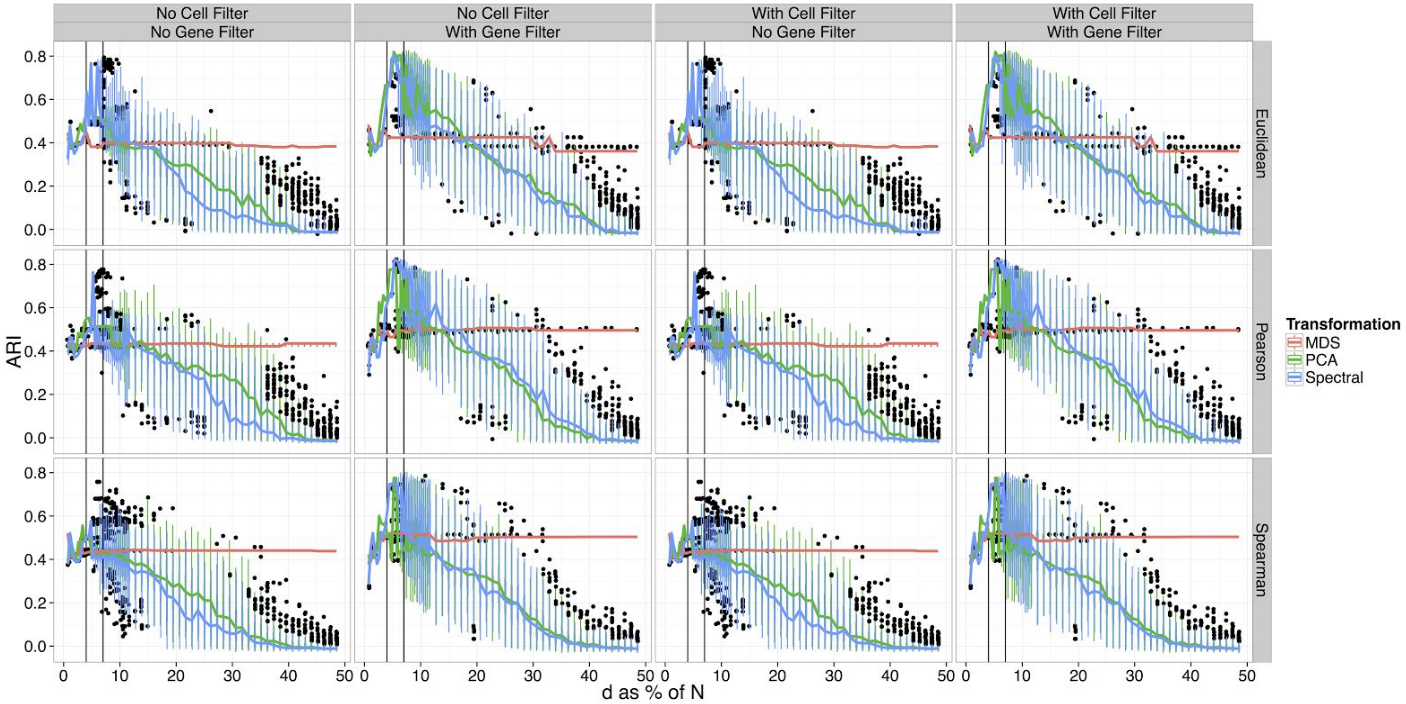

Figure S1. **Boxplots of 100 realizations of the SC3 clustering on the Deng dataset.** For clarity, lines are drawn through the medians of the boxplots. The x-axis shows the number of eigenvectors *d* of the transformed distance matrix as a percentage of the total number of cells *N* in each dataset. The black vertical lines correspond to *d* = 4% of *N* and *d* = 7% of *N* (*N* = 268). Dots represent outliers that are higher than the highest value (or lower than the lowest value) within 1.5 * IQR, where IQR is the inter-quartile range, or distance between the first and third quartiles.

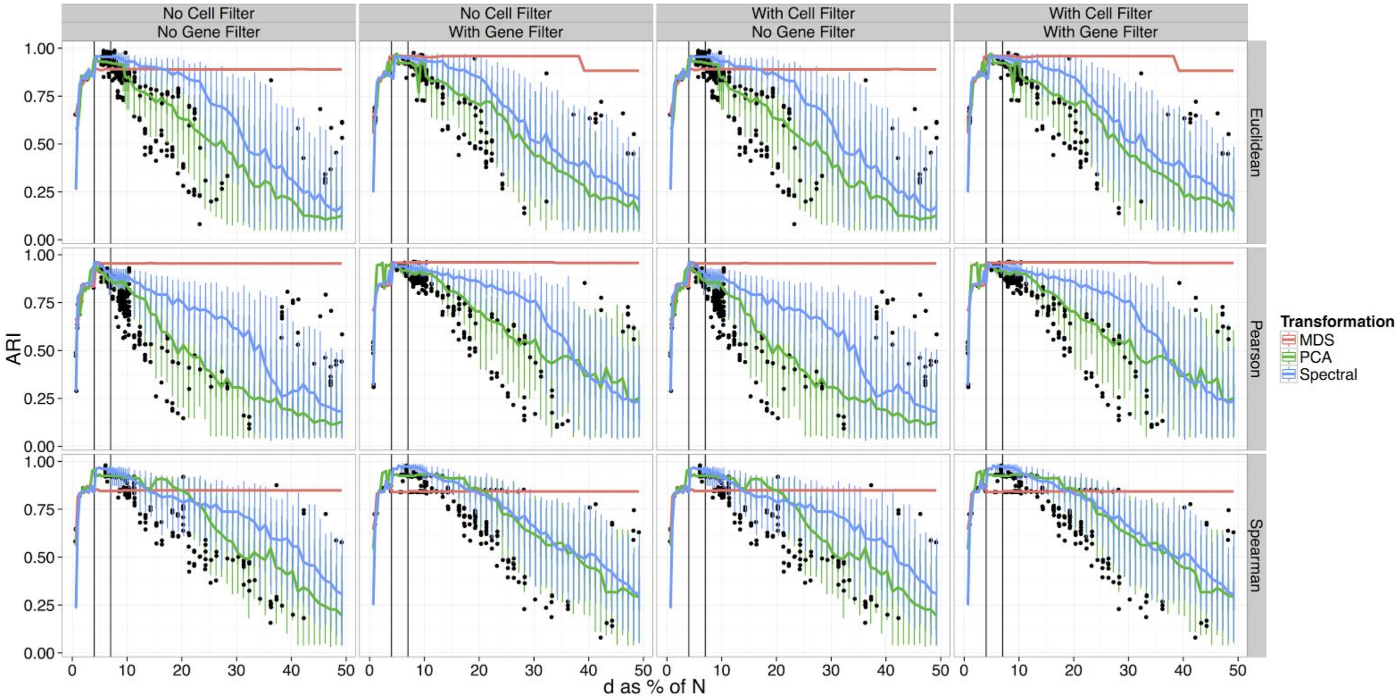

Figure S2. **Boxplots of 100 realizations of the SC3 clustering on the Pollen dataset.** For clarity, lines are drawn through the medians of the boxplots. The x-axis shows the number of eigenvectors *d* of the transformed distance matrix as a percentage of the total number of cells *N* in each dataset. The black vertical lines correspond to *d* = 4% of *N* and *d* = 7% of *N* (*N* = 301). Dots represent outliers that are higher than the highest value (or lower than the lowest value) within 1.5 * IQR, where IQR is the inter-quartile range, or distance between the first and third quartiles.

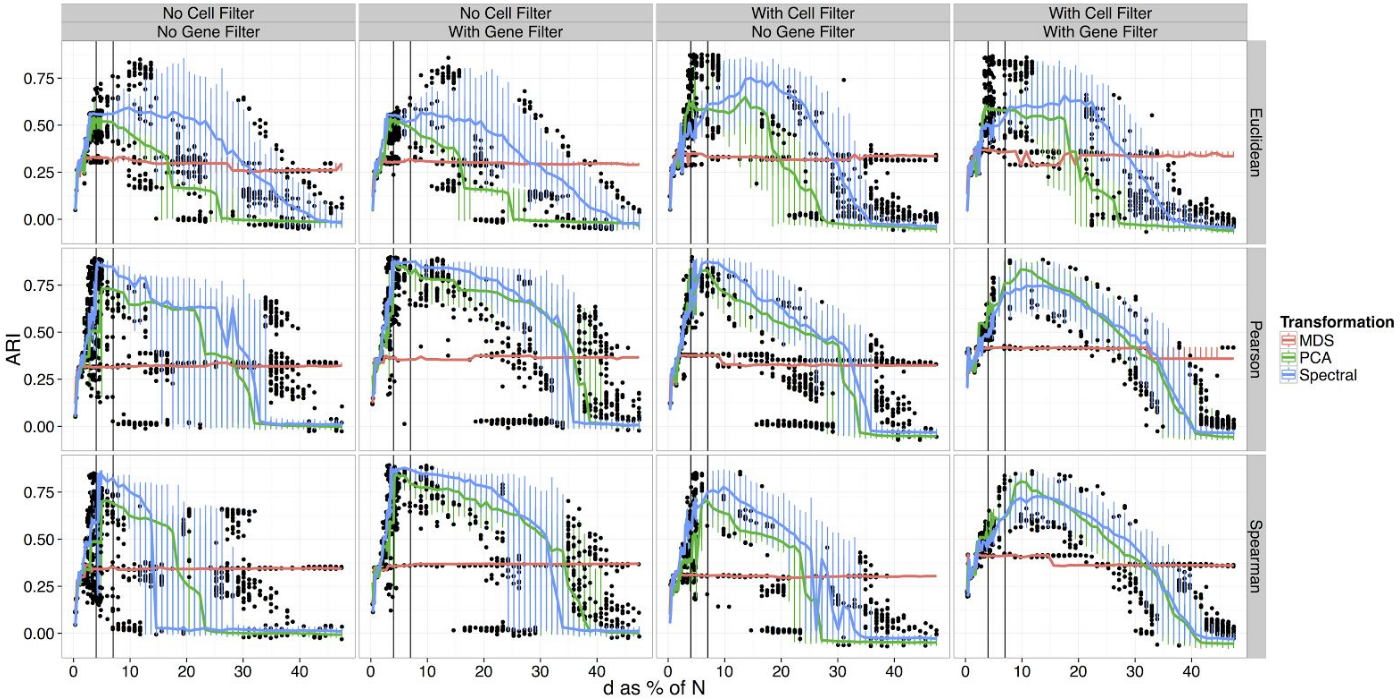

Figure S3. **Boxplots of 100 realizations of the SC3 clustering on the Usoskin dataset.** For clarity, lines are drawn through the medians of the boxplots. The x-axis shows the number of eigenvectors *d* of the transformed distance matrix as a percentage of the total number of cells *N* in each dataset. The black vertical lines correspond to *d* = 4% of *N* and *d* = 7% of *N* (*N* = 622). Dots represent outliers that are higher than the highest value (or lower than the lowest value) within 1.5 * IQR, where IQR is the inter-quartile range, or distance between the first and third quartiles.

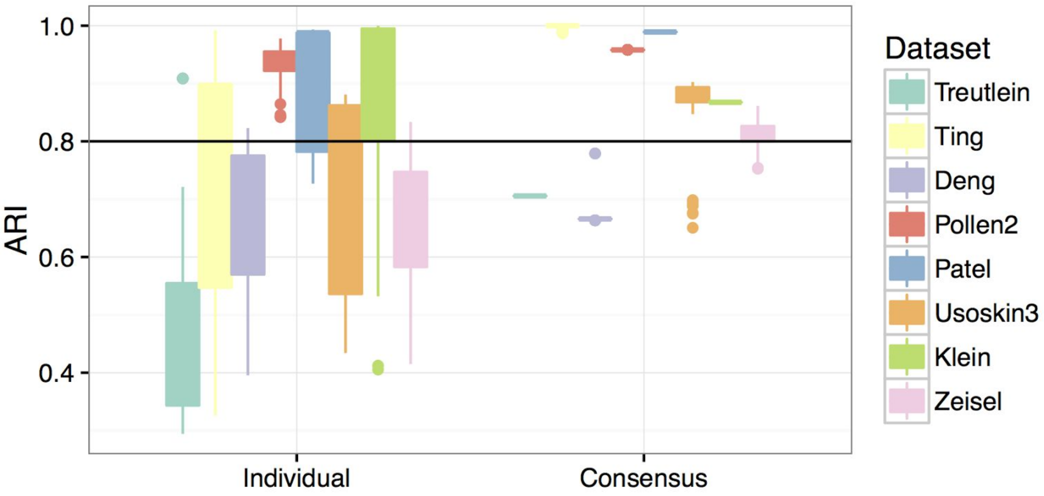

Figure S4. **Effect of consensus clustering on ARI.** Boxplots of 100 realizations of the SC3 clustering of all validation datasets (Fig. 1b). *Individual* corresponds to clustering without consensus approach. *Consensus* corresponds to the consensus clustering over the parameter set (see Methods for more details). The black line corresponds to ARI=0.8. Dots represent outliers that are higher than the highest value (or lower than the lowest value) within 1.5 * IQR, where IQR is the inter-quartile range, or distance between the first and third quartiles.

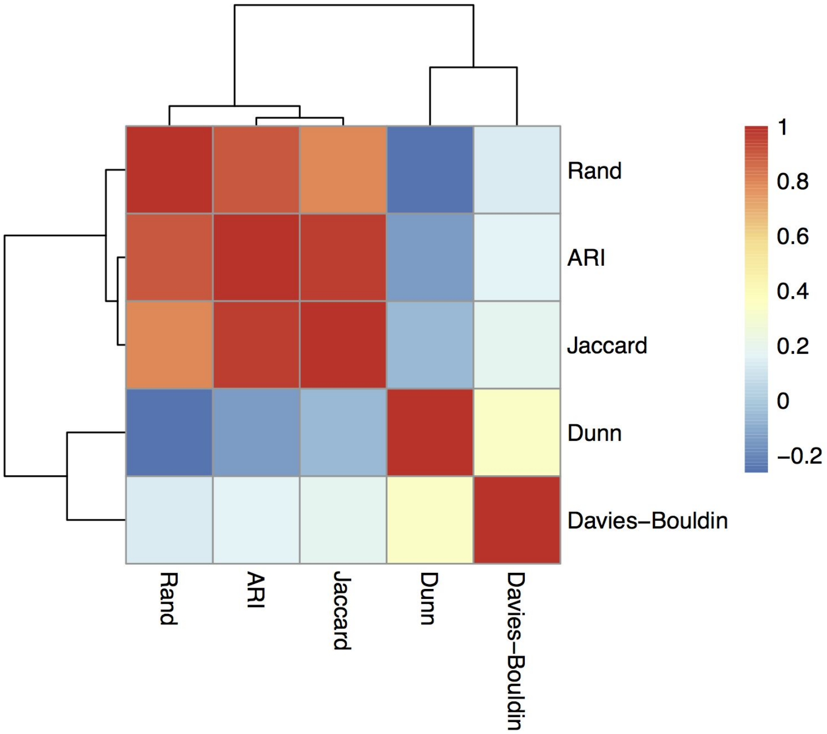

Figure S5. **Correlations between different external and internal measures of clustering.** Correlations are based on all results of SC3 clustering presented in Figs. S1-S3.

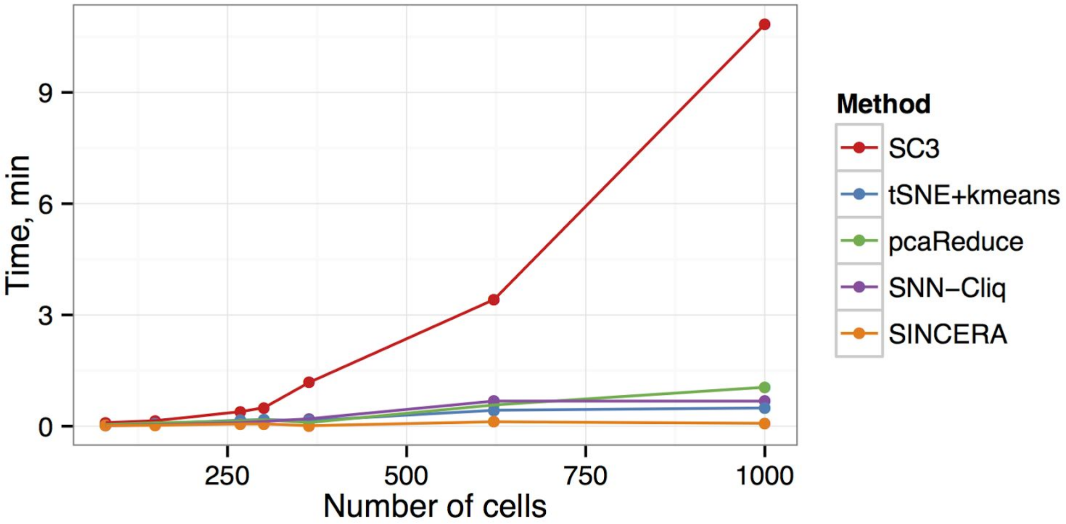

Figure S6. **Comparison of running times of SC3 with existing methods for different number of cells in an input expression matrix.** These measurements were performed on a MacBook Pro (Mid 2014), OS X Yosemite 10.10.5 with 2.8 GHz Intel Core i7 processor, 16 GB 1600 MHz DDR3 of RAM.

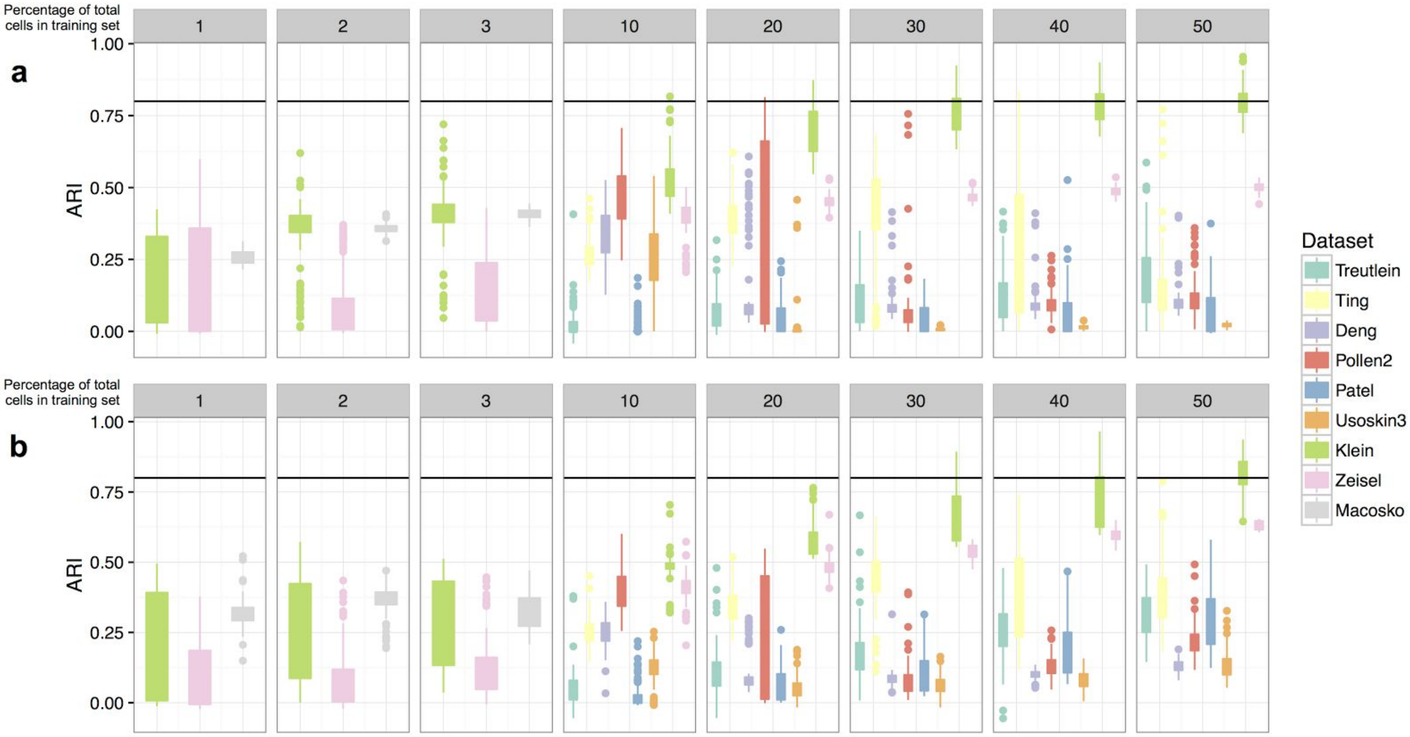

Figure S7. **Results for the hybrid clustering approach using a polynomial kernel.** The black line corresponds to ARI=0.8. Numbers in grey boxes correspond to the number of training cells as % of *N.* (**a**) ARI levels for SVM prediction when reference labels (provided by the authors) are used for training. (**b**) ARI levels for SVM prediction when labels calculated by SC3 are used for training. Dots represent outliers that are higher than the highest value (or lower than the lowest value) within 1.5 * IQR, where IQR is the inter-quartile range, or distance between the first and third quartiles.

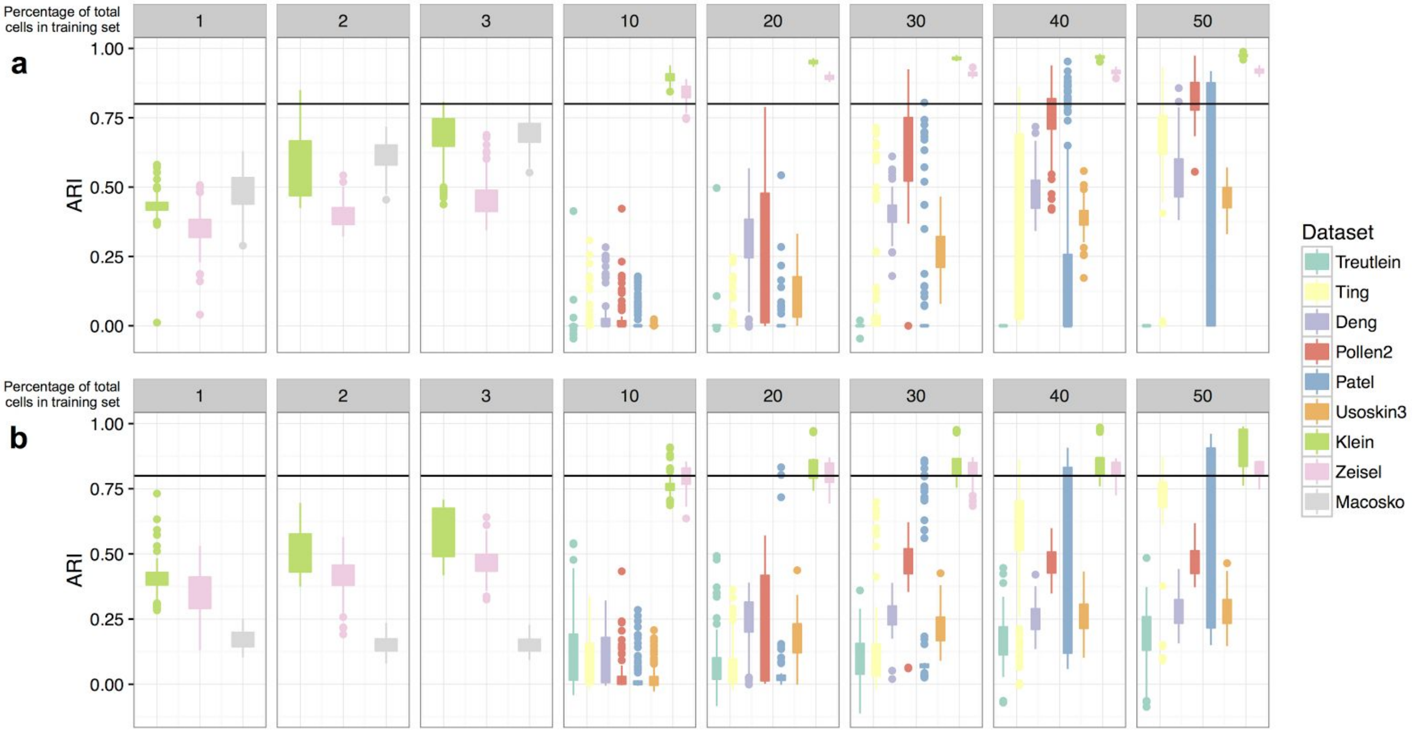

Figure S8. **Results for the hybrid clustering approach using a radial kernel.** The black line corresponds to ARI=0.8. Numbers in grey boxes correspond to the number of training cells as % of *N*. (**a**) ARI levels for SVM prediction when reference labels (provided by the authors) are used for training. (**b**) ARI levels for SVM prediction when labels calculated by SC3 are used for training. Dots represent outliers that are higher than the highest value (or lower than the lowest value) within 1.5 * IQR, where IQR is the inter-quartile range, or distance between the first and third quartiles.

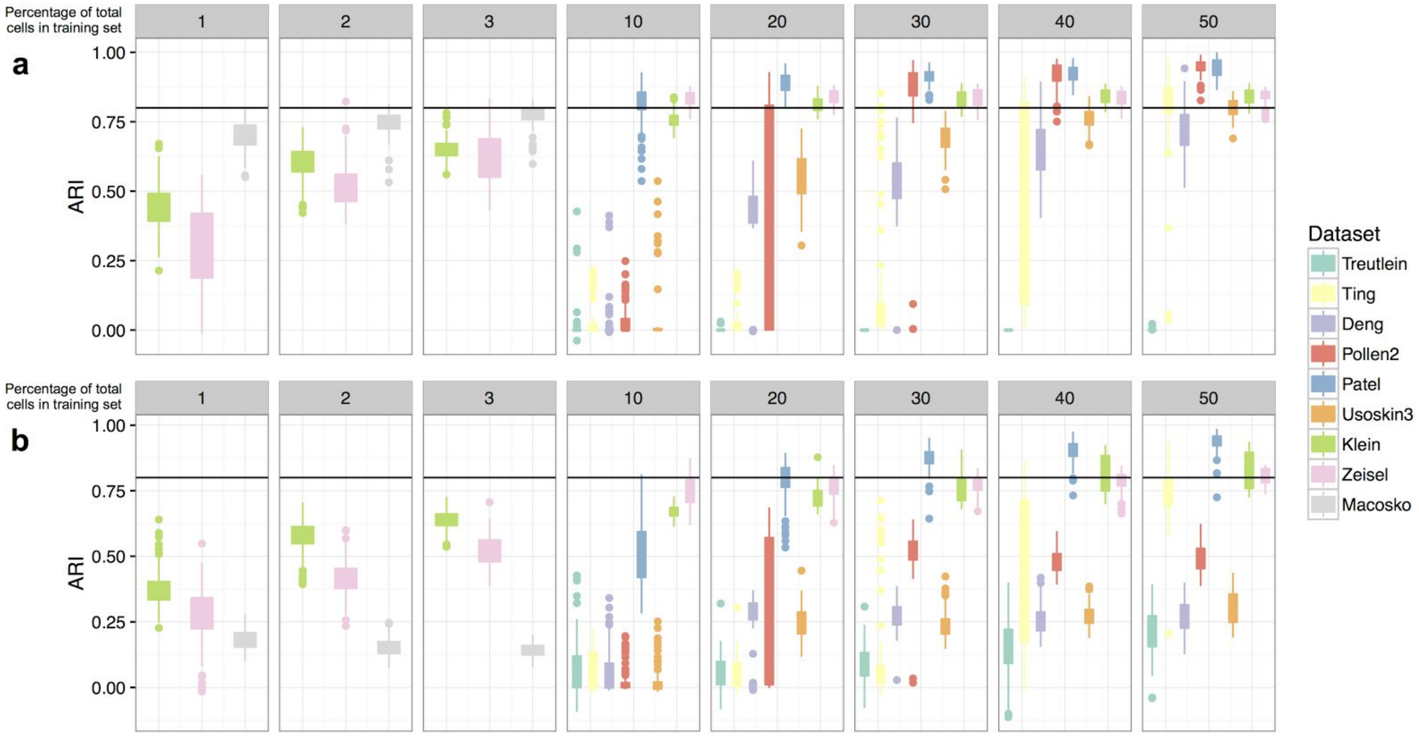

Figure S9. **Results for the hybrid clustering approach using a sigmoid kernel.** The black line corresponds to ARI=0.8. Numbers in grey boxes correspond to the number of training cells as % of *N.* (**a**) ARI levels for SVM prediction when reference labels (provided by the authors) are used for training. (**b**) ARI levels for SVM prediction when labels calculated by SC3 are used for training. Dots represent outliers that are higher than the highest value (or lower than the lowest value) within 1.5 * IQR, where IQR is the inter-quartile range, or distance between the first and third quartiles.

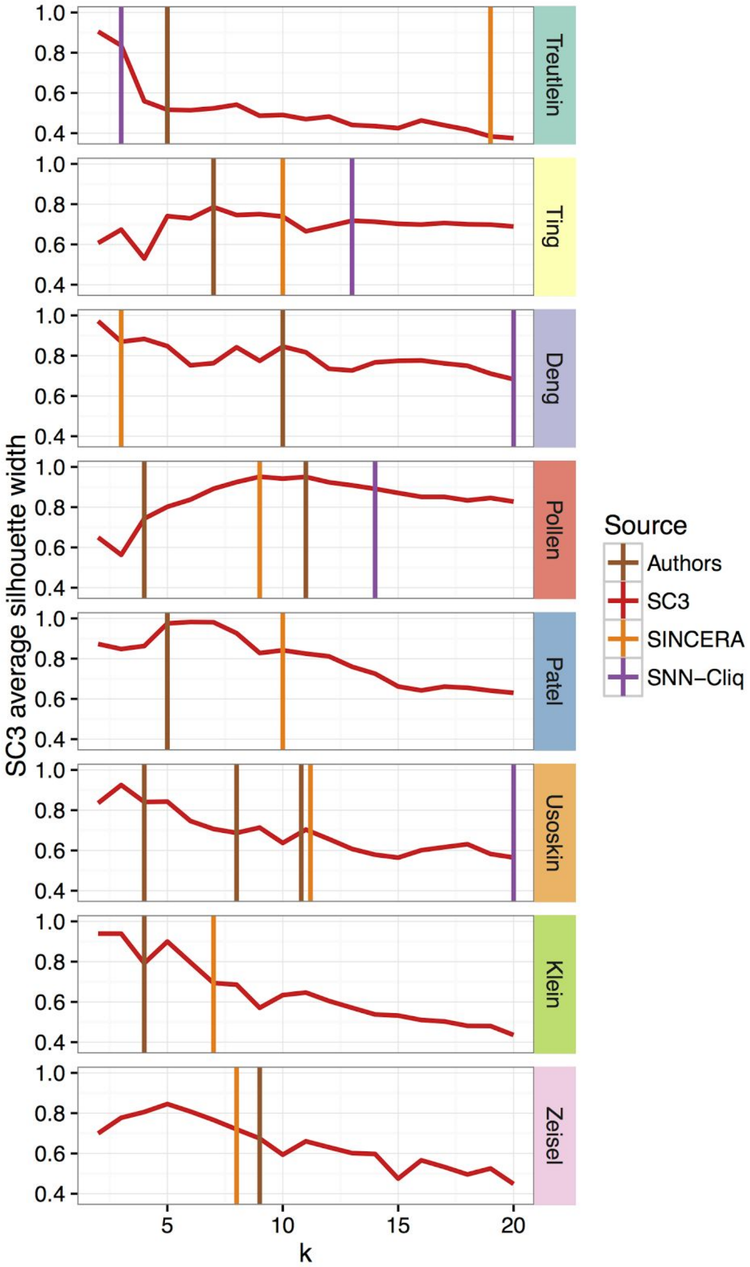

Figure S10. **Average silhouette index values for *k* in the range [2, 20] obtained by SC3.** The values of *k* suggested by the original authors, SNN-Cliq (for Patel, Klein and Zeisel the suggestion is >20) and SINCERA are indicated as vertical lines. Here only one plot per dataset is required to capture the different hierarchies.

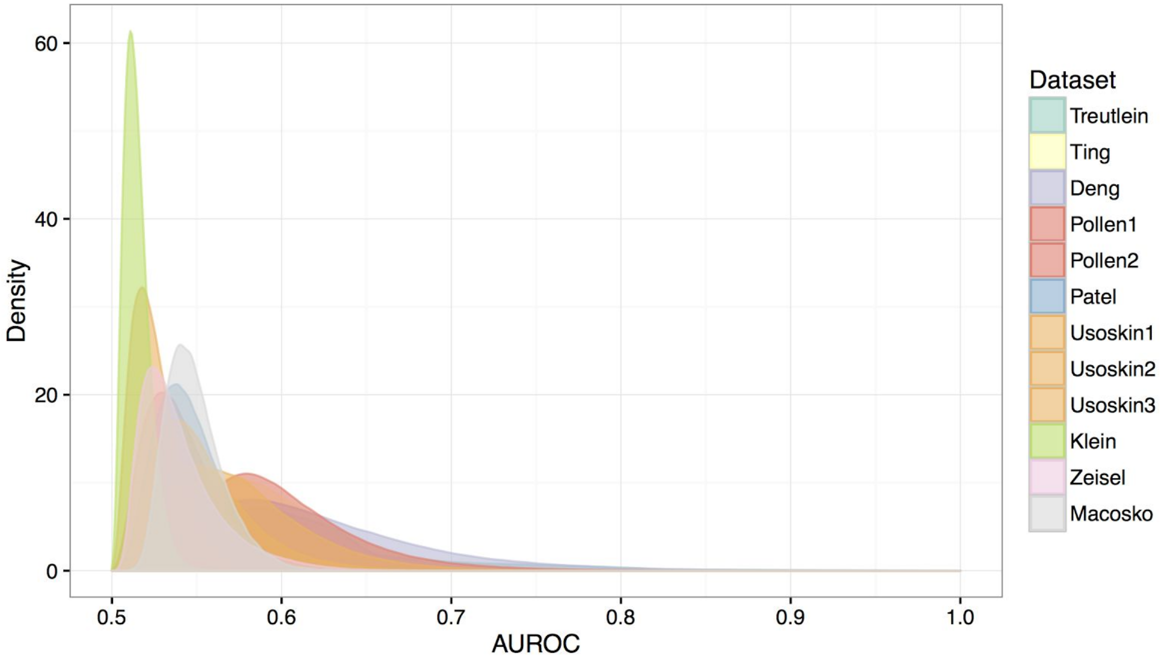

Figure S11. Density of distributions of AUROC obtained from merging of 100 calculations of marker genes using randomly shuffled assignments of reference labels (provided by the authors, see Methods).

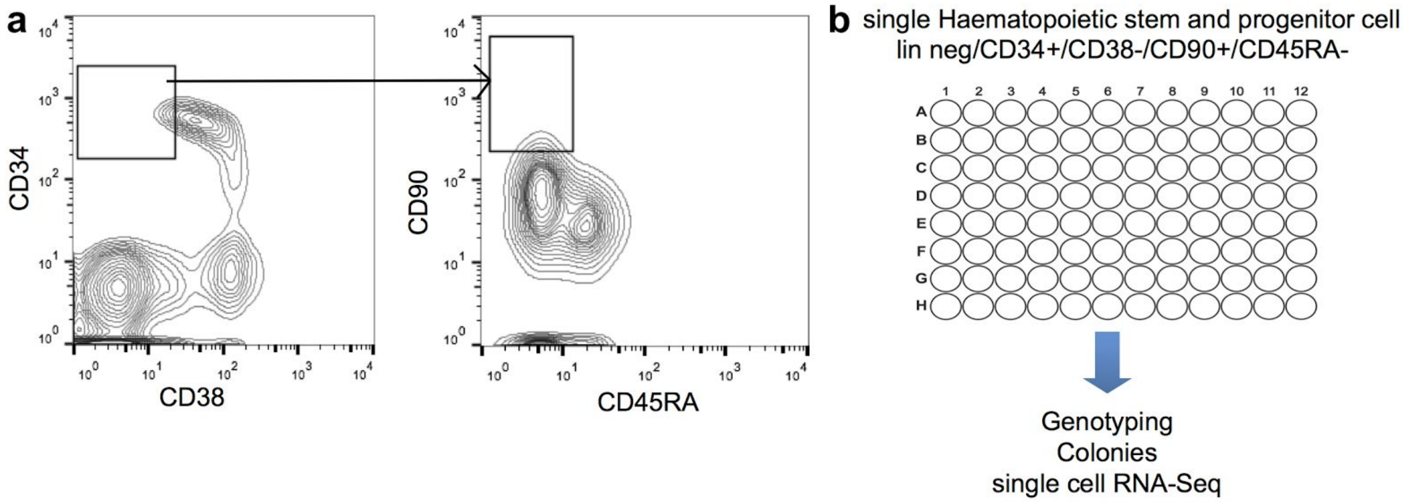

Figure S12. **Cell sorting procedure for patients.** (**a**) Contour plots describing the sorting strategy for isolating HSCs in patient 2 (the same was done for patient 1). CD34, CD38, CD90 and CD45RA expression is displayed using a log scale. (**b**) Lineage negative, CD34+/CD38-/CD90+/CD45RA- single cells were sorted into individual wells for scRNA-Seq or colony growth in cytokine cocktail allowing progenitor cell expansion.

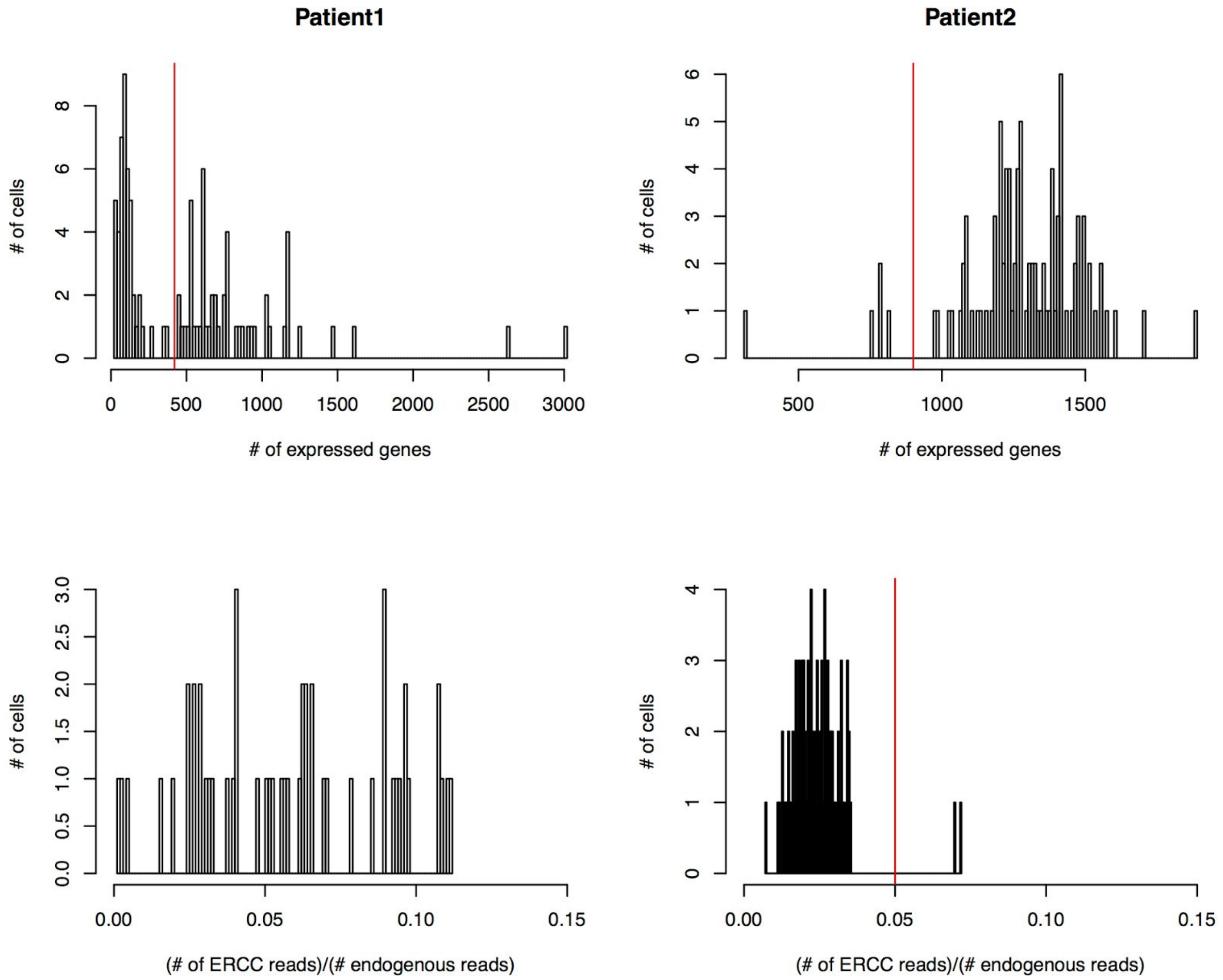

Figure S13. **Quality control of cells in the patient data.** (**a**) Number of cells with a given number of expressed genes in each patient. Cells on the left side of the red line were removed from further analysis as lowly expressed. (**b**) Number of cells with a given (# of ERCC reads)/(# endogenous reads) ratio in each patient. Cells on the right side of the red line were removed from further analysis as outliers.

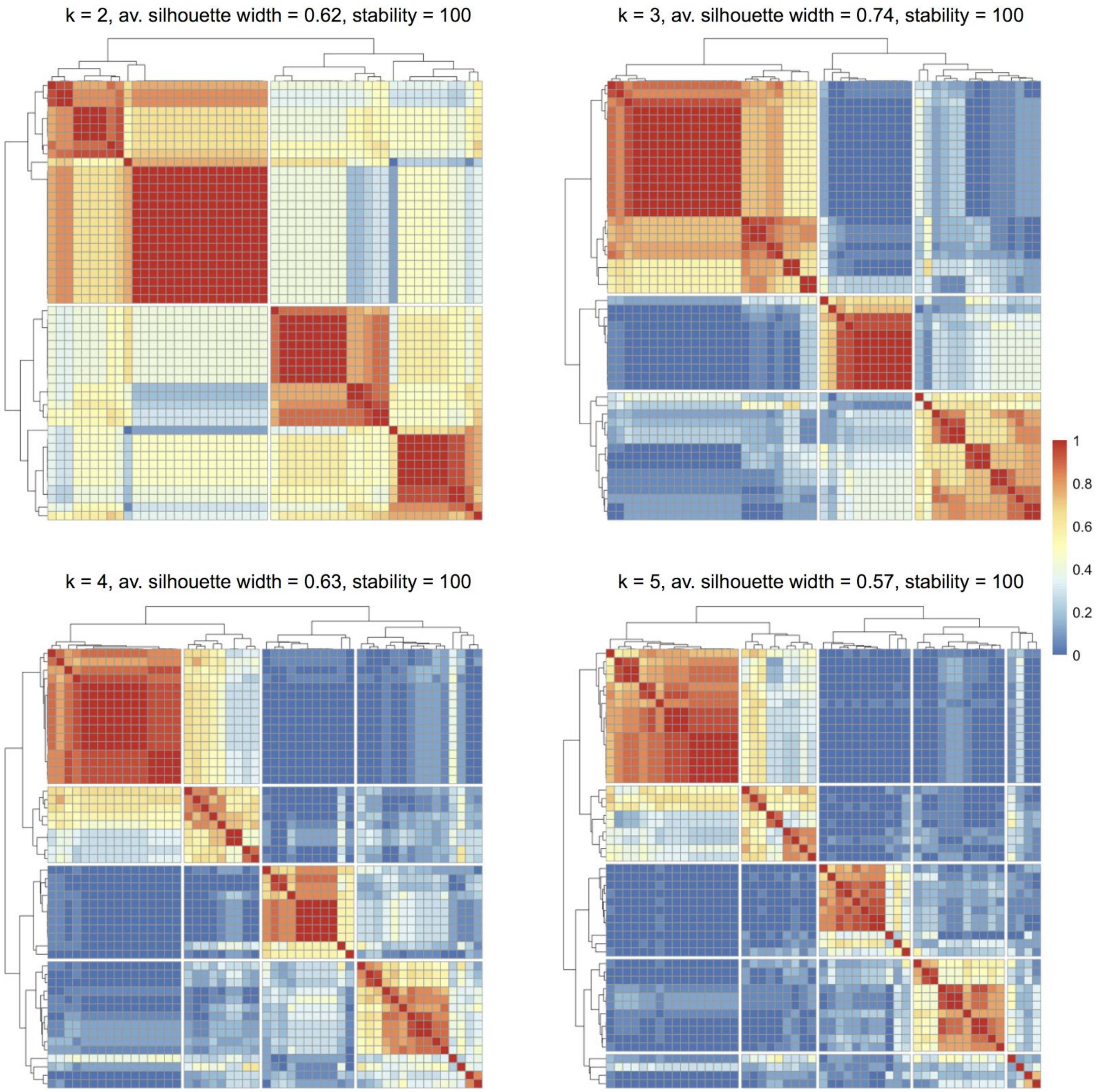

Figure S14. **Clustering of scRNA-seq data from patient 1.** Consensus matrices corresponding to different values of *k*. For average silhouette width and stability see Methods.

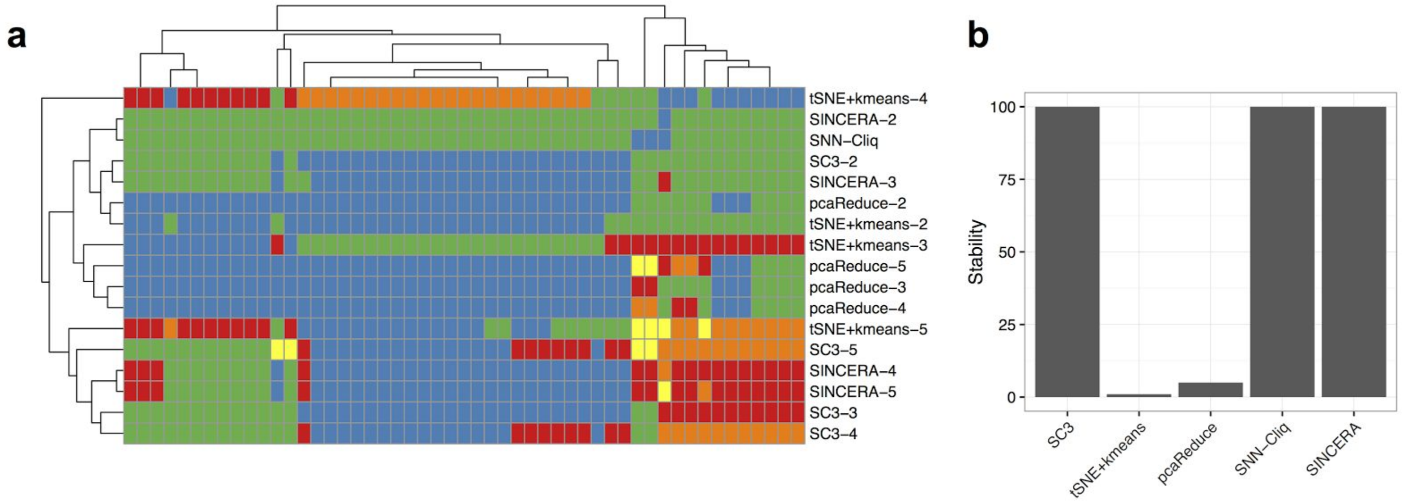

Figure S15. **Comparison of SC3 clustering to other methods based on scRNA-seq data from patient 1.** (**a**) Clustering solutions of all considered methods for *k* = 2 to *k* = 5. Colours correspond to different clusters. In each row colours of the clusters are assigned arbitrarily. (**b**) The stability (Methods) of the clusterings presented in (**a**).

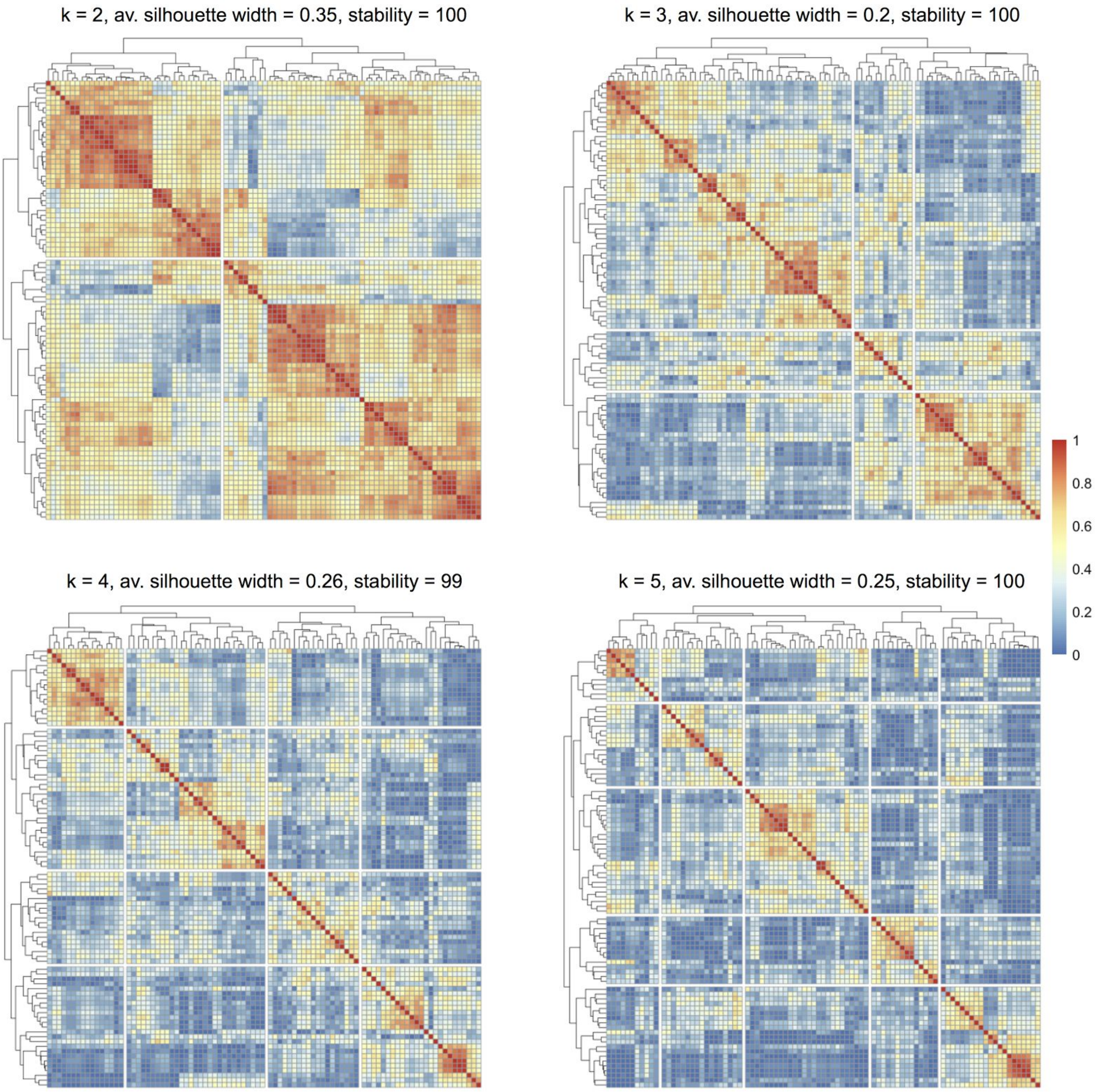

Figure S16. **Clustering of scRNA-seq data from patient 2.** Consensus matrices corresponding to different values of *k*. For average silhouette width and stability see Methods.

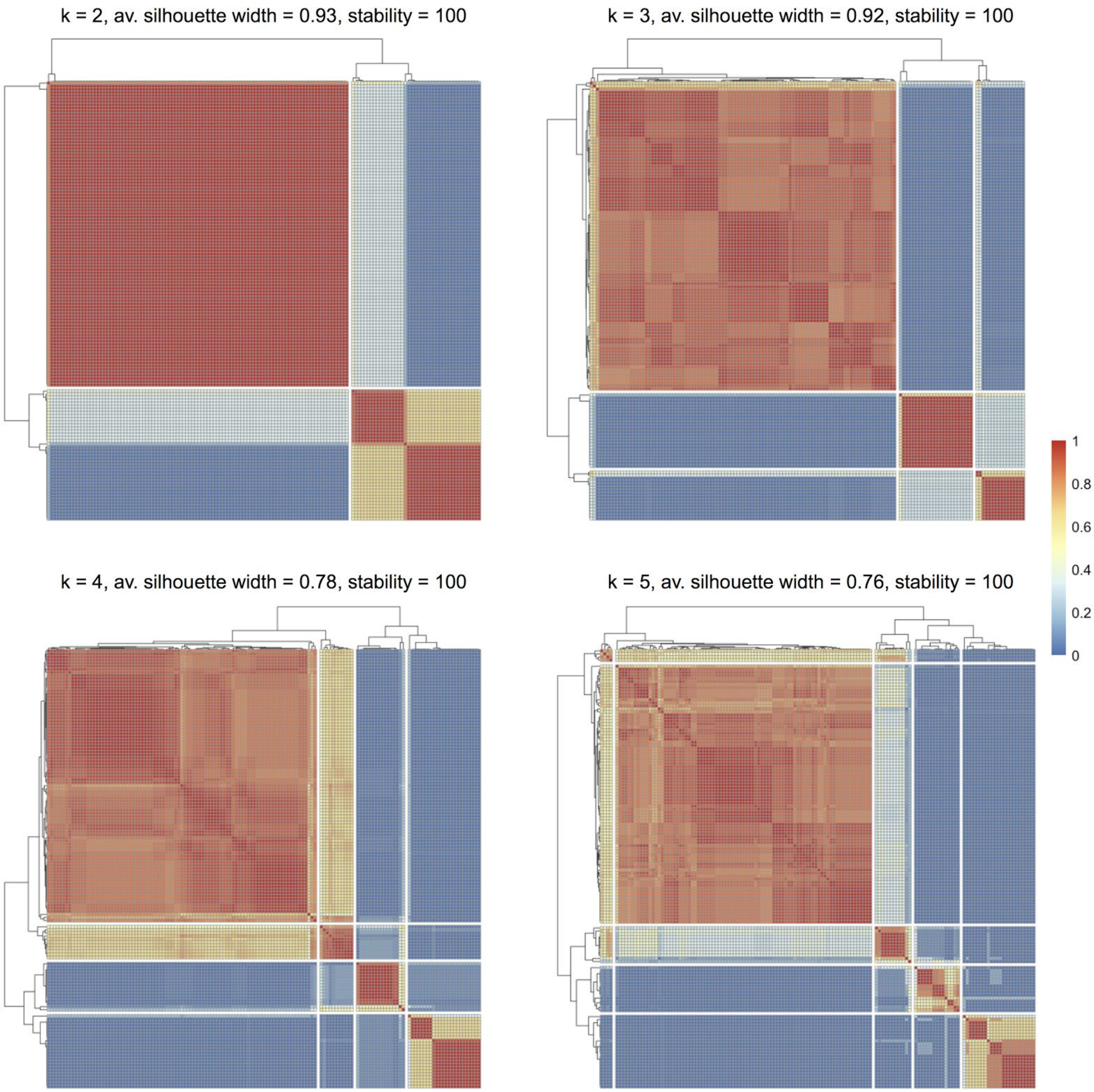

Figure S17. **Clustering of scRNA-seq data from combined patient 1 and patient 2 datasets.** Consensus matrices corresponding to different values of *k*. For average silhouette width and stability see Methods.

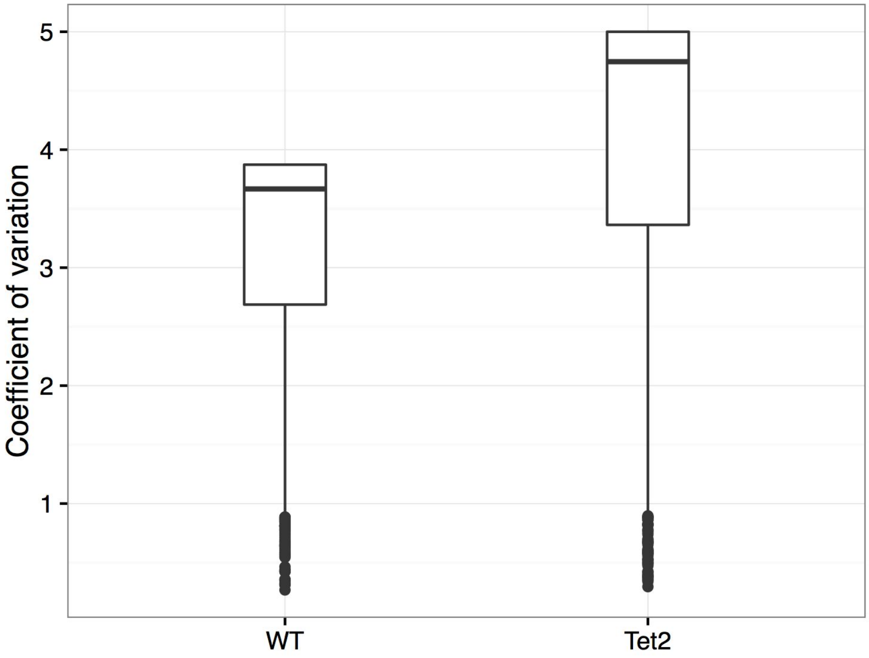

Figure S18. Comparison of the coefficient of variation of gene expression in Tet2 and WT subclones of patient 1.

Dataset	99% quantile of AUROC density distribution
Treutlein	0.83
Deng	0.82
Pollen2	0.74
Ting	0.72
Usoskin3	0.7
Usoskin2	0.65
Zeisel	0.62
Pollen1	0.61
Patel	0.6
Macosko	0.6
Usoskin1	0.57
Klein	0.54

Table S1. 99% quantiles of AUROC density distributions (Fig. S18) obtained from merging of 100 calculations of marker genes using randomly shuffled assignments of reference labels (provided by the authors, see Methods).

Table S2. SC3 output file containing all 3,500 identified marker genes from the Deng dataset.

Driver Mutations	patient ID	Gender	Diagnosis	Age at diagnosis	Disease duration at assay (years)	Therapy at assay
Tet2 c.3120_3121het _insA Jak2V617F	1	M	ET	75	12	hydroxycarbamide
Tet2 c.5447 T>A p.L1816X Jak2V617F	2	F	post-ET MF	78	14	pacritinib

Table S3. A summary of the patient information. ET, essential thrombocytosis; MF, myelofibrosis

Table S4. Marker genes for the comparison of patient 1 & 2

Pathway Name	Adjusted p-value	Gene Symbol	Gene description
RNA transport	0.013		
		EIF3F	eukaryotic translation initiation factor 3, subunit F
		EIF3C	eukaryotic translation initiation factor 3, subunit C
		NXF2B	nuclear RNA export factor 2B
		CASC3	cancer susceptibility candidate 3
		NUP54	nucleoporin 54kDa
Purine metabolism	0.013		
		ENTPD3	ectonucleoside triphosphate diphosphohydrolase 3
		NPR1	natriuretic peptide receptor A/guanylate cyclase A (atrionatriuretic peptide receptor A)
		AMPD3	adenosine monophosphate deaminase 3
		PDE1B	phosphodiesterase 1B, calmodulin-dependent
		PDE8B	phosphodiesterase 8B
Metabolic pathways	0.013		
		INPP5K	inositol polyphosphate-5-phosphatase K
		NDUFV2	NADH dehydrogenase (ubiquinone) flavoprotein 2, 24kDa
		GRHPR	glyoxylate reductase/hydroxypyruvate reductase
		AMPD3	adenosine monophosphate deaminase 3
		GFPT2	glutamine-fructose-6-phosphate transaminase 2
		PGP	phosphoglycolate phosphatase
		DGKZ	diacylglycerol kinase, zeta
		NDUFB6	NADH dehydrogenase (ubiquinone) 1 beta subcomplex, 6, 17kDa
		PIGP	phosphatidylinositol glycan anchor biosynthesis, class P
		CYP3A43	cytochrome P450, family 3, subfamily A, polypeptide 43
		MGAM	maltase-glucoamylase (alpha-glucosidase)
		TALDO1	transaldolase 1
		NOS3	nitric oxide synthase 3 (endothelial cell)
Alzheimer's disease	0.013		
		NDUFV2	NADH dehydrogenase (ubiquinone) flavoprotein 2, 24kDa
		ADAM17	ADAM metallopeptidase domain 17
		CACNA1C	calcium channel, voltage-dependent, L type, alpha 1C subunit
		NDUFB6	NADH dehydrogenase (ubiquinone) 1 beta subcomplex, 6, 17kDa
		LRP1	low density lipoprotein receptor-related protein 1
Amoebiasis	0.013		
		SERPINB6	serpin peptidase inhibitor, clade B (ovalbumin), member 6
		COL5A3	collagen, type V, alpha 3
		SERPINB1	serpin peptidase inhibitor, clade B (ovalbumin), member 1
		GNA15	guanine nucleotide binding protein (G protein), alpha 15 (Gq class)
Chemokine signaling pathway	0.013		
		GNB3	guanine nucleotide binding protein (G protein), beta polypeptide 3
		SHC1	SHC (Src homology 2 domain containing) transforming protein 1
		ELMO1	engulfment and cell motility 1
		NCF1	neutrophil cytosolic factor 1
		VAV2	vav 2 guanine nucleotide exchange factor
Retinol metabolism	0.0159		
		CYP3A43	cytochrome P450, family 3, subfamily A, polypeptide 43
		RETSAT	retinol saturase (all-trans-retinol 13,14-reductase)
		LRAT	lecithin retinol acyltransferase (phosphatidylcholine--retinol O-acyltransferase)
Glyoxylate and dicarboxylate metabolism	0.0159		
		GRHPR	glyoxylate reductase/hydroxypyruvate reductase
		PGP	phosphoglycolate phosphatase
mRNA surveillance pathway	0.0287		
		SMG5	smg-5 homolog, nonsense mediated mRNA decay factor (C. elegans)
		NXF2B	nuclear RNA export factor 2B
		CASC3	cancer susceptibility candidate 3
Calcium signaling pathway	0.0353		
		CACNA1C	calcium channel, voltage-dependent, L type, alpha 1C subunit
		PDE1B	phosphodiesterase 1B, calmodulin-dependent
		GNA15	guanine nucleotide binding protein (G protein), alpha 15 (Gq class)
		NOS3	nitric oxide synthase 3 (endothelial cell)
Focal adhesion	0.0485		
		COL5A3	collagen, type V, alpha 3
		SHC1	SHC (Src homology 2 domain containing) transforming protein 1

		FLT4	fms-related tyrosine kinase 4
		VAV2	vav 2 guanine nucleotide exchange factor
Leukocyte transendothelial migration	0.053		
		MMP2	matrix metallopeptidase 2 (gelatinase A, 72kDa gelatinase, 72kDa type IV collagenase)
		NCF1	neutrophil cytosolic factor 1
		VAV2	vav 2 guanine nucleotide exchange factor
Nucleotide excision repair	0.053		
		LIG1	ligase I, DNA, ATP-dependent
		CUL4B	cullin 4B
Notch signaling pathway	0.0559		
		ADAM17	ADAM metallopeptidase domain 17
		NOTCH4	notch 4
Leishmaniasis	0.0979		
		JAK1	Janus kinase 1
		NCF1	neutrophil cytosolic factor 1
Bacterial invasion of epithelial cells	0.0979		
		SHC1	SHC (Src homology 2 domain containing) transforming protein 1
		ELMO1	engulfment and cell motility 1
Adherens junction	0.0979		
		ACP1	acid phosphatase 1, soluble
		CSNK2B	casein kinase 2, beta polypeptide
Protein processing in endoplasmic reticulum	0.0979		
		UBQLN4	ubiquilin 4
		RRBP1	ribosome binding protein 1 homolog 180kDa (dog)
		HSPA1B	heat shock 70kDa protein 1B

Table S5. Enriched pathways from KEGG analysis for the marker genes from Table S4.

Pathway Name	Adjusted p-value	Gene Symbol	Gene description
AGE-RAGE pathway	0.001		
		MMP7	matrix metallopeptidase 7 (matrilysin, uterine)
		SHC1	SHC (Src homology 2 domain containing) transforming protein 1
		NCF1	neutrophil cytosolic factor 1
		MMP2	matrix metallopeptidase 2 (gelatinase A, 72kDa gelatinase, 72kDa type IV collagenase)
		NOS3	nitric oxide synthase 3 (endothelial cell)
G Protein Signaling Pathways	0.0013		
		GNB3	guanine nucleotide binding protein (G protein), beta polypeptide 3
		PDE1B	phosphodiesterase 1B, calmodulin-dependent
		GNA15	guanine nucleotide binding protein (G protein), alpha 15 (Gq class)
		AKAP11	A kinase (PRKA) anchor protein 11
		PDE8B	phosphodiesterase 8B
Nicotine Activity on Chromaffin Cells	0.0013		
		CACNA1C	calcium channel, voltage-dependent, L type, alpha 1C subunit
		CHRNA3	cholinergic receptor, nicotinic, alpha 3 (neuronal)
Matrix Metalloproteinases	0.004		
		BSG	basigin (Ok blood group)
		MMP7	matrix metallopeptidase 7 (matrilysin, uterine)
		MMP2	matrix metallopeptidase 2 (gelatinase A, 72kDa gelatinase, 72kDa type IV collagenase)
Leptin signaling pathway	0.004		
		CISH	cytokine inducible SH2-containing protein
		KPNA4	karyopherin alpha 4 (importin alpha 3)
		JAK1	Janus kinase 1
		SHC1	SHC (Src homology 2 domain containing) transforming protein 1
Prolactin Signaling Pathway	0.0053		
		CISH	cytokine inducible SH2-containing protein
		JAK1	Janus kinase 1
		SHC1	SHC (Src homology 2 domain containing) transforming protein 1
		VAV2	vav 2 guanine nucleotide exchange factor
Notch Signaling Pathway	0.0063		
		INPP5K	inositol polyphosphate-5-phosphatase K
		ADAM17	ADAM metallopeptidase domain 17
		NOTCH4	notch 4
IL-2 Signaling pathway	0.009		
		CISH	cytokine inducible SH2-containing protein
		JAK1	Janus kinase 1
		SHC1	SHC (Src homology 2 domain containing) transforming protein 1
angiogenesis overview	0.0236		
		SHC1	SHC (Src homology 2 domain containing) transforming protein 1
		MMP2	matrix metallopeptidase 2 (gelatinase A, 72kDa gelatinase, 72kDa type IV collagenase)
		NOS3	nitric oxide synthase 3 (endothelial cell)
Alzheimers Disease	0.0236		
		ADAM17	ADAM metallopeptidase domain 17
		CACNA1C	calcium channel, voltage-dependent, L type, alpha 1C subunit
		LRP1	low density lipoprotein receptor-related protein 1
Oncostatin M Signaling Pathway	0.0247		
		CISH	cytokine inducible SH2-containing protein
		JAK1	Janus kinase 1
		SHC1	SHC (Src homology 2 domain containing) transforming protein 1
EPO Receptor Signaling	0.0335		
		CISH	cytokine inducible SH2-containing protein
		SHC1	SHC (Src homology 2 domain containing) transforming protein 1
Ovarian Infertility Genes	0.0335		
		MSH5	mutS homolog 5 (E. coli)
		ZP3	zona pellucida glycoprotein 3 (sperm receptor)
Vitamin A and carotenoid metabolism	0.0515		
		RETSAT	retinol saturase (all-trans-retinol 13,14-reductase)

		LRAT	lecithin retinol acyltransferase (phosphatidylcholine--retinol O-acyltransferase)
TSLP Signaling Pathway	0.0515		
		CISH	cytokine inducible SH2-containing protein
		JAK1	Janus kinase 1
IL-5 signaling pathway	0.0515		
		JAK1	Janus kinase 1
		SHC1	SHC (Src homology 2 domain containing) transforming protein 1
Translation Factors	0.0522		
		EIF3F	eukaryotic translation initiation factor 3, subunit F
		EIF3C	eukaryotic translation initiation factor 3, subunit C
Endothelin	0.0545		
		GNA15	guanine nucleotide binding protein (G protein), alpha 15 (Gq class)
		NOS3	nitric oxide synthase 3 (endothelial cell)
Oxidative phosphorylation	0.0545		
		NDUFV2	NADH dehydrogenase (ubiquinone) flavoprotein 2, 24kDa
		NDUFB6	NADH dehydrogenase (ubiquinone) 1 beta subcomplex, 6, 17kDa
IL-4 signaling pathway	0.0545		
		JAK1	Janus kinase 1
		SHC1	SHC (Src homology 2 domain containing) transforming protein 1
IL-6 signaling pathway	0.0545		
		JAK1	Janus kinase 1
		SHC1	SHC (Src homology 2 domain containing) transforming protein 1
IL-3 Signaling Pathway	0.0545		
		JAK1	Janus kinase 1
		SHC1	SHC (Src homology 2 domain containing) transforming protein 1
Myometrial Relaxation and Contraction Pathways	0.0645		
		GNB3	guanine nucleotide binding protein (G protein), beta polypeptide 3
		DGKZ	diacylglycerol kinase, zeta
		NOS3	nitric oxide synthase 3 (endothelial cell)
Notch Signaling Pathway	0.0663		
		ADAM17	ADAM metallopeptidase domain 17
		NOTCH4	notch 4
EGF-EGFR Signaling Pathway	0.0688		
		JAK1	Janus kinase 1
		SHC1	SHC (Src homology 2 domain containing) transforming protein 1
		VAV2	vav 2 guanine nucleotide exchange factor
Integrated Pancreatic Cancer Pathway	0.0749		
		LIG1	ligase I, DNA, ATP-dependent
		RBL1	retinoblastoma-like 1 (p107)
		CD82	CD82 molecule

Table S6. Enriched pathways from Wikipathways analysis for the marker genes from Table S4.