Rapid growth of a hepatocellular carcinoma and the driving mutations revealed by cell-population genetic analysis of whole-genome data

Yong Tao[a,1], Jue Ruan[a,1], Shiou-Hwei Yeh[b,1], Xuemei Lu[a,c,1], Yu Wang[a,1], Weiwei Zhai[a,1], Jun Cai[a,1], Shaoping Ling[c], Qiang Gong[a], Zecheng Chong[a], Zhengzhong Qu[a], Qianqian Li[a], Jiang Liu[a], Jin Yang[c], Caihong Zheng[a], Changqing Zeng[a], Hurng-Yi Wang[b], Jing Zhang[a], Sheng-Han Wang[b], Lingtong Hao[c], Lili Dong[c], Wenjie Li[c], Min Sun[c], Wei Zou[c], Caixia Yu[c], Chaohua Li[c], Guojing Liu[a], Lan Jiang[a], Jin Xu[a], Huanwei Huang[a], Chunyan Li[a], Shuangli Mi[a], Bing Zhang[c], Baoxian Chen[c], Wenming Zhao[c], Songnian Hu[c], Shi-Mei Zhuang[d], Yang Shen[d], Suhua Shi[d], Christopher Brown[e], Kevin P. White[e], Ding-Shinn Chen[b,2], Pei-Jer Chen[b], and Chung-I Wu[a,f,2]

[a]Laboratory of Disease Genomics and Individualized Medicine, and [c]China Academy of Sciences Key Laboratory of Genome Sciences and Information, Beijing Institute of Genomics, Chinese Academy of Sciences, Beijing 100029, People's Republic of China; [b]Graduate Institute of Clinical Medicine and Hepatitis Research Center, National Taiwan University and Hospital, Taipei 106, Taiwan; [d]State Key Laboratory of Biocontrol, School of Life Science, SunYat-Sen University, Guangzhou 510275, People's Republic of China; and [e]Institute for Genomics and Systems Biology, and [f]Department of Ecology and Evolution, University of Chicago, Chicago, IL 60637

Contributed by Ding-Shinn Chen, June 3, 2011 (sent for review March 7, 2011)

We present the analysis of the evolution of tumors in a case of hepatocellular carcinoma. This case is particularly informative about cancer growth dynamics and the underlying driving mutations. We sampled nine different sections from three tumors and seven more sections from the adjacent nontumor tissues. Selected sections were subjected to exon as well as whole-genome sequencing. Putative somatic mutations were then individually validated across all 9 tumor and 7 nontumor sections. Among the mutations validated, 24 were amino acid changes; in addition, 22 large indels/copy number variants (>1 Mb) were detected. These somatic mutations define four evolutionary lineages among tumor cells. Separate evolution and expansion of these lineages were recent and rapid, each apparently having only one lineage-specific protein-coding mutation. Hence, by using a cell-population genetic definition, this approach identified three coding changes (CCNG1, P62, and an indel/fusion gene) as tumor driver mutations. These three mutations, affecting cell cycle control and apoptosis, are functionally distinct from mutations that accumulated earlier, many of which are involved in inflammation/immunity or cell anchoring. These distinct functions of mutations at different stages may reflect the genetic interactions underlying tumor growth.

cell genealogy | cellular evolution | foreground mutation

Tumorigenesis is generally believed to be the consequence of mutation accumulation, including single nucleotide substitutions, structural variations, and epigenetic changes, in somatic cells (1). A typical cancer may have thousands of somatic mutations, of which 10–100 may be in coding regions (2–7). A central issue in cancer genomics is then the dynamics of tumor growth in relation to the accumulation of these mutations. Given any individual case of cancer, the questions are hence: (*i*) how many adaptive mutations drive the tumor growth; (*ii*) how strongly each mutation drives the growth; and (*iii*) what their molecular nature is vis-à-vis that of the background mutations. To answer these questions, we treat each tumor as a population of cells and apply population genetic principles to infer adaptive mutations (8).

Cancer mutations are often divided into drivers and passengers (9). Driver mutations are those that contribute directly to tumorigenesis and their identification is crucial for understanding the molecular biology of cancers. An important issue is how driver mutations should be defined operationally. Candidate driver mutation in the literature often refers to coding changes in genes that are commonly mutated, for example, in multiple cases of hepatocellular carcinoma (HCC). Adaptive mutation proposed here is an alternative definition of candidate driver mutation, inferred from the dynamics of cell proliferation in its natural setting within a single patient.

In this report, we analyze a case of HCC, the fifth most common cancer worldwide, by such an approach. We regard HCC as particularly favorable for identifying candidate driver mutations for several reasons. First, liver resections from the surgery usually contain high yields of DNA from hepatocytes. Second, liver tissues regenerate, resulting in active cell turnover and an opportunity for a more clonal genealogical pattern. Third, previous studies including our own (10) suggest that different cases of HCC may exhibit a wide range of evolutionary dynamics, as their pathologies and anatomies vary extensively. Some of these cases should have the growth rate and pattern conducive for isolating the small number of adaptive mutations.

Results

Sequencing and Mutation Detection. The subject of this study was a female patient with chronic hepatitis B virus (HBV) infection, diagnosed with HCC at the age of 35. A pedunculate tumor (labeled "primary" in Fig. 1) was removed in the first surgery. This primary tumor was grade II to III HCC with prominent clear cell components. Fifteen months later, HCC recurrences were detected and the patient received a second surgery. Recurrent tumor 1 occurred in the regenerated liver at the site of the initial resection and a smaller recurrent tumor 2 was also identified at a second, nearby site. The case report and informed consent are presented in *SI Materials and Methods A1*.

The locations and sizes of patient samples are summarized in Fig. 1. In total, nine sample sections from the three tumors (T0–T6 from the primary tumor and R1/R2 from the two recurrent tumors) and seven sample sections from the adjacent nontumor tissues (N0, N1–N6; Fig. 1) were obtained. Examination of the pathological anatomy indicated that the proportion of hepatoma

Author contributions: S.-H.Y., X.L., D.-S.C., P.-J.C., and C.-I.W. designed research; Y.T., S.-H.Y., Q.G., Z.Q., J.Y., C. Zheng, H.-Y.W., S.-H.W., W.L., M.S., Chaohua Li, G.L., H.H., Chunyan Li, S.M., B.Z., B.C., S.-M.Z., and S.S. performed research; J.R., Y.W., W. Zhai, J.C., S.L., Q.G., Z.C., Q.L., J.Z., L.H., L.D., W. Zou, C.Y., L.J., J.X., W. Zhao, S.H., and Y.S. analyzed data; S.-H.Y. and P.-J.C. provided clinical samples; and Y.T., J.R., S.-H.Y., X.L., Y.W., W. Zhai, J.C., S.L., J.L., C. Zeng, C.B., K.P.W., D.-S.C., P.-J.C., and C.-I.W. wrote the paper.

The authors declare no conflict of interest.

[1]Y.T., J.R., S.-H.Y., X.L., Y.W., W. Z., and J.C. contributed equally to this work.

[2]To whom correspondence may be addressed. E-mail: chends@ntu.edu.tw or wuci@big.ac.cn.

This article contains supporting information online at www.pnas.org/lookup/suppl/doi:10.1073/pnas.1108715108/-/DCSupplemental.

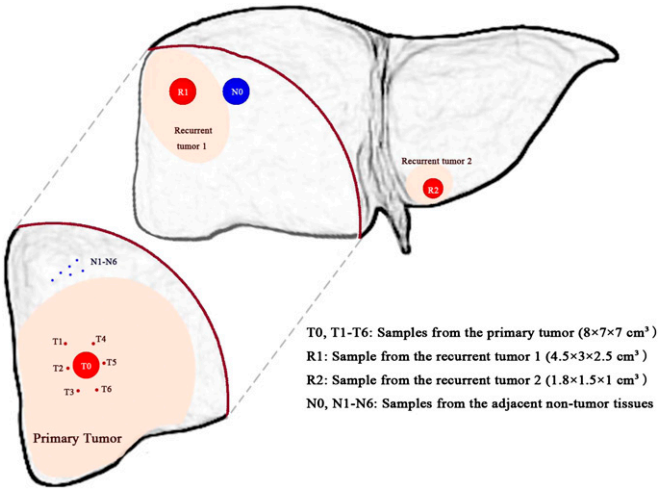


Fig. 1. The scheme of sampling from the HCC liver. The resected portion containing the primary tumor is drawn outside of the liver, as indicted by the dotted lines. From the primary tumor, one large section (T0, >50 mm^3) and six small sections (T0–T6, <5 mm^3, shown as dots) were taken. Six small sections (N1–N6) were also taken from the adjacent nontumor tissues. The recurrent tumors were detected and operated on 15 mo after the first surgery. The larger recurrent tumor resides in the regenerated portion of the liver and a section (R1, >50 mm^3) was taken from it, as well as a section (N0) from the adjacent nontumor tissue. Another section (R2) was taken from the smaller recurrent tumor. For more information, please see *Materials and Methods*.

cells in the tumor sections was 70–90%. This estimation is corroborated by the sequencing results presented below.

The samples of T0, R1, R2, and N0 were subjected to exon capture and Next Generation Sequencing (NGS) to a depth of 50–60×. In addition, R1 and N0 were subjected to whole-genome sequencing, which yielded a 20× coverage of uniquely mapped reads (*SI Materials and Methods A4* and *Results B1*). For R1 and N0, coverage of exon sequences thus reached an average depth of 70× . We chose the R1 section, instead of T0, for whole-genome sequencing, because the primary tumor and recurrent tumors occurred in the same region of the liver. Given the progression of events, we expected R1 to carry all T0 mutations as well as a few additional ones. This is indeed the case from the exon-capture data (*SI Materials and Methods A3*). Because all T0 mutations are represented by the R1 data, R1 cells are most likely the direct descendants of the primary tumor. Finally, the choice of nontumor liver sections (N0–N6), vis-a-vis a nonliver tissue, as a reference for inferring tumor-specific mutations has somewhat different consequences (*SI Materials and Methods A3*).

These sequencing data were used to select sites of somatic mutations for validation. The detailed procedures including site selection, validation accuracy, and false-positive and false-negative estimation are presented in *SI Materials and Methods A5.1* and Fig. S1. In brief, sites were chosen when the frequency of a candidate mutation was higher than a cutoff (often, but not always, at 30%) in the R1 or R2 section and zero in the normal section (referred to as T > N sites). The cutoff was chosen to include even marginal candidate sites so true sites would not be missed. False positives could then be screened out by validations. We allowed higher false positives than usual, obtaining an average validation rate of 50%. Validation was performed for all nine tumor and seven nontumor sections (Dataset S1). All T > N sites were subjected to Sequenom validation (MassARRAY MALDI-TOF MS system) and about one-half were subjected to further validation by PCR-NGS sequencing to an average depth of >8,000× . The validated mutation frequencies by Sequenom and PCR-NGS sequencing are in good agreement with the correlation coefficient ranging between 0.86 and 0.89 (*SI Results B2* and Dataset S1). In nontumor sections, mutant frequencies at T > N sites were too low to measure accurately by Sequenom; hence, only PCR-NGS data were used (*SI Results B2*).

Divergence Between Tumor (R1) and Nontumor (N0) Sections. We first present the accumulation of mutations in one tumor section (R1) relative to a nontumor section (N0). Data from other sections will follow later. Three classes of somatic mutations were considered.

First, we identified point mutations and small indels. In Table S1, 214 point (single nucleotide) mutations were validated, of which 193 were noncoding or synonymous (silent) and 21 were nonsynonymous. Fig. S2 shows examples of silent mutation frequencies. The list of criteria used in the filtering process is summarized in Table S2. On a whole-genome basis, the estimated rate of somatic point mutations is ~0.8 per Mb, or 2,500 mutations genome wide (*SI Materials and Methods A5.3*). This mutation density is close to the median value reported in the literature (2). The 21 nonsynonymous mutations we detected are close to the expected number based on the genome-wide density of 2,500 mutations. Table 1 lists these mutations individually, with their frequencies in N0, T0, R1, and R2 shown. The frequencies are usually <50%, as the sites are mostly heterozygous and the proportions of cancerous cells in the samples are 70–90% (see below). Table 1 also includes two small indels that cause a frame-shift in the coding region of the PIGF and TP53 genes (*SI Materials and Methods A6.1*).

Second, copy number variants (CNVs) and large chromosomal indels were identified. CNVs are a source of genetic diversity in many cancers (11, 12). A major class of CNVs is large chromosomal indels, which include duplications and deletions of chromosomal segments. It should be noted that polyploids are common in tumor cells and, in even normal hepatocytes, tetraploids are often observed (13–15). Hence, regionally averaged minor allele frequencies (MAFs) at heterozygous sites provide reliable estimates of local copy number. As expected, the MAF statistic is remarkably stable across chromosomes in nontumor sections (Fig. 2). In tumor sections, MAFs and read depth are both informative about chromosomal indels and are generally concordant (Fig. 2 and Fig. S3). A region on chromosome 6 (Fig. 2, marked by a red box) is an exception. It has the baseline copy number, but the MAFs deviate strongly from the average. A possible explanation is that this region has a 3:1 allele ratio instead of the baseline 2:2 for a tetraploid.

Given that the beginning and end of each chromosomal indel are characterized by abrupt transitions in both read depth and MAF (*SI Materials and Methods A7*), we used these data to identify all copy number breakpoints in the genome (Table S3). In total, 26 such breakpoints were identified from either MAF or read depth data. We conservatively considered breakpoints not jointly called by both data types to be false positives, as misinferences are common for smaller chromosomal indels. Using concordant breakpoints as a guide, we identified 19 chromosomal indels and three CNV regions (two on chromosome 5 and one on chromosome 11; Fig. S3) in the whole genome.

With the resolution of our analysis, all chromosomal indels of >1 Mb at >20% in frequency in the cell population should have been detected (*SI Materials and Methods A7*). Among those detected, Δ5q (Fig. 2 and Fig. S4) is of particular interest. The breakpoints of Δ5q fall in the introns of two genes, resulting in their truncation and fusion. The fused transcript can be detected by RT-PCR and will be referred to as the M10 mutation (Fig. S5). Because the impact of Δ5q on tumor growth could result from either the lower dosage of genes in the deleted region or the transcript at the breakpoints, we shall refer to this deletion as Δ5q (M10) whenever both properties are relevant.

Finally, 18 HBV integrations were detected. No insertion site was found in the coding regions. We chose four integration sites for PCR validation, two in the introns of coding genes (TPPP and SHANK2) and two in intergenic regions. The validation shows all four of them to be present in all tumor sections (T0–T6, R1, and

Table 1. Coding genes affected by somatic mutations in tumors

Gene name	Amino acid changes[1]	Mutation frequencies in (N0, R1, R2, T0)[2]	Mutation effect[3]	Description
Mutations polymorphic among tumors (M3 and M4 are foreground mutations; M1 and M2 are background mutations subsequently deleted by Δ5q)				
TMEM173 (M1)	276 Q->*	0.00, 0.00, **0.36**, 0.02		STING (stimulator of IFN genes); inflammation/defense/immunity
ANKHD1 (M2)	689 P->R	0.00, 0.00, **0.34**, 0.01	D	A scaffolding protein affecting leukemia-cell phenotype
CCNG1 (M3)	15 H->N	0.00, 0.00, **0.45**, 0.00	d	Cell cycle, G2/M arrest, a target of P53
P62 (M4)	258 D->G	0.00, **0.30**, 0.00, 0.00	D	Sequestosome-1; autophagy, apoptosis
Mutations in high frequencies in all tumor sections (Background mutations)				
TP53	151frame-shift	0.00, 0.56, 0.80, 0.72	D	tumor suppressor gene
DUOX2	519 F->S	0.00, 0.08, 0.28, 0.19	D	inflammation/defense/immunity
CYSLTR1	140 G->C	0.01, 0.31, 0.43, 0.34	d	inflammation/defense/immunity
CYBB	156 A->V	0.00, 0.23, 0.38, 0.34	0	inflammation/defense/immunity
PON3	248 E->*	0.00, 0.29, 0.43, 0.34		Hydrolyze lactone, inhibit oxidation, inflammation/defense
COL1A2	253 G->D	0.00, 0.16, 0.22, 0.19		Extracellular matrix, cell anchorage
COL4A6	497 P->A	0.00, 0.54, 0.60, 0.57		Extracellular matrix, cell anchorage
PIGF	19frame-shift	0.00, 0.16, 0.20, 0.19		Cell anchorage (frequency estimates less reliable)
NUP205	1771 S->I	0.00, 0.27, 0.34, 0.29	d	—
ATP13A3	539 V->G	0.00, 0.32, 0.43, 0.37	D	ATPase
ZNF541	1117 Q->H	0.00, 0.28, 0.41, 0.36		—
RPL12	68 Q->E	0.00, 0.31, 0.44, 0.38	0	—
HMGCS2	416 L->F	0.00, 0.92, 0.94, 0.89	0	Metabolic enzyme in mitochondria
GALNTL4	591 C->*	0.01, 0.76, 0.85, 0.85		—
KIAA1644	43 H->R	0.00, 0.13, 0.61, 0.29	d	—
C14orf28	78 R->W	0.00, 0.29, 0.44, 0.42	D	—
CXorf64	167 G->V	0.00, 0.31, 0.45, 0.35	d	—
RELN	824 I->F	0.04,0.38, 0.50, 0.44	0	Extracellular signal molecule
C17orf75	132 E->D	0.00, 0.35, 0.47, 0.41	0	—
Truncated/fused genes				
C5orf51 - CPEB4 (M10) Chr5:41951996 - Chr5: 173314849				The two genes were truncated and fused by Δ5q. The fused transcript can be detected, designated M10, and it is a foreground mutation.

1 *represents stop codon. *2.* Frequency of lineage specific mutation is boldfaced from the PCR-GAIIx data. *3.* The effect of mutation determined by the program PolyPhen-2. D, high confidence damaging effect; d, possible damaging effect; 0, low confidence damaging effect.

R2) and absent in nontumor sections (N0–N6). These insertions support a simple clonal-expansion model for these tumors. *SI Materials and Methods A2* provides further information.

Genetic Diversity Within and Between Tumors. Among the 214 point mutations shown in Table S1, 205 are observed at similar frequencies in all three tumors (Fig. S2*A* for examples). Only nine mutations, or 4.2%, were observed at very different frequencies among tumor sections (see Table 1 for the nonsynonymous ones). These mutations, polymorphic in the tumor tissue, are labeled M1–M9 in Fig. 3 and will be the basis on which the evolution of these tumors is analyzed in the next section (Fig. 3). Among the silent mutations, M5–M7 deserve a special note. As shown in Fig. S2*A*, these two mutations are absent in R2 and, interestingly, are unusually low in frequency in T3 and T6 (Fig. S2*C*).

Mutation frequencies are also used to gauge sample purity. We note in Fig. S2 that the frequency profiles are fairly consistent in the same sections, roughly in the order of R2 > T0 > R1. These differences likely reflect the different proportions of tumor cells that carry the mutation. We shall refer to this proportion as the composition index [= (the proportion of cancerous cells in the sample) × (the proportion of cancerous cells carrying the mutation)]. The composition index for R2, T0, and R1 is 0.88, 0.75, and 0.65, respectively (*SI Materials and Methods A8*), in accord with the pathology report of 70–90% hepatoma cells in the samples.

The 22 chromosomal indels/CNVs reported in Fig. S3 were initially observed in R1. We then surveyed the other eight tumor sections for their presence by genotyping germ-line heterozygous sites (the position of which is marked on the bottom of Fig. S3). MAFs across these sites indicated that Δ5q is the only chromosomal indel that is not present in every tumor section (*SI Results B3*). Indeed, Δ5q is completely missing in R2 and is in lower frequencies in T3 and T6 than in other tumor sections. Recall that the analysis of Fig. S2 has already found T3 and T6 to be somewhat differentiated from other T sections.

The polymorphism of Δ5q among the tumors raises an interesting question, as the three nonsynonymous mutations, M1–M3, all fall in the region spanned by Δ5q. These three mutations are common in R2 but absent in other tumor sections (Table 1). Hence, they could have occurred in R2, or, alternatively, in the common ancestors but were deleted by Δ5q in all other sections. From the analysis of the *SI Results B3*, M1 and M2 indeed occurred in the common ancestors but were deleted along with Δ5q, as shown in Fig. 3. M3, in contrast, occurred only in R2.

Evolution of the Tumors. The nine point mutations (M1–M9) together with Δ5q (M10) define four different cell lineages among the nine tumor sections (Fig. 3). Each tumor section contains cells from one single lineage, the exceptions being T3 and T6, which consist of mixed lineages. The table in the inset of Fig. 3 summarizes the pattern, as explained below. Two lineages of cells have the M1 and M2 mutations (or, more accurately, did not lose them as a result of Δ5q). The distinction between the two lineages is that the π_1 lineage has M3 (a nonsynonymous mutation in a cyclin G gene) and the π_0 lineage has a silent M8 mutation. The other two lineages, π_2 and π_3, both have M5–M8 and Δ5q (M10) mutations. The π_3 lineage, in addition, has M4 (a nonsynonymous mutation in the P62 gene) and M9. In this figure, cell lineages are drawn as triangles to denote their expansions from a single cell that acquired new mutations. In addition, the lineage from which tumor cells emerged is designated as π_0 in Fig. 3. Details of the phylogenetic reconstruction are given in *SI Results B5*.

There is hardly any doubt that these tumors and cell lineages are highly clonal. After all, more than 95% of somatic mutations, either coding or noncoding changes, are present in all tumor samples. As judged by the size of the π_0 lineage, the cell mass of these tumors generally remained small even when 95% of the

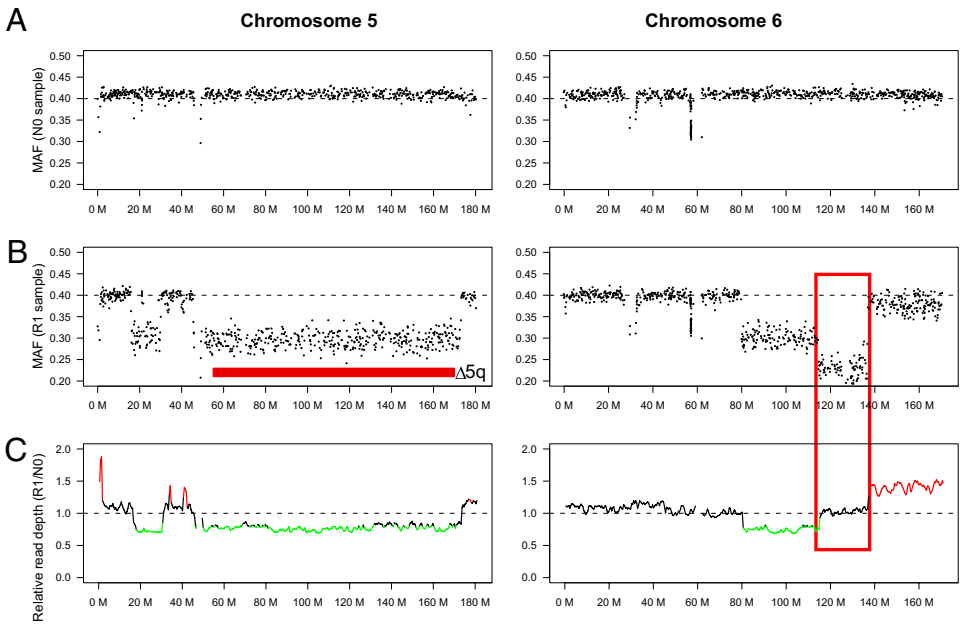


Fig. 2. Detection of large indels on chromosome 5 and 6 from sequence reads. (*A*) For each chromosome, shown are the minor allele frequency (MAF) at heterozygous site in the non-tumor tissue, N0. Each point represents the sum of 50 consecutive polymorphic sites. Non-tumor tissues do not appear to harbor large indels as the frequencies stay relatively constant across regions. (*B*) The corresponding frequencies in the R1 section. The contrast is clear since defined regions in R1 show characteristic reductions in MAFs. (*C*) Read depth is shown; red and green lines denote regions of unusually high or low read depth. There is substantial concordance between *B* and *C* in delineating regions of aberration. Since they are built on very different data, the concordance lends confidence to the interpretation of chromosomal indels. Two features are noteworthy as indicated by a red bar (a deletion, D5q) and a red box, respectively. The region marked by the red box has the average read depth but MAFs are aberrant there. A possible interpretation is that, in tetraploids, the two homologs exist in a ratio of 3:1, instead of 2:2.

mutations had accrued. The growth of the primary and R2 tumors associated with the last few mutations, as shown in Fig. 3, was hence very substantial. R1 deserves special mention, as it occurred in the regenerated liver. The progenitors of R1 are themselves aggressively growing cells of the primary tumor, as discussed before. However, 15 mo after the surgical removal of the primary tumor, the cells that predominated in the recurrent R1 all carried the M4 mutation, which was not even detectable in T0.

The various growth rates of these tumors raise a question of the designation of R2 as "recurrent." Because R2 and the big "primary" tumor (represented by T0–T6) bifurcated from a common lineage when the number of cancerous cells was still small, the late emergence of R2 was due to its slower growth. In fact, either one could have been the true primary tumor. We consider the latter a more likely candidate for the primary site not only because it was observed earlier but also because the least evolved $\boldsymbol{\pi_0}$ lineage can be found only in T3 and T6. The designation, however, affects neither the analysis nor the conclusion of this report.

Most tumor-associated mutations are found in all parts of the tumors. They accumulated in the normal cell lineage, shown as π_n in Fig. 3, and are referred to as background mutations. The remaining few mutations that are polymorphic between or within tumors are referred to as foreground mutations. Foreground mutations that are common in some part of the tumors but absent in other parts are most interesting. As stated above, if a foreground mutation in a gene-coding region is uniquely associated with a large section of tumor and its absence is associated with slower cell proliferation, then this mutation is considered adaptive in terms of the population genetics of cells.

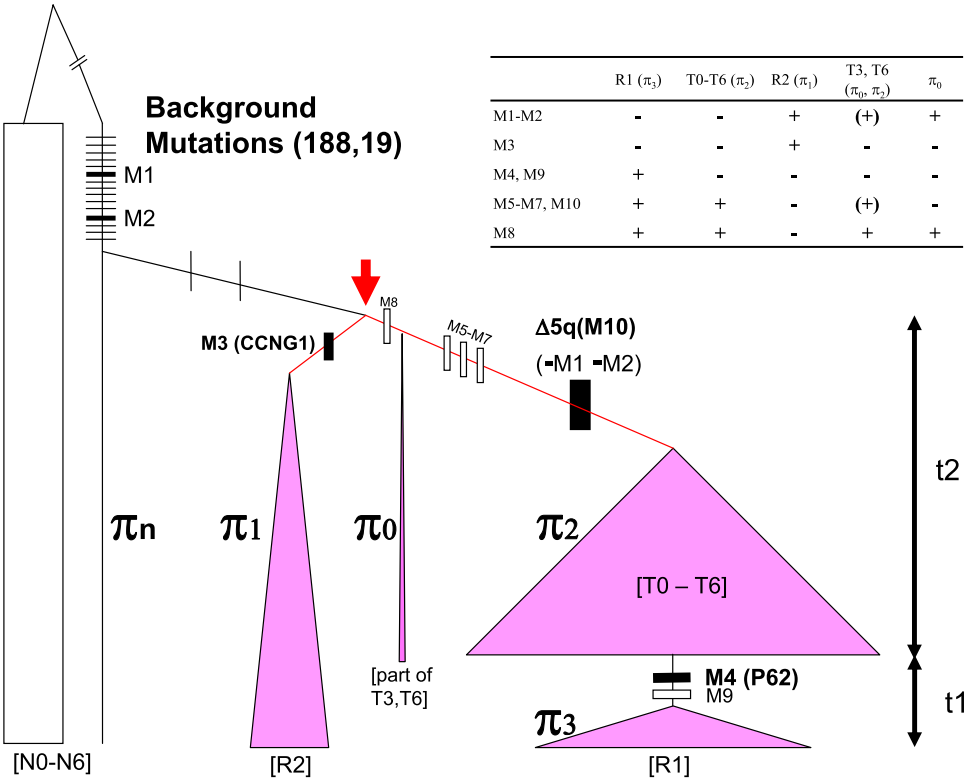
Background mutations. Among the 24 coding region mutations of Table 1, only 3 are foreground mutations, the genealogical patterns of which have been presented in the preceding paragraphs. We shall now describe the possible function of the remaining 21 background mutations and return to the functions of the 3 foreground mutations later.

Because the $\boldsymbol{\pi_0}$ lineage carries all of the background mutations without significant expansion, the background mutations by themselves appeared insufficient for cell proliferation. Nevertheless, some of these mutations may have "primed" cells for transformation, discussed below. One of the background mutations is in P53. Because nearly 30% of HCC have mutations in this gene, the observation is not unexpected. Four of the 21 background mutations in Table 1 affect genes related to inflammation, defense, and/or immunity. They are CYSLTR1 (cysteinyl leukotriene receptor 1), TMEM173, DUOX2 (dual oxidase 2), and CYBB (also called NOX2 for NADPH oxidase2). Recent studies have increasingly suggested a connection between inflammation, immunity, and cancer development (16–18). Most HCC cases in Asian populations, including the one reported here, are HBV-positive and arise following chronic inflammation of the liver (19).

Three genes affected by background mutations are related to cell anchoring and migration. In this study, the migration of cancerous or precancerous cells took place before the expansion of the $\boldsymbol{\pi_1}$ and $\boldsymbol{\pi_2}$ cell lineages. Proper anchoring can transduce signals through the integrin pathway to promote cell division. A step in cancer cell transformation is often the abolishment of this anchorage-dependent cell division (1). Collagens are an important component of this process and mutations in two collagen genes, COL1A2 (G253D) and COL4A6 (P497A), were found (Table 1). Both collagens have been reported to function in cell adhesion, migration, differentiation, and growth (20), and their disruption is associated with carcinomas (21). A third gene, PIGF (phosphatidylinositol glycan F), plays a role in cell-cell anchorage (22, 23).

The deletions of two background mutations, M1 and M2, in the $\boldsymbol{\pi_2}$ lineages by Δ5q merit some attention. Although these two mutations may be merely neutral mutations, it is also possible that they have played a role in the earlier phase of tumorigenesis but have become dispensable later. Both genes appear to have cancer-related functions (Table 1).

Foreground adaptive mutations. Among the coding region mutations of Table 1, only three are in the foreground and considered



	R1 (π_3)	T0-T6 (π_2)	R2 (π_1)	T3, T6 (π_0, π_2)	π_0
M1-M2	-	-	+	(+)	+
M3	-	-	+	-	-
M4, M9	+	-	-	-	-
M5-M7, M10	+	+	-	(+)	-
M8	+	+	-	+	+

Fig. 3. Evolution of the tumors inferred from the data of T0–T6, R1, and R2. The table in the inset shows the presence/absence, indicated by +/– , of each foreground mutation in the tumor sections. (+) denotes presence but at a lower frequency. The table defines the cell lineages. Below the red arrow are mutations accumulated during tumor growth. Red shade denotes tumor cell lineages (labeled as π_0–π_3). The closely related noncancerous cell lineage is labeled π_n. Sample sections, shown in brackets, are written beneath or inside the corresponding cell lineages. M1–M4 and M10 mutations affected amino acid sequences, as shown in Table 1. M5–M9 (□) are silent mutations in intergenic or intronic regions. The deletion Δ5q truncated and fused two genes at the breakpoints. This event is labeled M10. Δ5q also deleted two earlier mutations, M1 and M2. Time is marked by the length of the double arrows on the far right. t1 (=15 mo) is the time between the two surgeries. Among the life time collection of mutations, <5% occurred in the duration of t2. Above the red arrow are background mutations, 188 and 19 of which are silent and nonsynonymous, respectively.

adaptive. These three (M3, M4, and M10), together with a few silent mutations, delineate the cell lineages of Fig. 3. The π_0 lineage is represented by the least-evolved cancerous cells in our samples. π_0 also appears to have the fewest cells among all of the lineages, suggesting that the π_0 cells are less malignant than those in π_1 through π_3. The π_1 lineage is defined by M3 in CCNG-1 (Cyclin G1, H15N), which has a growth inhibitory activity linked to auxin response factor-tumor protein 53 (ARF-p53) and retinoblastoma protein (pRb) tumor suppressor pathways (24). Cyclin G1 is also a target of microRNA (miR)-122a, a microRNA frequently down-regulated in HCC (25). A mutation in CCNG1 has indeed been reported in renal cell carcinoma (7). The CCNG1 mutation marked the transition from the least proliferative cells of the π_0 lineage to the moderately aggressive cells of the π_1 lineage.

The π_2 lineage leads to the primary tumor and later to R1. Δ5q (M10) is the only known coding region mutation that marks the transition from π_0 to the aggressive π_2 and π_3 lineages. The breakpoints of Δ5q create a fused transcript, M10, which consists of the first five exons of C5orf51 and the 3′ end of CPEB4. The latter includes the last exon of CPEB4, inferred to have a frameshift, and the 3′ UTR. C5orf51 is known to be strongly expressed in the liver and highly conserved among mammals; CPEB4 has been implicated in mitotic control (26). Furthermore, the region spanned by Δ5q contains the APC gene and the 5q31 cluster of cytokines, both having been found to be lost in adenomas and carcinomas (27). Loss of heterozygosity in 5q has been reported to be correlated with cancer risk (28) and the histopathological grade of tumors and metastases (29, 30).

The π_3 lineage is defined by M4 (in the P62 gene). This rapidly proliferating lineage is a main constituent of R1. In the 15 mo after surgery that removed the primary tumor, R1 grew in the regenerated portion of the liver and reached a substantial size. In comparison, R2 is much smaller, even though it may have started earlier (as R1 could start growing only after the resection of the primary tumor). p62 is a multidomain signaling adaptor protein that affects autophagy, apoptosis, and cancer (31). Indeed, autophagy suppresses tumorigenesis with the elimination of p62 (32), which is implicated in the regulation of many targets, including MEK5, ERK, RIP, aPKC, and TRAF6 (31). Genetic ablation of p62 suppresses the appearance of ubiquitin-positive protein aggregates in hepatocytes (33). These findings link p62 activities to apoptosis and suggest that the modulation of p62 by autophagy might be relevant to tumorigenesis (32).

Finally, we should note that some normal samples can be informative about tumor evolution as well. For example, in the N3 section, the mutation frequency at many sites appears to be higher than those in other nontumor sections, but the difference is substantially larger in some sites than in others. Interestingly, M5 and M6 is unusually low in N3. If N3 contains advanced cancerous cells, the frequency profile should be even across sites. These observations suggest that N3 may contain precursor cancer cells at an earlier stage of evolution.

Discussion

In addition to the identities of somatic mutations, cancer genomic data can provide detailed information on how tumors grow in relation to the accumulation of mutations. A cell-population genetic analysis of tumors is not unlike the analysis of mutation accumulation in geographical populations of natural species like *E. coli* (34, 35) [and, to some extent, humans (36, 37) and *Drosophila* (38)]. Among the thousands of mutations accrued in each case, it is sometimes possible to identify a small number of adaptive mutations that drive cell proliferation. Furthermore, even noncoding mutations can be informative about how rapidly the tumors have grown. We should note that each individual case of cancer is informative on its own and the assumption of common mutations is not necessary.

In this case of HCC, the tumors remained small (judged by the size of the π0 lineage) late in cancer evolution, when all background mutations have already occurred (Fig. 3). If we use silent mutations to mark the divergence time between cell lineages, the ratio of foreground to background mutations is 5:188. For coding region mutations, three [CCNG1, P62, and Δ5q (M10)] are foreground changes among the 24 reported in Table 1. Thus, the evolutionary dynamics inferred from this study is a long process of accumulation

of background mutations, followed by the rapid spread of a relatively small number of (adaptive) foreground mutations.

Nonsynonymous mutations in the background and foreground fall into different functional categories. In this study, background mutations, including one in P53, did not directly cause cell proliferation, but some of them might have "primed" the cells to proliferate. Indeed, seven background mutations are in genes of inflammation/immunity or cell anchoring. In comparison, foreground mutations affect genes of cell cycle control and apoptosis. One might expect that, after the background mutations have laid the groundwork, foreground mutations should directly affect cell division and cell death. Hence, the functional division between background and foreground mutations appears to agree with this simple expectation.

The distinct functions between foreground and background mutations suggest that tumorigenesis may be driven by epistatic gene interactions. With epistasis, mutations of either kind alone may have a much weaker effect on tumor growth than the joint presence of background and foreground mutations. Such a genetic architecture is not uncommon for traits that have evolved over time (39). With that consideration, the best genetic background to test the functions of the three adaptive mutations would be that of the π_0 lineage, which has all of the background mutations. In a wild-type genetic background, it is possible that the three adaptive mutations may not impart cancer-causing phenotypes.

There are caveats, both specific and general, that need to be heeded. Specifically, we identified one, and only one, protein-coding mutation for each of the three proliferation events in Fig. 3. In *SI Results B5*, we present several lines of evidence that coding mutations should not have been missed. In addition, the depth of coverage, the cutoff used in choosing sites for validation, and the paucity of intermediate frequency mutations are also addressed. A more general caveat is that this case might be unusual and its level of genetic differentiation happens to be particularly suitable for identifying driver mutations. Indeed, the process of mutation accumulation and natural selection is likely to be highly stochastic. In some cases of tumor evolution, there might be little genetic diversity among all tumor cells if a powerful driver mutation has caused a strong "selective sweep" (40). The variation in the evolutionary dynamics may prove to be as informative about tumorigenesis as the common mutations. If that is true, this study would be a small step in elucidating that variation.

Materials and Methods

A 35-y-old woman with chronic HBV infection was diagnosed with HCC. Two tumor and one nontumor sections, R1, R2, and N0, were subjected to exon capture and SOLiD sequencing. R1 and N0, in addition, were subjected to whole-genome sequencing. Sequence reads were aligned to the reference human genome (NCBI36) using SOLiD Corona Lite and Burrows-Wheeler Aligner. Putative somatic mutations identified by sequencing were then validated by Sequenom genotyping and deep sequencing across all nine tumor and seven nontumor tissue sections. CNVs and chromosomal indels were identified using our in-house programs combining information from both read depth (coverage) and MAF at germ-line heterozygous sites. Full materials and methods used to generate this data set and results are provided in *SI Materials and Methods*.

ACKNOWLEDGMENTS. We thank Maynard Olson for advice. We also thank Andy Clark, Michelle LeBeau, Steve O'Brien, Julie Schneider, Ralph Weichselbaum, and Y.M. Jeng for comments and input. Taiwan Liver Cancer Network provided valuable assistance. This study was supported by National S&T Major Project of China Grants 2009ZX08010-017B and2009ZX08009-149B, National Natural Science Foundation of China Grants 30950006, 31000957, and 31071914, Chinese Academy of Sciences Grant KSCX1-YW-22, National Basic Research Program of China Grants 2011CB510101 and 2011CB510106, and an National Research Program for Genomic Medicine grant (to P.-J.C.).

1. Weinberg R (2007) *The Biology of Cancer* (Garland Science, New York).
2. Greenman C, et al. (2007) Patterns of somatic mutation in human cancer genomes. *Nature* 446:153–158.
3. Ley TJ, et al. (2008) DNA sequencing of a cytogenetically normal acute myeloid leukaemia genome. *Nature* 456:66–72.
4. Shah SP, et al. (2009) Mutational evolution in a lobular breast tumour profiled at single nucleotide resolution. *Nature* 461:809–813.
5. Pleasance ED, et al. (2010) A small-cell lung cancer genome with complex signatures of tobacco exposure. *Nature* 463:184–190.
6. Ding L, et al. (2010) Genome remodelling in a basal-like breast cancer metastasis and xenograft. *Nature* 464:999–1005.
7. Dalgliesh GL, et al. (2010) Systematic sequencing of renal carcinoma reveals inactivation of histone modifying genes. *Nature* 463:360–363.
8. Greaves M (2007) Darwinian medicine: a case for cancer. *Nat Rev Cancer* 7:213–221.
9. Segal NH, et al. (2008) Epitope landscape in breast and colorectal cancer. *Cancer Res* 68:889–892.
10. Chen YJ, et al. (2000) Chromosomal changes and clonality relationship between primary and recurrent hepatocellular carcinoma. *Gastroenterology* 119:431–440.
11. Navin N, et al. (2010) Inferring tumor progression from genomic heterogeneity. *Genome Res* 20:68–80.
12. Pollack JR, et al. (2002) Microarray analysis reveals a major direct role of DNA copy number alteration in the transcriptional program of human breast tumors. *Proc Natl Acad Sci USA* 99:12963–12968.
13. Kudryavtsev BN, Kudryavtseva MV, Sakuta GA, Stein GI (1993) Human hepatocyte polyploidization kinetics in the course of life cycle. *Virchows Arch B Cell Pathol Incl Mol Pathol* 64:387–393.
14. Guidotti JE, et al. (2003) Liver cell polyploidization: a pivotal role for binuclear hepatocytes. *J Biol Chem* 278:19095–19101.
15. Duncan AW, et al. (2010) The ploidy conveyor of mature hepatocytes as a source of genetic variation. *Nature* 467:707–710.
16. Grivennikov SI, Greten FR, Karin M (2010) Immunity, inflammation, and cancer. *Cell* 140:883–899.
17. Naugler WE, et al. (2007) Gender disparity in liver cancer due to sex differences in MyD88-dependent IL-6 production. *Science* 317:121–124.
18. Rebouissou S, et al. (2009) Frequent in-frame somatic deletions activate gp130 in inflammatory hepatocellular tumours. *Nature* 457:200–204.
19. Chen PJ, Chen DS (1999) Hepatitis B virus infection and hepatocellular carcinoma: molecular genetics and clinical perspectives. *Semin Liver Dis* 19:253–262.
20. Tanjore H, Kalluri R (2006) The role of type IV collagen and basement membranes in cancer progression and metastasis. *Am J Pathol* 168:715–717.
21. Ikeda K, et al. (2006) Loss of expression of type IV collagen alpha5 and alpha6 chains in colorectal cancer associated with the hypermethylation of their promoter region. *Am J Pathol* 168:856–865.
22. Ohishi K, et al. (1995) Structure and chromosomal localization of the GPI-anchor synthesis gene PIGF and its pseudogene psi PIGF. *Genomics* 29:804–807.
23. Takeda J, et al. (1993) Deficiency of the GPI anchor caused by a somatic mutation of the PIG-A gene in paroxysmal nocturnal hemoglobinuria. *Cell* 73:703–711.
24. Zhao L, et al. (2003) Cyclin G1 has growth inhibitory activity linked to the ARF-Mdm2-p53 and pRb tumor suppressor pathways. *Mol Cancer Res* 1:195–206.
25. Gramantieri L, et al. (2007) Cyclin G1 is a target of miR-122a, a microRNA frequently down-regulated in human hepatocellular carcinoma. *Cancer Res* 67:6092–6099.
26. Novoa I, Gallego J, Ferreira PG, Mendez R (2010) Mitotic cell-cycle progression is regulated by CPEB1 and CPEB4-dependent translational control. *Nat Cell Biol* 12: 447–456.
27. Vogelstein B, et al. (1988) Genetic alterations during colorectal-tumor development. *N Engl J Med* 319:525–532.
28. Johnson LG, et al. (2011) Risk of cervical cancer associated with allergies and polymorphisms in genes in the chromosome 5 cytokine cluster. *Cancer Epidemiol Biomarkers Prev* 20:199–207.
29. Morita R, et al. (1991) Common regions of deletion on chromosomes 5q, 6q, and 10q in renal cell carcinoma. *Cancer Res* 51:5817–5820.
30. Fong KM, Zimmerman PV, Smith PJ (1995) Tumor progression and loss of heterozygosity at 5q and 18q in non-small cell lung cancer. *Cancer Res* 55:220–223.
31. Moscat J, Diaz-Meco MT (2009) p62 at the crossroads of autophagy, apoptosis, and cancer. *Cell* 137:1001–1004.
32. Mathew R, et al. (2009) Autophagy suppresses tumorigenesis through elimination of p62. *Cell* 137:1062–1075.
33. Komatsu M, et al. (2007) Homeostatic levels of p62 control cytoplasmic inclusion body formation in autophagy-deficient mice. *Cell* 131:1149–1163.
34. Lenski R, Rose M, Simpson S, Tadler S (1991) Long-term experimental evolution in Escherichia coli. I. Adaptation and divergence during 2,000 generations. *Am Nat* 138: 1315–1341.
35. Cooper VS, Lenski RE (2000) The population genetics of ecological specialization in evolving Escherichia coli populations. *Nature* 407:736–739.
36. Tishkoff SA, et al. (2007) Convergent adaptation of human lactase persistence in Africa and Europe. *Nat Genet* 39:31–40.
37. Hamblin MT, Di Rienzo A (2000) Detection of the signature of natural selection in humans: evidence from the Duffy blood group locus. *Am J Hum Genet* 66:1669–1679.
38. Greenberg AJ, Moran JR, Coyne JA, Wu CI (2003) Ecological adaptation during incipient speciation revealed by precise gene replacement. *Science* 302:1754–1757.
39. Sun S, Ting CT, Wu CI (2004) The normal function of a speciation gene, Odysseus, and its hybrid sterility effect. *Science* 305:81–83.
40. Smith JM, Haigh J (1974) The hitch-hiking effect of a favourable gene. *Genet Res* 23: 23–35.

Supporting Information

Tao et al. 10.1073/pnas.1108715108

SI Materials and Methods

A1. Liver Samples and Patient Information. A 35-y-old woman with chronic HBV infection was diagnosed with hepatocellular carcinoma (HCC) in 1988 at National Taiwan University Hospital. The primary tumor that was on the right lobe of liver was well encapsulated and 8 × 7 × 7 cm in size. It was a grade II to III HCC with prominent clear cell components. The liver showed invasive and septal cirrhosis with macro- and micronodules. Fifteen months after the first hepatotomy when the primary tumor was resected, the patient developed recurrent intrahepatic tumors, among which one (R1) was 4.5 × 3 × 2.5 cm in size on the right lobe and the other (R2) was 1.8 × 1.5 × 1cm on the left lobe (Fig. 1). Informed consent for the whole genome sequencing was obtained from the Institutional Review Board of National Taiwan University Hospital.

For DNA preparation, nine tumor specimens were collected from the primary tumor (T0, T1–T6), as well as two recurrent tumors, R1 and R2. We also obtained seven nontumor sections as shown in Fig. 1. T1–T6 and N1–N6 are small tissue sections of 1–2 mm diameter in size. T0, R1, and R2 sections are larger and are ∼100 mg in weight. Less than 10 μg genomic DNA was obtained from T1–T6 and N1–N6 and 60–100 μg DNA were purified from T0, R1, and R2.

The pathology report prepared by the Hepatitis Research Center, National Taiwan University Hospital, indicated that the tumor sections contain 70–90% hepatoma cells. The percentages in the smaller T sections are usually ∼90%. The report is corroborated by the sequencing results on tumor-specific mutations.

A2. Additional Information on HCC. Liver cancer ranks fifth in prevalence among all cancers worldwide. HCC is more common in Asia than in other continents, partly due to the high rate of chronic HBV infection. About 30–50% of HCC patients with early diagnosis receive surgical treatments. However, HCC recurs at a rate ∼20–40% in the first 2 y. Some of these HCC recurrences receive second surgery (1, 2).

Although HCC incidence and HBV infection are strongly correlated, the correlation does not necessarily imply a causal relationship. HBV integrations have been shown to occur soon after infection, often in newborns or in young children (3, 4). Because it usually takes decades of chronic inflammations before HCC develops, the correlation might be mediated through inflammation, rather than direct HBV integrations into the genome (1). In this study, we consider the relationship between HBV integration and HCC evolution an unresolved question.

A3. Choice of Sample Sections. R1 and N0, representing the tumor and nontumor tissues, respectively, were subjected to exome and whole-genome sequencing. We chose R1, instead of T0–T6, for sequencing for the following reason. Because R1 was regenerated in the resected region where T sections used to be, the genealogical pattern (Fig. 3) suggests that R1 should have all of the fixed mutations in the T sections, but not vice versa. The M4 and M9 mutations are such examples. Rare mutations in the T sections might have been missed, but their relevance to tumorigenesis may be less compelling. As stated in the *Discussion*, the existence of high frequency mutations in T0 that are absent in R1 is now being tested, even though we hypothesize otherwise.

Reference (nontumor) tissues serve to define tumor-associated mutations. Choosing a more distantly related tissue than was done here (blood, for example), we might identify more somatic mutations shared by all liver cells including both cancerous and normal cells. Such mutations must occur early in embryogenesis and may not be highly germane to tumorigenesis. They are also likely to be quite rare. In this study, we chose the adjacent nontumor sections as reference and, hence, our study focused on mutations that differentiate cancerous and noncancerous hepatoma cells. By choosing this sample, we also had a chance of finding the nontumor cell lineage, π_n (Fig. 3), from which tumors eventually emerged. Figure S2*B* hints that the N3 sample may contain some cells from this lineage (see legends).

A4. Whole-Genome Sequencing and Exome Capturing and Sequencing. Both SOLiD mate-paired and Illumina GAIIx paired-end DNA libraries for recurrent tumor 1 (R1) and its adjacent nontumor tissue (N0) were constructed using paired-end sample prep kit (Illumina) and SOLiD mate-paired library construction kit (Life Technologies) according to standard manufacturer protocols. The SOLiD libraries were sequenced using SOLiD v3, which generated 2 × 35 bp and 2 × 50 bp mate-paired reads. SOLiD color space reads were mapped to the human reference genome (hg18) using the MAPREADS program in the Corona_Lite package (Life Technologies) allowing up to three mismatches for 35-bp reads and five mismatches for 50-bp reads. Sequencing of Illumina libraries was performed using GAIIx sequencer to generate 2 × 75 paired-end reads. Illumina reads were mapped to the human reference genome using BWA (5) with default parameters. In both platforms, only reads mapped uniquely to the genome were retained.

We carried out whole-exon capture using Agilent SureSelect Human All Exon Kit (Agilent) for two recurrent tumors (R1 and R2) and a nontumor tissue (N0). A total of 3 μg genomic DNA was sheared to generate fragments of the target range between 100 and 150 bp. DNA fragments were end-repaired, ligated with adapters, amplified, and selectively hybridized to capture probes according to the SureSelect Target Enrichment System for the Applied Biosystems SOLiD System Protocol (version 1.5; G3360-90001). The SureSelect fragment libraries were subjected to emulsion PCR and sequencing following standard SOLiD 3 protocols. For each sample, one-half of a SOLiD sequencing slide was used to generate 40–50 million reads.

A5. Detection of Mutations. ***A5.1 Selection of candidate point mutations.*** The criteria for selecting genome-wide putative R1-specific (high in R1, zero in N0) and N0-specific (high in N0, zero in R1) mutations are given in Table S2 and Fig. S1*A*. The collecting of sites was done in two separate batches. In batch 1, only SOLiD sequencing reads were used. In batch 2, SOLiD and GAIIx data were combined and used for selecting candidate sites. A list of criteria used in the filtering process is summarized in Table S2.

Candidate coding nonsynonymous point mutations were selected using both whole-genome sequencing data as well as exon-capture data. Fig. S1*B* presents the flowchart for picking and verifying tumor (R1) specific nonsynonymous mutations. The criteria used in this filtering are also summarized in Table S2. Note that the stringency in selecting nonsynonymous mutations is slightly lower compared with genome wide filtering (Table S2). Each candidate position was also subjected to manual checking. Sites with low read quality or mapping quality were filtered out.

We believe to have captured most nonsynonymous tumor-specific mutations for the following reasons. First, the average coverage of coding regions is 70×. Second, in choosing sites for validation of tumor-specific mutations, we set the cutoff in the mutation frequency from the sequencing data at 30%. Mutations

that are much lower in frequency in tumors may not be germane to tumorigenesis. Therefore, we only wish to address the intermediate frequency mutations that might have been missed. We reason that there may not be many such mutations. The cell population of each tumor section is likely to have experienced a recent selective sweep due to the fixation, or near fixation, of advantageous mutations (for example, CCNG1 in R2). As a result, the frequency spectrum of mutations in a local cell population is expected to be strongly bimodal, with mutations at either very high or very low frequencies (6, 7).

A5.2 Validation of point mutations by Sequenom and PCR-Next Generation Sequencing. Candidate point mutations that appeared as Tumor-specific or N0-specific were chosen for validation. Sequenom MassARRAY was used to validate all potential mutations. Genomic positions for all single nucleotide variants (SNVs) were retrieved using the HG18 build of the human genome and the University of California Santa Cruz (UCSC) genome annotation database. PCR and MassEXTEND primers for multiplexed assays are designed using the Sequenom MassARRAY Assay Design 3.1 software. MassEXTEND reaction and iPLEX Gold assay were carried out for primer extension and SNP allele detection. The allele-specific extension products of different masses were quantitatively analyzed using the MALDI-TOF mass spectrometer. Mutation calls and allele frequencies for each SNV site were determined using MassArray Typer 4.0 Analyzer according to the manufacturer's specifications. For frequency analysis, each sample was genotyped three times. Successful genotyping assays were defined those sites that can be successfully genotyped in all samples. In a calibration run, the observed mean + SD among 179 germ-line heterozygous sites is 0.503 ± 0.050, which compares favorably with the expected mean of 0.5.

Mutations in CCNG1 and P62 were further validated by Sanger sequencing on the ABI 3730xl sequencer. The primer sequences used are: CCNG1 forward primer 5′CAA CTT GTA GAA GGG AAA T-3′ and reverse primer 5′-CAA AGC CAA AGA ACT GA-3; P62 forward primer 5′-TGG GTT TGT ATC GTC TGG T-3′ and reverse primer 5′-GGT GCT GAG GAT GAG GC-3′.

Validation by PCR-Next Generation Sequencing (NGS) sequencing is described below. Amplicons for 107 T > N sites were designed to span the target mutation such that the mutation position could be reached with 81bp Illumina reads. The same amplicon primers were used for all sites. To control for amplification among all 16 samples, 10 randomly selected heterozygous germ-line amplicons were also amplified and sequenced together with the mutations. All amplicons were pooled. PCR amplicons were purified followed by library construction. The PCR fragments were treated with T4 DNA polymerase, T4 polynucleotide kinase, and Klenow DNA polymerase for end repairing, followed by treatment with Klenow fragment 3′–5′ exo and dATP to generate a protruding 3′ A for ligating with the adaptor carrying a six-base index. The indexed DNA samples were run on 2% agarose gels, and fragments of 170–180 bp were recovered and purified. DNA of 16 PCR pools with different indexes were mixed in an equal molar concentration and amplified by PCR for six cycles.

The average observed error rate in the flanking positions for the libraries was 0.0005/base. The mutant allele frequency of each somatic mutation was considered only if there were at least 100 reads from Illumina sequencing.

A5.3 Inference on somatic mutation rate among tumors. To infer the overall rate of tumor specific mutations, we restrict the analysis to a small fraction of sites with the highest quality. Thus, the observed number of tumor specific mutations represents a fraction of the total number of mutations in the tumor sample. We first calculated the efficiency (or probability) of a site being selected by our analysis pipeline. The number of mutations in the entire genome can then be estimated from the observed number divided by the efficiency.

Efficiency analysis There are three major criteria used in selecting tumor specific mutations. The total efficiency of our filtering process is the product of three individual efficiencies associated with each major criterion. These criteria are: (*i*) coverage in nontumor and tumor samples are not lower than 20; (*ii*) nonreference allele count in nontumor (N0) is zero; (*iii*) mutant coverage in SOLiD and GAIIx in our tumor samples are no less than two and the total mutant frequency is no less than 40%. Only batch two data were used for estimating the genome wide mutation number (Table S2).

i) With the required read depth of no less than 20 coverage in both N0 and R1 samples, we effectively only used 38% of the genome. In other words, our efficiency associated with the first criterion is ∼0.38.

ii) Our candidate sites are required to have no nonreference reads in N0 sample. In our data, 12.6% of the sites show at least one nonreference read. Because the percentage of heterozygous sites in a single individual is ∼1/1000, sites that appear to be polymorphic in our data are mostly due to artifacts from mapping and sequencing errors. Typically, sequencing and mapping errors are thought be random across the genome. Thus, efficiency for picking out a potential tumor specific mutation is 0.874.

iii) Our third criterion is that data from both SOLiD and GAIIx platforms have to result in at least two nonreference reads and total percentage of mutant reads is no less than 0.4 (Table S2). For a site with given mutant frequency, the probability of this site meeting our criteria can be calculated from binomial distribution. For example, a given site with frequency x and coverage n, the number of mutant reads is approximately binomial (n,x). Let's denote numbers of mutant reads in SOLiD and GAIIx data as m_{SOLiD} and m_{GAIIx} respectively. The probability of a site meeting our criteria can be expressed as Prob $[m_{SOLiD} \geq 2 \; ; \; m_{GAIIx} \geq 2; \; m_{SOLiD+mGAIIx} \geq 0.4 \times (n_{SOLiD} + n_{GAIIx})]$.

Because m_{SOLiD} follows binomial (n_{SOLID}, x), likewise for m_{GAIIx}, the above probability can be explicitly calculated. In this setting, we used the "0.5 × Composition Index" (see A8 of this *SI*) as the parameter value for x (0.324). Empirical observed value for n_{SOLiD} and n_{GAIIx} are used in the probability calculation. After all, the estimated efficiency with this step is 0.19.

Estimation of the total number of mutations With the results given above, the overall efficiency is $0.38 \times 0.87 \times 0.19 = 0.0628$. Because we found 158 tumor specific mutations (batch two data, Fig. S1) in the regions covered, the estimated total number of SNVs is 158/0.0628 ∼2,500. We also expect slightly fewer than 21 (∼2,500 × 1% × 0.83) nonsynonymous point mutations in this survey with two additional considerations: The sequencing depth suggests that 83% of the coding region was covered and slightly less than 1% of the genome consists of nonsynonymous sites.

A6. Detection of Other Simple Changes (Small Indels and HBV Integrations). ***A6.1 Small indel detection and validation.*** We used BWA (5) to align short reads to the reference genome and used Samtools package to call candidate indel variants. We count a region as indel if it has at least two reads called as an indel in tumor samples but no read in nontumor tissues. For tumor-specific small indels, we excluded all small indels that are in the UCSC dbSNP (dbSNP130), Database of Genomic Variants (8), the Asian genome (Yanhuang) (9) and the Korean genome (10).

A6.2 HBV integration and validation. We selected the paired reads whose one end mapped onto the human reference genome and the other end mapped onto the HBV reference genome. When HBV integrates into the human genome, multiple reads of this

type will cluster around a focal genomic region. Thus, the HBV insertion locations can be determined by the ends which were mapped onto the human genome.

A7. Detection of CNVs and chromosomal indels. A7.1 Read depth and minor allele frequency data We developed a method combining information from both read depth (coverage) and minor allele frequency (MAF) at germ-line heterozygous sites to detect chromosomal copy number variations.

We first called germ-line heterozygous sites by screening out those sites with at least two nonreference reads in both SOLiD and GAIIx data during N0 sample's whole genome sequencing. Candidate sites that overlap with dbSNP130 were extracted as our set of germ-line heterozygous sites for this individual. Afterward, to reduce variance in allele frequencies, we partitioned the set of SNPs along each chromosome into nonoverlapping bins of size 50. Within each bin, we merged 50 germ-line heterozygotes sites into one unit by taking averages of MAFs.

Pileup results from both SOLiD and GAIIx data were used to extract read depths (coverage) information. Breakpoints of indels were identified by detecting abrupt inflection point in both read depth and MAFs. GLAD (Gain and Loss Analysis of DNA) algorithm which employs the Adaptive Weights Smoothing procedure (11) was adopted for automatic detection of breakpoints. Both depth and MAF ratios (R1/N0) were used to identify breakpoints of chromosomal aneuploidy using GLAD algorithm, respectively (12). Overlapping breakpoints representing simultaneous change in depth and MAFs were retained as the candidates to be the final list of chromosome copy number variations.

A7.2 Large scale deletion detection using Mate-Paired reads We developed a unique statistical method for detecting large scale deletion using information from the library insert size. When a deletion event occurs, clustered reads of abnormal insert length will appear surrounding focal point of deletion. Assuming the library insert sizes associated with mate-paired reads follow a normal distribution $N(\mu, \sigma)$, a deletion event will lead to following patterns (1) aberrated library insert sizes will deviate from the original distribution but follow a new normal distribution with a shifted mean value $N(\mu + u', \sigma)$ (2). One side read from these aberrant paired reads would cluster at a local genomic region and span a genome segment of size approximately μ. The frequency of a deletion can be estimated as the proportion of reads with aberrant library insertion size over all of the reads in the spanning region.

To focus on aberrant mate-paired reads derived from somatic indels, we filtered out deletion mutations in public database (e.g., Human Genome Variation Database http://www.hgvbaseg2p.org/, as well as the Asian genome sequence). Top hits with high frequency in tumor samples but (nearly) absent in adjacent nontumor tissues were selected and verified using Sanger sequencing.

A7.3 Detection of C5orf51 and CPEB4 To detect the truncation and fusion transcript derived from C5orf51 and CPEB4, we performed RT with an oligo(dT) primer and total RNA isolated from T0 and N0 sections followed by PCR with forward primer: 5′- ATA TTG TTG TTT AGA CAT TAT CTG -3′, and reverse primer 5′- AAG TGA AGC CAA CTG TTT AG -3′. The primers for fusion gene detection were designed based on mate-paired reads.

A8. Calculation of the composition index Composition Index (CI) is defined as the proportion of cells that carry the mutant allele of interest within a sample. In other words, CI is a product of sample purity (i.e., that percentage of cells that are tumor cells) and the proportion of tumor cells carrying the mutation. For diploid cell populations, only one copy of the genome is mutated. (Hepatocyte stem cells are diploids.) Thus, the mutant allele frequency observed in a sample is CI × 0.5. We selected 84 validated sites that are not in regions of chromosomal indels (Fig. S3) and calculated the average frequency of these mutants across sites. CI is twice the mean mutant frequency. For R2, T0, and R1, the estimated CI is 0.88, 0.75, and 0.65, respectively.

SI Results

B1. Summary of Sequencing Data. For exon capture sequencing, we were able to get 2.0 billion, 2.3 billion, and 2.6 billion bp in N0, R1, and R2, respectively. After mapped short reads onto reference genome hg18, the average coverages in the coding region were 48.4×, 56.2×, and 60.3 × . For whole-genome sequencing, more than 100 billion bases (>36×) raw data and 19.6 × and 20.2 × mappable data are collected for both normal (N0) and cancer samples (R1) using SOLiD and GAIIx platforms.

B2. Simple Mutations. Candidate sites that are high in R1 and zero in N0 (T > N sites) were selected and validated by Sequenom as described in *Materials and Methods*. Mutant frequencies at these sites were measured by Sequenom with three experimental replications across the nine tumor and seven nontumor sections. The relative frequencies of these sites in R1 and N0 are presented in Fig. S2.

The frequencies of tumor-specific mutations in all sections are given in Dataset S1A. From the validation results of Dataset S1A, we could estimate the mutation rate for the whole genome, as described in A5.3 of this *SI*. The estimate is ~0.8/Mb.

In total, 101 T > N sites were also confirmed by PCR-NGS sequencing at an average depth of >8,000 × . The frequencies of tumor-specific mutations across the 16 sections, and the coverage for each mutation site are given in Dataset S1B. Mutation frequencies in the tumor sections estimated by Sequenom and PCR-NGS sequencing are in good agreement, and the correlation coefficient ranges from 0.86 to 0.89. The mutation frequencies at T > N sites in the N0 section are generally 0, as determined by PCR-NGS sequencing. For these sites, Sequenom sometimes give false positives, mostly <10%, in the N0 section.

Among the nonsynonymous point mutations detected in this study, only two are foreground mutations. They are CCNG1, which is private to R2, and P62 found only in R1. To further confirm their frequencies among tumor sections, we did Sanger sequencing to double check the variants in addition to second generation sequencing results. The pattern we observe by Sanger Sequencing is in agreement with the second generation sequencing and the Sequenom genotyping.

Through mate-paired reads, 18 genomic positions are found to have HBV integrations. They are distributed on chromosome 2, 4, 5, 9, 10, 11, 13, 14, and 19, and supported by 882 reads in total. Three of the 18 positions (chr5-727053, chr11-70069883, and chr11-69011259) were chosen for finer mapping of breakpoints by PCR and Sanger Sequencing. All three integration sites yield positive results and the junctional sequences will be provided upon request.

B3. CNVs and Chromosomal Indels. The inferred chromosomal indels are shown in Fig. S3 and the more refined physical locations of their breakpoints are listed in Table S3.

Δ5q heterogeneity among tumor sections. Among all of the indels detected in R1, only Δ5q (M10) has not been fixed among all tumor sections. The evidence presented below shows that it is missing in R2 and in a lower frequency in T3/T6 than in other tumor sections. In Fig. S4, we plotted the mutant allele frequencies at the germ-line heterozygous sites across the genome. If a genomic region bears no chromosomal deletions, mutant allele frequencies at these sites will stay ~0.5. When a deletion occurs, mutant allele frequencies in the focal segment will deviate from 0.5. We can see that mutant allele frequencies in the R2 sample at Δ5q stay close to 0.5, about the same as in the nontumor tissues (N0), indicating that Δ5q is absent or in very low frequency in R2. Similar to the top panel, the mutant allele

frequencies for these sites in T3 and T6 are also different from the rest of the primary tumor sections (T1, T2, T4, and T5). The frequencies in T3 and T6 are much closer to 0.5 in the region spanned by Δ5q, suggesting that T3/T6 samples has a much lower frequency of Δ5q. Across the entire genome, Δ5q is the only chromosomal indel differentiating various tumor sections.

The two truncated genes at both ends of Δ5q are fused into a new transcript, which can be detected by RT-PCR in R1 and T0. The fused structure, referred to as M10, is depicted in Fig. S5.

M1, M2, and M3 and their relationship to Δ5q. M1, M2, and M3 (Table 1) are the three mutations located in the region spanned by Δ5q. All three mutations are in high frequency in R2 (Dataset S1D) and low frequency in other tumor sections. Interestingly, the frequencies of M1 and M2 are much higher in T3 and T6 than in other T sections (Dataset S1D). Recall that the T3 and T6 sections contain cells that also lack the Δ5q deletion (Fig. S4, *Lower*), we hence suggest that M1/M2 are in fact background mutations common to all tumor samples. They remain in high frequency in R2 as well as some T3/T6 cells but were deleted along with Δ5q in R1 and other primary tumor sections (Fig. 3). In contrast, M3, also located in the region spanned by Δ5q, is invariant among T1–T6 sections suggesting that M3 is indeed an R2-specific mutation.

B4. Mutations in the nontumor sections. Candidate sites that are low in R1 but high in N0 (n > T sites) were selected and genotyped by Sequenom (*SI Materials and Methods A5* and Fig. S1). Mutant frequencies at these sites were also measured by Sequenom with three experimental replications across 16 samples. The frequencies of these nontumor-specific mutations are listed in Dataset S1C.

In comparison with that of the 194 tumor specific mutations, the frequency profile of N0 specific mutations show a very different pattern. These sites have a dense distribution of mutation frequency close to 0.5 in all nontumor tissues. In contrast, the mutation frequencies in the tumor samples, albeit <0.5, are never close to 0. The mutation frequency at these sites in the tumor samples falling between 0 and 0.5 can be explained by the following observation: Checking these sites along the genome, we found that all of the sites clustered within large chromosomal indels. The pattern can be seen most clearly at the bottom of each panel in Fig. S3. Each red tick represents the location of an n > T site chosen. Due to chromosomal indels, if the deleted copy bears the mutant allele, mutant frequencies at these sites would be <0.5 in the tumors but nearly 0.5 in the nontumor tissues. Hence, these n > T sites are actually germ-line heterozygotes.

B5. Identification of adaptive mutations. The question is how we infer that M3 (cyclin G), M4 (P62), and Δ5q(M10) (a large indel creating a fusion gene) are responsible for the proliferation of the π1–π3 lineages in Fig. 3 (note that the three lineages correspond roughly to sections from the R2, primary, and R1 tumors).

For the π1 and π2 lineages, they need to be contrasted with the π0 lineage (which consists of a small fraction of cells from the T3/T6 sections of the primary tumor; Fig. 3). If π1 and π2 did not have M3 and Δ5q(M10), respectively, they should have grown just as slowly as the π0 cell lineage with very few cells. The next question would naturally be whether π1 and π2 may have other mutations besides M3 and Δ5q(M10). Two lines of evidence suggest M3 and Δ5q(M10) are likely the only ones. First, with a total of 22 coding mutations, we expect fewer than one (22 × 3.5%) nonsynonymous mutation in the proliferative phase. Second, by exon capture and deep sequencing, we indeed found only one mutation when 85% of coding mutations should have been discovered. Hence, M3 and Δ5q(M10) are likely the only coding mutations in these lineages. Their known functions also fit well with the possible roles in proliferation.

The role of M4 in the proliferation of the π3 lineage requires a bit more data. After all, π2 cells, from which π3 cells emerged, are themselves aggressively growing cells. (Note that π3 cells proliferate 15 mo after π2 cells were surgically removed.) One could argue that M4 had nothing to do with the proliferation of π3 cells, even though M4 (p62) has been known to be important in tumorigenesis (13).

To address this issue, we resequenced the entire coding regions of the T0 section to look for mutations that might have existed in the primary tumor but are absent in the R1 sample. From the new data, we found that all nonsynonymous mutations present in the primary tumor are indeed found in R1 as well. This new observation shows that the π3 cells of R1 are the direct descendants of the π2 cells and only those π2 cells that acquired a new nonsynonymous mutation, M4, proliferated. Despite their initial larger number, cells without M4 did not proliferate after surgery.

1. Chen PJ, et al. (1989) Clonal origin of recurrent hepatocellular carcinomas. *Gastroenterology* 96:527–529.
2. Chen YJ, et al. (2000) Chromosomal changes and clonality relationship between primary and recurrent hepatocellular carcinoma. *Gastroenterology* 119:431–440.
3. Goto Y, Yoshida J, Kuzushima K, Terashima M, Morishima T (1993) Patterns of hepatitis B virus DNA integration in liver tissue of children with chronic infections. *J Pediatr Gastroenterol Nutr* 16:70–74.
4. Takada S, Gotoh Y, Hayashi S, Yoshida M, Koike K (1990) Structural rearrangement of integrated hepatitis B virus DNA as well as cellular flanking DNA is present in chronically infected hepatic tissues. *J Virol* 64:822–828.
5. Li H, Durbin R (2010) Fast and accurate long-read alignment with Burrows-Wheeler transform. *Bioinformatics* 26:589–595.
6. Fay JC, Wyckoff GJ, Wu CI (2001) Positive and negative selection on the human genome. *Genetics* 158:1227–1234.
7. Ewens W (1979) *Mathematical Population Genetics* (Springer, New York).
8. Zhang J, Feuk L, Duggan GE, Khaja R, Scherer SW (2006) Development of bioinformatics resources for display and analysis of copy number and other structural variants in the human genome. *Cytogenet Genome Res* 115:205–214.
9. Wang J, et al. (2008) The diploid genome sequence of an Asian individual. *Nature* 456:60–65.
10. Ahn SM, et al. (2009) The first Korean genome sequence and analysis: full genome sequencing for a socio-ethnic group. *Genome Res* 19:1622–1629.
11. Polzehl J, Spokoiny V (2000) Adaptive weights smoothing with applications to image restoration. *J R Stat Soc, B* 62:335–354.
12. Hupé P, Stransky N, Thiery JP, Radvanyi F, Barillot E (2004) Analysis of array CGH data: from signal ratio to gain and loss of DNA regions. *Bioinformatics* 20:3413–3422.
13. Moscat J, Diaz-Meco MT (2009) p62 at the crossroads of autophagy, apoptosis, and cancer. *Cell* 137:1001–1004.

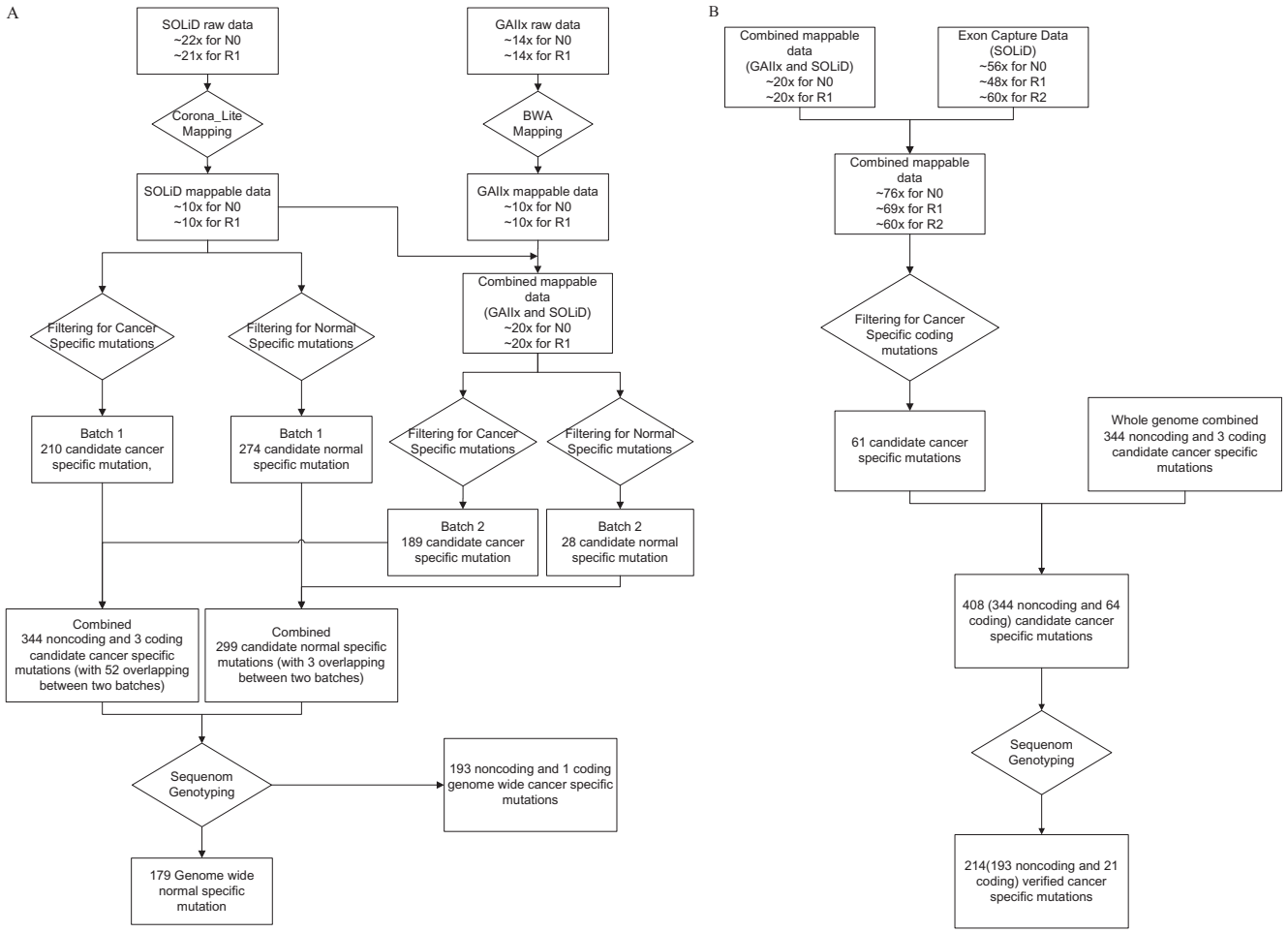


Fig. S1. The pipeline for point mutation detection. (*A*) Flowchart for filtering and verifying tumor (R1)/nontumor (N0) specific mutations. We picked two batches of tumor-specific mutations for validation during the data production progress. We used two platforms, SOLiD and GAIIx, to sequence the samples. The batch one sites were selected using only our SOLiD data. The batch two sites were gathered with all of the data from both SOLiD and GAIIx platforms. We picked mutations that are in high frequency at tumor sample R1 but absent in the adjacent nontumor sample N0, as candidates of tumor-specific mutations. Symmetrically, nontumor (N0) specific candidates were screened based on similar criteria (Table S2). Candidate sites were then individually genotyped using teh Sequenom platform. In total, 194 tumor-specific mutations and 179 nontumor-specific mutations were validated. (*B*) Data flow for picking tumor specific mutation combing genomic and exon capture data. To increase power in picking tumor-specific mutations in coding regions, we pooled genomic sequences together with exon capture reads and screened coding tumor-specific mutations based on less stringent criteria than whole-genome screening (Table S2). In total, 61 candidate tumor-specific sites were selected. Combined with results from genome-wide data (344 noncoding and 3 coding mutations), 408 sites (344 noncoding and 64 coding) were gathered in the end. After Sequenom validation, 214 tumor-specific mutations (193 noncoding and 21 coding) were verified.

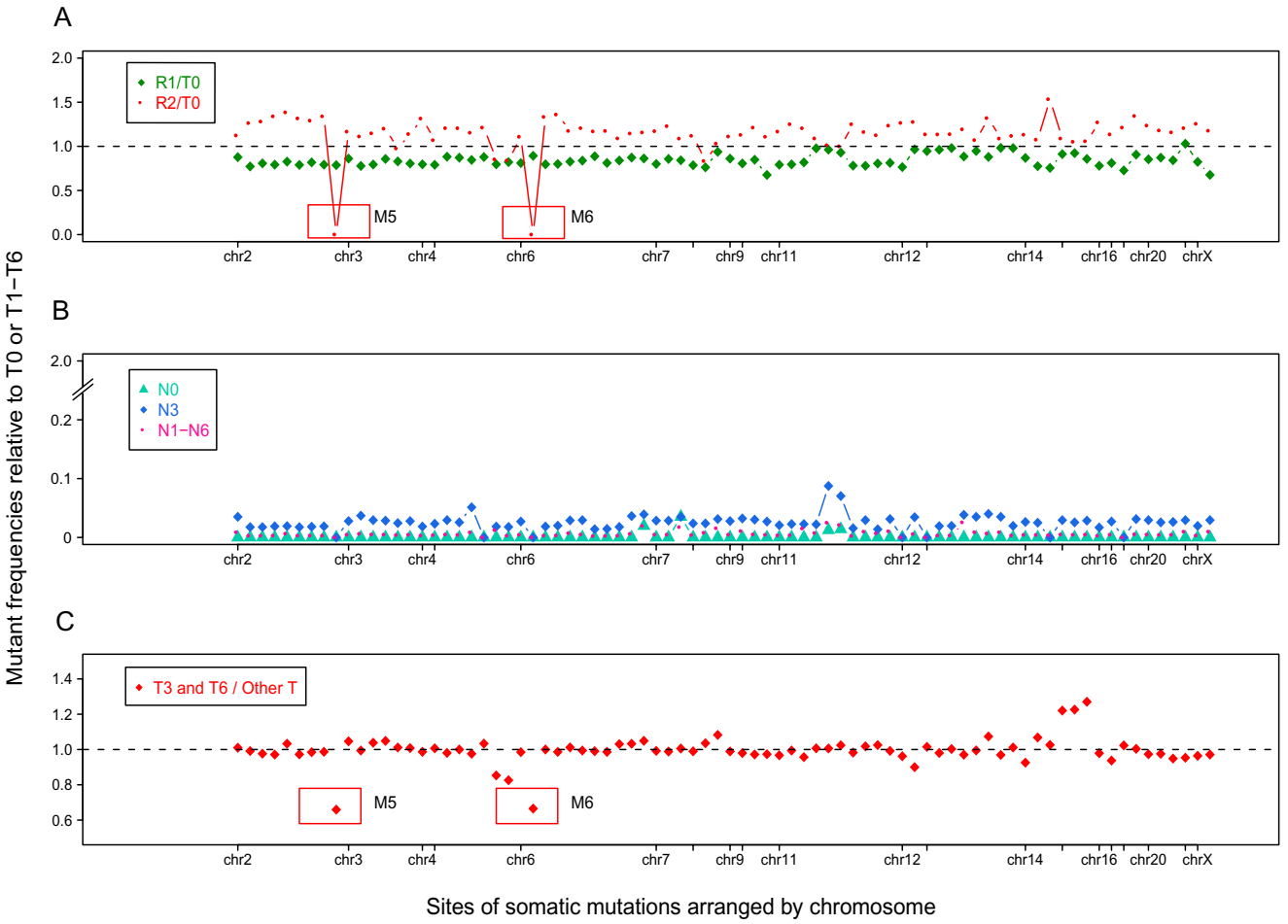


Fig. S2. Frequency profiles of somatic mutations at sites where the frequencies were high in the tumor and near zero in the nontumor tissues (T > N sites). These frequencies were validated by both Sequenom and PCR-NGS sequencing, which has an average coverage of >8,000× . The latter results are shown here. To avoid cluttering, only 86 of the 193 silent point mutations of Table S1 are shown. Sites are presented by chromosomes. Frequencies of tumor and normal sections are shown separately in the first two panels where the frequencies were normalized by those of the T0 section, which averaged 0.38 (Dataset S1A). In *A*, two mutations, labeled M5 and M6, were exceptionally low or absent in the R2 section. These same two mutations were also unusually low in frequency in the T3 and T6 sections, as shown in *C*. Note that the *y* axis in *B* is in a finer scale to show the near zero frequencies. In the N3 section, the mutation frequency at many sites appears to be higher than those in other nontumor sections.

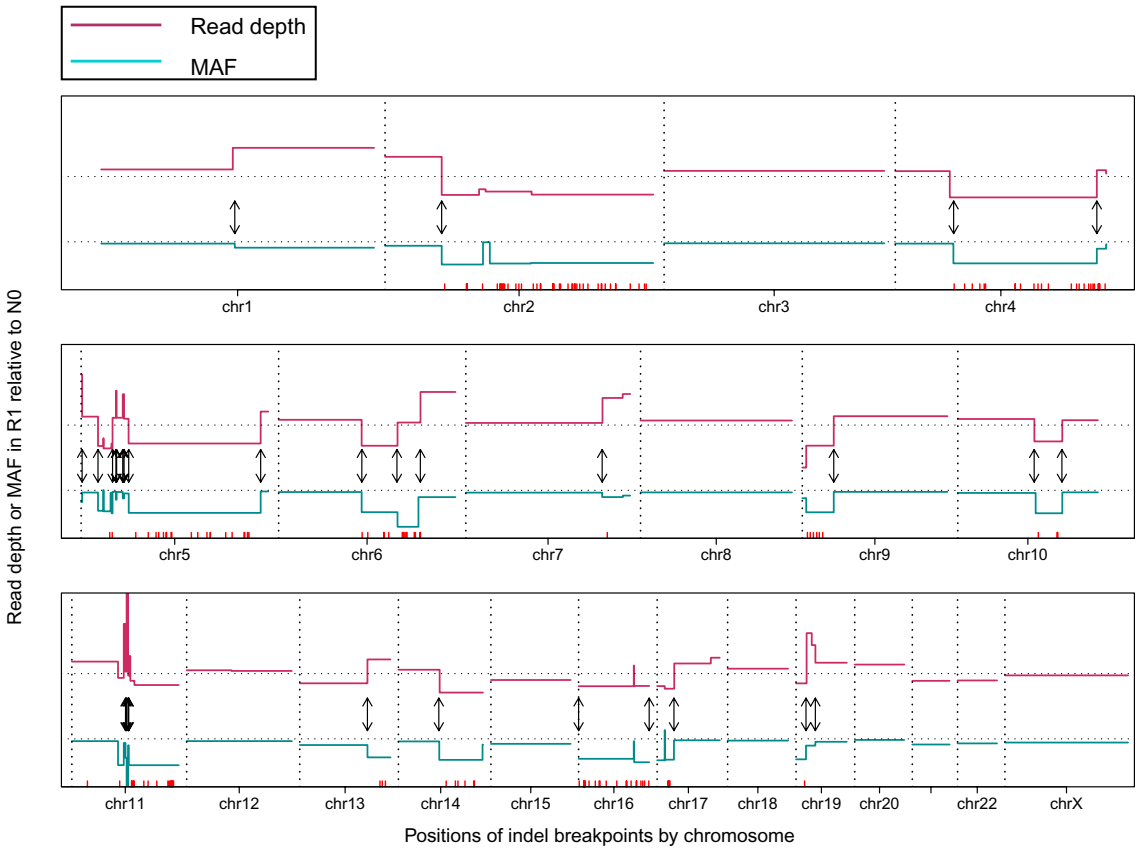


Fig. S3. Summary of chromosomal indels in the genome of R1. The *x* axis shows chromosomal location and *y* axis represents the read-depth (red line) and MAF (blue line) in R1 relative to those in N0. Breakpoints of indels are characterized by abrupt jumps in both read depth and MAFs. The GLAD algorithm, including the Adaptive Weights Smoothing procedure, was used to process the ratios between R1 and N0. Those that are jointly called by both types of data are marked by double-arrows and considered the true boundaries of chromosomal indels (Tables S1 and S3). Thick double-arrows represent two adjacent arrows too close to draw separately. The relative levels in read depth and MAFs suggest that some of the indels are not fixed in the tumor tissues. Sites shown at the bottom with red ticks are those used for the MAF analysis in other tumor sections.

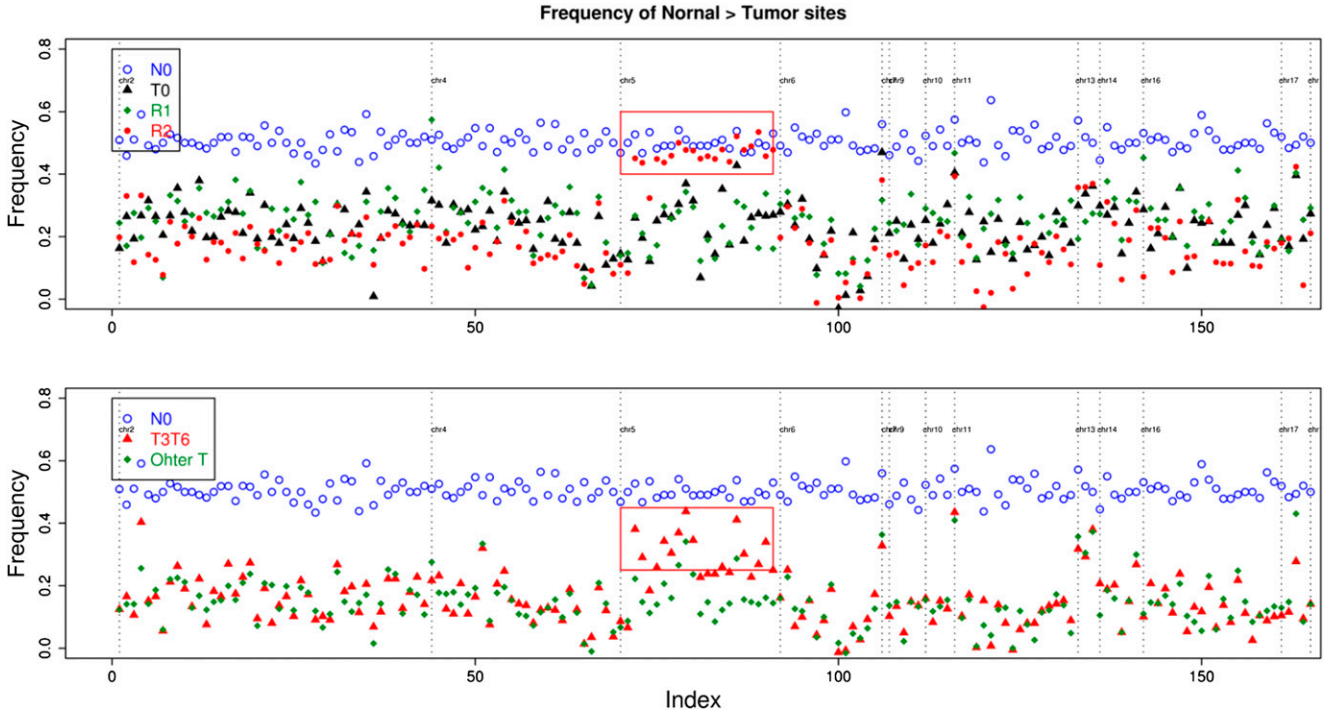


Fig. S4. Mutant frequency profiles for n > T sites across samples. The 179 n > T sites are all located in chromosomal deletions found in tumor samples. In the upper panel, we can see that mutant frequencies for Δ5q region in R2 sample are very similar to nontumor tissues. This is also the only chromosomal indel that shows this pattern. In the lower panel, Δ5q deletion is in lower frequencies in T3/T6 than in the other four tumor sections (T1,T2,T4, and T5) sections (*SI Results B3*).

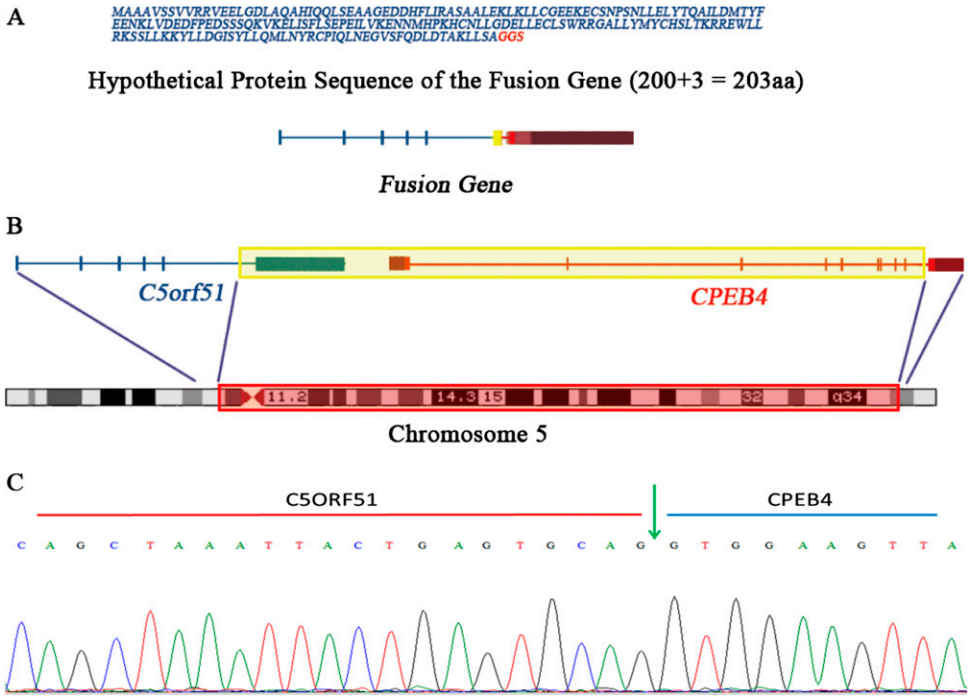


Fig. S5. A fusion gene of C5orf51 and CPEB4 from Δ5q. Hypothetical fusion protein resulted from Δ5q deletion are shown in the *A*. This fused transcript was confirmed by RT-PCR. The sequence result of RT-PCT is shown in *C*. The fused transcript is made of the last exon of CPEB4 together with a truncated C5orf51 gene. The truncated C5orf51 is missing its last exon. The CPEB4 gene only contributed three amino acids because of a frameshift (because exon boundary is not codon boundary) and a subsequent prestop codon in the last exon of CPEB4. Hypothetically, this fusion gene generated a protein of 203 amino acids. In *B*, the red block indicates the deleted genomic region (from 41,951,991 to 173,314,862) on chromosome 5. The yellow block indicates the deleted region from the two genes. The two break points of the deletion are in the fifth intron of C5orf51 and the ninth intron of CPEB4.

Table S1. Validated tumor-associated mutations

Type of changes	Count	Description
All point mutations	214	207: 7 are background: foreground mutations.
Nonsynonymous	21	See Table 2 for detail. ~83% of nonsynonymous sites in the genome are covered by the validation. Two are foreground mutations (M3 and M4, see Fig. 3).
Silent	193	Include synonymous, intergenic, intronic and UTR mutations. Five are foreground mutations (M5–M9, see Fig. 3).
UTR	11	—
Intronic	62	—
Intergenic	120	—
Frame-shift indels in coding	2	—
Truncated or fused genes	1	The breakpoints of the Δ5q deletion truncate and fuse two genes. This event is labeled M10 in Table 1 and Fig. 3.
HBV integration	18	None in the coding regions
CNVs and chromosomal indels	22	(see Figs. 2 and S3)

Table S2. Criteria for selecting point mutations

	Sample	Coverage	Mutant reads	Mutant frequency	Screen dbSNP
Batch 1 (Genome-wide)	R1	R1 ≥ 10	R1 ≥ 2	R1 ≥ 0.5	Yes
		N0 ≥ 10	N0 = 0	N0 = 0	
	N0	R1 ≥ 10	R1 = 0	R1 = 0	
		N0 ≥ 10	N0 ≥ 2	N0 ≥ 0.5	
Batch 2 (Genome-wide)	R1	R1 ≥ 10	R1(SOLiD) ≥ 2 R1(GAIIx) ≥ 2	R1 ≥ 0.4	
		N0 ≥ 10	N0 = 0	N0 = 0	
	N0	R1 ≥ 10	R1 = 0	R1 = 0	
		N0 ≥ 10	N0(SOLiD) ≥ 2 N0(GAIIx) ≥ 2	N0 ≥ 0.3	
Nonsynonymous	R1	R1 ≥ 10	R1 ≥ 2	R1 ≥ 0.3	
		N0 ≥ 10	N0 = 0*	N0 = 0	
	R2	R2 ≥ 10	R2 ≥ 2	R2 ≥ 0.3	
		N0 ≥ 10	N0 = 0*	N0 = 0	

We selected two batches of tumor-specific mutations for validation during the data production progress. The batch one sites were picked using only our SOLiD data. The batch two sites were gathered with data from both platforms (plus GAIIx). These criteria integrate three pieces of information from read coverage, the number of supported mutant read as well as mutant allele frequency.* We also allowed N0 have one mutant reads, but that mutant allele has to be different from the ones observed in R1 or R2 samples.

Table S3. Location and mutation type of chromosomal indels/CNVs based on R1 whole genome sequencing data

Chromosome	Start position (Mb)	End position (Mb)	Type	Gain or loss
1	121.1	247.3	CI	Gain
2	51.3	242	CI	Loss
4	52.6	182.6	CI	Loss
4	182.6	191.3	CI	Gain
5	16.3	30.3	CI	Loss
5	33.6	34.3	CNV	Gain
5	40.2	41.5	CNV	Gain
5	45.8	173.5	CI	Loss
6	80.2	114.9	CI	Loss
6	114.9	137.1	CI	Loss and gain
6	137.1	170.9	CI	Gain
7	131.7	158	CI	Gain
9	0	30.4	CI	Loss
10	75.2	101	CI	Loss
11	63.6	68.9	CI	Gain
11	72.6	134.5	CI	Loss
11	68.9	70.7	CNV	Gain
13	85.1	114.1	CI	Gain
14	51.9	106.4	CI	Loss
16	0	88.8	CI	Loss
17	0	21.2	CI	Loss
19	12.5	24.2	CI	Gain

Nineteen chromosome Indels and three CNVs were found. Their physical locations (second and third column) and their change in copy number (the last column) are also listed.

Other Supporting Information Files

Dataset S1 (XLS)