A novel virtual barcode strategy for accurate panel-wide variants calling in circulating tumor DNA

leilei wu
Shanghai jiaotong University
Qinfang Deng
Tongji University Affiliated Shanghai Pulmonary Hospital
Ze Xu
Smartquerier Biomedicine shanghai
Songwen Zhou
Tongji University Affiliated Shanghai Pulmonary Hospital
Chao Li(New Corresponding Author) (✉ lichao@smartquerier.com)
yixue Li
Shanghai Jiao Tong University School of Life Sciences and Biotechnology

Research article

Keywords: ctDNA, low-AF SNV, virtual barcode, panel-wide calling algorithm, stochastic noise, stereotypical noise

Posted Date: November 4th, 2019

DOI: https://doi.org/10.21203/rs.2.16704/v1

Version of Record: A version of this preprint was published on April 3rd, 2020. See the published version at https://doi.org/10.1186/s12859-020-3412-2.

Abstract

Background Hybrid capture-based next generation sequencing of DNA has been widely applied in the detection of circulating tumor DNA (ctDNA). Various methods have been proposed for ctDNA detection, while still a great challenging present for these low allelic fraction (AF) variants. In addition, no panel-wide calling algorithm is available, which hider the fully usage of ctDNA based ‘liquid biopsy’. Thus, we developed VBCALAVD in silico to overcome these limits.

Results Based on understanding of ctDNA fragmentation nature, a novel platform-independent virtual barcode strategy was established to eliminate random sequencing errors by clustering sequencing reads into virtual family. Polishing stereotypical background artifacts through constructing AF distributions. And additional three robust fine-tuning filters were obtained to eliminate stochastic mutant-family-level noises. The performance of our algorithm was validated using cell-free DNA reference standard samples (cfDNA RSDs) with AFs ranging from 0.1% to 5%. For the RSDs with AFs of 0.1%, 0.2%, 0.5%, 1% and 5%, the mean F1 scores were 0.27 (0~0.56), 0.77, 0.92, 0.926 (0.86~1.0) and 0.89 (0.75~1.0), respectively, which indicates that the proposed approach significantly outperforms the published algorithms. Also, no false positives were detected among 14 normal healthy cfDNA samples. Meanwhile, characteristics of mutant-family-level noise and quantity determinant of its divergence between samples with high and low templates were clearly depicted.

Conclusions Due to its good performance in the detection of low-AF mutations, our algorithm will greatly facilitate the noninvasive panel-wide detection of mutations in circulating tumor DNA (ctDNA) in clinical settings. The whole pipeline is available at https://github.com/zhaodalv/VBCALAVD.

Results

Performance of our virtual barcode

The performance of our virtual barcode was validated from three aspects using three Oncosmart2 UMI samples: 1) recovery rate of the real family from the UMI; 2) family contents; and 3) effectiveness in suppressing errors. Real family was clustered by UMI and genomic position. Virtual family was defined as reads that share the same start, template length and strand. We randomly selected genomic positions on our panel for 10 times (20,000 positions per random sample). The mean virtual family numbers were slightly lower than mean real family numbers (2730 vs. 2943) (Figure 2A, red bar vs. yellow bar), and a strong linear relationship was found between virtual and real family numbers among genomic positions in one random sample (Figure 2B; y = 1.1048x−73.82; R^2 = 1.0, 95% confidence interval (CI): 1.1038~1.1058; $P<10^{-40}$). The recovery rates for real families among the majority of the 20,000 positions ranged from 91.87% to 94.0% (Figure 2C; 92.98%±1.1%) and only a small proportion of reads with different UMI tags were mistakenly clustered by the virtual barcode. The incorrectly clustered family contents were investigated and 92.6% of these members were composed of two UMI tag families, 6.8% were three UMI tag families (Figure 2D). The incorrect clusters might introduce false negatives, particularly if the allele number of variants is extremely low. Thus, we compared f = 1.0 (here f is the ratio of the nonreference allele in a family) virtual family numbers with f = 1.0 real family numbers at six positive sites among three UMI samples. At the 0.1% level, five out of the six positive sites had equal family numbers and no false negatives were detected (Figure 2E), and consistency was found at the 1% and 5% levels (Figure S1). Additionally, AF values of six positive sites calculated from virtual-family-level showed good consistency with AF values from read-level and expected values (Figure S1). In decreasing noise aspect, efficiencies of the virtual barcode and UMI tag were the same supporting by similar mean fraction of panel-wide error-free genomic positions (Figure 2F; UMI tag: 84.44%±0.91%; virtual tag: 88.07% ±0.66%) and mean panel-wide error rates (UMI tag: $7.1*10^{-5}$±$0.3*10^{-5}$; virtual barcode: 5.9*10–5±0.5*10–5).

In conclusion, our virtual barcode was sufficiently robust to replace a real UMI tag and could become a universally applicable approach for reducing noise when sequencing cfDNA samples.

Subsequently, virtual barcode was applied for 30 BGs, the panel-wide error position percentage was significantly decreased in every BG (Figure 3A) and in turn, mean panel-wide error-free position percentage was improved by ~64.11%±12.9%. The ability of the method to decrease random read errors was further confirmed among six positive sites among the top 7 high-sequencing-depth control samples. The presence of random nonreference alleles in two or more samples at the positive site (Figure 3B), and nearly all of these alleles were decreased (Figure 3C). These results confirmed the good and stable performance of our virtual barcode for decreasing read-level stochastic noise.

Characteristics of mutant-family-level noise

A small proportion of error sites supported with f = 1.0 mutant families made the virtual barcode/UMI alone indistinguishable from real variants, and we denote this type of noise mutant-family-level noise (designated as f = 1.0 sites). Thus, additional robust filters are needed to improve the specificity of the proposed algorithm.

The profiles of mutant-family-level noise among 14 controls and 16 RSDs showed an interesting divergence. A significant linear relationship between the mean depth and error position percentage (Figure 3D) was remained at mutant-family-level in the RSDs (Figure 3E, green line; R^2 = 79.04%; P = $5.22*10^{-6}$; 95% CI: 0.059~0.107) but not among the controls (Figure 3E, red dots). This disagreement might be caused by input DNA quantities (virtual family numbers) and uneven depth/coverage. Through normalizing panel-wide virtual family numbers based on the coverage, family degree was obtained for every sample. Compared with controls, the median virtual family degree was significantly higher both in Oncosmart2 (2.49-fold, $2.26*10^{-5}$) and Oncosmart3 RSDs (1.88-fold, P = 0.007; Figure 3F). Based on the observation that the reciprocal of family degree could reflect panel-wide median virtual family size (Figure S2), controls had significantly larger overall virtual family size than RSDs (Figure 3G; P = $5.88*10^{-5}$), which in turn could give higher confident support for calculating f values and further decreasing random read-level noise (Figure 3E; Figure S2). The significantly larger family size in controls were caused by the significantly less template numbers than RSDs (P = 2.05*10^{-5}, Figure S2). Scatterplot clearly showed that high template numbers in RSDs caused significantly higher percentage of mutant-family level noise than controls (P = 6.25*10^{-8}; Figure 3H). This result indicated that using cfDNA data from normal healthy individuals with low templates as the background (20, 36) is not sufficient to cover all noises in samples with high templates under similar sequencing coverage. Thus, we combined controls with RSDs for following analysis.

According to relationship between sample occurrence and AF spectra (Figure S3), mutant-family-level noises were classified into two types: stereotypical (occurrence > = 6 BGs) and stochastic mutant-family-level noise. In total, we obtained 265 unique stereotypical variants (Figure 4A). The RSDs made a greater contribution than the controls in recovering stereotypical variants many of that happened only once in controls (Figure S4). As expected, 265 stereotypical noises were occurred stable as showing a significant linear relationship between 25 Oncosmart2 BGs and 529 Oncosmart2 cfDNA samples (Figure S3; P = 5.6*10^{-32}; 95% CI: 0.922~1.235; R^2 = 41.7). Further analysis of the occurrence rates of 121 shared noises (Figure 4A) showed a significant linear relationship with a higher R^2 value (Figure 4B; P = 2.66*10^{-12}; 95% CI: 1.019~1.308; R^2 = 67.8). Additionally, after polishing based on Oncosmart2, no stereotypical noises were found among five Oncosmart3 BGs at intersection region of the two panels (Table S2–2: False positive sites). Stereotypical noises are caused by many factors, such as DNA damage (37) and PCR errors (38), which had different substitution preference. The main substitution types of our stereotypical variants were that C>T/G>A, C>A/G>T, and A>G/T>C (71.05%, Figure 4C), and these were consistent with the substitution types from Oncosmart3 BGs (Table S2–4: Substitution frequency) and previously reported error profiles for ‘Kapa HF’ polymerase (38). The percentage of these six substitutions further increased to 84.297% in 121 shared sites, which demonstrated that these substitutions introduced by PCR errors were likely to occur universally (Figure 4B, Figure S3; R^2:67.8 vs. 41.7). These PCR-induced distortions are mainly caused by PCR stochasticity and polymerase errors (38, 39) and cannot be removed by UMI strategies only(20, 38).

Strategies for decreasing mutant-family-level noises

Based on clear understanding of characteristics of stereotypical noise, a filtered database needed to be constructed for the polishing of real mutations at these types of sites (265 polishing sites). Different from previously IDES proposed polishing method (20), we first obtained 10 best-fit candidate distributions from 529 Oncosmart2 cfDNA samples based on AIC, BIC, SEE, and R values, which was independently validated in Oncosmart1 samples. Then a comparison between the IDES construction step and our step was made (Figure S4). Finally, 265 stereotypical variants were polished by calculating cutoff AF values from best-fitted personalized distribution. Results showed that the ‘Johnsonsu’ distribution was the best-fitted distribution (Table 1; 26%) and AF cutoffs were shown in Table S3.

Compared with stereotypical noises, stochastic mutant- noises (designated as stochastic f = 1.0 site) were prone to low AF values, wide AF value spectra and unstable occurrence (Figure S3). Additional three fine-tuning filters were proposed based on appropriate specific features, namely, variant position in a segment, imbalanced singleton number and minimum template number requirement.

Specific variant position value (Ds value: 2 < = and > = 149) in f = 1.0 virtual family of stochastic mutant-family-level noise was obtained by comparison with Ds trajectories among f = 1.0 families from high AF sites, positive sites, mutant singletons and f<1.0 virtual families from genomic sites filtered by virtual barcode step (Figure S5). The virtual family that met the identified Ds value was defined as a false family. In every BG, the percentage of sites fully constituted with false family (false family ratio: FFR = 1.0) was calculated and showed orange bar in Figure 4D.

With respect to the variant singleton ratio, based on the observation that variant singleton numbers (ranging from 0 to 39) among stochastic mutant-family-level noises were significantly higher than variant singleton numbers among six positive sites, we hypothesized that for the real SNV site, the ratio of singleton numbers to f = 1.0 family numbers would fluctuate within a certain range. First, in panel level, singleton ratios of all BGs were less than 2.0 (Figure 4E). This singleton ratio was a general robust cutoff value that could well distinguish positive mutations, known mutations of Non-small-cell lung carcinoma (NSCLC) patients (40) and high AF variants from these stochastic family-level noise outliners (Figure S6). Second, in sample level, mean variant singleton ratios of high AF sites could reflect panel-wide singleton ratio, indicating that variant singleton ratio of real variants was fluctuated around panel-wide singleton ratio (Figure 4F). Thus, a sample-level strategy based on distribution of singleton ratios from high AF variants (AF> = 0.05) was applied as reference distribution (Figure S6). After false discovery rate (FDR) correction, a small number (blue bar) of extreme outliners with mean ratios ranging from 4.1~28.2 (orange bar) were removed (FDR< = 0.01; Figure 4G). Besides our method was relatively conservative and no outliner were found in samples with overall high/low singleton ratio (Figure S6), such as two tumor samples (Figure 4G). In conclusion, this filter could avoid over-recovery of variant singletons at genomic sites vulnerable to random noise.

Finally, template numbers were updated both for f = 1.0 numbers and variant singleton number. And updated template features were the most specific features (Figure S7). Based on this specific template feature, ROC curve was constructed for six positive sites at every AF level (Figure 4F), which showed an optimal tradeoff between sensitivity and specificity at a strict 99% confidence level.

Effectiveness of all the filters in improving the panel-wide calling efficacy

We systematically evaluated the effectiveness of each above-described three steps in the proposed approach With respect to reducing noise, the virtual barcode clustering step removed the majority of noise in both 14 Oncsmart2 controls (Figure 5A) and 11 Oncsmart2 RSDs (Figure 5B) The subsequent filters showed greater effectiveness of error reducing in RSDs versus controls (Figure 5B), indicating the necessity of theses filters for error reducing in high-template sample, such as various types of cancer. By combining all the filters, the false-positive sites were maintained at extremely low numbers (Figure 5C). We then calculated the sensitivity, PPV and F1 score of both our algorithm and an available calling algorithm at every level using 25 Oncosmart2 BGs (Figure S8). The results showed that the performance of our algorithm was significantly better than that of previously published calling software at every AF level from 0.1% to 5% (Figures 5D~5H). Additional validation of our algorithm using five Oncosmart3 RSDs proved the robustness of our algorithm at AF levels ranging from 0.1% to 5% (Figure S9; Table S2–1: Sensitivity).

A small number of false-positive sites were retained in the 25 Oncosmart2 BGs. From previous reference, we incorporated low-complexity (LC)(41) and short tandem region (STR)(42) into pipeline. Left in controls were SNP sites after annotation (Table S4). And left false-positive sites supported the “spreading-of-signal”(43) with the newer sequencing platform (HiSeq 3000/4000/X Ten) in the same sequencing lane (Table S5).

Discussion

Recently, several studies have been focusing on application of cfDNA fragmentation information in clinical settings (44–46). Traditional endogenous UMI of randomly sheared genomic DNA sequences in depressing noises has been technically validated. Due to the shearing process not entirely random, its usage is limited (25). However, in highly fragmentation nature cfDNA, this type of endogenous UMI as virtual barcode here was accommodated well supporting both by comprehensively validation here and application in our previous research(40). The downside of this step was that approximately 8% of the UMI was wrongly clustered by the virtual barcode, which was due to the fact that different ctDNA molecules have a certain probability of sharing the same virtual barcode (19).

This downside of our proposed method leads to a lower yield of usable families that might generate lower f = 1.0 supported family numbers for a candidate mutation, as shown by the lower f = 1.0 virtual family numbers compared with f = 1.0 real family numbers in Figure 2E, Figure S1. This downside did not have an effect on the sensitivity and PPV at any of the AF levels tested in this study, and thus, we did not further optimize this step of the algorithm. However, because this downside might have some effect in some cases, the value of the f parameter can be adjusted to minimize this effect. Besides, this step also can be affected by paralogous sequences. Reads in these regions tend to have lower mapping quality that is due to multiple alignments. Additionally, multiple mismatches (MM)(47) was another feature to avoid this effect.

For polishing step, unlike IDES, through large samples, we find the most fitted distribution of stereotypical noise under high depth. Meanwhile, best-fitted distributions also provided informative prior distribution for distribution construction with low sample size using Bayesian methods.

For variant singleton ratio filter, the hypothesis of this filter relies on panel-wide singleton ratio and sequencing depth (family degree). For sample with panel-wide single ratio lager than 2, it might be not necessary for this calculating process. For example, one exome data, most of its templates were singletons (Figure S6) that were the main virtual family form to support variants. Under this circumstance, overall variant singleton ratios were high among variants. Except for panel-wide singleton ratio, sequence depth is another factor. For tumor–70KB with extreme low sequence depth among all samples (Figure S2), its low family degree under low sequence depth lead a small proportion of singletons that caused overall low variant singleton ratios (Figure 4E: dark green dot; Figure S6). Although our method can intelligently recognize these samples, we though, there should be a sample level cutoff value to assess whether this sample needs calculating process of this filter and related precise sample level cutoff value need further detailed investigation among large series of family degree samples with different sequencing depth.

Conclusions

This study developed a novel calling algorithm for the accurate detection of somatic mutations with an AF as low as 0.1%. The algorithm introduces three noise-reduction strategies based on a comprehensive analysis of the source of different types of sequencing noise. The robustness of the strategies was well elaborated using 11 Oncosmart2 RSDs and 14 Oncosmart2 controls and validated with five Oncosmart3 RSDs. Our algorithm is independent of the platform and well suited for NGS data with or without an UMI. Due to its good performance for the detection of low-AF mutations, our algorithm will greatly facilitate the noninvasive panel-wide detection of mutations in cfDNA in clinical settings.

Material And Methods

Materials

In the present study, the following materials were included: 14 cfDNA samples (controls) from healthy individuals, 529 Oncosmart2 patient cfDNA, 104 Oncosmart1 patient cfDNA data and 2 tumor samples (one was from 70KB panel, another was whole exome data), 16 cfDNA reference standards (RSDs) (HD780) harboring six SNV positive sites with AF levels from 0.1% to 5% and corresponding 3 wild-type cfDNA control (HWT). Detailed sample descriptions and sample usages were provided in Supplementary Materials. After preprocessing, sample statistics were provided in Table S1.

Virtual barcode-based algorithm

The sequencing reads were clustered into virtual families according to the start, template length and strand. We validated the robustness and effectiveness of the virtual barcode compared with three UMI samples from three aspects: 1) recovery rate of the real family from the UMI; 2) family contents; and 3) effectiveness in suppressing errors. After validation, if both read1 (R1) and read2 (R2) from the sample template covered a genomic site, we further consolidated the read1 and read2 families. For a particular genomic site, if the bases from R1 and R2 were the same, only one read was retained in the corresponding virtual family; otherwise, both reads were discarded. The virtual barcode was then defined based on the start and template length. For a singleton, only the variant singleton was retained if the position had at least one virtual family with f = 1.0.

Construction of the polishing distribution

To establish a well-fitted distribution for stereotypical mutant-family-level noises (designated as stereotypical f = 1.0 site), we adopted a novel strategy consisting of two steps: 1) identifying candidate distributions from 529 Oncosmart2 cfDNA samples and validating the candidates in 104 Oncosmart1 cfDNA samples independently; and 2) constructing the best-fit distribution for a specific polishing site.

Additional finetuning filters

Based on comprehensive knowledge of the sources of stochastic mutant-family-level noises, three finetuning filters were introduced: 1) variant position in a segment, 2) imbalanced singleton number, 3) minimum template number requirement.

Detailed methods and illustrations of every part are provided in the Online Supplementary methods.

Abbreviations

allelic fraction

AF

single nucleotide variant

SNV

Cell-free DNA

cfDNA

Reference standard samples

RSDs

normal healthy cfDNA samples

controls

circulating tumor DNA

ctDNA

tumor mutation burden

TMB

Non-small-cell lung carcinoma

NSCLC

positive predictive values

PPV

background samples

BGs

95% confidence interval

95% CI

unique molecular identifier

UMI

wild-type cfDNA control

HWT

false family ratio

FFR

false discovery rate

FDR

Declarations

Ethics approval and consent to participate
Heathy person and NSCLC patients provided written informed consent before enrolment and all cfDNA samples were obtained under approval of Medical ethics committee of Shanghai Pulmonary Hospital.

Consent for publication

This manuscript contains no individual person’s data in any form.

Availability of data and materials

All data generated or analyzed during this study are included in this published article and its supplementary information files.

Competing interests
The authors declare that they have no competing interests.

Funding
This work was supported by the National Key R&D Program of China (No. 2017YFA0505500). The funding bodies played no role in the design of the study and collection, analysis, and interpretation of data and in writing the manuscript.

Author contributions YXL, CL, and SWZ conceived and supervised the research. LLW performed all bioinformatics analyses and wrote the code of whole pipleline. QFD prepared, collected the clinical patient cfDNA samples and provided clinical support. ZX constructed the NGS library, performed the sequencing analysis and provided technical support. LLW wrote the main text of the manuscript with contributions from all the authors. All the authors have read and approved the final manuscript.

Acknowledgement We appreciate Smartquerier Co., Ltd for provided designed panel.

References

1.Bamford S, Dawson E, Forbes S, Clements J, Pettett R, Dogan A, et al. The COSMIC (Catalogue of Somatic Mutations in Cancer) database and website. British Journal Of Cancer. 2004 06/08/online;91:355.

2.Poduri A, Evrony GD, Cai X, Walsh CA. Somatic mutation, genomic variation, and neurological disease. Science (New York, NY). 2013;341(6141):1237758-. PubMed PMID: 23828942.

3.Abbosh C, Birkbak NJ, Wilson GA, Jamal-Hanjani M, Constantin T, Salari R, et al. Phylogenetic ctDNA analysis depicts early-stage lung cancer evolution. Nature. 2017 04/26/online;545:446.

4.de Bruin EC, McGranahan N, Mitter R, Salm M, Wedge DC, Yates L, et al. Spatial and temporal diversity in genomic instability processes defines lung cancer evolution. Science. 2014 Oct 10;346(6206):251–6. PubMed PMID: 25301630. Pubmed Central PMCID: PMC4636050.

5.Chabon JJ, Simmons AD, Lovejoy AF, Esfahani MS, Newman AM, Haringsma HJ, et al. Circulating tumour DNA profiling reveals heterogeneity of EGFR inhibitor resistance mechanisms in lung cancer patients. Nature communications. 2016;7:11815.

6.Sequist LV, Martins RG, Spigel D, Grunberg SM, Spira A, Jänne PA, et al. First-line gefitinib in patients with advanced non-small-cell lung cancer harboring somatic EGFR mutations. Journal of clinical oncology. 2008;26(15):2442–9.

7.Thress KS, Paweletz CP, Felip E, Cho BC, Stetson D, Dougherty B, et al. Acquired EGFR C797S mutation mediates resistance to AZD9291 in non–small cell lung cancer harboring EGFR T790M. Nature medicine.

2015;21(6):560.

8.Cibulskis K, Lawrence MS, Carter SL, Sivachenko A, Jaffe D, Sougnez C, et al. Sensitive detection of somatic point mutations in impure and heterogeneous cancer samples. Nat Biotechnol. 2013 Mar;31(3):213–9. PubMed PMID: 23396013. Pubmed Central PMCID: PMC3833702.

9.Koboldt DC, Zhang Q, Larson DE, Shen D, McLellan MD, Lin L, et al. VarScan 2: somatic mutation and copy number alteration discovery in cancer by exome sequencing. Genome Res. 2012 Mar;22(3):568–76. PubMed PMID: 22300766. Pubmed Central PMCID: PMC3290792.

10.Kockan C, Hach F, Sarrafi I, Bell RH, McConeghy B, Beja K, et al. SiNVICT: ultra-sensitive detection of single nucleotide variants and indels in circulating tumour DNA. Bioinformatics. 2017 Jan 1;33(1):26–34. PubMed PMID: 27531099.

11.Lai Z, Markovets A, Ahdesmaki M, Chapman B, Hofmann O, McEwen R, et al. VarDict: a novel and versatile variant caller for next-generation sequencing in cancer research. Nucleic Acids Res. 2016 Jun 20;44(11):e108. PubMed PMID: 27060149. Pubmed Central PMCID: PMC4914105.

12.McKenna A, Hanna M, Banks E, Sivachenko A, Cibulskis K, Kernytsky A, et al. The Genome Analysis Toolkit: a MapReduce framework for analyzing next-generation DNA sequencing data. Genome Res. 2010 Sep;20(9):1297–303. PubMed PMID: 20644199. Pubmed Central PMCID: PMC2928508.

13.Xu H, DiCarlo J, Satya RV, Peng Q, Wang Y. Comparison of somatic mutation calling methods in amplicon and whole exome sequence data. BMC genomics. 2014;15(1):244.

14.Sandmann S, De Graaf AO, Karimi M, Van Der Reijden BA, Hellström-Lindberg E, Jansen JH, et al. Evaluating variant calling tools for non-matched next-generation sequencing data. Scientific reports. 2017;7:43169.

15.Spencer DH, Tyagi M, Vallania F, Bredemeyer AJ, Pfeifer JD, Mitra RD, et al. Performance of common analysis methods for detecting low-frequency single nucleotide variants in targeted next-generation sequence data. The Journal of Molecular Diagnostics. 2014;16(1):75–88.

16.Alioto TS, Buchhalter I, Derdak S, Hutter B, Eldridge MD, Hovig E, et al. A comprehensive assessment of somatic mutation detection in cancer using whole-genome sequencing. Nature communications. 2015;6:10001.

17.Krøigård AB, Thomassen M, Lænkholm A-V, Kruse TA, Larsen MJ. Evaluation of nine somatic variant callers for detection of somatic mutations in exome and targeted deep sequencing data. PLoS One. 2016;11(3):e0151664.

18.Remon J, Soria JC, Planchard D, Jovelet C, Pannet C, Lacroix L, et al. Liquid biopsies for molecular profiling of mutations in non-small cell lung cancer patients lacking tissue samples. AACR; 2016.

19.Newman AM, Bratman SV, To J, Wynne JF, Eclov NC, Modlin LA, et al. An ultrasensitive method for quantitating circulating tumor DNA with broad patient coverage. Nature medicine. 2014;20(5):548.

20.Newman AM, Lovejoy AF, Klass DM, Kurtz DM, Chabon JJ, Scherer F, et al. Integrated digital error suppression for improved detection of circulating tumor DNA. Nature biotechnology. 2016;34(5):547.

21.Zhang J, Fujimoto J, Zhang J, Wedge DC, Song X, Zhang J, et al. Intratumor heterogeneity in localized lung adenocarcinomas delineated by multiregion sequencing. Science. 2014 Oct 10;346(6206):256–9. PubMed PMID: 25301631. Pubmed Central PMCID: PMC4354858.

22.Schmitt MW, Fox EJ, Prindle MJ, Reid-Bayliss KS, True LD, Radich JP, et al. Sequencing small genomic targets with high efficiency and extreme accuracy. Nature methods. 2015;12(5):423.

23.Schmitt MW, Kennedy SR, Salk JJ, Fox EJ, Hiatt JB, Loeb LA. Detection of ultra-rare mutations by next-generation sequencing. Proceedings of the National Academy of Sciences. 2012;109(36):14508–13.

24.Shugay M, Zaretsky AR, Shagin DA, Shagina IA, Volchenkov IA, Shelenkov AA, et al. MAGERI: Computational pipeline for molecular-barcoded targeted resequencing. PLoS computational biology. 2017;13(5):e1005480.

25.Kinde I, Wu J, Papadopoulos N, Kinzler KW, Vogelstein B. Detection and quantification of rare mutations with massively parallel sequencing. Proc Natl Acad Sci U S A. 2011 2011/06//;108(23):9530–5. PubMed PMID: 21586637. eng.

26.Fang LT, Afshar PT, Chhibber A, Mohiyuddin M, Fan Y, Mu JC, et al. An ensemble approach to accurately detect somatic mutations using SomaticSeq. Genome biology. 2015;16(1):197.

27.Peng Q, Satya RV, Lewis M, Randad P, Wang Y. Reducing amplification artifacts in high multiplex amplicon sequencing by using molecular barcodes. BMC genomics. 2015;16(1):589.

28.Lanman RB, Mortimer SA, Zill OA, Sebisanovic D, Lopez R, Blau S, et al. Analytical and clinical validation of a digital sequencing panel for quantitative, highly accurate evaluation of cell-free circulating tumor DNA. PloS one. 2015;10(10):e0140712.

29.Schwarzenbach H, Müller V, Milde-Langosch K, Steinbach B, Pantel K. Evaluation of cell-free tumour DNA and RNA in patients with breast cancer and benign breast disease. Molecular BioSystems. 2011;7(10):2848–54.

30.Park J-L, Kim HJ, Choi BY, Lee H-C, Jang H-R, Song KS, et al. Quantitative analysis of cell-free DNA in the plasma of gastric cancer patients. Oncol Lett. 2012 2012/04//;3(4):921–6. PubMed PMID: 22741019. eng.

31.Szpechcinski A, Chorostowska-Wynimko J, Struniawski R, Kupis W, Rudzinski P, Langfort R, et al. Cell-free DNA levels in plasma of patients with non-small-cell lung cancer and inflammatory lung disease.

British journal of cancer. 2015;113(3):476–83. PubMed PMID: 26125447. Epub 06/30.

32.Schwarzenbach H, Stoehlmacher J, Pantel K, Goekkurt E. Detection and Monitoring of Cell-Free DNA in Blood of Patients with Colorectal Cancer. Annals of the New York Academy of Sciences. 2008;1137(1):190–6.

33.Gandara DR, Paul SM, Kowanetz M, Schleifman E, Zou W, Li Y, 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. 2018 2018/09/01;24(9):1441–8.

34.Chalmers ZR, Connelly CF, Fabrizio D, Gay L, Ali SM, Ennis R, et al. Analysis of 100,000 human cancer genomes reveals the landscape of tumor mutational burden. Genome medicine. 2017;9(1):34-. PubMed PMID: 28420421.

35.Benayed R, Offin M, Mullaney K, Sukhadia P, Rios K, Desmeules P, et al. High Yield of RNA Sequencing for Targetable Kinase Fusions in Lung Adenocarcinomas with No Mitogenic Driver Alteration Detected by DNA Sequencing and Low Tumor Mutation Burden. Clinical Cancer Research. 2019;25(15):4712–22.

36.Deng S, Lira M, Huang D, Wang K, Valdez C, Kinong J, et al. TNER: a novel background error suppression method for mutation detection in circulating tumor DNA. BMC Bioinformatics. 2018 2018/10/20;19(1):387.

37.Costello M, Pugh TJ, Fennell TJ, Stewart C, Lichtenstein L, Meldrim JC, et al. Discovery and characterization of artifactual mutations in deep coverage targeted capture sequencing data due to oxidative DNA damage during sample preparation. Nucleic acids research. 2013;41(6):e67-e.

38.Shagin DA, Shagina IA, Zaretsky AR, Barsova EV, Kelmanson IV, Lukyanov S, et al. A high-throughput assay for quantitative measurement of PCR errors. Scientific Reports. 2017 2017/06/02;7(1):2718.

39.Kebschull JM, Zador AM. Sources of PCR-induced distortions in high-throughput sequencing data sets. Nucleic acids research. 2015;43(21):e143-e.

40.Deng Q, Xie B, Wu L, Ji X, Li C, Feng L, et al. Competitive evolution of NSCLC tumor clones and the drug resistance mechanism of first-generation EGFR-TKIs in Chinese NSCLC patients. Heliyon. 2018;4(12):e01031.

41.Li H. Toward better understanding of artifacts in variant calling from high-coverage samples. Bioinformatics. 2014;30(20):2843–51.

42.Gymrek M, Golan D, Rosset S, Erlich Y. lobSTR: A short tandem repeat profiler for personal genomes. Genome research. 2012 2012/06//;22(6):1154–62. PubMed PMID: 22522390. eng.

43.Sinha R, Stanley G, Gulati GS, Ezran C, Travaglini KJ, Wei E, et al. Index switching causes “spreading-of-signal” among multiplexed samples in Illumina HiSeq 4000 DNA sequencing. BioRxiv. 2017:125724.

44.Mouliere F, Chandrananda D, Piskorz AM, Moore EK, Morris J, Ahlborn LB, et al. Enhanced detection of circulating tumor DNA by fragment size analysis. Science Translational Medicine. 2018;10(466):eaat4921.

45.Cristiano S, Leal A, Phallen J, Fiksel J, Adleff V, Bruhm DC, et al. Genome-wide cell-free DNA fragmentation in patients with cancer. Nature. 2019 2019/06/01;570(7761):385–9.

46.Mouliere F, Mair R, Chandrananda D, Marass F, Smith CG, Su J, et al. Detection of cell-free DNA fragmentation and copy number alterations in cerebrospinal fluid from glioma patients. EMBO Molecular Medicine. 2018:e9323.

47.Barnell EK, Ronning P, Campbell KM, Krysiak K, Ainscough BJ, Sheta LM, et al. Standard operating procedure for somatic variant refinement of sequencing data with paired tumor and normal samples. Genetics in Medicine. 2019 2019/04/01;21(4):972–81.

Table 1

Table 1: Information on the best distribution among 265 polishing sites.

Distributions	Best numbers	Percentage (%)	Mean sample size	Sample size range
Dweibull	11	4.15	21.181818	8~95
Lognorm	18	6.79	108.052632	19~354
Alpha	19	7.17	104.157895	8~475
Exponnorm	24	9.06	122.791667	21~545
Weibull_min	25	9.43	89.16	7~550
Nct	27	10.19	173.962963	9~514
Gamma	33	12.45	139.757576	8~525
Beta	39	14.72	139.794872	6~479
Johnsonsu	69	26.04	109.115942	8~529

Figures

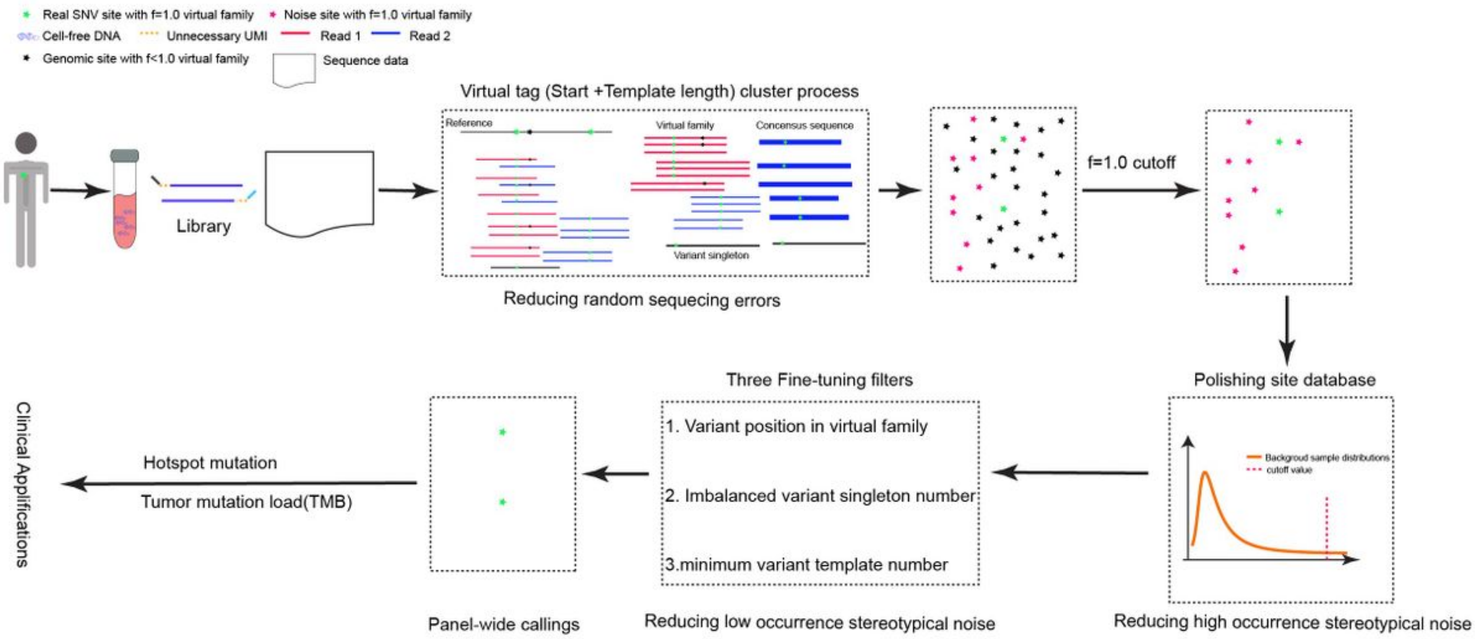
Real SNV site with f=1.0 virtual family
Noise site with f=1.0 virtual family
Cell-free DNA
Unnecessary UMI
Read 1
Read 2
Genomic site with f<1.0 virtual family
Sequence data
Library
Virtual tag (Start +Template length) cluster process
Reference
Virtual family
Concensus sequence
Variant singleton
Reducing random sequecing errors
f=1.0 cutoff
Polishing site database
Backgroud sample distributions
cutoff value
Reducing high occurrence stereotypical noise
Three Fine-tuning filters
1. Variant position in virtual family
2. Imbalanced variant singleton number
3.minimum variant template number
Reducing low occurrence stereotypical noise
Panel-wide callings
Hotspot mutation
Tumor mutation load(TMB)
Clinical Applifications

Figure 1

Flowchart of the novel virtual barcode-based calling algorithm.

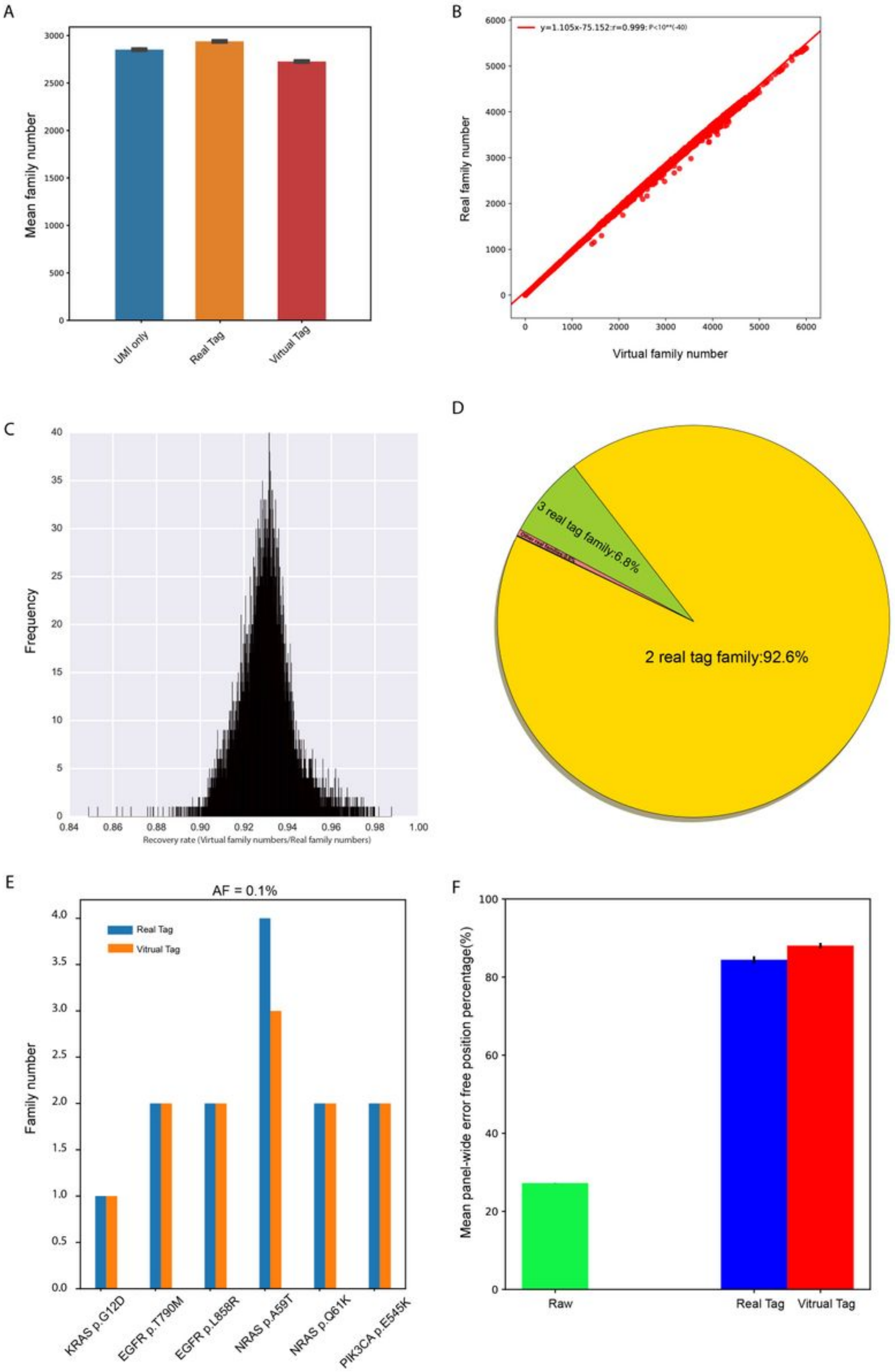

Figure 2

Comparison between the virtual barcode and a real barcode using 3 Oncosmart2 UMI samples. (A) Mean family numbers and corresponding SD from 10 random samples (20,000 genomic positions per sample) obtained using the UMI alone (blue bar), a real barcode (orange bar) and the virtual barcode (red bar). (B) Significant linear relationship between virtual family numbers and real family numbers for 20,000 genomic positions (R2=1.0). (C) Recovery rate distribution for real family numbers at 20,000 genomic positions. (D) Percentage of real barcodes among incorrectly assigned virtual barcode families. (E) Comparison between f=1.0 virtual family numbers (orange bar) and f=1.0 real family numbers (blue bar) among six positive sites in one UMI sample with an AF level of 0.1%. (F) Mean fraction and corresponding SD of the panel-wide error free position before (green bar) and after application of a real barcode (blue bar) and the virtual barcode (red bar).

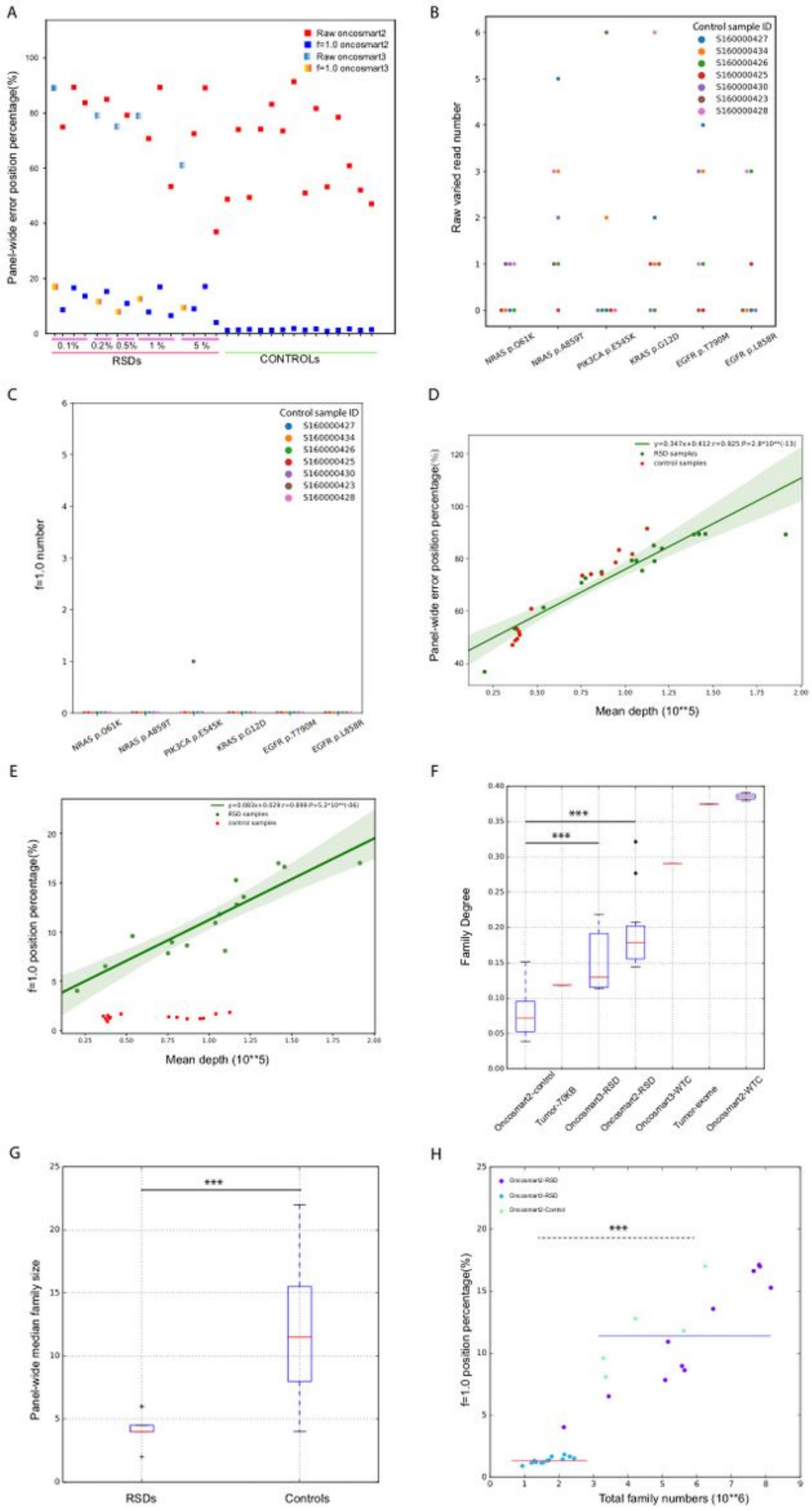

Figure 3

Noise profile among the 30 background samples (BGs) before and after application of the virtual barcode. (A) Panel-wide error position percentage in every BG before and after application of the virtual barcode (Oncosmart2 BGs: blue square to red square; Oncosmart3 BGs: gradient blue to gradient yellow). (B) Numbers of nonreference alleles among the top 7 high-sequence-depth controls at six positive sites. (C) Numbers of the f=1.0 virtual family at six positive sites among the top 7 high-sequence-depth controls. (D) Significant linear relationship between the panel-wide mean depth and the panel-wide error position percentage among 30 BGs (green dot: 16 RSDs; red dot: 14 controls; R2=0.856). (E) Relationship between the fraction of the error position with f=1.0 virtual family and the panel-wide mean depth among 30 BGs after application of the virtual barcode (RSDs, green dots; controls, red dots). A significant linear relationship was observed in the RSDs (R2=0.794; P=5.6*10-6). (F) Boxplot of family degree for 11 Oncosmart2 RSDs, 14 Oncosmart2 controls, HWT samples and tumor samples. Compared with controls, significant high family degree both in Oncosmart2 and Oncosmart3 RSDs; *** means P<0.001. (G) Boxplot of panel-wide median family size between controls and RSDs; *** means P<0.001. (H) Significantly higher error percentage after virtual family in high template RSDs (blue line) than in low template controls (red line).

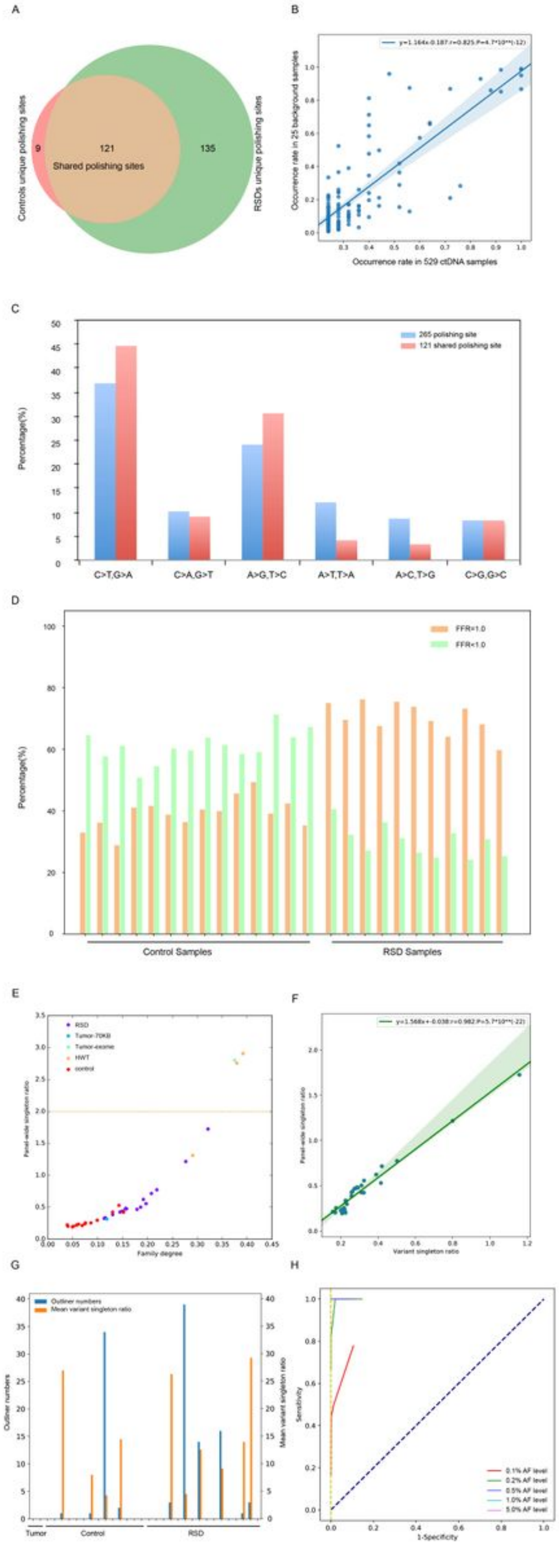

Figure 4

Stereotypical noise characteristics and effectiveness of finetuning filters. (A) Stereotypical site numbers from 14 Oncosmart2 controls and 11 Oncosmart2 RSDs: 121 shared sites among the controls and RSDs (brown region), nine sites from only the controls (red region), and 135 sites from only the RSDs (green region). (B) Significant linear relationship between the incidence rate in 25 BGs and the incidence rate in 529 Oncosmart2 cfDNA samples among 121 shared polishing sites (R2=0.6658). (C) Percentages of 12 substitution types among 265 polishing sites (blue bar) and 121 shared sites (red bar). (D) Fraction of positions that completely consisted of false families (orange bar) among genomic positions with at least one f=1.0 supported virtual families in every Oncosmart2 BG sample. (E) Direct relation between family degree and panel-wide singleton ratio among all samples (dash line represents 2.0). (F) Significant linear relationship (R=0.98; P=5.7*10-22) between panel-wide singleton ratio and mean variant singleton ratio from high AF sites (AF>=0.05) among 30 BGs. (G) Effectiveness of sample-level strategy to remove variant singleton ratio outliners at FDR<0.01 level for all samples; blue bar represents filtered numbers and orange bar represents corresponding mean variant singleton ratio. (H) ROC curve based on the optimal template feature (updated f=1.0 virtual family numbers plus updated variant singletons) at every AF level under a theoretical confidence level ranging from 80% to 99.5%.

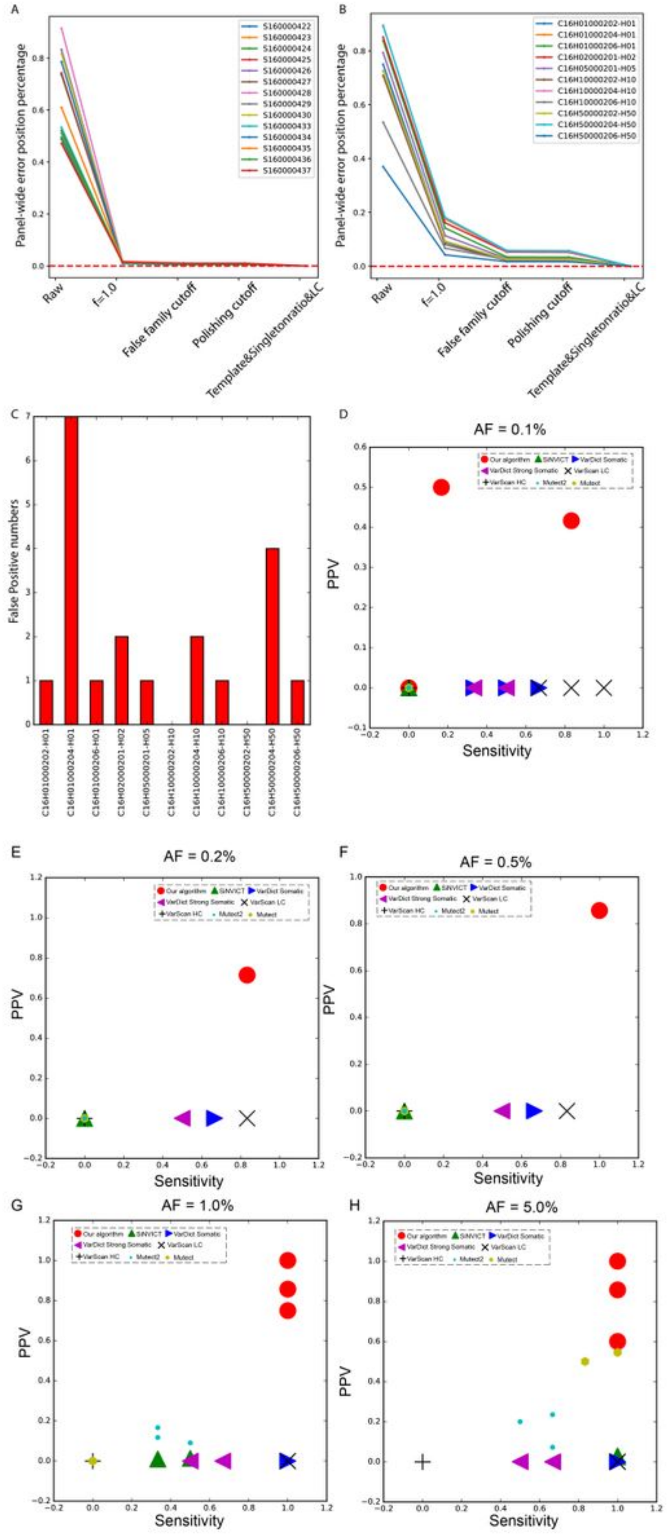

Figure 5

Systematic evaluation of the effectiveness of all filters on reducing panel-wide noise in 14 Oncosmart2 RSDs at AF levels from 0.1% to 5%. (A, B) Fraction of panel-wide error-free positions in the Oncosmart2 controls and RSDs obtained with each filter. (C) Numbers of false-positive sites retained among Oncosmart2 RSDs. (D~H) Panel-wide sensitivity and PPVs obtained with our algorithm (red circles) and five published calling algorithms using Oncosmart2 RSDs with AF values ranging from 0.1% to 5%.

Supplementary Files

This is a list of supplementary files associated with this preprint. Click to download.

- FigS7.jpg
- figS6.jpg
- FigS4.jpg
- FigureS201.jpg
- Supplemetary.docx
- TableS1.xlsx
- TableS5.xlsx
- TableS2.xlsx
- TableS3.xlsx
- TableS4.xlsx
- FigS1.jpg
- FigS9.jpg
- FigS5.jpg
- FigS8.jpg
- FigS3.jpg