Extraction of genome-wide reads distribution pattern through whole genome sequencing of cell-free DNA for tracking tissue-of-origin in cancer patients

Han Liang[1,*], Fuqiang Li[1,*,†], Sitan Qiao[1], Xinlan Zhou[1], Guoyun Xie[1], Xin Zhao[1], Kui Wu[1,†]

[1]BGI-Shenzhen, 518083 Shenzhen, China.
*These authors contributed equally to this work.
†To whom correspondence should be addressed E-mail: wukui@genomics.cn (K.W.); lifuqiang@genomics.cn (F.Q.L.)

Abstract

Somatic mosaicism exists among tissues widely and would mark circulating cell-free DNA (cfDNA) as DNA fragments released by lytic cells from distinct tissues into the blood. By investigating the alignment pattern of sequencing reads from whole genome sequencing on genomic DNA of different tissues, we found the reads distribution forms tissue-specific patterns on some regions, as a result of somatic mosaicism. We then utilized this indication to construct a tissue-of-origin mapping model and evaluated the predictive performance on WGS data from tissue and cfDNA. In total, 1,545 tissue samples involving 13 cancer types were included, and the performance of identification of tissue-of-origin achieved specificity of 82% and sensibility of 80%. Furthermore, a total of 30 cfDNA samples involving lung cancer, liver cancer, and healthy control were analyzed to indicate their nidus' tissue-of-origin with specificity and sensibility both at 87%. Our results show that reads distribution of whole genome sequencing could be used to identify the tissue-of-origin of cfDNA samples with high accuracy, suggesting the potential application of our model on early tumor detection and diagnosis.

Main

Somatic mosaicism exists among tissues widely (*1, 2*) and would mark circulating cell-free DNA (cfDNA) as DNA fragments released by lytic cells from distinct tissues into the blood. Two groups reported abnormal Copy Number Variations (CNVs) of cfDNA from pregnant women with tumors (*3, 4*). However, researchers could not determine the tumor's types with CNVs that were information loss when CNV calling. We developed a sensitive model to catch somatic mosaicism footprints from reads distribution of whole genome sequencing data directly.

To develop our model of tracking tissue-of-origin for circulating tumor DNA (ctDNA), we first investigated the alignment pattern of sequencing reads from whole genome sequencing on genomic DNA of 1,545 tissue samples involving 13 cancer types from the Pan-Cancer Analysis of Whole Genomes (PCAWG) project of the International Cancer Genome Consortium (ICGC) and The Cancer Genome Atlas (TCGA) (*5, 6*). All the cancer types contain more than 60 donors (**Table 1**). Our technology includes the following 4 major steps (**Methods**): 1) Count the reads' distribution on reference. Firstly, we divide reference into series length-fixed windows, the typical windows-length is 10K, an empirical value. For simplicity sake, we join all chromosomes together (Y excluded), and gained a chain of 257973 windows (existing windows spanning two adjacent chromosomes). Then count each window's mapped-inside-window reads by sample, gain the number of reads (NRs) sequences. 2) Search the frequent distribution patterns among the samples. This step is trying to summarize the landscape of samples of the same type with frequent patterns. The pattern refers to the relationships between windows by NRs, more/equal/less (**Figure 1a**). We think that only those near-by windows would influence each other effectively. Notably, one

pattern could involve several windows, if those windows' relationships are common in samples. As an example, assume exists such a pattern "3, 1, 2", the numbers in which point to the windows' indexes, means that for many samples, the NR of window 3 is more than that of window 1, and the latter is more than that of window 2. Here, we sort a pattern's windows by NRs to describe their relationship simply (**Figure 1b**). 3) Extract type-special patterns from frequent patterns. After the previous step, we gain a large number of frequent patterns by types. Patterns gained from one type of samples are frequent for that type, but it doesn't mean that all patterns are not frequent for other types. We need to extract those type-specific patterns. Here, we use the transform of fisher-exact-test p-value to measure how "specific" a pattern is for one type when compared with another. The transformed value is called pattern's weight. Extract patterns whose weights are up above a calculated threshold. Obviously, when describe a type-special pattern, we must point it out that from which type the pattern is gained, and with which type the pattern is extracted. 4) Identify a sample's type according to type-special patterns. Two types of samples generate two frequent pattern sets, the two frequent sets after filtered with each other's samples gain two paired type-special pattern sets. When we try to determine to which one of this two a type-unknown sample is probably belonged, we observe how much patterns from each type-special pattern set match the sample, and calculate the sum of matched pattern's weights by type. Here, we say a pattern "match" a sample if the windows' relationship described by the pattern is also valid for the sample. Compare the two weighted sums, the type which the bigger one stands for is the possible type. Obviously, if we need to deal with three or more types, we need to repeat the step 3 for the combines of every two types, and integrate all results to vote a final answer. Finally, we executed 5-fold cross validation on tissue samples and found our model achieve high specificity of 82% and high sensibility of 80% (**Figure 2a**).

To evaluate the performance of identification the tissue-of-origin of cfDNA samples, a total of 30 cfDNA samples involving lung cancer, liver cancer, and healthy control were analyzed by our model. The cfDNA samples were sequenced on BGIseq-500 with average 3X depth of coverage. We execute 10-fold cross validation on cfDNA samples and found our model achieve high specificity of 87% and high sensibility of 87% (**Figure 2b**). There are 4 misjudged samples out of 30 samples, 1 healthy control sample misjudged as liver cancer, 1 lung cancer samples misjudged as liver cancer and 2 liver cancer samples misjudged as healthy control and lung cancer.

Our model distinguishes the healthy control samples with high accuracy, which is very important in early tumor screening. Generally, cfDNA concentration of healthy control is significantly lower than that of cancer patient. It seems that we can determine whether an individual bears tumor according to their cfDNA concentration, however, it may be unreliable for early tumor patients. On the other hand, we sequence all cfDNA samples with the same sequencing depth of 3X~4X, which means that all samples have the approximate total reads numbers, eliminating the concentration differences.

We consider that CNV is not always sensitive enough to descript cfDNA features for the low tumor cell DNA concentration in cfDNA. The reads distribution forms tissue-specific patterns used in our model focusses on the relationship of windows with NRs. In theory, relationship of two windows about NRs of the same sample is not affected by the sequencing depth. In fact, it works quite well under low sequencing depth. Furthermore, the biological meaning of reads distribution pattern is still ambiguous, and still needs more exploring works. Understanding the biology meaning of pattern would help us more effectively improve our method, and help to discover cancer mechanism, promote cancer treatments.

Table 1. Detail of Cancer Type from PCAWG Project

Cancer Type from PCAWG Project	Cancer Type Abbreviation	Number of Donors
Bone Cancer - United Kingdom	BOCA-UK	76
Breast ER+ and HER2- Cancer - European Union/United Kingdom	BRCA-EU	79
Chronic Lymphocytic Leukemia - Spain	CLLE-ES	100
Esophageal Adenocarcinoma - United Kingdom	ESAD-UK	100
Liver Cancer - Japan	LIRI-JP	259
Malignant Lymphoma - Germany	MALY-DE	101
Skin Cancer - Australia	MELA-AU	70
Ovarian Cancer - Australia	OV-AU	73
Pancreatic Cancer - Canada	PACA-CA	148
Pancreatic Cancer Endocrine neoplasms - Australia	PAEN-AU	69
Pediatric Brain Cancer - Germany	PBCA-DE	251
Prostate Adenocarcinoma - Canada	PRAD-CA	124
Renal Cell Cancer - European Union/France	RECA-EU	95

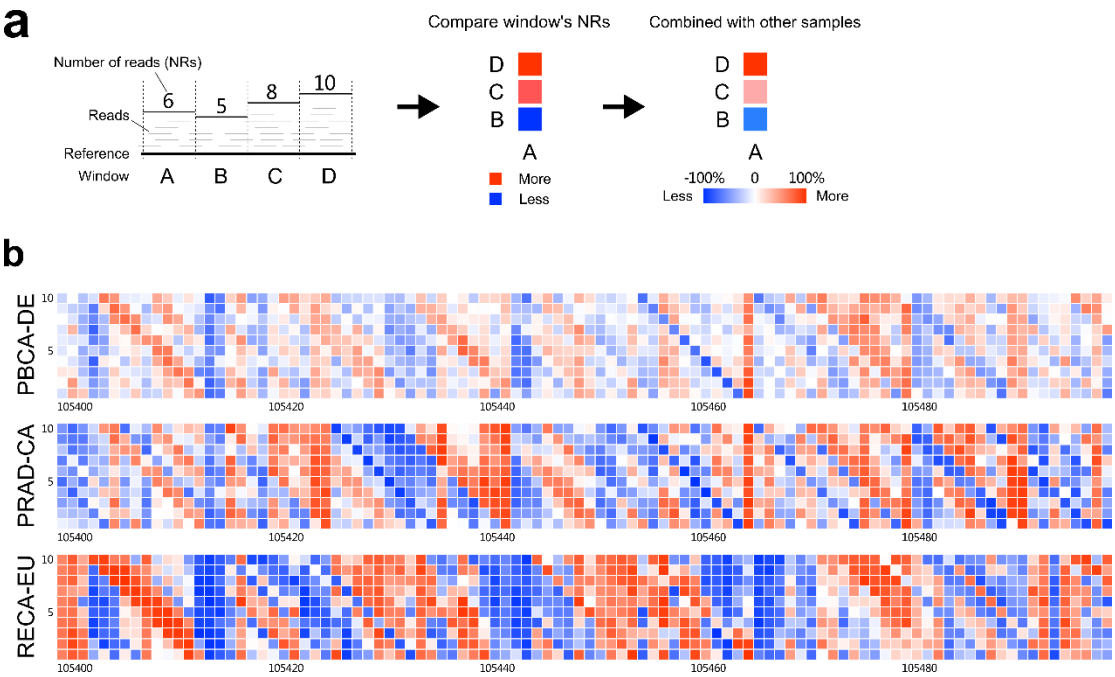

Figure 1. The reads distribution pattern on reference. a) The generation of reads distribution pattern. As what's shown, we divide the reference into 4 length-fixed windows, labeled A-D; count the number of reads (NRs) mapped on each window respectively; compared two windows by NRs to get the relationship, more/equal/less. When combined two windows' relationships of multiple samples, we use a percentage to represent the sample number difference between the NRs-of-A-more-than-B samples and the NRs-of-B-more-than-A samples (assume that the two windows were labeled A and B respectively). b) The reads distribution patterns of three types of

cfDNA samples. For the sake of simplicity, we joined all chromosomes (Y excluded) together, and obtain a long chain of 257973 windows. Here, we demonstrate the relationship of windows ranked 105400-105500, 100 windows in total, by group. Each window was compared with its 10 downstream windows.

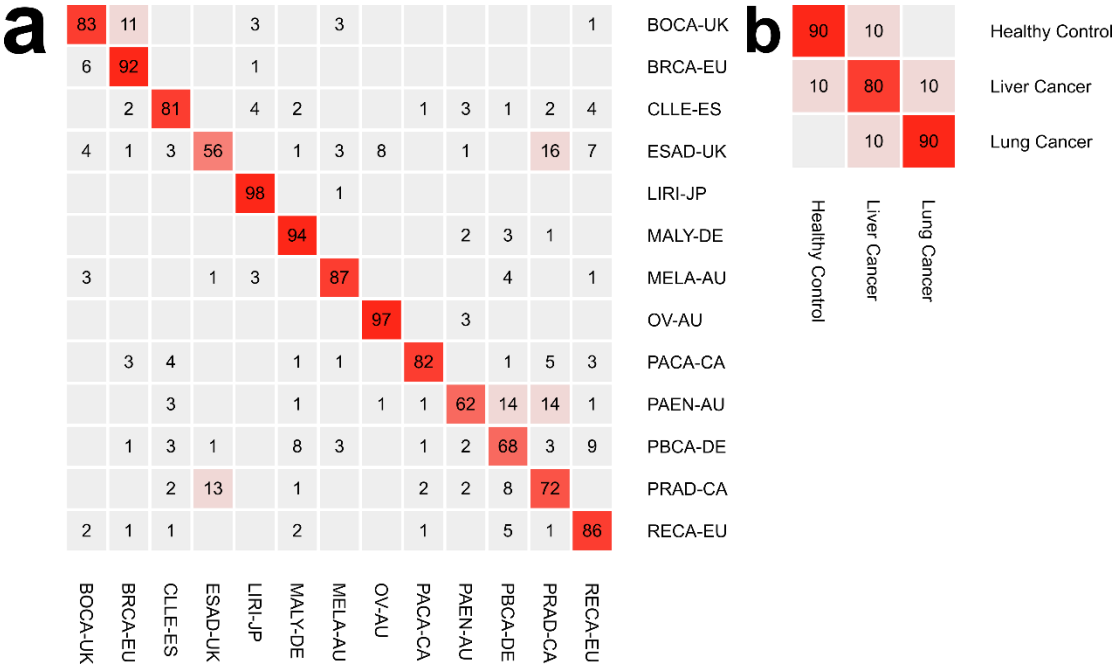

Figure 2. The predicted results of tissues and cfDNA samples. a) The results of tissue samples. This test involves 13 types of tissue samples, the figure shows the integration of 5-fold cross validation results. Rows stand for different types of samples, the right-side labels represent samples' real types, columns stand for predicted results, the bottom-side labels represented predicted types; the numbers inside cells represent the percentages of samples predicted as the predicted as the bottom-side labels in samples marked as the right-side labels. b) The results of cfDNA samples. This test involved 3 types of cfDNA samples, the figure shows the integration of 10-fold cross validation.

1. D. Freed, E. L. Stevens, J. Pevsner, Somatic mosaicism in the human genome. *Genes (Basel)* **5**, 1064-1094 (2014).
2. K. Yizhak *et al.*, RNA sequence analysis reveals macroscopic somatic clonal expansion across normal tissues. *Science* **364**, (2019).
3. D. W. Bianchi *et al.*, Noninvasive Prenatal Testing and Incidental Detection of Occult Maternal Malignancies. *JAMA* **314**, 162-169 (2015).
4. F. Amant *et al.*, Presymptomatic Identification of Cancers in Pregnant Women During Noninvasive Prenatal Testing. *JAMA Oncol* **1**, 814-819 (2015).
5. P. J. Campbell, G. Getz, J. M. Stuart, J. O. Korbel, L. D. Stein, Pan-cancer analysis of whole genomes. *bioRxiv*, (2017).
6. J. Zhang *et al.*, The International Cancer Genome Consortium Data Portal. *Nature biotechnology* **37**, 367-369 (2019).

ACKNOWLEDGMENTS

Funding: This work was supported by Science, Technology and Innovation Commission of Shenzhen Municipality under grant No. JCYJ20160531193931852.

Author contributions: K.W. and F.Q.L. conceived of the idea and supervised the work. H.L. developed the model of tracking tissue-of-origin for ctDNA. S.T.Q. and G.Y.X. performed standard pipeline of sequencing data. X.L.Z. and X.Z. performed experiments of sequencing. H.L., F.Q.L wrote the manuscript. K.W. contributed to drafting and revising the manuscript.

Data and materials availability: Alignment files of 1,545 tissue samples involving 13 cancer types from the Pan-Cancer Analysis of Whole Genomes (PCAWG) project were analyzed on the Cancer Genome Collaboratory, an academic compute cloud resource that allows researchers to run complex analysis operations across large ICGC cancer genome data sets. The sequencing data of 30 cfDNA samples have been deposited in the CNSA (https://db.cngb.org/cnsa/) of CNGBdb with accession code CNP0000680. The sequencing data of cfDNA samples are available on reasonable request. An up-to-date version of the analysis code, along with an up-to-date README, will available as a Github repository.

We are thankful to the production team of China National GeneBank, Shenzhen, China.

Methods

Count the reads' distribution on reference.

Here, we define the sorted windows indexes (SWI), which referring to a series indexes of windows divided on reference, those indexes are sorted according to numbers of window's mapped-inside-window reads, or NRs. In this paper, a SWI is considered as simplified reads distribution.

First of all, we need to count every sample's SWI, this step is described as the following:

1) Divide reference into series length-fixed windows, and labeled those windows with their indexes. For simplicity sake, we join all chromosomes together by the order of chromosome 1-22 & X (Y excluded). The window length is typical set as 10K, an empirical value, but sometime we would try another value in the range of 5K-50K.

2) Count each window's mapped-inside-window reads for each sample. The reads are mapped on reference after the standard short reads alignment. When a read spanning two windows, we consider the most-bases-located windows as its mapped window.

3) Get the SWI for each sample by sorting window-indexes according to NRs.

4) Repeat steps 2 and 3 until all samples are handled.

A sample produces a SWI, and SWIs gained from a sample set form a SWI set.

Search the frequent distribution patterns among SWIs.

The pattern is a series of order- sensitive numbers, which referring to windows indexes, e.g., (3, 1, 2), it is a mini-SWI. A pattern could content a series of numbers which are too many to search directly, we develop a model named splicing to find it:

If there are two patterns, one of which's tail section is the same with another pattern's head section, the operation of joining the former pattern and the latter pattern's remainder tail section, called splicing.

As an example, there are two patterns (1, 2) and (2, 3), the tail section of the former pattern is "2" which is the same with the latter pattern's head section "2", we will splice the former pattern (1, 2) with the latter pattern's remainder tail section "3" to gain a new longer pattern (1, 2, 3).

Actually, we only splice such two patterns whose indexes just content one different element with another to reduce computational complexity. For example, we could splice pattern (1, 2, 3) and (2, 3, 4) into a new pattern (1, 2, 3, 4), because the former and the latter are both just content one different element "1" and "4", respectively.

Splicing combines two shorter patterns would produce a longer one, and the shortest pattern just contents two indexes, called L2. We search L2s using the following formula:

$$L2(d, n)=(i, j) \quad i\epsilon N, j\epsilon N, 1\leq i\leq n, 1\leq j\leq n, |i-j|\leq d$$

Where n is the maximal index of windows, d is the maximal distance of two indexes of a L2. By experience, we set d as 40.

It is noteworthy that we just interest in frequent patterns, extra check is required. To determine whether a pattern is frequent, we check how much SWIs does this pattern

cover. Here, we say a pattern "cover" a SWI when the orders of windows contented by the pattern are the same with that in SWI. For example, we say pattern (3, 1, 2) covers SWI (4, 3, 1, 2) because the orders of "1", "2", "3" in pattern is the same with that in the SWI. Only if a pattern covered samples no less than a given threshold, will we consider it as a frequent pattern.

We try to gain all frequent patterns by splicing patterns iteratively, until no more frequent patterns are generated.

Frequent patterns gained from a sample set form a frequent pattern set.

Extract type-special patterns from frequent patterns

A frequent pattern is tissue-specific if there is significant difference between its coverages of two sample sets. We filter frequent patterns according to their coverages of two sample sets. When two pattern sets gained from two sample sets, after filtered with each other's sample set, gain two paired type-special pattern sets. To measure one frequent pattern's ability to distinguish two kinds of samples, we use the transform of Fisher-Exact-Test p-value as the pattern's weight. To calculate the p value, we need to check how much samples of each sample set are covered by a pattern. The transform formula is described as,

$$\mathrm{f(x)} = -\log(\max(1^{-100}, \mathrm{x}))$$

To judge a sample's possible type of the two types, we use the following formula,

$$\mathrm{score(S, P)} = \sum_{i=0}^{N} \begin{cases} weight(Pi), & if\ order(P_i, n_i) = order(S, n_i) \\ 0, & otherwise \end{cases}$$
$$n_i = index(P_i) \cap index(S)$$

Where S is a SWI extracted from the to-be-judged sample, P is one of two paired tissue-specific pattern sets, N is the total number of P's pattern(s), weight(x) is the weight of pattern x, P_i is the *i*-th pattern of P, order(x, y) is the order vector of index set y in pattern/SWI x, index(x) is the index set of pattern/SWI x.

Compare two scores of the two tissue-specific pattern sets, the type which the higher score stands for will be considered as the possible type.

The differences between the weighted sums of two specific pattern sets would enormous, making one score always be larger than another, we need to delete some patterns to balance them with the following formula,

$$\mathrm{balance}(\mathrm{C}_a, C_b | P_a, P_b, S_a, S_b)$$
$$= \underset{C_a, C_b}{\mathrm{argmin}}\, p(co(S_a, P_a, C_a, P_b, C_b), in(S_a, P_a, C_a, P_b, C_b), co(S_b, P_a, C_a, P_b, C_b), in(S_b, P_a, C_a, P_b, C_b))$$

Where C_a and C_b are two wanted factors. C_a is used as a weight-threshold to filter the tissue-specific pattern set P_a which is gained from sample set S_a; C_b, P_b and S_b are belonged to anther sample type. When a pattern set is filtered with a threshold, we will

delete all patterns whose weights are less than that threshold. The p(a, b, c, d) is the p value calculated using the Fisher exact test with factors a, b, c, d; co(s, a, b, c, d) is the number of samples judged correctly from sample-set S with pattern set a filtered with threshold b and pattern set c filtered with threshold d; in (s, a, b, c, d) is the number of samples misjudged with similar factors with the function co.
To solve this formula, we use the Expectation-Maximum like algorithm. At first, we set C_a as a reasonable random value, and find the best C_b under the given situation; then keep the C_b unchanged, and find the best C_a. This is an iterative process with the end condition of C_a and C_b never change again.
We use C_a and C_b to filter the two paired tissue-specific pattern sets, and update them.

Identify a sample's type according to type-special patterns

The way to identify a sample's type is introduced in the previous step, but just for two types. When try to judge a sample's type from N (N>2) candidates, we need repeat the previous step for combines of every two types. Obviously, for N types, there will be N(N-1)/2 combines. In this situation, every repeat will provide a possible answer, all these answers can vote a final one.