Science Advances
AAAS

advances.sciencemag.org/cgi/content/full/6/48/eaba4905/DC1

Supplementary Materials for

Germline genomic patterns are associated with cancer risk, oncogenic pathways, and clinical outcomes

Xue Xu, Yuan Zhou, Xiaowen Feng, Xiong Li, Mohammad Asad, Derek Li, Bo Liao*,
Jianqiang Li*, Qinghua Cui*, Edwin Wang*

*Corresponding author. Email: edwin.wang@ucalgary.ca (E.W.); cuiqinghua@hsc.pku.edu.cn (Q.C.); dragonbw@163.com (B.L.); lijq@szu.edu.cn (J.L.)

Published 27 November 2020, *Sci. Adv.* **6**, eaba4905 (2020)
DOI: 10.1126/sciadv.aba4905

The PDF file includes:

Supplementary Materials and Methods
Figs. S1 to S4
Tables S1 to S6
Legends for data S1 to S4

Other Supplementary Material for this manuscript includes the following:

(available at advances.sciencemag.org/cgi/content/full/6/48/eaba4905/DC1)

Data S1 to S4

Supplementary Materials

Supplementary Materials and Methods

The impact of germline variant calling methods on genomic pattern discovery

Variant calling method plays an important role in genomics. Here we examined two distinct variant calling setups, either using HaplotypeCaller or MuTect, to determine whether the genomic pattern discovery methodology was tolerant of this technical difference. We further tested two calling modes of HaplotypeCaller and found them comparable as for this study's purpose.

(1) MuTect v119 versus HaplotypeCaller. Using the same criteria for variant read depth and VAF described in the Materials and Methods, we examined VCF files of a set of a non-redundant set of 9,930 individuals provided by the TCGA legacy archive, where variants were called using MuTect (v119 as indicated in the file headers). Mutational catalogs generated from both datasets closely resembled each other (Supplementary Fig. S4B). Major discoveries derived from HaplotypeCaller's results reported in this study, including the CGGPs, the correlation between CGGP_E and smoking sensitivity, and the clinical outcomes between germline-variant-defined subgroups, were reproducible. These findings indicated that variant calling methods would not significantly affect the reported discoveries in this study.

(2) On HaplotypeCaller's two variant calling modes. HaplotypeCaller has two main setups: the Single Sample Mode, which processes each given sample separately, and the Jointed Mode. The Jointed Mode of the HaplotypeCaller provides more normalizations and normal-panel-based filters during variant calling. However, it requires unpractically large storage and computational resources for cohorts as large as the TCGA germline dataset (n=9,712) and non-cancer dataset (n=16,670) we used.

Alternatively, we examined if the Single Sample Mode and the Jointed Mode would generate similar mutational catalogs that would be the direct source input of germline genomic pattern discovery (see Materials and Methods for the definition of mutational catalogs). A total of 1,543 of 9,712 patients were randomly selected and processed using the HaplotypeCaller (GATK version 4.0.6.0; a different version was used due to un-fixed bugs in the implementation of parallel and data production in HaplotypeCaller 3.x's Jointed Mode) in either Jointed Mode or Single Sample Mode, generating two mutational catalogs. The two mutational catalogs closely resembled each other: For more than 98% of the patients, cosine similarities between their two mutational catalogs, defined as a vector of length 192, were higher than 0.995. All such cosine similarities were higher than 0.974 (Supplementary Fig. S4A). More importantly, our study focused on genomic patterns that represented enrichment of context-dependent variant sets, and mutational catalogs and germline genomic patterns were found more tolerant than the identification of individual variants. The genomic pattern discovery methodology showed its robustness for up to 30% disturbance of data points (see below, simulation tests). Based on these observations, we believed that Single Sample Mode and Jointed Mode could be considered equivalent with respect to our purpose.

Germline Variants of cancer patients: statistics of variant annotation

We annotated the germline variants obtained to provide an overview of the dataset, although variants were not discriminated based on their positional or functional annotations. Germline variants called using the HaplotypeCaller and then filtered with germline variant criteria (see Materials and Methods in the main text) were further annotated by the Variant Effect Predictor (VEP) from the Ensemble Tool (perl /path/to/bin/ensembl-vep/vep --offline --input_file /path/to/data/FILENAME -o /path/to/output/FILENAME.cadd --assembly GRCh38 --force_overwrite). VEP reported that 43.6% of all germline variants were missense mutations, whilst 54.5% were synonymous. 19.1% and 6.3% variants resided in 3' UTR regions and 5' UTR regions, respectively. Based on the annotation, we found 0.26% of the variants were stop-gaining mutations, 0.042% of the variants were stop-retaining, and 0.079% of the variants caused a loss of stop codon. Up to 0.27% of the variants were associated with splicing functions (0.14% were annotated to be splice acceptor variants, and 0.13% of the splice donor variants), while about 15.2% of the variants resided in splice regions. Note that none of these annotations affected our choice of germline variants (see Materials and Methods for germline variant criteria).

Filtering of somatic mutations

For the COSMIC signature analysis, to coordinate with the configuration of the previous study (*28*), the controlled somatic mutations called by VarScan2 were obtained from TCGA repository (current release at GDC, v12.0; https://gdc.cancer.gov/). No VAF criteria like that for the germline variants were set for tumor somatic variants. Instead, we only dropped any somatic variant that had read depth less than 20 or VAF less than 0.01 for better variant quality.

Variants were further filtered by VEP (Variant Effect Predictor) from the Ensemble Tool, using the command 'perl /path/to/bin/ensembl-vep/vep --offline --input_file /path/to/data/FILENAME -o /path/to/output/FILENAME.cadd --assembly GRCh38 --filter "DP > 19" --filter "AF>0.01" --force_overwrite'. Then variants flagged to have germline risks were dropped (i.e., flagged as germline_risk, panel_of_normals, alt_allele_in_normal in `FILTER`; only the following flags were accepted: 3_prime_UTR_variant, 5_prime_UTR_variant, NMD_transcript_variant, downstream_gene_variant, frameshift_variant, inframe_insertion, missense_variant, non_coding_transcript_exon_variant, non_coding_transcript_variant, splice_donor_variant, splice_region_variant, stop_gained, upstream_gene_variant).

It is also noteworthy that we directly used the recently updated MC3 somatic mutation dataset (v0.2.8, controlled version) without further filtration (*34*) when performing the oncogene analysis.

The robustness of the NMF method for the CGGP discovery

To test the robustness of the NMF method for the CGGP discovery, we conducted the followings:

(1) Simulation test A: random removal of signals. We down-sampled the variants from the VCF files of patient germline dataset to the point where 30% of data points were randomly removed:

(I) for each individual of patient germline dataset:

<i> let Ntarget be the target number of data points, and Ncurrent be the real total counts of profiles of the patient;

<ii> encode the patient's mutational profile to "reads": for example, profile of ACA>AAA, 100 counts will be conceived as 100 data points tagged as 'ACA>AAA'. In this way a virtual short read pool can be formed, where the number of "read s" will be Ncurrent;

<iii> draw Ntarget "reads" from the virtual pool, and decode the "reads" to called profiles.

(II) a down-sampled mutational catalog Viteration_i is formed;

(III) solve the Viteration_i.

The down-sampled catalogs and derived CGGPs were collected for 3,400 iterations. More than 80.2% of the resulted CGGPs closely resembled their real counterparts, measured by cosine similarities (threshold of 0.999).

(2) Simulation test B: manipulative noise. Random noise were introduced at the volume of roughly 30% of the total signal strength, modifying (i.e. substituting) the strength of affected signals. The process was implemented as below:

```
def bootstrapGenomes(X, seed=0, n=None):
    normX = X/np.sum(X, axis=0)
    bootstrap_all = []
    N = np.sum(X, axis=0)
    for i in range(X.shape[1]):
        tmp = normX[:, i]
        # 0.4 and 1.6 were chosen to achieve 30% signal strength difference
        tmp = tmp*np.random.uniform(0.4,1.6, len(tmp))
        tmp = tmp/np.sum(tmp)
        bootstrapX = np.random.multinomial(N[i], tmp)
        bootstrap_all.append(bootstrapX)
    bootstrap_all = np.asarray(bootstrap_all).T
        return bootstrap_all
```

5000 simulations were done. About 97.0% of the resulted CGGPs closely resembled their real counterparts, measured by cosine similarities (threshold of 0.999). The number was 85.7% when the threshold was set to 0.9995. The algorithm acting better on manipulative simulations was not a surprise. Alexendrov et al. reported that the NMF methodology would scale along with dataset size (*21*). Loss of signals in simulation 1 possible compromised pattern discoveries of NMF methodology slightly; in simulation 2, 30% of substituting random noises would not prevent NMF from finding the actual signals, therefore demonstrating the robustness.

(3) Germline genomic pattern discovery in the non-cancer dataset

Our algorithm performed robustly in the non-cancer dataset according to results from simulation tests (same procedures as described above). Germline patterns except CGGP_E were reproduced (cosine similarities were: 0.99, 0.98, 1.00, 0.97, 0.93 and 1.00, receptively, for CGGP_ A, B, C, D and F). Given that the non-cancer cohort mainly recruited individuals in a healthy state or with ails other than cancer, we think non-perfect reproduction is acceptable.

(4) Reproducibility of the CGGPs in the merged dataset of the non-cancer dataset and TCGA patient germline dataset. To further test the robustness of the NMF method, we applied it to the dataset which was the combination of the non-cancer dataset and patient germline dataset. All the 7 CGGPs were reproducible with cosine similarities of 1.00, 0.99, 1.00, 0.96, 1.00, 1.00, and 1.00, respectively. As it was suggested by the silhouette method for the germline dataset (see Materials and Methods), here k=7 were the best hyperparameter before overfitting.

In previous studies (*21, 23, 28*), the comparisons of somatic signatures were mainly examined for their strongest signal peaks or clusters of peaks, while a single peak alone was not of much interest. We considered the same rationale would hold for the germline genomic patterns. To capture both local and global features of genomic patterns, we pooled each germline pattern based on sequential contexts, rendering each (192, 1) vector to (24, 1) before calculating a cosine similarity. The cosine similarities reported in this study were very significant according to randomization tests. For each CGGP, we permutated their elements (a.k.a. the signal peaks) and calculated the similarity between the resulted vector and the original germline genomic pattern for 10,000 iterations. In randomization test, the average cosine similarity between permutations of each CGGP and their ground truth counterparts were 0.20, 0.090, 0.46, 0.25, 0.19, 0.09 and 0.35, respectively. Standard deviations were 0.058, 0.066, 0.039, 0.055, 0.058, 0.067 and 0.047, respectively. Compared to the randomization test results, cosine similarities between CGGPs reported in this study were significantly higher.

Examining the robustness of CGGPs by considering potential sequencing artifacts

Sequencing artifacts such as WES coverage depth, batches in library preparation and sequencing, and so on could become confounding factors for CGGPs. To rule out potential confounding factors, we tested the robustness the CGGPs using alternative sets of variants, samples, and conditions including:

(a) Removing variants in repeats and outlier samples that are likely to be associated with batch effects according to previous studies (named as the *DP20maksed conditio*n here): The genomic coordinates of repeats identified by RepeatMasker were from http://hgdownload.cse.ucsc.edu/goldenpath/hg38/database/rmsk.txt.gz, and the variants settled on the repeats regions were removed. Following the idea of Buckley et al. (25), the outlier samples for each self-reported ancestry in the principal component analysis (PCA) plots of common variants were deemed as the samples that were significantly influenced by batch effect and thus could be excluded. Firstly, the 1000 genome germline variants were downloaded from https://www.cog-genomics.org/plink/2.0/resources#1kg_phase3, and the variants with minor allele frequency > 0.01 were considered as common variants (filtration was done by vcftools). The PCA analysis of these common variants were performed by PLINK2 (version 20200314, with linkage disequilibrium flag of --indep-pairwise 50 10 0.2; using alternative r^2 thresholds like 0.1 or 0.5 produced fewer outliers). The eigenvectors of the first two principal components were extracted, and samples showing a 3-fold standard deviation from the center point of each ancestry sample set were considered as the outliers. As a result, 25 European ancestry samples and 3 Asian ancestry samples were excluded.

(b) Similar to the DP20masked condition, but further removing variants in low mappability regions. The low mappability regions were defined as genomic regions with the CRG-36 alignability score < 0.1. Because there were no pre-computed CRG-36 alignability scores for the hg38 genome, we re-calculated the scores using the GEM tool (version Linux-x86_64-core_2-20130406-045632) with following commands: gem-indexer -T 20 -c dna -i hg38 -o hg38_index --max-memory 'force-slow-algorithm'; gem-mappability -T 20 -I hg38_index.gem -l 36 -o hg38_k36; gem-2-wig -I hg38_index.gem -i hg38_k36.mappability -o hg38_k36; wigToBigWig hg38_k36.wig hg38_k36.sizes hg38_k36.bw. The germline variants overlapped with low mappability regions were detected by bwtool and discarded in this experiment condition.

(c) Similar to the DP20masked condition, but using an alternative read depth threshold. The alternative read depth thresholds (based on DP field annotation in VCF files) of 10, 15, and 25 were considered.

(d) Removing low-quality variants. Two variant quality filters based on VCF file information were considered: requiring QD>2 or requiring GQ>20.

(e) Considering subsets of samples from different sequencing centers, whole genome amplification protocol, or exome capture array platform. Following subsets of samples were considered: 1) those sequenced in Broad Institute (BI) sequencing center (6,018 samples); 2) those sequenced in other sequencing centers (3,694 samples (because the algorithm for identifying genomic patterns requires a large number of samples, we combined the samples from these centers); 3) those did not adopt whole genome amplification protocol before sequencing (8,512 samples); 4) those applied the most common ‘Custom V2 Exome Bait, 48 RXN X 16 tubes’ exome capture array (5,978 samples); 5) those did not apply this exome capture array (3,734 samples; we combined the samples from other platforms due to the same sample size requirement of the algorithm).

(f) Following the implication from Harris and Pritchard (*27*), splitting the dataset by population ancestry. 7,214, 898, 593, 1007 samples from European, African, Asian, and Other/Unknown ancestries were obtained, respectively. Because the NMF approach requires several thousands of samples to obtain stable results, we here only re-examined the samples from European ancestry.

(g) Using exon-defined strand-specific mutational profiles. The exon annotation of known genes was obtained from the UCSC genome browser. The strand attribution of a variant was in line with the strand annotations of the exon it overlapped with. Only the trinucleotide mutational profiles on the exons’ strand were considered in this experiment.

(h) Using the germline variants called by TCGA (Huang et al. (*16*)). All germline variants called by Huang et al. were included. To ensure the sample sets between two studies were comparable, only 9,521 samples which were also presented in the sample set of this study were considered.

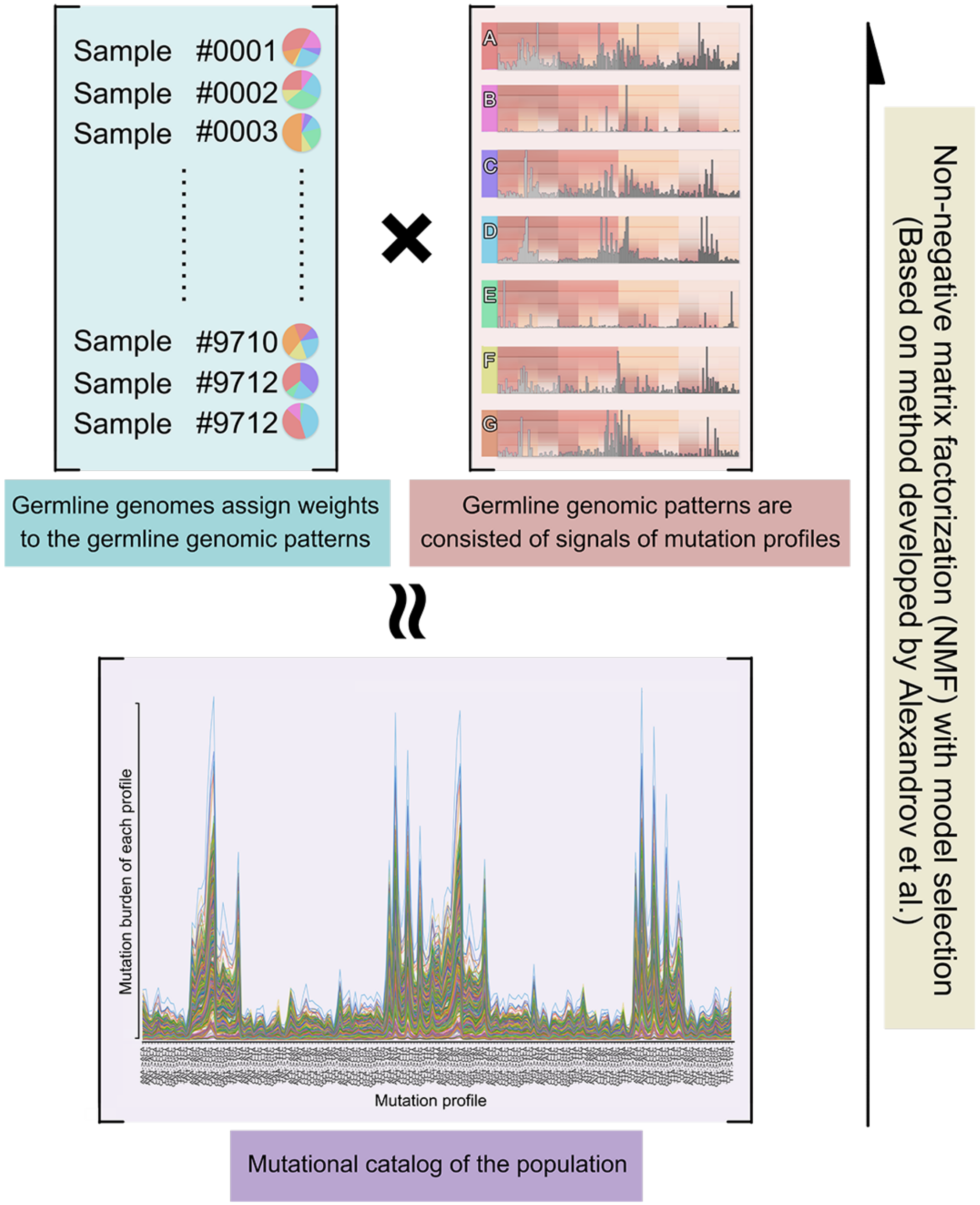
For each of the above experiment conditions, in addition to the cosine similarity (i.e., between a resulted CGGP and its corresponding-original CGGP) of each individual CGGP, two combined cosine similarities were also provided. The first approach was the collapsed cosine similarity, in which the resulted CGGPs and the corresponding-original CGGPs (i.e. the first matrix in Supplementary Data S1) were flattened into 192-dimensional vectors and then compared by calculating a cosine similarity of the two vectors. The second approach was the average cosine similarity, which simply took the average of the cosine similarities of individual CGGPs. Besides, to take care of sample variability and outliers, 1,000 bootstrap sample sets have been used to solve the CGGPs, and the resulted CGGPs were clustered by using affinity propagation algorithm in the scikit-learn Python package with default parameters, and the outliers distal to the cluster centers were removed (by requiring collapsed cosine similarity >0.98 to the cluster centers) before finding the best match to the original CGGPs. This step removed about 15% to 20% samples as the potential outliers. The cosine similarity with CGGPs obtained in the DP20masked condition was assessed in the same approach.

Assigning the CGGPs to a genome and converting CGGP's weighing factors to binary assignments

Weighing factors of a CGGP for an individual in their cohort (or an individual and a given set of CGGPs) can be obtained from the NMF methodology by solving the genomic patterns and converging (i.e., solving the NMF equation without updating pattern matrix), respectively. After normalization, each CGGP could be assigned to a number in the range 0 to 1, which represents an enrichment of the CGGP in the individual.

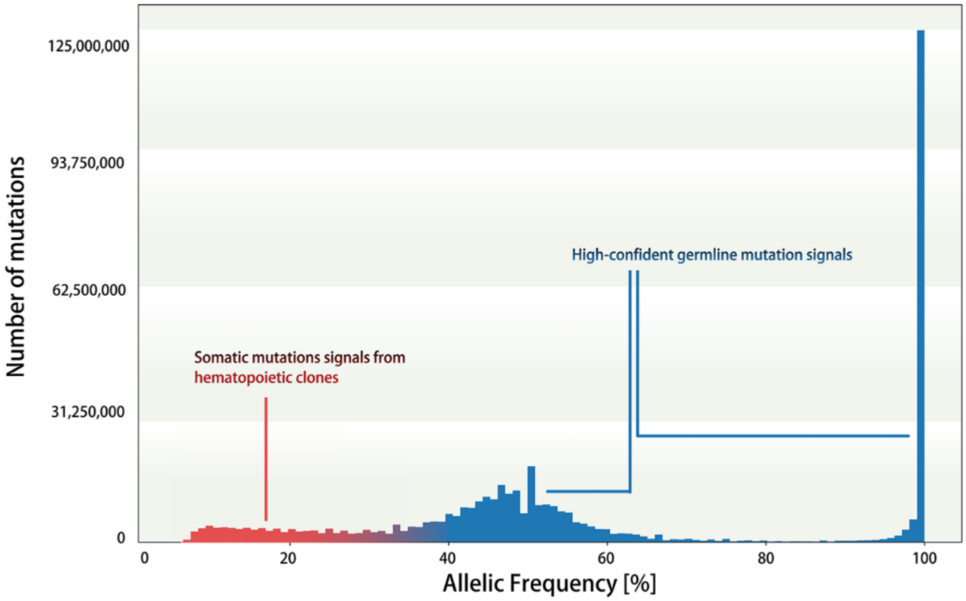
Alexandrov et al. (*22*) discovered 21 somatic mutational signatures and assigned each of them into each tumor by defining a weight threshold (by the criteria "more than 100 substitutions or more than 25% of all mutations in that sample"). Such a hard-coded threshold was not applicable for the CGGPs, because germline variants were due to not *de novo* mutations but inherited genetic heterogeneity, and the arbitrary threshold would barely make sense biologically in our case. Therefore, when performing the analyses based on the presence/absence partition of CGGPs, as described in the main text, we utilized the non-cancer cohort as a background to determine the weight thresholds for the CGGPs: for a given germline pattern X, a patient may carry (and potentially affected by) it if, and only if, the weighing factor of pattern X appears significantly higher than the weights given by non-cancer individuals; otherwise, we would conclude that the patient does possess a certain weight for pattern X, but not necessarily affected by it. The significance was modeled by the upper 95% confidence intervals of weights given by non-cancer individuals. By doing so, the numerical weights of the CGGPs were transformed into a binary format.

Supplementary Figures


Sample #0001
Sample #0002
Sample #0003
Sample #9710
Sample #9712
Sample #9712
A
B
C
D
E
F
G
Germline genomes assign weights to the germline genomic patterns
Germline genomic patterns are consisted of signals of mutation profiles
Mutation burden of each profile
Mutation profile
Mutational catalog of the population
Non-negative matrix factorization (NMF) with model selection
(Based on method developed by Alexandrov et al.)

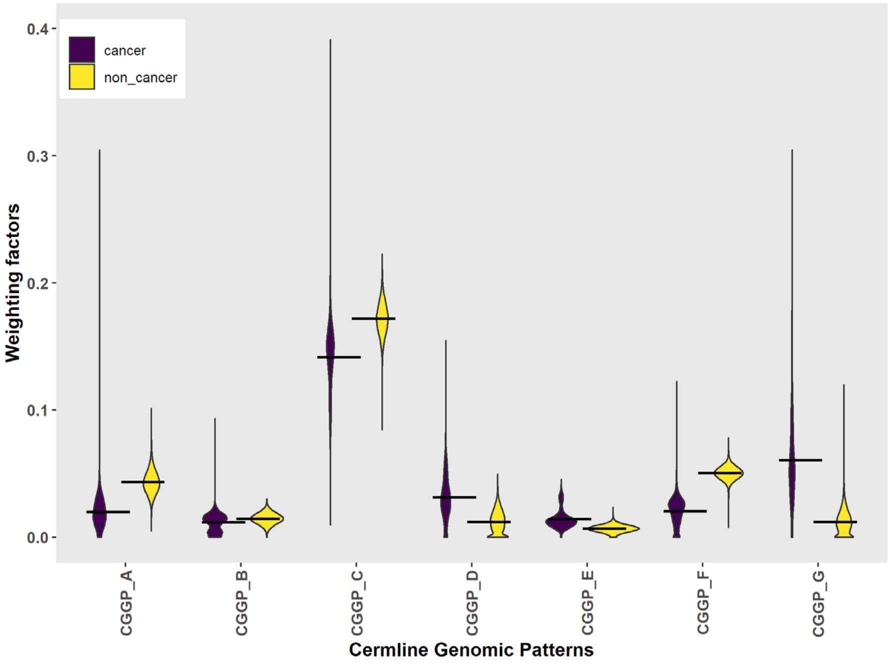
Supplementary Fig. S1. Deciphering of germline genomic patterns from germline mutational catalogs.

Germline variants were used to generate the germline mutational catalogs. A mutational catalog was of shape (192, number_of_samples), and further decomposed into pattern matrix and weight matrix of the shape (192, number_of_patterns) and (number_of_patterns, number_of_samples), respectively, by using the non-negative matrix factorization (NMF).



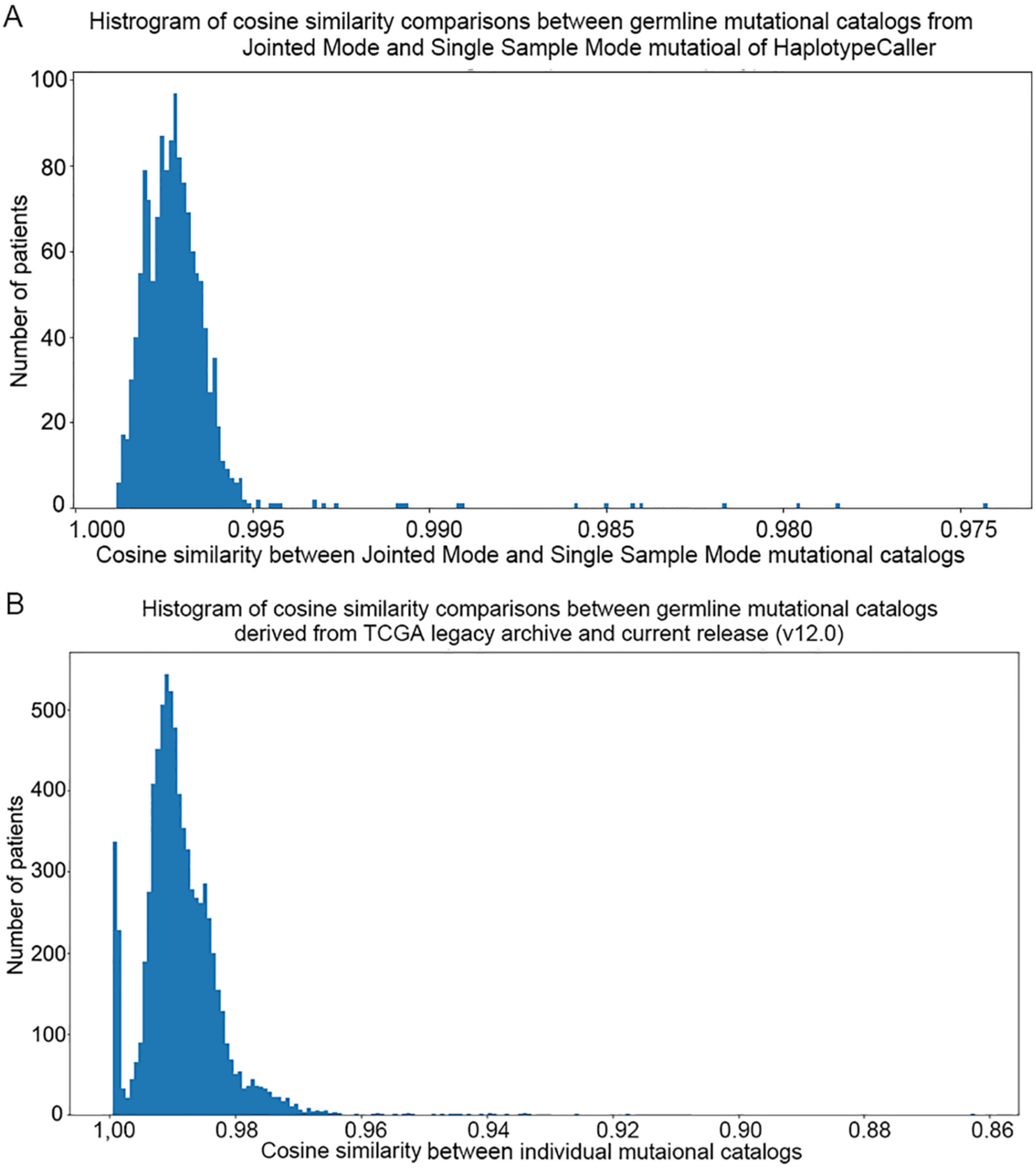
Supplementary Fig. S2. The distribution of variant allele frequencies in the cancer patients' germline genomes (n=9,712).

Expansion of hematopoietic stem cells (HSCs) could introduce somatic mutation contaminations in peripheral blood samples (i.e., buffy coats). The VAFs (variant allele frequencies) of the somatic mutations derived from HSCs were significantly departed from the main peaks around the theological VAF of germline variants (i.e. VAF = 0.5 or 1.0). Therefore, the variants which have such a VAF outside the intervals covered by these main peaks were considered as HSC' somatic mutation contaminations and therefore removed from the germline mutational catalogs.



Supplementary Fig. S3. Overall distribution of CGGP weighing factors between cancer and non-cancer samples.

Weighing factors assigned for each CGGP by either germline genomes of individuals from the TCGA dataset (n=9,712) or the non-cancer dataset (n=16,670) were illustrated as violin plots to show their distributions. Weighing factors from the non-cancer population were obtained through freezing the feature matrix (i.e. not updating values of CGGPs reported in this study) while resolving NMF. See Materials and Methods for details.



Histrogram of cosine similarity comparisons between germline mutational catalogs from Jointed Mode and Single Sample Mode mutatioal of HaplotypeCaller

Histogram of cosine similarity comparisons between germline mutational catalogs derived from TCGA legacy archive and current release (v12.0)

Supplementary Fig. S4. Comparison of mutational catalogs derived from different calling modes or different callers (represented by different TCGA releases). (A) Comparison of Single Sample Mode and Jointed Mode of HaplotypeCaller. Jointed Mode and Single Sample Mode of the HaplotypeCaller produced very similar mutational catalogs in randomly selected 1,543 patients. Limited by computational resources, here we examined only Chromosome 1. However, we believed that similar results should be obtained for other chromosomes since there was no significant and systemic bias between chromosomes in both exome-sequencing and variant calling. (B) Comparison of mutational catalogs derived from HaplotypeCaller and MuTect. Mutational catalogs were not strongly affected by different variant calling methods. Comparing patients' mutational catalogs derived from the BAM files of the TCGA current release (called by HaplotypeCaller) and the legacy archive dataset (called by MuTect v119, provided as-is), little difference was observed.

Supplementary Tables

Supplementary Table S1. The reproducibility of germline genomic patterns against the effects of potential sequencing artefacts

Experiment condition	Compared with[a]	Cosine similarity for particular CGGP							Sim_Collapsed[b]	Sim_Average[b]
		CGGP_A	CGGP_B	CGGP C	CGGP_D	CGGP_E	CGGP_F	CGGP_G		
Removing low quality variants (GQ<20)	Original	0.882	0.839	0.999	0.915	0.777	0.892	0.994	0.993	0.900
Removing low quality variants (QD<2)	Original	0.818	0.822	0.998	0.937	0.793	0.892	0.977	0.987	0.891
Removing repeats-related variants and outlier samples and keep other conditions as the same as those described for identifying genomic patterns in the main text (i.e. DP20masked condition, n = 9,684)	Original	0.903	0.803	0.989	0.893	0.896	0.894	0.984	0.961	0.909
Similar to the DP20masked condition but removing the variants in low mappability regions	Original	0.699	0.773	0.999	0.926	0.796	0.830	0.981	0.987	0.858
(Same as the above)	DP20masked	0.953	0.767	0.770	0.982	0.943	0.680	0.934	0.990	0.861
Similar to the DP20maksed condition, but with a read depth threshold of 10	Original	0.868	0.680	0.994	0.833	0.894	0.954	0.932	0.980	0.879
(Same as the above)	DP20masked	0.987	0.977	0.614	0.928	0.993	0.932	0.886	0.911	0.903
Similar to the DP20maksed condition, but with a read depth threshold of 15	Original	0.639	0.846	0.997	0.921	0.895	0.939	0.990	0.987	0.890
(Same as the above)	DP20masked	0.936	0.618	0.790	0.976	0.910	0.834	0.887	0.983	0.850
Similar to the DP20maksed condition, but with a read depth threshold of 25	Original	0.730	0.555	0.999	0.962	0.805	0.962	0.990	0.994	0.857
(Same as the above)	DP20masked	0.995	0.630	0.785	0.999	0.935	0.783	0.967	0.991	0.871
TCGA samples sequenced in the Broad Institute (BI) sequencing center, n=6,018	Original	0.708	0.812	0.999	0.901	0.821	0.868	0.975	0.989	0.869
TCGA samples sequenced in sequencing centers except the Broad Institute (BI), n = 3,694. The algorithm for identifying genomic patterns requires a large number of samples. Thus,	Original	0.784	0.635	0.988	0.886	0.833	0.841	0.955	0.964	0.846

we combined the samples from these centers										
A subset of randomly picked 3,694 samples from TCGA samples sequenced in Broad Institute (BI) [c]	Original	0.419	0.505	0.999	0.907	0.804	0.964	0.984	0.990	0.798
TCGA samples without whole genome amplification before sequencing, n = 8,512	Original	0.765	0.834	0.999	0.938	0.804	0.979	0.990	0.995	0.901
TCGA samples used Agilent ‘Custom V2 Exome Bait, 48 RXN X 16 tubes’ exome capture array before sequencing, n = 5,978	Original	0.755	0.819	0.999	0.899	0.814	0.927	0.980	0.990	0.885
TCGA samples used exome capture platforms excluding the Agilent ‘Custom V2 Exome Bait, 48 RXN X 16 tubes’ exome capture array before sequencing, n = 3,734	Original	0.592	0.770	0.996	0.918	0.700	0.573	0.946	0.974	0.785
A subset of randomly picked 3,734 samples from the set of Agilent ‘Custom V2 Exome Bait, 48 RXN X 16 tubes’ exome capture array before sequencing [c]	Original	0.522	0.493	0.998	0.838	0.870	0.832	0.908	0.984	0.780
Using exon-defined strand-specific mutational information	Original	0.580	0.753	0.992	0.947	0.794	0.829	0.952	0.989	0.835
Using germline variants from TCGA (Huang et al. 2018)	Original	0.772	0.763	0.990	0.611	0.817	0.848	0.942	0.947	0.820

[a] The CGGPs were compared either to the original CGGPs (denoted as ‘original’ in this column) or to the CGGPs resulted from the DP20masked condition (denoted as ‘DP20masked’ in this column).

[b] The collapsed cosine similarity was calculated by concatenating 7 CGGPs into one vector before calculating the similarity, while the average cosine similarity represents the average of cosine similarities of CGGP_A to _G.

[c] The condition was used to assess the influence of the sample size on solving CGGPs. The CGGPs from two conditions showed less similarities than the others: (1) TCGA samples sequenced in sequencing centers except the Broad Institute (BI), n = 3,694; and (2) TCGA samples used exome capture platforms excluding the Agilent ‘Custom V2 Exome Bait, 48 RXN X 16 tubes’ exome capture array before sequencing, n = 3,734. We noted that both conditions were performed with a relatively small sample size (n=~3,700). We were aware of that a larger sample size was necessary for obtaining stable genomic patterns. Here, to evaluate the influence of the sample size, we extended the same analysis by randomly picking 3,694 samples from TCGA samples sequenced in Broad Institute (BI) and 3,734 samples from the set of Agilent ‘Custom V2 Exome Bait, 48 RXN X 16 tubes' exome capture array before sequencing, respectively. Not surprisingly, the cosine similarities between the resulted genomic patterns and the original patterns were substantially reduced after the subsampling, indicating that sample sizes rather than batch effects affected the observed lower similarities mentioned above.

Supplementary Table S2. CGGPs impacted on somatic mutations of key oncogenes in tumors of European patients

CGGP	Gene symbol	P-value	Ratio[a]	FDR[b]
BLCA				
D	**AKAP9**	**0.01**	**1.20**	**0.06**
G	ARID1A	0.02	1.14	0.11
G	KMT2C	0.04	1.11	0.22
A	KMT2D	0.02	0.85	0.09
G	KMT2D	0.02	1.13	0.09
BRCA				
E	ARID1A	0.05	1.16	0.17
F	ARID1A	0.01	0.83	0.04
A	KMT2C	0.02	1.14	0.11
E	KMT2C	0.05	0.91	0.16
A	PTEN	1.24E-03	1.33	0.01
B	PTEN	0.04	0.84	0.12
B	TRRAP	0.01	0.85	0.08
B	ZFHX3	0.01	0.82	0.08
CESC				
B	ATRX	0.03	1.29	0.23
B	NF1	0.02	1.30	0.14
COAD				
D	**AKAP9**	**3.50E-03**	**1.21**	**0.02**
D	FAT1	0.01	1.21	0.04
D	**TRRAP**	**0.04**	**1.14**	**0.19**
GBM				
A	APC	0.02	0.78	0.15
G	**KMT2C**	**0.03**	**0.88**	**0.23**
HNSC				
B	**FAT1**	**0.03**	**0.88**	**0.14**
G	FAT1	0.04	0.90	0.14
G	PIK3CA	0.01	0.87	0.10
D	**TRRAP**	**0.02**	**1.15**	**0.14**
F	ZFHX3	0.03	0.88	0.22
KIRC				
A	ATM	1.21E-04	0.69	8.47E-04
G	ATM	0.02	1.18	0.07
B	**FAT1**	**0.03**	**1.28**	**0.19**
D	TP53	0.00	0.64	0.02
LGG				
D	APC	0.01	1.20	0.07
F	**APC**	**0.02**	**0.75**	**0.07**
B	**FAT1**	**0.03**	**1.28**	**0.19**
A	PIK3CA	0.02	1.17	0.14
LIHC				
B	ARID1A	0.05	1.32	0.16
C	ARID1A	0.03	1.25	0.16

C	ATM	0.03	1.21	0.24
D	KMT2C	0.04	1.17	0.22
A	**TP53**	**2.72E-03**	**0.77**	**0.02**
A	**TRRAP**	**0.03**	**1.21**	**0.22**
LUAD				
B	FAT4	0.04	1.16	0.13
C	FAT4	0.04	0.84	0.13
G	**KMT2C**	**0.01**	**0.88**	**0.08**
LUSC				
A	AKAP9	0.03	1.20	0.10
G	AKAP9	0.01	0.84	0.06
F	**APC**	**0.02**	**0.84**	**0.16**
B	GRIN2A	0.02	0.82	0.14
D	KMT2C	4.88E-03	0.84	0.03
B	KRAS	4.46E-03	0.65	0.03
D	KRAS	0.04	0.74	0.13
C	RNF213	0.03	1.21	0.19
C	TP53	0.01	1.36	0.10
B	TRRAP	0.04	1.16	0.25
OV				
A	AKAP9	0.04	0.85	0.15
F	AKAP9	0.02	1.15	0.15
G	FAT4	0.03	1.23	0.20
F	GRIN2A	0.01	0.83	0.10
E	KMT2D	3.73E-03	1.18	0.03
B	RNF213	0.01	1.16	0.05
G	RNF213	0.04	1.15	0.16
PAAD				
E	AKAP9	4.59E-03	1.25	0.03
B	APC	0.02	1.34	0.13
C	APC	0.04	1.25	0.13
B	**FAT1**	**0.02**	**0.63**	**0.10**
D	**FAT1**	**0.03**	**0.74**	**0.10**
E	KRAS	0.01	1.26	0.10
D	NF1	0.02	0.78	0.13
A	**TP53**	**0.01**	**0.81**	**0.07**
PCPG				
E	NF1	0.03	1.21	0.22
SKCM				
E	KMT2D	0.04	0.91	0.13
F	KMT2D	0.02	0.89	0.12
D	RNF213	0.03	1.12	0.18
STAD				
A	ATRX	0.04	0.87	0.14
D	ATRX	0.02	1.18	0.12
D	FAT4	0.03	1.14	0.22
A	PIK3CA	0.02	0.85	0.17
E	PTEN	0.01	0.79	0.05
B	TP53	0.05	0.86	0.16

E	TP53	0.03	0.87	0.16
E	TRRAP	0.03	0.85	0.21
THCA				
C	AKAP9	0.01	1.33	0.10
C	ATM	0.01	0.74	0.05
D	**FAT1**	**1.32E-03**	**0.63**	**0.01**
F	NF1	0.03	1.18	0.21
E	TRRAP	0.04	1.20	0.15
F	TRRAP	0.04	0.82	0.15
UCEC				
A	APC	0.01	1.22	0.06
A	KMT2D	0.01	1.22	0.04
A	**TRRAP**	**0.03**	**1.16**	**0.12**
G	TRRAP	0.01	0.87	0.07

[a] The ratio was calculated by comparing the mean of CGGP weights of the samples between a mutated gene group and the non-mutated gene group for a given gene.
[b] FDR-corrected p-values among each cancer type. The high confident associations whose empirical p-value<0.05 were highlighted in boldface.

Supplementary Table S3. Differential associations of CGGPs and their combinations between the germline genomes of cancer samples and those of non-cancer samples

CGGP	sample size of cancer patients	sample size of non-cancer patients	P-value	Left side of 95% CI	Right side of 95% CI	Odds ratio	FDR
BLCA							
A	325	3250	3.48E-89	6.25E-04	0.02	0.01	8.67E-89
B	325	3250	1.98E-05	1.50	3.18	2.17	2.09E-05
C	325	3250	1.79E-72	0.01	0.05	0.02	3.18E-72
D	325	3250	1.76E-74	23.80	127.22	50.22	3.29E-74
E	325	3250	6.37E-23	3.69	8.17	5.43	7.00E-23
F	325	3250	1.82E-105	1.25E-03	0.02	0.01	6.16E-105
G	325	3250	3.83E-66	10.27	24.63	15.60	6.05E-66
A_G	325	3250	6.35E-08	0.01	0.22	0.05	1.14E-07
B_C	325	3250	3.95E-10	0.02	0.21	0.08	7.68E-10
B_E	325	3250	2.55E-09	5.51	201.07	22.73	4.78E-09
B_F	325	3250	1.39E-12	0.01	0.16	0.05	2.86E-12
C_E	325	3250	1.46E-05	0.03	0.40	0.13	2.41E-05
D_E	325	3250	1.63E-53	49.49	1.06E+04	285.98	5.89E-53
E_G	325	3250	1.12E-44	20.42	107.98	43.83	3.70E-44
BRCA							
A	677	6770	1.06E-220	0.00	0.01	0.00	2.97E-219
B	677	6770	3.91E-225	0.00	0.01	0.00	1.46E-223
C	677	6770	2.17E-217	2.12E-03	0.01	0.01	4.86E-216
D	677	6770	2.45E-68	5.29	8.87	6.82	3.92E-68
E	677	6770	1.35E-198	76.37	551.92	179.86	2.51E-197

G	677	6770	7.48E-233	187.84	4.50E+15	1.05E+03	4.19E-231
C_D	677	6770	9.57E-11	0.04	0.25	0.11	1.87E-10
D_E	677	6770	2.68E-157	138.91	4.82E+03	512.61	1.00E-155
COAD							
A	213	2130	4.84E-50	0.01	0.06	0.03	6.02E-50
B	213	2130	4.15E-71	8.47E-05	0.02	3.09E-03	7.15E-71
D	213	2130	4.69E-61	27.67	273.47	73.09	6.56E-61
E	213	2130	9.97E-55	21.86	165.02	52.64	1.28E-54
A_B	213	2130	2.16E-57	6.10E-05	0.01	1.70E-03	8.14E-57
D_E	213	2130	4.49E-62	61.61	2.05E+03	237.43	2.01E-61
GBM							
A	337	3370	1.53E-86	0.01	0.04	0.02	3.64E-86
C	337	3370	6.06E-76	0.01	0.05	0.02	1.21E-75
D	337	3370	1.17E-25	3.41	6.60	4.71	1.31E-25
E	337	3370	4.04E-76	14.82	43.03	24.46	8.22E-76
B_C	337	3370	1.77E-13	4.44E-04	0.11	0.02	3.67E-13
B_D	337	3370	4.16E-03	1.25	3.76	2.16	6.32E-03
B_E	337	3370	2.58E-12	5.11	33.12	12.06	5.21E-12
C_D	337	3370	2.89E-06	1.04E-03	0.27	0.04	4.92E-06
D_E	337	3370	1.98E-57	16.13	50.79	27.80	7.58E-57
HNSC							
C	450	4500	3.53E-106	4.54E-03	0.03	0.01	1.27E-105
G	450	4500	2.20E-138	54.73	398.96	129.57	1.76E-137
B_C	450	4500	2.41E-24	2.13E-03	0.07	0.02	6.05E-24
B_G	450	4500	9.45E-30	22.45	5.10E+03	130.55	2.58E-29
KIRC							
A	278	2780	1.26E-85	2.65E-03	0.03	0.01	2.83E-85
B	278	2780	1.38E-14	0.13	0.33	0.21	1.50E-14
D	278	2780	6.34E-25	4.72	12.32	7.47	7.03E-25
KIRP							
B	205	2050	4.34E-26	0.06	0.18	0.11	4.91E-26
C	205	2050	3.76E-44	0.02	0.08	0.04	4.53E-44
E	205	2050	2.23E-61	37.84	1.18E+03	140.04	3.21E-61
LGG							
B	472	4720	6.11E-06	1.46	2.70	1.98	6.52E-06
C	472	4720	4.31E-116	2.14E-03	0.02	0.01	1.79E-115
E	472	4720	5.52E-91	19.43	64.78	33.94	1.44E-90
G	472	4720	7.85E-121	23.42	68.96	38.83	3.82E-120
B_C	472	4720	1.32E-17	0.01	0.12	0.04	2.89E-17
B_G	472	4720	1.23E-38	25.02	788.72	94.12	3.82E-38
E_G	472	4720	6.19E-111	103.93	1.60E+03	320.00	1.04E-109
LIHC							
A	186	1860	6.45E-61	9.77E-05	0.02	3.62E-03	8.92E-61
C	186	1860	7.82E-55	9.60E-04	0.03	0.01	1.02E-54
E	186	1860	2.55E-61	48.83	1.03E+04	278.54	3.62E-61
F	186	1860	6.16E-63	9.40E-05	0.02	3.47E-03	8.97E-63
G	186	1860	5.57E-58	36.20	1.16E+03	134.23	7.34E-58
E_G	186	1860	5.40E-61	110.17	1.64E+04	664.56	2.24E-60
LUAD							
B	390	3900	2.35E-04	1.31	2.53	1.81	2.46E-04
C	390	3900	2.13E-69	0.02	0.06	0.03	3.56E-69

D	390	3900	1.04E-105	40.95	301.29	97.14	3.63E-105
E	390	3900	1.04E-96	41.26	397.11	107.38	3.08E-96
F	390	3900	5.95E-134	6.10E-05	0.01	1.66E-03	4.44E-133
G	390	3900	1.93E-86	16.39	47.33	26.96	4.51E-86
B_C	390	3900	3.25E-14	4.43E-04	0.11	0.02	6.83E-14
B_D	390	3900	4.18E-35	29.45	6.63E+03	171.62	1.20E-34
B_G	390	3900	7.02E-26	13.59	210.69	42.05	1.80E-25
C_D	390	3900	1.04E-03	1.89	45.08	7.70	1.62E-03
LUSC							
B	337	3370	3.35E-04	1.32	2.72	1.89	3.48E-04
C	337	3370	2.07E-65	0.02	0.06	0.03	3.23E-65
D	337	3370	1.83E-76	22.62	107.44	45.68	3.86E-76
E	337	3370	4.45E-81	25.99	138.98	54.80	9.78E-81
F	337	3370	4.44E-115	6.10E-05	0.01	1.93E-03	1.77E-114
G	337	3370	9.20E-81	18.37	61.90	32.28	1.98E-80
B_C	337	3370	2.76E-09	0.04	0.27	0.11	5.15E-09
B_F	337	3370	3.27E-16	3.68E-04	0.09	0.01	7.03E-16
E_G	337	3370	1.69E-84	77.22	1.18E+03	238.37	1.24E-83
PAAD							
A	161	1610	7.36E-50	1.19E-04	0.03	4.48E-03	9.06E-50
C	161	1610	5.31E-41	1.49E-04	0.03	0.01	6.33E-41
D	161	1610	2.06E-44	33.67	7.36E+03	193.06	2.50E-44
E	161	1610	1.57E-27	10.81	84.22	26.47	1.80E-27
F	161	1610	7.40E-54	1.08E-04	0.02	4.06E-03	9.41E-54
G	161	1610	9.55E-35	10.09	42.35	19.54	1.13E-34
B_D	161	1610	2.67E-11	8.14	2.08E+03	50.62	5.31E-11
E_G	161	1610	8.67E-30	28.05	996.06	113.35	2.39E-29
SKCM							
A	446	4460	1.37E-89	0.03	0.07	0.05	3.48E-89
B	446	4460	4.30E-29	3.84	7.70	5.39	4.97E-29
C	446	4460	1.83E-88	0.02	0.05	0.03	4.46E-88
D	446	4460	5.16E-60	6.52	12.65	9.00	7.05E-60
E	446	4460	8.51E-30	3.69	7.08	5.07	9.92E-30
G	446	4460	6.10E-96	14.53	35.46	22.21	1.71E-95
A_B	446	4460	1.08E-03	0.18	0.68	0.36	1.68E-03
B_C	446	4460	7.10E-08	0.04	0.31	0.12	1.27E-07
C_E	446	4460	1.38E-06	0.08	0.42	0.20	2.39E-06
D_E	446	4460	4.52E-49	9.24	22.44	14.19	1.52E-48
E_G	446	4460	7.88E-63	34.22	268.08	84.15	3.58E-62
STAD							
A	277	2770	3.73E-86	7.32E-05	0.01	2.59E-03	8.53E-86
C	277	2770	1.42E-74	7.25E-04	0.02	0.01	2.70E-74
D	277	2770	1.42E-74	44.52	1.38E+03	163.92	2.70E-74
E	277	2770	1.33E-59	19.19	103.20	40.70	1.77E-59
G	277	2770	3.60E-76	23.65	112.86	47.88	7.48E-76
A_G	277	2770	3.39E-03	1.76E-03	0.60	0.08	5.20E-03
B_C	277	2770	1.33E-19	2.84E-04	0.07	0.01	2.98E-19
D_E	277	2770	9.28E-68	75.33	1.55E+04	436.44	4.95E-67
E_G	277	2770	7.14E-70	54.07	451.58	137.25	4.07E-69
THCA							
B	327	3270	6.19E-09	1.99	4.43	2.95	6.67E-09
C	327	3270	3.43E-64	0.01	0.05	0.03	5.13E-64

D	327	3270	2.52E-93	71.36	1.43E+04	398.59	6.73E-93
E	327	3270	7.54E-60	12.03	36.23	20.18	1.02E-59
F	327	3270	7.24E-101	3.33E-03	0.02	0.01	2.39E-100
G	327	3270	3.81E-64	10.15	24.75	15.54	5.62E-64
B_C	327	3270	3.55E-06	0.03	0.35	0.12	6.03E-06
B_F	327	3270	7.85E-10	0.01	0.20	0.06	1.50E-09
B_G	327	3270	2.37E-27	17.09	554.04	65.28	6.28E-27
C_D	327	3270	3.53E-03	1.75	705.05	15.01	5.39E-03
D_E	327	3270	2.16E-77	88.26	1.64E+04	508.64	1.42E-76
E_G	327	3270	4.51E-64	28.27	118.54	55.31	2.20E-63
UCEC							
D	357	3570	2.34E-50	6.39	13.29	9.12	2.94E-50

Note: Chi-square test (FDR < 0.25) was used to test if cancer and non-cancer samples could be distinguished by CGGPs or CGGP combinations.

Supplementary Table S4. CGGPs and their combinations in the germline genomes associated with the enrichment of COSMIC somatic signatures in tumors

CGGP	COSMIC	Odds ratio	P-value	95% CI	FDR
A	22	0.52	3.89E-09	0.41-0.65	1.83E-07
A	1	1.78	2.98E-07	1.42-2.23	7.74E-06
A	27	1.46	4.99E-04	1.18-1.82	8.28E-03
A_B	10	0.21	1.73E-05	0.09-0.46	1.63E-04
A_B	21	3.21	5.75E-04	1.58-6.81	3.69E-03
A_B	1	3.21	6.87E-04	1.56-6.82	4.15E-03
A_D	12	0.08	8.15E-13	0.04-0.18	4.62E-11
A_D	13	0.11	1.94E-10	0.05-0.25	6.58E-09
A_D	10	0.12	4.88E-08	0.04-0.29	1.04E-06
A_D	20	0.22	1.24E-05	0.10-0.47	1.24E-04
A_D	25	0.31	4.60E-04	0.15-0.62	3.17E-03
A_D	9	2.86	6.95E-04	1.49-5.55	4.15E-03
A_E	9	4.74	1.01E-04	2.05-11.19	8.62E-04
A_E	24	0.24	2.59E-04	0.10-0.54	1.95E-03
A_F	12	0.13	1.41E-10	0.06-0.26	5.32E-09
A_F	13	0.24	1.33E-06	0.13-0.45	1.83E-05
A_F	25	0.31	3.88E-05	0.17-0.57	3.47E-04
A_F	4	0.33	1.41E-04	0.18-0.60	1.11E-03
A_G	22	0.31	5.86E-10	0.21-0.46	1.66E-08
A_G	12	0.40	1.35E-06	0.27-0.59	1.83E-05
A_G	18	0.43	4.85E-06	0.29-0.63	6.11E-05
A_G	15	0.44	5.88E-06	0.30-0.64	6.66E-05
B_C	22	9.81	5.53E-06	3.18-34.05	6.48E-05
B_C	9	0.12	1.20E-05	0.04-0.35	1.23E-04
B_C	10	9.43	3.20E-05	2.77-42.12	2.94E-04
B_C	12	6.81	8.53E-05	2.30-23.38	7.43E-04

B_C	27	0.17	3.35E-04	0.05-0.50	2.47E-03
B_D	22	15.09	3.99E-12	6.27-39.22	1.69E-10
B_F	22	6.11	3.94E-17	3.87-9.73	2.68E-15
C_D	1	0.60	5.39E-04	0.45-0.81	3.52E-03
C_E	9	5.54	1.78E-33	4.11-7.51	6.05E-31
C_E	17	3.58	5.53E-19	2.67-4.83	6.27E-17
C_E	13	2.51	2.35E-10	1.87-3.38	7.27E-09
C_E	23	2.08	1.65E-07	1.57-2.76	2.99E-06
C_E	28	2.05	3.67E-07	1.54-2.73	5.94E-06
C_E	4	1.91	5.43E-06	1.44-2.54	6.48E-05
C_E	27	1.84	1.08E-05	1.39-2.44	1.18E-04
C_E	25	1.66	4.50E-04	1.24-2.22	3.17E-03
C_F	22	3.80	6.14E-09	2.35-6.22	1.61E-07
C_F	13	0.42	1.13E-04	0.26-0.67	9.34E-04
C_F	15	2.22	6.68E-04	1.37-3.61	4.13E-03
C_F	23	2.00	1.68E-03	1.28-3.14	9.51E-03
C_G	22	0.47	2.12E-08	0.36-0.62	4.81E-07
D	22	1.94	5.27E-09	1.55-2.44	1.83E-07
D	17	1.49	2.84E-04	1.20-1.87	5.90E-03
D	12	0.67	5.58E-04	0.54-0.85	8.28E-03
D_E	10	5.19	3.63E-18	3.49-7.77	3.09E-16
D_E	9	3.49	3.35E-12	2.40-5.11	1.63E-10
D_E	17	2.64	5.49E-08	1.83-3.81	1.10E-06
D_E	23	2.46	3.43E-07	1.72-3.55	5.83E-06
D_E	18	2.01	1.17E-04	1.39-2.92	9.47E-04
D_E	28	1.88	4.26E-04	1.31-2.71	3.09E-03
D_F	22	5.06	3.95E-24	3.62-7.11	6.71E-22
D_F	15	2.18	7.31E-07	1.59-3.01	1.08E-05
D_F	18	1.72	7.19E-04	1.25-2.38	4.21E-03
D_G	1	0.50	6.56E-04	0.33-0.76	4.13E-03
E	21	0.46	9.14E-12	0.37-0.58	9.50E-10
E_F	22	4.64	1.42E-08	2.61-8.40	3.45E-07
E_F	18	2.74	4.86E-04	1.52-5.00	3.24E-03
E_G	23	2.32	3.88E-07	1.65-3.27	5.99E-06
E_G	27	2.16	2.82E-06	1.55-3.02	3.69E-05
E_G	28	2.09	1.19E-05	1.49-2.95	1.23E-04
E_G	4	1.86	2.40E-04	1.32-2.63	1.86E-03
E_G	11	1.78	4.66E-04	1.27-2.49	3.17E-03
E_G	25	1.74	1.34E-03	1.23-2.47	7.70E-03
F	1	1.47	6.63E-04	1.17-1.84	8.62E-03
F_G	22	8.26	1.67E-07	3.33-22.81	2.99E-06
F_G	13	0.18	1.28E-05	0.07-0.43	1.24E-04

Note: Chi-square test (FDR < 0.25) was used to test the associations between COSMIC somatic mutational signatures and CGGPs (or CGGP combinations) in the germline genomes.

Supplementary Table S5. CGGPs and their combinations in the germline genomes associated with different cancer (sub)types

CGGP	Cancer type_1	Sample size_1	Cancer type_2	Sample size_2	P-value	95% CI	Odds ratio	FDR
A	LIHC	186	LUAD	390	1.81E-03	1.37-4.46	2.45	0.01
A	LIHC	186	BRCA	677	4.96E-03	1.24-3.76	2.14	0.03
A	LUAD	390	SKCM	446	0.01	0.40-0.91	0.61	0.10
A_B	GBM	337	BLCA	325	1.08E-04	0.04-0.42	0.14	2.27E-03
A_B	GBM	337	THCA	327	2.57E-04	0.06-0.47	0.17	5.40E-03
A_B	GBM	337	LUAD	390	9.58E-04	0.06-0.54	0.18	0.01
A_B	BLCA	325	LGG	472	1.07E-03	1.74-16.24	5.11	0.02
A_B	THCA	327	LGG	472	1.17E-03	1.72-15.62	4.99	0.02
A_B	LUAD	390	HNSC	450	2.73E-03	1.53-13.43	4.37	0.03
A_B	UCEC	357	THCA	327	3.83E-03	0.09-0.69	0.26	0.08
A_B	LUAD	390	UCEC	357	0.01	1.27-10.15	3.52	0.09
A_B	BLCA	325	UCEC	357	4.68E-03	1.39-12.16	4.01	0.10
A_C	TGCT	119	LGG	472	6.69E-05	3.49-1329.44	28.34	1.41E-03
A_C	BLCA	325	TGCT	119	3.48E-04	8.34E-04-0.35	0.04	7.31E-03
A_C	TGCT	119	LIHC	186	4.23E-04	3.33-487.15	29.20	8.88E-03
A_C	TGCT	119	STAD	277	7.25E-04	2.41-1017.88	20.88	0.02
A_C	PAAD	161	LIHC	186	2.76E-03	2.11-1425.14	24.46	0.06
A_D	COAD	213	KICH	58	2.70E-03	2.22-1314.68	23.80	0.06
A_D	GBM	337	COAD	213	6.05E-03	0.02-0.67	0.15	0.06
A_D	BRCA	677	COAD	213	4.01E-03	0.01-0.62	0.13	0.08
A_D	OV	348	COAD	213	6.70E-03	0.02-0.69	0.15	0.10
A_E	GBM	337	BRCA	677	6.81E-05	1.99-10.64	4.44	1.43E-03
A_E	LUSC	337	BRCA	677	2.18E-04	1.78-8.48	3.80	4.57E-03
A_E	GBM	337	UCEC	357	3.49E-04	1.77-9.98	4.12	7.32E-03
A_E	LIHC	186	UCEC	357	3.72E-04	2.02-21.91	6.20	7.80E-03
A_E	LIHC	186	LUAD	390	3.92E-04	2.21-31.19	7.58	8.23E-03
A_E	TGCT	119	UCEC	357	7.44E-04	2.44-856.29	18.80	9.41E-03
A_E	GBM	337	LUAD	390	1.02E-03	1.76-13.03	4.66	0.01
A_E	LIHC	186	BRCA	677	7.92E-04	1.78-19.44	5.33	0.02
A_E	TGCT	119	BRCA	677	1.67E-03	1.90-616.31	13.91	0.02
A_E	OV	348	LUAD	390	1.08E-03	1.68-10.87	4.17	0.02
A_E	LUSC	337	LUAD	390	9.83E-04	1.74-12.77	4.61	0.02
A_E	TGCT	119	LUAD	390	1.08E-03	2.18-816.11	17.47	0.02
A_E	TGCT	119	LGG	472	2.40E-03	1.83-602.06	13.47	0.03
A_E	GBM	337	THCA	327	3.93E-03	1.40-8.99	3.48	0.03
A_E	TGCT	119	THCA	327	1.41E-03	2.03-737.28	15.96	0.03
A_E	LUSC	337	UCEC	357	1.47E-03	1.53-8.17	3.48	0.03
A_E	OV	348	BRCA	677	3.74E-03	1.33-5.39	2.65	0.04
A_E	LUAD	390	STAD	277	3.81E-03	0.09-0.69	0.26	0.04
A_E	OV	348	UCEC	357	3.83E-03	1.36-6.72	2.99	0.04
A_E	TGCT	119	HNSC	450	4.87E-03	1.55-529.14	11.72	0.05
A_E	TGCT	119	KIRC	278	3.75E-03	1.69-612.98	13.27	0.05
A_E	GBM	337	HNSC	450	2.72E-03	1.41-8.18	3.31	0.06
A_E	GBM	337	COAD	213	3.74E-03	1.42-10.83	3.83	0.06
A_E	BRCA	677	STAD	277	3.54E-03	0.14-0.73	0.33	0.07

A_E	TGCT	119	COAD	213	4.75E-03	1.72-706.32	14.65	0.10
A_F	LUSC	337	HNSC	450	1.51E-03	1.55-9.84	3.85	0.03
A_F	LUSC	337	UCEC	357	5.42E-03	1.33-8.89	3.38	0.04
A_F	OV	348	PCPG	147	6.34E-03	0.03-0.65	0.16	0.07
B_C	LUAD	390	SKCM	446	1.72E-03	0.08-0.61	0.22	0.04
B_C	LUAD	390	STAD	277	9.15E-03	0.05-0.75	0.20	0.06
B_D	OV	348	THCA	327	3.20E-05	0.04-0.38	0.13	6.72E-04
B_D	GBM	337	THCA	327	1.86E-03	0.06-0.61	0.20	0.02
B_D	LUSC	337	THCA	327	9.77E-04	0.06-0.56	0.19	0.02
B_D	HNSC	450	THCA	327	2.18E-03	0.09-0.64	0.24	0.05
B_D	LUAD	390	THCA	327	2.18E-03	0.05-0.60	0.18	0.05
B_D	OV	348	UCEC	357	9.96E-03	0.07-0.76	0.24	0.07
B_D	OV	348	LGG	472	0.01	0.10-0.83	0.30	0.09
B_F	PRAD	413	LAML	120	6.56E-04	0.02-0.46	0.10	0.01
B_F	SKCM	446	LAML	120	1.11E-03	1.40E-03-0.47	0.06	0.02
B_F	BRCA	677	PCPG	147	1.62E-03	0.05-0.59	0.19	0.03
B_F	HNSC	450	PCPG	147	2.36E-03	0.04-0.61	0.17	0.04
B_F	CESC	205	LAML	120	3.73E-03	0.02-0.61	0.12	0.08
B_F	LAML	120	HNSC	450	4.35E-03	1.63-36.61	6.95	0.09
B_F	GBM	337	PCPG	147	0.01	0.06-0.83	0.24	0.10
B_G	OV	348	LUAD	390	1.70E-03	1.57-10.19	3.87	0.02
B_G	OV	348	PRAD	413	1.79E-03	1.53-9.53	3.71	0.02
B_G	OV	348	BLCA	325	3.53E-03	1.48-9.66	3.70	0.04
B_G	OV	348	LUSC	337	4.03E-03	1.39-9.25	3.49	0.04
B_G	GBM	337	BRCA	677	5.43E-03	1.29-5.42	2.61	0.05
B_G	OV	348	BRCA	677	0.01	1.19-5.44	2.51	0.07
B_G	OV	348	SKCM	446	7.28E-03	1.28-7.48	3.02	0.08
C_D	LUSC	337	UCEC	357	0.01	1.23-9.63	3.34	0.07
C_D	OV	348	LUAD	390	0.01	0.13-0.84	0.34	0.08
C_D	LUSC	337	SKCM	446	3.69E-03	1.38-7.83	3.23	0.08
C_D	LUAD	390	SKCM	446	8.46E-03	1.24-6.56	2.82	0.09
C_E	LUAD	390	HNSC	450	4.00E-03	0.11-0.72	0.29	0.03
C_E	LUSC	337	HNSC	450	4.23E-03	0.09-0.72	0.27	0.04
C_E	LUAD	390	UCEC	357	0.02	0.12-0.88	0.34	0.09
C_E	OV	348	LUSC	337	0.01	1.22-10.61	3.45	0.09
C_G	OV	348	SKCM	446	1.36E-11	0.03-0.18	0.07	2.86E-10
C_G	OV	348	BLCA	325	1.73E-08	0.03-0.24	0.09	3.63E-07
C_G	OV	348	BRCA	677	8.83E-08	0.06-0.32	0.14	1.86E-06
C_G	OV	348	UCEC	357	1.51E-07	0.02-0.23	0.08	3.16E-06
C_G	OV	348	LGG	472	2.14E-07	0.06-0.33	0.15	4.50E-06
C_G	OV	348	PCPG	147	2.06E-06	8.26E-03-0.23	0.05	4.33E-05
C_G	OV	348	STAD	277	3.53E-06	0.02-0.27	0.08	7.42E-05
C_G	OV	348	PRAD	413	4.80E-06	0.06-0.38	0.16	1.01E-04
C_G	OV	348	PAAD	161	9.47E-06	5.49E-04-0.21	0.03	1.99E-04
C_G	OV	348	KIRC	278	1.18E-05	0.04-0.35	0.12	2.48E-04
C_G	OV	348	CESC	205	4.25E-05	0.01-0.31	0.07	8.93E-04
C_G	OV	348	LUSC	337	5.11E-05	0.06-0.44	0.17	1.07E-03
C_G	OV	348	THCA	327	1.31E-03	0.09-0.62	0.24	0.01
C_G	GBM	337	SKCM	446	6.73E-04	0.10-0.59	0.24	0.01

C_G	OV	348	HNSC	450	6.85E-04	0.08-0.55	0.21	0.01
C_G	GBM	337	BLCA	325	1.72E-03	0.09-0.64	0.25	0.02
C_G	GBM	337	UCEC	357	3.35E-03	0.08-0.66	0.24	0.02
C_G	OV	348	LIHC	186	1.13E-03	0.04-0.53	0.16	0.02
C_G	GBM	337	PCPG	147	2.94E-03	0.05-0.63	0.18	0.03
C_G	OV	348	LUAD	390	4.49E-03	0.09-0.71	0.26	0.03
C_G	GBM	337	PRAD	413	2.05E-03	0.10-0.66	0.26	0.04
C_G	OV	348	KIRP	205	2.30E-03	0.07-0.60	0.21	0.05
C_G	GBM	337	PAAD	161	5.78E-03	0.05-0.70	0.20	0.06
C_G	OV	348	KICH	58	4.37E-03	1.11E-03-0.57	0.06	0.09
C_G	GBM	337	LGG	472	4.60E-03	0.13-0.75	0.32	0.10
C_G	OV	348	COAD	213	9.36E-03	0.07-0.72	0.22	0.10
C_G	OV	348	LAML	120	4.73E-03	0.02-0.65	0.14	0.10
C_G	SKCM	446	COAD	213	4.76E-03	1.33-7.86	3.19	0.10
D_E	GBM	337	UCEC	357	1.98E-03	0.06-0.63	0.21	0.02
D_E	UCEC	357	PCPG	147	1.55E-03	2.12-72.19	10.39	0.03
D_E	LUSC	337	UCEC	357	3.70E-03	0.07-0.68	0.22	0.04
D_E	GBM	337	KIRC	278	1.94E-03	8.55E-03-0.49	0.09	0.04
D_E	TGCT	119	KIRC	278	5.14E-03	4.93E-03-0.61	0.07	0.05
D_E	KIRC	278	LGG	472	3.49E-03	1.80-99.62	9.65	0.07
D_E	KIRC	278	PCPG	147	3.56E-03	1.89-99.80	11.58	0.07
D_E	LUAD	390	UCEC	357	0.01	0.08-0.82	0.26	0.09
D_E	GBM	337	BRCA	677	0.02	0.14-0.84	0.35	0.09
D_G	GBM	337	STAD	277	4.80E-04	2.35-126.14	12.29	0.01
D_G	GBM	337	BRCA	677	7.16E-03	1.29-9.71	3.47	0.05
D_G	GBM	337	PAAD	161	4.07E-03	1.68-100.57	9.41	0.06
D_G	PRAD	413	STAD	277	6.88E-03	1.46-66.62	6.92	0.07
E_F	GBM	337	UCEC	357	0.02	0.08-0.89	0.29	0.09
E_G	LUAD	390	HNSC	450	1.87E-04	0.04-0.46	0.15	3.93E-03
E_G	PRAD	413	HNSC	450	4.70E-04	0.05-0.50	0.17	9.87E-03
E_G	LUAD	390	STAD	277	3.02E-03	0.04-0.64	0.18	0.04
E_G	PRAD	413	STAD	277	3.88E-03	0.04-0.65	0.17	0.07
E_G	LUAD	390	UCEC	357	0.02	0.08-0.86	0.28	0.09
F_G	OV	348	BRCA	677	0.01	1.19-15.66	3.96	0.07
F_G	OV	348	LGG	472	9.75E-03	1.37-22.78	4.92	0.09

Note: Chi-square test (FDR < 0.25) was used to test if two cancer (sub)types could be distinguished by CGGPs or CGGP combinations.

Supplementary Table S6. Functional annotation of differentially expressed genes between the CGGP-defined subgroups of three cancer types

Cancer type	Subgroup	Differentially enriched Gene Ontology (GO) terms (FDR<0.05)
Breast (3 subgroups in total)	1 vs 3 (had significant survival differences)	Kinase, Nucleotide-binding, Transferase, protein phosphorylation,

	1 vs 2 (had significant survival differences)	Cell cycle, Mitosis, Mitotic nuclear division, Cell division, Centromere, Sister chromatid cohesion, Kinetochore
	2 vs 3	Ribonucleoprotein, Ribosomal protein, Ribosome (KEGG pathway hsa03010), rRNA processing, translation, SRP-dependent co-translational protein targeting to membrane, structural constituent of ribosome, ribosome, translational initiation, nuclear-transcribed mRNA catabolic process & nonsense-mediated decay, cytosolic large ribosomal subunit, viral transcription, cytosolic small ribosomal subunit; Spliceosome, mRNA splicing, mRNA processing; mitochondrial translational elongation/termination/ribosome/large ribosomal subunit; cell-cell adherence junction; RNA-binding
Kidney (3 subgroups in total)	1 vs 2 (had significant survival and histological differences)	ATP-binding, Nucleotide-binding, Kinase

Auxiliary Supplementary Materials

Supplementary Data S1. Seven cancer germline genomic patterns deciphered in this study. This file contains (a) the original CGGP, which is a float number matrix of shape (192, 7); (b) collapsed 96-component CGGPs, which is a float number matrix of shape (96, 7); (c) CGGPs derived from the DP20masked condition (i.e. removing repeats-related variants and outlier samples and keep other conditions as the same as those described for identifying genomic patterns in the main text), which is a float number matrix of shape (192,7); (d) the CGGs derived from European ancestry samples, which is a float number matrix of shape (192,7).

Supplementary Data S2. Per-sample CGGP weighing factor matrix of the TCGA dataset. This file contains (a) the per-sample CGGP weighing factor matrix re-solved by using the original CGGPs, which is a float number matrix of shape (9712,7); (b) the per-sample CGGP weighing factor matrix re-solved by using the DP20masked CGGPs, which is a float number matrix of shape (9712,7).

Supplementary Data S3. List of genes affected by CGGP_E. This file contains (a) the list of genes; (b) the associated functional enrichment analysis results.

Supplementary Data S4. Python code used to decipher germline genomic patterns.