Genomics Proteomics Bioinformatics 20 (2022) 177–191

Genomics Proteomics Bioinformatics

www.elsevier.com/locate/gpb
www.sciencedirect.com

ORIGINAL RESEARCH

Common Postzygotic Mutational Signatures in Healthy Adult Tissues Related to Embryonic Hypoxia

Yaqiang Hong [1,3,#], Dake Zhang [1,2,#], Xiangtian Zhou [4,5,#], Aili Chen [1,#], Amir Abliz [6,#], Jian Bai [1], Liang Wang [7,8], Qingtao Hu [1], Kenan Gong [1], Xiaonan Guan [1], Mengfei Liu [6], Xinchang Zheng [1,13], Shujuan Lai [1], Hongzhu Qu [9], Fuxin Zhao [4,5], Shuang Hao [1], Zhen Wu [7,8], Hong Cai [6], Shaoyan Hu [10], Yue Ma [11], Junting Zhang [7,8], Yang Ke [6], Qian-Fei Wang [1,13], Wei Chen [1,2,*], Changqing Zeng [1,12,13,*]

[1] *CAS Key Laboratory of Genomic and Precision Medicine, Beijing Institute of Genomics, Chinese Academy of Sciences and China National Center for Bioinformation, Beijing 100101, China*
[2] *Beijing Advanced Innovation Centre for Biomedical Engineering, Key Laboratory for Biomechanics and Mechanobiology of Ministry of Education, School of Biological Science and Medical Engineering, Beihang University, Beijing 100191, China*
[3] *Tsinghua-Peking Center for Life Sciences, School of Life Sciences, Tsinghua University, Beijing 100084, China*
[4] *School of Optometry and Ophthalmology and Eye Hospital, Wenzhou Medical University, Wenzhou 325035, China*
[5] *The State Key Laboratory of Optometry, Ophthalmology and Vision Science, Wenzhou 325035, China*
[6] *Key Laboratory of Carcinogenesis and Translational Research (MOE), Laboratory of Genetics, Peking University Cancer Hospital & Institute, Beijing 100142, China*
[7] *Skull Base and Brainstem Tumor Division, Department of Neurosurgery, Beijing Tian Tan Hospital, Capital Medical University, Beijing 100050, China*
[8] *China National Clinical Research Center for Neurological Diseases, Beijing 100050, China*
[9] *CAS Key Laboratory of Genome Sciences and Information, Beijing Institute of Genomics, Chinese Academy of Sciences and China National Center for Bioinformation, Beijing 100101, China*
[10] *Department of Hematology and Oncology, Children's Hospital of Soochow University, Suzhou 215025, China*
[11] *Institute of Biophysics, Chinese Academy of Sciences, Beijing 100101, China*
[12] *Collaborative Innovation Center for Genetics and Development, Shanghai 200438, China*
[13] *University of Chinese Academy of Sciences, Beijing 100049, China*

Received 23 June 2020; revised 31 August 2021; accepted 6 September 2021
Available online 5 October 2021

Handled by Leng Han

* Corresponding authors.
E-mail: chenw123@buaa.edu.cn (Chen W), czeng@big.ac.cn (Zeng C).
# Equal contribution.

Peer review under responsibility of Beijing Institute of Genomics, Chinese Academy of Sciences / China National Center for Bioinformation and Genetics Society of China.

https://doi.org/10.1016/j.gpb.2021.09.005

KEYWORDS

Postzygotic mutation;
Mutational signature;
Healthy individual;
Embryonic development;
Hypoxia

Abstract **Postzygotic mutations** are acquired in normal tissues throughout an individual's lifetime and hold clues for identifying mutagenic factors. Here, we investigated postzygotic mutation spectra of **healthy individuals** using optimized ultra-deep exome sequencing of the time-series samples from the same volunteer as well as the samples from different individuals. In blood, sperm, and muscle cells, we resolved three common types of **mutational signatures**. Signatures A and B represent clock-like mutational processes, and the polymorphisms of epigenetic regulation genes influence the proportion of signature B in mutation profiles. Notably, signature C, characterized by C > T transitions at GpCpN sites, tends to be a feature of diverse normal tissues. Mutations of this type are likely to occur early during **embryonic development**, supported by their relatively high allelic frequencies, presence in multiple tissues, and decrease in occurrence with age. Almost none of the public datasets for tumors feature this signature, except for 19.6% of samples of clear cell renal cell carcinoma with increased activation of the **hypoxia**-inducible factor 1 (HIF-1) signaling pathway. Moreover, the accumulation of signature C in the mutation profile was accelerated in a human embryonic stem cell line with drug-induced activation of HIF-1α. Thus, embryonic hypoxia may explain this novel signature across multiple normal tissues. Our study suggests that hypoxic condition in an early stage of embryonic development is a crucial factor inducing C > T transitions at GpCpN sites; and individuals' genetic background may also influence their postzygotic mutation profiles.

Introduction

After fertilization, most genomic mutations typically occur due to replication errors, DNA structural instabilities, or other endogenous and exogenous sources, resulting in genotypic and phenotypic heterogeneity among all types of cells in the body [1–3]. Mutations can be triggered by various environmental factors, producing characteristic patterns. The accumulation of somatic mutations results from chronic exposure to toxicity, regeneration, and clonal structures through the transition from health to disease [4–6]. Thus, the roles of somatic mutations have been widely explored in pathogenesis [7,8]. Moreover, in recent years, multiple cell clones have been found to have distinct genotypes, referred to as somatic mosaicism, resulting from lineage expansion in healthy tissues; these have drawn attention to the factors underlying certain disorders [9,10].

Tissue-specific processes and particular microenvironmental changes leave unique imprints in genomes [9,11]. With the advent of next-generation sequencing (NGS), the characteristics of multiple mutagenic processes have been revealed for the first time in tumors of various origins [7,11–13]. For instance, smoking results mainly in C > A transitions in lung cancers [14], while ultraviolet radiation leaves a footprint involving CC > TT dinucleotide substitutions in skin cancers [7]. A recent investigation has reported the mutation spectra of cultured adult stem cells (ASCs) derived from the liver that differ from those originating from the colon and small intestine [9]. Moreover, mutation spectra are influenced by the individual's genetic background. For example, breast cancer patients carrying the *BRCA1* or *BRCA2* germline mutations exhibit a specific mutational signature in their tumor genomes compared to patients carrying the *BRCA* wild types [11]. The confounding of different mutagenesis-related factors by the genetic background means that mutation accumulation patterns differ among tissues and individuals.

Two mutational signatures (signatures 1 and 5 in the Catalogue Of Somatic Mutations In Cancer; COSMIC) related to the deamination of methylated cytosines have been shown to accumulate with age in a broad range of cell types [2]. However, this does not seem to unfold at a steady pace. Specifically, the mutation rate per cell division varies during development [1]. *De novo* mutations in offsprings increase with the paternal age, and the accumulation rate in the gonads is estimated to be about two mutations per year [15]. More than two-fold variation differences have been observed between families, possibly influenced by germline methylation [1]. Hence, factors that influence mutagenic processes may differ due to various developmental demands, such as the activities of stem cells in tissue repair, exposure to environmental factors, and tissue-specific functions [7,9]. In addition, changes in the mutation profile of cultured cells also reflect the genetic drift that occurs during the clonal expansion of cell populations carrying multiple pre-existing mutations [16].

Most knowledge of somatic mutation has been obtained from the genomic analyses of cancer or noncancer diseases, animal models, and cultured cells. However, despite the importance of analyzing the appearance and subsequent effects of somatic mutations in normal tissues, studies of their mutation profiles remain limited. The hindrance is attributable to the difficulties in obtaining appropriate tissues from healthy individuals and the scarcity of cells carrying mutations [17–19]. Although great progress has been made in analyzing the somatic mutation profiles of various tissues, including the skin, liver, esophagus, and colon, our knowledge of the mutation spectrum and its dynamic nature in healthy individuals remains inadequate [6,10,20,21].

To obtain the somatic mutation spectra of healthy individuals in this study, we first conducted optimized ultra-deep exome sequencing (∼ 800× coverage) of blood samples in five trio families. One volunteer also provided time-series samples of blood, muscle, and sperm. By comparing the mutation profiles of the five trio families with another 50 samples, we found that certain single nucleotide polymorphisms (SNPs) residing in epigenetic regulators can explain the individual-specific mutation profiles in the population. We also identified a mutational signature characterized by C > T transitions at GpCpN sites that is specific to normal tissues. Somatic mutations in

cancer and an *in vitro* experiment further showed that hypoxia may be a trigger for such mutagenesis.

Results

Ultra-deep exome sequencing reveals postzygotic mutations in normal blood and sperm cells

We identified postzygotic mutations with ultra-deep exome sequencing (> 800× coverage), exhibiting increased sensitivity and accuracy due to a multiple-step optimization (Figure S1; File S1; see Materials and methods). First, over 4 years, we collected and analyzed two blood and three sperm samples from volunteer M0038 (Table S1). One *de novo* mutation was identified in all samples with the variant allele fraction (VAF) at 0.4 ± 0.02. Overall, 36 cross-tissue mutations, with VAFs of 0.002–0.434, were shared by at least one blood sample and one sperm sample (**Figure 1**A, Figure S2A; Table S2). These common mutations of relatively high VAFs may be carried by a group of cells contributing to forming multiple tissues in the early stages of embryonic development [22], and thus are likely to occur before tissue differentiation. For tissue-specific mutations, the two blood samples shared four common postzygotic mutations; the 35-year-old sperm sample shared 5 and 11 postzygotic mutations with the 34-year-old and 36-year-old

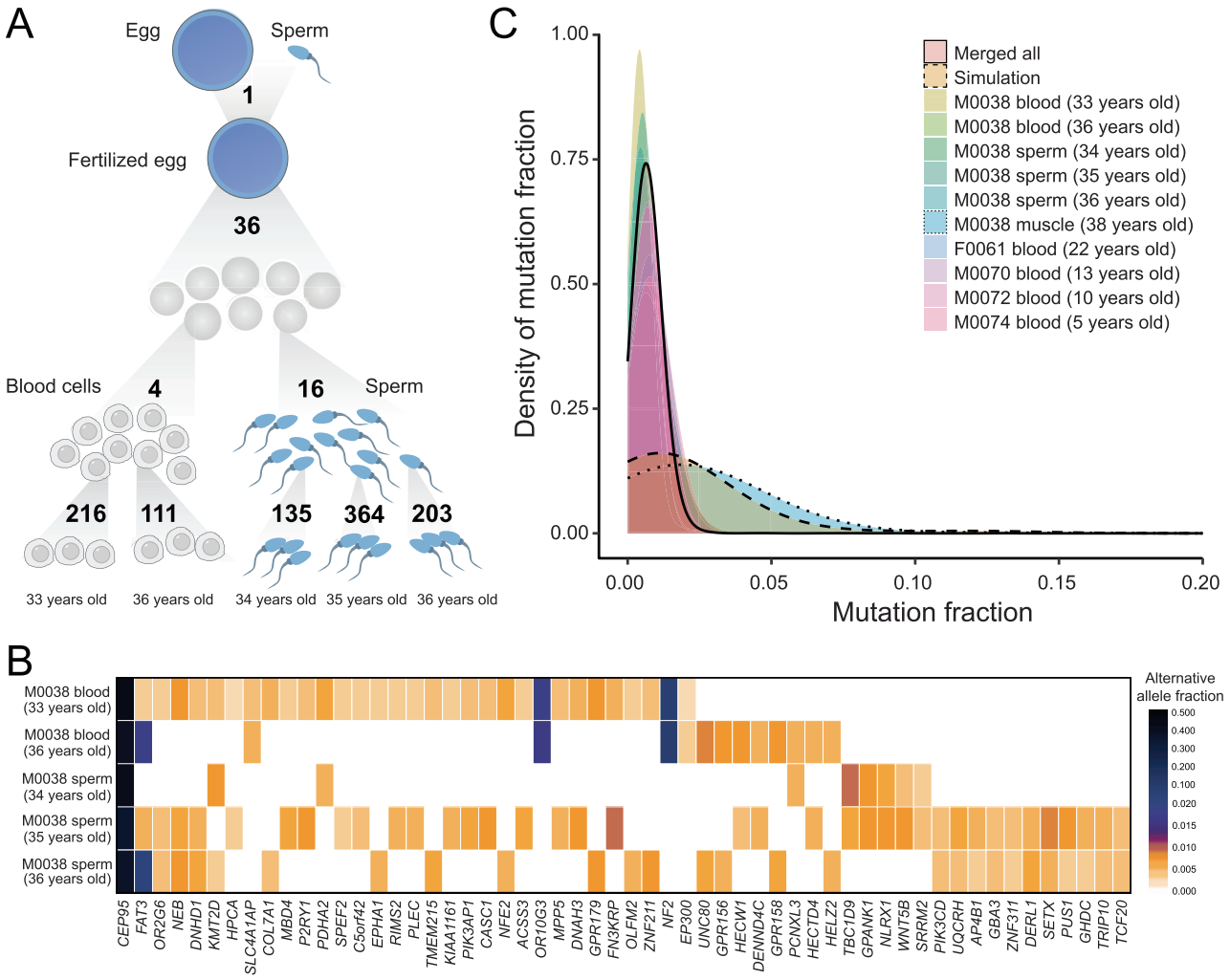

Figure 1 Postzygotic mutation profiling

A. Schematic diagram depicting the mutation accumulation among the time-series samples from individual M0038. The bottom shows the numbers of the specific mutations for each tested sample. One *de novo* mutation occurred before fertilization (top). In total, 36 postzygotic mutations were shared by at least one sperm sample and one blood sample. The 35-year-old sperm sample had 16 common postzygotic mutations with the 34-year-old (n = 5) and 36-year-old (n = 11) sperm samples, and the two blood samples had 4. **B.** Shared mutated genes in different samples of M0038. The X-axis indicates the mutated genes, and the Y-axis shows different samples. The scaled color represents the allele fraction of mutations. **C.** Density plot of the mutation fraction distribution. Data from different samples are shown in different colors.

sperm samples, respectively (Figure 1A and B). In particular, these common mutations had consistent VAFs across the samples.

We further used this approach to compare the postzygotic mutation profiles of five trio families (including M0038) (see Materials and methods; File S1). Overall, 3266 postzygotic mutations and 4 *de novo* mutations were identified (VAFs: 0.002–0.528) (Figure S3; Table S3). The validation rate for all mutations was above 85% using multiple methods (see Materials and methods; Files S1 and S2). Approximately 90% of the variants in all individuals had VAFs of less than 0.020 (Figure 1C), indicating that only a small subset of cells carried the mutations. These mutations had low recurrence rates. On average, only 2.7 (ranging between 0 and 7) mutations were shared by two individuals (Figure S2B), and none was found in more than two individuals (Table S4).

Significant enrichment of GpCpN and NpCpG postzygotic mutations in normal tissues

For each individual, we summarized the trinucleotide composition of all 96 substitution types according to the mutation and its two neighboring bases. As shown in **Figure 2**A, C > T transitions were enriched in all individuals, of which more than 90% were at GpCpN or NpCpG sites in both blood and sperm cells (Figure S4). For these two trinucleotide contexts, only individual M0038 had more GpCpN than NpCpG mutations in both types of samples collected over the 4 years, whereas the other individuals had more NpCpG mutations. This difference suggests the existence of distinct mutational processes among individuals.

C > T transitions at NpCpG sites commonly originate from age-related spontaneous deamination of methylated cytosines to thymines [2,7]. Nevertheless, the time-series samples from individual M0038 did not show the age-related increase of C > T transitions at NpCpG sites, with their proportions varying from 13% to 31% (Figure S5A and B). By contrast, the counts of C > T transitions at NpCpG sites slightly increased with age among the offsprings of the other four trio families, but this was without any significance (Figure S5C). In addition, the proportions of C > T transitions at GpCpN sites were consistent across all samples, with the highest rate being 29% and the lowest being 23% (Figure S5D).

These mutations also showed mutational strand asymmetries in the replication- and transcription-coupled DNA repair process (File S1). As shown in Figure 2B and Figure S6, during DNA replication, C > T transitions occurred more frequently in the left-replication regions of the genome, where the reference strand acts as the leading strand. More G > A transitions occurred in the right-replication regions, where the reference strand acts as the lagging strand. Mismatch repair (MMR) is more active in the lagging strand (right-replication) because of the high density of MMR signals in the lagging strand, including proliferating cell nuclear antigen (PCNA) and the 5′-ends of Okazaki fragments [23]. Therefore, mutations in the leading strand are more reflective of the mutagenesis

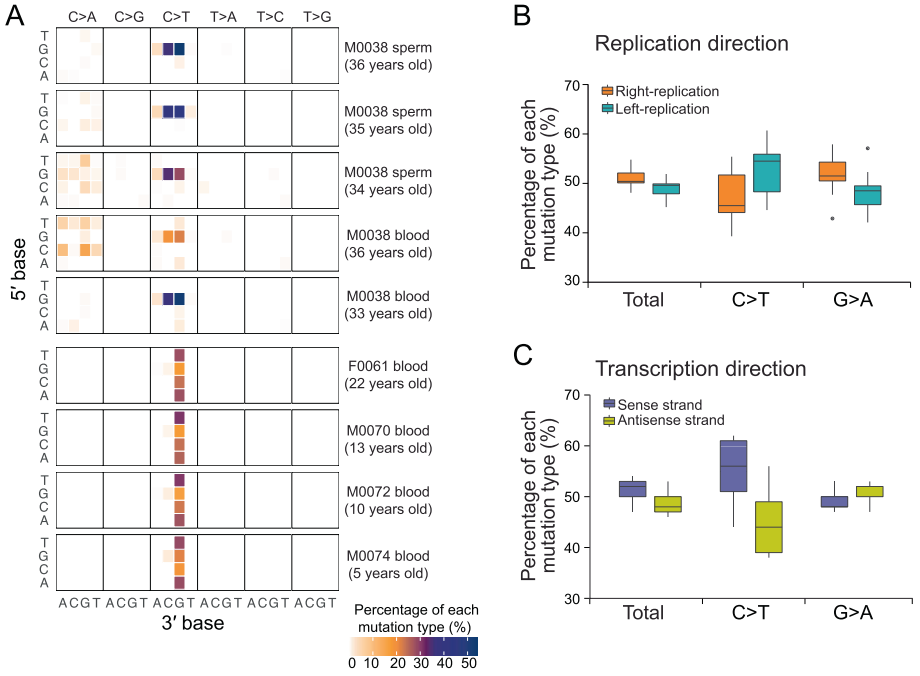

Figure 2 Patterns of postzygotic mutations in healthy individuals

A. Heatmap showing the percentages of each mutation type in healthy individuals. Each box represents a mutation type (top) with a 3′ flanking base (*X*-axis) and a 5′ flanking base (*Y*-axis). **B.** Strand asymmetry of C > T transitions according to the analyses of the replication direction. Box plot shows the percentages of total mutations (left), C > T transitions (middle), and G > A transitions (right) occurring in the right-replication regions (orange) or the left-replication regions (cyan). **C.** Strand asymmetry of C > T transitions according to the analyses of the transcription direction. Box plot shows the percentages of total mutations (left), C > T transitions (middle), and G > A transitions (right) occurring in the sense strand (purple) or the antisense strand (green).

process. During gene transcription, the genomic regions where the reference strand acts as the sense strand of mRNA exhibited a high density of C > T transitions. In contrast, the regions where the reference strand acts as the antisense strand exhibited a high density of G > A transitions (Figure 2C, Figure S7). According to the mechanism of transcription-coupled DNA repair, the RNA polymerase stalls at DNA lesions and triggers the assembly of repair complexes during transcription [24]. Hence, a mutation in the antisense strand will more likely be repaired and the mutation profile of the sense strand will better exhibit the mutagenesis. Together, mutational strand asymmetries in both replication and transcription indicate that the mutations identified by our methods are more likely to be spontaneous C > T transitions.

In addition, sperm and blood cells from M0038 exhibited consistent patterns of postzygotic mutations in the 96 mutation contexts and mutational strand asymmetries (Figure 2A, Figures S5–S7). Both tissues had higher levels of C > T transitions at GpCpN sites than at NpCpG sites, indicating the same mutational processes. In brief, the relatively constant VAFs of mutations in the time-series samples, the similar proportions of C > T transitions at GpCpN sites across samples from the same individual, and the evidence of mutational strand asymmetries were indicative of the reliability of the mutation profiles we observed.

The mutational signature with C > T at GpCpN commonly occurs in normal blood cells

To explore these mutation patterns in other normal tissues, we collected the deep exome sequencing data (> 200× coverage) of one muscle sample from individual M0038, paired normal blood cells of esophageal squamous-cell carcinoma (ESCC) and chordoma, and normal T lymphocytes and oral cells paired with acute myelocytic leukemia (AML) (**Figure 3**A). In addition, we collected and analyzed the targeted sequencing data of normal skin and single-cell sequencing data of neurons (Figure 3A). The enrichment of C > T transitions at GpCpN sites, particularly the GpCpC trinucleotides, was the most significant mutation feature in the paired normal blood cells, consistent with the observations in healthy individuals. Paired tumor samples did not show this feature in their mutation profiles. Nevertheless, normal solid tissues, both neuron and muscle, also displayed this kind of enrichment in their mutation profiles. These results strongly suggest that the C > T transitions at GpCpN sites commonly occur in normal cells.

Next, we merged the deep exome sequencing datasets of 55 normal tissue samples (see Materials and methods; File S1), and then resolved three unique mutational signatures A, B, and C using non-negative matrix factorization (NMF) (Figure 3B). Signature A is associated with the spontaneous deamination of methylated cytosines to thymines at NpCpG sites [7,11,25], and signature B is clock-like in which the number of mutations belonging to this type correlates with the age of the individual. In addition to these two known signatures, we resolved a signature C, characterized by C > T transitions at GpCpN sites, particularly the GpCpC trinucleotides. In particular, signature C is the major contributor to the somatic mutations detected in more than 30 normal samples (Figure 3C).

Epigenetic regulation may influence the proportion of mutational signatures in normal tissues

To probe for genetic factors contributing to these mutational processes, we further performed exome-wide association analyses with age as a covariate and the percentages of signatures A, B, and C as the quantitative values, respectively, in our 40 unrelated normal samples (see Materials and methods; File S1). In total, 12 SNPs located in 11 protein-coding genes were shown to correlate with the proportions of signature B ($P < 1 \times 10^{-10}$, permutation test; Figure 3D; **Table 1**), while no SNP was significantly associated with signature A or C. Among these, a missense variant in *NOTCH2* showed a significant association ($P = 4.61 \times 10^{-11}$, permutation test; Table 1). NOTCH2 is a key member of the NOTCH signaling pathway, which is important in metazoan development and tissue renewal. Its inter-cellular domain can act as a transcription factor regulating cell proliferation by controlling the expression of cyclin D1 [26,27]. Because the NOTCH signaling pathway regulates the G1/S cell cycle and signature B exhibits an age-related feature [2,28], the observed *NOTCH2* variation may affect the duration of cell proliferation, possibly explaining the proportion changes in signature B. Moreover, two associated genes, *PRDM9* and *KMT2C*, which encode zinc finger proteins that catalyze the trimethylation of histone H3 lysine 4 (H3K4me3) [29–31], were also identified. This result indicates that mutations in epigenetic regulators may influence the mutation profile of each individual in normal tissues.

Signature C may be a mutational type associated with embryonic development

Liquid tissues are polyclonal cell populations maintained by large amounts of active stem cells. Taking whole blood as an example, there may be 50,000–200,000 active stem cells of the haemato-lineage [32]. In our analyses, mutations in the bulk liquid tissues were supported by at least three unique DNA molecules (File S1). In other words, theoretically, the mutations should be shared by at least three stem cells. According to a previously constructed phylogenetic tree of single hematopoietic stem cells, these kinds of mutants are most probably generated during an early stage of haemato- or spermato-lineage development [32]. To determine the timing of mutant occurrence during development, we analyzed the somatic mutations in single B cells of newborns and older adults reported by Zhang and colleagues [33]. Using non-guided NMF, signatures A, B, and C were also resolved in the B cells of both newborns and older adults, indicating an underlying common mutation process during hematogenesis (Figure 3C). As expected, the age-related signature B was observed ($P = 0.0129$, Pearson's correlation). In particular, the signature C contributed to 20% of the mutations in the B cells of newborns and decreased significantly with aging to less than 5% in older adults ($P = 0.0076$, Pearson's correlation), indicating an age-related developmental pattern (Figure 3A and C).

Although a muscle biopsy is also a polyclonal sample without microanatomical structures, the amount of active stem cells in the sampling area is much less than in liquid tissues [34]. This was proven by the significantly higher VAFs in the muscle biopsy than those in the blood and sperm samples

(Figure 1C; $P < 2.2 \times 10^{-16}$, Kullback–Leibler divergence test). In the muscle biopsy, mutations generated during embryonic development (with relatively higher VAFs) were mixed with the newly occurred ones (with relatively lower VAFs). The mutations in the muscle biopsy were binned into high- and low-VAF groups with a cutoff of VAF = 0.025 to test the differences in mutational signature loads. As expected, the high-VAF group had a significantly higher proportion of signature C (0.185 *vs.* 0.06, Figure S8). This observation further supported the association between this mutation type and embryonic development.

Contribution of hypoxia to the occurrence of signature C

We obtained sequencing datasets from The Cancer Genome Atlas (TCGA; see Materials and methods), including 3827 samples from 19 types of tumors (Figure S9) [2,11,35]. For the somatic mutation profiles, only a small proportion of TCGA tumor samples (4%, 153/3827) exhibited high numbers of C > T transitions in the GpCpC context, of which 58 were clear cell renal cell carcinoma (CCRCC; Table S5). Notably, these 58 CCRCC samples (the high-GpCpC group) were distinguished from the remaining 273 CCRCC samples (the

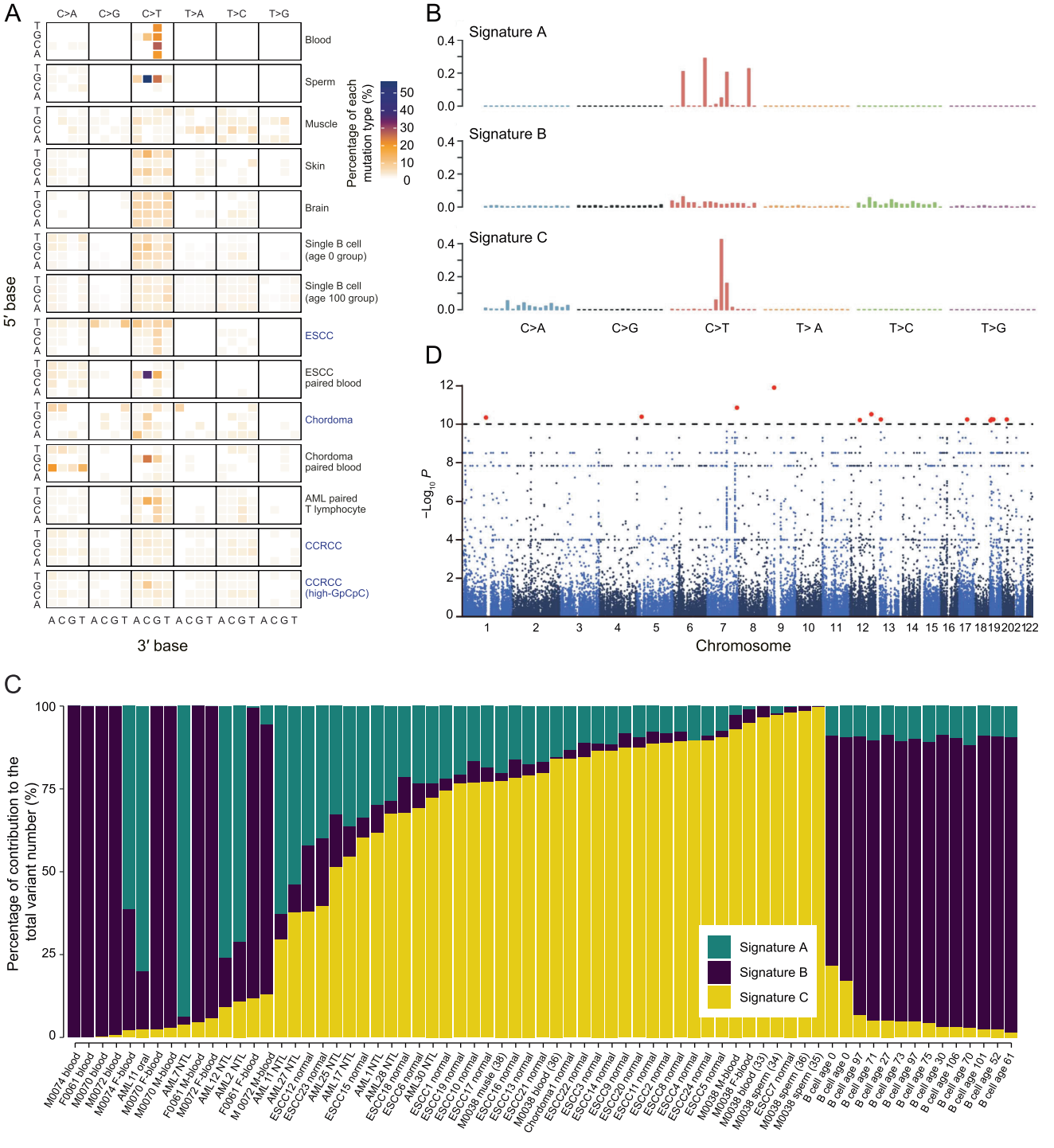

Table 1 Genes associated with the proportions of signature B

Chromosome	Position (GRCh37)	Effective allele	Major allele	Frequency	Frequency (East Asian)	HWE	Beta	*P* value	Adjust *P* value	Gene	Functional annotation
1	120,539,668	A	T	0.268	0.29*	0.02	0.41	4.61×10^{-11}	6.55×10^{-8}	*NOTCH2*	p.Thr196Ser (possibly_damaging)
5	23,527,777	T	C	0.109	0.06#	1	0.64	4.13×10^{-11}	1.26×10^{-7}	*PRDM9*	p.Tyr860=
7	151,962,309	A	G	0.329	0.08*	0.002	0.43	1.40×10^{-11}	6.55×10^{-8}	*KMT2C*	Intron variant
9	33,798,543	G	A	0.110	NA	1	0.67	1.28×10^{-12}	1.77×10^{-8}	*PRSS3*	p.Lys229Glu (benign)
12	53,865,349	G	T	0.110	0.17*	1	0.63	6.21×10^{-11}	1.40×10^{-7}	*PCBP2*	Intron variant
12	111,885,367	G	T	0.110	NA	1	0.64	3.09×10^{-11}	9.90×10^{-8}	*SH2B3*	Intron variant
13	25,671,429	T	G	0.195	0.35#	0.31	1.11	5.79×10^{-11}	1.36×10^{-7}	*PABPC3*	p.Val365Leu (benign)
17	44,850,996	C	A	0.195	0.16#	0.31	1.11	5.79×10^{-11}	1.36×10^{-7}	*WNT3*	Intron variant
19	3,586,698	G	C	0.110	0.0001#	1	0.63	6.51×10^{-11}	1.41×10^{-7}	*GIPC3*	Intron variant
19	9,012,789	T	C	0.110	NA	1	0.63	5.66×10^{-11}	1.36×10^{-7}	*MUC16*	p.Arg12885=
19	17,734,390	G	A	0.195	0#	0.31	1.11	5.79×10^{-11}	1.36×10^{-7}	*UNC13A*	Intron variant
20	26,094,525	C	T	0.195	0.08*	0.31	1.11	5.79×10^{-11}	1.36×10^{-7}	*NCOR1P1*	Noncoding Exon

Note: #, allele frequency aggregator; *, allele frequency from GnomAD; NA, not available; HWE, Hardy–Weinberger equilibrium.

low-GpCpC group) in the correlation analysis of mutation patterns (**Figure 4**A). The similarity of mutational profiles between the high-GpCpC group of CCRCC and the paired normal tissues of AML, chordoma, and ESCC indicates that same mutational processes are experienced by normal cells and some of CCRCC.

By comparing the expression profiles of the high-GpCpC and low-GpCpC groups of CCRCC, we resolved 145 differentially expressed genes ($P < 0.05$, chi-square test; Figure 4B, File S1; see Materials and methods). The most significant change in the high-GpCpC group was the increased transcription of protein phosphatase 1 regulatory subunit 12A (*PPP1R12A*; $P = 8.82 \times 10^{-5}$, Benjamini–Hochberg method; Figure 4C), which activates hypoxia-inducible factor-1α (HIF-1α) [36]. Moreover, the Hippo pathway, which is associated with the transcriptional response to hypoxia

Figure 3 C > T at GpCpN sites is a common feature of normal tissues and genetic factors may influence the mutational signature load in each individual

A. Heatmap of mutation proportions illustrates the enrichment of C > T at GpCpN in all types of normal cells (black) in both healthy subjects (upper) and patients (lower). Each box represents a mutation type (top) with a 3′ flanking base (*X*-axis) and a 5′ flanking base (*Y*-axis). The scaled color represents the percentage of each mutation type in total mutations. The mutation data presented from top to bottom are derived from the following sources: deep exome sequencing data of six blood samples from five healthy individuals in this study; deep exome sequencing data of three sperm samples from individual M0038; deep exome sequencing data of one muscle sample from individual M0038 (> 200× coverage; VAF = 0.021 ± 0.015); targeted sequencing data of 74 genes in 234 skin samples from four individuals (> 500× coverage; VAF = 0.042 ± 0.048) as reported by Martincorena et al. [5]; single-cell sequencing data of 36 single neurons from three individuals as reported by Lodato et al. [51]; single B cell sequencing data of individuals in "age 0" group (two sapmles at age 0) and "age 100" group (three samples at age 97, 101, and 106) as reported by Zhang et al. [33]; in-house exome sequencing data of 24 ESCC tumors and paired normal blood samples (both > 200× coverage; VAF = 0.175 ± 0.136 in tumors, VAF = 0.017 ± 0.009 in paired normal blood samples); in-house exome sequencing data of two chordoma tumors and paired normal blood samples (both > 250× coverage; VAF = 0.295 ± 0.200 in tumors, VAF = 0.017 ± 0.005 in paired normal blood samples); in-house exome sequencing data of 10 NTL samples and 1 normal oral sample that were paired with AML cells (> 200× coverage; VAF = 0.019 ± 0.008); exome sequencing data of 295 CCRCC samples from TCGA; and exome sequencing data of the 58 CCRCC samples with a high rate of C > T transitions at GpCpC sites among the 295 samples (termed the high-GpCpC group). **B.** Mutational signatures revealed by NMF in 55 normal samples. The 55 normal samples include 15 blood samples from five trio families, 3 sperm samples from individual M0038, 24 paired normal blood smaples for ESCC, 10 NTL samples and 1 normal oral sample paired with AML, and 1 paired normal blood sample for chordoma (the other one is excluded due to over 50% C > A mutations). The *X*-axis indicates the 96 trinucleotide mutation types, and the *Y*-axis shows the percentage of each kind of variant. Signature A involves the spontaneous deamination of methylated cytosines to thymines at NpCpG sites, signature B features the C > T and T > C transitions, and signature C features a mutational type characterized by C > T transitions at GpCpN sites. **C.** Varying proportions of the three signatures in 55 normal bulk samples and 14 single B cells. Each bar represents one sample in our analyses. The length of the color bar represents the percentage contribution of each signature among all variants. **D.** Manhattan plot of the whole-exome association analyses of germline variation with the proportion of signature B. The dashed line indicates the genome-wide significant loci with a threshold $P = 1 \times 10^{-10}$. SNPs that pass the threshold are labeled in red. VAF, variant allele fraction; ESCC, esophageal squamous-cell carcinoma; NTL, normal T lymphocyte; AML, acute myelocytic leukemia; CCRCC, clear cell renal cell carcinoma; TCGA, The Cancer Genome Atlas; NMF, non-negative matrix factorization; SNP, single nucleotide polymorphism; F/M-blood, blood sample from father or mother in a trio family.

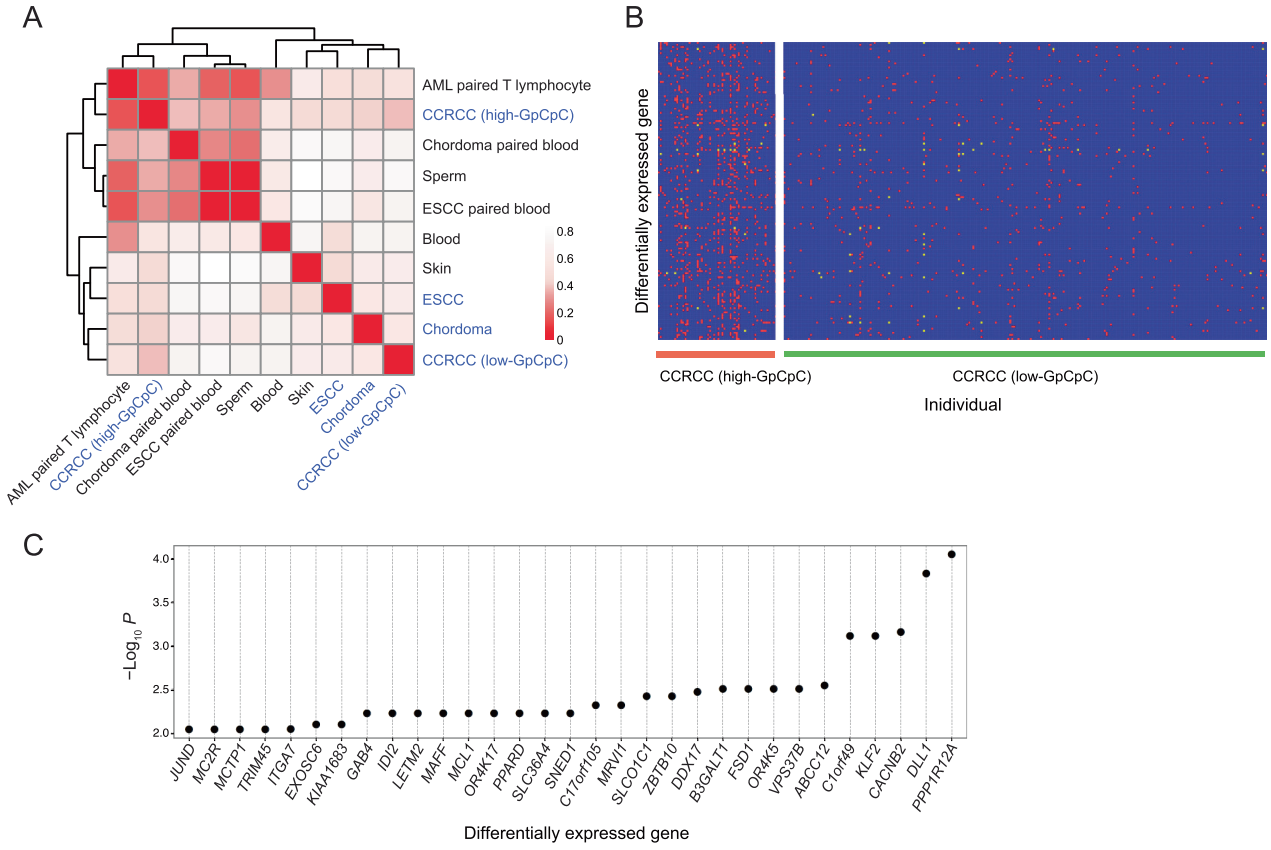

Figure 4 C > T at GpCpN is a common feature in parts of CCRCC samples with increased activation of the HIF-1 signaling pathway
A. Hierarchical clustering of the correlation matrix of normal (black font) and tumor (blue font) mutation profiles. Each box represents the correlation coefficient (r) of the column and row samples. The scaled color represents $1 - r$. CCRCC from TCGA, and AML and ESCC from in-house data were used for comparison. **B.** Differentially expressed genes in the high-GpCpC and low-GpCpC groups of CCRCC samples compared with their adjacent normal tissues in TCGA. The X-axis represents CCRCC samples, and the Y-axis represents differentially expressed genes. Red dots indicate up-regulated genes, yellow dots indicate down-regulated genes, and blue dots indicate genes with no significant change. $P < 0.05$, chi-square test. **C.** Differentially expressed genes ($P < 0.01$) in the high-GpCpC group. HIF-1, hypoxia-inducible factor 1.

[37–39], was significantly enriched ($P = 3.21 \times 10^{-5}$, Fisher's exact test). Gene-set enrichment analysis revealed an up-regulation of hypoxia-related genes ($P = 0.034$, $Q = 0.062$, permutation test; Figure S10), which further supported the increased activation of HIF-1α in the high-GpCpC group. In addition, a slightly higher mutation rate of von Hippel-Lindau tumor suppressor (*VHL*; 0.5 *vs.* 0.43), which encodes a protein involved in the ubiquitination and degradation of hypoxia-inducible factor proteins [40], was also observed in the high-GpCpC group (Table S6). Taken together, the increased activity of the HIF-1 signaling pathway could contribute to the high proportion of C > T transitions at GpCpC sites in these CCRCC samples.

To test the role of the HIF-1 signaling pathway, we treated the human embryonic stem cell (hESC) line WA07 (WiCell Research Institute) with ML228, a direct activator of the HIF-1 signaling pathway by stabilizing and activating the nuclear translocation of HIF-1α [41] (**Figure 5**A; File S1; see Materials and methods). In the first stage, WA07 cells were divided into two groups with ~ 1000 cells each. One group was treated with ML228 (0.125 μM) for 15 days, and the other was a mock-treated control group. In the second stage, 10 cells were randomly picked up from each group and expanded to ~ 1000 cells with or without ML228 treatment. The cells were then harvested for exome sequencing with barcoding in library construction (see Materials and methods; File S1). As expected, a significantly high proportion of C > T transitions at GpCpN sites was observed in ML228-treated cells compared to the control group (0.17 *vs.* 0.07, $P = 0.0091$, chi-square test; Figure 5B and C). According to their proportions, we divided all detected mutants into high-VAF mutations (VAF > 0.05, mainly originated in the first stage) and low-VAF mutations (VAF ≤ 0.05, mainly generated from the cell expansion process in the second stage). For both types of mutation, a higher accumulation of C > T transitions at GpCpN sites was observed in the ML228-treated cells than in the control ones (0.12 *vs.* 0.06 in the high-VAF group and 0.20 *vs.* 0.08 in the low-VAF group; Figure S11). These results demonstrate that activation of the HIF-1 signaling pathway is associated with C > T transitions at GpCpN sites.

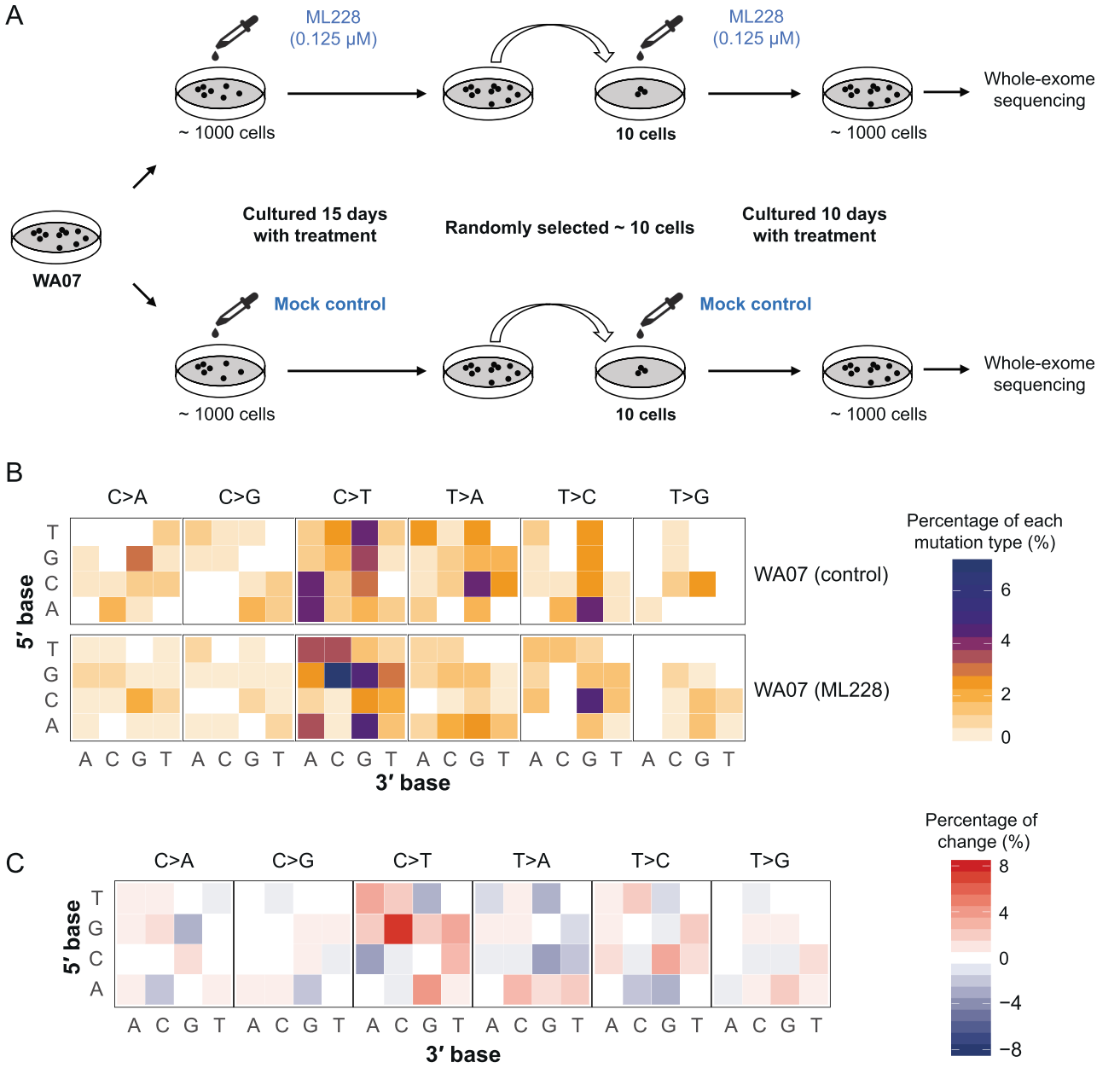

Figure 5 Activation of the HIF-1 signaling pathway leads to a high proportion of C > T transitions in the GpCpN context in hESCs
A. Two-stage treatment of WA07 cells with ML228 followed by exome sequencing with molecular barcoding. **B.** Somatic mutation profiles of WA07 cells after ML228 (lower panel) and mock control (upper panel) treatment. Each box represents a mutation type (top) with a 3′ flanking base (*X*-axis) and a 5′ flanking base (*Y*-axis). The scaled color represents the percentage of each mutation type in all mutations. **C.** The fluctuation of 96 mutation types upon the two-stage ML228 treatment compared to mock control treatment. Each box represents a mutation type (top) with a 3′ flanking base (*X*-axis) and a 5′ flanking base (*Y*-axis). The scaled color represents the percentage of fluctuation. Red indicates an increase in the ML228-treated group, and blue indicates a decrease in the ML228-treated group. hESC, human embryonic stem cell.

Discussion

In this study, using the postzygotic mutation profiles of healthy individuals from five trio families, we discovered a mutational signature characterized by C > T transitions at GpCpN trinucleotides as a major mutation type shared by blood and sperm cells. By interrogating sequencing data from normal muscle, paired normal samples in tumor studies, and single B cells, we found that this mutation pattern may be a hallmark trait of normal tissues. Interestingly, this mutation type was found in CCRCC samples with higher activation levels of the HIF-1 signaling pathway (data from public tumor datasets). The preponderance of this pathway prompted us to speculate that increases in the hypoxia status may trigger such mutations in healthy people. In *in vitro* validation, we observed accumulation of C > T transitions at GpCpN trinucleotides following HIF-1α activation in hESCs.

Patterns of low-VAF mutations in cell populations may be confounded by sequencing errors. Most sequencing artifacts are due to DNA damage during extraction and acoustic shearing [42,43]. We used several strategies to minimize declines in mutation authenticity. First, we largely repaired DNA damage before sequencing ($P < 0.05$; Figures S12 and S13; File S2). Second, we optimized the variant-calling method to get high-confidence calls of postzygotic mutations with a VAF around 0.005 (File S1). Third, to avoid mistaking inherited heterozygous variants as postzygotic mutations due to inaccurate allele fractions in NGS [44], we adopted a trio-based sequencing design (Figure S14). The validation of called variants with multiple methods ensured the robustness of our observations (Table S3). Here, we precluded the low-VAF variants in the offspring that also appeared in the parental genome with VAF > 1% (see Materials and methods). This may lead to the removal of recurrent somatic mutations. In each offspring, the number of removed variants using this criterion was 100–400, much higher than the number of either inter-individual shared mutations (2.7 mutations per individual pair, Table S4) or intra-individual shared mutations (57, Figure 1). In particular, 20%–57% of these variants are listed in dbSNP as common polymorphisms and thus are more likely to be germline mutations. Therefore, our filtering strategy might greatly improve the accuracy of somatic mutation calls, avoiding the interference caused by inherited variations.

Additionally, significant strand asymmetries, which are caused by the non-uniform MMR efficiency between different strands during replication or transcription, were observed among the identified mutations (Figure 2B and C). This feature further increases the reliability of these mutations because transcription- or replication-related MMR is unlikely to happen during the library construction and the sequencing process.

Moreover, three mutational signatures extracted from both our and public datasets further supported the reliability of our findings. Signatures A and B are known to arise from an age-related mutational process [2]; signature C, which was proposed to be promoted by hypoxic conditions in this study, poorly correlates with the known artifacts in COSMIC (Table S7). Several lines of evidence from public datasets proved that these pattern changes are not artifacts: signature C could be extracted from mutations in single B cells; the proportion of signature C in single B cells decreased with aging; signature C showed a high proportion in CCRCC with a more active HIF-1 signaling pathway. Most importantly, we successfully obtained signature C by inducing mutagenesis in hESCs (Figure 5), thus validating these findings.

Most of the mutations detected in our work were located at GC or CG sites. To assess the impact of the GC content on detection bias, we compared the sequencing coverage with the GC content. As shown in Figure S15, the GC content in the human exome was centered ~ 40%. However, the regions with a GC content of ~ 63% exhibited the highest sequencing coverage, which was 1.5 times higher than those with a GC content of 40%. This difference might lead to an underestimate of the proportion of mutations located in the non-GC regions, particularly for the low-VAF mutations. Because the major mutational signatures identified in our work were in the NpCpG and GpCpN regions, which had relatively better sequencing coverage, the GC bias issue should not influence the relative proportions of these two kinds of mutations.

We expected to find age-dependent cumulative increases in mutations. According to *in silico* simulation, there should be a significant increase in the number of mutations with VAF > 0.002 between 33 years old and 36 years old (197.26 ± 8.08 *vs.* 202.68 ± 8.36, $P = 5 \times 10^{-16}$, *t*-test). However, these increases only accounted for 2% of the total mutations, which may be masked by detection bias occurring in the experimental procedures. As described elsewhere, the sequencing of monoclonal tissues has shown that the detection bias is much larger than the accumulation process over 4 years [34]. This disadvantage also leads to the impossibility of constructing an accurate phylogeny tree for the time-series samples. Moreover, the trend of age-related accumulation could be seen among four other individuals whose ages ranged from 5 to 22 years old (Figure S6). Therefore, we concluded that enlarging the time span and increasing the sample size would help find the age-related accumulation of postzygotic mutations. Moreover, single-cell sequencing may help construct an accurate phylogeny tree based on postzygotic mutations.

To trace the mutagenesis process responsible for the generation of signature C, tumor samples from TCGA provided an interesting clue regarding how these substitutions may arise. Signature C can only be found in CCRCC, which was characterized by the activation of the HIF-1 signaling pathway [45]. Meanwhile, transcriptome analyses of the high-GpCpC group of CCRCC demonstrated that their HIF-1 signaling pathway was more active than in the low-GpCpC group (Figure S10). Moreover, a recent study of recurrent glioblastoma, which was featured by extremely hypoxic conditions in a tumor microenvironment [46], found that C > T transitions at GpCpC and GpCpT sites were enriched in its mutation spectrum (Figure S16; File S2) [47]. The induction of C > T transitions at the GpCpC sites by the HIF-1 signaling pathway was further validated in an oligoclonal culture of hESCs. By directly activating HIF-1α in oligoclonal hESCs using ML228 [41] and using a specially designed two-step cell culture experiment (Figure 5), we successfully observed the significant accumulation of C > T transitions at GpCpN sites.

Notably, after applying multiple stringent and efficient error correction procedures, a high proportion of mutations still had VAFs in the range of 0.01–0.05, indicating that they appear within 20 cell divisions after fertilization. The HIF-1 signaling pathway is crucial in oxygen sensing and mediates tissue adaptation to oxygen deficiency. This stress is critical during embryonic development [48–50]. We believe that hypoxia during embryonic development underlies the C > T transitions at GpCpN sites. In addition to our experimental validation using hESCs, C > T transitions at GpCpN sites can also be observed in normal neurons, in which the accumulation of C > T transitions at GpCpN sites is significantly higher than that of C > T transitions caused by the deamination of methylated cytosines at NpCpG sites ($P = 0.047$, *t*-test; Figure 3A, Figure S17; File S2) [51]. Because neuronal cell division stops after the neuroepithelial cells have differentiated into proper neurons, most mutations should occur during cortical neurogenesis, which is completed around 15 weeks post-conception [52].

The utilization of liquid samples of sperm and blood, and the bulk sequencing without further separation or cloning are two unique features of our sequencing strategy. These features allow us to capture those relatively common mutations occurring in each tissue in the early stages of a cell lineage.

This consideration also explains why we observed the same signatures from the mutations of muscle with high VAFs and the mutations from blood and sperm tissues. By contrast, previous studies on postzygotic mutations have mainly focused on cancer somatic mutations, organoid mutations, or *de novo* mutations [1,5,53]; most of these mutations are genomic changes specific to certain cell lineages across the entire life span. Therefore, in these mutation spectra, early events only account for a small proportion and are possibly overwhelmed by other mutations that occur later in development.

Postzygotic mutations from our sperm data showed that most of the mutations were C > T transitions, whereas the *de novo* mutations identified by previous research indicated that the proportions of C > T and T > C mutations were similar to one another [1]. We identified more C > T mutations because this mutagenesis process occurred in an early stage of the spermato-lineage. Moreover, T > C mutations may mainly originate from the GC-biased gene conversion during meiosis [54], which occurs in the late stage of gametogenesis. Therefore, in the final sperm cell population, T > C mutations may offset the effects caused by C > T mutations.

To find a link between the mutational signature C and the hypoxia conditions in hESCs, we used ML228, a direct activator of HIF-1α, instead of lowering the oxygen supply. Although physiologically lowering oxygen supply is sure to activate the HIF-1 signaling pathway, it also leads to an increase in reactive oxygen species and oxidative stress, which may cause the accumulation of oxidative DNA damage [55,56]. On the other hand, ML228, widely applied as a HIF activator, chelates iron and directly activates the HIF-1 signaling pathway without introducing the other known biological side effects [41].

One limitation of our cellular experiment is that it does not include a HIF-1α knockout/knockdown/inhibition condition to test the decrease in the proportion of signature C. It is also difficult to find a cell line with a sustainably activated HIF-1 signaling pathway. Although the CCRCC cell line Caki-1 may be a good choice, the major mutation type of this cell line is the C > T transition caused by deamination (COSMIC ID: COSS905963, https://cancer.sanger.ac.uk/cell_lines/). Therefore, this type of mutation may mask the contribution of the HIF-1 signaling pathway. The primary culture of tumor cells derived from CCRCC patients with high signature C loading may help solve this limitation, and should be performed in future studies.

In genetic association tests, we found that coding polymorphisms in *NOTCH2*, *KMT2C*, and *PRDM9* were associated with the individual load of mutational signature B. However, only three signatures were identified in our data. We cannot pinpoint whether these associations were from signature B or from signature A plus C, because signatures A and C shared a similar mutation type of C > T transitions. Moreover, the available deep sequencing datasets for the association test were only exonic. Future efforts are needed to reveal the contribution of polymorphisms in non-coding regions to the proportion of different mutational signatures. Nevertheless, these limitations did not influence our observation that polymorphisms in genes related to epigenetic regulation were associated with the individual load of specific mutational signatures. Although the load of signature C showed a significant divergence among individuals, we did not find any polymorphisms associated with it. Because this signature may be associated with the hypoxia condition during development, collecting phenotypes and environmental explorations during the fetus and infant ages of these participants will help pinpoint the origin of such divergence.

Another limitation of our work is the absence of postzygotic mutations from the human embryo to directly prove

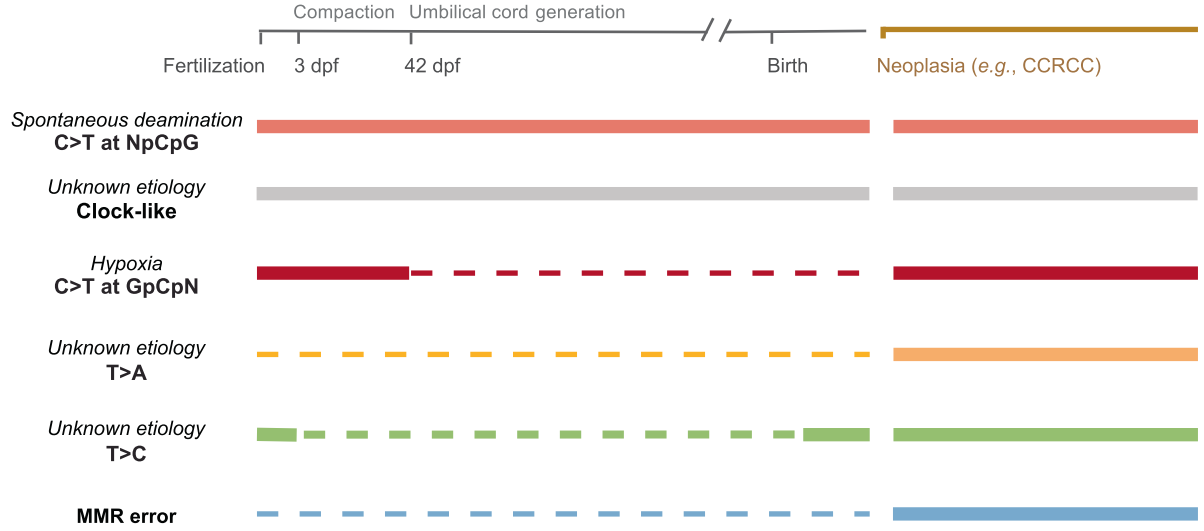

Figure 6 Proposed mutational processes during a lifespan and in CCRCC
After fertilization, the spontaneous deamination of methylated cytosines at NpCpG sites is the most common mutation type associated with age. In the early stages of embryonic development characterized by hypoxia, C > T transitions commonly occur at GpCpN sites. The enrichments of T > C transitions with unknown etiology are other mutational processes that occur during development. Regarding CCRCC development, the hypoxia-induced mutation process causes C > T transitions at GpCpN sites. Other mutation patterns in CCRCC include errors in MMR and T > A transversions via unknown mechanisms. The dashed lines represent a lack of supporting evidence in a given stage. dpf, days post-fertiliztion; MMR, mismatch repair.

the existence of signature C. Although the mutation profiles of single adult neurons provided indirect evidence (Figure S17) [51], data from embryonic tissues, particularly germ cells, which are important for understanding the feature of human germline mutations, should be investigated in future studies.

Finally, integrating somatic mutations in CCRCCs and single B cells as well as previously reported *de novo* mutations (Figure S18) with our results can help build a procedure for postzygotic mutation generation. This procedure would clarify the mutational signatures and the corresponding active mutational processes during embryonic and post-parturition development, as well as in tumor development (**Figure 6**). Across an individual's lifespan, C > T transitions at NpCpG trinucleotides due to spontaneous deamination (COSMIC signature 1 or signature A in our study) and another clock-like mutational type with unknown etiology (COSMIC signature 5 or signature B in our study) constantly occur after fertilization. During embryonic development, the hypoxic environment triggers the occurrence and accumulation of C > T transitions at GpCpN sites (signature C in our work). However, after birth, T > C transitions are generated in normal cells based on *de novo* mutations. In CCRCC, all types of mutational processes are present. In future investigations, samples of multiple normal tissues from one individual should help validate the molecular mechanism through which hypoxia induces an increased accumulation of mutations during embryonic development.

Materials and methods

Whole-exome sequencing

DNA samples were obtained from individuals aged 5–33, including four males and one female (Table S1). The details of the whole-exome sequencing analyses are summarized in File S1.

Postzygotic mutation detection

Using the error estimation model (detailed information is provided in File S1), postzygotic mutations in normal cells were detected by following several steps. Sequencing reads were aligned to the human reference genome build GRCh37 using the BWA algorithm [57] after removing the adapter segments and excluding the reads with low Q-scores (File S1). Uniquely mapped reads with less than three mismatched bases were processed using the error estimation model for all target regions, and variants with $P^m > P^e$ were qualified for subsequent analyses. Then, variants with more than 1% of reads supporting an alternative allele in either of the parents were filtered to remove the inherited variants. Due to the potential of misalignment, we only retained the variants excluded by the strict mask regions of the 1000 Genomes Project phase 1 [58].

Mutational signature analyses

Mutational signatures were analyzed based on the guidelines of the Wellcome Trust Sanger Institute [11,25]. We first calculated the percentages of the 96 possible mutated trinucleotides in each sample, which were identified according to the six classes of base substitutions and the 16 sequence contexts immediately 5′ and 3′ to the mutated base. The contexts of all mutations were extracted from the human reference genome build GRCh37. Then the mutational signatures in the selected samples were estimated using the NMF learning strategy. The appropriate number of mutational signatures was identified by calculating the reproducibility value and the reconstruction error for all samples. Finally, each mutational signature was displayed with the proportion of the 96 trinucleotides, and its contribution to each sample was estimated.

Cell culture and molecular barcoded whole-exome sequencing

WA07 (WiCell Research Institute, Madison, WI) cells were divided into two groups with ~ 1000 cells each and maintained in the human pluripotent stem cell chemical-defined medium (Catalog No. 400105, Baishou Biotechnology Co. Nanjing, China) according to the manufacturer's protocol. One group was treated with ML228 at 0.125 μM and the other group was treated with mock as the control. Both groups were cultured for 15 days, and the cells received fresh medium with/without ML228 every other day. Then, ~ 10 cells were randomly selected from each group and cultured in the aforementioned medium with/without ML228 (0.125 μM). Cells received fresh medium with/without ML228 every five days. Molecular barcoded whole-exome sequencing was performed on each group of cells after being expanded to ~ 1000 cells. The genomic DNA of cultured expanded WA07 cells was extracted with a QIAamp DNA Mini Kit (Catalog No. 51304, Qiagen, Hilden, Germany) following the manufacturer's protocols. Partition-barcoded libraries were prepared using the Chromium Exome Solution (Catalog No. 1000017, 10X Genomics, San Francisco, CA), and the exome target regions were enriched using SureSelect Human All Exon V5 Kit (Catalog No. G9448, Agilent, Santa Clara, CA) according to the manufacturer's protocols. Target-enriched libraries with molecular barcoding were subsequently sequenced on a HiSeq 4000 (Illumina, San Diego, CA) with 150 bp paired-end reads.

Mutation detection in hESCs

Exome sequencing data with molecular barcodes of WA07 cells were analyzed via Long Ranger (10X Genomics, San Francisco, CA). Then, the mutations that contained multiple molecular barcodes in the mismatched reads were kept. To reduce the false positive rate, we removed the mutations that contained two allele types in one molecular barcode at the site.

Ethical statement

This study was approved by the ethics committees of both Beijing Institute of Genomics, Chinese Academy of Sciences / China National Center for Bioinformation (No. 2016H006) and the Eye Hospital of Wenzhou Medical University, China (No. KYK [2015] 2), and it was conducted in accordance with the Declaration of Helsinki Principles. Individuals F0061, M0070, M0072, M0074, and their parents were enrolled at the Wenzhou Medical University and samples from M0038 and his parents were collected at Beijing Institute of Genomics, Chinese Academy of Sciences / China National Center for Bioinformation. All the participants were healthy and provided written informed consent.

Code availability

The software for postzygotic mutation calling is available at https://ngdc.cncb.ac.cn/biocode/ (BioCode: BT007245).

Data availability

All sequencing data generated in the current study are available in the Genome Sequence Archive [59] at the National Genomics Data Center, Beijing Institute of Genomics, Chinese Academy of Sciences / China National Center for Bioinformation (GSA: CRA000071), and are publicly accessible at https://ngdc.cncb.ac.cn/gsa.

CRediT author statement

Yaqiang Hong: Methodology, Software, Formal analysis, Data curation, Writing - original draft, Visualization. **Dake Zhang:** Investigation, Resources, Writing - review & editing, Visualization, Funding acquisition. **Xiangtian Zhou:** Validation, Investigation, Resources. **Aili Chen:** Software, Resources. **Amir Abliz:** Software, Resources. **Jian Bai:** Software. **Liang Wang:** Resources. **Qingtao Hu:** Software. **Kenan Gong:** Validation. **Xiaonan Guan:** Validation. **Mengfei Liu:** Resources. **Xinchang Zheng:** Resources. **Shujuan Lai:** Investigation. **Hongzhu Qu:** Software. **Fuxin Zhao:** Resources. **Shuang Hao:** Investigation. **Zhen Wu:** Resources. **Hong Cai:** Resources. **Shaoyan Hu:** Resources. **Yue Ma:** Resources. **Junting Zhang:** Resources. **Yang Ke:** Resources. **Qian-Fei Wang:** Resources. **Wei Chen:** Conceptualization, Methodology, Data curation, Writing - original draft, Project administration, Funding acquisition. **Changqing Zeng:** Writing - review & editing, Supervision, Project administration, Funding acquisition. All authors have read and approved the final manuscript

Competing interests

The authors have declared no competing interests.

Acknowledgments

This study was supported by the grants from the Strategic Priority Research Program of Chinese Academy of Sciences (Grant No. XDB13020500), the National Natural Science Foundation of China (NSFC) (Grant Nos. 91131905, 31471199, and 91631304), the Key Research Program of Chinese Academy of Sciences (Grant No. KJZD-EW-L14 to CZ), the NSFC (Grant Nos. 31440057 and 31701081 to WC), the 111 Project (Grant No. B13003 to WC and DZ), and the Innovation Promotion Association of Chinese Academy of Sciences (Grant Nos. 2016098 to DZ and 2019103 to AC). We gratefully thank Dr. Peter Reinach from Wenzhou Medical University and Charlesworth Author Services for editing and proofreading the manuscript. We thank Dr. Ian M. Campbell and Dr. Pawel Stankiewicz from Baylor College of Medicine for kindly providing R source code for building cell-division model. We thank Dr. Xinyu Zhang from Yale University for critical reading and comments on the manuscript. We also thank Dr. Caixia Guo from Beijing Institute of Genomics, Chinese Academy of Sciences / China National Center for Bioinformation for interpreting possible mechanisms underlying the occurrence of postzygotic mutations.

Supplementary material

Supplementary data to this article can be found online at https://doi.org/10.1016/j.gpb.2021.09.005.

ORCID

ORCID 0000-0002-6395-5037 (Yaqiang Hong)
ORCID 0000-0001-9508-8209 (Dake Zhang)
ORCID 0000-0002-1115-561X (Xiangtian Zhou)
ORCID 0000-0002-1751-1208 (Aili Chen)
ORCID 0000-0002-0557-4244 (Amir Abliz)
ORCID 0000-0001-5667-0618 (Jian Bai)
ORCID 0000-0003-3045-4903 (Liang Wang)
ORCID 0000-0002-8291-7116 (Qingtao Hu)
ORCID 0000-0002-0517-6078 (Kenan Gong)
ORCID 0000-0003-2614-6916 (Xiaonan Guan)
ORCID 0000-0001-6595-3855 (Mengfei Liu)
ORCID 0000-0001-5739-861X (Xinchang Zheng)
ORCID 0000-0003-1249-4258 (Shujuan Lai)
ORCID 0000-0001-7013-8409 (Hongzhu Qu)
ORCID 0000-0002-1006-6180 (Fuxin Zhao)
ORCID 0000-0001-7831-5940 (Shuang Hao)
ORCID 0000-0003-1292-5070 (Zhen Wu)
ORCID 0000-0002-3386-6957 (Shaoyan Hu)
ORCID 0000-0002-3922-594X (Yue Ma)
ORCID 0000-0003-3763-1095 (Junting Zhang)
ORCID 0000-0003-4394-8814 (Yang Ke)
ORCID 0000-0002-0086-2626 (Qian-Fei Wang)
ORCID 0000-0002-8087-7077 (Wei Chen)
ORCID 0000-0002-0037-1771 (Changqing Zeng)

References

[1] Rahbari R, Wuster A, Lindsay SJ, Hardwick RJ, Alexandrov LB, Al Turki S, et al. Timing, rates and spectra of human germline mutation. Nat Genet 2016;48:126–33.
[2] Alexandrov LB, Jones PH, Wedge DC, Sale JE, Campbell PJ, Nik-Zainal S, et al. Clock-like mutational processes in human somatic cells. Nat Genet 2015;47:1402–7.
[3] Hoeijmakers JHJ. Genome maintenance mechanisms for preventing cancer. Nature 2001;411:366–74.
[4] Stratton MR, Campbell PJ, Futreal PA. The cancer genome. Nature 2009;458:719–24.
[5] Martincorena I, Roshan A, Gerstung M, Ellis P, Van Loo P, McLaren S, et al. Tumor evolution. High burden and pervasive positive selection of somatic mutations in normal human skin. Science 2015;348:880–6.
[6] Brunner SF, Roberts ND, Wylie LA, Moore L, Aitken SJ, Davies SE, et al. Somatic mutations and clonal dynamics in healthy and cirrhotic human liver. Nature 2019;574:538–42.
[7] Helleday T, Eshtad S, Nik-Zainal S. Mechanisms underlying mutational signatures in human cancers. Nat Rev Genet 2014;15:585–98.

[8] Hart JR, Zhang Y, Liao L, Ueno L, Du L, Jonkers M, et al. The butterfly effect in cancer: a single base mutation can remodel the cell. Proc Natl Acad Sci U S A 2015;112:1131–6.
[9] Blokzijl F, de Ligt J, Jager M, Sasselli V, Roerink S, Sasaki N, et al. Tissue-specific mutation accumulation in human adult stem cells during life. Nature 2016;538:260–4.
[10] Martincorena I, Fowler JC, Wabik A, Lawson ARJ, Abascal F, Hall MWJ, et al. Somatic mutant clones colonize the human esophagus with age. Science 2018;362:911–7.
[11] Alexandrov LB, Nik-Zainal S, Wedge DC, Aparicio SA, Behjati S, Biankin AV, et al. Signatures of mutational processes in human cancer. Nature 2013;500:415–21.
[12] Alexandrov LB, Nik-Zainal S, Wedge DC, Campbell PJ, Stratton MR. Deciphering signatures of mutational processes operative in human cancer. Cell Rep 2013;3:246–59.
[13] Lee DD, Seung HS. Learning the parts of objects by non-negative matrix factorization. Nature 1999;401:788–91.
[14] Alexandrov LB, Ju YS, Haase K, Van Loo P, Martincorena I, Nik-Zainal S, et al. Mutational signatures associated with tobacco smoking in human cancer. Science 2016;354:618–22.
[15] Kong A, Frigge ML, Masson G, Besenbacher S, Sulem P, Magnusson G, et al. Rate of *de novo* mutations and the importance of father's age to disease risk. Nature 2012;488:471–5.
[16] Cai J, Miao X, Li Y, Smith C, Tsang K, Cheng L, et al. Whole-genome sequencing identifies genetic variances in culture-expanded human mesenchymal stem cells. Stem Cell Rep 2014;3:227–33.
[17] Krimmel JD, Schmitt MW, Harrell MI, Agnew KJ, Kennedy SR, Emond MJ, et al. Ultra-deep sequencing detects ovarian cancer cells in peritoneal fluid and reveals somatic *TP53* mutations in noncancerous tissues. Proc Natl Acad Sci U S A 2016;113:6005–10.
[18] Hoang ML, Kinde I, Tomasetti C, McMahon KW, Rosenquist TA, Grollman AP, et al. Genome-wide quantification of rare somatic mutations in normal human tissues using massively parallel sequencing. Proc Natl Acad Sci U S A 2016;113:9846–51.
[19] Navin N, Kendall J, Troge J, Andrews P, Rodgers L, McIndoo J, et al. Tumour evolution inferred by single-cell sequencing. Nature 2011;472:90–4.
[20] Forsberg LA, Gisselsson D, Dumanski JP. Mosaicism in health and disease - clones picking up speed. Nat Rev Genet 2017;18:128–42.
[21] Lee-Six H, Olafsson S, Ellis P, Osborne RJ, Sanders MA, Moore L, et al. The landscape of somatic mutation in normal colorectal epithelial cells. Nature 2019;574:532–7.
[22] Behjati S, Huch M, van Boxtel R, Karthaus W, Wedge DC, Tamuri AU, et al. Genome sequencing of normal cells reveals developmental lineages and mutational processes. Nature 2014;513:422–5.
[23] St Charles JA, Liberti SE, Williams JS, Lujan SA, Kunkel TA. Quantifying the contributions of base selectivity, proofreading and mismatch repair to nuclear DNA replication in *Saccharomyces cerevisiae*. DNA Repair (Amst) 2015;31:41–51.
[24] Strick TR, Savery NJ. Understanding bias in DNA repair. Proc Natl Acad Sci U S A 2017;114:2791–3.
[25] Nik-Zainal S, Alexandrov LB, Wedge DC, Van Loo P, Greenman CD, Raine K, et al. Mutational processes molding the genomes of 21 breast cancers. Cell 2012;149:979–93.
[26] Kopan R, Ilagan MX. The canonical Notch signaling pathway: unfolding the activation mechanism. Cell 2009;137:216–33.
[27] Das D, Lanner F, Main H, Andersson ER, Bergmann O, Sahlgren C, et al. Notch induces cyclin-D1-dependent proliferation during a specific temporal window of neural differentiation in ES cells. Dev Biol 2010;348:153–66.
[28] Joshi I, Minter LM, Telfer J, Demarest RM, Capobianco AJ, Aster JC, et al. Notch signaling mediates G1/S cell-cycle progression in T cells via cyclin D3 and its dependent kinases. Blood 2009;113:1689–98.
[29] Eram MS, Bustos SP, Lima-Fernandes E, Siarheyeva A, Senisterra G, Hajian T, et al. Trimethylation of histone H3 lysine 36 by human methyltransferase PRDM9 protein. J Biol Chem 2014;289:12177–88.
[30] Davies B, Hatton E, Altemose N, Hussin JG, Pratto F, Zhang G, et al. Re-engineering the zinc fingers of PRDM9 reverses hybrid sterility in mice. Nature 2016;530:171–6.
[31] Lee S, Lee DK, Dou Y, Lee J, Lee B, Kwak E, et al. Coactivator as a target gene specificity determinant for histone H3 lysine 4 methyltransferases. Proc Natl Acad Sci U S A 2006;103:15392–7.
[32] Lee-Six H, Øbro NF, Shepherd MS, Grossmann S, Dawson K, Belmonte M, et al. Population dynamics of normal human blood inferred from somatic mutations. Nature 2018;561:473–8.
[33] Zhang L, Dong X, Lee M, Maslov AY, Wang T, Vijg J. Single-cell whole-genome sequencing reveals the functional landscape of somatic mutations in B lymphocytes across the human lifespan. Proc Natl Acad Sci U S A 2019;116:9014–9.
[34] Moore L, Cagan A, Coorens THH, Neville MDC, Sanghvi R, Sanders MA, et al. The mutational landscape of human somatic and germline cells. Nature 2021;597:381–6.
[35] Forbes SA, Beare D, Gunasekaran P, Leung K, Bindal N, Boutselakis H, et al. COSMIC: exploring the world's knowledge of somatic mutations in human cancer. Nucleic Acids Res 2015;43:D805–11.
[36] Webb JD, Muranyi A, Pugh CW, Ratcliffe PJ, Coleman ML. MYPT1, the targeting subunit of smooth-muscle myosin phosphatase, is a substrate for the asparaginyl hydroxylase factor inhibiting hypoxia-inducible factor (FIH). Biochem J 2009;420:327–33.
[37] Saucedo LJ, Edgar BA. Filling out the Hippo pathway. Nat Rev Mol Cell Biol 2007;8:613–21.
[38] Pan D. The Hippo signaling pathway in development and cancer. Dev Cell 2010;19:491–505.
[39] Ma B, Chen Y, Chen L, Cheng H, Mu C, Li J, et al. Hypoxia regulates Hippo signalling through the SIAH2 ubiquitin E3 ligase. Nat Cell Biol 2015;17:95–103.
[40] Gossage L, Eisen T, Maher ER. *VHL*, the story of a tumour suppressor gene. Nat Rev Cancer 2015;15:55–64.
[41] Theriault JR, Felts AS, Bates BS, Perez JR, Palmer M, Gilbert SR, et al. Discovery of a new molecular probe ML228: an activator of the hypoxia inducible factor (HIF) pathway. Bioorg Med Chem Lett 2012;22:76–81.
[42] Chen L, Liu P, Evans TC, Ettwiller LM. DNA damage is a pervasive cause of sequencing errors, directly confounding variant identification. Science 2017;355:752–6.
[43] Costello M, Pugh TJ, Fennell TJ, Stewart C, Lichtenstein L, Meldrim JC, et al. Discovery and characterization of artifactual mutations in deep coverage targeted capture sequencing data due to oxidative DNA damage during sample preparation. Nucleic Acids Res 2013;41:e67.
[44] Salk JJ, Schmitt MW, Loeb LA. Enhancing the accuracy of next-generation sequencing for detecting rare and subclonal mutations. Nat Rev Genet 2018;19:269–85.
[45] Hypoxia SR. New connections. Nat Rev Cancer 2012;12:320.
[46] Li Z, Bao S, Wu Q, Wang H, Eyler C, Sathornsumetee S, et al. Hypoxia-inducible factors regulate tumorigenic capacity of glioma stem cells. Cancer Cell 2009;15:501–13.
[47] Wang J, Cazzato E, Ladewig E, Frattini V, Rosenbloom DI, Zairis S, et al. Clonal evolution of glioblastoma under therapy. Nat Genet 2016;48:768–76.
[48] Wang GL, Jiang BH, Rue EA, Semenza GL. Hypoxia-inducible factor 1 is a basic-helix-loop-helix-PAS heterodimer regulated by cellular O_2 tension. Proc Natl Acad Sci U S A 1995;92:5510–4.
[49] Simon MC, Keith B. The role of oxygen availability in embryonic development and stem cell function. Nat Rev Mol Cell Biol 2008;9:285–96.

[50] Dunwoodie SL. The role of hypoxia in development of the mammalian embryo. Dev Cell 2009;17:755–73.
[51] Lodato MA, Woodworth MB, Lee S, Evrony GD, Mehta BK, Karger A, et al. Somatic mutation in single human neurons tracks developmental and transcriptional history. Science 2015;350:94–8.
[52] Stiles J, Jernigan TL. The basics of brain development. Neuropsychol Rev 2010;20:327–48.
[53] Martincorena I, Campbell PJ. Somatic mutation in cancer and normal cells. Science 2015;349:1483–9.
[54] Marais G. Biased gene conversion: implications for genome and sex evolution. Trends Genet 2003;19:330–8.
[55] McGarry T, Biniecka M, Veale DJ, Fearon U. Hypoxia, oxidative stress and inflammation. Free Radic Biol Med 2018;125:15–24.
[56] Poetsch AR, Boulton SJ, Luscombe NM. Genomic landscape of oxidative DNA damage and repair reveals regioselective protection from mutagenesis. Genome Biol 2018;19:215.
[57] Li H, Durbin R. Fast and accurate long-read alignment with Burrows-Wheeler transform. Bioinformatics 2010;26:589–95.
[58] 1000 Genomes Project Consortium, Abecasis GR, Auton A, Brooks LD, DePristo MA, Durbin RM, et al. An integrated map of genetic variation from 1,092 human genomes. Nature 2012;491:56–65.
[59] Chen T, Chen X, Zhang S, Zhu J, Tang B, Wang A, et al. The Genome Sequence Archive Family: toward explosive data growth and diverse data types. Genomics Proteomics Bioinformatics 2021;19:578–83.