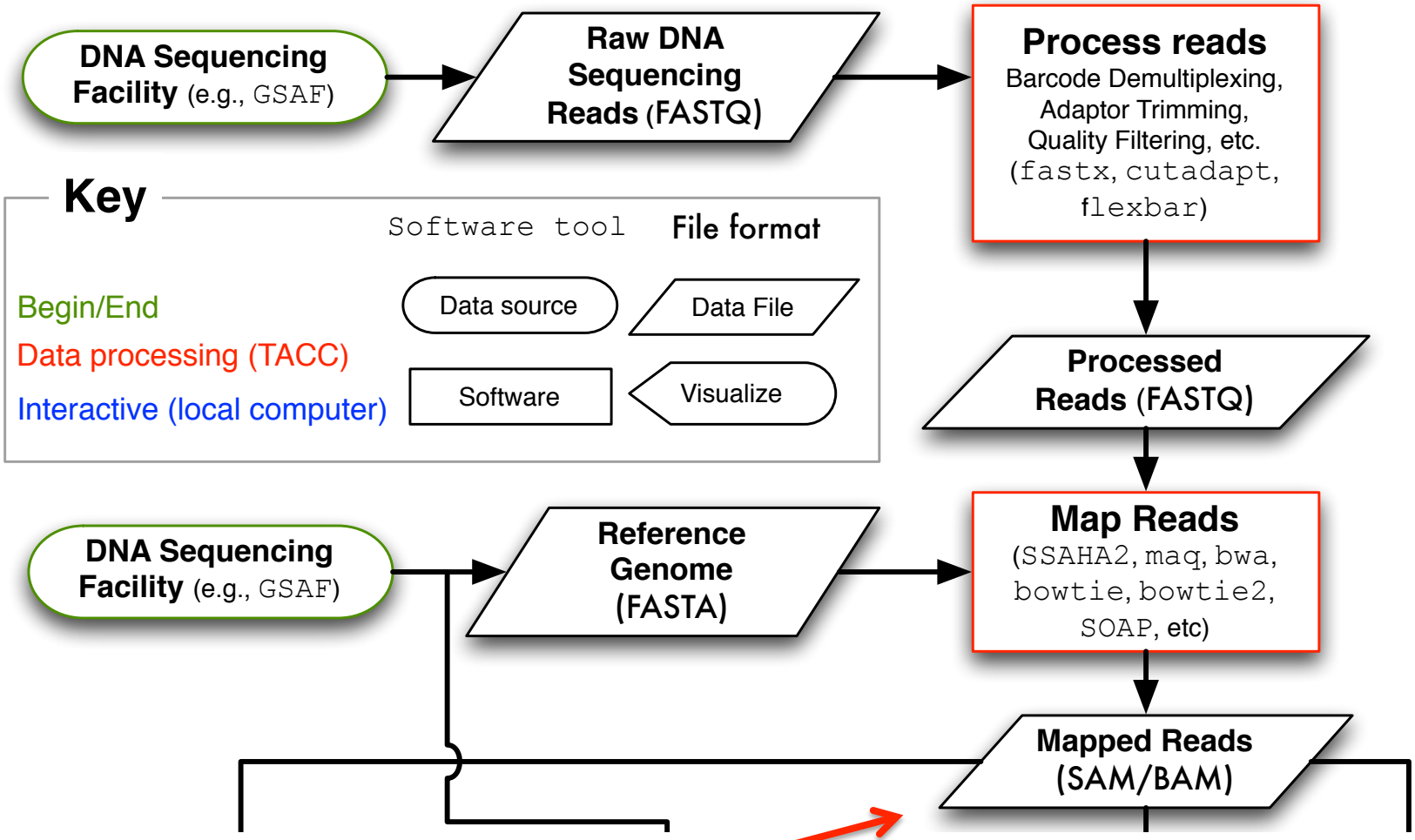
Introduction to Read Mapping

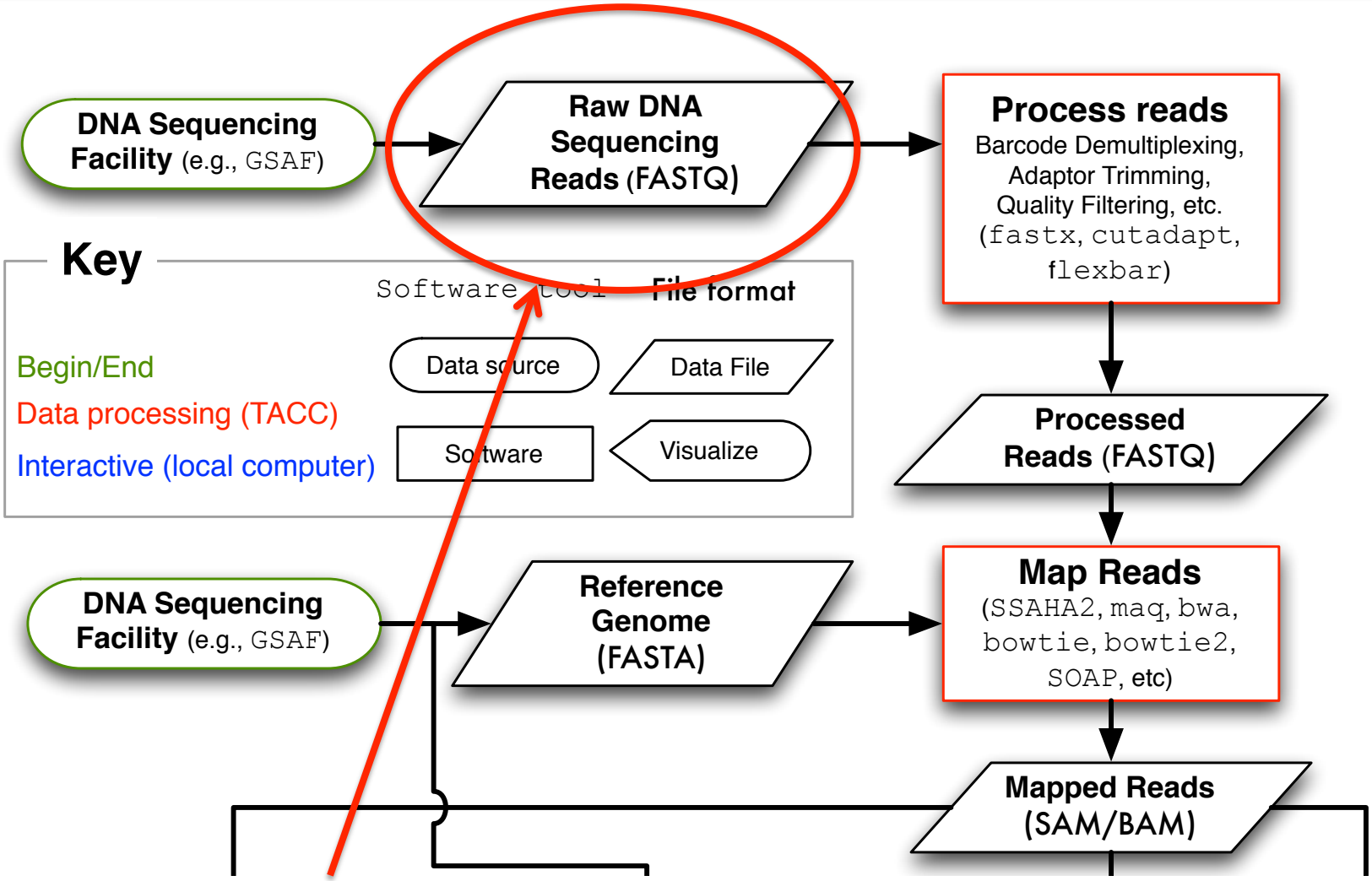



Drowned in next generation sequencing data
HELP!
http://biocomicals.blogspot.com

Basic steps of mapping reads

1. Read file quality control and processing
2. Build reference sequence index
3. Map DNA sequencing reads
   - Exact tool/approach depends on sequencing technology and DNA fragment library type
4. Convert result to SAM/BAM database
5. Application specific analysis...
   - These steps are common to any reference-based (opposed to *de novo*) data analysis.
   - We will use the mapped reads for variant calling.

Input: Raw DNA Sequencing Reads



You are here!

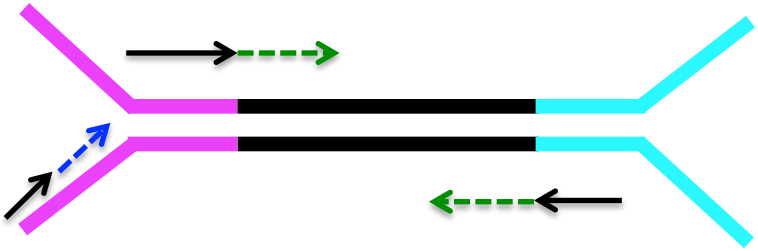
Read terminology

Fragment library (input DNA sample)

Library prep

Sequencing library
Double-stranded or Y- **adaptors** added

DNA sequencing



Barcode (6–12 bases) – so many samples can be run in one physical space (lane). Data is **demultiplexed**.

Primers

Reads (36–1000+ bases)

Types of Illumina fragment libraries

single-end

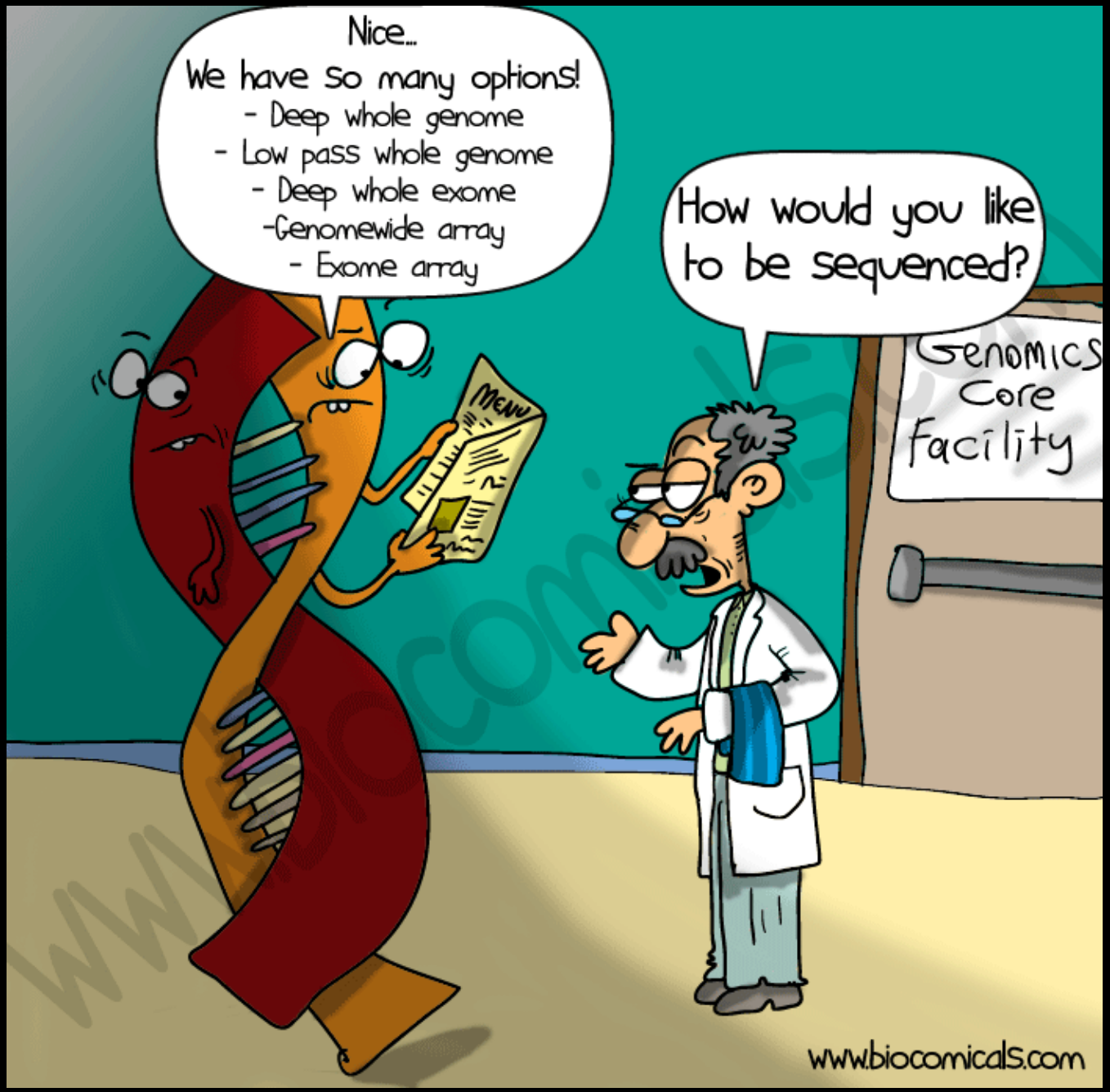
independent reads

paired-end

two inwardly oriented reads separated by ~200 nt

mate-paired

two outwardly oriented reads separated by ~3000 nt


Nice...
We have so many options!
- Deep whole genome
- Low pass whole genome
- Deep whole exome
-Genomewide array
- Exome array
MENU
How would you like to be sequenced?
Genomics Core Facility
www.biocomicals.com

Read file format

FASTQ Format

```
@HWI-EAS216_91209:1:2:454:192#0/1
GTTGATGAATTTCTCCAGCGCGAATTTGTGGGCT
+HWI-EAS216_91209:1:2:454:192#0/1
B@BBBBBB@BBBBAAAA>@AABA?BBBAAB??>A?
```

Line 1: @read name

Line 2: called base sequence

Line 3: +read name (optional after +)

Line 4: base quality scores

Deciphering base quality (Q) scores

http://www.asciitable.com/

```
Quality character   !"#$%&'()*+,-./0123456789:;<=>?@ABCDEFGHI
                    |         |         |         |         |
ASCII Value         33        43        53        63        73
Base Quality (Q)    0         10        20        30        40
```

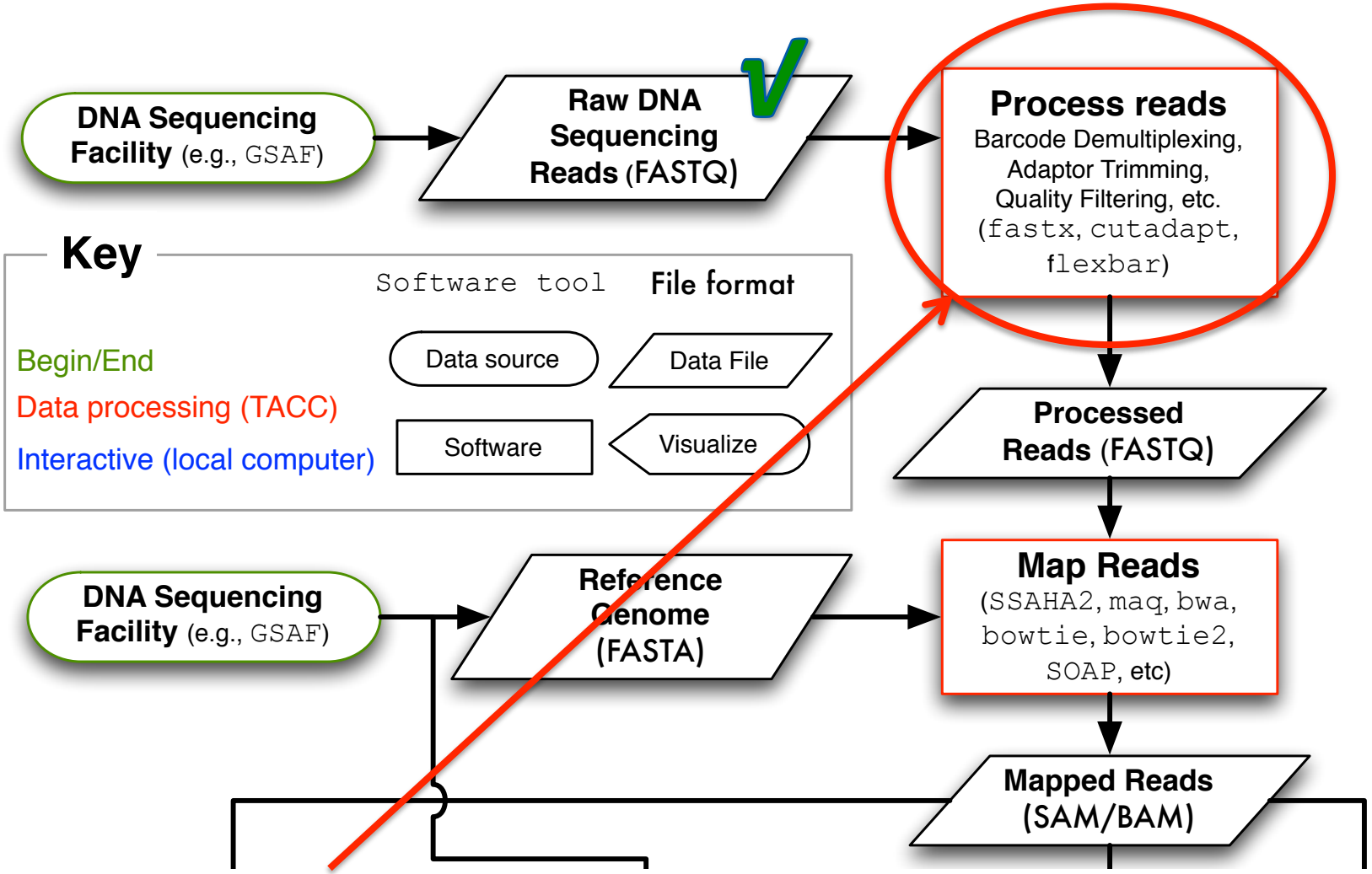
$$\text{Probability of Error} = 10^{-Q/10}$$

(This is a **Phred** score, also used for other types of qualities.)

* Very low quality scores can mean something special – Illumina Q ≤ 3 means something like: "I'm lost, you might want to stop believing sequencing cycles from here on out."

* In older FASTQ files, the formula and ASCII offset might differ. Consult: http://en.wikipedia.org/wiki/FASTQ_format

Input: Raw DNA Sequencing Reads



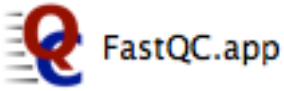
You are here!

Read sequence quality control

Garbage in = garbage out

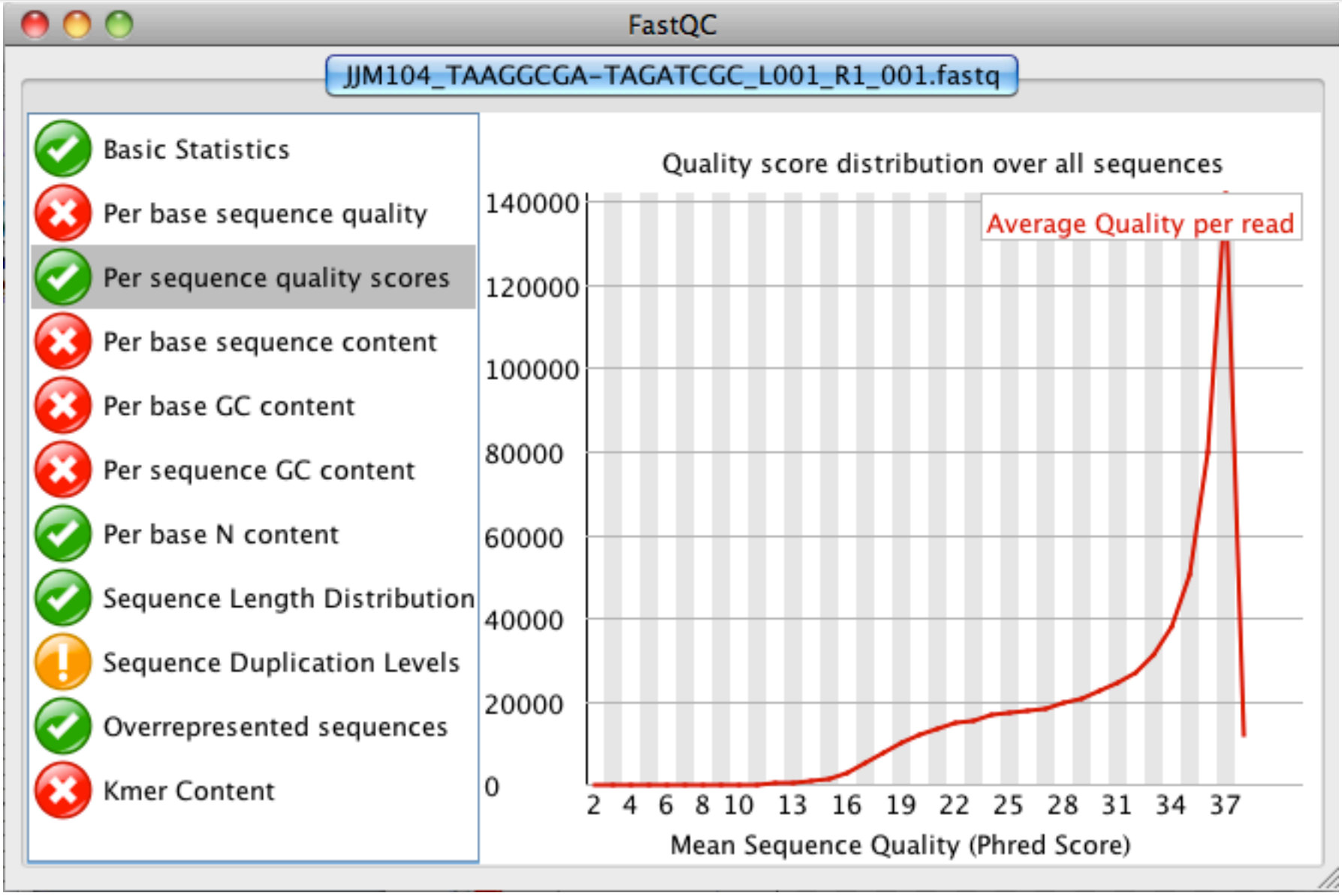
- Contaminated with other samples?
- Sample barcodes removed?
- Adaptor/bar codes trimmed?
  - Esp. important for MiSeq data
- Trim ends of reads with poor quality?
  - Less data but higher quality data
- Know your data
  - Paired reads? Relative orientations?
  - Technology specific concerns?
    - Error hotspots (e.g., 454 Indels, Illumina GGT)

Read quality control software

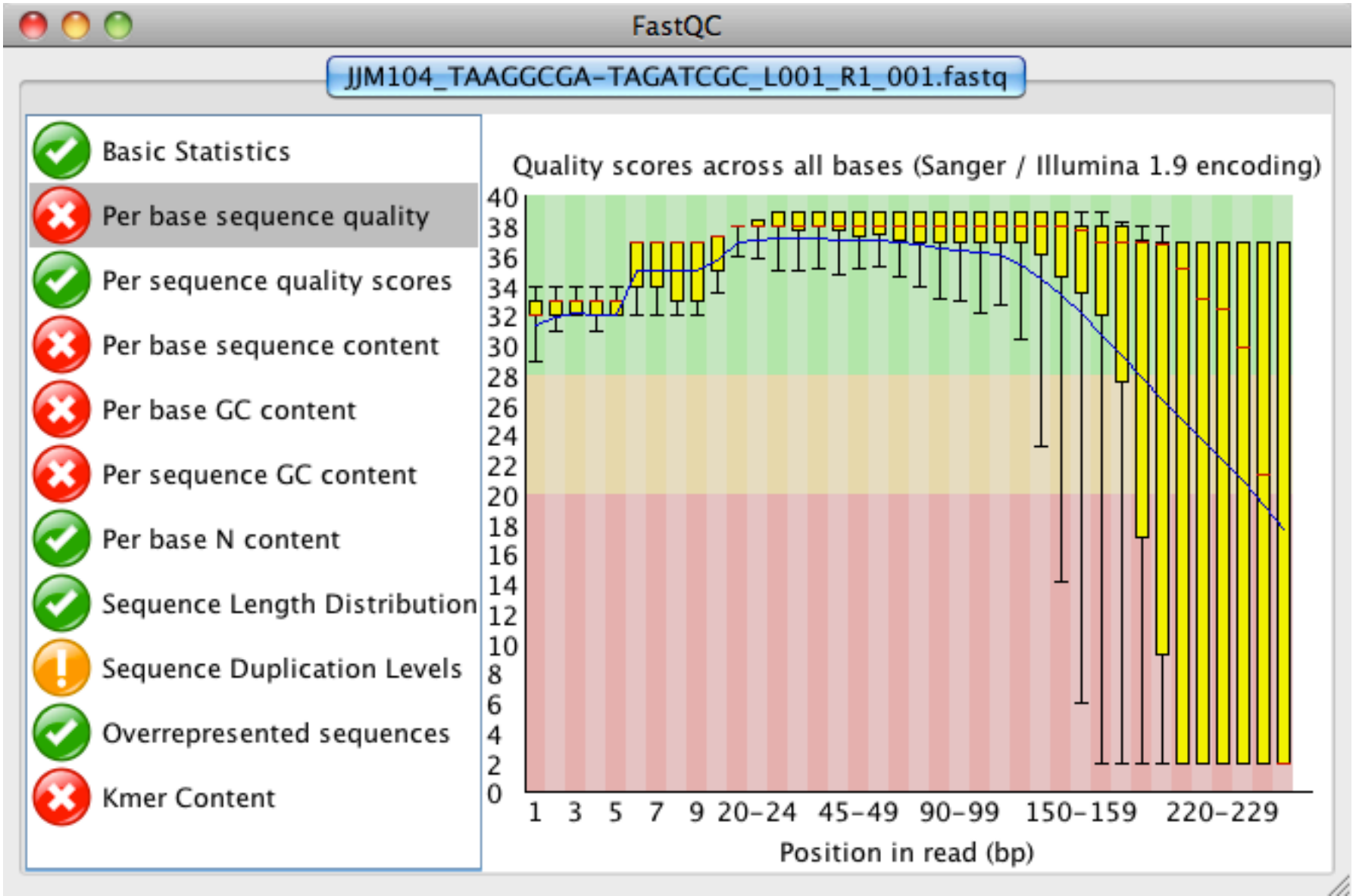
- FastQC is pretty much the only game in town
  - TACC module or run on your own computer
  - Generates nice graphical output 

- Do not be surprised if some criteria "fail" even for really good FASTQ data !!!
- Example FASTQ stats on the next two slides are for the 1st read of a paired-end 250-cycle MiSeq run of *E. coli* DNA.

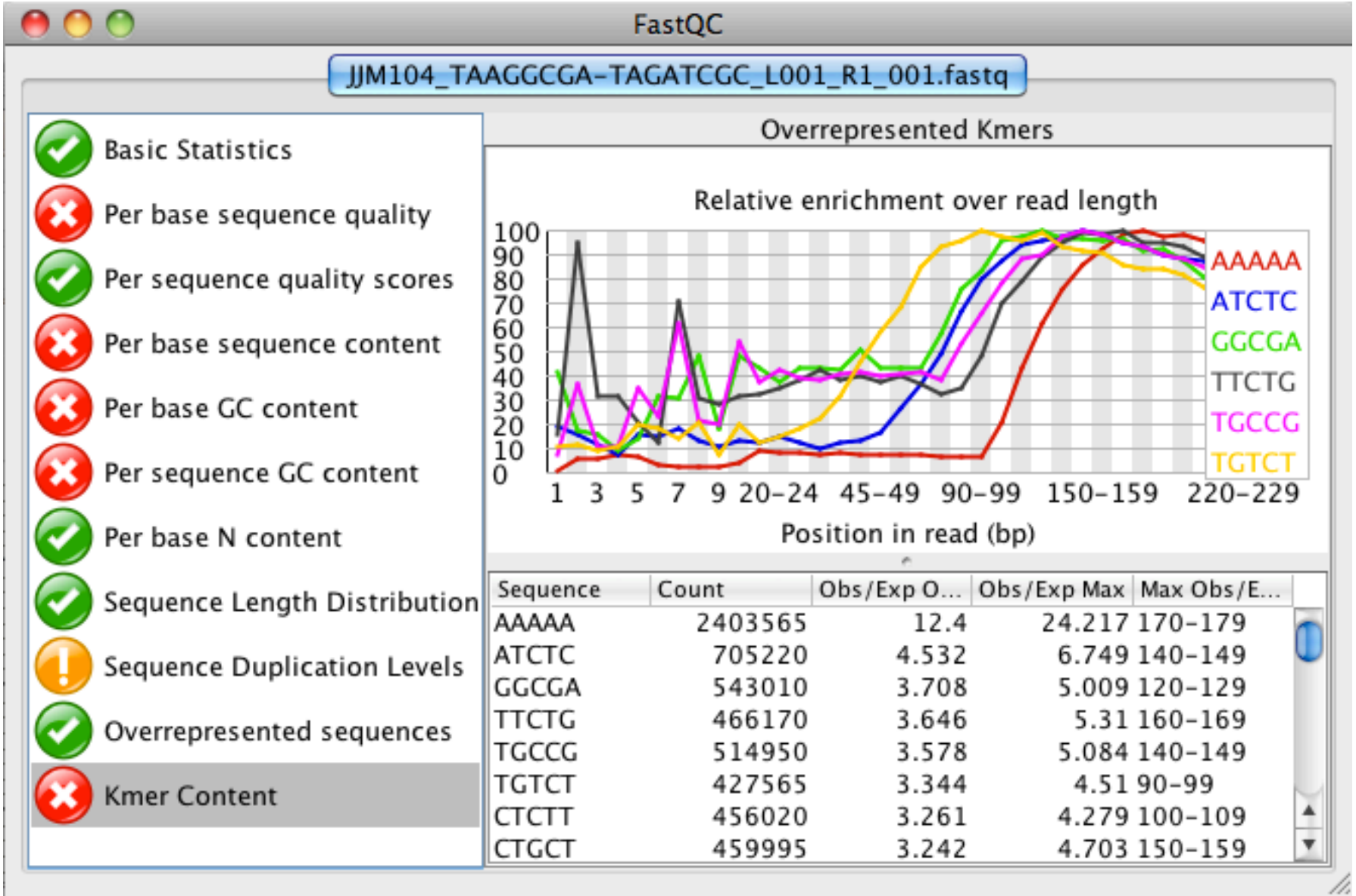
Illumina data example



Illumina data example



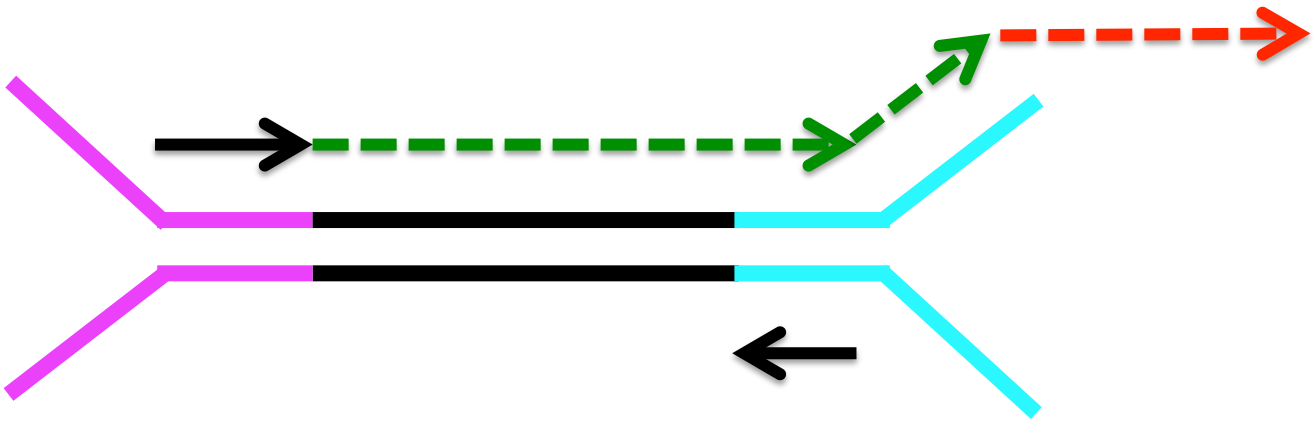
Illumina data example



Sequence	Count	Obs/Exp O...	Obs/Exp Max	Max Obs/E...
AAAAA	2403565	12.4	24.217	170-179
ATCTC	705220	4.532	6.749	140-149
GGCGA	543010	3.708	5.009	120-129
TTCTG	466170	3.646	5.31	160-169
TGCCG	514950	3.578	5.084	140-149
TGTCT	427565	3.344	4.51	90-99
CTCTT	456020	3.261	4.279	100-109
CTGCT	459995	3.242	4.703	150-159

Problem in this data set?

- Adaptor/bar codes trimmed?
  - Esp. important for MiSeq data
- DNA was sheared to smaller than the read length, so many reads extend past the end. They need their 3' ends trimmed of the adaptor and junk sequence.



Read processing software

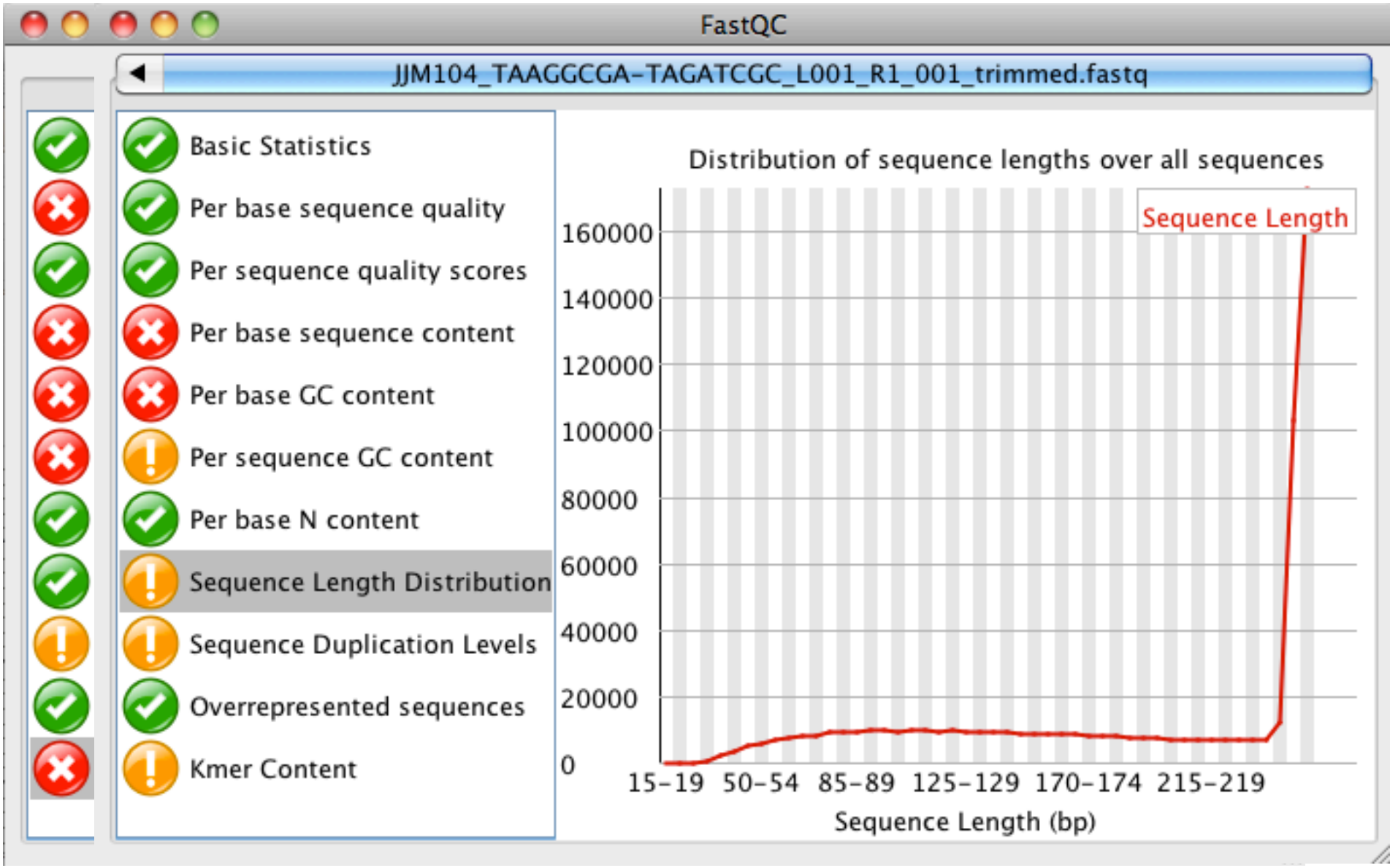
- You may see these aligners commonly used in published workflows:
  - FASTX toolkit, flexbar, cutadapt, trimmomatic
  - Most (except FASTX) not available as TACC modules, but they are installed in **`$BI/bin`**.
- Which to use depends on limitations:
  - Ease of use: cutadapt, trimmomatic
  - Control over matching adaptors: flexbar
  - Paired-end reads cutadapt, flexbar

Adaptor trimming example

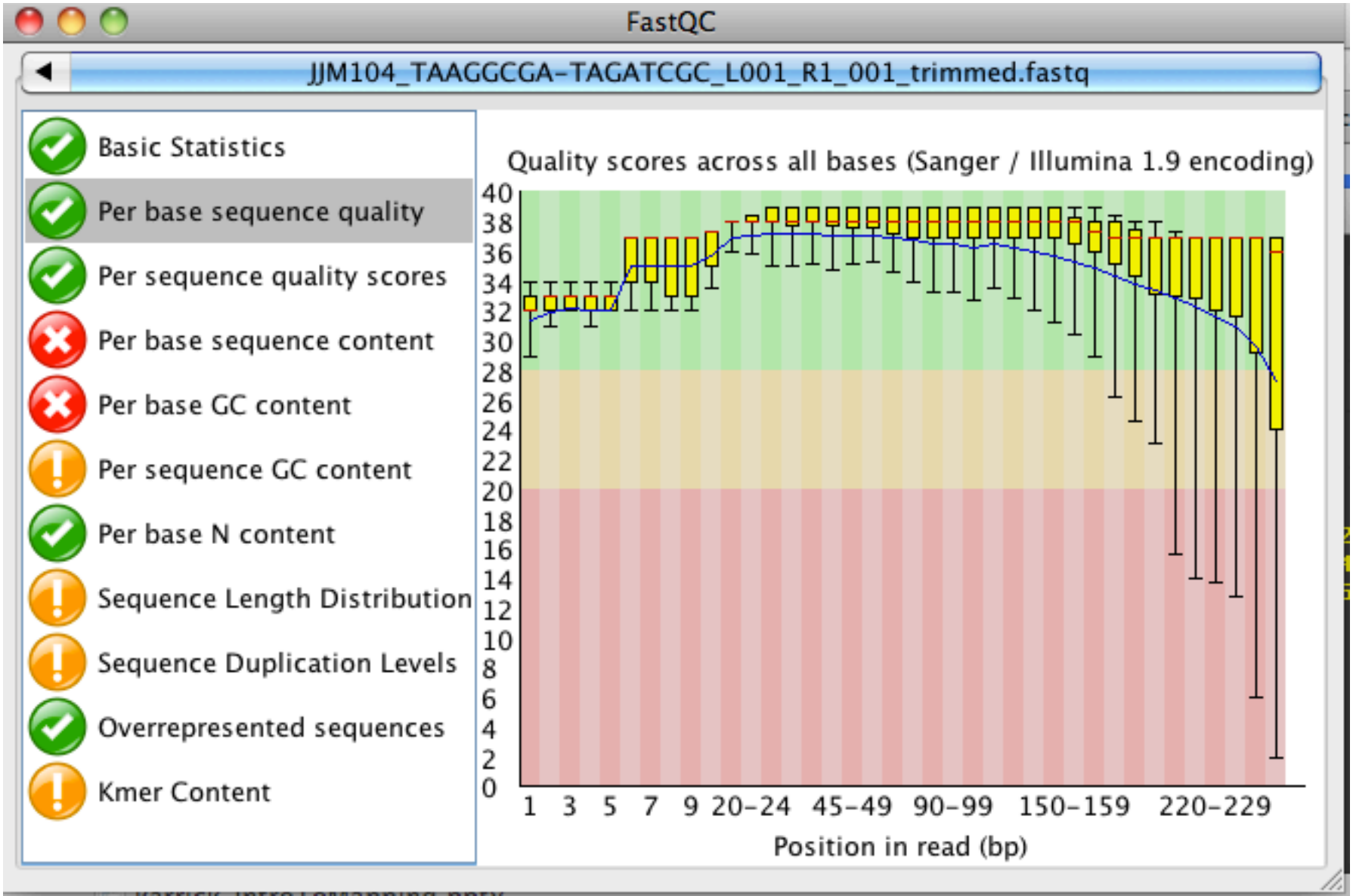
```
$ flexbar -n 12 -t JJM104_TAAGGCGA-TAGATCGC_L001_R1_001_trimmed
-f fastq -r JJM104_TAAGGCGA-TAGATCGC_L001_R1_001.fastq
-a illumina_nextera.fasta
```

```
Adapter removal statistics
==========================
Adapter:           Removed:
read1_end          456282
read2_end          1637

Min, max, mean and median adapter overlap: 1 / 71 / 45 / 66

Output file statistics
======================
Output read file:            JJM104_TAAGGCGA-TAGATCGC_L001_R1_001_trimmed.fastq
Discarded short reads:       7
Reads written to file:       630406

Flexbar completed adapter removal.
```

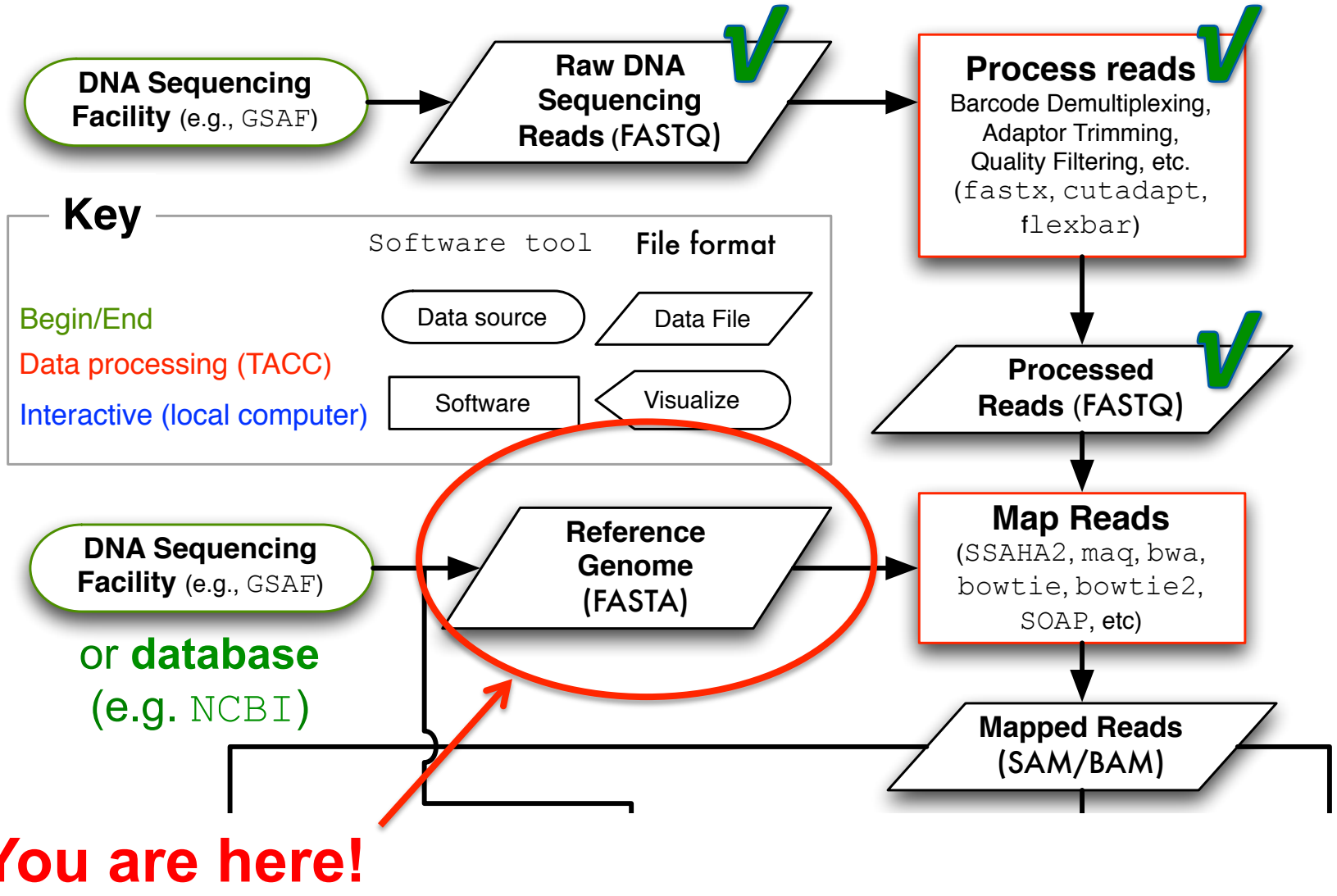
Processed Illumina data example



Processed Illumina data example



Input: Reference Genome



Finding a reference

- **Microbes (<10Mb)**: download FASTA containing in sequence and/or GenBank/EMBL/ GFF flat files encapsulating both sequence and features.
- **Macrobes (>100Mb)**: download specific consortium "build" of reference (Ex: hg19), consisting of FASTA, and various files used to construct a database of feature (PTT, GFF).
- **Non-model organisms**: build your own? *de novo assembly (outside scope of course)*

Reference considerations

- Is it appropriate to your study?
  - Close enough to your species?
  - Complete?
- Which version?
  - Make sure you use an agreed-on standard
- Does it contain repeats? What kinds?
  - Know this up front or you may be confused
- What annotations exist?
  - References lacking feature annotations are much more challenging to use

http://microbialgenomics.energy.gov

Reference sequences

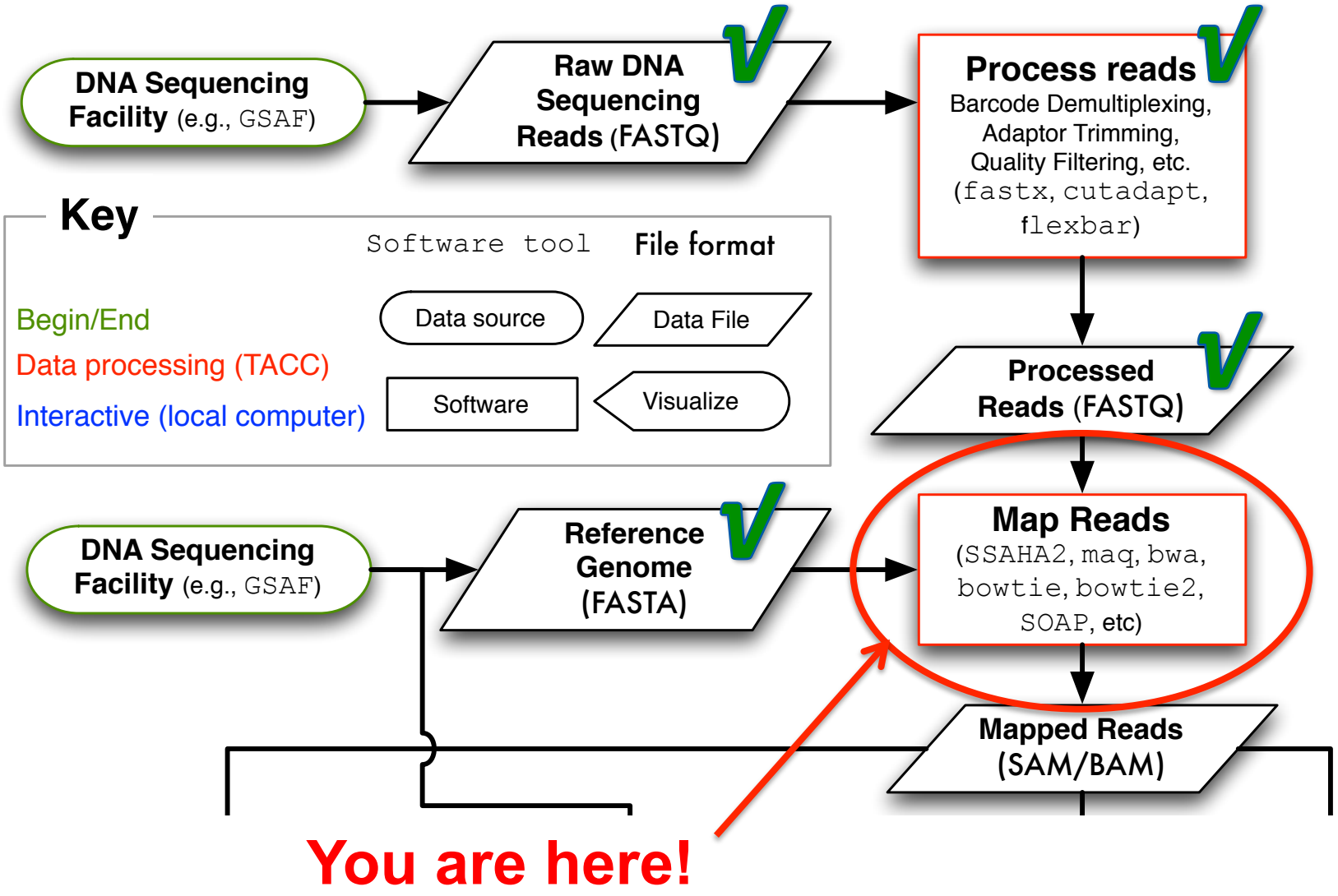
FASTA Format

```
>gi|254160123|ref|NC_012967.1| Escherichia coli B str. REL606
agcttttcattctgactgcaacgggcaatatgtctctgtgtggattaaaaaaagagtgtc
tgatagcagcttctgaactggttacctgccgtgagtaaattaaaatttattgacttagg
tcactaaatactttaaccaatataggcatagcgcacagacagataaaaattacagagtac
acaacatccatgaaacgcattagcaccaccattaccaccaccatcaccattaccacaggt
aacggtgcgggctgacgcgtacaggaaacacagaaaaaagcccgcacctgacagtgcggg
ctttttttcgaccaaaggtaacgaggtaacaaccatgcgagtgttgaagttcggcggta
....
```

Using complex reference sequence names is a common problem during analysis. Might rename:

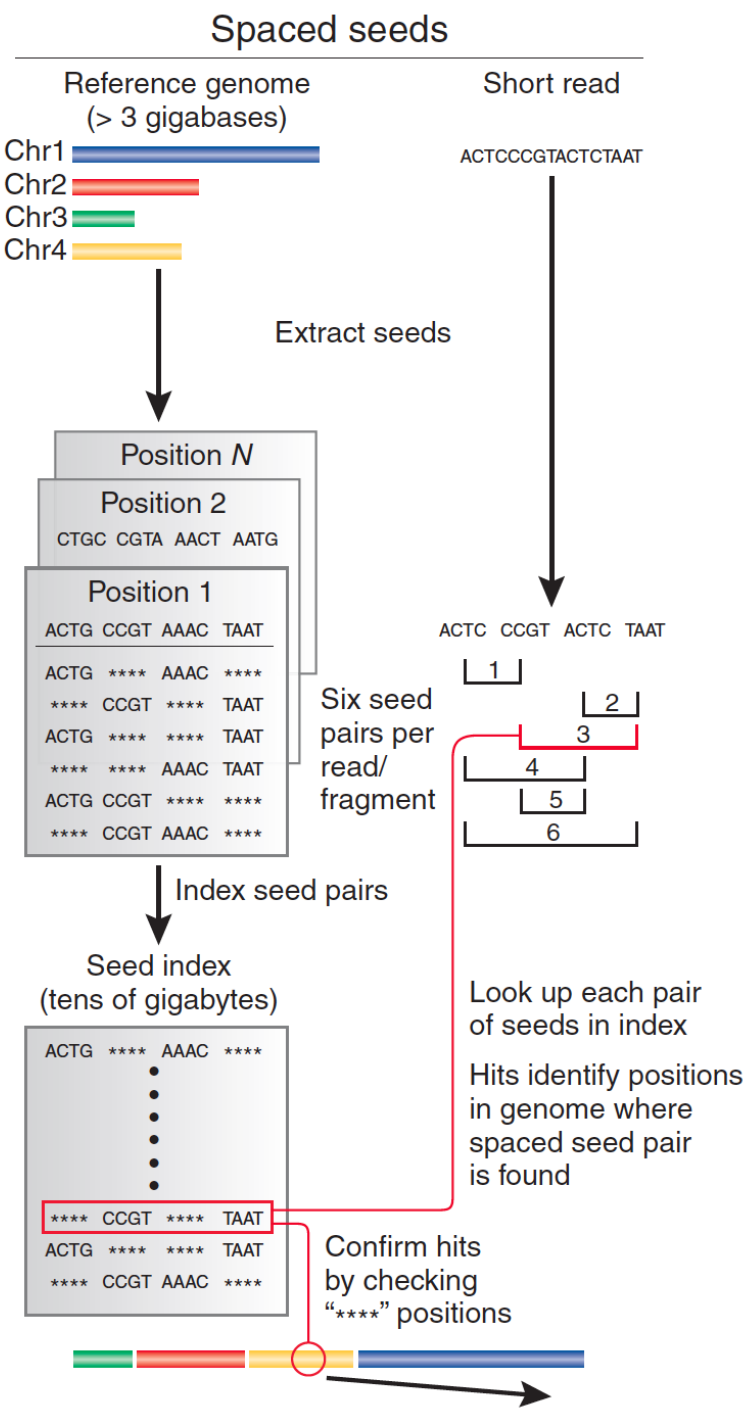
```
>REL606
agcttttcattctgactgcaacgggcaatatgtctctgtgtggattaaaaaaagagtgtc
tgatagcagcttctgaactggttacctgccgtgagtaaattaaaatttattgacttagg

....
```

Step: Map Reads



Mappers/Aligners

- Algorithms
  - Spaced-seed indexing
  - Burrows-Wheeler transform (BWT)
- Differences
  - Input data (read length, colorspace aware/useful)
  - Speed and scaleability (multithreading, GPUs)
  - Memory requirements (RAM, fat nodes)
  - Sensitivity: esp. indels (gaps)
  - Ease of installation and use. Development phase.
  - Uses base qualities? Outputs mapping scores?
  - Handles multiple matches, paired end matches
  - Configurability and transparency of options



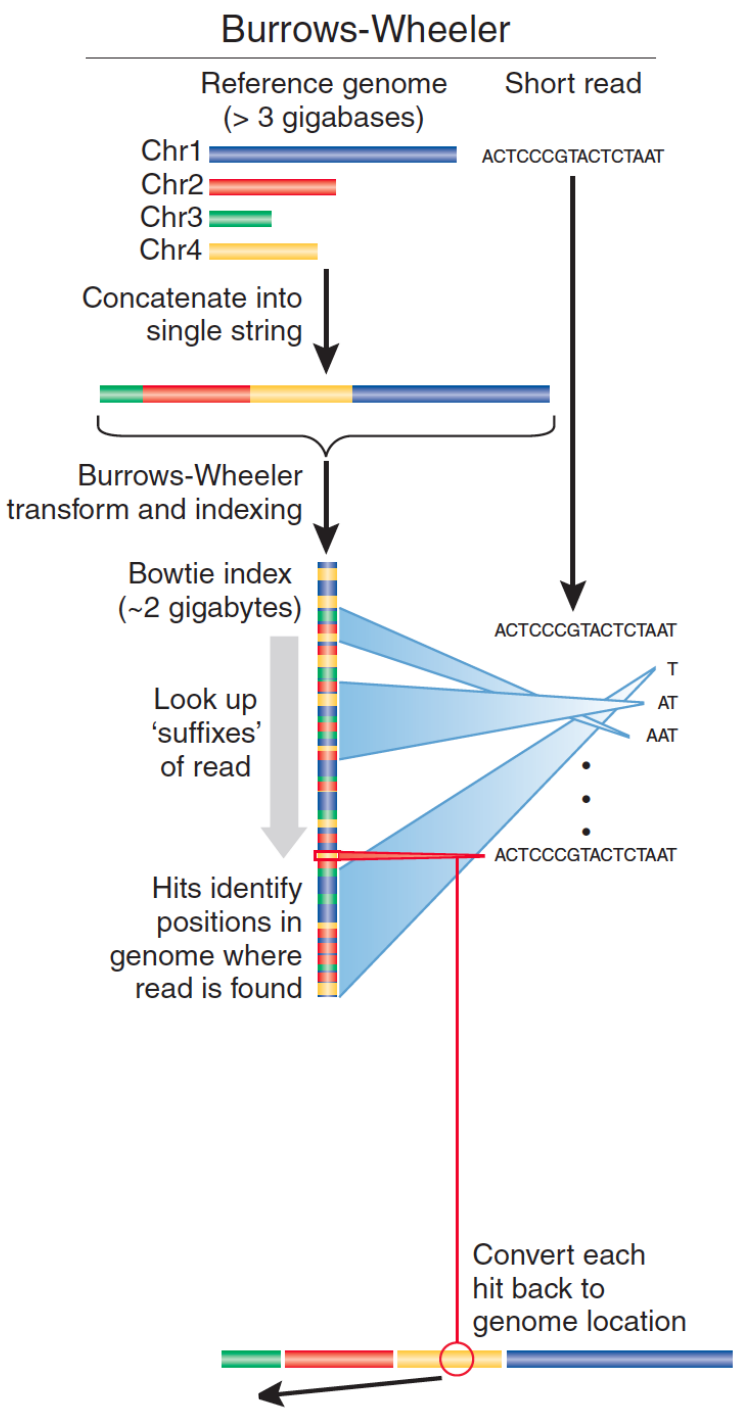
Hash table enables lookup of exact matches.

Subsequence	Reference Positions
ATAGCTAATCCAAA	2341, 2617264
ATAGCTAATCCAAT	
ATAGCTAATCCAAC	134, 13311, 732661,
ATAGCTATCCAAAG	
ATAGCTAATCCATA	
ATAGCTAATCCATT	3452
ATAGCTAATCCATC	
ATAGCTATCCAATG	234456673

Table is sorted and complete so you can jump immediately to matches. (But this can take a lot of memory.)

May include N bases, skip positions, etc.

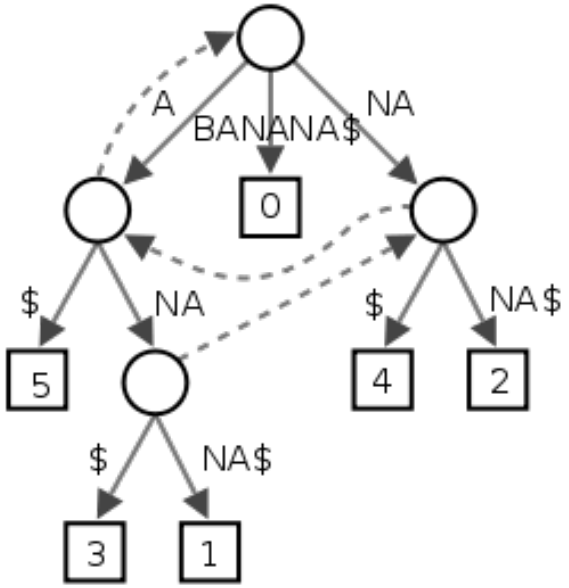
Trapnell, C. & Salzberg, S. L. How to map billions of short reads onto genomes. *Nature Biotech.* **27**, 455–457 (2009).



Burrows-Wheeler transform compresses sequence.

Input	SIX.MIXED.PIXIES.SIFT.SIXTY.PIXIE.DUST.BOXES
Output	TEXYDST.E.IXIXIXXSSMPPS.B..E.S.EUSFXDIIOIIIT

Suffix tree enables fast lookup of subsequences.



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

Exact matches at all positions below a node.

Trapnell, C. & Salzberg, S. L. How to map billions of short reads onto genomes. *Nature Biotech.* **27**, 455–457 (2009).

From Mapped Read to Alignment

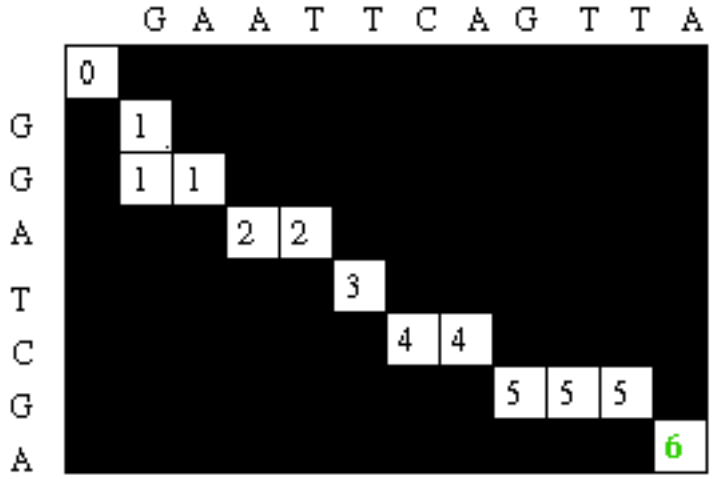
- **Mapping** determines a "seed" position where the read shares a subsequence with the reference. But, is this the best match?
- **Alignment** starts with the seed and determines how the read is best aligned on a base-by-base basis around the seed.

Seed→Alignment score→Mapping quality

Alignment

- Dynamic programming algorithm (Smith-Waterman | Needleman-Wunsch)

		G	A	A	T	T	C	A	G	T	T	A
	0	0	0	0	0	0	0	0	0	0	0	0
G	0	1	1	1	1	1	1	1	1	1	1	1
G	0	1	1	1	1	1	1	1	2	2	2	2
A	0	1	1	2	2	2	2	2	2	2	2	3
T	0	1	2	2	3	3	3	3	3	3	3	3
C	0	1	2	2	3	3	4	4	4	4	4	4
G	0	1	2	2	3	3	4	4	5	5	5	5
A	0	1	2	3	3	3	4	5	5	5	5	6




```
G _ A A T T C A G T T A
|   |   | |   |     |
G G _ A _ T C _ G _ _ A
```

- Various scoring schemes possible... (next slide)

Alignment Score

- Dynamic programming algorithm (Smith-Waterman | Needleman-Wunsch)
- **Alignment score** = Σ
  - match reward
  - base mismatch penalty
  - gap open penalty
  - gap extension penalty
  - rewards and penalties may be adjusted for quality scores of bases involved

Reference sequence

```
ATTTGCGATCGGATGAAGACGAA
||||||||||||||
ATTTGCGATCGGATGTTGACTTT

ATTTGCGATCGGATGAAGACG..AA
||||||||||||||XX|||XXX||
ATTTGCGATCGGATGTTGACTTTAA
```

- **Important: Local** versus **global** modes

Mapping Quality

Mapping quality – what is the probability that the read is correctly mapped to this location in the reference genome?

High **alignment** score ≠ high **mapping** quality.

Phred score: $P(\text{mismapped}) = 10^{-MQ/10}$

Types of DNA fragment libraries

single-end

independent reads

paired-end

two inwardly oriented reads separated by ~200 nt

mate-paired

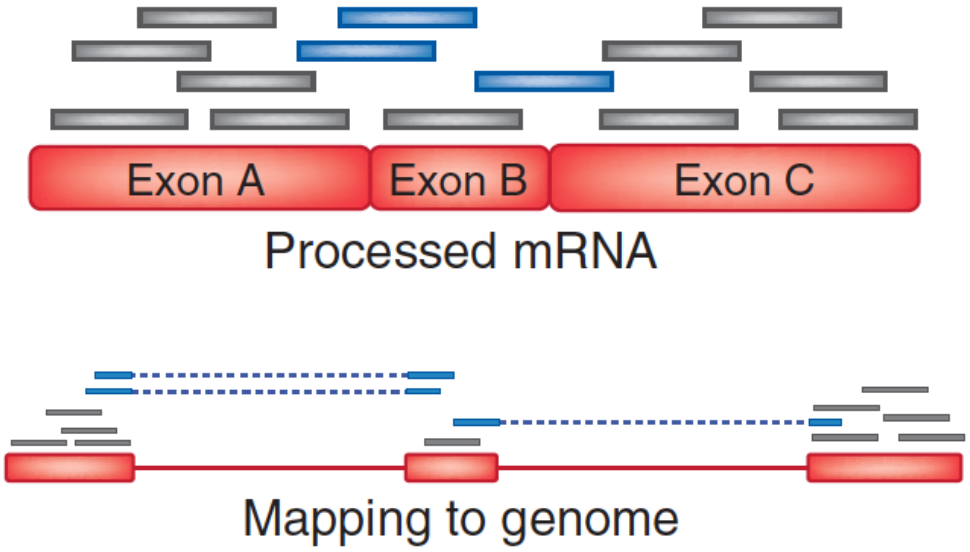
two outwardly oriented reads separated by ~3000 nt

Paired-end mapping (PEM)

- There is an expected insert size distribution based on the DNA fragment library.
- Mapping one read anchors the paired read to a specific location, even if the second read alone maps multiple places equally.
- Only one read in a pair might be mappable. (**singleton/orphan**)
- Both reads can map with an unexpected insert size or orientation (**discordant pair**)

Split-read alignment (SRA)

- Useful for predicting structural variants (or splice variants in RNA-seq, as pictured).
- Not many mappers do this directly, usually happens in a post-processing pipeline step.

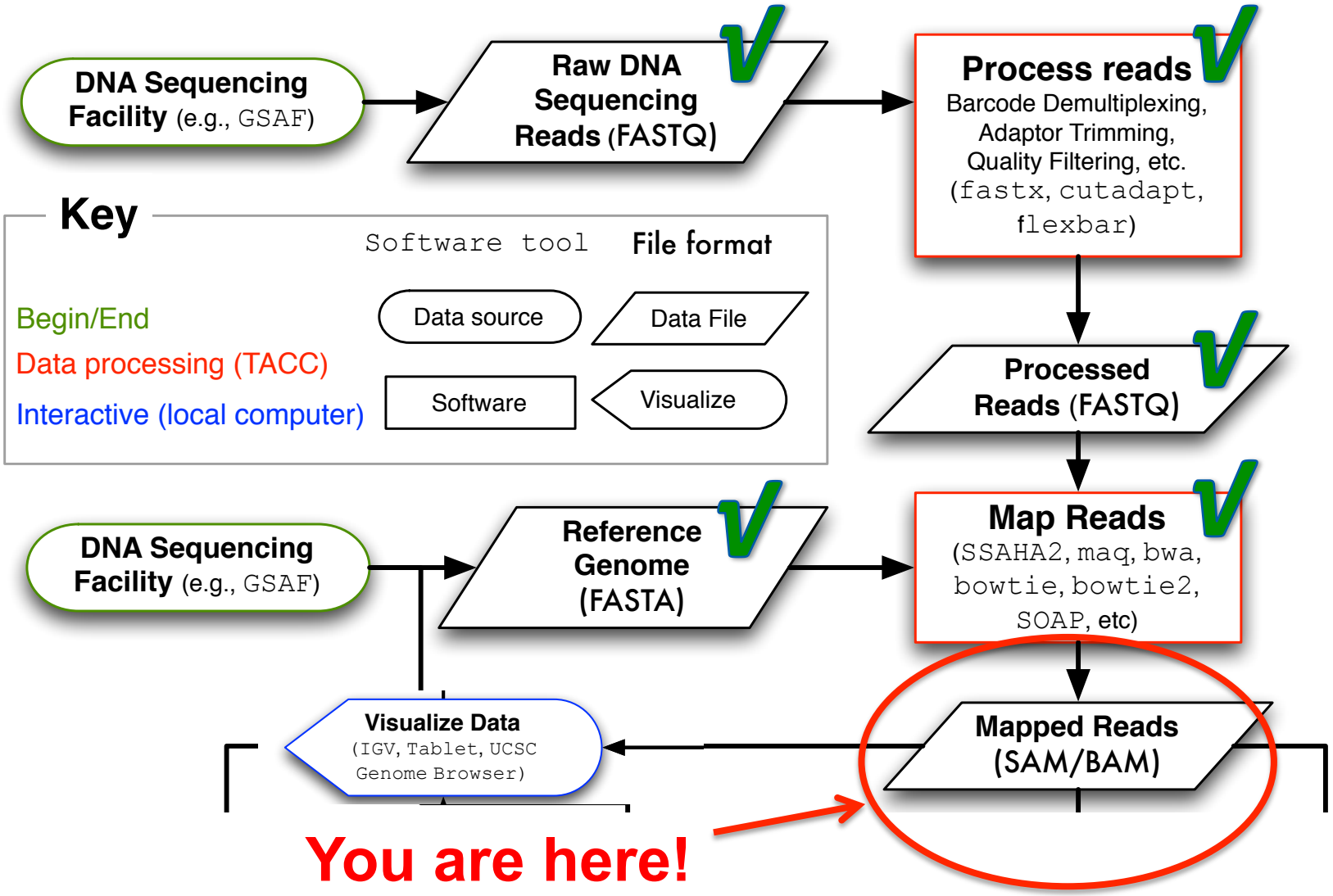


Trapnell, C. & Salzberg, S. L. How to map billions of short reads onto genomes. *Nature Biotech.* **27**, 455–457 (2009).

Read alignment software

- You may see these aligners commonly used in published workflows:
  - SSAHA2, maq, SHRiMP, BWA, bowtie, SOAP2
  - Available as TACC modules or easy to install
- Generally, we recommend using **bowtie2**
  - Best combination of speed and flexibility
  - Well documented and still actively developed
- Running nearly all aligners follows a very similar procedure: 1) build index 2) align

Output: Mapped Reads



SAM File Format

- Community flat file/database format that describes how reads align to a reference (and can also include unmapped reads).
- Can tag reads as being from different instrument runs / technologies / samples.
- Going forward you use the reference file and the SAM/BAM, no longer need the FASTQ.
- Tab delimited with fixed columns followed by arbitrary user-extendable key:data values.

SAM File Format

Two example SAM lines:

```
SRR030257.264529      99  NC_012967  1521  29  34M2S  =  1564  79
    CTGGCCATTATCTCGGTGGTAGGACATGGCATGCCC
    AAAAAA;AA;AAAAAA??A%.;?&'3735',()0*,
    XT:A:M  NM:i:3  SM:i:29  AM:i:29  XM:i:3  XO:i:0  XG:i:0  MD:Z:23T0G4T4

SRR030257.2669090     147 NC_012967  1521  60  36M    =  1458  -99
    CTGGCCATTATCTCGGTGGTAGGTGATGGTATGCGC
    <<9:<<AAAAAAAAAAAAAAAAAAAAAAAAAAAA
    XT:A:U  NM:i:0  SM:i:37  AM:i:37  X0:i:1  X1:i:0  XM:i:0  XO:i:0  XG:i:0  MD:Z:36
```

SAM File Format

SAM fixed fields:

http://samtools.sourceforge.net/

Col	Field	Type	Regexp/Range	Brief description
1	QNAME	String	`[!-?A-~]{1,255}`	Query template NAME
2	FLAG	Int	$[0,2^{16}-1]$	bitwise FLAG
3	RNAME	String	`\*|[!-()+-<>-~][!-~]*`	Reference sequence NAME
4	POS	Int	$[0,2^{29}-1]$	1-based leftmost mapping POSition
5	MAPQ	Int	$[0,2^{8}-1]$	MAPping Quality
6	CIGAR	String	`\*|([0-9]+[MIDNSHPX=])+`	CIGAR string
7	RNEXT	String	`\*|=|[!-()+-<>-~][!-~]*`	Ref. name of the mate/next segment
8	PNEXT	Int	$[0,2^{29}-1]$	Position of the mate/next segment
9	TLEN	Int	$[-2^{29}+1,2^{29}-1]$	observed Template LENgth
10	SEQ	String	`\*|[A-Za-z=.]+`	segment SEQuence
11	QUAL	String	`[!-~]+`	ASCII of Phred-scaled base QUALity+33

SRR030257.264529 99 NC_012967 1521 29 34M2S = 1564
79 CTGGCCATTATCTCGGTGGTAGGACATGGCATGCCC
AAAAAA;AA;AAAAAA??A%.;?&'3735',()0*,
XT:A:M NM:i:3 SM:i:29 AM:i:29 XM:i:3 XO:i:0 XG:i:0 MD:Z:23T0G4T4

Sometimes a CIGAR is a just a way of describing how a read is aligned...

```
Ref   CTGGCCATTATCTC--GGTGGTAGGACATGGCATGCCC
Read        aaATGTCGCGGTG.TAGGAggatcc
```

2S5M2I4M1D4M6S

	Op	BAM	Description
	M	0	alignment match (can be a sequence match or mismatch)
	I	1	insertion to the reference
	D	2	deletion from the reference
*	N	3	skipped region from the reference
	S	4	soft clipping (clipped sequences present in SEQ)
*	H	5	hard clipping (clipped sequences NOT present in SEQ)
*	P	6	padding (silent deletion from padded reference)
*	=	7	sequence match
*	X	8	sequence mismatch

Note: indels relative to reference

*Rarer / newer

CIGAR = "Concise Idiosyncratic Gapped Alignment Report"

BAM format

- "Human readable" text (SAM) and GZIP compressed binary (BAM) versions.
- BAM files can be **sorted** and **indexed**, so that all reads mapped to a given window of the reference genome can be retrieved rapidly (for display or processing).
- **SAMtools package** can calculate stats and perform basic genome variant calling. (available via a module on TACC)

Step: Evaluate Mapped Reads