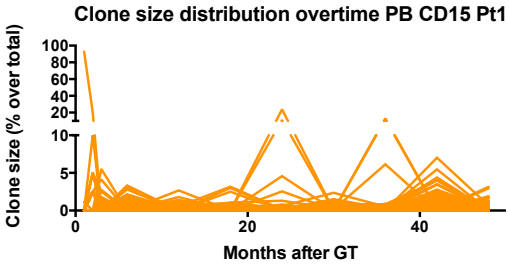
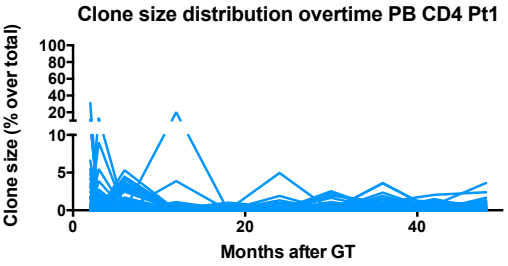
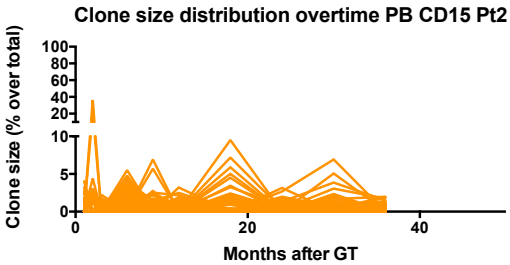
Supplemental Experimental Procedures

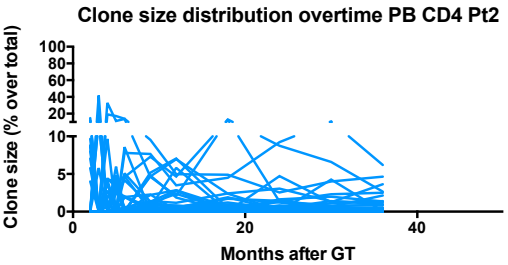
1. Longitudinal analysis of cell population diversity

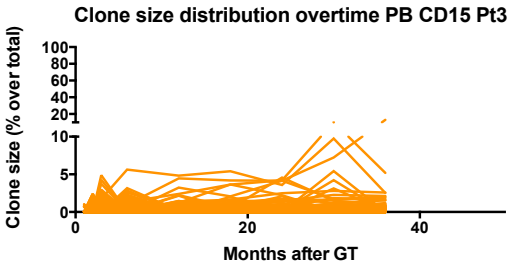
In order to analyze the diversity index for each cell population overtime a matrix of incidence *M* was generated where each row *r* represented an individual integration site (IS) while each column *c* an individual cell type/sample (e.g. CD15), source of sample (bone marrow = BM or peripheral blood = PB) and timepoint (e.g. 06 months after GT). Each entry of *M* contained the abundance of each *r* for each *c* in terms of sequencing reads. Estimation of diversity/entropy was performed through the R package Entropy (*http://cran.r-project.org/web/packages/entropy/index.html*) with *M* as input data for each patient**.** Five different methods for the calculation of diversity/entropy were applied to each *c* within *M* (ML=*maximum likelihood*; MM=*bias-corrected maximum likelihood*; SG=*Dirichlet Prior Bayesian Estimator;* Minimax=Dirichlet Prior Bayesian Estimator; CS=*Chao-Shen Entropy Estimator*) and the results were reported as longitudinal track for each cell type/sample and source in Figure 1A and Figure S1. The IS counts and clonal abundance values that are combined for the calculation of the population diversity are shown as a representative example relative to two populations (PB CD4+ and PB CD15+ cells) on the figures below reporting the clonal abundance as percentage of sequencing reads of PB CD15+ and PB CD4+ cells overtime and the number of unique IS collected in PB CD15+ and PB CD4+ cells overtime.

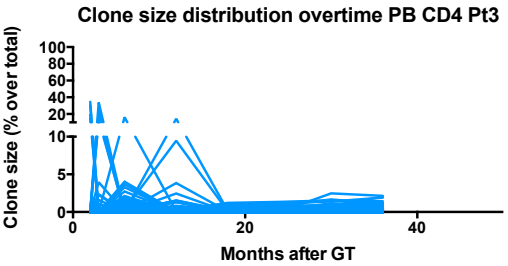

Clone size distribution overtime PB CD15 Pt1
Clone size (% over total)
100
80
60
40
20
10
5
0
0
20
40
Months after GT

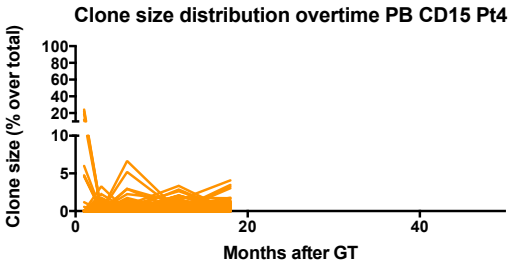

Clone size distribution overtime PB CD4 Pt1
Clone size (% over total)
100
80
60
40
20
10
5
0
0
20
40
Months after GT

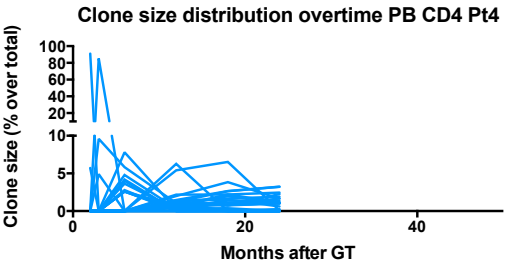

Clone size distribution overtime PB CD15 Pt2
Clone size (% over total)
100
80
60
40
20
10
5
0
0
20
40
Months after GT

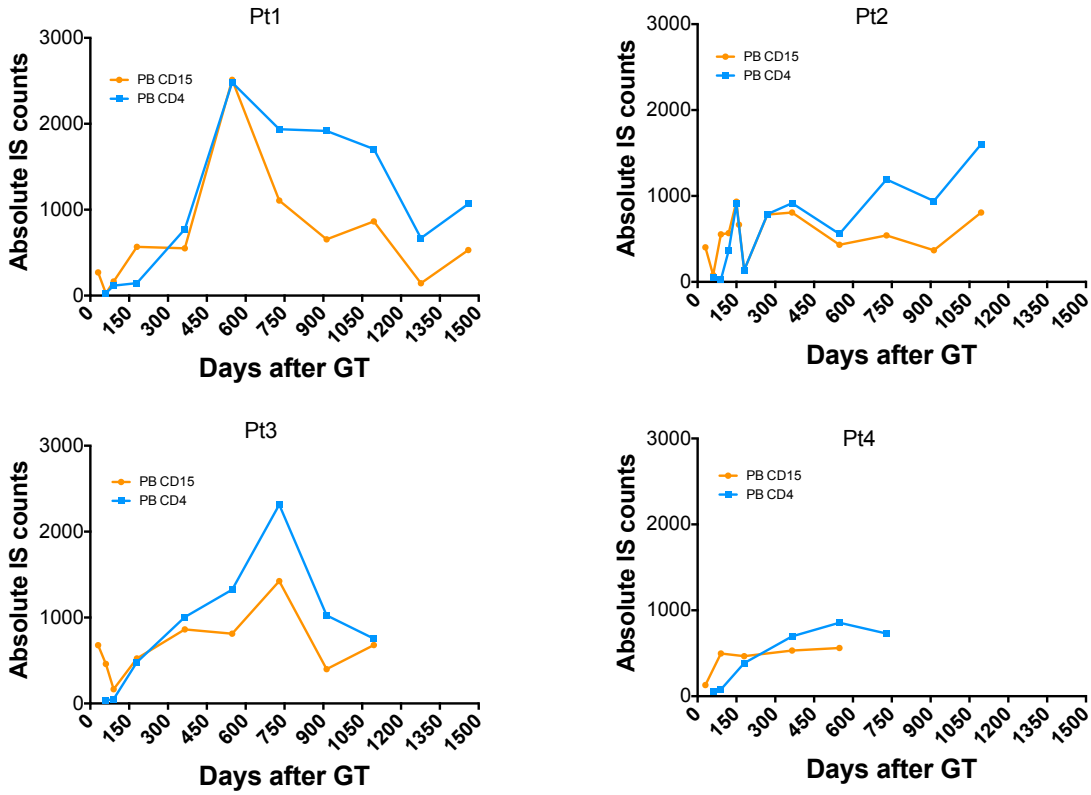

Clone size distribution overtime PB CD4 Pt2
Clone size (% over total)
100
80
60
40
20
10
5
0
0
20
40
Months after GT


Clone size distribution overtime PB CD15 Pt3
Clone size (% over total)
100
80
60
40
20
10
5
0
0
20
40
Months after GT


Clone size distribution overtime PB CD4 Pt3
Clone size (% over total)
100
80
60
40
20
10
5
0
0
20
40
Months after GT


Clone size distribution overtime PB CD15 Pt4
Clone size (% over total)
100
80
60
40
20
10
5
0
0
20
40
Months after GT


Clone size distribution overtime PB CD4 Pt4
Clone size (% over total)
100
80
60
40
20
10
5
0
0
20
40
Months after GT



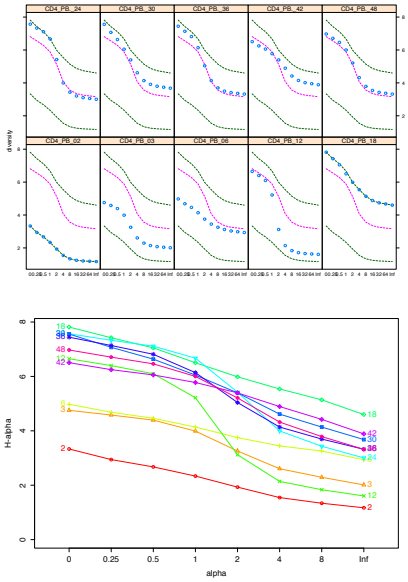
In order to extend our analysis to other methods of entropy evaluation we also calculated the Renyi diversity profile by the R package **BiodiversityR** (*http://cran.r-project.org/web/packages/BiodiversityR/index.html)* and reported the results for two different populations (CD4+ and CD15+ cells). The Rényi entropy of order α , where $\alpha > 0$ and $\neq 1$, is defined as $H_\alpha(X) = \frac{1}{1-\alpha} log(\sum p_i^\alpha)$. Here,is a discrete random variable with possible outcomes and corresponding probabilities for $i = 1, \dots, n$, and the logarithm is base 2. If the probabilities are $p_i = 1/n$for all $i = 1, \dots, n$, then all the Rényi entropies of the distribution are equal: $H_\alpha(X) = log(n)$. In general, for all discrete random variables X,$H_\alpha(X)$ is a non-increasing function in α (Rényi, 1961).

Renyi profiles are shown in the figures belo showing in blue dots the resulting curve for each timepoint and in dashed lines the lowest, median and highest diversity profile within the lineage analyzed. The less horizontal a profile is the

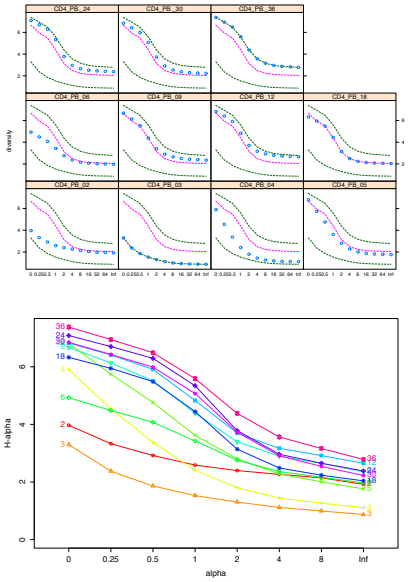
less evenly distributed are IS abundances. The profile values for $\alpha = 0$ provide information on richness. In particular the antilogarithm of values corresponding at $\alpha = 0$ is the value of each sample richness. Within Rényi profile different alpha values are considered progressively including more abundant species (namely IS). The profile of $\alpha = \infty$ contains the information on the proportion of the most abundant IS. In particular, profiles that are higher at $\alpha = \infty$ have a lower proportion of dominant IS. Importantly within specific alpha values are contained the values of different diversity indexes. The profile value for $\alpha = 1$ is the Shannon diversity index $H_1(X) = -\sum p_i \, log p_i$ for which the average of different calculations for each sample are reported longitudinally in Fig1 A. In order to verify whether the differences among lineages in terms of IS richness can significantly influence the diversity score we compared the diversity of subsets of identical sizes within each dataset belonging to each lineage. Accumulation curves were drawn on the basis of the average diversity value of 100 permutations from randomized sampling of data. Diversity has been compared for identical sample sizes according to the category “months after GT” reported as numeric values beside each line. Timepoints with more diverse profile have an higher line in the diagram.

Pt1

PB
CD4

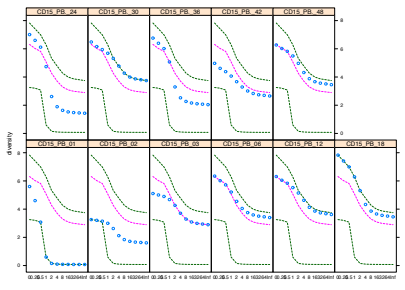


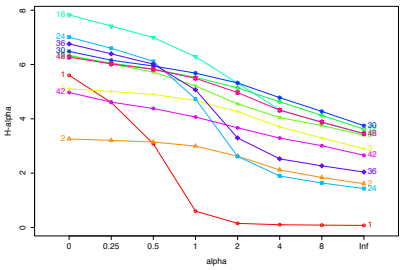
PB
CD15



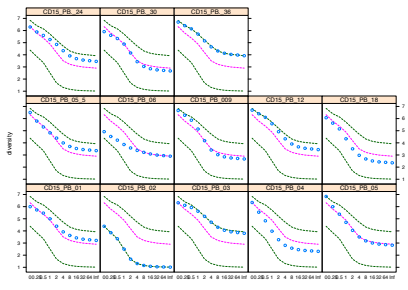
Pt2

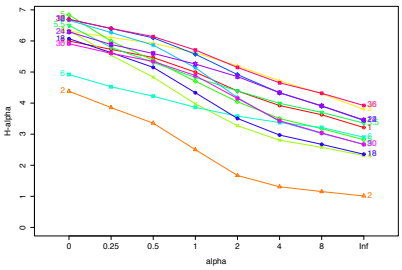
PB
CD4





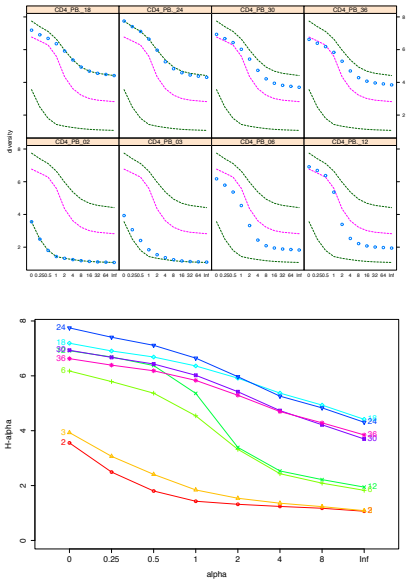
PB
CD15



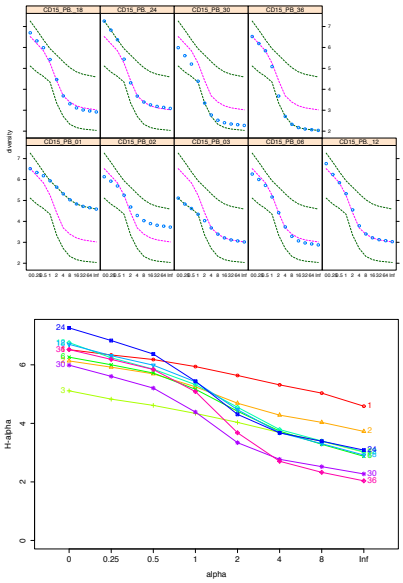


Pt3

PB
CD4

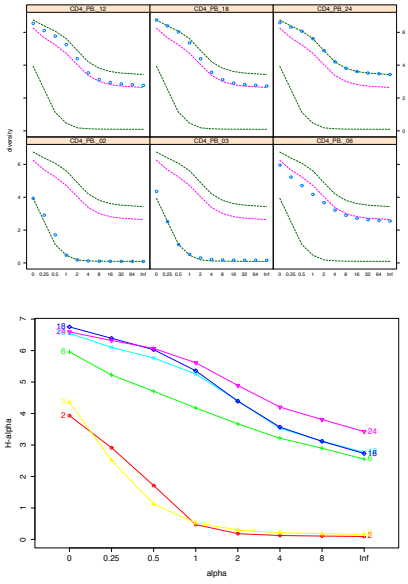


PB
CD15

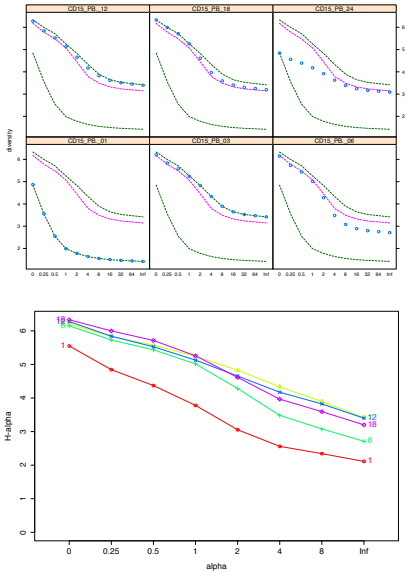


Pt4

PB
CD4



PB
CD15



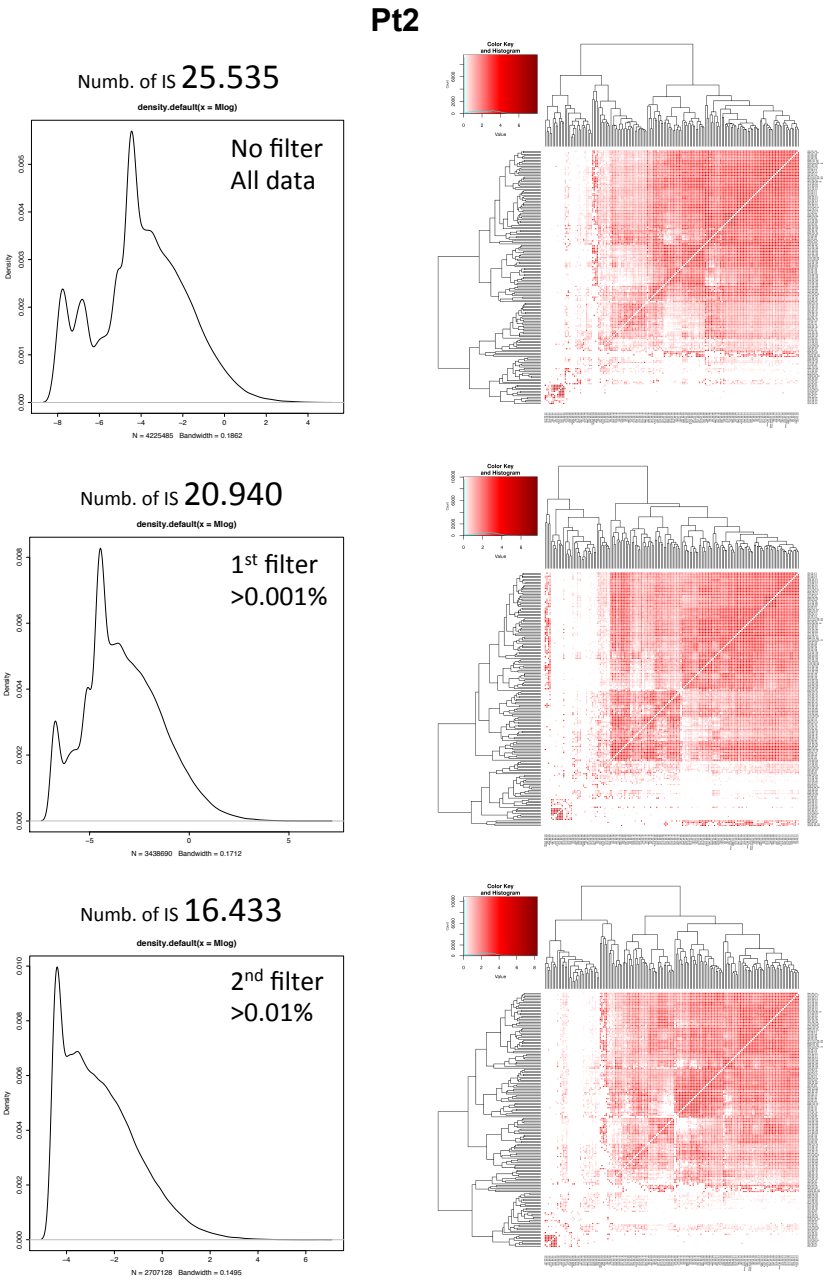
2- Positive association calculation through IS similarities among lineages

Heatmaps of supplementary figures 2,3 show positive association measurement of IS similarities among all lineages and time-points analyzed for each patient. We built incidence matrices M were rows$r$represented individual IS and column$c$represented individual cell types/samples and time-points. Each entry of M contained a binarized $0/1$ values for the presence absence of that given IS in that specific cell type and time-point. To investigate the presence/absence of dependence, we proceed as follows: given 2 generic columns of M, say i and$j$, in order to measure their association, we calculated the odds ratio (OR), corresponding to the ratio, $\mathrm{OR}_{ij} = \frac{P(IS_j=1 \vee IS_i=1)/P(IS_j=0 \vee IS_i=1)}{P(IS_j=1 \vee IS_i=0)/P(IS_j=0 \vee IS_i=0)}$.

As usual, the OR index takes value in $(1;\infty)$ if the columns are positively associated and in the range $(0;1)$ if negatively associated. This calculation had been performed pairwise, leading to a symmetric square matrix. For better visual representation, only positive $log_2(\mathrm{OR}_{ij})$ are plotted as heatmaps in supplementary figures 2,3. Heatmaps were drawn through the R package gplots (http://cran.r-project.org/web/packages/gplots/gplots.pdf) with hclust function for unsupervised clustering on average similarity. Color intensity is proportional to the values of $log_2(\mathrm{OR}_{ij})$ association coefficient.

To test whether IS similarities were significantly affected by the presence of IS with low relative abundance, as potential byproducts of cross-contamination or sequencing background, we tested the validity of this model applying progressive filters to the IS dataset. We calculated the relative contribution of each IS in terms of percentage of sequencing reads within each lineage and timepoint. We then plotted the density distribution of percentages on IS data for

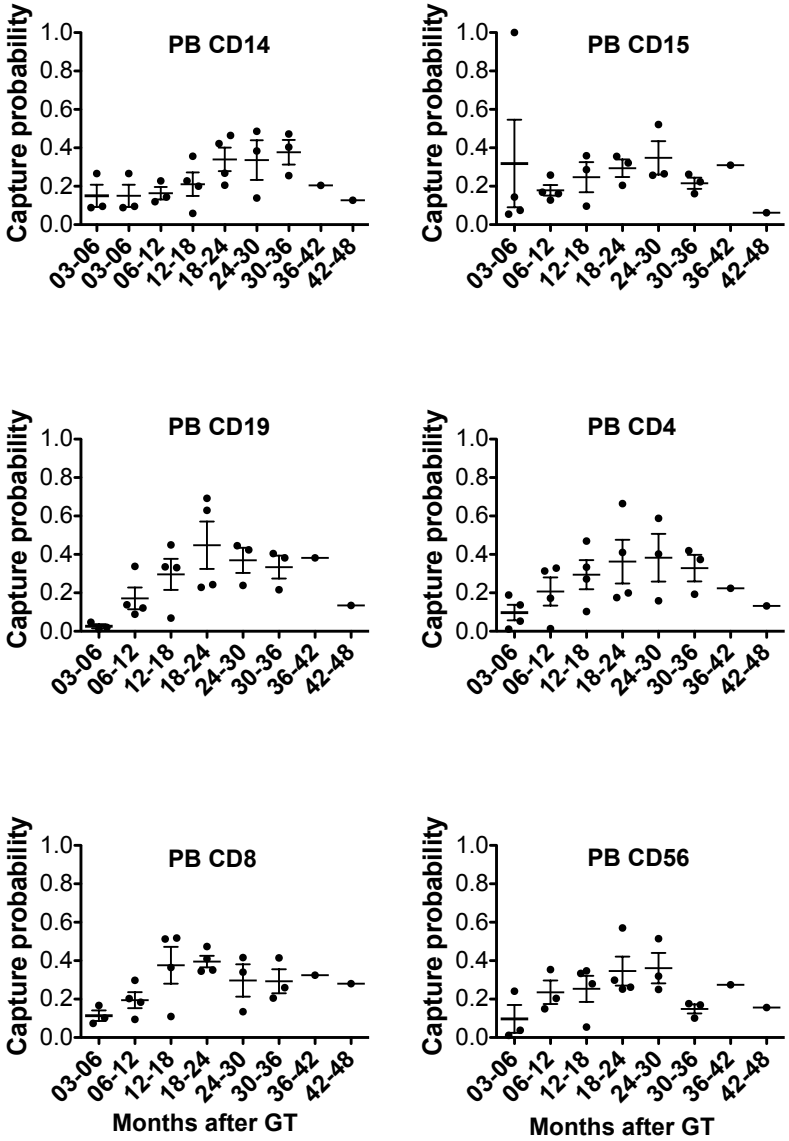
each patients. As exemplified on the following figure, we applied progressive filters to remove IS with low relative abundance and calculated and plotted $log_2(\mathrm{OR}_{ij})$ on the new filtered datasets. In the shown representative example the clustering of IS similarities was not substantially different even when, upon data filtering, the total number of IS was reduced by 9.102 IS (corresponding to 36% of total starting IS).



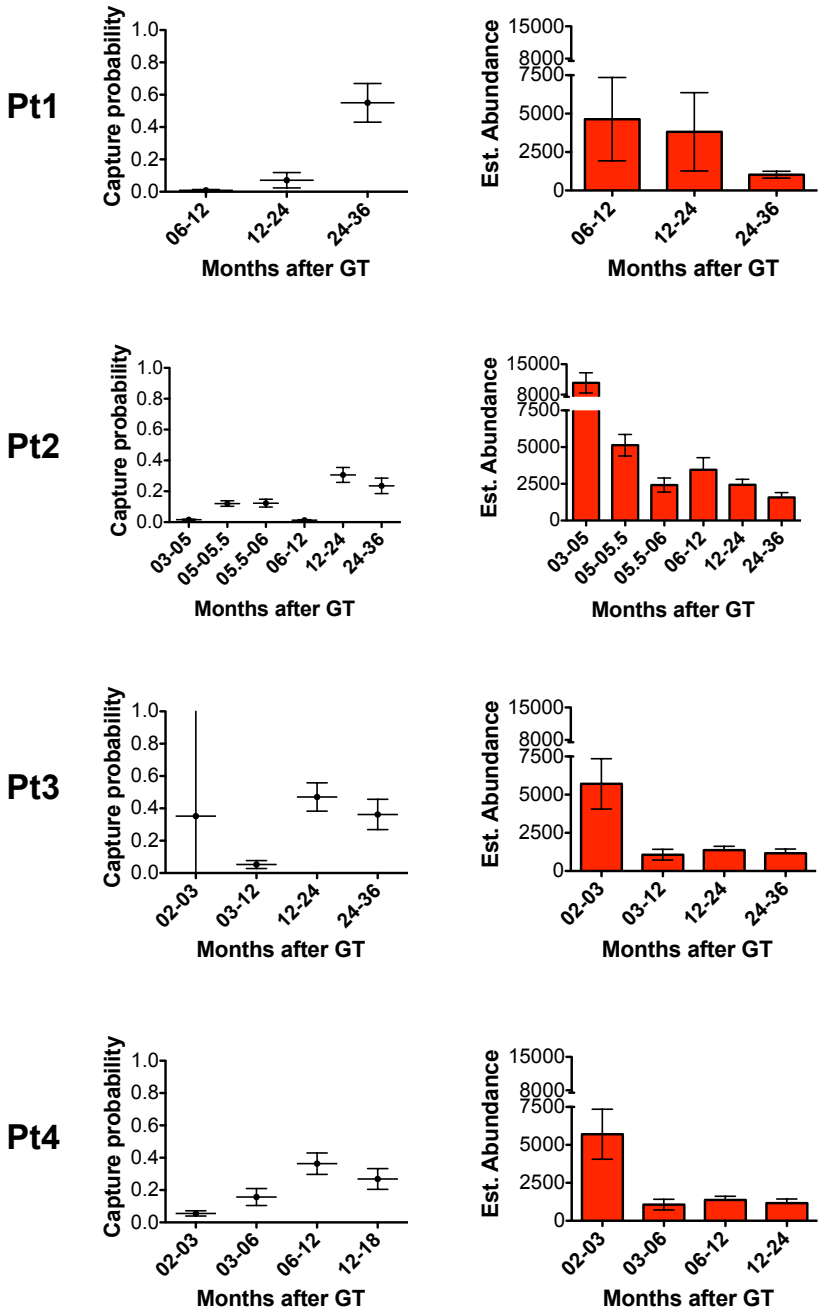
3. Estimation of population abundance by capture-recapture approach

The estimation of population abundance was performed by capture re-capture of identical IS within each cell type/sample over different timepoints. The input file for this analysis was a matrix of observed capture histories *M* where each column *c*represented a cell type and timepoint and each row*r*an individual IS. Abundance estimations and their standard errors were calculated for each timeframe within each cell type by the application to *M* of log-linear models for open populations through the *openp* function within the R package **Rcapture** (*http://cran.r-project.org/web/packages/Rcapture/Rcapture.pdf*).

Open populations capture re-capture models were chosen as they are based on several assumptions that we considered key to our IS-based tracking study. Indeed differently from previously exploited closed population models (Aiuti et al. 2013) this model assume that the clonal population can be subjected to "demographic" changes thus allowing taking into account clonal exhaustion or delayed activation of quiescent hematopoietic stem cell (HSC) clones. Capture probabilities and relative standard errors were modeled for each timeframe through the same R package. The results of these analyses are reported on Figure 1. Dotplots showing capture probabilities (y axis) of different PB lineages from four pooled WAS patients over different time windows represented as months after GT (x axis) are reported on the following figure.



The next figure shows dotplots displaying capture probabilities of BM CD34+ cells over time in four different WAS patients (left plots) over time. Estimated clonal abundance is shown in the bar graphs on the right over the same time windows.



4. Calculation of CD34+ population output overtime by IS sharing

In order to assess the output of CD34+ cells we calculated the percentage of IS shared by CD34+ cells with other lineages overtime. For each timepoint we grouped IS from multiple lineages into two main groups: Myeloid (composed by BM CD14+, CD15+, CD61+, Glycophorin+ cells and PB CD14+, CD15+ cells) and

Lymphoid (composed by BM CD19+, CD3+, CD56+ cells and PB CD19+, CD3+, CD4+, CD8+, CD56+ cells). We then calculated the percentage of IS collected from BM CD34+ cells at a given timepoint that were also detected in Myeloid, Lymphoid or both groups from the same timepoint. In order to account for potential cross-contaminations we also applied a filter to the analysis on the basis of the relative contribution of each IS as percentage of sequencing reads on total sequencing reads from BM CD34+ and Myeloid/Lymphoid datasets. A given IS was considered shared when the ratio between relative contributions in BM CD34+ cells vs Myeloid/Lymphoid cells of the same timepoint was below 10. When a IS was 10 fold more represented in one group as compared to the other it was assigned only to the first and not to the second. The results of these unfiltered and filtered analyses are reported in Figure 2 and in Supplementary Figure 6. As already observed for the filters applied for the analysis of positive correlation coefficient, we verified that filtering based on IS relative contribution did not substantially altered the overall trend of CD34+ IS sharing overtime. Thus, it was justifiable to exploit the whole datasets for the following analysis based on BM CD34+ IS sharing. In Figure 2B the CD34+ IS sharing is reported in more details. We built symmetric adjacency matrices M_{adj} were rows r and column c represented individual cell types/samples and timepoints. Each entry of M_{adj} contained the absolute number of IS shared between corresponding $c$and$r$. The diagonal of M_{adj} contained the number of IS belonging to each cell type/sample and timepoint. For this panel we isolated from M_{adj} the percentages of IS shared between CD34+ cells and the other cell subtypes of the same timepoint and produced individual matrices M_{out} with these values for each timepoint and patient. These data were used as input for drawing individual

plots through **Circos Table Viewer** software (http://mkweb.bcgsc.ca/tableviewer/), as reported in Figure 2B. Shared IS are reported in the "rainbow" area of the plot. The relative dimension of this area vs the red area represent the redundancy of IS detection in each lineage (dimension of the output) while the color variegation of the "rainbow" area is reflective of the number of lineages sharing IS with CD34+ cells (nature of the output).

5. Analysis of input from CD34+ cells to different lineages overtime

In figure 2 we analyzed the input that each lineage has received overtime from CD34+ cells in terms of shared IS. For this analysis we used, as input, data from the adiacency matrix M_{ad} (Figure 2B) used for the representation of CD34+ output (Figure 2B). Differently from what done for Figure 2B, where we considered how many IS from CD34+ cells were shared with other lineages of the same timepoint, we here evaluated how many IS from each lineage were shared with CD34+ cells of the same timepoint. This difference is exemplified on the table below. Here we reported, for the follow up 36 months after GT, in red the percentage of BMCD34+ vs PBCD14+ shared IS calculated on the total IS from BMCD34+ cells (CD34+ output, Figure 2B) and in blue the percentage of BMCD34+vsPBCD14+ shared IS calculated on the total IS from PBCD14+ cells (CD14+ input from CD34+ cells, Figure 2C).

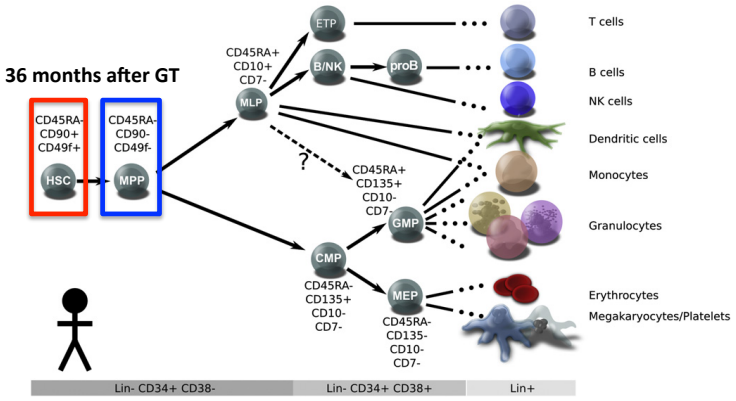
	CD34_BM_36	CD14_PB_36
CD14_PB_36	**24.5**	100.0
CD34_BM_36	100.0	**10.4**

We then built a matrix M_{input} containing in each column *c* BM CD34+ cells from a given patient and timepoint and in each row *r* a specific cell type (e.g. PB CD14+ cells). Each field of M_{input} contained the percentage of shared IS corresponding to

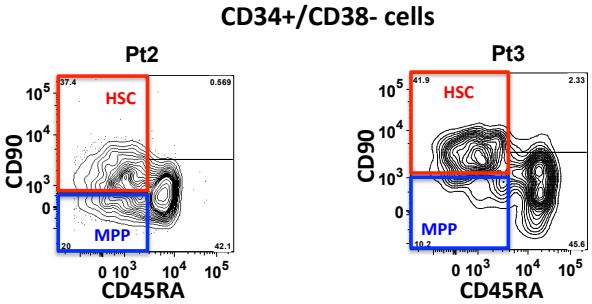
the input towards each lineage from CD34+ cells at a given timepoint. We then plot M_{input} as an heatmap through the R package **gplots** (*http://cran.r-project.org/web/packages/gplots/gplots.pdf)* with *hclust* function for unsupervised clustering on average similarity and color intensity corresponding to the relative value of input from CD34+.

6. Evaluation of HSC and MPP IS sharing overtime

The following figure shows a schematic representation of human hematopoietic hierarchy (top panel adapted from Doulatov et al. Cell Stem Cell 2012). HSC and MPP that were isolated at 36 months after GT from two WAS patients are highlighted by red and blue frames, respectively. The resulting cytofluorimetric contour plots show the expression of CD90 and CD45RA surface markers gated on Lin-/CD34+/CD38- populations are shown below. FACS-sorted HSC and MPP are highlighted in red and blue, respectively.



Adapted from Doulatov et al. *Cell Stem Cell* 2012



In order to represent the presence of shared IS between FACS sorted HSC and MPP and other ex vivo isolated cell types overtime, we generated an incidence matrix M_{prog} where each column c contained a different cell subtype, source and timepoint while each row r an individual IS found in HSC or MPP. Each entry of M_{prog} reported a sequencing reads value when that specific HSC or MPP IS on r was detected also in that given cell subtype/timepoint on c. The percentages of IS within HSC or MPP shared with at least one other cell subtype were reported in the stacked bar graph of Figure 3. Then, each sequencing reads value was converted in a "detection value" proportional to the overall level of detection for each IS overtime and over the lineages. Only shared IS were selected and ordered

within HSC and MPP groups from the more detected to the less detected in the other lineages and timepoints. A heatmap (reported in Figure 3) was created through the R package **gplots** (*http://cran.r-project.org/web/packages/gplots/gplots.pdf*) using these data as input with different color intensities reflecting different detection values. We then considered the abundance of the HSC or MPP shared IS within Whole BM, Whole PB, BM MNC and PBMC isolated at the same timepoint (36 months after GT) and we reported in Figure 3 the relative percentage of sequencing reads belonging to these IS as surrogate marker of the real time input by HSC or MPP towards BM and PB. In order to further quantify the level of HSC and MPP output overtime we calculated an additional "sharing score". For each timepoint we calculated the level of detection of each IS in Myeloid, Lymphoid T, B and NK cells. We considered that each group was composed by different cell subtypes that were not always evenly isolated for every timepoint. Thus in one timepoint for example the Lymphoid T group (CD3, CD4, CD8) could have been composed just by CD4 and CD8 T cells since CD3 were not purified. In order to normalize our results for this variability we calculated the maximum sharing score as number of IS x number of cell subtypes for each group in each timepoint. According to the data reported in the table below the maximum sharing score of HSC from Pt2 vs Lymphoid_T group at 5 months after GT was 6*2=12 while at 24 months after GT was 6*3=18. We then calculated the sharing score as percentage of the experimental score on the maximum score for HSC or MPP vs each subgroup in each timepoint and reported the results in tabular form as in the table below

Total cell types composing the subgroups													
Months after GT	*1*	*2*	*3*	*4*	*5*	*5.5*	*6*	*9*	*12*	*18*	*24*	*30*	*36*
Myeloid	**2**	**2**	**2**	**2**	**2**	**1**	**2**	**2**	**2**	**2**	**2**	**2**	**2**
Lymphoid_T	**1**	**3**	**2**	**2**	**2**	**1**	**2**	**3**	**3**	**3**	**3**	**3**	**3**
Lymphoid_B	**1**	**1**	**1**	**1**	**1**	**1**	**1**	**1**	**1**	**1**	**1**	**1**	**1**
Lymphoid_NK	**1**	**1**	**1**	**1**	**1**	**1**	**1**	**1**	**1**	**1**	**1**	**1**	**1**
	Total shared IS												
HSC	**6**												
MPP	**4**												

Sharing score values, shown in the next table in red, were plotted on the graph of Figure 2D.

HSC	*Months after GT*	*1*	*2*	*3*	*4*	*5*	*5.5*	*6*	*9*	*12*	*18*	*24*	*30*	*36*
experimental score	Myeloid	0	0	0	1	2	0	0	7	9	9	8	2	5
	Lymphoid_T	0	0	0	0	4	1	2	12	11	11	12	8	11
	Lymphoid_B	0	0	0	1	1	0	1	5	5	3	5	0	3
	Lymphoid_NK	0	0	0	0	0	0	2	5	4	4	4	2	3
Maximum score	**Myeloid**	**12**	**12**	**12**	**12**	**12**	**6**	**12**	**12**	**12**	**12**	**12**	**12**	**12**
	Lymphoid_T	**6**	**18**	**12**	**12**	**12**	**6**	**12**	**18**	**18**	**18**	**18**	**18**	**18**
	Lymphoid_B	**6**	**6**	**6**	**6**	**6**	**6**	**6**	**6**	**6**	**6**	**6**	**6**	**6**
	Lymphoid_NK	**6**	**6**	**6**	**6**	**6**	**6**	**6**	**6**	**6**	**6**	**6**	**6**	**6**
Sharing score	**Myeloid**	**0.0**	**0.0**	**0.0**	**8.3**	**16.7**	**0.0**	**0.0**	**58.3**	**75.0**	**75.0**	**66.7**	**16.7**	**41.7**
	Lymphoid_T	**0.0**	**0.0**	**0.0**	**0.0**	**33.3**	**16.7**	**16.7**	**66.7**	**61.1**	**61.1**	**66.7**	**44.4**	**61.1**
	Lymphoid_B	**0.0**	**0.0**	**0.0**	**16.7**	**16.7**	**0.0**	**16.7**	**83.3**	**83.3**	**50.0**	**83.3**	**0.0**	**50.0**
	Lymphoid_NK	**0.0**	**0.0**	**0.0**	**0.0**	**0.0**	**0.0**	**33.3**	**83.3**	**66.7**	**66.7**	**66.7**	**33.3**	**50.0**

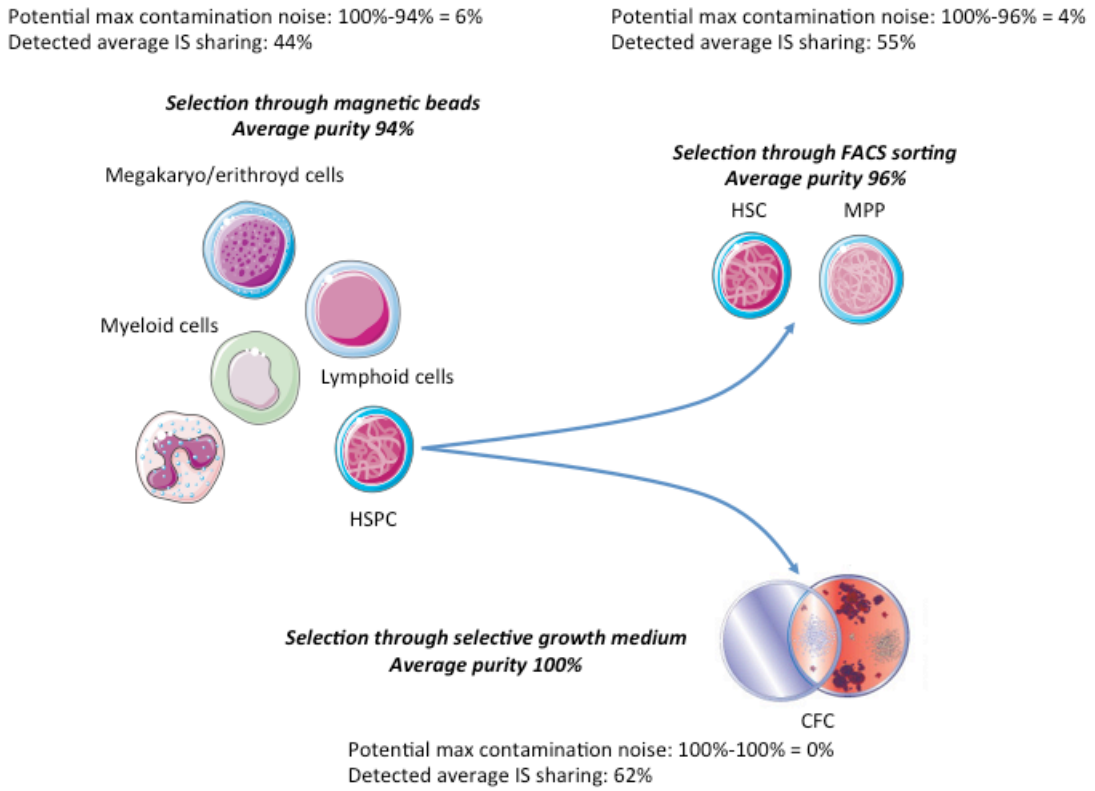
7. Analysis of sharing, entropy, number of IS and VCN for individual samples at the latest follow up.

In Figure 4A we evaluated different parameters on the IS datasets deriving from different lineages isolated from the latest follow up. We grouped the lineages as BM CD34+, BM Myeloid, PB Myeloid, BM Lymphoid and PB lymphoid with the same categories composition described in section 4 of supplementary material. The position of each dot in the plots corresponds to the value relative to a given lineage placed in its own category. Data from all patients are grouped together. Percentage of shared IS was calculated as the fraction of IS for each lineage that were shared with at least another lineage of the same timepoint (latest follow up). Number of IS and vector copy number were measured for each lineage as previously described (Aiuti et al. Science 2013). The entropy values were calculated for each lineage as described in section 1 of supplementary material.

8. Addressing cross-contamination and clonal quantification issues

Given the nature of the study, that relies largely on the identification and the tracking of identical IS among different timepoints and lineages, we devised several strategies to reduce as much as experimentally feasible and efficiently estimate the impact of cross-contamination among samples, which is an inherent constraint of all clonal tracking studies. A first source of potential contamination comes from the selection/purification of individual cell lineages from the patients' samples. As summarized in the figure below, IS were collected from different cell sources purified by different strategies. The average purity of cell lineages selected by magnetic beads was 94%. This high level of purity was achieved through a protocol that envisages that each cell type should undergo two sequential rounds of beads purification with its specific surface marker directed antibody. HSC and MPP were instead FACS sorted from purified CD34+ cells and yielded a purity range from 92% to 99% (average 96%) as assessed by re-analysis of sorted cells. In the case of CFC grown on methocult® from ex vivo purified CD34+ cells, the purity should be considered 100% as only HSPC could have generated colonies on these specific in vitro conditions. On the basis of these levels of purity and according to the average level of IS sharing detected among different cell types on independent analyses, we could estimate that the vast majority of shared IS (approximately 86% of shared IS from beads-selected cell types, 93% of shared IS from FACS sorted cell types, and 100% of shared IS from CFC; see captions on the figure) could have not been merely detected due to the occurrence of cross-contaminations among purified cell samples. Importantly, one should note that our estimates are conservative as the potential cell contaminants are also composed by a relevant fraction of vector-negative

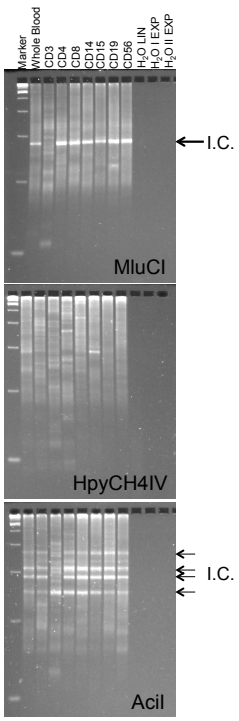
cells thus further reducing the risk of contaminating a selected cell type with unrelated IS.



A second source of cross-contamination among independent samples could arise after DNA extraction, LAM-PCR or high-throughput sequencing, where potential inter-sample contaminations could occur through aerosols or spillovers. We used different strategies to evaluate the impact of these events on our results.

Addressing the impact of contamination occurring upon DNA extraction is relatively straightforward. It should be indeed noted that identical integrants were detected not only among samples purified at the same time but consistently, with similar trends among patients, on genomic samples that were purified independently over different timepoints, with up to 3 years from one DNA isolation to the other. It is then extremely unlikely that DNA isolation had *per se* significantly impacted the outcome of this study.

One could still envisage that the several molecular manipulations performed during LAM-PCR procedure might be susceptible to inter-samples cross-contamination. To address this issue, for every round of LAM-PCR performed in the present study H_2O negative controls were added at each step of the procedure (1 for the linear amplification, 1 for the first exponential PCR and 1 for the second exponential PCR) and were processed in parallel to the other samples. Importantly, for all rounds of LAM-PCR and for all enzymes in use the 3 lines corresponding to negative controls were always clean and did not display any band, as shown on the following representative spreadex gel picture of (I.C. = internal control bands). This is a formal proof of the overall cleanliness of the procedure that we put in place for IS retrieval.



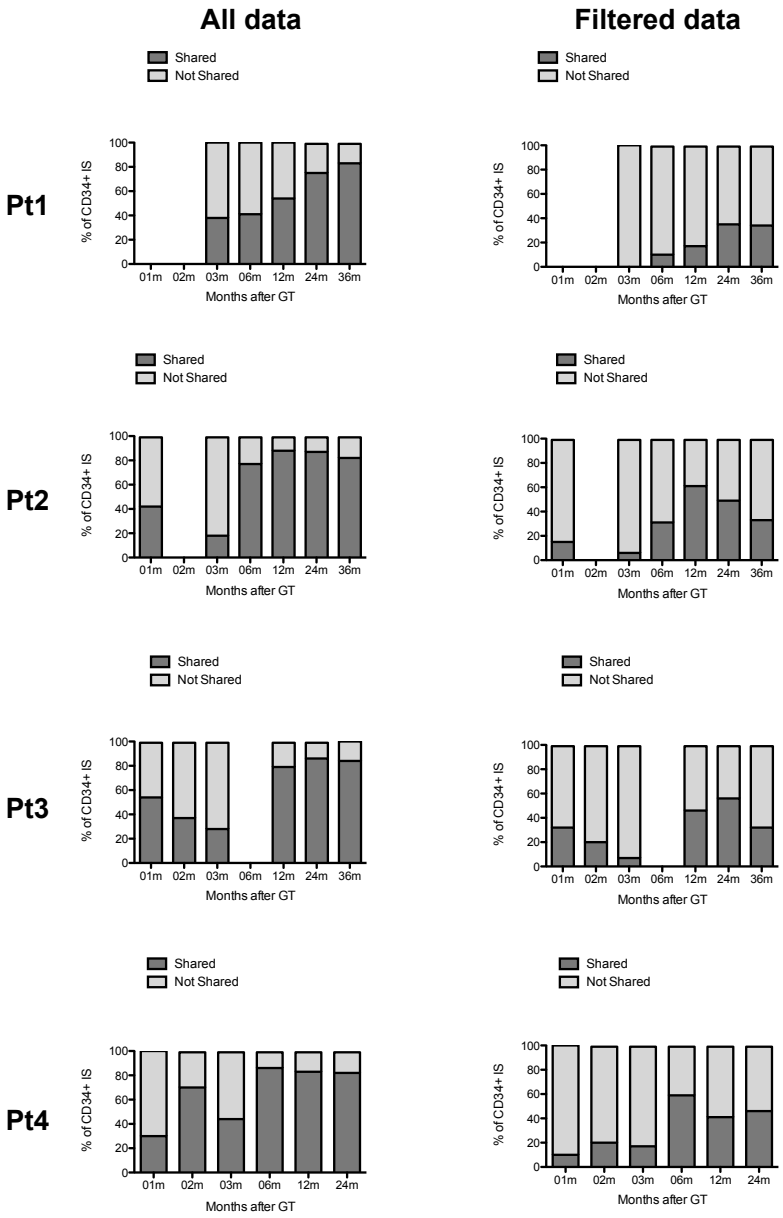
To further validate the purity of our strategy and to account for potential cross-contaminations occurring upon NGS processing (which could also most likely be the source of the collisions events observed among samples from independent

patients an filtered from our datasets as reported on Aiuti et al. Science 2013) we analyzed in details the sequencing reads on a controlled experiment where we collected integrations from individual clones with known vector content and IS localization. The results reported in the following table show that almost the totality of sequencing reads were associated to the expected integration sites belonging to each given clone with an average false discovery rate (FDR) of 0.66% and a precision ranging from 98.06% to 99.99%. These levels of accuracy were also highly reproducible as resulting from the analysis of 6 technical replicates of the same clone (clone1) generated independently starting from individual genomic DNA isolations. Notably, these sets of data also validate the precision of our bioinformatic pipeline currently in use for IS identification (manuscript in preparation) and the filters applied for removing inter-samples/patients collisions.

Clone	Known IS	Found IS	Chr	Position	Strand	Annotation	Replicates*	Read Count	Tot Reads	Precision	FDR**
Clone 1	1	1	chr8	8,866,487	+	ERI1	Replicate 1	70,190	70,241	99.93 %	0.073 %
							Replicate 2	103,027	103,706	99.35 %	0.655 %
							Replicate 3	57,268	57,729	99.20 %	0.799 %
							Replicate 4	17,939	18,293	98.06 %	1.869 %
							Replicate 5	68,975	68,991	99.98 %	0.023 %
							Replicate 6	30,096	30,575	98.43 %	1.567 %
Clone 2	4 and 6	5	chr2	73,762,397	-	ALMS1	Replicate 1	26,610	68,339	99.69 %	0.31 %
			chr11	64,537,167	-	SF1	Replicate 1	11,160			
			chr16	28,497,497	-	CLN3	Replicate 1	4,163			
			chr17	2,032,351	-	SMG6	Replicate 1	3,242			
			chr17	47,732,338	-	SPOP	Replicate 1	22,952			
Clone 3	1	1	chr3	52,306,696	-	WDR82	Replicate 1	95,803	95,815	99.99 %	0.012 %

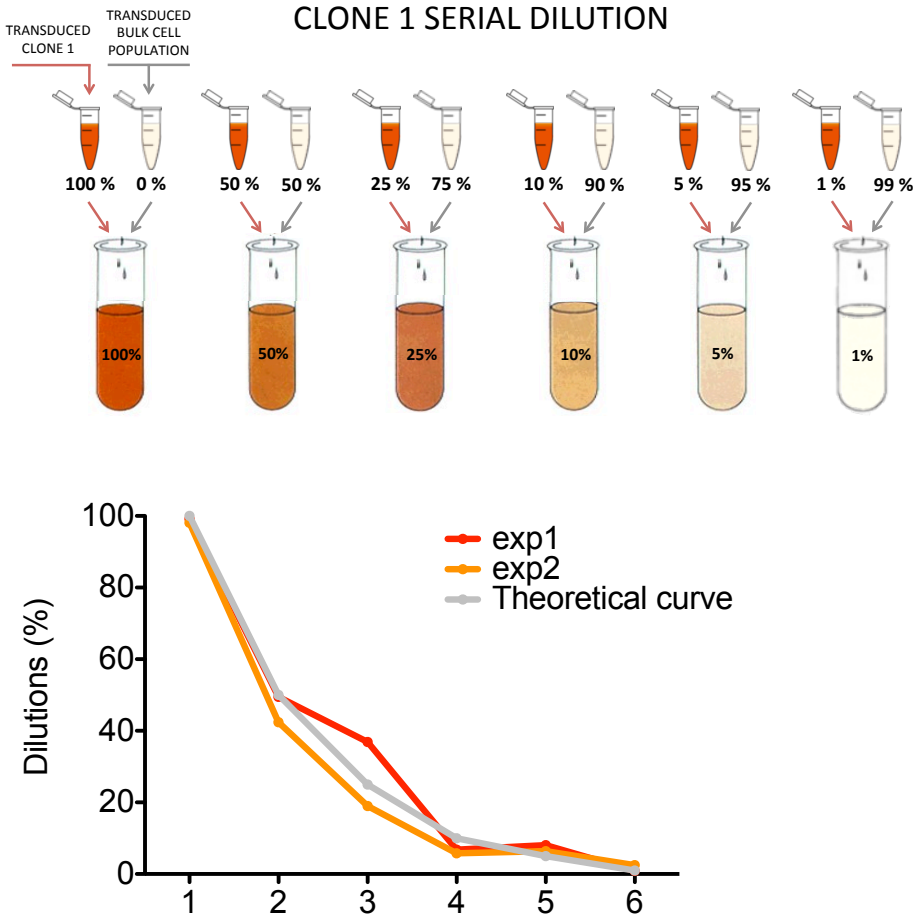
Having underlined that the issue of contamination was well considered and addressed from different perspectives, for specific analyses we introduced a further “stress test” on our datasets applying specific bioinformatics filters to account for potential unforeseen further contamination events (details described above on section 4 of supplementary materials). Importantly, even if this

stringent conditioning reduced substantially the levels of IS sharing, the overall overtime trend was maintained further validating the biological consistency of our data. The following figure shows bar charts displaying the percentage of IS from BM CD34+ cells shared with myeloid or lymphoid lineages (dark gray) from the same timepoint (months after GT) on four WAS patients (left panels). The same analyses performed on data filtered for contamination are shown in the bar charts on the right (see supplementary materials for details). Note that in the unfiltered analysis the extent of sharing reached very high proportions, but it became more difficult to distinguish between different biological output vs. contamination among samples.



A second confounding factor could have potentially affected the interpretation of our results in particular when analyzing individual clonal contributions and the diversity of the clonal repertoire. The use of sequencing reads as surrogate markers of clonal abundance is well established in the field but it is not exempt from technical biases linked to the several rounds of exponential amplifications employed during the LAM-PCR procedure. We are indeed aware of the fact that a given integration site could be more easily amplified than another regardless of

the initial representation of the corresponding clones in the original sample. The best way to evaluate the impact of these biases on clonal quantification is to perform serial dilutions of a clone with known IS on a bulk population containing thousands of unrelated IS. We performed such a validation on serial dilutions of clone 1 into genomic DNA isolated from in vitro BM CD34+ cells transduced with the same lentiviral vector. As shown in the figure below our data, obtained by means of relative representation of sequencing reads belonging to the IS of clone 1 (two independent experiments: red and orange lines), overlap quite well the theoretical dilution curve (grey line). Thus, quantification of IS through sequencing reads could indeed provide a good surrogate marker of clonal abundance.



Nevertheless, we recognize that the clonal quantification through sequencing reads will never be 100% accurate and it is still possible that the detection of certain IS could be more affected than other by PCR biases. The quantification of clones of small size could be particularly affected by these biases and the call for having observed a *bona fide* IS event could be difficult when this IS is associated with low reads counts. Addressing this issue, we were able to validate our results even after introducing a further "stress test" based on progressively removing the IS with lowest sequencing reads from our datasets, as shown in supplementary figure 3. One should also note that, conscious of the potential limitation of our technique, in the present study we never focus or made specific claims on the abundance of given individual clones but we instead discussed the changes observed overtime on the general clonal size distribution within each cell population underlining the presence of common trends among patients.

Despite potential technical constraints, making use of IS sharing and sequencing counts to estimate clonal relationships and relative abundance respectively, we generated a set of data consistent among patients and with a strong biological meaning. Actually, the concordance of our results with previous clonal tracking studies performed on animal models represents the ultimate validation of the approach used in this work for the tracking of the fate of engineered cells in humans.

9. Models of hematopoietic hierarchy

Methods used for testing the models schematically represented in Figure 4B

We placed the IS analysis within a Bayesian Network (BN) context to be able to handle complex dependence structure among several variables combining

information on conditional probability distributions and graphical models, and to provide evidence supporting specific hierarchical structures. We remark that this methodology is not only related to the validation of "assumed" biological structures but it can be employed also in more general settings to explore the whole space of dependencies.

One important advantage of the introduction of BN is to learn structure directly from data. The strictly hierarchical nature of the differentiation process and the absence of confounding factors in determining hypotheses of differentiation allows for an interpretation of the graph in "causal" terms, eluding one of the major hurdle in reading the dependence structure out of a graphical model. BNs are multivariate statistical models satisfying sets of (conditional) independence statements contained in a directed acyclic graph (DAG). A DAG is a pair $G = (V, E)$ where V is the set of nodes and $E$is the set of directed edges (arrows) between pairs of nodes. Each node represents a random variable, while missing arrows between nodes imply (conditional) independence between the corresponding variables. A directed graph is acyclic in the sense that it is forbidden to start from a node and, following arrows directions, go back to the starting node. The acyclicity is set to be in line with the hierarchical structure of the differentiation process. We remark that in reading graphical model an arrow connecting two variables, should not be interpreted as "biological" dependence, but rather as a statistical conditional dependence relation, that is a dependence in terms of conditional probability. In particular, conditional independence statements and absence of an arrow are connected by the Markov properties (Lauritzen (2001)). Moreover, the BN definition associates a factorization of the joint distribution that highlights the dependence structure of the variables (chain

rule), using as factors the conditional distributions attached to each node in the BN. In general, the chain rule for k categorical random variables states that:

$$p(X_1 = x_1, X_2 = x_2, \dots, X_k = x_k) = \prod \prod p\left(X_i | pa(X_i)\right)$$

where $pa(X_i)$ is the set of parents of node $X_i$and $p(X_i|pa(X_i))$ is the probability of $X_i = x_i$ given the set of values (configuration) assumed by the parents of X_i.

In the structural learning process, when the number of nodes (variables) is larger than 5, the number of possible network structures is so large that it is necessary to resort to appropriate search algorithms.

In general, structural learning can been carried out using either constraint based methods (such as the PC and the NPC algorithm) or score+search algorithms, both of them implemented in the software HUGIN (www.hugin.com).

Networks *goodness* can been compared by using the following two criterion:

- Bayesian Information Criterion (BIC)
- Receiving operating characteristic curve (ROC) and the related Area Under Curve (AUC) measure.

Bayesian Information Criterion (BIC) is given by:

$$BIC(DAG|data) = LogLik(DAG: data) - \frac{1}{2} Klogn$$

where K is the number of free parameters in the model.

The best model is associated with the highest value of the BIC score, indicating that we obtain the best log-likelihood corrected by the number of parameters to estimate (to avoid over-parametrization) and by the number observations. Direct BIC scores comparisons can be performed only within the same input dataset.

Moreover we remark that BIC can be directly interpreted in terms of posterior probability change. Indeed optimizing the BIC corresponds to optimizing the posterior model probability for a large number of observations and can be thought of as a way of selecting a model. We also recall that a *little change in BIC represents a big change in terms of posterior probability* as it can be seen by the following relation between BIC and posterior probability gain (Wit et al.2012):

$$W(M) = e^{-BIC(M)/2}$$

$$p(M_0|y) = \frac{W(M_0)}{\sum_{M \epsilon M} W(M)}$$

Different approaches are possible when learning a Bayesian network, ranging from "complete free" model, that means "no constrains" are assumed *a priori* and the conditional independence relations can be read off the DAG, to a "fixed-constrains" approach. The latter places BN in a more "confirmative" perspective since it can be used to first estimate the conditional probability distributions and then to provide statistical evidence in favor of one assumed biological structure with respect to a different one, by computing the associated BIC's.

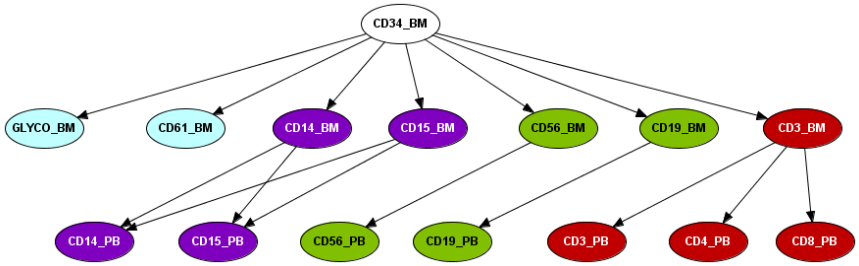
The ROC analysis helps validating the model since allows to inspect the performance of a given variable as a classifier for the data set. Hence the ROC curve has been used to evaluate whether a proposed model is accurate in predicting the bottom level variables (BM level) of the differentiation hierarchy.

The area under the ROC curve (AUC) can be used as a synthetic measure for the goodness of the network as a classifier, varying between 0.5 (a model with a null discriminant capacity) and 1 (a model with a perfect discriminant capacity).

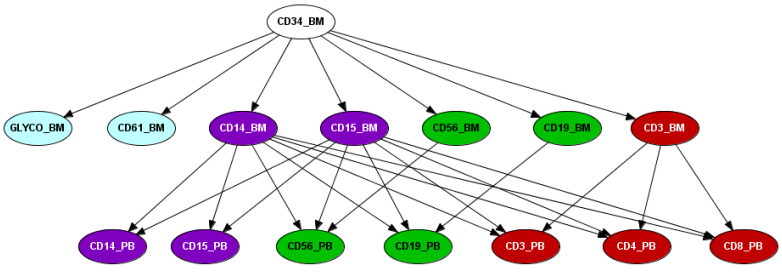
In this analysis we present results according to two different postulated chain models representing two alternative hematopoietic hierarchies schematized in

Figure 4B and reported in details here as in the figures below. We here represented with arrows positive dependences, defined as the increased probabilities of observing an IS in a downstream cell type with respect to chance, having observed the same IS in the upstream cell type.

Structure of conditional dependencies underlying the model 1 reported in Figure 4B



Structure of conditional dependencies underlying the model 2 reported in Figure 4B



Measures of *goodness* of Model 1 and Model 2 are provided in terms of both BIC and AUC on the table below (Area under the ROC curve (in black) and BIC (in blue and italic) for seven examined differentiation structures).

	AUC Model1	AUC Model2
CD14PB	0,61	0,61
CD15PB	0,61	0,62
CD56PB	0,50	0,54
CD19PB	0,54	0,56
CD3PB	0,58	0,61
CD4PB	0,51	0,57
CD8PB	0,58	0,62
BIC	*-103859*	*-103272*

Our results showed that model 2 is more informative both in terms of BIC and in terms of *discriminating capacity* (AUC).

Methods used for generating the model of Figure 4C

In this section we describe the model that has been generated to investigate the hematopoietic differentiation process in humans, in vivo in humans from gene therapy clinical trial data. A brief description of the underlying ideas, assumptions, model and inference procedure is given. At the beginning of the experiment, a sample of a patient's CD34+ cells is transduced, resulting in a pool of corrected cells that are unequivocally and permanently labeled by means of vector integration site coordinates. After re-infusion in the patient's body, corrected stem cells start to duplicate and differentiate in functionally more specialized cells, reconstituting the hematopoietic heritage of the patient. During the follow-up period, analyzing samples from patient bone marrow and peripheral blood by means of integration site analysis, we are able to observe the off-spring generated by a single transduced stem cell. Furthermore, using

sequence count data derived from a NGS platform, we can reasonably quantify the amount of cells, belonging to each of the 15 cell types (CT), of which 8 are from bone marrow (BM) and 7 from peripheral blood (PB) CTs, and that originated from the same progenitor. Samples are taken at 12, 24 and 36 months, collected from 3 patients, resulting in 37,637 clones tracked.

These data are interpreted as fixed time observations from a underlying latent 15-dimensional stochastic Markov counting process in continuous time. In other words, at time 0 at re-infusion of the transduced cells in the patient's body the counts of all CT are equal to 0, except for the unique CD34+ cell, for which is equal to 1. After a random, continuous time interval Δt an event occurs, changing the state of the process for each cell. Three different types of events are considered in our model: cell duplication, cell death and cell differentiation. These events can be described by means of a set of quasi-reactions and their associated transition rates. Cell duplication of CTi occurs with rate x_i α_i in the sample, where x_i is the number of cells of CTi and α_i is the rate of an individual cell duplication event of that type. Similarly, cell death of CTi is assumed to be happening with rate x_i δ_i. Most interestingly, cell differentiation changes a cell of type CTi in a cell of type CTj with rate x_i λ_"ij" . Clone dynamic evolution corresponds to a sequence of single events occurring to the cells belonging to the clone considered. Event rates αδ and Λ govern the underlying latent process and can be interpreted as the number of events of a particular type expected per time unit.

The aim is to estimate transition rates θ=(α,δ,Λ) based on the observed 37,637 clones and, most importantly, to identify which differentiation paths are most supported by the empirical data. In the analysis of biochemical reaction

networks (Purutcuoglu and Wit, 2007), various estimation strategies have been proposed. The above Markov counting process can be approximated by a diffusion process (Gillespie 1996). The continuous diffusion can be approximated by a discretized Euler-Maruyama approximation, which allows a straightforward local linear approximation approach. A constrained generalized least squares method allows us to estimate strictly transition rates.

Using this algorithm we are not only able to obtain positive rate estimates, but also to incorporate additional biological knowledge, such as the impossibility of peripheral blood cell differentiation into stem or pluripotent cells. This is straightforward and done simultaneously with estimation, setting some element of Θ to zero in the constraint list. The complete list of estimated rates (M) are reported on the table below with corresponding confidence interval.

	CD34	CD3	CD14	CD15	CD19	CD56	CD61	GLYCO	CD3	CD4	CD8	CD14	CD15	CD19	CD56
CD34	0.566 (0.556;0.577)	0.044 (0.041;0.047)	0.066 (0.062;0.069)	0.044 (0.041;0.047)	0.209 (0.205;0.213)	0.015 (0.012;0.017)	0.025 (0.022;0.028)	0.08 (0.077;0.083)	0.062 (0.059;0.065)	0.043 (0.04;0.046)	0.017 (0.015;0.02)	0.064 (0.061;0.067)	0.083 (0.08;0.086)	0.063 (0.06;0.066)	0.301 (0.297;0.305)
CD3	0	0.237 (0.233;0.24)	0.025 (0.024;0.026)	0.059 (0.058;0.061)	0.022 (0.021;0.023)	0.001 (0.001;0.002)	0.054 (0.052;0.055)	0.059 (0.058;0.061)	0.137 (0.135;0.139)	0.048 (0.047;0.05)	0.101 (0.099;0.103)	0.06 (0.058;0.061)	0.004 (0.003;0.004)	0.052 (0.051;0.054)	0.031 (0.03;0.033)
CD14	0	0.006 (0.003;0.008)	1.578 (1.564;1.591)	0.263 (0.258;0.268)	0.37 (0.365;0.375)	0.076 (0.073;0.08)	0.188 (0.184;0.193)	0.308 (0.303;0.313)	0.006 (0.004;0.009)	0.029 (0.026;0.032)	0 (0;0.003)	0.091 (0.086;0.095)	0.145 (0.141;0.15)	0.027 (0.024;0.031)	0.167 (0.163;0.171)
CD15	0	0.001 (0;0.002)	0.066 (0.064;0.069)	-0.274 (-0.28;-0.269)	0.002 (0;0.004)	0.014 (0.013;0.016)	0.041 (0.039;0.043)	0.077 (0.074;0.079)	0.007 (0.006;0.008)	0 (0;0.001)	0 (0;0.001)	0.053 (0.051;0.055)	0.089 (0.086;0.091)	0.066 (0.064;0.069)	0 (0;0.001)
CD19	0	0.101 (0.099;0.104)	0.029 (0.026;0.031)	0.022 (0.02;0.024)	-0.157 (-0.165;-0.15)	0.034 (0.032;0.036)	0.041 (0.039;0.043)	0.022 (0.02;0.025)	0.077 (0.075;0.08)	0.123 (0.121;0.126)	0.013 (0.011;0.015)	0.024 (0.022;0.027)	0.013 (0.011;0.016)	0.11 (0.108;0.113)	0.001 (0;0.003)
CD56	0	0.053 (0.051;0.056)	0.051 (0.049;0.054)	0.011 (0.009;0.013)	0.143 (0.14;0.146)	-0.016 (-0.023;-0.009)	0.003 (0.002;0.005)	0.052 (0.05;0.055)	0.013 (0.011;0.015)	0.026 (0.024;0.028)	0.004 (0.002;0.005)	0.023 (0.02;0.025)	0.029 (0.027;0.032)	0.109 (0.106;0.111)	0 (0;0.003)
CD61	0	0.03 (0.027;0.033)	0.025 (0.022;0.029)	0.04 (0.036;0.043)	0.044 (0.041;0.048)	0.004 (0.001;0.006)	-0.388 (-0.398;-0.377)	0.036 (0.032;0.039)	0.098 (0.094;0.101)	0.021 (0.018;0.024)	0.009 (0.007;0.011)	0.038 (0.035;0.041)	0.135 (0.131;0.14)	0.022 (0.019;0.025)	0 (0;0.002)
GLYCO	0	0 (0;0.002)	0.094 (0.091;0.097)	0.028 (0.026;0.03)	0.019 (0.016;0.021)	0.001 (0;0.003)	0.024 (0.022;0.027)	-0.601 (-0.61;-0.593)	0.005 (0.003;0.007)	0.001 (0;0.003)	0 (0;0.002)	0.063 (0.061;0.066)	0.05 (0.048;0.053)	0.001 (0;0.003)	0 (0;0.002)
CD3	0	0	0	0	0	0	0	0	-0.385 (-0.391;-0.378)	0	0	0	0	0	0
CD4	0	0	0	0	0	0	0	0	0	-0.4 (-0.407;-0.393)	0	0	0	0	0
CD8	0	0	0	0	0	0	0	0	0	0	-0.222 (-0.224;-0.219)	0	0	0	0
CD14	0	0	0	0	0	0	0	0	0	0	0	-0.67 (-0.676;-0.664)	0	0	0
CD15	0	0	0	0	0	0	0	0	0	0	0	0	-0.593 (-0.599;-0.588)	0	0
CD19	0	0	0	0	0	0	0	0	0	0	0	0	0	-0.613 (-0.618;-0.608)	0
CD56	0	0	0	0	0	0	0	0	0	0	0	0	0	0	-0.242 (-0.246;-0.238)

According to the informations available from the biology, a 3 levels hierarchy on cells lineages is assumed, corresponding to, from top to down, 1 HSC (CD34), 7 lineages in BM (CD3, CD14, CD15, CD19, CD56,CD61, GLYCO) and 7 lineages in PB (CD3, CD4, CD8, CD14, CD15, CD19, CD56). Regarding constraints on differentiation paths, HSC can differentiate in any other cell type in BM and PB; BM can be connected to any cell type in BM and PB level; PB lineages cannot differentiate. Entries on the main diagonal $M_{CT_i;CT_i}$ (highlighted in yellow) correspond to net duplication rates $\alpha_{CT_i} - \delta_{CT_i}$ (duplication rate – death rate). The generic off-diagonal $M_{CT_i;CT_i}$ elements represent the estimated

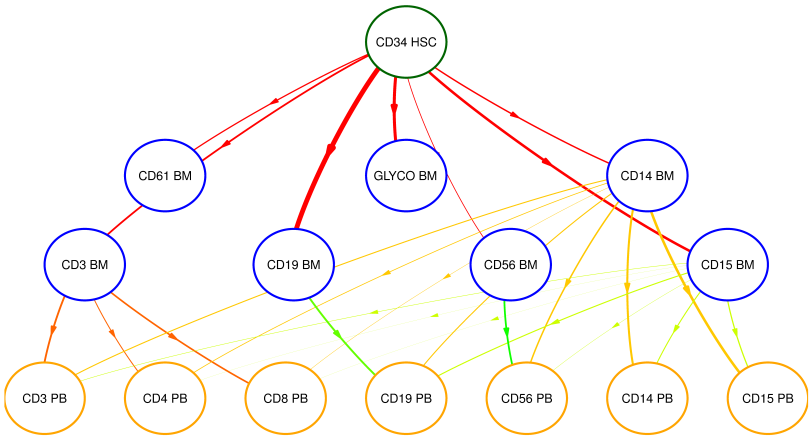
differentiation rates λ_{CT_i,CT_j}. All rates refer the expected number of events per year.

IS are known to be affected by various sources of possible noise and this may have an impact on the measurements of cells counts. To investigate the reliability of available data, part of treated HSCs sample has been analyzed by means of NGS immediately after the transduction. It is realistic to assume that in time interval elapsed, few events are likely to occur to transduced cells. Such experiment allow to measure the variability introduced by biological and technical protocols to clones known to be of size 1. Since experimental data present some outliers, we adopt the Median Absolute Deviation (MAD) statistics as dispersion measure, known to be more robust in these cases. Based on 3104 IS, a MAD statistic equal to 6.1 has been calculated. Using this information, a process noise estimation is calculated as σ^2=(1.48 x MAD)^2=81.5 (Rousseeuw1993). This information is incorporated in inference modifying estimation for parameters variance-covariance matrix.

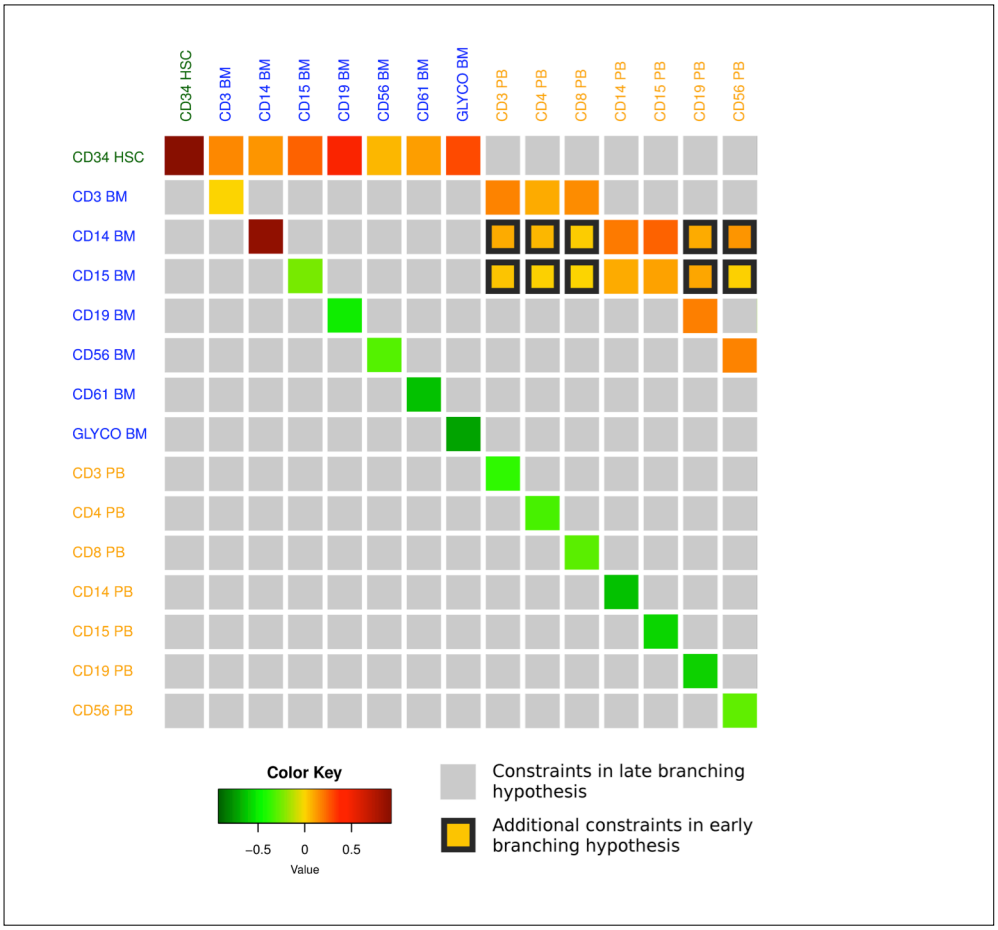
Contamination could also affect the procedure and is a more complicated problem to address from a methodological point of view. However, if it is realistic to assume that the total number of these misleading observations and associated reads counts are relatively small (see section 8 of supplementary material), their potential impact on final results is however mitigated by the constrained generalized least squares formulation and aforementioned variability estimations.

In order to take into account for differentiation rates positiveness, confidence intervals are obtained by means of a Monte Carlo technique, consisting in generating S=10000 random sample from a multivariate normal distribution

with mean vector and covariance matrix set equal to results obtained from constrained generalized least squares procedure. To display the results of our analysis, we performed a Principal Component Analysis (PCA) on M and then plotted the first 2 components. To further validate the biology consistency of this modeling approach and to compare the validity of the two alternatives hematopoiesis described in Figure 4B we set up a dynamic model according to late myeloid-lymphoid branching (early branching is nested within it) and estimate the remaining rates in order to verify which ones are less likely to be present (estimated as closed to 0). A network representation of the estimated model is showed on the figure below. Arrows correspond to differentiation rates estimated and their thickness is proportional to rates estimate.



The following figure gives a graphical representation of both rates' estimates and rates constrained, by means of an heatmap where parameters estimates and constraints induced by late myeloid-lymphoid branching are shown in grey squares and additional constraints due to early myeloid-lymphoid branching are colored squares with black edges.



The results derived from these additional analyses agree with the conclusion drawn from the BN analyses. The late branching schema seems to be more supported by the empirical data observed. However, since some differentiation paths rates estimates are closed to zero, we could also hypothesize that the true underlying structure could be an intermediate between the two. The analyses represent a prof of principle of the applicability of such stochastic dynamic model to study differentiation process in human, in vivo by means of IS analysis and demonstrate the flexibility of our tool to further refinement according to the availability of new biological information.

Additional references:

Rényi, Alfréd (1961). On measures of information and entropy. Proceedings of the fourth Berkeley Symposium on Mathematics, Statistics and Probability 1960. pp. 547–561.

Lauritzen, S. L. (2001). Causal inference from graphical models. Chapman and Hall/CRC Press, London/Boca Raton.

Wit, E, Romeijn, JW, Van den Heuvel, ER, (2012) An introduction to model uncertainty. Statistica Neerlandica, 66(3), p.5-21.

Purutcuoglu, V, Wit, EC (2008) Bayesian inference for the MAPK/ERK pathway by considering the dependency of the kinetic parameters. Bayesian Analysis, 3(4), pp.851-886.

Gillespie, D. (1996). The multivariate langevin and Fokker-Planck equations. American Journal of Physics, 64(10):1246–1257.

Rousseeuw, P. J. and Croux, C. (1993). Alternatives to the median absolute deviation. Journal of the American Statistical Association, 88(424):1273–1283.