Abstract 3372

Cell-free DNA (cfDNA) Fragment Length Patterns of Tumor- and Blood-derived Variants in Participants With and Without Cancer

2019 AACR Meeting
March 29–April 3, 2019
Atlanta, GA, USA

Earl Hubbell, PhD; Tara Maddala, PhD; Oliver Venn, DPhil; Eric Scott, PhD; Susan Tang, MS; Archana Shenoy, PhD; Alex Aravanis, MD, PhD

GRAIL, Inc., Menlo Park, CA

BACKGROUND

- Previous studies on transplanted tissue or single cancers have indicated that the fragment lengths of plasma-derived cfDNA reflect their respective source.
  - Specifically, non-hematopoietically-derived cfDNA molecules are shorter than those that are hematopoietically-derived,[1] and circulating tumor DNA (ctDNA) is shorter than normal cfDNA.[2,3]
- This has fueled research on the detection of tumor-derived mutations in cfDNA, commonly via whole-genome sequencing or PCR-based methods.[4,5]
  - Results, however, are often clouded by interfering (non-tumor-specific) somatic and clonal-hematopoiesis (CH)-derived mutations.[6,7]
- Given that CH increases with age,[8-10] and given the prevalence of cancer in the general population,[11] most individuals in a cancer screening population will have no tumor-derived alleles and mostly alleles from CH.
- To improve detection of non-metastatic tumors, there is a need for increased understanding about the nature of cfDNA variants derived from different sources.
- This analysis leverages data from the Circulating Cell-free Genome Atlas study (NCT02889978)—a prospective, multi-center, longitudinal observational study designed to develop a single blood test for multiple types of cancer across stages—to examine cfDNA variant fragment lengths across >10 tumor types and describe the nature of the associated cfDNA variants.

METHODS

Sample Processing

Plasma samples (N=1406) were evaluated from participants with cancer (n=845) and without cancer (n=561); the breakdown of cancer types is depicted in Table 1.

- cfDNA and genomic DNA from white blood cells (WBCs) were subjected to a high-intensity targeted sequencing panel (507 genes, 60000X) with error-correction; 533 samples also had matched tumor biopsy tissue that were subjected to whole-genome sequencing (30X).

Table 1. Sample Breakdown

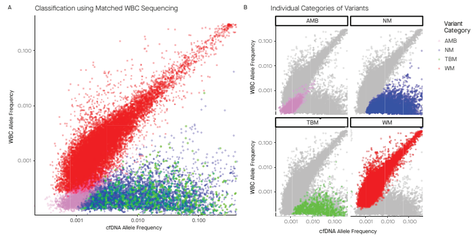
Group	N
Non-cancer	561
Lung	118
Breast	339
Prostate	69
Colorectal	45
Uterine	27
Pancreas	26
Renal	26
Esophageal	24
Lymphoma	22
Head/Neck	19
Ovarian	17
Remaining*	113

*Cancers with ≤15 samples each.

Variant Classification

- Somatic single-nucleotide variants (SNVs) that passed noise filters were identified and classified using the sequencing results into one of four categories:
  - Tumor biopsy-matched (TBM; present in cfDNA and biopsy)
  - WBC-matched (WM; present in cfDNA and WBC)
  - Non-matched (NM; low probability [P<0.01] of being WBC-derived)
  - Ambiguous (AMB; unidentifiable source).
- Classification of cancer versus non-cancer status was accomplished using a joint model between observed alternate cfDNA and WBC allele counts given depth (Figure 1); treating both cfDNA and WBC frequencies as joint observations from a pair of unknown true frequencies, we estimated the likelihood that the cfDNA was derived from a different source.
- The joint calling procedure combined a uniform prior on frequency with the observed counts for reference and alternate alleles to compute a posterior mean for the unknown true frequency conditional on the observed values. This posterior mean is always positive, and is used for plotting in the rest of this presentation.
- Biopsy-matched (TBM) variants were matched to variants detected in tissue samples by simple presence or absence at a location in the genome.
- "Ambiguous" (AMB) was assigned if the cfDNA frequency could not be determined to be above the WBS frequency with >99% probability, and no alternate alleles were found in the WBC; in this case, there was neither positive evidence for a WBC source, nor could the variant be excluded with sufficient confidence to be accurate.

Figure 1. Classification using Matched WBC Sequencing

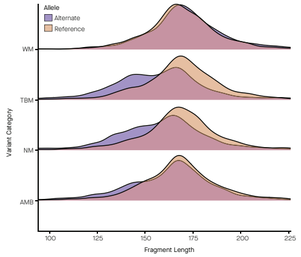


Plasma cfDNA allele frequencies (posterior mean) as determined by targeted panel sequencing are shown for each variant source (posterior mean is always positive allowing for log-scale plotting). Source is depicted by color (red: WBC-matched [WM]; green: tumor biopsy-matched [TBM]; pink: ambiguous [AMB]; blue: non-matched [NM]). To more clearly show each category, in (B) every SNV is plotted in gray as a background, and each category is then overplotted in a separate panel. Each dot represents a single SNV.

Statistical Modeling of Source Prediction Based on Fragment Lengths

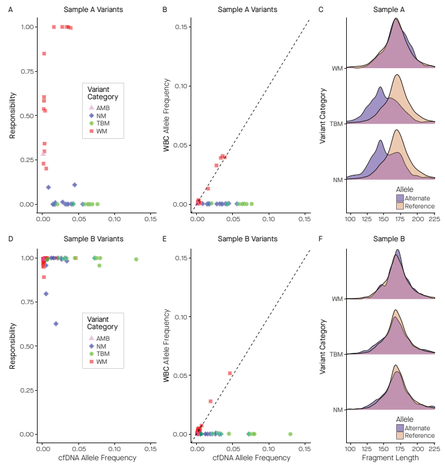
- In all samples, fragment lengths of molecules containing reference and alternate alleles for SNVs were recorded.
- A statistical model based on fragment lengths was built to predict the likelihood that an SNV belonged to a WBC-like source, without using the WBC sequencing results.
  - This statistical model was constructed as a mixture model: within each individual, a variant was either from a tumor-derived source or a blood-derived source.
    - Under the assumption that the variant is from a given source, the fragment lengths of molecules supporting that variant are each assigned a likelihood from that source distribution based on the density. Aggregating the likelihood over all fragments for a variant, we compared the total likelihood for the observed data coming from one source to the likelihood that the variant was derived from another source to estimate the likelihood that a variant was derived from one source or the other.
    - A latent variable representing the overall mixture probability within a sample (ie, the probability that a randomly selected variant comes from a given source) was constructed as part of the model, and individual variant cluster memberships (responsibilities) were computed by means of an Expectation Maximization algorithm run until convergence.
- Likelihoods of fragments of a given length from a given distribution were obtained from an estimated density of fragment lengths for each case. To establish a density for reference alleles, an Epanechnikov kernel was applied to the distribution of reference fragment lengths across samples to estimate density. For alternate alleles, a transformation of this density matching the observed typical distribution of alternate allele lengths in biopsy matched variants was generated: this avoided overfitting by restricting the degrees of freedom available in the density.
- Figure 2 depicts the four observed size distributions of the plasma DNA fragments:
  - Tumor biopsy-matched variants demonstrated the expected tumor-like shift to the left in the fragment length distribution.[2,3]
  - Interestingly, non-matched variants showed the same fragment length shift, suggesting that they are likely not noise, but rather may be variants related to the cancer that were not present in the particular biopsy sample.[12]
  - As expected, WBC-matched variants showed minimal shift in fragment length distribution.
  - Variants unable to be called (AMB) demonstrated intermediate fragment lengths.
- An illustration of the operation of the model is shown in Figure 3 using two participant examples: each variant in the given participant sample was plotted showing the frequency versus responsibility (source probability) for coming from the WBC-matched population of variants.
  - Individual variants of higher frequencies showed clear classification into categories, whereas lower frequency variants had intermediate responsibilities from the model.
- The participant shown in Figure 3A-C (metastatic esophageal cancer, age 61 years) shows the expected fragment length shift (Figure 3C).
- By contrast, in another participant (Figure 3D-3F; age 55 years, metastatic lung cancer) large differences in fragment length were not present (Figure 3F), limiting the ability to classify variants by means of fragment length within this individual.

Figure 2. Observed Fragment Length Distributions by Variant Category



Using the definitive classification derived from matched WBC and tumor tissue, the distribution of fragment lengths was plotted for each category. WBC matched variants had fragment lengths for both reference and alternate alleles, whereas tumor biopsy matched (TBM) variants showed an excess of shorter fragment lengths. Variants not matched to tumor biopsies showed the same shift, suggesting that they are also tumor derived. Variants with ambiguous assignment showed intermediate behavior, and thus were likely a mixture of types.

Figure 3. Classification within Individual Participant Samples



Examples of classification within individual participant samples: Participant A is depicted in A-C and Participant B in D-F. (A) Variants classified by fragment length into likely WM (responsibility near 1) and likely tumor derived (NM and TBM), responsibility near 0. Variants with very few alternate alleles were difficult to classify with certainty using fragment length; variants difficult to classify by fragment length were mostly resolved by matched WBC sequencing. (B) Variants showing WBC frequency matching. (C) Fragment length distributions by allele showing that within Sample A the distributions were very different by category. (D) Variants classified by fragment length into likely WM and likely tumor-derived. Note that within Sample B this yielded poor classification performance. (E) Variants showing WBC frequency matching. (F) Fragment length distributions by allele showing that within Sample B the distributions were not very different even for tumor biopsy-matched variants.
WM, WBC-matched; TBM, tumor biopsy-matched; AMB, ambiguous; NM, non-matched.

RESULTS

- A total of 21,604 SNVs were identified in the cancer and non-cancer samples: 4% were TBM, 68% WM, 19% NM, and 8% AMB (Table 2); the number of samples (non-mutually exclusive) that contributed to each category was 152, 1338, 499, and 761, respectively.

Table 2. Variant Characteristics

SNV Category, Sample Type	No. SNV Identified, n (%)	No. Samples with SNV (Total Samples)	Reference Allele Length, Median (SD)	Alternate Allele Length, Median (SD)
Tumor-matched	811 (4)	152 (1406)		
Cancer	811	152 (561)	167 (16.3)	156 (22.2)
Non-cancer	N/A	N/A	N/A	N/A
WBC-matched	14,788 (68)	1338 (1406)		
Cancer	9244	805 (561)	168 (16.3)	169 (14.8)
Non-cancer	5544	533 (845)	169 (14.8)	69 (14.8)
Non-matched	4197 (19)	499 (1406)		
Cancer	4071	400 (561)	167 (17.8)	158 (20.8)
Non-cancer	126	99 (845)	169 (16.3)	167 (17.8)
Ambiguous	1808 (8)	761 (1406)		
Cancer	1,322	497 (561)	166 (17.8)	164 (19.3)
Non-cancer	486	264 (845)	168 (14.8)	169 (14.8)

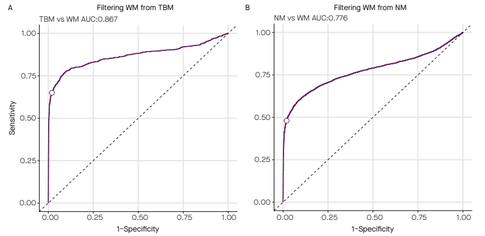
- Across SNV categories, the median (SD) length of fragments containing the reference allele was 167 (16.3). In samples derived from cancer participants, the median (SD) fragment lengths of alternate alleles were 156 (22.2; TBM), 169 (14.8; WM), 158 (20.8; NM), and 164 (19.3; AMB) (Table 2).
  - AMB and WM median SNV fragment lengths were similar to that of the reference allele, suggesting that fragment length shifts were minimal in SNVs derived from CH.
  - Fragment lengths of TBM and NM SNVs were similar; further, most NM SNVs came from cfDNA samples in the cancer cohort, suggesting that NM SNVs may be tumor-derived.
  - Most SNVs occurred in the WM category, which was expected in a population with a median (SD) age of 61 (12.2) due to age-related CH.[8-10]

References

1. Zheng YW1, Chan KC, Sun H, et al. Nonhematopoietically derived DNA is shorter than hematopoietically derived DNA in plasma: a transplantation model. *Clin Chem.* 2012 Mar;58(3):549-58.
2. Jiang P, Chan CW, Chan KC, et al. Lengthening and shortening of plasma DNA in cancer. *Proceedings of the National Academy of Sciences* Mar 2015, 112 (11) E1317-E1325.
3. Underhill HR, Kitzman JO, Hellwig S, et al. Fragment Length of Circulating Tumor DNA. *PLoS Genet.* 2016 Jul 18;12(7):e1006162.
4. Adalsteinsson VA, Ha G, Freeman SS, et al. Scalable whole-exome sequencing of cell-free DNA reveals high concordance with metastatic tumors. *Nat Commun.* 2017 Nov 6;8(1):1324.
5. Przybyl J, Chabon JJ, Spans L, et al. Combination Approach for Detecting Different Types of Alterations in Circulating Tumor DNA in Leiomyosarcoma. *Clin Cancer Res.* 2018 Jun 1;24(11):2688-2699.
6. Liu J, Chen X, Wang J, et al. Biological background of the genomic variations of cf-DNA in healthy individuals. *Annals of Oncology*, mdy513.
7. Hu Y, Ulrich BC, Supplee J, et al. False-Positive Plasma Genotyping Due to Clonal Hematopoiesis. *Clin Cancer Res.* 2018 Sep 15;24(18):4437-4443.
8. Genovese G, Kähler AK, Handsaker RE, et al. Clonal hematopoiesis and blood-cancer risk inferred from blood DNA sequence. *N Engl J Med.* 2014 Dec 25;371(26):2477-87.
9. Coombs CC, Zehir A, Devlin SM, et al. Therapy-Related Clonal Hematopoiesis in Patients with Non-hematologic Cancers Is Common and Associated with Adverse Clinical Outcomes. *Cell Stem Cell.* 2017 Sep 7;21(3):374-382.e4.
10. Jaiswal S, Fontanillas P, Flannick J, et al. Age-related clonal hematopoiesis associated with adverse outcomes. *N Engl J Med.* 2014 Dec 25;371(26):2488-98.
11. Surveillance, Epidemiology, and End Results (SEER) Program (www.seer.cancer.gov) SEER*Stat Database: Incidence - SEER 18 Regs Research Data + Hurricane Katrina Impacted Louisiana Cases, Nov 2017 Sub (2000-2015) - Linked To County Attributes - Total U.S., 1969-2016 Counties, National Cancer Institute, DCCPS, Surveillance Research Program, released April 2018, based on the November 2017 submission.
12. Gerlinger M, Rowan AJ, Horswell S, et al. Intratumor heterogeneity and branched evolution revealed by multiregion sequencing. *N Engl J Med.* 2012 Mar 8;366(10):883-892.

- The prediction model distinguished TBM from WM SNVs with an AUC of 0.87.
  - However, at a specificity of 98% (to match filtering based on WBC sequencing), false-negative rates were 35% (TBM; Figure 4A) and 52% (NM; Figure 4B).

Figure 4. Predictive Statistics for Distinguishing Tumor- Versus WBC-Derived Variants



Without white blood cell sequencing, WBC-matched variants are intermixed with other variants passing the noise filter. A) Using fragment length information, it is possible to partially classify WM variants from biopsy matched variants, however at high specificity, many biopsy matched variants are also lost. B) Similarly, the variants not matched in WBC and not matched to tumor can be partially classified by fragment length, but many are lost at high specificity.
WM, WBC-matched; TBM, tumor biopsy-matched; NM, non-matched.

CONCLUSIONS

- Characterizing the sources of cfDNA variants using high-depth, error-corrected sequencing (per-site error rate of <0.001) identified WBC-derived variants with low probability of error.
- By contrast, because most fragment length distributions from varied sources overlapped, fragment length alone did not strongly distinguish tumor-derived from WBC-derived variants.
- Therefore, to detect non-metastatic tumors, the lowest possible frequency of mutations needs to be analyzed reliably to find the lowest ctDNA fraction cancer individuals against this background.
- Together, these data suggest that source prediction based on fragment length alone is less robust than source assignment using individual-matched WBC sequencing, highlighting the importance of accounting for CH-derived SNVs when using targeted cfDNA-based approaches for cancer detection.