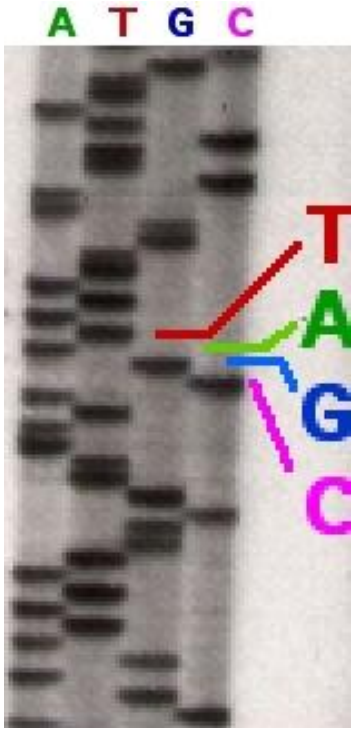
NGS applications and databases

Alla L Lapidus, Ph.D.

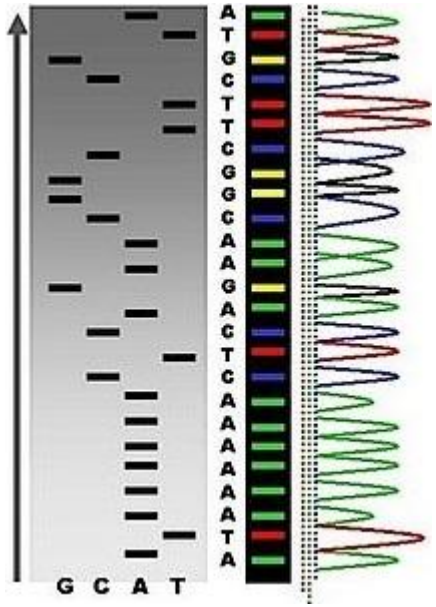
SPbAU, SPbSU,

St. Petersburg

Some history



Walter Gilbert and Allan Maxam at Harvard also developed sequencing methods. One of them is "DNA sequencing by chemical degradation“ – published in 1977. (radioactive, no cloning, purify fragment to be sequenced)



Frederick Sanger (MRC Centre, Cambridge, UK) - "DNA sequencing with chain-terminating inhibitors“ published in 1977 – radioactive but less toxic => was chosen and automated

Sanger vs NGS (1)

‘Sanger sequencing’ has been the only DNA sequencing method for 30 years but the need in greater sequencing throughput and more economical sequencing technology was obvious.

"REVOLUTIONARY GENOME SEQUENCING TECHNOLOGIES THE $1000 GENOME"

(Department of Health and Human Services (DHHS))

2004 - develop novel technologies that will enable extremely low-cost, high quality DNA sequencing

2009 - the cost to sequence an entire individual human genome to be $1,000 by the end of 2009 and the time required for sequencing less than one week

we are not their yet but very close – it is about $5000 these days

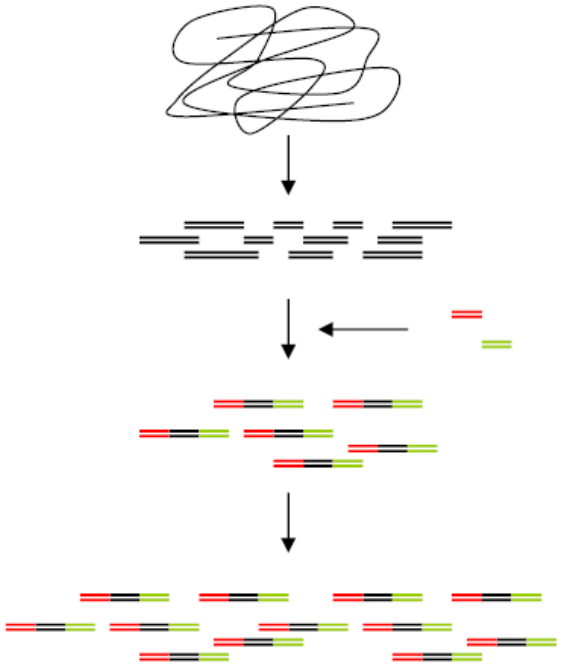
2012 - The NIH awarded $5.7 million in funding for research projects that explore ways to use genome sequencing in clinical care, and $800,000 to fund a coordinating center to support these studies.

Sequencing Technology at a Glance

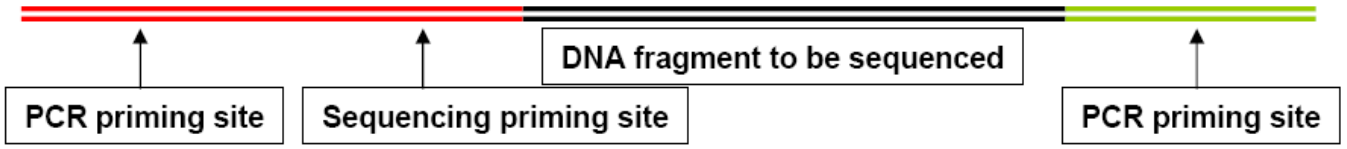
	Sanger	454/Roche	Solexa/Illumina	AB SOLiD	Helicos	Pacific Biosciences
Template preparation	Subcloning, DNA isolation from individual clone	DNA amplified by bead-based emulsion PCR	DNA amplified by solid phase bridge amplification	DNA amplified by bead-based emulsion PCR	Single molecule	Single molecule
Sequencing mechanism	Sequencing-by-synthesis on plate, Synthesis randomly terminated by ddNTP	Pyrosequencing DNA synthesis: one type of base at a time	in situ Sequencing-by-synthesis, DNA synthesis: one base at a time, Reversible terminator enables extension after detection	Sequencing-by-ligation	in situ Sequencing-by-synthesis, DNA synthesis: one base at a time, Reversible terminator enables extension after detection	in situ Sequencing-by-synthesis, DNA synthesis: continuous
Detection method	Detection by separation of fragments by size on instrument	Detection by measuring emitted light during sequencing reaction	Detection by excitation/emission of the label of incorporated nucleotide	Four-color two-base encoding system enables measurement error correction	Detection by excitation/emission of the label of incorporated nucleotide	Wave-guide technology. Real time detection while polymerase incorporates nt.

Library Preparation - new

- **DNA fragmentation using various methods**
  - Different platform requires different DNA size
  - 454: ~600 bp; Solexa: ~250 bp (Illumina)
- **Adaptor (ds) ligation**
  - no conventional cloning vector
- **Size selection**
- **Library amplification by PCR**
  - 454 does no need this
- **Quality assessment and quantification**



Final library fragment structure



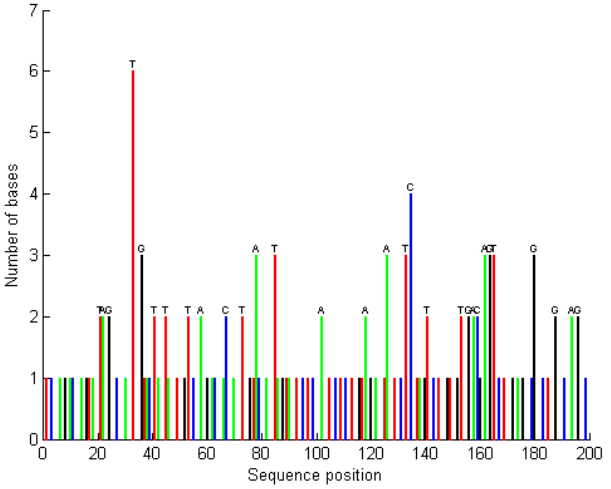
(Not for PacBio)

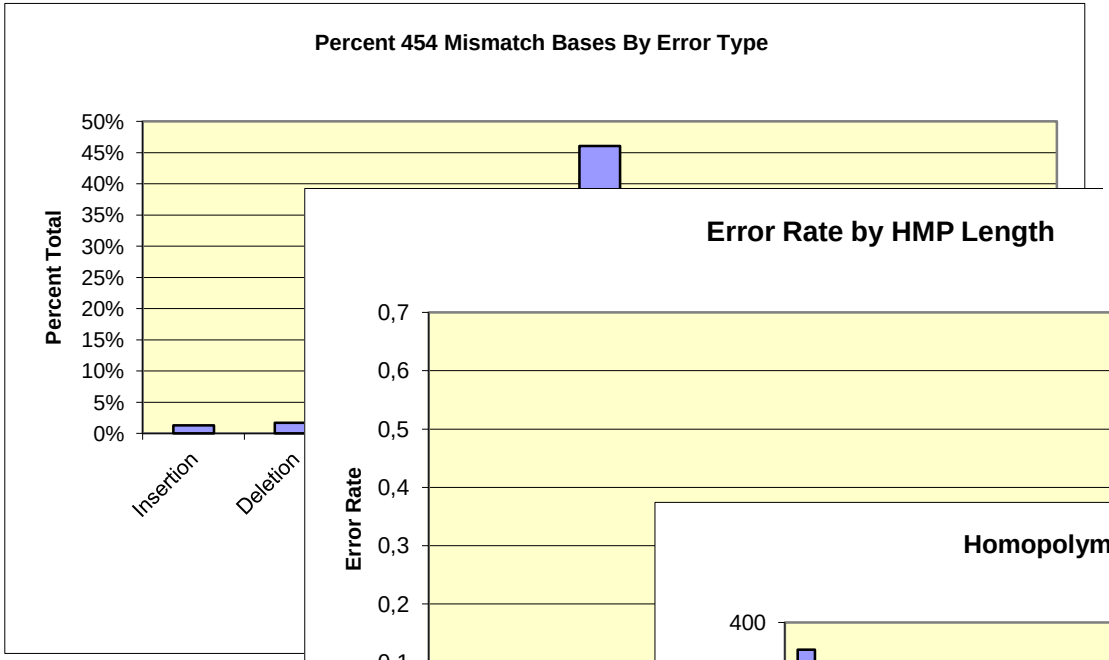
Sanger vs NGS.2

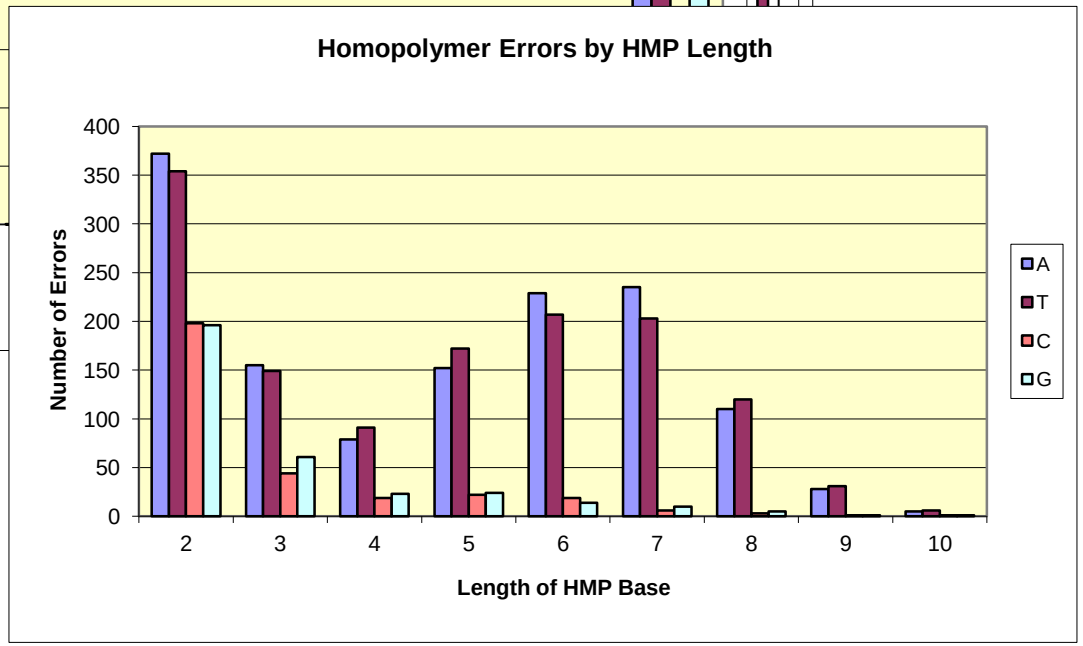
- NGS has the ability to process millions of sequence reads in parallel rather than 96 or 384 at a time (1/6 of the cost or even less)
- No clonning bias

Objections: fidelity, read length, infrastructure cost, handle large volum of data, need in the bioinformatics support

454 data quality



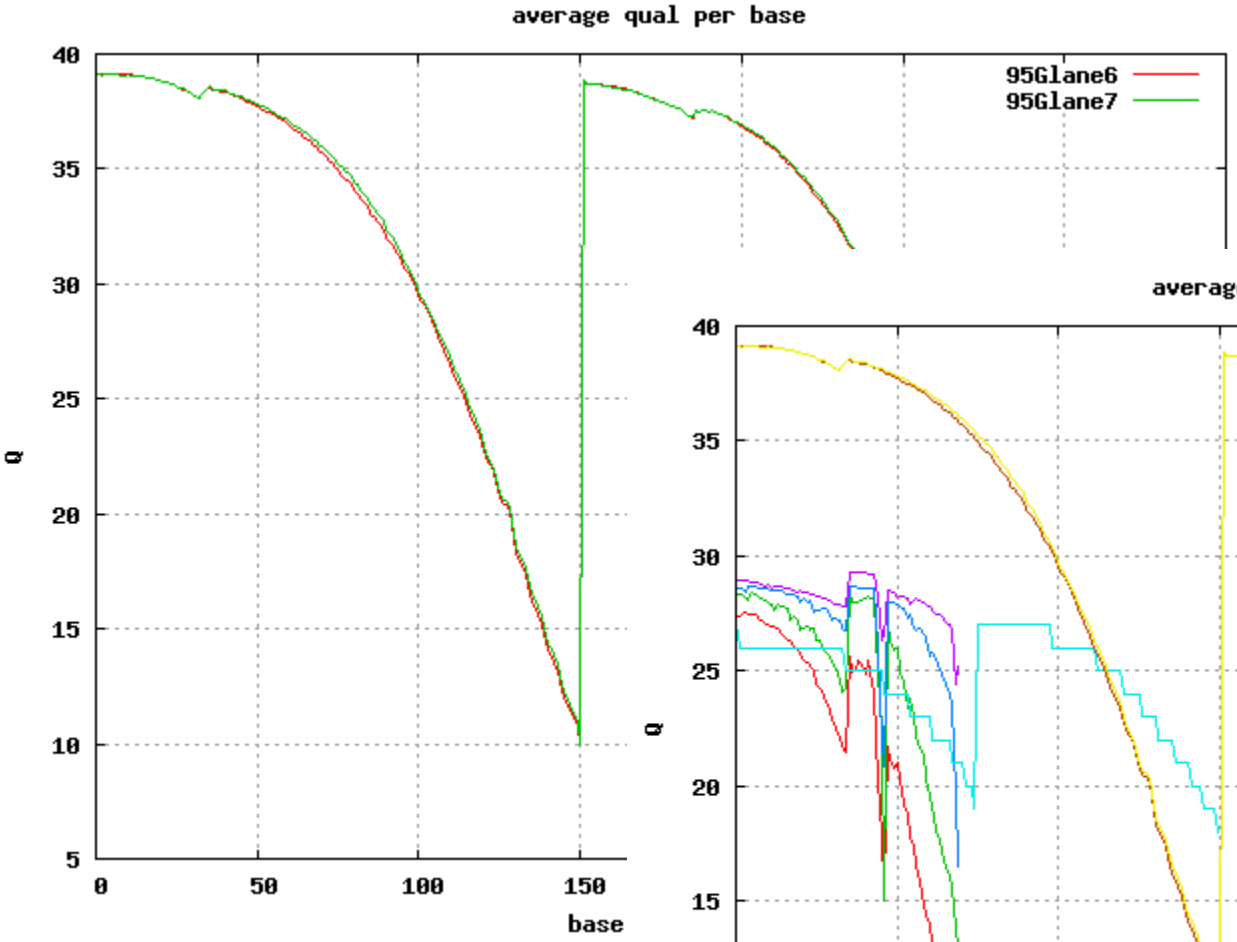


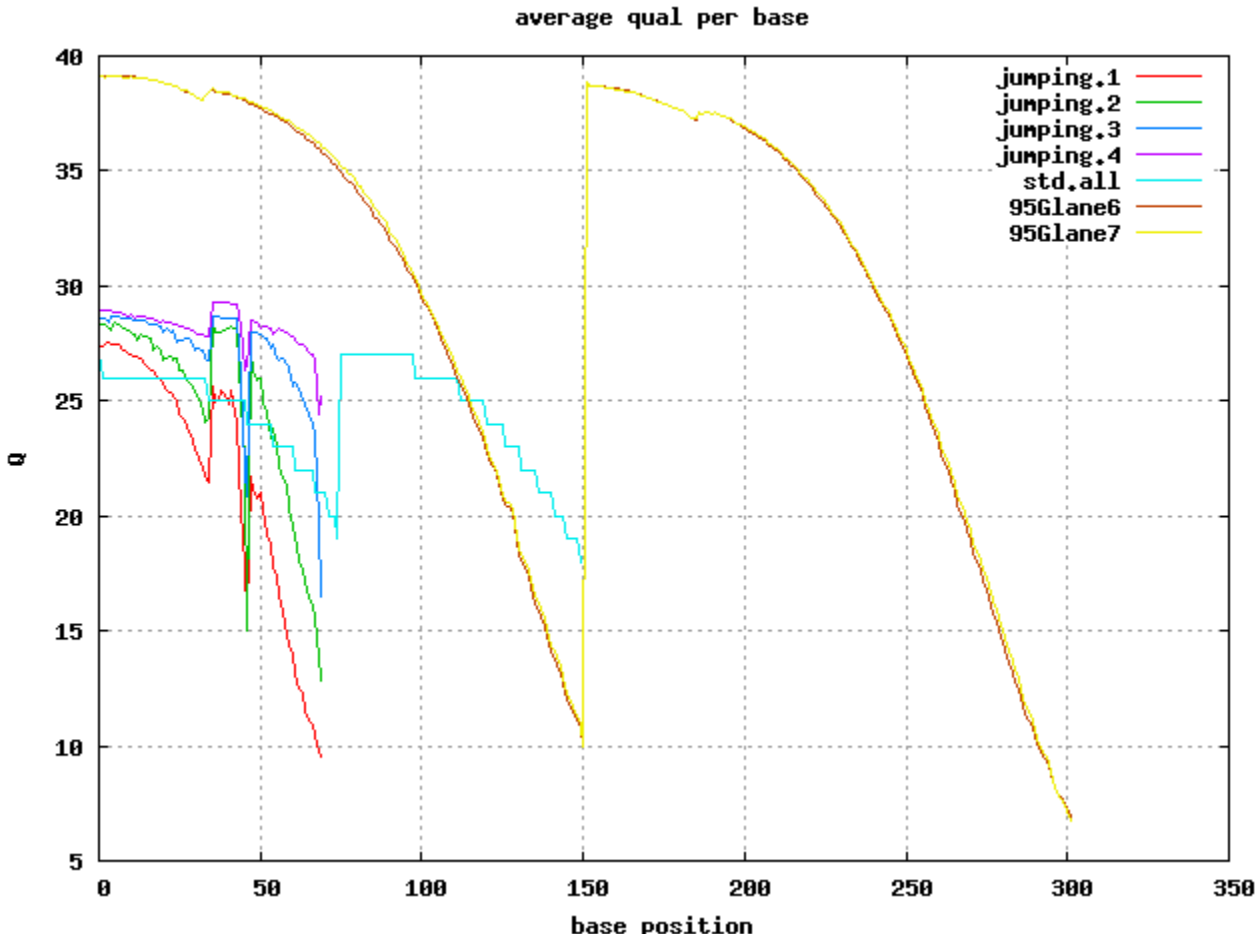


Illumina (MiSeq) data quality

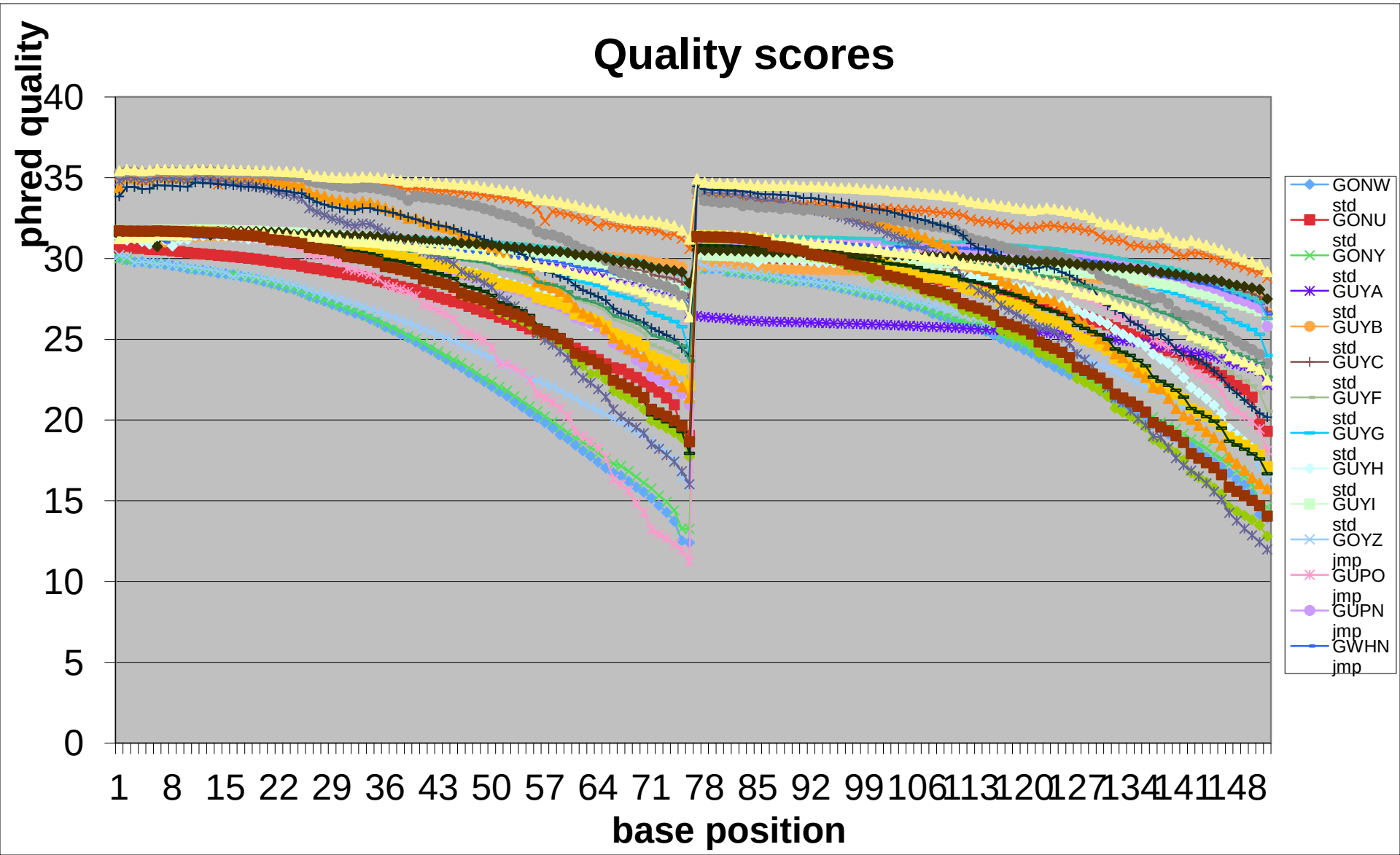
- Some bias in coverage when dealing with extremely high GC and AT-reach genomes
- Error rate is < 0.4%
- Produces errors after long (>20 bp) homopolymers
- Strand dependent errors due to GGC motif (followed by GC rich extension like GGG)

Illumina read quality





Illumina read quality by library

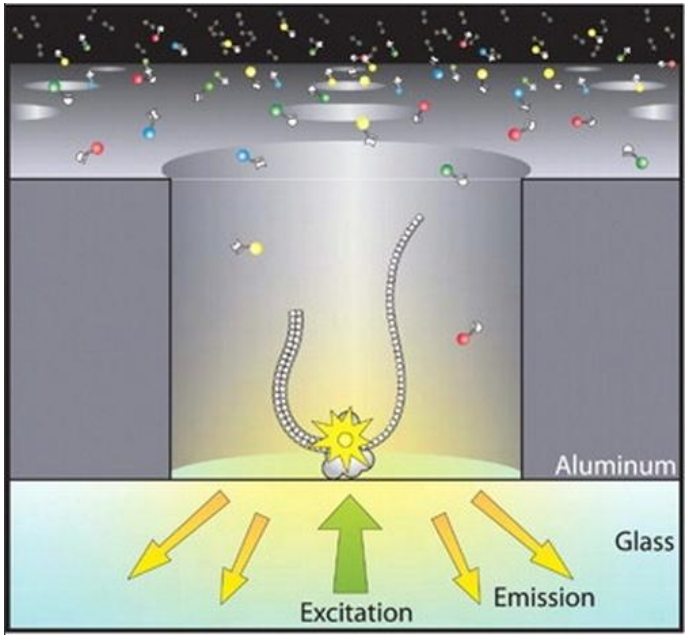


Ion Torrent (PGM) data quality

- Very biased coverage when dealing with extremely high AT-reach genomes
- Error rate is 1.78%
- Don’t generate reads for long (>14 base) homopolymer
- Can not predict the correct number of bases in homopolymers >8 bases long

PacBio read quality

- Some biased coverage towards high GC
- Error rate – 13% (!)
- The highest read length useful for *de novo* assembly scaffolding



Platform Specific errors

GGCGGG

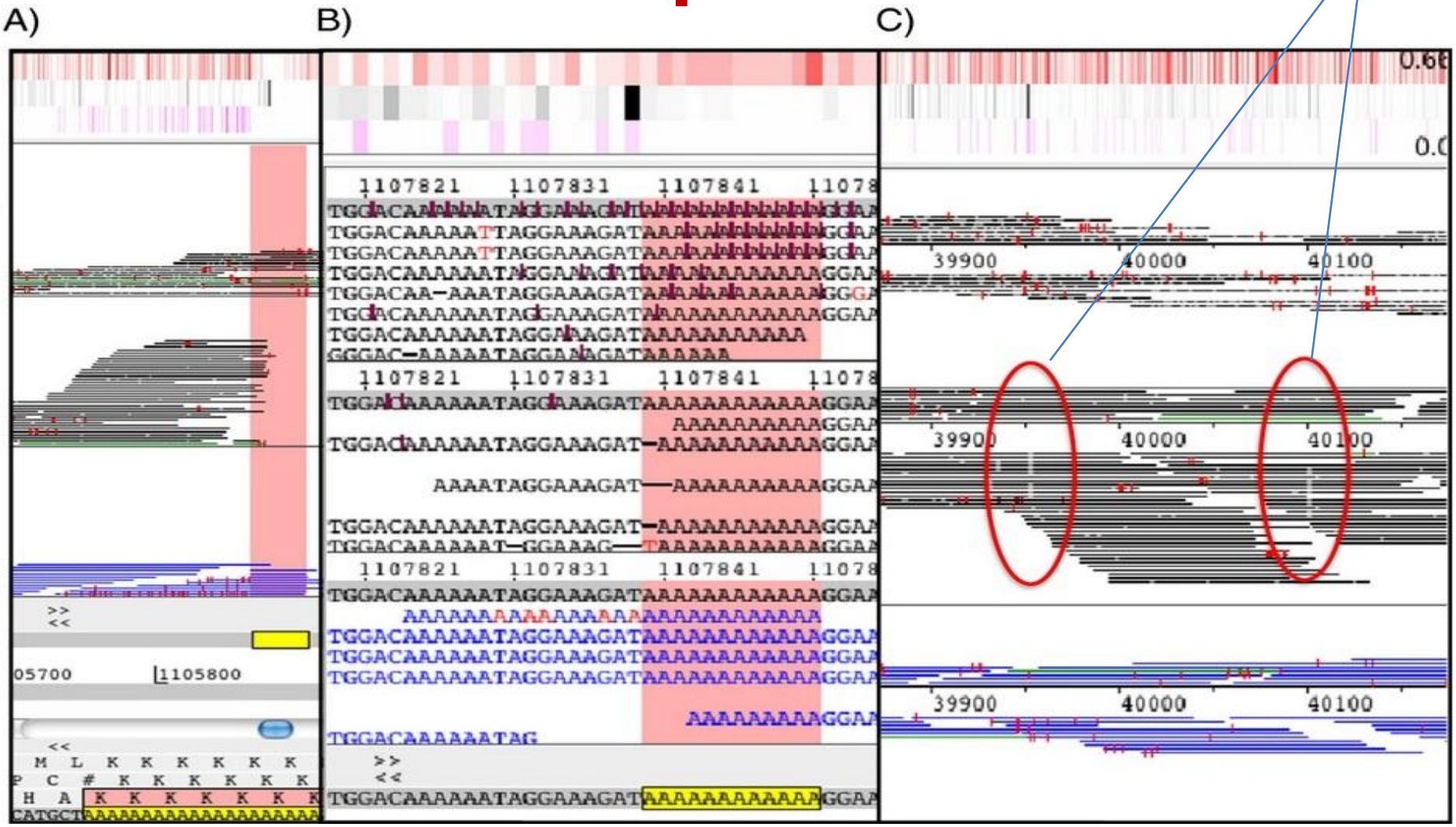
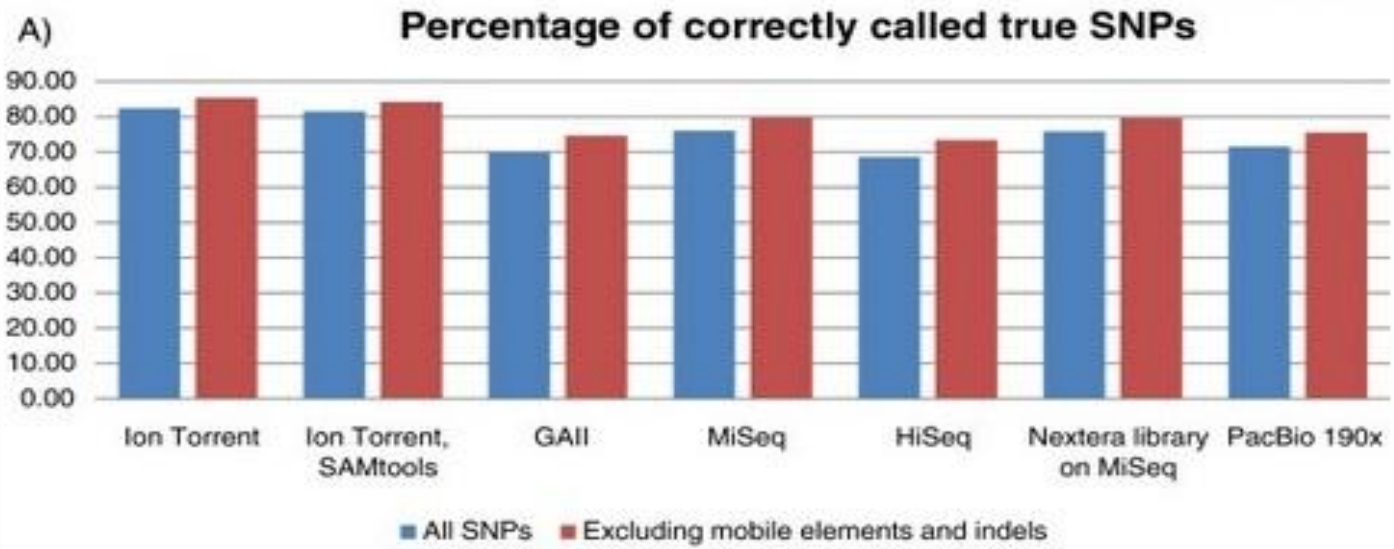
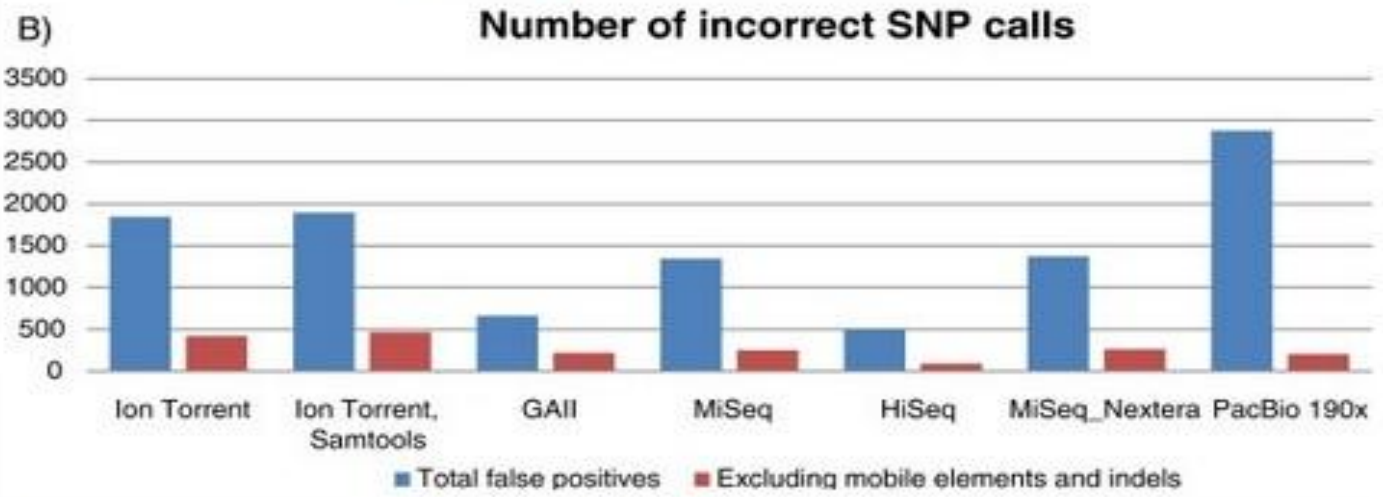


Illustration of platform-specific errors. The panels show Artemis BAM views with reads (horizontal bars) mapping to defined regions of chromosome 11 of *P. falciparum* from PacBio (P; top), Ion Torrent (I; middle) and MiSeq (M; bottom). Red vertical dashes are 1 base differences to the reference and white points are indels. **A**) Illustration of errors in Illumina data after a long homopolymer tract. Ion torrent data has a drop of coverage and multiple indels are visible in PacBio data. **B**) Example of errors associated with short homopolymer tracts. Multiple insertions are visible in the PacBio Data, deletions are observed in the PGM data and the MiSeq sequences read generally correct through the homopolymer tract. **C**) Example of strand specific deletions (red circles) observed in Ion Torrent data.

Quail *et al. BMC Genomics* 2012 **13**:341 doi:10.1186/1471-2164-13-341

Download authors' original image





Accuracy of SNP detection from the *S. aureus* datasets generated from each platform, compared against the reference genome of its close relative *S. aureus* USA300_FPR3757. Both the Torrent server variant calling pipeline and SAMtools were used for Ion Torrent data; SAMtools was used for Illumina data and SMRT portal pipeline for PacBio data. **A**) The percentage of SNPs detected using each platform overall (blue bar), and outside of repeats, indels and mobile genetic elements (red bar). **B**) The number of incorrect SNP calls for each platform overall (blue bar), and outside of repeats, indels and mobile genetic elements (red bar).

Quail *et al. BMC Genomics* 2012 **13**:341 doi:10.1186/1471-2164-13-341
Download authors' original image

Problems caused by

Sequencing approach:

-platform specific errors and features
-number of PCR steps
-enzyme used

Genome nature:

-ability to get sufficient amount of good quality gDNA
-single cell genomes
-GC content
-repeats (total number, variability, length)
-IS elements
-metagenomes

Applications

Microbial genomics

- *de novo sequencing*
- Re-sequencing previously published reference strains (Whole genome re-sequencing)
- Expand the number of available genomes
- Comparative studies

Ecology

- DNA mixtures from diverse ecosystems (Metagenomics)

Food, agriculture, forest

- Sequencing extremely large genomes, crop plants

Clinical Studies (personalysed medicin).

- Targeted sequencing (regions, genes, exomes)
- Microbioita
- Chip-seq: interactions protein-DNA
- Epigenomics
- Transriptom
- DNA Methylation

Pharmaceuticals

- **drug discovery**
- molecular basis of **drug resistance**
- **vaccine development**
- **disease diagnostics**

Ancient DNA

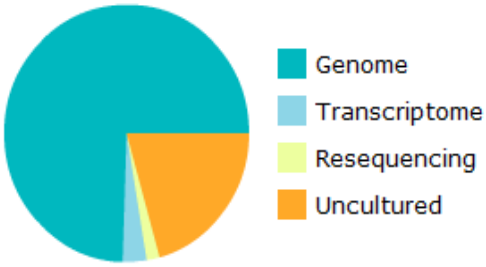
Forensic

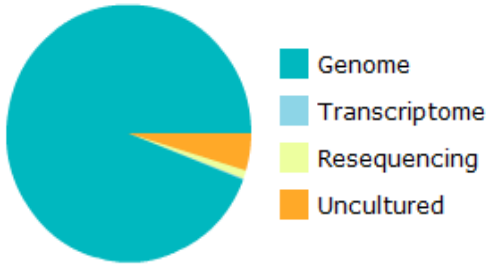
- ……

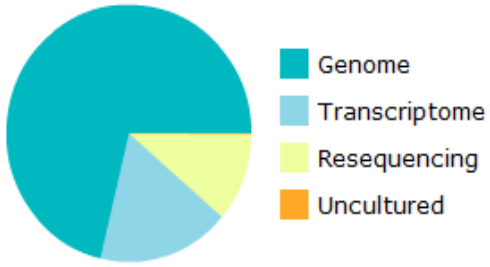
Welcome to the Genomes OnLine Database

PROJECT TYPE DISTRIBUTION					
A	ARCHAEA TOTAL: 632	Genome: 592	Transcriptome: 25	Resequencing: 14	Uncultured: 164
B	BACTERIA TOTAL: 22553	Genome: 22263	Transcriptome: 33	Resequencing: 255	Uncultured: 1099
E	EUKARYA TOTAL: 5665	Genome: 3936	EST/Transcriptome: 964	Resequencing: 617	Uncultured: 6

Archaeal Distribution	Bacterial Distribution	Eukaryotic Distribution



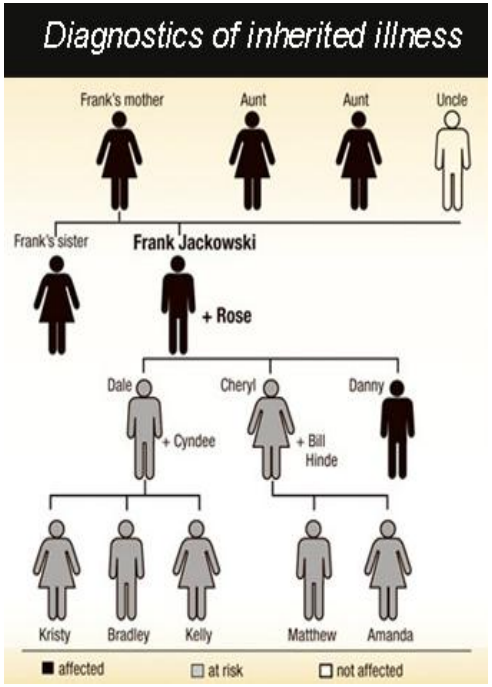




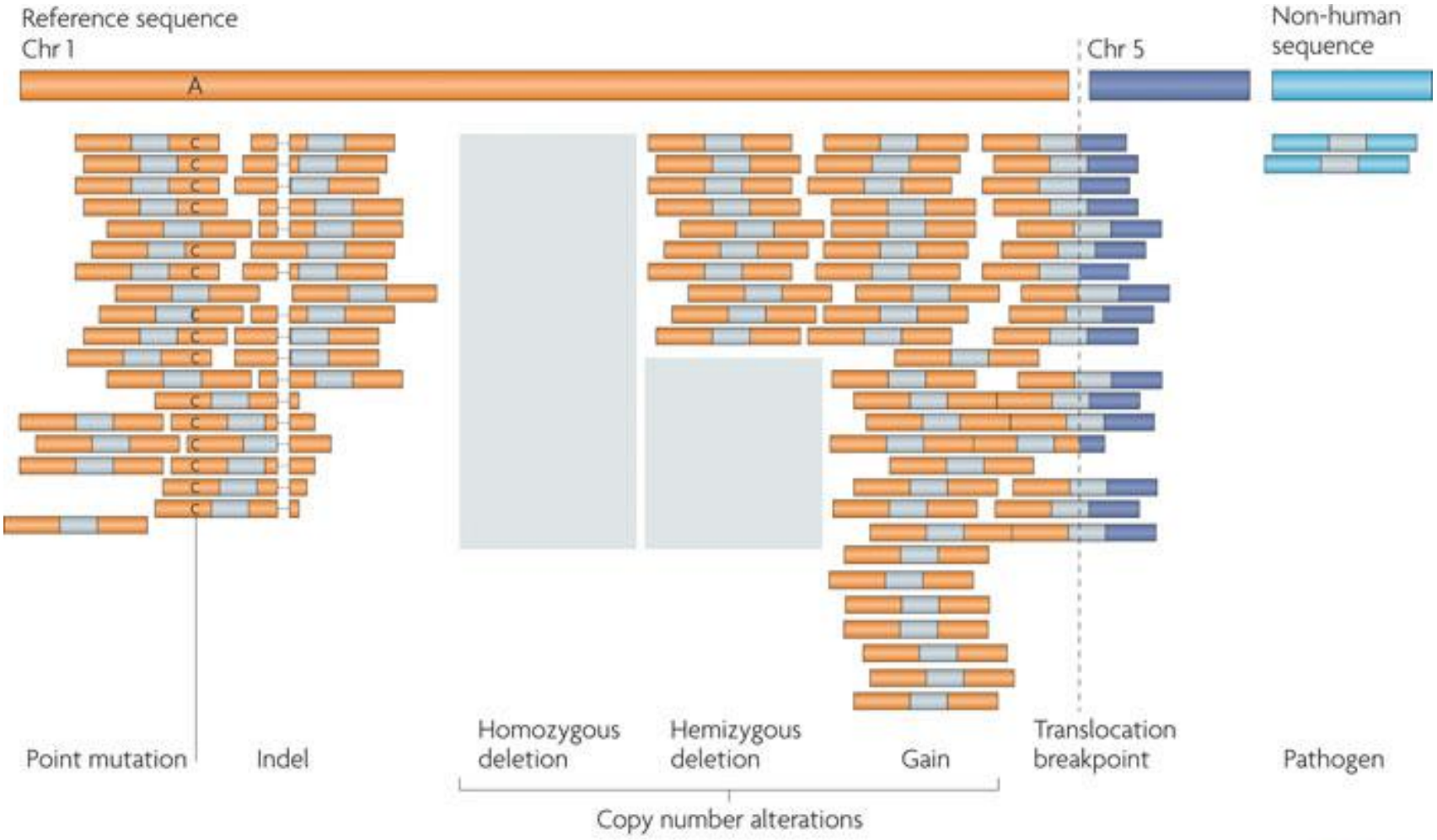
Back to GOLD

Molecular Medicine: let's move to Personalized medicine

- Improved diagnosis of disease
- Earlier detection of genetic predispositions to disease
- Rational drug design
- Gene therapy and control systems for drugs
- Pharmacogenomics "custom drugs"

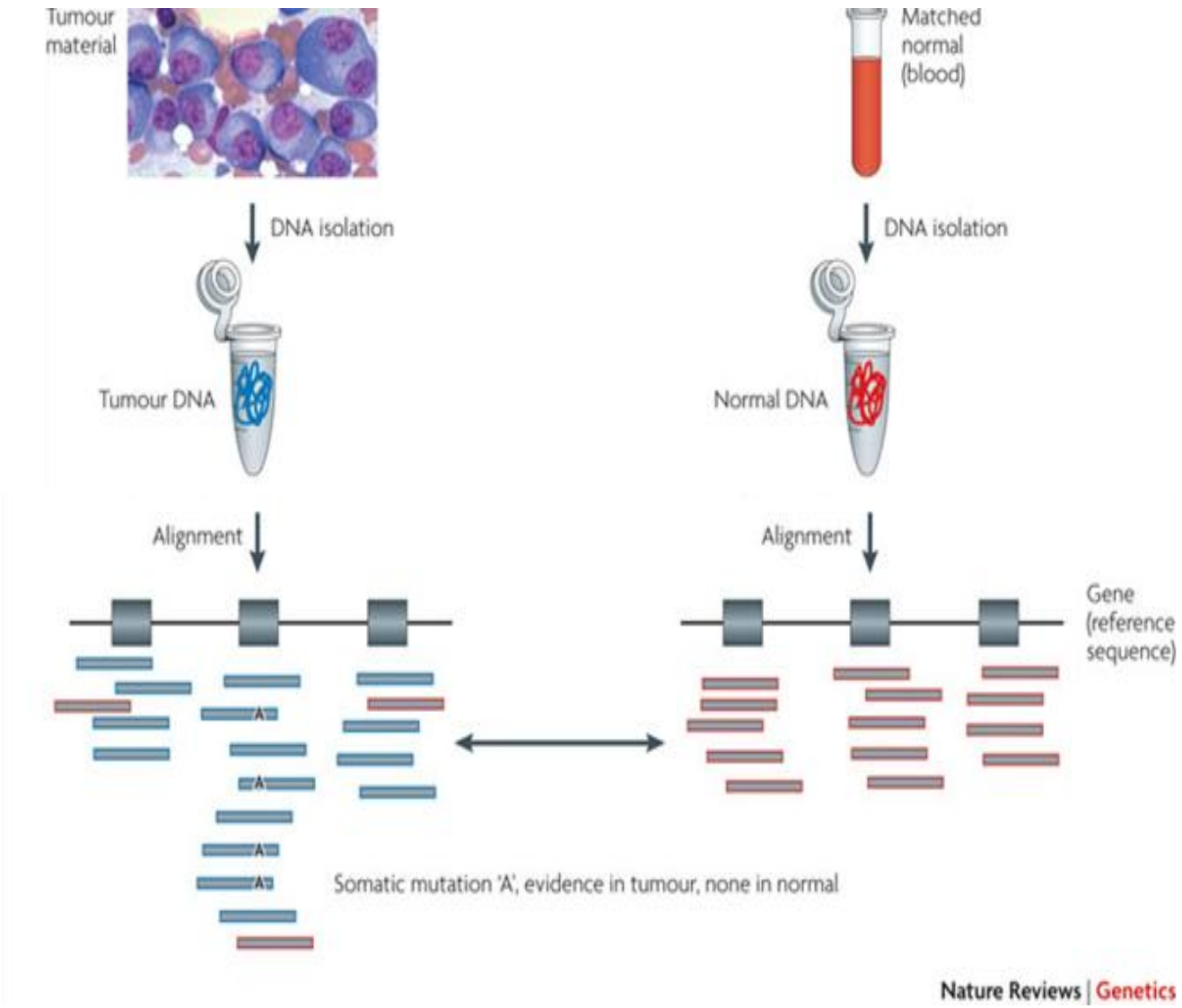


Cancer is a disease of genome alterations. Which alterations can be detected:



M. Meyerson, S.Gabriel, G.Getz, Nature Reviews Genetics 11, 685-696

Somatic mutations



Modified from: M. Meyerson, S.Gabriel, G.Getz, Nature Reviews Genetics 11, 685-696

COSMIC – Cancer Database

COSMIC
Catalogue of somatic mutations in cancer

- was designed to collect and display information on somatic mutations in cancer

Search Enter search here...

- Home
- About
- Download
- Publications
- News
- Contact
- Help
- FAQ

Search COSMIC v66

- Search
- By Gene
- By Sample

Some key features of COSMIC:

Contains information on publications, samples and mutations.in different cancer types.

Samples entered include benign neoplasms and other benign proliferations, in situ and invasive tumours, recurrences, metastases and cancer cell lines.

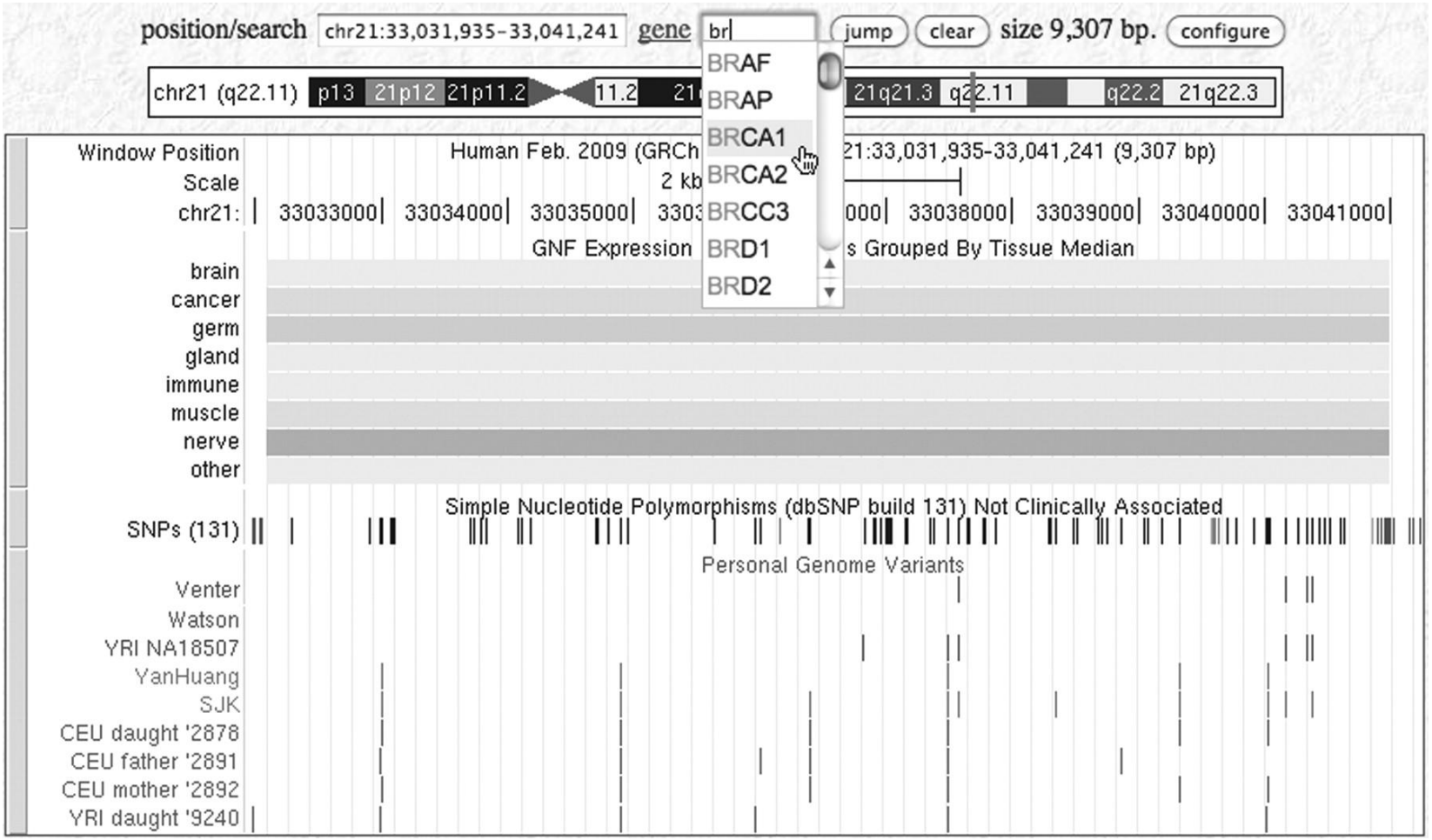
The mutation data and associated information is extracted from the primary literature and entered into the COSMIC database.
In order to provide a consistent view of the data a histology and tissue ontology has been created and all mutations are mapped to a single version of each gene.
The data can be queried by tissue, histology or gene and displayed as a graph, as a table or exported in various formats.

The UCSC Genome Browser database

The University of California, Santa Cruz Genome Browser (http://genome.ucsc.edu) -

offers online access to a database of genomic sequence and annotation data for a wide variety of organisms. The Browser also has many tools for visualizing, comparing and analyzing both publicly available and user-generated genomic data sets, aligning sequences and uploading user data.

Genome Browser display on the hg19 human assembly showing the gene search box in use.

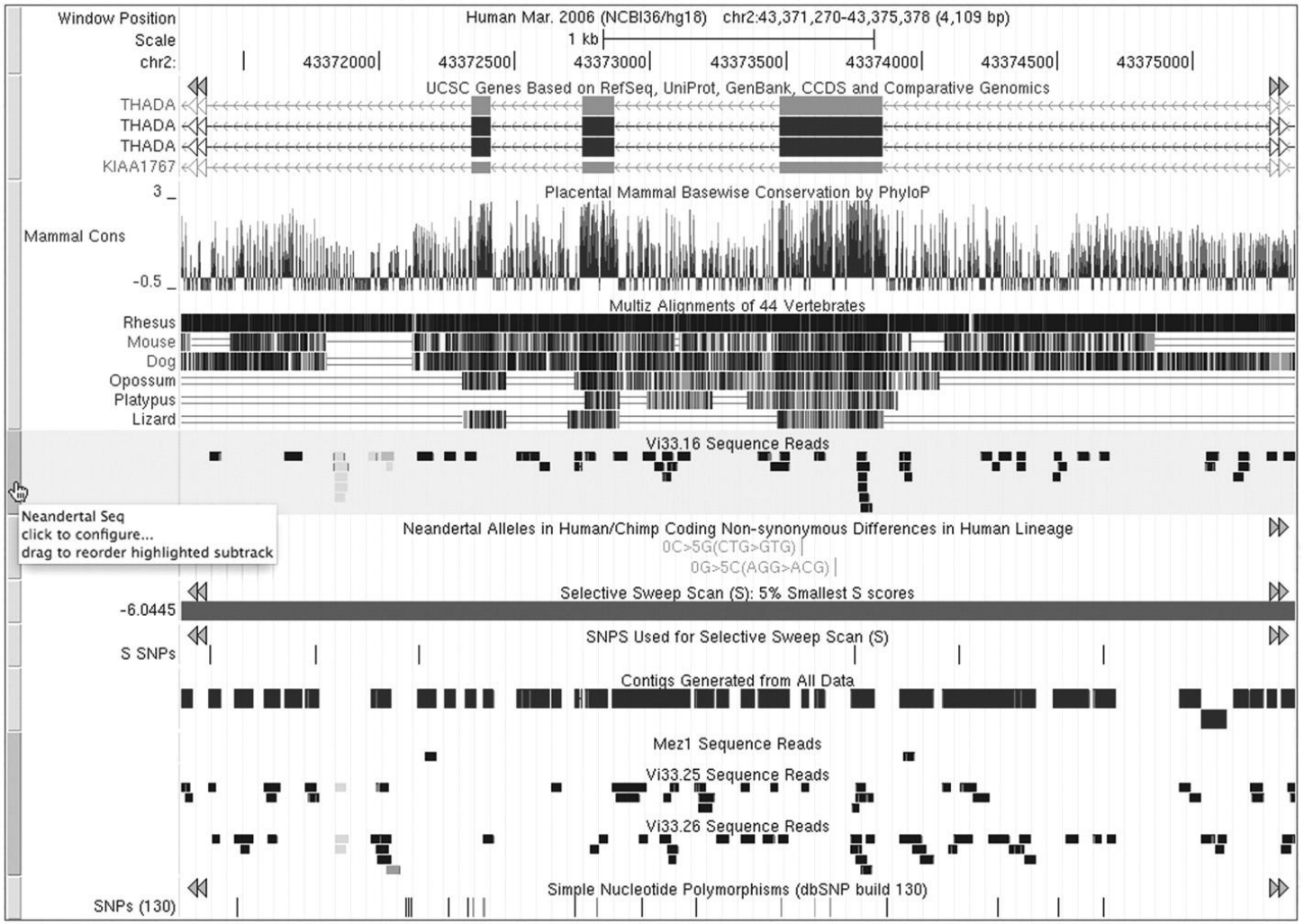


Fujita P A et al. Nucl. Acids Res. 2010;nar.gkq963

© The Author(s) 2010. Published by Oxford University Press.

Nucleic Acids Research

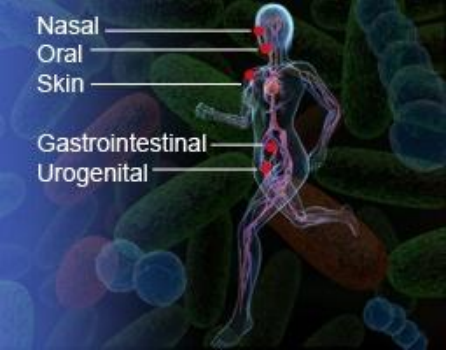
Genome Browser image on the hg18 human assembly showing the UCSC Genes, Conservation and Neandertal tracks (Human-Chimp coding differences, regions with the 5% lowest S, SNPs used to calculate S and alignments of Neandertal sequence reads).



Fujita P A et al. Nucl. Acids Res. 2010;nar.gkq963

© The Author(s) 2010. Published by Oxford University Press.

Nucleic Acids Research



Human microbiome

The human microbiome includes viruses, fungi and bacteria, their genes and their environmental interactions, and is known to influence human physiology.

There's very broad variation in these bacteria in different people and that severely limits our ability to create a "normal" microflora profile for comparison among healthy people and those with any kind of health issues.

Example 1: autistic children have microbiomes that differ from those of other kids. The strong correlation of gastrointestinal symptoms with autism severity indicates that children with more severe autism are likely to have more severe gastrointestinal symptoms and vice versa. It is possible that autism symptoms are exacerbated or even partially due to the underlying gastrointestinal problems.

Example 2: You are what you eat!

SILVA - Good for metagenome analysis

silva
comprehensive ribosomal RNA databases

Home Browser Search Aligner Download Documentation Projects FISH & Probes Shop Contact

SILVA

Welcome to the SILVA rRNA database project

A comprehensive on-line resource for quality checked and aligned ribosomal RNA sequence data.

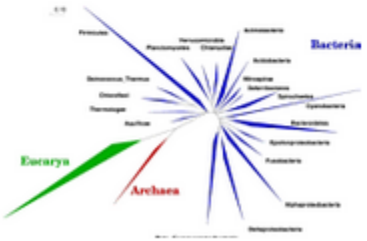
SILVA provides comprehensive, quality checked and regularly updated datasets of aligned small (16S/18S, SSU) and large subunit (23S/28S, LSU) ribosomal RNA (rRNA) sequences for all three domains of life (***Bacteria***, ***Archaea*** and ***Eukarya***).

SILVA are the official databases of the software package ARB.

For more background information → Click here

ARB

The software package ARB represents a graphically-oriented, fully-integrated package of cooperating software tools for handling and analysis of sequence information.



The ARB project has been started more than 15 years ago by Wolfgang Ludwig at the Technical University in Munich, Germany, see www.arb-home.de.

News

16.06.2013

Preview SILVA Release 115 Statistics

More than 4 Mio SSU and LSU sequences...

06.06.2013

Working towards SILVA release 115

Preparation of SILVA release 115 has started. SILVA 115 will be a full release with updated taxonomy and trees, as well as ARB files. The release is planned for July 2013.

03.03.2013

Meet ARB & SILVA at VAAM 2013

Talk to the ARB and SILVA developers at VAAM 2013 (10.03-13.03) in Bremen, Germany. Follow the link to see the sessions where you will find us.

18.02.2013

LTP 111 released

Version 111 of the "All Species Living Tree" (LTP) has been released. Check the project website (link above) for more information ...

go to Archive ->

SILVA 111 - full release

Artemis: *Genome Browser* - Welcome Trust Sanger Institute

wellcome trust sanger institute

Search for Enter search here...

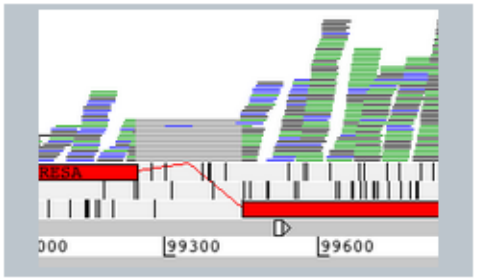
Home | Research | Scientific resources | Work & study | About us

Mouse | Zebrafish | Data | Software | Databases | Technologies | Talks & training

Artemis: Genome Browser and Annotation Tool

Artemis is a free genome browser and annotation tool that allows visualisation of sequence features, next generation data and the results of analyses within the context of the sequence, and also its six-frame translation.

Artemis is written in Java, and is available for UNIX, Macintosh and Windows systems. It can read EMBL and GENBANK database entries or sequence in FASTA, indexed FASTA or raw format. Other sequence features can be in EMBL, GENBANK or GFF format.



[Genome Research Limited]

Links

- ACT - a DNA sequence comparison viewer
- DNAPlotter - makes circular and linear interactive plots
- BamView - interactive display of read alignments in BAM data files

Information | Development | Download | FAQs | Chado | Courses | Contact

New to Artemis?

The Artemis manual explains how to install and run Artemis and what most parts of the program do. The FAQs may help if you are experiencing problems with Artemis. Also an Artemis poster gives an overview of browsing genomes and visualisation of next generation data in Artemis. There are also use case examples of browsing next generation sequence data.

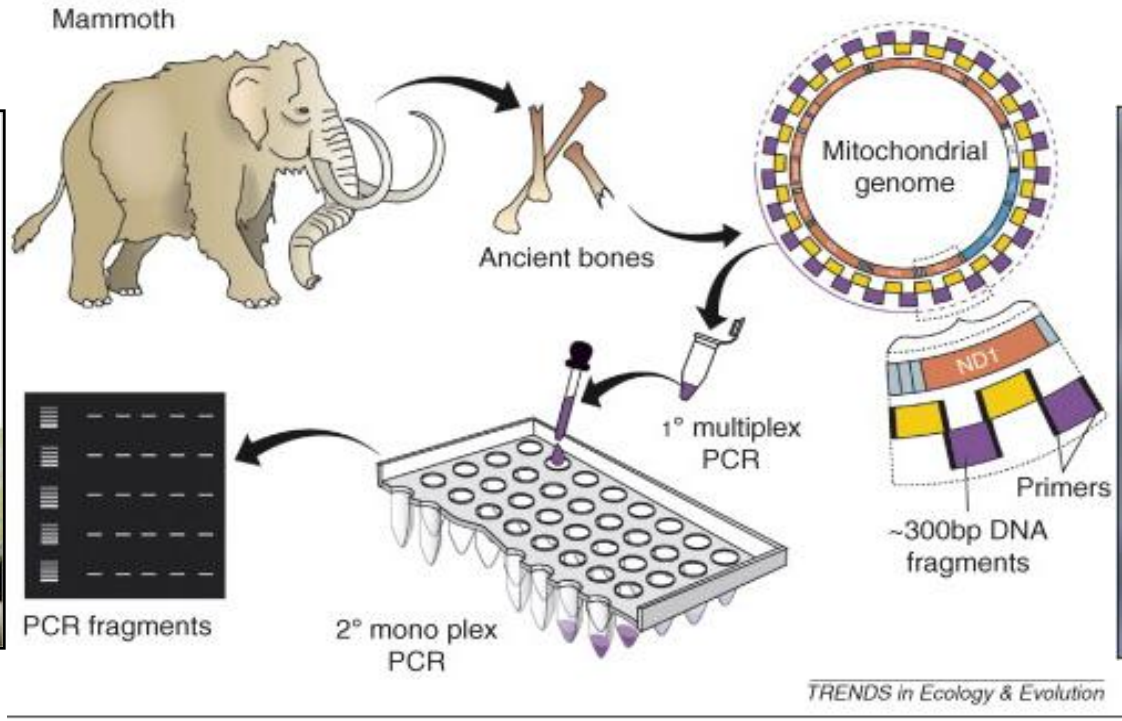
Full information about the latest release of Artemis can be found in the manual and the current release notes.

License

Artemis is free software and is distributed under the terms of the GNU General Public License. It should run on any system with a recent version of Java.

Ancient Genomes

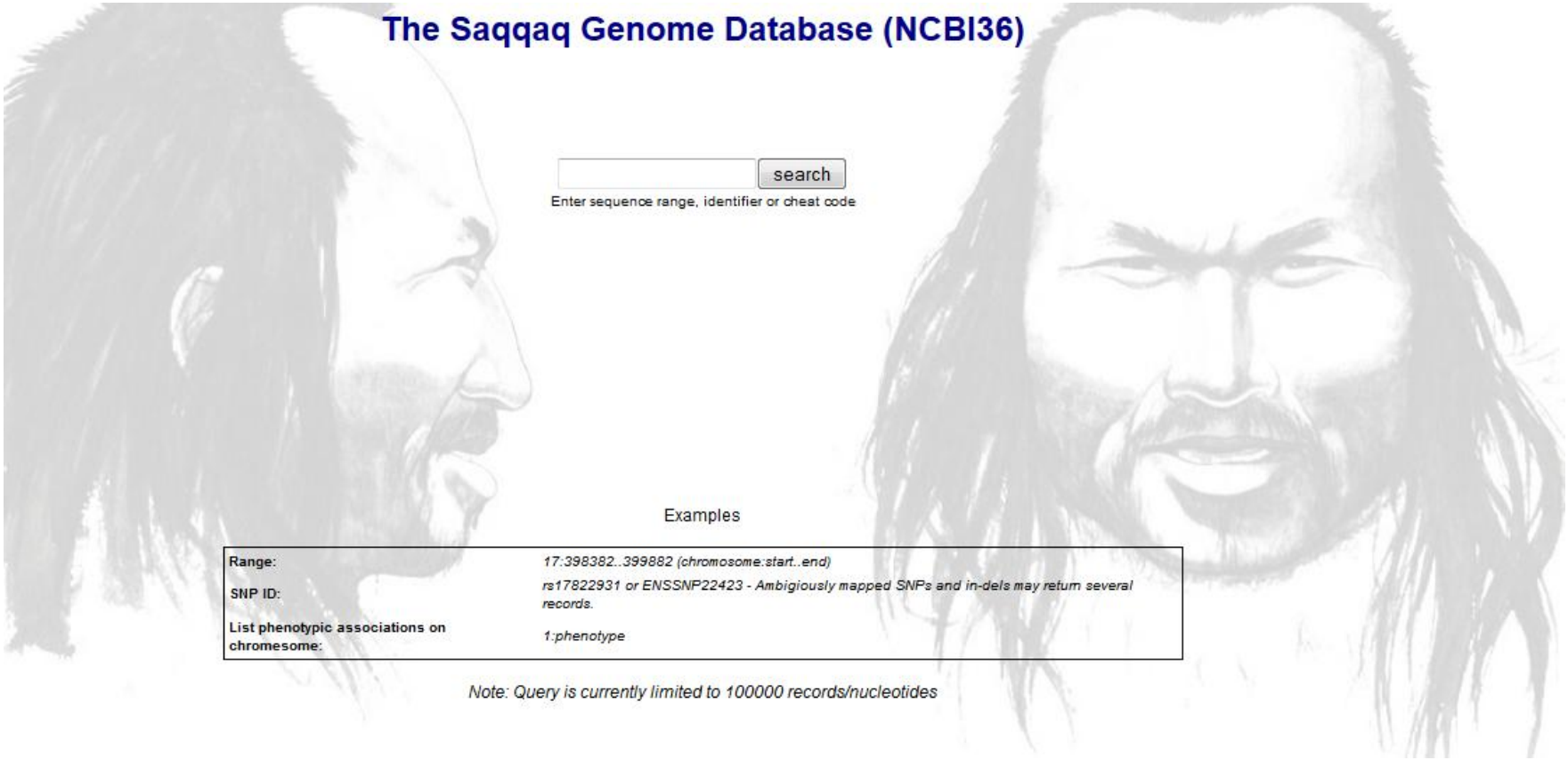
- **Degraded state of the sample → mitDNA sequencing**
- **Small amount of DNA**
- **Nuclear genomes of ancient remains: cave bear, mommoth, Neanderthal (10^6 bp)**



Problems: contamination modern humans and bacterial DNA

Ancient Genome DataBase: Saqqaq

The ancient genome database is an ongoing effort to build a genotype-phenotype catalogue and reference available ancient genome data to this. The saqqaq genome was the first ancient nuclear genome sequence at high coverage, which is what is currently referenced to in the database.



Forensic Genomics:

use of NGS for crime investigations and missing person identification, kinship testing and ancestry investigation

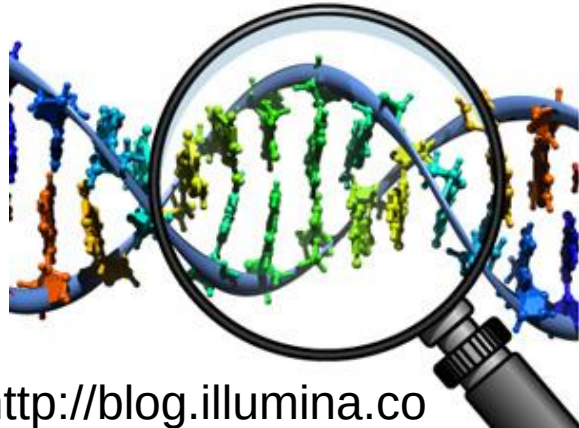
Task: reveal Forensic DNA evidences from tiny and highly mixed samples.

Study:

-short tandem repeat (STR) typing (repeating units of 2–6 nucleotides

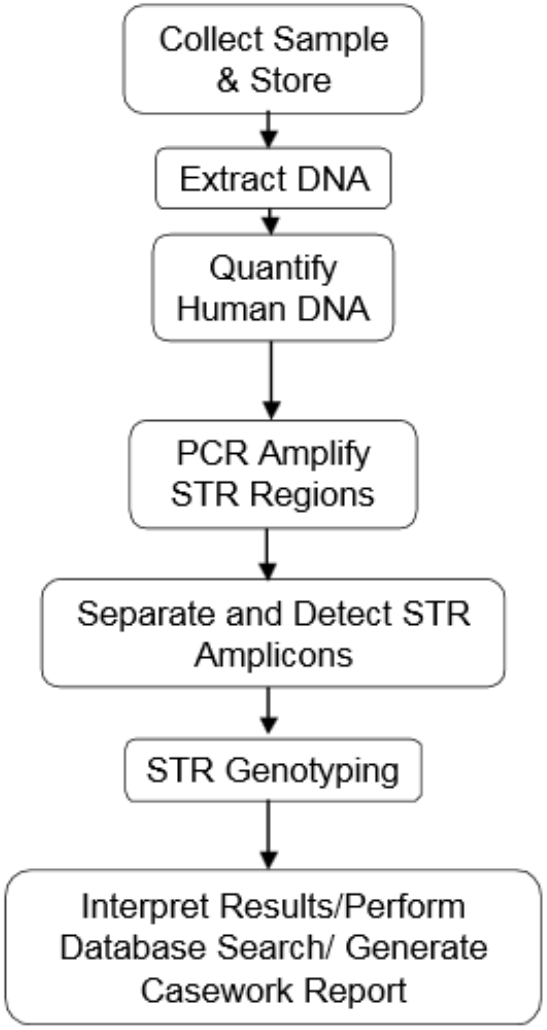
-mitochondrial DNA analysis,

-dense panels of single nucleotide polymorphisms (SNPs) offering



http://blog.illumina.co

Summary of process used for forensic DNA typing with STR marker



The human identity testing community has focused on 22 autosomal STR loci (Table and about a dozen Y-chromosome STR markers that are present in commercial kits.

The use of core sets of loci enables common information to be included in criminal DNA databases.

The chromosomal locations, repeat motifs, allele ranges, PCR product sizes, and random match probabilities for these common autosomal STR loci are stored in the database.

NCBI

- NCBI - Established in 1988 as a national resource for molecular biology information, NCBI creates public databases, conducts research in computational biology, develops software tools for analyzing genome data, and disseminates biomedical information - all for the better understanding of molecular processes affecting human health and disease.

NCBI - databases and services

- sequence databases

  GenBank, ESTs, SNPs, etc.
- PubMed - literature database
- Entrez

  http://www.ncbi.nlm.nih.gov/entrez/

  retrieval system connecting together plethora of databases including PubMed, genomes, ontologies
- **Blast** - basic local alignment search tool
- Science primer

  http://www.ncbi.nlm.nih.gov/About/primer/

  introductions into molecular biology and bioinformatics

EMBL

EBI - The European Bioinformatics Institute (EBI)is a non-profit academic organisation that forms part of the European Molecular Biology Laboratory (EMBL)

EBI databses

- EMBL nucleotide database - http://www.ebi.ac.uk/embl/
- UniProt (together with Expasy and PIR)
- ArrayExpress - http://www.ebi.ac.uk/arrayexpress/

  public repository for microarray data
- Ensembl - http://www.ensembl.org/ - genomes and annotation for metazoa

DNA Data Bank of Japan

DDBJ - DDBJ (DNA Data Bank of Japan) began DNA data bank activities in earnest in 1986 at the National Institute of Genetics (NIG).

DDBJ has been functioning as the international nucleotide sequence database in collaboration with EBI/EMBL and NCBI/GenBank.

DNA sequence records organismic evolution more directly than other biological materials and thus is invaluable not only for research in life sciences but also human welfare in general. The databases are, so to speak, a common treasure of human beings. With this in mind, we make the databases online accessible to anyone in the world.

DNA variations – SNPs, CNVs

- many projects set to deal with intra-species variation
  - dbSNP
    `http://www.ncbi.nlm.nih.gov/SNP/`
  - the SNP consortium
    `http://snp.cshl.org/`
  - haplotypes
    `http://www.hapmap.org/`
  - glovar - human variations
    `http://www.glovar.org/`
  - human variome
    `http://www.humanvariomeproject.org/`
  - general variomes
    `http://variome.net/`

Gene prediction

Open-source and on-line gene prediction

- Glimmer - bacteria, archea, viruses
  http://cbcb.umd.edu/software/glimmer/
- GlimmerHMM - eukaryotic genes
  http://cbcb.umd.edu/software/GlimmerHMM/
- GeneZilla (TIGRscan) - eukaryotic genes
  http://www.genezilla.org/
- GenScan - human genes
  http://genes.mit.edu/GENSCAN.html
- software lists
  http://www.genefinding.org/

Nucleic structures

RNAs and 3D nucleic structural databases

- 3D structures of nucleic acids

  RNABase - http://www.rnabase.org/

  NDB nucleic acids database- http://ndbserver.rutgers.edu/
- SCOR - structural classification of RNA - http://scor.berkeley.edu/

  RNA motifs, structures and interactions
- other databases

  Small RNA database - http://condor.bcm.tmc.edu/smallRNA/

  Noncoding RNA database - http://biobases.ibch.poznan.pl/ncRNA/

ExPASy (expert protein analysis system)

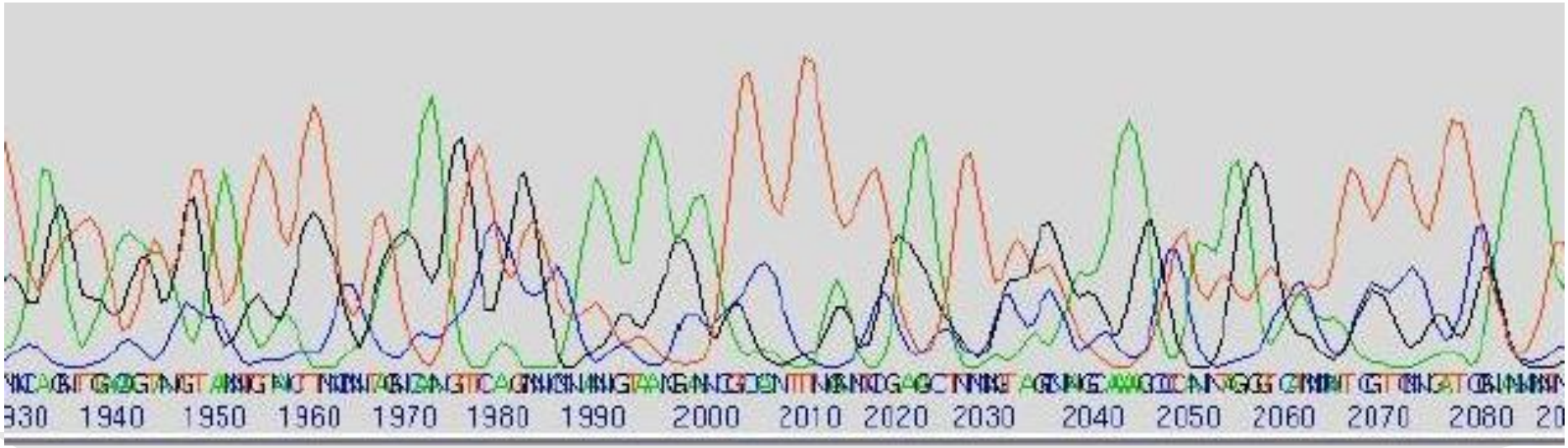
- UniProt - the universal protein resource
  http://www.expasy.uniprot.org/
  - knowledgebase, reference clusters, archives
- Swissprot - http://www.expasy.ch/sprot/
  - database of protein sequences together with annotations
  - structure and function of proteins
- prosite
  http://www.expasy.ch/prosite/
  - documentation on protein domains, folds, families

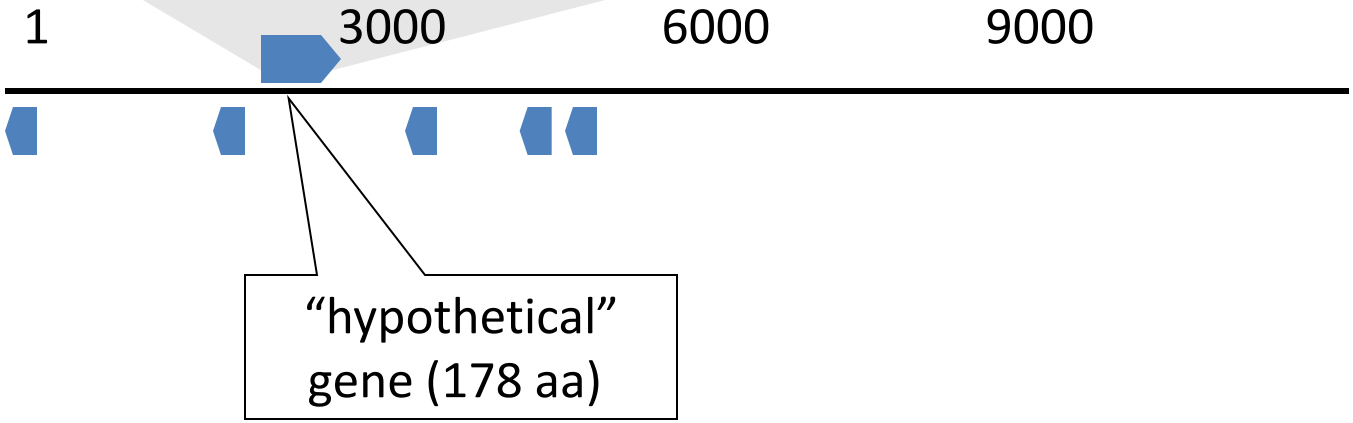
Moving from second-generation to third-generation sequencing strategies

Advanced Technological Approaches Generates Genomic Data Better, Faster, and Cheaper

In the next-gen sequencing arena, the focus over the past several years has been on technological advances, moving from second-generation (SGS) to third-generation sequencing strategies (TGS) and producing research instruments capable of delivering whole-genome sequences in parallel at increasing speed.

More recently, as read lengths and coverage continue to increase, throughputs rise, and costs decline, the expanding range of applications of NGS has taken center stage.


1930 1940 1950 1960 1970 1980 1990 2000 2010 2020 2030 2040 2050 2060 2070 2080


1
3000
6000
9000
“hypothetical” gene (178 aa)

Assembly problem -1

THANK YOU!