1  CHAPTER 1: Big Biology

1.1  Big Data

Big data is the buzzword today that is synonymously used to describe the immense volumes of data on the one hand, and to describe the science that allows one to mine the data to add value to businesses, individuals and society at large. The latter is also known as Data Science.

Large data in science and businesses are not new. What has changed in the last decade is the life like characteristics displayed by data because of increasing bandwidth on internet, impressive advances in technologies, and unimaginable widespread access to devices on everyone’s fingertips that can generate data. Decades ago big data generation was the luxury enjoyed by few big companies or governments, which used mainframe computers to extract value out of data by spending billions of dollars. Today more than 5 billion people are calling, texting, tweeting and browsing on mobile phones worldwide. The mass has the potential to generate valuable data for organizations/business to increase margins and/or add process efficiency.

Figure 1.1 Overview on BIG DATA

The sheer volume of data, the increasing velocity with which it is generated every minute, and the variety in the type of data has made it impossible to effectively process them with traditional applications. Here, the “Big Data” refers to tools and methods developed to process big data also, known as Data Science, is a field that encompasses anything related to data cleansing, preparation, and analysis. Put
simply, Data Science is an umbrella term for techniques used when trying to extract insights and information from big data. For example, Big Data can be used for tracking purchase patterns by customers to manage inventory, mine disparate data posted on Facebook for targeted advertising, manage web-information by Google to improve search efficiency, predicting trends to design successful products, predict weather, and reduce business risks.

Impressive advances in Data Science are now encouraging data generation itself as a strategy for businesses and research. In other words, generating big data has become a lucrative/profitable business attracting hundreds of millions of US dollars. With rising healthcare costs in the US, the urge for such efforts are nowhere as visible as in healthcare industry. What is really needed is to be able to predict drug efficacy, reduce cost of drugs, predict adverse drug reaction, predisposition to diseases, and early diagnostics for cancer based on the genetic profiling. And, this would require correlating genotype with diverse phenotypes.

1.2 Road to big biology

Early 1990’s could be considered the beginning of big biology when US government selected the Human Genome Project for funding over many other mega science projects including the famously known project called The Superconducting Super Collider project (SCSC). Given the funding constraint at that time, this shift is obvious. More emphasis was put on the benefit of any project to mankind. Under such an environment, who can underestimate the benefit of human genome sequencing to human health and prosperity.

Human Genome Project was funded for $3 billion to be completed in 15 years using an international consortium. What were we thinking? With technologies improving by leaps and bounds, fifteen years proved to be an overestimation. Furthermore, with benefits to mankind so obvious, HGP took a commercial angle. Healthy competition between the public and private sectors led to the completion of the HGP 5 years ahead of schedule. However, it took another decade to annotate the genome using other megaproject called ENCODE, which unearthed the importance and the extent of other aspects of the human genome that was unknown to the scientific community.

Soon after the HGP was announced, visionaries were already thinking of a day when genomes of every individual can be sequenced at affordable costs for use in personalized medicine and care. It was estimated that this feat is only possible when sequencing machines churn up an individual's genome at a cost of $1000. In 2004, NHGRI announced funding to companies that can develop technologies for sequencing human genome at the cost of $1000. Comparing this to the cost of $2.7 billion to sequence the reference human genome it was clear that it will take several incremental improvements to achieve this. Solexa (Purchased by Illumina), 454,
Applied Biosciences and other companies handsomely received funding from NIH to meet this milestones. In 2006, George Church made a case for Personal Genome Project (Church, 2005). In 2007, Baylor College of Medicine ceremoniously announced the first personal genome of James Watson sequenced using 454 technologies for a cost of $1 million. While this cost was 1000 fold higher than the goal of $1000, it was 3000 fold less than the original cost of sequencing the reference genome. It is worth mentioning that the 1000 fold reduction could be achieved in sheer 6 years, due to continued investment in this area.

In 2008 Knome Inc. (now Tute Genomics) offered personal genome for a sheer cost of $350,000 with real clients. The price of sequencing continued to plummet through a series of competing technologies including Helicos for sequencing a genome at the cost of $48,000 to Complete Genomics promising to have a technology that can do it at the cost of $5,000. Illumina started offering personalized genome sequencing services for $50,000/person in 2010. Through another roller coaster ride against newer competing products such as Ion Proton, Illumina launched HISeq machines capable of sequencing $1000 a genome not including the cost of the machine at $10 million and electricity cost that takes to run the machine.

### 1.3 Computers in biology

During the 50’s and early 60’s, right around the momentous discovery of DNA structure, DNA emerged as the keeper of all information required by an organism. These include the central dogma of molecular biology (Crick, 1970), the translational code from DNA triplets to amino acids (Ohno and Epplen, 1983), and the theory of molecular evolution (Kimura, 1979). With the parallel advances in computers, DNA became a string, transcription and translation became software tools that operated on DNA strings. As more genes were sequenced, databases of gene sequences appeared, which could be compared for evolutionary relationships using algorithms inherited from other disciplines developed to compare electromagnetic waves. Slowly but surely, both molecular biology and computational biology were born as indispensable sibling disciplines under biology.

The unidirectional flow of information from DNA to RNA to protein, which now can be accomplished in a computer, meant that developing DNA sequencing technologies is more critical than protein sequencing technologies, which predates DNA sequencing techniques to mid-1960s Maxim Gilbert (Maxam and Gilbert, 1977) followed by Sanger Sequencing in 1970s (Sanger et al., 1977). It took two decades to decipher the DNA sequence of an entire organism, *Haemophilus influenzae* in 1995 (Fleischmann et al., 1995). It is hard to believe that within the next two decades we would have sequenced the genomes of majority of the organism of economic interest including human, mouse, rice, maize, wheat, and hundreds more plants, mammals, fungi and microbes. How did this happen?
Below we provide a chronological history of how we reached from the era DNA structure to the ear of genome sequencing.

<table>
<thead>
<tr>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>DNA Structure</strong></td>
</tr>
<tr>
<td>1985: Human Genome Project pilot project launched</td>
</tr>
</tbody>
</table>

*Figure 1.2 Pre genomic era*
The Era of the Human Genome Project

**DNA Structure**
- 1988: Human Genome Project is initiated
- 1990: HGP plan is approved
- 1991: EST, expressed sequence tags

**Genomics Technology**
- 1992: Spotted oligonucleotide microarrays

**Computer Science**
- 1991: The World Wide Web is created at CERN

**Sequence Databases**
- 1988: National Center for Biotechnology Information (NCBI) is created by NIH
- 1990: BLAST sequence similarity searching by NCBI
- 1992: ILCPSS/Structure Meca6s launched by Henikoff and Henikoff

**Sequence analysis tools**
- 1994: Hidden Markov Model for protein sequence alignment by Hausler
- 1996: The PROSITE database is announced by Barroch
- 1997: Clustal_X for multiple sequence alignments
- 1998: HMMER suite of software by Eddy for identifying functional domains in proteins

**BIRTH OF BIOINFORMATICS AS A DISCIPLINE**
- 1995: Haemophilus influenzae genome is sequenced
- 1996: Yeast and Methanococcus jannaschii genome sequenced
- 1996: GeneChip oligonucleotide microarrays launched by Affymetrix
- 1995: tDNA microarrays
- 1996: Silicon Graphics launches Origen 2000, a high-end server
- 1996: Transfac database of transcription factor binding sites
- 1996: Specialized computer cluster to analyze sequence analysis by Tmple Logic
- 1997: Phrap database created by Eddy for use with HMMER
- 1997: Unigene cluster database of expressed sequence tags
- 1999: dbSNP database created at NCBI
- 2000: Whole Genome Shotgun sequencing
- 2000: Gene Expression Omnibus created by NCBI
- 2003: The official completion of the Human Genome Project

**Figure 1.3 Era of human genome project**
1.4 Biology joins Big Data Club

Impressive technological advances in sequencing relied on computers as much as on smart reagents. For example, sequencing short reads (shotgun sequencing), a key to plummeting costs, would not have been possible without computational techniques that can put them together in tandem to construct longer reads/contigs. Using simulated reads for human genome, Eugene Myers had demonstrated that shotgun sequencing methods can also work to assemble complex eukaryotic genomes (Weber and Myers, 1997).

![Figure 1.4 Plummeting cost of human genome](image)

The completion of the HGP five years ahead of schedule owes to advances in computational tools to make sense of short read. The beauty is that one could skip the laborious step of BAC cloning to place the reads in order. The usefulness of short reads taken by the private sector provided the foundation for the development of the first massively parallel next generation sequencing platform (NGS).

The advances in technologies continued to increase the sequencing throughput resulting in plummeting costs in sequencing (Figure 1.4). This was not without challenges. Now using the same amount of funding one could generate orders of magnitude more sequences. For example, according to the chart above, in 2011-2012 one could sequence 100 gigabases for $10,000, which could only produce 1 million bases in 2001. It is interesting to note that the price drop is not smooth or continuous. There is a sudden downward spiraling cost in 2007, suggesting a breakthrough in technology.

By the end of the first decade in this millennium, NIH and other major funding bodies across the world invested in data generation in anticipation that someday it will help human health and disease. Hundreds of thousands of samples/genomes were sequenced increasing the size of the data repositories (Figure 1.5).
The increase in sequencing throughput and investment in big data generation defied Moore’s law of growth in compute power (Figure 1.4). Data analysis cost continued to increase even as the cost of sequencing fell (Sboner et al., 2011). Sheer size of the files created by a single run became so huge that it was cheaper and faster to ship it on hard drives than transferring them over the internet. The table below lists the chronological order of events that took place since the launch of the HGP.

<table>
<thead>
<tr>
<th>Timeline</th>
<th>Task</th>
<th>Goal</th>
</tr>
</thead>
<tbody>
<tr>
<td>2001</td>
<td>Completion of the human genome sequencing</td>
<td>To decipher human proteome</td>
</tr>
<tr>
<td>2004</td>
<td>Launch of $1000 genome sequencing project</td>
<td>To sequence large number of individual genomes</td>
</tr>
<tr>
<td>2004</td>
<td>Biobanking consortia</td>
<td>To sequence cancer tissues</td>
</tr>
<tr>
<td>2007</td>
<td>Start of plummeting cost of sequencing</td>
<td>Plummeting costs in sequencing</td>
</tr>
<tr>
<td>2009</td>
<td>$145 million stimulus money for sequencing</td>
<td>To sequence samples in the biorepository</td>
</tr>
<tr>
<td>Year</td>
<td>Technology</td>
<td>Sequence</td>
</tr>
<tr>
<td>------</td>
<td>------------------------------------</td>
<td>---------------------------</td>
</tr>
<tr>
<td>2011</td>
<td>Creation of NCATS</td>
<td>To translate finding to clinic</td>
</tr>
<tr>
<td>2014</td>
<td>Cancer Genomics Cloud Pilot</td>
<td>To provide public access to sequence repositories</td>
</tr>
</tbody>
</table>

Table 1-1 Timeline for major achievements in genomics

<table>
<thead>
<tr>
<th>Year</th>
<th>Technology</th>
<th>Sequence</th>
<th>Tools</th>
</tr>
</thead>
<tbody>
<tr>
<td>2001</td>
<td>Human</td>
<td>Bowtie</td>
<td></td>
</tr>
<tr>
<td>2004</td>
<td>Mouse</td>
<td></td>
<td></td>
</tr>
<tr>
<td>2007</td>
<td>454/ Illumina acquires Solexa</td>
<td>Newbler</td>
<td></td>
</tr>
<tr>
<td>2008</td>
<td></td>
<td>Velvet</td>
<td></td>
</tr>
<tr>
<td>2009</td>
<td></td>
<td>SOAPdenovo</td>
<td></td>
</tr>
<tr>
<td>2011</td>
<td></td>
<td>QIIME</td>
<td></td>
</tr>
<tr>
<td>2012</td>
<td>Hiseq 2500, Miseq, Infinium HumanCore BeadChip</td>
<td>1 genome a day</td>
<td></td>
</tr>
<tr>
<td>2013</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>2014</td>
<td>Next-seq 500/Hiseq X ten</td>
<td>Affordable machine/$1000 a genome</td>
<td></td>
</tr>
<tr>
<td>2015</td>
<td>Hiseq 3000-4000, Pacific BioScience, PromethION, CHiCAGO</td>
<td>9000 human genomes per year</td>
<td></td>
</tr>
<tr>
<td>2016</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>2018</td>
<td>Nanopore direct RNA-seq</td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

Table 1-2 Advancements in genomics
1.5 Applications of Big Data in Human Disease

High throughput sequencing has a promise for every aspect of human health, disease and wellbeing. These may include cancer diagnostics, monitoring drug response, molecular breeding, soil management, revealing ancestry and management of ecosystem. Most of all, it has the potential to improve our understanding of biology and evolution, in general.

1.5.1 Cancer

Cancer is a heterogenous disease, a clonal disease and a genetic disease at the same time. Meaning, one drug will not cure all. In fact, lung cancer from two different individuals and two time points from the same individuals can display distinct molecular/mutation profiles. There is a demand towards understanding the genotype behind each cancer type using NGS. It is now possible to predict the efficacy of a drug to a given individual without trial and error method, which wastes a lot of precious time during which cancer continues to grow. Besides, the side effects from ineffective drugs is currently a major tragedy.

Revealing the genotype of a given cancer can help in multitude of ways. It can help design a proper treatment regimen, offer companion diagnostics for a given drug, and predict predisposition to some types of cancer. More recently, tumor DNAs in plasma can be used to monitor drug efficacy/resistance. Below I have provided some classic examples for these use case scenarios.

Therapeutics

Today we have a drug that is designed to target a subclass of leukemia that expresses BCR-ABL fusion genes (Kurzrock et al., 1986). Again, BRAF V600E mutation was found as a biomarker in tumor samples from > 30% of melanoma cases. Drugs that are sensitive to these mutations are now used to offer personalized treatment (Fisher and Larkin, 2012).

Diagnostics

Biomarkers are now used as companion diagnostics to offer personalized treatment regimen. HER2 expression level is a companion diagnostics for treatment of breast cancer with HERCEPTIN (Scheerens et al., 2017). More recently, PDL1 expression levels are correlated with response to PDL1 inhibitors (Abdel-Rahman, 2016). PCA3 is used as a diagnosis in urine samples for prostate cancer (Ploussard et al., 2010).
Prognostics

Prognostic markers are those that provide likely outcome of treatment or recurrences. PSA level is a good example of prognostic marker in prostate cancer (Collette et al., 2003). In HER2 positive breast cancer a mutation in PIK3CA gene is a prognostic marker (Pang et al., 2014).

Cell-free DNA

The ease of access to blood, urine, and saliva has triggered interest in using these as a surrogate tissue for cancer diagnosis, prognosis and design of personalized therapy. The potential of cell-free DNA as a surrogate tissue for primary and secondary cancer tissues has gained much attention recently. For example, the amount of tumor DNA in plasma, called DNA integrity, has also been explored as a potential diagnostic (Schwarzenbach et al., 2009). The higher the DNA integrity the more likely it is to predict cancer mutations or methylation status in primary tissue from cell-free DNA (Butler et al., 2015) (Warton and Samimi, 2015).

1.5.2 Rare genetic disorders

Rare genetic disorders are now addressable because of advances in NGS technologies. There are two types of genetic disorders. One is inherited and the other is caused by spontaneous defects in cell division during embryogenesis. Depending on the type of inheritance one can design NGS experiments to discover causative mutation. On the other hand, spontaneous defects from rare mutations are challenging to discover.

Inherited

Inherited disorders can be autosomal dominant or recessive, X-linked dominant or recessive. Sequencing the genomes of mother, father, affected children (proband) and unaffected children can reveal the mutation that causes autosomal recessive disorder for which the proband is homozygous. On the other hand, to discover the causative mutation for an autosomal dominant disorder, the affected parents and the affected child will share the mutation, which is missing in unaffected parents and siblings.

Congenital

Several types of trisomy are known to cause congenital disorder including Down syndrome and are easily identified by karyotyping. However, point spontaneous mutations causing congenital disorder such as defective heart can be a challenge. Genome sequencing allow for massively parallel sequencing of many regions helping in simultaneous identification of such mutations.
One can even use targeted gene panel focusing on only certain genes, which are known to be implicated in the disorder. With the current novel techniques like Hi-C one can even look at chromosomal level abnormalities with more details.

### 1.6 Genetic Elements of Interest as Biomarkers

Biomarker is a much generalized term used for any kind of variation in any of the biological molecules that either cause a phenotypic change or results from a phenotypic change. Sugar in urine is a classic biomarker for diagnosis of diabetes. Blood cholesterol level is another biomarker for high blood pressure. NGS offers another layer of biomarkers that allows measurement of small changes at the macromolecular levels, such as DNA and RNAs that either correlate with a phenotype or plays a role in eliciting the phenotype. Here, I will limit the types of changes within the various genetic elements that can result in changes in phenotypes. The flowchart below provides example of some genetic elements from genome, transcriptome and methylome. For example, CNVs/SNPs or SNVs within a genome can result in phenotypic changes. Correlating these changes with a given phenotype can lead to a potential biomarker.

![Flowchart of various genetic elements used as biomarker](image)

#### Figure 1.6 Various genetic elements used as biomarker

### 1.6.1 Chromosomal elements

A genome is the complete genetic makeup of an organism and genomics refers to the study of the structure, function, and evolution of the entire genome. A genome is usually made up of deoxyribonucleic acid (DNA) or ribonucleic acid (RNA), but not both. Although the terms genome and genome sequence are sometimes used synonymously, the former represents all of the biological information needed to build and maintain a living organism, whereas the latter is the biological information in the form of a sequence or consecutive DNA/RNA letters. Just like the words and sentences within this document, DNA/RNA sequences can be thought of as linear string of discrete elements, and since they are of the genome, they may be referred to as genomic elements.
Some changes in the genome can be disruptive of normal function leading to disease such as cancer and genetic disorder. These changes or disruptions can be used as biomarkers for diagnostic or therapeutic purposes. In plants, similar changes can lead to diverse traits. Translocation, aneuploidy, CNV, SNP, and other markers have been used as biomarkers of certain disease contexts.

**Translocation**

During meiosis or mitosis, recombination between parts of non-homologous chromosome may occur, causing two genes to get proximal to make aberrant proteins. There are two types of translocation. Balanced translocations are where no gain or loss of genetic material occurs. On the other hand unbalanced translocations result in extra or loss of genes. There are a number of maladies that are recorded as a result of translocation. The gene fusion BCR-ABL resulting from translocation is one of the biomarkers for leukemia. Among other examples translocation between chromosome 1 and 11 is implicated in Schizophrenia.

**Aneuploidy**

Aneuploidy is a chromosomal defect where there is abnormal number of chromosomes in the daughter cells during cell division. Aneuploidy has been associated with many human disorders. For example, extra chromosome 21 leads to Down syndrome.

**CNV**

A major contribution to the genome variability among individuals comes from deletions and duplications collectively termed copy number variations (CNVs). CNV basically involves unbalanced rearrangements that increase or decrease the DNA content. These defects can happen at the somatic level. Technically defining, the size of CNVs is larger than 50 bps, whereas smaller elements are known as insertions or deletions (INDELS). Certain oncogenic genes are duplicated several times in somatic tissue. One good example of that is HER2 amplification in breast cancer. Large increase in number of copies of HER2 leads to an increase in its expression level in breast cancer. TOP2A gene copy number variation and small deletion in Mtus1 gene are associated with a decreased risk of familial and high-risk breast cancer.

**SNP**

SNPs (single-nucleotide polymorphisms) are single nucleotide variations in DNA that determine susceptibility to developing some diseases including cancer. SNPs are usually germline and are known to predispose for several cancers. Tumor suppressor gene like TP53, BRAC1, EGFR are the well studies examples of SNPs which are linked with the predisposition of cancer. Single nucleotide mutations which are somatic are called as SNVs (single-nucleotide variations). Mutations in genes like PIK3CA and HER2 are the examples of somatic variations in breast cancer patients.
**Microsatellite**

Over the last few decades, the use of molecular markers has played an increasing role in rice breeding and genetics. Of the different types of molecular markers, microsatellites have been utilized most extensively, because they can be readily amplified by PCR and there are a large amount of allelic variation at each locus. Microsatellites are also known as simple sequence repeats (SSR), and they are typically composed of 1–6 nucleotide repeats. These markers are abundant, distributed throughout the genome and are highly polymorphic compared with other genetic markers, as well as being species-specific and co-dominant. For these reasons, they have become increasingly important genetic markers in rice breeding programs.

1.6.2 *Transcribed elements*

It is now known that majority of the genome transcribe, while only a small percentage codes for proteins. Hence, to understand the relationship between the genome and the functioning of cells, we need to study the expressed RNAs (mRNA or coding RNA) under every biological contexts. In multicellular organisms, nearly every cell contains the same genome. However, not every gene is transcriptionally active in every cell type, and different cells show different patterns of expression. These variations underlie the wide range of physical, biochemical, and developmental differences seen among various cells and tissues and may play a role in the difference between health and disease. Thus, by collecting and comparing transcriptomes of different types of cells or tissues, researchers can gain a deeper understanding of what constitutes a specific cell type and how changes in transcriptional activity may reflect or contribute to disease. Each gene may produce more than one variant of mRNA because of alternative splicing, RNA editing, or alternative transcription initiation and termination sites. The proportion of transcribed sequences that are non-protein-coding (non-coding RNA) appears to be greater in more complex organisms. Alternation in expression status of coding and non-coding RNA, generation of transcripts fusion, cryptic splicing of transcripts are few of the transcript features, which have the potential to be used as a biomarker.

**Coding gene expression**

Coding genes or coding RNA, encompass about 10% of our transcriptome. They are functional genes that translate to protein. Transcriptomic analyses examine the level of gene expression as the proxy of gene activity. Assessment of the gene expression profiles may explicate the molecular aspects of cancer progression, thus similarities and variations of gene expression profiles of cancer patients could be explored as potential biomarkers. For example, overexpression of *mTOR* and *c-erb-B2* may indicate tumor aggressiveness as a result of the presence of these genes in hepatic carcinogenesis (Kamel and Al-Amodi, 2017). Over amplification of androgen
receptor gene and c-MYC is used as a progressive biomarker for prostate cancer (Linja et al., 2001).

**Non-coding genes**

Non-coding RNAs are functional RNAs that do not code for a functional protein. These non-coding RNAs are majorly classified in two groups based on their length. Short non-coding RNAs, which are less than 200 nucleotide like miRNA, siRNA, etc and the long non-coding RNAs (lncRNAs) which are more than 200 nucleotide long. Recent transcriptome studies have shown that not only the coding gene expression is altered in normal and disease biology but even the non-coding RNA expression levels change. For example, miR-342 is a predictive biomarker for the response to tamoxifen in MCF-7 cell line. miR-342 is expressed only in cancer cells that are susceptible to tamoxifen therapy and blockage of its expression will lead to resistance to that chemotherapy (He et al., 2013). Expression of three miRNAs, miR-886-3p, miR-923, and miR-944, has been correlated with response to chemotherapy and survival in bladder cancer (Nordentoft et al., 2012). PCA3 is a well-studied example of a lncRNA that is highly specific for prostate tissue and overexpressed in prostate cancer and is indicative of aggressive tumors (Ploussard et al., 2010). Few other lncRNA like PCATs and CCATs are reported to be very specific to prostate (Prensner et al., 2011a) and colorectal cancer (Xie et al., 2015) respectively.

**Alternative splicing**

Inherited and acquired changes in pre-mRNA splicing have been documented to play a significant role in human disease development and it has been shown that many cancer-associated genes are regulated by alternative splicing (Ghigna et al., 2008). Loss of fidelity, variation of the splicing process, even controlled switching to specific splicing alternatives may occur during tumor progression and could play a major role in carcinogenesis (Venables et al., 2009). Cancer-specific splicing events have been reported at the mRNA level in colon, bladder, and prostate tissues, with diagnostic and prognostic implications (Climente-González et al., 2017). Splice events that affect the protein coding region of the mRNA give rise to proteins differing in sequence and activities; splicing within the noncoding regions can result in changes in regulatory elements, such as translation enhancers or RNA stability domains, which may dramatically influence protein expression. It has been shown that 30% of alternative splicing events occur in a tissue-specific manner. Cyclin D-binding myb-like transcription factor 1 (DMTF1) pre-mRNA splicing isoform, DMTF1β, is overexpressed in breast cancer and promotes mammary tumorigenesis (Inoue and Fry, 2016). Another well-known example is AR-V7, the alternate spliced isoform of androgen receptor in prostate cancer (Dehm and Tindall, 2011).
**Fusion gene**

Gene fusions are hybrid genes formed when two previously independent genes become juxtaposed. The fusion can result from structural rearrangements like translocations and deletions, transcription read-through of neighboring genes, or the trans- and cis-splicing of pre-mRNAs. Many gene fusions are associated with oncogenic properties, and often act as driver mutations in a wide array of cancer types (Latysheva and Babu, 2016). Gene fusions commonly exert their oncogenic influence by either deregulating one of the involved genes, forming a fusion protein with oncogenic functionality or inducing a loss of function. The prevalence of gene fusions varies widely between cancer types. Gene fusions were first discovered and occur frequently in 90% of lymphomas, over half of leukemias, and one third of soft tissue tumors (Latysheva and Babu, 2016). Chronic myelogenous leukemia (CML) is characterized by the Philadelphia chromosome (Ph) resulting from a balanced translocation between chromosome 9 and 22 which leads to the formation of the BCR/ABL fusion gene (O'Brien et al., 1994). Large number of studies have evaluated the molecular and cellular mechanisms contributing to CML, and a number of signaling pathways activated by BCR-ABL (Salesse and Verfaillie, 2002). In prostate cancer, one specific fusion (TMPRSS2-ERG) is the most common genetic alteration, being found in over 50% of patients (Gopalan et al., 2009). KIF5B-RET fusion, occurs in 1–2% of lung adenocarcinomas (Huang et al., 2016).

### 1.6.3 Epigenetic elements

Certain DNA modifications do not change the DNA sequence but can still affect gene activity. These modifications can regulate gene activity and are known as epigenetic changes. The epigenome is made up of chemical compounds and proteins that can attach to DNA and direct actions such as turning genes on or off, thus controlling the production of proteins in particular cells. When epigenomic compounds attach to DNA and modify its function, they are said to have "marked" the genome. These marks do not change the sequence of the DNA. Rather, they change the way cells use the DNA's instructions. They are either passed down from one generation to the next (germline) or are changed due to environmental influences, such as a person’s diet, exposure to pollutants, etc. The epigenome collectively is the set of chemical modifications to the DNA and DNA-associated proteins in the cell, which alter gene expression, and are heritable (via meiosis and mitosis). The modifications occur as a natural process of development and tissue differentiation, and can be altered in response to environmental exposures or disease.

The first type of mark, called DNA methylation, directly affects the DNA in a genome. In this process, proteins attach chemical tags called methyl groups to the bases of the DNA molecule in specific places. The methyl groups turn genes on or off by affecting interactions between the DNA and other proteins. The second kind of
mark, called histone modification, affects DNA indirectly. DNA in cells is wrapped around histone proteins, which form spool-like structures that enable DNA's very long molecules to be wound up neatly into chromosomes inside the cell nucleus. A variety of chemical tags can attach to histones. Other proteins in cells can detect these tags and determine whether that region of DNA should be used or ignored in that cell.

One of the hallmark examples of epigenetic modification is the X-inactivation. The inactive X chromosome (Xi) does not express the majority of its genes, unlike the active X chromosome (Xa). This is due to the silencing of the Xi by repressive heterochromatin, which compacts the Xi DNA and prevents the expression of most genes. Compared to the Xa, the Xi has high levels of DNA methylation, low levels of histone acetylation, low levels of histone H3 lysine-4 methylation, and high levels of histone H3 lysine-9 methylation and H3 lysine-27 methylation mark which is placed by the PRC2 complex recruited by Xist, all of which are associated with gene silencing (Chow et al., 2005). Additionally, a histone variant called macroH2A (H2AFY) is exclusively found on nucleosomes along the Xi (Costanzi and Pehrson, 1998).

Changes in the epigenome can switch on or off genes involved in cell growth or the immune response. These changes can lead to uncontrolled growth, a hallmark of cancer, or to a failure of the immune system to destroy tumors.

**Context-specific methylation**

Epigenetic marks include methylation of DNA at the cytosine residue of CpG dinucleotides and covalent modifications of amino acid residues within histone proteins that are responsible for the primary packaging of DNA. In the human genome, DNA methylation occurs almost exclusively at CpG dinucleotides. The cytosine residue of a CpG dinucleotide can be covalently modified by adding a methyl group to its carbon-5 atom resulting in 5-methylcytosine. The methyl group is transferred from S-adenosyl-l-methionine to a cytosine residue via DNA methyltransferases. CpG islands are often, but not exclusively, located at gene promoters, where the methylation status is generally correlated with transcriptional gene activity. Methylation of MGMT was shown to serves as a predictive biomarker for determining response of glioma and glioblastoma patients treated with the alkylating agent Temozolomide (Kim et al., 2016). Other DNA methylation based biomarkers include vimentin (VIM), septin 9 (SEPT9), and syndecan 2 (SDC2) for colorectal cancer, glutathione S-transferase pi 1 (GSTP1) for prostate cancer (Mikeska and Craig, 2014).

**Context-specific histone modification**

Covalent post-translational modifications of histone proteins are known to play an important role in chromatin remodeling and thereby in regulation of gene expression. It is majorly known to alter the local structural dynamics of chromatin to regulate the functioning of the genome, mostly by regulating its accessibility and
compactness. Disruption in the proper maintenance of these heritable epigenetic mechanisms can result in activation or inhibition of various critical cell-signaling pathways thus leading to disease states such as cancer. Epigenetic changes are reversible in nature and can be potentially restored back to their original state by epigenetic therapy. In cancer, histone H2A variants, H2A.1, H2A.Z and macroH2A have also been reported to express aberrantly (Monteiro et al., 2014). Also, histone proteins can undergo a variety of post transcription modifications some of which are methylation (me), acetylation (ac), ubiquitylation (ub), sumoylation (su) and phosphorylation (ph) on specific amino acid. Histones are also known to undergo homocysteinylation, crotonylation and glucosylation amongst others. The most well-known histone modification in cancer is the global loss of H4K16ac and H4K20me3. Loss of H3ac, H3K9me3 and H3S10ph is observed at the promoters of Sfrp2, Sfrp5 and Wnt5a in colon cancer (Khan et al., 2015).

1.7 High-throughput Data Generation Strategies

1.7.1 Microarray Technologies

“With close to gene-expression data from one million biological contexts in the public repositories, researchers can identify disease trends without ever having to enter a laboratory.” Monya Baker (Baker, 2012).

Although the excitement of NGS technologies and the promise of transcriptome sequencing, to both detect and discover novel genes, is becoming trendy, there is no disagreement on the usefulness of the millions of gene expression profiles in biomarker discovery. As rightly described by Monya Baker, perhaps, one can identify disease trends without entering a laboratory as demonstrated by researchers at Stanford (Kodama et al., 2012).

What are microarrays?

DNA microarrays are a collection of a large number of unique oligonucleotide, representing known genes, attached to a solid support, such as chip or glass slides, through specialized chemical bonds or non-bonding interactions. Although the principle of DNA microarray technology can be traced back to Southern Blotting techniques proposed by Edwin Southern in 1975 (Southern, 1975), it is not until the mid-1990s that the first DNA microarray, as we understand today, was designed and developed (Schena et al., 1995). The first generation of these arrays contained a large number of amplicons from PCR amplified genes spotted at specific locations on a glass slide through non-bonding interactions called cDNA arrays. With picomoles of identical DNA molecules on each spot, the spots acted like perfect baits/trap for the corresponding RNAs in a given sample. When these microarrays were washed with fluorescent labeled molecules, derived from mRNAs of a given biological contexts, the
corresponding baits lit-up indicating the level of expression of the respective RNA in the sample.

Based on the nature of the oligonucleotides and chemistry of attachment to the solid surface, several microarray companies emerged in the 1990s including Affymetrix, Illumina and Agilent, which fabricated some of the most widely used microarray platforms. It is fair to say that these companies made genome-wide gene expression profiling a common practice in genomics research and in clinical practice. Microarray technologies have also been widely used to detect SNPs (SNP arrays), measure exon level expression (exome arrays) and to identify transcriptionally active regions in the entire genome (titling array).

**Gene expression**

Microarray technology is the first to turn differential gene expression into a measurable parameter to describe biological contexts at systems level. Each cell in our body has the same genomic DNA. However, only a fraction of this is expressed in a given tissue/state at a given time. Since most human diseases are tissue specific, the change in expression levels of genes between normal and disease tissues has widely been used to reveal the disease status of a given tissue or to discover novel drug targets.

Genome-wide profiling of gene expression in samples from a large number of breast cancer patients has been useful in the identification of a molecular signature that can predict disease outcome and select candidates for adjuvant therapy (van ’t Veer et al., 2002). Taking advantage of the ability to quantitate gene expression, even change in the ratios of two genes has been useful to select patients for alternative treatment regimen (Ma et al., 2004).

There are currently three marketed assays, based on gene expression profiles, to predict treatment outcome in breast cancer (Marchionni et al., 2007). Of these efforts, the most reliable is the panel of 21 genes, 16 cancer and 5 controls, identified by Genomic Health that is being used to predict recurrence-free survival in patient treated for stage I and II breast cancer (Paik et al., 2004). Although these genes are discovered from genome-wide expression profiling of 25,000 human genes, the FDA approved test uses RT-PCR.

**SNP arrays**

There are now 50 million SNPs reported at the NCBI site (Sherry et al., 2001). Before NGS, the only technology to offer genome-wide SNP detection was microarray. Perhaps this scenario will continue (Gong et al., 2012). Many companies including Affymetrix and Illumina offer microarray platforms that can simultaneously detect 500,000 SNPs. These arrays have been used to derive signature that are predictive of clinical outcome (Tiu et al., 2009). A privately held personal genomics and biotechnology company called 23andMe (https://www.23andme.com/),
which offers autosomal DNA testing for ancestry and predisposition for more than 90 traits and conditions ranging from baldness to blindness using SNP genotyping.

**Exon/splice arrays**

Affymetrix also manufactured arrays containing a million feature representing majority of the known contiguously transcribed regions (exons) in the human genome. These are also called exome arrays. Since majority of genes are made up from 7 or 8 exons, these arrays helped identify exons that are missing in one sample compared to the other (Xi et al., 2008). This led to the discovery and measurement of splice variants of genes specific to a given context. Other types of microarray platforms used to measure splice variants used probes representing all known exon-exon junction of genes (Johnson et al., 2003) (Bingham et al., 2006).

**Tiling arrays**

Tiling arrays are one of the most ambitious applications of microarray technology. By fabricating microarrays that included oligonucleotides that are titled across the entire human genome, it was shown for the first time that much larger percentage of the human genome is actually transcribed than what would be the case if only protein coding regions are transcribed (Kampa et al., 2004). In Arabidopsis, 10 times more transcription activity was found using tiling arrays than was estimated by ESTs (Stolc et al., 2005). Tiling arrays are also useful for ChIP on chip experiments. DNAs released from immunoprecipitation can be hybridized to tiling arrays instead of an array with known probes; thus offering unbiased identification of transcription factor binding sites. This approach has been used to identify transcription factor binding sites in yeast, drosophila and human. Also, methyl-DNA immunoprecipitation followed by tiling array allowed DNA methylation mapping and measurement across the genome. In yet another application, by hybridizing the 1.5kb genome fragments from DNaseI digestion of the genome on to tiling arrays, one can discover sensitive regions of genome containing promoter signatures (Crawford et al., 2006).

**1.7.2 Next Generation Sequencing Technologies**

The affordability of generating genome-level sequencing by small laboratories is spawning diverse research applications limited only by the imaginations of the individual investigators. Some of the sequencing strategies used for various applications include de novo genome sequencing, genome resequencing, RNA sequencing, metagenomics, targeted sequencing, ChIP-seq and methyl-seq.
Figure 1.7 Application of Next Generation Sequencing

**Exome sequencing**

To keep the sequencing and analysis cost low, in some applications, only targeted regions of the genomes are sequenced. In exome sequencing only the exons from the entire genome are targeted and amplified for sequencing. By definition, exome sequencing is restricted for organisms with annotated reference genomes. Since the functional consequences of a non-synonymous (change in amino acids) variant can easily be deduced by sequencing only the coding regions (exons), exome sequencing approach is widely used in cancer to sift for driver mutations from millions of passenger mutations. For example, exome sequencing of head and neck cancer has revealed an inactivating mutation in NOTCH1 gene (Agrawal et al., 2011). Also, since exons constitute only a small fraction of the human genome, exome sequencing is a cost effective approach for finding causative mutations.

Roche, Illumina and Agilent both offer technologies for selective amplification of exons from human genome. Roche-Nimblegen offers SeqCap EZ technology to capture exomes and other customized regions of the genome (http://www.nimblegen.com/seqcap/). Agilent uses SureSelect, Illumina offers Nextera technology to select targeted regions of genomes including the human exome.

**Sequencing gene loci of interest**

Cancer is a genetic disease. In that, mutations that are advantageous to cancer are expected to be confined to oncogenes or tumor suppression factors. Although cancers of the same type can be heterogeneous, there is evidence to suggest that the
path to metastasis may be convergent. Drug ‘Sutent’ can be effective for treating leukemia and kidney cancer depending on the mutation profile in certain genes (Kolata, 2012). This being the case, it makes sense to sequence gene loci that are known to be implicated in cancer (Futreal et al., 2004).

Fluidigm offers a technology for selective amplification of a large number of gene loci simultaneously from a large number of samples (http://www.fluidigm.com/applications.html).

From the data analysis point of view, full genome sequencing can be safely divided into two major classes; de novo and resequencing. The main difference between the two strategies is the availability of a reference genome. While sequencing genomes for human diseases falls under the resequencing efforts, much of the plant, animal, microbe and insect genome sequencing efforts fall under the de novo category.

**De novo sequencing**

As the cost of sequencing become affordable for individual investigators, obtaining the genome sequence of the organisms of interest is no longer far-fetched. Many well-funded core sequencing facilities are sequencing the genomes of as many organisms as possible both to justify investment and to advance science. For lack of reference genomes of many species including zoo animals, plants and microbes, assembling the genomes using only short reads from NGS technologies remains a challenge. Multiple complementary platforms along with clever molecular biology techniques are used to assemble complex genomes from NGS data.

Widely used Illumina sequencing gives short reads, up to a few hundred base pairs in length. These short reads are typically not sufficient to assemble eukaryotic sequences at high contiguities. Currently two different technologies can generate long-read sequence data from single molecules at sufficient throughput. The sequencing technologies of Pacific Biosciences (PacBio) (Eid et al., 2009) and Oxford Nanopore produce reads of up to 20 kb on average, and though the reads of these technologies have high error rates of up to 15%, the accuracy of assembled sequences can be as accurate as the early gold standard reference genome sequences (Berlin et al., 2015; Koren and Phillippy, 2015; Quick et al., 2014). First assemblies of plant genomes exclusively based on PacBio sequences were published including assemblies of *Arabidopsis thaliana* with 38 contigs and an N50 of 11.2 Mb (Berlin et al., 2015) and of *Oropetium thomaeum*, a grass species having the smallest known grass genome of approximately 250 Mb, with 625 contigs and an N50 of 2.4 Mb (VanBuren et al., 2015).

Two recently introduced methods greatly improve the generation of scaffolding data with a promise to reconstruct entire chromosomes. (i) Optical mapping, was already invented at the end of last century (Schwartz et al., 1993), but
recent automation of this process has led to the development of commercial high-throughput platforms, such as the Irys system released by BioNano Genomics (Tang et al., 2015). Optical mapping generates fingerprints of DNA sequences of several hundred kb in size by imaging the locations of the restriction sites under light microscopes using fluorescent labels (Lam et al., 2012). Such individual fingerprints can be further assembled to construct genome-wide maps, which can then guide the order and orientation of sequence contigs. (ii) In 2015 Dovetail Genomics, introduced a technology called the Chicago (Putnam et al., 2016). This method is based on the Hi-C technology (sequencing of read pairs generated by proximity ligation of DNA in natural chromatin), but simplified this approach using in vitro reconstituted chromatin. Such data produces links between genomic regions that can be up to several hundred kb apart, and thus are useful for long-range scaffolding. Integration of such data has been shown to generate N50 values that can be as large as 30 Mb (Putnam et al., 2016).

**Genome resequencing**

The availability of high-quality reference human genome has spawned the era of medical sequencing. The genomes of individuals, sequenced using NGS, no longer require tedious assembly process for it to be useful in providing insight. In personalized medicine one is interested in identifying individual-specific variations, perhaps, suggesting a cause for a disease state or genetic predisposition. This can be achieved by simply mapping reads onto the reference genome. However, detecting individual-specific SNPs or SNVs from the large number of reads remain a challenge because these are buried among large number of sequencing errors. Elaborate statistical measures have been used to filter noise while identifying context-specific SNPs and SNVs (Wang et al., 2008b). It is not just the error in sequencing, but another major issue lies in calling for SNP’s using the current reference genome with the presence of minor alleles. The hg19 assembly is most well annotated reference assembly used in profiling variants. The underlying assumption being that hg19 harbors major alleles’ at all 3 billion positions. However, as more and more genomes of individuals from diverse ethnicity is being sequenced, such as the 1000 Genomes Project (1000 Genomes Project Consortium et al., 2015), it is becoming clear that significant positions on the hg19 reference assembly, which is the mosaic of genomes from six diverse individuals (Venter et al., 2001) do not represent major alleles at all three billion positions. This scenario is changing as genomes of more individuals are sequenced. In Chapter 6 we have shown that impact on SNP calling due to presence of minor alleles in hg19 and how one can improve it to have less false positive calls.

**RNA Seq**

The ability of NGS technologies to sequence the transcriptome (RNA Seq) of a given biological context is revolutionizing RNA research. Although this approach has been tried using conventional sequencing methods (EST sequencing), a window into
the entire RNA content within a biological context, would not have been possible without the spectacular advances in NGS technologies. With a single experiment, it is now possible to profile diverse RNA phenotypes to unprecedented depths expressed within a biological context. Thus, applications of RNA Seq include identification of transcript induced chimeras specific to cancer (Nacu et al., 2011), tissue-specific alternative splicing (Wang et al., 2008a), differential gene expression (Patil et al., April 25), tumor-specific translocations (Berger et al., 2010), allelic differences (Tuch et al., 2010), discovery of regulatory non-coding RNAs (Prensner et al., 2011a) and coding mutations.

As far as the NGS machines are concerned, transcriptome and genome sequencing are indistinguishable. To sequence a transcriptome, cDNAs representing cellular transcripts are used in place of genomic fragments to prepare the samples. The mapping and assembly strategies following NGS sequencing are also very similar to the genome sequencing discussed above. However, because of the diverse information content of cellular RNAs and because transcripts are stitched from discontinuous regions of the genome called exons, the bioinformatics strategy for RNA and genome sequencing diverge.

In plants and other organisms, when a reference genome is unknown, RNA Seq is often used to decipher all the genes that are differentially expressed between two biological contexts. For example, transcriptome from the leaves of a plant can be grown under various stress conditions to monitor molecular changes in response to stress (Délano-Frier et al., 2011). In these cases, short RNA reads need to be assembled into longer reads before the genes can be annotated or the two samples can be compared. Such an approach can also be used to construct full length context-specific transcripts even when a reference genome is unknown.

**Meta-genomics**

The size of the genome handled by NGS technologies is so large that it is no longer efficient to sequence individual microbes. Besides, to sequence individual microbes, each novel microbe needs to be identified and isolated, which is a laborious process. Small quantities of samples from the human gut (Qin et al., 2010a), soil (Riaz et al., 2008), and ocean water (Narasingarao et al., 2011) contains hundreds to thousands of distinct microbial species not yet isolated and characterized. Using NGS, one can generate large number of short DNA reads representing the collective genomes of all the microbes in the sample (Qin et al., 2010b) without the need to separate them. The genomes of individual organisms can then be separated and characterized using sophisticated computer-based assembly algorithms (Luo et al., 2012).
**ChIP Seq**

ChIP Seq technology is used to sequence regions of genomic DNA that binds to proteins. Again, this technology can be used only for organisms with highly annotated genomes. Transcription factor and other protein families are used to identify DNA binding motifs and their specific binding sites on the genome by chromatin immunoprecipitation. The process consists of cross-linking the proteins of interest with the DNA, digesting the unbound DNAs, and immunoprecipitating the bound complex using antibodies on a chip or magnetic bead. The RNA binding motifs for splicing factors could also be sequenced using this approach. This is called RIP Seq or CLIP Seq. ChIP Seq can be used in conjunction with genome or exome sequencing to arrive at context-specific inferences (Ozgül et al., 2011).

**Methyl Seq**

This approach is used to identify differential methylation states between two distinct biological contexts. The method involves digesting the chromosomal DNA with restriction enzymes HpaII and MspI. HpaII is a methyl sensitive enzyme, whereas MspI is a methyl insensitive enzyme. The combination can be used to first enrich the sample with hyper-methylated regions and to create fragments with one end of the sequence containing methylated sites. Applications of methyl Seq include identification of methylation pattern after stem cell differentiation (Brunner et al., 2009) or between tumor and normal tissues. As part of the ENCODE project, NGS data from a large number of methyl Seq experiments are available at the NCBI web site. One such data set, SRP000592, was generated to find the distinction between human embryonic stem cells and developing human fetal liver.
1.8 Algorithms in Big Biology

The deluge in data from NGS defied Moore’s law of growth in compute power, which predicted doubling of CPU power every 18 months. New and efficient algorithms are continuously required/developed to address and analyze huge datasets stemming from NGS.

1.8.1 Mapping short NGS reads to reference genome

Next-generation sequencing generally produces very short reads of 100-200 bases without any genomic context. Mapping short reads against a reference genome is typically the first step to assign chromosomal locus to these reads. For this one require to align/map these reads to the reference sequence. This means that one has to map tens to hundred million reads onto the reference genome. Unlike the previous generation tools, for the first time, the size of the query sequences are larger than the database; requiring indexing of not just the databases but of the query as well. This required novel algorithm to accomplish the task of mapping in a reasonable time with the existing computing power. On an average, a sequencing run produces millions of reads per sample. Traditional alignment algorithms such as BLAST or BLAT might take days or weeks to align this high number of reads onto a reference genome and are even computationally intensive. As a practical matter, the task of mapping billions of sequences to a genome calls for extraordinarily efficient algorithms, in which every bit of memory is used optimally or near optimally. New algorithms are built that exploit features of short DNA sequencing reads to map millions of reads per hour while minimizing both processing time and memory requirements.

Solution: Burrow Wheeler

Most of the current mapping/aligning tools uses a technique originally developed for compressing large files called the Burrows-Wheeler transform (Li and Durbin, 2009). Using this transform, the index for the reference genome fits into a small memory footprint in contrast to a spaced seed index, which may require a lot of gigabytes and yet reads can still be aligned efficiently. Reads are aligned one character at a time to the Burrows-Wheeler–transformed genome (Li and Durbin, 2010). Each successively aligned new character is followed by the removal of that character from the list. If the location where a read aligns perfectly cannot be found, the algorithm backtracks to a previous character of the read, makes a substitution and resumes the search. In effect, the Burrows-Wheeler transform enables to conquer the mapping problem by first solving a simple sub problem, align one character and then building on that solution to solve a slightly harder problem align two characters and then continuing on to three characters, and so on, until the entire read has been aligned.
Table 1-3 give the performance and sensitivity of Bowtie v0.9.6, SOAP v1.10, and Maq v0.6.6 when aligning 8.84 M reads (Langmead et al., 2009).

<table>
<thead>
<tr>
<th>Platform</th>
<th>CPU time</th>
<th>Wall clock time</th>
<th>Reads mapped per hour (millions)</th>
<th>Peak virtual memory footprint (megabytes)</th>
<th>Bowtie speed-up</th>
<th>Reads aligned (%)</th>
</tr>
</thead>
<tbody>
<tr>
<td>Bowtie v2 Server</td>
<td>15 m 7 s</td>
<td>15 m 41 s</td>
<td>33.8</td>
<td>1,149</td>
<td>-</td>
<td>67.4</td>
</tr>
<tr>
<td>SOAP</td>
<td>91 h 57 m 35 s</td>
<td>91 h 47 m 46 s</td>
<td>0.10</td>
<td>13,619</td>
<td>351×</td>
<td>67.3</td>
</tr>
<tr>
<td>Bowtie PC</td>
<td>16 m 41 s</td>
<td>17 m 57 s</td>
<td>29.5</td>
<td>1,353</td>
<td>-</td>
<td>71.9</td>
</tr>
<tr>
<td>Maq</td>
<td>17 h 46 m 35 s</td>
<td>17 h 53 m 7 s</td>
<td>0.49</td>
<td>804</td>
<td>59.8×</td>
<td>74.7</td>
</tr>
<tr>
<td>Bowtie Server</td>
<td>17 m 58 s</td>
<td>18 m 26 s</td>
<td>28.8</td>
<td>1,353</td>
<td>-</td>
<td>71.9</td>
</tr>
<tr>
<td>Maq</td>
<td>32 h 56 m 53 s</td>
<td>32 h 58 m 39 s</td>
<td>0.27</td>
<td>804</td>
<td>107×</td>
<td>74.7</td>
</tr>
</tbody>
</table>

Table 1-3 Bowtie alignment performance versus SOAP and Maq

**Challenges**

Another issue is sequencing errors or variations between the sequenced chromosomes and the reference genome. While mapping reads form a transcriptome experiment to a genome, problem of splice junction also is a major lacuna. As reads from an RNA Seq experiments will corresponds to exon-exon junction as well, such reads will mapping onto to a genome have to be split with large gaps in between them based on the length of the intron. This spliced mapping problem is further complicated by the possible presence of varied intron-sized gaps. To solve the splice junction issues we have certain splice-aligners, like TopHat (Trapnell et al., 2009). It uses BWT to identify exons that fully contain some of the reads, and then aligns the remaining reads to junctions between those exons.

**1.8.2 Assembling short NGS reads**

Genome assembly of an organism from short reads remains the major computational challenge. This is because billions of short reads need to be aligned to each other to find overlapping regions which are used in stitching up the genome together. This increase the scale of problem to billion by billion comparison requiring very high memory footprint.
The algorithm used to assembly genomes using long reads cannot be applied to short reads because of computational scale. Novel algorithms are required to address this problem. Solutions developed in graph theory has been applied to solve this problem. One such method is De Bruijn algorithm.

*de Bruijn*

The conventional methods used to assemble Sanger sequencing reads, which uses reads a nodes and overlap as bridges. The big shift provided by de bruijn algorithm is to convert overlap between reads as nodes and reads as bridges. This shift is critical because the number of reads are in billions compared to the overlaps, which depends on the size of the overlap. By controlling the size of the overlap one can fix the assembly problem on a smaller memory footprint. These overlaps are famously known as k-mer’s were K stands for the selected size of the overlaps. Thus De Bruijn graph offers an efficient way to represent a sequence in terms of its k-mer components. A de Bruijn graph is constructed from short reads, and then the genome is derived from the de Bruijn graph. The first step is to choose a k-mer size, and split the original sequence into its k-mer components. Then a directed graph is constructed by connecting pairs of k-mers with overlaps between the first k-1 nucleotides and the last k-1 nucleotides. The direction of arrow goes from the k-mer, whose last k-1 nucleotides are overlapping, to the k-mer, whose first k-1 nucleotides are overlapping (Compeau et al., 2011).

*Pain/challenges*

Since de Bruijn algorithm uses short read, repeating reads from distinct chromosomal locus cannot be distinguished. Due to short reads from repeat regions the concept of depth and breadth of sequencing gets merged. De Bruijn algorithm often treat the reads from repeat regions as redundant reads and compresses them during assembly; this remains one of the major drawback pertaining to short reads assembly.

While very short overlap increase the efficiency on the memory footprint of the algorithm. But very short overlap leads to branches and bubbles. For example, k-mer of 4 will allow a given reads to overlap with forming more than one optimal path for genome assembly.
Gene prediction is finding the location of protein coding regions on an assembled genome. Gene prediction in prokaryotic genomes is less difficult, due to the higher gene density and the absence of introns in their protein coding regions. In eukaryotic organisms it remains a challenge due to complex gene architecture. Two classes of methods are generally adopted for gene prediction includes homology based and ab initio prediction.

In homology based methods one compare the genome with existing protein databases to predict genes using conventional alignment methods like BLAST. This approach is based on the assumption that functional regions (exons) are more conserved evolutionarily than nonfunctional regions (intergenic or intronic regions). Once the similarity between a certain genomic regions is established, this information can be used to infer gene structure or function of that region. Local alignment and global alignment are two methods based on similarity searches. The most common local alignment tool is the BLAST family of programs, which detects sequence similarity to known genes. PROCRUSTES (Gelfand et al., 1996) and GeneWise (Birney and Durbin, 2000), uses global alignment of a homologous protein to translated ORFs in a genomic sequence for gene prediction.

*Ab initio* prediction relies on using gene structure as a template to annotate genes. *Ab initio* gene predictions rely on gene architecture information like splice sites, branch points, polypyrimidine tracts, start codons and stop codons, promoter site, TATA box. It also uses other information’s like codon usage that are unique to a species, and allow exons to be distinguished from the surrounding introns and intergenic regions.
Many algorithms are applied for modeling gene structure, such as Dynamic Programming, linear discriminant analysis, Linguist methods, Hidden Markov Model and Neural Network. Based on these models, a great number of ab initio gene prediction programs have been developed like GENESCAN (Burge and Karlin, 1997) and HMMGene (Krogh, 2000)

### 1.9 Data Repositories

One can now start answering biological questions without ever entering the lab. Both raw sequence data and meta data can be downloaded from a number of repositories. Following table give the most commonly used repositories for raw data.

<table>
<thead>
<tr>
<th>DNA DataBank of Japan (DDBJ)</th>
<th><a href="http://www.ddbj.nig.ac.jp">http://www.ddbj.nig.ac.jp</a></th>
<th>3.5 PB of raw data</th>
</tr>
</thead>
<tbody>
<tr>
<td>European Nucleotide Archive (ENA)</td>
<td><a href="http://www.ebi.ac.uk/ena">http://www.ebi.ac.uk/ena</a></td>
<td>1328.9 million assembled/annotated sequences available</td>
</tr>
<tr>
<td>EBI Metagenomics</td>
<td><a href="https://www.ebi.ac.uk/metagenomics/">https://www.ebi.ac.uk/metagenomics/</a></td>
<td>4037 metagenomes, 2975 16S rRNA amplicon datasets, 389 metatranscriptomes and 67 assemblies</td>
</tr>
<tr>
<td>NCBI Sequence Read Archive (SRA)</td>
<td><a href="http://www.ncbi.nlm.nih.gov/sra">www.ncbi.nlm.nih.gov/sra</a></td>
<td>Archives raw sequencing data and alignment information from high-throughput sequencing platforms, including Roche 454 GS System®, Illumina Genome Analyzer®, Applied Biosystems SOLID System®, Helicos Heliscope®, Complete Genomics®, and Pacific Biosciences SMRT®. As per March 2018, it has 6547681797553430 bases of open success data</td>
</tr>
</tbody>
</table>

*Table 1-4 NGS data repositories*
1.9.1 Meta/annotated data repositories

**TCGA**

The Cancer Genome Atlas is a collaboration between the National Cancer Institute (NCI) and National Human Genome Research Institute (NHGRI), and has generated comprehensive, multi-dimensional maps of the key genomic changes in 33 types of cancer. TCGA contains 2.5 petabytes of data describing tumor tissue and matched normal tissues from more than 11,000 patients. TCGA created a genomic, transcriptomics and epigenomic data analysis pipeline that can effectively collect, select, and analyze human tissues for genomic alterations, coding and non-coding RNA expression profiling and methylation status on a very large scale. This analyzed data is very useful for clinicians as they can just browse through to look at the say expression status for a particular gene in any cancer type, without a need for any specialized bioinformatics training.

**UCSC Browser**

The UCSC Genome Browser is an on-line genome browser. It is an interactive website offering access to genome sequence data from a variety of vertebrate and invertebrate species and major model organisms, integrated with a large collection of aligned annotations. The Browser is a graphical viewer optimized to support fast interactive performance and is an open-source, web-based tool suite built on top of a MySQL database for rapid visualization, examination, and querying of the data at many levels. UCSC browser accommodate genome sequences of all vertebrate species and selected invertebrates with a total of 46 species. The UCSC Genome Browser presents a diverse collection of annotation datasets (known as "tracks" and presented graphically), including mRNA alignments, mappings of DNA repeat elements, gene predictions, gene-expression data and disease-association data (representing the relationships of genes to diseases), dbSNP database, 1000 Genome project data, CNV, HapMap etc. The juxtaposition of the many types of data allow one to display exactly the combination of data that will answer specific questions for any position on the genome.

**NCBI**

The National Center for Biotechnology Information (NCBI) houses a series of databases relevant to biotechnology and biomedicine and is an important resource for bioinformatics tools and services. Major databases include GenBank for DNA sequences and PubMed, a bibliographic database for the biomedical literature.
The European Nucleotide Archive (ENA) provides a comprehensive record of the world's nucleotide sequencing information, covering raw sequencing data, sequence assembly information and functional annotation. It stores input data information (sample, experimental setup, machine configuration), output machine data (sequence traces, reads and quality scores) and interpreted information (assembly, mapping, functional annotation). One can find raw data, assembled sequences and annotation from small-scale sequencing efforts for multiple species and multiple conditions stored in EMBL.