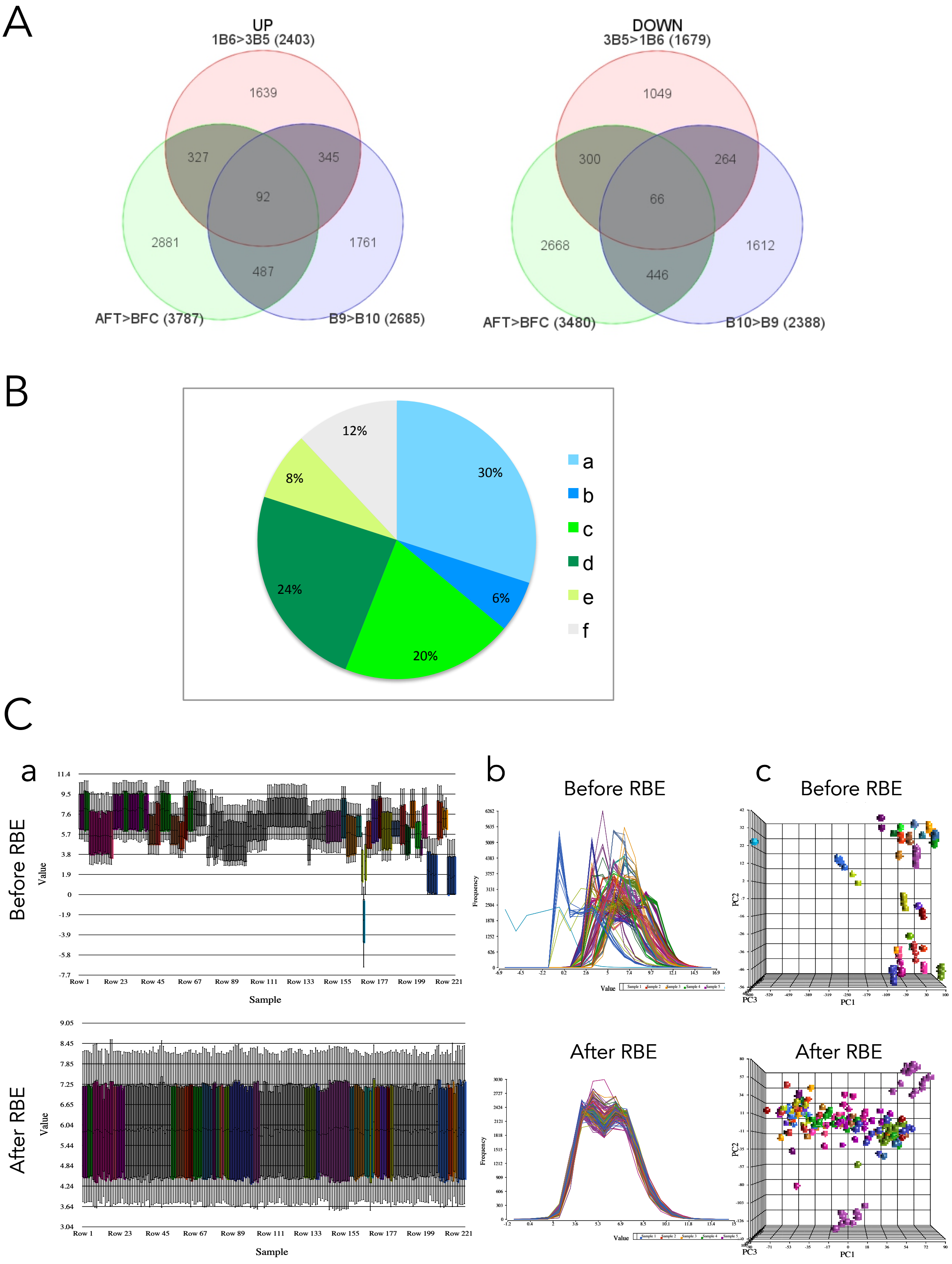
iScience, Volume *23*

Supplemental Information

Inferring Gene Networks in Bone Marrow Hematopoietic Stem Cell-Supporting Stromal Niche Populations

Christophe Desterke, Laurence Petit, Nadir Sella, Nathalie Chevallier, Vincent Cabeli, Laura Coquelin, Charles Durand, Robert A.J. Oostendorp, Hervé Isambert, Thierry Jaffredo, and Pierre Charbord



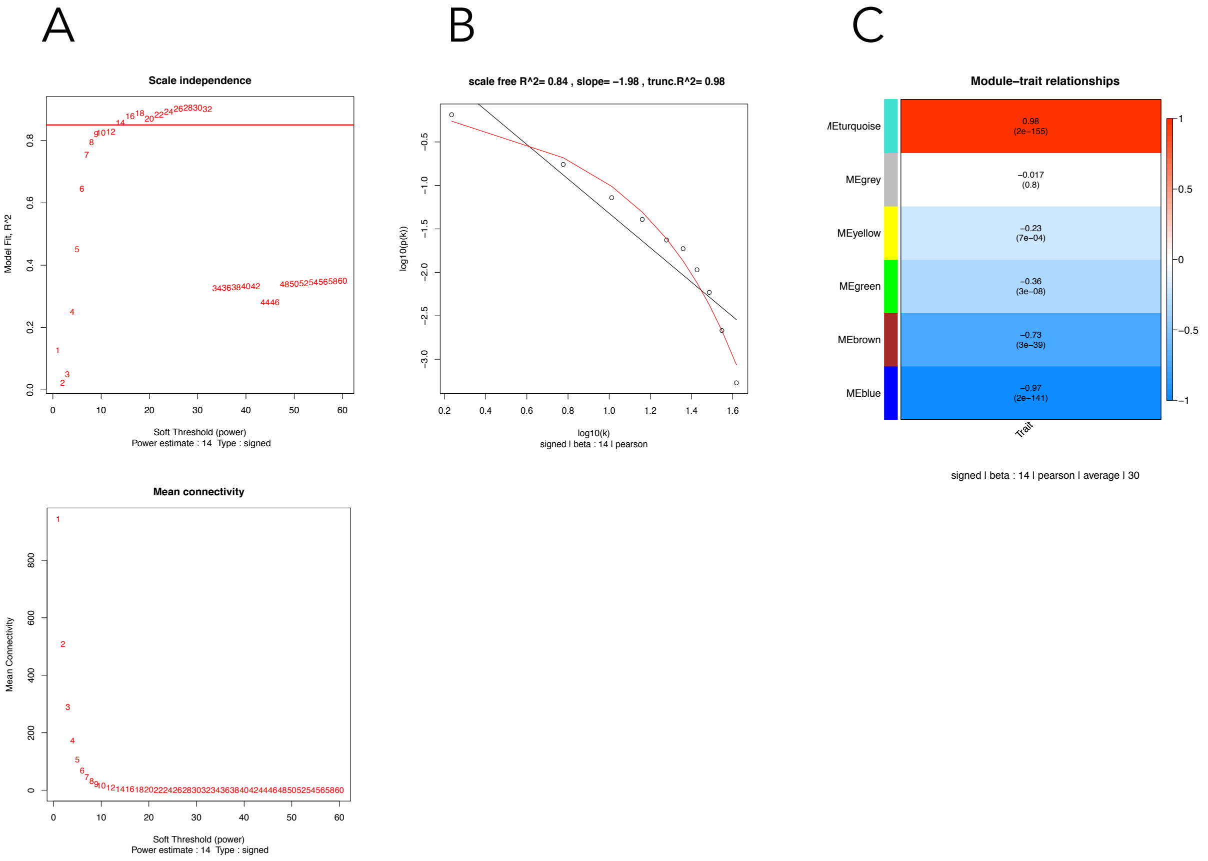
Supplemental Figure 1 (related to Figure 1): Study flowchart and cell cluster identification

A: Venn diagrams leading to the identification of the Set1 gene set. The gene sets obtained after ANOVA comparing transcriptomes of supportive vs. less or non supportive stromal lines from each developmental site were intersected. The resulting Venn diagrams for the genes UP- and DOWN-regulated in the supportive lines are shown on the left and right panels, respectively. Set 1 corresponds to the shaded areas. Abbreviations for the cell lines: 1B6: supportive AGM UG26.1B6;

3B5: less supportive AGM UG26.3B5; AFT: supportive FL AFT024; BFC: non-supportive FL BFC012; B9: supportive BM BMC9; B10: less supportive BM BMC10.

B: The relative relevance of the criteria evidencing the supportive capacity of stromal cells in the ensemble of articles corresponding to the datasets analyzed in this work. a) generation of hematopoietic colonies in co-culture of HSPCs with stromal cells, b) hematopoietic differentiation of embryonic stem cells co-cultured with stromal cells, c) *in vivo* co-localization of stromal cells expressing specific markers with HSCs, d) lineage reconstitution *in vivo* by HSCs co-cultured with stromal cells for variable time spans, e) reduction of the HSC pool *in vivo* after depletion of the stromal cell population, and f) characteristic phenotype.

C: Effect of Removal Batch Effect on Quality Control parameters using the matrix consisting in the 224 observations (samples) and the 15056 variables (entire transcriptome after dataset merging). Mean value and dispersion of the expression level of all genes in a sample (a), distribution of the gene expression levels (b), and representation of the samples in the PCA space (c).

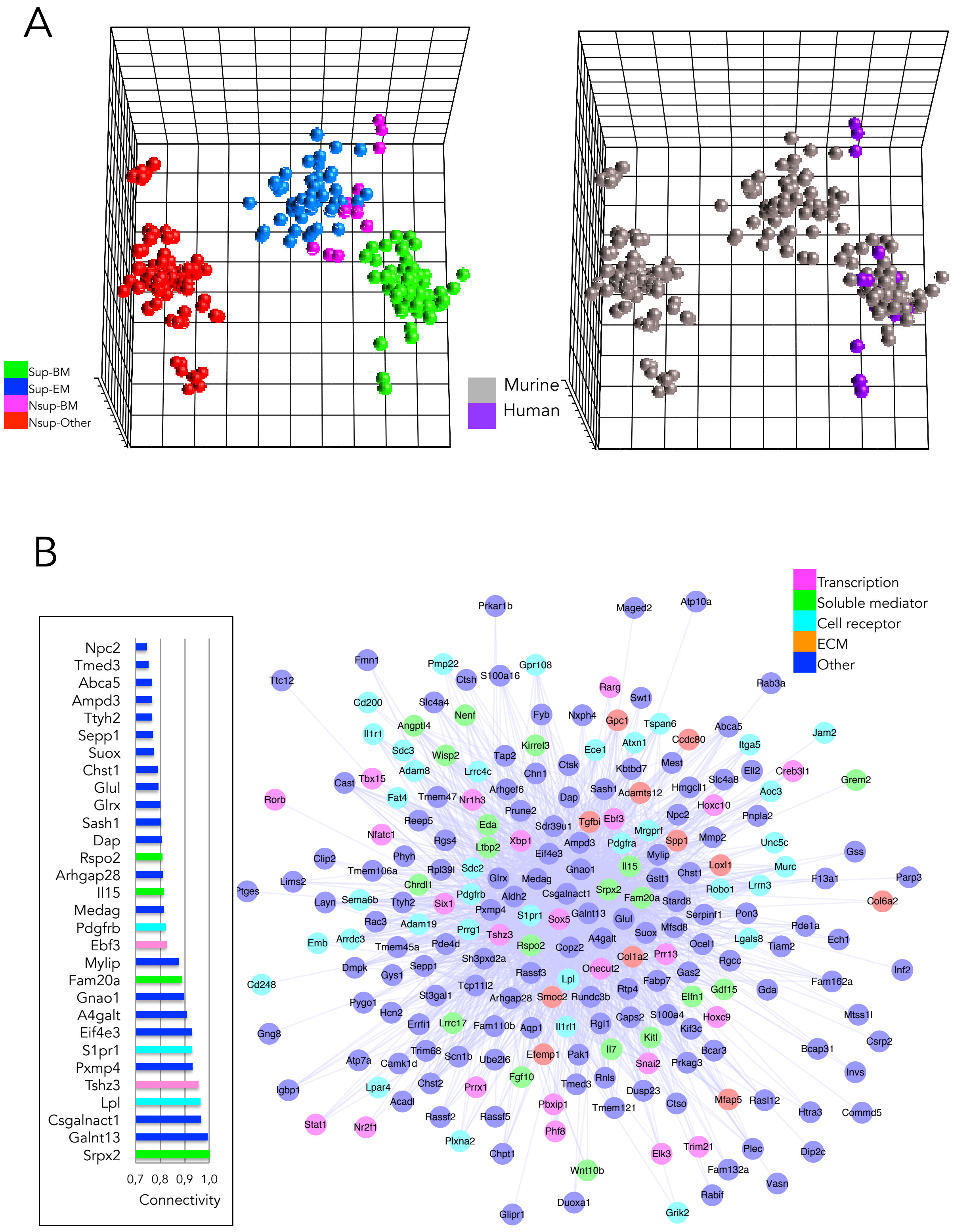


Supplemental Figure 2 (related to Figure 2): WGCNA of the 1869 Set 1 genes

*A: Choice of the power ß for the weighted adjacencies***.** Upper panel: scale free index fit (R^2) vs. ß power; ß=14 is chosen as the value where saturation is reached and above 0,8. Lower panel: connectivity distribution according to ß power.

B: Verification of the scale-free topology index for the chosen ß power value. Model fit for a power law (black) and an exponentially truncated power law (red).

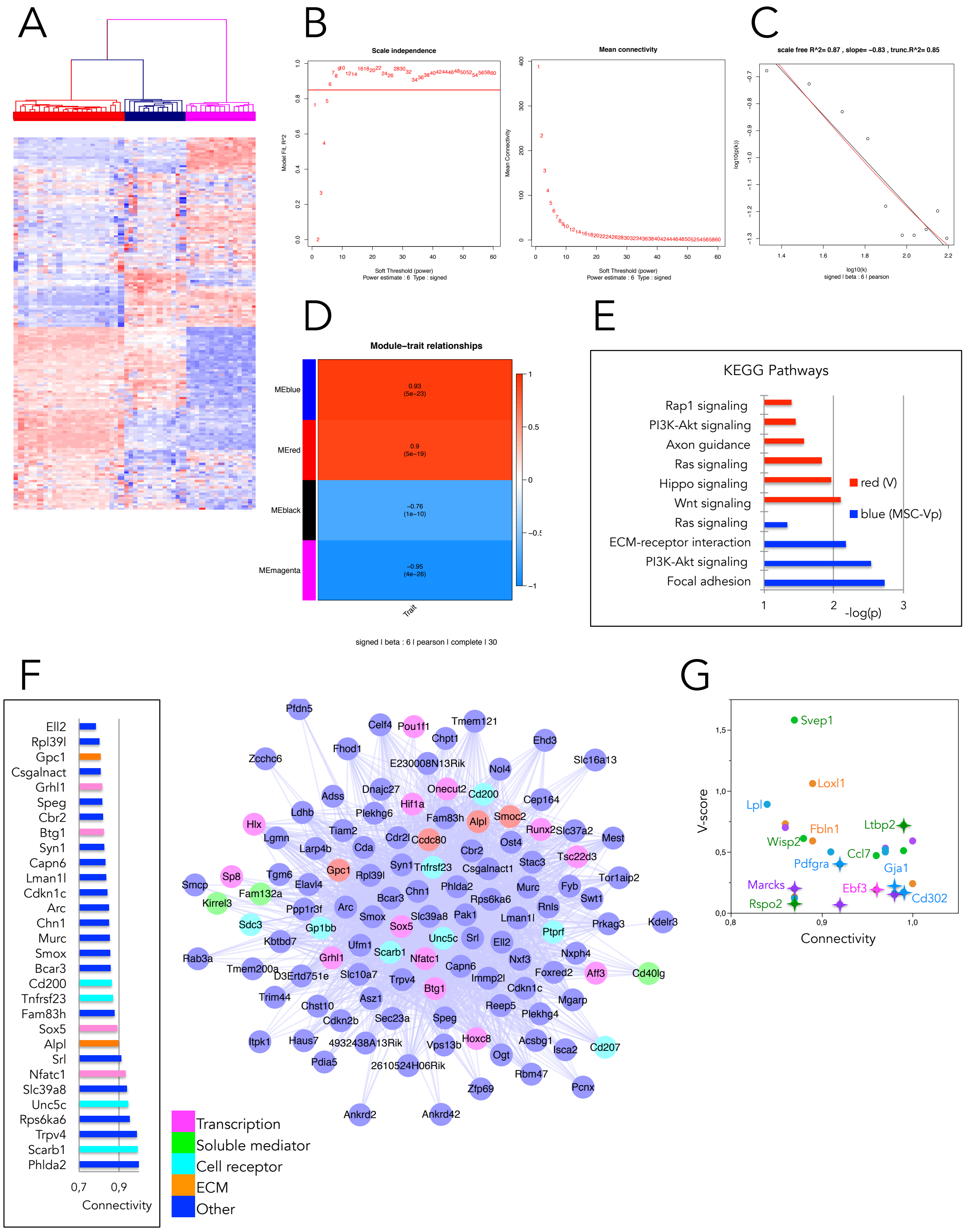
C: Module trait relationships. Pearson's correlation coefficient and corresponding p-value are given for each of the identified modules.



Supplemental Figure 3 (related to Figure 2): A similar gene network characterizes the BM stromal supportive capacity after inclusion of human samples

A: PCA using 248 samples as observations and 1619 genes as variables. The 3D score plot (PC1, PC1, PC3) is shown on the left panel (same sample colors as in Figure 1B). The right panel highlights the samples according to species (human vs. murine samples).

B: Cytoscape connectivity plot (right) and intra-modular connectivity bar plot of the Top-30 genes (left) for the turquoise module corresponding to the Sup-BM cluster.



Supplemental Figure 4 (related to Figure 3): The supportive capacity is implemented in four BM niche cell populations

A: Hierarchical clustering corresponding to O, V and MSC samples. Euclidian distances and Ward's linkage were used.

*B: Choice of the power ß for the weighted adjacencies***.** Left panel: scale free index fit (R^2) vs. ß power; ß=6 is chosen as the value where saturation is reached and above 0,8. Right panel:

connectivity distribution according to ß power.

C: Verification of the scale-free topology index for the chosen ß power value. Model fit for a power law (black) and an exponentially truncated power law (red).

D: Module trait relationships. Pearson's correlation coefficient and corresponding p-value are given for each of the identified modules.

E: Significant KEGG pathways corresponding to genes in V (red) and MSC-Vp (blue) populations.

F: Connectivity plots of genes belonging the O module. Right panel: Intra-modular connectivity bar plot of the Top-30 genes. Left panel: Cytoscape connectivity plot of the 115 most connected genes; genes were selected for representation above a threshold weight ≥ 0,3110.

G: Gene connectivity vs. PAM score scatter plot for the V and MSC-Vp modules. The genes belonging to the V module are shown as stars.

Transparent Methods

Bioinformatics

Data sets: The different samples used in this study are described in Supplemental Table 1. The retrieved transcriptomes were studied using Affymetrix microarrays (different platforms: Mouse 430 2.0, Mouse Gene 1.0 ST, Mouse Gene 2.0 ST, Mouse Exon 1.0 ST) or Illumina RNAseq. The data sets were merged in a single matrix. A normalization step was necessary to get rid of the covariate series (the different experiments), using the Partek 'Remove batch effect' (RBE) algorithm. Using the matrix consisting in the 224 observations (samples) and the 15056 variables (entire transcriptome after datasets merging) the results of this RBE are shown on three QC parameters: the mean value and dispersion of the expression level of all genes in a sample (A), the distribution of the gene expression levels (B), and the representation of the samples in the PCA space is shown on Supplemental Fig 1A. The RBE resulted in the alignment of mean value and dispersion of the expression level of all genes in the different samples, the normalization of the gene distributions, and the splitting of the different samples belonging to one dataset to the whole space.

PCA: It was performed on matrices indicated in the text using Partek and factominer R-package. P-value of group discrimination was evaluated by correlation on the first principal component axis (Le et al., 2008).

Discriminant analysis: To discriminate the best classifiers for the Sup-BM cluster, the learning machine algorithm 'Prediction Analysis for Microarrays' (PAMr) was applied using the 4 cell clusters as defined (Tibshirani et al., 2002). Positive score probes obtained by leave-one-out learning with cross validation for the Sup-BM cluster were retained and applied to a random forest algorithm (Breiman, 2001). Before starting the Random Forest analysis the dataset was split randomly in training set and validation set comprising 75% and 25% of the samples, respectively. The Random Forest "mtry" parameter was tuned with Caret R package on the training set in order to achieve the learning step with a minimum error of bagging. The Random Forest model was built on training, and supervised by implementing the Sup-BM cluster against all others pooled in one class. Subsequently, validation of the Random Forest model was performed with accurate prediction on the validation set. *Rspo2* expression probe, found highly predictive for the Sup-BM cluster, was used as quantitative predictor in LIMMA algorithm (Ritchie et al., 2015) in order to find correlated gene expression profile. Conjointly, the supervised Significance Analysis for Microarray (SAM) algorithm (Tusher et al., 2001) was applied, comparing the Sup-BM vs. other clusters: the threshold was fixed for a positive value of fold change superior to 2 and a minimal false discovery rate of 0,05. To perform unsupervised classification with Euclidean distances we retained the genes in the intersection of two gene sets, one corresponding to genes whose expression was correlated to that of *Rspo2*, and the other corresponding to genes with positive fold change in Sup-BM cluster according to SAM. Microarray expression heat-plot was drawn with made4 R-package (Culhane et al., 2005). Statistical significance tested for each gene was performed with Fisher one-way ANOVA followed by post-Hoc Tukey test. The PAMr algorithm was also applied to identify the genes that were predominantly expressed in each population.

Weighted Gene Correlation Network Analysis (WGCNA): It was performed according to the algorithm developed by Horvath and co-workers (Horvath, 2011; Langfelder and Horvath, 2008), using different gene sets as indicated in the text. The chosen values for essential input parameters for each gene set were chosen as giving in each case the simplest module definition defined by clustering using the Topological Overlap Matrix (TOM) and merged dynamic tree cut method. They were: 1) the power ß obtained from the scale-free fit model curve ($r^2 > 0,8$); 2) the correlation, in all cases Pearson's; 3) the adjacency, in all cases signed; 4) the linkage, average or complete; 5) the minimum module size of 30.

For the Cytoscape connectivity plot edges were selected above a given weight (intra-modular topological overlap measure) threshold. Graph layout was always force-directed.

Gene Set Enrichment Analysis (GSEA): GSEA aims at identifying the significant genes within a gene expression set contrasting two phenotypes. The expression set is compared to pre-established datasets corresponding to divers biological process. Datasets are retained if the distribution of their genes is shifted toward either phenotype. The statistical relevance of the shift is established by random permutations of the phenotypes (Liberzon et al., 2015; Subramanian et al., 2005). In this work we used the Sup-BM gene set as reference dataset.

Gene Ontology (GO): Selected genes from the gene sets were processed using DAVID (database for annotation, visualization and integrated discovery) (Huang da et al., 2009a, b).

Multivariate information-based inductive causation (miic) analysis

We used the miic algorithm which combines constraint-based and information-theoretic frameworks to reconstruct causal and non-causal networks from large scale datasets (Verny et al., 2017; Sella et al., 2018). Miic network predictions belong to the broad class of 'ancestral graphs' that include undirected, directed as well as bi-directed edges originating from latent common causes unobserved in the available dataset and represented by dashed edges in the networks of Figure 4C and Figure 4D. In brief, starting from a fully connected graph, miic iteratively removes dispensable edges, by uncovering significant information contributions from indirect paths. This amounts to progressively uncover the best supported conditional independencies (*i.e.* $I(X;Y/(Ai)) = 0$ implying no *XY* link in the underlying network) by iteratively 'taking off' the most significant indirect contributions of positive conditional 3-point information, $I(X;Y;A_k|_{k-1}) > 0$, from every 2-point (mutual) information, $I(X;Y)$, as,

$$I(X;Y|\{A_i\}_n) = I(X;Y) - I(X;Y;A_1) - I(X;Y;A_2|A_1) - \dots - I(X;Y;A_n|\{A_i\}_{n-1})$$

Miic also provides an edge-specific confidence assessment of retained edges, which are oriented based on the signature of causality in observational data. This has long been known to be associated to 'v-structures', $X \rightarrow Z \leftarrow Y$, between two mutually (or conditionally) independent variables, *X* and *Y*, connected to a third variable *Z*. Indeed, this entails the orientations of the v-structure as the edges *XZ* and *YZ* cannot be undirected, nor *Z* be a cause of *X* and *Y*, since these alternative graphical models would imply correlations in contradiction with the independence between *X* and *Y*.

Single cell analysis

Single cell analyses were performed in R software environment version 3.5.3 with package Seurat_2.3.4 that uses dependencies of packages Matrix_1.2-17, cowplot_0.9.4 and ggplot2_3.1.1. Digital matrix was built with LEPR principal cluster from GSE108891 stromal data comprising 1712 cells. Data were scaled and normalized on Unique Molecular Identifiers and gene counts before performing dimensional reduction with t-SNE analysis on the first forty principal axes. Expression of selected markers was plotted on the dimension 1 vs. dimension 2 t-SNE map.

Hematopoietic cell isolation and culture

Mice: Adult C57Bl/6 female mice (10-16 weeks old) were provided from Janvier Labs and maintained in the animal facility of the Laboratory of Developmental Biology according to institutional guidelines (experiment approved by the 'Charles. Darwin' ethical committee of the Sorbonne University). Femurs and tibias were flushed with DMEM (Gibco) supplemented with 10% FCS (Eurobio).

Cell sorting: BM lineage-negative cells were first isolated by depletion of hematopoietic lineage marker (Ter119, Gr1, B220, CD4, CD8, CD11b)-expressing cells using MACS columns (Miltenyi Biotec). Cells were then stained with PE-conjugated anti-Sca-1 and APC-conjugated anti-CD117 (c-kit). LSK cells were isolated using a FACS AriaIII (BD) cell sorter.

Colony-forming units: 250 LSK sorted cells were plated in 1 mL methyl-cellulose in IMDM (Gibco) supplemented with 30% FCS (Eurobio), 1% crystallized BSA (Sigma), and 10^{-4}M ß-mercaptoethanol (Sigma), in the presence of 20 ng/mL SCF, 10 ng/mL IL-3, 2 U/mL EPO and 20 ng/mL TPO (all from PromoKine). Tested RSPO2 (R&D) concentrations and/or 100 ng/mL WNT3A (Peprotech) were added. Samples were plated in duplicate 35-mm Petri dishes and incubated for 7 days at 37°C.

Culture of LSK cells in liquid medium: 200 LSK sorted cells were plated in U-bottom 96 well-plates in 100 µL X-Vivo 15 medium (Lonza) supplemented with 10% FCS, 5.10^{-5} M ß-mercaptoethanol (Sigma) and 10^{-4} M Methyl-ß-cyclodextrin (Sigma), in the presence of 20 ng/mL SCF, 20 ng/mL RSPO2and/or 100 ng/mL WNT3A. Five wells were seeded per condition and incubated for 12 days at 37°C, cells in each well were then counted using a Mallassez hemocytometer.

qRT-PCR: RNA from 40000 LSK sorted cells or incubated for 24h in conditions described above was isolated using RNeasy Micro Plus Kit (Qiagen). Reverse transcription was performed using iScript cDNA synthesis kit (Biorad). Relative quantitative PCR was performed using the SsoAdvanced Universal SYBR Green Supermix (Biorad) with gene-specific PCR primers on a PikoReal instrument (Thermo Scientific). After one step at 9°C for 3 minutes the samples were cycled 50 times (denaturation at 95°C for 25 seconds, annealing at 60°C for 30 seconds, and extension at 72°C for 20 seconds). Cq values were measured using the PikoReal software. Expression level as $2^{-\Delta Cq}$, ΔCq = Cq(gene of interest) - Cq(housekeeping gene), housekeeping gene being *Gapdh*. The following primer sequences (5' to 3') were used:
Gapdh fw : ATGGTGAAGGTCGGTGTGAA
Gapdh rv : AATGAAGGGGTCGTTGATGG
Axin2 fw: TGACTCTCCTTCCAGATCCCA
Axin2 rv: TGCCCACACTAGGCTGACA
Bmp4 fw: TTCCTGGTAACCGAATGCTGA
Bmp4 rv: CCTGAATCTCGGCGACTTTTT
Lgr6 fw: GAGGACGGCATCATGCTGTC
Lgr6 rv: GCTCCGTGAGGTTGTTCATACT
Tcf7 fw: AGCTTTCTCCACTCTACGAACA
Tcf7 rv: AATCCAGAGAGATCGGGGGTC

Immunohistochemistry

Femurs were incubated for 5h in formol before incubation in sucrose 30% at 4°C overnight. Bones embedded in OCT were cryosectioned on Superfrost+ slides and immediately fixed in acetone for 5min at -20°C. After drying, section were washed in distilled water (dH_2O) then antigen retrieval was performed by incubating sections in boiling 10 mM citrate buffer pH6 (Dako). After cooling down for 30 min at room temperature (RT), sections were washed in PBS then in PBS-0.1%Tween and blocked for 1h at RT in PBS-5%. Sections were then incubated with primary antibody (rabbit anti-RSPO2 (1/100, NBP2-38688 Novus); goat anti-LEPR (1/40, AF497 R&D systems); biotinylated Isolectin-b4 (1/200, B-1205 Vector)) in a humidified chamber overnight at 4°C. After 3 washes in PBS, slides were incubated with the anti-goat AF555 (1/500, Thermofisher A21432) for 30 min at RT then washed 3 times before adding the anti-rabbit AF488 (1/500, Jackson 111-545-144) and the streptavidin 647 (1/1000, Thermofisher S32357) for 30 min at RT. After washing, slides were incubated for 10min with DAPI (1/5000, Sigma), washed in dH_2O and mounted in Fluoromount (Thermofisher). Images were acquired using LSM800 confocal microscope (Zeiss) and processed with Fiji (version 1.47 v; National Institutes of Health, USA, https://imagej-nih-gov.gate2.inist.fr/ij/).

Statistical analysis

For experiments comparing one condition to another (e.g. LSK cells cultured in presence of Rspo2 + SCF vs. cultures with SCF only) results were analyzed using paired t-tests and expressed as means ± SEM.

Supplemental References

Culhane, A.C., Thioulouse, J., Perriere, G., and Higgins, D.G. (2005). MADE4: an R package for multivariate analysis of gene expression data. Bioinformatics *21*, 2789-2790.

Le, S., Josse, J., and Husson, F. (2008). FactoMineR: An R Package for Multivariate Analysis. Journal of Statistical Software *25*, 1-18.