**Chapter 5: Conclusion and Discussion**

**Conclusion**

In this PhD thesis, we challenge the preconception that PacBio CCS bases are inaccurate, and we claim that CCS bases are, in fact, sufficiently accurate for single molecule mutation detection.

To support this extraordinary claim, we accumulate extraordinary evidence to characterise the CCS sequencing process, identify sources of sequencing errors and empirically estimate the Q93 CCS base accuracy to between Q60 and Q90 depending on the substitution and the trinucleotide sequence context. CCS bases, hence, are a hundred thousand-fold to a million-fold more accurate than Illumina bases. In addition, we use samples with a single ongoing somatic mutational process to show that not only single molecule somatic mutation detection is possible, but also that the expected mutational pattern expected is directly observable from the called somatic mutations. Our approach is similar to how CHM1 and CHM13 cell-lines are used to assess heterozygous mutation calls can be used to assess and benchmark single molecule somatic mutation calls. DeepConsensus polished CCS reads, uncapped CCS BQ scores and CCS BQ score recalibration with partial order alignment between CCS and subreads from the same ZMW together indicate that pbccs assigns incorrect BQ score estimates, which is responsible for the false positive somatic mutation calls. We, here, have not explored whether library errors are a source of false positive substitutions, but we believe that CCS library preparation could be optimised to reduce library errors and further improve single molecule somatic mutation call sensitivity and specificity similar to how the Nanoseq protocol improves the duplex protocol to improve somatic mutation call sensitivity and specificity. Using our understanding, we develop and benchmark himut that enables single molecule somatic mutation calls with PacBio CCS reads and himut is available as a Python package under MIT open license at <https://github.com/sjin09/himut.git>.

We have discussed the advantages and disadvantages of PacBio SMRT sequencing platform. Before the introduction of circular consensus sequencing, PacBio optimised for read length instead of base accuracy and offered continuous long read sequencing with average read length between 5kb and 20kb and error rate of 10-15\%. CLR reads, hence, were limited to de novo assembly and germline structural variation detection. The advent of CCS reads, however, is a instrumental/monumental momenet in human genomics on multiple-levels. We never had a readout of genetic sequences at this accuracy at this scale with this level of base accuracy. CCS reads have an average read accuracy of Q20 and above, but CCS reads have base accuracy between 1 and 93 with a nominal error rate of 1 error per 5 billion bases. To date, there has not been an independent assessment of PacBio CCS base accuracy except for data described in this PhD thesis. We estimate the empirical error rate of Q93 CCS bases to be between Q60 and Q95 and the error rate is dependent on the substitution and the trinucleotide sequence context. In addition, PacBio has informed us that they use a dinucleotide sequence context hidden markov model for consensus sequence generation and base accuracy estimation, and the limited observation of sequence context might be responsible for the erroneous base accuracy estimation. Moreover, we were able to recover mutational pattern that was more consistent with the gold-standard mutational pattern from the sample when we recalibrated the base quality scores, providing further evidence that base quality scores are erroneously calculated for each base for each trinucleotide sequence context. It is unclear whether how the erroneous bases are introduced to the CCS reads and these erroneous bases must be introduced upstream of the sequencing process or be a result of systematic sequencing error, but a better consensus sequence algorithm will be able to address this problem in the future. We, furthermore, observed that somatic mutations called from shorter CCS reads have a higher number of false positive mutations than that called from longer CCS reads. Our hypothesis is that template with read-of-insert will have higher number of full passes and hence, more bases will be assigned Q93 base quality score, increasing the likelihood that erroenous library errors are assigned a high base quality score. In addition, we have observed in one of our sperm samples and in some of the DToL samples where Blue Pippin based size selection prior to CCS library preparation will introduce DNA damage to the template DNA such that C>T mutations are elevated in the overall mutation call. For a damage introduced upstream of CCS library preparation to have Q93, the DNA damage must be repaired such that the DNA base on both the forward and reverse strand is erroneously repaired. We hypothesied that \*\* might be responsible for this type of erroneous DNA damage repair. Hence, a combination of library errors and consensus sequencing errors are present currently in the CCS reads. Since himut relies on base quality score as one of the features of single molecule somatic mutation calling, the increase in the proportion of bases with Q93 bases leads to distortions in the number of absolute number of called mutations and decreases sensitivity.

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.

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.

Wright’s laws and Moore’s law should enable PacBio to achieve economies of scale at an exponential speed and the future that we dream of might be closer than we anticipate.

**Discussion**

*See things not as they are, but as they might be* [J. Robert Oppenheimer]

Library errors, sequencing errors are absent and where input DNA requirement is not a constraint towards sequencing.

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.

Full-length

Transcriptome and proteome per cell

With base modifications

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.

**SMRT sequencing: the last DNA sequencing platform**

“*Nothing is more powerful than an idea whose time has come*” [Victor Hugo]

Illumina platform was the sequencer of choice for most researchers and clinicians, and we were able to deliver the promise of genomics with continued decrease in compute, storage, and sequencing costs to greater and greater number of people. Illumina sequencing cost has decreased faster than Moore’s law from XXX to XXX, but the rate at which sequencing cost has decreased had slowed in recent years (Figure XX). In addition, the read length and base accuracy of Illumina hasn’t changed marginally, the only noticeable change/innovation has been in the throughput per lane. There is a limit to the knowledge that can be gained with marginal increase in number of genomes sequenced with Illumina sequencing platform. This is demonstrable from ~30% rare genetic disease diagnosis rate with Illumina platform and the need to develop new protocols to study single-cell genomic and transcriptomic heterogeneity. And without competition, Illumina has not reduced their sequencing costs to maintain their profit and operating margin [Figure X]. We can conclude that for new technologies and new approaches are required to have a better understanding and to advance human genomics.

Third-generation sequencing or single molecule sequencing from ONT and PacBio was a hard sell for most consumers. The throughput was lower, error rate higher and sequencing costs was higher, and the read length was not substantially better than that from Illumina either. In the last decade, however, the both ONT and PacBio have substantially increased throughput, decreasing per base sequencing cost, and improved upon the base accuracy and the longer read length (>10kb-100kb) have started to interest scientists to revisit the problem of de novo assembly algorithms, structural variation detection and construction of high-quality plant and animal genomes. In addition, PacBio started to optimise their library preparation to optimise for read base accuracy instead of read length by increasing DNA polymerase processivity and keeping the read length constant.

The author, here, believes that PacBio SMRT platform could be the last DNA sequencing platform. The PacBio SMRT platform has the potential to be the cheapest and the most accurate and scalable sequencing platform in the market and PacBio has demonstrated excellence in execution and delivered on their promises. PacBio long reads have improved in base accuracy rate from Q10 to Q90 in the last decade, improved throughput CLR throughput from XXX to XXX and CCS throughput from XXX to XXX with the introduction of Revio, which delivers whole-human genome at $1000, a competitive price considering that CCS reads can be used for de novo assembly, haplotype phasing, 5mC detection, somatic mutation detection and structural variation. (the versatile applications of CCS reads). Our research suggest that PacBio SMRT platform will be able to increase exponentially in the future as well with increase in the number of ZMWs per SMRTcell and increase in the read-of-insert-length. Our research also suggests that DNA polymerase processivity is no longer the bottleneck to obtaining Q90 bases and that CCS base quality score estimate is responsible for obtaining correct/incorrect BQ score estimates and hence, read-of-insert length can be further increased (Figure XX). The way in which the number of ZMWs per SMRTcell is increased is similar to how the number of transistors is increased per semiconductor chip and improvements in fabrications technologies from TSMC, ASML, Lam Research, Applied Materials have pushed the limits of what is possible. Furthermore, the acquisition of circulomics and optimization of CCS library preparation reduces the HMW DNA input requirements and in the future, we expect we can run SMRT sequencing from picograms of DNA. The trajectory of their improvement follows the improvements made on the Illumina platform (Figure XX).

The question is, hence, not whether PacBio SMRT platform is useful, but whether what will we do with reads produced from the PacBio SMRT platform.

The higher baser accuracy reduces the need to obtain higher sequence coverage to have the confidence with which the base is called.

In comparison to the traditional next-generation sequencing methods, CCS reads have longer read length, is free from PCR amplification and has higher base accuracy. Despite these limitations, PacBio CCS reads outperform on every metric from read length, base accuracy, number of applications from a single run compared to short reads from next-generation sequencing except for per base sequencing cost. This, however, is a limitation that PacBio as a company can overcome through a number of ways: i) the number of ZMWs per SMRTcell can be increased and ii) the average read-of-insert length can be incresed per template molecule. PacBio has increased the number of ZMWS per SMRTcell from XX ZMWs in XXXX to 8 million ZMWs to XXXX. In addition, the average read-of-insert length for CCS sequencing has increased from 10kb in 2019 to 20kb to 2021. Morever, if PacBio is further able to increase the processivitiy of DNA polymerase through further protein engineering or DNA polymerase evolution, they will be able to choose between longer average read-of-insert lnegth or increase in base accuracy through increases in the number of passes per template. I would assume that PacBio will choose to increase the read-of-insert length instead of base accuracy as base accuracy is certaintly sufficiently high at the moment for most practical purposes and higher than what is offered through NGS platforms. In addition, our research suggests that PacBio CCS base accuracy problem should be resolved not through increase in the number of passes per read, but through better design of their conensus sequence algorithm. Recently, Google released deepConsensus algorithm to polish CCS reads based on alignment of subreads from the same ZMW to the CCS reads and to recalibrate the base quality scores. Deepconsensus, currently, cannot be applied towards all the CCS reads produced from SMRTcell and instead must be applied a subset of CCS reads for an average user. In addition, deepConsensus fails to estimate the base accuracy of the reads properly and the base accuracy estimates are too pessimistic, ranging from Q1 to Q50, which is below our empirical estimate between Q60 and Q90 for Q93 bases. In addition, if somatic mutations are called from CCS reads with polished with deepConsensus using Q50 bases, we are not able to obtain a mutational pattern that is expected from the sample.

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.

*People don't have ideas. Ideas have people.* [Carl Jung]

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.

%% history of sequencing

%% sanger sequencing

%% cheapter Illumina sequencing

%% solexa

%% rolling circle amplification based approaches

%% Oxford Nanopore Technologies (legal dispute between ONT and PacBio)

%% the advent of PacBio CCS sequencing

%% the limitations of PacBio CCS sequencing

%% the advantages of PacBio CCS sequencing

%% alternative methods towards single molecule somatic mutation

%% PacBio CCS reads are more accurate than duplex reads, but less accurate than nanorate reads

%% PacBio CCS sequencing could increase throughput and thereby lower the cost per base sequencing by increasing the read-of-insert length and increasing the number of ZMWs per SMRTcell.

%% PacBio have approached this in the past and there is no reason why it should not happen

%% Similar to transisoter trechnology per CPU chip

%% whole-genome CCS sequenncing allows users to perform de novo assembly, 5mC detection, somatic mutation detection, germline mutation detection from a single run, providing the most comprehensive set of both genetic and epigenetic information to scientsits

%% to obtain similar set of depth and breadth of information using Illumina sequencing would cost more and provide data that has less resolution.

%% in addition, obtaining some information requires arduous laboratory procedures or modified library protocols to increase the base quality scores

%% the deceleration in the cost of sequencing from Illumina as Illumina dominated the sequencing market

%% challengers: ONT, PACB, BGI, another company ... cant' remember

%% the longer reads from third-generation sequencing platforms allows, despite the average lower base accuracy, confident placement of reads relative to the reference genome

%% infinity from illumina is based on clever library prep, but it is not a true single molecule sequencing

%% higher base accuracy allows lower sequence coverage to call germline mutations as less evidence is required to have equal confidence in calling germline mutations

%% similarily, if the base accuracy is sufficiently high that sequencing errors can be distinguished from somatic mutations, a mismatch between a single read and the reference genome is a true mutation instead of a sequencing error unless that mismatch was created through DNA damage during library preparation or incorrect repair of DNA damage during library preparation.