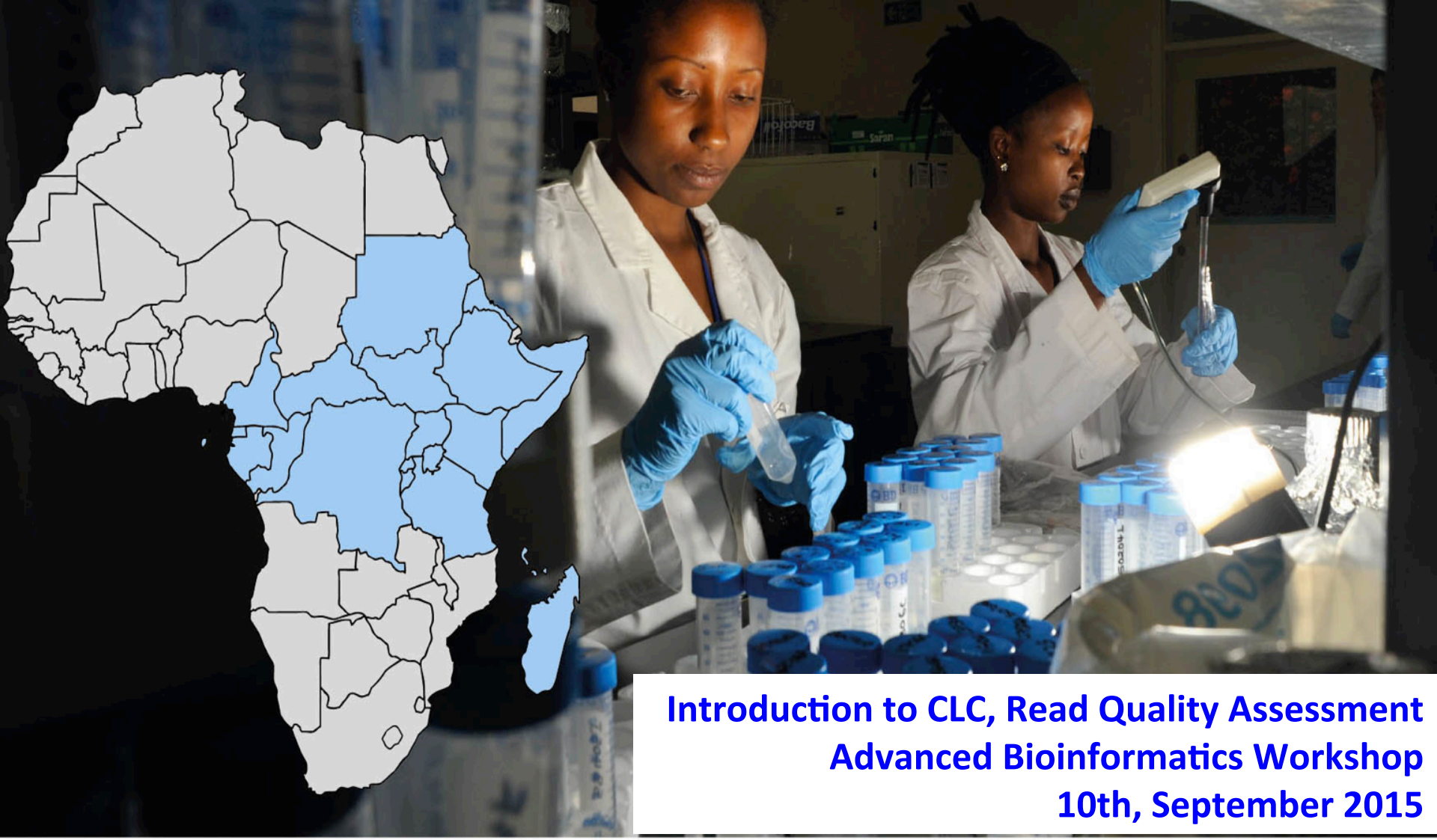


Introduction to CLC, Read Quality Assessment
Advanced Bioinformatics Workshop
10th, September 2015

Joyce Njoki Nzioki
BecA-ILRI Hub, Nairobi, Kenya
http://hub.africabiosciences.org/
http://www.Ilri.org/
j.n.njuguna@cgiar.org

ILRI
INTERNATIONAL LIVESTOCK RESEARCH INSTITUTE

African Union

NEPAD
THE NEW PARTNERSHIP FOR AFRICA'S DEVELOPMENT
A PROGRAMME OF THE AFRICAN UNION

biosciences
eastern and central africa

CLC Genomics Workbench

CLC Main Workbench is a software package that supports analysis of sequence data

Functions include:

- ✓ **Basic Nucleotide and Protein Sequence analysis**
- ✓ **High Throughput Analysis**
- ✓ **Expression Analysis**
- ✓ **Alignment and Phylogeny**
- ✓ **Epigenomics**
- ✓ **etc**

Sanger Sequenced Data

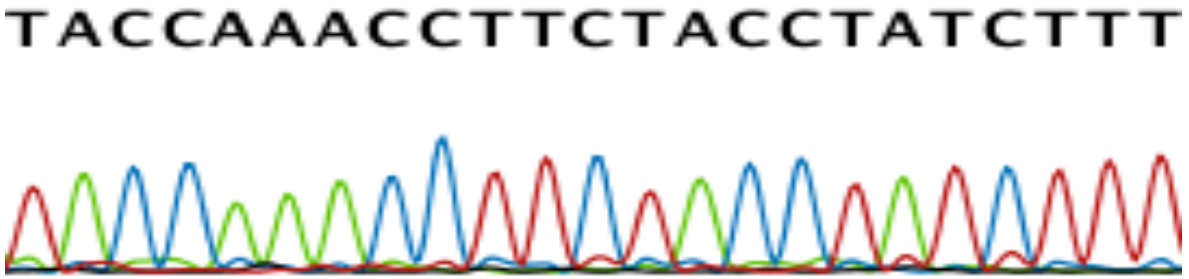
- ✓ You can view your sequences data by opening the sequence files (trace files) extension .ab1 /.abi

- ✓ NOTE: In order to obtain good sequencing results, you MUST download and examine your sequencing chromatogram. If you are using just the text data, you could be publishing data that is completely invalid!

- ✓ Software used for viewing include: CLC bio, BioEdit, TracerView

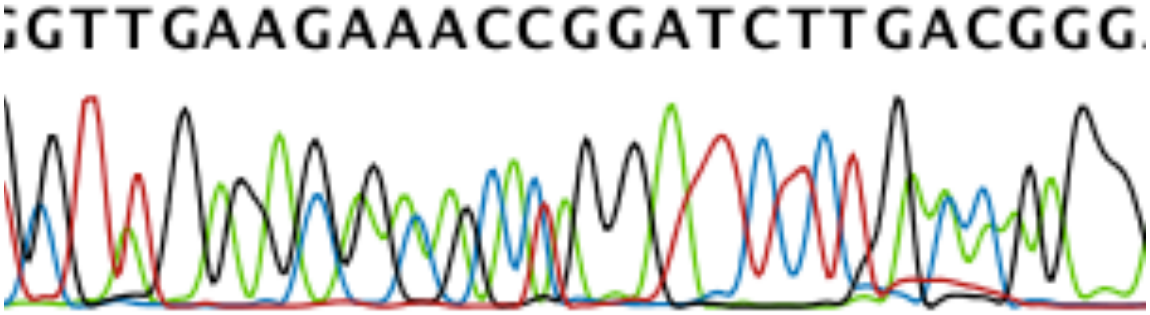
Trouble shoot sequenced data "the good"

- Good quality peaks are smooth, distinct or well formed, evenly spaced and with little baseline noise



Trouble shoot sequenced data "the bad"

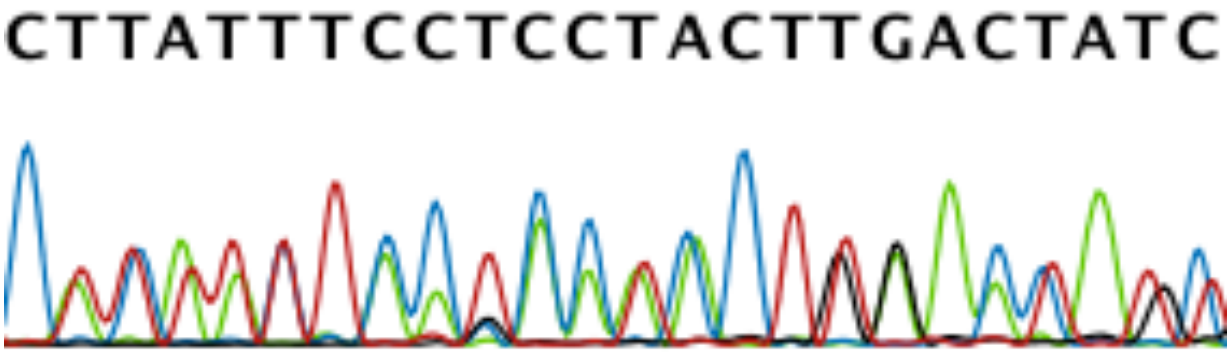
- ✓ A failed sequencing reaction: the chromatographs look messy, many 'N's in the sequence.
- ✓ Non-usable sequenced data: can be due to low concentration of DNA template, none or wrong primer added.



Trouble shoot sequenced data "double peaks"

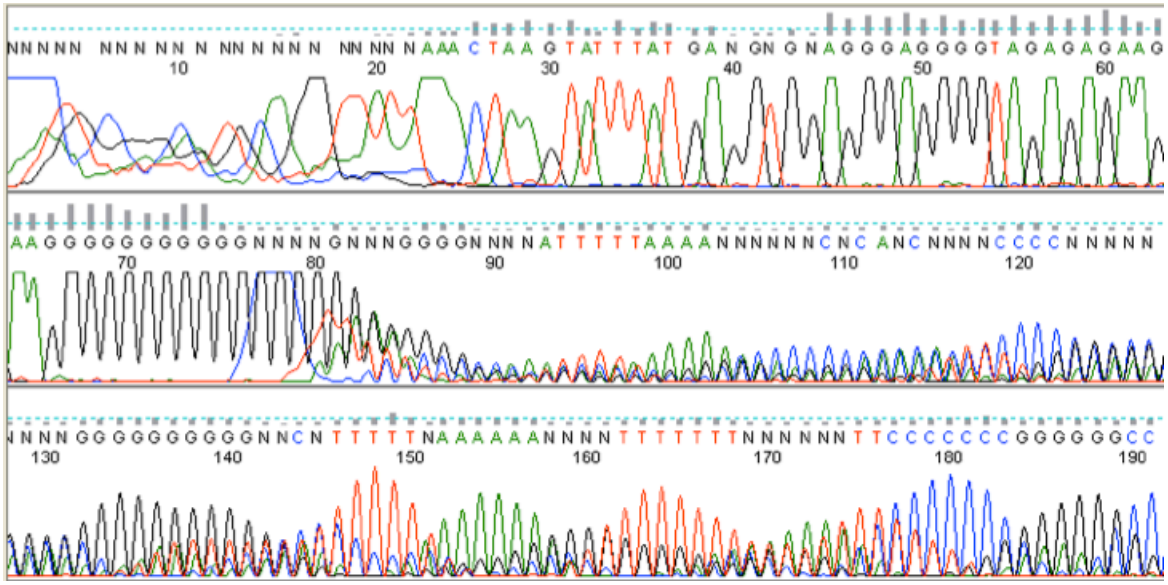
- ✓ Double peaks: multiple peaks of same or different length at the same position; this is due to clone contamination, heterozygous position (SNP), contaminated PCR reaction
- ✓ Can be corrected using degenerate codes;

  N (a c t g) , Y (c t), R (a g)



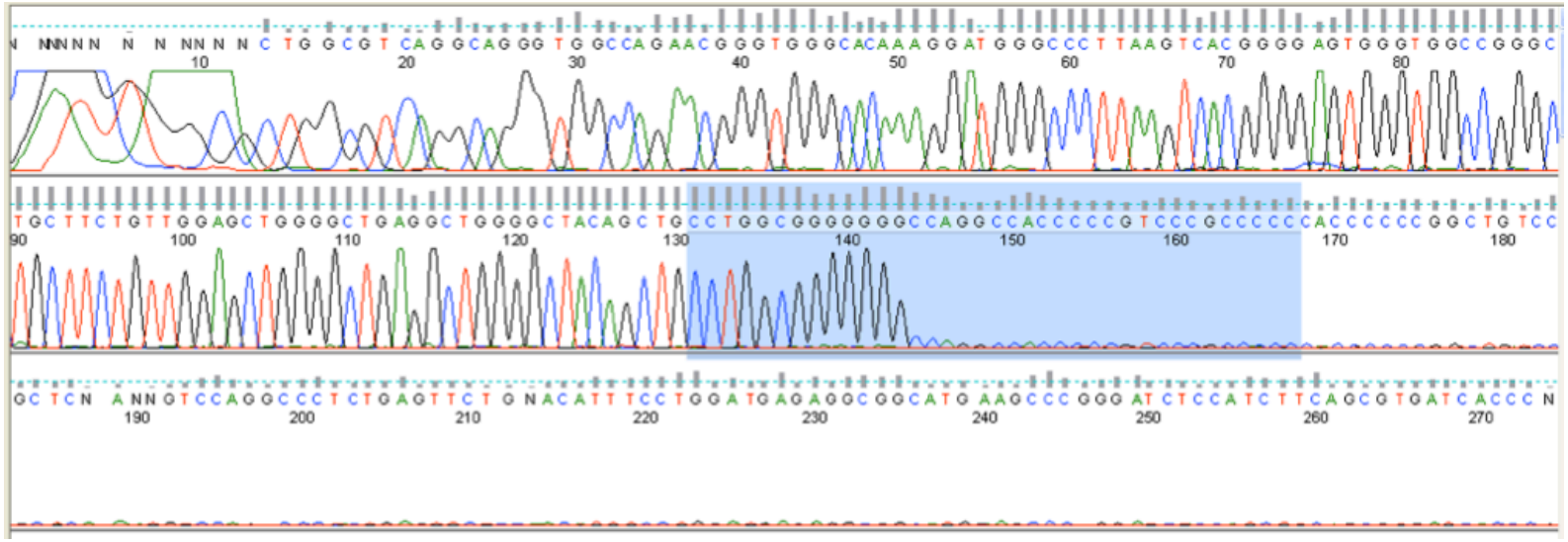
Trouble shoot sequenced data "stuttering"

- ✓ Sequence data quality is poor after stretches of 7 or more nucleotides of the same base. This is due to polymerase slippage during DNA synthesis, it's a limitation of sanger



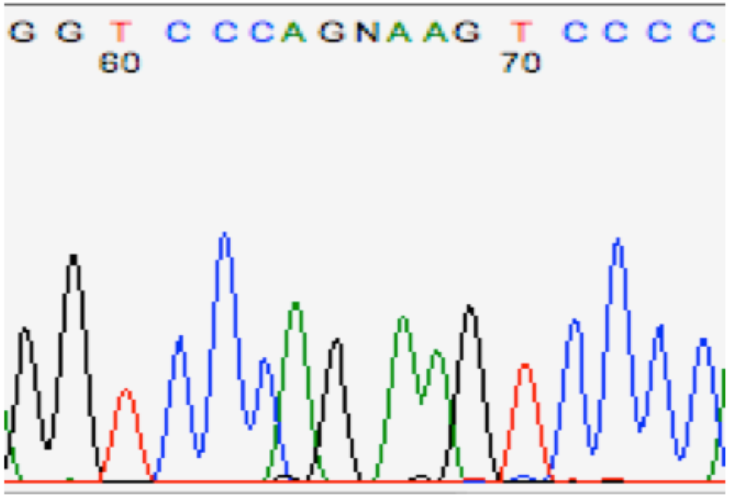
Trouble shoot sequenced data "drop off"

- ✓ The DNA sequence suddenly stops or peak intensely drops off substantially. This is caused by secondary structures like hairpin loops or GC/GT rich regions.



Trouble shoot sequenced data "mis-called bases"

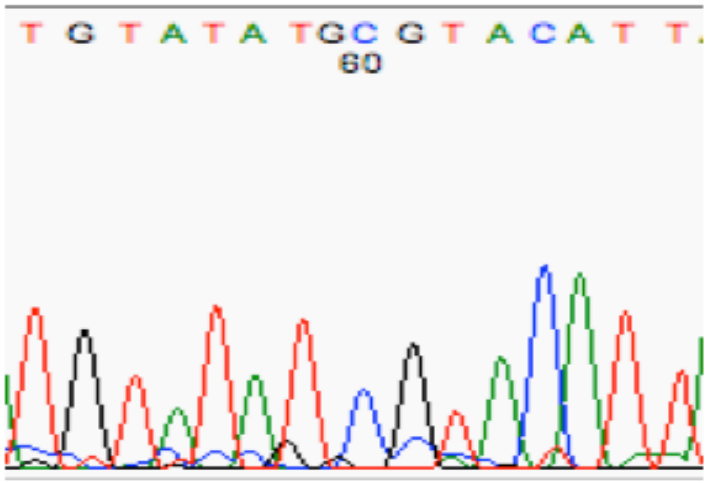
- ✓ Nucleotides that have been erroneously inserted into a sequence will appear oddly spaced relative to their neighboring bases



'N' called in the space between the G-A pair. That is an erroneous call; there is no missing base 'N' at that position.

Trouble shoot sequenced data "mis-called bases"

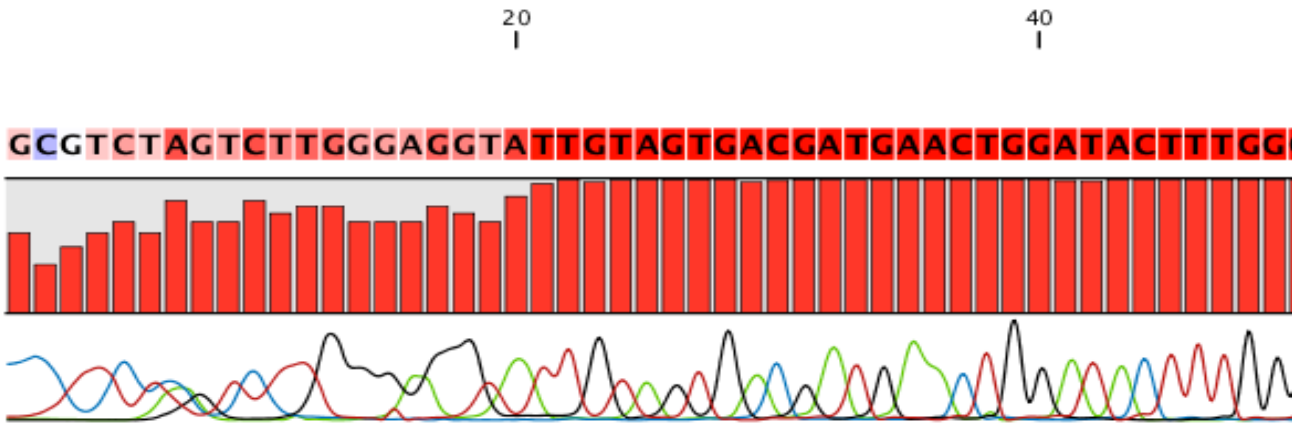
- ✓ Nucleotides that have been erroneously inserted into a sequence will appear oddly spaced relative to their neighboring bases



Note the real T peak (nt 58) and the real C peak (nt 60), with the G barely visible between them. Despite it size, the baseline-noise G peak was picked as if it were real.

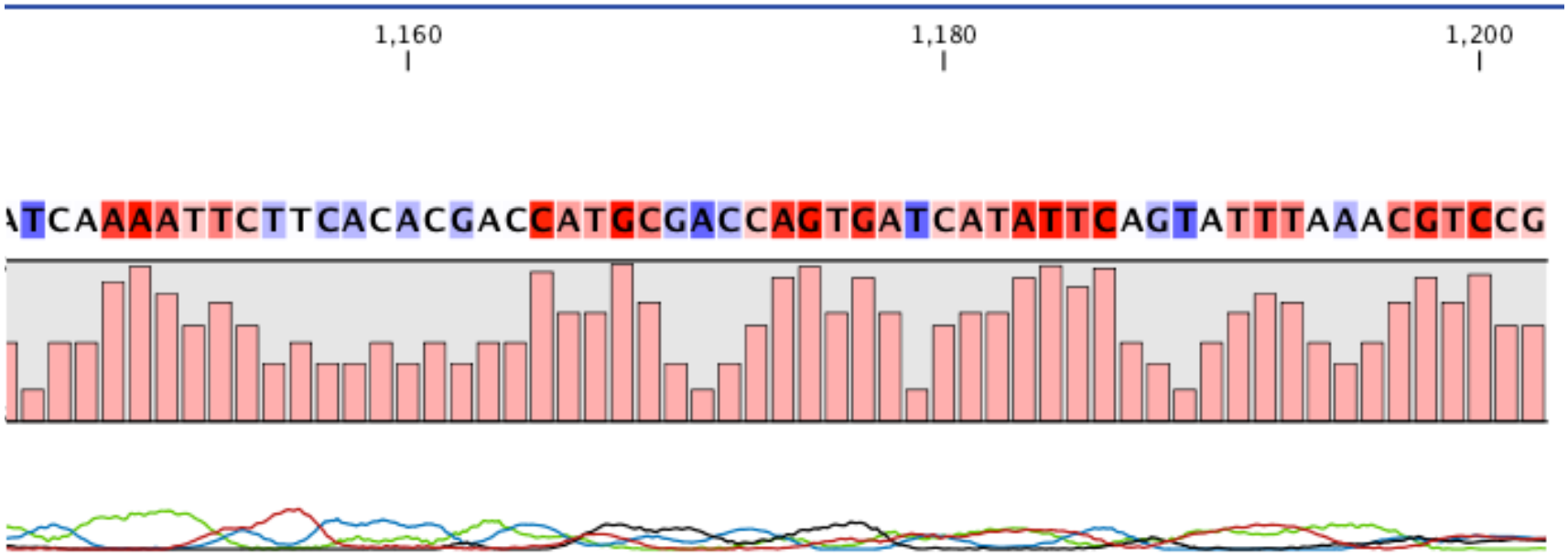
Trim 3’ and 5’ ends

At 5’ end sequences don’t start of very clearly till about bases 20-30 bases. Due to non-fully activated taq polymerase / poor termination near the primer



Trim 3’ and 5’ ends

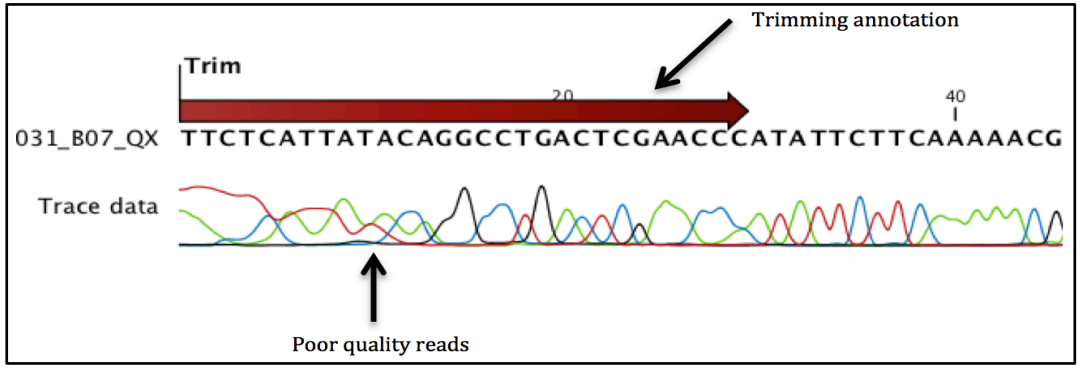
At 5’ end towards the end base 500-800 the quality will degrade as well. due to diminishing bases.



Quality Control using CLC

- ✓ The first step in sequence analysis is to check the quality of reads and trim sequences where need be to eliminate poor quality or vector contamination.

- ✓ When the trimming is done the parts of the sequences that are trimmed are not actually removed but trim annotations are saved to the sequences. These annotated sections are ignored in further analysis.



NGS Data Quality Control

NGS-techniques greatly enhance in depth analysis of DNA samples, they however introduce additional error sources. Thus it is important to pass sequences through quality control / assessment

Error modes

Each technology has unique error modes, depending on the physico-chemical processes involved in the whole sequencing life cycle (not just the base calling)

Contaminating sequence within reads

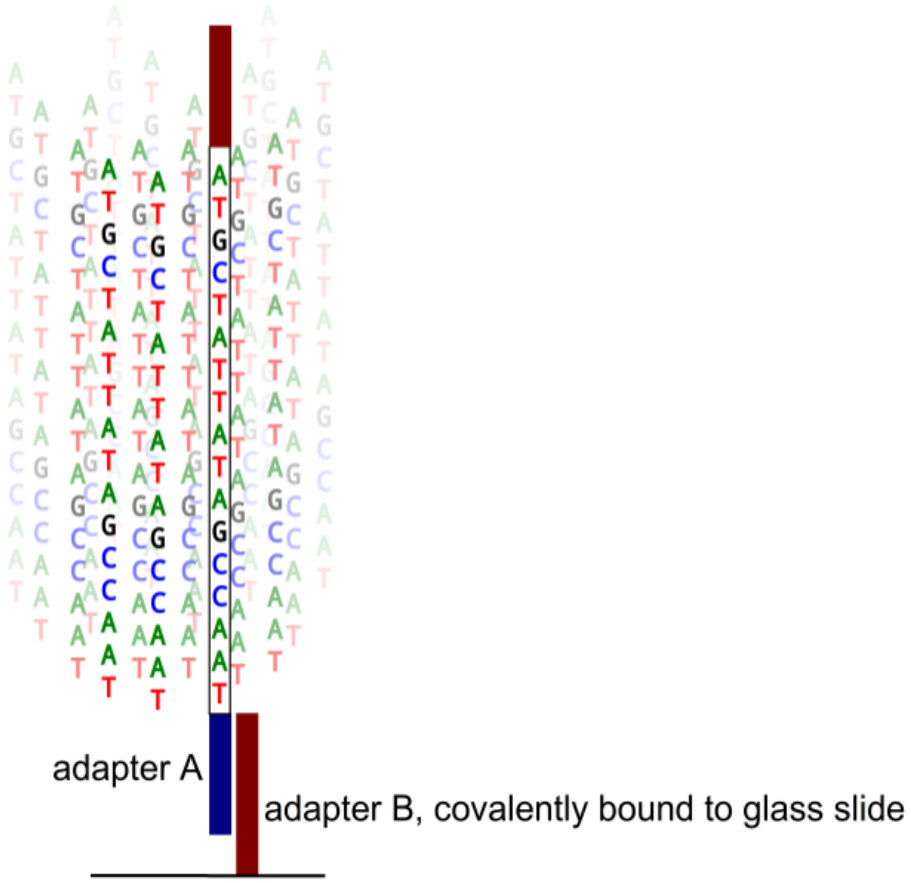
- Adapters

Poor quality sequence

- Substitutions, indels, errors

Illumina

3'-end noise



Illumina

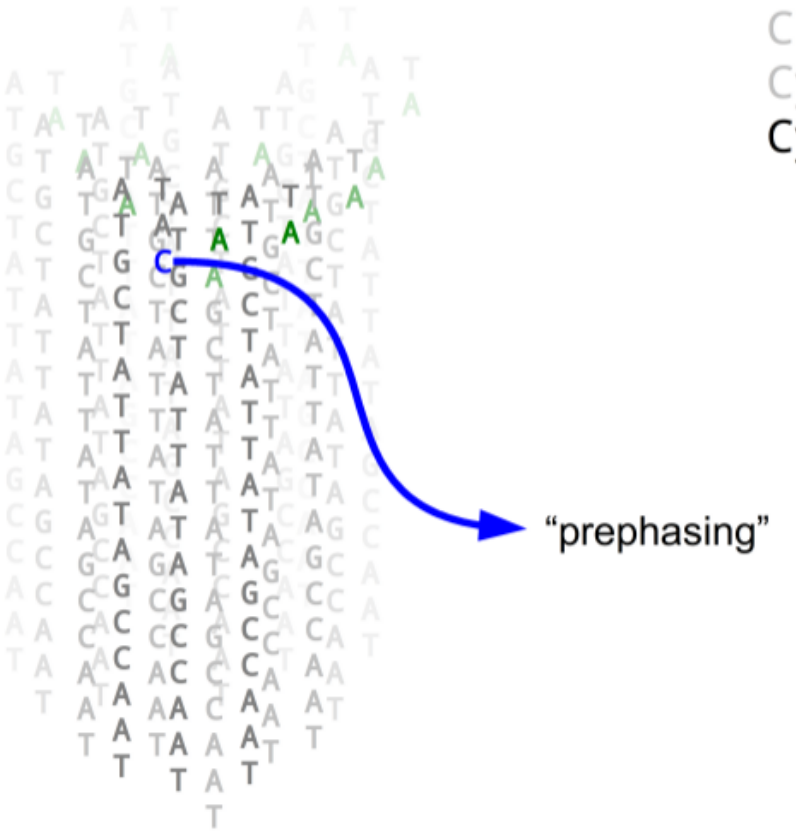
3'-end noise

Cluster generation

Cycle 1 *read as*: T

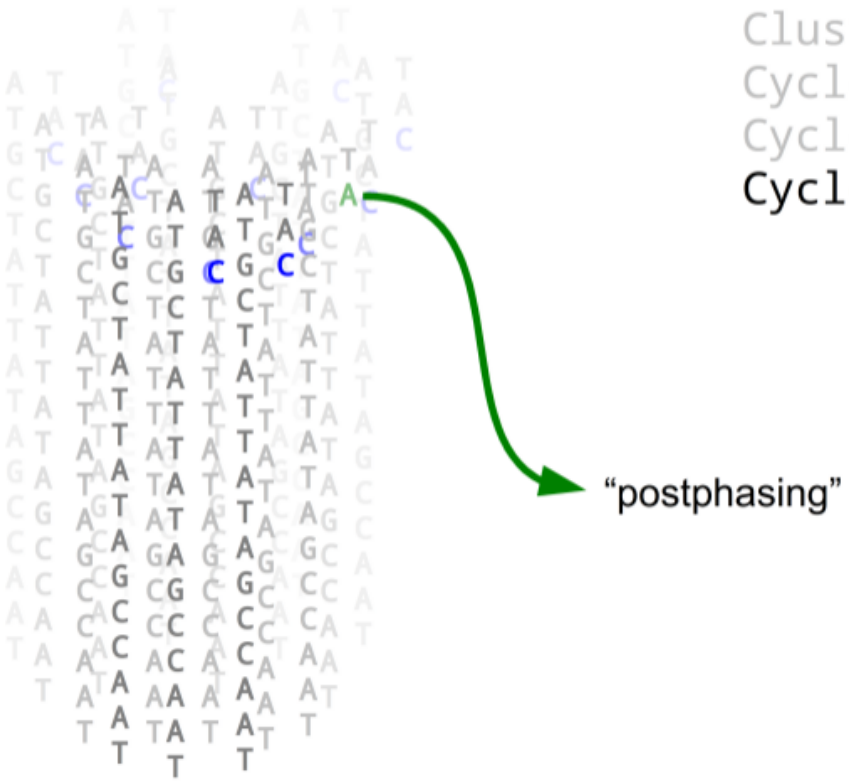
Illumina

3'-end noise



Illumina

3'-end noise



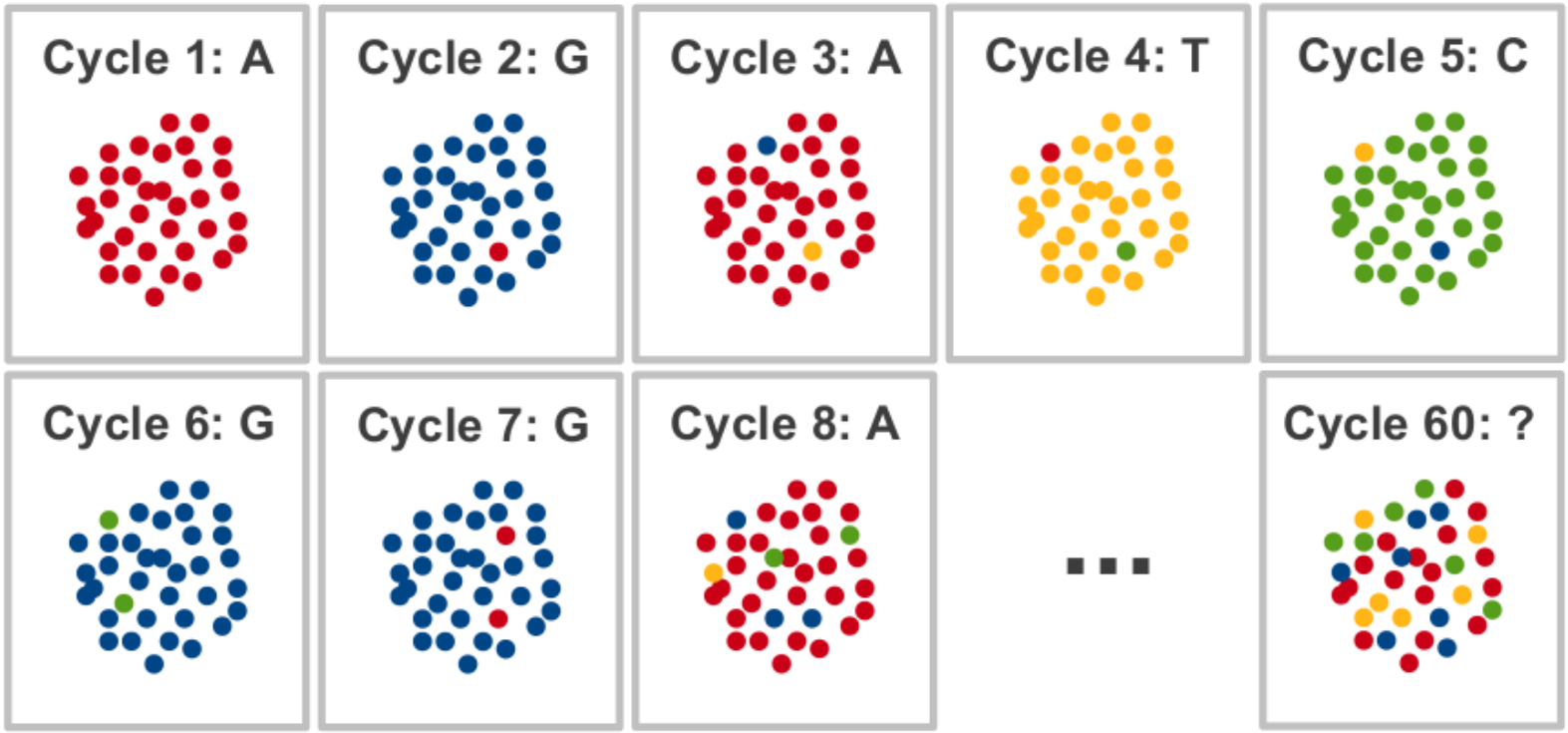
Illumina

3'-end noise

```
Cluster generation
Cycle 1   read as:   T
Cycle 2   read as:   A
Cycle 3   read as:   C
Cycle 4   read as:   G
Cycle 5   read as:   A
Cycle 6   read as:   T
Cycle 7   read as:   A
Cycle 8   read as:   A
Cycle 9   read as:   T
Cycle 10  read as:   A
Cycle 11  read as:   ?
Cycle 12  read as:   ?
Cycle 13  read as:   ?
Cycle 14  read as:   ?
Cycle 15  read as:   ?
Cycle 16  read as:   ?
```

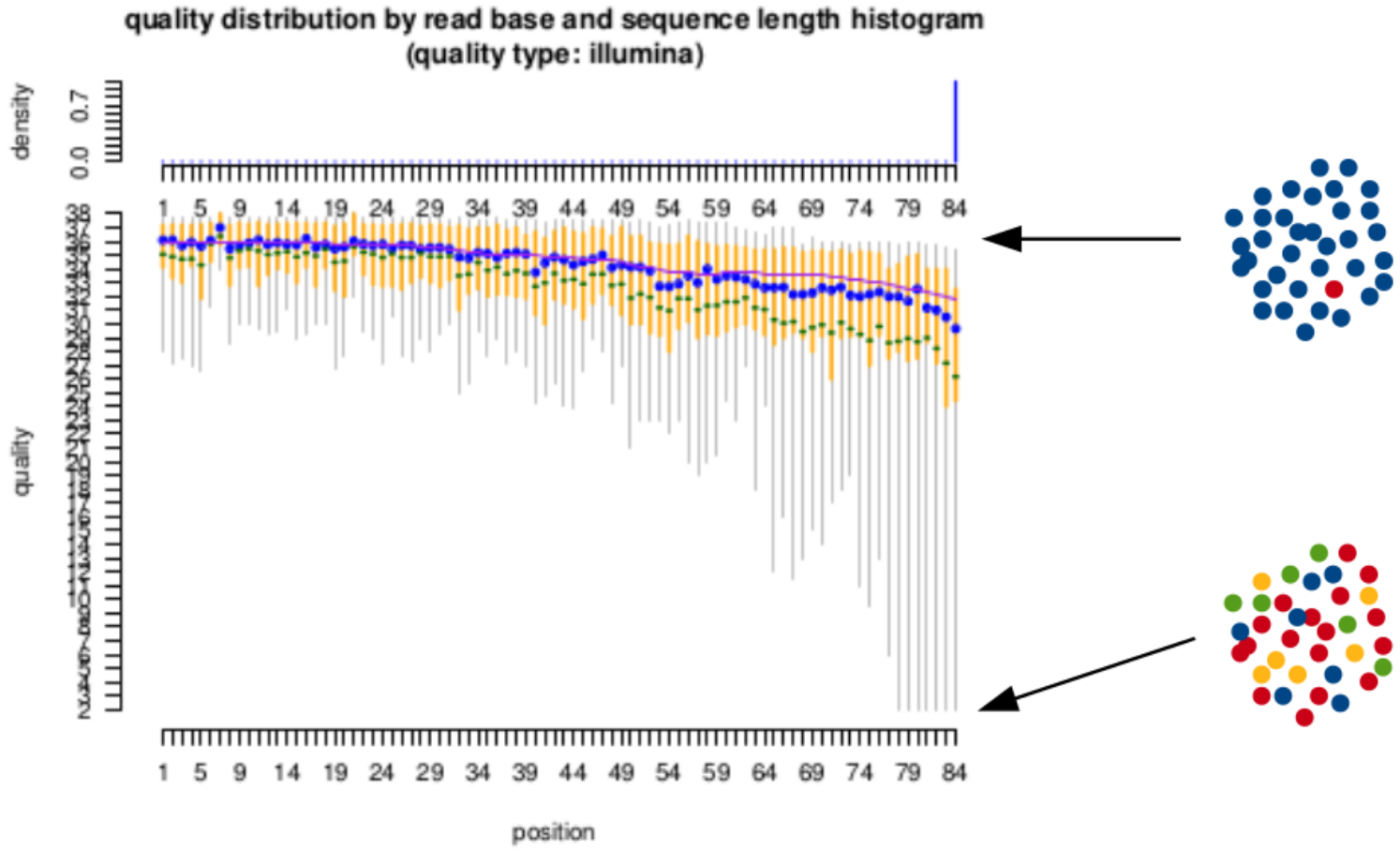
Illumina

3'-end noise



Illumina

3'-end noise



Base qualities

- Sanger-standard: ascii (phred + 33)
- Solexa: ~ascii (phred + 64)
- Illumina pipelines starting with v 1.3: ascii (phred + 64)
- Illumina pipelines starting with v 1.8: ascii (phred + 33)

http://en.wikipedia.org/wiki/FASTQ_format

$$(\text{probability of error}) = 10^{-(\text{phred score})/10}$$

Phred quality score	Probability that the base is called wrong	Accuracy of the base call
10	1 in 10	90%
20	1 in 100	99%
30	1 in 1,000	99.9%
40	1 in 10,000	99.99%
50	1 in 100,000	99.999%

Base Quality in the FASTQ format

Base Quality in the FASTQ format

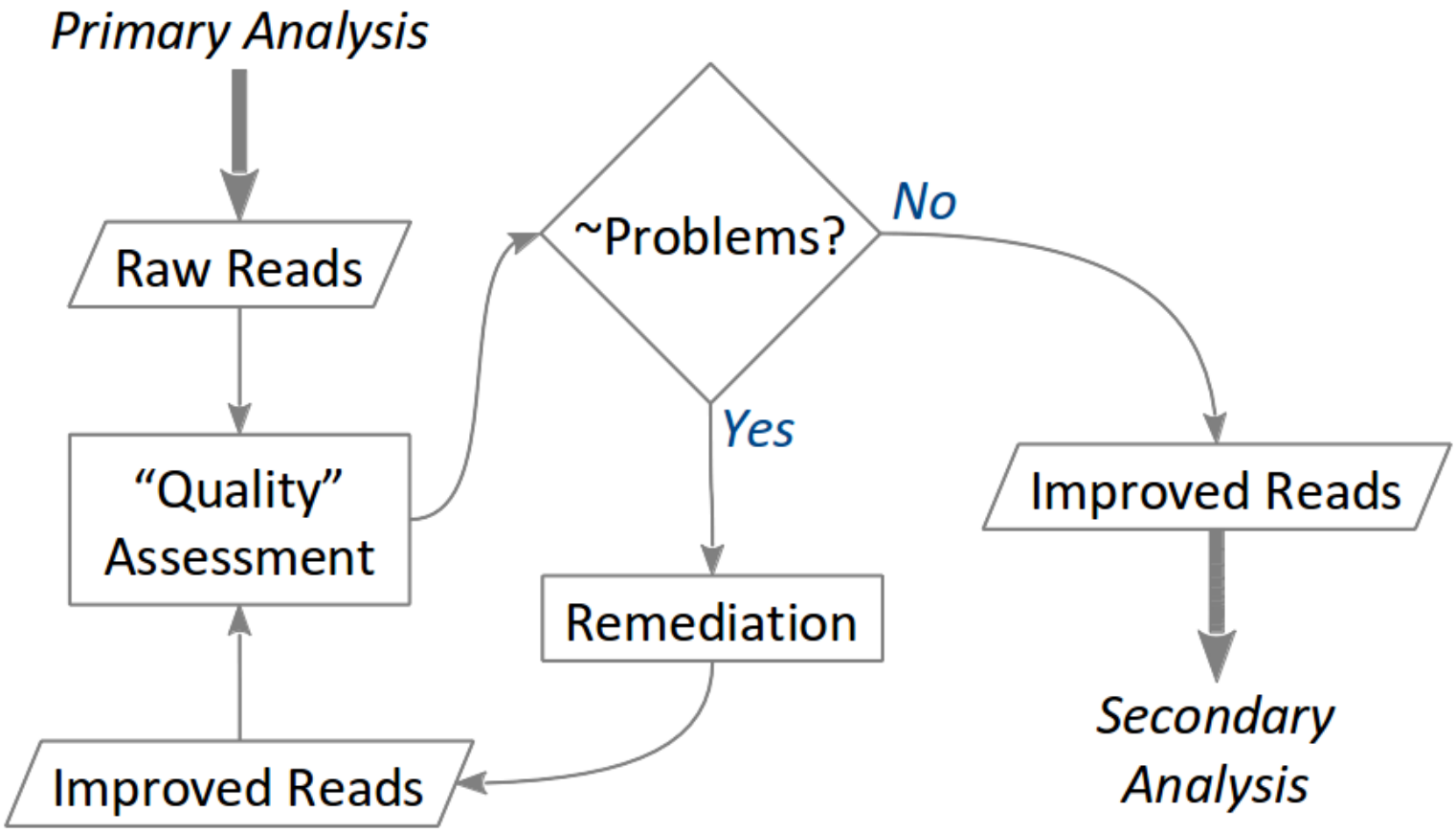
Sanger, Illumina v1.3 to 1.7 (ASCII_BASE=64)

Q	ASCII	P	Q	ASCII	P	Q	ASCII	P	Q	ASCII	P
1	A	0.79433	12	L	0.06310	23	W	0.00501	34	b	0.00040
2	B	0.63096	13	M	0.05012	24	X	0.00398	35	c	0.00032
3	C	0.50119	14	N	0.03981	25	Y	0.00316	36	d	0.00025
4	D	0.39811	15	O	0.03162	26	Z	0.00251	37	e	0.00020
5	E	0.31623	16	P	0.02512	27	[	0.00200	38	f	0.00016
6	F	0.25119	17	Q	0.01995	28	\	0.00158	39	g	0.00013
7	G	0.19953	18	R	0.01585	29	]	0.00126	40	h	0.00010
8	H	0.15849	19	S	0.01259	30	^	0.00100			
9	I	0.12589	20	T	0.01000	31	_	0.00079			
10	J	0.10000	21	U	0.00794	32	`	0.00063			
11	K	0.07943	22	V	0.00631	33	a	0.00050			

Illumina v1.8 and later (ASCII_BASE=33)

Q	ASCII	P	Q	ASCII	P	Q	ASCII	P	Q	ASCII	P
1	"	0.79433	12	-	0.06310	23	8	0.00501	34	C	0.00040
2	#	0.63096	13	.	0.05012	24	9	0.00398	35	D	0.00032
3	$	0.50119	14	/	0.03981	25	:	0.00316	36	E	0.00025
4	%	0.39811	15	0	0.03162	26	;	0.00251	37	F	0.00020
5	&	0.31623	16	1	0.02512	27	<	0.00200	38	G	0.00016
6	'	0.25119	17	2	0.01995	28	=	0.00158	39	H	0.00013
7	(	0.19953	18	3	0.01585	29	>	0.00126	40	I	0.00010
8	)	0.15849	19	4	0.01259	30	?	0.00100	41	J	0.00008
9	*	0.12589	20	5	0.01000	31	@	0.00079			
10	+	0.10000	21	6	0.00794	32	A	0.00063			
11	,	0.07943	22	7	0.00631	33	B	0.00050			

Data "grooming"



QC CLC Genomics Workbench

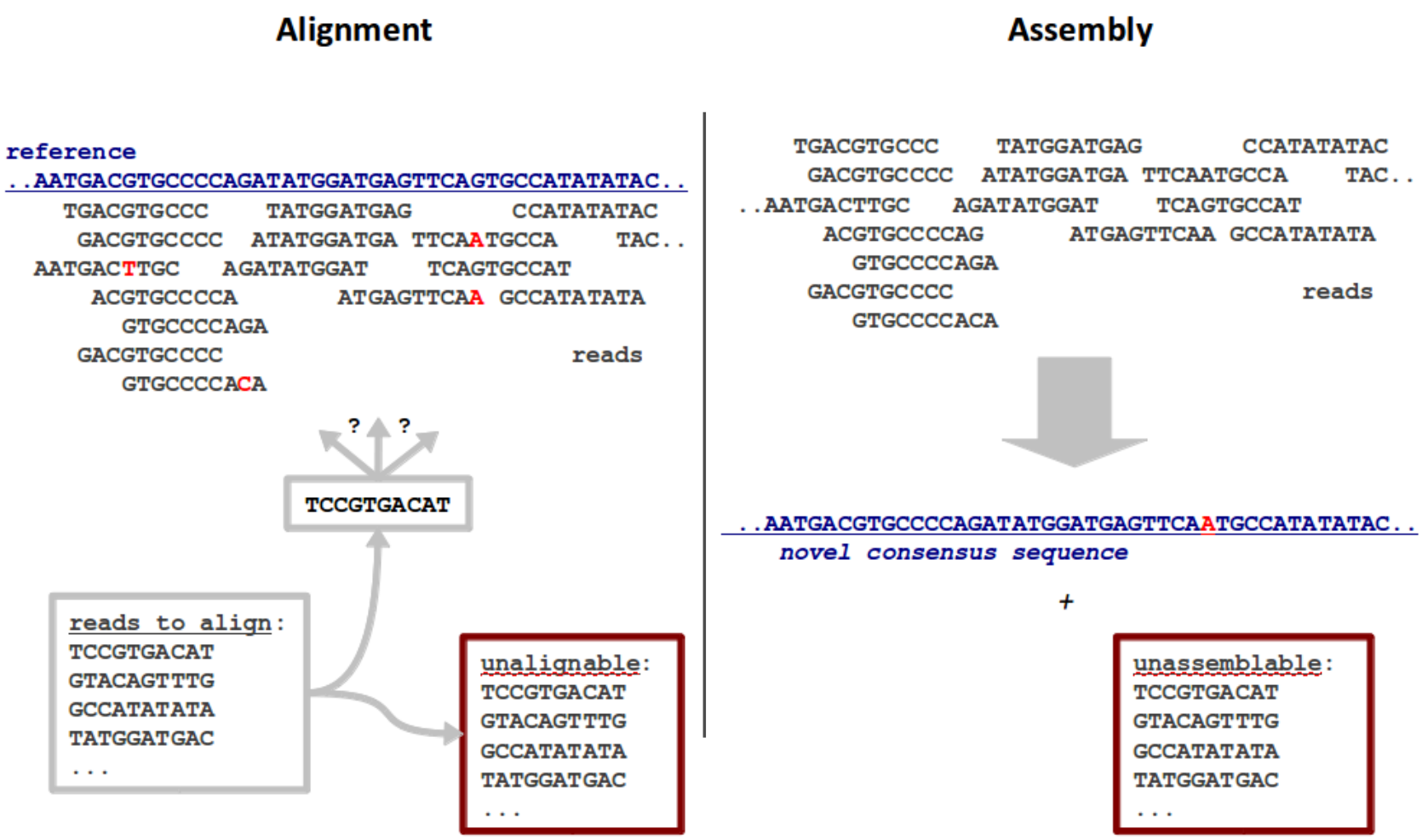
This tool asses the **below quality indicators**

- Sequence-read lengths and base coverage
- Nucleotide contributions and base ambiguities
- Quality scores as emitted by the base caller
- Over represents sequences and hints suggesting contamination events
- Adaptor contamination
- Example QC report

Trimming Sequences

- There are a number of ways to trim your sequences prior to downstream analysis
- Quality trimming based on quality scores
- Ambiguity trimming to trim off “N’s”.
- Adapter trimming
- Base trim to remove specific number of bases to either 3’ or 5’ ends of the reads
- Length trimming to remover reads shorter than a specific threshold

From reads to *molecules*



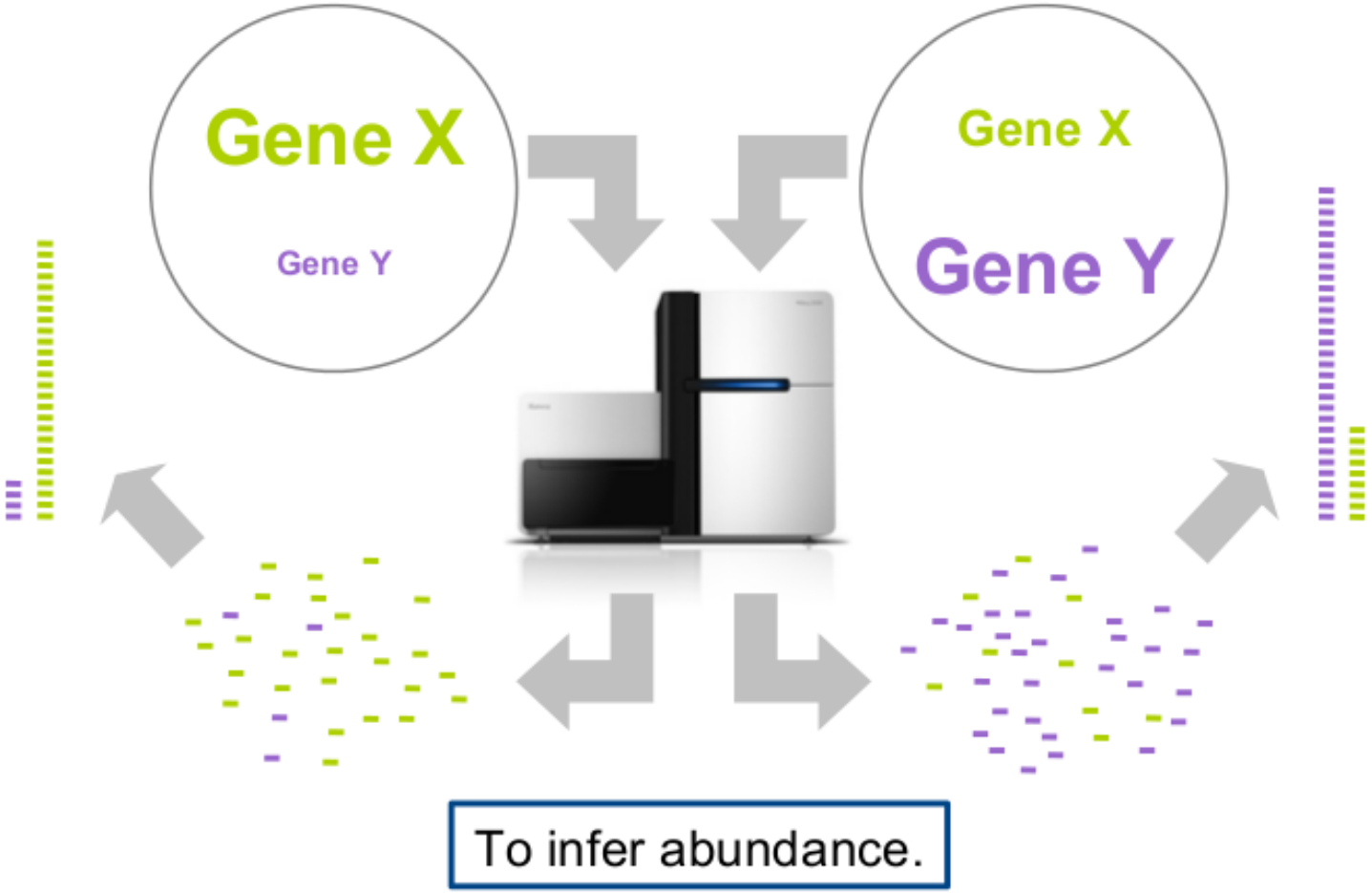
Why align?

Individual A → **Individual B**

```
ATGATAGCATCGTCGGGTGTCTGCTCAATAATAGTGCCGTATCATGCTGGTGTTATAATCGCCGCATGACATGATCAATGG
                                CAATAAAAGTGCCGTATCATGCTGGTGTTACAATCGCCGCA
                                          CGTATCATGCTGGTGTTACAATCGCCGCATGACATGATCAATGG
                                TGTCTGCTCAATAAAAGTGCCGTATCATGCTGGTGTTACAATC
                          ATCGTCGGGTGTCTGCTCAATAAAAGTGCCGTATCATG--GGTGTTATAA
                                CTCAATAAGAGTGCCGTATCATG--GGTGTTATAATCGCCGCA
                                                    GTTATAATCGCCGCATGACATGATCAATGG
```

To measure variation.

Why align?



The End

Acknowledge Joe Fass (UC Davis) for some of the slides

Thank you