Current Trends in Human Genome Sequencing and Data Analyses; 
A step towards personalized medicine

Dr. Asif Ullah Khan
Assistant Professor
Institute of Basic Medical Sciences (IBMS),
Khyber Medical University, PDA Building,
Block IV, Phase 5, Hayatabad,
Peshawar, Pakistan

Genome

• Life is specified by genomes (Entire DNA content of an organism)
• Genome include all the biological information require to build and maintain a living organism.
Genome Sequencing

- Genome sequencing is figuring out the order of DNA nucleotides, or bases, in a genome.
- Reading the blueprint of life's chemical alphabet.

Why from Gene To Genome Sequencing

No More Junk DNA
Human Genome Sequencing

- Human Genome Project began in October 1990.
- Complete human genome was acquired in 2003.
- Sequencing was performed in research centers of US, UK, Japan, France and Germany.
- The main goal was to understand the genetic make up of entire human genome.

Human Genome Project: Initial findings

- Approximately, 3 billion chemical base pairs make up human DNA
- Approximately, 30,000 genes were identified in human genome.
- Information were stored in databases.
The Human Genome Project (and others)

• Potential benefits

Molecular medicine
To explore the mysteries of human development and disease
– > Improved diagnosis of disease
– > Disease gene identification will lead to more accurate diagnosis
– > Earlier detection of genetic predispositions to disease
– > Will be able to assess risk for certain diseases, e.g. cancer, Type II diabetes, heart diseases
– > Rational drug design
– > Drugs designed to target specific gene products that cause diseases.
– > > Gene therapy
– > Replacement of defective genes for certain diseases
– > > Pharmacogenomics "custom drugs"
– > Deal with effect of genetic variations on drug efficacy and response

The Human Genome Project (and others)

• Potential benefits
  – Bioarchaeology, evolution, and human migrations
  • Our genomes preserve incredible ancient record of our ancestors that reveals human population sizes dating all the way back to before humans even existed.
  • Study migration of different population groups based on female genetic inheritance.
  • Study mutations on the Y chromosome to trace lineage and migration of males.
The Human Genome Project (and others)

- Potential benefits
  - DNA forensics (identification)
    - Identify potential suspects whose DNA may match evidence left at crime scenes.
    - Exonerate persons wrongly accused of crimes.
    - Establish paternity and other family relationships.

Genome Sequencing Strategies
Some basic Aspects: The Sanger Sequencing
(First Generation 1980)

- DNA is fragmented
- Cloned to a plasmid vector
- Cyclic sequencing reaction
- Separation by electrophoresis
- Readout with fluorescent tags

Sequencing Revolution in the form of Next Generation

2008  2012

Capillary electrophoresis based Sanger Sequencing

Next generation Sequencing
The Next Generation DNA sequencing

- Demand for faster, affordable DNA sequencing has led to the development of so-called "next generation" sequencing technologies.
- These technologies are delivering DNA sequencing at unprecedented speed, thereby enabling impressive scientific achievements and novel biological applications.
- To date, these technologies have been applied in a variety of contexts,
  - Whole-genome sequencing
  - Targeted resequencing
  - Transcriptome analysis
  - Discovery of transcription factor binding sites
  - Discoveries of small Non-coding RNAs.

Current Commercially available Next Generation DNA sequencing platform

- Roche 454 technology
- The Illumina/Solexa genome analyzer
- ABI SoliD
- Helicos tSMS Technology
Basic Principle of Next Generation DNA Sequencing Technologies

- Based on Sequencing by Synthesis principle
- Fragmenting DNA and adapter ligation.
- Sequence fragments 36 – 400 bp (sequence read)
- Map fragments to human reference sequence.
- Call DNA variants, i.e. SNPs, Indels, Structural Variations (Bioinformatics)

Sequencing Costs

The National Human Genome Research Institute (NHGRI)
http://www.genome.gov/sequencingcosts/
## Sequencing Landscape-2009

<table>
<thead>
<tr>
<th>Year</th>
<th>Sequence</th>
<th>Technology</th>
<th>Reagent Cost ($)</th>
<th>Runs</th>
<th>Coverage</th>
<th>Authors</th>
</tr>
</thead>
<tbody>
<tr>
<td>2001</td>
<td>Reference</td>
<td>Capillary (ABI)</td>
<td>300,000,000</td>
<td>4.0</td>
<td>251</td>
<td></td>
</tr>
<tr>
<td>2001</td>
<td>Reference</td>
<td>Capillary (ABI)</td>
<td>100,000,000</td>
<td>100,000</td>
<td>5.0</td>
<td>274</td>
</tr>
<tr>
<td>2007</td>
<td>C. Venter</td>
<td>Capillary (ABI)</td>
<td>10,000,000</td>
<td>100,000</td>
<td>7.5</td>
<td>31</td>
</tr>
<tr>
<td>2008</td>
<td>J. Watson</td>
<td>Roche (454)</td>
<td>2,000,000</td>
<td>234</td>
<td>7.4</td>
<td>27</td>
</tr>
<tr>
<td>2008</td>
<td>AML Patient</td>
<td>Illumina GA</td>
<td>1,000,000</td>
<td>98</td>
<td>33.0</td>
<td>48</td>
</tr>
<tr>
<td>2008</td>
<td>Y. Huang (CHB)</td>
<td>Illumina GA</td>
<td>500,000</td>
<td>35</td>
<td>36.0</td>
<td>77</td>
</tr>
<tr>
<td>2008</td>
<td>YRI (NA18507)</td>
<td>Illumina GA/Solid</td>
<td>250,000</td>
<td>40</td>
<td>40.6/17.9</td>
<td>196</td>
</tr>
<tr>
<td>2009</td>
<td>S-J Kim (KOR)</td>
<td>Illumina GA</td>
<td>NA</td>
<td>29.0</td>
<td>21</td>
<td></td>
</tr>
<tr>
<td>2009</td>
<td>AK1 (KOR)</td>
<td>Illumina GA</td>
<td>200,000</td>
<td>-</td>
<td>27.8</td>
<td>45</td>
</tr>
<tr>
<td>2009</td>
<td>S. Quake</td>
<td>Helicos</td>
<td>48,000</td>
<td>4</td>
<td>28.0</td>
<td>3</td>
</tr>
<tr>
<td>2009</td>
<td>NA07022</td>
<td>Complete Genomics</td>
<td>20,000</td>
<td>1</td>
<td>65.0</td>
<td>65</td>
</tr>
</tbody>
</table>

## Sequence Read Length of Available NGS Platforms

<table>
<thead>
<tr>
<th>Platform</th>
<th>Min read Length</th>
<th>Max read length</th>
</tr>
</thead>
<tbody>
<tr>
<td>454 Roche GS FLX Titanium</td>
<td>70</td>
<td>400</td>
</tr>
<tr>
<td>Illumina GA</td>
<td>30</td>
<td>81</td>
</tr>
<tr>
<td>Illumina GA II</td>
<td>26</td>
<td>160</td>
</tr>
<tr>
<td>Illumina HiSeq</td>
<td>50</td>
<td>102</td>
</tr>
<tr>
<td>ABI Solid System 2.0</td>
<td>25</td>
<td>35</td>
</tr>
<tr>
<td>ABI Solid System 2.5</td>
<td>50</td>
<td>50</td>
</tr>
<tr>
<td>Helicos (tSMS)</td>
<td>20</td>
<td>25</td>
</tr>
</tbody>
</table>
Sequencing Centres

![Sequencing Centres Diagram]

Two Days Workshop; May 14-15, 2012 "Bioinformatics in Medicine"
Venue: Institute of Basic Medical Sciences, Khyber Medical University Peshawar

Basic Bioinformatics Concept; Denovo Reads

Denovo assembling

- Sequence Reads
- Contigs
- Scaffolds

Two Days Workshop; May 14-15, 2012 "Bioinformatics in Medicine"
Venue: Institute of Basic Medical Sciences, Khyber Medical University Peshawar
Basic Bioinformatics Concept; Genome Mapping

Advance Bioinformatics

Denovo Read Assembling
- Velvet (Zerbino et al., 2008)
- EULER (Pevzner et al., 2001)
- ALLPATHS (Butler et al., 2008)
- SoapDenovo (Li et al., 2010)

MANY MORE!!!!!!

Reads Mapping/Alignment
- MAQ (Li H et al., 2008)
- SOAP3 (Liu et al., 2012)
- Bowtie (Mane et al., 2011)
- BFAST (Homer et al., 2009)

MANY MORE!!!!!!

Databases
- Ensembl genome Browser (Spudich et al., 2010)
- 1K genome Browser (http://browser.1000genomes.org/index.html)
- UCSC Databases (http://genome.ucsc.edu/)
- Personal DB
Basic Bioinformatics Analyses

SNPs in NGS

Indels in NGS

SVs/CNV from NGS

Bioinformatics Journals; Tools for particular applications

Oxford Journal, Bioinformatics

Briefings in Bioinformatics

Bioinformatics and Systems Biology

Two Days Workshop; May 14-15, 2012 "Bioinformatics in Medicine"
Venue: Institute of Basic Medical Sciences, Khyber Medical University Peshawar
PERSONAL HUMAN GENOMES

- **First Human genome was consensus mosaic sequence (2001-03)**
- **The Venter Genome (2007) (Sanger technology, 7.5x)**
  - 1.2 M novel variants compare to reference HG
  - 74% of these variants account for small indels and larger CNVs
  - 4,107 non-synonymous variants
  - Heterozygous variants in 10,208 genes.
  - Heterozygous mutations in genes associated with coronary artery disease, hypertension, and myocardial infarction (i.e. Because of his family history).
- **The Watson Genome (2008) (Roche 454, 7.4x)**
  - Showed significant number of SNPs, Indels & CNVs
  - Important aspect was the identification of more deletion events than insertions (i.e. 2.3:1 ratio)
  - Significant numbers of 300-350 bp indels
  - Most of coding indels were heterozygous
  - Genome showed 23 large benign CNVs (i.e. size range of 26 Kb-1.6 Mb)

**African Genomes (2008) (Illumina + ABI SoliD, 40.6x)**
- The African individual genome identified homozygous SNPs associated cancer susceptibility.
- Genome sequences of Southern-African indigenous groups showed an average difference of 1.2 nucleotides per one Kb (i.e. average inter-individual variation of 1.0 nucleotide per kilobase observed in European).
- Variants in the SLC24A5 gene associated with skin color and increased production of melanin were observed in African population.
- Interestingly, homozygosity for a VDR and ACTN3 alleles associated with increased bone mineral density, muscle power performance and sprint were found in majority of these individuals.
- Heterozygous for allele CLCNKB encoding a chloride channel, that has greater ability to reabsorb chloride ions from renal glomerulus (advantageous in the desert habitat).
PERSONAL HUMAN GENOMES

- **Asian Genome**
  - First Asian genome (YH, Han Chinese individual) (Illumina tech., 36.0×) published in 2008.
  - Heterozygous mutation in the *GJB2* gene responsible for autosomal recessive deafness (population prone to deafness).
  - Increased risk for Alzheimer diseases.
  - Two Korean individual (Illumina, 28.9× and 27.8×) genome revealed known SNPs associated with variable risk of developing certain types of cancer, diabetes, or Alzheimer disease.

---

1000 Genomes Project

1000 Genomes
A Deep Catalog of Human Genetic Variation
Sanger Institute 1 K genome Project

1000 Genomes Project

<table>
<thead>
<tr>
<th>Phase</th>
<th>Samples</th>
<th>Coverage</th>
</tr>
</thead>
<tbody>
<tr>
<td>Pilot 1</td>
<td>179 (3 population)</td>
<td>2-6X</td>
</tr>
<tr>
<td>Pilot 2</td>
<td>6 (2 trios)</td>
<td>30-60X</td>
</tr>
<tr>
<td>Pilot 3</td>
<td>8,140 (exons) 906 (genes) (697 individuals) (3 major population)</td>
<td>50X</td>
</tr>
</tbody>
</table>

Characterize all common human genetic variation with a minor allele frequency (MAF) ≥ 0.5%.

1 K Genomes Browser Home page

(http://browser.1000genomes.org/index.html)
1 K Genomes Browser; Transcripts detail of Genes

1 K Genomes Browser; Variations detail in genes
Identification of damaging non-damaging SNPs

1 K Genomes Browser; Allelic Frequencies
Lessen from available Human Genomes


- The human genome is highly variable.
  - Each personal genome differs from the reference human assembly in \( \sim 3.5 \) million SNPs and \( \sim 1000 \) large (>500 bp) SVs.
- On average, individual genome contain 20,000–25,000 coding variants, of which 9,000–11,000 are non-synonymous and a slightly higher number are synonymous.
- SNPs call at homozygous position require 15x coverage depth while at heterozygous position require 30x.
- SNPs are more frequent in autosomes than in the sex chromosomes.
- Bias SNPs and Indels occurrences;
  - Occurrence is enriched in the first and last exons of genes.
  - Favoring multiples of three indels in order not to disrupt the reading frame.

- Enrich non-synonymous SNPs occurrence in some genes;
  - Genes with functions associated with environmental adaptation, such as those involved in sensory functions (e.g., olfactory and taste receptors) or immunological functions and signal transduction (e.g., GPCRs) seem to be enriched for non-synonymous SNPs. For example, it is well recognized that some of the genes that vary the most in humans are those for olfactory receptors.

- Large repetitive elements and high CNVs in human genome
  - In each genome sequenced, there have been megabases of DNA sequence that cannot be mapped to the reference genome assembly. This sequence is enriched for repeated elements but also contains functional elements including genes, many of which are known to be relevant to environmental perception and adaptation.
Personalized medicine

- Personalized medicine is a medical model that implies the patient’s genetic makeup information to make individualized treatment decisions.
- A paradigm shift in Healthcare
- The goal of personalized medicine is to treat each patient with the best possible therapy.

“Promise” of Personalized Medicine

- Promises of Personal Genomics:
  - Genetic Diagnosis
  - Accurate Disease Prediction
  - Disease Monitoring
  - Personalized Treatment
  - Gene Environment Interactions
  - Understand Biology
- Potential for Abuse
  - Discrimination
    - Sexual
    - Racial
    - Physical
    - Intelligence/aptitude
  - Suitability for employment
  - Suitability for insurance
Pediatric Tumors Made Personal

A mixed collection of relatively rare but often deadly pediatric tumors are collectively known as small round blue cell tumors (SRBCT) for precisely the reason one might imagine. Examined under a microscope after routine processing, bone marrow biopsies from cancers including neuroblastoma, Ewing sarcoma, rhabdomyosarcoma, and lymphoma appear as small, blue, and round cells. Despite some distinguishing molecular markers to guide them, oncologists can, on occasion, find it hard to diagnose these tumors specifically. Javed Khan, M.D., Head of the Oncogenics Section of CCR’s Pediatric Oncology Branch, has been using genomic approaches to study pediatric cancers for several years. He is now poised to launch an ambitious multicenter project to use comprehensive genomic data to guide the individualized treatment of children with advanced solid tumors.

Javed Khan et al., CCR connections, Volume 4, No. 1, 2010

Two Days Workshop; May 14-15, 2012 “Bioinformatics in Medicine”
Venue: Institute of Basic Medical Sciences, Khyber Medical University Peshawar
The Future
Developing of electronic molecular analysis system

Nanopore Sequencing
Oxford Nanopore Technologies Ltd.
http://www.nanoporetech.com/
- Minion Oxford Nanotechnology 2012

Two Days Workshop; May 14-15, 2012 “Bioinformatics in Medicine”
Venue: Institute of Basic Medical Sciences, Khyber Medical University Peshawar

Conclusion

- Development of next generation DNA sequencing technologies have greatly facilitated the rapid acquisition of complete human genome and transcriptome (exome) sequencing in a cost effective way.

- Personal genome sequencing are providing genetic basis for variability in drug efficacy and toxicity in different population/individuals and it will eventually become an instrument of common medical practice.

- Understanding of personal genetic makeup & genome profile of patients via genome sequencing & bioinformatics analyses are providing backbone information towards differential diagnosis and therapies.
Concluding aspects regarding Bioinformatics

- A hub of bioinformatics is involved in designing tools,
  - To compare billions of nucleotides of human genome (i.e. genome mapping) in order to identify specific genomic alterations.
  - To assemble millions of genome reads into contigs/scaffolds (i.e. denovo assembling)
  - To identify different type of genomic alterations i.e. mutation/polyorphism (homo, hetero, synonymous, non-synonymous, damaging, non- damaging) in different genomic loci (i.e. regulatory (CRE, non-coding RNAs), exons, introns and intergenic spacers regions).
- To identify the effect of particular mutations on 3D structures of proteins (i.e. Molecular Protein Modeling).
- To examine ligands & proteins/enzymes interaction aspects (i.e. Molecular Docking, simulation and drug discovery programs).

References

Our M.Phil/PhD Group Members

An effort
Dedicated to
All Pakistani
Scientists &
Researchers!!!

LONG LIVE PAKISTAN
Thanks ......