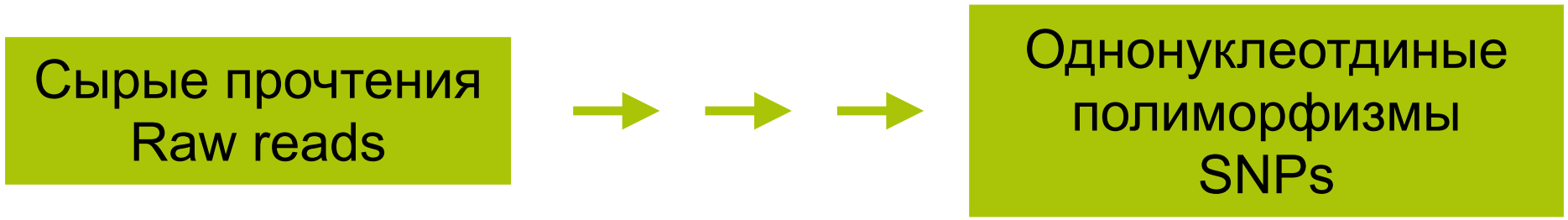
From raw reads to SNP

От сырых ридов к однонуклеотидным полиморфизмам

Skoltech

Алексей Замалутдинов, аспирант Центра Агротехнологий, Сколтех

Our pipeline



Skoltech

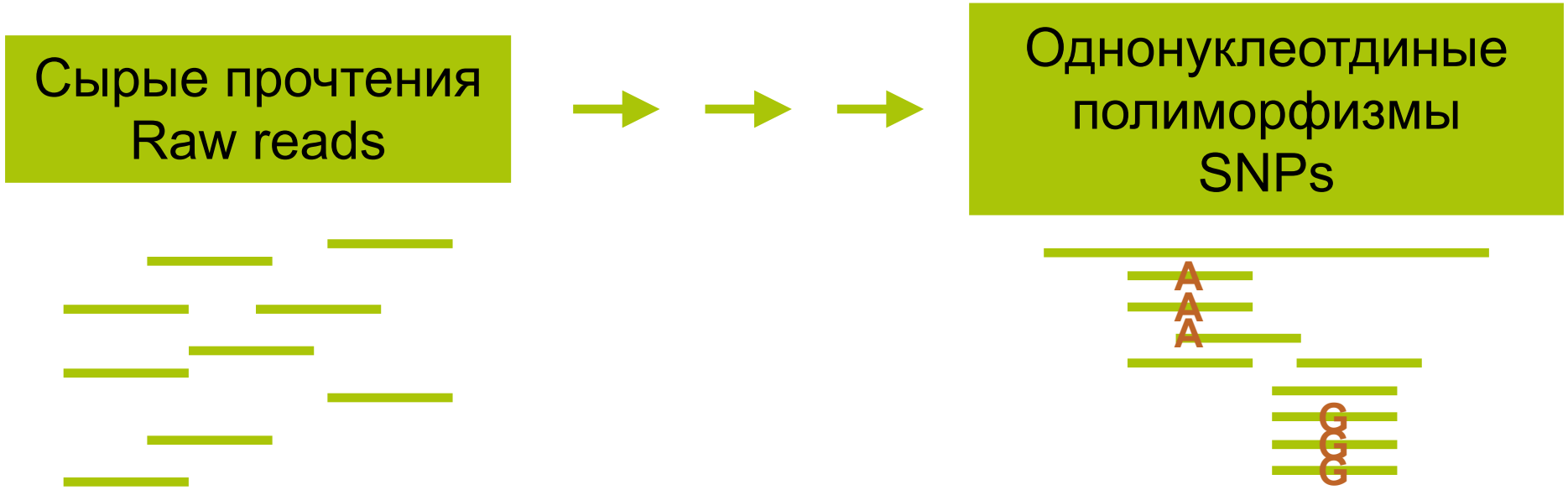
Our pipeline

Сырые прочтения
Raw reads

Однонуклеотдиные
полиморфизмы
SNPs

Skoltech

Our pipeline



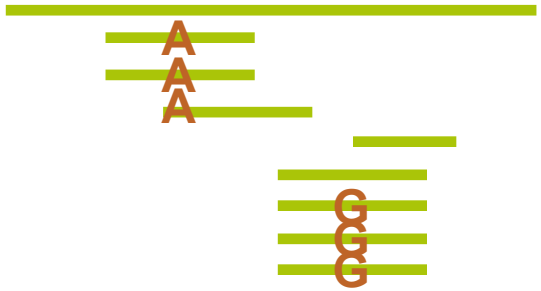
Skoltech

Our pipeline

Сырые прочтения
Raw reads

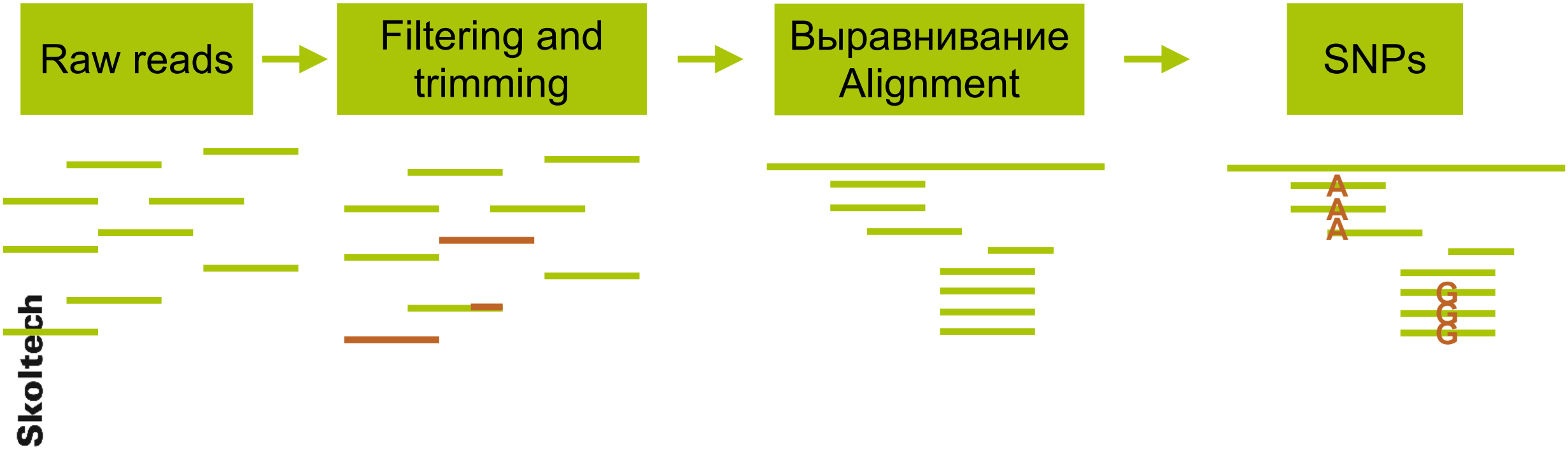
Фильтрация и тримминг
Filtering and trimming

Однонуклеотдиные
полиморфизмы
SNPs



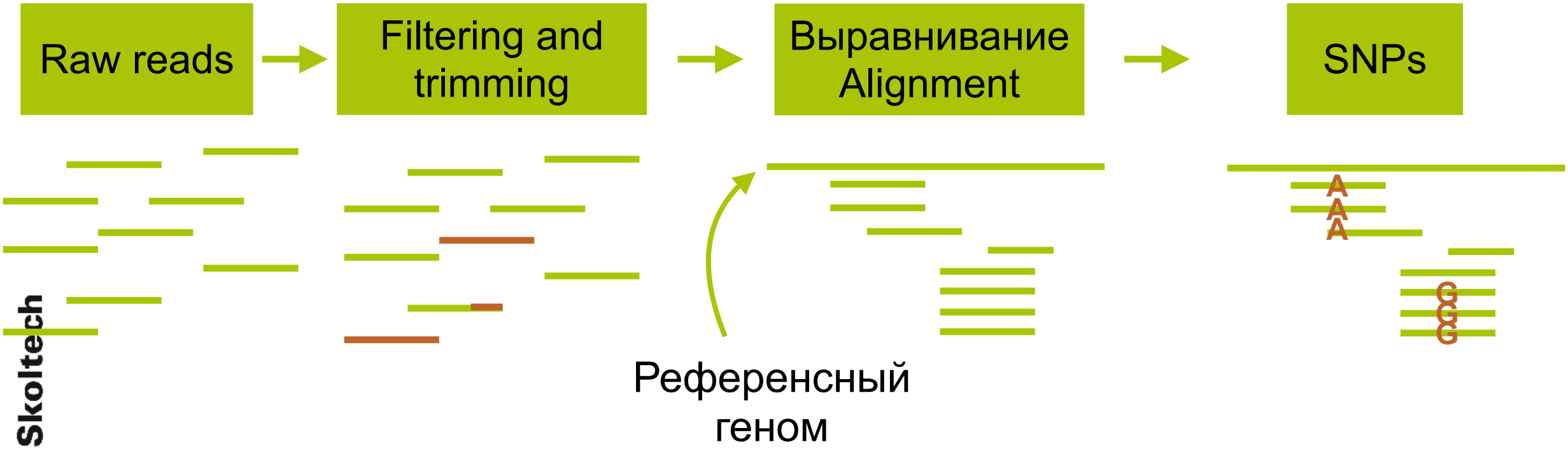
Skoltech

Our pipeline

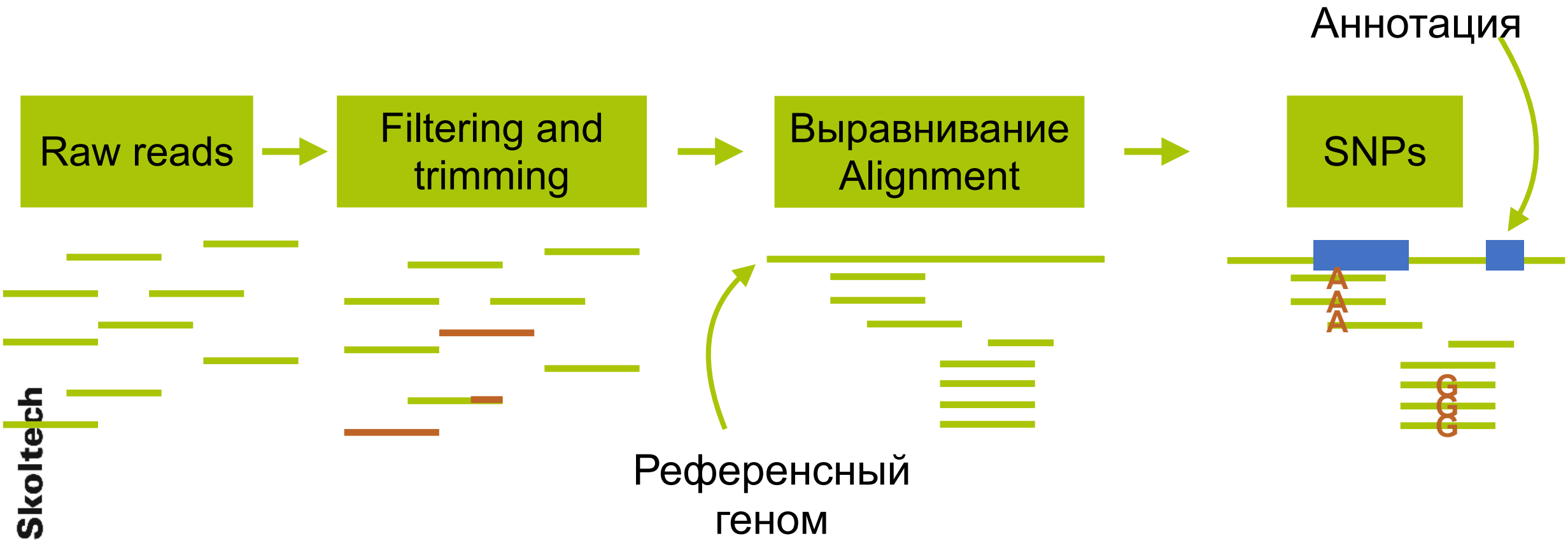


Skoltech

Our pipeline

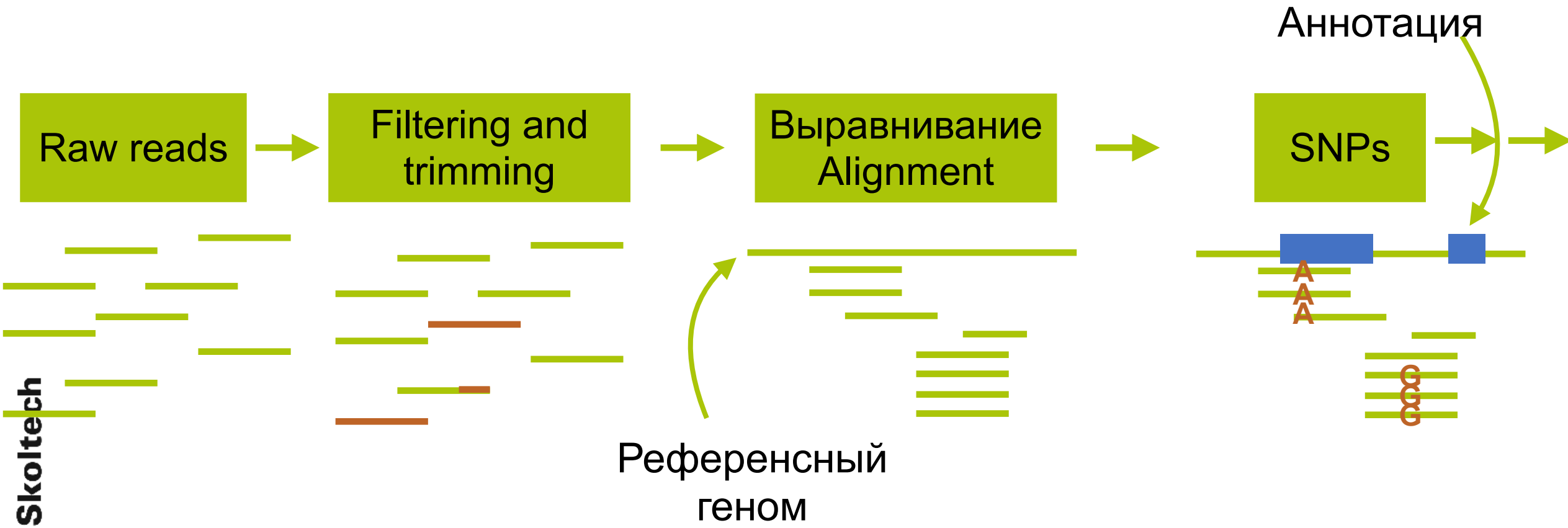


Our pipeline



Skoltech

Our pipeline



Как выглядят «сырые риды». FASTQ

```
@EAS139:136:FC706VJ:2:2104:15343:197393 1:Y:18:ATCACG
GATTTGGGGTTCAAAGCAGTATCGATCAAATAGTAAATCCATTTGTTCAACTCACAGTTT
+
!''*((((***+))%%%++)(%%%%).1***-+*''))**55CCF>>>>>>CCCCCCC65
```

ASCII_BASE=33 Illumina, Ion Torrent, PacBio and Sanger

Q	P_error	ASCII	Q	P_error	ASCII	Q	P_error	ASCII	Q	P_error	ASCII
0	1.00000	33 !	11	0.07943	44 ,	22	0.00631	55 7	33	0.00050	66 B
1	0.79433	34 "	12	0.06310	45 -	23	0.00501	56 8	34	0.00040	67 C
2	0.63096	35 #	13	0.05012	46 .	24	0.00398	57 9	35	0.00032	68 D
3	0.50119	36 $	14	0.03981	47 /	25	0.00316	58 :	36	0.00025	69 E
4	0.39811	37 %	15	0.03162	48 0	26	0.00251	59 ;	37	0.00020	70 F
5	0.31623	38 &	16	0.02512	49 1	27	0.00200	60 <	38	0.00016	71 G
6	0.25119	39 '	17	0.01995	50 2	28	0.00158	61 =	39	0.00013	72 H
7	0.19953	40 (	18	0.01585	51 3	29	0.00126	62 >	40	0.00010	73 I
8	0.15849	41)	19	0.01259	52 4	30	0.00100	63 ?	41	0.00008	74 J
9	0.12589	42 *	20	0.01000	53 5	31	0.00079	64 @	42	0.00006	75 K
10	0.10000	43 +	21	0.00794	54 6	32	0.00063	65 A			

$$E = 10^{-\left(\frac{Q}{10}\right)}$$

Как выглядят «сырые риды». FASTQ

```
@EAS139:136:FC706VJ:2:2104:15343:197393 1:Y:18:ATCACG
GATTTGGGGTTCAAAGCAGTATCGATCAAATAGTAAATCCATTTGTTCAACTCACAGTTT
+
!''*((((***+))%%%++)(%%%%).1***-+*''))**55CCF>>>>>>CCCCCCC65
```

ASCII_BASE=33 Illumina, Ion Torrent, PacBio and Sanger

Q	P_error	ASCII	Q	P_error	ASCII	Q	P_error	ASCII	Q	P_error	ASCII
0	1.00000	33 !	11	0.07943	44 ,	22	0.00631	55 7	33	0.00050	66 B
1	0.79433	34 "	12	0.06310	45 -	23	0.00501	56 8	34	0.00040	67 C
2	0.63096	35 #	13	0.05012	46 .	24	0.00398	57 9	35	0.00032	68 D
3	0.50119	36 $	14	0.03981	47 /	25	0.00316	58 :	36	0.00025	69 E
4	0.39811	37 %	15	0.03162	48 0	26	0.00251	59 ;	37	0.00020	70 F
5	0.31623	38 &	16	0.02512	49 1	27	0.00200	60 <	38	0.00016	71 G
6	0.25119	39 '	17	0.01995	50 2	28	0.00158	61 =	39	0.00013	72 H
7	0.19953	40 (	18	0.01585	51 3	29	0.00126	62 >	40	0.00010	73 I
8	0.15849	41)	19	0.01259	52 4	30	0.00100	63 ?	41	0.00008	74 J
9	0.12589	42 *	20	0.01000	53 5	31	0.00079	64 @	42	0.00006	75 K
10	0.10000	43 +	21	0.00794	54 6	32	0.00063	65 A			

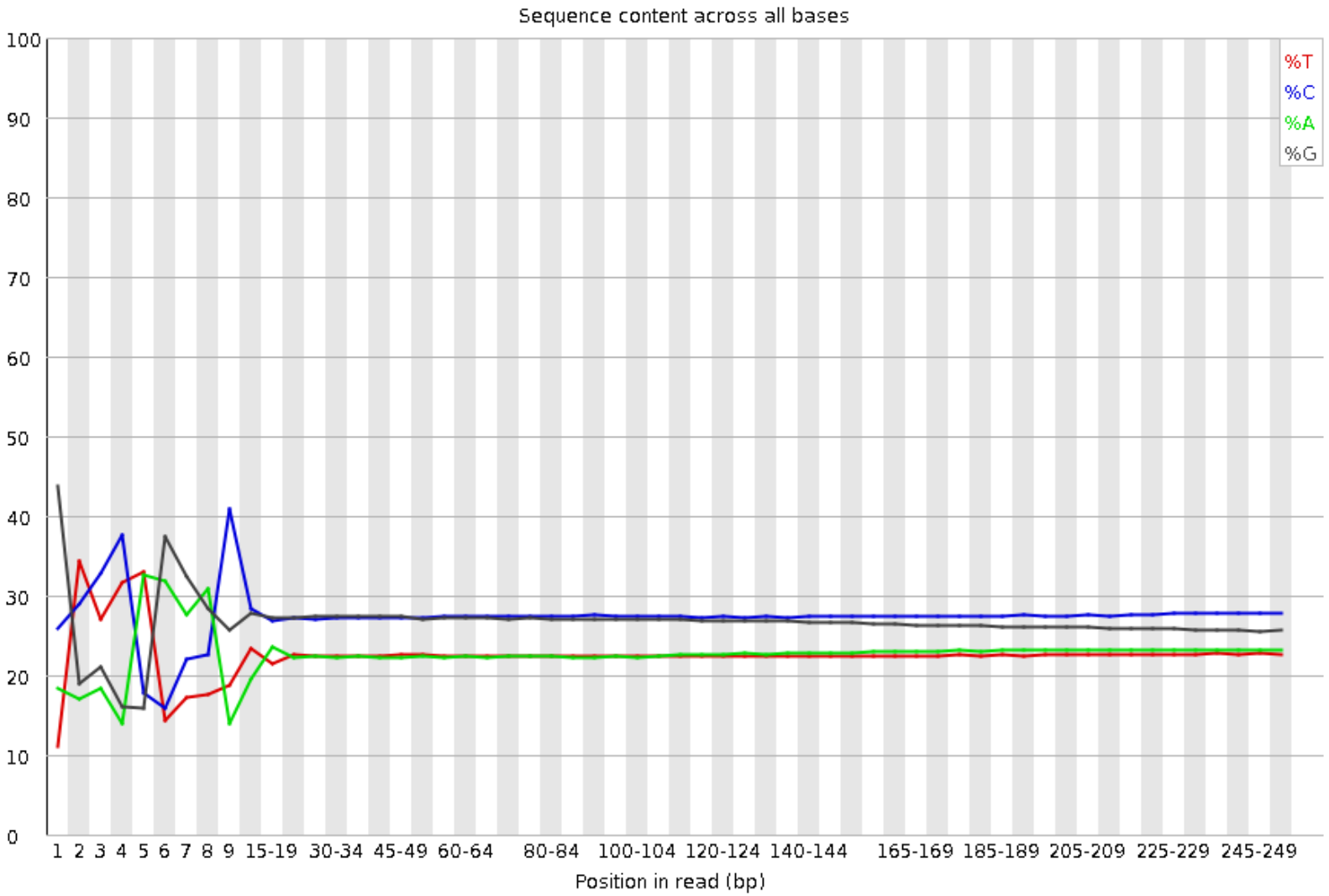
$$E = 10^{-\left(\frac{Q}{10}\right)}$$

Как выглядят «сырые риды». FASTQ

Phred Quality Score	Error	Accuracy (1 - Error)
10	1/10 = 10%	90%
20	1/100 = 1%	99%
30	1/1000 = 0.1%	99.9%
40	1/10000 = 0.01%	99.99%
50	1/100000 = 0.001%	99.999%
60	1/1000000 = 0.0001%	99.9999%

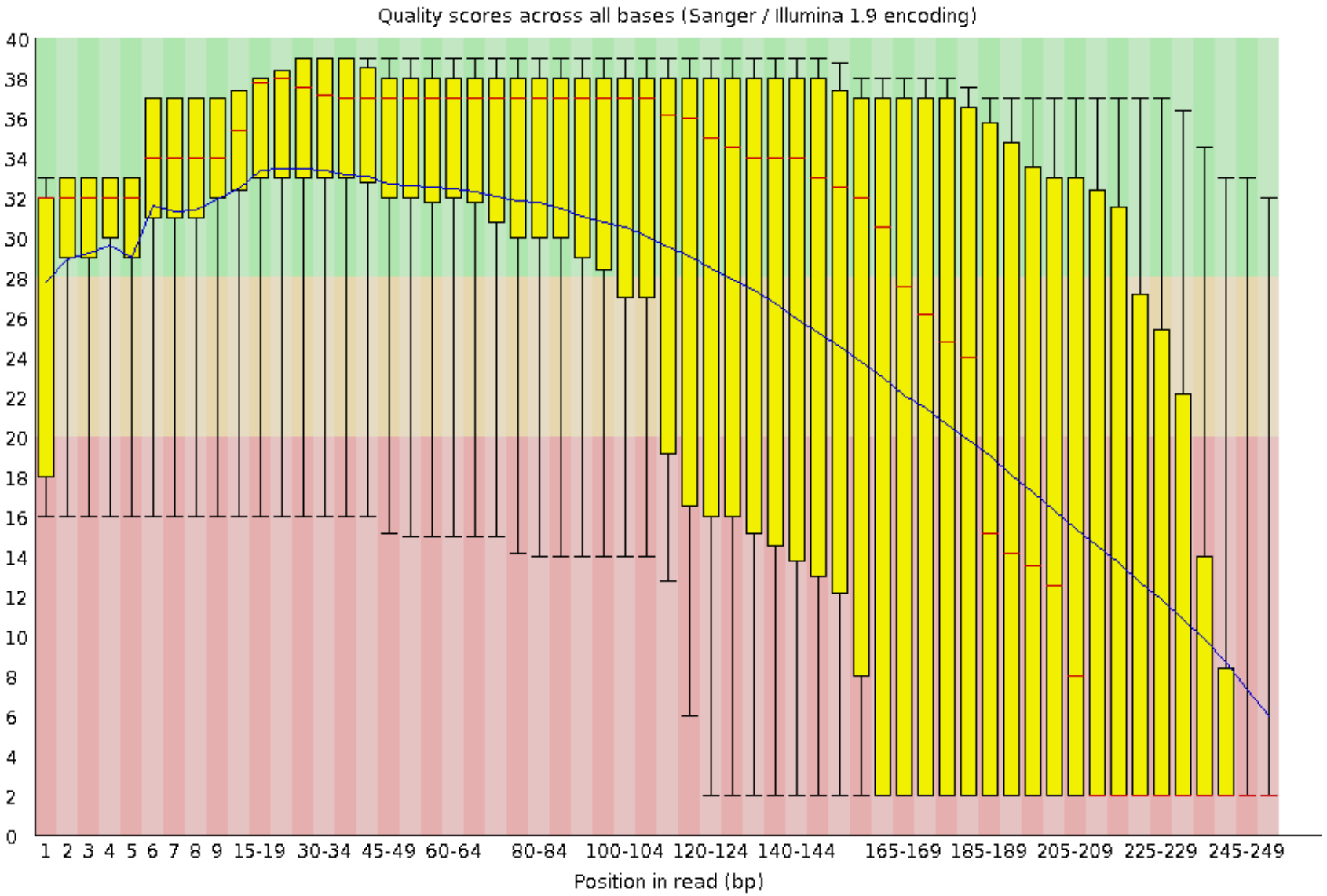
$$E = 10^{-\left(\frac{Q}{10}\right)}$$

Примеры плохого качества FastQC

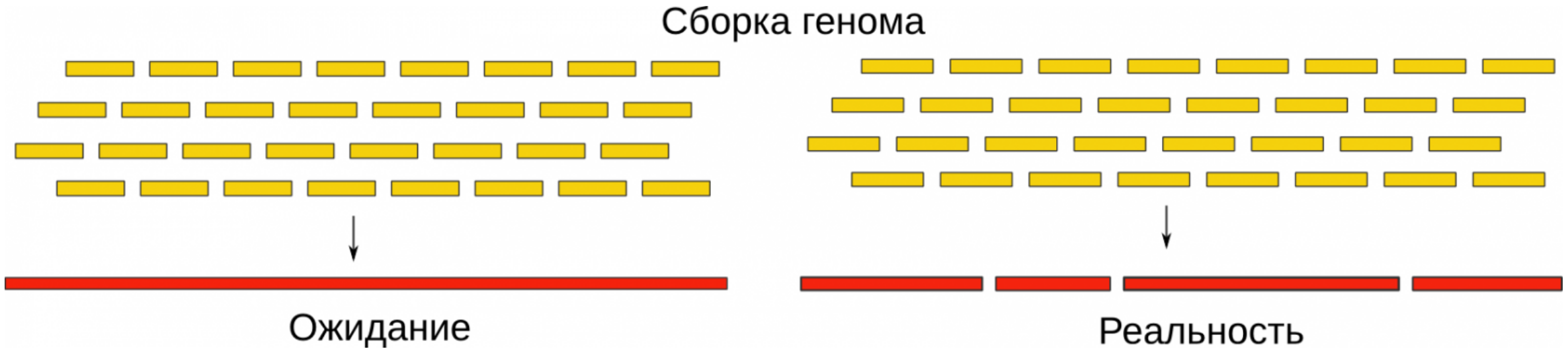


Skoltech

Примеры плохого качества FastQC

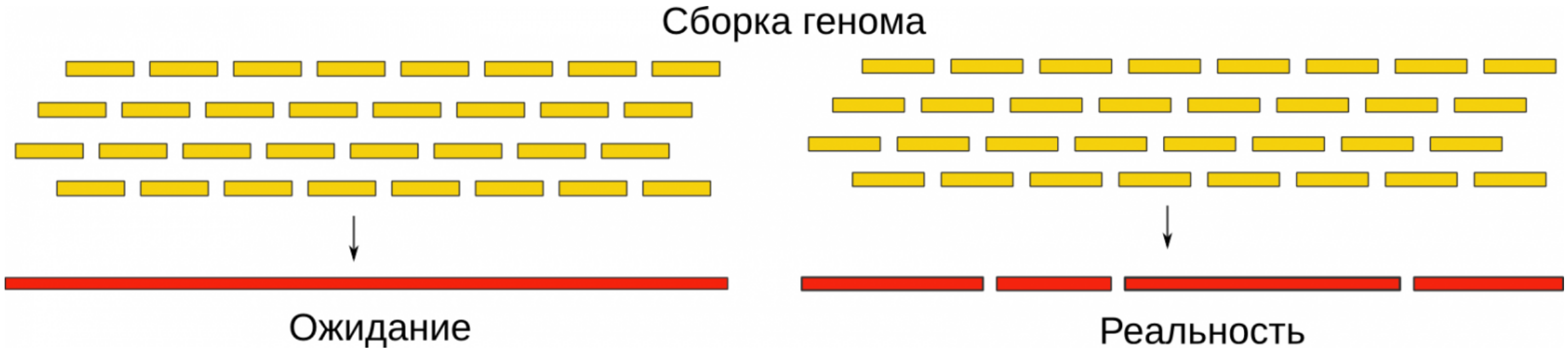


Сборка генома Genome assembly



Skoltech

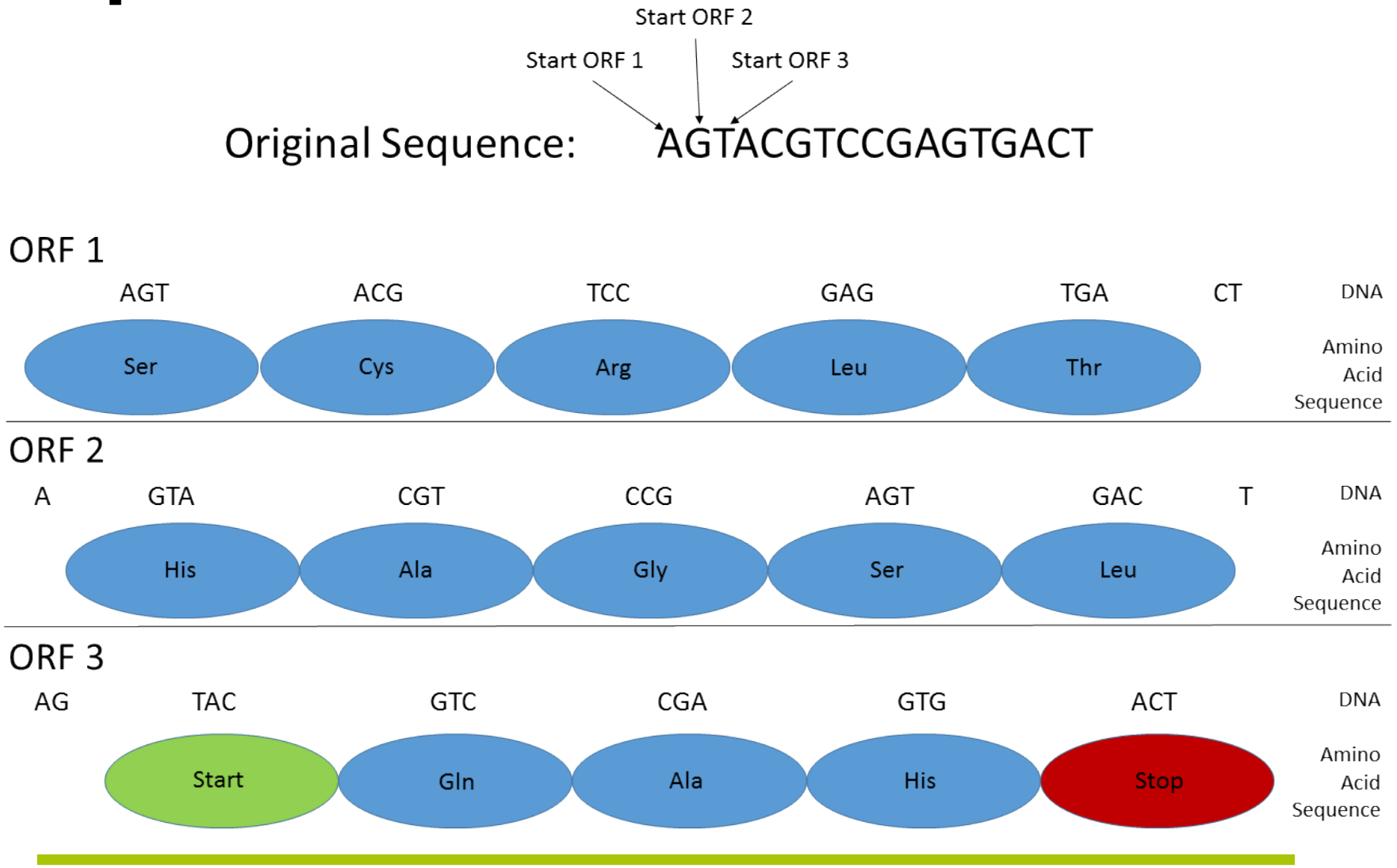
Сборка генома Genome assembly



Парные риды

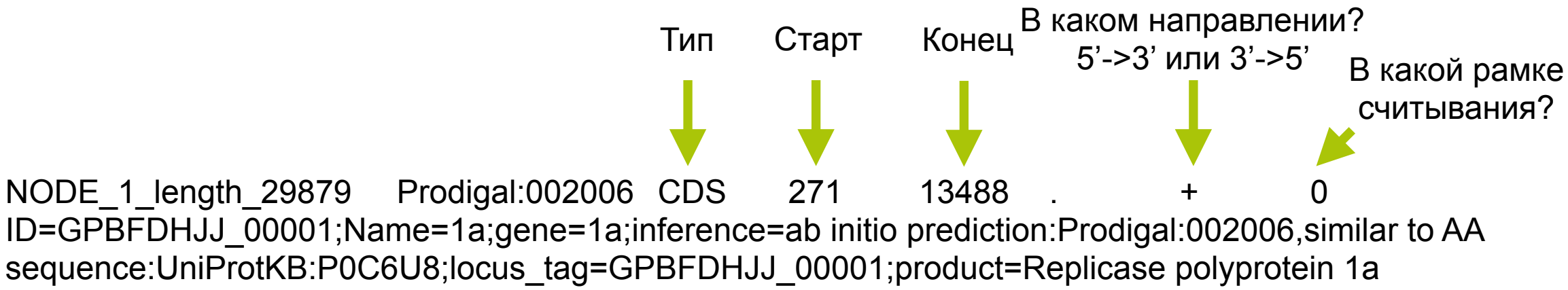
Skoltech

Аннотирование генома



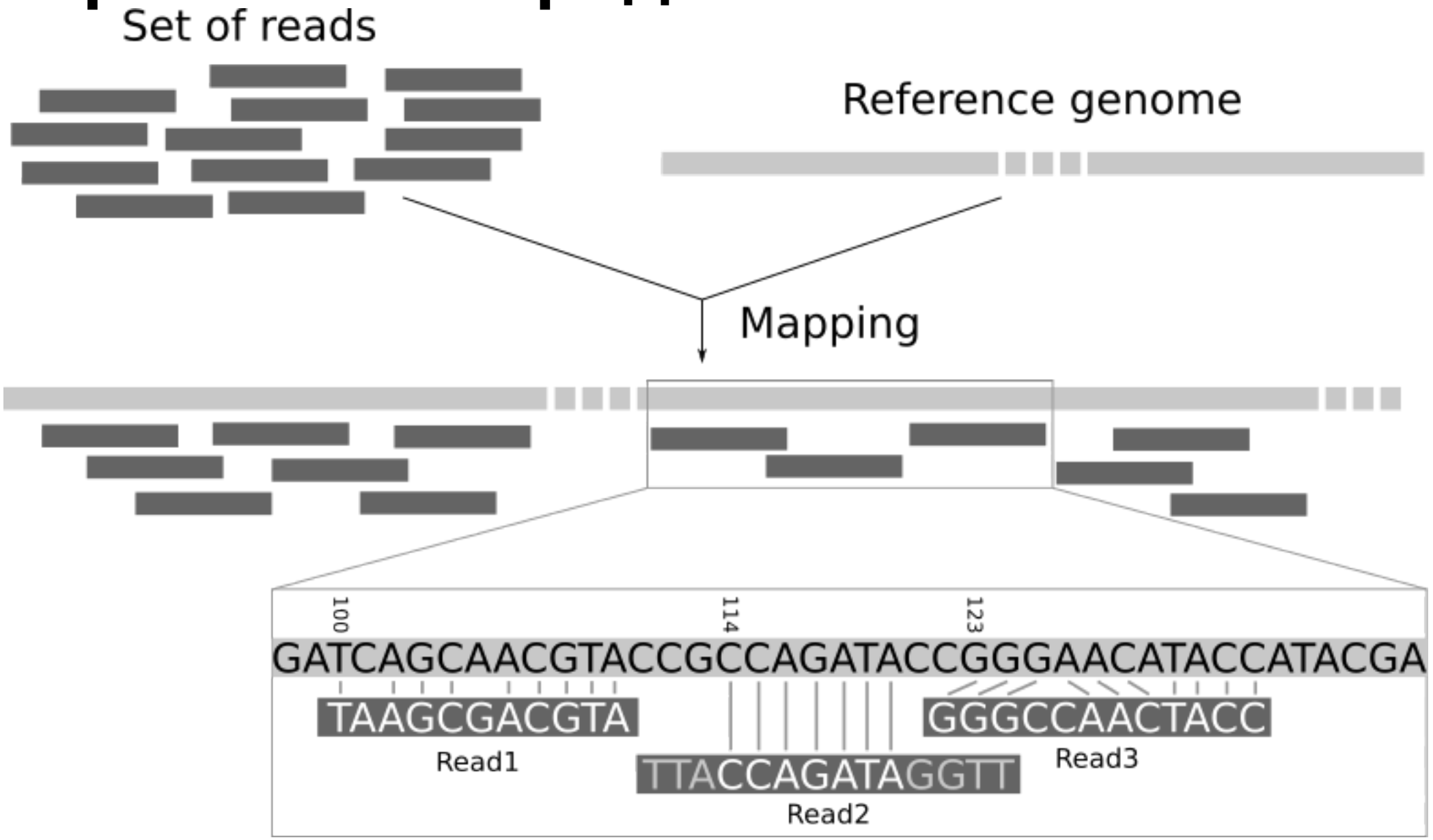
Search in NCBI

Аннотирование генома



Skoltech

Выравнивание ридов

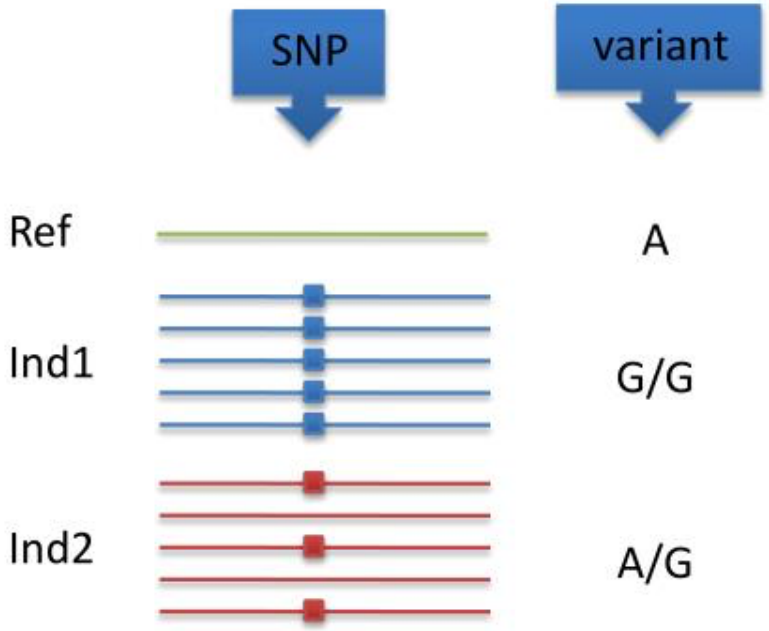


Выравнивание ридов. SAM формат

SRR10971381.512 83 NC_045512.2 3248 60 143M = 2969 -422
GAGCACAATCAGACAACTACTATTCAAACAATTGTTGAGGTTCAACCTCAATTAGAGATGGAACTAACACCAGTTGTT
CAGACTATTGAAGTGAATAGTTTTAGTGGTTATTTAAAACTTACTGACAATGTATACATTAAAAA F/=/FFFAA/A/
FFFAFFFF/AFFAFAAFFAFFF//F=FF//FFA=/FFFF//FFFFFFA/FFF//FFAFF//FAFF/A/F/A/FFFAF6FF6F/F/F//F/AFFFA6F/
FFFFFFA/FFFFF6AFFFFF/FFFFFFFFF/F6 NM:i:2 MD:Z:3G61T77 MC:Z:142M AS:i:134

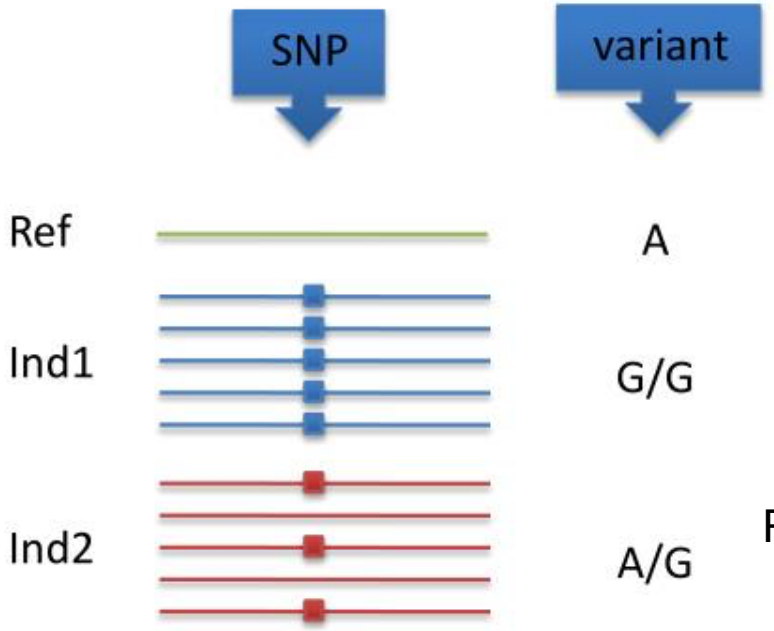
Skoltech

SNP calling. VCF формат



Skoltech

SNP calling. VCF формат



Как организована информация об образце?

Информация по позиции

Reference allele

Alternative allele

Информация об образце

```
#CHROM          POS     ID      REF     ALT     QUAL    FILTER  INFO    FORMAT  Sample1
contig_length_29869     205     .       G       T       .       PASS
ADP=6883;WT=0;HET=0;HOM=1;NC=0          GT:GQ:SDP:DP:RD:AD:FREQ:PVAL:RBQ:ABQ:RDF:RDR:ADF:ADR
1/1:255:6883:6883:2:6879:99,97%:0E0:25:35:2:0:6879:0
```

Skoltech

Our pipeline