Shared regulation and functional relevance of local gene co-expression revealed by single cell analysis

Diogo M. Ribeiro^{1,2*}, Chaymae Ziyani^{1,2}, Olivier Delaneau^{1,2*}

¹Department of Computational Biology, University of Lausanne, Lausanne, Switzerland,

²Swiss Institute of Bioinformatics (SIB), Lausanne, Switzerland

*co-corresponding authors: diogo.ribeiro@unil.ch, olivier.delaneau@unil.ch

Abstract

Most human genes are co-expressed with a nearby gene. Yet, previous studies only reported this extensive local gene co-expression using bulk RNA-seq. Here, we leverage single cell datasets in >85 individuals to identify gene co-expression across cells, unbiased by cell type heterogeneity and benefiting from the co-occurrence of transcription events in single cells. We discover thousands of co-expressed genes in two cell types and (i) compare single cell to bulk RNA-seq in identifying local gene co-expression, (ii) show that many co-expressed genes – but not the majority – are composed of functionally-related genes and (iii) provide evidence that these genes are transcribed synchronously and their co-expression is maintained up to the protein level. Finally, we identify gene-enhancer associations using multimodal single cell data, which reveal that >95% of co-expressed gene pairs share regulatory elements. Our in-depth view of local gene co-expression and regulatory element co-activity advances our understanding of the shared regulatory architecture between genes.

Introduction

The expression of genes is regulated in space and time through the action of various cis-regulatory elements such as promoters, insulators and enhancers¹⁻³. These elements play an important role in buffering and fine-tuning gene expression in response to stress conditions, differentiation cues and cell states^{4,5}. To achieve a tight control and robust expression level, genes are usually regulated by multiple enhancers, eventually with redundant action⁶ as well as multiple target genes^{7,8}. Indeed, neighboring genes frequently exhibit similar behavior in terms of expression level⁹⁻¹¹. This local gene co-expression is more pronounced in the immediate vicinity of a gene (e.g. <100 kb) but can occur at longer distances and regardless of the transcriptional orientation or shared functionality^{12,13}. In particular, we have previously shown that as many as 59% genes are co-expressed with a nearby gene (within 1Mb) across 49 GTEx tissues⁹. However, these observations came from studies measuring bulk gene expression in tissues, which entails several limitations. In particular, as these measurements

are averages across many sampled cells, the correlation between nearby genes does not necessarily represent co-expression in the same cell. Moreover, tissues contain multiple cell types, which may mask the detection of cell-type specific co-expression^{14,15}.

Single cell analysis has multiple advantages over bulk analysis for addressing the molecular circuitry between gene (co-)expression and regulatory elements by (i) being able to reduce cell type heterogeneity, or even study one specific cell type, (ii) producing measurements per cell and thus getting closer in time to the transcription event and (iii) allow to detect co-expression events per individual (i.e. a single genetic background), and thus not affected by linkage disequilibrium (LD). Moreover, with the advent of multimodal single cell datasets^{16,17}, chromatin accessibility and gene expression levels can be measured in the same exact cells, which allows exploring the regulatory elements affecting gene expression at a very high resolution.

Here, we provide an in-depth view of local gene co-expression and regulatory element co-activity using single cell data in two cell lines (iPSC, LCLs). Namely, we (i) confirm the widespread local co-expression of thousands of gene pairs at the single cell level, (ii) compare single cell to bulk RNA-seq in identifying local gene co-expression, (iii) explore the synchronous transcription of co-expressed genes and their maintenance up to the protein level and (iv) identify enhancers involved in local gene co-expression by analysing single cell RNA-seq and ATAC-seq data performed on the same cells. Our study improves the understanding of the shared regulatory architecture between genes as well as their regulators, which is a prerequisite for understanding their implication in complex traits and disease.

Results

Local gene co-expression is widespread in single cells

To identify locally co-expressed gene pairs (COPs) using single cell data, we adapted a method developed previously⁹ to handle gene expression measurements across thousands of single cells (Fig. 1). Briefly, this method generates genome-wide maps of local gene co-expression from gene expression quantifications by looking at the correlation between genes across cells of the same cell line. For each gene, all other genes in a cis window of 1 Mb around the gene transcription start site (TSS) are tested for having higher than expected expression Pearson correlation. We control for a maximum false discovery rate (FDR) of 5% by comparing the observed correlation to expected correlation values under the null obtained by shuffling expression values 1000 times (see Methods). Only autosomal protein-coding genes were assessed and gene pairs with high cross-mappability, i.e. the extent to which reads from one gene are mapped to the other gene¹⁸, were excluded.

First, we applied this method to gene expression quantifications for undifferentiated induced pluripotent stem cells (iPSC) from 87 individuals from the HipSci consortium¹⁹ for which both single cell RNA-seq data (SmartSeq2, 7440 cells in total) and bulk RNA-seq data were available^{20,21}. As the single cell data was obtained in batches which comprised sets of 4 to 6 individuals, to account for potential batch effects, we performed COP identification on a per individual and per experiment basis, as recommended by the data providers²¹ (see Methods). Then, for each individual, we obtained the union of COPs stemming from the different experiments (Supplementary Fig. 1). We discovered between 4 and 442 COPs per individual (mean and standard deviation, 113.3 COPs $\pm$ 102), the number of COPs identified per individual being strongly correlated with the number of cells available for each individual (Fig. 2a, Spearman R = 0.88, p-value = $4.2e^{-29}$), similar to what was observed for the GTEx dataset and tissue sample sizes^{9,22}. Across the 87 individuals, we obtained 3877 distinct COPs from single cell data, out of 254,647 gene pairs tested (Supplementary Table 1). Of these, we found 613 COPs present in 2 to 5 individuals and 377 COPs in more than 5 individuals (Fig. 2b). The modest number of COPs found in common between multiple individuals may suggest a high individual and cellular specificity of COPs but may also stem from the lack of power in detecting COPs in individuals with a low number of cells available. Yet, when comparing COP replication rates between experiments of the same individual to experiments from other individuals, we found COP replication to be clearly higher than expected by chance (23.8% versus 2.8%, Wilcoxon test p-value = $3e^{-14}$, Supplementary Fig. 2). Moreover, the 4797 distinct genes composing the 3877 COPs are found widespread across the genome, with between 18.1% and 31.3% of genes per chromosome associated with at least one COP (Supplementary Fig. 3).

In addition to scCOPs mapped for iPSCs across individuals, we also identified 2589 scCOPs from 26,589 cells of a single human lymphoblastoid cell line (LCL, GM12878, see Methods) using the SHARE-seq single cell dataset¹⁷, a multimodal dataset further explored in the next sections. Although different single cell methodologies and sequencing depths were used between the iPSC and LCL COP discovery experiments, the discovery of several thousand LCL scCOPs in a single individual is likely linked to the availability of many thousands of cells and may serve as an example for what may be obtained per individual with large datasets. Together with the iPSC results, the extensive discovery of COPs using single cell data demonstrates that local gene co-expression is present for a large proportion of genes and seems to vary across individuals.

Fig. 1 Summary of single cell and bulk local gene co-expression detection across 87 individuals. Using single cell data we identify scCOPs per individual based on measuring the gene expression correlation across all cells of the same individual. Using bulk data (right part of the plot) we identify bulkCOPs by correlating the expression levels of nearby genes across individuals.

Single cell COP identification discovers unique biology

Next, using bulk RNA-seq data for the same set of 87 individuals, we identified 3705 bulkCOPs by correlating gene expression across individuals (Fig. 1, see Methods). This compares to 3877 scCOPs identified in the same samples, with 313 COPs found in both datasets (Fig. 2c). Of note, 311 out of 377 scCOPs replicated across 5 or more individuals were not found with bulk data, yet, many of those are clearly functionally related genes, such as the histone 1 gene cluster (Supplementary Table 1). Indeed, although the set of COPs differ between single cell and bulk datasets, we found an enrichment for both scCOPs and bulkCOP genes to belong in the same pathway ($OR > 1.7$, $p\text{-value} < 1.3e^{-27}$, Fig. 2d) and the same protein complex ($OR > 11.5$, $p\text{-value} < 1.9e^{-22}$). Moreover, both scCOPs and bulkCOPs are enriched in previously identified⁹ GTEx COPs which are conserved across $>50\%$ tissues ($OR = 6.2$ to 15.6 , $p\text{-value} < 6.1e^{-42}$). These enrichments greatly increase for the scCOPs replicated across 5 or more individuals ($OR = 5.5$ to 82.4 , $p\text{-value} < 1.4e^{-36}$). We confirmed that the scCOP enrichments are not driven by the COPs that are in common with bulk data (Supplementary Fig. 4). In addition, these enrichments are also replicated on the 2589 scCOPs mapped for LCLs ($OR = 1.9$ to 5.7 , $p\text{-value} < 4.0e^{-10}$, Supplementary Fig. 5). Although highly enriched, the overlap between scCOPs and functionally-related gene pairs ranged between 1.3% (same complex) and 16.1% of the

scCOPs (same pathway, Fig. 2d), i.e. functionally-related genes only represent a minority of all co-expressed gene pairs.

The fact that scCOPs and bulkCOPs identify different sets of COPs, yet both show strong functional enrichments, suggests that the single cell COP identification could be complementary to bulk COP identification in discovering novel and biologically-relevant gene co-expression. Out of the 4797 distinct genes present in scCOPs, 1587 genes (33%) are also present in bulkCOPs (Supplementary Fig. 6a). When performing functional enrichments with gProfiler²³ (see Methods) for genes in scCOPs and for genes in bulkCOPs, we observe several enriched biological process terms in common between the two, in particular those related to transcription such as “gene expression” (GO:0010467, adjusted p-value < 2.3×10^{-9}) and “cellular metabolic process” (GO:0044237, adjusted p-value < 2.2×10^{-10}) (Supplementary Table 2, Supplementary Fig. 6b). However, while the majority of bulkCOP significant enrichments are also found with scCOPs (65 out of 87), scCOPs revealed an additional 547 significant enrichments (Supplementary Table 2, Supplementary Fig. 6c). These enrichments included terms related to protein translation, such as “translation” (GO:0006412, adjusted p-value = 3.9×10^{-26}), “protein-containing complex subunit organization” (GO:0043933, adjusted p-value = 1.0×10^{-26}) and “Ribosome” (KEGG, adjusted p-value = 2.3×10^{-21}). Moreover, scCOPs are also highly enriched in “Cell Cycle” (Reactome, adjusted p-value = 3.3×10^{-11}) and “Cellular responses to stress” (Reactome, adjusted p-value = 5.1×10^{-21}). The expression of cellular stress response genes due to single cell RNA-seq preparation has previously been reported^{24,25}. While this can be problematic in studies of differential gene expression under multiple conditions, here we measure local gene co-expression, which is largely independent of expression levels. Thus, changes in cell states can be seen as an opportunity of discovering novel co-expression events. Indeed, gene expression differences due to cell cycle phases have been previously studied in this dataset²¹. We have thus discovered COPs per cell cycle phase, by first annotating each of the 7440 cells with its most likely cell cycle phase (see Methods) and then separately identifying COPs per individual from cells of each of the G1 (697 cells, 291 COPs), S (3371 cells, 1786 COPs) and G2M phases (3372 cells, 2483 COPs). We again observed that COP identification is dependent on the sample size for each individual-experiment (Supplementary Fig. 7), which likely explains the modest overlap in COPs between phases (e.g. 300 COPs overlapping between S and G2M phases, Supplementary Fig. 8a). However, the enrichment of COPs in genes belonging to the same pathway, protein complex or conserved COPs shows to be very consistent across all phases (Supplementary Fig. 8b). Likewise, many enriched GO terms are present across the 3 cell cycle phases analysed (Supplementary Table 3), including for “Ribosome” (KEGG, adjusted p-value = 6.1×10^{-7}) and “Cellular responses to stress” (Reactome, adjusted p-value = 1.2×10^{-6}), indicating that while COP identification is very much dependent on the available data, the functional enrichments of COP genes are robust. Overall, these results evidence the benefit of using single cell

data to identify COPs in addition to bulk data and the consistent enrichments in pathways and protein complexes reiterate the functional usefulness of local gene co-expression in cells.

Fig. 2 Features of single cell local gene co-expression (a) number of cells per individual and number of COPs mapped. Fit line corresponds to a linear regression model with 95% confidence intervals; (b) distribution of the percentage of individuals in which COPs are present. The y-axis is log₁₀ scaled. The inner plot counts how many COPs in 1, 2 to 5 (exclusive) and 5 or more individuals; (c) total number of COPs detected with bulk data (bulkCOPs) and single cell data (scCOPs, union across individuals). Numbers in green represent COPs found from both bulk and single cell data; (d) Fisher's exact test odds ratio enrichment (and 95% confidence interval) for the pair of genes in COPs to belong to the same gene pathway, protein complex or in the set of COPs conserved across GTEx tissues. "scCOPs ≥ 5 " are a subset of COPs that are found across 5 or more individuals. X-axis is log-scaled, but values shown are before transformation. The right part of the plot denotes the percentage of COPs in each functional annotation;

Local gene co-expression is kept up from nascent RNA to protein levels

In a bid to determine if local gene co-expression not only occurs in the same cells but actually stems from gene transcription at the same time, we analysed gene transcription initiation from a nuclear run-on dataset (GRO-seq) publicly available for the GM12878 LCL cell line²⁶. In practice, we evaluated whether the nascent transcription of both genes in 2589 scCOPs mapped for LCLs is observed, by correlating the number of reads mapping to TSSs of gene pairs (see Methods). Indeed, we found significant read number correlation for COP gene pairs (Spearman $R = 0.22$, p -value = $2.8e^{-4}$, Fig. 3a). To determine if this correlation is higher than expected by the genomic proximity between COP gene TSSs, we produced a control set of 2589 gene pairs that is not co-expressed but closely matches the distance between COPs' TSSs, which we named 'non-COPs' (see Methods). Importantly, in this set of distance-matched non-COP gene pairs we found no read correlation (Spearman $R = 0.001$, p -value = 0.99, Supplementary Fig. 9). Given that scCOPs were discovered through their concerted expression in the same cell, this finding suggests that genes in a COP are transcribed at the same time and cell, i.e. that local gene co-expression may be synchronous.

Next, to address whether the local gene co-expression observed would be relevant at the cellular level, we used bulk proteomics data (MS/MS) from Mirauta *et al.* 2020²⁷ which is available for 42 out of the 87 HipSci consortium iPSC cell lines studied here (68 out of 152 individual-experiment combinations). This allowed us to assess whether local gene co-expression correlation can be reproduced as protein intensity correlation (total of 9013 genes with protein intensities, see Methods). We found that the 2577 distinct scCOPs discovered in the matching cell lines often display correlated protein intensities, with a mean Spearman correlation of 0.29 across 58 individual-experiment combinations with more than 10 COPs (Fig. 3b). We observed no positive correlation for 2577 distance-matched non-COPs (Spearman $R = -0.004$), indicating that protein intensities correlations are exclusive to COPs (Wilcoxon test p -value = $3.1e^{-7}$). Moreover, this correlation was not present when shuffling gene pair labels for COPs and non-COPs on each individual-experiment (Supplementary Fig. 10a). Furthermore, when averaging intensities across all the 42 individuals a significant correlation in intensity levels is also observed for scCOPs but not for distance-matched non-COPs (scCOPs Spearman $R = 0.15$ p -value = $4e^{-13}$, non-COPs Spearman $R = -0.05$, p -value = 0.2, Supplementary Fig. 10b). This was also observed when considering bulkCOPs (COPs Spearman $R = 0.19$ p -value = $2e^{-11}$, non-COPs Spearman $R = 0.04$, p -value = 0.2, Supplementary Fig. 10c). These results demonstrate that the co-expression of nearby genes often leads to similar protein abundance levels, perhaps unsurprisingly, given the demonstrated functional relatedness between the genes in a COP and their need to ensure co-expression stability.

Fig. 3 GRO-seq and proteomics correlation in scCOPs (a) GRO-seq read correlation for LCL scCOP genes (N = 264). Reads mapping to the TSS positions of each gene were considered. Gene pairs with missing data in at least one of the genes were excluded. The number of reads across all genes in COPs and non-COPs was ranked prior to plotting. Two genes sharing the same number of reads share the same rank; (b) correlation of protein intensities for each of 58 individual-experiments with iPSC proteomics data. The two-tailed Wilcoxon test refers to the comparison of correlation values between iPSC scCOPs and non-COPs for all 58 individual-experiments. The length of the box corresponds to the interquartile range (IQR) with the centre line corresponding to the median, the upper and lower whiskers represent the largest or lowest value no further than $1.5 \times \text{IQR}$ from the third and first quartile, respectively.

Enhancer regulation of local gene co-expression

To further explore the regulatory mechanisms leading to gene co-expression we utilized publicly available multimodal SHARE-seq data, which simultaneously profiles gene expression (scRNA-seq) and open chromatin (scATAC-seq) on the same single cells¹⁷. In particular, we used the data available for a human LCL (GM12878), for which 24,844 cells with both scRNA-seq and scATAC-seq are available (see Methods). Focusing on ATAC-seq peaks overlapping known LCL enhancer regions from the EpiMap repository²⁸, we correlated the activity of enhancers with the expression of nearby genes ($\pm 1\text{Mb}$ window of gene TSS, Fig. 4a, see Methods). Out of 350,182 gene-enhancer pairs tested, 32,883 (9.4%) were determined as significant gene-enhancer associations (FDR < 5%, Spearman correlation > 0.05, Supplementary Table 4). As expected, significant gene-enhancer associations are more often found at close distances between the gene TSS and the enhancer, yet, in 83% of associations the gene TSS and enhancers are >100kb apart (Supplementary Fig. 11). Importantly, we found significant correlations between the gene-enhancer associations and two recent orthogonal datasets of gene-enhancer associations in LCLs: (i) EpiMap²⁸, based on gene expression and epigenetic modification measurements (Spearman $R = 0.18$, $p\text{-value} < 2.2e^{-16}$) and (ii)

activity-by-contact (ABC) model⁸, based on CRISPR perturbations (Spearman $R = 0.06$, p -value 3.9×10^{-15} , Supplementary Fig. 12a,b). Of note, these correlation levels are higher than what is obtained when comparing the EpiMap and ABC model orthogonal methods (Spearman $R = 0.04$, p -value $= 1.8 \times 10^{-6}$, Supplementary Fig. 12c), evidencing their orthogonal discovery of gene-enhancer associations and the lack of a gold standard for evaluating such associations. To further confirm the validity of our gene-enhancer associations, we analysed Hi-C data (5kb resolution) for LCLs²⁹. We find that the correlation level of gene-enhancer associations corresponds to higher Hi-C contact intensities (Spearman $R = 0.13$, p -value $< 2.2 \times 10^{-16}$, Supplementary Fig. 13a). Importantly, this correlation was not observed for distance-matched control regions (see Methods, Supplementary Fig. 13b). Indeed, 22,102 (67.2%) out of the 32,883 significant gene-enhancer associations displayed higher Hi-C contacts than expected by their distance (Supplementary Fig. 13c).

Next, using the 32,883 gene-enhancer associations identified, we explored the role of enhancer sharing in local gene co-expression. For this, we first identified enhancers associated with both genes of a gene pair ($\text{corr} > 0.05$, $\text{FDR} < 5\%$, see Methods). Notably, we found that 95.6% of the 2589 COPs identified share at least one enhancer (mean of 6.2 enhancers shared, range: 0 to 26, Fig. 4b,c), significantly more than for 2589 distance-matched non-COPs, where 32.6% shared at least one enhancer (mean of 0.9 enhancers shared, Fisher's exact test $\text{OR} = 2.9$, p -value $= 2 \times 10^{-114}$). These results were not driven by a difference in the number of enhancers tested for associations between COPs (21.6 enhancers) and non-COPs (21.5 enhancers, Wilcoxon test p -value $= 0.6$, Supplementary Fig. 14). Remarkably, using the same Hi-C dataset as before, we found that 53.1% of COPs are associated with at least one enhancer displaying high Hi-C contacts to both genes (Fig. 4d, intensities higher than the 75th quantile, see Methods), indicating that these enhancers are likely to interact with both gene's TSSs. Indeed, 76.8% COPs share enhancers with higher Hi-C contacts than distance-matched control regions (Supplementary Fig. 15). Regulatory element sharing between genes in COPs was also observed when considering all ATAC-seq peaks within $\pm 100\text{Kb}$ of each gene TSS (regardless of overlap with known enhancers), with 76.7% COPs sharing peaks compared to 25.1% for non-COPs (Supplementary Fig. 16). Overall, the widespread usage of shared regulatory elements shows to be an hallmark of local gene co-expression and may be a key mechanism in ensuring that the transcription of nearby genes is similar and results in stable co-expression.

Fig. 4 Identification of enhancers linked to local gene co-expression (a) SHARE-seq and identification of gene-enhancer associations and enhancer sharing, based on correlating gene expression and chromatin activity on the same cells between nearby genes and enhancer regions (+/-1 Mb from gene TSS). Shared enhancers are identified as having a significant association (FDR 5%, Spearman correlation > 0.05) with multiple nearby genes; (b) percentage of COPs and non-COPs sharing at least one enhancer, i.e. enhancer significantly associated with both genes in the pair; (c) number of significantly associated enhancers per COP and non-COP. Two-tailed Wilcoxon test was performed between COP and non-COP numbers of enhancers. The length of the box corresponds to the IQR with the centre line corresponding to the median, the upper and lower whiskers represent the largest or lowest value no further than $1.5 \times \text{IQR}$ from the third and first quartile, respectively; (d) number of COPs and non-COPs with Hi-C support (e.g. both enhancer-gene1 and enhancer-gene2 having Hi-C contact higher than the 75th quantile). Note that non-COPs are less likely to share enhancers and thus a smaller number of gene pairs is liable to have Hi-C support.

Discussion

Previous work described the co-expression of nearby genes through the co-variation of expression levels across individuals for a multitude of tissues^{9,11}. However, this work has been achieved with bulk RNA-seq, often in primary tissues, in which the heterogeneity of cell types can confound gene co-expression measurements. Indeed, various studies have shown that the variation in cell-type abundance between samples of the same tissue can affect gene co-expression measurements, i.e. these patterns may reflect the differential expression between cell types of a tissue and even mask robust cell-type-specific co-expression patterns^{15,30}. In addition, gene co-expression measured across individuals leverages differences in their genetics and it's unclear if this co-expression could also be detected independently of genetic regulation, i.e. if it could be detected in a single individual with a fixed genetic background. Moreover, by measuring transcription events in a single cell rather than using gene expression averages, single cell measurements may be better suited for the study of local gene co-expression than bulk measurements. Here, leveraging the inter-cellular variability of expression levels present in single cell RNA-seq datasets, we confirmed the widespread presence of thousands of co-expressed gene pairs across the genome in two specific cell types (iPSC and LCL). Indeed, we discovered more distinct COPs (3877 iPSC COPs) with single cell data than with bulk data (3705 COPs) for the same set of 87 individuals. Interestingly, only a fraction of these COPs were discovered from both datasets even though the original material came from the same cell lines. Besides technical differences in the library preparation, experimental design, sequencing and processing of bulk and single cell datasets, obtaining disparate COPs between these approaches is also expected from other points of view. On one hand, when using bulk data to estimate gene co-expression across individuals we leverage inter-individual variability in expression levels, which may be controlled by genetic variations. On the other hand, using single cell data and leveraging inter-cellular variability, the genomes of each cell are the same and gene expression variability may stem from different cell states. These two systems for detecting gene co-expression (perturbation from genetic variants versus perturbation from cell state) are clearly different and thus should be expected to reveal different co-expression events. Our study confirms this view, by finding a similar global pattern of local gene co-expression between bulk and single cell COPs (e.g. both are strongly enriched for being in the same pathway), yet each approach finding complementary sets of COPs with different functional enrichments (e.g. scCOP enrichment in cell cycle genes).

A benefit of using single cell data across multiple individuals is the ability to study the individual-specificity of local gene co-expression. Our results suggest that COPs may be highly individual-specific, however, the limited number of cells per each individual impacts the ability to discover COPs, with only a handful of COPs identified for individuals with less than 20 Smartseq2 cells but with 2589 COPs discovered with >25,000 SHARE-seq cells. Future studies including a

higher number of cells per individual as well as more individuals, such as the single-cell eQTLGen consortium³¹, will unlock the ability to further explore the interindividual variability of gene co-expression and may even reveal differences between groups of individuals (e.g. based on sex or certain experimental conditions).

Multimodal single cell methods performing scRNA-seq and scATAC-seq in the same cells have proven useful in connecting regulatory elements with their target gene expression^{17,32,33}. This approach is orthogonal to those using bulk RNA-seq data, such as EpiMap, as the activity of gene expression and regulatory activity is directly observed and correlated in the same cells. Here, we have characterised gene-enhancer links using multimodal data, finding them enriched in Hi-C contacts and the deep resolution of multimodal data allowed us to explore the regulatory circuitry between local gene co-expression and nearby regulatory elements. While the action of multiple regulatory elements on the same gene has often been described^{34,35}, the action of a regulatory region on multiple genes is seldom explored¹⁷. Notably, here we found that the vast majority of co-expressed gene pairs (>95%) share at least one regulatory region, much more than expected by chance, indicating this may be a key mechanism for achieving gene co-expression and that the action of enhancers may often be pleiotropic. Further work could unveil not only which regulatory elements are involved in regulating each gene(s) but also provide detail on the actual set of regulatory elements that may work together, by finding evidence for the co-activity of these elements in the same cell.

COPs derived from single cell data are evidence of co-expression in the same cells. For a deeper understanding on the path of gene co-expression, from start to endpoint, we explored datasets of nascent transcription (GRO-seq) and proteomics. We provided evidence that local gene co-expression may stem from the synchronous transcription of genes. Strikingly, despite the known poor correlation between mRNA and protein levels³⁶, we observed that gene co-expression leads to matched protein levels. Together with our previous findings regarding (i) functional relatedness of COPs and (ii) the close matching between expression levels and expression variation of co-expressed genes⁹, we can stipulate that this consistency in gene and protein levels may derive from the need for cells to closely match quantities of functional related proteins. Overall, our results lead us to conceive a model in which (1) genes may be arranged in the genome due to functional relatedness and/or the need to keep a similar level of expression, (2) nearby co-expressed genes are transcribed synchronously, (3) the gene co-expression is kept up to the protein level and (4) this is achieved due to their proximity and sharing of regulatory elements. The potential benefit of such a system could include the minimisation of expression noise and maintenance of appropriate protein stoichiometry³⁷⁻³⁹.

Our work serves as further evidence for the widespread co-expression of nearby genes in addition to providing clues regarding the elements involved in their co-expression. With the unravelling of large projects such as the Human Cell Atlas⁴⁰, which will contain single-cell RNA-seq complemented with

chromatin, protein and spatial information across cell types, the future promises vast datasets in which to further explore the shared regulatory architecture of gene (co-)expression.

Methods

Single cell COP identification in iPSCs

Single cell RNA-seq was produced by Cuomo *et al.* 2020²⁰ for a subset of cell lines from the HipSci consortium¹⁹. Previously processed and quality controlled log2-transformed TPM gene expression measurements in iPSC (undifferentiated cells, day 0) for 87 individuals were obtained from Cuomo *et al.* 2021²¹. These gene expression measurements were further normalised by scran⁴¹ and transformed to match a normal distribution $N(0,1)$. Single cell COPs were identified using the previously described method for COP identification in Ribeiro *et al.* 2021⁹. To exclude potential batch effects, COPs were identified for each individual-experiment combination (total of 152 combinations), as recommended by Cuomo *et al.* 2021²¹. Gene names were annotated with genomic coordinates (hg19), gene types and Ensembl gene IDs from Gencode v19. Only protein-coding genes were tested for co-expression ($N = 18,943$). In each individual-experiment gene expression matrix, genes without any expression variability across the cells (e.g. only zeroes) were excluded. Genes in non-autosomes (including mitochondrial genes) or the MHC region (chr6:29500000-33600000) were also excluded. We measured co-expression of all genes within a cis-window of 1Mb (based on TSS coordinates) and used 1000 permutations to determine observed versus expected empirical p-values, as done before⁹. Positively correlated COPs were determined by having a Benjamini-Hochberg FDR $< 5\%$ (the minimum correlation across COPs is 0.248). Three out 152 individual-experiment combinations were excluded for having a ratio of number of COPs per number of cells above the 75th quartile + interquartile range (IQR) $\times 3$. Finally, 14,176 gene pairs (out of 268,417 gene pairs within 1Mb window) with cross-mappability score > 10 were excluded (75mer Exon, 36mer UTR, 2 mismatch, symmetric mean)¹⁸. In addition, COPs with correlation > 0.99 (i.e. near perfect correlation) were excluded as these are likely to be artifacts. In most analyses, COPs from the different experiments of the same individual were combined together as their union.

Single cell COPs were also identified for each cell cycle phase. For this, the “CellCycleScoring” function of the Seurat 4.0 R package⁴² was used to predict the cell cycle phase for each of the 7440 cells. Then, a matrix was produced with cells of each cell cycle phase (G1, S and G2M) and COP identification for each was performed as described above.

Bulk COP identification in iPSCs

Bulk RNA-seq was produced by the HipSci consortium¹⁹. Processed and quality controlled gene expression measurements were obtained from Cuomo *et al.* 2021²¹ for the same set of 87 individuals for which single cell data was available. As recommended by Cuomo *et al.* 2021 to exclude confounding factors, the first 15 PCA principal components from the expression matrix were regressed out using QTLtools⁴³. COPs were identified by correlating expression levels across all 87 individuals, as previously done⁹. A total of 244,341 gene pairs were tested for co-expression. All parameters and post-processing (1Mb window size, 1000 permutations, 5% FDR, only coding genes, mappability filters) were performed identical as for single cell COP identification.

Single cell COP identification in LCLs

Preprocessed gene counts from single cell RNA-seq for GM12878 LCL (GSM4156603, rep3) produced by Ma *et al.* 2020¹⁷ were obtained through GEO (GSE140203). This included 26,434 genes and 26,589 cells. Cells where <300 or >7,500 genes were expressed had been previously removed. As done for the iPSC dataset, genomic coordinates (hg19) and Ensembl gene IDs were obtained from Gencode v19 and non-protein-coding genes were excluded, as well as genes in non-autosomes or in the MHC region. Moreover, genes expressed in less than 100 cells were removed here, resulting in keeping a total of 10,821 genes. Lastly, given that >76% of non-zero gene counts were 1s, the gene expression matrix was binarised. Using this matrix, COPs were identified as previously (1000 permutations, 1Mb window, cross-mappability score < 10, correlation ≤ 0.99), with the exception that a minimum correlation value of 0.2 was used in addition to statistical significance (FDR<5%).

Creation of control non-COP dataset

To control for distance effects in local gene pair co-expression, a control set of non-co-expressed genes (non-COPs) was built for each individual-experiment combination, as well as for the set of all distinct COPs found across individuals in the following manner. For each COP, a gene pair that is not co-expressed (non-COP) in any individual but is matched for the distance between TSSs observed in the COP is randomly sampled. The maximum distance discrepancy allowed between a COP and a matching non-COP is 100 bp and non-COPs are picked randomly without replacement. Given the large number of non-COPs available (>90% of all gene pairs tested), a matching non-COP was found for every COP. The same procedure was applied independently for single cell and bulk COPs.

Enrichment to functionally related gene pairs and annotation terms

The overrepresentation of COPs as functional-related gene pairs was assessed with one-way Fisher's Exact tests to: (i) genes belonging to the same biological pathway (N = 24,544), gathered from

KEGG^{44,45} and Reactome⁴⁶ through the Ensembl v98 BioMart data mining tool⁴⁷ (25 May 2020); (ii) genes belonging to the same human protein complex (N = 350), gathered from the CORUM 3.0 database⁴⁸ and hu.MAP⁴⁹ (20-April-2020). UniprotKB IDs were converted to Ensembl IDs with the Uniprot ID mapping tool⁵⁰. In addition, enrichment tests were performed for a set of COPs conserved across >50% GTEx tissues in which they are expressed (N = 2441) from Ribeiro *et al.* 2021⁹. The background set of gene pairs used included the union of all gene pairs tested for scCOPs and bulkCOPs (N = 259,834).

Enrichments to annotation terms for genes in bulkCOPs and scCOPs (union) were performed with the gprofiler2 R package^{23,51} (December 2021). For this, a multi-query search was performed for the “GO:BP”, “KEGG” and “REACTOME” sources. As the background set, the union of all gene pairs tested for bulkCOPs and scCOPs was used. The results for “GO:BP” were further summarised using REVIGO⁵² using the parameters “semantic similarity”, “homo sapiens” and “Small (0.5)” return list and plotted with the “TreeMap” R package.

GRO-seq data and COPs

Nascent RNA GRO-seq information for the GM12878 LCL was obtained from Core *et al.* 2014²⁶ bigWig files. These were converted to bedgraph format using UCSC utils⁵³ and bedtools intersect was used to combine gene TSS (Gencode v19⁵⁴, matching for genomic strand) with the number of GRO-seq reads mapped to the TSS position. A total of 5432 genes had GRO-seq reads in their TSS position. Finally, the number of reads between the two genes in LCL COPs was correlated (Spearman correlation). For a comparison, the same was performed for non-COP gene pairs. Gene pairs with missing data for at least one of the two genes were excluded from the correlation.

Proteomics data and COPs

Quantitative proteomic data (Tandem Mass Tag Mass Spectrometry) for 202 iPSC lines derived from 151 individuals of the HipSci consortium samples was collected from Mirauta *et al.* 2020²⁷. This consisted of a processed matrix of protein isoform intensities across individuals, of which 42 are included in the set of 87 individuals assessed here. To measure correlation of protein intensities among COPs and non-COPs, first the protein isoform intensities were converted into ‘gene-based intensities’ by summing all the protein isoform intensities from each gene. Protein intensities for 9013 genes were obtained in this manner. Next, the intensities for COP gene pairs (and non-COP gene pairs, separately) were correlated (Spearman correlation) per individual-experiment combination, thus producing a correlation value for COPs and a correlation value for non-COPs for each individual-experiment. To ensure the availability of enough data for correlation, only individual-experiment combinations with >10 COPs were considered (a total of 58 out of 68 possible combinations). Gene pairs with missing data for at least one of the two genes were excluded from the

correlation. As a control, the genes across the pairs of each individual-experiment COPs/non-COPs were shuffled once.

Gene-enhancer associations using SHARE-seq

To identify enhancer regions associated with nearby gene expression, processed and quality-controlled ATAC-seq peaks were retrieved from Ma *et al.* 2020¹⁷ (GSM4156592, rep3, 507,307 peaks across 67,418 cells). Of these, only 24,844 cells that also had gene expression measurements were kept. GM12878-specific enhancer annotations from the EpiMap repository²⁸ for hg19 were obtained (18-state chromHMM models). Only regions of genic enhancers (EnhG1, EnhG2) and active enhancers (EnhA1, EnhA2) were considered as enhancer regions. Successive enhancers were merged using bedtools (v2.29.2) *merge* command with default parameters (i.e. only merging “book-ended” features), leading to 33,776 distinct enhancer regions. ATAC-seq peaks were then intersected with these enhancer regions using bedtools intersect with the -F 0.5 parameters, thus requiring that at least 50% of the peak overlaps an enhancer region, resulting in 6,443,451 enhancer-cell combinations (17,765 distinct enhancer regions). Finally, gene expression and open chromatin activity measurements (binarised) were integrated for the same cells and enhancer regions within +/-1Mb of a gene TSS were tested for association with the gene through Pearson correlation, in a total of 350,182 tests performed. Only protein-coding genes in non-autosomal chromosomes were considered (MHC also excluded). Gene-enhancer pairs with correlation coefficient > 0.05 and FDR < 5% were identified as significant gene-enhancer associations (32,883 associations). This correlation coefficient was chosen in order to retain a relatively high number of associations while still providing a biological signal. Enhancers significantly associated with both genes of COPs and non-COPs were identified as shared enhancers. The analysis using ATAC-seq peaks (within 100kb of gene TSS) instead of overlapping with enhancer regions was performed in the same manner as above.

Comparison to EpiMap and ABC model gene-enhancer associations

To evaluate the gene-enhancer associations identified here, sets of LCL-specific gene-enhancer maps from the EpiMap repository (links_by_group.lymphoblastoid.tsv.gz)²⁸ and the activity-by-contact (ABC) model⁸ (AllPredictions.AvgHiC.ABC0.015.minus150.ForABCPaperV3.txt) were obtained. The ABC model file was processed to obtain only data for “GM12878-Roadmap” CellType entries and gene names were converted to Ensembl gene IDs using the gprofiler2 R package (gconvert function)²³. EpiMap and ABC model enhancers were separately intersected with the enhancer regions produced here using bedtools intersect with -wa -wb parameters. Matching gene-enhancer pairs between datasets (N = 99,911 gene-enhancer pairs between EpiMap and SHARE-seq, N = 19,801 between ABC model and SHARE-seq) were obtained and the correlation value from the SHARE-seq

dataset (no filter) was correlated with EpiMap scores and ABC scores, respectively. As a comparison, EpiMap scores and ABC model scores for matching gene-enhancer pairs were also correlated.

Hi-C support of gene-enhancer and COP-enhancer associations

Hi-C data the GM12878 cell line (LCL) at 5 kb resolution was obtained from Rao *et al.* 2014²⁹. KR normalised (MAPQG0) 5 Kb bins encompassing the TSSs coordinate of gene and midpoint of enhancer regions was obtained through custom Python scripts. Normalised Hi-C contacts were log2-transformed. Using this, gene-enhancer association strength was correlated with Hi-C contacts through Spearman correlation. Gene-enhancer associations with Hi-C contacts above the 75% quantile across all tested gene-enhancer pairs were determined as supported by Hi-C. Missing data (genes or enhancers without Hi-C data) was replaced with 0. As a control, for each gene-enhancer pair, another control pair composed of the gene TSS and an ‘enhancer’ region on the opposite up- or down-stream location in respect to the gene TSS was produced (e.g. if an enhancer is 1000bp upstream of the gene TSS, the matching control region is 1000bp downstream of the gene TSS).

Data Availability

The COPs and enhancer-gene associations produced where are available for download as supplementary tables and through the LoCOP public database (<https://glcoex.unil.ch/>).

Code Availability

The code used for analysis is available under: https://github.com/diogomribeiro/sc_cop

Acknowledgements

O.D. and D.M.R. have been funded by a Swiss National Science Foundation (SNSF) project grant (PP00P3_176977). D.M.R. has also been funded by the European Union’s Horizon 2020 research and innovation programme under the Marie Skłodowska-Curie grant agreement No 885998. The funders had no role in study design, data collection and analysis, decision to publish, or preparation of the manuscript.

Author Contributions

D.M.R. performed the experiments, analysed the data and wrote the manuscript. C.Z. helped with the gene-enhancer association analysis. O.D. supervised the study and revised the manuscript.

Competing Interests

The authors declare no competing interests.

Supplementary Information

Supplementary Fig. 1 Number of COPs identified per individual-experiment. The y-axis represents the individual and is sorted by total number of COPs, the experiment is color-coded. Multiple experiments are present for each individual. Note that the same COP can be identified in multiple experiments of the same individual.

Supplementary Fig. 2 Percentage of COPs replicated across multiple experiments of the same individual.

The COPs of each individual-experiment are tested for replication across all other experiments of the same individual. Only 48 out of 87 individuals with more than one experiment were considered. The expected values were calculated by comparing the COPs to a sample with the same number of COPs as those in the other experiments, but drawn from experiments of different individuals. For instance, if calculating the replication rate of 10 COPs of a certain individual-experiment against 100 other COPs from other experiments of the same individual, these 100 COPs will instead be withdrawn from the pool of all COPs present across individuals. The length of the box corresponds to the interquartile range (IQR) with the centre line corresponding to the median, the upper and lower whiskers represent the largest or lowest value no further than $1.5 \times \text{IQR}$ from the third and first quartile, respectively.

Supplementary Fig. 3 Genes in COPs across chromosomes. Red represents a co-expressed gene, black a non-co-expressed gene. Genes are indexed by their genomic position in the chromosome. The percentages on the right side of each plot denote the percentage of genes that are co-expressed per chromosome.

Supplementary Fig. 4 scCOP enrichment for the same gene pathway, protein complex or GTEx conserved COPs separated by identification in bulk. “scCOPs >=5” are the subset of scCOPs found across 5 or more individuals. Odds ratios were calculated with one-sided Fisher’s Exact tests and the 95% confidence interval is shown. X-axis is log-scaled, but values shown are before transformation. The right part of the plot denotes the percentage of COPs in each functional annotation.

Supplementary Fig. 5 COP enrichment for the same gene pathway, protein complex or GTEx conserved COPs for SHARE-seq LCL COPs. Odds ratios were calculated with one-sided Fisher’s Exact tests and the 95% confidence interval is shown. X-axis is log-scaled, but values shown are before transformation. As a comparison, results for COPs identified for GTEx LCLs (Cells EBV-transformed lymphocytes) from Ribeiro *et al.* 2021⁹ are also shown used. The right part of the plot denotes the percentage of COPs in each functional annotation.

Supplementary Fig. 7 Number of cells per individual and number of COPs mapped per cell cycle phase (a) G1 phase, (b) S phase and (c) G2M phase.

Supplementary Fig. 8 Cell cycle phase COP overlap and functional enrichment (a) COP overlap between the G1, S and G2M phases (Venny v2.1, <https://bioinfogp.cnb.csic.es/tools/venny>); (b) COP enrichment for the same gene pathway, protein complex or GTEx conserved COPs for each cell cycle phase. ‘All phases’ represent the COP identification with all available cells. Odds ratios were calculated with one-sided Fisher’s Exact tests and the 95% confidence interval is shown. X-axis is log-scaled, but values shown are before transformation. The right part of the plot denotes the percentage of COPs in each functional annotation.

Supplementary Fig. 9 GRO-seq read correlation for non-COPs. N = 150. Reads mapping to the TSS positions of each gene were considered. Gene pairs with missing data in at least one of the genes were excluded. The number of reads across all genes in COPs and non-COPs was ranked prior to plotting. Two genes sharing the same number of reads share the same rank.

Supplementary Fig. 10 Proteomics support for COPs and non-COPs (a) gene pair correlation of protein intensities for each of 58 individual-experiments with proteomics data when shuffling gene pair labels for COPs and non-COPs, separately, on each individual-experiment. Missing data was replaced with 0. The length of the box corresponds to the IQR with the centre line corresponding to the median, the upper and lower whiskers represent the largest or lowest value no further than $1.5 \times \text{IQR}$ from the third and first quartile, respectively. (b) correlation of protein intensities averaged across 42 individuals for all distinct scCOPs identified in those individuals and their matching non-COPs; (c) protein intensity correlation for bulk COPs and non-COPs across all individual-experiments. Fit lines correspond to a linear regression model with 95% confidence intervals.

Supplementary Fig. 11 Absolute distance distribution between significantly associated and other gene-enhancer pairs tested. The midpoint of the enhancer and the gene TSS were used for the distance calculation. Significant associations were defined as a Pearson correlation FDR < 5% and correlation > 0.05.

Supplementary Fig. 12 Correlation between gene-enhancer association scores between several approaches. (a) SHARE-seq (produced here) versus EpiMap (N = 99,911); (b) SHARE-seq versus ABC model (N = 19,024); (c) EpiMap versus ABC model (N = 18,375). For each comparison, only gene-enhancer pairs assessed in both associations were considered. Values were quantile normalised (mean = 0, sd = 1). No score or correlation filters were applied.

Supplementary Fig. 13 Hi-C support of gene-enhancer associations. (a) Hi-C contact intensities per gene-enhancer association correlation; (b) Hi-C contact intensities per gene-enhancer association correlations when considering control ‘enhancer’ regions on the opposite up- or down-stream location in respect to the gene TSS; (c) Hi-C contact intensity difference between real and control regions. A shift of the distribution to the right (above 0) represents higher Hi-C contacts in the real data compared to control. Missing data (genes or enhancers without Hi-C data) was replaced with 0.

Supplementary Fig. 14 Boxplot of the number of enhancers tested per COP and non-COP. The length of the box corresponds to the IQR with the centre line corresponding to the median, the upper and lower whiskers represent the largest or lowest value no further than $1.5 \times \text{IQR}$ from the third and first quartile, respectively. Outlier values are not shown.

Supplementary Fig. 15 Hi-C contact intensity difference between real and control regions. Control regions are the opposite up- or down-stream location in respect to the gene TSS. The mean Hi-C contacts between the enhancer and both genes in a COP were used. A shift of the distribution to the right (above 0) represents higher Hi-C contacts in the real data compared to control. Missing data (genes or enhancers without Hi-C data) was replaced with 0.

Supplementary Fig. 16 Percentage of COPs and non-COPs sharing at least one peak. A total of 31,403 significant gene-peak associations were identified (correlation > 0.05, FDR < 5%).

Bibliography

1. Delaneau, O. et al. Chromatin three-dimensional interactions mediate genetic effects on gene expression. *Science* 364, (2019).
2. Klemm, S. L., Shipony, Z. & Greenleaf, W. J. Chromatin accessibility and the regulatory epigenome. *Nature Reviews Genetics* vol. 20 207–220 (2019).
3. Andersson, R. & Sandelin, A. Determinants of enhancer and promoter activities of regulatory elements. *Nat. Rev. Genet.* 21, 71–87 (2020).
4. Vihervaara, A., Duarte, F. M. & Lis, J. T. Molecular mechanisms driving transcriptional stress responses. *Nat. Rev. Genet.* 19, 385–397 (2018).
5. Strober, B. J. et al. Dynamic genetic regulation of gene expression during cellular differentiation. *Science* vol. 364 1287–1290 (2019).
6. Kvon, E. Z., Waymack, R., Gad, M. & Wunderlich, Z. Enhancer redundancy in development and disease. *Nat. Rev. Genet.* 22, 324–336 (2021).
7. Panigrahi, A. & O’Malley, B. W. Mechanisms of enhancer action: the known and the unknown. *Genome Biol.* 22, 1–30 (2021).
8. Nasser, J. et al. Genome-wide enhancer maps link risk variants to disease genes. *Nature* 593, 238–243 (2021).
9. Ribeiro, D. M. et al. The molecular basis, genetic control and pleiotropic effects of local gene co-expression. *Nat. Commun.* 12, 4842 (2021).
10. Hurst, L. D., Pál, C. & Lercher, M. J. The evolutionary dynamics of eukaryotic gene order. *Nature Reviews Genetics* vol. 5 299–310 (2004).
11. Soler-Oliva, M. E., Guerrero-Martínez, J. A., Bachetti, V. & Reyes, J. C. Analysis of the relationship between coexpression domains and chromatin 3D organization. *PLoS Comput. Biol.*

- 13, e1005708 (2017).
12. Ghanbarian, A. T. & Hurst, L. D. Neighboring Genes Show Correlated Evolution in Gene Expression. *Mol. Biol. Evol.* 32, 1748–1766 (2015).
13. Ebisuya, M., Yamamoto, T., Nakajima, M. & Nishida, E. Ripples from neighbouring transcription. *Nat. Cell Biol.* 10, 1106–1113 (2008).
14. Donovan, M. K. R., D’Antonio-Chronowska, A., D’Antonio, M. & Frazer, K. A. Cellular deconvolution of GTEx tissues powers discovery of disease and cell-type associated regulatory variants. *Nat. Commun.* 11, 955 (2020).
15. Farahbod, M. & Pavlidis, P. Untangling the effects of cellular composition on coexpression analysis. *Genome Res.* 30, 849–859 (2020).
16. Zhu, C., Preissl, S. & Ren, B. Single-cell multimodal omics: the power of many. *Nat. Methods* 17, 11–14 (2020).
17. Ma, S. et al. Chromatin Potential Identified by Shared Single-Cell Profiling of RNA and Chromatin. *Cell* 183, 1103–1116.e20 (2020).
18. Saha, A. & Battle, A. False positives in trans-eQTL and co-expression analyses arising from RNA-sequencing alignment errors. *F1000Res.* 7, 1860 (2018).
19. Kilpinen, H. et al. Common genetic variation drives molecular heterogeneity in human iPSCs. *Nature* 546, 370–375 (2017).
20. Cuomo, A. S. E. et al. Single-cell RNA-sequencing of differentiating iPS cells reveals dynamic genetic effects on gene expression. *Nat. Commun.* 11, 810 (2020).
21. Cuomo, A. S. E. et al. Optimizing expression quantitative trait locus mapping workflows for single-cell studies. *Genome Biol.* 22, 188 (2021).
22. Consortium, T. G. & The GTEx Consortium. The GTEx Consortium atlas of genetic regulatory

- effects across human tissues. *Science* vol. 369 1318–1330 (2020).
23. Raudvere, U. et al. g:Profiler: a web server for functional enrichment analysis and conversions of gene lists (2019 update). *Nucleic Acids Res.* 47, W191–W198 (2019).
24. Wu, H., Kirita, Y., Donnelly, E. L. & Humphreys, B. D. Advantages of Single-Nucleus over Single-Cell RNA Sequencing of Adult Kidney: Rare Cell Types and Novel Cell States Revealed in Fibrosis. *J. Am. Soc. Nephrol.* 30, 23–32 (2019).
25. Chen, Z., Yang, Z., Yuan, X., Zhang, X. & Hao, P. scSensitiveGeneDefine: sensitive gene detection in single-cell RNA sequencing data by Shannon entropy. *BMC Bioinformatics* 22, 211 (2021).
26. Core, L. J. et al. Analysis of nascent RNA identifies a unified architecture of initiation regions at mammalian promoters and enhancers. *Nat. Genet.* 46, 1311–1320 (2014).
27. Mirauta, B. A. et al. Population-scale proteome variation in human induced pluripotent stem cells. *Elife* 9, (2020).
28. Boix, C. A., James, B. T., Park, Y. P., Meuleman, W. & Kellis, M. Regulatory genomic circuitry of human disease loci by integrative epigenomics. *Nature* 590, 300–307 (2021).
29. Rao, S. S. P. et al. A 3D map of the human genome at kilobase resolution reveals principles of chromatin looping. *Cell* 159, 1665–1680 (2014).
30. McCall, M. N., Illei, P. B. & Halushka, M. K. Complex Sources of Variation in Tissue Expression Data: Analysis of the GTEx Lung Transcriptome. *Am. J. Hum. Genet.* 99, 624–635 (2016).
31. van der Wijst, M. et al. The single-cell eQTLGen consortium. *Elife* 9, (2020).
32. Mimitou, E. P. et al. Scalable, multimodal profiling of chromatin accessibility, gene expression and protein levels in single cells. *Nat. Biotechnol.* 39, 1246–1258 (2021).

33. Hu, X., Hu, Y., Wu, F., Leung, R. W. T. & Qin, J. Integration of single-cell multi-omics for gene regulatory network inference. *Comput. Struct. Biotechnol. J.* 18, 1925–1938 (2020).
34. Hong, J.-W., Hendrix, D. A. & Levine, M. S. Shadow enhancers as a source of evolutionary novelty. *Science* 321, 1314 (2008).
35. Osterwalder, M. et al. Enhancer redundancy provides phenotypic robustness in mammalian development. *Nature* 554, 239–243 (2018).
36. Vogel, C. & Marcotte, E. M. Insights into the regulation of protein abundance from proteomic and transcriptomic analyses. *Nature Reviews Genetics* vol. 13 227–232 (2012).
37. Raser, J. M. & O’Shea, E. K. Noise in gene expression: origins, consequences, and control. *Science* 309, 2010–2013 (2005).
38. Goldschmidt, Y. et al. Control of relative timing and stoichiometry by a master regulator. *PLoS One* 10, e0127339 (2015).
39. Taggart, J. C., Zaubner, H., Selbach, M., Li, G.-W. & McShane, E. Keeping the Proportions of Protein Complex Components in Check. *Cell Syst* 10, 125–132 (2020).
40. Lindeboom, R. G. H., Regev, A. & Teichmann, S. A. Towards a Human Cell Atlas: Taking Notes from the Past. *Trends Genet.* 37, 625–630 (2021).
41. Lun, A. T. L., McCarthy, D. J. & Marioni, J. C. A step-by-step workflow for low-level analysis of single-cell RNA-seq data with Bioconductor. *F1000Res.* 5, 2122 (2016).
42. Hao, Y. et al. Integrated analysis of multimodal single-cell data. *Cell* 184, 3573–3587.e29 (2021).
43. Delaneau, O. et al. A complete tool set for molecular QTL discovery and analysis. *Nat. Commun.* 8, 15452 (2017).
44. Kowalczyk, M. S. et al. Single-cell RNA-seq reveals changes in cell cycle and differentiation programs upon aging of hematopoietic stem cells. *Genome Res.* 25, 1860–1872 (2015).

45. Kanehisa, M. KEGG: Kyoto Encyclopedia of Genes and Genomes. *Nucleic Acids Research* vol. 28 27–30 (2000).
46. Jassal, B. et al. The reactome pathway knowledgebase. *Nucleic Acids Res.* 48, D498–D503 (2020).
47. Yates, A. D. et al. Ensembl 2020. *Nucleic Acids Res.* 48, D682–D688 (2020).
48. Giurgiu, M. et al. CORUM: the comprehensive resource of mammalian protein complexes-2019. *Nucleic Acids Res.* 47, D559–D563 (2019).
49. Drew, K. et al. Integration of over 9,000 mass spectrometry experiments builds a global map of human protein complexes. *Mol. Syst. Biol.* 13, 932 (2017).
50. UniProt Consortium. UniProt: a worldwide hub of protein knowledge. *Nucleic Acids Res.* 47, D506–D515 (2019).
51. Kolberg, L., Raudvere, U., Kuzmin, I., Vilo, J. & Peterson, H. gprofiler2 -- an R package for gene list functional enrichment analysis and namespace conversion toolset g:Profiler. *F1000Res.* 9, (2020).
52. Supek, F., Bošnjak, M., Škunca, N. & Šmuc, T. REVIGO summarizes and visualizes long lists of gene ontology terms. *PLoS One* 6, e21800 (2011).
53. Lee, B. T. et al. The UCSC Genome Browser database: 2022 update. *Nucleic Acids Res.* (2021) doi:10.1093/nar/gkab959.
54. Frankish, A. et al. GENCODE reference annotation for the human and mouse genomes. *Nucleic Acids Res.* 47, D766–D773 (2019).