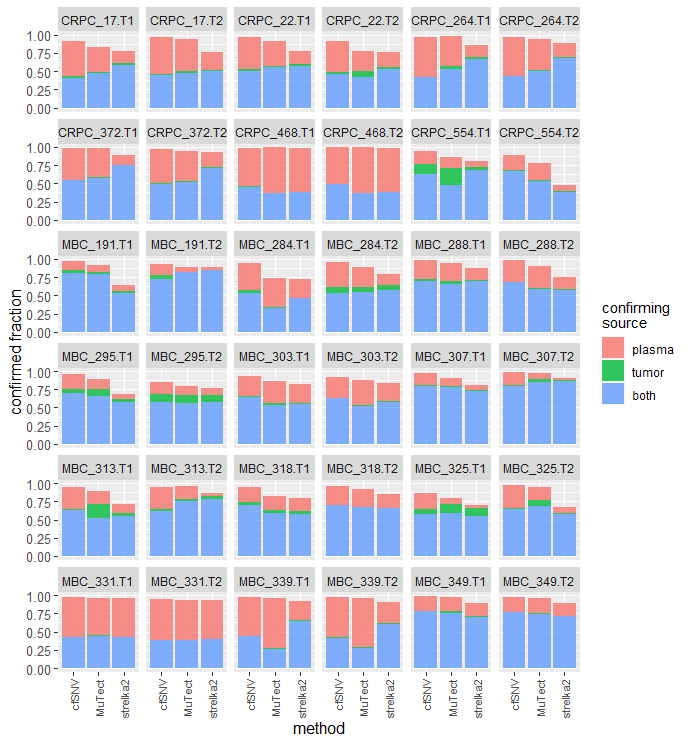
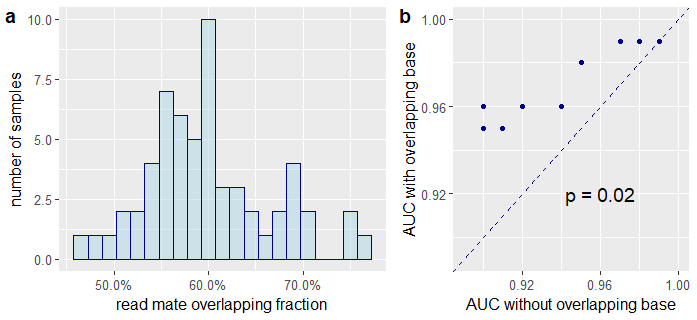
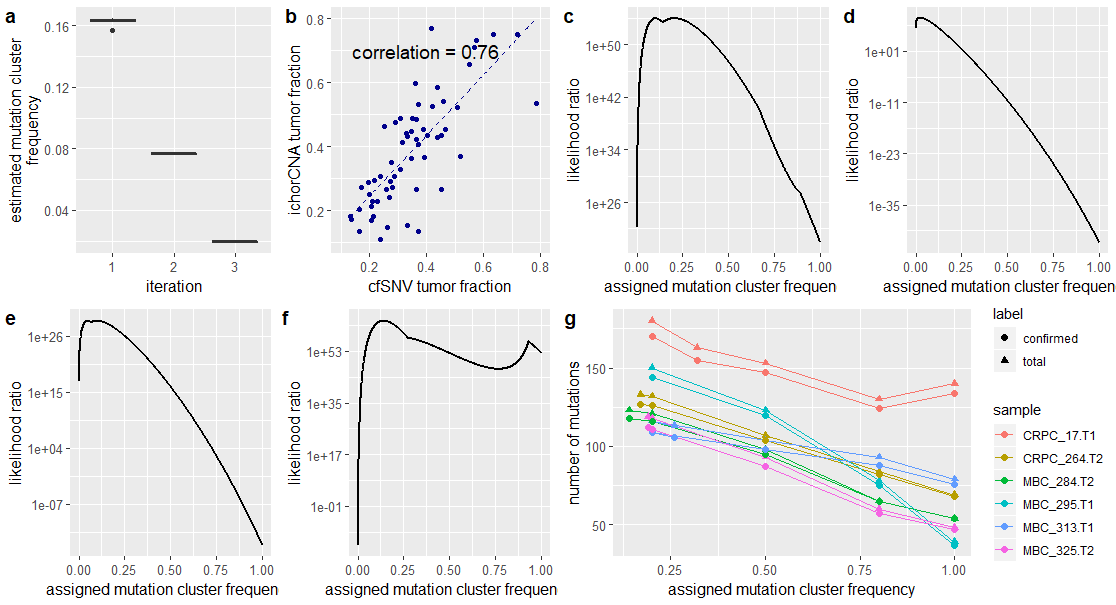
**Supplementary Figures**



**Supplementary Fig. 1 |** **Confirmed fractions of somatic SNVs from *cfSNV*, *MuTect* and*Strelka2*, along with confirming sources.**



**Supplementary Fig. 2 | Quantifying the existence and impact of overlapping read mates in cfDNA sequencing data.** **a,** Fraction of merged overlapping read mates in 59 cfDNA whole exome sequencing samples from metastatic cancer patients. **b,** Comparison of AUC metrics from classifiers trained on overlapping read pairs and non-overlapping read pairs on 12 testing samples.



**Supplementary Fig. 3 | Simulation on the estimated mutation cluster frequency**. **a,** Performance of mutation cluster frequency estimation in the first experiment with simulated samples containing purely synthetic mutations inserted at known VAF levels (20%, 8% and 2%). Each box in this plot shows the estimated mutation cluster frequency for the synthetic mutation cluster at the same VAF level in five independent simulation samples. **b,** Performance of mutation cluster frequency estimation in the third experiment with cfDNA data. The graph demonstrates the correlation between the tumor fractions estimated by *cfSNV* and *ichorCNA* on different sequencing experiments using the same cfDNA samples. **c-f,** The likelihood ratio plot of a simulated mutation with VAF 0.1 (**c**), 0.01 (**d**), 0.05 (**e**), and 0.2 (**f**), under varied mutation cluster frequencies. **g,** The number of mutations detected using different mutation cluster frequencies on four plasma samples whose significant mutation clusters have prevalence ≤ 20%.The left most point on each line showed the number of mutations detected at the estimated mutation cluster frequency.

A close up of a white wall

Description automatically generated

**Supplementary Fig. 4 | Difference between the tumor-genotype fractions estimated in the first round and the second round (refined by mutation candidates).**

A screenshot of a computer

Description automatically generated

**Supplementary Fig. 5 | Case studies of mutations rescued from the standard clustered read position filter.** Panel 1 and Panel 2 are plasma samples at two different time points. Panel 3 is blood normal sample. Panel 4 is tumor biopsy sample. The position between the two dashed vertical lines is the variant position. In the first plasma sample (panel 1) the variant base in three of five supporting reads clustered at the same location, so this position was filtered by the standard clustered read position. However, in the second plasma sample (panel 2) and the tumor biopsy sample (panel 4), there was no clustered read position event at this position, and it was detected as a mutation. Therefore this clustered read position event in panel 1 is likely due to non-random fragmentation other than misalignment.

**A close up of a map

Description automatically generated**

**Supplementary Fig. 6 | ROC curves for random forest classifiers in out-of-sample tests.** **(a,c)** trained on overlapping reads only, and **(b,d)** trained on non-overlapping reads only. The test data are taken from either **(a,b)** metastatic breast cancer patients or **(c,d)** metastatic prostate cancer patients. The classifiers were trained using data derived from only WES data of cfDNA sample from a single patient (patient MBC\_315, sample SRR6708941). Each independent testing dataset (from one patient) has its own ROC curve. The numbers in parentheses are area under curve (AUC) metrics.

**Supplementary Tables**

|  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- |
| **ID** | | **Error** | | **Variant** | | **Total** |
| **Patient ID** | **Sample ID** | **No Overlap** | **Overlap** | **No Overlap** | **Overlap** |
| CRPC\_17 | SRR6708976 | 53021 | 16105 | 2730098 | 633412 | 3432636 |
| CRPC\_22 | SRR6708978 | 85005 | 24760 | 1937389 | 403481 | 2450635 |
| CRPC\_22 | SRR6708979 | 38423 | 10160 | 1263267 | 301639 | 1613489 |
| CRPC\_264 | SRR6708961 | 6164 | 3673 | 2396790 | 647633 | 3054260 |
| CRPC\_264 | SRR6708962 | 4358 | 2659 | 1842613 | 495319 | 2344949 |
| MBC\_284 | SRR6708924 | 5788 | 3163 | 5169493 | 1219561 | 6398005 |
| MBC\_284 | SRR6708925 | 13245 | 6507 | 3110798 | 842080 | 3972630 |
| MBC\_288 | SRR6708927 | 12224 | 7294 | 3230182 | 1066400 | 4316100 |
| MBC\_288 | SRR6708928 | 194 | 222 | 3230182 | 1066400 | 4296998 |
| MBC\_303 | SRR6708935 | 22679 | 11741 | 5165262 | 1006855 | 6206537 |
| MBC\_303 | SRR6708936 | 9067 | 3399 | 5165262 | 1006855 | 6184583 |
| MBC\_315 | SRR6708920 | 14762 | 9181 | 482160 | 118771 | 624874 |

**Supplementary Table 1 | Sample IDs and number of testing reads extracted.** Patient IDs follow the naming convention in reference [36], while sample IDs are the SRA accession IDs of the sample.

|  |  |  |  |
| --- | --- | --- | --- |
| **Filter** | **Description and default thresholds** | **Pass** | **Hold** |
| Strand bias | Removes false positives caused by context-specific or systematic sequencing errors. These are recognized by observing an abnormal number of variant alleles in a single direction of reads. We test for strand bias by calculating the binomial probability that variant alleles are only observed in a single direction of reads. The parameter used in the binomial distribution is the strand ratio, calculated from reference supporting alleles. Candidates are rejected if the binomial probability is less than 0.05. This threshold is equivalent to saying that when variant alleles are observed from both directions, the ratio between forward variant alleles and reverse variant alleles must be in the range [7, 1/7] to pass this filter. | pass | pass |
| Variant frequency | Remove false positives caused by tri-allelic sites or random sequencing errors. We compare the number of variant supporting reads and the number of non-germline reads. If the fraction of variant supporting reads in all non-germline reads is less than 0.8, then the candidate is rejected. | pass | pass |
| Mapping quality | Remove false positives caused by context-specific or location-specific misalignments, so that reads aligned to the candidate site have a lower mapping quality in general. Candidates always pass the filter when there are enough reads (n > 20) with high mapping quality (phred > 10). Otherwise, candidates are rejected if there are more than 3 reads with low mapping quality (phred < 5) and the number of total reads is less than 20. Candidates are also rejected if the fraction of reads with low mapping quality is greater than 0.4, or if the median mapping quality at the position is low (phred < 10). Finally, candidates meeting none of these criteria pass the filter. | pass | pass |
| Variant allele mapping quality | Remove false positives caused by misalignment, where reads aligned to the candidate have a lower mapping quality in general. Candidates pass the filter directly when there are enough reads (n > 3) with high mapping quality (phred > 10). Candidates are rejected if there are more than 3 reads with low mapping quality (phred < 5) and the number of total reads is less than 20. Candidates are also rejected if the fraction of reads with low mapping quality is greater than 0.4, or if median mapping quality at the position is low (phred < 10). Otherwise, candidates pass the filter. | pass | pass |
| Variant base quality | Remove false positives caused by incorrect base calls. Candidates are rejected when the median base call error probability of variants is more than 7 times the median base call error probability of reference bases. Candidates are also rejected if the number of high-quality variant bases (phred > 23) is fewer than 3. | pass | pass |
| Supporting fragments | Mark candidates with strong evidence. Candidates with more than three supporting reads are marked as having strong evidence. | pass | - |
| Tumor coverage | Remove false positives caused by inadequate sequencing. Candidates are rejected if they have ≤ 10x coverage in plasma. Candidates are marked as having a low-confidence VAF if they have coverage > 10x and ≤ 50x in plasma. | pass | pass |
| Normal coverage | Remove false positives caused by inadequate sequencing of the matched germline blood sample. Candidates are rejected if their coverage is ≤ 7 in the germline blood sample. | pass | pass |
| Nearby repeats | Remove false positives caused by misalignment of nearby repeats. Candidates are rejected if they are within a repeat region annotated by *RepeatMasker*. | pass | pass |
| Nearby indels | Remove false positives caused by misalignment of nearby indels. Candidates are rejected if they are within a distance of 5 base pairs from an indel. Indels are marked by the alignment tool *(BWA*), and collected for use in this filter if there are ≥ 3 reads supporting an indel at the position or if the fraction of reads supporting an indel is greater than 0.02. | pass | pass |
| Binomial VAF test | Remove false positives with low confidence given the current global tumor fraction. Candidates are rejected if the binomial probability of observing the number of variant supporting reads is less than 0.1. | pass | - |
| Public databases | Remove germline variants by rejecting candidates present in a public germline database (dbSNP). | pass | pass |
| Co-occurrence of candidates | Remove false positives associated with misalignment. Candidates are rejected if they always co-occur with other candidates on the variant supporting reads, or their position on the reads are always the same. | pass | pass |

**Supplementary Table 2 | Description of site-level post-filtration criteria and thresholds.**

|  |  |
| --- | --- |
| **Category from iterative detection** | **Passing criteria** |
| pass | number of variant supporting reads > 5 |
| pass | number of variant supporting reads > 3 and binomial probability of observing more variant supporting reads given the tumor fraction < 0.6 |
| hold | number of variant supporting reads > 12 |
| hold | number of variant supporting reads > 5 and binomial probability of observing more variant supporting reads given the tumor fraction < 0.6 |

**Supplementary Table 3 | Description of read-level post-filtration criteria and thresholds.**

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
|  | **Sites** | **Total read pairs** | **Overlapping read pairs** | **Non-overlapping read pairs** |
| Variant SRR6708941 | 33355 | 4397873 | 772283 | 3625590 |
| Variant SRR6708920 | 33355 | 600931 | 118771 | 482160 |
| Error SRR6708941 | 41903 | 59170 | 21325 | 37845 |
| Error SRR6708920 | 23148 | 23943 | 9181 | 14762 |

**Supplementary Table 4 | Training data for the random forest model.** Reads from two experiments of the same plasma sample were labeled as containing true variants or sequencing errors. Reads from SRR6708941 were used for training, while reads from SRR6708920 were only used for validating the model.

|  |  |  |  |
| --- | --- | --- | --- |
| **Feature description** | **Feature type** | **Non-Overlapping model** | **Overlapping model** |
| base call in a 7-bp window centered on the query site from the read | categorical | yes | yes |
| base quality in a 7-bp window centered on the query site from the read | numerical | yes | yes |
| CIGAR information in a 7-bp window centered on the query site from the read | categorical | yes | yes |
| Occurrence CIGAR operators from the read | Boolean | yes | yes |
| mapping quality of the read pair | numerical | yes | yes |
| distance to the nearest indel on the read pair | numerical | yes | yes |
| whether the query site was contained in a homopolymer with size >= 5 | Boolean | yes | yes |
| insertion sizes of the read pair | numerical | yes | yes |
| mapping flags of the read and the mate | categorical | yes | yes |
| base call in a 7-bp window centered on the query site from the mate | categorical | no | yes |
| base quality in a 7-bp window centered on the query site from the mate | numerical | no | yes |
| CIGcentered on the query site from the mate | categorical | no | yes |

**Supplementary Table 5 | Extracted features from read pairs for the random forest models.** The column “non-overlapping model” indicates which features are used in the random forest model for purifying non-overlapping read pairs. The column “overlapping model” indicates which features are used in the model for overlapping read pairs.