Assessing reliability of intra-tumor heterogeneity estimates from single sample whole exome sequencing data

Judith Abécassis[1,2,3], Anne-Sophie Hamy[1], Cécile Laurent[1], Benjamin Sadacca[1,4], Hélène Bonsang-Kitzis[1,5], Fabien Reyal[1,5], Jean-Philippe Vert[6,2*]

[1] Institut Curie, PSL Research University, Translational Research Department, INSERM, U932 Immunity and Cancer, Residual Tumor & Response to Treatment Laboratory (RT2Lab), F-75248, Paris, France
[2] MINES ParisTech, PSL Research University, CBIO - Centre for Computational Biology, F-75006 Paris, France
[3] Institut Curie, PSL Research University, INSERM, U900, F-75005 Paris, France
[4] Institut de Mathématiques de Toulouse, UMR5219 Université de Toulouse, CNRS UPS IMT, Toulouse, France
[5] Department of Surgery, Institut Curie, F-75248 Paris, France
[6] Google Brain, F-75009 Paris, France.

Abstract

Tumors are made of evolving and heterogeneous populations of cells which arise from successive appearance and expansion of subclonal populations, following acquisition of mutations conferring them a selective advantage. Those subclonal populations can be sensitive or resistant to different treatments, and provide information about tumor aetiology and future evolution. Hence, it is important to be able to assess the level of heterogeneity of tumors with high reliability for clinical applications.

In the past few years, a large number of methods have been proposed to estimate intra-tumor heterogeneity from whole exome sequencing (WES) data, but the accuracy and robustness of these methods on real data remains elusive. Here we systematically apply and compare 6 computational methods to estimate tumor heterogeneity on 1,697 WES samples from the cancer genome atlas (TCGA) covering 3 cancer types (breast invasive carcinoma, bladder urothelial carcinoma, and head and neck squamous cell carcinoma), and two distinct input mutation sets. We observe significant differences between the estimates produced by different methods, and identify several likely confounding factors in heterogeneity assessment for the different methods. We further show that the prognostic value of tumor heterogeneity for survival prediction is limited in those datasets, and find no evidence that it improves over prognosis based on other clinical variables.

In conclusion, heterogeneity inference from WES data on a single sample, and its use in cancer prognosis, should be considered with caution. Other approaches to assess intra-tumoral heterogeneity such as those based on multiple samples may be preferable for clinical applications.

Introduction

Cancer is characterized by the presence of cells growing and dividing without proper control. In the 1970s, Nowell and colleagues suggested that tumor cells follow evolutionary principles, as any other biological population able to acquire heritable transformations [1]. This evolutionary framework has proven very useful in deepening our understanding of cancer aetiology [2].

A consequence of this progressive accumulation of mutations is intra-tumor heterogeneity. Indeed, when a new mutation occurs in a tumor cell and provides an evolutionary advantage, this cell tends to have a higher

*Correspondance: jpvert@google.com

probability to survive and divide, hence seeding a new clonal population [3]. This new clone may supersede the whole tumor population, or coexist along it. This process results in a tumor made of a mosaic of clones. Next generation sequencing (NGS), in particular whole exome and whole genome sequencing (WES, WGS), can provide new insights into the heterogeneity and evolution of tumors. Indeed, early mutations shared among all cancer cells should be detected in more sequencing reads than mutations acquired later by only a fraction of the tumor cells. Thus it may be possible to estimate the intra-tumor heterogeneity (ITH) and reconstruct the clonal history of tumors from WES or WGS data, as reviewed by [3, 4, 5], and many computational methods have been developed for that purpose [6, 7, 8, 9]. We collectively refer to these methods as "ITH methods" in the following. Subclonal reconstruction from single cell sequencing has emerged as a new field, simplifying part of the inference problem, but raising other issues, related to technical limitations (high dropout rate) and high cost, possibly a limitation to the availability of large cohorts [10, 11, 12, 3].

Previous studies have reported that a large proportion of tumors are heterogeneous [13, 14, 15, 16], with various consequences for the patient. In particular, high ITH has been associated with treatment resistance and poor prognosis [17]. However, those results rely mostly on very detailed case studies involving only a small number of patients, with favorable experimental settings such as high coverage targeted sequencing on top of NGS, multiple sample collection (multi-site or longitudinal studies) [18, 19, 20] or even single-cell sequencing [21]. In the perspective of large-scale application in a clinical context, one needs to consider more accessible data with respect to cost and invasiveness for the patient, like moderate coverage WES on one sample per patient. A precise evaluation of existing ITH methods in this setting is needed to determine whether they allow us to find distinguishable patterns of heterogeneity and evolution of clinical relevance. Several large scale analyses have attempted to depict the evolutionary landscape of ITH in several cancer types [2], and to assess the prognostic power of ITH. In particular, using data from the cancer genome atlas (TCGA), a significant association between ITH and overall survival was found in at least one of the three studies [14, 13, 22] for 9 cancer types: breast invasive carcinoma (BRCA), kidney renal clear cell carcinoma (KIRC), brain lower grade glioma (LGG), prostate adenocarcinoma (PRAD), glioblastoma multiforme (GBM), head and neck squamous cell carcinoma (HNSC), ovarian serous cystadenocarcinoma (OV), uterine corpus endometrial carcinoma (UCEC), and colon adenocarcinoma (COAD). However, 5 of them were considered in another study with no significant result. In other cancer types, 2 studies consistently found no significant results for 3 cancer types: bladder urothelial carcinoma (BLAC), lung squamous cell carcinoma (LUSC) and stomach adenocarcinoma (STAD), and all 3 studies found no significant results for lung adenocarcinoma (LUAD) nor for skin cutaneous melanoma (SKM). A possible explanation for this discrepancy is that the studies base their analyses on different computational pipelines, from variant calling to ITH estimation, leading to different and sometimes contradictory results [22].

To clarify the robustness and consistency of different ITH methods, we perform a systematic benchmark of 18 computational pipelines for ITH estimates from a single WES sample per patient (combining 2 ways to call mutations, and 2 methods to assess copy number variations (only 3 out of 4 combinations were tested) with 6 ITH methods), using data from 1,697 patients with three types of cancer from the TCGA database (BRCA, BLCA, HNSC). We selected these cancer types following conclusions of Morris et al. [13], since HNSC, BRCA and BLCA are characterized by respectively high, intermediate and absence of prognostic power of ITH. We show that most existing ITH methods are very sensitive to the choice of mutations and copy number variations called, and that they can give very inconsistent results between each other. We highlight in particular that some methods are influenced by confounding factors such as tumor purity or mutation load. Finally, we show that although ITH measured by some computational pipelines have a weak prognostic power on some cancer types, the prognosis signal is not robust across methods and cancer types, and is confounded with informations available in standard clinical data. To further characterize those inconsistencies, we report results for ITH methods on 7 WES samples associated with single cell sequencing allowing to have an estimate of the ground truth. As a conclusion, we suggest that results of ITH analysis from single sample WES data with current computational pipelines should be manipulated with caution, and that more robust methods or protocols are likely to be needed for clinical applications.

Materials and methods

Data

We downloaded data from the GDC data portal `https://portal.gdc.cancer.gov/` for 3 cancer types (BLCA - 351 patients, BRCA - 904 patients, HNSC - 442 patients). We gathered annotated somatic mutations, both raw variant calling output, whose access is restricted and public mutations, from the new unified TCGA pipeline `https://docs.gdc.cancer.gov/Data/Bioinformatics_Pipelines/DNA_Seq_Variant_Calling_Pipeline/`, with alignment to the GRCh38 assembly, and variant calling using 4 variant callers: MuSe, Mutect2, VarScan2 and SomaticSniper. Instructions for download can be found in the companion Github repository (`https://github.com/judithabk6/ITH_TCGA`). RNAseq data used to compute immune signatures were downloaded through TCGABiolinks [23], and we downloaded clinical data from the CBIO portal [24].

Copy number calling and purity estimation

We obtained copy number alterations (CNA) data from the ASCAT complete results on TCGA data partly reported on the COSMIC database [25, 26]. We then converted ASCAT results on hg19 to GRCh38 coordinates using the `segment_liftover` Python package [27]. ASCAT results also provide an estimate of purity, which we used as input to ITH methods when possible. Other purity measures are available [28]; however we selected the ASCAT estimate to ensure consistency with CNV data.

The calls of allele-specific copy number and purity from ABSOLUTE [29] were downloaded from the GDC data portal `https://gdc.cancer.gov/about-data/publications/pancanatlas` on August 18th 2019. They were converted to GRCh38 as the ones from ASCAT.

Variant calling filtering

Variant calling is known to be a challenging problem. It is common practice to filter variant callers output, as ITH methods are deemed to be highly sensitive to false positive single nucleotide variants (SNVs). We filtered out indels from the public dataset, and considered the union of the 4 variant callers output SNVs. For the protected data, we also removed indels, and then filtered SNVs on the `FILTER` columns output by the variant caller ("PASS" only VarScan2, SomaticSniper, "PASS" or "panel_of_normals" for Mutect2, and "Tier1" to "Tier5" for MuSe). In addition, for all variant callers, we removed SNVs with a frequency in 1000 genomes or Exac greater than 0.01, except if the SNV was reported in COSMIC. A coverage filter was added, and we kept SNVs with at least 6 reads at the position in the normal sample, of which 1 maximum reports the alternative nucleotide (or with a variant allele frequency (VAF) <0.01), and for the tumor sample, at least 8 reads covering the position, of which at least 3 reporting the variant, or a VAF>0.2. The relative amount of excluded SNVs from protected to public SNV sets varied significantly between the 3 cancer types (see Table S1). All annotations are the ones downloaded from the TCGA, using VEP v84, and GENCODE v.22, sift v.5.2.2, ESP v.20141103, polyphen v.2.2.2, dbSNP v.146, Ensembl genebuild v.2014-07, Ensembl regbuild v.13.0, HGMD public v.20154, ClinVar v.201601. We further denote the filtered raw mutation set as "Protected SNVs" and the other one, which is publicly available, as "Public SNVs"

ITH methods

Published methods

We consider four published ITH methods: SciClone [7], PhyloWGS [8], PyClone [6] and EXPANDS [9]. In addition, we consider the MATH score [30] as a simple indicator of ITH, as well as a baseline ITH method described below. All computations were stopped after running 15 hours. This threshold was chosen to get results for most samples ($> 95\%$ when time was the limiting factor) for most methods while saving computational resources. Mean and standard deviation (std) of runtimes were computed for each method with each input mutation set separately. All parameters used for each method are detailed in the companion public Github repository containing all the commands `https://github.com/judithabk6/ITH_TCGA`. To ensure comparison, the runtimes were only performed on runs with ASCAT copy number calls.

We performed post-treatment to keep only clones with at least 5 SNVs, except for samples in which all clones were under 5 SNVs when all clones were considered. After running each ITH method we extracted 5 features to characterize ITH in a sample: the number of clones, the proportion of SNVs that belong the the major clone, the minimal cellular prevalence of a subclone, the Shannon index of the clonal distribution, and the cellular prevalence of the largest clone in terms of number of SNVs.

Consensus (CSR)

We computed a consensus of several ITH methods using the open source package CSR available at `https://github.com/kaixiany/CSR`. This method relies on matrix factorization to output a consensus clustering. We computed two separate consensus (for protected and public data), using as input the results of PyClone, SciClone, PhyloWGS, EXPANDS and baseline. MATH estimates were not well suited for the consensus. For each run, we ran matrix factorization for a maximum of 500 seconds.

Clinical variables

For each cancer type, we collected clinical variables from the CBIO Portal according to the following conditions: (i) categorical variables were one-hot encoded, and each level was kept if it involved at least 50 patients, and at most 50 patients had another level of the same variable; (ii) we kept numerical variables available for every patient; and (iii) in addition, we only kept the variables (if numerical) or the levels (categorical) which were significantly associated with overall survival by a single-variable cox model estimated with the Python package `lifelines` [31] after Benjamini-Hochberg correction for multiple hypothesis testing [32]. Tables S2, S3, and S4 summarize the clinical variables retained for each cancer type.

Survival regression

Model

To estimate the prognosis power of a set of features, we use a survival SVM model [33]. Survival SVM maximizes a concave relaxation of the concordance between the predicted survival ranks and the original observed survival, regularized by a Euclidean norm penalty. Formally, given a training set of n patients with survival information $(\mathbf{x}_i, y_i, \delta_i)_{i=1,\dots,n}$, where $\mathbf{x}_i \in \mathbb{R}^p$ is a vector of p features for patient i, $y_i \in \mathbb{R}$ is the time, and $\delta_i \in \{0,1\}$ indicates the event ($\delta_i = 1$) or censoring ($\delta_i = 0$), a survival SVM learns a linear score of the form $f(\mathbf{x}) = \mathbf{w}^\top \mathbf{x}$ for any new patient represented by features $\mathbf{x} \in \mathbb{R}^p$ by solving:

$$\min_{\mathbf{w}} \mathbf{w}^\top \mathbf{w} + \alpha \sum_{i,j \in \mathcal{P}} \max(0, 1 - (\mathbf{w}^\top \mathbf{x}_i - \mathbf{w}^\top \mathbf{x}_j))^2 \,,$$

where $\mathcal{P} = \{(i,j) \in [1,n]^2 \mid y_i \geq y_j \wedge \delta_j = 1\}$ is the set of pairs of patients (i,j) which are comparable, that is, for which we are certain that patient i lived longer than patient j. Intuitively, the loss penalizes the cases where patient i survives longer than patient j but the opposite is predicted by the model. For all computations, we used the function `FastSurvivalSVM` in the Python Package `scikit-survival` [34], with default parameters. The model was trained and tested using a 5-fold cross-validation procedure.

Evaluation procedure

To assess the accuracy of a survival regression model, we use the concordance index (CI) between the predicted score and the true survival information on a cohort with survival information. Given such a cohort $(\mathbf{x}_i, y_i, \delta_i)_{i=1,\dots,n}$, the CI measures how concordant the predicted survival times $s_i = f(\mathbf{x}_i)$ are with the observed survival times y_i for comparable pairs of patients:

$$CI = \frac{1}{|\mathcal{P}|} \sum_{i,j \in \mathcal{P}} I(s_j - s_i) \,,$$

$$\text{with } I(u) = \begin{cases} 1 & \text{if } u > 0 \,, \\ \frac{1}{2} & \text{if } u = 0 \,, \\ 0 & \text{otherwise.} \end{cases}$$

In practice, we compute an approximation of CI with the function `concordance.index` from the R package `survcomp` [35], using the `noether` method [36], and the associated one-sided test to compare CI to 0.5, which is the mean CI obtained with a random predictor. To compare CI's of different methods, we use a paired Student t-test for dependent samples implemented in the function `cindex.comp` from the same package. In both test settings, we aggregate p-values from each of the five cross-validation folds using the Fisher method from Python package `statsmodels`, and apply a Benjamini-Hochberg correction [32] to correct for multiple testing.

Immune signatures

We normalized RNAseq raw count data using a variance stabilizing transformation (VST) implemented in the Deseq2 R package [37], treating each cancer type separately. We mapped genes from Bindea et al. [38] to Ensembl GeneIds present in the TCGA matrix using EntrezId match table downloaded from Biomart [39] on March 26th 2018. Out of 681 EntrezId (577 unique), 31 (24 unique) were not matched to an Ensembl Id with associated gene expression in the TCGA RNAseq data. Each signature was then computed by averaging the VST output value for the relevant Ensembl Id for each TCGA sample. The resulting signatures we used can be found as Supplementary Table S5. For analysis purposes, we use the complementary to the maximal value in the cohort so that the content in immune cells varies in the same direction as tumor purity and remains a positive quantity. We denote those new variables with the prefix `inv`, e.g., for patient i in the BRCA cohort we define

$$\texttt{inv_T_cells}_i = \left(\max_{j \in \text{BRCA patients}} \texttt{T_cells}_j \right) - \texttt{T_cells}_i \,,$$

where $\texttt{T_cells}_i$ represents the signature for T cells estimated as explained above.

Correlations

We assessed correlations using Pearson's correlation coefficient. We computed the associated significance (for the null hypothesis that the correlation coefficient is 0) using the `scipy.stats.pearsonr` function, and we corrected the significance for multiple testing using the Benjamini Hochberg procedure at $FDR \leq 0.05$.

Comparison metrics

In addition to the correlations of the number of clones between methods, we have implemented three metrics derived from [40] to compare ITH methods together:

Score1B measures the adequacy between one number of clones J_1 and another number of clones J_2. It is computed as $\frac{J_1+1-\min(J_1+1,|J_2-J_1|)}{J_1+1}$.

Score1C is the Wasserstein distance between two clusterings, defined by the CCFs of the different clones and their associated weights (proportion of mutations), implemented as the function `stats.wasserstein_distance` in the Python package `scipy`.

Score2A measures the correlation between two binary co-clustering matrices in a vector form, M_1 and M_2. It is the average of 3 correlation coefficients:

- **Pearson correlation coefficient** $PCC = \frac{\text{Cov}(M_1,M_2)}{\sigma_{M_1},\sigma_{M_2}}$, implemented as the function `pearsonr` in the Python package `scipy`,
- **Matthews correlation coefficient** $\text{MCC} = \frac{TP \times TN - FP \times FN}{\sqrt{(TP+FP)(TP+FN)(TN+FP)(TN+FN)}}$, implemented as the function `metrics.matthews_corrcoef` in the Python package `scikit-learn`,
- **V-measure** is the harmonic mean of a homogeneity score that quantifies the fact that each cluster contains only members of a single class, and a completeness score measuring if all members of a given class are assigned to the same cluster [41]; here the classes are the true clustering. We used the function `v_measure_score` in the Python package `scikit-learn`.

Before averaging, all those scores were rescaled between 0 and 1 using the score of the minimal score between two "bad scenarios": all mutations are in the same cluster, or all mutations are in their own cluster ($M_{pred} = \mathbf{1}_{N\times N}$ or $M_{pred} = \mathbb{I}_{N\times N}$).

All scores as asymmetrical and were hence computed twice. In the case of score2A, only the mutations present in the two reconstructions were considered.

WES and single cell paired dataset

Data availability and preprocessing

The raw data for 7 normal-tumor WES samples analyzed jointly with matching single cell sequencing [42] were downloaded from the NCBI SRA platform `https://www.ncbi.nlm.nih.gov/sra` and processed into fastq format using the tool `fastq-dump` for the two acute lymphoblastic leukemia (ALL) patients (accession numbers: SRR1517761, SRR1517762, SRR1517763, SRR1517764) [43], or directly downloaded in the fastq format from the EBI ENA platform `https://www.ebi.ac.uk/ena` for the Triple Negative Breast Cancer patient (TNBC) [44] (accession number: SRR1163508 and SRR1298936), and the two samples (primary tumor and liver metastasis) from the two colorectal cancer patients (CRC) [45] (accession number: SRR3472566, SRR3472567, SRR3472569, SRR3472571, SRR3472796, SRR3472798, SRR3472799, SRR3472800).

All normal-tumor pairs underwent a pipeline of analysis including alignment with BWA-MEM [46] with options "-k 19 -T 30 -M", filtering of reads based on target intersection, mapping quality and PCR duplicates removal, using Picard [47], Bedtools [48] and Samtools [49], and preprocess using GATK [50] for local re-alignment around indels, and base score recalibration. Variant calling was performed using Mutect2 [51], and variants filtered under the same rules as used for the TCGA (only "PASS" variants, and minimal covering rules), and copy number assessed with Facets [52]. SNVs used in the analysis with B-SCITE [42], passing the covering filters but not recovered by this pipeline were added to the final variant list. Those variants and the copy number profile were then passed to PyClone, SciClone, PhyloWGS and Expands for ITH deconvolution.

Evaluation metrics

To measure the accuracy of subclonal reconstructions from the WES data only using different methods, we compared these reconstructions to the reconstruction obtained by B-SCITE using both WES and single cell sequencing [42]. To quantify the similarity of the different reconstruction results, we compared the number of clones, and for the common mutations, the metric 2A, used in [40] and redefined above.

Results

Assessing ITH on TCGA samples

We collected somatic mutation information from 1,697 TCGA patients with BLCA ($n = 351$), BRCA ($n = 904$), and HNSC ($n = 442$). We selected these three cancer types following conclusions of Morris et al. [13], since HNSC, BRCA and BLCA are characterized by respectively high (hazard ratio, HR=3.75, p=0.007 in multivariate Cox model), intermediate (HR=2.5, p=0.15) and absence (HR=1.05, p=0.91) of prognostic power of ITH. For each patient, we collected two sets of mutations based respectively on protected and public SNV sets. The protected set corresponds to raw variant calling outputs, with an extra filtering step described in Methods. The public set corresponds to publicly available SNV calls, filtered from the raw variant calling outputs to only retain somatic mutations with very high confidence, in order to ensure patients' anonymity. Supplementary Table S1 summarizes some statistics on the number of mutations per sample for each cancer type.

We assess ITH in each sample using 6 representative computational methods: PyClone [6], SciClone [7], PhyloWGS [8] EXPANDS [9], the mutant-allele tumor heterogeneity (MATH) score [30], and CSR [16], a method providing a consensus of all of the above results (except MATH which is not compatible, see Methods). Table 1 summarizes some important properties of the different methods, which might be helpful for designing future studies and selecting the appropriate tool. All methods but MATH take as input the CNA information in addition to a set of somatic mutation VAFs. PyClone and PhyloWGS also take purity

as input. All input has to be pre-computed by third-party approaches. While MATH is a single quantitative measure of ITH based on differences in the mutant-allele fractions among mutated loci, all 6 other methods produce more details such as the number of subclones and their respective proportions in the tumor. In particular, PhyloWGS outputs a lineage tree connecting the subclones.

Method	**CNA as input**	**Purity as input**	**Outputs tree(s)**	**Reference**	**Mean (std) runtime protected in seconds**	**Mean (std) runtime public in seconds**	**Success rate (protected)**	**Success rate (public)**
MATH	no	no	no	[30]	$<< 1$	$<< 1$	100%	100%
EXPANDS	yes	no	no	[9]	891 (604)	267 (258)	89%	**71%**
PyClone	yes	yes	no	[6]	7,035 (8,464)	1,414 (1,415)	95%	99%
SciClone	yes	no	no	[7]	62 (48)	41 (51)	92%	**78%**
PhyloWGS	yes	yes	yes	[8]	13,258 (9,058)	4,730 (4,139)	95%	97%

Table 1: Main characteristics of ITH methods tested. The mean runtime is the mean time to process a TCGA sample. The success rate is the fraction of TCGA samples for which the method produced an output without error, with ASCAT calls as input only. The MATH score was computed in one step for all samples, using a table containing all mutations for all samples; the operation lasted 3.21s (std. 47.6 ms) for the protected dataset, and 3.39s (std. 11ms) for the public dataset. All time measurements were measured on a single cluster node with a 2.2 GHz processor and 3GB of RAM.

We tested each method three times: on each sample for the two mutation sets combined with ASCAT calls for purity and copy number, and combined with ABSOLUTE calls for the protected mutation set. We observed that some methods failed to produce an output on some samples, for different reasons (see success rate for each method in Table 1). EXPANDS produces an error for 30% of the samples, mostly for tumors with high purity or very few CNAs. SciClone fails to provide an output for samples with an insufficient number of SNV in regions without CNA or LOH event. PyClone and PhyloWGS non completion cases were caused by a too long runtime.

As shown in Figure 1, there is little overlap between the samples where each method fails. Out of 1,697 initial TCGA samples, all methods produced an output for the three runs on only 686 samples (296 BRCA, 178 BLCA, 212 HNSC). Those failure cases unveil indications of each method's limitations, in particular EXPANDS and SciClone. In the following we restrict our analysis to those 686 samples. One can note that there is more difference between public and protected results for BRCA samples; this is expected as the number of mutations in those two sets is more different for this cancer type, as shown in S1.

In addition to failures, we observed that the runtime varies significantly between methods (Table 1). As shown on Figure 2, the run time of different ITH methods increases with the number of somatic mutations. PyClone and PhyloWGS runtime rises very quickly with the number of mutations in tumor sample, which can be a limitation for applications to heavily mutated tumors.

Methods quantifying ITH exhibit inconsistent results

As a first evaluation of ITH methods in the absence of ground truth, we assess the agreement between methods, with a focus on the number of clones. Each method except MATH outputs an estimated number S of subclonal populations, ranging from $S = 1$ for an homogeneous, clonal tumor to any positive number for an heterogeneous one. Figure 3 presents the distribution of estimated clonality among all samples for each approach and each SNV set, and each copy number calling method. We observe large differences between methods, as well as between SNV sets: for instance, over all samples, the percentage of estimated clonal tumors ($S = 1$) varies from 4% (for PhyloWGS on protected data) to 57 % (for PyClone on public data). Moreover, the number of estimated populations can vary strikingly with the mutation set used, but not really with the different input copy number. There is a clear trend among all methods to yield higher ITH estimates with the protected mutation set. PhyloWGS and EXPANDS (and CSR) are the only methods that detect ITH in almost all tested samples with the protected mutation set.

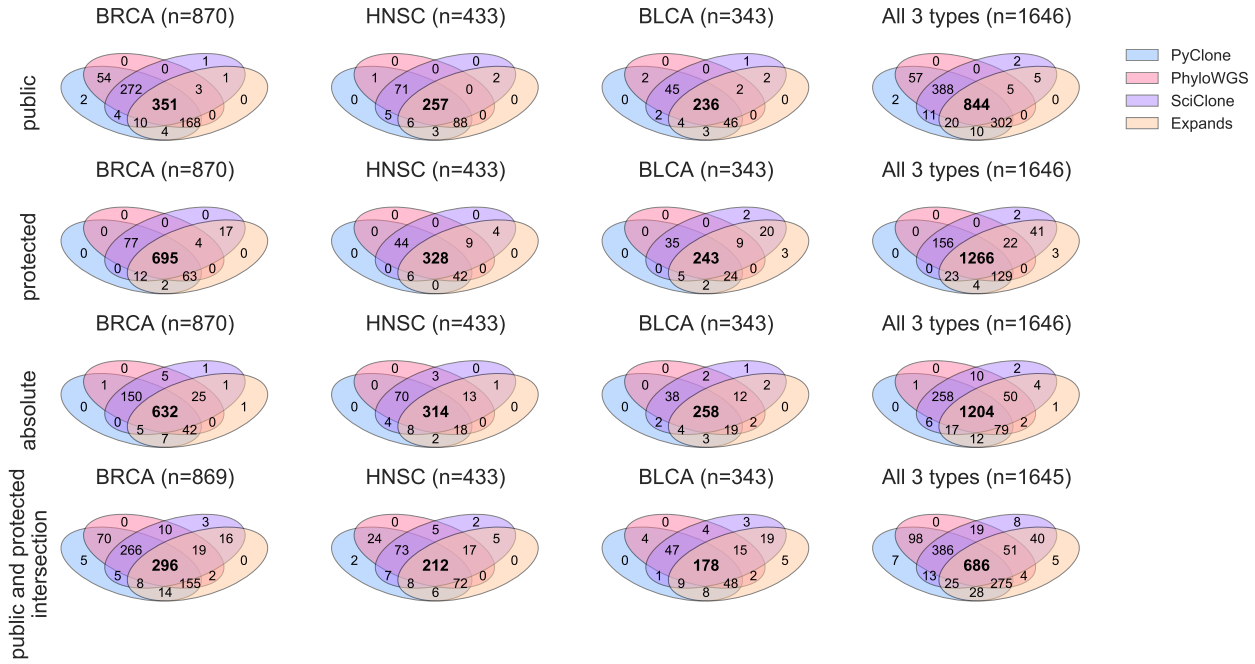

Figure 1: Intersection of successful runs among the 4 considered ITH methods. The upper venn diagrams concern runs with the public input SNV set, the second line with the protected, and the third the overall intersection, as results with both sets are necessary for a proper and rigorous comparison.

Another way to compare methods is to consider correlations (Pearson's r) between the estimated numbers of populations. This allows us to include the MATH score in the evaluation, considering it as an increasing function of heterogeneity just like the number of populations. In addition, we add to the comparison 5 measures directly extracted from the NGS analysis, namely, the number of mutations in the protected and in the public sets, the percentage of non-diploid cells (estimated by ASCAT and ABSOLUTE), the purity (estimated by ASCAT and ABSOLUTE), and the *inv_T_cell* (estimated from gene expression signatures). Results are presented in Figure 4.

Although a clear and consistent message is hard to extract, a few general trends seem to emerge. First, there is a tendency of results to be more similar for different methods with the same input mutation set, in particular for BRCA, where results for PyClone, SciClone, PhyloWGS and CSR are grouped together for each input set. Second, the really unexpected result is to observe that ITH results with the same input can be uncorrelated, and even significantly negatively correlated. Third, we observe two groups of methods that remain more similar across all three cancer types: EXPANDS and MATH score on the one hand, and PhyloWGS, PyClone, SciClone on the other hand. Those results can be related to the methods themselves. Indeed, PyClone, PhyloWGS and SciClone all define a probabilistic model explaining all observations of copy-number and read counts, based on a mixture model. They differ by the exact nature of the model (choice of distributions, exact definition of parameters), but they have similar structures. SciClone is different from PyClone and PhyloWGS in two ways: it only relies on mutations that are in regions without copy number alterations; and it does not correct for tumor purity. It is therefore not surprising that PyClone and PhyloWGS yield similar results, and that SciClone is a bit more different. CSR performs a consensus of all obtained clusterings; since 3 methods out of 4 have similar results, CSR might be biased towards those 3 methods. Expands makes similar assumptions as PyClone and PhyloWGS. However, the estimation process is very different: Expands estimates a distribution of read number for each position, and then clusters those distributions, while PyClone, SciClone and PhyloWGS attempt to find a common distribution for a group of mutations. The MATH score has an entirely different rationale as it simply ignores CNVs.Similar trends are observed when comparing the methods based on other pairwise comparison metrics (see Supplementary Figure S2-S4).

Regarding potential confounding variables, previous studies have reported a correlation between MATH

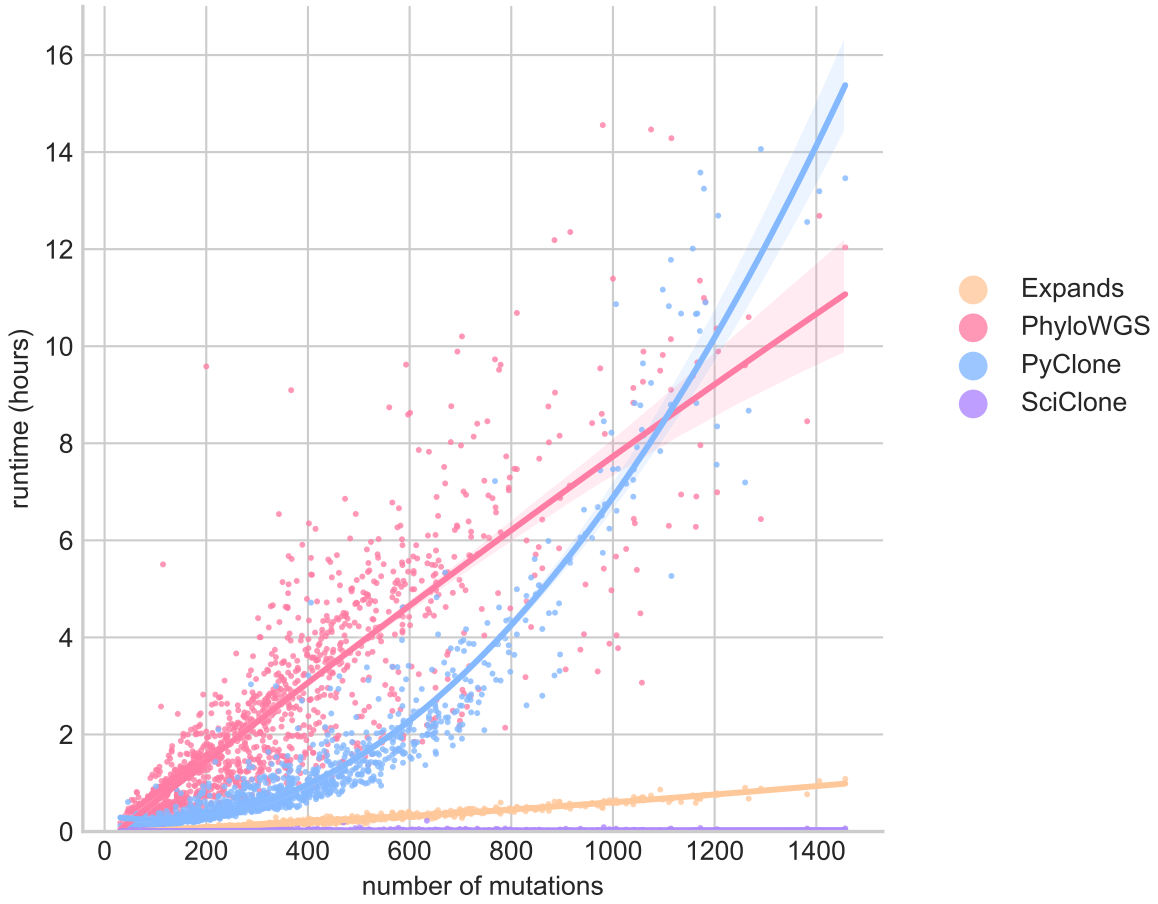

Figure 2: Runtime of the different ITH methods as a function of the number of mutations in each sample. Lines represent second degree polynomial fit with shaded regions are 95% confidence-intervals

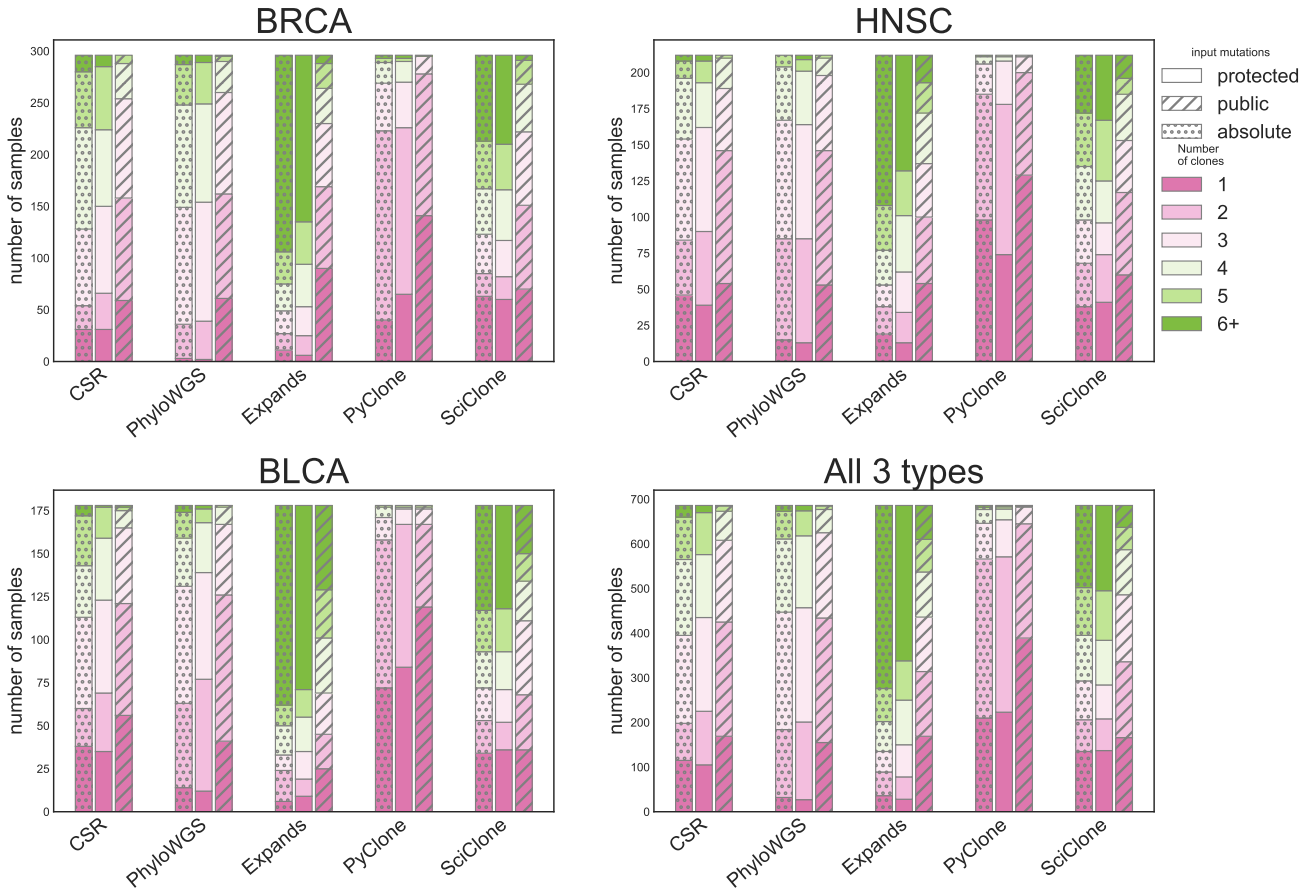

Figure 3: Distribution of number of clones called by ITH methods with public and protected mutations sets as inputs. Distribution of the number of subclones for the tested ITH methods, and 2 alternative input mutation sets for samples in the different cancer types and 2 different copy number methods for the protected mutation set. MATH could not be included in this analysis as this method does not estimate a number of clones.

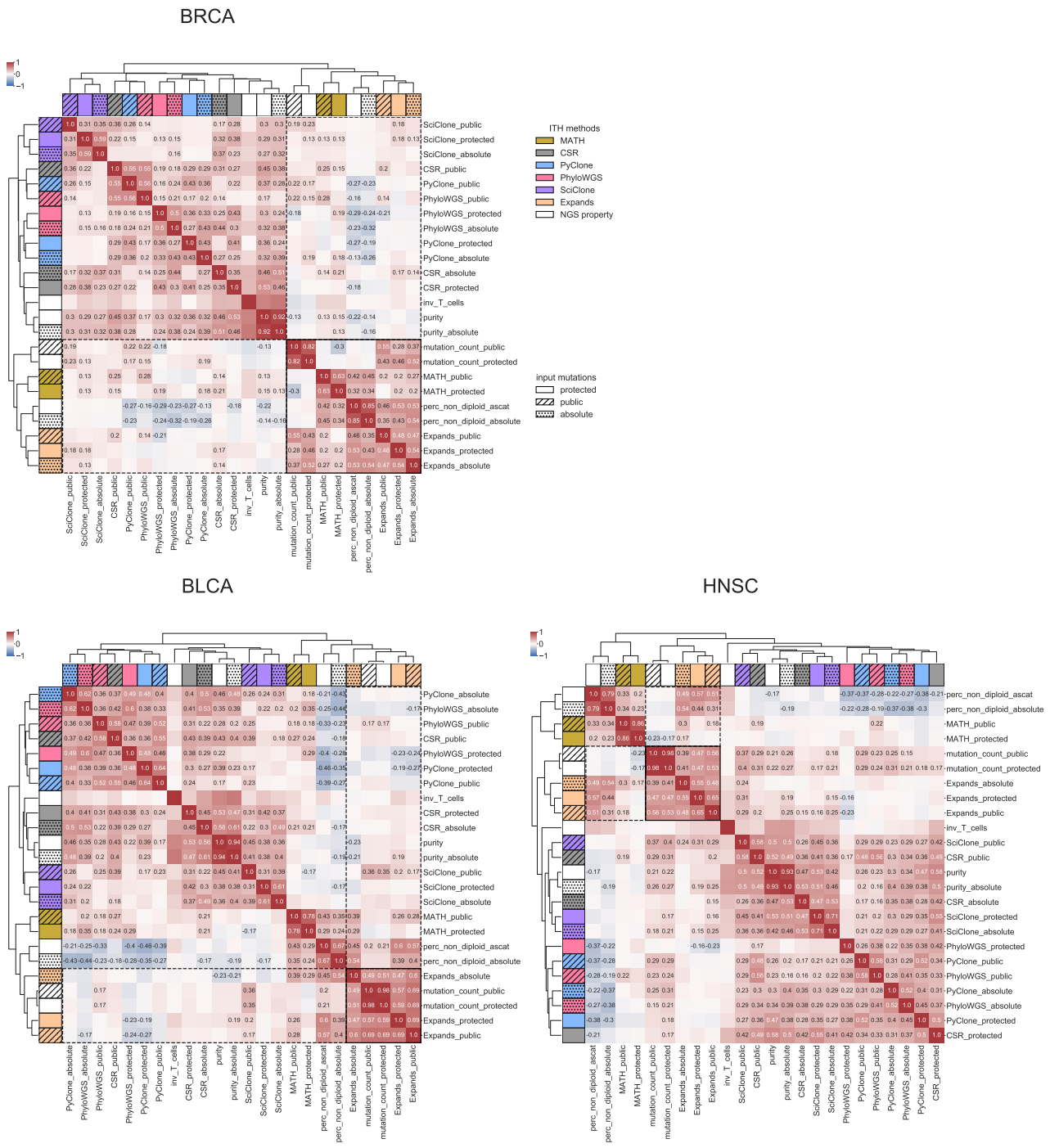

Figure 4: Correlation between various measures of ITH (MATH score, and number of subclones for the other methods), and other potential confounding variables measured using WES and trancriptomics data. Row and color label represent the method used, with white for the genomic measures not involving ITH. Hatches correspond to public mutation sets. Heatmap colors represent the value of the Pearson's r, which is written numerically whenever it is significantly different from 0 ($FDR \leq 0.05$ after Benjamini Hochberg correction for multiple tests). We can observe clustering tendencies stable across the 3 cancer types. One of them is highlighted in black lines.

score and CNA abundance [53, 22, 54], or between purity and ITH, as ITH methods were initially designed to refine purity estimation [29], and we observe similar behaviors. Association with immune infiltration has also been considered [54], though it is worth noting that immune infiltration and tumor purity are not independent, as immune cells are not cancerous. Each group of ITH methods is highly correlated to distinct genomic metrics, mutation load and CNV abundance (`perc_non_diploid`) for the first group (MATH, Expands), and purity (and the opposite of immune cells infiltration (`inv_T_cells`)) for the latter (PyClone, SciClone, PhyloWGS CSR). This might be indicative of systematic biases in the different methods, rather than biological strong signal as previously reported. Indeed, the strength and direction of all correlations vary between the two groups of ITH methods, and is hence hardly reliable or interpretable in terms of clinically actionable information without more data.

Similar results are obtained on an independent dataset of 7 samples from 5 patients where both WES and single cell sequencing was performed. In this dataset, subclonal reconstruction was performed by the method B-SCITE [42] that uses both bulk sequencing and single cell sequencing as input, and provides the most accurate representation possible. To further illustrate the behavior of ITH methods, we have compared results obtained for each sample separately to the B-SCITE result. To evaluate the concordance of each reconstruction to the B-SCITE reconstruction, we compare the number of clones, and the score2A from [40] that evaluates the co-clustering of mutations. The other metrics considered in [40] focus on the distance between the true and reconstructed cancer cell fraction distributions (score 1C), but in this setting, the ground truth does not provide a true CCF distribution estimate, and on the phylogenetic relationships between clones (score 3), but only PhyloWGS provides a tree among the considered methods. For this evaluation, we have left the true estimate for PyClone that provides a lot (several dozens) of clones with a single mutation. The input to ITH methods we have used results from variant calling on the bulk WES data, whereas the input to B-SCITE is more restrictive, and focuses on mutations detected both in the WES and in the single cells; the score2A is computed on the common mutations. Results are presented in Table 2. As observed on the TCGA, different methods based on WES data exhibit very different estimates of the number of clones, and none is very close to the estimates of B-SCITE using WES and single-cell data. In terms of clone composition, PyClone is the closest to B-SCITE in terms of score2A correlation in four out or seven samples, although the score2A values remain very modest.

sample	nb clones bscite	nb clones pyclone	nb clones sciclone	nb clones phylowgs	nb clones expands	nb clones CSR	score2A pyclone	score2A sciclone	score2A phylowgs	score2A expands	score2A CSR	MATH score	nb mutations WES	nb mutations metric
CO8_colon	10	20	2	4	6	4	0.000	0.211	0.000	0.146	**0.387**	82.733	272	12
CO8_liver	10	81	6	5	6	3	**0.254**	0.248	0.218	0.241	0.185	80.062	486	17
BRCA_wang_TNBC	13	166	7	3	9	5	**0.465**	0.193	0.269	0.227	0.296	53.939	1458	7
ALL_gawad_P2	7	6	2	4	NA	NA	**0.315**	0.303	0.231	NA	NA	54.918	115	15
CO5_liver	5	29	6	5	3	3	0.060	0.197	0.000	0.143	**0.319**	66.151	271	11
CO5_colon	5	8	6	5	2	3	0.000	**0.180**	0.000	0.093	0.031	70.680	305	10
ALL_gawad_P1	6	18	2	5	NA	NA	**0.345**	0.039	0.138	NA	NA	71.163	110	20

Table 2: Results on the single cell-WES dataset.

ITH is a weak and non robust prognosis factor

To test the prognostic power of each ITH quantification method, we collected survival information for the 686 patients on which all ITH methods ran successfully, and assessed how each ITH method allows to predict survival. Since all ITH methods except MATH output several features related to ITH, we did not test each feature individually but instead estimated a combined score for each method with a survival SVM model (see Methods). More precisely, we extract 5 features from each ITH method: the number of subclonal populations, the proportion of SNVs that belong the the major clone, the minimal cellular prevalence of a subclone, the Shannon index of the clonal distribution, and the cellular prevalence of the largest clone in terms of number of SNVs that enable to distinguish several evolutionary patterns, like early (star-like evolution) or late (tree with a long trunk) clonal diversification (see Methods). We evaluate the performance of each score by 5-fold cross-validation, and prognostic power is assessed on the test fold by computing the concordance index between the SVM prediction and the true patient survival. For MATH, a single feature is computed, so this procedure simply evaluates the concordance index of the MATH score with survival. In addition, we consider a model where all features of all methods (i.e., a total of $6 \times 5 + 1 = 26$ features) are combined together.

Figure 5 shows the results for each cancer type, each method, and each set of mutations used.

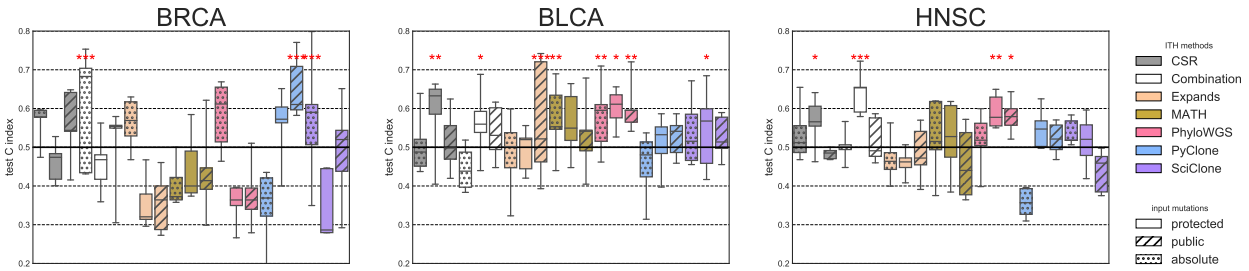

Figure 5: **Prognostic power of ITH measured using different ITH method and input mutation set combination.** 0.5 corresponds to a random prediction, and stars indicate statistical significance (p-value < 0.001: ***, < 0.01: **, < 0.05: *). Results are presented for 3 cancer types (BRCA, BLCA and HNSC from left to right).

Overall, we observe at least one method achieving significant survival prediction in each cancer type. The combined model is significantly prognostic with both protected and public sets in all three cancer types. Among the three cancer types, in the best case, however, the median concordance index on the test sets barely reaches 0.6 (except with the combination with absolute copy number in BRCA, but with an important variance), which remains modest for any clinical use. This suggests that there may be a weak prognostic signal captured by ITH measurement, but it can not be observed consistently with a single method and a single variant and copy number calling pipeline in the three cancer types, illustrating the frailty of obtained results. The combined model seems to be a robust alternative, as when it is significant, it has a concordance index in the range of the best performing single method; however the case of BRCA seems particular, as many methods perform worse than random.

Some authors [14, 55] have suggested a non-linear relation between survival and ITH, as very high ITH might be damaging for the tumor, while moderate ITH would be associated with aggressive tumors and prone to treatment resistance. To test this hypothesis in our framework we added squared features to the survival model, allowing second order polynomial relations between ITH and survival. However, this did not significantly impact the results (Supplementary S1). Indeed, after multiple test correction, only PyClone with the protected mutation set and ABSOLUTE copy number in BRCA prognostic power is increased by adding the squared features ($p = 0.027$, paired t-test), but both CI indexes remain below 0.5. We also assessed whether the relatively poor performance of the different methods was due to the difficulty to learn a prognostic score combining 5 features from limited amounts of training samples, by assessing the prognostic ability of a single feature: the number of clones. A significant improvement was obtained for 7and a significant decrease in performance in 3 of the 36 tested settings (4 methods, 2 mutation sets, 2 copy number methods, 3 cancer types). This suggests that the complexity of the model (polynomial of order 2 instead of linear) and the choice of ITH features have little influence on the results. This might be related to the fact that very little signal can be detected in the first place.

ITH prognosis signal is redundant with other known factors

We have established that in some cases, ITH may exhibit weak prognostic power. It is then very important to assess whether it is complementary to already available prognostic features, like clinical characteristics. To answer this question, we consider relevant clinical features, as described in Methods.

Figure 6 presents a comparison between different prediction settings: clinical features without any clonality and clonality associated with clinical features. In all cases, clinical features alone have a significant prognostic power (median CI=0.79 for BRCA, 0.65 for BLCA, 0.65 for HNSC). More importantly, when we combine each ITH feature set with clinical features, we observe no significant improvement over clinical features alone. This suggests that the weak prognostic signal captured by ITH measures is in fact redundant with already available clinical factors.

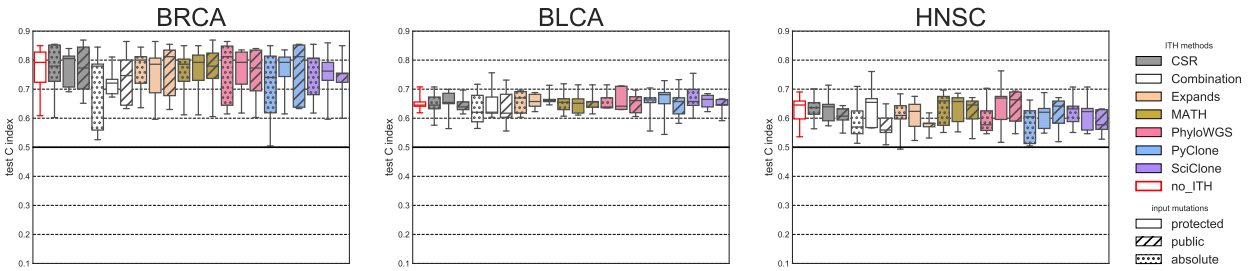

Figure 6: Prognostic power of ITH-derived features compared to other prognostic factors (686 patients in total). ITH-derived features are used in association with clinical features to predict overall survival. Left-most boxplots (with red contour lines) represent results using clinical variables alone, without any ITH, to serve as reference.

Discussion

Comparison to similar studies

Previous findings report divergent prognostic power for ITH in several pan cancer studies [14, 13, 22]. Andor et al [14] analyzed 1,165 patients across 12 cancer types from the TCGA, and found an overall prognostic power by considering all types together, and suggested that this effect might be nonlinear, with a trade-off between ITH and overall survival [55]. However, the association between the number of subclones and overall survival was significant with EXPANDS, but not with PyClone results, and no significant association was detected when considering each cancer type separately, except for gliomas. This might be due to the small number of cases of each type (between 33 and 166). Morris et al. [13] considered 3,383 patients of 9 cancer types from the TCGA and found significant association between the number of subclones found by PyClone in 5 types: HNSC, BRCA, KIRC, LGG, and PRAD. Noorbakhsh et al. [22] studied 4,722 patients from 11 types from the TCGA, and found significant prognostic power in 4 types using MATH score and distinct input mutation sets from different variant callers. They obtain significant prognostic association for all variant calling results in only one cancer type: UCEC, and already report some lack of robustness in the results. We have been further in testing up to 7 ITH methods with 2 alternative input mutation sets, in addition to the combination of all methods, and found no significant association, either for the same framework in all considered cancer types, nor for the same cancer type with all frameworks. We have also tested more powerful polynomial models to account for a potential nonlinear relationship, and results were inconclusive. This is an important distinction, because mutation calling can be made robust by additional experiments (targeted sequencing on WXS or WGS candidates), but our results highlight intrinsic limitations of ITH methods.

Considering results in details, there are discrepancies that should be discussed. For BRCA, conclusions are more discordant: Morris et al. [13] found significant results, Noorbakhsh et al. [22] did not, and in more specialized studies like METABRIC [53], significant association was found when considering only the upper and lower quartile of MATH score for ER+ tumors. For BLCA, contradictory conclusions were also drawn, as previous studies [14, 13] found no prognostic power and we have with some ITH methods. There are several explanations: each study considered a distinct subset of patients, with a distinct pipeline for calling mutations and measure ITH. This instability with respect to patient selection has been confirmed by our study. All of those studies, including ours, observed ITH prognostic relevance in HNSC. Good prognostic power for HNSC and BLCA might be an indication that the importance of ITH for cancer aetiology differs across cancer types.

Can we truly measure ITH?

Beyond the question of the prognostic power of ITH, our results challenge the very fact that ITH can be measured accurately with one WES sample per patient. Up to 30 methods have been developed to tackle ITH detection and quantification from NGS data in tumor samples, and new ones are still being developed [56]. This analysis has focused on relatively early but among the most widely used ITH methods in order to provide valuable insight on the degree of reliability of provided results. Indeed results presented here show

that there is a very weak correlation (and sometimes even a significant negative correlation) between results obtained with different methods on the same patients. Another source of inconsistency is that ITH methods rely on results from previous analysis steps, in particular variant calling. Indeed, all ITH methods rely on the distribution of SNV frequencies, in association or not with structural variants (also called by a variety of dedicated methods). This has already been discussed by Noorbakhsh et al. [22] for MATH score computation. We show here that this issue is not limited to the MATH score. Some authors have suggested that being very restrictive in variant calling, even resorting to targeted deep sequencing to experimentally validate SNVs [6], would exhibit less noisy results. Here we have not observed any evidence that ITH methods estimated more robust results with a restricted input mutation set (i.e. the public mutation set in this study). Overall, lack of agreement between the different ITH measures is a real concern, indicating again that ITH is probably not very accurate. A similar conclusion was recently and independently reached by [57].

Beyond the methods used for ITH inference, the data might also be questioned. Being able to measure ITH to one sample WES with moderate sequencing depth is tempting for future clinical application where the cost and the inconvenience of multiple samples for patients should be limited [3], but it may be unrealistic, as the true heterogeneity of a tumor can be missed by a single biopsy. However, more complex experimental settings have allowed more convincing findings in the field of tumor evolution [58, 59], and it may be necessary to further evaluate lack of accuracy due to undersampling from the whole tumor or to use of WES instead of WGS, and the impact of sequencing depth. A recent and broad analysis of ITH with one WGS sample per patient [16] partially answers as the authors could detect ITH in almost every patient, and conduct interesting further analyses as they had confidence in the robustness of ITH estimates. Most published methods are able to account for multiple samples from the same patient, either sampled at different times or from different regions of the tumor. However, for extension to WGS analysis, our work highlights limitations with respect to the computation time for high numbers of mutation as input.

Association with survival, link with other variables

It is tempting to formulate the hypothesis that higher association with patient survival is a sign of higher accuracy. We have already mentioned some technical issues associated with the setting of one sample WES per patient, as even without measure issues, ITH might just be under-represented in the sample compared to the whole tumor [60]. Another limitation is that this does not represent a dynamic measure. For instance a tumor can be clonal because it is not very aggressive, or on the contrary this might be the result of a selective sweep after a phase of new clonal expansion. Moreover, several authors discuss the consequences and the interplay of the presence of distinct subclonal populations, in terms of cooperation [61, 62], competition [63, 64], or even neutral evolution [65, 66]. Hence, the same level of ITH might uncover very diverse situations, and may not be a prognostic factor by itself.

Moreover, the dataset used in this survival analysis has some particularities: the TCGA has selected patients with criteria allowing high sequencing quality, and ITH analysis itself has further eliminated tumors with no or very high CNA abundance, which may also bias results. Finally, absence of prognosis power in one dataset does not constitute a formal proof that ITH is not associated with survival.

Besides, ITH is likely to be influenced and to interplay with other external factors including tumor micro-environment, immune response, nutrient availability. Recent work has tried to set a full framework for analysis including many factors [67]. However, in the case of the TCGA, not all those variables are measurable, but some might be included in further work. In this line of thought, earlier results exhibited correlation of ITH with other factors like CNA abundance, sample purity, immune infiltration [53, 54, 68, 13]. Our results show that the strength (and even direction in the case of CNA abundance and mutation load) of correlation between those factors and ITH varies between the different tested ITH measures. This again calls for further and more detailed analysis, as results show ambiguity and lack of robustness.

Can we build a gold standard dataset for benchmark?

The main difficulty of ITH estimation is to assess the accuracy of the results. In this work, we have considered two possibilities. The first one on data from the TCGA is to work without any ground truth proxy and measure other features of accuracy: robustness, agreement of results obtained by different methods and association with other clinical variables. The obtained results suggest that the considered ITH methods are

relatively robust to changes in the copy number input, but very sensitive to the input mutations. The last two options are more difficult to work with, as one method could be in disagreement with all the others but still provide the most accurate result, and absence or presence of association between ITH and other clinical or genomic variables can be either due to a real biological signal or be an artifact (or bias) of the method. Though the goal of this study is not to provide a formal evaluation of the considered method, the results on the TCGA provide information on systematic trends of each method, and the level of confidence to expect when applying ITH methods.

A second possibility is to try and obtain a proxy for the ground truth. This can be done using single cell sequencing in addition to the bulk sequencing. Though suffering from other issues, single cell sequencing provides true associations or exclusions of mutations, and hence constraints the subclonal reconstruction [42]. However, a large number of cells is necessary. In the 7 samples considered in this study, only a subset of the mutations identified in the bulk sequencing were also identified in single cells, limiting the representativity and the relevance of the extracted accuracy measures. A second possibility is to rely on several samples from the same tumor to obtain a better ground truth to compare to the result obtained with one sample. However, each sample is a priori heterogeneous itself, requiring a first multi-sample deconvolution. This first step can be challenging, as it is thought that multi-sample reconstruction is subject to a larger statistical bias compared to single sample reconstruction [69], and the accuracy of this first step will be critical in the final results. A final possibility is to rely on simulated data, which have the major drawback to not be necessarily representative of the true biological data, as recently highlighted for ITH in [69], that point to an aspect of the input data so far overlooked by the community.

Acknowledgments

The authors declare no potential conflicts of interest. We thank Peter Van Loo and Kerstin Haase for sharing ASCAT results on the TCGA, and Christoffer Flensburg for his help with the liftover of ASCAT results. We thank the authors of "Integrative inference of subclonal tumor evolution from single-cell and bulk sequencing data" for help in adapting their results and reprocess the data, and Elodie Girard for her help with NGS data processing. We also thank Alice Schoenauer Sebag for helpful discussions. The results shown here are based upon data generated by the TCGA Research Network: `http://cancergenome.nih.gov/`, under authorization for project 10569 (dbGaP).

References

[1] Peter C Nowell. The clonal evolution of tumor cell populations. *Science*, 194(4260):23–28, 1976.

[2] Moritz Gerstung, Clemency Jolly, Ignaty Leshchiner, Stefan C Dentro, Kaixian Yu, Maxime Tarabichi, Amit Deshwar, and Jeff Wintersinger. The evolutionary history of 2,658 cancers. *bioRxiv*, 2017.

[3] Stefan C. Dentro, David C. Wedge, and Peter Van Loo. Principles of Reconstructing the Subclonal Architecture of Cancers. *Cold Spring Harbor perspectives in medicine*, 7(8):a026625, 2017.

[4] Niko Beerenwinkel, Roland F. Schwarz, Moritz Gerstung, and Florian Markowetz. Cancer evolution: Mathematical models and computational inference. *Systematic Biology*, 64(1):e1–e25, 2015.

[5] Russell Schwartz and Alejandro A. Schäffer. The evolution of tumour phylogenetics: Principles and practice. *Nature Reviews Genetics*, 18(4):213–229, 2017.

[6] Andrew Roth, Jaswinder Khattra, Damian Yap, Adrian Wan, Emma Laks, Justina Biele, Gavin Ha, Samuel Aparicio, Alexandre Bouchard-Côté, and Sohrab P. Shah. PyClone: Statistical inference of clonal population structure in cancer. *Nature Methods*, 11(4):396–398, 2014.

[7] Christopher A. Miller, Brian S. White, Nathan D. Dees, Malachi Griffith, John S. Welch, Obi L. Griffith, Ravi Vij, Michael H. Tomasson, Timothy A. Graubert, Matthew J. Walter, Matthew J. Ellis, William Schierding, John F. DiPersio, Timothy J. Ley, Elaine R. Mardis, Richard K. Wilson, and Li Ding. SciClone: Inferring Clonal Architecture and Tracking the Spatial and Temporal Patterns of Tumor Evolution. *PLoS Computational Biology*, 10(8):e1003665, 2014.

[8] Amit G. Deshwar, Shankar Vembu, Christina K. Yung, Gun Ho Jang, Lincoln Stein, and Quaid Morris. PhyloWGS: Reconstructing subclonal composition and evolution from whole-genome sequencing of tumors. *Genome Biology*, 16(1):1–20, 2015.

[9] Noemi Andor, Julie V. Harness, Sabine Müller, Hans W. Mewes, and Claudia Petritsch. Expands: Expanding ploidy and allele frequency on nested subpopulations. *Bioinformatics*, 30(1):50–60, 2014.

[10] Katharina Jahn, Jack Kuipers, and Niko Beerenwinkel. Tree inference for single-cell data. *Genome Biology*, 17(1):86, 2016.

[11] Alexander Davis and Nicholas E. Navin. Computing tumor trees from single cells. *Genome Biology*, 17(1):1–4, 2016.

[12] Simone Ciccolella, Mauricio Soto Gomez, Murray Patterson, Gianluca Della Vedova, Iman Hajirasouliha, and Paola Bonizzoni. Inferring Cancer Progression from Single-cell Sequencing while Allowing Mutation Losses. *bioRxiv*, page 268243, 2018.

[13] Luc G.T. Morris, Nadeem Riaz, Alexis Desrichard, Yasin enbabaolu, A. Ari Hakimi, Vladimir Makarov, Jorge S. Reis-Filho, and Timothy A. Chan. Pan-cancer analysis of intratumor heterogeneity as a prognostic determinant of survival. *Oncotarget*, 7(9), 2016.

[14] Noemi Andor, Trevor A. Graham, Marnix Jansen, Li C. Xia, C. Athena Aktipis, Claudia Petritsch, Hanlee P. Ji, and Carlo C. Maley. Pan-cancer analysis of the extent and consequences of intratumor heterogeneity. *Nature Medicine*, 22(1):105–113, 2016.

[15] Nicholas McGranahan and Charles Swanton. Clonal Heterogeneity and Tumor Evolution: Past, Present, and the Future. *Cell*, 168(4):613–628, 2017.

[16] Stefan C Dentro, Ignaty Leshchiner, Kerstin Haase, Maxime Tarabichi, Jeff Wintersinger, Amit G. Deshwar, Kaixian Yu, Yulia Rubanova, Geoff Macintyre, Ignacio Vazquez-Garcia, Kortine Kleinheinz, Dimitri G. Livitz, Salem Malikic, Nilgun Donmez, Subhajit Sengupta, Jonas Demeulemeester, Pavana Anur, Clemency Jolly, Marek Cmero, Daniel Rosebrock, Steven Schumacher, Yu Fan, Matthew Fittall, Ruben M. Drews, Xiaotong Yao, Juhee Lee, Matthias Schlesner, Hongtu Zhu, David J. Adams, Gad Getz, Paul C. Boutros, Marcin Imielinski, Rameen Beroukhim, S. Cenk Sahinalp, Yuan Ji, Martin Peifer, Inigo Martincorena, Florian Markowetz, Ville Mustonen, Ke Yuan, Moritz Gerstung, Paul T. Spellman, Wenyi Wang, Quaid D. Morris, David C. Wedge, Peter Van Loo, PCAWG Evolution Group, Heterogeneity Working, and PCAWG Network. Portraits of genetic intra-tumour heterogeneity and subclonal selection across cancer types. *bioRxiv*, page 312041, 2018.

[17] Ibiayi Dagogo-Jack and Alice T. Shaw. Tumour heterogeneity and resistance to cancer therapies. *Nature Reviews Clinical Oncology*, 15(2):81–94, 2018.

[18] Serena Nik-Zainal, Peter Van Loo, David C. Wedge, Ludmil B. Alexandrov, Christopher D. Greenman, King Wai Lau, Keiran Raine, David Jones, John Marshall, Manasa Ramakrishna, Adam Shlien, Susanna L. Cooke, Jonathan Hinton, Andrew Menzies, Lucy A. Stebbings, Catherine Leroy, Mingming Jia, Richard Rance, Laura J. Mudie, Stephen J. Gamble, Philip J. Stephens, Stuart McLaren, Patrick S. Tarpey, Elli Papaemmanuil, Helen R. Davies, Ignacio Varela, David J. McBride, Graham R. Bignell, Kenric Leung, Adam P. Butler, Jon W. Teague, Sancha Martin, Goran Jönsson, Odette Mariani, Sandrine Boyault, Penelope Miron, Aquila Fatima, Anita Langerod, Samuel A.J.R. Aparicio, Andrew Tutt, Anieta M. Sieuwerts, Åke Borg, Gilles Thomas, Anne Vincent Salomon, Andrea L. Richardson, Anne Lise Borresen-Dale, P. Andrew Futreal, Michael R. Stratton, and Peter J. Campbell. The life history of 21 breast cancers. *Cell*, 149(5):994–1007, 2012.

[19] Marco Gerlinger, Stuart Horswell, James Larkin, Andrew J. Rowan, Max P. Salm, Ignacio Varela, Rosalie Fisher, Nicholas Mcgranahan, Nicholas Matthews, Claudio R. Santos, Pierre Martinez, Benjamin Phillimore, Sharmin Begum, Adam Rabinowitz, Bradley Spencer-Dene, Sakshi Gulati, Paul A. Bates, Gordon Stamp, Lisa Pickering, Martin Gore, David L. Nicol, Steven Hazell, P. Andrew Futreal, Aengus Stewart, and Charles Swanton. Genomic architecture and evolution of clear cell renal cell carcinomas defined by multiregion sequencing. *Nature Genetics*, 46(3):225–233, 2014.

[20] Nicholas E. Navin. Tumor evolution in response to chemotherapy: Phenotype versus genotype. *Cell Reports*, 6(3):417–419, 2014.

[21] Nicholas Navin, Jude Kendall, Jennifer Troge, Peter Andrews, Linda Rodgers, Jeanne McIndoo, Kerry Cook, Asya Stepansky, Dan Levy, Diane Esposito, Lakshmi Muthuswamy, Alex Krasnitz, W. Richard McCombie, James Hicks, and Michael Wigler. Tumour evolution inferred by single-cell sequencing. *Nature*, 472(7341):90–95, 2011.

[22] Javad Noorbakhsh, Hyunsoo Kim, Sandeep Namburi, and Jeffrey H. Chuang. Distribution-based measures of tumor heterogeneity are sensitive to mutation calling and lack strong clinical predictive power. *Scientific Reports*, 8(1):1–12, 2018.

[23] Antonio Colaprico, Tiago C. Silva, Catharina Olsen, Luciano Garofano, Claudia Cava, Davide Garolini, Thais S. Sabedot, Tathiane M. Malta, Stefano M. Pagnotta, Isabella Castiglioni, Michele Ceccarelli, Gianluca Bontempi, and Houtan Noushmehr. TCGAbiolinks: An R/Bioconductor package for integrative analysis of TCGA data. *Nucleic Acids Research*, 44(8):e71, 2016.

[24] Jianjiong Gao, Bülent Arman Aksoy, Ugur Dogrusoz, Gideon Dresdner, Benjamin Gross, S. Onur Sumer, Yichao Sun, Anders Jacobsen, Rileen Sinha, Erik Larsson, Ethan Cerami, Chris Sander, and Nikolaus Schultz. Integrative analysis of complex cancer genomics and clinical profiles using the cBioPortal. *Science Signaling*, 6(269), 2013.

[25] Iñigo Martincorena, Keiran M. Raine, Moritz Gerstung, Kevin J. Dawson, Kerstin Haase, Peter Van Loo, Helen Davies, Michael R Stratton, and Peter J Campbell. Universal Patterns of Selection in Cancer and Somatic Tissues. *Cell*, 171(5):1029–1041.e21, 2017.

[26] Simon A. Forbes, David Beare, Harry Boutselakis, Sally Bamford, Nidhi Bindal, John Tate, Charlotte G. Cole, Sari Ward, Elisabeth Dawson, Laura Ponting, Raymund Stefancsik, Bhavana Harsha, Chai YinKok, Mingming Jia, Harry Jubb, Zbyslaw Sondka, Sam Thompson, Tisham De, and Peter J. Campbell. COSMIC: Somatic cancer genetics at high-resolution. *Nucleic Acids Research*, 45(D1):D777–D783, 2017.

[27] Bo Gao, Qingyao Huang, and Michael Baudis. segment_liftover : a Python tool to convert segments between genome assemblies. *F1000Research*, 7:319, jun 2018.

[28] Dvir Aran, Marina Sirota, and Atul J Butte. Systematic pan-cancer analysis of tumour purity. *Nature communications*, 6:8971, 2015.

[29] Scott L. Carter, Kristian Cibulskis, Elena Helman, Aaron McKenna, Hui Shen, Travis Zack, Peter W. Laird, Robert C. Onofrio, Wendy Winckler, Barbara A. Weir, Rameen Beroukhim, David Pellman, Douglas A. Levine, Eric S. Lander, Matthew Meyerson, and Gad Getz. Absolute quantification of somatic DNA alterations in human cancer. *Nature Biotechnology*, 30(5):413–421, 2012.

[30] Edmund A. Mroz and James W. Rocco. MATH, a novel measure of intratumor genetic heterogeneity, is high in poor-outcome classes of head and neck squamous cell carcinoma. *Oral Oncology*, 49(3):211–215, 2013.

[31] Cameron Davidson-Pilon, Jonas Kalderstam, Paul Zivich, Ben Kuhn, Andrew Fiore-Gartland, Luis Moneda, Gabriel, Daniel WIlson, Alex Parij, Kyle Stark, Steven Anton, Lilian Besson, Jona, Harsh Gadgil, Dave Golland, Sean Hussey, Javad Noorbakhsh, Andreas Klintberg, Joanne Jordan, Jeff Rose, Isaac Slavitt, Eric Martin, Eduardo Ochoa, Dylan Albrecht, dhuynh, Denis Zgonjanin, Daniel Chen, Chris Fournier, Arturo, and Andr F. Rendeiro. Camdavidsonpilon/lifelines: v0.20.0, March 2019.

[32] Yoav Benjamini and Yosef Hochberg. Controlling the false discovery rate: a practical and powerful approach to multiple testing. *Journal of the Royal Statistical Society*, 57(1):289–300, 1995.

[33] Vanya Van Belle, Kristiaan Pelckmans, Sabine Van Huffel, and Johan A.K. Suykens. Support vector methods for survival analysis: A comparison between ranking and regression approaches. *Artificial Intelligence in Medicine*, 53(2):107–118, 2011.

[34] Sebastian Pölsterl, Pankaj Gupta, Lichao Wang, Sailesh Conjeti, Amin Katouzian, and Nassir Navab. Heterogeneous ensembles for predicting survival of metastatic, castrate-resistant prostate cancer patients. *F1000Research*, 5:2676, 2017.

[35] Markus S. Schröder, Aedín C. Culhane, John Quackenbush, and Benjamin Haibe-Kains. survcomp: An R/Bioconductor package for performance assessment and comparison of survival models. *Bioinformatics*, 27(22):3206–3208, 2011.

[36] Michael J Pencina and Ralph B. D'Agostino. OverallC as a measure of discrimination in survival analysis: model specific population value and confidence interval estimation. *Statistics in Medicine*, 23(13):2109–2123, jul 2004.

[37] Michael I. Love, Wolfgang Huber, and Simon Anders. Moderated estimation of fold change and dispersion for RNA-seq data with DESeq2. *Genome Biology*, 15(12), 2014.

[38] Gabriela Bindea, Bernhard Mlecnik, Marie Tosolini, Amos Kirilovsky, Maximilian Waldner, Anna C. Obenauf, Helen Angell, Tessa Fredriksen, Lucie Lafontaine, Anne Berger, Patrick Bruneval, Wolf Herman Fridman, Christoph Becker, Franck Pagès, Michael R. Speicher, Zlatko Trajanoski, and Jér Ôme Galon. Spatiotemporal dynamics of intratumoral immune cells reveal the immune landscape in human cancer. *Immunity*, 39(4):782–795, 2013.

[39] Daniel R. Zerbino, Premanand Achuthan, Wasiu Akanni, M. Ridwan Amode, Daniel Barrell, Jyothish Bhai, Konstantinos Billis, Carla Cummins, Astrid Gall, Carlos García Girón, Laurent Gil, Leo Gordon, Leanne Haggerty, Erin Haskell, Thibaut Hourlier, Osagie G. Izuogu, Sophie H. Janacek, Thomas Juettemann, Jimmy Kiang To, Matthew R. Laird, Ilias Lavidas, Zhicheng Liu, Jane E. Loveland, Thomas Maurel, William McLaren, Benjamin Moore, Jonathan Mudge, Daniel N. Murphy, Victoria Newman, Michael Nuhn, Denye Ogeh, Chuang Kee Ong, Anne Parker, Mateus Patricio, Harpreet Singh Riat, Helen Schuilenburg, Dan Sheppard, Helen Sparrow, Kieron Taylor, Anja Thormann, Alessandro Vullo, Brandon Walts, Amonida Zadissa, Adam Frankish, Sarah E. Hunt, Myrto Kostadima, Nicholas Langridge, Fergal J. Martin, Matthieu Muffato, Emily Perry, Magali Ruffier, Dan M. Staines, Stephen J. Trevanion, Bronwen L. Aken, Fiona Cunningham, Andrew Yates, and Paul Flicek. Ensembl 2018. *Nucleic Acids Research*, 46(D1):D754–D761, 2018.

[40] Adriana Salcedo, Maxime Tarabichi, Shadrielle Melijah G Espiritu, Amit G Deshwar, Alex Buchanan, Christopher M Lalansingh, P Christine, and Catalina V Anghel. Creating Standards for Evaluating Tumour Subclonal Reconstruction. 2018.

[41] Andrew Rosenberg and Julia Hirschberg. V-measure: A conditional entropy-based external cluster evaluation measure. *Proceedings of the Joint Conference on Empirical Methods in Natural Language Processing and Computational Natural Language (EMNLP-CoNLL'07)*, 1(June):410–420, 2007.

[42] Salem Malikic, Katharina Jahn, Jack Kuipers, S. Cenk Sahinalp, and Niko Beerenwinkel. Integrative inference of subclonal tumour evolution from single-cell and bulk sequencing data. *Nature Communications*, 10(1):1–12, 2019.

[43] Charles Gawad, Winston Koh, and Stephen R. Quake. Dissecting the clonal origins of childhood acute lymphoblastic leukemia by single-cell genomics. *Proceedings of the National Academy of Sciences of the United States of America*, 111(50):17947–17952, 2014.

[44] Yong Wang, Jill Waters, Marco L. Leung, Anna Unruh, Whijae Roh, Xiuqing Shi, Ken Chen, Paul Scheet, Selina Vattathil, Han Liang, Asha Multani, Hong Zhang, Rui Zhao, Franziska Michor, Funda Meric-Bernstam, and Nicholas E. Navin. Clonal evolution in breast cancer revealed by single nucleus genome sequencing. *Nature*, 512(7513):155–160, 2014.

[45] Marco L. Leung, Alexander Davis, Ruli Gao, Anna Casasent, Yong Wang, Emi Sei, Eduardo Vilar, Dipen Maru, Scott Kopetz, and Nicholas E. Navin. Single-cell DNA sequencing reveals a latedissemination model in metastatic colorectal cancer. *Genome Research*, 27(8):1287–1299, 2017.

[46] Heng Li and Richard Durbin. Fast and accurate short read alignment with Burrows Wheeler transform. 25(14):1754–1760, 2009.

[47] Broad Institute. Picard tools. `http://broadinstitute.github.io/picard/`.

[48] Aaron R Quinlan and Ira M Hall. BEDTools : a flexible suite of utilities for comparing genomic features. 26(6):841–842, 2010.

[49] Heng Li, Bob Handsaker, Alec Wysoker, Tim Fennell, Jue Ruan, Nils Homer, Gabor Marth, Goncalo Abecasis, Richard Durbin, Genome Project Data, and The Sam. The Sequence Alignment / Map format and SAMtools. 25(16):2078–2079, 2009.

[50] Aaron Mckenna, Matthew Hanna, Eric Banks, Andrey Sivachenko, Kristian Cibulskis, Andrew Kernytsky, Kiran Garimella, David Altshuler, Stacey Gabriel, Mark Daly, and Mark A Depristo. The Genome Analysis Toolkit : A MapReduce framework for analyzing next-generation DNA sequencing data. pages 1297–1303, 2010.

[51] Kristian Cibulskis, Michael S. Lawrence, Scott L. Carter, Andrey Sivachenko, David Jaffe, Carrie Sougnez, Stacey Gabriel, Matthew Meyerson, Eric S. Lander, and Gad Getz. Sensitive detection of somatic point mutations in impure and heterogeneous cancer samples. *Nature Biotechnology*, 31(3):213–219, 2013.

[52] Ronglai Shen and Venkatraman E Seshan. FACETS : allele-specific copy number and clonal heterogeneity analysis tool for high-throughput DNA sequencing. (8):1–9, 2016.

[53] Bernard Pereira, Suet Feung Chin, Oscar M. Rueda, Hans Kristian Moen Vollan, Elena Provenzano, Helen A. Bardwell, Michelle Pugh, Linda Jones, Roslin Russell, Stephen John Sammut, Dana W.Y. Tsui, Bin Liu, Sarah Jane Dawson, Jean Abraham, Helen Northen, John F. Peden, Abhik Mukherjee, Gulisa Turashvili, Andrew R. Green, Steve McKinney, Arusha Oloumi, Sohrab Shah, Nitzan Rosenfeld, Leigh Murphy, David R. Bentley, Ian O. Ellis, Arnie Purushotham, Sarah E. Pinder, Anne Lise BØrresen-Dale, Helena M. Earl, Paul D. Pharoah, Mark T. Ross, Samuel Aparicio, and Carlos Caldas. The somatic mutation profiles of 2,433 breast cancers refines their genomic and transcriptomic landscapes. *Nature Communications*, 7(May):11479, 2016.

[54] Thomas Karn, Tingting Jiang, Christos Hatzis, Nicole Sänger, Ahmed El-Balat, Achim Rody, Uwe Holtrich, Sven Becker, Giampaolo Bianchini, and Lajos Pusztai. Association Between Genomic Metrics and Immune Infiltration in Triple-Negative Breast Cancer. *JAMA Oncology*, 3(12):1707–1711, 2017.

[55] Subramanian Venkatesan and Charles Swanton. Tumor Evolutionary Principles: How Intratumor Heterogeneity Influences Cancer Treatment and Outcome. *American Society of Clinical Oncology Educational Book*, 36:e141–e149, 2016.

[56] Jesse Eaton, Jingyi Wang, and Russell Schwartz. Deconvolution and phylogeny inference of structural variations in tumor genomic samples. *Bioinformatics*, 34(13):i357–i365, 2018.

[57] Vinayak Bhandari, Lydia Y Liu, Adriana Salcedo, Shadrielle M G Espiritu, Quaid D Morris, and Paul C Boutros. The Inter and Intra-Tumoural Heterogeneity of Subclonal Reconstruction. *bioRxiv*, 2018.

[58] Samra Turajlic and Charles Swanton. TRACERx Renal: tracking renal cancer evolution through therapy. *Nature Reviews Urology*, 14(10):575–576, jul 2017.

[59] Charissa Kim, Ruli Gao, Emi Sei, Rachel Brandt, Johan Hartman, Thomas Hatschek, Nicola Crosetto, Theodoros Foukakis, and Nicholas E. Navin. Chemoresistance Evolution in Triple-Negative Breast Cancer Delineated by Single-Cell Sequencing. *Cell*, 173(4):879–893.e13, 2018.

[60] Weiwei Shi, Charlotte K.Y. Ng, Raymond S Lim, Tingting Jiang, Sushant Kumar, Xiaotong Li, Vikram B. Wali, Salvatore Piscuoglio, Mark B. Gerstein, Anees Chagpar, Britta Weigelt, Lajos Pusztai, Jorge S. Reis-Filho, and Christos Hatzis. Reliability of Whole-Exome Sequencing for Assessing Intratumor Genetic Heterogeneity. 2018.

[61] Hengbo Zhou, Deepika Neelakantan, and Heide L. Ford. Clonal cooperativity in heterogenous cancers. *Seminars in Cell & Developmental Biology*, 64:79–89, apr 2017.

[62] Nicholas McGranahan, Francesco Favero, Elza C. De Bruin, Nicolai Juul Birkbak, Zoltan Szallasi, and Charles Swanton. Clonal status of actionable driver events and the timing of mutational processes in cancer evolution. *Science Translational Medicine*, 7(283):283ra54–283ra54, 2015.

[63] Jonathan J. Keats, Marta Chesi, Jan B. Egan, Victoria M. Garbitt, Stephen E. Palmer, Esteban Braggio, Scott Van Wier, Patrick R. Blackburn, Angela S. Baker, Angela Dispenzieri, Shaji Kumar, S. Vincent Rajkumar, John D. Carpten, Michael Barrett, Rafael Fonseca, A. Keith Stewart, and P. Leif Bergsagel. Clonal competition with alternating dominance in multiple myeloma. *Blood*, 120(5):1067–1076, aug 2012.

[64] Jacob Scott and Andriy Marusyk. Somatic clonal evolution: A selection-centric perspective. *Biochimica et Biophysica Acta - Reviews on Cancer*, 1867(2):139–150, 2017.

[65] William C.H. Cross, Trevor A. Graham, and Nicholas A. Wright. New paradigms in clonal evolution: punctuated equilibrium in cancer. *Journal of Pathology*, 240(2):126–136, 2016.

[66] Andrea Sottoriva, Chris P. Barnes, and Trevor A. Graham. Catch my drift? Making sense of genomic intra-tumour heterogeneity. *Biochimica et Biophysica Acta (BBA) - Reviews on Cancer*, 1867(2):95–100, apr 2017.

[67] Carlo C. Maley, Athena Aktipis, Trevor A. Graham, Andrea Sottoriva, Amy M. Boddy, Michalina Janiszewska, Ariosto S. Silva, Marco Gerlinger, Yinyin Yuan, Kenneth J. Pienta, Karen S. Anderson, Robert Gatenby, Charles Swanton, David Posada, Chung-I Wu, Joshua D. Schiffman, E. Shelley Hwang, Kornelia Polyak, Alexander R. A. Anderson, Joel S. Brown, Mel Greaves, and Darryl Shibata. Classifying the evolutionary and ecological features of neoplasms. *Nature Reviews Cancer*, 17(10):605–619, sep 2017.

[68] Anton Safonov, Tingting Jiang, Giampaolo Bianchini, Balázs Gyrffy, Thomas Karn, Christos Hatzis, and Lajos Pusztai. Immune Gene Expression Is Associated with Genomic Aberrations in Breast Cancer. *Cancer Research*, 77(12):3317–3324, 2017.

[69] Giulio Caravagna, Timon Heide, Marc Williams, Luis Zapata, Daniel Nichol, Ketevan Chkhaidze, George D Cresswell, Benjamin Werner, Ahmet Acar, Chris P Barnes, Guido Sanguinetti, A Trevor, and Andrea Sottoriva. Model-based tumor subclonal reconstruction. pages 1–31, 2019.

Supplementary information

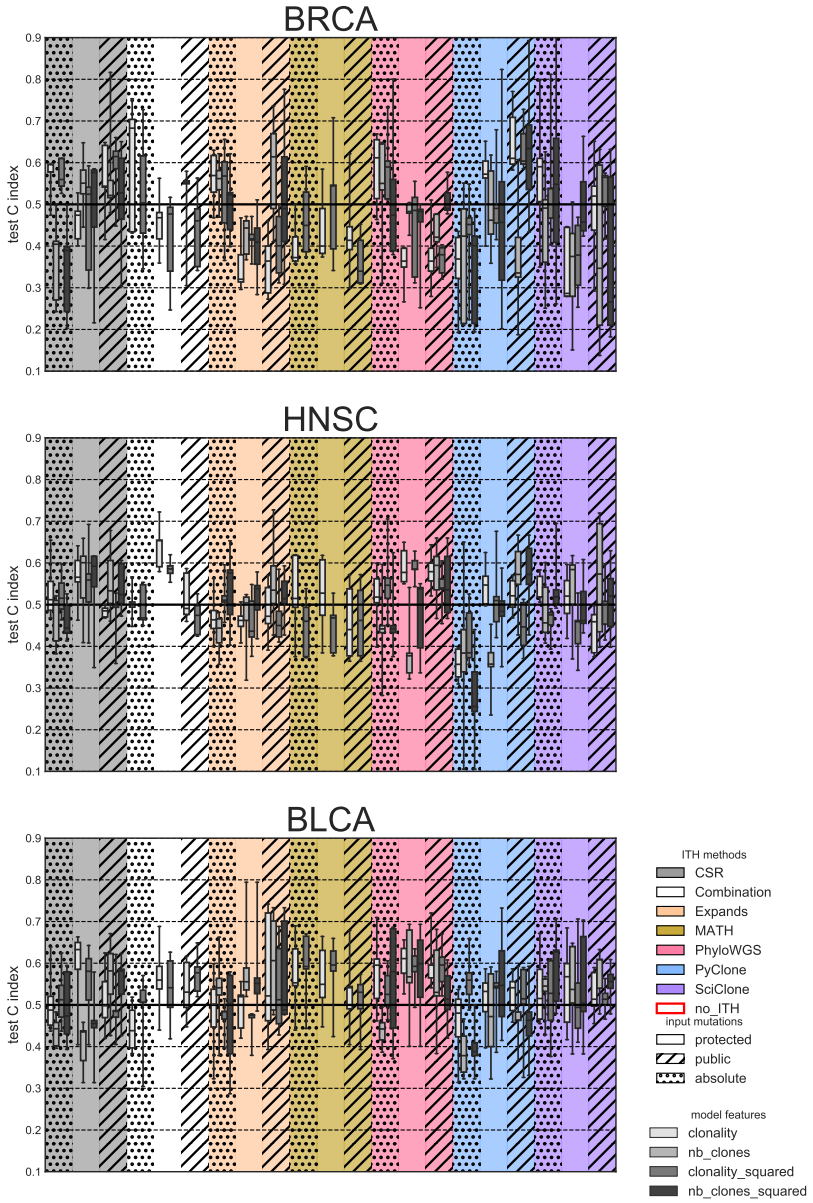

Figure S1: Prognostic power of diverse combination of ITH-derived features, on the three cancer types (respectively BRCA, HNSC and BLCA from top to bottom). In each plot, the background color indicates the ITH method used. Each method is tested on protected or public mutations (hashed). For each method, we assess the ability to predict survival with a survival SVM using 4 sets of features: (i) the number of clones alone, (ii) the five custom features which include the number of clones, and (iii) and (iv) the concatenations of features in (i) and (ii) with their squares, to account for possible nonlinear quadratic effects. We observe no clear trend of one of the two sets performs systematically better than the other, and the squared features have not significantly improved results either.

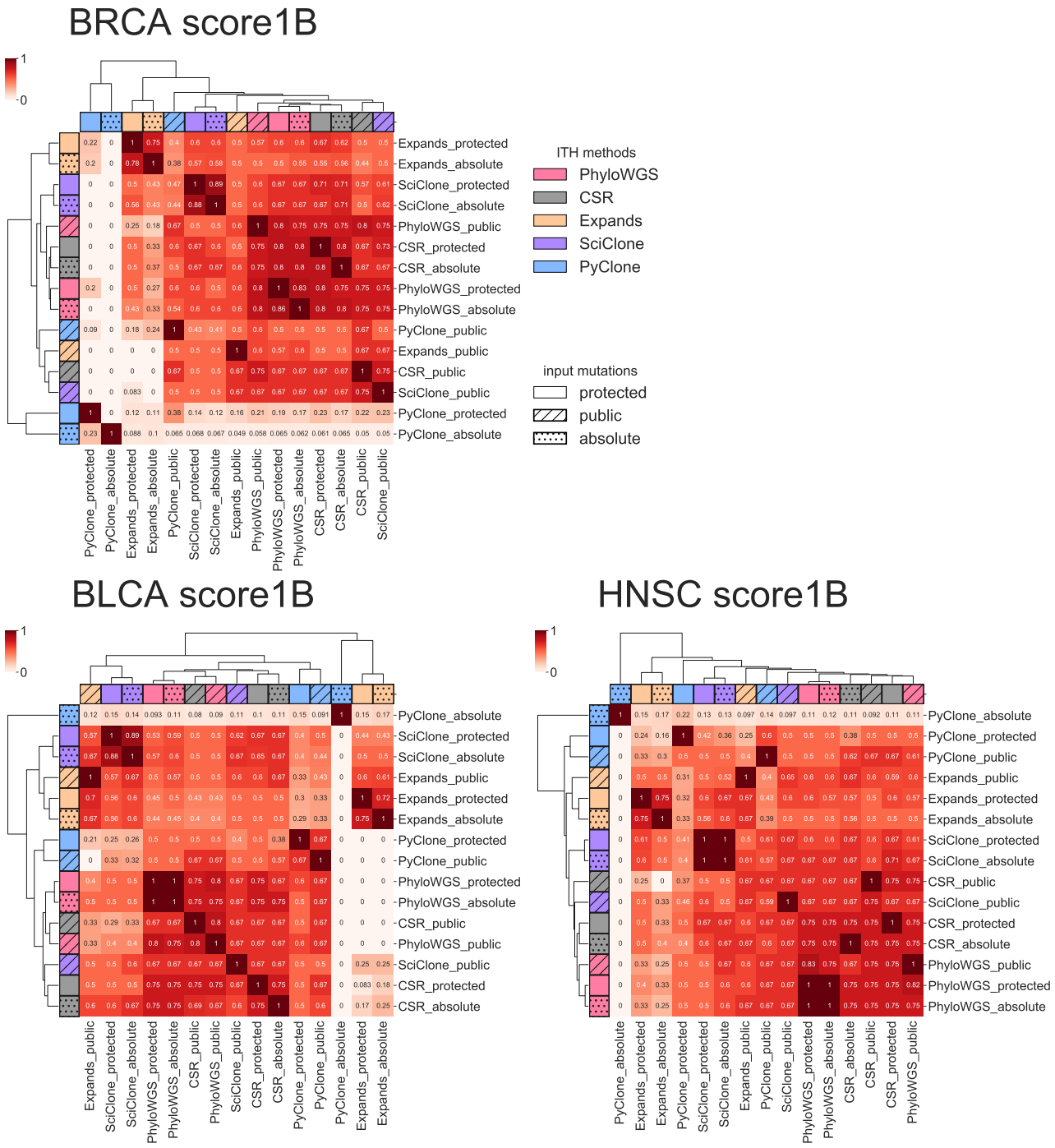

Figure S2: Pairwise computation of score1B for the different ITH methods and inputs. Score1B is a metric designed in [40] penalizes differences between the number of clones inferred in each case in a symmetric way (only the difference matters, either more or fewer clones are detected), following the formula $\frac{J_1+1-\min(J_1+1,|J_2-J_1|)}{J_1+1}$, with J_1 and J_2 the numbers of clones found by each method. The score was computed for all patients, and this heatmap represents the median score. We observe a particular feature of PyClone, which tends to find a lot (sometimes several dozens) of clones with only one mutation. They were discarded when comparing the number of clones, but not for the computation of metric 1B to ensure consistency with the other metrics.

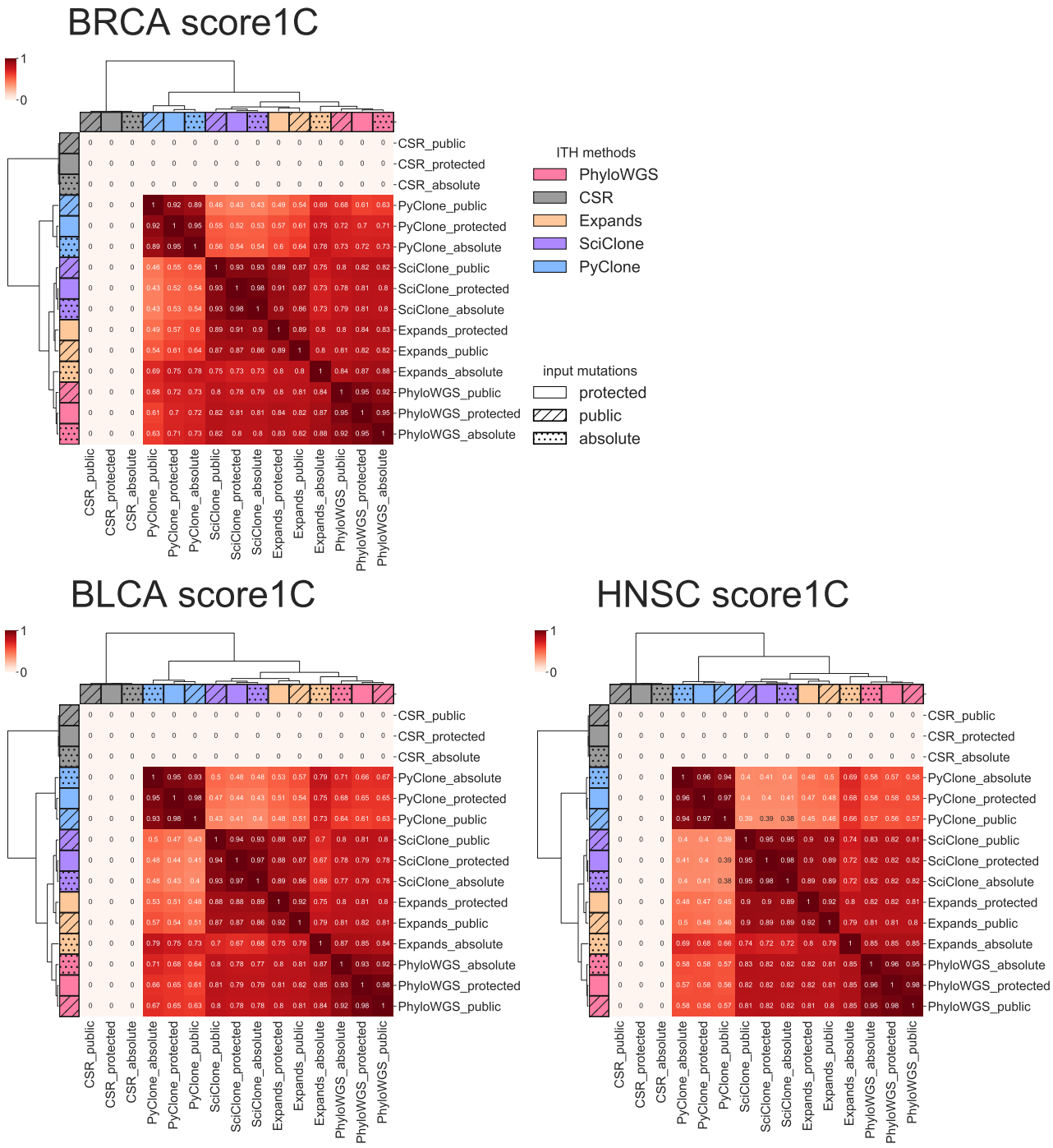

Figure S3: Pairwise computation of score1C for the different ITH methods and inputs. Score1C is a metric designed in [40] that represents the Wasserstein distance between the cancer cell fraction (CCF) distribution resulting from each clone's mean CCF and number of mutations. Due to the number of single-mutation clones of PyClone, the resulting distribution is quite different from the other cases. As CSR only takes as input the mutation attribution to clones by other methods, without taking into account their CCF, we did not compute score1C for that method. The score was computed for all patients, and this heatmap represents the median score.

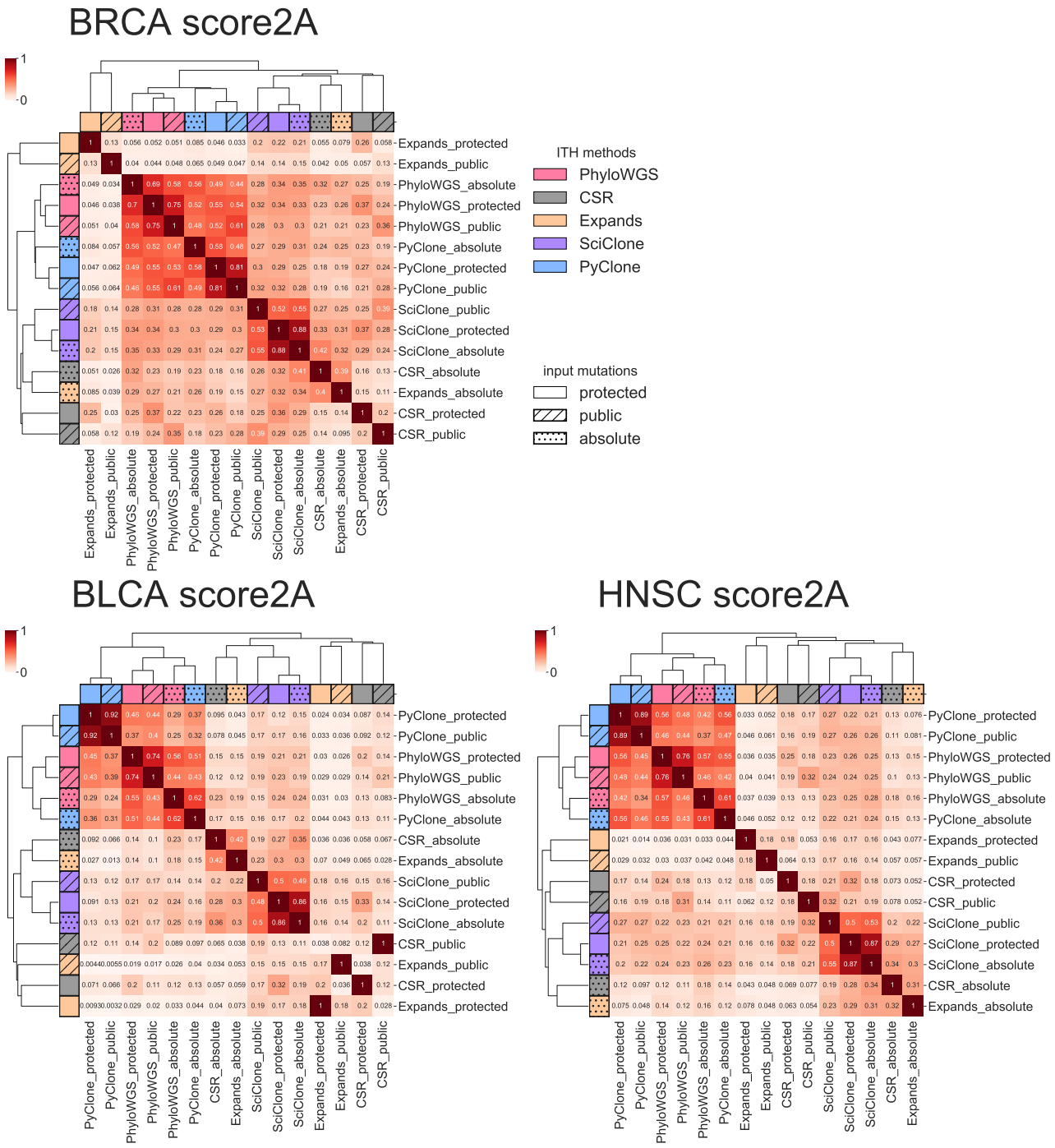

Figure S4: Pairwise computation of score2A for the different ITH methods and inputs. Score2A is a metric designed in [40] that assesses the similarity of the mutation clustering resulting from subclonal reconstruction (see Methods for details). We recover the previously observed pattern that PyClone and PhyloWGS are the closest methods. The score was computed for all patients, and this heatmap represents the median score.

	BRCA	BLCA	HNSC
number of samples	962	351	445
Protected mutations			
average	530	735	456
std	1,189	890	544
min	79	61	31
median	342	525	340
max	21,821	12,774	7,941
Public mutations			
average	121	352	202
std	375	426	271
min	1	2	1
median	63	241	140
max	7,919	5,478	3,935

Table S1: Summary statistics of the number of protected and public mutations per sample for BRCA, BLCA and HNSC samples. The protected set corresponds to raw variant calling outputs. The public set corresponds to publicly available SNV calls.

Variable	Hazard ratio	P-value	Corrected P-value
age_at_diagnosis	1.032023	4.328538e-05	0.000577
PROSPECTIVE_COLLECTION_NO	0.885701	4.431127e-01	0.631390
PROSPECTIVE_COLLECTION_YES	1.129050	4.431127e-01	0.631390
RETROSPECTIVE_COLLECTION_NO	1.138738	4.114412e-01	0.631390
RETROSPECTIVE_COLLECTION_YES	0.885701	4.431127e-01	0.631390
SEX_Female	1.114329	5.131653e-01	0.651638
SEX_Male	0.897401	5.131653e-01	0.651638
RACE_WHITE	1.144930	5.054529e-01	0.651638
HISTORY_OTHER_MALIGNANCY_No	1.100403	5.660149e-01	0.696634
HISTORY_OTHER_MALIGNANCY_Yes	0.908758	5.660149e-01	0.696634
NONINVASIVE_BLADDER_HISTORY_NO	0.878835	3.891088e-01	0.610367
NONINVASIVE_BLADDER_HISTORY_YES	0.868563	4.940541e-01	0.651638
NONINVASIVE_BLADDER_HISTORY_[Not Available]	1.298013	1.113381e-01	0.329891
NONINVASIVE_BLADDER_CA_TX_TYPE_[Not Applicable]	0.878835	3.891088e-01	0.610367
NONINVASIVE_BLADDER_CA_TX_TYPE_[Not Available]	1.225435	2.146545e-01	0.490639
TX_90DAYS_POST_RESECTION_[Not Applicable]	0.878835	3.891088e-01	0.610367
TX_90DAYS_POST_RESECTION_[Not Available]	1.329106	6.724015e-02	0.244510
TX_COMPLETE_RESPONSE_[Not Applicable]	0.878835	3.891088e-01	0.610367
TX_COMPLETE_RESPONSE_[Not Available]	1.183784	2.774378e-01	0.539748
TX_INDUCTION_COURSES_INDICATOR_[Not Applicable]	0.878835	3.891088e-01	0.610367
TX_INDUCTION_COURSES_INDICATOR_[Not Available]	1.234697	1.748396e-01	0.436750
TX_MAINTENANCE_COURSES_INDICATOR_[Not Applicable]	0.878835	3.891088e-01	0.610367
TX_MAINTENANCE_COURSES_INDICATOR_[Not Available]	1.233463	1.768826e-01	0.436750
OCCUPATION_CURRENT_Retired	0.917140	6.255270e-01	0.733401
OCCUPATION_CURRENT_[Not Available]	1.271108	1.467738e-01	0.404893
OCCUPATION_CURRENT_retired	0.960885	8.446695e-01	0.889126
OCCUPATION_PRIMARY_[Not Available]	1.295025	9.419336e-02	0.301419
OCCUPATION_PRIMARY_CHEMICAL_EXPOSURE_[Not Avail...	1.765620	4.091600e-03	0.032733

OCCUPATION_PRIMARY_INDUSTRY_[Not Available]	1.160094	3.247413e-01	0.590439
FAMILY_HISTORY_CANCER_RELATIONSHIP_[Not Available]	1.226713	1.856187e-01	0.436750
FAMILY_HISTORY_CANCER_TYPE_[Not Available]	1.226713	1.856187e-01	0.436750
RADIATION_TREATMENT_ADJUVANT_NO	0.712057	2.632089e-02	0.131604
RADIATION_TREATMENT_ADJUVANT_[Not Available]	1.354145	5.230845e-02	0.210985
PHARMACEUTICAL_TX_ADJUVANT_NO	0.986814	9.296012e-01	0.941368
PHARMACEUTICAL_TX_ADJUVANT_YES	0.676359	4.539278e-02	0.201746
PHARMACEUTICAL_TX_ADJUVANT_[Not Available]	1.387004	3.427365e-02	0.161288
HISTOLOGICAL_SUBTYPE_Non-Papillary	1.364821	7.158132e-02	0.248979
HISTOLOGICAL_SUBTYPE_Papillary	0.673148	2.630851e-02	0.131604
METHOD_OF_INITIAL_SAMPLE_PROCUREMENT_Other meth...	0.813722	4.457771e-01	0.631390
METHOD_OF_INITIAL_SAMPLE_PROCUREMENT_Transureth...	1.234871	2.692175e-01	0.539748
METHOD_OF_INITIAL_SAMPLE_PROCUREMENT_OTHER_[Not...	1.290044	3.459972e-01	0.610367
AJCC_STAGING_EDITION_6th	0.931002	6.427039e-01	0.733401
AJCC_STAGING_EDITION_7th	1.061071	7.031650e-01	0.760178
ANGIOLYMPHATIC_INVASION_NO	0.485392	3.934074e-05	0.000577
ANGIOLYMPHATIC_INVASION_YES	1.818002	6.992848e-05	0.000799
ANGIOLYMPHATIC_INVASION_[Not Available]	1.282364	1.626703e-01	0.433787
LYMPH_NODES_EXAMINED_NO	0.861162	4.914651e-01	0.651638
LYMPH_NODES_EXAMINED_YES	1.144075	4.498650e-01	0.631390
EXTRACAPSULAR_EXTENSION_NO	1.101253	5.846915e-01	0.708717
EXTRACAPSULAR_EXTENSION_YES	1.411088	5.538363e-02	0.210985
EXTRACAPSULAR_EXTENSION_[Not Available]	0.748263	5.291460e-02	0.210985
EXTRACAPSULAR_EXTENSION_PRESENT_[Not Available]	0.804289	2.901143e-01	0.539748
METASTATIC_SITE_Lymph node only	1.687527	6.731283e-03	0.041832
METASTATIC_SITE_None	0.641938	5.729977e-03	0.041673
METASTATIC_SITE_[Not Available]	1.050773	7.417814e-01	0.791234
AJCC_PATHOLOGIC_TUMOR_STAGE_Stage II	0.457221	3.329395e-05	0.000577
AJCC_PATHOLOGIC_TUMOR_STAGE_Stage III	0.840677	2.814664e-01	0.539748
AJCC_PATHOLOGIC_TUMOR_STAGE_Stage IV	2.264256	4.728912e-08	0.000004
INCIDENTAL_PROSTATE_CANCER_NO	1.002834	9.852841e-01	0.985284
INCIDENTAL_PROSTATE_CANCER_YES	0.926590	6.685831e-01	0.742870
INCIDENTAL_PROSTATE_CANCER_[Not Available]	1.107850	6.188245e-01	0.733401
AJCC_INCIDENTAL_PROSTATE_CANCER_[Not Available]	1.135748	4.866109e-01	0.651638
PRIMARY_SITE_Bladder - NOS	0.839975	2.465470e-01	0.533075
CLIN_T_STAGE_T2	1.072482	6.867865e-01	0.752643
CLIN_T_STAGE_[Not Available]	0.801825	1.408015e-01	0.402290
ICD_10_C67.2	0.981491	9.276451e-01	0.941368
ICD_10_C67.9	0.934338	6.508937e-01	0.733401
ICD_O_3_HISTOLOGY_8120/3	1.264392	2.765995e-01	0.539748
ICD_O_3_HISTOLOGY_8130/3	0.790498	2.838725e-01	0.539748
ICD_O_3_SITE_C67.2	0.981491	9.276451e-01	0.941368
ICD_O_3_SITE_C67.9	0.934338	6.508937e-01	0.733401
TISSUE_SOURCE_SITE_DK	0.563206	2.178870e-02	0.124507
TISSUE_SOURCE_SITE_XF	1.263603	2.246627e-01	0.499250
AJCC_TUMOR_PATHOLOGIC_PT_simple_T2	0.515628	4.470852e-04	0.004471
AJCC_TUMOR_PATHOLOGIC_PT_simple_T3	1.271201	1.094625e-01	0.329891
AJCC_TUMOR_PATHOLOGIC_PT_simple_T4	1.880901	9.411876e-04	0.008366
AJCC_NODES_PATHOLOGIC_PN_simple_N0	0.450186	1.450583e-07	0.000006
AJCC_NODES_PATHOLOGIC_PN_simple_N2	2.130213	5.222366e-06	0.000139
AJCC_METASTASIS_PATHOLOGIC_PM_simple_M0	0.663211	6.797682e-03	0.041832
AJCC_METASTASIS_PATHOLOGIC_PM_simple_MX	1.305408	7.476644e-02	0.249221

Table S2: Clinical variables significance for single-variable cox model for BLCA (409 patients). Variable significantly associated with survival are shaded.

Variable	Hazard ratio	P-value	Corrected P-value
age_at_diagnosis	1.032118	4.566898e-07	1.077788e-05
PROSPECTIVE_COLLECTION_NO	0.602422	4.530180e-02	1.243166e-01
PROSPECTIVE_COLLECTION_YES	1.686136	3.888498e-02	1.176520e-01
RETROSPECTIVE_COLLECTION_NO	1.686136	3.888498e-02	1.176520e-01
RETROSPECTIVE_COLLECTION_YES	0.602422	4.530180e-02	1.243166e-01
MENOPAUSE_STATUS_Post (prior bilateral ovariect...	1.324224	9.572180e-02	2.026486e-01
MENOPAUSE_STATUS_Pre (<6 months since LMP AND n...	0.431889	8.809889e-04	5.887468e-03
MENOPAUSE_STATUS_[Not Available]	2.943808	9.817809e-07	1.930836e-05
RACE_BLACK OR AFRICAN AMERICAN	1.215466	3.390936e-01	5.275821e-01
RACE_WHITE	0.825045	2.987240e-01	4.828689e-01
RACE_[Not Available]	1.274439	5.361448e-01	6.741716e-01
ETHNICITY_NOT HISPANIC OR LATINO	2.207632	1.176302e-02	5.552148e-02
ETHNICITY_[Not Available]	0.603636	1.259999e-01	2.447704e-01
HISTORY_OTHER_MALIGNANCY_No	0.669493	2.469654e-01	4.047489e-01
HISTORY_OTHER_MALIGNANCY_Yes	1.501028	2.411757e-01	4.016651e-01
RADIATION_TREATMENT_ADJUVANT_NO	0.825072	6.454644e-01	7.693415e-01
RADIATION_TREATMENT_ADJUVANT_YES	1.184551	6.240900e-01	7.525332e-01
RADIATION_TREATMENT_ADJUVANT_[Not Available]	0.991267	9.744419e-01	9.936324e-01
PHARMACEUTICAL_TX_ADJUVANT_YES	0.893189	7.194462e-01	8.162947e-01
PHARMACEUTICAL_TX_ADJUVANT_[Not Available]	1.001332	9.961223e-01	9.961223e-01
METHOD_OF_INITIAL_SAMPLE_PROCUREMENT_Core needl...	0.545734	3.371119e-04	2.841372e-03
METHOD_OF_INITIAL_SAMPLE_PROCUREMENT_Fine needl...	1.669895	1.358907e-02	5.938927e-02
METHOD_OF_INITIAL_SAMPLE_PROCUREMENT_Other meth...	1.009070	9.765300e-01	9.936324e-01
METHOD_OF_INITIAL_SAMPLE_PROCUREMENT_Tumor rese...	0.740284	3.397987e-01	5.275821e-01
METHOD_OF_INITIAL_SAMPLE_PROCUREMENT_[Not Avail...	1.944929	3.630026e-02	1.157684e-01
METHOD_OF_INITIAL_SAMPLE_PROCUREMENT_OTHER_[Not...	0.994946	9.868232e-01	9.952576e-01
SURGICAL_PROCEDURE_FIRST_Lumpectomy	0.690083	9.704508e-02	2.026486e-01
SURGICAL_PROCEDURE_FIRST_Modified Radical Maste...	1.476196	2.170723e-02	8.004541e-02
SURGICAL_PROCEDURE_FIRST_Other	0.882243	5.027363e-01	6.448140e-01
SURGICAL_PROCEDURE_FIRST_Simple Mastectomy	0.657380	9.788958e-02	2.026486e-01
SURGICAL_PROCEDURE_FIRST_[Not Available]	2.480055	2.729308e-03	1.463901e-02
FIRST_SURGICAL_PROCEDURE_OTHER_Surgical Resection	1.534372	3.107861e-01	4.955778e-01
FIRST_SURGICAL_PROCEDURE_OTHER_[Not Available]	1.084261	6.601793e-01	7.790115e-01
PATH_MARGIN_Negative	0.328745	2.487080e-11	1.467377e-09
PATH_MARGIN_Positive	1.456750	1.265338e-01	2.447704e-01
PATH_MARGIN_[Not Available]	4.295169	1.280242e-14	1.510685e-12
SURGERY_FOR_POSITIVE_MARGINS_[Not Available]	0.881873	7.014933e-01	8.141488e-01
MARGIN_STATUS_REEXCISION_Negative	1.189978	5.498957e-01	6.830283e-01
MARGIN_STATUS_REEXCISION_[Not Available]	0.788031	3.970790e-01	5.714063e-01
STAGING_SYSTEM_Axillary lymph node dissection a...	1.147628	4.478192e-01	6.047185e-01
STAGING_SYSTEM_Sentinel lymph node biopsy plus ...	0.471890	9.803200e-04	6.088303e-03
STAGING_SYSTEM_Sentinel node biopsy alone	0.517762	9.212199e-03	4.529331e-02
STAGING_SYSTEM_[Not Available]	2.306054	2.789049e-06	3.502342e-05
MICROMET_DETECTION_BY_IHC_NO	0.983191	9.205373e-01	9.860025e-01
MICROMET_DETECTION_BY_IHC_YES	0.330809	1.172534e-05	1.257809e-04
MICROMET_DETECTION_BY_IHC_[Not Available]	2.187550	2.968086e-06	3.502342e-05
LYMPH_NODES_EXAMINED_YES	1.127012	4.842736e-01	6.279592e-01

LYMPH_NODES_EXAMINED_[Not Available]	0.789955	1.767459e-01	3.160002e-01
AJCC_STAGING_EDITION_5th	2.522972	2.425021e-06	3.502342e-05
AJCC_STAGING_EDITION_6th	0.391254	3.279076e-07	9.673274e-06
AJCC_STAGING_EDITION_7th	1.553473	5.971021e-02	1.455199e-01
AJCC_STAGING_EDITION_[Not Available]	0.905993	7.048542e-01	8.141488e-01
AJCC_PATHOLOGIC_TUMOR_STAGE_Stage I	0.623674	1.050290e-01	2.136796e-01
AJCC_PATHOLOGIC_TUMOR_STAGE_Stage IA	0.246046	1.650590e-02	6.956059e-02
AJCC_PATHOLOGIC_TUMOR_STAGE_Stage IIA	0.697084	6.472348e-02	1.510602e-01
AJCC_PATHOLOGIC_TUMOR_STAGE_Stage IIB	0.828285	3.501038e-01	5.365227e-01
AJCC_PATHOLOGIC_TUMOR_STAGE_Stage IIIA	1.293391	2.416799e-01	4.016651e-01
AJCC_PATHOLOGIC_TUMOR_STAGE_Stage IIIC	1.898479	6.528872e-02	1.510602e-01
ER_STATUS_BY_IHC_Negative	1.272320	1.906432e-01	3.308221e-01
ER_STATUS_BY_IHC_Positive	0.662619	1.727163e-02	7.027765e-02
ER_STATUS_IHC_PERCENT_POSITIVE_90-99%	0.370064	2.554685e-03	1.435490e-02
ER_STATUS_IHC_PERCENT_POSITIVE_¡10%	0.771901	4.520339e-01	6.047185e-01
ER_STATUS_IHC_PERCENT_POSITIVE_[Not Available]	2.276830	1.936838e-05	1.758053e-04
ER_POSITIVITY_SCALE_USED_3 Point Scale	0.369684	5.031369e-02	1.349322e-01
ER_POSITIVITY_SCALE_USED_[Not Available]	2.435285	3.311614e-02	1.116487e-01
ER_POSITIVITY_SCALE_OTHER_[Not Available]	1.595260	1.827359e-01	3.218334e-01
BRACHYTHERAPY_TOTAL_DOSE_POINT_A_[Not Available]	0.994099	9.767911e-01	9.936324e-01
PR_STATUS_BY_IHC_Negative	1.257745	1.746637e-01	3.160002e-01
PR_STATUS_BY_IHC_Positive	0.717066	4.336782e-02	1.243166e-01
PR_STATUS_IHC_PERCENT_POSITIVE_90-99%	0.082244	1.281943e-02	5.818049e-02
PR_STATUS_IHC_PERCENT_POSITIVE_¡10%	0.987074	9.609962e-01	9.936324e-01
PR_STATUS_IHC_PERCENT_POSITIVE_[Not Available]	2.440652	1.832734e-05	1.758053e-04
PR_POSITIVITY_SCALE_USED_3 Point Scale	0.401865	7.305251e-02	1.626452e-01
PR_POSITIVITY_SCALE_USED_[Not Available]	2.235079	5.428402e-02	1.411211e-01
PR_POSITIVITY_IHC_INTENSITY_SCORE_3+	0.127355	4.012324e-02	1.183636e-01
PR_POSITIVITY_IHC_INTENSITY_SCORE_[Not Available]	3.001282	2.553857e-03	1.435490e-02
PR_POSITIVITY_SCALE_OTHER_[Not Available]	1.327000	3.966578e-01	5.714063e-01
PR_POSITIVITY_DEFINE_METHOD_[Not Available]	1.008404	9.676018e-01	9.936324e-01
IHC_HER2_Equivocal	0.601537	5.547120e-02	1.411211e-01
IHC_HER2_Negative	0.791443	1.711526e-01	3.155626e-01
IHC_HER2_Positive	1.504160	7.472326e-02	1.632842e-01
IHC_HER2_[Not Available]	1.460324	3.585853e-02	1.157684e-01
HER2_IHC_PERCENT_POSITIVE_¡10%	0.408510	2.110215e-02	8.004541e-02
HER2_IHC_PERCENT_POSITIVE_[Not Available]	1.864321	2.281919e-02	8.159588e-02
HER2_POSITIVITY_METHOD_TEXT_[Not Available]	2.593860	6.042777e-02	1.455199e-01
HER2_FISH_STATUS_Negative	0.739080	1.150908e-01	2.301816e-01
HER2_FISH_STATUS_Positive	0.880007	7.106553e-01	8.141488e-01
HER2_FISH_STATUS_[Not Available]	1.385428	6.801151e-02	1.543338e-01
HER2_COPY_NUMBER_[Not Available]	1.645350	1.487810e-01	2.831638e-01
CENT17_COPY_NUMBER_[Not Available]	1.548415	2.047798e-01	3.502032e-01
PRIMARY_SITE_Left	1.945749	7.591896e-04	5.599023e-03
PRIMARY_SITE_Left Upper Inner Quadrant	0.734898	4.268913e-01	6.047185e-01
PRIMARY_SITE_Left Upper Outer Quadrant	0.826630	4.397503e-01	6.047185e-01
PRIMARY_SITE_Right	0.892605	6.249852e-01	7.525332e-01
PRIMARY_SITE_Right Upper Outer Quadrant	1.138461	6.005181e-01	7.381368e-01
HISTOLOGICAL_DIAGNOSIS_Infiltrating Ductal Carc...	0.969817	8.637923e-01	9.351146e-01
HISTOLOGICAL_DIAGNOSIS_Infiltrating Lobular Car...	0.823077	3.912033e-01	5.714063e-01
ICD_O_3_HISTOLOGY_8500/3	0.949261	7.692879e-01	8.563771e-01
ICD_O_3_HISTOLOGY_8520/3	0.871304	5.370519e-01	6.741716e-01
METASTATIC_TUMOR_INDICATOR_NO	0.641384	2.830420e-02	9.823221e-02
METASTATIC_TUMOR_INDICATOR_[Not Available]	1.149966	4.502190e-01	6.047185e-01

TISSUE_SOURCE_SITE_A2	0.816065	4.561012e-01	6.047185e-01
TISSUE_SOURCE_SITE_A8	1.386201	4.396658e-01	6.047185e-01
TISSUE_SOURCE_SITE_AR	0.426279	7.036257e-03	3.609906e-02
TISSUE_SOURCE_SITE_B6	1.053316	8.202897e-01	9.046185e-01
TISSUE_SOURCE_SITE_BH	2.772661	1.247824e-08	4.908109e-07
TISSUE_SOURCE_SITE_D8	0.947428	9.275108e-01	9.860025e-01
TISSUE_SOURCE_SITE_E2	0.555754	1.612908e-01	3.021003e-01
TISSUE_SOURCE_SITE_E9	1.350339	4.798622e-01	6.279592e-01
AJCC_TUMOR_PATHOLOGIC_PT_simple_T1	0.686002	5.620924e-02	1.411211e-01
AJCC_TUMOR_PATHOLOGIC_PT_simple_T2	0.951778	7.620026e-01	8.563458e-01
AJCC_TUMOR_PATHOLOGIC_PT_simple_T3	1.220149	3.652398e-01	5.525423e-01
AJCC_NODES_PATHOLOGIC_PN_simple_N0	0.428539	2.290188e-06	3.502342e-05
AJCC_NODES_PATHOLOGIC_PN_simple_N1	1.159871	3.741624e-01	5.588754e-01
AJCC_NODES_PATHOLOGIC_PN_simple_N2	1.714021	1.986213e-02	7.812439e-02
AJCC_NODES_PATHOLOGIC_PN_simple_N3	2.564601	6.195610e-04	4.873880e-03
AJCC_METASTASIS_PATHOLOGIC_PM_simple_M0	0.509375	8.980883e-04	5.887468e-03
AJCC_METASTASIS_PATHOLOGIC_PM_simple_MX	1.062431	8.307274e-01	9.076466e-01

Table S3: Clinical variables significance for single-variable cox model for BRCA (1080 patients). Variable significantly associated with survival are shaded.

Variable	Hazard ratio	P-value	Corrected P-value
INITIAL_PATHOLOGIC_DX_YEAR	0.944473	0.000002	0.000085
age_at_diagnosis	1.019742	0.001594	0.012355
PRIMARY_SITE_Floor of mouth	1.633189	0.009866	0.048599
PRIMARY_SITE_Larynx	0.870495	0.392949	0.589424
PRIMARY_SITE_Oral Cavity	1.216594	0.257959	0.436185
PRIMARY_SITE_Oral Tongue	1.027363	0.864976	0.935381
LATERALITY_Left	0.997223	0.985683	0.985683
LATERALITY_Right	0.596169	0.001229	0.010391
LATERALITY_[Not Available]	1.487693	0.003515	0.021791
PROSPECTIVE_COLLECTION_NO	1.275968	0.136430	0.309463
PROSPECTIVE_COLLECTION_YES	0.796014	0.163125	0.337124
RETROSPECTIVE_COLLECTION_NO	0.796014	0.163125	0.337124
RETROSPECTIVE_COLLECTION_YES	1.275968	0.136430	0.309463
SEX_Female	1.382324	0.024394	0.087255
SEX_Male	0.723420	0.024394	0.087255
RACE_WHITE	0.742605	0.111804	0.273433
ETHNICITY_NOT HISPANIC OR LATINO	0.853619	0.414453	0.611812
LYMPH_NODE_NECK_DISSECTION_INDICATOR_NO	1.236683	0.203896	0.403454
LYMPH_NODE_NECK_DISSECTION_INDICATOR_YES	0.814286	0.215391	0.412907
LYMPH_NODE_DISSECTION_METHOD_Functional (Limite...	1.086460	0.540168	0.749786
LYMPH_NODE_DISSECTION_METHOD_Modified Radical N...	0.741422	0.093637	0.241895
LYMPH_NODE_DISSECTION_METHOD_[Not Available]	1.090921	0.583249	0.763974
LYMPH_NODES_EXAMINED_NO	1.470161	0.061855	0.179766
LYMPH_NODES_EXAMINED_YES	0.837805	0.290092	0.481761
PATH_MARGIN_Negative	0.777605	0.079280	0.216854
PATH_MARGIN_Positive	1.727697	0.002205	0.014645
PATH_MARGIN_[Not Available]	0.699698	0.158253	0.337124
AJCC_STAGING_EDITION_6th	1.010909	0.944167	0.955798
AJCC_STAGING_EDITION_7th	0.973421	0.855035	0.935381
AJCC_PATHOLOGIC_TUMOR_STAGE_Stage II	0.668524	0.054444	0.168777
AJCC_PATHOLOGIC_TUMOR_STAGE_Stage III	0.726051	0.114665	0.273433
AJCC_PATHOLOGIC_TUMOR_STAGE_Stage IVA	1.532143	0.001807	0.012924
AJCC_PATHOLOGIC_TUMOR_STAGE_[Not Available]	0.962185	0.849093	0.935381
EXTRACAPSULAR_SPREAD_PATHOLOGIC_Microscopic Ext...	2.254494	0.000001	0.000085
EXTRACAPSULAR_SPREAD_PATHOLOGIC_No Extranodal E...	0.548915	0.000017	0.000392
EXTRACAPSULAR_SPREAD_PATHOLOGIC_[Not Available]	0.990168	0.945521	0.955798
GRADE_G1	0.657592	0.059833	0.179498
GRADE_G2	1.431246	0.010999	0.050448
GRADE_G3	0.957962	0.783689	0.915850
ANGIOLYMPHATIC_INVASION_NO	0.693371	0.009929	0.048599
ANGIOLYMPHATIC_INVASION_YES	1.484568	0.008174	0.047509
ANGIOLYMPHATIC_INVASION_[Not Available]	1.057948	0.692475	0.847371
PERINEURAL_INVASION_NO	0.516015	0.000030	0.000565
PERINEURAL_INVASION_YES	1.588730	0.000670	0.006230
PERINEURAL_INVASION_[Not Available]	1.147166	0.339620	0.535334
HPV_STATUS_P16_Negative	0.856744	0.477818	0.673289
HPV_STATUS_P16_[Not Available]	1.469555	0.052450	0.168202
HPV_STATUS_ISH_Negative	0.842503	0.448772	0.642089
HPV_STATUS_ISH_[Not Available]	1.564703	0.043660	0.145013
TOBACCO_SMOKING_HISTORY_INDICATOR_1	0.853226	0.347479	0.538592

TOBACCO_SMOKING_HISTORY_INDICATOR_2	1.349280	0.033815	0.116475
TOBACCO_SMOKING_HISTORY_INDICATOR_3	0.708845	0.088463	0.235059
TOBACCO_SMOKING_HISTORY_INDICATOR_4	0.949831	0.733565	0.874635
ALCOHOL_HISTORY_DOCUMENTED_NO	1.064729	0.660954	0.841584
ALCOHOL_HISTORY_DOCUMENTED_YES	0.941649	0.669647	0.841584
RADIATION_TREATMENT_ADJUVANT_NO	1.062557	0.787828	0.915850
RADIATION_TREATMENT_ADJUVANT_YES	0.905136	0.549419	0.751412
RADIATION_TREATMENT_ADJUVANT_[Not Available]	1.055628	0.712535	0.860594
PHARMACEUTICAL_TX_ADJUVANT_NO	0.774111	0.144766	0.320553
PHARMACEUTICAL_TX_ADJUVANT_YES	1.454739	0.067970	0.191551
PHARMACEUTICAL_TX_ADJUVANT_[Not Available]	1.032903	0.825634	0.935381
CLIN_N_STAGE_N0	0.853501	0.240473	0.436185
CLIN_N_STAGE_N1	0.915126	0.621348	0.802575
CLIN_N_STAGE_N2b	0.980360	0.916449	0.955798
CLIN_T_STAGE_T2	0.838379	0.251047	0.436185
CLIN_T_STAGE_T3	1.269859	0.104375	0.262348
CLIN_T_STAGE_T4a	1.014869	0.918479	0.955798
CLINICAL_STAGE_Stage II	0.872315	0.435405	0.632698
CLINICAL_STAGE_Stage III	0.978740	0.898501	0.955798
CLINICAL_STAGE_Stage IVA	1.138939	0.338664	0.535334
ICD_10_C02.9	0.970829	0.852993	0.935381
ICD_10_C04.9	2.068995	0.000137	0.001823
ICD_10_C14.8	1.218218	0.254981	0.436185
ICD_10_C32.9	0.914839	0.578172	0.763974
ICD_O_3_HISTOLOGY_8070/3	1.344115	0.221993	0.412907
ICD_O_3_HISTOLOGY_8071/3	0.898413	0.682330	0.846089
ICD_O_3_SITE_C02.9	0.970829	0.852993	0.935381
ICD_O_3_SITE_C04.9	2.068995	0.000137	0.001823
ICD_O_3_SITE_C14.8	1.218218	0.254981	0.436185
ICD_O_3_SITE_C32.9	0.914839	0.578172	0.763974
TISSUE_SOURCE_SITE_CN	1.213414	0.297016	0.484604
TISSUE_SOURCE_SITE_CR	0.431810	0.009643	0.048599
TISSUE_SOURCE_SITE_CV	1.663811	0.000385	0.003978
AJCC_TUMOR_PATHOLOGIC_PT_simple_T2	0.659617	0.011391	0.050448
AJCC_TUMOR_PATHOLOGIC_PT_simple_T3	1.459494	0.016649	0.070379
AJCC_TUMOR_PATHOLOGIC_PT_simple_T4	1.388501	0.017779	0.071889
AJCC_NODES_PATHOLOGIC_PN_simple_N0	0.556349	0.000185	0.002146
AJCC_NODES_PATHOLOGIC_PN_simple_N1	0.576861	0.022426	0.086899
AJCC_NODES_PATHOLOGIC_PN_simple_N2	1.834907	0.000009	0.000292
AJCC_NODES_PATHOLOGIC_PN_simple_NX	1.165686	0.381317	0.581352
AJCC_METASTASIS_PATHOLOGIC_PM_simple_M0	0.818485	0.176511	0.356858
AJCC_METASTASIS_PATHOLOGIC_PM_simple_MX	0.979428	0.928250	0.955798
AJCC_METASTASIS_PATHOLOGIC_PM_simple_[N	1.185902	0.220529	0.412907

Table S4: Clinical variables significance for single-variable cox model for HNSC (526 patients). Variable significantly associated with survival are shaded.

CellType	Symbol	AffymetrixID	EntrezGene	Gene Ensembl ID
B cells	MS4A1	217418_x_at	931	ENSG00000156738
B cells	TCL1A	209995_s_at	8115	ENSG00000100721
B cells	MS4A1	210356_x_at	931	ENSG00000156738
B cells	TCL1A	39318_at	8115	ENSG00000100721
B cells	HLA-DOB	205671_s_at	3112	ENSG00000241106
B cells	HLA-DOB	205671_s_at	3112	ENSG00000239457
B cells	HLA-DOB	205671_s_at	3112	ENSG00000243496
B cells	HLA-DOB	205671_s_at	3112	ENSG00000241386
B cells	HLA-DOB	205671_s_at	3112	ENSG00000241910
B cells	HLA-DOB	205671_s_at	3112	ENSG00000243612
B cells	PNOC	205901_at	5368	ENSG00000168081
B cells	KIAA0125	206478_at	9834	ENSG00000277059
B cells	KIAA0125	206478_at	9834	ENSG00000226777
B cells	CD19	206398_s_at	930	ENSG00000177455
B cells	CR2	205544_s_at	1380	ENSG00000117322
B cells	IGHG1	213674_x_at	3500	NaN
B cells	FCRL2	221239_s_at	79368	ENSG00000132704
B cells	BLK	206255_at	640	ENSG00000136573
B cells	IGHG1	222285_at	3500	NaN
B cells	COCH	205229_s_at	1690	ENSG00000100473
B cells	OSBPL10	219073_s_at	114884	ENSG00000144645
B cells	IGHA1	215118_s_at	3493	NaN
B cells	TNFRSF17	206641_at	608	ENSG00000048462
B cells	ABCB4	207819_s_at	5244	ENSG00000005471
B cells	BLNK	207655_s_at	29760	ENSG00000095585
B cells	GLDC	204836_at	2731	ENSG00000178445
B cells	MEF2C	209200_at	4208	ENSG00000081189
B cells	MEF2C	209199_s_at	4208	ENSG00000081189
B cells	IGHM	209374_s_at	3507	NaN
B cells	FAM30A	220377_at	29064	NaN
B cells	SPIB	205861_at	6689	ENSG00000269404
B cells	BCL11A	219497_s_at	53335	ENSG00000119866
B cells	GNG7	206896_s_at	2788	ENSG00000176533
B cells	IGKC	215217_at	3514	NaN
B cells	CD72	215925_s_at	971	ENSG00000137101
B cells	MICAL3	212715_s_at	57553	ENSG00000243156
B cells	BCL11A	210347_s_at	53335	ENSG00000119866
B cells	BACH2	221234_s_at	60468	ENSG00000112182
B cells	IGL@	217138_x_at	3535	NaN
B cells	CCR9	207445_s_at	10803	ENSG00000173585
B cells	QRSL1	218948_at	55278	ENSG00000130348
B cells	DTNB	215565_at	1838	ENSG00000138101
B cells	HLA-DQA1	212671_s_at	3117	ENSG00000236418
B cells	HLA-DQA1	212671_s_at	3117	ENSG00000232062
B cells	HLA-DQA1	212671_s_at	3117	ENSG00000223793
B cells	HLA-DQA1	212671_s_at	3117	ENSG00000231526
B cells	HLA-DQA1	212671_s_at	3117	ENSG00000257473
B cells	HLA-DQA1	212671_s_at	3117	ENSG00000228284
B cells	HLA-DQA1	212671_s_at	3117	ENSG00000225890
B cells	HLA-DQA1	212671_s_at	3117	ENSG00000231823
B cells	HLA-DQA1	212671_s_at	3117	ENSG00000206301
B cells	HLA-DQA1	212671_s_at	3117	ENSG00000206305

B cells	HLA-DQA1	212671_s_at	3117	ENSG00000225103
B cells	HLA-DQA1	212671_s_at	3117	ENSG00000233192
B cells	HLA-DQA1	212671_s_at	3117	ENSG00000196735
B cells	HLA-DQA1	212671_s_at	3117	ENSG00000237541
B cells	SCN3A	210432_s_at	6328	ENSG00000153253
B cells	QRSL1	218949_s_at	55278	ENSG00000130348
B cells	SLC15A2	205316_at	6565	ENSG00000163406
T cells	PRKCQ	210038_at	5588	ENSG00000065675
T cells	CD3D	213539_at	915	ENSG00000167286
T cells	CD3G	206804_at	917	ENSG00000160654
T cells	CD28	206545_at	940	ENSG00000178562
T cells	LCK	204891_s_at	3932	ENSG00000182866
T cells	TRAT1	217147_s_at	50852	ENSG00000163519
T cells	PRKCQ	210039_s_at	5588	ENSG00000065675
T cells	BCL11B	219528_s_at	64919	ENSG00000127152
T cells	CD2	205831_at	914	ENSG00000116824
T cells	LCK	204890_s_at	3932	ENSG00000182866
T cells	TRBC1	213193_x_at	28639	NaN
T cells	TRBC1	210915_x_at	28639	NaN
T cells	TRA@	209670_at	28755	NaN
T cells	ITM2A	202747_s_at	9452	ENSG00000078596
T cells	SH2D1A	210116_at	4068	ENSG00000183918
T cells	CD6	213958_at	923	ENSG00000013725
T cells	CD96	206761_at	10225	ENSG00000153283
T cells	NCALD	211685_s_at	83988	ENSG00000104490
T cells	GIMAP5	218805_at	55340	ENSG00000196329
T cells	TRA@	209671_x_at	6955	NaN
T cells	CD3E	205456_at	916	ENSG00000198851
T cells	SKAP1	205790_at	8631	ENSG00000141293
T cells	TRA@	213830_at	6955	NaN
T cells	TRA@	216191_s_at	6955	NaN
T helper cells	ICOS	210439_at	29851	ENSG00000163600
T helper cells	LRBA	214109_at	987	ENSG00000198589
T helper cells	ITM2A	202746_at	9452	ENSG00000078596
T helper cells	FAM111A	218248_at	63901	ENSG00000166801
T helper cells	PHF10	219126_at	55274	ENSG00000130024
T helper cells	NUP107	218768_at	57122	ENSG00000111581
T helper cells	SEC24C	202361_at	9632	ENSG00000176986
T helper cells	NAP1L4	201414_s_at	4676	ENSG00000205531
T helper cells	NAP1L4	201414_s_at	4676	ENSG00000273562
T helper cells	BATF	205965_at	10538	ENSG00000156127
T helper cells	ASF1A	203428_s_at	25842	ENSG00000111875
T helper cells	FRYL	212546_s_at	285527	ENSG00000075539
T helper cells	FUSIP1	213594_x_at	10772	ENSG00000188529
T helper cells	TRA@	215524_x_at	10730	ENSG00000136758
T helper cells	TRA@	217412_at	6955	NaN
T helper cells	RPA1	201528_at	6117	ENSG00000132383
T helper cells	UBE2L3	200683_s_at	7332	ENSG00000185651
T helper cells	ANP32B	201306_s_at	10541	ENSG00000136938
T helper cells	DDX50	221699_s_at	79009	ENSG00000107625
T helper cells	C13orf34	219544_at	79866	ENSG00000136122
T helper cells	PPP2R5C	213305_s_at	5527	ENSG00000078304
T helper cells	SLC25A12	203340_s_at	8604	ENSG00000115840
T helper cells	ATF2	205446_s_at	1386	ENSG00000115966

T helper cells	CD28	211856_x_at	940	ENSG00000178562
T helper cells	GOLGA8A	208798_x_at	23015	ENSG00000175265
Tcm	CDC14A	210441_at	8556	ENSG00000079335
Tcm	ATM	208442_s_at	472	ENSG00000149311
Tcm	USP9Y	206624_at	8287	ENSG00000114374
Tcm	PCNX	215175_at	22990	ENSG00000100731
Tcm	ATM	210858_x_at	472	ENSG00000149311
Tcm	FOXP1	215221_at	27086	ENSG00000114861
Tcm	KLF12	206965_at	11278	ENSG00000118922
Tcm	ST3GAL1	215874_at	6482	ENSG00000008513
Tcm	INPP4B	215864_at	8821	ENSG00000109452
Tcm	CASP8	207686_s_at	841	ENSG00000064012
Tcm	MLL	216624_s_at	4297	ENSG00000118058
Tcm	PCM1	209997_x_at	5108	ENSG00000078674
Tcm	RP11-74E24.2	205787_x_at	441155	ENSG00000215817
Tcm	PHC3	215521_at	80012	ENSG00000173889
Tcm	NFATC3	210556_at	4775	ENSG00000072736
Tcm	LOC202134	215133_s_at	202134	ENSG00000182230
Tcm	TIMM8A	210800_at	1678	ENSG00000126953
Tcm	ATF7IP	216197_at	55729	ENSG00000171681
Tcm	REPS1	215201_at	85021	ENSG00000135597
Tcm	PSPC1	215083_at	55269	ENSG00000121390
Tcm	RPP38	215743_at	9397	ENSG00000152465
Tcm	HNRPH1	213472_at	3187	ENSG00000169045
Tcm	STX16	221638_s_at	8675	ENSG00000124222
Tcm	CYLD	214272_at	1540	ENSG00000083799
Tcm	SNRPN	216850_at	6638	ENSG00000128739
Tcm	TRAF3IP3	215275_at	80342	ENSG00000009790
Tcm	NEFL	221805_at	4747	ENSG00000277586
Tcm	POLR2J2	217610_at	246721	ENSG00000228049
Tcm	AQP3	203747_at	360	ENSG00000165272
Tcm	CG030	215105_at	116828	ENSG00000281026
Tcm	PDXDC2	215920_s_at	283970	ENSG00000196696
Tcm	CLUAP1	204576_s_at	23059	ENSG00000103351
Tcm	DOCK9	215041_s_at	23348	ENSG00000088387
Tcm	CYorf15B	214131_at	84663	NaN
Tcm	CREBZF	213584_s_at	58487	ENSG00000137504
Tcm	CEP68	207971_s_at	23177	ENSG00000011523
Tcm	TXK	206828_at	7294	ENSG00000074966
Tcm	SLC7A6	203578_s_at	9057	ENSG00000103064
Tcm	FYB	205285_s_at	2533	ENSG00000082074
Tcm	MAP3K1	214786_at	4214	ENSG00000095015
Tem	TRA@	217397_at	6955	NaN
Tem	PRKY	206279_at	5616	ENSG00000099725
Tem	VIL2	217230_at	7430	ENSG00000092820
Tem	GDPD5	32502_at	81544	ENSG00000158555
Tem	CCR2	206978_at	1231	NaN
Tem	MEFV	208262_x_at	4210	ENSG00000103313
Tem	C7orf54	210109_at	27099	ENSG00000279078
Tem	FLI1	210786_s_at	2313	ENSG00000151702
Tem	TBC1D5	201815_s_at	9779	ENSG00000131374
Tem	DDX17	208719_s_at	10521	ENSG00000100201
Tem	AKT3	212609_s_at	10000	ENSG00000275199
Tem	AKT3	212609_s_at	10000	ENSG00000117020

Tem	EWSR1	211825_s_at	2130	ENSG00000182944
Tem	TBCD	201759_at	6904	ENSG00000278759
Tem	TBCD	201759_at	6904	ENSG00000141556
Tem	CCR2	207794_at	1231	NaN
Tem	NFATC4	205897_at	4776	ENSG00000100968
Tem	LTK	207106_s_at	4058	ENSG00000062524
Th1 cells	IFNG	210354_at	3458	ENSG00000111537
Th1 cells	LTA	206975_at	4049	ENSG00000238130
Th1 cells	LTA	206975_at	4049	ENSG00000173503
Th1 cells	LTA	206975_at	4049	ENSG00000231408
Th1 cells	LTA	206975_at	4049	ENSG00000226979
Th1 cells	LTA	206975_at	4049	ENSG00000230279
Th1 cells	LTA	206975_at	4049	ENSG00000226275
Th1 cells	LTA	206975_at	4049	ENSG00000223919
Th1 cells	APBB2	213419_at	323	ENSG00000163697
Th1 cells	DOK5	214844_s_at	55816	ENSG00000101134
Th1 cells	IL12RB2	206999_at	3595	ENSG00000081985
Th1 cells	APBB2	40148_at	323	ENSG00000163697
Th1 cells	APOD	201525_at	347	ENSG00000189058
Th1 cells	ZBTB32	220118_at	27033	ENSG00000011590
Th1 cells	CD38	205692_s_at	952	ENSG00000004468
Th1 cells	CSF2	210229_s_at	1437	ENSG00000164400
Th1 cells	CTLA4	221331_x_at	1493	ENSG00000163599
Th1 cells	CD70	206508_at	970	ENSG00000125726
Th1 cells	DPP4	211478_s_at	1803	ENSG00000197635
Th1 cells	EGFL6	219454_at	25975	ENSG00000198759
Th1 cells	BST2	201641_at	684	ENSG00000130303
Th1 cells	DUSP5	209457_at	1847	ENSG00000138166
Th1 cells	LRP8	205282_at	7804	ENSG00000157193
Th1 cells	IL22	221165_s_at	50616	ENSG00000127318
Th1 cells	DGKI	206806_at	9162	ENSG00000157680
Th1 cells	CCL4	204103_at	6351	ENSG00000275824
Th1 cells	CCL4	204103_at	6351	ENSG00000275302
Th1 cells	CCL4	204103_at	6351	ENSG00000277943
Th1 cells	DPP4	203716_s_at	1803	ENSG00000197635
Th1 cells	GGT1	211417_x_at	2678	ENSG00000100031
Th1 cells	LRRN3	209840_s_at	54674	ENSG00000173114
Th1 cells	SYNGR3	205691_at	9143	ENSG00000127561
Th1 cells	ATP9A	212062_at	10079	ENSG00000054793
Th1 cells	BTG3	205548_s_at	10950	ENSG00000281484
Th1 cells	BTG3	205548_s_at	10950	ENSG00000154640
Th1 cells	CMAH	210571_s_at	8418	NaN
Th1 cells	HBEGF	38037_at	1839	ENSG00000113070
Th1 cells	SGCB	205120_s_at	6443	ENSG00000163069
Th2 cells	PMCH	206942_s_at	5367	ENSG00000183395
Th2 cells	AHI1	220841_s_at	54806	ENSG00000135541
Th2 cells	PTGIS	208131_s_at	5740	ENSG00000124212
Th2 cells	AHI1	220842_at	54806	ENSG00000135541
Th2 cells	CXCR6	211469_s_at	10663	ENSG00000172215
Th2 cells	EVI5	209717_at	7813	ENSG00000067208
Th2 cells	AHI1	221569_at	54806	ENSG00000135541
Th2 cells	IL26	221111_at	55801	ENSG00000111536
Th2 cells	MB	204179_at	4151	ENSG00000198125
Th2 cells	NEIL3	219502_at	55247	ENSG00000109674

Th2 cells	GSTA4	202967_at	2941	ENSG00000170899
Th2 cells	PHEX	210617_at	5251	ENSG00000102174
Th2 cells	SMAD2	203076_s_at	4087	ENSG00000175387
Th2 cells	CENPF	209172_s_at	1063	ENSG00000117724
Th2 cells	ANK1	208353_x_at	286	ENSG00000029534
Th2 cells	ADCY1	213245_at	107	ENSG00000164742
Th2 cells	AI582773	214373_at	728210	NaN
Th2 cells	LAIR2	207509_s_at	3904	ENSG00000277335
Th2 cells	LAIR2	207509_s_at	3904	ENSG00000274084
Th2 cells	LAIR2	207509_s_at	3904	ENSG00000167618
Th2 cells	LAIR2	207509_s_at	3904	ENSG00000275819
Th2 cells	SNRPD1	202691_at	6632	ENSG00000167088
Th2 cells	CXCR6	206974_at	10663	ENSG00000172215
Th2 cells	MICAL2	212472_at	9645	ENSG00000133816
Th2 cells	DHFR	202534_x_at	1719	ENSG00000228716
Th2 cells	SMAD2	203077_s_at	4087	ENSG00000175387
Th2 cells	WDHD1	204728_s_at	11169	ENSG00000198554
Th2 cells	BIRC5	210334_x_at	332	ENSG00000089685
Th2 cells	DHFR	48808_at	1719	ENSG00000228716
Th2 cells	SLC39A14	212110_at	23516	ENSG00000104635
Th2 cells	HELLS	220085_at	3070	ENSG00000119969
Th2 cells	LIMA1	217892_s_at	51474	ENSG00000050405
Th2 cells	CDC25C	205167_s_at	995	ENSG00000158402
Th2 cells	CDC7	204510_at	8317	ENSG00000097046
Th2 cells	GATA3	209602_s_at	2625	ENSG00000107485
TFH	CHI3L2	213060_s_at	1117	ENSG00000064886
TFH	CXCL13	205242_at	10563	ENSG00000156234
TFH	MYO7A	33197_at	4647	ENSG00000137474
TFH	CHGB	204260_at	1114	ENSG00000089199
TFH	MYO7A	208189_s_at	4647	ENSG00000137474
TFH	ICA1	210547_x_at	3382	ENSG00000003147
TFH	HEY1	218839_at	23462	ENSG00000164683
TFH	CDK5R1	204995_at	8851	ENSG00000176749
TFH	ST8SIA1	210073_at	6489	ENSG00000111728
TFH	PDCD1	207634_at	5133	ENSG00000276977
TFH	PDCD1	207634_at	5133	ENSG00000188389
TFH	BLR1	216734_s_at	643	ENSG00000160683
TFH	KIAA1324	221874_at	57535	ENSG00000116299
TFH	PVALB	205336_at	5816	ENSG00000100362
TFH	PVALB	205336_at	5816	ENSG00000274665
TFH	ICA1	207949_s_at	3382	ENSG00000003147
TFH	TSHR	210055_at	7253	ENSG00000165409
TFH	C18orf1	209574_s_at	753	ENSG00000168675
TFH	HEY1	44783_s_at	23462	ENSG00000164683
TFH	TOX	204529_s_at	9760	ENSG00000198846
TFH	BLR1	206126_at	643	ENSG00000160683
TFH	SLC7A10	220868_s_at	56301	ENSG00000130876
TFH	SMAD1	210993_s_at	4086	ENSG00000170365
TFH	POMT1	218476_at	10585	ENSG00000130714
TFH	PASK	216945_x_at	23178	ENSG00000115687
TFH	MKL2	218259_at	57496	ENSG00000186260
TFH	PTPN13	204201_s_at	5783	ENSG00000163629
TFH	PASK	213534_s_at	23178	ENSG00000115687
TFH	KCNK5	219615_s_at	8645	ENSG00000164626

TFH	C18orf1	207996_s_at	753	ENSG00000168675
TFH	ZNF764	57516_at	92595	ENSG00000169951
TFH	MAF	206363_at	4094	ENSG00000178573
TFH	MYO6	210480_s_at	4646	ENSG00000196586
TFH	SIRPG	220485_s_at	55423	ENSG00000089012
TFH	THADA	54632_at	63892	ENSG00000115970
TFH	THADA	220212_s_at	63892	ENSG00000115970
TFH	MAGEH1	218573_at	28986	ENSG00000187601
TFH	B3GAT1	219521_at	27087	ENSG00000109956
TFH	MAF	209348_s_at	4094	ENSG00000178573
TFH	SH3TC1	219256_s_at	54436	ENSG00000125089
TFH	HIST1H4K	214463_x_at	8362	ENSG00000273542
TFH	STK39	202786_at	27347	ENSG00000198648
Th17 cells	IL17A	208402_at	3605	ENSG00000112115
Th17 cells	IL17A	216876_s_at	3605	ENSG00000112115
Th17 cells	IL17RA	205707_at	23765	ENSG00000177663
Th17 cells	RORC	206419_at	6097	ENSG00000143365
TReg	FOXP3	221333_at	50943	ENSG00000049768
TReg	FOXP3	221334_s_at	50943	ENSG00000049768
CD8 T cells	CD8B	207979_s_at	926	ENSG00000172116
CD8 T cells	CD8A	205758_at	925	ENSG00000153563
CD8 T cells	CD8B	215332_s_at	926	ENSG00000172116
CD8 T cells	PF4	206390_x_at	5196	ENSG00000163737
CD8 T cells	PRR5	47069_at	55615	ENSG00000186654
CD8 T cells	SF1	210172_at	7536	ENSG00000168066
CD8 T cells	LIME1	219541_at	54923	ENSG00000203896
CD8 T cells	DNAJB1	200664_s_at	3337	ENSG00000132002
CD8 T cells	ARHGAP8	219168_s_at	55615	ENSG00000186654
CD8 T cells	GZMM	207460_at	3004	ENSG00000197540
CD8 T cells	SLC16A7	207057_at	9194	ENSG00000118596
CD8 T cells	SFRS7	213649_at	6432	ENSG00000115875
CD8 T cells	APBA2	209871_s_at	321	ENSG00000034053
CD8 T cells	APBA2	209871_s_at	321	ENSG00000276495
CD8 T cells	C4orf15	210054_at	79441	ENSG00000214367
CD8 T cells	LEPROTL1	202595_s_at	23484	ENSG00000104660
CD8 T cells	ZFP36L2	201367_s_at	678	ENSG00000152518
CD8 T cells	GADD45A	203725_at	1647	ENSG00000116717
CD8 T cells	ZFP36L2	201369_s_at	678	ENSG00000152518
CD8 T cells	MYST3	216361_s_at	7994	ENSG00000083168
CD8 T cells	ZEB1	208078_s_at	6935	ENSG00000148516
CD8 T cells	ZNF609	212620_at	23060	ENSG00000180357
CD8 T cells	C12orf47	64432_at	51275	ENSG00000234608
CD8 T cells	THUMPD1	206555_s_at	55623	ENSG00000066654
CD8 T cells	VAMP2	201557_at	6844	ENSG00000220205
CD8 T cells	ZNF91	206059_at	7644	ENSG00000167232
CD8 T cells	ZNF22	218006_s_at	7570	ENSG00000165512
CD8 T cells	TMC6	214958_s_at	11322	ENSG00000141524
CD8 T cells	DNAJB1	200666_s_at	3337	ENSG00000132002
CD8 T cells	FLT3LG	210607_at	2323	ENSG00000090554
CD8 T cells	CDKN2AIP	218929_at	55602	ENSG00000168564
CD8 T cells	TSC22D3	207001_x_at	1831	ENSG00000157514
CD8 T cells	TBCC	202495_at	6903	ENSG00000124659
CD8 T cells	RBM3	208319_s_at	5935	ENSG00000102317
CD8 T cells	ABT1	218405_at	29777	ENSG00000146109

CD8 T cells	C19orf6	212574_x_at	91304	ENSG00000182087
CD8 T cells	CAMLG	203538_at	819	ENSG00000164615
CD8 T cells	PPP1R2	202165_at	5504	ENSG00000184203
CD8 T cells	AES	217729_s_at	166	ENSG00000104964
CD8 T cells	KLF9	203543_s_at	687	ENSG00000119138
CD8 T cells	PRF1	214617_at	5551	ENSG00000180644
Tgd	TRD@	217143_s_at	6964	NaN
Tgd	TARP	211144_x_at	445347	ENSG00000211689
Tgd	C1orf61	205103_at	10485	ENSG00000125462
Tgd	TRGV9	209813_x_at	6983	NaN
Tgd	CD160	207840_at	11126	ENSG00000117281
Tgd	TARP	216920_s_at	445347	ENSG00000211689
Tgd	FEZ1	203562_at	9638	ENSG00000149557
Cytotoxic cells	KLRD1	210606_x_at	3824	ENSG00000134539
Cytotoxic cells	KLRF1	220646_s_at	51348	ENSG00000150045
Cytotoxic cells	GNLY	37145_at	10578	ENSG00000115523
Cytotoxic cells	GNLY	205495_s_at	10578	ENSG00000115523
Cytotoxic cells	CTSW	214450_at	1521	ENSG00000172543
Cytotoxic cells	KLRB1	214470_at	3820	ENSG00000111796
Cytotoxic cells	KLRD1	207795_s_at	3824	ENSG00000134539
Cytotoxic cells	KLRK1	205821_at	22914	ENSG00000213809
Cytotoxic cells	NKG7	213915_at	4818	ENSG00000105374
Cytotoxic cells	GZMH	210321_at	2999	ENSG00000100450
Cytotoxic cells	KLRD1	207796_x_at	3824	ENSG00000134539
Cytotoxic cells	SIGIRR	218921_at	59307	ENSG00000185187
Cytotoxic cells	ZBTB16	205883_at	7704	ENSG00000109906
Cytotoxic cells	RUNX3	204198_s_at	864	ENSG00000020633
Cytotoxic cells	APOL3	221087_s_at	80833	ENSG00000128284
Cytotoxic cells	RORA	210426_x_at	6095	ENSG00000069667
Cytotoxic cells	APBA2	209870_s_at	321	ENSG00000034053
Cytotoxic cells	APBA2	209870_s_at	321	ENSG00000276495
Cytotoxic cells	SIGIRR	52940_at	59307	ENSG00000185187
Cytotoxic cells	WHDC1L1	213908_at	339005	NaN
Cytotoxic cells	DUSP2	204794_at	1844	ENSG00000158050
Cytotoxic cells	GZMA	205488_at	3001	ENSG00000145649
NK cells	LOC643313	211050_x_at	643313	NaN
NK cells	GAGE2	207739_s_at	2574	ENSG00000236362
NK cells	ZNF747	206180_x_at	65988	ENSG00000169955
NK cells	XCL1	206366_x_at	6375	ENSG00000143184
NK cells	XCL2	214567_s_at	6846	ENSG00000143185
NK cells	AF107846	217058_at	2778	ENSG00000087460
NK cells	SLC30A5	220181_x_at	64924	ENSG00000145740
NK cells	NM_014114	220691_at	259230	ENSG00000198964
NK cells	MCM3AP	215582_x_at	8888	ENSG00000160294
NK cells	TBXA2R	207555_s_at	6915	ENSG00000006638
NK cells	CDC5L	209057_x_at	988	ENSG00000096401
NK cells	LOC730096	215182_x_at	730096	NaN
NK cells	FUT5	211225_at	2527	ENSG00000130383
NK cells	FGF18	206986_at	8817	ENSG00000156427
NK cells	MRC2	209280_at	9902	ENSG00000011028
NK cells	RP5-886K2.1	208014_x_at	27308	NaN
NK cells	SPN	216981_x_at	6693	ENSG00000197471
NK cells	PSMD4	210459_at	5710	ENSG00000159352
NK cells	PRX	220024_s_at	57716	ENSG00000105227

NK cells	FZR1	209415_at	51343	ENSG00000105325
NK cells	ZNF205	206416_at	7755	ENSG00000122386
NK cells	AL080130	212972_x_at	323	ENSG00000163697
NK cells	ZNF528	215019_x_at	84436	ENSG00000167555
NK cells	MAPRE3	203842_s_at	22924	ENSG00000084764
NK cells	BCL2	207004_at	596	ENSG00000171791
NK cells	NM_017616	221068_at	25959	ENSG00000197256
NK cells	ARL6IP2	217580_x_at	64225	ENSG00000119787
NK cells	SPN	206057_x_at	6693	ENSG00000197471
NK cells	FZR1	211865_s_at	51343	ENSG00000105325
NK cells	PDLIM4	214174_s_at	8572	ENSG00000131435
NK cells	NM_014274	206827_s_at	55503	ENSG00000165125
NK cells	NM_014274	206827_s_at	55503	ENSG00000276971
NK cells	LDB3	216888_at	11155	ENSG00000122367
NK cells	ADARB1	209979_at	104	ENSG00000197381
NK cells	SMEK1	215607_x_at	55671	ENSG00000100796
NK cells	TCTN2	206438_x_at	79867	ENSG00000168778
NK cells	TINAGL1	219058_x_at	64129	ENSG00000142910
NK cells	IGFBP5	203426_s_at	3488	ENSG00000115461
NK cells	ALDH1B1	209646_x_at	219	ENSG00000137124
NK cells	NCR1	217095_x_at	9437	ENSG00000278362
NK cells	NCR1	217095_x_at	9437	ENSG00000275156
NK cells	NCR1	217095_x_at	9437	ENSG00000277442
NK cells	NCR1	217095_x_at	9437	ENSG00000276450
NK cells	NCR1	217095_x_at	9437	ENSG00000278025
NK cells	NCR1	217095_x_at	9437	ENSG00000189430
NK cells	NCR1	217095_x_at	9437	ENSG00000275521
NK cells	NCR1	217095_x_at	9437	ENSG00000277334
NK cells	NCR1	217095_x_at	9437	ENSG00000275637
NK cells	NCR1	217095_x_at	9437	ENSG00000273535
NK cells	NCR1	217095_x_at	9437	ENSG00000277824
NK cells	NCR1	217095_x_at	9437	ENSG00000275822
NK cells	NCR1	217095_x_at	9437	ENSG00000274053
NK cells	NCR1	217095_x_at	9437	ENSG00000273506
NK cells	NCR1	217095_x_at	9437	ENSG00000277629
NK cells	NCR1	217095_x_at	9437	ENSG00000273916
NK cells	NCR1	217095_x_at	9437	ENSG00000284113
NK cells	NCR1	217095_x_at	9437	ENSG00000284208
NK cells	NCR1	217088_s_at	9437	ENSG00000278362
NK cells	NCR1	217088_s_at	9437	ENSG00000275156
NK cells	NCR1	217088_s_at	9437	ENSG00000277442
NK cells	NCR1	217088_s_at	9437	ENSG00000276450
NK cells	NCR1	217088_s_at	9437	ENSG00000278025
NK cells	NCR1	217088_s_at	9437	ENSG00000189430
NK cells	NCR1	217088_s_at	9437	ENSG00000275521
NK cells	NCR1	217088_s_at	9437	ENSG00000277334
NK cells	NCR1	217088_s_at	9437	ENSG00000275637
NK cells	NCR1	217088_s_at	9437	ENSG00000273535
NK cells	NCR1	217088_s_at	9437	ENSG00000277824
NK cells	NCR1	217088_s_at	9437	ENSG00000275822
NK cells	NCR1	217088_s_at	9437	ENSG00000274053
NK cells	NCR1	217088_s_at	9437	ENSG00000273506
NK cells	NCR1	217088_s_at	9437	ENSG00000277629
NK cells	NCR1	217088_s_at	9437	ENSG00000273916

NK cells	NCR1	217088_s_at	9437	ENSG00000284113
NK cells	NCR1	217088_s_at	9437	ENSG00000284208
NK cells	NCR1	207860_at	9437	ENSG00000278362
NK cells	NCR1	207860_at	9437	ENSG00000275156
NK cells	NCR1	207860_at	9437	ENSG00000277442
NK cells	NCR1	207860_at	9437	ENSG00000276450
NK cells	NCR1	207860_at	9437	ENSG00000278025
NK cells	NCR1	207860_at	9437	ENSG00000189430
NK cells	NCR1	207860_at	9437	ENSG00000275521
NK cells	NCR1	207860_at	9437	ENSG00000277334
NK cells	NCR1	207860_at	9437	ENSG00000275637
NK cells	NCR1	207860_at	9437	ENSG00000273535
NK cells	NCR1	207860_at	9437	ENSG00000277824
NK cells	NCR1	207860_at	9437	ENSG00000275822
NK cells	NCR1	207860_at	9437	ENSG00000274053
NK cells	NCR1	207860_at	9437	ENSG00000273506
NK cells	NCR1	207860_at	9437	ENSG00000277629
NK cells	NCR1	207860_at	9437	ENSG00000273916
NK cells	NCR1	207860_at	9437	ENSG00000284113
NK cells	NCR1	207860_at	9437	ENSG00000284208
NK CD56dim cells	KIR3DL2	207314_x_at	3812	ENSG00000278656
NK CD56dim cells	KIR3DL2	207314_x_at	3812	ENSG00000278361
NK CD56dim cells	KIR3DL2	207314_x_at	3812	ENSG00000277982
NK CD56dim cells	KIR3DL2	207314_x_at	3812	ENSG00000275626
NK CD56dim cells	KIR3DL2	207314_x_at	3812	ENSG00000276357
NK CD56dim cells	KIR3DL2	207314_x_at	3812	ENSG00000275511
NK CD56dim cells	KIR3DL2	207314_x_at	3812	ENSG00000273735
NK CD56dim cells	KIR3DL2	207314_x_at	3812	ENSG00000275262
NK CD56dim cells	KIR3DL2	207314_x_at	3812	ENSG00000278442
NK CD56dim cells	KIR3DL2	207314_x_at	3812	ENSG00000278707
NK CD56dim cells	KIR3DL2	207314_x_at	3812	ENSG00000278758
NK CD56dim cells	KIR3DL2	207314_x_at	3812	ENSG00000277181
NK CD56dim cells	KIR3DL2	207314_x_at	3812	ENSG00000275838
NK CD56dim cells	KIR3DL2	207314_x_at	3812	ENSG00000276739
NK CD56dim cells	KIR3DL2	207314_x_at	3812	ENSG00000277709
NK CD56dim cells	KIR3DL2	207314_x_at	3812	ENSG00000273911
NK CD56dim cells	KIR3DL2	207314_x_at	3812	ENSG00000276004
NK CD56dim cells	KIR3DL2	207314_x_at	3812	ENSG00000278403
NK CD56dim cells	KIR3DL2	207314_x_at	3812	ENSG00000275083
NK CD56dim cells	KIR3DL2	207314_x_at	3812	ENSG00000278809
NK CD56dim cells	KIR3DL2	207314_x_at	3812	ENSG00000278726
NK CD56dim cells	KIR3DL2	207314_x_at	3812	ENSG00000275629
NK CD56dim cells	KIR3DL2	207314_x_at	3812	ENSG00000276882
NK CD56dim cells	KIR3DL2	207314_x_at	3812	ENSG00000240403
NK CD56dim cells	KIR3DL2	207314_x_at	3812	ENSG00000278710
NK CD56dim cells	KIR3DL2	207314_x_at	3812	ENSG00000274722
NK CD56dim cells	KIR3DL2	207314_x_at	3812	ENSG00000275416
NK CD56dim cells	KIR3DL2	207314_x_at	3812	ENSG00000278474
NK CD56dim cells	KIR3DL2	207314_x_at	3812	ENSG00000275566
NK CD56dim cells	KIR3DL2	207314_x_at	3812	ENSG00000278850
NK CD56dim cells	KIR3DL2	207314_x_at	3812	ENSG00000276424
NK CD56dim cells	KIR3DL2	207314_x_at	3812	ENSG00000284295
NK CD56dim cells	KIR3DL2	207314_x_at	3812	ENSG00000284384
NK CD56dim cells	KIR3DL2	207314_x_at	3812	ENSG00000284466

NK CD56dim cells	KIR3DL2	207314_x_at	3812	ENSG00000284101
NK CD56dim cells	KIR3DL2	207314_x_at	3812	ENSG00000284213
NK CD56dim cells	KIR3DL2	207314_x_at	3812	ENSG00000284046
NK CD56dim cells	KIR3DL2	207314_x_at	3812	ENSG00000284063
NK CD56dim cells	KIR3DL2	207314_x_at	3812	ENSG00000283975
NK CD56dim cells	KIR3DL2	207314_x_at	3812	ENSG00000284528
NK CD56dim cells	KIR3DL2	207314_x_at	3812	ENSG00000284053
NK CD56dim cells	KIR3DL2	207314_x_at	3812	ENSG00000284192
NK CD56dim cells	KIR3DL2	207314_x_at	3812	ENSG00000283951
NK CD56dim cells	KIR3DL2	216907_x_at	3812	ENSG00000278656
NK CD56dim cells	KIR3DL2	216907_x_at	3812	ENSG00000278361
NK CD56dim cells	KIR3DL2	216907_x_at	3812	ENSG00000277982
NK CD56dim cells	KIR3DL2	216907_x_at	3812	ENSG00000275626
NK CD56dim cells	KIR3DL2	216907_x_at	3812	ENSG00000276357
NK CD56dim cells	KIR3DL2	216907_x_at	3812	ENSG00000275511
NK CD56dim cells	KIR3DL2	216907_x_at	3812	ENSG00000273735
NK CD56dim cells	KIR3DL2	216907_x_at	3812	ENSG00000275262
NK CD56dim cells	KIR3DL2	216907_x_at	3812	ENSG00000278442
NK CD56dim cells	KIR3DL2	216907_x_at	3812	ENSG00000278707
NK CD56dim cells	KIR3DL2	216907_x_at	3812	ENSG00000278758
NK CD56dim cells	KIR3DL2	216907_x_at	3812	ENSG00000277181
NK CD56dim cells	KIR3DL2	216907_x_at	3812	ENSG00000275838
NK CD56dim cells	KIR3DL2	216907_x_at	3812	ENSG00000276739
NK CD56dim cells	KIR3DL2	216907_x_at	3812	ENSG00000277709
NK CD56dim cells	KIR3DL2	216907_x_at	3812	ENSG00000273911
NK CD56dim cells	KIR3DL2	216907_x_at	3812	ENSG00000276004
NK CD56dim cells	KIR3DL2	216907_x_at	3812	ENSG00000278403
NK CD56dim cells	KIR3DL2	216907_x_at	3812	ENSG00000275083
NK CD56dim cells	KIR3DL2	216907_x_at	3812	ENSG00000278809
NK CD56dim cells	KIR3DL2	216907_x_at	3812	ENSG00000278726
NK CD56dim cells	KIR3DL2	216907_x_at	3812	ENSG00000275629
NK CD56dim cells	KIR3DL2	216907_x_at	3812	ENSG00000276882
NK CD56dim cells	KIR3DL2	216907_x_at	3812	ENSG00000240403
NK CD56dim cells	KIR3DL2	216907_x_at	3812	ENSG00000278710
NK CD56dim cells	KIR3DL2	216907_x_at	3812	ENSG00000274722
NK CD56dim cells	KIR3DL2	216907_x_at	3812	ENSG00000275416
NK CD56dim cells	KIR3DL2	216907_x_at	3812	ENSG00000278474
NK CD56dim cells	KIR3DL2	216907_x_at	3812	ENSG00000275566
NK CD56dim cells	KIR3DL2	216907_x_at	3812	ENSG00000278850
NK CD56dim cells	KIR3DL2	216907_x_at	3812	ENSG00000276424
NK CD56dim cells	KIR3DL2	216907_x_at	3812	ENSG00000284295
NK CD56dim cells	KIR3DL2	216907_x_at	3812	ENSG00000284384
NK CD56dim cells	KIR3DL2	216907_x_at	3812	ENSG00000284466
NK CD56dim cells	KIR3DL2	216907_x_at	3812	ENSG00000284101
NK CD56dim cells	KIR3DL2	216907_x_at	3812	ENSG00000284213
NK CD56dim cells	KIR3DL2	216907_x_at	3812	ENSG00000284046
NK CD56dim cells	KIR3DL2	216907_x_at	3812	ENSG00000284063
NK CD56dim cells	KIR3DL2	216907_x_at	3812	ENSG00000283975
NK CD56dim cells	KIR3DL2	216907_x_at	3812	ENSG00000284528
NK CD56dim cells	KIR3DL2	216907_x_at	3812	ENSG00000284053
NK CD56dim cells	KIR3DL2	216907_x_at	3812	ENSG00000284192
NK CD56dim cells	KIR3DL2	216907_x_at	3812	ENSG00000283951
NK CD56dim cells	SPON2	218638_s_at	10417	ENSG00000159674
NK CD56dim cells	KIR2DL3	208179_x_at	3804	ENSG00000275623

NK CD56dim cells	KIR2DL3	208179_x_at	3804	ENSG00000275008
NK CD56dim cells	KIR2DL3	208179_x_at	3804	ENSG00000274952
NK CD56dim cells	KIR2DL3	208179_x_at	3804	ENSG00000276459
NK CD56dim cells	KIR2DL3	208179_x_at	3804	ENSG00000277554
NK CD56dim cells	KIR2DL3	208179_x_at	3804	ENSG00000276590
NK CD56dim cells	KIR2DL3	208179_x_at	3804	ENSG00000274410
NK CD56dim cells	KIR2DL3	208179_x_at	3804	ENSG00000275658
NK CD56dim cells	KIR2DL3	208179_x_at	3804	ENSG00000274830
NK CD56dim cells	KIR2DL3	208179_x_at	3804	ENSG00000278327
NK CD56dim cells	KIR2DL3	208179_x_at	3804	ENSG00000274402
NK CD56dim cells	KIR2DL3	208179_x_at	3804	ENSG00000277484
NK CD56dim cells	KIR2DL3	208179_x_at	3804	ENSG00000277317
NK CD56dim cells	KIR2DL3	208179_x_at	3804	ENSG00000273887
NK CD56dim cells	KIR2DL3	208179_x_at	3804	ENSG00000243772
NK CD56dim cells	KIR2DL3	208179_x_at	3804	ENSG00000274108
NK CD56dim cells	KIR2DL3	208179_x_at	3804	ENSG00000273947
NK CD56dim cells	KIR2DL3	208179_x_at	3804	ENSG00000277924
NK CD56dim cells	KIR2DL3	208179_x_at	3804	ENSG00000278369
NK CD56dim cells	KIR2DL3	208179_x_at	3804	ENSG00000276218
NK CD56dim cells	KIR2DL3	208179_x_at	3804	ENSG00000284132
NK CD56dim cells	KIR2DL3	208179_x_at	3804	ENSG00000284504
NK CD56dim cells	KIR2DL3	208179_x_at	3804	ENSG00000284510
NK CD56dim cells	KIR2DL3	208179_x_at	3804	ENSG00000283996
NK CD56dim cells	KIR2DL3	208179_x_at	3804	ENSG00000283708
NK CD56dim cells	KIR2DL3	208179_x_at	3804	ENSG00000283702
NK CD56dim cells	KIR2DL3	208179_x_at	3804	ENSG00000283790
NK CD56dim cells	KIR2DL3	208179_x_at	3804	ENSG00000284236
NK CD56dim cells	KIR2DL3	208179_x_at	3804	ENSG00000284044
NK CD56dim cells	KIR2DL3	208179_x_at	3804	ENSG00000284333
NK CD56dim cells	KIR2DL3	208179_x_at	3804	ENSG00000284241
NK CD56dim cells	GZMB	210164_at	3002	ENSG00000100453
NK CD56dim cells	KIR3DS1	211389_x_at	3813	ENSG00000274283
NK CD56dim cells	KIR3DS1	211389_x_at	3813	ENSG00000275608
NK CD56dim cells	KIR3DS1	211389_x_at	3813	ENSG00000276534
NK CD56dim cells	KIR3DS1	211389_x_at	3813	ENSG00000275434
NK CD56dim cells	KIR3DS1	211389_x_at	3813	ENSG00000276498
NK CD56dim cells	KIR3DS1	211389_x_at	3813	ENSG00000275037
NK CD56dim cells	KIR3DS1	211389_x_at	3813	ENSG00000274465
NK CD56dim cells	KIR3DL1	211687_x_at	3811	ENSG00000274146
NK CD56dim cells	KIR3DL1	211687_x_at	3811	ENSG00000277175
NK CD56dim cells	KIR3DL1	211687_x_at	3811	ENSG00000274920
NK CD56dim cells	KIR3DL1	211687_x_at	3811	ENSG00000278427
NK CD56dim cells	KIR3DL1	211687_x_at	3811	ENSG00000278856
NK CD56dim cells	KIR3DL1	211687_x_at	3811	ENSG00000273518
NK CD56dim cells	KIR3DL1	211687_x_at	3811	ENSG00000276501
NK CD56dim cells	KIR3DL1	211687_x_at	3811	ENSG00000277272
NK CD56dim cells	KIR3DL1	211687_x_at	3811	ENSG00000274036
NK CD56dim cells	KIR3DL1	211687_x_at	3811	ENSG00000274948
NK CD56dim cells	KIR3DL1	211687_x_at	3811	ENSG00000275717
NK CD56dim cells	KIR3DL1	211687_x_at	3811	ENSG00000275486
NK CD56dim cells	KIR3DL1	211687_x_at	3811	ENSG00000276379
NK CD56dim cells	KIR3DL1	211687_x_at	3811	ENSG00000275288
NK CD56dim cells	KIR3DL1	211687_x_at	3811	ENSG00000275786
NK CD56dim cells	KIR3DL1	211687_x_at	3811	ENSG00000167633

NK CD56dim cells	KIR3DL1	211687_x_at	3811	ENSG00000278368
NK CD56dim cells	KIR3DL1	211687_x_at	3811	ENSG00000276423
NK CD56dim cells	KIR3DL1	211687_x_at	3811	ENSG00000278079
NK CD56dim cells	KIR3DL1	211687_x_at	3811	ENSG00000275545
NK CD56dim cells	KIR3DL1	211687_x_at	3811	ENSG00000275659
NK CD56dim cells	KIR3DL1	211687_x_at	3811	ENSG00000273775
NK CD56dim cells	KIR3DL1	211687_x_at	3811	ENSG00000276329
NK CD56dim cells	KIR3DL1	211687_x_at	3811	ENSG00000284426
NK CD56dim cells	KIR3DL1	211687_x_at	3811	ENSG00000283954
NK CD56dim cells	KIR3DL1	211687_x_at	3811	ENSG00000283729
NK CD56dim cells	KIR3DL1	211687_x_at	3811	ENSG00000284589
NK CD56dim cells	KIR3DL1	211687_x_at	3811	ENSG00000284177
NK CD56dim cells	KIR3DL1	211687_x_at	3811	ENSG00000283731
NK CD56dim cells	KIR3DL1	211687_x_at	3811	ENSG00000284093
NK CD56dim cells	KIR3DL1	211687_x_at	3811	ENSG00000283827
NK CD56dim cells	FLJ20699	216434_at	55020	ENSG00000075234
NK CD56dim cells	TMEPAI	217875_s_at	56937	ENSG00000124225
NK CD56dim cells	KIR3DL2	211688_x_at	3812	ENSG00000278656
NK CD56dim cells	KIR3DL2	211688_x_at	3812	ENSG00000278361
NK CD56dim cells	KIR3DL2	211688_x_at	3812	ENSG00000277982
NK CD56dim cells	KIR3DL2	211688_x_at	3812	ENSG00000275626
NK CD56dim cells	KIR3DL2	211688_x_at	3812	ENSG00000276357
NK CD56dim cells	KIR3DL2	211688_x_at	3812	ENSG00000275511
NK CD56dim cells	KIR3DL2	211688_x_at	3812	ENSG00000273735
NK CD56dim cells	KIR3DL2	211688_x_at	3812	ENSG00000275262
NK CD56dim cells	KIR3DL2	211688_x_at	3812	ENSG00000278442
NK CD56dim cells	KIR3DL2	211688_x_at	3812	ENSG00000278707
NK CD56dim cells	KIR3DL2	211688_x_at	3812	ENSG00000278758
NK CD56dim cells	KIR3DL2	211688_x_at	3812	ENSG00000277181
NK CD56dim cells	KIR3DL2	211688_x_at	3812	ENSG00000275838
NK CD56dim cells	KIR3DL2	211688_x_at	3812	ENSG00000276739
NK CD56dim cells	KIR3DL2	211688_x_at	3812	ENSG00000277709
NK CD56dim cells	KIR3DL2	211688_x_at	3812	ENSG00000273911
NK CD56dim cells	KIR3DL2	211688_x_at	3812	ENSG00000276004
NK CD56dim cells	KIR3DL2	211688_x_at	3812	ENSG00000278403
NK CD56dim cells	KIR3DL2	211688_x_at	3812	ENSG00000275083
NK CD56dim cells	KIR3DL2	211688_x_at	3812	ENSG00000278809
NK CD56dim cells	KIR3DL2	211688_x_at	3812	ENSG00000278726
NK CD56dim cells	KIR3DL2	211688_x_at	3812	ENSG00000275629
NK CD56dim cells	KIR3DL2	211688_x_at	3812	ENSG00000276882
NK CD56dim cells	KIR3DL2	211688_x_at	3812	ENSG00000240403
NK CD56dim cells	KIR3DL2	211688_x_at	3812	ENSG00000278710
NK CD56dim cells	KIR3DL2	211688_x_at	3812	ENSG00000274722
NK CD56dim cells	KIR3DL2	211688_x_at	3812	ENSG00000275416
NK CD56dim cells	KIR3DL2	211688_x_at	3812	ENSG00000278474
NK CD56dim cells	KIR3DL2	211688_x_at	3812	ENSG00000275566
NK CD56dim cells	KIR3DL2	211688_x_at	3812	ENSG00000278850
NK CD56dim cells	KIR3DL2	211688_x_at	3812	ENSG00000276424
NK CD56dim cells	KIR3DL2	211688_x_at	3812	ENSG00000284295
NK CD56dim cells	KIR3DL2	211688_x_at	3812	ENSG00000284384
NK CD56dim cells	KIR3DL2	211688_x_at	3812	ENSG00000284466
NK CD56dim cells	KIR3DL2	211688_x_at	3812	ENSG00000284101
NK CD56dim cells	KIR3DL2	211688_x_at	3812	ENSG00000284213
NK CD56dim cells	KIR3DL2	211688_x_at	3812	ENSG00000284046

NK CD56dim cells	KIR3DL2	211688_x_at	3812	ENSG00000284063
NK CD56dim cells	KIR3DL2	211688_x_at	3812	ENSG00000283975
NK CD56dim cells	KIR3DL2	211688_x_at	3812	ENSG00000284528
NK CD56dim cells	KIR3DL2	211688_x_at	3812	ENSG00000284053
NK CD56dim cells	KIR3DL2	211688_x_at	3812	ENSG00000284192
NK CD56dim cells	KIR3DL2	211688_x_at	3812	ENSG00000283951
NK CD56dim cells	IL21R	221658_s_at	50615	ENSG00000103522
NK CD56dim cells	KIR3DL2	207313_x_at	3812	ENSG00000278656
NK CD56dim cells	KIR3DL2	207313_x_at	3812	ENSG00000278361
NK CD56dim cells	KIR3DL2	207313_x_at	3812	ENSG00000277982
NK CD56dim cells	KIR3DL2	207313_x_at	3812	ENSG00000275626
NK CD56dim cells	KIR3DL2	207313_x_at	3812	ENSG00000276357
NK CD56dim cells	KIR3DL2	207313_x_at	3812	ENSG00000275511
NK CD56dim cells	KIR3DL2	207313_x_at	3812	ENSG00000273735
NK CD56dim cells	KIR3DL2	207313_x_at	3812	ENSG00000275262
NK CD56dim cells	KIR3DL2	207313_x_at	3812	ENSG00000278442
NK CD56dim cells	KIR3DL2	207313_x_at	3812	ENSG00000278707
NK CD56dim cells	KIR3DL2	207313_x_at	3812	ENSG00000278758
NK CD56dim cells	KIR3DL2	207313_x_at	3812	ENSG00000277181
NK CD56dim cells	KIR3DL2	207313_x_at	3812	ENSG00000275838
NK CD56dim cells	KIR3DL2	207313_x_at	3812	ENSG00000276739
NK CD56dim cells	KIR3DL2	207313_x_at	3812	ENSG00000277709
NK CD56dim cells	KIR3DL2	207313_x_at	3812	ENSG00000273911
NK CD56dim cells	KIR3DL2	207313_x_at	3812	ENSG00000276004
NK CD56dim cells	KIR3DL2	207313_x_at	3812	ENSG00000278403
NK CD56dim cells	KIR3DL2	207313_x_at	3812	ENSG00000275083
NK CD56dim cells	KIR3DL2	207313_x_at	3812	ENSG00000278809
NK CD56dim cells	KIR3DL2	207313_x_at	3812	ENSG00000278726
NK CD56dim cells	KIR3DL2	207313_x_at	3812	ENSG00000275629
NK CD56dim cells	KIR3DL2	207313_x_at	3812	ENSG00000276882
NK CD56dim cells	KIR3DL2	207313_x_at	3812	ENSG00000240403
NK CD56dim cells	KIR3DL2	207313_x_at	3812	ENSG00000278710
NK CD56dim cells	KIR3DL2	207313_x_at	3812	ENSG00000274722
NK CD56dim cells	KIR3DL2	207313_x_at	3812	ENSG00000275416
NK CD56dim cells	KIR3DL2	207313_x_at	3812	ENSG00000278474
NK CD56dim cells	KIR3DL2	207313_x_at	3812	ENSG00000275566
NK CD56dim cells	KIR3DL2	207313_x_at	3812	ENSG00000278850
NK CD56dim cells	KIR3DL2	207313_x_at	3812	ENSG00000276424
NK CD56dim cells	KIR3DL2	207313_x_at	3812	ENSG00000284295
NK CD56dim cells	KIR3DL2	207313_x_at	3812	ENSG00000284384
NK CD56dim cells	KIR3DL2	207313_x_at	3812	ENSG00000284466
NK CD56dim cells	KIR3DL2	207313_x_at	3812	ENSG00000284101
NK CD56dim cells	KIR3DL2	207313_x_at	3812	ENSG00000284213
NK CD56dim cells	KIR3DL2	207313_x_at	3812	ENSG00000284046
NK CD56dim cells	KIR3DL2	207313_x_at	3812	ENSG00000284063
NK CD56dim cells	KIR3DL2	207313_x_at	3812	ENSG00000283975
NK CD56dim cells	KIR3DL2	207313_x_at	3812	ENSG00000284528
NK CD56dim cells	KIR3DL2	207313_x_at	3812	ENSG00000284053
NK CD56dim cells	KIR3DL2	207313_x_at	3812	ENSG00000284192
NK CD56dim cells	KIR3DL2	207313_x_at	3812	ENSG00000283951
NK CD56dim cells	KIR3DL3	216676_x_at	115653	ENSG00000278490
NK CD56dim cells	KIR3DL3	216676_x_at	115653	ENSG00000274763
NK CD56dim cells	KIR3DL3	216676_x_at	115653	ENSG00000275513
NK CD56dim cells	KIR3DL3	216676_x_at	115653	ENSG00000275062

NK CD56dim cells	KIR3DL3	216676_x_at	115653	ENSG00000278723
NK CD56dim cells	KIR3DL3	216676_x_at	115653	ENSG00000274254
NK CD56dim cells	KIR3DL3	216676_x_at	115653	ENSG00000274639
NK CD56dim cells	KIR3DL3	216676_x_at	115653	ENSG00000276086
NK CD56dim cells	KIR3DL3	216676_x_at	115653	ENSG00000274511
NK CD56dim cells	KIR3DL3	216676_x_at	115653	ENSG00000276433
NK CD56dim cells	KIR3DL3	216676_x_at	115653	ENSG00000276328
NK CD56dim cells	KIR3DL3	216676_x_at	115653	ENSG00000277620
NK CD56dim cells	KIR3DL3	216676_x_at	115653	ENSG00000274786
NK CD56dim cells	KIR3DL3	216676_x_at	115653	ENSG00000276196
NK CD56dim cells	KIR3DL3	216676_x_at	115653	ENSG00000275172
NK CD56dim cells	KIR3DL3	216676_x_at	115653	ENSG00000276572
NK CD56dim cells	KIR3DL3	216676_x_at	115653	ENSG00000276875
NK CD56dim cells	KIR3DL3	216676_x_at	115653	ENSG00000274556
NK CD56dim cells	KIR3DL3	216676_x_at	115653	ENSG00000276806
NK CD56dim cells	KIR3DL3	216676_x_at	115653	ENSG00000276084
NK CD56dim cells	KIR3DL3	216676_x_at	115653	ENSG00000274394
NK CD56dim cells	KIR3DL3	216676_x_at	115653	ENSG00000274696
NK CD56dim cells	KIR3DL3	216676_x_at	115653	ENSG00000276930
NK CD56dim cells	KIR3DL3	216676_x_at	115653	ENSG00000277028
NK CD56dim cells	KIR3DL3	216676_x_at	115653	ENSG00000277552
NK CD56dim cells	KIR3DL3	216676_x_at	115653	ENSG00000242019
NK CD56dim cells	KIR3DL3	216676_x_at	115653	ENSG00000277392
NK CD56dim cells	KIR3DL3	216676_x_at	115653	ENSG00000274480
NK CD56dim cells	KIR3DL3	216676_x_at	115653	ENSG00000275433
NK CD56dim cells	KIR3DL3	216676_x_at	115653	ENSG00000278729
NK CD56dim cells	KIR3DL3	216676_x_at	115653	ENSG00000277596
NK CD56dim cells	KIR3DL3	216676_x_at	115653	ENSG00000273502
NK CD56dim cells	KIR3DL3	216676_x_at	115653	ENSG00000274724
NK CD56dim cells	KIR3DL3	216676_x_at	115653	ENSG00000283875
NK CD56dim cells	KIR3DL3	216676_x_at	115653	ENSG00000284104
NK CD56dim cells	KIR3DL3	216676_x_at	115653	ENSG00000284371
NK CD56dim cells	KIR3DL3	216676_x_at	115653	ENSG00000283823
NK CD56dim cells	KIR3DL3	216676_x_at	115653	ENSG00000283915
NK CD56dim cells	KIR3DL3	216676_x_at	115653	ENSG00000284086
NK CD56dim cells	KIR3DL3	216676_x_at	115653	ENSG00000283966
NK CD56dim cells	KIR3DL3	216676_x_at	115653	ENSG00000284480
NK CD56dim cells	KIR3DL3	216676_x_at	115653	ENSG00000284127
NK CD56dim cells	KIR2DS5	208203_x_at	3810	ENSG00000274739
NK CD56dim cells	KIR2DS5	208203_x_at	3810	ENSG00000276676
NK CD56dim cells	KIR2DS2	211532_x_at	3807	NaN
NK CD56dim cells	GTF3C1	202320_at	2975	ENSG00000077235
NK CD56dim cells	KIR2DS1	216552_x_at	3806	ENSG00000278304
NK CD56dim cells	KIR2DS1	216552_x_at	3806	ENSG00000273603
NK CD56dim cells	KIR2DS1	216552_x_at	3806	ENSG00000273517
NK CD56dim cells	KIR2DS1	216552_x_at	3806	ENSG00000278120
NK CD56dim cells	KIR2DS1	216552_x_at	3806	ENSG00000275421
NK CD56dim cells	KIR2DS1	216552_x_at	3806	ENSG00000275306
NK CD56dim cells	KIR2DS1	216552_x_at	3806	ENSG00000276327
NK CD56dim cells	EDG8	221417_x_at	53637	ENSG00000180739
NK CD56bright cells	DUSP4	204014_at	1846	ENSG00000120875
NK CD56bright cells	RRAD	204803_s_at	6236	ENSG00000166592
NK CD56bright cells	XCL1	206365_at	6375	ENSG00000143184
NK CD56bright cells	PLA2G6	215938_s_at	8398	ENSG00000184381

NK CD56bright cells	PLA2G6	204691_x_at	8398	ENSG00000184381
NK CD56bright cells	NIBP	221672_s_at	83696	ENSG00000167632
NK CD56bright cells	FOXJ1	205906_at	2302	ENSG00000129654
NK CD56bright cells	03/06/09	215908_at	10299	ENSG00000145495
NK CD56bright cells	DUSP4	204015_s_at	1846	ENSG00000120875
NK CD56bright cells	PLA2G6	210647_x_at	8398	ENSG00000184381
NK CD56bright cells	MADD	38398_at	8567	ENSG00000110514
NK CD56bright cells	BG255923	215409_at	254531	ENSG00000176454
NK CD56bright cells	MPPED1	206436_at	758	ENSG00000186732
NK CD56bright cells	MUC3B	214676_x_at	57876	NaN
DC	CD209	207277_at	30835	ENSG00000090659
DC	CCL17	207900_at	6361	ENSG00000102970
DC	HSD11B1	205404_at	3290	ENSG00000117594
DC	CCL13	206407_s_at	6357	ENSG00000181374
DC	CCL22	207861_at	6367	ENSG00000102962
DC	PPFIBP2	212841_s_at	8495	ENSG00000166387
DC	NPR1	32625_at	4881	ENSG00000169418
iDC	CD1B	206749_at	910	ENSG00000158485
iDC	VASH1	203940_s_at	22846	ENSG00000071246
iDC	F13A1	203305_at	2162	ENSG00000124491
iDC	CD1E	215784_at	913	ENSG00000158488
iDC	MMP12	204580_at	4321	ENSG00000262406
iDC	FABP4	203980_at	2167	ENSG00000170323
iDC	CLEC10A	206682_at	10462	ENSG00000132514
iDC	SYT17	205613_at	51760	ENSG00000103528
iDC	MS4A6A	219666_at	64231	ENSG00000110077
iDC	CTNS	204925_at	1497	ENSG00000040531
iDC	GUCA1A	206062_at	2978	ENSG00000048545
iDC	CARD9	220162_s_at	64170	ENSG00000187796
iDC	CD1E	208592_s_at	913	ENSG00000158488
iDC	ABCG2	209735_at	9429	ENSG00000118777
iDC	CD1A	210325_at	909	ENSG00000158477
iDC	PPARG	208510_s_at	5468	ENSG00000132170
iDC	RAP1GAP	203911_at	5909	ENSG00000076864
iDC	SLC7A8	216604_s_at	23428	ENSG00000092068
iDC	GSTT1	203815_at	2952	ENSG00000277656
iDC	NM_021941	218019_s_at	8566	ENSG00000160209
iDC	FZD2	210220_at	2535	ENSG00000180340
iDC	CSF1R	203104_at	1436	ENSG00000182578
iDC	HS3ST2	219697_at	9956	ENSG00000122254
iDC	CH25H	206932_at	9023	ENSG00000138135
iDC	LMAN2L	221274_s_at	81562	ENSG00000114988
iDC	SLC26A6	221572_s_at	65010	ENSG00000225697
iDC	BLVRB	202201_at	645	ENSG00000090013
iDC	NUDT9	218375_at	53343	ENSG00000170502
iDC	PREP	204117_at	5550	ENSG00000085377
iDC	TM7SF4	221266_s_at	81501	ENSG00000164935
iDC	TACSTD2	202286_s_at	4070	ENSG00000184292
iDC	CD1C	205987_at	911	ENSG00000158481
aDC	CCL1	207533_at	6346	ENSG00000108702
aDC	EBI3	219424_at	10148	ENSG00000105246
aDC	INDO	210029_at	3620	ENSG00000131203
aDC	LAMP3	205569_at	27074	ENSG00000078081
aDC	OAS3	218400_at	4940	ENSG00000111331

pDC	IL3RA	206148_at	3563	ENSG00000185291
Eosinophils	IL5RA	211517_s_at	3568	ENSG00000091181
Eosinophils	KCNH2	205262_at	3757	ENSG00000055118
Eosinophils	TKTL1	216370_s_at	8277	ENSG00000007350
Eosinophils	IL5RA	210744_s_at	3568	ENSG00000091181
Eosinophils	EMR1	207111_at	2015	ENSG00000174837
Eosinophils	KCNH2	210036_s_at	3757	ENSG00000055118
Eosinophils	CCR3	208304_at	1232	ENSG00000183625
Eosinophils	ACACB	49452_at	32	ENSG00000076555
Eosinophils	THBS1	201108_s_at	7057	ENSG00000137801
Eosinophils	GALC	211810_s_at	2581	ENSG00000054983
Eosinophils	TKTL1	214183_s_at	8277	ENSG00000007350
Eosinophils	RNU2	210230_at	728965	NaN
Eosinophils	CLC	206207_at	1178	ENSG00000105205
Eosinophils	THBS1	201109_s_at	7057	ENSG00000137801
Eosinophils	HIST1H1C	209398_at	3006	ENSG00000187837
Eosinophils	CYSLTR2	220813_at	57105	ENSG00000152207
Eosinophils	HRH4	221170_at	59340	ENSG00000134489
Eosinophils	RNASE2	206111_at	6036	ENSG00000169385
Eosinophils	CAT	211922_s_at	847	ENSG00000121691
Eosinophils	LRP5L	214873_at	91355	ENSG00000100068
Eosinophils	SYNJ1	207594_s_at	8867	ENSG00000159082
Eosinophils	SYNJ1	212990_at	8867	ENSG00000159082
Eosinophils	THBS4	204776_at	7060	ENSG00000113296
Eosinophils	GPR44	206361_at	11251	ENSG00000183134
Eosinophils	KBTBD11	204301_at	9920	ENSG00000176595
Eosinophils	KBTBD11	204301_at	9920	ENSG00000273645
Eosinophils	HES1	203394_s_at	3280	ENSG00000114315
Eosinophils	ABHD2	205566_at	11057	ENSG00000140526
Eosinophils	TIPARP	212665_at	25976	ENSG00000163659
Eosinophils	SMPD3	219695_at	55512	ENSG00000103056
Eosinophils	MYO15B	59375_at	80022	NaN
Eosinophils	TGIF1	203313_s_at	7050	ENSG00000177426
Eosinophils	RRP12	216360_x_at	23223	ENSG00000052749
Eosinophils	ACACB	43427_at	32	ENSG00000076555
Eosinophils	IGSF2	207167_at	9398	ENSG00000134256
Eosinophils	HES1	203395_s_at	3280	ENSG00000114315
Eosinophils	RCOR3	218344_s_at	55758	ENSG00000117625
Eosinophils	EPN2	203463_s_at	22905	ENSG00000072134
Eosinophils	C9orf156	222195_s_at	51531	ENSG00000136932
Eosinophils	SIAH1	202981_x_at	6477	ENSG00000196470
Eosinophils	ACACB	221928_at	32	ENSG00000076555
Macrophages	MARCO	205819_at	8685	ENSG00000019169
Macrophages	CXCL5	214974_x_at	6374	ENSG00000163735
Macrophages	SCG5	203889_at	6447	ENSG00000166922
Macrophages	SCG5	203889_at	6447	ENSG00000277614
Macrophages	SCG5	203889_at	6447	ENSG00000281931
Macrophages	SULT1C2	205342_s_at	6819	ENSG00000198203
Macrophages	SULT1C2	211470_s_at	6819	ENSG00000198203
Macrophages	MSR1	214770_at	4481	ENSG00000038945
Macrophages	CTSK	202450_s_at	1513	ENSG00000143387
Macrophages	PTGDS	212187_x_at	5730	ENSG00000107317
Macrophages	COLEC12	221019_s_at	81035	ENSG00000158270
Macrophages	GPC4	204984_at	2239	ENSG00000076716

Macrophages	MSR1	208423_s_at	4481	ENSG00000038945
Macrophages	PCOLCE2	219295_s_at	26577	ENSG00000163710
Macrophages	CHIT1	208168_s_at	1118	ENSG00000133063
Macrophages	PTGDS	211748_x_at	5730	ENSG00000107317
Macrophages	KAL1	205206_at	3730	ENSG00000011201
Macrophages	CLEC5A	219890_at	23601	ENSG00000258227
Macrophages	GPC4	204983_s_at	2239	ENSG00000076716
Macrophages	ME1	204058_at	4199	ENSG00000065833
Macrophages	DNASE2B	220380_at	58511	ENSG00000137976
Macrophages	CCL7	208075_s_at	6354	ENSG00000108688
Macrophages	FN1	214701_s_at	2335	ENSG00000115414
Macrophages	CD163	203645_s_at	9332	ENSG00000177575
Macrophages	GM2A	215891_s_at	2760	ENSG00000196743
Macrophages	SCARB2	201647_s_at	950	ENSG00000138760
Macrophages	BCAT1	214452_at	586	ENSG00000060982
Macrophages	BCAT1	214390_s_at	586	ENSG00000060982
Macrophages	RAI14	202052_s_at	26064	ENSG00000039560
Macrophages	MSR1	211887_x_at	4481	ENSG00000038945
Macrophages	COL8A2	52651_at	1296	ENSG00000171812
Macrophages	CD163	215049_x_at	9332	ENSG00000177575
Macrophages	APOE	203381_s_at	348	ENSG00000130203
Macrophages	CHI3L1	209396_s_at	1116	ENSG00000133048
Macrophages	ATG7	218673_s_at	10533	ENSG00000197548
Macrophages	CD84	211190_x_at	8832	ENSG00000066294
Macrophages	FDX1	203646_at	2230	ENSG00000137714
Macrophages	MS4A4A	219607_s_at	51338	ENSG00000110079
Macrophages	SGMS1	212989_at	259230	ENSG00000198964
Macrophages	EMP1	201324_at	2012	ENSG00000134531
Macrophages	CYBB	203922_s_at	1536	ENSG00000165168
Macrophages	CD68	203507_at	968	ENSG00000129226
Mast cells	PRG2	211743_s_at	5553	ENSG00000186652
Mast cells	CTSG	205653_at	1511	ENSG00000100448
Mast cells	TPSAB1	215382_x_at	7177	ENSG00000172236
Mast cells	SLC18A2	205857_at	6571	ENSG00000165646
Mast cells	TPSAB1	205683_x_at	7177	ENSG00000172236
Mast cells	MS4A2	207497_s_at	2206	ENSG00000149534
Mast cells	CPA3	205624_at	1359	ENSG00000163751
Mast cells	TPSB2	207134_x_at	64499	ENSG00000197253
Mast cells	TPSAB1	216474_x_at	7177	ENSG00000172236
Mast cells	NM_003293	207741_x_at	64499	ENSG00000197253
Mast cells	TPSAB1	210084_x_at	7177	ENSG00000172236
Mast cells	MS4A2	207496_at	2206	ENSG00000149534
Mast cells	TPSAB1	217023_x_at	7177	ENSG00000172236
Mast cells	GATA2	209710_at	2624	ENSG00000179348
Mast cells	HDC	207067_s_at	3067	ENSG00000140287
Mast cells	LOH11CR2A	210102_at	4013	ENSG00000110002
Mast cells	SIGLEC6	210796_x_at	946	ENSG00000105492
Mast cells	ELA2	206871_at	1991	ENSG00000277571
Mast cells	ELA2	206871_at	1991	ENSG00000197561
Mast cells	LOH11CR2A	205011_at	4013	ENSG00000110002
Mast cells	CMA1	214533_at	1215	ENSG00000092009
Mast cells	SIGLEC6	206520_x_at	946	ENSG00000105492
Mast cells	PGDS	206726_at	27306	ENSG00000163106
Mast cells	MLPH	218211_s_at	79083	ENSG00000115648

Mast cells	ADCYAP1	206281_at	116	ENSG00000141433
Mast cells	SIGLEC6	206519_x_at	946	ENSG00000105492
Mast cells	SLC24A3	57588_at	57419	ENSG00000185052
Mast cells	CALB2	205428_s_at	794	ENSG00000172137
Mast cells	CALB2	205428_s_at	794	ENSG00000282830
Mast cells	SLC24A3	219090_at	57419	ENSG00000185052
Mast cells	KIT	205051_s_at	3815	ENSG00000157404
Mast cells	TAL1	206283_s_at	6886	ENSG00000162367
Mast cells	ABCC4	203196_at	10257	ENSG00000125257
Mast cells	PPM1H	212686_at	57460	ENSG00000111110
Mast cells	MAOB	204041_at	4129	ENSG00000069535
Mast cells	HPGD	211549_s_at	3248	ENSG00000164120
Mast cells	SCG2	204035_at	7857	ENSG00000171951
Mast cells	PTGS1	205127_at	5742	ENSG00000095303
Mast cells	CEACAM8	206676_at	1088	ENSG00000124469
Mast cells	MPO	203949_at	4353	ENSG00000005381
Mast cells	NR0B1	206645_s_at	190	ENSG00000169297
Mast cells	LOC339524	215039_at	339524	ENSG00000267272
Neutrophils	CSF3R	203591_s_at	1441	ENSG00000119535
Neutrophils	CYP4F3	206515_at	4051	ENSG00000186529
Neutrophils	VNN3	220528_at	55350	ENSG00000093134
Neutrophils	FPRL1	210773_s_at	2358	ENSG00000171049
Neutrophils	KCNJ15	216782_at	3772	ENSG00000157551
Neutrophils	MME	203434_s_at	4311	ENSG00000196549
Neutrophils	IL8RA	207094_at	3577	ENSG00000163464
Neutrophils	IL8RB	207008_at	3579	ENSG00000180871
Neutrophils	MME	203435_s_at	4311	ENSG00000196549
Neutrophils	FCGR3B	204007_at	2215	ENSG00000162747
Neutrophils	DYSF	218660_at	8291	ENSG00000135636
Neutrophils	KCNJ15	211806_s_at	3772	ENSG00000157551
Neutrophils	FCAR	211816_x_at	2204	ENSG00000278415
Neutrophils	FCAR	211816_x_at	2204	ENSG00000274580
Neutrophils	FCAR	211816_x_at	2204	ENSG00000275136
Neutrophils	FCAR	211816_x_at	2204	ENSG00000275970
Neutrophils	FCAR	211816_x_at	2204	ENSG00000276985
Neutrophils	FCAR	211816_x_at	2204	ENSG00000186431
Neutrophils	FCAR	211816_x_at	2204	ENSG00000276858
Neutrophils	FCAR	211816_x_at	2204	ENSG00000275269
Neutrophils	FCAR	211816_x_at	2204	ENSG00000273738
Neutrophils	FCAR	211816_x_at	2204	ENSG00000275564
Neutrophils	FCAR	211816_x_at	2204	ENSG00000283953
Neutrophils	FCAR	211816_x_at	2204	ENSG00000283750
Neutrophils	FCAR	211816_x_at	2204	ENSG00000284245
Neutrophils	FCAR	211816_x_at	2204	ENSG00000284004
Neutrophils	FCAR	211816_x_at	2204	ENSG00000284061
Neutrophils	FCAR	211307_s_at	2204	ENSG00000278415
Neutrophils	FCAR	211307_s_at	2204	ENSG00000274580
Neutrophils	FCAR	211307_s_at	2204	ENSG00000275136
Neutrophils	FCAR	211307_s_at	2204	ENSG00000275970
Neutrophils	FCAR	211307_s_at	2204	ENSG00000276985
Neutrophils	FCAR	211307_s_at	2204	ENSG00000186431
Neutrophils	FCAR	211307_s_at	2204	ENSG00000276858
Neutrophils	FCAR	211307_s_at	2204	ENSG00000275269
Neutrophils	FCAR	211307_s_at	2204	ENSG00000273738

Neutrophils	FCAR	211307_s_at	2204	ENSG00000275564
Neutrophils	FCAR	211307_s_at	2204	ENSG00000283953
Neutrophils	FCAR	211307_s_at	2204	ENSG00000283750
Neutrophils	FCAR	211307_s_at	2204	ENSG00000284245
Neutrophils	FCAR	211307_s_at	2204	ENSG00000284004
Neutrophils	FCAR	211307_s_at	2204	ENSG00000284061
Neutrophils	CEACAM3	210789_x_at	1084	ENSG00000170956
Neutrophils	FPRL1	210772_at	2358	ENSG00000171049
Neutrophils	HIST1H2BC	214455_at	8347	ENSG00000180596
Neutrophils	HPSE	219403_s_at	10855	ENSG00000173083
Neutrophils	FLJ11151	218610_s_at	55313	ENSG00000103381
Neutrophils	CREB5	205931_s_at	9586	ENSG00000146592
Neutrophils	S100A12	205863_at	6283	ENSG00000163221
Neutrophils	FCGR3B	204006_s_at	2215	ENSG00000162747
Neutrophils	TNFRSF10C	211163_s_at	8794	ENSG00000173535
Neutrophils	SLC22A4	205896_at	6583	ENSG00000197208
Neutrophils	KIAA0329	204307_at	9895	ENSG00000196663
Neutrophils	SLC25A37	218136_s_at	51312	ENSG00000147454
Neutrophils	BST1	205715_at	683	ENSG00000109743
Neutrophils	FCAR	207674_at	2204	ENSG00000278415
Neutrophils	FCAR	207674_at	2204	ENSG00000274580
Neutrophils	FCAR	207674_at	2204	ENSG00000275136
Neutrophils	FCAR	207674_at	2204	ENSG00000275970
Neutrophils	FCAR	207674_at	2204	ENSG00000276985
Neutrophils	FCAR	207674_at	2204	ENSG00000186431
Neutrophils	FCAR	207674_at	2204	ENSG00000276858
Neutrophils	FCAR	207674_at	2204	ENSG00000275269
Neutrophils	FCAR	207674_at	2204	ENSG00000273738
Neutrophils	FCAR	207674_at	2204	ENSG00000275564
Neutrophils	FCAR	207674_at	2204	ENSG00000283953
Neutrophils	FCAR	207674_at	2204	ENSG00000283750
Neutrophils	FCAR	207674_at	2204	ENSG00000284245
Neutrophils	FCAR	207674_at	2204	ENSG00000284004
Neutrophils	FCAR	207674_at	2204	ENSG00000284061
Neutrophils	CEACAM3	208052_x_at	1084	ENSG00000170956
Neutrophils	CRISPLD2	221541_at	83716	ENSG00000103196
Neutrophils	TNFRSF10C	206222_at	8794	ENSG00000173535
Neutrophils	G0S2	213524_s_at	50486	ENSG00000123689
Neutrophils	SIGLEC5	220000_at	8778	ENSG00000105501
Neutrophils	CD93	202878_s_at	22918	ENSG00000125810
Neutrophils	MGAM	206522_at	8972	ENSG00000257335
Neutrophils	MGAM	206522_at	8972	ENSG00000282607
Neutrophils	ALPL	215783_s_at	249	ENSG00000162551
Neutrophils	FPR1	205119_s_at	2357	ENSG00000171051
Neutrophils	CD93	202877_s_at	22918	ENSG00000125810
Neutrophils	PDE4B	222326_at	5142	ENSG00000184588
Neutrophils	LILRB2	210146_x_at	10288	ENSG00000274513
Neutrophils	LILRB2	210146_x_at	10288	ENSG00000131042
Neutrophils	LILRB2	210146_x_at	10288	ENSG00000275463
Neutrophils	LILRB2	210146_x_at	10288	ENSG00000277751
Neutrophils	LILRB2	210146_x_at	10288	ENSG00000276146
SW480 cancer cells	KRT5	201820_at	3852	ENSG00000186081
SW480 cancer cells	RBP1	203423_at	5947	ENSG00000114115
SW480 cancer cells	TRIM29	202504_at	23650	ENSG00000137699

SW480 cancer cells	DEFA5	207529_at	1670	ENSG00000164816
SW480 cancer cells	BMP4	211518_s_at	652	ENSG00000125378
SW480 cancer cells	EEF1A2	204540_at	1917	ENSG00000101210
SW480 cancer cells	VSNL1	203797_at	7447	ENSG00000163032
SW480 cancer cells	ASPSCR1	218908_at	79058	ENSG00000169696
SW480 cancer cells	IGF2	202409_at	3481	ENSG00000167244
SW480 cancer cells	IGF2	202409_at	3481	ENSG00000129965
SW480 cancer cells	MFAP2	203417_at	4237	ENSG00000117122
SW480 cancer cells	FGF3	214571_at	2248	ENSG00000186895
SW480 cancer cells	S100A2	204268_at	6273	ENSG00000196754
SW480 cancer cells	INHBB	205258_at	3625	ENSG00000163083
SW480 cancer cells	JAG2	209784_s_at	3714	ENSG00000184916
SW480 cancer cells	LOC89944	213713_s_at	89944	ENSG00000149328
SW480 cancer cells	BAMBI	203304_at	25805	ENSG00000095739
SW480 cancer cells	JAG2	32137_at	3714	ENSG00000184916
SW480 cancer cells	BMP7	209591_s_at	655	ENSG00000101144
SW480 cancer cells	RPP25	219143_s_at	54913	ENSG00000178718
SW480 cancer cells	RHOD	209885_at	29984	ENSG00000173156
SW480 cancer cells	DHRS2	214079_at	10202	ENSG00000100867
SW480 cancer cells	ITGB4	204989_s_at	3691	ENSG00000132470
SW480 cancer cells	NTSR1	207360_s_at	4923	ENSG00000101188
SW480 cancer cells	STRA6	221701_s_at	64220	ENSG00000137868
SW480 cancer cells	SLC1A5	208916_at	6510	ENSG00000105281
SW480 cancer cells	VSNL1	203798_s_at	7447	ENSG00000163032
SW480 cancer cells	FKBP4	200894_s_at	2288	ENSG00000004478
SW480 cancer cells	S100A3	206027_at	6274	ENSG00000188015
SW480 cancer cells	TEAD4	41037_at	7004	ENSG00000197905
SW480 cancer cells	KLK6	204733_at	5653	ENSG00000167755
SW480 cancer cells	CCND1	208711_s_at	595	ENSG00000110092
SW480 cancer cells	SLC27A5	219733_s_at	10998	ENSG00000083807
SW480 cancer cells	HOXA9	214651_s_at	3205	ENSG00000078399
SW480 cancer cells	F12	205774_at	2161	ENSG00000131187
SW480 cancer cells	LRFN4	219491_at	78999	ENSG00000173621
SW480 cancer cells	NM_024609	218678_at	10763	ENSG00000132688
SW480 cancer cells	SLC6A8	210854_x_at	6535	ENSG00000130821
SW480 cancer cells	PCTK1	207239_s_at	5127	ENSG00000102225
SW480 cancer cells	SLC6A8	213843_x_at	6535	ENSG00000130821
SW480 cancer cells	KRT13	207935_s_at	3860	ENSG00000171401
Normal mucosa	TSPAN8	203824_at	7103	ENSG00000127324
Normal mucosa	LGALS4	204272_at	3960	ENSG00000171747
Normal mucosa	LGALS4	204272_at	3960	ENSG00000282992
Normal mucosa	DCN	201893_x_at	1634	ENSG00000011465
Normal mucosa	COL3A1	215076_s_at	1281	ENSG00000168542
Normal mucosa	COL3A1	201852_x_at	1281	ENSG00000168542
Normal mucosa	CEACAM5	201884_at	1048	ENSG00000105388
Normal mucosa	TAGLN	205547_s_at	6876	ENSG00000149591
Normal mucosa	KRT20	213953_at	54474	ENSG00000263057
Normal mucosa	KRT20	213953_at	54474	ENSG00000171431
Normal mucosa	DCN	211896_s_at	1634	ENSG00000011465
Normal mucosa	MYH11	201497_x_at	4629	ENSG00000133392
Normal mucosa	MYH11	201497_x_at	4629	ENSG00000276480
Normal mucosa	FXYD3	202489_s_at	5349	ENSG00000089356
Normal mucosa	ACTG2	202274_at	72	ENSG00000163017
Normal mucosa	MYLK	202555_s_at	4638	ENSG00000065534

Normal mucosa	TPM1	210987_x_at	7168	ENSG00000140416
Normal mucosa	CDH17	209847_at	1015	ENSG00000079112
Normal mucosa	NFIB	209289_at	4781	ENSG00000147862
Normal mucosa	MGP	202291_s_at	4256	ENSG00000111341
Normal mucosa	SPARCL1	200795_at	8404	ENSG00000152583
Normal mucosa	RGS5	209071_s_at	8490	ENSG00000232995
Normal mucosa	RGS5	209071_s_at	8490	ENSG00000143248
Normal mucosa	MYH11	207961_x_at	4629	ENSG00000133392
Normal mucosa	MYH11	207961_x_at	4629	ENSG00000276480
Normal mucosa	PPAP2B	212226_s_at	8613	ENSG00000162407
Normal mucosa	COL3A1	211161_s_at	1281	ENSG00000168542
Normal mucosa	IGFBP7	201162_at	3490	ENSG00000163453
Normal mucosa	PPAP2B	209355_s_at	8613	ENSG00000162407
Normal mucosa	CALD1	201616_s_at	800	ENSG00000122786
Normal mucosa	C1R	212067_s_at	715	ENSG00000159403
Normal mucosa	CALD1	212077_at	800	ENSG00000122786
Normal mucosa	AGR2	209173_at	10551	ENSG00000106541
Normal mucosa	GNA11	564_at	2767	ENSG00000088256
Normal mucosa	HEPH	203903_s_at	9843	ENSG00000089472
Normal mucosa	GNA11	213944_x_at	2767	ENSG00000088256
Normal mucosa	GNG12	212294_at	55970	ENSG00000172380
Normal mucosa	ADH1B	209612_s_at	125	ENSG00000196616
Normal mucosa	TPM1	206116_s_at	7168	ENSG00000140416
Normal mucosa	CRIM1	202551_s_at	51232	ENSG00000277354
Normal mucosa	CRIM1	202551_s_at	51232	ENSG00000150938
Normal mucosa	FBLN1	202994_s_at	2192	ENSG00000077942
Normal mucosa	IGFBP5	211959_at	3488	ENSG00000115461
Normal mucosa	LAMA4	202202_s_at	3910	ENSG00000112769
Normal mucosa	CAV1	212097_at	857	ENSG00000105974
Normal mucosa	WASL	205809_s_at	8976	ENSG00000106299
Blood vessels	CDH5	204677_at	1003	ENSG00000179776
Lymph vessels	PDPN	204879_at	10630	ENSG00000162493
Lymph vessels	VEGFC	209946_at	7424	ENSG00000150630
Lymph vessels	FIGF	206742_at	2277	ENSG00000165197

Table S5: Signatures adapted from Bindea et al. [38]. Genes Id were matched from tables available at `https://github.com/judithabk6/ITH_TCGA/tree/master/external_data`