Current Trends in Human Genome Sequencing and Data Analyses;

A step towards personalized medicine

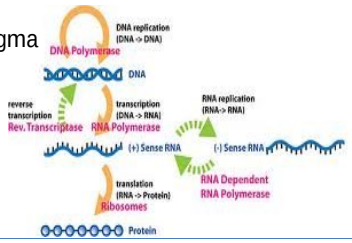
Dr. Asif Ullah Khan

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.

Central Dogma of Life



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)

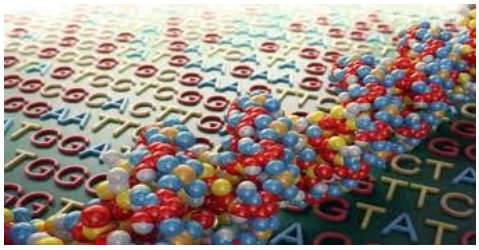
- Potential benefits
  - Bioarchaeology, evolution, and human migrations
  - Our genomes preserve incredible ancient record of our ancestors that reveals human population sizes dating all the way back to before humans even existed.
    - Study migration of different population groups based on female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

The Human Genome Project (and others)

- Potential benefits

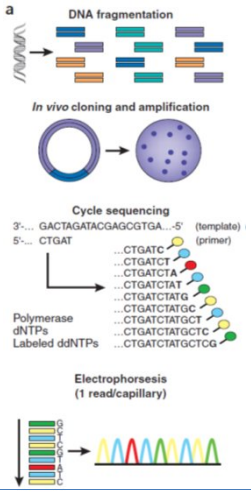
  DNA forensics (identification)
  - Identify potential suspects whose DNA may match evidence left at crime scenes.
  - Exonerate persons wrongly accused of crimes.
  - Establish paternity and other family relationships.

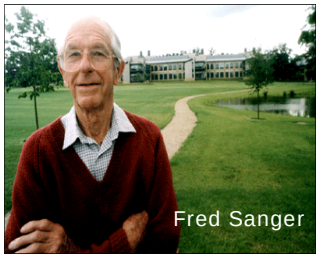
Genome Sequencing Strategies



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





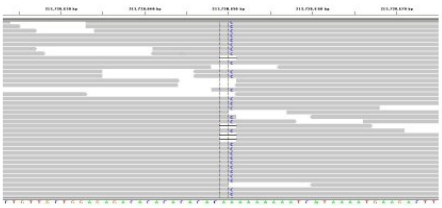
The Nobel Prize in Chemistry 1980

Sequencing Revolution in the form of Next Generation

2008

Capillary electrophoresis based Sanger Sequencing

2012



Next generation Sequencing

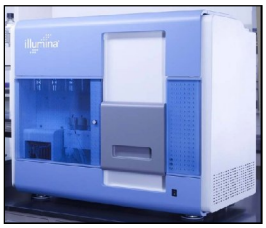
The Next Generation DNA sequencing

- 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

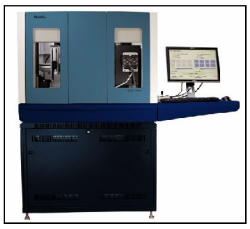
  &
  - Discoveries of small Non-coding RNAs.

Current Commercially available Next Generation DNA sequencing platform

Roche 454 technology



The Illumina/Solexa genome analyzer



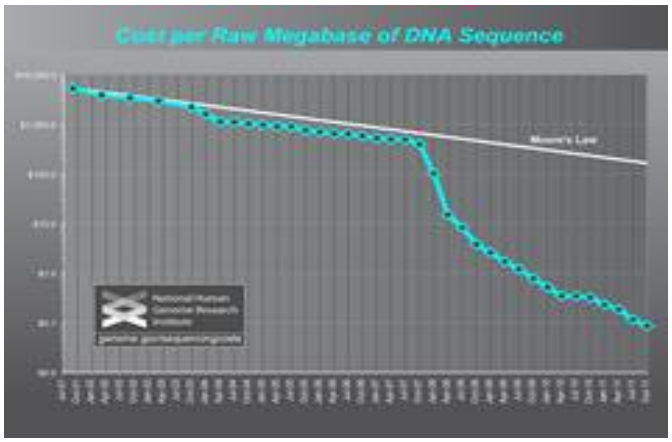
ABI SoliD

Helicos tSMS Technology

Basic Principle of Next Generation DNA Sequencing Technologies

- Based on Sequencing by Synthesis principle
- Fragmenting DNA and adapator ligation.
- Sequence fragments 36 – 400 bp (séquence 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

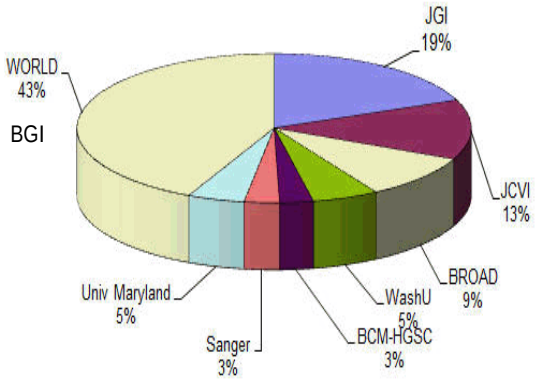
Year	Sequence	Technology	Reagent Cost ($)	Runs	Coverage	Authors
2001	Reference	Capillary (ABI)	300,000,000		4.0	251
2001	Reference	Capillary (ABI)	100,000,000	100,000	5.0	274
2007	C. Venter	Capillary (ABI)	10,000,000	100,000	7.5	31
2008	J. Watson	Roche (454)	2,000,000	234	7.4	27
2008	AML Patient	Illumina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

17

Sequence Read Length of Available NGS Platforms

Platform	Min read Length	Max read length
454 Roche GS FLX Titanium	70	400
Illumina GA	30	81
Illumina GA II	26	160
Illumina HiSeq	50	102
ABI Solid System 2.0	25	35
ABI Solid System 2.5	50	50
Helicos (tSMS)	20	25

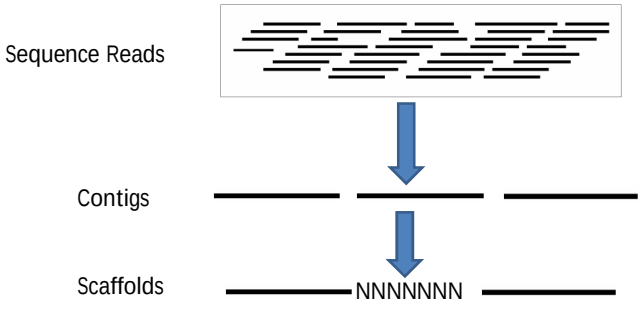
Sequencing Centres



Two Days Workshop; May 14-15 , 2012 "Bioinformatics in Medicine"
Venue: Institute of Basic Medical Sciences, Khyber Medical University Peshawar 19

Basic Bioinformatics Concept; Denovo Reads Assembling

Denovo assembling



Two Days Workshop; May 14-15 , 2012 "Bioinformatics in Medicine"
Venue: Institute of Basic Medical Sciences, Khyber Medical University Peshawar 20

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)

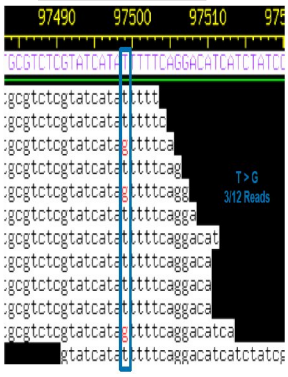
MANY MORE!!!!!!!

Databases

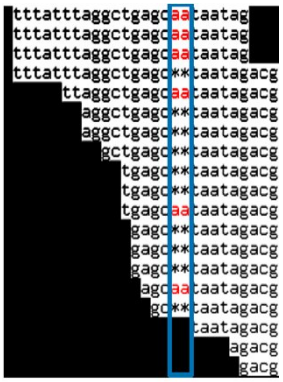
- Ensamble genome Browser (Spudich et al., 2010)
- 1 K genome Browser (http://browser.1000genomes.org/index.html)
- UCGC Databases (http://genome.ucsc.edu/)
- Personal DB

Basic Bioinformatics Analyses

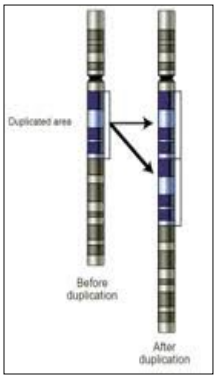
SNPs in NGS



Indels in NGS



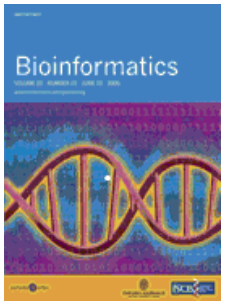
SVs/CNV from NGS



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

23

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

24

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)

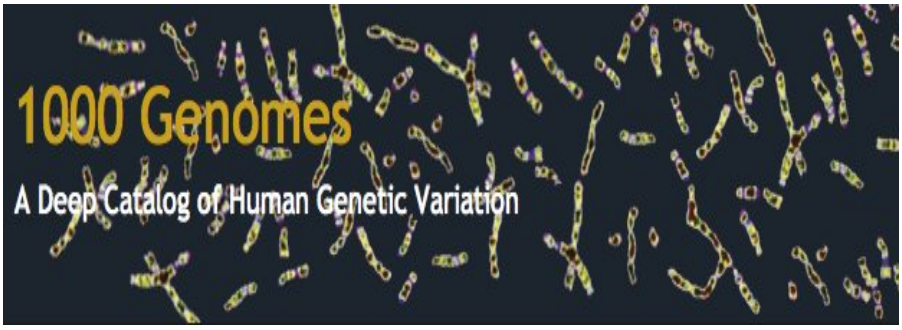
PERSONAL HUMAN GENOMES

- African Genomes (2008) (Illumina + ABI SoliD, 40.6×)
  - The African individual genome Identified homozygous SNPs associated cancer susceptibility.
  - Genome sequences of Southern-African indigenous groups showed 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).

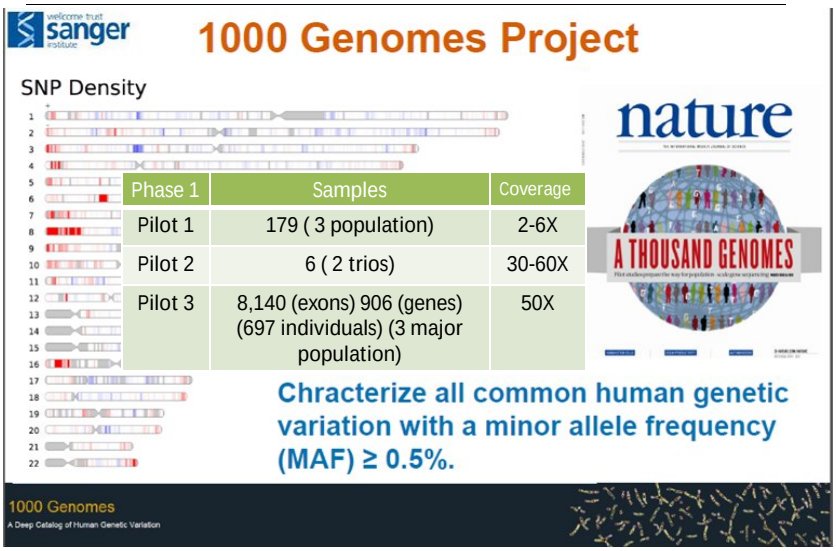
PERSONAL HUMAN GENOMES

- 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 reveled known SNPs associated with variable risk of developing certain types of cancer, diabetes, or Alzheimer disease.

1000 Genomes Project



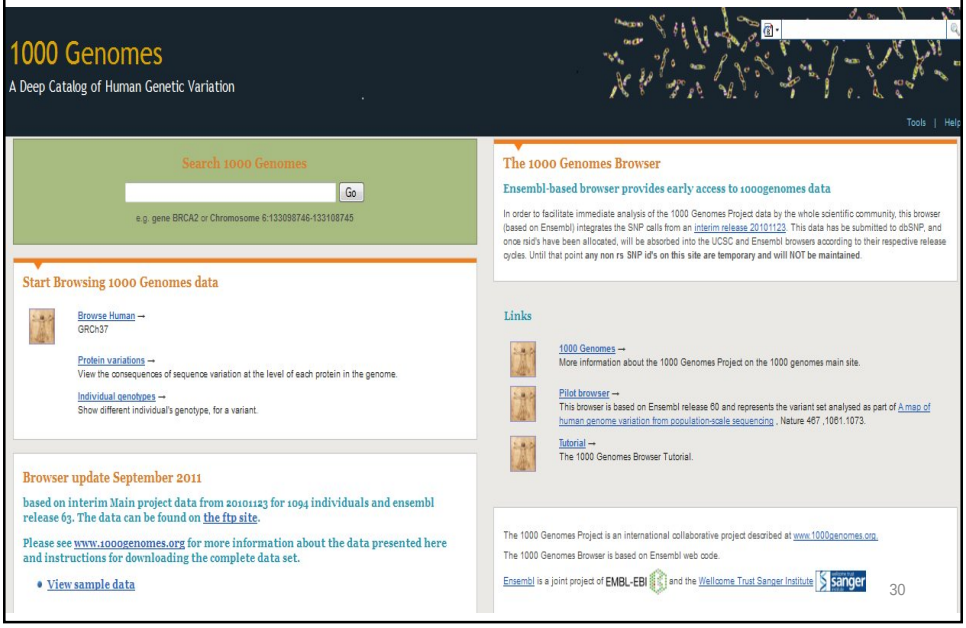
Sanger Institute 1 K genome Project



Phase 1	Samples	Coverage
Pilot 1	179 (3 population)	2-6X
Pilot 2	6 (2 trios)	30-60X
Pilot 3	8,140 (exons) 906 (genes) (697 individuals) (3 major population)	50X

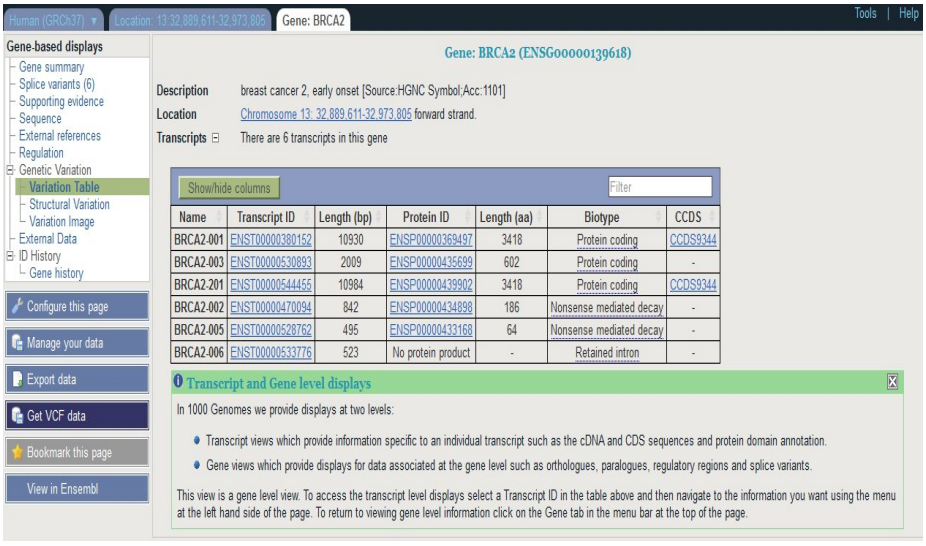
29

1 K Genomes Browser Home page
(http://browser.1000genomes.org/index.html)



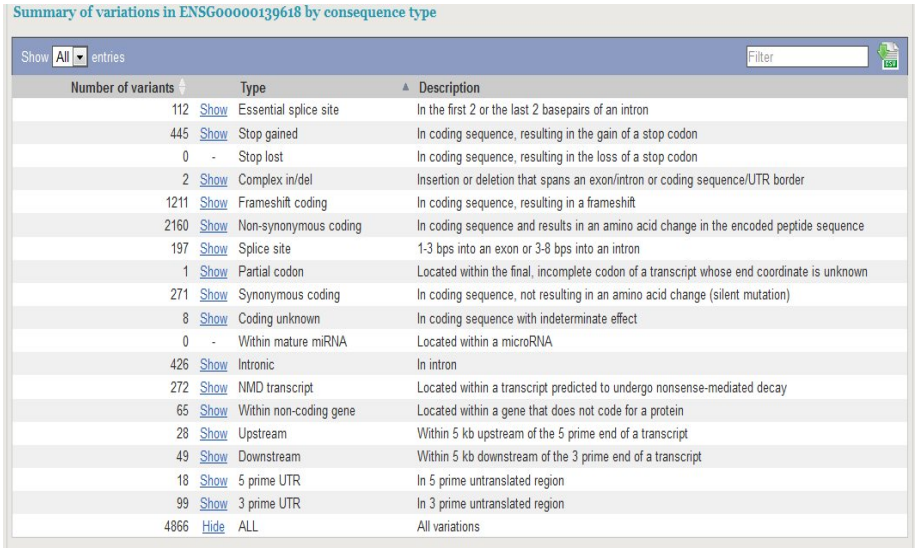
30

1 K Genomes Browser; Transcripts detail of Genes



31

1 K Genomes Browser; Variations detail in genes



32

Identification of damaging non-damaging SNPs

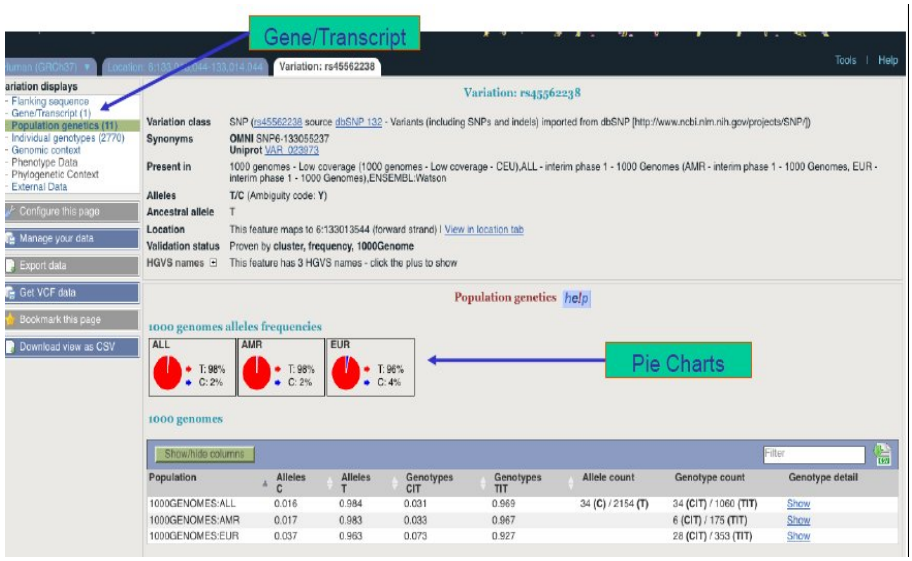
ALL variants

[back to top]

Show All entries | Show/hide columns | Filter

ID	Chr: bp	Alleles,	HGVS name(s)	Class	Source	Validation	Type	Amino Acid	AA co-ordinate	SIFT	PolyPhen	Transcript
						1000Genome						
rs76874770	13:32890587	C/T	ENST00000530893.1:c.-11C>T	SNP	dbSNP	cluster, 1000Genome	5 prime UTR	-	-	-	-	ENST00000530893
rs76874770	13:32890587	C/T	ENST00000544455.1:c.-11C>T	SNP	dbSNP	cluster, 1000Genome	5 prime UTR	-	-	-	-	ENST00000544455
rs80358547	13:32890599	T/G	ENST00000380152.3:c.2T>G ENSP00000369497.3:p.Met1?	SNP	dbSNP	precious	Non-synonymous coding	M/R	1 (2)	deleterious	probably damaging	ENST00000380152
rs80358547	13:32890599	T/G	ENST00000530893.1:c.2T>G ENSP00000435699.1:p.Met1?	SNP	dbSNP	precious	Non-synonymous coding	M/R	1 (2)	deleterious	probably damaging	ENST00000530893
rs80358547	13:32890599	T/G	ENST00000544455.1:c.2T>G ENSP00000439902.1:p.Met1?	SNP	dbSNP	precious	Non-synonymous coding	M/R	1 (2)	deleterious	probably damaging	ENST00000544455
rs80359418	13:32890600	G/-	ENST00000380152.3:c.3delG ENSP00000369497.3:p.Met1?	deletion	dbSNP	precious	Frameshift coding		1 (3)	-	-	ENST00000380152
rs80358650	13:32890600	G/A/T	ENST00000380152.3:c.3G>A ENSP00000369497.3:p.Met1?	SNP	dbSNP	cluster, precious	Non-synonymous coding	M/I	1 (3)	deleterious	probably damaging	ENST00000380152
rs80358650	13:32890600	G/A/T	ENST00000380152.3:c.3G>T ENSP00000369497.3:p.Met1?	SNP	dbSNP	cluster, precious	Non-synonymous coding	M/I	1 (3)	deleterious	probably damaging	ENST00000380152
rs80359418	13:32890600	G/-	ENST00000530893.1:c.3delG ENSP00000435699.1:p.Met1?	deletion	dbSNP	precious	Frameshift coding		1 (3)	-	-	ENST00000530893
rs80358650	13:32890600	G/A/T	ENST00000530893.1:c.3G>A ENSP00000435699.1:p.Met1?	SNP	dbSNP	cluster, precious	Non-synonymous coding	M/I	1 (3)	deleterious	probably damaging	ENST00000530893
rs80358650	13:32890600	G/A/T	ENST00000530893.1:c.3G>T ENSP00000435699.1:p.Met1?	SNP	dbSNP	cluster, precious	Non-synonymous coding	M/I	1 (3)	deleterious	probably damaging	ENST00000530893
rs80359418	13:32890600	G/-	ENST00000544455.1:c.3delG ENSP00000439902.1:p.Met1?	deletion	dbSNP	precious	Frameshift coding		1 (3)	-	-	ENST00000544455
rs80358650	13:32890600	G/A/T	ENST00000544455.1:c.3G>A ENSP00000439902.1:p.Met1?	SNP	dbSNP	cluster, precious	Non-synonymous coding	M/I	1 (3)	deleterious	probably damaging	ENST00000544455

1 K Genomes Browser; Allelic Frequencies



34

Lessen from available Human Genomes

[Human Genome Sequencing in Health and Disease; *Annu. Rev. Med. 2012. 63:35–61*]

- The human genome is highly variable.
- Each personal genome differs from the reference human assembly in ~3.5 million SNPs and ~ 1000 large (*>500 bp) SVs.*
- On average, individual genome contain20,000–25,000 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.

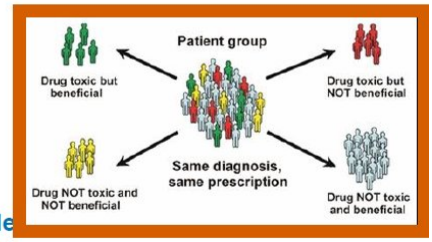
- Enrich non-synonymous SNPs occurrence in some genes;
- Genes with functions associated with environmental adaptation, such as those involved in sensory functions (e.g., olfactory and taste receptors) or immunological functions and signal transduction (e.g., GPCRs) seem to be enriched for non-synonymous 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



A News from Center of Cancer Research (CCR)

National Cancer Institute, NIH, Bethesda, Maryland.

Pediatric Tumors Made Personal

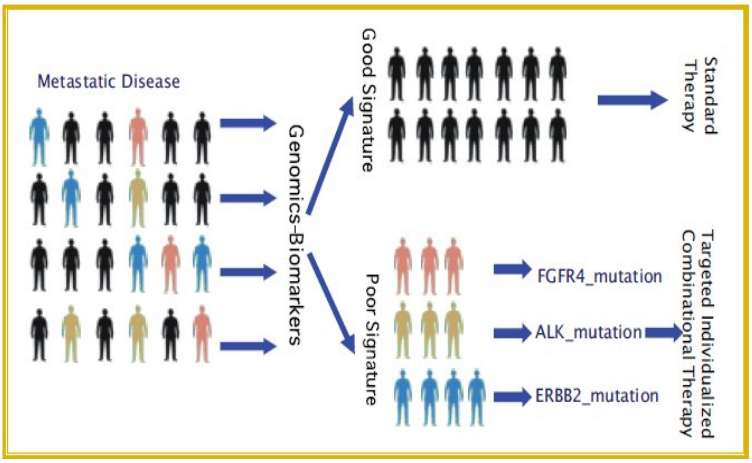
A mixed collection of relatively rare but often deadly pediatric tumors are collectively known as small round blue cell tumors (SRBCT) for precisely the reason one might imagine. Examined under a microscope after routine processing, bone marrow biopsies from cancers including neuroblastoma, Ewing sarcoma, 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 Oncogenomics 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.

Tapping Gene Expression

Khan is a strong believer in the power pediatric solid tumors, we were one of the first to use microarrays to find can use the rules it learns to predict new cases.

CCR connections, 4, (1) 2010.

Pediatric Tumors Made Personal



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

The Future

Developing of electronic molecular analysis system

Nanopore Sequencing

Oxford Nanopore Technologies Ltd.

http://www.nanoporetech.com/

➢ Minion Oxford Nanotechnology 2012

"I.B.M. Joins Pursuit of $1,000 Personal Genome"

New York Times October 5, 2009.

Conclusion

- Development of next generation DNA sequencing technologies have greatly facilitated the rapid acquirement of complete human genome and trancriptome (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/polymorphisim (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

1. Zerbino DR, Birney E. Velvet: algorithms for de novo short read assembly using de Bruijn graphs. Genome Res 2008;18:821–9.
2. Pevzner PA, Tang H. Fragment assembly with double-barreled data. Bioinformatics 2001;17(Suppl 1):S225–33.
3. Butler J, MacCallum I, Kleber M, et al. ALLPATHS: de novo assembly of whole- genome shotgun microreads. Genome Res 2008;18:810–20.
4. Li R, Zhu H, Ruan J, Qian W, Fang X, Shi Z, Li Y, Li S, Shan G, Kristiansen K, Li S, Yang H, Wang J, Wang J (2010) De novo assembly of human genomes with massively parallel short read sequencing. *Genome Res.*, 20(2): 265-272.
5. Mane SP, Modise T, Sobral BW (2011) Analysis of high-throughput sequencing data. Methods Mol Biol. 2011;678:1-11.
6. Homer N, Merriman B, Nelson SF. BFAST: an alignment tool for large scale genome resequencing. PLoS One. 2009 Nov 11;4(11):e7767.
7. Li H, Ruan J, Durbin R (2008) Mapping short DNA sequencing reads and calling variants using mapping quality scores. Genome Res 18: 1851–1858.
8. Liu CM, Wong T, Wu E, Luo R, Yiu SM, Li Y, Wang B, Yu C, Chu X, Zhao K, Li R, Lam TW (2012) SOAP3: ultra-fast GPU-based parallel alignment tool for short reads. Bioinformatics. 28(6):878-9.
9. Spudich GM, Fernández-Suárez XM. Touring Ensembl: a practical guide to genome browsing. BMC Genomics. 2010 May 11;11:295.
10. Javed Khan et al. Pediatric Tumors Made Personal. CCR *connections, 4, (1) 2010.*
11. Mardis ER. 2008. The impact of next-generation sequencing technology on genetics. Trends in Genetics 24:133-141.
12. Gonzaga-Jauregui C, Lupski JR, Gibbs RA (2012) Human genome sequencing in health and disease. Annu Rev Med. 63:35-61.

Our M.Phil/PhDGroup Members

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

LONG LIVE PAKISTAN

46

Thanks……..