**\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 X 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 is 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 distinctly 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 assembly contiguity and accuracy, 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 consequence 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. Mutational signature analysis identifies the mutational processes that have acted upon the genome and calculates the contribution of each identified mutational process to the total mutation burden. If the mutational processes and associated mutational signatures in the genome of interest are known, it is possible to calculate the posterior probability that a given somatic mutation originates from either a biological process or a CCS error process (eq \ref{}).

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 lower sequence coverage is needed to call the same number of germline mutations and that the number of called somatic mutations increases linearly with sequence coverage.

In contrast, Illumina’s sequencing by synthesis approach has several disadvantages that limits 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 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 extraction, input requirements for CCS library preparation and sequence coverage of the sample becomes the limiting factor to identifying and studying somatic mutagenesis across all tissues and all species.

Himut currently does not consider matched tumour-normal sequencing for somatic mutation detection, but this would be the natural next step as the number of matched tumour-normal samples sequenced with the SMRT platform is expected to increase with the introduction of the Revio instrument.

Despite this limitation, himut will still enable researchers to rapidly screen for mutational signatures from bulk normal tissue without arduous experiments such as LCM or single-cell clone expansion sequencing, identification of environmental mutagenesis such as exposure to aristolochic acids\cite{} across different locations and populations, lineage trace embryonic and tumour development through accurate detection of mosaic and somatic mutations, respectively.

In the interim, I believe that a wider range of somatic mutation detection will be possible with the benchmarking approach I have established where a sample with a known double base substitution and indel somatic mutational process is sequenced and used to fine-tune 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. Although the exact mechanism that converts UV light DNA damage to DNA mutation is unknown, CPD deamination has been suggested as one of the mechanisms generating C>T mutations (SBS7abc) and CC>TT mutations (DBS1) \cite{}. Cisplatin, a commonly used chemotherapy drug, forms inter-strand DNA crosslinks to prevent DNA replication, which induces cell cycle arrest and apoptosis. Cisplatin produces a unique mutational signature where a single T insertion is introduced downstream of GG dinucleotides \cite{}, which is attributed to nucleotide excision repair of 1-3d(GpXpG) intra-strand cisplatin adducts \cite{}

**\subsubsection{Ultra-fast somatic mutation detection}**

**\subsubsection{Public health}**

**\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 is 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, thereafter, can be estimated by dividing the mutation burden per cell by the tissue-specific somatic mutation rate.

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

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.

To date, somatic mutation detection in normal tissues and tumours have focused on analysing sub-clonal or clonal somatic mutations that are fixed in a group of cells above the limit of detection threshold. The unique capability of the SMRT platform to both generate single-strand consensus sequences (SSCS) and CCS reads, along with the ability to detect epigenetic modifications \cite{Vong2019-bi, Tse2021-or} and somatic mutations at a single-molecule resolution using CCS reads, presents an exciting opportunity to detect and analyse somatic mutational process from the beginning to the end.

The DNA damage and repair process associated with SBS1 mutational signature, for example, is amenable to further qualitative and quantitative examination through SMRT sequencing. 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.

estimate the nonlinear contribution of DNA damage, repair, and mutation fixation process to the SBS1 mutational signature.

Mutational signature is a mathematical abstraction of these three inter-dependent processes (\ref{}) and describes the probability that a given somatic mutational processes will introduce a mutation at a specific sequence context.

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

Here, I also hypothesise that CCS read length and base accuracy can be leveraged for meiotic and mitotic recombination product detection.

genome-wide gene conversion and crossover detection resulting from meiotic and mitotic recombination.

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

In contrast, CCS sequencing of bulk sperm samples will enable haplotype phasing without parental data, genome-wide detection of recombination products and determination of PRDM9 allele specific recombination hotspots.

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.

**\subsection{Tree of Life}**

**\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.

Despite these limitations, as HMW DNA input requirements for CCS library preparation decreases and as sequence throughput and sequencing cost decreases, I believe that PacBio CCS sequencing might be the last DNA sequencing platform to dominate the sequencing market.

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.

A thought experiment

We are still early.

It might be possible to obtain sequence all of life within my lifetime and study/measure evolution in real time.

Intelligence is equally distributed, and resources are unequally distributed. The unequal distribution of resources has been another factor that slows the understanding of all life on planet Earth.

During my bioinformatics career, PacBio has managed to improve their read base quality score a million-fold to a billion-fold while doubling the read length. In addition, what has traditionally required super-computers and international efforts to de novo assemble human genomes can now be done with a powerful laptop in a matter of hours thanks to new algorithms that makes the NP-hard problem de novo assembly problem to a more local problem that take advantage of the read length and base accuracy of the CCS reads and thanks to increase in the processing power of each semiconductor chip. The ability to cluster and phase reads based on their hetSNP and long-range information provided by Hi-C reads. We might be at the inflection point where we will be able to observe a Cambrian explosion in the number of new species studied.

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

To have no stone unturned.

When the author whole-genome sequence analysis with Illumina reads, I cannot help but feel that I have not explored all that could be explored and that there might be something missing in the data that cannot be explored like the dark matter in the universe, which we know to exist, which we don’t have any idea of its content. PacBio CCS reads resolves this issue.

**\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.

I imagine a future where we will be able to telomere-to-telomere sequence haplotype phased genome of a cell at a penny per cell and de novo assemblies are not required to infer the genome of the cell. In addition, the base accuracy will be so accurate that we can believe that every base is always representative of the underlying sequence.

And where we will not be aligning reads to the reference genome for variant calling, but when we will be performing comparative genomics between the genome of a single cell and that of the reference genome to study cellular heterogeneity and the collective impact on phenotype, wirings of a single cell, fine-tune the genotype to phenotype relationship and have a systematic engineering approach to understanding life across all species.