

Supplementary Note

Demonstration on Demuxlet paper data

In order to demonstrate souporecell on an external and widely used benchmark dataset, we downloaded the three overlapping mixtures from the demuxlet paper⁶. Sample A contains a mixture of four donors' PBMCs, Sample B contains a mixture of four different donors' PBMCs, and Sample C contains a mixture of all 8 donors' PBMCs. We synthetically combined this data into a single dataset and clustered with souporecell. Supp Fig 4a shows that the resulting clusters either contain cells from Sample A or Sample B, but not both as is expected from this experimental setup. We also show that the first cluster of the doublet assignments are also largely consistent with this experimental design (Supp Fig 4b).

Deconvolution of overlapping mixtures

To enable identification of which cluster is which individual using the overlapping mixture experimental design outlined in Table 1, we provide a tool `shared_samples.py` which takes as input two souporecell output directories and the number of samples which are shared. It compares the sum of squared differences of the allele fraction of confident (>95% confident genotype call in all clusters) shared variant calls between clusters in the two experiments and outputs the best matches for the number of shared samples. We tested this using multiple different runs of souporecell on the synthetic mixture of 5 HipSci cell lines with 6% doublets and 5% ambient RNA and gave both as input to the `shared_samples.py` tool and it correctly assigned the clusters in one run to the clusters in the second experiment which corresponded to the same samples. We also ran souporecell on the three demuxlet datasets separately and ran the `shared_samples.py` tool on Sample A vs Sample C and Sample B vs Sample C and it confidently identified the non-overlapping clusters in Sample C which correspond to A and B.

Demonstration on 21 donor sample

We demonstrate that souporecell is capable of demultiplexing many donors by creating a synthetic mixture of 21 different individuals, which given the current recommendations from 10x on cells per run would be a high-end number of donors to multiplex. To generate this 21-donor mix, we used the 5 HipSci samples described in Fig 2 and added to them 16 PBMC samples obtained from the Human Cell Atlas Census of Immune Cells²⁵. From each dataset we randomly selected 1000 cells with at least 4000 UMIs and simulated 10% doublets and 2.5% ambient RNA by altering the cell barcodes, as described above. We clustered these with souporecell and the software correctly identifies 1690 of the 2100 synthetic doublets. A further 69 cells were unassigned, and in total we have an ARI of 0.95. Excluding all doublets the ARI is 0.98. We find a total of 134/16800 singletons misassigned where 129 of them are CB8 cells assigned to the CB3 cluster. We show later that this is likely because the CB8 sample is contaminated by another (non CB3) donor. Supp Figure 5 shows the UMAP projection of the normalized cluster log likelihood matrix. It is clear that souporecell is able to handle at least 21 distinct donors and accurately assign cluster identities to the majority of cells.

Because this error type accounted for >95% of singleton errors, we suspected this may be due to contamination. We repeated this experiment with several of the replicates of the CB8 donor and found consistent results. We then made a synthetic mixture of CB3 and CB8 in order to determine if this was due to the large number of donors and it was not. We still found that roughly 20% of CB8 cells would cluster with CB3, but if given 3 clusters, all of those cells formed their own cluster. This made us suspect

that the CB8 sample was contaminated with cells from a different (non CB3) donor. We created an elbow plot and a PCA of the normalized cell x cluster log likelihood matrix, and both support this hypothesis (Supp Fig 6). In communications with the Broad Institute, they confirm that the genomics platform reported 19% contamination with this sample, perhaps from maternal blood as this was a cord blood sample.

Downsampling experiments for cells and UMIs

In order to explore the regime for which it is still possible to accurately demultiplex mixed samples, we used our synthetically mixed 5 HipSci samples and downsampled UMIs (Supp Fig 6a) and cell (Supp Fig 6b) and report the ARI versus the ground truth. We find that while overall clustering remains good, cell assignment accuracy decreases below 800 median UMI per cell and that accuracy remains high down to an average of 40 cells per cluster (Supp Fig 6b).

Software versions

Initial data analysis was done by cellranger v2.1.1 as input to souporecell, demuxlet, vireo, and scSplit

Souporcell

<https://github.com/wheaton5/souporcell>

We provide a singularity container build encapsulating all requirements for souporecell as well as a singularity definition file to recreate this container. The following are the software versions for all software excluding software required to build the system.

Freebayes - v1.3.1-17-gaa2ace8

Pyvcf - 0.6.7

Pysam - 0.15.3

Numpy - 1.17.0

Scipy - 1.3.0

Tensorflow - 1.14.0

Pystan - 2.17.1.0

Pyfasta - 0.5.2

Htslib - 1.9

Samtools - 1.9

Bcftools - 1.9

Vartrix - 1.1.3

Minimap2 - 2.7-r654

Bedtools - 2.28.0

Demuxlet

<https://github.com/statgen/demuxlet>

git hash 85dca0a4d648d18e6b240a2298672394fe10c6e6

Vireo

Cardelino R package version 0.3.8

cellSNP (<https://github.com/huangyh09/cellSNP> version 0.1.6)

scSplit

Freebayes - v1.3.1-17-gaa2ace8

<https://github.com/jon-xu/scSplit> git commit hash 52face6a4c1b291651bdf9b56328d168c7cb1fa6

Supp Table 1: Sample metrics.

Sample	Cells	Median UMI/cell	NOTES
Mixture1	4925	25155	HipSci
Mixture2	4832	24913	HipSci
Mixture3	5144	24807	HipSci
euts	4859	25088	HipSci
nufh	6781	17254	HipSci
babz	12299	11343	HipSci
oaqd	7107	18315	HipSci
ieki	6586	21969	HipSci
SyntheticMixture	7073	18793	synthetic HipSci mixture (5% ambient RNA, 6% doublets)
Placenta1	3835	18415	Maternal Fetal
Placenta2	3968	18046	Maternal Fetal
Decidua1	2119	23075	Maternal Fetal
Plasmodium1	2608	995	<i>Plasmodium falciparum</i>
Plasmodium2	1893	762	<i>Plasmodium falciparum</i>
Plasmodium3	2293	1126	<i>Plasmodium falciparum</i>

Number of cells and median UMI/cell for each sample including one representative synthetic mixture.

Supp Table 2: Hipscli clustering.

	method	mixture1	mixture2	mixture3	synthetic mixture
ARI vs demuxlet single best (excluding doublets called by each tool)	souporcell	1	1	1	n/a
	vireo	1	1	1	n/a
	scSplit	0.97	0.97	0.98	n/a
ARI vs truth (including doublets)	souporcell	n/a	n/a	n/a	0.99
	demuxlet	n/a	n/a	n/a	0.76
	vireo	n/a	n/a	n/a	0.98
	scSplit	n/a	n/a	n/a	0.94
ambient RNA	souporcell	2.70%	3%	2.70%	6.70%
	truth	n/a	n/a	n/a	5%
doublets	souporcell	338 (6.8%)	313 (6.4%)	323 (6.2%)	420 (6.0%)
	demuxlet	1537 (31)	1811 (37.5%)	1596 (31.2%)	1329 (18.7%)
	vireo	306 (6.2%)	287 (5.9%)	285 (5.5%)	311 (4.4%)
	scSplit	127 (2.6%)	107 (2.2)	102 (2%)	89 (1.2%)

	truth	n/a	n/a	n/a	451 (6.4%)
euts	souporcell	1033	1068	1177	829
	demuxlet	790	743	922	799
	vireo	1039	1074	1185	876
	scSplit	1076	1102	1211	856
	truth	n/a	n/a	n/a	817
nufh	souporcell	659	663	691	1192
	demuxlet	317	264	330	1037
	vireo	660	665	694	1211
	scSplit	707	721	766	1271
	truth	n/a	n/a	n/a	1183
babz	souporcell	739	727	731	2239
	demuxlet	522	459	504	1591
	vireo	742	734	740	2241
	scSplit	818	742	806	2263
	truth	n/a	n/a	n/a	2242
oaqd	souporcell	1680	1469	1592	1250
	demuxlet	1341	1172	1345	1227
	vireo	1590	1481	1609	1261
	scSplit	1587	1501	1600	1353
	truth	n/a	n/a	n/a	1247
ieki	souporcell	586	591	630	1138
	demuxlet	418	383	447	1090
	vireo	586	590	630	1160
	scSplit	610	659	659	1241
	truth	n/a	n/a	n/a	1133
unassigned	souporcell	2	1	2	5
	demuxlet	0	0	0	0
	vireo	2	1	1	13
	scSplit	0	0	0	0

Evaluation of clustering metrics for each tool on HipSci experimental mixtures as well as one representative synthetic mixture with 5% ambient RNA and 6% doublets.

Supp Table 3: Maternal/Fetal clustering.

	method	placenta1	placenta2	decidua1
--	--------	-----------	-----------	----------

ARI vs demuxlet single best (excluding doublets called by each tool)	souporcell	0.96	0.96	0.93
	vireo	0	0.18	0.3
	scSplit	0.03	0	0
ambient RNA	souporcell	4%	3.80%	6.80%
cells evaluated	souporcell	3835	3968	2119
	demuxlet	3805	3941	2115
	vireo	3835	3968	2119
	scSplit	3835	3968	2119
doublets	souporcell	49 (1.2%)	40 (1%)	3 (0.14%)
	demuxlet	189 (4.9%)	168 (4.2%)	112 (5.3%)
	vireo	660 (17.2%)	543 (13.7%)	229 (10.8%)
	scSplit	1705 (44%)	2297 (58%)	1145 (54%)
Maternal	souporcell	217	232	1944
	demuxlet	198	201	1863
	vireo	924	822	1240
	scSplit	1063	675	542
Fetal	souporcell	3316	3432	151
	demuxlet	3418	3572	140
	vireo	1427	1822	440
	scSplit	1067	996	432
unassigned	souporcell	253	274	21
	demuxlet	0	0	0
	vireo	824	781	210
	scSplit	0	0	0

A comparison of the clustering performance of each tool on the maternal/fetal data.

Supp Table 4: Plasmodium clustering.

	method	Plasmodium1	Plasmodium2	Plasmodium3*
ARI vs demuxlet single best (excluding doublets called by each tool)	souporcell	1	0.99	0.99
	vireo	0.95	0.51	0.24
	sc_split	0.6	0.46	0.45
ambient RNA	souporcell	2.4%	3.7%	6.8%

doublets	sourporcell	92 (3.5%)	175 (9.2%)	41 (1.8%)
	demuxlet	621 (23.8%)	297 (15.7%)	283 (12.3%)
	vireo	241 (9.2%)	340 (18%)	423 (18.4%)
	sc_split	338 (13%)	272 (14.4%)	463 (20%)
3D7	sourporcell	1079	705	1805
	demuxlet	951	703	1649
	vireo	1040	476	883 (in 2 clusters)
	sc_split	1131 (in 2 clusters)	784 (in 2 clusters)	1390 (in 2 cluster)
GB4	sourporcell	392	268	166 (including the SenTh strains)
	demuxlet	331	242	73
	vireo	336 (in 2 clusters)	198	n/a
	sc_split	389	281 (cluster contains both GB4 and SenTh015)	n/a
7G8	sourporcell	273	201	281
	demuxlet	223	183	229
	vireo	292	294 (cluster contains both 7G8 and SenTh028)	435 (including all non-3D7 strains)
	sc_split	237	311	440 (including all non-3D7 strains)
SenTh011	sourporcell	239	159	n/a
	demuxlet	128	139	11
	vireo	229	117	n/a
	sc_split	323	254 (cluster contains both Senth011 and a smattering of other cells)	n/a
SenTh015	sourporcell	215	149	n/a
	demuxlet	163	121	42
	vireo	no cluster represents	no cluster clearly	n/a

			represents this strain	
	sc_split	190		n/a
SenTh028	sourporcell	323	232	n/a
	demuxlet	190	208	5
	vireo	320	no cluster clearly represents this strain	n/a
	sc_split	no cluster represents	split across many clusters	n/a
unassigned	sourporcell	6	23	0
	demuxlet	0	0	0
	vireo	150	335	551
	sc_split	0	0	0

A comparison of the clustering performance of each tool on the Plasmodium data.