Genomics 103 (2014) 189–203

Contents lists available at ScienceDirect

Genomics

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

Decoding complex patterns of genomic rearrangement in hepatocellular carcinoma☆

Julio Fernandez-Banet [a], Nikki P. Lee [b], Kin Tak Chan [b], Huan Gao [c], Xiao Liu [c,d], Wing-Kin Sung [e,f], Winnie Tan [b], Sheung Tat Fan [b], Ronnie T. Poon [b], Shiyong Li [c], Keith Ching [a], Paul A. Rejto [a], Mao Mao [a,g], Zhengyan Kan [a,]*

[a] Pfizer Oncology, San Diego, CA, USA
[b] Department of Surgery, University of Hong Kong, Hong Kong, China
[c] BGI-Shenzhen, Shenzhen, China
[d] Department of Biology, University of Copenhagen, Copenhagen, Denmark
[e] School of Computing, National University of Singapore, Singapore
[f] Computational and Systems Biology, Genome Institute of Singapore, Singapore
[g] Asian Cancer Research Group, Inc., Wilmington, DE, USA

ARTICLE INFO

Article history:
Received 12 October 2013
Accepted 11 January 2014
Available online 21 January 2014

Keywords:
Genomic rearrangement
Hepatocellular carcinoma
Whole-genome sequencing
Copy number variation

ABSTRACT

Elucidating the molecular basis of hepatocellular carcinoma (HCC) is crucial to developing targeted diagnostics and therapies for this deadly disease. The landscape of somatic genomic rearrangements (GRs), which can lead to oncogenic gene fusions, remains poorly characterized in HCC. We have predicted 4314 GRs including large-scale insertions, deletions, inversions and translocations based on the whole-genome sequencing data for 88 primary HCC tumor/non-tumor tissues. We identified chromothripsis in 5 HCC genomes (5.7%) recurrently affecting chromosomal arms 1q and 8q. Albumin (*ALB*) was found to harbor GRs, deactivating mutations and deletions in 10% of cohort. Integrative analysis identified a pattern of paired intra-chromosomal translocations flanking focal amplifications and asymmetrical patterns of copy number variation flanking breakpoints of translocations. Furthermore, we predicted 260 gene fusions which frequently result in aberrant over-expression of the 3′ genes in tumors and validated 18 gene fusions, including recurrent fusion (2/88) of *ABCB11* and *LRP2*.

1. Introduction

Hepatocellular carcinoma (HCC) is the major histological subtype of liver cancer, the third leading cause of cancer mortality worldwide with high prevalence in Asia and sub-Saharan Africa. Hepatitis B virus (HBV) infection is believed to cause the majority of HCCs while other etiological factors include hepatitis C viral (HCV) infection, alcoholism and aflatoxin B1 exposure [1]. Characterization of the molecular pathogenesis of HCC could have a major impact on the diagnosis and treatment of this disease with few effective therapies [2,3]. Significant progress has been made to uncover genetic aberrations in HCCs [4], including identification of mutations in p53 (*TP53*) and β-catenin (*CTNNB1*), amplifications of *MYC*, *FGF19* and cyclin D1 (*CCND1*), over-expression of ErbB and cMet receptors, and HBV integrations into the *TERT* and *KMT2B* gene loci. Recent next generation sequencing studies [5] have further implicated chromatin remodeling pathway genes such as *ARID2*, *ARID1A* and *KMT2C* as potential drivers of HCC carcinogenesis. However, the genomic landscape of somatic genomic rearrangement (GR) remains poorly characterized in HCC. Somatic genomic rearrangement is known to induce oncogenic gene fusions such as *TMPRSS2-ETS* in prostate cancer [6] and *EML4-ALK* in non-small cell lung cancer [7]. The advent of whole-genome and transcriptome sequencing provides opportunities to comprehensively characterize large-scale and complex genomic variations at single base-pair resolution [8,9]. Here we report a comprehensive study of somatic genomic rearrangements and gene fusions in HCC based on whole genome sequencing (WGS) of a cohort of 88 tumors and matched normal samples.

2. Materials and methods

2.1. Whole-genome sequencing

Liver tumor and matched adjacent non-tumor tissues were collected with written informed consents from 88 Chinese HCC patients who received surgical treatments at Hong Kong Queen Mary Hospital as previously described [10]. Approval for the use of clinical specimens for research was obtained from Institutional Review Board of the University of Hong Kong/Hospital Authority Hong Kong West Cluster (HKU/HA HKW IRB). The vast majority (92%, $n = 81$) of patients in

☆ This is an open-access article distributed under the terms of the Creative Commons Attribution-NonCommercial-No Derivative Works License, which permits non-commercial use, distribution, and reproduction in any medium, provided the original author and source are credited.

* Corresponding author at: 10770 Science Center Drive, San Diego, CA 92121, USA. Fax: +1 8585264214.
E-mail address: zhengyan.kan@pfizer.com (Z. Kan).

0888-7543/$ – see front matter © 2014 The Authors. Published by Elsevier Inc. All rights reserved.
http://dx.doi.org/10.1016/j.ygeno.2014.01.003

this cohort were HBV carriers suffering from chronic hepatitis B or cirrhosis. WGS libraries of two different insert sizes (170-bp and 800-bp) were constructed from each sample and sequenced in 2:3 ratio on the Hiseq 2000 sequencers according to manufacturer's instructions (Illumina) [10]. The average depth of base pair coverage was 36X except for three tumor/normal pairs sequenced at 100X coverage. 90-bp paired-end reads were aligned to the hg19 reference genome, and somatic SNVs were called by SOAPsnv [11]. We used SegSeq [12] to identify copy number segments. Somatic mutation and CNV predictions [13], HBV integration site analysis [10] and gene expression profiling [14] were previously described.

2.2. Somatic GR detection and filtering

We developed a somatic GR detection and annotation pipeline consisting of four major steps (Supplementary Fig. 1). (1) Raw WGS reads were aligned to the reference human genome (hg19) by the Burrows-Wheeler Aligner (BWA) [15]. (2) The alignment outputs were screened, and soft-clipped sequences were extracted and analyzed using CREST [16], run in tumor and non-tumor samples independently. (3) GR calls were filtered as germline events if there was an exact match in the coordinates of the breakpoints with GRs identified in the matched or any other non-tumor samples. We also filtered GRs with at least one breakpoint matching a known germline event reported in DGV [17], GRs with breakpoint located within <1 kb from a gap region in genome assembly and GRs with breakpoint falling into a repeat-masked region. A GR event was called somatic only if it passes above filtering criteria and there is sufficient read coverage (≥3) at the genomic region corresponding to each GR breakpoint in the matched non-tumor sample. (4) RefSeq transcript dataset [18] was used to annotate the remaining somatic candidates. For each gene, a reference transcript was defined as the transcript having the longest protein-coding sequence. GR breakpoints were annotated based on locations relative to the reference transcript of the affected gene as "intronic", "exonic", "intergenic" or "promoter", <1 kb upstream of transcription start site.

2.3. Gene fusion annotation

GR events can fuse together sequences from disparate gene loci to form gene fusions. To define a candidate gene fusion, we required transcriptional directions of the partner genes to agree in fused sequences (Supplementary Fig. 2). A gene fusion event was classified as "coding" if both breakpoints reside within the coding regions of affected genes, "UTR" if one or both breakpoints are located in the UTR or "promoter" if breakpoint at the 5′ or 3′ gene is located within the promoter region. Frame conservation status was evaluated for "coding" gene fusions. A fusion was classified as "frame-shift" if it alters the translation frame of the 3′ partner gene based on reference transcript for each of the fusion genes. "Frame-shift" fusion sequences were translated into the frame that maximally conserves the protein sequence of the 3′ gene. Alternative Methionine residues that could represent new translation initiation sites were identified. The protein domain composition in the fusion product sequence was analyzed using NCBI's conserved domain database and search tool [19].

2.4. RNA-seq experiment

Total RNA isolated with TRIzol reagent was treated with RNase-free DNaseI(New England BioLabs) at 37 °C for 10 min. The Dynabeads mRNA Purification Kit (Life Technologies) was used to isolate mRNA from the total RNA samples. The mRNA was chemically fragmented by divalent cations and converted into single-stranded cDNA using random hexamer primers and SuperscriptIIreverse transcriptase (Life Technologies). The second strand was generated to create double-stranded cDNA using RNase H (Enzymatics) and DNA polymeraseI. The cDNA product was purified by Ampure beads XP (Beckman). After converting the overhangs into blunt ends using T4 DNA polymerase and Klenow DNA polymerase, an "A" base was added to the 3′ end of the DNA fragments by the polymerase activity of Klenow fragment. Sequencing adapters were subsequently ligated to the cDNA fragment ends using T4 DNA Ligase (Enzymatics). Fragments of ~200 bps were selected by Ampure beads XP (Beckman) and enriched by 12 cycles of PCR. PCR products were sequenced by Hiseq 2000 (Illumina) according to manufacturer's instructions.

2.5. RNA-seq data analysis

Reads that contain adapter sequences, ≥10% unknown bases or ≥ 50% low quality bases (quality score ≤5) were removed before analysis. Filtered reads are mapped to reference genome (hg19) using SOAP2 [11] (http://soap.genomics.org.cn/). For the 90-bp reads, ≤5 mismatches are allowed in the alignment. The gene expression level is calculated using the RPKM method [20]:

$$RPKM = \frac{\frac{10^6 C}{NL}}{10^3}$$

C: number of reads uniquely aligned to gene of interest; N: number of reads uniquely aligned to all genes; L: gene length in bps. To assess RNA-seq support for gene fusion predictions, we performed read coverage analysis on gene fusions identified in 9 samples with RNA-seq data available. Each exon of the fusion gene was divided into 100-bp windows, and the RPKM values for each window were calculated. For cases where an exon was split by a GR breakpoint, 100-bp windows were derived for each exon segment independently. The RNA-seq read coverage flanking the GR breakpoint as well as coverage in tumors vs. matched non-tumors was compared to identify anomalous expression patterns indicative of gene fusion.

2.6. GR simulation

We repeated the following process for 1000 iterations to generate simulated GR events using the set of 4314 somatic GRs as seed, keeping the number of events per sample constant for each iteration. First, chromosome and coordinates of the observed GR breakpoints were randomized. For inter-chromosomal events, both breakpoints are randomized. For intra-chromosomal events, only one of the breakpoints was randomized, and the other breakpoint was kept in the same distance as observed with a correction applied to fit within the chromosome. The mitochondrial (MT) chromosome was excluded from this step. Simulated GRs were annotated in the same way as described previously.

2.7. Integrative analysis of GR and CNV patterns

Predicted CNV segments were filtered of segments shorter than 500 bps. We define "breakpoint juxtaposition" as an event where a GR breakpoint falls within 100 bps of the start or end coordinates of a CNV segment. The percentage of GR breakpoints were calculated for each GR type and shown in Fig. 5a. Both CNV segments upstream and downstream of the GR breakpoint were counted to calculate the relative distribution of copy gain/loss statuses for CNVs juxtaposed with a specific GR type.

The copy number profile of the 200-kb region flanking translocation breakpoints on both sides was derived from read coverage (cov) of tumor and matched non-tumor samples. The "mpileup" utility from the samtools package [21] was used to fetch the coverage in 100-bp windows, and the copy number (CN) for each window is calculated as the following.

$$CN = 2^{*}\left(\frac{cov_{tumor}}{cov_{non-tumor}}\right)$$