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

Judith Abécassis, Anne-Sophie Hamy, Cécile Laurent, Benjamin Sadacca, Hélène Bonsang-Kitzis, Fabien Reyal, Jean-Philippe Vert

PLOS ONE

RESEARCH ARTICLE

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[2,6]*

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

* jpvert@google.com

OPEN ACCESS

Citation: Abécassis J, Hamy A-S, Laurent C, Sadacca B, Bonsang-Kitzis H, Reyal F, et al. (2019) Assessing reliability of intra-tumor heterogeneity estimates from single sample whole exome sequencing data. PLoS ONE 14(11): e0224143. https://doi.org/10.1371/journal.pone.0224143

Editor: Kyle Ellrott, Oregon Health & Science University, UNITED STATES

Received: March 14, 2019

Accepted: October 7, 2019

Published: November 7, 2019

Data Availability Statement: The data underlying the results presented in the study are available from The Cancer Genome Atlas https://tcga-data.nci.nih.gov/. The code to reproduce all results is available at https://github.com/judithabk6/ITH_TCGA.

Funding: The author(s) received no specific funding for this work. JPV is employed by Google France. The funder provided support in the form of salaries for author JPV but did not have any additional role in the study design, data collection

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,

and analysis, decision to publish, or preparation of the manuscript. The specific role of this author is articulated in the 'author contributions' section.

Competing interests: JPV is employed by Google France. This does not alter our adherence to PLOS ONE policies on sharing data and materials.

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 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 [3–5], and many computational methods have been developed for that purpose [6–9]. 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 [3, 10–12].

Previous studies have reported that a large proportion of tumors are heterogeneous [13–16], 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) [18–20] 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 [13, 14, 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 S1 Table). 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 sub-clone, 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]. S2, S3 and S4 Tables 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 $(\mathbf{x}_i, y_i, \delta_i)_{i=1,\ldots,n}$, 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 $(\mathbf{x}_i, y_i, \delta_i)_{i=1,\ldots,n}$, the CI measures how concordant the predicted survival times $s_i = f(\mathbf{x}_i)$ are with the observed survival times $\boldsymbol{y_i}$ 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 S5 Table. 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 realignment 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. S1 Table 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

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.

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%

https://doi.org/10.1371/journal.pone.0224143.t001

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.

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 Fig 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 Table.

In addition to failures, we observed that the runtime varies significantly between methods (Table 1). As shown on Fig 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

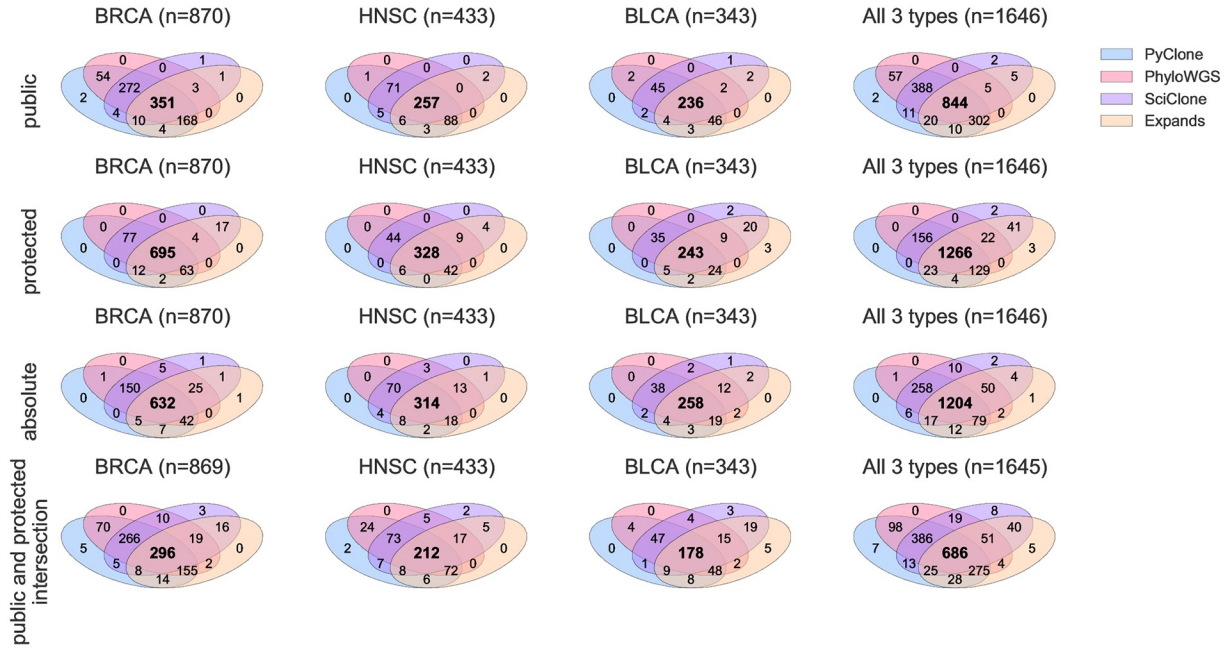

Fig 1. Intersection of successful runs among the 4 considered ITH methods. The upper venn diagrams concern runs with the public input SNV set, the second line with the protected, and the third the overall intersection, as results with both sets are necessary for a proper and rigorous comparison.

https://doi.org/10.1371/journal.pone.0224143.g001

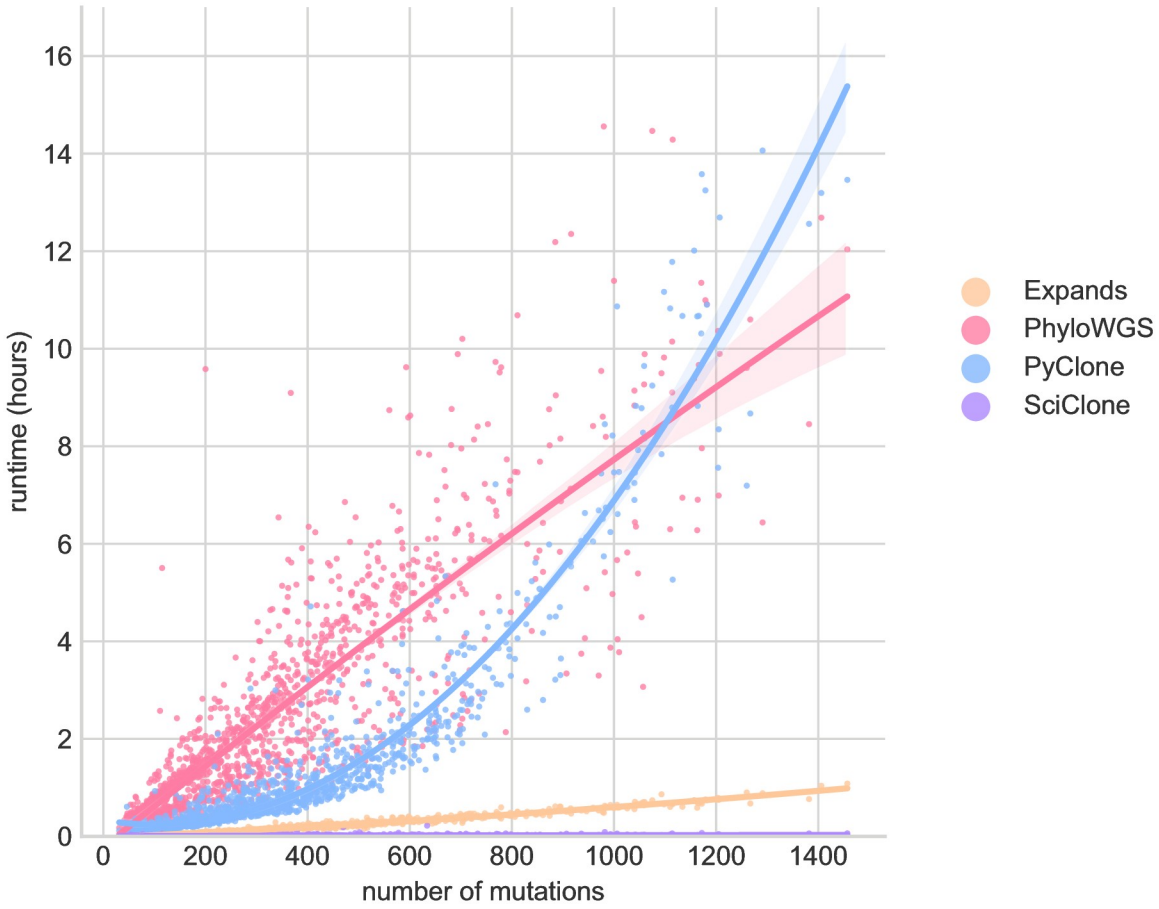

Fig 2. Runtime of the different ITH methods as a function of the number of mutations in each sample. Lines represent second degree polynomial fit with shaded regions are 95% confidence-intervals.

https://doi.org/10.1371/journal.pone.0224143.g002

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. Fig 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.

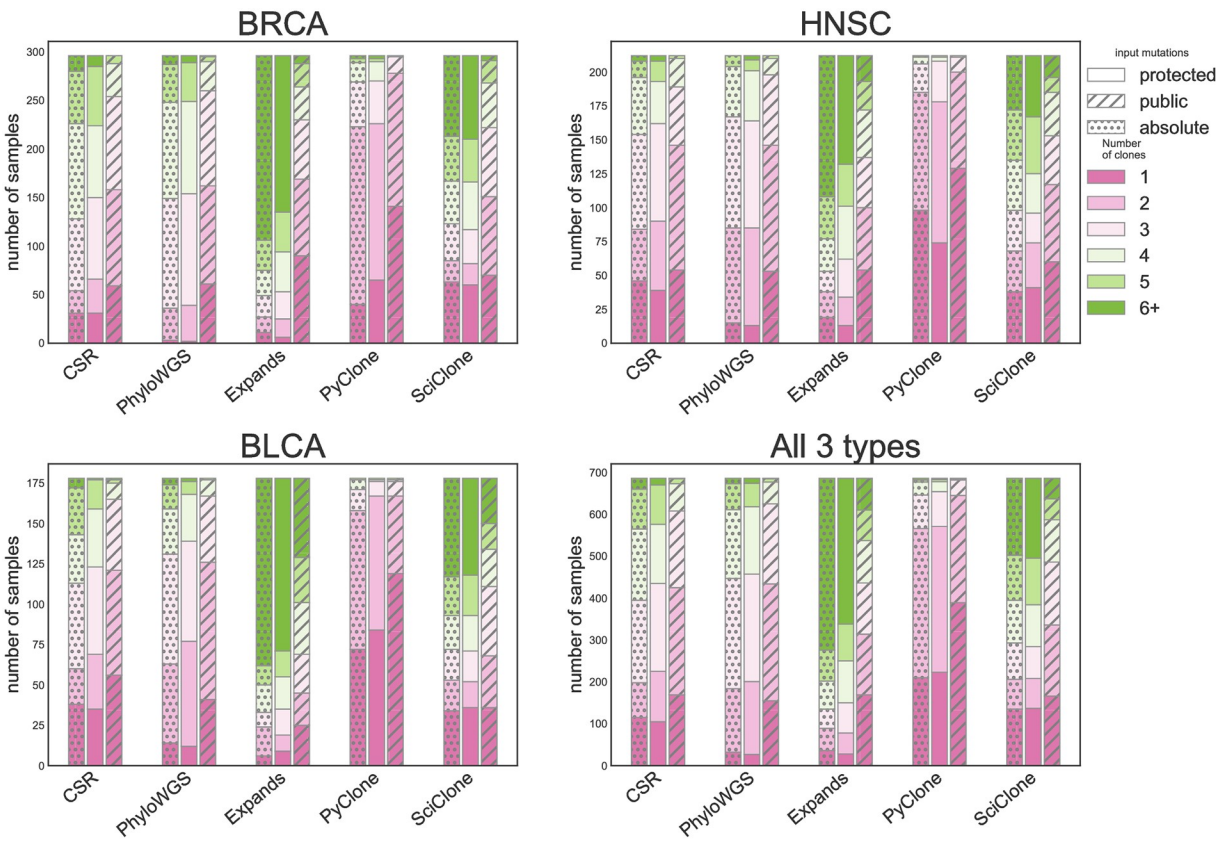

Fig 3. Distribution of number of clones called by ITH methods with public and protected mutations sets as inputs. Distribution of the number of subclones for the tested ITH methods, and 2 alternative input mutation sets for samples in the different cancer types and 2 different copy number methods for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

https://doi.org/10.1371/journal.pone.0224143.g003

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 Fig 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

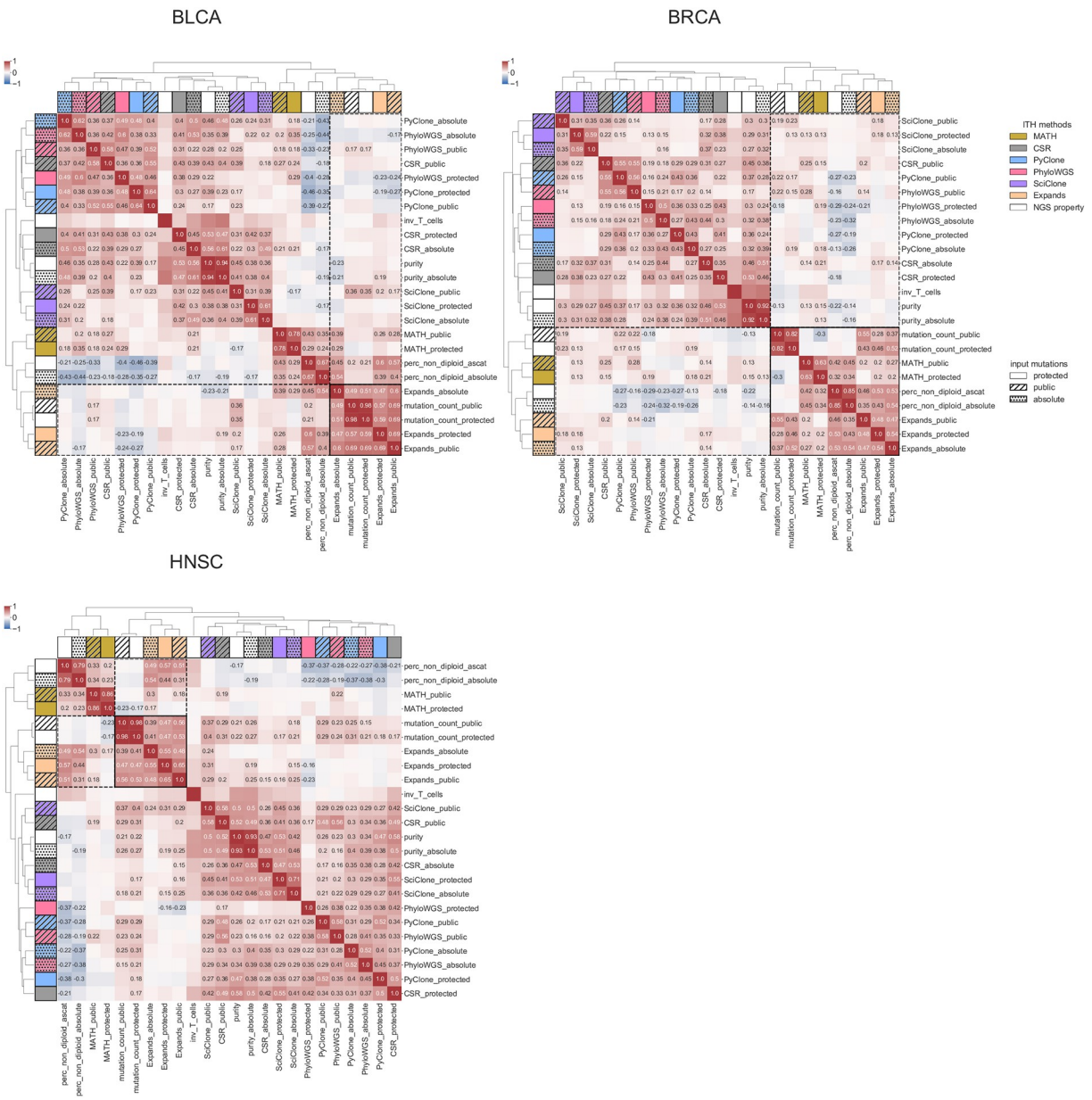

Fig 4. Correlation between various measures of ITH (MATH score, and number of subclones for the other methods), and other potential confounding variables measured using WES and trancriptomics data. Row and color label represent the method used, with white for the genomic measures not involving ITH. Hatches 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.

https://doi.org/10.1371/journal.pone.0224143.g004

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 S2–S4 Figs).

Regarding potential confounding variables, previous studies have reported a correlation between MATH score and CNA abundance [22, 53, 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.

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

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

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

https://doi.org/10.1371/journal.pone.0224143.t002

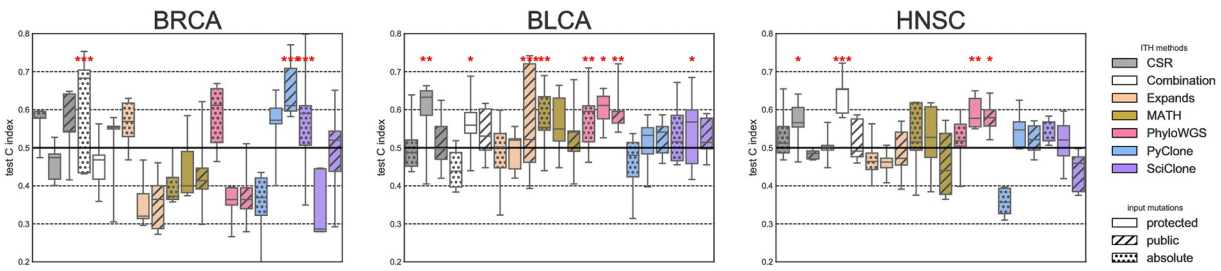

Fig 5. Prognostic power of ITH measured using different ITH method and input mutation set combination. 0.5 corresponds to a random prediction, and stars indicate statistical significance (p-value <0.001: ***, <0.01: **, <0.05: *). Results are presented for 3 cancer types (BRCA, BLCA and HNSC from left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

https://doi.org/10.1371/journal.pone.0224143.g005

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.

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

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 (S1 Fig). 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.

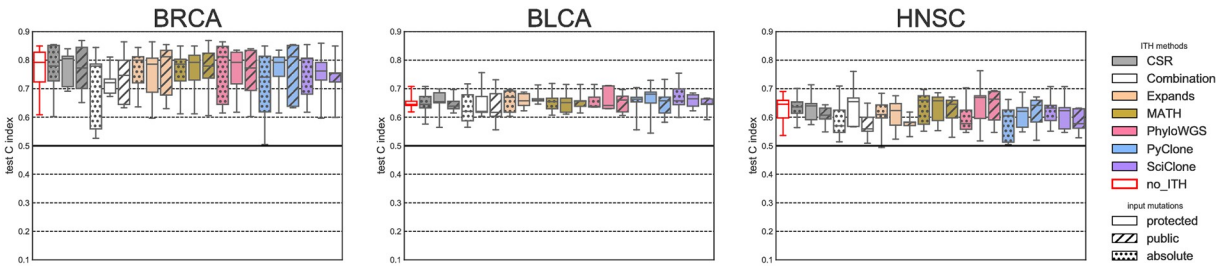

Fig 6. Prognostic power of ITH-derived features compared to other prognostic factors (686 patients in total). ITH-derived features are used in association with clinical features to predict overall survival. Left-most boxplots (with red contour lines) represent results using clinical variables alone,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

https://doi.org/10.1371/journal.pone.0224143.g006

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.

Fig 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.

Discussion

Comparison to similar studies

Previous findings report divergent prognostic power for ITH in several pan cancer studies [13, 14, 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 [13, 14] 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 [13, 53, 54, 68]. 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.

Supporting information

S1 Fig. 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.
(TIF)

S2 Fig. 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.
(TIF)

S3 Fig. 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.
(TIF)

S4 Fig. 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.
(TIF)

S1 Table. Characteristics of input mutation sets. 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.
(PDF)

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

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

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

S5 Table. 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.
(PDF)

Acknowledgments

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 Niko Beerenwickel for help with single-cell data analysis, 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).

Author Contributions

Conceptualization: Judith Abécassis, Fabien Reyal, Jean-Philippe Vert.

Data curation: Judith Abécassis, Cécile Laurent, Benjamin Sadacca, Hélène Bonsang-Kitzis.

Investigation: Judith Abécassis, Anne-Sophie Hamy.

Methodology: Judith Abécassis, Anne-Sophie Hamy, Fabien Reyal, Jean-Philippe Vert.

Software: Judith Abécassis.

Supervision: Fabien Reyal, Jean-Philippe Vert.

Visualization: Judith Abécassis.

Writing – original draft: Judith Abécassis, Jean-Philippe Vert.

Writing – review & editing: Judith Abécassis, Jean-Philippe Vert.

References

1. Nowell PC. The clonal evolution of tumor cell populations. Science. 1976; 194(4260):23–28.
2. Gerstung M, Jolly C, Leshchiner I, Dentro SC, Yu K, Tarabichi M, et al. The evolutionary history of 2,658 cancers. bioRxiv. 2017.
3. Dentro SC, Wedge DC, Van Loo P. Principles of Reconstructing the Subclonal Architecture of Cancers. Cold Spring Harbor perspectives in medicine. 2017; 7(8):a026625. https://doi.org/10.1101/cshperspect.a026625 PMID: 28270531

4. Beerenwinkel N, Schwarz RF, Gerstung M, Markowetz F. Cancer evolution: Mathematical models and computational inference. Systematic Biology. 2015; 64(1):e1–e25. https://doi.org/10.1093/sysbio/syu081 PMID: 25293804
5. Schwartz R, Schäffer AA. The evolution of tumour phylogenetics: Principles and practice. Nature Reviews Genetics. 2017; 18(4):213–229. https://doi.org/10.1038/nrg.2016.170 PMID: 28190876
6. Roth A, Khattra J, Yap D, Wan A, Laks E, Biele J, et al. PyClone: Statistical inference of clonal population structure in cancer. Nature Methods. 2014; 11(4):396–398. https://doi.org/10.1038/nmeth.2883 PMID: 24633410
7. Miller CA, White BS, Dees ND, Griffith M, Welch JS, Griffith OL, et al. SciClone: Inferring Clonal Architecture and Tracking the Spatial and Temporal Patterns of Tumor Evolution. PLoS Computational Biology. 2014; 10(8):e1003665. https://doi.org/10.1371/journal.pcbi.1003665 PMID: 25102416
8. Deshwar AG, Vembu S, Yung CK, Jang GH, Stein L, Morris Q. PhyloWGS: Reconstructing subclonal composition and evolution from whole-genome sequencing of tumors. Genome Biology. 2015; 16(1):1–20. https://doi.org/10.1186/s13059-015-0602-8
9. Andor N, Harness JV, Müller S, Mewes HW, Petritsch C. Expands: Expanding ploidy and allele frequency on nested subpopulations. Bioinformatics. 2014; 30(1):50–60. https://doi.org/10.1093/bioinformatics/btt622 PMID: 24177718
10. Jahn K, Kuipers J, Beerenwinkel N. Tree inference for single-cell data. Genome Biology. 2016; 17(1):86. https://doi.org/10.1186/s13059-016-0936-x PMID: 27149953
11. Davis A, Navin NE. Computing tumor trees from single cells. Genome Biology. 2016; 17(1):1–4. https://doi.org/10.1186/s13059-016-0987-z
12. Ciccolella S, Soto Gomez M, Patterson M, Della Vedova G, Hajirasouliha I, Bonizzoni P. Inferring Cancer Progression from Single-cell Sequencing while Allowing Mutation Losses. bioRxiv. 2018; p. 268243.
13. Morris LGT, Riaz N, Desrichard A, Şenbabaoğlu Y, Hakimi AA, Makarov V, et al. Pan-cancer analysis of intratumor heterogeneity as a prognostic determinant of survival. Oncotarget. 2016; 7(9). https://doi.org/10.18632/oncotarget.7067
14. Andor N, Graham TA, Jansen M, Xia LC, Aktipis CA, Petritsch C, et al. Pan-cancer analysis of the extent and consequences of intratumor heterogeneity. Nature Medicine. 2016; 22(1):105–113. https://doi.org/10.1038/nm.3984 PMID: 26618723
15. McGranahan N, Swanton C. Clonal Heterogeneity and Tumor Evolution: Past, Present, and the Future. Cell. 2017; 168(4):613–628. https://doi.org/10.1016/j.cell.2017.01.018 PMID: 28187284
16. Dentro SC, Leshchiner I, Haase K, Tarabichi M, Wintersinger J, Deshwar AG, et al. Portraits of genetic intra-tumour heterogeneity and subclonal selection across cancer types. bioRxiv. 2018; p. 312041.
17. Dagogo-Jack I, Shaw AT. Tumour heterogeneity and resistance to cancer therapies. Nature Reviews Clinical Oncology. 2018; 15(2):81–94. https://doi.org/10.1038/nrclinonc.2017.166 PMID: 29115304
18. Nik-Zainal S, Van Loo P, Wedge DC, Alexandrov LB, Greenman CD, Lau KW, et al. The life history of 21 breast cancers. Cell. 2012; 149(5):994–1007. https://doi.org/10.1016/j.cell.2012.04.023 PMID: 22608083
19. Gerlinger M, Horswell S, Larkin J, Rowan AJ, Salm MP, Varela I, et al. Genomic architecture and evolution of clear cell renal cell carcinomas defined by multiregion sequencing. Nature Genetics. 2014; 46(3):225–233. https://doi.org/10.1038/ng.2891 PMID: 24487277
20. Navin NE. Tumor evolution in response to chemotherapy: Phenotype versus genotype. Cell Reports. 2014; 6(3):417–419. https://doi.org/10.1016/j.celrep.2014.01.035 PMID: 24529750
21. Navin N, Kendall J, Troge J, Andrews P, Rodgers L, McIndoo J, et al. Tumour evolution inferred by single-cell sequencing. Nature. 2011; 472(7341):90–95. https://doi.org/10.1038/nature09807 PMID: 21399628
22. Noorbakhsh J, Kim H, Namburi S, Chuang JH. Distribution-based measures of tumor heterogeneity are sensitive to mutation calling and lack strong clinical predictive power. Scientific Reports. 2018; 8(1):1–12. https://doi.org/10.1038/s41598-018-29154-7
23. Colaprico A, Silva TC, Olsen C, Garofano L, Cava C, Garolini D, et al. TCGAbiolinks: An R/Bioconductor package for integrative analysis of TCGA data. Nucleic Acids Research. 2016; 44(8):e71. https://doi.org/10.1093/nar/gkv1507 PMID: 26704973
24. Gao J, Aksoy BA, Dogrusoz U, Dresdner G, Gross B, Sumer SO, et al. Integrative analysis of complex cancer genomics and clinical profiles using the cBioPortal. Science Signaling. 2013; 6(269). https://doi.org/10.1126/scisignal.2004088
25. Martincorena I, Raine KM, Gerstung M, Dawson KJ, Haase K, Van Loo P, et al. Universal Patterns of Selection in Cancer and Somatic Tissues. Cell. 2017; 171(5):1029–1041.e21. https://doi.org/10.1016/j.cell.2017.09.042 PMID: 29056346

26. Forbes SA, Beare D, Boutselakis H, Bamford S, Bindal N, Tate J, et al. COSMIC: Somatic cancer genetics at high-resolution. Nucleic Acids Research. 2017; 45(D1):D777–D783. https://doi.org/10.1093/nar/gkw1121 PMID: 27899578

27. Gao B, Huang Q, Baudis M. segment_liftover: a Python tool to convert segments between genome assemblies. F1000Research. 2018; 7:319. https://doi.org/10.12688/f1000research.14148.1 PMID: 29946440

28. Aran D, Sirota M, Butte AJ. Systematic pan-cancer analysis of tumour purity. Nature communications. 2015; 6:8971. https://doi.org/10.1038/ncomms9971 PMID: 26634437

29. Carter SL, Cibulskis K, Helman E, McKenna A, Shen H, Zack T, et al. Absolute quantification of somatic DNA alterations in human cancer. Nature Biotechnology. 2012; 30(5):413–421. https://doi.org/10.1038/nbt.2203 PMID: 22544022

30. Mroz EA, Rocco JW. MATH, a novel measure of intratumor genetic heterogeneity, is high in poor-outcome classes of head and neck squamous cell carcinoma. Oral Oncology. 2013; 49(3):211–215. https://doi.org/10.1016/j.oraloncology.2012.09.007 PMID: 23079694

31. Davidson-Pilon C, Kalderstam J, Zivich P, Kuhn B, Fiore-Gartland A, Moneda L, et al. CamDavidsonPilon/lifelines: v0.20.0; 2019. Available from: https://doi.org/10.5281/zenodo.2584900.

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

33. Van Belle V, Pelckmans K, Van Huffel S, Suykens JAK. Support vector methods for survival analysis: A comparison between ranking and regression approaches. Artificial Intelligence in Medicine. 2011; 53(2):107–118. https://doi.org/10.1016/j.artmed.2011.06.006 PMID: 21821401

34. Pölsterl S, Gupta P, Wang L, Conjeti S, Katouzian A, Navab N. Heterogeneous ensembles for predicting survival of metastatic, castrate-resistant prostate cancer patients. F1000Research. 2017; 5:2676. https://doi.org/10.12688/f1000research.8231.2

35. Schröder MS, Culhane AC, Quackenbush J, Haibe-Kains B. survcomp: An R/Bioconductor package for performance assessment and comparison of survival models. Bioinformatics. 2011; 27(22):3206–3208. https://doi.org/10.1093/bioinformatics/btr511 PMID: 21903630

36. Pencina MJ, D'Agostino RB. OverallC as a measure of discrimination in survival analysis: model specific population value and confidence interval estimation. Statistics in Medicine. 2004; 23(13):2109–2123. https://doi.org/10.1002/sim.1802 PMID: 15211606

37. Love MI, Huber W, Anders S. Moderated estimation of fold change and dispersion for RNA-seq data with DESeq2. Genome Biology. 2014; 15(12). https://doi.org/10.1186/s13059-014-0550-8

38. Bindea G, Mlecnik B, Tosolini M, Kirilovsky A, Waldner M, Obenauf AC, et al. Spatiotemporal dynamics of intratumoral immune cells reveal the immune landscape in human cancer. Immunity. 2013; 39(4):782–795. https://doi.org/10.1016/j.immuni.2013.10.003 PMID: 24138885

39. Zerbino DR, Achuthan P, Akanni W, Amode MR, Barrell D, Bhai J, et al. Ensembl 2018. Nucleic Acids Research. 2018; 46(D1):D754–D761. https://doi.org/10.1093/nar/gkx1098 PMID: 29155950

40. Salcedo A, Tarabichi M, Espiritu SMG, Deshwar AG, Buchanan A, Lalansingh CM, et al. Creating Standards for Evaluating Tumour Subclonal Reconstruction. 2018.

41. Rosenberg A, Hirschberg J. 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). 2007; 1(June):410–420.

42. Malikic S, Jahn K, Kuipers J, Sahinalp SC, Beerenwinkel N. Integrative inference of subclonal tumour evolution from single-cell and bulk sequencing data. Nature Communications. 2019; 10(1):1–12. https://doi.org/10.1038/s41467-019-10737-5

43. Gawad C, Koh W, Quake SR. 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. 2014; 111(50):17947–17952. https://doi.org/10.1073/pnas.1420822111 PMID: 25425670

44. Wang Y, Waters J, Leung ML, Unruh A, Roh W, Shi X, et al. Clonal evolution in breast cancer revealed by single nucleus genome sequencing. Nature. 2014; 512(7513):155–160. https://doi.org/10.1038/nature13600 PMID: 25079324

45. Leung ML, Davis A, Gao R, Casasent A, Wang Y, Sei E, et al. Single-cell DNA sequencing reveals a latedissemination model in metastatic colorectal cancer. Genome Research. 2017; 27(8):1287–1299. https://doi.org/10.1101/gr.209973.116 PMID: 28546418

46. Li H, Durbin R. Fast and accurate short read alignment with Burrows—Wheeler transform. 2009; 25(14):1754–1760.

47. Institute B. Picard Tools;. http://broadinstitute.github.io/picard/.

48. Quinlan AR, Hall IM. BEDTools: a flexible suite of utilities for comparing genomic features. 2010; 26 (6):841–842.

49. Li H, Handsaker B, Wysoker A, Fennell T, Ruan J, Homer N, et al. The Sequence Alignment / Map format and SAMtools. 2009; 25(16):2078–2079.

50. Mckenna A, Hanna M, Banks E, Sivachenko A, Cibulskis K, Kernytsky A, et al. The Genome Analysis Toolkit: A MapReduce framework for analyzing next-generation DNA sequencing data. 2010; p. 1297–1303.

51. Cibulskis K, Lawrence MS, Carter SL, Sivachenko A, Jaffe D, Sougnez C, et al. Sensitive detection of somatic point mutations in impure and heterogeneous cancer samples. Nature Biotechnology. 2013; 31(3):213–219. https://doi.org/10.1038/nbt.2514 PMID: 23396013

52. Shen R, Seshan VE. FACETS: allele-specific copy number and clonal heterogeneity analysis tool for high-throughput DNA sequencing. 2016;(8):1–9.

53. Pereira B, Chin SF, Rueda OM, Vollan HKM, Provenzano E, Bardwell HA, et al. The somatic mutation profiles of 2,433 breast cancers refines their genomic and transcriptomic landscapes. Nature Communications. 2016; 7(May):11479. https://doi.org/10.1038/ncomms11479 PMID: 27161491

54. Karn T, Jiang T, Hatzis C, Sänger N, El-Balat A, Rody A, et al. Association Between Genomic Metrics and Immune Infiltration in Triple-Negative Breast Cancer. JAMA Oncology. 2017; 3(12):1707–1711. https://doi.org/10.1001/jamaoncol.2017.2140 PMID: 28750120

55. Venkatesan S, Swanton C. Tumor Evolutionary Principles: How Intratumor Heterogeneity Influences Cancer Treatment and Outcome. American Society of Clinical Oncology Educational Book. 2016; 36: e141–e149. https://doi.org/10.1200/EDBK_158930

56. Eaton J, Wang J, Schwartz R. Deconvolution and phylogeny inference of structural variations in tumor genomic samples. Bioinformatics. 2018; 34(13):i357–i365. https://doi.org/10.1093/bioinformatics/bty270 PMID: 29950001

57. Bhandari V, Liu LY, Salcedo A, Espiritu SMG, Morris QD, Boutros PC. The Inter and Intra-Tumoural Heterogeneity of Subclonal Reconstruction. bioRxiv. 2018.

58. Turajlic S, Swanton C. TRACERx Renal: tracking renal cancer evolution through therapy. Nature Reviews Urology. 2017; 14(10):575–576. https://doi.org/10.1038/nrurol.2017.112

59. Kim C, Gao R, Sei E, Brandt R, Hartman J, Hatschek T, et al. Chemoresistance Evolution in Triple-Negative Breast Cancer Delineated by Single-Cell Sequencing. Cell. 2018; 173(4):879–893.e13. https://doi.org/10.1016/j.cell.2018.03.041 PMID: 29681456

60. Shi W, Ng CKY, Lim RS, Jiang T, Kumar S, Li X, et al. Reliability of Whole-Exome Sequencing for Assessing Intratumor Genetic Heterogeneity; 2018. Available from: https://doi.org/10.1016/j.celrep.2018.10.046.

61. Zhou H, Neelakantan D, Ford HL. Clonal cooperativity in heterogenous cancers. Seminars in Cell & Developmental Biology. 2017; 64:79–89. https://doi.org/10.1016/j.semcdb.2016.08.028

62. McGranahan N, Favero F, De Bruin EC, Birkbak NJ, Szallasi Z, Swanton C. Clonal status of actionable driver events and the timing of mutational processes in cancer evolution. Science Translational Medicine. 2015; 7(283):283ra54–283ra54. https://doi.org/10.1126/scitranslmed.aaa1408 PMID: 25877892

63. Keats JJ, Chesi M, Egan JB, Garbitt VM, Palmer SE, Braggio E, et al. Clonal competition with alternating dominance in multiple myeloma. Blood. 2012; 120(5):1067–1076. https://doi.org/10.1182/blood-2012-01-405985 PMID: 22498740

64. Scott J, Marusyk A. Somatic clonal evolution: A selection-centric perspective. Biochimica et Biophysica Acta—Reviews on Cancer. 2017; 1867(2):139–150. https://doi.org/10.1016/j.bbcan.2017.01.006 PMID: 28161395

65. Cross WCH, Graham TA, Wright NA. New paradigms in clonal evolution: punctuated equilibrium in cancer. Journal of Pathology. 2016; 240(2):126–136. https://doi.org/10.1002/path.4757 PMID: 27282810

66. Sottoriva A, Barnes CP, Graham TA. Catch my drift? Making sense of genomic intra-tumour heterogeneity. Biochimica et Biophysica Acta (BBA)—Reviews on Cancer. 2017; 1867(2):95–100. https://doi.org/10.1016/j.bbcan.2016.12.003

67. Maley CC, Aktipis A, Graham TA, Sottoriva A, Boddy AM, Janiszewska M, et al. Classifying the evolutionary and ecological features of neoplasms. Nature Reviews Cancer. 2017; 17(10):605–619. https://doi.org/10.1038/nrc.2017.69 PMID: 28912577

68. Safonov A, Jiang T, Bianchini G, Győrffy B, Karn T, Hatzis C, et al. Immune Gene Expression Is Associated with Genomic Aberrations in Breast Cancer. Cancer Research. 2017; 77(12):3317–3324. https://doi.org/10.1158/0008-5472.CAN-16-3478 PMID: 28428277

69. Caravagna G, Heide T, Williams M, Zapata L, Nichol D, Chkhaidze K, et al. Model-based tumor subclonal reconstruction. 2019; p. 1–31.