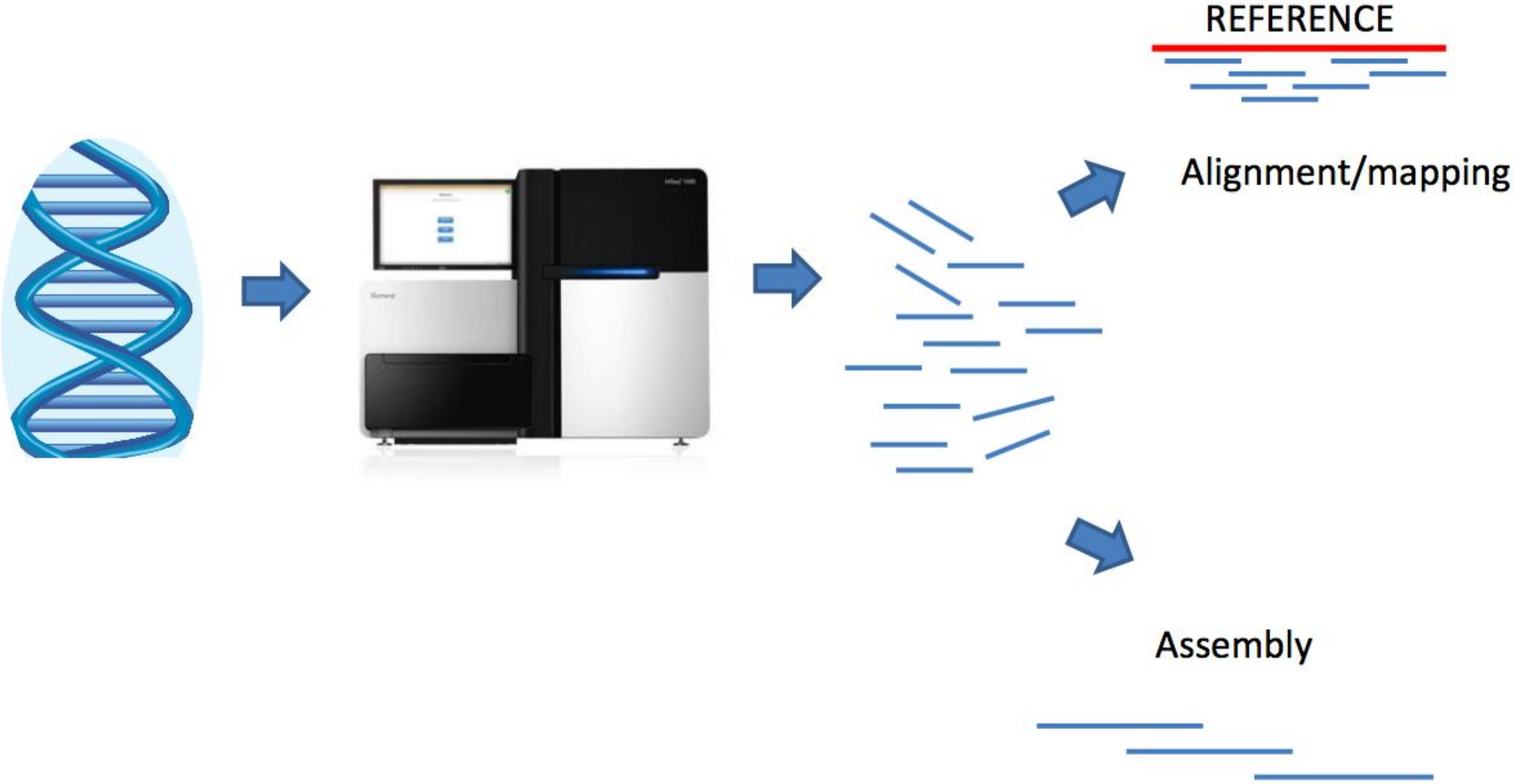
BINOM
Bioinformatics accelerated

Недостатки, ограничения и ошибки NGS

Андрей Афанасьев

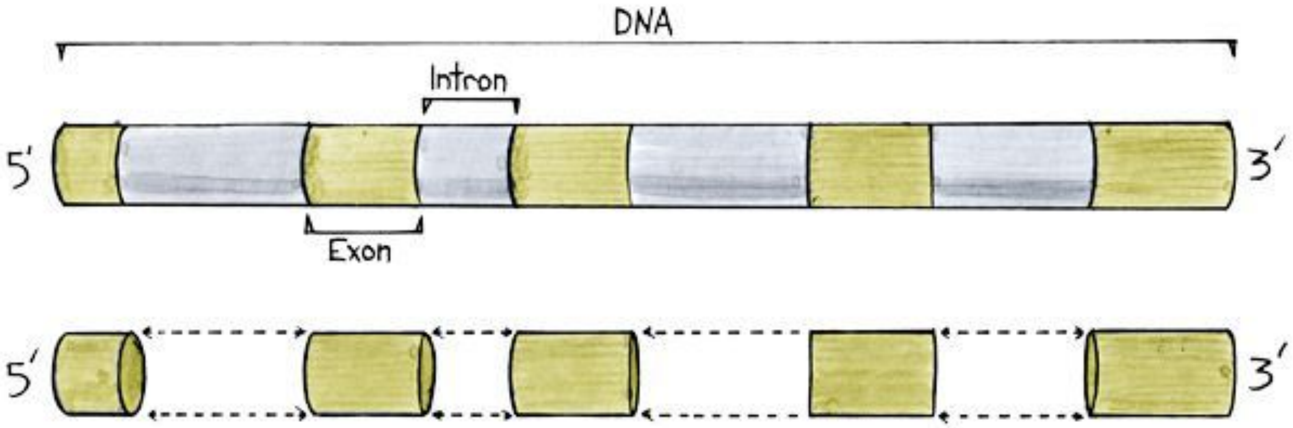


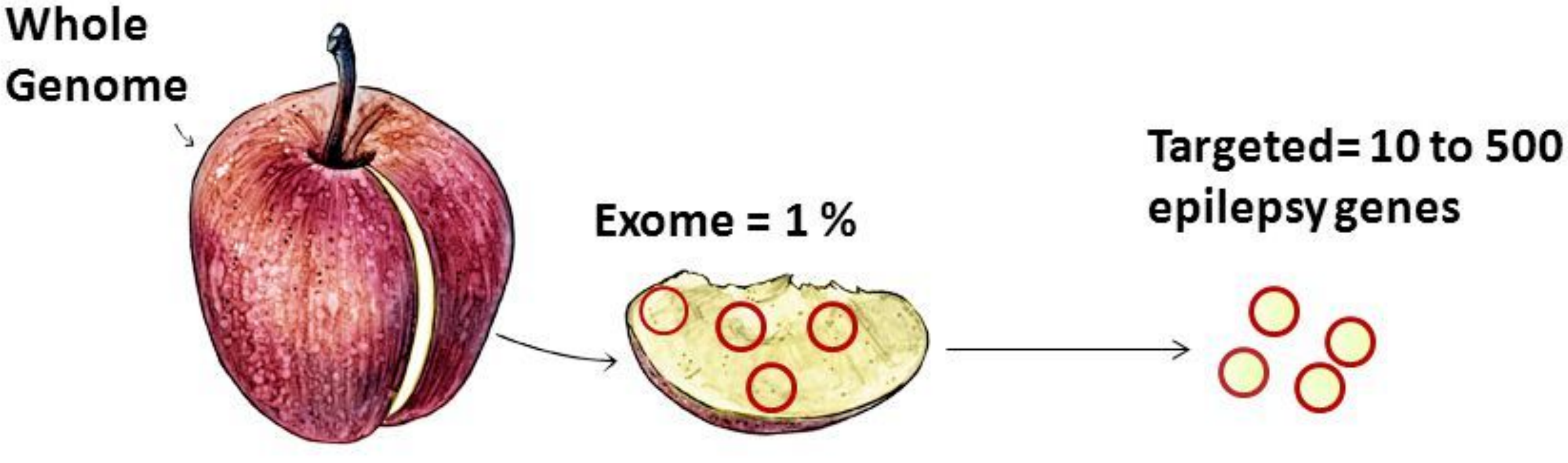
REFERENCE

Alignment/mapping

Assembly

Неисчерпывающие списки генов в наборах обогащения.



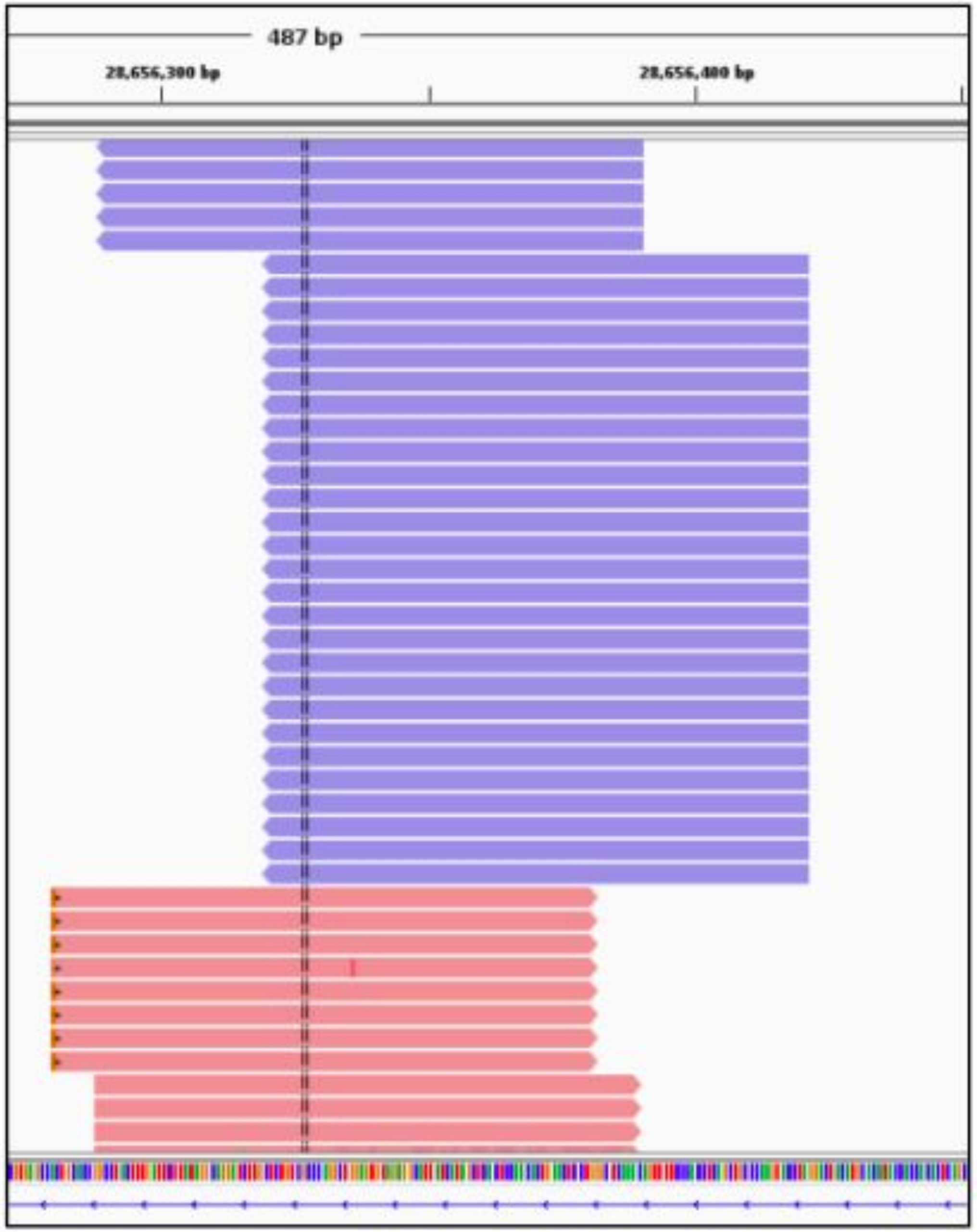


Пробоподготовка

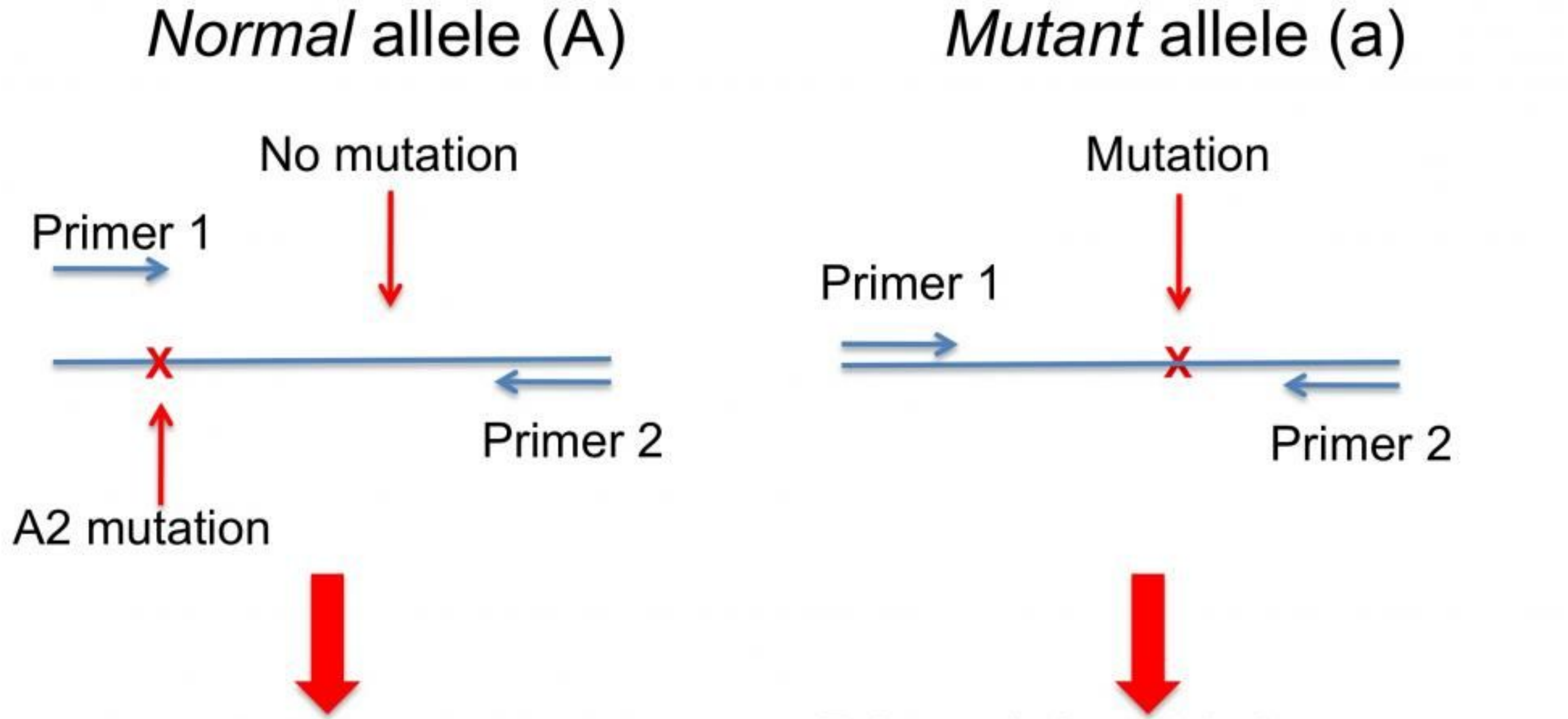
миллион способов накосячить

Дубликаты

- ПЦР-дубликаты из пробоподготовки
- Оптические дубликаты: дважды прочитанный секвенатором кластер



Потеря аллеля



Normal allele (A)

No mutation

Primer 1

Primer 2

A2 mutation

Mutant allele (a)

Mutation

Primer 1

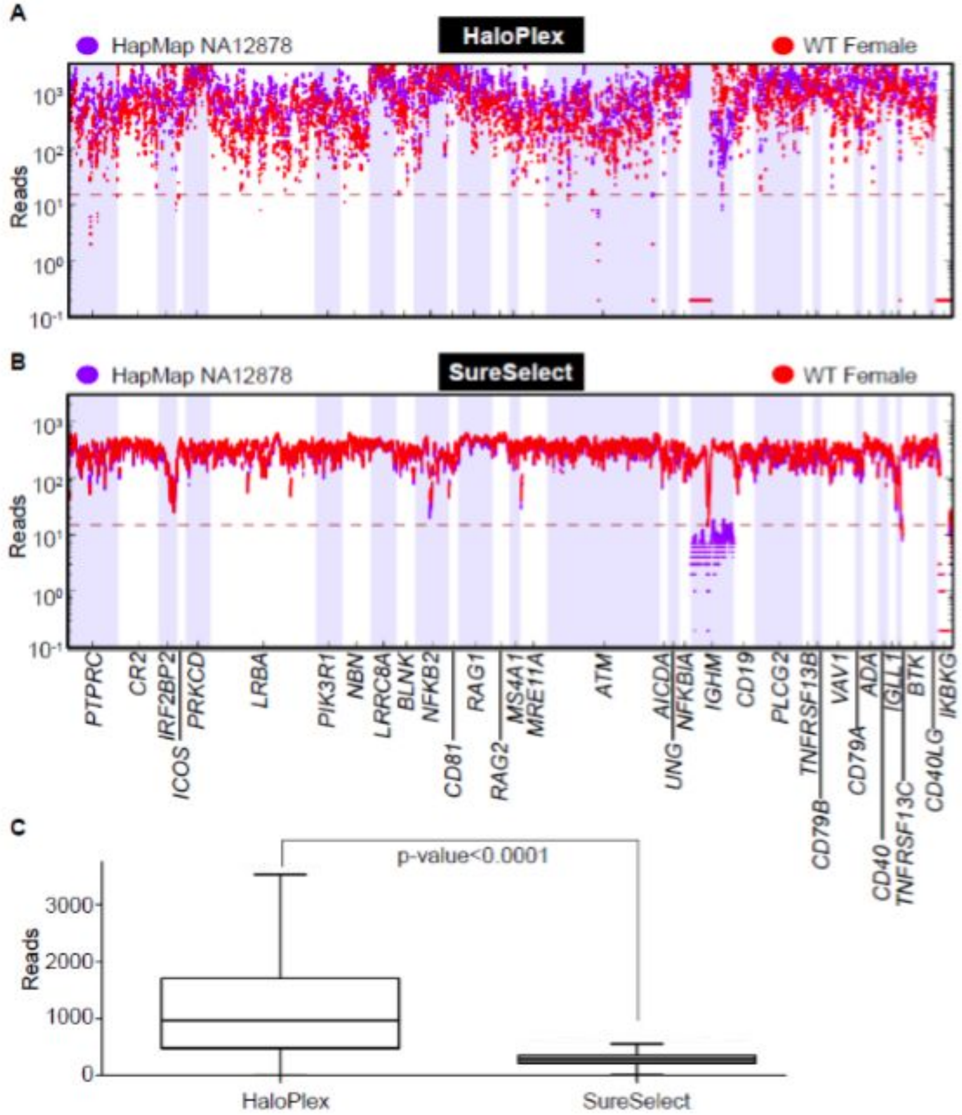
Primer 2

Primer 1 does not bind,
PCR **fails** and *normal*
allele is **not** detected

Primer 1 does bind,
PCR works and *mutant*
allele **is** detected

Дыры в покрытии

15 Read Cutoff



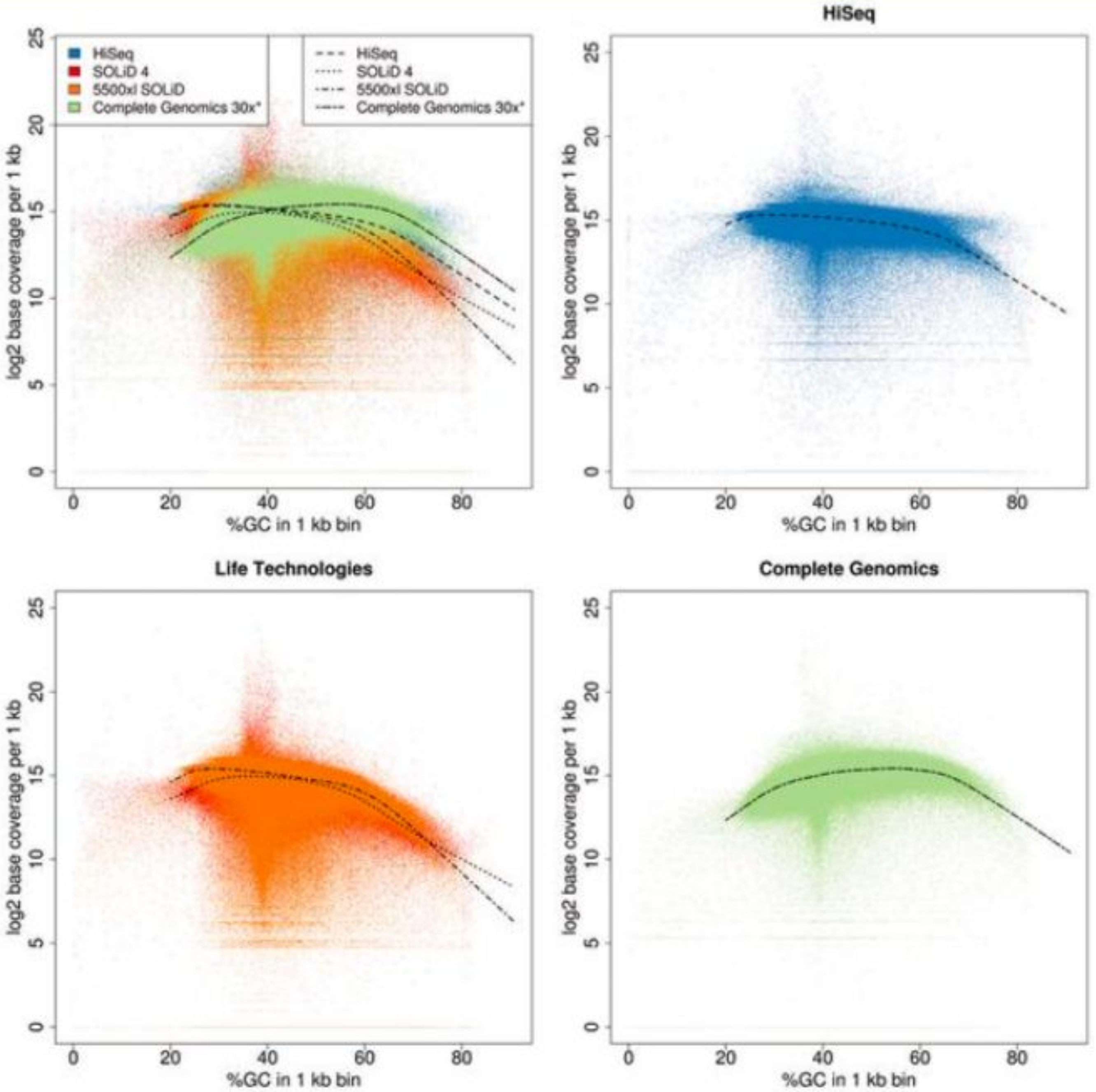
15 Read Cutoff

Haloplex Libraries Show Greater Read Number Variation

Lower Read Number Variation in SureSelect Libraries Allows Greater Library Pooling

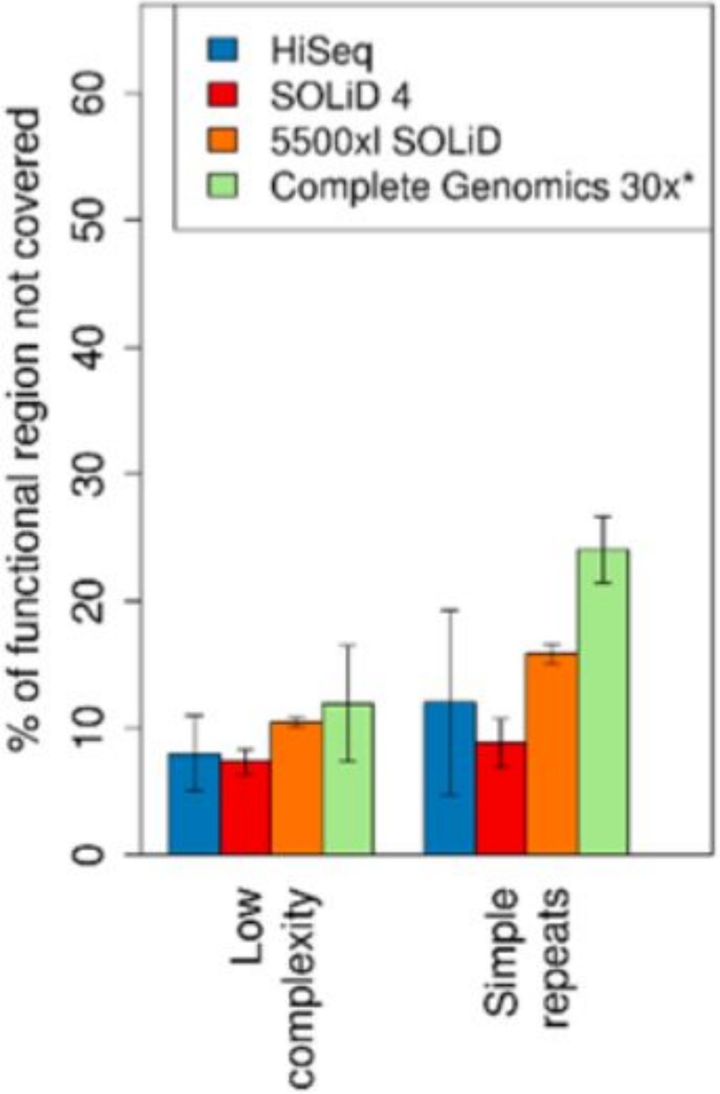
Беспощадная химия

GC-контент



Rieber N, et al. (2013) PLoS ONE 8(6): e66621.

Повторы



Rieber N, et al. (2013) PLoS ONE 8(6): e66621.

Приборные ошибки

Приборные ошибки

	Процент ошибок (%)
454	~1
Illumina	~0,1
Ion torrent	~1
PACBIO RSII	~10

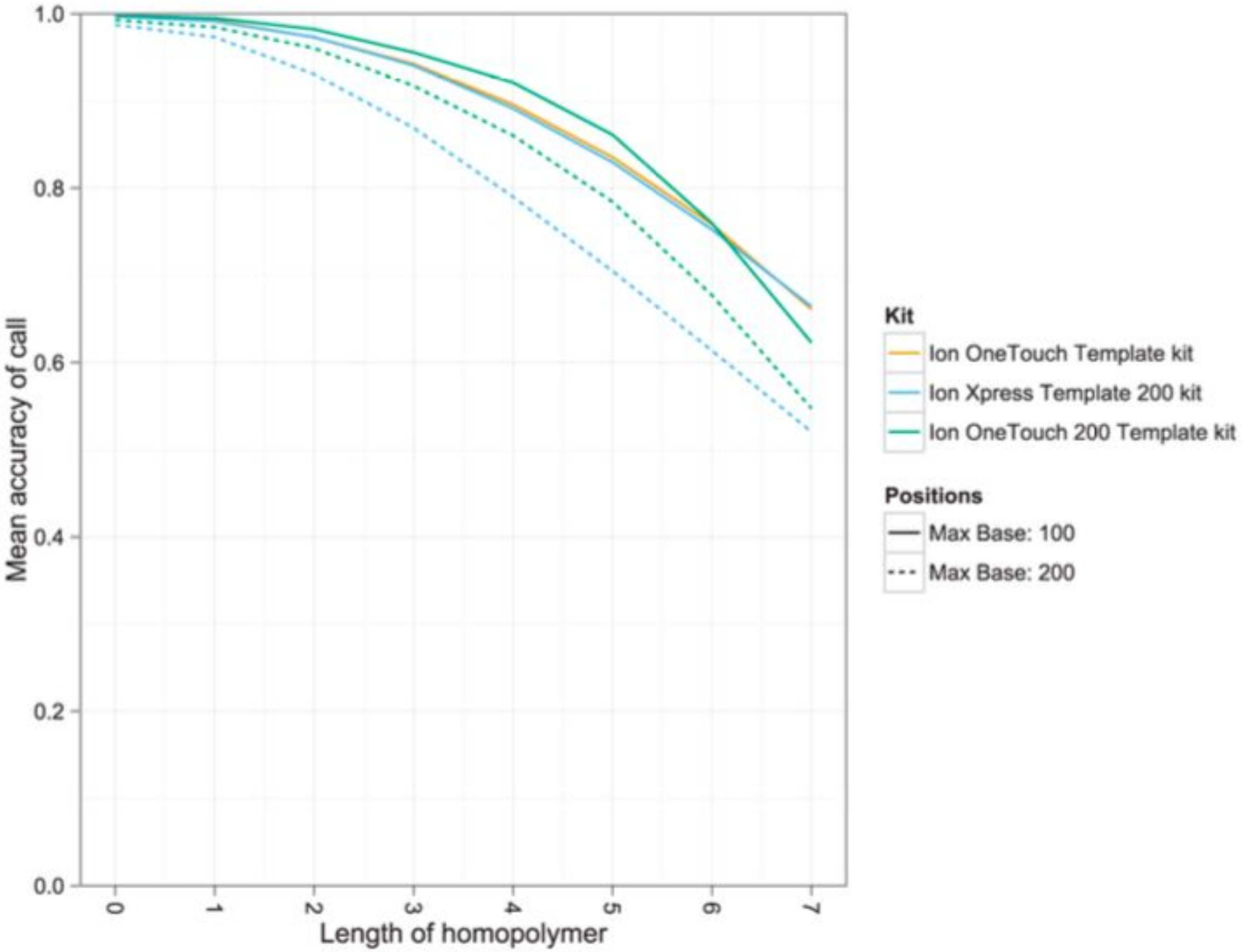
Ошибки на гомополимерах

AAAAAAAA

TTTTTTTT

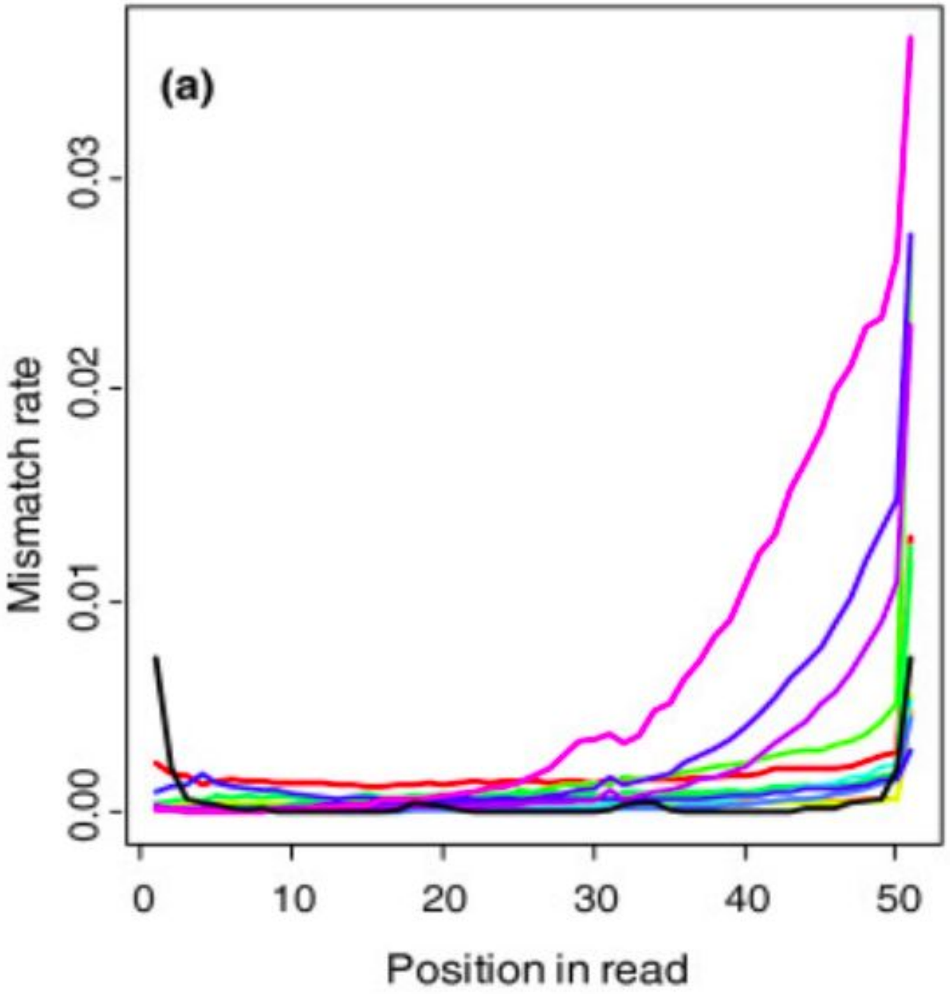
CCCCCCCC

GGGGGGG



Bragg et al PLoS Comput Biol. 2013 Apr;9(4):e1003031.

Падение качества прочтения к концу рида



- Kircher et al. Genome Biology 2009 10:R83

Все ошибаются по-своему

Ion Torrent	Гомополимеры
PacBio	Высокий процент ошибок (зато случайный)
Illumina	Зависимость от GC-контента
Complete Genomics	Неравномерное покрытие

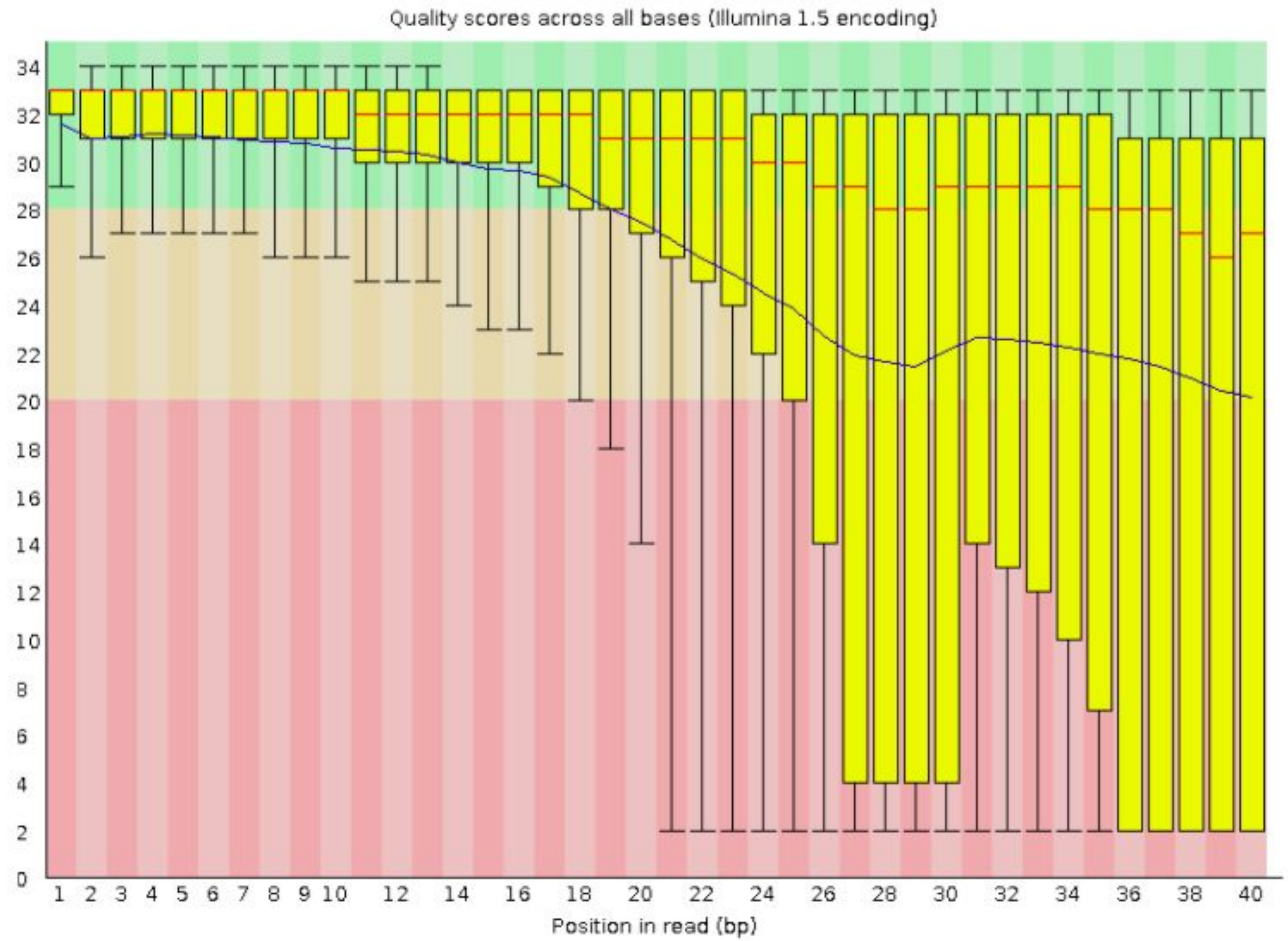
Контроль качества прочтений

Base qualities

Good

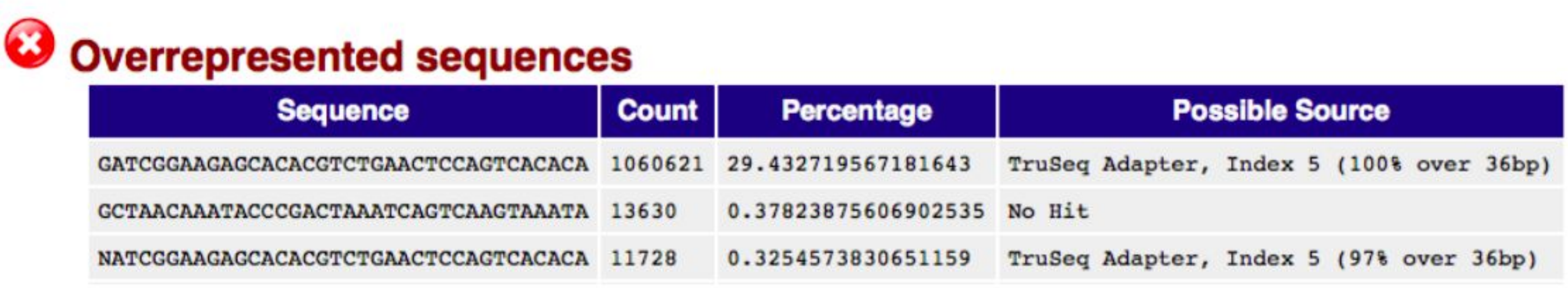
Okay

Bad



http://www.bioinformatics.babraham.ac.uk/projects/fastqc/bad_sequence_fastqc.html

Загрязнение адаптерами

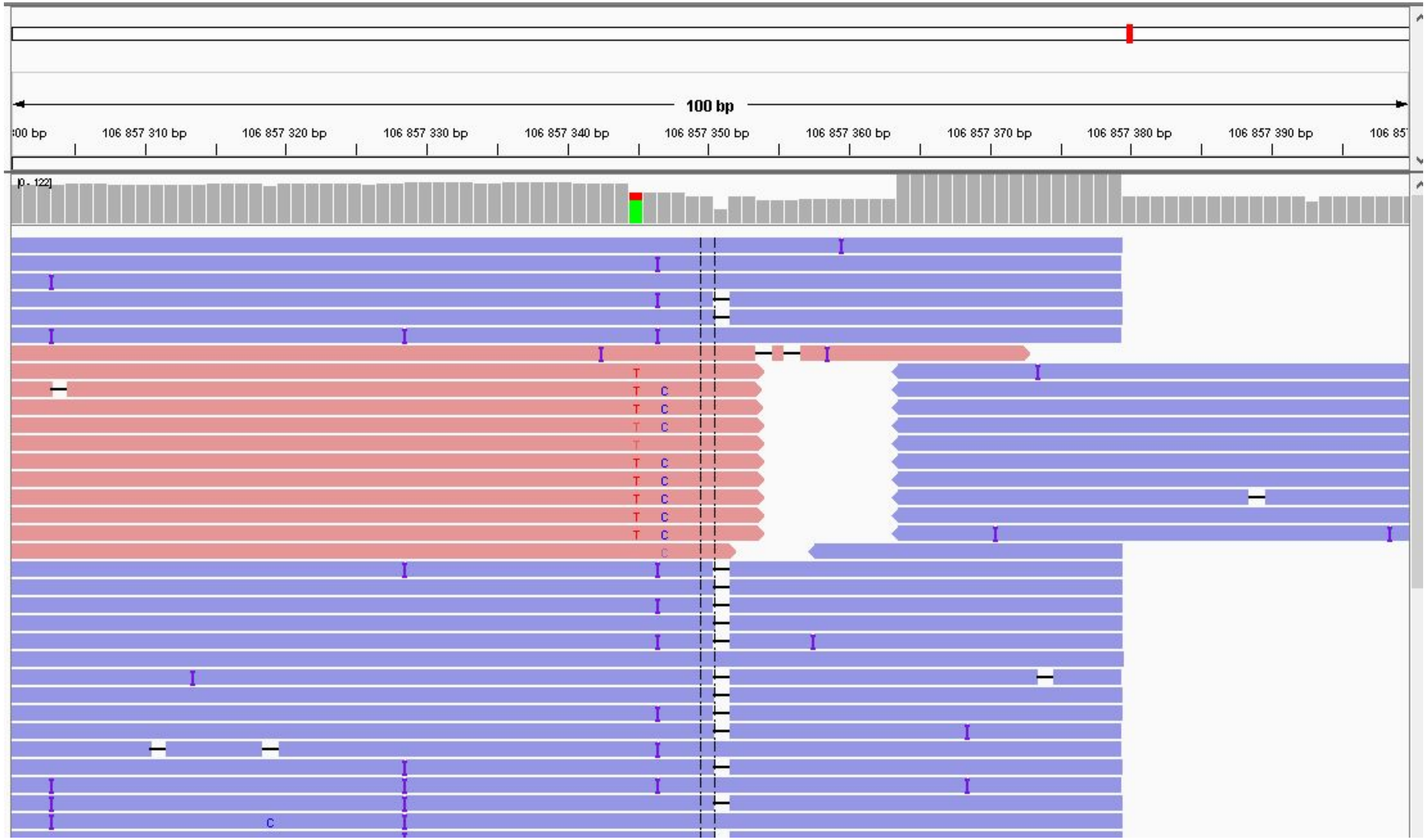


Overrepresented sequences

Sequence	Count	Percentage	Possible Source
GATCGGAAGAGCACACGTCTGAACTCCAGTCACACA	1060621	29.432719567181643	TruSeq Adapter, Index 5 (100% over 36bp)
GCTAACAAATACCCGACTAAATCAGTCAAGTAAATA	13630	0.37823875606902535	No Hit
NATCGGAAGAGCACACGTCTGAACTCCAGTCACACA	11728	0.3254573830651159	TruSeq Adapter, Index 5 (97% over 36bp)

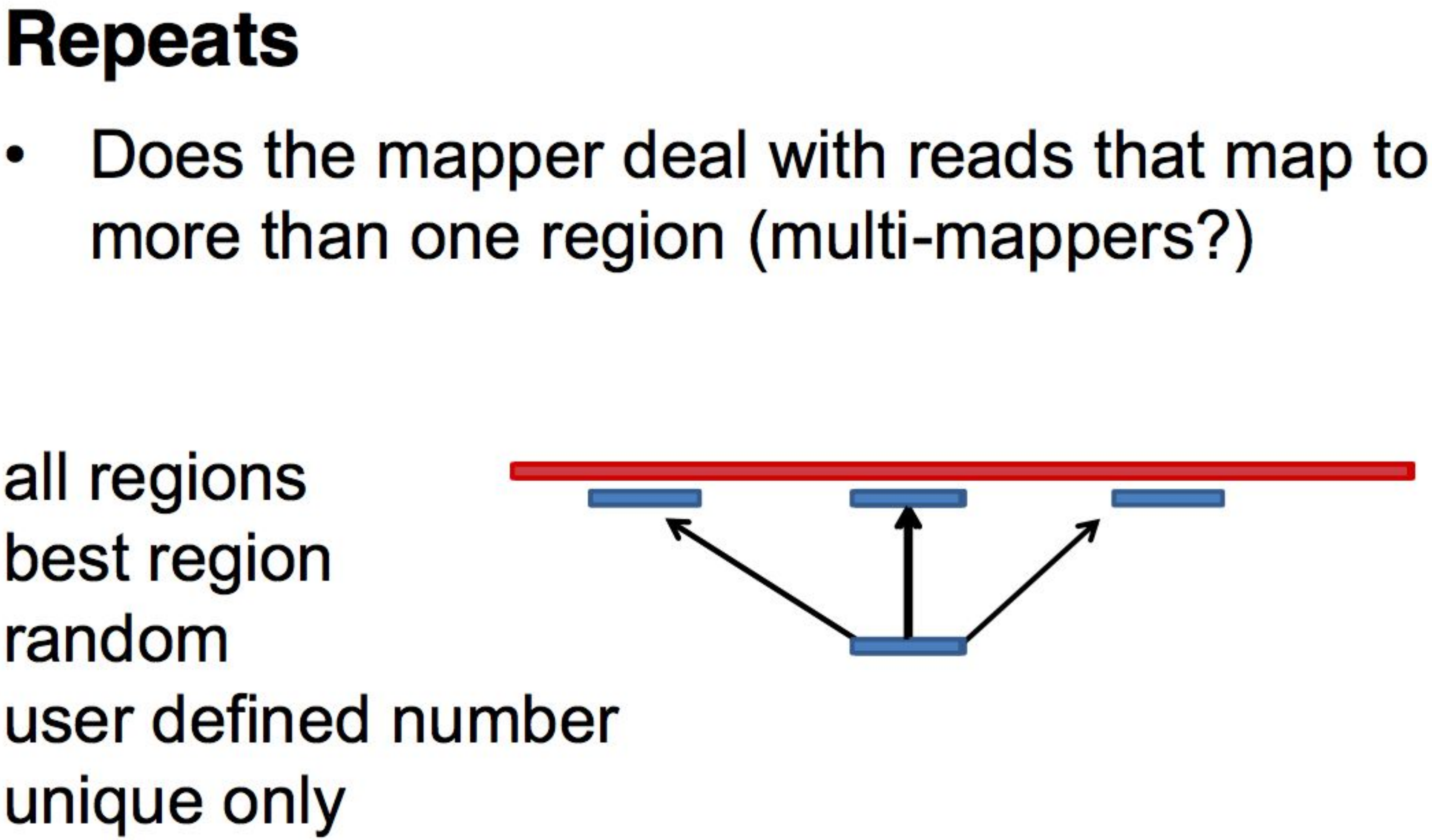
http://www.bioinformatics.babraham.ac.uk/projects/fastqc/small_rna_fastqc.html

Strand bias


100 bp
:00 bp
106 857 310 bp
106 857 320 bp
106 857 330 bp
106 857 340 bp
106 857 350 bp
106 857 360 bp
106 857 370 bp
106 857 380 bp
106 857 390 bp
106 85
[0 - 122]
T
C

Ошибки выравнивания

Неоднозначность выравнивания

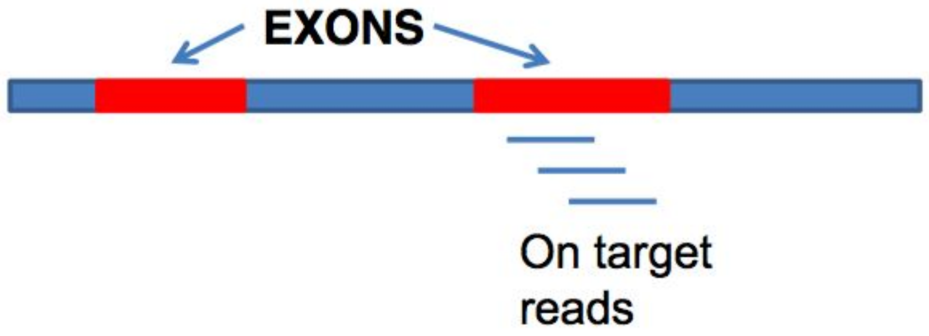


Repeats

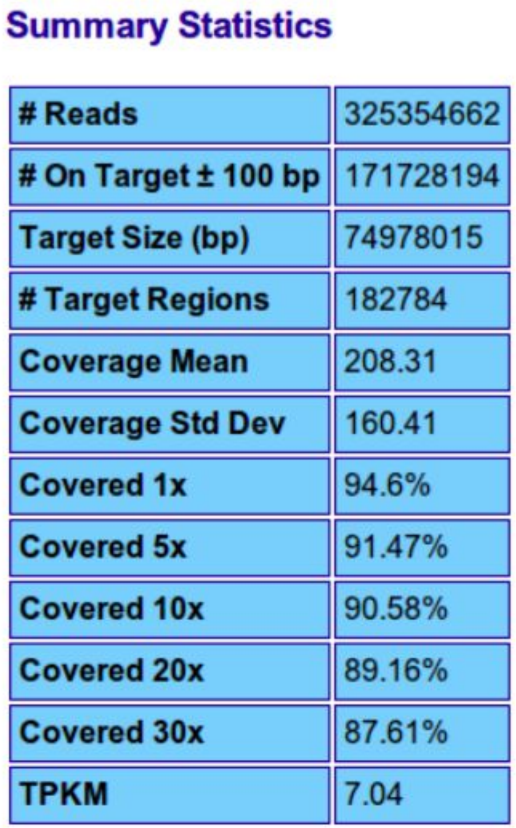
- Does the mapper deal with reads that map to more than one region (multi-mappers?)

all regions
best region
random
user defined number
unique only

Картирование за пределы таргета



NGSrich:

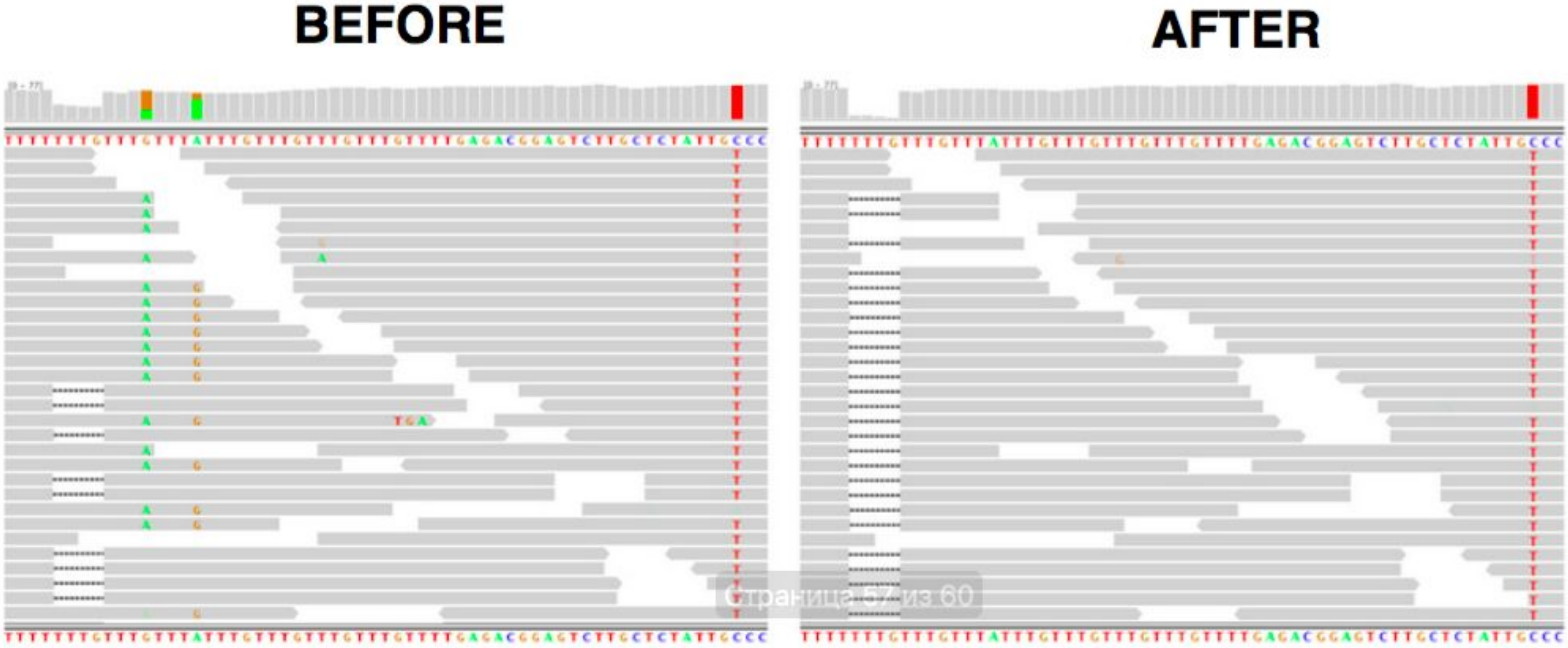


Summary Statistics

# Reads	325354662
# On Target ± 100 bp	171728194
Target Size (bp)	74978015
# Target Regions	182784
Coverage Mean	208.31
Coverage Std Dev	160.41
Covered 1x	94.6%
Covered 5x	91.47%
Covered 10x	90.58%
Covered 20x	89.16%
Covered 30x	87.61%
TPKM	7.04

http://sourceforge.net/projects/ngsrich/

Перевыравнивание инделов

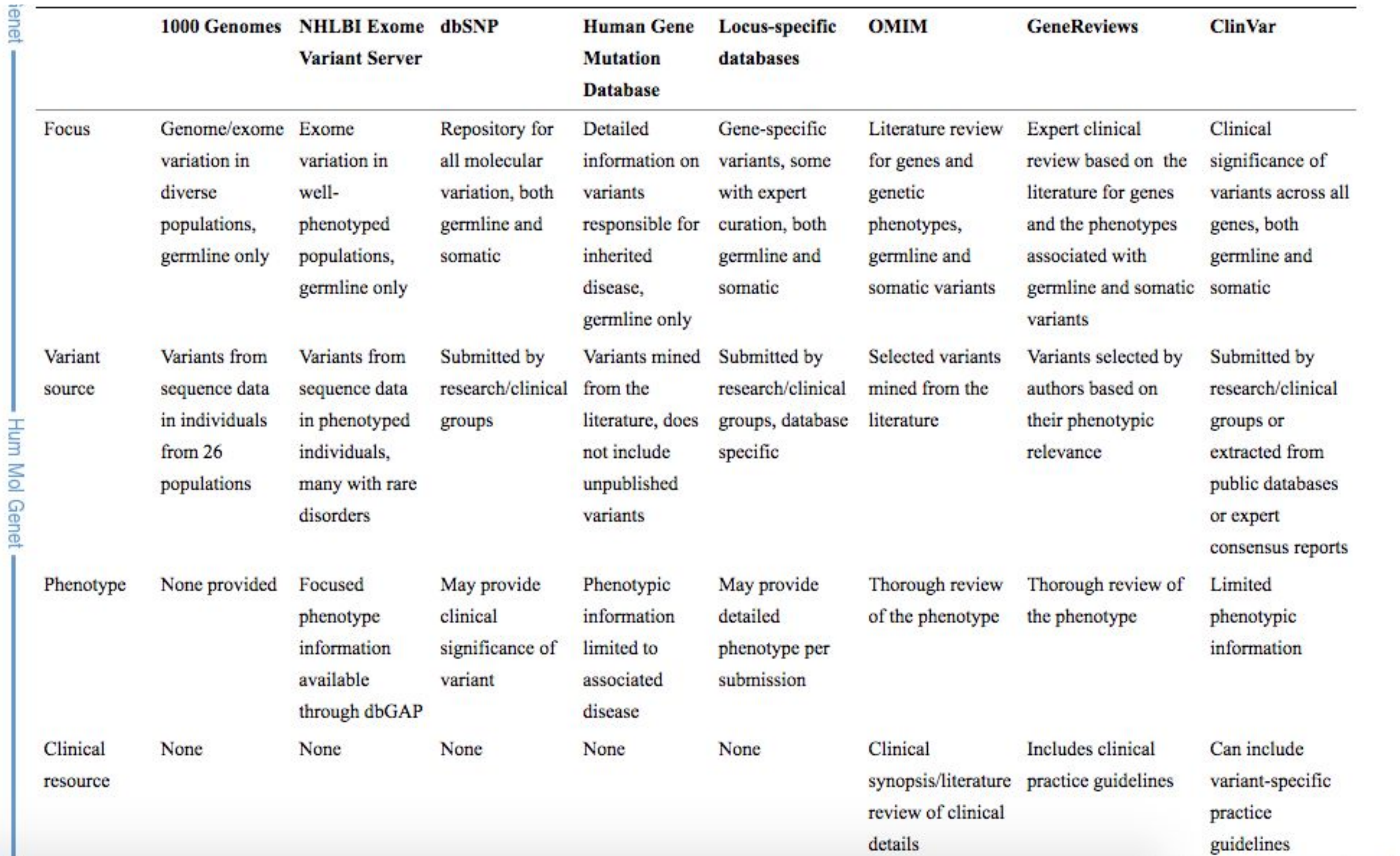


https://www.broadinstitute.org/gatk/

Ошибки интерпретации

PMC full text: Hum Mol Genet. 2013 Oct 15; 22(R1): R27–R31.
Published online 2013 Aug 19. doi: 10.1093/hmg/ddt384
Copyright/License ▶ Request permission to reuse

Противоречивые базы данных


ienet
Hum Mol Genet

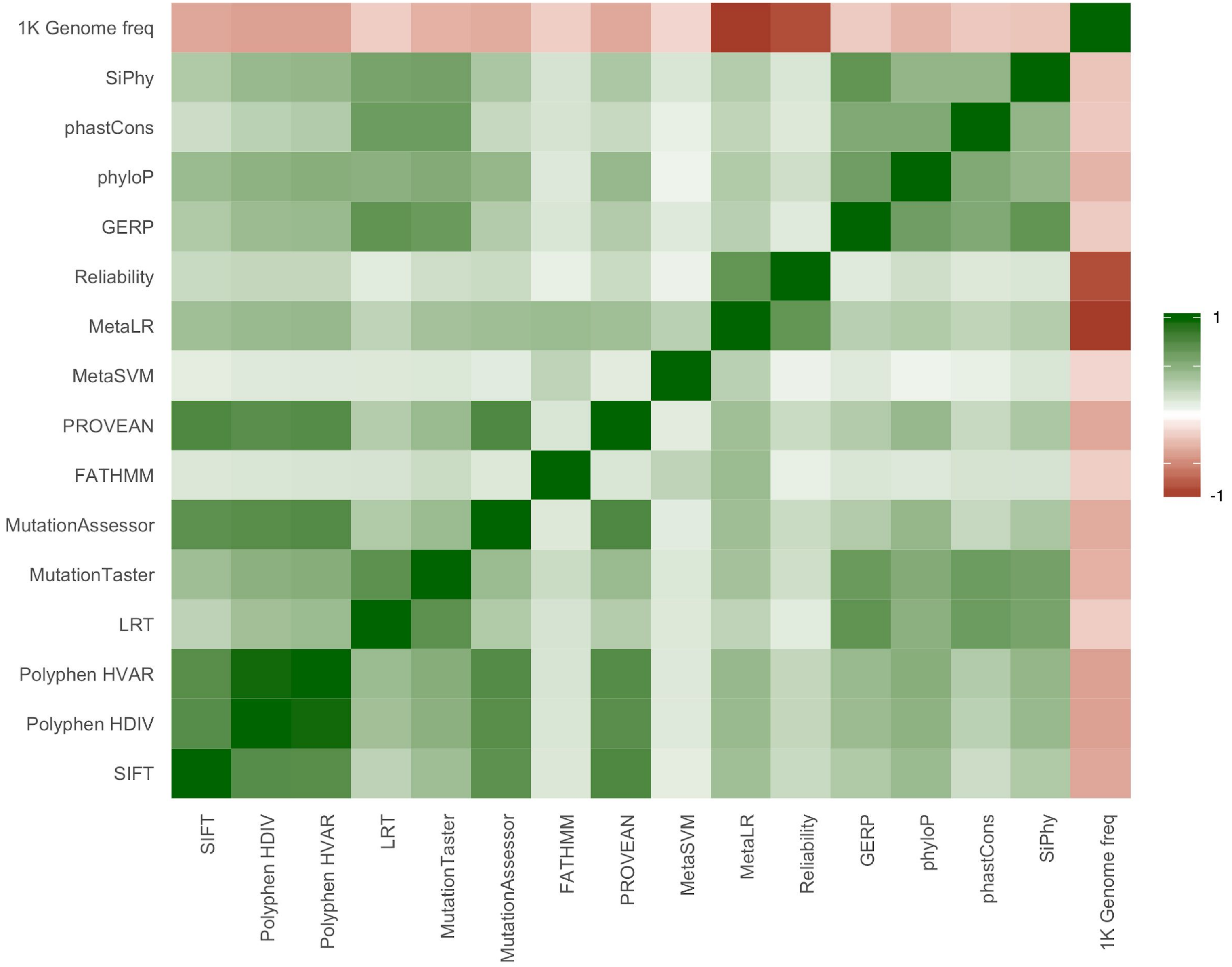
	1000 Genomes	NHLBI Exome Variant Server	dbSNP	Human Gene Mutation Database	Locus-specific databases	OMIM	GeneReviews	ClinVar
Focus	Genome/exome variation in diverse populations, germline only	Exome variation in well-phenotyped populations, germline only	Repository for all molecular variation, both germline and somatic	Detailed information on variants responsible for inherited disease, germline only	Gene-specific variants, some with expert curation, both germline and somatic	Literature review for genes and genetic phenotypes, germline and somatic variants	Expert clinical review based on the literature for genes and the phenotypes associated with germline and somatic variants	Clinical significance of variants across all genes, both germline and somatic
Variant source	Variants from sequence data in individuals from 26 populations	Variants from sequence data in phenotyped individuals, many with rare disorders	Submitted by research/clinical groups	Variants mined from the literature, does not include unpublished variants	Submitted by research/clinical groups, database specific	Selected variants mined from the literature	Variants selected by authors based on their phenotypic relevance	Submitted by research/clinical groups or extracted from public databases or expert consensus reports
Phenotype	None provided	Focused phenotype information available through dbGAP	May provide clinical significance of variant	Phenotypic information limited to associated disease	May provide detailed phenotype per submission	Thorough review of the phenotype	Thorough review of the phenotype	Limited phenotypic information
Clinical resource	None	None	None	None	None	Clinical synopsis/literature review of clinical details	Includes clinical practice guidelines	Can include variant-specific practice guidelines

ClinVar

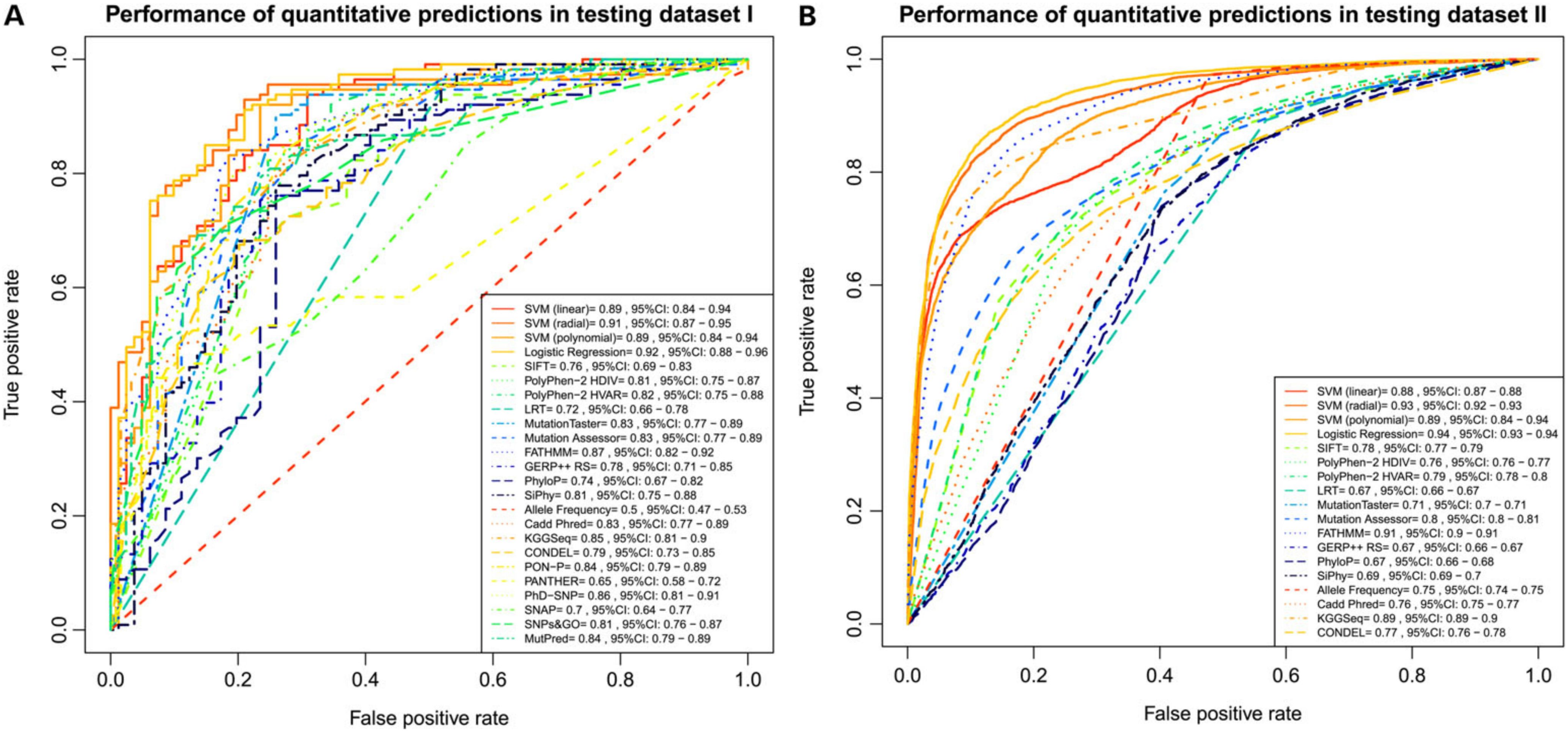
- As of May 4, 2015 (according to the *NEJM* report), ClinVar contains 172,055 variant submissions across 22,864 genes from 314 submitters—35 of which have deposited more than 50 genetic variants with medical interpretation into ClinVar.
- More than 118,000 of the unique variants have clinical interpretations, though 21% of those interpretations are clinical question marks—variants of uncertain significance.
- Only 11% of the variants with clinical interpretations have been submitted by more than one lab, the first step in arriving at a consensus. For 17% of those, the interpretations do not agree.

Missense prediction	ConSurf	http://consurftest.tau.ac.il	Evolutionary conservation
	FATHMM	http://fathmm.biocompute.org.uk	Evolutionary conservation
	MutationAssessor	http://mutationassessor.org	Evolutionary conservation
	PANTHER	http://www.pantherdb.org/tools/csnpScoreForm.jsp	Evolutionary conservation
	PhD-SNP	http://snps.biofold.org/phd-snp/phd-snp.html	Evolutionary conservation
	SIFT	http://sift.jcvi.org	Evolutionary conservation
	SNPs&GO	http://snps-and-go.biocomp.unibo.it/snps-and-go	Protein structure/function
	Align GVGD	http://agvgd.iarc.fr/agvgd_input.php	Protein structure/function and evolutionary conservation
	MAPP	http://mendel.stanford.edu/SidowLab/downloads/MAPP/index.html	Protein structure/function and evolutionary conservation
	MutationTaster	http://www.mutationtaster.org	Protein structure/function and evolutionary conservation
	MutPred	http://mutpred.mutdb.org	Protein structure/function and evolutionary conservation
	PolyPhen-2	http://genetics.bwh.harvard.edu/pph2	Protein structure/function and evolutionary conservation
	PROVEAN	http://provean.jcvi.org/index.php	Alignment and measurement of similarity between variant sequence and protein sequence homolog
	nsSNPAnalyzer	http://snpanalyzer.uthsc.edu	Multiple sequence alignment and protein structure analysis
	Condel	http://bg.upf.edu/fannsdb/	Combines SIFT, PolyPhen-2, and MutationAssessor
	CADD	http://cadd.gs.washington.edu	Contrasts annotations of fixed/nearly fixed derived alleles in humans with simulated variants
Nucleotide conservation prediction	GERP	http://mendel.stanford.edu/sidowlab/downloads/gerp/index.html	Genomic evolutionary rate profiling
	PhastCons	http://compgen.bscb.cornell.edu/phast/	Conservation scoring and identification of conserved elements
	PhyloP	http://compgen.bscb.cornell.edu/phast/	
		http://compgen.bscb.cornell.edu/phast/help-pages/phyloP.txt	Alignment and phylogenetic trees: Computation of *P* values for conservation or acceleration, either lineage-specific or across all branches

Такие противоречивые скоры

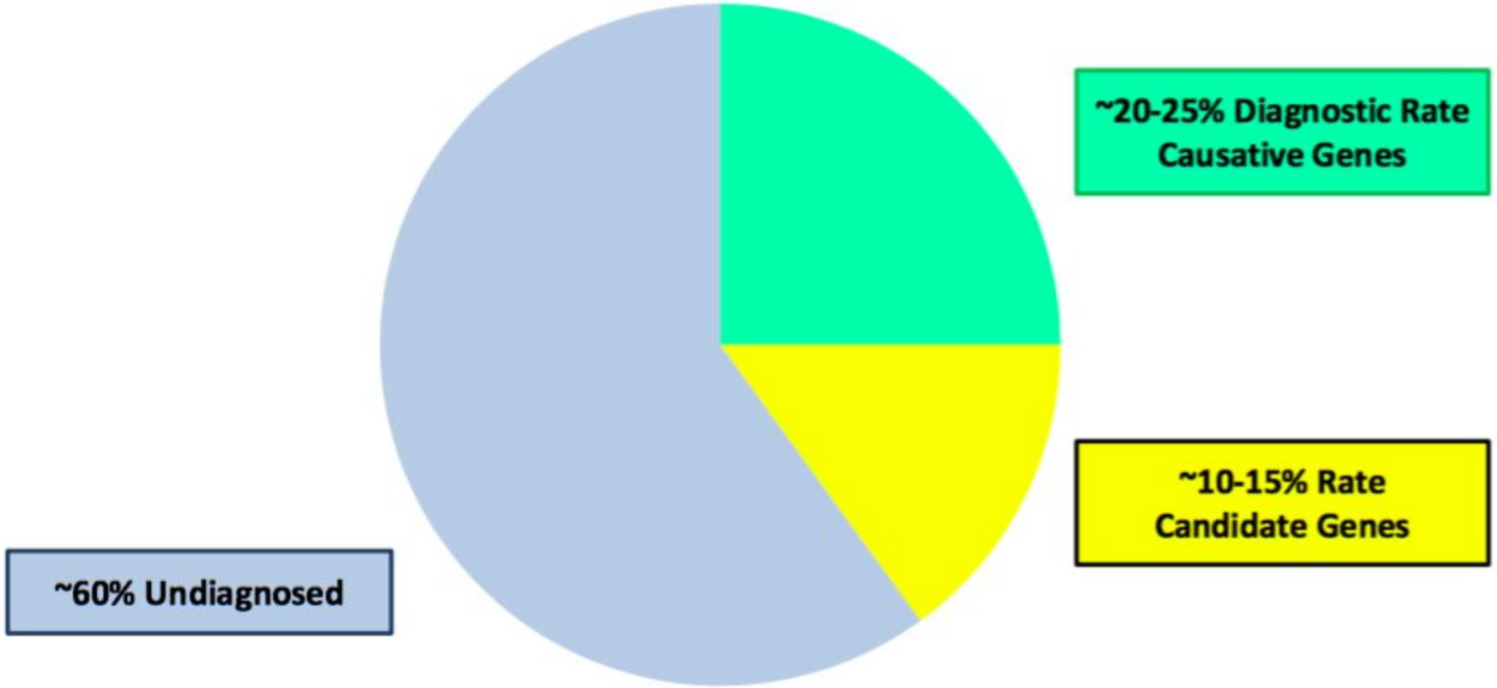


Такие противоречивые скоры



Фундаментальный недостаток знаний

Current Exome Diagnostic Yield in Undiagnosed Disorders*



*NIH UDP (ASHG 2011)
*Baylor (Yang et al NEJM 2013 and JAMA 2014)
*FORGE Canada (Sawyer et al Hum Mut 2014)
*UCLA (Lee et al JAMA 2014)
*ARUP Laboratories

ТАКИЕ ДЕЛА