An effective bioinformatic quality control method for NGS

Sijung Yun, Ph.D.
NIDDK/Yotta Biomed, LLC.

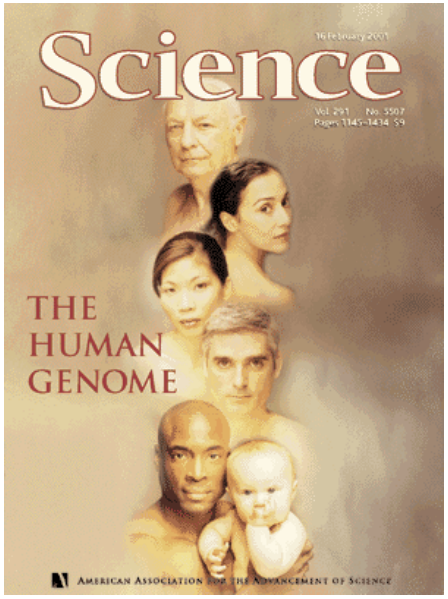
Next generation sequencing (NGS)
Sanger sequencing vs. NGS

Sanger sequencing

- Human genome project
- 1990 ~ 2003
- ~3 billion dollars

NGS for a human genome

- Illumina's Xten
- ~3 days (Q1 of 2014)
- ~$1500 including library prep



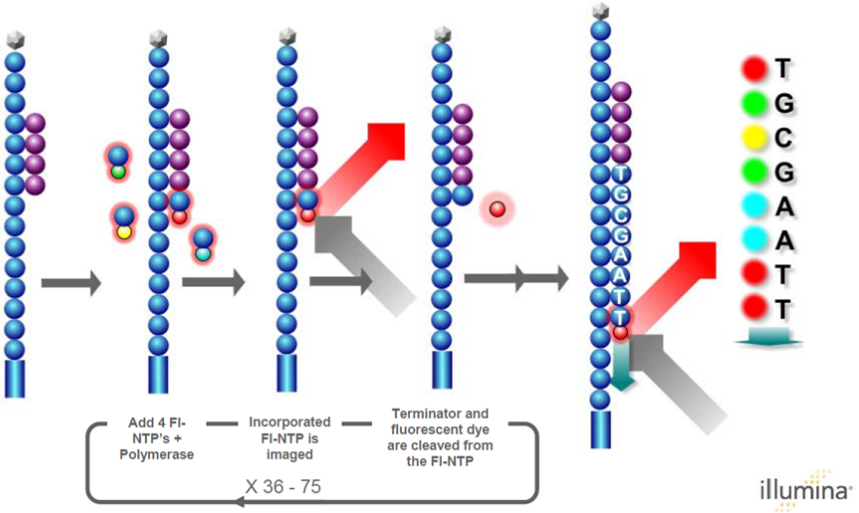
Second generation sequencers Illumina

Sequencers

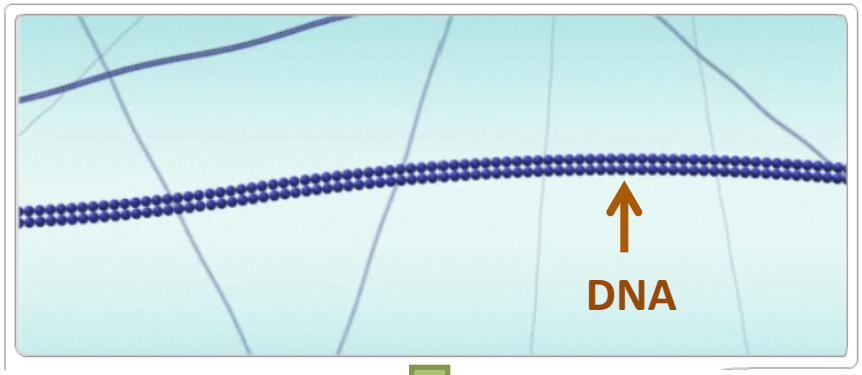
- **Xten (Numbers are for reference only, contact the company for details)**
  - 320 human genome sequencing per week (0.5 hr/human genome)
  - Available from the first quarter of 2014
- **HiSeq-2500**
  - 27 hr/human genome in 30x
  - Available since 2012
- **NextSeq 500**
  - 29 hr/human genome in 30x
  - Available from January 14th, 2014
- **MiSeq**
  - Appropriate for exom sequencing (5GB/day) ➔ ~18 days for 30x human whole genome sequencing

Key Technologies

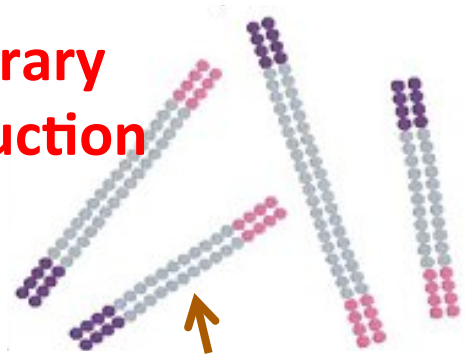
- Sequencing by synthesis
- Bridge amplification

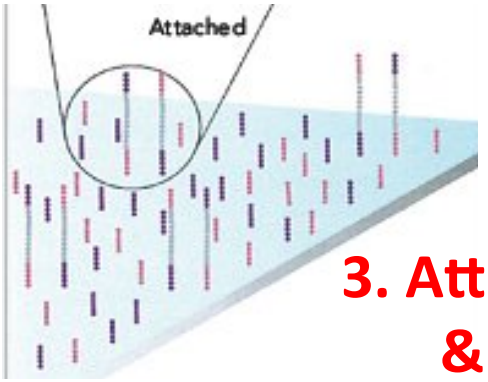


How does Illumina's HiSeq work?

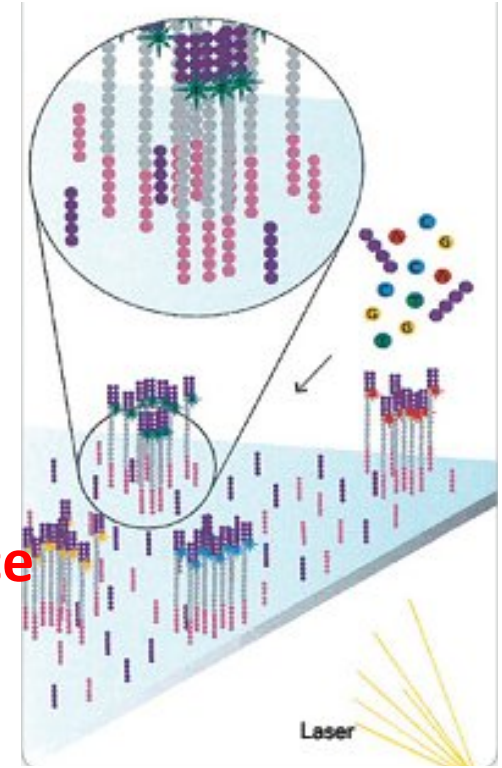


2. Library construction





3. Attach & amplify

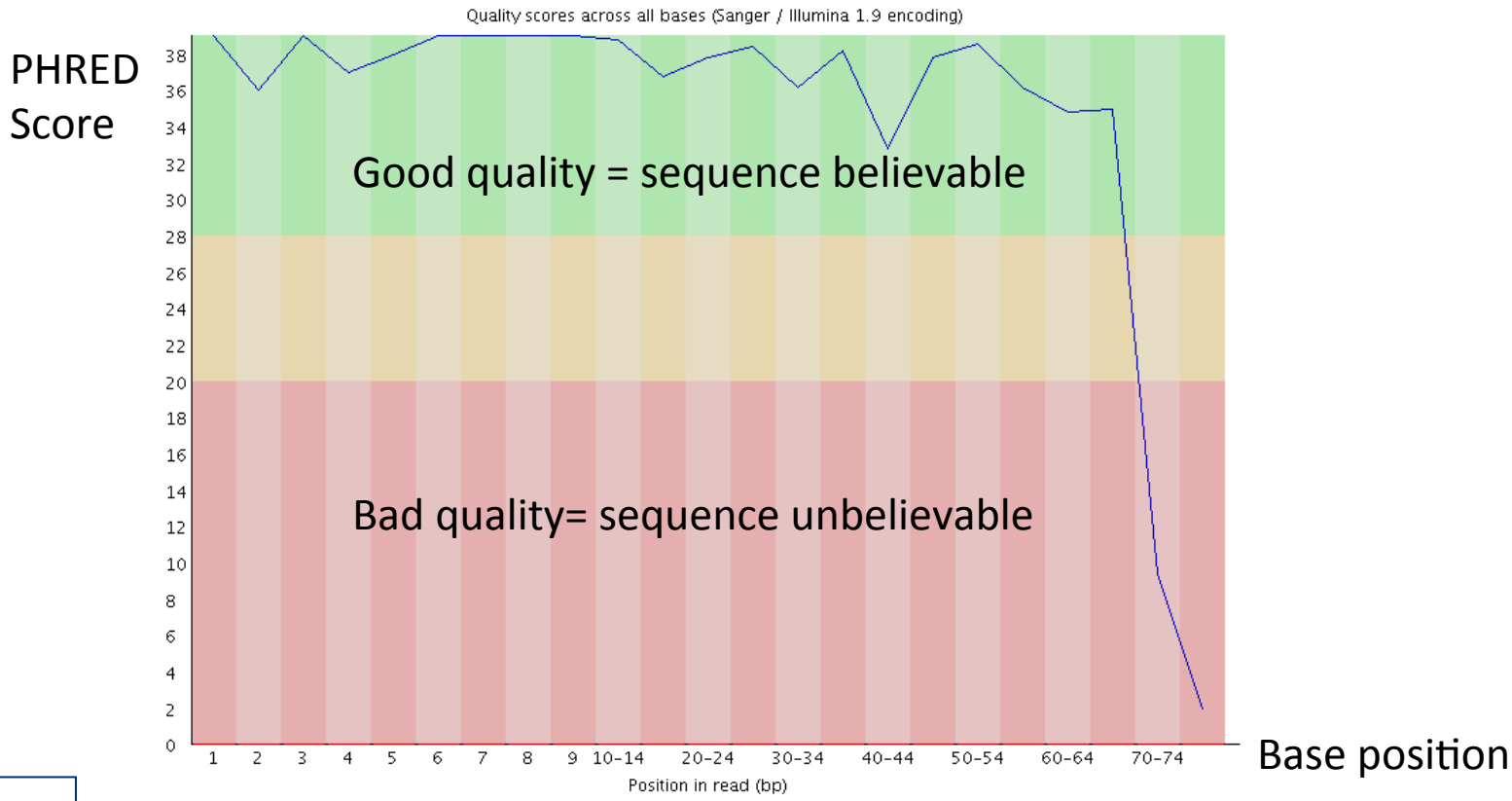


4. Sequence by synthesis

Next Generation Sequencer
&
low quality base calls

- Even the most recent NGS sequencers generate base calls with high error rate!

(Example) Quality of a read



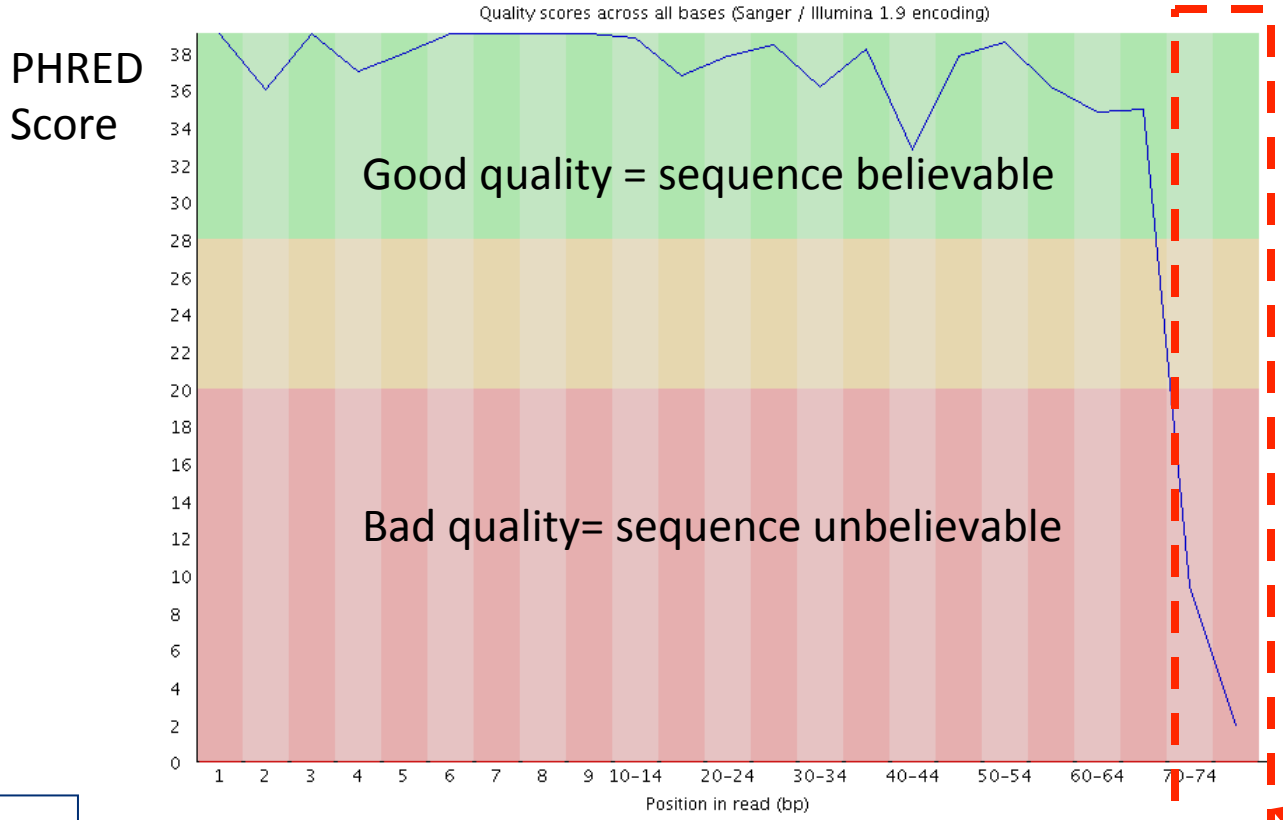
Sequence

TTGTTTAAGATGAGTACTAGATTGCTGTGCCGAAACCATACAGTTCAACATATTAATTCGTAACACGTTTAAAANC

Quality

HEHFGHHHHGHHHHEHHCEHBHHHHHHEHHHH=EGHHHE@F=F@FHFFHHHHFH@HHHBHHEF8HHH<@H#####

(Example) Quality of a read



Base position

Sequence

TTGTTTAAGATGAGTACTAGATTGCTGTGCCGAAACCATACAGTTCAACATATTAATTCGTAACACGTTTAAAANC

Quality

HEHFGHHHHGHHHHEHHCEHBHHHHHHEHHHH=EGHHHE@F=F@FHFFHHHHFH@HHHBHHEF8HHH<@H######

What would you do with the base calls of high error rate? ➔ Trimming

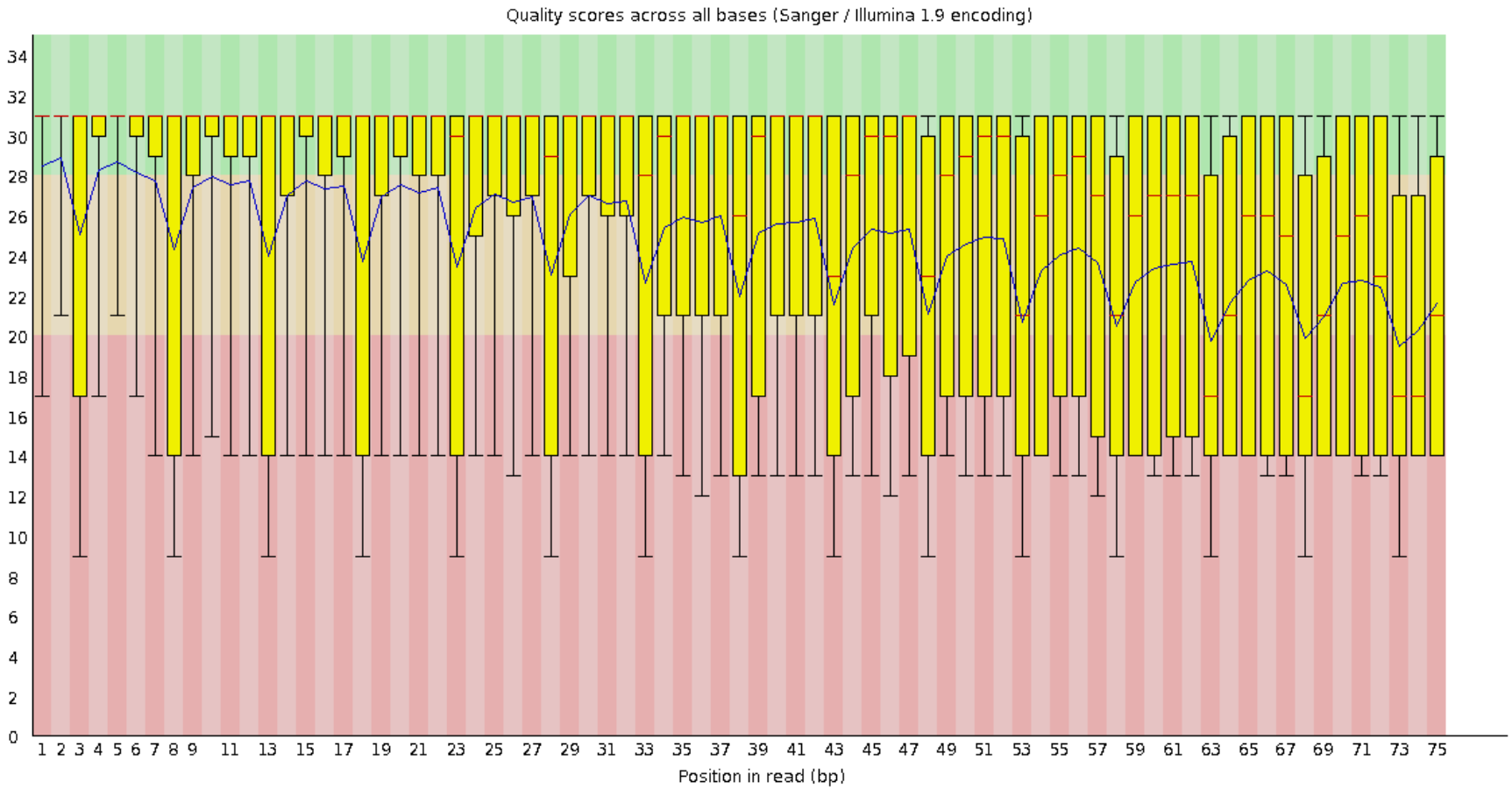
- Currently, trim them off!

```
@HWI-EAS179_0001:5:1:9:1271#0/1
TTGTTTAAGATGAGTACTAGATTGCTGTGCCGAAACCATACAGTTCAACATATTAATTCGTAACACGTTTAAAANC
+HWI-EAS179_0001:5:1:9:1271#0/1
HEHFGHHHHGHHHHEHHCEHBHHHHHHEHHHH=EGHHHE@F=F@FHFFHHHHFH@HHHBHHEF8HHH<@H######
```

Trimming

```
@HWI-EAS179_0001:5:1:9:1271#0/1
TTGTTTAAGATGAGTACTAGATTGCTGTGCCGAAACCATACAGTTCAACATATTAATTCGTAACACGTTTAAAANC
+HWI-EAS179_0001:5:1:9:1271#0/1
HEHFGHHHHGHHHHEHHCEHBHHHHHHEHHHH=EGHHHE@F=F@FHFFHHHHFH@HHHBHHEF8HHH<@H######
```

How to trim if you have multiple regions of low quality?



ALTERNATIVE WAY?

➔ MASKING

Masking

- Masking: Substitute a base with low quality as ‘N’

```
@HWI-EAS179_0001:5:1:9:1271#0/1
TTGTTTAAGATGAGTACTAGATTGCTGTGCCGAAACCATACAGTTCAACATATTAATTCGTAACACGTTTAAAANC
+HWI-EAS179_0001:5:1:9:1271#0/1
HEHFGHHHHGHHHHEHHCEHBHHHHHHEHHHH=EGHHHE@F=F@FHFFHHHHFH@HHHBHHEF8HHH<@H######
```

Masking

```
@HWI-EAS179_0001:5:1:9:1271#0/1
TTGTTTAAGATGAGTACTAGATTGCTGTGCCGAAACCATACAGTTCAACATATTAATTCGTAACACGTTTNNNNNN
+HWI-EAS179_0001:5:1:9:1271#0/1
HEHFGHHHHGHHHHEHHCEHBHHHHHHEHHHH=EGHHHE@F=F@FHFFHHHHFH@HHHBHHEF8HHH<@H######
```

(Q) WHICH ONE IS BETTER?

MASKING OR TRIMMING?

A recently accepted manuscript at BMC Bioinformatics

Masking as an effective quality control method for next-generation sequencing data analysis

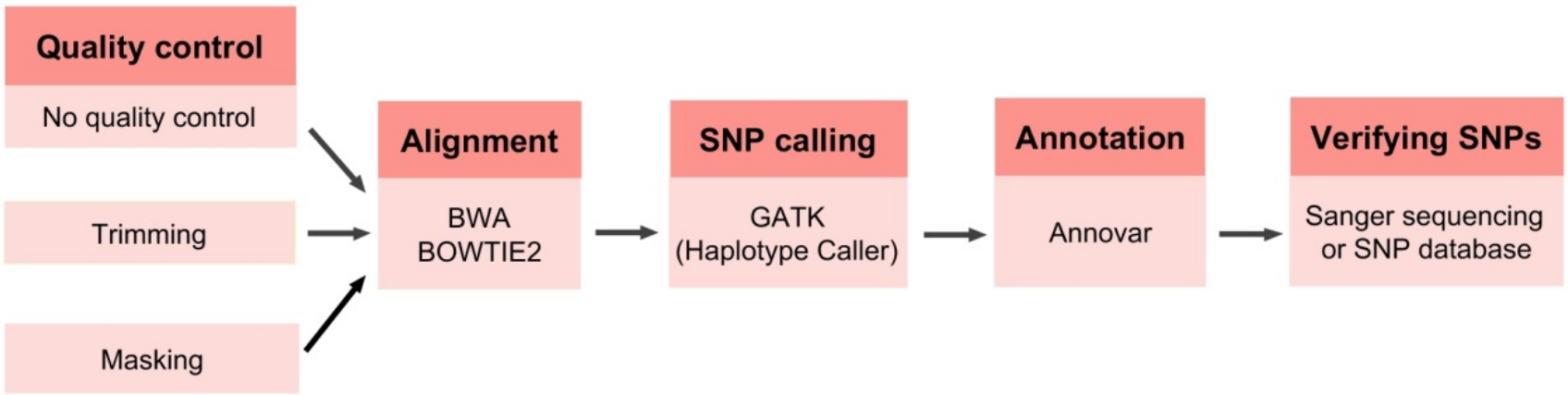
Sajung Yun[1], Sijung Yun[2§]

Abstract

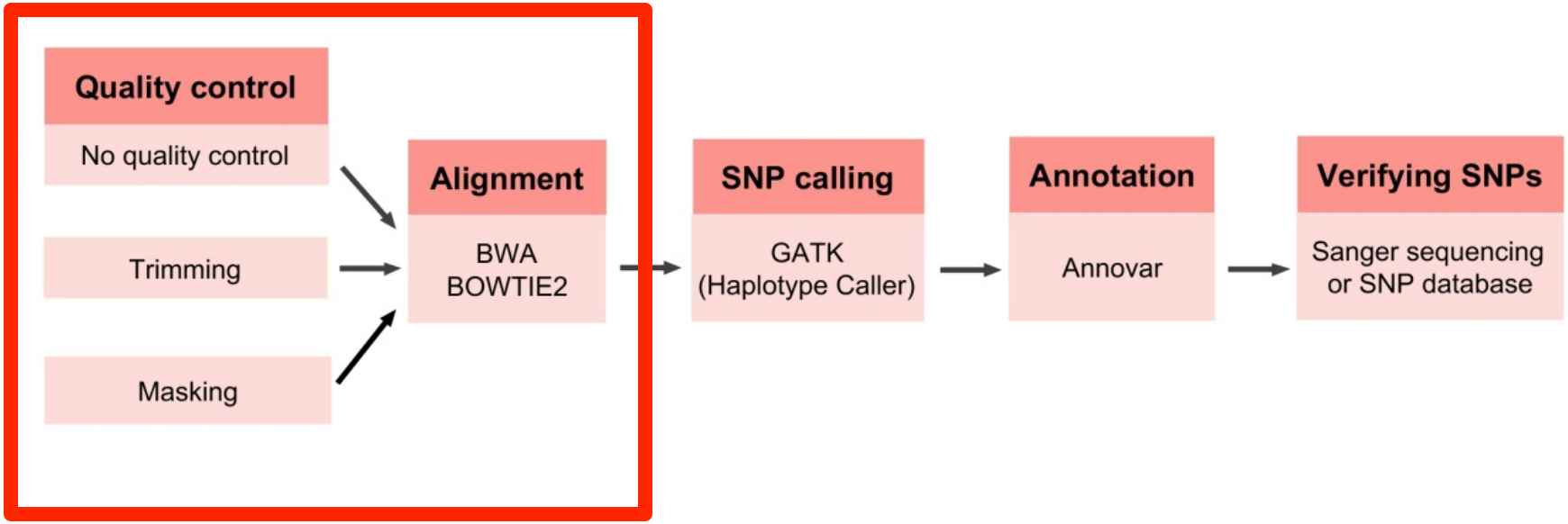
Next generation sequencing produces base calls with low quality scores that can affect the accuracy of identifying simple nucleotide variation calls, including single nucleotide polymorphisms and small insertions and deletions. Here we compare the effectiveness of two data preprocessing methods, masking and trimming, and the accuracy of simple nucleotide variation calls on whole-genome sequence data from *Caenorhabditis elegans*. Masking substitutes low quality base calls with 'N's (undetermined bases), whereas trimming removes low quality bases that results in a shorter read lengths.

…

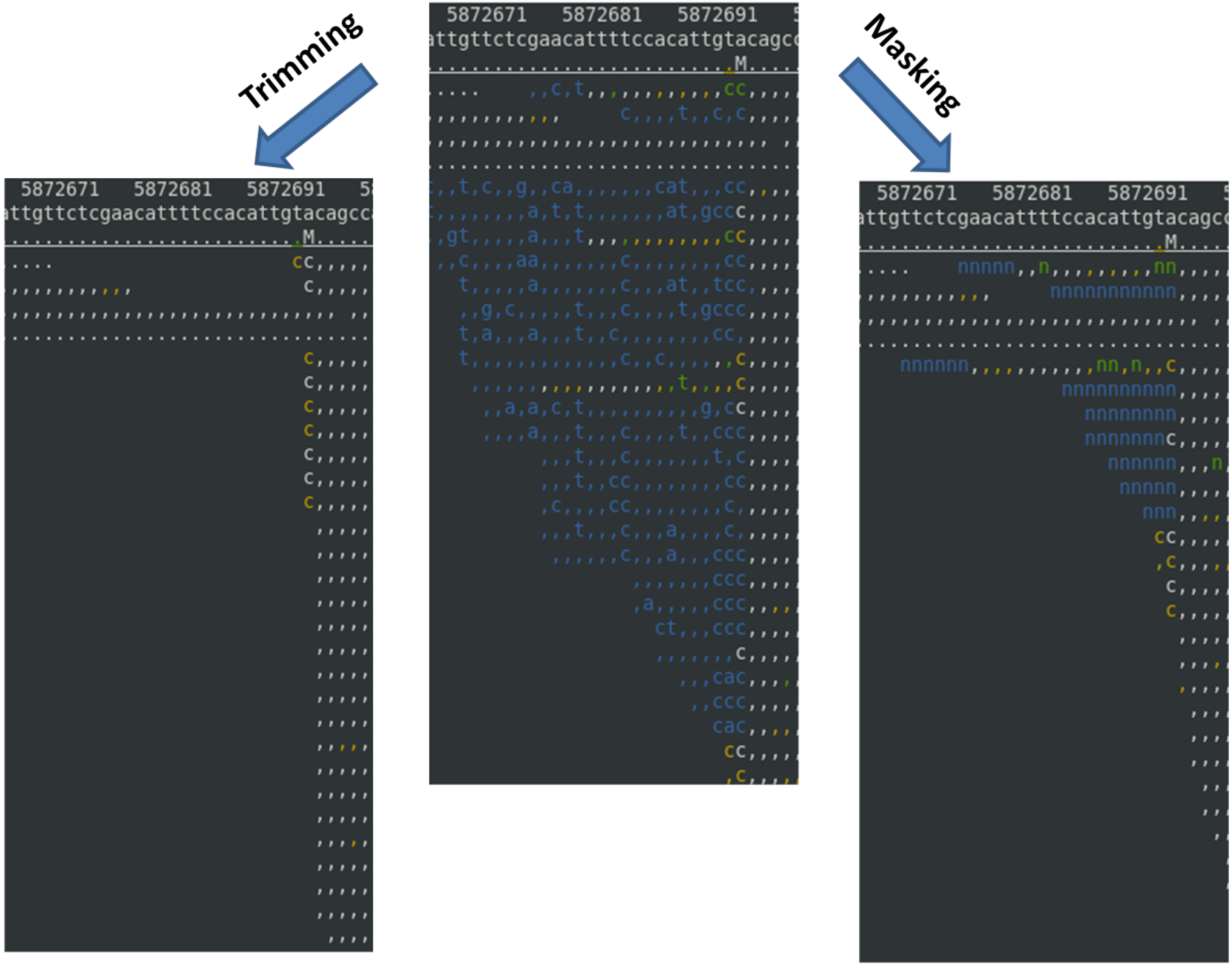
Study design



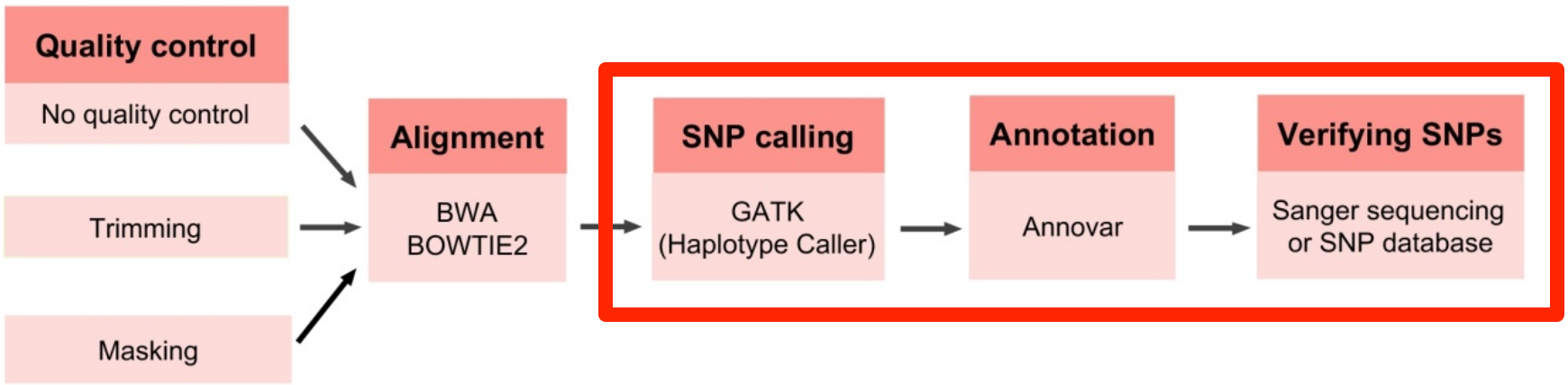
Study design



No quality control



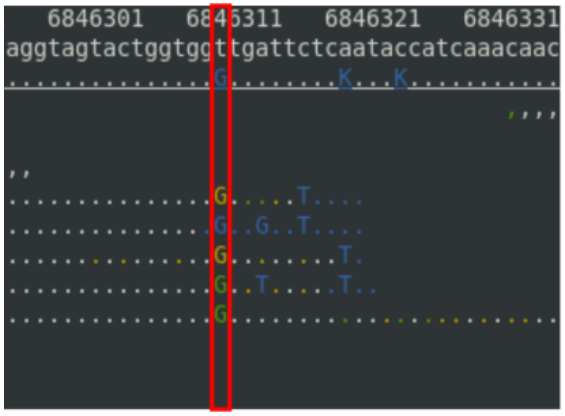
Study design



Masking (0% FP) vs. Trimming (2.3% FP)

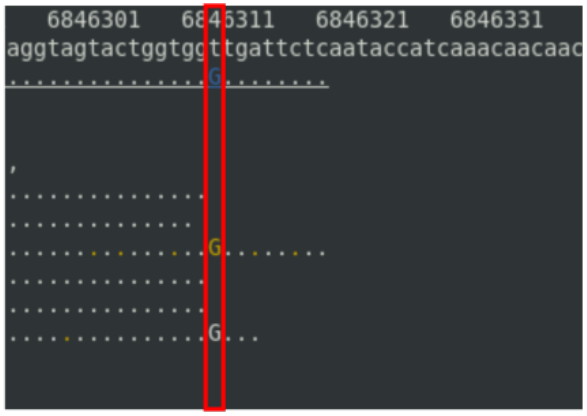
Aligners	No quality control				Trimming				Masking			
	TP	FP	TN	FN	TP	FP	TN	FN	TP	FP	**TN**	**FN**
BWA	35	2 (2.3 %)	84	2	35	2 (2.3 %)	84	2	35	0 (0 %)	86	2
BOWTIE2	35	6 (7.0 %)	80	2	35	2 (2.3 %)	84	2	35	0 (0 %)	86	2

Masking was correct not to call as a SNP



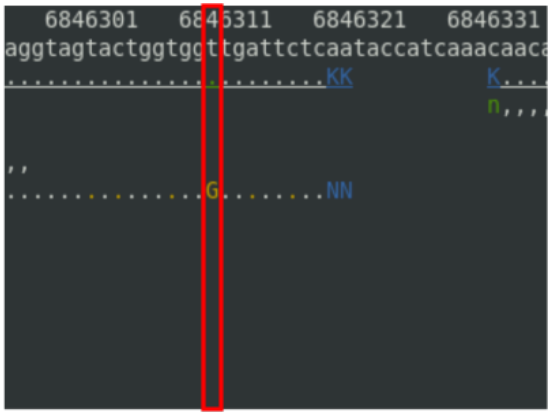
(a)

No pre-processing



(b)

Trimming



(c)

Masking

THEREFORE, MASKING WAS EFFECTIVE IN REDUCING 'FALSE POSITIVE RATE' IN SNP CALLING

HOW ABOUT FALSE NEGATIVE RATE?

Masking & Trimming:
No effect in false negative rate

Aligners	No quality control				Trimming				Masking			
	TP	FP	TN	FN	TP	FP	TN	FN	TP	FP	**TN**	**FN**
BWA	35	2 (2.3 %)	84	2	35	2 (2.3 %)	84	2	35	0 (0 %)	86	2
BOWTIE2	35	6 (7.0 %)	80	2	35	2 (2.3 %)	84	2	35	0 (0 %)	86	2

Masking & Trimming: No effect in false negative rate

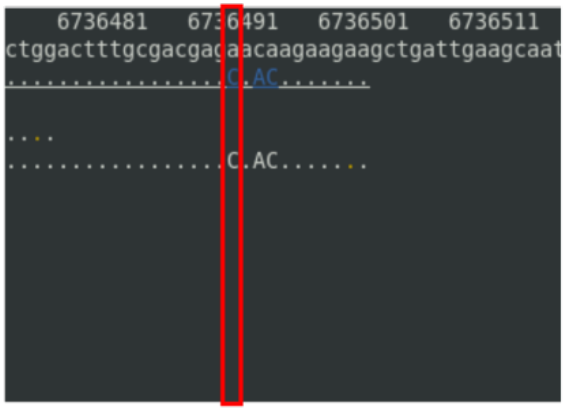
Aligners	No quality control	Trimming	Masking
BWA	104,858 (6.4 %)	105,355 (6.0 %)	104,775 (6.5 %)
BOWTIE2	104,422 (6.8 %)	104,470 (6.8 %)	104,297 (6.9 %)

P-values from χ^2 test for significance > 0.05

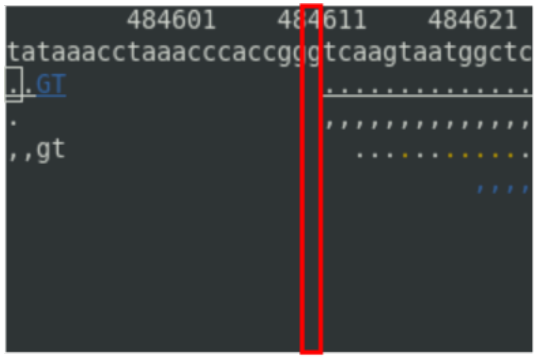
WHY ARE PRE-PROCESSING METHODS NOT EFFECTIVE IN REDUCING FALSE NEGATIVE RATES?

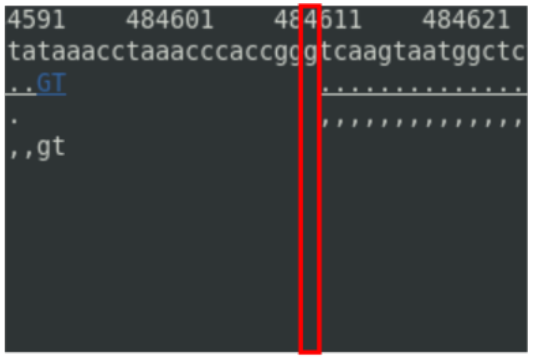
False negative due to low coverage!

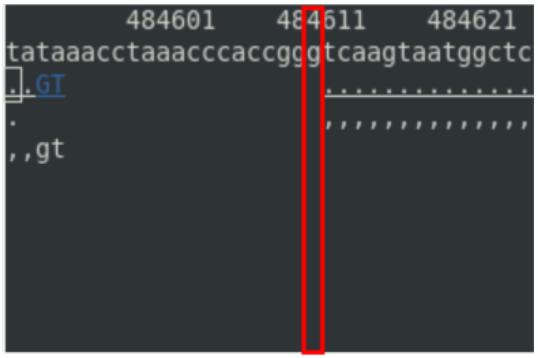
(Example 1)



No pre-processing **Trimming** **Masking**







(Example 2)

HOW ABOUT INDELS?

No effect for Indels

Aligners	*No quality control*		*Trimming*		*Masking*	
	True-Positive	**False-Negative**	**True-Positive**	**False-Negative**	**True-Positive**	**False-Negative**
BWA	20	2	20	2	20	2
BOWTIE2	20	2	20	2	20	2

IS THE LOCAL ALIGNMENT MORE EFFECTIVE THAN PRE-PROCESSING?

Pre-processing works better than local alignment in false positives

	No quality control	Trimming	Masking
Global alignment	7.0 %	2.3 %	0 %
Local alignment	5.8 %	2.3 %	0 %

Summary

- Take home message
  - Masking is more effective in reducing false positive rate in SNP calling than trimming

- Minor findings
  - Pre-processing (masking or trimming) not effective in reducing
    - false negative rate
    - Indel discovery
  - Pre-processing more effective than local alignment in reducing false positive rate

convert ascii33 to error probability

front

$$Q_{\mathrm{PHRED}} = -10 \times \log_{10}(P_e)$$

BITS VIB
Bioinformatics Training and Service Facility

hg18-total-sequenced=2'858'034'764 (UCSC)

Char (q)	Dec	Q	error probability	%correct	1-error in # bases	# errors in 2.85Gb
!	33	0	1.00E+00	0.000%	1	2,858,034,764
"	34	1	7.94E-01	20.567%	1	2,270,217,709
#	35	2	6.31E-01	36.904%	2	1,803,298,025
$	36	3	5.01E-01	49.881%	2	1,432,410,537
%	37	4	3.98E-01	60.189%	3	1,137,804,133
&	38	5	3.16E-01	68.377%	3	903,789,949
`	39	6	2.51E-01	74.881%	4	717,905,874
(	40	7	2.00E-01	80.047%	5	570,252,906
)	41	8	1.58E-01	84.151%	6	452,967,984
*	42	9	1.26E-01	87.411%	8	359,805,259
+	43	10	1.00E-01	90.000%	10	285,803,476
,	44	11	7.94E-02	92.057%	13	227,021,771
-	45	12	6.31E-02	93.690%	16	180,329,803
.	46	13	5.01E-02	94.988%	20	143,241,054
/	47	14	3.98E-02	96.019%	25	113,780,413
0	48	15	3.16E-02	96.838%	32	90,378,995
1	49	16	2.51E-02	97.488%	40	71,790,587
2	50	17	2.00E-02	98.005%	50	57,025,291
3	51	18	1.58E-02	98.415%	63	45,296,798
4	52	19	1.26E-02	98.741%	79	35,980,526
5	53	20	1.00E-02	99.000%	100	28,580,348
6	54	21	7.94E-03	99.206%	126	22,702,177
7	55	22	6.31E-03	99.369%	158	18,032,980
8	56	23	5.01E-03	99.499%	200	14,324,105
9	57	24	3.98E-03	99.602%	251	11,378,041
:	58	25	3.16E-03	99.684%	316	9,037,899
;	59	26	2.51E-03	99.749%	398	7,179,059
<	60	27	2.00E-03	99.800%	501	5,702,529
=	61	28	1.58E-03	99.842%	631	4,529,680
>	62	29	1.26E-03	99.874%	794	3,598,053
?	63	30	1.00E-03	99.900%	1,000	2,858,035
@	64	31	7.94E-04	99.921%	1,259	2,270,218
A	65	32	6.31E-04	99.937%	1,585	1,803,298
B	66	33	5.01E-04	99.950%	1,995	1,432,411
C	67	34	3.98E-04	99.960%	2,512	1,137,804
D	68	35	3.16E-04	99.968%	3,162	903,790
E	69	36	2.51E-04	99.975%	3,981	717,906
F	70	37	2.00E-04	99.980%	5,012	570,253
G	71	38	1.58E-04	99.984%	6,310	452,968
H	72	39	1.26E-04	99.987%	7,943	359,805
I	73	40	1.00E-04	99.990%	10,000	285,803
J	74	41	7.94E-05	99.992%	12,589	227,022
K	75	42	6.31E-05	99.994%	15,849	180,330
L	76	43	5.01E-05	99.995%	19,953	143,241
M	77	44	3.98E-05	99.996%	25,119	113,780
N	78	45	3.16E-05	99.997%	31,623	90,379
O	79	46	2.51E-05	99.997%	39,811	71,791

back

Char (q)	Dec	Q	error probability	%correct	1-error in # bases	# errors in 2.85Gb
P	80	47	2.00E-05	99.998%	50,119	57,025
Q	81	48	1.58E-05	99.998%	63,096	45,297
R	82	49	1.26E-05	99.999%	79,433	35,981
S	83	50	1.00E-05	99.999%	100,000	28,580
T	84	51	7.94E-06	99.999%	125,893	22,702
U	85	52	6.31E-06	99.999%	158,489	18,033
V	86	53	5.01E-06	99.999%	199,526	14,324
W	87	54	3.98E-06	100.000%	251,189	11,378
X	88	55	3.16E-06	100.000%	316,228	9,038
Y	89	56	2.51E-06	100.000%	398,107	7,179
Z	90	57	2.00E-06	100.000%	501,187	5,703
[	91	58	1.58E-06	100.000%	630,957	4,530
\	92	59	1.26E-06	100.000%	794,328	3,598
]	93	60	1.00E-06	100.000%	1,000,000	2,858
^	94	61	7.94E-07	100.000%	1,258,925	2,270
_	95	62	6.31E-07	100.000%	1,584,893	1,803
`	96	63	5.01E-07	100.000%	1,995,262	1,432
a	97	64	3.98E-07	100.000%	2,511,886	1,138
b	98	65	3.16E-07	100.000%	3,162,278	904
c	99	66	2.51E-07	100.000%	3,981,072	718
d	100	67	2.00E-07	100.000%	5,011,872	570
e	101	68	1.58E-07	100.000%	6,309,573	453
f	102	69	1.26E-07	100.000%	7,943,282	360
g	103	70	1.00E-07	100.000%	10,000,000	286
h	104	71	7.94E-08	100.000%	12,589,254	227
i	105	72	6.31E-08	100.000%	15,848,932	180
j	106	73	5.01E-08	100.000%	19,952,623	143
k	107	74	3.98E-08	100.000%	25,118,864	114
l	108	75	3.16E-08	100.000%	31,622,777	90
m	109	76	2.51E-08	100.000%	39,810,717	72
n	110	77	2.00E-08	100.000%	50,118,723	57
o	111	78	1.58E-08	100.000%	63,095,734	45
p	112	79	1.26E-08	100.000%	79,432,823	36
q	113	80	1.00E-08	100.000%	100,000,000	29
r	114	81	7.94E-09	100.000%	125,892,541	23
s	115	82	6.31E-09	100.000%	158,489,319	18
t	116	83	5.01E-09	100.000%	199,526,231	14
u	117	84	3.98E-09	100.000%	251,188,643	11
v	118	85	3.16E-09	100.000%	316,227,766	9
w	119	86	2.51E-09	100.000%	398,107,171	7
x	120	87	2.00E-09	100.000%	501,187,234	6
y	121	88	1.58E-09	100.000%	630,957,344	5
z	122	89	1.26E-09	100.000%	794,328,235	4
{	123	90	1.00E-09	100.000%	1,000,000,000	3
|	124	91	7.94E-10	100.000%	1,258,925,412	2
}	125	92	6.31E-10	100.000%	1,584,893,192	2
~	126	93	5.01E-10	100.000%	1,995,262,315	1

DNA Level

Whole genome resequencing (WGS)
- Discover the genetic variations in a genome-wide range.

Exome Seq
- Discover the causative, susceptibility loci
- Discover rare/novel variants
- More economical and efficient

Target Region Seq
- Find the novel variants or validate the candidate variants in the target regions

Genotyping
- SNP and CNV detection in a genome-wide range
- Customized array for personal usage which is more flexible
- Validation of candidate pathogenetic genes or loci in large amount of samples

Single Cell Seq
- Genetic variation research at single cell level
- Explore cancer cells evolution during tumor progression

RNA Level

Transcriptome Seq
- Comprehensive analysis of differential gene expression
- Discover novel genes
- RNA editing analysis(such as alternative splicing, cSNP, gene fusion, etc)

RNA-Seq（Quantification）
- Precise quantification of gene expression analysis that is suitable for large samples
- Discover disease-related functional genes

Small RNA Seq
- Gene expression analysis of miRNA
- Gene regulatory networks and targets study of mi RNA
- Discover disease-specific biomarkers

Non-coding RNA Seq
- Identify novel non-coding RNA
- Discover disease-specific biomarkers

Cell Line Seq
- Obtain a clear and comprehensive genetic patterns of the cell lines
- Obtain mutation information of high accuracy

Epigenetic Level

Whole Genome Bisulfite Seq (WGBS)
- DNA methylation research at whole genome-wide level
- High accuracy and high resolution(single-based)

MeDIP Seq
- Based on immunoprecipitation for methylated DNA enrichment
- Whole genome-wide DNA methylation research and cost-effective

RRBS Seq
- Methylation analysis of promoter regions with substantial genome coverage
- Based on enzyme digestion and bisulfite treatment
- Good repeatability

ChIP Seq
- Genome-wide protein-DNA interaction studies
- Higher resolution, more precise and abundant than ChIP-chip

Protein Level

Proteome Profiling
- Analyze the component of protein mixtures
- Obtain comprehensive information of protein category, metabolic pathways, etc

Quantitative Proteomics
- Fast and accurate protein differential analysis for multiple samples

Modification Proteomics
- Fast and comprehensive analysis of protein modification spectrum for multiple samples

Target Proteomics
- Based on the technology of Multiple Reaction Monitoring(MRM)
- Validate the discovered biomarkers
- Identify protein modification and low abundant proteins