UNIVERSITY OF CALIFORNIA

Los Angeles

Disease detection and monitoring from plasma cell-free DNA

A dissertation submitted in partial satisfaction
of the requirements for the degree
Doctor of Philosophy in Bioinformatics

by

Shuo Li

2020

ABSTRACT OF THE DISSERTATION

Disease detection and monitoring
from plasma cell-free DNA

by

Shuo Li
Doctor of Philosophy in Bioinformatics
University of California, Los Angeles, 2020
Professor Xianghong Jasmine Zhou, Chair

With the noninvasiveness of sample collection and the comprehensiveness of the DNA profile from various tissues, plasma cell-free DNA (cfDNA) has attracted enormous attention for many applications, including disease-related marker identification, disease detection, and disease monitoring. However, since cfDNA is a mixture of disease-related DNA in an overwhelming pool of DNA from normal cells, the weak disease signal poses a major challenge for these applications. Current methods usually employ traditional error suppression for genomic DNA samples and deep sequencing on small panels, which limit their performance. A fundamental and yet underdeveloped task for these applications is the precise and sensitive calling of somatic single nucleotide variants (SNVs) from cfDNA. We present *cfSNV*, a somatic SNV detection method designed specifically for cfDNA that incorporates multilayer error suppression and hierarchical mutation calling. The accurate and sensitive identification of disease-related markers can provide a reliable foundation for disease monitoring, which is essential for assessing the effectiveness of treatment. We provide a novel cancer monitoring approach, *OncoMonitor*, which comprehensively analyzes tumor mutations and sensitively detects minimal residual disease, cancer recurrence, secondary disease, and cancer progression with longitudinal cfDNA samples. Further leveraging the information in cfDNA samples, we developed a workflow using the microbiome composition in cfDNA for disease detection, which provides complementary disease evidence to current human-origin cfDNA-based methods. In summary, this work uses statistical methods and machine learning models to address the current limitations in mutation detection and disease monitoring in cfDNA and provide complementary information for disease detection.

The dissertation of Shuo Li is approved.

Wenyuan Li

Frank Alber

Yingnian Wu

Xinshu Grace Xiao

Xianghong Jasmine Zhou, Committee Chair

University of California, Los Angeles

2020

This dissertation is dedicated to my beloved family
especially to my parents Chunmei Li and Yudong Li
to my parents-in-law Shuang Yu and Ping Wang
to my dearest husband Fangzhou Wang.

TABLE OF CONTENTS

LIST OF FIGURES

LIST OF TABLES

ACKNOWLEDGMENTS

My doctorial journey is one of the most memorable experiences in my life. It would not have been possible and joyful for me to complete this journey without the support from my advisor, mentors, colleagues, friends, and family members.

My most sincere appreciation goes to Professor Xianghong Jasmine Zhou for being a thoughtful, caring, inspiring, and enthusiastic advisor. Her encouragement built up my confidence and her constant support prepared me to become a better scientist. From her intelligence and guidance, I learned to form clear, meaningful and impactful questions, to make proper and practical plans for a research project, to identify key points in a scientific problem, to justify, appreciate and highlight the novelties of the contribution, and especially to present a research project in a clear and structural way, and lots more. Apart from the invaluable guidance and discussions, her life stories and her helpful advice diminished my confusion and anxiety for the future and will keep inspiring me in my future career and life. I am so grateful to have been under her guidance for the past five years.

I would like to thank Dr. Wenyuan Li, Dr. Frank Alber, Dr. Yingnian Wu, and Dr. Xinshu Grace Xiao for serving on my qualification and thesis committee. Their insightful comments and suggestions deepened my understanding of my projects. More specifically, I would like to thank Dr. Wenyuan Li for all his great supports, including suggestions to the method development and the presentation of all my projects, Dr. Frank Alber for constructive comments on my manuscripts, Dr. Xinshu Grace Xiao for discussions about capability and limitation of the error suppression strategy in *cfSNV*, and Dr. Yingnian Wu for discussion about the machine learning models for the error suppression in *cfSNV*.

I would also like to thank all my colleagues and collaborators. Without your help, this thesis cannot be possible. Dr. Zorawar Noor and Dr. Edward Garon are amazing people to work with. They provided precious patient data, lent me their expertise in cancer oncology and immunotherapy, and helped me interpret the biological insights of our analysis.

They have been so supportive and responsive for all my questions. I particularly want to thank Dr. Wing Hung Wong for his kind support and essential comments on the *cfSNV* manuscript. I want to thank Dr. Rui Liu and Dr. Pei Chen. They helped me a lot on all the computing works of the cfDNA microbiome analysis on sepsis patients. I would also like to thank Qingjiao Li and Mary Stackpole for all the inspiring discussion and helps in my daily research life. I also want to express my gratitude to Dr. Weihua Benny Zeng, Dr. Zuyang Yuan, and Dr. Xiaohui Ni. They helped me perform experiments for my projects and gave me valuable suggestions as biological and medical experts.

A special thank goes to Mandy McWeeney, our former student affair official, and Gene Gray, our current student affair official. Mandy's professional assistance made my transfer from USC so smooth and efficient. I sincerely appreciate the helps from Mandy and Gene to all my questions.

I also want to thank many friends at USC and UCLA: Zhu Liu, Wenbo Chen, Ziye Wang, Mengxi Yu, Yingfei Wang, Yuanbin Wang, and Ning Wang for the joy we shared together and for always being so helpful and supportive.

I especially would like to thank my family. My parents, Yudong Li and Chunmei Li, deserve special thanks for the continuous and unconditional love and support they gave me throughout my entire life. I am forever indebted to my parents for giving me the experiences that made me who I am. I would like to show my great gratitude to my dearest husband Fangzhou Wang. I always feel blessed to meet him in that lovely morning, when our love story began. My doctorate journey has not been easy, and so do our lives. I am so grateful for his bearing my mood swings, my anxieties, and my fears, and still loving me the same. I am so grateful for his always being by my side and working hard every day. I am so grateful for his believing in me and always letting me be his top priority.

Chapter 2 is a version of **Li, Shuo**, Zorawar Noor, Weihua Zeng, Mary Stackpole, Xiaohui Ni, Zuyang Yuan, Yonggang Zhou, Wing Hung Wong, Vatche G. Agopian, Steven M. Dubinett, Frank Alber, Wenyuan Li, Edward B. Garon, and Xianghong Zhou. "Sensitive

detection of tumor mutations from blood and its application to immunotherapy prognosis." ZN is supported by UCLA Tumor Immunology Training Grant (USHHS Ruth L. Kirschstein Institutional National Research Service Award T32CA009120). EBG is supported by NCI RO1CA208403, NCI P30CA0160142, CIRM CLIN2-10784. This work is supported by NCI U01CA230705, NCI R01CA246329, NCI U01CA237711.

Chapter 3 is a version of **Li, Shuo**, Weihua Zeng, Xiaohui Ni, Yonggang Zhou, Mary Stackpole, Zorawar Noor, Zuyang Yuan, Edward B. Garon, Steven M. Dubinett, Wenyuan Li, Xianghong Jasmine Zhou "*OncoMonitor*: monitoring cancer recurrence and detecting MRD by exome-wide mutation analysis of cell-free DNA." ZN is supported by UCLA Tumor Immunology Training Grant (USHHS Ruth L. Kirschstein Institutional National Research Service Award T32CA009120). EBG is supported by NCI RO1CA208403, NCI P30CA0160142, CIRM CLIN2-10784. This work is supported by NCI U01CA230705, NCI R01CA246329, NCI U01CA237711.

Chapter 4 is a combination of Chen, Pei, **Li, Shuo**, Wenyuan Li, Jie Ren, Fengzhu Sun, Rui Liu, and Xianghong Jasmine Zhou. "Rapid diagnosis and comprehensive bacteria profiling of sepsis based on cell-free DNA" and collaboration work with Mary Stackpole. The first part of this work is supported by a UCLA start-up package for XJZ; National Natural Science Foundation of China (Nos. 11771152, 11901203, 11971176); Guangdong Basic and Applied Basic Research Foundation (2019B151502062); National Institutes of Health Grant (Nos. R01CA246329, U01CA237711); China Postdoctoral Science Foundation funded project (No. 2019M662895); the Fundamental Research Funds for the Central Universities (2019MS111). The second part of the work is supported by NCI U01CA230705, NCI R01CA246329, NCI U01CA237711.

VITA

2019 Ph.D. Candidate, Bioinformatics, University of California Los Angeles.

2017 M.S., Computer Science, University of Southern California.

2015 B.S., Statistics, School of Mathematical Sciences, Peking University, Beijing, China.

PUBLICATIONS

PEER-REVIEWED PUBLICATIONS

Chen, Pei, **Li, Shuo**, Wenyuan Li, Jie Ren, Fengzhu Sun, Rui Liu, and Xianghong Jasmine Zhou. "Rapid diagnosis and comprehensive bacteria profiling of sepsis based on cell-free DNA." ***Journal of Translational Medicine*** 18, no. 1 (2020): 1-10.

Li, Shuo, Xialiang Dou, Ruiqi Gao, Xinzhou Ge, Minping Qian, and Lin Wan. "A remark on copy number variation detection methods." ***PloS one*** 13, no. 4 (2018): e0196226. **Li, Shuo**, Zorawar Noor, Weihua Zeng, Mary L. Stackpole, Xiaohui Ni, Zuyang Yuan, Yonggang Zhou, Wing Hung Wong, Vatche G. Agopian, Steven M. Dubinett, Frank Alber, Wenyuan Li, Edward B. Garon, and Xianghong Zhou. "Sensitive detection of tumor mutations from blood and its application to immunotherapy prognosis." ***in review***.

Li, Shuo, Weihua Zeng, Xiaohui Ni, Yonggang Zhou, Mary L. Stackpole, Zorawar Noor, Zuyang Yuan, Edward B. Garon, Steven M. Dubinett, Wenyuan Li, Xianghong Jasmine Zhou "*OncoMonitor*: monitoring cancer recurrence and detecting MRD by exome-wide mutation analysis of cell-free DNA." ***in preparation***.

CHAPTER 1

Introduction

Plasma cell-free DNA (DNA) is a degraded DNA fragment released into the blood. These DNA fragments are derived from dying human cells and microorganisms from different tissues [GDP17] [LNZ16] [SJC15] [SJC15]. In particular, disease-related tissues, such as tumor tissues, also release DNA to the bloodstream [GDP17] [SJC15]. Therefore, cfDNA comprises a thorough profile of DNA from various body sites, including those that are representative of diseases. With the noninvasiveness of sample collection and the comprehensiveness of the DNA profile, cfDNA delivers the possibility of taking repeated blood samples and consequently tracing the changes in cfDNA during the natural course of diseases or during treatment [SHP11]. Given its great potential, cfDNA is treated as a possible surrogate for invasive or time-consuming sample collection methods, and a wide range of applications have been developed for disease diagnosis and monitoring, especially for cancer [ZBF18] [CLW18] [KLC17] [LLK18] [FMP12] and infectious diseases [GSG16] [BTR19].

However, a major challenge for detecting the disease-related signal from cfDNA is the often very low fraction of these DNA fragments from disease-related tissues in the overwhelming pool of DNA from normal cells [ABW17]. To detect the signal, previous methods usually (1) rely on traditional error suppression strategies, which fail to accommodate the cfDNA-specific properties, or (2) deep sequencing on small panels [NBT14] [CCL17] [ABW17] [GSW15] [TWT16] [MCS19], which therefore limit the genomic coverage of these methods. Due to the inter-individual and intra-individual differences in diseases, such as tumor heterogeneity, focusing on small regions of the genome might naturally lead to over-looking disease-related signals outside the targeted range. To address the aforementioned limitations, we present a set of computational methods for analyzing medium-depth cfDNA sequencing data covering a wide range of genomes (e.g., whole genome, whole exome, and whole methylome). Our study mainly focuses on cancer and infectious diseases, such as sepsis. These methods incorporate a number of cfDNA-specific properties and thus enable (1) the sensitive and accurate identification of tumor somatic single nucleotide variants (SNVs), (2) a comprehensive analysis for cancer monitoring, and (3) the delivery of complementary evidence for disease diagnosis (cancer and sepsis) from microbe-origin cfDNA.

In Chapter 2, we present a new somatic SNV caller for cfDNA from cancer patients, named *cfSNV*, which provides hierarchical mutation profiling and multilayer error suppression, including error suppression in read mates, site-level error filtration and read-level error filtration. We validated the performance of *cfSNV* in both simulation data and real cancer patient data. It achieves high precision and sensitivity in cfDNA samples that have both low tumor purity and a highly heterogeneous clonal landscape. As an example application, in this study, we demonstrate that applying *cfSNV* to cfDNA whole-exome sequencing (WES) data allows a new promising biomarker (truncal-bTMB) for immunotherapy prognosis by simultaneously capturing both the tumor mutation burden and clonal structure information. Compared to existing methods, *cfSNV* can dramatically reduce the required sequencing depth for profiling given genomic regions, therefore reducing the sequencing cost and further making WES of cfDNA a viable option.

In Chapter 3, we present a new cancer monitoring approach, *OncoMonitor*, based on cfDNA standard WES data, which comprehensively monitors cancer by analyzing both the mutations in pretreatment/surgical samples and those in newly emerging tumor clones. We demonstrate that our method achieves a sensitive and specific detection of recurrence and secondary disease in simulated plasma samples with low tumor fractions. Specifically, in a cohort of non-small-cell lung cancer patients, we show that our method can detect comprehensive tumor changes for response prediction, which cannot be achieved by previous methods based only on mutations in pretreatment/surgical samples.

In Chapter 4, we present a workflow using the microbiome composition in cfDNA for disease detection. With cfDNA sequencing, we trained a random forest model based on the microbial composition of healthy individuals and patients. Specifically, as examples, we focused on rapid sepsis diagnosis and noninvasive cancer detection. For rapid sepsis diagnosis, we applied the workflow to cfDNA whole-genome sequencing (WGS) data from sepsis patients and healthy individuals and evaluated the performance of the random forest model. Then, we analyzed the co-occurrence network of the candidate pathogens and showed the characteristics of the abundant pathogens, which could be further utilized to guide therapies. For cancer detection, we applied the workflow to cfMethyl-seq data from cancer patients and noncancer individuals and showed the ability of our workflow to discriminate cancer and noncancer individuals and classify cancer patients with different tissues of origin. We further validated the microbes with the top importance in the random forest model with statistical tests and findings in previous studies.

CHAPTER 2

Sensitive detection of tumor mutations from blood and its application to immunotherapy prognosis

2.1 Introduction

Cell-free DNA (cfDNA) in blood has received enormous attention thanks to its clinical utility as a surrogate for tumor biopsy, especially in cases where the latter is unavailable or insufficient [VYF14]. A tissue biopsy is invasive by nature, and is only extracted from a single site. In contrast, cfDNA in blood can be obtained noninvasively, and provides a comprehensive landscape of the heterogeneous genetic alterations in tumors. Hence, a wide range of cfDNA-based applications have been developed to detect cancer [ZBF18] [CLW18] [KLC17] [LLK18], locate tumors in the body [KLC17] [FMP12], select the best therapy [RAC19] [GPK18], and monitor treatment [FMP12] [CVD17] [CCC18]. All these applications depend upon an indispensable, yet underdeveloped task: precise and sensitive calling of somatic single nucleotide variations (SNV) from cfDNA sequencing data. This task is challenging to conventional SNV callers because somatic mutations in cfDNA generally have low allele frequency. This property follows from the major hallmarks of cfDNA: (1) cfDNA is a mixture of DNA fragments from both normal and tumor cells, and in most cancer patients the fraction of tumor-derived cfDNA is extremely low ($< 1\%$ for most early-stage cancer patients [ABW17] and $< 10\%$ even for some metastatic patients [CWF18]). Therefore, almost all somatic mutations in tumor-derived cfDNA have much lower allele frequencies than in solid tumors. (2) cfDNA comes from the entire volume of a tumor, and from every tumor present in a patient, so it provides complete information on clonal and subclonal mutations, while subclonal mutations generally have lower allele frequencies than clonal mutations.

To conquer these challenges in cfDNA data, some efforts have been made on the experimental technologies and the computational error filtration to optimize the variant calling on targeted deep-sequencing data [MAS19] [KMK18] [WDX20]. Despite the encouraging progress, existing methods are not sufficiently equipped to handle this complicated scenario, especially on the medium-coverage sequencing data such as whole-exome sequencing (WES). Specifically, they are lacking in three aspects: (1) They do not automatically account for the low fraction of tumor-derived cfDNA or variability due to the tumor clonal hierarchy in the context of mutation calling, though clonality has been considered in other studies [ABW17]. A few SNV callers (e.g., *MuTect* [CLC13]) try to handle the issue of tumor impurity, but even these cannot robustly and sensitively detect mutations with variant allele frequency (VAF) $< 5\%$ [CLC13]. One mutation caller [DJB19] integrated clonal information to improve somatic mutation calling, but this method required extra user input of the clonal hierarchy. (2) They rely on post-filtration steps that require reliable estimation of site-level statistics (e.g. strand bias and averaged base quality). However, robust estimates are challenging to obtain for low-frequency cfDNA mutations, due to insufficient variant supporting reads, and become even more difficult for WES, which does not permit deep sequencing in terms of sequencing cost. (3) They do not exploit two key features of cfDNA, namely short fragment size ($\sim$ 166 bp on average) and non-random fragmentation [JCC15] [JST18], which we prove in this study to be very useful for enhancing the detection performance.

Therefore, we have developed a new cfDNA SNV caller named *cfSNV*. This is the ***first*** algorithm to comprehensively address the cfDNA-specific challenges and opportunities mentioned above. Taking advantage of modern statistical models and machine learning approaches, *cfSNV* provides hierarchical mutation profiling and multi-layer error suppression, including error suppression in read mates, site-level error filtration and read-level error filtration. It achieves high precision and sensitivity in cfDNA samples that have both low tumor purity and highly heterogeneous clonal landscape, even for medium-coverage sequencing data such as WES, in a purely computational fashion without attachment to specific experimental technologies. In both simulated and real patient data, as shown in Figure 2.2, *cfSNV* vastly outperforms existing tools, showing tens of times increase in sensitivity in detecting mutations with low allele frequency while maintaining high precision. Up to now, existing efforts on SNV detection in cfDNA rely on specifically designed experiments (e.g. barcode-based sequencing [NBT14] [NLK16] [MAS19] [KMK18] [WDX20]) with ultra-deep sequencing, which, therefore, are only effective on small gene panels. *cfSNV* can dramatically reduce the required sequencing depth for profiling given genomic regions and therefore bring down the cost by magnitudes, and further make the Whole-Exome Sequencing of cfDNA a viable option. As an example application, in this study we demonstrate that applying *cfSNV* to cfDNA WES data allows a new promising biomarker (truncal-bTMB) for immunotherapy prognosis, by simultaneously capturing both the tumor mutation burden and clonal structure information.

2.2 Results

2.2.1 *cfSNV*: A new computational framework for calling SNVs from cfDNA

We developed the *cfSNV* framework (Figure 2.1c) by introducing five new techniques (Figure2.1b) into the standard SNV calling workflow (Figure 2.1a). Each of the five techniques either overcomes a specific challenge of cfDNA or takes advantage of a specific feature of cfDNA. The challenges and features are:

1. *Short fragments*: the fragment length distribution of cfDNA peaks at 166bp. Therefore, paired-end sequencing (usually 150 bp for a read) usually results in a large fraction of overlapping read mates, which can be used to suppress sequencing errors (Figure 2.1b(1) and Figure 2.1c(i)). This error-correction step is performed before the standard

data preprocessing.

2. *Mixed nature*: the cfDNA found in blood from cancer patients generally consists of a small amount of tumor-derived cfDNA among an overwhelming majority of cfDNA from normal cells. By incorporating the germline data of white blood cells (WBCs) from the same subject, we can fit a joint-genotype model that precisely describes this mixture. Specifically, we model the triplet (g_T, g_N, g_W) of genotypes, among which g_T and g_N actually describes the mixed nature of cfDNA by representing the genotypes of **T**umor-derived cfDNA, **N**ormal cfDNA respectively, while g_W represents the genotype of the matched **W**BC DNA for the reference purpose. The modeling of cfDNA is performed by first aggregating reads from mutation hotspots in order to robustly estimate the fraction of tumor-derived cfDNA, which then serves as a parameter in the joint-genotype model for the probabilistic deconvolution of tumor-derived and normal reads in a specific locus. Note that the fraction of tumor-derived cfDNA is usually low, therefore it cannot be precisely estimated at a single locus due to the limited tumor-derived reads falling onto the locus. Aggregating reads of multiple potential mutation loci allows more robust estimate of the tumor cfDNA fraction.

3. *Heterogeneous clonal compositions*: unlike tissue biopsies, a blood sample includes DNA fragments from all tumor sites, so it covers the full range of clonal and subclonal mutations [ABW17] [MDP15]. However, admitting a heterogeneous cfDNA clonal composition poses a great challenge to existing methods. A statistical model capable of fitting the data from clonal mutations, inevitably sacrifices accuracy for subclonal mutations using the same parameters, which however has been practiced in all existing methods. To address this challenge, we can take advantage of the fact that the mutations associated with a given clone have similar allele frequency in cfDNA. The mutations are therefore naturally clustered according to the clonal hierarchy [ABW17] [MDP15]. This fact permits us to develop a “divide-and-conquer” algorithm (Figure 2.1c(ii)) that first automatically groups the mutations of the highest and similar fre-

quencies into a cluster, then estimates parameters that best fit the data of the cluster. We then remove these detected mutations, and repeatedly perform the same operation to identify the next most frequent mutation cluster. In other words, this algorithm intelligently and iteratively searches for the best parameters of the cfDNA joint-genotype statistical model (Figure 2.1c(ii.a)) to detect and model the cluster of mutations with the highest frequency in the cfDNA sample (Figure 2.1c(ii.b)), then removing its loci and data. The process repeats, detecting the next most frequent mutation cluster at each iteration (Figure 2.1c(ii.d)), until no more mutations are detected with confidence. Therefore, we can profile the cfDNA mutation hierarchy in terms of mutation frequencies.

4. *Non-random fragmentation*: cfDNA fragments have preferred start and end positions [JST18], so true mutations could cluster at certain positions on the supporting reads. Conventional tools which assume randomly fragmented genomic DNA tend to classify mutation candidates with clustered positions on reads as misalignment artifacts, therefore eliminating them [CLC13]. Consequently, the true mutations in cfDNA samples could be removed by this artifact filter in the conventional tools. We remove this artifact filter to keep true cfDNA mutations, while building a new filter to jointly analyze the positions of multiple nearby mutation candidates and precisely remove cfDNA misalignment artifacts (Figure 2.1b(4) and Figure 2.1c(ii.c)). The new filter successfully rescued $1 \sim 16$ mutations (median 6.8) per subject that would have been discarded by conventional methods.

5. *Confusion between sequence errors and low-frequency mutations*: When the tumor-derived cfDNA fraction is low, sequencing errors impair the detection sensitivity. We get around the problem of low signal-to-noise ratio for individual alleles by developing a machine learning approach to accurately distinguish true variants from sequencing errors for individual reads. The algorithm exploits a variety of contextual information from the region surrounding the target allele (Figure 2.1b(5) and Figure 2.1c(iii)) to

provide an accurate prediction. The detailed workflow is illustrated in Figure 2.7 and described in Methods.

2.2.2 Validation of ***cfSNV*** on simulation data

To evaluate the performance of *cfSNV* in calling low-frequency somatic mutations, we tested the method on simulated data. To generate the dataset, a set of predefined somatic SNVs were added to the simulation data, the mixture of the cfDNA sequencing data from 8 cancer patients (around 2200x, see section 2.4). To avoid the interference of the somatic mutations and the germline mutations in individual cfDNA samples, we carefully removed reads contained these mutations (see section 2.4). We used eight variant allele frequencies (VAF) ranging from 0.1% to 8% for the SNVs, in order to simulate tumor heterogeneity in patients plasma (see section 2.4). Mutations called at positions other than the ground-truth SNVs were regarded as false positives. We compared *cfSNV* with two established SNV callers, *MuTect* and *Strelka2*, which were designed for solid tumor tissue samples but have been utilized in studies on cfDNA samples . The results of the test show that *cfSNV* far outperforms the two competing methods for all ground-truth mutations (Table 1a). Specifically, *cfSNV* achieves much higher sensitivity (64.0%) than *MuTect* (20.3%), *Strelka2* (25.6%), and *Strelka2* with disabled filters (32.7%), while maintaining very high precision (100.0% vs. 99.2%, 100.0%, and 11.6% respectively). When looking at low-frequency mutations specifically, the contrast between *cfSNV* and other methods is even stronger (Table 2.1b and Figure 2.2a). In this sub-population, *cfSNV* detected 39.7% $\sim$ 74.6% of mutations with VAFs of 0.1% $\sim$ 1% respectively, whereas most competing methods detected zero mutations. Overall, without sacrificing precision, *cfSNV* showed 3.2 and 2.5 times increase in sensitivity of all somatic SNVs, and 14.2 and 106.5 times increase in sensitivity of somatic SNVs with allele frequency $< 1\%$ comparing to *MuTect* and *Strelka2* respectively.

2.2.3 Validation of *cfSNV* on patient data

Next, we tested the ability of *cfSNV* to call somatic mutations on patient data. We collected WES data of samples obtained from six metastatic prostate cancer (castrate-resistant prostate cancer, CRPC) and twelve metastatic breast cancer (MBC) patients [AHF17] (see section 2.4). For each patient, we collected a metastatic tumor biopsy sample, a WBC sample, and two plasma cfDNA samples. The cfDNA samples were drawn at two different time points after the patients were diagnosed as metastatic, with time gaps in the range $14 \sim 138$ days (Table 2.8). We compare the different SNV callers in terms of the confirmation rate, defined as the fraction of mutations detected in one cfDNA sample that are also confirmed to be present in either the matched tumor tissue or the other cfDNA sample. Following a recent study [AHF17], we confirm the presence of a mutation by the number of variant supporting reads from the raw sequencing data (i.e. supported by ≥ 3 variant reads, see Methods) [AHF17]. The confirmed mutations in the matched tumor tissue are regarded as true positives. As a single tumor biopsy sample cannot profile all tumor clones in a metastatic cancer patient, we also regarded mutations present in both plasma samples but absent in the tumor biopsy as true positives. Thus, this confirmation rate is basically the same as the precision on the patient data. We performed the evaluation in the following two steps. First, we tested the confirmation rate of *cfSNV* across different samples. We applied *cfSNV* to the 18 cfDNA samples of the initial time point to obtain a baseline mutation set for calculating the confirmation rate. We validated the truncal and branch mutations detected. A mutation is defined as "truncal" if its VAF is above 60% of the average VAF of the five most frequent mutations in the sample; otherwise, it is "branch" (Methods). Averaged across all 18 subjects, 97.7% and 76.7% of truncal mutations are confirmed in the later cfDNA sample and the tumor biopsy of the same subject, respectively. 93.2% and 62.1% of branch mutations are confirmed in the later cfDNA sample and the tumor biopsy of the same subject respectively (Figure 2.7). The confirmation rates are similar if we instead use mutations detected in the 18 later cfDNA samples as a baseline (Figure 2.7, 96.5% and 78.6% for truncal mutations, 93.0% and 59.9% for branch mutations in the earlier cfDNA sample and the tumor biopsy respectively). We observed that the larger the time gap between the two blood draws, the lower the confirmation rate of branch mutations between the two cfDNA samples (Pearsons correlation between the time gap and the confirmation rate = 0.57, p = 0.0003, see Figure 2.17 and Table 2.8). This trend was not observed for truncal mutations (Pearsons correlation between the time gap and the confirmation rate = 0.08, p = 0.651, see Figure 2.17 and Table 2.8). This observed trend implied that the mutation landscape of cfDNA could change with time, especially for branch mutations. Second, we compare *cfSNV* with competing methods (MuTect and *Strelka2*) on the same samples in terms of the confirmation rate. Although these metastatic plasma samples with a high tumor fraction (ranging from 13% to 79%) are not the best scenario to demonstrate the power of *cfSNV* (as majority of mutations have VAF $> 10\%$, see Figure 2.19), still *cfSNV* outperformed both methods, achieving the highest precision (confirmation rate) in 33 out of 36 samples (Figure 2.3a). For the remaining 3 samples, *cfSNV*s precision was only marginally lower than the highest precision (by 0.2%, 0.9%, and 1.2%). In fact, the lower the VAF of mutations, the more power exhibited by *cfSNV* compared to other methods (Figure 2.2b). Strikingly, at VAF of 1%, 3%, and 5%, *cfSNV* yielded 100.0%, and 8.4% higher precision and identified $+\infty$, 9.9, and 1.8 times more confirmed mutations than *MuTect*(no mutations detected below 2% using default *MuTect*); *cfSNV* yielded 82.4%, 53.6%, abd 39/4% higher precision and identified 31, 5.8, and 3.9 times more confirmed mutations than *Strelka2* (Figure 2.3b and Figure 2.2b). Across all VAF range, on average *cfSNV* yielded 5% and 14% higher precision (Figure 2.3a) and detected 1.6 and 2.0 times more confirmed mutations (Figure 2.3a), respectively, demonstrating an overall higher precision and sensitivity. Note that all three methods have consistently higher confirmation rates in the second plasma sample than the matched tumor tissue sample, implying that plasma cfDNA offers a more comprehensive coverage of tumor mutations than a single tumor biopsy for metastatic cancer patients. Therefore, whenever multifocal sampling of tumors from a metastatic cancer patient is infeasible, cfDNA is a viable alternative to obtain comprehensive mutation profiles.

2.2.4 Experimental analysis of five new techniques

Here, we quantitatively assess how each of the five new techniques impacts the performance of *cfSNV*.

1. **Suppression of sequencing errors using overlaps of read mates.** The pair-end sequencing of cfDNA results in significant overlaps in the read mates. For example, in 95% of 59 cfDNA samples collected from Adalsteinsson et al. [AHF17], > 50% of read mates overlap (Figure 2.9 and Table 2.7). Our result shows that using overlapping read pairs, combined with a machine learning approach (see point (v) below and Figure 2.1b(5)), can greatly facilitate the detection of true mutations while rejecting sequencing errors. Specifically, we compare the models with and without using the overlapping read information, the AUC performance averaged across 36 independent test datasets (cfDNA samples from Adalsteinsson et al. [AHF17]) shows significant improvement (one-sided Wilcoxon rank sum test p-value = 3.38e-8, Figure 2.9).

2. **Enhance mutation detection by the joint-genotype model that allows for cluster-focused mutation calling.** As aforementioned, a model cannot use the same parameter to best fit both clonal and subclonal mutations that have distinct allele frequencies. We therefore introduce the "divide-and-conquer" strategy to first train the model to detect only mutations of the cluster with the highest frequency, and then remove loci of these detected mutations, and repeat the same procedure for the next most frequent mutation cluster. The key component of this iterative process is our joint-genotype model that supports the cluster-focused mutation calling. Specifically, the model has a parameter of describing how frequent the mutation cluster is (denoted as θ) and this parameter allows the model to best fit the data of only those mutations in this cluster, not all the mutations of the heterogeneous landscape. Therefore, we as-

sess the model by answering two questions: (1) Can θ estimated by our method reflect the VAFs of the mutations in the most frequent mutation cluster? We designed three experiments to answer this question, using simulated data with synthetic mutations, simulation data obtained by mixing real sample data with a known dilution ratio, and real cfDNA data. In the first experiment, we generated sequencing data with three groups of synthetic mutations: one mutation cluster with a VAF of 20%, one cluster with a VAF of 8%, and one with a VAF of 2% (see section 2.4). Our method not only automatically identifies the most frequent cluster and estimates its VAF, but also finds the other two clusters in subsequent iterations (Figure 2.10). In the second experiment, we subsampled and mixed sequencing reads from WBC and primary tumor biopsy samples, both taken from the same cancer patient (Methods). The tumor fraction, which is estimated by the frequency θ of the most frequent mutation cluster in these mixed samples, correlates very strongly (Pearsons correlation = 0.99) with the ground-truth mixing dilution (Figure 2.4a) across the study population. In the third experiment, we used data from two independent sequencing experiments (WES and WGS) on the same cfDNA sample from cancer patients. Specifically, we compare the tumor fraction estimated by *cfSNV* on WES to that estimated by *ichorCNA* on WGS. This result, shown in Figure 2.10, also confirms that our method accurately estimates the frequency of the major mutation clusters. (2) Does accurately estimating the mutation cluster frequency θ enhance mutation detection? We generated simulated sequencing data with a list of predefined θ values, from 0% to 100%, and observed the optimal θ that fits the joint-genotype model. Our performance metric is the model-to-data fitness ratio, defined as the ratio between the likelihoods of correct and incorrect joint genotypes (see section 2.4). A higher ratio means that the model is a better fit, so the mutation is more likely to be identified. Our result shows that any given mutation is best fit by the model when θ takes on a value close to the mutations frequency (Figure 2.10). In addition, when comparing the fitness of the model with and without θ (i.e., comparing the two likelihood ratios), we find that the smaller a mutations VAF, the larger the difference (e.g., the model-to-data fitness ratio is 40 times higher with θ present, for VAF$< 5\%$). This relationship indicates that an accurate θ estimate significantly enhances the detection power for low-frequency mutations (Figure 2.4b). Furthermore, we used cfDNA samples whose frequent mutation clusters have low frequency ($< 20\%$) to further confirm this conclusion (Figure 2.10). More mutations were detected when the assigned θ approached the true value of the mutation cluster frequency.

3. **Enhance the sensitivity of mutation detection by an iterative process.** We compared two versions of *cfSNV*, with and without the iterative process, on real data: four cfDNA samples whose frequent mutation clusters have low frequency ($< 20\%$ estimated from *cfSNV* and *ichorCNA*). With the iterative process, *cfSNV* detected 1.41 to 1.73 times more confirmed mutations (true positives) than *cfSNV* without the iterative process (Figure 2.4c). Both versions had high precision: namely, 95.3% and 95.0% for *cfSNV* with and without the iterative process respectively (Figure 2.4c).

4. **High confirmation rate of rescued mutations by cfDNA-specific post-filtration.** Compared with the conventional post-filtration strategy, which models the distribution of variant-base positions on reads, our new filtration strategy rescues 1 16 mutations (6.8 on average) per sample among the 36 plasma samples in this study. In 69.4% (26) of the samples, 100% of the rescued mutations are confirmed in either the matched tumor biopsy or the other plasma sample (Figure 2.4d).

5. **Machine learning approach to distinguish true mutations from sequencing errors in cfDNA reads.** The independent data used to test the machine learning model are data from 12 MBC and 6 CRPC patients. We hand-labeled read pairs containing high-confidence mutations or sequencing errors, and applied the random forest classifier (Methods). Our method achieves an average AUC-ROC of 0.95 over the MBC cfDNA samples (Figure 2.4e and Figure 2.11) and an average AUC-ROC of

0.94 over the CRPC cfDNA samples (Figure 2.4f and Figure 2.11). This result shows that our machine learning model can distinguish true mutations from sequencing errors with high accuracy at the level of individual reads. It implies that our machine learning model is non-specific to tumor types, and can easily be generalized to include samples from many kinds of tumors.

2.2.5 Application to predict the outcome of anti-PD-1 treatment: a new bTMB measure

Cancer immunotherapies, which activate a patients own immune system to kill tumor, have remarkably improved the clinical outcome of a subset of patients with non-small-cell lung cancer (NSCLC) [RHS15]. To better predict the therapy response and identify patients with potential clinical benefit, tumor mutational burden (TMB) based on solid tumor biopsies, which measures the extent of nonsynonymous genetic changes of the tumors, has been studied and utilized as a biomarker in various cancer types [RHS15] [SMM14] [MMG18], including NSCLC. In addition to the work on TMB, recent studies [GPK18] [WDC19] have shown that blood-based tumor mutational burden (bTMB) is an attractive alternative to tissue-based TMB due to three advantages: (1) noninvasiveness, (2) more comprehensive mutation coverage (by cfDNA) than a single-site tumor biopsy, and (3) the VAFs of mutations in cfDNA reflect their clonality in tumors. It has also been shown that in solid tumor samples, high truncal neoantigen load and low intra-tumor heterogeneity more significantly associate with longer progression-free survival (PFS) than total neoantigen load alone [MFR16] [WBP19]. Advantage (3) allows the inference of the clonality of tumor-derived mutations from cfDNA, and thus improves the prognosis. To fully exploit advantages (2) and (3), profiling of cfDNA with a broad genomic coverage (e.g. whole exome) is needed. However, due to the lack of efficient tools to accurately call SNV from cfDNA using medium-coverage WES data (e.g. 200x), all current bTMB methods [GPK18] [WDC19] use small gene panels (¡600) in order to perform deep sequencing (e.g. ¿ 5000x). Small panels can only sparsely sample the total mutation landscape, so the resulting estimates of TMB or bTMB are influenced by population and sampling variation [FGP19]. In contrast, *cfSNV* enables sensitive and precise mutation profiling in even medium-depth sequencing data, thus allowing us to fully profile the mutation landscape as well as benefit from all the other advantages offered by cfDNA. Specifically, we exploit the clonality information in cfDNA to develop a new immunotherapy prognosis metric, truncal-bTMB, which uses only truncal mutations called by *cfSNV* from the WES profiling of cfDNA samples (Methods). We applied this new metric to predict the outcomes of anti-PD-1 treatment, and achieved superior performance compared with bTMB and TMB. To comprehensively evaluate the predictive power of the measures bTMB and truncal-bTMB (facilitated by our powerful tool *cfSNV*), we studied a cohort of 30 non-small-cell lung cancer patients who received anti-PD-1 treatment (pembrolizumab). Blood samples were drawn from these patients before their treatment. All cfDNA samples were sequenced with WES. First, we compared bTMB based on different mutation callers (MuTect, *Strelka2* and *cfSNV*). We split the 30 patients into two groups using the population median [RHS15] of the respective truncal-bTMB metric (the distribution shown in Figure 2.18), which we call the high-burden (>median) and low-burden (≤median) groups, and evaluate how Kaplan-Meier survival curves of the progression-free survival time (PFS) differ between the two groups. The truncal-bTMB calculated based on *cfSNV* mutation calls had the most significant one-sided log-rank p-value (Figure 2.5a-c), 0.015 (cfSNV) vs. 0.225 (Strelka2) and 0.322 (MuTect), implying that the truncal-bTMB derived from *cfSNV* has the highest power for predicting patients with longer PFS. We further show that the truncal-bTMB metric is a more powerful predictor than the bTMB metric, for which the PFS association is less significant (Figure 2.5d-f), 0.097 (cfSNV) vs. 0.369 (Strelka2) and 0.446 (MuTect), although *cfSNV* mutation calls again yielded the best predictor. Note that using any of the three callers, truncal-bTMB always offers better predictive power than bTMB, indicating that combining mutation clonality and intra-tumor heterogeneity improves predictive power. Interestingly, comparing the three variant calling methods, the disagreement of the high/low burden group assignment concentrated on the samples with estimated tumor fraction lower than 20%, indicating that those samples contributed most to the superior performance of *cfSNV*. This is consistent with the aforementioned major strength of *cfSNV* in sensitively and precisely calling mutations in samples with low tumor fraction. Furthermore, we compared tumor-derived TMB with bTMB and truncal-bTMB on a subset of 14 patients, for whom the tumor biopsies were available. Again, *cfSNV*-derived truncal-bTMB had the best performance in predicting outcomes (Figure 2.8) also in this cohort, where the one-sided log-rank test p-values are 0.028 for truncal-bTMB, 0.280 for TMB, and 0.067 for bTMB with *cfSNV*, respectively. In this cohort, *cfSNV*-derived trunctal-bTMB showed the best performance in predicting the PFS outcome, as the truncal-bTMB values gave the most significant p-value between the high-burden group and the low burden group. From the survival analysis, the high truncal-bTMB in the plasma cfDNA was associated with the improved progression-free survival. Therefore, even though our analysis was based on a small cohort of NSCLC patients, our proposed new measure, by exploiting the unique advantages of cfDNA using *cfSNV*, provides a promising prognosis indicator for anti-PD-1 immunotherapy on NSCLC patients.

2.3 Discussion

We presented a new computational framework, *cfSNV*, that sensitively detects low-frequency somatic SNVs in cfDNA sequencing data. *cfSNV* is equipped with a series of innovative techniques to address cfDNA-specific challenges (i.e., mixed tumor-derived/normal cfDNA, low tumor-derived cfDNA fraction, high heterogeneity, and non-random fragmentation) and take advantage of cfDNA-specific features (high rate of overlapping reads, complete coverage of the mutation landscape). Specifically, (1) we designed a joint-genotype statistical model, parametrized by the mutation cluster frequency, to probabilistically deconvolute the mixture of tumor-derived and normal reads in cfDNA data; (2) we developed an iterative approach to detect clusters of mutations with different variant allele frequencies; (3) we designed a data pre-processing step that exploits the overlapping read mates caused by short cfDNA fragments to improve data quality; (4) we developed a new procedure for filtering misalignment errors that accounts for the non-random fragmentation pattern of cfDNA; and (5) we developed a machine learning approach that incorporates the sequencing context to filter errors at the level of individual reads. Equipped with the new techniques and special considerations for cfDNA, we have shown *cfSNV* outperforms the existing methods in terms of overall precision and sensitivity. The cancer patients of this study are metastatic, so their plasma cfDNA has high fractions of tumor-derived cfDNA and carry many high-frequency mutations that can be usually detected by all conventional methods. For these high-frequency mutations, *cfSNV* can still achieve the best performance. Especially, for those low-frequency mutations, *cfSNV* achieves the sensitivity >10 times higher than competing methods, without sacrificing precision, not only in the real patient data but also in the simulation data. These results demonstrate that *cfSNV* could provide high-quality discovery of both low- and high-frequency mutations even in the medium-depth sequencing data, such as WES data. *cfSNV* is a general computational framework, applicable to medium- or deep-coverage cfDNA sequencing data. While the existing efforts address the challenge of low tumor-content in cfDNA by ultra-deep sequencing of a limited number of loci, the power of *cfSNV* can significantly reduce the required sequencing depth for profiling given genomic regions, and therefore the cost of the current cfDNA clinical tests. On the other hand, coping up with the ever-increasing demand of large gene panels, *cfSNV*s power allows cfDNA medium-depth WES to be used in a wide variety of clinical applications. Here we presented an example application of cfDNA WES that offers a novel and effective immunotherapy response measure (truncal-bTMB) by exploiting a comprehensive coverage of the clonal mutation landscape in cfDNA. We believe that *cfSNV* will greatly facilitate cfDNA-based therapy prognosis and longitudinal monitoring.

2.4 Methods

2.4.1 Data collection

We collected WES data of 42 metastatic cancer patients from two sources: (1) the data of 41 patients were obtained from Adalsteinsson et al. [AHF17] under dbGaP accession code phs001417.v1.p1. Each patients data include a WBC sample, a tumor biopsy sample, and one or two plasma cfDNA samples. Among the 41 patients, 18 have two plasma cfDNA samples. A patient (MBC_315) has her cfDNA sample sequenced with both WES and deep WGS. (2) The data of one patient was obtained from Butler et al. [BJP15] (European Nucleotide Archive accession numbers ERS700858, ERS700859, ERS700860, and ERS700861). The data include a white blood cell sample, a primary breast cancer biopsy sample, a metastatic liver biopsy sample, and a plasma cfDNA sample. We also collected samples from 30 lung cancer patients and generated our own WES data as described below.

2.4.2 Human subjects

We collected blood samples, tumor biopsy samples and white blood cell samples from 30 non-small-cell lung cancer patients from KEYNOTE-001 [GRH15] and KEYNOTE-010 [HBK16], who all provided informed consent for research use. The blood and tissue collection was described in the full protocol of KEYNOTE-001 and KEYNOTE-010. The project was approved by the Institutional Review Boards (IRBs) of University of California, Los Angeles (IRB# 12-001891, IRB# 11-003066, and IRB# 13-00394).

2.4.3 Genomic DNA whole exome sequencing (WES) library construction

The WBC and tissue samples underwent multiplexed paired-end whole-exome sequencing (WES) to a target depth of 100-150x on HiSeq 2000/3000 (Illumina, San Diego, CA) performed by the UCLA Technology Center for Genomics & Bioinformatics. Macrodissection was not performed. DNA isolation was performed with DNeasy Blood & Tissue Kit (Qiagen, Germany); exon capture and library preparation used the KAPA HyperPrep Kit and Nimblegen SeqCap EZ Human Exome Library v3.0 (Roche, Switzerland).

2.4.4 Plasma cfDNA whole exome sequencing (WES) library construction

For each of the 30 non-small-cell lung cancer patients, venipuncture was performed by trained phlebotomists such as nurses or medical assistants. Blood tubes were centrifuged at 1,800g for 20 min at room temperature and plasma supernatant was isolated within 2 hours of collection. Samples were stored at -80C until use. Then, cfDNA was extracted from their plasma samples using the QIAamp circulating nucleic acid kit from QIAGEN (Germantown, MD). The cfDNA WES library was constructed with the SureSelect XT HS kit from Agilent Technologies (Santa Clara, CA) according to the manufacturers protocol. No molecular barcodes were used in the sequencing libraries. In brief, 10ng of cfDNA was used as input material. After end repair/dA-tailing of cfDNA, the adaptor was ligated. The ligation product was purified with Ampure XP beads (Beckman-Coulter, Atlanta, GA) and the adaptor-ligated library was amplified with index primer in 10-cycle PCR. The amplified library was purified again with Ampure XP beads, and the amount of amplified DNA was measured using the Qubit 1xdsDNA HS assay kit (ThermoFisher, Waltham, MA). 700-1000 ng of DNA sample was hybridized to the capture library and pulled down by streptavidin-coated beads. After washing the beads, the DNA library captured on the beads was re-amplified with 10-cycle PCR. The final libraries were purified by Ampure XP beads. The library concentration was measured by Qubit, and the quality was further examined with Agilent Bioanalyzer before the final step of 2x150bp paired-end sequencing by Genewiz (South Plainfield, NJ), at an average coverage of 200.

2.4.5 The workflow of *cfSNV*

cfSNV takes the plasma DNA and germline DNA sequencing data of a patient as inputs, and detects SNVs using the three-step process described below (Figure 2.6). The outputs at the end of the pipeline are the detected mutations and the tumor fraction.

2.4.5.1 Data preprocessing.

A short cfDNA fragment (size peak 166 base pairs) usually has the overlapping read mates in the paired-end sequencing data and this cfDNA feature poses two data preprocessing challenges: double-counting the overlapping regions and biasing variant allele frequencies. Simply discarding overlapping regions [CLC13] [MHB10] [DBP11] would waste a large amount of sequencing data. Actually, these overlapping regions provide the opportunity to detect and suppress sequencing errors as two copies of the original DNA template are available. Therefore, in addition to the standard data preprocessing steps of alignment, deduplication, local realignment, and base quality recalibration, we perform an additional step: merging overlapping read mates. This new step is performed before the standard preprocessing pipeline (Figure 2.6) for addressing two challenges and utilizing the emerging opportunity from the overlapping regions. It corrects the read counts in overlapping regions, thereby removing the bias in variant allele frequencies from double-counting, and also detects sequencing errors by comparing the context of the two cfDNA copies in the overlapping region. Specifically, inconsistent bases in the overlapping region are corrected to the base call with higher quality, while consistent bases are confirmed and assigned a high base quality. This step is implemented by *FLASh* [MS11]. Those read mates that are overlapping are merged as single-end reads, while the rest of read pairs are treated as paired-end reads. The parameters for *FLASh* were adjusted to accommodate the typical fragment lengths of cfDNA and read lengths in sequencing data. We aligned paired-end reads and single-end reads separately to the hg19 human reference genome. We used *bwa mem* [LD09] to align the reads, and *samtools* [LHW09] to sort them. Then we used *picard tools* [Ins16] *MarkDuplicates* to remove duplicate reads resulting from PCR amplification. After this step, we added read group information to the bam file using *picard tools AddOrReplaceReadGroups*, and realigned reads around indels using *GATK* [MHB10] [DBP11]. The target regions in realignment were identified through *GATK RealignerTargetCreator*, then reads around target regions were realigned using *GATK IndelRealigner*. Finally, base quality scores were recalibrated using *GATK BaseRecalibrator* and *PrintReads*.

2.4.5.2 Iterative process of detecting mutation candidates

As illustrated in Figure 2.6, this process repeats a sequence of four steps until no more mutation candidates are detected with confidence. In each complete iterative round, a mutation cluster is determined.

- *(Step 1) Estimating the mutation cluster frequency θ of the most frequent mutation cluster.* As the frequency of mutations in cfDNA are naturally clustered to the clonal hierarchy [ABW17] [MDP15], we defined a mutation cluster as a group of mutations with similar variant allele frequencies. The mutation cluster frequency θ is defined as the fraction of cfDNA carrying the mutations in the cluster, out of all cfDNA mapped to the same genomic positions. Due to the low amount of tumor-derived cfDNA in blood, individual sites may be covered by a very small number of tumor-derived cfDNA reads (or none), leading to highly uncertain estimates of the tumor-derived cfDNA fraction. Therefore, we aggregate tumor-derived signal from multiple sites to obtain a robust estimation. The first step is to identify sites across the genome that are highly likely to be mutated (called hotspots). Specifically, a locus is selected as a hotspot if it meets the following criteria: (a) both matched germline DNA and cfDNA sequencing data have adequate coverage (30 for germline, 80 for cfDNA in this study); (b) bases at the locus in matched germline DNA data contain only reference alleles; (c) the average sequencing error probability is less than the variants observed frequency; (d) reads in both matched germline DNA and cfDNA data have high mapping quality (≥ 20); (e) no strong strand bias is observed; and (f) enough variant supporting reads are observed in the cfDNA data (≥ 3). All hotspots are ranked by read coverage, VAF, and the counts of variant alleles in matched germline DNA data. Next, we estimated θ by maximizing the likelihood of observing the data at all hotspots $\mathrm{P}(\mathbf{X}|\theta)$, where $\mathbf{X} = (X_1, X_2, \cdots, X_r, \cdots)$ is the cfDNA sequencing data and X_r represents all the information (such as sequence and base qualities) contained in a single read r. For each locus, we assume that reads are independently sampled from a cfDNA joint-genotype model that is denoted by the triplet $G = (g_{\mathrm{T}}, g_{\mathrm{N}}, g_{\mathrm{W}})$ where the subscripts N, T and W refer to normal cfDNA, tumor-derived cfDNA and WBC DNA respectively. However, only the normal cfDNA genotype g_{N} and tumor-derived cfDNA genotype g_{T} are utilized in this step, because the WBC genotype g_{W} is already controlled by hotspot selection (criterion b). All three genotypes are used in *(Step 2)* and *(Step 3)*, described below. Specifically, g_{W} is essential in the later step of the process to remove germline mutations and WBC-derived somatic mutations (clonal hematopoiesis). Based on the independence assumption of reads, the likelihood of θ at a hotspot is calculated as the product of the probabilities of observing individual reads covering the hotspot, given the parameter θ. We express this relation as follows:

$$\mathrm{P}(\mathbf{X}|\theta) = \prod_{r \in \mathbf{R}_{\mathrm{H}}} \mathrm{P}(X_r|\theta) = \prod_{r \in \mathbf{R}_{\mathrm{H}}} \sum_{G_r} \mathrm{P}(X_r|G_r, \theta)\mathrm{P}(G_r),$$

where $\mathbf{R}_{\mathrm{H}}$ is the pool of reads covering a selected hotspot and G_r is the joint genotype at the hotspot covered by a read r. Note that sometimes a read r may cover multiple hotspots, so G_r could be the combination of all hotspots covered by read r. Since an individual read is sequenced from either tumor-derived cfDNA (with probability θ) or normal cfDNA (with probability $1 - \theta$), the likelihood of observing this read can be calculated using a probabilistic mixture model that describes the presence of two subpopulations:

$$\mathrm{P}(X_r|G_r,\theta) = \theta\mathrm{P}(X_r|g_{\mathrm{T}_r}) + (1-\theta)\mathrm{P}(X_r|g_{\mathrm{N}_r}),$$

where g_{T_r} and g_{N_r} are the tumor-derived and normal cfDNA genotypes of the hotspot on read r. The information contained in an aligned read r (X_r) consists of base calls, base qualities and mapping qualities at hotspots in the read. So we can expand $\mathrm{P}(X_r|g_{\mathrm{T}_r})$ as follows:

$$\mathrm{P}(X_r|g_{\mathrm{T}_r}) = \mathrm{P}(B_r|g_{\mathrm{T}_r})$$

and

$$\mathrm{P}(X_r|g_{\mathrm{N}_r}) = \mathrm{P}(B_r|g_{\mathrm{N}_r}),$$

where B_r represents base calls at the hotspot on read r. The base quality and the mapping quality are embedded in the probability of sequencing error described below. The probability of error ϵ is calculated from the mapping quality m and the base quality q, as $1-(1-10^{-\frac{m}{10}})(1-10^{-\frac{q}{10}})$. Assuming that all sequencing error directions have the same probability, the probability of observing a base call given genotype g can be calculated from the probability of error ϵ. So we have

$$\mathrm{P}(\mathbf{A}|g) = \begin{cases} 1-\epsilon, & \text{if } g = \mathbf{AA}, \\ \frac{1}{2}(1-\epsilon)+\frac{1}{6}\epsilon, & \text{if } g = \mathbf{AB}, \\ \frac{1}{3}\epsilon, & \text{if } g = \mathbf{BB}, \end{cases}$$

where $\mathbf{A}$ and $\mathbf{B}$ are the reference and non-reference alleles respectively. Based on the above formulation, an estimation of the mutation cluster frequency θ can be achieved by optimizing the likelihood $\mathrm{P}(\mathbf{X}|\theta)$ via the Expectation-Maximization (EM) algorithm or a simple grid search.

- *(Step 2) Predicting somatic mutation candidates using the joint genotype.* After obtaining θ, we can determine the variant status of a genomic position by finding the joint genotype that optimizes the posterior probability of reads at that position. As illustrated in Figure 2.6(ii), for a given locus, we collected all reads that are aligned to the locus in both cfDNA data and the matched germline DNA data, then computed the posterior probability of each joint genotype from the observed reads. This probability can be modeled by a mixture model similar to that aforementioned in *(Step 1)*. Subsequently, the joint genotype with the highest posterior probability is adopted as the prediction result at the locus. Somatic mutation candidates are then selected by following the inferred joint genotype. In this step, we used the matched germline data $\mathbf{X}_{\mathrm{W}}$ from WBC and the cfDNA data $\mathbf{X}_{\mathrm{P}}$ from plasma cfDNA, consisting of normal cfDNA and tumor-derived cfDNA. For a specific locus, its joint genotype is determined as G_{MAP}, the joint genotype that maximizes the posterior probability given the observed data and θ:

$$G_{\mathrm{MAP}} = \arg\max \mathrm{P}(G|\mathbf{X}_{\mathrm{W}}, \mathbf{X}_{\mathrm{P}}, \theta).$$

Using Bayes theorem, we have

$$\mathrm{P}(G|\mathbf{X}_{\mathrm{W}}, \mathbf{X}_{\mathrm{P}}, \theta) \propto \mathrm{P}(\mathbf{X}_{\mathrm{W}}, \mathbf{X}_{\mathrm{P}}|G, \theta)P(G)$$

The probability of observing the data is the product of the probability of observing individual reads. So we have

$$\mathrm{P}(\mathbf{X}_{\mathrm{W}}, \mathbf{X}_{\mathrm{P}}|G, \theta) = \mathrm{P}(\mathbf{X}_{\mathrm{W}}|g_{\mathrm{W}})\mathrm{P}(\mathbf{X}_{\mathrm{P}}|g_{\mathrm{N}}, g_{\mathrm{T}}, \theta),$$

$$\mathrm{P}(\mathbf{X}_{\mathrm{P}}|g_{\mathrm{N}}, g_{\mathrm{T}}, \theta) = \prod_{r} \mathrm{P}(X_r|g_{\mathrm{N}}, g_{\mathrm{T}}, \theta),$$

and

$$\mathrm{P}(\mathbf{X}_{\mathrm{W}}|g_{\mathrm{W}}) = \prod_{r} \mathrm{P}(X_r|g_{\mathrm{W}}),$$

where X_r stands for a single read r. In the same way we calculate the likelihood of a given θ, we decompose $\mathrm{P}(X_r|g_{\mathrm{N}}, g_{\mathrm{T}}, \theta)$ and $\mathrm{P}(X_r|g_{\mathrm{W}})$ and get

$$\mathrm{P}(G|\mathbf{X}_{\mathrm{W}}, \mathbf{X}_{\mathrm{P}}, \theta) \propto \mathrm{P}(G) \prod_{r} [(1-\theta)\mathrm{P}(X_r|g_{\mathrm{N}}) + \theta\mathrm{P}(X_r|g_{\mathrm{T}})] \prod_{r'} \mathrm{P}(X_{r'}|g_{\mathrm{W}})$$

As the majority of normal cfDNA comes from WBCs, we set the prior distribution of the joint genotype G as

$$\mathrm{P}(G) = \mathrm{P}(g_{\mathrm{N}}, g_{\mathrm{T}}, g_{\mathrm{W}}) = \begin{cases} \mathrm{P}(g_{\mathrm{N}}, g_{\mathrm{T}}), & \text{if } g_{\mathrm{W}} = g_{\mathrm{N}}, \\ 0, & \text{otherwise.} \end{cases}$$

The joint distribution of the component $(g_{\mathrm{N}}, g_{\mathrm{T}})$ in joint genotype G has been defined in JointSNVMix [RDM12]. It can also be calculated from public databases. Based on above formulation, the joint genotype can be determined for every locus. By comparing the three components of the joint genotype with the highest posterior probability, then we can determine whether the locus is a somatic mutation, a germline mutation, or a loss of heterozygosity (LOH) site. The somatic mutation loci are input as mutation candidates in the next filtration steps. The above model is actually a probabilistic deconvolution of the normal and tumor signals in cfDNA. By incorporating the matched germline data (WBC) and the mutation cluster frequency θ, we separate the tumor-derived cfDNA from the total cfDNA at individual somatic SNV candidates, and thus enhance mutation detection (as shown in section 2.2.4 (2)).

- *(Step 3) Site-level filtration.* To reduce false positives from mutation candidates, we investigated a set of site-level statistics in raw data and *FLASh*-processed data (i.e., both single-end reads from merged overlapping read pairs, and paired-end read pairs without overlapping regions). The site-level statistics used here include averaged base quality, averaged mapping quality, strand bias, depth of coverage, and nearby sequencing context (e.g. repeats and indels). Detailed descriptions and default thresholds for these site-level filters are listed in Table 2.3. One essential filter to determine the mutation candidates in this iterative round is the binomial VAF test. It removes the mutation candidates whose VAF is not likely to be observed based on the current mutation cluster frequency. With the joint-genotype model and the binomial VAF test, the VAF of the mutation candidates in this iteration is around the estimated mutation cluster frequency, and thus these mutation candidates can form a cluster. Based on the results from all filters, each mutation candidate is sorted into one of three categories: “pass”, “hold”, or “reject”. Candidates in the “pass” category pass all filters, so they are very likely to be mutations. Candidates in the “hold” category fail some non-essential filters, so we cannot determine whether they are mutations at this step. Candidates in the “reject” category fail at least one essential filter (e.g. averaged base quality), so they are regarded as false positives and removed from further analysis. The requirements for a variant to be classified as either “pass or “hold, are listed in Table 2.3.

- *Iterating (Steps 1-3) to refine the mutation cluster frequency estimate.* After *(Step 3)*, we select hotspots from the mutation candidates in the “pass” category to refine the θ estimation in *(Step 1)*. By repeating *(Steps 1-3)* for the same mutation cluster, we obtain a stable frequency estimate and a group of mutation candidates for this cluster. Convergence is reached when the difference between two consecutive θ estimations is less than 0.01. In our experiments with simulation data, convergence is usually reached after only two rounds (Figure 2.12). Thus, with just one iteration of *(Steps 1-3)*, we already accurately capture the most frequent mutation cluster. In fact, our software offers both options: a quick version that performs only one round of estimation and candidate detection for each cluster, and a slow version that iterates until convergence for each mutation cluster.

- *(Step 4) Output and removal candidates from data.* After obtaining somatic mutation candidates from the most frequent mutation cluster, we output the mutation candidates in the “pass” and “hold” categories from (Step 3). Thus the mutation cluster at this iterative round is determined. Then we remove the loci and data of these sites from the cfDNA data. After removal, we continue iterating from *(Step 1)* to identify the next most frequent mutation cluster.

- *Termination criterion.* Mutation clusters are detected one at a time, in the decreasing order of their frequency in cfDNA. The process terminates until no mutation candidates are found in *(Step 4)* (i.e., the "pass" and "hold" categories are empty).

2.4.5.3 Error filtration at the read level

Site-level statistics provide some information on the difference between sequencing errors and true mutations, but are not adequate for error filtration in cfDNA. Due to the low tumor fraction and high heterogeneity of cfDNA, site-level frequency estimates are uncertain and unreliable for mutations with only a few supporting reads. To reduce the number of false positives among mutation candidates, we developed a machine learning filter to eliminate reads with sequencing errors at candidate sites and remove SNV candidates whose count of "confirmed" supporting reads fails to pass a threshold (see details in Table 2.4). Specifically, for each mutation candidate, we classify each of its supporting reads with a random forest model in order to distinguish sequencing errors from true variants. This model combines a variety of features (Table 2.6) and automatically discovers statistical relationships among the features that reflect sequencing errors. It is worth noting that read pair statistics (e.g. fragment length and features of the read mate) are always among the most informative features of the random forest model. Since this error filtration method is applied at the read level, it vastly improves the precision of detecting low-frequency somatic mutations. Although this read-level filter can be performed at any step of the method (e.g., after alignment or during the iterations), we prefer to perform it at the end of the *cfSNV* workflow in order to save computing time and resources. Generally, the later this step is performed, the fewer sequencing reads need to be inspected for errors, and thus the less-consuming time is needed for *cfSNV*. Practically, based on our hands-on experience of the real data, the times of inspecting read-level errors in the beginning of the process is reduced 50 times if it is performed at the end of the process: that is, for each read that needs to be inspected at the end of the process, at least 50 reads would need to be inspected at the beginning.

To train the random forest model, we used four WES sequencing datasets from the same cancer patient (MBC_315): two cfDNA sequencing datasets, a WBC sequencing dataset, and a tumor biopsy sequencing dataset. As the two cfDNA sequencing datasets were obtained from the same cfDNA sample, we can treat them as technical replicates and label their read pairs by their concordance. The training data are the supporting cfDNA read pairs at known mutation/error sites, and labeled as containing mutations or errors. Mutation sites are defined as the collection of common germline mutations detected using *Strelka2 germline* [KSH18] from all four datasets. In addition, common somatic mutations were detected using *Strelka2 somatic* [KSH18] and *MuTect* [CLC13] from two cfDNA-WBC pairs (cfDNA data vs. WBC data) and one tumor-WBC pair (tumor data vs. WBC data). Error sites are defined as sufficiently covered sites (>80x) with only one high-quality non-reference read (base quality $\geq$ 20 and mapping quality $\geq$ 40) in all four datasets. All labeled read pairs were extracted from raw cfDNA data using *picard tools FilterSamReads* (Table 2.5). Different features were extracted from the overlapping read pairs and the non-overlapping read pairs (Table 2.6). All categorical features were expanded using one-hot encoding method. We used the parameters of the random forest model as follows: (1) the number of decision trees is 100, (2) the maximum tree depth is 10, (3) imbalanced classes were handled by setting the class weights with option "balanced", and (4) other parameters were left at their default values. Two random forest classifiers (for overlapping read pairs and non-overlapping read pairs) were trained on read pairs extracted from the WES data (SRR6708941) using RandomForestClassifier from the python library *sklearn* [PVG11]. Read pairs from SRR6708920 were only used for validating the model. The trained classifiers are saved in the *cfSNV* code package (https://zhoulab.dgsom.ucla.edu/pages/cfSNV).

2.4.6 Truncal-bTMB measure

Somatic SNVs are annotated using snpEff. Nonsynonymous mutations and high-impact mutations are treated the same in snpEff results. Mutations from *Strelka2* were filtered if their VAF in the matched normal is greater than 1%. As the mutations VAF in cfDNA reflects the clonality of a mutation, we treat a mutation as truncal mutation if its VAF is greater than a threshold; otherwise it is a branch mutation. The threshold is defined as 60% of the average VAF of the 5 most frequent mutations. The truncal-bTMB measure can then be calculated as the sum of the normalized VAFs of all truncal nonsynonymous mutations.

$$\text{truncal-bTMB} = \frac{\sum \text{VAF of truncal mutations}}{\sum \frac{\text{highest 5 VAF}}{5}}.$$

2.4.7 Additional validation data for random forest classifier

To further test the random forest classifiers, we generated data from other patients with metastatic breast or prostate cancer (Table 2.2). For each patient, we obtained WES data of a WBC sample, a tumor biopsy sample, and plasma samples from two different time points. To generate the testing data and label the individual reads, we used the same procedure as described in section 2.4.5.3 for producing the training data.

2.4.8 Simulation with *BAMSurgeon* to evaluate precision and sensitivity

To evaluate the performance of *cfSNV*, we employed *BAMSurgeon* [EHH15] to generate simulation data by inserting individual mutations at different allele frequencies. The input to *BAMSurgeon* was a pool of cfDNA DNA data from eight cancer patients (MBC_333, MBC_336, MBC_292, CRPC_531, MBC_284, CRPC_525, MBC_303, and MBC_335) [AHF17]. Before mixing the eight cfDNA samples, to avoid the potential interference of the germline and somatic mutations in the individual cfDNA samples, we removed the reads covering these positions. The germline mutations were identified using a standard pipeline (GATK HaplotypeCaller) from individual samples; the somatic mutations were identified using *cf-*

SNV, *MuTect* and *Strelka2* from the individual cfDNA samples and their matched WBC samples. Three methods were used in the somatic mutation removal to avoid potential bias introduced in this step. Sequencing reads in the individual data were removed if they fell in a 200bp region centered at any germiline/somatic mutations (upstream 100bp and downstream 100bp). Then the eight individual cfDNA samples were merged. The mean target coverage of the pooled sample reached 2200x. The *BAMSurgeon* program attempted to insert 1000 somatic SNVs with different variant allele frequencies: 100 at 8%, 100 at 5%, 100 at 3%, 100 at 1%, 100 at 0.8%,100 at 0.5%, 200 at 0.3%, and 200 at 0.1%. A total of 581 mutations were successfully inserted. The other 419 mutations failed to insert into the sequencing data because their assigned VAF was incompatible with the sequencing depth in the original data, e.g. 1% VAF among 10 reads. We evaluated the performance of *cfSNV*, *MuTect* (disabling the contamination filter and testing different levels of the tumor_lod parameter) and *Strelka2* (default parameters, with enabled and disabled filters) on this simulation dataset by comparing the ground truth to the final variant reports. *MuTect* performed best when "tumor_lod" was set to 6, so we only report its results for this setting.

2.4.9 Mutation concordance between tumor biopsy and plasma samples

To validate our method on real data, we examined mutation concordance between a tumor biopsy sample and the plasma samples. This analysis involves twelve patients with metastatic breast cancer and six patients with metastatic prostate cancer [AHF17]. Each patient had a tumor biopsy sample, a WBC sample, and plasma samples from two different time points, all processed with WES. Mutations called from one plasma sample were checked in the raw sequencing data of the matched tumor biopsy sample and the other plasma sample. A somatic SNV is confirmed if there are at least three reads supporting the variant allele in the matched tumor biopsy sample or at least three reads supporting in the other plasma sample. A somatic SNV is not confirmed when the mutation has power at least 0.9 and fewer than 3 alternative reads [AHF17].

2.4.10 Comparison with *MuTect* and *Strelka2* on real cfDNA data

We compared our method to two state-of-the-art methods, *MuTect* and *Strelka2*. The same validation analysis was conducted for both methods on the same samples. Both tools were run with their default parameters unless otherwise noted in the text. The same confirmation process described in section 2.4.9 was conducted for somatic SNVs detected by *MuTect* and *Strelka2*.

2.4.11 Calculation of TMB and bTMB

For tissue biopsy samples, we called their somatic SNVs using *Strelka2*. The mutations were annotated using snpEff [CPW12]. TMB was calculated as the number of nonsynonymous SNVs. For plasma samples, we called somatic mutations using *MuTect*, *Strelka2* or *cfSNV*, and annotated them using snpEff. Mutations from *Strelka2* were filtered if their VAF in the matched normal is greater than 1%. We calculated traditional bTMB as the count of all nonsynonymous mutations with VAF ≥ 0.15.

2.4.12 Simulation with *BAMSurgeon* to evaluate the accuracy of the intelligent search of the most frequent mutation cluster

We used *BAMSurgeon* to generate simulation data. The input to *BAMSurgeon* was the WBC sequencing data from MBC_299. The program attempted to insert 300 mutations at three different VAF levels: 50 mutations at 20%, 150 mutations at 8%, and 100 mutations at 2%. Five simulated samples with the same settings were generated.

2.4.13 Generating spike-in simulation data to validate the mutation cluster frequency estimates

To evaluate the accuracy of our mutation cluster frequency estimation, we generated spike-in simulation data by mixing the primary tumor sequencing data (ERS700859) and the WBC sequencing data (ERS700858) of a metastatic breast cancer patient, at varying concentrations of cfDNA reads (from 2% to 20% in eight steps). Five independent mixtures are generated at every concentration. Each spike-in sample contains a total number of randomly sampled reads equivalent to 170x coverage of the targeted regions. The coverage of the targeted regions is limited by the number of sequencing reads in the original data.

2.4.14 Impact of the mutation cluster frequency on the model-to-data fitness at a single simulated mutation

The model-to-data fitness is evaluated using the likelihood ratio L_θ, the ratio between the maximum likelihood of a somatic-mutation joint genotype (i.e., homozygous and heterozygous genotypes) and the maximum likelihood of a non-somatic-mutation joint genotype (other joint genotypes) given an θ. Since we screened mutation candidates based on the joint genotype estimated at each position, this likelihood ratio reflects the ability of *cfSNV* to detect a somatic mutation candidate. We explored the theoretical properties of this likelihood ratio using simulated mutations, which consist of randomly generated base quality values, mapping quality values and a corresponding list of base calls reflecting the VAF. To compare the fitness of the model with and without θ, we calculated the value of $\frac{L_\theta}{L_1}$.

2.4.15 Impact of the mutation cluster frequency on real patient data

To test the impact of estimated mutation cluster frequency on real patient data, we selected four samples whose frequent mutation clusters have low frequency $< 20\%$ estimated from *cfSNV* and *ichorCNA*. We performed *cfSNV* on the four samples using both a predetermined value of θ (0.2, 0.5, 0.8, and 1.0) and the estimated θ of the most frequent mutation cluster in the sample. When we set θ as 1.0, the candidate screening model is the same as the regular joint genotype model for solid tumor samples, which is equivalent to a model that does not incorporate the estimated mutation cluster frequency. In this simulation, we also disabled the iterative procedure to converge on the best value of θ, so the candidate screening only took place at the given θ.

2.4.16 Rescuing mutations from conventional post-filtration

The "clustered read position", defined as positions with the alternative alleles being clustered at a constant distance from the start and end of the read alignment [CLC13], is regarded as hallmarks of misalignment artifacts. Because of the existence of the preferred start and end positions, the start and end sites of reads at some mutations tend to cluster together, and thus the position of the alternative alleles on these reads tend to cluster together. Therefore, cfDNA preferred start and end positions may make the true somatic mutations look like misalignment false positives with "clustered read position". To rescue these mutations, we removed the conventional clustered read position filter entirely. Instead, to remove misalignment artifacts, we implemented a new filter that simultaneously checks the co-occurrence of candidates and mismatch positions on the reads with alternative alleles (variant supporting reads), instead of purely relying on the "clustered read position" of a single mutation. If multiple candidates and mismatch positions exclusively co-occur on the variant supporting reads, we regard them as artifacts from misalignment (Table 2.3). A mutation is called "rescued" if it is reported by *cfSNV* but would be filtered by conventional methods due to the clustered read position. For each rescued mutation, the same confirmation process described in section 2.4.9 was conducted. The fraction of confirmed rescued mutations among all rescued mutations was calculated for every sample. Indeed, we were able to confirm that for some rescued mutations, the variant bases are more clustered in cfDNA reads than in solid tumor samples (Figure 2.13), validating our rationale.

2.4.17 Data availability

Sequencing data have been deposited into EGA under accession code EGAS00001004373.

2.4.18 Code availability

cfSNV can be obtained at https://zhoulab.dgsom.ucla.edu/pages/cfSNV.

Sub-table a. Performance of *cfSNV*, evaluated based on all ground-truth mutations (581)

	cfSNV	***MuTect***	***Strelka2***	***Strelka2 (filters disabled)***
# predicted positives	386	119	149	1643
# true positives	386	118	149	190
# false positives	0	1	0	1453
Sensitivity	64.0%	20.3%	25.6%	32.7%
Precision	100.0%	99.2%	100.0%	11.6%

Sub-table b. Sensitivity of *cfSNV* for mutations at different VAFs

VAF	**# Mutations**	**Sensitivity**			
		cfSNV	***MuTect***	***Strelka2***	***Strelka2 (filters disabled)***
0.10%	116	39.7% (46)	3.4% (4)	0.9% (1)	2.6% (3)
0.30%	116	50.9% (59)	3.4% (4)	0.9% (1)	4.3% (5)
0.50%	58	53.4% (31)	1.7% (1)	0.0% (0)	3.4% (2)
0.80%	57	57.9% (33)	3.5% (2)	0.0% (0)	5.3% (3)
1%	59	74.6% (44)	6.8% (4)	0.0% (0)	3.4% (2)
3%	61	91.8% (56)	41.0% (25)	72.1% (44)	100.0% (61)
5%	53	84.9% (45)	62.3% (33)	88.7% (47)	100.0% (53)

8%	61	95.1% (58)	73.8% (45)	91.8% (56)	100.0% (61)
Total	581	64.0% (386)	20.3% (118)	25.6% (149)	32.7% (149)

Table 2.1: Validation of *cfSNV* on simulation data. Note that because the sequencing errors in the simulation data were less complicated than those in real data, all three methods achieved comparably high precisions. Therefore, here we focused on the comparison of the sensitivities. *Strelka2* had no false positives because its background scoring model uses a high cutoff and hence sacrifices sensitivity. To make a fair comparison with the other methods, we disabled the filters of *Strelka2* in the last column.

Patient ID	Sample ID	Error No Overlap	Error Overlap	Variant No Overlap	Variant Overlap
CRPC_17	SRR6708977	495000	178567	1796509	411606
CRPC_17	SRR6708976	52494	16194	2723733	612499
CRPC_22	SRR6708979	35705	9218	1265653	294224
CRPC_22	SRR6708978	81928	23566	1935107	397078
CRPC_264	SRR6708961	46534	20939	2195553	577018
CRPC_264	SRR6708962	19780	9723	1699284	441753
CRPC_372	SRR6708965	190119	78577	2281386	559180
CRPC_372	SRR6708966	10442	6567	3391413	820014
CRPC_468	SRR6708970	26671	18198	3508685	880805
CRPC_468	SRR6708971	36245	24241	3477492	878561
CRPC_554	SRR6708975	114429	34938	3770965	699477
CRPC_554	SRR6708974	674340	213989	5252738	973627
MBC_191	SRR6708921	210	138	1249492	405881
MBC_191	SRR6708922	1081490	532325	2593057	808558
MBC_284	SRR6708924	23606	11451	4890905	1150491

MBC_284	SRR6708925	156905	62434	2842976	760736
MBC_288	SRR6708927	744548	334731	2957088	950078
MBC_288	SRR6708928	422	486	1934183	742452
MBC_295	SRR6708931	107341	60058	2206167	454093
MBC_295	SRR6708932	2157991	1079179	2892371	687378
MBC_303	SRR6708935	30212	16417	5075765	962065
MBC_303	SRR6708936	189828	114320	2105320	537091
MBC_307	SRR6708937	12730	6766	3274014	664581
MBC_307	SRR6708938	1050679	423942	3800586	907059
MBC_313	SRR6708939	18338	9232	3836598	834756
MBC_313	SRR6708940	1598064	785133	3548035	1056629
MBC_318	SRR6708943	7786	6301	3190647	800833
MBC_318	SRR6708944	504353	424018	2702102	1007410
MBC_325	SRR6708947	426575	295975	2884043	942507
MBC_325	SRR6708948	143538	90622	3336840	921844
MBC_339	SRR6708955	348650	119469	3888221	767321
MBC_339	SRR6708956	340479	114771	3661371	710946
MBC_349	SRR6708957	168731	56234	3864562	860056
MBC_349	SRR6708958	159512	54131	3734487	864607
MBC_331	SRR6708950	479579	274047	4272310	1004417
MBC_331	SRR6708951	924421	587408	2885284	730299

Table 2.2: Sample IDs and number of testing reads extracted. Patient IDs follow the naming convention in reference [AHF17], 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

Table 2.3: 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

Table 2.4: 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

Table 2.5: 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
CIGAR information in a 7-bp window centered on the query site from the mate	categorical	no	yes

Table 2.6: 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.

Sample ID	Sequencing protocol	Total analyzed pairs	Overlapping pairs	Non-overlapping pairs	Overlapping rate
CRPC_161.T1	PE100	133713461	76706008	57007453	57.37%
CRPC_17.T2	PE100	112912255	68762359	44149896	60.90%
CRPC_17.T1	PE100	231379063	116648407	114730656	50.41%
CRPC_22.T2	PE100	69719765	40838217	28881548	58.57%
CRPC_22.T1	PE100	50922107	34962567	15959540	68.66%
CRPC_264.T1	PE100	94489440	49437464	45051976	52.32%
CRPC_264.T2	PE100	76285866	39931380	36354486	52.34%
CRPC_342.T1	PE100	77800434	39983987	37816447	51.39%
CRPC_362.T1	PE100	144733728	90831593	53902135	62.76%
CRPC_372.T1	PE100	95419408	46729963	48689445	48.97%
CRPC_372.T2	PE100	141469008	85545590	55923418	60.47%
CRPC_444.T1	PE100	99823352	58835498	40987854	58.94%
CRPC_463.T1	PE100	64410200	34965310	29444890	54.29%
CRPC_466.T1	PE100	86867442	40093027	46774415	46.15%
CRPC_468.T1	PE100	127076900	73725114	53351786	58.02%
CRPC_468.T2	PE100	125959241	69753527	56205714	55.38%
CRPC_525.T1	PE100	181507284	98384770	83122514	54.20%
CRPC_531.T1	PE100	191525002	110656948	80868054	57.78%
CRPC_554.T2	PE100	172555255	93394611	79160644	54.12%
CRPC_554.T1	PE100	115826043	69821170	46004873	60.28%
MBC_191.T1	PE100	49434221	37173220	12261001	75.20%
MBC_191.T2	PE100	104346049	58194183	46151866	55.77%
MBC_217.T1	PE100	45819537	31484308	14335229	68.71%

MBC_284.T1	PE100	187766846	115972627	71794219	61.76%
MBC_284.T2	PE100	105253472	57912040	47341432	55.02%
MBC_287.T1	PE100	72621076	47398698	25222378	65.27%
MBC_288.T1	PE100	109914978	60307802	49607176	54.87%
MBC_288.T2	PE100	71159260	49173622	21985638	69.10%
MBC_291.T1	PE100	125487153	79432588	46054565	63.30%
MBC_292.T1	PE100	169264847	109287196	59977651	64.57%
MBC_295.T1	PE100	72290837	40661840	31628997	56.25%
MBC_295.T2	PE100	95321907	45817182	49504725	48.07%
MBC_299.T1	PE100	145495818	89226402	56269416	61.33%
MBC_301.T1	PE100	139127700	80183853	58943847	57.63%
MBC_303.T1	PE100	172833875	103629229	69204646	59.96%
MBC_303.T2	PE100	70238383	39263198	30975185	55.90%
MBC_307.T1	PE100	107376735	63850659	43526076	59.46%
MBC_307.T2	PE100	147719840	83423006	64296834	56.47%
MBC_313.T1	PE100	129906737	75086942	54819795	57.80%
MBC_313.T2	PE100	145333019	83038161	62294858	57.14%
MBC_321.T1	PE100	233457402	125507698	107949704	53.76%
MBC_330.T1	PE100	131880037	83083092	48796945	63.00%
MBC_325.T1	PE100	113079532	67731787	45347745	59.90%
MBC_325.T2	PE100	139083004	94962794	44120210	68.28%
MBC_335.T1	PE100	176603512	113906872	62696640	64.50%
MBC_336.T1	PE100	188162046	143807711	44354335	76.43%
MBC_331.T2	PE100	106639523	60706708	45932815	56.93%
MBC_331.T1	PE100	153265273	92462423	60802850	60.33%
MBC_333.T1	PE100	295079422	221021645	74057777	74.90%

MBC_339.T1	PE100	122625928	72668646	49957282	59.26%
MBC_349.T2	PE100	123144300	87299997	35844303	70.89%
MBC_339.T2	PE100	118773552	68464597	50308955	57.64%
MBC_349.T1	PE100	135793407	96439191	39354216	71.02%
MBC_8.T1	PE100	102102243	61739698	40362545	60.47%
MBC_315.T1	PE100	106654039	63979866	42674173	59.99%
MBC_318.T1	PE100	97363967	65247299	32116668	67.01%
MBC_317.T1	PE100	89449800	53075788	36374012	59.34%
MBC_318.T2	PE100	109788289	73645143	36143146	67.08%
MBC_320.T1	PE100	169238177	94487319	74750858	55.83%
1129838	PE150	90493972	60668969	29825003	67.04%
3397799	PE150	98954989	62899357	36055632	63.56%
3736900	PE150	101233709	63795666	37438043	63.02%
4193384	PE150	92216457	56592576	35623881	61.37%
4258357	PE150	77145937	51554475	25591462	66.83%
4325774	PE150	112462632	69365854	43096778	61.68%
4492669	PE150	92722035	62999558	29722477	67.94%
4496246	PE150	53272682	36793304	16479378	69.07%
4514025	PE150	116818638	73998376	42820262	63.34%
4528560	PE150	73080877	47761989	25318888	65.35%
4532964	PE150	95269655	62118982	33150673	65.20%
4536877	PE150	84576575	53857762	30718813	63.68%
4545410	PE150	105795081	72662041	33133040	68.68%
4561279	PE150	125089513	80524552	44564961	64.37%
4563728	PE150	31166243	22383526	8782717	71.82%
4583975	PE150	117726231	69731179	47995052	59.23%

4599369	PE150	100124756	63787421	36337335	63.71%
2163573	PE150	94058875	59201901	34856974	62.94%
4650336	PE150	92495802	54860077	37635725	59.31%
2510880	PE150	70282777	41196303	29086474	58.62%
4390360	PE150	84718207	53665469	31052738	63.35%
4471067	PE150	84266052	47593768	36672284	56.48%
4566326	PE150	97320644	55114291	42206353	56.63%
4582920	PE150	59732944	33126627	26606317	55.46%
4612584	PE150	95861797	57164638	38697159	59.63%
4562675	PE150	87382996	49998703	37384293	57.22%
2450596	PE150	81592146	49147981	32444165	60.24%
4335068	PE150	87083450	47799248	39284202	54.89%
4526552	PE150	93363300	63106874	30256426	67.59%
4411770	PE150	56819893	37203845	19616048	65.48%
4580642	PE150	113976400	76878551	37097849	67.45%
4637842	PE150	69840351	44686904	25153447	63.98%

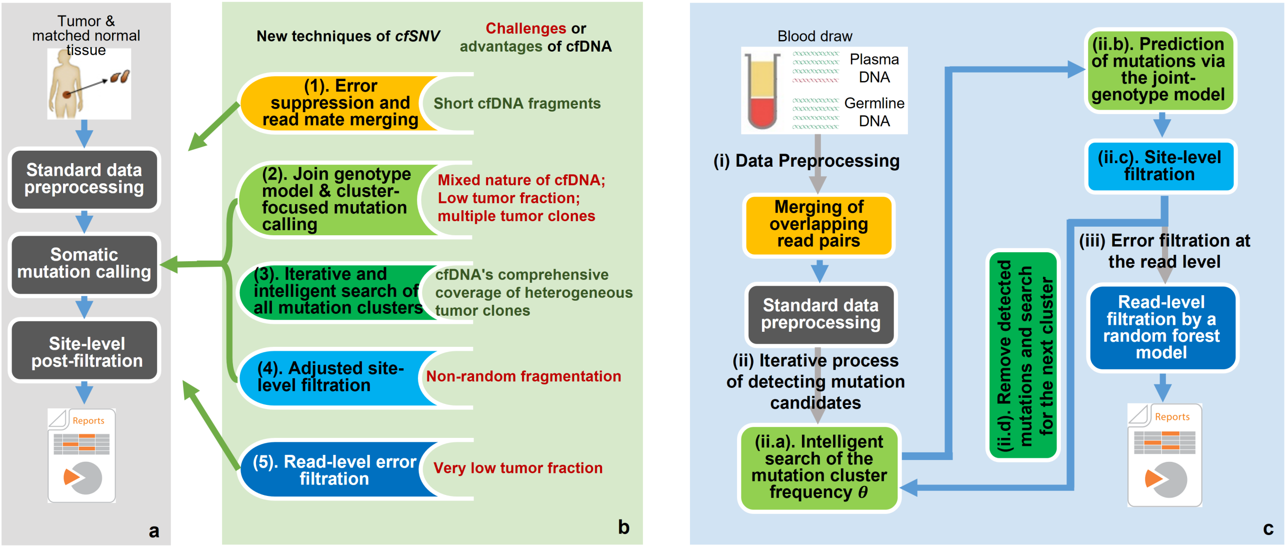
Table 2.7: Statistics of overlapping read pairs in the cfDNA samples.

Sample ID	Time gap (days)	Truncal mutation confirmation rate	Branch mutation confirmation rate
CRPC_17.T2	24	100.00%	96.70%
CRPC_17.T1	24	100.00%	90.00%
CRPC_22.T2	21	93.10%	88.20%
CRPC_22.T1	21	100.00%	92.20%
CRPC_264.T1	28	100.00%	97.00%
CRPC_264.T2	28	100.00%	96.30%

CRPC_372.T1	62	100.00%	98.50%
CRPC_372.T2	62	100.00%	95.20%
CRPC_468.T1	14	99.00%	98.90%
CRPC_468.T2	14	99.00%	99.60%
CRPC_554.T2	138	96.30%	85.10%
CRPC_554.T1	138	100.00%	79.20%
MBC_191.T1	71	97.80%	93.20%
MBC_191.T2	71	72.70%	89.40%
MBC_284.T1	51	100.00%	93.30%
MBC_284.T2	51	100.00%	91.90%
MBC_288.T1	39	90.30%	97.60%
MBC_288.T2	39	97.40%	98.10%
MBC_295.T1	56	100.00%	91.10%
MBC_295.T2	56	92.90%	77.60%
MBC_303.T1	55	95.90%	91.40%
MBC_303.T2	55	100.00%	89.30%
MBC_307.T1	42	100.00%	96.40%
MBC_307.T2	42	96.20%	96.20%
MBC_313.T1	21	90.90%	93.10%
MBC_313.T2	21	92.90%	91.80%
MBC_318.T1	35	100.00%	90.90%
MBC_318.T2	35	100.00%	95.40%
MBC_325.T1	56	91.70%	81.10%
MBC_325.T2	56	100.00%	96.00%
MBC_331.T1	22	93.80%	97.50%
MBC_331.T2	22	96.40%	95.00%

MBC_339.T1	37	100.00%	97.60%
MBC_339.T2	37	100.00%	96.00%
MBC_349.T1	37	100.00%	98.10%
MBC_349.T2	37	100.00%	97.10%

Table 2.8: The plasma confirmation rate of the truncal and branch mutations in the validation patient cfDNA data and the time gap of the plasma collection between two time points.


Tumor & matched normal tissue
Standard data preprocessing
Somatic mutation calling
Site-level post-filtration
Reports
a
New techniques of cfSNV
Challenges or advantages of cfDNA
(1). Error suppression and read mate merging
Short cfDNA fragments
(2). Join genotype model & cluster-focused mutation calling
Mixed nature of cfDNA; Low tumor fraction; multiple tumor clones
(3). Iterative and intelligent search of all mutation clusters
cfDNA's comprehensive coverage of heterogeneous tumor clones
(4). Adjusted site-level filtration
Non-random fragmentation
(5). Read-level error filtration
Very low tumor fraction
b
Blood draw
Plasma DNA
Germline DNA
(i) Data Preprocessing
Merging of overlapping read pairs
Standard data preprocessing
(ii) Iterative process of detecting mutation candidates
(ii.a). Intelligent search of the mutation cluster frequency θ
(ii.b). Prediction of mutations via the joint-genotype model
(ii.c). Site-level filtration
(ii.d). Remove detected mutations and search for the next cluster
(iii) Error filtration at the read level
Read-level filtration by a random forest model
Reports
c

Conventional SNV caller **Novelty of *cfSNV*** ***cfSNV***

Figure 2.1: *cfSNV* framework and its novel techniques. **a.** The workflow of conventional SNV callers takes the genomic data of a tumor and its matched normal tissue as inputs. **b.** Five new techniques introduced to *cfSNV* that modify the standard workflow. **c.** Full workflow of *cfSNV*. *cfSNV* takes plasma DNA and germline DNA sequencing data as inputs. It first merges overlapping read pairs in cfDNA sequencing data. Next, we apply standard data preprocessing tools. An iterative procedure then detects mutation clusters and estimates their frequencies θ based on multiple, automatically selected hotspots. Each iteration determines joint genotypes across sequencing regions to predict somatic SNV candidates, and masks the mutation candidates before proceeding. After all clusters and mutation candidates have been detected, a random forest classifier identifies raw read pairs with sequencing errors. Finally, somatic SNVs are reported only if enough variant supporting read pairs pass the random forest screening. The background color of steps in c corresponds to the feature listed in **b**.

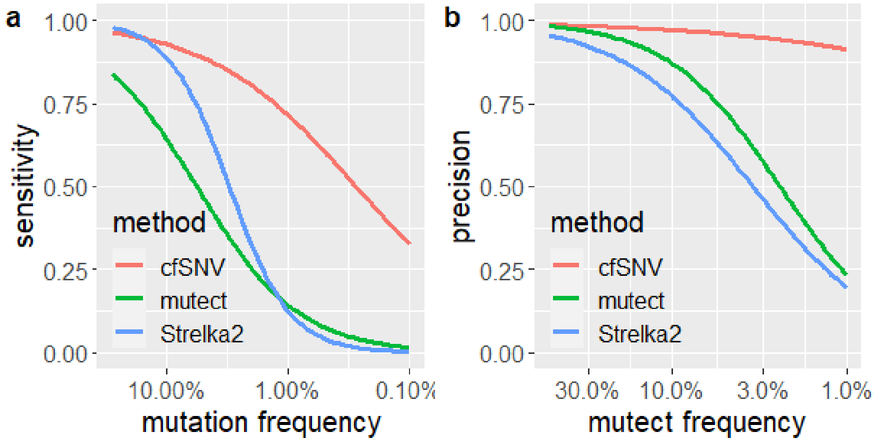


Figure 2.2: *cfSNV* outperforms competing methods in sensitivity and precision, especially for low-frequency mutations. **a.** The sensitivity of three variant calling methods on simulation data as a function of VAF for *cfSNV*, *MuTect* and *Strelka2*. Mutations were grouped based on their simulated VAF, and the sensitivity at each simulated VAF level was calculated separately. The precision of all three methods remained at comparable and high level (Table 2.1). **b.** The precision of three variant calling methods on patient data as a function of VAF. Mutations detected from all samples were grouped based on their rounded VAF (two decimal places). The precision at each VAF level was estimated by the confirmation rate. The sensitivity of patient data cannot be quantified because of the unknown ground truth, but *cfSNV* detected the most true positive mutations. All curves were fitted using logit functions.

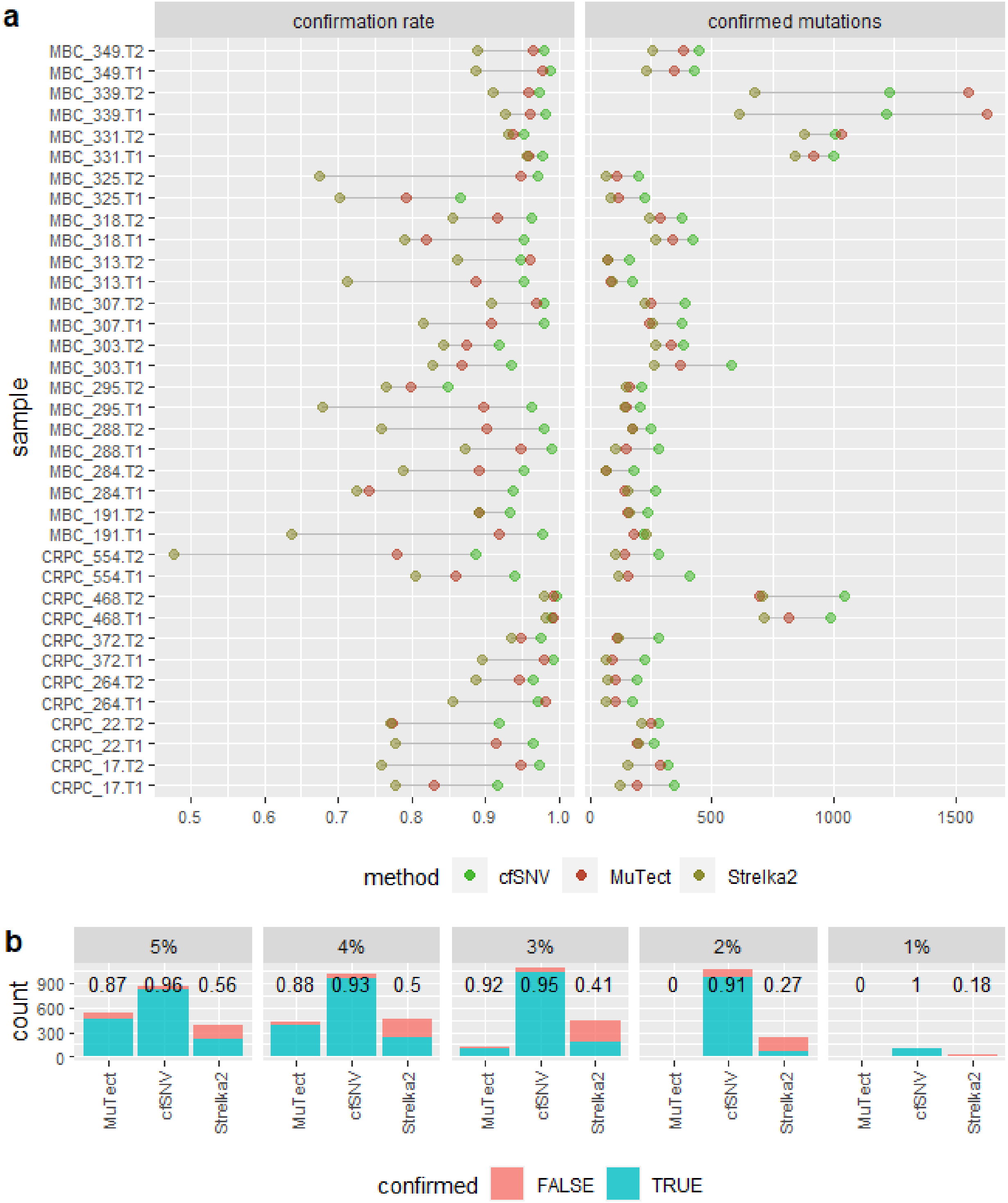

a
confirmation rate
confirmed mutations
sample
MBC_349.T2
MBC_349.T1
MBC_339.T2
MBC_339.T1
MBC_331.T2
MBC_331.T1
MBC_325.T2
MBC_325.T1
MBC_318.T2
MBC_318.T1
MBC_313.T2
MBC_313.T1
MBC_307.T2
MBC_307.T1
MBC_303.T2
MBC_303.T1
MBC_295.T2
MBC_295.T1
MBC_288.T2
MBC_288.T1
MBC_284.T2
MBC_284.T1
MBC_191.T2
MBC_191.T1
CRPC_554.T2
CRPC_554.T1
CRPC_468.T2
CRPC_468.T1
CRPC_372.T2
CRPC_372.T1
CRPC_264.T2
CRPC_264.T1
CRPC_22.T2
CRPC_22.T1
CRPC_17.T2
CRPC_17.T1
0.5
0.6
0.7
0.8
0.9
1.0
0
500
1000
1500
method
cfSNV
MuTect
Strelka2
b
5%
4%
3%
2%
1%
count
900
600
300
0
0.87
0.96
0.56
0.88
0.93
0.5
0.92
0.95
0.41
0
0.91
0.27
0
1
0.18
MuTect
cfSNV
Strelka2
confirmed
FALSE
TRUE

Figure 2.3: Somatic SNV calling on cfDNA sequencing samples from cancer patients. a, Total number of confirmed mutations and precision (confirmation rate) using *cfSNV*, *MuTect* and *Strelka2*. The precision is the number of confirmed mutations divided by the total number of reported mutations. In the sample name, “T1” and “T2” indicate the first time point and the second time point of blood plasma samples respectively. b, The total number of low-frequency variants and their confirmation status found by *cfSNV*, *MuTect* and *Strelka2* from all plasma samples. Low-frequency variants are divided into five groups according to their rounded VAF, and the number of confirmed and unconfirmed mutations for each variant group are plotted in five subfigures for comparing between our method and two competing methods. The number at the top of each bar indicates the precision (the confirmation rate).

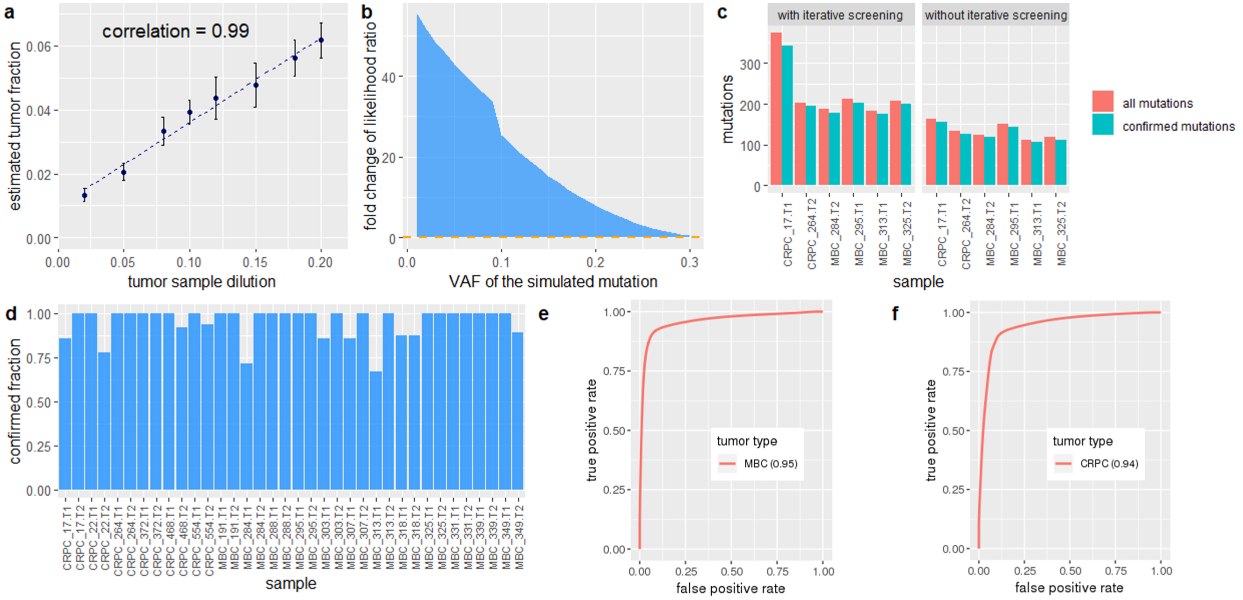


Figure 2.4: Experimental analysis of five new techniques. **a.** Performance of mutation cluster frequency estimation in terms of the correlation between the estimated tumor fraction and the true dilution ratio. This experiment uses simulated data based on WES of a single patient, with dilution ratios ranging from 2% to 20%. The points are the means ± s.d. of five independently generated datasets at each dilution. **b.** the fold change in the likelihood ratio between *cfSNV* models with and without a step to estimate the mutation cluster frequency, based on simulated mutations at different VAFs. **c.** Number of mutations detected with and without the iterative screening procedure. **d.** Confirmation rate of rescued mutations after adjusting conventional site-level post-filtration. **e-f.** Performance of read-level variant classifier on testing data. **e.** The averaged ROC of applying the classifier to labeled data taken from 24 cfDNA sequencing samples of 12 metastatic breast cancer patients. **f.** The averaged ROC of applying the classifier to labeled data taken from 12 cfDNA sequencing samples of 6 metastatic prostate cancer patients. The numbers in parentheses indicate the area under curve (AUC) metric.

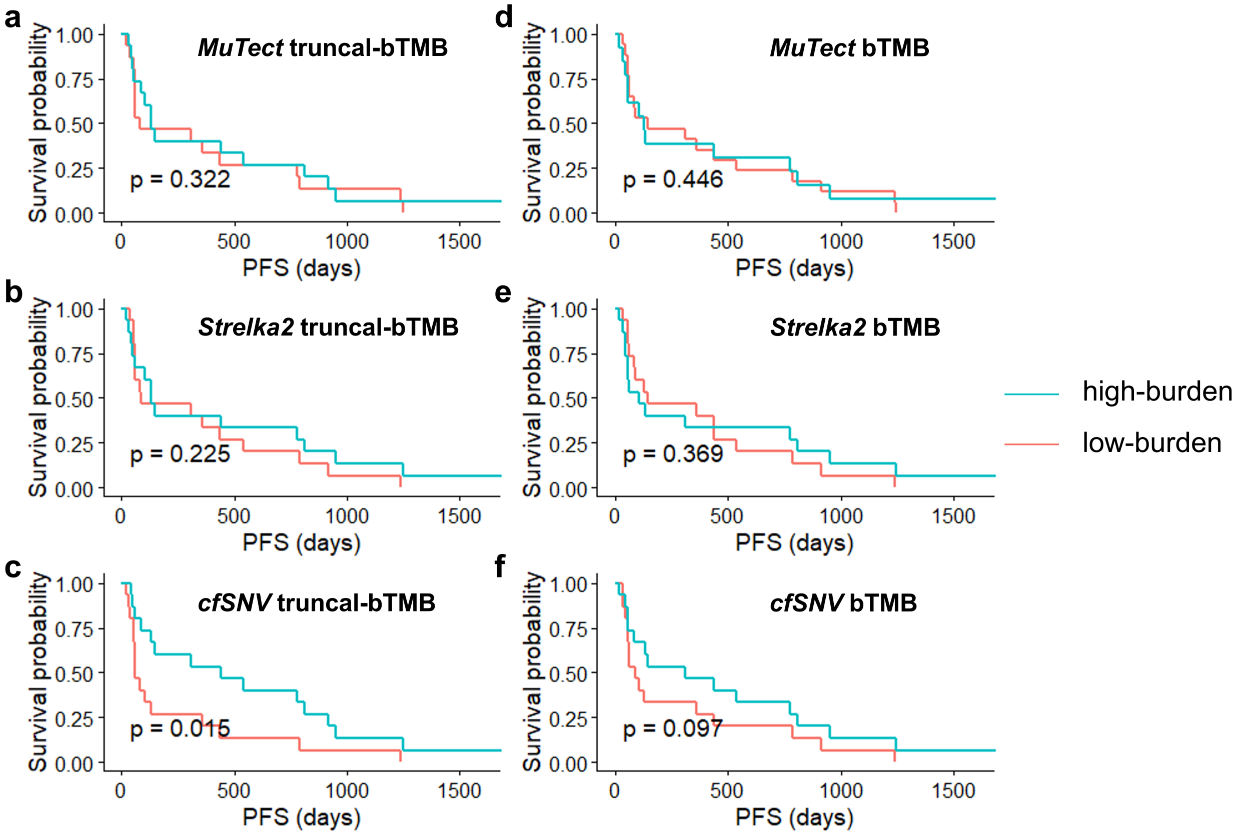


Figure 2.5: Kaplan-Meier curves for progression-free survival (PFS) on the pre-treatment cfDNA sequencing data of 30 advanced non-small cell lung cancer patients. **a-c**, Kaplan-Meier curves based on truncal-bTMB calculated using *MuTect*, *Strelka2* and *cfSNV*. The high-burden and low-burden groups in each plot are defined by the median value of the measure: *MuTect* (**a**, Hazard Ratio (HR)=0.839, 95% confidence interval (CI) [0.403, 1.747]), *Strelka2* (**b**, HR=0.745, 95% CI [0.352, 1.581]), or *cfSNV* (**c**, HR=0.438, 95% CI [0.205, 0.938]). **d-f**, Kaplan-Meier curves based on bTMB calculated using *MuTect*, *Strelka2* and *cfSNV*. The high-burden and low-burden groups in each plot are defined by the median value of the measure: *MuTect* (**d**, HR=0.948, 95% CI [0.451, 1.990]), *Strelka2* (**e**, HR=0.883, 95% CI [0.415, 1.880]), or *cfSNV* (**f**, HR=0.611, 95% CI [0.288, 1.295]).

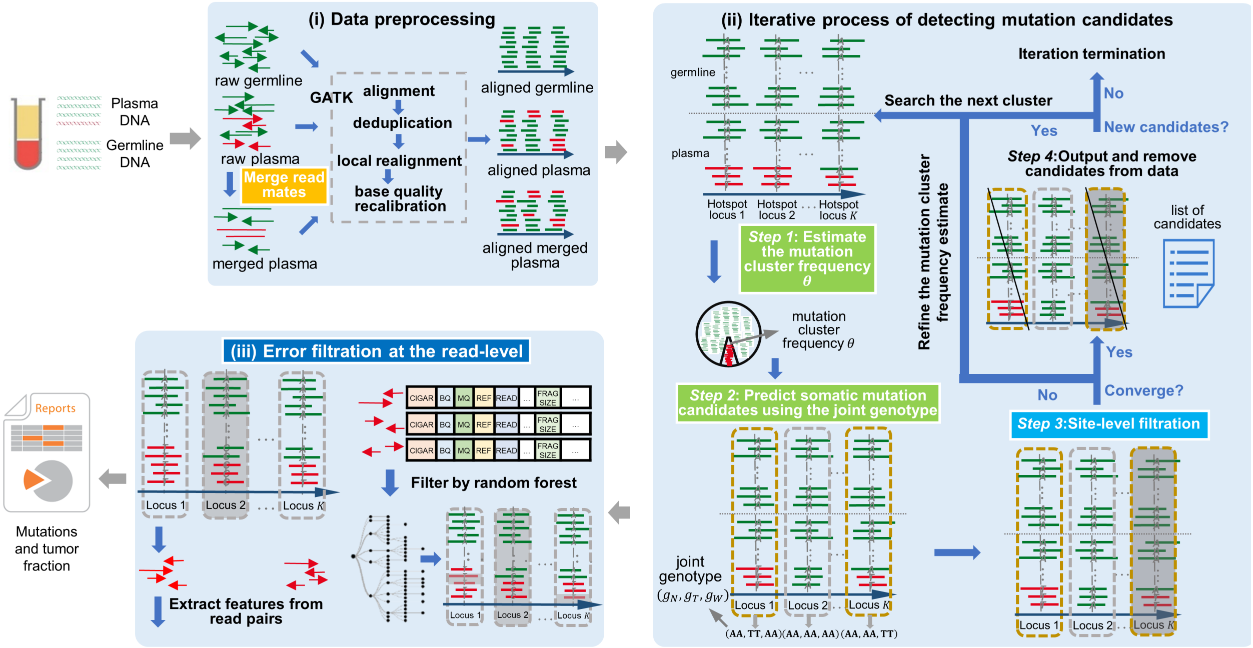

Plasma DNA
Germline DNA
(i) Data preprocessing
raw germline
raw plasma
Merge read mates
merged plasma
GATK
alignment
deduplication
local realignment
base quality recalibration
aligned germline
aligned plasma
aligned merged plasma
(ii) Iterative process of detecting mutation candidates
germline
plasma
Hotspot locus 1
Hotspot locus 2
Hotspot locus K
Step 1: Estimate the mutation cluster frequency θ
mutation cluster frequency θ
Step 2: Predict somatic mutation candidates using the joint genotype
joint genotype
(g_N, g_T, g_W)
Locus 1
Locus 2
Locus K
(AA, TT, AA)(AA, AA, AA)(AA, AA, TT)
Step 3:Site-level filtration
Converge?
No
Yes
Step 4:Output and remove candidates from data
list of candidates
New candidates?
Yes
No
Iteration termination
Search the next cluster
Refine the mutation cluster frequency estimate
(iii) Error filtration at the read-level
Extract features from read pairs
CIGAR
BQ
MQ
REF
READ
FRAG SIZE
Filter by random forest
Reports
Mutations and tumor fraction

Figure 2.6: Workflow of *cfSNV*. *cfSNV* takes plasma DNA and germline DNA sequencing data as inputs. It first merges overlapping read mates in cfDNA sequencing data. The reads are processed using the *GATK* pipeline. After these steps, an iterative procedure estimates the mutation cluster frequency θ based on a set of carefully selected sample-specific hotspots. Each iteration step determines the joint tumor-normal genotypes across sequencing regions, then eliminates somatic SNV candidates that fail essential filters based on site-level statistics (Methods). Mutation candidates are used as hotspot sites to refine θ and candidate detection until the frequency converges. SNV candidates from the previous iteration are output and masked before the next iteration. After all candidates are detected, a random forest classifier identifies raw read pairs with sequencing errors. Finally, somatic SNVs are reported only if enough variant supporting read pairs passed the random forest screening.

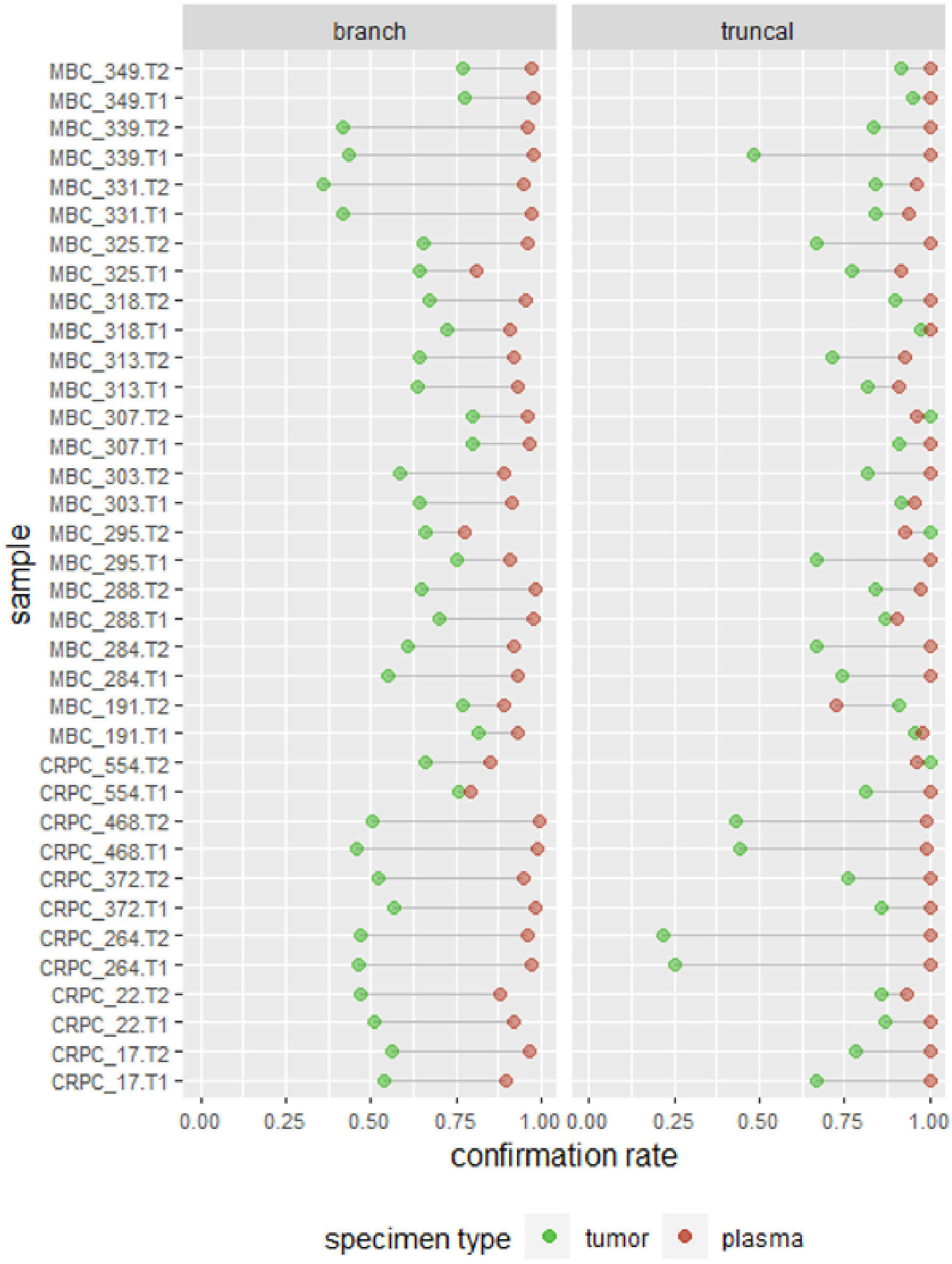


Figure 2.7: Fraction of confirmed truncal mutations and branch mutations detected by *cfSNV* on patient data. Mutations found in cfDNA sequencing data were validated by variant supporting read counts, either in cfDNA sequencing data from the other plasma sample or in genomic DNA sequencing data from a tumor biopsy sample collected from the same patient. The clonality of mutations was determined by their relative VAFs.

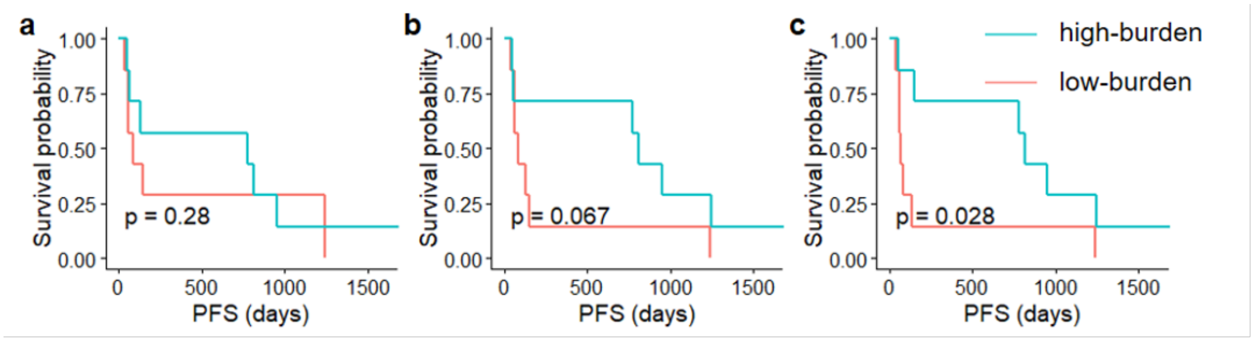


Figure 2.8: Kaplan-Meier curves for progression-free survival (PFS) on advanced non-small cell lung cancer patients. **a-c**, PFS for 14 patients with both tumor biopsy and pre-treatment cfDNA sequencing data. The high-burden and low-burden groups in each plot are defined by the median value of the measure: TMB (**a**, HR=0.721, 95% CI [0.239, 2.173]), bTMB (**b**, HR=0.411, 95% CI [0.124, 1.355]), or truncal-bTMB (**c**, HR=0.326, 95% CI [0.098, 1.079]).

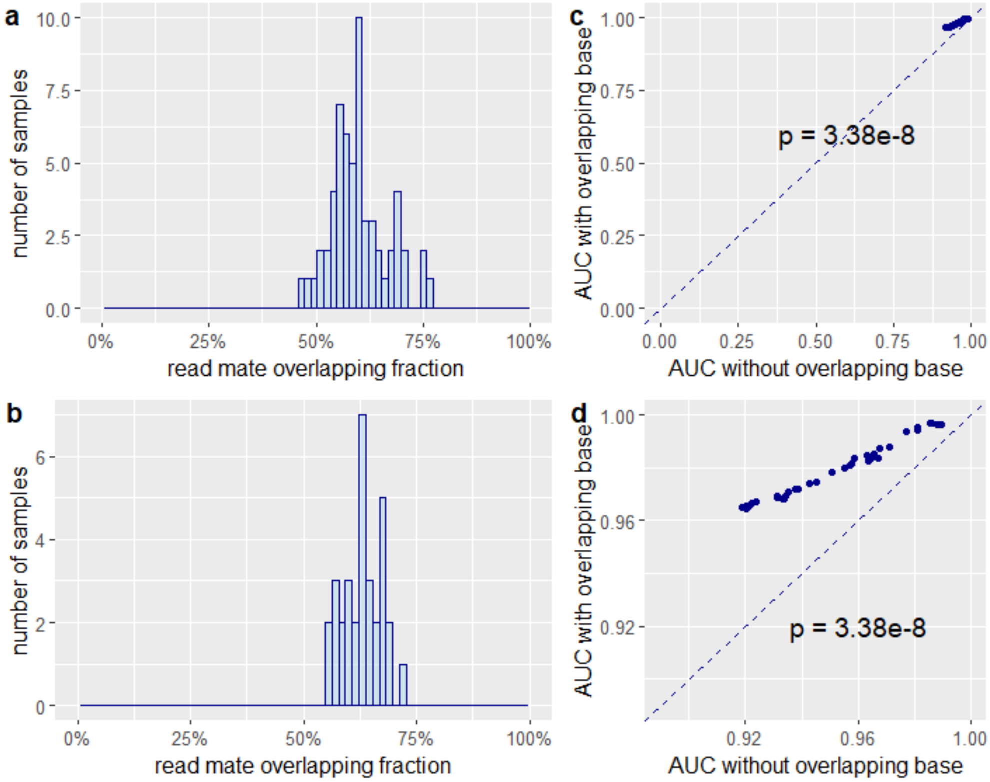


Figure 2.9: Quantifying the existence and impact of overlapping read mates in cfDNA sequencing data. **a**, Fraction of merged overlapping read mates in 59 cfDNA whole exome sequencing samples from metastatic cancer patients (paired-end 2x150bp). **b**, Fraction of merged overlapping read mates in 30 cfDNA whole exome sequencing samples from NSCLC patients (paired-end 2x150bp). **c**,Comparison of AUC metrics from classifiers trained on overlapping read pairs and non-overlapping read pairs on 36 testing samples. **d**, A zoom of **c**.

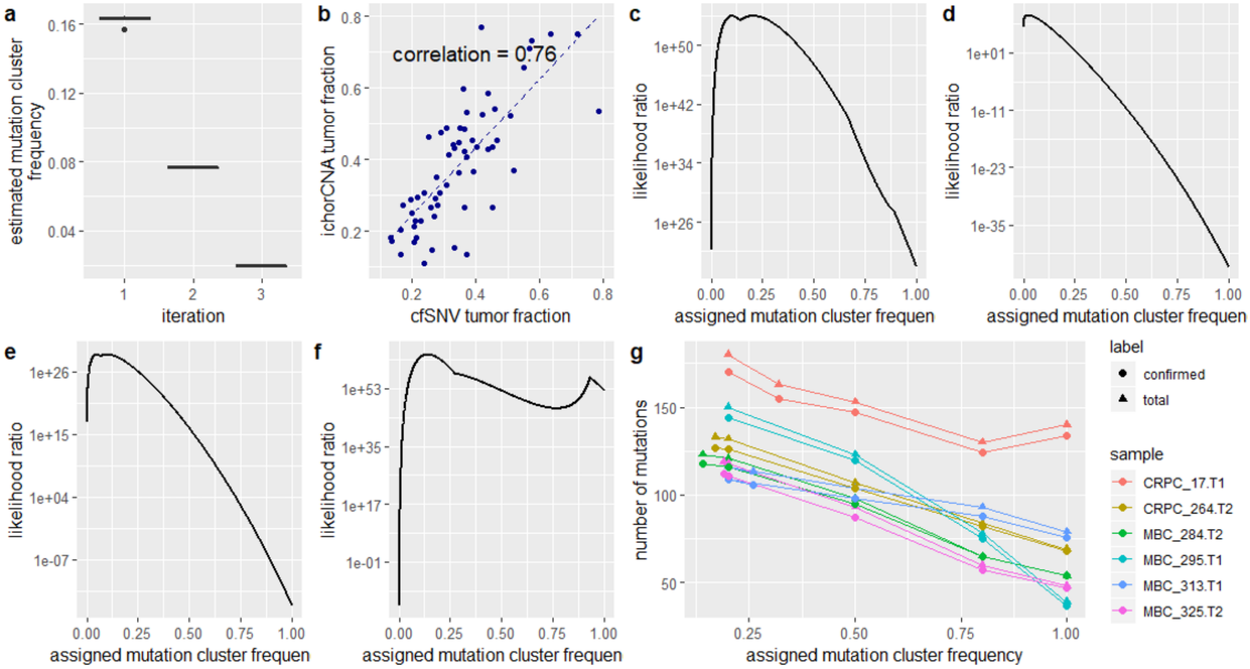


Figure 2.10: Simulation on the estimated mutation cluster frequency. **a**, Performance of mutation cluster frequency estimation in the first experiment with simulated samples containing purely synthetic mutations inserted at known VAF levels (20%, 8% and 2%). Each box in this plot shows the estimated 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 $\leq 20\%$.The left most point on each line showed the number of mutations detected at the estimated mutation cluster frequency.

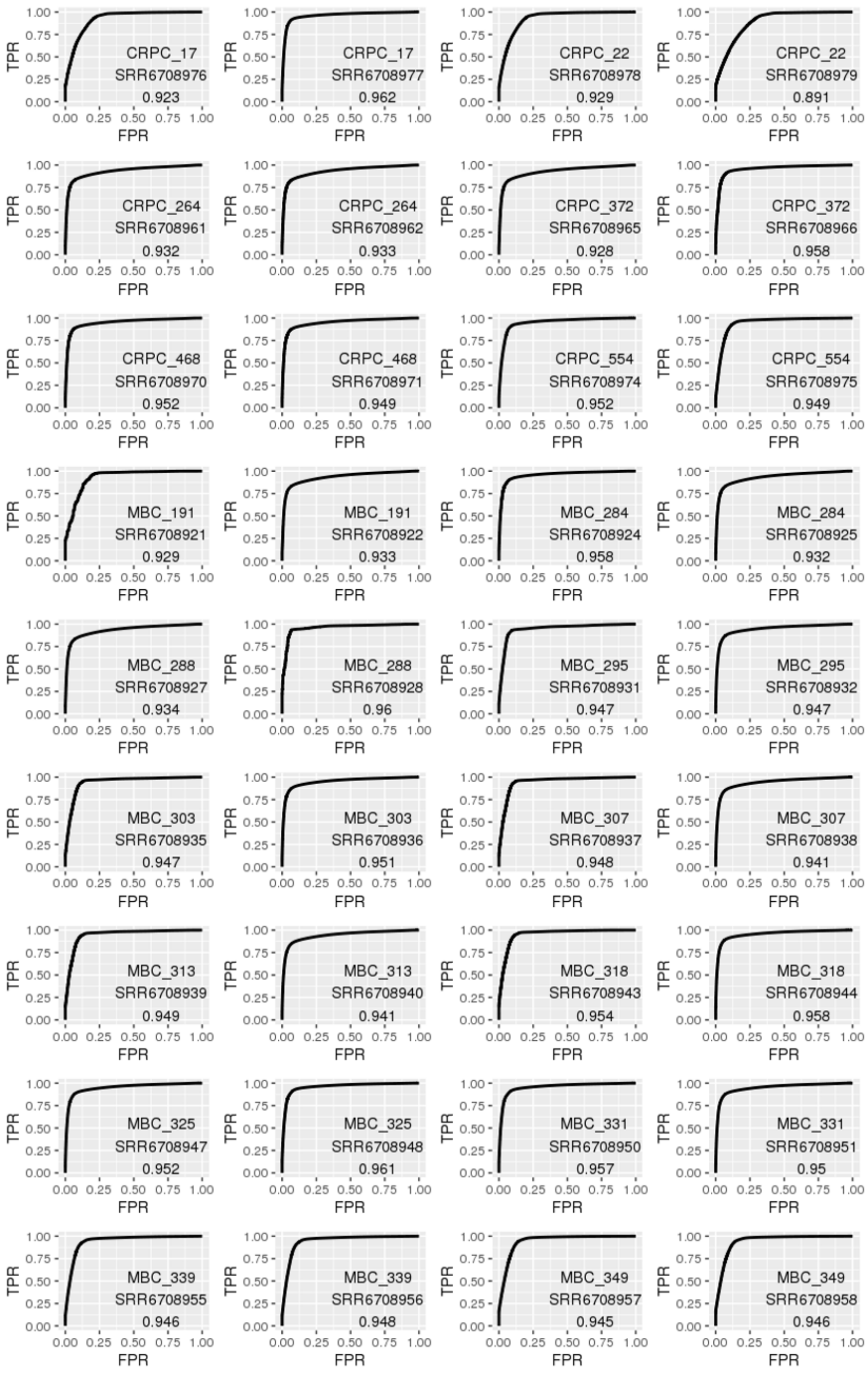


Figure 2.11: ROC curves for random forest classifiers on all read pairs in out-of-sample tests. The classifiers were trained using data derived from only WES data of cfDNA sample from a single patient (patient MBC_315, sample SRR6708941). Each independent testing dataset (from one patient) has its own ROC curve. The numbers in parentheses are area under curve (AUC) metrics.

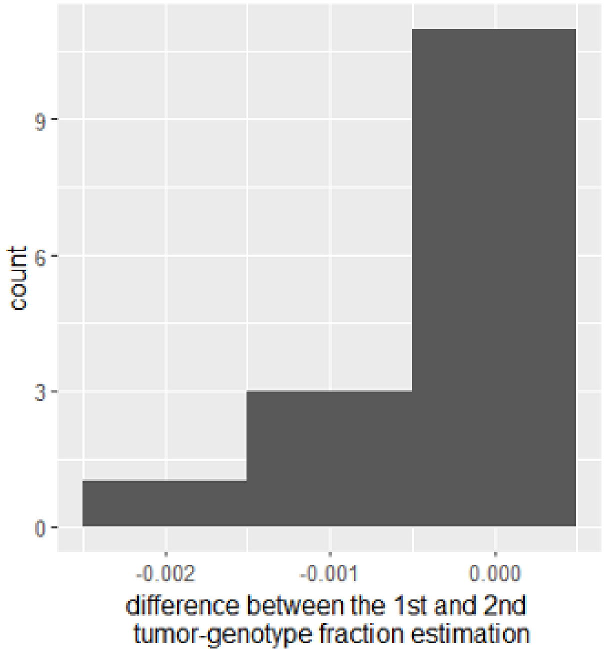


Figure 2.12: Difference between the tumor-genotype fractions estimated in the first round and the second round (refined by mutation candidates).

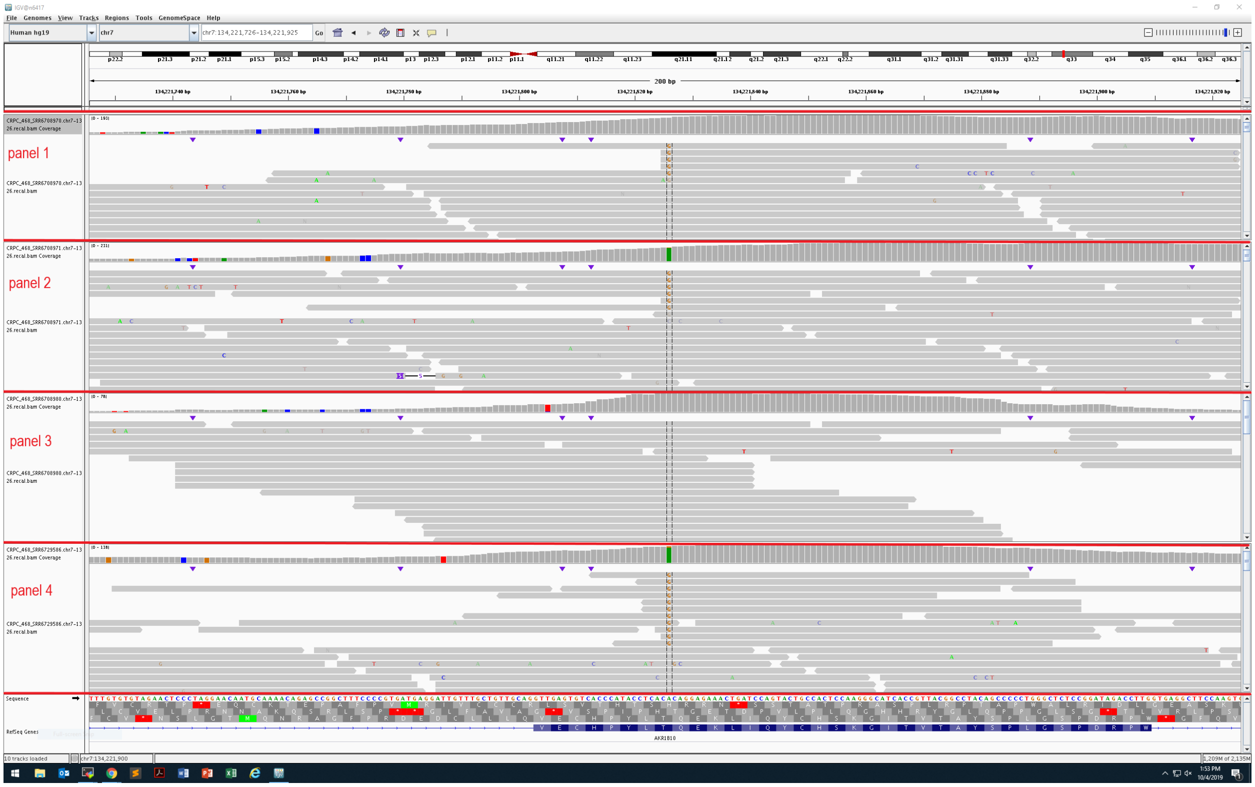

IGV@n6417
File Genomes View Tracks Regions Tools GenomeSpace Help
Human hg19
chr7
chr7:134,221,726-134,221,925
Go
200 bp
134,221,740 bp
134,221,760 bp
134,221,780 bp
CRPC_468_SRR6708970.chr7-1326.recal.bam Coverage
panel 1
CRPC_468_SRR6708970.chr7-1326.recal.bam
CRPC_468_SRR6708971.chr7-1326.recal.bam Coverage
panel 2
CRPC_468_SRR6708971.chr7-1326.recal.bam
CRPC_468_SRR6708980.chr7-1326.recal.bam Coverage
panel 3
CRPC_468_SRR6708980.chr7-1326.recal.bam
CRPC_468_SRR6729586.chr7-1326.recal.bam Coverage
panel 4
CRPC_468_SRR6729586.chr7-1326.recal.bam
Sequence
RefSeq Genes
AKR1B10
10 tracks loaded

Figure 2.13: 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.

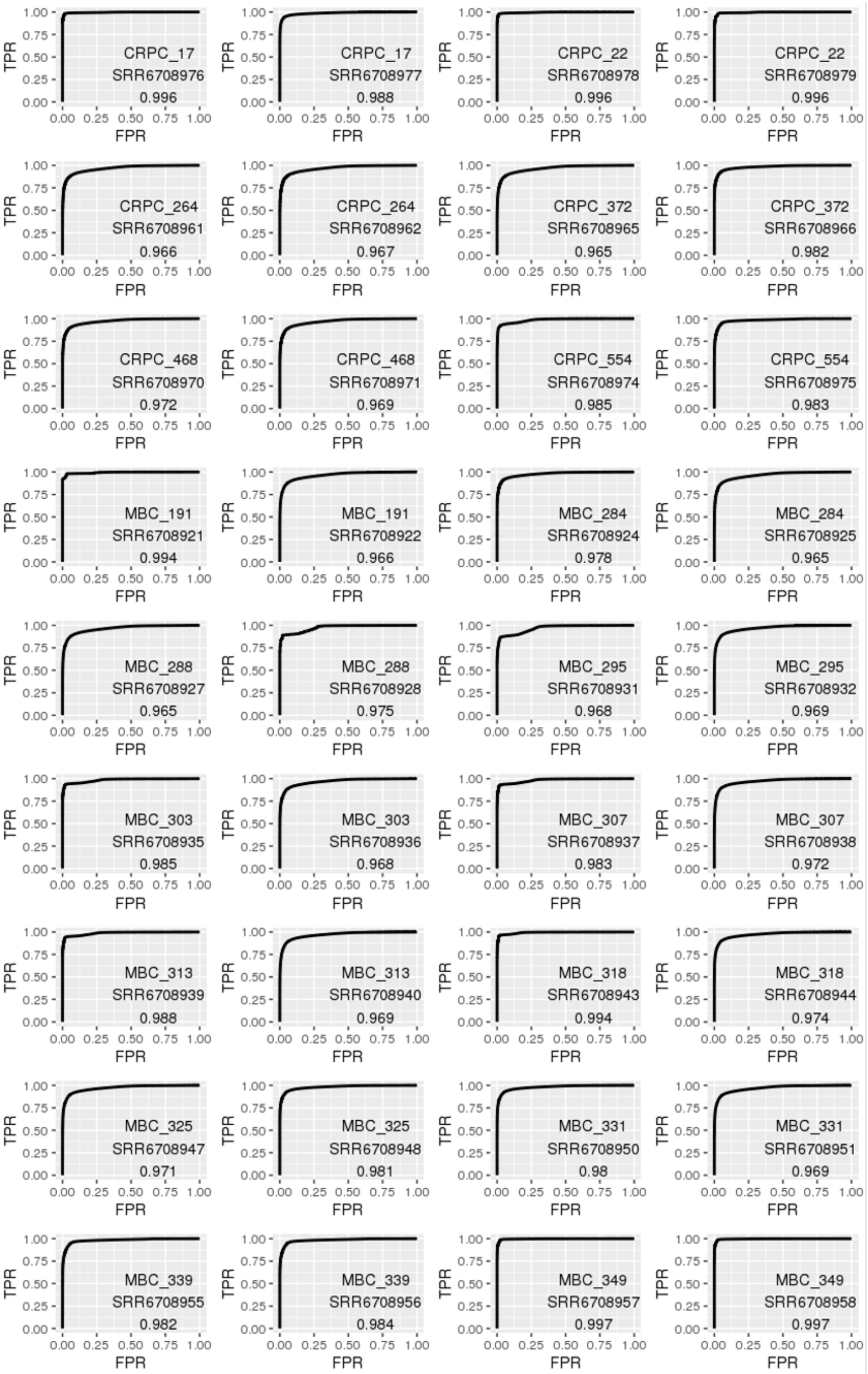


Figure 2.14: ROC curves for random forest classifiers on overlapping read pairs only in out-of-sample tests. The classifiers were trained using data derived from only WES data of cfDNA sample from a single patient (patient MBC_315, sample SRR6708941). Each independent testing dataset (from one patient) has its own ROC curve. The numbers in parentheses are area under curve (AUC) metrics.

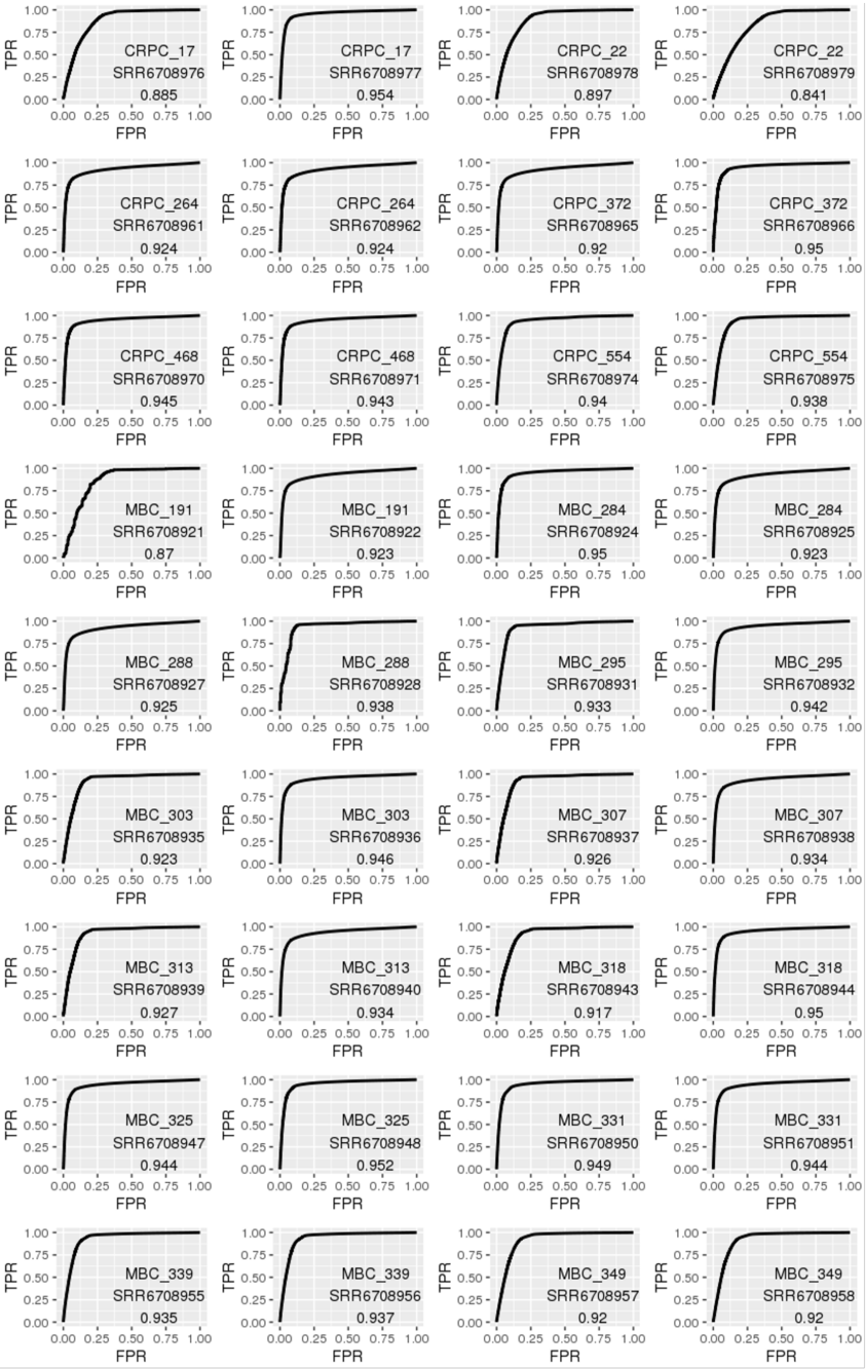


Figure 2.15: ROC curves for random forest classifiers on non-overlapping read pairs only in out-of-sample tests. The classifiers were trained using data derived from only WES data of cfDNA sample from a single patient (patient MBC_315, sample SRR6708941). Each independent testing dataset (from one patient) has its own ROC curve. The numbers in parentheses are area under curve (AUC) metrics.

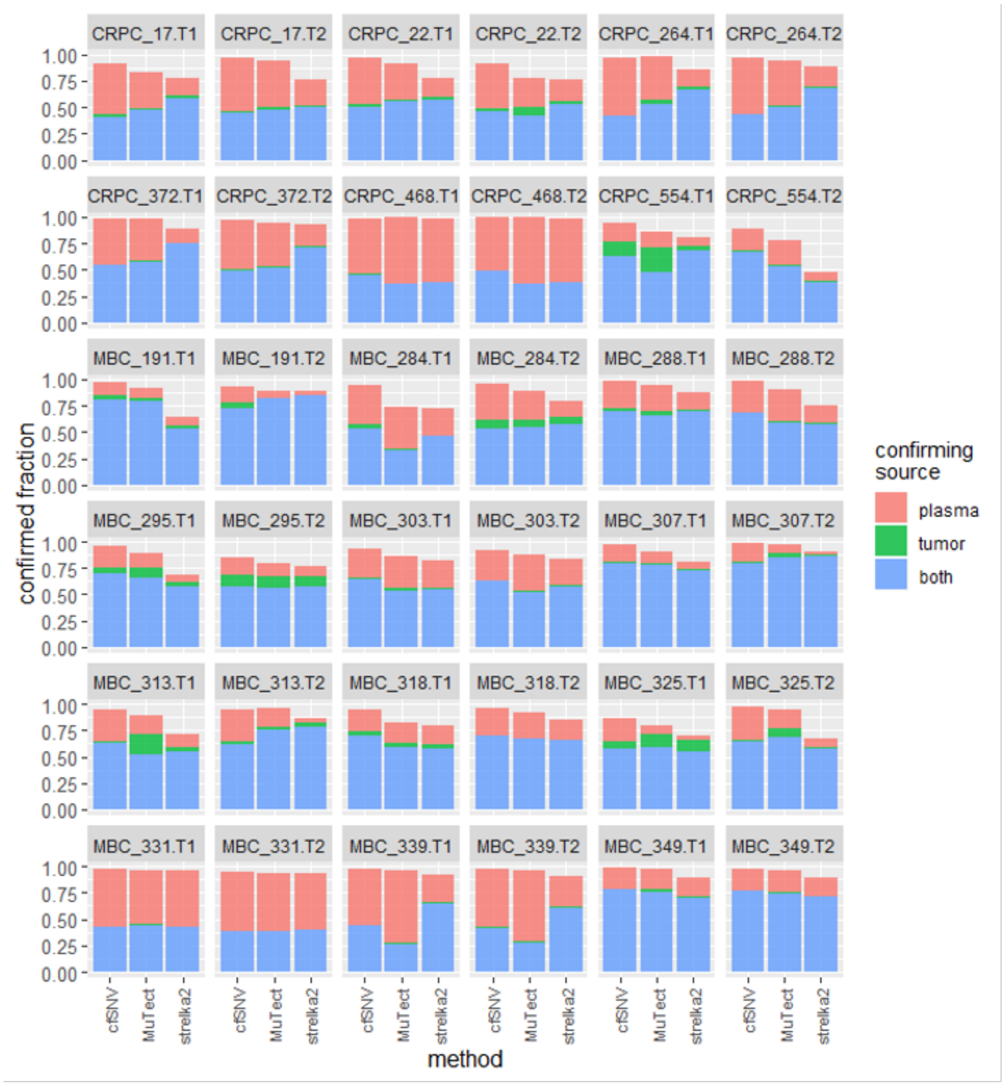


Figure 2.16: Confirmed fractions of somatic SNVs from *cfSNV*, *MuTect* andStrelka2, along with confirming sources.

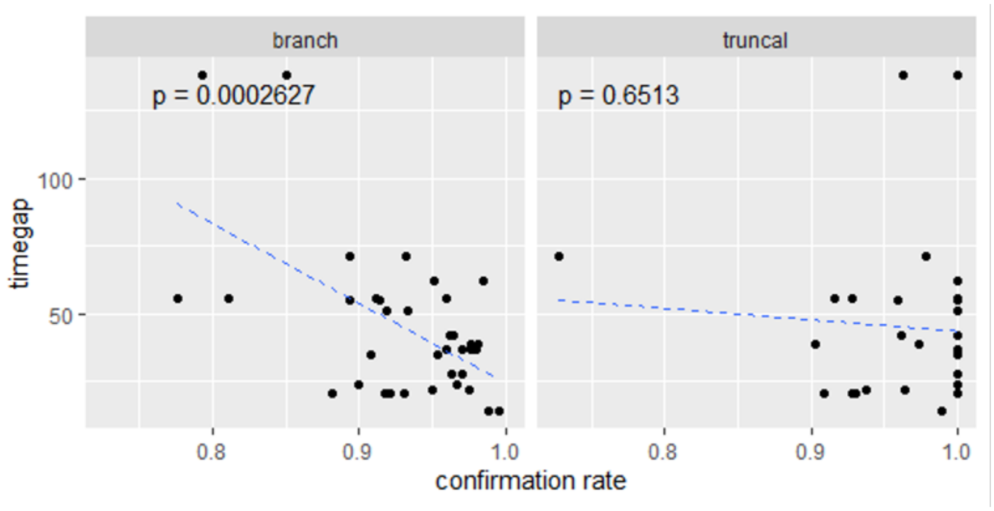


Figure 2.17: Correlation on the truncal and branch mutation confirmation rates in plasma samples with respect to the sample collecting time gap.

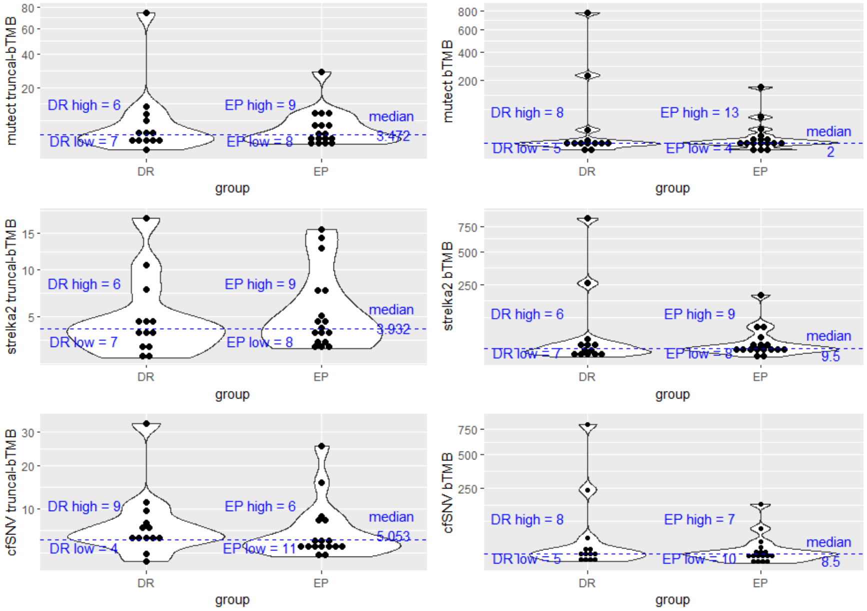


Figure 2.18: Distribution of bTMB and truncal-bTMB in the 30 NSCLC patients. The durable responders (DR, PFS > 9 months) and early progressors (EP, PFS < 6 months) are defined based on the outcome of the patients, i.e. progression-free survival. The cutoff for the two patient groups based on bTMB and truncal-bTMB is marked as the dashed blue line.

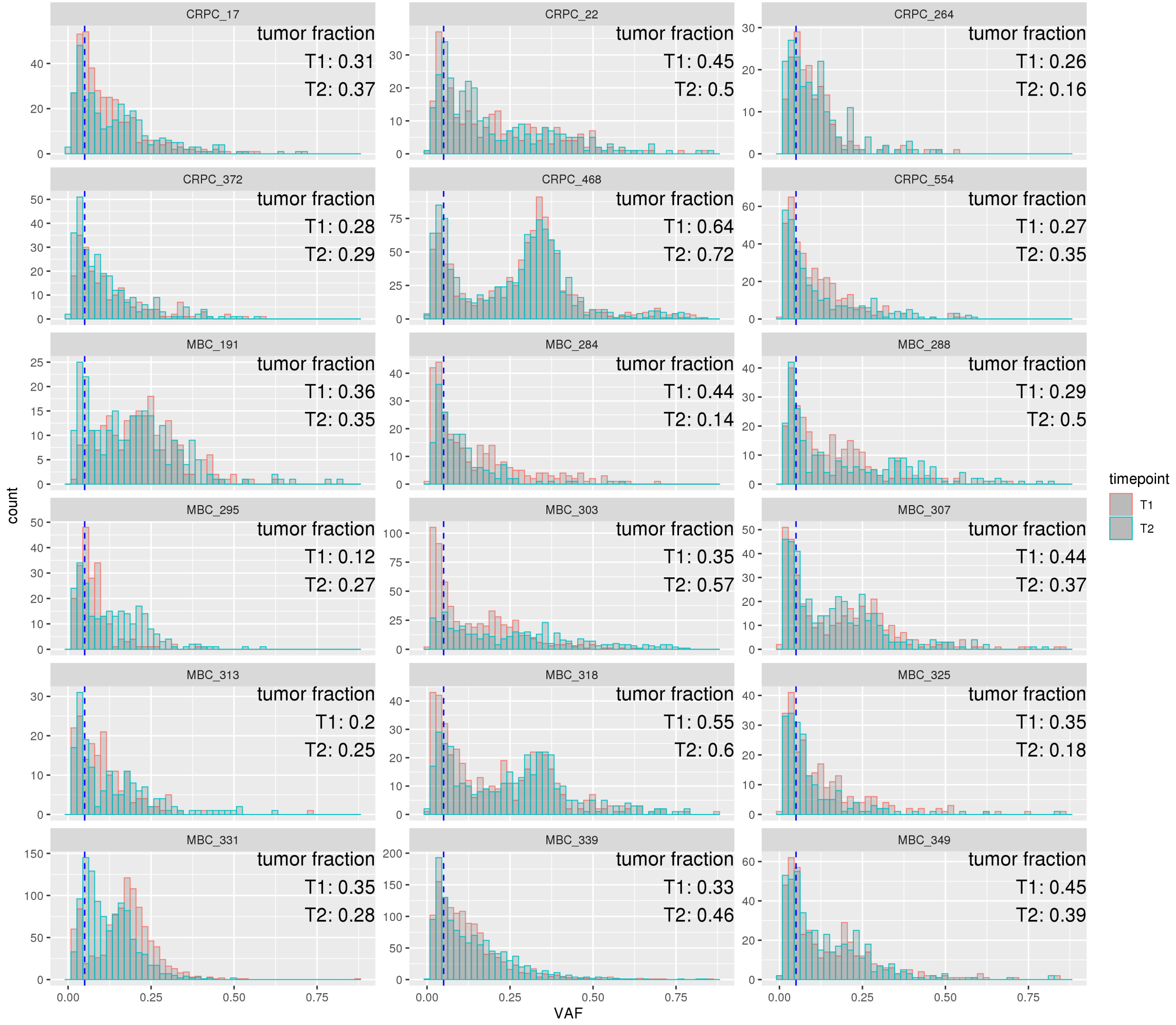


Figure 2.19: Distribution of variant allele frequency of somatic mutations detected by *cfSNV* in each sample in the validation patient cfDNA data. The dashed line marks VAF = 0.05.

CHAPTER 3

OncoMonitor: noninvasive monitoring of MRD and progression by comprehensive tumor mutation analysis in plasma cfDNA

3.1 Introduction

Despite the rapid development of cancer treatment, a large fraction of cancer patients develop recurrence, resistance, or progression during or after treatment [MLG18]. Even with the surgical removal of tumors, there could still be minimal residual disease (MRD), which is associated with an increased likelihood of the disease returning after treatment [CCL17]. Thus, monitoring cancer patients for the early detection of MRD, cancer recurrence and progression is essential to assess the response and detect relapse. This in turn could facilitate early intervention and the personalization of adjuvant therapies and most importantly improve the quality of life of cancer patients [CCL17] [KR14]. Although cancer monitoring is clinically important, it often requires sequential sampling of the tumor from the patient, which poses a difficult challenge toward traditional tumor biopsy. In this context, liquid biopsy provides attractive options, especially the option of using cell-free DNA (cfDNA) in blood, because blood can be obtained noninvasively, and tumor DNA in cfDNA can provide comprehensive genetic profiles of heterogeneous tumors [MDP15].

However, a major challenge associated with cfDNA-based cancer monitoring is the often very low tumor content in cfDNA. The fraction of tumor DNA can be as low as 0.1% in a cfDNA sample from cancer patients receiving treatment or with MRD [ABW17]. Previous studies on cancer monitoring used deep sequencing on a small mutation panel to discover the weak tumor signal in the plasma [CCL17] [ABW17] [GSW15] [TWT16] [MCS19]. However, there are inevitable limitations to these methods: (1) due to the cost of deep sequencing, these small panels track only a limited number of known mutations, e.g., common mutations in cancer, or in the case of personalized panels, mutations identified from the pretreatment/surgery tumor sample of the same patient; (2) personalized panels usually require a labor-intensive experimental design; (3) these panels cannot detect newly emerging tumors, e.g., secondary disease, because they cover a narrow genomic region; and (4) these panels usually require a cohort of noncancer individuals to set cutoffs, which could result in implicit systemic bias from both interindividual variations and interexperimental differences. Recently, two studies [WHG20] [ZSS20] presented cancer monitoring methods using whole-genome sequencing, but they have not yet addressed all the issues discussed above (issues 3 and 4) and focused only on mutations from pretreatment/surgery tumor samples.

In this study, we developed a new cancer monitoring approach, *OncoMonitor*, based on cfDNA standard whole-exome sequencing (WES). It addresses all the aforementioned limitations of existing methods. Specifically, it can be used to comprehensively monitor cancer by analyzing the mutations in both pretreatment/surgery samples and newly emerging tumor clones. By combining statistical methods and machine learning models, our method provides sensitive and unbiased detection of cancer recurrence and MRD by (1) integrating all clonal somatic mutations in the whole exome and (2) sample-specific modeling background noise distribution in the cfDNA sequencing data. Furthermore, our method permits the detection of secondary disease via the de novo detection of newly emerging tumor mutations and the detection of progression via a comprehensive analysis of tumor mutations in post-treatment/surgery plasma samples. Previous methods limit their focus to a few mutations detected from pretreatment/surgery tumor samples, so they can only draw conclusions with respect to the pretreatment/surgery tumor profiles, not tumor evolution. However, approximately 30% of patients with no detectable recurrence or MRD develop secondary disease [AFG20]. With broad sequencing coverage of the genome and a comprehensive analysis of mutations, our method can identify these previously undiagnosed patients in a timely manner when the tumor fraction in the plasma is still low, provide a thorough view of their tumor status, and enable early intervention and personalization of treatment. In this study, we demonstrate that our method achieves sensitive and specific detection of recurrence and secondary disease in plasma samples with low tumor fractions. Specifically, in a cohort of non-small-cell lung cancer patients, we show that our method can provide comprehensive tumor changes for response prediction, which cannot be achieved by previous methods based only on mutations in pretreatment/surgery samples.

In this paper, we have developed a new cancer monitoring approach, *OncoMonitor*, based on cfDNA standard whole-exome sequencing (WES). It addresses all the aforementioned limitations of existing methods. Specifically, it can comprehensively monitor cancer by analyzing both the mutations in the pre-treatment/surgery samples and those in the newly emerging tumor clones. Combining statistical methods and machine learning models, our method provides sensitive and unbiased detection of cancer recurrence and MRD by (1) integration of all clonal somatic mutations on the whole exome and (2) sample-specific modeling of background noise distribution in the cfDNA sequencing data. Furthermore, our method permits the detection of secondary primary diseases by de novo detection of newly emerging tumor mutations, and also enables the detection of progression by comprehensive analysis of tumor mutations in the post-treatment/surgery plasma samples. Previous methods limit their focus on a few mutations detected from pre-treatment/surgery tumor samples, so they can only draw conclusions with respect to the pre-treatment/surgery tumor profiles, in spite of tumor evolution. However, around 30% of the patients with no detectable recurrence or MRD have a second primary disease [AFG20]. With broad sequencing coverage on the genome and comprehensive analysis of mutations, our method can timely identify these previously undiagnosed patients when the tumor fraction in the plasma is still low, provide thorough view of their tumor status, and enable early intervention and personalization of treatment. In this study, we demonstrate that our method achieves sensitive and specific detection of recurrence and second primary disease in the plasma samples with low tumor fraction. Specifically, on a cohort of non-small-cell lung cancer patients, we show that our method can provide comprehensive tumor changes for response prediction, which cannot be achieved by previous methods based only on mutations in the pre-treatment/surgery samples.

3.2 Results

3.2.1 Comprehensive and personalized cancer monitoring from plasma cfDNA

We present a new cancer monitoring method (Figure 3.1a and Figure 3.1b), *OncoMonitor*, to tackle the limitations of previous methods by analyzing both pretreatment/surgery tumor mutations and the newly emerging mutations in posttreatment/surgery samples. We developed four major techniques to achieve comprehensive and sensitive detection of tumor-derived cfDNA. Specifically, we collect a plasma sample, a tumor sample (optional), and a matched white blood cell (WBC) sample from a patient before the treatment/surgery to select markers (i.e., mutations) that are specific to the pretreatment/surgery tumor profile. In the posttreatment/surgery plasma samples, selected pretreatment/surgery tumor markers are tracked, and newly emerging somatic mutations are identified to comprehensively monitor the tumor.

1. **Integrate all clonal tumor mutations from the pretreatment/surgery samples.** Tumor mutations change as tumors evolve, so somatic mutations in pretreatment/surgery samples may disappear in posttreatment/surgery samples. Instead of using tumor mutations from a small panel, we fully utilize the broad genome coverage of the WES data of the pretreatment/surgery samples and identify clonal somatic mutations, which appear in all cancer cells and have high variant allele frequencies (VAFs)

in the plasma [MDP15]. Compared to arbitrary tumor mutations, these mutations are more likely to appear in posttreatment/surgery samples and thus are the most informative for cancer monitoring in posttreatment/surgery samples [ABW17]. To overcome the low tumor fraction in WES data, our method aggregates variant supporting reads at all clonal somatic mutations in pretreatment/surgery samples to track the tumor signal (for details, see section 3.4 and Figure 3.7). Specifically, we quantify the tumor content using the integrated variant allele frequency (IVAF), calculated as the sum of variant supporting reads divided by the sum of all reads at the personalized marker sites. The IVAF indicates the fraction of high-confidence tumor DNA in all cfDNA fragments, so it is treated as the estimated tumor fraction in this study.

2. **Suppress sequencing errors at the read level with a random forest model.** While the tumor reads at a large number of mutations are integrated to amplify the tumor signal, sequencing errors also accumulate. Therefore, we suppress sequencing errors and enhance the signal-to-noise ratio by differentiating the reads containing sequencing errors from those containing true variants with a random forest model (for details, see section 3.4). The model incorporates various information from reads, including the sequencing context, alignment status, mapping quality, base quality, and fragment length, which have been shown to be differential between tumor-derived and non-tumor-derived cfDNA [JCC15] [MR15]. With the random forest model, we classify all variant supporting reads at the personalized marker sites (i.e., the clonal somatic mutation positions) as containing true variants or sequencing errors. Only the reads with true variants are counted as variant supporting reads.

3. **Predict recurrence or MRD from sample-specific background noise distribution.** To predict whether a patient has recurrence or MRD, we need to distinguish the tumor signal from background noise (e.g., sequencing errors) in the plasma sample. Previous studies usually compared the postsurgery/treatment sample of a patient with a cohort of samples from healthy individuals. Because the difference among samples

and experiments is difficult to account for, this kind of comparison might introduce implicit bias to the prediction, and the resulting cutoffs would be difficult to generalize to other experimental protocols. To avoid potential bias, we built a background noise distribution by calculating the IVAF from random genomic positions in the same sample (Figure 3.1b; for details, see section 3.4). Therefore, this background noise distribution represents the error rates in this specific sequencing experiment. The presence of recurrence or MRD can be determined by the p-value of the observed IVAF at the true marker sites given the sample-specific background noise distribution, i.e., the fraction of random samplings with a large IVAF (for details, see section 3.4).

4. **Detect tumor evolution from de novo-identified newly emerging tumor mutations.** Previously described methods for cancer monitoring focus mainly on a small mutation panel, so it is difficult to detect tumor evolution, especially secondary disease. Taking advantage of the WES data with broad genome coverage, our method performs de novo mutation identification to track tumor changes and detect secondary diseases. We utilize *cfSNV* [LNZ20], a sensitive and accurate somatic mutation caller we previously developed to detect somatic mutations between posttreatment/surgery plasma samples and matched WBC samples. The mutation caller, cfSNV, particularly accommodates essential cfDNA-specific properties, including a low tumor fraction, short and nonrandomly fragmented DNA, and heterogeneous tumor content. It addresses the low tumor fraction and the tumor heterogeneity in cfDNA by iterative and hierarchical mutation profiling and ensures a low false-positive rate by multilayer error suppression. From the mutation calling results from cfDNA, the presence of secondary diseases and tumor changes is predicted by the de novo-detected mutations and the corresponding tumor fraction (IVAF, for details, see section 3.4).

3.2.2 Training a random forest model to suppress sequencing errors

When the tumor fraction is low, sequencing errors impair the detection of tumor signals in plasma cfDNA. To increase the signal-to-noise ratio in cfDNA sequencing data, we developed a random forest model to accurately distinguish true variants and sequencing errors for individual reads. The classification of true cancer mutations from sequencing artifacts at the read level has been previously utilized to predict mutations and detect cancer and MRD [ZSS20] [KZS18], though their implementation differs. Our error suppression model incorporates various information, including the sequencing context, alignment status, and quality score, in individual reads. Specifically, in this study, all data were generated from paired-end sequencing, so the read pair can provide additional information on the original cfDNA fragment, such as quality scores from the read mate and fragment length (Table 3.1). To fully utilize the information from the paired-end sequencing data, here we treat a read pair as a unit in the error suppression model. To train the model, we use the sequencing data of two plasma cfDNA samples (collected at two different time points), the matched WBC sample, and the tumor biopsy sample from each of 18 patients with advanced cancer (12 with metastatic breast cancer (MBC) and 6 with metastatic prostate cancer (CRPC)). To build the training data, we label the read pairs based on the consistent mutation calling results across the four samples from the same patient (for details, see section 3.4 and Figure 3.8). The random forest model is then evaluated using leave-one-out cross-validation (for details, see section 3.4). On all validation datasets, the random forest model can accurately distinguish sequencing errors from true variants (average AUC = 0.95, 95% confidence interval [CI] = 0.9496-0.9503, Figure 3.2a, Figure 3.9 and 3.10). By incorporating the random forest model into our cancer monitoring method, the increased signal-to-noise ratio can largely improve the detection of recurrence and MRD in samples with a low tumor fraction (Figure 3.2b and Figure 3.2c) based on our *in silico* spike-in simulation. In particular, in the samples with a 0.025% tumor fraction, the AUC increased ¿ 20%, and the sensitivity increased ¿ 50% (cutoff p-value = 0.05 of background noise distribution) after employing this model. Hence, the random forest model can accurately distinguish true mutations from sequencing errors at the read level and thus facilitate the sensitive detection of weak tumor signals in plasma samples by suppressing sequencing errors.

3.2.3 Detection of cancer recurrence and MRD in the simulation data

To evaluate the performance of our cancer monitoring method, we test our method with the *in silico* spike-in simulation data. We use the sequencing data from 12 MBC and 6 CRPC patients [AHF17]. Each patient has sequencing data from two plasma cfDNA samples (collected at two different time points), the matched WBC sample, and the tumor biopsy sample. To demonstrate the sensitivity of the method, we generated an *in silico* dilution series by mixing the plasma cfDNA samples at the second time point and the matched WBC samples from the 18 MBC and CRPC patients at varying concentrations of cfDNA reads (theoretical tumor fraction ranging from 0.001% to 0.768%, with median 0.105%; for details, see section 3.4 and Figure 3.3a). To test the specificity of the method, we generated 0% dilution samples by subsampling the original WBC samples (for details, see section 3.4 and Figure 3.3b). In the 0% dilution samples, all reads are derived from the original WBC sample, so theoretically, the tumor fraction is 0%, i.e., these simulated samples are from patients who achieved complete remission. For each dilution, five independent random samples with three theoretical depths of coverage (50x, 100x, and 200x) are generated. In total, 968 simulated recurrence samples are generated with positive dilutions, and 150 complete remission samples are generated. Tens to hundreds of clonal somatic mutations (ranging from 49 to 674, with a median of 94) are identified using the pretreatment/surgery plasma and WBC samples.

By applying our monitoring pipeline, we observe increased detection performance with increasing sequencing depth (Figure 3.4a and Figure 3.4b). The trend is as expected because the higher the sequencing depth is, the more tumor DNA fragments might be captured. Specifically, we achieve an AUC $> 95\%$ when tumor fraction is $\geq 0.025\%$ at 200x coverage (Figure 3.4a, with $> 95\%$ sensitivity and 95% specificity (cutoff p-value $= 0.05$ of background noise distribution, Figure 3.4b). This indicates a low detection limit of our method using only 200x whole-exome sequencing data

3.2.4 Detection of secondary disease in the simulation data

Regarding the detection of secondary disease, pretreatment/surgery plasma and tumor biopsy samples cannot provide effective tumor markers; thus, we need to perform de novo SNV detection using posttreatment/surgery plasma samples. To simulate this scenario, we generate an *in silico* dilution series by mixing the plasma samples at the second time point and the matched WBC samples from the 12 MBC and 6 CRPC patients [AHF17] at varying concentrations of cfDNA reads (theoretical tumor fraction ranging from 0.331% to 7.680%, with a median of 2.617%; for details, see section 3.4 and Figure 3.3c). At each dilution, the simulation data are generated at the theoretical depth of coverage 200x. Five independent mixtures are generated at every dilution. To evaluate the specificity of the method, the samples from patients who achieved complete remission generated for recurrence detection are reused. In total, 70 simulated recurrence samples are generated with positive dilutions, and 50 complete remission samples at 200x are reused.

For each pair of simulated plasma and simulated WBC samples, we use *cfSNV* to identify somatic mutations. We use the sum of the tumor fraction at the detected mutations and the number of detected mutations as a prediction score for secondary disease. A patient is predicted to have secondary disease if a large tumor fraction (IVAF $\geq 0.1\%$) and a number of novel mutations (≥ 2) are detected. The AUC is calculated based on the prediction score of all complete remission samples and the simulation samples with positive dilutions of the cfDNA at a specific tumor fraction (see section 3.4). We achieve an AUC $> 80\%$ when the IVAF $\geq 0.1\%$ at 200x coverage (Figure 3.5a), with a sensitivity of approximately 75% and a specificity of approximately 100% (Figure 3.5b). The sensitivity for the detection of secondary disease is lower than that of recurrence and MRD because to confirm secondary disease, novel somatic mutations need to be identified. The detection of novel somatic mutations requires more variant supporting reads at a single position, so the advantage of integrating personalized markers is limited. In summary, our method can be used to sensitively detect secondary disease in plasma samples with a low tumor fraction.

3.2.5 Monitoring non-small-cell lung cancer patients on anti-PD-1 immunotherapy through cfDNA

Cancer immunotherapy, which activates a patients own immune system to fight cancer, has remarkably improved the clinical outcome of a subset of patients with non-small-cell lung cancer (NSCLC) [RHS15]. Despite encouraging clinical improvements, the majority of patients eventually develop resistance and fail to respond to therapy [HTZ20]. Therefore, it is essential to closely monitor the response of patients and identify early their potential need for alternative treatment. However, since resistance may be associated with tumor evolution [SHW17], monitoring requires consideration of the comprehensive tumor profile in the plasma sample during treatment instead of only the pretreatment tumor profile. Our method, which uses sequencing data covering the whole exome, not only covers mutations in the pretreatment samples but also detects newly emerging tumor mutations during treatment. Hence, it allows us to track tumor evolution as well as major tumor clones pretreatment, which may serve as indicators of a patients response.

We applied our cancer monitoring method to plasma cfDNA samples from a cohort of nine non-small-cell lung cancer patients who received anti-PD-1 immunotherapy. Among the nine patients, five are durable responders whose progression-free survival (PFS) is longer than 12 months. The other four patients are early progressors whose PFS is shorter than 6 months. Plasma cfDNA samples were collected at 0 weeks (baseline), 6 weeks and 12 weeks from each patient. The tumor biopsy sample and the matched WBC sample were collected at baseline.

The tumor fraction (IVAF) is calculated using the pretreatment tumor mutations and the newly emerging tumor mutations. The two tumor fractions are different due to possible changes in tumor-derived somatic mutations in plasma cfDNA during treatment. In general, we observe a decreasing or low tumor fraction in the majority of the durable responders and an elevated or high tumor fraction in the early progressors (Figure 3.6). However, in the early progressor 3736900, the tumor fraction calculated using the pretreatment tumor mutations remains at a low level during immunotherapy treatment, but the tumor fraction calculated from the newly emerging tumor mutations shows the opposite trend. This implies a potential clonality change during treatment. The responding clone might have shrunk, but the other clones grew. As our method does not consider newly emerging mutations, the actual trend of the tumors during immunotherapy would be concealed by the changes in the major clone at baseline, which could mislead further treatment. Therefore, by using our cancer monitoring method, we can closely track the change in tumor fraction and mutation clonality in the plasma sample and therefore enable timely treatment guidance.

3.3 Discussion

Cancer monitoring is essential to assess the effectiveness of treatment and thus improve the quality of life of cancer patients. Unlike traditional tumor biopsy, plasma cfDNA provides a unique opportunity for the noninvasive continuous monitoring of cancer patients, but the often very low tumor content in cfDNA is still a major challenge. The current cfDNA-based methods usually rely on deep sequencing a small gene panel to overcome the low tumor content and the low input amount of cfDNA, which limit their power to detect evolving tumors. Therefore, we aimed to develop a new cfDNA-based cancer monitoring method that can effectively and sensitively track changes in tumors, detect cancer recurrence/MRD, and identify the presence of secondary disease despite tumor evolution. We present a new computation method, *OncoMonitor*, for cancer monitoring using cfDNA WES data to tackle the limitations of previous methods. Taking advantage of the wide genome coverage of WES data, our method (1) integrates a large number of clonal tumor mutations identified from pretreatment surgery samples to overcome the challenge of the low tumor fraction in cfDNA, (2) suppresses sequencing errors at the read level with an accurate random forest model to further enhance the tumor signal, (3) builds sample-specific background noise distributions to predict recurrence and MRD to avoid interference from interindividual variations and interexperimental effects, and (4) detects tumor changes, especially secondary disease and progression, from de novo-identified newly emerging tumor mutations.

Combining these techniques, our method achieves sensitive and specific detection of recurrence, MRD and secondary disease. Our method can be used to detect recurrence in a sample with a 0.025% tumor fraction with $> 95\%$ sensitivity and 95% specificity and secondary disease in a sample with a 0.1% tumor fraction with approximate 75% sensitivity and 100% specificity. Since the performance of the method increases with a larger sequencing depth, its performance could be further improved. As an application, we show that in the monitoring of immunotherapy treatment in NSCLC patients, our method achieves accurate and comprehensive monitoring of the changes in a tumor during treatment, which cannot be performed by previous methods focusing only on mutations from pretreatment/surgery tumor samples.

This study has notable limitations. First, our method was validated and evaluated on *in silico* spike-in simulation data and a limited number of NSCLC patients only. To address this limitation, we generated simulation data by considering factors in real cases, including tumor evolution and sampling randomness. For example, simulated plasma samples with positive tumor fractions are generated with subsampling the original plasma sample from the second time point, which already contains a different tumor profile compared to the sample at baseline. Nevertheless, we acknowledge that real cases of MRD, recurrence and secondary disease could be more complicated. Applying our method to larger datasets would enable a more comprehensive evaluation and possible optimization of parameter selection. Second, similar to the tumor fraction calculation in the study by Wan JC et al. [WHG20], tumor evolution during or after treatment/surgery and the random sampling effects in plasma collection could result in a low tumor fraction (IVAF) because some pretreatment/surgery clonal mutations are not detected in posttreatment/surgery samples. Longitudinal monitoring of the same patient would not change the relative trend of tumor changes, but care must be taken when comparing the tumor fraction (IVAF) from our method with that from other methods.

Our results suggest that *OncoMonitor* may provide actionable information and treatment guidance for patients. Our method mainly utilizes point mutations as cancer markers. Next, more cancer-specific features in cfDNA can be incorporated. Recent studies have discovered that copy number variations, fragment length, and jagged ends of cfDNA are all associated with tumor-derived cfDNA. In our random forest model, we incorporated the fragment length of the DNA fragments to discriminate the true variants from sequencing errors. By integrating other features, *OncoMonitor* may further empower cancer monitoring.

3.4 Methods

3.4.1 Data preprocessing

Both genomic DNA sequencing data and cfDNA sequencing data were preprocessed using the same procedure. Raw sequencing data (FASTQ files) were aligned to the hg19 reference genome by *bwa mem* [LD09] and sorted by *samtools* [LHW09]. Then, duplicated reads from PCR amplification were identified and removed by *picard tools MarkDuplicates* [Ins16]. After this step, read group information was added to the bam file using *picard tools AddOrReplaceReadGroups*, and reads were realigned around indels using *GATK RealignerTargetCreator* and IndelRealigner [PRD17] [ACH13]. After realignment, base quality scores were recalibrated using *GATK BaseRecalibrator* and PrintReads. All tools in the data preprocessing were used under their default settings. After data preprocessing, the resulting bam files were used as inputs for mutation detection and MRD detection.

3.4.2 Predicting MRD using tumor-derived somatic mutations in pretreatment/surgery samples

We predicted the presence of MRD by tracking the cfDNA fragments containing tumor-derived somatic mutations (i.e., tumor-derived cfDNA fragments). Due to the low tumor fraction in the plasma samples from patients with MRD, clonal mutations (see section 3.4.3) in a wide range of genomic regions were integrated. To avoid the accumulating of sequencing errors in the integration of mutations, we suppressed the sequencing errors (i.e., filtered sequencing reads with nonreference alleles caused by sequencing errors) by employing a machine learning model (see section 3.4.6), which can accurately classify sequencing reads with sequencing errors or true mutations. Then, the level of tumor-derived cfDNA fragments was compared with the background noise distribution generated from the same plasma sample (see section 3.4.4 and 3.4.5) by a permutation test. If the tumor-derived cfDNA fragments are significantly more abundant than the background noise in the sample (p-value ≤ 0.05), the patient is predicted as having MRD. If no MRD is detected, the follow-up sample from the patient is examined for the presence of secondary disease (see section 3.4.9).

3.4.3 Identification clonal mutations in the pretreatment/surgery plasma sample

The presence of tumor-derived somatic mutations in plasma is usually treated as a reliable tumor marker to confirm the presence of cancer. However, not all tumor-derived somatic mutations are equally effective because subclonal mutations have a lower observed allele frequency than clonal mutations [ABW17]. To overcome the challenge of a low tumor content in the plasma samples of patients with MRD, we integrated tumor-derived somatic mutations in a wide range of the genome (e.g., whole exome). The integration accumulated not only tumor-derived signals but also sequencing errors. Therefore, it is essential to select effective tumor-derived somatic mutations for predicting MRD. If a mutation is selected and less likely to be observed in the plasma, the mutation is more likely to contribute only noise to the prediction of MRD. Therefore, to monitor tumor development in patients, we used all clonal somatic mutations in the pretreatment/surgery plasma and matched WBC samples as tumor markers. Tumor-derived somatic mutations were detected using *cfSNV* [LNZ20]. The detected mutations were then filtered if there was at least one variant supporting the read in the matched WBC sample. A mutation was clonal if its VAF was greater than 25% of the average of the highest five VAFs in the sample [SLA18].

3.4.4 Identification of mutations and CHIP positions

To accurately estimate the background noise in a sequencing experiment, it is essential to remove the interference of the nonreference alleles at the germline mutations, somatic mutations, and CHIP positions. Otherwise, these nonreference alleles can cause largely overestimated levels of nonreference alleles from sequencing errors (i.e., background noise). Therefore, we identified germline mutations in the pretreatment/surgery plasma sample, the tumor biopsy sample, and the matched WBC sample from the same patient using *GATK HaplotypeCaller* and *Strelka2 Germline* [KSH18] using default settings. *GATK HaplotypeCaller* was applied to the plasma sample, the tumor biopsy sample, and the WBC sample individually; *Strelka2 Germline* was applied to the plasma-WBC sample pair and the tumor biopsy-WBC sample pair separately. Somatic mutations were detected from the tumor biopsy sample and the matched WBC sample as a tumor-normal pair using *MuTect* [CLC13] and *Strelka2 Somatic* under default settings. The CHIP positions were identified from pileup files generated using *samtools mpileup*. If a nonmutation position has ≥ 3 variant supporting reads or a VAF $> 1\%$ in the matched WBC sample, the position is regarded as a CHIP position. All the identified germline mutations, somatic mutations and CHIP positions were excluded in the step of building background noise distribution.

3.4.5 Building background noise distribution using random genomic locations

The presence of variant supporting reads at tumor-derived somatic mutations alone cannot be used to determine the presence of MRD because the variant supporting reads could be caused by sequencing errors. Therefore, to quantify the sequencing error level, we built a background noise distribution directly from the exact same plasma sample we used to monitor MRD. Unlike using a panel of normal samples from other sources, we can avoid potential biases from interindividual and interexperimental differences by quantifying the background noise using the same sample. A background noise distribution is generated for a specific size of tumor markers used for monitoring MRD. For a given set of tumor markers of size n, n positions are randomly selected from the targeted genomic region (e.g., whole exome), and the mutations and CHIP positions are excluded. Thus, ideally, all read pairs with nonreference alleles at these n positions are expected to be from sequencing errors, so the observed frequency of these reads represents the background noise level. The sequencing read pairs containing nonreference alleles at these n positions are extracted and input into the sequencing noise suppression model. Then, the observed frequency of the nonreference allele (i.e., integrated variant allele frequency) is calculated as the fraction of the sequencing read pairs, which are classified by the model as containing true mutations, among all the read pairs aligned to the n positions. We repeated the random sampling of n positions and calculated the observed frequency of nonreference alleles K times. A background noise distribution was built from the K observed frequency of nonreference alleles at random n positions. By comparing the integrated variant allele frequency at the tumor markers θ (selected clonal mutations) with the background noise distribution, a p-value can be calculated as the rank of θ among the K background integrated variant allele frequencies. If the p-value is ≥ 0.05, the patient is regarded as having MRD. Based on our simulation, there is a minor difference when $K = 100$, 500, or 1000. In our simulation, we set K to 100.

3.4.6 Machine learning model for suppressing sequencing errors

Although weak tumor signals in plasma samples can be amplified by integrating the variant supporting reads across a large genomic region, sequencing errors can also accumulate and possibly confound the tumor signal. Moreover, because of the low fraction of tumor DNA, the variant supporting reads at a single mutation are not sufficient to provide a robust and accurate estimation of site-level statistics (e.g., strand bias and average base quality) for error removal. Therefore, we developed a machine learning filter to eliminate reads with sequencing errors. Specifically, for a group of genomic positions (tumor markers or random positions), we classify the variant supporting reads with a random forest model to distinguish sequencing errors from true variants. This kind of sequencing error classification has been previously utilized in mutation prediction, cancer detection and MRD detection [ZSS20] [KZS18], though their implementation differs. Our model combines a variety of features (Table 3.1) and automatically discovers statistical relationships among the features that reflect sequencing errors. Since all data in this study were generated from paired-end sequencing, in the following section, we detail the model for paired-end reads, but the principle can also be applied to single-end reads.

To train the random forest model, we used whole-exome sequencing data from 18 patients: 12 with metastatic breast cancer (MBC) and 6 with metastatic prostate cancer (CRPC) [AHF17]. Each patient had four samples sequenced: two plasma cfDNA samples (collected at two different time points), a WBC sample, and a tumor biopsy sample. The training data represent the supporting cfDNA read pairs at known mutation (error) sites and are predicted to contain mutations (errors). Mutation sites are defined as the collection of common germline mutations detected using *Strelka2 Germline* from all four datasets. In addition, common somatic mutations between two cfDNA-WBC pairs (cfDNA data vs. WBC data) and one tumor-WBC pair (tumor data vs. WBC data) were detected using *Strelka2 Somatic* and *MuTect*. Error sites are defined as sufficiently covered sites ($>$ 80x) with only one high-quality nonreference read (base quality $\geq$ 20 and mapping quality $\geq$ 40) in all four datasets. All high-quality labeled read pairs (base quality ≥ 30 and mapping quality ≥ 40) were extracted from raw cfDNA data using *picard tools FilterSamReads*. Multiple read pairs may be extracted from a single locus, but these read pairs are similar and might cause redundancy in the training and testing data. Therefore, we solved the redundancy problem by retaining only one read pair per locus (Table 3.2). Different features were extracted from the overlapping read pairs and the nonoverlapping read pairs (Table 3.1). All categorical features were expanded using the one-hot encoding method. The parameters of the random forest model used were as follows: (1) the number of decision trees was 100, (2) the maximum tree depth was 50, (3) imbalanced classes were addressed by setting the class weights to "balanced", and (4) other parameters were left at their default values. Two random forest classifiers (for overlapping read pairs and nonoverlapping read pairs) were trained on the extracted read pairs.

To validate the performance of the random forest model, leave-one-out cross-validation was performed. For each patient, the labeled read pairs from the 17 other patients were used to train the model, while the labeled read pairs from this patient were used to test the model (results shown in Figure 3.9). The simulation of MRD detection also used the leave-one-out model to avoid data leakage. As an independent testing set, a group of non-small-cell lung cancer patients (12 patients each with 3 samples) with sequential plasma cfDNA samples was used. The read pairs in these cfDNA samples were labeled in the same manner as described above. Then, these labeled read pairs were used as independent testing data for the random forest model trained by the data generated from the 12 MBC and 6 CRPC patients (results shown in Figure 3.10).

3.4.7 Simulation of recurrence and MRD detection by tracking clonal somatic mutations in pretreatment/surgery plasma samples

To demonstrate the sensitivity of the MRD detection pipeline by IVAR, we generated an *in silico* dilution series by mixing the plasma cfDNA samples at the second time point and the matched WBC samples for each of the 18 MBC and CRPC patients at varying concentrations of cfDNA reads (from 0.01% to 1%: 0.01%, 0.05%, 0.1%, 0.3%, 0.5%, 0.8%, and 1%) using *samtools view* and *samtools merge*. The theoretical tumor fraction of these simulation samples was calculated as the product of the original tumor fraction in the cfDNA sample and the dilution. The theoretical tumor fraction ranges from 0.001% to 0.768%, with a median of 0.105%. Note that the theoretical tumor fraction is usually an overestimation of the true tumor fraction because of random sampling and the imperfect on-target rate. Five independent mixtures were generated at every concentration and at a theoretical coverage of 200x, 100x or 50x on the WES targeted regions. Since read sampling is random, it is possible that there is no variant supporting read in a positive-dilution sample. Thus, we removed those positive-dilution samples with no variant supporting read at the personalized markers. In this simulation, the original matched WBC samples, the original cfDNA samples at the first time point, and the original tumor biopsy samples were used as the WBC samples, the pretreatment/surgery cfDNA samples, and the tumor biopsy samples, respectively (Figure 3.3). The *in silico* dilution series was used as the follow-up plasma samples.

To evaluate the specificity of the MRD detection pipeline, we generated subsamples from the original WBC samples. Therefore, these subsamples were expected to have no tumor DNA. For each WBC sample from the 12 MBC and 6 CRPC patients, five subsamples were generated, with reads theoretically equivalent to 200x, 100x, and 50x coverage of the targeted regions. These subsamples were used as the follow-up plasma samples. The original cfDNA samples at the first time point and the original tumor biopsy samples were used as the pretreatment/surgery cfDNA samples and the tumor biopsy samples, respectively. To avoid potential data leakage in this simulation, we used another subsample of the original WBC samples at a sampling rate of 95% (Figure 3.3). Therefore, in this simulation, we preserved some randomness between the WC samples and the follow-up plasma samples, which reflects real cases.

In total, 968 simulated recurrence samples were generated with positive dilutions, and

150 complete remission samples were generated. The performance metrics (AUC, sensitivity, and specificity) were evaluated on positive-dilution samples grouped by the tumor fraction at a 0.005% step size and all zero-dilution samples (the samples with WBC reads only).

3.4.8 Calculation of the integrated variant allele frequency

To quantify tumor DNA across multiple loci, we calculated the "integrated variant allele frequency" as the fraction of the sequencing read pairs, which were classified by the model as containing true mutations by the model, in all the read pairs aligned to the loci. The IVAF indicates the fraction of high-confidence tumor DNA in all cfDNA fragments, so it is treated as the estimated tumor fraction in this study.

3.4.9 Detection of secondary disease

For a patient with secondary disease, the tumor markers identified from pretreatment/surgery samples are not effective, so detecting novel mutations is important for secondary disease detection. Since data from a wide range of genomic regions are used in the MRD detection pipeline, these data provide opportunities for novel somatic mutation detection. We employ *cfSNV*, which is a sensitive and accurate somatic mutation caller we previously developed, to detect somatic mutations between the posttreatment/surgery plasma sample and the matched WBC sample. The mutation caller, *cfSNV*, particularly accommodates essential cfDNA-specific properties, including a low tumor fraction, short and nonrandomly fragmented DNA, and heterogeneous tumor content. It addresses the low tumor fraction and the tumor heterogeneity in cfDNA by iterative and hierarchical mutation profiling and ensures a low false-positive rate by multilayer error suppression. From the mutation calling results from cfDNA, secondary disease is detected based on a prediction score, which is calculated as the sum of the tumor fraction at the detected somatic mutations and the number of the detected somatic mutations. The performance metrics (AUC, sensitivity, and specificity) are evaluated on the prediction score of the positive-dilution samples grouped by tumor fraction at a 0.1% step size and all zero-dilution samples. The sensitivity and specificity were evaluated at a prediction score of 2.001, i.e., $\geq$ 0.1% tumor fraction and ≥ 2 detected mutations. In other words, a patient is predicted to have secondary disease if a large tumor fraction (tumor fraction calculated at the novel mutations $\geq$ 0.1%) and ≥ 2 novel mutations are detected.

3.4.10 Simulation of secondary disease detection

To evaluate the sensitivity of the method for secondary disease detection, we generated an *in silico* dilution series by mixing the plasma cfDNA samples at the second time point and the matched WBC samples from the 18 MBC and CRPC patients at varying concentrations of cfDNA reads (from 1% to 10%: 1%, 3%, 5%, 8%, and 10%) using *samtools view* and *samtools merge*. The theoretical tumor fraction of these simulation samples was calculated as the product of the original tumor fraction in the cfDNA sample and the dilution. The theoretical tumor fraction ranged from 0.331% to 7.680%, with a median of 2.617%. Note that the theoretical tumor fraction is usually an overestimation of the true tumor fraction because of random sampling and the imperfect on-target rate. Each spike-in sample contained a total number of randomly sampled reads theoretically equivalent to 200x coverage of the targeted regions. Five independent mixtures were generated at every concentration. In this simulation, the original matched WBC samples were used as the WBC samples. The *in silico* dilution series was used as the follow-up plasma samples. To demonstrate the specificity of the method, we utilized the zero-dilution simulation data generated for MRD detection. Because the pretreatment/surgery plasma samples and tumor biopsy samples cannot provide effective tumor markers for secondary disease, no original plasma samples from the first time point or original tumor biopsy samples were used in this simulation (Figure 3.3). The performance metrics (AUC, sensitivity, and specificity) are evaluated on the positive-dilution samples grouped by the tumor fraction at a 0.1% step size and all zero-dilution samples. In total, 70 simulated recurrence samples were generated with positive dilutions, and 50 complete remission samples at 200x were reused.

Feature description	Feature type	Nonoverlapping 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 a 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
CIGAR information in a 7-bp window centered on the query site from the mate	categorical	no	yes

Table 3.1: Extracted features from read pairs for the random forest models. The column "nonoverlapping model" indicates which features are used in the random forest model to filter nonoverlapping read pairs. The column "overlapping model" indicates which features are used in the model for overlapping read pairs.

Patient ID	Sample ID	Error No Overlap	Error Overlap	Variant No Overlap	Variant Overlap
CRPC_17	SRR6708976	15680	151	28570	24730
CRPC_17	SRR6708977	10789	1358	28185	25619
CRPC_22	SRR6708978	16282	144	28622	24474
CRPC_22	SRR6708979	6927	97	27933	23840
CRPC_264	SRR6708961	509	225	27052	24457
CRPC_264	SRR6708962	319	99	27122	23835
CRPC_372	SRR6708965	1646	584	26069	23315
CRPC_372	SRR6708966	396	92	27065	24682
CRPC_468	SRR6708970	504	198	32441	29473
CRPC_468	SRR6708971	510	239	31484	27138
CRPC_554	SRR6708974	45617	845	33894	30638
CRPC_554	SRR6708975	9434	435	33687	30218
MBC_191	SRR6708921	39	3	33692	30424
MBC_191	SRR6708922	12953	5129	32565	28987
MBC_284	SRR6708924	594	124	32020	26885
MBC_284	SRR6708925	1726	612	32400	29515
MBC_288	SRR6708927	9658	3366	32654	29815
MBC_288	SRR6708928	38	4	32852	28932
MBC_295	SRR6708931	8895	370	33720	28268

MBC_295	SRR6708932	34007	10810	32476	27214
MBC_303	SRR6708935	3353	117	34204	31599
MBC_303	SRR6708936	3144	1126	31067	26469
MBC_307	SRR6708937	1383	40	33354	29336
MBC_307	SRR6708938	21055	5974	32663	29378
MBC_313	SRR6708939	2292	70	33820	30856
MBC_313	SRR6708940	21503	8790	31725	28822
MBC_318	SRR6708943	1330	51	33391	30548
MBC_318	SRR6708944	16734	5850	32684	29834
MBC_325	SRR6708947	5099	2431	29623	27158
MBC_325	SRR6708948	6120	1521	32499	29882
MBC_331	SRR6708950	14813	4030	32628	29610
MBC_331	SRR6708951	17448	5969	32055	28088
MBC_339	SRR6708955	28526	1651	34122	30393
MBC_339	SRR6708956	28212	1446	34088	29869
MBC_349	SRR6708957	17462	532	33951	31605
MBC_349	SRR6708958	19876	436	34059	31567
NSCLC_1129838	241	1	0	34188	34003
NSCLC_1129838	714	0	1	32863	33395
NSCLC_1129838	729	4	2	34186	34068
NSCLC_2163573	706	1750	1390	34237	34357
NSCLC_2163573	717	2065	1607	34356	34337
NSCLC_2163573	732	1582	1516	34168	34349
NSCLC_3736900	751	48627	19237	34227	34139
NSCLC_3736900	760	38567	12011	33316	33900
NSCLC_3736900	768	54542	7655	34522	33087

NSCLC_4492669	643	24672	9970	32821	33588
NSCLC_4492669	704	19354	6022	33109	33642
NSCLC_4492669	715	24949	5439	34582	34155
NSCLC_4496246	620	34884	12337	32299	33056
NSCLC_4496246	636	38017	12905	33568	33659
NSCLC_4496246	644	40591	12438	33622	33614
NSCLC_4528560	640	31073	17751	33310	34140
NSCLC_4528560	647	76658	22143	35122	34883
NSCLC_4528560	708	45394	17274	23596	23540
NSCLC_4536877	650	24	11	33873	33887
NSCLC_4536877	712	46	9	33248	33635
NSCLC_4536877	725	22	9	33244	33738
NSCLC_4563728	705	1071	422	31756	33222
NSCLC_4563728	716	7433	2075	34431	34373
NSCLC_4563728	730	13020	2695	34688	34389
NSCLC_4566326	711	36161	12359	30874	30758
NSCLC_4566326	723	22368	9583	30620	30743
NSCLC_4566326	737	23331	10328	30772	30731
NSCLC_4582920	720	10	3	33874	33952
NSCLC_4582920	734	8	2	33891	33908
NSCLC_4582920	743	10	5	34162	33872
NSCLC_4599369	726	9345	4080	34467	34522
NSCLC_4599369	739	28379	19286	34878	34570
NSCLC_4599369	747	17077	14692	33156	33922
NSCLC_4650336	762	8296	7250	34574	34618
NSCLC_4650336	770	7877	8024	34535	34626

NSCLC_4650336	785	10825	11328	34547	34633

Table 3.2: Sample IDs and the number of labeled read pairs for training and testing on the random forest model. For the MBC patients and the CRPC patients, the patient IDs follow the naming convention in [AHF17], while the sample IDs are the SRA accession IDs of the sample.

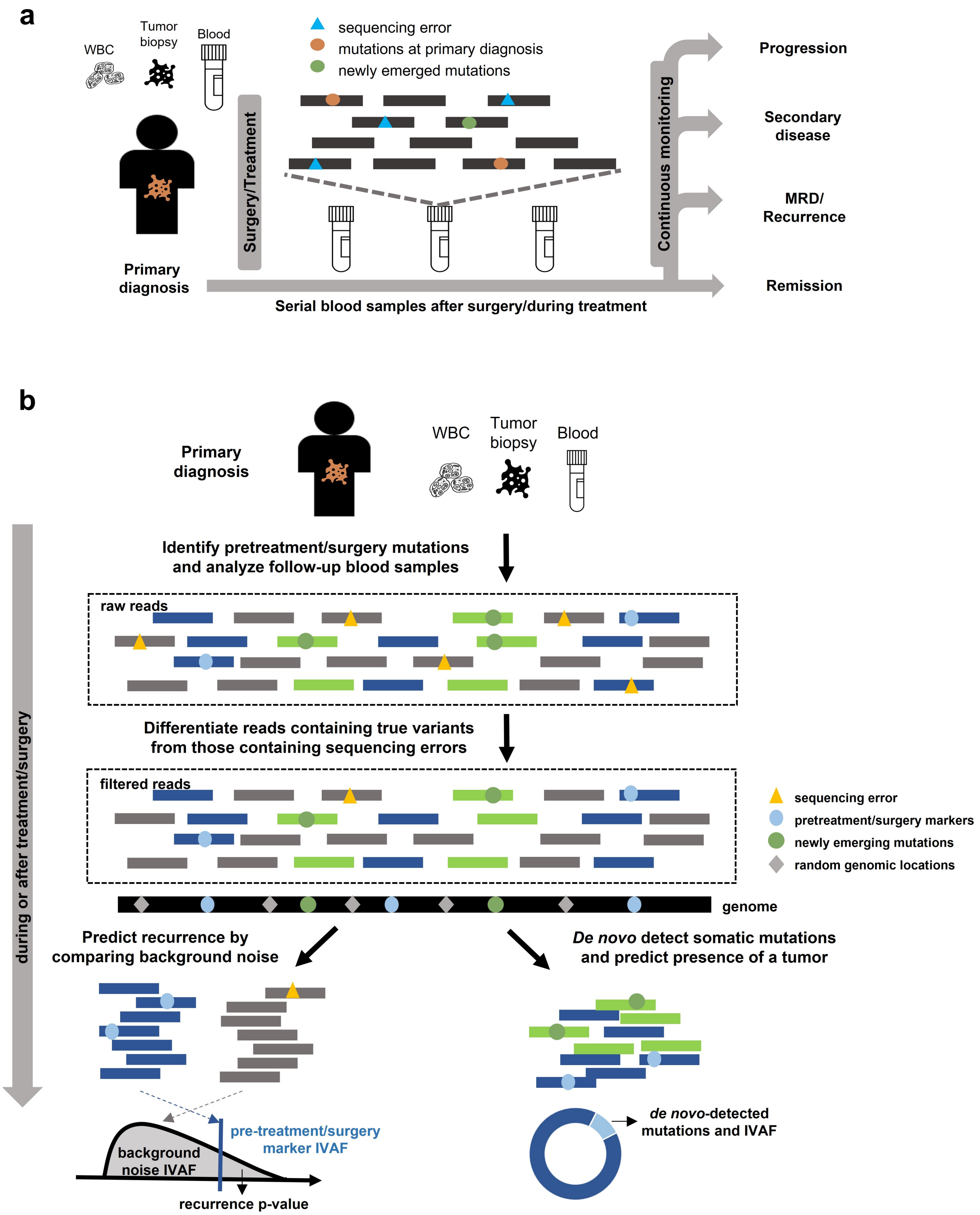

a
WBC
Tumor biopsy
Blood
sequencing error
mutations at primary diagnosis
newly emerged mutations
Surgery/Treatment
Continuous monitoring
Progression
Secondary disease
MRD/ Recurrence
Remission
Primary diagnosis
Serial blood samples after surgery/during treatment
b
Primary diagnosis
WBC
Tumor biopsy
Blood
during or after treatment/surgery
Identify pretreatment/surgery mutations and analyze follow-up blood samples
raw reads
Differentiate reads containing true variants from those containing sequencing errors
filtered reads
sequencing error
pretreatment/surgery markers
newly emerging mutations
random genomic locations
genome
Predict recurrence by comparing background noise
De novo detect somatic mutations and predict presence of a tumor
background noise IVAF
pre-treatment/surgery marker IVAF
recurrence p-value
de novo-detected mutations and IVAF

Figure 3.1: Cancer monitoring in cfDNA samples by tracking pretreatment/surgery tumor mutations and newly emerging tumor mutations. (**a**) Illustration of sample collections for cfDNA-based cancer monitoring. At primary diagnosis, the tumor biopsy, plasma sample (blood), and matched white blood cell (WBC) sample are collected to generate the pretreatment/surgery tumor profile. Serial blood samples are collected to monitor tumor evolution and detect recurrence/MRD during or after treatment/surgery. (**b**) *OncoMonitor* workflow. In the pretreatment/surgery samples, clonal tumor mutations are identified for tumor tracking in the posttreatment/surgery samples. Given a posttreatment/surgery plasma sample, the integrated variant allele frequency (IVAF) is calculated at the selected pretreatment/surgery tumor mutations and compared to a sample-specific background noise distribution generated by randomly sampled positions to predict recurrence/MRD. Furthermore, somatic mutations are de novo detected using cfSNV between the posttreatment/surgery plasma and WBC samples. The presence of secondary diseases and tumor changes is predicted by the de novo-detected mutations and the corresponding IVAF.

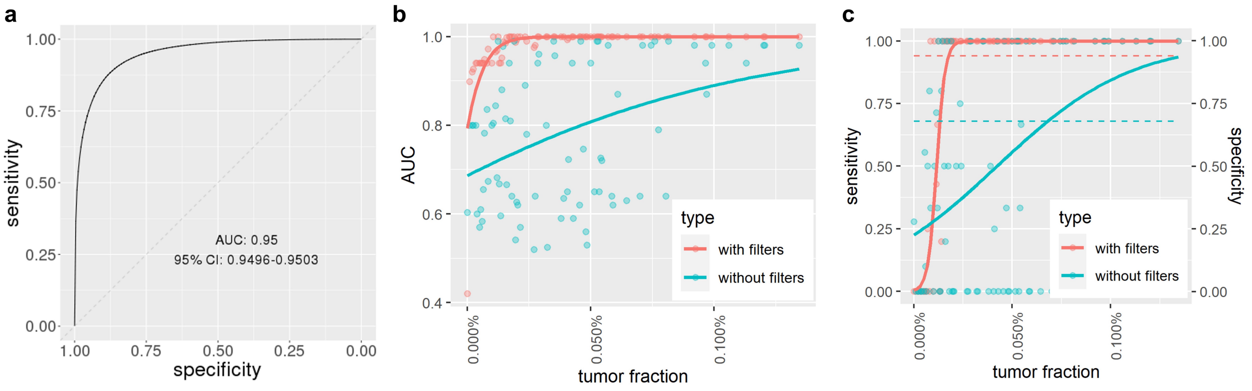


Figure 3.2: Performance of the random forest model and the improved tumor detection power with the model using simulation data. (**a**) Receiver operating characteristic (ROC) curve of the random forest model in all 36 leave-one-out cross-validation sets (two validation sets for each of the 18 patients). The 95% confidence interval (95% CI) is indicated in the figure. (**b**) The area under the ROC curve (AUC) of recurrence/MRD detection with and without the random forest model in the *in silico* spike-in samples with different tumor fractions. The circles and error bars indicate the average AUC. The solid lines show the smoothed performance fitted with logit functions. (**c**) The sensitivity and specificity of recurrence/MRD detection with and without the random forest model in the *in silico* spike-in samples with different tumor fractions. The circles show the sensitivity using a cutoff p-value = 0.05 of background noise distribution. The dashed lines show the specificity using a cutoff p-value = 0.05 of background noise distribution. The solid lines show the smoothed performance fitted with logit functions.

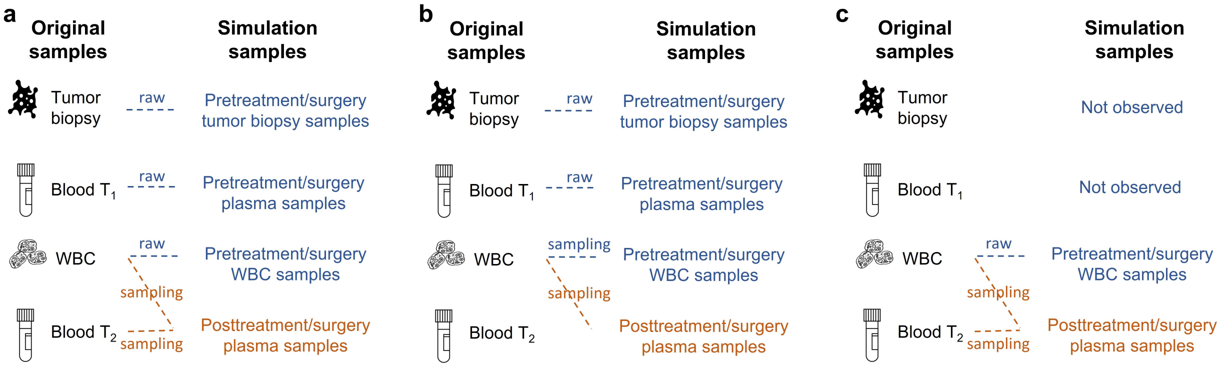

a
Original samples
Simulation samples
Tumor biopsy
raw
Pretreatment/surgery tumor biopsy samples
Blood T_1
raw
Pretreatment/surgery plasma samples
WBC
raw
Pretreatment/surgery WBC samples
sampling
Blood T_2
sampling
Posttreatment/surgery plasma samples
b
Original samples
Simulation samples
Tumor biopsy
raw
Pretreatment/surgery tumor biopsy samples
Blood T_1
raw
Pretreatment/surgery plasma samples
WBC
sampling
Pretreatment/surgery WBC samples
sampling
Blood T_2
Posttreatment/surgery plasma samples
c
Original samples
Simulation samples
Tumor biopsy
Not observed
Blood T_1
Not observed
WBC
raw
Pretreatment/surgery WBC samples
sampling
Blood T_2
sampling
Posttreatment/surgery plasma samples

Figure 3.3: Sample generation settings for the *in silico* spike-in simulation. (**a**) Simulation samples with positive cfDNA dilutions for recurrence/MRD detection, i.e., recurrence/MRD cases. The original tumor biopsy sample, the original plasma sample at the first time point (T1), and the original WBC sample are directly used as the pretreatment/surgery samples in the simulation. The original WBC sample and the original plasma sample at the second time point (T2) are mixed at known dilutions. The mixed samples are used as the posttreatment/surgery plasma sample. (**b**) Simulation samples with zero cfDNA dilutions, i.e., complete remission cases. The original tumor biopsy sample and the original plasma sample at the first time point (T1) are used directly as the pretreatment/surgery samples in the simulation. Two random samplings of the original WBC sample are used as the pretreatment/surgery WBC sample and the posttreatment/surgery plasma sample. (**c**) Simulation samples with positive cfDNA dilutions as used for secondary disease detection. The original WBC sample is directly used as the pretreatment/surgery WBC sample in the simulation. The original WBC sample and the original plasma sample at the second time point (T2) are mixed at known dilutions. The mixed samples are used as the posttreatment/surgery plasma sample. In the simulation, there is no pretreatment/surgery tumor biopsy sample or plasma sample.

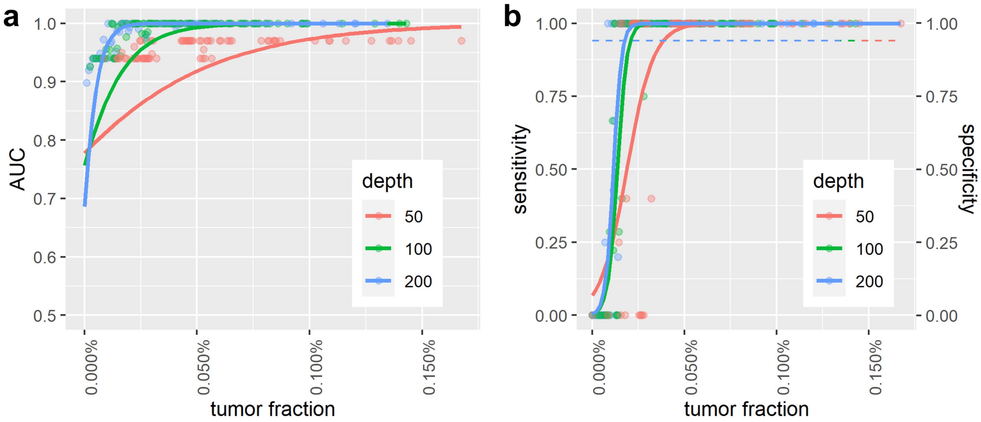


Figure 3.4: Performance of cancer recurrence and MRD detection using the simulation data. (**a**) AUCs of the *in silico* spike-in samples with different tumor fractions and different sequencing depths. The circles represent the average AUC. The solid lines are the smoothed performance fitted with logit functions. (**b**) Sensitivity and specificity of the *in silico* spike-in samples with different tumor fractions and different sequencing depths. The circles show the sensitivity using a cutoff p-value = 0.05 of background noise distribution. The dashed lines show the specificity using a cutoff p-value = 0.05 of background noise distribution. The solid lines show the smoothed performance fitted with logit functions.

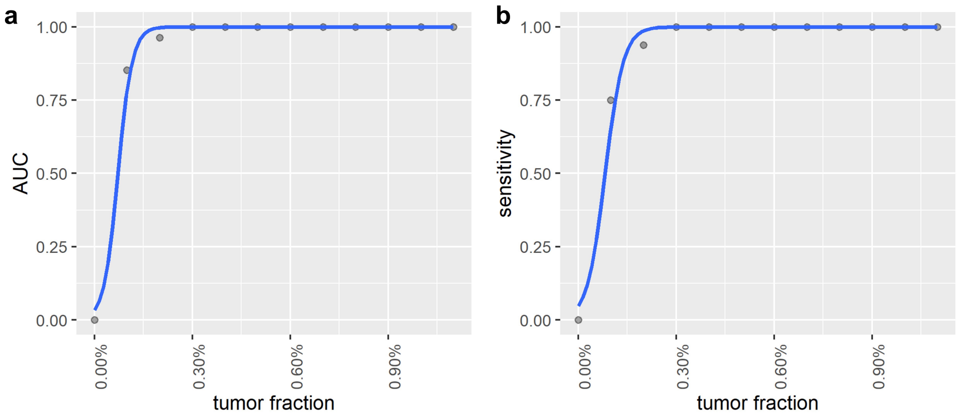


Figure 3.5: Performance of secondary disease detection with the simulation data. (**a**) AUC of the *in silico* spike-in samples with different tumor fractions at 200x sequencing depth. The circles represent the average AUC. The solid lines are the smoothed performance fitted with logit functions. (**b**) The sensitivity at the specificity of approximately 100% in the *in silico* spike-in samples with different tumor fractions at a sequencing depth of 200x. The circles show the sensitivity. The solid lines show the smoothed performance fitted with logit functions.

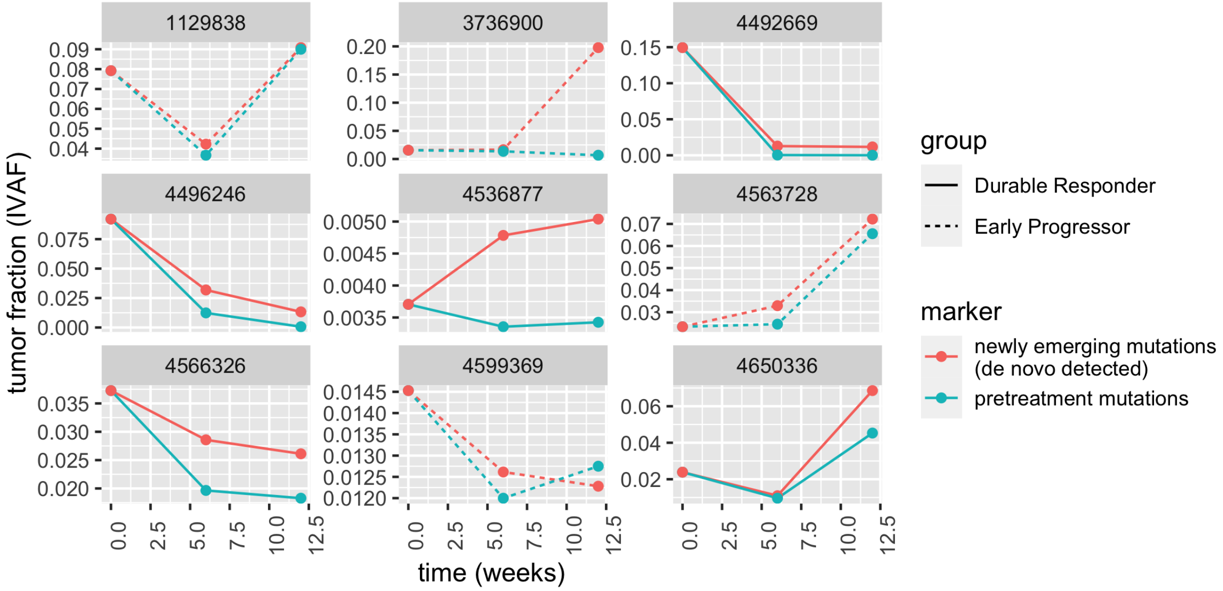


Figure 3.6: Longitudinal cfDNA monitoring in NSCLC patients who received anti-PD-1 immunotherapy. The lines show the tumor fraction in cfDNA during treatment.

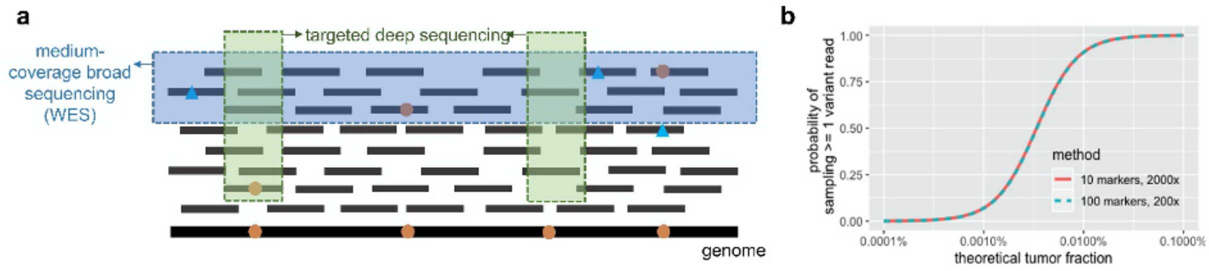


Figure 3.7: The difference in tumor signal detection between targeted deep sequencing data and medium-coverage broad sequencing data. (**a**) Illustration of observed sequencing reads as a sample from the pool of cfDNA. The blue box indicates the observed reads from medium-coverage broad sequencing, while the green box indicates the observed reads from targeted deep sequencing. (**b**) The theoretical detection probability of tracking 10 markers at 2000x and 100 markers at 200x. The probability of sampling ≥ 1 variant read is determined by a binomial distribution over all markers given a fixed tumor fraction, which is the probability of a read from tumor cells.

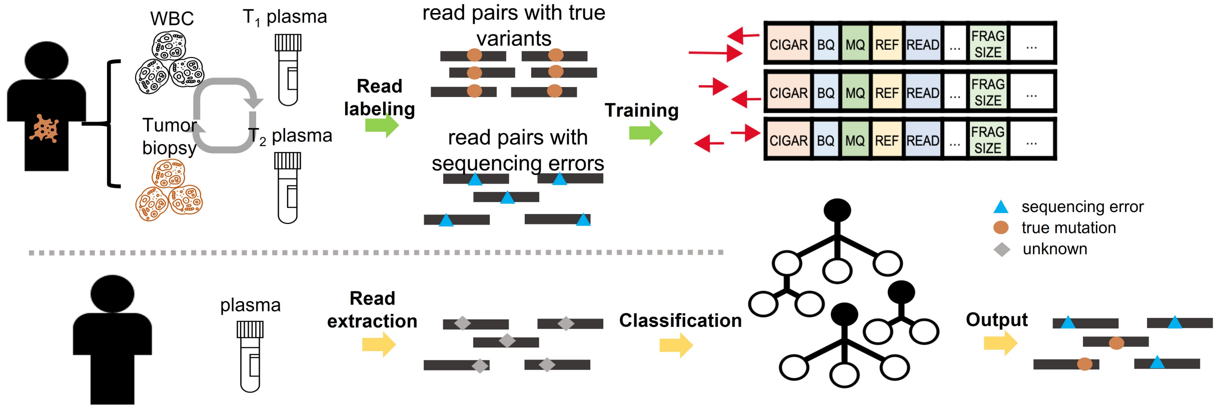


Figure 3.8: Training data extraction and utilization of the random forest model for suppressing sequencing errors at the read level. The upper panel shows the training data extraction workflow (for details, see 3.4). True variant positions (somatic and germline mutations) and sequencing error positions are identified by comparing the WBC sample, the tumor biopsy sample and two plasma samples from the same patient. Read pairs with nonreference bases at these identified positions are extracted and labeled “true variants” and “sequencing errors”, respectively. Then, various features are extracted from each read pair, and these data are used as training and testing data for the random forest model. The lower panel shows the utilization of the random forest model. Given a posttreatment/surgery sample, the features from the read pairs at given loci are extracted from the sequencing data and classified as containing a “sequencing error” or a “true variant”.

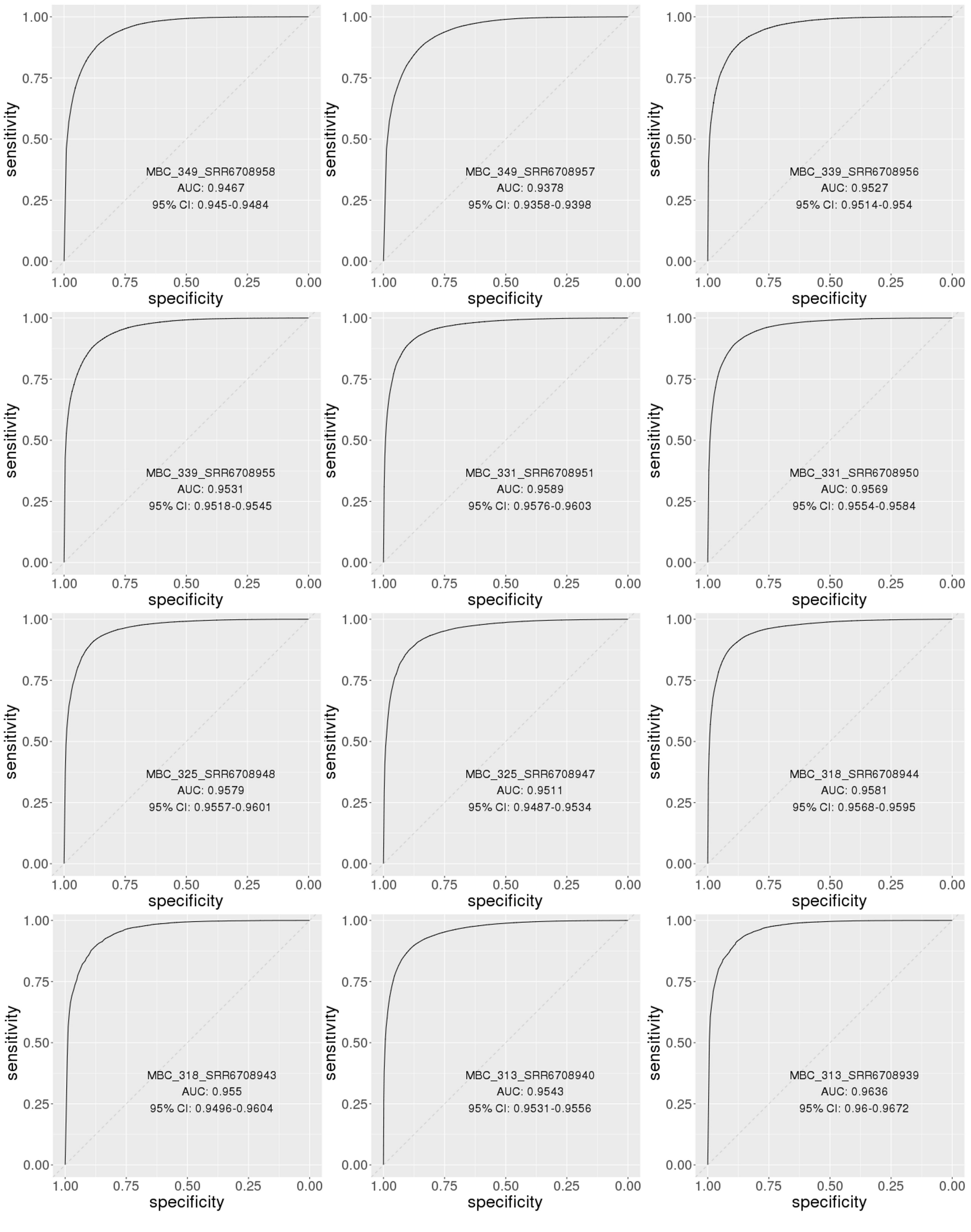


Figure 3.9: Performance of the random forest model in the 36 leave-one-out cross-validation sets. The receiver operating characteristic (ROC) curve of the random forest model in the 36 leave-one-out cross-validation sets (Part 1).

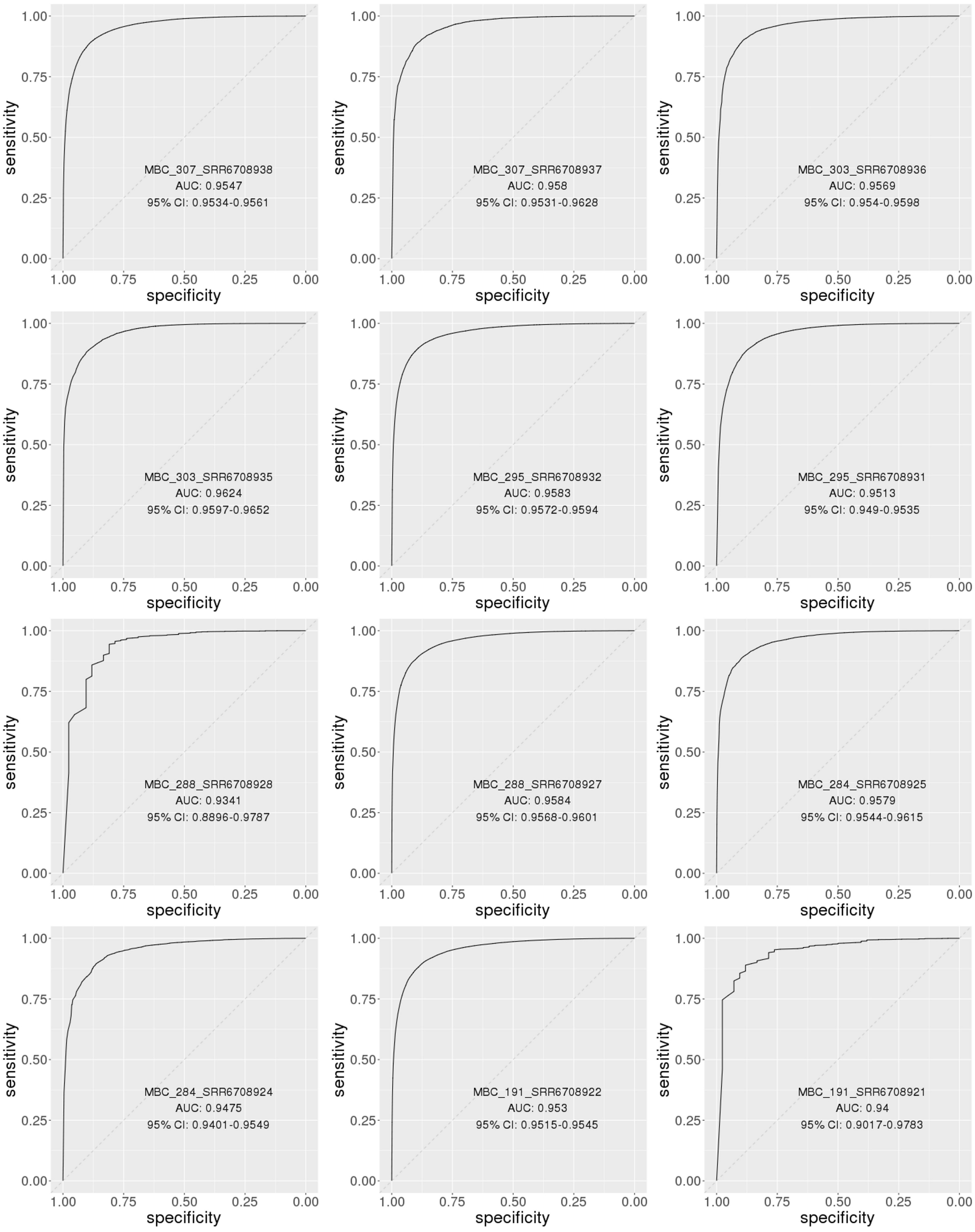


Figure 3.9: (Continued) Performance of the random forest model in the 36 leave-one-out cross-validation sets. The receiver operating characteristic (ROC) curve of the random forest model in the 36 leave-one-out cross-validation sets (Part 2).

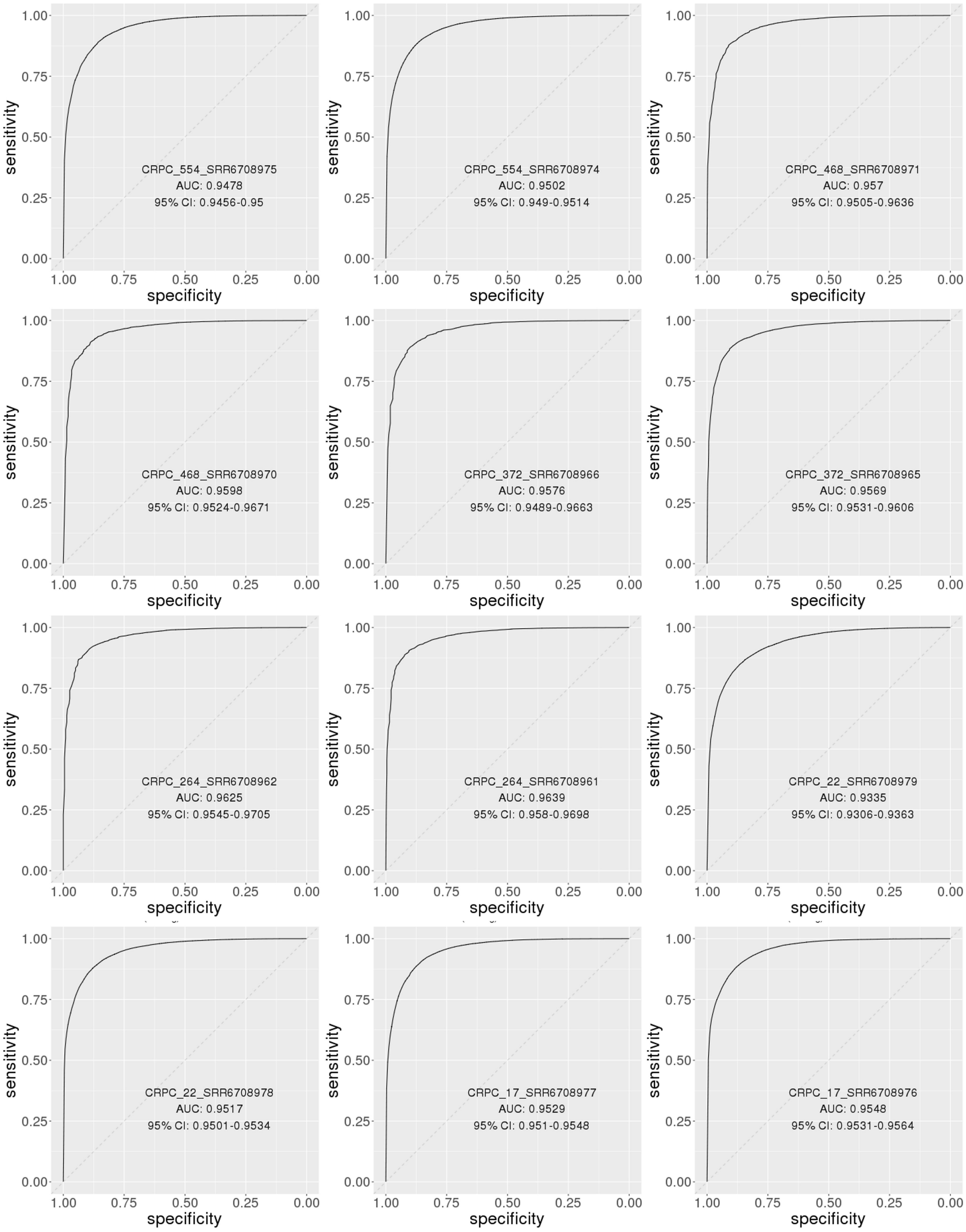


Figure 3.9: (Continued) Performance of the random forest model in the 36 leave-one-out cross-validation sets. The receiver operating characteristic (ROC) curve of the random forest model in the 36 leave-one-out cross-validation sets (Part 3).

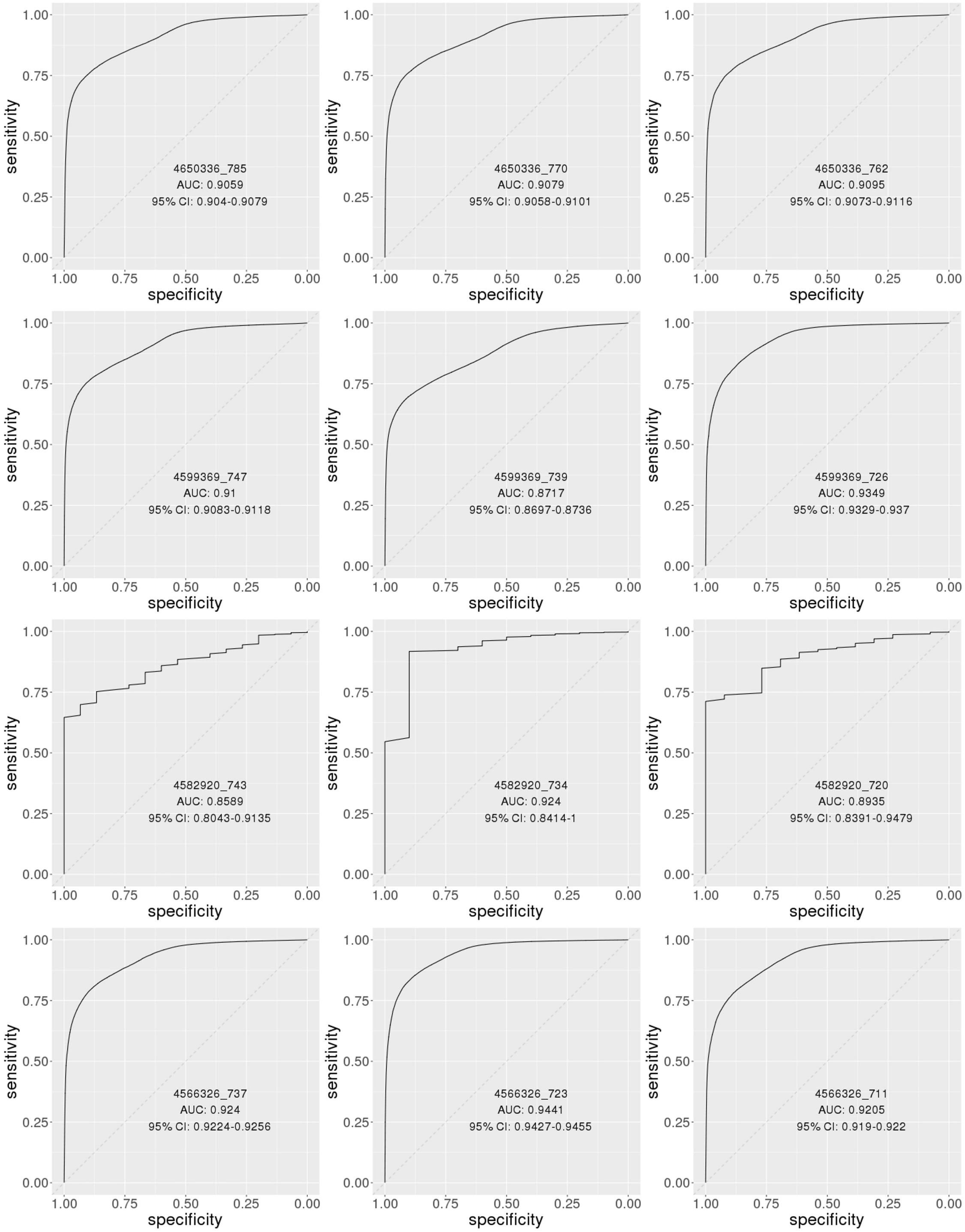


Figure 3.10: Performance of the random forest model in the 36 independent testing datasets from 12 NSCLC patients. The receiver operating characteristic (ROC) curve of the random forest model in the 36 independent datasets. (Part 1).

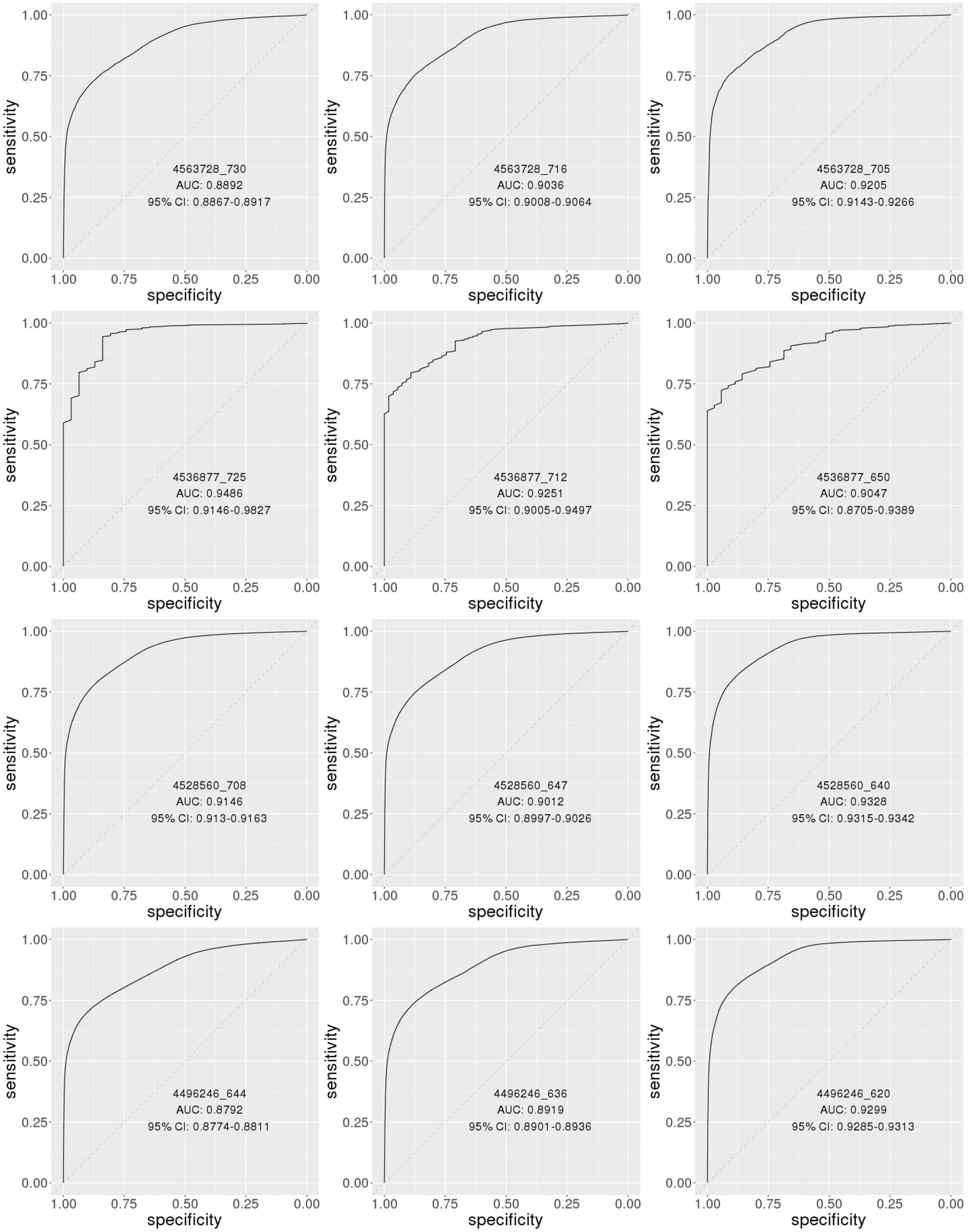


Figure 3.10: (Continued) Performance of the random forest model in the 36 independent testing datasets from 12 NSCLC patients. The receiver operating characteristic (ROC) curve of the random forest model in the 36 independent datasets (Part 2).

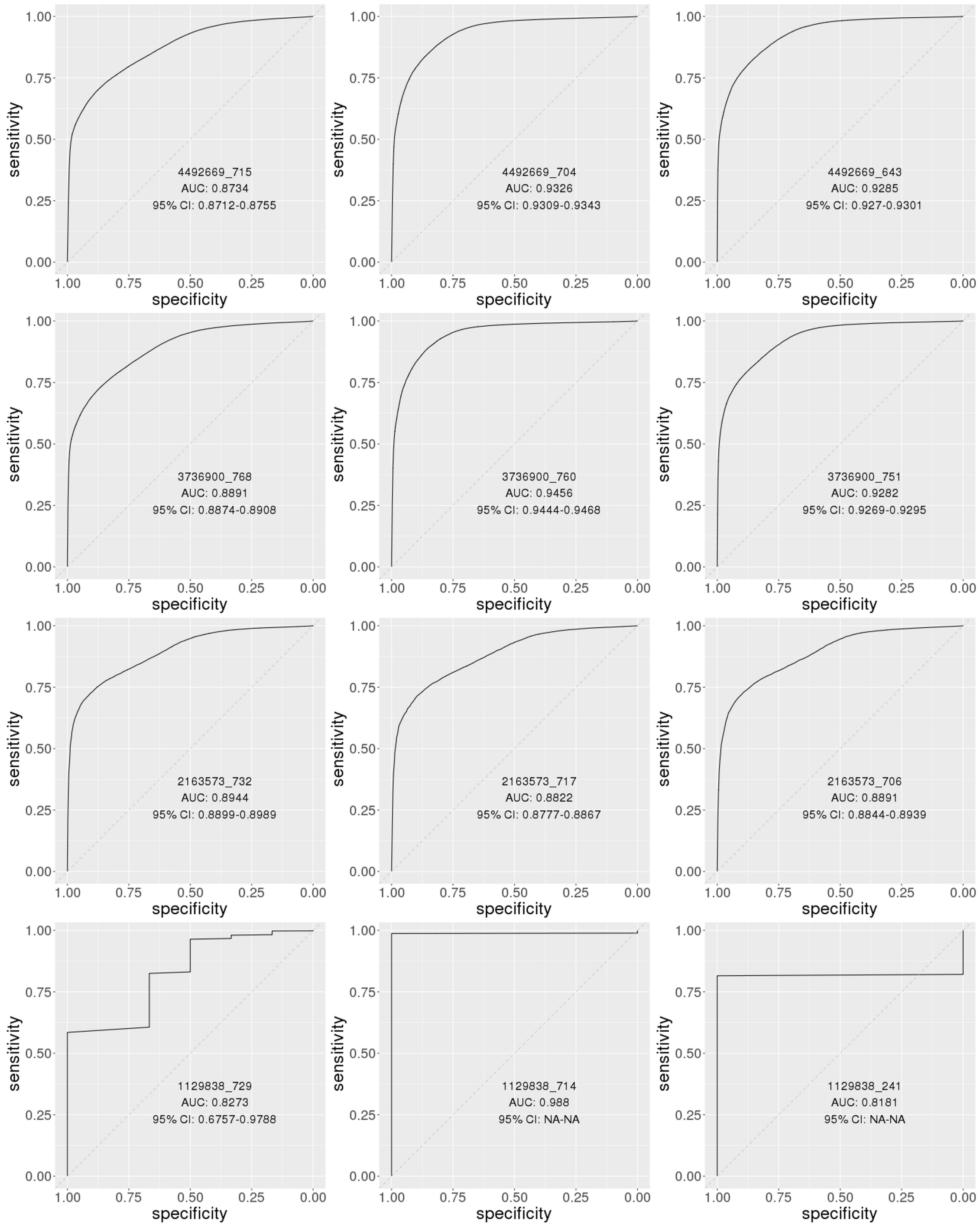


Figure 3.10: (Continued) Performance of the random forest model in the 36 independent testing datasets from 12 NSCLC patients. The receiver operating characteristic (ROC) curve of the random forest model in the 36 independent datasets (Part 3).

CHAPTER 4

Disease detection with the microbiome profile in the plasma cfDNA: sepsis and cancer

4.1 Introduction

The human microbiota consists of tens of trillions of cells living in or on each person [TLH07]. With the development of high-throughput sequencing technology, the human microbiota has been found to substantially contribute to human diseases, such as sepsis [HW17], obesity [LNQ13] [THY09], cirrhosis [QYL14] [CTH19], liver cancer [JJ19] [MB19], and stomach cancer [BMS14] [PB02]. Given its importance, efforts have been devoted to disease diagnosis and classification based on the human microbiota [SJS20] [BTR19] [PKZ20]. Currently, studies usually focus on microbes at body sites from which samples are relatively easy to obtain, such as the skin [APD13] and gut [KDN11], and sometimes require microbial cultivation [KKC03]. The sampling site limitation has restricted the range of diseases that can be studied, and the need for cultivation poses further constraints on microbe identification [WDL14]. With the discovery of cell-free microbial DNA in plasma, liquid biopsy offers a potential solution to comprehensively profile the microbiota in the human body, as microbes in nearly all tissues can release DNA into the blood [HCF18] [CBL19] [BHL20] [HLS20]. Therefore, be combining this method with high-throughput sequencing technology, microbes from various tissues, including unculturable microbes, can be profiled from plasma cell-free DNA (cfDNA).

In addition to the comprehensiveness of the microbial profile in the plasma, cell-free microbial DNA has advantages in the detection and classification of specific diseases, such as sepsis and cancer. Sepsis is a life-threatening emergency condition arising from various infections of human tissues, and it is vital to quickly and accurately detect the causative microbes to ensure prompt treatment. According to recent guidelines, deploying appropriate antibiotic therapy as early as possible (preferably within 1 hour) is crucial for sepsis patients [DLR13]. However, the standard procedure of pathogen detection requires blood culturing for a significant period of time (up to 5 days) [DLR13] [VBL15]. Therefore, a more rapid approach for diagnosing sepsis and comprehensively profiling the microbiome is urgently required. By sequencing plasma cfDNA, we could promptly profile different potential pathogens in the blood and thus save precious time for the patients. In addition to sepsis, which is directly caused by microbial infection, the human microbiota has been indicated to substantially impact some types of cancer [JJ19] [MB19] [BMS14] [PB02] [DBP17]. In particular, recent studies have shown that the microbiome in the tissue and blood is specific to cancer types and can discriminate samples from healthy individuals and samples from patients with multiple types of cancers [PKZ20]. Considering the widely utilized human methylome of cfDNA, the microbiota could possibly provide complementary information to assist the current noninvasive cancer detection and location methods.

In this study, we aimed to develop a workflow using the microbiome composition in cfDNA for disease detection. For this purpose, cfDNA was isolated and sequenced from the blood samples of healthy and diseased cohorts. Then, a random forest model was trained based on the microbiome composition of the healthy individuals and patients. Specifically, as examples, we focused on rapid sepsis diagnosis and noninvasive cancer detection. For rapid sepsis diagnosis, our method achieved an area under the ROC curve (AUC) of 93% based on whole-genome sequencing data. The cooccurrence network of the candidate pathogens showed the characteristics of the abundant pathogens and could be further utilized to guide therapies. For cancer detection, our method achieved an AUC of 93% for patients with colon (COAD), lung (LUAD and LUSC), liver (LIHC), and stomach (STAD) cancers based on the cfMethyl-seq data. In addition to cancer detection, the microbiome composition has the potential to distinguish patients with different cancer types with an accuracy of 0.63 among the five cancer types. By functional analysis of the microbes with the most importance in the random forest model, we showed that most of these microbes were significantly differentially abundant among cancer types and between cancer patients and healthy individuals. The microbes with high importance in the model were consistent with the findings in previous studies.

4.2 Results

4.2.1 Rapid sepsis diagnosis based on the cfDNA microbiome from WGS data

Following the procedures shown in Figure 1a and 1b, we developed a two-step approach for rapid sepsis diagnosis. First, we identified 3546 bacterial species through alignment and classification of cfDNA sequencing reads from 118 healthy and 38 sepsis samples [GSG16] [UHS16]. These samples were randomly partitioned into two groups: 103 samples (78 healthy samples and 25 sepsis samples) for training and 53 samples (40 healthy samples and 13 sepsis samples) for testing. For each species, we fit a beta distribution based on the bacterial abundance vector with 78 elements from the healthy training samples. Then, the 25 abundances from the sepsis training samples were tested one by one against the beta distribution to generate 25 p-values. A species was considered a candidate pathogen if at least one P-value was smaller than 0.01. By this filtering procedure, approximately 200 candidate pathogenic bacteria were selected. Figure 2 shows some examples of these candidate pathogens, with the bacterial abundances showing significantly different distributions between healthy and sepsis samples.

Second, based only on the observed abundances of the candidate pathogenic bacteria, we trained a random forest model with balanced subsampling to generate an accurate classifier. Finally, we used this classifier to test the other one-third of normal and sepsis samples reserved for this purpose. The above pipeline was repeated 1000 times through bootstrapping. As shown in Figure 3a, the average out-of-bag error (OOB error) was 0.16 when there were a sufficiently large number of decision trees (>100). The performance of the diagnosis strategy was satisfactory, with an average AUC of 0.93, sensitivity of 0.91 and specificity of 0.83. As an alternative, we also tried a logistic regression approach (average AUC 0.77, sensitivity of 0.71 and specificity of 0.80) (Figure 3b). The logistic regression model was based on a principal component analysis of all candidate species abundances, keeping the first 25 components. A ranked list of the candidate bacterial species with respect to their importance in the random forest model is provided in Supplementary Table 1. For the validation of an independent dataset, all 118 healthy and 38 sepsis samples were used as the training set, and samples from [BTR19] (No. PRJNA507824) were set as an independent validation set. The AUC shows that the proposed method also performed well in the independent dataset (Figure 3c).

4.2.2 Functional analysis of microbes from sepsis patients and healthy individuals

Using the bacterial abundance matrices from 78 healthy and 25 sepsis samples for training, we constructed two bacterial cooccurrence networks (Figure 4a). Each network contains 224 nodes, representing the 224 candidate pathogenic bacteria that were selected for having significantly different abundance distributions between healthy and sepsis samples. As mentioned above, blood can contain cfDNA fragments released by the bacteria inhabiting all human body sites. Thus, we expect the cooccurrence networks of healthy and sepsis samples to include some associations among “harmless” species that are generally not involved in sepsis. To focus on sepsis-specific associations, we generated a differential network by excluding all association patterns also found in the healthy cooccurrence network from the sepsis cooccurrence network (Figure 4a). We found 19 clusters (Figure 4b) of species in the differential network, which are the strongly connected components visible in Figure 4a. In the 25 sepsis samples, all the species in a cluster were strongly correlated in terms of their abundance levels. Detailed cluster information is provided in Supplementary Table 2.

To analyze the biological features of the clusters, we characterized the species in each cluster according to three aspects: respiration mode, metabolic habitat, and growth rate.

First, among all the candidate pathogen species, 35.52%, 3.66%, and 52.12% were anaerobic, aerobic, and facultative, respectively (the remaining 8.7% were unknown). Most of the clusters showed similarity in terms of respiration mode: 9 clusters exhibited a preference for facultative species (clusters 3, 5, 6, 10, 14, 15, 16, 17 and 19), and 7 clusters exhibited a preference for anaerobic species (clusters 1, 2, 7, 11, 12, 13 and 18). The few anaerobic species in the sample did not dominate any cluster.

Second, before causing infection in blood, these bacteria usually originate in specialized metabolic environments. Bacterial metabolic habitats are divided into 4 types: host-associated, terrestrial, aquatic, and diverse. The species in clusters 3, 4, 5, 9, 14, 15, 17, 18, and 19 were mainly host-associated, the species in cluster 10 were mainly terrestrial, the species in cluster 3 were mainly aquatic, and clusters 1, 6, 7, 10, 12, 13, and 16 contained species from diverse metabolic environments.

Third, bacterial growth was significantly correlated with metabolic variability and the level of cohabitation. Analysis of the doubling-time data led to the important finding that variations in the expression levels of genes involved in translation and transcription influenced the growth rate [Roc04] [CR06]. We partitioned the clusters into two groups according to the doubling time of their member species: "fast"- and "slow"-growing clusters are those whose median duplication time is shorter or longer, respectively, than the mean over all species by at least one standard deviation [FKB09]. The median doubling time for species distributed in clusters 6, 7, 11 and 13 was larger than 1 (fast-growing clusters), while the doubling time for members in clusters 1, 3, 4, 5, 15, and 16 was smaller than 0.6 (slow-growing clusters). Notably, fast growth rates are typical of species that exhibit ecological diversity, so the identification of "fast" clusters is consistent with the metabolic habitats analyzed in the previous paragraph.

4.2.3 Cancer detection and localization based on the cfDNA microbiome from cfMethyl-seq data

As shown in Figure 1a and 1c, we developed a two-step approach for rapid sepsis diagnosis. First, we aligned nonhuman reads from cfMethyl-seq data of 204 healthy individuals and 280 cancer patients to a hand-curated microbe database (see section 4.4). These samples were randomly partitioned into three groups: 310 samples (from 210 cancer patients and 100 individuals without cancer) for training, 103 samples (from 70 cancer patients and 33 individuals without cancer) for testing, and 30 samples (from 30 individuals without cancer) for normalization of the microbial abundance. For a robust performance evaluation, we repeated this split scheme 10 times and reported the average prediction performance.

Second, we trained a random forest model for cancer detection, i.e., classification of plasma samples from cancer patients and healthy individuals. Our model achieved an AUROC of 0.93 (aggregated from prediction results from 10 random splits, with a 95% confidence interval [0.91, 0.95], Figure 5) on the testing set.

Third, we trained a random forest model for cancer location, i.e., tissue of origin classification of plasma samples from cancer patients with different cancer types. We used the same strategy to evaluate the performance of cancer location prediction on 280 cfDNA samples from the five cancer types. Among the 10 runs, we achieved an average accuracy of 0.63 (standard deviation 0.06) on the testing sets (Table 1). Specifically, the average prediction accuracy of COAD/LIHC/LUAD/LUSC/STAD was 63/49/81/51/28% (Table 1).

Our results suggest the potential of the cfDNA microbiome in cancer detection and location.

4.2.4 Functional analysis of microbes from cancer patients and individuals without cancer

Given the discriminative power of the microbial profiles, we analyzed the abundance of the 200 most important microbes in the pancancer classifier. We performed Mann-Whitney U tests for each of the 200 microbes in the plasma samples from 204 healthy individuals and 280 cancer patients. For each of the 200 most important microbes in the cancer location classifier, we performed Kruskal-Wallis tests on the plasma samples of cancer patients with the five different cancer types (67 COAD, 49 LUSC, 77 LUAD, 47 LIHC, and 40 STAD). A large proportion of the 200 most important microbes showed significantly differential abundances between healthy individuals and cancer patients (73% in the cancer detection classifier, Figure 6a) or among different cancer types (48% in the cancer location classifier, Figure 6b).

In addition to the difference in the abundance of the important microbes in the plasma samples, we also analyzed the abundance of these microbes in 164 solid tumor samples (26 COAD, 32 LUSC, 42 LUAD, 26 LIHC, and 38 STAD). For each of the top 200 microbes in the cancer location classifier, we performed Kruskal-Wallis tests on the solid tumor samples from the five cancer types. The abundance of 38% of the microbes was significantly different among the solid tumor samples from the five cancer types (Figure 6c). We found that the microbes in the plasma were partially consistent with the microbes in the solid tumor. For example, eight *Bacteroides* species were found to be overabundant (Kruskal-Wallis test, $p = 0$) in tumor samples from colon cancer patients. One of the eight species also showed significant differences (the Kruskal-Wallis test, $p = 0.022$) among the plasma samples from patients with the five cancer types, while the other species did not show strong differences among plasma samples. One *Pseudomonas* species also showed significant differences in both the plasma samples and the solid tumor samples (Kruskal-Wallis test, $p = 0.019$ in the plasma samples and $p = 0.005$ in the solid tumor samples). Both *Bacteroides spp.* and *Pseudomonas spp.* were previously found to be enriched in mucosal samples and tumor samples from colorectal cancer patients [ZRR14] [FLB17] [SLJ12]. A species from the *Comamonadaceae* family, which is overabundant in other cancer types (e.g., breast cancer [UGB16]), also showed significant differences in both the plasma samples and the solid tumor samples (Kruskal-Wallis test, p = 0.037 in the plasma samples and p = 0.0002 in the solid tumor samples). Therefore, our analysis indicated that the abundance of the important microbes in the plasma samples was partially consistent with that in the solid tumor samples and the previous findings in the microbiome analysis from cancer patients. As many tissues can release DNA to the bloodstream, the difference between the plasma samples and the solid tumor samples might result from the various tissue sources of fragmented microbial DNA in the blood.

4.3 Discussion

We developed a workflow for disease detection using the microbiome composition in cfDNA. Specifically, we focused on rapid sepsis diagnosis and noninvasive cancer detection. Following the general workflow, we developed an approach for sepsis diagnosis and pathogen identification using cfDNA sequencing data mapped to bacterial genomes. This approach does not require cultivation, greatly enhancing the efficiency of diagnosis, with an AUC of 93%, and shortening the estimated turn-around time to approximately a day. Thus, it overcomes the limitations of the current culture-based diagnosis methods and fulfills the urgent need for timely diagnosis for sepsis patients. In addition to sepsis, we trained random forest models for noninvasive cancer detection and location using the cfMethyl-seq data. We showed that the microbiome composition derived from cfDNA can achieve an AUC of 93% for cancer detection and an accuracy of 0.62 for cancer location. As an orthogonal data source, the microbiome composition from cfDNA may further enhance the current noninvasive cancer diagnosis methods, which are mainly built on the human genome or epigenome.

Furthermore, functional analysis of the candidate pathogens in sepsis patients suggests the potential to guide therapy selection based on the detected pathogen clusters, which show clear characteristics in terms of respiration mode, metabolic habitat, and growth rate. The microbes with high importance in the random forest model of cancer detection and location are consistent with the previous findings and show statistically significant differences between cancer patients and healthy individuals or among different cancer types. Despite the good performance that we observed for rapid sepsis diagnosis and noninvasive cancer detection, further investigation of the relationships among the cfDNA-derived microbiome, the disease-related tissue-specific microbiome and the origin of the observed fragmented microbial DNA is needed. Nevertheless, the results indicate that the microbiome composition identified by plasma cfDNA analysis could provide extra information about the human host. The microbiome information in cfDNA could possibly offer additional evidence for cfDNA-based disease diagnosis, which is now accepted as a promising, noninvasive tool in disease detection.

4.4 Methods

4.4.1 Sample collection and processing

The cfDNA whole-genome sequencing (WGS) data used in this study were taken from 38 sepsis and 118 healthy samples. The raw sequencing reads were derived from two previously published data sources: 38 sepsis and 15 healthy samples from the European Nucleotide Archive (ENA, study No.: PRJEB13247 [GSG16], 103 healthy samples from the European Genome-phenome Archive (EGA, accession No. EGAS00001001754 [UHS16]), and 165 asymptomatic samples and 187 symptomatic samples from the European Nucleotide Archive (ENA, study 3, No. PRJNA507824 [BTR19]). Samples from both studies were taken from plasma and then sequenced by whole-genome and single-end sequencing.

The cfDNA cfMethyl-seq sequencing data used in this study were taken from 280 cancer patients and 163 individuals without cancer. The plasma samples of cancer patients were either collected at UCLA hospitals (i.e., Ronald Reagan UCLA Medical Center, UCLA Medical Plaza, or UCLA Santa Monica hospital) or purchased from BioPartners Inc. All plasma samples of cirrhotic patients without cancer were collected from patients at UCLA hospitals. All plasma samples of individuals without cirrhosis and cancer were collected from UCLAs Institute for Precision Health or purchased from BioPartners, Inc. (Woodland Hills, CA), or BioChain Institute, Inc. (Newark, CA). All solid normal and tumor tissue samples were either collected from the UCLA Translational Pathology Core Laboratory or purchased from Biopartners, Inc., Biochain Institute, Inc., Origene, Inc., or the Gundersen Health System. The enrollment criteria at UCLA hospitals were as follows: (1) at least 18 years old, (2) able to give consent, and (3) either not a cancer patient or diagnosed with colon cancer, liver cancer, lung cancer, or stomach cancer. The institutional review board (IRB) of the University of California, Los Angeles, approved the study. We obtained informed consent from the patients. cfDNA was extracted from plasma samples with a Qiagen QIAamp Circulating Nucleic Acid Kit (Catalog# 55114, Germantown, MD) by following the manufacturers protocol. The amount of starting material was 2-10 ml of plasma for healthy controls and 1-4 ml of plasma for cancer samples. The solid-tissue gDNA samples were extracted with a Qiagen Blood and Tissue Kit (Catalog# 69506). Approximately 100-200 ng of tissue was used to extract gDNA from each sample. The cfMethyl-seq libraries were constructed for all cfDNA samples; the RRBS libraries were constructed for all tissue gDNA samples.

4.4.2 Removal of human-like reads from WGS and cfMethyl-seq data

Because the cfDNA samples were processed with different protocols, namely, WGS and cfMethyl-seq, the sequencing data were processed in different ways.

For the WGS data, the raw reads from ENA (PRJEB13247 and PRJNA507824) were cleaned of human-like reads and reads with low complexity stretches using the NextGenMap tool. For the EGA data (EGAS00001001754), the raw sequencing reads were preprocessed to remove human and human-like reads using the fast alignment program *Bowtie2* [LS12]. The raw sequencing reads were aligned to the human genome sequence (UCSC Human hg19 reference genome). All mapped reads were regarded as human-like and removed. The remaining unmapped reads were of nonhuman origin and were used for microbiome analyses.

For the cfMethyl-seq data, the cfDNA was bisulfite-converted, so the unmethylated Cs were converted to Ts in the sequencing data. The raw sequencing reads were aligned to the human genome sequence (UCSC Human hg19 reference genome) using *Bismark* [KA11], which is a widely used aligner for bisulfite sequencing data. The unmapped reads were of nonhuman origin and were used for microbiome analyses.

4.4.3 Read alignment and microbe abundance quantification for WGS data

The nonhuman sequencing reads were aligned to a microbial genome sequence database using Centrifuge [KSB16], an open-source microbial classification engine that enables rapid and accurate labeling of reads and quantification of species. Specifically, the mapping was based on a database of compressed microbial sequences provided by Centrifuge (https://ccb.jhu.edu/software/centrif

Traversing up a taxonomic tree, Centrifuge maps reads to taxon nodes and assigns a "species abundance" to each taxonomic category. The abundances are the estimated fractions $\alpha = (\alpha_1, \alpha_2, \cdots, \alpha_S)$ that maximize a likelihood function, i.e.,

$$\alpha = \arg\max_{\alpha}(L),$$

with the likelihood L given by

$$L(\alpha) = \prod_{i=1}^{R} \sum_{j=1}^{S} (\frac{\alpha_j l_j}{\sum_{k=1}^{S} \alpha_k l_k} C_{ij}).$$

R is the number of reads, S is the number of species, α_j is the abundance of species j ($\sum_{j=1}^{S} \alpha_j = 1$, $0 < \alpha_j < 1$), and l_j is the average length of the genomes of species j. The coefficient C_{ij} is 1 if read i is classified as species j and 0 otherwise. The abundance vector α is obtained through an expectation maximization (EM) procedure.

4.4.4 Identification of candidate pathogenic bacteria for sepsis

To detect an abnormal bacterial abundance in a cfDNA sample, we need to first establish the background distribution of abundances under healthy conditions. We fit the expected abundance of each species in healthy samples with a beta distribution (this is a family of continuous probability distributions defined on the interval [0, 1] and parametrized by two positive parameters). Specifically, for each bacterial species j, its observed abundance values across a training set of healthy samples were used to fit a species-specific beta distribution defined by the parameters aj and bj.

To determine whether bacterial species j is a candidate pathogen, we compare the abundance value α_j from a new sample (healthy or sepsis) to the beta distribution. Specifically, we calculate the probability P to observe an abundance higher than α_j assuming that the sample is healthy:

$$\mathrm{P}(x \geq \alpha_j | a, b) = \frac{\int_{\alpha_j}^{1} u^{a_j - 1}(1-u)^{b_j - 1} du}{\int_{0}^{1} u^{a_j - 1}(1-u)^{b_j - 1} du}.$$

If P is very small, then we can reject the hypothesis that the observed abundance of this bacterial species in this sample is produced by the beta distribution determined under healthy conditions and hence conclude that the abundance of this species is abnormally high and that it is a candidate pathogen for sepsis. A bacterial species is classified as a candidate pathogen for sepsis in our study if it meets this condition for at least one of the sepsis samples.

4.4.5 Read alignment and microbe abundance quantification for cfMethyl-seq data

Since sequencing reads from cfMethyl-seq are bisulfite-converted, standard tools designed for abundance quantification in metagenomics are not suitable for cfMethyl-seq data. Methylation in microbes is different from methylation in humans, but a fraction of CpG sites were observed to be methylated in microbial genomes [WCB13]. Thus, we did not assume fully unmethylated microbial genomes in the cfDNA. To fully consider the potential bisulfite conversion in the microbial genomes, sequencing reads that did not align to the human genome were mapped against a hand-curated database, which contained a total of 1017 bacterial genomes, 1 archaeal genome, 453 eukaryotic genomes and 149 viral genomes from the literature. The 1620 microbial genomes were merged into 13 large FASTA files to save computing costs during alignment and prepared by *bismark_genome_preparation* before alignment. The unmapped reads were aligned to the merged microbial genome using Bismark with default parameters. *UMI-Grinder* (https://github.com/FelixKrueger/Umi-Grinder) was used to remove PCR duplicates based on the UMIs written into the read names during the UMI reformatting step and the mapping location. After alignment, uniquely mapped reads were counted for each microbial genome. The abundance of a microbe in a cfDNA sample was calculated as the uniquely mapped read count divided by the total number of sequencing reads in the sample and the size of the microbial genome. The abundance was then scaled by multiplying a large integer (10^9) to avoid small floats. We then normalized each abundance by standardizing it with the 30 reference cfDNA samples from individuals without cancer, i.e., $\frac{(\tau_{\text{sample}}-\mu_{\text{reference}})}{\sigma_{\text{reference}}}$, where τ_{sample} is the abundance of a bacterial or viral genome in the cfDNA sample, and $\mu_{\text{reference}}$ and $\sigma_{\text{reference}}$ are the mean and standard deviation of the abundance in the same microbial genome among the 30 reference cfDNA samples from individuals without cancer.

4.4.6 Random forest models for sepsis diagnosis, cancer diagnosis and cancer typing

Random forest is an effective classification method that generates many binary decision trees and aggregates their results.

To distinguish sepsis patients and healthy individuals, we trained a random forest model on the abundance of the candidate pathogens for sepsis. The WGS samples were randomly partitioned into two groups: 103 samples (78 healthy samples and 25 sepsis samples) for training, 53 samples (40 healthy samples and 13 sepsis samples) for testing, and 352 samples (165 asymptomatic samples and 187 symptomatic samples) for independent validation. Due to the imbalanced sizes of the healthy and sepsis samples, a traditional random forest may yield biased predictions. Therefore, we employed repeated balanced subsampling to build our sepsis diagnosis model. Specifically, a random forest model was trained using *sklearn* [PVG11] on the training data with the following parameters: (1) bootstrapping was enabled; (2) class weights were balanced; (3) in each bootstrap subsampling, the maximum sample size was 30; (4) the number of decision trees was 500; and (5) other parameters were set as the default values. The final decision of the random forest is reached by aggregating the decisions of each tree with the majority vote. The trained random forest was then evaluated on the independent validation samples.

To distinguish cancer patients from individuals without cancer and predict their tumor tissue of origin, we trained two random forest models on the abundance of microbes. The cfMethyl-seq samples were randomly partitioned into three groups: 310 samples (from 210 cancer patients and 100 noncancer individuals) for training, 103 samples (from 70 cancer patients and 33 noncancer individuals) for testing, and 30 samples (from 30 noncancer individuals) for the normalization of the microbial abundance. The random partition was performed 10 times to obtain a stable evaluation of the performance. To detect cancer patients, we trained a random forest model using *sklearn* on the training data on each set of partitions with the following parameters: (1) class weights were balanced; (2) the number of decision trees was 2000; (3) the number of variables randomly selected at each split was $\sqrt{\text{number of features}}$; and (4) other parameters were set as default values. To classify the tumor tissue of origin, we trained a random forest model for multiclass classification using *sklearn* on the training data on each set of partitions with the following parameters: (1) class weights were balanced; (2) the number of decision trees was 2000; (3) the number of variables randomly selected at each split was $\sqrt{\text{number of features}}$; and (4) other parameters were set as default values. For every plasma sample from cancer patients, the random forest model calculated a prediction score for each of the five cancer types, i.e., COAD, LUAD, LUSC, LIHC, and STAD. A prediction for tissue of origin is made if the ratio between the max prediction score and the second max prediction score is greater than 1.5; otherwise, the sample is considered indecisive.

The testing data were left untouched for independent validation of the performance.

4.4.7 Construction of cooccurrence networks for sepsis patients and healthy individuals

The bacterial DNA fragments in human blood may be shed from many species [KCK17]. These living bacteria are naturally present throughout the human body, from skin to viscera, and even in environments previously considered sterile, such as blood in circulation [PAM14]. It is of great importance to understand how DNA fragments from different species with different habitats come together. Strong inter-taxa associations in the data may indicate a community (even including different domains of life, such as bacteria and archaea) originating in a common niche space or perhaps direct symbioses between community members. Such information is particularly valuable in environments in which the basic ecology and life history strategies of many microbial taxa remain unknown. In addition, exploration of the cooccurrence patterns between different microorganisms can help identify potential biotic interactions, habitat affinities, or shared physiologies that could guide the development of more focused studies or experimental settings [BBC12].

A cooccurrence network is a visualization of relationships among entities that usually appear together. For example, it can be used to study the distribution of biotic populations [WBS14], to predict cancer risk [ZW18] or to analyze text collections [GGG12]. We constructed a cfDNA-based bacterial cooccurrence network, where two species are considered to cooccur if their abundances estimated from cfDNA are strongly correlated. Each node in the network represents a bacterial species, while each edge represents a cooccurring relationship.

To construct a bacterial cooccurrence network, we first generated two matrices: (1) the observed abundance matrix O (with n species, m samples) and (2) the expected abundance matrix N (also with n species, m samples). The latter is filled at each located sample as predicted by a regional species distribution model (e.g., a leave-one-out LOESS model [BBC12]). An $n \times n$ covariance matrix is calculated from either O or N by comparing rows (i.e., the abundances of 2 species across all samples). From the inverse of this covariance matrix (Σ), the partial correlation C_{ij} between a pair of bacterial species is calculated as follows:

$$C_{ij}(M) = \frac{-\Sigma_{ij}^{-1}(M)}{\sqrt{\Sigma_{ii}^{-1}(M)\Sigma_{jj}^{-1}(M)}},$$

where M is an $n \times m$ input matrix (O or N).

Both $C(O)$ and $C(N)$ were computed based on the above equation. Then, the standard effect of the correlation between O and N was calculated by rescaling $C(O)$ and $C(N)$. Finally, significant associations were found by calculating the p-value of the correlation coefficient for each pair of species i and j, with the null hypothesis being that the observations are uncorrelated. Finally, our cooccurrence network was generated by placing edges between each pair of bacterial species with a significant link. The detailed algorithm of network construction is described in [MBS16].

Split ID	Overall	COAD	LUAD	LUSC	LIHC	STAD
1	0.53	0.50	0.50	0.77	0.33	0.00
2	0.61	0.78	0.40	0.79	0.50	0.00
3	0.70	0.50	0.71	1.00	0.67	0.40
4	0.61	0.43	1.00	0.57	0.50	0.00
5	0.72	1.00	0.33	0.77	0.60	0.60
6	0.67	0.33	0.50	0.88	0.50	0.50
7	0.63	0.67	0.50	0.75	0.71	0.33

8	0.64	0.75	0.50	0.75	0.40	0.50
9	0.56	0.57	0.17	0.87	0.50	0.29
10	0.64	0.78	0.33	1.00	0.40	0.20

Table 4.1: Testing accuracy of the cancer location model on the ten random splits. The second column shows the overall accuracy of all five cancer types. The third to seventh columns show the accuracy for individual cancer types, which is the fraction of patients with correctly predicted specific cancer type.

Order	Bacteria species	MeanDecreaseGini
1	*Enterococcus faecium*	0.87608067
2	*Escherichia coli*	0.71117135
3	*Pseudomonas sp. TKP*	0.66447362
4	*Delftia tsuruhatensis*	0.65595001
5	*Xanthomonas campestris*	0.37769903
6	*Pseudomonas aeruginosa*	0.33547929
7	*Bacteroides fragilis*	0.30124156
8	*Pseudomonas pseudoalcaligenes*	0.25519889
9	*Gardnerella vaginalis*	0.24256536
10	*Staphylococcus aureus*	0.24053659
11	*Pseudomonas putida*	0.23099539
12	*Streptococcus mitis*	0.21879577
13	*Propionibacterium sp. oral taxon 193*	0.2152918
14	*Lawsonella clevelandensis*	0.21400124
15	*Mycoplasma mycoides*	0.21099248
16	*Lactobacillus sakei*	0.20126349
17	*Acinetobacter baumannii*	0.19489872

18	*Pseudomonas mendocina*	0.19331255
19	*Delftia sp. Cs1-4*	0.17978586
20	*Bifidobacterium adolescentis*	0.17706793
21	*Streptococcus salivarius*	0.17694939
22	*Rhodococcus erythropolis*	0.16457406
23	*Comamonas testosteroni*	0.16025665
24	*Enterobacter cloacae*	0.1599143
25	*Klebsiella pneumoniae*	0.1584747
26	*Thermus scotoductus*	0.15744758
27	*Rothia mucilaginosa*	0.14895304
28	*Cupriavidus metallidurans*	0.14692427
29	*Rothia dentocariosa*	0.14353751
30	*Alicycliphilus denitrificans*	0.1392072
31	*Roseburia hominis*	0.13707822
32	*Elizabethkingia miricola*	0.13525933
33	*Acidovorax sp. RAC01*	0.13150666
34	*Staphylococcus pasteuri*	0.12650772
35	*Streptococcus parasanguinis*	0.12640767
36	*Bacteroides ovatus*	0.12301837
37	*Burkholderia cepacia*	0.11336059
38	*Sphingomonas sp. MM-1*	0.11327717
39	*Bradyrhizobium sp. S23321*	0.11002664
40	*Paraburkholderia phytofirmans*	0.10998147
41	*Pediococcus pentosaceus*	0.10624074
42	*Cutibacterium avidum*	0.10558094
43	*Pseudomonas antarctica*	0.10112312

44	*Streptococcus pneumoniae*	0.10109002
45	*Klebsiella sp. LTGPAF-6F*	0.09607979
46	*Thermus parvatiensis*	0.09532804
47	*Staphylococcus haemolyticus*	0.09349563
48	*Bacillus cereus*	0.09160156
49	*Thermus thermophilus*	0.09159278
50	*beta proteobacterium CB*	0.09100694
51	*Leuconostoc mesenteroides*	0.08929793
52	*Pseudomonas stutzeri*	0.08780254
53	*Burkholderia sp. OLGA172*	0.08748431
54	*[Eubacterium] eligens*	0.08709432
55	*Staphylococcus saprophyticus*	0.08453405
56	*Bacteroides thetaiotaomicron*	0.08329281
57	*Enterobacter sp. HK169*	0.08150616
58	*Janthinobacterium sp. 1_2014MBL_MicDiv*	0.08144096
59	*Acinetobacter sp. TTH0-4*	0.08094619
60	*Kytococcus sedentarius*	0.08091638
61	*Shigella boydii*	0.0784422
62	*Cupriavidus necator*	0.0709098
63	*Staphylococcus argenteus*	0.06985492
64	*Geobacillus sp. 12AMOR1*	0.06619285
65	*Leuconostoc carnosum*	0.06254964
66	*Veillonella parvula*	0.06240773
67	*Alistipes finegoldii*	0.0619031
68	*Klebsiella oxytoca*	0.05933975
69	*Pseudomonas resinovorans*	0.05539547

70	*Collimonas fungivorans*	0.05425206
71	*Ralstonia insidiosa*	0.05422794
72	*Corynebacterium diphtheriae*	0.05386443
73	*Ruminococcus bicirculans*	0.0536176
74	*Limnohabitans sp. 103DPR2*	0.05199756
75	*Haemophilus parainfluenzae*	0.05197863
76	*Finegoldia magna*	0.05057758
77	*Sphingobium sp. EP60837*	0.04996478
78	*Sphingopyxis fribergensis*	0.04992504
79	*Haemophilus influenzae*	0.04974396
80	*Shewanella sp. ANA-3*	0.04705159
81	*Streptococcus gordonii*	0.045204
82	*Lactobacillus johnsonii*	0.04479736
83	*Enterobacter cloacae complex Hoffmann cluster III*	0.0440336
84	*Sphingobium baderi*	0.04153171
85	*Pseudomonas mandelii*	0.04068203
86	*Streptococcus sp. I-P16*	0.04061352
87	*Prevotella denticola*	0.04037215
88	*Flavobacterium sp. PK15*	0.04015331
89	*Enterobacter asburiae*	0.04012392
90	*Corynebacterium ureicelerivorans*	0.03939868
91	*Cronobacter sakazakii*	0.0393335
92	*Citrobacter koseri*	0.03923981
93	*Streptococcus sp. oral taxon 431*	0.03712368
94	*Azospira oryzae*	0.03674209

95	*Pseudomonas sp. CCOS 191*	0.03654554
96	*Rhodopseudomonas palustris*	0.03575288
97	*Sphingobium sp. RAC03*	0.03457213
98	*Enterobacter hormaechei*	0.03409861
99	*Parabacteroides distasonis*	0.03332963
100	*Streptococcus sanguinis*	0.03274637
101	*Pseudomonas trivialis*	0.03242779
102	*Janthinobacterium agaricidamnosum*	0.0323105
103	*Enterococcus faecalis*	0.03132254
104	*Streptococcus oralis*	0.03108336
105	*Filifactor alocis*	0.03067699
106	*Xanthomonas axonopodis*	0.03062822
107	*Ramlibacter tataouinensis*	0.03051396
108	*Acinetobacter nosocomialis*	0.02943133
109	*Shewanella frigidimarina*	0.02893095
110	*Burkholderia seminalis*	0.02891773
111	*Carnobacterium maltaromaticum*	0.02824715
112	*Bacteroides cellulosilyticus*	0.02764164
113	*Bradyrhizobium oligotrophicum*	0.02761387
114	*Propionibacterium freudenreichii*	0.02739786
115	*Burkholderia multivorans*	0.02731146
116	*Barnesiella viscericola*	0.02720064
117	*Lactobacillus buchneri*	0.02719418
118	*Ralstonia pickettii*	0.0269362
119	*Shigella dysenteriae*	0.02646106
120	*Fusobacterium nucleatum*	0.0264442

121	*Corynebacterium aurimucosum*	0.02634571
122	*Methylobacterium sp. C1*	0.02595359
123	*Leuconostoc gelidum*	0.02594945
124	*Corynebacterium kroppenstedtii*	0.02590795
125	*Polynucleobacter asymbioticus*	0.0257555
126	*Pseudoalteromonas luteoviolacea*	0.02470581
127	*Verminephrobacter eiseniae*	0.02461531
128	*Acinetobacter pittii*	0.02431786
129	*Caulobacter segnis*	0.02362301
130	*Bifidobacterium bifidum*	0.02255195
131	*Weissella cibaria*	0.02249466
132	*Variovorax paradoxus*	0.02247406
133	*Shigella flexneri*	0.02227876
134	*Streptococcus sp. A12*	0.02226548
135	*Akkermansia muciniphila*	0.0218752
136	*Brevundimonas subvibrioides*	0.02153242
137	*Bacillus pseudofirmus*	0.02091149
138	*Streptococcus pseudopneumoniae*	0.02031545
139	*Staphylococcus equorum*	0.02009038
140	*Morganella morganii*	0.01982859
141	*Lactobacillus acidophilus*	0.0194956
142	*Bradyrhizobium icense*	0.01915556
143	*Pseudomonas balearica*	0.01906781
144	*Thiomonas intermedia*	0.01892311
145	*Streptococcus intermedius*	0.01886052
146	*Corynebacterium singulare*	0.01731943

147	*Rhodoluna lacicola*	0.01617393
148	*Sodalis glossinidius*	0.01557276
149	*Dechloromonas aromatica*	0.01504293
150	*Eggerthella lenta*	0.01411821
151	*Streptococcus sp. I-G2*	0.01403761
152	*Burkholderia ambifaria*	0.01323175
153	*Bacillus bombysepticus*	0.01309759
154	*Pseudomonas alcaligenes*	0.01203509
155	*Raoultella ornithinolytica*	0.01188065
156	*Streptococcus constellatus*	0.01176466
157	*Pseudomonas rhizosphaerae*	0.01173927
158	*Corynebacterium urealyticum*	0.0117
159	*Lactobacillus brevis*	0.01132764
160	*Salmonella enterica*	0.01126597
161	*Paracoccus denitrificans*	0.0111644
162	*Pseudarthrobacter phenanthrenivorans*	0.01040193
163	*Escherichia fergusonii*	0.01032787
164	*Leclercia adecarboxylata*	0.00922222
165	*Burkholderia sp. RPE64*	0.00887778
166	*Pseudomonas sp. FGI182*	0.00868713
167	*Burkholderia sp. CCGE1003*	0.00857079
168	*Pseudomonas plecoglossicida*	0.00845887
169	*Enterobacter xiangfangensis*	0.00834994
170	*Shewanella sp. MR-4*	0.00832222
171	*Psychrobacter cryohalolentis*	0.00815079
172	*Burkholderia sp. KJ006*	0.00796854

173	*Thioalkalivibrio sulfidiphilus*	0.0079
174	*Pectobacterium carotovorum*	0.00777794
175	*Thermobispora bispora*	0.00738713
176	*Ornithobacterium rhinotracheale*	0.00711111
177	*Lachnoclostridium sp. YL32*	0.00702814
178	*Enterobacter kobei*	0.00654545
179	*Odoribacter splanchnicus*	0.00615327
180	*Nitrosomonas europaea*	0.00599394
181	*Acidipropionibacterium acidipropionici*	0.0059404
182	*Pseudomonas parafulva*	0.00593333
183	*Sphingorhabdus sp. M41*	0.00589286
184	*Leptotrichia buccalis*	0.00563636
185	*Neisseria gonorrhoeae*	0.00555128
186	*Enterobacter sp. FY-07*	0.00550327
187	*Corynebacterium efficiens*	0.00451128
188	*Parvimonas micra*	0.00451128
189	*Kosakonia sacchari*	0.00409091
190	*Bacillus licheniformis*	0.00394805
191	*Arcanobacterium haemolyticum*	0.00371429
192	*Peptoniphilus sp. 1-1*	0.0036359
193	*Mobiluncus curtisii*	0.00355556
194	*Bifidobacterium angulatum*	0.0035
195	*Anaerococcus prevotii*	0.00346609
196	*Pseudomonas denitrificans*	0.00330882
197	*Shewanella sp. MR-7*	0.00321905
198	*Bradyrhizobium sp. CCGE-LA001*	0.0032

199	*Caldicellulosiruptor lactoaceticus*	0.00316484
200	*Mycobacterium vanbaalenii*	0.003
201	*Campylobacter concisus*	0.00294545
202	*Aequorivita sublithincola*	0.00290909
203	*Streptococcus dysgalactiae*	0.00288235
204	*Treponema denticola*	0.00276491
205	*Corynebacterium uterequi*	0.00274286
206	*Pasteurella multocida*	0.00232727
207	*Lactobacillus plantarum*	0.00220915
208	*Flavobacterium branchiophilum*	0.00213333
209	*Bradyrhizobium sp. ORS 278*	0.00201667
210	*Thermus aquaticus*	0.002
211	*Leuconostoc citreum*	0
212	*Acinetobacter equi*	0
213	*Leuconostoc lactis*	0
214	*Renibacterium salmoninarum*	0
215	*Bacillus sp. OxB-1*	0
216	*Kocuria flava*	0
217	*Thalassolituus oleivorans*	0
218	*Isoptericola variabilis*	0
219	*Cloacibacillus porcorum*	0
220	*Burkholderia sp. CCGE1001*	0
221	*Parageobacillus thermoglucosidans*	0
222	*Exiguobacterium sibiricum*	0
223	*Pseudomonas oryzihabitans*	0
224	*Mycoplasma hominis*	0

Table 4.2: Importance of candidate bacterial species by Random Forest.

Cluster index	Bacteria	Possible antibiotics
Cluster 1	*Bifidobacterium adolescentis, Bacteroides thetaiotaomicron, Odoribacter splanchnicus, Bifidobacterium angulatum, Ornithobacterium rhinotracheale, Lactobacillus brevis, Pseudomonas rhizosphaerae, Flavobacterium branchiophilum, Bacillus pseudofirmus, Pasteurella multocida*	aminoglycoside, ciprofloxacin, raxibacuma, anti-pseudomonal penicillins such as ticarcillin, doxycycline, bordetella pertussis, metronidazole, penicillin
Cluster 2	*Sphingomonas sanxanigenens, Akkermansia muciniphila, Novosphingobium pentaromativorans, Sphingomonas wittichii, Pseudarthrobacter phenanthrenivorans, Serratia marcescens, Acidithiobacillus caldus, Achromobacter xylosoxidans, Cyanobium gracile*	aminoglycoside, cefotaxime, gentamicin, Anti-pseudomonal penicillins such as ticarcillin
Cluster 3	*Staphylococcus aureus, Sphingomonas sp. MM-1, Sphingobium sp. EP60837, Staphylococcus equorum, Klebsiella sp. LTGPAF-6F, Streptococcus gordonii, Prevotella denticola, Bradyrhizobium sp. ORS 278*	penicillin G, ciprofloxacin, 3rd generation cephalosporin, TMP/SMX, methicillin, oxacillin, vancomycin, nafcillin

Cluster 4	*Escherichia coli, Pseudomonas aeruginosa, Ralstonia pickettii, Burkholderia multivorans, Burkholderia ambifaria, Cronobacter sakazakii, Burkholderia sp. CCGE1003*	aminoglycoside, cefotaxime, gentamicin, Anti-pseudomonal penicillins such as ticarcillin, streptomycin, piperacillin
Cluster 5	*Propionibacterium freudenreichii, Streptococcus mutans, Rhizobium sp. IRBG74, Brevundimonas sp. GW460-12-10-14-LB2, Staphylococcus lugdunensis, Prevotella melaninogenica, Corynebacterium diphtheriae, Bacillus sp. ABP14, Polynucleobacter necessarius, Polyangium brachysporum, Enterococcus casseliflavus, Bacillus bombysepticus*	aminoglycoside, ciprofloxacin, penicillin G, Raxibacuma, vancomycin, doxycycline, TMP/SMX, ampicillin, penicillin, erythromycin
Cluster 6	*Barnesiella viscericola, Leuconostoc citreum, Corynebacterium singulare, Thermobispora bispora, Streptococcus dysgalactiae, Pseudomonas sp. FGI182, Thermus aquaticus, Renibacterium salmoninarum, Treponema denticola*	aminoglycoside, penicillin G, Anti-pseudomonal penicillins such as ticarcillin, doxycycline, erythromycin, penicillin
Cluster 7	*Corynebacterium ureicelerivorans, Corynebacterium kroppenstedtii, Nitrosomonas europaea, Arcanobacterium haemolyticum, Thioalkalivibrio sulfidiphilus, Pseudomonas parafulva*	aminoglycoside, Anti-pseudomonal penicillins such as ticarcillin, erythromycin, penicillin

Cluster 8	*Enterobacter hormaechei, Enterobacter xiangfangensis, Enterobacter kobei, Sodalis glossinidius, Enterobacter sp. FY-07*	amoxicillin and clavulanic acid + gentamicin/ciprofloxacin, second/third generation cephalosporin excluding ceftazidime gentamicin/ciprofloxacin, piperacillin/tazobactam
Cluster 9	*Mycoplasma mycoides, Bacteroides fragilis, Cupriavidus necator, Carnobacterium maltaromaticum, Lachnoclostridium sp. YL32, Verminephrobacter eiseniae, Sphingorhabdus sp. M41, Pseudoalteromonas luteoviolacea*	doxycycline, metronidazole, bordetella pertussis, erythromycin
Cluster 10	*Acinetobacter oleivorans, Leptothrix cholodnii, Yersinia intermedia, Thiomonas intermedia, Caulobacter sp. K31, Shewanella putrefaciens, Nitrobacter winogradskyi*	streptomycin, tetracyclin, penicillin, chloramphenicol, aminoglycosides
Cluster 11	*Ruminococcus bicirculans, Pseudomonas plecoglossicida, Pseudomonas denitrificans, Corynebacterium urealyticum, Shewanella sp. MR-7, Mycobacterium vanbaalenii, Thalassolituus oleivorans*	rifampicin, Anti-pseudomonal penicillins such as ticarcillin, isoniazid, aminoglycoside, pyrazinamide, ethambutol, erythromycin, penicillin
Cluster 12	*Bifidobacterium bifidum, Lactobacillus acidophilus, Bacillus cereus, Anaerococcus prevotii, Parvimonas micra, Pectobacterium carotovorum*	Doxycycline, Ciprofloxacin, penicillin, Raxibacuma

Cluster 13	*Pseudomonas balearica, Pseudomonas trivialis, Salmonella enterica, Pseudomonas alcaligenes, Burkholderia sp. RPE64, Aequorivita sublithincola*	Anti-pseudomonal penicillins such as ticarcillin, ciprofloxacin, ceftriaxone, aminoglycoside, TMP/SMX, azithromycin
Cluster 14	*Leuconostoc carnosum, Lactobacillus fermentum, Sphingopyxis alaskensis, Streptococcus anginosus, Leuconostoc sp. C2*	penicillin G
Cluster 15	*Eubacterium eligens, Psychrobacter cryohalolentis, Eggerthella lenta, Escherichia fergusonii, Peptoniphilus sp. 1-1*	cefotaxime, gentamicin
Cluster 16	*Thermus thermophilus, Enterococcus faecalis, Corynebacterium aurimucosum, Acinetobacter equi, Bacillus sp. OxB-1*	aminoglycoside, Ciprofloxacin, vancomycin, Doxycycline, erythromycin, ampicillin, penicillin, Raxibacuma
Cluster 17	*Pseudomonas mandelii, Streptococcus constellatus, Caldicellulosiruptor lactoaceticus, Leptotrichia buccalis*	Anti-pseudomonal penicillins such as ticarcillin, penicillin G, aminoglycoside
Cluster 18	*Xanthomonas campestris, Methylophaga frappieri, Leuconostoc lactis, Xanthomonas alfalfae*	second/third generation cephalosporin and metronidazole +/- gentamicin
Cluster 19	*Weissella cibaria, Streptococcus sp. I-P16, Enterobacter sp. HK169, Xanthomonas axonopodis*	penicillin G, amoxicillin and clavulanic acid +/- gentamicin

Table 4.3: The cluster information of the bacterial co-occurrence network.

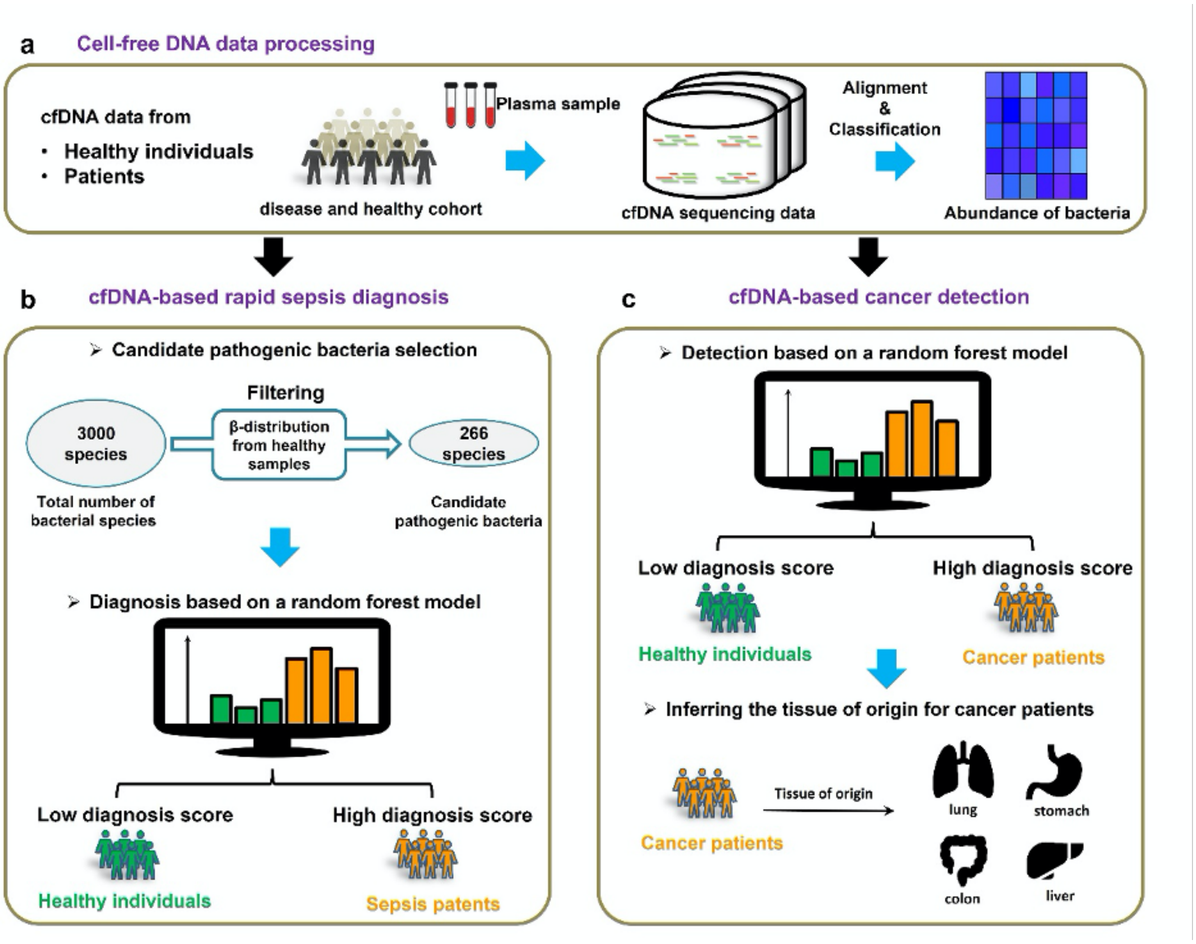

a
Cell-free DNA data processing
cfDNA data from
• Healthy individuals
• Patients
Plasma sample
disease and healthy cohort
cfDNA sequencing data
Alignment
&
Classification
Abundance of bacteria
b
cfDNA-based rapid sepsis diagnosis
➢ Candidate pathogenic bacteria selection
Filtering
3000 species
β-distribution from healthy samples
266 species
Total number of bacterial species
Candidate pathogenic bacteria
➢ Diagnosis based on a random forest model
Low diagnosis score
High diagnosis score
Healthy individuals
Sepsis patents
c
cfDNA-based cancer detection
➢ Detection based on a random forest model
Low diagnosis score
High diagnosis score
Healthy individuals
Cancer patients
➢ Inferring the tissue of origin for cancer patients
Tissue of origin
Cancer patients
lung
stomach
colon
liver

Figure 4.1: An illustration of our approach to rapid sepsis diagnosis and cancer detection based on cell-free DNA (cfDNA). (**a**) Illustration of cfDNA data processing pipeline. All human reads were removed from the cfDNA sequencing data by read alignment to the reference human genome. Nonhuman reads were assigned to different microbial genomes based on the sequence context. From the read assignment, the normalized abundance of the microbes was calculated. (**b**) Illustration of the rapid sepsis diagnosis approach. Our diagnosis strategy is a two-step procedure based solely on cfDNA from blood. First, we select candidate pathogenic bacterial species through statistical analysis (see section 4.4). Second, a random forest model is used to calculate a diagnosis score for each sample. (**c**) Illustration of the cancer detection approach. A random forest model is used to distinguish cancer patients and healthy individuals. For the cancer patients, a second random forest model is used to calculate the prediction score for the cancer type (tissue of origin, see section 4.4).

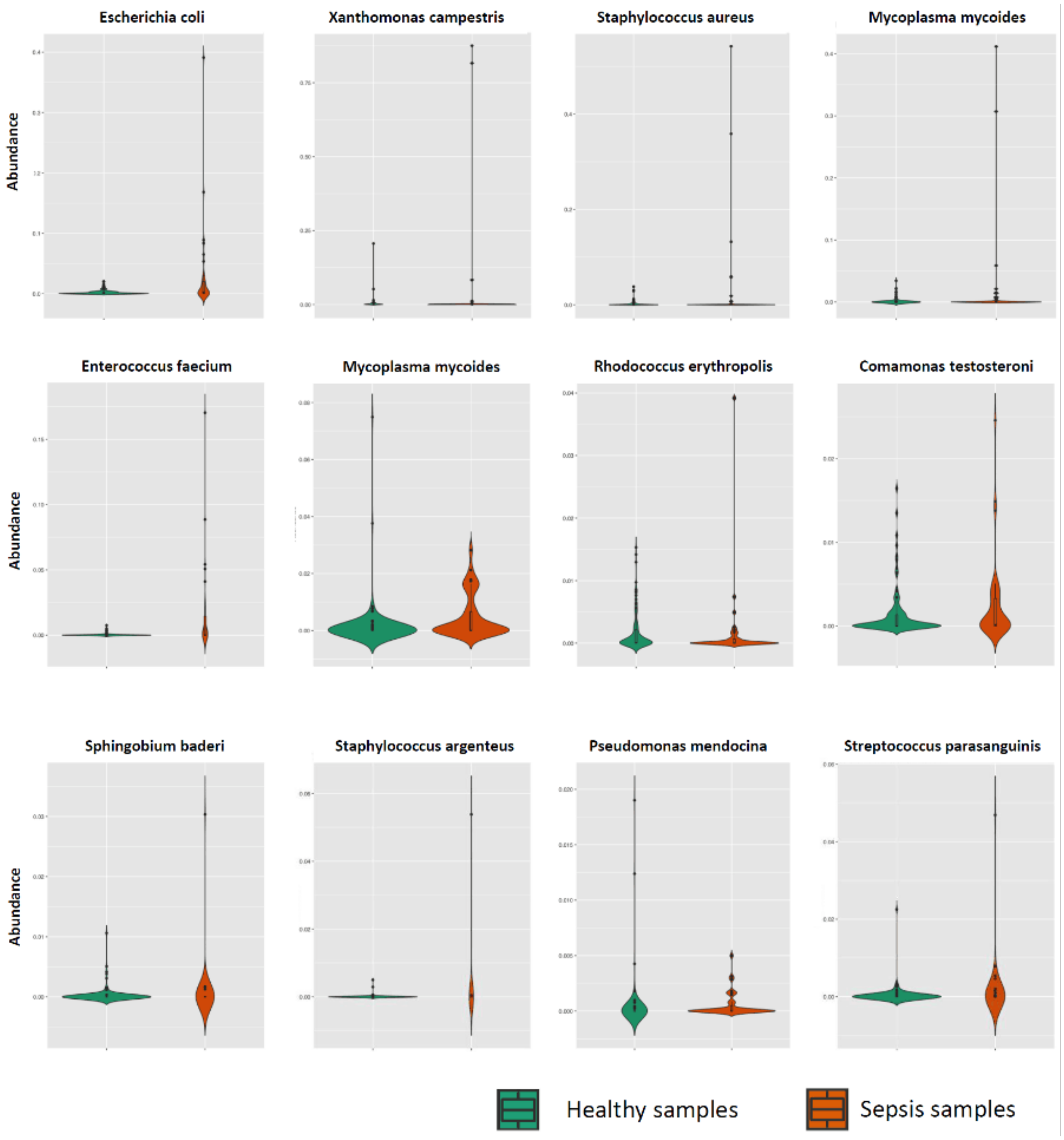


Figure 4.2: Differential abundances of some candidate pathogenic bacterial species in heathy and sepsis samples. The distributions of bacterial abundances for 12 candidate pathogens are visualized as violin plots.

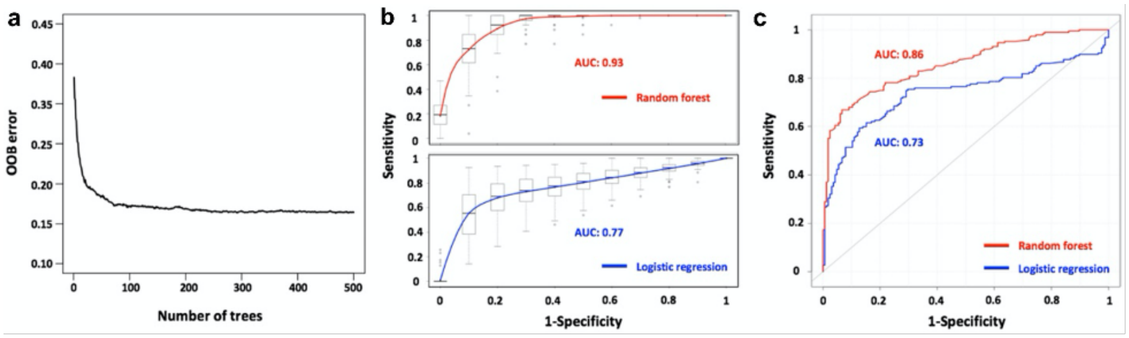


Figure 4.3: Performance of a random forest classifier with balanced subsampling for identifying sepsis samples and healthy samples. (**a**) The out-of-bag error converges to 0.16 if the number of decision trees is over 100. (**b**) The average AUC curves for our diagnosis strategy and a logistic regression scheme based on one-third of the samples reserved for testing the model. (**c**) The AUC curves of our diagnosis strategy (red) and a logistic regression scheme (blue) based on an independent validation dataset for validating the proposed algorithm.

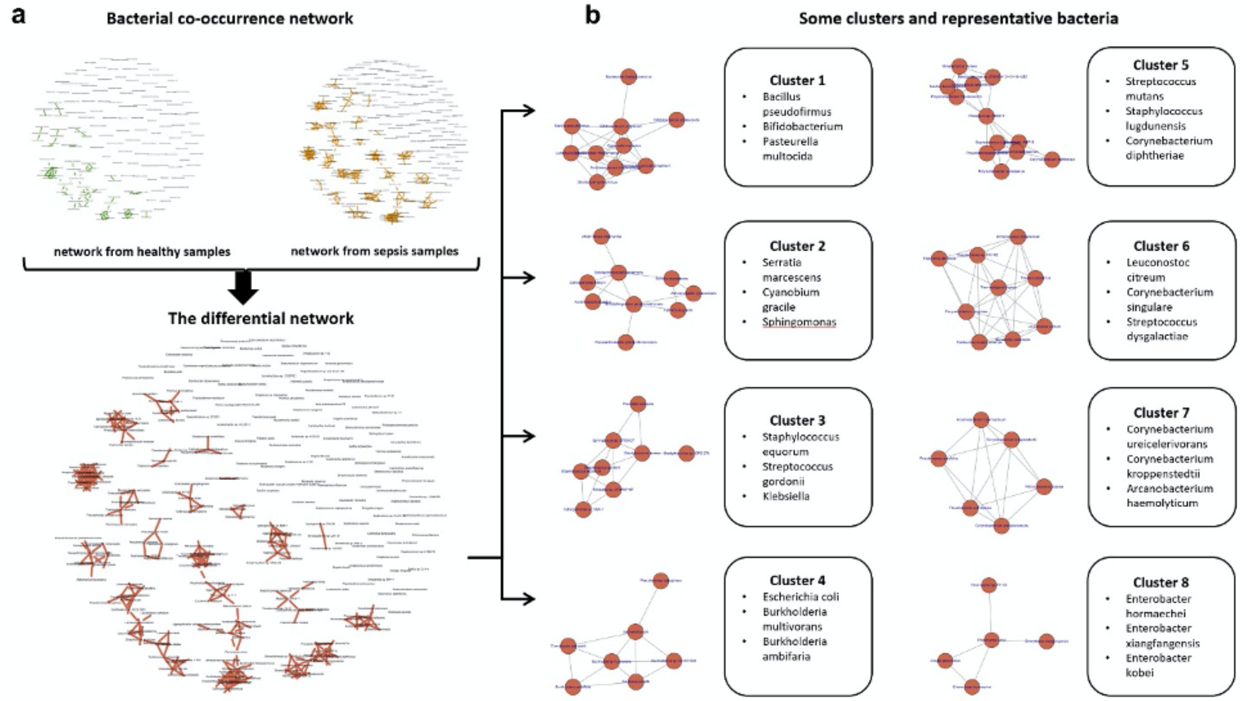

a
Bacterial co-occurrence network
network from healthy samples
network from sepsis samples
The differential network
b
Some clusters and representative bacteria
Cluster 1
• Bacillus pseudofirmus
• Bifidobacterium
• Pasteurella multocida
Cluster 2
• Serratia marcescens
• Cyanobium gracile
• Sphingomonas
Cluster 3
• Staphylococcus equorum
• Streptococcus gordonii
• Klebsiella
Cluster 4
• Escherichia coli
• Burkholderia multivorans
• Burkholderia ambifaria
Cluster 5
• Streptococcus mutans
• Staphylococcus lugdunensis
• Corynebacterium diphtheriae
Cluster 6
• Leuconostoc citreum
• Corynebacterium singulare
• Streptococcus dysgalactiae
Cluster 7
• Corynebacterium ureicelerivorans
• Corynebacterium kroppenstedtii
• Arcanobacterium haemolyticum
Cluster 8
• Enterobacter hormaechei
• Enterobacter xiangfangensis
• Enterobacter kobei

Figure 4.4: Bacterial cooccurrence networks constructed on the basis of cfDNA data from normal and sepsis samples. (**a**) The differential cooccurrence network describing associations between species that are only observed in the sepsis samples. (**b**) A partial list of clusters (connected components) from the differential network. For each cluster, the representative bacteria are listed.

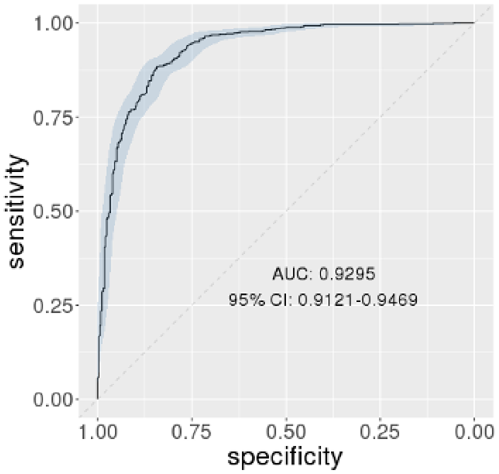


Figure 4.5: Performance of the cancer detection model. The receiver operating characteristic (ROC) curve of the random forest model in testing sets of all ten random splits. The blue band shows the confidence interval of the ROC curve.

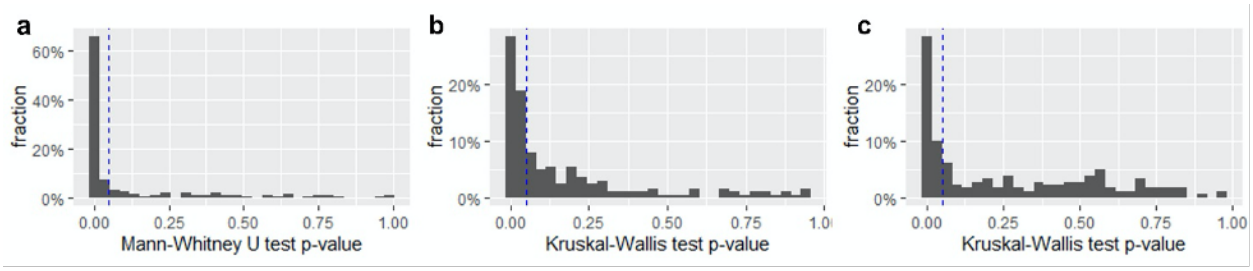


Figure 4.6: Histograms of p-values in statistical tests of the abundance of the 200 most important microbes from the cancer detection classifier and the cancer location classifier. (**a**) The histogram of p-values from the Mann-Whitney U test of the abundance of the 200 most important microbes in the plasma samples between healthy individuals and cancer patients. (**b**) The histogram of p-values from the Kruskal-Wallis test of the abundance of the 200 most important microbes in the plasma samples from patients with the five cancer types. (**c**) The histogram of p-values from the Kruskal-Wallis test of the abundance of the 200 most important microbes in the solid tumor samples from patients with the five cancer types.

REFERENCES

[ABW17] Christopher Abbosh, Nicolai J Birkbak, Gareth A Wilson, Mariam Jamal-Hanjani, Tudor Constantin, Raheleh Salari, John Le Quesne, David A Moore, Selvaraju Veeriah, Rachel Rosenthal, et al. "Phylogenetic ctDNA analysis depicts early-stage lung cancer evolution." *Nature*, **545**(7655):446–451, 2017.

[ACH13] Geraldine A Van der Auwera, Mauricio O Carneiro, Christopher Hartl, Ryan Poplin, Guillermo Del Angel, Ami Levy-Moonshine, Tadeusz Jordan, Khalid Shakir, David Roazen, Joel Thibault, et al. "From FastQ data to high-confidence variant calls: the genome analysis toolkit best practices pipeline." *Current protocols in bioinformatics*, **43**(1):11–10, 2013.

[AFG20] Chris Abbosh, Alexander Frankell, Aaron Garnett, Thomas Harrison, Morgan Weichert, Abel Licon, Selvaraju Veeriah, Bob Daber, Mike Moreau, Adrian Chesh, et al. "Abstract CT023: Phylogenetic tracking and minimal residual disease detection using ctDNA in early-stage NSCLC: A lung TRACERx study.", 2020.

[AHF17] Viktor A Adalsteinsson, Gavin Ha, Samuel S Freeman, Atish D Choudhury, Daniel G Stover, Heather A Parsons, Gregory Gydush, Sarah C Reed, Denisse Rotem, Justin Rhoades, et al. "Scalable whole-exome sequencing of cell-free DNA reveals high concordance with metastatic tumors." *Nature communications*, **8**(1):1–13, 2017.

[APD13] Alexander V Alekseyenko, Guillermo I Perez-Perez, Aieska De Souza, Bruce Strober, Zhan Gao, Monika Bihan, Kelvin Li, Barbara A Methé, and Martin J Blaser. "Community differentiation of the cutaneous microbiota in psoriasis." *Microbiome*, **1**(1):31, 2013.

[BBC12] Albert Barberán, Scott T Bates, Emilio O Casamayor, and Noah Fierer. "Using network analysis to explore co-occurrence patterns in soil microbial communities." *The ISME journal*, **6**(2):343–351, 2012.

[BHL20] Sara L Rassoulian Barrett, Elizabeth A Holmes, Dustin R Long, Ryan C Shean, Gilbert E Bautista, Sumedha Ravishankar, Vikas Peddu, Brad T Cookson, Pradeep K Singh, Alexander L Greninger, et al. "Cell free DNA from respiratory pathogens is detectable in the blood plasma of Cystic Fibrosis patients." *Scientific Reports*, **10**(1):1–6, 2020.

[BJP15] Timothy M Butler, Katherine Johnson-Camacho, Myron Peto, Nicholas J Wang, Tara A Macey, James E Korkola, Theresa M Koppie, Christopher L Corless, Joe W Gray, and Paul T Spellman. "Exome sequencing of cell-free DNA from metastatic cancer patients identifies clinically actionable mutations distinct from primary disease." *PloS one*, **10**(8):e0136407, 2015.

[BMS14] Kyle M Brawner, Casey D Morrow, and Phillip D Smith. "Gastric microbiome and gastric cancer." *Cancer journal (Sudbury, Mass.)*, **20**(3):211, 2014.

[BTR19] Timothy A Blauwkamp, Simone Thair, Michael J Rosen, Lily Blair, Martin S Lindner, Igor D Vilfan, Trupti Kawli, Fred C Christians, Shivkumar Venkatasubrahmanyam, Gregory D Wall, et al. "Analytical and clinical validation of a microbial cell-free DNA sequencing test for infectious disease." *Nature microbiology*, **4**(4):663–674, 2019.

[CBL19] Alexandre Pellan Cheng, Philip Burnham, John Richard Lee, Matthew Pellan Cheng, Manikkam Suthanthiran, Darshana Dadhania, and Iwijn De Vlaminck. "A cell-free DNA metagenomic sequencing assay that integrates the host injury response to infection." *Proceedings of the National Academy of Sciences*, **116**(37):18738–18744, 2019.

[CCC18] Mathieu Chicard, Leo Colmet-Daage, Nathalie Clement, Adrien Danzon, Mylène Bohec, Virginie Bernard, Sylvain Baulande, Angela Bellini, Paul Deveau, Gaëlle Pierron, et al. "Whole-exome sequencing of cell-free DNA reveals temporo-spatial heterogeneity and identifies treatment-resistant clones in neuroblastoma." *Clinical Cancer Research*, **24**(4):939–949, 2018.

[CCL17] Aadel A Chaudhuri, Jacob J Chabon, Alexander F Lovejoy, Aaron M Newman, Henning Stehr, Tej D Azad, Michael S Khodadoust, Mohammad Shahrokh Esfahani, Chih Long Liu, Li Zhou, et al. "Early detection of molecular residual disease in localized lung cancer by circulating tumor DNA profiling." *Cancer discovery*, **7**(12):1394–1403, 2017.

[CLC13] Kristian Cibulskis, Michael S Lawrence, Scott L Carter, Andrey Sivachenko, David Jaffe, Carrie Sougnez, Stacey Gabriel, Matthew Meyerson, Eric S Lander, and Gad Getz. "Sensitive detection of somatic point mutations in impure and heterogeneous cancer samples." *Nature biotechnology*, **31**(3):213–219, 2013.

[CLW18] Joshua D Cohen, Lu Li, Yuxuan Wang, Christopher Thoburn, Bahman Afsari, Ludmila Danilova, Christopher Douville, Ammar A Javed, Fay Wong, Austin Mattox, et al. "Detection and localization of surgically resectable cancers with a multi-analyte blood test." *Science*, **359**(6378):926–930, 2018.

[CPW12] Pablo Cingolani, Adrian Platts, Le Lily Wang, Melissa Coon, Tung Nguyen, Luan Wang, Susan J Land, Xiangyi Lu, and Douglas M Ruden. "A program for annotating and predicting the effects of single nucleotide polymorphisms, SnpEff: SNPs in the genome of Drosophila melanogaster strain w1118; iso-2; iso-3." *Fly*, **6**(2):80–92, 2012.

[CR06] Etienne Couturier and Eduardo PC Rocha. "Replication-associated gene dosage effects shape the genomes of fast-growing bacteria but only for transcription and translation genes." *Molecular microbiology*, **59**(5):1506–1518, 2006.

[CTH19] Cyrielle Caussy, Anupriya Tripathi, Greg Humphrey, Shirin Bassirian, Seema Singh, Claire Faulkner, Ricki Bettencourt, Emily Rizo, Lisa Richards, Zhenjiang Z Xu, et al. "A gut microbiome signature for cirrhosis due to nonalcoholic fatty liver disease." *Nature communications*, **10**(1):1–9, 2019.

[CVD17] Anthony Cutts, Oliver Venn, Alexander Dilthey, Avinash Gupta, Dimitris Vavoulis, Helene Dreau, Mark Middleton, Gil McVean, Jenny C Taylor, and Anna Schuh. "Characterisation of the changing genomic landscape of metastatic melanoma using cell free DNA." *NPJ genomic medicine*, **2**(1):1–8, 2017.

[CWF18] Atish D Choudhury, Lillian Werner, Edoardo Francini, Xiao X Wei, Gavin Ha, Samuel S Freeman, Justin Rhoades, Sarah C Reed, Gregory Gydush, Denisse Rotem, et al. "Tumor fraction in cell-free DNA as a biomarker in prostate cancer." *JCI insight*, **3**(21), 2018.

[DBP11] Mark A DePristo, Eric Banks, Ryan Poplin, Kiran V Garimella, Jared R Maguire, Christopher Hartl, Anthony A Philippakis, Guillermo Del Angel, Manuel A Rivas, Matt Hanna, et al. "A framework for variation discovery and genotyping using next-generation DNA sequencing data." *Nature genetics*, **43**(5):491, 2011.

[DBP17] Amiran Dzutsev, Jonathan H Badger, Ernesto Perez-Chanona, Soumen Roy, Rosalba Salcedo, Carolyne K Smith, and Giorgio Trinchieri. "Microbes and cancer." *Annual review of immunology*, **35**:199–228, 2017.

[DJB19] Fatemeh Dorri, Sean Jewell, Alexandre Bouchard-Côté, and Sohrab P Shah. "Somatic mutation detection and classification through probabilistic integration of clonal population information." *Communications biology*, **2**(1):1–10, 2019.

[DLR13] R Phillip Dellinger, Mitchell M Levy, Andrew Rhodes, Djillali Annane, Herwig Gerlach, Steven M Opal, Jonathan E Sevransky, Charles L Sprung, Ivor S Douglas, Roman Jaeschke, et al. "Surviving Sepsis Campaign: international guidelines for management of severe sepsis and septic shock, 2012." *Intensive care medicine*, **39**(2):165–228, 2013.

[EHH15] Adam D Ewing, Kathleen E Houlahan, Yin Hu, Kyle Ellrott, Cristian Caloian, Takafumi N Yamaguchi, J Christopher Bare, Christine P'ng, Daryl Waggott, Veronica Y Sabelnykova, et al. "Combining tumor genome simulation with crowdsourcing to benchmark somatic single-nucleotide-variant detection." *Nature methods*, **12**(7):623–630, 2015.

[FGP19] Laura Fancello, Sara Gandini, Pier Giuseppe Pelicci, and Luca Mazzarella. "Tumor mutational burden quantification from targeted gene panels: major advancements and challenges." *Journal for immunotherapy of cancer*, **7**(1):183, 2019.

[FKB09] Shiri Freilich, Anat Kreimer, Elhanan Borenstein, Nir Yosef, Roded Sharan, Uri Gophna, and Eytan Ruppin. “Metabolic-network-driven analysis of bacterial ecological strategies.” *Genome biology*, **10**(6):R61, 2009.

[FLB17] Burkhardt Flemer, Denise B Lynch, Jillian MR Brown, Ian B Jeffery, Feargal J Ryan, Marcus J Claesson, Micheal O’Riordain, Fergus Shanahan, and Paul W O’Toole. “Tumour-associated and non-tumour-associated microbiota in colorectal cancer.” *Gut*, **66**(4):633–643, 2017.

[FMP12] Tim Forshew, Muhammed Murtaza, Christine Parkinson, Davina Gale, Dana WY Tsui, Fiona Kaper, Sarah-Jane Dawson, Anna M Piskorz, Mercedes Jimenez-Linan, David Bentley, et al. “Noninvasive identification and monitoring of cancer mutations by targeted deep sequencing of plasma DNA.” *Science translational medicine*, **4**(136):136ra68–136ra68, 2012.

[GDP17] Shicheng Guo, Dinh Diep, Nongluk Plongthongkum, Ho-Lim Fung, Kang Zhang, and Kun Zhang. “Identification of methylation haplotype blocks aids in deconvolution of heterogeneous tissue samples and tumor tissue-of-origin mapping from plasma DNA.” *Nature genetics*, **49**(4):635–642, 2017.

[GGG12] Emil Gegov, Alexander Gegov, Fernand Gobet, Mark Atherton, Daniel Freudenthal, and Julian Pine. “Cognitive modelling of language acquisition with complex networks.” In *Computational intelligence*, pp. 95–106. Nova Science Publishers, 2012.

[GPK18] David R Gandara, Sarah M Paul, Marcin Kowanetz, Erica Schleifman, Wei Zou, Yan Li, Achim Rittmeyer, Louis Fehrenbacher, Geoff Otto, Christine Malboeuf, et al. “Blood-based tumor mutational burden as a predictor of clinical benefit in non-small-cell lung cancer patients treated with atezolizumab.” *Nature medicine*, **24**(9):1441–1448, 2018.

[GRH15] Edward B Garon, Naiyer A Rizvi, Rina Hui, Natasha Leighl, Ani S Balmanoukian, Joseph Paul Eder, Amita Patnaik, Charu Aggarwal, Matthew Gubens, Leora Horn, et al. “Pembrolizumab for the treatment of non-small-cell lung cancer.” *New England Journal of Medicine*, **372**(21):2018–2028, 2015.

[GSG16] Silke Grumaz, Philip Stevens, Christian Grumaz, Sebastian O Decker, Markus A Weigand, Stefan Hofer, Thorsten Brenner, Arndt von Haeseler, and Kai Sohn. “Next-generation sequencing diagnostics of bacteremia in septic patients.” *Genome medicine*, **8**(1):1–13, 2016.

[GSW15] Isaac Garcia-Murillas, Gaia Schiavon, Britta Weigelt, Charlotte Ng, Sarah Hrebien, Rosalind J Cutts, Maggie Cheang, Peter Osin, Ashutosh Nerurkar, Iwanka Kozarewa, et al. “Mutation tracking in circulating tumor DNA predicts relapse in early breast cancer.” *Science translational medicine*, **7**(302):302ra133–302ra133, 2015.

[HBK16] Roy S Herbst, Paul Baas, Dong-Wan Kim, Enriqueta Felip, José L Pérez-Gracia, Ji-Youn Han, Julian Molina, Joo-Hang Kim, Catherine Dubos Arvis, Myung-Ju Ahn, et al. “Pembrolizumab versus docetaxel for previously treated, PD-L1-positive, advanced non-small-cell lung cancer (KEYNOTE-010): a randomised controlled trial.” *The Lancet*, **387**(10027):1540–1550, 2016.

[HCF18] Yu-Feng Huang, Yen-Ju Chen, Tan-Chi Fan, Nai-Chuan Chang, Yi-Jie Chen, Mohit K Midha, Tzu-Han Chen, Hsiao-Hsiang Yang, Yu-Tai Wang, L Yu Alice, et al. “Analysis of microbial sequences in plasma cell-free DNA for early-onset breast cancer patients and healthy females.” *BMC medical genomics*, **11**(1):16, 2018.

[HLS20] Dongsheng Han, Rui Li, Jiping Shi, Ping Tan, Rui Zhang, and Jinming Li. “Liquid biopsy for infectious diseases: a focus on microbial cell-free DNA sequencing.” *Theranostics*, **10**(12):5501, 2020.

[HTZ20] Lena Horvath, Bernard Thienpont, Liyun Zhao, Dominik Wolf, and Andreas Pircher. “Overcoming immunotherapy resistance in non-small cell lung cancer (NSCLC)-novel approaches and future outlook.” *Molecular Cancer*, **19**(1):1–15, 2020.

[HW17] Bastiaan W Haak and W Joost Wiersinga. “The role of the gut microbiota in sepsis.” *The Lancet Gastroenterology & Hepatology*, **2**(2):135–143, 2017.

[Ins16] Broad Institute. “Picard tools.”, 2016.

[JCC15] Peiyong Jiang, Carol WM Chan, KC Allen Chan, Suk Hang Cheng, John Wong, Vincent Wai-Sun Wong, Grace LH Wong, Stephen L Chan, Tony SK Mok, Henry LY Chan, et al. “Lengthening and shortening of plasma DNA in hepatocellular carcinoma patients.” *Proceedings of the National Academy of Sciences*, **112**(11):E1317–E1325, 2015.

[JJ19] Baolei Jia and Che Ok Jeon. “Promotion and induction of liver cancer by gut microbiome-mediated modulation of bile acids.” *PLoS pathogens*, **15**(9):e1007954, 2019.

[JST18] Peiyong Jiang, Kun Sun, Yu K Tong, Suk Hang Cheng, Timothy HT Cheng, Macy MS Heung, John Wong, Vincent WS Wong, Henry LY Chan, KC Allen Chan, et al. “Preferred end coordinates and somatic variants as signatures of circulating tumor DNA associated with hepatocellular carcinoma.” *Proceedings of the National Academy of Sciences*, **115**(46):E10925–E10933, 2018.

[KA11] Felix Krueger and Simon R Andrews. "Bismark: a flexible aligner and methylation caller for Bisulfite-Seq applications." *bioinformatics*, **27**(11):1571–1572, 2011.

[KCK17] Mark Kowarsky, Joan Camunas-Soler, Michael Kertesz, Iwijn De Vlaminck, Winston Koh, Wenying Pan, Lance Martin, Norma F Neff, Jennifer Okamoto, Ronald J Wong, et al. "Numerous uncharacterized and highly divergent microbes which colonize humans are revealed by circulating cell-free DNA." *Proceedings of the National Academy of Sciences*, **114**(36):9623–9628, 2017.

[KDN11] James M Kinross, Ara W Darzi, and Jeremy K Nicholson. "Gut microbiome-host interactions in health and disease." *Genome medicine*, **3**(3):14, 2011.

[KKC03] Esko Kankuri, Tapio Kurki, Petteri Carlson, and Vilho Hiilesmaa. "Incidence, treatment and outcome of peripartum sepsis." *Acta obstetricia et gynecologica Scandinavica*, **82**(8):730–735, 2003.

[KLC17] Shuli Kang, Qingjiao Li, Quan Chen, Yonggang Zhou, Stacy Park, Gina Lee, Brandon Grimes, Kostyantyn Krysan, Min Yu, Wei Wang, et al. "CancerLocator: non-invasive cancer diagnosis and tissue-of-origin prediction using methylation profiles of cell-free DNA." *Genome biology*, **18**(1):1–12, 2017.

[KMK18] Nick Kamps-Hughes, Andrew McUsic, Laurie Kurihara, Timothy T Harkins, Prithwish Pal, Claire Ray, and Cristian Ionescu-Zanetti. "ERASE-Seq: leveraging replicate measurements to enhance ultralow frequency variant detection in NGS data." *PloS one*, **13**(4):e0195272, 2018.

[KR14] Shaji K Kumar and S Vincent Rajkumar. "The current status of minimal residual disease assessment in myeloma." *Leukemia*, **28**(2):239–240, 2014.

[KSB16] Daehwan Kim, Li Song, Florian P Breitwieser, and Steven L Salzberg. "Centrifuge: rapid and sensitive classification of metagenomic sequences." *Genome research*, **26**(12):1721–1729, 2016.

[KSH18] Sangtae Kim, Konrad Scheffler, Aaron L Halpern, Mitchell A Bekritsky, Eunho Noh, Morten Källberg, Xiaoyu Chen, Yeonbin Kim, Doruk Beyter, Peter Krusche, et al. "Strelka2: fast and accurate calling of germline and somatic variants." *Nature methods*, **15**(8):591–594, 2018.

[KZS18] Steven T Kothen-Hill, Asaf Zviran, Rafael C Schulman, Sunil Deochand, Federico Gaiti, Dillon Maloney, Kevin Y Huang, Will Liao, Nicolas Robine, Nathaniel D Omans, et al. "Deep learning mutation prediction enables early stage lung cancer detection in liquid biopsy." 2018.

[LD09] Heng Li and Richard Durbin. "Fast and accurate short read alignment with Burrows–Wheeler transform." *bioinformatics*, **25**(14):1754–1760, 2009.

[LHW09] Heng Li, Bob Handsaker, Alec Wysoker, Tim Fennell, Jue Ruan, Nils Homer, Gabor Marth, Goncalo Abecasis, and Richard Durbin. "The sequence alignment/map format and SAMtools." *Bioinformatics*, **25**(16):2078–2079, 2009.

[LLK18] Wenyuan Li, Qingjiao Li, Shuli Kang, Mary Same, Yonggang Zhou, Carol Sun, Chun-Chi Liu, Lea Matsuoka, Linda Sher, Wing Hung Wong, et al. "CancerDetector: ultrasensitive and non-invasive cancer detection at the resolution of individual reads using cell-free DNA methylation sequencing data." *Nucleic acids research*, **46**(15):e89–e89, 2018.

[LNQ13] Emmanuelle Le Chatelier, Trine Nielsen, Junjie Qin, Edi Prifti, Falk Hildebrand, Gwen Falony, Mathieu Almeida, Manimozhiyan Arumugam, Jean-Michel Batto, Sean Kennedy, et al. "Richness of human gut microbiome correlates with metabolic markers." *Nature*, **500**(7464):541–546, 2013.

[LNZ16] Roni Lehmann-Werman, Daniel Neiman, Hai Zemmour, Joshua Moss, Judith Magenheim, Adi Vaknin-Dembinsky, Sten Rubertsson, Bengt Nellgård, Kaj Blennow, Henrik Zetterberg, et al. "Identification of tissue-specific cell death using methylation patterns of circulating DNA." *Proceedings of the National Academy of Sciences*, **113**(13):E1826–E1834, 2016.

[LNZ20] Shuo Li, Zorawar Noor, Weihua Zeng, Xiaohui Ni, Zuyang Yuan, Frank Alber, Wenyuan Li, Edward B Garon, Xianghong Zhou, Wing Hung Wong, et al. "Abstract LB-247: Sensitive detection of tumor mutations from blood and its application to immunotherapy prognosis.", 2020.

[LS12] Ben Langmead and Steven L Salzberg. "Fast gapped-read alignment with Bowtie 2." *Nature methods*, **9**(4):357, 2012.

[MAS19] Kei Mizuno, Shusuke Akamatsu, Takayuki Sumiyoshi, Jing Hao Wong, Masashi Fujita, Kazuaki Maejima, Kaoru Nakano, Atushi Ono, Hiroshi Aikata, Masaki Ueno, et al. "eVIDENCE: a practical variant filtering for low-frequency variants detection in cell-free DNA." *Scientific reports*, **9**(1):1–11, 2019.

[MB19] Kosuke Mima and Hideo Baba. "The gut microbiome, antitumor immunity, and liver cancer." *Hepatobiliary surgery and nutrition*, **8**(1):67, 2019.

[MBS16] Naia Morueta-Holme, Benjamin Blonder, Brody Sandel, Brian J McGill, Robert K Peet, Jeffrey E Ott, Cyrille Violle, Brian J Enquist, Peter M Jørgensen, and Jens-Christian Svenning. "A network approach for inferring species associations from co-occurrence data." *Ecography*, **39**(12):1139–1150, 2016.

[MCS19] Bradon R McDonald, Tania Contente-Cuomo, Stephen-John Sammut, Ahuva Odenheimer-Bergman, Brenda Ernst, Nieves Perdigones, Suet-Feung Chin, Maria Farooq, Rosa Mejia, Patricia A Cronin, et al. "Personalized circulating tumor

DNA analysis to detect residual disease after neoadjuvant therapy in breast cancer." *Science translational medicine*, **11**(504):eaax7392, 2019.

[MDP15] Muhammed Murtaza, Sarah-Jane Dawson, Katherine Pogrebniak, Oscar M Rueda, Elena Provenzano, John Grant, Suet-Feung Chin, Dana WY Tsui, Francesco Marass, Davina Gale, et al. "Multifocal clonal evolution characterized using circulating tumour DNA in a case of metastatic breast cancer." *Nature communications*, **6**(1):1–6, 2015.

[MFR16] Nicholas McGranahan, Andrew JS Furness, Rachel Rosenthal, Sofie Ramskov, Rikke Lyngaa, Sunil Kumar Saini, Mariam Jamal-Hanjani, Gareth A Wilson, Nicolai J Birkbak, Crispin T Hiley, et al. "Clonal neoantigens elicit T cell immunoreactivity and sensitivity to immune checkpoint blockade." *Science*, **351**(6280):1463–1469, 2016.

[MHB10] Aaron McKenna, Matthew Hanna, Eric Banks, Andrey Sivachenko, Kristian Cibulskis, Andrew Kernytsky, Kiran Garimella, David Altshuler, Stacey Gabriel, Mark Daly, et al. "The Genome Analysis Toolkit: a MapReduce framework for analyzing next-generation DNA sequencing data." *Genome research*, **20**(9):1297–1303, 2010.

[MLG18] David A Mahvi, Rong Liu, Mark W Grinstaff, Yolonda L Colson, and Chandrajit P Raut. "Local cancer recurrence: the realities, challenges, and opportunities for new therapies." *CA: a cancer journal for clinicians*, **68**(6):488–505, 2018.

[MMG18] Diana Miao, Claire A Margolis, Wenhua Gao, Martin H Voss, Wei Li, Dylan J Martini, Craig Norton, Dominick Bossé, Stephanie M Wankowicz, Dana Cullen, et al. "Genomic correlates of response to immune checkpoint therapies in clear cell renal cell carcinoma." *Science*, **359**(6377):801–806, 2018.

[MR15] Florent Mouliere and Nitzan Rosenfeld. "Circulating tumor-derived DNA is shorter than somatic DNA in plasma." *Proceedings of the National Academy of Sciences*, **112**(11):3178–3179, 2015.

[MS11] Tanja Magoč and Steven L Salzberg. "FLASH: fast length adjustment of short reads to improve genome assemblies." *Bioinformatics*, **27**(21):2957–2963, 2011.

[NBT14] Aaron M Newman, Scott V Bratman, Jacqueline To, Jacob F Wynne, Neville CW Eclov, Leslie A Modlin, Chih Long Liu, Joel W Neal, Heather A Wakelee, Robert E Merritt, et al. "An ultrasensitive method for quantitating circulating tumor DNA with broad patient coverage." *Nature medicine*, **20**(5):548–554, 2014.

[NLK16] Aaron M Newman, Alexander F Lovejoy, Daniel M Klass, David M Kurtz, Jacob J Chabon, Florian Scherer, Henning Stehr, Chih Long Liu, Scott V Bratman, Carmen Say, et al. “Integrated digital error suppression for improved detection of circulating tumor DNA.” *Nature biotechnology*, **34**(5):547–555, 2016.

[PAM14] Amy D Proal, Paul J Albert, and Trevor G Marshall. “Inflammatory disease and the human microbiome.” *Discovery medicine*, (17):257–265, 2014.

[PB02] Richard M Peek and Martin J Blaser. “Helicobacter pylori and gastrointestinal tract adenocarcinomas.” *Nature Reviews Cancer*, **2**(1):28–37, 2002.

[PKZ20] Gregory D Poore, Evguenia Kopylova, Qiyun Zhu, Carolina Carpenter, Serena Fraraccio, Stephen Wandro, Tomasz Kosciolek, Stefan Janssen, Jessica Metcalf, Se Jin Song, et al. “Microbiome analyses of blood and tissues suggest cancer diagnostic approach.” *Nature*, **579**(7800):567–574, 2020.

[PRD17] Ryan Poplin, Valentin Ruano-Rubio, Mark A DePristo, Tim J Fennell, Mauricio O Carneiro, Geraldine A Van der Auwera, David E Kling, Laura D Gauthier, Ami Levy-Moonshine, David Roazen, et al. “Scaling accurate genetic variant discovery to tens of thousands of samples.” *BioRxiv*, p. 201178, 2017.

[PVG11] Fabian Pedregosa, Gaël Varoquaux, Alexandre Gramfort, Vincent Michel, Bertrand Thirion, Olivier Grisel, Mathieu Blondel, Peter Prettenhofer, Ron Weiss, Vincent Dubourg, et al. “Scikit-learn: Machine learning in Python.” *the Journal of machine Learning research*, **12**:2825–2830, 2011.

[QYL14] Nan Qin, Fengling Yang, Ang Li, Edi Prifti, Yanfei Chen, Li Shao, Jing Guo, Emmanuelle Le Chatelier, Jian Yao, Lingjiao Wu, et al. “Alterations of the human gut microbiome in liver cirrhosis.” *Nature*, **513**(7516):59–64, 2014.

[RAC19] Dominic G Rothwell, Mahmood Ayub, Natalie Cook, Fiona Thistlethwaite, Louise Carter, Emma Dean, Nigel Smith, Shaun Villa, Joanne Dransfield, Alexandra Clipson, et al. “Utility of ctDNA to support patient selection for early phase clinical trials: the TARGET study.” *Nature medicine*, **25**(5):738–743, 2019.

[RDM12] Andrew Roth, Jiarui Ding, Ryan Morin, Anamaria Crisan, Gavin Ha, Ryan Giuliany, Ali Bashashati, Martin Hirst, Gulisa Turashvili, Arusha Oloumi, et al. “JointSNVMix: a probabilistic model for accurate detection of somatic mutations in normal/tumour paired next-generation sequencing data.” *Bioinformatics*, **28**(7):907–913, 2012.

[RHS15] Naiyer A Rizvi, Matthew D Hellmann, Alexandra Snyder, Pia Kvistborg, Vladimir Makarov, Jonathan J Havel, William Lee, Jianda Yuan, Phillip Wong, Teresa S Ho, et al. “Mutational landscape determines sensitivity to PD-1 blockade in non–small cell lung cancer.” *Science*, **348**(6230):124–128, 2015.

[Roc04] Eduardo PC Rocha. “Codon usage bias from tRNA’s point of view: redundancy, specialization, and efficient decoding for translation optimization.” *Genome research*, **14**(11):2279–2286, 2004.

[SHP11] Heidi Schwarzenbach, Dave SB Hoon, and Klaus Pantel. “Cell-free nucleic acids as biomarkers in cancer patients.” *Nature Reviews Cancer*, **11**(6):426–437, 2011.

[SHW17] Padmanee Sharma, Siwen Hu-Lieskovan, Jennifer A Wargo, and Antoni Ribas. “Primary, adaptive, and acquired resistance to cancer immunotherapy.” *Cell*, **168**(4):707–723, 2017.

[SJC15] Kun Sun, Peiyong Jiang, KC Allen Chan, John Wong, Yvonne KY Cheng, Raymond HS Liang, Wai-kong Chan, Edmond SK Ma, Stephen L Chan, Suk Hang Cheng, et al. “Plasma DNA tissue mapping by genome-wide methylation sequencing for noninvasive prenatal, cancer, and transplantation assessments.” *Proceedings of the National Academy of Sciences*, **112**(40):E5503–E5512, 2015.

[SJS20] Xiaoquan Su, Gongchao Jing, Zheng Sun, Lu Liu, Zhenjiang Xu, Daniel McDonald, Zengbin Wang, Honglei Wang, Antonio Gonzalez, Yufeng Zhang, et al. “Multiple-Disease Detection and Classification across Cohorts via Microbiome Search.” *Msystems*, **5**(2), 2020.

[SLA18] John H Strickler, Jonathan M Loree, Leanne G Ahronian, Aparna R Parikh, Donna Niedzwiecki, Allan Andresson Lima Pereira, Matthew McKinney, W Michael Korn, Chloe E Atreya, Kimberly C Banks, et al. “Genomic landscape of cell-free DNA in patients with colorectal cancer.” *Cancer discovery*, **8**(2):164–173, 2018.

[SLJ12] Nina Sanapareddy, Ryan M Legge, Biljana Jovov, Amber McCoy, Lauren Burcal, Felix Araujo-Perez, Thomas A Randall, Joseph Galanko, Andrew Benson, Robert S Sandler, et al. “Increased rectal microbial richness is associated with the presence of colorectal adenomas in humans.” *The ISME journal*, **6**(10):1858–1868, 2012.

[SMM14] Alexandra Snyder, Vladimir Makarov, Taha Merghoub, Jianda Yuan, Jesse M Zaretsky, Alexis Desrichard, Logan A Walsh, Michael A Postow, Phillip Wong, Teresa S Ho, et al. “Genetic basis for clinical response to CTLA-4 blockade in melanoma.” *New England Journal of Medicine*, **371**(23):2189–2199, 2014.

[THY09] Peter J Turnbaugh, Micah Hamady, Tanya Yatsunenko, Brandi L Cantarel, Alexis Duncan, Ruth E Ley, Mitchell L Sogin, William J Jones, Bruce A Roe, Jason P Affourtit, et al. “A core gut microbiome in obese and lean twins.” *nature*, **457**(7228):480–484, 2009.

[TLH07] Peter J Turnbaugh, Ruth E Ley, Micah Hamady, Claire M Fraser-Liggett, Rob Knight, and Jeffrey I Gordon. "The human microbiome project." *Nature*, **449**(7164):804–810, 2007.

[TWT16] Jeanne Tie, Yuxuan Wang, Cristian Tomasetti, Lu Li, Simeon Springer, Isaac Kinde, Natalie Silliman, Mark Tacey, Hui-Li Wong, Michael Christie, et al. "Circulating tumor DNA analysis detects minimal residual disease and predicts recurrence in patients with stage II colon cancer." *Science translational medicine*, **8**(346):346ra92–346ra92, 2016.

[UGB16] Camilla Urbaniak, Gregory B Gloor, Muriel Brackstone, Leslie Scott, Mark Tangney, and Gregor Reid. "The microbiota of breast tissue and its association with breast cancer." *Applied and environmental microbiology*, **82**(16):5039–5048, 2016.

[UHS16] Peter Ulz, Ellen Heitzer, and Michael R Speicher. "Co-occurrence of MYC amplification and TP53 mutations in human cancer." *Nature genetics*, **48**(2):104–106, 2016.

[VBL15] Jean-Louis Vincent, David Brealey, Nicolas Libert, Nour Elhouda Abidi, Michael ODwyer, Kai Zacharowski, Malgorzata Mikaszewska-Sokolewicz, Jacques Schrenzel, François Simon, Mark Wilks, et al. "Rapid diagnosis of infection in the critically ill, a multicenter study of molecular detection in bloodstream infections, pneumonia, and sterile site infections." *Critical care medicine*, **43**(11):2283, 2015.

[VYF14] Paul A VanderLaan, Norihiro Yamaguchi, Erik Folch, David H Boucher, Michael S Kent, Sidharta P Gangadharan, Adnan Majid, Michael A Goldstein, Mark S Huberman, Olivier N Kocher, et al. "Success and failure rates of tumor genotyping techniques in routine pathological samples with non-small-cell lung cancer." *Lung cancer*, **84**(1):39–44, 2014.

[WBP19] Yochai Wolf, Osnat Bartok, Sushant Patkar, Gitit Bar Eli, Sapir Cohen, Kevin Litchfield, Ronen Levy, Alejandro Jiménez-Sánchez, Sophie Trabish, Joo Sang Lee, et al. "UVB-induced tumor heterogeneity diminishes immune response in melanoma." *Cell*, **179**(1):219–235, 2019.

[WBS14] Stefanie Widder, Katharina Besemer, Gabriel A Singer, Serena Ceola, Enrico Bertuzzo, Christopher Quince, William T Sloan, Andrea Rinaldo, and Tom J Battin. "Fluvial network organization imprints on microbial co-occurrence networks." *Proceedings of the National Academy of Sciences*, **111**(35):12799–12804, 2014.

[WCB13] Marek Wojciechowski, Honorata Czapinska, and Matthias Bochtler. "CpG underrepresentation and the bacterial CpG-specific DNA methyltransferase M. MpeI." *Proceedings of the National Academy of Sciences*, **110**(1):105–110, 2013.

[WDC19] Zhijie Wang, Jianchun Duan, Shangli Cai, Miao Han, Hua Dong, Jun Zhao, Bo Zhu, Shuhang Wang, Minglei Zhuo, Jianguo Sun, et al. "Assessment of blood tumor mutational burden as a potential biomarker for immunotherapy in patients with non–small cell lung cancer with use of a next-generation sequencing cancer gene panel." *JAMA oncology*, **5**(5):696–702, 2019.

[WDL14] Alan W Walker, Sylvia H Duncan, Petra Louis, and Harry J Flint. "Phylogeny, culturing, and metagenomics of the human gut microbiota." *Trends in microbiology*, **22**(5):267–274, 2014.

[WDX20] Leilei Wu, Qinfang Deng, Ze Xu, Songwen Zhou, Chao Li, and Yi-Xue Li. "A novel virtual barcode strategy for accurate panel-wide variant calling in circulating tumor DNA." *BMC bioinformatics*, **21**:1–13, 2020.

[WHG20] Jonathan CM Wan, Katrin Heider, Davina Gale, Suzanne Murphy, Eyal Fisher, Florent Mouliere, Andrea Ruiz-Valdepenas, Angela Santonja, James Morris, Dineika Chandrananda, et al. "ctDNA monitoring using patient-specific sequencing and integration of variant reads." *Science translational medicine*, **12**(548), 2020.

[ZBF18] Oliver A Zill, Kimberly C Banks, Stephen R Fairclough, Stefanie A Mortimer, James V Vowles, Reza Mokhtari, David R Gandara, Philip C Mack, Justin I Odegaard, Rebecca J Nagy, et al. "The landscape of actionable genomic alterations in cell-free circulating tumor DNA from 21,807 advanced cancer patients." *Clinical Cancer Research*, **24**(15):3528–3538, 2018.

[ZRR14] Joseph P Zackular, Mary AM Rogers, Mack T Ruffin, and Patrick D Schloss. "The human gut microbiome as a screening tool for colorectal cancer." *Cancer prevention research*, **7**(11):1112–1121, 2014.

[ZSS20] Asaf Zviran, Rafael C Schulman, Minita Shah, Steven TK Hill, Sunil Deochand, Cole C Khamnei, Dillon Maloney, Kristofer Patel, Will Liao, Adam J Widman, et al. "Genome-wide cell-free DNA mutational integration enables ultra-sensitive cancer monitoring." *Nature Medicine*, pp. 1–11, 2020.

[ZW18] Jinfeng Zou and Edwin Wang. "eTumorRisk, an algorithm predicts cancer risk based on comutated gene networks in an individuals germline genome." *bioRxiv*, p. 393090, 2018.