nature research

https://doi.org/10.1038/s41556-020-0529-5

Supplementary information

AP-1 imprints a reversible transcriptional programme of senescent cells

In the format provided by the authors and unedited

SUPPLEMENTAL MATERIAL AND METHODS

Expression microarray pre-processing

Raw Affymetrix HTA 2.0 array intensity data were analyzed using open-source Bioconductor packages on R. The quiescence and the RAS-OIS time series data were normalized together (2 conditions, 2 biological replicates per condition, 6 time points per replicates) using the robust multi-array average normalization approach implemented in the *oligo* package. Internal control probe sets were removed and average expression deciles over time-points were independently defined for each treatment. Probes whose average expression was lower than the 4^{th} expression decile in both conditions were removed for subsequent analyses. To remove sources of variation and account for batch effects, data were finally corrected with the *sva* package. To recover as much annotation information as possible, we combined Affymetrix HTA 2.0 annotations provided by Affymetrix and Ensembl through the packages *hta20sttranscriptcluster.db* and *biomaRt*. Principal component analysis and bi-clustering based on Pearson's correlation and Ward's aggregation criterion were used to confirm consistency between biological replicates and experimental conditions at each step of the pre-processing.

Self-organizing maps (SOM)

Normalized log-scaled and filtered expression values were processed using the unsupervised machine learning method implemented in *oposSOM*[18] to train a self-organizing map. This algorithm applies a neural network algorithm to project high dimensional data onto a two-dimensional visualization space. In this application, we used a two-dimensional grid of size 60 x 60 metagenes of rectangular topology. The SOM portraits were then plotted using a logarithmic fold-change scale. To define

modules of co-expressed meta-genes, we used a clustering approach relying on distance matrix and implemented in *oposSOM*. Briefly, clusters of gene expression were determined based on the patterns of the distance map which visualizes the mean Euclidean distance of each SOM unit to its adjacent neighbors. This clustering algorithm – referred to as D-clustering – finds the SOM units referring to local maxima of their mean distance with respect to their neighbors. These pixels form halos edging the relevant clusters in the respective distance map and enable robust determination of feature clusters in the SOM. We finally performed a gene set over-representation analysis in each cluster considering the Molecular Signature Database (MSigDB) hallmark gene sets using a right-tail modified Fisher's exact test and the hypergeometric distribution to provide *p*-value.

Information theory – derived metrics

To evaluate transcriptome diversity and specialization, we used an approach based on information theory as described in [19].

Gene expression unsupervised clustering

Probes constitutive of the RAS-OIS specific transcriptomic signature were clustered using the weighted gene correlated network analysis approach implemented in the *WGCNA* R package[49]. Standard WGCNA parameters were used for the analysis, with the exceptions of soft-thresholding power, which was defined using methods described by and set at 18. The 7 co-expressed probe clusters identified were further functionally characterized using gene set over-representation tests. The same approach as previously described for the SOM-defined clusters was used.

Histone modification ChIP-seq data processing

Reads were cleaned and trimmed using *fastq-mcf* from the *ea-utils* suite v1.1.2 to remove adapters, low quality bases and reads, and discard reads shorter than 25 bp after filtering. Reads were then aligned to the human reference genome (hg19) with *bowtie* v1.1.1 using best matches parameters (bowtie -v 2 -m 1 --best --strata). Alignment files were further processed with *samtools* v1.2 and *PicardTools* v1.130 to flag PCR and optical duplicates and remove alignments located in Encode blacklisted regions. Fragment size was estimated *in silico* for each library using *spp* v1.10.1. Genome-wide consistency between replicates was checked using custom R scripts. Enriched regions were identified for each replicate independently with *MACS* v2.1.0 with non-IPed genomic DNA as a control (macs2 callpeak --nomodel --shiftsize --shift-control --gsize hs -p 1e-1). These relaxed peak lists were then processed through the irreproducible discovery rate (IDR) pipeline[50] to generate an optimal and reproducible set of peaks for each histone modification and each time point.

ATAC-seq data processing

Paired-ends reads were cropped to 100bp with *trimmomatic* v0.36[51] and cleaned using *cutadapt* v1.8.3[52] to remove Nextera adapters, low quality bases and reads, and discard reads shorter than 25 bp after filtering. Fragments were then aligned to the human reference genome (hg19) using *bowtie2* v2.2.3 discarding inconsistent pairs and considering a maximum insert size of 2kb (bowtie2 -N 0 --no-mixed --no-discordant --minins 30 --maxins 2000). Alignment files were further processed with *samtools* v1.2 and *PicardTools* v1.130 to flag PCR and optical duplicates and remove alignments located in Encode blacklisted regions. Accessible regions were identified using *MACS2*

v2.1.0 without control (macs2 callpeak --gsize hs -p 1e-3). These relaxed peak lists were then processed through the irreproducible discovery rate (IDR) pipeline to generate an optimal and reproducible set of peaks for each time point.

Normalized ATAC-seq and ChIP-seq signal tracks

After verifying the consistency between biological replicates, time points and data type using *deepTools*[53], alignments related to biological replicates for a given assay and a given time point were combined. We then binned the genome in 200bp non-overlapping windows and generated genome-wide read count matrices for each assay independently. These matrices were finally quantile normalized with custom R script and further used to generate genome-wide signal tracts.

Histone modification ChIP-seq and ATAC-seq differential analysis

After assessing library saturation using *preseqR*, alignment and peak data were imported and pre-processed in R using the *DiffBind* package[54]. Briefly, for a given histone modification type, we first defined the global reproducible peak set as the union of each time-specific reproducible peak sets defined previously. We then counted the number of reads mapping inside each of these intervals at each time point and for each replicate. The raw count matrix was then normalized for sequencing depth using a non-linear full quantile normalization as implemented in the *EDASeq* package[55]. To remove sources of unwanted variation and consider batch effects, data were finally corrected with the *RUVSeq*[56] package considering 2 surrogate variables. Differential analyses for count data were performed using *edgeR*[57] considering time and batch in the design matrix, by fitting a negative binomial generalized log-linear model to the read counts for

each peak. Peaks were finally annotated using *ChIPpeakAnno* considering annotations provided by Ensembl v86.

Chromatin state differential analysis

To quantify and define combinatorial chromatin state dynamics in space and time, we analyzed histone modification combinations with the *chromstaR* package[58]. Briefly, after partitioning the genome into 100bp non-overlapping bins and counting the number of reads mapping into each bin at each time point and for each histone modification, this algorithm relies on a univariate Hidden Markov Model (HMM) with two hidden states (unmodified, modified). This HMM is used to fit the parameters of the two-component mixture of zero-inflated negative binomial distribution considered to model read counts for every ChIP-seq experiments. A multivariate HMM is then used to assign every bin in the genome to one of the multivariate components considering $2^{(3 \text{ time points x } 4 \text{ histone modifications})}$ possible states. To limit computational burden and focus on accurate differences, the analysis was run in differential mode with a 100bp resolution (*i.e.* smaller than a single nucleosome), such that every mark is first analyzed separately with all conditions combined while the full combinatorial state dynamics is rebuilt by combining the differential calls obtained for the four marks. We finally filtered out differential calls not overlapping with any histone modification and ATAC-seq reproducible peaks. To properly associate histone modification combinations with biologically meaningful mnemonics, we made an extensive comparison between the binning we obtained in WI38 fibroblasts undergoing RAS-OIS and IMR90 fetal lung fibroblasts chromatin states described in the scope of the Epigenomic Roadmap consortium. To test for association between changes in chromatin states through time

and gene expression modules we ran a correspondence analysis. Briefly, genomic loci experiencing changes in chromatin states through time were first associated to the nearest gene. We then specifically focused on loci associated to genes belonging to any expression module and built a two-way contingency table summarizing the number of transition in states (considering all possible combinations) occurring in each expression module, further used as an input for a correspondence analysis using *FactoMineR*[59]. The significance of association between the two qualitative variables (transition in state and module) was assed using a χ^2 test. Results of the CA were visualized using a row-metric-preserving contribution asymmetric biplot and filtering for the top contributing and well-projected (squared cosine > 0.5) changes in chromatin states.

Motif enrichment analysis in active enhancers

For each time point independently, we defined the set of active enhancers as the overlap between H3K4me1, H3K27ac and ATAC-seq reproducible peaks using *bedtools*[60]. We then ran 3 independent motif enrichment analyses with *homer* v4.9 using default parameters.

Gene expression time series analysis

Normalized log-scaled and filtered expression data related to the quiescence and the OIS time series were further considered for differential analysis with *limma*[61]. To define an RAS-OIS specific transcriptomic signature, we proceeded in three steps, each relying on linear mixed model cubic B-splines, as nonlinear response patterns are commonly encountered in time course biological data. For each probe, and each treatment the expression was modeled as follow:

$$y = \beta_0 + \beta_1 x + \beta_2 x^2 + \beta_3 x^3 + \sum_{k=0}^{K-1} \gamma_k (x - \xi_k)^3 + \varepsilon$$

$$\text{with } (x - \xi_k) = \begin{cases} 1 if x \leq \xi \\ x - \xi if x > \xi \end{cases}$$

where $\beta_0$is the average probe expression over all samples in a given condition, β_{1-3} the model coefficients, K the number of knots, ξ_k the k^{th} knot and ε the error term. First, we defined probes responding over time to RASV12 induction. Second, we considered all together the quiescence and the RAS-OIS time series, as well as the interaction between time and treatment, and defined probes responding to one or the other treatment over time, as well as probes responding differently between the two treatments at any time point. We finally defined the set of probes responding consistently to both treatment and time and removed these probes from the global set of probes responding to RASV12 induction defined at the first step. Moderated *F*-statistics that combine the empirical Bayes moderated *t*-statistics for all contrasts into an overall test of significance for each probe were used to assess the significance of the observed expression changes. At any step of this workflow, *p*-values were corrected for multiple testing using the FDR approach for a stringent significance level of 0.005. For validation purposes, we compressed the RAS-OIS time-series to achieve a volcano plot representation. To this end, we computed the maximal absolute $\log_2$ fold change in expression in the RAS-OIS time series considering T_0 as the reference and selected up and down regulated probes using an absolute $\log_2$ fold change cutoff at 1.2 and a corrected *p*-value cutoff of 0.005. We then build a scatter-plot plotting the $\log_{10}$ significance versus $\log_2$ fold-change on the y and x axes, respectively. Probes responding consistently to both ER: RASV12 induction and quiescence were finally over-plotted.

Correlation and multidimensional analyses

To highlight differences in expression profiles between quiescence and RAS-OIS through time, we used multi-dimensional scaling plot representing leading fold change, which is defined as the root-mean-square average of the log-fold-changes for the genes best distinguishing each pair of samples. To quantify the evolution of transcriptomic variability and noise through time, we looked at the gene expression density distributions for all possible pairs of treated *vs* T_0 transcriptomes. Distributions were estimated using kernel density estimation of all genes' expression in the i^{th} T_0 transcriptome and the j^{th} treated transcriptome. We also computed Pearson's correlation for each of these combinations. The Pearson's correlation between two transcriptomes, X and Y containing n gene expressions, is obtained by $R(X,Y) = \sum_{i=1}^{n}(x_i - \mu_X)\,(y_i - \mu_Y)/(\sigma_X\sigma_Y)$, where x_i and y_i are the i^{th} observation in the vectors X and Y respectively, μ_X and μ_Y the average values of each transcriptome, and σ_X and σ_Y, the corresponding standard deviations.

Transcription factor footprinting

All transcription factor Position-Weight Matrices (PWM) representing eukaryote transcription factors were downloaded from the JASPAR database and used as an input for PIQ[20] to predict transcription factor binding sites from the genome sequence on down-sampled ATAC-seq alignments. For each motif, we retained only binding sites that were within the reproducible ATAC-seq peaks and passed the default purity cut-off (70%). We then computed pairwise PWM similarities based on Pearson's correlation,

and clustered together PWMs sharing more than 90% similarity, defining a set of 310 non-redundant and distinct PWMs. The Pearson's correlation between two PWM P^1 and P^2 of length l was defined as:

$$r(P^1, P^2) = \frac{1}{l} \times \sum_{i=1}^{l} \frac{\sum_{b \in \{A,C,G,T\}} \left(P_{i,b}^1 - 0.25\right)\left(P_{i,b}^2 - 0.25\right)}{\sqrt{\sum_{b \in \{A,C,G,T\}} \left(P_{i,b}^1 - 0.25\right)^2 \times \sum_{b \in \{A,C,G,T\}} \left(P_{i,b}^2 - 0.25\right)^2}}$$

We further combined the bound instances identified with PIQ according to the PWM clustering.

Transcription factor metrics

For each transcription factor, we computed the chromatin-opening index (COI), the motif dependence and the chromatin dependence (CD) following the approach described in [20].

Validation of PIQ predictions through ChIP-seq

To compare PIQ prediction with RELA, JUN and FOSL2 ChIP-seq data, we first used the approach suggested in [20], computing how many of the total ChIP-seq peaks are overlapping with any potential factor motif (since ChIP-Seq peaks can result from co-factor binding, and methods such as digital genomic footprinting are factor agnostic). We then used a more sophisticated approach aiming at correlating the ChIP-seq signal intensity with the bound / unbound status at PWM matches. For a given transcription factor (cJUN, FOSL2 or RELA,), we first considered all the PWM matches located inside ATAC-seq reproducible peaks, we selected all the PWM matches assigned with a purity score > 0.7 (the threshold used to define "bound" instances), and then randomly

selected 3 times more PWM matches assigned to a purity score < 0.7 (considered as "unbound" instances) to obtain a global set containing 25% / 75% of bound / unbound instances for each TF. The selected regions were extended up to 2kb (1kb in each direction, from the middle of the match), and the 2kb intervals were binned in one hundred 20bp windows. We computed the normalized ChIP-seq and ATAC-seq signal inside each bin. The windows were finally ranked according to the summed ChIP-seq signal in the 10 most central bins (200bp). We finally run a set enrichment analysis with the *fgsea* package to assess whether bound / unbound PWM matches were enriched / depleted along this ranking and computed the enrichment score (ES, positive when bound instances are enriched for highest ChIP-seq signals, negative when unbound instances are depleted for highest ChIP-seq signals) and *p*-values which revealed the strength of the correlation. We performed 1,000 permutations to obtain *p*-values.

Transcription factor co-binding

For every cluster of PWM and time-point independently, we first removed all the bound instances identified outside enhancers. The remaining bound instances for all PWM were then combined for every time point using GEM *regulatory module discovery*[25] setting at 500 bp the minimal distance for merging nearby TF bound instances into co-binding regions and at 3 the minimum number of TF bound instances in a co-binding region.

Global pairwise co-binding heatmap. At this step, we obtained a set of contingency matrices M_{mt} of dimension $n_{mt} \times j$ with *i* the number of co-binding regions for the transcriptomic module *m* at the time point *t* and *j* = 310 PWM clusters, for each time point and each transcriptomic module. We then generated module- and time- specific

normalized pairwise co-binding matrices C_{mt} by computing the normalized cross-product of matrices M_{mt} defined as:

$$C_{mt} = \frac{M_{mt} \times {}^{t}M_{mt}}{\sum_t \sum_m \sum_j a_{tmj}} \times 10^6$$

with a_{tmj} the number of bound instances for the PWM clusters *j*, in transcriptomic module *m*, at the time point *t*. To get a global picture of pairwise co-binding, we summed these matrices and tested for each combination of PWM clusters A and B whether the overlap between bound instances for A and B was significant using a hyper-geometric test defined as:

$$p(Q, M, n, k) = \sum_{m=k}^{min\{k,B\}} \frac{\binom{M}{m}\binom{Q-M}{n-m}}{\binom{Q}{m}}$$

where *Q* is the overall number of regions in the universe, *M* is the number of regions bound by A, *n* is the number of regions bound by B, and *k* the total number of regions bound by A and B. The *p*-values were further corrected for multiple testing using the Bonferroni strategy. We finally clustered the co-binding occurrence matrix using Ward's aggregation criterion and projected corresponding corrected *q*-values on this clustering.

Pair-wise co-binding circos plots. To generate the co-binding circos plots, we used the global time- and, module-specific pair-wise normalized co-binding matrix C_{mt} described above, after a logarithmic transformation. For each time-point and module independently, we selected the top 500 interactions based on their occurrence *N*. The images were generated using the *Circos* suite[62].

Identification of TF regulatory modules

We used the data-sets generated using GEM regulatory module discovery described above. We applied a Hierarchical Dirichlet Process topic model which automatically determines the number of topics from the data, with the hyperparameter for the topic Dirichlet distribution set at 0.1 (encoding the assumption that most of the topics contains a few TFs) and the maximum number of iterations set at 2000. The lexicon usage for each time point and each transcriptomic was explored using a multiple factor analysis (MFA) with the R package *FactoMineR,* and lexicons were further selected based on their goodness of representation on the 3 first components (squared cosine > 0.5).

TF properties

With the aim of characterizing the binding properties of each TF, we computed the dynamicity, the total number of bound regions, the fraction of bound regions in enhancers and the fraction of bound regions before stimulation.

Dynamicity. We quantified the dynamicity of a TF accordingly to the following expression:

$$d(A) = \frac{\sum \frac{n_t(A)}{TR_t}}{\sum \frac{t_t(A)}{TR_t}}$$

where $d(A)$ is the dynamicity of TF A; $n_t(A)$ is the number of regions bound by A for the first time at time point t ; $t_t(A)$ is the number of regions bound by A at time point t and TR_t is the number of regions bound by any TF in time point t. The factor TR_t was added to the expression to account for differences in the number of reads sequenced by the ATAC-seq protocol and normalizes the number of regions bound by TF A based on the

number of bound regions detected at its corresponding time point. Notice that, if all samples have the same amount of TF binding events, this expression is reduced to the quotient of the sum of the regions first bound at each time point by the sum of all regions bound by the TF at each time point. By using this definition, the function $d(A)$ maps the activity of a TF to the interval $\left[\frac{1}{N_t}, 1\right]$, where N_t is the number of time points in the timecourse and is higher as the TF binds to previously not bound regions or leaves already bound regions. In the case of a TF that, for every time point, leaves all its previous bound regions and binds to only regions not previously bound, the numerator will be identical to the denominator, leading to $d(A) = 1$. Alternatively, if a TF remains on the same regions it has bound at $t = 0$, then $\sum n_t = n_0$ and $\sum t_t = N_t * n_0$, resulting in $d(A) = \frac{1}{N_t}$. One can observe that, if the same region is bound by TF A in different time points, it will contribute once to the numerator of the expression, while it will contribute to the denominator once for each time point it has been bound to.

Total number of bound regions. The number of bound regions was calculated by the following the expression:

$$R(A) = \frac{\sum \frac{n_t(A)}{TR_t} \times \sum TR_t}{N_t}$$

where $R(A)$ is the normalized number of bound regions by TF A during the timecourse and $n_t(A)$, TR_t and N_t are defined as above. The first factor is a normalized sum of the regions bound by TF A, counting each region only once. The second factor scales the result by the mean of the number of regions bound by all TFs on each day.

TF percentage of binding at enhancers. The ratio of binding at enhancers, relative to all cis regulatory regions, was assessed by:

$$P_E(A) = \frac{R_E(A)}{R_E(A) + R_P(A)}$$

where $P_E(A)$ is the percentage of bound regions in enhancers for TF A, $R_E(A)$ is the number of regions bound by TF A marked as enhancers and $R_P(A)$ is the number of regions bound by TF A marked as promoters.

TF prestimulation binding. For each TF, we computed the ratio of regions bound at T_0, relative to the number of regions bound during the whole timecourse. We used the following definition for the prestimulation binding factor for each TF:

$$p(A) = \frac{\frac{n_{D0}(A)}{TR_{D0}}}{\sum \frac{n_t(A)}{TR_t}}$$

where $p(A)$ corresponds to the prestimulation binding of TF A and $n_t(A)$ and TR_t are defined as above. The numerator of this expression corresponds to the normalized number of regions bound by TF A at t = T_0, while the denominator is the normalized number of regions bound by TF A during the whole timecourse. Notice the denominator also corresponds to factor $R(A)$ before scaling.

Hierarchical transcription factor network

In order to assess the TF chromatin binding hierarchy, i.e. TFs required for the binding of a specific TF, we generated a network for each gene module depicting the precedence of TF chromatin binding. The algorithms mentioned were implemented in R and all networks were visualized in CytoScape[63].

Computing precedence relationships. The edges in the generated networks represent the precedence relationship of TFs: an oriented edge from TF A to TF B, represented as

(A, B), means that A was present in at least 30 % of the cis-regulatory regions bound by B at the same instant or before[29]. To account for the difference in the number of reads sequenced for each sample in the ATAC-seq, we normalized the number of regions bound based on the first day they appeared. The weight of an edge from A to B is given by:

$$w_{A \to B} = \frac{\sum \frac{R_t(A,B)}{R_t}}{\sum \frac{R_t(B)}{R_t}}$$

where $R_t(B)$ stands for the number of regions first bound by TF B at time point t ; $R_t(A,B)$, for the number of regions first bound by TF B at time point t that were bound by TF A at time point t or before; and R_t represents the total number of regions bound by any TF in time point t. In order to handle the networks, we used the *igraph* package.

Network simplification. Aiming to analyze the hierarchical relationship of TFs and simplify the interpretation of the network, we performed two operations over each gene module network: Vertex Sort and transitive reduction (TR)[64]. Briefly, the vertex sort algorithm assigns two parameters for each node in the network: the distance, in edges, between the node and the bottom of the network; and the distance between the node and the top of the network. Combined, those parameters allow for the topological ordering of the network, which consists in listing its nodes such that nodes at the top precede downstream nodes. We then defined the 'top layer' as the set of nodes with lowest distance to the top of the network, i.e., nodes that have no incoming edges or nodes that assemble a strongly connected component (SCC) with all upstream nodes. Analogously, the 'bottom layer' was defined as the set of nodes with lowest distance to the bottom of the network, i.e., nodes with no outgoing edges or that form a SCC with all

downstream nodes. The ’core layer’ comprises nodes that link top layer and bottom layer. Nodes in the core layer that are exactly one edge from both top and bottom layers constitute the ’single-level core layer’, while nodes that link top and bottom through paths composed of more than one edge form the ’multi-level core layer’. The result of this procedure for each gene module can be seen in Figure 6a and supplementary data. The TR, in turn, simplifies the network visualization by generating the network with the smallest number of edges that keeps the reachability of the original network.

Network validation. We validated our approach by comparing the network produced when applying our method to the ChIP-seq data produced by [29]. Transcription factor ChIP-seq peak files were retrieved from Gene Expression Omnibus (GSE36099, 23 TFs, and 4 time points; note that RUNX1 and ATF4 were discarded from the analysis since one and three time points, respectively, were missing on GEO for those TFs) and preprocessed as previously described to generated time resolved co-binding matrices, further used as an input for our networking algorithm. We computed the precedence relationships among TFs and generated the TF binding hierarchy networks for visualization. We produced TF hierarchy network using two metrics [29]: sensitivity and specificity. Sensitivity is calculated as the ratio of edges described in this study over the edge number sum for both networks. Specificity is defined as the ratio of the number of edges that were described to not exist in the network produced by our software over the number of edges described to not occur in any of both studies.

Proportion of incoming edges based on the classification of the TF source node. Aiming to assess the hierarchy of TFs accordingly to their chromatin dependence and chromatin-opening index, we computed the number of edges connecting the sets of all TFs with a given classification for each gene module. We then divided those values by

the number of edges that target TFs with a specific classification. Hence, the proportion of incoming edges based on TF classification is given by:

$$P_{C1\to C2} = \frac{|W_{C1\to C2}|}{\sum|W_{K\to C2}|}$$

where $P_{C1\to C2}$ is the proportion of edges from nodes with classification C1 to nodes with classification C2; $W_{C1\to C2}$ is the set of edges from nodes with classification C1 to nodes with classification C2; K can represent either pioneer, settler or migrant and $|\cdot|$ means the cardinality of a set, *i.e.* the number of elements it contains.

We assessed the classification precedence significance for TF interaction with a hypergeometric test. We consider the sample space as all possible oriented edges in a network with the same number of nodes for each classification as the hierarchy network for a given transcriptional module. Formally:

$$p_{C1\to C2}(E, E_{C1\to C2}, W, W_{C1\to C2}) = \sum_{x=|W_{C1\to C2}|}^{|W|} \frac{\binom{E_{C1\to C2}}{x}\binom{E-E_{C1\to C2}}{|W|-x}}{\binom{E}{|W|}}$$

Where $E$is the number of edges on the sample space network, i.e., a fully connected network with the same number of nodes as the TF hierarchy network for a given transcriptional module (excluding self-loops),$E_{C1\to C2}$is the number of edges from TFs with classification $C1$to TFs with classification $C2$ in the sample space network, $|W|$is the number of edges on the TF hierarchy network for a given transcriptional module and $|W_{C1\to C2}|$is the number of edges in the same network connecting TFs with classification $C1$to TFs with classification $C2$.

Network visualization. In order to visualize the network, we exported the adjacency matrices in the R environment to CytoScape using the CyREST API[65]. The networks' layout and style were automated with the help of packages RCy3[66] and RJSONIO.

Network mining

With the purpose of identifying key TFs in the transition to the senescent phenotype, we analyzed the TF binding characteristics with their relative location in the chromatin binding hierarchy networks for each gene module. The figures illustrating this analysis were generated with the help of the ggplot2 R package.

TF classification. For each network relative to a transcriptional gene module, the number of TF classified as either pioneer, settler or migrant was calculated for each layer, with the subdivision of the core layer as 'multi-level' and 'single-level' (see "*Network simplification*"). The overrepresentation of TFs with a specific classification in a given layer was evaluated by using a hypergeometric test. We calculated the *p*-value given by:

$$p(K, N, n, k) = \sum \frac{\binom{K}{x}\binom{N-K}{n-x}}{\binom{N}{n}}$$

where *K* is the number of TFs with a certain classification in the whole network, *N* is the number of TFs in the network; *n* is the number of TFs that belong to a specific layer and *k* is the number of TFs that belong to the same layer and have the referred classification. The *p*-values were corrected for multiple testing with FDR and a corrected $p = 0.05$ was considered an indicative of enrichment for that specific classification in the corresponding layer.

TF dynamicity. For each network, relative to a transcriptional gene module, we compared the distribution of the dynamicity of TFs belonging to a certain layer with the distribution of the dynamicity of TFs belonging to the rest of the network. We used the dynamicity index defined previously for each TF, considering only the regions marked as

enhancers belonging only to the gene module relative to the network. For each layer in the network, we applied the Kolmogorov-Smirnov test to compare the TF dynamicity distribution for the chosen layer with the dynamicity distribution relative to the TFs belonging to three other layers in the respective network. To account for multiple hypothesis testing, we also performed an FDR correction, considering values of $p = 0.05$ as an indicative of statistical significance.

TF number of binding regions. We performed the same analysis as described in the previous section ("*TF dynamicity*") for the number of bound regions defined in section "*Total number of bound regions*", instead of the dynamicity index.

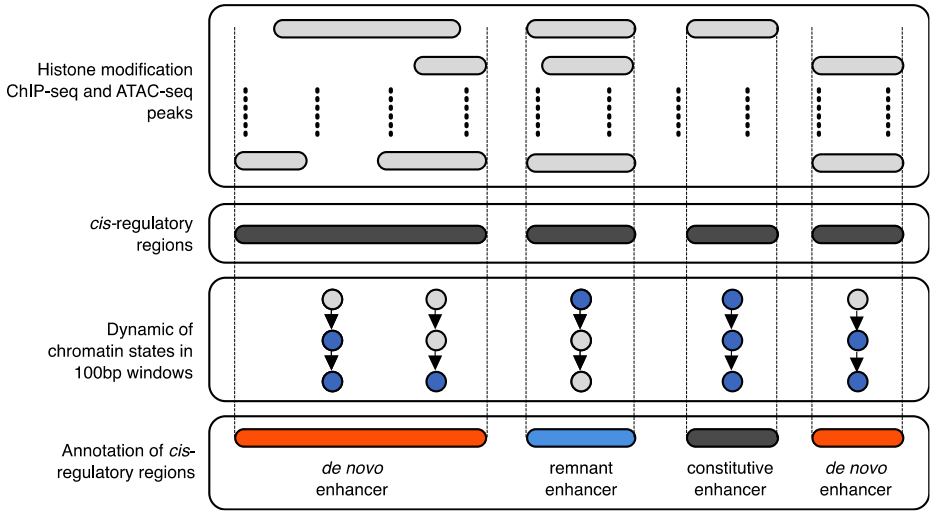
TF binding characteristics and transcriptional modules. In order to characterize the binding activity of each TF for the different gene modules, we ranked them accordingly to their dynamicity and their number of bound regions. Both parameters for each gene module are shown inExtended Data 6e , which was generated with the *ComplexHeatmap*[67] and *circlize*[68] R packages. We used the mean of the ratio dynamicity - number of bound regions to order the TFs. We assessed the significance of pioneer (respectively, migrant) TF enrichment at the top (respectively, bottom) of the ranked clustered list by employing a set enrichment analysis implemented in the package fgsea.

TF chromatin binding hierarchy networks overlap. To analyze the similarity between the networks for different transcriptional gene modules, we generated a 7-set Euler diagram, where each set contains the edges present in the TF hierarchy network relative to a gene module. Edges in two different networks are considered equal if they link nodes corresponding to the same TFs in their respective networks. We used the package Vennerable to compute the intersections of all possible network combinations and to

create the Euler diagram in Extended Data 7a-d. In this figure, the area of each region is proportional to the number of edges shared by the networks corresponding to the sets that contain the referred region and was calculated using the Chow-Ruskey algorithm[69]. A Euler diagram is similar to a Venn diagram, with the difference that the area of a region representing a set is proportional to the number of elements in the set.

Analysis of *de novo* and remnant enhancers

To track combinatorial chromatin state dynamics in space and time, we integrated histone modification ChIP-seq signals at a sub-nucleosomal resolution considering non-overlapping 100bp windows genome-wide using chromstaR (see above), which converts quantitative ChIP-seq data to qualitative chromatin states. For subsequent analysis, since these 100bp windows can be either isolated or organized in stretches experiencing consistent changes in states, we summarized the information at a higher level, and linked them with the histone modification peaks identified using the more classical ChIP-seq and ATAC-seq peak-calling approach. Briefly, after merging all the peaks identified for all the time-points, for all the histone modification and for the ATAC-seq data sets defining *cis*-regulatory regions, we determined the overlap between "poised ", "*de novo*", "remnant" or "constitutive enhancers"-flagged 100bp windows. When an overlap was found, the entire *cis*-regulatory regions were annotated according to the 100bp window it is overlapping with. This operation rendered a list of annotated *cis*-regulatory regions with *de novo*, constitutive, poised or remnant enhancer elements. We finally considered the center +/- 10kb of these elements.


Histone modification ChIP-seq and ATAC-seq peaks
cis-regulatory regions
Dynamic of chromatin states in 100bp windows
Annotation of cis-regulatory regions
de novo enhancer
remnant enhancer
constitutive enhancer
de novo enhancer

References

49. Langfelder, P. & Horvath, S. WGCNA: an R package for weighted correlation network analysis. *BMC Bioinformatics* **9**, 559 (2008).

50. Landt, S.G. *et al.* ChIP-seq guidelines and practices of the ENCODE and modENCODE consortia. *Genome Res* **22**, 1813-1831 (2012).

51. Bolger, A.M., Lohse, M. & Usadel, B. Trimmomatic: a flexible trimmer for Illumina sequence data. *Bioinformatics* **30**, 2114-2120 (2014).

52. Martin, M. Cutadapt removes adapter sequences from high-throughput sequencing reads. *2011* **17**, 3 (2011).

53. Ramirez, F. *et al.* deepTools2: a next generation web server for deep-sequencing data analysis. *Nucleic Acids Res* **44**, W160-165 (2016).

54. Ross-Innes, C.S. *et al.* Differential oestrogen receptor binding is associated with clinical outcome in breast cancer. *Nature* **481**, 389-393 (2012).

55. Risso, D., Schwartz, K., Sherlock, G. & Dudoit, S. GC-content normalization for RNA-Seq data. *BMC Bioinformatics* **12**, 480 (2011).

56. Risso, D., Ngai, J., Speed, T.P. & Dudoit, S. Normalization of RNA-seq data using factor analysis of control genes or samples. *Nat Biotechnol* **32**, 896-902 (2014).

57. Robinson, M.D., McCarthy, D.J. & Smyth, G.K. edgeR: a Bioconductor package for differential expression analysis of digital gene expression data. *Bioinformatics* **26**, 139-140 (2010).

58. Taudt, A., Nguyen, M. A., Heinig, M., Johannes, F. & Colome-Tatche, M. chromstaR: Tracking combinatorial chromatin state dynamics in space and time. *bioRxiv* (2016).

59. Lê, S., Josse, J. & Husson, F. FactoMineR: An R Package for Multivariate Analysis. *2008* **25**, 18 (2008).

60. Quinlan, A.R. & Hall, I.M. BEDTools: a flexible suite of utilities for comparing genomic features. *Bioinformatics* **26**, 841-842 (2010).

61. Ritchie, M.E. *et al.* limma powers differential expression analyses for RNA-sequencing and microarray studies. *Nucleic Acids Res* **43**, e47 (2015).

62. Krzywinski, M. *et al.* Circos: an information aesthetic for comparative genomics. *Genome Res* **19**, 1639-1645 (2009).

63. Shannon, P. *et al.* Cytoscape: a software environment for integrated models of biomolecular interaction networks. *Genome Res* **13**, 2498-2504 (2003).

64. Aho, A.V., Garey, M.R. & Ullman, J.D. The Transitive Reduction of a Directed Graph. *SIAM Journal on Computing* **1**, 131-137 (1972).

65. Ono, K., Muetze, T., Kolishovski, G., Shannon, P. & Demchak, B. CyREST: Turbocharging Cytoscape Access for External Tools via a RESTful API. *F1000Res* **4**, 478 (2015).

66. Shannon, P.T., Grimes, M., Kutlu, B., Bot, J.J. & Galas, D.J. RCytoscape: tools for exploratory network analysis. *BMC Bioinformatics* **14**, 217 (2013).

67. Gu, Z., Eils, R. & Schlesner, M. Complex heatmaps reveal patterns and correlations in multidimensional genomic data. *Bioinformatics* **32**, 2847-2849 (2016).

68. Gu, Z., Gu, L., Eils, R., Schlesner, M. & Brors, B. circlize Implements and enhances circular visualization in R. *Bioinformatics* **30**, 2811-2812 (2014).

69. Chow, S. & Ruskey, F. 466-477 (Springer Berlin Heidelberg, Berlin, Heidelberg; 2004).