**\chapter{Germline and somatic mutational processes across the tree of life}**

**\section{Introduction}**

Somatic mutations can occur in cells at all stages of life and in all tissues.

**\subsection{The Darwin Tree of Life Project}**

**\subsection{CCS sequencing and \textit{de novo} assembly}**

**\section{Results}**

**\subsection{Somatic mutation detection}**

**\subsubsection{\textit{Phorcus lineatus} somatic mutation rate}**

To evaluate whether somatic mutation detection in non-human samples is possible with himut, \textit{P. lineatus} samples of different ages (3, 5 and 15) were sequenced and analysed (Methods, Table \ref{}). As somatic mutational process is a continuous process throughout life, mutation burden increases as a function of age. If himut is successful at calling somatic mutations in non-human samples, mutation burden in \textit{P. lineatus} should, therefore, also increase with age and should allow us to calculate the somatic substitution rate. If age and mutation burden are not positively correlated, additional factors not investigated in chapter 2 must be affecting the somatic mutation detection sensitivity or specificity.

The CCS read length of \textit{P. lineatus} samples is 2 to 3 times shorter on average than that of the cancer cell line and normal blood granulocyte samples described in chapter 2. If DNAP processivity is a constant in sequencing of both \textit{P. lineatus} and human samples, the number of subreads per CCS read length is expected to be higher for the \textit{P. lineatus} samples. As expected, the number of subreads per CCS read in \textit{P. lineatus} is 2 to 3 times that in the human samples. The higher number of subreads per CCS read length is reflected in the higher proportion of CCS bases with Q93 BQ score, which increases both the number of bases from which somatic mutation could have been called from and the number of called somatic mutations (Table \ref{}).

In chapter 2, I established that inaccurate estimation of BQ score is one of the main causes of false positive somatic mutation detection. The higher number of subreads per CCS read increases the proportion of CCS bases with Q93 BQ score and inflates the mutation burden of samples with shorter CCS read length. To address the differences in average CCS read length between the samples, CCS reads are re-generated such that the number of subreads per CCS read is a constant and age is the only variable (Methods). In contrast to the previous result, the same downstream sequence analysis with the newly generated CCS reads shows a linear relationship between age and mutation burden and somatic substitution rate in \textit{P. lineatus} foot muscle is calculated to be 40 somatic mutations per cell per year, demonstrating that himut is applicable in non-human samples.

**\subsubsection{Somatic mutation burden}**

In contrast to the somatic mutation detection in human samples, somatic mutation detection in eukaryotic samples from the DToL project requires a VCF file with germline mutations to calculate the heterozygosity and genotype prior for germline mutation detection.

Here, himut was adapted to call somatic mutations

somatic mutations were called in 518 eukaryotic species from the DToL project

**\subsubsection{Somatic mutational signatures}**

**\subsection{Germline mutation detection}**

**\subsubsection{Germline mutation burden}**

**\subsubsection{Transition to transversion ratio}**

**\subsubsection{Germline mutational signatures}**

**\section{Conclusion}**

**\section{Materials and Methods}**

**\subsection{CCS library preparation, sequencing and \textit{de novo} assembly}**

To date, the DToL consortium has collected, prepared, and sequenced approximately $\sim$3000 eukaryotic samples in Great Britain and Ireland. In addition, reference genomes for around $600 eukaryotic species have been assembled and made available to the public, which is accompanied by a genome note that details the process from sample acquisition to chromosome-length scaffold construction.

The DToL project initially used a combination of sequencing (CLR, CCS and linked reads) and scaffolding (e.g. Hi-C reads and BioNano genome maps) technologies to generate chromosome-length scaffolds. The DToL project currently uses HiFiAdapterFilter \cite{} to remove CCS reads with adapter sequences, either hifiasm \cite{} or hicanu \cite{} for \text{de novo} assembly of contigs from CCS reads, purgedups \cite{} to remove haplotype duplication, arrow \cite{} for contig polishing and SALSA \cite{} to order and orient contigs into chromosome-length scaffolds with Hi-C reads. If both the parent and child was sequenced, trio-canu was used to generate haplotype phased contigs. The chromosome-length scaffolds are, thereafter, manually curated using a Hi-C contact matrix to identify and correct misassemblies and to perform additional scaffolding where appropriate. If transcriptome data was available through either RNA or isoform sequencing, gene annotation was also performed in collaboration with the EMBL-EBI eukaryotic annotation team. The specific method and algorithm described here is subject to change with updates to the sequencing method, \textit{de novo} assembly and scaffolding algorithm.

**\subsection{\textit{Phorcus lineatus} somatic mutation rate measurement}**

To calculate the somatic mutation rate of \textit{P. lineatus} (thick top shell), samples of different ages (3, 5 and 15) were collected from Plymouth, UK. Collaborators at the Marine Biological Association (MBA) determined the age of the samples from growth marks on the shells of samples. As recommended, a bench-mounted vice was first used to crush the shell and to carefully separate the sample from the shell (personal communication with Robert Mrowicki at the MBA). In addition, disposable scalpels were used during the dissection to prevent cross-contamination between the samples. HMW DNA was subsequently extracted from the foot muscle using the Circulomics Nanobind Tissue Big DNA Kit (SKU 102-302-100). CCS libraries were prepared following the low-input CCS library preparation protocol () and BluePippin system () was used to size select CCS libraries prior to sequencing.

CCS BQ score is a function of the number of supporting subreads and the concordance between the CCS base and subread bases. The DNAP processivity and CCS read length determine the number of full-length subreads per CCS read, which in turn influences the number of Q93 CCS bases from which potential somatic mutations can be identified. To account for the differences in the number of subreads per CCS read for each ZMW, the raw subreads BAM file was parsed using a custom script to select 10 full-length subreads per ZMW. The script calculates the median subread length for each ZMW and considers subreads between 0.9 times the median subread length and 1.1 times the median subread length as full-length subreads. The processed subreads BAM file was, thereafter, provided as an input to the pbccs algorithm to re-generate CCS reads. CCS reads were subsequently processed as described in chapter 2 and below. Except for the \textit{P. lineatus} somatic mutation rate measurement, CCS reads generated with default pbccs parameters for used for the rest of the analysis.

**\subsection{Germline and somatic mutation detection}**

As detailed in chapter 2, CCS reads with adapter sequences were identified using HiFiAdapterFilt \cite{} and subsequently discarded. In addition, if ultra-low input CCS library preparation protocol was used for CCS generation and if this was documented, CCS reads were also excluded from downstream sequence analysis (this information, however, was not always available). CCS reads were, thereafter, aligned to the assembled reference genomes using minimap2 \cite{} and primary alignments were selected, sorted and merged into a single BAM file using samtools \cite{}. Germline mutations were called using deepvariant \cite{} and provided as an input to himut for phasing of hetSNPs and somatic mutation detection.

Somatic mutations were detected, and mutation burden was calculated using himut.

The default parameters for somatic mutation detection in non-human samples are different from that in human samples. In addition, himut was used to phase hetSNPs to construct haplotype blocks and to enable haplotype phased somatic mutation detection where applicable. Somatic mutation detection was again restricted to the autosomes of the reference genome.

As CCS reads and reference genomes are derived from the same sample, homozygous germline mutations indicate assembly errors and analysis of germline mutations are restricted to heterozygous mutations for samples with a diploid genome.

The detection of somatic mutations across the tree of life followed a similar approach to the one described in chapter 2, but with minor modifications. When somatic mutations were called from DToL eukaryotic species, a VCF file containing germline mutations was supplied to himut to calculate heterozygosity ($\theta$) and the genotype prior $P(G)$. In addition, because a single sample was sequenced per species and as population-scale sequencing studies has not been performed for these species, PoN VCF file could not be generated and VCF file with common SNPs were not available for distinguishing false positive substitutions arising from systematic errors and gDNA contamination, respectively. However, given that CCS library and reference genome originate from the same sample, false positive substitutions arising from alignment errors should be minimal and CCS reads resulting from gDNA contamination should be excluded from the analysis based on their sequence identity.

**\subsection{Mutational signature extraction and analysis}**

As described in chapter 1, there are 6 substitution types (C>A, C>G, C>T, T>A, T>C, T>G) in the pyrimidine context and 16 trinucleotide sequence contexts for each substitution class, creating the canonical SBS96 classification system. Since the ancestral allele is known for somatic mutations, the SBS96 classification system is often used to categorise somatic substitutions. In contrast, because the ancestral allele is unknown for germline mutations, the SBS52 classification system is used for germline substitution classification.

Here, I describe the SBS52 classification system and how the SBS96 classification system is transformed into the SBS52 classification system. The need for the SBS52 classification system arises from the fact that certain germline substitutions are indistinguishable from one another because the reference base cannot be assumed to be the ancestral allele; as the reference genome is sequenced and assembled from a randomly sampled individual, the haplotype containing the germline mutation could have also been the reference sequence. For instance, a C>A substitution in the AAA trinucleotide sequence context on the forward strand cannot be distinguished from a T>G substitution in the TTT trinucleotide sequence context on the reverse strand. Similarly, C>T substitutions cannot be differentiated from T>C substitutions. In addition, a C>G (T>A) substitution in a certain trinucleotide sequence context is interchangeable with another C>G (T>A) substitution in a different trinucleotide sequence context. Organised in Table \ref{} is the complete transformation of the SBS96 classification system into the SBS52 classification system.

After categorising germline and somatic substitutions based on the SBS52 and SBS96 classification system, SBS52 and SBS96 counts were processed as described below for \textit{de novo} mutational signature extraction using HDP \cite{}

**\subsubsection{SBS96 mutational signature extraction}**

As detailed in chapter 2, SBS96 counts were normalised according to the number of callable CCS bases, callable reference bases and the trinucleotide distribution in the autosomes of the reference genome. Somatic SBS96 counts in each species is a linear combination of true positive substitutions from somatic mutational processes and false positive substitutions from library, sequencing, and software errors.

The number of false positive substitutions for each SBS96 classification, however, can be estimated from the substitution and trinucleotide sequence context dependent CCS error rate, which was determined in chapter 2, and the number of CCS trinucleotides from which somatic mutations could have been potentially called. These estimates can then be subtracted to obtain SBS96 counts where true positive substitution counts is better presented. If the same CCS library preparation protocol was not used for the DToL and the cord blood granulocyte sample, the estimation and subtraction of false positive substitutions from each SBS96 category may not be as effective in improving the signal-to-noise ratio. If a different protocol, for example, was used to extract HMW DNA and prepare a CCS library, false positive substitutions could be generated from an uncharacterized error process distinct from that identified in chapter 2.

After speciation, ongoing somatic mutational process(es) in germline stem cells depletes the trinucleotide sequence context it acts upon and shapes the trinucleotide distribution.To account for differences in trinucleotide distribution in each species, SBS96 counts are further normalised such that each trinucleotide sequence context equally contributes to the total SBS96 count (eq \ref{}). This normalisation further increases the signal-to-noise ratio of somatic mutational processes, particularly those with a higher somatic mutation rate and that are shared between the somatic and germline cells.

Before \textit{de novo} mutational signature extraction, normalised SBS96 counts from each species are organised into a single matrix, samples with less than 100 somatic mutations are removed from the matrix and the total somatic mutation count for each species is normalised to the median somatic mutation count. After mutational signature extraction, each mutational signature was inspected for the following qualities to distinguish mutational signatures arising from ongoing somatic mutational processes in the sample or from library and sequencing errors upstream of sequence analysis.

\begin{enumerate}

\item The mutational signature is similar to those found in the COSMIC mutational signature database.

\item The mutational signature is present in another species in the same phyla

\item The mutational signature has biological replicates (e.g. idPlaAlba and xgPhoLine).

\item The mutational signature has transcriptional-strand bias

\item The mutational signature is similar to the germline mutational spectrum.

\item The attribution of the mutational signature to the total mutation burden in the sample

\item The number of somatic mutations associated with the mutational signature is not a multiple of the number of germline mutations.

\end{enumerate}

**\subsubsection{Independent biological replication of mutational signatures}**

To confirm that identified mutational signatures are the result of a biological process and not stochastic errors,

ten,

three , ten idSyrPipi, were sequenced and analysed as described to call somatic mutations. Cosine similarity between the mutational spectrum of the samples were measured and similarity measurements were visualised through a heatmap to show that a unique somatic mutational process is operational in each species and that samples from same species share same somatic mutational process. Samples with sequence coverage below 30 and somatic mutation count below 100 were excluded from the analysis.

**\subsubsection{Mutational signatures with transcriptional-strand bias}**

 Single point substitutions are oriented based on their pyrimidine base and the strand of the reference genome. When a gene is found on the reference strand an A:T > T:A substitution in the footprint of the gene is classified as transcribed T > A (example indicated by circle) while a C:G > G:C substitution in the footprint of the gene is classified as un-transcribed C > G (example indicated by star). Mutations outside of the footprints of genes are classified as non-transcribed (example indicated by square). Classification of single base substitutions is shown both in regard to SBS-24 and SBS-384

The mutational classifications described above provide a detailed characterization of mutational patterns of single base substitutions, doublet base substitutions, and small insertions and deletions. Nevertheless, these classifications can be further elaborated by incorporating additional features. Strand bias is one commonly used feature that we and others have incorporated in prior analyses [[13](https://bmcgenomics.biomedcentral.com/articles/10.1186/s12864-019-6041-2#ref-CR13),[14](https://bmcgenomics.biomedcentral.com/articles/10.1186/s12864-019-6041-2#ref-CR14),[15](https://bmcgenomics.biomedcentral.com/articles/10.1186/s12864-019-6041-2#ref-CR15), [22](https://bmcgenomics.biomedcentral.com/articles/10.1186/s12864-019-6041-2#ref-CR22)]. While one cannot distinguish the strand of a mutation, one expects that mutations from the same type will be equally distributed across the two DNA strands. For example, given a mutational process that causes purely C:G > T:A mutations and a long repetitive sequence 5′-CGCGCGCGCGCGCGCGCCG-3′ on the reference genome, one would expect to see an equal number of C > T and G > A mutations. However, in many cases an asymmetric number of mutations are observed due to either one of the strands being preferentially repaired or one of the strands having a higher propensity for being damaged. Common examples of strand bias are transcription strand bias in which transcription-couple nucleotide excision repair (TC-NER) fixes DNA damage on one strand as part of the transcriptional process [[30](https://bmcgenomics.biomedcentral.com/articles/10.1186/s12864-019-6041-2#ref-CR30)] and replicational strand bias in which the DNA replication process may result in preferential mutagenesis of one of the strands [[31](https://bmcgenomics.biomedcentral.com/articles/10.1186/s12864-019-6041-2#ref-CR31)]. Strand bias can be measured by orienting mutations based on the reference strand. In the above-mentioned example, observing exclusively C > A mutations (and no G > A mutations) in the reference genome sequence 5′-CGCGCGCGCGCGCGCGCCG-3′ may mean that: (i) the guanine on the reference strand is protected; (ii) the cytosine on the reference strand is preferentially damaged; (iii) the guanine on the non-reference strand is preferentially damaged; (iv) the cytosine on the non-reference strand is protected; or (v) a combination of the previous four examples. In principle, a strand bias reveals additional strand-specific molecular mechanisms related to DNA damage, repair, and mutagenesis.

SigProfilerMatrixGenerator provides a standard support for examining transcriptional strand bias for single base substitutions, doublet base substitutions, and small indels. The tool evaluates whether a mutation occurs on the transcribed or the non-transcribed strand of well-annotated protein coding genes of a reference genome. Mutations found in the transcribed regions of the genome are further subclassified as: (i) transcribed, (ii) un-transcribed, (iii) bi-directional, or (iv) unknown. In all cases, mutations are oriented based on the reference strand and their pyrimidine context.

To sub-classify mutations based on their transcriptional strand bias, we consider the pyrimidine orientation with respect to the locations of well-annotated protein coding genes on a genome. For instance, when the coding strand (i.e., the strand containing the coding sequence of a gene; also known as the un-transcribed strand) matches the reference strand, a T:A > A:T will be reported as an untranscribed T > A (abbreviated as U:T > A; Fig. [2](https://bmcgenomics.biomedcentral.com/articles/10.1186/s12864-019-6041-2#Fig2)). In this case, the template strand (i.e., the strand NOT containing the coding sequence of a gene; also known as the transcribed strand) will be complementary to the reference strand and a G:C > C:G mutation will be reported as a transcribed C > G (abbreviated as T:C > G; Fig. [2](https://bmcgenomics.biomedcentral.com/articles/10.1186/s12864-019-6041-2#Fig2)). In rare cases, both strands of a genomic region code for a gene. Such mutations are annotated as bidirectional based on their pyrimidine context. For example, both a T:A > C:G and a A:T > G:C mutations in regions of bidirectional transcription will both be annotated as a bidirectional T > C (abbreviated as B:T > C). The outlined notations are applicable when describing mutations that are located within the transcribed regions of the genome. When a mutation is located outside of these regions, it will be classified as non-transcribed. For example, both a C:G > T:A and a G:C > A:T mutations in non-transcribed regions will be annotated as a non-transcribed C > T (abbreviated as N:C > T).

**\subsubsection{SBS52 mutational signature extraction}**

To \textit{de novo} extract mutational signatures from germline mutations, germline and normalised somatic SBS52 counts from 518 eukaryotic species were organised into a single matrix and each SBS52 count was further normalised such that each trinucleotide sequence context contributes equally to the SBS52 count (eq. \ref{}). In contrast to the SBS96 classification system, the SBS52 classification system has 26 trinucleotide sequence contexts where the middle base is a pyrimidine base.

Through mutational signature extraction from normalised germline and somatic SBS52 counts and downstream mutational signature analysis, ancestral alleles of germline mutations were recovered and the contribution of somatic mutational processes to the germline mutational spectrum was also measured.

**\subsection{Timing the emergence of somatic mutational processes}**

The phylogenetic relationship between 518 species and the time of speciation was inferred from phylogenetic tree available at <http://www.timetree.org> and relevant information from academic literature. The birth of new species with new somatic mutational processes was used to time the emergence of new somatic mutational processes. The time at which the new somatic mutational process is estimated to have emerged will have to be updated with the ongoing efforts from the DToL consortium.