Sun et al. BMC Genomics 2013, 14(Suppl 2):S7
http://www.biomedcentral.com/1471-2164/14/S2/S7

BMC Genomics

RESEARCH Open Access

iSeeRNA: identification of long intergenic non-coding RNA transcripts from transcriptome sequencing data

Kun Sun[1,2], Xiaona Chen[1,3], Peiyong Jiang[1,2], Xiaofeng Song[4*], Huating Wang[1,3*], Hao Sun[1,2*]

From ISCB-Asia 2012
Shenzhen, China. 17-19 December 2012

Abstract

Background: Long intergenic non-coding RNAs (lincRNAs) are emerging as a novel class of non-coding RNAs and potent gene regulators. High-throughput RNA-sequencing combined with *de novo* assembly promises quantity discovery of novel transcripts. However, the identification of lincRNAs from thousands of assembled transcripts is still challenging due to the difficulties of separating them from protein coding transcripts (PCTs).

Results: We have implemented iSeeRNA, a support vector machine (SVM)-based classifier for the identification of lincRNAs. iSeeRNA shows better performance compared to other software. A public available webserver for iSeeRNA is also provided for small size dataset.

Conclusions: iSeeRNA demonstrates high prediction accuracy and runs several magnitudes faster than other similar programs. It can be integrated into the transcriptome data analysis pipelines or run as a web server, thus offering a valuable tool for lincRNA study.

Background

Over the past decade, evidence from numerous high-throughput genomic platforms reveals that even though less than 2% of the mammalian genome encodes proteins, a significant fraction can be transcribed into different complex families of non-coding RNAs (ncRNAs) [1-4]. Other than microRNAs and other families of small non-coding RNAs, long non-coding RNAs (lncRNAs, >200nt) are emerging as potent regulators of gene expression [5]. Originally identified by Guttman et al. [6] from four mouse cell types using chromatin state maps as a subtype of lncRNAs, long intergenic non-coding RNAs (lincRNAs), are discrete transcriptional unit intervening known protein-coding loci. Recent studies demonstrate the functional significance of lincRNAs. However, it remains a daunting task to identify all the lincRNAs existent in various biological processes and systems.

Whole transcriptome sequencing, known as RNA-Seq, offers the promise of rapid comprehensive discovery of novel genes and transcripts [7]. With the *de novo* assembly software such as Cufflinks [8] and Scripture [6], a large set of novel assemblies can be obtained from RNA-Seq data. Several programs have been used to facilitate the cataloging of lincRNAs from RNA-Seq assemblies. For example, Li et al. [9] used Codon Substitution Frequency (CSF) score [10] to identify lincRNAs from *de novo* assembled transcripts in chicken skeletal muscle. Pauli et al. [11] took advantage of PhyloCSF score [12] followed by other filtering steps to identify lincRNAs expressed during zebrafish embryogenesis. Cabili et al. [13] also used PhyloCSF program to eliminate the *de novo* assembled transcripts with positive coding potential and identified ~8200 lincRNA loci in 24 human tissues. However, the extremely high computational times demanded by PhyloCSF, may become the bottleneck for handling millions of assemblies generated from high throughput sequencing. Furthermore,

* Correspondence: xfsong@nuaa.edu.cn; huating.wang@cuhk.edu.hk; haosun@cuhk.edu.hk
[1]Li Ka Shing Institute of Health Sciences, The Chinese University of Hong Kong, Shatin, New Territories, Hong Kong SAR, China
[4]Department of Biomedical Engineering, Nanjing University of Aeronautics and Astronautics, Nanjing 210016, China
Full list of author information is available at the end of the article

The EMBO Journal (2013) 32, 2575–2588
www.embojournal.org

TRANSPARENT PROCESS

THE EMBO JOURNAL

Genome-wide survey by ChIP-seq reveals YY1 regulation of lincRNAs in skeletal myogenesis

Leina Lu[1,3], Kun Sun[1,3], Xiaona Chen[2], Yu Zhao[2], Lijun Wang[2], Liang Zhou[2], Hao Sun[1,*] and Huating Wang[2,*]

[1]Department of Chemical Pathology, Li Ka Shing Institute of Health Sciences, Prince of Wales Hospital, The Chinese University of Hong Kong, Hong Kong, China and [2]Department of Obstetrics and Gynaecology, Li Ka Shing Institute of Health Sciences, Prince of Wales Hospital, The Chinese University of Hong Kong, Hong Kong, China

Skeletal muscle differentiation is orchestrated by a network of transcription factors, epigenetic regulators, and non-coding RNAs. The transcription factor Yin Yang 1 (YY1) silences multiple target genes in myoblasts (MBs) by recruiting Ezh2 (Enhancer of Zeste Homologue2). To elucidate genome-wide YY1 binding in MBs, we performed chromatin immunoprecipitation (ChIP)-seq and found 1820 specific binding sites in MBs with a large portion residing in intergenic regions. Detailed analysis demonstrated that YY1 acts as an activator for many loci in addition to its known repressor function. No significant co-occupancy was found between YY1 and Ezh2, suggesting an additional Ezh2-independent function for YY1 in MBs. Further analysis of intergenic binding sites showed that YY1 potentially regulates dozens of large intergenic non-coding RNAs (lincRNAs), whose function in myogenesis is under-explored. We characterized a novel muscle-associated lincRNA (Yam-1) that is positively regulated by YY1. Yam-1 is downregulated upon differentiation and acts as an inhibitor of myogenesis. We demonstrated that Yam-1 functions through *in cis* regulation of miR-715, which in turn targets Wnt7b. Our findings not only provide the first genome-wide picture of YY1 association in muscle cells, but also uncover the functional role of lincRNA Yam-1.

The EMBO Journal (2013) **32,** 2575–2588. doi:10.1038/emboj.2013.182; Published online 13 August 2013
Subject Categories: chromatin & transcription; RNA; differentiation & death
Keywords: ChIP-seq; lincRNA; myogenesis; PRC2; YY1; Yam-1

Introduction

Normal skeletal muscle growth as well as the regeneration of damaged muscle fibres in post-embryonic life are attributed to satellite cells (muscle stem cells), which are characterized by the expression of paired-box transcription factor 7 (Pax7) and when activated, become immature muscle cells or myoblasts (MBs) that will proliferate and differentiate (Buckingham, 2006). The molecular characterization of the late stages of myogenesis has been studied in-depth in a mouse C2C12 MB cell line. The formation of mature muscle proceeds with the exit of MBs from the cell cycle, the expression of muscle-specific genes, and the suppression of genes that are specific to other cell lineages and tissues (Buckingham, 2006). A major portion of our understanding of myogenic differentiation is focused at the level of transcription, orchestrated by a complex network of muscle-specific transcription factors (TFs), including MyoD family (MyoD, Myf5, Myogenin, and MRF4), and MEF2 family (MEF2A-D). These factors coordinate with other transcriptional regulators in a stage-specific manner to activate the differentiation program by inducing or repressing gene transcriptions (Sabourin and Rudnicki, 2000). In addition to TFs, epigenetic regulators exert an important layer of transcriptional control (Perdiguero *et al*, 2009). The interplays between TFs and these regulators on muscle loci constitute an essential part of the regulatory networks.

Yin Yang 1 (YY1) is a multifunctional TF, which regulates various processes of development and differentiation (Gordon *et al*, 2006). It is highly expressed in proliferating MBs and gradually downregulated when differentiation starts, playing essential roles in the transcriptional regulation of myogenesis (Wang *et al*, 2007). Most of the previous work from our group and others identified YY1 as an epigenetic repressor of multiple muscle genes (Caretti *et al*, 2004; Wang *et al*, 2007, 2008) although YY1 is well known to play dual roles in either repressing or activating the transcription depending on the cofactors that it recruits (Deng *et al*, 2010). In proliferating MBs, activated by NF-κB signalling, YY1 functions to repress muscle loci by recruiting histone methyltransferase Ezh2 (Enhancer of Zeste Homolog2) containing Polycomb repressive complex 2 (PRC2) to target loci, which causes trimethylation of lysine 27 of histone 3 (H3K27me3) to silence transcription. When myogenesis ensues, YY1 and the repressive complex are replaced by MyoD/PCAF/SRF complex that activates gene expression. Identified YY1 repressive targets include not only muscle structural genes, such as Myosin heavy chain (MyHC), Troponin, and alpha skeletal actin (α-Actin), but also several muscle relevant miRNAs, miR-29, miR-1, miR-133, and miR-206 (Wang *et al*, 2008; Lu *et al*, 2012). The interplays between YY1 and muscle loci thus constitute essential components of myogenic network. Recently, we also showed that YY1 regulates a large number of additional miRNAs uncovered by genome-wide computational prediction (Lu *et al*, 2012). Deregulation of the YY1-miRNA circuitries could lead to aberrant myogenic differentiation, which contributes to pathogenesis of Rhabdomyosarcoma and Duchenne muscular dystrophy (Wang *et al*, 2008, 2012; Zhou *et al*, 2012a). Moreover, YY1/PRC2 was

*Corresponding authors. H Sun, Department of Chemical Pathology, Li Ka Shing Institute of Health Sciences, Prince of Wales Hospital, The Chinese University of Hong Kong, Hong Kong, NT, China. Tel.: +852 3763 6048; Fax: +852 2636 5090; E-mail: haosun@cuhk.edu.hk or H Wang, Department of Obstetrics and Gynaecology, The Chinese University of Hong Kong, Room 507A, Li Ka Shing Institute of Health Sciences, Prince of Wales Hospital, The Chinese University of Hong Kong, Hong Kong, NT, China. Tel.: +852 3763 6047; Fax: +852 2632 0008; E-mail: huating.wang@cuhk.edu.hk
[3]These authors contributed equally to this work.

Received: 15 November 2012; accepted: 9 July 2013; published online: 13 August 2013

OPEN ACCESS Freely available online

PLOS ONE

Sebnif: An Integrated Bioinformatics Pipeline for the Identification of Novel Large Intergenic Noncoding RNAs (lincRNAs) - Application in Human Skeletal Muscle Cells

Kun Sun[1], **Yu Zhao**[2], **Huating Wang**[2], **Hao Sun**[1]*

1 Department of Chemical Pathology, Li Ka Shing Institute of Health Sciences, Prince of Wales Hospital, The Chinese University of Hong Kong, Hong Kong SAR, China, **2** Department of Obstetrics and Gynaecology, Li Ka Shing Institute of Health Sciences, Prince of Wales Hospital, The Chinese University of Hong Kong, Hong Kong SAR, China

Abstract

Ab initio assembly of transcriptome sequencing data has been widely used to identify large intergenic non-coding RNAs (lincRNAs), a novel class of gene regulators involved in many biological processes. To differentiate real lincRNA transcripts from thousands of assembly artifacts, a series of filtering steps such as filters of transcript length, expression level and coding potential, need to be applied. However, an easy-to-use and publicly available bioinformatics pipeline that integrates these filters is not yet available. Hence, we implemented sebnif, an integrative bioinformatics pipeline to facilitate the discovery of *bona fide* novel lincRNAs that are suitable for further functional characterization. Specifically, sebnif is the only pipeline that implements an algorithm for identifying high-quality single-exonic lincRNAs that were often omitted in many studies. To demonstrate the usage of sebnif, we applied it on a real biological RNA-seq dataset from Human Skeletal Muscle Cells (HSkMC) and built a novel lincRNA catalog containing 917 highly reliable lincRNAs. Sebnif is available at http://sunlab.lihs.cuhk.edu.hk/sebnif/.

Citation: Sun K, Zhao Y, Wang H, Sun H (2014) Sebnif: An Integrated Bioinformatics Pipeline for the Identification of Novel Large Intergenic Noncoding RNAs (lincRNAs) - Application in Human Skeletal Muscle Cells. PLoS ONE 9(1): e84500. doi:10.1371/journal.pone.0084500

Editor: Yu Xue, Huazhong University of Science and Technology, China

Received September 5, 2013; **Accepted** November 15, 2013; **Published** January 6, 2014

Funding: The work described in this paper was substantially supported by the General Research Funds (GRF) to HW and HS from the Research Grants Council (RGC) of the Hong Kong Special Administrative Region, China (Project Code: CUHK 476310 to HW and CUHK 473211 to HS). The funders had no role in study design, data collection and analysis, decision to publish, or preparation of the manuscript.

Competing Interests: The authors have declared that no competing interests exist.

* E-mail: haosun@cuhk.edu.hk.

Introduction

Recent advances in transcriptome sequencing have led to the identification of many lincRNA transcripts (>200 nucleotides) [1,2,3] that localize in the intergenic region of protein coding genes (mRNAs). These transcripts have very weak or no coding potential for any protein products; their expression levels are generally lower than that of mRNAs thus are often mistakenly considered as transcriptional noises; many of them are transcribed by Polymerase II (Pol II) and spliced like mRNAs while a significant portion of them remain as single-exonic transcripts [3,4]. Emerging evidence suggests that lincRNAs are functional transcripts in various biological systems under different physiological and pathological conditions. In addition, the number of lincRNAs in mammalian species is estimated to be at least twice the number of mRNAs [5] with the majority of them are still undiscovered. Therefore, fervent efforts are being made in identifying novel lincRNAs in various biological systems.

Whole genome transcriptome sequencing, also known as RNA-seq, coupled with *ab initio* assembly has become an effective approach to discover novel lincRNAs [6]. To this end, RNAs are converted to cDNAs and subjected to high throughput sequencing; the obtained raw reads are then aligned to a reference genome and compared to known gene annotations to generate a list of novel transcripts. However, a high portion of the assembled transcripts are artifacts from genomic contamination or alignment bias, which could be falsely identified as novel lincRNAs. Therefore, the key issue is how to discriminate *bona fide* novel lincRNA transcripts from thousands of assembly artifacts. A widely used approach is to apply several filters, such as filters of transcript length, expression level and coding potential, to remove these artifacts step by step [1,7,8]. This multi-filtering approach has been proven effective in discovering thousands of novel multi-exonic lincRNAs in various systems [1,7,8,9]. But a large number of single-exonic transcripts were often discarded simply due to the lack of effective ways to discriminate them from thousands of the assembled artifacts. On the other hand, more and more studies have demonstrated that single-exonic lincRNAs are indeed functional. Well-characterized examples include MALAT1 [10], NEAT1 [11], Xist [12], HOTAIR [13] and Yam-1 [14]. Therefore, single-exonic transcripts should be considered as an important subclass in lincRNA families; and algorithms towards identification of unknown single-exonic lincRNA transcripts need to be developed. Furthermore, a bioinformatics pipeline, which integrates these filtering steps, is not yet publicly available. To fill these gaps, we designed and implemented an integrative bioinformatics pipeline named sebnif (Self-Estimation Based Novel LincRNA Filtering pipeline) to facilitate the identification of both multi- and single-exonic lincRNAs. To illustrate its usage and performance, we applied it on a RNA-seq dataset from Human Skeletal Muscle Cells (HSkMC) to build a lincRNA catalog. Further analysis of these

Genomics Data 2 (2014) 89–91

Contents lists available at ScienceDirect

Genomics Data

journal homepage: http://www.journals.elsevier.com/genomics-data/

Data in Brief

Genome-wide profiling of YY1 binding sites during skeletal myogenesis

Kun Sun [a], Leina Lu [b], Huating Wang [b], Hao Sun [a,*]

[a] Department of Chemical Pathology, Li Ka Shing Institute of Health Sciences, The Chinese University of Hong Kong, Shatin, New Territories, Hong Kong SAR, China
[b] Department of Obstetrics and Gynaecology, Li Ka Shing Institute of Health Sciences, The Chinese University of Hong Kong, Shatin, New Territories, Hong Kong SAR, China

ARTICLE INFO

Article history:
Received 28 April 2014
Received in revised form 14 May 2014
Accepted 15 May 2014
Available online 23 May 2014

Keywords:
YY1
ChIP-seq
Myogenesis
LincRNA

ABSTRACT

Skeletal muscle differentiation is regulated by a network of transcription factors, epigenetic regulators and non-coding RNAs. We have recently performed ChIP-seq experiments to explore the genome-wide binding of transcription factor YY1 in skeletal muscle cells. Our results identified thousands of YY1 binding peaks, underscoring its multifaceted functions in muscle cells. In particular, we identified a very high proportion of YY1 binding peaks residing in the intergenic regions, which led to the discovery of some novel lincRNAs under YY1 regulation. Here we describe the details of the ChIP-seq experiments and data analysis procedures associated with the study published by Lu et al. in the EMBO Journal in 2013 [1].

Specifications	
Organism/cell line/tissue	*Mus musculus/C2C12*
Sex	*NA*
Sequencer or array type	*Illumina Hiseq 2000, Illumina GA IIx*
Data format	*Raw data: FASTQ files* *Processed data: BEDGRAPH, TXT*
Experimental factors	*Myoblast vs myotube*
Experimental features	*Using ChIP-seq, we generated genome-wide maps of YY1 in skeletal myoblasts and myotubes with biological replicates. We found that a large proportion of the binding sites reside in the intergenic regions; therefore, many lincRNAs are regulated by YY1.*
Consent	*NA*
Sample source location	*Manassas, VA, USA*

Direct link to deposited data

http://www.ncbi.nlm.nih.gov/geo/query/acc.cgi?acc=GSE45875

Experimental Design, Materials and Methods

Cell culture

Mouse C2C12 myoblast cell line was purchased from American Type Culture Collection (ATCC). The myoblasts were maintained in a growth medium (DMEM, 10%FBS and 1% penicillin/streptomycin), and induced to myotubes by culturing in a differentiation medium (DMEM, 2% horse serum and 1% penicillin/streptomycin).

ChIP assays and sequencing experiments

ChIP assays were performed as previously described [2,3]. About 2×10^7 C2C12 cells and 5 μg of antibodies were used in one immunoprecipitation. The antibodies include YY1 #1 (Santa Cruz Biotechnology, Cat# SC-1703, rabbit polyclonal), YY1 #2 (Abcam, Cat# AB58066, mouse monoclonal), Ezh2 (Cell Signaling, MA, USA, Cat# AC22), trimethyl-histone H3-K27 (Millipore, Cat# 07-449), trimethyl-histone H3-K4 (Millipore, Cat# 07-473), or normal mouse IgG (Santa Cruz Biotechnology, Cat# SC-2025) as a negative control.

For library construction, we used a protocol as described before [4]. Briefly, the immunoprecipitated DNA (~10 ng) were end-repaired, and A-nucleotide overhangs were then added, followed by adapter ligation, PCR enrichment, size selection and purification. The purified DNA library products were evaluated using Bioanalyzer (Agilent) and SYBR qPCR and diluted to 10 nM for sequencing on Illumina Hi-seq 2000 sequencer (YY1) (pair-end with 50 bp) or Illumina Genome Analyzer II sequencer (Ezh2, H3K27me3 and H3K4me3) (pair-end with 36 bp). Technical replicates were prepared by sequencing the same library twice. A data analysis pipeline CASAVA 1.8 (Illumina) was employed to perform the initial bioinformatic analysis (base calling). Table 1 lists all the experiments that we had performed. For MB YY1, we performed two biological replicates with the antibody SC-1703 and a third biological replicate with a second antibody AB58066. We also performed two technical replicates for each antibody (run 1 and run 2).

* Corresponding author at: Department of Chemical Pathology, The Chinese University of Hong Kong, Room 503A, Li Ka Shing Institute of Health Sciences, Prince of Wales Hospital, The Chinese University of Hong Kong, Shatin, New Territories, Hong Kong SAR, China. Tel.: +852 3763 6048; fax: +852 3763 6033.
E-mail address: haosun@cuhk.edu.hk (H. Sun).

http://dx.doi.org/10.1016/j.gdata.2014.05.008

Methods 69 (2014) 230–236

Contents lists available at ScienceDirect

Methods

journal homepage: www.elsevier.com/locate/ymeth

Online Diagnosis System: A webserver for analysis of Sanger sequencing-based genetic testing data

Kun Sun [a,b], Yuet-Ping Yuen [a], Huating Wang [b,c], Hao Sun [a,b,*]

[a] Department of Chemical Pathology, Prince of Wales Hospital, The Chinese University of Hong Kong, Hong Kong Special Administrative Region
[b] Li Ka Shing Institute of Health Sciences, Prince of Wales Hospital, The Chinese University of Hong Kong, Hong Kong Special Administrative Region
[c] Department of Obstetrics and Gynaecology, Prince of Wales Hospital, The Chinese University of Hong Kong, Hong Kong Special Administrative Region

ARTICLE INFO

Article history:
Received 27 March 2014
Revised 25 June 2014
Accepted 5 July 2014
Available online 22 July 2014

Keywords:
Genetic testing
Sanger sequencing
Small nucleotide variation

ABSTRACT

Sanger sequencing is a well-established molecular technique for diagnosis of genetic diseases. In these tests, DNA sequencers produce vast amounts of data that need to be examined and annotated within a short period of time. To achieve this goal, an online bioinformatics platform that can automate the process is essential. However, to date, there is no such integrated bioinformatics platform available. To fulfill this gap, we developed the Online Diagnosis System (ODS), which is a freely available webserver and supports the commonly used file format of Sanger sequencing data. ODS seamlessly integrates base calling, single nucleotide variation (SNV) identification, and SNV annotation into one single platform. It also allows laboratorians to manually inspect the quality of the identified SNVs in the final report. ODS can significantly reduce the data analysis time therefore allows Sanger sequencing-based genetic testing to be finished in a timely manner. ODS is freely available at http://sunlab.lihs.cuhk.edu.hk/ODS/.

1. Introduction

Molecular genetic testing studies single genes or short length DNA sequences to identify the single nucleotide variations (SNVs) or mutations that lead to a genetic disorder. The results of a genetic test can confirm or rule out a suspected genetic condition. Currently, more than 2000 genetic tests are in clinical use, and more are being developed. Detailed information can be found at the National Institutes of Health (NIH) website (http://www.nlm.nih.gov/medlineplus/genetictesting.html).

Sanger sequencing is one of the DNA sequencing method by selectively incorporating chain-terminating dideoxynucleotides by DNA polymerase during in vitro DNA replication [1–3]. Sanger sequencing-based genetic testing has been widely used in genetic testing laboratories. The analysis of Sanger sequencing data for genetic testing typically involves several steps: (*i*) processing of the raw sequencing data; (*ii*) identifying SNVs; and (*iii*) annotating the identified SNVs to facilitate the evaluation of their effects on tested gene.

The data analysis process is usually the bottleneck of the entire testing process because it involves manual inspection of raw data (i.e., electropherograms), SNV identification, and extracting annotations for each identified SNV from different sources. Therefore, reducing the time needed to interpret the results of each sequencing tracing is critical for a practical genetic test. Seamlessly integration and automation of the Sanger sequencing data analysis, including raw data analysis, quality control, file handling, variant identification, and annotation, will largely improve the turnaround time for the genetic testing thus ultimately reduce the costs of those tests. To this end, we designed and implemented the Online Diagnosis System (ODS), a freely accessible webserver for analyzing of Sanger sequencing data for speeding up the data analysis and interpretation process. We demonstrated the usage and the performance of ODS on two disease diagnosis scenarios and the results indicate that this webserver can be used to streamline analysis of clinical Sanger sequencing data with the ability to accurately interrogate electropherograms for base calling, identify SNVs, and provide comprehensive SNV annotations, therefore facilitate the clinician to finish the genetic testing in a timely manner.

2. Methods

2.1. Workflow of ODS

ODS is an integrated, easy-to-use bioinformatics platform for Sanger sequencing-based genetic testing. It consists of the following three major components (Fig. 1):

* Corresponding author at: Department of Chemical Pathology, Li Ka Shing Institute of Health Sciences, Prince of Wales Hospital, The Chinese University of Hong Kong, Hong Kong Special Administrative Region. Fax: +852 37636033.
E-mail address: haosun@cuhk.edu.hk (H. Sun).

http://dx.doi.org/10.1016/j.ymeth.2014.07.002
1046-2023/

Plasma DNA tissue mapping by genome-wide methylation sequencing for noninvasive prenatal, cancer, and transplantation assessments

Kun Sun[a,b,1], Peiyong Jiang[a,b,1], K. C. Allen Chan[a,b,c,1], John Wong[d], Yvonne K. Y. Cheng[e], Raymond H. S. Liang[f], Wai-kong Chan[g], Edmond S. K. Ma[g], Stephen L. Chan[h], Suk Hang Cheng[a,b], Rebecca W. Y. Chan[a,b], Yu K. Tong[a,b], Simon S. M. Ng[d], Raymond S. M. Wong[i,j], David S. C. Hui[i], Tse Ngong Leung[k], Tak Y. Leung[e], Paul B. S. Lai[c,d], Rossa W. K. Chiu[a,b], and Yuk Ming Dennis Lo[a,b,c,2]

[a]Li Ka Shing Institute of Health Sciences, The Chinese University of Hong Kong, Shatin, New Territories, Hong Kong SAR, China; [b]Department of Chemical Pathology, The Chinese University of Hong Kong, Prince of Wales Hospital, Shatin, New Territories, Hong Kong SAR, China; [c]State Key Laboratory in Oncology in South China, The Chinese University of Hong Kong, Prince of Wales Hospital, Shatin, New Territories, Hong Kong SAR, China; [d]Department of Surgery, The Chinese University of Hong Kong, Prince of Wales Hospital, Shatin, New Territories, Hong Kong SAR, China; [e]Department of Obstetrics and Gynaecology, The Chinese University of Hong Kong, Prince of Wales Hospital, Shatin, New Territories, Hong Kong SAR, China; [f]Comprehensive Oncology Centre, Hong Kong Sanatorium & Hospital, Hong Kong SAR, China; [g]Department of Pathology, Hong Kong Sanatorium & Hospital, Hong Kong SAR, China; [h]Department of Clinical Oncology, The Chinese University of Hong Kong, Prince of Wales Hospital, Shatin, New Territories, Hong Kong SAR, China; [i]Department of Medicine and Therapeutics, The Chinese University of Hong Kong, Prince of Wales Hospital, Shatin, New Territories, Hong Kong SAR, China; [j]Sir Y.K. Pao Centre for Cancer, The Chinese University of Hong Kong, Hong Kong SAR, China; and [k]Obstetrics and Gynaecology Centre, Hong Kong Sanatorium & Hospital, Hong Kong SAR, China

Contributed by Yuk Ming Dennis Lo, May 4, 2015 (sent for review April 24, 2015; reviewed by Frederik Banch Clausen)

Plasma consists of DNA released from multiple tissues within the body. Using genome-wide bisulfite sequencing of plasma DNA and deconvolution of the sequencing data with reference to methylation profiles of different tissues, we developed a general approach for studying the major tissue contributors to the circulating DNA pool. We tested this method in pregnant women, patients with hepatocellular carcinoma, and subjects following bone marrow and liver transplantation. In most subjects, white blood cells were the predominant contributors to the circulating DNA pool. The placental contributions in the plasma of pregnant women correlated with the proportional contributions as revealed by fetal-specific genetic markers. The graft-derived contributions to the plasma in the transplant recipients correlated with those determined using donor-specific genetic markers. Patients with hepatocellular carcinoma showed elevated plasma DNA contributions from the liver, which correlated with measurements made using tumor-associated copy number aberrations. In hepatocellular carcinoma patients and in pregnant women exhibiting copy number aberrations in plasma, comparison of methylation deconvolution results using genomic regions with different copy number status pinpointed the tissue type responsible for the aberrations. In a pregnant woman diagnosed as having follicular lymphoma during pregnancy, methylation deconvolution indicated a grossly elevated contribution from B cells into the plasma DNA pool and localized B cells as the origin of the copy number aberrations observed in plasma. This method may serve as a powerful tool for assessing a wide range of physiological and pathological conditions based on the identification of perturbed proportional contributions of different tissues into plasma.

noninvasive prenatal testing | circulating tumor DNA | liquid biopsy | transplantation monitoring | epigenetics

There is much recent interest in the diagnostic applications of cell-free DNA in plasma. Cell-free fetal DNA has been found in the plasma of pregnant women (1). Its detection has made noninvasive prenatal testing, most notably for chromosomal aneuploidies, a clinical reality (2–7). Tumor-derived DNA has been found in the plasma of cancer patients (8–12), offering the possibility of performing "liquid biopsy" for cancer assessment and monitoring. Following organ transplantation, donor-derived DNA from the transplanted organs has been detected in the plasma of the recipients (13) and has been used for monitoring graft rejection (14).

Plasma DNA is generally regarded as consisting of a mixture of DNA released from cells from different tissues of the body. Through the analysis of genetic differences between the minor and major background circulating DNA species, researchers have shown that a number of bodily organs made contributions to the plasma DNA pool. For example, studies on pregnant cases in which the fetus and placenta exhibit different karyotypes have demonstrated that the placenta is the origin of the cell-free fetal DNA detectable in the maternal circulation (15, 16). The detection of tumor-associated genetic alterations has allowed the detection of tumor DNA originating from cancer at different body organs in plasma (17). The detection of donor-derived genetic signatures in the plasma of patients following bone

Significance

Plasma consists of DNA released from multiple tissues within the body. Using genome-wide bisulfite sequencing of plasma DNA, we obtained a bird's eye view of the identities and contributions of these tissues to the circulating DNA pool. The tissue contributors and their relative proportions are identified by a bioinformatics deconvolution process that draws reference from DNA methylation signatures representative of each tissue type. We validated this approach in pregnant women, cancer patients, and transplant recipients. This method also allows one to identify the tissue of origin of genomic aberrations observed in plasma DNA. This approach has numerous research and diagnostic applications in prenatal testing, oncology, transplantation monitoring, and other fields.

Author contributions: K.C.A.C., R.W.K.C., and Y.M.D.L. designed research; K.S., P.J., K.C.A.C., S.H.C., R.W.Y.C., and Y.K.T. performed research; K.S., P.J., K.C.A.C., R.W.K.C., and Y.M.D.L. analyzed data; K.S., P.J., K.C.A.C., R.W.K.C., and Y.M.D.L. wrote the paper; J.W., Y.K.Y.C., R.H.S.L., S.L.C., S.S.M.N., R.S.M.W., D.S.C.H., T.N.L., T.Y.L., and P.B.S.L. recruited patients and analyzed clinical data; and W.-k.C. and E.S.K.M. provided specimens and analyzed clinical data.

Reviewers included: F.B.C., Copenhagen University Hospital.

The authors declare no conflict of interest.

Freely available online through the PNAS open access option.

Data deposition: Sequence data for the subjects studied in this work who consented to data archiving have been deposited in the European Genome-Phenome Archive (EGA), www.ebi.ac.uk/ega/, hosted by the European Bioinformatics Institute (accession no. EGAS00001001219).

[1]K.S., P.J., and K.C.A.C. contributed equally to this work.

[2]To whom correspondence should be addressed. Email: loym@cuhk.edu.hk.

This article contains supporting information online at www.pnas.org/lookup/suppl/doi:10.1073/pnas.1508736112/-/DCSupplemental.

EXPERT REVIEWS

The impact of digital DNA counting technologies on noninvasive prenatal testing

Expert Rev. Mol. Diagn. 15(10), 1261–1268 (2015)

Kun Sun, Peiyong Jiang and KC Allen Chan*

Department of Chemical Pathology, The Chinese University of Hong Kong, Prince of Wales Hospital, Shatin, New Territories, Hong Kong SAR, China
**Author for correspondence:*
Tel.: +852 26 32 35 89
Fax: +852 26 36 50 90
allen@cuhk.edu.hk

The discovery of cell-free DNA molecules in maternal plasma has opened up numerous opportunities for noninvasive prenatal testing. The advent of new digital counting technologies, including digital polymerase chain reaction and massive parallel sequencing, has provided the opportunity to quantify the cell-free DNA molecules in maternal plasma in an unprecedentedly precise manner. Powered by these technologies, prenatal testing of different kinds of hereditary conditions, ranging from monogenic diseases to chromosome aneuploidies, has been shown to be possible through the analysis of maternal plasma DNA. Discussed here are the principles of the approaches used in the noninvasive testing of different fetal conditions, with an emphasis on the impact that different digital DNA counting strategies have made on the development of these tests.

Keywords: cell-free fetal DNA • digital PCR • massive parallel sequencing • maternal plasma • NIPT

Prenatal diagnosis is an essential part of modern obstetric care. Conventional methods for obtaining fetal genetic materials for prenatal diagnosis, including amniocentesis and chorionic villus sampling (CVS), carry risk of fetal loss. In 1997, Lo *et al.* discovered that fetal DNA is present in the cell-free plasma of pregnant women by demonstrating that chromosome Y sequences could be detected in the plasma of pregnant women carrying male fetus using polymerase chain reaction (PCR) but not detectable in the plasma of those carrying female fetus [1]. This discovery has provided the scientific foundation of noninvasive prenatal tests (NIPT) based on the analysis of fetal DNA in maternal plasma.

The early applications of the analysis of fetal DNA in maternal plasma focused on the qualitative detection of paternal-specific sequences, sequences that are present in the father's but not the mother's genome. The presence of such sequences in the maternal plasma implies that the fetus has inherited the sequence from the father. One important early clinical application based on this strategy is the prenatal detection of rhesus D blood group incompatibility. The detection of the RHD gene sequences in a rhesus D-negative pregnant women is used to indicate a rhesus D-positive fetus [2]. Noninvasive prenatal rhesus D testing is possible at as early as 5 weeks of gestation, and the overall accuracy has been reported to range from 95.7 to 99.8% [3–6].

To date, the qualitative analysis for paternal sequences has been applied for the noninvasive prenatal testing of a wide spectrum of other paternally inherited autosomal dominant conditions, for example Huntington's disease, myotonic dystrophy and achondroplasia [7–9]. However, the same qualitative approach fails to provide a definitive answer for conditions with other inheritance patterns, including maternally inherited autosomal dominant conditions, autosomal recessive conditions and chromosomal aneuploidies. In these conditions, the target sequences, namely the maternally derived mutations or the aneuploid chromosomes, would be present in the maternal plasma regardless of whether the fetus is affected or not. More complex algorithms require precise quantification of different targets in maternal plasma to detect these conditions. The accurate quantification of these targets is made possible with the recent advances in digital DNA counting technologies. In this review, the different diagnostic approaches and the digital DNA counting technologies that enable these analyses are discussed.

 10.1586/14737159.2015.1084227 ISSN 1473-7159

nature COMMUNICATIONS

ARTICLE

Received 15 Apr 2015 | Accepted 28 Oct 2015 | Published 11 Dec 2015

DOI: 10.1038/ncomms10026

Linc-YY1 promotes myogenic differentiation and muscle regeneration through an interaction with the transcription factor YY1

Liang Zhou[1,*,†], Kun Sun[1,2,*], Yu Zhao[1,3,*], Suyang Zhang[1,2], Xuecong Wang[4], Yuying Li[1,2], Leina Lu[1], Xiaona Chen[1], Fengyuan Chen[1,2], Xichen Bao[5], Xihua Zhu[5], Lijun Wang[1], Ling-Yin Tang[3], Miguel A. Esteban[5], Chi-Chiu Wang[3], Ralf Jauch[4], Hao Sun[1,2] & Huating Wang[1,6]

Little is known how lincRNAs are involved in skeletal myogenesis. Here we describe the discovery of *Linc-YY1* from the promoter of the transcription factor (TF) Yin Yang 1 (*YY1*) gene. We demonstrate that *Linc-YY1* is dynamically regulated during myogenesis *in vitro* and *in vivo*. Gain or loss of function of *Linc-YY1* in C2C12 myoblasts or muscle satellite cells alters myogenic differentiation and in injured muscles has an impact on the course of regeneration. *Linc-YY1* interacts with YY1 through its middle domain, to evict YY1/Polycomb repressive complex (PRC2) from target promoters, thus activating the gene expression *in trans*. In addition, *Linc-YY1* also regulates PRC2-independent function of YY1. Finally, we identify a human *Linc-YY1* orthologue with conserved function and show that many human and mouse TF genes are associated with lincRNAs that may modulate their activity. Altogether, we show that *Linc-YY1* regulates skeletal myogenesis and uncover a previously unappreciated mechanism of gene regulation by lincRNA.

[1] Li Ka Shing Institute of Health Sciences, The Chinese University of Hong Kong, Hong Kong, China. [2] Department of Chemical Pathology, Prince of Wales Hospital, Li Ka Shing Institute of Health Sciences, Chinese University of Hong Kong, Hong Kong, China. [3] Department of Obstetrics and Gynaecology, Prince of Wales Hospital, The Chinese University of Hong Kong, Hong Kong, China. [4] Genome Regulation Laboratory, Drug Discovery Pipeline, Key Laboratory of Regenerative Biology, Guangdong Provincial Key Laboratory of Stem Cell and Regenerative Medicine, South China Institute for Stem Cell Biology and Regenerative Medicine, Guangzhou Institutes of Biomedicine and Health, Chinese Academy of Sciences, Guangzhou, Guangdong 510530, China. [5] Laboratory of Chromatin and Human Disease, Key Laboratory of Regenerative Biology, South China Institute for Stem Cell Biology and Regenerative Medicine, Guangzhou Institutes of Biomedicine and Health, Chinese Academy of Sciences, Guangzhou, Guangdong 510530, China. [6] Department of Orthopedics and Traumatology, Prince of Wales Hospital, Li Ka Shing Institute of Health Sciences, The Chinese University of Hong Kong, Hong Kong, China. * These authors contributed equally to this work. † Present address: Department of Radiation Medicine, Guangdong Provincial Key Laboratory of Tropical Disease Research, School of Public Health and Tropical Medicine, Southern Medical University, Guangzhou, Guangdong 510515, PR China. Correspondence and requests for materials should be addressed to H.W. (email: huating.wang@cuhk.edu.hk) or H.S. (email: haosun@cuhk.edu.hk).

Genomics Data 7 (2016) 247–249

Contents lists available at ScienceDirect

Genomics Data

journal homepage: www.elsevier.com/locate/gdata

Data in Brief

Genome-wide RNA-seq and ChIP-seq reveal Linc-YY1 function in regulating YY1/PRC2 activity during skeletal myogenesis

Kun Sun [a,b], Liang Zhou [b], Yu Zhao [b], Huating Wang [b,c,*], Hao Sun [a,b,*]

[a] Department of Chemical Pathology, The Chinese University of Hong Kong, Hong Kong, China
[b] Li Ka Shing Institute of Health Sciences, The Chinese University of Hong Kong, Hong Kong, China
[c] Department of Orthopaedics and Traumatology, The Chinese University of Hong Kong, Hong Kong, China

ARTICLE INFO

Article history:
Received 19 January 2016
Received in revised form 30 January 2016
Accepted 31 January 2016
Available online 2 February 2016

Keywords:
Linc-YY1
Myogenesis
RNA-seq
ChIP-seq
C2C12 cells

ABSTRACT

Little is known how lincRNAs are involved in skeletal myogenesis. Here we describe the discovery and functional annotation of Linc-YY1, a novel lincRNA originating from the promoter of the transcription factor (TF) Yin Yang 1 (YY1). Starting from whole transcriptome shotgun sequencing (a.k.a. RNA-seq) data from muscle C2C12 cells, a series of bioinformatics analysis was applied towards the identification of hundreds of high-confidence novel lincRNAs. Genome-wide approaches were then employed to demonstrate that Linc-YY1 functions to promote myogenesis through associating with YY1 and regulating YY1/PRC2 transcriptional activity *in trans*. Here we describe the details of the ChIP-seq, RNA-seq experiments, and data analysis procedures associated with the study published by Zhou and colleagues in the Nature Communications Journal in 2015 Zhou et al. (2015) [1]. The data was deposited on NCBI's Gene Expression Omnibus (GEO, http://www.ncbi.nlm.nih.gov/geo/) with accession number GSE74049.

Specifications	
Organism/cell line/tissue	*Mus musculus*/C2C12
Sex	NA
Sequencer or array type	Illumina HiSeq 1500, Illumina Hiseq 2000, Illumina GA IIx
Data format	Raw data: FASTQ files Processed data: BEDGRAPH, TXT
Experimental factors	Myoblast vs myotube
Experimental features	Using RNA-seq, we identified hundreds of high confidence novel lincRNAs in skeletal muscle. Among these lincRNAs, one near the YY1 transcription factor caused our attention, which we named Linc-YY1. We then used ChIP-seq and RNA-seq to investigate the function of this novel lincRNA during myogenesis.
Consent	NA
Sample source location	Manassas, VA, USA

1. Direct link to deposited data

http://www.ncbi.nlm.nih.gov/geo/query/acc.cgi?acc=GSE74049.

* Corresponding authors at: Li Ka Shing Institute of Health Sciences, The Chinese University of Hong Kong, Hong Kong, China.
E-mail addresses: huating.wang@cuhk.edu.hk (H. Wang), haosun@cuhk.edu.hk (H. Sun).

2. Experimental design, materials, and methods

2.1. Cell culture

Mouse C2C12 myoblasts cell line was purchased from American Type Culture Collection (ATCC). The myoblasts were maintain in growth medium (DMEM, 10%FBS and 1% Penicillin/Streptomycin), and induced to myotubes by culturing in differentiation medium (DMEM, 2% horse serum and 1% Penicillin/Streptomycin).

2.2. ChIP assays and sequencing experiments

ChIP assays were performed as previously described [2,3]. About 2×10^7 C2C12 cells and 5 μg of antibodies were used in one immunoprecipitation. The antibodies include YY1 (Santa Cruz Biotechnology), Ezh2 (Active Motif), Eed (Millipore), trimethyl-histone H3-K27 (Millipore), or normal mouse IgG (Santa Cruz Biotechnology) as a negative control.

For library construction, we used a protocol as described before [4]. Briefly, the immunoprecipitated DNA (~10 ng) were end-repaired, and A-nucleotide overhangs were then added, followed by adapter ligation, PCR enrichment, size selection and purification. The purified DNA library products were evaluated using Bioanalyzer (Agilent) and SYBR qPCR and diluted to 10 nM for sequencing on Illumina Hi-seq 2000 sequencer (YY1) (pair-end with 50 bp) or Illumina Genome Analyzer II sequencer (Ezh2, Eed and H3K27me3) (pair-end with 36 bp).

http://dx.doi.org/10.1016/j.gdata.2016.01.022

Second generation noninvasive fetal genome analysis reveals de novo mutations, single-base parental inheritance, and preferred DNA ends

K. C. Allen Chan[a,b,1], Peiyong Jiang[a,b,1], Kun Sun[a,b,1], Yvonne K. Y. Cheng[c], Yu K. Tong[a,b], Suk Hang Cheng[a,b], Ada I. C. Wong[a,b], Irena Hudecova[a,b], Tak Y. Leung[c], Rossa W. K. Chiu[a,b,2], and Yuk Ming Dennis Lo[a,b,2]

[a]Li Ka Shing Institute of Health Sciences, The Chinese University of Hong Kong, Shatin, New Territories, Hong Kong; [b]Department of Chemical Pathology, The Chinese University of Hong Kong, Prince of Wales Hospital, Shatin, New Territories, Hong Kong; and [c]Department of Obstetrics and Gynaecology, The Chinese University of Hong Kong, Prince of Wales Hospital, Shatin, New Territories, Hong Kong

Contributed by Yuk Ming Dennis Lo, October 4, 2016 (sent for review August 24, 2016; reviewed by Mary E. Norton and Erik A. Sistermans)

Plasma DNA obtained from a pregnant woman was sequenced to a depth of 270× haploid genome coverage. Comparing the maternal plasma DNA sequencing data with the parental genomic DNA data and using a series of bioinformatics filters, fetal de novo mutations were detected at a sensitivity of 85% and a positive predictive value of 74%. These results represent a 169-fold improvement in the positive predictive value over previous attempts. Improvements in the interpretation of the sequence information of every base position in the genome allowed us to interrogate the maternal inheritance of the fetus for 618,271 of 656,676 (94.2%) heterozygous SNPs within the maternal genome. The fetal genotype at each of these sites was deduced individually, unlike previously, where the inheritance was determined for a collection of sites within a haplotype. These results represent a 90-fold enhancement in the resolution in determining the fetus's maternal inheritance. Selected genomic locations were more likely to be found at the ends of plasma DNA molecules. We found that a subset of such preferred ends exhibited selectivity for fetal- or maternal-derived DNA in maternal plasma. The ratio of the number of maternal plasma DNA molecules with fetal preferred ends to those with maternal preferred ends showed a correlation with the fetal DNA fraction. Finally, this second generation approach for noninvasive fetal whole-genome analysis was validated in a pregnancy diagnosed with cardiofaciocutaneous syndrome with maternal plasma DNA sequenced to 195× coverage. The causative de novo *BRAF* mutation was successfully detected through the maternal plasma DNA analysis.

noninvasive prenatal testing | massively parallel sequencing | DNA fragmentation patterns

The discovery of cell-free fetal DNA in maternal plasma has enabled the development of noninvasive prenatal testing (NIPT) (1). Over the last few years, NIPT has been implemented globally for the noninvasive prenatal investigation of fetal chromosomal aneuploidies (2–5). With higher depth of sequencing and improved bioinformatics analyses, NIPT has now been extended to the detection of a variety of subchromosomal aberrations (6, 7). Further expanding the applications of NIPT, we showed in 2010 that it was possible to deduce the fetal genome by deep sequencing of maternal plasma (8). Work by Fan et al. (9) and Kitzman et al. (10) confirmed these results. In these previous efforts, the depths of maternal plasma DNA sequencing ranged from 52.7× to 78× haploid human genome coverages (8–10).

There are a number of limitations in these previous studies. For example, Kitzman et al. (10) explored the possibility of detecting fetal de novo mutations on a genome-wide level from the maternal plasma DNA sequencing data. In one variation of bioinformatics analysis, they found 2.5×10^7 candidate fetal de novo mutation sites in the plasma DNA sequencing data. Only 39 of these were true fetal de novo mutations. Because the studied fetus had a total of 44 de novo mutations, the positive predictive value (PPV) was 0.000156%, and the sensitivity was 88.6%. With additional refinement in bioinformatics analysis, Kitzman et al. (10) improved the PPV to 0.438%, although the sensitivity was reduced to 38.6%. These data, thus, indicate the enormous challenge of detecting fetal de novo mutations on a genome-wide scale using NIPT. In particular, dramatic improvement in the PPV would be needed for such an approach to be clinically practical.

A second area that needs improvement concerns the detection of sequences that the fetus has inherited from its mother. Previous efforts in elucidating the maternal inheritance of the fetus on a genome-wide scale have generally used a haplotype-based strategy, which has been referred to as the relative haplotype dosage (RHDO) approach (8–10). Hence, for a pregnant woman who has two haplotypes in a particular chromosomal region, she would pass one of these onto her fetus. Because her plasma contains a

Significance

We explored the limit of noninvasive prenatal testing by performing genome-wide sequencing of maternal plasma DNA at 195× and 270× haploid genome coverages. Combined with the use of a series of bioinformatics filters, fetal de novo mutations could be detected with a positive predictive value that was two orders of magnitude higher than previously reported. A de novo *BRAF* mutation was noninvasively detected in a case with cardiofaciocutaneous syndrome. The maternal inheritance of the fetus could be ascertained on a genome-wide level without the use of maternal haplotypes, hence greatly increasing the resolution of such analysis. Finally, we showed that certain genomic locations were overrepresented at the ends of plasma DNA fragments with fetal or maternal selectivity.

Author contributions: K.C.A.C., R.W.K.C., and Y.M.D.L. designed research; K.C.A.C., P.J., K.S., Y.K.Y.C., Y.K.T., S.H.C., A.I.C.W., I.H., and T.Y.L. performed research; K.C.A.C., P.J., K.S., R.W.K.C., and Y.M.D.L. analyzed data; Y.K.Y.C. and T.Y.L. recruited subjects and analyzed clinical data; and K.C.A.C., P.J., R.W.K.C., and Y.M.D.L. wrote the paper.

Reviewers: M.E.N., University of California, San Francisco; and E.A.S., VU University Medical Center.

Conflict of interest statement: R.W.K.C. and Y.M.D.L. received research support from Sequenom, Inc. R.W.K.C. and Y.M.D.L. were consultants to Sequenom, Inc. K.C.A.C., R.W.K.C., and Y.M.D.L. hold equities in Sequenom, Inc. K.C.A.C., R.W.K.C., and Y.M.D.L. are founders of Xcelom and Cirina. K.C.A.C., P.J., and R.W.K.C. are consultants to Xcelom. P.J. is a consultant to Cirina. K.C.A.C., P.J., R.W.K.C., and Y.M.D.L. have filed patent applications (PCT/CN2016/073753 and PCT/CN2016/091531) based on the data generated from this work, which have been licensed to Cirina.

Freely available online through the PNAS open access option.

Data deposition: The sequence data for the subjects studied in this work who had consented to data archiving have been deposited in the European Genome-Phenome Archive (EGA), https://www.ebi.ac.uk/ega/, hosted by the European Bioinformatics Institute (EBI; accession no. EGAS00001001882).

See Commentary on page 14173.

[1]K.C.A.C., P.J., and K.S. contributed equally to this work.

[2]To whom correspondence may be addressed. Email: rossachiu@cuhk.edu.hk or loym@cuhk.edu.hk.

This article contains supporting information online at www.pnas.org/lookup/suppl/doi:10.1073/pnas.1615800113/-/DCSupplemental.

ORIGINAL ARTICLE

COFFEE: control-free noninvasive fetal chromosomal examination using maternal plasma DNA

Kun Sun[1,2], K. C. Allen Chan[1,2], Irena Hudecova[1,2], Rossa W. K. Chiu[1,2], Y. M. Dennis Lo[1,2] and Peiyong Jiang[1,2]*

[1]Centre for Research into Circulating Fetal Nucleic Acids, Li Ka Shing Institute of Health Sciences, The Chinese University of Hong Kong, Shatin, NT, Hong Kong
[2]Department of Chemical Pathology, The Chinese University of Hong Kong, Prince of Wales Hospital, Shatin, NT, Hong Kong
*Correspondence to: Peiyong Jiang. E-mail: jiangpeiyong@cuhk.edu.hk

ABSTRACT

Objective The aim of this study is to develop an approach for analyzing plasma DNA sequencing data for noninvasive fetal chromosomal aneuploidy testing that does not require the comparison with control samples or a series of selected genomic regions.

Results We developed the control-free noninvasive fetal chromosomal examination (COFFEE) algorithm by utilizing the size differences between the fetally derived and maternally derived DNA molecules in maternal plasma. We applied COFFEE on three datasets generated in different experimental settings. COFFEE showed 100% accuracy in trisomy 21 testing on these datasets. In contrast, samples analyzed using an existing control-based *z*-score method would introduce a false-positive result because of batch-to-batch variation, when the tested samples were analyzed using control samples from other batches. We believe that COFFEE is useful for enhancing the cost-effectiveness of noninvasive fetal chromosomal aneuploidy testing particularly in laboratories with small caseloads. Source code and testing datasets for COFFEE are available for download at http://www.cuhk.edu.hk/med/cpy/Research/COFFEE/.

Conclusion Control-free noninvasive fetal chromosomal examination is demonstrated to be a versatile data analysis approach and could enhance the application of noninvasive fetal chromosomal aneuploidy detection. © 2017 John Wiley & Sons, Ltd.

Funding sources: This work was supported by the Hong Kong research grants Council Theme-based Research Scheme (T12-403/15-N) and the Vice Chancellor's One-off Discretionary Fund of The Chinese University of Hong Kong (VCF2014021). Y.M.D.L. is supported by an endowed chair from the Li Ka Shing Foundation. Conflicts of interest: R.W.K.C. and Y.M.D.L. received research support from Sequenom, Inc. R.W.K.C. and Y.M.D.L. were consultants to Sequenom, Inc. R.W.K.C., K.C.A.C., and Y.M.D.L. hold equities in Sequenom, Inc. K.C.A.C., R.W.K.C., and Y.M.D.L. are founders of Xcelom and Cirina. K.C.A.C., R.W.K.C., and P.J. are consultants to Xcelom. P.J. is a consultant to Cirina. K.S., K.C.A.C., R.W.K.C., Y.M.D.L., and P.J. have filed patent applications (US Provisional Application No. 62/410,143) related to this work.

INTRODUCTION

The presence of cell-free fetal DNA in maternal plasma of pregnant women was first reported in 1997,[1] offering numerous possibilities for noninvasive prenatal testing. Examples include fetal sex determination for sex-linked diseases,[1] fetal *RHD* genotyping for pregnancies in RhD-negative mothers,[2] chromosomal aneuploidy detection,[3–8] and fetal monogenic disorder detection.[9–13] Noninvasive fetal chromosomal aneuploidy testing using massively parallel sequencing has been rapidly translated into routine clinical services worldwide.[4,14,15] Aneuploid fetuses with an extra copy of one chromosome (e.g. trisomy 21) could be detected noninvasively by counting the small increase in plasma DNA molecules from the affected chromosome.[4,16,17] On the other hand, because fetally derived DNA molecules are shorter than the maternally derived DNA molecules,[10,13,18] the increased portion of DNA from the extra copy of a fetal chromosome would shorten the size distribution derived from that chromosome. On the basis of this principle, Yu *et al.* demonstrated that fetal chromosomal aneuploidy testing could also be achieved with high accuracy by analyzing the sizes of DNA in maternal plasma.[19]

For most existing algorithms, the tested case needs to be compared with a group of control samples to determine whether the count or size of plasma DNA would suggest the tested pregnancy to involve an unaffected fetus or not.[14,15] Because different sequencers or library preparation kits would introduce uneven genome coverage, mostly caused by guanine–cytosine (GC)-content biases,[16,20] some researchers have reservation about sharing data from control samples across different batches. On the other hand, the use of batch-specific controls would increase the cost and reduce the throughput of the analysis. Therefore, algorithms that do not require control samples would be operationally useful. Such

SCIENTIFIC REPORTS

OPEN

mTFkb: a knowledgebase for fundamental annotation of mouse transcription factors

Received: 1 December 2016
Accepted: 12 April 2017
Published online: 08 June 2017

Kun Sun[1,2], Huating Wang[1,3] & Hao Sun[1,2]

Transcription factors (TFs) are well-known important regulators in cell biology and tissue development. However, in mouse, one of the most widely-used model species, currently the vast majority of the known TFs have not been functionally studied due to the lack of sufficient annotations. To this end, we collected and analyzed the whole transcriptome sequencing data from more than 30 major mouse tissues and used the expression profiles to annotate the TFs. We found that the expression patterns of the TFs are highly correlated with the histology of the tissue types thus can be used to infer the potential functions of the TFs. Furthermore, we found that as many as 30% TFs display tissue-specific expression pattern, and these tissue-specific TFs are among the key TFs in their corresponding tissues. We also observed signals of divergent transcription associated with many TFs with unique expression pattern. Lastly, we have integrated all the data, our analysis results as well as various annotation resources to build a web-based database named mTFkb freely accessible at http://www.myogenesisdb.org/mTFkb/. We believe that mTFkb could serve as a useful and valuable resource for TF studies in mouse.

Transcription factors (TFs) are a family of proteins that could bind to specific DNA sequences, usually in enhancer or promoter regions, to regulate the expression of target genes, either positively (as an activator) or negatively (as a repressor)[1–3]. In human, around 8% of the total genes encode TFs[4]. TFs are found to be highly conserved among most of the organisms. For instance, the numbers of annotated TFs in human (Homo Sapiens) and mouse (Mus Musculus) are similar[5] and most of them are conserved between these two species. This highly conserved characteristic suggests that TFs are among the fundamental proteins for normal cellular functions[6]. Therefore, there is ongoing interest in the functional investigation of TFs. They are known essential regulators in normal cell function and tissue development. For instance, MyoD (Myogenic Differentiation 1) and Myf5 (Myogenic factor 5) play key roles in the development of limb and skeletal muscle[7,8]. Furthermore, TFs that are key to guide cell differentiation and tissue development are discovered to interact with regulatory DNA elements such as enhancers and promoters[3,9]. Recent studies also showed that key TFs could establish super-enhancers, clusters of enhancers with high activity, which are essential in controlling cell identity and disease[10,11]. In addition, more and more studies demonstrated the successful reprogramming of somatic cells using a "cocktail" containing key TFs of the target cell type[12]. Very interestingly, emerging reports demonstrated the biological phenomenon of divergent transcription from the promoters of TFs[13,14], which could be helpful in deciphering its significance and functional mechanism[14,15]. For instance, our group has recently discovered a novel long noncoding RNA, Linc-Yy1, which is transcribed from ~2 kb upstream of the Yy1 (Yin Yang 1) gene and serves as an important regulator of mouse skeletal myoblast differentiation through interaction with the Yy1 transcription factor[14]. Collectively, the existing studies reinforced that the TFs are among the most important regulators affecting the identity of cell/tissue type through diversified mechanisms of actions; it is thus imperative to identify the key TFs that are critical for the development of certain tissues.

Knowing their functional significance, however, most of the known TFs have yet to be characterized[16]. Existing studies in human found that the TFs are expressed in a tissue-dependent manner hence the expression pattern of the TFs across various tissues is closely correlated with their functions and could be used to mine the key TFs for the tissues[16–19]. Similar study however is still lacking in mouse, warranting the creation of a public

[1]Li Ka Shing Institute of Health Sciences, The Chinese University of Hong Kong, Hong Kong SAR, China. [2]Department of Chemical Pathology, The Chinese University of Hong Kong, Hong Kong SAR, China. [3]Department of Orthopaedics and Traumatology, The Chinese University of Hong Kong, Hong Kong SAR, China. Correspondence and requests for materials should be addressed to H.W. (email: huating.wang@cuhk.edu.hk) or H.S. (email: haosun@cuhk.edu.hk)

Clinical Chemistry 63:10
1614–1623 (2017)

DNA of Erythroid Origin Is Present in Human Plasma and Informs the Types of Anemia

W.K. Jacky Lam,[1,2†] Wanxia Gai,[1,2†] Kun Sun,[1,2†] Raymond S.M. Wong,[3] Rebecca W.Y. Chan,[1,2] Peiyong Jiang,[1,2] Natalie P.H. Chan,[4] Winnie W.I. Hui,[1,2] Anthony W.H. Chan,[4] Cheuk-Chun Szeto,[3] Siew C. Ng,[3] Man-Fai Law,[3] K.C. Allen Chan,[1,2] Rossa W.K. Chiu,[1,2] and Y.M. Dennis Lo[1,2*]

BACKGROUND: There is much interest in the tissue of origin of circulating DNA in plasma. Data generated using DNA methylation markers have suggested that hematopoietic cells of white cell lineages are important contributors to the circulating DNA pool. However, it is not known whether cells of the erythroid lineage would also release DNA into the plasma.

METHODS: Using high-resolution methylation profiles of erythroblasts and other tissue types, 3 genomic loci were found to be hypomethylated in erythroblasts but hypermethylated in other cell types. We developed digital PCR assays for measuring erythroid DNA using the differentially methylated region for each locus.

RESULTS: Based on the methylation marker in the ferrochelatase gene, erythroid DNA represented a median of 30.1% of the plasma DNA of healthy subjects. In subjects with anemia of different etiologies, quantitative analysis of circulating erythroid DNA could reflect the erythropoietic activity in the bone marrow. For patients with reduced erythropoietic activity, as exemplified by aplastic anemia, the percentage of circulating erythroid DNA was decreased. For patients with increased but ineffective erythropoiesis, as exemplified by β-thalassemia major, the percentage was increased. In addition, the plasma concentration of erythroid DNA was found to correlate with treatment response in aplastic anemia and iron deficiency anemia. Plasma DNA analysis using digital PCR assays targeting the other 2 differentially methylated regions showed similar findings.

CONCLUSIONS: Erythroid DNA is a hitherto unrecognized major component of the circulating DNA pool and is a noninvasive biomarker for differential diagnosis and monitoring of anemia.

Plasma DNA is an increasingly pursued analyte for molecular diagnostics. There are ongoing research studies on its clinical applications, especially in noninvasive prenatal testing *(1–7)* and oncology *(8–12)*. Despite a wide variety of clinical applications, the tissue origin of circulating DNA is not completely understood.

It has been shown that circulating DNA is predominantly released from hematopoietic cells using sex-mismatched bone marrow transplantation as model systems *(13, 14)*. Kun et al. recently demonstrated that a substantial proportion of plasma DNA has methylation signatures of neutrophils and lymphocytes *(15)*. However, there is currently no information regarding whether DNA of erythroid origin might also be detectable in plasma. Red blood cells are the largest population of hematopoietic cells in blood. Mature red blood cells in humans do not have a nucleus. It is during the enucleation step that erythroblasts lose their nuclei and mature into reticulocytes in the bone marrow *(16)*. The nuclear material of the erythroblasts is phagocytosed and degraded by the marrow macrophages in the erythroblastic islands *(17)*. We postulate that some of the degraded DNA material of the erythroid lineage from the bone marrow would be released into the circulation.

We further propose that one can identify methylation signatures of DNA from cells of erythroid origin and use such signatures to see if erythroid DNA might be detectable in human plasma. High-resolution reference methylomes of different tissues and hematopoietic cell types have become publicly available through collaborative projects including

[1] Li Ka Shing Institute of Health Sciences, The Chinese University of Hong Kong, Shatin, New Territories, Hong Kong SAR, China; [2] Department of Chemical Pathology, The Chinese University of Hong Kong, Prince of Wales Hospital, Shatin, New Territories, Hong Kong SAR, China; [3] Department of Medicine and Therapeutics, The Chinese University of Hong Kong, Prince of Wales Hospital, Shatin, New Territories, Hong Kong SAR, China; [4] Department of Anatomical and Cellular Pathology, The Chinese University of Hong Kong, Prince of Wales Hospital, Shatin, New Territories, Hong Kong SAR, China.

* Address correspondence to this author at: Department of Chemical Pathology, The Chinese University of Hong Kong, Prince of Wales Hospital, 30–32 Ngan Shing Street, Shatin, New Territories, Hong Kong SAR, China. Fax +852-26365090; e-mail loym@cuhk.edu.hk.

† W.K. Jacky Lam, Wanxia Gai, and Kun Sun contributed equally to this work, and all 3 should be considered as first authors.

Received February 2, 2017; accepted July 20, 2017.
Previously published online at DOI: 10.1373/clinchem.2017.272401

Bioinformatics, 33(21), 2017, 3495–3496
doi: 10.1093/bioinformatics/btx505
Advance Access Publication Date: 8 August 2017
Applications Note

Systems biology

BSviewer: a genotype-preserving, nucleotide-level visualizer for bisulfite sequencing data

Kun Sun[1,2], Fiona F. M. Lun[1,2], Peiyong Jiang[1,2] and Hao Sun[1,2,*]

[1]Li Ka Shing Institute of Health Sciences and [2]Department of Chemical Pathology, The Chinese University of Hong Kong, Prince of Wales Hospital, New Territories, Hong Kong, SAR, China

*To whom correspondence should be addressed.
Associate Editor: Bonnie Berger

Received on May 25, 2017; revised on July 23, 2017; editorial decision on August 3, 2017; accepted on August 7, 2017

Abstract

Motivation: The bisulfite sequencing technology has been widely used to study the DNA methylation profile in many species. However, most of the current visualization tools for bisulfite sequencing data only provide high-level views (i.e. overall methylation densities) while miss the methylation dynamics at nucleotide level. Meanwhile, they also focus on CpG sites while omit other information (such as genotypes on SNP sites) which could be helpful for interpreting the methylation pattern of the data. A bioinformatics tool that visualizes the methylation statuses at nucleotide level and preserves the most essential information of the sequencing data is thus valuable and needed.
Results: We have developed BSviewer, a lightweight nucleotide-level visualization tool for bisulfite sequencing data. Using an imprinting gene as an example, we show that BSviewer could be specifically helpful for interpreting the data with allele-specific DNA methylation pattern.
Availability and implementation: BSviewer is implemented in Perl and runs on most GNU/Linux platforms. Source code and testing dataset are freely available at http://sunlab.cpy.cuhk.edu.hk/BSviewer/.
Contact: haosun@cuhk.edu.hk

1 Introduction

Recently, the development of bisulfite sequencing technology (Lister *et al.*, 2009) has largely facilitated the studies on genomewide DNA methylation profiling. Although plenty of bioinformatics software has been developed to analyze and interpret the bisulfite sequencing data, the visualization tools currently available do not support the bisulfite sequencing data analysis well. For example, universal data visualization platforms like UCSC Genome Browser (Kent *et al.*, 2002), IGV (Integrative Genomics Viewer) (Thorvaldsdottir *et al.*, 2013) and EpiViz (Chelaru *et al.*, 2014) provide interactive interfaces and analysis utilities while they are mostly suitable for viewing the data at high level (i.e. overall methylation densities for CpG sites or genomic regions) while missed methylation dynamics at the nucleotide level; other tools like MethylViewer (Pardo *et al.*, 2011), CpGviewer (Carr *et al.*, 2007) and MethPat (Wong *et al.*, 2016), on the other hand, allow the users to view the methylation pattern at single nucleotide resolution as well as the methylation pattern within one single read, but they only focus on CpG sites while omit other useful information in the sequencing data. In fact, besides the methylation statuses of the CpG sites, the sequencing reads contain much additional information (e.g. genotypes) which is helpful for interpreting the data. For instance, the imprinting genes are known to have a parent-of-origin-specific DNA methylation pattern, therefore the genotype information is critical to infer the parent-of-origin of the DNA molecules for interpreting the imprinting pattern. However, using the current visualization tools, one could only observe an intermediate DNA methylation signal (i.e. overall DNA methylation level) while miss the allele-specific DNA methylation pattern. A novel visualization tool that could preserve the most essential information in the sequencing data (e.g. genotype) along with the methylation statuses of CpG sites thus could be valuable for bisulfite sequencing data interpretation.

In this work, we have developed a bioinformatics tool, named BSviewer, specifically designed for visualizing bisulfite sequencing data.

Received: 30 November 2017 | Revised: 5 January 2018 | Accepted: 5 January 2018

DOI: 10.1002/pd.5214

ORIGINAL ARTICLE

WILEY PRENATAL DIAGNOSIS

Noninvasive reconstruction of placental methylome from maternal plasma DNA: Potential for prenatal testing and monitoring

Kun Sun[1,2] | Fiona M.F. Lun[1,2] | Tak Y. Leung[3] | Rossa W.K. Chiu[1,2] | Y.M. Dennis Lo[1,2] | Hao Sun[1,2]

[1] Li Ka Shing Institute of Health Sciences, The Chinese University of Hong Kong, Shatin, Hong Kong

[2] Department of Chemical Pathology, The Chinese University of Hong Kong, Shatin, Hong Kong

[3] Department of Obstetrics and Gynaecology, The Chinese University of Hong Kong, Shatin, Hong Kong

Correspondence
Hao Sun, Rm 503, Li Ka Shing Institute of Health Sciences, Prince of Wales Hospital, Shatin, Hong Kong.
Email: haosun@cuhk.edu.hk

Funding information
Li Ka Shing Foundation; The Chinese University of Hong Kong, Grant/Award Numbers: VCF2014021 and 1907307; Hong Kong Research Grants Council Theme-Based Research Scheme, Grant/Award Number: T12-403/15-N

Abstract

Objective: During human pregnancy, the DNA methylation of placental tissue is highly relevant to the normal growth and development of the fetus; therefore, methylomic analysis of the placental tissue possesses high research and clinical value in prenatal testing and monitoring. Thus, our aim is to develop an approach for reconstruction of the placental methylome, which should be completely noninvasive and achieve high accuracy and resolution.

Results: We propose a novel size-based algorithm, FEtal MEthylome Reconstructor (FEMER), to noninvasively reconstruct the placental methylome by genomewide bisulfite sequencing and size-based analysis of maternal plasma DNA. By applying FEMER on a real clinical dataset, we demonstrate that FEMER achieves both high accuracy and resolution, thus provides a high-quality view of the placental methylome from maternal plasma DNA. FEtal MEthylome Reconstructor could also predict the DNA methylation profile of CpG islands with high accuracy, thus shows potential in monitoring of key genes involved in placental/fetal development. Source code and testing datasets for FEMER are available at http://sunlab.cpy.cuhk.edu.hk/FEMER/.

Conclusion: FEtal MEthylome Reconstructor could enhance the noninvasive fetal/placental methylomic analysis and facilitate its application in prenatal testing and monitoring.

1 | INTRODUCTION

Human pregnancy is a complex while strictly orchestrated process, during which the placenta plays an essential role for the normal growth and development of the fetus. Among the many biological factors that control the development and function of the placenta, DNA methylation is known to serve as a key epigenetic regulator.[1] In the mammalian genome, DNA methylation mostly appears in CpG dinucleotides and is known to be associated with gene repression.[2] Compared to most human somatic tissues, the methylome of placental tissue shows a number of notable characteristics, such as genomewide hypomethylation and specific imprinting patterns.[3-5] Dysregulation of the placental methylome is associated with adverse placental morphology and birth outcome, including growth restriction and spontaneous abortion.[6-8] Thus, methylomic analysis of the placenta is of great research interest and potential clinical value. Currently, there are several approaches for obtaining methylomic information of the placenta.[9] However, most of these approaches require invasive sampling of placental tissue, thus is associated with risks to the fetus. Furthermore, their invasive nature makes these approaches infeasible for use in monitoring the dynamic changes during different gestation periods. Noninvasive approaches are thus much favored.

The discovery of circulating cell-free DNA of fetal origin in the plasma of pregnant women[10] opens up the possibility of plasma DNA-based noninvasive prenatal testing (NIPT). Recent studies have demonstrated the high clinical value of cell-free fetal DNA in NIPT, such as fetal aneuploidy detection.[11] The cell-free fetal DNA molecules in maternal plasma mostly originate from the placental tissue,[12-14] thus can potentially be used for examining the placental methylome in a noninvasive manner. Recently, genomewide bisulfite sequencing (BS-seq) technology[15,16] has largely facilitated the methylomic studies. In BS-seq experiments, DNA is treated with sodium bisulfite to convert cytosines (Cs) into uracils (Us), while methylcytosines remain unmodified. During polymerase chain reaction, the Us are amplified as thymines (Ts); therefore, by comparing the modified DNA with the original sequence, the methylation states

Size-tagged preferred ends in maternal plasma DNA shed light on the production mechanism and show utility in noninvasive prenatal testing

Kun Sun[a,b], Peiyong Jiang[a,b], Ada I. C. Wong[a,b], Yvonne K. Y. Cheng[c], Suk Hang Cheng[a,b], Haiqiang Zhang[a,b], K. C. Allen Chan[a,b], Tak Y. Leung[c], Rossa W. K. Chiu[a,b], and Y. M. Dennis Lo[a,b,1]

[a]Li Ka Shing Institute of Health Sciences, The Chinese University of Hong Kong, Shatin, New Territories, Hong Kong SAR, China; [b]Department of Chemical Pathology, Prince of Wales Hospital, The Chinese University of Hong Kong, Shatin, New Territories, Hong Kong SAR, China; and [c]Department of Obstetrics and Gynaecology, Prince of Wales Hospital, The Chinese University of Hong Kong, Shatin, New Territories, Hong Kong SAR, China

Contributed by Y. M. Dennis Lo, April 24, 2018 (sent for review March 13, 2018; reviewed by Erik A. Sistermans and Michael R. Speicher)

Cell-free DNA in human plasma is nonrandomly fragmented and reflects genomewide nucleosomal organization. Previous studies had demonstrated tissue-specific preferred end sites in plasma DNA of pregnant women. In this study, we performed integrative analysis of preferred end sites with the size characteristics of plasma DNA fragments. We mined the preferred end sites in short and long plasma DNA molecules separately and found that these "size-tagged" ends showed improved accuracy in fetal DNA fraction estimation and enhanced noninvasive fetal trisomy 21 testing. Further analysis revealed that the fetal and maternal preferred ends were generated from different locations within the nucleosomal structure. Hence, fetal DNA was frequently cut within the nucleosome core while maternal DNA was mostly cut within the linker region. We further demonstrated that the nucleosome accessibility in placental cells was higher than that for white blood cells, which might explain the difference in the cutting positions and the shortness of fetal DNA in maternal plasma. Interestingly, short and long size-tagged ends were also observable in the plasma of nonpregnant healthy subjects and demonstrated size differences similar to those in the pregnant samples. Because the nonpregnant samples did not contain fetal DNA, the data suggested that the interrelationship of preferred DNA ends, chromatin accessibility, and plasma DNA size profile is likely a general one, extending beyond the context of pregnancy. Plasma DNA fragment end patterns have thus shed light on production mechanisms and show utility in future developments in plasma DNA-based noninvasive molecular diagnostics.

circulating cell-free DNA | prenatal diagnosis | nucleosome structure | size-based molecular diagnostics | liquid biopsy

There is global interest in adopting circulating cell-free DNA analysis in human plasma for molecular diagnostics and monitoring. The discoveries of fetal DNA in the plasma of pregnant women (1), donor-specific DNA in organ-transplantation patients (2), and tumor-derived DNA in cancer patients (3) have enabled technologies for noninvasive prenatal testing, cancer liquid biopsies, transplant monitoring, and organ damage assessment (4–8). Despite the numerous clinical applications, however, the biological characteristics of the plasma DNA have received much less research attention.

It has been demonstrated that plasma DNA is not randomly fragmented. High-resolution plasma DNA size profiling revealed a predominant peak at 166 bp and a 10-bp periodicity below 150 bp (9). This size profile has been proposed to be closely related to the nucleosomal structure (9). In this regard, the nucleosome is composed of an octamer of four core histone proteins (forming a "nucleosome core" wrapped by 147 bp of DNA with a ~10-bp helical repeat), linker histones, and linker DNA (mean size around 20 bp; size varies from 0 to 80 bp) (10). Furthermore, the fetal DNA in maternal plasma [mostly originating from placental tissues (11)] has been found to be shorter than the maternal DNA [mostly originating from the hematopoietic system (12–14)]. The size differences in the fetal and maternal DNA molecules had been utilized in noninvasive prenatal testing, allowing fetal DNA fraction estimation, fetal chromosomal aneuploidy detection, and fetal methylome analysis (15–19). However, the mechanistic basis for this relative shortening of circulating fetal DNA is still poorly understood (9, 14, 20).

Recent studies further explored the ending pattern of plasma DNA. Ultradeep sequencing of plasma DNA in pregnant women revealed the existence of fetal- and maternal-specific preferred end sites (21). Although these preferred end sites demonstrated potential for noninvasive prenatal testing, the molecular basis for their existence is largely unknown. In addition, plasma DNA is believed to be released from apoptotic cells (22), suggesting that the fragmentation pattern is correlated with the nucleosomal structure and chromatin states (23–26). It is thus worthwhile to

Significance

Cell-free DNA molecules in the plasma of pregnant women exhibit nonrandom fragmentation with preferred end sites. We studied if such preferred end sites might bear any relationship with fragment lengths of plasma DNA. Short and long plasma DNA molecules were associated with different preferred DNA end sites. Analysis of size-tagged preferred ends could be used for measuring fetal DNA fraction and for facilitating fetal trisomy 21 detection. Fetal preferred end sites were generally located in the nucleosome cores, while the maternal ones were located in the linker regions. This conceptual framework provides an explanation of the relative shortness of fetal DNA in maternal plasma and brings us closer to understanding the biological mechanisms that influence plasma DNA fragmentation.

Author contributions: K.S., K.C.A.C., R.W.K.C., and Y.M.D.L. designed research; K.S., P.J., A.I.C.W., S.H.C., and H.Z. performed research; K.S., P.J., Y.K.Y.C., K.C.A.C., T.Y.L., R.W.K.C., and Y.M.D.L. analyzed data; Y.K.Y.C. and T.Y.L. recruited subjects; and K.S., R.W.K.C., and Y.M.D.L. wrote the paper.

Reviewers: E.A.S., VU University Medical Center; and M.R.S., Medical University of Graz.

Conflict of interest statement: K.C.A.C., R.W.K.C., and Y.M.D.L. hold equities in DRA and Grail, are consultants to Grail, and receive research funding from Grail/Cirina. P.J. is a consultant to Xcelom and Grail. Y.M.D.L. is a scientific cofounder and a member of the scientific advisory board for Grail.

Data deposition: Sequence data for the subjects studied in this work who had consented to data archiving have been deposited at the European Genome-Phenome Archive, https://www.ebi.ac.uk/ega/, hosted by the European Bioinformatics Institute (accession no. EGAS00001002831).

[1]To whom correspondence should be addressed. Email: loym@cuhk.edu.hk.

This article contains supporting information online at www.pnas.org/lookup/suppl/doi:10.1073/pnas.1804134115/-/DCSupplemental.

Published online May 14, 2018.

Preferred end coordinates and somatic variants as signatures of circulating tumor DNA associated with hepatocellular carcinoma

Peiyong Jiang[a,b,1], Kun Sun[a,b,1], Yu K. Tong[a,b], Suk Hang Cheng[a,b], Timothy H. T. Cheng[a,b], Macy M. S. Heung[a,b], John Wong[c], Vincent W. S. Wong[d,e], Henry L. Y. Chan[d,e], K. C. Allen Chan[a,b,f], Y. M. Dennis Lo[a,b,f,2], and Rossa W. K. Chiu[a,b,2]

[a]Li Ka Shing Institute of Health Sciences, The Chinese University of Hong Kong, Shatin, New Territories, Hong Kong SAR, China; [b]Department of Chemical Pathology, The Chinese University of Hong Kong, Prince of Wales Hospital, Shatin, New Territories, Hong Kong SAR, China; [c]Department of Surgery, The Chinese University of Hong Kong, Prince of Wales Hospital, Shatin, New Territories, Hong Kong SAR, China; [d]Department of Medicine and Therapeutics, The Chinese University of Hong Kong, Prince of Wales Hospital, Shatin, New Territories, Hong Kong; [e]Institute of Digestive Diseases, The Chinese University of Hong Kong, Prince of Wales Hospital, Shatin, New Territories, Hong Kong; and [f]State Key Laboratory in Translational Oncology, The Chinese University of Hong Kong, Prince of Wales Hospital, Shatin, New Territories, Hong Kong SAR, China

Contributed by Y. M. Dennis Lo, September 24, 2018 (sent for review August 28, 2018; reviewed by Ellen Heitzer and David Sidransky)

Circulating tumor-derived cell-free DNA (ctDNA) analysis offers an attractive noninvasive means for detection and monitoring of cancers. Evidence for the presence of cancer is dependent on the ability to detect features in the peripheral circulation that are deemed as cancer-associated. We explored approaches to improve the chance of detecting the presence of cancer based on sequence information present on ctDNA molecules. We developed an approach to detect the total pool of somatic mutations. We then investigated if there existed a class of ctDNA signature in the form of preferred plasma DNA end coordinates. Cell-free DNA fragmentation is a nonrandom process. Using plasma samples obtained from liver transplant recipients, we showed that liver contributed cell-free DNA molecules ended more frequently at certain genomic coordinates than the nonliver-derived molecules. The abundance of plasma DNA molecules with these liver-associated ends correlated with the liver DNA fractions in the plasma samples. Studying the DNA end characteristics in plasma of patients with hepatocellular carcinoma and chronic hepatitis B, we showed that there were millions of tumor-associated plasma DNA end coordinates in the genome. Abundance of plasma DNA molecules with tumor-associated DNA ends correlated with the tumor DNA fractions even in plasma samples of hepatocellular carcinoma patients that were subjected to shallow-depth sequencing analysis. Plasma DNA end coordinates may therefore serve as hallmarks of ctDNA that could be sampled readily and, hence, may improve the cost-effectiveness of liquid biopsy assessment.

tumor-associated preferred ends | liver-associated preferred ends | tumor-derived cell-free DNA | hepatocellular carcinoma | transplantation

DNA molecules released by malignant cells are present in the peripheral circulation of cancer patients, providing noninvasive access to such genetic material. Circulating tumor-derived cell-free DNA (ctDNA) analysis has been utilized as a liquid biopsy for the management of cancer. Liquid biopsies may serve as a surrogate of invasive biopsies because ctDNA molecules harbor molecular features that are associated with cancers. ctDNA features that have been characterized include somatic mutations, cancer-associated viral sequences, copy number aberrations, and differential DNA methylation signatures (1–5). We envision that the most direct means to detect cancer using a cell-free DNA sample may be through the detection of cancer-associated hallmarks that are physically present on cell-free DNA molecules. Such DNA molecules would be deemed as informative signals for the presence of cancer. Among the methods mentioned above, one disadvantage of approaches based on the detection of viral nucleic acids is that not all cancers are associated with viral infections. However, the use of DNA methylation analyses require performance of additional laboratory steps such as bisulfite conversion. As a result, many research groups have focused on the detection of somatic mutations in plasma (4, 6–8).

For the purpose of early cancer detection (1, 4, 9, 10), testing approaches need to be able to detect biomarkers that are broadly represented among the majority of cancer cases in the target population without a priori information of the tumor genetic

Significance

Cell-free DNA fragmentation is a nonrandom process. We showed that cell-free DNA fragments with ends at certain genomic coordinates had higher likelihoods of being derived from hepatocellular carcinoma. Other coordinates were associated with cell-free DNA molecules originating from the liver. Quantitative assessment of cell-free DNA molecules bearing these respective groups of end signatures correlated with the amounts of tumor-derived or liver-derived DNA in plasma. There were millions of tumor-associated plasma DNA end coordinates across the genome. Due to their high prevalence, they were more readily detectable than somatic mutations as a cancer signature in plasma. Hence, detection of tumor-associated plasma DNA ends may offer a cost-effective means of capturing evidence for the presence of cancer through liquid biopsy assessment.

Author contributions: K.C.A.C., Y.M.D.L., and R.W.K.C. designed research; P.J., K.S., Y.K.T., S.H.C., T.H.T.C., M.M.S.H., K.C.A.C., Y.M.D.L., and R.W.K.C. performed research; P.J., K.S., J.W., V.W.S.W., H.L.Y.C., K.C.A.C., Y.M.D.L., and R.W.K.C. analyzed data; P.J., Y.M.D.L., and R.W.K.C. wrote the paper; and T.H.T.C., J.W., V.W.S.W., and H.L.Y.C. recruited subjects and analyzed clinical data.

Reviewers: E.H., Medical University of Graz; and D.S., Johns Hopkins University School of Medicine.

Conflict of interest statement: The authors declare a conflict of interest. K.C.A.C., R.W.K.C., and Y.M.D.L. hold equities in Grail. P.J., K.C.A.C., R.W.K.C., and Y.M.D.L. are consultants to Grail. K.C.A.C., R.W.K.C., and Y.M.D.L. receive research funding from Grail. Y.M.D.L. is a scientific cofounder of and serves on the scientific advisory board of Grail. P.J., K.C.A.C., R.W.K.C., and Y.M.D.L. have filed patent applications based on the data generated from this work. Patent royalties are received from Grail, Illumina, Sequenom, and Xcelom.

Data deposition: Sequence data for the subjects studied in this work who had consented to data archiving have been deposited at the European Genome-Phenome Archive (EGA), https://www.ebi.ac.uk/ega/, hosted by the European Bioinformatics Institute (EBI) (accession no. EGAS00001003160).

[1]P.J. and K.S. contributed equally to this work.

[2]To whom correspondence may be addressed. Email: loym@cuhk.edu.hk or rossachiu@cuhk.edu.hk.

This article contains supporting information online at www.pnas.org/lookup/suppl/doi:10.1073/pnas.1814616115/-/DCSupplemental.

Published online October 29, 2018.

Bioinformatics, 34(23), 2018, 4129–4130
doi: 10.1093/bioinformatics/bty524
Advance Access Publication Date: 27 June 2018
Applications Note

Systems biology

GeneCT: a generalizable cancerous status and tissue origin classifier for pan-cancer biopsies

Kun Sun[1,2,*], Jiguang Wang[3], Huating Wang[1,4] and Hao Sun[1,2,*]

[1]Li Ka Shing Institute of Health Sciences and [2]Department of Chemical Pathology, The Chinese University of Hong Kong, Hong Kong, SAR, China, [3]Divison of Life Science and Department of Chemical and Biological Engineering, The Hong Kong University of Science and Technology, Hong Kong, SAR, China and [4]Department of Orthopaedics and Traumatology, The Chinese University of Hong Kong, Hong Kong, SAR, China

*To whom correspondence should be addressed.

Associate Editor: Bonnie Berger

Received on March 29, 2018; revised on June 3, 2018; editorial decision on June 25, 2018; accepted on June 26, 2018

Abstract

Motivation: Tissue biopsy is commonly used in cancer diagnosis and molecular studies. However, advanced skills are required for determining cancerous status of biopsies and tissue origin of tumor for cancerous ones. Correct classification is essential for downstream experiment design and result interpretation, especially in molecular cancer studies. Methods for accurate classification of cancerous status and tissue origin for pan-cancer biopsies are thus urgently needed.

Results: We developed a deep learning-based classifier, named GeneCT, for predicting cancerous status and tissue origin of pan-cancer biopsies. GeneCT showed high performance on pan-cancer datasets from various sources and outperformed existing tools. We believe that GeneCT can potentially facilitate cancer diagnosis, tumor origin determination and molecular cancer studies.

Availability and implementation: GeneCT is implemented in Perl/R and supported on GNU/Linux platforms. Source code, testing data and webserver are freely available at http://sunlab.cpy.cuhk.edu.hk/GeneCT/.

Contact: sunkun@cuhk.edu.hk or haosun@cuhk.edu.hk

Supplementary information: Supplementary data are available at *Bioinformatics* online.

1 Introduction

Tissue biopsy is a widely used technique for cancer diagnosis and molecular cancer studies. Conventionally, a skilled pathologist is required to perform physical and histological analysis of the cells to determine the cancerous status and/or tissue origin of the biopsy, which is costly, time-consuming and demands specific knowledge and experience. The correct classification is critical for downstream experimental design and data interpretation. Emerging studies have built classifiers for predicting cancerous status and tissue origin of pan-cancer biopsies using molecular data. For instance, gene expression based classifiers have demonstrated high accuracy in the task (Jung *et al.*, 2015; Pal *et al.*, 2014, 2015; Wei *et al.*, 2014). However, most of these studies heavily rely on cancer/tissue-type specific biomarkers mined from existing data which limits the applications on cancer types that are not included in the training dataset when building the classifiers. In addition, most of the existing classifiers have not been tested by datasets generated from different sources, which may suffer from low consistency due to variance in protocols/platforms used in library preparation (SEQC Consortium, 2014). To this end, in this study, we developed a novel cancerous status and tissue origin classifier, named GeneCT (Generalizable Cancerous-status and Tissue-of-origin classifier), for pan-cancer biopsies by utilizing gene expression profiles and harnessing the power of deep learning. Compared to general machine learning, deep learning requires a higher number of training samples and computing resources while resulting in a higher performance classifier (Jurtz *et al.*, 2017), which was demonstrated in various biological studies such as skin cancer (Esteva *et al.*, 2017) and eye disease (Kermany *et al.*, 2018) predictions using medical images. In this study, we demonstrated that GeneCT showed high performance on pan-cancer datasets from various sources and outperformed existing methods. GeneCT thus could serve as a useful tool for analysis of tissue biopsies, especially in molecular cancer studies.