**Thesis Objectives**

The history of science is riddled with examples where theory, technology, and serendipitous discovery drives science. The advent of Illumina short reads and continued decrease in per base sequencing cost has accelerated our understanding of human evolution and migration patterns \cite{}, helped identify pathogenic mutations in Mendelian diseases \cite{}, enabled the analysis of driver mutation and transcriptome landscape in thousands of cancer genomes \cite{}.

The inability to generate contiguous and complete reference genomes, however, with Illumina short reads and the prohibitively expensive cost of BAC clone library preparation and hierarchical shotgun sequencing has thwarted our efforts to understand genetic variation in non-model organisms \cite{}. The advent of high-throughput long-read sequencing and genome mapping technologies \cite{}, improvements in base accuracy of long reads \cite{} and development of algorithms that take advantage of the longer read length and long-range genomic interactions \cite{} has brought new enthusiasm to sequence and assemble high-quality reference genomes for all of life \cite {}.

*It's not an experiment if you know it's going to work*. [Jeff Bezos]

The DToL project has generated an extraordinary public resource that comprises CCS reads, linked reads, Hi-C reads, high-quality chromosome-length scaffolds, and associated gene annotations. Comparative genomics in linear and three-dimensional space and population genetic studies with the newly assembled reference genomes will undoubtedly enhance our understanding of the process of speciation and evolution. Here, we instead aspired to better understand the mutational process operational in each species and in each kingdom.

To determine the germline and somatic mutational process across the Tree of Life, we considered the following:

\begin{enumerate}}

\item Based on the similarities between the duplex sequencing \cite{} and CCS library preparation protocol \cite{}, we hypothesized that CCS reads might have sufficient base accuracy for ultra-rare somatic mutation and potentially single molecule somatic mutation detection.

\item CCS reads are reported to have a predicted accuracy above Q20, but their base accuracies have not been independently examined.

\item Somatic mutation detection algorithm needs to distinguish somatic mutations from sequencing, alignment and systematic bioinformatic errors. We, unfortunately, cannot differentiate somatic mutations from library errors unless there are upstream modifications to the library preparation protocol.

\item Using samples with single ongoing somatic mutational process and mutational signature analysis, we can demonstrate that CCS reads have sufficient or insufficient base accuracy for single molecule somatic mutation detection and determine the parameters that influence sensitivity and specificity.

\item If the sample in question has either high mutation rate or high mutation burden, the expected and the correct mutational spectrum will be observable from the validation and test data sets, respectively.

\end{enumerate}

In short, we aimed to measure the CCS error rate, assess whether CCS bases have sufficient base accuracy for single molecule somatic mutation detection, develop a method to detect somatic mutations where a single read alignment supports the mismatch between the sample and the reference genome and apply the method to understand germline and somatic mutational processes across the Tree of Life.

%% \subsection{Somatic mutations}

%% \subsection{Mutational signatures and mutational processes}

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

%% enodgenous and exogenous somatic mutation

%% DNA damage, repair, fixation

%% envrionment

%% DNA polymerase infidelity, germline mutations

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

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

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

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

%% first protein

%% first riboenzyme?

%% first unicellular organism %% first in the sea

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

%% fusions, meiotic recombination, plant recombination?

%% non-hologous end joining

%% transcripion-coupled repair

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

%% deleterious, postivie,

%% linked by DNA

%% entropy to submission

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

%% Complexity that

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

%% first multicellular-organism

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

%% different types of DNA polymerases, redundancies

%% Oswald Avery: amino acids, greater number of combinations, genetic sequence as the transforming substance

%% Rosalind, Watson: Structure of DNA

%% what happened from the discovery of the structure of the DNA to the human genome project?

%% in humans

%% c-elegans? other species?

%% The Tree of Life is connected through genetic sequence

%% DNA is the puzzle that links us all

%% since inception, birth, somatic mutations starts to accumualte

%% fertilsiation for most organisms

%% cellcular division for unicellular organisms

%% depending on the timing and the type of tissue in which the somatic mutations occur somatic mutations are inherited to the daughter cells or the next generation

%% some mutations result in speciation

%% some mutations lead to survival of fittest

%% some mutations have a large consequence, recombination, structural variations

%% the study of mutations across the Tree of Life has been limited by the cost of reference genome construction and the availiabilty of reference genomes for population genetics and for comparative genomics.

%% the cost of reference genome construction has been prohibitively high

%% the human geome project, for example, cost 3 billion dollars, a dollar per base.

%% international collaboration, multiple sequencing centers with thousands of people

%% multiple-years

%% physical-maps %% fish %% restriction-enzyme based

%% YACs

%% fosmid 50kb-200kb

%% bacterial artificial chromosome clone 100kb fragments

%% gaps, missing sequences, acrocentric chromosomes, large sections of chromosome Y

%% unplaced, unlocalised chromsomes and contigs

%% placement of contigs, scaffolding of contigs

%% Sanger di-deoxy sequencing, limited to 500bp to 1000bp

%% Solexa and Illumina sequencing by synthesis

%% de brujin graph based assemblies are short, fragmented and incomplete

%% high-throughput, relatively high accuracy of short-reads

%% de novo assembly quality is a function of read depth, base accuracy, read length and complexity/repetitiveness of the target genome, %% Eric Lander

%% assemblies/genomes are abundant with sequences that are longer than Illumina read length: SINE, LINEs, repeat expansions, segmental duplications

%% longer read length is required to trasverse the repetitive sequence and uniquely locate/place the read amongst other reads, reads are collapsed into contigs in the face of high repetitive sequences

%% scaffolding technologies: mate-pair sequencing with longer-read inserts insufficient and not scalable

%% assembly and comparative genomics didn't improve in the last decade

%% cost was high, and the effort did not yield sufficiently meaningful assembly results

%% initially Single-molecule sequencing from Oxford Nanopore Technologies and Pacific Biosciences were also inaccurate and the read length were not magnitude of orders longer, low throughput

%% continuous long read sequencing from Pacific Biosciences, 10-15kb in read length with 10-15% error rate, the errors were thought to be random, free of amplification bias

%% sufficiently long enough to trasverse repeats, however not sufficient to distinguish between unique copies of segmental duplications

%% used to reconstruct missing sequences in the human reference genome %% eichler

%% updates in the human reference genome %% tina

%% CHM1 and CHM13 seuqencing to identify structural variations

%% pathogenic mutations/repeat expansions

%% ONT for chrY centromere sequencing

%% alpha-satelitte expansion

%% usefulness of haploid genomes

%% T2T consortium, for example, recently, completed the end-to-end assembly of CHM13 genome

%% high-throughput chromatin conformation capture sequencing (Hi-C), similar to mate-pair sequencing in concept, but across the whole-genome

%% 3C job-dekker, loops, configurations

%% originally used to study the three-dimensional genome configuration

%% chromosomes self-aggregate

%% end of one chromosome is in more contact with the end of the same chromosome than another chromosome

%% what about contacts between paternal and maternal haplotype of the same chromosome?

%% sequences in close proximity are in contact with each other more

%% conctact matrix can be used to discern correct assemblies from misassemblies

%% order and orient contigs %% matrix inversion, %% techniques from linear algebra

%% manually curate scaffolding and correct assemblies

%% studying the genomes from the Tree of Life provides snapshots of environments that the genomes were under through space and time

%% events that might have spurred natural selection, speciation and radiation

%% timed the emergence of species, but never timed the emergence of unique somatic mutational processes across time and space

%% assembly: assumption: haploid genome

%% Pacific Biosciences circular consensus sequencing, increase in the number of ZMWs per SMRTcell from 1 million to 8 million, circular consensus sequencing instead of continuous long read sequencing

%% increase in DNA polymerase processivity, continuous long-read sequencing perhaps once or twice per molecule, circular consensu sequencing: 8 to 16 times per molecule

%% because the errors are thought to be random, highly accurate circular consensus sequence generation is possible

%% estimated to have accuracy between Q20 and Q30

%% assemblies produced from PacBio CCS reads have accuracy between Q50 and Q60.

%% massive incerase in the contiguity and completeness and assembly of the genome

%% time to complete the genome

%% cost to complete the genome

%% thousands of scientists to handful of scientists

%% except for the most complex genome

%% significant upgrade in the quality of the genome compared to that produced from short reads

%% also comparable to that produced through the human genome project

%% or small organisms or unicellular organims with limited DNA %% low-input protocol makes this possible albeit with errors introduced during PCR amplification %% bias towards sampling of reads or amplification of dna molecules before library preparation

%% the number of eukaryotic species sequenced and assemblies with PacBio sequencing increased dramatically since the introduction of long-read sequencing

%% uncovering the evolutionary history of these species

%% Methods to study somatic mutations in cancer

%% the reasons to study cancer

%% somatic mtuational processes in cancer

%% mutational patterns, mutational signatures

%% tumour and matched normal

%% technical limitations of short reads

%% sub-cloncal

%% minute fraction

%% Methods to study somatic mutations in normal tissues

%% single-cell PCR amplification and sequencing

%% single-cell clone expansion and sequencing

%% duplex sequencing, nanorate sequencing

%% laser-capture and microdissection and sequencing of clonal tissues

%% driver mutations

%% drug resistance development

%% evolutionary history of cancers

%% developmental biology

%% lineage-tracing

%% Wellcome Trust Sanger Institute has initiated the Darwin Tree of Life project to sequence approximately ~66,000 eukaryotic species in the and the primary mode of sequencing is CCS sequencing, hi-c sequencing

%% to sequence and assemble the samples with CCS sequencing, scaffold the samples with Hi-C reads and to curate the scaffolded assemblies through manual inspection of the contact matrix

%% We and others have hypothesized the potential for CCS sequencing for somatic mutation detection

%% Nanorate sequencing, blunt-end restriction enzyme digestion, DNA nicks, dideoxy nucleic acid, DNA damage during sonication %% to preserve the native DNA molecule and to sequence the DNA molecule

%% We noticed the high simliarity between duplex sequencing and CCS sequencing and hypothesized that CCS sequencing might have sufficient base accuracy for single molecule somatic mutation detection, if we can distinguish highly accurate bases from that resulting from library errors, alignment errors and sequencing errors and systematic errors. artefacts that cannot be removed

%% Other mammalian species with shorter life span have higher somatic mutation rate such that at the terminal stages of life, the species in question have same mutation burden at the time of death

%% Peto's paradox

%% resequencing studies have enabled the identification of germline mutational process, somatic mutational process in humans

%% the study of other species have been limited to date

%% c-elegans? %% what are other species?

%% if our hypothesis is true, we conjectured that we will able to detect somatic mutations across the Tree of Life, reveal somatic mutational processes active in the species, time the emergence of somatic mutational processes and attribute the contribution of somatic mutational processes to the germline mutational process, %% environmental mutagenesis

%% in Chatper 2, we demonstrate that PacBio CCS base accuracy is sufficiently accurate to call and study single molecule somatic single-base-substitution across species

%% sequence samples with a single dominant somatic mutational process

%% know the mutational signature or have gold-standard mutational signature for the sample generated from single-cell clone expansion and sequencing

%%

%% somatic mutation detection from a single read alignment to the reference genome

%% if we were to call every mismatch between the read and reference genome, we will be able to call all somatic mutations at the cost of high false positive rate

%% Oxidative DNA damage

%% typically requires a normal sample to distinguish between germline and somatic mutations

%% typically requires multiple reads to suppport the somatic single base substitution

%% VCF file produced from somatic mutation callers are the sum of library errors, systematic errors, sequencing errors, alignment errors, %% reference bias?

%% unresolved errors

%% if we are able call somatic mutations from a single read alignment to the reference genome, we are not only able to reduce the cost of sequencing, but also do germline mutation calling from reduced read depth

%% 30X sequence coverage required to call heterozygous mutations %% reference

%% problems with PacBio CCS sequencing: incomplete removal of adapter sequences, chimeric sequences resulting from problems with adapater sequence calling, fragmer and concatmer

%% reads significantly shorter and longer than the read-of-insert length

%% empirically estimate the PacBio CCS base accuracy

%% PacBio CCS base accuracy has not been measured yet, PacBio CCS base also cannot be measured with exisiting sequencing technologies with lower base accuracy

%% in Chapter 3, confirm that our method is applicable to other eukaryotic species, we use the newly developed method to study somatic mutational processes across the ~400 eukaryotic species sequenced the Darwin Tree of Life project, attempt to understand both the germline and somatic mutational processes across species, identify potential sources of environemtnal mutagenesis

%% phorcus lineatus: age

%% insects: life cycle of insects (choleoptera)

%% mutation burden of insects with metamorphosis and without metamorphosis

%% germline mutational process

%% somatic mutational process

%% environmental mutagenesis

There are two types of third-generation sequencing: one from Pacific Biosciences and one from Oxford Nanopore Technologies and both companies attempt to sequence single molecule of DNA, in contrast to the sequencing by synthesis approach. These approaches, previsouly, had an error rate ranging from 20\% to 40\% depending on the library chemistry and the base caller version. PacBio introduced circular consensus sequencing in 2010, but circular consensus sequencing could not be adopted for mass-adoption as DNA polymerase for SMRT sequencing didn't have sufficient processivitiy to read long read-of-insert multiple times. Instead, PacBio offered continuous long read (CLR) sequencing to its customers which maximized for read length instead of average read accuracy. CLR reads typically have 10-15\% error rate, but is free from PCR amplification, the errors are thought to be randomly introduced and CLR reads have read length that is 100-fold longer than that from short reads. CLR reads, hence, was adopted for de novo assembly of complex genomes that could not be assembled with short reads and for structural variation detection. The longer read length enables the read alignment software to confidently assign the location of the reads relative to the reference genome as unique sequences are flanking repetitive sequences. Germline structural variation detection with long reads doubles the average number of structural variations discovered per genome compared to that from short reads and improves the diagnostic yield of rare genetic disease detection from 30\% to 80\%. The lower base accuracy and cost of SMRT sequencing, however, limited the wider adoption of PacBio SMRT sequencing except for one-off de novo assembly projects and clinical sequencing of patients with rare genetic diseases. PacBio, however, successfully engineered DNA polymerases with increased processivitiy and was further able to improve their circular consensus sequencing method such that read-of-insert with average read length of 10kb to 20kb can be read multiple times and because the errors are introduced randomly to each single-strand sequence templates, consensus sequence algorithms can take advantage of the complementary nature of double-stranded DNA to produce circular consensus sequences with average read accuracy greater than Q20.

%% in Chapter 4 and 5, we use the unique combination of long read length and base accuracy of PacBio CCS reads to study both meiotic and mitotic recombniation, respectively.

%% in Chapter 2 and Chapter 3, we demonstrate that PacBio CCS reads have sufficient read length and base accuracy for single molecule somatic single-base substitution agnostic of clonality and species.

%% to explore the unexplored phenomena of meiotic recombniation through Sperm PacBio CCS sequencing

%% diffences to previous attempts to understand meiotic recombination through trio sequencing and sperm-typing

%% gene conversions requires the detection of chimeric dna molecules with both maternal and paternal sequences

%% meiotic event generates 2 recombinant products and 2 wild type molecules

%% crossover leads to the generation of dna molecule with a stretch of paternal hetsnps followed by a stretch of maternal hetsnps and vice versa

%% gene conversion leads to the generation of a dna molecule where paternal hetsnps is flanked by maternal hetsnps (and vice versa)

%% complex recombinant product with resulting from both crossover and gene conversion

%% on average, there 1 SNP per 1000bp

%% requires long-range PCR products to detect

%% hotspots

%% coldspots

%% meiotic recombination product requires reads that can span multiple hetsnps and requires sufficient base accuracy to determine that hetsnp switch is a result of a biological event rather than a sequencing error.

%% in addition, meiotic recombination can be a source of mutagenic event

%% PacBio CCS reads have sufficient base accuracy to detect single molecule recombination events and associated mutations

%% recombniationi might not be a perfect/clean

%% mutational process that generates de novo single-base substitution seems to be driven by clock-like somatic mutational processes (SBS1 and SBS5)

%% mitotic gene conversion can be a source of oncogenic mechanism in somatic cells

%% simliar to meiotic recombination, products from mitotic recombination cannot be detected with short reads due to the technical limiations of the Illumina platform

%% sequenced Bloom syndrome patient samples with defects in DNA double-strand break damage repair process

%% known to have gene conersions or loss of heterozygous caused by gene conversions

%% perfect sample to assess the differences in mitotic and meiotic recombniation and gene conversions

%% mitotic gene conversions are thought to be longer in length

%% in Chapter 6,

%% the benefits of PacBio CCS sequencing

%% the last DNA sequencing platform

%% requires significantly less sequencing coverage than short reads to detect the same number of mutations

%% can detect small SNPs, indels, structural variations, 5mC from the same platform

%% with the development of himut, CCS reads can be also used to detect somatic mutations, gene conversion and crossovers from the same sample.

%% potentially other base modifications caused by environmental exposure, chemotherapeutics in the future

%% Moore's law: the number of transisitors per semiconductor has doubled, the distance at which the electrons has to be moved has shorteneed

%% the cost of sequencing per base was decreasing at a faster speed than Moore's law and many has anticipated that we might have a $100 genome, if the development had continued

%% stagnation in development, and Illumina monopoly status, financialisation, stock buybacks instead of research and develompent

%% increase in the number of ZMWs per SMRTcell, PacBio has achieved 8-fold improvement in throughput

%% increase in the read-of-insert length, doubling, stabiltiy of the circular template molecule

%% direct-engineering, directed-natural selection

%% increase in DNA polymerase processivity can increase either the read-of-insert length or the base accuracy of the same read-of-insert length

%% improvement in HMW DNA extraction, from the smallest organism

%% past Illumina platform generation has also required high DNA concentration

%% improvements in circular consensus sequence calling process can lead to the better discernment of adapter sequences from

%% PacBio CCS sequencing offers an alternative method for DNA sequencing with potential to improve throughput and base accuracy at a faster rate than that from Illumina unless Illumina profit margin compresses

%% PacBio CCS sequencing will be cheaper, more accurate, have higher throughput than Illumina sequencing

%% Illumina might compete in terms of price, but the wealth of information that is delivered from PacBio will be immense %% adoption curve

%% the cumulative improvement will us to better understand all of life

%% \subsection{The

PacBio introduced circular consensus sequencing in 2010, but circular consensus sequencing could not be adopted for mass-adoption as DNA polymerase for SMRT sequencing didn't have sufficient processivitiy to read long read-of-insert multiple times. Instead, PacBio offered continuous long read (CLR) sequencing to its customers which maximized for read length instead of average read accuracy. CLR reads typically have 10-15\% error rate, but is free from PCR amplification, the errors are thought to be randomly introduced and CLR reads have read length that is 100-fold longer than that from short reads. CLR reads, hence, was adopted for de novo assembly of complex genomes that could not be assembled with short reads and for structural variation detection.

The longer read length enables the read alignment software to confidently assign the location of the reads relative to the reference genome as unique sequences are flanking repetitive sequences.

Germline structural variation detection with long reads doubles the average number of structural variations discovered per genome compared to that from short reads and improves the diagnostic yield of rare genetic disease detection from 30\% to 80\%. The lower base accuracy and cost of SMRT sequencing, however, limited the wider adoption of PacBio SMRT sequencing except for one-off de novo assembly projects and clinical sequencing of patients with rare genetic diseases. PacBio, however, successfully engineered DNA polymerases with increased processivitiy and was further able to improve their circular consensus sequencing method such that read-of-insert with average read length of 10kb to 20kb can be read multiple times and because the errors are introduced randomly to each single-strand sequence templates, consensus sequence algorithms can take advantage of the complementary nature of double-stranded DNA to produce circular consensus sequences with average read accuracy greater than Q20.