**Chapter 3: Germline and somatic mutational processes across the Darwin Tree of Life**

**Introduction**

The Tree of Life encapsulates biological entities with 5 billion years of history on Earth, extinct species, survivors, and their descendants. The genomes of a select number of species, deemed biologically important, have been sequenced and assembled [ref,ref,ref, Drosophila melanogaster, C.elegans, Zebrafish, Mouse]. Homo sapiens, as a matter of fact, is one leaf in the Tree of Life and an unknown number of leaves remains to be studied. The completion of the human genome project and ramifications of the human genome project is undoubtedly a monumental moment in human genomics, but we far from studying and understanding the question “What is Life?” “What constitutes Life on Planet Earth”.

Contracts, expands, fuses, inverts, rearranges, inserts, deletes, substitutes and copies and pastes, recombines, and the combination of all the above mechanisms to change the genome.

A number of factors has thwarted our efforts to understand species across the Tree of Life. These factors include the sequencing cost, read length, base accuracy, and computational costs, genome sequence complexity and ploidy.

The human genome project cost approximately 3 billion dollars, equivalent to dollar per base pair and required colossal effort requiring international collaboration across major sequencing institutions. Despite the gargantuan effort to physically map and assemble individual BAC clones, the human reference genome had missing sequences, unplaced and unlocalized scaffolds with unknown locations on the human reference genome. The p-arm of acrocentric chromosomes and centromeric sequences of every chromosome, for example, remains unassembled because of their highly repetitive sequence content. The centromeric sequence in the latest human reference genome grch38, hence, is modelled and is not a true representative of the underlying sequence. In addition, the palindromic sequences in chromosome Y makes chromosome Y particularly difficult to assemble and the high degree of similarity between chromosome X and chromosome Y because of X-degenerate and X-transposed sequences. The human reference genome required the advent of new sequencing technologies with higher base accuracy and longer read length to correct misassemblies and minor errors and a new generation of human reference genome [ref, ref].

Segmental duplications defined as non-repetitive sequences with >90% sequence homology between multiple copies makes de novo assembly an extremely difficult problem.

The whole-genome shotgun sequencing approach and assembly approach with Illumina short reads might be scalable, but the assembly produced from this approach has been incomplete and uninformative and not suitable for population genetic analysis. The JCVI genome, for example, created from ~500bp Sanger reads are devoid of segmental duplications.

The human reference genome is undoubtedly the most accurate mammalian reference genome and required a colossal effort to generate BAC clones, to determine the location of BAC clones through physical mapping and determining the BAC clones for minimal tiling path generation.

There were initially two competing approaches for human genome construction: whole-genome shotgun sequencing by JCVI and minimum tiling path by the NCBI?

The advent of PacBio CCS and ONT sequencing has been a game-changer/monumental/pivotal moment for de novo assembly and Hi-C based scaffolding has been a game changer for generating chromosome-length scaffolds. Hi-C reads, were originally used for interrogating the 3D structure of the genome, to understand how the genome is folded and tightly packed. Illumina mate-pair sequencing with different insert-sizes have been used for order and orienting contigs. Similarly, Hi-C reads can be thought of as mate-pair sequencing with read-insert sizes ranging from 100 to chromosome-length insert sizes. Hi-C reads were first used by XXX for scaffolding by XX. In the 3D space, sequences that are closer in linear space is also closer in 3D space and further linear distance, the further the sequences are also in 3D space. In other words, sequences that are in proximity are more likely to be together in 3D space and vice versa [ref]. In addition, DNA derived from the same chromosome are in more contact with each other. Chromosomes are isolated in 3D space. Using these features, assembled contigs can be clustered to chromosomes and contigs can be ordered and oriented. In addition, aberrant Hi-C read signals can be used to detect misassemblies and to identify regions that needs to be separated. BioNano Genome mapping also has enabled high-throughput physical mapping of the genome at scale, but as sequence information is not provided by optical genome mapping and does not provide additional structural information that is different from chromosome-length scaffolds produced from contigs and Hi-C reads, long-read and Hi-C sequencing based de novo assembly is the method of choice for most large-scale de novo assembly projects. In addition, Hi-C contact matrix against the assembled chromosome-length scaffold can be manually inspected through visualisation to identify misassemblies and to correct misassembliess. In addition, the assembly graph constructed from pairwise read alignment can also be visualised to inspect problematic assembly regions.

These advances have enabled T2T-assembly of CHM13 haploid genome and T2T assembly of microbial genome can be routinely done with ease.

The ability to generate highly contiguous and highly complete chromosome-length scaffolds inspired many groups to revive genome assembly projects to revisit the problem of understanding our relatives in the Tree of Life. The Darwin Tree of Life project at the Wellcome Sanger Institute aims to sequence and assemble high-quality reference genomes for 66,000 eukaryotic species from \*Britain and Island\* with the most recent sequencing technologies. The DToL project initially used a combination of CLR, linked reads, BioNano genome maps and Hi-C reads to construct chromosome length scaffolds, but the combination of CCS and Hi-C reads have become the sequencing method of choice to construct reference genomes of different eukaryotic species. We hypothesized that our method for single molecule somatic mutation detection in human samples agnostic of clonality will be applicable towards somatic mutation detection across species agnostic of species. The understanding of somatic mutagenesis process in non-human samples have been limited to date and we thought this would be an opportunity to study both germline and somatic mutational in non-human samples, and in many species for the first time, to understand the evolutionary relationship of different mutational processes and the emergence and convergence of different mutational process across time.

This opportunity allows us to answer/address many questions that could not be addressed to date. This opportunity allows us to have an attack vector with which the question can be interrogated. What is the somatic mutation rate of different species? How has somatic mutation rate changed during the millions of years of evolution? Why do certain species don’t have cancer? Why is there no relationship between the number of cells per species and the incidence of cancer for each species? How has other species evolved to protect their genome integrity and how has DNA damage and repair mechanisms evolved to protect the DNA from hostile environment?

%% \subsection{Somatic mutations}

%% \subsection{Mutational signatures and mutational processes}

%% darwin pondered the unit of inheritance (the physical material and the mechanism responsible for changing the physical material)

%% enodgenous and exogenous somatic mutation

%% DNA damage, repair, fixation

%% envrionment

%% DNA polymerase infidelity, germline mutations

%% importance of somatic mutation detection, lineage tracing, driver mutations

%% a harsh environment, insult to the DNA, necessary to repair DNA damage

%% type of DNA damage: single-base substitution,

%% what is universal about DNA? codon, degenerate, universal, 64 codons, stop-codon, start-codons

%% first protein

%% first riboenzyme?

%% first unicellular organism %% first in the sea

%% first lipid-bilayer that separates order from disorder, control of passage of molecules across a semi-permeable membrane

%% fusions, meiotic recombination, plant recombination?

%% non-hologous end joining

%% transcripion-coupled repair

%% Selection Pressure & Natural Selection & Survival of the fittest

%% deleterious, postivie,

%% linked by DNA

%% entropy to submission

%% Scientists still have not figured out how the first unicellar organism has arisen

%% Complexity that

%% DNA replication, DNA polymerase fidelity, DNA polymerase error rate, as a source of first mutations

%% first multicellular-organism

%% DNA nicks, DNA double-strand breaks, cyclo-butane dimer, UV light, chemicals

%% different types of DNA polymerases, redundancies

Relatives in the tree of life

A subset of the branches in the trees of life

Select number of leaves on the tree of life has been studied

which in turn was limited by the sequencing cost and the technical limitations of the next-generation sequencing platform.

**Materials and Methods**

**CCS library preparation and sequencing**

**De novo assembly, scaffolding and curation**

Darwin Tree of Life project members assembled, scaffolded, and curated the reference genomes. The specific method used is dependent on the species and the availability of data, but the method is similar across species. Contigs were generated using either hifiasm [ref] or hicanu [ref], and misassemblies were detected and purged with purgedup [ref]. If parental data was available, trio-canu was used to construct haplotype phased assemblies. The contigs were ordered and oriented using Hi-C reads and scaffolds were polished with Arrow to close gaps and to obtain a more accurate consensus sequence of the assembly. The chromosome-length scaffolds are, thereafter, manually inspected with Hi-C contact matrix to identify remaining misassemblies, to correct misassemblies and to scaffold contigs where there is sufficient Hi-C signal to connect, order and orient the remaining unplaced and unlocalised contigs. If RNA-seq or Isoform-seq was available, EBI gene annotation pipeline was used to obtain gene annotations from each reference genome. The de novo assembly is an ongoing process with improvements in sequence data and assembly algorithms and the method is subject to change with changes in availability of sequence data and assembly algorithms.

**Phorcus lineatus preparation**

To obtain the foot muscle of P. lineatus, the shell was cracked open and carefully the foot muscle was obtained. We dissected the foot muscle of the P.lineatus and sent the sample for HMW DNA extraction using circulomics HMW DNA extraction kit. Insufficient HMW DNA was obtained through shearing and hence, Blue Pippin size selection was performed to size select the library.

**CCS read alignment and germline and somatic mutation detection**

CCS reads were aligned to the human reference genome (b37 and grch38) with minimap2 (version --) with the parameters “” [ref] and primary alignments were compressed, merged, and sorted with samtools (version --)[ref]. Germline SNPs and indels were detected with deepvariant (version --) and germline hetSNPs were haplotype phased with himut (version 1.0.0). Somatic SBS were also identified with himut (version 1.0.0) with minor modifications to enable somatic mutation detection agnostic of species. *Before somatic mutation detection, himut loads the deepvariant VCF file to calculate the germline heterozygosity prior and uses the prior for subsequent germline mutation detection and to distinguish germline mutation from somatic mutation.*

**HDP mutational signature extraction**

**Mutation signature analysis**

**Results**

**DToL project**

The DToL project aims to sequence and assemble ~2000 species in phase 1 of the project. To date, chromosome-length scaffolds of ~600 eukaryotic species have been sequenced and assembled (Methods), of which ~ number of species were CCS sequenced. The assemblies and the sequence data are publicly available. Thanks to the read length and base accuracy of CCS reads, contigs have a high contig N50 (Figure XX) and Hi-C reads enable the construction of chromosome-length scaffolds and the scaffold N50 is limited by the chromosome length. In addition, the assemblies typically have Q50-Q60 base accuracy, comparable to the base accuracy of the human reference genome [ref]. The assembly statistics for each species and the reference genome accession number is summarised in Table XX.

Of which XXX number of samples had diploid genomes. We excluded polyploid samples from the analysis.

**Somatic mutation detection and evaluation**

As the CCS read and the reference genome is derived from the same sample, homozygous mutations should be reflected in the reference genome, and any mutation detected must be either a heterozygous mutation, a somatic mutation, or an assembly error. In addition, As the CCS read and the reference genome is derived from the same sample, false positive substitutions originating from alignment errors should be significantly reduced.

Different samples have different heterozygosity and hetSNPs can be easily mistaken as somatic mutations. To confirm that our method is applicable to non-human samples, we obtained Phorcus lineatus samples with different ages (3 samples each from the 3-, 5-, 10- and 15-year-old) to confirm the linear relationship between time and mutation burden per cell. We calculated the somatic mutation rate of Phorcus lineatus to be XXX per cell per year (Figure X). The tight bound on the linear relationship between time and mutation burden per cell gave us the confidence that our sample is applicable to all species.

The sequencing summary statistics for P. lineatus is summarised in Table XX. As the read-of-insert size decreases, the number of subreads per CCS read increases (Table). As the number of subreads per CCS read increases, CCS read should have higher proportion of CCS bases with Q93 bases. We, however, were aware from uncapped CCS BQ scores that increase in the number of supporting subreads does not necessarily lead to more accurate BQ scores. We sub-selected 10 full-length subreads from each productive ZMW and re-generated the CCS reads such that all the samples shared the same constant for comparison (Methods).

The age of the samples is unknown and hence, somatic mutation rate cannot be calculated per species basis, but we can make some reasonable assumptions based on the life cycle of the species in question to estimate the somatic mutation rate of each species. We have excluded insects that undergoes metamorphosis from the calculation of somatic mutation rate as the embryonic stem cells which grows into the larvae and adult cells are distinct and separated earlier in the life cycle of the insect [ref]. We identified XXX number of somatic mutations across XXX number of species and discovered X number of mutational signatures from the somatic mutations with unknown aetiology.

**Mutational signature analysis**

As the CCS read and the reference genome is derived from the same sample, homozygous mutations are assembly errors that were not polished, and heterozygous mutations are the true mutations (Methods).

We observed a high concordance between the germline and somatic mutational process, suggesting that the somatic mutational processes we discovered is an endogenous somatic mutational process much like the clock-like mutational process SBS1 and SBS5 in human samples. The detected somatic mutational signature could explain much of the germline mutational process in many of the species (Figure XX).

We found SBS1 and SBS5 mutational signature to be common in birds and mammals. We, however, also discovered SBSX in killer whales. Interestingly, SBS1 was not found in any other species while SBS5-like signature was commonly found in other species. We still do not know the aetiology of SBS5 and the presence of SBS5 in non-dividing somatic cells suggesting that DNA replication is not the driver of SBS5 and SBS5 might be a composite of multiple different mutational processes [ref]. Our data suggest the combination of mutational process that produces SBS5 might be an ancestral one as it is shared by species separated by hundreds of millions of years of evolution.

In contrast, there were some species where the germline mutational process and somatic mutational process were distinct from one another. The pitfall of our experimental design is that only one sample is available from each species, and we can be confident of the identified mutational signature unless the mutational signature is observed in multiple species of the sample family or if multiple samples from the same species is sequenced. We hypothesized that environmental mutagenesis might be responsible for the observed mutational spectra as it has a strong transcriptional strand bias and a strong preference for a specific trinucleotide sequence context (Figure XX). To confirm that this environmental mutagenesis is common in this species, we collected and analysed a number of additional hoverflies (Figure XX) and compared the mutational spectrum of species where multiple samples are available (Figure XX). The species could be clustered based on the similarity of the mutational pattern observed in each species (Figure XX).

Germline mutational processes are typically studied in the context of TiTv ratio to measure the ratio of mutations that are purine mutations to pyrimidine mutations. Human germline mutations typically have a TiTv ratio of 2.0-2.1. If the mutation process was truly random, the TiTv ratio would be 0.5, but because spontaneous deamination of 5mC to thymine is the common germline mutational process in humans, transitions are more frequent than transversions. TiTv ratio, hence, can give us an indication of what might be the frequent germline mutational process in other species (Figure XX).

We studied the germline mutational process in the light of somatic mutational process that we discovered. To compare the two mutational processes, we compressed the SBS96 into SBS48 for comparison as the ancestral allele is unknown for the germline mutation while the ancestral allele is known for the somatic mutation. To comparison revealed that much of the germline mutational process can be explained by the detected somatic mutational process while the remaining germline mutational process might be originating from mutagenesis associated with recombination or other unknown factors in each individual.

In addition, we discovered new PacBio artefact signatures independent of that one discovered and discussed in Chapter 2. The discovered PacBio artefact signature, we believe to be from library errors.

**Germline and somatic mutational processes**

**TiTv ratio**

**Conclusion**

We discover XX number of mutational signatures previously undiscovered in previous studies and XX number of mutational signatures absent in database.

**Discussion**

We expected short-lived insects to have the highest somatic mutation rate, but in contrast to our assumption, many of the insects, especially insects belonging to the lepidtopera family has the lowest mutation burden. XX family which diverged from the lepidoptera family XXX mya ago, however, seems to experience increase in mutation burden with age. The difference between the two families is that while lepidoptera has a metamorphosis stage while XX family does not. In addition, the lepdioptera, coleptera, XX and XX that undergoes metamorphosis account for 80% of the insects, suggesting that insects that undergo metamorphosis has an evolutionary advantage against insects that does not. We conjecture that metamorphosis allows adult insects to have limited exposure to the DNA damage that might have accumulated during the larvae stage and that the imaginal disc that developed into the adult insect might be protected from DNA damage like the gametes in human samples [ref]. In addition, placenta is reported to have higher somatic mutation rate and higher number of chromosomal alterations as a tissue that is useful only for a limited amount of time [ref]. Similarly, caterpillars or young larvae stage of the insect might accrue more somatic mutations and chromosomal alterations. To confirm our hypothesis, gDNA from chrysalis and the adult insect of the same individual could be acquired and sequenced.

%% Samples without somatic mutations

%% Somatic theory of aging

%% Life cycle of Insects

%% Environmental mutagenesis

%% number of eukaryotic species from the DToL project

%% Peto's paradox

%% limited study of somatic mutational processes in other non-human samples: C-elegans and Drosophila melanogaster

%% limited study of germline mutational processes

%% Life cycle of animals

%% recipe for evolution: mutation, natural selection, speciation,

%% cold blood

%% environment, habitat

%% endogenous, exogenous processes

%% somatic mutation: DNA damage, repair and fixation %% clonality

%% germline and somatic mutational processes could be shared

%% Darwin Tree of Life project aims to sequence and assemble 66,000 species in Britain and Island

%% somatic theory of aging: somatic mutations increase with age and the accumulation of somatic mutations impairs cellular functions

Genetic components/modules for synthetic biology purposes and applications