

RESEARCH

Open Access

The rate and spectrum of mosaic mutations during embryogenesis revealed by RNA sequencing of 49 tissues



Francesc Muyas^{1,2,3*}, Luis Zapata^{2,4}, Roderic Guigó^{2,3} and Stephan Ossowski^{1,2,3*} 

Abstract

Background: Mosaic mutations acquired during early embryogenesis can lead to severe early-onset genetic disorders and cancer predisposition, but are often undetectable in blood samples. The rate and mutational spectrum of embryonic mosaic mutations (EMMs) have only been studied in few tissues, and their contribution to genetic disorders is unknown. Therefore, we investigated how frequent mosaic mutations occur during embryogenesis across all germ layers and tissues.

Methods: Mosaic mutation detection in 49 normal tissues from 570 individuals (Genotype-Tissue Expression (GTEx) cohort) was performed using a newly developed multi-tissue, multi-individual variant calling approach for RNA-seq data. Our method allows for reliable identification of EMMs and the developmental stage during which they appeared.

Results: The analysis of EMMs in 570 individuals revealed that newborns on average harbor 0.5–1 EMMs in the exome affecting multiple organs (1.3230×10^{-8} per nucleotide per individual), a similar frequency as reported for germline de novo mutations. Our multi-tissue, multi-individual study design allowed us to distinguish mosaic mutations acquired during different stages of embryogenesis and adult life, as well as to provide insights into the rate and spectrum of mosaic mutations. We observed that EMMs are dominated by a mutational signature associated with spontaneous deamination of methylated cytosines and the number of cell divisions. After birth, cells continue to accumulate somatic mutations, which can lead to the development of cancer. Investigation of the mutational spectrum of the gastrointestinal tract revealed a mutational pattern associated with the food-borne carcinogen aflatoxin, a signature that has so far only been reported in liver cancer.

Conclusions: In summary, our multi-tissue, multi-individual study reveals a surprisingly high number of embryonic mosaic mutations in coding regions, implying novel hypotheses and diagnostic procedures for investigating genetic causes of disease and cancer predisposition.

Keywords: Genetic mosaicism, Human embryogenesis, Mosaic mutation rate

* Correspondence: francesc.muyas-remolar@med.uni-tuebingen.de;
Stephan.ossowski@med.uni-tuebingen.de

¹Institute of Medical Genetics and Applied Genomics, University of Tübingen, Tübingen, Germany

Full list of author information is available at the end of the article



© The Author(s). 2020 **Open Access** This article is licensed under a Creative Commons Attribution 4.0 International License, which permits use, sharing, adaptation, distribution and reproduction in any medium or format, as long as you give appropriate credit to the original author(s) and the source, provide a link to the Creative Commons licence, and indicate if changes were made. The images or other third party material in this article are included in the article's Creative Commons licence, unless indicated otherwise in a credit line to the material. If material is not included in the article's Creative Commons licence and your intended use is not permitted by statutory regulation or exceeds the permitted use, you will need to obtain permission directly from the copyright holder. To view a copy of this licence, visit <http://creativecommons.org/licenses/by/4.0/>. The Creative Commons Public Domain Dedication waiver (<http://creativecommons.org/publicdomain/zero/1.0/>) applies to the data made available in this article, unless otherwise stated in a credit line to the data.

Background

Genetic mosaicism describes the co-existence of genetically different cell populations in an individual developing from a single fertilized egg [1–3]. Mosaicism has been associated with a broad range of genetic diseases [4], including neurological disorders [5, 6], brain malformation and overgrowth syndromes [7, 8], autism spectrum disorders [9], and cancer predisposition syndromes [10, 11]. Mosaicism can lead to genetic disorders that are embryonic lethal when occurring in germ cells [12], or result in a milder phenotype than a constitutive mutation [13]. The timing of mutations during embryogenesis (e.g., cleavage, blastulation, implantation, gastrulation, neurulation, and organogenesis) influences the fraction of affected cells and organs in the organism [4, 14]. Moreover, when occurring during gametogenesis, mosaic mutations can be passed on constitutionally to multiple offspring [3].

As expected, mosaic mutations are found in the form of single nucleotide variants (SNVs), insertions and deletions (indels), and copy number variants (CNVs) and have been studied using array technology [15] as well as next-generation sequencing (NGS) [16, 17]. A SNP array-based study of the Children’s Hospital of Philadelphia found that 17% of the diagnosed cases were caused by mosaic aneuploidies [18]. Acuna-Hidalgo and colleagues suggested that around 7% of presumed germline de novo mutations are in fact post-zygotic mosaic mutations [17]. Using whole-genome sequencing of normal blood from 241 adults, Ju et al. [19] estimated that approximately three mutations are accumulated per cell division during early embryogenesis. However, despite their potential importance for human disease, previous studies of mosaic mutations have focused on only one or few tissues or organs, e.g., using whole-exome sequencing data of brain tissues [20] or blood [17]. Therefore, a comprehensive view of mosaic mutations arising during embryogenesis, including their rate and mutational spectrum, is missing. Here, we exploit 10,097 RNA-seq samples from 49 different tissues and 570 individuals of the Genotype-Tissue Expression (GTEx) cohort [21] to uncover the rate and spectrum of mosaic mutations acquired post-zygotically during early embryogenesis.

Methods

Samples

In this study, we used release 7 of the Genotype-Tissue Expression (GTEx) [21, 22] project (dbGaP accession phs000424.v7.p2) [23], including RNA-seq data for 49 tissues from 570 individuals. We included only individuals for which whole-genome sequencing (WGS) data was available (necessary for distinguishing somatic from germline variants) and for which at least 8 tissues were analyzed by RNA-seq. Furthermore, we only included tissues for which RNA-seq data from at least 25 donors

was available. Filtering by these criteria resulted in RNA-seq data from 10,097 samples distributed over 570 individuals and 49 tissues (Additional file 1). Additional QC and filtering steps were performed depending on the specific analysis, as detailed below.

Pipeline for somatic variant prediction in RNA-seq data

Reads were aligned using STAR (version v2.4.2a, parameters see Additional file 2: Table S1) against the human reference genome (GRCh37), and the resulting BAM files were post-processed in order to remove alignment artifacts. PCR duplicates were marked using Picard (version 2.10.1), and reads mapping to different exons were split using SplitNCigar (part of GATK 3.7 package). Additionally, reads not overlapping with annotated human exons (ENSEMBL GRCh37 release 95) or aligning to immunoglobulin genes (potentially hyper-mutated) were removed from downstream analysis. Furthermore, reads aligning with mapping quality lower than 255, more than one gap opening, or more than 3 mismatches were filtered. Finally, in order to avoid systematic alignment errors at the extremes of the reads (which also include the “inner ends” of reads split across introns, i.e., breakpoints of spliced reads), we trimmed the first and last 4 bases from each read-end or read-breakpoint (BamUtil version 1.0.14).

Using the post-processed BAM files, we computed a three-dimensional genotype array (variant \times tissue \times individual) for all positions found to have a significant alternative allele call in at least one sample. This algorithm consists of two main steps:

Step 1: Single sample variant calling. First, base counts are obtained with *samtools mpileup* (version 1.3.1) followed by post-processing using custom scripts (Python and R scripts). We modeled the error rate (ER) distribution for each sample using a beta-binomial distribution. Counts of alternative (non-reference) reads at homozygous-reference positions (germline) are distributed following a binomial distribution with parameter P (error rate), which is a random variable that follows a beta distribution with parameters α and β .

Alternative counts \sim Bin (Coverage, error rate)

Error rate \sim Beta(α, β)

As the error rate differs depending on the nucleotide change (for example due to DNA oxidation artifacts affecting only a specific base), we modeled error distributions independently for each possible nucleotide change (A>C, A>T, A>G, C>A, C>T, C>G). Finally, we identified all sites showing alternative allele counts significantly deviating from the ER distribution after FDR correction. Additional filtering criteria were applied for

each site, including a minimum alternative allele count of 4 (each having at least base quality of 20), minimum read coverage of 10, alternative calls presented in forward and reverse strand following the same distribution as for reference counts (i.e., no strand bias), variant allele frequency (VAF) greater or equal to 5%, and minimum distance of 20 bp between variable sites in the same sample.

Step 2: Multi-sample re-calling of all potentially variable sites across all individuals and tissues is performed using a custom algorithm in order to build the three-dimensional genotype array. To this end, sites passing step 1 as significant in at least one sample were evaluated in each sample using the beta-binomial distribution as described for single samples, but with less stringent post-filtering criteria (i.e., without strand bias test and minimum required distance between variants), resulting in one of four possible filter states per sample: NO_EXPRESSION, HOM_REF, LOW_QUALITY or PASS. Furthermore, the exact reference-like and alternative allele counts are stored in the coordinates \times tissue \times individual array.

A random forest model for multi-tissue, multi-individual germline and somatic variant calling from RNA-seq data

We next aimed at training a random forest classifier distinguishing true from false positive variant calls in RNA-seq data. To this end, we selected 40 cases studied as part of the ICGC Chronic Lymphocytic Leukemia project [24, 25], for which whole-exome sequencing (WES) data for tumor and normal sample and RNA-seq data for tumor samples are available (see Additional file 3). RNA-seq-based variant calling was performed as described above for GTEx samples. Additionally, we obtained the reference and alternative allele counts from tumor and normal WES data for all putative calls identified in RNA-seq data. Finally, we used the WES data to predict high-quality germline and somatic variant calls using GATK HaplotypeCaller and MuTect2 as described before [26, 27].

Next, variants identified in RNA-seq data were randomly split into training and test sets for RF model training and testing, with the restrictions that:

- Training and test sets contain a similar number of true and false events according to WES data
- Training and test sets have a uniform distribution of variant allele frequencies, except for variants with $VAF < 10\%$, which were doubled (in order to increase the sensitivity of the RF for low VAF)

In addition, a set of non-overlapping high-quality calls from WES data was incorporated in the training and test sets. We labeled as true variants any site with $VAF \geq 5\%$

and at least 2 reads supporting the alternative allele in WES data, and all other sites as false variants. This procedure resulted in training and test data sets of 2402 sites each.

To train the RF model (R *randomForest* package) for distinguishing true and false positive variants (germline or somatic) called in RNA-seq data, we included as features (a) alternative allele count, (b) coverage, (c) VAF, (d) strand bias, (e) blacklisted genes [28], and (f) average alternative base quality. As this model, termed *RF-RNAmut* from here on, returned a response value between 0 and 1 for detecting calls, we chose our cutoff based on the maximum F1 score in the training set (cutoff = 0.19). Sites with response values exceeding 0.19 were labeled as high confident variants. To finally generate the somatic mutation call set and to remove systematic calling errors, we filtered variants if (1) they were recurrently called in RNA-seq data of multiple individuals, (2) their population allele frequency in GnomAD or 1000GP was greater than 1%, (3) they overlapped with repetitive elements annotated by Repeat Masker, (4) they overlapped with low complexity regions, (5) they were flagged as likely systematic analysis error by ABB [27], or (6) they overlapped with a known RNA editing site [29–31].

We measured the performance (precision and recall) of *RF-RNAmut* + Filter on identifying (a) germline and (b) somatic variant calls using the test set, following the same procedure as described above. To calculate precision, we considered as true or false positive calls those variants which were found in RNA-seq data and matched or not matched with tumor WES data, respectively. For calculating the false negative rate, we considered high-quality calls identified by MuTect2 in tumor-normal paired WES analysis that were not found in RNA-seq data. For benchmarking purposes, we only analyzed regions overlapping between RNA-seq (with more than 10x read coverage in annotated exons) and the WES enrichment kit (Agilent SureSelect 71Mb). Again, non-exonic regions, known editing sites, and immunoglobulin genes were ignored.

To demonstrate the gain in performance (precision) when using the RF variant filter and to validate that the RF model was not over-fitted to the training data, we trained 500 RF models on permuted training data. To this end, we permuted the labels (true, false) of the training set while keeping the other data (features) unchanged. Performances of the 500 permutation test models and the original *RF-RNAmut* model were plotted in a histogram for visual comparison (Additional file 4: Fig. S1).

Identification of mosaic mutations in the GTEx cohort

In order to obtain true mosaic variant calls for the GTEx cohort, we first removed all germline variants detected by WGS analysis in any individual (GATK HaplotypeCaller)

from the 3D genotype array (Fig. 1). Additionally, we removed any site for which the minor allele frequency in the population was greater or equal than 1% in GNOMAD or 1000GP. Furthermore, we removed all variants present in expressed tissues of all individuals, as they likely represent systematic errors, RNA editing sites, or germline de novo mutations. To further deplete calls produced by RNA editing events (mainly A > I, less frequently C > U), we ignored known editing sites described in the literature (<http://lilab.stanford.edu/> [29]), found in the Darned database (<https://darned.ucc.ie/download/> [30]), or identified by the GTEx consortium (<http://srv00.recas.ba.infn.it/atlas/pubs.html> - REDiportal [31]).

Next, we removed sites, which recurrently exhibit low-quality (LQ) calls across multiple individuals, which are likely systematic sequencing or alignment errors. Moreover, we filtered out positions labeled as systematic errors by ABB [27]. Additionally, we removed any variant overlapping with low complexity regions or repeat regions annotated by *Repeat Masker*. Finally, as we did not expect mosaic mutations to be highly recurrent in different individuals, we removed sites called in more than 2 individuals of our cohort.

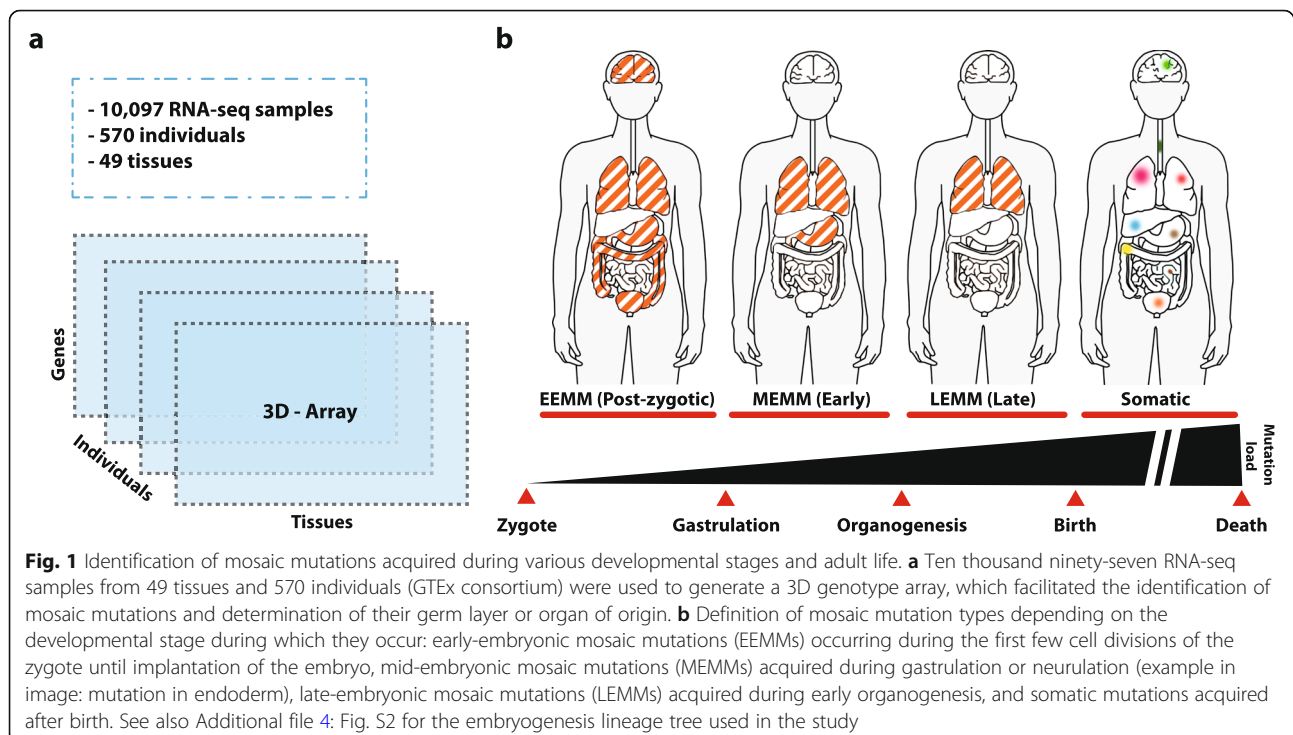
Identification of early- (EEMMs) and mid-embryonic mosaic mutations (MEMMs)

In order to identify mosaic mutations acquired during early embryogenesis (cleavage, blastulation, gastrulation, neurulation, and early organogenesis), we contrasted the somatic calls in the 3D genotype array with a lineage tree

of human embryogenesis and tissue development including the 49 tissues studied here (Fig. 1, Additional file 4: Fig. S2) [32]. In this part of the analysis, only individuals with 10 or more tissues sequenced with at least two germ layers represented by 2 sequenced tissues were included in the analysis (526 individuals, see Additional file 1). This procedure allowed us to identify mosaic mutations affecting at least two tissues, whose origin could be unambiguously mapped to a specific stage of development and/or primary germ layer.

Mosaic mutations identified in both the ectoderm and mesendoderm branches having zygote as most likely ancestral node, i.e., variants likely originating from the first few divisions of the zygote (cleavage, blastulation, implantation stages), were defined as early-embryonic mosaic mutations (EEMMs). In order to avoid detection of de novo germline variants as EEMMs, we only considered variants with VAF less than 0.35 that were not found in all expressed tissues of an individual. Importantly, the adrenal gland, which is comprised of cells originating from the ectoderm (medulla) and cells originating from the mesoderm (cortex) [33], was excluded from this analysis in order to avoid overestimation of early embryonic mutations.

The remaining mutations found in at least two tissues of an individual were defined as mid-embryonic mosaic mutations (MEMMs) if (1) their most likely ancestral node was not zygote, (2) they were only observed in either the ectoderm or the mesendoderm sub-tree, and (3) their appearance in the lineage tree was coherent. Contradictory (non-coherent) mutation patterns were



defined as alternative alleles, which were observed in far-apart nodes in the tree, but which were undetectable in any node close to the affected tissues. In other words, mosaic mutations that required the assumption that they had occurred multiple times independently in different cells of the same individual were not considered coherent and were removed.

Finally, we defined late-embryonic mosaic mutations (LEMMs) as those mutations that are restricted to one tissue/organ, but likely occurred early during organogenesis. To this end, we considered variants found in a single tissue per individual, supported by 5 or more reads and with VAF of ≥ 0.2 . This procedure cannot distinguish mosaic mutations acquired during late embryogenesis (organogenesis) from mutations in clonal expansions acquired after birth. We therefore excluded somatic variants from tissues known to have detectable clonal expansions such as the sun-exposed skin, esophagus-mucosa, and whole blood.

Estimating the rate of mosaic mutations during embryogenesis

Reliable detection of mosaic mutations in a gene using RNA-seq data and definition of the mutation's origin in the lineage tree requires high gene expression in a majority of tissues of an individual. In order to estimate the rate of mosaic mutations, we therefore focused on genes that are highly and constitutively expressed in most of the analyzed tissues. Given a large enough pool of constitutively expressed genes, we can subsequently extrapolate mutation rates to the whole exome or genome, as suggested previously for measuring genome-wide tumor mutation burden (TMB) using small cancer gene panels [34]. We used four different thresholds to define sets of constitutively expressed genes. For each set, we independently estimated the rate of mosaic mutations, to ultimately evaluate the robustness of our approach by comparing the four estimates. The following definitions were used to define constitutively expressed genes:

1. Genes with TPM ≥ 5 in more than 75% of all total samples (7630 genes)
2. Genes with TPM ≥ 10 in more than 75% of the total samples (5231 genes)
3. Genes with COV ≥ 20 in more than 75% of the total samples (6888 genes)
4. Genes with COV ≥ 30 in more than 75% of the total samples (5370 genes)

(TPM = transcripts per kilobase per million, COV = average read coverage across a gene)

Next, we obtained all mosaic variants identified in a given set of constitutively expressed genes and calculated

the number of mutations per base and individual relative to the total length of the interrogated region. Finally, we extrapolated this value to the approximate total length of all coding exons (45 Mbp) in order to calculate the number of mosaic coding mutations expected on average for a newborn child. The procedure was independently performed for EEMMs and MEMMs.

For LEMMs, which were defined as tissue-specific, we considered any gene highly expressed in a given tissue of an individual (i.e., a sample). We normalized the number of mutations per base and individual relative to the interrogated region for a given sample and extrapolated this value to the approximate total length of all coding exons (45 Mbp). Due to their similarity with mutations in clonal expansions, the rates of LEMMs per exome per individual are likely overestimated.

Tissue-specific somatic mutation rates

In order to study somatic mutations acquired after birth, the rate of somatic mutations, signatures of selection, and mutation spectra in a tissue-specific manner, we performed somatic variant calling using *RF-RNAmut* without the restrictions applied for the detection of embryonic mosaic mutations. Here, we only considered somatic mutations identified in exactly one tissue per individual in order to minimize the number of mosaic mutations acquired before birth in this set. First, we performed samples-wise quality control (Additional file 4: Fig. S3) and excluded samples with the following characteristics:

- PCR duplicate rates in the top 5%
- Outliers for the number of callable sites (top and bottom 1% per tissue). We considered a site as *callable* if the read coverage was ≥ 10
- Outliers for RIN (bottom 1% per tissue)
- Outliers for mutation rate (top 1% per tissue)
- Samples obtained from cell culture (cells-EBV-transformed_lymphocytes, cells-transformed_fibroblasts)
- Individuals affected by cancer

In order to improve the statistical power, we removed tissues with less than 50 high-quality samples from downstream analysis (affecting only the kidney with 38 high-quality samples, see Additional file 4: Fig. S3d), resulting in 8351 samples from 46 tissues and 558 individuals. We calculated the somatic mutation rate based on the number of identified somatic mutations divided by the callable sites per sample. As quality control revealed a strong influence of technical confounders (PCR duplicate rate, RIN, average coverage, sequencing center) on the number of detectable mutations, we used a linear regression model to estimate and subtract technical biases. The linear regression model uses the following variables:

$$\begin{aligned} \text{Mutation rate} \sim & \text{duplicates} + \text{cohort} + \text{RIN} \\ & + \text{TRISCHD} + \text{DP}_{\text{median}} \\ & + \varepsilon \text{ (mutRate residuals)} \end{aligned}$$

We understand mutRate residuals (ε) as the variability of the observed (raw) mutational rate, which is not explained by non-biological (technical) features such as PCR duplicate rates, cohort, or RIN. In order to assess the effect of age and tissue on mutation rates, we assessed the relation of the remaining variability (mutRate residuals) and the age of an individual at death, separately for each tissue, using a Spearman's rank correlation test (all p values were corrected with FDR).

Mutational signatures

Mutational signatures were computed using the R package *deconstructSigs* [35], and only signature weights greater than 0.1 were shown in plots.

For computing mutational signatures of embryonic mosaic mutations, all individuals were considered for which at least 10 tissues were sequenced. For the calculation of signatures of somatic mutations acquired during the lifespan, only individuals older than 60 years were included in the analysis in order to increase the number of mutations related to mutagenic processes. Again, we focused on mutations found in exons due to the limited RNA-seq coverage in intergenic and introic regions. We obtained mutational signatures for each tissue separately, as well as for groups of tissues based on predominant environmental exposures, with a specific focus on:

- Sun-exposed skin
- Non-sun-exposed skin
- Exposure to mutagens in food: colon, esophagus-mucosa, small intestine, liver, and stomach
- Brain tissues: brain-anterior_cingulate_cortex_BA24, brain-hippocampus, brain-substantia_nigra, brain-caudate_basal_ganglia, brain-cerebellar_hemisphere, brain-frontal_cortex_BA9, brain-spinal_cord_cervical_c-1, brain-amygdala, brain-cortex, brain-cerebellum, brain-hypothalamus, brain-nucleus_accumbens_basal_ganglia, brain-putamen_basal_ganglia

Identifying signatures of positive selection in cancer genes using dN/dS

To estimate the extent of selection acting on somatic mutations in healthy tissues, we used the SSB-dN/dS method [36], which calculates the trinucleotide-corrected ratio of nonsynonymous to synonymous mutations from NGS data. Somatic mutations identified by *RF-RNAmut* were

annotated using the variant effect predictor (VEP). To increase statistical power, we only considered constitutively expressed genes having more than 5 TPM in at least 75% of patients for a focal tissue. We computed SSB-dN/dS in each tissue separately, and in the pan-tissue combinations listed above, using 192 parameters for nucleotide bias correction (correcting for mutation bias in all possible triplets on forward and reverse strand). However, we only computed dN/dS values for those tissues having at least 3 non-silent or silent somatic mutations in the analyzed genes. In addition to the exome-wide dN/dS provided in the output of the SSB-dN/dS method, we calculated the global dN/dS for 198 cancer genes [37] and 995 essential genes [36]. Finally, we focused on NOTCH1 and TP53 genes in order to replicate the findings of strong positive selection described recently [37–40].

Results

Somatic variant calling in RNA-seq data

Somatic variant detection using RNA-seq data is challenging, especially if subclonal mutations with allele fractions as low as 5% are of interest [40]. We therefore developed a highly accurate multi-sample variant calling procedure, which models nucleotide-specific errors, removes germline variants and confounders such as RNA editing sites, and generates a multi-individual, multi-tissue array of variant calls (3D genotype array) by re-genotyping potentially variable sites across thousands of GTEx RNA-seq samples (Fig. 1a). Although several methods for RNA-seq-based mutation detection exist (SEURAT [41], RADIA [42], VaDiR [43], or RNA-MuTect [40]), our method is the first to apply a multi-sample variant detection design concurrently taking into account multiple tissues across hundreds of individuals. This novel approach permitted us to (1) reliably distinguish somatic mutations from germline variants and post-transcriptional modifications, (2) distinguish embryonic mosaic mutations (EMMs) from germline de novo and adult somatic mutations, (3) achieve high sensitivity for detecting all tissues of a person harboring a specific mosaic mutation, (4) estimate the time point and germ layer at which an EMM occurred, and (5) establish a reliable estimate of embryonic mosaic mutation frequencies across a large cohort.

To further improve the specificity of our method, we trained a random forest classifier (*RF-RNAmut*) distinguishing true from false mutation calls. We used whole-exome sequencing (WES) and RNA-seq data from the ICGC Chronic Lymphocytic Leukaemia project [24] for generating training and independent test data sets (the “Methods” section, Additional file 3). High confidence somatic variant calls with >0.15 VAF in tumor WES data were identified in RNA-seq with 71% sensitivity and 85% precision (comparable to the performance of the method described by Yizhak et al. [40] with sensitivity

and precision of 0.72 and 0.87, respectively), and sensitivity was positively correlated with VAF (Additional file 2: Table S2). A comparison of *RF-RNAmut* to random forest models trained on permuted data sets indicated that *RF-RNAmut* increases the precision of the “raw” SNV calls from 54 to 85% (the “Methods” section, Additional file 5 and Additional file 4: Fig. S1). Germline variants found in tumor and normal WES data were identified in RNA-seq data with 86% sensitivity and 95% precision (however, germline variants are of no interest for this study).

Rate and spectrum of early mosaic mutations during embryogenesis

In order to identify mosaic mutations acquired during embryonic development, we computed the 3D genotype array for 9704 samples of the GTEx cohort comprising 526 cancer-free individuals and 49 tissues (see the “Methods” section and Additional file 1 for sample selection criteria). We contrasted the 3D genotype array with the embryogenesis lineage tree (Additional file 4: Fig. S2, Additional files 6 and 7 for detailed calls) to identify the most likely germ layer or tissue of origin of each mutation. We first removed variants occurring in all expressed tissues with average VAF greater than 0.35, as they might constitute *de novo* germline variants. Then, we defined three types of embryonic mosaic mutations (EMMs): early- (pre-implantation), mid- (gastrulation and neurulation), and late- (organogenesis) embryonic mosaic mutations (EEMMs, MEMMs, LEMMs in Fig. 1b). EEMMs appeared during the first few divisions of the zygote (cleavage, blastulation, implantation) and therefore are present in the ectoderm and mesendoderm (mesoderm and/or endoderm). MEMMs are mutations found in at least two tissues of the same individual that originate from the same germ layer. We define LEMMs as mutations present in a large cell fraction of a single organ, which are not the consequence of somatic clonal expansions. Finally, we also screened for postnatal and adult somatic mutations in the transcriptome of all cancer-free individuals.

To minimize false negatives, we focused our analysis on housekeeping genes constitutively expressed in the majority of tissues and samples (7630 genes with TPM > 5 in at least 75% of tissues). After strict filtering (the “Methods” section), we identified 58 putative EEMMs and 37 MEMMs in 7630 constitutively expressed genes. We estimated a rate of 8.1164×10^{-9} (CI (95%) = $[7.0973 \times 10^{-9}$ to $9.1292 \times 10^{-9}]$) EEMMs and a rate of 5.1166×10^{-9} (CI (95%) = $[4.5592 \times 10^{-9}$ to $5.6740 \times 10^{-9}]$) MEMMs per nucleotide and individual for exonic regions. Following an approach for extrapolating tumor mutation burden (TMB) from gene panels to exomes (45-Mbp exonic regions) [34], we estimated a mean of 0.37 exonic EEMMs (Fig. 2a) and 0.23 exonic MEMMs (Fig. 2b) per individual

(0.44 and 0.275 when correcting for precision and sensitivity of our variant calling algorithm). Using different thresholds for constitutively expressed genes only marginally affected the estimated rate of EEMMs or MEMMs (Fig. 2a, b, Additional file 2: Table S3). We also observed no correlation between the embryonic expression levels of the 7630 selected genes (based on Yan et al. [44]) and EEMM mutational rates ($R^2 = 0.009$ and p value = 0.94, Additional file 5 and Additional file 4: Fig. S4), indicating that transcription-coupled repair efficiency at different expression levels had no measurable effect on the estimation of mutational rates. We have furthermore tested if the level of immune cell infiltration in different tissues biased the variant allele frequencies of detectable mutations and thereby the estimate of EMM rates (Additional file 5). We found no correlation between the VAF of EMMs and the fraction of immune cells in their respective tissue (p value = 0.648, Pearson correlation’s test, Additional file 4: Fig. S5), evidencing that infiltration of hematopoietic cells did not measurably bias our results.

A recent study by Ju et al. [19] used whole-genome sequencing of blood samples from 241 individuals to estimate that approximately three mutations are acquired per cell per cell division during early embryogenesis. Extrapolation of this estimate to the expected mutation burden after three to four divisions of the zygote (Additional file 5 and Additional file 4: Fig. S6) results in approximately 0.3 to 0.6 exonic mutations per individual, an estimate that is reassuringly similar to our estimate of 0.44 early embryonic mosaic mutations per individual.

On average, a specific EEMM was detectable in 63.6% of the tissues of an individual expressing the respective gene, consistent with the assumption that they arose during the first divisions of the zygote. Interestingly, only 41% of EEMMs in genes expressed in blood were detectable in blood samples, which could be explained by the asymmetric cell doubling model (unequal contribution of early-embryonic cells to adult somatic tissues) suggested by Ju et al. [19]. Hence, a large fraction of mosaic mutations would be missed by blood-based genetic diagnostic tests. As expected, we observed a positive correlation between the variant allele fraction of EEMMs/MEMMs and the number of tissues supporting the variant ($Rho = 0.56$; p value = 3.24×10^{-9}). Moreover, mutations occurring earlier in development also showed a greater proportion of cells carrying the variant ($Rho = -0.39$; p value = 7.83×10^{-5} , Additional file 4: Fig. S7).

The combined rate of EEMMs and MEMMs of 1.32×10^{-8} is comparable to the estimated rate of *de novo* germline mutations reported in the literature [3, 17], ranging from 1.0 to 1.8×10^{-8} per nucleotide per generation (44 to 82 mutations per genome [3], or ~0.5–1 mutations per exome (45 Mbp) per individual). Recently, several genetic disease studies indicated that more than

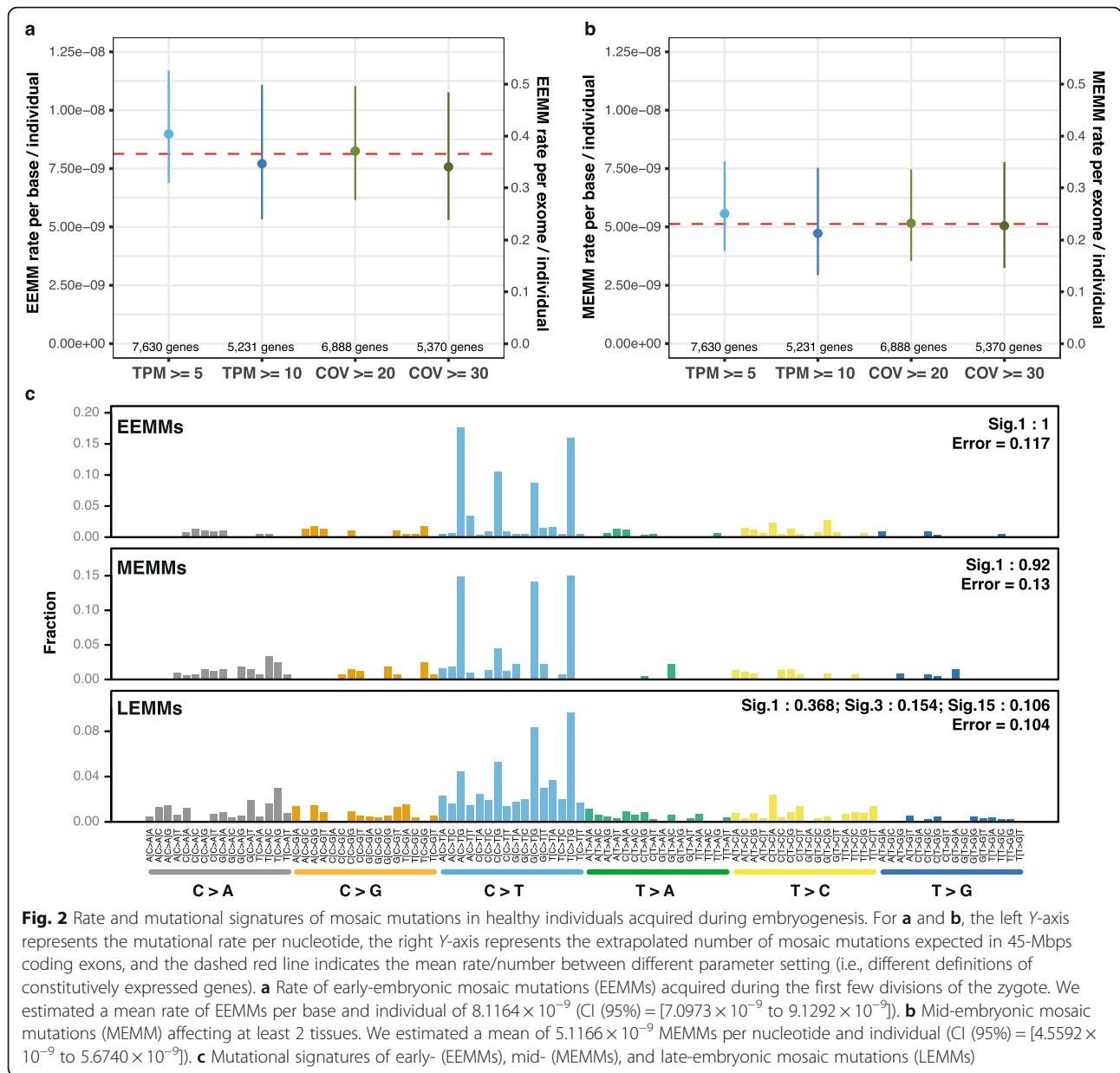


Fig. 2 Rate and mutational signatures of mosaic mutations in healthy individuals acquired during embryogenesis. For **a** and **b**, the left Y-axis represents the mutational rate per nucleotide, the right Y-axis represents the extrapolated number of mosaic mutations expected in 45-Mbps coding exons, and the dashed red line indicates the mean rate/number between different parameter setting (i.e., different definitions of constitutively expressed genes). **a** Rate of early-embryonic mosaic mutations (EEMMs) acquired during the first few divisions of the zygote. We estimated a mean rate of EEMMs per base and individual of 8.1164×10^{-9} (CI (95%) = $[7.0973 \times 10^{-9}$ to $9.1292 \times 10^{-9}]$). **b** Mid-embryonic mosaic mutations (MEMM) affecting at least 2 tissues. We estimated a mean of 5.1166×10^{-9} MEMMs per nucleotide and individual (CI (95%) = $[4.5592 \times 10^{-9}$ to $5.6740 \times 10^{-9}]$). **c** Mutational signatures of early- (EEMMs), mid- (MEMMs), and late-embryonic mosaic mutations (LEMMs)

50% of sporadic cases can be explained by de novo germline mutations [3, 5]. Consequently, embryonic mosaic mutations are similarly likely to explain a significant fraction of sporadic genetic disease cases, and a substantial fraction of germline de novo variants identified in blood are potentially post-zygotic mutations. Moreover, we likely underestimated the rate of EEMMs and MEMMs due to factors such as allele-specific expression, nonsense-mediated decay, and more effective transcription-coupled repair in highly expressed genes. As most of the disease-causing mosaic mutations cannot be detected by sequencing blood-derived DNA, these variants have likely been missed in past studies and could explain a substantial part of the missing heritability.

In order to identify the most likely processes causing early- and mid-embryonic mosaic mutations, we investigated their mutational signatures. We found that a large fraction of EEMMs and MEMMs (1 and 0.92) could be explained by Signature 1 [45–47] (Fig. 2c). signature 1 is thought to be the result of an endogenous mutational process initiated by spontaneous deamination of 5-methylcytosine leading to C>T transitions at CpG dinucleotides and likely reflects a cell-cycle-dependent mutational clock [46]. Hence, our findings indicate that most early- and mid-embryonic mosaic mutations occur spontaneously with very limited contributions from exposure to environmental factors or other endogenous processes. Furthermore, our results clearly distinguish early mosaic mutations from

germline de novo mutations, which are dominated by signature 5 characterized by A>G transitions [3].

Late-embryonic mosaic mutations arising during organogenesis

Our definitions of EEMMs and MEMMs prevent the identification of organ-specific mutations acquired during organogenesis. We therefore screened for late-embryonic mosaic mutations (LEMMs, Fig. 1b), which we defined as tissue-specific mutations at high cell fraction ($VAF \geq 0.2$). Here, we excluded tissues previously shown to be affected by clonal expansion of mutated cells such as the esophagus-mucosa, sun-exposed skin [34, 37–40, 48], and whole blood [14, 49], which also showed the highest somatic mutation rates in our analysis (Additional file 4: Fig. S8). We identified 377 mutations across all individuals, considering any gene expressed in at least one tissue (Additional file 8), resulting in an estimate of 2.44×10^{-9} (CI [0.95] = $[1.86 \times 10^{-9} - 3.03 \times 10^{-9}]$) LEMMs per nucleotide per tissue per individual, and extrapolating to 0.11 (CI [0.95] = $[0.084 - 0.137]$) mutations per exome per tissue. Notably, the average rate of LEMMs (2.23×10^{-9}) for brain tissues closely resembled the estimates by Wei et al. [20] (2.55×10^{-9}) obtained using WES data of brain tissues. In sum across all 43 examined tissues, we estimated 4.7 LEMMs per exome per individual.

Due to the incompleteness of the GTEx tissue matrix and variable expression levels of genes across tissues, it is not possible to ascertain if exactly one tissue is affected by a mosaic mutation. However, examination of expression levels for 100 randomly selected single-tissue mutations revealed that for most of the mutations a sufficient fraction of tissues showed enough high expression to determine that they did not occur before neurulation (Additional file 5 and Additional file 4: Fig. S9).

LEMMs are indistinguishable from mutations in clonal expansions acquired after birth [37, 38, 40, 48], and the rate of LEMMs is therefore likely overestimated. Nonetheless, our results indicate that organ-specific mosaic mutations arising during organogenesis could significantly contribute to the phenomenon of missing heritability in rare genetic diseases as well as cancer predisposition.

Rate and mutational signatures of tissue-specific somatic mutations

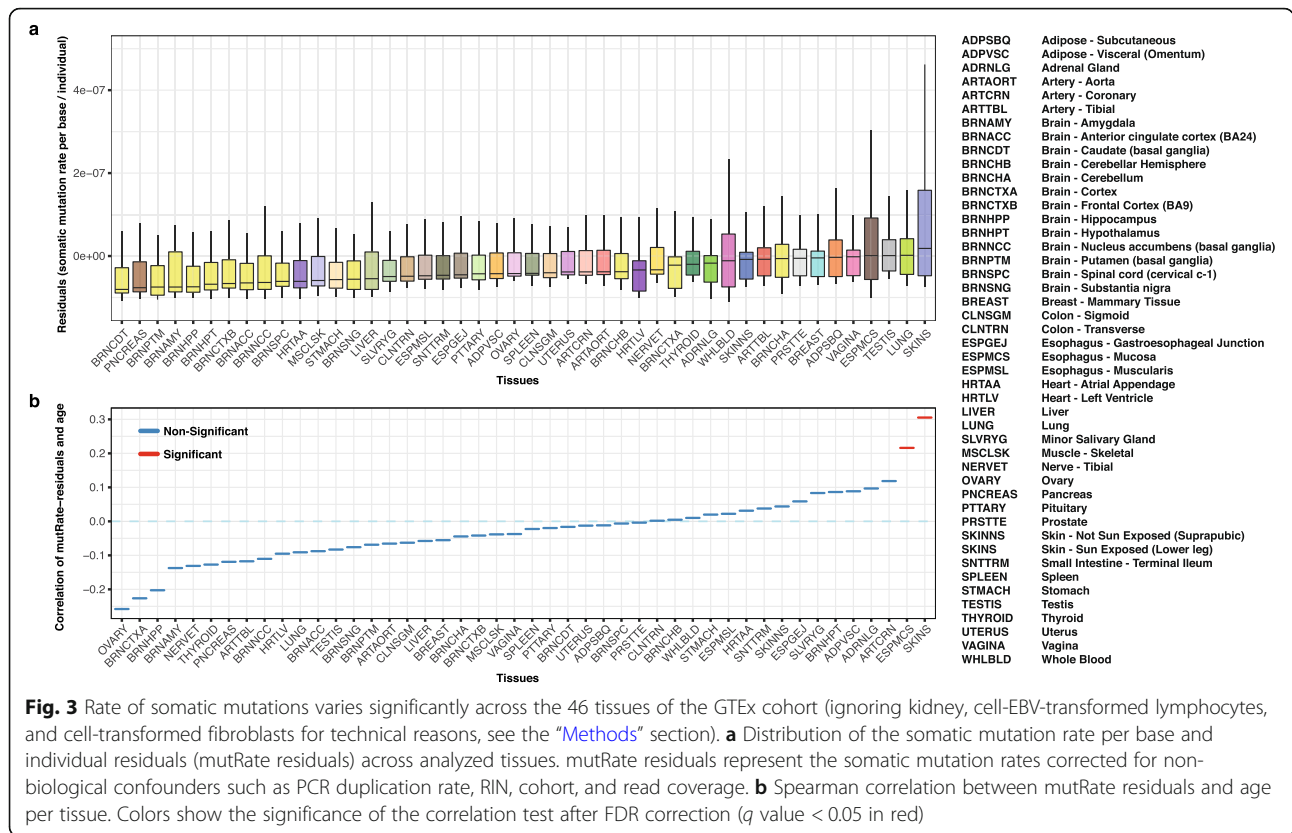
To identify other mutation processes leading to the accumulation of somatic mutations during adult life, we next studied mutational signatures across all tissue-specific somatic variants identified in the GTEx cohort. Considering only variants with $VAF \geq 0.05$, we identified 8780 somatic mutations in 8351 samples representing 46 tissues (the “Methods” section, Additional file 4: Fig. S3, and Additional files 6 and 7 for call set details). We observed lower power to detect somatic SNVs in lowly

expressed genes (TPM < 10) likely due to lack of coverage (Additional file 4: Fig. S10a–b) and a negative correlation between read coverage and VAFs of detectable mutations (Spearman $R = -0.81$, p value < 10^{-16} ; see Additional file 5 and Additional file 4: Fig. S10c). However, considering only genes with TPM > 10, we observed no significant correlation between gene expression and the fraction of mutated genes (Additional file 4: Fig. S10b). After removal of technical confounders (PCR duplicate rates, RIN, TRISCHD, coverage, laboratory), we observed the highest mutation burden for the sun-exposed skin, lung, testis, esophagus-mucosa, and vagina (Fig. 3a, Additional file 4: Fig. S11, Additional file 9). Our results confirmed the previous finding presented in Yizhak et al. [40] that the skin, lung, and esophagus are the tissues with the highest average number of mutations, likely explained by the constant exposure to environmental factors such as UV radiation, air pollution, smoking, and food. As expected, the sun-exposed skin showed significantly higher mutation burden than the non-sun-exposed skin, while brain tissues showed, in general, the lowest somatic mutation burden. We observed that the mean numbers of somatic mutations per sample highly correlated between the two studies for all analyzed tissues (Additional file 4: Fig. S12, Pearson $R = 0.92$, p value = 1.46×10^{-10}).

Finally, we tested if residual mutation rates were related with the age of individuals for each tissue individually (Fig. 3b). Only two tissues showed a significant association between age and mutational rates (after FDR correction), namely the sun-exposed skin ($Rho = 0.31$; $qval = 1.19 \times 10^{-7}$) and esophagus-mucosa ($Rho = 0.22$; $qval = 2.82 \times 10^{-3}$), confirming previously reported results [37–40]. Using dN/dS as a measure of selection, we observed a lack of selection in highly expressed genes at a pan-tissue level (dN/dS = 0.98, CI [95] = $[0.92 - 1.06]$). However, when focusing on cancer genes, we observed a strong positive selection for the sun-exposed skin and esophagus-mucosa (Additional file 4: Fig. S13). Mutations in *NOTCH1* and *TP53* disproportionately contributed to the high dN/dS values and showed the highest overall mutation rates. *NOTCH1* showed stronger positive selection than *TP53* in both esophagus-mucosa and skin sun-exposed (dN/dS of 8.46 vs. 4.57 and dN/dS of 4.01 vs. 2.85, respectively, Additional file 2: Table S4). Interestingly, we did not find a positive selection of these two genes in any other tissues, and no other gene reached significance in any of the tissues.

Aflatoxin mutational signature in organs of the dietary tract

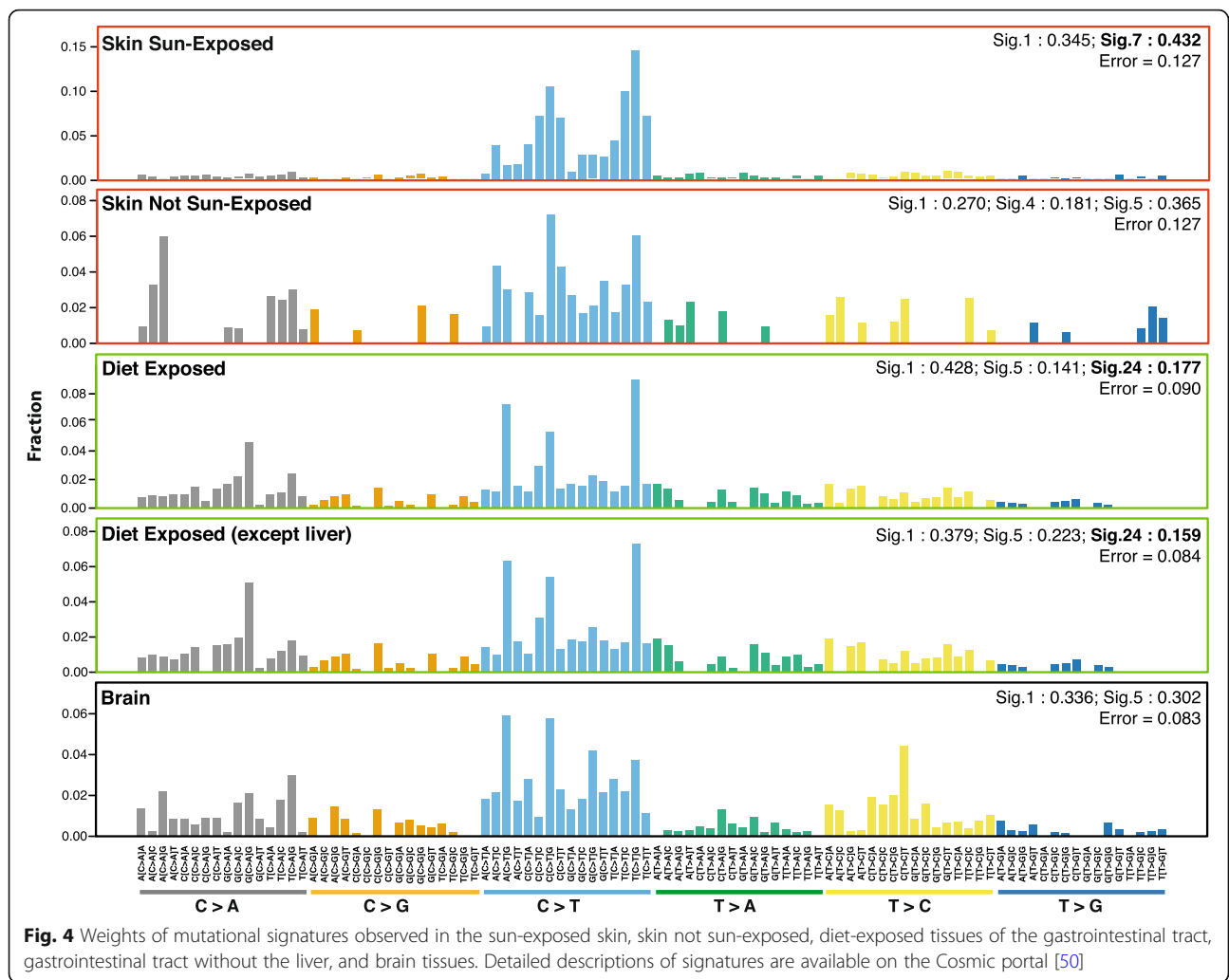
Previous studies have analyzed the spectrum of somatic mutations in the healthy esophagus and skin [37–40], identifying mutational signatures [45] 1, 5, and 7 [50]. Our analysis of mutational signatures for patients who



died at advanced age (≥ 60 years old) revealed that ultraviolet light (UV) exposure (signature 7) was predominant in the sun-exposed skin, while it was absent from the non-sun-exposed skin (Fig. 4). Our observations confirm the results of previous studies on healthy skin samples [37, 51], which showed a highly similar distribution of nucleotide substitutions and a strong prevalence for C>T mutations characteristic of UV-radiation damage (Additional file 4: Fig. S14). Interestingly, studies of the mutational signatures found in healthy tissues forming the gastrointestinal tract (GI tract) are lacking, although the constant exposure to food likely leads to a particular mutational spectrum. We therefore performed a pan-gastrointestinal-tract mutational signature analysis considering the colon, esophagus-mucosa, liver, small intestine, and stomach. Apart from signatures 1 and 5, which are frequently observed in most tissues, we found a signature explained by the mutagenic effect of dietary aflatoxin (signature 24). The aflatoxin signature explained a fraction of 0.18 of the mutational spectrum in the tissues of the GI tract (Fig. 4, Additional file 10). Furthermore, we saw a strong enrichment of the characteristic CGN > CTN mutations not observed in any other tissue. Finally, we observed that the aflatoxin signature is significantly stronger in older individuals (age > 60) than in younger individuals (age < 45) across all organs of the

GI tract (two-way Mann-Whitney-Wilcoxon’s test, p value < 0.01, Additional file 4: Fig. S15).

Aflatoxin B₁ (AFB1) is a potent mutagen and carcinogen typically found in grains contaminated with the food spoilage fungus, *Aspergillus flavus*. Dietary exposure to AFB1 is a known risk factor for human hepatocellular carcinoma (HCC), the third leading cause of cancer death worldwide. One of aflatoxin degradation products, the metabolite exoepoxide, forms a covalent bond with guanyl N7 (AFB1-N7-Gua), ultimately leading to G>T mutations during replication. Consistently, signature 24 has previously been found in a subset of liver cancers [52, 53], but has not been reported for other cancer entities. We therefore tested if the observed enrichment of signature 24 was solely introduced by a strong mutagenic effect in the liver. On the contrary, when excluding the liver from the analysis, the aflatoxin signature was still found at a similar level, explaining a fraction of close to 0.16 of the mutational spectrum. These results indicate that aflatoxin-related mutations are frequent in all tissues of the gastrointestinal tract and might play a role in the development of cancer in several organs. Indeed, evidence for the involvement of aflatoxin in gallbladder cancer (an organ of the GI tract) has recently been published by Koshiol et al. [54] and reported in the latest signature analysis of the Cosmic cohort (<https://cancer.sanger.ac.uk/cosmic/signatures>), supporting our hypothesis.



Discussion

The accumulation of DNA mutations during life is inevitable, despite the many cellular mechanisms involved in the preservation of genome integrity. In this study, we presented a novel analysis strategy using RNA-seq data of multiple tissues per individual to identify mosaic mutations occurring during various stages of embryo development. Using the human embryonic lineage tree, we approximated the time point of the mutation events as well as the affected germ layer or developing organ. We demonstrated how to distinguish, to some extent, embryonic mosaic mutations from de novo germline mutations and somatic mutations in clonal expansions acquired after birth.

Analyzing RNA sequence data from 49 tissues and 570 patients, we found that newborns on average harbor 0.5–1 mosaic mutation in coding exons affecting multiple tissues and organs, and likely an even larger number of organ-specific coding mutations. Post-zygotic and early-embryonic

mosaic mutation patterns are dominated by signature 1, which is associated with aging and cell division. Hence, they largely result from spontaneous deamination of methylated cytosines without showing any influence of external mutagens. Moreover, our estimates suggest that embryonic mosaic mutations are as frequent as germline de novo mutations and could explain a substantial fraction of unresolved cases of sporadic and rare genetic diseases, as well as play a role in cancer predisposition.

The recognition of a widespread and under-recognized role of mosaic mutations in genetic disease would have many implications for genetic diagnostics procedures [55]. We have furthermore demonstrated that a substantial fraction of EMMs is not detectable in blood cells, a finding which has important implications for clinical diagnostics, as samples from the affected tissue are often unavailable. Instead, sequencing of circulating cell-free DNA (liquid biopsy), which has been successfully applied for the detection of somatic mutations in solid

tumor tissues [56–58] and healthy individuals [59–61], could be an unbiased approach for the detection of embryonic mosaic mutations causing rare genetic diseases.

Interestingly, our method also revealed a strong signature of the food poison aflatoxin detectable in all organs of the dietary tract. Aflatoxin mutations have previously been associated to liver cancer. Our results indicate that the role of aflatoxins in cancer development might be more widespread than previously appreciated, affecting the mutation spectrum of tumors in the colon, esophagus-mucosa, liver, small intestine, and stomach.

Conclusions

In this study, based on a multi-tissue, multi-individual analysis, we found a surprisingly high number of embryonic mosaic mutations in exonic regions of healthy individuals, implying novel hypotheses and diagnostic procedures for investigating genetic causes of disease, cancer predisposition, and aging.

Supplementary information

Supplementary information accompanies this paper at <https://doi.org/10.1186/s13073-020-00746-1>.

Additional file 1. GTEx donor IDs. GTEx release v7 samples used for different analyses performed in this study.

Additional file 2: Supplementary tables. This document contains additional supporting evidences presented as supplemental tables (Table S1–S4).

Additional file 3 Sample information for the chronic lymphocytic leukemia cohort used for training and benchmarking of *RF-RNAmut*.

Additional file 4: Supplementary figures. This document contains additional supporting evidences presented as supplemental figures (Fig. S1–S15).

Additional file 5. Supplementary methods and extended results.

Additional file 6. 3D-genotype array containing all multi-sample, multi-tissue somatic and mosaic calls. Calls passing all quality filters are represented with the alternative allele (A,C,T,G), reference calls are represented as "." and tissues without expression are labeled as NA. Variants overlapping with any set of constitutively expressed genes are labeled as *TRUE* in the respective columns *Highly_expressed_genes_TPM5*, *Highly_expressed_genes_TPM10*, *Highly_expressed_genes_COV20* or *Highly_expressed_genes_COV30*.

Additional file 7. Somatic variant calling information for all somatic mutations identified in this study.

Additional file 8. Number of late-embryonic mosaic mutations (LEMMs) per tissue and individual. Donors with 0 LEMMs were not shown in this table.

Additional file 9. Somatic mutation rate (SMR) per nucleotide and individual. The column *Somatic_Mutation_Rate_Residuals* provides the residuals of the SMR obtained after removal of batch effects using linear regression.

Additional file 10. Signature weights obtained in the mutational signature analysis performed in tissue-specific somatic mutations.

Abbreviations

AFB1: Aflatoxin B₁; CI: Confidence interval; CLL: Chronic lymphocytic leukemia; CNV: Copy number variant; COV: Read coverage; DP: Depth of coverage; EEMM: Early-embryonic mosaic mutation; EMM: Embryonic mosaic mutation; ER: Error rate; FDR: False discovery rate; GATK: The Genome

Analysis Toolkit; GI: Gastrointestinal; GTEx: Genotype-Tissue Expression; HCC: Human hepatocellular carcinoma; Indel: Insertion or deletion; LEMM: Late-embryonic mosaic mutation; MEMM: Mid-embryonic mosaic mutation; NGS: Next-generation sequencing; RF: Random forest; RIN: RNA integrity number; SNV: Single nucleotide variant; TMB: Tumor mutation burden; TPM: Transcripts per million; UV: Ultraviolet light; VAF: Variant allele frequency; VEP: Variant effect predictor; WES: Whole-exome sequencing

Acknowledgements

We would like to thank the donors and their families for their generous gifts of organ donation for transplantation and tissue donations for the GTEx research study.

Authors' contributions

SO, RG, and FM designed the study. FM developed bioinformatics methods and/or performed statistical analysis for mosaic mutation detection, mutational signature analysis, and estimation of mutation rates. LZ performed the evolutionary selection dN/dS analysis. FM and LZ generated the figures. SO and FM wrote the paper with the help of all authors. The authors read and approved the final manuscript.

Funding

The Genotype-Tissue Expression (GTEx) project was supported by the Common Fund of the Office of the Director of the National Institutes of Health (<https://commonfund.nih.gov/GTEx>). This project has received funding from the European Union's H2020 research and innovation programme under grant agreement no. 635290 (PanCanRisk). We acknowledge support by the Faculty of Medicine of the University of Tübingen, the Spanish Ministry of Economy and Competitiveness, "Centro de Excelencia Severo Ochoa 2013-2017," SEV-2012-0208, and the CERCA Programme/Generalitat de Catalunya. LZ is supported by the European Union's Horizon 2020 research and innovation programme under the Marie Skłodowska-Curie Research Fellowship scheme (846614).

Availability of data and materials

The data supporting the conclusions of this article were obtained from the Genotype-Tissue Expression (GTEx) portal [22] and are available in the dbGaP repository, accession phs000424.v7.p2 [23]. GTEx sample IDs used in this study are provided in Additional file 1.

Sequencing data and variants' calls for CLL samples are available at the European Genome-Phenome Archive (EGA), which is hosted by the EBI and the CRG, under accession number EGAS00000000092 [24, 25]. CLL sample IDs used in this study are provided in Additional file 3.

The embryonic expression levels used in this study are available as supplementary information in Yan et al. [44]. Other sets of somatic mutations used for evaluation purposes in this study are available as supplementary information in Yizhak et al. [40], Martincorena et al. [37], and Saini et al. [51]. Somatic and mosaic mutations detected in this study are included in this manuscript (Additional files 6 and 7).

Software used to process the RNA-seq data for somatic and mosaic variant calling is available at <https://github.com/Francesc-Muyas/RnaMosaicMutationFinder> [62].

Ethics approval and consent to participate

Not applicable

Consent for publication

Not applicable

Competing interests

The authors declare that they have no competing interests.

Author details

¹Institute of Medical Genetics and Applied Genomics, University of Tübingen, Tübingen, Germany. ²Center for Genomic Regulation, The Barcelona Institute of Science and Technology, Barcelona, Spain. ³Universitat Pompeu Fabra (UPF), Barcelona, Spain. ⁴Centre for Evolution and Cancer, The Institute of Cancer Research, London, UK.

Received: 3 February 2020 Accepted: 8 May 2020

Published online: 27 May 2020

References

1. Youssoufian H, Peyeritz RE. Mechanisms and consequences of somatic mosaicism in humans. *Nat Rev Genet.* 2002;3:748–58.
2. Biesecker LG, Spinner NB. A genomic view of mosaicism and human disease. *Nat Rev Genet.* 2013;14:307–20 Nature Publishing Group.
3. Acuna-Hidalgo R, Veltman JA, Hoischen A. New insights into the generation and role of de novo mutations in health and disease. *Genome Biol.* 2016;17:241 BioMed Central.
4. Campbell IM, Shaw CA, Stankiewicz P, Lupski JR. Somatic mosaicism: implications for disease and transmission genetics. *Trends Genet.* 2015;31:382–92.
5. Poduri A, Evrony GD, Cai X, Walsh CA. Somatic mutation, genomic variation, and neurological disease. *Science.* 2013;341:1237758 American Association for the Advancement of Science.
6. Halvorsen M, Petrovski S, Shellhaas R, Tang Y, Crandall L, Goldstein D, et al. Mosaic mutations in early-onset genetic diseases. *Genet Med.* 2016;18:746–9.
7. Rivière JB, Mirzaa GM, O'Roak BJ, Beddaoui M, Alcantara D, Conway RL, et al. De novo germline and postzygotic mutations in AKT3, PIK3R2 and PIK3CA cause a spectrum of related megalencephaly syndromes. *Nat Genet.* 2012;44:934–40 Nature Publishing Group.
8. Lindhurst M, Teer JK, Sapp JC, Johnston JJ, Ph D, Finn EM, et al. A mosaic activating mutation in AKT1 associated with the Proteus syndrome. *Genome Res.* 2011;365(7):611–9.
9. Yurov YB, Vorsanova SG, Iourov IY, Demidova IA, Beresheva AK, Kravetz VS, et al. Unexplained autism is frequently associated with low-level mosaic aneuploidy. *J Med Genet.* 2007;44:521–5.
10. Prochazkova K, Pavlikova K, Minarik M, Sumerauer D, Kodet R, Sedlacek Z. Somatic TP53 mutation mosaicism in a patient with Li-Fraumeni syndrome. *Am J Med Genet Part A.* 2009;149:206–11.
11. Ruark E, Snape K, Humburg P, Loveday C, Bajrami I, Brough R, et al. Mosaic PPM1D mutations are associated with predisposition to breast and ovarian cancer. *Nature.* 2013;493:406–10 Nature Publishing Group.
12. Happle R. Lethal genes surviving by mosaicism: a possible explanation for sporadic birth defects involving the skin. *J Am Acad Dermatol.* 1987;16:899–906.
13. PLANT KE, Boye E, Green PM, Vetrie D, Flinter FA. Somatic mosaicism associated with a mild Alport syndrome phenotype. *J Med Genet.* 2000;37:238–9.
14. Acuna-Hidalgo R, Sengul H, Steehouwer M, van de Vorst M, Vermeulen SH, Kiemeneij LALM, et al. Ultra-sensitive sequencing identifies high prevalence of clonal hematopoiesis-associated mutations throughout adult life. *Am J Hum Genet.* 2017;101:50–64 Elsevier.
15. Pham J, Shaw C, Pursley A, Hixson P, Sampath S, Roney E, et al. Somatic mosaicism detected by exon-targeted, high-resolution aCGH in 10 362 consecutive cases. *Eur J Hum Genet.* 2014;22:969–78 Nature Publishing Group.
16. 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 Nature Publishing Group.
17. 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 The American Society of Human Genetics.
18. Conlin LK, Thiel BD, Bonnemenn CG, Medne L, Ernst LM, Zackai EH, et al. Mechanisms of mosaicism, chimerism and uniparental disomy identified by single nucleotide polymorphism array analysis. *Hum Mol Genet.* 2010;19:1263–75.
19. Ju YS, Martincorena I, Gerstung M, Petljak M, Alexandrov LB, Rahbari R, et al. Somatic mutations reveal asymmetric cellular dynamics in the early human embryo. *Nature.* 2017;543:714–8 Nature Publishing Group.
20. Wei W, Keogh MJ, Aryaman J, Golder Z, Kullar PJ, Wilson I, et al. Frequency and signature of somatic variants in 1461 human brain exomes. *Genet Med.* 2018;1 Nature Publishing Group.
21. Lonsdale J, Thomas J, Salvatore M, Phillips R, Lo E, Shad S, et al. The Genotype-Tissue Expression (GTEx) project. *Nat Genet.* 2013;45:580–5.
22. GTEx Portal. <https://gtexportal.org/>. Accessed 8 Nov 2019.
23. GTEx Consortium. Genotype-Tissue Expression (GTEx) data (v7). dbGaP repository, project id phs000424.v7.p2 (https://www.ncbi.nlm.nih.gov/projects/gap/cgi-bin/study.cgi?study_id=phs000424.v7.p2).
24. Puente XS, Beà S, Valdés-Mas R, Villamor N, Gutiérrez-Abril J, Martín-Subero JI, et al. Non-coding recurrent mutations in chronic lymphocytic leukaemia. *Nature.* 2015;526:519–24 Nature Publishing Group.
25. Puente XS, Beà S, Valdés-Mas R, Villamor N, Gutiérrez-Abril J, Martín-Subero JI, et al. Non-coding recurrent mutations in chronic lymphocytic leukaemia. *Nature.* European Genome-Phenome Archive (EGA), accession number EGAS00000000092. (<https://www.ebi.ac.uk/ega/studies/EGAS00000000092>).
26. Zapata L, Susak H, Drechsel O, Friedländer MR, Estivill X, Ossowski S. Signatures of positive selection reveal a universal role of chromatin modifiers as cancer driver genes. *Sci Rep.* 2017;7:13124.
27. Muyas F, Bosio M, Puig A, Susak H, Domènech L, Escaramis G, et al. Allele balance bias identifies systematic genotyping errors and false disease associations. *Hum Mutat.* 2019;40:115–26.
28. Fuentes Fajardo KV, Adams D, Mason CE, Sincan M, Tiff C, Toro C, et al. Detecting false-positive signals in exome sequencing. *Hum Mutat.* 2012;33:609–13 John Wiley & Sons, Ltd.
29. Tan MH, Li Q, Shanmugam R, Piskol R, Kohler J, Young AN, et al. Dynamic landscape and regulation of RNA editing in mammals. *Nature.* 2017;550:249–54.
30. Kiran AM, O'Mahony JJ, Sanjeev K, Baranov PV. Darned in 2013: inclusion of model organisms and linking with Wikipedia. *Nucleic Acids Res.* 2012;41:D258–61.
31. Picardi E, D'Erchia AM, Lo Giudice C, Pesole G. REDportal: a comprehensive database of A-to-I RNA editing events in humans. *Nucleic Acids Res.* 2017;45:D750–7.
32. Yu Y, Xu T, Yu Y, Hao P, Li X. Association of tissue lineage and gene expression: conservatively and differentially expressed genes define common and special functions of tissues. *BMC Bioinformatics.* 2010;11(Suppl 1):S1 BioMed Central.
33. Xing Y, Lerario AM, Rainey W, Hammer GD. Development of adrenal cortex zonation. *Endocrinol Metab Clin North Am.* 2015;44(2):243–74.
34. Chalmers ZR, Connelly CF, Fabrizio D, Gay L, Ali SM, Ennis R, et al. Analysis of 100,000 human cancer genomes reveals the landscape of tumor mutational burden. *Genome Med.* 2017;9:34 BioMed Central.
35. Rosenthal R, McGranahan N, Herrero J, Taylor BS, Swanton C. DeconstructSigs: delineating mutational processes in single tumors distinguishes DNA repair deficiencies and patterns of carcinoma evolution. *Genome Biol.* 2016;17:31 BioMed Central.
36. Zapata L, Pich O, Serrano L, Kondrashov FA, Ossowski S, Schaefer MH. Negative selection in tumor genome evolution acts on essential cellular functions and the immunopeptidome. *Genome Biol.* 2018;19:67.
37. 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 Europe PMC Funders.* 2015;348:880–6.
38. Martincorena I, Fowler JC, Wabik A, ARJ L, Abascal F, Hall MWJ, et al. Somatic mutant clones colonize the human esophagus with age. *Science.* 2018;362(6417):911–17.
39. Yokoyama A, Kakiuchi N, Yoshizato T, Nannya Y, Suzuki H, Takeuchi Y, et al. Age-related remodelling of oesophageal epithelia by mutated cancer drivers. *Nature.* 2019;565:312–7 Nature Publishing Group.
40. Yizhak K, Aguet F, Kim J, Hess JM, Kübler K, Grimbsy J, et al. RNA sequence analysis reveals macroscopic somatic clonal expansion across normal tissues. *Science.* 2019;364:eaaw0726 American Association for the Advancement of Science.
41. Christoforides A, Carpen JD, Weiss GJ, Demeure MJ, Von Hoff DD, Craig DW. Identification of somatic mutations in cancer through Bayesian-based analysis of sequenced genome pairs. *BMC Genomics.* 2013;14:302.
42. Radenbaugh AJ, Ma S, Ewing A, Stuart JM, Collisson EA, Zhu J, et al. RADIA: RNA and DNA integrated analysis for somatic mutation detection. *PLoS One.* 2014;9(11):e111516.
43. Neums L, Suenaga S, Beyerlein P, Anders S, Koestler D, Mariani A, et al. VaDir: an integrated approach to variant detection in RNA. *Gigascience.* 2018;7:1–13 Oxford University Press.
44. Yan L, Yang M, Guo H, Yang L, Wu J, Li R, et al. Single-cell RNA-Seq profiling of human preimplantation embryos and embryonic stem cells. *Nat Struct Mol Biol.* 2013;20:1131–9 Nature Publishing Group.

45. Alexandrov LB, Nik-Zainal S, Wedge DC, Aparicio SAJR, Behjati S, Biankin AV, et al. Signatures of mutational processes in human cancer. *Nature*. 2013;500:415–21 Europe PMC Funders.
46. 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 Europe PMC Funders.
47. Nik-Zainal S, Davies H, Staaf J, Ramakrishna M, Glodzik D, Zou X, et al. Landscape of somatic mutations in 560 breast cancer whole-genome sequences. *Nature*. 2016;534:47–54.
48. García-Nieto PE, Morrison AJ, Fraser HB. The somatic mutation landscape of the human body. *Genome Biol*. 2019;20:298 BioMed Central Ltd.
49. Watson CJ, Papula AL, Poon GYP, Wong WH, Young AL, Druley TE, et al. The evolutionary dynamics and fitness landscape of clonal hematopoiesis. *Science*. 2020;367:1449–54.
50. Signatures of mutational processes in human cancer. https://cancer.sanger.ac.uk/cosmic/signatures_v2. Accessed 8 Nov 2019.
51. Saini N, Roberts SA, Klimczak LJ, Chan K, Grimm SA, Dai S, et al. The impact of environmental and endogenous damage on somatic mutation load in human skin fibroblasts. *PLoS Genet*. 2016;12(10):e1006385.
52. Chawanthayatham S, Valentine CC, Fedeles BI, Fox EJ, Loeb LA, Levine SS, et al. Mutational spectra of aflatoxin B 1 in vivo establish biomarkers of exposure for human hepatocellular carcinoma. *Proc Natl Acad Sci*. 2017;114:E3101–9.
53. Zhang W, He H, Zang M, Wu Q, Zhao H, Lu LL, et al. Genetic features of aflatoxin-associated hepatocellular carcinoma. *Gastroenterology*. 2017;153:249–262.e2 Elsevier, Inc.
54. Koshiol J, Gao YT, Dean M, Egner P, Nepal C, Jones K, et al. Association of aflatoxin and gallbladder cancer. *Gastroenterology*. 2017;153:488–494.e1 W. B. Saunders.
55. Lupski JR. Genetics. Genome mosaicism—one human, multiple genomes. *Science*. 2013;341:358–9.
56. Wan JCM, Massie C, Garcia-Corbacho J, Mouliere F, Brenton JD, Caldas C, et al. Liquid biopsies come of age: towards implementation of circulating tumour DNA. *Nat Rev Cancer*. 2017;17:223–38.
57. Christensen E, Birkenkamp-Demtröder K, Sethi H, Shchegrova S, Salari R, Nordentoft I, et al. Early detection of metastatic relapse and monitoring of therapeutic efficacy by ultra-deep sequencing of plasma cell-free DNA in patients with urothelial bladder carcinoma. *J Clin Oncol*. 2019;37:1547–57 American Society of Clinical Oncology.
58. Newman AM, Lovejoy AF, Klass DM, Kurtz DM, Chabon JJ, Scherer F, et al. Integrated digital error suppression for improved detection of circulating tumor DNA. *Nat Biotechnol*. 2016; Article in press.
59. Alborelli I, Generali D, Jermann P, Cappelletti MR, Ferrero G, Scaggiante B, et al. Cell-free DNA analysis in healthy individuals by next-generation sequencing: a proof of concept and technical validation study. *Cell Death Dis*. 2019;10(7):534.
60. Meddeb R, Dache ZAA, Thezenas S, Otandault A, Tanos R, Pastor B, et al. Quantifying circulating cell-free DNA in humans. *Sci Rep*. 2019;9(1):5220.
61. Spindler KLG, Appelt AL, Pallisgaard N, Andersen RF, Brandslund I, Jakobsen A. Cell-free DNA in healthy individuals, noncancerous disease and strong prognostic value in colorectal cancer. *Int J Cancer*. 2014;135:2984–91 Wiley-Liss Inc.
62. Muyas F, Zapata L, Guigó R, Ossowski S. RnaMosaicMutationFinder. Github. 2019. <https://github.com/Francesc-Muyas/RnaMosaicMutationFinder>.

Publisher's Note

Springer Nature remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.

Ready to submit your research? Choose BMC and benefit from:

- fast, convenient online submission
- thorough peer review by experienced researchers in your field
- rapid publication on acceptance
- support for research data, including large and complex data types
- gold Open Access which fosters wider collaboration and increased citations
- maximum visibility for your research: over 100M website views per year

At BMC, research is always in progress.

Learn more biomedcentral.com/submissions

