**\chapter{Discussion}**

**\section{Summary of findings}**

In this PhD thesis, I challenge the preconception that long reads from single-molecule sequencing technologies are inaccurate.

In chapter 2, I hypothesise that CCS bases are, in fact, the most accurate among commercially available sequencing platforms, and I develop a tool, himut, to leverage CCS read length and base accuracy for single molecule somatic mutation detection. I benchmark himut’s performance using samples where each sample has a distinct somatic mutational process and where a single somatic mutational process is responsible for newly acquired somatic mutations. The introduction of himut enables researchers to call somatic mutations, in addition to germline mutation and base modification detection from a single human genome using a single SMRTcell and the Revio sequencing instrument.

In chapter 3, I use CCS reads and high-quality reference from a range of eukaryotic species from the DToL project to study somatic mutagenesis across the tree of life. Until recently, our understanding of somatic mutational processes has been limited to species where high-quality reference genomes are available such as \textit{H. sapiens} and model organisms such as \textit{C. elegans} \cite{} and \textit{M. musculus} \cite{}.

To confirm that himut is applicable in non-human samples, I called somatic mutations and calculated the mutation burden of \textit{P. lineatus} with different ages (3, 5 and 15). The mutation burden per cell increased with age in a clock-like fashion at a rate of 40 substitutions per cell per year, demonstrating that himut is applicable in non-human samples.

Thereafter, I use himut to detect somatic mutations in approximately 600 eukaryotic species from the DToL project. I discovered XX number of mutational signatures where SB1 and SBS5-like mutational signatures were previously reported through mutational signature analysis of somatic mutations in cancers \cite{} and where X number of mutational signatures (SBSX1,SBSX2, SBSX3) were new mutational signatures with an unknown aetiology. In addition, under the assumption that these mutational processes are conserved across time, I was also able to time the emergence of some of these mutational processes according to their phylogenetic origin and relationship. The ubiquitous presence of SBS1 mutational signature in the fungi, animalia and plantae kingdom, for example, suggests that methylation of CG dinucleotides is as old as eukaryotes themselves.

Germline mutation is the product of somatic mutations in germ cells and inheritance of these \textit{de novo} mutations from one generation to the next before and after speciation. As the ancestral allele of germline mutation cannot be determined without sequence alignment with outgroup species, germline mutational processes often cannot be determined without \textit{de novo} mutation detection through trio-sequencing. I, however, was able to use the newly extracted somatic mutational signatures to determine the germline mutational process in each species and the relative contribution of each germline mutational process in shaping the sequence context.

**\section{Limitations}**

**\subsection{CCS errors}**

CCS library preparation, sequencing and consensus sequence generation algorithm is currently not optimised to produce CCS reads where CCS bases are assigned the correct base-specific error probabilities. I limited the analysis to Q93 CCS bases as library errors and sequencing errors are unlikely to create substitution errors on both strands of the DNA and for these errors to be propagated to all the subreads during SMRT sequencing. The experimental design, hence, restricts the analysis of errors to cases where error probabilities of Q93 CCS bases are inaccurate or to rare cases where the combination of library and sequencing errors are pervasive in CCS reads and underlying subreads.

In chapter 2, I assess Q93 CCS base accuracy using CCS reads from normal cord blood granulocytes where few somatic mutations are present. I empirically estimate that Q93 CCS base substitution error rate ranges from Q60 to Q90 depending on the substitution and the trinucleotide sequence context. In addition, I show that false positive substitutions are derived from inaccurate base accuracy estimates. What deserves the most attention is that accurate $\sim$Q90 CCS bases can be produced for all trinucleotide sequence contexts if there are enough subreads and if correct error probabilities for subread bases are used for consensus sequence generation.

In chapter 3, I observed a somatic mutational spectrum from several species where 1) the number of called somatic mutations were greater than that from germline mutations, 2) the somatic mutational spectrum was noticeably different from the germline mutational spectrum, and finally 3) the somatic mutation spectrum was shared between phylogenetically unrelated species. In this PhD thesis, I do not investigate the origin of this somatic mutational spectrum or the downstream consequences to germline mutation detection and to assembly quality, but I hypothesise library errors to be the primary source of this erroneous somatic mutational spectrum as CCS library preparation is the only common factor in all the samples exhibiting this issue.

**\subsection{Single-molecule somatic mutation detection}**

In contrast to single-molecule resolution somatic mutation detection using duplex reads from the nanorate sequencing protocol, himut cannot ascertain at all trinucleotide sequence contexts whether an individual substitution, where a single read supports the mismatch between the read and the reference genome, is an error or a somatic mutation. In a clinical setting, where himut might be used to detect the earliest transformation of a normal tissue to a neoplastic tissue or to monitor tumour regression and relapse after treatment, the accuracy of every somatic mutation call is critical in helping the clinician arrive at the correct clinical interpretation. Any false-positive or false-negative mutation call could have serious consequences for the patient’s treatment and prognosis.

Multiple mutational processes act on the genome at any given time and they can generate the same sequence-context specific mutation. Given a catalogue of somatic mutations from multiple samples, mutational signature analysis identifies the mutational signature in each sample and the contribution of each mutational signature to the mutation burden of the sample (eq. \ref{eq:1}). Each mutational signature represents the probability that a specific somatic mutational process will produce a somatic mutation in a specific sequence context.

\begin{align}

\begin{split}

M &\approx PE \label{eq:1} \\

\begin{bmatrix}

    m^{1}\_{1} & \dots & m^{1}\_{j} \\

    \vdots & \ddots & \vdots \\

    m^{96}\_{1} & \dots & m^{96}\_{j} \\

\end{bmatrix} &\approx

\begin{bmatrix}

    p^{1}\_{1} & \dots & p^{1}\_{s} \\

    \vdots & \ddots & \vdots \\

    p^{96}\_{1} & \dots & p^{96}\_{s} \\

\end{bmatrix} \times

\begin{bmatrix}

    e^{1}\_{1} & \dots & e^{s}\_{j} \\

    \vdots & \ddots & \vdots \\

    e^{s}\_{1} & \dots & e^{s}\_{j} \\

\end{bmatrix}

\end{split}

\end{align}

where $M$ is the somatic mutation catalogue matrix with mutation type as rows and samples as columns. $P$ is the mutational signature matrix with mutation types as rows and signatures as columns. $E$ is the exposure matrix with signatures as rows and samples as columns. Here, I use the mutation types as defined by the SBS96 classification system for illustration purposes.

If the mutational processes and associated mutational signatures in the genome of interest are known, it is possible to calculate the probability that a given somatic mutation $m$ in sample $j$ originates signature $s$ can be estimated (eq. \ref{eq:2}).

\begin{equation} \label{eq:2}

P(m,s) = \frac{p^{i}\_{s} \times e^{s}\_{j}}{\sum^{n}\_{s=1}p^{i}\_{s} \times e^{s}\_{j}}

\end{equation}

This approach previously was used to determine SBS16 mutational signature, a signature associated with alcohol consumption, as the main source of somatic mutations in CTNNB1 gene in hepatocellular carcinoma \cite{}. Until the generation of error-free CCS bases, these posterior probabilities can serve as a measure of confidence for individual somatic mutations where single molecule somatic mutations are called from bulk normal tissue.

**\section{Future directions}**

\textit{See things not as they are, but as they might be}

\begin{flushright} [J. Robert Oppenheimer] \end{flushright}

In the imminent future, I conjecture that CCS library errors and inaccurate CCS BQ score estimation will be properly addressed and that the majority (>50\%) of CCS bases will have $\sim$Q90 base accuracy. Here, I discuss the potential opportunities following this development.

**\subsection{Single-molecule real-time sequencing}**

\textit{Nothing is more powerful than an idea whose time has come}

\begin{flushright} [Victor Hugo] \end{flushright}

As discussed in chapter 1, CCS sequence throughput and sequencing cost per base is a function of the number of ZMWs and the read-of-insert length of the SMRTbell template. As demonstrated in chapter 2 and chapter 3, CCS base accuracy is a function of subread error rate and the number of subreads per CCS read, under the assumption that there are no new errors are introduced during consensus sequence generation. Here, I make some informed predictions about the forthcoming advancements in the SMRT sequencing platform based on observations made in this PhD thesis.

I expect the number of ZMWs per SMRTcell to double every two to three years like how Moore’s Law predicts the number of transistors per chip to double every two years. As the number of ZMWs per SMRTcell increases exponentially, sequencing cost per base is expected to decrease exponentially as well (Figure \ref{}). Moore’s law has continued for approximately $\sim$50 years and similar performance increases can be expected from SMRTcell as well. In addition, as CCS sequence throughput is directly proportional to CCS read length, the rate at which CCS sequence throughput increases could also exceed all our expectations due to the combined effect of parallel increases in both CCS read length and number of ZMWS per SMRTcell.

To make accurate predictions about future technological advances, I must also consider Wright’s Law, a companion of Moore’s Law. Wright’s Law, also known as experience curve effect, states that for every cumulative doubling of units produced, costs will fall by a constant percentage. As discussed above, the rapid decrease in CCS sequencing costs will accelerate the adoption of CCS sequencing and when economies of scale is achieved, the positive flywheel effect will be unstoppable (Figure \ref{}).

DNAP processivity, the rate at which DNAP synthesises a new strand of DNA, is another crucial factor that determines CCS base accuracy and sequence throughput. If subread error rate is between 10\% and 15\%, CCS generation typically requires at least 10 subreads per CCS read to generate CCS bases with Q93 base accuracy. DNAP processivity determines the number of subreads per CCS read and hence, increasing DNAP processivity translates to increasing CCS read length. If, for example, DNAP processivity is doubled, CCS read length can also be doubled without sacrifices in base accuracy. DNAP’s biological limit, hence, will be the theoretical limit of CCS read length. In addition, DNAP replication error rate is not an obstacle to $\sim$Q90 CCS base generation at all trinucleotide sequence contexts as demonstrated in chapter 2. One interesting ramification of improved base accuracy is that pooled and non-barcodes samples can be sequenced together to detect common germline SNPs as lower sequence coverage is needed to call the germline mutations with confidence. In addition, as the number of somatic mutations increases linearly with sequence coverage, CCS sequence coverage will not determine the confidence with which germline mutations are detected, but the number of somatic mutations that are detected from the sample.

In contrast, Illumina’s sequencing by synthesis approach has several disadvantages that limit improvements in read length and base accuracy. In each Illumina sequencing cycle, the rate at which a growing DNA becomes asynchronous with the rest of DNA fragments from the same cluster increases, resulting in a reduced signal-to-noise ratio as sequencing progresses \cite{}. This technical limitation places a ceiling on Illumina read length and is responsible for decline in per-base sequence quality towards the end of the read \cite{}. In addition, CCS sequence throughput is a polynomial function with read length and number of ZMWs per SMRTcell as input while Illumina sequence throughput is a linear function with number of clusters per flow cell as the only input. Consequently, CCS sequencing possesses a greater potential for improving sequence throughput. Considering the fact that CCS reads enables \textit{de novo} assembly and simultaneous detection of haplotype phased somatic and germline mutations and epigenetic modifications, I believe that CCS sequencing will be the primary DNA sequencing method in clinics and research in the imminent future.

**\subsection{Somatic mutation detection}**

Here, I discuss the potential future improvements to himut and its applications. In the future, when error-free native DNA CCS library preparation is possible and when CCS BQ scores are correctly calibrated, HMW DNA input requirements and sequence coverage are the only limiting factors for the examination of somatic mutagenesis across all species and all tissues.

In the interim, I believe that a wider range of somatic mutation detection will be possible with the benchmarking approach I have established. A sample with a known double base substitution and indel somatic mutational process, for example, can be sequenced to optimise the pbccs algorithm and improve himut sensitivity and specificity. UV light, for example, induces the photoexcitation and dimerisation of adjacent pyrimidines into cyclobutane pyrimidine dimer (CPD) and 6-4 photoproduct. CPD deamination has been suggested as one of the mechanisms generating C>T mutations (SBS7abc) and CC>TT mutations (DBS1) \cite{}. In addition, exposure to cisplatin, a commonly used chemotherapy drug, generates a unique indel mutational signature characterised by the introduction of a single T insertion downstream of GG dinucleotide \cite{}, which is thought to arise from nucleotide excision repair of 1-3d(GpXpG) intra-strand cisplatin adducts \cite{}.

**\subsubsection{Public health}**

The All of Us (AoU) research program aims to sequence the genomes and collect electronic health records (EHR) data of at least one million individuals in the United States from under-represented demographic categories to accelerate biomedical research \cite{}. In addition, the Human Genome Structural Variation Consortium (HGSVC) has actively adopted CCS sequencing to detect and analyse haplotype phased structural variations, recalcitrant to detection using short read sequencing, from various ethnic groups \cite{}. Genomics England is also considering the use of long-read sequencing to enhance the diagnosis rate of rare genetic diseases \cite{}.

The AoU research program and other population-scale sequencing efforts can use himut to investigate somatic mutagenesis across all their samples where CCS reads are available, just as I have used himut to study somatic mutational processes across the tree of life. The sheer number of samples sequenced under the AoU research program will enable the discovery of new mutational signatures resulting from environmental mutagenesis, DNA damage and mismatch repair deficiencies, as well as their possible combinations. Moreover, AoU research program can also leverage the EHR records (e.g. age of the sample, geographical location, dietary and drinking habits and drug prescription history) to develop and evaluate hypotheses about the etiology of the newly discovered mutational signatures.

The tobacco smoking mutational signature (SBS4) is a canonical example of a mutational signature where exogenous exposure to a mutagen (tobacco carcinogen) is responsible for somatic mutagenesis. The elevation of mutation burden attributable to SBS4 in smokers compared to non-smokers suggests that lung cancer, linked to tobacco smoking, is a preventable disease \cite{}. Aristolochic acid (AA) consumption, often through traditional Chinese medicine, is an under-recognised source of somatic mutagenesis (SBS22) and is a major contributor to endemic Balkan nephropathy \cite{} and urinary tract urothelial carcinoma in Taiwan \cite{}. The discovery of somatic mutagenesis resulting from inadvertent or involuntary exposure to carcinogens, hence, might be one of the most intriguing outcomes of population-scale CCS sequencing efforts.

**\subsubsection{DNA forensics}**

DNA fingerprinting is often used in criminal investigations to determine whether the reference DNA sample from the crime scene and DNA sample from the suspect is derived from the same individual \cite{}. If the genetic sequence of two random individuals are compared, 99.9\% of their genetic sequences are estimated to be identical \cite{}. The number of core repeat motifs in variable number tandem repeat (VNTR) loci, however, is unique to each individual and DNA fingerprinting leverages variation in VNTR loci as a unique genetic fingerprint to compare and match DNA samples from different individuals.

The age of the sample is another unique biomarker that can facilitate the identification of an individual. Cells accrue somatic mutations in a clock-like fashion. Haematopoietic stem cells, for example, acquire 16.8 substitutions per cell per year \cite{} while sperm cells with the lowest somatic mutation rate accumulate 2.9 substitutions per cell per year \cite{}. In addition to the ability to call somatic mutations, himut can also calculate the mutation burden per cell, based on the number of somatic mutations and the number of reference bases and CCS bases from which the mutations were detected. The age of the sample at the time it was collected, therefore, can be estimated by dividing the mutation burden per cell by the tissue-specific somatic mutation rate. The age of the sample in question, thereafter, can either help investigators narrow the number of suspects or free innocent individuals.

**\subsection{Strand-specific somatic mutation detection}**

To date, somatic mutation detection in normal tissues and tumours with next-generation sequencing have focused on analysing sub-clonal or clonal somatic mutations that are fixed in a group of cells above the limit of detection threshold. Single-molecule resolution and strand-specific base modification somatic mutation detection has the potential to enhance our understanding of somatic mutagenesis.

As described in chapter 1, DNAP sequences both the forward and reverse strand of the SMRTbell template multiple times through rolling circle replication. The pbccs algorithm leverages the redundancies and complementary base pairing between the forward and reverse strand subreads to generate CCS reads. As demonstrated in chapter 2, CCS reads have sufficient base accuracy for single molecule somatic mutation detection and as described in a previous publication, single-molecule resolution 5mC detection is also possible from CCS DNAP kinetics \cite{Vong2019-bi, Tse2021-or}. The pbccs algorithm can also generate single-strand consensus sequence (SSCS) reads from the forward and reverse strand subreads. Strand-specific somatic mutation and base modification detection with SSCS reads presents an exciting opportunity to analyse somatic mutations from their beginning to their end.

Somatic mutation is a three-step process: 1) DNA damage, mutation, or modification from endogenous or exogenous sources, 2) failure to detect and repair the DNA damage or mutation, and 3) fixation, persistence of DNA mutation in daughter cells through genetic drift or selection. In a population of DNA molecules, there will be a group of wild type DNA molecules, a group of DNA molecules with DNA damage, mutation or modification, a group of DNA molecules undergoing DNA damage repair and a group of DNA molecules with new somatic mutations.

The DNA damage and repair process associated with SBS1 mutational signature, for example, is amenable to further qualitative and quantitative examination through this approach. The spontaneous deamination of 5mC to thymine results in a TG:GC mismatch and results in C>T somatic mutation at a CG dinucleotide if left unrepaired by the mismatch repair (MMR) pathway. If both strands of the double-stranded DNA molecule are sequenced, TG dinucleotide will be present on the strand where deamination has happened and GC dinucleotide with methylation will be present on the complementary strand. SSCS reads from SMRT sequencing enable the detection of TG:GC mismatches and associated hemi-methylation (Figure \ref{}). In addition, CCS reads allow the estimation of the number of methylated CG dinucleotides where deamination could have happened and the number of CG dinucleotides where somatic mutations have occurred (Figure \ref{}). If the same tissue is sequenced at multiple different timepoints, the gain and loss of somatic mutations in the population can also be studied (Figure \ref{}).

If successful, we will be able to measure the \textit{in vivo} deamination rate from the number of TG:GC mismatch and the number of GC dinucleotides, and compare it against the \textit{in vitro} deamination rate of $5.8x10^{-13}$ per 5mC per second at 37°C \cite{}. In addition, TG:GC mismatch repair efficiency and fidelity can also be measured under wild-type and mutant conditions. MutS$\alpha$, for example, is critical in recognising the TG:GC mismatch and initiating DNA damage repair. MutS$\alpha$ deficiency, therefore, elevates the number of C>T somatic mutations. Similarly, cross-examination of both SSCS and CCS reads and associated DNAP kinetics can also be used to better understand the C>T (SBS2) somatic mutations resulting from APOBEC-dependent deamination of cytosine to uracil.

**\subsubsection{Decomposition of a mutational signature}**

Single-molecule resolution and strand-specific base modification somatic mutation detection, most importantly, creates an opportunity to gain greater insights into the dynamics of somatic mutational processes. Each somatic mutational process leaves a characteristic imprint to the genome and mutational signatures represent the probability that a specific somatic mutational process will produce a somatic mutation in a specific sequence context. Given a catalogue of somatic mutations from multiple samples, mutational signature analysis identifies the mutational signature in each sample and the contribution of each mutational signature to the mutation burden of the sample (eq \ref{}).

Since each mutational signature is a cumulative result of DNA damage, mutation or modification, failure to repair the DNA damage or mismatch, and persistence of the mutation in bulk tissue, each mutational signature can be re-defined as such (eq. \ref{eq:3})

\begin{align}

\begin{split}

\alpha D \cdot \beta R \cdot \gamma F &\approx P\_{i} \label{eq:3} \\

\alpha \begin{bmatrix}

    d^{1}\_{1}  \\

    \vdots &  \\

    d^{96}\_{1}  \\

\end{bmatrix}

\beta \begin{bmatrix}

    r^{1}\_{1} \\

    \vdots &  \\

    r^{96}\_{1} \\

\end{bmatrix}

\gamma \begin{bmatrix}

    f^{1}\_{1}  \\

    \vdots &  \\

    f^{96}\_{1}  \\

\end{bmatrix} &\approx

\begin{bmatrix}

    p^{1}\_{1} \\

    \vdots \\

    p^{96}\_{1} \\

\end{bmatrix}

\end{split}

\end{align}

where $D$ is the DNA damage matrix and each element is a probability that a specific sequence context will be damaged. $R$ is DNA damage repair matrix and each element is a probability that DNA damage in a specific sequence context will be repaired. $F$ is DNA mutation fixation matrix and each element is a probability that the mutation type will be fixed in the population. $\alpha$, $\beta$, $\gamma$ are scalar values that represent genetic and environmental factors that modulate the somatic mutational process. In addition, $R$ could be further decomposed into multiple subcomponents where each matrix represents a different DNA damage repair pathway specific to the DNA damage (eq. \ref{eq:4})

\begin{equation} \label{eq:4}

\alpha \begin{bmatrix}

    d^{1}\_{1}  \\

    \vdots &  \\

    d^{96}\_{1}  \\

\end{bmatrix} \

\beta \left[\begin{bmatrix}

    r^{1}\_{i} \\

    \vdots &  \\

    r^{96}\_{i} \\

\end{bmatrix}

\begin{bmatrix}

    r^{1}\_{i+1} \\

    \vdots &  \\

    r^{96}\_{i+1} \\

\end{bmatrix} \ldots

\begin{bmatrix}

    r^{1}\_{n} \\

    \vdots &  \\

    r^{96}\_{n} \\

\end{bmatrix}\right]

\gamma \begin{bmatrix}

    f^{1}\_{1}  \\

    \vdots &  \\

    f^{96}\_{1}  \\

\end{bmatrix} \approx

\begin{bmatrix}

    p^{1}\_{1} \\

    \vdots \\

    p^{96}\_{1} \\

\end{bmatrix}

\end{equation}

The decomposition of a mutational signature into their individual components and subcomponents should enable us to have a greater understanding of the nonlinear relationship between the components and the mechanisms underlying somatic mutagenesis.

**\subsection{Gene conversion and crossover detection}**

Here, I also hypothesise that CCS read length and base accuracy can be leveraged to detect gene conversion and crossover events generated during meiotic and mitotic recombination. Gene conversion (Fig\ref{}) and crossover (CO) (Fig \ref{}) arise from the non-reciprocal and reciprocal exchange of genetic material during double-strand break (DSB) repair through homologous recombination. Gene conversions are also referred to as non-crossovers (NCO), but gene conversions and crossovers are not mutually exclusive events \cite{}.

In germ cells, meiotic recombination is an essential process that generates new combinations of alleles that serve as the foundation for adaptation and speciation through natural selection, an advantage for sexually reproducing organisms. In addition, the formation of at least one chiasma per pair of homologous chromosomes ensures proper segregation of chromosomes in anaphase I of meiosis. Improper chromosome segregation can result in aneuploid gametes with abnormal numbers of chromosomes. If DSB repair is not repaired in somatic cells, DNA damage response initiates programmed cell death. It is worth noting that DSB repair during meiotic recombination generates new allele combinations and contributes to genetic diversity, while DSB repair during mitotic recombination can result in the loss of heterozygosity (LOH). Furthermore, meiotic DSBs are deliberately introduced through the concerted action of PRDM9 and SPO11 to initiate meiotic recombination \cite{} while mitotic DSBs are inadvertently generated from both endogenous (e.g. reactive oxygen species) and exogenous factors (e.g. ionising radiation) \cite{}.

Two distinct types of recombination products — [crossovers](https://www.nature.com/articles/nrg3573#Glos2) and [non-crossovers](https://www.nature.com/articles/nrg3573#Glos3) — are generated by recombination between homologues. Crossover recombination products, which consist of reciprocal exchanges between homologues, result in the exchange of alleles over large chromosomal intervals. Conversely, non-crossovers, which are thought to outnumber crossovers by a factor of ten in mice and humans, involve only a unidirectional transfer of genetic information (or [gene conversion](https://www.nature.com/articles/nrg3573#Glos4)) over short intervals and therefore have only a limited, local effect on genetic diversity

An understanding of the genomic landscape of human recombination rate variation would facilitate the efficient design and analysis of disease association studies and greatly improve inferences from polymorphism data about selection and human demographic history.

. Fine-scale recombination rate estimates would also provide a new route to understanding the molecular mechanisms underlying human recombination. Current approaches cannot provide this information: Pedigree studies do not have the required resolution, whereas sperm analyses can only detect recombination rate variation in males and are impracticable for studies on chromosomal scales.

Patterns of genetic diversity and LD are shaped by many factors ([*15*](https://www.science.org/doi/10.1126/science.1092500?url_ver=Z39.88-2003&rfr_id=ori:rid:crossref.org&rfr_dat=cr_pub%20%200pubmed#core-REF15)), mutation, recombination, selection, population demography, and genetic drift.

Finally, it can provide insights into the evolutionary history of different populations and the mechanisms that have shaped their genetic diversity over time.

ecombination rates in higher eukaryotes are generally measured using (1) genetic comparisons between parents and offspring (e.g., using genotype or sequence data), (2) genotyping or sequencing of single or pooled sperm (i.e., potential gametes), or (3) indirect estimation via statistical methods that quantify the relationship between linkage disequilibrium (LD) and recombination.

Each of these three approaches involves tradeoffs regarding cost/effort and the breadth and depth of information they can provide. In particular, only pedigree-based studies provide both sex-specific recombination estimates and direct identification of both crossover (CO) and noncrossover (NCO) recombination events, but they are more difficult to conduct due to sample acquisition challenges.

Maternal reciprocal translocation Paternal reciprocal translocation

This process breaks linkage disequilibrium, contributes to genetic exchange across generations and provides genetic diversity for natural selections ([Coop and Przeworski, 2007](https://www.ncbi.nlm.nih.gov/pmc/articles/PMC6956785/#btz503-B4)).

Because meiotic recombination involves rejoining of two homologous chromosomes, the two homologous chromosome fragments need to be distinguished in order to identify the crossover events. The process to resolve the haplotype of homologous chromosomes is called phasing. Traditional phasing methods rely on certain genotyped SNP markers to discover crossover events. For example, the identity by descent method ([Lee et al., 2011](https://www.ncbi.nlm.nih.gov/pmc/articles/PMC6956785/#btz503-B15)) uses a three-generation dataset and selects candidate SNPs which are heterozygous in a parent, but homozygous in at least one grandparent. The identity by state method ([Lee et al., 2011](https://www.ncbi.nlm.nih.gov/pmc/articles/PMC6956785/#btz503-B15)) uses a two-generation pedigree and compares the patterns of SNPs between siblings. Although genotyping-based approaches have shown great success in revealing crossover patterns in human populations, the requirement for grandparent or sibling genomes is not applicable to a typical trio dataset. In addition, the resolution of crossover breakpoint regions is over 100 kb ([Lee et al., 2011](https://www.ncbi.nlm.nih.gov/pmc/articles/PMC6956785/#btz503-B15)). Even with large related populations and high-density SNP arrays, only 20% of crossover events can be narrowed down to 30 kb genomic regions ([Coop et al., 2008](https://www.ncbi.nlm.nih.gov/pmc/articles/PMC6956785/#btz503-B5)).

During mammalian meiosis, double-strand breaks are deliberately made throughout the genome and then repaired, leading to the exchange of genetic material between copies of chromosomes.

Locations of crossovers were determined from haplotype phase transitions in parent–proband pairs based on informative markers where the parents were heterozygous. Initially, the location was determined by the 2 closest heterozygous markers to the cross- over, providing upper and lower bounds for the location.

Sequence context

Population structure and recombination amp/ genetic maps

The maps of human recombination

The most detailed maps of human recombination have been generated by means of computational inference of recombination rates from patterns of linkage disequilibrium (LD) in the human population ([*12*](https://www.science.org/doi/10.1126/science.1256442#core-R12)–[*15*](https://www.science.org/doi/10.1126/science.1256442#core-R15)).

These maps do not, however, provide information about recombination rates in individuals.

Recently, high-resolution recombination maps were inferred from high-density single-nucleotide polymorphism (SNP) data using linkage disequilibrium (LD) patterns that capture historical recombination events[1](https://www.nature.com/articles/nature09525#ref-CR1),[2](https://www.nature.com/articles/nature09525#ref-CR2).  The use of these maps has been demonstrated by the identification of recombination hotspots[2](https://www.nature.com/articles/nature09525#ref-CR2) and associated motifs[3](https://www.nature.com/articles/nature09525#ref-CR3), and the discovery that the *PRDM9* gene affects the proportion of recombinations occurring at hotspots[4](https://www.nature.com/articles/nature09525#ref-CR4),[5](https://www.nature.com/articles/nature09525#ref-CR5),[6](https://www.nature.com/articles/nature09525#ref-CR6).  However, these maps provide no information about individual or sex differences. Moreover, locus-specific demographic factors like natural selection[7](https://www.nature.com/articles/nature09525#ref-CR7) can bias LD-based estimates of recombination rate. Existing genetic maps based on family data avoid these shortcomings[8](https://www.nature.com/articles/nature09525#ref-CR8), but their resolution is limited by relatively few meioses and a low density of markers.

Current approaches to study the early steps of meiotic recombination in humans primarily detect genetic crossovers, only one of the possible outcomes of DSB repair.

To perform a large, family-based recombination study, one challenge is to phase the genotypes of the parents when the grandparents are not genotyped. One solution is to use genotyped nuclear families with two or more offspring, which in essence uses the children to phase the parents.

 The data only allowed us to assign a recombination to the region spanned by the two closest flanking heterozygous markers in the parent ([Fig. 1](https://www.nature.com/articles/nature09525#Fig1)).

The nature and scale of recombination rate variation are largely unknown for most species. In humans, pedigree analysis has documented variation at the chromosomal level, and sperm studies have identified specific hotspots in which crossing-over events cluster.

In meiosis, a DSB can be repaired as either a crossover or as a noncrossover ([Fig. 6A](https://www.science.org/doi/10.1126/science.1256442#F6)) ([*26*](https://www.science.org/doi/10.1126/science.1256442#core-R26)). Because the proportion of DSBs resolved as crossovers might vary from hotspot to hotspot ([*11*](https://www.science.org/doi/10.1126/science.1256442#core-R11)), the frequency of crossing over need not necessarily reflect the DSB formation rate. We thus asked whether the crossover landscape is largely shaped by variation in crossover/noncrossover resolution or whether it is mostly determined by the DSB frequency.

But, in fact, it is not known how widespread hotspots are in the human genome, neither is the magnitude of rate differences, nor the physical scales over which this occurs, known.

An understanding of the genomic landscape of human recombination rate variation would facilitate the efficient design and analysis of disease association studies and greatly improve inferences from polymorphism data about selection and human demographic history.

. Fine-scale recombination rate estimates would also provide a new route to understanding the molecular mechanisms underlying human recombination. Current approaches cannot provide this information: Pedigree studies do not have the required resolution, whereas sperm analyses can only detect recombination rate variation in males and are impracticable for studies on chromosomal scales.

Patterns of genetic diversity and LD are shaped by many factors ([*15*](https://www.science.org/doi/10.1126/science.1092500?url_ver=Z39.88-2003&rfr_id=ori:rid:crossref.org&rfr_dat=cr_pub%20%200pubmed#core-REF15)), mutation, recombination, selection, population demography, and genetic drift.

 Meiotic recombination is initiated from double-strand breaks (DSBs). DSBs occur more frequently in regions of the genome termed hotspots, and a small subset eventually gives rise to crossovers, a reciprocal exchange of large pieces between homologous chromosomes.

The majority of DSBs do not lead to crossovers but end as localized transfers of short segments between homologous chromosomes or sister chromatids, observable as gene conversions when the segment includes a heterozygous marker. Crossovers co-occurring with distal gene conversions are known as complex crossovers.

Current meiotic recombination maps either have limited resolution or the events cannot be resolved to an individual level. The detection of recombination and de novo mutations (DNMs) requires genetic data on a proband and its parents, and a fine resolution of these events is possible only with whole-genome sequence data. Whole-genome sequencing and DNA microarray data allowed us to identify crossovers and DNMs in families at a high resolution. We resolved crossovers at an individual level, allowing us to examine variation in crossover patterns between individuals, analyzing which crossovers are complex and how crossover patterns are influenced by age, sex, sequence variants, and epigenomic factors. It is known that the mutation rate is increased near crossovers, but the rate of DNMs near crossovers has been characterized only indirectly or at a small scale.

Recombination is initiated through the formation of double-strand breaks (DSBs) catalyzed by SPO11 ([*3*](https://www.science.org/doi/10.1126/science.aau1043?url_ver=Z39.88-2003&rfr_id=ori:rid:crossref.org&rfr_dat=cr_pub%20%200pubmed#core-R3), [*4*](https://www.science.org/doi/10.1126/science.aau1043?url_ver=Z39.88-2003&rfr_id=ori:rid:crossref.org&rfr_dat=cr_pub%20%200pubmed#core-R4)). A large number of DSBs are generated across the genome ([*5*](https://www.science.org/doi/10.1126/science.aau1043?url_ver=Z39.88-2003&rfr_id=ori:rid:crossref.org&rfr_dat=cr_pub%20%200pubmed#core-R5)), of which only a small subset eventually gives rise to crossovers, which result in the exchange of sequences flanking the crossover point and yield recombinant chromosomes ([*6*](https://www.science.org/doi/10.1126/science.aau1043?url_ver=Z39.88-2003&rfr_id=ori:rid:crossref.org&rfr_dat=cr_pub%20%200pubmed#core-R6)).

We constructed a sequence-level genetic map pinpointing the locations of crossovers in individuals to a subkilobase resolution. The locations of crossovers were determined from haplotype phase transitions in parent-offspring pairs, at loci where the parent is heterozygous ([Fig. 1A](https://www.science.org/doi/10.1126/science.aau1043?url_ver=Z39.88-2003&rfr_id=ori:rid:crossref.org&rfr_dat=cr_pub%20%200pubmed#F1)). The sequence data resolution allows us to identify the crossovers that are accompanied by distal gene conversion events, hereinafter referred to as complex crossovers. The subkilobase resolution of crossovers provides an opportunity to examine the correlation of any sequence attribute with crossovers, analyze differences between individuals, and determine whether crossovers occur in an age-dependent manner

As the resolution of crossovers is inherently limited by the interval between heterozygous markers and as the WGS data used in this study capture almost all such markers, we expect that future genetic maps may yield only minor improvements in resolution.

bulk

Sperm,

Eggs

The process of gene conversion can occur during homologous recombination, particularly when there is an unequal crossover between the two homologous chromosomes. In this process, one chromosome donates a longer segment of DNA to the other chromosome, and the recipient chromosome may replace some or all of its original sequence with the donor sequence. The result is a hybrid chromosome that contains a mixture of genetic material from both parents.

Gene conversion can also occur spontaneously as a result of errors in DNA replication or repair, which can cause mismatches between the two DNA strands. These mismatches can be corrected by the cellular machinery that recognizes and repairs DNA damage, leading to the replacement of one allele with a copy of the other allele.

Gene conversion plays a role in the evolution of species by promoting genetic diversity and facilitating the spread of advantageous mutations. However, it can also contribute to genetic disorders and diseases, particularly when it occurs in genes that are involved in DNA repair or other critical cellular processes.

A crossover is a type of genetic recombination that occurs during meiosis, the process by which diploid cells divide and produce haploid gametes such as sperm and eggs. During meiosis, homologous chromosomes pair up and exchange segments of DNA, resulting in the production of genetically diverse offspring.

Trio and sperm sequencing has often been used to detect approximately one meiotic recombination product per child per chromosome \cite{} and to determine meiotic recombination rate at a specific target locus \cite{}, respectively.

Illumina short read sequencing alone cannot detect recombination products as short reads do not provide long-range haplotype information and base accuracy is not sufficient to determine whether the change in heterozygous SNPs is from sequencing errors or from recombination.

Gene conversion was first used to describe a phenomenon where there is a non-Mendelian segregation of heterozygous markers.

Illumina sequencing alone cannot detect recombination products as long-range haplotype phasing is required to

Previously, to detect gene conversions and crossover, a trio-sequencing was done or sperm-typing was done. Trio-sequencing, however, can only capture 1 meiotic event per chromosome per child while sperm-typing is restricted to a known hotspot. Our approach, however, assesses gene conversions and crossovers across the genome where there is sufficient sequence coverage and hetSNP density to haplotype phase the target region.

We tackled another original question to assess the genome-wide meiotic and mitotic recombination products in sperm samples and Bloom syndrome patient samples and compare and contrast characteristics of meiotic and mitotic recombination. Gene conversions and crossover detection requires long-range haplotype phasing of hetSNPs and individual reads to detect recombinant products that contains both maternal and paternal hetSNPs. The standard Illumina reads, unfortunately, cannot be used haplotype phase multiple hetSNPs at a time while CCS reads with their longer read length and is able to span multiple hetSNPs. CCS reads also have sufficient base accuracy to have confidence that the hetSNP flip is a result of not sequencing error, but a biological process. We successfully demonstrate that not only single molecule somatic single-base-substitution detection is possible, but also that single molecule gene conversion and crossover detection is possible with CCS reads. The detected gene conversion and crossovers are located on known meiotic recombination hotspots.

Double strand break (DSB) repair is a mechanism by which cells repair breaks that occur in both strands of DNA, which can be caused by endogenous processes such as DNA replication and recombination, or exogenous factors such as radiation and chemical damage.

n HR, the broken ends of the DNA are resected by nucleases to generate single-stranded DNA (ssDNA) overhangs, which are then coated with the protein complex RAD51. RAD51 forms a nucleoprotein filament that searches for and pairs with a homologous sequence in an undamaged sister chromatid or a homologous chromosome. The ssDNA then serves as a template for DNA synthesis to repair the break, resulting in accurate repair with no loss or gain of genetic material.

The PRDM9 protein contains several functional domains, including a PR domain, which is responsible for regulating gene expression by modifying chromatin structure, and a DNA-binding domain consisting of multiple C2H2 zinc finger motifs. The zinc fingers are responsible for recognizing and binding to specific DNA sequences, known as hotspots, that are prone to crossing over during meiosis.

The structure of the PRDM9 gene and protein is highly variable between different species, with different numbers and arrangements of zinc finger motifs and hotspots. This diversity is thought to contribute to the evolution of species by generating new combinations of alleles through recombination during meiosis.

Recognition and processing of the break: A double-strand break is recognized by the DNA damage response machinery, which recruits proteins that cleave the DNA ends and generate single-stranded DNA (ssDNA) overhangs.

Strand invasion: The ssDNA overhangs are coated with the recombinase enzyme RAD51, which forms a nucleoprotein filament that searches for a homologous sequence in the undamaged sister chromatid or a homologous chromosome. The filament displaces one of the strands of the homologous duplex DNA and forms a D-loop structure.

DNA synthesis and branch migration: DNA synthesis initiates at the 3' end of the invading ssDNA strand, using the complementary strand of the homologous chromosome as a template. The DNA synthesis proceeds in the 5' to 3' direction, extending the invading strand and displacing the complementary strand from the D-loop. The newly synthesized strand can then anneal with the complementary strand of the other end of the break, forming a repaired DNA molecule.

Resolution and ligation: The Holliday junction, which is formed as a result of the strand invasion, can be resolved by several mechanisms, including branch migration and resolution by nucleases. The strands are then ligated together to form a repaired DNA molecule.

Homologous recombination is a highly accurate repair mechanism, and it plays a critical role in the maintenance of genome stability, particularly during meiosis, where it ensures proper chromosome segregation and genetic diversity. Defects in HR can lead to various genetic disorders and diseases, including cancer.

Gene conversion is a non-reciprocal transfer of genetic information between two DNA sequences that are similar or identical, such as two alleles of a gene. It occurs when one sequence serves as a template for DNA repair, leading to the replacement of the other sequence with a copy of the template.

The process of gene conversion can occur during homologous recombination, particularly when there is an unequal crossover between the two homologous chromosomes. In this process, one chromosome donates a longer segment of DNA to the other chromosome, and the recipient chromosome may replace some or all of its original sequence with the donor sequence. The result is a hybrid chromosome that contains a mixture of genetic material from both parents.

Gene conversion can also occur spontaneously as a result of errors in DNA replication or repair, which can cause mismatches between the two DNA strands. These mismatches can be corrected by the cellular machinery that recognizes and repairs DNA damage, leading to the replacement of one allele with a copy of the other allele.

Gene conversion plays a role in the evolution of species by promoting genetic diversity and facilitating the spread of advantageous mutations. However, it can also contribute to genetic disorders and diseases, particularly when it occurs in genes that are involved in DNA repair or other critical cellular processes.

A crossover is a type of genetic recombination that occurs during meiosis, the process by which diploid cells divide and produce haploid gametes such as sperm and eggs. During meiosis, homologous chromosomes pair up and exchange segments of DNA, resulting in the production of genetically diverse offspring.

The process of crossover begins when the homologous chromosomes align and form a complex called a synaptonemal complex. Within this complex, the DNA molecules of the two homologous chromosomes break at the same position, creating a double-strand break (DSB) in each chromosome. The two ends of each DSB then invade the homologous chromosome, pairing with a complementary segment of DNA on the other chromosome.

At this point, the DNA strands can exchange segments, resulting in the exchange of genetic material between the two homologous chromosomes. The exchange of genetic material is facilitated by enzymes such as recombinases, which can cleave and rejoin the DNA strands in a process known as homologous recombination.

Once the exchange is complete, the two homologous chromosomes can separate and move to opposite poles of the cell, leading to the production of haploid daughter cells with new combinations of genetic material. Crossover events are relatively rare, occurring at a frequency of about one to three crossovers per pair of homologous chromosomes in humans.

Crossover is an important mechanism for generating genetic diversity within a population, as it creates new combinations of alleles and can promote the spread of advantageous mutations. However, defects in the crossover process can lead to genetic disorders and diseases, including infertility and certain types of cancer.

enetic mapping is a classic technique used to determine the location and frequency of crossovers between genes. This method involves analyzing the segregation of genetic markers, such as single nucleotide polymorphisms (SNPs), in families or populations. By examining the co-inheritance of these markers with different alleles of a gene, researchers can infer the frequency and location of crossovers.

Sperm typing is a molecular biology technique used to detect meiotic recombination events in individual sperm cells. This technique involves isolating and analyzing the DNA from individual sperm cells, which allows researchers to directly observe the genetic makeup of each individual sperm and infer the location and frequency of meiotic crossovers.

The process of sperm typing involves several steps. First, individual sperm cells are isolated from a semen sample or obtained from the testes of experimental animals. The DNA from each individual sperm cell is then extracted and amplified using PCR, typically with primers that target specific regions of the genome that are known to undergo meiotic recombination.

After amplification, the PCR products are analyzed using a variety of techniques, such as gel electrophoresis, DNA sequencing, or microarray analysis, to detect and quantify the occurrence of recombination events. For example, gel electrophoresis can be used to visualize the size and number of PCR products, which correspond to different recombination events. DNA sequencing can be used to directly identify the location and frequency of recombination events at the level of individual nucleotides.

Sperm typing has several advantages over other methods for detecting meiotic recombination, such as genetic mapping and cytogenetics. It allows for the direct observation of individual recombination events in a high-throughput and quantitative manner, and it is not limited by the size or complexity of the genome. This technique has been used to study recombination events in a variety of organisms, including humans, mice, and fruit flies, and has provided insights into the mechanisms and evolution of meiotic recombination.

**\subsection{Tree of Life}**

**\subsubsection{LUCA}**

**\subsubsection{Wildlife conservation}**

**\subsubsection{Embryonic development}**

**\subsubsection{Chloroplast DNA somatic mutations}**

**\subsection{Archaea and prokaryotes}**

Our understanding of germline and somatic mutational processes of non-human species has been limited to date. The availability of both CCS reads and high-quality reference genomes from the Darwin Tree of Life project creates an opportunity to study both germline and somatic mutational processes. We used himut to call somatic mutations across the DToL eukaryotic species, discover XX number of mutational signatures, of which XX were distinct from known COSMIC mutational signatures, indicating the presence of distinct DNA damage and repair process operational in other species. In XX% of species, germline and somatic mutational process were analysed to be similar like how clock-like mutational processes (SBS1 and SBS5) are responsible for germline mutagenesis in sperms and oocytes. In addition, some of these endogenous somatic mutational processes were shared in insects, which are known to have diverged 450 million years ago (mya), suggesting the mutational signature that we have discovered might be an ancient somatic mutational process or that these insects independently developed the same mutational process. Mother Nature, however, often doesn’t change if there is an existing solution unless there is immense selection pressure and the author believes that the mutational process has been conserved across insects.

In XX% of species (hoverflies), however, germline mutational process and somatic mutational process were discordant and with strong transcription-bias, potentially suggesting environmental mutagenesis might be responsible for the observed somatic mutations. XX, XX, XX and XX insects undergo metamorphosis from caterpillar to adult insect and imaginal discs develop into adult insects. We, conjecture, that the absence of somatic mutations in some of the adult insects that undergo metamorphosis to the fact that larvae form and the adult insects are derived from independent embryonic stem cells. The adult insect is derived from the imaginal disc, which remains inactive under the metamorphosis in the chrysalis stage. Hence, somatic mutation that might have accumulated during the young larvae stage will not be passed on to the adult insect and the adult insect will be able to pass on their genome with limited DNA damage. The absence of somatic mutations in lepidoptera, however, might also be confounded with the short lifespan of the adult insects. It is interesting, however, that insects that undergo metamorphosis account for 80% of the insect population [ref] and there must have been a selective advantage to undergo metamorphosis despite the vulnerability that it might pose to the insect.

Based on our understanding of CCS characteristics, we attempted to search for genomic events that could not be captured with short read sequencing and that could, however, be captured PacBio CCS sequencing. We hypothesised that PacBio CCS reads will also have sufficient base accuracy to detect gene conversions and crossovers from both sperm during meiotic recombination, granulocytes from Bloom syndrome patients and normal individuals during mitotic recombination. Gene conversion and crossover detection necessitates haplotype phasing of multiple kilobases and detection of haplotype rearrangement that might occur in a single sperm or a single cell.

If we had the correct phylogenetic relationship between all species and mutational processes of all species on Earth, could we model and infer the mutational process of extinct species? Could we model and infer the mutational process of LUCA? Could we even derive the genome sequence of LUCA?

If life existed outside of Earth, what might be the mutational process responsible for speciation on other planets? How has Nature on other planets create new species? What is the creative process that Nature uses to create new species? Mutations are the paints that Nature uses to draw the Canvas.

We will be able to determine the ancestral mutational processes that shaped our genomes and the selection and evolution of mutational processes in light of different selection pressures that different environments applied our ancestors. As a consequence, we will also be able to determine the average fidelity of the DNA damage and repair process of all the species.

We don’t know what might be the carrier of information that preserves the biological constraints of life might be on other planets.

The DToL project has sequenced ~600 of 66,000 eukaryotic species in Britain and … As the number

Kimura hypothesises that genetic drift would have been major driver of evolution and we would be happy to test this hypothesis.

The nucleotide composition of also extinct species.

We might be closer than we think on answering the question “What is Life” succinctly proposed by Erwin Schrodinger on XXXX at Dublin.

**\section{Concluding remarks}**

\textit{People don't have ideas. Ideas have people}

\begin{flushright} [Carl Jung] \end{flushright}

I first encountered SMRT sequencing as a bioinformatics scientist at Macrogen in 2015, and the potential of single molecule sequencing was immediately apparent to me. Since then, much of this promise has been realised, while some opportunities remain untapped. At the time, the number of algorithms that could analyse and return meaning results from reads with high error rate was very limited. CLR reads with average read length (>10kb) longer than the common repeats in the human genome, and the ability to \textit{de novo} assemble multi-megabase contigs \cite {} and detect structural variations at nucleotide-resolution \cite{} started to garner excitement from scientists who were frustrated with the limitations of short-read sequencing \cite{}. The simultaneous development of the 3D-DNA Hi-C scaffolding algorithm \cite{}, which can correct assembly errors and order and orient contigs into chromosome-length scaffolds, completely transformed the time and cost to sequence and assemble high-quality reference genomes.

I was particularly excited to explore the unexplored biological phenomenon with SMRT sequencing as a PhD candidate at the Wellcome Sanger Institute. Peter inspired me to evaluate the potential of CCS reads for somatic mutation detection across the tree of life and to design a somatic mutation detection algorithm that will be agnostic of tissue and species. I was successfully able to develop and benchmark himut, with great patience from both Peter and Richard, to detect and analyse somatic mutational processes across different eukaryotic species. During my PhD, I have been constantly amazed by the conservation and the diversification of somatic mutational processes across different kingdoms and phyla, despite the millions of years of evolution that have separated them. In this PhD thesis, I have barely scratched the surface of the somatic mutation landscape in the tree of life and much remains to be discovered.

I have concluded that factors that prevent the adoption of CCS sequencing are technical problems where solutions exist. As discussed in chapter 2, almost error-free CCS bases can be generated and as conjectured in this chapter, CCS sequencing cost and HMW DNA input requirement for CCS library preparation will no longer be a limitation to research. The exponential increase in the number of ZMWs per SMRTcell and the read-of-insert length will be the primary factors driving the increase in sequence throughput and decrease in per-base sequencing cost. The present sequencing methods necessitates a specific DNA input requirement to sequence the genome multiple times and thereby enable the detection of mutations with greater confidence despite the presence of sequencing errors. If DNAP processivity improves to enable CCS library preparation from longer read-of-insert and if CCS base accuracy improves to be error-free, only a single read will be required from each haplotype for germline and somatic mutation detection and epigenetic modification identification, drastically lowering the HMW DNA input requirements for CCS library preparation. I believe that we are witnessing a historic moment where error-free sequencing will be feasible at a fraction of current sequencing costs and where it will be possible to interrogate the genetic, epigenetic, and transcriptomic information of all forms of life.