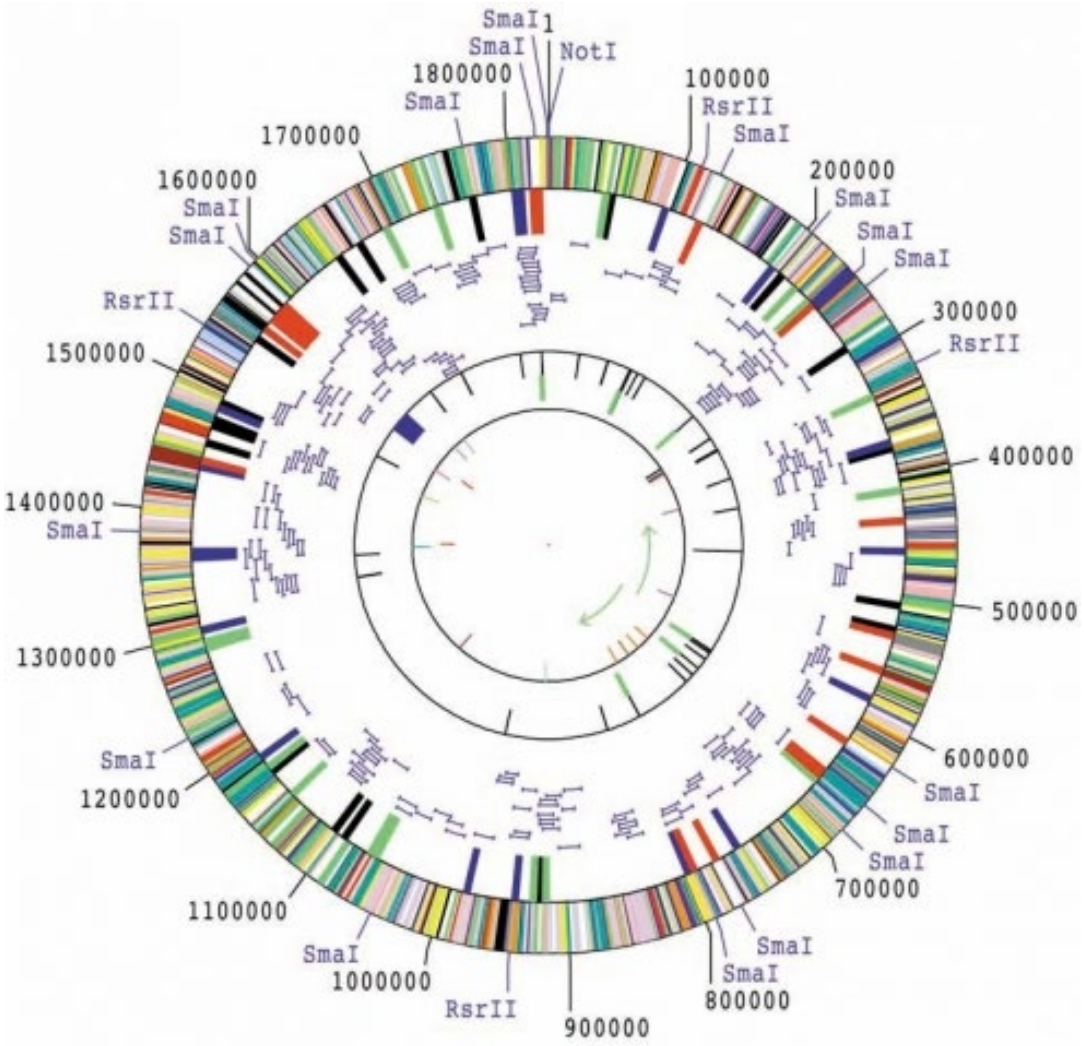
DNA sequencing, accuracy and errors

Linda van der Graaf – van Bloois

Faculty of Veterinary Medicine

Department of Biomolecular Health Sciences

First microbial genome



Fleischmann et al Science. 1995 Jul 28;269(5223):496-512.

1995

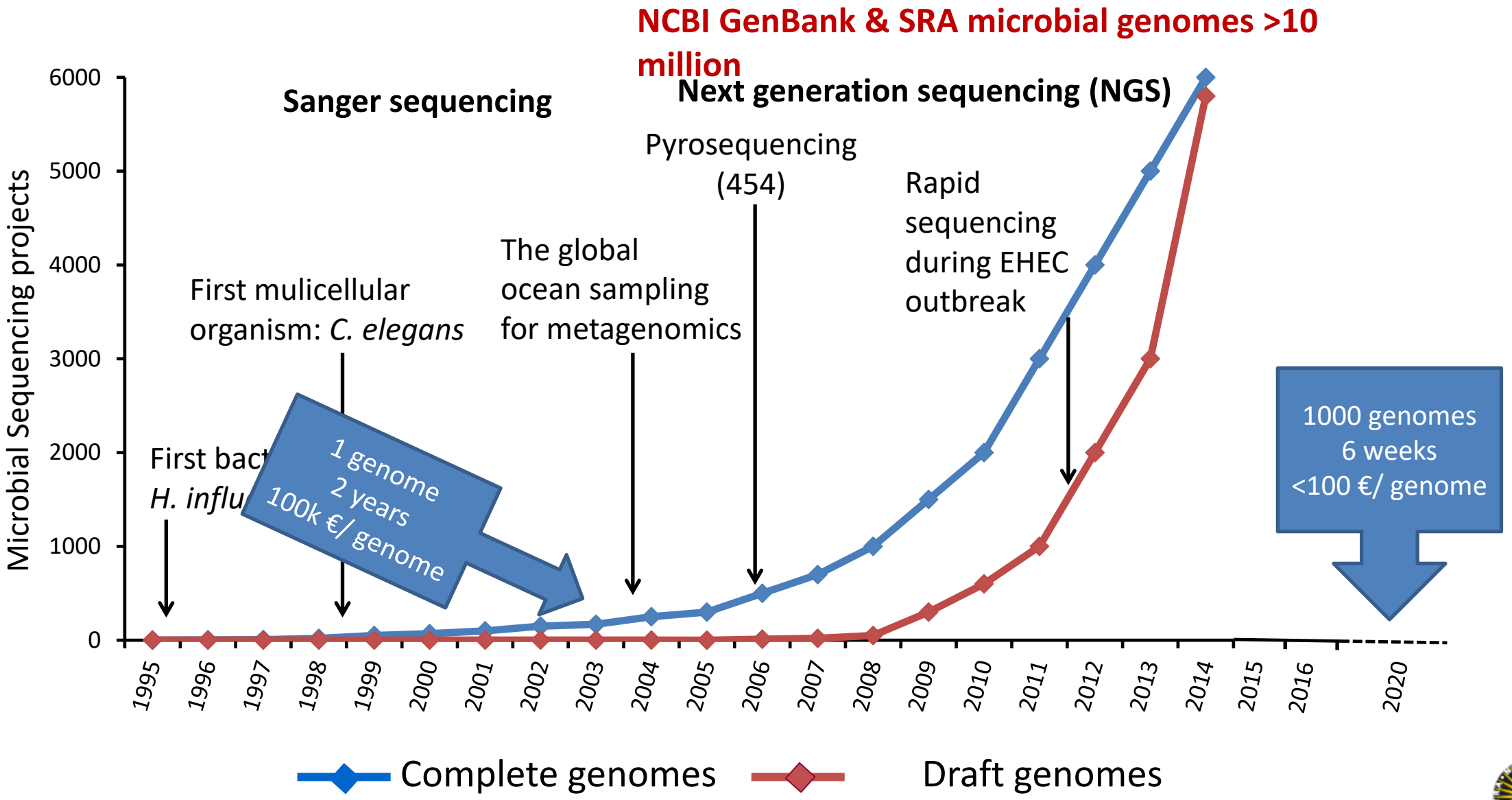
Entirely done by Sanger sequencing

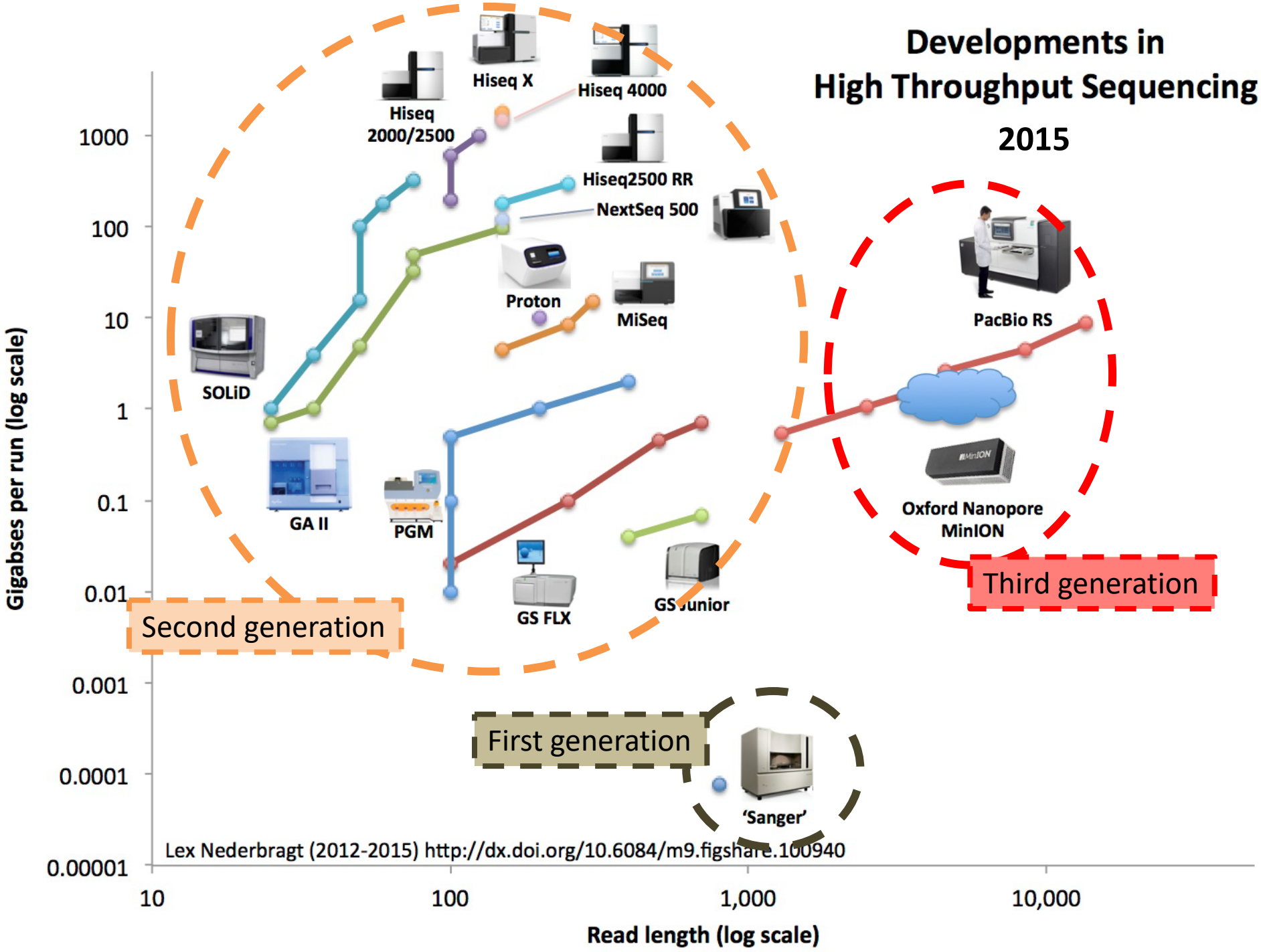
Estimated cost: $0.48/base

> 1M€ per genome in todays money

Haemophilus influenzae

Microbial sequencing projects




Developments in
High Throughput Sequencing
2015
Hiseq X
Hiseq 4000
Hiseq
2000/2500
Hiseq2500 RR
NextSeq 500
Proton
MiSeq
SOLiD
GA II
PGM
GS FLX
GS Junior
PacBio RS
Oxford Nanopore
MinION
Second generation
Third generation
First generation
'Sanger'
Lex Nederbragt (2012-2015) http://dx.doi.org/10.6084/m9.figshare.100940
Gigabses per run (log scale)
1000
100
10
1
0.1
0.01
0.001
0.0001
0.00001
10
100
1,000
10,000
Read length (log scale)

Second generation (Short-read) sequencing methods

Roche 454

Discontinued

1st next-generation system: pyrosequencing

IonTorrent

Ion Gene S5

Ion PGM

Ion Proton

cheap, fast, sequence errors, 200-400 bp reads

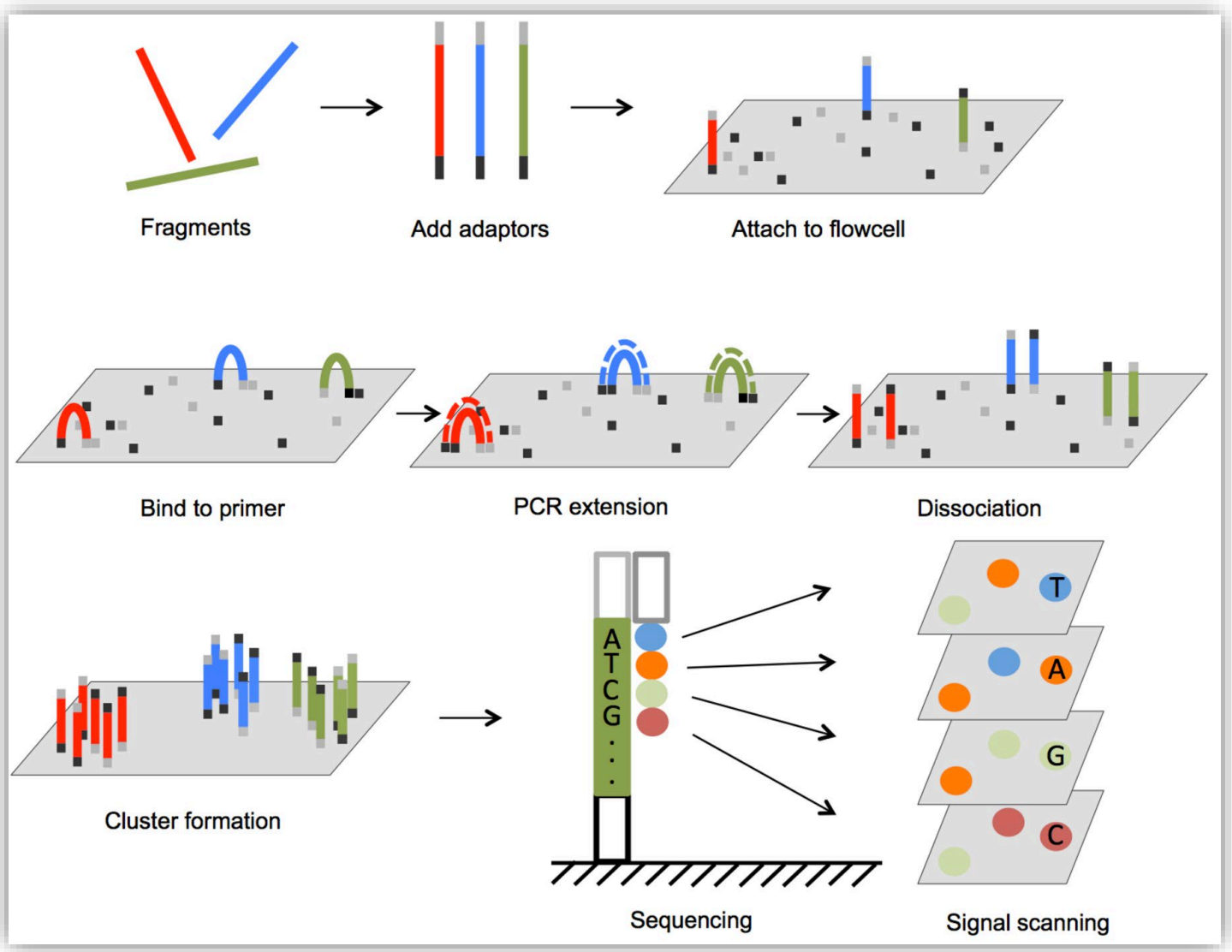
Illumina

ISeq – MiniSeq - MiSeq

NextSeq – HiSeq – HiSeqX - NovaSeq

cheap, precise, slow, high-throughput, 50-300 bp reads

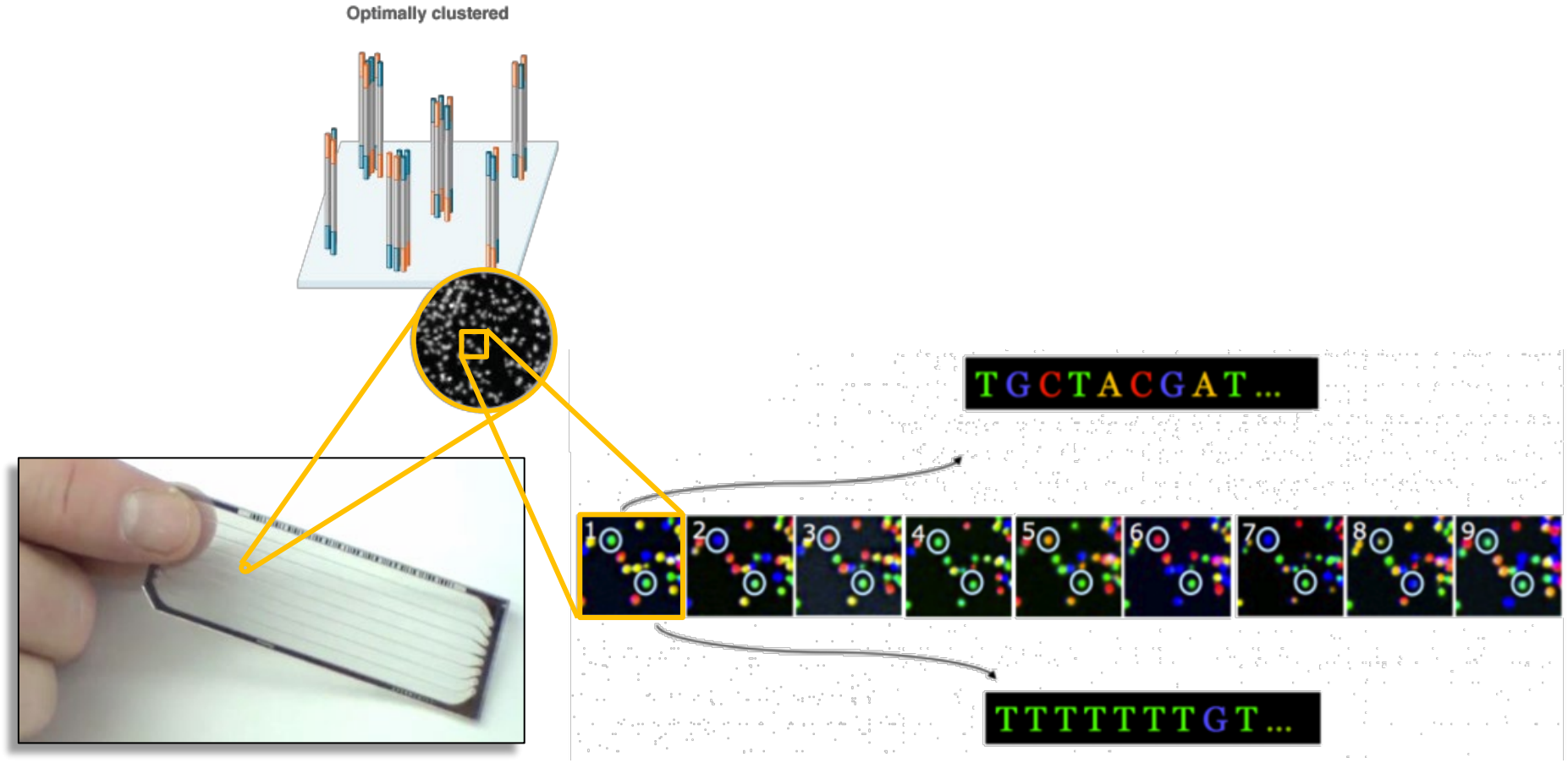
Illumina sequencing



DNA sequencing depends on "reading" A/C/G/T signal

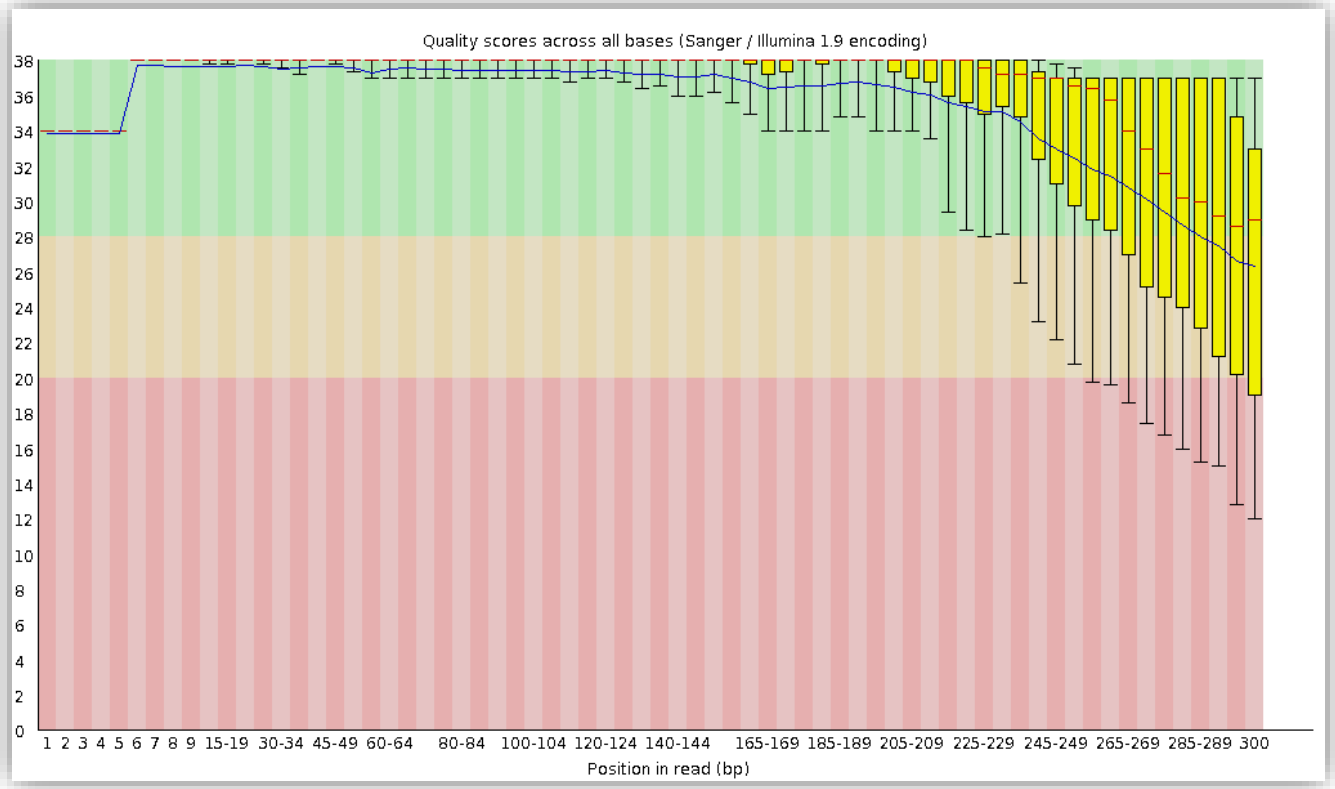
Signals: differently colored fluorophores

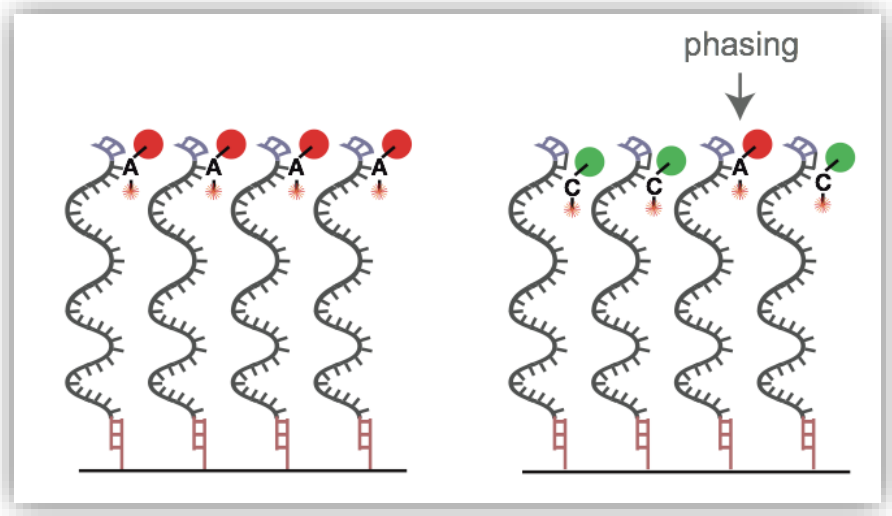
Illumina sequencing



Color signal of a cluster is not always 100% unambiguous

Illumina sequencing – quality scores bases





Phasing:

- The blocker of a nucleotide is not correctly removed after signal detection

- A nucleotide has a defect terminator cap (prephasing) and two nucleotides can bind in one cycle

Main reason for decreasing sequence quality:

- Phasing
- Decrease of fluorescent signal over time

Quality scores

- Phred 10: 10^{-1} chance that the base is wrong
  - 90% accuracy; 10% error rate
- Phred 20: 10^{-2} chance that the base is wrong
  - 99% accuracy ; 1% error rate
- Phred 30: 10^{-3} chance that the base is wrong
  - 99.9% accuracy ; 0.1% error rate

Phred Quality Score	Probability of incorrect base call	Base call accuracy
10	1 in 10	90%
20	1 in 100	99%
30	1 in 1000	99.9%

Phred score in ASCII text

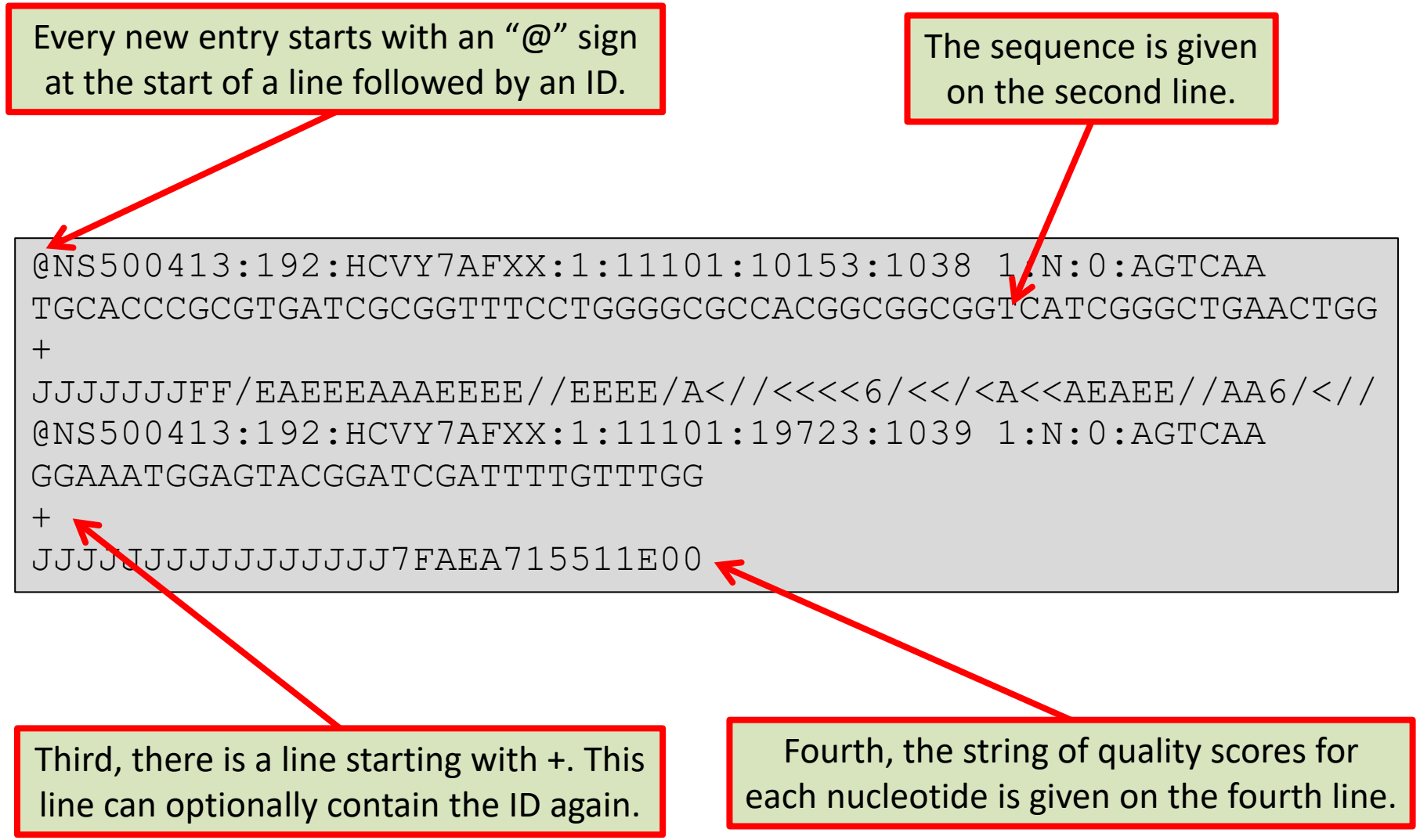
Fastq quality score = Phred score -> converted to ASCII text

ASCII TABLE

Decimal	Hex	Char	Decimal	Hex	Char	Decimal	Hex	Char	Decimal	Hex	Char
0	0	[NULL]	32	20	[SPACE]	64	40	@	96	60	`
1	1	[START OF HEADING]	33	21	!	65	41	A	97	61	a
2	2	[START OF TEXT]	34	22	"	66	42	B	98	62	b
3	3	[END OF TEXT]	35	23	#	67	43	C	99	63	c
4	4	[END OF TRANSMISSION]	36	24	$	68	44	D	100	64	d
5	5	[ENQUIRY]	37	25	%	69	45	E	101	65	e
6	6	[ACKNOWLEDGE]	38	26	&	70	46	F	102	66	f
7	7	[BELL]	39	27	'	71	47	G	103	67	g
8	8	[BACKSPACE]	40	28	(	72	48	H	104	68	h
9	9	[HORIZONTAL TAB]	41	29	)	73	49	I	105	69	i
10	A	[LINE FEED]	42	2A	*	74	4A	J	106	6A	j
11	B	[VERTICAL TAB]	43	2B	+	75	4B	K	107	6B	k
12	C	[FORM FEED]	44	2C	,	76	4C	L	108	6C	l
13	D	[CARRIAGE RETURN]	45	2D	-	77	4D	M	109	6D	m
14	E	[SHIFT OUT]	46	2E	.	78	4E	N	110	6E	n
15	F	[SHIFT IN]	47	2F	/	79	4F	O	111	6F	o
16	10	[DATA LINK ESCAPE]	48	30	0	80	50	P	112	70	p
17	11	[DEVICE CONTROL 1]	49	31	1	81	51	Q	113	71	q
18	12	[DEVICE CONTROL 2]	50	32	2	82	52	R	114	72	r
19	13	[DEVICE CONTROL 3]	51	33	3	83	53	S	115	73	s
20	14	[DEVICE CONTROL 4]	52	34	4	84	54	T	116	74	t
21	15	[NEGATIVE ACKNOWLEDGE]	53	35	5	85	55	U	117	75	u
22	16	[SYNCHRONOUS IDLE]	54	36	6	86	56	V	118	76	v
23	17	[END OF TRANS. BLOCK]	55	37	7	87	57	W	119	77	w
24	18	[CANCEL]	56	38	8	88	58	X	120	78	x
25	19	[END OF MEDIUM]	57	39	9	89	59	Y	121	79	y
26	1A	[SUBSTITUTE]	58	3A	:	90	5A	Z	122	7A	z
27	1B	[ESCAPE]	59	3B	;	91	5B	[	123	7B	{
28	1C	[FILE SEPARATOR]	60	3C	<	92	5C	\	124	7C	|
29	1D	[GROUP SEPARATOR]	61	3D	=	93	5D	]	125	7D	}
30	1E	[RECORD SEPARATOR]	62	3E	>	94	5E	^	126	7E	~
31	1F	[UNIT SEPARATOR]	63	3F	?	95	5F	_	127	7F	[DEL]

Fastq files

- Sequence and quality scores are stored in Fastq files



Phred scores in fastq files

```
@NS500413:192:HCVY7AFXX:1:111(
TGCACCCGCGTGATCGCGGTTTCCTGGGGC
+
JJJJJJJFF/EAEEEAAAEEEE//EEEE/A
@NS500413:192:HCVY7AFXX:1:111(
GGAAATGGAGTACGGATCGATTTTGTTTGC
+
JJJJJJJJJJJJJJJJJ7FAEA715511E0(
```

Phred score per base:

J = 74

74 – 33 = 41 → phred score

ASCII TABLE

Decimal	Hex	Char	Decimal	Hex	Char	Decimal	Hex	Char	Decimal	Hex	Char
0	0	[NULL]	32	20	[SPACE]	64	40	@	96	60	`
1	1	[START OF HEADING]	33	21	!	65	41	A	97	61	a
2	2	[START OF TEXT]	34	22	"	66	42	B	98	62	b
3	3	[END OF TEXT]	35	23	#	67	43	C	99	63	c
4	4	[END OF TRANSMISSION]	36	24	$	68	44	D	100	64	d
5	5	[ENQUIRY]	37	25	%	69	45	E	101	65	e
6	6	[ACKNOWLEDGE]	38	26	&	70	46	F	102	66	f
7	7	[BELL]	39	27	'	71	47	G	103	67	g
8	8	[BACKSPACE]	40	28	(	72	48	H	104	68	h
9	9	[HORIZONTAL TAB]	41	29	)	73	49	I	105	69	i
10	A	[LINE FEED]	42	2A	*	74	4A	J	106	6A	j
11	B	[VERTICAL TAB]	43	2B	+	75	4B	K	107	6B	k
12	C	[FORM FEED]	44	2C	,	76	4C	L	108	6C	l
13	D	[CARRIAGE RETURN]	45	2D	-	77	4D	M	109	6D	m
14	E	[SHIFT OUT]	46	2E	.	78	4E	N	110	6E	n
15	F	[SHIFT IN]	47	2F	/	79	4F	O	111	6F	o
16	10	[DATA LINK ESCAPE]	48	30	0	80	50	P	112	70	p
17	11	[DEVICE CONTROL 1]	49	31	1	81	51	Q	113	71	q
18	12	[DEVICE CONTROL 2]	50	32	2	82	52	R	114	72	r
19	13	[DEVICE CONTROL 3]	51	33	3	83	53	S	115	73	s
20	14	[DEVICE CONTROL 4]	52	34	4	84	54	T	116	74	t
21	15	[NEGATIVE ACKNOWLEDGE]	53	35	5	85	55	U	117	75	u
22	16	[SYNCHRONOUS IDLE]	54	36	6	86	56	V	118	76	v
23	17	[END OF TRANS. BLOCK]	55	37	7	87	57	W	119	77	w
24	18	[CANCEL]	56	38	8	88	58	X	120	78	x
25	19	[END OF MEDIUM]	57	39	9	89	59	Y	121	79	y
26	1A	[SUBSTITUTE]	58	3A	:	90	5A	Z	122	7A	z
27	1B	[ESCAPE]	59	3B	;	91	5B	[	123	7B	{
28	1C	[FILE SEPARATOR]	60	3C	<	92	5C	\	124	7C	|
29	1D	[GROUP SEPARATOR]	61	3D	=	93	5D	]	125	7D	}
30	1E	[RECORD SEPARATOR]	62	3E	>	94	5E	^	126	7E	~
31	1F	[UNIT SEPARATOR]	63	3F	?	95	5F	_	127	7F	[DEL]

Fastq files

The ID may contain information such as the run and the position of the cluster on the sequencing slide.

Paired sequence and quality lines should be equally long.

```
@NS500413:192:HCVY7AFXX:1:11101:10153:1038 1:N:0:AGTCAA
TGCACCCGCGTGATCGCGGTTTCCTGGGGCGCCACGGCGGCGGTCATCGGGCTGAACTGG
+
JJJJJJJFF/EAEEEAAAEEEE//EEEE/A<//<<<<6/<</<A<<AEAEE//AA6/<//
@NS500413:192:HCVY7AFXX:1:11101:19723:1039 1:N:0:AGTCAA
GGAAATGGAGTACGGATCGATTTTGTTTGG
+
JJJJJJJJJJJJJJJJJ7FAEA715511E00
```

Illumina reads are all the same length when they come off the sequencer. If the sequences have different lengths, they have probably been processed.

Sometimes the sample barcode is also listed.

Fasta files

- Sequences are stored in Fasta files
- Fasta files are plain text files (open e.g. in 

)

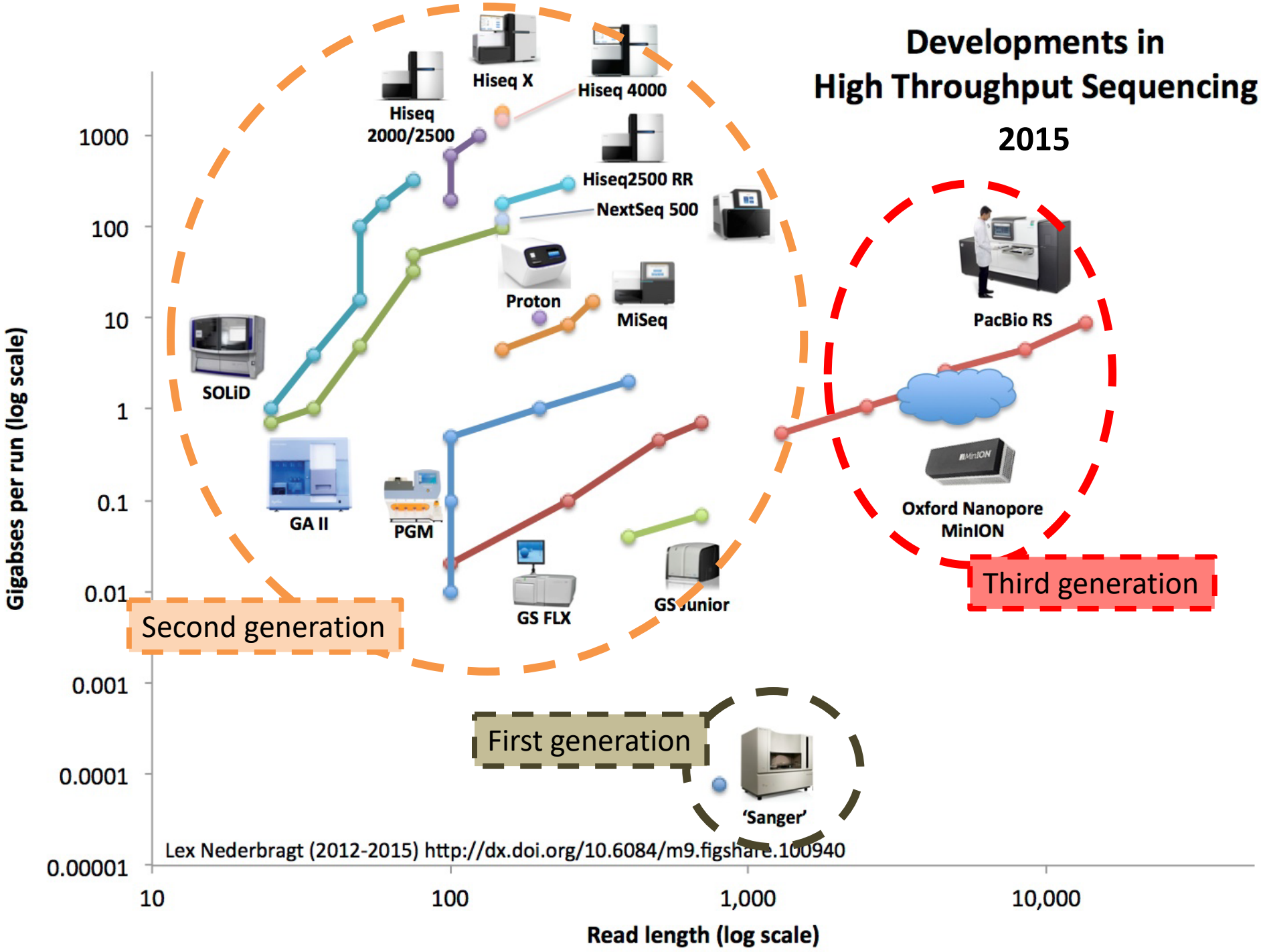
Every new entry starts with a “>” sign at the start of a line followed by an ID

Each ID is unique in the file. The ID is the part until a space, so no spaces allowed in the ID itself

There can be extra attributes or comments on the same line, but after the first space

```
>dna_sequence_A length=36 source=human_gut
CCGATCATATGACTAGCATGTCGACTAGCATTTAGA
>dna_sequence_B length=43 source=human_gut
GACTAGCATGCAATCGCGATAGCTATCGACTAGCATTTAGCGA
>dna_sequence_C length=86 source=human_gut
GCGGCGGCTAACGCATCGATCTTTGTACGATGATTGGCGGCGGC
TATTATGCATTGGGAATGCATCGATCGACATCGATCGAAGCTAT
```

While the description has to be on a single line, the sequence can be on one or more lines until the next “>” at the start of a line

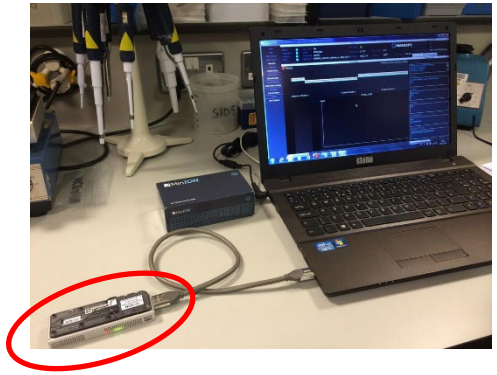

Developments in
High Throughput Sequencing
2015
Hiseq X
Hiseq 4000
Hiseq
2000/2500
Hiseq2500 RR
NextSeq 500
Proton
MiSeq
SOLiD
GA II
PGM
GS FLX
GS Junior
PacBio RS
Oxford Nanopore
MinION
Second generation
Third generation
First generation
'Sanger'
Lex Nederbragt (2012-2015) http://dx.doi.org/10.6084/m9.figshare.100940
Gigabses per run (log scale)
1000
100
10
1
0.1
0.01
0.001
0.0001
0.00001
10
100
1,000
10,000
Read length (log scale)

Third generation (Long-read) sequencing methods

Pacific Biosystems
(PacBio)

Long reads (20 kb)
Perfect for genome assembly
Moderate to high error rate

Oxford Nanopore
Technologies



Tiny (minION)
Long reads (>50 kb)
Perfect for genome assembly
High error rate

GridION

PromethION

SmidgION

Oxford Nanopore

i. Open channel

ii. DNA with molecular motor captured by nanopore

iii. Translocation of 5’ adaptor

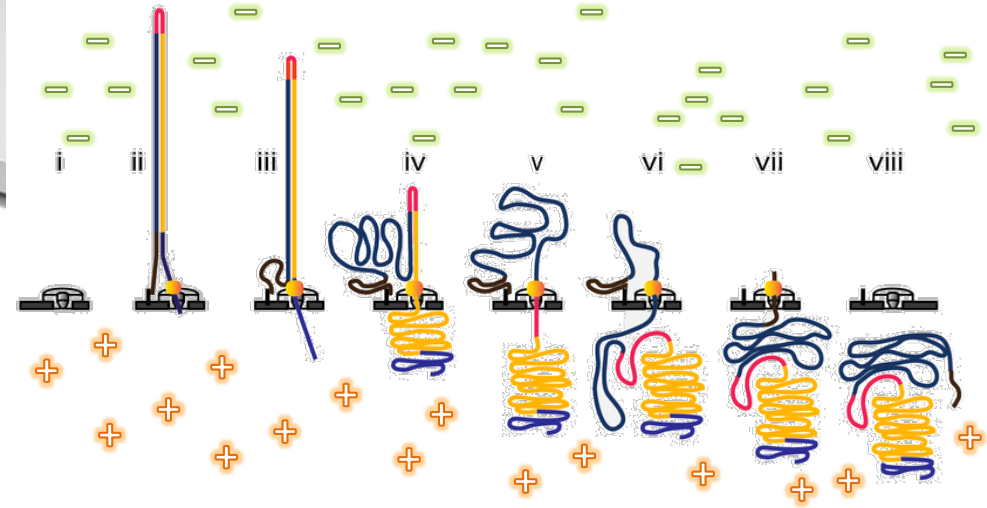
iv. Translocation of template strand

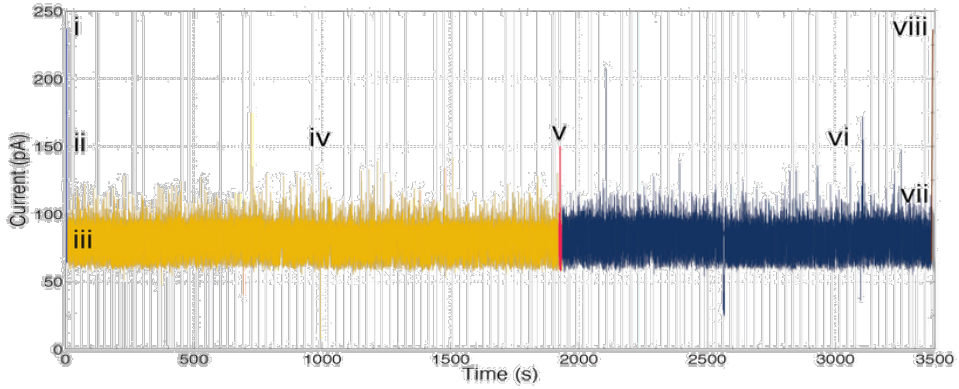
v. Translocation of hairpin

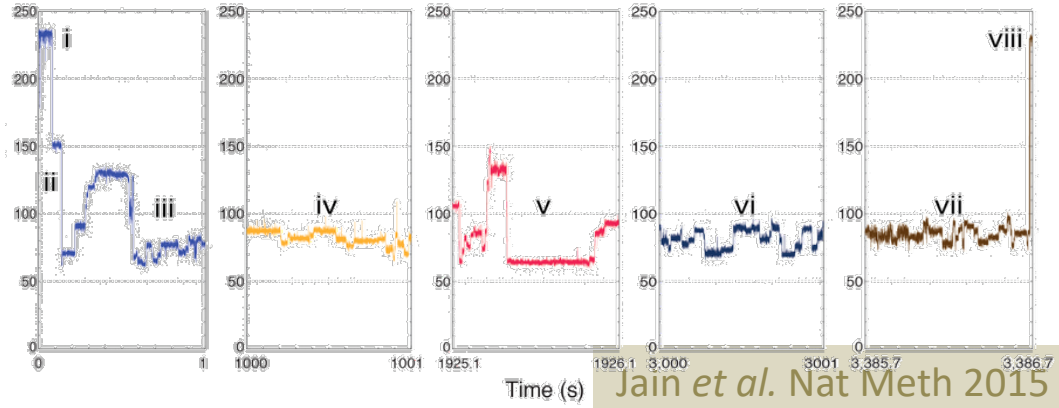
vi. Translocation of complement strand

vii. Translocation of 3’ adaptor

viii. Open channel

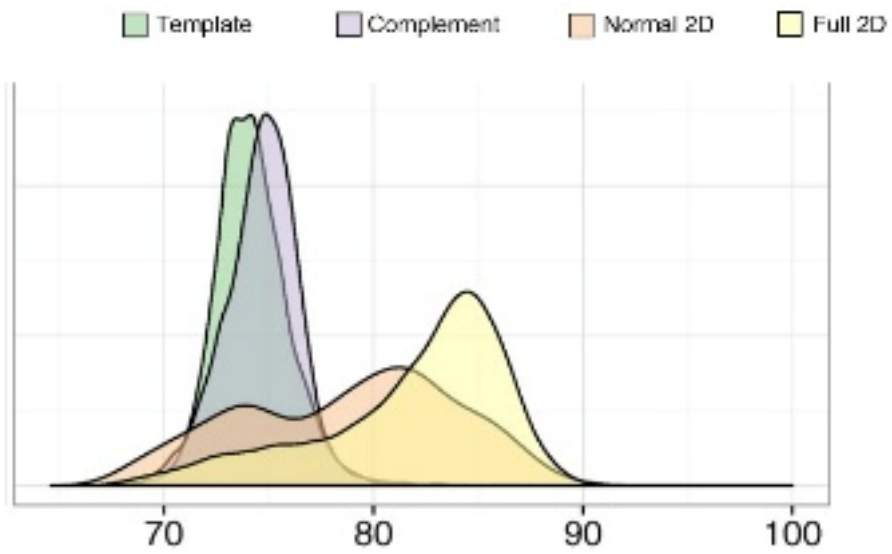


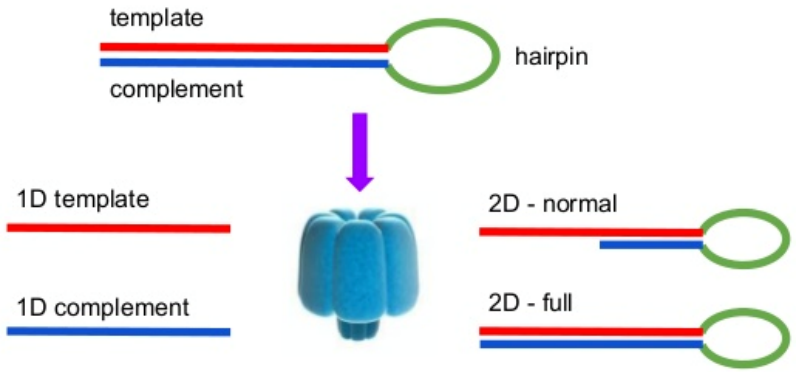




Jain *et al.* Nat Meth 2015

Nanopore accuracy

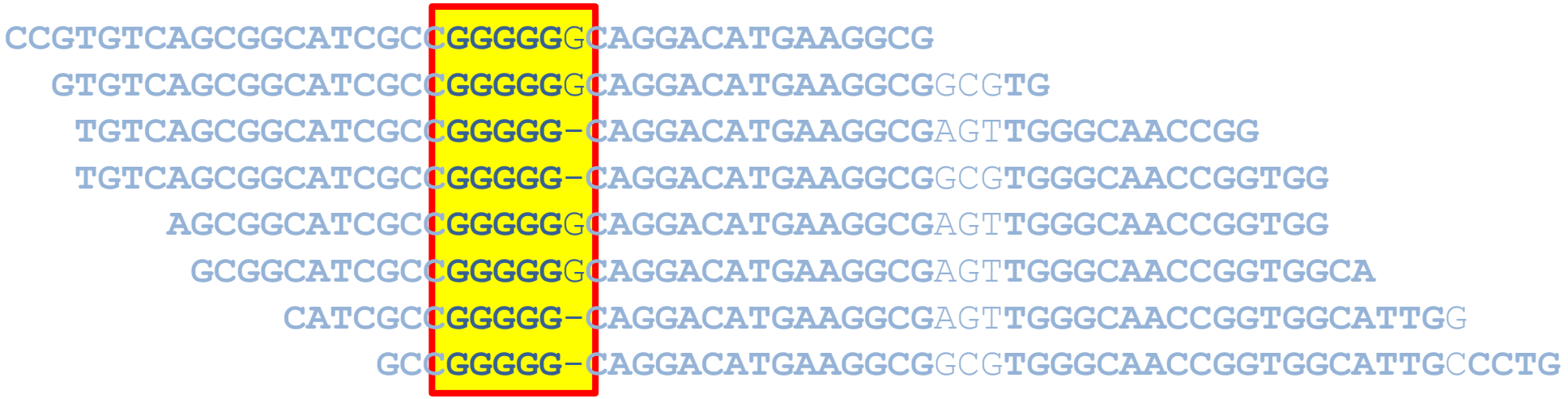




- By themselves, Nanopore sequences have relatively low accuracy (1D)
- By measuring double stranded DNA, the same sequence can be measured twice (2D)
- This decreases the error rate / increases accuracy

Seemann, SlideShare 2015

Homopolymer errors

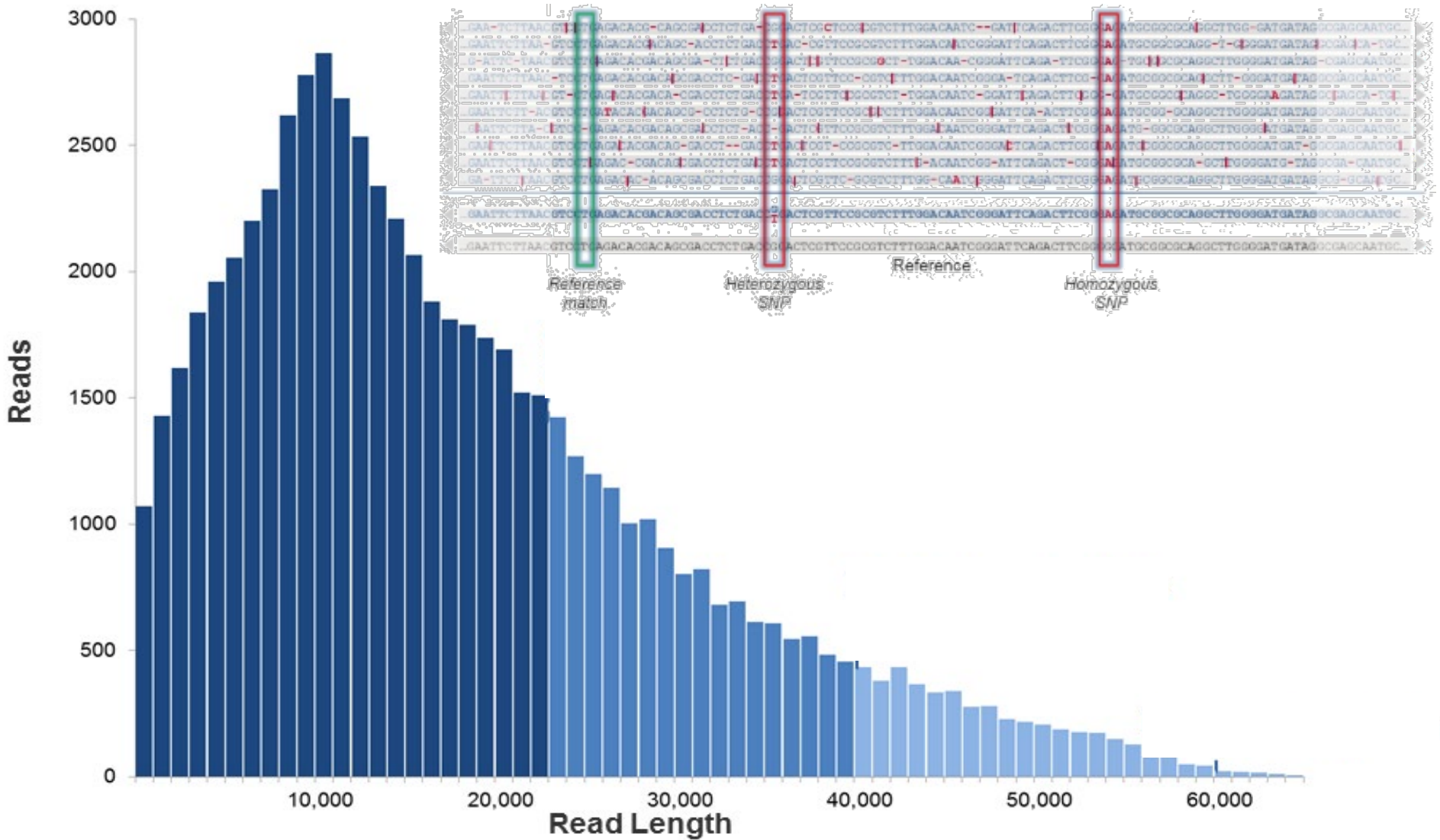


- Homopolymer: stretch of identical nucleotides
- Homopolymer error: length is misidentified
  - When longer than 3-5 nucleotides
  - Problem with long-read sequencing

- Longer homopolymer tracks: more errors

Consensus correction

- Long reads (Nanopore, Pacbio) can be corrected with shorter reads, or with contigs (e.g. Illumina)



Summary

- Second generation sequencing (Illumina)
- Sequencing errors
- Phred scores, fastq files
- Third generation sequencing (Nanopore)
- Homopolymer errors
- Consensus correction