Calling DNA variants (SNV, SV, CNV), limitations from sequencing technologies

Victor Guryev

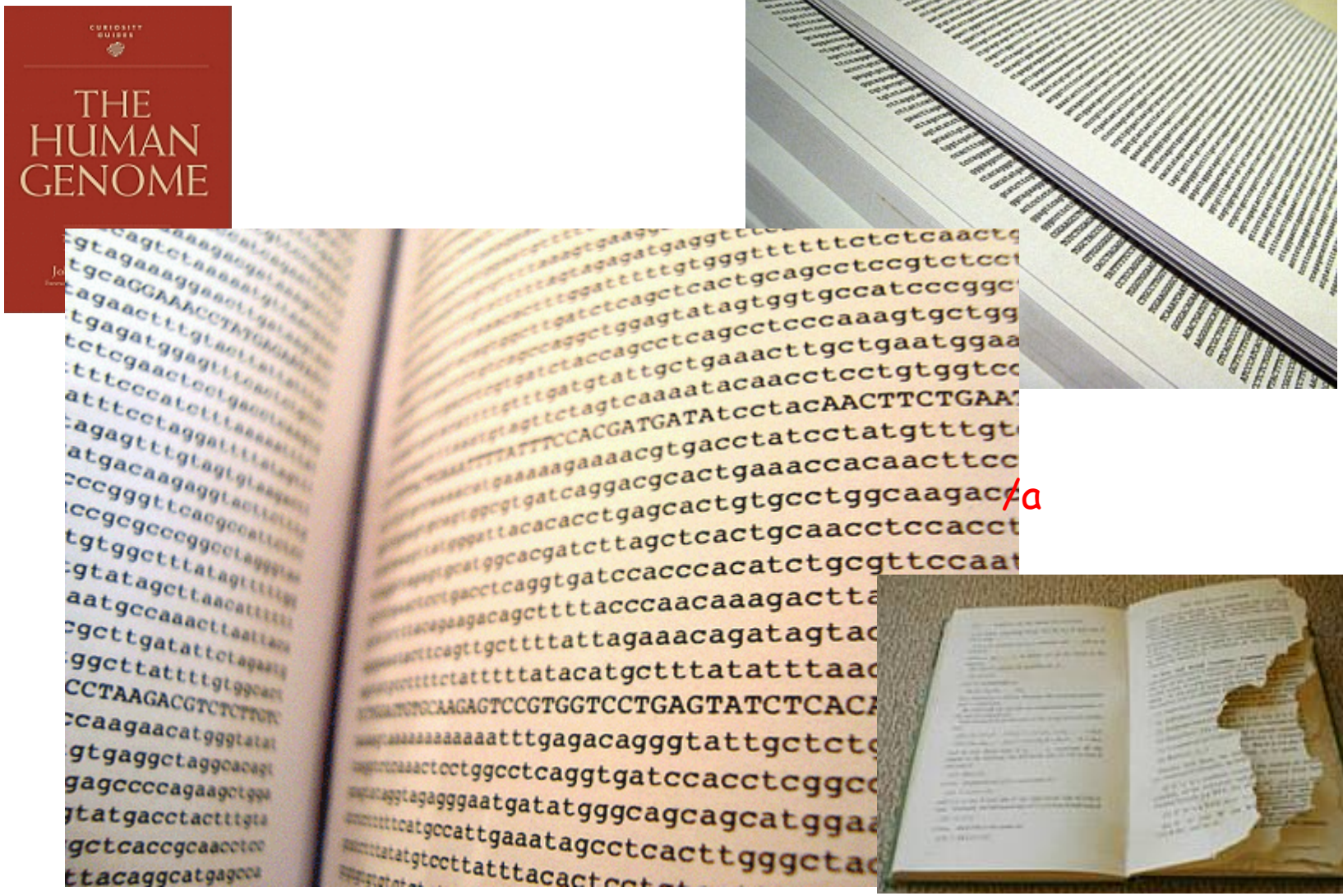
April 2, 2019

International Postgraduate Course on Variant Effect Prediction

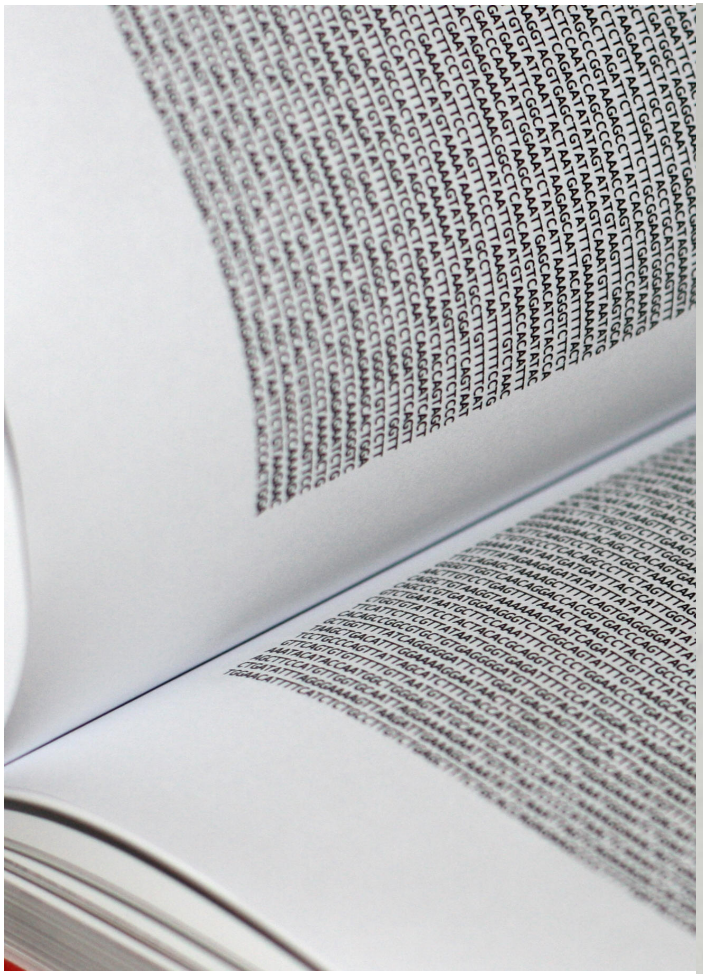
Avans School, Breda

ERIBA

Our genomes: base and structural variants



How do we get our NGS genomes?



ERIBA

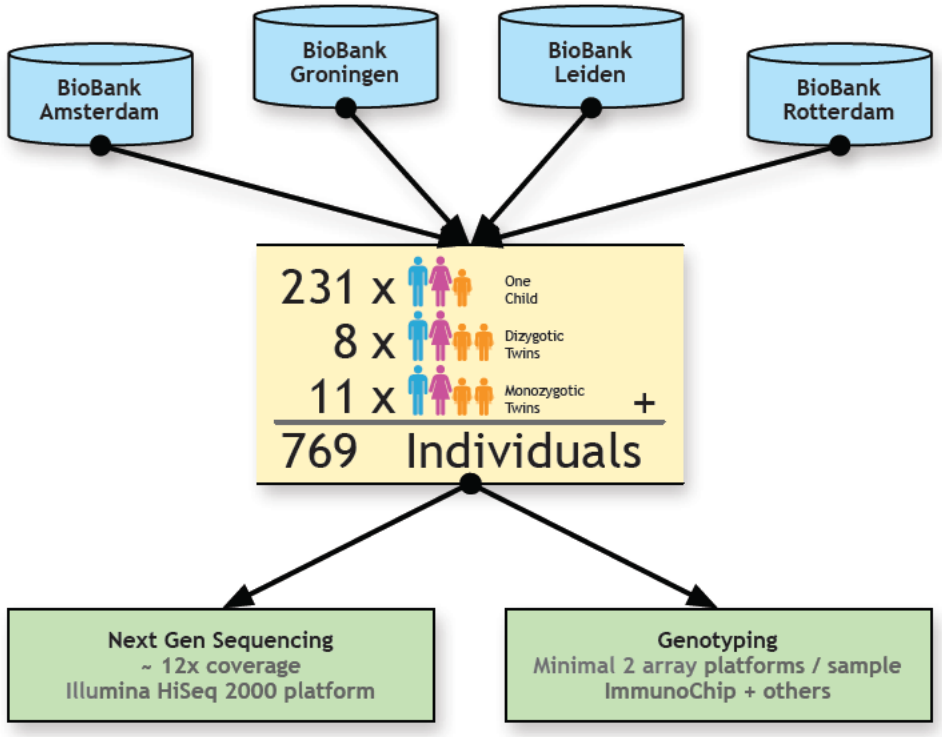
1000 genomes project (1kG)

Low coverage whole genome and deep exome sequencing of 2,500 individuals to discover 95% of variants at 1% frequency

ERIBA

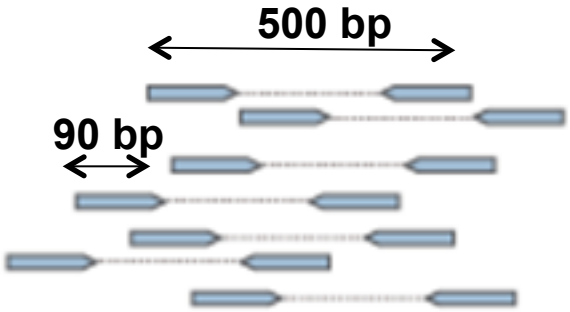
Small variants: The 1000 Genomes Project Consortium, 2015. Nature 526:68-74
Structural variants: Sudmant et al, 2015. Nature 526:75-81

Genome of the Netherlands (GoNL)



Position paper: Boomsma et al, 2013
Small variants: Francioli et al, 2014
Structural variants: Hehir-Kwa et al, 2016

Median base coverage: 12x



	1000 G	GoNL
DNA source	Cell lines	Blood
Coverage	3-4x	>12x
Data generation	Mult. platforms	BGI/Illumina
Population	Multiple, unrelated	Dutch only, trios, twins
Phenotype info	None	Multiple

NGS platforms

Illumina
HiSeq 2500
NextSeq
NovaSeq

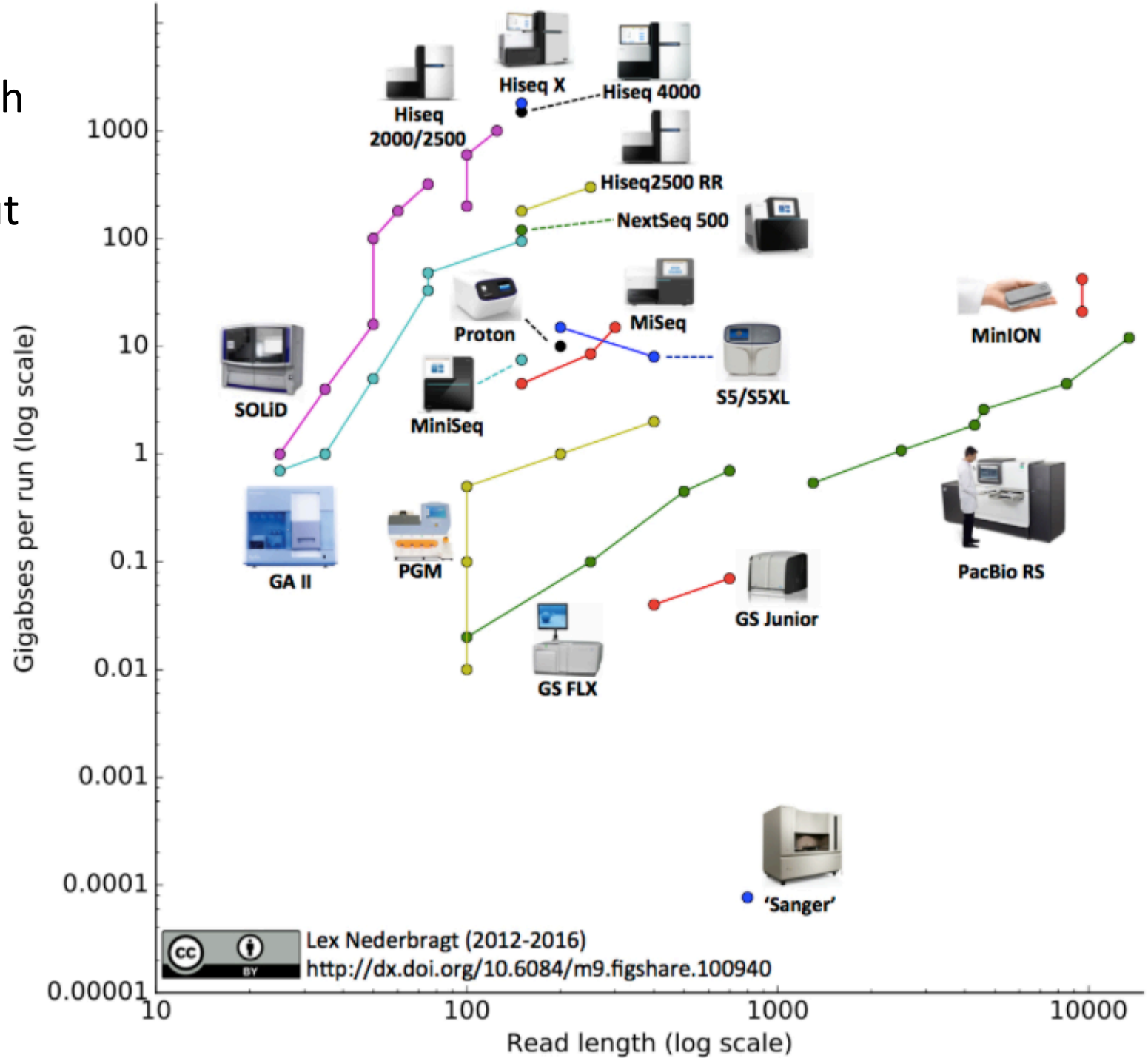
Thermo Fisher
Ion Torrent
Ion Proton

PacBio
RS II
Sequel

Oxford Nanopore
MinION
PromethION

ERIBA

Read length
vs
throughput



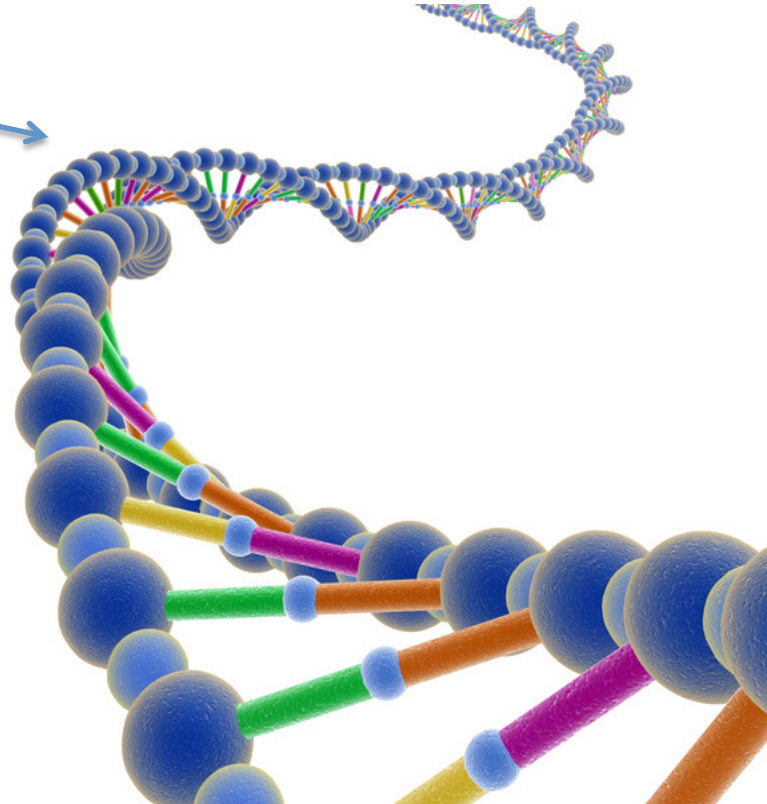
ERIBA

Combining platforms is power

Long low-quality reads

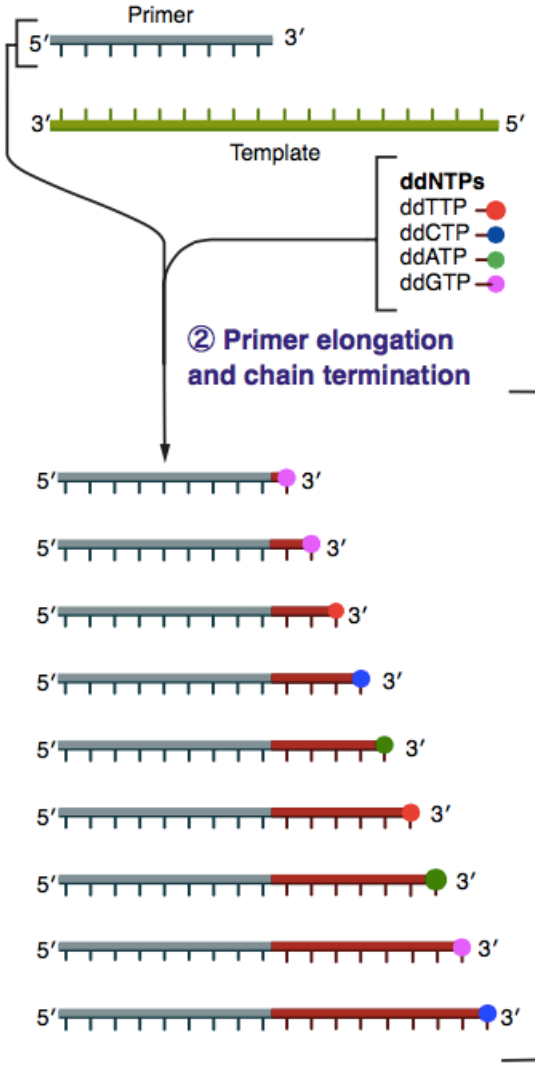
Short **high-quality** reads

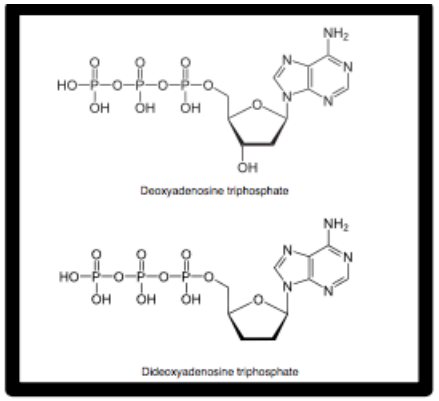
Long high-quality reads



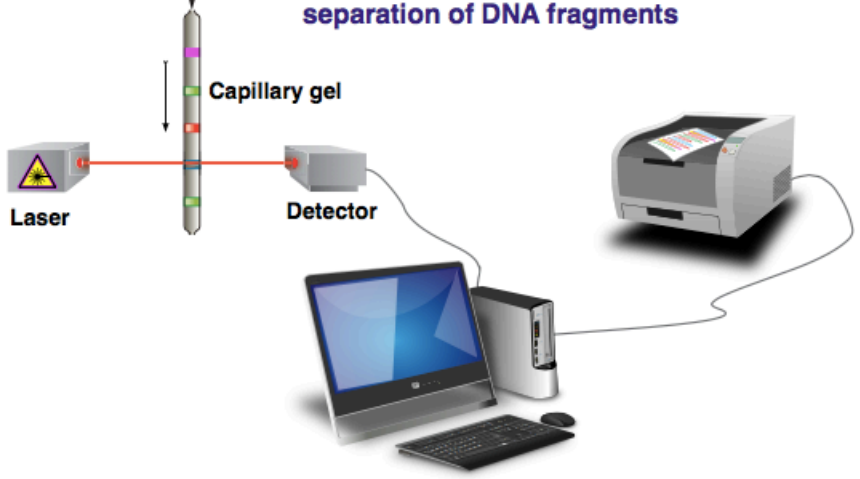
Most common way: sequencing by synthesis

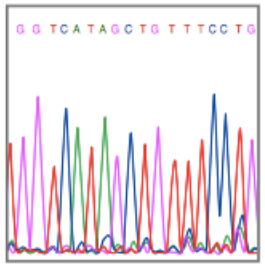
① Reaction mixture
- Primer and DNA template
- DNA polymerase
- ddNTPs with flourochromes
- dNTPs (dATP, dCTP, dGTP, and dTTP)





③ Capillary gel electrophoresis separation of DNA fragments





Chromatograph

④ Laser detection of flourochromes and computational sequence analysis

Quality measures

Formula: Phred value = $-10 \times \text{Log}_{10}(\text{Error\_rate})$

What does it mean in practice:

- Quality 0 = Call cannot be made (base = N)
- Quality 10 = 10% error (90% confidence)
- Quality 20 = 1% error (99% confidence)
- Quality 30 = 0.1% error (99.9% confidence)

Data storing in Sanger sequencing times

File A01.fasta

```
>A01
CAGCATGCTATCGTAGTCGTAGTCGTAGCTGTACATGCTGATCTGATGCT
GATCGTAGCTGATGCTCTAGTACGTGCATGGTCATGTGATCGGGTACGCA
TGTCATGTC
```

File A01.qual

```
>A01
23 24 26 38 31 11 27 28 25 28 22 25 27 28 36 27 32
22 33 23 27 16 40 33 18 28 28 24 25 20 26 26 37 31
10 21 27 16 36 28 32 22 27 26 28 37 30 9 28 27 26
36 29 8 33 23 37 30 9 37 30 9 34 26 32 22 28 28 28
22 33 23 28 31 21 28 26 33 23 28 27 28 28 28 21 25
37 33 16 34 28 25 28 37 33 17 28 28 27 34 27 25 30
25 26 24 34 27 34 27 23 28 36 32 14 24 28 27 27 23
```

Encoding base quality values

Char	Value	Char	Value	Char	Value	Char	Value
null	000	¶	020	(	040	=	061
☺	001	§	021	)	041	>	062
☻	002	▬	022	*	042	? q=30	063
♥	003	↕	023	+ q=10	043	@	064
♦	004	↑	024	,	044	A	065
♣	005	↓	025	-	045	B	066
♠	006	→	026	.	046	C	067
•	007	←	027	/	047	D	068
◘	008	∟	028	0	048	E	069
○	009	↔	029	1	049	F	070
◙	010	▲	030	2	050	G	071
♂	011		031	3	051	H	072
♀	012	▼		4	052	I q=40	073
♪	013	space	032	5 q=20	053	J	074
♫	014	! q=0	033	6	054	K	075
		"	034	7	055	L	076
☼	015	#	035	8	056	M	077
►	016	$	036	9	057	N	078
◄	017	%	037	:	058	O	079
↕	018	&	038	;	059	P	080
‼	019	'	039	<	060	Q	081

ERIBA

FASTQ file format – raw, unaligned reads

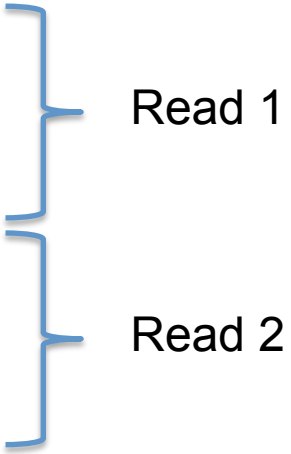
A variant of FASTA format, but with quality information

Every sequence entry has 4 lines:

Sequence name (FlowcellID/Instrument, Lane, Tile and position)
Sequence
+ (may be followed by read name again)
Quality information

Example:

```
@HWIEAS210R_0016:5:1:5510:19182
AAACTGCGCTCTCAAAAGGAGTGTTCAACTCCGTGAGT
+
DDDDDDCEDDDFEADDDECDDBEACADCCDBAFEDADD
@HWIEAS210R_0016:5:1:5510:14732
ACACTGCTTGCTATGCATGTATGAATTCCTCTGGGAAA
+
FFFFFEFFFEFFFFFFFFFFFFEFFFFFFFFFFCDDDD
```



ERIBA

Alignment algorithms, pre-NGS

Smith-Waterman algorithm
for local alignments

Dynamic programming,
Finds most optimal alignment

		C	O	E	L	A	C	A	N	T	H
	0	0	0	0	0	0	0	0	0	0	0
P	0	0	0	0	0	0	0	0	0	0	0
E	0	0	0	1	0	0	0	0	0	0	0
L	0	0	0	0	↖2	←1	0	0	0	0	0
I	0	0	0	0	↑1	↖1	0	0	0	0	0
C	0	1	0	0	0	0	↖2	0	0	0	0
A	0	0	0	0	0	1	0	↖3	←2	←1	0
N	0	0	0	0	0	0	0	↑1	↖4	←3	←2

Basic **L**ocal **A**lignment **S**earch **T**ool
BLASTing against big genomes takes a lot of CPU time:
100 short sequences vs human genome, default parameters 20 CPU minutes

NGS run would take hundreds of CPU years

```
COELACANTH
||x|||
PELICAN
```

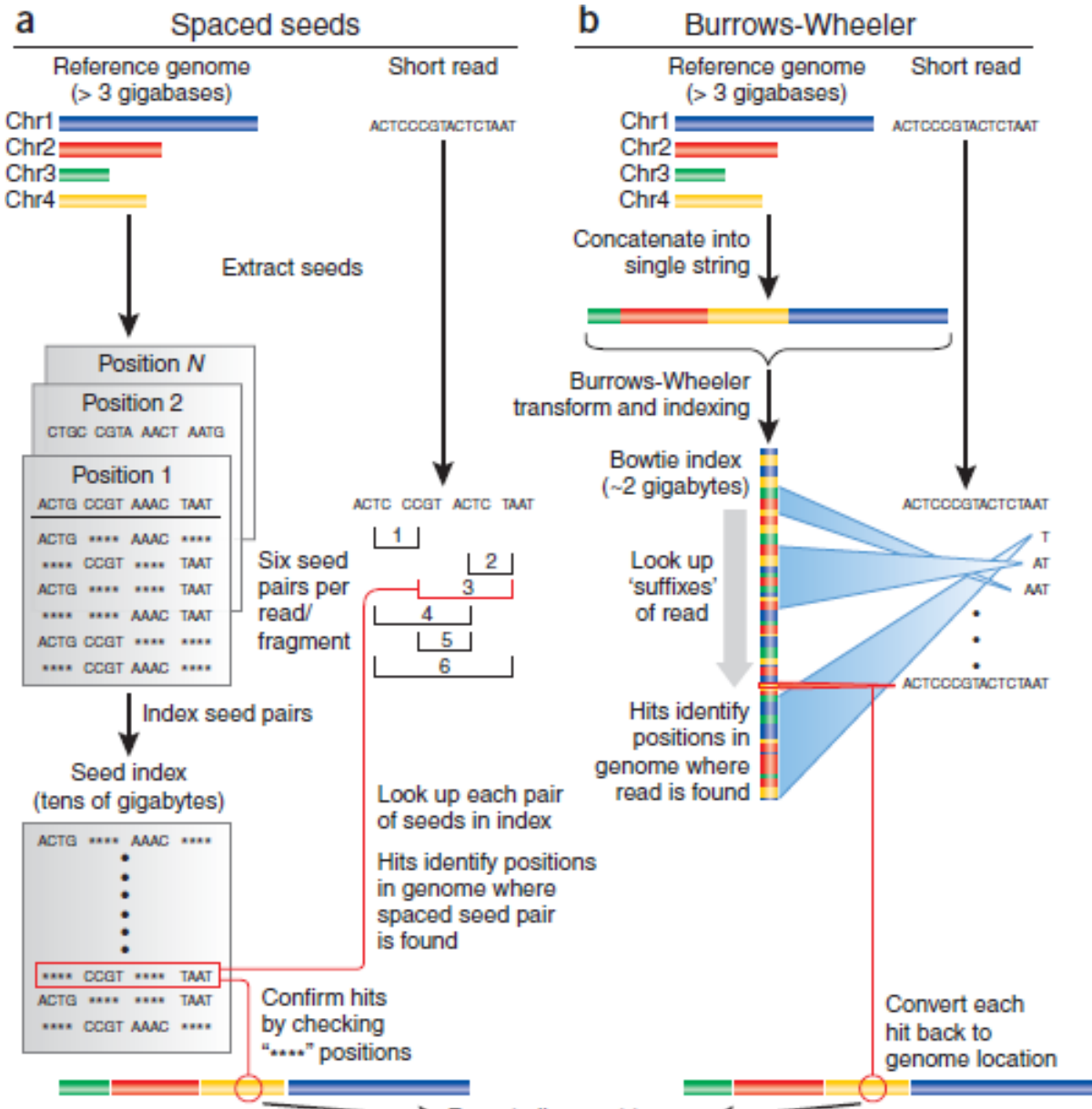
Modifications for faster mapping

MEGABLAST – a greedy algorithm for highly similar sequences, concatenates query sequences before mapping, hence > 10 times faster than BLAST, can be used for 454 runs (~5 CPU days)

Indexing with Hash Tables
SSAHA

Indexing Genome (BFAST, PASS)
Indexing Reads (MAQ, RMAP, SHRiMP)

Burrows-Wheeler Transform
(Bowtie, BWA)



[Trapnell & Salzberg, 2009]

Aligner choice: platform, application, reference

BLASR
BWA mem

PacBio

BWA Bowtie 2
CLC-Bio, Stampy

Illumina

TMAP

Ion Torrent / Proton

LAST, BLAST
BWA mem

Nanopore

ERIBA

Sample BLAST output

PREDICTED: Homo sapiens fibroblast growth factor receptor 3 (FGFR3), transcript variant X4, mRNA

Sequence ID: ref|XM_006713871.1| **Length:** 4309 **Number of Matches:** 2

Range 1: 2702 to 3124 GenBank Graphics ▼ Next Match ▲ Previous Match

Score	Expect	Identities	Gaps	Strand
743 bits(402)	0.0	417/423(99%)	5/423(1%)	Plus/Plus

```
Query  108   GGCCACTGGTCCCCAACAATGTGAGGGGTCCCTAGCAGCCCACCCTGCTGCTGGTGCACA  167
             ||||||||||||||||||||||||||||||||||||||||||||||||||||||||||||
Sbjct  2702  GGCCACTGGTCCCCAACAATGTGAGGGGTCCCTAGCAGCCCACCCTGCTGCTGGTGCACA  2761

Query  168   GCCACTCCCCGGCATGAGACTCAGTGCAGATGGAGAGACAGCTACACAGAGCTTTGGTCt  227
             ||||||||||||||||||||||||||||||||||||||||||||||||||||||||||||
Sbjct  2762  GCCACTCCCCGGCATGAGACTCAGTGCAGATGGAGAGACAGCTACACAGAGCTTTGGTCT  2821

Query  228   gtgtgtgtgtgtgtgc--gtgtgtgtgtgtgtgtgcacatccgcgtgtgcctgtgtgcgt  285
             ||||||||||||||||  ||||||||||||||||||||||||||||||||||||||||||
Sbjct  2822  GTGTGTGTGTGTGTGCGTGTGTGTGTGTGTGTGTGCACATCCGCGTGTGCCTGTGTGCGT  2881

Query  286   gcGCATCTTGCCTCCAGGTGCAGAGGTACCCTGGGTGTCCCCGCTGCTGTGCAACGGTCT  345
             ||||||||||||||||||||||||||||||||||||||||||||||||||||||||||||
Sbjct  2882  GCGCATCTTGCCTCCAGGTGCAGAGGTACCCTGGGTGTCCCCGCTGCTGTGCAACGGTCT  2941

Query  346   CCTGACTGGTGCTGCAGCACCGAGGGGCCTTTGTTCTGGGGGGACCCAGTGCAGAATGTA  405
             ||||||||||||||||||||||||||||||||||||||||||||||||||||||||||||
Sbjct  2942  CCTGACTGGTGCTGCAGCACCGAGGGGCCTTTGTTCTGGGGGGACCCAGTGCAGAATGTA  3001

Query  406   AGTGGGCCCACCCGGTGGGA-CCCCGTGGGGCAGGGAGCTGGGCCCGACATGGCT-CGGC  463
             |||||||||||||||||||| |||||||||||||||||||||||||||||||| ||||
Sbjct  3002  AGTGGGCCCACCCGGTGGGACCCCCGTGGGGCAGGGAGCTGGGCCCGACATGGCTCCGGC  3061

Query  464   CTCTGCCTTTGCACCACGGGACATCACAGGGTGCG-CTCGGCCCCTCCCACACCCAAAGC  522
             |||||||||||||||||||||||||||||||||| | ||||||||||||||||||||||||
Sbjct  3062  CTCTGCCTTTGCACCACGGGACATCACAGGGTGGGCCTCGGCCCCTCCCACACCCAAAGC  3121

Query  523   TGA  525
             |||
Sbjct  3122  TGA  3124
```

SAM format

Read ID | Chr | MapQ | Chr2 | TLen | Flag | Position | CIGAR | Position2

```
HWI-D00182:112:H9YLGADXX:1:1101:1566:2036 83 x 98641293 42 25M = 98641213 -105
```

Sequence | Quality | Aln.score | Ns | Mismatches

```
TGCTGATTAAGATGGTACATTTCTN IIFIIIIIIIIIFFFFBFFFFF<0# AS:i:-1 XN:i:0 XM:i:1
```

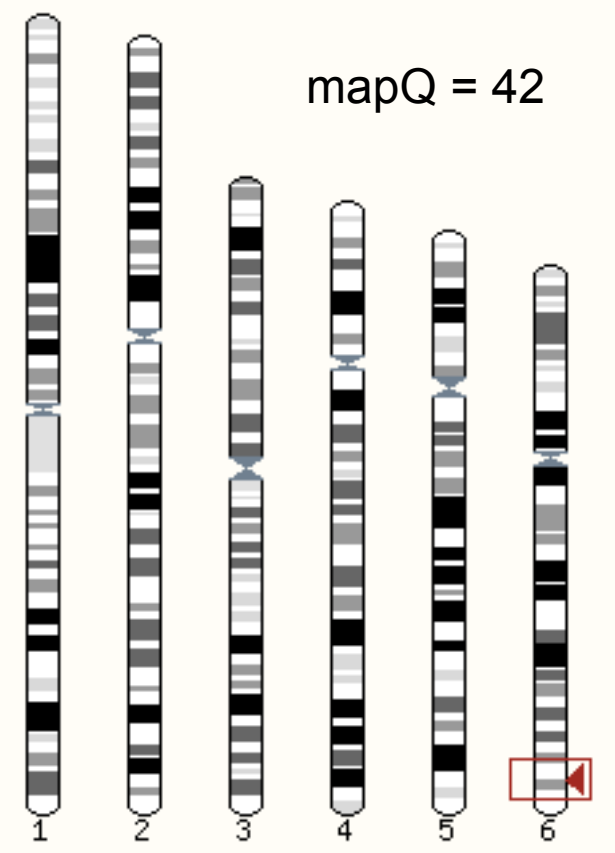
GapO | Gaps | Alignment

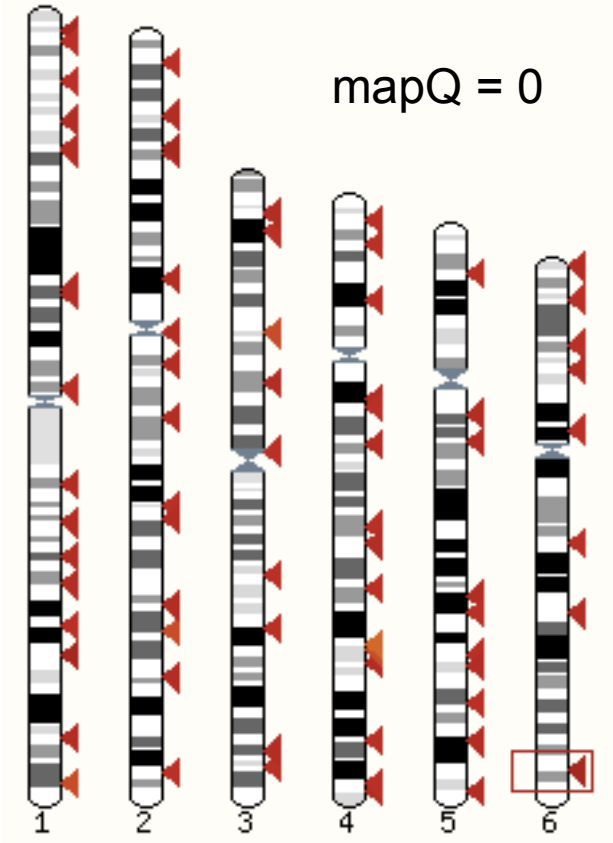
```
XO:i:0 XG:i:0 NM:i:1 MD:Z:24A0 YS:i:0 YT:Z:CP
```

```
HWI-D00182:112:H9YLGADXX:1:1101:1566:
105 TCCTGTTTCTAATAATAACTCTGCA BBBFFFF
XO:i:0 XG:i:0 NM:i:0 MD:Z:25 YS:i:-1

HWI-D00182:112:H9YLGADXX:1:1101:1570:
-112 GGGCTGTTTATCTATCCACCTACCT IIIIII
XM:i:0 XO:i:0 XG:i:0 NM:i:0 MD:Z:25 Y

HWI-D00182:112:H9YLGADXX:1:1101:1570:
112 AGGTGGTCTTCATGTAAAGACAAGG BBBFFFF
XM:i:0 XO:i:0 XG:i:0 NM:i:0 MD:Z:25 Y

HWI-D00182:112:H9YLGADXX:1:1101:1593:
-159 CAGTAGCATACCATTATATAGTGTT IIIIIF
XO:i:0 XG:i:0 NM:i:0 MD:Z:25 YS:i:0 Y
```

the read is paired in sequencing	1
the read is mapped in a proper pair	2
the query sequence itself is unmapped	4
the mate is unmapped	8
strand of the query (1 for reverse)	16
strand of the mate	32
the read is the first read in a pair	64
the read is the second read in a pair	128
the alignment is not primary	256
the read fails platform/vendor quality checks	512
the read is either a PCR or an optical duplicate	1024

Mapping quality





Mapping quality is defined similarly as base quality:

Mapq = 10	10% mapping error rate
Mapq = 20	1% mapping error rate
Mapq = 30	0.1% mapping error rate

When **multiple equally good** mapping locations are possible, mappers **randomly** select one of them

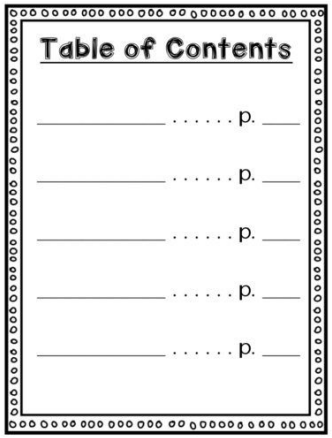
Therefore it is a good idea to disregard or account for reads with low mapQ during variant calling

Binary version, sorting and indexing

BAM – Binary SAM – compressed
much smaller file, same info

Sorted BAM – reads are ordered
by chromosome and genomic positions

Index (BAI) – needed for fast access
to any chromosomal location



All format conversion is done via **SAMTools** program
(http://samtools.sourceforge.net)

CRAM format

How much disk space does 1 base need?

1. Which base is it? (G,A,T or C), 4 options => 2 bits
2. What is quality of that base (0 .. 63) 64 options => 6 bits

8 bits (1 byte) per base.
30x human genome coverage = 9 x 10^{10} bytes = 90 Gb

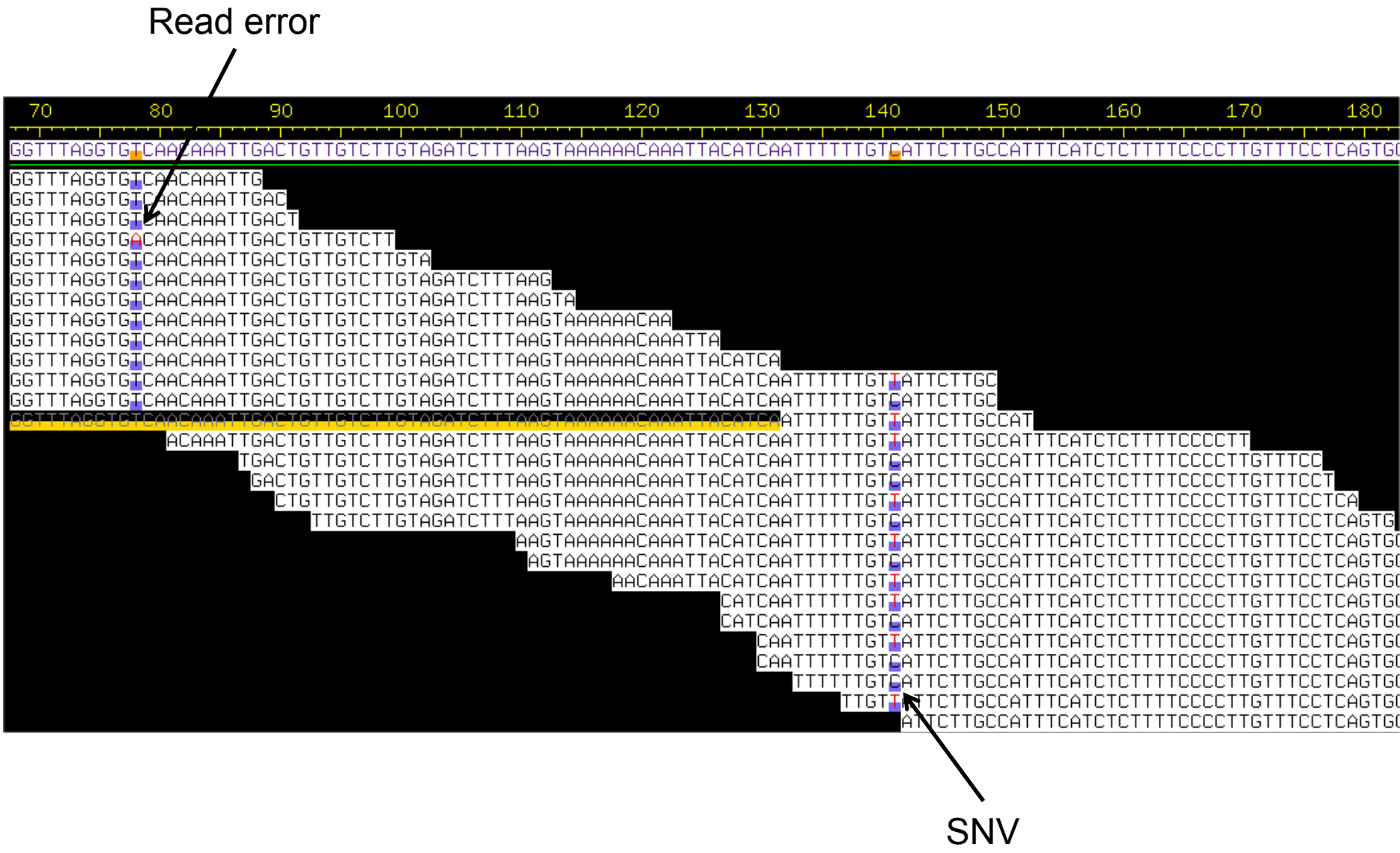
Reference based compression

```
Chr1: GGTGGAGCGCGCCGCCACGGACCACGGGCGGGCTGGCGGGCGAGCGG
Read1:     GAGCGCGCCGCCACGGACCACGGGCGGG
Read2:                   CACGGACCACGGACGGGCTGGCGGGCGAGCGG
```

Controlled loss of quality information

Low quality 0-7
Medium quality: 8-16
High quality 17-24
…

Variant calling: pileup



ERIBA

Small variant calling: pileup files

Chr	Pos	Ref	Cov	Bases	Qualities
10	264	C	9	.,,,$,,,,^F.	DBCAHBCBD
10	265	A	14	.-2TG,,,,,,,.^F.^F.^F.^F.^F,	AHJDGGDCGGCDGD
10	266	T	14	*gg,ggggGGGGGg	HAJBHDABHDJABD
10	267	G	16	*,,,,,,,,,,.,^F.^F,	+AHBCJKH>BDDIJKJ

Legend

G	Non-reference base (G) at plus strand
g	Non-reference base (G) at minus strand
. (dot)	Reference base at plus strand
, (comma)	Reference base at minus strand
^F	First base of read, read mapped with quality “F”
$	Last base of read
.-2TG	Next two bases (TG) are missing
,+1a	(insertion of base ‘A’ after this
*	gap in alignment (due to deletion)

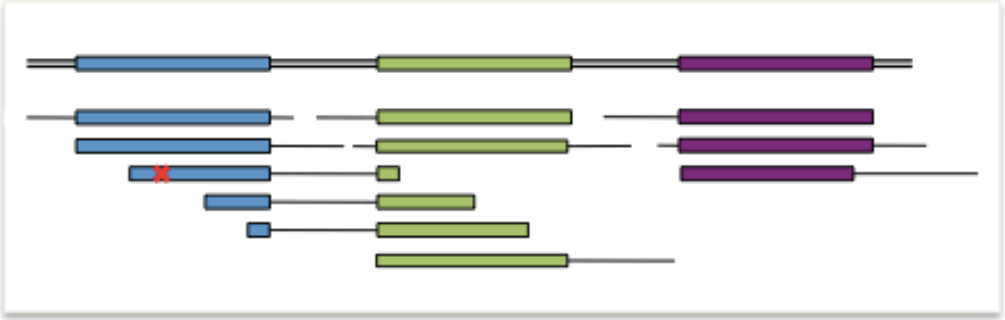
Marking/removing the PCR duplicates

✖ = sequencing error propagated in duplicates

FP variant call (bad)

Tools: BAMutil, Picard

After marking duplicates, the GATK will only see :



... and thus be more likely to make the right call

[from GATK manual]

Calling of Small Variants

Simple Variant Calling

Command:

```
samtools mpileup -uf ref.fa HG00512_merged.bam | bcftools view −vcg - >calls.vcf
```

Make pileup

Call variants

Output (calls.vcf file)

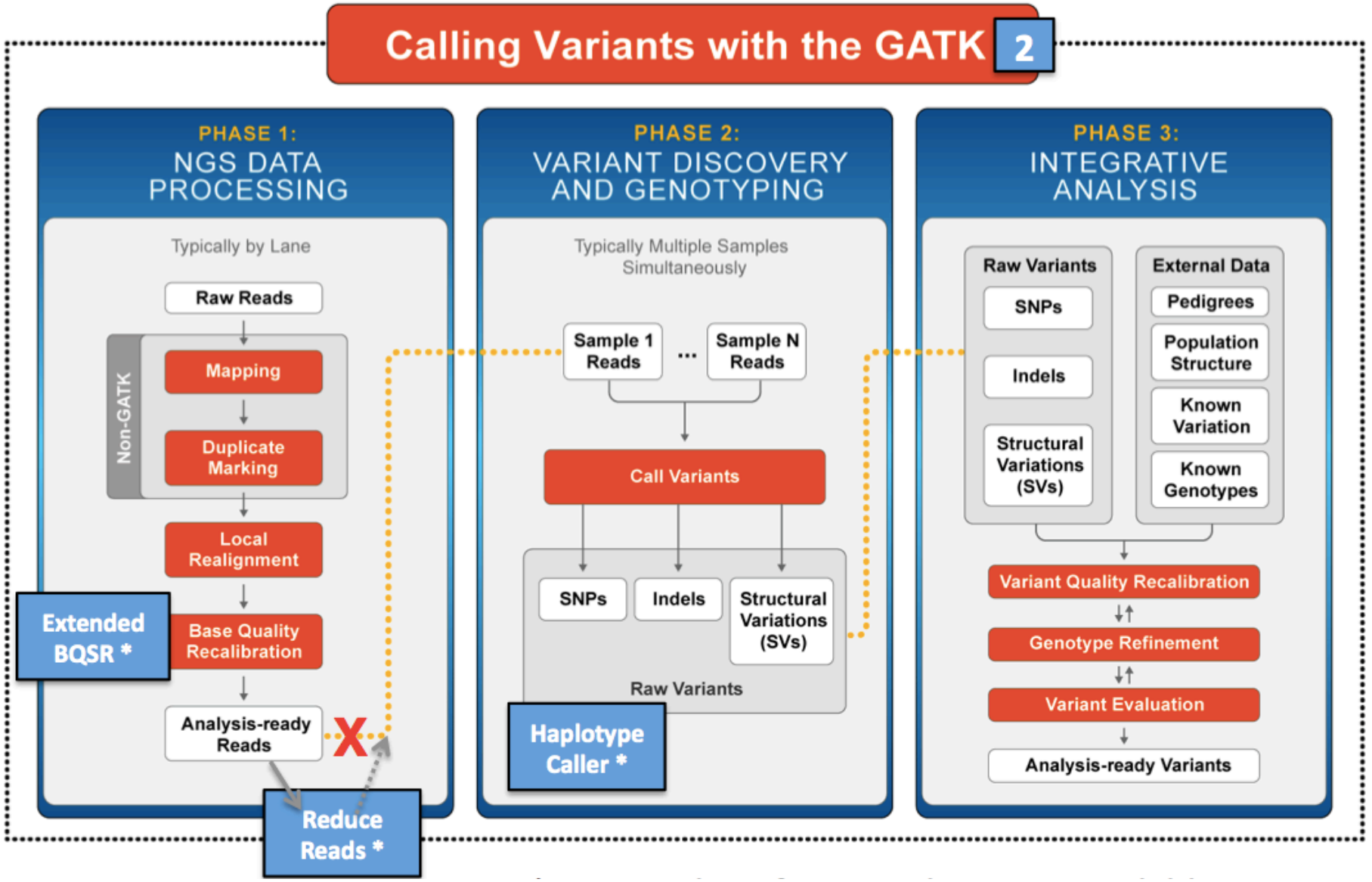
```
#CHR POS  ID REF    ALT   QUAL FILTER     INFO                                FORMAT   HG00512
Chr1 10439 . ACCCCT ACCCT 28.5 .      INDEL;DP=140;VDB=0.0157;AF1=0.5;AC1=1… GT:PL:GQ 0/1:66,0,59:61
Chr1 10442 . CCT     CTCT  6.5 .      INDEL;DP=128;VDB=0.0144;AF1=0.5;AC1=1… GT:PL:GQ 0/1:43,0,68:44
Chr1 13868 . A          G 3.55 .      DP=13;VDB=0.0275;AF1=1;AC1=2…       GT:PL:GQ 0/1:31,3,0:4
Chr1 15903 . GCC     GCCC 5.44 .      INDEL;DP=4;VDB=0.0251;AF1=1;AC1=2…  GT:PL:GQ 1/1:44,12,0:12
Chr1 16257 . G          C 8.64 .      DP=18;VDB=0.0280;AF1=0.5;AC1=1…     GT:PL:GQ 0/1:38,0,198:40
Chr1 16298 . C          T  7.8 .      DP=10;VDB=0.0280;AF1=0.5;AC1=1…     GT:PL:GQ 0/1:37,0,105:39
```

gatk talks

Genome Analysis ToolKit

BROAD INSTITUTE

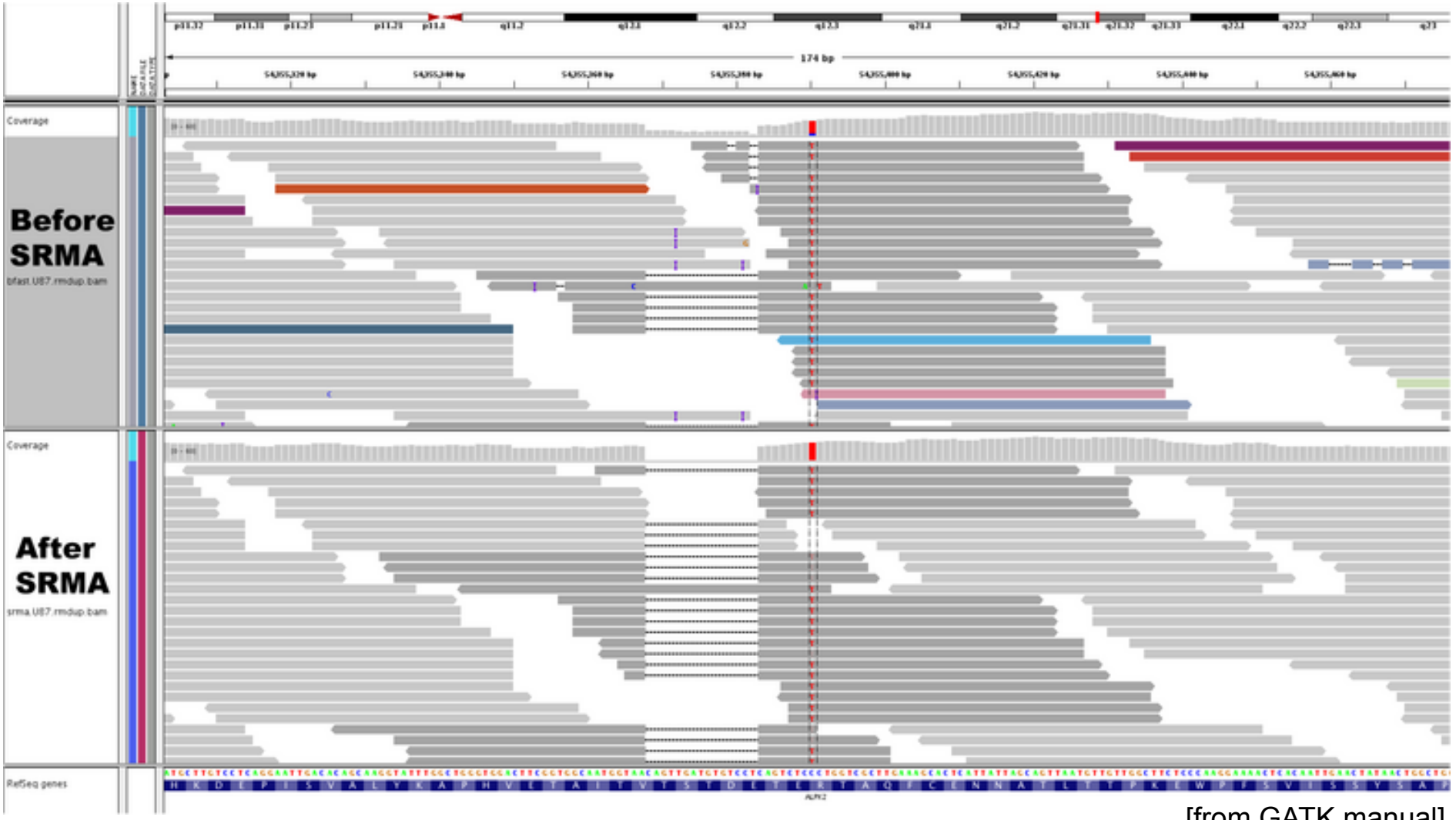
Genome Analysis ToolKit (GATK)



[from GATK manual]

Local Realignment (around indels)

```
Reference  …CATGCAGACTAAAAAAACACCATGCATCATCACTA…
Read1      …CATGCAGACTAAA----CACCATGCATCAT
Read2      …CATGCAGACTAAACA
```



[from GATK manual]

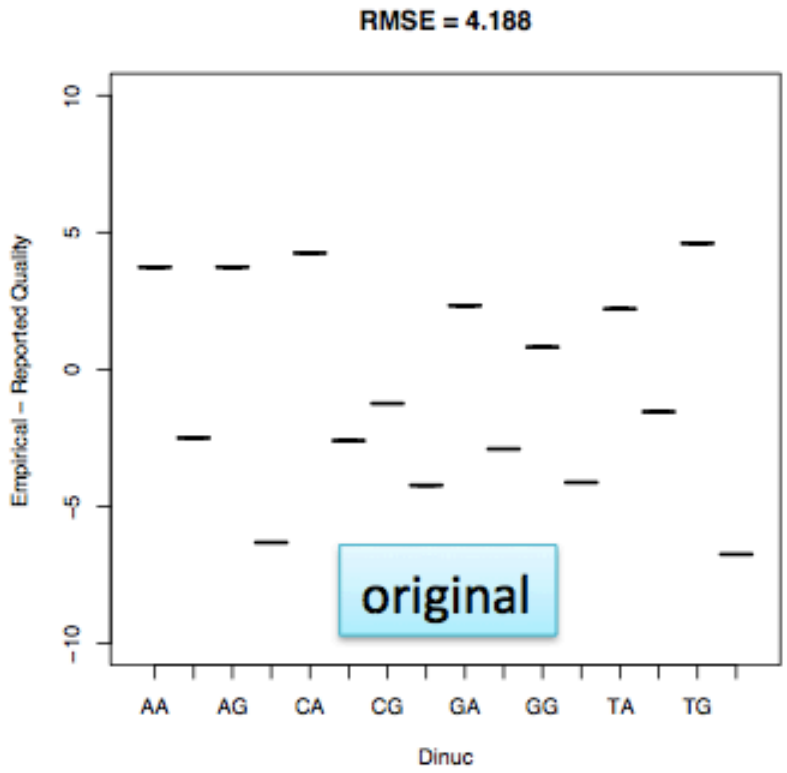
Why recalibrate qualities?

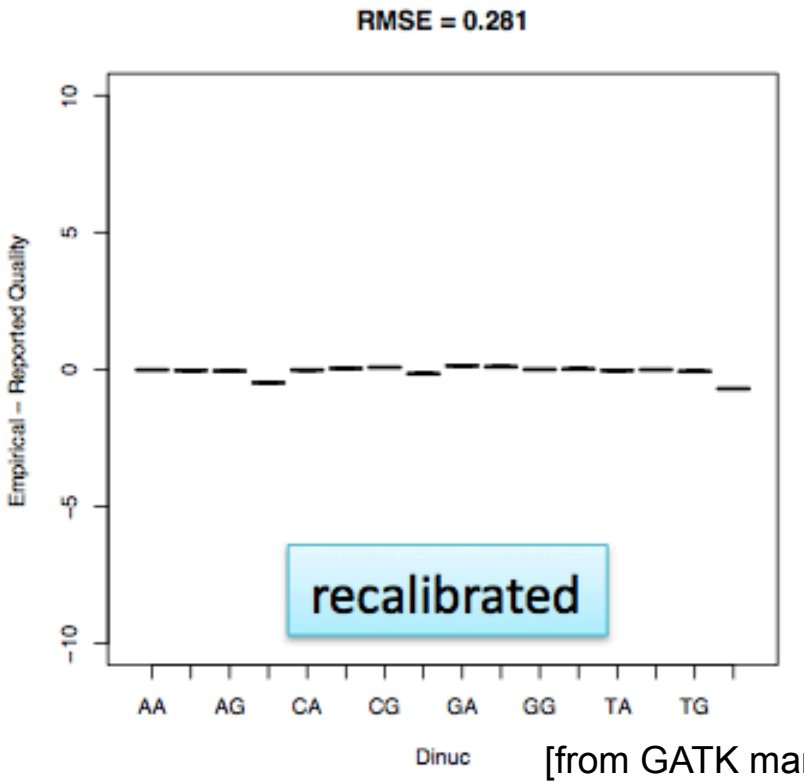
[from GATK manual]

Quality scores reported by sequencers are inaccurate and biased

- Quality scores are critical for all downstream analysis
- Systematic biases are a major contributor to bad calls

Example: Bias in the qualities reported depending on nucleotide context





[from GATK manual]

GATK variant calling approaches

- UnifiedGenotyper

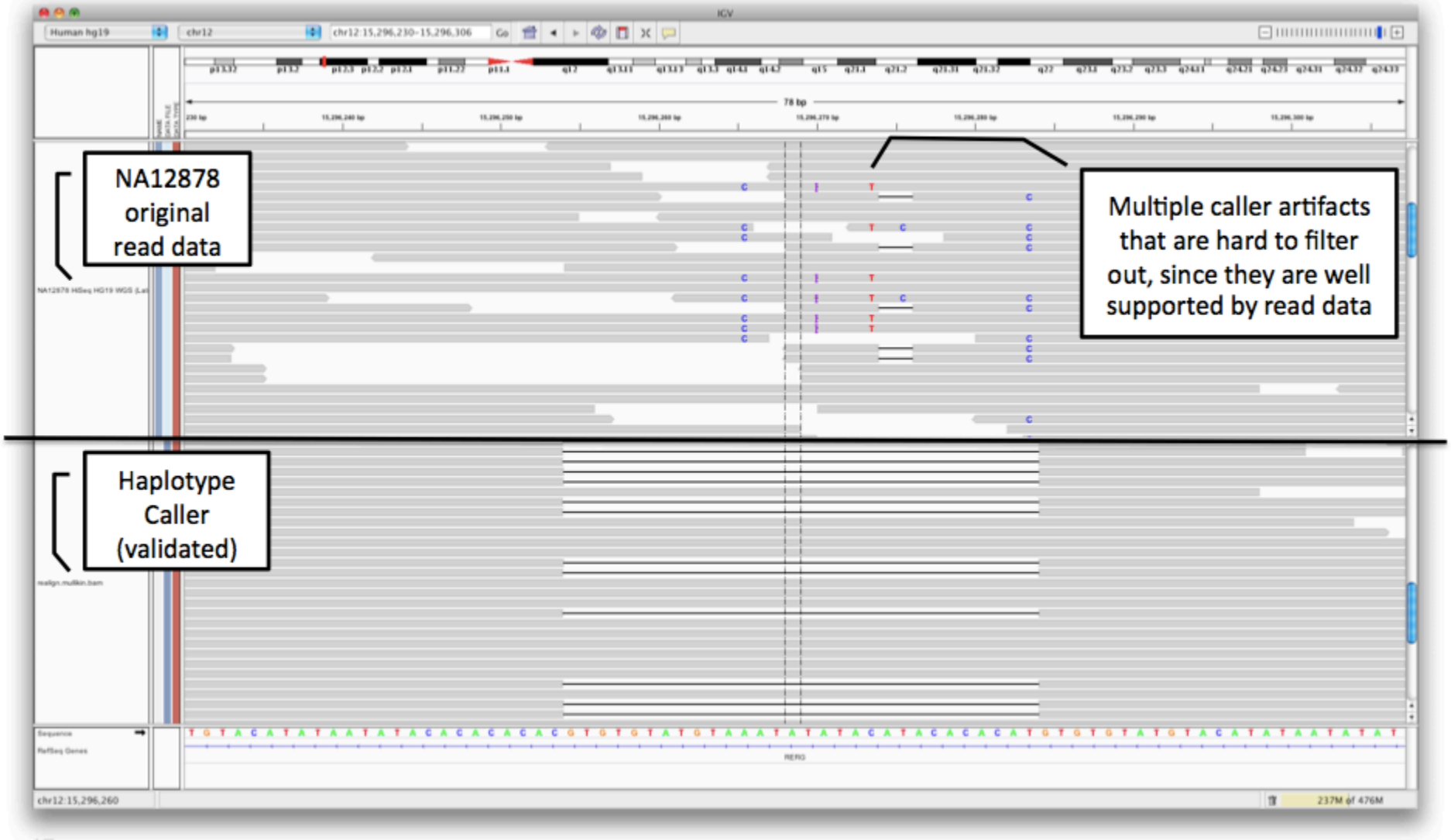
  Call SNPs and indels separately by considering each variant locus independently

- HaplotypeCaller

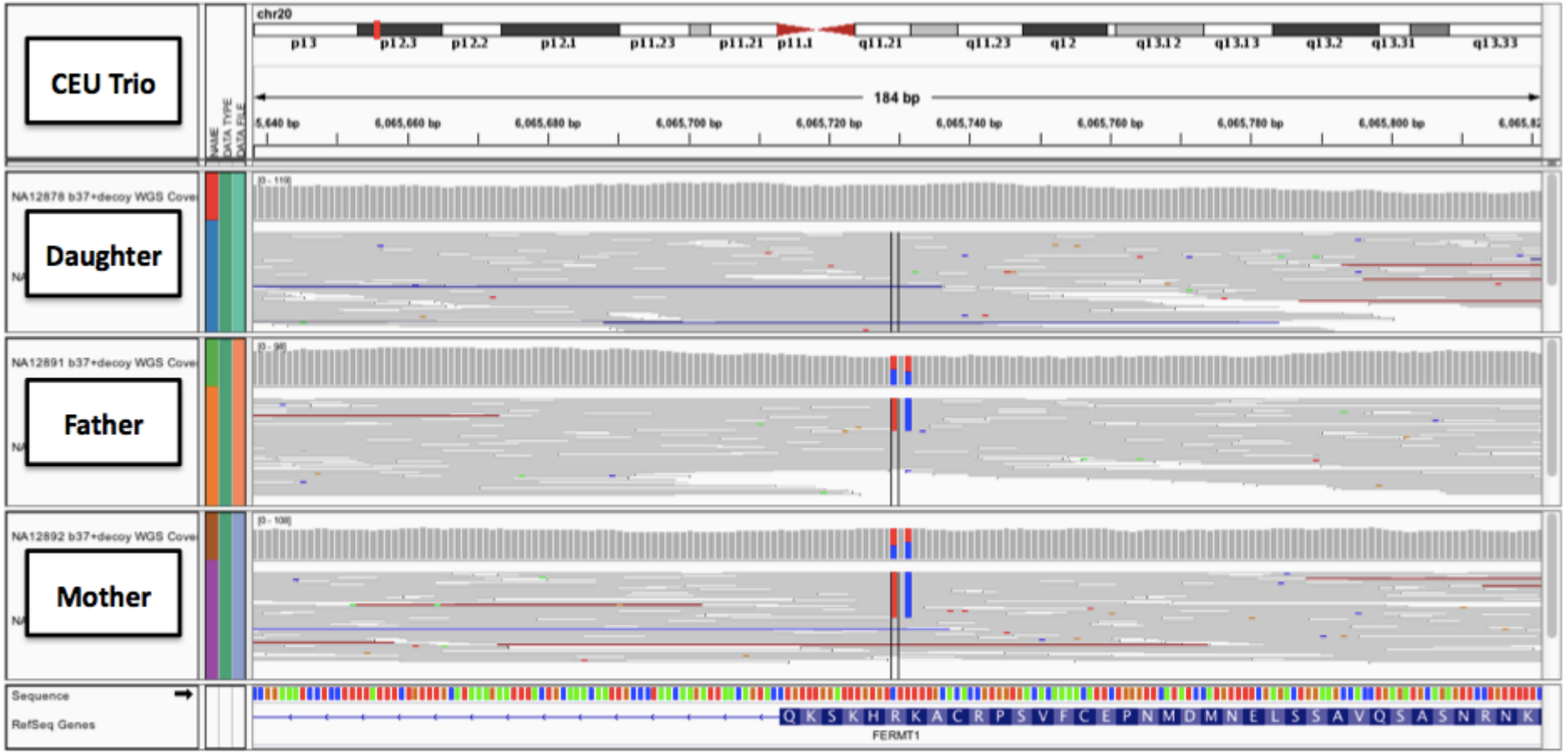
  Call SNPs, indels, and some SVs simultaneously by performing a local *de-novo* assembly

[from GATK manual]

Artificial SNVs caused by indel and recovered by local assembly



Added bonus: phasing (e.g. for distinguishing compound heterozygotes)



[from GATK manual]

Variant Quality Score Recalibration (VQSR)

We know true variants (e.g. validated/common dbSNP polymorphisms) and know parameters of each predicted variant. Can we learn what distinguished true variant from a false positive?

VCF record for an A/G SNP at 22:49582364

```
22 49582364    .    A    G    198.96    .
   AB=0.67;
   AC=3;
   AF=0.50;
   AN=6;
   DP=87;
   Dels=0.00;
   HRun=1;
   MQ=71.31;
   MQ0=22;
   QD=2.29;
   SB=-31.76
   GT:DP:GQ    0/1:12:99.00    0/1:11:89.43    0/1:28:37.78
```

INFO field

AC	No. chromosomes carrying alt allele	AB	Allele balance of ref/alt in hets
AN	Total no. of chromosomes	HRun	Length of longest contiguous homopolymer
AF	Allele frequency	MQ	RMS MAPQ of all reads
DP	Depth of coverage	MQ0	No. of MAPQ 0 reads at locus
QD	QUAL score over depth	SB	Estimated strand bias score

[from GATK manual]

Larger variants?

This way we can detect SNVs and short indels (up to 20-50 bp long)

What about other variants?

SV classes and detection methods

Structural Genome Variations (SVs)
ABCD

Copy-number variants

Deletion
ABD

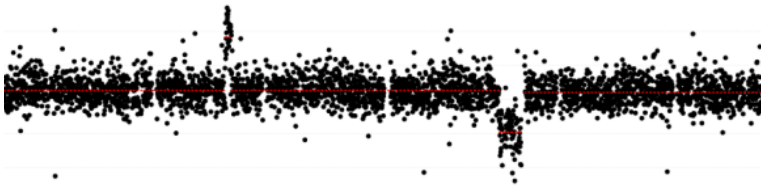
Duplication
ABCCCD

Copy-balanced variants

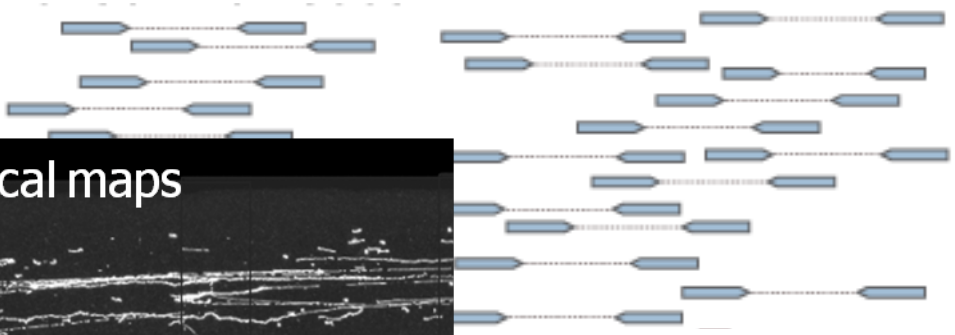
Inversion
ADCB

Translocation
AB **CD**

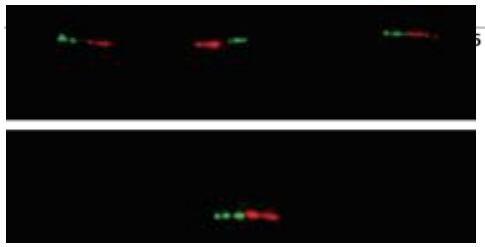
aCGH

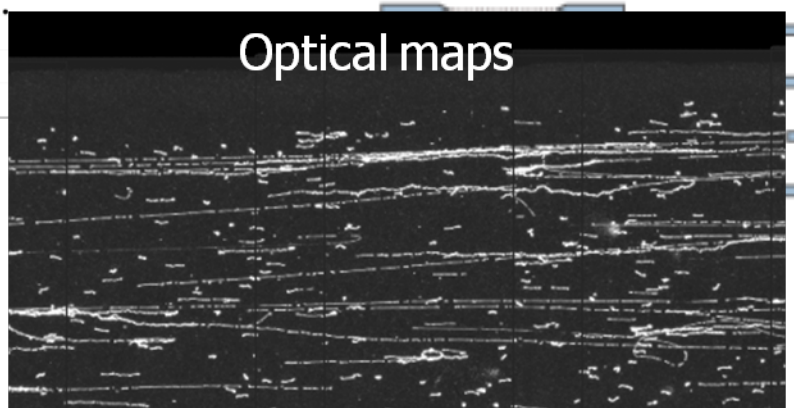


Di-tag fosmid and NGS sequencing



Fibre-FISH





ERIBA

Advantages of paired-end sequencing

1) Twice as many bases per slide !

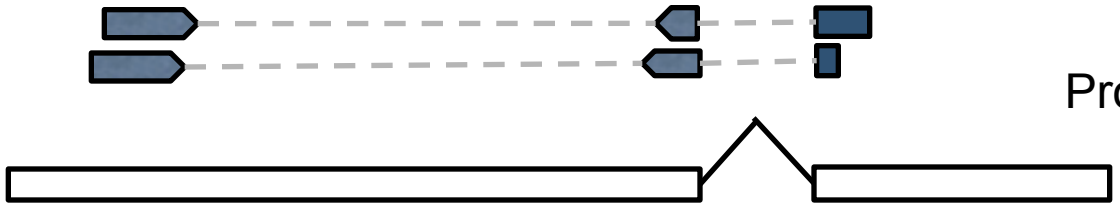
2) Structural information !!!

Molecular haplotyping (phasing)

Genome assembly

Structural variants

Better repeat coverage

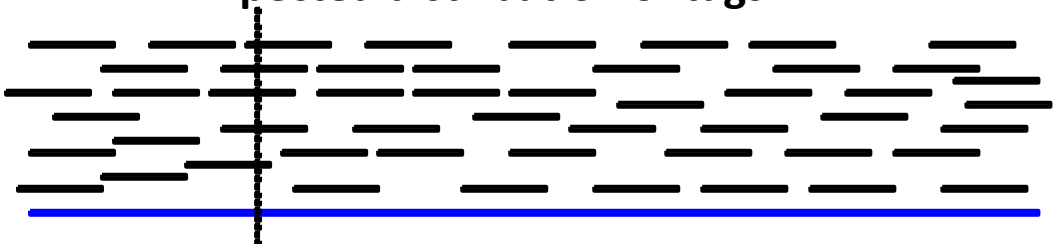


Profiling of transcript isoforms

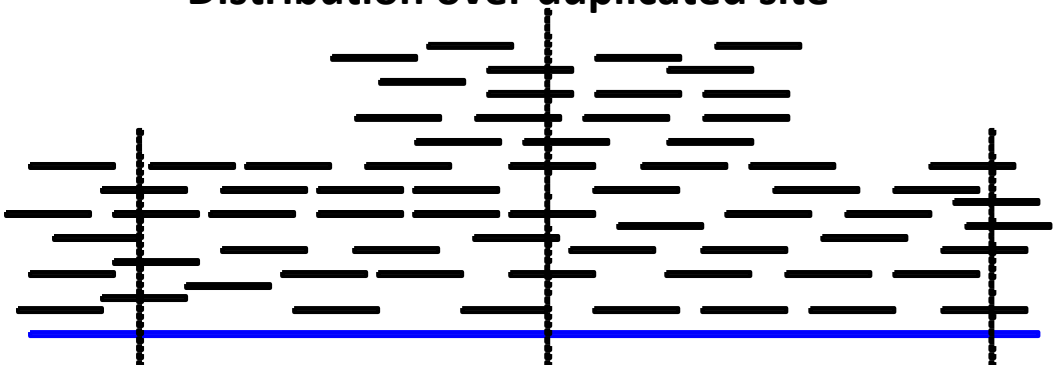
ERIBA

Method 1: Read depth analysis (RD)

Expected distribution of tags



Distribution over duplicated site



Scope:
Copy-number changes

Tool examples:

CNV-Seq (Xie &Tammi 2009)

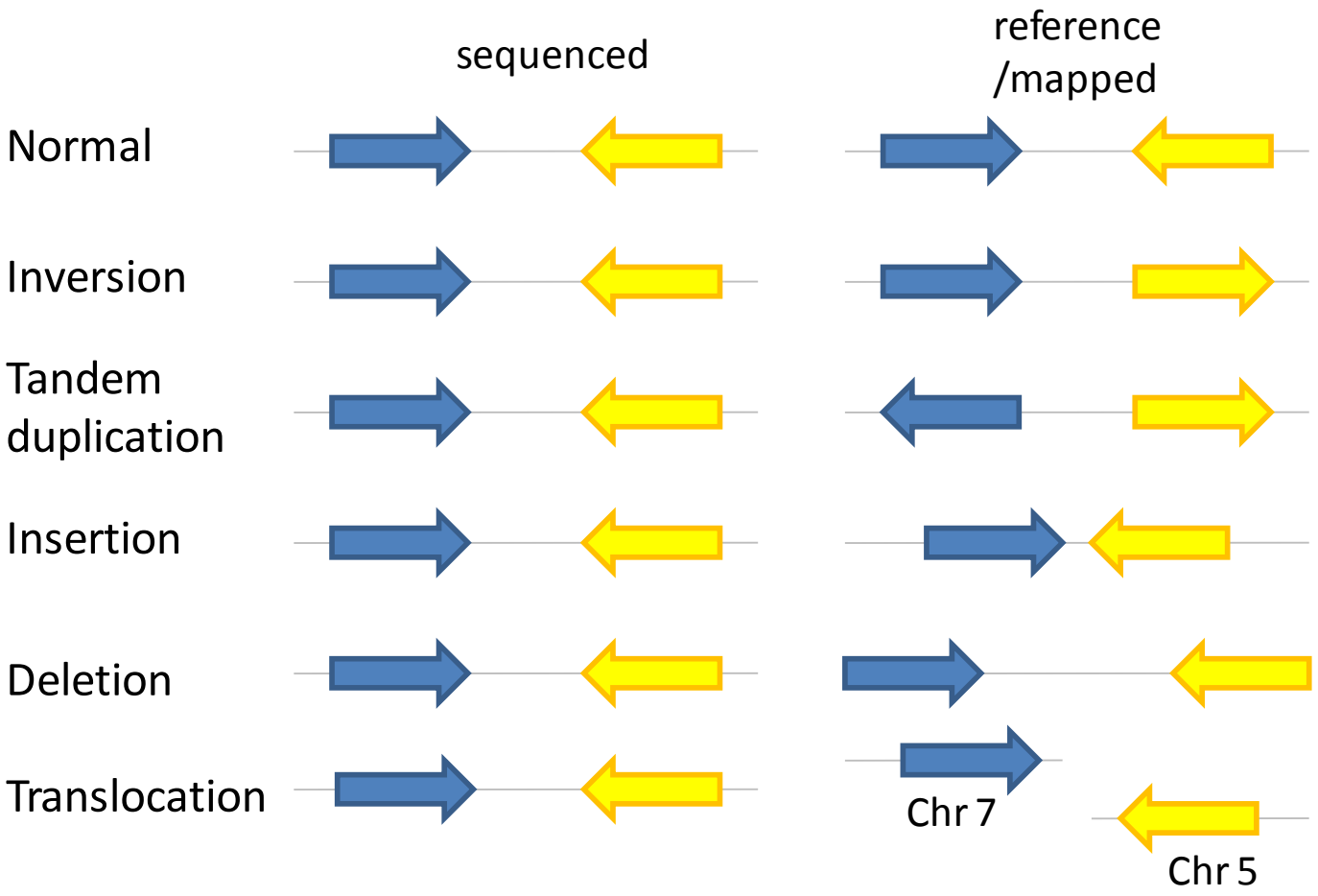
CNVnator (Abyzov et al, 2011)

SegSeq (Chiang et al, 2009)

DWAC-Seq (our tool)

ERIBA

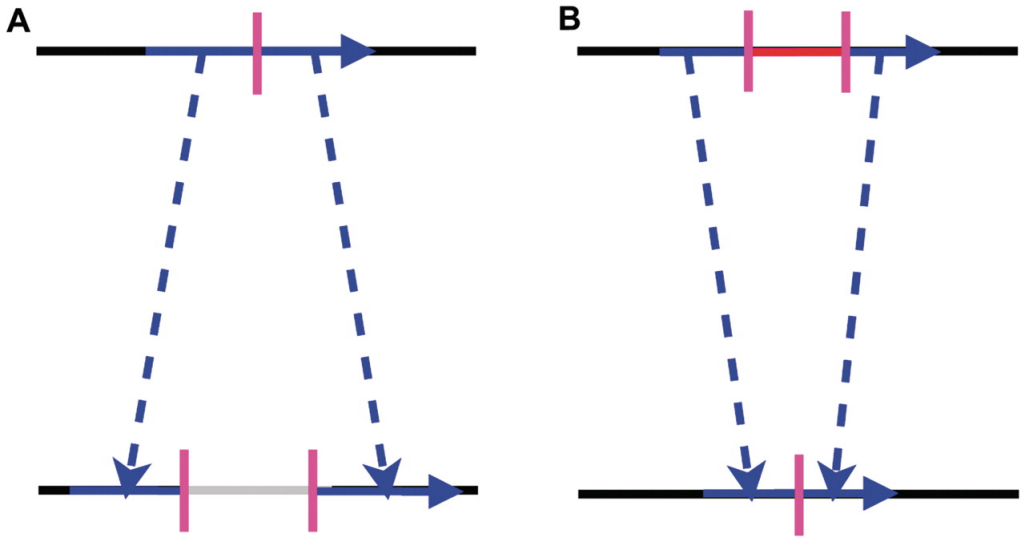
Method 2: Discordant pairs (DP)



Scope: copy-number and copy-neutral SV at resolution close to base-pair

Tool exampless: Breakdancer (Chen et al, 2009); 123SV (our tool)

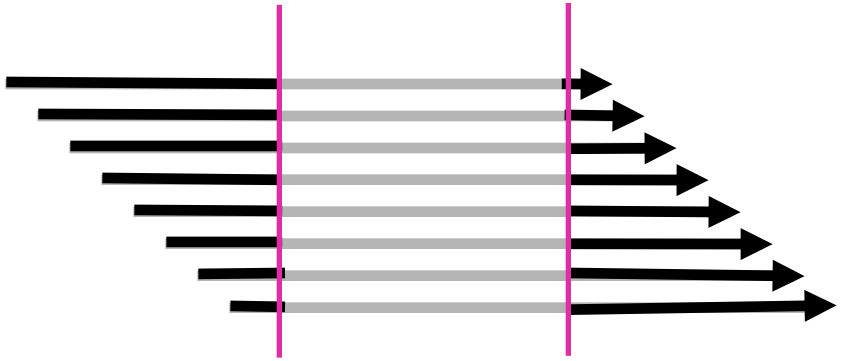
Method 3: Split-read mapping (SR)



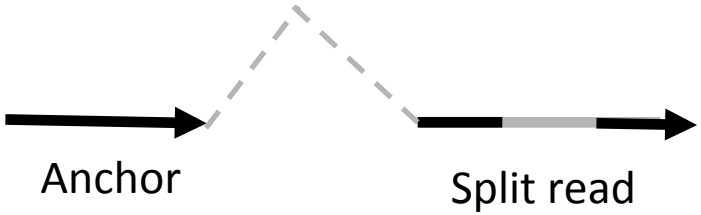
Scope: prediction of copy-number and copy-neutral SV at nucleotide resolution

Tool examples:
Pindel (Ye et al, 2009)
SRiC (Zhang et al, 2011)

Evidence from multiple reads

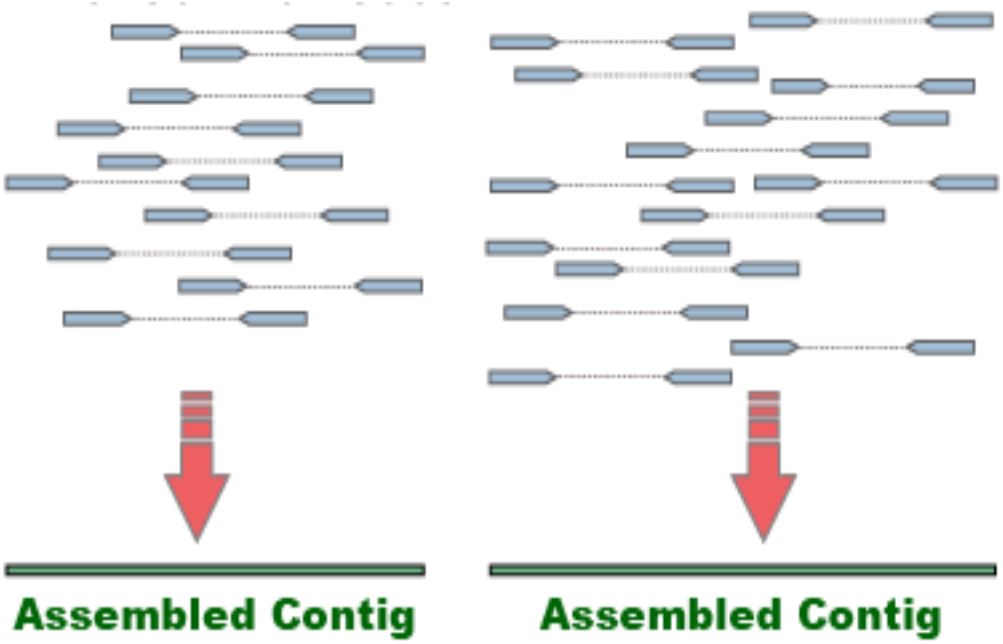


Advantage of **paired reads**



Unmapped reads are good candidates for split-mapping

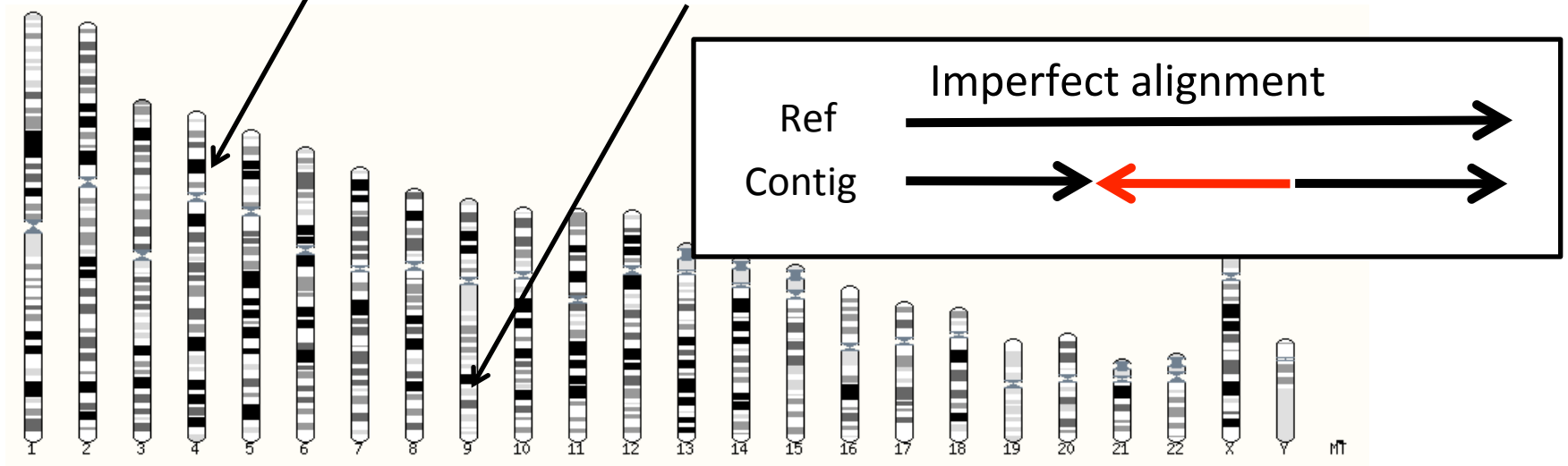
Method 4: Genome assembly (AS)



Scope: various types of SVs including large inserts

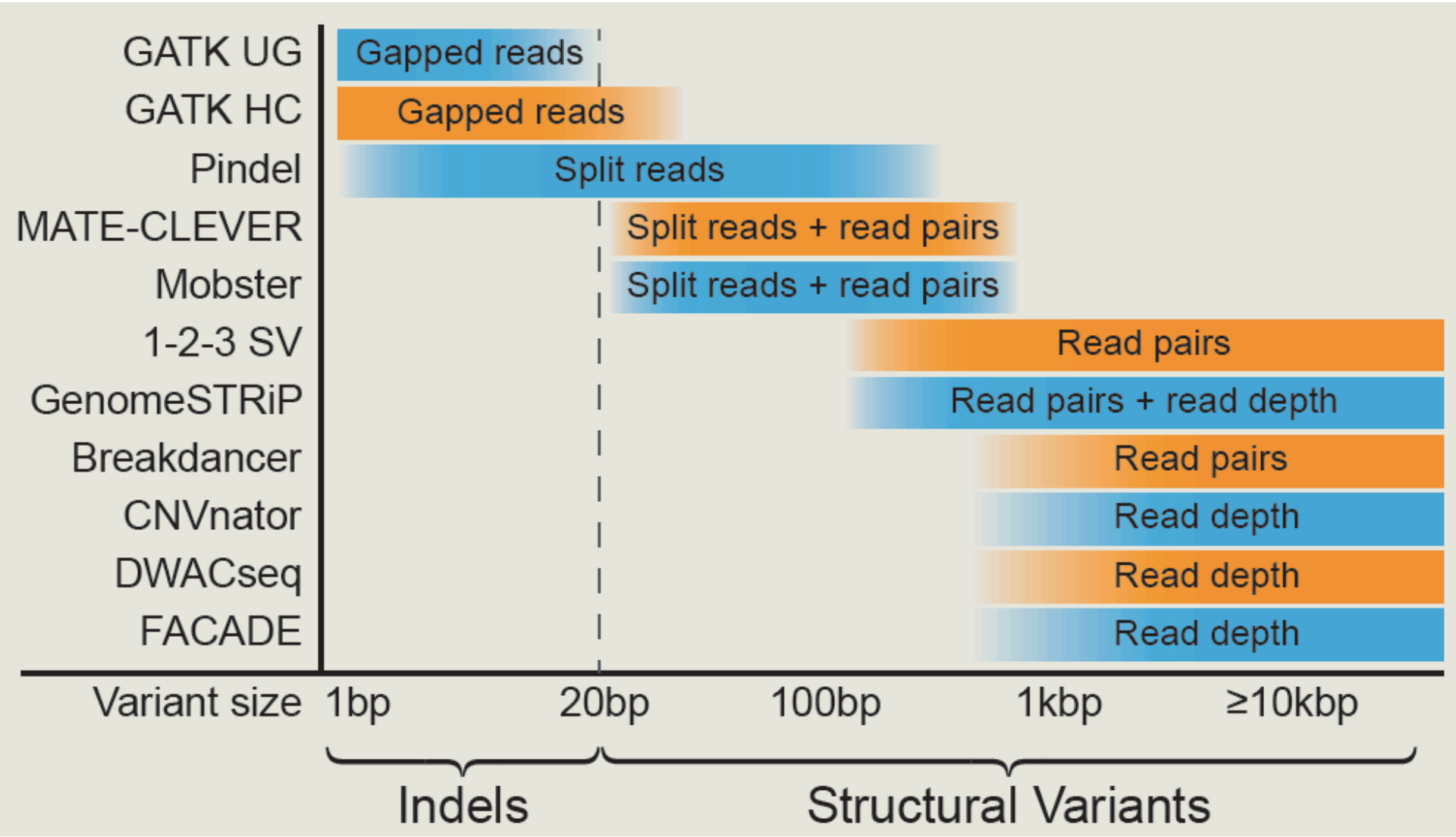
Tool examples: *de novo* assemblers SOAPdenovo, ABYSS, Allpaths-LG

BLAST/BLAT/BWASW search for comparison of contigs and genome reference



ERIBA

GoNL pipeline for SV discovery



ERIBA

Multi-method approaches to SV discovery

PINDEL (http://gmt.genome.wustl.edu/packages/pindel/)
Split-read mapping (very specific for short and mid-size variants)

DELLY (https://github.com/dellytools/delly)
Discordant read and split-read methods

LUMPY-SV (https://github.com/arq5x/lumpy-sv)
Multi-method tool

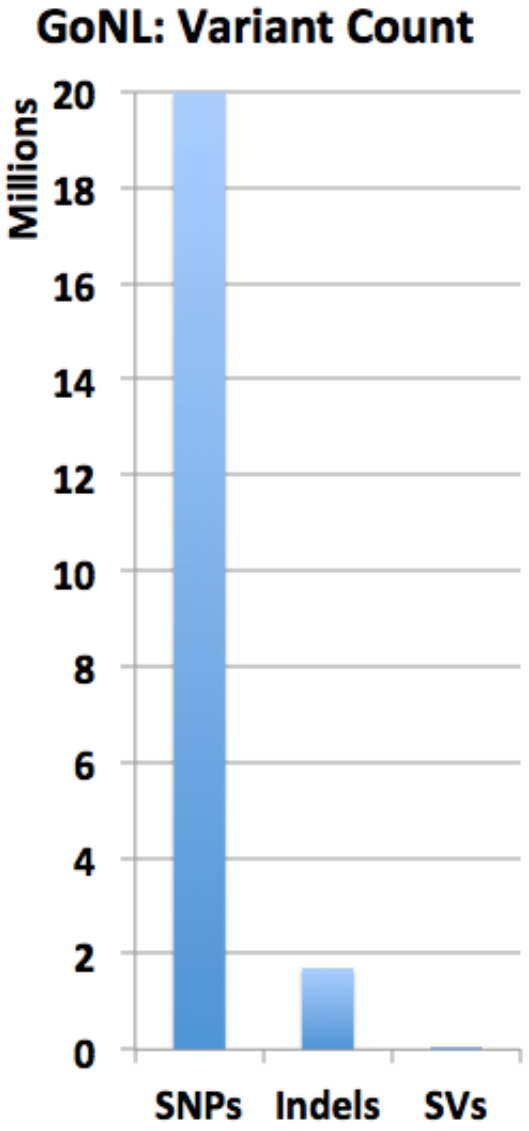
SURVIVOR, MetaSV, Parliament − creating consensus or multi-sample callset

Parliament2 − run multiple tools (Breakdancer, BreakSeq, CNVnator, Delly, Lumpy, Manta) and create consensus callset (using SURVIVOR)
Also available as docker container

ERIBA

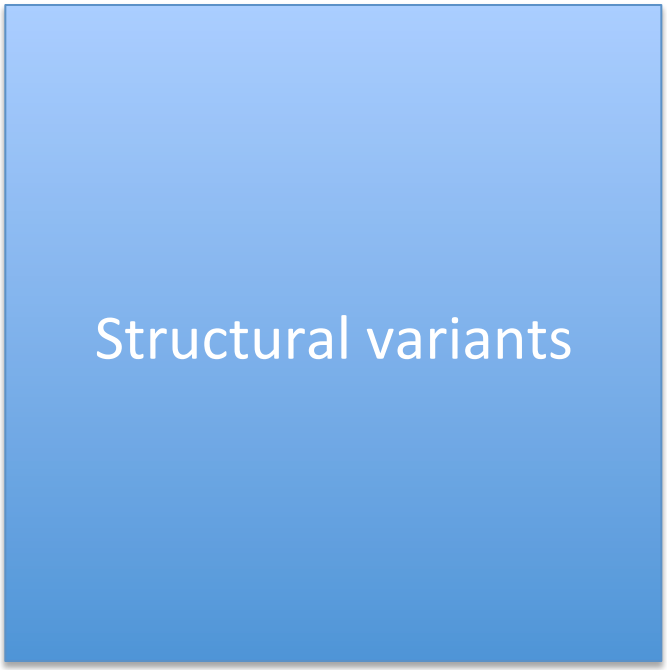
docker

Impact of Structural Variants



GoNL: Bases affected

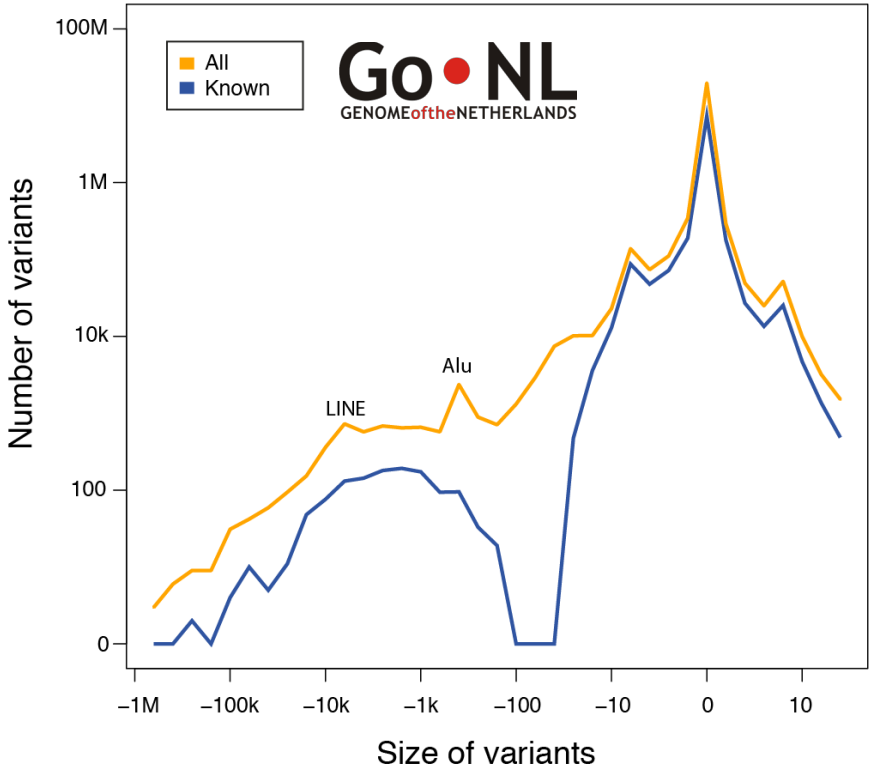
Variant type	Megabases
SNVs	20.4
Indels	4.3
SVs	75.3



ERIBA

Genome sequencing: what do we get?

GoNL variant list	
SNPs	20.4 M
Short indels 1-20 bp	1.7M
Deletions 20-99 bp	31.5k
Deletions 100+ bp	20k
Mobile Element Insertions	13k
Insertions	2,2k
Duplications	1,8k
Inversions	90
Interchromosomal events	60



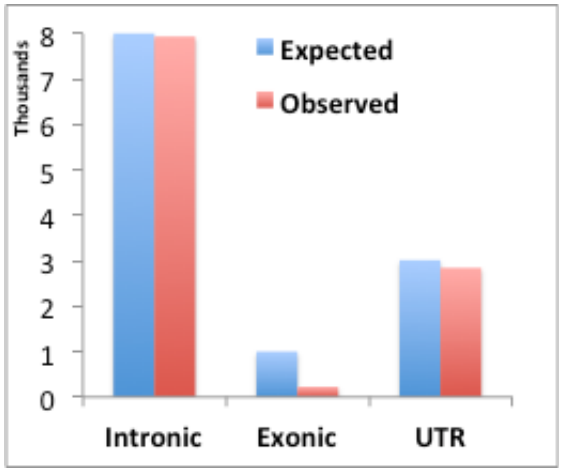
Per individual genome (compared to reference genome)

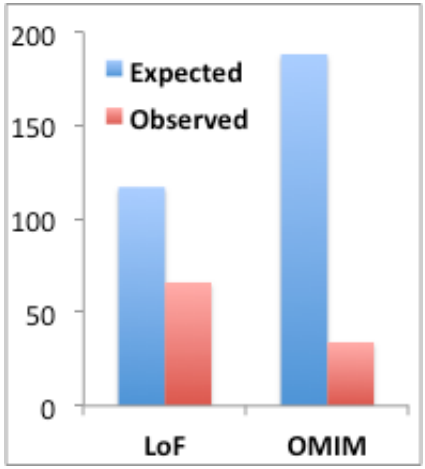
3.7M SNPs

360k short indels (1-20bp)

5.2k medium deletions (20 – 100 bp)

3.3k large deletions (100+ bp)

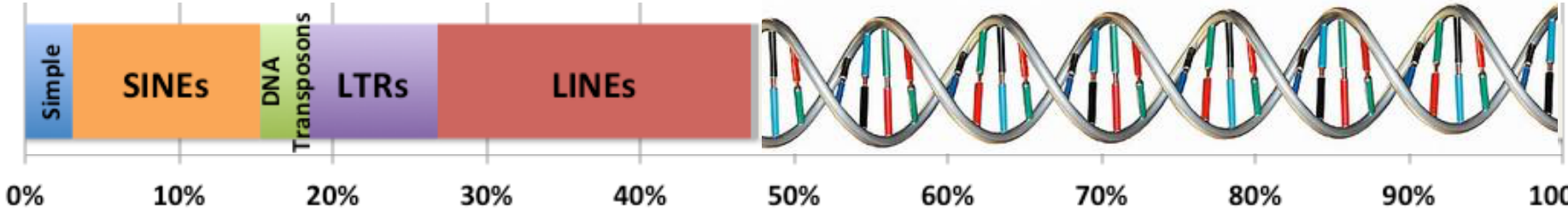


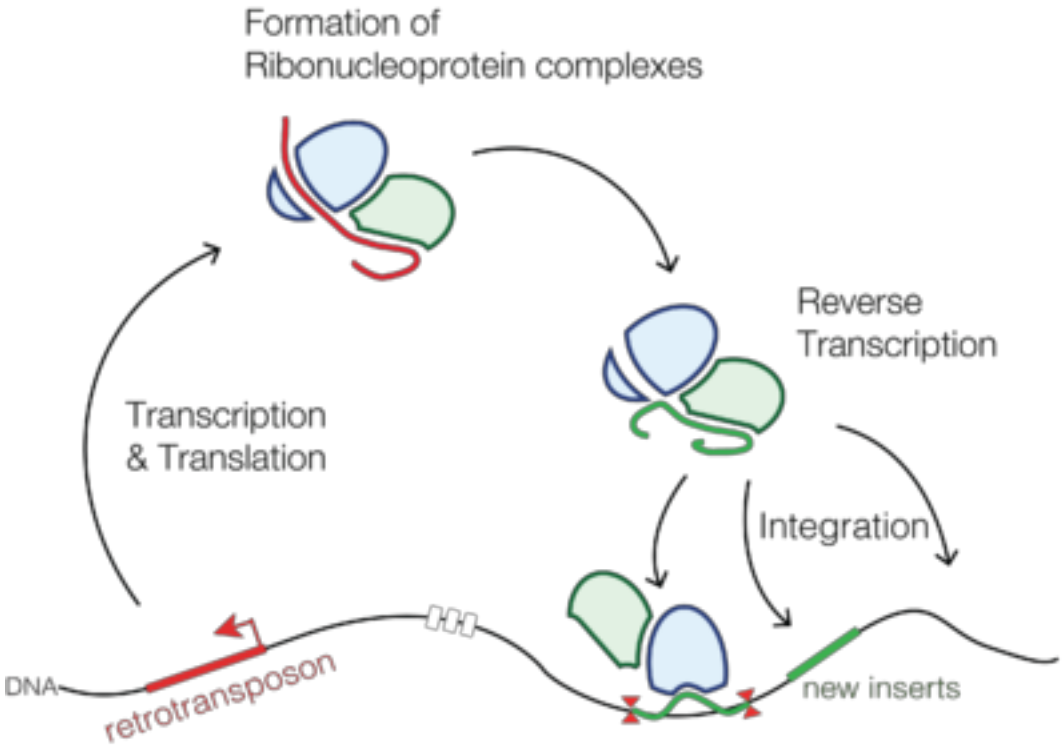


ERIBA

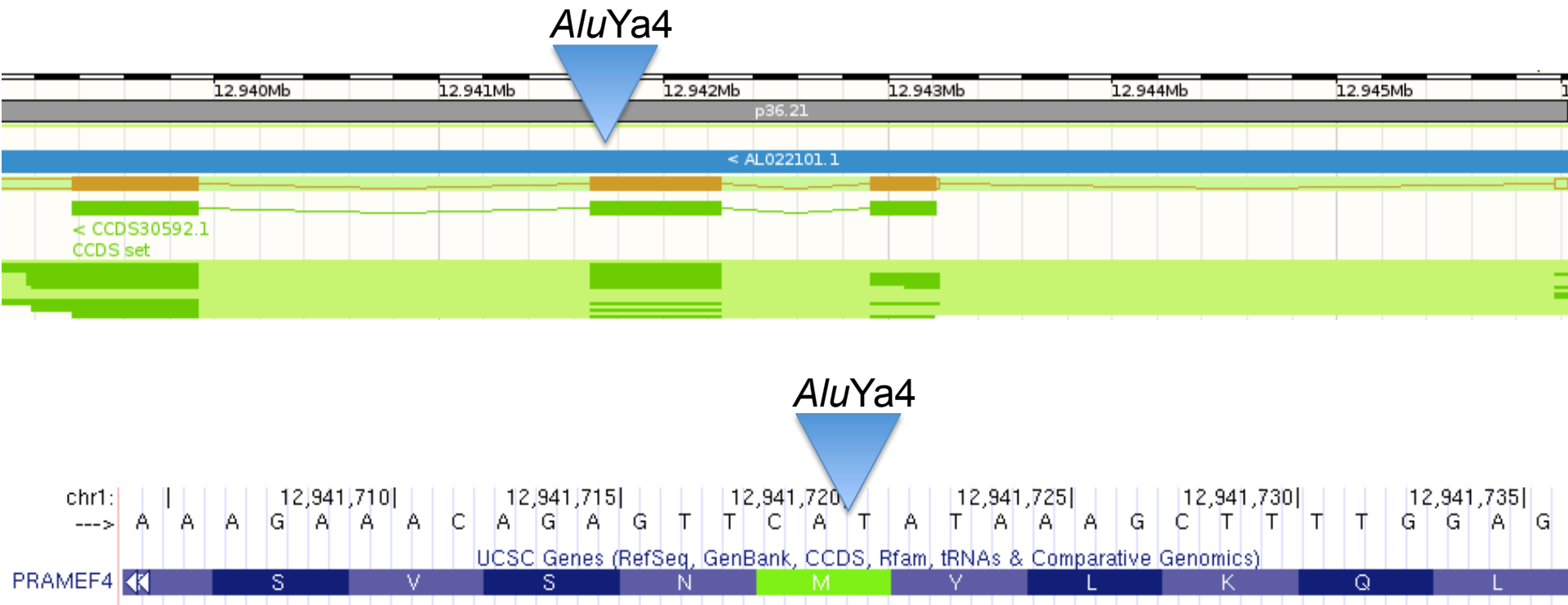
Variable ‘junk’ DNA

Repeat content in human genome





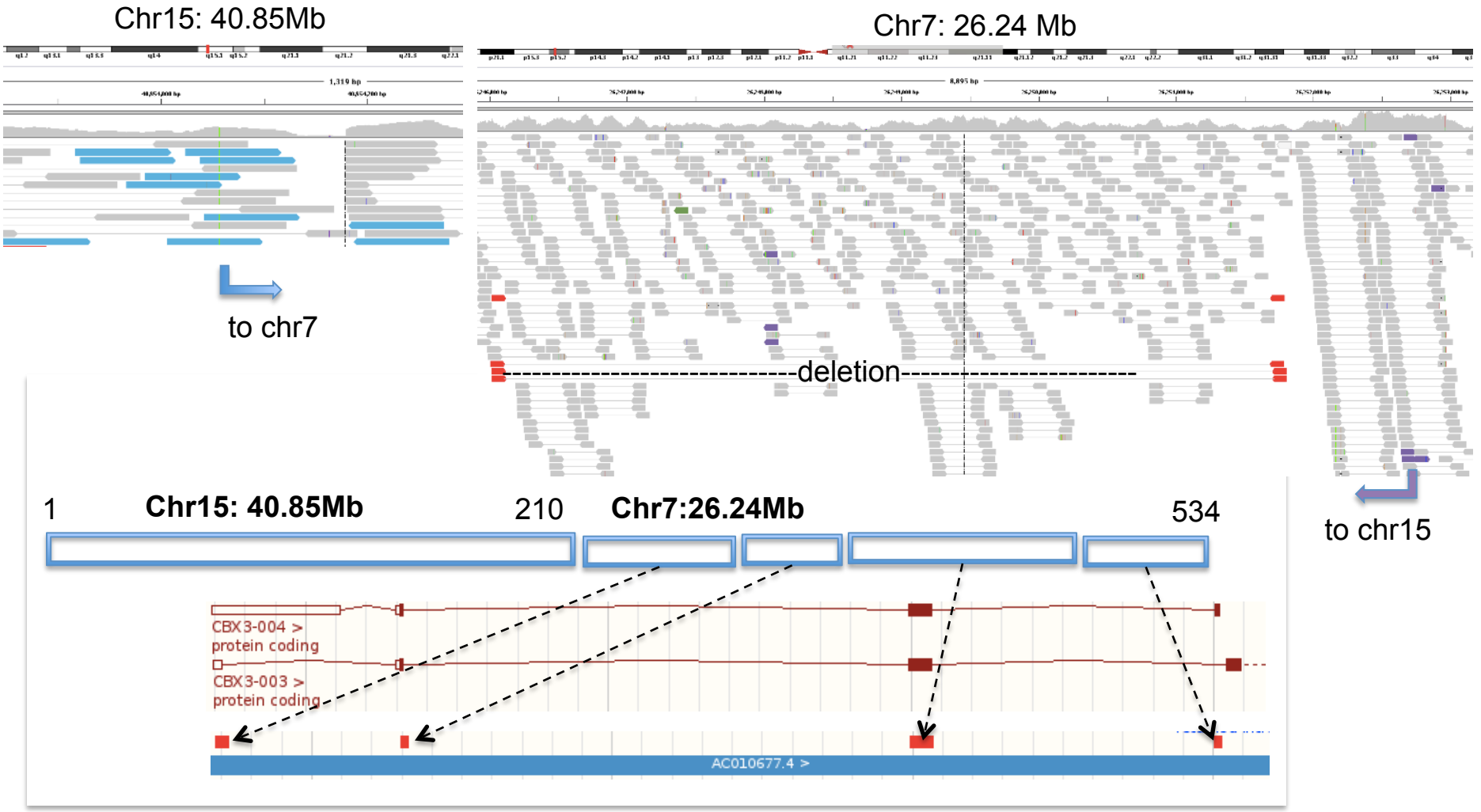
*Alu*Ya4 insertion in PRAMEF4 gene



PRAME Family member 4
In constitutive exon
Observed in 21 samples
Mutations in gene are associated with melanoma

[Hehir-Kwa et al., 2016]

Complex variants: gene retrotransposition insertion polymorphism (GRIP)



Mechanism: (retro)transposition

[Hehir-Kwa et al., 2016]

Prevalence: GoNL about 40 cases

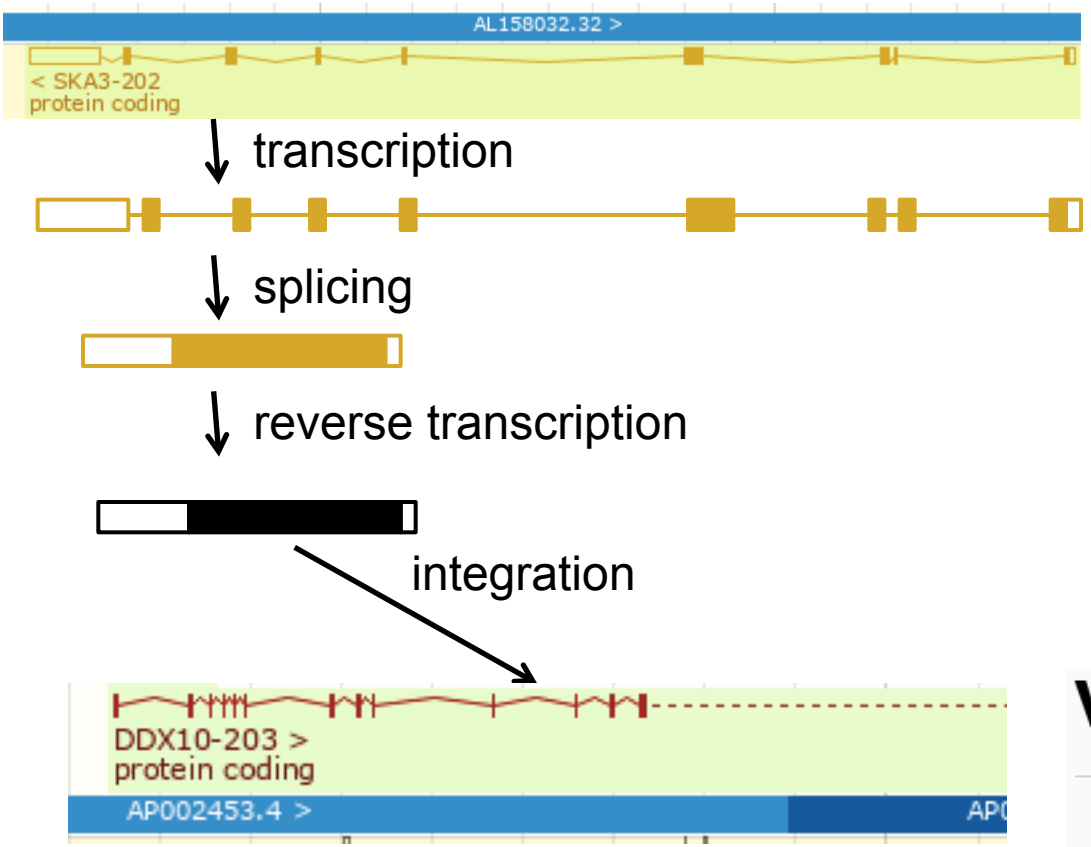
Tools: Discordant pairs (1-2-3-SV)

“Knock-outs” in our genome

SKA3 Gene (Protein Coding) ★

Spindle And Kinetochore Associated Complex Subunit 3

Chr13



Chr11

Variant: 13-21746600-TG-T

	Exomes	Genomes	Total
Filter	RF	RF	
Allele Count	20346	954	21300
Allele Number	199986	26322	226308
Allele Frequency	0.1017	0.03624	0.09412

frameshift

- SKA3
  - ENST00000314759 * p.Gln70LysfsTer7 LoF: High-confidence Flag: PHYLOCSF_WEAK
  - ENST00000400018 p.Gln70LysfsTer7 LoF: High-confidence Flag: PHYLOCSF_WEAK

stop gained

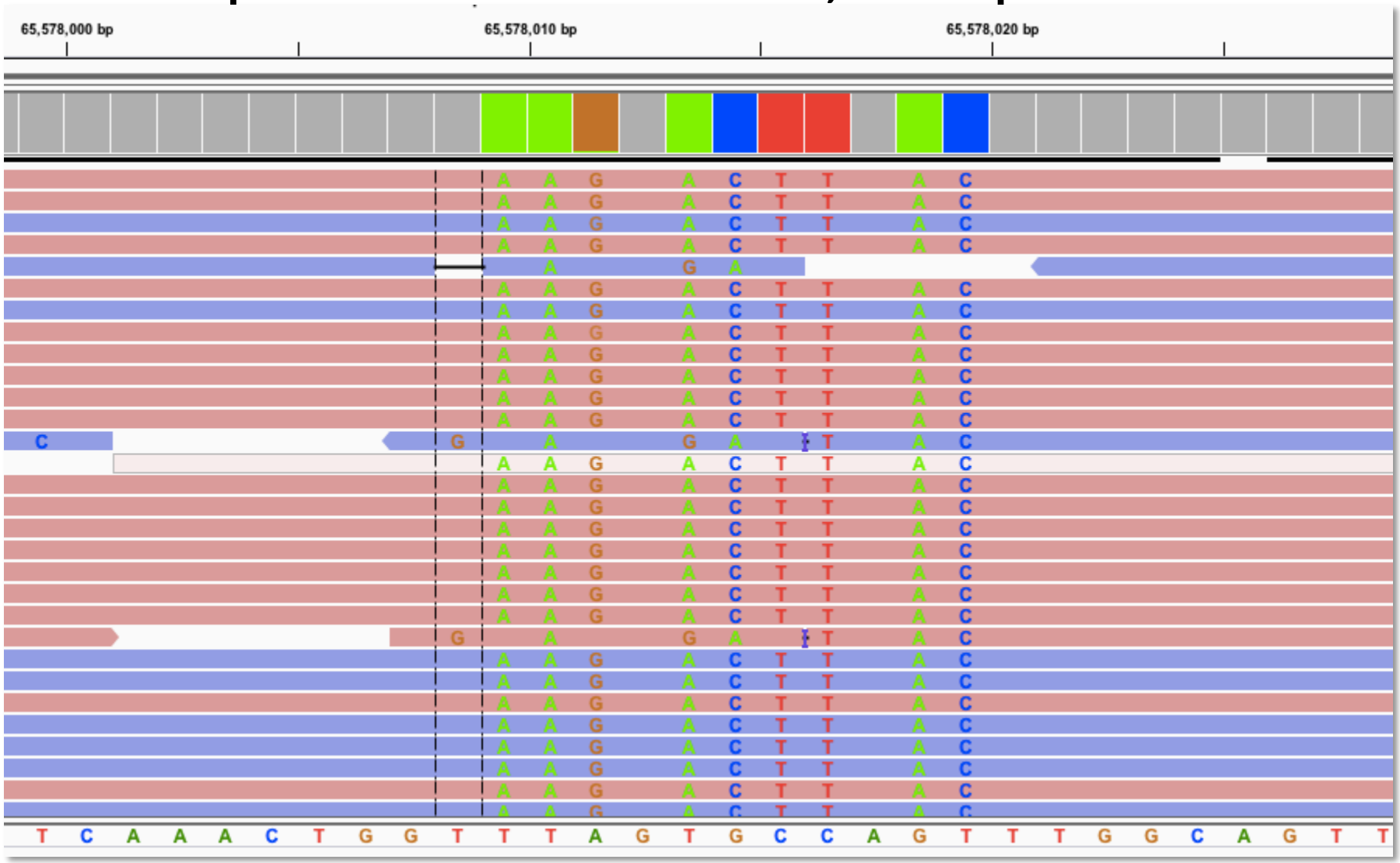
- SKA3
  - ENST00000314759 * p.Arg27Ter LoF: High-confidence
  - ENST00000400018 p.Arg27Ter LoF: High-confidence

Variant: 13-21750538-G-A

	Exomes	Genomes	Total
Filter	RF	RF	
Allele Count	13679	1315	14994
Allele Number	155602	27232	182834
Allele Frequency	0.08791	0.04829	0.08201

ERIBA

Complex variants: MNPs, complex indels



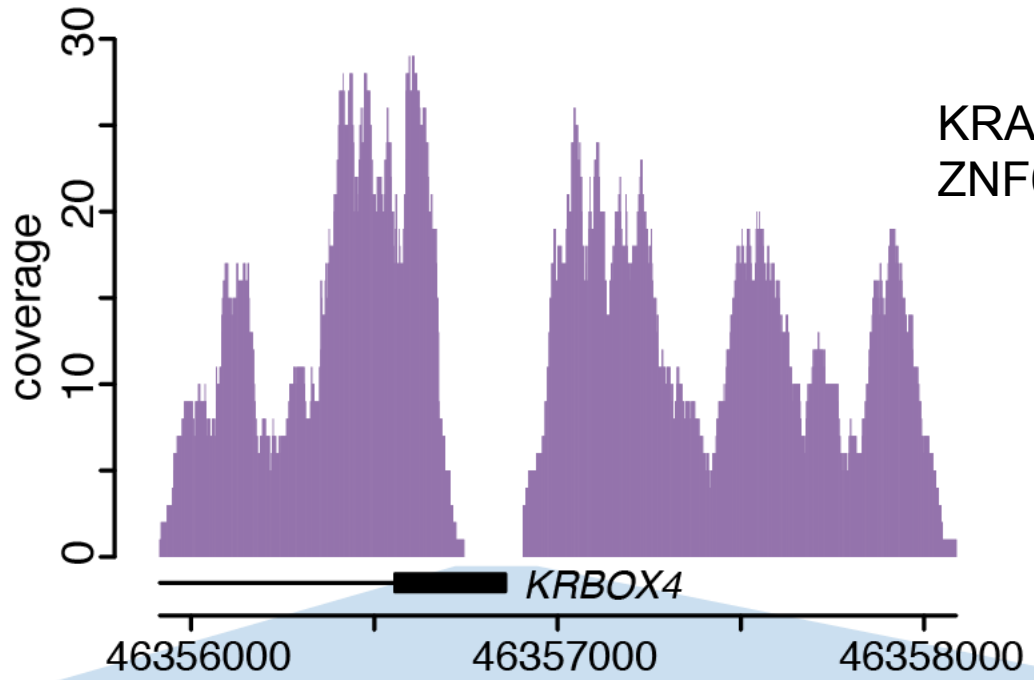
Mechanism: polymerase errors **Tool example**: GATK Haplotype Caller

Prevalence: ~3% of all indels are non-simple

[Hehir-Kwa et al., 2016]

ERIBA

Complex variants



KRAB box domain containing 4, aka ZNF673, transcription regulator

```
...ATCTATCCCAAG...                          ...TGATCTTGCTGT...  reference
...ATCTATCCCAAG|CCATTTACCAAT...   ...AGAAAGAAGGGT|TGATCTTGCTGT...  clipped read
              CCATTTACCAAT...     ...AGAAAGAAGGGT               replacing sequence
```

chr1:199325670

```
CTTTAAAAAGAAGTAGTATGGTATATAATCACAAGA
CTTTAAAAAGAAGAAGTATAATCACAAGAAATCCTG
```

chr2:228554844

```
AATAACAGGAGTAAGTCCTTACTTATCAATCTAAAC
AATAACAGGAGTTTAGTTTATCAATCTAAACTCTTC
```

chr3:31723929

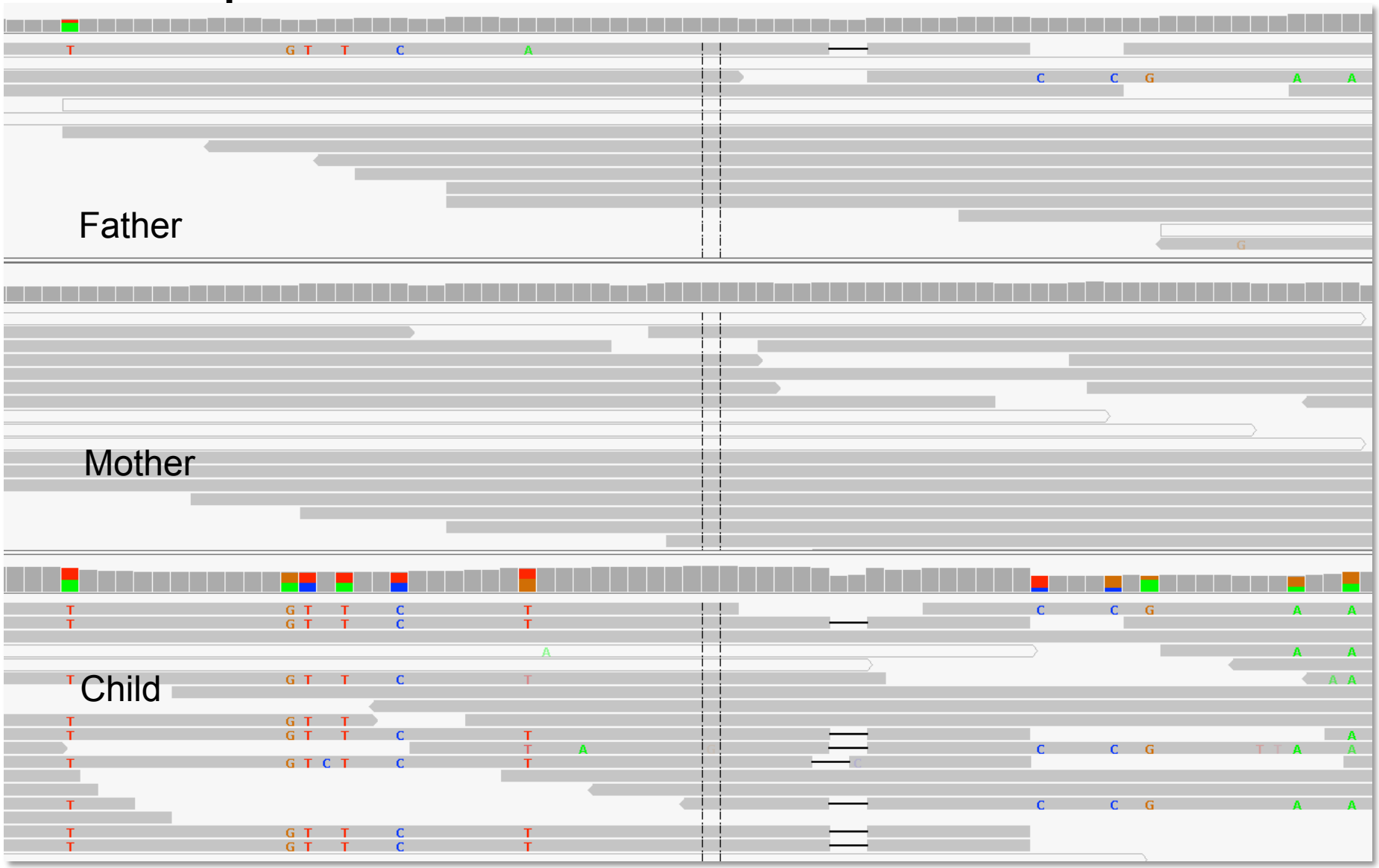
```
TCCAGCCACATGGGCTTCCTGTCTCCATCCCCGTGC
TCCAGCCACATGGGCTGGAGCCCATGTCTCCATCCC
```

chr3:108389073

```
ATATTCACAGAGGTTTACAGTCATCACCACTATCTA
ATATTCACAGAGGTTTACAGTGAATAAACTATCTAT
```

[Hehir-Kwa et al., 2016]

Complex variants: Non-allelic conversion



Mechanism: gene conversion

Tool example: assembly, discordant pairs

Prevalence: currently only several cases

ERIBA

“Dark matter” DNA

Onvindbare genen bezorgen de wetenschap kopzorgen

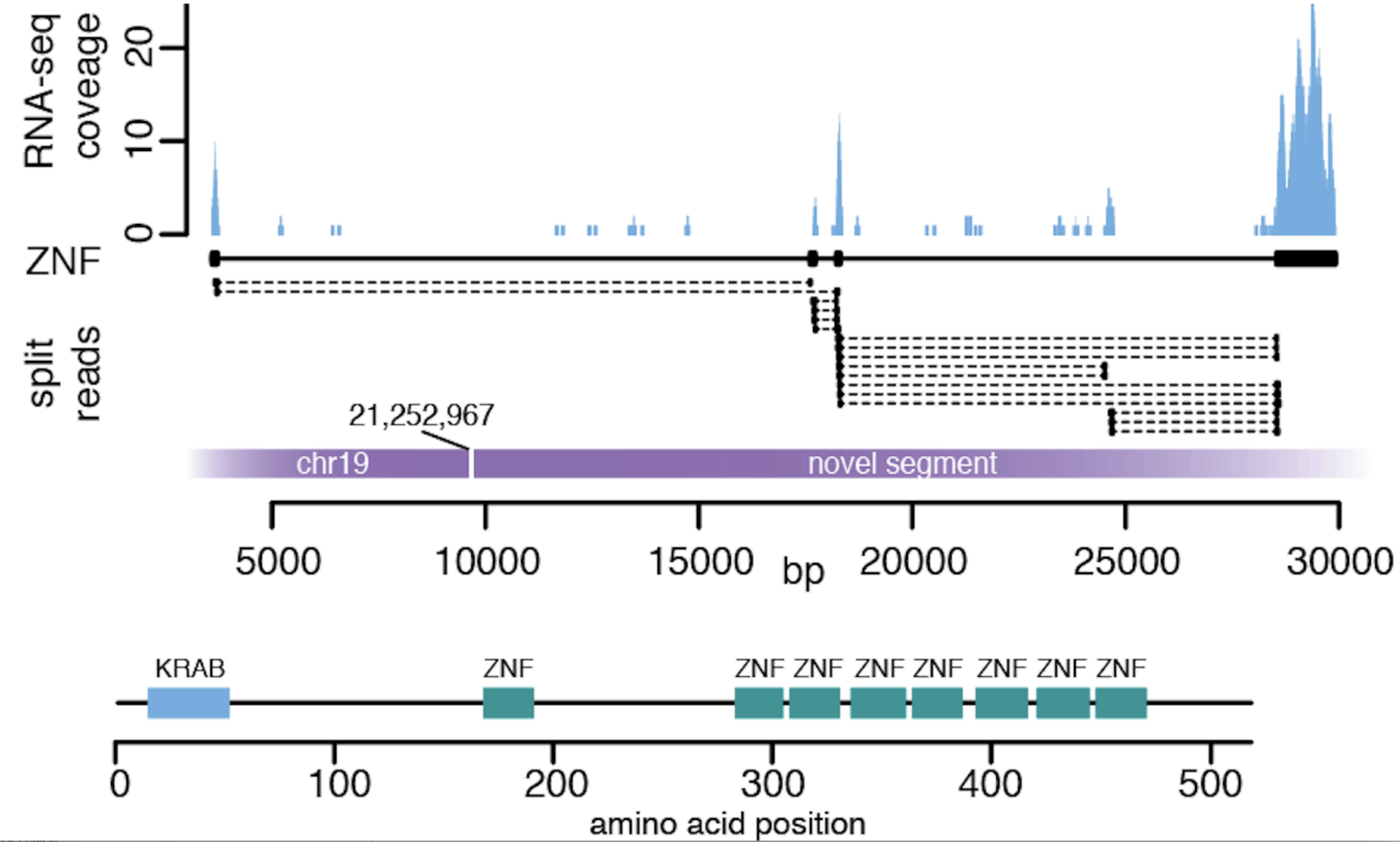
KIJK-REDACTIE - 12 FEBRUARI 2019

Steeds vaker stuiten wetenschappers op gaten in het genetisch materiaal van mensen en dieren. Net als de ruimte lijkt DNA ‘dark matter’ te bevatten.

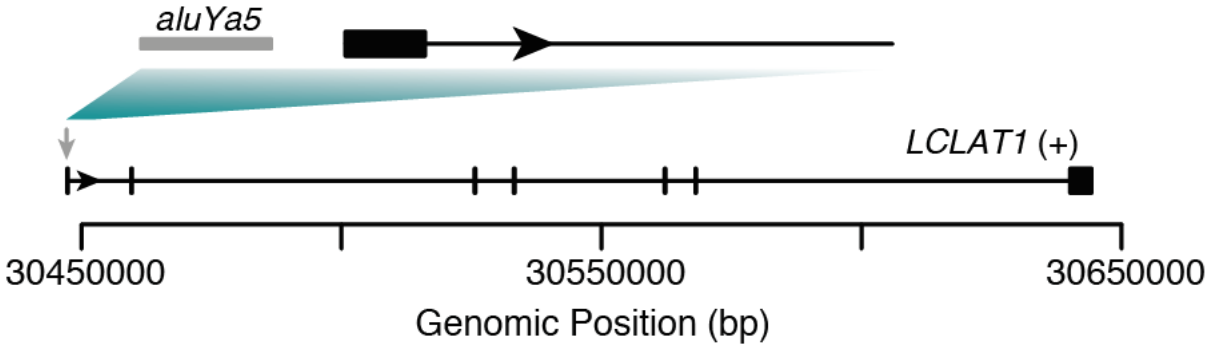
New genomic segments

Set	Characteristcs	Segments	Size, bp	Remark
(Nearly-) Fixed	Seen in >95% of individuals	1,517	2,378,724	
Common	Seen in 5-95% of individuals	7,236	4,101,713	
Rare	Seen in <5% of individuals	2,250	714,279	
Male-specific	Segment is seen in males, but not in females	222	211,118	Parts of Y chromosome
Herpesvirus	Matches to HHV6a, HHV6b	112	163,383	CIHHV-6A (1 mother + 1 kid) CIHHV-6B (1 mother + 1 kid)
Bacterial	Matches to Bacterial DNA	3,439	13,067,381	Contamination, Samples: 12, Different titer, Lanes: 36/36
Plant-like1	Matches to plant DNA	159	165,371	Contamination, Samples: 4, Lanes: 6/15
Rat-like	Rat mtDNA, Alpha-, Getah-virus	23	13,520	Contamination, Sample: 1, Lanes: 1/3
Mouse-like	Mouse sequences	38	14,821	Contamination, Samples: multiple, Lanes: multiple
Plant-like2	Matches to plant DNA	89	28,394	Contamination. Sample:1, Lanes: 1/3
TOTAL		**15k**	**20.86 Mb**	

New segments



Change in expression level



Phenotype, disease and trait annotations associated with variants in this gene

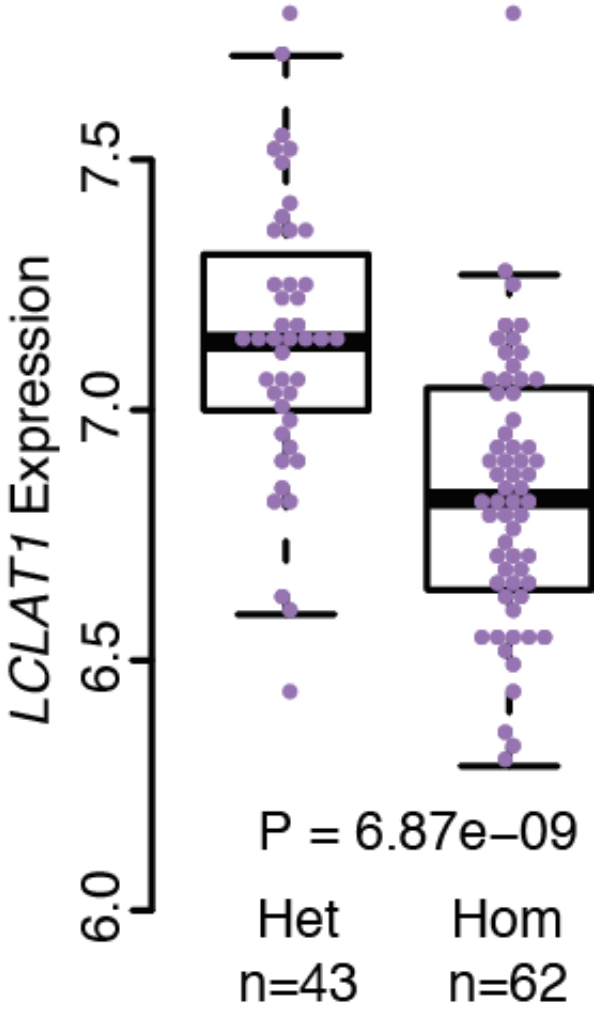
Filter

Phenotype, disease and trait	Source(s)	Genomic locations	Biomart	Number of variants	Show/hide details
ALL variants with a phenotype annotation	-			4	Show
Albumins	dbGaP	View on Karyotype	-	1	Show
Alcoholism	dbGaP	View on Karyotype	-	2	Show
Heart failure	dbGaP	View on Karyotype	-	1	Show

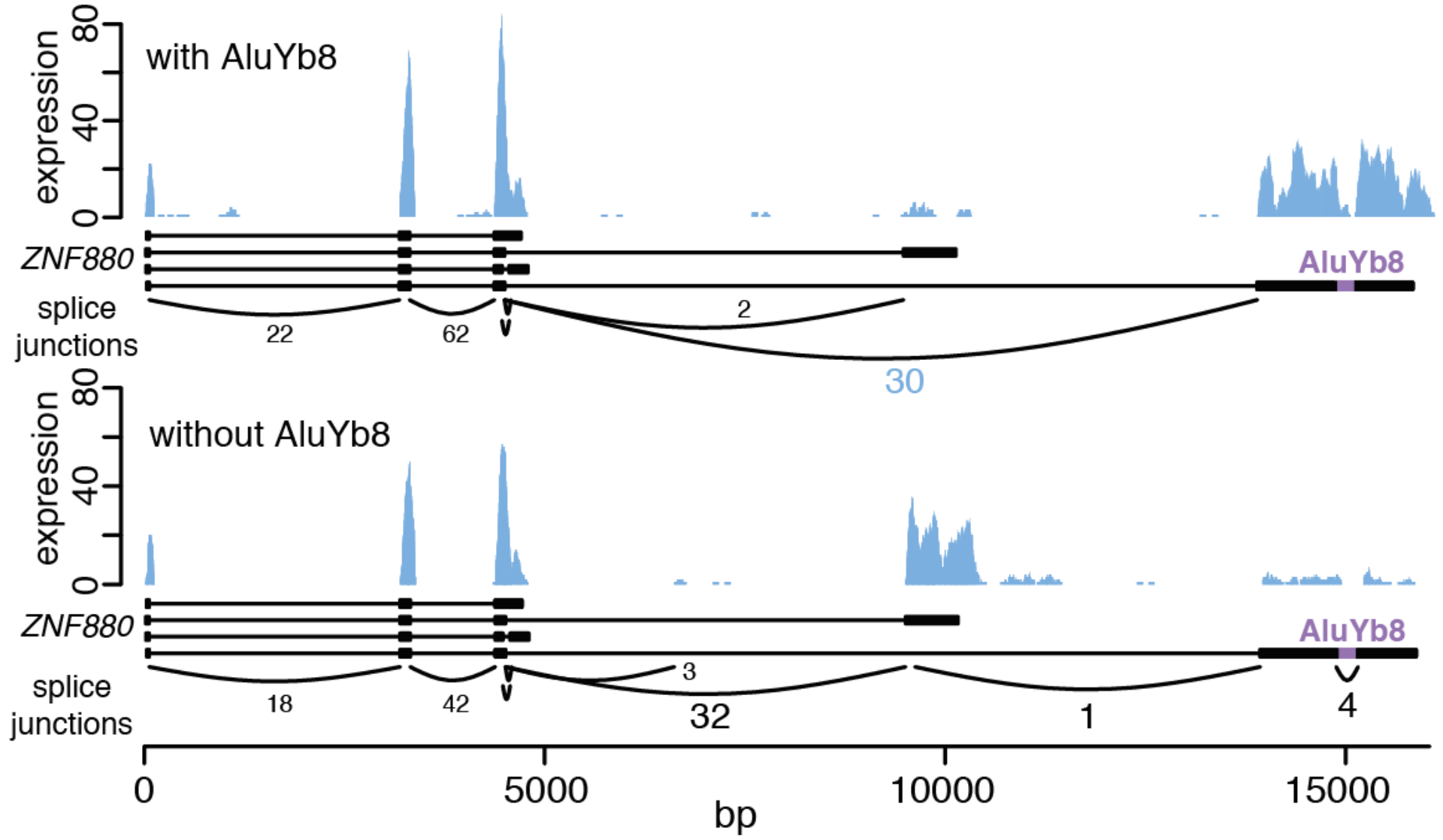
Phenotype, disease and trait annotations associated orthologues of this gene in other species

Filter

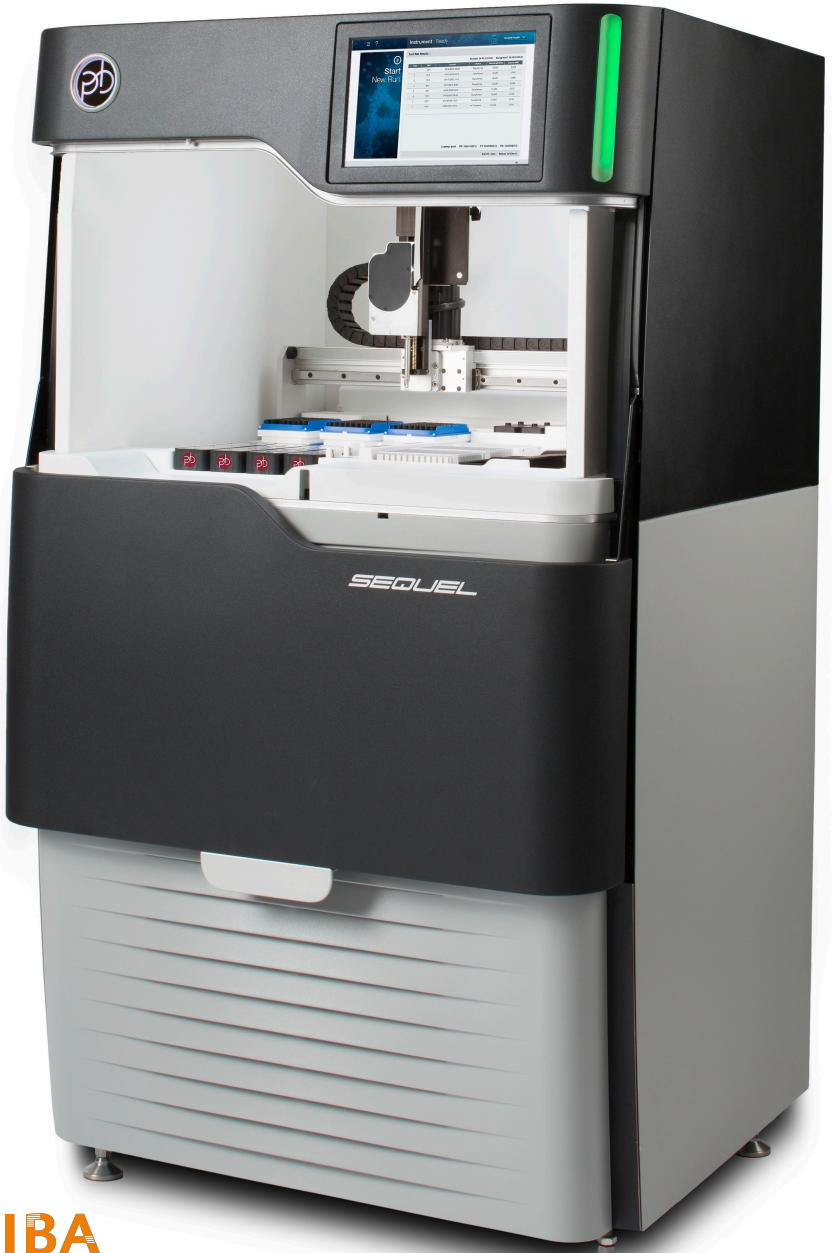
Phenotype, disease and trait	Source	Species	Gene
blood vessel development, disrupted	ZFIN	Zebrafish (*Danio rerio*)	ENSDARG00000103320 lclat1
blood vasculature, decreased amount	ZFIN	Zebrafish (*Danio rerio*)	ENSDARG00000103320 lclat1
blood vessel endothelial cell, decreased amount	ZFIN	Zebrafish (*Danio rerio*)	ENSDARG00000103320 lclat1
blood cell, decreased amount	ZFIN	Zebrafish (*Danio rerio*)	ENSDARG00000103320 lclat1
endothelial cell development, disrupted	ZFIN	Zebrafish (*Danio rerio*)	ENSDARG00000103320 lclat1
endocardium, aplastic	ZFIN	Zebrafish (*Danio rerio*)	ENSDARG00000103320 lclat1



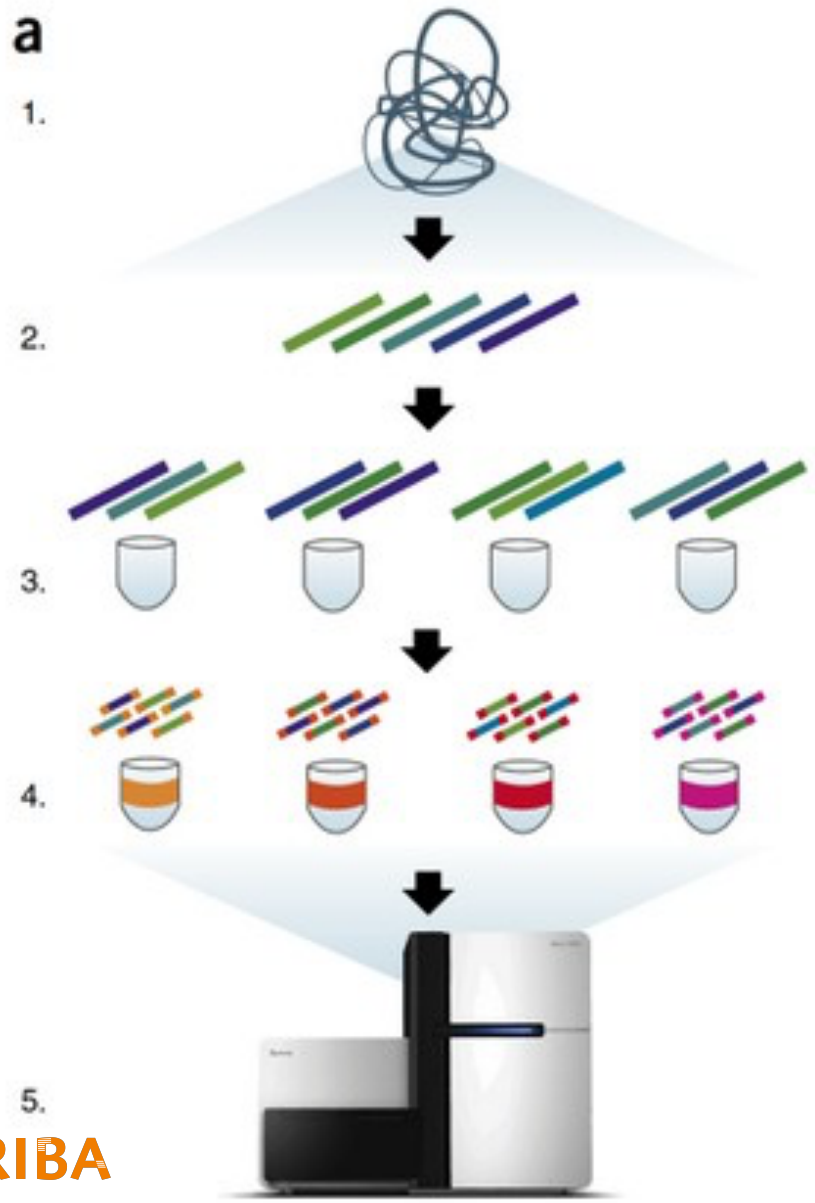
Change in transcript structure

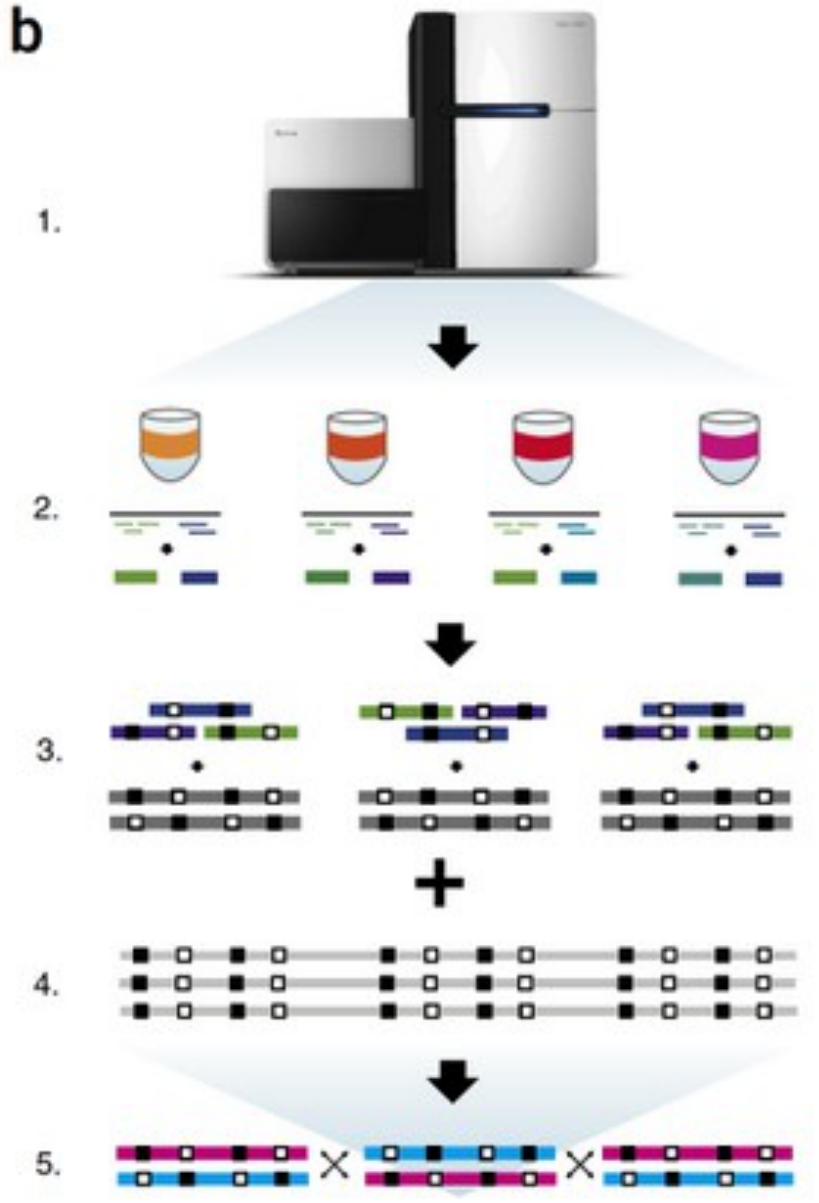


PacBio and Oxford Nano: true long reads



Moleculo, 10xGenomics: synthetic long reads





ERIBA

Take home message: importance of SVs

Variant type	Human Vs Chimp	Common Variants AF > 5%	Rare variants	Individual/ family-specific	*De novo* Variants (avg per kid)	Somatic, ageing-related
Single Base Changes	1.23% of genome	5.948 Mb	6.625 Mb	6,989 Mb	45 bp	?
Structural	3% of genome	10.916 Mb	28.507 Mb	43,317 Mb	4,084 bp	?
SNV:CNV ratio	**1 : 2**	**1 : 2**	**1 : 4**	**1 : 6**	**1 : 91**	**1 : ?**

[Chimp genome consortium, 2005] [Hehir-Kwa, ⋯ Guryev, 2016] [Kloosterman, ⋯ Guryev, 2015]

ERIBA

Acknowledgements

GoNL SV Team

Victor Guryev	*UMCG*
Wigard Kloosterman	*UMCU*
Laurent C. Francioli	*UMCU*
Jayne Y. Hehir-Kwa	*UMCN*
Djie Tjwan Thung	*UMCN*
Tobias Marschall	*CWI/MPI*
Alexander Schoenhuth	*CWI*
Matthijs Moed	*LUMC*
Eric-Wubbo Lameijer	*LUMC*
Abdel Abdellaoui	*VU*
Slavik Koval	*EMC/LUMC*
Joep de Ligt	*UMCN*
Najaf Amin	*EMC*
Freerk van Dijk	*UMCG*
Lennart Karssen	*EM/Polyomica*
Leon Mei	*LUMC*
Kai Ye	*LUMC/WASHU*

GoNL steering committee

Paul de Bakker	*UMCU*
Dorret Boomsma	*VU*
Cornelia van Duin	*EMC*
Gert-Jan van Ommen	*LUMC*
Eline Slagboom	*LUMC*
Morris Swertz	*UMCG*
Cisca Wijmenga	*UMCG*

University of Washington
Fereydoun Hormozdiari
Evan E. Eichler

BGI Shenzen
Jun Wang

ERIBA, RuG, UMC Groningen
Diana Spierings
Marianna Bevova
Rene Wardenaar
Tristan de Jong
Peter Lansdorp

university of groningen

University Medical Center Groningen

Title

ERIBA