A Study of Cell-free DNA Fragmentation Pattern and Its Application in DNA Sample Type Classification

Shifu Chen*†‡, Ming Liu†, Xiaoni Zhang†, Renwen Long†, Yixing, Wang†, Yue Han†, Shiwei Zhang†, Mingyan Xu†, Jia Gu*§
*Shenzhen Institutes of Advanced Technology, Chinese Academy of Sciences, Shenzhen, China.
†HaploX Biotechnology, Shenzhen, China.
‡University of Chinese Academy of Sciences, Beijing, China.
§Corresponding Author.

Abstract—Plasma cell-free DNA (cfDNA) has certain fragmentation patterns, which can bring non-random base content curves of the sequencing data's beginning cycles. We studied the patterns and found that we could determine whether a sample is cfDNA or not by just looking into the first 10 cycles of its base content curves. We analysed 3189 FastQ files, including 1442 cfDNA, 1234 genomic DNA, 507 FFPE tumour DNA and 6 urinary cfDNA. By deep analysing these data, we find the patterns are stable enough to distinguish cfDNA from other kinds of DNA samples. Based on this finding, we build classification models to recognise cfDNA samples by their sequencing data. Pattern recognition models are then trained with different classification algorithms like k-nearest neighbours (KNN), random forest and support vector machine (SVM). The result of 1000 iteration .632+ bootstrapping shows that all these classifiers can give an average accuracy higher than 98%, indicating that the cfDNA patterns are unique and can make the dataset highly separable. The best result is obtained using random forest classifier with a 99.89% average accuracy ($\sigma = 0.00068$). A tool called CfdnaPattern (http://github.com/OpenGene/CfdnaPattern) has been developed to train the model and to predict whether a sample is cfDNA or not.

Index Terms—Cell free DNA, Liquid Biopsy, Fragmentation, Pattern Recognition

1 Introduction

The extracellular DNA fragment, which is called cell-free DNA (cfDNA), was first found in the human plasma by Mandel and Metais [1], and then found in other body fluids, like urine [2], pleural effusion [3] and cerebrospinal fluid [4]. For healthy people, cfDNA is mainly released from cell apoptosis [5], and partly released from necrosis [6] and active cell release [5]. But for tumour patients, it is known that tumour cells can release large amount of DNA carrying lots of mutation information from tumour cells, which is called circulating tumour DNA (ctDNA) .

As techniques like next generation sequencing (NGS) become cheaper and better, genetic testing using cfDNA samples goes popular and has been applied into clinical environments. After the boom of non-invasive prenatal testing (NIPT) [7] using cell-free fetal DNA, tumour genetics testing, which mainly relies on the analysis of ctDNA [8], [9], is considered as a much bigger opportunity. According to previous studies [29], the variations detected in tumour patient's cfDNA usually have highly consistency with the variations detected from the same patient's tumour tissue samples. CtDNA testing is usually non-invasive, highly available, truly dynamic, and is able to profile tumour heterogeneity. These features make ctDNA testing more feasible than tissue testing for tumour genetics diagnosis, which plays a key role for personalised tumour treatment, tumour monitoring and screening. A new term, liquid biopsy, was created for tumour genetics testing using cell-free DNA, and was presented as the top of 10 breakthrough technologies in 2015, by MIT Technology Review Press.

NGS is considered as the most common and powerful technique to discover DNA alterations in liquid biopsy. Lots of studies reported the methods of detecting genetic variations from cfDNA sequencing data, while some other studies reported hundreds of cases using mutation information found in cfDNA sequencing to guide tumour treatment. However, only few studies paid attention to the data and sequence characteristics of cfDNA, such as length distribution [10], break positions, and especially the fragmentation patterns [11].

For cfDNA length distributions, previous studies reported there is a dominant peak around 166bp, and two small peaks around 350bp and 510bp. These three peaks are roughly the size of DNA with mono-, di-, and tri-nucleosome. Shorter than 166bp, there are small peaks near 152bp, 143bp, 133bp, 122bp, 112bp and 102bp, which show a 10 bases periodicity, same as one turn of a double helix DNA in nucleosome packaging. These cfDNA size distributions obviously reflect the nucleosome guided fragmentation patterns [10].

The cfDNA fragmentation patterns were first reported by Chandrananda at one nucleotide resolution in 2014 [11]. When he investigated the sequencing coverage bias of cfDNA from pregnant women, he found some high frequency 10-nucleotide motifs on either side of cfDNA

fragments. Specially, he found the first two bases of the cfDNA at cleavage site could determine most of the other 8 bases. His further study in 2015 indicated these fragmentation patterns were related to the non-random biological cleavage over chromosomes. The 10 positions on either side of the DNA cleavage site show consistent patterns with preference of specific nucleotides for nucleosomal cores and linker regions [12]. A further study [13] stated that the non-random cfDNA pattern could reflect its chromatin features, such as epigenetic landscapes and gene expressions. In a latest study [14] using deep sequencing of cfDNA from healthy people and tumour patients, a dense and genome wide map of nucleosome occupancy was constructed to distinguish the original cell types of cfDNA . These works show that cfDNA fragmentation patterns have the potential to be a novel biomarker for disease diagnosis.

This paper also focuses on the study and application of cfDNA fragmentation patterns. We will present our analysis result of cfDNA fragmentation patterns from more than one thousand samples, including cfDNA, genomic DNA and FFPE samples. Same as previous studies, we found the fragmentation patterns were non-random, and we confirmed they were unique and stable. These features give the fragmentation patterns ability to identify if a sequenced sample is cfDNA, or not cfDNA. Using pattern recognition technologies, we build a cfDNA classifier just using the fragmentation patterns extracted from the sequence data, and our evaluation result showed it could achieve about 99.89% accuracy by average ($\sigma = 0.00068$).

We developed an open-source tool CfdnaPattern for training this cfDNA classification model. Interestingly, we found our tool could be used to scan regular sequencing FastQ files to detect if there exists the possibilities that cfDNA samples are marked or treated as non-cfDNA samples, or vice versa. This feature makes CfdnaPattern useful for regular data auditing to detect the messing up of cfDNA and other not-cfDNA samples.

2 METHODS

The data used to train and validate the models are a part of recent sequencing output from HaploX Biotechnology. 3189 FastQ files, including 1442 cfDNA, 1234 genomic DNA (gDNA), 507 FFPE tumour DNA and 6 urinary cfDNA, are gathered into this dataset. Some cfDNA and gDNA samples are paired, which means a pair of cfDNA and gDNA are both extracted from same single tube of blood. After centrifugation of the blood sample, cfDNA is extracted from plasma, while gDNA is extracted from blood cells. Most samples are sequenced after DNA target capture using gene panels, and few samples are sequenced in a whole exome or whole genome wide.

2.1 Experiment description

The collected blood samples were processed within 2 hours, and were centrifuged at 1600 g for 10 minutes at 4°C to separate the blood cells and plasma. Next, plasma was transferred into a new 10 mL tubes and centrifuged at 16000 g for 10 minutes at 4°C to remove residual cells. Then they were stored at -80 °C.

cfDNA was extracted with the Serum / Plasma Circulating DNA Kit (Tiangen) according to the manufacturer's protocol from 2 mL plasma [15]. gDNA was extracted with the TIANamp Blood DNA Kit (Tiangen) according to the manufacture's protocol from 2 mL blood cell fraction [16]. FFPE DNA was extracted with the GeneRead DNA FFPE Kit (Qiagen) according to the manufacturer?s protocol from 10 mm thick sections of FFPE tissue blocks . GeneRead FFPE purification process introduced in the UNG enzymatic treatment could dramatically eliminate the artifactual G>T or G>A mutations [17], [18].

The concentration of purified DNA was determined by a Qubit dsDNA HS assay (Invitrogen), Qubit-quantified genomic DNA (6 ug; non-amplified) and FFPE DNA (2 ug; non-amplified) was digested to 100-250bp fragments using the NEBNext dsDNA Fragmentase (NEB, M0348L) [19], and the cfDNA was directly used for library construction. Libraries were constructed using enzymatic reagents from KAPA Library Preparation kits (KAPA Biosciences) according to protocols [20]. NimbleGen SeqCap EZ Choice(Roche) was used for hybridization-based enrichment according to the manufacturer's protocol [21].

2.2 Feature selection

Initially the fragment length and fragmentation patterns of cfDNA are both considered as features to be extracted from cfDNA sequencing data. The cfDNA fragment length has certain distribution, such its peak is usually near 166bp [31]. This characteristic makes cfDNA length distribution a very good feature for modelling cfDNA classifiers. Fig.1 shows a typical curve of cfDNA fragment length distribution. However, the calculation of DNA template length distribution requires pair-end sequencing, so this feature will be not available for single-end sequencing data. Secondly the DNA length distribution can be affected by the sequence trimming operation, which is a commonly applied before in data preprocessing stage. Furthermore, we usually need to do the time-consuming alignment process before we can calculate the insert size, this will take more computation resource and can be slow. For above reasons, we gave up the idea of using cfDNA fragment length distribution as a feature of the classifier model.

To make our cfDNA classification model to be more universal for different sequencing methods, we switched to using only DNA fragmentation patterns as the features. According to previous studies [11], the cfDNA fragmentation patterns locate in the first 10 base pairs. We did statistics of the selected 1442 cfDNA FastQ files, and the results also support the same conclusion. The $ATCG$ ratio of the first 10 cycles are not flat, but highly homological across different samples. For each FastQ file, we counted the $ATCG$ base numbers in first 10 cycles, calculated their percentages, and then stored these $ATCG$ base content ratios of 10 cycles in a 40-element vector to be the feature. Fig.2 shows the mean ratio curves of $ATCG$ bases and the quartile values at every cycle.

To confirm whether this 40-element feature is good to classify different kinds of DNA data, we first conducted principle component analysis (PCA) [30] to see if the dataset is separable by this feature. Fig. 3 shows the PCA result with

This article has been accepted for publication in a future issue of this journal, but has not been fully edited. Content may change prior to final publication. Citation information: DOI 10.1109/TCBB.2017.2723388, IEEE/ACM Transactions on Computational Biology and Bioinformatics

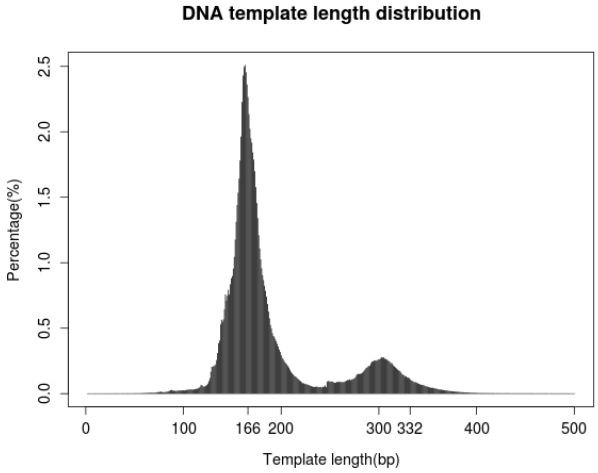

Fig. 1. The cfDNA template length distribution calculated from one of our sequencing files. DNA fragments are not filtered by length before library preparation, so we can find many long reads in this figure. 166bp and 332bp are marked as ticks in the X-axis and we can find they match the highest and secondary peaks.

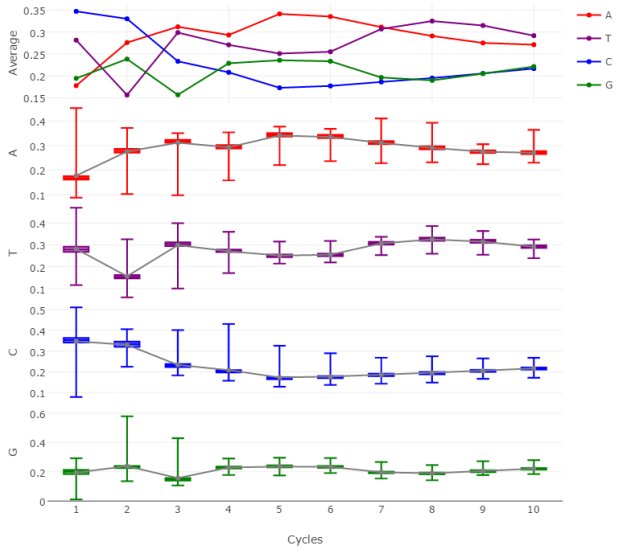

Fig. 2. The cfDNA fragmentation patterns. The $Average$ figure in the top gives the average base content percentage curves of $ATCG$ bases. Under the average figure, there are four individual curves for each kind of base, with the mean values presented by the grey lines, the quantile values presented by the box, and the maximum or minimum values presented by the dashes on the top and bottom. From this figure we can find all cycles of each base type have very compact quantile ranges, which means the distribution variations are relative little and the patterns are consistent.

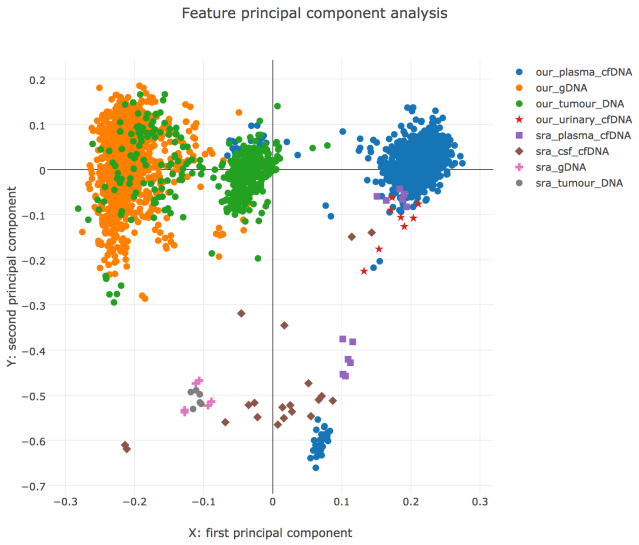

Fig. 3. The PCA result with this 40-element feature for all data. By transforming with the PCA result, each 40D vector is reduced to a 2D vector. X-axis is the first principle component and Y-axis is the second principle component. We can find cfDNA can be mostly separated by even using only the first principle component. The samples sra_xxx are downloaded from NCBI SRA, and sra_csf_cfDNA means cerebrospinal fluid cfDNA data downloaded from NCBI SRA.

only top 2 principle components are kept. From Fig. 3, we can find this feature is good enough to separate cfDNA and non-cfDNA data.

By looking to the base content curves of thousands of cfDNA, we found the base content ratio fluctuation is a relative stable value. For each base b, let R_n^b denotes its base content ratio at cycle n, we can define its base content fluctuation at this cycle as:

$$F_n^b = \begin{cases} -1, & R_n^b < R_{n+1}^b \\ 0, & R_n^b = R_{n+1}^b \\ 1, & R_n^b > R_{n+1}^b \end{cases} \quad (1)$$

Since there are $ATCG$ four different bases, if we count the first N cycles, we will have a $4 \times (N-1)$ dimension fluctuation vector $(F_1^A, F_1^T, F_1^C, F_1^G, ..., F_{N-1}^A, F_{N-1}^T, F_{N-1}^C, F_{N-1}^G)$. We can then calculate the vector $\overline{F_{cfdna}}$ as the mean base content ratio fluctuation of cfDNA, and $\overline{F_{gdna}}$ as the corresponding value of gDNA . For a sample S, we can calculate the angle of its base content ratio fluctuation F_S to $\overline{F_{cfdna}}$ and $\overline{F_{gdna}}$ by:

$$\begin{cases} A_{cfdna} = arccos \dfrac{F_S . \overline{F_{cfdna}}}{|F_S| \times |\overline{F_{cfdna}}|} \\ \\ A_{gdna} = arccos \dfrac{F_S . \overline{F_{gdna}}}{|F_S| \times |\overline{F_{gdna}}|} \end{cases} \quad (2)$$

We then visualised all samples' (A_{cfdna}, A_{gdna}) with different cycle number settings to see whether this angle value can be used to separate cfDNA from other DNA samples. Fig. 4 shows the data plotting result with cycle = 5, 6, 8, 10, from which we can find that this feature is able to separate the dataset if cycle ≥ 6.

This article has been accepted for publication in a future issue of this journal, but has not been fully edited. Content may change prior to final publication. Citation information: DOI 10.1109/TCBB.2017.2723388, IEEE/ACM Transactions on Computational Biology and Bioinformatics

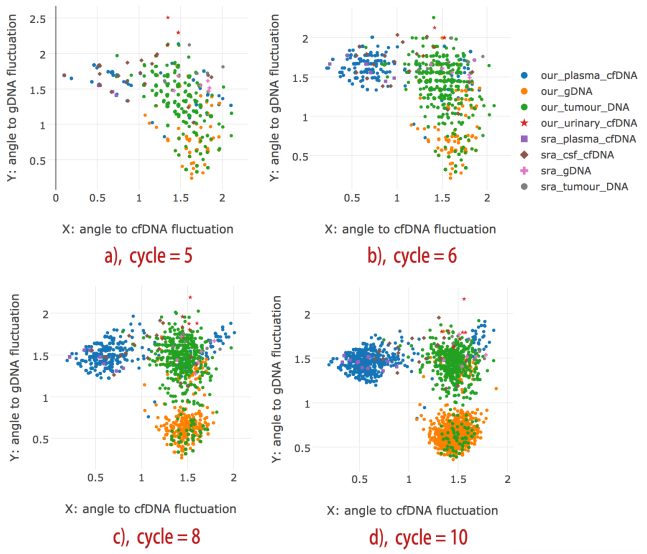

Fig. 4. The ploting result of all samples' (A_{cfdna}, A_{gdna}) with cycle = 5, 6, 8, 10. Each point is a sample, with its X-axis value is A_{cfdna}, which means the angle to $\overline{F_{cfdna}}$ in radian, while its Y-axis value is A_{gdna}, which means the angle to $\overline{F_{gdna}}$ in radian. We can find cfDNA can be mostly separated if cycle $\geq$ 6. Particularly, the 12 plasma sample data downloaded from NCBI SRA (purple squares) are well clustered with the plasma sample data sequenced in our lab (blue circles), which is better than the PCA clustering result.

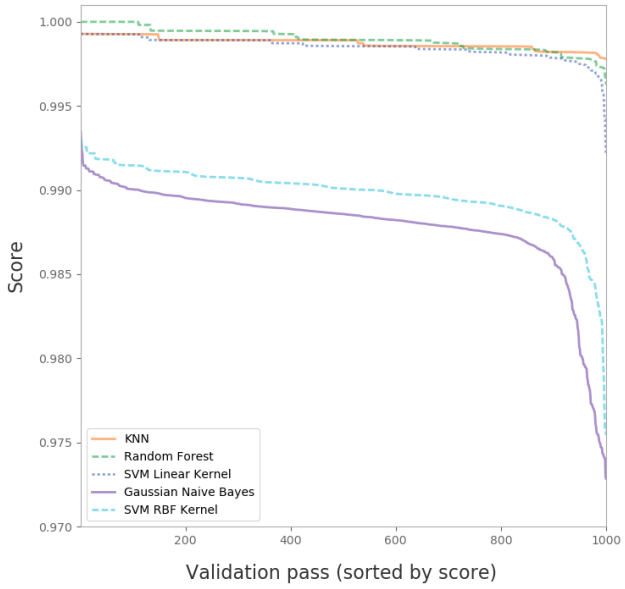

Fig. 5. The benchmark results of cross validation with 1000 iteration bootstrapping using different algorithms. Y-axis values are the sorted accuracy score, 1.0 means 100% accuracy. From this figure, we can find that random forest ($treeNumber = 20$) gives the best performance (average accuracy = 99.89%), KNN ($K = 8, weights = uniform$) and linear SVM ($norm = L2, loss = hinge$) give comparable performance, while nonlinear SVM ($kernel = rbf, degree = 3$) and GNB give the worst. LibD3C classifier achieved 99.1% of weighted accuracy with 5-fold cross validation, but its result is not shown here since it requires nontrivial efforts to support bootstrapping and .632+ rules.

By looking to the result of PCA analysis and base content ratio fluctuation, we confirm this 40-element feature is a good candidate to classify cfDNA samples and other DNA samples. Then we can train classification models based on this feature.

2.3 Model training and validating

To train this cfDNA classification model, we tried different supervised learning algorithms including k-nearest neighbours (KNN), SVM with both linear and radial basis function (RBF) kernels, Gaussian Naive Bayes [23] and some ensemble algorithms like random forest [24] and libD3C [25].

We implemented bootstrapping [26] and applied .632+ rules [27] to evaluate the model errors. 1000 iterations of bootstrapping were performed, within each iteration the dataset was split into training set and validating set using random sampling with replacement. Model error in each iteration was evaluated using .632+ rule, and the mean error and mean accuracy were immediately obtained after all 1000 iterations were done. We can then simply compare the performance of different algorithms using the mean accuracy obtained from bootstrapping evaluation. To directly visualise the performance of different algorithms, we sorted the scores and plot them in a same figure, which is shown in Fig.5.

2.4 Software Implementation

This project is developed in Python and the core classification functions, including training and predicting utilise the well documented machine learning library scikit-learn [22]. Feature extraction is the most time-consuming process of the whole pipeline. We both tried using all or just a part of the reads in FastQ files to calculate the 40-element feature vector, and found they only produced very little difference. So we only count 10,000 reads by default to make this program ultra-fast. Typically it only takes a few minutes to do both feature extraction, training and validating on this 2362 files. Once the feature extraction process is done, its result is cached as a JSON file. Loading this JSON file to do training and validating again only takes a few seconds.

The CfdnaPattern tool is available at github (https://github.com/OpenGene/CfdnaPattern) with good maintenance. We have uploaded our pre-trained model for convenient use to classify cfDNA and not-cfDNA sample data. But be aware that the patterns of cfDNA sequencing data can be slightly different for whole genome sequencing, whole exome sequencing and other target sequencing. And the gene panels used to do target capturing can also affect the cfDNA fragmentation patterns. So it is preferred for users to train their own models if they have enough data. But commonly the pre-trained model is good and accurate enough to handle most classification tasks.

3 Results

We first used this tool to test other plasma cfDNA and genomic DNA samples outside the training and validating datasets, and it gave near 100% accuracy for classifying hundreds of samples. After we confirmed the performance of our model, we started to use this tool to scan our every new FastQ file, and found ever one special case that CfdnaPattern prediction result was not consistent with the sample

This article has been accepted for publication in a future issue of this journal, but has not been fully edited. Content may change prior to final publication. Citation information: DOI 10.1109/TCBB.2017.2723388, IEEE/ACM Transactions on Computational Biology and Bioinformatics

registry tables. In this case, a sample registered as cfDNA was predicted as not-cfDNA, and its paired genomic DNA was predicted as cfDNA. After careful checking the experiment processes, we confirmed the prediction was correct, and the cfDNA and genomic DNA samples were actually swapped by using incorrect adapters. This story indicates that CfdnaPattern tool can be used for regular sequencing data auditing, to prevent the happening of messing up of DNA types by incorrect experiment or wrong regents. The ability to check correctness of experiment and data analysis processes is extremely important for providing a high quality sequencing and analysing service, especially for those applications in clinical environments.

Then we tested cfDNA data from other cell-free fluids rather than plasma. First we sequenced 5 urinary cfDNA samples with target capturing, and we found their patterns are similar, but different from the patterns of plasma cfDNA. Then we downloaded some cerebrospinal fluid cfDNA sequencing data from NCBI SRA (accession numbers: SRR2496749 SRR2496739, SRR2496735, SRR2496731, SRR2496709, SRR2496702, SRR2496699, SRR2496693, SRR1656605, SRR1654347) [28], along with some plasma cfDNA data (accession numbers: SRR2496737, SRR2496711, SRR3706309, SRR3706280, SRR3706298), some germline genomic DNA data (accession numbers: SRR2496740, SRR2496716, SRR2496710) and some tumour DNA data (accession numbers: SRR2496722, SRR2496689, SRR2496713). We tried to predict if our model can differentiate the cfDNA samples and not-cfDNA correctly. The result showed that CfdnaPattern tool could correctly classified plasma cfDNA as cfDNA, genomic and tumour DNA as not-cfDNA. Visualised figures showed the patterns of cerebrospinal fluid cfDNA samples are varying and different from the patterns of plasma cfDNA, but the result showed that most of them still can be classified as cfDNA sample. We evaluated 44 items downloaded from NCBI SRA, and the result reported only 1 item wrongly predicted, which gave a prediction accuracy of 97.73%. Fig.6 plots some cerebrospinal fluid cfDNA and plasma pattern together.

To analyse this cfDNA fragmentation pattern more deeply, we split the original sequencing data by different criterions. The first immediate thought is to split aligned cfDNA sequencing data by chromosomes and try to find whether different chromosomes have different patterns. We got a result that the patterns of the 22 autosomes and 2 sex chromosomes were similar, and all these chromosome data could be recognised as cfDNA. We also separated the reads into forward reads and reverse reads to see if the patterns could be different but still got a negative result. Since mitochondria are not inside the nucleus and are not protected by nucleosome, we supposed that the cfDNA came from mitochondria might not have the same patterns as autosomes. We did find the patterns of mitochondria are a bit different, but the difference was much less than we expected. This result indicates that the chromosomes and mitochondria share major mechanisms of cleaving and producing cfDNA. Fig.7 shows a comparison of the patterns of a whole genome sequencing data and the its subset of mitochondrial DNA.

To figure out if DNA fragments with different lengths

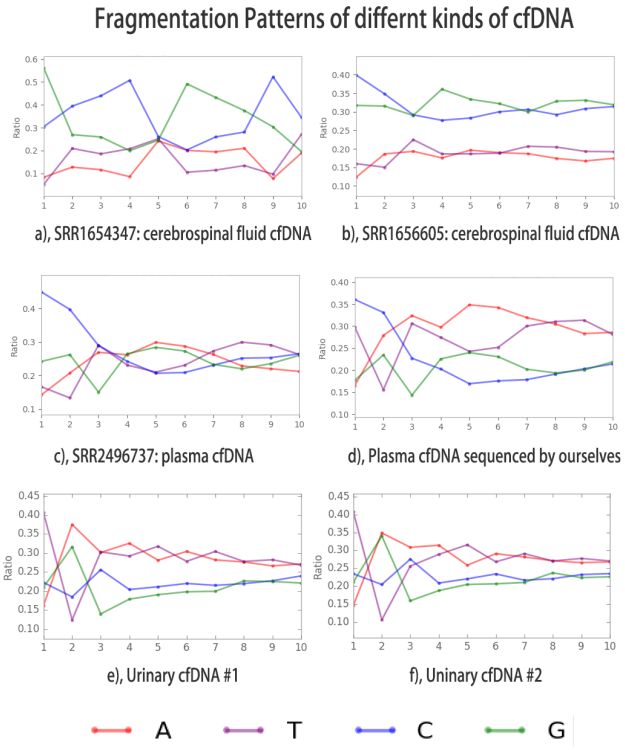

Fig. 6. The fragmentation patterns of different cfDNA samples. Sub-figure a, b are target sequencing with cerebrospinal fluid cfDNA, sub-figure c and d are target sequencing with plasma cfDNA, sub-figure e and f are target sequencing with urinary cfDNA. Data of sub-figure a, b and c are downloaded from NCBI SRA, which are produced by same project (accession number: PRJNA266729) and uploaded by same research group [28]. Data of sub-figure d, e and f are sequenced by ourselves. From this figure, we can find the patterns of cerebrospinal fluid cfDNA vary a lot, while the patterns of urinary cfDNA are similar. We can also find some differences between the two plasma cfDNA samples, with one downloaded from NCBI SRA, and the other sequenced by ourselves. The differences, which may be caused by different experiment processing methods, is relative small so both samples can be classified as plasma.

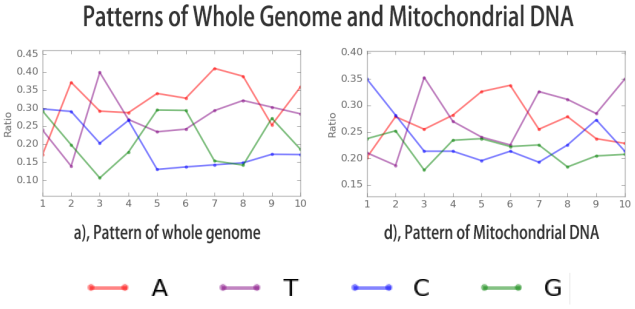

Fig. 7. The fragmentation patterns of whole genome sequencing data and its mitochondrial DNA subset. Although mitochondria are not inside the nucleus, we can still find the patterns are very similar.

can have different patterns, we split the original sequencing data by sequencing DNA template length. In our experiment, we separated the pair-end sequencing reads by every 10 bp step, which means reads of 150-159 bp and 160-169 bp will be categorised into different subsets. We studied the patterns of every subsets and tried to recognise them as cfDNA using our tool. The result showed different length subsets shared the same patterns, and most of them could be successfully recognised as cfDNA, even for those in short length range of 60-69 bp, or long length range of 290-299 bp. After the cfDNA degradation phenomenon is known [13], a reasonable speculation is: for those short cfDNA fragments in fresh blood, are they mainly produced by such degradation of long cfDNA fragments when they are circulating? It seems that our result doesn't support this speculation. Since if it is true, the short cfDNA fragments should lose their fragmentation patterns.

4 Discussion

In summary, the fragmentation patterns of cell-free DNA were carefully studied, and a classification model was built based on these patterns to discriminate cfDNA or non-cfDNA sequencing data. A tool called CfdnaPattern is developed for training and benchmarking the cfDNA classifiers using different algorithms, and the chosen classifier achieved an average of 99.89% accuracy ($\sigma = 0.00068$) in the cross validation process.

Deep analysis of different chromosomes gave similar patterns for autosomes and sex chromosomes. Comparing to the patterns of whole genome sequencing, the patterns of mitochondrial DNA were found closer than supposed, suggesting that all chromosomes and mitochondria may share same mechanisms to produce cfDNA. The cfDNA fragmentation patterns of long or short cell-free DNA fragments also have similar patterns, which indicates the short DNA fragments are mainly from the release from cells, not produced by the degradation of longer fragments when they are circulating.

Since the genomic DNA has different patterns from cfDNA, we can imagine that the data patterns will be a combination of cfDNA and gDNA patterns if large amount of gDNA get mixed into cfDNA sample. This idea gives a possible way to check if hemolysis happened during the blood storage or shipping stages. For pair-end sequencing, the read length distribution can be also used as an important feature for such classification tasks. We will try to implement this hemolysis checking function once we collect enough data from samples with hemolysis for training our models. This feature will be also included in CfdnaPattern tool once it is finished.

Some other kinds of cfDNA also worth a study. We analysed the pattern from urinary cfDNA and cerebrospinal fluid cfDNA, we found urinary cfDNA were with certain patterns while we didn't find any stable patterns for cerebrospinal fluid cfDNA. Other kinds of cfDNA, like pleural fluid cfDNA may also have certain patterns, but remain to be examined. Since the data of cerebrospinal fluid cfDNA are downloaded from NCBI SRA, we currently cannot determine whether the difference and instability of cerebrospinal fluid cfDNA patterns are caused by different cells, or caused by different experiment processing methods.

As one of the most important liquid biopsy technologies, cfDNA sequencing will play a more important role in applications like cancer diagnosis, monitoring and screening. For such applications, the correctness of experiments and data analysis pipelines should be guaranteed. This tool proposed in this paper is helpful for fast analysis of cfDNA sequencing data to check sample type identity. And in future, we will extend its functions to give more support to cfDNA sequencing quality control and data auditing, which are indispensable processes to provide steady sequencing service.

Acknowledgments

This study was financed partially by National Science Foundation of China (NSFC Project No.61472411), Technology Development and Creative Design Program of Nanshan Shenzhen (Project No. KC2015JSJS0028A), and Special Funds for Future Industries of Shenzhen (Project No. JSGG20160229123927512).

References

[1] Mandel P, Metais P. (1948). Les acides nucleiques du plasma sanguin chez l'homme [in French]. C R Seances Soc Biol Fil 142:241-243.
[2] otezatu I, Serdyuk O, Potapova G, Shelepov V, Alechina R, Molyaka Y, Anan'ev V, Bazin I, Garin A, Narimanov M, Melkonyan H, Umansky S, Lichtenstein AV. (2000). Genetic analysis of DNA excreted in urine: a new approach for detecting specific genomic DNA sequences from cells dying in an organism. Clin Chem 46:1078-1084.
[3] Sriram KB, Relan V, Clarke BE, Duhig EE, Windsor MN, Matar KS, et al. (2012). Pleural fluid cell-free DNA integrity index to identify cytologically negative malignant pleural effusions including mesotheliomas. BMC Cancer 12:428.
[4] Liimatainen SP, Jylhv J, Raitanen J, Peltola JT, Hurme MA. (2013). The concentration of cell-free DNA in focal epilepsy. Epilepsy Res 105(3):292-8
[5] Stroun M, Lyautey J, Lederrey C, Olson-Sand A, Anker P. (2001). About the possible origin and mechanism of circulating DNA: Apoptosis and active DNA release. Clin Chim Acta 313(1-2):139-42.
[6] Jahr S, Hentze H, Englisch S, Hardt D, Fackelmayer FO, Hesch RD, et al (2001). DNA fragments in the blood plasma of cancer patients: quantitations and evidence for their origin from apoptotic and necrotic cells. Cancer Res 61(4):1659-65
[7] Chiu RWK, Chan KCA, Gao Y, Lau VYM, Zheng W, et al. (2008). Noninvasive prenatal diagnosis of fetal chromosomal aneuploidy by massively parallel genomic sequencing of DNA in maternal plasma. Proc Natl Acad Sci U S A 105: 20458-20463.
[8] Diehl F, Schmidt K, Choti MA, et al (2008). Circulating mutant DNA to assess tumor dynamics. Nat Med 14:985-990.
[9] Atamaniuk J, Kopecky C, Skoupy S, et al. (2012). Apoptotic cell-free DNA promotes inflammation in haemodialysis patients. Nephrol Dial Transplant 27:902-905.
[10] Fan HC, Blumenfeld YJ, Chitkara U, Hudgins L, Quake SR. (2010). Analysis of the size distributions of fetal and maternal cell-free DNA by paired-end sequencing. Clin Chem 56(8):1279-86
[11] Chandrananda D, Thorne NP, Ganesamoorthy D, Bruno DL, Benjamini Y, et al.. Investigating and Correcting Plasma DNA Sequencing Coverage Bias to Enhance Aneuploidy Discovery. PLoS ONE 9(1): e86993.
[12] Chandrananda D, Thorne NP, Bahlo M. (2015). High-resolution characterization of sequence signatures due to non-random cleavage of cell-free DNA. BMC Med Genomics 8:29.
[13] Ivanov et al. (2015). Non-random fragmentation patterns in circulating cell-free DNA reflect epigenetic regulation. BMC Genomics 16(Suppl 13):S1.

[14] Snyder MW, Kircher M, Hill AJ, Daza RM, Shendure J. (2016). Cell-free DNA Comprises an In Vivo Nucleosome Footprint that Informs Its Tissues-Of-Origin. Cell 164(1-2):57-68.
[15] TIANGEN Serum / Plasma Circulating DNA Kit (DP140113).
[16] TIANamp Blood DNA Kit Handbook.
[17] GeneRead DNA FFPE Handbook 03/2014
[18] Do, H., Wong, S.Q., Li, J., and Dobrovic, A., et al. (2013). Reducing sequence artifacts in amplicon-based massively parallel sequencing of formalin-fixed paraffin-embedded DNA by enzymatic depletion of uracil-containing templates. Clinical Chemistry 59:9, 18
[19] Digestion with NEBNext dsDNA Fragmentase (M0348) protocols
[20] KAPA LTP Library Preparation Kit Illumina Platforms KR0453 v5.16
[21] SeqCap EZ HyperCap Workflow for Target Enrichment data sheet v1.0
[22] Pedregosa F, Varoquaux G, Gramfort A, Michel V, Thirion B, Grisel O, Vanderplas J. (2011). Scikit-learn: Machine learning in Python. Journal of Machine Learning Research, 12(Oct), 2825-2830.
[23] John GH, Langley P. (1995). Estimating continuous distributions in Bayesian classifiers. InProceedings of the Eleventh conference on Uncertainty in artificial intelligence (pp. 338-345). Morgan Kaufmann Publishers Inc.
[24] Daz-Uriarte R, De Andres S. A. (2006). Gene selection and classification of microarray data using random forest. BMC bioinformatics, 7(1), 1.
[25] Lin C, Chen W, Qiu C, Wu Y, Krishnan S, Zou Q (2014). LibD3C: ensemble classifiers with a clustering and dynamic selection strategy. Neurocomputing, 123, 424-435.
[26] Efron B. (1979). Bootstrap Methods: Another Look at the Jackknife. Annals of Statistics, 7(1), 1-26.
[27] Efron B, Tibshirani R. (2012). Improvements on Cross-Validation: The 632+ Bootstrap Method. Journal of the American Statistical Association, 92(438), 548-560.
[28] Hyman DM, Solit DB, Arcila ME, Cheng DT, Sabbatini P, Baselga J, Berger MF, Ladanyi M. (2015). Precision medicine at Memorial Sloan Kettering Cancer Center: clinical next-generation sequencing enabling next-generation targeted therapy trials. Drug discovery today;20(12):1422-8.
[29] Jiang R, Lu YT, Ho H, Li B, Chen JF, Lin M, Li F, Wu K, Wu H, Lichterman J, Wan H. (2015). A comparison of isolated circulating tumor cells and tissue biopsies using whole-genome sequencing in prostate cancer. Oncotarget;6(42):44781.
[30] Wold S, Esbensen K and Geladi P, 1987. Principal component analysis. Chemometrics and intelligent laboratory systems, 2(1-3), pp.37-52.
[31] Lo YD, Chan KA, Sun H, Chen EZ, Jiang P, Lun FM, Zheng YW, Leung TY, Lau TK, Cantor CR, Chiu RW. Maternal plasma DNA sequencing reveals the genome-wide genetic and mutational profile of the fetus. Science translational medicine. 2010 Dec 8;2(61).