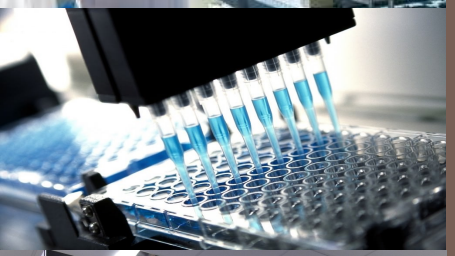
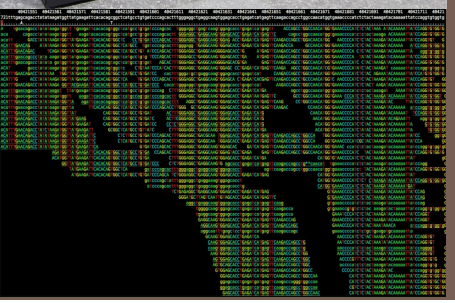




CS3110/CS5110
Bioinformatics

Next-generation Sequencing (NGS)

By Jamie Alnasir PhD candidate
PhD supervisor Dr. Hugh Shanahan

Overview of this lecture

- Contents
  - Recap on Molecular Biology – basic definitions
  - Transcriptomics experiments – datasize
  - Nucleic acid sequencing – how it's done
  - Massively parallel Next-generation Sequencing (NGS)
  - RNA-sequencing (RNA-seq) – estimating expression
  - RNA-seq file formats – annotation and aligned reads files
  - Analyses of RNA-seq. "bigdata" – cluster computing, cloud

Life is dictated by Nucleic acids

- **Recall central dogma of Molecular Biology:**
  - **DNA → RNA → Protein**
- Life, as we know it, could not exist without nucleic aids or their protein products
  - **DNA** and **RNA** - large **self-replicating** biomolecules known as polymers, comprised of monomers, i.e. A,T,G,C, and U
  - Constitute **Genome** and **Transcriptome** of all living organisms
  - Used to **determine *what* proteins** are synthesised
  - Designating letters for the chemical monomer “unit” structure facilitates digitisation, this is known as **Sequencing**

Genomes and Transcriptomes

- **Genome**
  - **complete set of genetic material** (DNA, or RNA in viruses).
  - consists of genes, exons (coding regions) and introns (non-coding regions, operons in viruses.
- **Transcriptome (transcribed from the Genome)**
  - **sum total of all the messenger RNA** (mRNA) molecules **expressed** from the genes of that organism.
  - **highly dynamic**, complex and **in a constant state of flux**
  - accommodates the cells **constantly changing** requirement
    - a result of intra- and extra-cellular stimuli as well as pathology (disease processes).
  - **gene expression** studies the products of transcription (mRNAs)

Transcriptomics – quantifying gene expression

- **Analysis of transcriptome yields insights into cellular function**
  - Alternative **splicing isoforms** (different gene variants from splicing)
  - **post-transcriptional modifications** (processes applied to RNA)
  - **disease processes**
  - **expression** (of genes)

- **Gene expression - quantification of mRNA transcript molecules – a measure of how many transcripts are being produced**
  - dictates the **amount of particular proteins** in a cell
  - protein products of transcription **determine activity** in the cell
  - quantities **reflect the balance of synthetic and degradative metabolic pathways** in the cell
  - **enzymes** that carry out intra-cellular chemistry are proteins themselves!
  - two main techniques: **Microarrays** and **RNA-seq**
  - **Measurement is tricky – *we'll come to that…***

Nucleic acids = complex, "bigdata"

- **Consider Transcriptome of *H. sapiens***
  - estimated to comprise of approximately 3.72×10^{13} cells [1]
  - mean gene size of ~ 28 kilobase pairs [2]
  - approximately 300,000 RNA molecules per cell [1]

- **a full representation of the transcriptome comprises of…**
  - approximately 3.1×10^{23} (28,000 x 300,000 x 3.7×10^{13}) RNA bases.
  - Sequencing of the transcriptome (and genome) therefore generates large amounts of very complicated raw experimental sequencing data

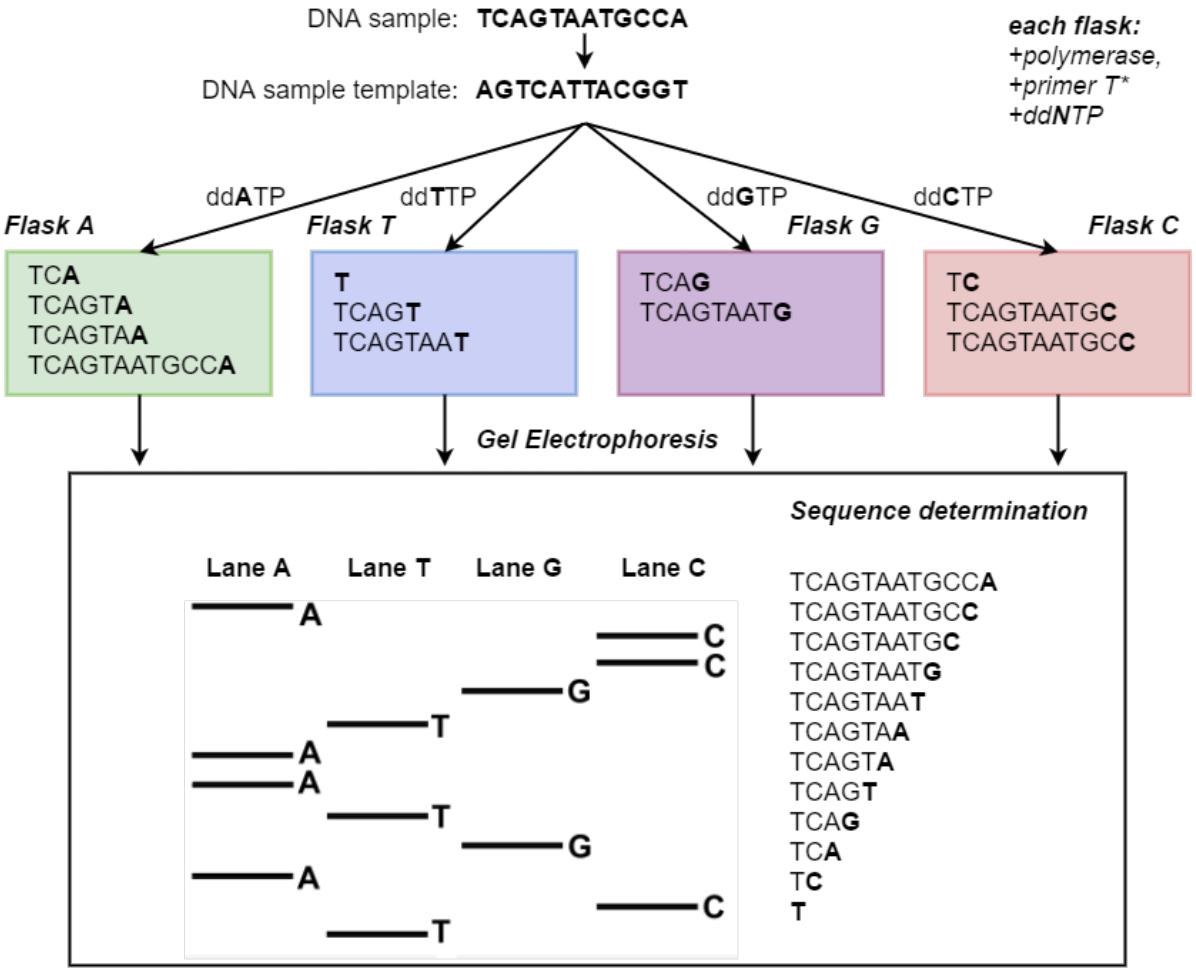
Many sequencing experiments = "bigdata"

- Many sequencing experiments are being carried out worldwide
- Much of the data is deposited in **public databases** such as:
  - **Sequence Read Archive (SRA)**
  - **Gene Expression Omnibus (GEO)**
  - **ArrayExpress**
- This **data is growing**, looks to outpace astronomical data: [3]
  - In just a year (from 2010 to July 2011) the amount of data deposited in just the SRA increased by an order of magnitude from 10 tera (10^{12}) bases to surpass 100 tera bases
- SRA currently holds over 9 peta bases (9.377×10^{15}) of seq. data[4]
- Bigdata – characterised by the three Vs: *volume, velocity, variety*

- **How to process this "bigdata"?** – *we will come to that…*

How are nucleic acids sequenced?

- Original chemical method devised by Frederick Sanger[5], 1977



Complementary template is made from DNA sample.

A primer is added to start the reaction. Polymerase enzyme carries out all the reactions.

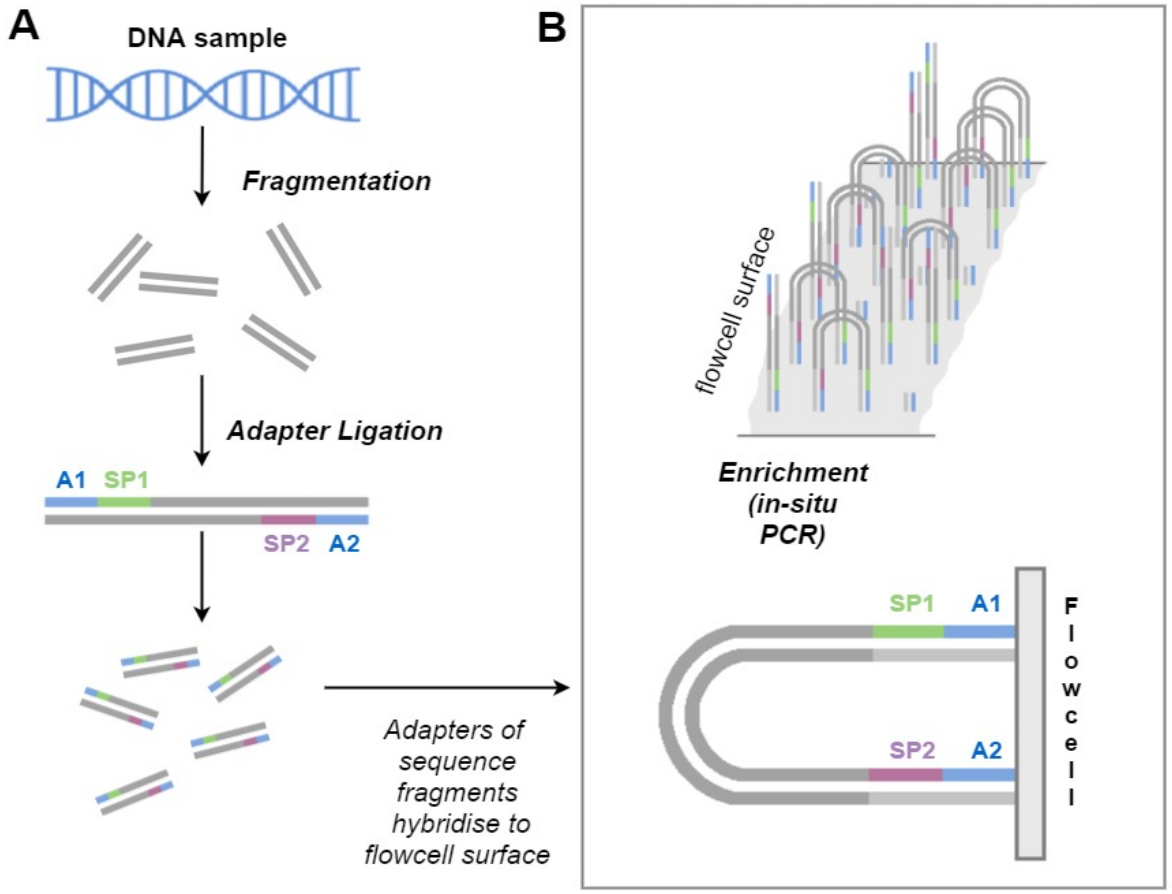
Normal nucleotides (d**N**TPs) are added to each flask, where N is the letter A,T,G,C

Modified nucleotides (**ddN**TPs) are added to flask. These modified nucleotides i.e. ddATP will terminate the sequence after their letter.

Fragments are size separated in different lanes on a gel to elucidate sequence.

Massively parallel Next-generation seq.

- New NGS methods massively parallelise[6] sequencing process



DNA sample is fragmented to size optimal for the machine (current machines are limited to reading short fragment lengths)

Synthetic “known sequences”, called adapters are ligated (joined) to unknown sequence fragments.

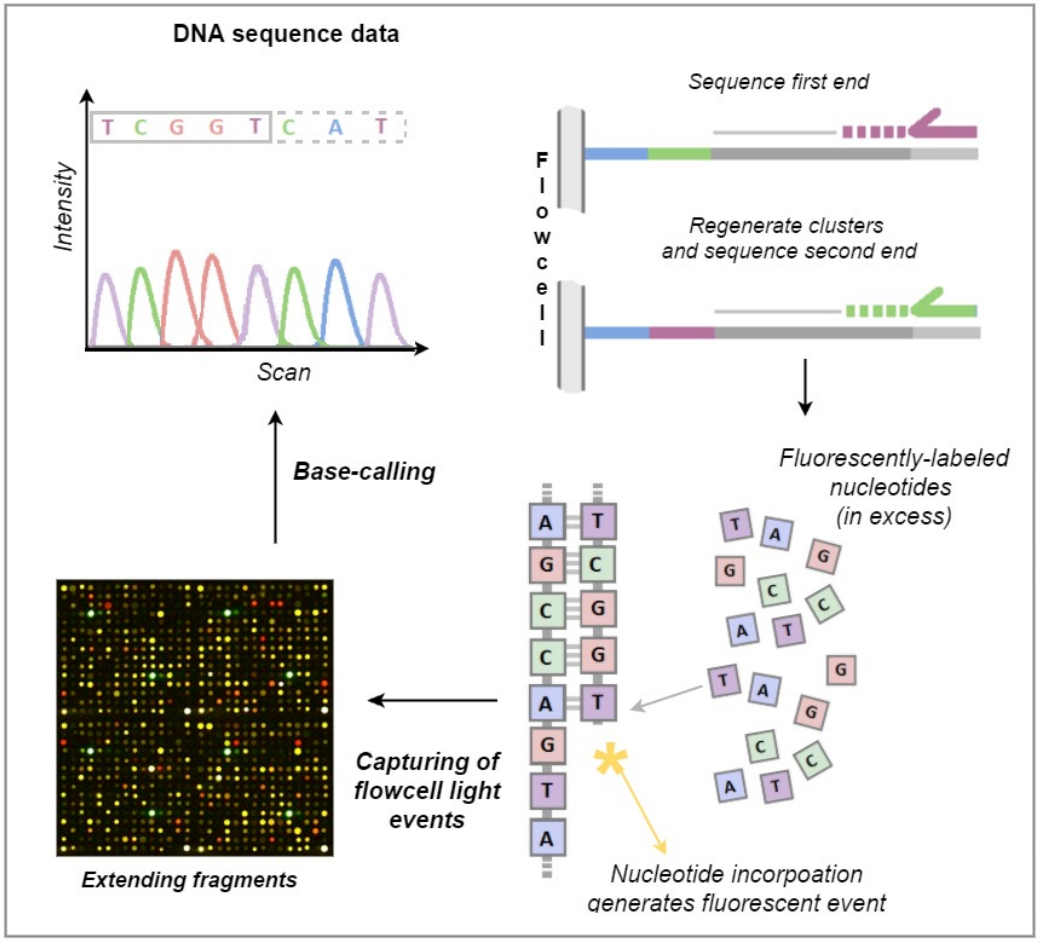
Adapters are complementary to short DNA fragments etched onto flowcell surface using photolithography. The adapter terminated fragments join to the flowcell.

Fragments are then replicated using PCR amplification method …. ***continued*** ….

Molecular bias introduced in these steps inevitable, affects measurement

Massively parallel Next-generation seq.

- Reading the clusters of sequences and base-calling



After PCR amplification, the adapter terminated unknown seq. fragments form colonies known as clusters of the same unknown sequence. Necessary to gain sufficient signal (nucleotides are tiny!)

An excess of fluorescently labelled nucleotides are washed over the flowcell.

These labelled nucleotides fluoresce when they undergo the chemical reaction that joins them to their complementary strand.

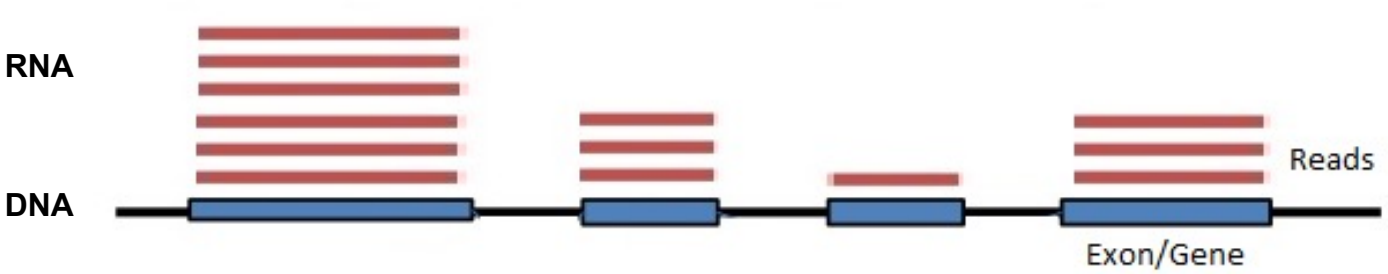
This light event is captured. The wavelength of the light is used to determine which nucleotides added, therefore revealing sequence known as “base-calling”.

RNA-seq

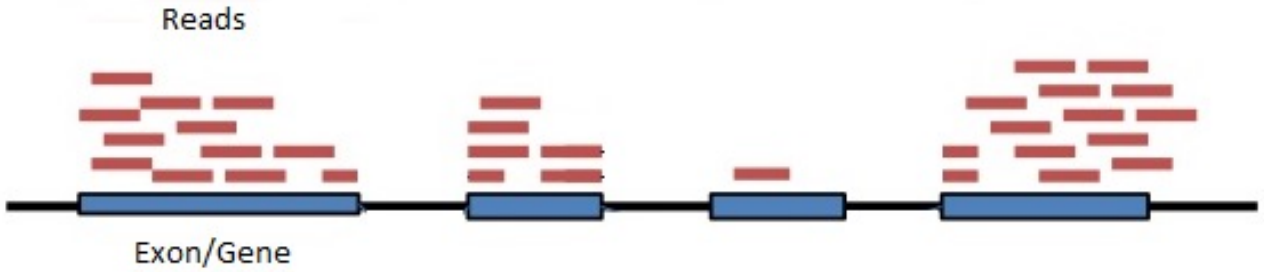
- **RNA-seq**
  - **overcomes limits of Microarrays** (which use fixed probe sequences) and uses Next-generation massively parallel tech.
  - **direct sequencing of RNA transcripts** from the sample
    - can study all the types of RNA [7]
      - Total RNA, pre-mRNA (before splicing), ncRNA (non-coding RNA), mRNA
  - uses a **Reference Genome** (for the species under investigation)
    - contains boundaries of where exons of genes begin/end
    - RNA fragments mapped to these regions (using alignment)
      - NB: Transcriptome assembly is impractical because of the short length and quantity of fragments
  - **gene expression** is measured by counting the mRNA fragments (called *reads*) that map to a given exon/gene…

RNA-seq read fragment distribution

- ***in an ideal world…***
  - there would be **no biases** and **uniform coverage** of long **contiguous reads** along a gene/exon…



- ***in reality…***
  - Sequencing machine can **only read short fragments**
  - some **fragments under/over-expressed due to biases**



RNA-seq expression estimates

- **Normalising the expression estimates**
  - RPKM (Reads per Kilobase Million)[8] is used to normalise the counts of mapped fragments:

$$RPKM = \frac{10^9 \times C}{NL}$$

C = count of mappable reads on the feature (i.e. on the exon)
N = number of mappable reads in experiment (in millions)
L = sum of exons (in Kb)

- **Improving expression estimates and understanding bias**
  - Bias in microarray data is well characterised
  - Bias in RNA-seq data is rather complicated because of the preparatory steps as well as sequencing of small fragments
  - Types of bias:
    - **Sequence specific:** certain bases problematic to prep and sequence
    - **positional bias:** abnormal i.e. predictable base frequencies at specific positions of the sequence fragments

RNA-seq file formats

- **Genome annotation file format (GTF/GFF)**
  - contains the boundary information for genomic features (exons, start codons, coding sequences etc…) for a species of organism

```
chr1    hg38_refGene    start_codon     67127238        67127240    ...
chr1    hg38_refGene    exon            67127166        67127257    ...
chr1    hg38_refGene    exon            67131142        67131227    ...
chr1    hg38_refGene    exon            67134930        67134971    ...
chr1    hg38_refGene    stop_codon      67093580        67093582    ...
```

  - Fields
    - seqname - name of the chromosome,
    - feature - i.e. exon, CDS (coding sequence), start/stop codons
    - feature start, feature end
    - … and some other fields, not shown in image …
      - score, strand + (forward) or - (reverse).
      - frame, attribute

RNA-seq file formats

- **Aligned reads (sequence alignment map) file format (SAM)**
  - contains the mapped RNA-seq reads
  - output produced by the *sequence alignment program*

```
HWI-ST1145:74:C101DACXX:7:1102:4284:73714       16      chr20   190930  3       100M    *       0       0
       CCGTGTTTAAAGGTGGATGCGGTCACCTTCCCAGCTAGGCTTAGGGATTCTTAGTTGGCCTAGGAAATCCAGCTAGTCCTGTCTCTCAGTCCCCCCTCT
C    BBDCCDDCCDDDDCDDDDDDCDCCCDBC?DDDDDDDDDDDDDDDCCDCDDDDDDDDDDCCCCEDDDC?DDDDDDDDDDDDDDDDDDDDDBDHFFFFDC@@
    AS:i:-15        XM:i:3  XO:i:0  XG:i:0  MD:Z:55C20C13A9 NM:i:3  NH:i:2  CC:Z:=  CP:i:55352714   HI:i:0
HWI-ST1145:74:C101DACXX:7:1114:2759:41961       16      chr20   193953  50      100M    *       0       0
       TGCTGGATCATCTGGTTAGTGGCTTCTGACTCAGAGGACCTTCGTCCCCTGGGGCAGTGGACCTTCCAGTGATTCCCCTGACATAAGGGGCATGGACGA
G    DCDDDDEDDDDDDDCDDDDDDDCCCDDDCDDDDDEEC>DFFFEJJJJJIGJJJJIHGBHHGJIJJJJJJGJJJIJJJJJIHJJJJJJHHHHHFFFFFCCC
    AS:i:-16        XM:i:3  XO:i:0  XG:i:0  MD:Z:60G16T18T3 NM:i:3  NH:i:1
```

  - Fields
    - read name, bitwise flag for pairing/strand
    - reference name, mapping position, mapping quality, CIGAR string
    - reference for the mate, position of the mate, distance between paired reads
    - read sequence, read quality

Processing RNA-seq data - Cluster computing

- transcriptomics datasets are typically 100s of gigabytes
- production wrt Moore's law -- cheaper to sequence than to store!
- often necessary to perform analysis across a number of datasets
- typically data is processed through "Pipelines" – a series of analysis programs which are run on the data

- **Cluster computing often used for RNA-seq data**
  - Traditionally in-house clusters
  - Now, cloud service providers have emerged like AWS, Azure
    - Leasing of virtual machines, scalable to requirements
    - Cloud providers offer access to large public scientific databases
    - Bioinformatics projects migrating to the cloud [9]
- **More to be done to make best use of distributed tech.** [10]
  - i.e. Spark, MapReduce etc… this is my research focus!

References

1. Velculescu, Madden et al. (1999). Analysis of human transcriptomes. Nature genetics, 23(4):387-388.
2. Venter, J. C., Adams, et al. (2001). The sequence of the human genome. science, 291(5507):1304-1351
3. Stephens, Z. D., Lee, S. Y. et al. (2015). Big data: Astronomical or genomical? PLoS Biol, 13(7):1-11.
4. SRA (Sequence Read Archive), S. R. (2017). Overview of the Sequence Read Archive (SRA). https://trace.ncbi.nlm.nih.gov/Traces/sra/sra.cgi/
5. Sanger, F., Nicklen, S., and Coulson, A. R. (1977). Dna sequencing with chain-terminating inhibitors. Proceedings of the National Academy of Sciences, 74(12):5463-5467.
6. Metzker, M. L. (2010). Sequencing technologies - the next generation. Nature reviews. Genetics, 11(1):31-46.
7. Park, J. W., Tokheim, C., Shen, S., and Xing, Y. (2013). Identifying differential alternative splicing events from rna sequencing data using rnaseq-mats. Deep Sequencing Data Analysis, pages 171-179.
8. Mortazavi, A., Williams, B. A., McCue, K., Schae er, L., and Wold, B. (2008). Mapping and quantifying mammalian transcriptomes by rna-seq. Nature methods, 5(7):621-628.
9. Stein, L. D. (2010). The case for cloud computing in genome informatics. Genome biology, 11(5):207.
10. Taylor, Ronald C. "An overview of the Hadoop/MapReduce/HBase framework and its current applications in bioinformatics." *BMC bioinformatics* 11.12 (2010): S1.