



OPEN Monitoring the rate and variability of somatic genomic alterations using long-read sequencing

Xingyao Chen¹, Hagai Ligumsky¹, Charlie Ambrose¹, Denisse Sibrian¹, Brian Tran¹, Daoud Arif¹, Olga Castellanos¹, Darren Kessner¹, Hanyi Luo¹, Mukta Ubale¹, Abigail Coleman¹, Vaidhyanathan Mahaganapathy¹, Thomas J. Jönsson¹, Reva K. Basho^{1,2}, Jerry S. H. Lee^{1,3,4,5}, Naim Matasci^{1,6}✉ & David B. Agus^{1,3,7}✉

Cancer initiation occurs when a cell acquires and accumulates mutations in genes involved in the regulation of cell processes: each cell division throughout a person's life introduces novel mutations in the cells' DNA and under normal circumstances, the body is primed to prevent those from leading to cancer. Occasionally, a subset of those mutations escapes those safeguards and might eventually result in the emergence of the disease. To understand the dynamics of accumulation of somatic mutations, we have performed longitudinal whole genome sequencing of DNA obtained from whole blood from healthy individuals and cancer patients using Oxford Nanopore Technologies' Long Read Sequencing. Here we show that the number of somatic single nucleotide variants detected increases with their age and that for specific mutational processes, changes can be detected within months. We computed aggregated metrics for unique participants at each timepoint across types of variants (based on single based substitution molecular signatures) and identified patterns of change both over an individual's lifespan (age) and over the sampling period (months). This study showcases the suitability of long read sequencing of blood DNA for detecting coarse-grained differences over time and enable future development of "state of the system" personalized prevention programs.

Throughout the aging process, healthy cells accumulate somatic mutations due to intrinsic and extrinsic forces, a process that varies between both individuals and tissues¹. Studies have repeatedly shown that mutations involved in the regulation of cell processes are associated with the development of cancer, sometimes years before the appearance of the disease^{2–5}. For example, mutations present in both healthy skin and neighboring cancerous lesions sharing a mutational signature associated with UV damage have been demonstrated to accumulate with age^{6,7}. The human body is primed to prevent the accumulation of these mutations and the formation of cancer through critical cellular regulation processes^{8,9}. However, when the body is under acute or chronic stress, these safeguarding processes may become less dependable, and the affected tissues may be more susceptible to developing cancer¹⁰. Decades of research have revealed deficient mechanisms of DNA repair and replication in mutagenesis, yet surprisingly little is known about the rate of accumulation of somatic mutations caused by normal cellular processes over an individual's lifetime and its fluctuations in response to intrinsic or extrinsic factors.

Clonal hematopoiesis (CH) can result from mutations caused by environmental triggers (e.g. smoking or cytotoxic therapies), errors in DNA replication, or a combination of both extrinsic stress and intrinsic repair rate and which yield a competitive growth advantage leading to detectable clonally expanded populations of hematopoietic stem cells and peripheral blood¹¹. While CH can result in hematologic malignancy, cooperative mutations in additional genes are required to induce malignant transformation¹². Population-scale data have revealed that on average, ~ 1.3 somatic exonic mutations are acquired per hematopoietic stem cell per decade¹³. In addition, as hematopoietic cells travel to different tissues of the body, affecting homeostasis among a number

¹Ellison Medical Institute, Los Angeles, CA, USA. ²Providence Saint John's Cancer Institute, Santa Monica, CA, USA.

³Department of Medicine, Keck School of Medicine, University of Southern California, Los Angeles, CA, USA. ⁴Mork Family Department of Chemical Engineering and Materials Science, Viterbi School of Engineering, University of Southern California, Los Angeles, CA, USA. ⁵Department of Quantitative and Computational Biology, Dornsife College of Letters, Arts and Sciences, University of Southern California, Los Angeles, CA, USA. ⁶Department of Computer Science, Harvey Mudd College, Claremont, CA, USA. ⁷Alfred E. Mann Department of Biomedical Engineering, Viterbi School of Engineering, University of Southern California, Los Angeles, CA, USA. ✉email: nmatasci@emila.org; agus@emila.org

of different organ systems, recent studies have demonstrated the potential for these altered immune cells to impact risk for cardiovascular disease, diabetes, cancer, and other diseases¹⁴. Though CH was found to be associated with aging, most studies investigating CH and its clinical significance captures only a snapshot of the somatic genome landscape in time, making it hard to evaluate the dynamics of somatic genomes and its effect on human health. Longitudinal studies can elucidate pattern of clonal expansion¹⁵ and provide insights into the origin of particular cell populations¹⁶. Therefore, investigating the rate of accumulation of the somatic alterations, its dynamics and its association with clinical manifestations may improve understanding of disease onset, uncovering a new paradigm for monitoring general health. Initial efforts of sequencing peripheral blood DNA in a cohort of 12,380 Swedish patients was among the first to identify a large fraction of CH carriers that did not have known disease driving mutations¹⁷, and subsequently, similar results reported in ~11,262 Icelandic patients using short-read sequencing¹⁸. These studies emphasize the need to better understand the process and rate of accumulation of somatic mutations and their effect on general health.

The introduction of advanced molecular profiling platforms enabled researchers to start addressing the contribution of somatic mutation rates and of endogenous and exogenous mutational processes across normal tissues to disease phenotypes. For example, a study using long-read sequencing (LRS) of whole blood and heart tissue from ~3622 patients provided preliminary evidence that application of LRS can further our understanding of the effects of sequence variation on disease phenotype, particularly genomic regions thus far inaccessible using short-read sequencing alone¹⁹.

These studies, however, fail to assess the dynamic aspect of mutation accumulation and whether the rate of accumulation is related to environmental factors. Longitudinal sampling can help improve our understanding of the dynamics of mutational changes overtime, and ultimately the impact of alterations on the development of disease. The use of LRS technology provides several advantages over traditional Next Generation approaches²⁰, including delivering richer data (methylation, structural variants) and more extensive coverage in hard-to-sequence regions²¹. While it is currently still not cost competitive, the ease of use makes it highly suitable for medium scale sequencing studies where laboratory conditions require tight control. Moreover, the possibility to computationally determine specific regions of the genome be included (or excluded) available with Oxford Nanopore Technologies make it especially attractive for designing flexible assays that target specific areas of the genome.

Thanks to the convergence of technology and molecular knowledge we can now start characterizing the accumulation of somatic mutations and measuring its rate of change over time to serve as a surrogate for the state of the system of an individual, and to inform clinical course trajectories through this monitoring, and ultimately to evaluate useful interventions for changing their outcome.

In this study, we hypothesize that an individual's rate of accumulation of mutations, in particular somatic single nucleotide variants (sSNVs), can be measured over time using LRS from whole blood. Data generated and analysis of variability will be used to develop a model and inform statistical plans for future validation studies, and correlation to cancer risk determination. Improved understanding of which cohorts of patients could benefit from targeted molecular profiling could enable a precision-medicine approach to cancer risk reduction.

Results

Long read sequencing can be used to identify and characterize low frequency single nucleotide variants in a contrived sample

To determine whether our approach can capture and characterize low-frequency single nucleotide variants, we performed a contrived sample experiment in which increasing amounts of DNA from a well characterized cancer cell line (BT-474) were spiked into a known genomic DNA background (NIST Genome in a Bottle reference sample NA12878/HG001). BT-474 is a breast cancer-derived cell line characterized by APOBEC activity and displaying an enrichment of single nucleotide variants matching the mutational signatures SBS2 and SBS13²². At 10% w/w BT-474 DNA concentration, our pipeline was able to recover an average of 2133 SNVs (SEM 263), corresponding to 2.9% of the number observed in a pure sample of BT-474 DNA sequenced on the same instrument (71,931 SNVs common to two replicates). At 5% and 1% BT-474 concentration, the proportions recovered are 1.5% and 1.2% respectively, indicating a linear recovery between 5 and 10% and a drop below accurate quantification below 5%. We also observed a significant linear relationship between the recall (true positives over sum of true positives and false negatives) and the percentage of exogenous DNA (Fig. 1A, linear regression model: slope=0.06, $t=5.00$, p value=5.4 E-4, $n=12$). While the overall pattern demonstrates our ability to detect lower frequency variants, only a small proportion of the actual variants was recovered. This is likely due to the filtering step in our analysis pipeline (Supplementary Figure S2), which is designed to aggressively filter out inherited (germline) variants and low-quality variants.

Because of the presence of specific strong and distinct signatures, compared to other cell lines with more diffuse mutational patterns, we can use this cell line to test our ability not only to recover specific signature at low frequencies. At BT-474 DNA concentration above 1%, our signature analysis shows significant enrichment of signatures SBS2 and SBS13 whereas all other signatures remain undetected or at much lower level, with the exception of SBS1 and SBS5 (Fig. 1B), which are ubiquitous, clock-like mutational signature²³. For the two BT-474 specific signatures SBS2 and SBS13²², we see a strong linear relationship between the ratio of each signature and the concentration of BT-474 DNA (Fig. 1C,D, SBS2: slope=0.023, $t=6.06$, p value=1.2 E-4; SBS13: slope=0.030, $t=6.92$, $p=4.1$ E-5; $n=12$ for all linear regression models). In contrast, while SBS1 and SBS5 were also observed in BT-474²² the relative proportions of these two signatures do not change depending on the amount of BT-474 DNA spiked in due to its ubiquitous nature and presence in the background genome (Fig. 1E,F, SBS1: slope= - 0.61E-4, $t= - 0.59$, p value=0.57; SBS5: slope= - 0.0033, $t= - 0.30$, $p=0.77$; $n=12$ for all linear regression models).

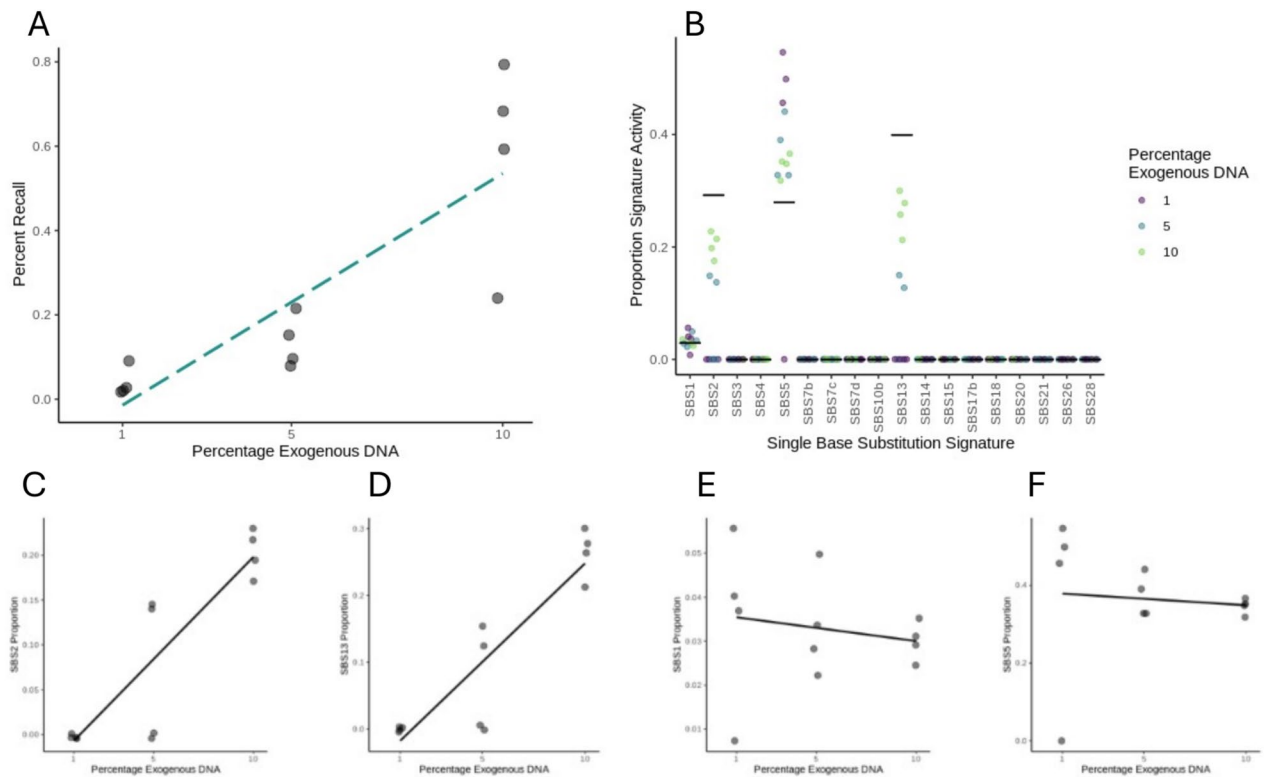


Fig. 1. Exogenous DNA from a well characterized breast cancer cell line (BT-474) spiked into a known genomic background at increasing concentrations can be recovered by our approach (A), preserving the mutational signatures present in the original cell line (B, horizontal lines representing the signature proportion in²²). The cell-line specific signatures also correlate linearly with the amount of exogenous call line DNA (C, D). Ubiquitous mutational signatures that are non-specific of the breast cancer cell line do not show a linear correlation with the amount of exogenous DNA (E, F).

Aggressive filtering results in high-confidence somatic SNVs

Given the challenge present in identifying and distinguishing somatic variants from germline variants and sequencing artifacts, we applied a set of filtering steps to our initial set of SNV to progressively enrich for somatic SNVs (Supplementary Figure S1). In total, our characterization pipeline identified 425,459,468 SNVs across 277 samples corresponding to 103 individuals (mean SNVs per sample 1,477,289, SD 311,645), corresponding to an average depth of 31x (sample-level depth of coverage can be found in sample detail table as part of the online data). We expect these to be composed of bona-fide somatic variants as well as germline variants and sequencing errors and artifacts. The aggregate allele frequency spectrum (Fig. 2, top panel) is compatible with this expectation with a strong peak around 100% frequencies caused by sequencing errors and a broad peak around 50% composed of heterozygous germline variants affected by sampling noise. After applying our filtering, we are left with 211,530 high-confidence sSNVs (mean sSNVs per sample 801, SD 227) with an allele frequency distribution centered around 20% (Fig. 2, bottom panel). The unfiltered set of SNV displayed an almost uniform distribution of mutational signatures with limited variation across individuals and just three signatures accounting for almost 100% of all observed SNV (mean \pm SD: SBS5 94% \pm 2.60%, SBS1 4% \pm 0.04% and SBS54 2% \pm 2.61%). The most prevalent signatures (SBS5 and SBS1) are known to be ubiquitous across all cells and clock-like whereas SBS54 is known to be associated with sequencing artifacts and germline contamination in somatic studies²³.

High-confidence somatic SNVs display different mutational signatures from sequencing artifacts.

To verify that our high confidence sSNVs are not in fact sequencing artifacts, we compared their mutational signature to the signature derived from false positive SNVs detected in our reference HG002 samples. As HG002 is a very well characterized genome, all (germline) SNV variants are known, which allows us to infer that virtually all detected SNVs are sequencing errors and artifacts. Because our starting material is native DNA that does not undergo PCR amplification, we do not expect PCR bias or PCR errors to be a major source of errors. We applied TensorSignatures²⁴ to the set of high confidence sSNVs obtained from the reference genome after applying our filtering (i.e. false positive set). This set showed a different profile from all the subject samples (Fig. 3), with significant overrepresentation of certain signatures (e.g. TS06, TS11, TS13, TS15 and TS16) and absence of multiple signatures observed across subject samples (e.g. TS03, TS05, TS07, TS08, TS09, TS12, TS17, TS18 and TS19). Additionally, the relative proportions of the different signatures are highly variable across subjects and fairly uniform across the different replicate sequencing runs of HG002.

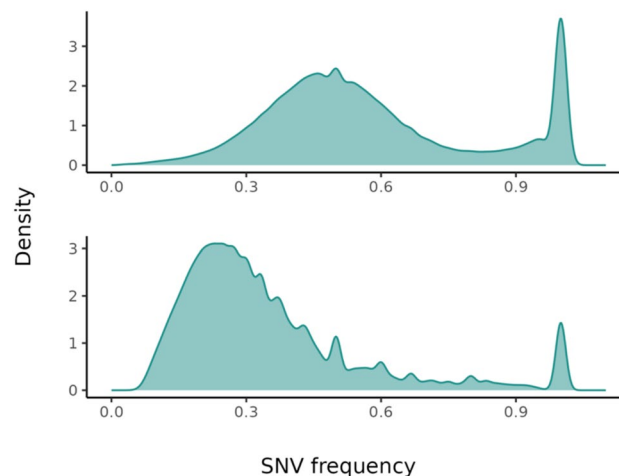


Fig. 2. Density plots of SNV allele frequencies before filtration (top) and after aggressive filtration (bottom) to enrich for somatic variants. The large majority of all SNVs (99.95%) are removed by the filtering step. The unfiltered SNVs show a typical distribution of frequencies associated with germline variants (broad peak around 50% frequency associated with heterozygous variants and a narrow peak around 100% comprising homozygous variants and sequencing errors), whereas the filtered SNVs are enriched for low-intermediate frequencies (10–40%).

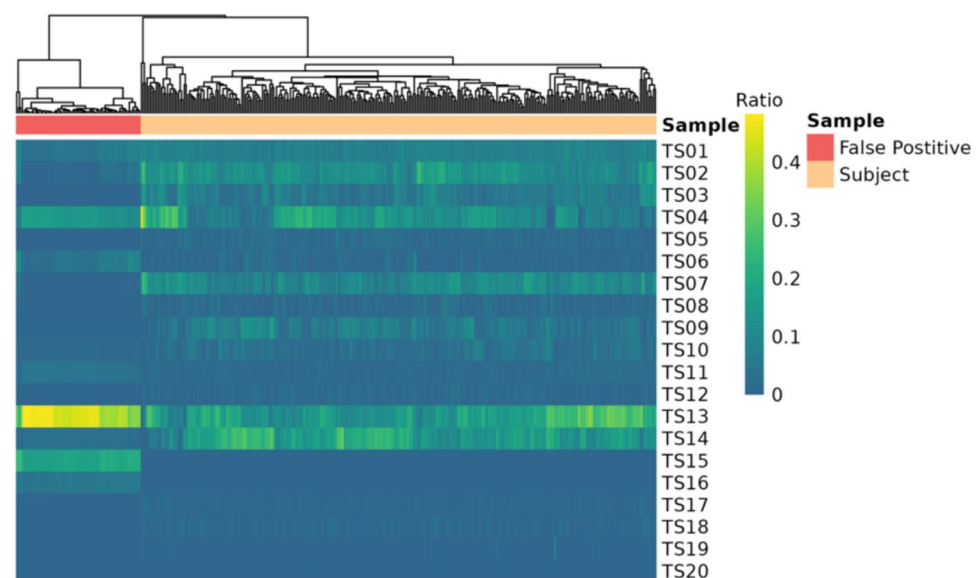


Fig. 3. Heatmap of TensorSignature associated with sequencing errors (HG002 false positives, in red) versus study subjects after filtration (in salmon) shows distinct signatures between the two groups, indicating that the somatic SNVs observed in the subjects are not caused by sequencing errors and artifacts. The color scale corresponds to the proportion of that signature present across all extracted signatures in that particular sample.

Somatic SNVs are enriched in certain functional regions of the genome

Due to different selective pressures, different functional regions of the genome are expected to accumulate mutations at different rates. This is a well-known phenomenon at evolutionary times scales, but it has been observed for somatic mutations as well¹. We measured the median allele frequency of all high confidence sSNVs for each individual across several functional categories (Fig. 4). As expected, the more functionally constrained region encompassing the mRNA encoding portion of the genome display the lowest median frequency, reflecting the fitness effect of deleterious variants on the population of cells carrying such variants. Intronic and intergenic regions show a higher median frequency. Silent variants (i.e. nucleotide substitutions in the protein coding sequence that do not change the encoded amino acid) show the highest allele frequency, as expected given that they are functionally neutral. Surprisingly, on average, missense mutations (i.e. nucleotide substitutions that

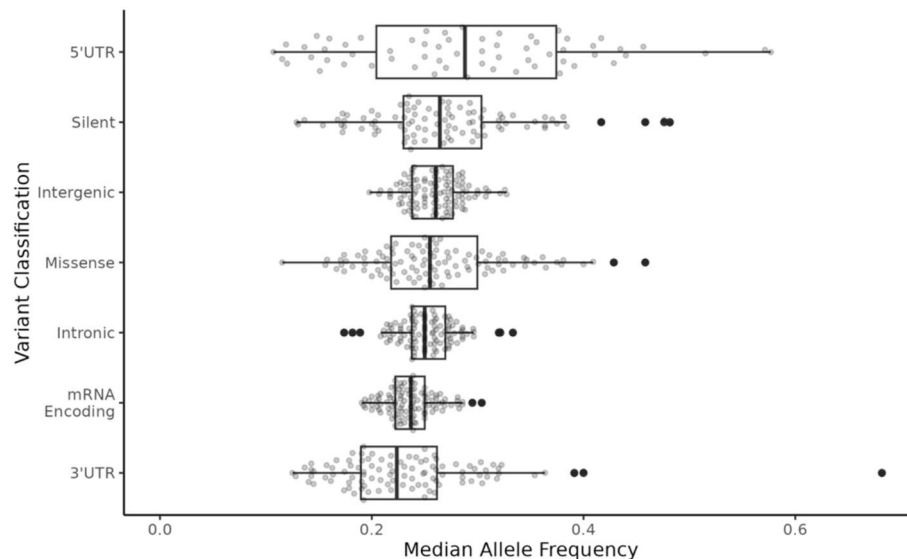


Fig. 4. Median sSNV allele frequency across different functional parts of the genome and mutational effects shows the highest frequency for 5' untranslated region (UTR) followed by silent sSNVs and by sSNVs found in intergenic regions, which is compatible with the overall lack of functional constraints. Conversely, sSNVs found in the mRNA coding region of the genome (defined here as the aggregate of 5' and 3' UTR and protein coding sequences) show the lowest median allele frequencies, compatible with the strongest functional constraints. UTRs, silent mutation and missense mutations show the broadest spread of frequencies due to the smaller overall number of sSNVs falling in those classes. This is especially prevalent for 5' UTR due to the average short length of those regions in humans²⁵. Each point represents the class median for the sample. Note that UTRs, silent and missense variants are also included in the calculations for the mRNA Encoding class. Center line: median; box limits: 1st and 3rd quartiles; whiskers 1.5× interquartile range; outliers: filled black points.

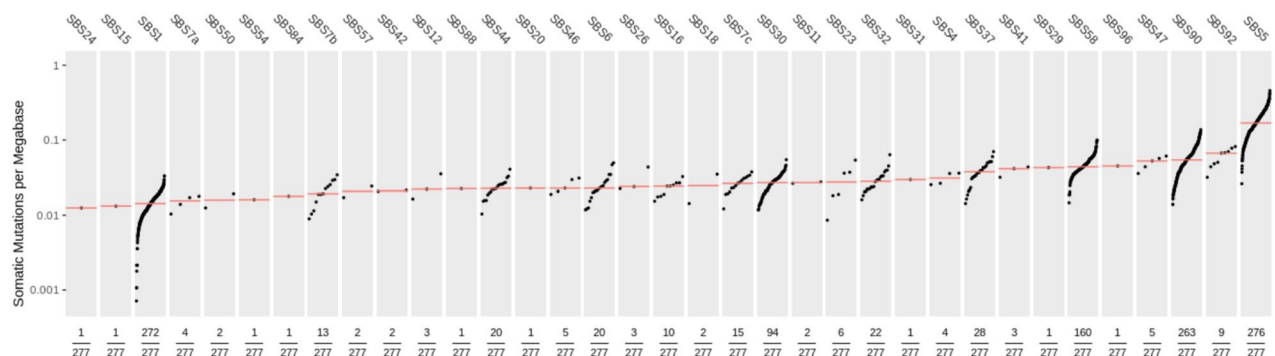


Fig. 5. Distribution of COSMIC SBS signatures identified in our subject cohort. Each point represents one sample. Red line indicates the average across all samples. Few signatures are near ubiquitous (SBS1, SBS5 and SBS90) whereas most signatures are found in subsets of samples.

change the encoded amino acid), do not display particularly low allele frequencies, but instead have a large variability across individuals.

Interpretable mutational signatures differ among individuals

While TensorSignatures²⁴ are useful for identifying novel patterns, they have poor interpretability and provide limited insights into the underlying mutational processes. Conversely, COSMIC's Single Base Substitution (SBSs) mutational signatures have been linked to various mutational processes and exposures^{23,26}. To further attempt to better understand the sources and processes associated with the sSNVs identified from whole blood DNA sequencing, we classified them into known COSMIC SBSs (Fig. 5). SBS1 (clock-like, ubiquitous signature, associated with 5-methylcytosine deamination), SBS5 (clock-like ubiquitous signature of unknown etiology) are found in most samples (272/277 and 276/277, respectively). Some others SBSs appear only in a sizable but limited subset of samples: for example, SBS7a/7b/7c are associated with UV exposure damage, SBS6, SBS30 and SBS44 are associated with defects in DNA repair machinery. It is also possible to identify specific mutational

sources: for example, SBS4 and SBS92 are both associated with tobacco smoke and are found in multiple samples. An unexpected finding is the near ubiquitous presence of SBS90, found in 263 samples out of 277 and absent from the set of sequencing errors and artifacts obtained from HG002, which has been previously associated with exposure to duocarmycin, a potent anti-cancer drug^{23,26}. As we have no evidence that any of the subjects in the study was previously treated with or exposed to duocarmycin, we hypothesize that the presence of this signature is a technical artifact, rather than being of biological origin. We were able to demonstrate this by performing a small experiment in which we run our sequencing pipeline on a HG002 sample as if it were a subject sample, using a different HG002 dataset obtain from a different run as a batch control. In that case, the SBS90 signature was also found in the HG002 sample, confirming its ubiquitous origin. Overall, this approach demonstrates the possibility of further characterizing the relative impact of different mutational process and possible exposures on the total accumulation of somatic variants over an individual's lifetime.

The number of somatic SNVs increases over a person's lifetime

The number of sSNVs is known to increase over a person lifetime, however, this has primarily been observed by analyzing the RNA of autopsy material¹ or DNA from biopsies^{6,27}, or by laborious lab experiments²⁸. Recently, an analysis of genomes obtained from the peripheral blood of over 43,000 individuals revealed a collection of sites that are recurrently mutated in a time dependent fashion, meaning that in older individuals more sites are mutated²⁹. To address the question of whether we can detect a time dependent accumulation of sSNVs over an individual's lifetime, we modeled the corrected number of observed sSNVs against the individual's age (at first draw). We identify a strong age dependent relationship (slope = 2.64, p value = 0.0027; linear mixed effect model with multiple draws per subject, subject ID as a random variable, $n = 277$) corresponding to an accumulation of approximately 2.6 sSNVs per year of life (Fig. 6). This rate does not appear to be dependent on the clinical status of the subjects, i.e. we did not detect any significant differences in the accumulation rates between the cohorts. However, because our cohort was not designed to address this questions, further studies with appropriate power and design will be needed.

Different mutational processes accumulate at different rates

We were further interested to determine whether changes in mutation accumulation rate could be observed at shorter time interval. Given the overall rate of accumulation of sSNVs and the measurement noise, our ability to detect short-time (i.e. within months) changes is limited. Nevertheless, we hypothesized that because the different mutational processes are occurring at different rates, there might be specific signatures that might be better at capturing rapid changes compared to others. For every mutational signature collected, we performed a linear regression on the corrected counts of the signature over time and looked at the distribution of slopes from the linear model. As expected, for most signatures, the distribution of regression slopes across individuals are roughly symmetrical and centered across 0, indicating no temporal accumulation effect. However, for several signatures we observe an overall skew towards positive numbers Fig. 7A. In particular, we see a significant bias towards positive slopes for SBS5 (Fig. 7B), a signature known to be associated with age^{23,26}. Similarly, SBS1 (Fig. 7C), a signature associated with the spontaneous deamination of 5-methylcytosine but not correlated with SBS5^{23,26,30} also shows a skew towards positive slope values. Overall, these findings are compatible with the mutational processes associated with these signatures and support the hypothesis that our approach is able to capture mutation accumulation in real time.

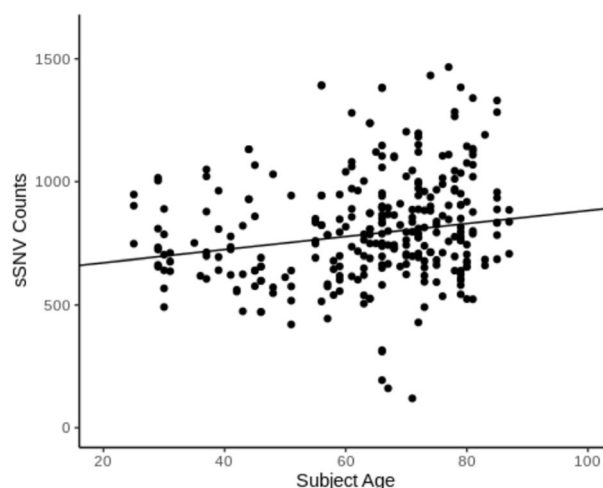


Fig. 6. Relationship between the number of observed somatic SNVs and the subject age at first draw. A significant linear correlation can be determined, corresponding to approximately 27 sSNVs accumulated per decade of life.

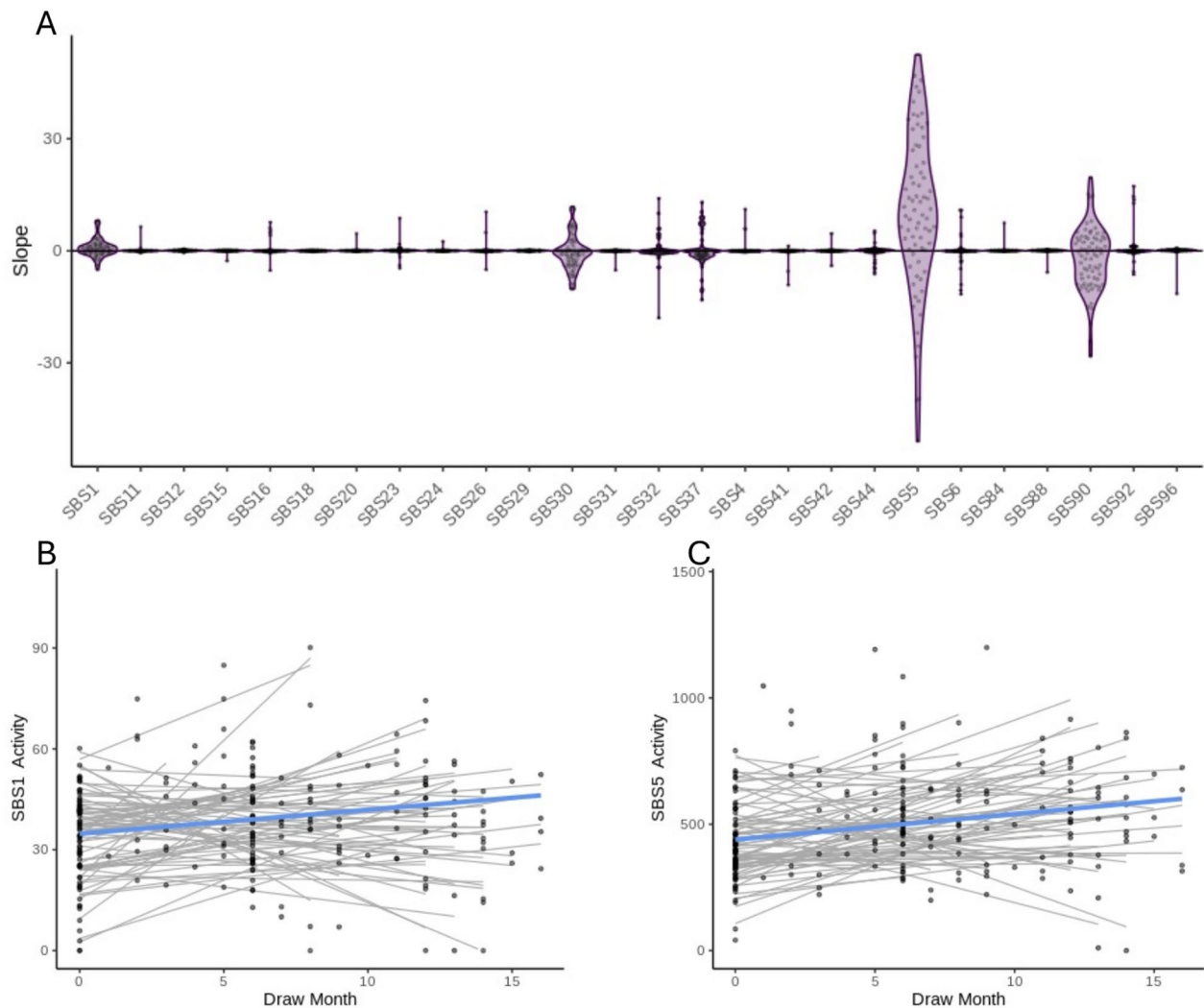


Fig. 7. Detection of changes in specific mutational signatures over multiple draws. To identify whether any signature showed a tendency to increase over the span of our sampling (3 draws over 9–18 months) we created a linear model of the activity of each signature versus the number of months after first draw. Under a model of no accumulation, a distribution of the slopes of those models to be symmetrical and centered around zero would be expected, which was observed for the majority of all signatures (A). However, for SBS1 and SBS5 a significant bias towards positive slopes was observed, indicating an accumulation of sSNVs associated with those mutational processes over the period of the study (B, C, individual subject slopes in grey, average across all subjects in blue. SBS1: mean slope = 0.73, p value = 0.011, $t = 2.62$, 1-sample student t -test; SBS5: mean slope = 10.11, p value = $1.19\text{E}-4$, $t = 4.07$, 1-sample student t -test).

Discussion

Here we report the results of a study aimed to detect and characterize somatic single nucleotide variants from whole blood and measure their rate of accumulation over time.

The accumulation of mutations and the rate of their occurrence are affected by the efficacy of DNA repair mechanisms, which are key molecular mechanisms involved in maintaining cellular homeostasis and normal cellular function. The ability to monitor changes in such a fundamental molecular mechanism over time might help us better understand the processes that result in disease, ultimately allowing the assessment and monitoring of an individual's risk for developing cancer and other diseases.

We used LRS to monitor the rate and variability of somatic genomic alterations over time. This experimental approach required special attention and rigorous monitoring of pre-analytical variables to enable reliable individual-level detection of genomic changes across the whole genome. We established a workflow for serial whole genome LRS from individual's whole blood (Supplementary Figure S2) and were able to quantify and characterize the rate of accumulation of sSNVs across both a long time (lifetime) and short time (months) interval. Using this approach, we were able to identify changes in specific genomic signatures on an individual-level basis and differences in the rate of their accumulation over time.

While these initial results are encouraging, this study has several important limitations: as an exploratory study, we have not yet optimized our assay to amplify the signal and decrease the noise, potentially by focusing

on specific region of the genome, which would allow for more accurate and precise measurements. Our approach is designed to detect a coarse, aggregate signal: individual somatic variants linked to, or indicative of causative effects cannot be identified using our analysis pipeline. In this study we opted to exclude areas of the genomes associated with known pathogenic variants, which might be relevant in the context of exploring emergence of diseases. We have also used samples of convenience to determine potential pre-analytical variations, which limits our ability to further delve into the clinical implication of our findings and might limit its generalizability.

This study opens a new avenue of research to investigate the effects of a variety of intrinsic and extrinsic factors on the dynamics of the human somatic genome landscape and set the grounds for evaluating preventative and therapeutic interventions to mitigate disease risk. These findings support further investigation to validate and assess the clinical significance of the genomic changes identified and to introduce a powerful technology of third-generation sequencing into wide-scale clinical utility, enabling easy longitudinal monitoring of genomic changes toward the development of “state of the system” personalized cancer prevention programs. In addition, looking at outliers, rates of change higher and lower than mean, may yield important biologic insights.

Overall, we believe that this approach can provide a coarse estimate of the “state of the system” with respect to the body’s ability to manage, repair and purge DNA mutations. Based on these data we are planning to refine our experimental and analytical approach to improve our sensitivity and specificity and decrease costs and turnaround time. The use of direct, electronic analysis of DNA and adaptive sequencing, using Oxford Nanopore’s technology also provides unique opportunities as it allows us to potentially target specific regions of relevance with the goal of further refining our ability to detect biologically meaningful changes. Moreover, the ability to integrate other sources of molecular information from the same starting DNA such as methylation or novel structural variants enables us to include additional measurements that will help us further validate and characterize the observed changes, work that is currently ongoing. We are now planning a larger, confirmatory study to be performed with the goal of validating our initial findings and to address and evaluate the significance of these findings in a clinically relevant setting.

Methods

Study Design

This study is a non-interventional prospective pilot study recruiting participants for the collection of blood samples over time for whole genome sequencing.

Participants

We enrolled for the study a total of 103 male and female adult participants (Supplementary Table S1. Participants details available as online data). Patients treated at the outpatient clinic at the Ellison Medical Institute (EMI) were approached and offered to participate in the study. In addition, EMI staff members were also offered the chance to voluntarily participate in the study. The study cohort includes active cancer patients defined as the presence of local or advanced disease in the last 6 month (regardless of treatment status), patients with a history of cancer that were definitely treated and are disease free for at least 6 months, and individuals without a history of cancer. Individuals without symptoms of active infection at enrollment were allowed to participate. The study was approved by Western Institutional Review Board-Copernicus Group (IRB study number: 1328455). All research was performed in accordance with relevant guidelines and regulations and informed consent was obtained from all study participants.

Clinical and demographic data

In accordance with the approved protocol, participants’ demographics were recorded at the time of the first draw. This information was anecdotal only if the participant is not a patient. For the clinic patients, clinical data was also gathered if available from medical record review including clinical/pathologic staging (if applicable), results of imaging studies (if applicable), results of diagnostic blood tests, treatment history, current treatments, current medications, and participant demographics. For the non-patients composed of the EMI staff volunteers taking part in the study, clinical data was collected using designated questionnaires capturing medical conditions including current and past medical history, allergies, medications, herbal or vitamin supplements, family history of cancer and/or chronic diseases.

Blood collection

Informed Consent covers the collection of up to 10ml of blood for this study at each visit. For patients, efforts were made to obtain research blood samples immediately following venipuncture or vascular access performed as part of standard-of-care procedures. Blood was collected into K2EDTA tubes, put in ice and then aliquoted and frozen in -80°C . Participants had multiple specimens collected over the duration of this study. The interval of blood draws varies between the study participants, as active cancer patients, could receive a study draw (up to 10ml) every scheduled visit, optimally at least 2 draws during cytotoxic therapy, whereas participants not receiving treatment (individuals with a history of cancer being followed at the clinic and participant without any history of cancer) may receive a study draw (up to 10ml) every 6 to 12 months. Specimens collected at each visit were labeled with an Ellison Biospecimen ID and linked to a deidentified Ellison ID according to the study protocol.

DNA source

DNA samples sequenced in this study were isolated from whole blood. DNA from whole blood was extracted using either the MagAttract HMW DNA Kit (Qiagen) or the Quick-DNA MagBead Plus Kit (Zymo) following the manufacturer’s protocol (sample processing details available as online data). Isolated DNA sample concentrations were quantified using a Qubit 2.0 Fluorometer (Life Technologies) and purity was assessed using

NanoDrop One (Thermo Scientific). Lymphoblastoid cell line purified DNA for the reference genomes HG001 and HG002 were purchased from Coriell (IDs NA12878 and NA24385, respectively). High molecular weight DNA for the BT-474 human breast carcinoma cell line was purchased from ATTC (ID: HTB-20D).

Library preparation and DNA sequencing

After extraction, DNA was fragmented using FastPrep-96 (MP Biomedicals) and fragment size assessed on a 4150 TapeStation (Agilent). Sequencing libraries were generated using the SQK-LSK110 ligation kit from Oxford Nanopore Technologies (ONT). Sample input DNA was determined based on QC analysis (described below) to ensure consistency and quality of sequencing. For sequencing, samples libraries were loaded onto PromethION R9.4.1 flow cells following ONT's standard operating procedures. Sequencing was performed on the PromethION device for 55 h and data acquisition time was recorded. Reference genome HG002 was sequenced every time in the same batch with the sequencing of clinical samples.

Basecalling

Sequencing and data acquisition were performed using ONT's MinKNOW v. 22.08.06, 22.10.7 and 23.07.12. Raw signal values were processed in real time using guppy (v. 6.2.7, 6.3.9 and 7.1.4) with a high accuracy model and mapped to the reference chromosomes of the human genome v GRCh38.p13. Adaptive sequencing functionality was used to exclude gene regions harboring pathogenic and likely pathogenic variants according to ClinGen³¹, downloaded September 6, 2022) or in the ACMG list of genes with reportable secondary findings³² (ACMG SF v3.1) including 10,000 bp upstream and downstream of the gene start sites (filter file in BED format available as online data).

Somatic variant calling

A custom-developed pipeline was developed and implemented using the Nextflow workflow engine³³, based on the nf-core sarek pipeline^{34,35}. In order to enrich for putative SNVs, ClairS³⁶, a deep-learning based somatic variant caller for LRS was used followed by multiple filtering steps to remove artifacts and likely germline mutations (Supplementary Figure S1). In order to identify hard to sequence areas and capture potential batch effects, a well characterized DNA reference genome sample (HG002) was processed in parallel and sequenced alongside each batch of subject samples. The same ClairS pipeline was used on the reference genome to identify all variants and after removing the known germline variants, used to identify sequencing artifacts and errors. The artifact identified in this step were then removed from the list of putative sample SNVs. Next, variant sites with base quality scores lower than Q15 are removed as well as any SNV with frequencies above 40% (likely subject-specific germline variants). Finally, we compared our putative somatic variants to a list of known SNPs present across human populations (gnomAD³⁷ v. 3.1.2,) and removed any matching sites.

Additionally, a set of locations-based filters were applied to eliminate putative SNVs occurring in sequence regions known to be challenging to sequence (Centromeres;³⁸ Repetitive regions;³⁹ ENCODE Blacklist;⁴⁰) as well as regions that were empirically determined to have excessive noise level as determined a running window of mean coverage. SNVs that had a depth of coverage greater than 1.5 times the mean depth of coverage for that chromosome were excluded from the analysis. We also removed all variants in gene regions associated with reportable finding to minimize the risk of revealing relevant medical information.

These filtering steps reduced the average number of putative SNV per subject from 1.5 million to approximately 800 high-confidence sSNVs per subject.

Annotation

Raw SNVs were annotated with all the information necessary to conduct variant-level filtration: population-wide allele frequencies were obtained from gnomAD³⁷ v3. GATK v4.4.0 VariantAnnotator⁴¹ was used for annotating for filtration (see Somatic Variant Calling and Supplementary Figure S1). High confidence sSNVs were annotated with GATK 4.4.0 Funcotator⁴¹ (prepackaged data source v1.7.20200521s) for downstream functional analyses.

Batch correction

Differences in average high confidence sSNV count were observed between different Ligations Sequencing Kit (LSK) lot numbers. To adjust for batch effects due to LSK lots while accounting for the confounding effect of coverage, we used the standardize method from the Batchma R package⁴², where batch is set to a discrete variable of LSK lot number and confounder is set to median coverage. Batch correction results were validated by diagnosing the parameters of the fitted linear regression models.

Identification of mutational signatures

TensorSignatures²⁴, a multi-dimensional tensor factorization framework was used for characterizing high-confidence sSNVs in terms of the underlying mutational signatures and associated processes. Additionally, mutational signatures associated with sequencing errors and artifacts were generated from the false positive SNVs obtained from the HG002 reference genome sequencing. To this aim, we applied TensorSignatures refit with the predefined Pan-Cancer Analysis of Whole Genomes (PCAWG) signatures to the HG002 false positive catalog, which results in a set of falsely discovered signatures and the corresponding exposure matrix. We then analyzed the falsely discovered exposure matrix to calculate false discovery rates for each of the 20 PCAWG signatures. From the false discovery rates, we built an error model that adjusts an arbitrary PCAWG exposure matrix based on the false discovery rates into a corrected exposure matrix.

To characterize subject samples, we applied TensorSignatures refit with the PCAWG signatures to each subject's high-confidence sSNV catalogue and applied the error model to adjust the output and calculate a false

discovery rate (FDR)-adjusted exposure matrix. The FDR-adjusted exposure matrix was used to model and track subject mutational signatures over time.

For interpretable signatures, COSMIC's Single Base Substitution signatures^{23,26} were assigned to each high confidence sSNV and the counts of each SBS class in each sample was normalized to the total number of high-confidence sSNVs to account for difference in total number of sSNVs (SBS proportion).

Statistical analysis

Statistical analyses were performed using R (v. 4.4.1)⁴³ with packages lme4 (v. 1.1–28)⁴⁴ and lqmm (v. 1.5.8)⁴⁵, and python (v. 3.10.12)⁴⁶.

Data availability

The high confidence somatic variant calls for study participants, reference genomes and spike-in experiments, sample processing details and subject demographic information are available in Zenodo with the identifier: <https://doi.org/10.5281/zenodo.14399981>. The raw sequences and unfiltered variant calls are not publicly available due to privacy reasons to prevent identification of study participants.

Code availability

The nf-core/sarek derived deployment ClairS used to characterize somatic variants is available from GitHub at https://github.com/eitm-org/sarek_pipeline. The custom code used to process and analyze the data is available from GitHub at https://github.com/eitm-org/state_manuscript_analyses/.

Received: 31 January 2025; Accepted: 7 May 2025

Published online: 11 June 2025

References

- García-Nieto, P. E., Morrison, A. J. & Fraser, H. B. The somatic mutation landscape of the human body. *Genome Biol.* **20**, 298 (2019).
- Helleday, T., Eshtad, S. & Nik-Zainal, S. Mechanisms underlying mutational signatures in human cancers. *Nat. Rev. Genet.* **15**, 585–598 (2014).
- Martincorena, I. et al. Universal patterns of selection in cancer and somatic tissues. *Cell* **171**, 1029–1041.e21 (2017).
- García-Nieto, P. E. et al. Carcinogen susceptibility is regulated by genome architecture and predicts cancer mutagenesis. *EMBO J.* **36**, 2829–2843 (2017).
- Franco, I. et al. Whole genome DNA sequencing provides an atlas of somatic mutagenesis in healthy human cells and identifies a tumor-prone cell type. *Genome Biol.* **20**, 285 (2019).
- Martincorena, I. et al. High burden and pervasive positive selection of somatic mutations in normal human skin. *Science* **348**, 880–886 (2015).
- Tang, J. et al. The genomic landscapes of individual melanocytes from human skin. *Nature* **586**, 600–605 (2020).
- Figueroa-González, G. & Pérez-Plasencia, C. Strategies for the evaluation of DNA damage and repair mechanisms in cancer. *Oncol. Lett* **13**, 3982–3988 (2017).
- Abbas, S. et al. Mutational signature dynamics shaping the evolution of oesophageal adenocarcinoma. *Nat. Commun.* **14**, 4239 (2023).
- Fitzgerald, D. M., Hastings, P. J. & Rosenberg, S. M. Stress-induced mutagenesis: Implications in cancer and drug resistance. *Annu. Rev. Cancer Biol.* **1**, 119–140 (2017).
- King, K. Y., Huang, Y., Nakada, D. & Goodell, M. A. Environmental influences on clonal hematopoiesis. *Exp. Hematol.* **83**, 66–73 (2020).
- Warren, J. T. & Link, D. C. Clonal hematopoiesis and risk for hematologic malignancy. *Blood* **136**, 1599–1605 (2020).
- Welch, J. S. et al. The origin and evolution of mutations in acute myeloid leukemia. *Cell* **150**, 264–278 (2012).
- Fuster, J. J. et al. Clonal hematopoiesis associated with TET2 deficiency accelerates atherosclerosis development in mice. *Science* **355**, 842–847 (2017).
- Uddin, M. M. et al. Longitudinal profiling of clonal hematopoiesis provides insight into clonal dynamics. *Immun Ageing* **19**, 23 (2022).
- Hagiwara, K. et al. Dynamics of age- versus therapy-related clonal hematopoiesis in long-term survivors of pediatric cancer. *Cancer Discov.* **13**, 844–857 (2023).
- Genovese, G. et al. Clonal hematopoiesis and blood-cancer risk inferred from blood DNA sequence. *N. Engl. J. Med.* **371**, 2477–2487 (2014).
- Zink, F. et al. Clonal hematopoiesis, with and without candidate driver mutations, is common in the elderly. *Blood* **130**, 742–752 (2017).
- Beyter, D. et al. Long-read sequencing of 3,622 Icelanders provides insight into the role of structural variants in human diseases and other traits. *Nat. Genet.* **53**, 779–786 (2021).
- Marx, V. Method of the year: long-read sequencing. *Nat. Methods* **20**, 6–11 (2023).
- Aganezov, S. et al. A complete reference genome improves analysis of human genetic variation. *Science* **376**, eabl3533 (2022).
- Petljak, M. et al. Characterizing mutational signatures in human cancer cell lines reveals episodic APOBEC mutagenesis. *Cell* **176**, 1282–1294.e20 (2019).
- Alexandrov, L. B. et al. Signatures of mutational processes in human cancer. *Nature* **500**, 415–421 (2013).
- Vöhringer, H., Hoeck, A. V., Cuppen, E. & Gerstung, M. Learning mutational signatures and their multidimensional genomic properties with TensorSignatures. *Nat. Commun.* **12**, 3628 (2021).
- Pesole, G. et al. Structural and functional features of eukaryotic mRNA untranslated regions. *Gene* **276**, 73–81 (2001).
- Sondka, Z. et al. COSMIC: A curated database of somatic variants and clinical data for cancer. *Nucleic Acids Res.* **52**, D1210–D1217 (2024).
- Yokoyama, A. et al. Age-related remodelling of oesophageal epithelia by mutated cancer drivers. *Nature* **565**, 312–317 (2019).
- Olafsson, S. & Anderson, C. A. Somatic mutations provide important and unique insights into the biology of complex diseases. *Trends Genet.* **37**, 872–881 (2021).
- Weinstock, J. S. et al. The genetic determinants of recurrent somatic mutations in 43,693 blood genomes. *Sci. Adv.* **9**, eabm4945 (2023).
- Nik-Zainal, S. et al. Mutational processes molding the genomes of 21 breast cancers. *Cell* **149**, 979–993 (2012).
- Rehm, H. L. et al. ClinGen—The clinical genome resource. *N. Engl. J. Med.* **372**, 2235–2242 (2015).

32. Miller, D. T. et al. ACMG SF v3.1 list for reporting of secondary findings in clinical exome and genome sequencing: A policy statement of the American College of Medical Genetics and Genomics (ACMG). *Genet. Med.* **24**, 1407–1414 (2022).
33. Di Tommaso, P. et al. Nextflow enables reproducible computational workflows. *Nat. Biotechnol.* **35**, 316–319 (2017).
34. Ewels, P. A. et al. The nf-core framework for community-curated bioinformatics pipelines. *Nat. Biotechnol.* **38**, 276–278 (2020).
35. Garcia, M. et al. Sarek: A portable workflow for whole-genome sequencing analysis of germline and somatic variants. *F1000Res* **9**, 63 (2020).
36. Zheng, Z. et al. ClairS: A deep-learning method for long-read somatic small variant calling. <https://doi.org/10.1101/2023.08.17.553778> (2023).
37. Chen, S. et al. A genomic mutational constraint map using variation in 76,156 human genomes. *Nature* **625**, 92–100 (2024).
38. Miga, K. H. et al. Centromere reference models for human chromosomes X and Y satellite arrays. *Genome Res.* **24**, 697–707 (2014).
39. Jurka, J. Repbase Update: A database and an electronic journal of repetitive elements. *Trends Genet.* **16**, 418–420 (2000).
40. Amemiya, H. M., Kundaje, A. & Boyle, A. P. The ENCODE blacklist: identification of problematic regions of the genome. *Sci. Rep.* **9**, 9354 (2019).
41. McKenna, A. et al. The Genome Analysis Toolkit: a MapReduce framework for analyzing next-generation DNA sequencing data. *Genome Res.* **20**, 1297–1303 (2010).
42. Stopsack, K. & Gerke, T. batchtm: Batch effect adjustments. (2021).
43. R Core Team. R: A Language and Environment for Statistical Computing, R Foundation for Statistical Computing (2024).
44. Bates, D. et al. lme4: Linear Mixed-Effects Models using 'Eigen' and S4 (2024).
45. Geraci, M. lqmm: Linear Quantile Mixed Models (2022).
46. Python Software Foundation. The Python Language Reference (2024).

Acknowledgements

We would like to thank Ann Barker, Shannon Mumenthaler, Mitchell Gross, Jonathan Katz, Sissel Juul and Claire Attwooll for input and feedback. We are grateful to Androo Markham and Sean Evans at Oxford Nanopore Technologies for their technical expertise and to Nolan Ung, Frank Callari, Toan Nguyen, Martin Purschke and the facilities and lab operations teams at EMI for their help and support. We would also like to acknowledge Jillian Infusino for her help managing this project. We are indebted to Gordon Sanghera for championing this project and to Oxford Nanopore Technologies for contributing equipment, reagents and support. We are thankful to the two anonymous reviewers for their input and comments. Hag.L was supported by the Rothschild Fellowships for Physician-Researchers.

Author contributions

Study conceptualization: D.B.A., N.M., J.S.H.L., Hag.L and R.K.B. Data curation: Hag.L., B.T., D.A., D.K and T.J.J. Formal analysis: N.M., X.C., Han.L., M.U., V.M. and A.C. Investigation: X.C., C.A., D.S., Han.L., M.U., V.M. and A.C. Methodology: D.B.A., N.M., X.C., C.A., D.S. and T.J.J. Project administration: D.B.A., N.M., J.S.H.L., O.C. and T.J.J. Resources: D.B.A. Software: X.C., D.K., Han.L., M.U., V.M. and A.C. Supervision: N.M., D.B.A., J.S.H.L., R.K.B. and O.C. Visualization: N.M., X.C., M.U. and A.C. Writing (original draft): N.M., Hag.L. and X.C. Writing (reviewing and editing): all authors. Funding acquisition: D.B.A.

Declarations

Competing interests

All authors except Hag.L., D.A. and Han.L. are employees of the Ellison Institute, LLC. D.B.A. is a scientific advisor with equity interests in Oxford Nanopore Technologies. D.B.A. and N.M. are inventors on a provisional patent (US Provisional Patent Application no.: 63/581,553) related to this work. Oxford Nanopore Technologies contributed equipment, materials, reagents and technical support. The remaining authors declare no competing interests.

Additional information

Supplementary Information The online version contains supplementary material available at <https://doi.org/10.1038/s41598-025-01690-z>.

Correspondence and requests for materials should be addressed to N.M. or D.B.A.

Reprints and permissions information is available at www.nature.com/reprints.

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

Open Access This article is licensed under a Creative Commons Attribution-NonCommercial-NoDerivatives 4.0 International License, which permits any non-commercial use, sharing, 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 you modified the licensed material. You do not have permission under this licence to share adapted material derived from this article or parts of it. 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-nc-nd/4.0/>.

© The Author(s) 2025