nature genetics

High-resolution characterization of a hepatocellular carcinoma genome

Yasushi Totoki[1], Kenji Tatsuno[2], Shogo Yamamoto[2], Yasuhito Arai[1], Fumie Hosoda[1], Shumpei Ishikawa[3], Shuichi Tsutsumi[2], Kohtaro Sonoda[2], Hirohiko Totsuka[4], Takuya Shirakihara[1], Hiromi Sakamoto[4], Linghua Wang[2], Hidenori Ojima[5], Kazuaki Shimada[6], Tomoo Kosuge[6], Takuji Okusaka[7], Kazuto Kato[8], Jun Kusuda[9], Teruhiko Yoshida[4], Hiroyuki Aburatani[2] & Tatsuhiro Shibata[1]

Hepatocellular carcinoma, one of the most common virus-associated cancers, is the third most frequent cause of cancer-related death worldwide[1]. By massively parallel sequencing[2] of a primary hepatitis C virus–positive hepatocellular carcinoma (36× coverage) and matched lymphocytes (>28× coverage) from the same individual, we identified more than 11,000 somatic substitutions of the tumor genome that showed predominance of T>C/A>G transition and a decrease of the T>C substitution on the transcribed strand, suggesting preferential DNA repair. Gene annotation enrichment analysis[3] of 63 validated non-synonymous substitutions revealed enrichment of phosphoproteins. We further validated 22 chromosomal rearrangements, generating four fusion transcripts that had altered transcriptional regulation (*BCORL1-ELF4*) or promoter activity. Whole-exome sequencing[4,5] at a higher sequence depth (>76× coverage) revealed a *TSC1* nonsense substitution in a subpopulation of the tumor cells. This first high-resolution characterization of a virus-associated cancer genome identified previously uncharacterized mutation patterns, intra-chromosomal rearrangements and fusion genes, as well as genetic heterogeneity within the tumor.

We sequenced short-insert (250 bp, on average) genomic libraries of a primary hepatitis C virus (HCV)–positive hepatocellular carcinoma (HCC) and lymphocytes from a Japanese male (**Supplementary Fig. 1**) using the Illumina GAIIx sequencer with 50-bp paired-end reads. After alignment to the human reference genome and removal of PCR duplications, we obtained high-quality nucleotide sequences covering 102.5 Gb of the tumor genome (35.9× coverage) and 80.2 Gb (28.1× coverage) of the lymphocyte genome (**Supplementary Table 1**). The sequenced reads covered 99.69% (tumor) and 99.79% (lymphocyte) of the human reference genome. We identified 3,023,587 germline variations in the lymphocyte genome, approximately 90% of which were found in the dbSNP database, and 2,939,032 nucleotide variations in the tumor genome (a proportion of the variation was lost as a result of chromosomal alterations in the tumor genome). Comparison of the tumor and lymphocyte genomes revealed 11,731 somatically acquired nucleotide changes in the tumor genome (**Table 1**).

The prevalence of somatic substitutions was significantly less in the genic (intronic, non-coding exon and coding exon) regions relative to the intergenic regions (**Fig. 1a**, left), which could be partially explained by negative selection of lethal mutations in the gene regions or by the existence of specific molecules responsible for the repair of transcribed regions[6]. There was no significant difference in the prevalence of somatic substitutions between those of non-coding and coding exons (**Fig. 1a**, left), whereas the prevalence of germline variation was significantly decreased in the coding exons (**Fig. 1a**, right). Additionally, the ratio of non-synonymous to synonymous somatic substitutions (63/18 = 3.5) in the tumor genome was significantly higher than that of germline variations (9,573/10,552 = 0.91; $P < 0.0001$) but was not significantly different from that expected by chance (3.36; $P = 0.91$). This result suggests that an increase in negative selection of somatic substitution on the coding exons is weaker than that of germline variation. An alternative, but not mutually exclusive, explanation is that positive selection, which benefits the survival of tumor cells, partially occurs on the coding exons. The distribution of somatic substitutions revealed the dominance of T>C/A>G and C>T/G>A transitions (**Fig. 1b**). Sequence context preference was evident in some nucleotide substitutions. The C>T transition occurred significantly at CpG sites (15%; $P < 0.0001$), whereas the T>C transition occurred frequently at ApT sites (40%; $P < 0.0001$) (**Supplementary Fig. 2**). Only the T>C/A>G transition was significantly ($P = 0.01$) lower in the coding exons relative to the intergenic

[1]Division of Cancer Genomics, National Cancer Center Research Institute, Chuo-ku, Tokyo, Japan. [2]Genome Science Division, Research Center for Advanced Science and Technology, University of Tokyo, Meguro-ku, Tokyo, Japan. [3]Department of Pathology, Graduate School of Medicine, University of Tokyo, Bunkyo-ku, Tokyo, Japan. [4]Division of Genetics, National Cancer Center Research Institute, Chuo-ku, Tokyo, Japan. [5]Division of Molecular Pathology, National Cancer Center Research Institute, Chuo-ku, Tokyo, Japan. [6]Hepatobiliary and Pancreatic Surgery Division, National Cancer Center Hospital, Chuo-ku, Tokyo, Japan. [7]Hepatobiliary and Pancreatic Oncology Division, National Cancer Center Hospital, Chuo-ku, Tokyo, Japan. [8]Institute for Research in Humanities, Graduate School of Biostudies, Institute for Integrated Cell-Material Sciences, Kyoto University, Kyoto, Japan. [9]National Institute of Biomedical Innovation, Ibaraki, Osaka, Japan. Correspondence should be addressed to T. Shibata (tashibat@ncc.go.jp).

Received 21 July 2010; accepted 14 March 2011; published online 17 April 2011; doi:10.1038/ng.804

Table 1 Somatically acquired alterations in a liver cancer genome

Type of change	Number	Percentage
Substitutions	11,731	100.0
Coding	81	0.7
Nonsense	1	<0.1
Missense	62	0.5
Synonymous	18	0.2
Non-coding	120	1.0
UTR	83	0.7
Pseudogene	23	0.2
ncRNA	19	0.2
Intronic	4,001	34.1
Splice site	2	<0.1
Other	3,999	34.1
Intergenic	7,529	64.2
Small insertions and deletions	670	100.0
Coding	7	1.0
Non-coding	9	1.3
UTR	8	1.2
Pseudogene	0	0.0
ncRNA	2	0.3
Intronic	249	37.2
Splice site	0	0.0
Other	249	37.2
Intergenic	405	60.4
Rearrangements	22	100.0
Intrachromosomal	21	95.5
Deletions	11	50.0
Inversions	9	40.9
Tandem duplications	1	4.5
Interchromosomal	1	4.5

In 'non-coding' categories, some mutations have been classified into two subgroups. Four substitutions were classified as both UTR and non-coding RNA. One substitution was classified as both a pseudogene and non-coding RNA. One indel was classified as both UTR and non-coding RNA. UTR, untranslated region; ncRNA, non-coding RNA.

regions (**Fig. 1c**), and the C>T/G>A transition was more frequent in the coding exons relative to the intronic and non-coding exon regions, partly due to the higher GC content of coding exons and the higher frequency of CpG methylation. There were fewer T>C transitions on the transcribed strands than on the untranscribed strands ($P < 0.0001$) (**Fig. 1d**), and we observed no statistically significant differences for other substitutions.

We detected 90 somatic substitutions in protein-coding regions, 81 (including 63 non-synonymous substitutions) of which were validated as somatic alterations by Sanger sequencing of both the tumor and lymphocyte genomes (**Tables 1**,**2** and **Supplementary Fig. 3**). Of the remaining nine substitutions, three could not be amplified by PCR, four could not be sequenced due to the surrounding repetitive sequences and two could not be validated, likely because they were located within highly homologous segmental duplications or processed pseudogene regions. We also found evidence for 670 small somatic insertions and deletions, and all seven that are located in protein-coding regions were validated (**Tables 1** and **2**, **Supplementary Fig. 13**). These somatic alterations included mutations of two well-known tumor suppressor genes for HCC (*TP53* and *AXIN1*) and five genes (*ADAM22*, *JAK2*, *KHDRBS2*, *NEK8* and *TRRAP*) that have been found to be mutated in other cancers[7]. Gene annotation enrichment analysis[3] of the non-synonymous somatic mutations revealed significant overrepresentation of genes encoding phosphorproteins ($P = 0.0017$) and those with bipartite nuclear localization signals ($P = 0.029$) (**Supplementary Table 2**). Further re-sequencing of the exons containing potentially deleterious mutations in 96 additional pairs of primary HCC and non-cancerous liver and 21 HCC cell lines revealed two mutations (resulting in p.Phe190Leu and p.Gln212X, of which only the latter was proven to be somatic) in *LRRC30* (**Supplementary Fig. 4**). *LRRC30* contains nine repeats of a leucine-rich domain of unknown function, and all validated mutations changed the well-conserved amino acid in these repeats or produced a truncated protein.

We predicted 33 somatic rearrangements, 22 of which were validated by Sanger sequencing of the breakpoints in both the tumor and lymphocyte genomes (**Table 3**). Most of the rearrangements were intrachromosomal and occurred at the boundaries of copy number change (**Supplementary Fig. 5**). In particular, nine structural aberrations were clustered in the region of 11q12.2–11q13.4, generating a complex pattern of chromosomal amplification and loss (**Supplementary Fig. 6**). RT-PCR and sequencing analysis of the tumor and matched non-cancerous liver tissue validated four somatic fusion transcripts generated by rearrangements: the *BCORL1-ELF4* and *CTNND1-STX5* fusion genes by intra-chromosomal inversions (Xq25 and 11q12, respectively), the *VCL-ADK* fusion gene by an interstitial deletion in 10q22 (**Supplementary Fig. 7**) and the *CABP2-LOC645332* fusion gene by a tandem duplication in 11q13 (**Supplementary Fig. 8**). The *BCORL1-ELF4* chimeric transcript combining exons 1–11 of *BCORL1* and exon 8 of *ELF4* encodes an in-frame fusion protein (**Fig. 2a**,**b**). Quantitative RT-PCR revealed increased (>sixfold) expression of fusion transcripts in the tumor relative to wild-type *BCORL1* and *ELF4* gene expression in the non-cancerous liver (data not shown). *BCORL1* associates with CtBP and class II histone deacetylases and functions as a transcriptional repressor[8], and *ELF4* encodes a transcriptional activator[9,10] (**Fig. 2b**). We expressed *BCORL1*, *ELF4* and the chimera *BCORL1-ELF4* as Gal4-DBD fusion proteins and evaluated their transcriptional activities using a luciferase reporter assay. The chimeric protein had reduced repression activity compared to wild-type *BCORL1* (**Fig. 2c**). For the *CTNND1-STX5* fusion gene, the combination of non-coding exon 1 of *CTNND1* and exons 3–11 of *STX5* resulted in the deletion of 96 amino acids at the terminal end of *STX5* and increased (>twofold) *STX5* gene expression in the tumor,

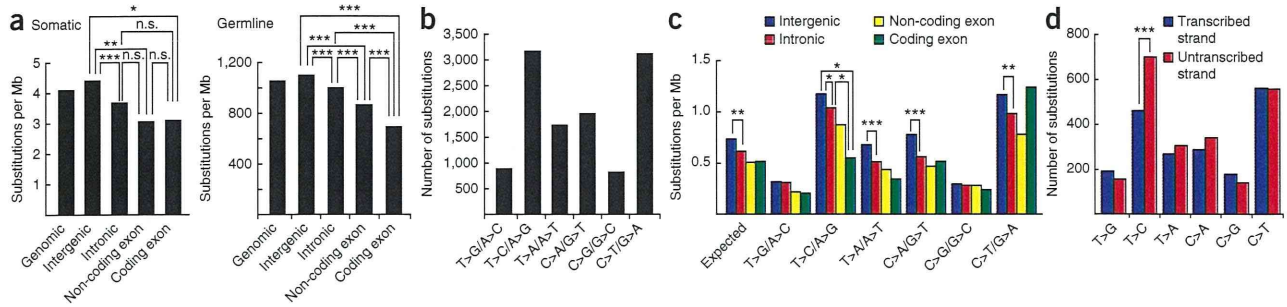

Figure 1 Somatic substitution pattern of the liver cancer genome. (**a**) Prevalence of somatic and germline substitutions in different genome regions. (**b**) Number of each type of somatic substitution in the liver cancer genome. (**c**) Prevalence of each type of somatic substitution in different genome regions. (**d**) Number of each type of somatic substitution on the transcribed and untranscribed strands. $*P < 0.05$, $**P < 0.01$, $***P < 0.0001$.

Table 2 Validated somatic non-synonymous substitutions and small indels in coding regions of a liver cancer genome

Gene	Chr.	Strand	Position	Allele change	Amino acid change	Copy number	Mutant allele (%) in whole-genome sequencing	Mutant allele (%) in whole-exome sequencing	Expression ratio (T/N)	Functional
PLEKHG5	1	–	6,452,224	G>T	Asp>Tyr	N	49.0	27.7	1.86	Deleterious
KIAA1026	1	+	15,294,007	C>A	Ala>Glu	N	45.7	nd	0.15	Tolerated
MYCL1	1	–	40,139,080	T>G	Phe>Cys	N	54.5	nd	1.93	Tolerated
PDE4B	1	+	66,231,185	C>A	Ala>Glu	N	57.1	42.9	0.83	Tolerated
CLCC1	1	–	109,284,236	A>G	Tyr>Cys	N	33.3	39.3	1.61	Deleterious
CNRIP1	2	–	68,397,833	C>T	Thr>Met	N	40.0	33.3	1.39	Deleterious
ANKRD36	2	+	97,181,397	A>G	Lys>Glu	N	17.8	nd	9.49	Tolerated
UBR3	2	+	170,511,073	A>C	Glu>Asp	N	57.1	nd	18.10	Tolerated
CUL3	2	–	225,070,790	G>A	Ser>Asn	N	42.9	52.8	12.80	Tolerated
COPS7B	2	+	232,369,129	A>G	Ile>Val	N	44.4	41.5	1.82	Tolerated
RAF1	3	–	12,625,811	A>G	Asn>Ser	N	40.0	50.0	2.31	Tolerated
ITIH3	3	+	52,813,002	A>G	Met>Val	N	43.9	nd	1.25	Deleterious
ERC2	3	–	56,148,636	G>C	Glu>Gln	N	40.0	nd	1.33	Tolerated
TBC1D23	3	+	101,496,868	del AAG	Deletion (E)	N	14.8	nd	4.90	na
ATR	3	–	143,671,657	del AT	Deletion (frame shift)	N	20.0	nd	4.49	na
SLC7A14	3	–	171,701,666	G>A	Ser>Asn	N	52.8	46.3	2.19	Deleterious
PCDH7	4	+	30,333,134	G>A	Arg>His	N	47.1	47.8	1.74	Tolerated
FAM13A	4	–	89,872,188	A>T	His>Leu	N	52.0	47.4	0.85	Tolerated
MFSD8	4	–	129,090,435	A>T	Met>Leu	Loss	62.5	74.3	1.15	Tolerated
DMGDH	5	–	78,375,996	T>A	Leu>Gln	N	50.0	37.6	3.04	Tolerated
PCDHA13	5	+	140,244,063	C>T	Pro>Ser	N	45.1	34.8	na	Deleterious
CCDC99	5	+	168,960,950	T>G	Ser>Arg	N	37.1	39.4	13.30	Deleterious
GABBR1	6	–	29,706,345	C>T	Thr>Met	N	42.0	37.8	0.59	Tolerated
CSNK2B	6	+	31,745,659	A>T	Ser>Cys	N	37.3	nd	1.41	Deleterious
MOCS1	6	–	40,003,210	G>T	Ser>Ile	N	34.4	nd	1.54	Tolerated
GTPBP2	6	–	43,699,685	A>T	Glu>Val	N	58.0	56.3	1.36	Tolerated
KHDRBS2	6	–	62,662,692	G>T	Arg>Leu	N	34.1	nd	0.88	Deleterious
SLC29A4	7	+	5,303,324	A>T	His>Leu	N	43.8	nd	7.00	Deleterious
TMEM195	7	–	15,567,887	C>G	Pro>Ala	N	41.2	38.3	1.03	Deleterious
RFC2	7	–	73,302,032	A>T	Glu>Asp	N	26.0	41.9	1.09	Tolerated
ADAM22	7	+	87,653,951	A>T	Arg>Trp	N	41.2	39.1	0.55	Deleterious
TRRAP	7	+	98,417,359	G>T	Trp>Leu	N	39.0	nd	2.07	Deleterious
XRCC2	7	–	151,977,231	G>A	Arg>Gln	N	56.2	36.5	4.18	Deleterious
MTDH	8	+	98,781,211	G>T	Val>Phe	N	33.3	46.9	14.40	Tolerated
SLA	8	–	134,141,539	C>A	Pro>Thr	N	43.6	nd	1.18	Deleterious
JAK2	9	+	5,045,703	T>G	Ile>Ser	Loss	100.0	84.2	4.84	Tolerated
NTRK2	9	+	86,532,391	G>A	Ala>Thr	Loss	90.0	85.9	0.84	Tolerated
TSC1	9	–	134,767,848	C>T	Arg>stop	Loss	13.3	13.0	1.85	Deleterious
CREM	10	+	35,496,706	A>G	Glu>Gly	N	44.8	42.3	3.28	Tolerated
C10orf95	10	–	104,200,839	T>C	Cys>Arg	N	39.7	nd	3.05	Tolerated
PSTK	10	+	124,730,061	C>T	Leu>Phe	N	53.6	nd	6.94	Deleterious
ATHL1	11	+	283,903	C>T	Ala>Val	N	40.9	26.8	1.12	Tolerated
MUC5B	11	+	1,213,214	G>T	Val>Leu	N	33.8	nd	0.83	Tolerated
DENND5A	11	–	9,181,879	C>T	Pro>Ser	N	21.4	29.9	2.43	Deleterious
GIF	11	–	59,369,438	C>T	Thr>Ile	AMP (3)	29.2	nd	0.83	Tolerated
STIP1	11	+	63,719,763	G>A	Glu>Lys	Loss	66.7	nd	1.28	Tolerated
FAT3	11	+	91,727,805	C>G	Thr>Ser	Loss	73.1	nd	na	Tolerated
PTMS	12	+	6,749,421	A>G	Glu>Gly	Loss	55.0	nd	0.56	Tolerated
ARID2	12	+	44,530,716	ins T	Insertion (frame shift)	N	31.9	nd	2.35	na
C12orf51	12	–	111,134,825	del CCTGCCACGTCA	Deletion (GDVA)	N	21.6	nd	1.44	Torelated
RBM19	12	–	112,868,641	C>T	Pro>Leu	N	49.3	42.2	1.32	Deleterious
AACS	12	+	124,142,015	G>T	Gly>Val	N	34.9	26.0	1.75	Deleterious
KHNYN	14	+	23,971,333	del CCT	Deletion (L)	N	24.1	nd	2.17	Tolerated
NOVA1	14	–	25,987,233	A>T	Leu>Phe	N	36.7	38.1	0.91	Tolerated
LTBP2	14	–	74,045,780	G>A	Gly>Glu	N	38.1	nd	3.43	Deleterious
CYFIP1	15	+	20,498,517	C>T	Ala>Val	N	55.1	41.4	1.88	Deleterious
GABRB3	15	–	24,357,328	G>T	Met>Ile	N	39.4	43.4	0.15	Tolerated
EID1	15	+	46,957,688	C>G	Ser>Cys	N	40.4	nd	8.60	Deleterious
HCN4	15	–	71,402,254	G>A	Arg>His	N	43.6	nd	0.61	Tolerated
AKAP13	15	+	84,060,152	del T	Deletion (frame shift)	N	34.5	nd	0.88	na
AXIN1	16	–	287,910	C>T	Arg>stop	Loss	78.7	nd	0.94	Deleterious
LITAF	16	–	11,554,943	del G	Deletion (frame shift)	Loss	61.3	nd	0.97	na
TP53	17	–	7,518,985	G>T	Val>Leu	Loss	78.0	73.1	0.06	Deleterious
NEK8	17	+	24,092,271	G>A	Gly>Asp	N	36.7	39.1	1.44	Deleterious
CPD	17	+	25,773,820	A>G	Tyr>Cys	N	47.1	52.3	2.28	Deleterious
LRRC30	18	+	7,221,594	C>G	Ser>Cys	N	52.0	45.6	na	Deleterious
ZNF560	19	–	9,439,794	A>C	Ile>Leu	N	58.8	48.3	0.86	Tolerated
SCRT2	20	–	593,073	T>A	Tyr>Asn	N	53.7	nd	0.51	Deleterious
USP25	21	+	16,119,227	C>T	Thr>Met	N	44.4	nd	13.00	Deleterious
USP25	21	+	16,125,626	A>C	Glu>Asp	N	35.3	38.1	na	Tolerated
ARVCF	22	–	18,341,717	C>G	Ser>Cys	N	53.0	50.0	1.30	Deleterious
USP26	X	–	131,988,824	T>C	Leu>Pro	AMP (4)	93.8	94.4	0.85	Tolerated

Except for *ANKRD36* and *TSC1*, all 63 somatic non-synonymous substitutions were predicted by whole-genome sequencing and in-house informatics method using stringent analysis criteria (Online Methods). One somatic missense substitution in *ANKRD36* was predicted under less stringent criteria. One somatic nonsense substitution in *TSC1* was predicted only by whole-exome sequencing. Chr., chromosome; N, copy neutral; AMP, amplicon; nd, not detected; na, not applicable.

Table 3 Validated somatic structural alterations in a liver cancer genome

Type	Chr. A	Break point A	CNV (Chr. A)	Chr. B	Break point B	CNV (Chr. B)	Intervening sequence	Associated genes	Fusion genes
Deletion	3	111,866,468	BCNC	3	111,868,894	BCNC	0		
Deletion	4	57,529,004	BCNC	4	57,530,452	BCNC	0	*C4orf14* (exon 4 is deleted)	
Deletion	4	92,895,135	BCNC	4	93,151,201	BCNC	0		
Deletion	5	18,130,563	BCNC	5	18,133,946	BCNC	(+) 29bp		
Deletion	6	90,130,109	BCNC	6	90,819,100	BCNC	0	*LYRM2, ANKRD6, BACH2, MDN1, CASP8AP2, RRAGD, GJA10*	
Deletion	7	69,321,043	N	7	69,404,639	N	0	*AUTS2*	
Deletion	9	132,763,157	BCNC	9	132,764,920	BCNC	0		
Deletion	10	75,477,784	BCNC	10	75,956,310	BCNC	(+) 1 bp	*AP3M1, VCL, ADK*	*VCL, ADK*
Deletion	11	67,126,436	BCNC	11	68,254,241	BCNC	0	*SUV420H1, SAPS3, ACY3, ALDH3B2, CHKA, TCIRG1, LRP5, GAL, ALDH3B1, TBX10, NDUFV1, UNC93B1, NUDT8, C11orf24*	
Deletion	15	47,394,203	BCNC	15	47,467,920	BCNC	0	*GALK2, C15orf33*	
Deletion	17	15,902,440	BCNC	17	16,056,159	BCNC	0	*NCOR1* (homozygous deletion)	
Inversion	4	60,946,299	N	4	60,947,151	N	0		
Inversion	4	172,703,199	Loss	4	172,706,239	Loss	(+) 4bp		
Inversion	11	57,305,269	BCNC	11	62,352,275	BCNC	0	*CTNND1* (UTR), *STX5*	*CTNND1, STX5*
Inversion	11	57,770,822	BCNC	11	67,133,985	BCNC	0	*NDUFV1*	
Inversion	11	62,309,952	BCNC	11	70,746,006	BCNC	0	*TAF6L*	
Inversion	11	69,067,231	AMP	11	69,317,424	AMP	0		
Inversion	11	69,093,978	AMP	11	69,098,117	AMP	0		
Inversion	11	69,871,206	AMP	11	69,877,391	AMP	(+) 6bp	*PPFIA1*	
Inversion	X	129,015,072	N	X	129,029,501	BCNC	(+) 23bp	*BCORL1, ELF4*	*BCORL1, ELF4*
Inversion	X	129,016,981	N	X	129,031,425	BCNC	0	*BCORL1, ELF4*	*BCORL1, ELF4*
Tandem duplication	11	67,043,308	BCNC	11	67,318,685	BCNC	0	*ACY3, ALDH3B2, GSTP1, TBX10, NDUFV1, NUDT8, CABP2, LOC645332*	*CABP2, LOC645332*
Translocation	11	69,316,960	AMP	X	129,030,346	BCNC	0	*ELF4*	

The inversions at Xq25 occurred from one rearrangement event and the total number of inversion is counted as nine. Chr., chromosome; BCNC, boundary of copy number change; N, copy neutral; AMP, amplicon.

which harbors only the rearranged allele (**Fig. 2d** and **Supplementary Fig. 9**). We screened for the presence of these four chimera transcripts by RT-PCR, but we detected no recurrent fusion event in 47 cases of primary HCC, possibly due to the low frequency of these rearrangements in HCC or because of the technical difficulty in detecting all variant fusion transcripts.

We also sequenced the whole exomes of the same samples using an in-solution gene enrichment system[5] (**Fig. 3a**). Capture probes for whole-exome sequencing were designed to cover the protein coding exons using the consensus coding sequences, excluding highly homologous regions. The average coverage of the whole exome sequences (41.3 Mb in total) was about twice (76.8× for HCC and 74.3× for lymphocytes) that of the whole genome sequences and had one twelfth of the total sequence amount (8.9 Gb for HCC and 8.6 Gb for lymphocyte) (**Supplementary Table 3**). Whole-exome sequencing detected 47 non-synonymous somatic substitutions, 40 of which were validated by Sanger sequencing. Among the validated substitutions, a nonsense substitution (p.Arg785X) in *TSC1*, located in the hemizygous region (9q34), was not detected by whole-genome sequencing (**Fig. 3b**). Capillary sequencing validated the same substitution with a very low

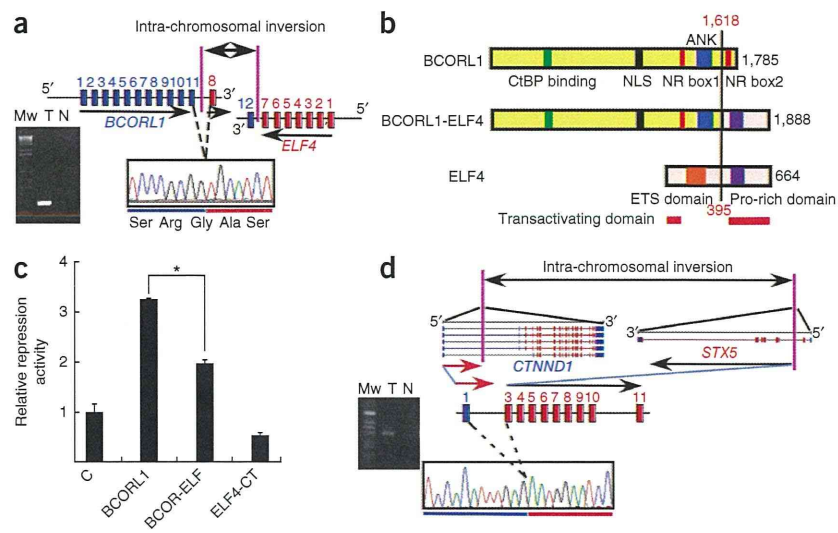

Figure 2 Characterization of rearrangements in liver cancer. (**a**) Top, schematic representation of the intra-chromosomal inversion at Xq25. Bottom left, RT-PCR analysis of the fused *BCORL1-ELF4* transcript in tumor (T) and non-cancerous liver (N) tissues. We detected no *ELF4-BCORL1* transcript (data not shown). Bottom right, sequence chromatography of the fusion transcript revealed an in-frame protein. Mw, molecular marker. (**b**) Schematic representation of the BCORL1-ELF4 fusion protein. BCORL1 (top) contains a CtBP1 binding domain (PXDLS sequence), a binuclear localization signal (NLS), two LXXLL nuclear receptor recruitment motifs (NR box) and tandem ankyrin repeats (ANK). ELF4 (bottom) contains an ETS (E Twenty Six) DNA binding domain and a proline-rich domain. Transactivating domains are indicated by the red bars[16]. The BCORL1-ELF4 chimeric protein includes most of BCORL1 (1–1,618 amino acids) lacking the NR box2 and the carboxyl-terminal portion of ELF4 containing the proline-rich domain. The number of amino acids is indicated on the right. (**c**) Wild-type BCORL1, ELF4-CT (395–664 amino acids) and the BCORL1-ELF4 chimera were expressed as Gal4-DBD fusion proteins, and their relative transcriptional activities were compared to the Gal4-DBD protein (C) as shown. (**d**) Characterization of the *CTNND1-STX5* fusion gene. Bottom left, RTPCR analysis of the fused *CTNND1-STX5* transcript in tumor (T) and non-cancerous liver tissue (N). Bottom right, sequence chromatography of the fusion transcript. Data is the mean ± s.d. ($n = 3$). *$P < 0.001$.

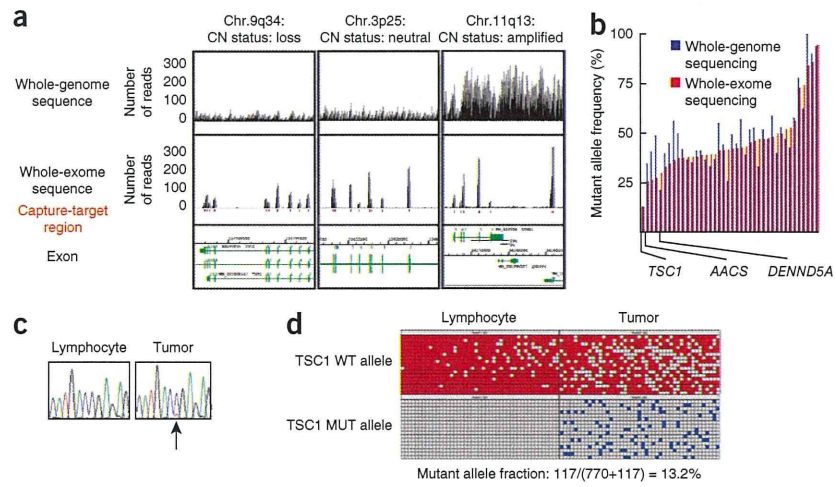

Figure 3 Intra-tumoral genetic heterogeneity detected by exon-capture sequencing. (**a**) Specific enrichment and high sequence coverage of the target genome regions indicated by the sequence viewer (copy number (CN) status is shown above). The distribution and number of reads (black, forward read; gray, reverse read) from whole-genome sequencing (top) and whole-exome sequencing (middle) are shown. The location of the capture target regions (red box) and the exons (green box) along the genome are shown at the bottom. Note that the number of reads is dependent on copy number status. (**b**) Mutant allele frequency detected by whole-genome sequencing and whole-exome sequencing. *TSC1*, *AACS* (whose heterogeneity is shown in **Supplementary Fig. 10**) and *DENND5A* are indicated. (**c**) *TSC1* mutation in the liver cancer subpopulation. Sequence chromatograms of *TSC1* in lymphocytes and whole-tumor tissue are shown. Note the small peak for the mutant T allele (indicated by the arrow) in the tumor DNA. (**d**) Determination of mutant *TSC1* allele frequency by digital PCR genotyping. WT, wild type; MUT, mutant.

signal peak (**Fig. 3c**), and digital genotyping showed that 13.2% of the tumor alleles harbored this substitution (**Fig. 3d**), suggesting that this substitution occurred in a minor population of cancer cells. Whole-exome sequencing missed 25 non-synonymous somatic substitutions that were detected by whole-genome sequencing. These missed substitutions were located in regions where sequence coverage was low or where further optimization of the probe design was required.

The number of non-synonymous somatic substitutions validated in this HCC (63) was greater than those for acute myeloid leukemia[11] (10), basal-like breast cancer[12] (22), lobular carcinoma[13] (32), glioblastoma multiforme[14] (32) and pancreatic cancer[15] (43) but is in the range of those previously reported for colorectal[16] (70) and breast[16] (88) cancer. We have shown that the pattern of somatic substitutions in a HCV-associated HCC genome is different (predominance of T>C, especially at ApT sites, and C>T, especially at CpG sites) compared to smoking-related[17,18] and ultraviolet light–related[6] cancers. Preferential C>T/G>A transition may partly be due to the higher frequency of CpG methylation in the genome sequence and is a common form of mutation in cancers[19]. Therefore, the T>C/A>G transition could be a characteristic mutational signature of HCV-associated cancer, which would be consistent with a previous observation that HCV induces error-prone DNA polymerases that preferentially cause the T>C/A>G mutation[20]. It is also possible that this mutation pattern is independent of viral infection and is organ specific, as a comparable substitution spectrum has been reported in renal cancer[19]. Additionally, only T>C changes, but not C>T changes, were effectively repaired on the transcribed strand. Similar enhanced transcription-coupled repair on preferentially acquired substitutions has been reported in other cancers[6,17,18] and could be a common phenomenon in cancer mutation.

Because single-molecule sequencing has the capability to detect every individual somatic event in parallel, higher sequence coverage will enable us to clarify the intra-tumoral heterogeneity that is associated with diverse aspects of clinical behavior such as metastasis[21]. The *TSC1* complex, which is inactivated in a subpopulation of tumors, negatively regulates the mammalian target of rapamycin signaling, which is an important oncogenic pathway related to the growth, metabolism and stemness of cancer cells[22,23], and could be a promising molecular therapeutic target in HCC progression[24].

URLs. International Cancer Genome Consortium, http://www.icgc.org/; Catalogue of Somatic Mutations in Cancer, http://www.sanger.ac.uk/genetics/CGP/cosmic/; BLASTN, ftp://ftp.ncbi.nlm.nih.gov/blast/executables/release/LATEST.

METHODS

Methods and any associated references are available in the online version of the paper at http://www.nature.com/naturegenetics/.

Note: Supplementary information is available on the Nature Genetics website.

ACKNOWLEDGMENTS

We thank K.K. Khanna (The Queensland Institute of Medical Research) for providing a human *BCORL1* cDNA clone; T.D. Taylor (RIKEN) for comments on the manuscript; T. Urushidate, S. Ohashi, S. Ohnami, A. Kokubu, N. Okada, K. Shiina, H. Meguro and K. Nakano for their excellent technical assistance. This work was supported by the Program for Promotion of Fundamental Studies in Health Sciences of the National Institute of Biomedical Innovation (NIBIO), Japan, and the Industrial Technology Research Grant Program from the New Energy and Industrial Technology Development Organization (NEDO), Japan. This study is associated with the International Cancer Genome Consortium (ICGC), and the mutation data were deposited at and released from the ICGC web site.

AUTHOR CONTRIBUTIONS

The study was designed by T. Shibata, H.A., T.Y. and J.K. Sequencing and data analyses were conducted by Y.T., K.T., S.Y., S.T., K. Sonoda and H.T. Allele typing and copy number analyses were performed by H.S. and S.I. Other molecular studies were done by Y.A., F.H., T. Shirakihara, and L.W.; H.O., K. Shimada, T.K., T.O. and K.K. coordinated collection of clinical sample and information. The manuscript was written by Y.T., T. Shibata, K.T., S.Y., H.A. and T.Y.

COMPETING FINANCIAL INTERESTS

The authors declare no competing financial interests.

Published online at http://www.nature.com/naturegenetics /.
Reprints and permissions information is available online at http://npg.nature.com/reprintsandpermissions/ .

1. El-Serag, H.B. & Rudolph, K.L. Hepatocellular carcinoma: epidemiology and molecular carcinogenesis. *Gastroenterology* **132**, 2557–2576 (2007).
2. Bentley, D.R. *et al.* Accurate whole human genome sequencing using reversible terminator chemistry. *Nature* **456**, 53–59 (2008).
3. Huang, W., Sherman, B.T. & Lempicki, R.A. Systematic and integrative analysis of large gene lists using DAVID Bioinformatics Resources. *Nat. Protoc.* **4**, 44–57 (2009).
4. Ng, S.B. *et al.* Targeted capture and massively parallel sequencing of 12 human exomes. *Nature* **461**, 272–276 (2009).

5. Gnirke, A. *et al.* Solution hybrid selection with ultra-long oligonucleotides for massively parallel targeted sequencing. *Nat. Biotechnol.* **27**, 182–189 (2009).
6. Pleasance, E.D. *et al.* A comprehensive catalogue of somatic mutations from a human cancer genome. *Nature* **463**, 191–196 (2010).
7. Xiang, Z. *et al.* Identification of somatic *JAK1* mutations in patients with acute myeloid leukemia. *Blood* **111**, 4809–4812 (2008).
8. Pagan, J.K. *et al.* A novel corepressor, BCoR-L1, represses transcription through an interaction with CtBP. *J. Biol. Chem.* **282**, 15248–15257 (2002).
9. Miyazaki, Y., Sun, X., Uchida, H., Zhang, J. & Nimer, S. MEF, a novel transcription factor with an Elf-1 like DNA binding domain but distinct transcriptional activating properties. *Oncogene* **13**, 1721–1729 (1996).
10. Suico, M.A. *et al.* Functional dissection of the ETS transcription factor MEF. *Biochim. Biophys. Acta* **1577**, 113–120 (2002).
11. Mardis, E.R. *et al.* Recurring mutations found by sequencing an acute myeloid leukemia genome. *N. Engl. J. Med.* **361**, 1058–1066 (2009).
12. Ding, L. *et al.* Genome remodelling in a basal-like breast cancer metastasis and xenograft. *Nature* **464**, 999–1005 (2010).
13. Shah, S.P. *et al.* Mutation evolution in a lobular breast tumour profiled at single nucleotide resolution. *Nature* **461**, 809–813 (2009).
14. Parsons, D.W. *et al.* An integrated genomic analysis of human glioblastoma multiforme. *Science* **321**, 1807–1812 (2008).
15. Jones, S. *et al.* Core signaling pathways in human pancreatic cancers revealed by global genomic analyses. *Science* **321**, 1801–1806 (2008).
16. Wood, L.D. *et al.* The genomic landscapes of human breast and colorectal cancers. *Science* **318**, 1108–1113 (2007).
17. Pleasance, E.D. *et al.* A small-cell lung cancer genome with complex signatures of tobacco exposure. *Nature* **463**, 184–190 (2010).
18. Lee, W. *et al.* The mutation spectrum revealed by paired genome sequences from a lung cancer patient. *Nature* **465**, 473–477 (2010).
19. Greenman, C. *et al.* Patterns of somatic mutation in human cancer genomes. *Nature* **446**, 153–158 (2007).
20. Machida, K. *et al.* Hepatitis C virus induces a mutator phenotype: enhanced mutations of immunoglobulin and protooncogenes. *Proc. Natl. Acad. Sci. USA* **101**, 4262–4267 (2004).
21. Kim, M.Y. *et al.* Tumor self-seeding by circulating cancer cells. *Cell* **139**, 1315–1326 (2009).
22. Guertin, D.A. & Sabatini, D.M. Defining the role of mTOR in cancer. *Cancer Cell* **12**, 9–22 (2007).
23. Yilmaz, O.H. *et al.* Pten dependence distinguishes haematopoietic stem cells from leukaemia-initiating cells. *Nature* **441**, 475–482 (2006).
24. Meric-Bernstam, F. & Gonzalez-Angulo, A.M. Targeting the mTOR signaling network for cancer therapy. *J. Clin. Oncol.* **27**, 2278–2287 (2009).

ONLINE METHODS

Whole-genome sequencing. High molecular weight DNA was extracted from freshly frozen tumor tissue and lymphocytes. DNA was fragmented using an ultrasonic solubilizer (Covaris) using a combination of quick bursts (20% duty, 5 intensity with 200 cycles per burst for 5 s) and sonication (10% duty, 5 intensity with 200 cycles per burst for 120 s) for the short fragment DNA library. DNA of the appropriate size was gel purified to exclude any inappropriate DNA fusions during library construction. The short fragment DNA libraries were generated using a paired-end DNA sample prep kit (Illumina) following the manufacturer's protocols. The concentration of the libraries was quantified using a Bioanalyzer (Agilent Technologies); 4~8 pM/lane of DNA was applied to the flow cell, and paired-end sequencing was performed using the GAIIx sequencer (Illumina).

Whole-exome capture sequencing. Whole-exome capture sequencing was performed using the SureSelect Target Enrichment System (Agilent Technologies) in accordance with the manufacturer's protocol with slight modifications. Briefly, the same Illumina sequence libraries as those prepared for the whole-genome sequence were amplified with six cycles of PCR, and then 500 ng of the amplified libraries was hybridized with the capture probes for 24 h. The hybridized sequence libraries were collected and further amplified with 14 cycles of PCR. We generated 51-nucleotide–long paired-end reads using the GAIIx sequencer (Illumina). We used five lanes of a paired-end flow cell for each sample.

Bioinformatics (Supplementary Fig. 11). *Sequence alignment to the human genome and removal of PCR duplications.* Paired-end reads were aligned to the human reference genome (hg18, NCBI Build 36.1) using Burrows-Wheeler Aligner (BWA) (version 0.4.9)[25]. Because there were duplicated reads which were generated during the PCR amplification process, paired-end reads that aligned to the same genomic positions were removed using SAMtools (version 0.1.5c)[26] and a program developed in house. We removed 12.5% (14.6/117.1 Gbp) of the aligned reads for tumor and 7.1% (6.1/86.3 Gbp) for lymphocytes.

Detection of somatic single nucleotide variations (SNVs) **(Supplementary Fig. 12).** Based on the genotyping data from two SNP arrays, appropriate thresholds for base quality, mapping quality and frequency of non-reference alleles were determined to obtain the highest confidence calls for SNV detection (**Supplementary Table 4**). To predict somatic SNVs, the alignment results were classified, and three datasets were constructed. Dataset 1 included paired-end reads with both ends aligned uniquely and with proper spacing and orientation. Dataset 2 included paired-end reads that aligned uniquely for at least one read and with proper spacing and orientation of the reads. Dataset 3 included dataset 2 and paired-end reads for which both ends aligned uniquely but with improper spacing or orientation or both. Dataset 1 likely contains false positive somatic SNVs because of the low sequence depth of the lymphocyte genome, and dataset 3 likely contains false positives due to misalignments of the sequence reads. To reduce the number of false positives, the following filters were applied to these three datasets, and concordant somatic SNVs among the three datasets were selected: (i) a mapping quality score of 20 was used as a cutoff value for read selection; (ii) base quality scores of 10 and 15 were used as cutoff values for base selection for the tumor and lymphocyte genomes, respectively; (iii) SNVs were selected when the frequency of the non-reference allele was at least 15% in the tumor genome and 5% in the lymphocyte genome; (iv) SNVs located within 5 bp from a potential insertion or deletion were discarded; (v) SNVs with a root mean square mapping quality score of the reads covering the SNV less than 40 were discarded; (vi) when there were three or more SNVs within any 10-bp window, all of them were discarded; (vii) SNVs with a consensus quality score less than 20 as calculated by SAMtools (version 0.1.5c) were discarded; (viii) when a base with a consensus quality score less than 20 was located within 3-bp on either side of a SNV, the SNV was discarded; (ix) for the tumor genome, SNVs found in at least two sequence reads with the same SNV were selected; (x) for the lymphocyte genome, SNVs covered by at least six sequence reads were selected; and (xi) the repetitive regions within 1 Mb of a centromeric or telomeric sequence gap were excluded. By comparing the predicted nucleotide variations in the tumor and lymphocyte genomes, somatic SNVs which occurred only in the tumor genome were identified. If somatic SNVs were not covered in the lymphocyte genome by at least six sequence reads, they were discarded.

Using this approach, 66 non-synonymous and 24 synonymous somatic SNVs in protein-coding regions were predicted. These 90 substitutions were examined by Sanger sequencing of both the tumor and lymphocyte genomes, and 81 of them were validated as somatic mutations. Of the remaining nine substitutions, three could not be amplified by PCR, four could not be sequenced because of the surrounding repetitive sequences, and two could not be validated likely because they were located in highly homologous segmentally duplicated or processed pseudogene regions, suggesting a high prediction accuracy (specificity, 81/83 = 97.6%) for our approach for detecting somatic SNVs in protein-coding regions. An additional 36 non-synonymous somatic SNVs were also predicted using only dataset 3 and filtering methods (i–iv) (less stringent filtering condition). Five of these SNVs were not validated and 30 of them were found to be germline variations by Sanger sequencing, and only the one remaining was validated as a somatic mutation. These findings suggest that our filtering method (stringent condition) effectively removed false-positive somatic SNVs.

Detection of somatic structural alterations. To detect structural alterations, paired-end reads for which both ends aligned uniquely to the human reference genome, but with improper spacing or orientation or both, were used. First, paired-end reads were selected based on the following filtering conditions: (i) sequence reads with mapping quality scores greater than 37; and (ii) sequence reads aligned with two mismatches or less.

Rearrangements were then identified using the following analytical conditions: (i) 'clusters' which included reads aligned within the maximum insert distance were constructed from the forward and reverse alignments, respectively (two reads were allocated to the same cluster if their end positions were not further apart than the maximum insert distance); (ii) clusters whose distance between the leftmost and rightmost reads were greater than the maximum insert distance were discarded; (iii) paired-end reads were selected if one end sequence was allocated in the 'forward cluster' and the other end was allocated in the 'reverse cluster' (we called these 'forward cluster and reverse cluster' paired clusters); (iv) if a cluster overlapped another cluster, all of the overlapping paired-clusters were discarded; (v) for the tumor genome, rearrangements (paired-clusters) predicted by at least four paired-end reads which included at least one paired-end read perfectly matched to the human reference genome were selected; and (vi) for the lymphocyte genome, rearrangements (paired clusters) predicted by at least one paired-end read were selected. By comparing the predicted rearrangements in the tumor and lymphocyte genomes, somatic rearrangements that were only detected in the tumor genome were identified.

Lastly, rearrangements predicted due to variations in the analyzed genomes were removed. For this analysis, paired-end reads contained in paired clusters were aligned to the human reference genome using the BLASTN program (see URLs). If one end sequence was aligned to the region of paired clusters (the flanking region of the rearrangement breakpoint) and the other end was aligned with proper spacing and orientation, the rearrangement was removed. An expectation value of 1,000 was used as a cutoff value for BLASTN so that paired-end reads with low similarity to the human reference genome could also be aligned.

Using this method, 33 somatic rearrangements were predicted and 22 of these were validated by Sanger sequencing of the rearrangement breakpoints in both the tumor and lymphocyte genomes.

Exome capture sequence analysis. To analyze the capture sequencing data, the Illumina sequencing pipeline version 1.4 and in-house programs were used. The sequence reads were mapped to the human reference sequence (NCBI Build 36.3) using GERALD (Illumina), and only high-quality ('pass filter') reads with base-call quality scores more than ten were used for SNV detection.

doi:10.1038/ng.804

SNVs were determined using the frequency (>20%) of the highest non-reference base call with a read depth greater than 20×.

Other molecular analyses. SNP genotyping and copy number detection were determined using the Affymetrix Mapping 500K Array, the Agilent Human Genome CGH microarray and the Illumina Human 610-Quad BeadChip system. Gene expression levels of the tumor were measured using the Agilent Whole Human Genome Oligo Microarray. Wild-type and mutant allele frequencies were determined using the Digital PCR system.

Detailed experimental methods and additional bioinformatics procedures are described in **Supplementary Note**. The somatic substitutions and insertions/deletions found are listed in **Supplementary Tables 5–9**.

25. Li, H. & Durbin, R. Fast and accurate short read alignment with Burrows-Wheeler Transform. *Bioinformatics* **25**, 1754–1760 (2009).
26. Li, H. *et al.* The Sequence alignment/map (SAM) format and SAMtools. *Bioinformatics* **25**, 2078–2079 (2009).