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], Stephan Ossowski[1,2,3*]

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

***Corresponding authors:**
Stephan Ossowski (Stephan.ossowski@med.uni-tuebingen.de)
Francesc Muyas (francesc.muyas-remolar@med.uni-tuebingen.de)

Abstract

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 selected tissues and their contribution to genetic disorders is unknown. For this reason, we investigated how frequent mosaic mutations occur during embryogenesis across all germ layers and tissues. Using RNA sequencing (RNA-seq) data from the Genotype-Tissue Expression (GTEx) cohort comprising 49 tissues and 570 individuals, we found that new-borns on average harbour 0.5 - 1 post-zygotic and early embryonic mosaic mutations in coding exons affecting multiple tissues and organs (rate of 1.3225×10^{-8} per nucleotide per individual). We further observed that EMMs are dominated by a mutational signature associated with spontaneous deamination of methylated cytosines and the number of cell divisions. Our findings suggest that EMMs are as frequent as germline *de novo* mutations, could therefore explain a substantial fraction of unsolved sporadic disease entities, and might play a previously underappreciated role in cancer predisposition.

After birth, cells continue to accumulate somatic mutations, which can lead to the development of cancer if key functions such as cell cycle control are affected. 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.

In summary, the analysis of multiple tissues per individual allowed us to distinguish mosaic mutations acquired during different stages of embryogenesis and life. Our results show that embryonic mosaic mutations are frequent and likely play a role in many unsolved genetic disease cases. Hence, their detection needs to be an indispensable part of clinical diagnostics.

Introduction

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 SNParray-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]. 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[19] or blood[17]. Therefore, a comprehensive view of mosaic mutations arising during embryogenesis, including their rate and mutational spectrum, is missing. Here, we exploit RNA sequencing data from 49 different tissues from 570 individuals of the Genotype-Tissue Expression (GTEx) cohort[20] to uncover the rate and spectrum of mosaic mutations acquired post-zygotic during early embryogenesis.

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[21]. We therefore developed a highly accurate multi-sample variant calling procedure leveraging the large number of GTEx tissues and individuals. Our method models errors using a beta binomial distribution in a nucleotide specific manor, removes germline variants and confounders such as RNA editing sites and generates a multi-individual, multi-tissue call matrix (termed 3D-genotype matrix) by re-genotyping potentially variable sites across thousands of RNA-seq samples. Finally, we trained a random forest classifier (*RF-RNAmut*) to distinguish true from false variant calls using whole exome sequencing (WES) and RNA-seq data from the ICGC Chronic Lymphocytic Leukaemia project[22] as training, test and benchmarking data (see Online Methods for details

and Supp. Table 1 for samples). Germline variants found in tumour and normal WES data were identified in RNA-seq data with 86% sensitivity and 95% precision. High confidence somatic variant calls with >0.15 VAF in tumour WES data were identified in RNA-seq with 71% sensitivity and 85% precision. Sensitivity of our method for somatic variants is positively correlated with the VAF found in WES data (Supp. Table 2 shows results for minimum VAF ranging from 0.05 to 0.2).

Rate and spectrum of early mosaic mutations during embryogenesis

In order to identify mosaic mutations acquired during embryonic development, we computed the 3D genotype matrix for 9,704 samples of the GTEx cohort comprising 526 cancer-free individuals and 49 different tissues. We contrasted the 3D genotype matrix with the embryogenesis lineage tree (Supp. Fig. 1 and Supp. Table 3) to identify the most likely stage and tissue or germ layer of origin of each mutation. Mosaic mutations identified in both the Ectoderm and Mesendoderm, which likely occurred during the first few divisions of the zygote (cleavage, blastulation, implantation), were defined as postzygotic mosaic mutations (PMMs). Variants occurring in all expressed tissues with an average VAF greater 0.35 were excluded from PMM prediction, as they might constitute *de novo* germline variants. Mutations found in at least two tissues of the same individual, for which exactly one tissue or germ layer of origin could be determined in the lineage tree that is not zygote, were defined as early embryonic mosaic mutations (EEMMs). EEMMs most likely occurred during neurolation, gastrulation and early organogenesis, but after implantation. Contradictory calls, i.e. mutations requiring the hypothesis to have been independently acquired by more than one tissue in different branches of the lineage tree, were removed as likely false positives (see Online Methods for details on PMM and EMM definition).

To minimise false negatives, we focused our analysis on housekeeping genes constitutively expressed in the majority of tissues and samples (7,630 genes with TPM > 5 in at least 75% of tissues, see figure 1 and Supp. Table 4 for other thresholds). After strict filtering of germline variants found per individual by WES or in the population (GnomAD), as well as predicted or known RNA editing sites we identified 58 putative PMMs and 37 EEMMS in 7630 constitutively expressed genes. We estimated a rate of 8.1164×10^{-9} (**CI (95%)** = [8.1018×10^{-9} to 8.1309×10^{-9}]) postzygotic mosaic mutations per nucleotide and individual and a rate of 5.1089×10^{-9} (**CI (95%)** = [5.0977×10^{-9} to 5.1200×10^{-9}]) early embryonic mosaic mutations per nucleotide and individual for human coding regions. Following an approach for extrapolating tumour mutation burden (TMB) from gene panels to exomes (45Mbp coding region)[23], we estimated a mean of 0.37 coding PMMs (Fig. 1A) and 0.23 coding EEMMs (Fig 1B) 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 PMMs or EEMMs (Figure 1A and B).

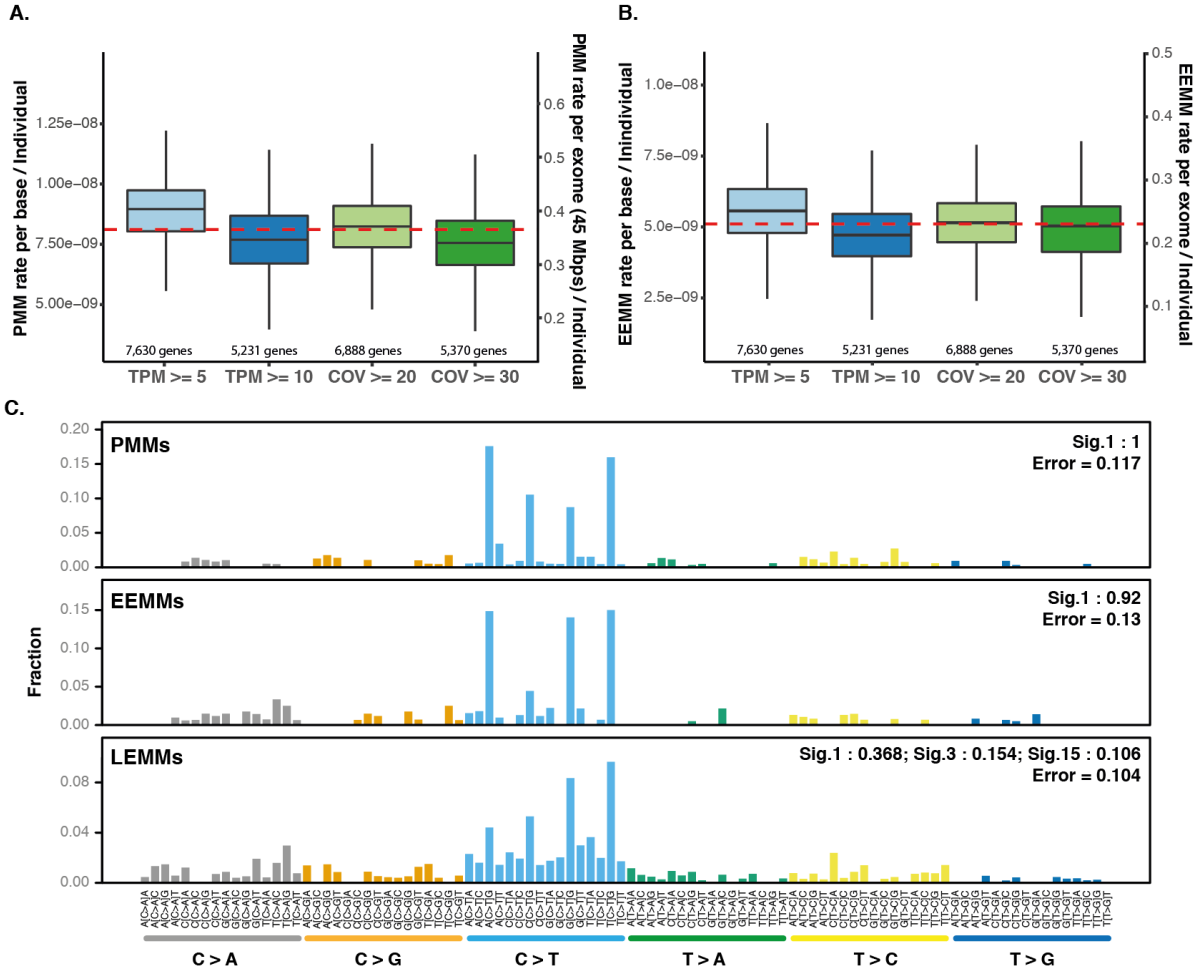

Figure 1. Rate, spectrum and mutational signatures of mosaic mutations in healthy individuals acquired during embryogenesis. For A and B, the left Y-axis represents the mutational rate per nucleotide, the right Y-axis represents the extrapolated number of mosaic mutations expected in 45 Mbps coding exons, 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 postzygotic mosaic mutations (PMM) acquired during the first few divisions of the zygote. We estimated a mean rate of PMMs per base and individual of 8.1164 x 10^{-9} (**CI (95%)** = [8.1018 x 10^{-9} to 8.1309 x 10^{-9}]). B) Early embryonic mosaic mutations (EEMM) affecting at least 2 tissues. We estimated a mean of 5.1089 x 10^{-9} EEMMs per nucleotide and individual ((**CI (95%)** = [5.0977 x 10^{-9} to 5.1200 x 10^{-9}]). C) Mutational signature of postzygotic (PMMs), early embryonic (EEMMs) and late embryonic mosaic mutations (LEMMs).

On average, a specific coding PMM was detectable in 63.6% of the tissues of an individual, which expressed the respective gene, consistent with the assumption that they arose during the first divisions of the zygote. Interestingly, only 41% of PMMs in genes expressed in blood were detectable in blood samples, 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 and the number of tissues supporting the variant (Spearman rank correlation Rho=0.4944; p-value=0.0019). Moreover, mutations occurring earlier in development also showed a greater proportion of cells carrying the variant (Rho=-0.33; p-value=0.046, Supp. Fig. 2).

Our estimated rate of postzygotic and early embryonic mosaic mutations of 1.3225x 10^{-8} (rate of EMMs + PMMs) is surprisingly close 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, or ~0.5 – 1 mutations per 45Mbp coding exome per individual). Recently, several rare disease studies based on WES indicated that between more than 50% of sporadic cases can be explained by *de novo* germline mutations, especially in common neurodevelopmental disorders[3,5]. Consequently, embryonic mosaic mutations are similarly likely to explain a significant fraction of sporadic rare disease cases, and a substantial fraction of *de novo* variants predicted by parent-child trio analysis of using blood are potentially postzygotic mutations. Moreover, we potentially underestimate the rate of embryonic mosaic mutations due to factors such as the limited sensitivity of RNA-based somatic variant calling, allele specific expression, nonsense-mediated decay, more effective transcription-coupled repair in highly expressed genes, and the limited number of tissues studied in GTEx. 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 clinical diagnostics, and could explain a substantial part of the missing heritability. In summary, our results indicate that a significant fraction of the undiagnosed cases of sporadic genetic diseases could be caused by postzygotic or early embryonic mosaic mutations.

In order to identify the most likely processes causing mosaic mutations during embryogenesis we investigated their mutational signatures. We found that a large fraction of PMMs and EEMMs (1 and 0.92) could be explained by Signature 1[24–26] (figure 1B). 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 is often associated with aging and the number of cell divisions[24]. Hence, our findings indicate that most postzygotic and early embryonic mosaic mutations occur spontaneously with very limited contributions from exposure to environmental factors or other endogenous processes (such as APOBEC). Furthermore, our results clearly distinguish early mosaic mutations from germline *de novo* mutations, which are dominated by Signature 5 characterised by A>G transitions[3].

Late embryonic mosaic mutations arising during organogenesis

Our definitions of EEMMs requiring their identification in at least two tissue types prevents identification of organ-specific mutations occurring during organogenesis, which is starting during weeks 3 to 8 of embryogenesis depending on the organ. We therefore screened for late embryonic mosaic mutations (LEMMs), which we defined as variants identified only in one tissue type but which were present in a large fraction of cells in that tissue (VAF >= 0.2). Here we excluded tissues previously shown to be affected by clonal expansion of mutated cells such as esophagus-mucosa, sun-exposed skin[21,23,27–29] and whole blood[14]. We identified

377 mutations across all individuals, considering any gene expressed in at least one tissue (Supp. Table 5), resulting in an estimate of 2.44 x 10^{-9} (CI [0.95] = [1.86 x 10^{-9} - 3.03 x 10^{-9}]) LEMMs per nucleotide per tissue per individual, and extrapolating to 0.11 (CI [0.95] = [0.084 - 0.137]) mutations per coding exome (45Mbps) per tissue. We found substantial differences between tissues, with brain tissues showing the lowest rates of late mosaic mutations (Supp. Fig 3). Notably, the average rate of LEMMs (2.23 x 10^{-9}) for brain tissues closely resembles the estimate by Wei *et al.* 2018[30] (2.55 x 10^{-9}) obtained using WES data of brain tissues. In sum across all 43 examined tissues we estimate 4.7 LEMMs per coding exome per individual. However, LEMMs are indistinguishable from mutations in clonal expansions acquired after birth[21,27,29], and the rate of LEMMs is possibly 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 disease-relevant somatic mutation processes in the GTEx cohort, we next studied mutational signatures across all tissue-specific somatic variants identified by the RF model. We first aimed at reproducing somatic mutation rates, signatures of positive selection and various mutational processes published recently by Martincorena *et al.* and Yizhak *et al*[21,27,29]*.* Considering only variants with VAF >= 0.05, *RF-RNAmut* identified 8,780 somatic mutations in 8,351 samples representing 46 tissues (see Online Methods and Supp. Fig. 4 for details on sample QC, batch effect removal and somatic variant calling). Estimated mutation rates per tissue ordered by nucleotide per individual are shown in Supp. Fig. 5 and closely resembles the order obtained by Yizhak *et al.* on a previous release of GTEx[21]. After removal of technical confounders (PCR duplicate rates, RIN, TRISCHD, coverage, sequencing center) using a linear regression model we observed the highest mutation rates for sun-exposed skin, lung, testis, esophagus-mucosa and vagina (Figure 2A), as previously reported[21]. As expected, sun-exposed skin showed significantly higher mutation rates than non-sun-exposed skin, while brain tissues showed the lowest somatic mutation rates. Finally, we tested if residual mutation rates were related with age of individuals at death for each tissue individually (Fig. 2B). Only two tissues showed a significant association between age and mutational rates (after FDR correction), namely sun-exposed skin (Rho = 0.31; qval = 1.19 x 10-7) and esophagus-mucosa (Rho = 0.22; qval = 2.82 x 10-3), as previously reported[27–29].

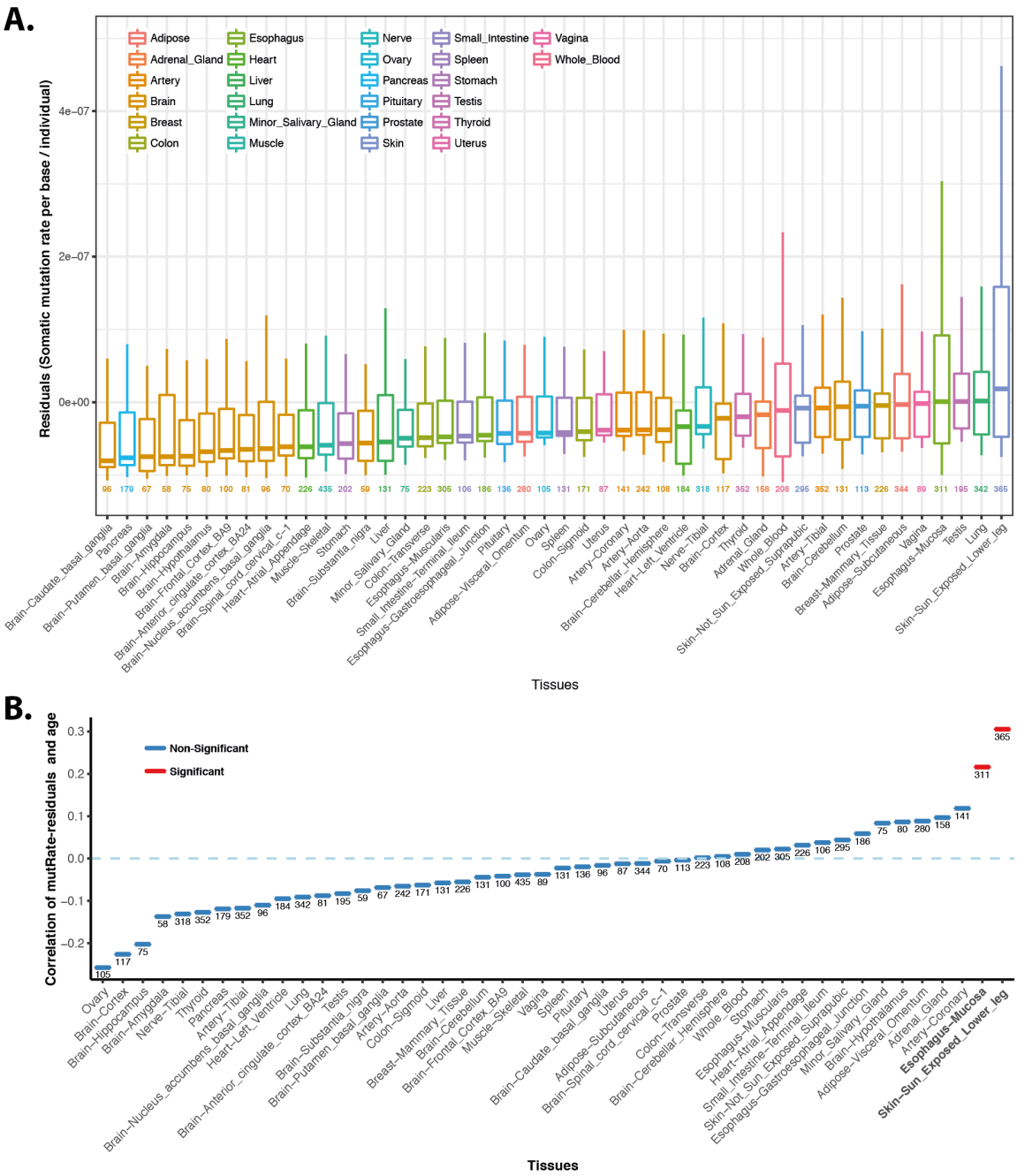

Figure 2. Rate of somatic mutations varies significantly across the 46 tissues of the GTEx cohort (ignoring Kidney, Cells-EBV-transformed lymphocytes and Cells-transformed fibroblasts for technical reasons, see Methods). A) Distribution of residuals of the somatic mutation rate per base and individual residuals (mutRate-residuals) across analysed tissues. MutRate-residuals represent the somatic mutation rates corrected for non-biological confounders such as PCR duplication rate, RIN, cohort, and read coverage. B) Spearman correlation between mutRate-residuals and age per tissue. Colours show the significance of the correlation test after FDR correction (q-value < 0.05 in red). Numbers below bars represent the number of analysed individuals per tissue.

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 strong positive selection for sun-exposed skin and esophagus-

mucosa (Supp. Fig. 6). Mutations in *NOTCH1* and *TP53* disproportionally contributed to the high dN/dS value and showed the highest overall mutation rate. *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, see Supp. Table 6). Interestingly, we did not find 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 healthy esophagus and skin[21,27–29], identifying mutational signatures[24] 1, 5 and 7 (see https://cancer.sanger.ac.uk/cosmic/signatures/SBS). 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, while this signature was absent from non-sun-exposed skin (figure 3). Interestingly, studies of the mutational signatures found in healthy tissues forming the dietary tract are lacking, although the constant exposure to food likely leads to a particular mutational spectrum. We therefore performed a pan-dietary-tract mutational signature analysis considering colon, esophagus-mucosa, liver, small intestine and stomach. Apart from signatures 1 and 5, which are frequently observed in most cancer 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 gastrointestinal tract (figure 3). Furthermore, we saw a strong enrichment of the characteristic CGN > CTN mutations not observed in any other tissue.

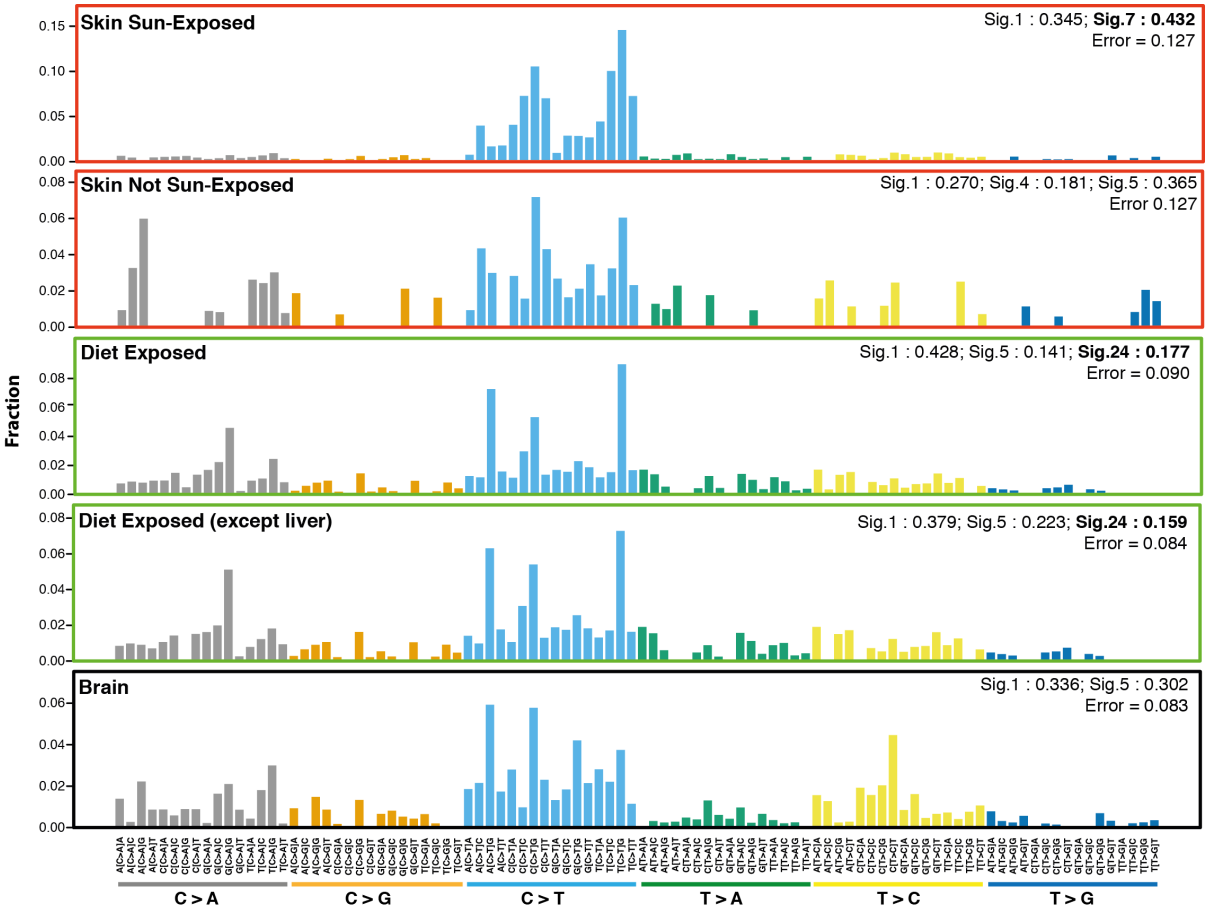

Figure 3. Mutational signatures observed in (A) sun-exposed skin, (B) non-sun-exposed skin, (C) diet exposed tissues, (D) diet exposed tissues without liver, and (E) brain tissues (considered control tissues with low exposure to environmental factors). Detailed descriptions of signatures are available at https://cancer.sanger.ac.uk/cosmic/signatures/SBS/.

Aflatoxin B_1 (AFB1) is a potent mutagen and carcinogen typically found in grains contaminated with the food spoilage fungus, *Aspergillus flavus*. Dietary exposure to aflatoxin B_1 (AFB1) is a known risk factor for human hepatocellular carcinoma (HCC), the third leading cause of cancer death worldwide. Upon ingestion, it is rapidly absorbed from the digestive tract, transported to the liver, and absorbed by liver cells[31]. One of the degradation products, the metabolite exo-epoxide, 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[32,33], but has not been reported for other cancer entities. We therefore tested, if the observed enrichment of signature 24 is solely introduced by a strong mutagenic effect in the liver. To the contrary, even excluding liver from the signature analysis, the aflatoxin signature was still revealed at a similar level, explaining a fraction of close to 0.16 of the mutational spectrum. In comparison, we did not identify signature 24 in any other tissue, e.g. brain tissues or skin (figure 3). Our results indicate that aflatoxins are responsible for a substantial fraction of somatic mutations in all tissues that are in contact with food, such as colon, esophagus-

mucosa, liver, small intestine and stomach, and might be responsible for the development of cancers in several organs of the dietary tract.

Discussion

We presented a novel analysis strategy using RNA-seq data of multiple tissues of an individual to identify mosaic mutations occurring during various stages of embryo development. Our method can approximate the time point of the event as well as the affected germ layer or developing organ. We demonstrate how to distinguish, to some extent, embryonic mosaic mutations from *de novo* germline mutations and somatic mutations in clonal expansions acquired after birth.

Analysing RNA sequence data from 49 tissues and 570 patients we found that new-borns on average harbour 0.5 - 1 mosaic mutation in coding exons affecting multiple tissues and organs, and likely an even larger number of organ-specific coding mutations. We further observed that postzygotic 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. 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-recognised role of mosaic mutations in genetic disease would have many implications for genetic diagnostics procedures[34]. We have demonstrated that a substantial fraction of postzygotic, early and late embryonic mutations (PMMs, EEMMs and LEMMs) identified in the GTEx cohort are not detectable in blood cells. Samples from the tissue of origin of a genetic disease are often not available, hence other diagnostic methods should be considered. Circulating cell free DNA could be an unbiased source for detecting mosaic mutations, using cfDNA-sequencing methods ('liquid biopsy'), which have been successfully applied in cancer diagnostics for several years.

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 tumours in colon, esophagus-mucosa, liver, small intestine and stomach.

Methods

Samples

In this study we used release 7 of the Genotype-Tissue Expression (GTEx)[20] project (dbGaP accession phs000424.v7.p2), including RNA-seq data for 49 tissues from 570 individuals. We included only individuals for which Whole Exome Sequencing (WES) 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 (Supp. Table 3). 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 2.7) against the human reference genome (GRCh37) and the resulting BAM files were post-processed in order to remove alignment artefacts. 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-matrix (variant x tissue x 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) and afterwards post-processed 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.

$$\text{Alternative counts} \sim Bin(\text{Coverage, error rate})$$
$$\text{Error rate} \sim Beta(\alpha, \beta)$$

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 matrix. 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, resulting in one of four possible filter states per sample: NO_EXPRESSION, NO_CALL, LOW_QUALITY or PASS. Furthermore, the exact reference-like and alternative allele counts are stored in the Coordinates x Tissue x Individual matrix.

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, for which whole exome sequencing (WES) data for tumor and normal sample and RNA-seq data for tumor samples are available. 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 predicted high quality somatic and germline variant calls using GATK HaplotypeCaller and MuTect2 as described before[35,36].

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 set 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 sensitivity of the RF for low VAF)

In addition, a set of non-overlapping high quality calls from WES data was incorporated in training and test sets. We labeled as true variants any site with VAF >= 5% and at least 2 alternative reads in WES data, and all other sites as false variants. This procedure resulted in training and test datasets 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[37], 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 labelled 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) were flagged as likely systematic analysis error by ABB[36], or (6) they overlapped with a known RNA editing site[38–40].

We measured the performance (precision and recall) of the 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.

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 WES analysis in any individual (GATK HaplotypeCaller) from the 3D genotype matrix. 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 or RNA editing sites. 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/[38], found in the Darned database (https://darned.ucc.ie/download/[39]) or identified by the GTEx consortium (http://srv00.recas.ba.infn.it/atlas/pubs.html - REDIportal[40]).

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[41]. 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 postzygotic mosaic mutations (PMMs) and early embryonic mosaic mutations (EEMMs)

In order to identify mosaic mutations acquired during early embryogenesis (cleavage, blastulation, gastrulation, neurolation and early organogenesis) we contrasted the somatic calls in the 3D genotype matrix with a lineage tree of human embryogenesis and tissue development including the 49 tissues studied here (Supp. Fig. 1)[42]. 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 Supp. Table 3). 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 postzygotic mosaic mutations (PMMs). In order to avoid detection of *de novo* germline variants as PMMs we only considered variants with VAF less than 0.35 that were not found in all expressed tissues of an individual.

The remaining mutations found in at least two tissues of an individual were defined as early embryonic mosaic mutations (EEMMs) 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 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[23]. 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 (7,630 genes).
2. Genes with TPM >= 10 in more than 75 % of the total samples (5,231 genes).
3. Genes with COV >= 20 in than 75 % of the total samples (6,888 genes).
4. Genes with COV >= 30 in more than 75 % of the total samples (5,370 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 calculates 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 (45Mbp) in order to calculate the number of mosaic coding mutations expected on average for a newborn child. The procedure was independently performed for PMMs and EEMMs.

On the other side, for LEMMs, as they are tissue and individual specific, we checked the tissue-individual expression for each sample. Then we calculated the number of mutations per base and individual relative to the tissue-individual interrogated region and extrapolated this value to the approximate total length of all coding exons (45Mbp). 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, signature of selection, as well as mutation spectra in a tissue specific manor, 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 only in one tissue per patient in order to minimize the number of mosaic mutations acquired before birth in this set. First, we performed samples-wise quality control (Supp. Fig. 4) 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 kidney with 38 high quality samples, see Supp. Figure 4D), resulting in 8,351 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:

$$\text{Mutation Rate} \sim \text{Duplicates} + \text{Cohort} + \text{RIN} + \text{TRISCHD} + DP_{median} + \varepsilon \text{ (mutRate residuals)}$$

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*[43] 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 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 founds in exons of protein-coding genes due to the limited RNA-seq coverage outside of genes. 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[44], which calculates the trinucleotide-corrected ratio of nonsynonymous to synonymous mutations from NGS data[44]. Somatic mutations identified by *RF-RNAmut* were annotated using 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-tissues 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[29] and 995 essential genes[44]. Finally, we focused on NOTCH1 and TP53 genes in order to replicate the findings of strong positive selection described recently[21,27–29].

Funding

Funding The study was funded by the Faculty of Medicine of the University of Tübingen, the Spanish Ministry of Economy and Competitiveness, the Centro de Excelencia Severo Ochoa, the CERCA Programme.

References

1. Youssoufian, H. & Pyeritz, R. E. Mechanisms and consequences of somatic mosaicism in humans. *Nat. Rev. Genet.* **3,** 748–758 (2002).
2. Biesecker, L. G. & Spinner, N. B. A genomic view of mosaicism and human disease. *Nat. Rev. Genet.* **14,** 307–320 (2013).
3. Acuna-Hidalgo, R., Veltman, J. A. & Hoischen, A. New insights into the generation and role of de novo mutations in health and disease. *Genome Biol.* **17,** 241 (2016).
4. Campbell, I. M., Shaw, C. A., Stankiewicz, P. & Lupski, J. R. Somatic mosaicism: implications for disease and transmission genetics. *Trends Genet.* **31,** 382–392 (2015).
5. Poduri, A., Evrony, G. D., Cai, X. & Walsh, C. A. Somatic mutation, genomic variation, and neurological disease. *Science* **341,** 1237758 (2013).
6. Halvorsen, M. *et al.* Mosaic mutations in early-onset genetic diseases. *Genet. Med.* **18,** 746–749 (2016).
7. Rivière, J. B. *et al.* De novo germline and postzygotic mutations in AKT3, PIK3R2 and PIK3CA cause a spectrum of related megalencephaly syndromes. *Nat. Genet.* **44,** 934–940 (2012).
8. Lindhurst, M. *et al.* A Mosaic Activating Mutation in. *Genome Res.* 611–619 (2011).
9. Yurov, Y. B. *et al.* Unexplained autism is frequently associated with low-level mosaic aneuploidy. *J. Med. Genet.* **44,** 521–525 (2007).
10. Prochazkova, K. *et al.* Somatic TP53 mutation mosaicism in a patient with Li-Fraumeni syndrome. *Am. J. Med. Genet. Part A* **149,** 206–211 (2009).
11. Ruark, E. *et al.* Mosaic PPM1D mutations are associated with predisposition to breast and ovarian cancer. *Nature* **493,** 406–410 (2013).
12. Happle, R. Lethal genes surviving by mosaicism: a possible explanation for sporadic birth defects involving the skin. *J. Am. Acad. Dermatol.* **16,** 899–906 (1987).
13. PLANT, K. E., Boye, E., Green, P. M., Vetrie, D. & Flinter, F. A. Somatic mosaicism associated with a mild Alport syndrome phenotype. *J. Med. Genet.* **37,** 238–239 (2000).
14. Acuna-Hidalgo, R. *et al.* Ultra-sensitive Sequencing Identifies High Prevalence of Clonal Hematopoiesis-Associated Mutations throughout Adult Life. *Am. J. Hum. Genet.* **101,** 50–64 (2017).
15. Pham, J. *et al.* Somatic mosaicism detected by exon-targeted, high-resolution aCGH in 10 362 consecutive cases. *Eur. J. Hum. Genet.* **22,** 969–978 (2014).
16. Huang, A. Y. *et al.* Postzygotic single-nucleotide mosaicisms in whole-genome sequences of clinically unremarkable individuals. *Cell Res.* **24,** 1311–1327 (2014).
17. Acuna-Hidalgo, R. *et al.* Post-zygotic Point Mutations Are an Underrecognized Source of de Novo Genomic Variation. *Am. J. Hum. Genet.* **97,** 67–74 (2015).
18. Conlin, L. K. *et al.* Mechanisms of mosaicism, chimerism and uniparental disomy identified by single nucleotide polymorphism array analysis. *Hum. Mol. Genet.* **19,** 1263–1275 (2010).
19. Wei, W. *et al.* Frequency and signature of somatic variants in 1461 human brain exomes. *Genet. Med.* (2018). doi:10.1038/s41436-018-0274-3
20. Lonsdale, J. *et al.* The Genotype-Tissue Expression (GTEx) project. *Nat. Genet.* **45,** 580–585 (2013).
21. Yizhak, K. *et al.* RNA sequence analysis reveals macroscopic somatic clonal expansion

across normal tissues. *Science* **364,** eaaw0726 (2019).
22. Puente, X. S. *et al.* Non-coding recurrent mutations in chronic lymphocytic leukaemia. *Nature* **526,** 519–524 (2015).
23. Chalmers, Z. R. *et al.* Analysis of 100,000 human cancer genomes reveals the landscape of tumor mutational burden. *Genome Med.* **9,** 34 (2017).
24. Alexandrov, L. B. *et al.* Signatures of mutational processes in human cancer. *Nature* **500,** 415–21 (2013).
25. Alexandrov, L. B. *et al.* Clock-like mutational processes in human somatic cells. *Nat. Genet.* **47,** 1402–7 (2015).
26. Nik-Zainal, S. *et al.* Landscape of somatic mutations in 560 breast cancer whole-genome sequences. *Nature* **534,** 47–54 (2016).
27. Martincorena, I. *et al.* Somatic mutant clones colonize the human esophagus with age. *Science (80-.).* eaau3879 (2018). doi:10.1126/SCIENCE.AAU3879
28. Yokoyama, A. *et al.* Age-related remodelling of oesophageal epithelia by mutated cancer drivers. *Nature* **565,** 312–317 (2019).
29. Martincorena, I. *et al.* Tumor evolution. High burden and pervasive positive selection of somatic mutations in normal human skin. *Science* **348,** 880–6 (2015).
30. Wei, W. *et al.* Frequency and signature of somatic variants in 1461 human brain exomes. *Genet. Med.* 1 (2018). doi:10.1038/s41436-018-0274-3
31. Kumar, P., Mahato, D. K., Kamle, M., Mohanta, T. K. & Kang, S. G. Aflatoxins: A Global Concern for Food Safety, Human Health and Their Management. *Front. Microbiol.* **7,** 2170 (2016).
32. Chawanthayatham, S. *et al.* Mutational spectra of aflatoxin B 1 in vivo establish biomarkers of exposure for human hepatocellular carcinoma . *Proc. Natl. Acad. Sci.* **114,** E3101–E3109 (2017).
33. Zhang, W. *et al.* Genetic Features of Aflatoxin-Associated Hepatocellular Carcinoma. *Gastroenterology* **153,** 249–262.e2 (2017).
34. Lupski, J. R. Genetics. Genome mosaicism--one human, multiple genomes. *Science* **341,** 358–9 (2013).
35. Zapata, L. *et al.* Signatures of positive selection reveal a universal role of chromatin modifiers as cancer driver genes. *Sci. Rep.* **7,** 13124 (2017).
36. Muyas, F. *et al.* Allele balance bias identifies systematic genotyping errors and false disease associations. *Hum. Mutat.* **40,** 115–126 (2019).
37. Fuentes Fajardo, K. V. *et al.* Detecting false-positive signals in exome sequencing. *Hum. Mutat.* **33,** 609–613 (2012).
38. Tan, M. H. *et al.* Dynamic landscape and regulation of RNA editing in mammals. *Nature* **550,** 249–254 (2017).
39. Kiran, A. M., O'Mahony, J. J., Sanjeev, K. & Baranov, P. V. Darned in 2013: inclusion of model organisms and linking with Wikipedia. *Nucleic Acids Res.* **41,** D258–D261 (2012).
40. Picardi, E., D'Erchia, A. M., Lo Giudice, C. & Pesole, G. REDIportal: a comprehensive database of A-to-I RNA editing events in humans. *Nucleic Acids Res.* **45,** D750–D757 (2017).
41. Muyas, F. *et al.* Allele balance bias identifies systematic genotyping errors and false disease associations. *Hum. Mutat.* **40,** 115–126 (2019).
42. 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* **11 Suppl 1,** S1 (2010).
43. Rosenthal, R., McGranahan, N., Herrero, J., Taylor, B. S. & Swanton, C. DeconstructSigs: delineating mutational processes in single tumors distinguishes DNA repair deficiencies and patterns of carcinoma evolution. *Genome Biol.* **17,** 31 (2016).
44. Zapata, L. *et al.* Negative selection in tumor genome evolution acts on essential cellular functions and the immunopeptidome. *Genome Biol.* **19,** 67 (2018).