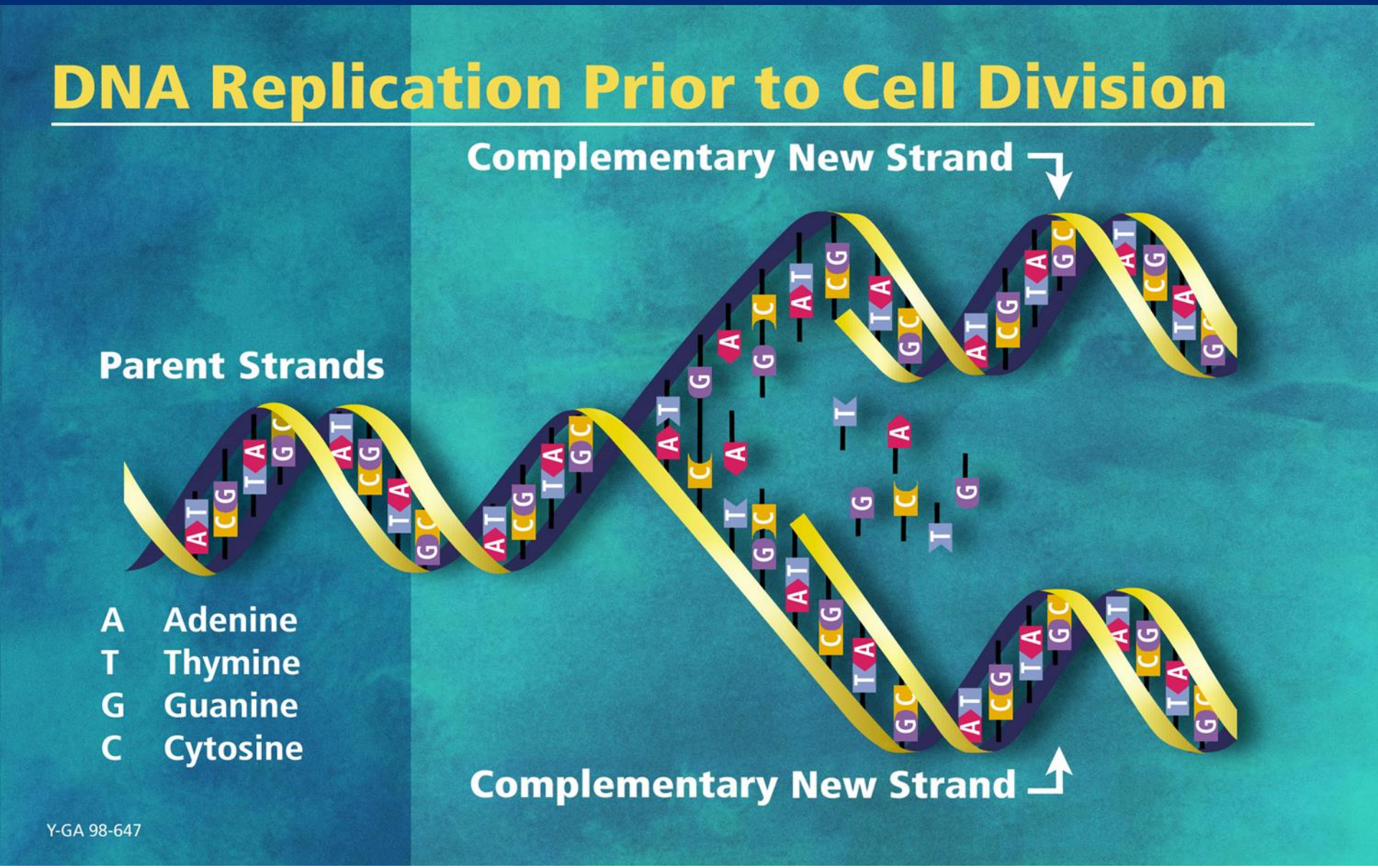
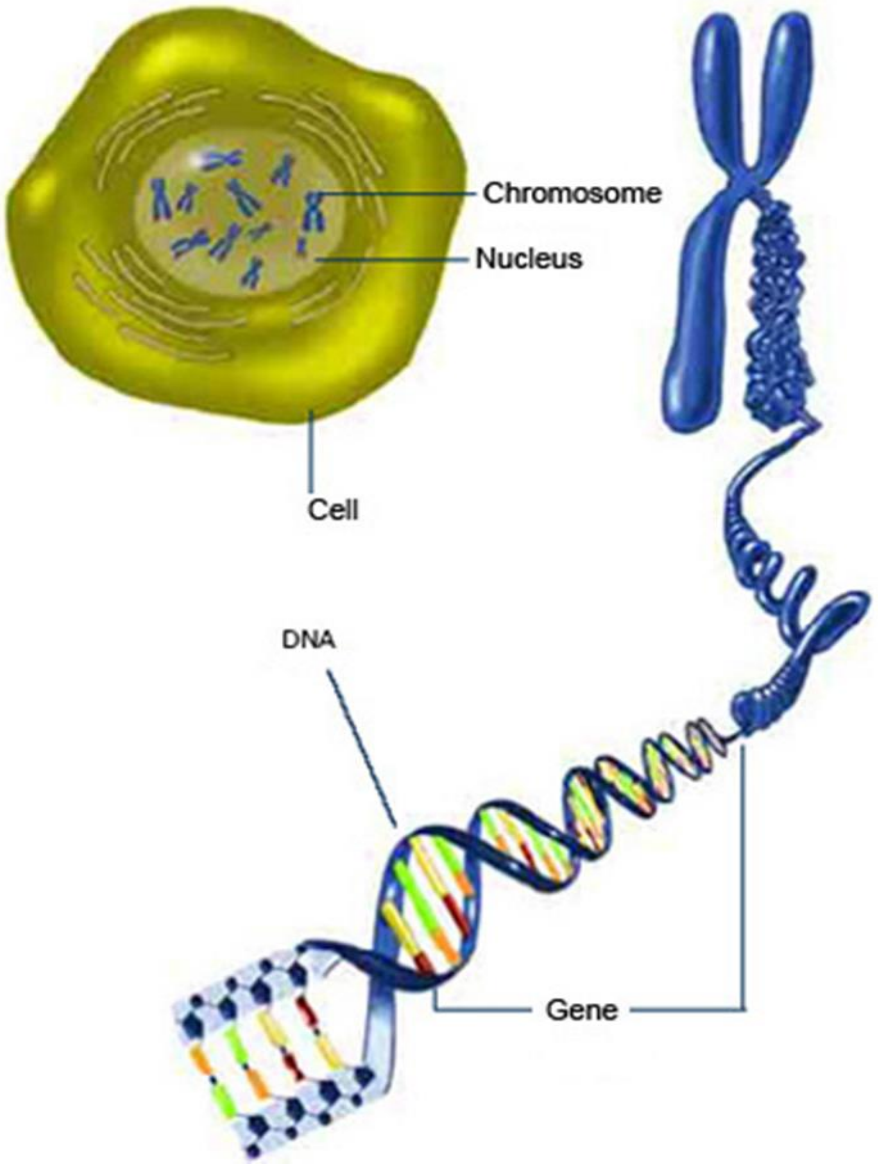
Fundamental Concepts of Sequencing

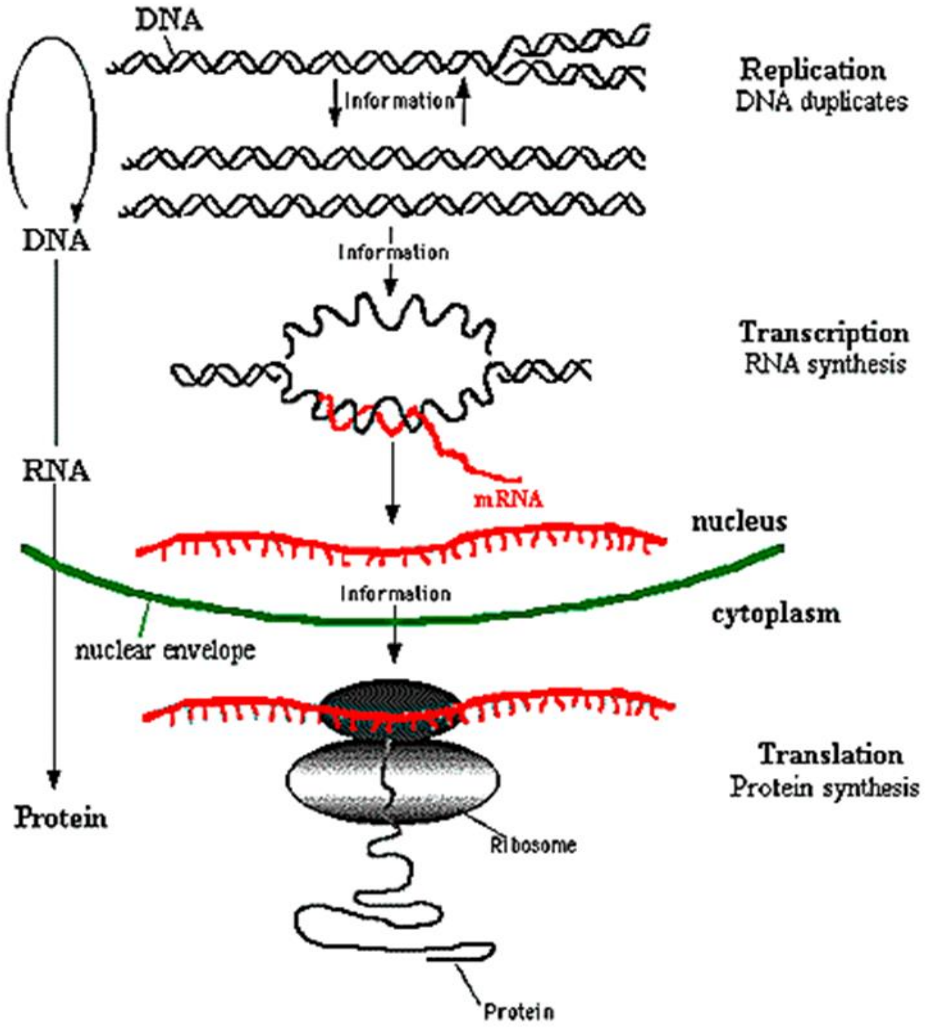
Carlos A. Saavedra-Matiz, MD
Newborn Screening Program
Wadsworth Center
New York State Department of Health

February 25, 2020
APHL-CDC

DNA Replication Prior to Cell Division

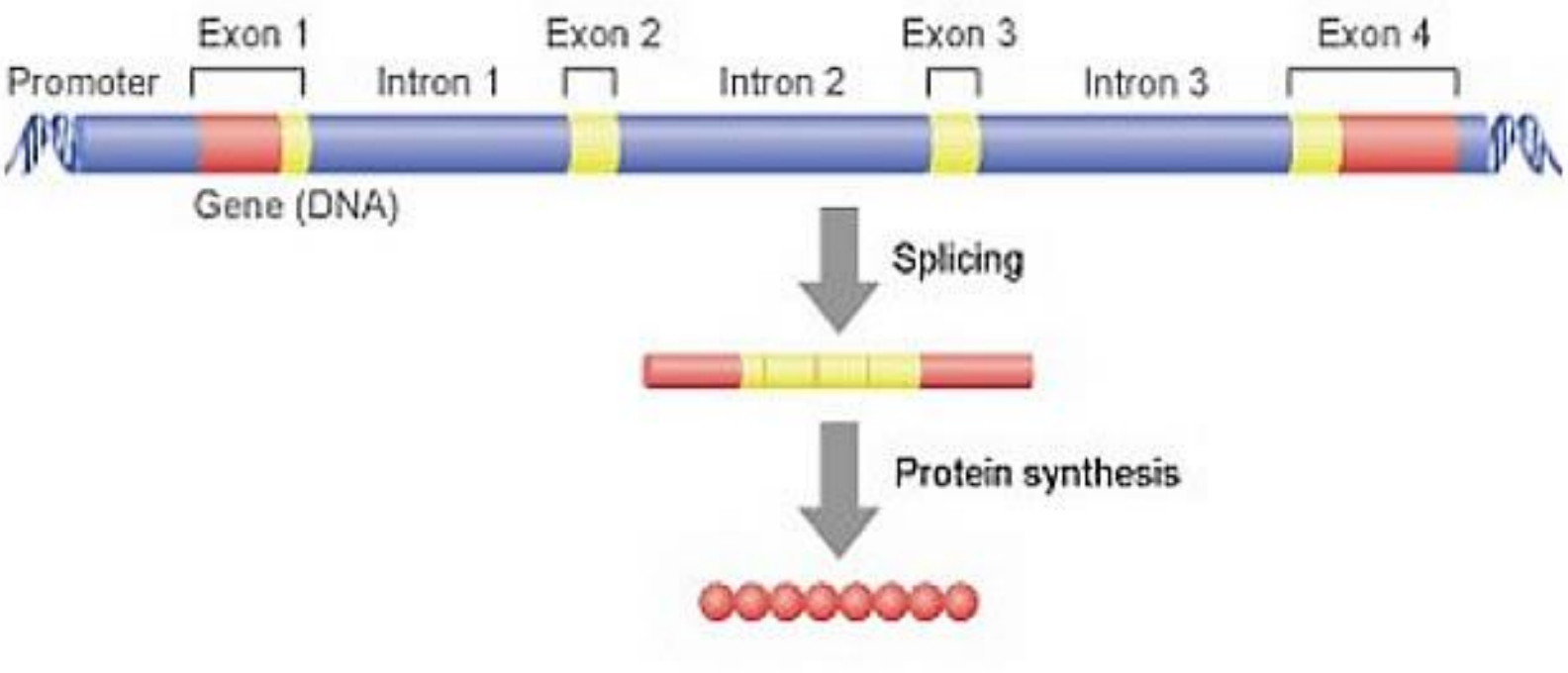




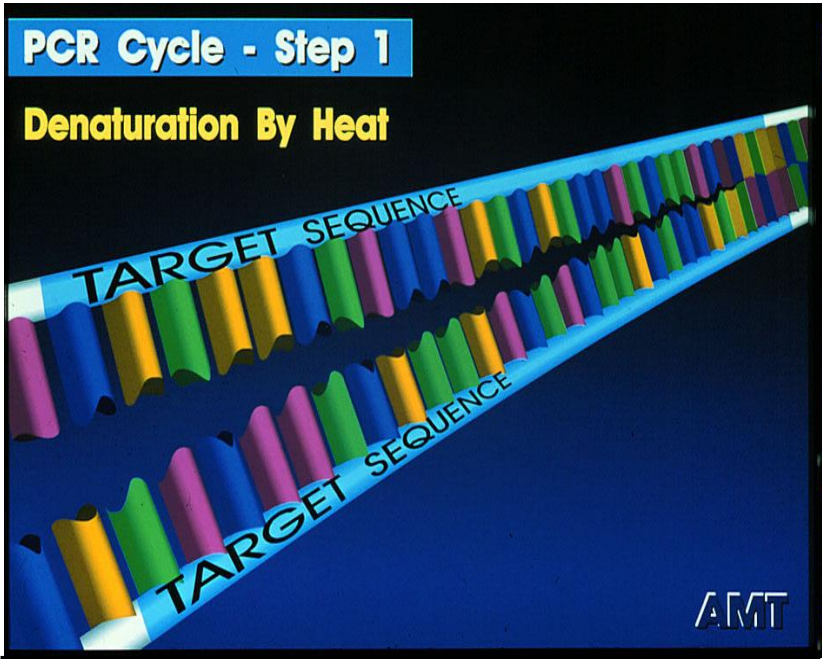


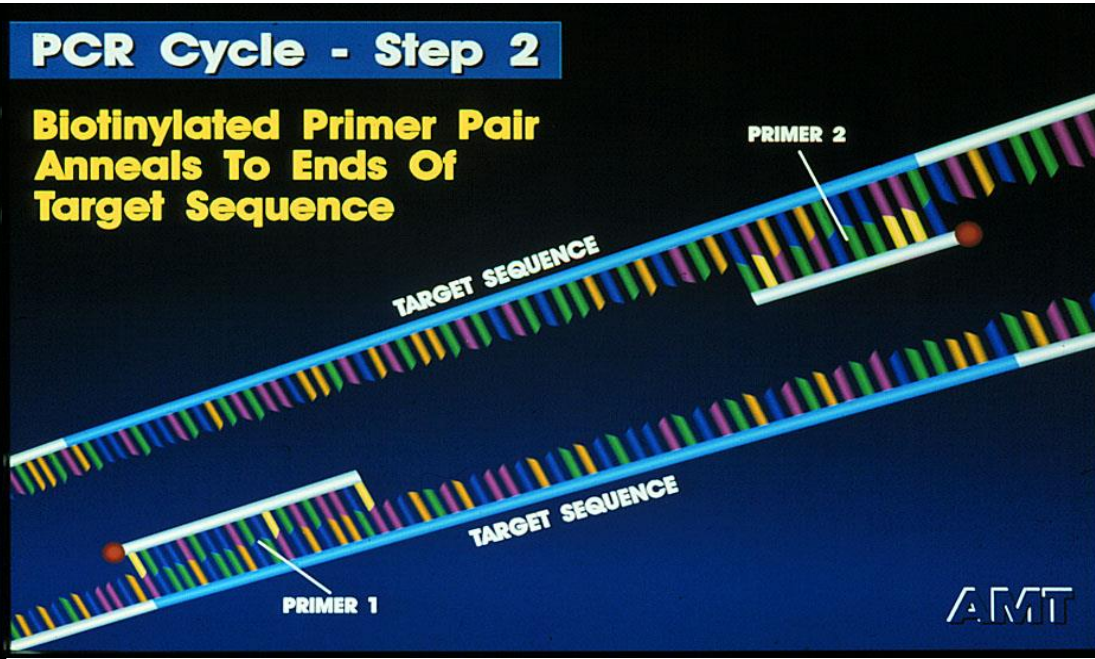
The Central Dogma of Molecular Biology

Gene = DNA Instructions for A Specific RNA/Protein



POLYMERASE CHAIN REACTION

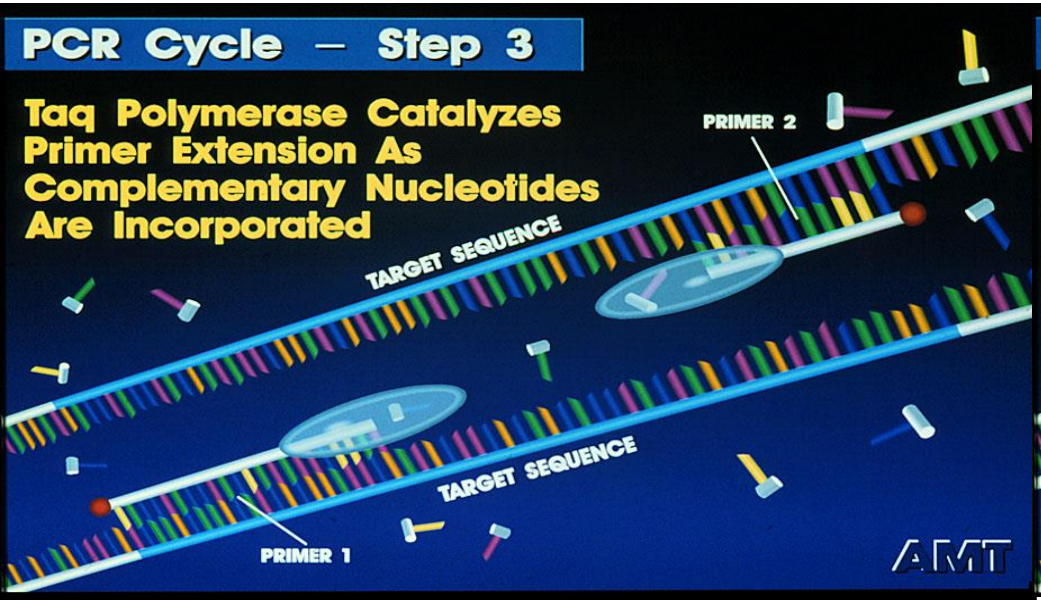




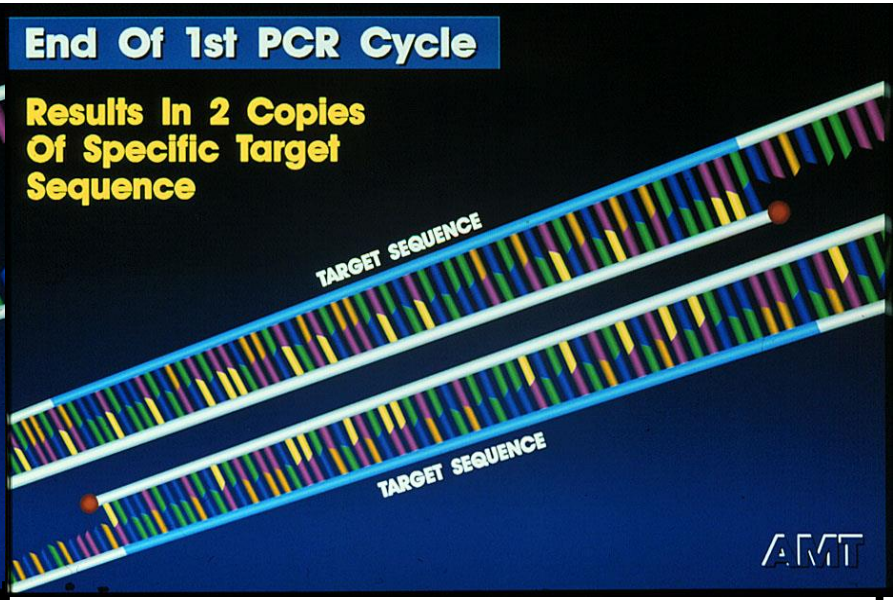
- Heat denaturation (95 degrees C).
- Provides sufficient energy to break hydrogen bonds.

- Primer annealing 55-65 degrees C.
- Takes advantage of base complementarity.

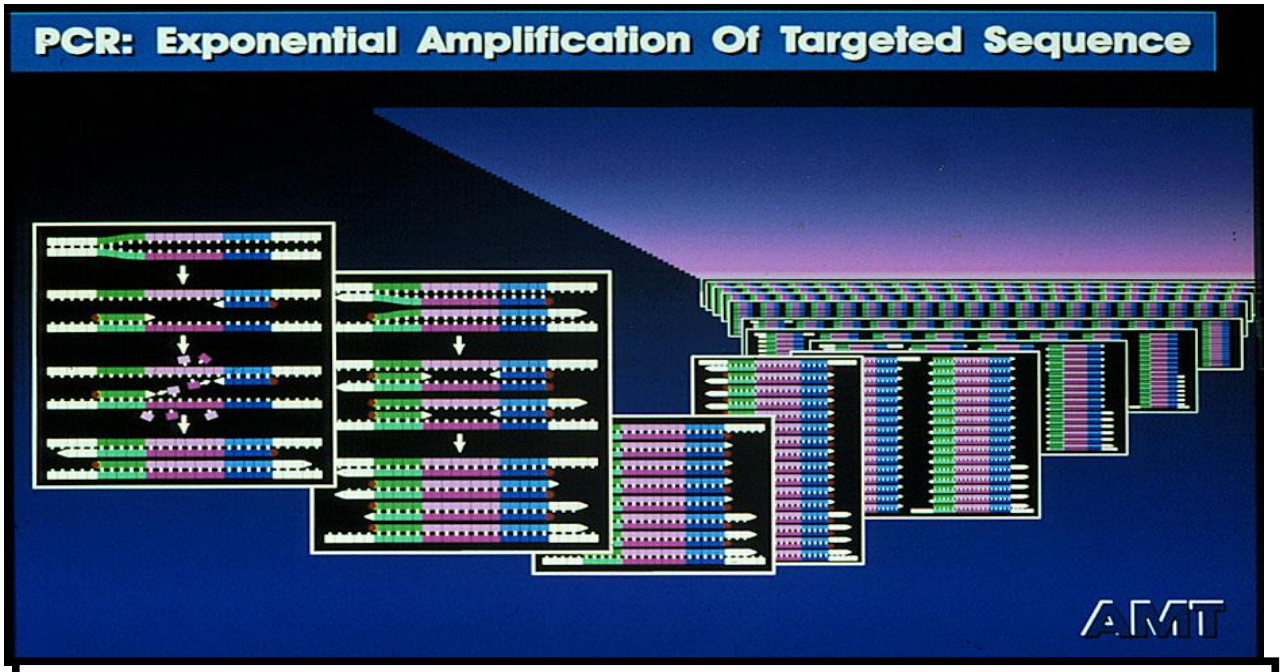
POLYMERASE CHAIN REACTION



- Polymerization at 72 degrees C.
- Requires availability of Thermus aquaticus DNA polymerase (heat stable).

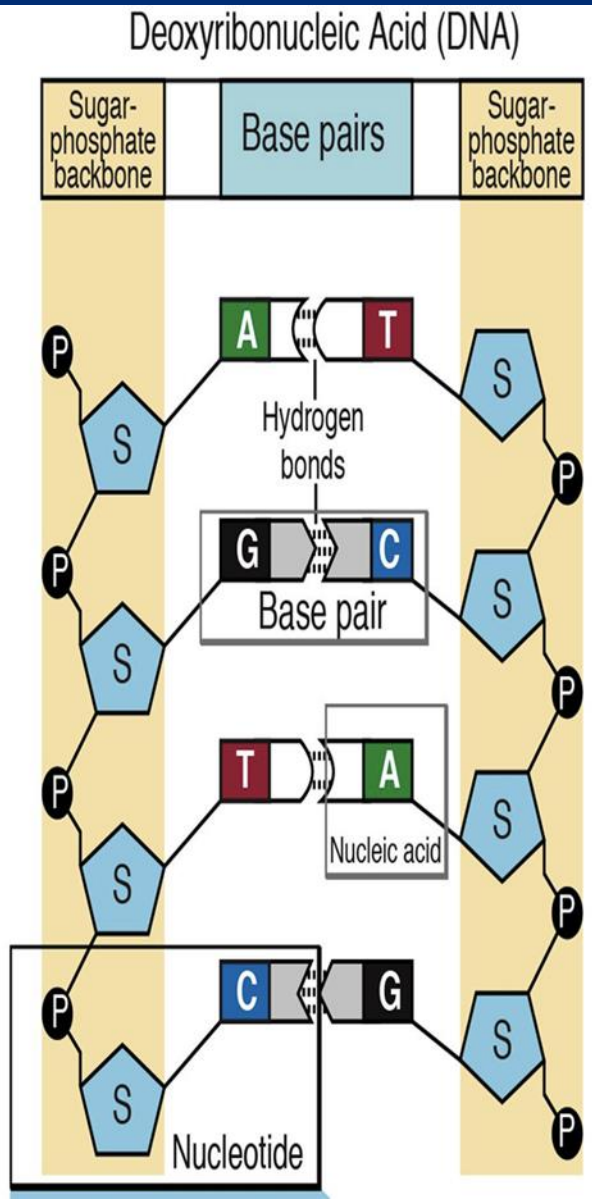


- After the first cycle, there are 2 copies of the original double helix.
- Continue cycling….

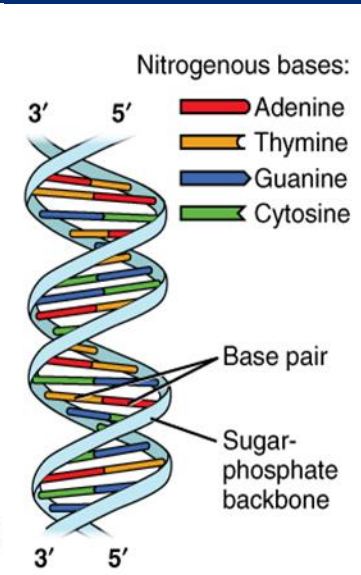


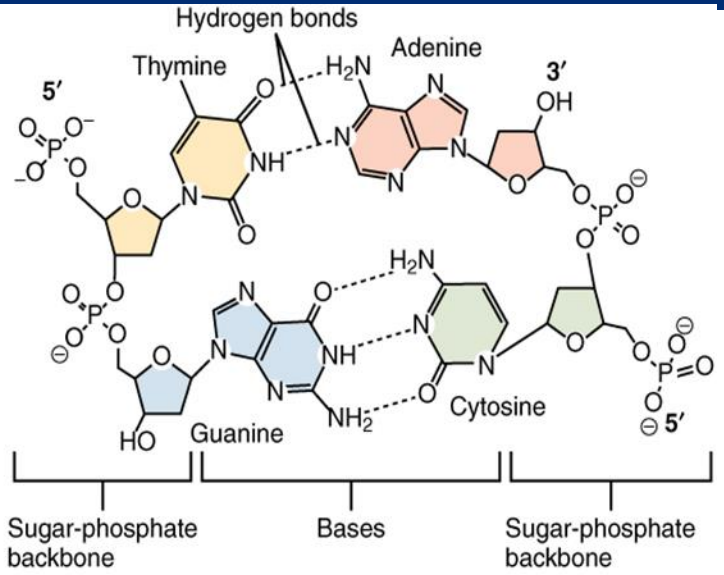
POLYMERASE CHAIN REACTION

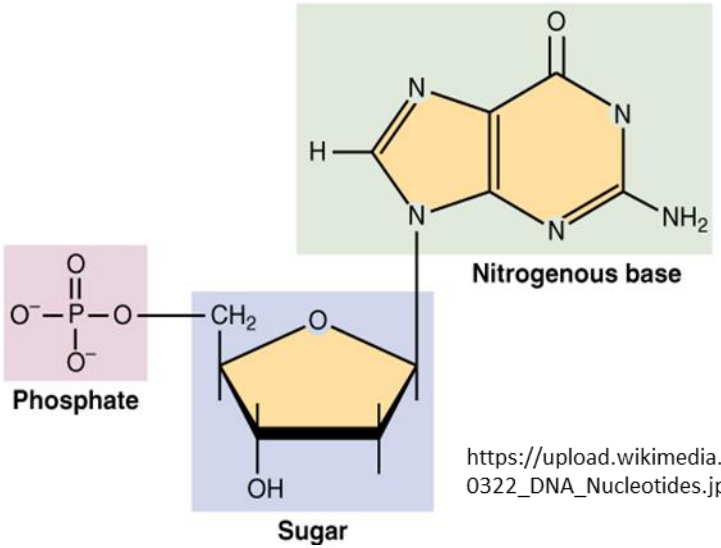
- **And so on…for 30-35 cycles**
- **Result is billion-fold amplification of target.**



https://knowgenetics.org/nucleotides-and-bases/



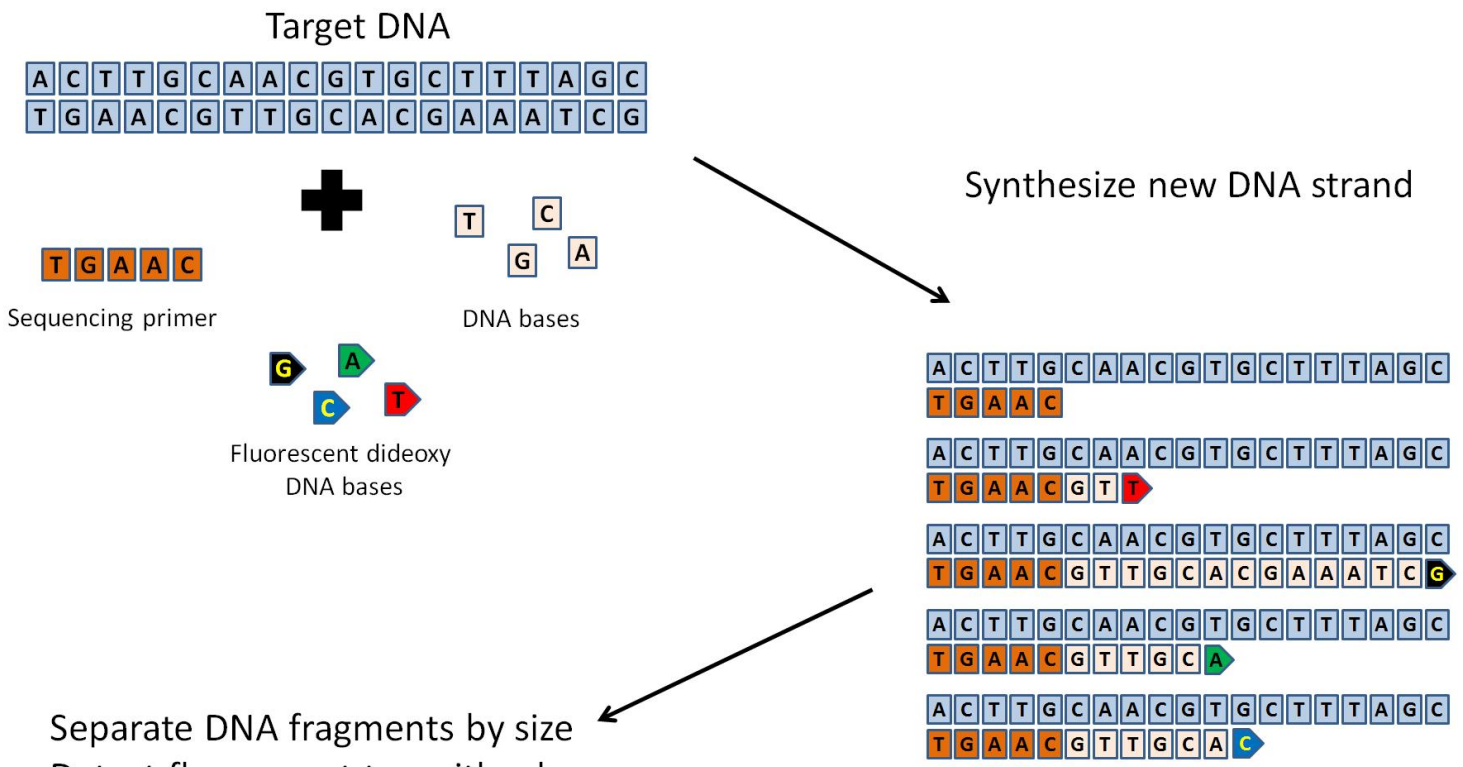




https://upload.wikimedia.org/wikipedia/commons/d/d3/0322_DNA_Nucleotides.jpg

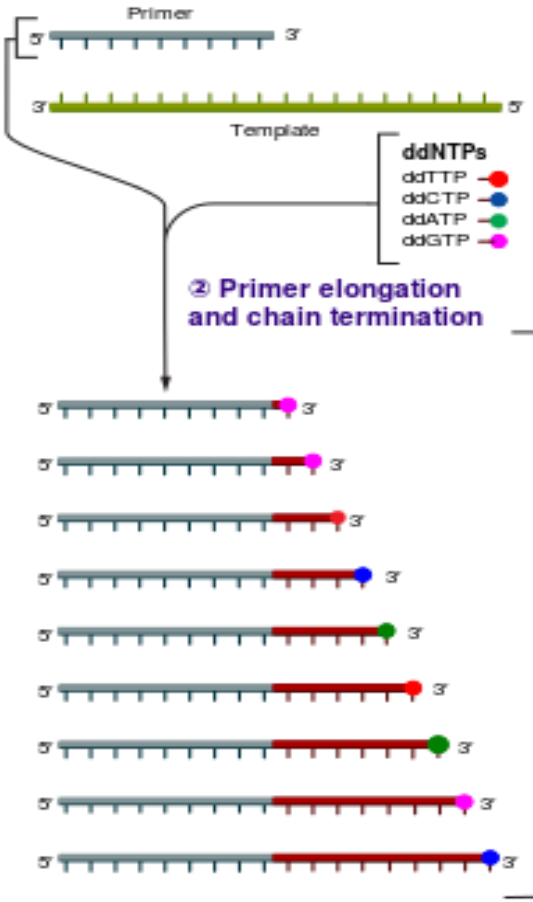
SEQUENCING: Determine the order of nucleotides in a DNA or RNA strand

Sanger

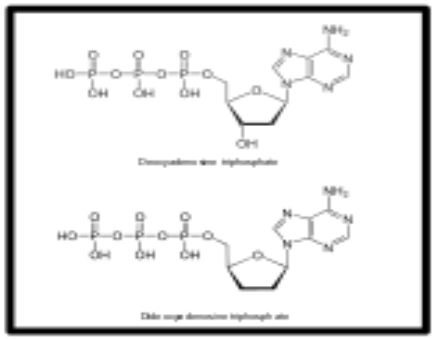


Sanger Sequencing

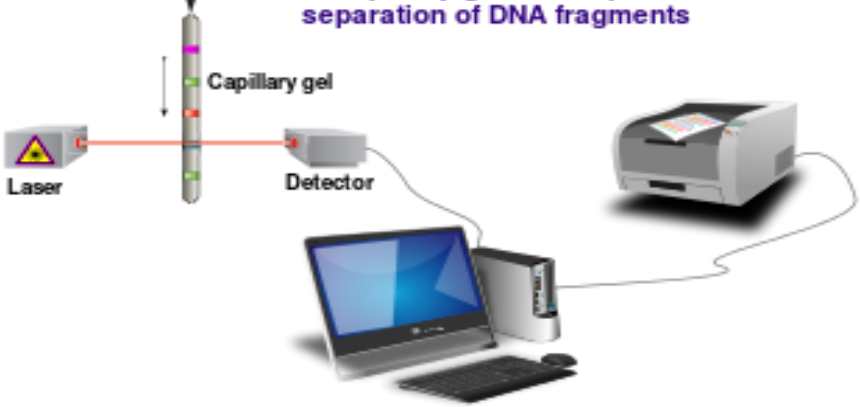
① Reaction mixture
- Primer and DNA template
- DNA polymerase
- ddNTPs with flourochromes
- dNTPs (dATP, dCTP, dGTP, and dTTP)

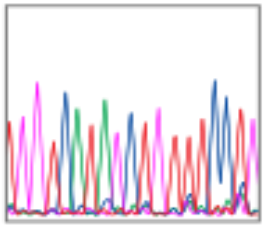


② Primer elongation and chain termination



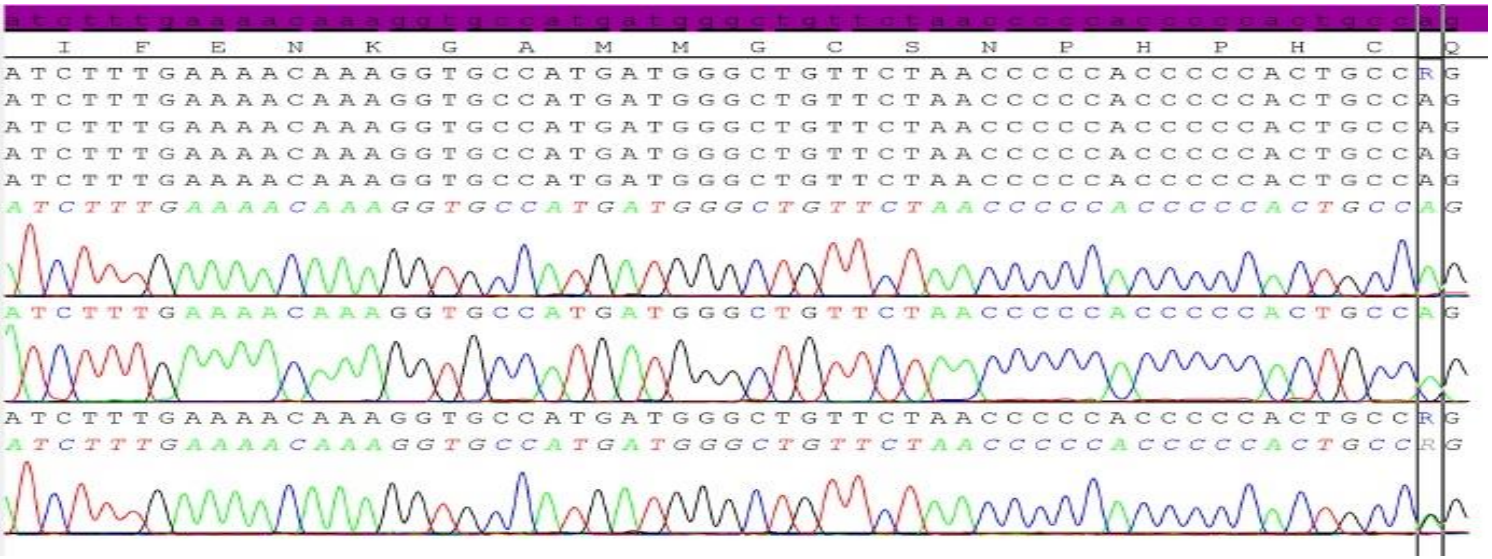
③ Capillary gel electrophoresis separation of DNA fragments





④ Laser detection of flourochromes and computational sequence analysis

GALT Genotyping – Moving to Sequencing



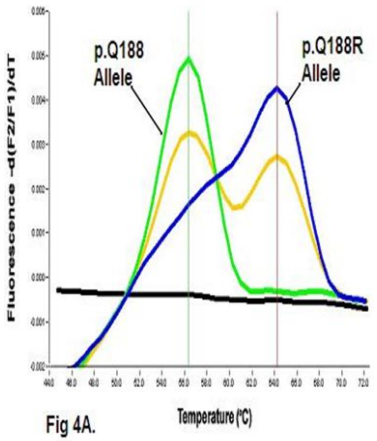
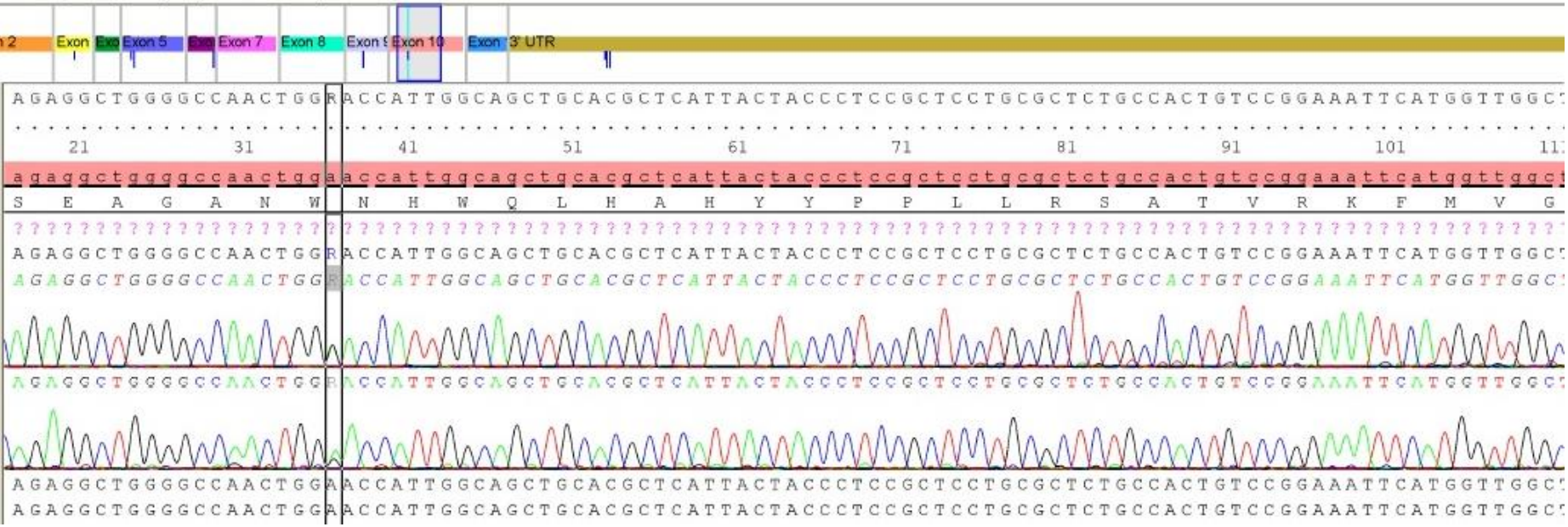


Fig 4A.

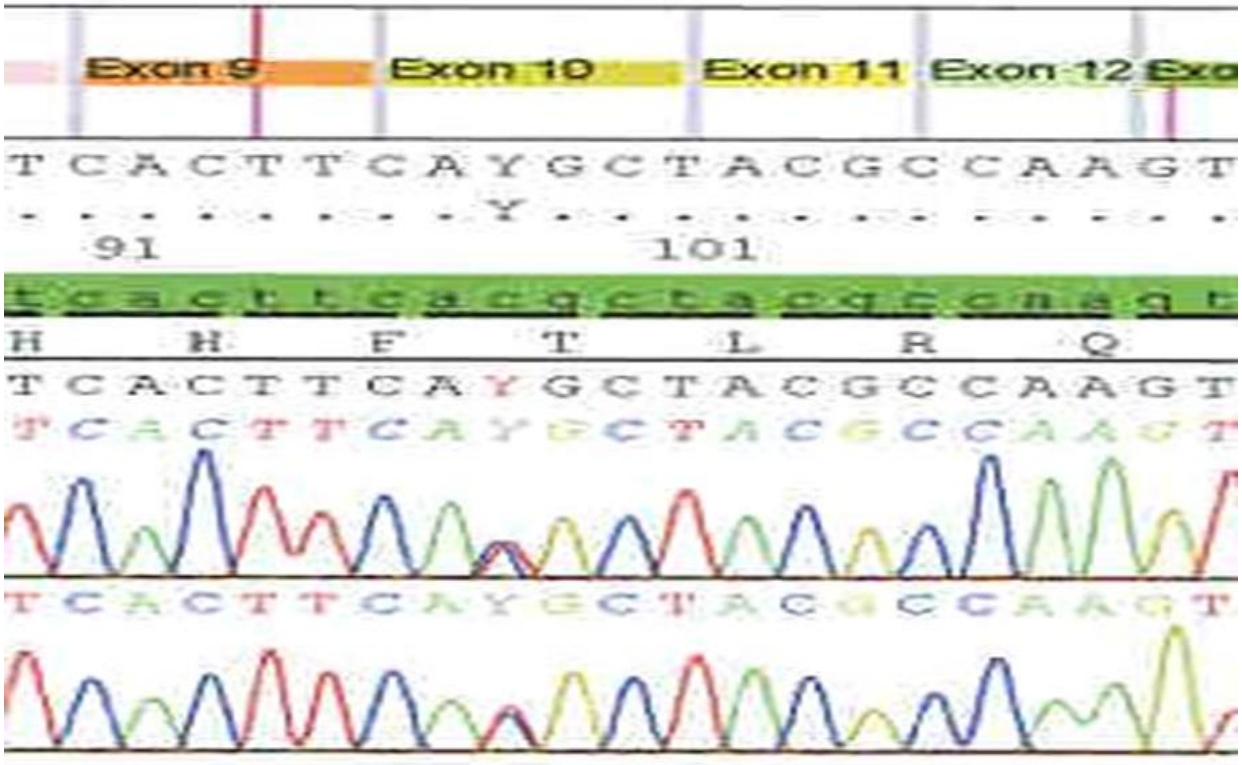
c.563A>G (p.Gln188Arg)



c.940A>G (p.Asn314Asp) Duarte

Molecular Analysis of GALC

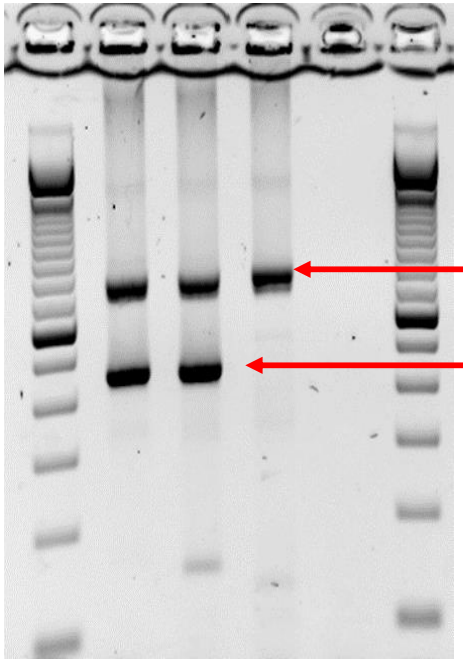
- Reduce number of false positive screens
- Predict phenotype (?)
- Method:
  - Agarose gel for 2 common deletions
    - 30 kb & 7 kb
  - **Sequence all 17 exons and Promoter Region**



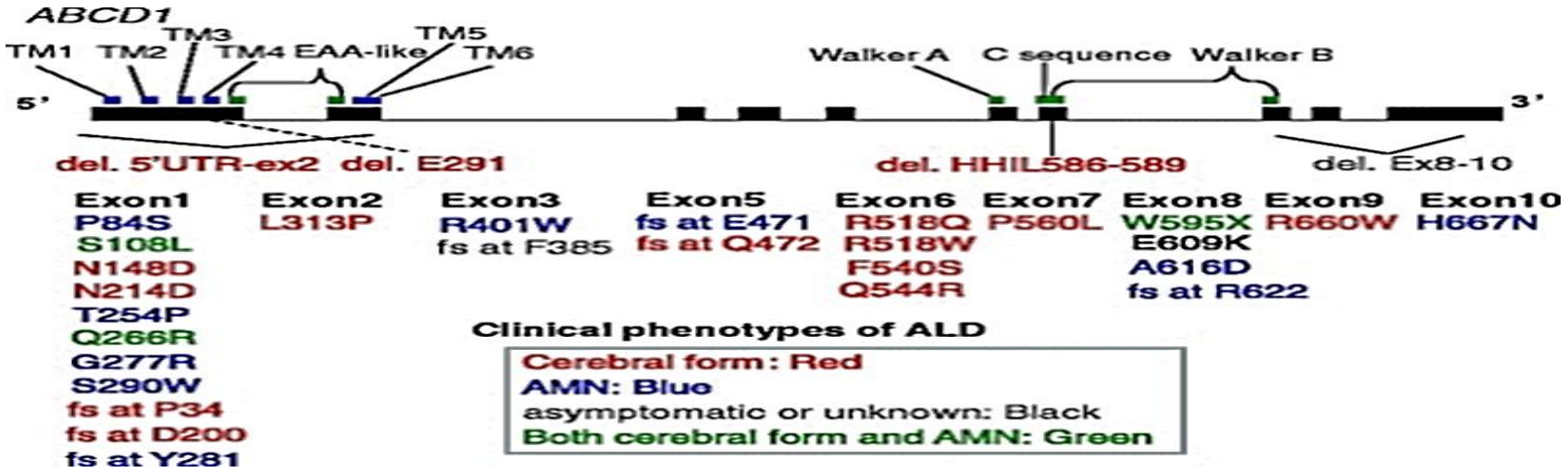
c.1586C>T (p.Thr529Met)

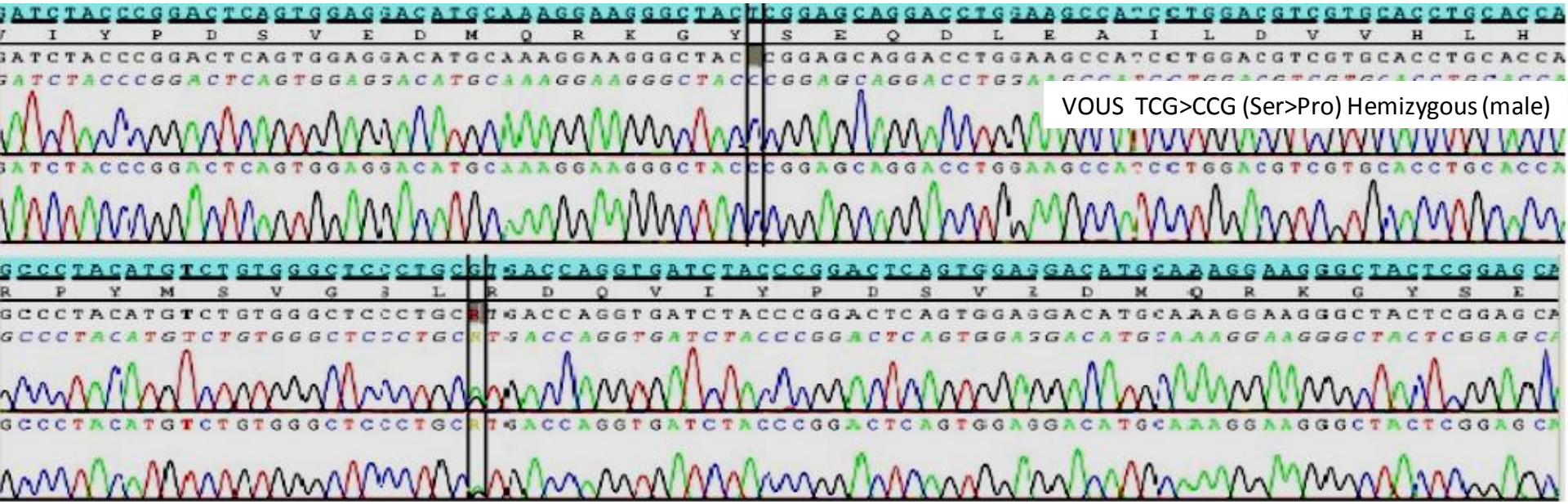
acg > atg

30 kb deletion

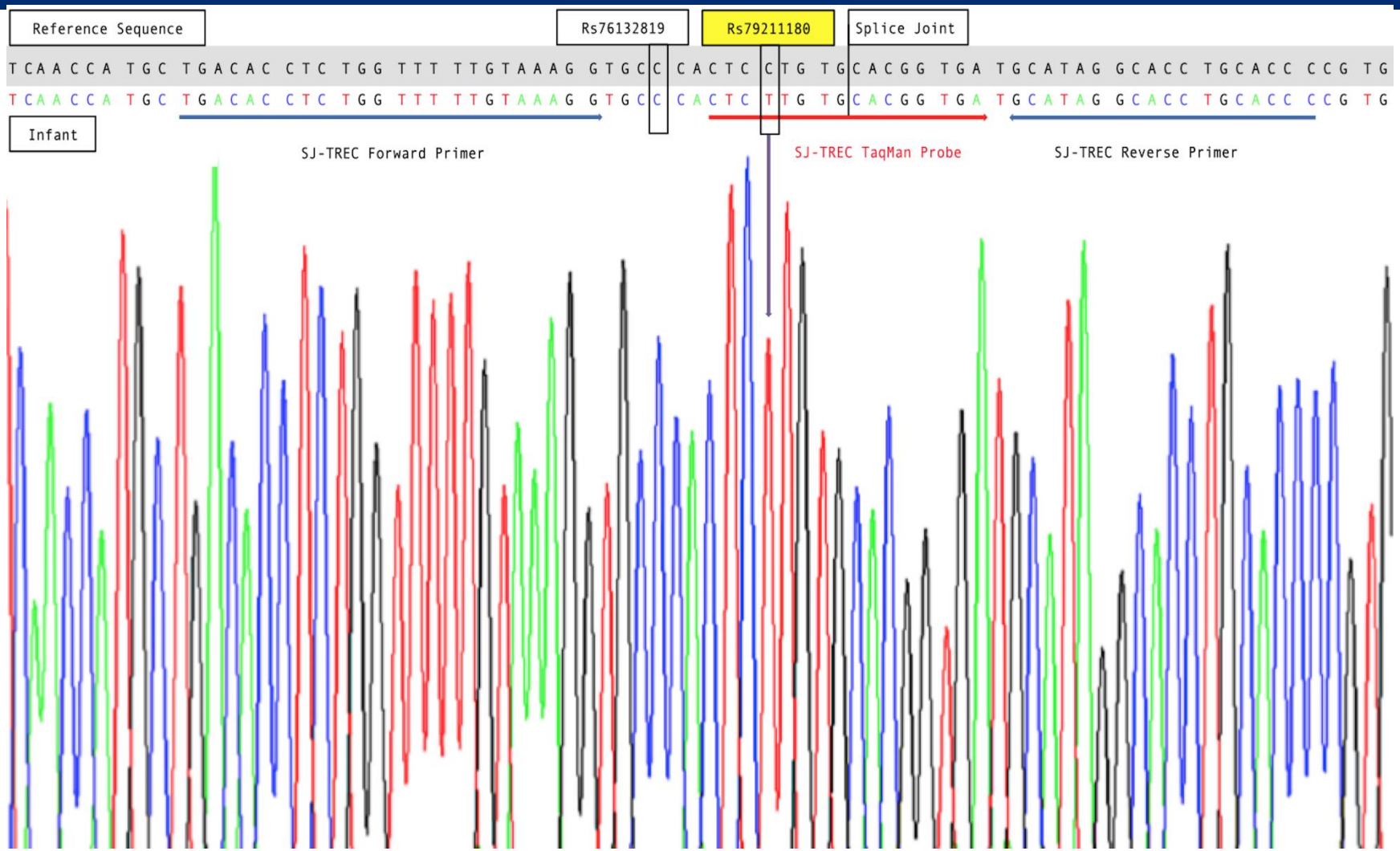


Known Variants in the ALD Gene - *ABCD1*





c.1661 G>A (p.Arg554His) Heterozygous (female). Childhood cerebral & adult onset X-ALD



Title: A Single Nucleotide Polymorphism in the T-Cell Receptor Excision Circle.

Gans MD, Saavedra-Matiz CA, Bernstein L.

J Allergy Clin Immunol Pract. 2019 Aug 17. pii: S2213-2198(19)30720-2. doi: 10.1016/j.jaip.2019.08.014. [Epub ahead of print] No abstract available.

PMID:31430593

The Next Wave of DNA Sequencing

Buzzwords

- "Massively parallel" sequencing
- "High-throughput" sequencing
- "Ultra high-throughput" sequencing
- "Next generation" sequencing (NGS)
- "Second generation" sequencing

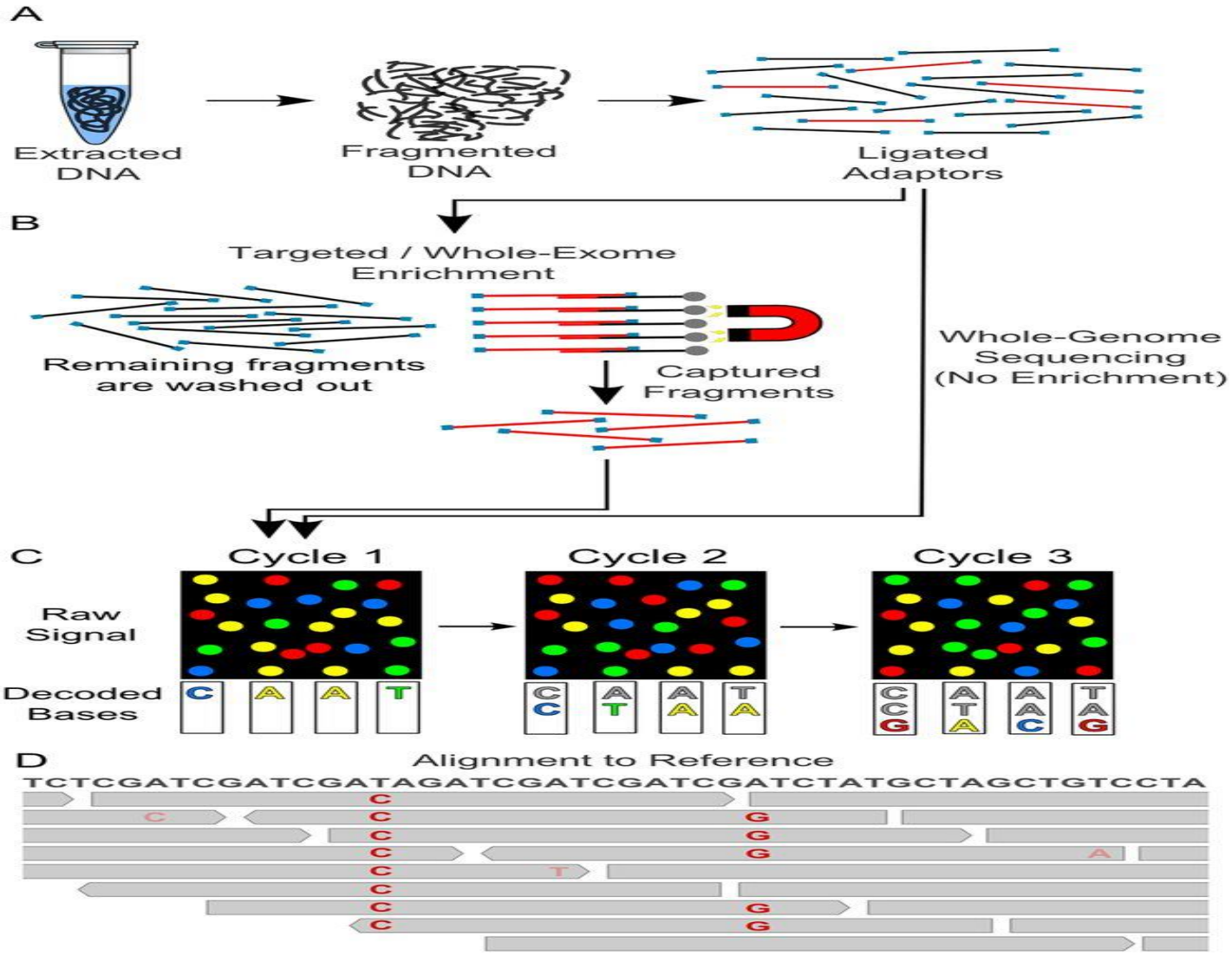
- **2005: 454 (Roche)**
- **2006: Solexa (Illumina)**
- **2007: ABI/SOLiD (Life Technologies)**
- **2010: Complete Genomics**
- **2011: Pacific Biosciences**
- **2010: Ion Torrent (Life Technologies)**
- **2015: Oxford Nanopore Technologies**

Next Generation Sequencing

- Utility
  - Sequence of a large gene
  - Sequence of several genes
  - Sequencing many people simultaneously

- Disorders
  - Cystic Fibrosis – large gene (44 exons, 10 -18 kb coding)
  - Severe Combined Immunodeficiency – 39 genes (55)

Steps in next-generation sequencing.

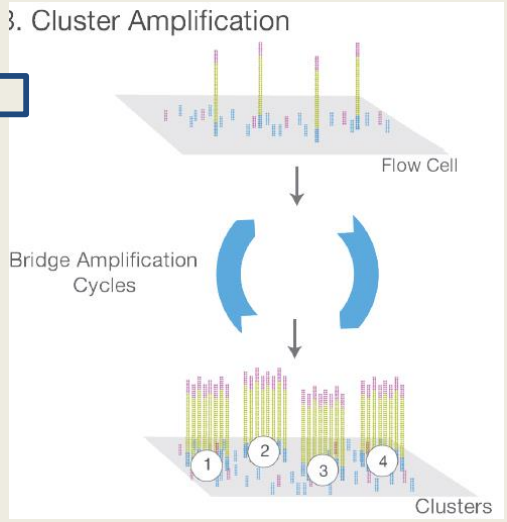


Ricardo Parolin Schnekenberg, and Andrea H Németh Arch Dis Child 2014;99:284-290

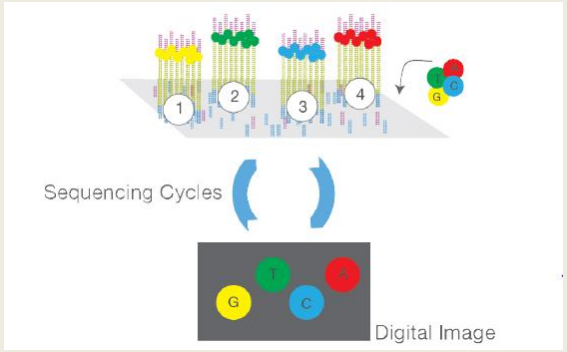
Ilumina Sequencing by Synthesis Workflow

Library Preparation – prepares target DNA with indices and sequence adaptors

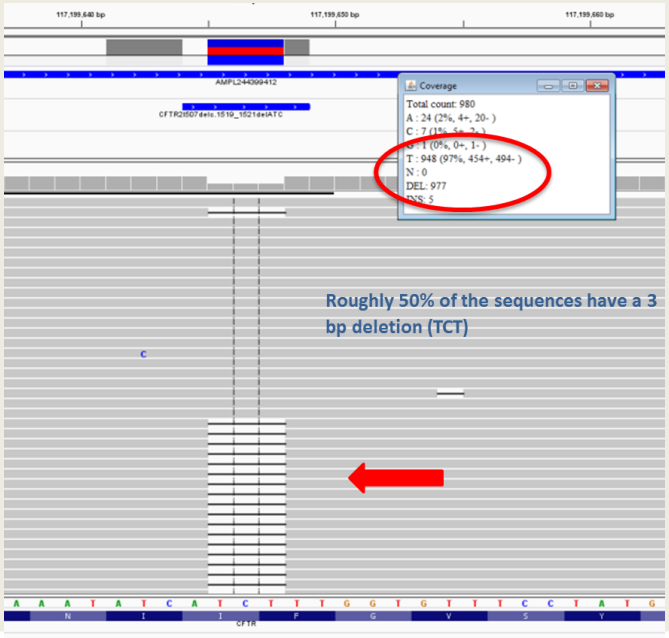
Illumina MiSeq Flow cell



Amplification of target regions on flow cell



DNA sequencing – a single fluorescent nucleotide is added, an image is taken, and the process repeats

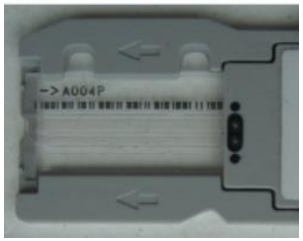


DNA sequence from each target is aligned

Images except sequence are from © Illumina with permission

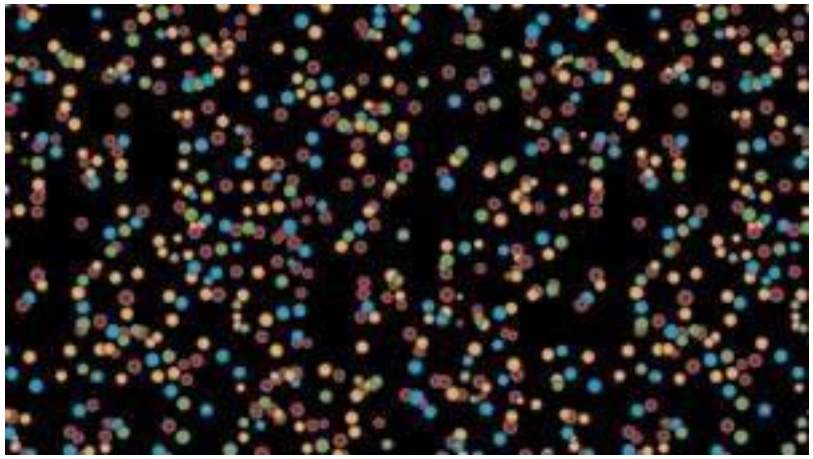
NGS Sequencing

- PCR amplify sample (opt.)
- Immobilize and amplify single molecules on a solid surface

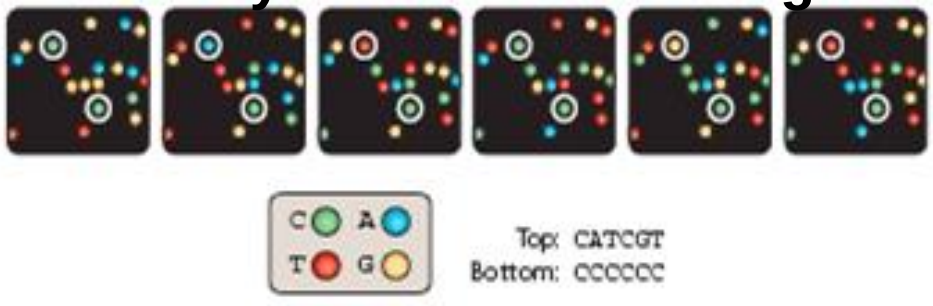


- Reversible terminator sequencing with 4 color dye-labelled nucleotides

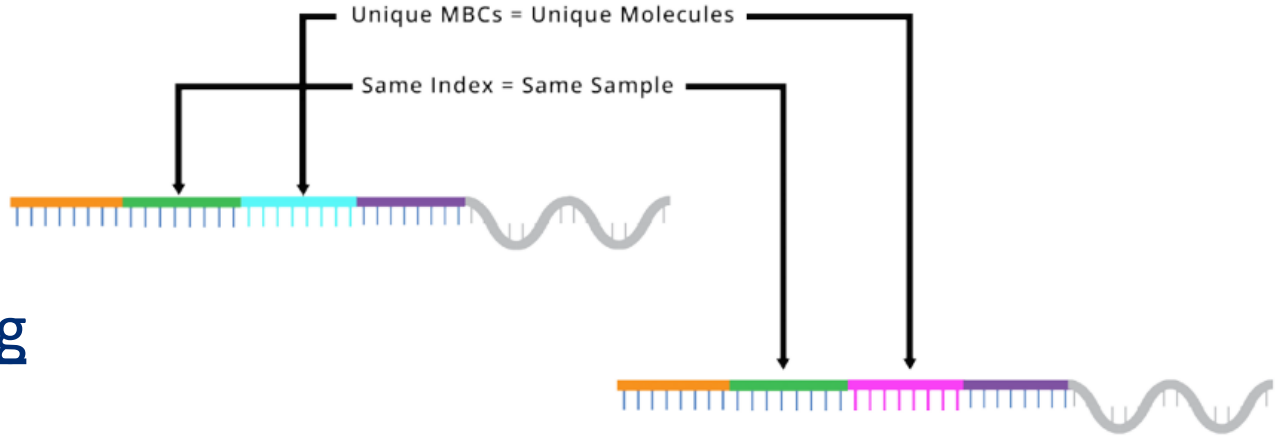
4 different images merged



6 cycles w/ base-calling



Barcoding



Sample Barcoding

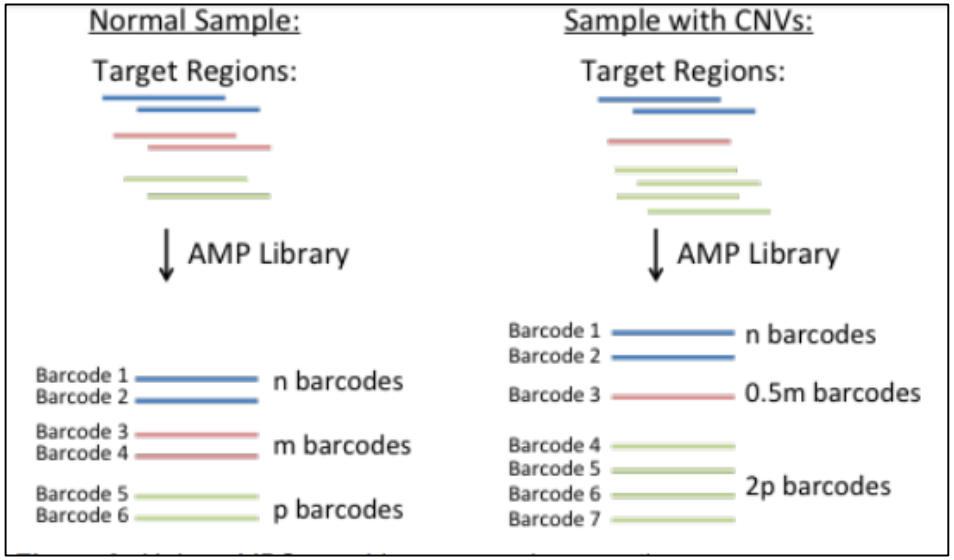
- Think of it as multiple Sanger sequencing reactions being imaged simultaneously.
- Saves time and money.

Molecular Barcoding (MBCs)

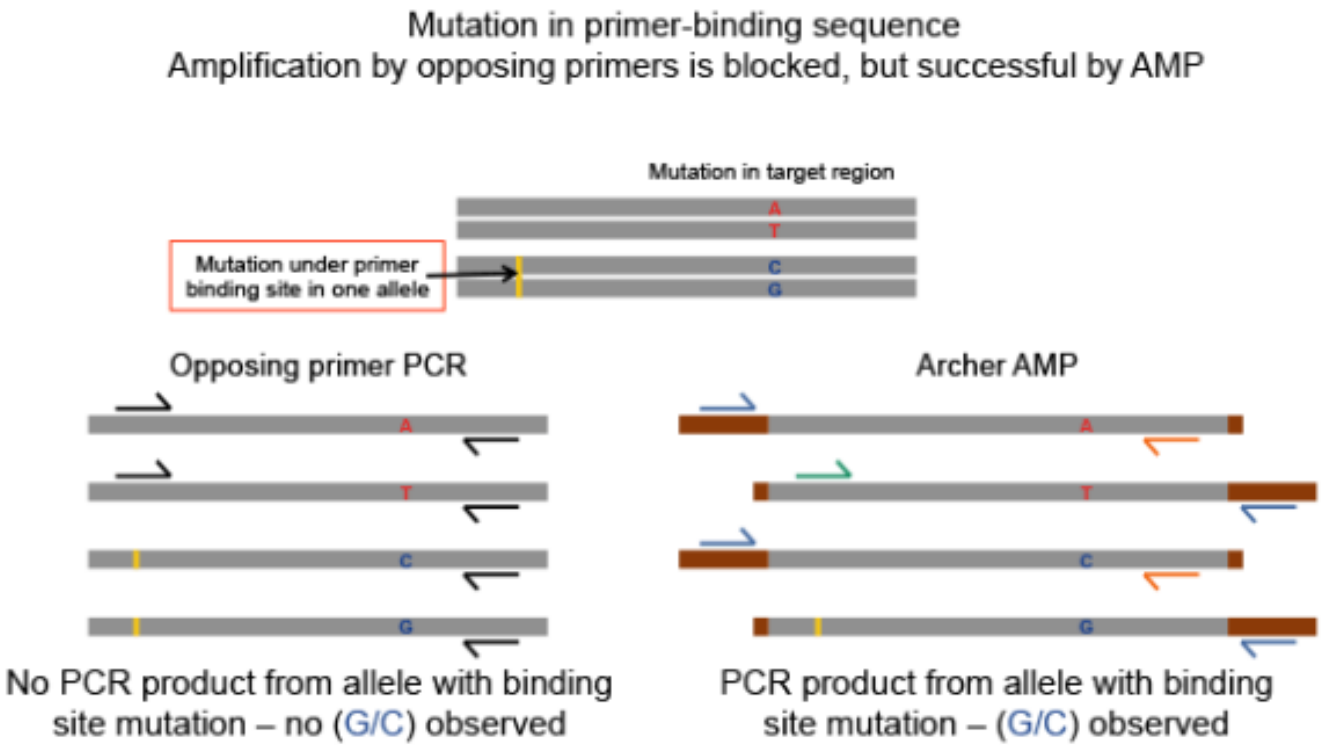
- Identify unique DNA molecules.
- Error correction.
- Copy-number variant (CNV) detection.

NYS Multi-Gene Panel

- Developing a multi-gene panel (VariantPlex assay).
  - Anchored Multiplex PCR (AMP)
  - MBCs for detection of exon-level deletions/ duplications



VariantPlex Assay



- Anchored Multiplex PCR (AMP)
- Minimum recommended DNA input 10 ng

Gene	Condition
MCCC1	(3-methylcrotonyl CoA carboxylase deficiency)
MCCC2	(3-methylcrotonyl CoA carboxylase deficiency)
HBB	(hemoglobinopathies)
HBA1	(alpha-thalassemia)
HBA2	(alpha-thalassemia)
SLC22A5	(primary carnitine deficiency)
GALT	(galactosemia)
IDUA	(mucopolysaccharidosis type 1)
GAMT	(guanidinoacetate methyltransferase deficiency)
ACADVL	(very long-chain acyl CoA dehydrogenase deficiency)
ACADM	(medium long-chain acyl CoA dehydrogenase deficiency)
ETFA	(multiple acyl-CoA carboxylase deficiency)
ETFB	(multiple acyl-CoA carboxylase deficiency)
ETFDH	(multiple acyl-CoA carboxylase deficiency)
GAA	(Pompe disease)
ABCD1	(X-Adrenoleukodystrophy)
GALC	(Krabbe disease)
CPT2	(carnitine palmitoyltransferase deficiency, type 2)
HMGCL	HMG
HSD17B10	2M3HBA
AUH	3MGA

39 – Gene Panel

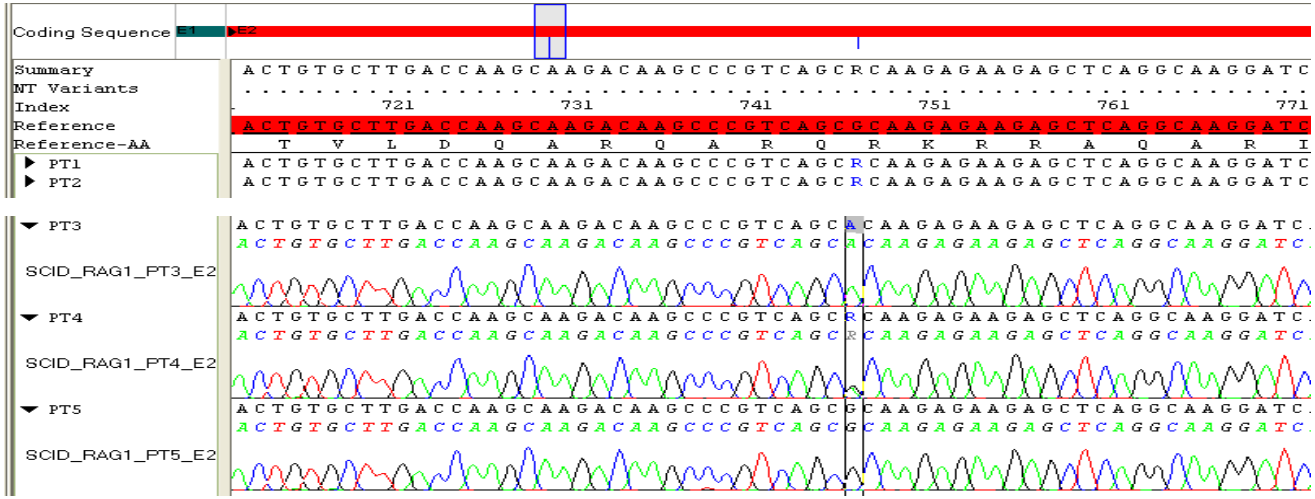
CORO1A	CD3Z	ATM	LIG4	DOCK2
PRKDC	DCLRE1C	CHD7	NHEJ1	GATA2
TBX1	IL2RG	MTHFD1	RAC2	IGHM
ADA	IL7RA	MTR	CD3G	BTK
AK2	JAK3	RMRP	STAT5B	CD40LG
CD3D	RAG1	SLC46A1	ZAP70	WAS
CD3E	RAG2	DOCK8	PNP	DKC1
CD45	FOXN1	NBN	BLNK	

Genes involved in T-cell and B-cell deficiencies and syndromes

Adding 16

Sanger vs Next Generation Sequencing

Single Sequence vs. Parallel Sequence



Sanger sequencing

NGS sequencing

BIOINFORMATICS **APPLICATIONS NOTE** *Vol. 25 no. 16 2009, pages 2078–2079 doi:10.1093/bioinformatics/btp352*

Sequence analysis

The Sequence Alignment/Map format and SAMtools

Heng Li[1,†], Bob Handsaker[2,†], Alec Wysoker[2], Tim Fennell[2], Jue Ruan[3], Nils Homer[4], Gabor Marth[5], Goncalo Abecasis[6], Richard Durbin[1,*] and 1000 Genome Project Data Processing Subgroup[7]

NEW YORK STATE Department of Health | Wadsworth Center

Targeted Next-generation Sequencing Overview

Sample QC

- DNA extraction and purification
- Quantification (optional)

Library Prep

Ampliseq Custom Amplicon (Illumina)	AmpliSeq (ThermoFisher – Ion)	VariantPlex (ArcherDX)	Accel-Amplicon (Swift)
Multiplex PCR	Multiplex PCR	Anchored multiplex PCR	Multiplex PCR

Purify Library

- Remove unincorporated primers, residual DNA and contaminates
- Quantify & normalize libraries
  - qPCR Quantify – check if library preparation worked and allows for normalization across samples
  - Normalization beads – as an alternative to quantification of each library, bead purification can capture a specific amount of library per sample

MiSeq or Ion Run

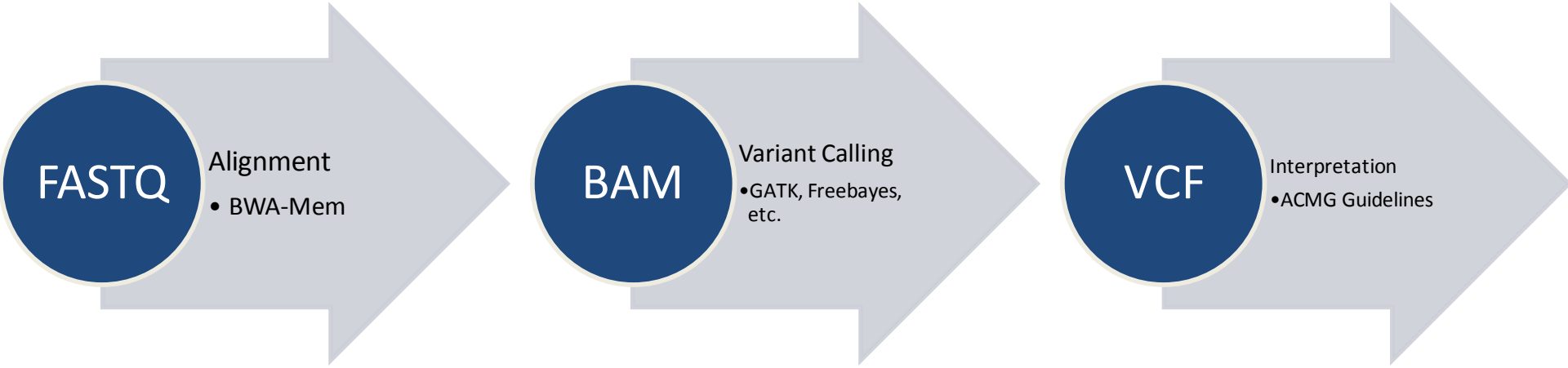
- Generate sequence data - FASTQ files
- Demultiplex - sort reads by sample barcode(s) and output to a file specific for that sample

Analysis

AmpliSeq for Illumina (Illumina)	AmpliSeq (ThermoFisher – Ion)	VariantPlex (ArcherDX)	Accel-Amplicon (Swift)
MiSeq Reporter	Torrent Server	Archer Analysis	VarSome (Cloud based)

Each company offer some basic analysis software with their instrument or panels. Some have more functionality than others.

NGS Analysis Workflow (Pipeline)



Pathogenic

NM_000033.3(ABCD1):c.1390C>T (p.Arg464Ter) – Heterozygous

NM_000033.3(ABCD1):c.1771C>T (p.Arg591Trp) - Heterozygous

VOUS

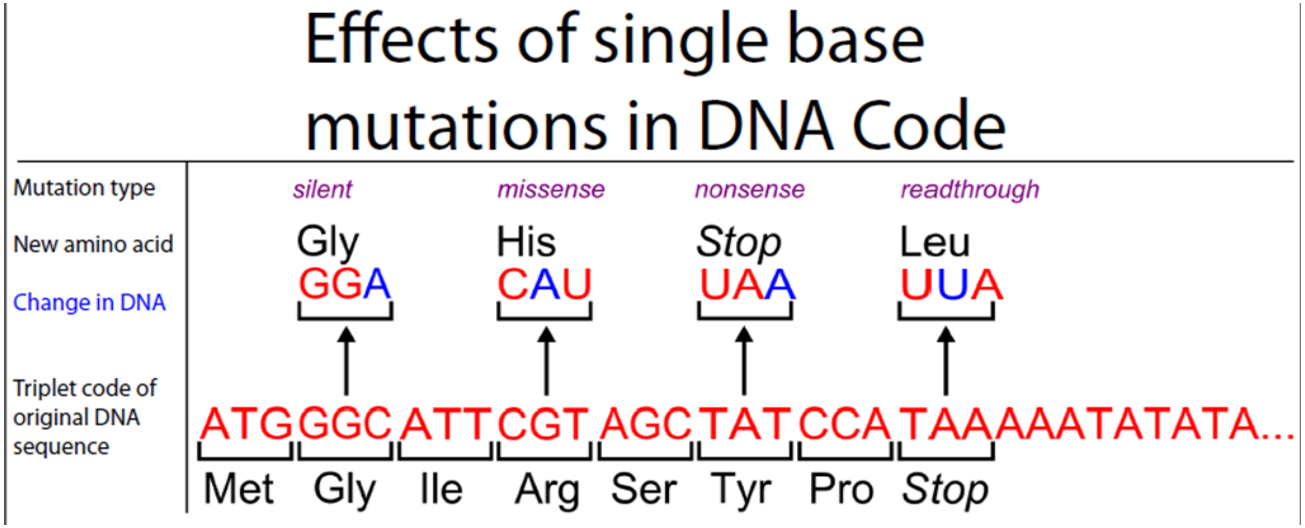
NM_000033.3(ABCD1):c.1780+4G>A - Heterozygous

Two NGS Platforms Used for SCID Sequencing

Illumina MiSeq – ~~TruSeq Custom Amplicon~~/ AS panel

ThermoFisher Chef/S5 AmpliSeq (AS) panel

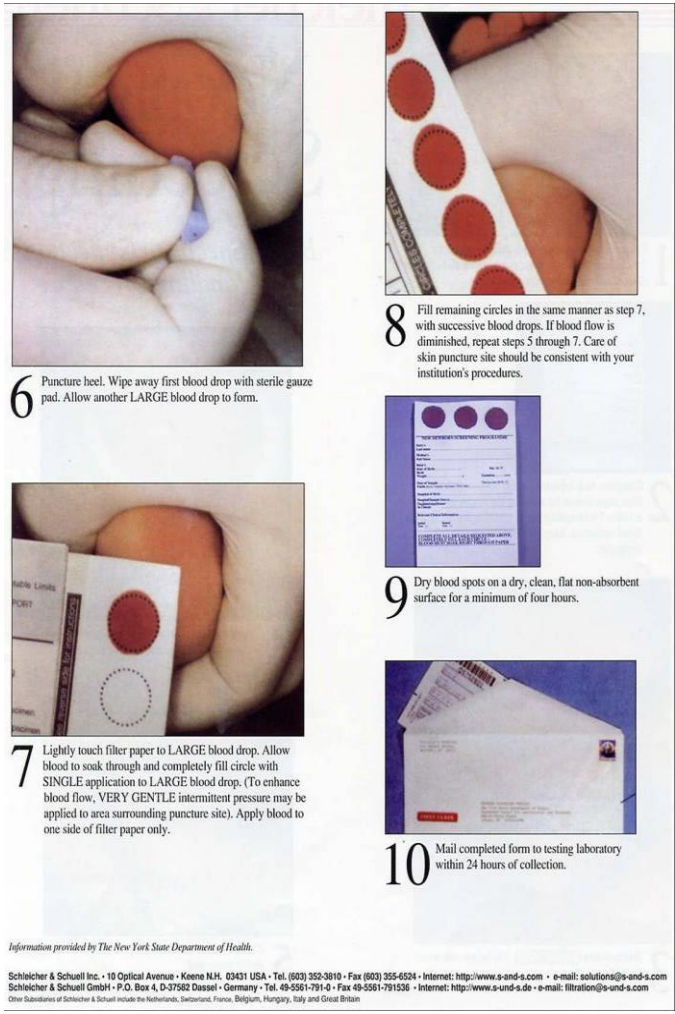
Types of Variants



Dried Blood Spot (DBS) "Guthrie" Card

- S&S® 903™ Cotton Paper
- 3.1 mm DBS ~ 3.1 uL whole blood.
- Benefits:
  - Rapid absorption (~ 10 seconds)
  - Easy transportation
  - Blood constituents "easily" eluted
  - Typeable DNA from 3-15 years.

- > 2500 articles in PubMed
- > 14 Million Google "hits"



NEW YORK STATE | Department of Health | Wadsworth Center

Clin Chem 59:7; 1045-1051 (2013)

Cost-Effective and Scalable DNA Extraction Method from Dried Blood Spots.

Saavedra-Matiz CA, Isabelle JT, Biski CK, Duva SJ, Sweeney ML, Parker AL, Young AJ, Diantonio LL, Krein LM, Nichols MJ, Caggana M.

Newborn Screening Program, Division of Genetics, Wadsworth Center, New York State Department of Health, Albany, NY.

Abstract

BACKGROUND: Dried blood spot (DBS) samples have been widely used in newborn screening (NBS) for the early identification of disease to facilitate the presymptomatic treatment of congenital diseases in newborns. As molecular genetics knowledge and technology progresses, there is an increased demand on NBS programs for molecular testing and a need to establish reliable, low-cost methods to perform those analyses. Here we report a flexible, cost-efficient, high-throughput DNA extraction method from DBS adaptable to small- and large-scale screening settings.METHODS: Genomic DNA (g.DNA) was extracted from single 3-mm diameter DBS by the sequential use of red cell lysis, detergent-alkaline, and acid-neutralizing buffers routinely used in whole blood and plant tissue DNA extractions. We performed PCR amplification of several genomic regions using standard PCR conditions and detection methods (agarose gel, melting-curve analysis, TaqMan-based assays). Amplicons were confirmed by BigDye® Terminator cycle sequencing and compared with reference sequences.RESULTS: High-quality g.DNA was extracted from hundreds of DBS, as proven by mutation detection of several human genes on multiple platforms. Manual and automated extraction protocols were validated. Quantification of g.DNA by Oligreen® fluorescent nucleic acid stain demonstrated a normal population distribution closely corresponding with white blood cell counts detected in newborn populations.CONCLUSIONS: High-quality, amplifiable g.DNA is extractable from DBSs. Our method is adaptable, reliable, and scalable to low- and high-throughput NBS at low cost ($0.10/sample). This method is routinely used for molecular testing in the New York State NBS program.

Clin Chem 59:7; 1011-1013 (2013) **Editorials**

Newborn Screening by Sequence and the Road Ahead.

Sondheimer N.

Department of Pediatrics, University of Pennsylvania, and Section of Biochemical Genetics, Children's Hospital Philadelphia, Philadelphia, PA.

Molecular Testing in Newborn Screening

- **Primary** molecular test
- When no other assay is available
  - Severe Combined Immunodeficiency (**SCID**)
  - Spinal Muscular Atrophy **(SMA)**

- **Second tier** molecular test
- **Increase sensitivity or specificity of primary assay**
  - Cystic Fibrosis (CF), Krabbe, **Pompe**, **MPSI**, etc.

- **Clarify an ambiguous result**
  - Hemoglobinopathies

- **Supplemental "Just in Time" assay**
  - Galactosemia, **X-ALD** (if not Pathogenic Variants > other PBD), MCAD

Second/Third tier DNA Sequencing Clinical Assessment vs Supplemental

- Pompe/MPS I/Krabbe/Cystic Fibrosis: Sanger sequencing of gene
  - Reduces referrals – contributes to the clinical assessment
  - Can provide information on classic infantile forms phenotype
  - Report detected

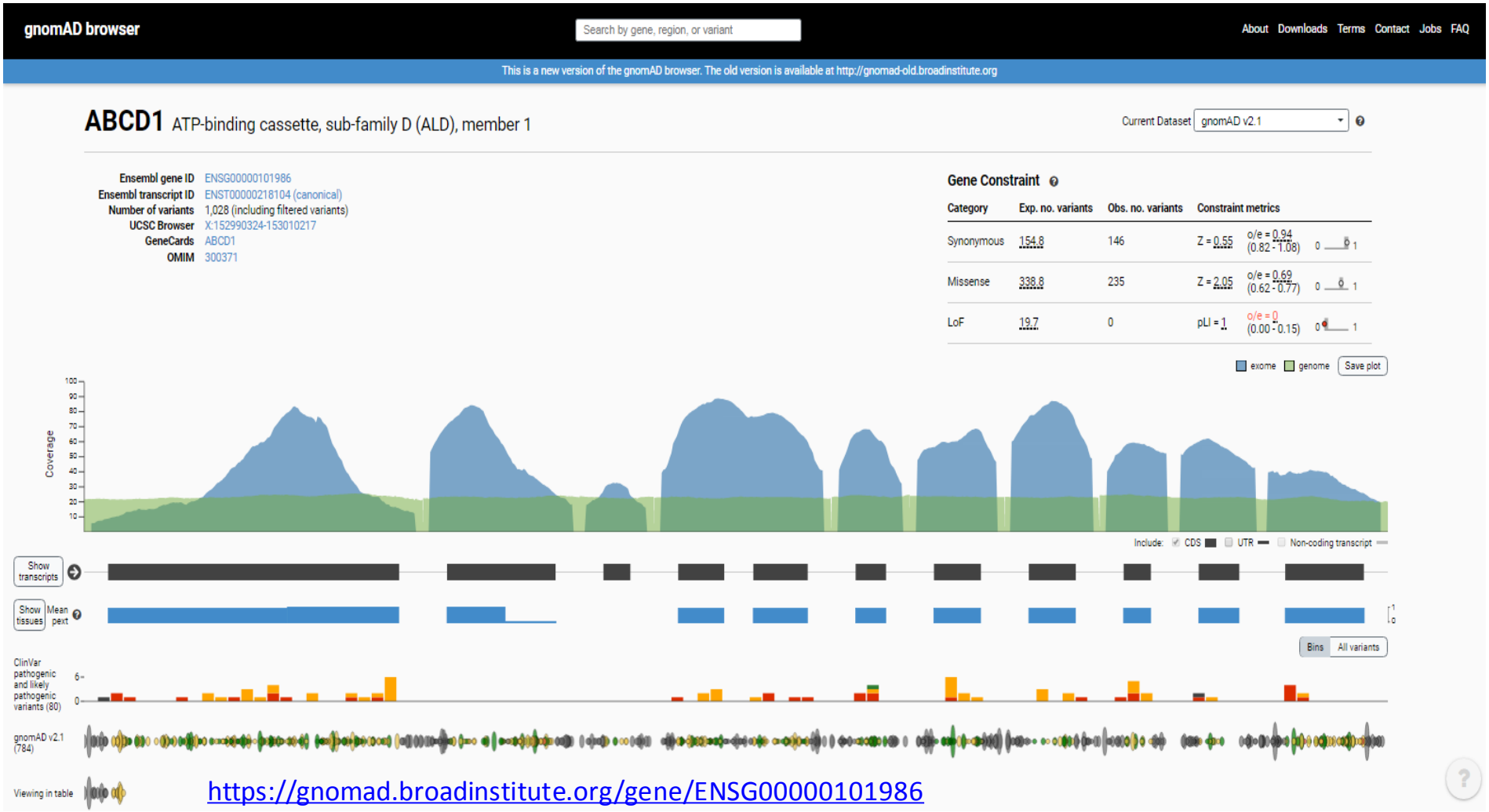
- ALD: full sequencing of ABCD1 gene
  - Not intended to reduce referrals – supplemental
  - Determine if females are ALD carriers
  - Determine if males have mutation
  - if no mutation, consider other PGD
  - For ABCD1, genotype (nor marker concentrations) do not correlate with phenotype!!

Tools to Classify DNA Sequence Variants

- Review of primary literature
- Databases
  - GnomAD - http://gnomad.broadinstitute.org/
  - ClinVar - https://www.ncbi.nlm.nih.gov/clinvar/
  - ExAC Browser - http://exac.broadinstitute.org/
  - Gene specific databases
    - MPS-1
      - http://mps1-database.org/
    - ALD
      - http://www.x-ald.nl/
    - Pompe
      - https://www.erasmusmc.nl/klinische_genetica/research/Iijnen/pompe_center/moleculaire_aspecten/?lang=en

Searching gnomAD Database for Your Variant

gnomAD – Genome Aggregation Database

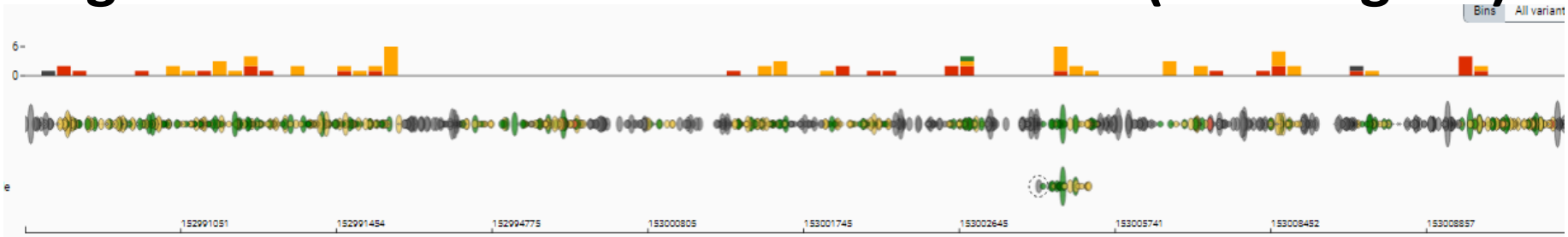


https://gnomad.broadinstitute.org/gene/ENSG00000101986

gnomAD – 125,748 exome sequences and 15,708 whole-genome sequences from 141,456 unrelated individuals

NEW YORK STATE Department of Health | Wadsworth Center

gnomAD Variant Search Results for ALD (*ABCD1* gene)



LoF only | Missense only | Synonymous only | Other only

Include filtered variants
SNPs
Indels

Search variant table

Export variants to CSV

† denotes a consequence that is for a non-canonical transcript

Variant ID	Source	Consequence	Annotation	Flags	Allele Count	Allele Number	Allele Frequency	Number of Homozygotes	Number Hemizygot
X-153005554-A-G	E	c.1497A>G(p.=)	synonymous		1	181555	5.508e-6	0	0
X-153005575-A-T	E	c.1518A>T(p.=)	synonymous		3	182241	1.646e-5	0	1
X-153005578-C-T	E G	c.1521C>T(p.=)	synonymous		5	204151	2.449e-5	0	2
X-153005583-A-G	E	p.Asn509Ser	missense		1	182301	5.485e-6	0	0
X-153005590-C-T	E	c.1533C>T(p.=)	synonymous		5	182133	2.745e-5	0	0
X-153005593-C-G	E	c.1536C>G(p.=)	synonymous		1	182294	5.486e-6	0	0
X-153005596-G-A	E G	c.1539G>A(p.=)	synonymous		2	204350	9.787e-6	0	1
X-153005599-C-T	E	c.1542C>T(p.=)	synonymous		3	182341	1.645e-5	0	1
X-153005605-G-A	E G	c.1548G>A(p.=)	synonymous		19244	204205	9.424e-2	636	7522
X-153005608-C-T	E	c.1551C>T(p.=)	synonymous		2	182410	1.096e-5	0	0
X-153005610-G-A	E	p.Arg518Gln	missense		1	182400	5.482e-6	0	0
X-153005624-C-A	G	p.Leu523Ile	missense		1	21901	4.566e-5	0	1
X-153005638-C-T	E G	c.1581C>T(p.=)	synonymous		44	204577	2.151e-4	0	13
X-153005639-G-A	E G	p.Gly528Ser	missense		8	204648	3.909e-5	0	1
X-153005645-G-A	E	p.Val530Met	missense		1	182794	5.471e-6	0	0
X-153005654-A-C	E	p.Lys533Gln	missense		1	182895	5.468e-6	0	0
X-153005668-G-C	E	p.Gln537His	missense		2	182857	1.094e-5	0	1
X-153005669-C-T	E	p.Arg538Cys	missense		1	182791	5.471e-6	0	0
X-153005670-G-A	E	p.Arg538His	missense		1	182755	5.472e-6	0	1

https://gnomad.broadinstitute.org/gene/ENSG00000101986

gnomAD Variant Population Frequency Results

- Request additional information

Annotations

This variant falls on 3 transcript(s) in 2 gene(s).

missense

- ABCD1
  - ENST00000218104 *
    HGVSp: p.Arg518Gln
    Polyphen: probably_damaging
    SIFT: deleterious
  - ENST00000443684
    HGVSp: p.Arg186Gln
    Polyphen: probably_damaging
    SIFT: deleterious

intron

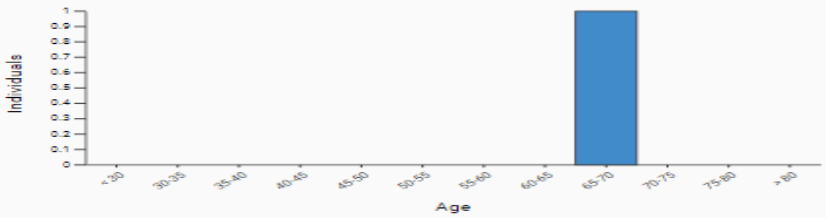
- U52111.14
  - ENST00000434284 *

Population Frequencies

Population	Allele Count	Allele Number	Number of Homozygotes	Number of Hemizygotes	Allele Frequency
European (non-Finnish)	1	81371	0	0	0.00001229
African	0	13154	0	0	0.000
Latino	0	27317	0	0	0.000
Ashkenazi Jewish	0	7394	0	0	0.000
East Asian	0	13845	0	0	0.000
European (Finnish)	0	15907	0	0	0.000
Other	0	4506	0	0	0.000
South Asian	0	18906	0	0	0.000
Total	**1**	**182400**	**0**	**0**	**0.000005482**

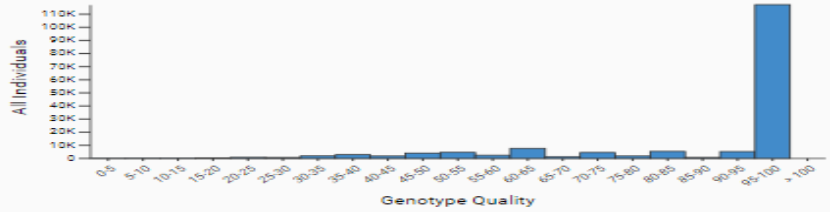
Include: Exomes Genomes

Age Distribution



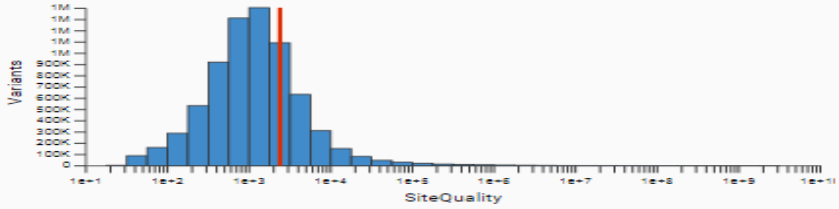
Heterozygotes Homozygotes

Genotype Quality Metrics



All Variant Carriers | Genotype Quality | Exomes Genomes

Site Quality Metrics



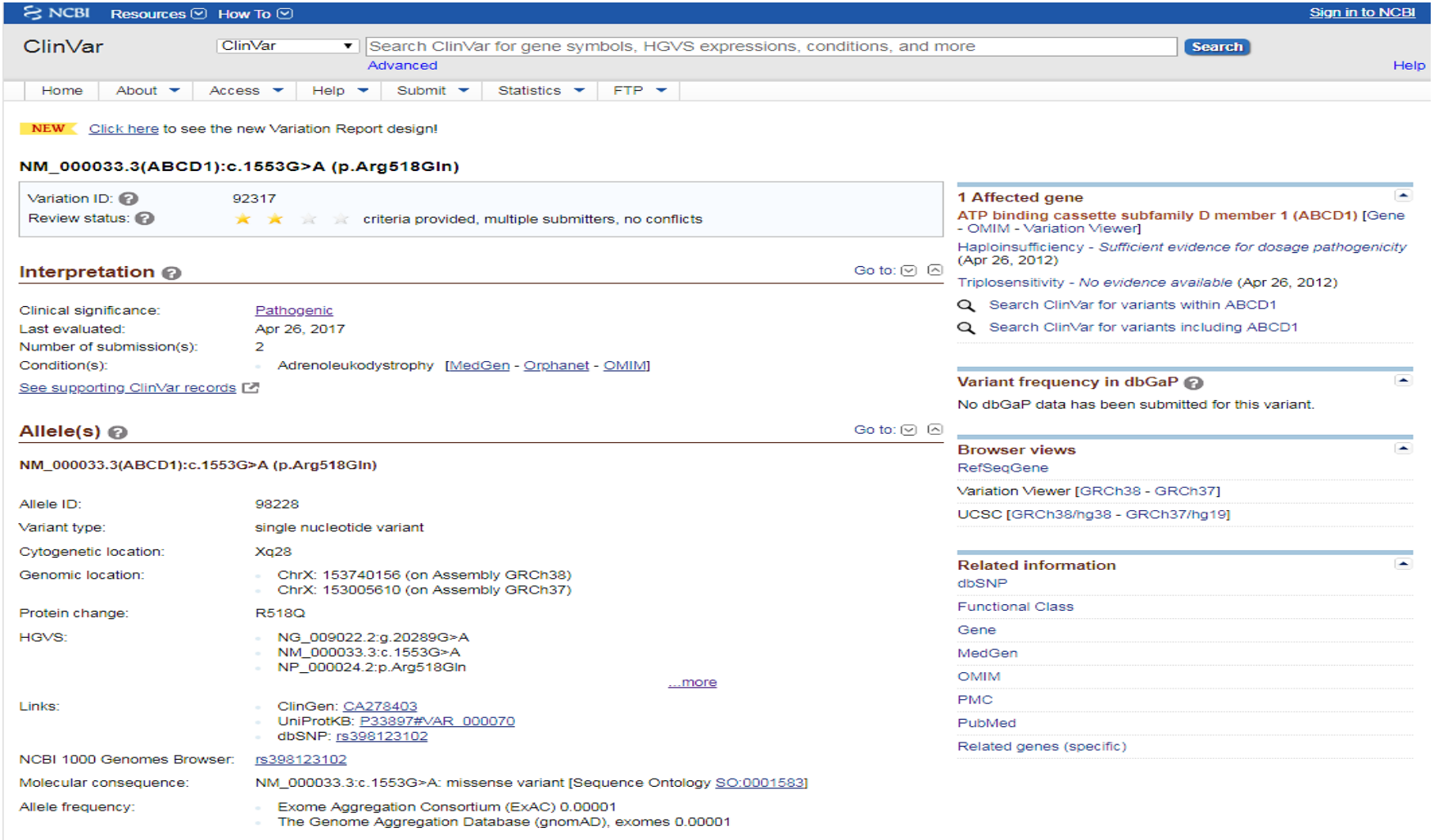
SiteQuality (2388) | Exomes Genomes

Note: These are site-level quality metrics, they may be unpredictable for multi-allelic sites

https://gnomad.broadinstitute.org/variant/X-153005610-G-A

ClinVar Results of ALD Variant

ClinVar – Database with Clinical and Genomic Information



NCBI Resources How To Sign in to NCBI

ClinVar | ClinVar | Search ClinVar for gene symbols, HGVS expressions, conditions, and more | Search | Advanced | Help

Home | About | Access | Help | Submit | Statistics | FTP

NEW Click here to see the new Variation Report design!

NM_000033.3(ABCD1):c.1553G>A (p.Arg518Gln)

Variation ID: 92317
Review status: ★★☆☆ criteria provided, multiple submitters, no conflicts

Interpretation Go to:

Clinical significance: Pathogenic
Last evaluated: Apr 26, 2017
Number of submission(s): 2
Condition(s): Adrenoleukodystrophy [MedGen - Orphanet - OMIM]
See supporting ClinVar records

Allele(s) Go to:

NM_000033.3(ABCD1):c.1553G>A (p.Arg518Gln)

Allele ID: 98228
Variant type: single nucleotide variant
Cytogenetic location: Xq28
Genomic location: ChrX: 153740156 (on Assembly GRCh38)
ChrX: 153005610 (on Assembly GRCh37)
Protein change: R518Q
HGVS: NG_009022.2:g.20289G>A
NM_000033.3:c.1553G>A
NP_000024.2:p.Arg518Gln
...more
Links: ClinGen: CA278403
UniProtKB: P33897#VAR_000070
dbSNP: rs398123102
NCBI 1000 Genomes Browser: rs398123102
Molecular consequence: NM_000033.3:c.1553G>A: missense variant [Sequence Ontology SO:0001583]
Allele frequency: Exome Aggregation Consortium (ExAC) 0.00001
The Genome Aggregation Database (gnomAD), exomes 0.00001

1 Affected gene
ATP binding cassette subfamily D member 1 (ABCD1) [Gene - OMIM - Variation Viewer]
Haploinsufficiency - *Sufficient evidence for dosage pathogenicity* (Apr 26, 2012)
Triplosensitivity - *No evidence available* (Apr 26, 2012)
Search ClinVar for variants within ABCD1
Search ClinVar for variants including ABCD1

Variant frequency in dbGaP
No dbGaP data has been submitted for this variant.

Browser views
RefSeqGene
Variation Viewer [GRCh38 - GRCh37]
UCSC [GRCh38/hg38 - GRCh37/hg19]

Related information
dbSNP
Functional Class
Gene
MedGen
OMIM
PMC
PubMed
Related genes (specific)

https://www.ncbi.nlm.nih.gov/clinvar/variation/92317/

Challenges of Sequencing

- **Major Challenge: Determining pathogenicity**
- **ACMG defined 5 categories to classify variants:**
  - **Known pathogenic**
  - **Likely to be pathogenic**
  - **Unknown significance**
  - **Likely to be benign**
  - **Benign**
- **Knowledge accruing daily, however the medical impact of most variants is unknown**

Acknowledgments

New York State Newborn Screening Program

Michele Caggana, ScD, FACMG
Colleen Stevens, PhD
Denise Kay, PhD
Robert Sicko, BS
DNA Lab Group

New York State Department of Health
Wadsworth Center
New York State Specialty Care Centers

APHL

CDC

Mi familia.