NGS в медицинской генетике

Школа анализа данных

Применение NGS-технологий в медицинской генетике. Эффективность. Ограничения

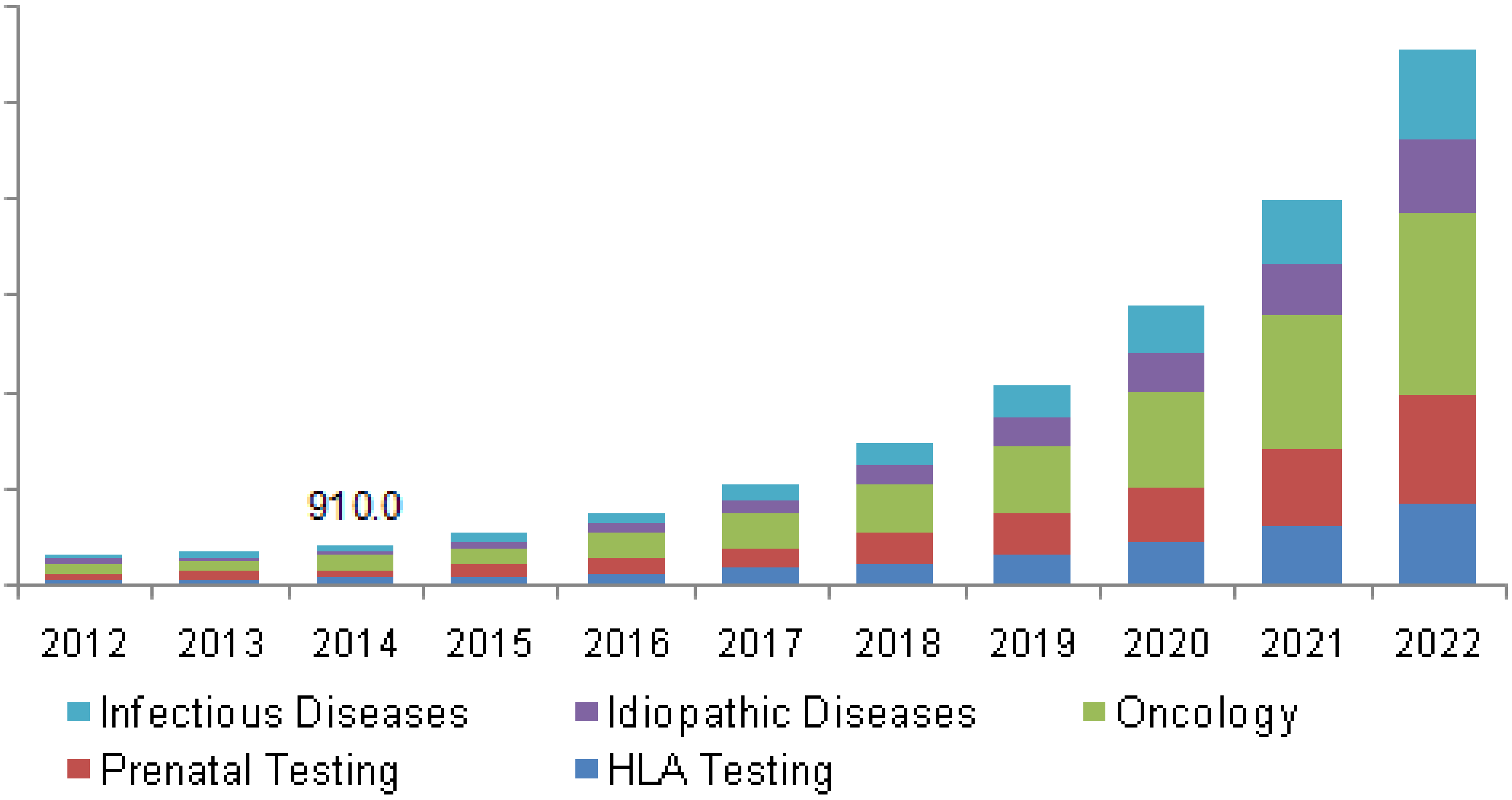
Скоблов Михаил Юрьевич

заведующий лабораторией функциональной геномики

ФГБНУ "Медико-генетический научный центр"

Осенняя школа MGNGS'2019

U.S. next generation sequencing market, by application, 2012-2022 (USD Million)



…ТИЧЕСКИЙ
…ЫЙ ЦЕНТР
МЕДИКО - ГЕНЕТИЧЕСКИЙ
НАУЧНЫЙ ЦЕНТР

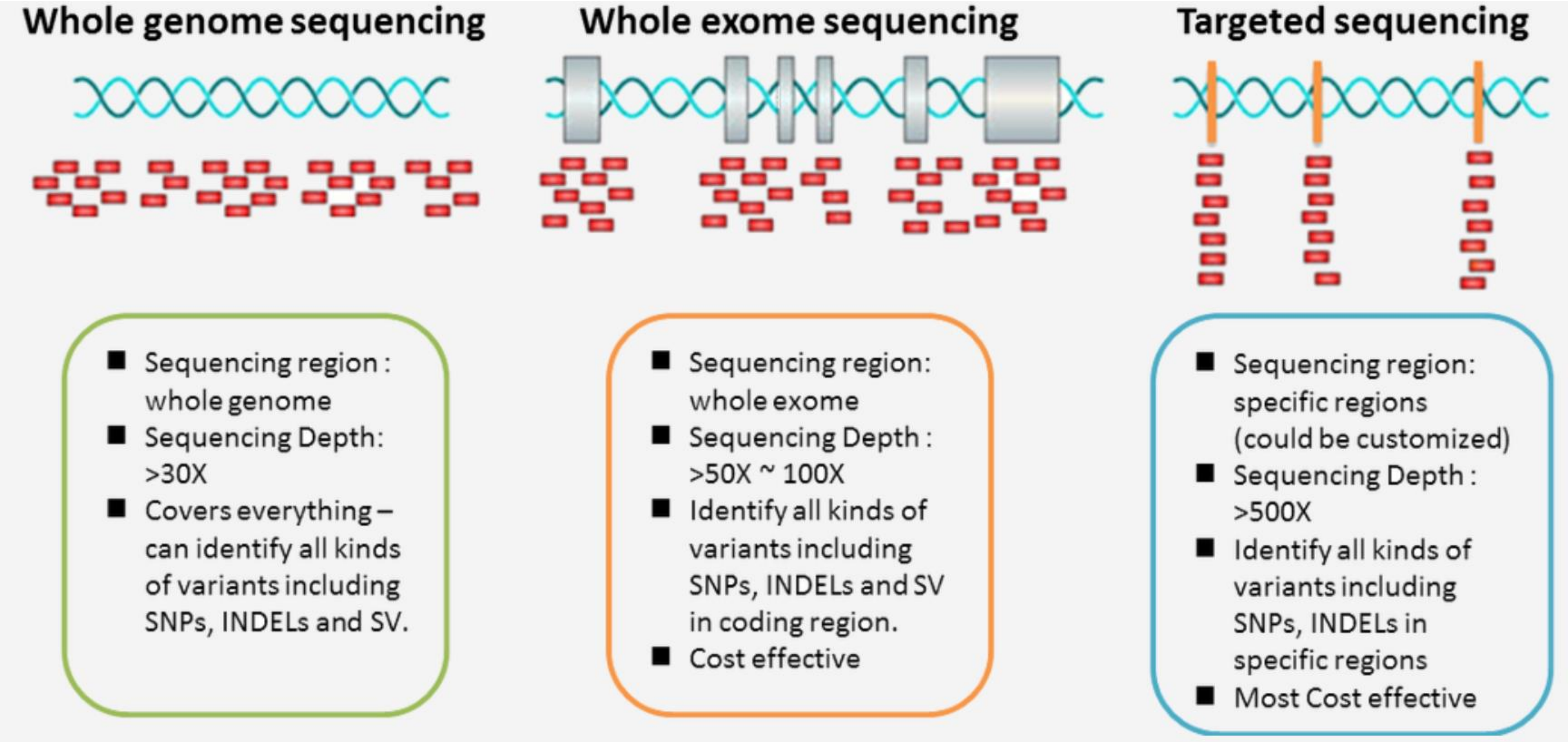
Из 13 тысяч молекулярных анализов в год:

- 20 геномов
- 700 экзомов
- 2000 панелей
- 10 тысяч секвенирований по Сэнгеру

- **NGS** – **N**ext **G**eneration **S**equencing

- **MPS** - **M**assively **P**arallel **S**equencing

Применение NGS в медицинской генетике



Whole genome sequencing

- Sequencing region : whole genome
- Sequencing Depth: >30X
- Covers everything – can identify all kinds of variants including SNPs, INDELs and SV.

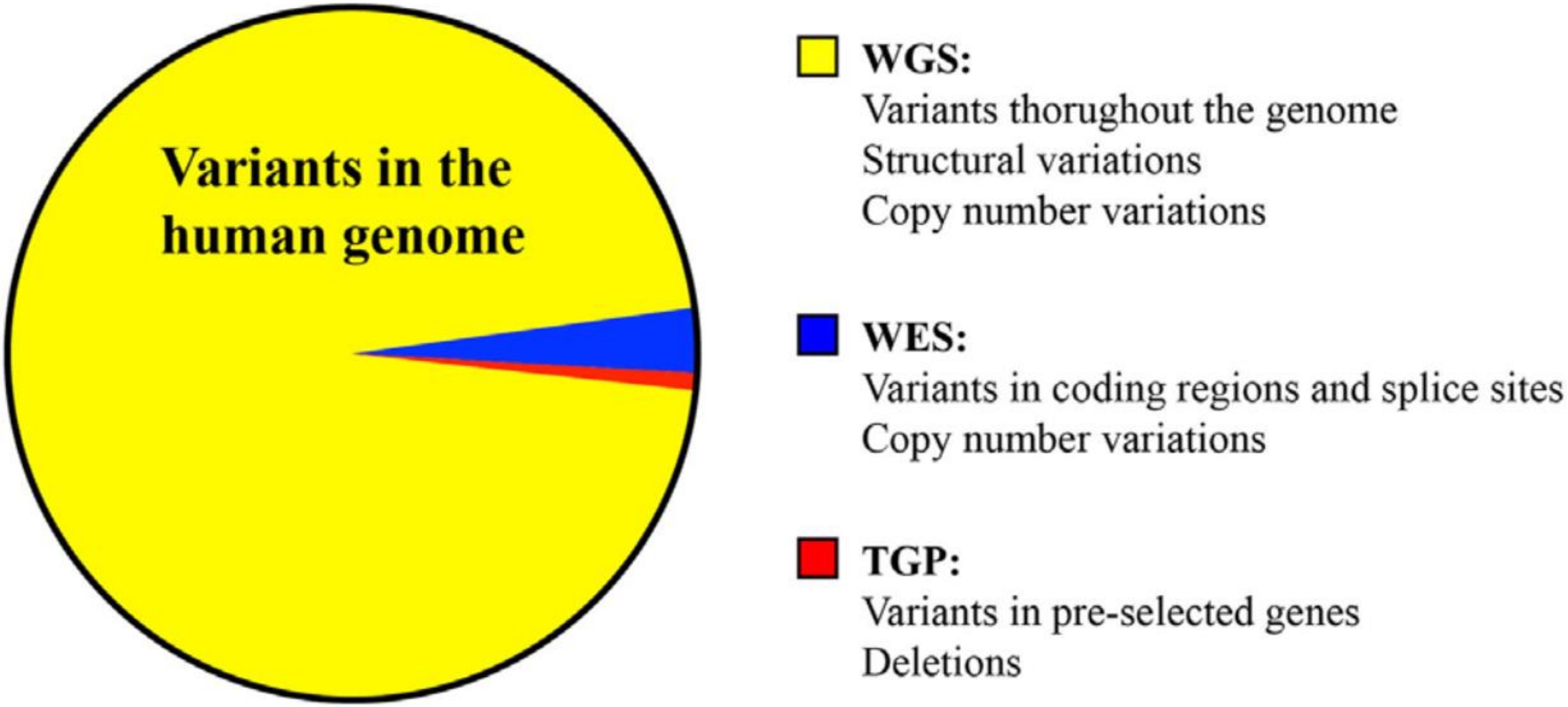
Whole exome sequencing

- Sequencing region: whole exome
- Sequencing Depth : >50X ~ 100X
- Identify all kinds of variants including SNPs, INDELs and SV in coding region.
- Cost effective

Targeted sequencing

- Sequencing region: specific regions (could be customized)
- Sequencing Depth : >500X
- Identify all kinds of variants including SNPs, INDELs in specific regions
- Most Cost effective

Применение NGS в медицинской генетике

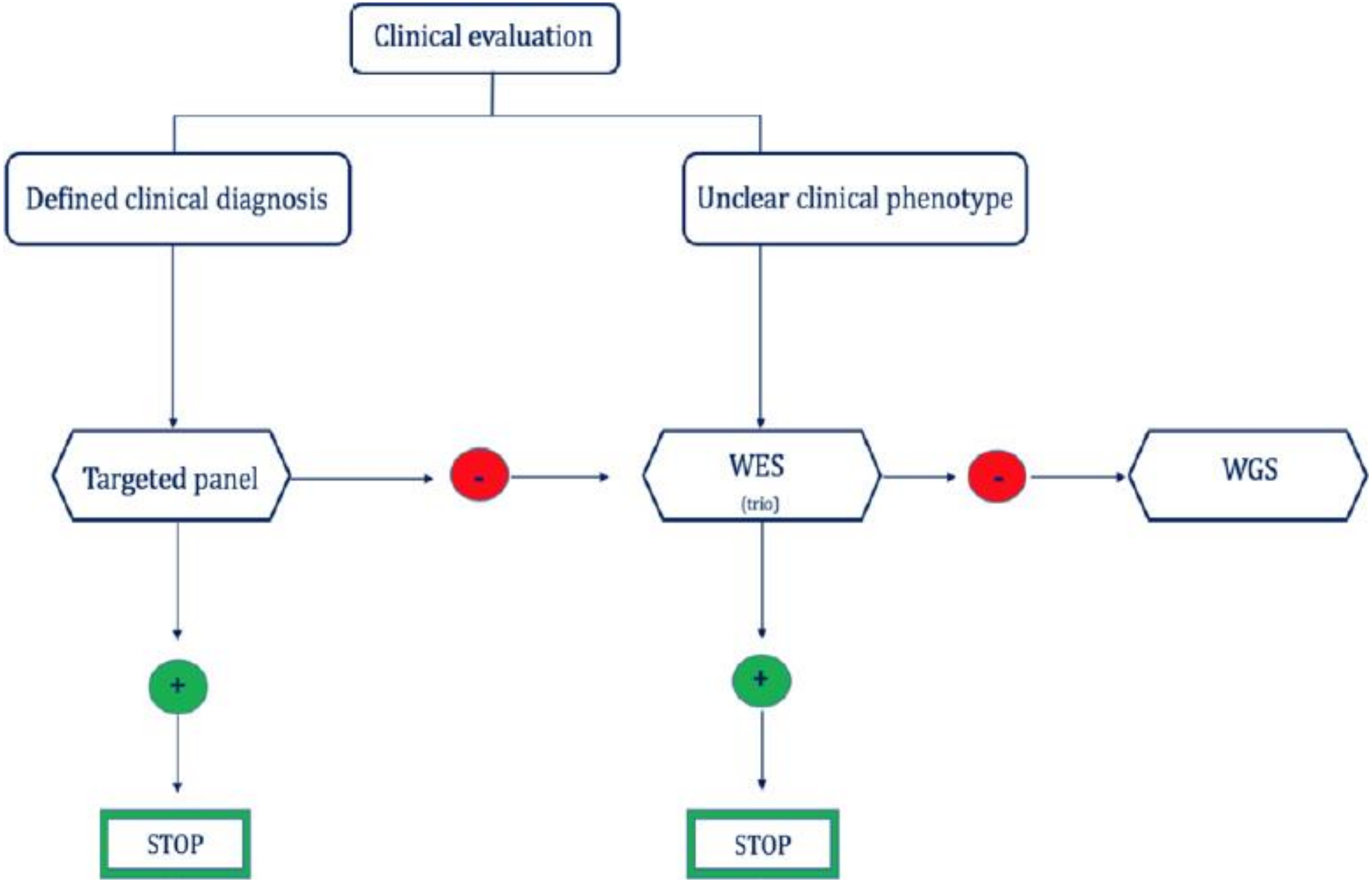


🟨 **WGS:**
Variants thorughout the genome
Structural variations
Copy number variations

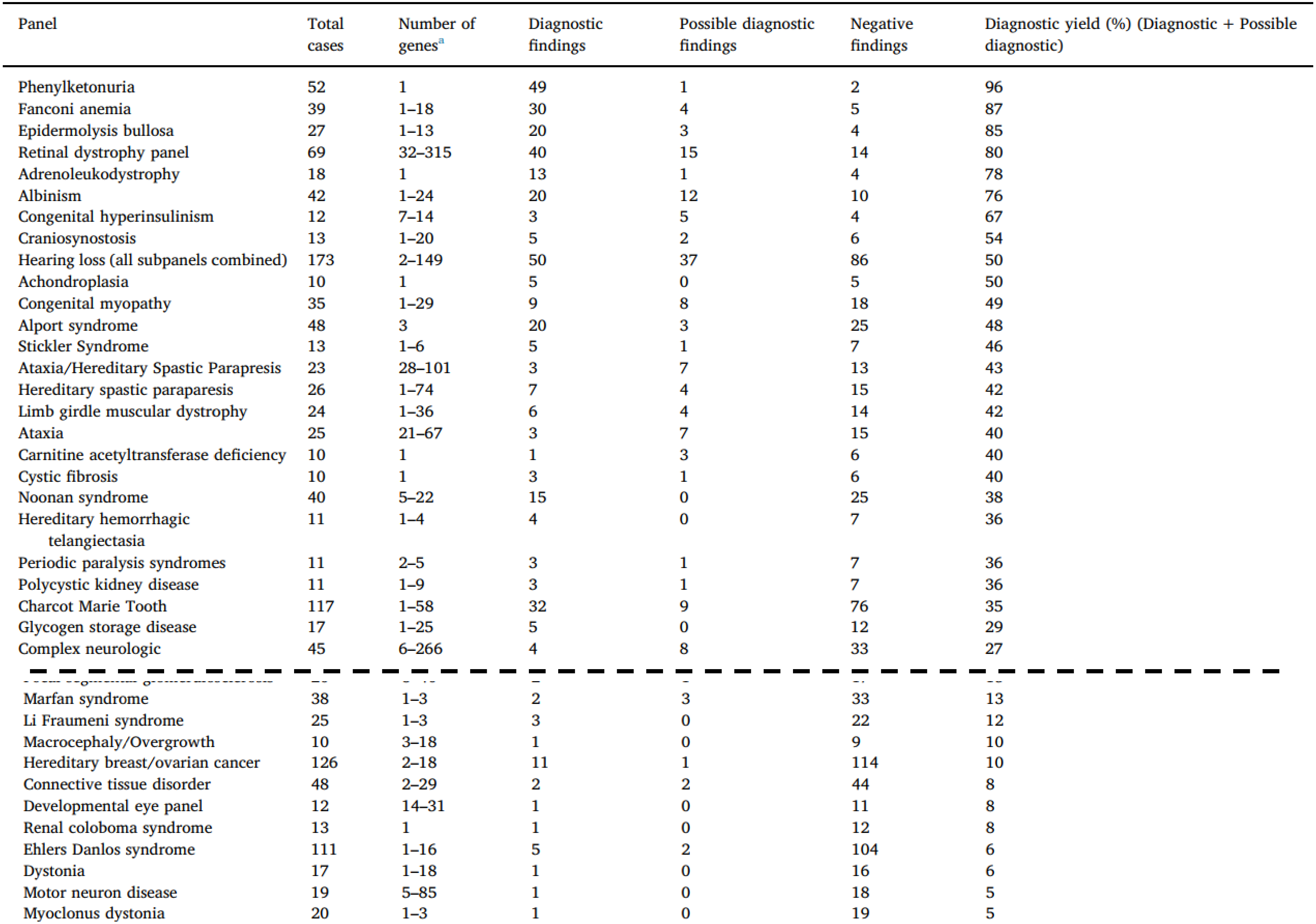
🟦 **WES:**
Variants in coding regions and splice sites
Copy number variations

🟥 **TGP:**
Variants in pre-selected genes
Deletions

Применение NGS в медицинской генетике

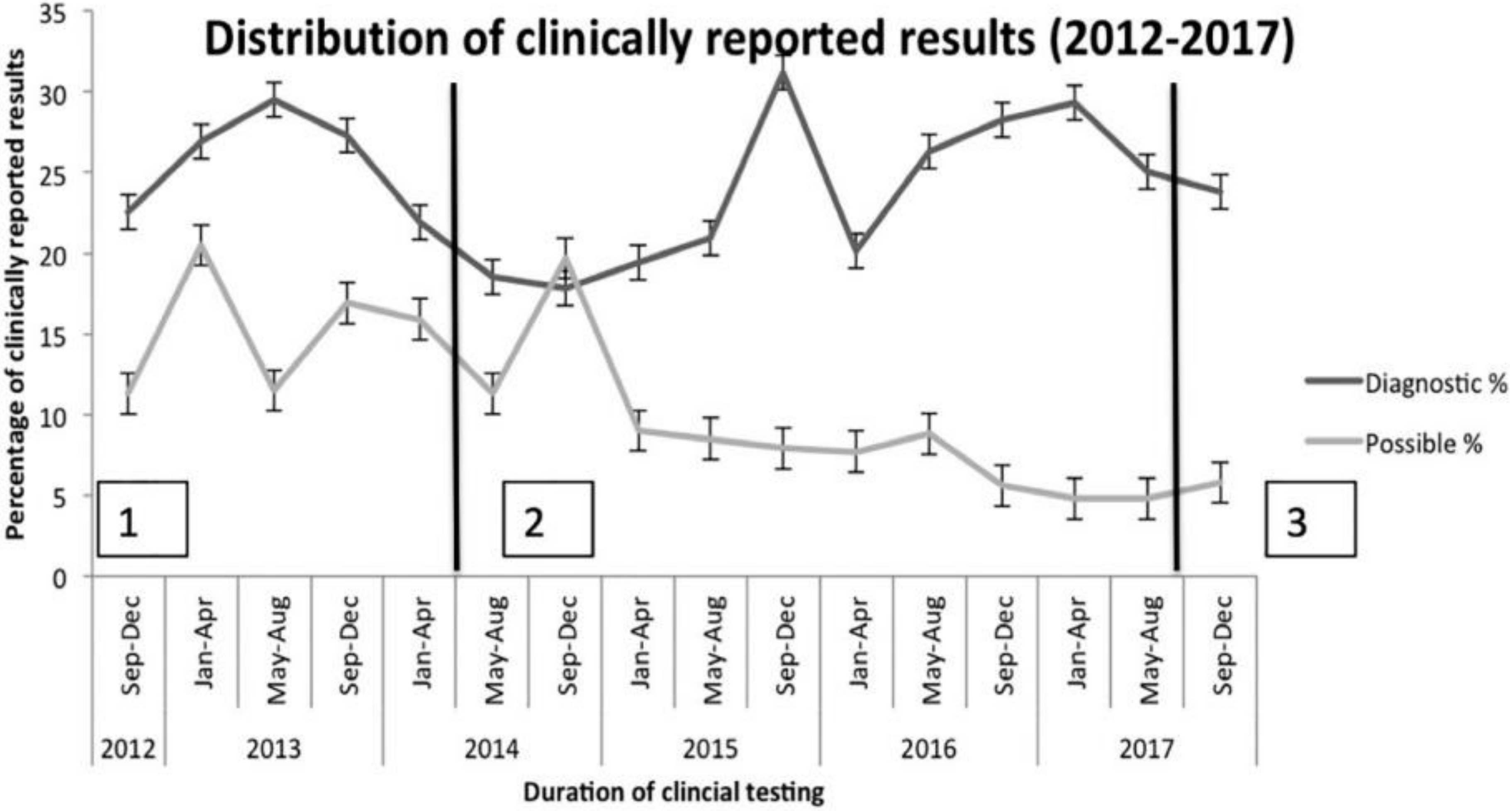


Diagnostic yield for commonly ordered gene panels (n ≥ 10 samples) at the University of Minnesota from 2012 to 2017.



Panel	Total cases	Number of genes[a]	Diagnostic findings	Possible diagnostic findings	Negative findings	Diagnostic yield (%) (Diagnostic + Possible diagnostic)
Phenylketonuria	52	1	49	1	2	96
Fanconi anemia	39	1–18	30	4	5	87
Epidermolysis bullosa	27	1–13	20	3	4	85
Retinal dystrophy panel	69	32–315	40	15	14	80
Adrenoleukodystrophy	18	1	13	1	4	78
Albinism	42	1–24	20	12	10	76
Congenital hyperinsulinism	12	7–14	3	5	4	67
Craniosynostosis	13	1–20	5	2	6	54
Hearing loss (all subpanels combined)	173	2–149	50	37	86	50
Achondroplasia	10	1	5	0	5	50
Congenital myopathy	35	1–29	9	8	18	49
Alport syndrome	48	3	20	3	25	48
Stickler Syndrome	13	1–6	5	1	7	46
Ataxia/Hereditary Spastic Parapresis	23	28–101	3	7	13	43
Hereditary spastic paraparesis	26	1–74	7	4	15	42
Limb girdle muscular dystrophy	24	1–36	6	4	14	42
Ataxia	25	21–67	3	7	15	40
Carnitine acetyltransferase deficiency	10	1	1	3	6	40
Cystic fibrosis	10	1	3	1	6	40
Noonan syndrome	40	5–22	15	0	25	38
Hereditary hemorrhagic telangiectasia	11	1–4	4	0	7	36
Periodic paralysis syndromes	11	2–5	3	1	7	36
Polycystic kidney disease	11	1–9	3	1	7	36
Charcot Marie Tooth	117	1–58	32	9	76	35
Glycogen storage disease	17	1–25	5	0	12	29
Complex neurologic	45	6–266	4	8	33	27
Marfan syndrome	38	1–3	2	3	33	13
Li Fraumeni syndrome	25	1–3	3	0	22	12
Macrocephaly/Overgrowth	10	3–18	1	0	9	10
Hereditary breast/ovarian cancer	126	2–18	11	1	114	10
Connective tissue disorder	48	2–29	2	2	44	8
Developmental eye panel	12	14–31	1	0	11	8
Renal coloboma syndrome	13	1	1	0	12	8
Ehlers Danlos syndrome	111	1–16	5	2	104	6
Dystonia	17	1–18	1	0	16	6
Motor neuron disease	19	5–85	1	0	18	5
Myoclonus dystonia	20	1–3	1	0	19	5

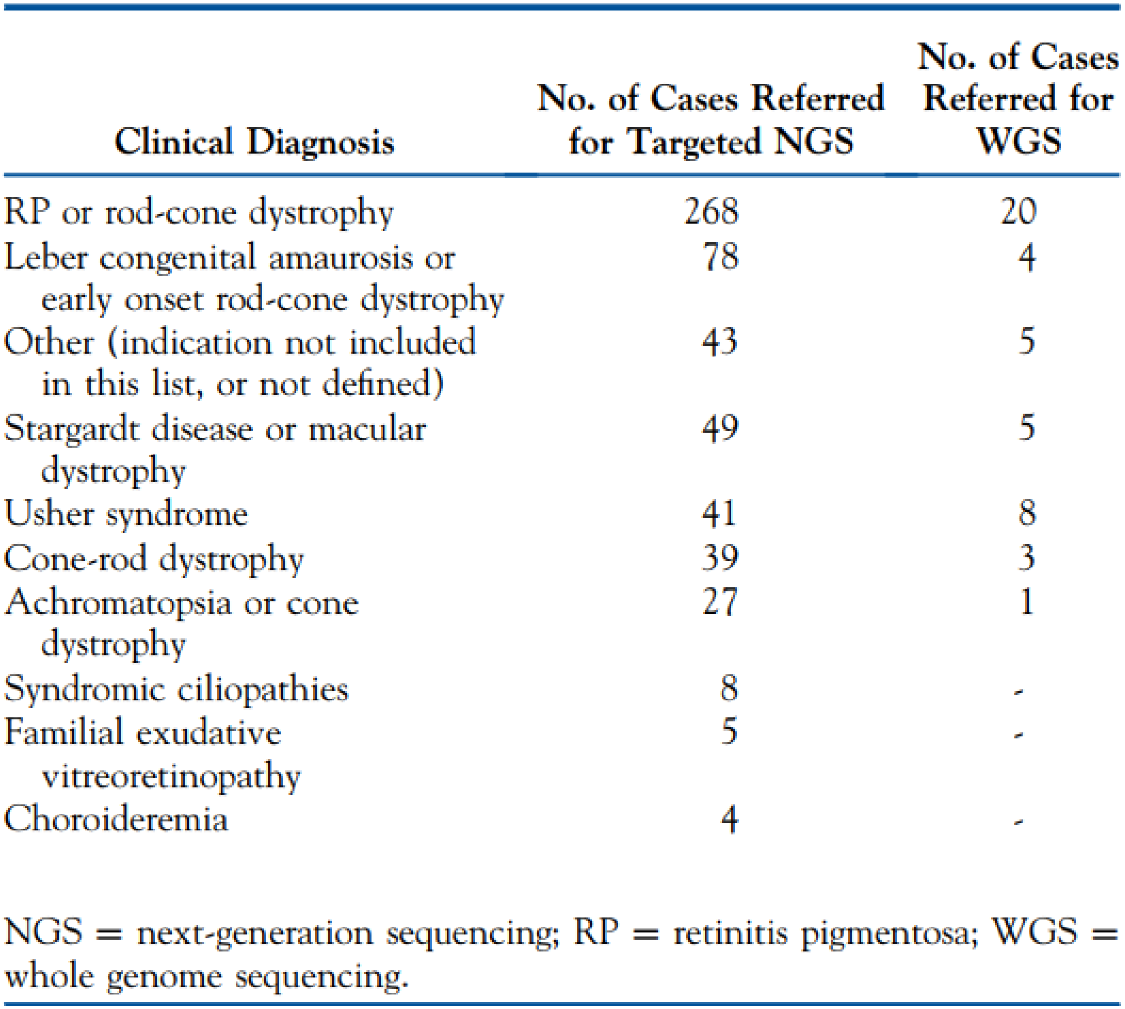
https://doi.org/10.1016/j.ymgmr.2019.100464



- The proportion of samples reported with diagnostic findings and as negative remained relatively stable over time while there was a decrease in the number of samples reported with possible diagnostic findings from 16.3% in 2013 to 5.13% in 2017.

Whole Genome Sequencing Increases Molecular Diagnostic Yield Compared with Current Diagnostic Testing for Inherited Retinal Disease

- 562 individuals underwent clinical analysis of genetic variation within 105 genes known to underpin IRD.
- A subset of 46 of 562 patients underwent WGS, and we compared mutation detection rates and molecular diagnostic yields.

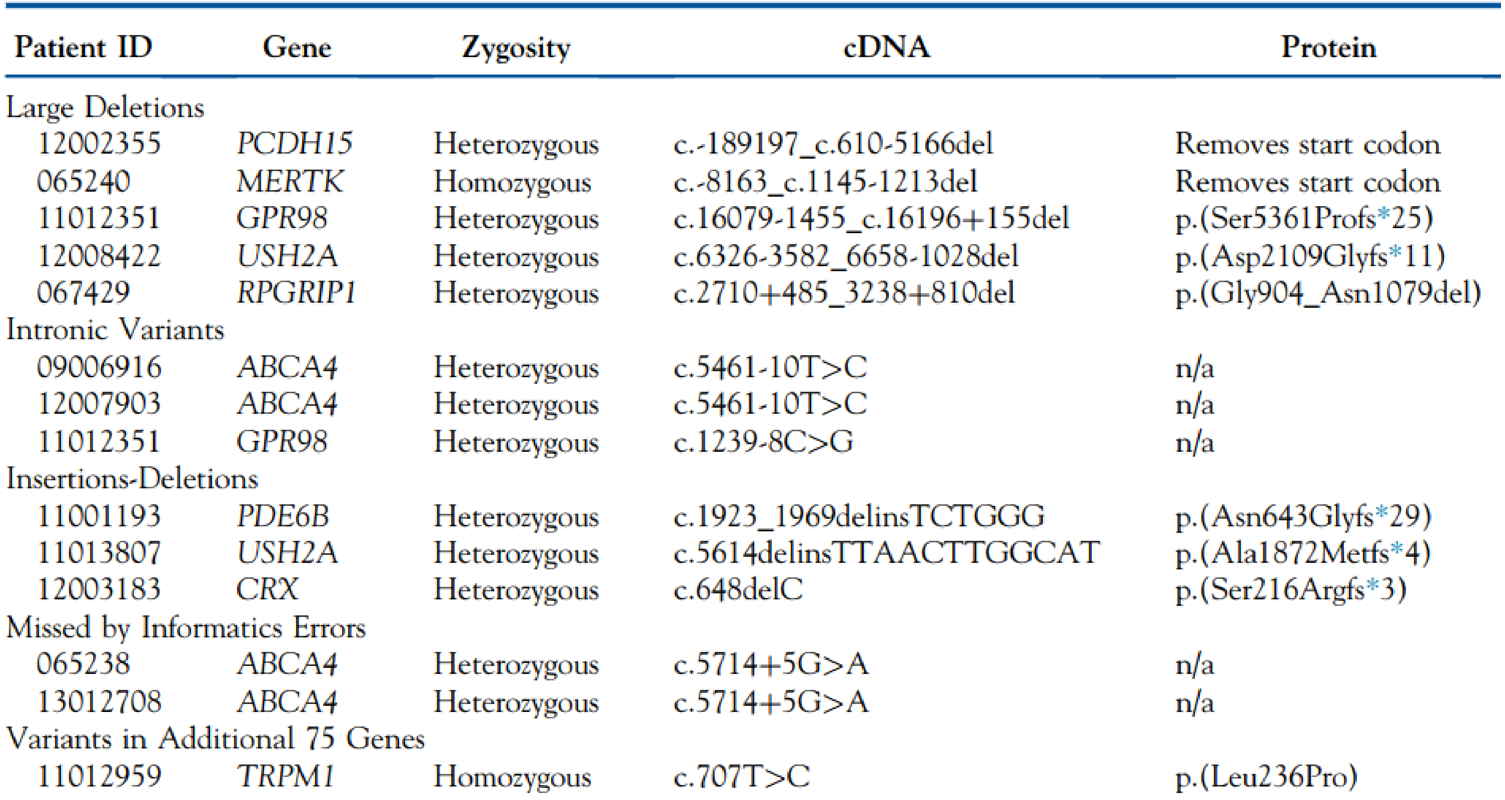


Clinical Diagnosis	No. of Cases Referred for Targeted NGS	No. of Cases Referred for WGS
RP or rod-cone dystrophy	268	20
Leber congenital amaurosis or early onset rod-cone dystrophy	78	4
Other (indication not included in this list, or not defined)	43	5
Stargardt disease or macular dystrophy	49	5
Usher syndrome	41	8
Cone-rod dystrophy	39	3
Achromatopsia or cone dystrophy	27	1
Syndromic ciliopathies	8	-
Familial exudative vitreoretinopathy	5	-
Choroideremia	4	-

NGS = next-generation sequencing; RP = retinitis pigmentosa; WGS = whole genome sequencing.

Whole Genome Sequencing Increases Molecular Diagnostic Yield Compared with Current Diagnostic Testing for Inherited Retinal Disease

- WGS identified 14 clinically relevant genetic variants through WGS that had not been identified by NGS diagnostic testing for the 46 individuals with IRD.
- Weighted estimates, accounting for population structure, suggest that WGS methods could result in an overall 29% uplift in diagnostic yield.

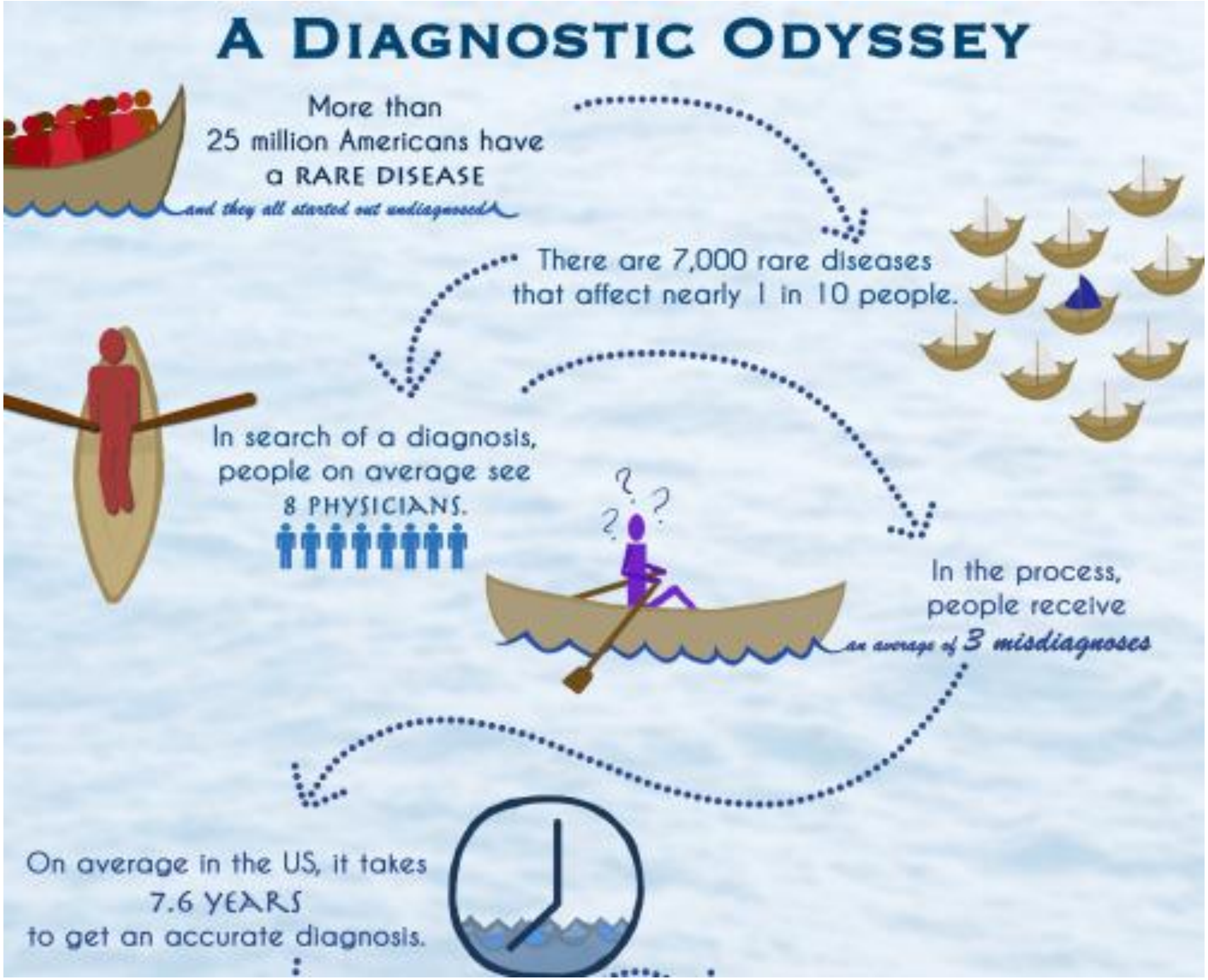


Patient ID	Gene	Zygosity	cDNA	Protein
Large Deletions				
12002355	*PCDH15*	Heterozygous	c.-189197_c.610-5166del	Removes start codon
065240	*MERTK*	Homozygous	c.-8163_c.1145-1213del	Removes start codon
11012351	*GPR98*	Heterozygous	c.16079-1455_c.16196+155del	p.(Ser5361Profs*25)
12008422	*USH2A*	Heterozygous	c.6326-3582_6658-1028del	p.(Asp2109Glyfs*11)
067429	*RPGRIP1*	Heterozygous	c.2710+485_3238+810del	p.(Gly904_Asn1079del)
Intronic Variants				
09006916	*ABCA4*	Heterozygous	c.5461-10T>C	n/a
12007903	*ABCA4*	Heterozygous	c.5461-10T>C	n/a
11012351	*GPR98*	Heterozygous	c.1239-8C>G	n/a
Insertions-Deletions				
11001193	*PDE6B*	Heterozygous	c.1923_1969delinsTCTGGG	p.(Asn643Glyfs*29)
11013807	*USH2A*	Heterozygous	c.5614delinsTTAACTTGGCAT	p.(Ala1872Metfs*4)
12003183	*CRX*	Heterozygous	c.648delC	p.(Ser216Argfs*3)
Missed by Informatics Errors				
065238	*ABCA4*	Heterozygous	c.5714+5G>A	n/a
13012708	*ABCA4*	Heterozygous	c.5714+5G>A	n/a
Variants in Additional 75 Genes				
11012959	*TRPM1*	Homozygous	c.707T>C	p.(Leu236Pro)

Полный экзом vs полный геном

	Полный экзом	Полный геном
Область исследование	95% код. части генома (1-2% от всего генома)	98% всего генома
Варианты из HGMD, которые не будут обнаружены	2,1%	0,3%
CNV	Возможна по анализу покрытия. Необходимо подтверждение другим методом	Разрешение точных границ перестроек выше ХМА высокого разрешения.
Другие перестройки	-	+
Некодирующая область	-	+
мтДНК	-	С прочтением 2000Х
Экспансия тринуклеотидных повторов	-	Возможен в некоторых случаях

A Diagnostic Odyssey



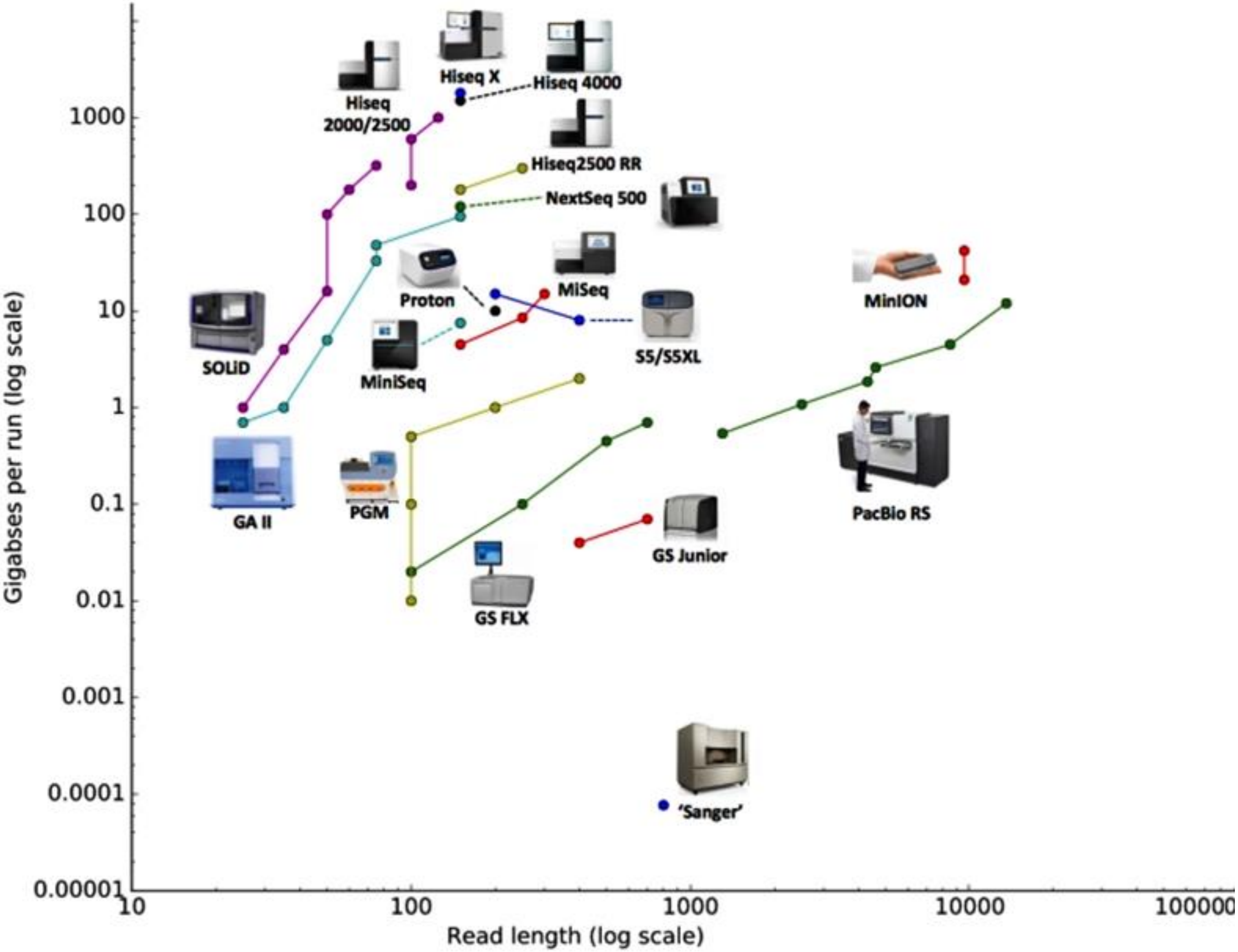
More than
25 million Americans have
a RARE DISEASE
and they all started out undiagnosed

There are 7,000 rare diseases
that affect nearly 1 in 10 people.

In search of a diagnosis,
people on average see
8 PHYSICIANS.

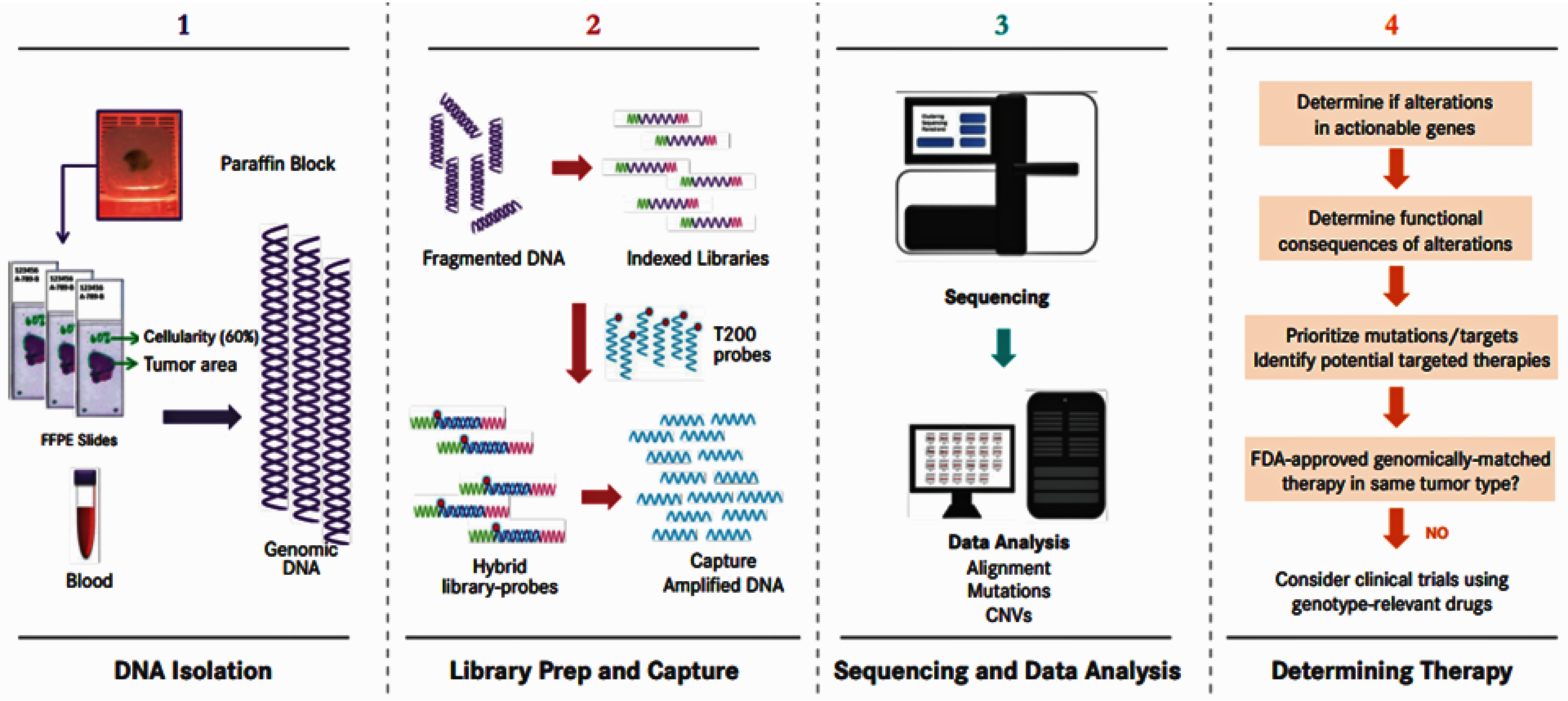
In the process,
people receive
an average of 3 misdiagnoses

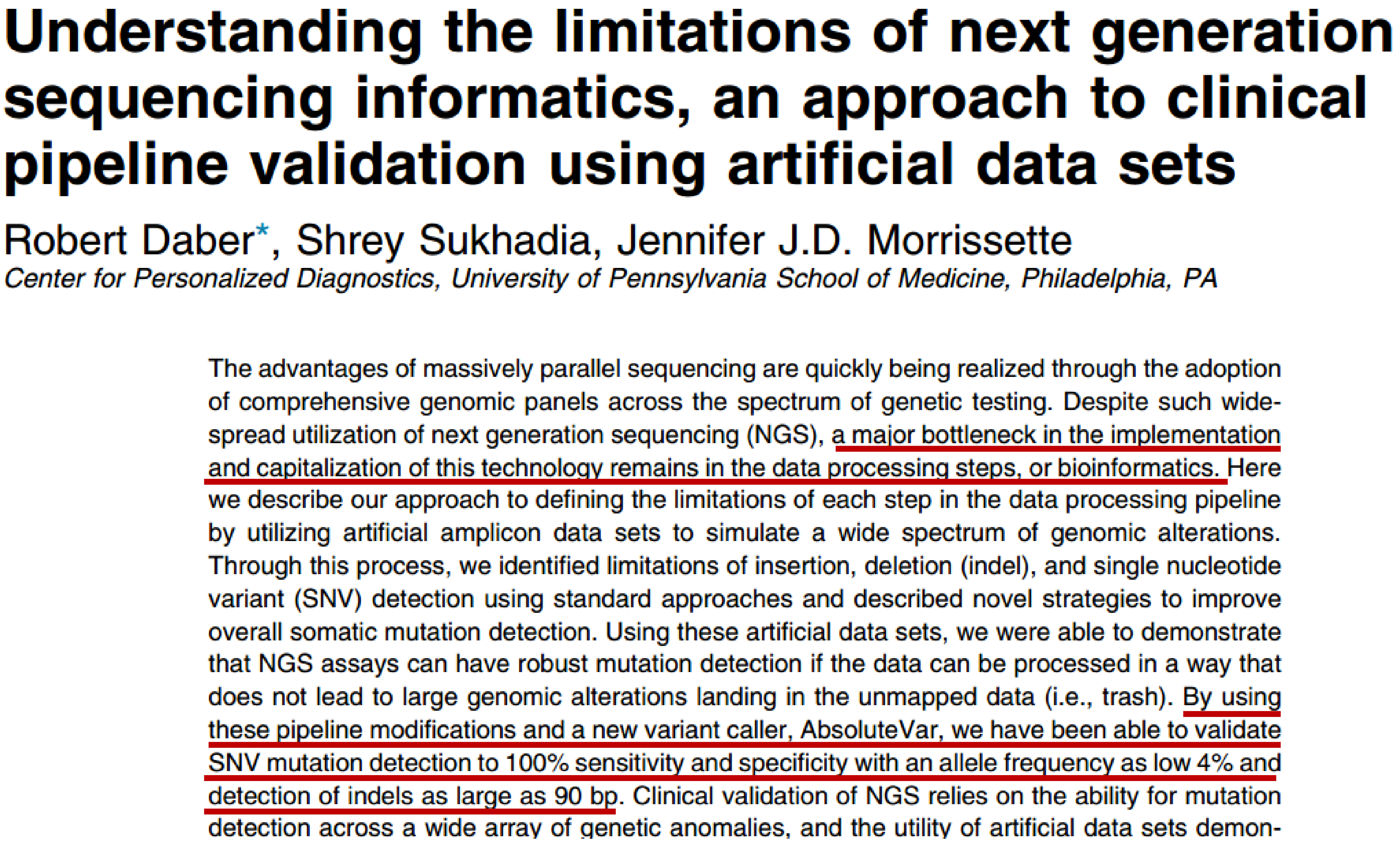
On average in the US, it takes
7.6 YEARS
to get an accurate diagnosis.



Hiseq X
Hiseq 4000
Hiseq 2000/2500
Hiseq2500 RR
NextSeq 500
Proton
MiSeq
MinION
SOLiD
MiniSeq
S5/S5XL
GA II
PGM
PacBio RS
GS Junior
GS FLX
'Sanger'
Gigabses per run (log scale)
1000
100
10
1
0.1
0.01
0.001
0.0001
0.00001
10
100
1000
10000
100000
Read length (log scale)

Схема процесса NGS анализа


1
Paraffin Block
Cellularity (60%)
Tumor area
FFPE Slides
Blood
Genomic DNA
DNA Isolation
2
Fragmented DNA
Indexed Libraries
T200 probes
Hybrid library-probes
Capture Amplified DNA
Library Prep and Capture
3
Sequencing
Data Analysis
Alignment
Mutations
CNVs
Sequencing and Data Analysis
4
Determine if alterations in actionable genes
Determine functional consequences of alterations
Prioritize mutations/targets
Identify potential targeted therapies
FDA-approved genomically-matched therapy in same tumor type?
NO
Consider clinical trials using genotype-relevant drugs
Determining Therapy



Understanding the limitations of next generation sequencing informatics, an approach to clinical pipeline validation using artificial data sets

Robert Daber*, Shrey Sukhadia, Jennifer J.D. Morrissette

Center for Personalized Diagnostics, University of Pennsylvania School of Medicine, Philadelphia, PA

The advantages of massively parallel sequencing are quickly being realized through the adoption of comprehensive genomic panels across the spectrum of genetic testing. Despite such widespread utilization of next generation sequencing (NGS), a major bottleneck in the implementation and capitalization of this technology remains in the data processing steps, or bioinformatics. Here we describe our approach to defining the limitations of each step in the data processing pipeline by utilizing artificial amplicon data sets to simulate a wide spectrum of genomic alterations. Through this process, we identified limitations of insertion, deletion (indel), and single nucleotide variant (SNV) detection using standard approaches and described novel strategies to improve overall somatic mutation detection. Using these artificial data sets, we were able to demonstrate that NGS assays can have robust mutation detection if the data can be processed in a way that does not lead to large genomic alterations landing in the unmapped data (i.e., trash). By using these pipeline modifications and a new variant caller, AbsoluteVar, we have been able to validate SNV mutation detection to 100% sensitivity and specificity with an allele frequency as low 4% and detection of indels as large as 90 bp. Clinical validation of NGS relies on the ability for mutation detection across a wide array of genetic anomalies, and the utility of artificial data sets demon-

Data
analysis
Shift
Del
Pr

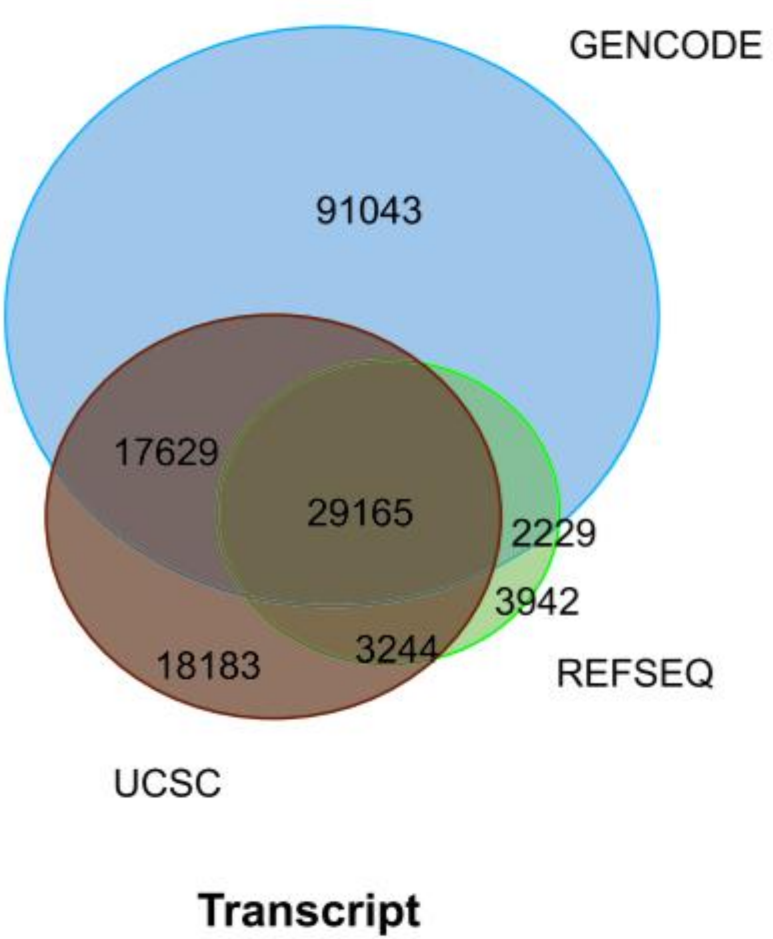
Massive amount of sequence data

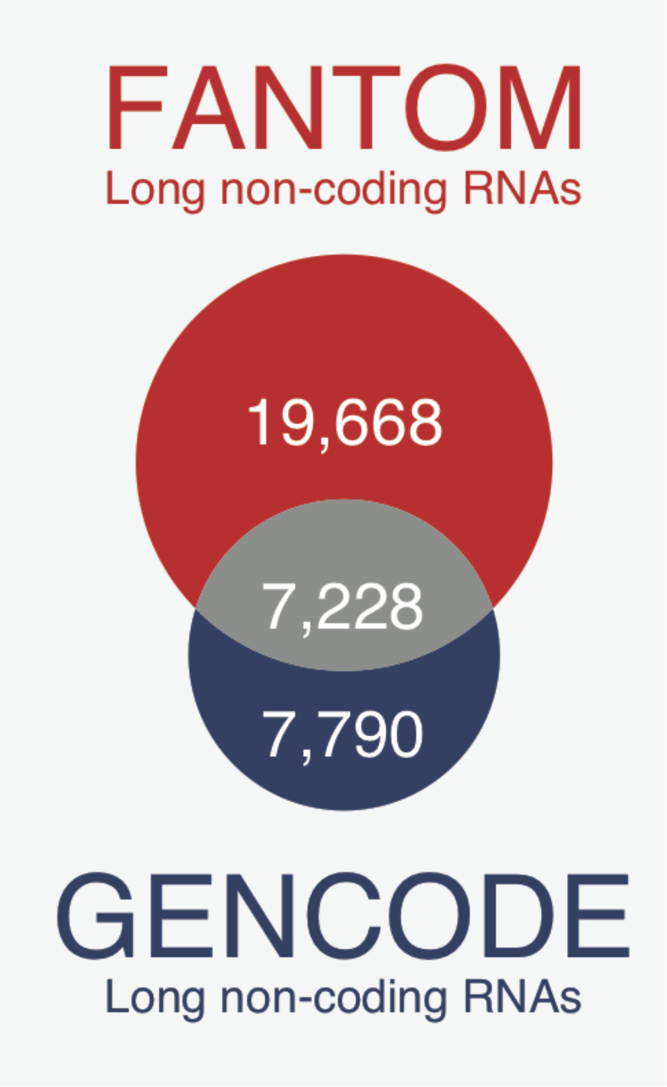
NGS machnies

Программы и алгоритмы для аннотации генов

	RefSeq	Ensembl	Gencode (v. 29)
Total No of Genes	**6118**	**57365**	**58721**
Protein-coding genes	20216	20418	19940
Long non-coding RNA genes	18533	15014	16066
Small non-coding RNA genes		4871	7577
Pseudogenes	16435	15195	14505

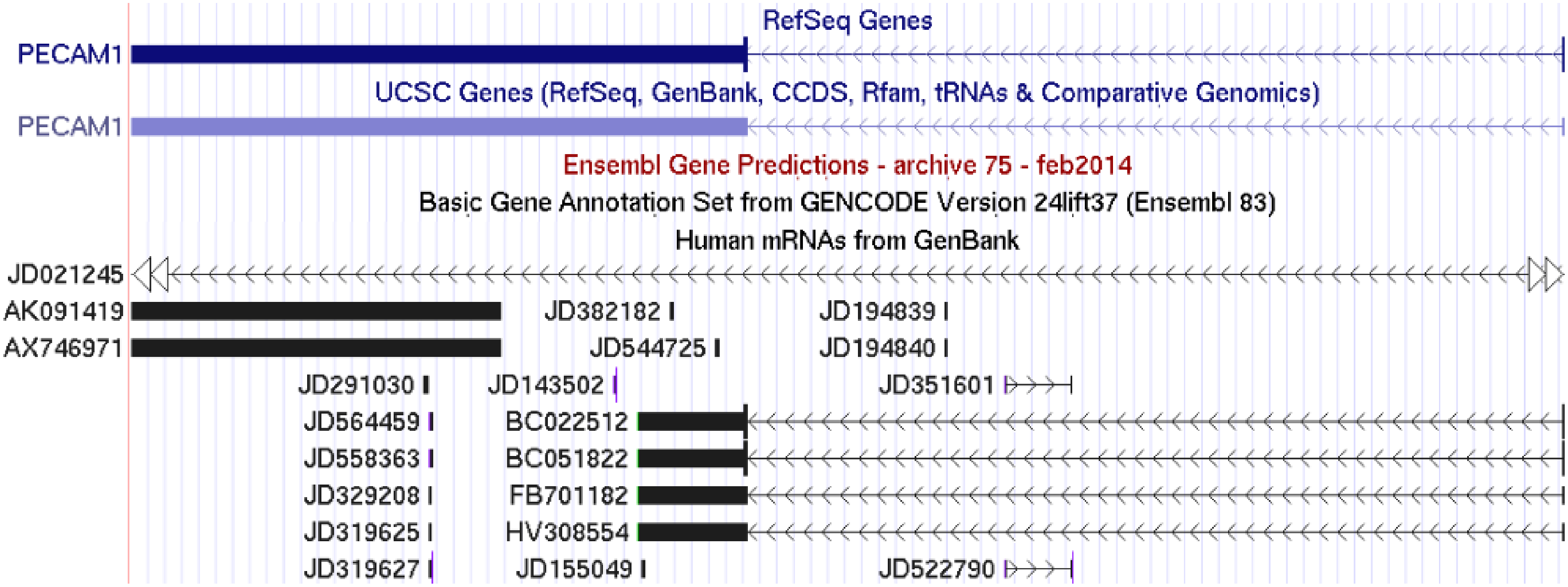
Сравнение разных аннотаций генов



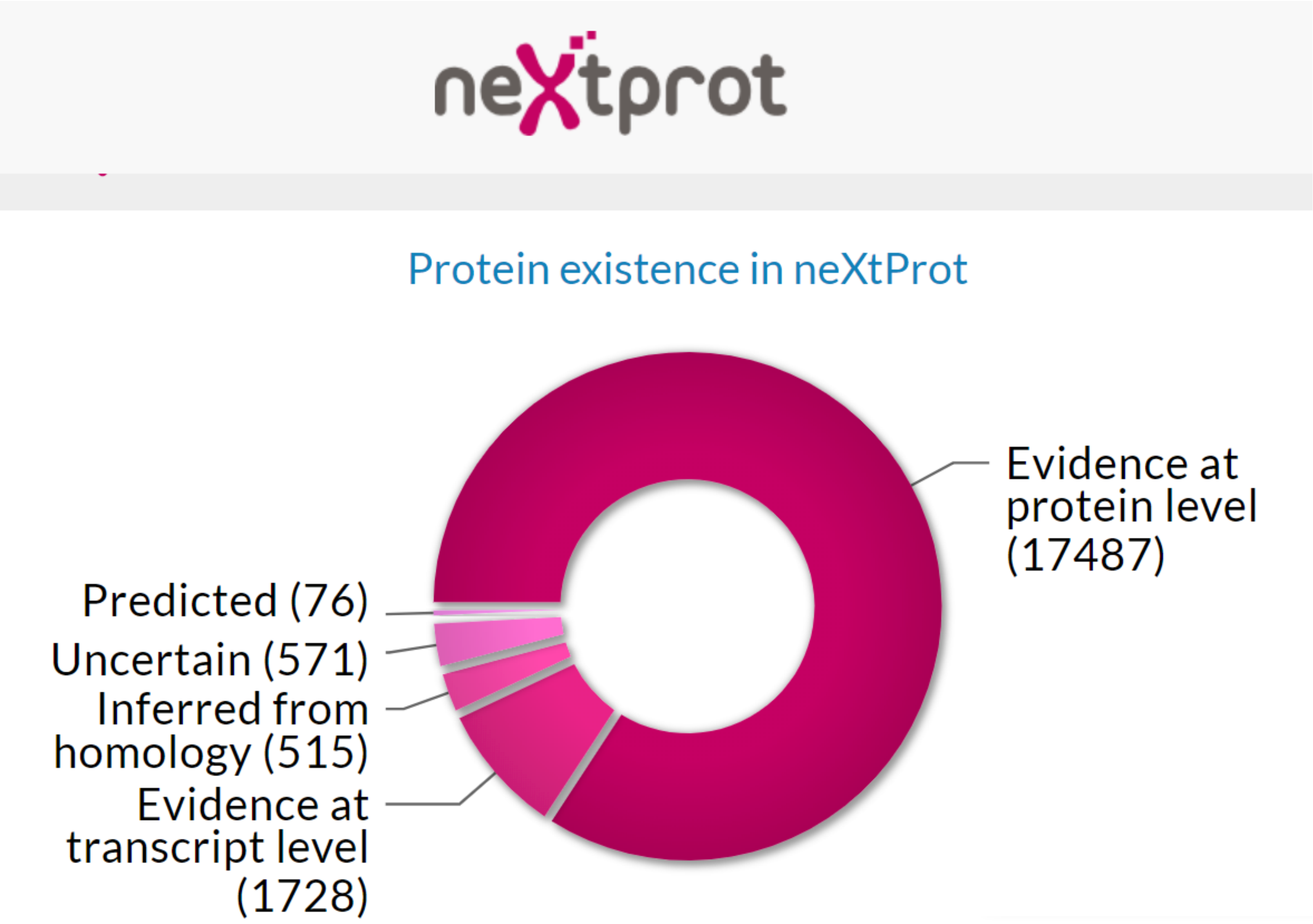


doi: 10.1101/gr.135350.111

Пример разницы аннотаций генов

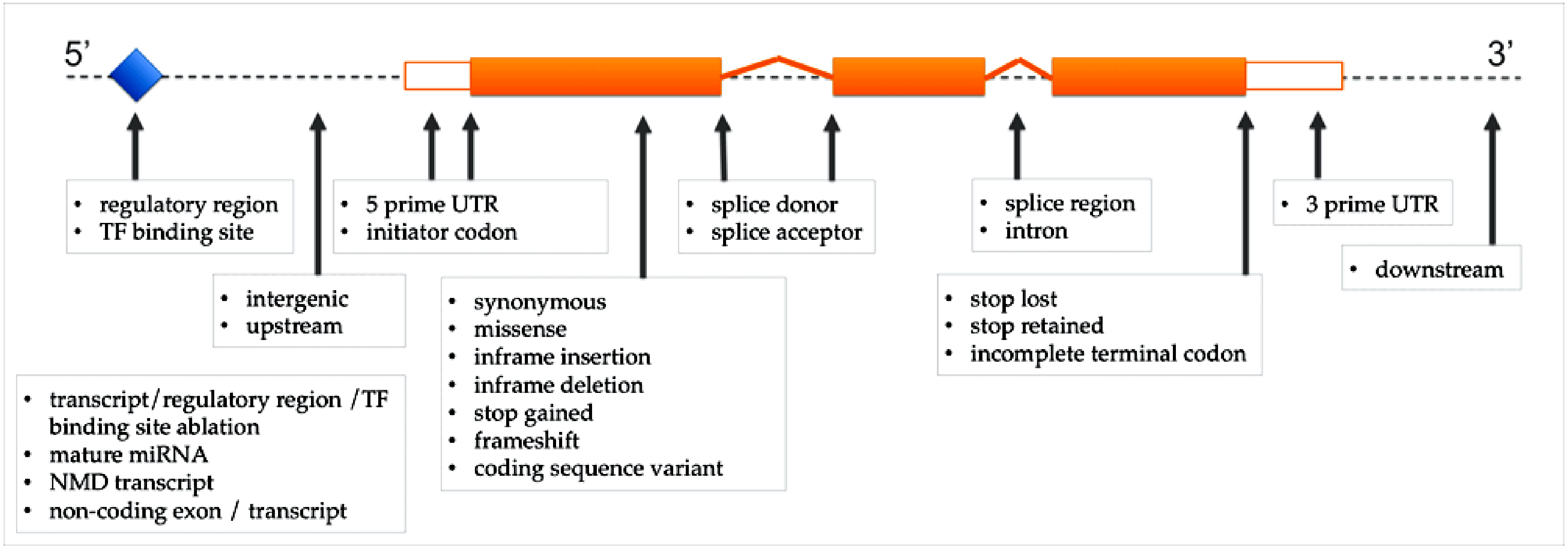


Реализация кодирующего потенциала РНК



Protein existence in neXtProt

Где могут быть патогенные варианты?



- regulatory region
- TF binding site

- intergenic
- upstream

- transcript/regulatory region /TF binding site ablation
- mature miRNA
- NMD transcript
- non-coding exon / transcript

- 5 prime UTR
- initiator codon

- synonymous
- missense
- inframe insertion
- inframe deletion
- stop gained
- frameshift
- coding sequence variant

- splice donor
- splice acceptor

- splice region
- intron

- stop lost
- stop retained
- incomplete terminal codon

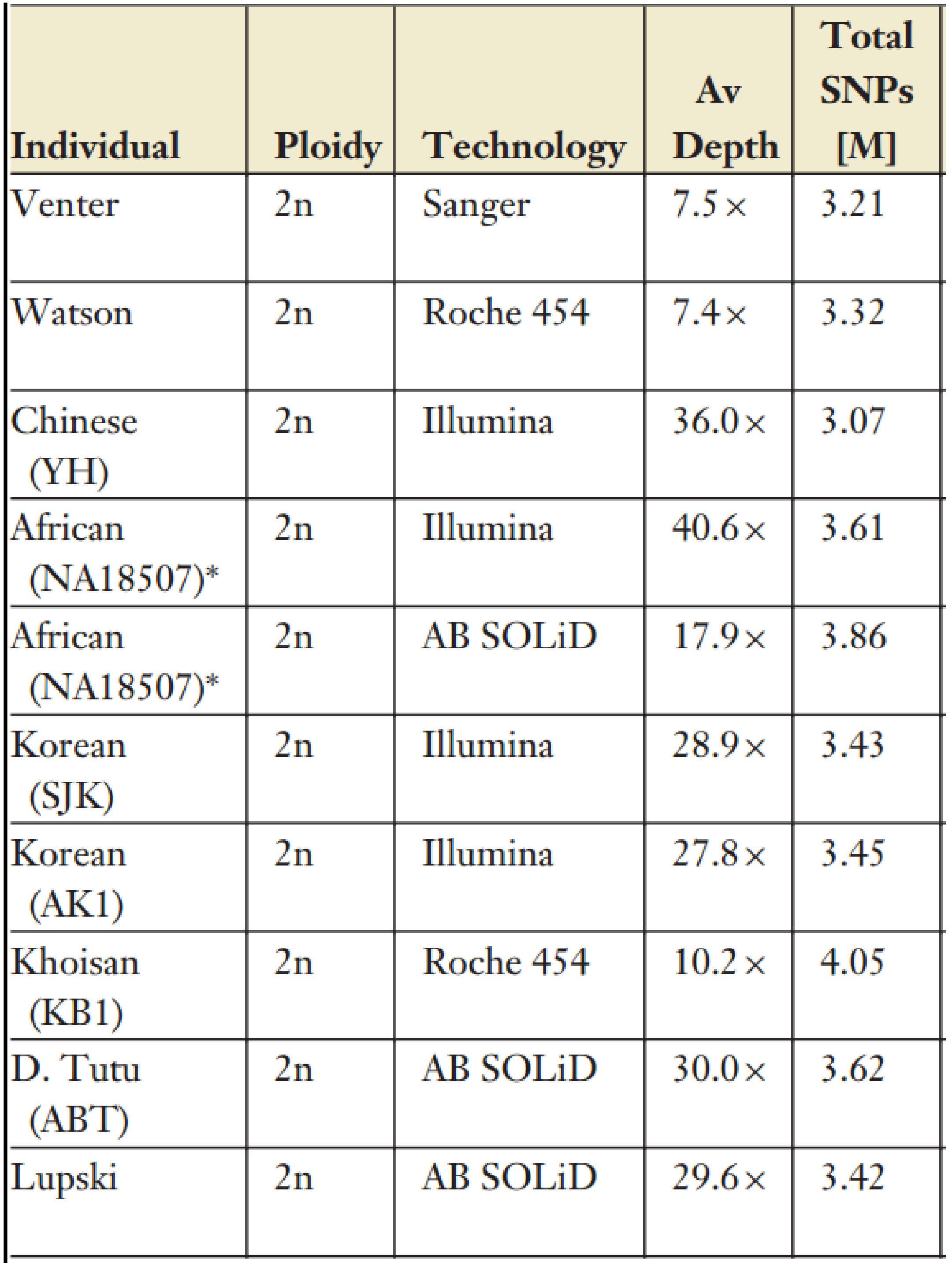
- 3 prime UTR

- downstream

dante labs

Черная пятница: 178 евро

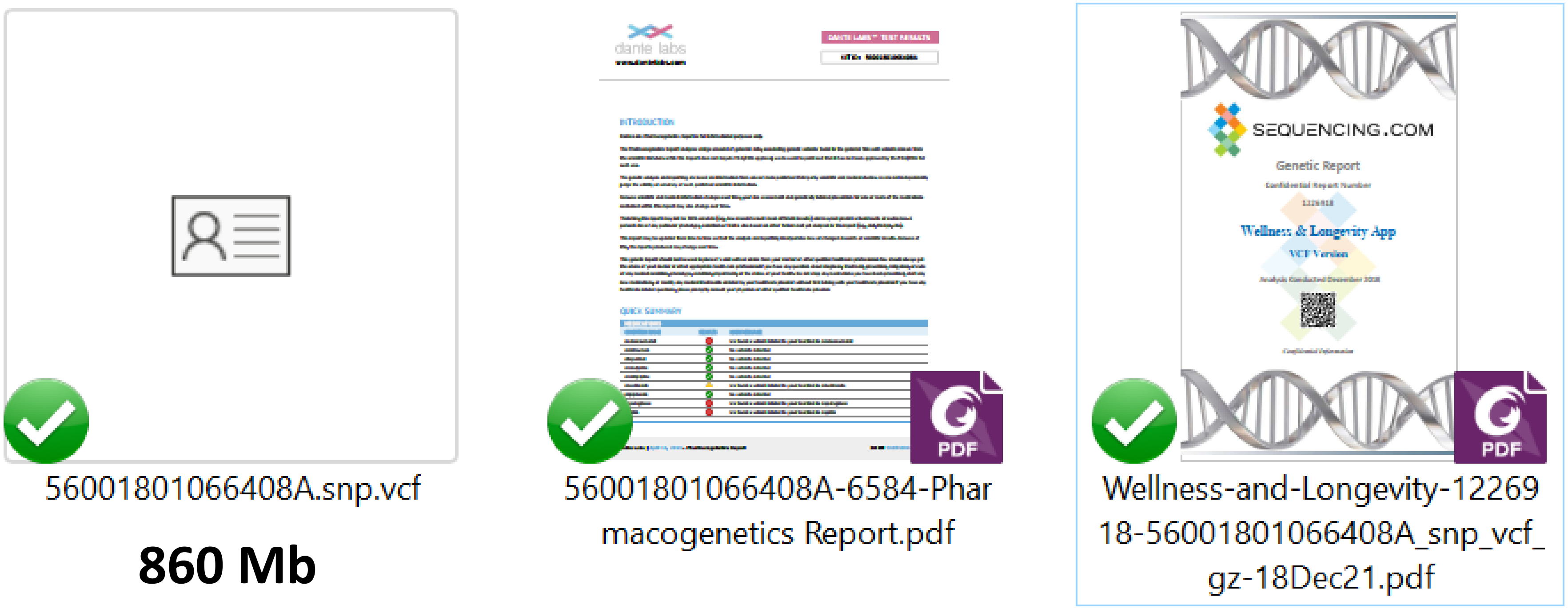
Сравнение результатов секвенирования индивидуальных геномов человека



Individual	Ploidy	Technology	Av Depth	Total SNPs [M]
Venter	2n	Sanger	7.5×	3.21
Watson	2n	Roche 454	7.4×	3.32
Chinese (YH)	2n	Illumina	36.0×	3.07
African (NA18507)*	2n	Illumina	40.6×	3.61
African (NA18507)*	2n	AB SOLiD	17.9×	3.86
Korean (SJK)	2n	Illumina	28.9×	3.43
Korean (AK1)	2n	Illumina	27.8×	3.45
Khoisan (KB1)	2n	Roche 454	10.2×	4.05
D. Tutu (ABT)	2n	AB SOLiD	30.0×	3.62
Lupski	2n	AB SOLiD	29.6×	3.42

Gonzaga-Jauregui C, Lupski JR, Gibbs RA.
Human genome sequencing in health and disease.
Annu Rev Med. 2012;63:35-61.

3’415’465 персональных вариаций



56001801066408A.snp.vcf

56001801066408A-6584-Pharmacogenetics Report.pdf

Wellness-and-Longevity-1226918-56001801066408A_snp_vcf_gz-18Dec21.pdf

860 Mb

Genetic Report

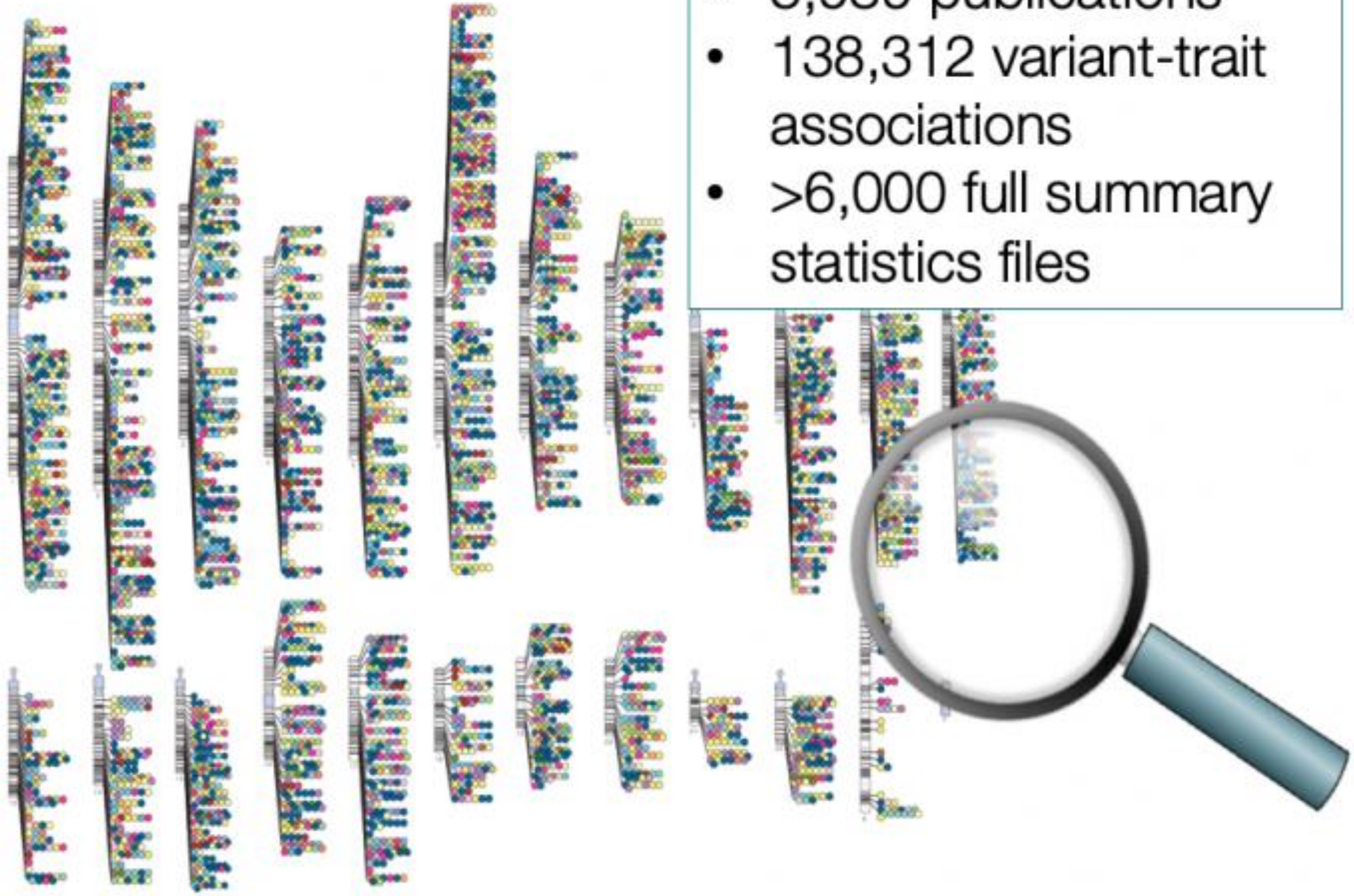
Confidential Report Number

1226918

Wellness & Longevity App

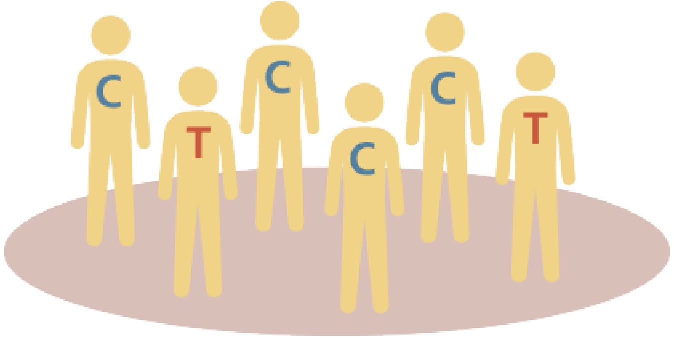
- The Report analyzes a large amount of genomic data, associating genetic variants found in the **genomic files** with variants known from the scientific literature. While this Report does not require FDA/EMA approval, we do want to point out that it has not been approved by the FDA/EMA for such use.
- We do not independently judge the validity or accuracy of such published scientific information.
- Because scientific and medical information changes over time, your risk assessment and genetically tailored prevention for one or more of the medications contained within this report may also change over time.
- Therefore, this report may not be 100% accurate (e.g., new research could mean different results) and may not predict actual results or outcomes.

GWAS Catalog

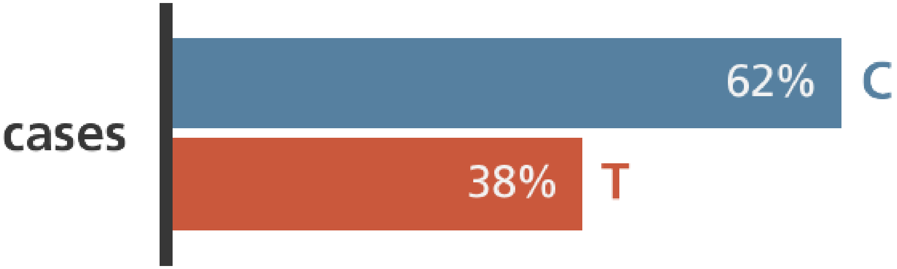


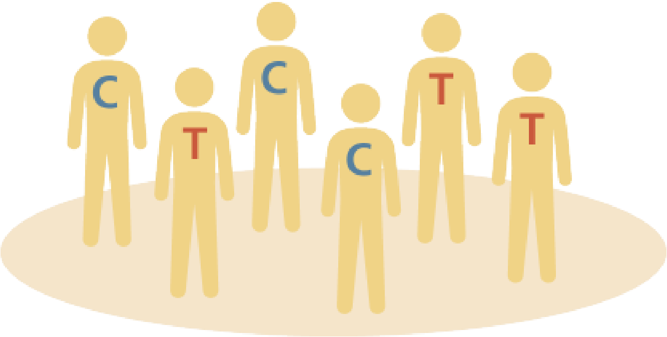
As of May 2019

- 3,989 publications
- 138,312 variant-trait associations
- >6,000 full summary statistics files

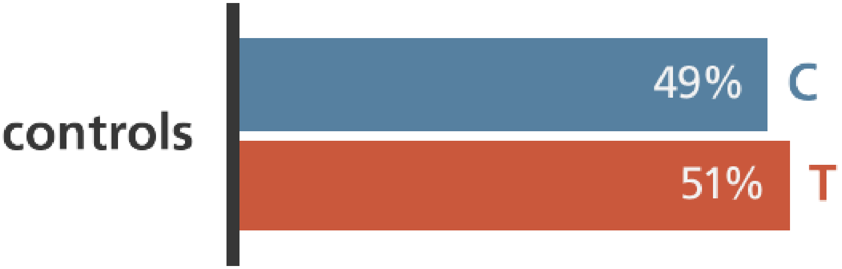


cases (n=1,000)
people with heart disease

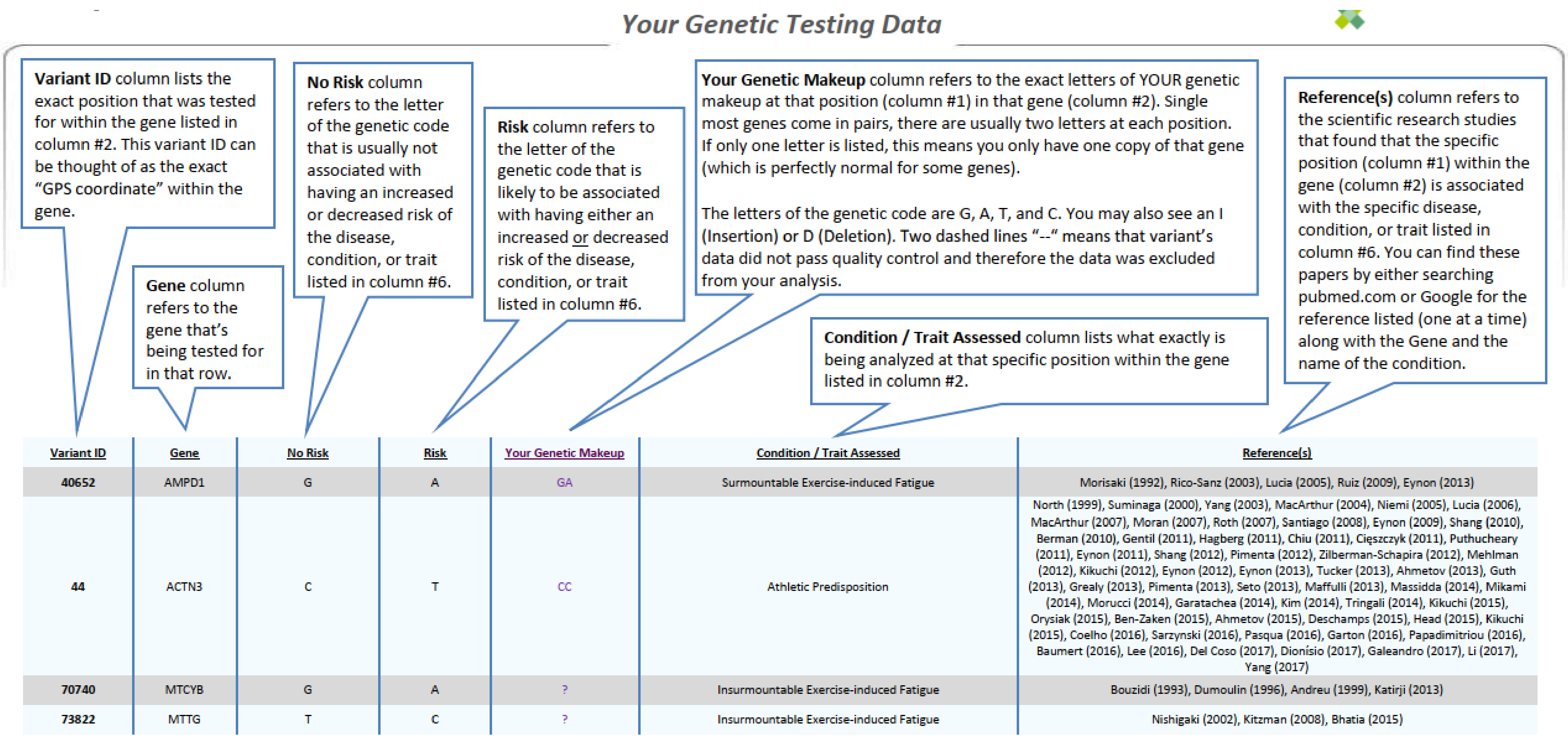




controls (n=1,000)
people without heart disease



«Genetic Report» на 163 страницах



Your Genetic Testing Data

Variant ID column lists the exact position that was tested for within the gene listed in column #2. This variant ID can be thought of as the exact "GPS coordinate" within the gene.

Gene column refers to the gene that's being tested for in that row.

No Risk column refers to the letter of the genetic code that is usually not associated with having an increased or decreased risk of the disease, condition, or trait listed in column #6.

Risk column refers to the letter of the genetic code that is likely to be associated with having either an increased or decreased risk of the disease, condition, or trait listed in column #6.

Your Genetic Makeup column refers to the exact letters of YOUR genetic makeup at that position (column #1) in that gene (column #2). Single most genes come in pairs, there are usually two letters at each position. If only one letter is listed, this means you only have one copy of that gene (which is perfectly normal for some genes).

The letters of the genetic code are G, A, T, and C. You may also see an I (Insertion) or D (Deletion). Two dashed lines "--" means that variant's data did not pass quality control and therefore the data was excluded from your analysis.

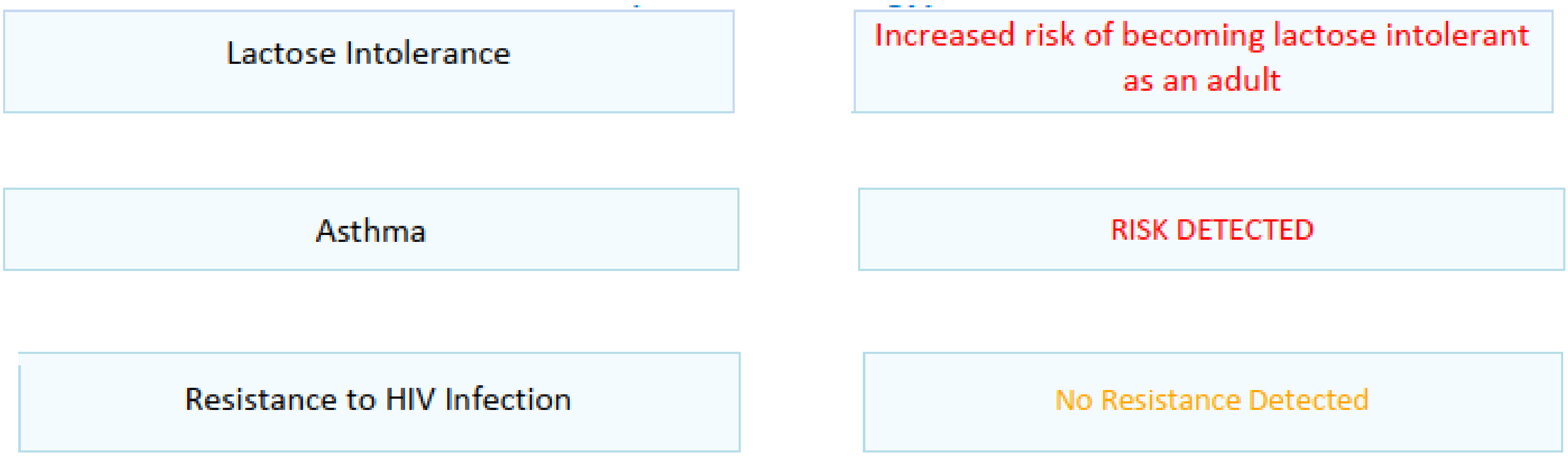
Condition / Trait Assessed column lists what exactly is being analyzed at that specific position within the gene listed in column #2.

Reference(s) column refers to the scientific research studies that found that the specific position (column #1) within the gene (column #2) is associated with the specific disease, condition, or trait listed in column #6. You can find these papers by either searching pubmed.com or Google for the reference listed (one at a time) along with the Gene and the name of the condition.

Variant ID	Gene	No Risk	Risk	Your Genetic Makeup	Condition / Trait Assessed	Reference(s)
40652	AMPD1	G	A	GA	Surmountable Exercise-induced Fatigue	Morisaki (1992), Rico-Sanz (2003), Lucia (2005), Ruiz (2009), Eynon (2013)
44	ACTN3	C	T	CC	Athletic Predisposition	North (1999), Suminaga (2000), Yang (2003), MacArthur (2004), Niemi (2005), Lucia (2006), MacArthur (2007), Moran (2007), Roth (2007), Santiago (2008), Eynon (2009), Shang (2010), Berman (2010), Gentil (2011), Hagberg (2011), Chiu (2011), Cięszczyk (2011), Puthucheary (2011), Eynon (2011), Shang (2012), Pimenta (2012), Zilberman-Schapira (2012), Mehlman (2012), Kikuchi (2012), Eynon (2012), Eynon (2013), Tucker (2013), Ahmetov (2013), Guth (2013), Grealy (2013), Pimenta (2013), Seto (2013), Maffulli (2013), Massidda (2014), Mikami (2014), Morucci (2014), Garatachea (2014), Kim (2014), Tringali (2014), Kikuchi (2015), Orysiak (2015), Ben-Zaken (2015), Ahmetov (2015), Deschamps (2015), Head (2015), Kikuchi (2015), Coelho (2016), Sarzynski (2016), Pasqua (2016), Garton (2016), Papadimitriou (2016), Baumert (2016), Lee (2016), Del Coso (2017), Dionísio (2017), Galeandro (2017), Li (2017), Yang (2017)
70740	MTCYB	G	A	?	Insurmountable Exercise-induced Fatigue	Bouzidi (1993), Dumoulin (1996), Andreu (1999), Katirji (2013)
73822	MTTG	T	C	?	Insurmountable Exercise-induced Fatigue	Nishigaki (2002), Kitzman (2008), Bhatia (2015)

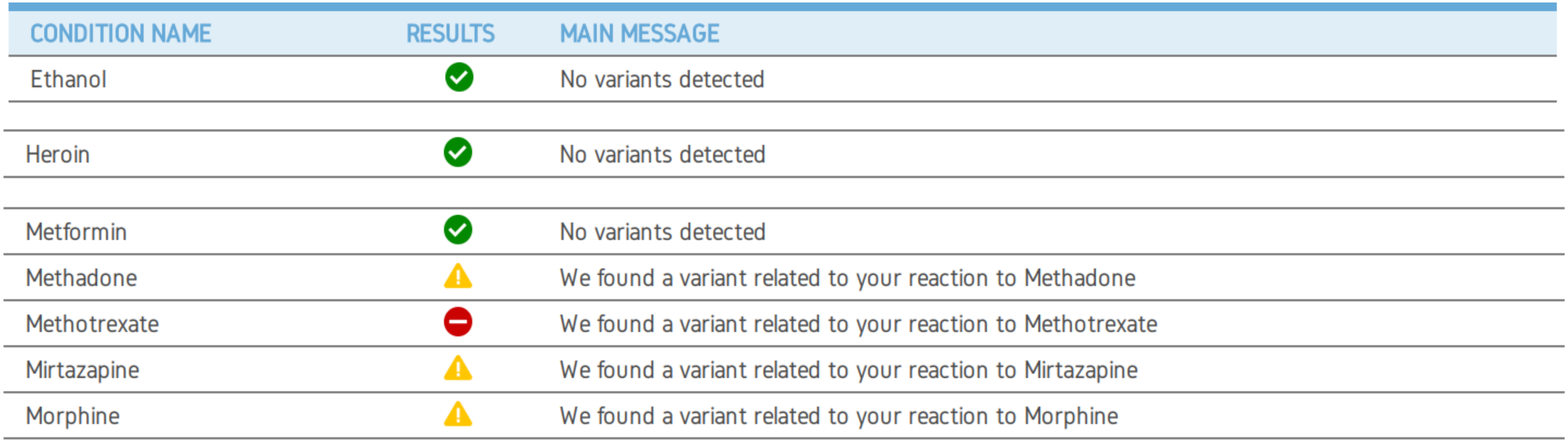
- 3'450 описанных вариаций

«Genetic Report» на 163 страницах



Lactose Intolerance	Increased risk of becoming lactose intolerant as an adult
Asthma	RISK DETECTED
Resistance to HIV Infection	No Resistance Detected

«Genetic Report» на 163 страницах



CONDITION NAME	RESULTS	MAIN MESSAGE
Ethanol	✓	No variants detected
Heroin	✓	No variants detected
Metformin	✓	No variants detected
Methadone	⚠	We found a variant related to your reaction to Methadone
Methotrexate	⛔	We found a variant related to your reaction to Methotrexate
Mirtazapine	⚠	We found a variant related to your reaction to Mirtazapine
Morphine	⚠	We found a variant related to your reaction to Morphine

Результаты данного исследования могут быть правильно интерпретированы только врачом-генетиком.