

# UC San Diego

## UC San Diego Previously Published Works

### Title

MosaicBase: A Knowledgebase of Postzygotic Mosaic Variants in Noncancer Disease-related and Healthy Human Individuals

### Permalink

<https://escholarship.org/uc/item/20x304f9>

### Journal

Genomics Proteomics & Bioinformatics, 18(2)

### ISSN

1672-0229

### Authors

Yang, Xiaoxu  
Yang, Changhong  
Zheng, Xianing  
et al.

### Publication Date

2020-04-01

### DOI

10.1016/j.gpb.2020.05.002

Peer reviewed



## DATABASE

# MosaicBase: A Knowledgebase of Postzygotic Mosaic Variants in Noncancer Disease-related and Healthy Human Individuals



Xiaoxu Yang<sup>1,#</sup>, Changhong Yang<sup>2,3,4,#</sup>, Xianing Zheng<sup>4,#</sup>, Luoxing Xiong<sup>5</sup>, Yutian Tao<sup>4,6</sup>, Meng Wang<sup>1</sup>, Adam Yongxin Ye<sup>1,5</sup>, Qixi Wu<sup>7</sup>, Yanmei Dou<sup>1</sup>, Junyu Luo<sup>4</sup>, Liping Wei<sup>1,\*</sup>, August Yue Huang<sup>1,\*</sup>

<sup>1</sup> Center for Bioinformatics, State Key Laboratory of Protein and Plant Gene Research, School of Life Sciences, Peking University, Beijing 100871, China

<sup>2</sup> Department of Bioinformatics, Chongqing Medical University, Chongqing 400016, China

<sup>3</sup> College of Life Sciences, Beijing Normal University, Beijing 100875, China

<sup>4</sup> National Institute of Biological Sciences, Beijing 102206, China

<sup>5</sup> Peking-Tsinghua Center for Life Sciences (CLS), Academy for Advanced Interdisciplinary Studies, Peking University, Beijing 100871, China

<sup>6</sup> Chinese Academy of Medical Sciences and Peking Union Medical College, Beijing 100730, China

<sup>7</sup> School of Life Sciences, Peking University, Beijing 100871, China

Received 9 March 2019; revised 18 March 2020; accepted 31 May 2020

Available online 8 September 2020

Handled by Qian-Fei Wang

## KEYWORDS

Postzygotic;  
Mosaicism;  
Noncancer;  
Mutation;  
MosaicBase

**Abstract** Mosaic variants resulting from **postzygotic** mutations are prevalent in the human genome and play important roles in human diseases. However, except for cancer-related variants, there is no collection of postzygotic mosaic variants in **noncancer** disease-related and healthy individuals. Here, we present **MosaicBase**, a comprehensive database that includes 6698 mosaic variants related to 266 noncancer diseases and 27,991 mosaic variants identified in 422 healthy individuals. Genomic and phenotypic information of each variant was manually extracted and curated from 383 publications. MosaicBase supports the query of variants with Online Mendelian Inheritance in Man (OMIM) entries, genomic coordinates, gene symbols, or Entrez IDs. We also provide an integrated genome browser for users to easily access mosaic variants and their related annotations for any genomic region. By analyzing the variants collected in MosaicBase, we find that mosaic variants that directly

\* Corresponding authors.

E-mail: [weilp@mail.cbi.pku.edu.cn](mailto:weilp@mail.cbi.pku.edu.cn) (Wei L), [huangy@mail.cbi.pku.edu.cn](mailto:huangy@mail.cbi.pku.edu.cn) (Huang AY).

# Equal contribution.

Peer review under responsibility of Beijing Institute of Genomics, Chinese Academy of Sciences and Genetics Society of China.

<https://doi.org/10.1016/j.gpb.2020.05.002>

1672-0229 © 2020 The Authors. Published by Elsevier B.V. and Science Press on behalf of Beijing Institute of Genomics, Chinese Academy of Sciences and Genetics Society of China.

This is an open access article under the CC BY license (<http://creativecommons.org/licenses/by/4.0/>).

contribute to disease phenotype show features distinct from those of variants in individuals with mild or no phenotypes, in terms of their genomic distribution, **mutation** signatures, and fraction of mutant cells. MosaicBase will not only assist clinicians in genetic counseling and diagnosis but also provide a useful resource to understand the genomic baseline of postzygotic mutations in the general human population. MosaicBase is publicly available at <http://mosaicbase.com/> or <http://49.4.21.8:8000>.

## Introduction

Genomic mosaicism results from postzygotic mutations that arise during embryonic development, tissue self-renewal [1], aging processes [2], or exposure to other DNA-damaging circumstances [3]. Unlike *de novo* or inherited germline variants that affect every cell in the carrier individual [4], postzygotic mosaic variants only affect a portion of cells or cell populations, and their mutant allelic fractions (MAFs) should be less than 50% [5]. If a postzygotic mutation affects germ cells [6], the mutant allele may theoretically be transmitted to offspring, which is the major source of genetic variations in the human population [7].

Postzygotic mosaic variants have previously been demonstrated to be directly responsible for the etiology of cancer [8,9] and an increasing number of other Mendelian or complex diseases, including epilepsy-related neurodevelopment disorders [10], Costello syndrome [11], autism spectrum disorders [12,13], and intellectual disability [14]. On the other hand, pathogenic genetic variants inherited from detectable parental mosaicism have been demonstrated to be an important source of monogenic genetic disorders, including Noonan syndrome [15], Marfan syndrome [16], Dravet syndrome [17], and complex disorders, including autism [18] and intellectual disability [19]. The MAF of a mosaic variant has been reported to be directly related to the carrier's phenotype [20,21] and to be associated with the recurrence risk in children [5].

With the rapid advances in next-generation sequencing (NGS) technologies, tens of thousands of postzygotic mosaic single-nucleotide variants (SNVs) and insertions/deletions (indels) have been identified in the genomes of human individuals [3,22,23]. However, except for cancer-related variants that have been collected by databases such as the Catalogue of Somatic Mutations in Cancer (COSMIC) [24] and somatic mutations impacting microRNA function in cancer (Soma-miR) [25], there is no integrated database focusing on mosaic variants in noncancer disease-related and healthy individuals.

Here, we present MosaicBase. To our best knowledge, MosaicBase is the first knowledgebase of mosaic SNVs and indels identified in patients with noncancer diseases and their parents or grandparents as well as healthy individuals. MosaicBase currently contains 34,689 validated mosaic variants that have been manually curated from 383 publications. MosaicBase has further integrated comprehensive genomic and phenotypic information about each variant and its carrier. It provides multi-scale information about disease-related mosaic variants for genetic counseling and molecular diagnosis, as well as the genomic background of mosaic variants in general population.

## Database implementation

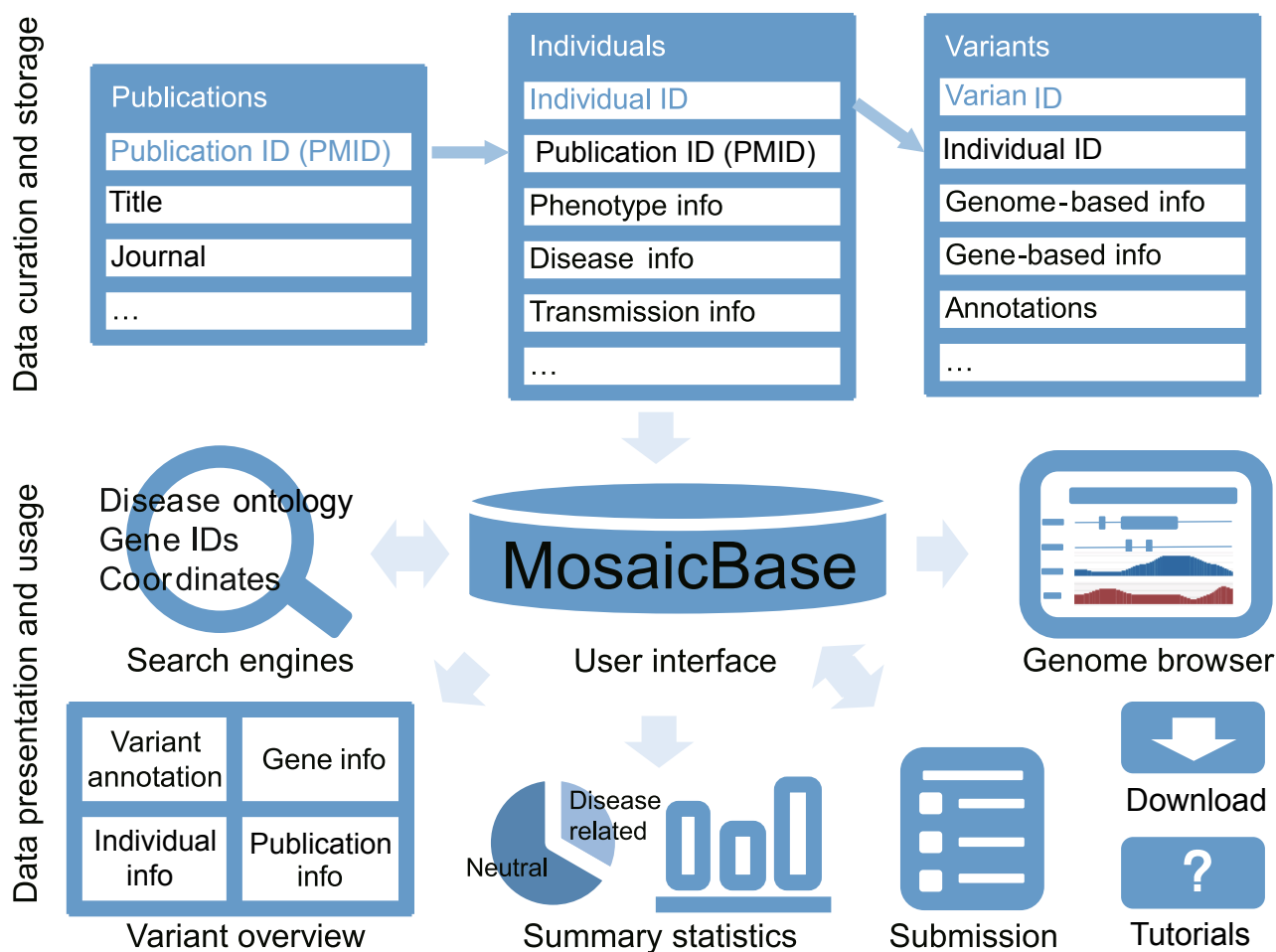
### Framework of MosaicBase

An overview of the framework of MosaicBase is shown in **Figure 1**. MosaicBase consists of two logical parts: the database and server as the backend, and the user interface as the frontend. Structured data based on three main relational tables were established in the backend of MosaicBase. The storage and maintenance of the database were implemented with SQLite v3. The frontend of MosaicBase provides a user-friendly interface written in Personal home page Hypertext Preprocessor (PHP), JavaScript, HyperText Markup Language (HTML), and Cascading Style Sheets (CSS), with Django applications.

MosaicBase incorporates two different search modes (basic mode and ontology-based mode) to help users browse the database. The information for each mosaic variant has been summarized from the publication and individual levels to the gene and variant levels. A built-in genome browser is provided to visualize variants. The statistical summaries and detailed tutorials for MosaicBase are available on the main page. MosaicBase further provides an online submission system to encourage the community to contribute to the database.

### Data collection, processing, and annotation

We queried against the PubMed database using keywords including “mosaic”, “mosaicism”, “post-zygotic”, “somatic”, “sequencing” (see the full query string in File S1), and excluded publications about cancer-related mosaic mutations or studies on non-human organisms by examining the titles and abstracts. For more than 1000 search results, we scrutinized the main text as well as supplemental information to confirm the relevance of each publication. After this process, 383 journal research articles about mosaic SNVs and indels in noncancer individuals that were published between Jan 1989 and May 2018 were collected into MosaicBase. For each article, data fields for the publication, individual, and variation information were extracted and saved into tables in the backend (**Figure 1**). For studies involving single-cell technologies, only the validated or high-confidence postzygotic mosaic SNVs were collected. For the table of variation information, we further integrated genomic annotations generated by ANNOtate VARIation (ANNOVAR) [26], including population allele frequency from dbSNP (version 137) [27] and gnomAD (genome; version 2.0.1) [28], risk scores such as Combined Annotation-Dependent Depletion (CADD) scores (version 1.30) [29] and Eigen scores [30], functional predictions by Functional Analysis through Hidden Markov Models (FATHMM) [31], Sorting Intolerant From Tolerant (SIFT)



**Figure 1** Overview of the data collection, storage, and visualization of MosaicBase

[32], iFish2 [33], DeFine [34], conservation prediction by GERP++ [35] and PhyloP [36], and annotations in COSMIC [37]. A detailed description of different fields and data types required in each field is listed in Tables S1, S2, and S3. The transcript-based variation information was confirmed using Mutalyzer following the suggestions from the Human Genome Variation Society (HGVS) [38]. Genomic coordinates were provided according to the human reference genome from University of California at Santa Cruz (UCSC) hg19/GRCh37 as well as hg38/GRCh38.

#### Statistical analysis and visualization of mosaic variants

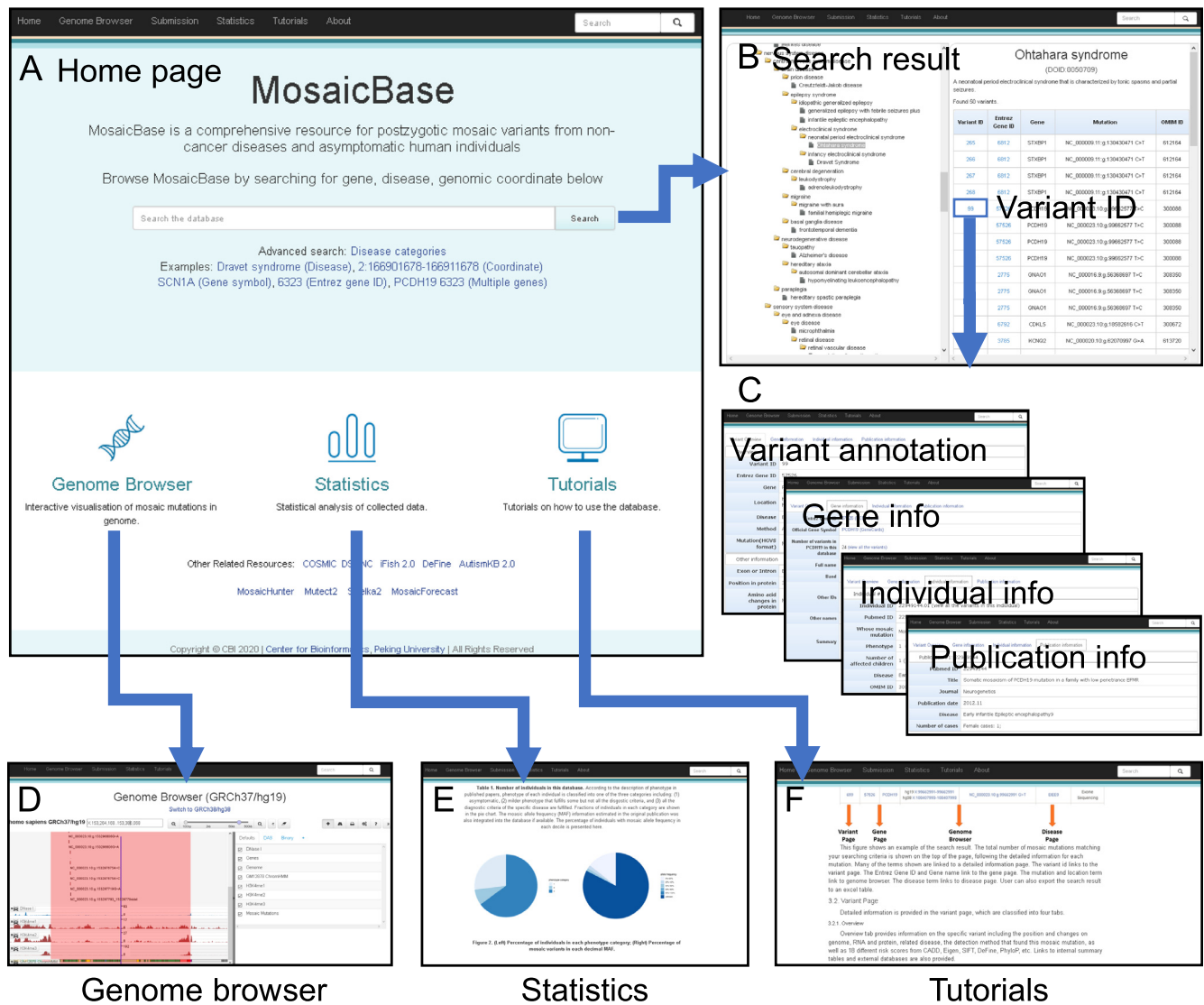
Mutation signature analysis has been widely used in cancer studies to elaborate the etiology of somatic mosaic variants, by decomposing the matrix of tri-nucleotide context into cancer-related signatures. In this study, the signature of non-cancer mosaic variants was analyzed by Mutalisk [39], and the maximum likelihood estimation of proportions for each mutation signature was performed based on a greedy algorithm. For each variant group, we further tested whether its genomic density within each 1-Mb interval was linearly correlated with the GC content, DNase I hypersensitive (DHS) regions, replication timing, and histone modification profiles

measured in the GM12878 cell line [39]. A genome browser based on the Dalliace platform [40] was implemented to interactively visualize mosaic variants. Circos [41] was utilized to show the genomic distribution of mosaic variants.

## Web interface

### User interface and functions

We incorporate two search modes in MosaicBase. The basic search mode provided on the main page recognizes search terms based on the name of diseases, the range of genomic coordinates, gene symbols, or Entrez Gene IDs (Figure 2A), in which the search engine is comparable with space-delimited multiple search terms. The result page of the basic search mode displays variant summary information according to the categories of search terms, and search results can then be downloaded as an .xls format table. We also include an ontology-based search mode as an advanced option in MosaicBase. With this mode, users can browse the mosaic variants relevant to a specific disease or disease category according to the Disease Ontology [42]. A brief summary of the description of the disease or disease category is provided along with a sum-



**Figure 2** Screenshots of MosaicBase

**A.** The main page provides the search modes and multiple links to different utilities of the database. **B.** Disease ontology-based advanced search page and an example of a result table. **C.** The variant pages from the basic search results; these pages provide information about each variant and its corresponding gene, individual and publication annotation, the individuals carrying the same variant, as well as the publications describing the variant. **D.** Integrated genome browser to visualize mosaic variants with genetic and epigenetic annotations. **E.** Summary statistics of the publications, mutational spectrum, and individuals collected in MosaicBase. **F.** Detailed tutorials for the introduction, data presentation, and usage of MosaicBase.

mary table of all the related mosaic variants collected in MosaicBase (Figure 2B).

Detailed information about each mosaic variant is summarized in four different panels in MosaicBase. These include the overview panel, the gene information panel, the individual information panel, and the publication information panel (Figure 2C). In the overview panel, we provide the genomic information as well as the identification and validation methodologies for each variant. In the gene information panel, we annotate each gene by its Entrez Gene ID, official gene symbol and alternative names, number of reported mosaic variants in this gene, Vega ID, OMIM ID, Human Genome Organisation Gene Nomenclature Committee (HGNC) ID, Ensembl ID, and a brief gene summary. In addition, we also

summarize all the collected mosaic variants from the same gene and provide links for gene annotation from external databases. In the individual information panel, we classify the phenotypes of the individual carrying the mosaic variant and display the information according to the original descriptions in the publication. The severity of phenotype collected in MosaicBase is defined as “1” if the carrier was healthy, “2” if the carrier had a mild phenotype but did not fulfill all the diagnostic criteria for a specific disease or characterized syndrome, and “3” if the carrier fulfilled all the clinical diagnostic criteria for a specific disease. In the publication information panel, we summarize the title, journal, sample, and additional information about the publication which reports the mosaic variant.

MosaicBase integrates a build-in genome browser to provide convenient interactive data visualization for the mosaic variants (Figure 2D). In addition to the built-in tracks about genetic and epigenetic annotations, such as DHS sites and H3K4me predictions, MosaicBase also allows the user to import customized tracks from URLs, UCSC-style track hubs, or to upload files in a UCSC-style genome browser track format. The URLs for tracks of Ensembl Gene and methylated DNA immunoprecipitation sequencing (MeDIP-seq) data are provided as examples, and a help page providing detailed guidance is also available by clicking the question mark in the top-left panel of the genome browser. MosaicBase further provides users with an application that can generate publication-quality scalable vector graphic (SVG) files from the control panel of the genome browser.

MosaicBase included a “Statistics” page to show a summary of all the collected mosaic variants (Figure 2E) and a “Tutorials” page (Figure 2F) with detailed introductions about the database, as well as its search modes, data presentation, and genome browser. We also implement an online submission system that allows users to submit mosaic variants from newly published or uncollected publications. Such variants will be manually examined by our team and integrated into MosaicBase with scheduled updates.

### Statistical analysis of noncancer mosaic variants

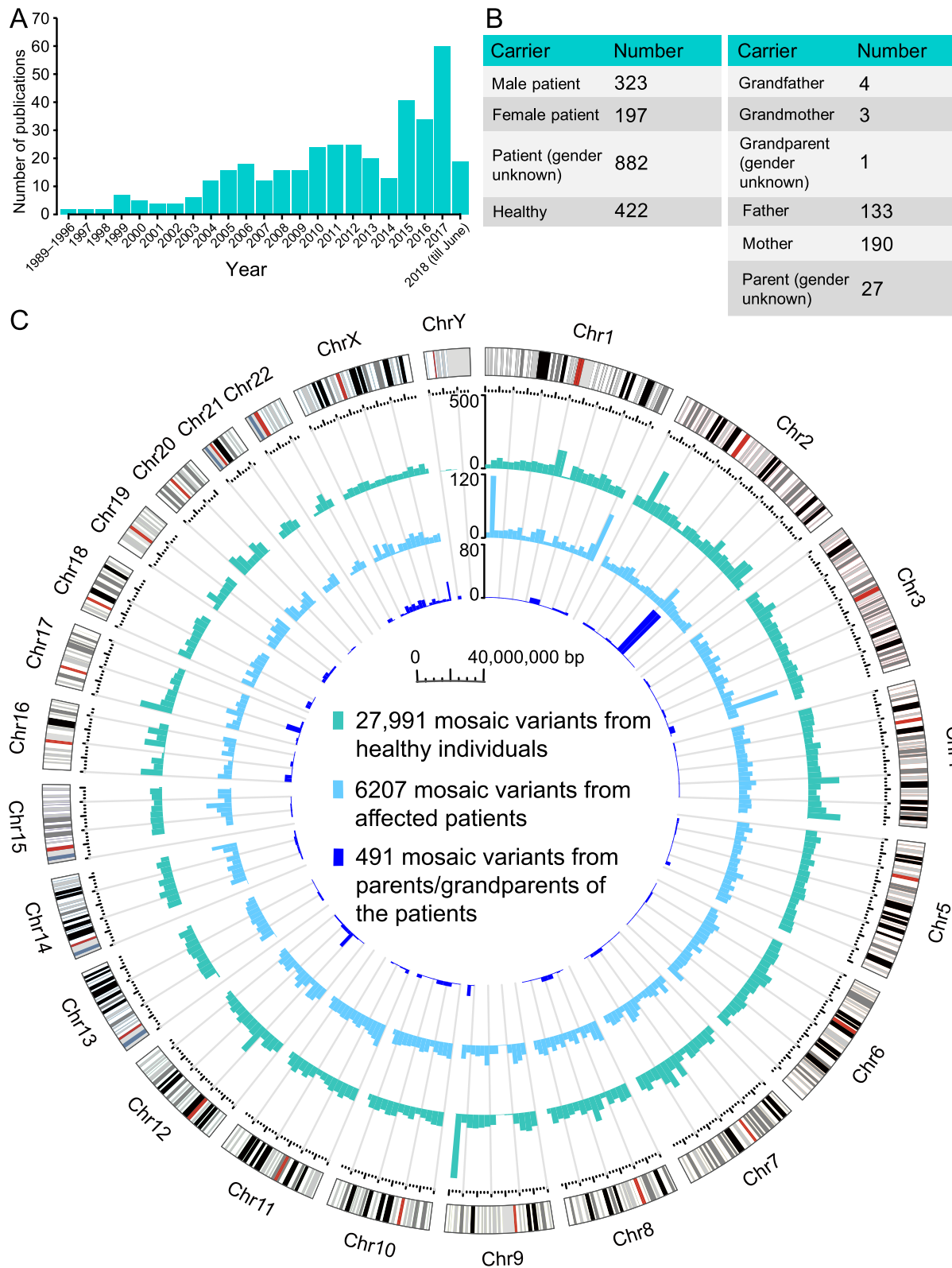
MosaicBase currently includes 383 journal research articles, letters, and clinical genetic reports about noncancer postzygotic mosaic variants that were published from 1989 to 2018 (Figure 3A), with an accelerated accumulation of relevant publications boosted by the recent advances in NGS technologies. After manually extracting the mosaic variants reported in each publication, we thoroughly compiled 34,689 mosaic variants from 2182 human individuals, including 6698 disease-related variants from 3638 genes related to 266 noncancer diseases from 1402 patients and 358 parents or grandparents of the patients, as well as 27,991 apparently neutral variants identified from 422 healthy individuals (Figure 3B and Table S4). Specifically, two types of disease-related mosaic variants have been collected in MosaicBase: (1) 6207 mosaic variants that directly contribute to the disease phenotype in 1402 patients (323 men, 197 women, and 882 cases with gender information not available from the original publications); and (2) 491 mosaic variants identified from 358 parents or grandparents (137 men, 193 women, and 28 cases with gender information not available from the original publications) of the probands who had transmitted the mosaic allele to their offspring for a heterozygous genotype that lead to disease phenotypes (Figure 3B). The collected mosaic variants are classified into three groups according to the origin of the variants described in the original publications. Variants from healthy individuals are termed the “healthy” group; variants from patients fulfilling the full diagnostic criteria of a specific disease are termed the “patients” group; and variants from parents/grandparents of the patients are termed the “parents/grandparents” group. As shown in Figure 3C, mosaic variants are generally distributed across all the autosomes and X chromosomes. Parental mosaic variants are clustered in the *SCN1A* gene on

chromosome 2, which results from the well-studied parental mosaic cases for Dravet syndrome [17,20]. The underrepresentation of mosaic variants in the Y chromosome might be explained by its low gene density and the technical challenge of detecting mosaic variants in haplotype chromosomes.

To study whether mosaic variants from different groups of individuals have distinct genomic characteristics, we calculated their correlation with various genomic regulation features, including GC content, DHS positions, and epigenetic modifications. Because the vast majority of mosaic variants have been identified from peripheral blood or saliva samples, genomic regulation annotations of the lymphoblastoid cell line GM12878 was used in this analysis. Common germline variants annotated in dbSNP 137 with allele frequency higher than 10% (“dbSNP” group) serve as a control. According to the Pearson correlation coefficients between the signal intensities of genomic features and the density of variants with a window size of 1 Mb across the genome [43], we find that the mosaic variants contributing directly to the disease phenotype (“patients” group) are more positively correlated with such genomic features than the mosaic variants of other groups (Figure 4A).

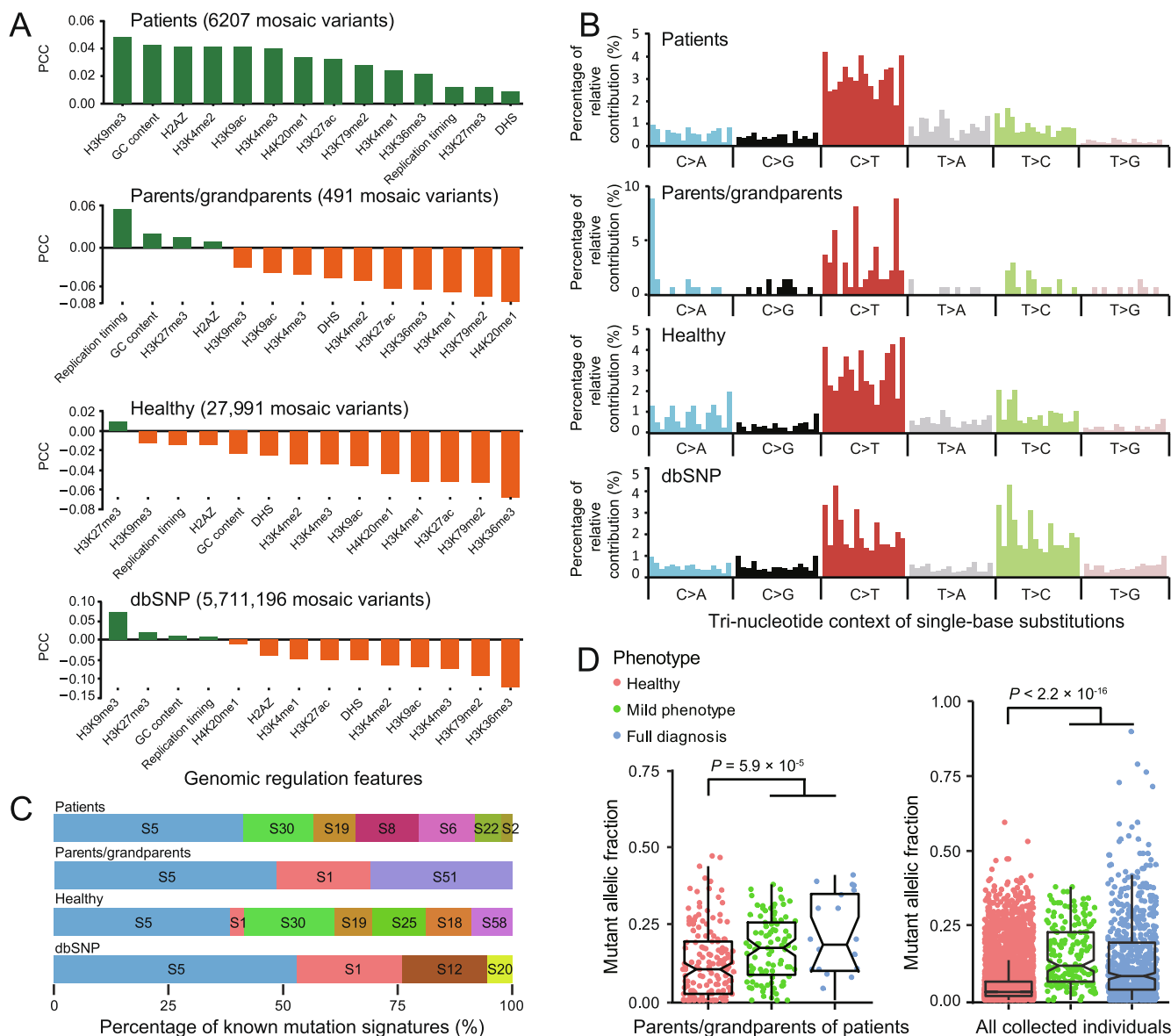
Next, we examined the mutation spectrum of the collected mosaic variants. Similar to inherited germline variants [44] and somatic variants reported in cancer studies [45], C > T is the most predominant type for mosaic variants (Figure 4B). We then extracted the tri-nucleotide genomic context of each variant and decomposed the matrix into mutation signatures previously identified in various types of cancers (<https://cancer.sanger.ac.uk/cosmic/signatures>). Mutation signature analysis further revealed that over 50% of the mosaic variants can be decomposed into the combination of cancer signatures 1, 5, and 30, whereas the remaining mosaic variants consist of signatures 2, 6, 8, 12, 18, 19, 20, 22, 25, 51, and 58 (Figure 4C). Signatures 1 and 5 result from the age-related process of spontaneous or enzymatic deamination of 5-methylcytosine to thymine; signatures 18 and 30 result from deficient base excision repair [46]; signature 2 indicates the activation of AID/APOBEC cytidine deaminase; signatures 6 and 20 are associated with defective DNA mismatch repair; signature 22 is associated with aristolochic acid exposure. The etiology of signatures 8, 12, 19, 25 are unknown, whereas signatures 51 and 58 are potential sequencing artefacts. Detailed descriptions of the signatures are provided in File S1.

To explore the general relationship between the MAF of a mosaic variant and its carrier’s phenotype, we extracted the allele fraction and phenotypic severity information for each mosaic variant in MosaicBase. Among parents and grandparents of patients (“parents/grandparents” group), we observed that the mosaic variants from carriers manifesting mild or full disease phenotypes had significantly higher MAFs than those from carriers without any disease phenotypes ( $P = 5.9 \times 10^{-5}$ ; two-tailed Mann–Whitney  $U$  test with continuity correction, Figure 4D), which is in accordance with previous observations [18,20,47]. When we considered mosaic variants in all the collected individuals, the difference became even more significant ( $P < 2.2 \times 10^{-16}$ ; two-tailed Mann–Whitney  $U$  test, Figure 4D). These results highlight the importance of the MAF information of mosaic variants in clinical applications such as genetic counseling.



**Figure 3** Statistics about the publication, individual, and variant data collected in MosaicBase

**A.** Number of relevant publications from 1989 to June 2018 collected from PubMed. Data before 1997 were condensed. Query terms and inclusion/exclusion criteria for literature acquisition are described in File S1. **B.** Summary of different categories of mosaic carriers. Left: Number of patients (grouped by gender) and healthy individuals. Right: Number of parents and grandparents (grouped by gender). **C.** Genomic distribution of mosaic variants. Chromosomal bands are illustrated in the outer circle with centromeres in red. Histograms show the count of mosaic variants for 1-Mb interval in each inner circle. Genomic coordinates and color codes of the categories are shown in the center.



**Figure 4** Genomic features of mosaic variants collected in MosaicBase

**A.** PCC of the density of mosaic variants and genomic features. Genomic density within each 1-Mb interval was linearly correlated with GC content, DHS regions, replication timing, and histone modification profiles measured in the GM12878 cell line. **B.** Tri-nucleotide genomic context of mosaic variants. **C.** Estimated contribution of mutation signatures for mosaic variants. 60 single-base substitution signatures and artefact signatures from COSMIC were considered, and a linear regression model was used to estimate the proportion of signatures. **D.** Mutant allele fraction of 491 mosaic variants in 358 individuals from parents/grandparents group (left) and 34,689 variants in all the 2182 individuals (including 1402 noncancer patients, 358 parents/grandparents of the patients, and 422 healthy individuals) collected in MosaicBase (right). Significantly higher mosaic allele fractions were observed in individuals with disease phenotypes ( $P = 5.9 \times 10^{-5}$  and  $P < 2.2 \times 10^{-16}$ ; two-tailed Mann–Whitney  $U$  test). In panels A–C, 5,711,196 common germline variants with population allele frequency  $\geq 0.1$  in dbSNP (version 137) were shown for comparison. H2AZ, histone 2A.Z variant; DHS, DNase I hypersensitive; PCC, Pearson correlation coefficient.

## Discussion

MosaicBase currently contains 34,689 mosaic SNVs and indels identified from patients with noncancer diseases and their parents or grandparents, as well as from healthy individuals, with rich information at the publication, individual, gene, and variant levels. The user-friendly interface of allows users to access MosaicBase by multiple searching methods and the integrated genome browser.

The pathogenic contribution of mosaic variants to noncancer diseases has been increasingly recognized in the past few years. MosaicBase provides genetic and phenotypic information about 6698 disease-related mosaic variants in 266 noncancer diseases. This database may help clinicians understand the pathogenesis and inheritance of mosaic variants and shed new light on future clinical applications, such as genetic counseling and diagnosis. On the other hand, the collection of 27,991 mosaic variants identified in healthy individuals could



be useful for understanding the genomic baseline of postzygotic mutations in the general human population. MosaicBase also integrates risk prediction from multiple computational tools for each variant. Unlike germline variants which are present in all cells of the carriers, mosaic variants are only present in a fraction of cells, in which the level of mosaic fraction can be an additional factor contributing to variant pathogenicity [18,20]. In the future, with the increasing number of relevant studies, we would expect a well-benchmarked scoring system specifically designed for predicting the deleterious probability of mosaic variants.

Of the 34,689 mosaic variants collected in MosaicBase, only 0.7%–8.7% are present in large-scale population polymorphism databases (Table S5). If we only consider common SNPs with population allele frequency (AF) higher than 0.01, the overlapping proportion further reduces to 0.1%–0.7%. This suggests that MosaicBase provides a unique set of human genetic variants which have been overlooked in previous genomic studies. Indeed, these apparently benign variants that are generated *de novo* show characteristics distinct from those of the variants that directly contribute to a disease phenotype, and also different from polymorphisms that are fixed in population under selective pressure (Figure 4). The data collected by MosaicBase will encourage researchers to reanalyze existing NGS data of human diseases by mosaic variant calling tools, such as MosaicHunter [48], Mutect2 [49], and Strelka [50], to identify previously ignored disease causative variants.

In the future, our team will update MosaicBase regularly by collecting and reviewing new publications in PubMed and publications submitted through our online submission system. After each update, we will update the statistics and release update reports on the website. We plan to further improve the user interface of MosaicBase and add new analysis tools based on feedback from the community.

## Data availability

MosaicBase is publicly available at <http://mosaicbase.com/> or <http://49.4.21.8:8000>.

## CRediT author statement

**Xiaoxu Yang:** Conceptualization, Data curation, Formal analysis, Software, Methodology, Visualization, Writing - original draft, Writing - review & editing, Project administration. **Changhong Yang:** Data curation, Software, Methodology, Formal analysis, Methodology, Visualization, Writing - original draft. **Xianing Zheng:** Data curation, Software, Methodology, Formal analysis, Visualization, Writing - original draft. **Luoxing Xiong:** Data curation. **Yutian Tao:** Data curation. **Meng Wang:** Software. **Adam Yongxin Ye:** Software. **Qixi Wu:** Data curation, Supervision. **Yanmei Dou:** Data curation. **Junyu Luo:** Data curation. **Liping Wei:** Conceptualization, Resources, Writing - review & editing, Funding acquisition, Supervision. **August Yue Huang:** Conceptualization, Supervision, Software, Writing - original draft, Writing - review & editing, Supervision, Project administration. All authors read and approved the final manuscript.

## Competing interests

The authors declare no competing interests.

## Acknowledgments

This work was supported by grants from the National Natural Science Foundation of China (Grant No. 31530092) and the Ministry of Science and Technology of China (Grant No. 2015AA020108) awarded to LW. We thank Dr. Lei Kong, Dr. Ge Gao, and Mr. Dechang Yang from the Centre for Bioinformatics, Peking University, for their help in setting up the hardware environments and maintaining the MosaicBase server. We thank Drs. Sijin Cheng and Jinpu Jin for their suggestions on the database.

## Supplementary material

Supplementary data to this article can be found online at <https://doi.org/10.1016/j.gpb.2020.05.002>.

## ORCID

0000-0003-0219-0023 (Yang, X)  
 0000-0002-7573-3765 (Yang, C)  
 0000-0002-3302-1241 (Zheng, X)  
 0000-0002-8884-0594 (Xiong, L)  
 0000-0002-9360-0771 (Tao, Y)  
 0000-0002-1072-7073 (Wang, M)  
 0000-0002-1542-0740 (Ye, YA)  
 0000-0002-9959-2629 (Wu, Q)  
 0000-0002-9328-1731 (Dou, Y)  
 0000-0002-4679-004X (Luo, J)  
 0000-0002-1795-8755 (Wei, L)  
 0000-0002-0416-2854 (Huang, YA)

## References

- [1] Huang AY, Yang X, Wang S, Zheng X, Wu Q, Ye AY, et al. Distinctive types of postzygotic single-nucleotide mosaicism in healthy individuals revealed by genome-wide profiling of multiple organs. *PLoS Genet* 2018;14:e1007395.1–19.
- [2] Holstege H, Pfeiffer W, Sie D, Hulsman M, Nicholas TJ, Lee CC, et al. Somatic mutations found in the healthy blood compartment of a 115-yr-old woman demonstrate oligoclonal hematopoiesis. *Genome Res* 2014;24:733–42.
- [3] 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.
- [4] Freed D, Stevens EL, Pevsner J. Somatic mosaicism in the human genome. *Genes (Basel)* 2014;5:1064–94.
- [5] Ye AY, Dou Y, Yang X, Wang S, Huang AY, Wei L. A model for postzygotic mosaicism quantifies the allele fraction drift, mutation rate, and contribution to *de novo* mutations. *Genome Res* 2018;28:1–9.
- [6] Biesecker LG, Spinner NB. A genomic view of mosaicism and human disease. *Nat Rev Genet* 2013;14:307–20.
- [7] Poduri A, Evrony GD, Cai X, Walsh CA. Somatic mutation, genomic variation, and neurological disease. *Science* 2013;341:1237758.

- [8] Lawrence MS, Stojanov P, Mermel CH, Robinson JT, Garraway LA, Golub TR, et al. Discovery and saturation analysis of cancer genes across 21 tumour types. *Nature* 2014;505:49–501.
- [9] Mari F, Azimonti S, Bertani I, Bolognese F, Colombo E, Caselli R, et al. *CDKL5* belongs to the same molecular pathway of *MeCP2* and it is responsible for the early-onset seizure variant of Rett syndrome. *Hum Mol Genet* 2005;14:1935–46.
- [10] Stosser MB, Lindy AS, Butler E, Retterer K, Piccirillo-Stosser CM, Richard G, et al. High frequency of mosaic pathogenic variants in genes causing epilepsy-related neurodevelopmental disorders. *Genet Med* 2017;24:1–8.
- [11] Gripp KW, Stables DL, Nicholson L, Hoffman JD, Sol-Church K. Somatic mosaicism for an *HRAS* mutation causes Costello syndrome. *Am J Med Genet A* 2006;140:2163–9.
- [12] Freed D, Pevsner J. The contribution of mosaic variants to autism spectrum disorder. *PLoS Genet* 2016;12:e1006245.1–20.
- [13] Krupp DR, Barnard RA, Duffourd Y, Evans SA, Mulqueen RM, Bernier R, et al. Exonic mosaic mutations contribute risk for autism spectrum disorder. *Am J Hum Genet* 2017;101:369–90.
- [14] Gilissen C, Hehir-Kwa JY, Thung DT, van de Vorst M, van Bon BW, Willemsen MH, et al. Genome sequencing identifies major causes of severe intellectual disability. *Nature* 2014;511:344–7.
- [15] Tartaglia M, Cordeddu V, Chang H, Shaw A, Kalidas K, Crosby A, et al. Paternal germline origin and sex-ratio distortion in transmission of *PTPN11* mutations in Noonan syndrome. *Am J Hum Genet* 2004;75:492–7.
- [16] Tekin M, Cengiz FB, Ayberkin E, Kendirli T, Fitoz S, Tutar E, et al. Familial neonatal Marfan syndrome due to parental mosaicism of a missense mutation in the *FBN1* gene. *Am J Med Genet A* 2007;143A:875–80.
- [17] Xu X, Yang X, Wu Q, Liu A, Yang X, Ye AY, et al. Amplicon resequencing identified parental mosaicism for approximately 10% of “*de novo*” *SCN1A* mutations in children with Dravet syndrome. *Hum Mutat* 2015;36:861–72.
- [18] Dou Y, Yang X, Li Z, Wang S, Zhang Z, Ye AY, et al. Postzygotic single-nucleotide mosaicism contribute to the etiology of autism spectrum disorder and autistic traits and the origin of mutations. *Hum Mutat* 2017;38:1002–13.
- [19] Acuna-Hidalgo R, Bo T, Kwint MP, van de Vorst M, Pinelli M, Veltman JA, et al. Post-zygotic point mutations are an underrecognized source of *de novo* genomic variation. *Am J Hum Genet* 2015;97:67–74.
- [20] Yang X, Liu A, Xu X, Yang X, Zeng Q, Ye AY, et al. Genomic mosaicism in paternal sperm and multiple parental tissues in a Dravet syndrome cohort. *Sci Rep* 2017;7:15677.
- [21] de Lange IM, Koudijs MJ, van’t Slot R, Gunning B, Sonsma ACM, van Gemert L, et al. Mosaicism of *de novo* pathogenic *SCN1A* variants in epilepsy is a frequent phenomenon that correlates with variable phenotypes. *Epilepsia* 2018;59:690–703.
- [22] Huang AY, Xu X, Ye AY, Wu Q, Yan L, Zhao B, et al. Postzygotic single-nucleotide mosaicism in whole-genome sequences of clinically unremarkable individuals. *Cell Res* 2014;24:1311–27.
- [23] Vijg J, Dong X, Zhang L. A high-fidelity method for genomic sequencing of single somatic cells reveals a very high mutational burden. *Exp Biol Med (Maywood)* 2017;242:1318–24.
- [24] Forbes SA, Beare D, Boutselakis H, Bamford S, Bindal N, Tate J, et al. COSMIC: somatic cancer genetics at high-resolution. *Nucleic Acids Res* 2017;45:D777–83.
- [25] Bhattacharya A, Ziebarth JD, Cui Y. SomamiR: a database for somatic mutations impacting microRNA function in cancer. *Nucleic Acids Res* 2013;41:D977–82.
- [26] Wang K, Li M, Hakonarson H. ANNOVAR: functional annotation of genetic variants from high-throughput sequencing data. *Nucleic Acids Res* 2010;38:e164.
- [27] Sherry ST, Ward MH, Kholodov M, Baker J, Phan L, Smigielski EM, et al. dbSNP: the NCBI database of genetic variation. *Nucleic Acids Res* 2001;29:308–11.
- [28] Karczewski KJ, Francioli LC, Tiao G, Cummings BB, Alföldi J, Wang Q, et al. The mutational constraint spectrum quantified from variation in 141,456 humans. *Nature* 2020;581:434–43.
- [29] Kircher M, Witten DM, Jain P, O’Roak BJ, Cooper GM, Shendure J. A general framework for estimating the relative pathogenicity of human genetic variants. *Nat Genet* 2014;46:310–5.
- [30] Ionita-Laza I, McCallum K, Xu B, Buxbaum JD. A spectral approach integrating functional genomic annotations for coding and noncoding variants. *Nat Genet* 2016;48:214–20.
- [31] Shihab HA, Gough J, Cooper DN, Stenson PD, Barker GL, Edwards KJ, et al. Predicting the functional, molecular, and phenotypic consequences of amino acid substitutions using hidden Markov models. *Hum Mutat* 2013;34:57–65.
- [32] Ng PC, Henikoff S. SIFT: predicting amino acid changes that affect protein function. *Nucleic Acids Res* 2003;31:3812–4.
- [33] Wang M, Wei L. iFish: predicting the pathogenicity of human nonsynonymous variants using gene-specific/family-specific attributes and classifiers. *Sci Rep* 2016;6:31321.
- [34] Wang M, Tai C, Weinan E, Wei L. DeFine: deep convolutional neural networks accurately quantify intensities of transcription factor-DNA binding and facilitate evaluation of functional non-coding variants. *Nucleic Acids Res* 2018;46:e69.
- [35] Davydov EV, Goode DL, Sirota M, Cooper GM, Sidow A, Batzoglou S. Identifying a high fraction of the human genome to be under selective constraint using GERP++. *PLoS Comput Biol* 2010;6:e1001025.
- [36] Pollard KS, Hubisz MJ, Rosenbloom KR, Siepel A. Detection of nonneutral substitution rates on mammalian phylogenies. *Genome Res* 2010;20:110–21.
- [37] Reva B, Antipin Y, Sander C. Predicting the functional impact of protein mutations: application to cancer genomics. *Nucleic Acids Res* 2011;39:e118.
- [38] Wildeman M, van Ophuizen E, den Dunnen JT, Taschner PE. Improving sequence variant descriptions in mutation databases and literature using the Mutalyzer sequence variation nomenclature checker. *Hum Mutat* 2008;29:6–13.
- [39] Lee J, Lee AJ, Lee JK, Park J, Kwon Y, Park S, et al. Mutalisk: a web-based somatic MUTation AnaLysis toolKit for genomic, transcriptional and epigenomic signatures. *Nucleic Acids Res* 2018;46:W102–8.
- [40] Down TA, Piipari M, Hubbard TJ. Dalliance: interactive genome viewing on the web. *Bioinformatics* 2011;27:889–90.
- [41] Krzywinski M, Schein J, Birol I, Connors J, Gascoyne R, Horsman D, et al. Circos: an information aesthetic for comparative genomics. *Genome Res* 2009;19:1639–45.
- [42] Kibbe WA, Arze C, Felix V, Mittra E, Bolton E, Fu G, et al. Disease Ontology 2015 update: an expanded and updated database of human diseases for linking biomedical knowledge through disease data. *Nucleic Acids Res* 2015;43:D1071–8.
- [43] Schuster-Bockler B, Lehner B. Chromatin organization is a major influence on regional mutation rates in human cancer cells. *Nature* 2012;488:504–7.
- [44] Conrad DF, Keebler JE, DePristo MA, Lindsay SJ, Zhang Y, Casals F, et al. Variation in genome-wide mutation rates within and between human families. *Nat Genet* 2011;43:712–4.
- [45] 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.
- [46] Helleday T, Eshtad S, Nik-Zainal S. Mechanisms underlying mutational signatures in human cancers. *Nat Rev Genet* 2014;15:585–98.

- 
- [47] Yang X, Gao H, Zhang J, Xu X, Liu X, Wu X, et al. *ATP1A3* mutations and genotype-phenotype correlation of alternating hemiplegia of childhood in Chinese patients. *PLoS One* 2014;9:e97274.
- [48] Huang AY, Zhang Z, Ye AY, Dou Y, Yan L, Yang X, et al. MosaicHunter: accurate detection of postzygotic single-nucleotide mosaicism through next-generation sequencing of unpaired, trio, and paired samples. *Nucleic Acids Res* 2017;45:e76.
- [49] Cibulskis K, Lawrence MS, Carter SL, Sivachenko A, Jaffe D, Sougnez C, et al. Sensitive detection of somatic point mutations in impure and heterogeneous cancer samples. *Nat Biotechnol* 2013;31:213–9.
- [50] Kim S, Scheffler K, Halpern AL, Bekritsky MA, Noh E, Kallberg M, et al. Strelka2: fast and accurate calling of germline and somatic variants. *Nat Methods* 2018;15:591–4.