Clinical Bioinformatics

Concepts & Practice

Dr. Vivek Gopalan
Director, Bioinformatics and Data Engineering
ClairLabs.ai - https://clairlabs.ai/

About me – Vivek Gopalan

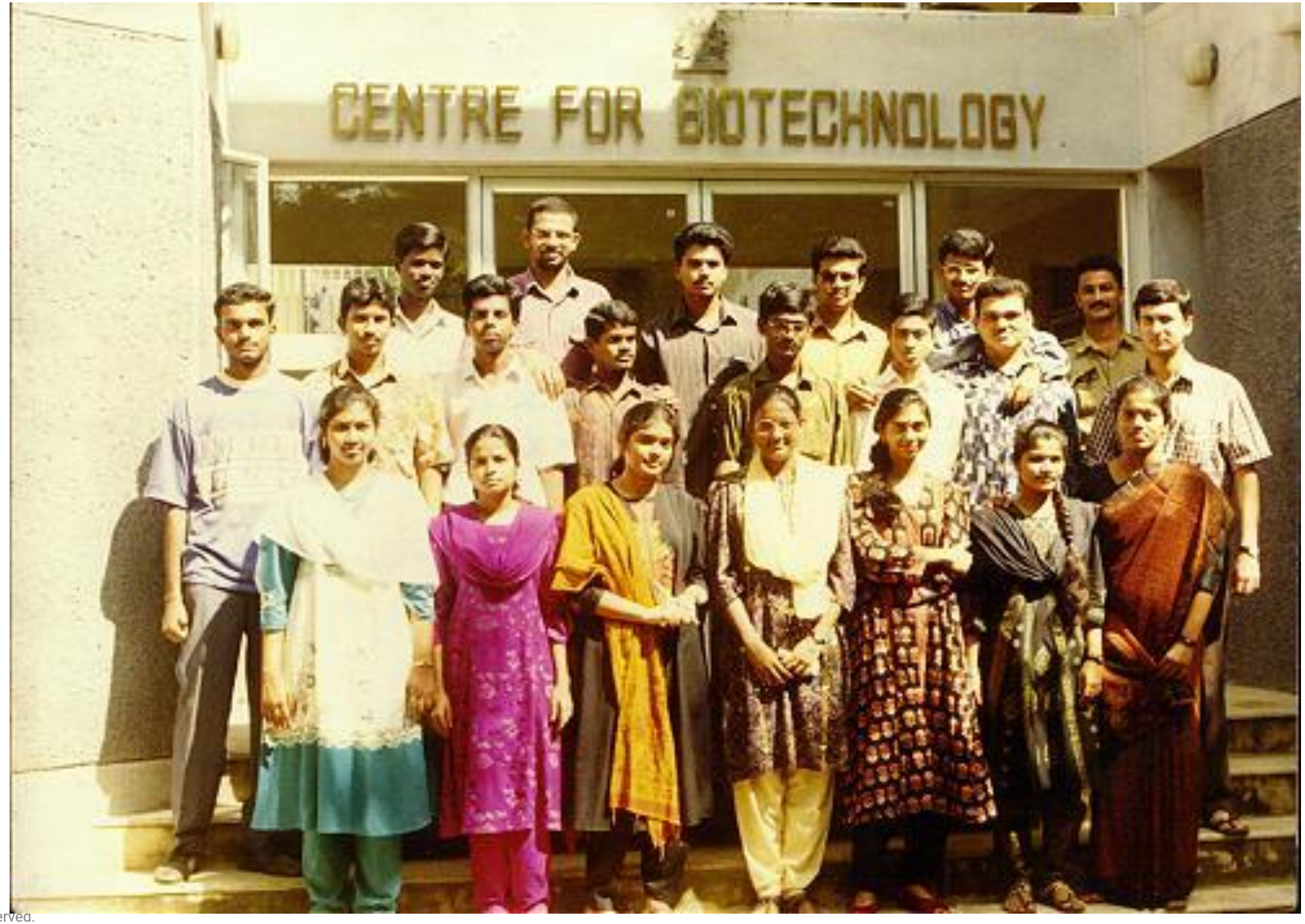
- Working and managing Bioinformatics projects since 2001..
- **Director, Bioinformatics and Data Engineering, ClairLabs AI** since Oct 2025
- ***Senior Manager, Novo Nordisk, India*** *from Sept 2021 till May 2025*
- ***Director, Clinical Bioinformatics and Clinical Software*** at MedGenome Labs Ltd., India for 5.5 years
- ***Sr. Bioinformatics Scientific Staff*** at Qiagen Silicon Valley for 3 years
- ***Bioinformatics Lead*** at Lockheed HLS R&D team for 2 years
- ***Scientific Infrastructure Team Lead*** (contractor) at BCBB/NIAID/NIH for 5 years
- ***Research Associate*** at University Of Maryland and performed **Postdoctoral research** on Computational Molecular Evolution (2005 to 2007)
- ***PhD (Bioinformatics)*** at National University of Singapore, Singapore (2001 to 2005)
- As ***Research Scientist*** at Vittal Mallya Scientific Research Foundation (VMSRF), Bangalore, India (1997 to 2001)
- ***B.Tech.(Industrial Biotechnology)*** at Anna University (AC tech Campus), India (1993 to 1997)

NIH Director Award

NIAID Merit Award

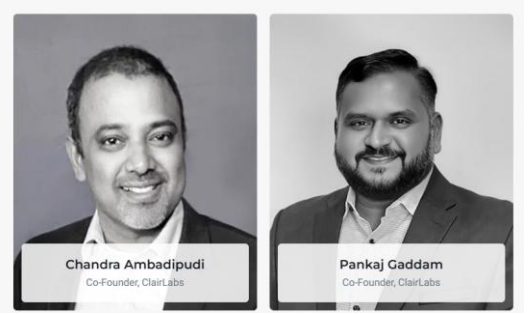
Partner for the Start-up company SeekQuence Inc (Acquired by DoctorForMe in 2018, I got 5% of the total asset). Worked in stealth mode with [Dr. Wei Zhou] for 3 years.

1993 – CBT batch



About ClairLabs.ai

Advancing science with AI and Data Innovation



- Launched in 2024
- Specialized in ***lab information management systems, data management, and bioinformatics solutions*** for genomic diagnostics
  - Cloud and AI technologies

Vision

Clairlabs' vision is to be a leading provider of technology-driven solutions that accelerate scientific research in the life sciences industry. The company aims to leverage the power of artificial intelligence (AI) and a global talent pool to empower life sciences companies to bring innovative treatments and therapies to market faster.

Mission

- **AI-powered platforms and services** drug discovery and clinical trials.
- Enable strategic **Global Competency Center** and innovation hubs in shared services models.
- Foster a c**ulture of innovation, collaboration, and excellence**, with talent thriving on advancement of scientific research.

https://clairlabs.ai/

Genomic diagnostics assay

Why genetic assay was not available 20 years back?

Clinical diagnostics assay

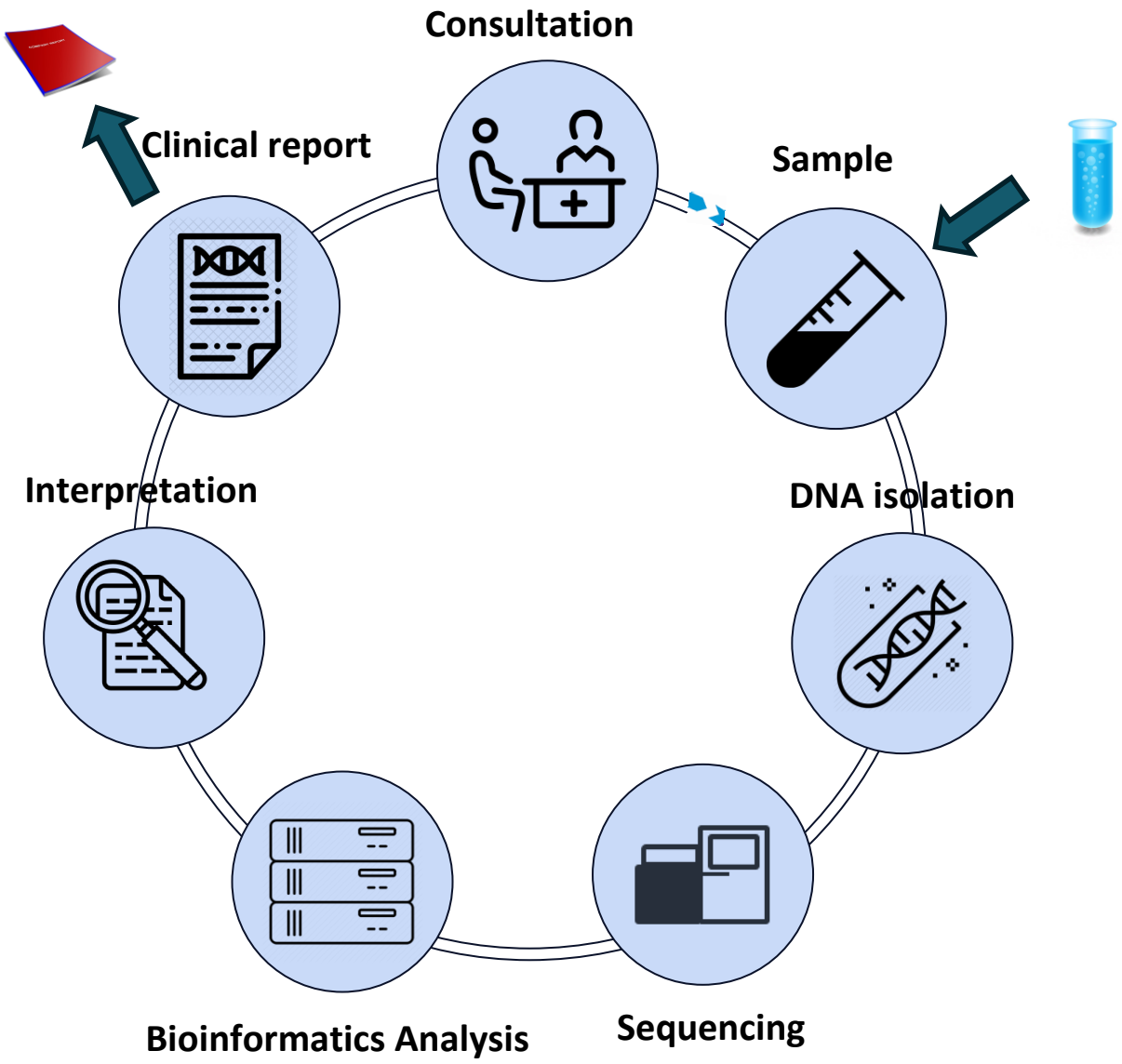
- “Qualitatively assess or quantitatively measure the presence, amount, or functional activity of a target entity (the analyte) in the patients sample”
  - Blood, Saliva, Urine
- Aids doctors
  - To differentially diagnosis the disease for the patients
  - To monitor the disease progress for the patients
- The final result is a one or two pages report for the doctors to interpret
  - Contains actual value of the analyte, reference range, interpretation

Genomic or Genetic Diagnostics assay

Type of Diagnostics assay: Molecular Diagnostics Assay

“Simultaneous sequencing of thousands of genes and quantitatively identifying the underlying molecular level defects (biomarkers) that are associated with the hereditary diseases”

Genetic diagnostics assay



Clinical sample

ABC Diagnostics Lab

Patient details

Name: Mr. Z	**Age:** 15 years	**Sex:** Male
Date and Time Collected: 10th Feb 2018 7:55am	**Physicians name:** Dr. Y	**Fasting:** Yes

Clinical Symptoms

Jaundice, history of abdominal distension, younger brother died due to liver disease.

Results

PATHOGENIC VARIANT CAUSATIVE OF THE REPORTED PHENOTYPE WAS IDENTIFIED

Gene (Transcript)	Location	Variant	Zygosity	Disease (OMIM)	Inheritance	Classification
ATP7B (-) (ENST00000242839)	Exon 13	c.2997dupC **(p.Gly1000ArgfsTer28)**	Homozygous	Wilson disease	Autosomal recessive	Pathogenic

The *ATP7B* gene is 100% covered in this assay.

Variant interpretation and Clinical Correlation

A homozygous single base pair duplication in exon 13 of the ***ATP7B*** gene **(chr13:52520483dupG; Depth: 100x)** that results in a frameshift and premature truncation of the protein 28 amino acids

Clinical report
(3 weeks time)

PATHOGENIC VARIANT CAUSATIVE OF THE REPORTED PHENOTYPE WAS IDENTIFIED

Gene (Transcript)	Location	Variant	Zygosity	Disease (OMIM)	Inheritance	Classification
ATP7B (-) (ENST00000242839)	Exon 13	c.2997dupC **(p.Gly1000ArgfsTer28)**	Homozygous	Wilson disease	Autosomal recessive	Pathogenic

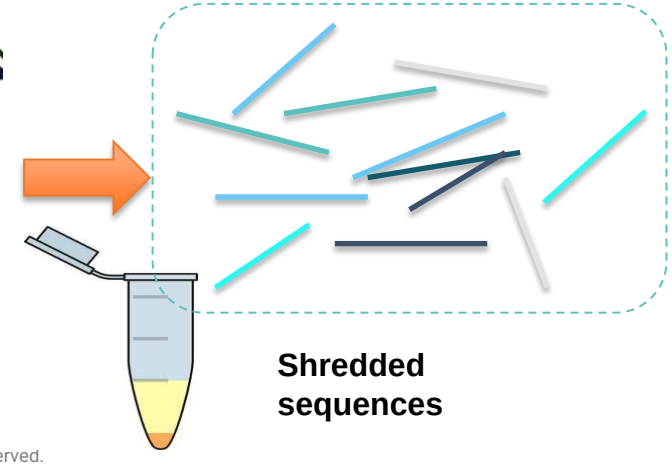
Driving force of Genomic diagnostics assay

Why genetic assay was not available 20 years back?

Driving force for genetic diagnostics tests – 1. Technology

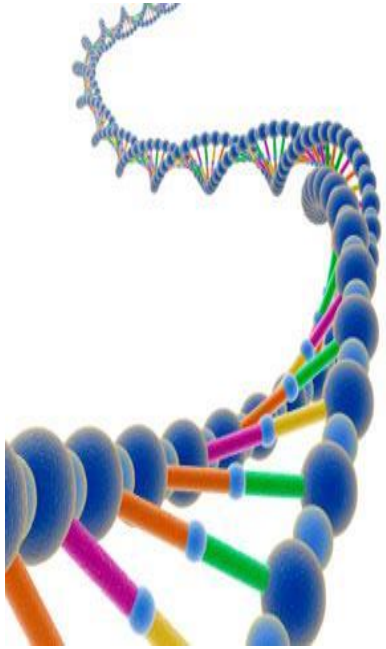
- **N**ext **G**eneration **S**equencing Technology
  - High throughput genome sequencing technology that breaks the genomes in small pieces and **accurately** sequence the fragments in **massively parallel reactions**
  - First introduced in 2004 by Roche Pharma. Currently Illumina's & Life technologies sequencers are widely used

Genomic Sequence



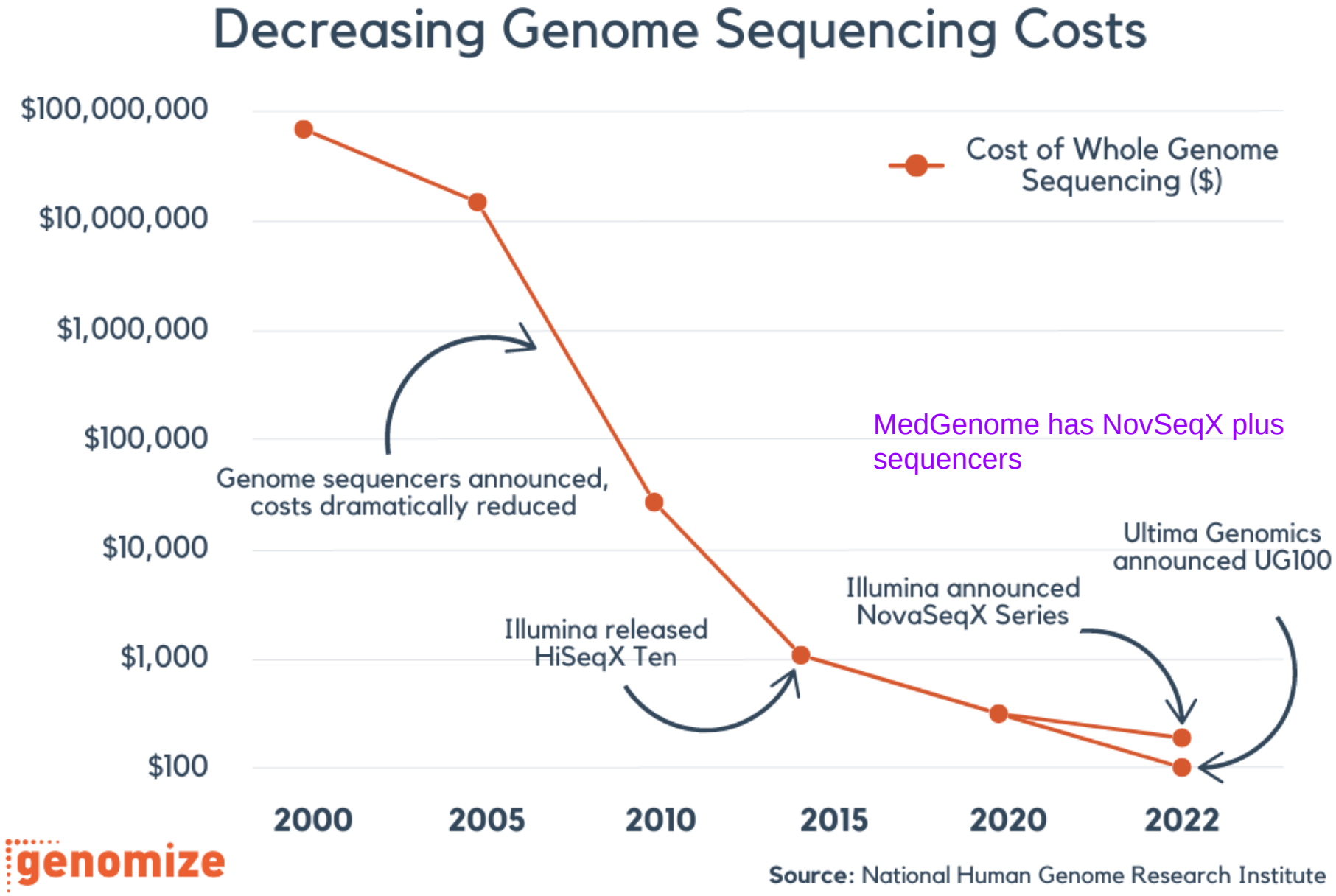
Shredded sequences

NGS sequencing - Massively parallel sequencing



ATGCGATAAG
ATGCGATAAG
GAGGCGATAAG
GTGCCCTAAG
TTGAAATAAG

Driving force for genetic diagnostic test – 2. Affordability



https://genomize.com/genome-sequencing-costs-effects-on-clinical-genetics/

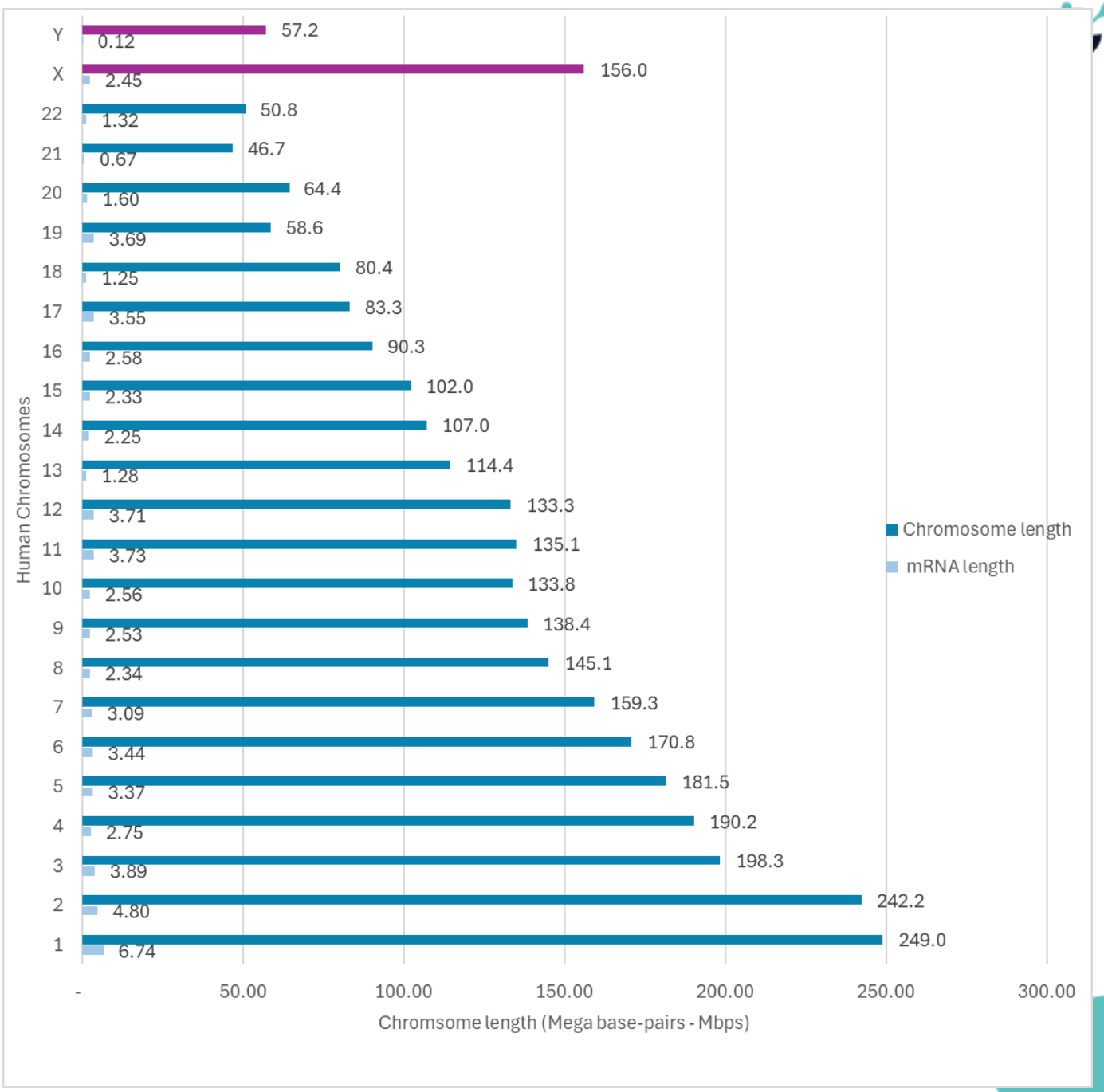
NGS sequencing - NovaSeq X plus

Flow cell type	25B (Max capacity)
No. of flow cells per run	2
Sequencing type	2 x 150 bps
Max amount of DNA sequenced	~16 Terra bps in 48 hours (16000 Gbps)
Human genomes (30X)	2 x 64
Exomes (100X)	2 x 750
Clinical Exomes per flow cell	2 x ~1500

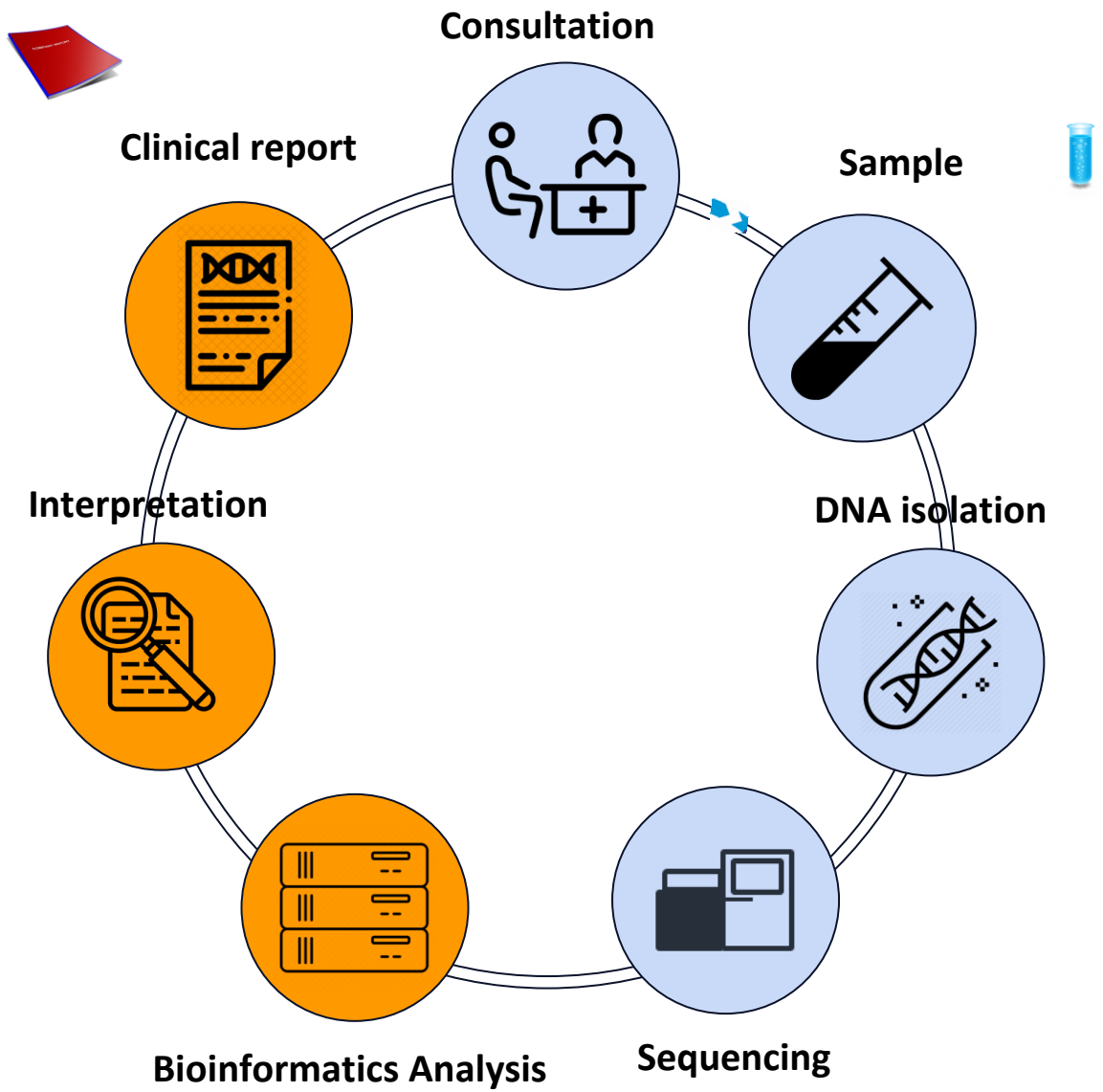
16,000,000,000,000 bases per run

Human Genome, whole Exome, Clinical Exome

- Whole Genome
  - 23 pairs of chromosomes
    - 22 pairs autosomes
    - XY - Male
    - XX - Female
  - 3 billion base-pairs of DNA - 3000 Mbps
- Whole Exome
  - All protein coding genes : ~20,000 genes
  - 60 Mbps (2% of Whole Genome) of exons
  - Represents ~60K protein
- Clinical Exome
  - All genes that are associated with known rare diseases : ~6K genes
  - 30 Mbps (1% of the Whole genome) of exons
  - Represents ~20K proteins



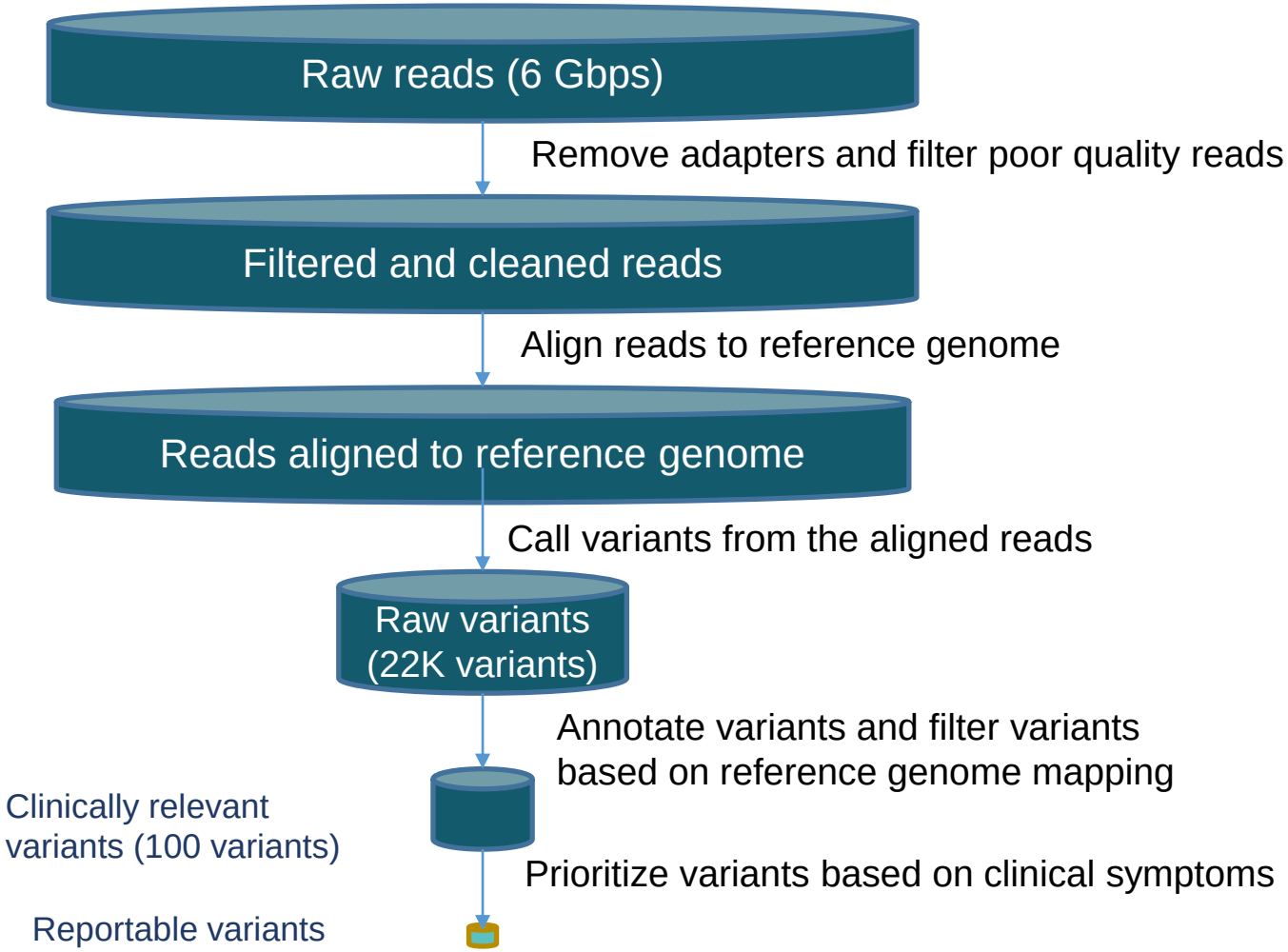
What is Clinical bioinformatics

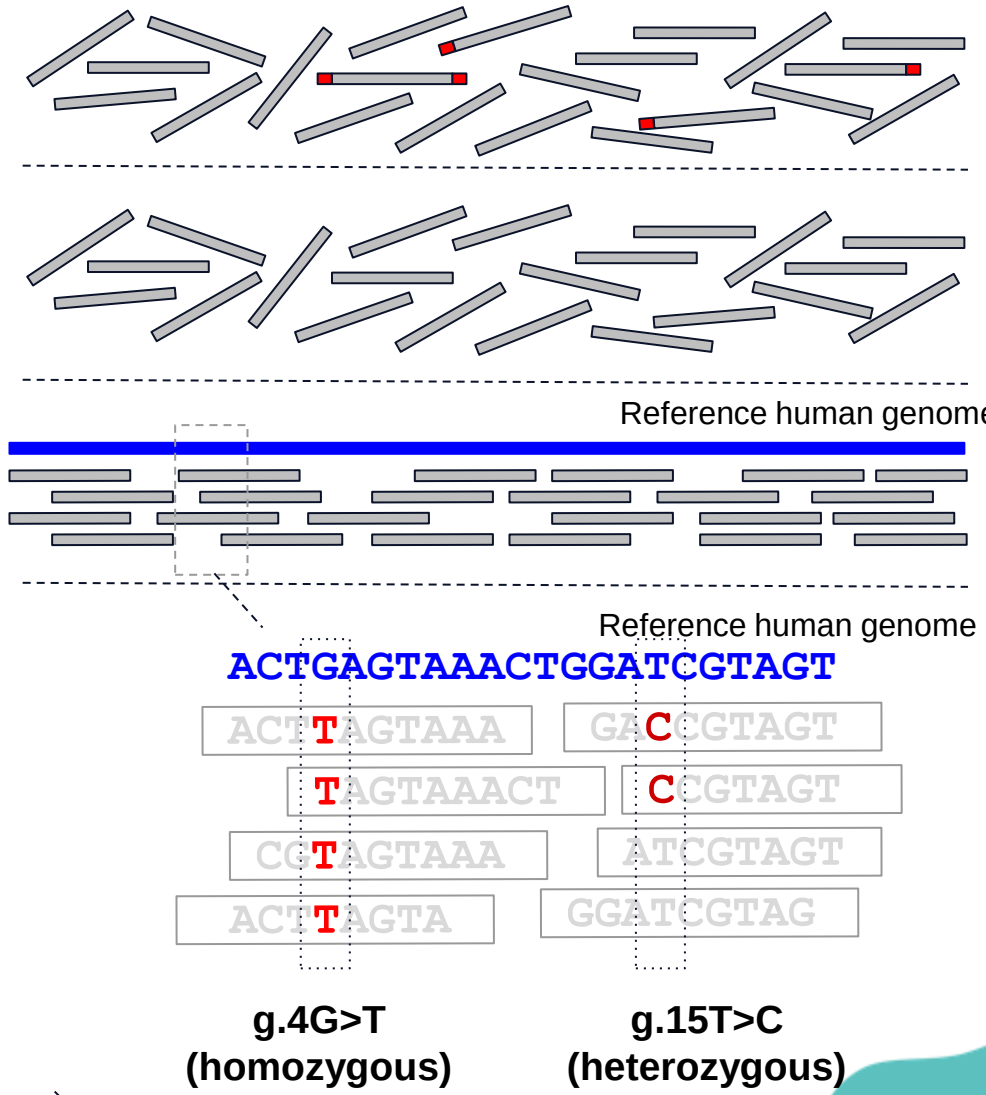


Clinical dry lab process

- Informatics component of the NGS or Genetic Diagnostic assay

Clinical Bioinformatics workflow - principle





Clinical report guidelines

1. ACMG guidelines
   a. Standard for Genetic diagnostics report
   b. Rules for associating variants to genetic diseases
2. ACMG Variant classification - based on quantitative principles
   a. Pathogenic
   b. Likely Pathogenic
   c. Variant of Uncertain Significance (VUS)
   d. Likely Benign
   e. Benign

Pathogenic (P)	Causative for clinical phenotype
Likely pathogenic (LP)	Probably causative for clinical pheno
Uncertain significance (VUS)	Unknown effect on clinical phenotyp
Likely benign (LB)	Probably not causative for clinical ph
Benign (B)	Not causative for clinical phenotype
Pathogenic with family data (PF)	Evidence shows causative for clinica insertion/deletion variants)

ACMG STANDARDS AND GUIDELINES | Genetics inMedicine

Standards and guidelines for the interpretation of sequence variants: a joint consensus recommendation of the American College of Medical Genetics and Genomics and the Association for Molecular Pathology

Sue Richards, PhD[1], Nazneen Aziz, PhD[2,16], Sherri Bale, PhD[3], David Bick, MD[4], Soma Das, PhD[5], Julie Gastier-Foster, PhD[6,7,8], Wayne W. Grody, MD, PhD[9,10,11], Madhuri Hegde, PhD[12], Elaine Lyon, PhD[13], Elaine Spector, PhD[14], Karl Voelkerding, MD[13] and Heidi L. Rehm, PhD[15]; on behalf of the ACMG Laboratory Quality Assurance Committee

	Benign		Pathogenic			
	Strong	**Supporting**	**Supporting**	**Moderate**	**Strong**	**Very Strong**
Population Data	MAF is too high for disorder *BA1/BS1* **OR** observation in controls inconsistent with disease penetrance *BS2*			Absent in population databases *PM2*	Prevalence in affecteds statistically increased over controls *PS4*	
Computational And Predictive Data		Multiple lines of computational evidence suggest no impact on gene /gene product *BP4* Missense in gene where only truncating cause disease *BP1* Silent variant with non predicted splice impact *BP7*	Multiple lines of computational evidence support a deleterious effect on the gene /gene product *PP3*	Novel missense change at an amino acid residue where a different pathogenic missense change has been seen before *PM5* Protein length changing variant *PM4*	Same amino acid change as an established pathogenic variant *PS1*	Predicted null variant in a gene where LOF is a known mechanism of disease *PVS1*
Functional Data	Well-established functional studies show no deleterious effect *BS3*		Missense in gene with low rate of benign missense variants and path. missenses common *PP2*	Mutational hot spot or well-studied functional domain without benign variation *PM1*	Well-established functional studies show a deleterious effect *PS3*	
Segregation Data	Non-segregation with disease *BS4*		Co-segregation with disease in multiple affected family members *PP1*	Increased segregation data →		
De novo Data				*De novo* (without paternity & maternity confirmed) *PM6*	*De novo* (paternity & maternity confirmed) *PS2*	
Allelic Data		Observed in *trans* with a dominant variant *BP2* Observed in *cis* with a pathogenic variant *BP2*		For recessive disorders, detected in *trans* with a pathogenic variant *PM3*		
Other Database		Reputable source w/out shared data = benign *BP6*	Reputable source = pathogenic *PP5*			
Other Data		Found in case with an alternate cause *BP5*	Patient's phenotype or FH highly specific for gene *PP4*			

Good practise pipeline link: https://gatk.broadinstitute.org/hc/en-us

Clinical bioinfo pipeline steps (Germline)

Directly stored NFS

Demux *bcl2fastq*

Adapter trimming *fastmcq*

Alignment to reference *BWA*

Duplicate removal *Picard, GATK*

Realignment removal *GATK,*

Recalibration *GATK*

Variant Calling *GATK*

Variant annotation *SNPeff/VEP*

SV & CNV *BreakDancer, Pindel, Contra,*

Variant annotation *Custom scripts*

Demux metrics *Custom scripts*

Read metrics *Custom scripts*

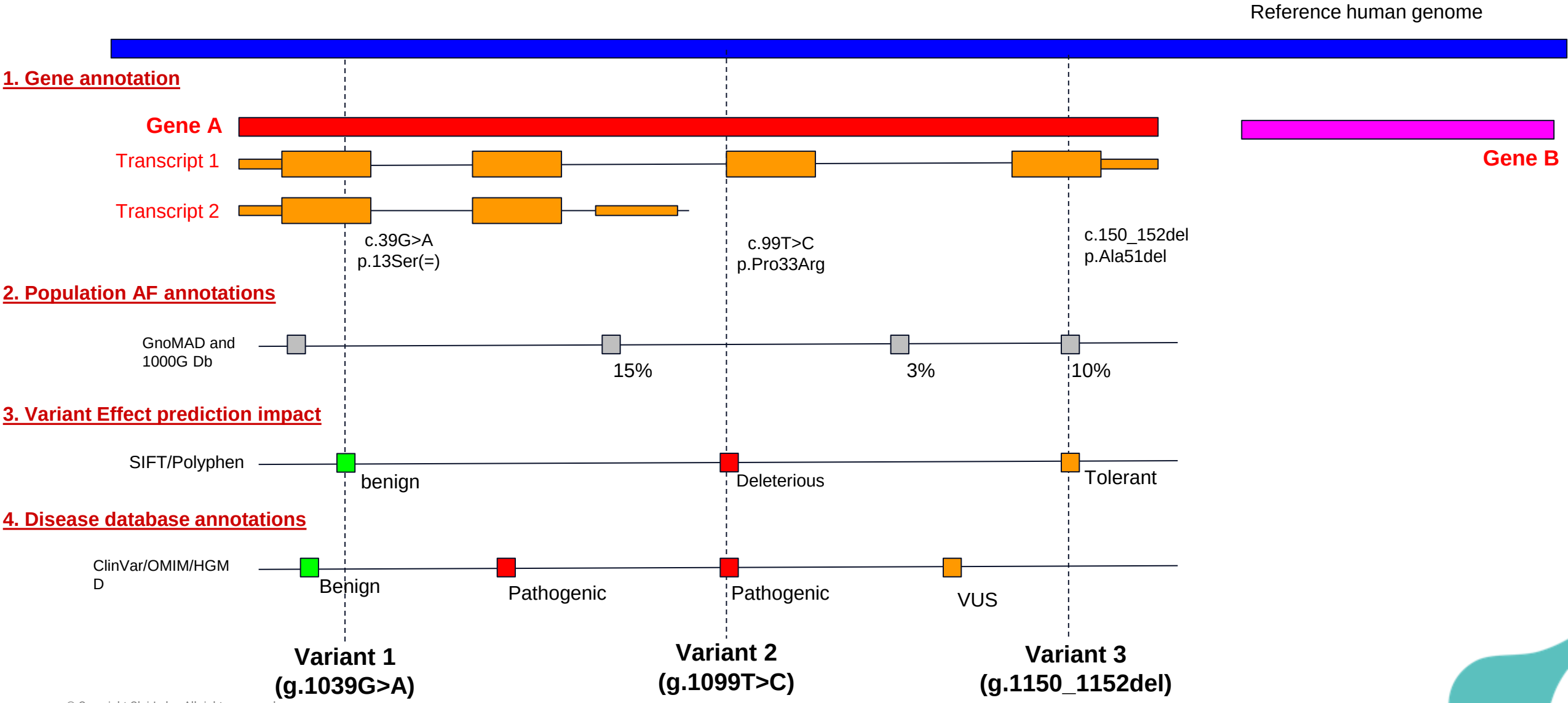
Alignment metrics *samtools*

Coverage metrics *bedtools*

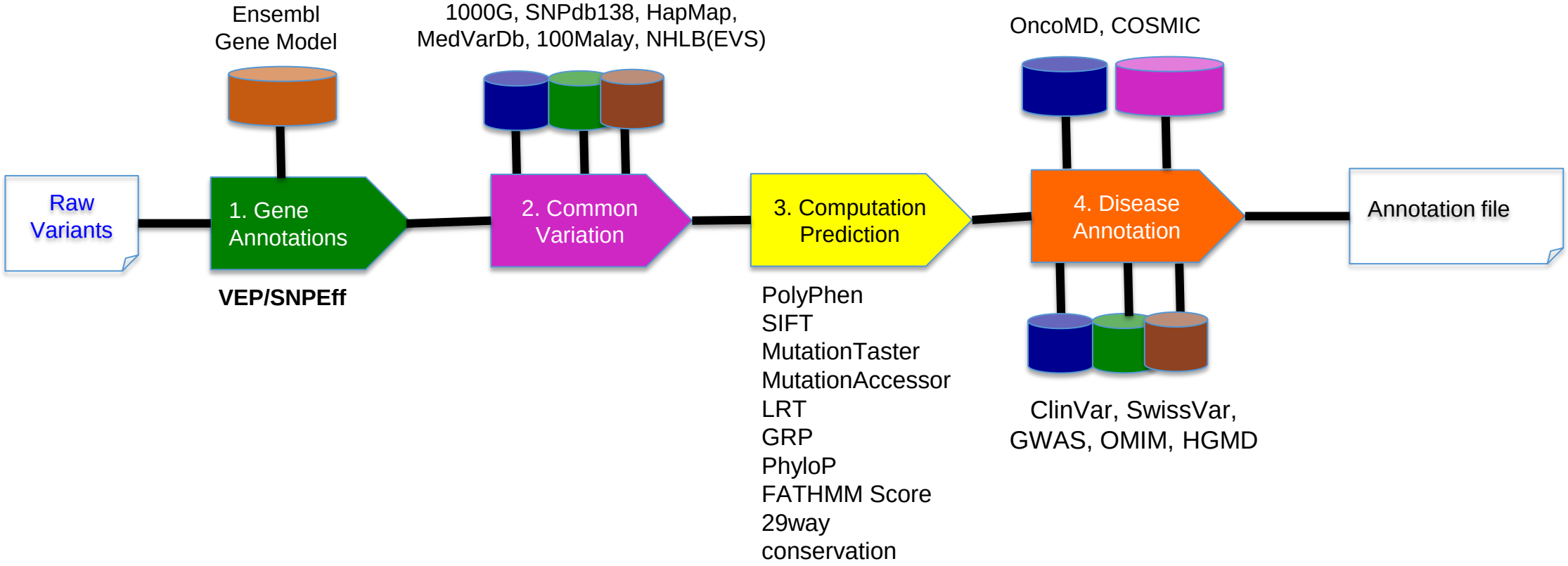
Variant metrics *bedtools*

QC report

Annotation of variants - principle

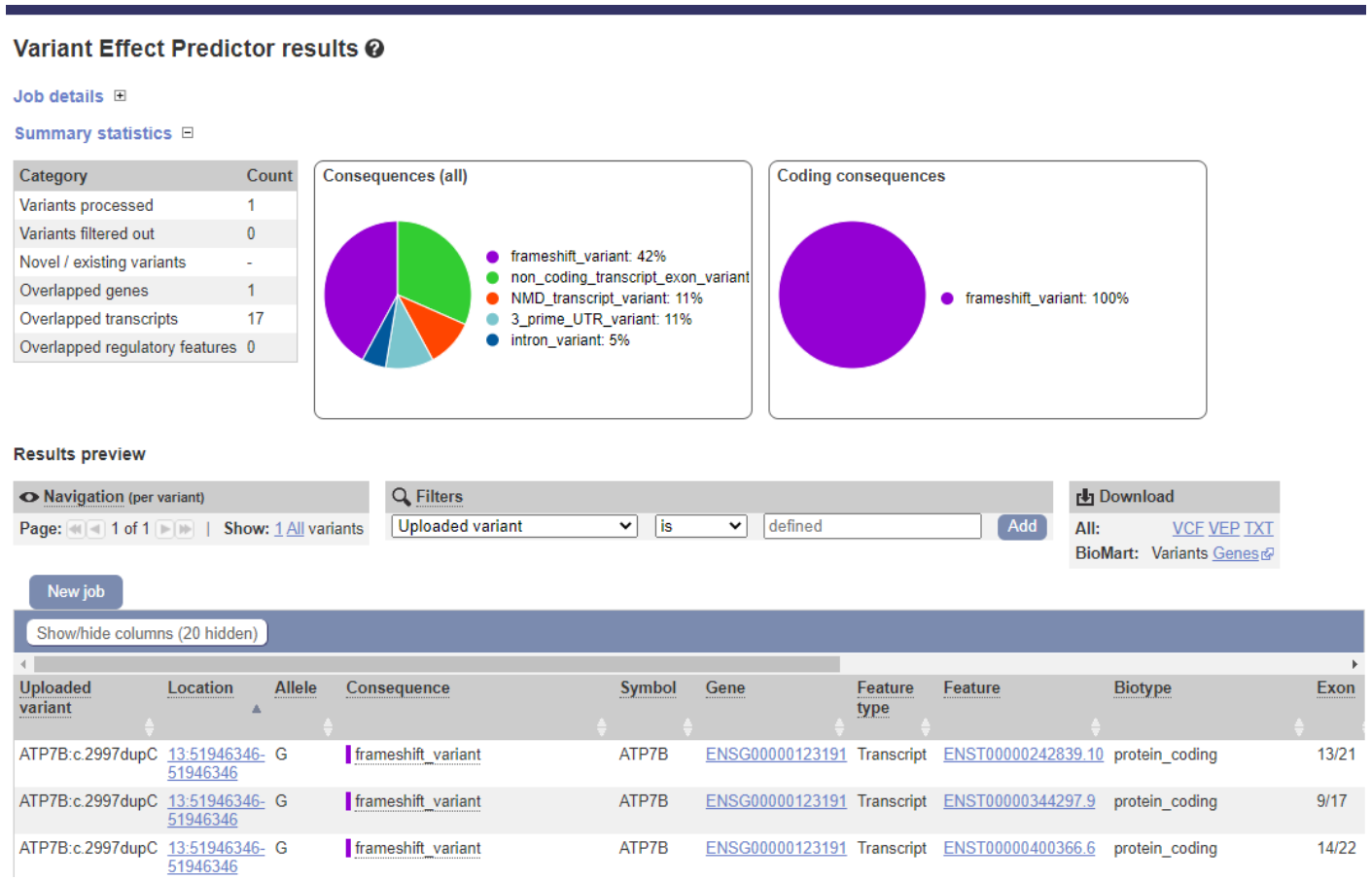


Annotation pipeline - content



Annotation sources - VEP

1. Ensembl's VEP - https://asia.ensembl.org/info/docs/tools/vep/index.html
2. Query: **ATP7B:c.2997dupC or 13:g.51946347dup**



Variant Effect Predictor results

Job details

Summary statistics

Category	Count
Variants processed	1
Variants filtered out	0
Novel / existing variants	-
Overlapped genes	1
Overlapped transcripts	17
Overlapped regulatory features	0

Results preview

Navigation (per variant) Page: 1 of 1 | Show: 1 All variants

Filters: Uploaded variant | is | defined | Add

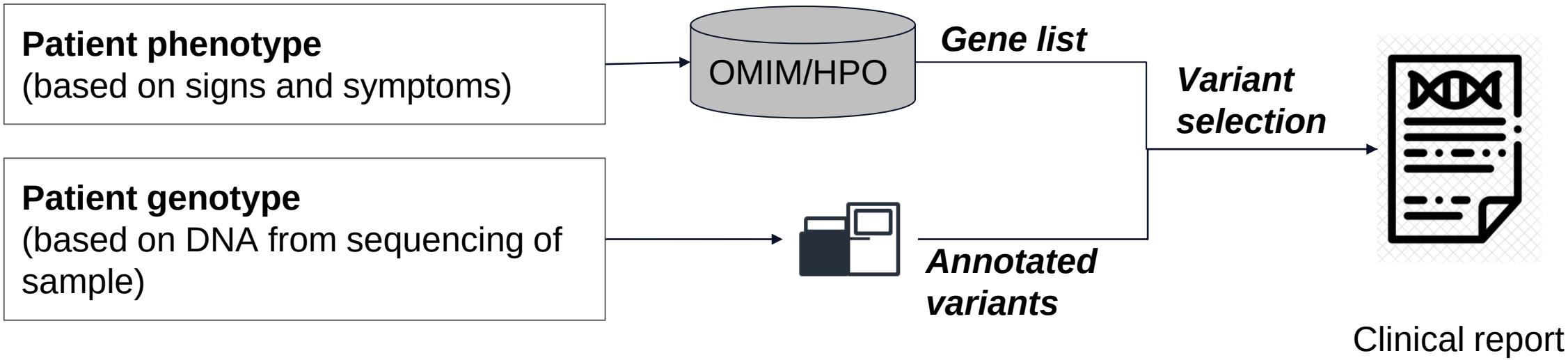
Download — All: VCF VEP TXT; BioMart: Variants Genes

New job

Show/hide columns (20 hidden)

Uploaded variant	Location	Allele	Consequence	Symbol	Gene	Feature type	Feature	Biotype	Exon
ATP7B:c.2997dupC	13:51946346-51946346	G	frameshift_variant	ATP7B	ENSG00000123191	Transcript	ENST00000242839.10	protein_coding	13/21
ATP7B:c.2997dupC	13:51946346-51946346	G	frameshift_variant	ATP7B	ENSG00000123191	Transcript	ENST00000344297.9	protein_coding	9/17
ATP7B:c.2997dupC	13:51946346-51946346	G	frameshift_variant	ATP7B	ENSG00000123191	Transcript	ENST00000400366.6	protein_coding	14/22

Genetic report generation - Principle



#277900
Table of Contents

WD HEPATOLENTICULAR DEGENERATION

Phenotype-Gene Relationships

Location	Phenotype	Phenotype MIM number	Inheritance	Phenotype mapping key	Gene/Locus	Gene/Locus MIM number
13q14.3	Wilson disease	277900	AR	3	ATP7B	606882

Clinical Synopsis PheneGene Graphics

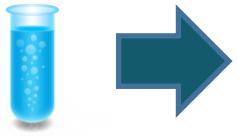
TEXT

A number sign (#) is used with this entry because Wilson disease (WND) is caused by homozygous or compound heterozygous mutation in the ATP7B gene (606882) on chromosome 13q14.

Description

Wilson disease is an autosomal recessive disorder characterized by dramatic build-up of intracellular hepatic copper with subsequent hepatic and neurologic abnormalities.

De Bie et al. (2007) provided a detailed review of the molecular pathogenesis of Wilson disease.



Clinical sample

ABC Diagnostics Lab

Patient details

Name: Mr. Z	**Age:** 15 years	**Sex:** Male
Date and Time Collected: 10th Feb 2018 7:55am	**Physicians name:** Dr. Y	**Fasting:** Yes

Clinical Symptoms

Jaundice, history of abdominal distension, younger brother died due to liver disease.

Results

PATHOGENIC VARIANT CAUSATIVE OF THE REPORTED PHENOTYPE WAS IDENTIFIED

Gene (Transcript)	Location	Variant	Zygosity	Disease (OMIM)	Inheritance	Classification
ATP7B (-) (ENST00000242839)	Exon 13	c.2997dupC **(p.Gly1000ArgfsTer28)**	Homozygous	Wilson disease	Autosomal recessive	Pathogenic

The *ATP7B* gene is 100% covered in this assay.

Variant interpretation and Clinical Correlation

A homozygous single base pair duplication in exon 13 of the ***ATP7B*** gene **(chr13:52520483dupG; Depth: 100x)** that results in a frameshift and premature truncation of the protein 28 amino acids downstream to codon 1000 **(p.Gly1000ArgfsTer28; ENST00000242839)** was detected (Table).

Wilson disease (OMIM#277900) is caused by mutation in the *ATP7B* gene (OMIM*606882). This disorder is primarily characterized by dramatic build-up of intracellular hepatic copper with subsequent hepatic and neurologic abnormalities [10]. The observed variation has previously been reported (as c.2996insC) in patients affected with Wilson disease [23]. The *ATP7B* variant has not been reported in the 1000 genomes, ExAC and our internal databases respectively. The *in silico* prediction# of the variant is damaging by MutationTaster2. The reference region is conserved across species.

Based on the above evidence, ***this ATP7B variation is classified as a pathogenic variant and has to be carefully correlated with the clinical symptoms.***

Clinical report
(3 weeks time)

Operation pipeline

Clinical samples (~1000 samples per day)

Sample collection and storage room

Customer support room

Blood handling room

DNA preparation room

Amplification and probe selection room

Sequencing machine room

Clinical Bioinformatics analysis room

Clinical Bioinformatics reporting room

Customer support room

Clinical report (3 weeks time)

Clinical Bioinformatics operations – core tasks

1. Run data-intensive informatics pipeline for **1000s** of samples as batch in **production** mode
   - Prioritize analysis for samples that have shorter report time (Emergency samples)
   - Keep track of the analysis status – re run analysis that failed because of technical issues
   - Keep track of any exception in analysis, if there is any
2. Perform QC and approve each of every sample in bulk
   - Keep track of QC failed samples and report promptly to the lab team for reprocessing
   - Check for any systematic errors – trending of errors
   - Validation of the pipeline results
3. Develop and optimize new pipelines - new assays
4. Manage data sources and content for variant annotations
5. Provide Bioinformatics support for customers and internal teams
   1. Variant confirmation
   2. Annotation discussions

Clinical Bioinformatics operations - requirements

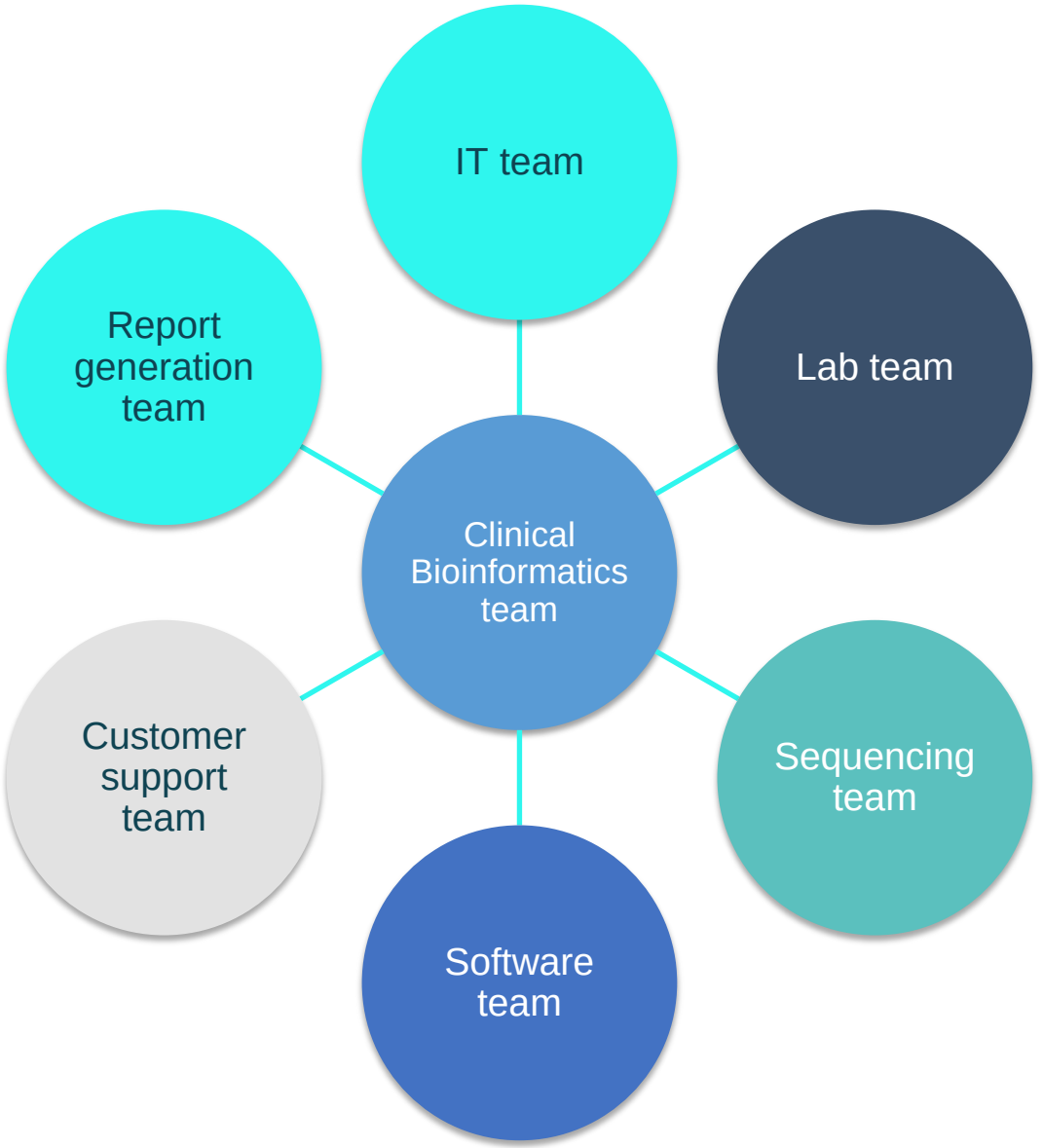
Assumptions : 1500-2000 clinical exome sample analyses per day
(4 hours to perform analysis; ~10Gbps sequence data; ~5 Gb file size read data)

- Need dedicated infrastructure components to analysis samples
  1. Storage
     - 1500 x 5 = 7.5 Tb of raw data file per week == ~55 Tb per week
     - **~2 Pb Storage should be planned for every year**
  2. Analysis servers
     - 4 hours per clinical sample
     - For ~10,000 samples analyses per week

Qualified Team

- Need dedicated and structured clinical bioinformatics team
  1. Analysis group
     - Regular analysis of clinical servers in batch and track issues with analysis
     - Perform integration of new components to the pipeline
     - Monitor the analysis servers performance and issues
  2. QC group
     - Perform QC and approve regular clinical sample samples
     - Add more QC tests and metrics for approval
  3. Development & build group
     - Enhancement of the pipeline code and manage versions of the pipeline
     - Automation of analysis and QC process
     - Upgrade software and data resources used for annotations
  4. Data management group
     - Manage analysis files in the storage system– archival, deletion
     - Perform and track data sharing requests
  5. Validation group
     - Manage documentations and records on operation tasks
     - Develop policies for process; Validate new assays

Clinical bioinformatics team interactions



About Impactomics

Life sciences and healthcare teams process unprecedented volumes of genomic, clinical, and real-world data. Yet fragmented tooling, regulatory friction, and legacy infrastructure slow the path from signal to validated action.

ClairLabs brings you Impactomics – integrating multi-omics NGS, bioinformatics, agentic AI, and cloud engineering with enterprise data governance. Transform raw NGS data into actionable clinical insights for rare diseases, oncology, and precision medicine. Accelerate scientific outcomes through FASTQ, audit-ready reports, and more to bridge the gap between sequencing and clinical decision-making – with the speed, accuracy, and compliance rigor for modern health and diagnostics demand.

Book a Demo

https://clairlabs.ai/impactomics

Summary

1. Clinical Bioinformatics is one of the key component of the NGS-based Diagnostics Assay
2. Production grade operations are required to perform analysis of each patient sample
   - Process level errors are not tolerated; Error can result in wrong treatment or misdiagnosis
3. Dedicated bioinformatics with focus on analysis, QC and development is required to handle large volume of samples from NGS systems
   - Repeatable, reproducible, reliable and trackable results should be obtained from each patient samples

NGS-based diagnostics assay - challenges

- Clinically validity is established for many hereditary diseases, but not all
  - The assay should be able to differentiate between the disease conditions and healthy controls
- Clinical usability is established for many hereditary diseases, but not all
  - The assay should enable doctors to improve the patient’s health conditions by providing treatment

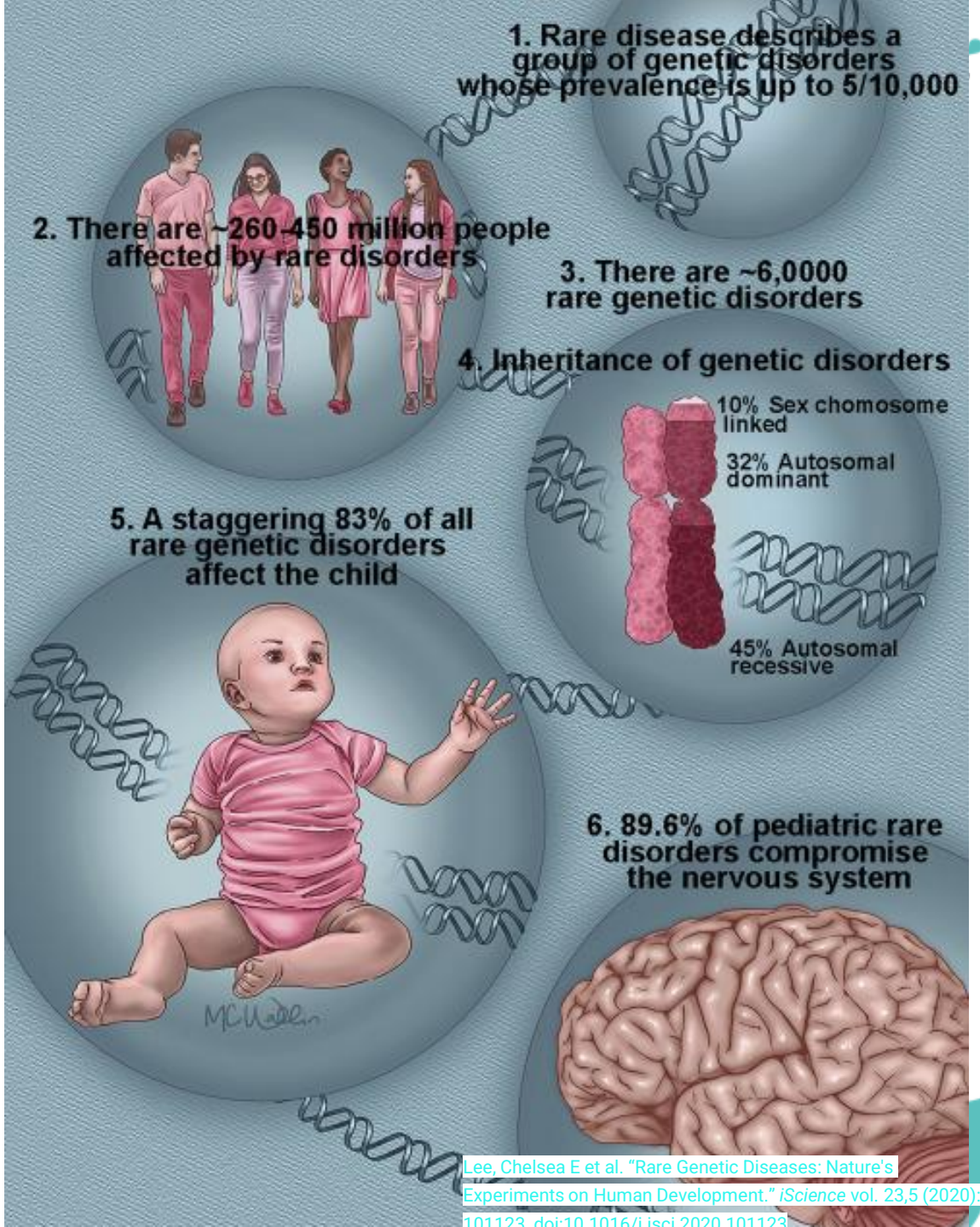
AIRLABS

Thank you

Rare diseases

- Diseases 1 in 2000 in population
- >80% are **genetic** - caused by one or two changes in the single gene or chromosomal abnormality
- 250-450 million people are affected
- ~6000 rare diseases
- Cystic fibrosis, Spinal Muscular Atrophy, Osteopetrosis, Retinitis pigmentosa, etc
- **Not** like common diseases
    - More influenced by environment and associated with more genes
    - Diabetes, Hypertension etc

Genetics : Study of **heredity** — how traits and diseases are **passed from parents to offspring**.



Lee, Chelsea E et al. "Rare Genetic Diseases: Nature's Experiments on Human Development." *iScience* vol. 23,5 (2020): 101123. doi:10.1016/j.isci.2020.101123