**Chapter 1**

"Let there be light", Genesis 1:3

%% According to Aristotle's theory of heredity, males are responsible for providing the "active element" that gives life to the offspring and makes sure that it is a male or female, and females provide the nutrients to the offspring.

%% <https://thelampstand.com.au/the-amazing-living-cell-a-model-for-christs-ecclesia-wrong-part-2/>

We can study physics and chemistry anywhere in the universe, but we can only study biology on planet Earth.

description of the nature that changes

earth was the centre of the universe

nature does not change

In science, attempting to describe the natural world

Steve jobs pancreatic cancer sequencing: how many dollars

if physics studies atoms

the monomeric unit for biology is the cell

for the chemist, element in the periodic table is the fundamental unit of interest

physical, chemical, biological constraint that

Is dictated bt the genome

the complete sequencing of mtDNA

purgedups

polyploid genomes

no stones left unturned

where de novo assembly becomes the relifct of the past

and full-length genome sequencing becomes the norm

intelligence is equally distributed

but resources are not

isoforms

gene conversion and emthyaltion

thigs don’t exisnt in isolation

sexual selection

angle of attack

absence of reference bias

software standardisation and best practices

repeat expansions and small me,lcule therapy

shaping the genome

and how differeent species have built in error tolerance mechanisms

robust

replicates

reproduces

speciates

precision and accuracy

Higgs boson

growth

Division

sjin.lee09@icloud.com:

see more of reality

**Bible, cell structure, nucleus**

**The Origins of Life**

%% Error tolerance

%% replication: RNA, RNA as a genetic information and as a catalyst, Ribozyme (ref)

%% metabolism: dissociation, association, assimilation (hydrolysis), polymerisation, monomers, polymers

%% gene, the unit of inheritance

%% the first gametes, the first sex cell, sex

%% asexual reproduction

%% sexual reproduction

%% horizontal gene transfer

%% the transition from sea to earth/land

%% unicellular organism to multicellular organism

%% Cambrian explosion: separation of embryo and adult structure

%% vertebrate, %% invertebrate

%% symbiosis between eukaryotic cell and mitochondria

%% ATP as the unit of energy [ATP was discovered in 1929 by Karl Lohmann]

%% DNA as the genetic apparatus

%% The last universal common ancestor (LUCA)

%% Central Dogma

%% messenger RNA, transfer RNA

%% the first fertilisation

%% each transition, new adaptation, the ability to fly,

%% natural selection

%% DNA as a decentralised record-keeper, bookkeeper, recording errors and innovations in real time

%% Genetic drift

%% thermodynamic decay

%% The word cell from the latin world cellula, which means room.

Contrary to the origin of life as laid out in the bible, the first life probably originated under the sea near thermal vents and Garden of Eden must have been under the sea [ref] and if Garden of Eden existed, it must have been under the sea. There are multiple theories to how the first cell came to be: Epigen theory, Oparin theory, and XXX theory. The author of this PhD thesis subscribes to the view that the first cell would have required multiple innovations to arrive the last universal common ancestor (LUCA) and subscribes to the evidence that potentially suggests that capacity to perform replication and metabolism didn’t’ happen at once, but that the ability to metabolise came first and that the ability to replicate came second.

RNA is conceived to the be the first carrier of genetic information and the existence of ribozymes, RNA molecules capable of catalysing chemical reactions, suggests that RNA molecules could have self-replicated. RNA, however, could not be a stable source of genetic information as the presence of additional oxygen group compared to DNA made RNA more susceptible towards chemical degradation. The derivation of LUCA would have required multiple steps. First, the formation of semi-permeable phospholipid bilayer would have been required to separate the intercellular and extracellular environment, the first instance of physiochemical? separation of order and disorder. Amino acids are easier to produce under the conditions of the early stages of earth (no oxygen, etc, etc). Thereafter, random Brownian motion would have allowed for polymerisation of monomers to an active catalytic component capable of catalysing chemical reactions [ref] and the cells that have the most efficient catalytic machineries would have grown in volume and the pressure on the plasma membrane would have naturally divided the cells such that the cells with the most productivity catalytic machinery would have dominated the environment.

This first cell would, at some point, would have merged and evolved together with another cell that had RNA as genetic information and would have enabled the production of nucleotides and polymerisation of nucleotides to nucleic acids and genetic replication. In the history of eukaryotic cells, cells aggregated to form multicellular organisms with tissues with specific functions. Oxygenation of the planet happened \*\* how many years ago. The first eukaryotic cell is estimated to have been formed 2 billion years ago. The eukaryotic cell is characterised by the membrane-bound organelles and the presence of mitochondria and/or chloroplast that functions as a source of energy and producer of energy, respectively. There have been multiple milestones, thereafter, from the transition of organisms from sea to land,

\*\*What was the impact of oxygenation of Earth?\*

\*\*Change in diet\*

\*\*Change in

Since the start of time, entropy has been increasing following the second law of thermodynamics and biological systems have emerged to reduce or maintain entropy using energy. Phospholipid permeable membrane was the first spontaneous invention that separated order from disorder and allowed for the movement of molecules between the extracellular and intercellular environment and for the emergence of primordial cell. It is uncertain whether the first cell had both the capacity to replicate itself or whether had the capacity to catalyse chemical reaction first. In a prebiotic environment, amino acids can be created in a reducing environment if sufficient energy in the form of ionizing radiation, ultra-violet light, is introduced into a gaseous atmosphere containing methane, ..., ... and ... [ref] and nucleotide bases are thought to be harder to spontaneously create in a prebiotic environment [ref]. Despite the uncertainty in how the first cell arose, the first prokaryotic organism is thought to have arose ~XX billion years ago and the first eukaryotic organism is thought to have arisen approximately 2 billion years ago [ref]. Once the first cell was created, selection pressure and natural selection acted upon these cells to create the first multicellular organism and these multicellular organisms evolved to create multiple different species that is best adapted to the environment surrounding them. Mutations play a central role in creating new innovations that allows for individual species to better adapt to the environment and to produce progenitors that inherit the mutations.

%% What is Life by Schrodinger

%%

Structure of the DNA double helix

It is now widely accepted truth that DNA is the unit of inheritance, and that DNA has a double-helix structure and that the structure of the DNA drives many of the important chemical reactions in the cells such DNA replication and transcription. In addition, sequencing technologies has become cheap enough such that clinical sequencing is routine enough to be able to detect the mutations that is responsible for disease and to understand the mutations that confer selective growth advantage to cancer genomes and amazingly, the cost of sequencing is still decreasing, and new sequencing technologies are emerging to differentiate itself from short reads produced from next-generation sequencing platform. These widely accepted truth, however, were only enabled by giants who reimagined what was possible and who were willing to against the norm.

We must have wondered about the physical material that is responsible for the unit of inheritance from ancient times [ref]. [Greeks, Romans, Bible], Gregor Mendel is thought to be the father of modern genetics and provided the theoretical framework for the study of genetics with his famous experiment where the studied he inheritance of Peas's traits to their descendent in 1866X. Mendel carefully cross-breaded peas with different traits to discover that traits were inherited with a fixed ratio, also known as Mendelian ratio, and how certain traits are governed by dominant and recessive alleles. His experiment revealed how the physical material that is responsible for unit of inheritance must be separated into gametes and randomly united during fertilisation to determine the phenotype of the progenies and that the factors responsible for the phenotypic differences must be located independent of each other. These two rules are referred principle of segregation and principle of independent assortment.

Amino acids were initially proposed as the physical material responsible for inheritance as the number of amino acids and different varieties of proteins that could be created from different combinations of amino acids could could potentialyl explain the complexity of a living organism and DNA was thought to be too simple to be able to encode the complexity of a living organism. It was not until the Oswald Avery's experiment in XXXX that demonstrated the DNA to be the physical material responsible for the transformation of R-strain bacteria to S-strain bacteria and despite the evidence, DNA was not believed to be physical material for unit of inheritance. The next race started with the aim of discovering the structure of the DNA and there were many potential protagonists who could have discovered the structure of the DNA, but James Watson and Francis Crick, then post-doctoral fellow and PhD student at the laboratory of molecular biology, respectively, were the first to the race in 1954 [ref]. Despite the initial scepticism of how DNA could be the unit of inheritance and how DNA could be responsible for the complexity of an organism, the mechanisms of the central dogma was slowly revealed. Series of discoveries following the discovery of the structure of the DNA has cemented the importance of DNA as the central unit responsible for directing cellular behaviours and determining phenotypes and encoding the software to produce proteins, the hardware that is responsible for catalysing chemical reactions within the cell. Despite their simplicity, methods for DNA sequencing was designed later than that for amino acid sequencing. Frederick Sanger and Walter Gilbert came up with Sanger dideoxy sequencing and Maxam-Gilbert sequencing, respectively, to determine the nucleotide monomer that constitutes the given nucleic acid. Sanger was able to determine the genetic sequence of XXXX and XXXX using Sanger dideoxy sequencing for the first time. The Sanger dideoxy sequencing was more amenable to sequencing at scale and was adopted for the Human Genome Project (HGP) as the primary sequencing instrument and Sanger reads produced from ABI had an average read length of 500bp to 1000bp and had an average base accuracy between Q20 and Q50.

%% Hunt Morgan, Drosophila, Mutations

%% Linkage disequilibrium

%% Craig Venter,

%% 20th century

%% 21st century

%% minimal viable cell

%% Herbert Boyer, cloning, 1st generation

%% genetic and genomic engineering, 2nd generation, CRISPR-CAS, prime editing, base editing

%%

Lens\*

Philosophers have questioned the origin of life for eons and as precision and accuracy of the tools that we use to study the world around us increases, we have improved our description of Nature.

Telescope to study the heavens and the microscope to study cells.

**The human genome project (HGP)**

%% physical mapping

%% FISH

%% BAC clones

%% YACs

%% sanger dideoxy sequencing

%% international collaboration

%% alternative loci

%% gaps, missing sequences, acrocentric chromosomes, chromosome Y, unplaced and unlocalised contigs,

%% mate-pair sequencing with longer read inserts

%% hi-c sequencing

%% alternative locis (genes)

%% expansions

%% collapse

%% small, and large

%% alpha-satellites centromeric regions were modelled

%% manually curate

The Human Genome Project was initiated to sequence and assemble the human reference genome that would standardise the genetics and genomics studies to a single reference genome. There were two approaches towards the human reference genome construction: one was the hierarchical shotgun sequencing and assembly strategy and the other was whole-genome shotgun sequencing and assembly approach. The human reference genome constructed from the former approach is still the human reference genome used in most genetic and genomics studies and is the bedrock of genomic medicine revolution [ref]. The availability of the human reference genome together with sequencing-by-synthesis approach from Solexa, now Illumina, revolutionised the field of human genetics and enabled population-scale studies of genetic diseases and cancers [ref]. Population-structure, human history, discovery of somatic mutations that confer selective growth advantage to the tumour cell, the identification of mutations that leads to genetic diseases. In addition, scientists have developed clever ways to modify library protocol upstream of Illumina adapter ligation to enable the study of epigenomes, base modifications, transcriptome of bulk tissue and more, recently, the advent of high-throughput chromatin conformation capture sequencing has enabled the study of the three-dimensional configuration of the genome and the nucleotide sequences are organised into regular repeating patterns.

%% single cell %% why was hierarchical shotgun strategy more accurate?

**Illumina sequencing and past competitors**

%% Ion-torrent

%% 454

%% sequencing cost, moore’s law

%% economics of scale

%% innovation on top of next-generation sequencing

%% ancient genomics

%% iteration of illumine sequencing machines: GAII

%% hi-c sequencing: loops configurations, topologically associating domain (TADs), A/B compartments

%% chip sequencing

%% 3C sequencing (Job Dekker)

The technical limitations of Illumina sequencing (base accuracy and short read length), however, has been the bottleneck for improving rare genetic disease diagnostics yield, detecting rare somatic mutations and constructing high-quality reference genomes for non-human species. De novo assembly of other species, previously, have been attempted using de brugjin graph based de novo assembly algorithms with short reads, but assemblies produced from short reads were highly fragmented and incomplete. In addition, scaffolding strategies often did not provide sufficient long-range information to produce chromosome-level pseudomolecules and as a result, these assemblies provided incomplete information for comparative genomics purposes. Hence, assemblies produced from short reads often have collapsed repeats or contigs that cannot be placed accurately. To construct complete assemblies, reads need to be longer than the repeats of the target genome such that the reads can traverse the repetitive regions and optimally have unique sequences flanking the repetitive sequences such that the read can be placed in the assembly graph unambiguously. Not all repetitive sequences are repetitive. There are unique class of repeats called segmental duplications, which doesn't have a classical repetitive sequence, has a unique sequence, but is duplicated across the many parts of the genome and are thought to be important in driving evolution and these segmental duplications are typically defined as sections greater than 1kb with sequence similarity above 90\% to other regions of the genome. To distinguish segmental duplications from one another, reads also need to have high base accuracy to be able to distinguish closest segmental duplications from one another. Long-reads from third-generation sequencing technologies such as Oxford Nanopore Technologies (ONT) and Pacific Biosciences (PacBio) provide an alternative towards improving the rare genetic diagnostics yield and improving the reference genome qualities in terms of both completeness and contiguity. Long-reads produced from third-generation sequencing platforms were orders-of-magnitude longer than that from the Illumina platform but had a much higher error rate; 10-15\% error rate for continuous long reads (CLR) from PacBio and 20-35\% error rate for ONT reads. Because of these high error rates, higher sequencing costs (lower yield per dollar) and insufficient improvement in read length, these platforms had limited use except for rare cases for real-time monitoring of ... and de novo assembly of plants and animal genomes..., and detection of pathogenic mutations that could not be detected with short reads [ref, ref]. Despite high error rate, the longer read length enabled the detection of structural variations that could not be previously detected with short reads, doubled the number of structural variations that can be detected from a typical human genome compared to the human reference genome. The longer read length allowed for the de novo assembly of BAC clones to hierarchically assemble missing sequences, also known as gaps, in the human reference genome, which have been problematic to assemble before and reveal human-specific gene duplications.

%% population-genetics

%% population genomics

%% increase in the number of whole-genome sequenced with illumine sequencing

%% cancer genomics, driver mutation, mutational signatures

%% clinical sequencing

%% tumour evolution

%% liquid biopsy

The human genome project is estimated to have cost 3 billion dollars, equivalent to 1 dollar per base pair. And as technology becomes more ubiquitous and democratised, we have constantly shifted/moved from studying one individual to studying the group. We initially focused on studying a single individual and as sequencing cost has decreased, population genomic studies and the history of differences and going back in time to study our lineage. What is common and different.

Thanks to new technologies.

A wave of standardisation to create file formats that is universally accepted across the community.

**Third generation sequencing**

**Nanopore sequencing**

%% UCSC

%% ONT

%% blocked

%% limit

**PacBio SMRT sequencing**

%% Cornell

%% P0, P1, P2

%% Zero-mode wave guide

**Continuous long read sequencing**

These companies have improved their library preparation protocol and base callers to improve the base accuracy. PacBio, for example, came up with circular consensus sequencing protocol in 2014, but this protocol had limited use commercially until 2018 because of insufficient DNA polymerase processivity. CLR reads from the first Sequel Platform, hence, didn’t produce CCS reads, but produced the read-out of either the forward or the reverse the strands as a CLR reads and in rare instances, the DNA polymerase would read both forward and reverse strand, but the number of times the forward and reverse strand read was read was insufficient to generate an accurate consensus sequence. CLR reads, with an average read length of 10kb, produced from the RSII platform was, however, ~100-fold longer than that produced from Illumina reads and was widely adopted for de novo assembly of more contiguous and complete plant and animal genomes and inspired the development of new de novo assembly algorithms that is optimised for the longer read length and lower base accuracy [ref]. The contigs produced from these long reads were orders of magnitude longer and complete than that produced from Illumina short reads and combined with scaffolding methods such as optical maps or Hi-C reads, could produce chromosome-length scaffolds. The development of complementary scaffolding technologies and methods has allowed the rapid production of chromosome-length scaffolds [ref]. In addition, as the PacBio sequencing errors are thought to be introduced to the read randomly, the errors do not occur frequently in the same position and hence, highly accurate consensus sequence for the contigs could be achieved with an error rate between Q40-Q50. In addition, these longer reads enabled the discovery of structural variations that could not be detected with short reads. These longer reads were especially useful in detecting SINE and LINE insertions and deletions that could not be accurately detected with short reads [ref] and other non-repetitive non-reference structural variations with ethnic differences [ref]. These longer reads doubles the number of structural variations that can be detected from the comparison of a typical human genome to the reference genome and provides the identity of the structural variations at nucleotide resolution, in comparison with short reads that can only provide nucleotide resolution for small indels and that infers the size and position of the structural variation based on the presence of discordant reads, read features such as soft-clipped reads, and sequence coverage [ref]. The ability to determine the specific sequence at an insertion site is important as the sequence can determine the severity of disease [ref]. Moreover, instead of performing transcriptome-sequencing and de novo assembly of RNA reads to recover the transcript sequence, Isoform-sequencing allows for direct full-length sequencing of the transcriptome, enabling the discovery of new tissue-specific isoforms and isoforms is disease-specific [ref, ref]. The BAC clone assembly and tiling of gaps in the human reference genome. The use of ONT platform for full-length sequencing of BAC clones and assembly of human chromosome Y centromere.

%% repeat-sequencing

%% pathogenic mutations

**Circular consensus sequencing**

The lower base accuracy and higher sequencing cost, however, limited the use of PacBio long-reads to de novo assembly and occasional discovery of pathogenic mutations in clinical sequencing settings. Together with the release of PacBio Sequel II platform, PacBio released SMRTcell with increased number of ZMWs from 1 million ZMWs to 8 million ZMWs and substantially increased the DNA polymerase processivity, which allowed for multiple sequencing of both the forward and reverse strand of the circulate template and construction of highly accurate consensus sequence, also known as circular consensus sequence (CCS) or high-fidelity (HiFi), reads. These CCS reads have an average read accuracy between Q20 and Q30 and have base quality scores ranging from Q1 to Q93. The higher base accuracy combined with longer read length enabled the construction of even more complete and contiguous assemblies with less sequence coverage [ref]. In addition, the higher base accuracy allows for de novo assembly algorithms to distinguish reads coming from two different haplotypes such that haplotype-resolved assemblies can be produced [ref, ref]. If de novo assembly algorithms are complemented with trio-sequencing, reads can be binned to paternal and maternal haplotypes to construct even more contiguous haplotype-resolved genomes and de novo assembly algorithms no longer must assume that the source of DNA is from a haploid genome. If trio-sequencing is not available, haplotagging can be alternatively performed to achieve haplotype-resolved assemblies.

%% de novo assembly (OLC, string graph, haplotype phased)

%% T2T

%% pangenomes

%% scaffolding methods

%% haplotype phasing

%% deepvariant

%% imprinting

%% compound heterozygous

%% methylation

%% %% 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

%% pairwise sequencing

%% increase in compute power

In PacBio SMRT sequencing, DNA polymerase processivity and DNA damage repair during CCS library preparation, potentially is the most important factor for SMRT sequencing. DNA polymerase processivity determines the length of the template used for CCS library preparation and the base accuracy of the CCS reads. In addition, if DNA damage repair isn’t’ performed with sufficient fidelity, remaining DNA damage in the template can lead to earlier than expected termination of the SMRT sequencing process as DNA polymerase in the ZMW cannot perform DNA synthesis beyond the bulky DNA damage.

%% subread-throughput/single-strand read-throughput

**The Darwin Tree of Life (DToL) project**

These technological improvements in sequencing technologies, scaffolding technologies, cost of sequencing and de novo assembly algorithms have prompted many international collaborations to sequence and assemble and provide high-quality reference genomes of plant and animal genomes. The Wellcome Sanger Institute also initiated the 25 genomes project and the Darwin Tree of Life project to sequence and assemble all 66,000 eukaryotic species in Britain and Ireland [ref]. The DToL project has considered a number of sequencing technologies, but circular consensus sequencing has been adopted as the primary sequencing technology for de novo assembly.

**Germline and somatic mutation calling across the Tree of Life**

%% We and other [ref]

%% questions, peto’s paradox

%% speciation

%% evolution

%% mechanism that generates non-sense mutations, sense mutations, etc,

**Thesis objectives**

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

*The first principle is that you must not fool yourself, and you are the easiest person to fool*. [Richard Feynman]

Based on our understanding of duplex sequencing methods and the recently developed nanorate sequencing protocol, a derivative of the duplex sequencing protocol [ref, ref], and considering the similarities between two sequencing methods, we hypothesized that CCS reads might be as accurate or more accurate as duplex reads and can be used for single molecule somatic mutation calling. Both methods take advantage of the complementary information in the double-stranded DNA molecule to identify library errors and sequencing errors and to produce a highly accurate double-stranded consensus sequence (Fig 1). In the duplex sequencing protocol, a unique molecular identifier (UMI), consisting of 8 to 12 nucleotides, is added to the double-stranded DNA before the separation of the DNA molecule into single-strands for PCR amplification. Illumina adapters are added to the PCR amplified DNA, library is sequenced, and single strand reads belonging to the same DNA molecule is identified through their UMI. Because the forward and reverse strand is complementary and because DNA polymerase introduces errors infrequently during PCR amplification, library error and PCR amplification error should be present in one or two of the molecules, but not in all the molecules and most molecules should be consistent with each other such that the correct base can be ascertained. [bottleneck sequencing]. In contrast to the duplex sequencing protocol, CCS library preparation protocol doesn’t require a UMI as a single circular template is loaded to a zero-mode-waveguide (ZMW), DNA polymerase attaches to the template and sequences the circular template in a manner similar to rolling-circle-amplification to sequence both the forward and reverse strand multiple times until DNA polymerase terminates DNA synthesis [ref]. During DNA synthesis, DNA polymerase incorporates fluorescently labelled nucleotides and the fluorescence that is emitted during DNA synthesis is captured and fluorescence signal is converted to a nucleotide base. Thereafter, PacBio circular consensus sequence (pbccs) algorithm is used to generate CCS reads and for base quality score estimation.

If CCS reads have sufficient base quality score to enable single molecule somatic mutation calling and if DToL project generates high-quality reference genomes from these reads, we thought that we have an unparalleled opportunity to investigate the somatic mutational processes of non-human species at scale (discussed in Chapter 3). To enable, the study of somatic mutational processes in non-human species, we sequenced positive and negative control samples to assess the feasibility of single molecule somatic mutation detection with PacBio CCS reads, demonstrate that CCS reads indeed have sufficient base quality score for single molecule somatic mutation detection and develop and benchmark a method to call somatic mutations agnostic of clonality and species (discussed in Chapter 2). Based on our understanding of unique characteristics of CCS reads, we also sequenced and analysed granulocytes from Bloom syndrome patients and sperm samples of different ages to detect and analyse non-crossover and crossover resulting from mitotic and meiotic recombination at scale and across the whole genome (discussed in Chapter 4).

%% opportune to perform population-genomics on other species

%% understand evolution through time and space

Here, we take advantage of CCS read length and high base accuracy to detect mutations at variant allele frequencies that was not possible before and to detect mutations resulting from double-strand break repair that could not be detected with short-read sequencing technologies.

%% \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 cost of Reference genomes as a bottleneck}

%\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\* %Second Section \*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*

%% \section{Reference genomes}

%% \subsection{De novo assembly}

%% bacterial artificial chromosomes

%% yeast artificial chromosomes

%% \subsection{Short-read sequencing}

%% de brujin raph

%% \subsection{Long-read sequencing}

%% linked-read sequencing

%% single-moelcule sequencing

%% oxford-nanpore technologies

%% Pacific Biosciences circular consensus sequencing

%% overlap-layout consensus

%% string graph

%% falcon

%% haplotype-phased

%% \subsection{Haplotype tagging}

%% \subsection{High-throughput chromatin conformation capture sequencing}

%% optical-mapping

%\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\* % Third Section \*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*

%% \section{Resequencing}

%% \subsection{Germline mutation detection}

%% \subsection{Somatic mutation detection}

%% \subsection{Somatic mutation detection in cancer}

%% \subsection{Somatic mutation detection in normal tissues}

%% \subsubsection{Single-cell expansion and sequencing}

%% \subsubsection{Laser-capture microdissection and sequencing}

%% \subsubsection{Single-cell DNA PCR amplification and sequencing}

%% \subsubsection{Duplex sequencing}

%% \subsubsection{Pacific Biosciences circular consensus sequencing}

%\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\* % Fourth Section \*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*

%% \section{Darwin Tree of Life project}

%\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\* % Fifth Section \*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*

%% \section{Sperm sequencing for meiotic recombination product investigation}

%\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\* % Sixth Section \*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*

%% \section{Bloom syndrome patient sample sequencing for mitotic gene conversion detection}

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.