Forward-reverse mutation cycles between stages of cancer development

Taobo Hu[1,9], Yogesh Kumar[1,9], Iram Shazia[1,9], Shen-Jia Duan[2,9], Yi Li[3], Lei Chen[4], Jin-Fei Chen[5], Rong Yin[6], Ava Kwong[7], Gilberto Ka-Kit Leung[7], Wai-Kin Mat[1], Zhenggang Wu[1], Xi Long[1], Cheuk-Hin Chan[1], Peggy Lee[1], Siu-Kin Ng[1], Timothy Y. C. Ho[1], Jianfeng Yang[1], Xiaofan Ding[1], Shui-Ying Tsang[1], Xuqing Zhou[1], Dan-Hua Zhang[2], the International Cancer Genome Consortium, En-Xiang Zhou[2], Lin Xu[6], Wai-Sang Poon[3], Hong-Yang Wang[4,10] and Hong Xue[1,8,10]

[1]Division of Life Science, Applied Genomics Centre and Centre for Statistical Science, Hong Kong University of Science and Technology, Clear Water Bay, Hong Kong;

[2]Department of General Surgery, The Second Xiangya Hospital, Central South University, Changsha, Hunan, China;

[3]Department of Surgery, The Chinese University of Hong Kong, Hong Kong;

[4]Eastern Hepatobiliary Surgery Institute, Second Military Medical University, Shanghai, China;

[5]Department of Oncology, Nanjing First Hospital, and Collaborative Innovation Center for Cancer Personalized Medicine, Nanjing Medical University, Nanjing, China;

[6]Jiangsu Key Laboratory of Cancer Molecular Biology and Translational Medicine, Jiangsu Cancer Hospital, Nanjing, China;

[7]Division of Neurosurgery, Department of Surgery, Li Ka Shing Faculty of Medicine, University of Hong Kong, Queen Mary Hospital, 102 Pokfulam Road, Hong Kong, China;

[8]School of Basic Medical Sciences and Clinical Pharmacology, China Pharmaceutical University, Nanjing, China

[9] These authors contributed equally to this work

[10] These authors jointly directed this work. Correspondence should be addressed to

Hong Xue
Division of Life Science
Hong Kong University of Science and Technology
Clear Water Bay, Hong Kong
hxue@ust.hk

Keywords

Single nucleotide variation; copy number variation; interstitial loss-of-heterozygosity; pre-cancer mutations; clonal evolution

Abstract

Earlier, prominent occurrences of interstitial loss-of-heterozygosities (LOHs) were found in different cancers as a type of single-nucleotide-variations (SNVs), at rates far exceeding those of the commonly investigated gain-of-heterozygosities (GOHs) type of SNVs. Herein, such co-occurrences of LOHs and GOHs were confirmed in 102 cases of four cancer types analyzed with three different next-generation sequencing platforms, comparing non-tumor, paratumor, and tumor tissues with white-blood-cell controls; and in 246 pan-cancer cases of whole-genome tumor-control pairs. Unexpectedly, large numbers of SNVs enriched with CG>TG GOHs and copy-number-variations (CNVs) proximal to these GOHs were detected in the non-tumor tissues, which were extensively reversed in paratumors showing prominent TG>CG LOHs with proximal CNVs, and less so in tumors to form forward-reverse mutation cycles. Lineage effects in the reversions, likely resulting from directional selection, supported a sequential rather than parallel mode of evolution as described in a 'Stage Specific Populations' model of cancer development.

Introduction

The progressive development of cancer has been investigated at the cytochemical and genetic levels[1-6], and the existence of early premalignant stages in the process has been suggested by precancer manifestations in terms of DNA, gene expression, protein and microscopic changes[7-14]. Genomic analysis has played an increasingly important role toward delineation of mutational events[15,16], and a recent study by us has led to the finding of massive single-base interstitial loss-of-heterozygosities (LOHs) in various types of cancers, pointing to widespread intersister chromosomal gene conversions arising from defective DNA double-strand break (DSB) repair in tumor development[17]. Thus the interplay between the gain-of-heterozygosity (GOH) and LOH types of single nucleotide variations (SNVs) requires examination along with copy number variations (CNVs) as mutational elements of cancer regarding the extent of possible reversion of forward mutations during cancer development.

In the present study, non-tumor, paratumor, primary tumor and metastatic tumor tissues in different types of solid tumors were compared with same-patient white-blood-cell controls based on massive parallel sequencing. Somatic mutations in both directions, *i.e.* GOHs, LOHs, CNV gains and CNV losses were examined residue-by-residue and window-by-window in order to detect the presence of mutation reversals in the development of cancer cells and to assess their biological significance. The results obtained suggested that forward-reverse cycles of mutations along with directional selection are important determinants of stage-specific cell population profiles in cancer development.

Results

Genotypic changes in non-tumor and paratumor tissues

White-blood-cells (B), tumor tissue (T), paratumor tissue (P) immediately adjacent to tumor, and more remote non-tumor tissue (N) were collected in twelve same-patient tetra-sample cases consisted of four breast carcinomas (BRCA), five stomach adenocarcinomas (STAD) and three hepatocellular carcinomas (LIHC) (Supplementary Table 1), and subjected to DNA analysis using the AluScan platform based on inter-Alu polymerase chain reaction (PCR) followed by massively parallel sequencing (MPS) as described in Methods. The genotype of a base residue was referred as a major allele (M) when it matched the genotype on human reference genome hg19, or as a minor allele (m) when there was no match, thereby enabling the identification of changes in the form of MM-to-Mm GOH ('GOH-M'), mm-to-Mm GOH ('GOH-m'), Mm-to-MM LOH ('LOH-M') or Mm-to-mm LOH ('LOH-m')[17]. Figs. 1a and 1b show the total number of residue-by-residue changes in the N-, P- or T-sample genomes relative to B-sample, *viz.* ΔNB, ΔPB or ΔTB respectively in terms of GOH-M, GOH-m and LOH (sum of instances of LOH-M and LOH-m). Since the numbers of GOH-M, GOH-m and LOH mutations were higher in ΔNB than in ΔTB, and comparable in ΔPB and ΔTB, both the N-sample and P-sample cells had to be regarded as premalignant or early malignant cells despite their normal morphology and expression of immunohistochemistry (IHC) markers, in contrast with T-sample cells showing enlarged nuclei (Fig. 1c) and reduced expression of IHC markers (Supplementary Table 1). Notably, a forward-reverse mutation cycle (FR-cycle) was commonly observed in the patch diagrams in Fig. 1b: GOHs were strongly reversed by LOHs (*e.g.*, G1 mutations reversed by L1 mutations), and LOHs strongly reversed by GOHs (*e.g.*, L14 reversed by G12). Moreover, the LOH mutations of Mm residues displayed pronounced lineage effects, favoring the formation of MM residues when the Mm residues themselves were derived from MM residues, but favored the formation of mm residues when the Mm residues were derived from mm residues (as in the LOH mutations highlighted by yellow triangles in the figure).

Notably, in Fig. 1b right panel, partition of germline Mm residues via LOH steps L13 and L14 yielded a greater MM/mm product ratio than partition via LOH steps L15 and L16, and greater still than partition via LOH steps L21 and L22, although in each instance MM products exceeded mm products (Fig. 2a). Since all these three successive partitions emanated from germline Mm residues, their diminishing MM/mm ratios could not be the consequence of lineage effects. Instead, because MM residues in the genome generally have been optimized for growth by prolonged evolution, they tend to be favored over mm residues. Accordingly, the finding of [L13/L14] > [L15/L16] > [L21/L22] plausibly could be the result of a longer period of positive selection for MM over mm products of LOH in the N-stage cells relative to P-stage cells, and likewise P-stage cells relative to T-stage cells, suggesting that MM was generally preferred over mm for optimized growth of cells.

When the trinucleotide-based mutational profile method[18] was employed to classify the GOHs and LOHs observed in the B-N-T-P tetra samples into the C>A, C>G, C>T, T>A, T>C and T>G groups, the results showed that C>T and T>C mutations were particularly prominent among both GOHs and LOHs, in keeping with the general expectation that transitions exceed transversions in SNVs (Fig. 1d). The C>T GOHs among the ΔNB changes displayed peak frequencies at the four NCG triplets, conforming to the 'Signature 1A' (marked by four solid arrowheads) common to cancers, and likely ascribable to the contribution of spontaneous deaminations of 5-methyl-cytosine at methylated CpG to form thymidine[18,19]. These deaminations would also explain the ~50% greater occurrence of C>T GOHs than T>C GOHs in the ΔNB changes. In support of this, Fig. 2b shows that although there were less CG dimers than other dimers among AluScan captured as well as whole-genome sequences (Supplementary Fig. 1), more CG dimers underwent SNV mutations than any other dimers. In Fig. 1d, all SNV frequency columns in the ΔNB tier were represented by a solid segment and an open segment: the mutations in the solid segments were reversed in the next ΔPN tier, whereas the open segments were unreversed. Both the C>T and T>C GOHs show large solid segments indicating their extensive reversals in the ΔPN changes; since the T>C LOHs in the ΔPN tier were mostly reversal of the C>T GOHs in the ΔNB tier, these T>C LOHs were likewise more abundant than ΔPN C>T LOHs, and showed four NTG peaks (marked by open-arrows), which may be referred to as a 'Signature 1A' LOH feature.

Fig. 2c summarizes the forward and reverse mutation occurrences in the B-N-P-T samples: more SNVs and CNVs occurred in N-stage (*viz.* sum of Types II, III, IVa and IVb patterns) than in P- and T-stages combined (*viz.* Type-I). Reversals of N-stage SNVs and CNVs (*viz.* sum of Types IVa and IVb patterns) were common, amounting to ~70% of N-stage SNVs or ~40% of N-stage CNVs; and far more of such reversals took place in P-stage (Type-IVa) than in T-stage (Type-IVb).

When another seventeen B-N-T trio sample sets consisted of one BRCA, two LIHC and fourteen non-small cell lung cancers (NSCLC) were analyzed with respect to

GOH and LOH changes in the N- and T-stage cells relative to B-stage cells (Fig. 3), the results obtained showed the same regularities as the B-N-P-T tetra samples: the B-genomes displayed much higher LOH (L5, L6) rates and GOH-m (G3, G4) rates than GOH-M (G1, G2) rates, strong lineage effects in LOH-partitions between MM and mm products (highlighted by yellow triangles), and prominent FR-cycles *viz.* L1 reversing G1, L4 reversing G3, G5 reversing L5, and G6 reversing L6.

Genotypic changes in primary and metastatic tumors

Figs. 4 and 5 compare the mutations observed in five cancer groups based on same-patient N-stage, T-stage and metastatic-stage (M-stage) samples: (i) AluScan group of two N-T-M trio sets analyzed with AluScan sequencing; (ii) WGS-Liver-M group of four trio sets of liver-to-lung metastasis analyzed by Ouyang *et al*[20] using whole genome sequencing (WGS); and sixty-seven trio sets involving brain metastases analyzed with whole exome sequencing (WES) by Brastianos *et al*[21], which were separated into (iii) thirty-eight WES-Non-Lung cancers, (iv) six WES-NSCLC-L (L = low in C>A GOHs) cancers and (v) twenty-three WES-NSCLC-H (H = high in C>A GOHs) cancers. Although the five N-T-M trio groups compared in Fig. 4a were analyzed using variously the AluScan, WES and WGS platforms, the ratios of the [ΔTN]/N and [ΔMN]/N counts both indicated that the rates of LOH far surpassed the rates of GOH-m, which in turn far surpassed the rates of GOH-M. All five groups also displayed pronounced lineage effects in Fig. 4b in the partitions of LOH mutations of Mm residues between MM and mm products (highlighted by yellow triangles).

In Fig. 5a, the relative prominences of ΔTN GOHs, ΔTN LOHs, ΔMT GOHs and ΔMT LOHs varied among the five different cancer groups. This could arise in part from biological dissimilarities between the sequences analyzed on the different platforms on account of their varied sequence coverages of the genome. The SNV sites observed in the five groups displayed non-identical distributions among the Genic, Proximal and Distal sequence zones[22], as well as non-identical replication timings during the cell cycle (Fig. 5b). The proportion of ΔTN GOHs that became reversed in the ΔMT changes, marked by solid segments of the GOH frequency bars in the ΔTN tiers, were highest in the AluScan group, also quite high in the WES-NSCLC-L group, modest in the WES-Non-Lung group and lowest in the WGS-Liver-M and WES-NSCLC-H groups, even though the WES-NSCLC-L, WES-Non-Lung and WES-NSCLC-H groups were all analyzed based on the WES platform[21].

The WES-NSCLC-H group was unique in its display of particularly eminent C>A GOHs. Previously, C>A transversions were linked to polycyclic aromatic hydrocarbons[23] and acrolein[24] in tobacco smoke. The 23 WES-NSCLC-H samples were derived entirely from smokers, in accord with smoking being a significant factor for their elevated C>A GOHs. However, the WES-NSCLC-L samples with much more subdued C>A GOHs included two non-smokers and four smokers, suggesting that smoking or high C>A GOHs could play a less important carcinogenic role in a minority of smokers.

Whole genome sequencing confirmed the abundance of interstitial LOHs

A total of 246 tumor-control pairs from International Cancer Genome Consortium (ICGC) collection of WGS data[25] were analyzed to yield LOH and GOH types of SNVs in each paired-samples. These included a panel of sixty-three pan-cancer cases (Pilot-63) (Fig. 6a), and twenty-two intrahepatic cholangiocarcinoma (ICC), eighty-six LIHC and seventy-five NSCLC cases (Fig. 6b), showing prominent LOHs in each instance. For the Pilot-63 dataset, the different ΔTB mutation counts in T-stage cells relative to B-stage cells (Fig. 6a left panel) yielded a rate ratio of 4,300 between LOH and GOH-M, which was comparable to the 5,400, 2,700 and 5,300 rate ratios observed in Fig. 1, Fig. 3 and earlier in Ref. 17 respectively, indicating a vastly greater rate of LOH than GOH in the cancer cells in all four instances. As well, in all four instances, the three mutation rates remain in the same order of $R_{LOH} > R_{GOH\text{-}m} > R_{GOH\text{-}M}$, with LOH rate being the highest. The massive interstitial LOH rates observed earlier based on AluScan data were thus confirmed by the ICGC Pilot-63 WGS dataset.

Evidence from mutational profiles for gene conversions in LOH production

For the ΔTB SNVs of Pilot-63 (Fig. 6a), the alteration-group plots of mutational profile in Fig. 6c revealed strikingly similar heights for the opposing C>T (blue) and T>C (pink) LOH bars in the right panel, but not for the opposing C>T (blue) and T>C (pink) GOH bars in the left panel, consistent with excess LOHs than GOHs. The context-group plots in Supplementary Fig. 2a show comparable rates for all pairs of opposing LOHs in contrast to the unequal rates for the opposing GOHs. The mutation-rate diagrams in Fig. 6d show the actual rates for all opposing GOH pairs and LOH pairs. These results are summarized in Fig. 6e, where over 61% of the rate ratios of opposing substitutions were greater than two for the GOHs (blue bars), but 100% were between one and two for the LOHs (striped red bars), supporting different underlining mechanisms for GOH and LOH. This divergence of the rate ratios between opposing GOHs and opposing LOHs was in accord with our proposal that the LOHs in cancer cells resulted mainly from DSB repairs through gene conversions[17]. In a DSB at a heterozygous C/T, gene conversion could yield either a C/C or T/T homozygous position at comparable rates, depending on which homologous chromatid bears the DSB. On the other hand, because GOHs depend on point mutations rather than gene conversions, this comparable-rate constraint would not apply to GOHs.

Distances between SNVs and recurrent CNVs

That the N-stage SNVs and CNVs in the B-N-P-T tetra samples both underwent active reversions, and more in P-stage than in T-stage (Fig. 2c) suggest some form of possible correlation between these two types of mutations. This was supported by Fig. 2d which shows that the sites of C>T GOHs with NCG context occurring in the ΔNB changes, and T>C LOHs with NTG context occurring in the ΔPN changes, were located particularly close to recurrent CNVs compared to mutations with other contexts or in other stages of change, $p < 10^{-7}$. Furthermore, these two groups of SNVs declined in old age (Fig. 2e), in resemblance to the decrease of global DNA methylation in old age[26]. The correlation between somatic CNVs with CpGe and MeMRE (Fig. 2f), the increased SNVs at CpG sites (Fig. 2b) and the high tendency of methylated CpG conversion to TpG[27], also pointed to some SNV-CNV relationships in the CNV process merits investigation.

Frequency classes of serial CNV changes

In the B-N-P-T tetra-sample cases, the status of any CNV in the N-, P- and T-stages could be CN-unaltered (U), CN-gain (G) or CN-loss (L) relative to its status in B-stage. Arranging in serial order, the CN-status found in the N-P-T stages (Supplementary Tables 5) yielded twenty-six different serial orders, and their frequencies fell into three classes (Figs. 7a and 7b). In the LUG order, for example, each CNV site was CN-loss in N-stage, CN-unaltered in P-stage and CN-gain in T-stage, and the total number of sequence windows in the B-N-P-T sequences analyzed that exhibited such an LUG order made up the frequency on the y-axis of Fig. 7a. The three frequency classes separated by vertical dashed lines in the Fig. were:

(I) *Class I (U=2)*: comprising six different orders, where a U status occurred in two of the N-, P- and T-stages.

(II) *Class II (U≤1)*: comprising eight different orders, where the U status occurred in no more than one of the N-, P- and T-stages.

(III) *Class III (Disadvantaged)*: comprising twelve different orders, where ten of them (*viz.* outside of LUG and GUL) included an abrupt double-dose change directly from G to L, or L to G in the order.

The plausible basis for these different classes could be straightforward: the CNV orders in Class I entailed minimal copy-number departures from the starting B-stage, and were therefore well tolerated; in comparison, the Class II of CNV orders incurred greater departures from U and were less well tolerated. Every CNV order in the Class III involved at least one double-dose change jumping either from G to L or from L to G between two successive stages of cancer development, a distinct disadvantage that led to their lowest frequencies.

The double-dose disadvantage explained the low frequencies of GLU, LGU, UGL, ULG, LGL, GLG, LGG, GGL, GLL and LLG, but not the low frequencies of LUG and GUL which fell into Class III even though they did not incur any double-dose copy number changes, in contrast to GUG and LUL which belonged to the more abundant Class II. The contrast indicates that lineage effects were important not only to LOH-partitions (Figs. 1b, 3b and 4b), but also to the frequencies of different CNV orders. In GUG, the G status of T-stage cells constituted a reversion to the G status of N-stage cells. In LUL, the L status of T-stage cells likewise constituted a reversion to the L status of the N-stage cells. Thus, both these reversions were favored by lineage effects, allowing GUG and LUL to join Class II even though they each incurred two CNV-status changes. In contrast, lineage effects acted against LUG and GUL, because the CNV-status of the T-stage cells in these cases was not a restoration of the CNV-status of the N-stage cells, thereby explaining their diminished frequencies. These lineage effects were observable when different sizes of sequence windows were employed for CNV identification: as shown in Fig. 7c, both the quotients (Q-value) of GUG/GUL and LUL/LUG greatly exceeded unity (marked by dashed red line), yielding significant lineage effects of $p < 10^{-16}$ for all sizes of sequence windows ranging from 50-500kb.

Discussion

In the present study, genomic sequence analysis employing the AluScan platform revealed that the paratumor region at <2cm from tumor's edge and the non-tumor tissue region at >2cm from tumor (see Methods) contained numerous GOH and LOH types of SNVs along with CNVs that pointed to these regions as premalignant or early malignant stages in cancer development despite their apparently normal cell morphologies. Residue-by-residue tracing of these mutations through the B-, N-, P- and T-stages pointed to much higher LOH rates and GOH-m rates than GOH-M rates (Fig. 1a) in accord with our earlier findings[17]. In addition, prevalence of FR mutation cycles was observed with strong lineage effects in the partition of LOHs between MM and mm products (Fig. 1b), and in the steeply unequal CNV frequencies between GUG and GUL and between LUL and LUG (Fig. 7c). These findings based on the B-N-P-T tetra samples, and supported by those from the B-N-T trio samples (Fig. 3), the N-T-M trio samples based on AluScan, WES and WGS (Figs. 4 and 5) and ICGC B-T paired samples based on WGS (Fig. 6) pointed to a cancer development sequence in the order of B-N-P-T, as represented in the Stage Specific Population (SSP) Model in Fig. 8a, in accord with the greater distance of the N-region than the P-region from the tumor. In this SSP model, while each of the N-, P- and T-stage cell populations could comprise multiple cell clones, a majority of the cell clones within the same stage would display largely similar mutational and morphological characteristics.

Fig. 8b shows that although the twelve B-N-P-T tetra-sample cases were derived from three types of solid tumors, they all displayed a largely similar mutational profile: the SNVs found at N-stage (σ_N) comprised different proportions of GOHs (red) and

LOHs (orange), and underwent substantial reversions in P-stage cells, followed by a much smaller number of additional reversions occurring in the T-stage cells. Altogether, that more N-stage SNVs were reversed in P-stage than in T-stage in all twelve samples amounted to a highly non-random observation ($p = 3.1\times10^{-12}$), in confirmation of the stage-specific difference in mutational activities between P- and T-stage cells. Overall, the numbers of forward SNVs decreased in P-stage compared to N-stage, and decreased even further in T-stage. The forward N-stage GOHs and LOHs underwent strong reversals in the P-stage, and far less so in T-stage. On the other hand, the forward P-stage GOHs and LOHs were strongly reversed in T-stage (Fig. 8c). The huge difference between the numbers of forward N-stage GOHs and LOHs reversed in P-stage and the numbers of forward N-stage GOHs and LOHs reversed in T-stage clearly attested to the highly dissimilar mutational activities of the N-, P- and T-stage cells. On account of the importance of microenvironment in carcinogenesis[6], such stage-specific characteristics could arise at least in part from the sharing of a common microenvironment by all cell clones within the same N-, P- or T-stage. The finding of an N-stage cell population characterized by high GOH type of SNVs and normal cell morphology (Fig. 8a) was consisted with the elevated SNV prevalences by 27-fold ($p < 0.001$) or 36-fold ($p < 0.0001$) observed in normal kidney cortices of subjects that were smokers or exposed to the environmental carcinogen aristolochic acid respectively[28], and pointed to the propensity of N-stage cells to carcinogenesis.

The pronounced reversions of N-stage SNVs in P-stage, P-stage SNVs in T-stage, and T-stage SNVs in M-stage (Figs. 1b and 3b) suggest that FR-cycle between successive development stages could be a common cellular evolution strategy for adjustment to the microenvironmental changes encountered during stage transition. In this regard, it is notable that, when a deviant *Drosophila melanogaster* population induced by extreme starvation was allowed to readapt to the ancestral culture environment, reversions of SNPs back to ancestral allele genotypes over fifty generations of evolution amounted to about 50%[29], comparable to the average level of SNV-reversions exhibited by P-stage cells in the form of Type-IVa changes in Fig. 2c.

In any event, the wide occurrence of FR-cycles over the N-, P- and T-stages suggests that these stages of cancer development could be burdened with excessive deleterious mutations that require alleviation through reverse mutations to ensure optimal growths. This could help explain the findings of transformed cells multiplying more slowly at low population density than untransformed cells, *e.g.*, upon transformation of a C3H/10T1/2CL8 fibroblast cell line derived from C3H mouse embryos by 3-methylcholanthrene, the transformed cells exhibited a saturation density 2-3 times that of untransformed cells, but generation times of 22 and 27 hours, *viz.* 40-70% longer than the 15.5 hours for the untransformed cells[30]. Likewise, NIH 3T3 cells displayed retarded growth at low density and increased saturation density preceding the formation of transformed loci[31], while the increased density attained might stem from reduced contact inhibition, and the longer generation times could be the result of excessive deleterious mutations.

The *FRMD4A* gene[32] was mutated in four BRCA cases and three STAD cases in the twelve tetra-sample cases, and *CAGE1* for cancer antigen 1[33] was mutated in BRCA, STAD as well as LIHC samples. SNVs recurrent in four out of twelve cases were detected for sixteen different genes (Fig. 8d). Based on the GOTERM and INTERPRO databases, pathway enrichment analysis show that SNVs in the B-N-P-T tetra samples were frequent in the cell-adhesion and protein kinase pathways (Fig. 8e). It was striking that all of the N-stage mutations in the cadherin N-terminal domain family persisted un-reversely throughout the P- and T-stages, pointing to the importance of this family of cell adhesion molecules at multiple stages of cancer development.

In conclusion, because cancers are driven by mutations, the nature of the mutations in the evolving cancer cells furnishes an appropriate basis for delineating the major stages of carcinogenesis. In the present study, the mutational profiles of the cell populations in the N-, P- and T-stage samples showed that N-stage cells harbored large numbers of SNV mutations, more GOHs than LOHs, which were enriched with NCG>NTG type of GOHs with associated CNVs. The P-stage cells displayed, relative to N-stage cells, more LOHs than GOHs. A major fraction of their LOHs represented reversals of the forward GOH mutations found in N-stage, and were enriched with NTG>NCG type of LOHs with associated CNVs. In the T-stage cells, the ratio between LOHs and GOHs was even higher than P-stage cells. At T-stage, there were numerous reversals of P-stage mutations but far fewer reversals of N-stage mutations. Such distinct mutational profiles of the different stages gave rise to FR-cycles of mutational changes between N- and P-stages and between P- and T-stages in a sequential SSP model of cancer development that was supported by pronounced lineage effects on the partition of LOHs between MM and mm products (yellow highlighted in Figs. 1b, 3b and 4b), as well as the relative frequencies of different orders of copy number changes between the N-, P- and T-stages (Fig. 7a), which could not be explained by early divergence of different cell clones that subsequently engaged in parallel evolution. While mutations increased diversity and contributed to the formation of stage specific cell populations, directional selection made evident by MM-over-mm preference in LOHs stemming from germline heterozygous sites and the reversal of SNVs and CNVs to their original genotypes, likely on account of the optimization of cell growth and multiplication, was essential in shaping the distinctive mutational landscapes of different stages of cancer development. In this basis, the possibility arises that early N- and P-stage cells which have not yet accomplished their requisite mutation reversals might be relatively deficient in growth and replication vigor, in which case it could be advantageous to target therapeutic interventions at these early stages of cells before they have accomplished their mutation reversals to become fully malignant, therapy-resistant cancers.

Methods

Tumor purity and histology

Tumor purity in all B-N-P-T tetra and B-N-T trio samples was estimated using VarScan software[34] and ‘absCNseq’ R package[35]. The ‘my.res.list’ function of absCNseq was applied with the following parameters: alpha.min = 0.2, alpha.max = 1, tau.min = 1.5, tau.max = 5, min.sol.freq = 0, min.seg.len = 0, qmax = 7, lamda = 0.5.

For histological and immunohistochemical staining (Fig. 1c and Supplementary Table 1), samples were taken from the tumor, the adjacent paratumor region (< 2cm from tumor), and the non-tumor region (> 2cm from tumor) of a BRCA patient. The samples were fixed in 4% paraformaldehyde, dehydrated, embedded in paraffin, sectioned and subjected to standard hematoxylin and eosin (HE) staining. Immunohistochemical staining for estrogen receptor (ER), progesterone receptor (PR) and human epidermal growth factor receptor-2 (HER2) were conducted following the conventional procedures as described[36].

DNA samples for AluScan sequencing

Written informed consent was obtained from each patient who participated in this study. Subject recruitment and sample collection were approved by the institutional ethics review boards of Hong Kong University of Science and Technology, Second Xiangya Hospital of Changsha, Chinese University of Hong Kong, Second Military Medical University of Shanghai, First Hospital of Nanjing, Jiangsu Cancer Hospital, and University of Hong Kong. DNA extraction and AluScan sequencing library preparation were performed as described previously[17,37]. White-blood-cells were treated as representative of germline controls in keeping with the recommendation by the Cancer Genome Atlas (TCGA) project[38]. The N- and P-stage tissues included and subjected to AluScan sequencing in this study were obtained as follows: N-stage tissue was collected at > 2cm from the edge of tumor in the vicinity of tumor, and P-stage tissue was collected at < 2cm from the edge of tumor. The AluScan cancer cases, designated as B-N-P-T, B-N-T or N-T-M sample-sets, were listed in Supplementary Table 1 with demographical and clinical information.

AluScan sequencing

AluScans of genomic regions flanked by Alu repetitive sequences were obtained by means of inter-Alu PCR as described[17,37], employing both Head-type and Tail-type Alu consensus-based primers to ensure capture of a vast number of inter-Alu amplicons. In brief, a 25-µl PCR reaction mixture contained 2 µl Bioline 10× NH4 buffer (160 mM ammonium sulfate, 670 mM Tris-HCl, pH 8.8, 0.1% stabilizer; www.bioline.com), 3 mM $MgCl_2$, 0.15 mM dNTP mix, 1 unit Taq polymerase, 0.1 µg DNA sample, and 0.075 µM each of the four following Alu consensus sequence-based PCR primers:

AluY278T18 (5’-GAGCGAGACTCCGTCTCA-3’);

AluY66H21 (5’-TGGTCTCGATCTCCTGACCTC-3’);

R12A/267 (5’-AGCGAGACTCCG-3’);

L12A/8 (5’-TGAGCCACCGCG-3’).

PCR was carried out at 95 °C, 5 min for DNA denaturation, followed by 30 cycles each of 30 s at 95 °C, 30 s at 50 °C, and 5 min at 72 °C, plus finally another 7 min at 72 °C. Amplicons were purified with ethanol precipitation, sequenced on the Illumina HiSeq platform at Beijing Genomics Institute (Shenzhen, China) and mapped to the hg19 reference human genome for downstream bioinformatic analysis.

WGS and WES raw data

Whole-genome sequencing (WGS) data generated from tumor-blood paired samples with the Illumina system by the International Cancer Genome Consortium (ICGC) and the Cancer Genome Atlas (TCGA) were downloaded in fastq format with permission (https://www.synapse.org/#!Synapse:syn2887117). These included the Pilot-63 set, and eighty-six hepatocellular carcinoma (LIHC), seventy-five non-small-cell lung cancer (NSCLC) and twenty-two intrahepatic cholangiocarcinoma (ICC) cases (Supplementary Table 10) with information accessible through the ICGC Data Portal (https://dcc.icgc.org). In addition, raw WGS data generated by Ouyang *et al*[20] from four hepatitis B positive LIHC patients having pulmonary metastasis were obtained along with data from same-patient liver tissue controls, and included in the N-T-M trio sample analysis as the WGS-Liver-M subset. Moreover, raw data of whole-exome sequencing (WES) from sixty-seven brain metastatic cancer patients were obtained from Brastianos *et al*[21]. Tissues sampled at > 2cm from the edge of tumors were used as normal control tissues in Ouyang *et al*[20] and Brastianos *et al*[21], and treated as non-tumor, or N-stage, samples in this study.

SNV calling

For the paired-end sequencing reads generated on the Illumina platform by the AluScan, WGS or WES methods, bioinformatics analysis including alignment, sorting, recalibration, realignment and removal of duplicates using BWA (Burrows-Wheeler Aligner, version 0.6.1)[39], SAMtools (Sequence Alignment/Map, version 0.1.18)[40] and GATK (Genome Analysis Tool-Kit, version 3.5) were performed for identification of single nucleotide variations (SNVs) according to the standard framework[41] as described previously[17,37]. The 'UnifiedGenotyper' module of GATK was employed for genotyping of SNVs. Only genomic sequence regions with enough coverage, *i.e.*, read depth > 8, were included in the analysis, and the following parameters were applied to filtrate for SNVs of different genotypes: major allele frequency ≥ 95% for the 'MM' loci; major allele frequency ≥ 30% and ≤ 70%, and QD ≥ 4 for 'Mm' or 'mn' loci; and minor allele frequency ≥ 95% and QD ≥ 20 for 'mm' loci. Strand bias estimated using Fisher's Exact Test (FS) was employed to ensure FS value ≤ 20 for both heterozygous 'Mm' or 'mn' loci and homozygous 'MM' or 'mm' loci.

For cancer cases with more than two samples from each patient, *i.e.*, the B-N-P-T tetra-sample set of twelve cases, the B-N-T trio-sample set of seventeen cases and the N-T-M trio-sample set of seventy-three cases, the above mentioned calling of SNVs were first performed for each sample of each case in the multiple sample sets. For each of the multiple-sample cases, only nucleotide positions conformed to all the above SNV calling criteria in every samples of the same patient were included in further analysis. Sites not covered in further analysis were arising from either lack of sequencing reads or failure to meet filtering criteria in any one of the samples of the same patient.

Mutational profiles of SNVs

Mutational profiles of SNVs were analyzed following the procedure developed by Alexandrov *et al*[18]. For each SNV site, the SomaticSignatures package[42] under R environment was employed to determine its preceding and following bases. The results were unnormalized for the observed trinucleotide frequencies in the human genome. The resulting mutation frequency profiles were illustrated in three different graphical presentations, *i.e.*, the alteration-group plot, context-group plot and mutation-rate diagram. Custom R scripts for drawing the three different presentations are available at GitHub website (https://github.com/hutaobo/ProfilePlots).

CNV calling and identification of recurrent CNVs

From AluScan data, the AluScanCNV software[43] was employed to call paired CNVs between B- and N-stage (ΔNB), between N- and P-stage (ΔPN), between P- and T-stage (ΔTP), between B- and P-stage (ΔPB), as well as between B- and T-stage (ΔTB) samples of the same patient in the B-N-P-T tetra-sample sets of twelve cases, using fixed window sizes ranging from 50- to 500-kb. The ΔNB, ΔPB and ΔTB were arranged sequentially to yield the twenty-six possible serial orders shown in Figs. 7a and 7b. To identify recurrent CNVs, all CNVs found in any sequence window of any of the twelve tetra-sample cases at any stages, including ΔNB, ΔPN and ΔTP, were aggregated. CNVs located in the recently identified Distal Zones[22] were removed from further analysis to reduce background noise introduced by less informative windows in the human genome. Only sequence windows where CNV was detected in six or more of the twelve patients were considered to harbor a recurrent CNV.

Co-localization of CNVT and CpGe/MeMRE

CpGe and MeMRE entries were downloaded from UCSC Genome Browser as described[22], and somatic CNV (CNVT) entries classified as "Copy Number Variants" were downloaded from COSMIC database (http://grch37-cancer.sanger.ac.uk/cosmic/download). The human genome was divided into tandem 2000-bp windows, and the average densities of CNVT breakpoints and base pairs in CpGe or MeMRE in each window were calculated. Thereupon the windows with zero CpGe or MeMRE density were removed to avoid error caused by missing data, and the remaining windows were separated into ten groups based on the percentile of CpGe or MeMRE density. Finally, the average CNVT breakpoint densities in the groups were plotted against the percentile CpGe or MeMRE density.

Mutation enrichment in genes and pathways

The results of variant analysis of AluScan data of the twelve tetra-sample cases were uploaded to BioMart of the Ensembl database to generate a list of their gene contents under R environment using the 'biomaRt' R package[44]. For the 'getBM' function, 'chromosome_name', 'start_position', 'end_position', 'external_gene_name', 'ensembl_gene_id' and 'description' were selected as attributes, with 'chromosomal_region' filter type, 'sublist' filter value, 'ENSEMBL_MART_ENSEMBL' biomart type, and 'grch37.ensembl.org' host. The resultant gene list was uploaded to DAVID Bioinformatics Resources 6.7[45] using 'Functional Annotation Tool' to obtain three lists of mutation enriched functional groups and pathways as annotated in the three databases GOTERM, InterPro and KEGG, respectively, with mutated genes specified for each groups and pathways. Only those functionally annotated groups and pathways yielding Bonferroni corrected *p*-value, Benjamini corrected *p*-value as well as FDR *q*-value all less than 0.05 were considered statistically significant.

Statistical analysis and data visualization

Statistical analyses were performed using R software (http://www.r-project.org). The significance probability (*p*) values were calculated by the two-tailed t.test or chisq.test functions in R, and the Pearson correlation coefficients (*r*) were calculated by the cor function in R. Figures were drawn using the ggplot2, lattice or ellipse package under R environment, except for Fig. 7b which was drawn using the Circos program[46].

Disclosure of potential conflicts of interest

The authors declare no competing financial or non-financial interests.

Author contributions

HX conceived and initiated the study; SJD, YL, LC, JFC, RY, AK, GKKL, DHZ, EXZ, LX, WSP and HYW organized and collected the clinical samples and data; TH, YK, IS, WKM, ZW, XL, CHC, PL, SKN, TYCH, JY, XD, SYT and XZ analyzed the samples and data; and HX, TH and SYT wrote the paper. The International Cancer Genome Consortium (ICGC) which led a worldwide cancer genomics collaboration including a project on pan-cancer analysis of whole genomes (PCAWG), contributed tumor-control sample pairs collected and sequenced by PCAWG to the present study.

Acknowledgments

The study was supported by grants to H. Xue from University Grants Council of Hong Kong SAR (ITS/113/15FP, VPRDO09/10.SC08, VPRDO14SC01, DG14SC02, SRFI11SC06 and SRFI11SC06PG), and grants to J. F. Chen (National 973 Basic Research Program of China, No. 2013CB911300; National Natural Science Foundation of China, No. 81272469; and Natural Science Foundation of Jiangsu Province special clinical project, No. BL2012016). Y. Kumar was a recipient of International Ph.D. Studentship from Hong Kong University of Science and Technology. X. Long was a recipient of Hong Kong Ph.D. Fellowship from Government of Hong Kong SAR. F.W. Pun was recipient of Research Fellowship from HKUST Jockey Club Institute of Advanced Study. Hututa Technologies Limited assisted with the computation facilities.

References

1. Nowell, P.C. The clonal evolution of tumor cell populations. *Science* **194**, 23-28 (1976).

2. Oosterhuis, J.W. & Looijenga, L.H. Testicular germ-cell tumours in a broader perspective. *Nat. Rev. Cancer* **5**, 210-222 (2005).

3. Bateman, C.M. *et al.* Acquisition of genome-wide copy number alterations in monozygotic twins with acute lymphoblastic leukemia. *Blood* **115**, 3553-3558 (2010).

4. Andor, N. *et al.* Pan-cancer analysis of the extent and consequences of intratumor heterogeneity. *Nat. Med.* **22**, 105-113 (2016).

5. Ding, L. *et al.* Clonal evolution in relapsed acute myeloid leukaemia revealed by whole-genome sequencing. *Nature* **481**, 506-510 (2012).

6. McGranahan, N. & Swanton, C. Clonal heterogeneity and tumor evolution: past, present, and the future. *Cell* **168**, 613-628 (2017).

7. Roncalli, M., Borzio, M., Brando, B., Colloredo, G. & Servida, E. Abnormal DNA content in liver-cell dysplasia: a flow cytometric study. *Int. J. Cancer* **44**, 204-207 (1989).

8. Fearon, E.R. & Vogelstein, B. A genetic model for colorectal tumorigenesis. *Cell* **61**, 759-767 (1990).

9. Libbrecht, L., Desmet, V., Van Damme, B. & Roskams, T. The immunohistochemical phenotype of dysplastic foci in human liver: correlation with putative progenitor cells. *J. Hepatol.* **33**, 76-84 (2000).

10. Wang, G.S., Wang, M.W., Wu, B.Y., You, W.D. & Yang, X.Y. A novel gene, GCRG224, is differentially expressed in human gastric mucosa. *World J. Gastroenterol.* **9**, 30-34 (2003).

11. Gorgoulis, V.G. *et al.* Activation of the DNA damage checkpoint and genomic instability in human precancerous lesions. *Nature* **434**, 907-913 (2005).

12. Cooper, C.S. *et al.* Analysis of the genetic phylogeny of multifocal prostate cancer identifies multiple independent clonal expansions in neoplastic and morphologically normal prostate tissue. *Nat. Genet.* **47**, 367-372 (2015).

13. Shen, L.J. *et al.* Detection of HBV, PCNA and GST-pi in hepatocellular carcinoma and chronic liver diseases. *World J. Gastroenterol.* **9**, 459-462 (2003).

14. Yadav, V.K., DeGregori, J. & De, S. The landscape of somatic mutations in protein coding genes in apparently benign human tissues carries signatures of relaxed purifying selection. *Nucleic Acids Res.* **44**, 2075-2084 (2016).

15. Beerenwinkel, N. *et al.* Genetic progression and the waiting time to cancer. *PLoS Comput. Biol.* **3**, e225 (2007).

16. Kuhner, M.K., Kostadinov, R. & Reid, B.J. Limitations of the driver/passenger model in cancer prevention. *Cancer Prev. Res. (Phila.)* **9**, 335-338 (2016).

17. Kumar, Y. *et al.* Massive interstitial copy-neutral loss-of-heterozygosity as evidence for cancer being a disease of the DNA-damage response. *BMC Med. Genomics* **8**, 42 (2015).

18. Alexandrov, L.B. *et al.* Signatures of mutational processes in human cancer. *Nature* **500**, 415-421 (2013).

19. Helleday, T., Eshtad, S. & Nik-Zainal, S. Mechanisms underlying mutational signatures in human cancers. *Nat. Rev. Genet.* **15**, 585-598 (2014).

20. Ouyang, L. *et al.* Whole-genome sequencing of matched primary and metastatic hepatocellular carcinomas. *BMC Med. Genomics* **7**, 2 (2014).

21. Brastianos, P.K. *et al.* Genomic characterization of brain metastases reveals branched evolution and potential therapeutic targets. *Cancer Discovery* **5**, 1164-1177 (2015).

22. Ng, S.K. *et al.* Feature co-localization landscape of the human genome. *Sci. Rep.* **6**, 20650 (2016).

23. Pfeifer, G.P. *et al.* Tobacco smoke carcinogens, DNA damage and p53 mutations in smoking-associated cancers. *Oncogene* **21**, 7435-7451 (2002).

24. Feng, Z., Hu, W., Hu, Y. & Tang, M.S. Acrolein is a major cigarette-related lung cancer agent: Preferential binding at p53 mutational hotspots and inhibition of DNA repair. *Proc. Natl. Acad. Sci. USA* **103**, 15404-15409 (2006).

25. International Cancer Genome, C. *et al.* International network of cancer genome projects. *Nature* **464**, 993-998 (2010).

26. Tsang, S.Y. *et al.* Variation of global DNA methylation levels with age and in autistic children. *Hum. Genomics* **10**, 31 (2016).

27. Poole, A., Penny, D. & Sjoberg, B.M. Confounded cytosine! Tinkering and the evolution of DNA. *Nat. Rev. Mol. Cell Biol.* **2**, 147-151 (2001).

28. Hoang, M.L. *et al.* Genome-wide quantification of rare somatic mutations in normal human tissues using massively parallel sequencing. *Proc. Natl. Acad. Sci. USA* **113**, 9846-9851 (2016).

29. Teotonio, H., Chelo, I.M., Bradic, M., Rose, M.R. & Long, A.D. Experimental evolution reveals natural selection on standing genetic variation. *Nat. Genet.* **41**, 251-257 (2009).

30. Reznikoff, C.A., Bertram, J.S., Brankow, D.W. & Heidelberger, C. Quantitative and qualitative studies of chemical transformation of cloned C3H mouse

embryo cells sensitive to postconfluence inhibition of cell division. *Cancer Res.* **33**, 3239-3249 (1973).

31. Rubin, H., Yao, A. & Chow, M. Heritable, population-wide damage to cells as the driving-force of neoplastic transformation. *Proc. Natl. Acad. Sci. USA* **92**, 4843-4847 (1995).

32. Goldie, S.J. *et al.* FRMD4A upregulation in human squamous cell carcinoma promotes tumor growth and metastasis and is associated with poor prognosis. *Cancer Res.* **72**, 3424-3436 (2012).

33. Kunze, E. & Schlott, T. High frequency of promoter methylation of the 14-3-3 sigma and CAGE-1 genes, but lack of hypermethylation of the caveolin-1 gene, in primary adenocarcinomas and signet ring cell carcinomas of the urinary bladder. *Int. J. Mol. Med.* **20**, 557-563 (2007).

34. Koboldt, D.C. *et al.* VarScan 2: somatic mutation and copy number alteration discovery in cancer by exome sequencing. *Genome Res.* **22**, 568-576 (2012).

35. Bao, L., Pu, M. & Messer, K. AbsCN-seq: a statistical method to estimate tumor purity, ploidy and absolute copy numbers from next-generation sequencing data. *Bioinformatics* **30**, 1056-1063 (2014).

36. Yu, X. *et al.* Neoalbaconol inhibits angiogenesis and tumor growth by suppressing EGFR-mediated VEGF production. *Mol. Carcinog.* **56**, 1414-1426 (2017).

37. Mei, L. *et al.* AluScan: a method for genome-wide scanning of sequence and structure variations in the human genome. *BMC Genomics* **12**, 564 (2011).

38. Cancer Genome Atlas Research, N. *et al.* The Cancer Genome Atlas Pan-Cancer analysis project. *Nat. Genet.* **45**, 1113-1120 (2013).

39. Li, H. & Durbin, R. Fast and accurate short read alignment with Burrows-Wheeler transform. *Bioinformatics* **25**, 1754-1760 (2009).

40. Li, H. *et al.* The Sequence Alignment/Map format and SAMtools. *Bioinformatics* **25**, 2078-2079 (2009).

41. McKenna, A. *et al.* The Genome Analysis Toolkit: a MapReduce framework for analyzing next-generation DNA sequencing data. *Genome Res.* **20**, 1297-1303 (2010).

42. Gehring, J.S., Fischer, B., Lawrence, M. & Huber, W. SomaticSignatures: inferring mutational signatures from single-nucleotide variants. *Bioinformatics* **31**, 3673-3675 (2015).

43. Yang, J.F. *et al.* Copy number variation analysis based on AluScan sequences. *J. Clin. Bioinforma.* **4**, 15 (2014).

44. Durinck, S., Spellman, P.T., Birney, E. & Huber, W. Mapping identifiers for the integration of genomic datasets with the R/Bioconductor package biomaRt. *Nat. Protoc.* **4**, 1184-1191 (2009).

45. Huang, D.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).

46. Krzywinski, M. *et al.* Circos: an information aesthetic for comparative genomics. *Genome Res.* **19**, 1639-1645 (2009).

47. Hansen, R.S. *et al.* Sequencing newly replicated DNA reveals widespread plasticity in human replication timing. *Proc. Natl. Acad. Sci. USA* **107**, 139-144 (2010).

Figure legends

Figure 1. SNV mutations in B-N-P-T tetra samples. (**a**) Genotypic changes at N-, P- or T-stage of samples. The numbers of genotypic changes in N-, P- or T-stage sequences relative to B-stage sequences are represented by ΔNB, ΔPB and ΔTB respectively. LOH represents the sum of LOH-M and LOH-m changes. The twelve B-N-P-T cases consisted of four breast carcinomas (BRCA), five stomach adenocarcinomas (STAD) and three hepatocellular carcinomas (LIHC) analyzed using AluScan sequencing (Supplementary Table 1). (**b**) Patch-diagrams tracing SNVs between the B-, N-, P- and T-samples originating from MM, mm or Mm residues in B-samples. Mutation rate is indicated below each LOH step (L1, L2, *etc.*) or GOH step (G1, G2, *etc.*). (**c**) Micrographs of N-stage tissue (left), P-stage tissue (middle) and T-stage tissue (right) in one of the representative BRCA B-N-P-T tetra samples. Magnification in each instance was 400. (**d**) Mutational profiles for the ΔNB, ΔPN and ΔTP SNV changes as numbered in the patch diagrams in Part **b**. The profiles are separated into the C>A, C>G, C>T, T>A, T>C and T>G types, where C>A includes both the C-to-A and the complementary G-to-T mutations, *etc*. Within each type, the 16 possible kinds of sequence contexts are indicated on an expanded scale on the x-axis, and the total number of SNVs observed for each kind of trinucleotide sequence contexts is represented by a vertical bar. In each vertical bar in the ΔNB tier, the solid segment represents the SNVs that were reversed in the next ΔPN tier, *e.g.* C>T GOHs being reversed by T>C LOHs, whereas the open segment represents the unreversed SNVs. Subgroups of contexts are compartmentalized by vertical dashed lines. Since residues of minor bases different from “m” were rare in the samples analyzed, mutations involving them are listed in Supplementary Tables 3 and 4 but not shown in the patch diagrams in Figs. 1b, 3b and 4b. M: major allele; m: minor allele; GOH-M:

MM-to-Mm mutation; GOH-m: mm-to-Mm mutation; LOH-M: Mm-to-MM mutation; LOH-m: Mm-to-mm mutation.

Figure 2. Properties of mutations in B-N-P-T tetra samples. (**a**) M-over-m preference in LOHs. LOH-M/LOH-m ratios for LOHs arising at germline Mm positions are shown for N-, P- and T-stages (data from Fig. 1b right panel). (**b**) Dinucleotides in genomic sequences captured by AluScan (blue bars), and SNVs found at the first base of dinucleotides (red bars). (**c**) Percentile of different types of sequential SNV and CNV changes in twelve tetra-sample cases of BRCA, STAD and LIHC. Type-I (green) B=N, *viz.* no SNV (or CNV) found in ΔNB. Type-II (yellow) B≠N=T, *viz.* same SNV (or CNV) found in ΔNB and ΔTB. Type-III (blue) B≠N≠T and B≠T, *viz.* altered in ΔNB and ΔTN, and also in ΔTB. Type-IVa (red) B≠N≠P=T and B=T, *viz.* altered in ΔNB and ΔPN, but not in ΔTP or ΔTB. Type-IVb (purple) B≠N=P≠T and B=T, *viz.* altered in ΔNB and ΔTP, but not in ΔPN or ΔTB. See Supplementary Tables 3 and 5 for detailed numbers of SNVs and CNVs at different stages. (**d**) Average distances between GOHs (left panel) or LOHs (right panel) and their nearest recurrent CNVs (rCNVs). Left panel shows that on average CG>TG GOHs among the ΔNB changes was closer to their nearest rCNVs than it was the case with eleven other kinds of C>T GOHs, with the five horizontal bars indicating how much closer in terms of *p* values. Likewise, right panel shows that on average TG>CG LOHs among the ΔPN changes was closer to their nearest rCNVs than it was the case with eleven other kinds of LOHs, with the five horizontal bars indicating how much closer in terms of *p* values. (**e**) Correlation between patient’s age at diagnosis and percentage of CG>TG GOHs among all GOHs for ΔNB (left panel), or percentage of TG>CG LOHs among all LOHs for ΔPN (right panel). (**f**) Correlation of numbers of somatic CNV breakpoints found in tumors (CNVT) from COSMIC database with either evolutionarily conserved CpG-rich regions (CpGe) (left panel), or unmethylated CpG-rich regions (MeMRE), from UCSC Table Browser database[22] (right panel). *p*: significance probability; *r*: Pearson correlation coefficient (See Fig. 1 for abbreviations)

Figure 3. SNV mutations in B-N-T trio samples. (**a**) Genotypic changes in N- or T-stage of samples. The numbers of genotypic changes in N- or T-stage sequences relative to B-stage sequences are represented by ΔNB and ΔTB respectively. The seventeen B-N-T trio samples consisted of one BRCA, two LIHC and fourteen non-small cell lung cancer (NSCLC) cases. (**b**) Patch-diagrams tracing SNVs between the B-, N- and T-samples. (**c**) Mutational profiles for the ΔNB and ΔTN SNV changes as numbered in the patch diagrams in Part **b**. In each vertical bar in the ΔNB tier, the solid segment represents the SNVs that were reversed in the ΔTN tier, whereas the open segment indicates the unreversed SNVs. (See Supplementary Tables 6 and 7 for details of SNVs, and Fig. 1 for abbreviations)

Figure 4. SNV mutations in N-T-M trio samples. (**a**) Genotypic changes in T- or M-stage sequences. The numbers of genotypic changes in T- or M-stage sequences relative to N-stage sequences in each of five case groups are represented by ΔTN and ΔMN respectively. The five case groups include AluScan group of two N-T-M trio sets analyzed with AluScan sequencing; WGS-Liver-M group of four trio sets of liver-to-lung metastasis analyzed by whole genome sequencing (WGS); and sixty-seven trio sets involving brain metastases analyzed with whole exome sequencing (WES) which were separated into thirty-eight WES-Non-Lung cancers, six WES-

NSCLC-L (L = low in C>A GOHs) cancers and twenty-three WES-NSCLC-H (H = high in C>A GOHs) cancers. In the trio sets of WGS-Liver-M, WES-Non-Lung, WES-NSCLC-L and WES-NSCLC-H, because non-tumor tissues were sampled at > 2cm from tumor's edge as controls instead of blood cells, the samples were designated as 'N-stage' for comparability with Figs. 1a and 3a. (**b**) Patch-diagrams tracing SNVs between the N-, T- and M-samples. (See Supplementary Tables 8 and 9 for details of SNVs). M: metastatic tumor. (Liver = hepatocellular carcinoma; NSCLC = non-small-cell lung cancer; Non-Lung = all 12 types of solid tumors in Ref. 21 except lung adenocarcinomas and lung squamous carcinomas; See Fig. 1 for abbreviations)

Figure 5. Properties of SNVs in N-T-M trio samples. (**a**) Mutational profiles for the ΔTN and ΔMT SNV changes as numbered in the patch diagrams in Fig. 4b. In each vertical bar in the ΔTN tier, the solid segment represents the SNVs that were reversed in the ΔMT tier, whereas the open segment indicates the unreversed SNVs. (**b**) Zonal distribution and replication timing of the SNVs in the ΔTN tier and the ΔMT tier in five case groups. Left panels: zonal classification of SNV (GOH and LOH) sites determined as described[22]. Right panels: replication timing scores derived from ENCODE at UCSC based on the "Repli-chip" method[47]. (See Figs. 1 and 4 for abbreviations)

Figure 6. Properties of ICGC-whole genome sequencing of tumor (T)-blood (B) pairs. (**a**) Left panel: numbers of different genotypic changes between T- and B-stage sequences (ΔTB) in ICGC Pilot-sixty-three pan-cancer samples. Right panel: rates of different types of SNV mutations. For Pilot-63 samples, the rates were derived from WGS ΔTB/B values in the left panel; for B-N-P-T AluScan samples, from ΔTB/B values in Fig. 1a; for B-N-T AluScan samples, from ΔTB/B values in Fig. 3a; for Ref. 17 samples, from AluScan results on six types of cancers[17]. (**b**) Scatter plot of numbers of GOHs and LOHs in twenty-two intrahepatic cholangiocarcinoma (green triangles), eighty-six LIHC (red circles) and seventy-five NSCLC (blue squares) pairs of samples (Supplementary Table 10). (**c**) Alteration-group plots of mutational profiles for GOHs and LOHs between B- and T-stages in Pilot-63 samples. Each plot was grouped using the six alteration types: C>A, C>G, C>T, T>A, T>C, and T>G. (**d**) Representative mutation-rate diagrams, GOH-3 (left panel) and LOH-3 (right panel), were generated from the ICGC Pilot-63 dataset (See additional nine pairs of diagrams in Supplementary Fig. 2). In these diagrams based on triplet duplexes, the rates of opposing mutations from Part **c** were paired and labeled on bidirectional arrowheads with same color-codes as in Part **c**, *e.g.*, the rates of ACG>ATG GOH and ATG>ACG GOH changes (arrowhead-marked in Part **c** left panel) were dissimilar (229 in blue versus 95 in pink), but the rates of LOH changes of the same triplet duplexes were similar (1617 in blue versus 1558 in pink, and arrowhead-marked in Part **c** right panel). (**e**) Distribution of the observed rate ratios between opposing GOH pairs (the blue columns) and LOH pairs (the striped red column). (See Fig. 1 for abbreviations)

Figure 7. Serial orders of copy number (CN) changes in B-N-P-T tetra samples. (**a**) Frequencies of three classes of sequential orders of CN changes. Total numbers of CNV windows conformed to each serial order of changes are plotted out for all twenty-six possible orders. The CN-status of 500-kb windows of N-, P- or T-stage was determined relative to B-stage for twelve B-N-P-T tetra samples as described in Methods. The total numbers of windows that conformed to the different possible serial orders of CN changes (Supplementary Table 11) provided the basis for their partition into three frequency classes (separated by two vertical dashed lines). U stands for CN-unaltered, G for CN-gain and L for CN-loss. (**b**) Circos diagram showing the chromosomal distribution of different CNV serial orders. The twenty-six different orders are arranged from the rim, showing chromosome banding, inward in accordance to their left-to-right arrangement in Part **a**, *viz.* ULU (blue) in the outermost ring followed by LUU (red), UGU (blue) *etc.*, where orders with a CN-status of U at N-stage are colored blue, and orders with a CN-status of G or L at N-stage colored red. Classes I-III are separated by dashed circles in black. (**c**) GUG/GUL and LUL/LUG frequency ratios determined using different window sizes. Without lineage effects, these ratios would be expected to yield Q=1 (as represented by the red dashed line). In contrast, all the Q-values of GUG/GUL (blue) and LUL/LUG (red) observed in sequence windows varying from 50-kb to 500-kb significantly exceeded unity (*, $p < 10^{-16}$). (See Fig. 1 for abbreviations, B, N, P and T)

Figure 8. Mutational features of different stages of cancer development. (**a**) Stage Specific Population (SSP) model of cancer development. (**b**) Reversal of N-stage SNVs at P- and T-stages in twelve tetra-sample cases of BRCA, STAD and LIHC. Colored sectors represent N-stage GOHs (red), N-stage LOHs (orange), P-stage reversals of N-stage SNVs (light green), and T-stage reversals of N-stage SNVs (dark green), respectively. σ_N shows the total number of SNVs occurred at N-stage compared to B-stage cells with relative proportions of GOH and LOH shown by colored sectors. σ_P and σ_T show the proportions of N-stage SNVs unreversed and reversed at P-stage and T-stage, as color-coded sectors. For N-stage SNVs, P-stage reversion significantly exceeded T-stage reversion (p = 3.1×10^{-12}). (**c**) Forward and reverse SNVs in N-, P- and T-stage samples. *Left*: N-stage SNVs comprising forward GOHs and LOHs (solid red). *Middle*: P-stage SNVs comprising forward GOHs and LOHs (solid green); and GOHs and LOHs that reversed N-stage LOHs and GOHs respectively (striped red). *Right*: T-stage SNVs comprising forward GOHs and LOHs (solid blue); and GOHs and LOHs that reversed P-stage LOHs and GOHs (striped green), and N-stage LOHs and GOHs (striped red highlighted by arrows), respectively. See Supplementary Table 3 for data used in this plot. (**d**) Frequently mutated genes in the twelve tetra-sample cases. Each mutation found in a BRCA (orange), STAD (green) or LIHC (blue) sample is represented by a square listed with the gene. (**e**) Pathway enrichments of mutations in the twelve tetra-sample cases. The horizontal bars showing numbers of mutated genes are arranged from top down in the order of increasing Bonferroni corrected p-values for pathways in GOTERM and INTERPRO databases, respectively. The % of genes in the pathway that displayed one or more forward SNVs at N-, P- or T-stage is indicated under ΔNB, ΔPB and ΔTB respectively. The % of N-stage SNVs that were reversed subsequently at P- or T-stage was indicated by (R_{NB}). The % of P-stage SNVs that were reversed subsequently at T-stage was indicated by (R_{PB}). (See Supplementary Tables 12 and 13 for mutated genes and pathways, and Fig. 1 for abbreviations)

Supplementary Information

Supplementary Table 1. Information on 105 samples analyzed by AluScan.

Supplementary Table 2. Tumor purities of B-N-P-T tetra and B-N-T trio samples estimated by absCN-seq.

Supplementary Table 3. Summary of SNV mutations in B-N-P-T tetra samples.

Supplementary Table 4. The exact residue-by-residue SNV mutations in each sample of the B-N-P-T tetra-sample cases.

Supplementary Table 5. Summary of CNV mutations in B-N-P-T tetra samples.

Supplementary Table 6. Summary of SNV mutations in B-N-T trio samples.

Supplementary Table 7. The exact residue-by-residue SNV mutations in each sample of the B-N-T trio-sample cases.

Supplementary Table 8. Summary of SNV mutations in N-T-M trio samples. (**a**) AluScan; (**b**) WES-Non-Lung; (**c**) WES-NSCLC-L; (**d**) WES-NSCLC-H; and (**e**) WGS-Liver-M.

Supplementary Table 9. The exact residue-by-residue SNV mutations in each sample of the AluScan N-T-M trio-sample cases.

Supplementary Table 10. Numbers of GOHs and LOHs in 22 ICC, 86 LIHC and 75 NSCLC samples.

Supplementary Table 11. The exact window-by-window CNV mutations in each sample of the B-N-P-T tetra-sample cases.

Supplementary Table 12. List of genes harboring SNV mutations in B-N-P-T tetra-sample cases.

Supplementary Table 13. SNV mutations enriched pathways and genes in B-N-P-T tetra-sample cases.

Supplementary Figure 1. Total numbers of different dinucleotide sites in the human genome. Numbers of CG as well as other fifteen types of dinucleotides in human reference genome hg19 are plotted out.

Supplementary Figure 2. Mutational profiles of Pilot-63 samples from ICGC analyzed by WGS. (**a**) Context-group plots of GOH and LOH types of SNVs. Each plot is arranged by the ten context groups: A.A, C.C, A.G, C.A, A.C, G.A, C.G, A.T, T.A and G.C, designating the different immediate 5' and 3' flanking nucleotides. The opposing GOH mutations (left panel) or opposing LOH mutations (right panel) are placed side-by-side, *e.g.*, by pairing the ATG>ACG GOH (pink bar) with the ACG>ATG GOH (blue bar) in Section-3 of left panel, and similarly pairing the ATG>ACG LOH (pink bar) with the ACG>ATG LOH (blue bar) in Section-3 of right panel (marked by arrowheads and color-coded as in Fig. 6c). (**b**) Mutation-rate diagrams for GOHs. Each of the ten diagrams of triplet duplexes correspond to a context group, labeled 1-10 as in Part **a**. The mutation rates of opposing GOH mutations are labeled on double-headed arrows, except for the single-headed curved arrows in groups 7-10, where the two sequences are identical in a triple duplex. Each double-headed arrows is accompanied by two color-coded mutation rates that correspond to the heights of color-coded bars in Part **a**, *e.g.*, in context group 1, the conversion of double-stranded ACA/TGT to AAA/TTT is associated with a mutation rate of 162, colored red to correspond to the red C>A bar with A.A context in the left panel of Part **a**; whereas the opposing conversion of AAA/TTT to ACA/TGT is associated with a mutation rate of 641, colored orange to correspond to the orange A.C bar with A.A context in the left panel of Part **a**. (**c**) Mutation-rate diagrams for LOHs. The arrows employed are similar to those in Part **b**. All arrows in Parts **b** and **c** are shown as dashed lines for transitions (TSs) or solid lines for transversions (TVs). In the ten diagrams in Part **b** or Part **c**, the boxed TS/TV ratio given for each diagram represents the ratio pertaining to all the TS and TV mutations in the diagram, *e.g.*, in Diagram-1 of Part **b**, TS equals the sum of the four TS rates in the diagram, and TV the sum of the eight TV rates, yielding TS/TV = 1430/2804 = 0.51. The different rates in the diagrams in Parts **b** and **c** are color-coded as in Part **a**.

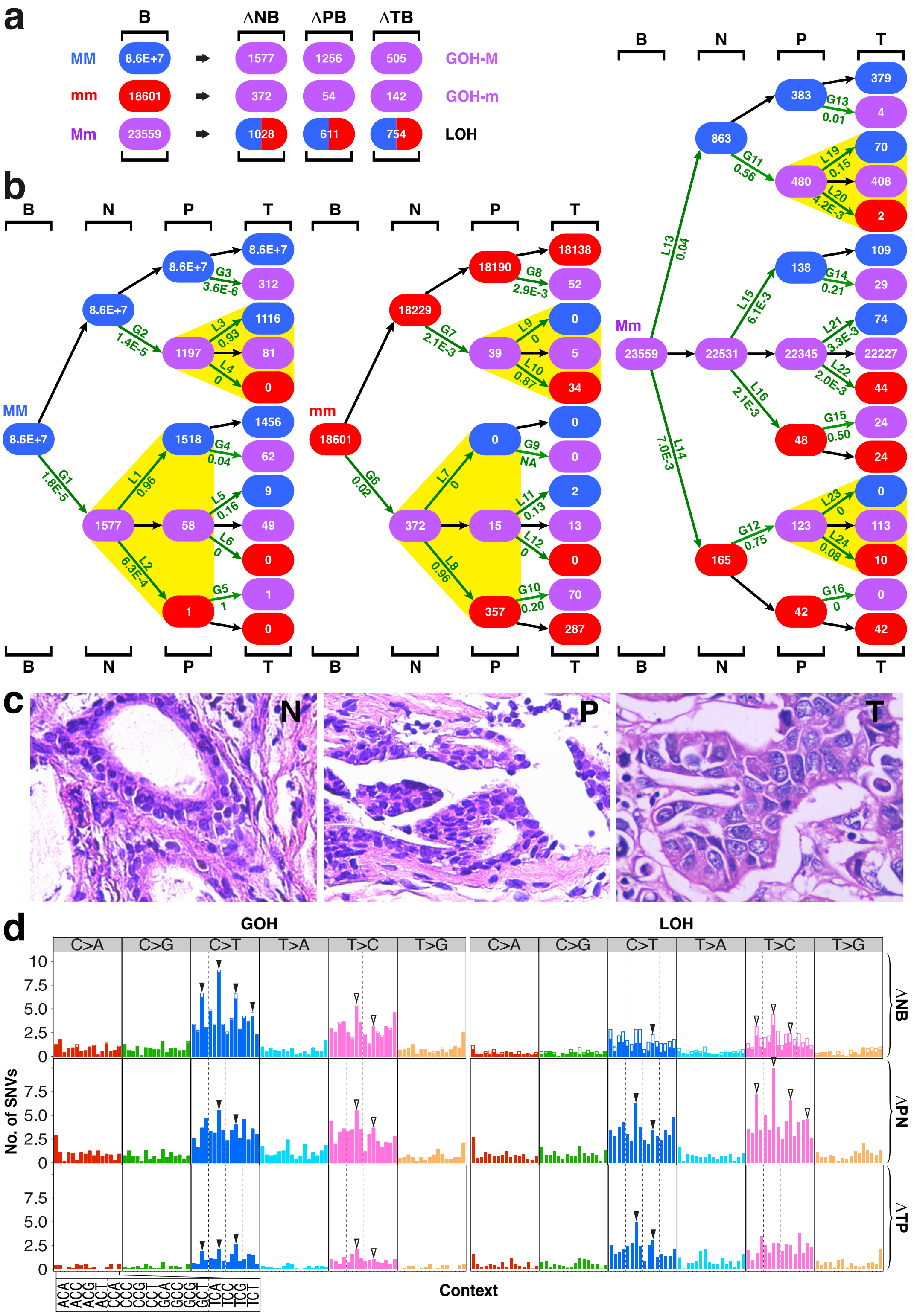
a
B
ΔNB
ΔPB
ΔTB
MM
8.6E+7
1577
1256
505
GOH-M
mm
18601
372
54
142
GOH-m
Mm
23559
1028
611
754
LOH
b
B
N
P
T
MM
8.6E+7
8.6E+7
8.6E+7
8.6E+7
G3 3.6E-6
312
G2 1.4E-5
1197
L3 0.93
1116
81
L4 0
0
G1 1.8E-5
1577
L1 0.96
1518
1456
G4 0.04
62
58
L5 0.16
9
49
L6 0
0
L2 6.3E-4
1
G5 1
1
0
mm
18601
18229
18190
18138
G8 2.9E-3
52
G7 2.1E-3
39
L9 0
0
5
L10 0.87
34
G6 0.02
372
L7 0
0
0
G9 NA
0
15
L11 0.13
2
13
L12 0
0
L8 0.96
357
G10 0.20
70
287
Mm
23559
L13 0.04
863
383
379
G13 0.01
4
G11 0.56
480
L19 0.15
70
408
L20 4.2E-3
2
22531
L15 6.1E-3
138
109
G14 0.21
29
22345
L21 3.3E-3
74
22227
L22 2.0E-3
44
L16 2.1E-3
48
G15 0.50
24
24
L14 7.0E-3
165
G12 0.75
123
L23 0
0
113
L24 0.08
10
42
G16 0
0
42
c
N
P
T
d
GOH
LOH
C>A
C>G
C>T
T>A
T>C
T>G
ΔNB
ΔPN
ΔTP
No. of SNVs
0
2.5
5.0
7.5
10
Context
ACA
ACC
ACG
ACT
CCA
CCC
CCG
CCT
GCA
GCC
GCG
GCT
TCA
TCC
TCG
TCT

Figure 1

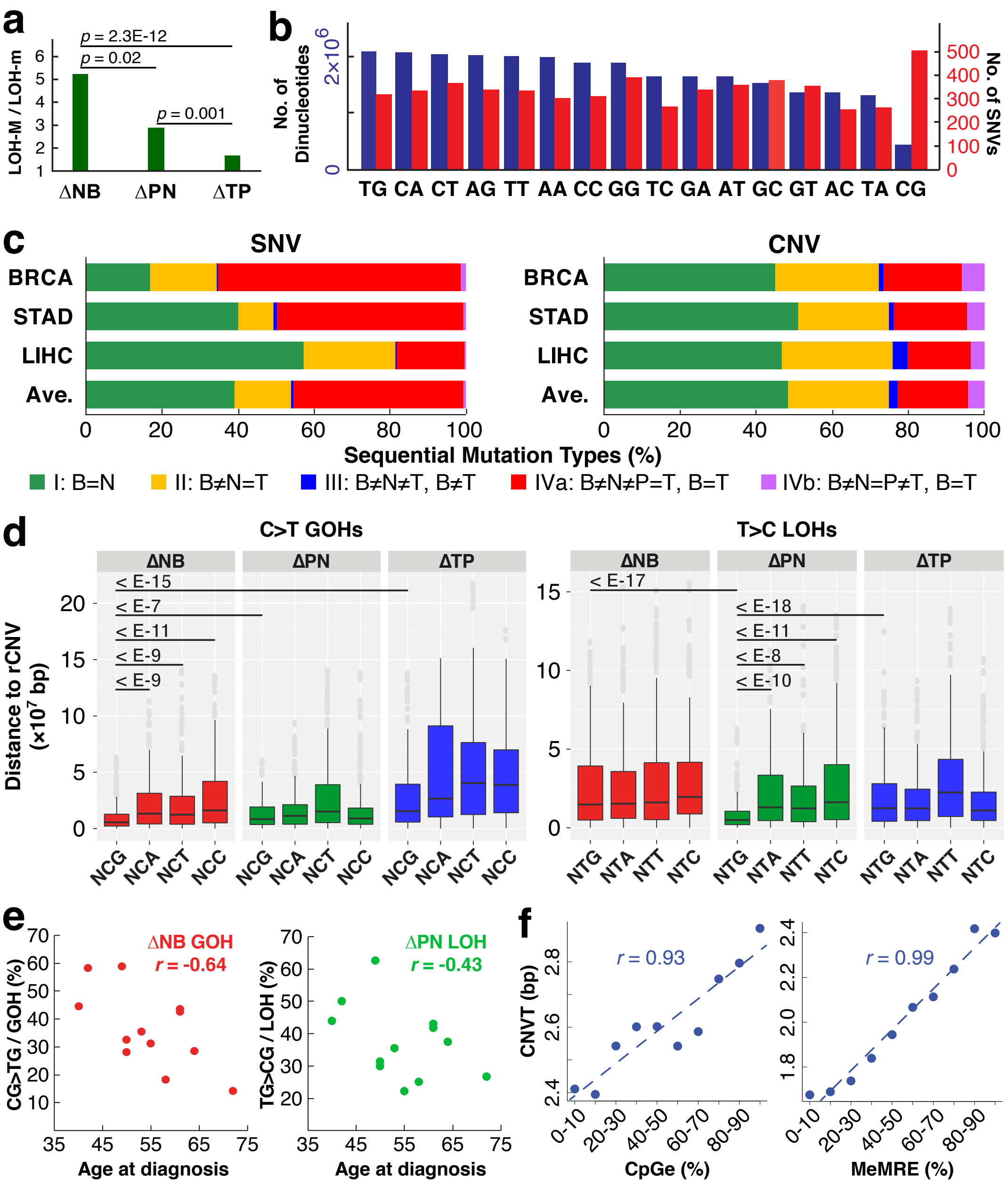
a
LOH-M / LOH-m
p = 2.3E-12
p = 0.02
p = 0.001
ΔNB
ΔPN
ΔTP
b
No. of Dinucleotides
2×10^6
0
No. of SNVs
500
400
300
200
100
0
TG CA CT AG TT AA CC GG TC GA AT GC GT AC TA CG
c
SNV
CNV
BRCA
STAD
LIHC
Ave.
0 20 40 60 80 100
Sequential Mutation Types (%)
I: B=N
II: B≠N=T
III: B≠N≠T, B≠T
IVa: B≠N≠P=T, B=T
IVb: B≠N=P≠T, B=T
d
C>T GOHs
T>C LOHs
ΔNB
ΔPN
ΔTP
< E-15
< E-7
< E-11
< E-9
< E-9
< E-17
< E-18
< E-11
< E-8
< E-10
Distance to rCNV (×10^7 bp)
NCG NCA NCT NCC
NTG NTA NTT NTC
e
ΔNB GOH
r = -0.64
ΔPN LOH
r = -0.43
CG>TG / GOH (%)
TG>CG / LOH (%)
Age at diagnosis
f
r = 0.93
r = 0.99
CNVT (bp)
CpGe (%)
MeMRE (%)
0-10 20-30 40-50 60-70 80-90

Figure 2

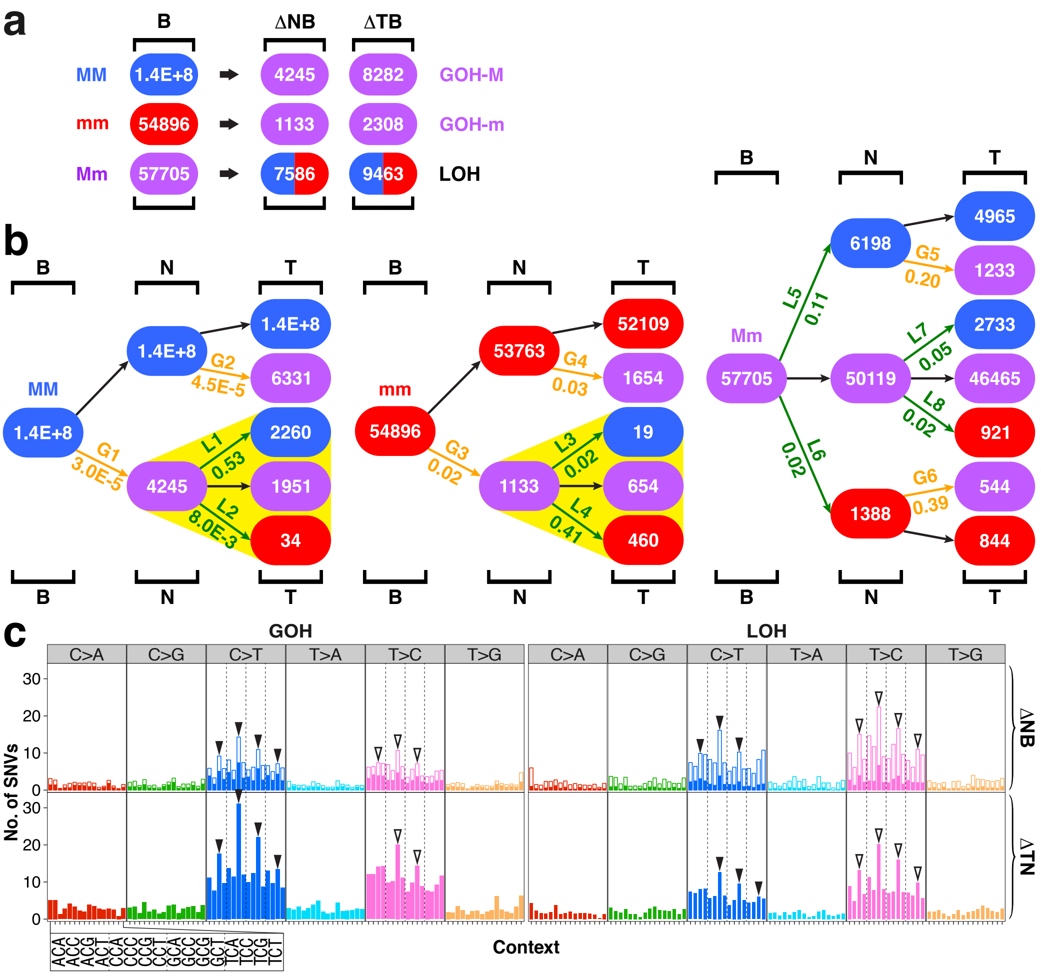
a
B
ΔNB
ΔTB
MM
1.4E+8
4245
8282
GOH-M
mm
54896
1133
2308
GOH-m
Mm
57705
7586
9463
LOH
b
B
N
T
MM
1.4E+8
1.4E+8
1.4E+8
G2
4.5E-5
6331
G1
3.0E-5
4245
L1
0.53
2260
1951
L2
8.0E-3
34
mm
54896
53763
52109
G4
0.03
1654
G3
0.02
1133
L3
0.02
19
654
L4
0.41
460
Mm
57705
L5
0.11
6198
4965
G5
0.20
1233
50119
L7
0.05
2733
46465
L8
0.02
921
L6
0.02
1388
G6
0.39
544
844
c
GOH
LOH
C>A
C>G
C>T
T>A
T>C
T>G
No. of SNVs
0
10
20
30
ΔNB
ΔTN
ACA
ACC
ACG
ACT
CCA
CCC
CCG
CCT
GCA
GCC
GCG
GCT
TCA
TCC
TCG
TCT
Context

Figure 3

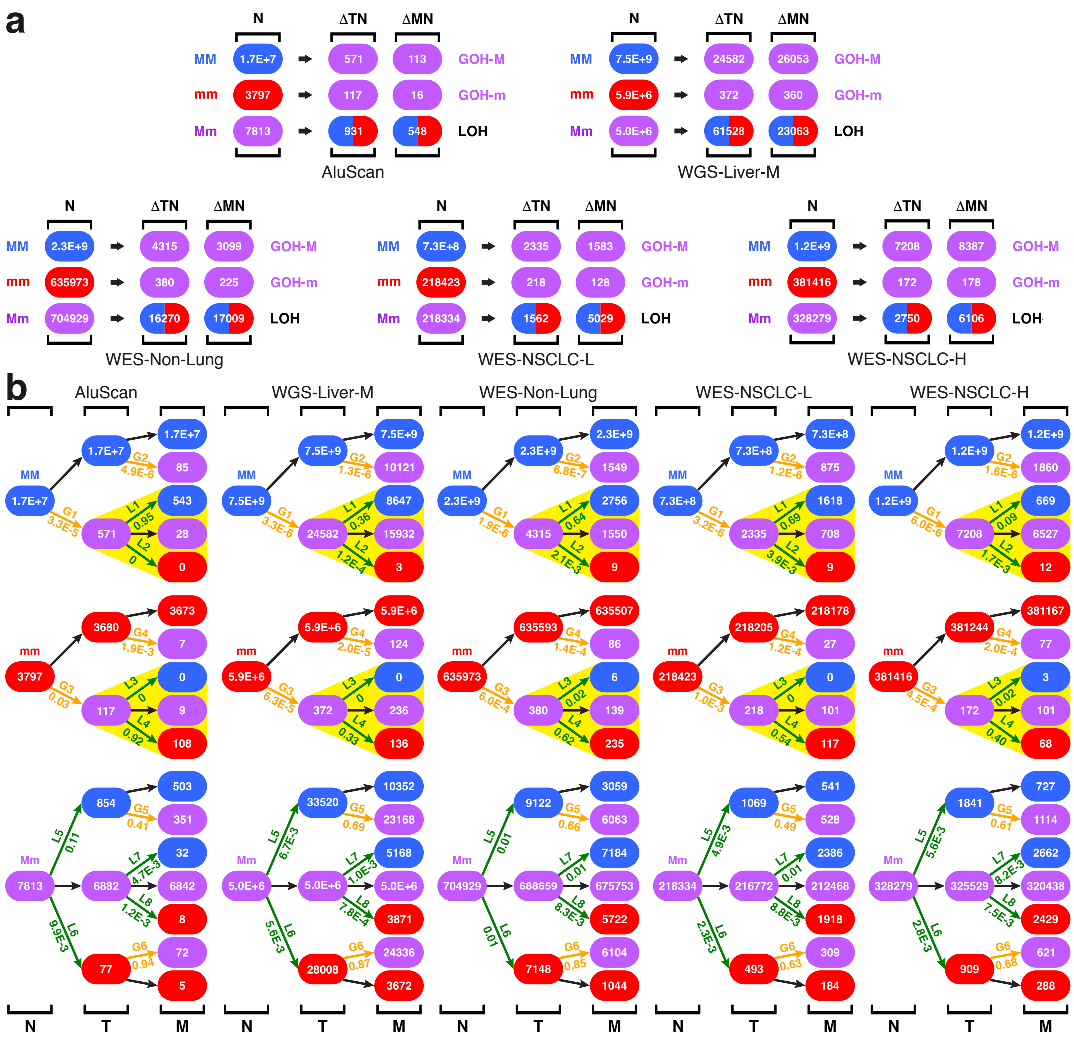
a
N
ΔTN
ΔMN
MM 1.7E+7 571 113 GOH-M
mm 3797 117 16 GOH-m
Mm 7813 931 548 LOH
AluScan
MM 7.5E+9 24582 26053 GOH-M
mm 5.9E+6 372 360 GOH-m
Mm 5.0E+6 61528 23063 LOH
WGS-Liver-M
MM 2.3E+9 4315 3099 GOH-M
mm 635973 380 225 GOH-m
Mm 704929 16270 17009 LOH
WES-Non-Lung
MM 7.3E+8 2335 1583 GOH-M
mm 218423 218 128 GOH-m
Mm 218334 1562 5029 LOH
WES-NSCLC-L
MM 1.2E+9 7208 8387 GOH-M
mm 381416 172 178 GOH-m
Mm 328279 2750 6106 LOH
WES-NSCLC-H
b
AluScan
MM 1.7E+7 1.7E+7 1.7E+7 G2 4.9E-6 85 G1 3.3E-5 571 L1 0.95 543 28 L2 0 0
mm 3797 3680 3673 G4 1.9E-3 7 G3 0.03 117 L3 0 0 9 L4 0.92 108
Mm 7813 L5 0.11 854 503 G5 0.41 351 6882 L7 4.7E-3 32 6842 L8 1.2E-3 8 L6 9.9E-3 77 G6 0.94 72 5
WGS-Liver-M
MM 7.5E+9 7.5E+9 7.5E+9 G2 1.3E-6 10121 G1 3.3E-6 24582 L1 0.36 8647 15932 L2 1.2E-4 3
mm 5.9E+6 5.9E+6 5.9E+6 G4 2.0E-5 124 G3 6.3E-5 372 L3 0 0 236 L4 0.33 136
Mm 5.0E+6 L5 6.7E-3 33520 10352 G5 0.69 23168 5.0E+6 L7 1.0E-3 5168 5.0E+6 L8 7.8E-4 3871 L6 5.6E-3 28008 G6 0.87 24336 3672
WES-Non-Lung
MM 2.3E+9 2.3E+9 2.3E+9 G2 6.8E-7 1549 G1 1.9E-6 4315 L1 0.64 2756 1550 L2 2.1E-3 9
mm 635973 635593 635507 G4 1.4E-4 86 G3 6.0E-4 380 L3 0.02 6 139 L4 0.62 235
Mm 704929 L5 0.01 9122 3059 G5 0.66 6063 688659 L7 0.01 7184 675753 L8 8.3E-3 5722 L6 0.01 7148 G6 0.85 6104 1044
WES-NSCLC-L
MM 7.3E+8 7.3E+8 7.3E+8 G2 1.2E-6 875 G1 3.2E-6 2335 L1 0.69 1618 708 L2 3.9E-3 9
mm 218423 218205 218178 G4 1.2E-4 27 G3 1.0E-3 218 L3 0 0 101 L4 0.54 117
Mm 218334 L5 4.9E-3 1069 541 G5 0.49 528 216772 L7 0.01 2386 212468 L8 8.8E-3 1918 L6 2.3E-3 493 G6 0.63 309 184
WES-NSCLC-H
MM 1.2E+9 1.2E+9 1.2E+9 G2 1.6E-6 1860 G1 6.0E-6 7208 L1 0.09 669 6527 L2 1.7E-3 12
mm 381416 381244 381167 G4 2.0E-4 77 G3 4.5E-4 172 L3 0.02 3 101 L4 0.40 68
Mm 328279 L5 5.6E-3 1841 727 G5 0.61 1114 325529 L7 8.2E-3 2662 320438 L8 7.5E-3 2429 L6 2.8E-3 909 G6 0.68 621 288
N T M

Figure 4

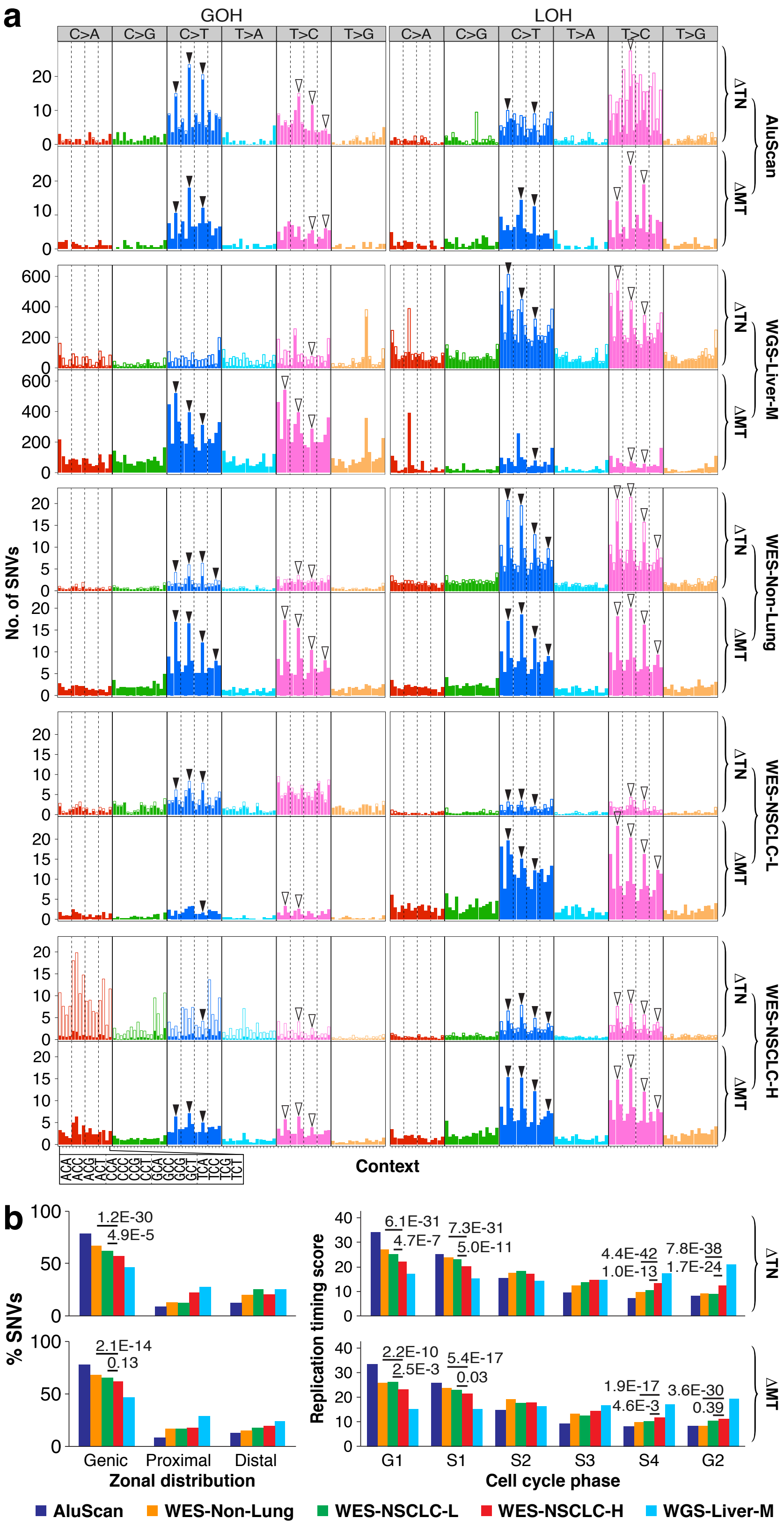
a
GOH
LOH
C>A
C>G
C>T
T>A
T>C
T>G
ΔTN
ΔMT
AluScan
WGS-Liver-M
WES-Non-Lung
WES-NSCLC-L
WES-NSCLC-H
No. of SNVs
0
5
10
15
20
200
400
600
ACA
ACC
ACG
ACT
CCA
CCC
CCG
CCT
GCA
GCC
GCG
GCT
TCA
TCC
TCG
TCT
Context
b
% SNVs
0
50
100
1.2E-30
4.9E-5
2.1E-14
0.13
Genic
Proximal
Distal
Zonal distribution
Replication timing score
0
10
20
30
40
6.1E-31
4.7E-7
7.3E-31
5.0E-11
4.4E-42
1.0E-13
7.8E-38
1.7E-24
2.2E-10
2.5E-3
5.4E-17
0.03
1.9E-17
4.6E-3
3.6E-30
0.39
G1
S1
S2
S3
S4
G2
Cell cycle phase
AluScan
WES-Non-Lung
WES-NSCLC-L
WES-NSCLC-H
WGS-Liver-M

Figure 5

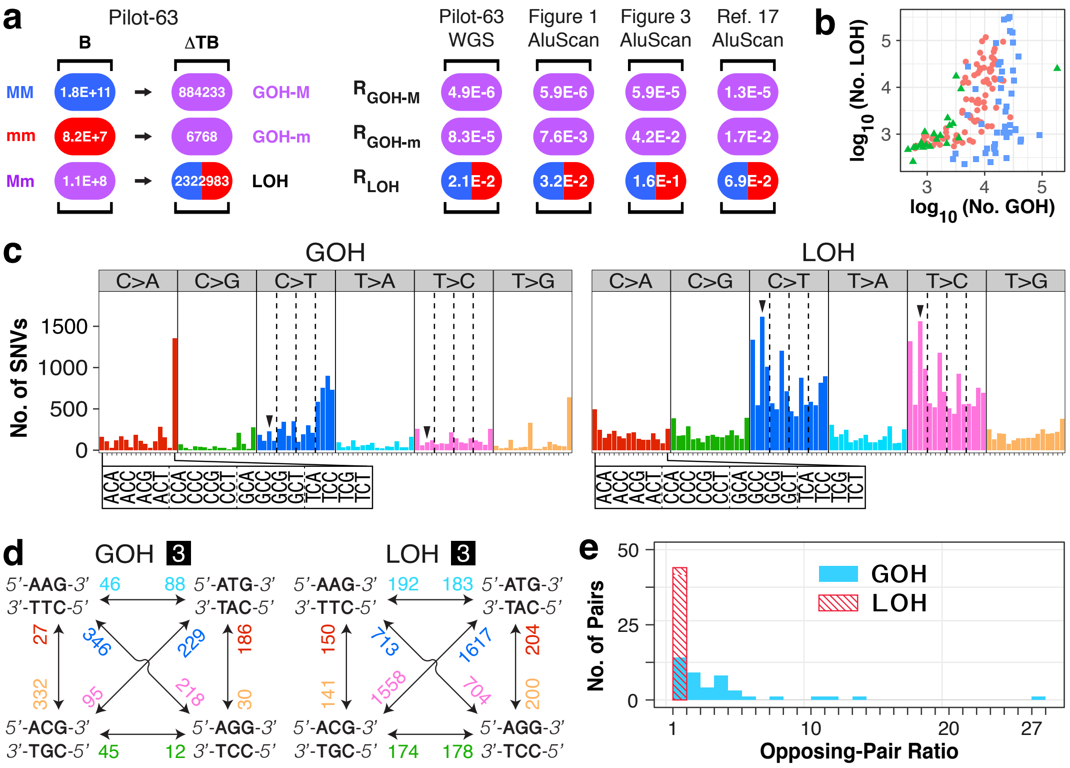
a
Pilot-63
B
ΔTB
MM
1.8E+11
884233
GOH-M
mm
8.2E+7
6768
GOH-m
Mm
1.1E+8
2322983
LOH
Pilot-63 WGS
Figure 1 AluScan
Figure 3 AluScan
Ref. 17 AluScan
R_GOH-M 4.9E-6 5.9E-6 5.9E-5 1.3E-5
R_GOH-m 8.3E-5 7.6E-3 4.2E-2 1.7E-2
R_LOH 2.1E-2 3.2E-2 1.6E-1 6.9E-2
b
log10 (No. LOH)
log10 (No. GOH)
3
4
5
c
GOH
LOH
C>A
C>G
C>T
T>A
T>C
T>G
No. of SNVs
0
500
1000
1500
ACA
ACC
ACG
ACT
CCA
CCC
CCG
CCT
GCA
GCC
GCG
GCT
TCA
TCC
TCG
TCT
d
GOH 3
LOH 3
5′-AAG-3′ 3′-TTC-5′
5′-ATG-3′ 3′-TAC-5′
5′-ACG-3′ 3′-TGC-5′
5′-AGG-3′ 3′-TCC-5′
46
88
27
346
229
186
332
95
218
30
45
12
192
183
150
713
1617
204
141
1558
704
200
174
178
e
No. of Pairs
0
25
50
GOH
LOH
1
10
20
27
Opposing-Pair Ratio

Figure 6

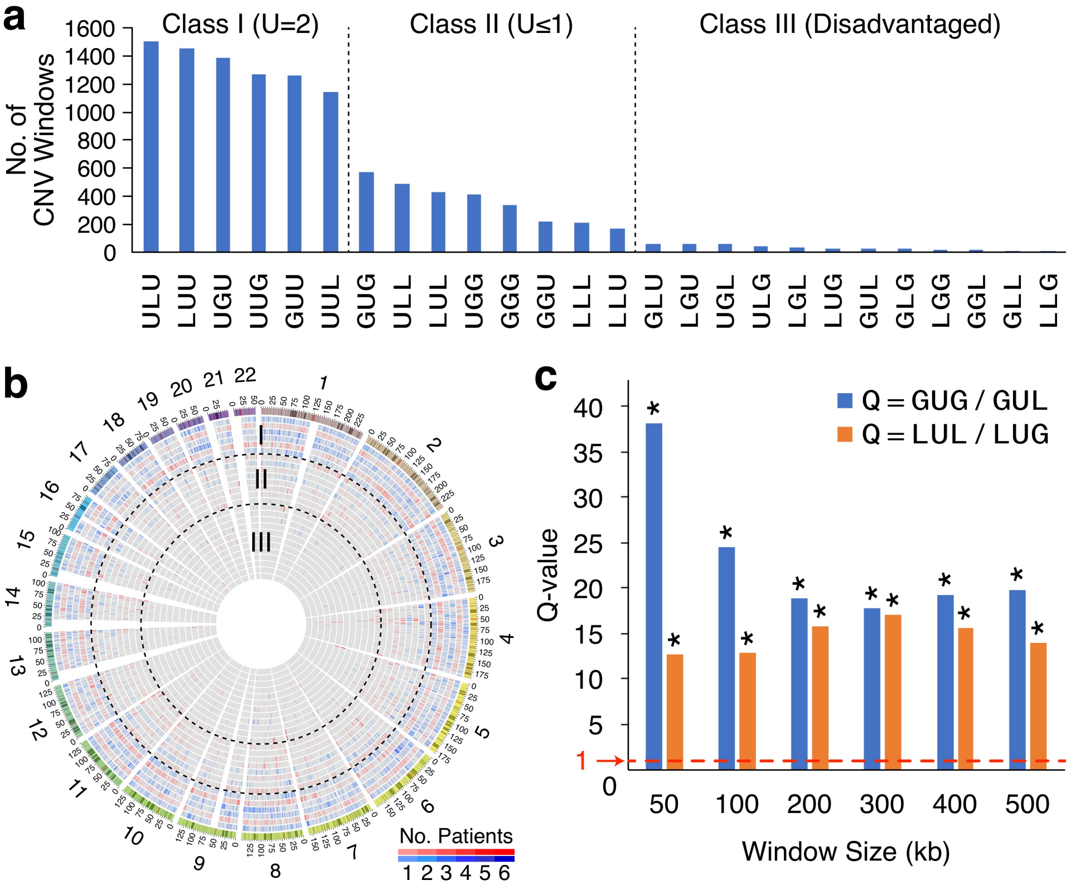
a
Class I (U=2)
Class II (U≤1)
Class III (Disadvantaged)
No. of CNV Windows
1600
1400
1200
1000
800
600
400
200
0
ULU
LUU
UGU
UUG
GUU
UUL
GUG
ULL
LUL
UGG
GGG
GGU
LLL
LLU
GLU
LGU
UGL
ULG
LGL
LUG
GUL
GLG
LGG
GGL
GLL
LLG
b
1
2
3
4
5
6
7
8
9
10
11
12
13
14
15
16
17
18
19
20
21
22
I
II
III
No. Patients
1 2 3 4 5 6
c
Q = GUG / GUL
Q = LUL / LUG
Q-value
40
35
30
25
20
15
10
5
1 →
0
*
50
100
200
300
400
500
Window Size (kb)

Figure 7

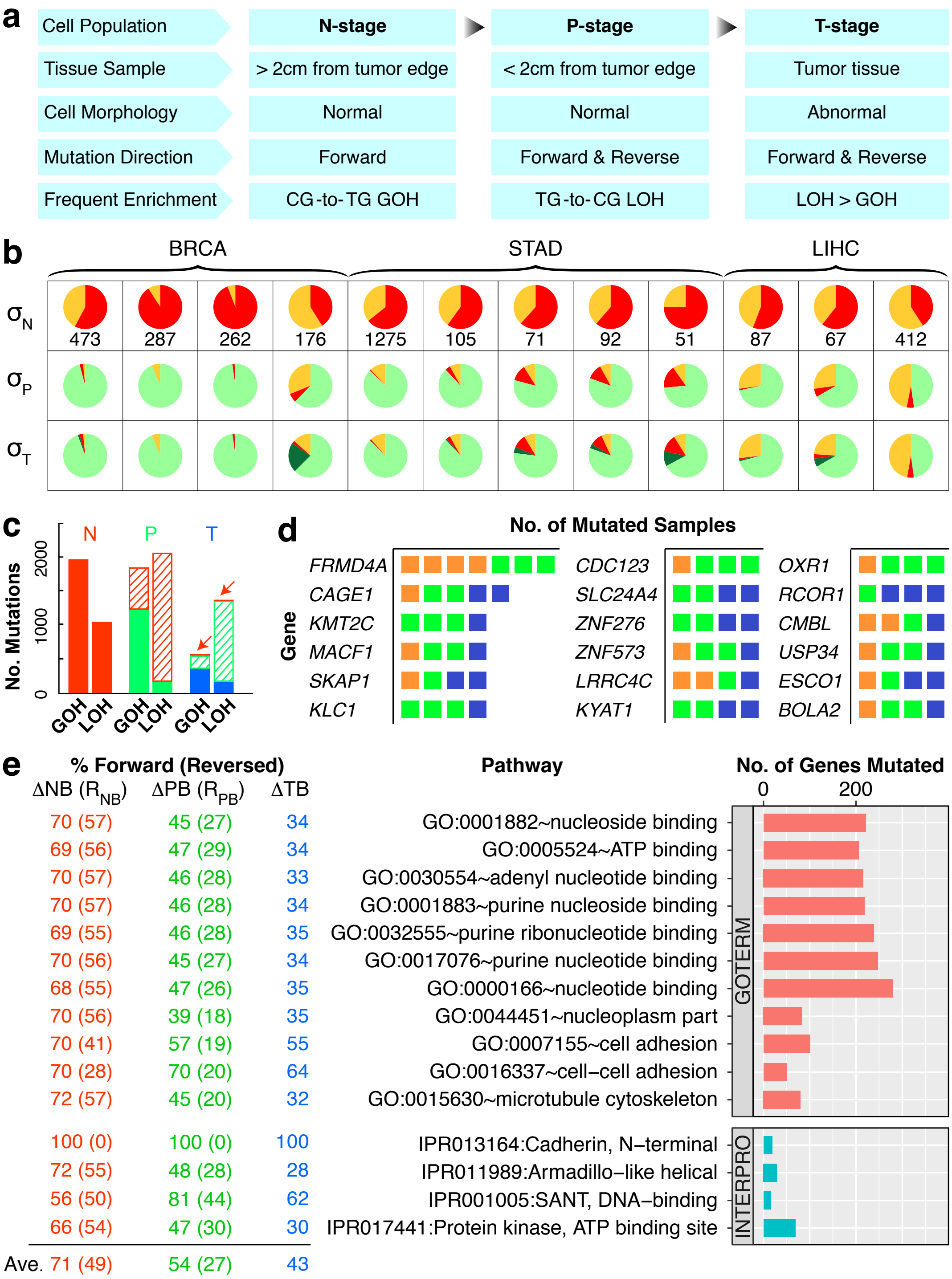
a
Cell Population | N-stage | P-stage | T-stage
Tissue Sample | > 2cm from tumor edge | < 2cm from tumor edge | Tumor tissue
Cell Morphology | Normal | Normal | Abnormal
Mutation Direction | Forward | Forward & Reverse | Forward & Reverse
Frequent Enrichment | CG-to-TG GOH | TG-to-CG LOH | LOH > GOH
b
BRCA
STAD
LIHC
σN
σP
σT
473
287
262
176
1275
105
71
92
51
87
67
412
c
No. Mutations
N
P
T
0
1000
2000
GOH
LOH
GOH
LOH
GOH
LOH
d
No. of Mutated Samples
Gene
FRMD4A
CAGE1
KMT2C
MACF1
SKAP1
KLC1
CDC123
SLC24A4
ZNF276
ZNF573
LRRC4C
KYAT1
OXR1
RCOR1
CMBL
USP34
ESCO1
BOLA2
e
% Forward (Reversed)
ΔNB (RNB) ΔPB (RPB) ΔTB
Pathway
No. of Genes Mutated
0
200
70 (57) 45 (27) 34 GO:0001882~nucleoside binding
69 (56) 47 (29) 34 GO:0005524~ATP binding
70 (57) 46 (28) 33 GO:0030554~adenyl nucleotide binding
70 (57) 46 (28) 34 GO:0001883~purine nucleoside binding
69 (55) 46 (28) 35 GO:0032555~purine ribonucleotide binding
70 (56) 45 (27) 34 GO:0017076~purine nucleotide binding
68 (55) 47 (26) 35 GO:0000166~nucleotide binding
70 (56) 39 (18) 35 GO:0044451~nucleoplasm part
70 (41) 57 (19) 55 GO:0007155~cell adhesion
70 (28) 70 (20) 64 GO:0016337~cell−cell adhesion
72 (57) 45 (20) 32 GO:0015630~microtubule cytoskeleton
GOTERM
100 (0) 100 (0) 100 IPR013164:Cadherin, N−terminal
72 (55) 48 (28) 28 IPR011989:Armadillo−like helical
56 (50) 81 (44) 62 IPR001005:SANT, DNA−binding
66 (54) 47 (30) 30 IPR017441:Protein kinase, ATP binding site
INTERPRO
Ave. 71 (49) 54 (27) 43

Figure 8