Клиническая биоинформатика
vs
биоинформатика

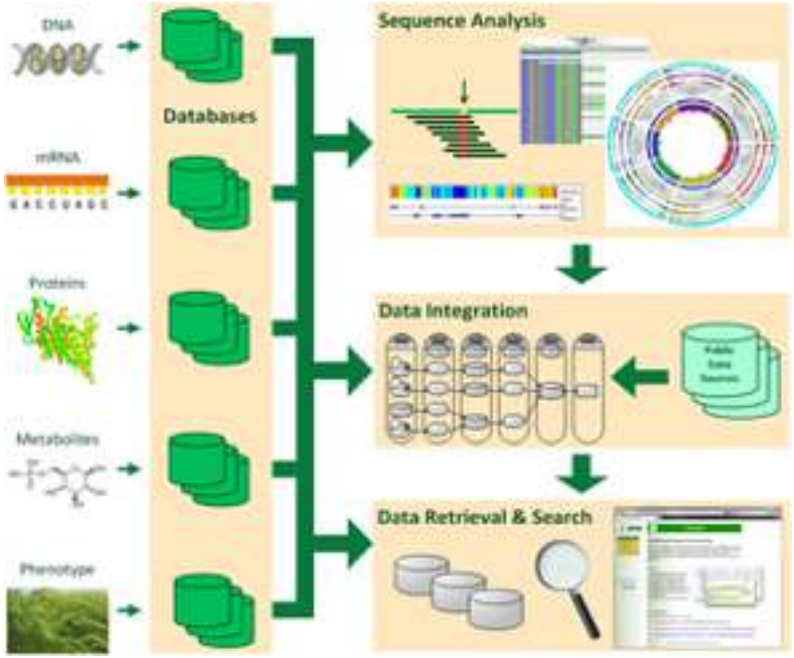
Alla L Lapidus, Ph.D.

What is Bioinformatics

Bioinformatics is a science of biological data that:

- develops algorithms and software for biological tasks
- curates, analyses and stores data related to biology
- organizes and distributes knowledge of living organisms functionality

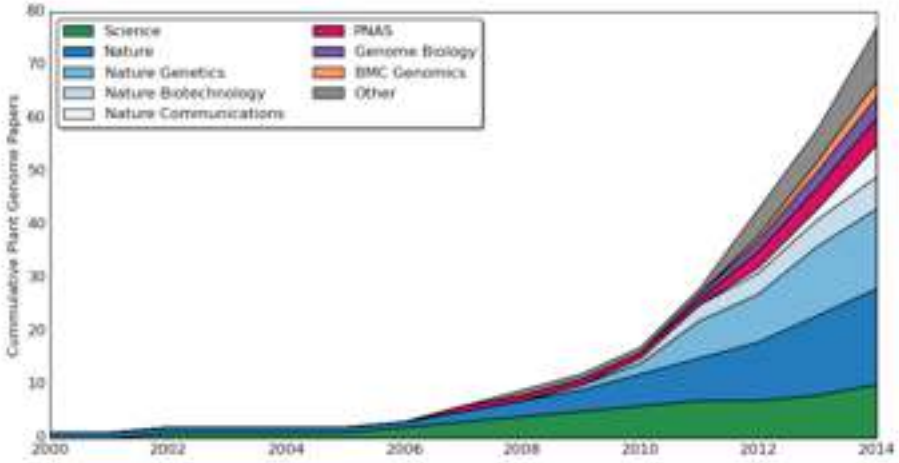
What data does bioinformatics deal with?



DNA, RNA, and protein sequences
Molecular Structures
Expression Data
Data storage
Data transfer
Data Bases

Bibliographic Data

The number of scientific articles has increased dramatically in the last few decades, due to the increasing number of research projects and genome sequencing programs. These articles are organized in public databases available online



Why bioinformatics is important

Shift from experimental laboratory bench only to
Incorporation of computers into the research process

The greatest challenges

- to make sense of the wealth of data that has been produced at genome, transcriptome, proteome and metabalome levels

- intelligent and efficient storage of huge amount of data generated

- to provide easy and reliable access to this data

- creation of tools to allow the extraction of meaningful biological information

- apply bioinformatics tools to reduce time and cost in molecular marker development, drug development etc

Why bioinformatics is important

The genomic era has seen a massive explosion in the amount of biological information due to huge advances in the fields of
molecular biology and new sequencing technologies

Depends on access to high performance computing capabilities

Instruments on the market vs applications

Single gene assays

Multiplexed hotspots

Multi gene panels

Whole exome

Whole genome

ABI

Sequenom MassArray

IonTorrent PGM

MiSeq Illumina

HiSeq Illumina

Instruments on the market vs applications

“A space-based DNA sequencer could identify microbes, diagnose diseases and understand crew member health, and potentially help detect DNA-based life elsewhere in the solar system”

http://www.nasa.gov/

Bioinformatics is being used in the following fields

- Molecular medicine
- Personalised medicine
- Preventative medicine
- Gene therapy
- Drug development
- Antibiotic resistance
- Microbial genome applications
- Evolutionary studies
- Forensic analysis
- Waste cleanup
- Climate change Studies
- Alternative energy sources
- Bio-weapon creation
- Biotechnology
- Crop improvement
- Insect resistance
- Improve nutritional quality
- Development of Drought resistant varieties
- Vetinary Science

Clinical application of bioinformatics => Clinical Bioinformatics (CBI)

- helps to understand molecular mechanisms and potential therapies for human diseases

- integrates molecular and clinical data to accelerate the translation of knowledge discovery into effective treatment and personalized medicine.

- is aimed at providing methods and tools to support
  clinicians and researchers

Особенности биоинформатики, исследующей данные человека

Повышенные требования к:

- качеству экспериментальных данных
- уровню ответственности
- простоте и надежности создаваемых программ и аналитических подходов

Этические аспекты

Bioinformatics tools in Medical Research

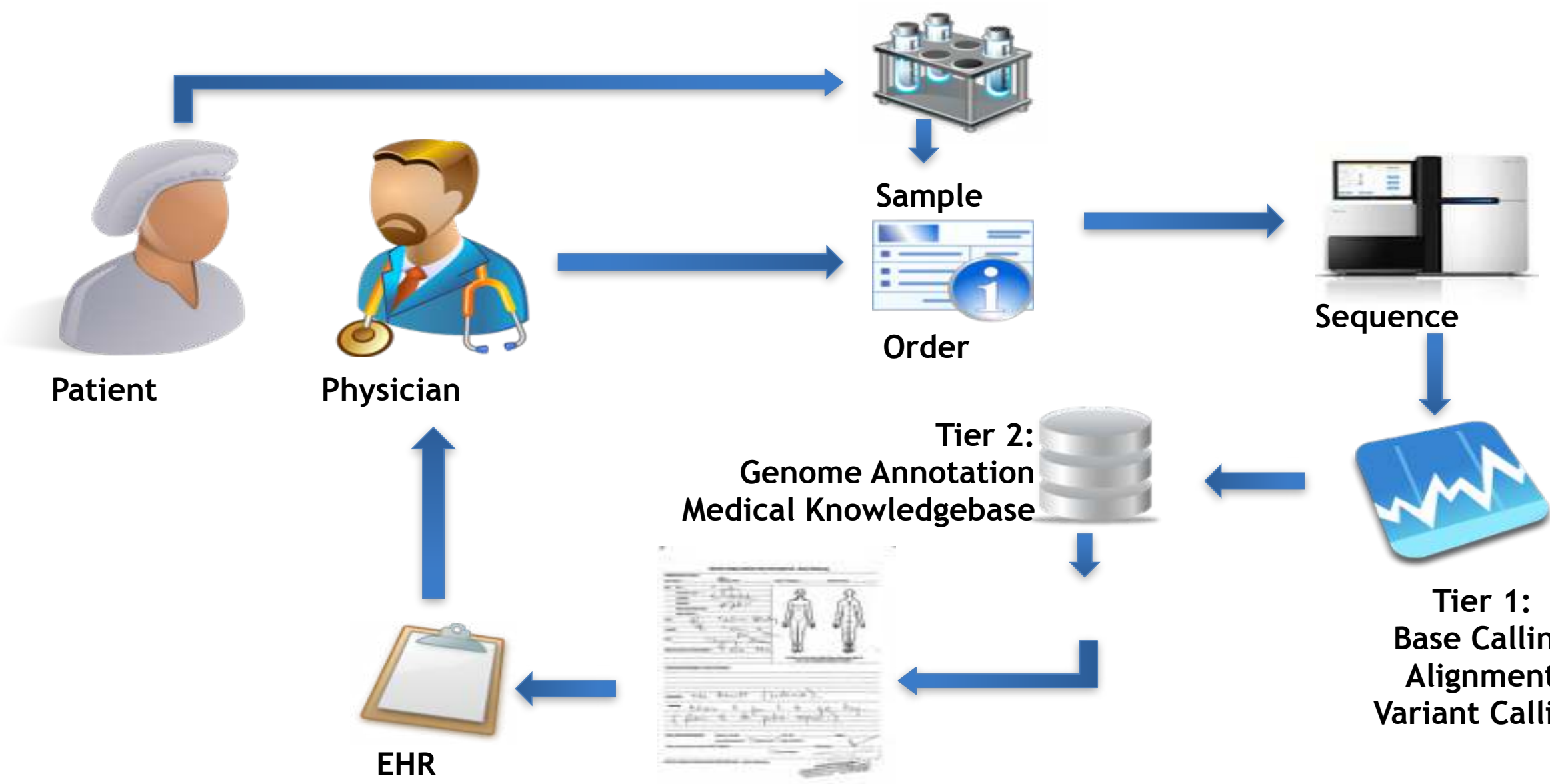
Help

- to compare genetic and genomic data
- with understanding of various molecules that are amenable for the disease
- to analyze and document the biological systems and pathways
- to combine clinical data from patients in Electronic Medical Reports
- can affirm potential biomarkers and clinical phenotypes that allow researchers to develop experimental strategies using selected patient

Bioinformatics and Health Informatics

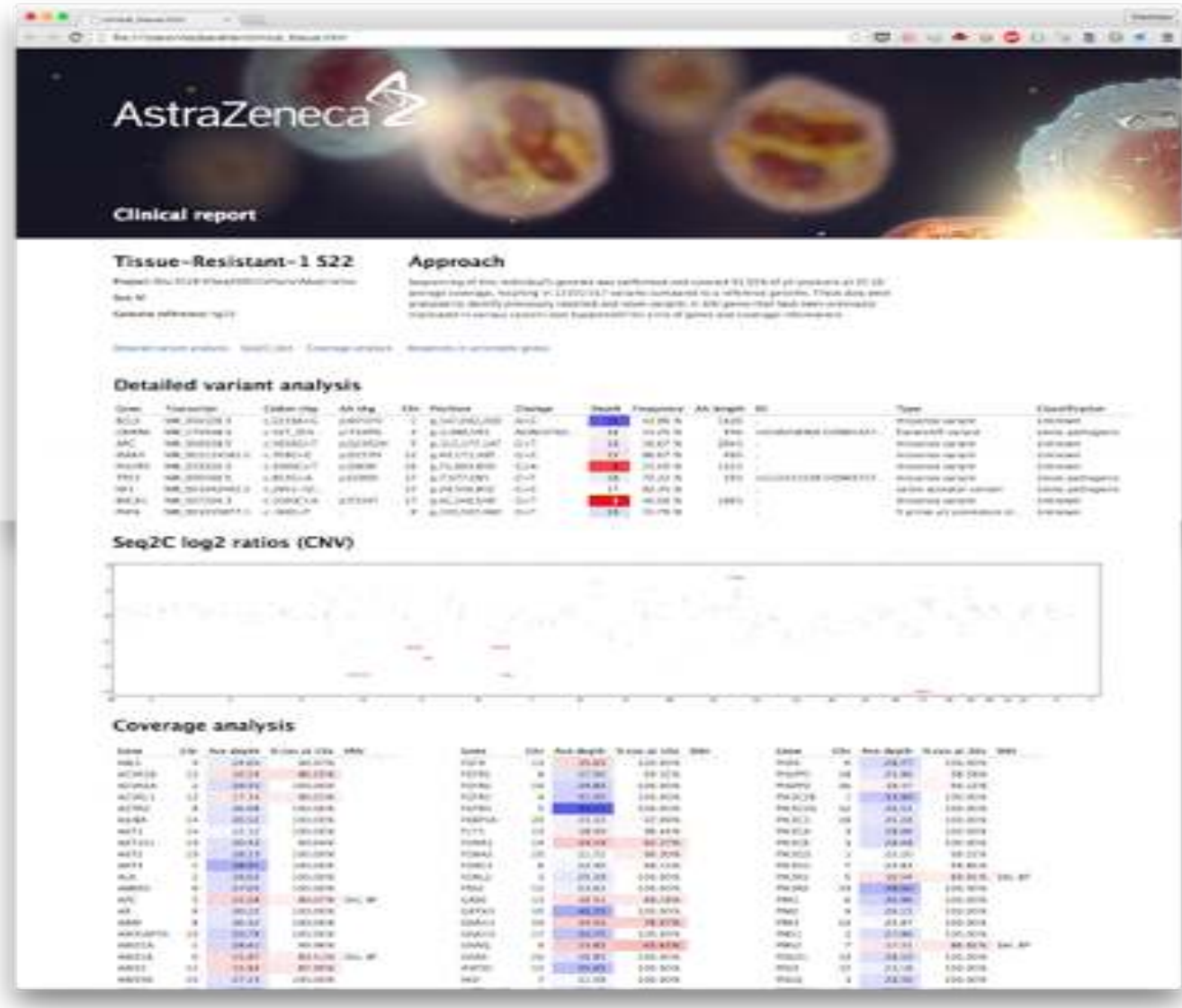
If bioinformatics is the study of the flow of information in biological sciences,
Health Informatics is the study of the information in patient care

Medicine: Informatics pipeline workflow



Автоматическое составление клинического отчета

1. Информация о пациенте
2. Метод секвенирования
3. Найденные мутации и их влияние на фенотип
4. Качество покрытия ключевых для онкологии генов
5. Вариация копийности экзонов
6. Открытые исследования, где встречались такие же мутации:
   - Инф. о пациенте и диагнозе
   - Лекарства
   - Существующие клеточные линии
   - Ассоц. биомаркеры



Tools

Aligners (Bwa, MAQ, Bowtie, Eland, MOSAIK, SHRiMP, SOAP2 etc)

Mutation calling (SamTools, CASAVA, VarScan, etc)

Indel calling (Pindel, GATK, CLC, etc)

Visualization (MapView, Maqview, consed, CLC, Consed)

Annotation (SeattleSNP website,UCSC *Genome* Browser, Ensembl…)

GATK

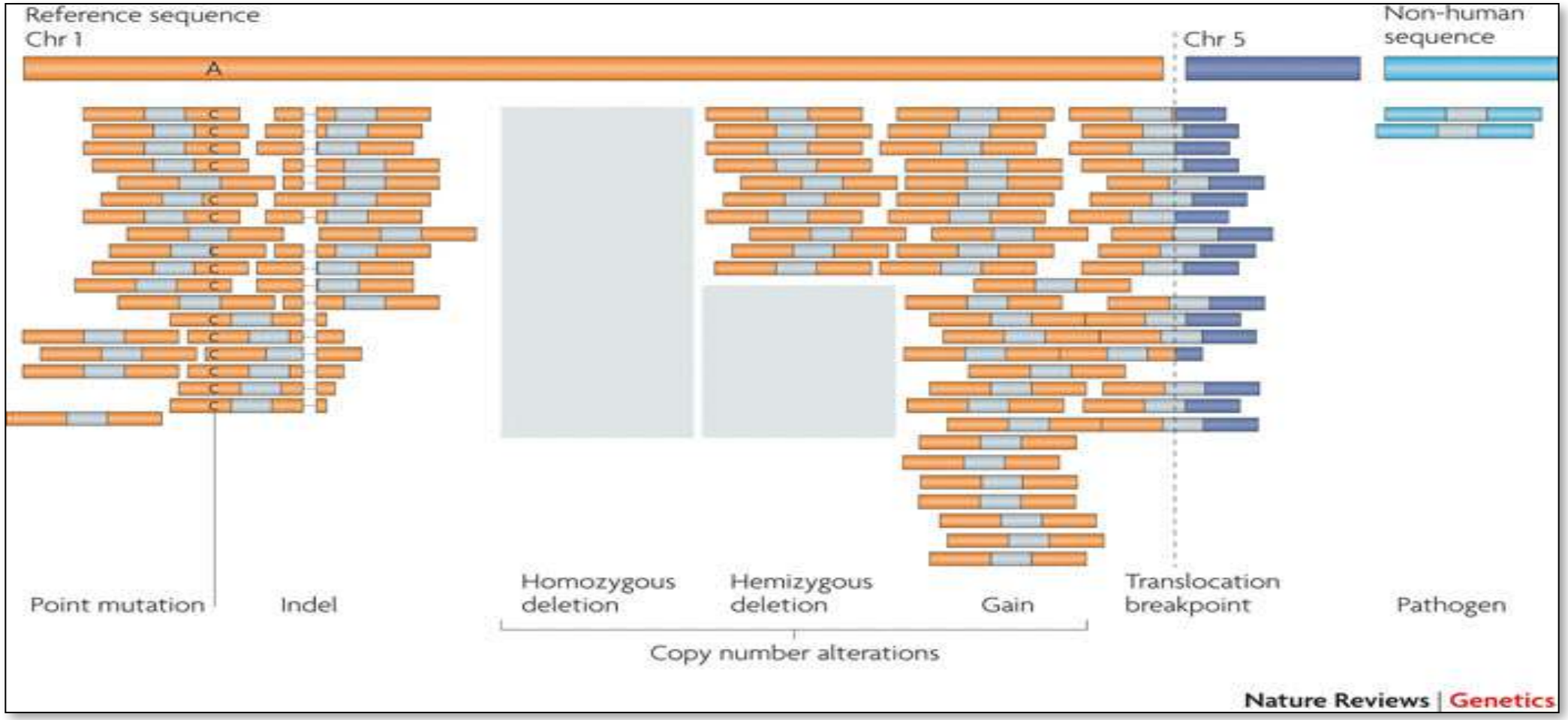
Genome Analysis Toolkit

Variant Discovery in High-Throughput Sequencing Data

Developed by the Data Science and Data Engineering group at the Broad Institute, the toolkit offers a wide variety of tools with a primary focus on variant discovery and genotyping. Its powerful processing engine and high-performance computing features make it capable of taking on projects of any size.

https://www.broadinstitute.org/gatk/

Genome alterations



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

Few Mutations Can Make a Big Difference

Different people have slightly different genomes:
on average, roughly 1 mutation in 1000 nucleotides.

The 1 in 1000 nucleotides difference accounts for height, high cholesterol susceptibility, and 1000s of genetic diseases.

CTGATGATGGACTACGCTACTACTGCTAGCTGTATTACGATCAGCTACC
ACATCGTAGCTACGATGCATTAGCAAGCTATCGATCGATCGATCGATTAT
CTACGATCGATCGATCGATCACTATACGAGCTACTACGTACGTACGATC
GCGGGACTATTATCGACTACAGATAAAACATGCTAGTACAACAGTATACA
TAGCTGCGGGATACGATTAGCTAATAGCTGACGATATCCGAT

CTGATGATGGACTACGCTACTACTGCTAGCTGTATTACGATCAGCTACA
ACATCGTAGCTACGATGCATTAGCAAGCTATCGATCGATCGATCGATTAT
CTACGATCGATCGATCGATCACTATACGAGCTACTACGTACGTACGATC
GCGTGACTATTATCGACTACAGATGAAACATGCTAGTACAACAGTATACA
TAGCTGCGGGATACGATTAGCTAATAGCTGACGATATCCGAT

Some challenges in cancer genome sequencing and analysis

Some genomic alterations in cancer are predominantly found at low frequency in clinical samples

Tools with low level of false-positive detections are needed

Normal tissue should be sequenced to identify rare somatic mutations

Whole genome sequence and de-novo assemblies are needed to analyze chromosomal rearrangements and somatic mutations in non-coding and not annotated regions (costly!)

Gapped reference for read alignment (genome finishing is in progress)

Cancer genomes are highly diverse and complex and vary significantly from normal genomes and from each other (mutation frequency, genome structure)

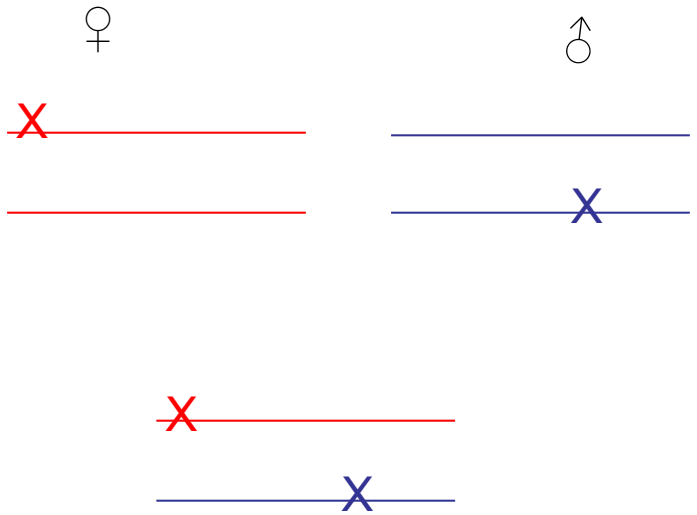
“Noisy” cancer DNA (non-tumor contamination; heterogeneity within the tumor)

Identification of Disease-Causing Mutations Using Exome-Wide Sequencing: Combined Immunodeficiency Disease (SCID)

Goal: identify candidate genes with disease-causing mutations

Samples: Mother, Father, affected child, unaffected siblings(2)

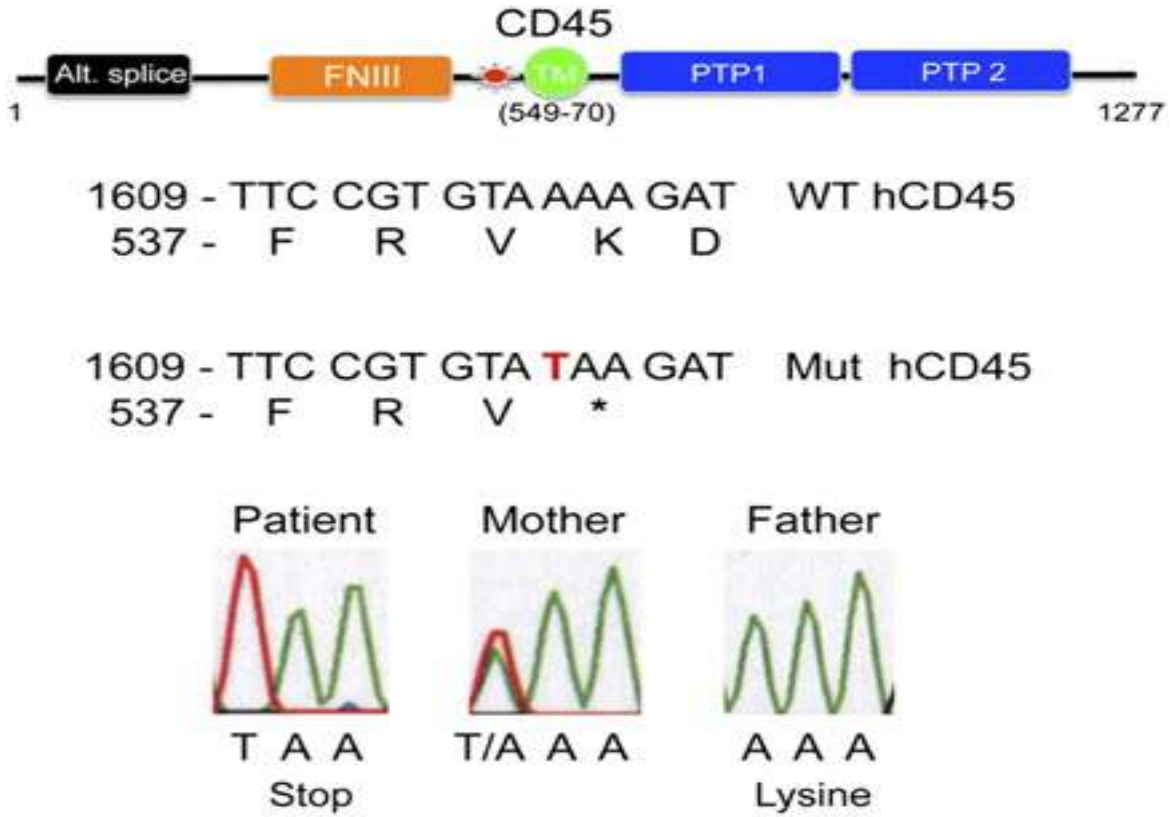
TOOLS: *Samtools, CLCBio, GATK* used to detect SNPs and Indels:



Gene (Sam/ GATK)
ANXA8L2
C7orf28B
FAM22D
FOXD4L1
FRG1
LRP2
MTUS1
OR52A1
PRAMEF4
SEZ6L

Gene (CLC)
CNTNAP3B
FAM22A
FAM22D
LOC653269
LRP2
PRAMEF4
TEKT4

Sequence analysis of the CD45 alleles of the patient and parents.

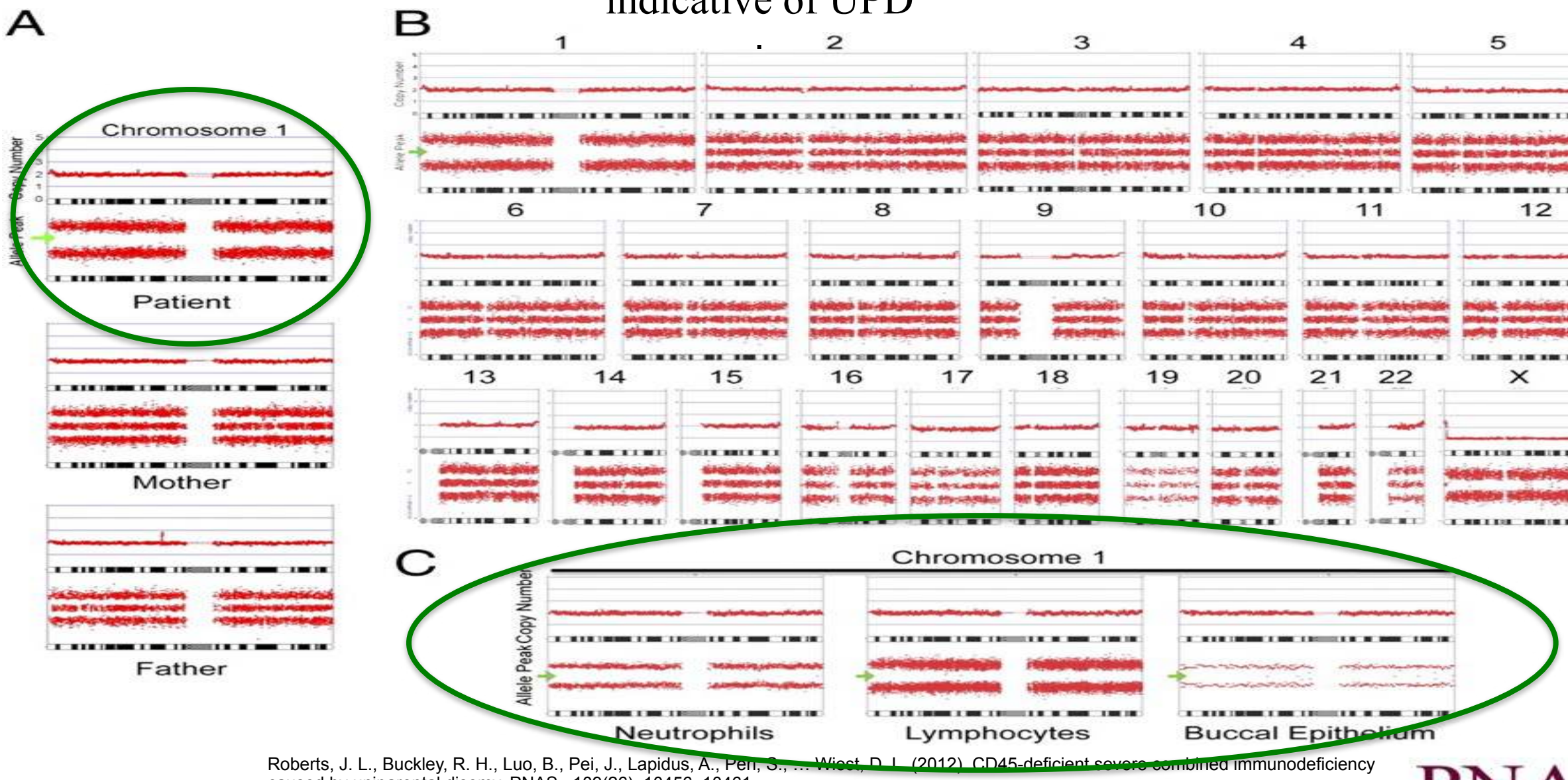


CD45 gene codes tyrosine phosphatase

SCID = Combined Immunodeficiency Disease.

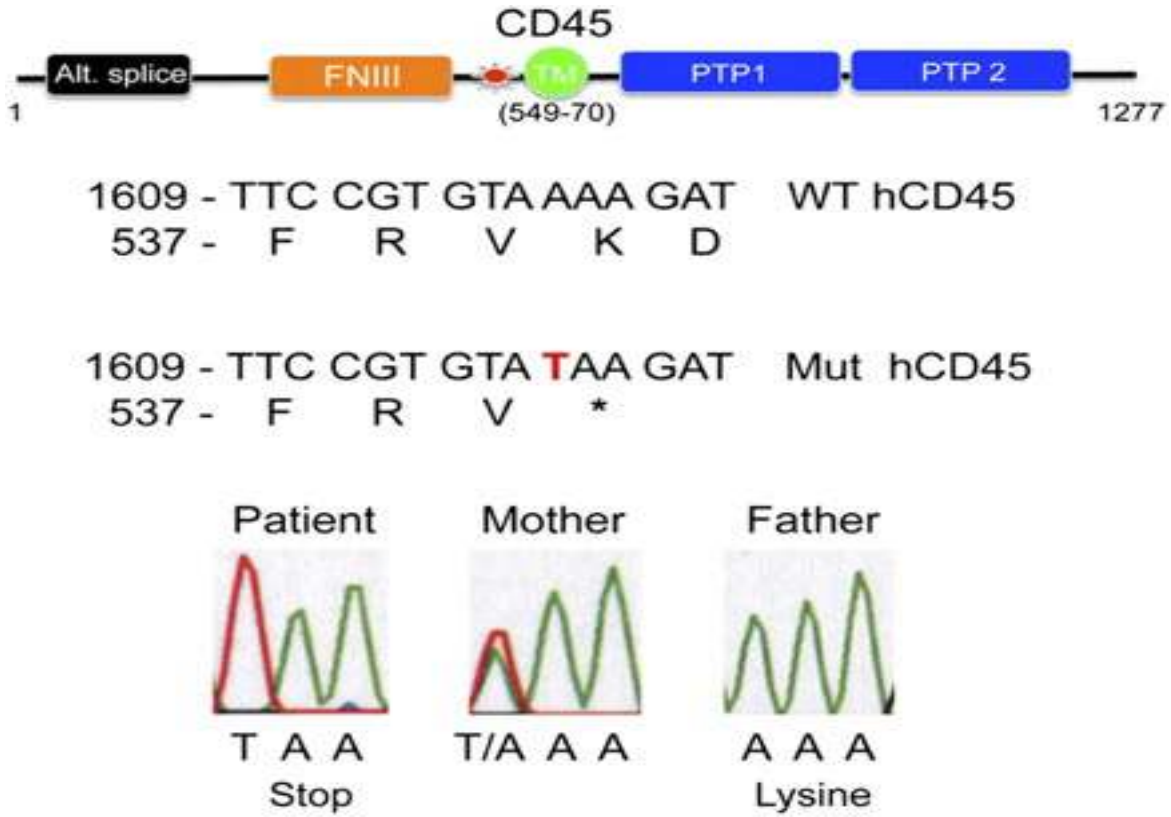
Roberts, J. L., Buckley, R. H., Luo, B., Pei, J., Lapidus, A., Peri, S., … Wiest, D. L. (2012). CD45-deficient severe combined immunodeficiency caused by uniparental disomy. PNAS, 109(26), 10456–10461.

The copy number panels reveal two copies of each gene on Chr1, but show only two bands, with loss of the Middle (heterozygous) band (arrow) across the entire chromosome, indicative of UPD



Roberts, J. L., Buckley, R. H., Luo, B., Pei, J., Lapidus, A., Peri, S., ... Wiest, D. L. (2012). CD45-deficient severe combined immunodeficiency caused by uniparental disomy. PNAS, 109(26), 10456–10461.

Sequence analysis of the CD45 alleles of the patient and parents



CD45 gene codes tyrosine phosphatase

SCID = Combined Immunodeficiency Disease.

Roberts, J. L., Buckley, R. H., Luo, B., Pei, J., Lapidus, A., Peri, S., … Wiest, D. L. (2012). CD45-deficient severe combined immunodeficiency caused by uniparental disomy. PNAS, 109(26), 10456–10461.

Search for new antibiotics

Устойчивость к антибиотикам из научной и клинической давно переросла в экономическую.
В год в отдельной стране число инфекций, вызванных устойчивыми к антибиотикам бактериями, исчисляется миллионами.
В одних только Соединенных Штатах 23 000 смертей/год

Современные методы секвенирования ДНК

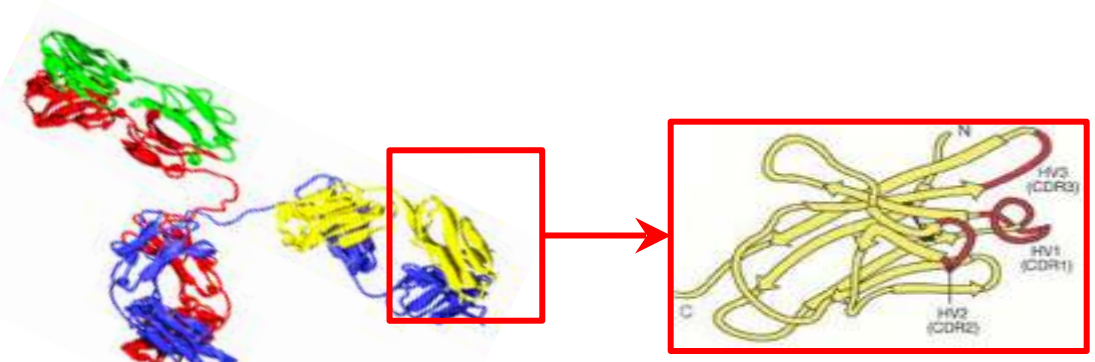
конкретные генетически е маркеры обнаружива ются в течение нескольких часов/дней

устанавлива ются прямые связи с источниками вспышек и помогают отслеживать их распростран ение

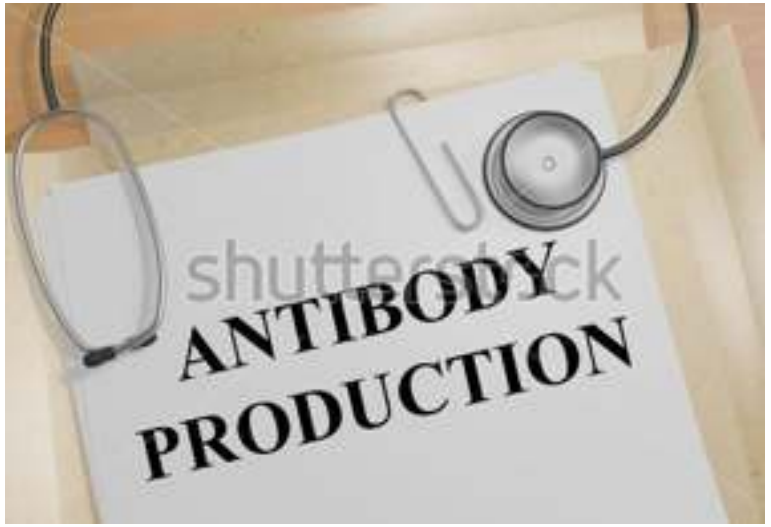
- Острая необходимость в инновационных методах диагностики
- ПОТРЕБНОСТЬ В СОЗДАНИИ АНТИБИОТИКОВ НОВОГО ПОКОЛЕНИЯ

Antibody drugs

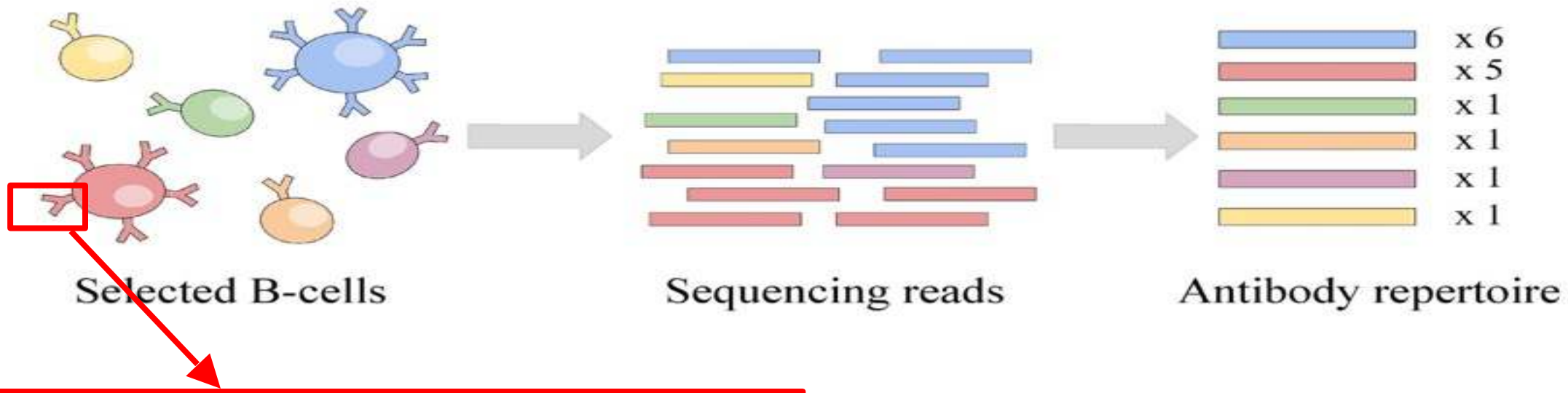
Antibodies are produced by immune system to neutralize antigens

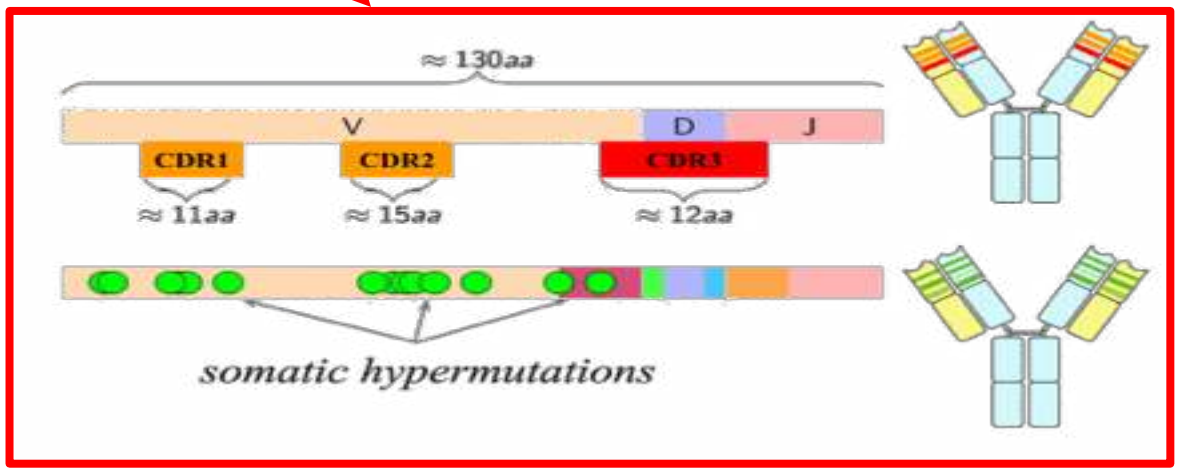


In 2013, global sale revenue for antibody drugs was $75 billion



Antibody repertoire





Construction of antibody repertoire from NGS data and its analysis are important steps in design of antibody drugs and clinical studies

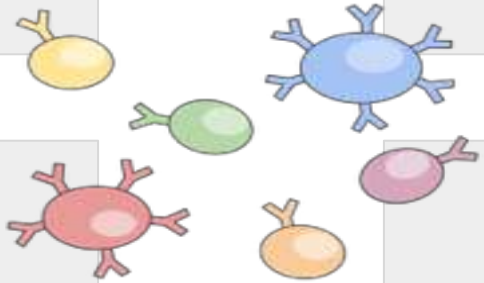
Immunoinformatics projects (SPbU – Center for Algorithmic Biotechnology)

IgRepertoireConstructor

tool for construction of antibody repertoire and using mass spectra

IgAnalyzer & IgQUAST

tools for “quality” assessment of antibody repertoire



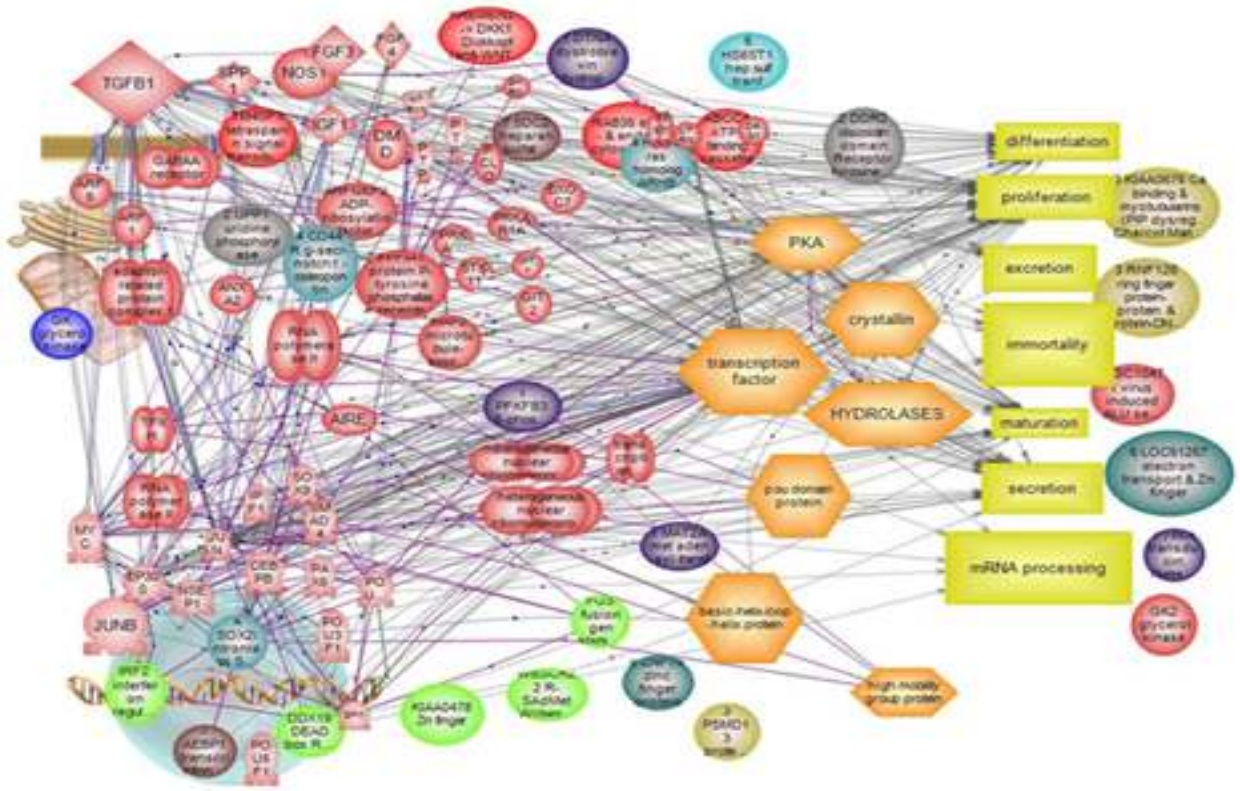
IgSimulator

versatile immunosequencing simulator

AntEvolo

tool for construction of clonal trees and analysis of SHMs

To improve drug discovery we need to discover (read "develop") efficient bioinformatics algorithms and approaches for



target identification

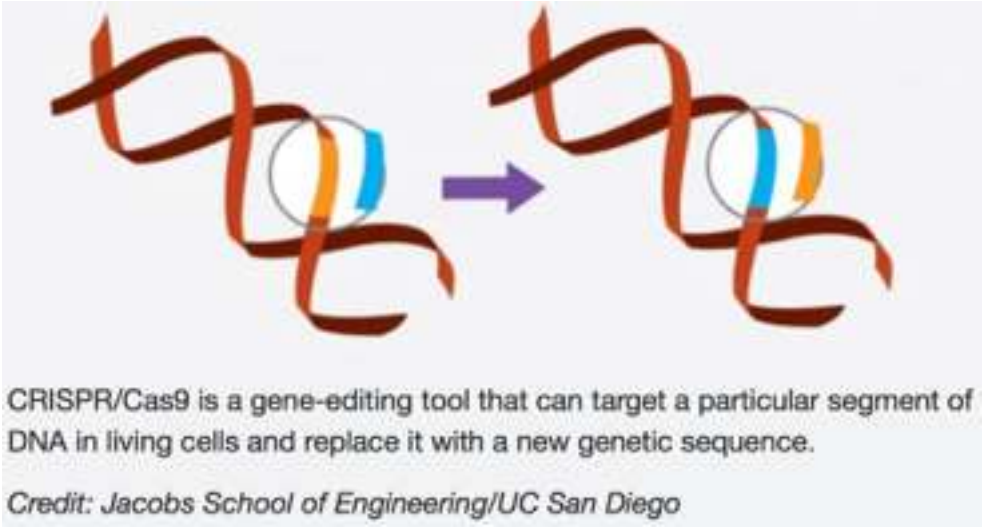
target validation

lead identification

lead optimization

Gene editing – new path to gene therapy

"**CRISPR** (*Clustered Regularly Interspaced Short Palindromic Repeats)* and its associated protein, **Cas9**, provide sequence-specific adaptive immunity in bacteria and archaea by integrating short viral DNA sequences in the host cell genome, allowing the cell to remember, recognize, and clear infections"



CRISPR/Cas9 is a gene-editing tool that can target a particular segment of DNA in living cells and replace it with a new genetic sequence.

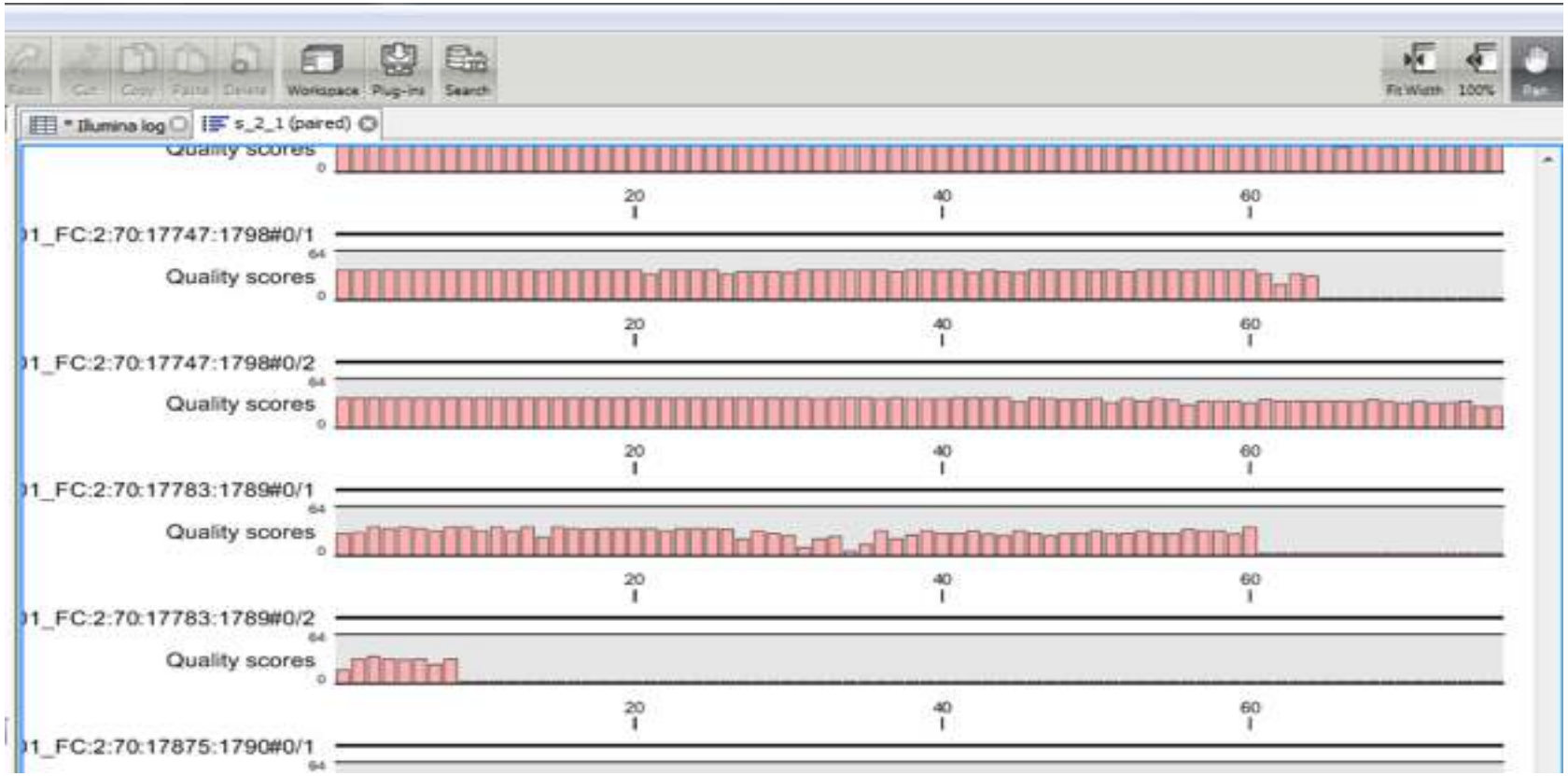
Credit: Jacobs School of Engineering/UC San Diego

sgRNA Scorer - an interactive software for users to find guide RNAs that are predicted to be highly specific and highly active for their gene targets

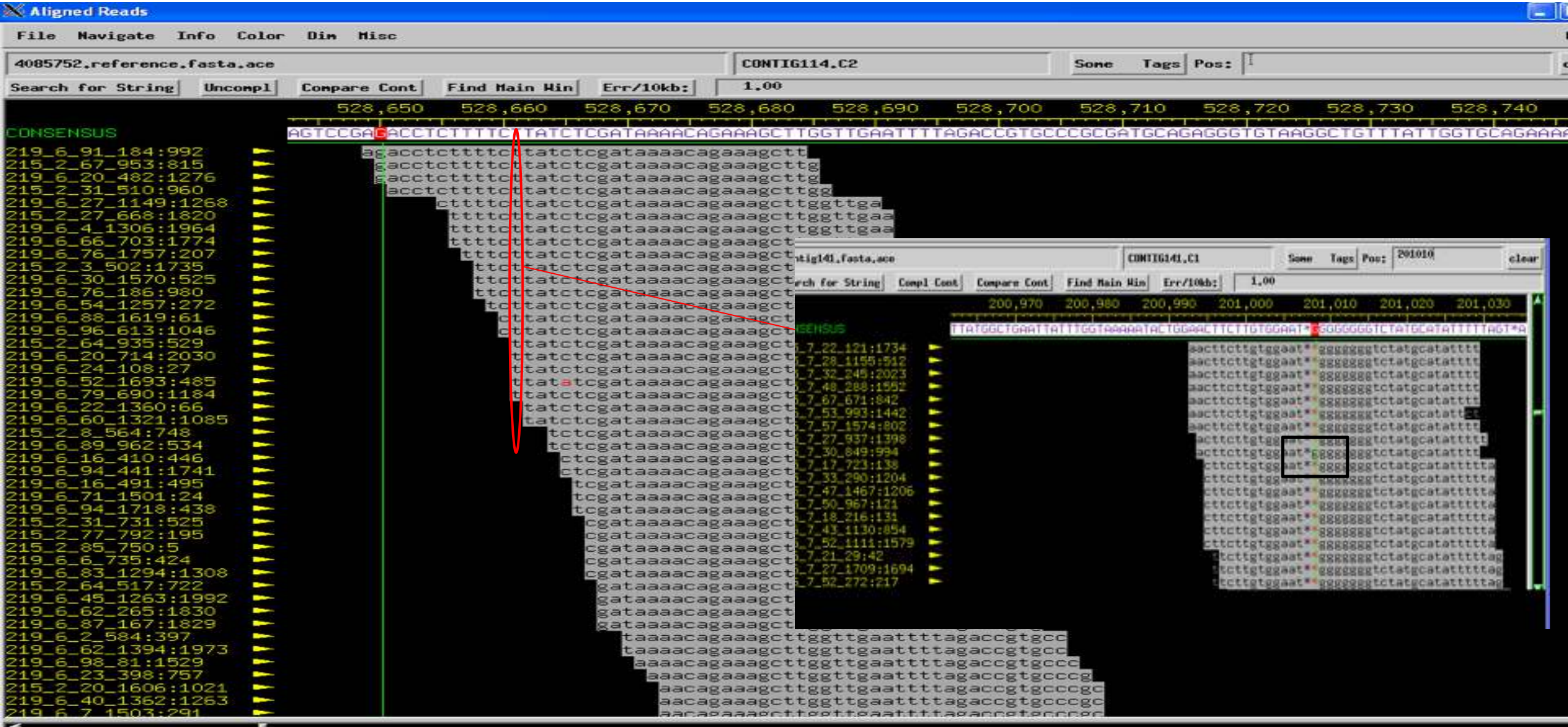
http://biotechconnection-losangeles.org/2015/07/01/crispr-cas9-genome-editing-technology-part-i-mechanism-applications-and-limitations/

QC: visualization

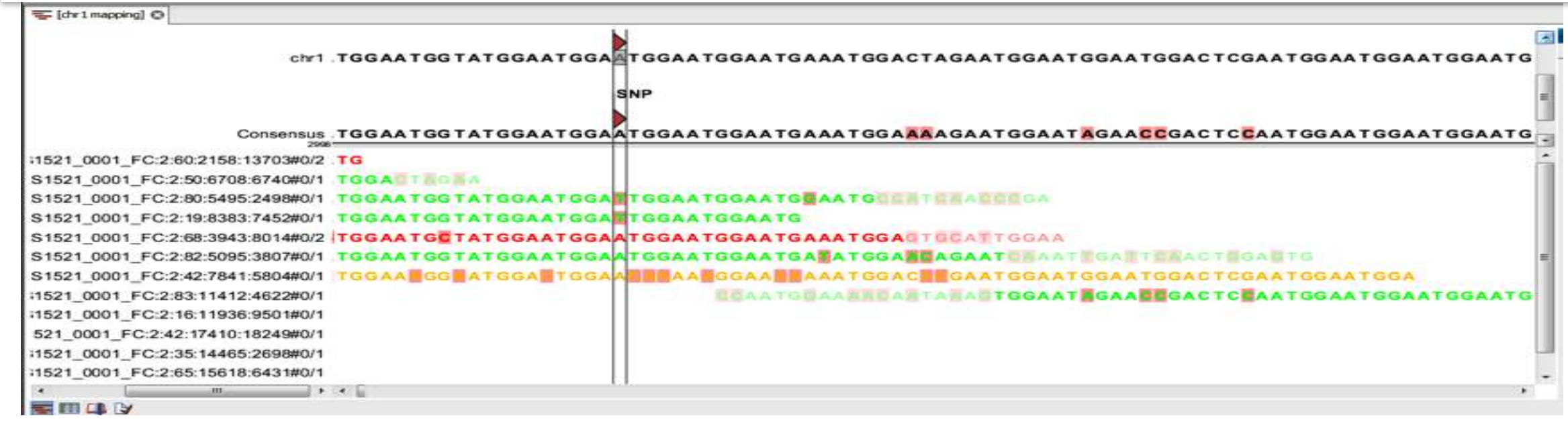
CLC



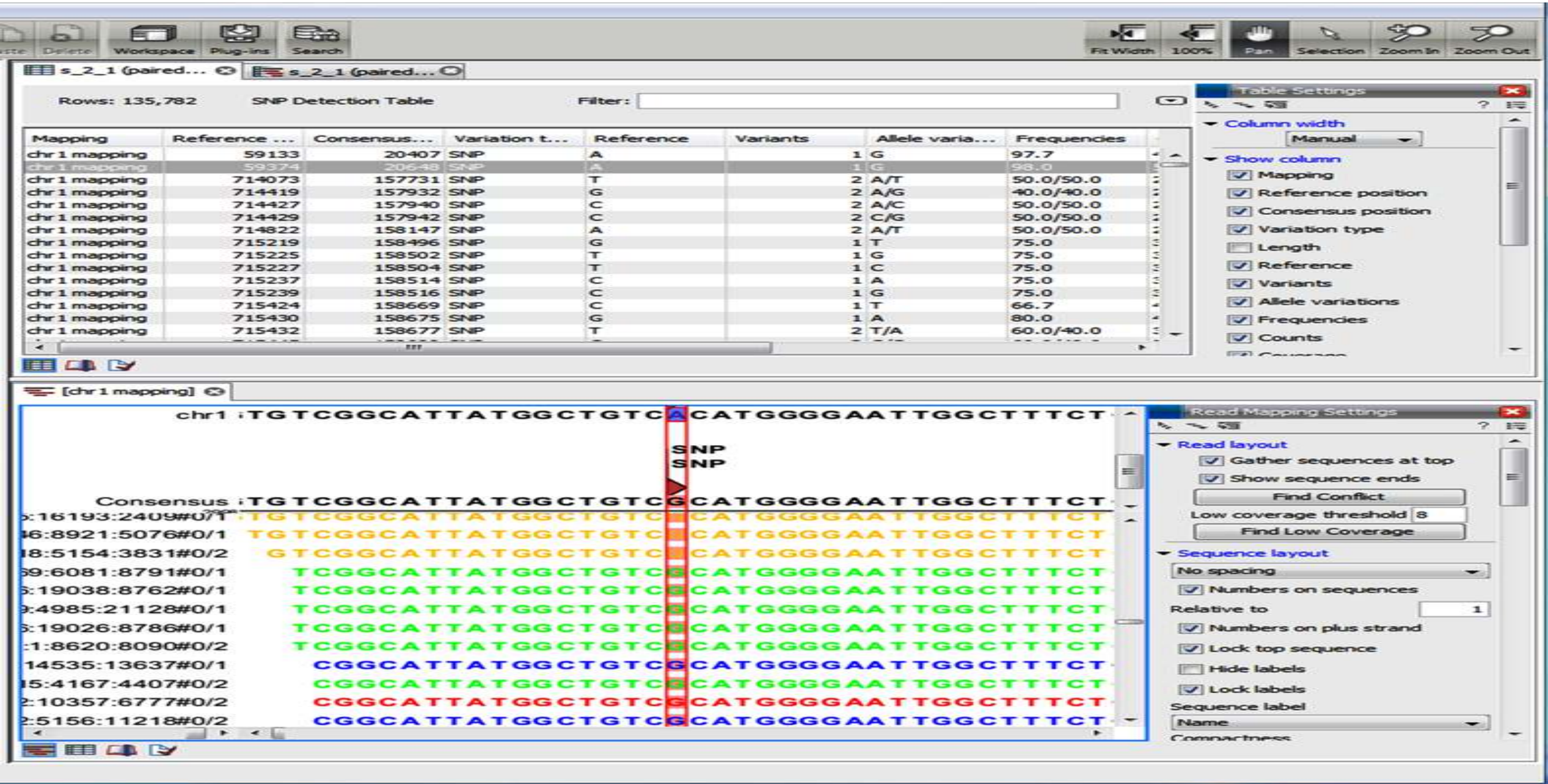
QC: duplications



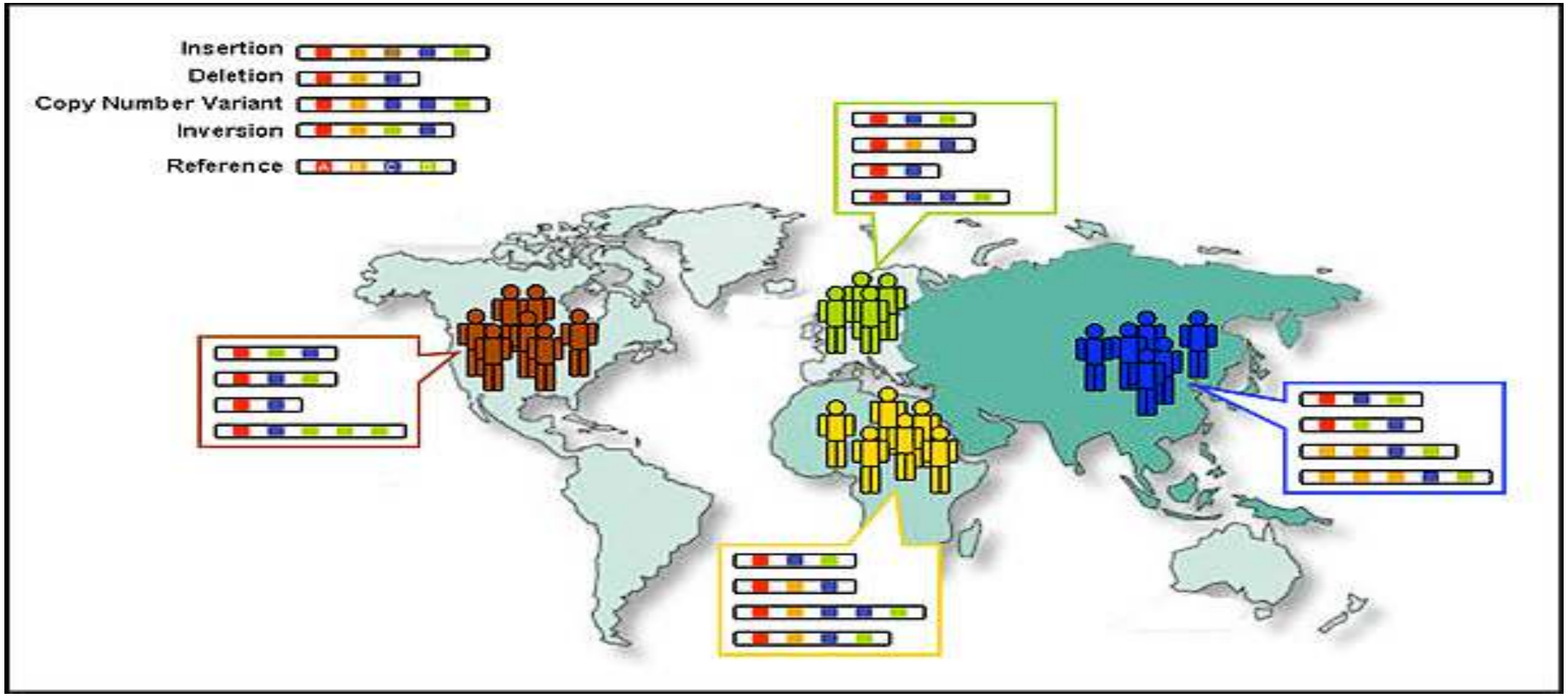
QC: Alignment quality = f (read quality)



SNP detection



1000 Genomes Project



Changes in the number and order of genes (A-D) create genetic diversity within and between populations.

https://en.wikipedia.org/wiki/1000_Genomes_Project

Sequence of Personal Genomes?

2010:

Nicholas Volker became the first human being to be saved by genome sequencing.

Doctors could not diagnose his condition; he went through dozens of surgeries.

Sequencing revealed a rare mutation in a *XIAP* gene linked to a defect in his immune system.

This led doctors to use immunotherapy, which saved the child.

Bottlenecks in bioinformatics

- education of biologists in the use of advanced computing tools
- the recruitment of computer scientists into the field of bioinformatics
- the limited availability of developed databases of biological information
- need for efficient and intelligent search engines for complex databases
- need for innovative NGS pipelines

Key segments of the Bioinformatics market by:

Platforms, Tools, and Services

Platforms

- ✓ Sequence Manipulation
- ✓ Sequence Alignment
- ✓ Structural Analysis
- ✓ Sequence Analysis

Tools

- ✓ General Knowledge
- ✓ Specialized Knowledge

Services

- ✓ Data analysis
- ✓ Sequencing services
- ✓ Database & management
- ✓ Other services

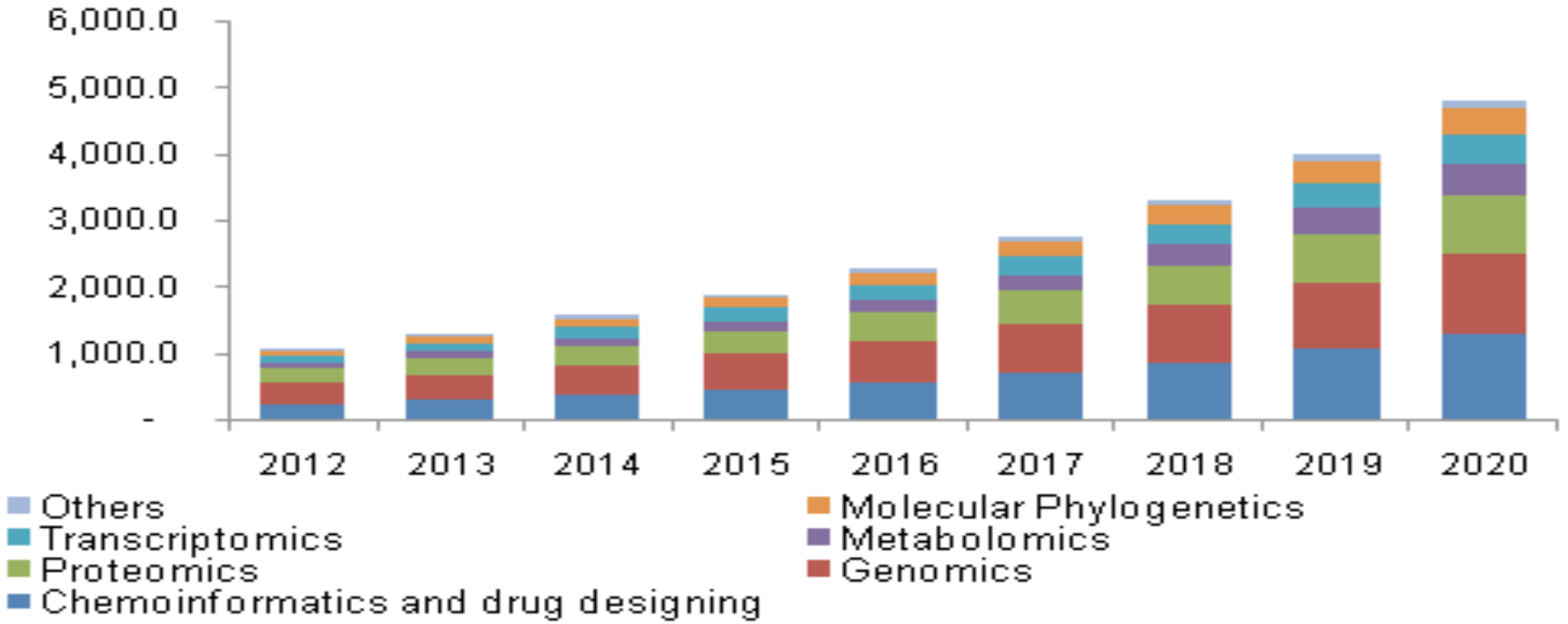
Applications

- ✓ **Preventive medicine**
- ✓ **Molecular medicine**
- ✓ **Gene therapy**
- ✓ **Drug development**
- ✓ **Others**

Geography

- ✓ North America (35%)
- ✓ Europe
- ✓ Asia-Pacific
- ✓ Rest of the World

European bioinformatics market, by product, 2012-2020 (USD Million)



Application Insights

Genomic was the leading applications in 2013 accounting for approximately USD 1.0 billion of the global revenue. Growing demand for pharmacogenomics in drug development and sequence screening and introduction of technological advancements aimed at managing large sets of genomic data are key factors accounting for this large share.

Chemo-informatics is expected to gain market share and grow at a CAGR of 23.5% during the forecast period wherein increasing demand for biomarker discovery and development is the primary reason accounting for this expected growth over the forecast period.

Proteomics has experienced extensive R&D investment and bioinformatics plays an integral role in analyzing and managing the resultant data. Also, this makes it simpler to handle heterogeneous and large data sets, introduce new algorithms and improve the knowledge discovery procedure.

http://www.grandviewresearch.com/industry-analysis/bioinformatics-industry

THANK YOU!

piterlabs@gmail.com