Next-generation sequencing

Lecture 3

Homework Assignment (1)

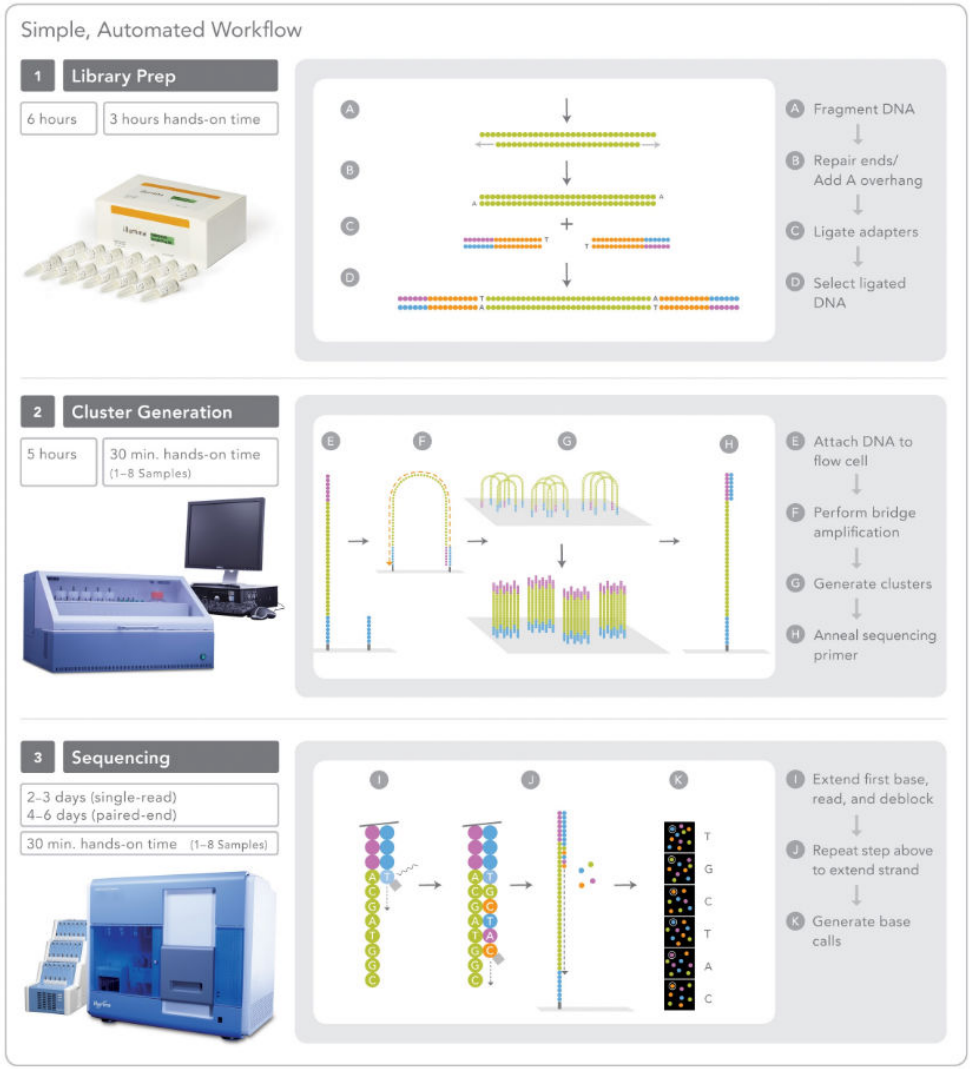
- Do you have a computer? Yes (all)
- If yes, what operating system does this computer have? Windows and Mac OS
- Do you have permission to install a software on this computer? Yes (all)
- Do you have experience in programming? No (most)
- Do you know how to use R? No (most)
- If no, how difficult is it for you to master this software? Positive (all)

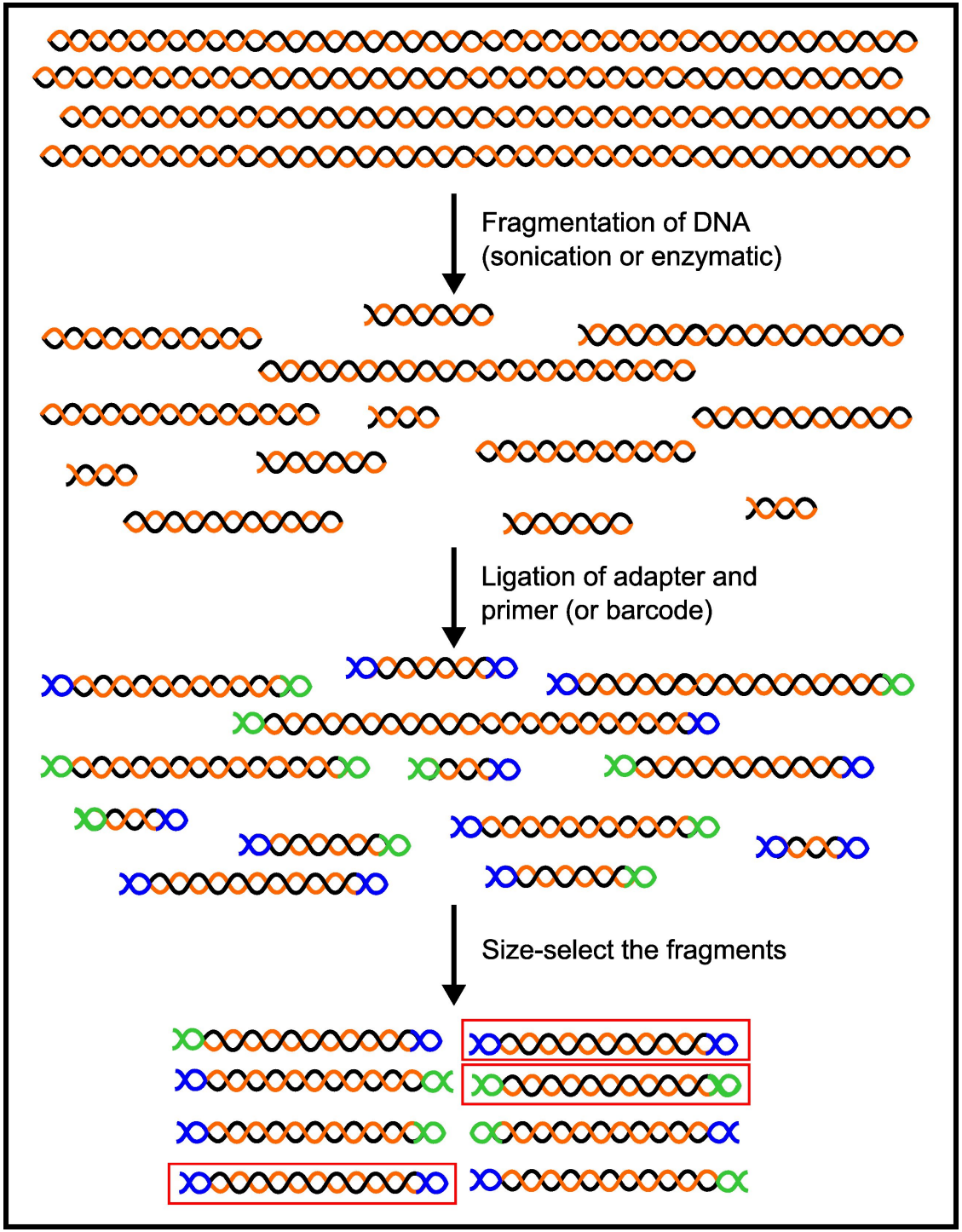
NGS

- Introduction to the background
- NGS workflow and accuracy
- Data format
- Assembly
- RNA-seq
  - Aligner
  - Analysis tools
  - Applications, such as MiRNA
- Chip-seq
  - Applications

Work flow

- Library preparation: fragmenting, end polish, ligation of adaptors, size selection.
- Amplification: emPCR and solid phase amplification.
- Sequencing and imaging
- Data analysis





1. fragmenting the DNA (sonication, nebulization, or shearing)
2. DNA repair and end polishing (blunt end, phosphorylated end that is ready for ligation)
3. platform-specific adaptor ligation.
4. Size-select

	Roche/454	SOLiD	Hi-Seq 2000	Pacific Biosci RS
Amplification	emPCR on bead surface	emPCR on bead surface	Enzymatic amplification on glass surface	NA
Sequencing	Pyrosequencing, Polymerase-mediated incorporation of unlabelled nucleotides	Sequencing by ligation, Ligase-mediated addition of 2-base encoded fluorescent oligonucleotides	Cyclic reversible termination, Polymerase-mediated incorporation of end-blocked fluorescent nucleotides	Real time sequencing. Polymerase-mediated incorporation of terminal phosphate labelled fluorescent nucleotides
Detection	Light emitted from secondary reactions initiated by release of PPi	Fluorescent emission from ligated dye-labelled oligonucleotides	Fluorescent emission from incorporated dye-labelled nucleotides	Real time detection of fluorescent dye in polymerase active site during incorporation
Error model	Substitution errors rare, insertion/ deletion errors at homopolymers	End of read substitution errors	End of read substitution errors	Random insertion/ deletion errors
Read length	400 bp	75 bp	150 bp	>1,000 bp

Mardis, Nature 2011

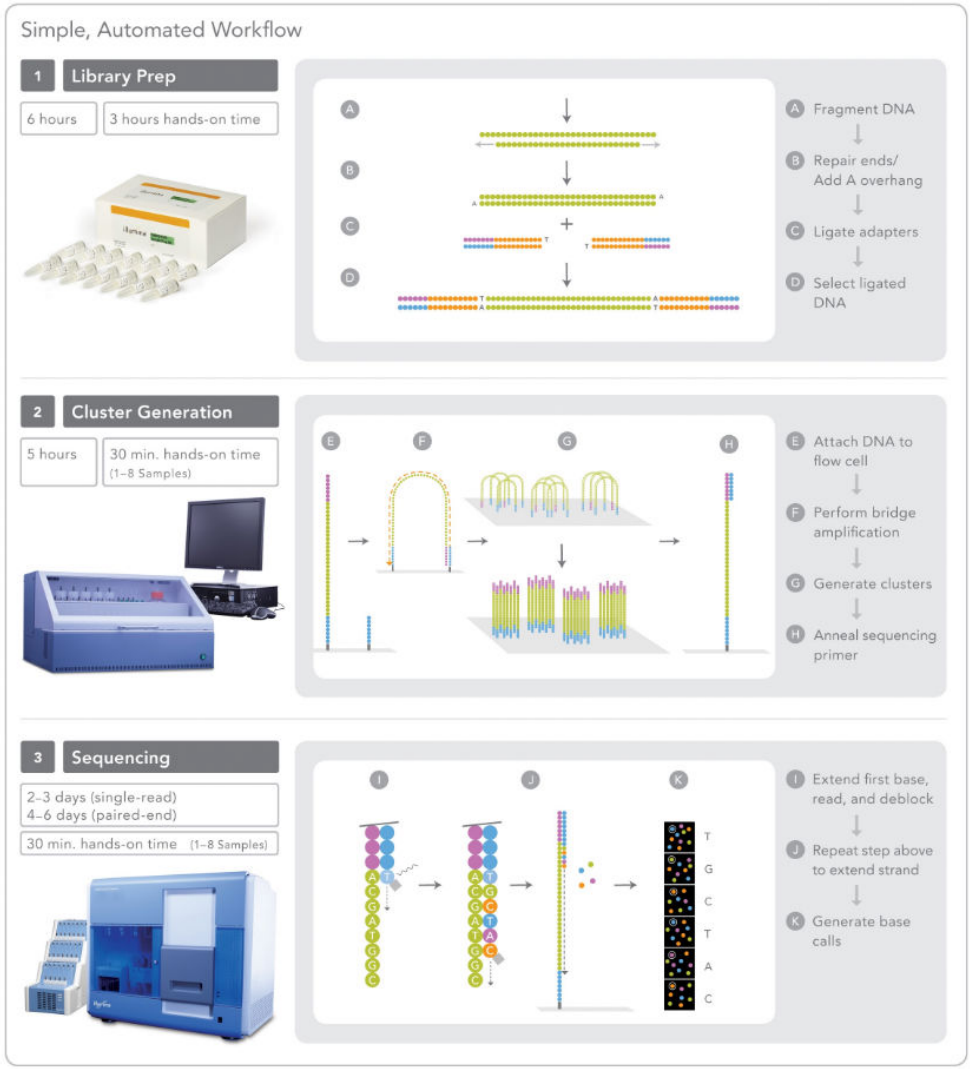
Accuracy

- base quality drops along read
  Sanger > SOLiD > Illumina > 454 > Helicos

- Issue for Roche 454:
  39% of errors are homopolymers

Work flow

- Library preparation: fragmenting, end polish, ligation of adaptors, size selection.
- Amplification: emPCR and solid phase amplification.
- Sequencing and imaging
- Data analysis



NGS

- Introduction to the background
- NGS workflow and accuracy
- Data format, quality control, data management
- Assembly
- RNA-seq
  - Aligner
  - Analysis tools
  - Applications, such as MiRNA
- Chip-seq
  - Applications

Data format: fastq

- FASTQ format was originally developed at the Wellcome Trust Sanger Institute, and used for Sanger Method.
- FASTQ is a text-based format for storing both a biological sequence in a plain text file.
- FASTQ has recently become the *de facto* standard for storing the output of high throughput sequencing instruments.
- FASTQ has both the sequence and an associated per base quality score.
- Both the sequence letter and quality score are encoded with a single ASCII character for brevity.
- The FASTQ format has become widely used as a simple interchange file format.
- Lacks any formal definition to date, and exists in some incompatible variants.

Data format: fastq

Example of one read (Illumina):

```
@HWI-EAS209_0006_FC706VJ:5:58:5894:21141#ATCACG/1
TTAATTGGTAAATAAATCTCCTAATAGCTTAGATNTTACCTTNNNNNNNNNNTAGTTTCTTGAGATTTGTT
+HWI-EAS209_0006_FC706VJ:5:58:5894:21141#ATCACG/1
efcfffffcfeefffcffffffddf`feed]`]_Ba_^__[YBBBBBBBBBBRTT\]][]dddd`ddd^dd
```

Line 1: “@” + identifier

Line 2: sequence

Line 3: “+” + identifier (optional)

Line 4: phred-based quality scores

Data format: fastq

- '@' title line. This is a free format field with no length limit, and allowing arbitrary annotation or comments to be included.
- the sequence line(s). Similar in the FASTA format, it can be line wrapped. No explicit limitation on the characters expected. White space such as tabs or spaces is NOT permitted.
- '+' line. Originally this is a full repeat of the title line text. however, by common usage, this is optional and the '+' line can contain just this one character, reducing the file size significantly.
- quality line(s). Like the seq lines, it can be wrapped. These use a subset of the ASCII printable characters to represent sequence quality. The quality string must be equal in length to the sequence string.

Illumina sequence identifiers

@HWUSI-EAS100R:6:73:941:1973#0/1

HWUSI-EAS100R	the unique instrument name
6	flowcell lane
73	tile number within the flowcell lane
941	'x'-coordinate of the cluster within the tile
1973	'y'-coordinate of the cluster within the tile
#0	index number for a multiplexed sample (0 for no indexing)
/1	the member of a pair, /1 or /2 (paired-end or mate-pair reads only)

NCBI format

```
@SRR001666.1 071112_SLXA-EAS1_s_7:5:1:817:345 length=36
GGGTGATGGCCGCTGCCGATGGCGTCAAATCCCACC
+SRR001666.1 071112_SLXA-EAS1_s_7:5:1:817:345 length=36
IIIIIIIIIIIIIIIIIIIIIIIIIIIIII9IG9IC
```

- Sequence read achieve ID
- holds the original identifier from sequencer
- the read length

Example (1)

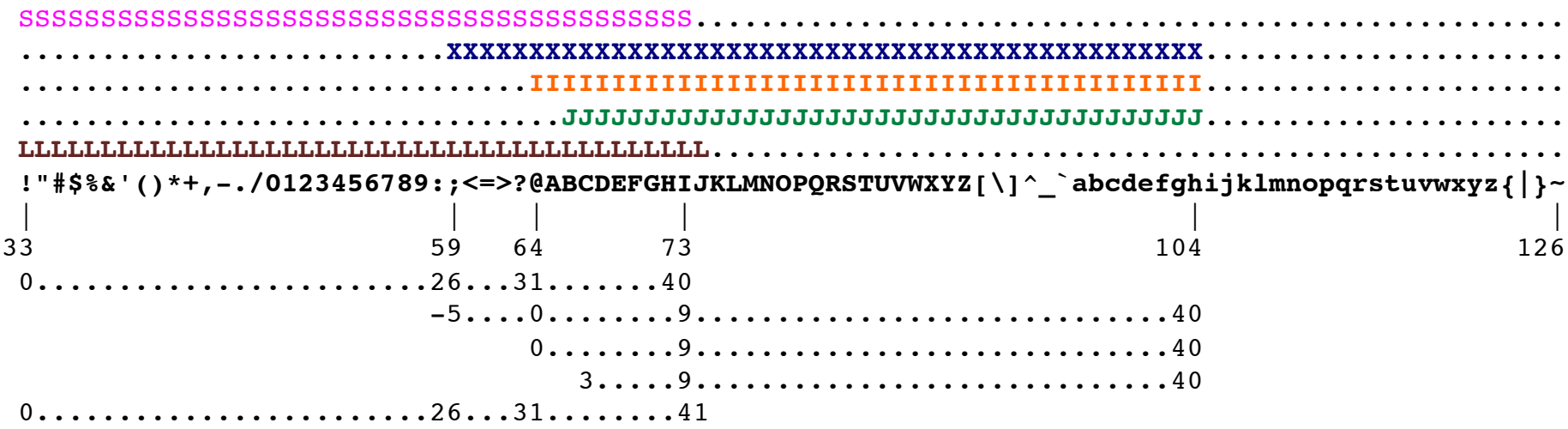
@FAKE0007 Original version has lower case unambiguous DNA with PHRED scores from 0 to 40 inclusive (in that order)
ACGTACGTACGTACGTACGTACGTACGTACGTACGTACGTA
+
!"#$%&'()*+,-./0123456789:;<=>?@ABCDEFGHI
@FAKE0008 Original version has mixed case unambiguous DNA with PHRED scores from 0 to 40 inclusive (in that order)
gTcatAGcgTcatAGcgTcatAGcgTcatAGcgTcatAGcg
+
!"#$%&'()*+,-./0123456789:;<=>?@ABCDEFGHI
@FAKE0009 Original version has lower case unambiguous DNA with PHRED scores from 0 to 40 inclusive (in that order)
tcagtcagtcagtcagtcagtcagtcagtcagtcagtcagt
+
!"#$%&'()*+,-./0123456789:;<=>?@ABCDEFGHI
@FAKE0010 Original version has mixed case ambiguous DNA and PHRED scores of 40, 30, 20, 10 (cycled)
gatcrywsmkhbvdnGATCRYWSMKHBVDN
+
I?5+I?5+I?5+I?5+I?5+I?5+I?5+I?5+I?

Examples, error format

```
@SLXA-B3_649_FC8437_R1_1_1_610_79
GATGTGCAATACCTTTGTAGAGGAA
+SLXA-B3_649_FC8437_R1_1_1_610_79
YYYYYYYYYYYYYYYYYYWYWYYSU
@SLXA-B3_649_FC8437_R1_1_1_397_389
GGTTTGAGAAAGAGAAATGAGATAA
+SLXA-B3_649_FC8437_R1_1_1_397_389
YYYYYYYYYWYYYYWWYYYWYWYWW
@SLXA-B3_649_FC8437_R1_1_1_850_123
GAGGGTGTTGATCATGATGATGGCG
YYYYYYYYYYYYYWYYWYYSYYYSY
@SLXA-B3_649_FC8437_R1_1_1_362_549
GGAAACAAAGTTTTTCTCAACATAG
+SLXA-B3_649_FC8437_R1_1_1_362_549
YYYYYYYYYYYYYYYYYYWWWYWY
@SLXA-B3_649_FC8437_R1_1_1_183_714
GTATTATTTAATGGCATACACTCAA
+SLXA-B3_649_FC8437_R1_1_1_183_714
YYYYYYYYYYWYYYYWYWWUWWWQQ
```

```
@SLXA-B3_649_FC8437_R1_1_1_610_79
GATGTGCAATACCTTTGTAGAGGAA
+SLXA-B3_649_FC8437_R1_1_1_610_79

@SLXA-B3_649_FC8437_R1_1_1_397_389
GGTTTGAGAAAGAGAAATGAGATAA
+SLXA-B3_649_FC8437_R1_1_1_397_389

@SLXA-B3_649_FC8437_R1_1_1_850_123
GAGGGTGTTGATCATGATGATGGC
+SLXA-B3_649_FC8437_R1_1_1_850_123

@SLXA-B3_649_FC8437_R1_1_1_362_549
GGAAACAAAGTTTTTCTCAACATAG
+SLXA-B3_649_FC8437_R1_1_1_362_549

@SLXA-B3_649_FC8437_R1_1_1_183_714
GTATTATTTAATGGCATACACTCAA
+SLXA-B3_649_FC8437_R1_1_1_183_714
```

phred-based quality scores



S - Sanger Phred+33, raw reads typically (0, 40)
X - Solexa Solexa+64, raw reads typically (-5, 40)
I - Illumina 1.3+ Phred+64, raw reads typically (0, 40)
J - Illumina 1.5+ Phred+64, raw reads typically (3, 40)
L - Illumina 1.8+ Phred+33, raw reads typically (0, 41)

Wikipedia

ASCII

Dec	Hex	Name	Char	Ctrl-char	Dec	Hex	Char	Dec	Hex	Char	Dec	Hex	Char
0	0	Null	NUL	CTRL-@	32	20	Space	64	40	@	96	60	`
1	1	Start of heading	SOH	CTRL-A	33	21	!	65	41	A	97	61	a
2	2	Start of text	STX	CTRL-B	34	22	"	66	42	B	98	62	b
3	3	End of text	ETX	CTRL-C	35	23	#	67	43	C	99	63	c
4	4	End of xmit	EOT	CTRL-D	36	24	$	68	44	D	100	64	d
5	5	Enquiry	ENQ	CTRL-E	37	25	%	69	45	E	101	65	e
6	6	Acknowledge	ACK	CTRL-F	38	26	&	70	46	F	102	66	f
7	7	Bell	BEL	CTRL-G	39	27	'	71	47	G	103	67	g
8	8	Backspace	BS	CTRL-H	40	28	(	72	48	H	104	68	h
9	9	Horizontal tab	HT	CTRL-I	41	29	)	73	49	I	105	69	i
10	0A	Line feed	LF	CTRL-J	42	2A	*	74	4A	J	106	6A	j
11	0B	Vertical tab	VT	CTRL-K	43	2B	+	75	4B	K	107	6B	k
12	0C	Form feed	FF	CTRL-L	44	2C	,	76	4C	L	108	6C	l
13	0D	Carriage feed	CR	CTRL-M	45	2D	-	77	4D	M	109	6D	m
14	0E	Shift out	SO	CTRL-N	46	2E	.	78	4E	N	110	6E	n
15	0F	Shift in	SI	CTRL-O	47	2F	/	79	4F	O	111	6F	o
16	10	Data line escape	DLE	CTRL-P	48	30	0	80	50	P	112	70	p
17	11	Device control 1	DC1	CTRL-Q	49	31	1	81	51	Q	113	71	q
18	12	Device control 2	DC2	CTRL-R	50	32	2	82	52	R	114	72	r
19	13	Device control 3	DC3	CTRL-S	51	33	3	83	53	S	115	73	s
20	14	Device control 4	DC4	CTRL-T	52	34	4	84	54	T	116	74	t
21	15	Neg acknowledge	NAK	CTRL-U	53	35	5	85	55	U	117	75	u
22	16	Synchronous idle	SYN	CTRL-V	54	36	6	86	56	V	118	76	v
23	17	End of xmit block	ETB	CTRL-W	55	37	7	87	57	W	119	77	w
24	18	Cancel	CAN	CTRL-X	56	38	8	88	58	X	120	78	x
25	19	End of medium	EM	CTRL-Y	57	39	9	89	59	Y	121	79	y
26	1A	Substitute	SUB	CTRL-Z	58	3A	:	90	5A	Z	122	7A	z
27	1B	Escape	ESC	CTRL-[	59	3B	;	91	5B	[	123	7B	{
28	1C	File separator	FS	CTRL-\	60	3C	<	92	5C	\	124	7C	|
29	1D	Group separator	GS	CTRL-]	61	3D	=	93	5D	]	125	7D	}
30	1E	Record separator	RS	CTRL-^	62	3E	>	94	5E	^	126	7E	~
31	1F	Unit separator	US	CTRL-_	63	3F	?	95	5F	_	127	7F	DEL

Quality scores

	ASCII range	offset	Quality Score (Q)	Quality type
Sanger	33-126	33	0 to 93 (ASCII-33)	PHRED
Solexa\early Illumina	59-126	64	-5 to 62 (ASCII-64)	Solexa
Illumina 1.3+	64-126	64	0 to 62 (ASCII-64)	PHRED

Quality Score

- A quality value *Q* is an integer mapping of *p* (i.e., the error probability that the corresponding base call is incorrect. p is the small, the better).
- Two different equations have been in use:
  - Phred quality score

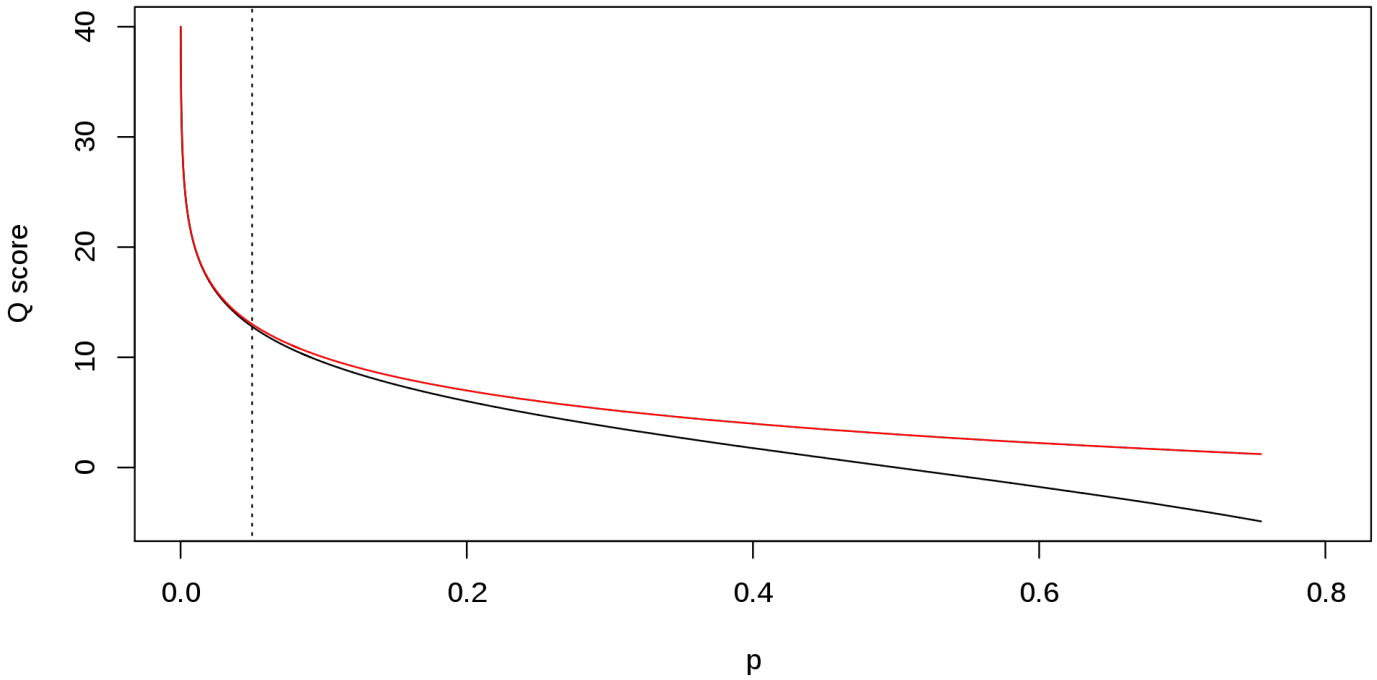
$$p = 10^{\frac{-Q}{10}}$$

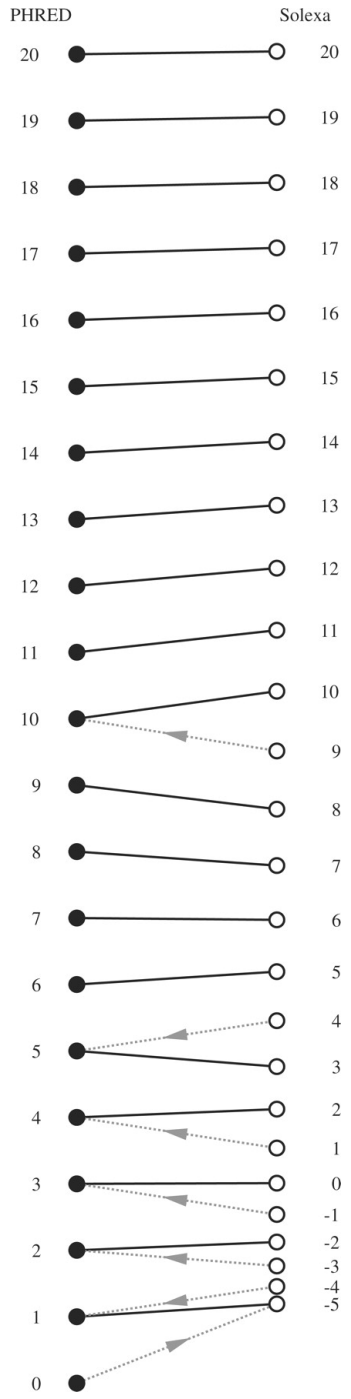
  - Solexa quality score

$$\frac{p}{1-p} = 10^{\frac{-Q}{10}}$$

Ewing B, Green P. (1998) Genome Res. 8(3):186-194.

Quality Score





Relationship between Quality scores (Q) and *p* using different equations: PHRED (red) and Solexa (black).
The vertical dotted line indicates p = 0.05.

Some issues for fastq

- The lack of ownership of this emerging standard by the Sanger Institute contributed greatly to later confusion, such as the phred scores. Users need to figure out which *version* of the Solexa/Illumina pipeline was used.
- Lacks any formal definition to date, and exists in some incompatible variants.
- The '@' and '+' characters have dual usage as line markers or anywhere within the quality string.

Cock et al. Nucleic Acids Res. 2010; 38(6):1767-71.

Some issues for fastq

- The '@' and '+' characters have dual usage as line markers or anywhere within the quality string.

```
@FAKE0005 Original version has PHRED scores from 0 to 62 inclusive (in that
order)
ACGTACGTACGTACGTACGTACGTACGTACGTACGTACGTACGTACGTACGTACGTACG
+
@ABCDEFGHIJKLMNOPQRSTUVWXYZ[\]^_`abcdefghijklmnopqrstuvwxyz{|}~
@FAKE0006 Original version has PHRED scores from 62 to 0 inclusive (in that
order)
GCATGCATGCATGCATGCATGCATGCATGCATGCATGCATGCATGCATGCATGCATGCA
+
~}|{zyxwvutsrqponmlkjihgfedcba`_^]\[ZYXWVUTSRQPONMLKJIHGFEDCBA@
```

Cock et al. Nucleic Acids Res. 2010; 38(6):1767-71.

Manipulate FASTQ files

- fastx_toolkit (http://hannonlab.cshl.edu/fastx_toolkit/)
  - The FASTX-Toolkit is a collection of command line tools for Short-Reads FASTA/FASTQ files preprocessing.
  - "fastq_quality_converter" program can convert Illumina to Sanger
- MAQ (http://maq.sourceforge.net)
  - stands for *Mapping and Assembly with Quality* It builds assembly by mapping short reads to reference sequences.
  - can convert from Solexa to Sanger

SFF file format for 454

- **Standard flowgram format** (SFF) is a binary file format used to encode results from the 454 plat form.
- Need special tool to display.
- Can be converted to FASTQ format
  - sff2fastq (https://github.com/indraniel/sff2fastq)
  - sff_extract (http://bioinf.comav.upv.es/sff_extract/)

Sequence quality control

- Is this good sequence? (essential!)
- Different platforms can introduce varied level of sequence reads error.
- There can be significant lab-to-lab, batch-to-batch and even within chip/slide variations.
- Errors significantly effect the quality of downstream analysis.

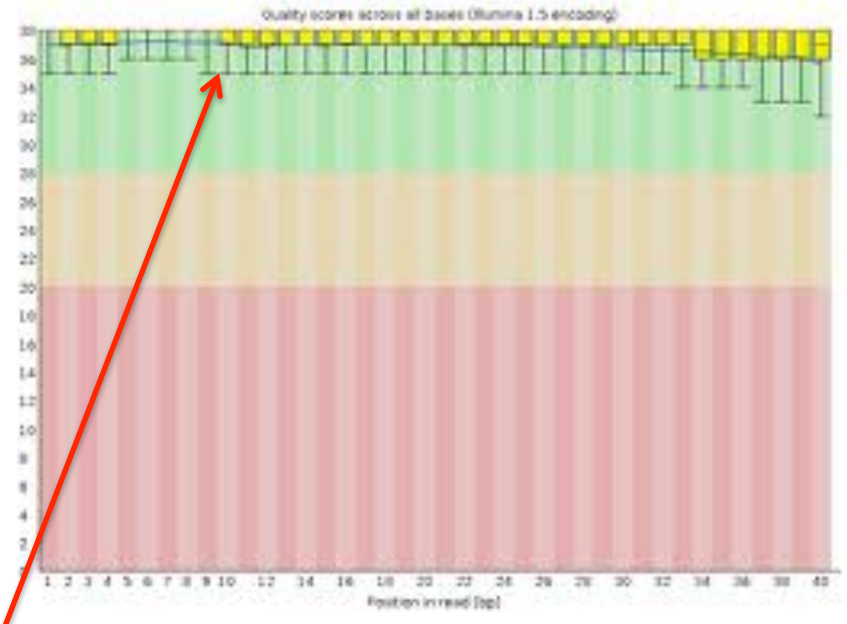
Sequence quality control

per base per sequence quality
(for one read only)

Good

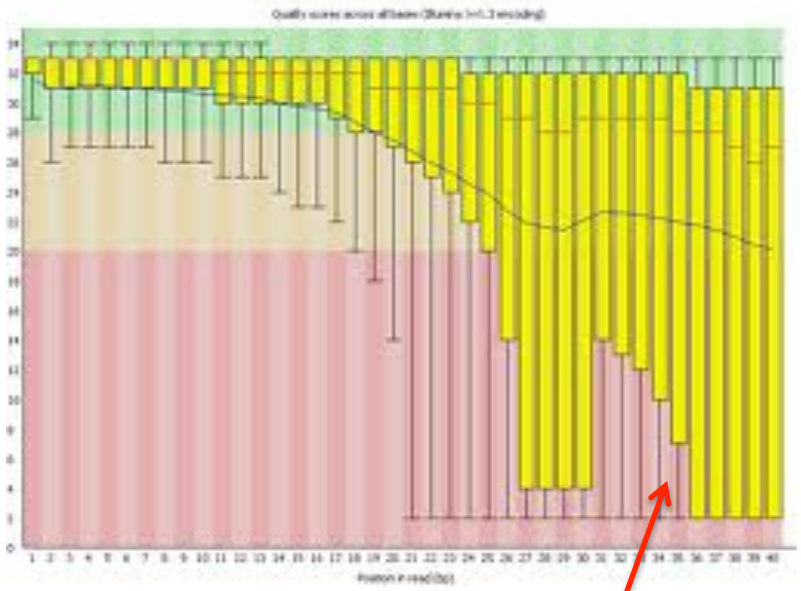
Bad

Q



Position in a read (bp)

Very high score, indicating low error possibility.



Position in a read (bp)

Very low score, indicating high error possibility.

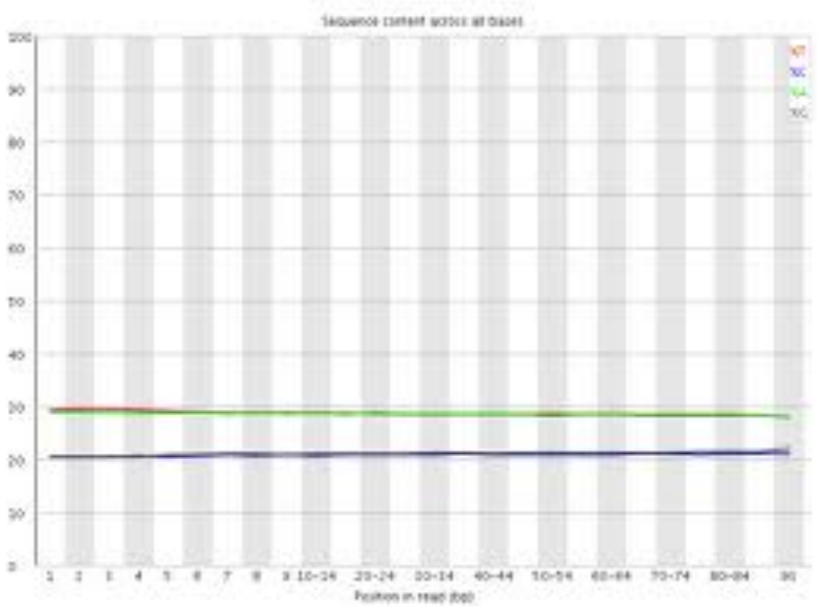
Sequence quality control

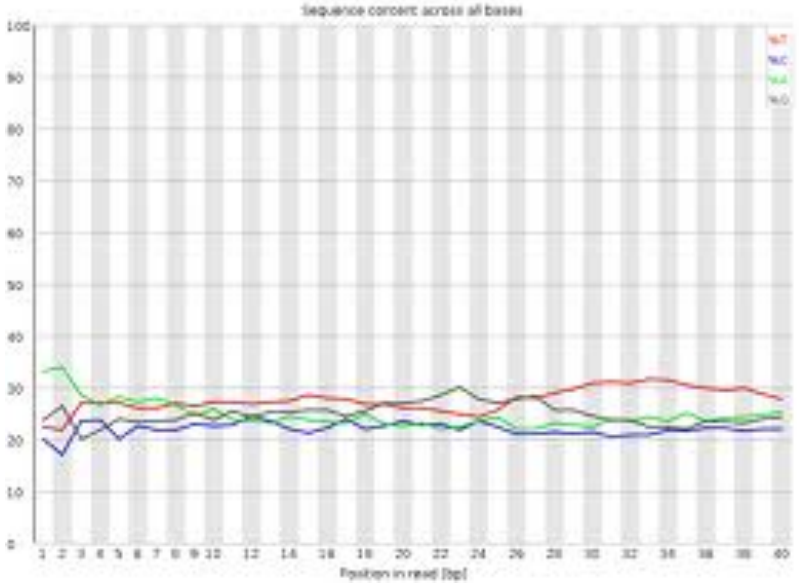
per base average sequence content

Good

Bad

Average Q





Position in read (bp)

Position in read (bp)

```
R1  8er09j3*j(f09diD4^2
R2  {fj0kf9k4;w44{d3@83
R3  C9!&;8r:4#0djr)3.{|
```

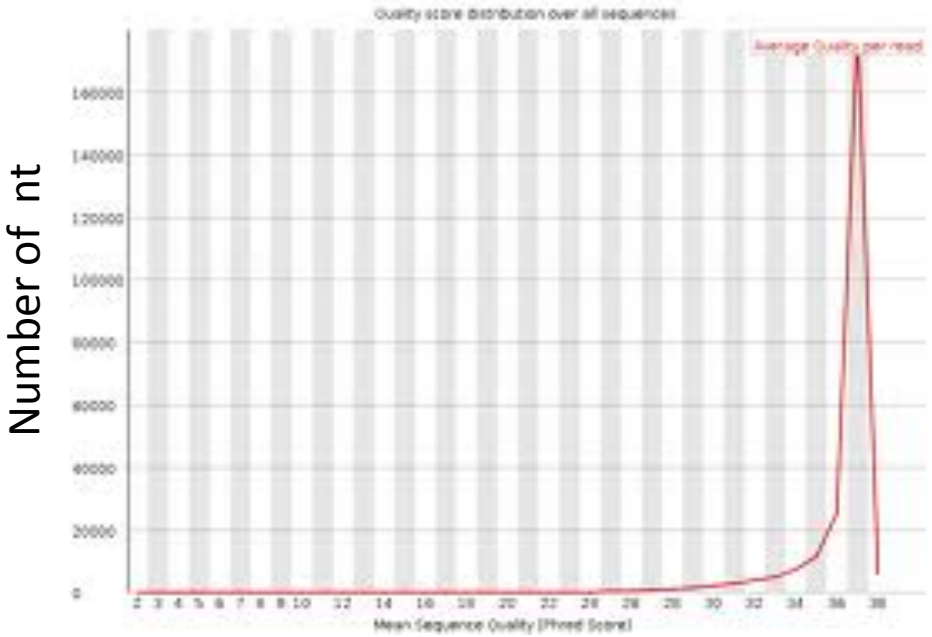
- Low average Q scores.
- large fluctuations.
- Overall, reads have low quality.

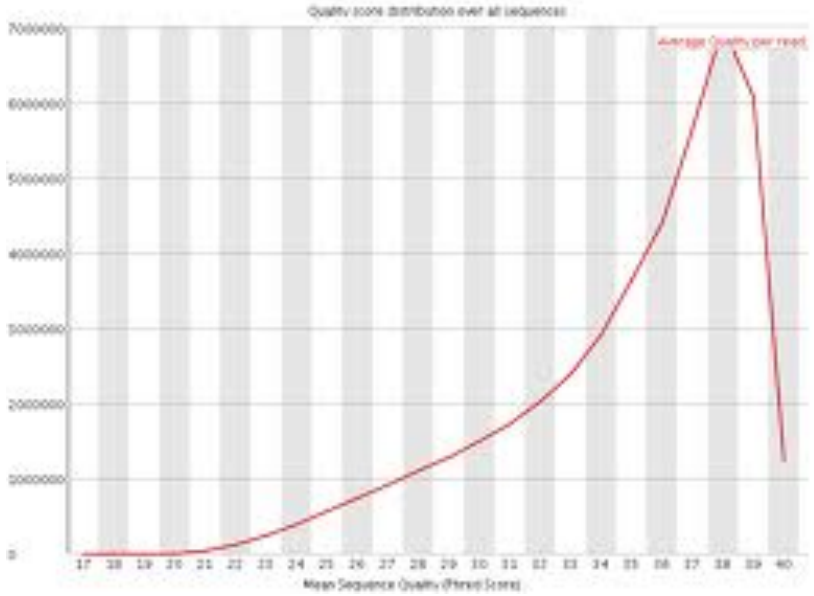
Sequence quality control

Distribution of phred scores in all positions of all reads

Good

Bad





Q (Phred Score)

Most bas calls have high quality scores

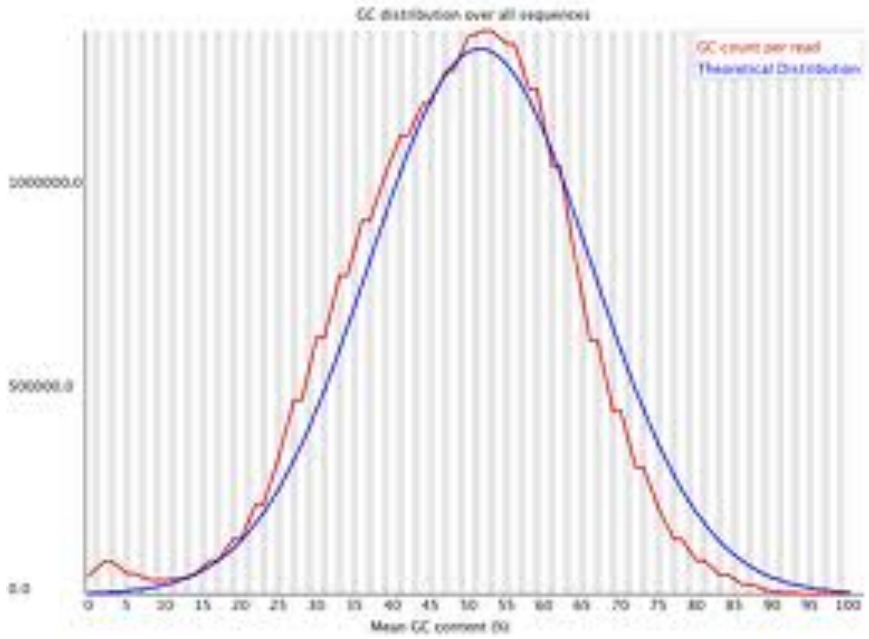
Q (Phred Score)

- Quality scores have a wide range
- Many base calls have low scores
- Overall, reads have low quality

Sequence quality control

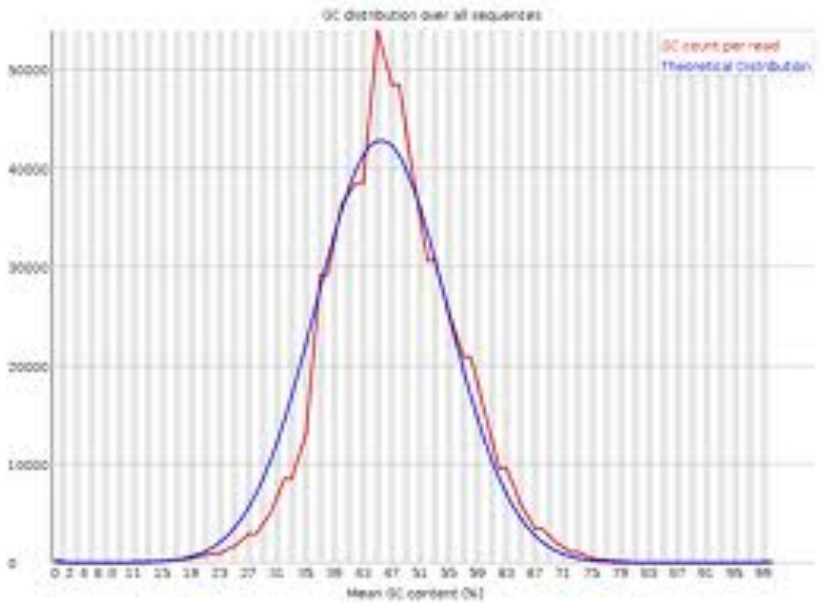
Per sequence GC content

Good



Mean GC content per read (%)

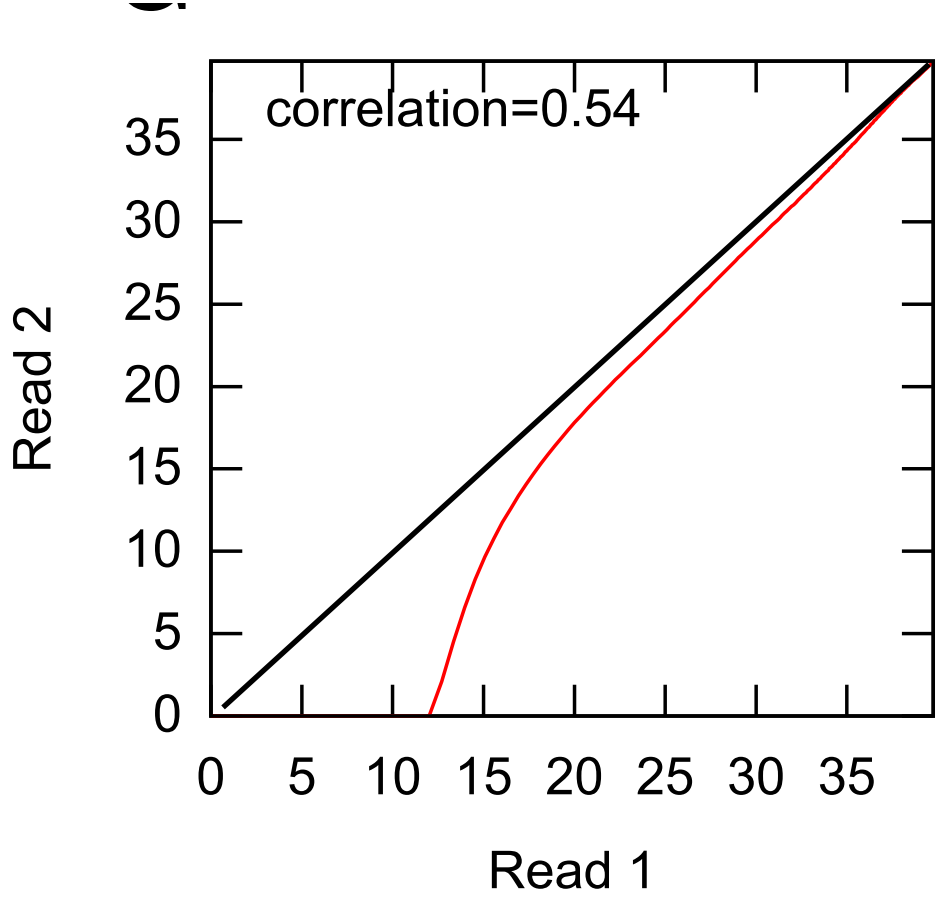
Bad



Mean GC content per read (%)

Distribution of GC contents over all reads

Sequence quality control



Low correlation indicates low quality.

Quality control

- remove reads from problematic tiles that may not be reliable due to sequencing chip quality
- remove reads with low quality, such as mean Q score < 20 for illumina RNA-seq.
- remove low quality bases at two ends of the reads until the quality score reaches a given threshold, such as mean Q score = 20.
- remove short reads.
  - FastQC tool (http://www.bioinformatics.bbsrc.ac.uk/projects/fastqc/)
  - NGSQC: Cross-Platform Quality Analysis Pipeline for Deep Sequencing Data.
    - http://brainarray.mbni.med.umich.edu/brainarray/ngsqc/
  - HTQC: a fast quality control toolkit for Illumina sequencing data
    - https://sourceforge.net/projects/htqc

Data Management

- Raw data are large
  - 5Gb to 20Gb per lane.
  - Do not try to open it with Windows notepad.
  - Convert fastq file to Processed data (e.g., BAM files) are manageable.
- Whole-genome sequencing:
  - A 30X coverage genome pair (tumor/normal): ~500GB
  - 50 genome pairs: ~25TB
- We need high-performance, replicated storage
  - ~$700/TB; but non-redundant storage: $200/TB
  - to be kept for > 36 months?

Transfer data

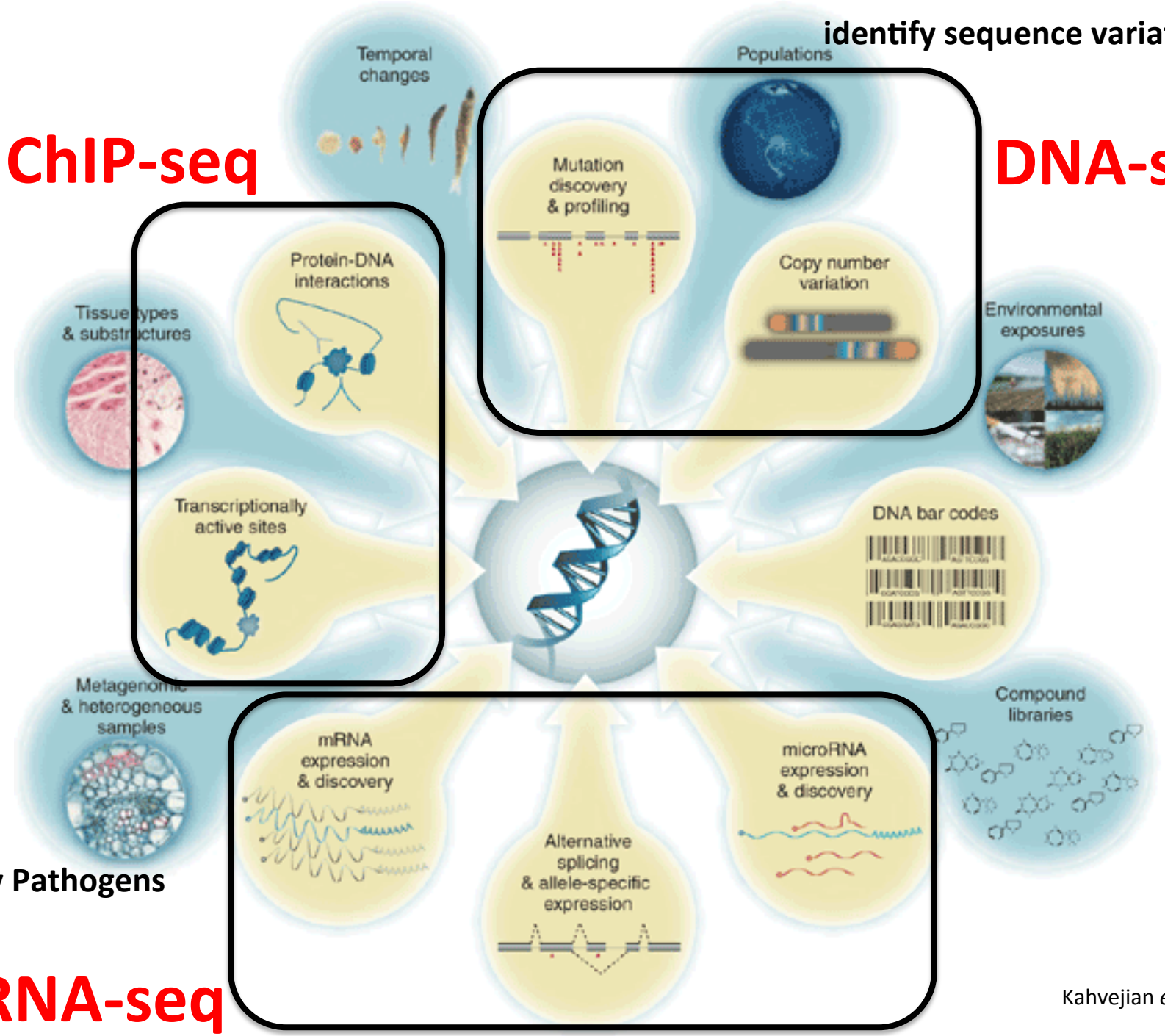
- FASTQ files will be compressed with gzip prior to delivery.
- It is difficult to download data via http or ftp (15Mb/s, about 1 terabyte per day).
- A commercial software/protocol is become popular
  - Aspera “next-generation file transport”
  - transfer protocol that leverages existing WAN infrastructure and commodity hardware to achieve speeds that are up to hundreds of times faster than FTP and HTTP.
  - This can give 400-800Mb/s. It takes 30 seconds to move a 24-GB data file.
  - 1000 Genomes Project, BGI etc. applied this method for their data transport.

Nature 498, 255–260 (2013)

NGS

- Introduction to the background
- NGS workflow and accuracy
- Data format, quality control, data management
- Assembly
- RNA-seq
  - Aligner
  - Analysis tools
  - Applications, such as MiRNA
- Chip-seq
  - Applications

identify sequence variations



Identify Pathogens

Kahvejian *et al*, 2008

Goals of Assembly

- Reconstruction of unknown genomes
  - viruses, bacteria
  - Eucaryotes (individual genomes)
  - metagenomics (environmental samples, microbial communities)
- RNA-Seq => transcriptomes
  - for organisms without reference genomes
  - novel transcripts
  - fusion genes
  - viral integration
- Local assembly for detection of insertions and genomic rearrangements
  - unmapped reads from WGS
  - transposable elements
  - viral integration

Next-generation sequencing

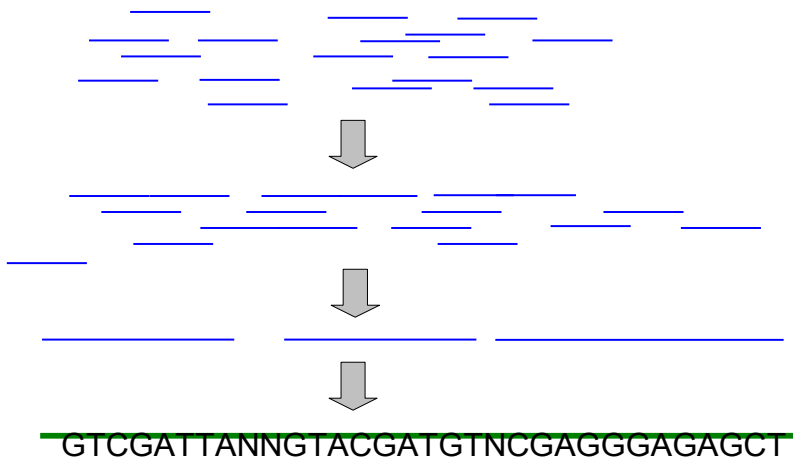
- Much higher throughput (1-4gbps / day)
- Lower cost / base pair
- Inherent ability to do paired-end (mate-pair) sequencing

Assembly

- **sequence assembly** refers to aligning and merging fragments of a much longer DNA sequence in order to reconstruct the original sequence.
- Taking many copies of a book, passing each of them through a shredder with a different cutter, and piecing the text of the book back together just by looking at the shredded pieces.
- Combines short sequencing reads into contigs based on sequence similarity and overlap between reads.
- Find the shortest common sequence of a set of reads

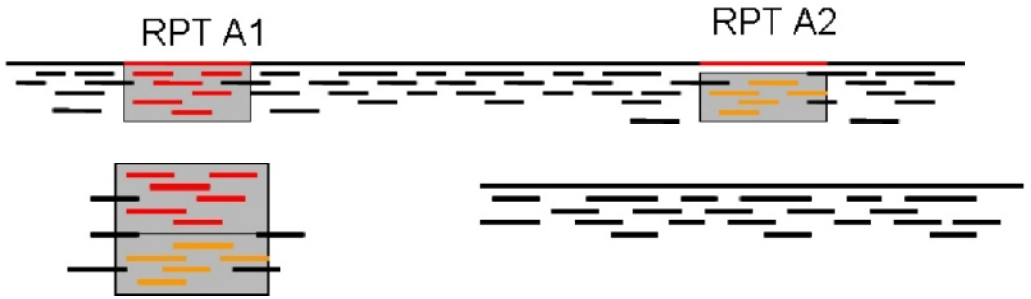
Assembly

- **Spanner**: single read that spans a repeat instance with sufficient unique sequence on either side of the repeat.
- **Contig**: contiguous sequence formed by several overlapping reads with **no gaps**.
- **Supercontig (scaffold)**: ordered and oriented set of contigs, usually by mate pairs. Relative distance known => fill gaps between contigs with "NNNNNNNNN…"
- **Consensus sequence**: sequence derived from the multiple alignment of reads in a contig.



Challenges for Assembly

- Repeats or similar parts in the genome. The reads originating from different copies of a repeat appear identical to the assembler and cause assembly errors.



- Non-random shearing
- Lose DNA fragments during library preparation
- Bias during amplification
- Very short fragment lengths (25-200bps)
- High error rate

Assembly

- Assembly algorithms
- De novo whole genome assembling strategies
- Mapping assembling strategies