonsense-mediated decay in C1, and no enrichment in C6 (Figure S4).

DISCUSSION

When Cubism Meets Genomics

The timing is right in the era of doing genomic data science. A demanding issue in doing big genomic data science is how to identify and compare the occurrence of the patterns from multidimensional and multiparametric data. These data come in a heterogeneous form, making it difficult to integrate information of different types and sources. We address this issue following the Picasso's cubism philosophy. We devise a suite of maps that are able to capture a wide range of data shapes under a single framework. We suggest the mapping of the input data onto a 2D hyperspace oriented along the first two axes, for example, identified by PCA, and the resulting data point cloud directs the choice of the map shape (as illustrated in Figures S1 and S3). We recognize the possibility of such empirical observation resulting in no match with any shape supported currently (shown in Figures 1A–1I); in this situation, it is advisable to consider the supra-hexagonal map owing to its perfect symmetry. In addition to the shape, the maps also have AI allowing for data modeling in a self-organizing manner.

Challenges and Opportunities of Regulatory Genomics

In the human body, connections between gene promoters and regulatory elements control cell-type specificity. This context-specific control requires mapping of such connections for every cell type. Experimentally, it poses great challenges to data generators. When compared with identifying chromatin interactions (Javierre et al., 2016), technologies make it much easier to generate (have generated) regulatory elements such as TF binding sites and DHSs for most cell types. Computationally, it is achievable by integrating available context-specific regulatory elements and their target genes into a less context-specific scaffold; one such effort is to assimilate knowledge of enhancer-target connections (Fishilevich et al., 2017). Building on such computational opportunities and also as a proof of principle, we have utilized the enhancer-target knowledge combined with context-specific regulatory genomic data to estimate gene-centric targeting profiles of TFs and DHSs. The self-organized modeling and representation of gene targeting profiles between TFs and/or DHSs make it straightforward for effective comparisons.

Analytical Advantages of ATLAS

ATLAS is the first realization of analytical cubism in regulatory genomics. More importantly, it enables comparisons, both visually intuitive and scientifically insightful. We demonstrate the use of ATLAS to compare regulatory genomic data of different types and involving different cell types. In the illustrated use cases, we show that modeling TF binding/targeting within a cell type defines inter-TF taxonomy; comparing multilayer regulatory data in the same cell type reveals the targeting relationships between the TF binding events and chromatin accessibility; and linking TFs to CRISPR identifies many TF binding events for essential genes. All these are achieved in a transparent way, rather than in the black box. The added value of the landscape-guided correlation is that such correlation can be intuitively visualized, for example, the correlation shown in Figures 4A and 4B. This value is useful particularly at the exploratory stage of multilayer genomic data analysis. Furthermore, reproducible use cases with 2D visuals enable effective data-driven and visual-aided exploratory analysis.

Future Directions of ATLAS

Other than the analytical powers currently supported by ATLAS, we plan to build a resource connecting regulatory elements to putative target genes, under contexts of different levels (from the generic to the system-/organ-specific and to the lineage-specific level). Future efforts will also focus on a user-friendly web interface targeting users who are less familiar with the R environment. We anticipate that the self-organized representation and comparison offered by ATLAS will be of great use at the exploratory stage of regulatory genomic data analysis. We also anticipate that ATLAS, freely available to and reproducible by the scientific community, will aid in the use of big data produced from genomics consortia to address the big questions.

METHODS

All methods can be found in the accompanying Transparent Methods supplemental file.

SUPPLEMENTAL INFORMATION

Supplemental Information includes Transparent Methods, four figures, and four tables and can be found with this article online at https://doi.org/10.1016/j.isci.2018.04.017.

ACKNOWLEDGMENTS

This work was funded by the National Natural Science Foundation Grants of China (31301041 to H.F., 81530003 and 81770153 to K.W.) and by the National Key Research and Development Program of China (2016YFC0902800).

AUTHOR CONTRIBUTIONS

H.F. and K.W. designed the study. H.F. developed the method and the software. H.F. drafted the paper. K.W. reviewed and edited the paper. H.F. and K.W. obtained the funding.

DECLARATION OF INTERESTS

The authors declare no competing interests.

Received: January 26, 2018
Revised: March 28, 2018
Accepted: April 20, 2018
Published: May 25, 2018

REFERENCES

Allman, E.L., Painter, H.J., Samra, J., and Carrasquilla, M. (2016). Metabolomic profiling of the malaria box reveals antimalarial target. Antimicrob. Agents Chemother. *60*, 6635–6649.

Biddie, S.C., John, S., Sabo, P.J., Thurman, R.E., Johnson, T.A., Schiltz, R.L., Miranda, T.B., Sung, M.H., Trump, S., Lightman, S.L., et al. (2011). Transcription factor AP1 potentiates chromatin accessibility and glucocorticoid receptor binding. Mol. Cell *43*, 145–155.

de Lima Morais, D.A., Fang, H., Rackham, O.J., Wilson, D., Pethica, R., Chothia, C., and Gough, J. (2011). SUPERFAMILY 1.75 including a domain-centric gene ontology method. Nucleic Acids Res. *39*, D427–D434.

Fabregat, A., Jupe, S., Matthews, L., Sidiropoulos, K., Gillespie, M., Garapati, P., Haw, R., Jassal, B., Korninger, F., May, B., et al. (2018). The Reactome pathway Knowledgebase. Nucleic Acids Res. *46*, D649–D655.

Fang, H., and Gough, J. (2013). dcGO: database of domain-centric ontologies on functions, phenotypes, diseases and more. Nucleic Acids Res. *41*, D536–D544.

Fang, H., and Gough, J. (2014a). supraHex: an R/Bioconductor package for tabular omics data analysis using a supra-hexagonal map. Biochem. Biophys. Res. Commun. *443*, 285–289.

Fang, H., and Gough, J. (2014b). The 'dnet' approach promotes emerging research on cancer patient survival. Genome Med. *6*, 64.

Fishilevich, S., Nudel, R., Rappaport, N., Hadar, R., Plaschkes, I., Stein, T.I., Rosen, N., Kohn, A., Twik, M., Safran, M., et al. (2017). GeneHancer: genome-wide integration of enhancers and target genes in GeneCards. Database (Oxford) *2017*, 1–17.

Javierre, B.M., Burren, O.S., Wilder, S.P., Kreuzhuber, R., Hill, S.M., Sewitz, S., Cairns, J., Wingett, S.W., Várnai, C., Thiecke, M.J., et al. (2016). Lineage-specific genome architecture links enhancers and non-coding disease variants to target gene promoters. Cell *167*, 1369–1384.e19.

Jiang, L., Hindmarch, C.C.T., Rogers, M., Campbell, C., Waterfall, C., Coghill, J., Mathieson, P.W., and Welsh, G.I. (2016). RNA sequencing analysis of human podocytes reveals glucocorticoid regulated gene networks targeting non-immune pathways. Sci. Rep. *6*, 35671.

Marecki, S., McCarthy, K.M., and Nikolajczyk, B.S. (2004). PU.1 as a chromatin accessibility factor for immunoglobulin genes. Mol. Immunol. *40*, 723–731.

Marx, V. (2013). Biology: the big challenges of big data. Nature *498*, 255–260.

Paradis, E., Claude, J., and Strimmer, K. (2004). APE: analyses of phylogenetics and evolution in R language. Bioinformatics *20*, 289–290.

Sabò, A., and Amati, B. (2014). Genome recognition by MYC. Cold Spring Harb. Perspect. Med. *4*, 1–14.

The ENCODE Project Consortium (2012). An integrated encyclopedia of DNA elements in the human genome. Nature *489*, 57–74.

Thurman, R.E., Rynes, E., Humbert, R., Vierstra, J., Maurano, M.T., Haugen, E., Sheffield, N.C., Stergachis, A.B., Wang, H., Vernot, B., et al. (2012). The accessible chromatin landscape of the human genome. Nature *489*, 75–82.

Wang, T., Birsoy, K., Hughes, N.W., Krupczak, M., Post, Y., Wei, J.J., Eric, S., and Sabatini, D.M. (2015). Identification and characterization of essential genes in the human genome. Science *350*, 1096–1101.

Zhang, J., and Fang, H. (2012). Using self-organizing maps to visualize, filter and cluster multidimensional bio-omics data. In Developments and Applications of Self-organizing Maps, M. Johnsson, ed. (InTech.). https://doi.org/10.5772/51702.

ISCI, Volume *3*

Supplemental Information

Regulatory Genomic Data Cubism

Hai Fang and Kankan Wang

Supplemental Figures

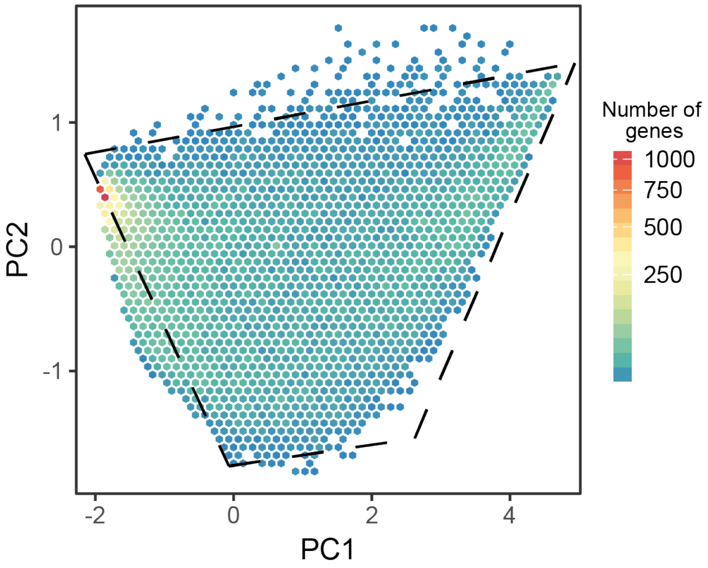


Figure S1. Principle component analysis of gene-centric TF association scores in K562, Related to Figure 3.

Dots along the first two component axes are genes, collectively forming a ladder-like shape.

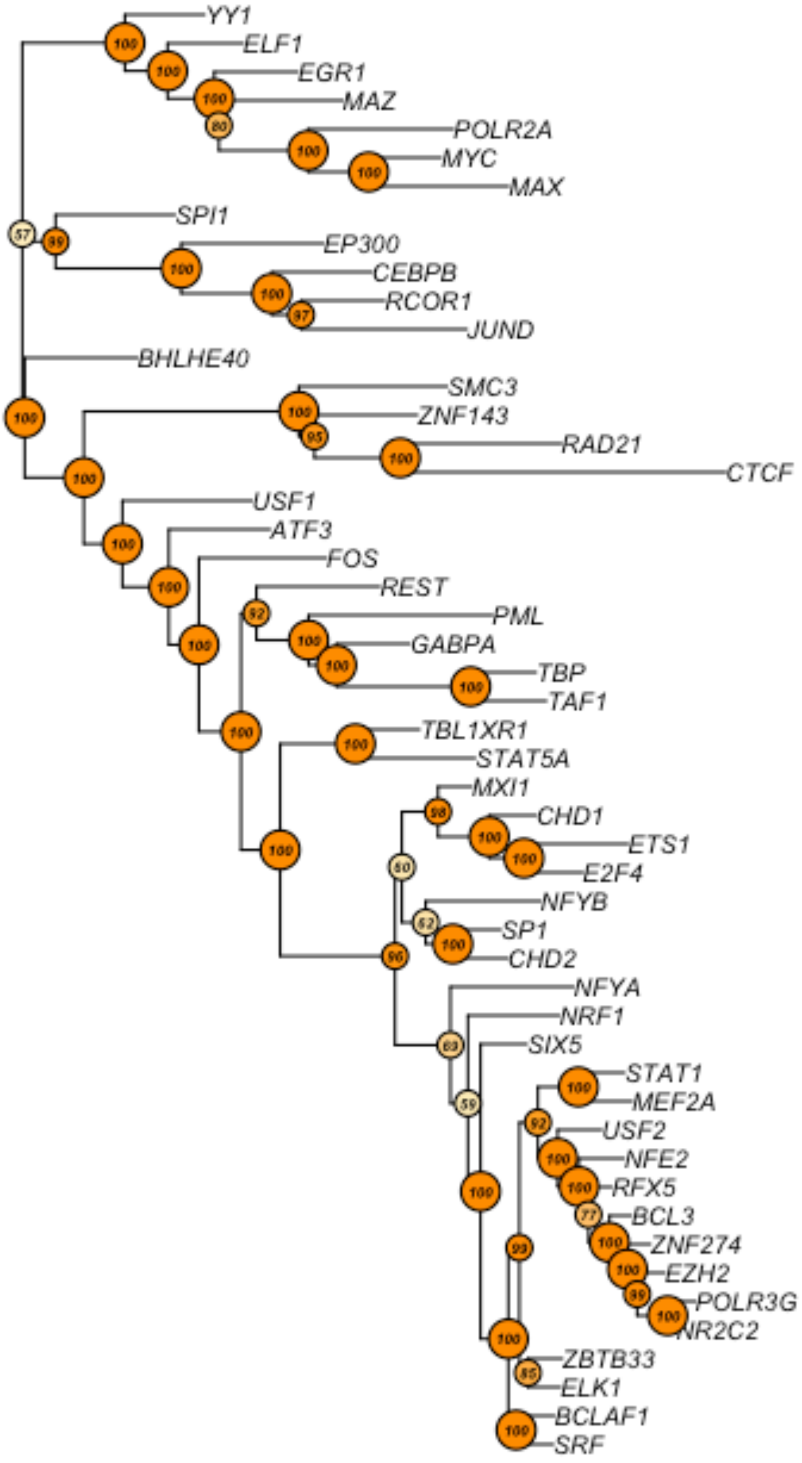


Figure S2. A neighbor-joining tree built based on inter-TF pairwise distance, Related to Figure 3.

The number in the circle is the bootstrap value (i.e. the confidence value in support for the tree branching).

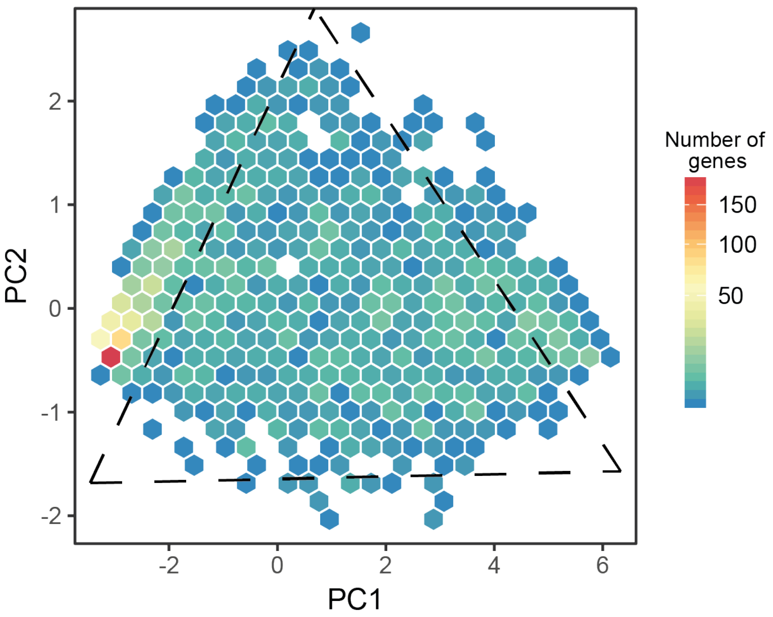


Figure S3. Principle component analysis of TF association scores for CRISPR genes in K562, Related to Figure 5.

Dots along the first two component axes are genes, collectively forming a triangle-like shape.

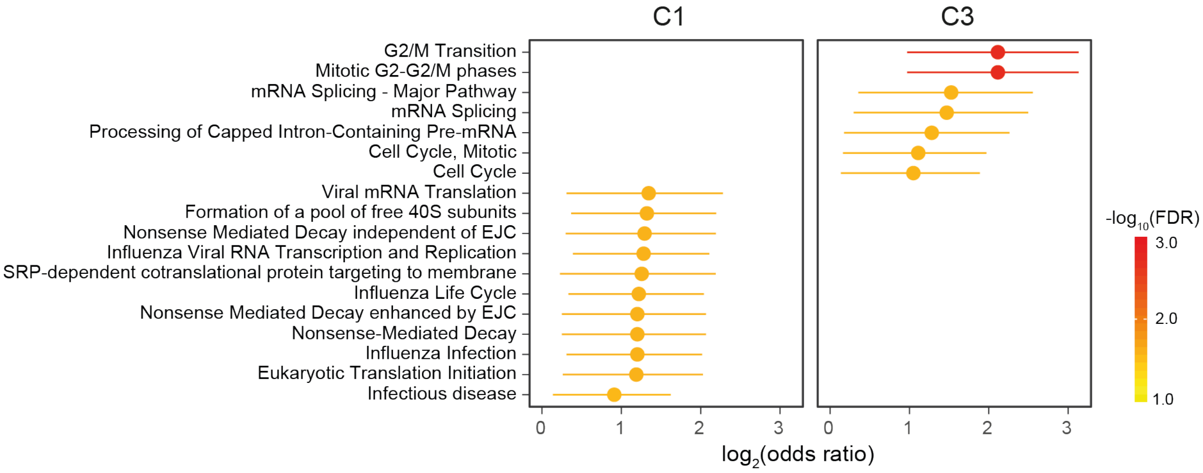


Figure S4. Reactome pathways enriched in gene clusters, Related to Figure 5.

Odds ratio (and 95% confidence interval) based on Fisher’s exact test.

Supplemental Tables

Table S1. Gene-centric association scores for TFs in K562, Related to Figure 3.

Table S2. Gene-centric association scores for DHSs in K562, Related to Figure 3.

Table S3. Gene-centric association scores for TFs in GM12878, Related to Figure 4.

Table S4. Gene clusters identified from the learned TF map together with CRISPR scores in K562, Related to Figure 5.

Transparent Methods

Implementation of cubism via the self-organising learning algorithm

We implemented a self-organising learning algorithm enabling a range of maps (Figures 1A-1I) learned from input data. The learned map was associated with the codebook matrix. The choice of the map shape is empirically informed by looking at the structure of input data, depicted as points on the 2D hyperspace orientated along the first two principle axes of the data distribution. For example, principle component analysis (PCA) of gene-TF association scores in K562 revealed a ladder-like shape of the data distribution (Figure S1); based on this observation, a ladder-shaped map was selected to model this TF targeting dataset. In addition to the shape, the number of hexagons was empirically determined, approximated by this heuristic equation:

$$N \geq C \times \sqrt{n}, \text{ (Equation 1)}$$

where n is the number of input training data vectors (here the number of genes), C for the constant value controlling the map size (5 for the big map, 3 for the intermediate map, and 1 for the small map), and N for the number of hexagons. The actual value of N is automatically fine-tuned so that hexagons form exactly the map grid as indicated (e.g. the ladder-shaped map).

In the map, a node i is represented by two types of vectors: the location vector on the 2-dimensional map grid ($\vec{r}_i \in \Re^2$), and the codebook vector in the K-dimensional hyperspace ($\vec{m}_i \in \Re^K$). The learning consists of two steps.

In the first step, the winner map node w is chosen where its codebook vector is closest to the input training vector $\vec{x}$:

$$\|\vec{x} - \vec{m}_w\| = \min_i \left\{ \|\vec{x} - \vec{m}_i\| \right\}, i = 1, \cdots, N, \text{ (Equation 2)}$$

where $\vec{x}$ is an input training vector of the gene, $\vec{m}_i$ for a codebook vector of node *i*, and $\vec{m}_w$ for the codebook vector of the winner node *w*.

In the second step, the winner node *w* and its neighbors are updated by moving towards the input training vector $\vec{x}$:

$$\vec{m}_i(t+1) = \vec{m}_i(t) + \alpha(t)h_{wi}(t)\left[\vec{x}(t) - \vec{m}_i(t)\right], \text{(Equation 3)}$$

$$h_{wi}(t) = \exp\left(-\frac{\left\|\vec{r}_w - \vec{r}_i\right\|^2}{2\sigma^2(t)}\right), \text{(Equation 4)}$$

where $\alpha(t)$ is the learning rate at training time *t*, $h_{wi}(t)$ for a Gaussian kernel function centered on the winner node *w*, $\sigma(t)$ for the width of the kernel at training time *t*, and $\vec{r}_w$ and $\vec{r}_i$ respectively for the location vectors of map node *w* and *i*.

Visualisation of and gene cluster identification from the learned map

The codebook matrix associated with the learned map was used for visualisation as the landscape. For example, the learned TF map was visualised providing a TF-specific view of the binding/targeting profile, collectively forming TF landscape. We identified gene clusters from the learned map in which nodes of similar patterns are configured together. We used a region-growing algorithm to partition the learned map into gene clusters, each of which is continuous over the map. In brief, this algorithm first identifies a set of seed nodes, each of which has local minima of distance matrix (between map nodes):

$$f(\vec{m}_i, S_i) \leq f(\vec{m}_j, S_j), \forall j \in S_i, \text{ (Equation 5)}$$

$$f(\vec{m}_i, S_i) = median\left\{\left\|\vec{m}_i - \vec{m}_k\right\|, k \in S_i\right\}, \text{(Equation 6)}$$

where $\vec{m}$ is the code vector, S_i and S_j respectively for the sets of neighboring map nodes *i* and *j* have, and $f(\vec{m}_i, S_i)$ for the median distance between the map node *i* and its neighboring map nodes S_i. Then, for each seed node, assign a gene cluster; it starts with seed nodes and competes for unassigned neighboring nodes (with the shortest distance) iteratively until all the nodes are assigned.

Fusion of additional data into the learned map

The fusion of additional (non-training) data onto a learned map enables correlation between the input data (used for map learning) and the additional data (used for map fusion). The map fused with an additional data produces a fused map, which is associated with a fused codebook matrix capturing inherent relationships of this additional data with the input training data (as a reference). As before, this fused codebook matrix can be used for landscape visualisation. The similarity between input data and additional data was calculated based on the codebook matrix (associated with the learned/fused map), measured as Pearson's correlation coefficient. The per-cluster summary was calculated from the fused map.

Analysis using multilayer regulatory genomic data

We obtained genome-wide binding information on peaks per TF from ENCODE (The ENCODE Project Consortium, 2012). In K562, a total of 100 TFs were assayed, amongst of which 51 TFs were also assayed in GM12878. We used TF binding data in K562 as input data for the learning (i.e. as a reference), upon which TF binding data in GM12878 were fused for comparison. For comparisons involving data of different types produced in the same cell line, we also obtained genome-wide DHSs in K562 from ENCODE (Thurman et al., 2012), used as additional data for fusion.

Gene-centric association scoring from TF binding sites or DHSs

We obtained genome-wide integration of enhancers and their target genes from GeneCards (Fishilevich et al., 2017). Each enhancer has a confidence score S_e, and each enhancer-target connection has a link score $S_{e \to g}$ quantifying the strength linking an enhancer to a target gene. Such knowledge is less context-specific, representing unified links between enhancers and target genes. We utilised the unified enhancer-target link knowledge to estimate and score target genes from an input list of TF binding sites or DHSs (context-specific). Given genome-wide sites bound by a TF (or genome-wide DHSs), an association score AS_g for a target gene g was formulated as follows:

$$x_g = \max_{e \in \Omega} \left[S_g \times S_{e \to g} \right], \text{(Equation 7)}$$

$$AS_g = eCDF(x_g), \text{(Equation 8)}$$

where Ω stands for collections of enhancers (bound by the TF or overlapped with DHSs) the gene g links to, *max* denotes maximum scoring scheme to give a conservative estimate, and *eCDF* is empirical cumulative distribution function used to scale the raw score x_g into AS_g(ranging 0 and 1).

Analysis using datasets from CRISPR

We obtained 2,181 genes (FDR <0.05) required for survival in K562 according to the CRISPR-based screen (Wang et al., 2015). These genes together with CRISPR scores were compared with TF target genes in K562 (calculated using the equations 7 and 8 above), that is, TF gene-centric association scores used for map learning and CRISPR scores for map fusion.

Construction of a TF tree

For the TF map in K562, the codebook matrix was used to construct a TF tree using a neighbour-joining algorithm (Paradis et al., 2004) and using the distance metric (calculated as $1-\rho$, where ρ is Pearson's correlation coefficient).

Extraction of TF domain architectures

Domain architectures for TFs were extracted via the dcGO Predictor Batch Query using UniProt ID (the longest protein per TF) (Fang and Gough, 2013).

Enrichment analysis

Enrichment analysis was performed using the dnet package (Fang and Gough, 2014) to identify enriched SCOP structural domains (version 1.75) (de Lima Morais et al., 2011) and enriched Reactome pathways (version 63) (Fabregat et al., 2018).

Code availability

All codes are accessible at http://suprahex.r-forge.r-project.org/ATLAS.html, in which all showcases described above are reproducible following step-by-step instructions. At the time of writing (March 2018), it requires R (version 3.4.3 or higher), and the packages supraHex (version 1.13.3 or higher) and GenomicRanges (1.30.0 or higher).

Supplemental references

Fabregat, A., Jupe, S., Matthews, L., Sidiropoulos, K., Gillespie, M., Garapati, P., Haw, R., Jassal, B., Korninger, F., May, B., et al. (2018). The Reactome Pathway Knowledgebase. Nucleic Acids Res. *46*, D649–D655.

Fang, H., and Gough, J. (2013). dcGO: database of domain-centric ontologies on functions, phenotypes, diseases and more. Nucleic Acids Res. *41*, D536-44.

Fang, H., and Gough, J. (2014). The ‘dnet’ approach promotes emerging research on cancer patient survival. Genome Med. *6*, 64.

Fishilevich, S., Nudel, R., Rappaport, N., Hadar, R., Plaschkes, I., Stein, T.I., Rosen, N., Kohn, A., Twik, M., Safran, M., et al. (2017). GeneHancer : genome-wide integration of enhancers and target genes in GeneCards. Database *2017*, 1–17.

de Lima Morais, D.A., Fang, H., Rackham, O.J., Wilson, D., Pethica, R., Chothia, C., and Gough, J. (2011). SUPERFAMILY 1.75 including a domain-centric gene ontology method. Nucleic Acids Res *39*, D427-34.

Paradis, E., Claude, J., and Strimmer, K. (2004). APE: analyses of phylogenetics and evolution in R language. Bioinformatics *20*, 289–290.

The ENCODE Project Consortium (2012). An integrated encyclopedia of DNA elements in the human genome. Nature *489*, 57–74.

Thurman, R.E., Rynes, E., Humbert, R., Vierstra, J., Maurano, M.T., Haugen, E., Sheffield, N.C., Stergachis, A.B., Wang, H., Vernot, B., et al. (2012). The accessible chromatin landscape of the human genome. Nature *489*, 75–82.

Wang, T., Birsoy, K., Hughes, N.W., Krupczak, M., Post, Y., Wei, J.J., Eric, S., and Sabatini, D.M. (2015). Identification and characterization of essential genes in the human genome. Science (80-.). *350*, 1096–1101.