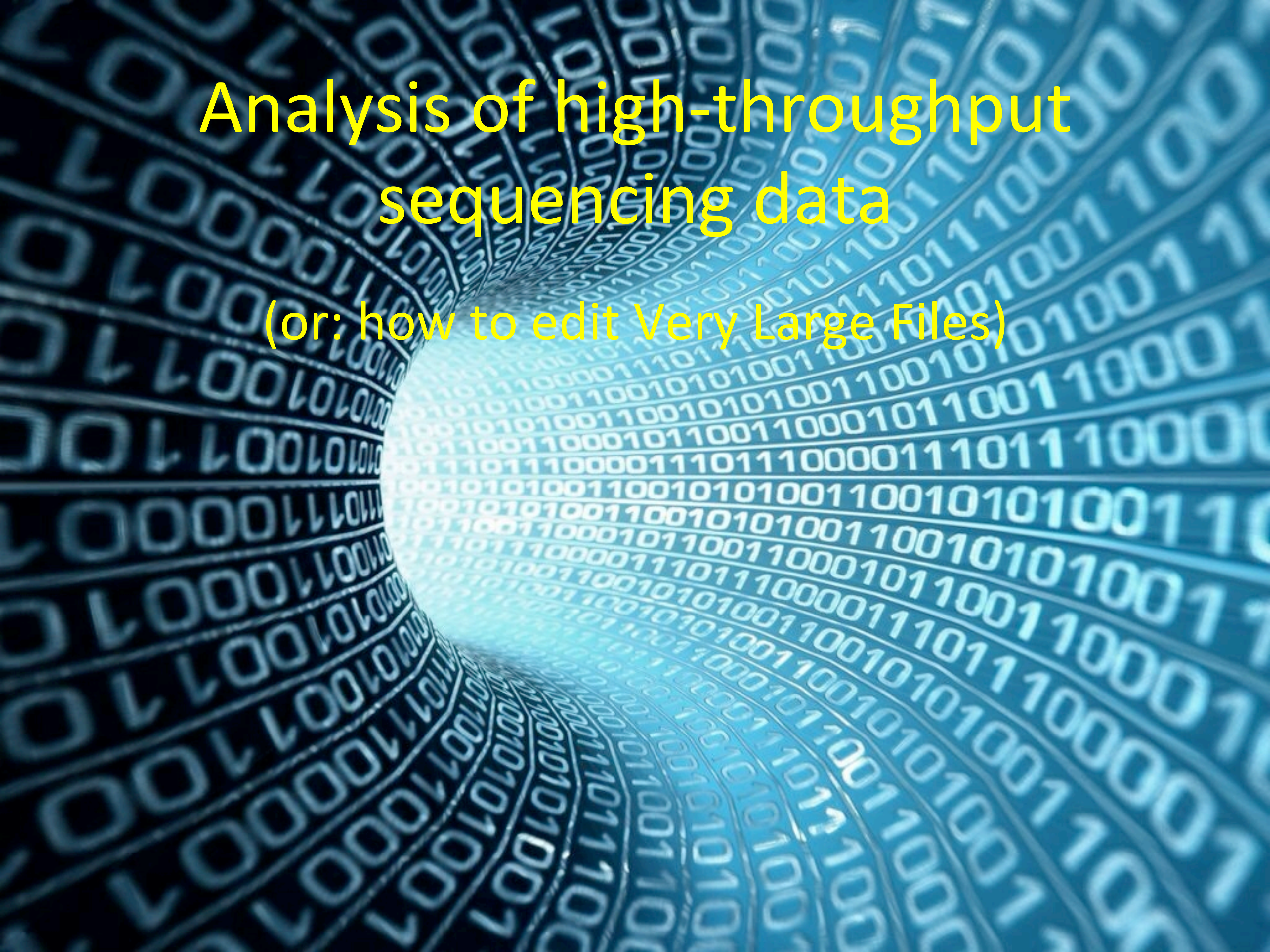
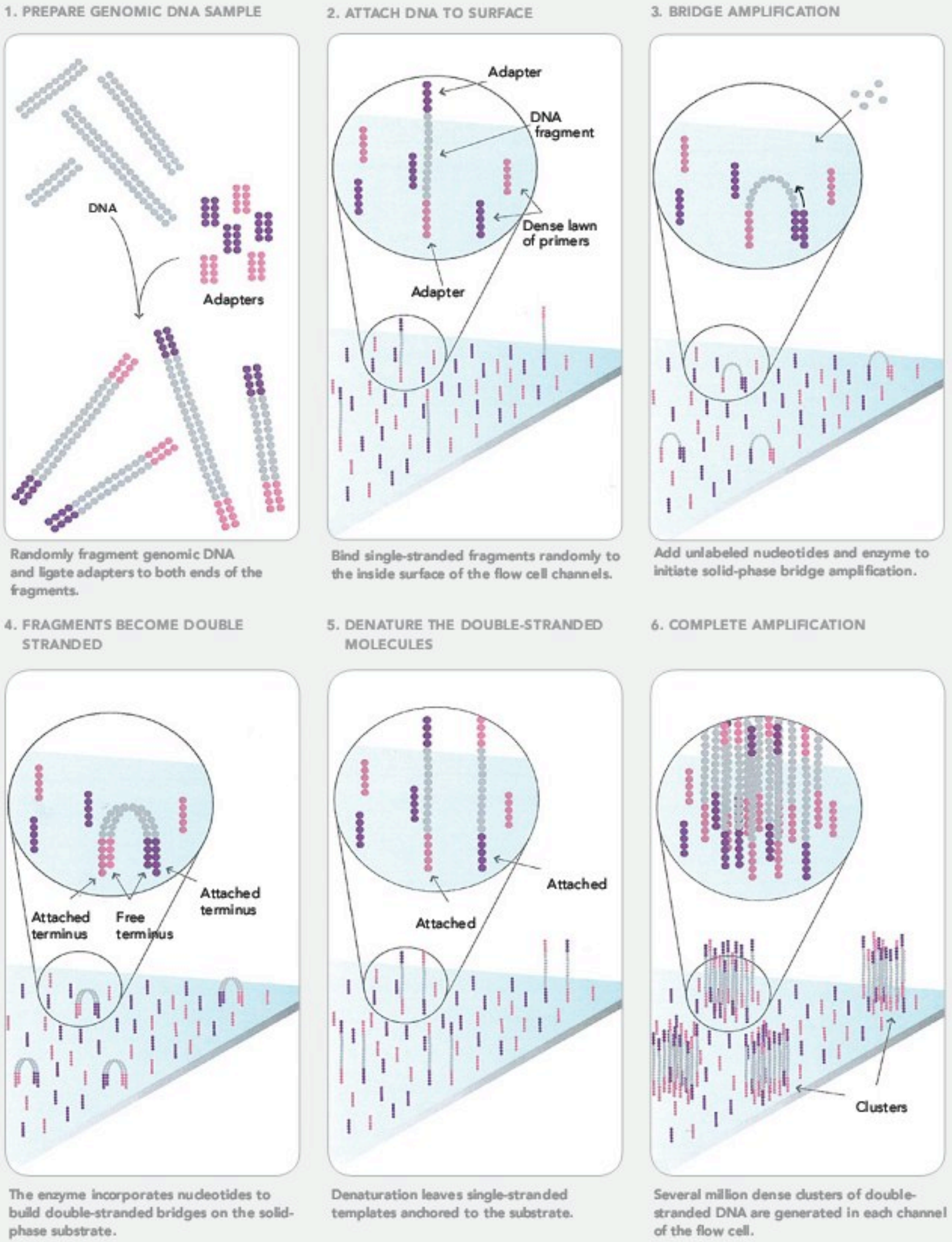
Analysis of high-throughput sequencing data

(or: how to edit Very Large Files)



Cluster generation



1. PREPARE GENOMIC DNA SAMPLE

Randomly fragment genomic DNA and ligate adapters to both ends of the fragments.

2. ATTACH DNA TO SURFACE

Bind single-stranded fragments randomly to the inside surface of the flow cell channels.

3. BRIDGE AMPLIFICATION

Add unlabeled nucleotides and enzyme to initiate solid-phase bridge amplification.

4. FRAGMENTS BECOME DOUBLE STRANDED

The enzyme incorporates nucleotides to build double-stranded bridges on the solid-phase substrate.

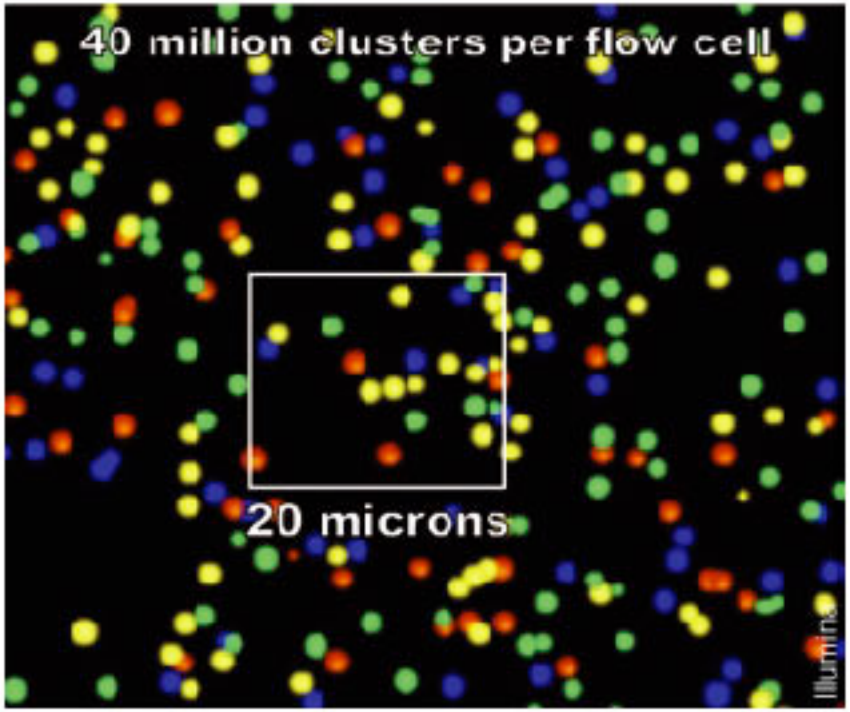
5. DENATURE THE DOUBLE-STRANDED MOLECULES

Denaturation leaves single-stranded templates anchored to the substrate.

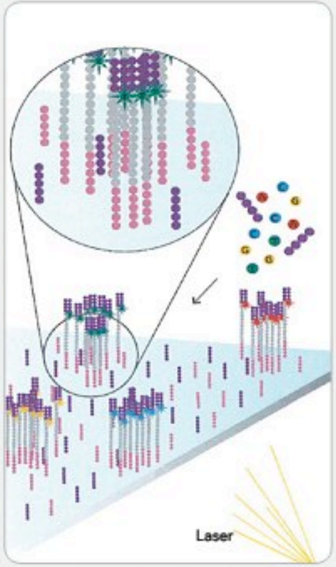
6. COMPLETE AMPLIFICATION

Several million dense clusters of double-stranded DNA are generated in each channel of the flow cell.

Sequencing

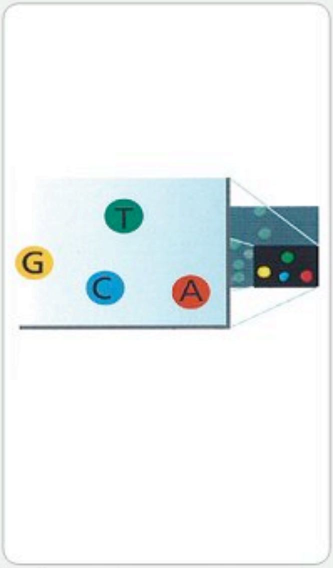


7. DETERMINE FIRST BASE



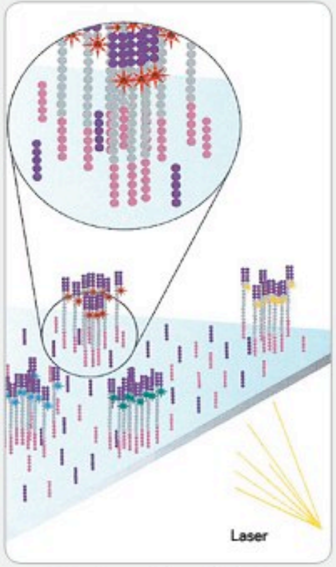
First chemistry cycle: to initiate the first sequencing cycle, add all four labeled reversible terminators, primers and DNA polymerase enzyme to the flow cell.

8. IMAGE FIRST BASE



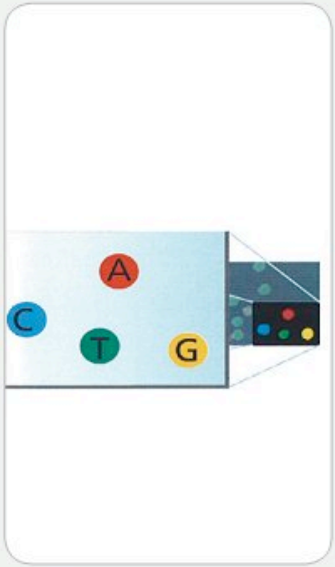
After laser excitation, capture the image of emitted fluorescence from each cluster on the flow cell. Record the identity of the first base for each cluster.

9. DETERMINE SECOND BASE



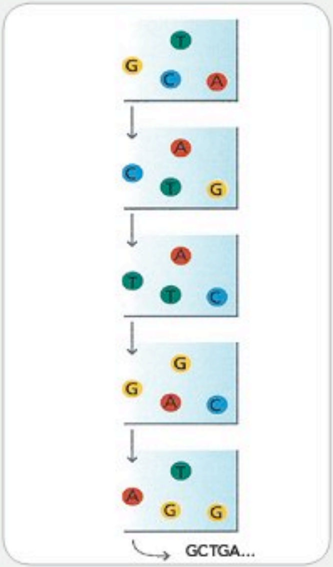
Second chemistry cycle: to initiate the next sequencing cycle, add all four labeled reversible terminators and enzyme to the flow cell.

10. IMAGE SECOND CHEMISTRY CYCLE



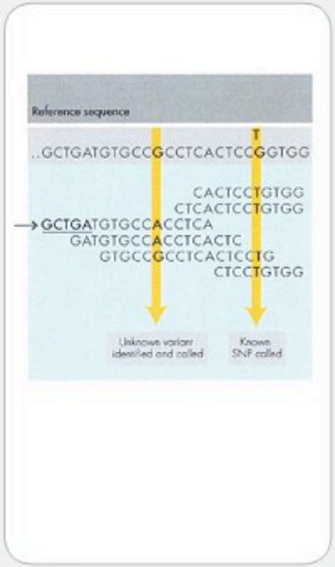
After laser excitation, collect the image data as before. Record the identity of the second base for each cluster.

11. SEQUENCE READS OVER MULTIPLE CHEMISTRY CYCLES



Repeat cycles of sequencing to determine the sequence of bases in a given fragment a single base at time.

12. ALIGN DATA



Align data, compare to a reference, and identify sequence differences.

Now what?

- Download FASTQ files from the sequencing center.
- VERY LARGE – do not try to open in a text editor.

```
@HWI-ST141_0365:2:1101:2983:2114#TTAGGC/1
TGAGAATGTGAAGGCCAAGATCCAAGACAAGGAAGGGATTCCCCCAGACC
+HWI-ST141_0365:2:1101:2983:2114#TTAGGC/1
Y^_ccZZa\\\b``Rbfb`Z`^`eaRbReJYbHYYbH^c]YXa^__^acc
@HWI-ST141_0365:2:1101:2965:2155#TTAGGC/1
GGCATCTGACAGTTGTATTTGAGATGGTATGCCACACACCTAGTTAAGTA
+HWI-ST141_0365:2:1101:2965:2155#TTAGGC/1
_bbceeeeeeegeghhdefihhacgggf`eggggghhgiiiiiihaedhhfd
@HWI-ST141_0365:2:1101:2886:2184#TTAGGC/1
CTCCAACATAGCTGAACGTTGATACAGATCTACAAAAATAATGAAATGAT
+HWI-ST141_0365:2:1101:2886:2184#TTAGGC/1
bbbeeeeeegggggiiiiiiiihiiiiihihihiiiiiiiiiiihiiiih
@HWI-ST141_0365:2:1101:2778:2215#TTAGGC/1
CGAAAACCACTTCTCAGGGAGGGAGAGTGCATACATGTTGTTC
+HWI-ST141_0365:2:1101:2778:2215#TTAGGC/1
bbbeeeeeegggggiihiiifgiiehfh`gghiiiiiiiiiii
@HWI-ST141_0365:2:1101:2956:2216#TTAGGC/1
CATGGATGCTCTCAAAGTGTTGTCTGATATGGGCTACTACATCGAGGACA
+HWI-ST141_0365:2:1101:2956:2216#TTAGGC/1
_bbeeeeeegggggiiifbgeghfgihagffgghfgifhiifcbgaffcfg
@HWI-ST141_0365:2:1101:2985:2249#TTAGGC/1
GAGAAACCTCCGACACTGGCTG
+HWI-ST141_0365:2:1101:2985:2249#TTAGGC/1
Bbbeeeeeeggggfhhhhhhhh
……………………………………………………………………
…………………………………………………………………….
```

FASTQ files

1 read = 4 lines

```
@HWI-ST141_0365:2:1101:2983:2114#TTAGGC/1
TGAGAATGTGAAGGCCAAGATCCAAGACAAGGAAGGGATTCCCCCAGACC
+HWI-ST141_0365:2:1101:2983:2114#TTAGGC/1
Y^_ccZZa\\\b``Rbfb`Z`^`eaRbReJYbHYYbH^c]YXa^__^acc
@HWI-ST141_0365:2:1101:2965:2155#TTAGGC/1
GGCATCTGACAGTTGTATTTGAGATGGTATGCCACACACCTAGTTAAGTA
+HWI-ST141_0365:2:1101:2965:2155#TTAGGC/1
_bbceeeeeeegeghhdefihhacgggf`eggggghhgiiiiiihaedhhfd
@HWI-ST141_0365:2:1101:2886:2184#TTAGGC/1
CTCCAACATAGCTGAACGTTGATACAGATCTACAAAAATAATGAAATGAT
+HWI-ST141_0365:2:1101:2886:2184#TTAGGC/1
bbbeeeeegggggiiiiiiiihiiiiihihihiiiiiiiiiiiihiiiih
@HWI-ST141_0365:2:1101:2778:2215#TTAGGC/1
CGAAAACCACTTCTCAGGGAGGGAGAGTGCATACATGTTGTTC
+HWI-ST141_0365:2:1101:2778:2215#TTAGGC/1
bbbeeeeegggggiihiiifgiiehfh`gghiiiiiiiiiiii
@HWI-ST141_0365:2:1101:2956:2216#TTAGGC/1
CATGGATGCTCTCAAAGTGTTGTCTGATATGGGCTACTACATCGAGGACA
+HWI-ST141_0365:2:1101:2956:2216#TTAGGC/1
_bbeeeeeeggggggiiifbgeghfgihagffgghfgifhiifcbgaffcfg
@HWI-ST141_0365:2:1101:2985:2249#TTAGGC/1
GAGAAACCTCCGACACTGGCTG
+HWI-ST141_0365:2:1101:2985:2249#TTAGGC/1
Bbbeeeeeggggfhhhhhhhhh
………………………………………………………………
………………………………………………………………
```

FASTQ files

```
@HWI-ST141_0365:2:1101:2983:2114#TTAGGC/1
TGAGAATGTGAAGGCCAAGATCCAAGACAAGGAAGGGATTCCCCCAGACC
+HWI-ST141_0365:2:1101:2983:2114#TTAGGC/1
Y^_ccZZa\\\b``Rbfb`Z`^`eaRbReJYbHYYbH^c]YXa^__^acc
```

- First line is an IDENTIFIER.
- Starts with @, then instrument name, flowcell lane, tile number and flowcell x,y coordinates.
- Ends with the barcode sequence and pair # (for paired-end sequencing).

FASTQ files

```
@HWI-ST141_0365:2:1101:2983:2114#TTAGGC/1
TGAGAATGTGAAGGCCAAGATCCAAGACAAGGAAGGGATTCCCCCAGACC
+HWI-ST141_0365:2:1101:2983:2114#TTAGGC/1
Y^_ccZZa\\\b``Rbfb`Z`^`eaRbReJYbHYYbH^c]YXa^__^acc
```

- Second line is the READ SEQUENCE

FASTQ files

```
@HWI-ST141_0365:2:1101:2983:2114#TTAGGC/1
TGAGAATGTGAAGGCCAAGATCCAAGACAAGGAAGGGATTCCCCCAGACC
+HWI-ST141_0365:2:1101:2983:2114#TTAGGC/1
Y^_ccZZa\\\b``Rbfb`Z`^`eaRbReJYbHYYbH^c]YXa^__^acc
```

- Third line is the same IDENTIFIER, now starting with +
  (for additional information)

FASTQ files

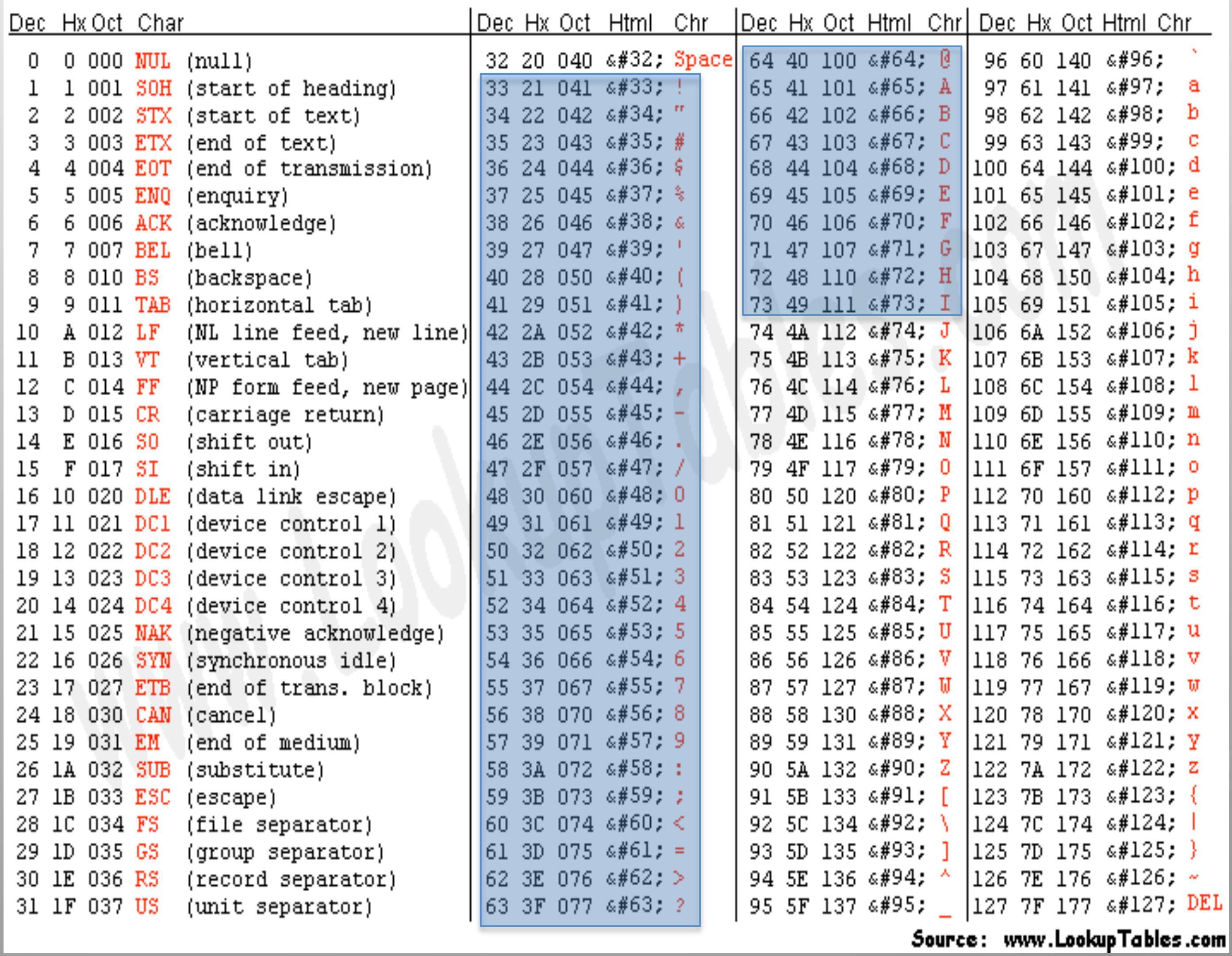
```
@HWI-ST141_0365:2:1101:2983:2114#TTAGGC/1
TGAGAATGTGAAGGCCAAGATCCAAGACAAGGAAGGGATTCCCCCAGACC
+HWI-ST141_0365:2:1101:2983:2114#TTAGGC/1
Y^_ccZZa\\\b``Rbfb`Z`^`eaRbReJYbHYYbH^c]YXa^__^acc
```

- Fourth line is Phred scale QUALITY SCORES.
- Based on the computer ASCII numbers.
- Range from 0 to 40.
- Q = -10 log P, where P is the probability of incorrect base call.

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 %
40	1 in 10000	99.99 %

FASTQ files

- All keyboard characters represent an ASCII number.
- Saves space, e.g. U (1 byte) instead of 85 (2 bytes).
- For Illumina data, ASCII 33 (!) = 0, ASCII 73 (I) = 40

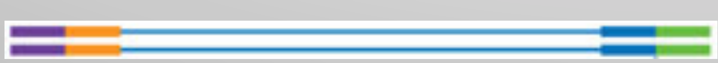


Dec	Hx	Oct	Char	Dec	Hx	Oct	Html	Chr	Dec	Hx	Oct	Html	Chr	Dec	Hx	Oct	Html	Chr
0	0	000	NUL (null)	32	20	040	&#32;	Space	64	40	100	&#64;	@	96	60	140	&#96;	`
1	1	001	SOH (start of heading)	33	21	041	&#33;	!	65	41	101	&#65;	A	97	61	141	&#97;	a
2	2	002	STX (start of text)	34	22	042	&#34;	"	66	42	102	&#66;	B	98	62	142	&#98;	b
3	3	003	ETX (end of text)	35	23	043	&#35;	#	67	43	103	&#67;	C	99	63	143	&#99;	c
4	4	004	EOT (end of transmission)	36	24	044	&#36;	$	68	44	104	&#68;	D	100	64	144	&#100;	d
5	5	005	ENQ (enquiry)	37	25	045	&#37;	%	69	45	105	&#69;	E	101	65	145	&#101;	e
6	6	006	ACK (acknowledge)	38	26	046	&#38;	&	70	46	106	&#70;	F	102	66	146	&#102;	f
7	7	007	BEL (bell)	39	27	047	&#39;	'	71	47	107	&#71;	G	103	67	147	&#103;	g
8	8	010	BS (backspace)	40	28	050	&#40;	(	72	48	110	&#72;	H	104	68	150	&#104;	h
9	9	011	TAB (horizontal tab)	41	29	051	&#41;	)	73	49	111	&#73;	I	105	69	151	&#105;	i
10	A	012	LF (NL line feed, new line)	42	2A	052	&#42;	*	74	4A	112	&#74;	J	106	6A	152	&#106;	j
11	B	013	VT (vertical tab)	43	2B	053	&#43;	+	75	4B	113	&#75;	K	107	6B	153	&#107;	k
12	C	014	FF (NP form feed, new page)	44	2C	054	&#44;	,	76	4C	114	&#76;	L	108	6C	154	&#108;	l
13	D	015	CR (carriage return)	45	2D	055	&#45;	-	77	4D	115	&#77;	M	109	6D	155	&#109;	m
14	E	016	SO (shift out)	46	2E	056	&#46;	.	78	4E	116	&#78;	N	110	6E	156	&#110;	n
15	F	017	SI (shift in)	47	2F	057	&#47;	/	79	4F	117	&#79;	O	111	6F	157	&#111;	o
16	10	020	DLE (data link escape)	48	30	060	&#48;	0	80	50	120	&#80;	P	112	70	160	&#112;	p
17	11	021	DC1 (device control 1)	49	31	061	&#49;	1	81	51	121	&#81;	Q	113	71	161	&#113;	q
18	12	022	DC2 (device control 2)	50	32	062	&#50;	2	82	52	122	&#82;	R	114	72	162	&#114;	r
19	13	023	DC3 (device control 3)	51	33	063	&#51;	3	83	53	123	&#83;	S	115	73	163	&#115;	s
20	14	024	DC4 (device control 4)	52	34	064	&#52;	4	84	54	124	&#84;	T	116	74	164	&#116;	t
21	15	025	NAK (negative acknowledge)	53	35	065	&#53;	5	85	55	125	&#85;	U	117	75	165	&#117;	u
22	16	026	SYN (synchronous idle)	54	36	066	&#54;	6	86	56	126	&#86;	V	118	76	166	&#118;	v
23	17	027	ETB (end of trans. block)	55	37	067	&#55;	7	87	57	127	&#87;	W	119	77	167	&#119;	w
24	18	030	CAN (cancel)	56	38	070	&#56;	8	88	58	130	&#88;	X	120	78	170	&#120;	x
25	19	031	EM (end of medium)	57	39	071	&#57;	9	89	59	131	&#89;	Y	121	79	171	&#121;	y
26	1A	032	SUB (substitute)	58	3A	072	&#58;	:	90	5A	132	&#90;	Z	122	7A	172	&#122;	z
27	1B	033	ESC (escape)	59	3B	073	&#59;	;	91	5B	133	&#91;	[	123	7B	173	&#123;	{
28	1C	034	FS (file separator)	60	3C	074	&#60;	<	92	5C	134	&#92;	\	124	7C	174	&#124;	|
29	1D	035	GS (group separator)	61	3D	075	&#61;	=	93	5D	135	&#93;	]	125	7D	175	&#125;	}
30	1E	036	RS (record separator)	62	3E	076	&#62;	>	94	5E	136	&#94;	^	126	7E	176	&#126;	~
31	1F	037	US (unit separator)	63	3F	077	&#63;	?	95	5F	137	&#95;	_	127	7F	177	&#127;	DEL

Source: www.LookupTables.com

Quality control processing of raw data

- We need to make sure that no low quality reads are in the data
- Also that no adapter sequences are left in the dataset

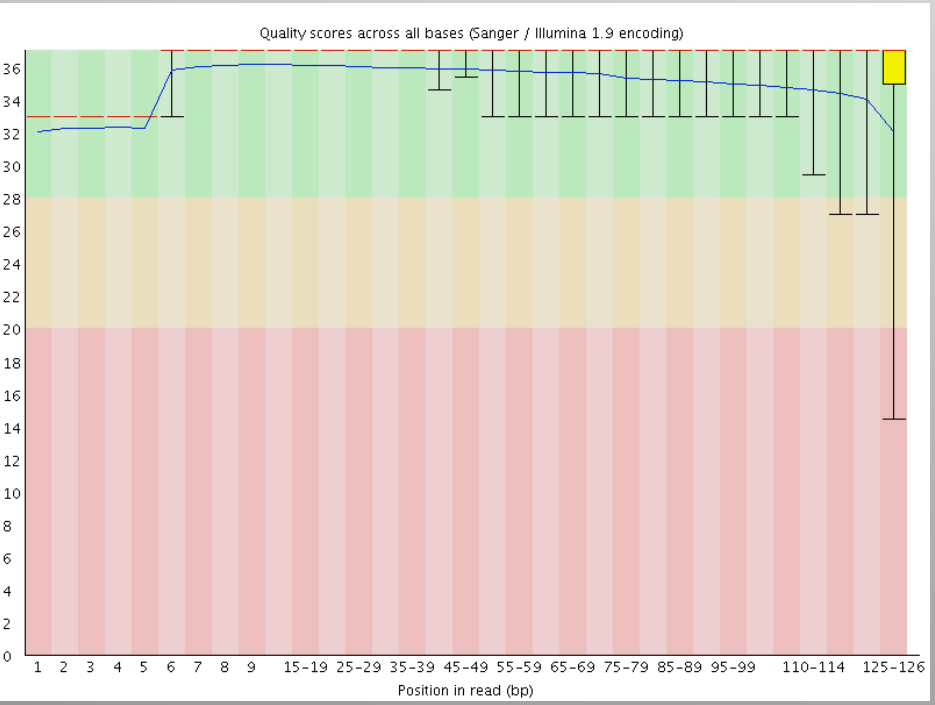


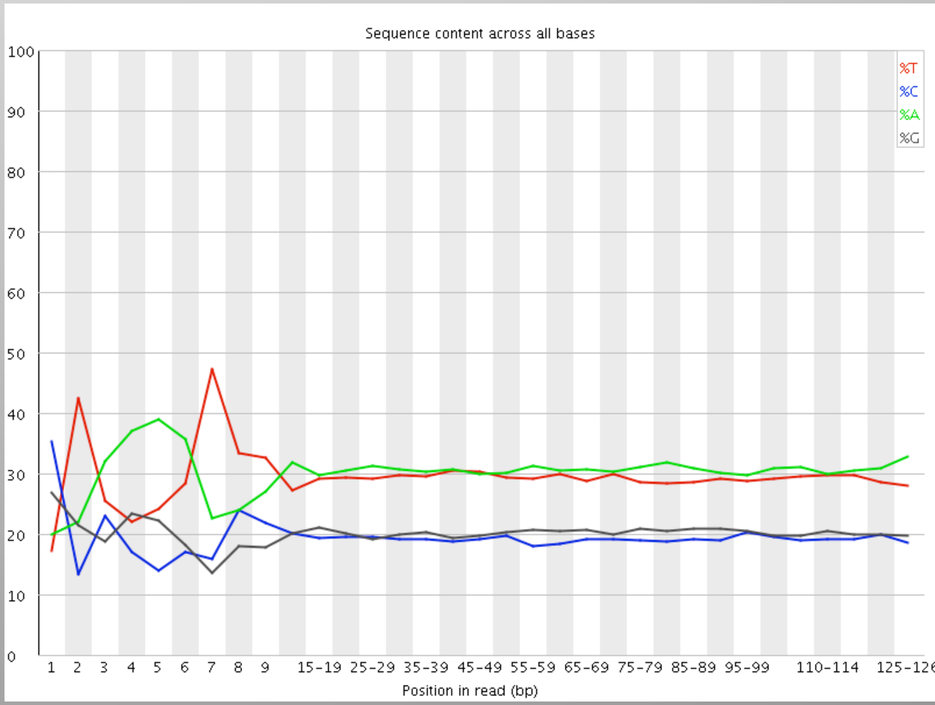
Adapter cDNA Adapter

- We also want to study the quality score distribution (should be high) and nucleotide distribution (should be random), and the fraction of duplicate reads (should be low).

fastQC

Quick way to visualize raw data, automatically generates html report file.





The FASTX toolkit

FASTQ-to-FASTA converter
Convert FASTQ files to FASTA files.

FASTQ Information
Chart Quality Statistics and Nucleotide Distribution

FASTQ/A Collapser
Collapsing identical sequences in a FASTQ/A file into a single sequence (while maintaining reads counts)

FASTQ/A Trimmer
Shortening reads in a FASTQ or FASTQ files (removing barcodes or noise).

FASTQ/A Renamer
Renames the sequence identifiers in FASTQ/A file.

FASTQ/A Clipper
Removing sequencing adapters / linkers

The FASTX toolkit

FASTQ/A Reverse-Complement
Producing the Reverse-complement of each sequence in a FASTQ/FASTA file.

FASTQ/A Barcode splitter
Splitting a FASTQ/FASTA files containing multiple samples

FASTA Formatter
changes the width of sequences line in a FASTA file

FASTA Nucleotide Changer
Converts FASTA sequences from/to RNA/DNA

FASTQ Quality Filter
Filters sequences based on quality

FASTQ Quality Trimmer
Trims (cuts) sequences based on quality

FASTQ Masker
Masks nucleotides with 'N' (or other character) based on quality

The FASTX toolkit

2 run modes:

Through command-line interface on your computer

- No queues or server downtime to worry about
- No large data transfers

Through graphical interface on the GALAXY web portal

- Avoid “complicated” command structure
- No need for plotting software (e.g. gnuplot)
- Potentially large processing power through the cloud.

GALAXY

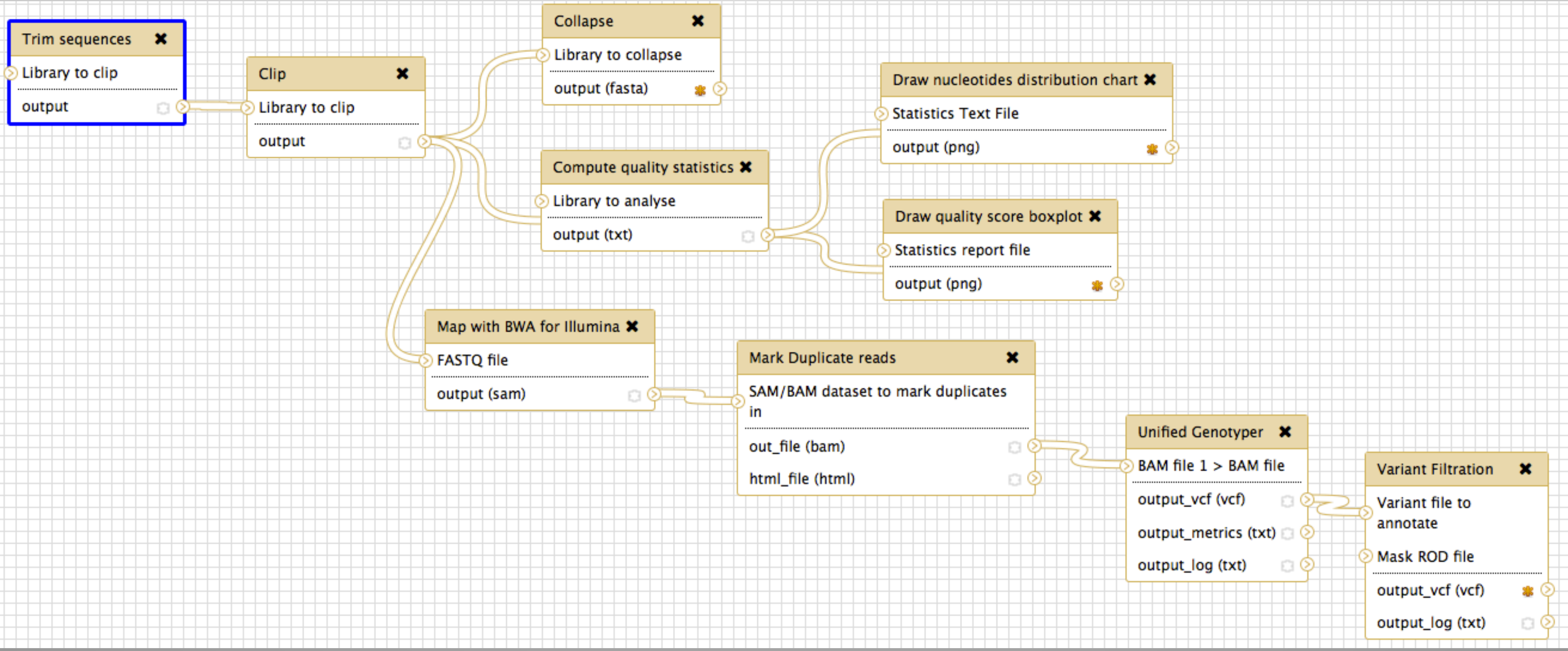
https://usegalaxy.org/

- Easy to make and share customized workflows for data processing
- Embeds other software within the site (e.g. fastx toolkit, SAMtools, Picard, GATK)
- Now available on the Amazon cloud for $

- However, no software for *de novo* assembly / clustering available by default
- And you still need to understand what the parameters *mean*.

GALAXY

https://usegalaxy.org/u/pierredewit/w/test



Trim sequences ✖
Library to clip
output

Clip ✖
Library to clip
output

Collapse ✖
Library to collapse
output (fasta)

Compute quality statistics ✖
Library to analyse
output (txt)

Draw nucleotides distribution chart ✖
Statistics Text File
output (png)

Draw quality score boxplot ✖
Statistics report file
output (png)

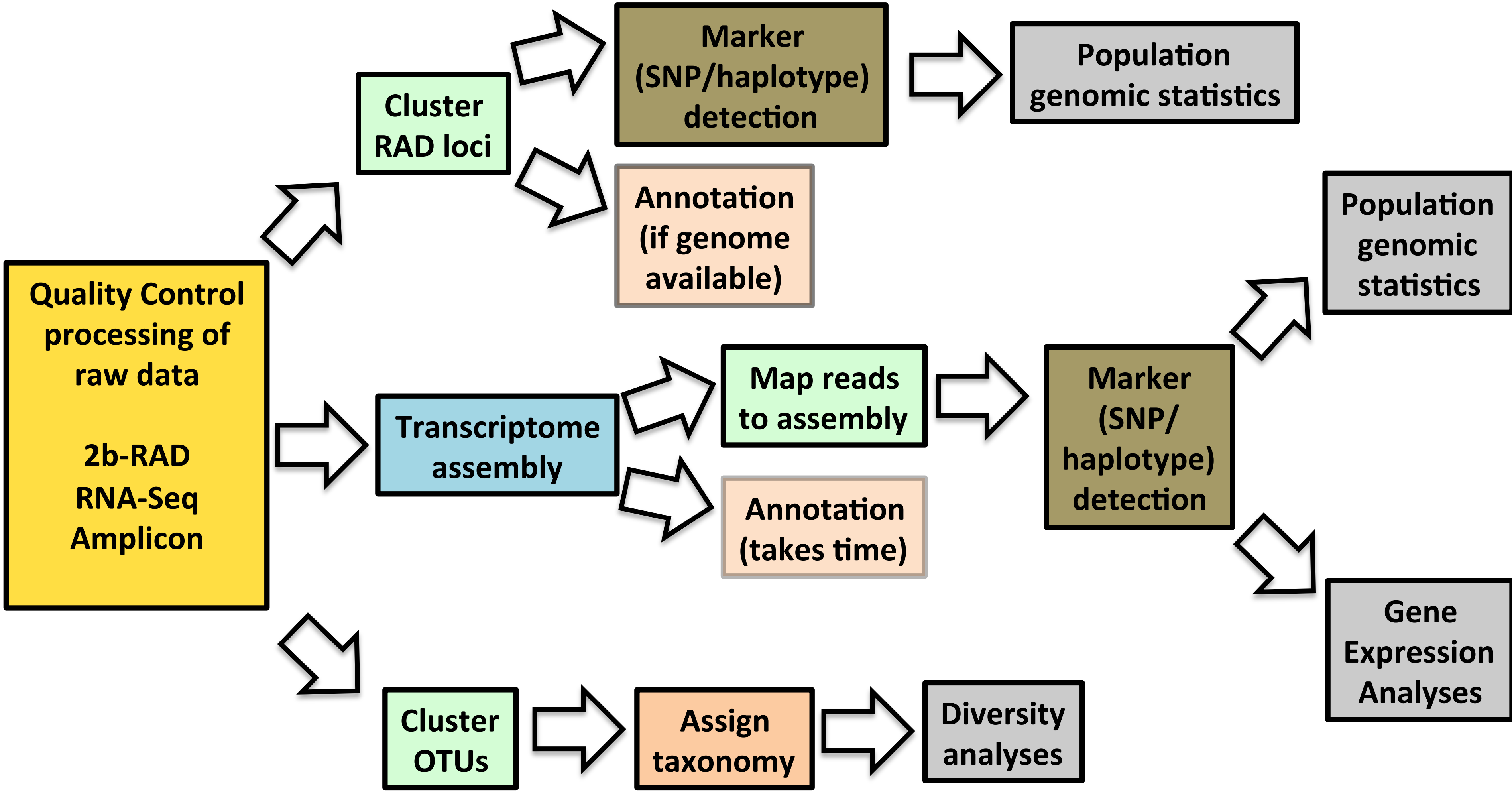
Map with BWA for Illumina ✖
FASTQ file
output (sam)

Mark Duplicate reads ✖
SAM/BAM dataset to mark duplicates in
out_file (bam)
html_file (html)

Unified Genotyper ✖
BAM file 1 > BAM file
output_vcf (vcf)
output_metrics (txt)
output_log (txt)

Variant Filtration ✖
Variant file to annotate
Mask ROD file
output_vcf (vcf)
output_log (txt)

Data processing steps for the next four days



Quality Control processing of raw data

2b-RAD
RNA-Seq
Amplicon

Cluster RAD loci

Marker (SNP/haplotype) detection

Population genomic statistics

Annotation (if genome available)

Transcriptome assembly

Map reads to assembly

Annotation (takes time)

Marker (SNP/ haplotype) detection

Population genomic statistics

Gene Expression Analyses

Cluster OTUs

Assign taxonomy

Diversity analyses