GeTallele: a method for integrative analysis and visualization of DNA and RNA allele frequencies

Piotr Słowiński1,2,*, Muzi Li2,#, Paula Restrepo2,3, Nawaf Alomran2, Liam F. Spurr2,4,5, Christian Miller2, Krasimira Tsaneva-Atanasova1,6, and Anelia Horvath2,7,8

1Department of Mathematics, College of Engineering, Mathematics and Physical Sciences, Living Systems Institute, Translational Research Exchange @ Exeter and EPSRC Centre for Predictive Modelling in Healthcare, University of Exeter, EX4 4QJ Exeter, UK, 2McCutcheon Genomics and Proteomics Center, School of Medicine and Health Sciences, The George Washington University, 20037 Washington, DC, USA, 3Department of Genetics & Genomics Sciences, Icahn School of Medicine at Mount Sinai, New York, NY 10029, 4Cancer Program, Broad Institute of MIT and Harvard, Cambridge, MA, 5Medical Oncology, Dana-Farber Cancer Institute, Boston, MA, 6Department of Bioinformatics and Mathematical Modelling, Institute of Biophysics and Biomedical Engineering, Bulgarian Academy of Sciences, 1113 Sofia, Bulgaria, 7Department of Pharmacology and Physiology, School of Medicine and Health Sciences, The George Washington University, 20037 Washington, DC, USA, 8Department of Biochemistry and Molecular Medicine, School of Medicine and Health Sciences, The George Washington University, 20037 Washington, DC, USA.

* To whom correspondence should be addressed.
# Equal contribution

Abstract

Background: Asymmetric allele expression typically indicates functional and/or structural features associated with the underlying genetic variants. When integrated, RNA and DNA allele frequencies can reveal patterns characteristic of a wide-range of biological traits, including ploidy changes, genome admixture, allele-specific expression and gene-dosage transcriptional response.

Results: To assess RNA and DNA allele frequencies from matched sequencing datasets, we introduce a method for generating model distributions of variant allele frequencies (VAF) with a given variant read probability. In contrast to other methods, based on whole sequences or single SNV, proposed methodology uses continuous multi-SNV genomic regions. The methodology is implemented in a GeTallele toolbox that provides a suite of functions for integrative analysis, statistical assessment and visualization of Genome and Transcriptome allele frequencies. Using model VAF probabilities, GeTallele allows estimation and comparison of variant read probabilities (VAF distributions) in a sequencing dataset. We demonstrate this functionality across cancer DNA and RNA sequencing datasets.

Conclusion: Based on our evaluation, variant read probabilities can serve as a dependable indicator to assess gene and chromosomal allele asymmetries and to aid calls of genomic events in matched sequencing RNA and DNA datasets.

Contact: P.M.Slowinski@exeter.ac.uk

1 Introduction

RNA and DNA carry and present genetic variation in related yet distinct manners; the differences being informative of DNA and structural traits. In diploid organisms, an important measure of genetic variation is the allele frequency, which can be measured from both genome (DNA) and transcriptome (RNA) sequencing data (encoded and expressed allele frequency, respectively). Differential DNA-RNA allele frequencies are associated with a variety of biological processes, such as genome admixture or allele-specific transcriptional regulation (Ferreira, et al., 2016; Ha, et al., 2012; Han, et al., 2015; Movassagh, et al., 2016; Shah, et al., 2012).

Most of the RNA-DNA allele comparisons from sequencing have been approached at nucleotide level, where it has proven to be highly informative for determining the alleles’ functionality (Ferreira, et al., 2016; Ha, et al., 2012; Han, et al., 2015; Macaulay, et al., 2016; Morin, et al., 2013; Movassagh, et al., 2016; Reuter, et al., 2016; Shah, et al., 2012; Shi, et al., 2016; Shlien, et al., 2016; The, et al., 2012; Yang, et al., 2016). Comparatively, integration of allele signals at the molecular level, as derived from linear DNA and RNA carriers, is less explored due to challenges presented by short sequencing length. The different molecular nature of RNA and DNA also leads to limited compatibility of the sequencing output.

Herein, we address some of the above challenges by employing a mathematical model to assess differences between RNA- and DNA- variant allele frequencies (VAF) at single nucleotide variant (SNV) positions in the genome. To do that, we introduce GeTallele: a toolbox that provides a suit of functions for integrative analysis and visualization of Genome (VAF genomic) and Transcriptome (VAF transcriptome) allele frequencies along
genes and chromosomes. Using VAF, GeTallele infers possible copy number alterations (CNAs), and, at positions of interest, mathematically and statistically compares VAF\textsubscript{RNA} and VAF\textsubscript{DNA} distributions. Furthermore, GeTallele supports visualization of the allele distribution at a desired resolution—from nucleotide to genome. In contrast to other CNA-modelling methods based on statistical models, GeTallele is based on a mechanistic model of VAF distributions.

2 Results

We demonstrate GeTallele’s functionality using sequencing DNA and RNA datasets from paired normal and tumour tissue obtained from 72 female patients with breast invasive carcinoma (BRCA) from The Cancer Genome Atlas (TCGA). Each dataset contains four matched sequencing sets: normal exome (Nex), normal transcriptome (Ntr), tumour exome (Tex), and tumour transcriptome (Ttr). The raw sequencing data were processed as previously described (Movassagh, et al., 2016) to generate the inputs for GetAllele.

The overall workflow of GetAllele is shown on Figure 1. As an input, GeTallele requires the absolute number of sequencing reads bearing the variant and reference nucleotide in each single-nucleotide variant (SNV) position across the four matched datasets. For each sample, to select SNV positions for analysis, we start with the high quality heterozygous SNV calls on the normal exome (Li, et al., 2009). In each of these positions we estimate the counts of the variant and reference reads (n\textsubscript{VAR} and n\textsubscript{REF}, respectively) across the 4 matching datasets, and retain for further analyses only positions covered by a minimum total (variant and reference) required unique sequencing reads across. This threshold is flexible and is required to ensure that only sufficiently covered positions will be analysed; it is set to 3 in the herein presented results.

From each of the 4 matched datasets GetAllele estimates VAF based on n\textsubscript{VAR} and n\textsubscript{REF} covering the positions of interest: VAF = n\textsubscript{VAR}/(n\textsubscript{VAR}+n\textsubscript{REF}). An example of genome-wide VAF values estimated from Tex, and their corresponding distribution of VAF values is shown in Figure 2.
Fig. 1. GeTallele and visualisation of VAF data. A Toolbox description. B Visualisation of the whole dataset on the level of genome using Circos plot (blue – normal exome, cyan – normal transcriptome, orange - tumour exome, yellow - tumour transcriptome). C - F show in details VAF_Tex and VAF_Ttr values of chromosome 1; C - F Visualization of the VAF values with fitted variant probability (vPR – see Section 3.1 and Figure 2) values at the level of chromosome (D), custom genome region (E) and gene (F), for the chromosome level shown also are CNA values (C). Panel D shows that there are two segments with different VAF distributions. Panel C shows that change in the CNA is concurrent with the change in the VAF distributions. Tex - tumour exome (orange); Ttr - tumour transcriptome (yellow).
Analysis of the VAF\textsubscript{RNA} and VAF\textsubscript{DNA} in the GetTallele is based on comparing probability of observing a given VAF value at various positions of interest. Estimation of the variant allele probability (v\textsubscript{PR}) is implemented using a data-driven mathematical model of distribution of the VAF values and is the core functionality of the GetTallele.

2.1 Estimation of variant probability v\textsubscript{PR}

2.1.1 Variant probability

GetAllele aggregates frequency counts of a VAF sample into variant probability, v\textsubscript{PR} – probability of observing a variant allele. The v\textsubscript{PR} is a single informative value indicative of copy number variation. The v\textsubscript{PR} is a parameter that describes the genomic event that through the sequencing process was transformed into a specific distribution of VAF values found in the signal. For example, in VAF\textsubscript{DNA} from a diploid genome, variant probability v\textsubscript{PR}=0.5 (meaning that both alleles are equally probable) corresponds to a true allelic ratio of 1:1 for heterozygote sites. For heterozygote sites in the DNA from diploid monoclonal sample, the corresponding tumour VAF\textsubscript{DNA} is expected to have the following interpretations: v\textsubscript{PR}=1 or v\textsubscript{PR}=0 corresponds to a monoallelic status resulting from a deletion, and v\textsubscript{PR}=0.8 (or 0.2), 0.75 (or 0.25), 0.67 (or 0.33) correspond to allele-specific tetra-, tri-, and duplication of the variant-bearing allele, respectively.

The v\textsubscript{PR} of the VAF\textsubscript{RNA} is interpreted as follows. In positions corresponding to DNA heterozygote sites, alleles not preferentially targeted by regulatory traits are expected to have expression rates with variant probability v\textsubscript{PR}=0.5, which (by default) scale with the DNA allele distribution. Differences between VAF\textsubscript{DNA} and VAF\textsubscript{RNA} values are observed in special cases of transcriptional regulation where one of the alleles is preferentially transcribed over the other. In the absence of allele-preferential transcription, VAF\textsubscript{DNA} and VAF\textsubscript{RNA} are anticipated to have similar v\textsubscript{PR} across both diploid (normal) and aneuploid (affected by CNAs) genomic regions. Consequently, VAF\textsubscript{DNA} and VAF\textsubscript{RNA} are expected to synchronously switch between allelic patterns along the chromosomes, with the switches indicating break points of DNA deletions or amplifications.

Since we observed that DNA and RNA signals have different distributions of total reads and also that the distributions of total reads vary between participants, the model VAF distributions are generated individually for each sequencing signal and each participant.

To estimate v\textsubscript{PR} in the signals, GetTallele first generates model VAF distributions and then uses the earth mover’s distance (EMD) (Kantorovich and Rubinstein, 1958; Levina and Bickel, 2001) to fit them to the data. To generate a model VAF distribution with a given variant probability, v\textsubscript{PR}, GetTallele, bootstraps 10000 values of the total reads (sum of the variant and reference reads; n\textsubscript{VAR} + n\textsubscript{REF}) from the analysed signal in the dataset. It then uses binomial pseudorandom number generator to get number of successes for given number of total reads and a given value of v\textsubscript{PR} (implemented in the Matlab function binornd). The v\textsubscript{PR} is the probability of success and generated number of successes is interpreted as an n\textsubscript{VAR}. Since the model v\textsubscript{PR} can take any value, it can correspond to a single genomic event as well as any combination of genomic events in any mixture of normal and tumour populations (See section 2.1.3).
The analysis presented in the paper uses 51 model VAF distributions with vPR values that vary from 0.5 to 1 with step (increment of) 0.01. The model VAF distributions are parametrized using only vPR ≥ 0.5, however, to generate them we use vPR and its symmetric counterpart 1−vPR. The process of generating model VAF distributions and examples of model and real VAF distributions with different values of vPR are illustrated in Fig. 3.

![Diagram](image)

Fig. 3. Model and real VAF distributions. A Description of the model used to generate model VAF distributions, Bin(n,p) stands for binomial distribution with parameters n (number of trials) and p (probability of success). B - F Model VAF distributions for different values of vPR. Panels B and F show additionally distributions of VAFTEX for the two windows shown in Figure 1D.

2.1.2 vPR estimation

Earth mover’s distance (EMD) is a metric for quantifying differences between probability distributions (Kantorovich and Rubinstein, 1958; Levina and Bickel, 2001) and in the case of univariate distributions it can be computed as:

$$\text{EMD} (PDF_1, PDF_2) = \int_2 |CDF_1(x) - CDF_2(x)| dx.$$
Here, PDF₁ and PDF₂ are two probability density functions and CDF₁ and CDF₂ are their respective cumulative distribution functions. Z is the support of the PDFs (i.e., set of all the possible values of the random variables described by them). Because VAFs are defined as simple fractions with values between 0 and 1, their support is given by a Farey sequence (Hardy, et al., 2008) of order n; n is the highest denominator in the sequence. For example, Farey sequence of order 2 is 0, 1/2, 1 and Farey sequence of order 3 is 0, 1/3, 1/2, 2/3, 1. GeTallele uses a Farey sequence of order 1000 as the support Z for all computations involving EMD.

To estimate \( \nu_{PR} \), GeTallele computes EMD between the distribution of the VAF values of each signal in the window and the 51 model VAF distributions (i.e., observed vs modelled VAF), the estimate is given by the \( \nu_{PR} \) of the model VAF distribution that is closest to the VAF distribution in the window. Examples of VAF distributions with fitted model VAF distributions are shown in Figure 3A and D. The dependence of the confidence intervals of the estimation on the number of VAF values in a window is illustrated in Fig. 4, which clearly demonstrates that the larger the number of VAFs in the chosen window the higher the accuracy of the estimate.

### Fig. 4. Confidence intervals for artificial samples with different numbers of VAFs.

Each confidence interval is based on estimation of \( \nu_{PR} \) in 1000 randomly generated samples with a fixed \( \nu_{PR} \) (True value). Light grey bar is 95% confidence interval (950 samples lay within this interval), dark grey bar is 50% confidence interval (500 samples lay within this interval), red dot is median value.

#### 2.1.3 \( \nu_{PR} \) values in mixtures of normal and tumour populations

Since the \( \nu_{PR} \) can take any value between 0.5 and 1 it can correspond to a single genomic event as well as any combination of genomic events in any mixture of normal and tumour populations. A mixture \( \nu_{PR} \) value that corresponds to a combination of genomic events can be computed using the following expression:

\[
\nu_{PR} = \frac{\sum_{\text{events}} \sum_{\text{events}} e_{\text{VAR}} \cdot p_{\text{PL}}}{\sum_{\text{events}} \sum_{\text{events}} e_{\text{VAR}} \cdot p_{\text{PL}} + \sum_{\text{events}} \sum_{\text{events}} e_{\text{REF}} \cdot p_{\text{PL}}}
\]

Where \( e_{\text{VAR}} \) and \( e_{\text{REF}} \) are the multiplicities of variant and reference alleles and \( p_{\text{PL}} \) is a proportion of one of the populations. For heterozygote sites \( e_{\text{VAR}} = 1 \) and \( e_{\text{REF}} = 1 \), for deletions \( e_{\text{VAR}} = 0 \) or \( e_{\text{REF}} = 0 \), for du-, tri- and
tetraplications eVAR or eREF can be equal to 2, 3 or 4, respectively. The sum of proportions pNL over the populations is equal 1. For example, for a mixture of 1 normal (N, pN=0.44) and 2 tumour populations (T1, pT1=0.39 and T2, pT2=0.17), T1 with deletion and T2 with deletion the mixture vPR value can be computed as follows:

\[
v_{PR} = \frac{p_N \cdot B + p_{T1} \cdot B + p_{T2} \cdot B}{p_N \cdot (A + B) + p_{T1} \cdot (0 + B) + p_{T2} \cdot (0 + B)}
\]

\[
= \frac{0.44 \cdot 1 + 0.39 \cdot 1 + 0.17 \cdot 1}{0.44 \cdot (1 + 1) + 0.39 \cdot (1 + 0) + 0.17 \cdot (1 + 0)} = 0.694.
\]

By comparing the vPR values estimated from data with possible mixture vPR values we propose to estimate sample purity and its clonal composition. To this end, we first generate a full set of proportions of all the population in the mixture with step (increment of) 0.01 and compute all the possible vPR values that each of the mixtures could produce. For step 0.01: two populations (1 tumour) give 99 proportions, three populations (2 tumours) give 4851 proportions, four populations (3 tumours) give 156849 proportions. The matrices with mixture vPR values for each proportion, vary from 2x2, for two populations with deletions, to 35x35 for four populations with all events up to tetraplications. Then, we run an exhaustive approximate search over all the matrices with mixture vPR values over all the proportions. The search is approximate because the estimated vPR values have limited accuracy and because we consider only discrete values of proportions. In the analysis we define a match between estimated and mixture vPR values if they differ <0.009. The search returns a large number of admissible mixtures that could produce the estimated vPR values. This process is illustrated in Fig. 5.

To represent the admissible mixtures graphically we use ternary plots. Ternary plot allows to illustrate composition of three components using just two dimensions. The composition, represented by ratios of the three components which sum to a constant, is depicted as point inside or on the edge of an equilateral triangle. If the point is on the edges, the composition has only two components. To facilitate interpretation of the ternary plots, we also plot the grid lines that are parallel to the sides of the triangle. These gridlines indicate the directions of constant ratios of the components. Along such direction the ratio of one of the components is fixed and only the other two ratios vary. Examples of visualisation of admissible mixtures on ternary plots are shown in Figs. 5 and 6.
**Estimated v_{PR} values:**

|       | 0.62 | 0.63 | 0.69 |

**Proportions and matrices with mixture v_{PR} values**

|       |       |       |       |
|-------|-------|-------|-------|
| T1    | T1:0  | 0.621 | 0.695 |
|       | T1:1  | 0.379 | 0.5   |
|       | T1:BB | 0.305 | 0.418 |
|       | T1:BB | 0.5   | 0.5   |

**Admissible mixtures**

The dot is at the edge because there are two populations, T2=0.

**Fig. 5. Mixtures admissible by the v_{PR} values estimated from data.** To uncover mixtures that could produce the three estimated v_{PR} values we perform an exhaustive approximate search of all the possible v_{PR} values produced by any mixture of the populations with a given set of genetic events. In each case we generate a full set of proportions with a given step (e.g., 0.01) and compute all the possible v_{PR} values that such a mixture could produce. In the illustrated cases: 2 populations (1 tumour) could produce the estimated v_{PR} values through a deletion (estimated v_{PR}=0.62 and v_{PR}=0.63) and via deletion of one allele and duplication of another (estimated v_{PR}=0.69); 3 populations (2 tumours) could produce the estimated v_{PR} values through a deletion in one of the tumour populations (estimated v_{PR}=0.62 and v_{PR}=0.63) and via deletion in both of the tumour populations (estimated v_{PR}=0.69). The 2 populations case admits a single mixture and the 3 populations allow 9 mixtures with similar compositions. The admissible mixtures are depicted on the ternary plots, red circle indicates solution corresponding to the presented matrix. We exclude mixture v_{PR} values that result from deletion of both the variant and reference alleles (empty fields in the matrices).

To facilitate analysis of the admissible mixtures returned by the search procedure we introduce mixture complexity. Mixture complexity is a measure that increases with number of populations as well as with variety of genetic events. From the simplest mixture of 1 normal and 1 tumour population in which only deletions are possible to a model with 1 normal and multiple tumour populations where each can have deletions, and any level of multiplications. In practice, we set the limit at 3 tumour populations and tetraplications. Mixture complexity helps to group and visualise admissible mixtures. Mixtures with higher complexity allow more possible v_{PR} values, meaning that it is easier to find the match with the estimated v_{PR} values and the number of admissible mixtures increases (see Figure 6). We, further, observe that proportion of normal population, p_{N}, increases with a number of clonal tumour populations included in the model mixture and that, generally, p_{N} stays constant with increasing variety of genetic events, for a fixed number of clonal tumour populations.
Fig. 6. Admissible mixtures for increasing mixture complexity. A shows admissible mixtures for 3 different mixtures with increasing complexity. The simplest mixture (mixture with the lowest number of components and the simplest set of genetic events) is shown within the grey frame. On each ternary plot, the admissible mixtures are indicated by grey dots. The green axis indicates proportion of the normal population (N), the yellow axis indicates proportion of the 1st tumour population (T1), the blue axis shows proportion of the 2nd tumour population (T2) or sum of the 2nd and 3rd tumour populations (T2+T3). B Schematic representation of increasing complexity of the mixture models. From a mixture of 1 normal and 1 tumour population in which only deletion is possible to a model with 1 normal and 3 tumour populations and each can have deletions, du-, tri- and tetraplications.

2.2 Analysis of RNA-DNA relationships

GetAllele is readily applicable to assess RNA-DNA relationships between normal and tumour sequencing signals derived from the same sample/individual (matched datasets). As a proof of concept, we assessed matched normal and tumour exome and transcriptome sequencing data of 72 breast carcinoma (BRCA) datasets with pre-assessed copy-number and genome admixture estimation acquired through TCGA (Supplementary Table 1). For these datasets, purity and genome admixture has been assessed using at least three of the following five approaches: ESTIMATE, ABSOLUTE, LUMP, IHC, and the Consensus Purity Estimation (CPE)(Aran, et al., 2015; Carter, et al., 2012; Katkovnik, et al., 2002; Pagès, et al., 2010; Yoshihara, et al., 2013; Zheng, et al., 2014). In addition, on the same datasets we applied THetA – a popular tool for assessing CNA and admixture from sequencing data (Oesper, et al., 2013; Oesper, et al., 2014).

2.2.1 Segmentation results

Segmentation of the data, based on the tumour exome signal, resulted in 2697 windows across the 72 datasets. We excluded from further analysis 294 windows where either tumour exome or transcriptome had $v_{PR}>0.58$ but their VAF distribution could not be differentiated from the model VAF distributions with $v_{PR}=0.5$ ($p>1e-5$, Kolmogorov Smirnov test, equivalent to Bonferroni FWER correction for 100000 comparisons). The 294 excluded windows correspond to 4% of the data in terms of number of base pairs in the windows and 4% of all the available data points; i.e. they are short and contain only few VAF values. In the remaining 2403 windows, we systematically examined the similarity between corresponding VAF (tumour exome), VAF (tumour transcriptome) and CNA. We obtained several distinct patterns of coordinated RNA-DNA allelic behaviour as well as correlations with CNA data.
In 60% of all analysed windows the distributions of VAF_{TEX} and VAF_{TTR} were statistically concordant (had p>1e-5, Kolmogorov Smirnov test), and in 40% they were statistically discordant (p<1e-5, Kolmogorov Smirnov test). In 2 windows VAF_{TEX} and VAF_{TTR} had the same v_{PR} but had statistically different distributions (p<1e-5, Kolmogorov Smirnov test), we consider such windows as concordant; Kolmogorov-Smirnov test is very sensitive for differences between distributions, v_{PR} fitting is more robust. In the vast majority of the discordant windows v_{PR} of the VAF_{TTR}, v_{PR,TTR} was higher than v_{PR} of the VAF_{TEX}, v_{PR,TEX} (only in 21 out of 959 discordant windows v_{PR,TTR} was lower than v_{PR,TEX}).

2.2.2 Concurrence of segmentation based on DNA and RNA

We next analysed the concurrence between windows resulting from independent segmentations of the dataset based on the tumour exome and transcriptome signals in the datasets (2697 and 3605 windows, respectively, across all the samples). We first assessed chromosome-wise alignment of the start and end points of the windows. In 45% of the chromosomes both VAF_{TEX} and VAF_{TTR} signals produce a single window that contains the whole chromosome. In 33% of chromosomes both signals produced multiple windows. These windows are well aligned, with 90% of the break-points within 7% difference in terms of the number of data points in the chromosome (Q50=0.02%, Q75=2% of data points in the chromosome). The probability of observing such an alignment by chance is smaller than p=1e-5 (100,000 bootstrap samples with breaking points assigned randomly in all the individual chromosomes where both signals produced multiple windows).

In 22% of the chromosome windows based on VAF_{TEX} and VAF_{TTR} the signals were positionally discordant – one signal produced a single window containing whole chromosome while the other produced multiple windows.

To compare the v_{PR} values in the 55% of chromosomes where at least one signal produced more than one window, we computed chromosome-wise mean absolute error (MAE) between the v_{PR} in two sets of windows. To account for different start and end points of the windows we interpolated the v_{PR} values (nearest neighbour interpolation) at each data point in the chromosome. We separately compared the v_{PR,TEX} and v_{PR,TTR} values. The alignment in terms of MAE is very good, v_{PR,TEX} agreed perfectly in 11% of the chromosomes and had the percentiles of MAE equal to Q50= 0.012, Q75= 0.022 and Q97.5= 0.047, while v_{PR,TTR} agreed perfectly in 8% but had slightly higher percentiles of MAE Q50= 0.019, Q75= 0.034 and Q97.5= 0.07. v_{PR,TEX} and v_{PR,TTR} values had MAE=0 simultaneously in 4% of the chromosomes. Probability of observing such values of MAE by chance is smaller than p=1e-3 (1000 random assignments of v_{PR,TEX} and v_{PR,TTR} values to windows in the 873 chromosomes where at least one signal had more than one window). It is noteworthy that MAE Q97.5<0.07 is comparable with the confidence interval of single v_{PR} estimate based on 50 VAF values. In other words, both signals in a sample (Tex and Ttr) give very similar results in terms of windows’ segmentation and estimated values of the v_{PR}. Albeit, segmentation of VAF_{TTR} generates a higher number of windows. The higher number of VAF_{TTR} windows indicates that transcriptional regulation occurs at a smaller scale than allelic status changes in DNA. Figure 7 shows examples of concurrence between windows based on VAF_{TEX} and VAF_{TTR} signals in a positionally concordant chromosome (both signals produced multiple windows).
2.2.3 Correlation between \( v_{PR} \) and CNA

Finally, we assess the correlations between \( v_{PR} \) and CNA in the individual datasets. We separately computed correlations for deletions and amplifications. In order to separate deletions and amplifications, for each data set we found \( \text{CNA}_{\text{MIN}} \), value of the CNA in the range -0.3 to 0.3 that had the smallest corresponding \( v_{PR,TEX} \). To account for observed variability of the CNA values near the \( \text{CNA}_{\text{MIN}} \), we set the threshold for amplifications to \( \text{CNA}_A = \text{CNA}_{\text{MIN}} - 0.05 \), and for deletions we set it to \( \text{CNA}_D = \text{CNA}_{\text{MIN}} + 0.05 \).

For \( \text{VAF}_{\text{TEX}} \) we observed significant correlations with negative trend between \( v_{PR,TEX} \) and \( \text{CNA} \leq \text{CNA}_D \) in 57 datasets and with positive trend between \( v_{PR,TEX} \) and \( \text{CNA} \geq \text{CNA}_A \) in 39 datasets (\( p_{FDR} < 0.05 \), Pearson’s correlation with Benjamini-Hochberg multiple comparison correction for 72 samples). For \( \text{VAF}_{\text{TTR}} \) we observed significant correlations with negative trend between \( v_{PR,TTR} \) and \( \text{CNA} \leq \text{CNA}_D \) in 62 datasets and with positive trend between \( v_{PR,TTR} \) and \( \text{CNA} \geq \text{CNA}_A \) in 33 datasets (\( p_{FDR} < 0.05 \), Pearson correlation with Benjamini-Hochberg correction). These correlations indicate that although \( v_{PR,TEX} \) as well as \( v_{PR,TTR} \) values capture information contained in CNA, they do not differentiate between positive and negative values of the CNA.
normal exome estimate et al., 2010; Yoshihara, et al., 2013; Zheng, et al., 2014). To observe that distributions of the VAF (with median of 0.18) between VR,TEX values are noisy, only some correlations are statistically significant. PR,TTR values (circles) follow closely the VR,TEX (squares) indicating concordance of the VAF,TEX and VAF,TTR distributions. Only correlations for CNA ≤ CNA0 are statistically significant.

Figure 8 shows four typical patterns of correlation between the CNA and VR values observed in the data. In Figure 8A all the values of CNA are close to CNA,MN. In Figure 8B the relationship between CNA and VR is noisy, only correlation between VR,TTR and CNA ≤ CNA0 are statistically significant (rTEX,CNA,D = -0.29, pFDR = 0.063; rTEX,CNA,A = -0.38, pFDR = 0.012; rTEX,CNA,TEX = 0.14, pFDR = 0.58; rTEX,CNA,A = 0.19, pFDR = 0.47; Pearson’s correlation with Benjamini Hochberg multiple comparison correction for 72 samples). In Figure 8C all the correlations are statistically significant, VR,TEX values (circles) follow closely the VR,TEX (squares) indicating that in most of the windows distributions of the VAF,TEX and VAF,TTR are concordant (rTEX,CNA,D = -0.91, pFDR<1e-10; rTEX,CNA,D = -0.96, pFDR<1e-10; rTEX,CNA,A = 0.92, pFDR<1e-10; rTEX,CNA,TEX = 0.95, pFDR<1e-10). In Figure 8D correlations between VR,TEX, VR,TTR and CNA ≤ CNA0 are statistically significant, but there is a big difference (with median of 0.18) between VR,TEX and VR,TTR values, indicating that in most of the windows the distributions of the VAF,TEX and VAF,TTR in this dataset are discordant (rTEX,CNA,D = -0.44, pFDR = 0.047; rTEX,CNA,D = -0.64, pFDR = 0.0017; rTEX,CNA,A = 0.44 pFDR = 0.16; rTEX,CNA,TEX = 0.28, pFDR = 0.41). In many of the datasets we observe that the VR,TTR values are higher than the corresponding VR,TEX values (median VR,TTR-VR,TEX = 0.03). Correlations between VR and CNA in all datasets are shown in the Supplementary Figure 1.

2.3 VR based purity estimation
To demonstrate a practical application of estimating the admissible mixtures by means of comparing the estimated VR,TEX values with the mixture VR values we next use them to estimate purity of the samples. To this end we compared the VR based purity (VBP) estimates with ESTIMATE, ABSOLUTE, LUMP, IHC, and the Consensus Purity Estimation (CPE)(Arora, et al., 2015; Carter, et al., 2012; Katkovnik, et al., 2002; Pagès, et al., 2010; Yoshihara, et al., 2013; Zheng, et al., 2014).

To obtain the VBP estimate we used VR,TEX values. We, first, selected the VR,TEX values that: 1. are estimated with high confidence, i.e. are based on at least 50 data points; 2. are most likely heterozygous in normal exome, i.e., have a corresponding VR value in normal exome VR,NEX <0.59; 3. most likely have VR,TEX >
0.5, i.e., their p-value for comparison with $v_{PR,TEX} = 0.5$ is very small $p < 1e^{-5}$ (Kolmogorov-Smirnov test). Next, we used the selected $v_{PR,TEX}$ values to find all admissible mixtures (with 1 to 3 tumour populations and allowing for all events, from deletions to tetraplications). To estimate the VBP, out of all the admissible mixtures we chose these with lowest mixture complexity and among these mixtures we take one with the highest $p_N$ (proportion of the normal population). The VBP, percentage of tumour populations in the sample, is then given as $1 - p_N$. Such approach provides rather conservative estimates of VBP (the smallest $1 - p_N$). However, GetAllele can be extended to offer alternative methods of employing the admissible mixtures to estimate VBP. Development, analysis and comparison of alternative VBP estimation methods is beyond scope of the current paper.

Figure 9A shows violin plots of all considered $1 - p_N$ values and (x) indicates the smallest value taken as a VBP estimate. In two of the datasets we could not estimate the purity due to lack of suitable $v_{PR,TEX}$ values. The VBP estimates show the best agreement with ABSOLUTE method ($r = 0.76$, $p < 3.4e^{-14}$, Pearson’s correlation, Fig. 9B2). We suppose that this is because the ABSOLUTE method is based on copy number distributions, and our analysis (Section 2.2.3) revealed high correlations between the CNAs and $v_{PR,TEX}$ values. Similarly to the ABSOLUTE method, VBP estimates are generally lower than the other purity estimates (ESTIMATE, LUMP, IHC, CPE); see Fig. 9B1-B5.

The approach presented in this section differs from other methods for inferring genomic mixture composition (e.g. SciClone (Miller, et al., 2014), PyClone (Roth, et al., 2014) or TPES (Locallo, et al., 2019)) in that it is based on changes of the alleles multiplicity estimated along continuous multi-SNV genomic regions. The $v_{PR,TEX}$ values used to estimate admissible mixtures are based on windows with at least 50 VAF values which extend over millions of base pairs. In this way, the presented method is complementary to the SciClone and PyClone methods that use VAFs of somatic mutations.
Fig. 9. Illustration of purity estimation based on model mixtures and \( v_{PR} \). Comparison of the Estimate (EST), Absolute (ABS), LUMP, IHC, and the consensus purity estimate (CPE) methods with \( v_{PR} \)-based purity (VBP). A violin plots show distributions of purity based on all the admissible proportions of the normal population \( x \) indicates the lowest value selected as the most conservative estimate; colours corresponding to the different methods are indicated in B1-5. B1-5 Correlation of the VBP with individual methods; coloured line indicates best fit linear trend. C Matrix showing significant \( p<0.05 \) Pearson’s correlation coefficients between all tested methods.

### Discussion

Potential for integrative analysis of RNA and DNA sequence data is facilitated by the growing availability of RNA and DNA sequencing datasets and by the technological advances now enabling simultaneous RNA and DNA sequencing from the same source (Macaulay, et al., 2016; Reuter, et al., 2016; The, et al., 2012). However, integrative RNA and DNA analyses are challenged by limited compatibility between RNA and DNA datasets and high technical variance of the sequencing-produced signals. Our approach – GeTallele – addresses the compatibility by restricting the analyses to confidently co-covered DNA and RNA regions, and the high variability - through quantification of differences between the DNA and RNA signals.

In contrast to other methods based on statistical modelling, GeTallele is based on a mechanistic model of VAF distributions that depends on the distribution of total reads (extracted from data) and the \( v_{PR} \) parameter. The brute-force simplicity and transparency of the presented methodology is one of its biggest advantages.
Additionally, in contrast to other methods, to analyse sequencing data GeTallele uses data segments that include multiple adjacent SNVs. The proposed mechanistic model indicates that due to probabilistic nature of the reads that estimates of genomic events based on continuous multi-SNV regions are intrinsically more robust than estimates based on a single SNV. Finally, GeTallele offers a unified pipeline to evaluate and visualise DNA-RNA whereas many of the other methods cited herein are used in conjunction with each other to achieve a similar result.

Using GeTallele, we detected several relationships between DNA-RNA allele frequencies and biological processes. First, in chromosomes affected by deletions and amplifications, VAF_{RNA} and VAF_{DNA} showed highly concordant breakpoint calls. This indicates that VAF_{RNA} alone can serve as preliminary indicator for deletions and amplifications, which can facilitate the applications of RNA-sequencing analysis on the large and constantly growing collections of transcriptome sequencing data. Second, we showcased that v_{Pr,TEX} can be used to estimate sample composition (in terms of proportions of normal and tumour populations) and as a result its purity. Once the mixture composition is decided by the user, GeTallele allow to further interrogate genetic events in each population at a specific data segment.

4 Conclusions

Based on our results, variant probability v_{Pr} can serve as a dependable indicator to assess gene and chromosomal allele asymmetries and to aid calls of genomic events in matched sequencing RNA and DNA datasets. Methods for estimating and analysing v_{Pr} values are implemented in a GeTallele toolbox. GeTallele provides a singular suit of functions for integrative analysis, statistical assessment and visualization of the observed patterns at desired resolution, including chromosome, gene, or custom genome region.

5 Methods

5.1 Samples

The GeTallele was developed using sequencing datasets from paired normal and tumour tissue obtained from 72 female patients with breast invasive carcinoma (BRCA) from The Cancer Genome Atlas (TCGA). Each of the 72 datasets contains four matched sequencing datasets: normal exome (Nex), normal transcriptome (Ntr), tumour exome (Tex), and tumour transcriptome (Ttr). In addition, we required each tumour sample to have at least three of the following five purity estimates - Estimate, Absolute, LUMP, IHC, and the consensus purity estimate (CPE), (Supplementary Table 1). Finally, each sample was required to have CNA estimation (genomic segment means based on Genome-Wide-SNPv6 hybridization array) (Aran, et al., 2015; Carter, et al., 2012; Katkovnik, et al., 2002; Pagès, et al., 2010; Yoshihara, et al., 2013; Zheng, et al., 2014).

5.2 Data processing

All datasets were generated through paired-end sequencing on an Illumina HiSeq platform. The human genome reference (hg38)-aligned sequencing reads (Binary Alignment Maps, .bams) were downloaded from the Genomic Data Commons Data Portal (https://portal.gdc.cancer.gov/) and processed downstream through an in-house pipeline. After variant call (Li, 2011), the RNA and DNA alignments, together with the variant lists were processed through the read count module of the package RNA2DNAlign (Movassagh, et al., 2016), to produce variant and reference sequencing read counts for all the variant positions in all four sequencing signals (normal exome, normal transcriptome, tumour exome and tumour transcriptome). Selected read count assessments were visually examined using Integrative Genomics Viewer (Thorvaldsdóttir, et al., 2013).

5.3 Data segmentation

To analyse variant allele frequencies (VAF) at genome-wide level, GeTallele first divides the VAF dataset into a set of non-overlapping windows along the chromosomes. Segmentation of the dataset into windows is based on a sequencing signal chosen out of all the available datasets in the aggregated aligned VAF dataset (one out of four in the presented analysis).

To partition the data into the windows GeTallele uses a parametric global method, which detects the breakpoints in the signals using its mean, as implemented in the Matlab function findchangepts (Killick, et al., 2012; Lavielle, 2005) or in R (Killick and Eckley, 2014). In each window, the VAF values of the chosen signal have a mean that is different from the mean in the adjacent windows. Sensitivity of breakpoint detection can be controlled using parameter MinThreshold, in the presented analysis it was set to 0.2. For
segmentation and analysis (without loss of generality) we transform all the original VAF values to $VAF = |VAF - 0.5| + 0.5$.

### 5.4 Statistics

To test statistical significance, GeTallele uses parametric and non-parametric methods and statistical tests (Corder and Foreman, 2014; Hollander, et al., 2013). Namely, to compare distributions of the variant allele frequencies (VAF) we use Kolmogorov–Smirnov test (examples of VAF distributions are depicted in Figs. 2 and 3). To study concurrence of windows, we use permutation/ bootstrap tests. To test relations between $v_{PR}$ and copy number alterations (CNA) we use Pearson’s correlation coefficient.

To account for multiple comparisons, we set the probability for rejecting the null hypothesis at $p < 1 \times 10^{-5}$, which corresponds to Bonferroni (Dunn, 1961) family-wise error rate (FWER) correction against 100000 comparisons. We use a fixed value, rather than other approaches, to ensure better consistency and reproducibility of the results. Alternatively, we apply Benjamini and Hochberg (Benjamini and Hochberg, 1995) false discovery rate (FDR) correction with a probability of accepting false positive results $p_{FDR} < 0.05$. We specify the method used in the text when reporting the results.

### List of abbreviations

- BRCA – breast invasive carcinoma,
- CDF – cumulative distribution function,
- CNA – copy number alterations,
- CNA$_D$ – copy number alterations corresponding to deletions (see section 2.2.3),
- CNA$_A$ – copy number alterations corresponding to amplifications (see section 2.2.3),
- CPE – consensus purity estimate,
- DNA – genome,
- EMD – earth mover’s distance
- FWER – family-wise error rate,
- FDR – false discovery rate,
- MEA – mean absolute error
- Nex – normal exome,
- Ntr – normal transcriptome,
- $r_{TEX,CNA,D}$ – Pearson’s correlation coefficient between $v_{PR,TEX}$ and CNA$_D$
- $r_{TEX,CNA,A}$ – Pearson’s correlation coefficient between $v_{PR,TEX}$ and CNA$_A$
- $r_{TTR,CNA,D}$ – Pearson’s correlation coefficient between $v_{PR,TTR}$ and CNA$_D$
- $r_{TTR,CNA,A}$ – Pearson’s correlation coefficient between $v_{PR,TTR}$ and CNA$_A$
- $p_{FDR}$ – $p$-value after multiple comparisons Benjamini and Hochberg false discovery rate correction,
- PDF – probability density function,
- $Q_N$ (e.g. $Q_{50}$) – N-th percentile
- RNA – transcriptome
- SNV – single-nucleotide variant
- TCGA – the cancer genome atlas,
- Tex – tumour exome,
- Tr – tumour transcriptome,
- VAF – variant allele frequency,
- VAF$_{TEX}$ – variant allele frequency in tumour exome sequence,
- VAF$_{TTR}$ – variant allele frequency in tumour transcriptome sequence,
- VBP – $v_{PR}$ based purity,
- $v_{PR}$ – variant probability,
- $v_{PR,TEX}$ – variant probability estimated from tumour exome sequence,
- $v_{PR,TTR}$ – variant probability estimated from tumour transcriptome sequence,

### Declarations

- **Ethics approval and consent to participate:** Not applicable
- **Consent for publication:** Not applicable
- **Availability of data and materials:** The datasets used and/or analysed during the current study are available from the corresponding author on reasonable request. GeTallele is implemented as a Matlab
• Competing interests: The authors declare that they have no competing interests.
• Funding: This work was supported by McComick Genomic and Proteomic Center (MGPC), The George Washington University; [MGPC_PG2018 to AH]. Work of PS was generously supported by the Wellcome Trust Institutional Strategic Support Award [204909/Z/16/Z]. KTA gratefully acknowledges the financial support of the EPSRC via grant EP/N014391/1.
• Authors’ contributions: PS, ML, PR, NA, LFS, CM, KTA, AH conception and design of the work; ML data acquisition; PS data analysis; PS, ML, PR, LFS, KTA, AH interpretation of data; PS the creation of new software used in the work; PS, ML, PR, LFS, KTA, AH have drafted the work or substantively revised it; all authors approved the submitted version. All authors agreed both to be personally accountable for the author’s own contributions and to ensure that questions related to the accuracy or integrity of any part of the work, even ones in which the author was not personally involved, are appropriately investigated, resolved, and the resolution documented in the literature.
• Acknowledgements: Not applicable

References
Aran, D., Sirota, M. and Butte, A.J. Systematic pan-cancer analysis of tumour purity. Nature communications 2015;6:9971.
Benjamini, Y. and Hochberg, Y. Controlling the False Discovery Rate - a Practical and Powerful Approach to Multiple Testing. J Roy Stat Soc B Met 1995;57(1):289–300.
Carter, S., et al. Absolute quantification of somatic DNA alterations in human cancer. Nature biotechnology 2012;30(5):413–421.
Corder, G.W. and Foreman, D.J. Nonparametric Statistics. John Wiley & Sons; 2014.
Dunn, O.J. Multiple Comparisons among Means. J Am Stat Assoc 1961;56(293):52–64.
Ferreira, E., Shaw, D.M. and Oddo, S. Identification of learning-induced changes in protein networks in the hippocampi of a mouse model of Alzheimer’s disease. Translational psychiatry 2016;6(7):e849.
Ha, G., et al. Integrative analysis of genome-wide loss of heterozygosity and monoaecic expression at nucleotide resolution reveals disrupted pathways in triple-negative breast cancer. Genome research 2012;22(10):1995-2007.
Han, L., et al. Alternative applications for distinct RNA sequencing strategies. Briefings in bioinformatics 2011;12(2):178-188.
Hardy, G.H., et al. An introduction to the theory of numbers. Oxford ; New York: Oxford University Press; 2008.
Hollander, M., Wolfe, D.A. and Chicken, E. Nonparametric Statistical Methods. John Wiley & Sons; 2013.
Kantorovich, L.V. and Rubinstein, G.S. On a space of completely additive functions. Vestnik Leningrad. Univ 1958;13(7):52-69.
Levina, E. and Bickel, P. The Earth Mover’s distance is the Mallows distance: some insights from statistics. In, IEEE International Conference on Computer Vision 2001. p. 251-256.
Li, H. A statistical framework for SNP calling, mutation discovery, association mapping and population genetical parameter estimation from sequencing data. Bioinformatics 2011;27(21):2987-2993.
Li, H., et al. The Sequence Alignment/Map format and SAMtools. Bioinformatics 2009;25(16):2078-2079.
Locallo, A., et al. TPES: tumor purity estimation from SNVs. Bioinformatics 2019;35(21):4433-4435.
Macaulay, I.C., et al. Separation and parallel sequencing of the genomes and transcriptomes of single cells using G&T-seq. Nature protocols 2016;11(11):2081-2093.
Miller CA, White BS, Dees ND, Griffith M, Welch JS, 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
Morin, R.D., et al. Mutational and structural analysis of diffuse large B-cell lymphoma using whole-genome sequencing. Blood 2013;122(7):1256-1265.
Movassagh, M., et al. RNA2DNAAlign: nucleotide resolution allele alignments through quantitative assessment of RNA and DNA paired sequencing data. Nucleic acids research 2016;44(22):e161.
Newson, R. Parameters behind “nonparametric” statistics: Kendall’s tau,Somers’ D and median differences. Stata Journal 2002;2(1):45-64.
Oesper, L., Mahmoody, A. and Raphael, B.J. ThetA: inferring intra-tumor heterogeneity from high-throughput DNA sequencing data. Genome biology 2013;14(7):P80.
Oesper, L., Satas, G. and Raphael, B.J. Quantifying tumor heterogeneity in whole-genome and whole-exome sequencing data. Bioinformatics (Oxford, England) 2014;30(24):3532-3540.
Page, F., et al. Immune infiltration in human tumors: a prognostic factor that should not be ignored. Oncogene 2010;29(8):1093-1102.
Reuter, J.A., et al. Simul-seq: combined DNA and RNA sequencing for whole-genome and transcriptome profiling. Nature methods 2016;13(11):953-958.
Roth, A., Khattra, J., Yap, D. et al. PyClone: statistical inference of clonal population structure in cancer. Nat Methods 2014; 11, 396–398. https://doi.org/10.1038/nmeth.2883
Shah, S.P., et al. The clonal and mutational evolution spectrum of primary triple-negative breast cancers. Nature 2012;486(7403):395-399.
Shi, L., et al. Long-read sequencing and de novo assembly of a Chinese genome. Nature communications 2016;7:12065.
Shlien, A., et al. Direct Transcriptional Consequences of Somatic Mutation in Breast Cancer. Cell reports 2016;16(7):2032-2046.
The, E.P.C., et al. An integrated encyclopedia of DNA elements in the human genome. Nature 2012;489:57.
Thorvaldsdóttir, H., Robinson, J.T. and Mesirov, J.P. Integrative Genomics Viewer (IGV): high-performance genomics data visualization and exploration. Briefings in bioinformatics 2013;14(2):178-192.
Yang, S., et al. An Integrated Approach for RNA-seq Data Normalization. Cancer informatics 2016;15:129-141.
Yoshihara, K., et al. Inferring tumour purity and stromal and immune cell admixture from expression data. Nature communications 2013;4:2612.
Zheng, X., et al. MethylPurify: tumor purity deconvolution and differential methylation detection from single tumor DNA methylomes. Genome biology 2014;15(8):419.
Supplementary Fig. 1. Illustration of the correlations between vPR and CNA. Orange squares vPR,TEX, yellow circles vPR,TTR. Lines, least-squares fitted trends for significant correlations (orange correlation with vPR,TEX, yellow correlation with vPR,TTR). Black, vPR for CNA-min±0.05. Title format Number of the dataset: #sig number of significant correlations in the dataset.
# Supplementary Table 1. Datasets, signals and purity estimates.

| # | TCGA BRCA datasets | EST | ABS | LUMP | IHC | CPE |
|---|--------------------|-----|-----|------|-----|-----|
| 1 | 001_Nex_BRC_A_TCGA-BH-A1FC-11A_413b80f6-66cf-4992-804a-f045e38ce6ef | 0.7615 | 0.49 | 0.6693 | 0.9 | 0.6517 |
| 2 | 002_Nex_BRC_A_TCGA-BH-A0B5-11A_c724807c-d38d-4582-8238-833937b6acc | 0.7539 | 0.5 | 0.8944 | 0.575 | 0.6566 |
| 3 | 003_Nex_BRC_A_TCGA-BH-A0B1-11A_a99888bb-909a-4363-bf73-7050c17f063 | 0.7386 | 0.37 | 0.8453 | 0.7 | 0.7458 |
| 4 | 004_Nex_BRC_A_TCGA-E2-A158-11A_5ff36f7-8189-f6be-b02b-28464c451c39 | 0.9534 | NaN | NaN | 0.8 | 0.7799 |
| 5 | 005_Nex_BRC_A_TCGA-A7-A133-11A_bd7fe6f7-2713-4ded-a8ca-3c737bd918 | 0.909 | 0.83 | 0.9772 | 0.85 | 0.9184 |
| 6 | 006_Nex_BRC_A_TCGA-A208-11A_645bb78f-1942-4ece-973b-4a75f9626f55 | 0.5951 | 0.31 | 0.6877 | 0.6 | 0.5642 |
| 7 | 007_Nex_BRC_A_TCGA-A206-11A_d7ebe21b-40e1-99c7-72231b703a4 | 0.6835 | 0.25 | 0.667 | 0.6 | 0.6121 |
| 8 | 008_Nex_BRC_A_TCGA-BH-A0AY-11A_2eecc325-3973-483b-b35b-52aa5ea9bc | 0.6376 | 0.42 | NaN | 0.75 | 0.8077 |
| 9 | 009_Nex_BRC_A_TCGA-A18U-11A_b96d6c2-fa16-4d56-9bc3-416585f1d319 | 0.7499 | 0.68 | NaN | 0.75 | 0.8077 |
| 10 | 010_Nex_BRC_A_TCGA-A22F-11A_714ee11b-7b41-99d3-457883a07e0f | 0.5705 | NaN | 0.688 | 0.8 | 0.6667 |
| 11 | 011_Nex_BRC_A_TCGA-A22F-11A_45c217d2-22ed-418b-97fc-7104e11e4ac1 | 0.6814 | 0.4 | NaN | 0.5 | 0.5779 |
| 12 | 012_Nex_BRC_A_TCGA-A0BQ-11A_a5bddd-16-8cb-4787-b0e1-4c09f6790cc5 | 0.7944 | 0.4895 | 0.872 | 0.7278 |
| 13 | 013_Nex_BRC_A_TCGA-A0BQ-11A_27138831-1865-4a6a-bd70-5872592cb49 | 0.8571 | 0.87 | 0.9827 | 0.85 | 0.9342 |
| 14 | 014_Nex_BRC_A_TCGA-A0AU-11A_15483d36-ad2-4771-a991-8a8435e6fda | 0.765 | 0.46 | 0.8525 | 0.775 | 0.825 |
| 15 | 015_Nex_BRC_A_TCGA-A18S-11A_9ed6d2a-2e9e-4dd4-9063-8043fa0d6e3c | 0.8948 | 0.89 | NaN | 0.85 | 0.8675 |
| 062 | Nex_BRCA_TCGA-BH-A1EO-11A | 0.90308930-c3e5-47bb-bcbe-58aeae77d3f80b4-83b0-7f1496cf29ae | 0.5849 | 0.48 | 0.8854 | 0.9 | 0.7493 |
| 064 | Nex_BRCA_TCGA-BH-A0DT-11A | 0.6640-a1d79-4ed7-9491-c500b8183d9a | 0.7462 | 0.42 | NaN | 0.55 | 0.6659 |
| 066 | Nex_BRCA_TCGA-BH-A209-11A | 0.5806-b610-fl-9369-270b-4e7b-9fe8-3a9e-529baa | 0.5979 | 0.24 | 0.6243 | 0.6 | 0.4645 |
| 068 | Nex_BRCA_TCGA-GI-A2C8-11A | 0.8364-b612-11c7-4422-a2e5-cac9-9846e777c | 0.601 | 0.48 | 0.7846 | 0.85 | 0.6998 |
| 070 | Nex_BRCA_TCGA-BH-A18L-11A | 0.7a0100-ced7-7f80-4f5b-98ae-7c738e87bbd3 | 0.927 | 0.81 | NaN | 0.8 | 0.9113 |
| 072 | Nex_BRCA_TCGA-E2-A15M-11A | b2138d6a-519f-4961-bf1c-0f6b35b5d888 | 0.4911 | 0.28 | NaN | 0.8 | 0.4285 |