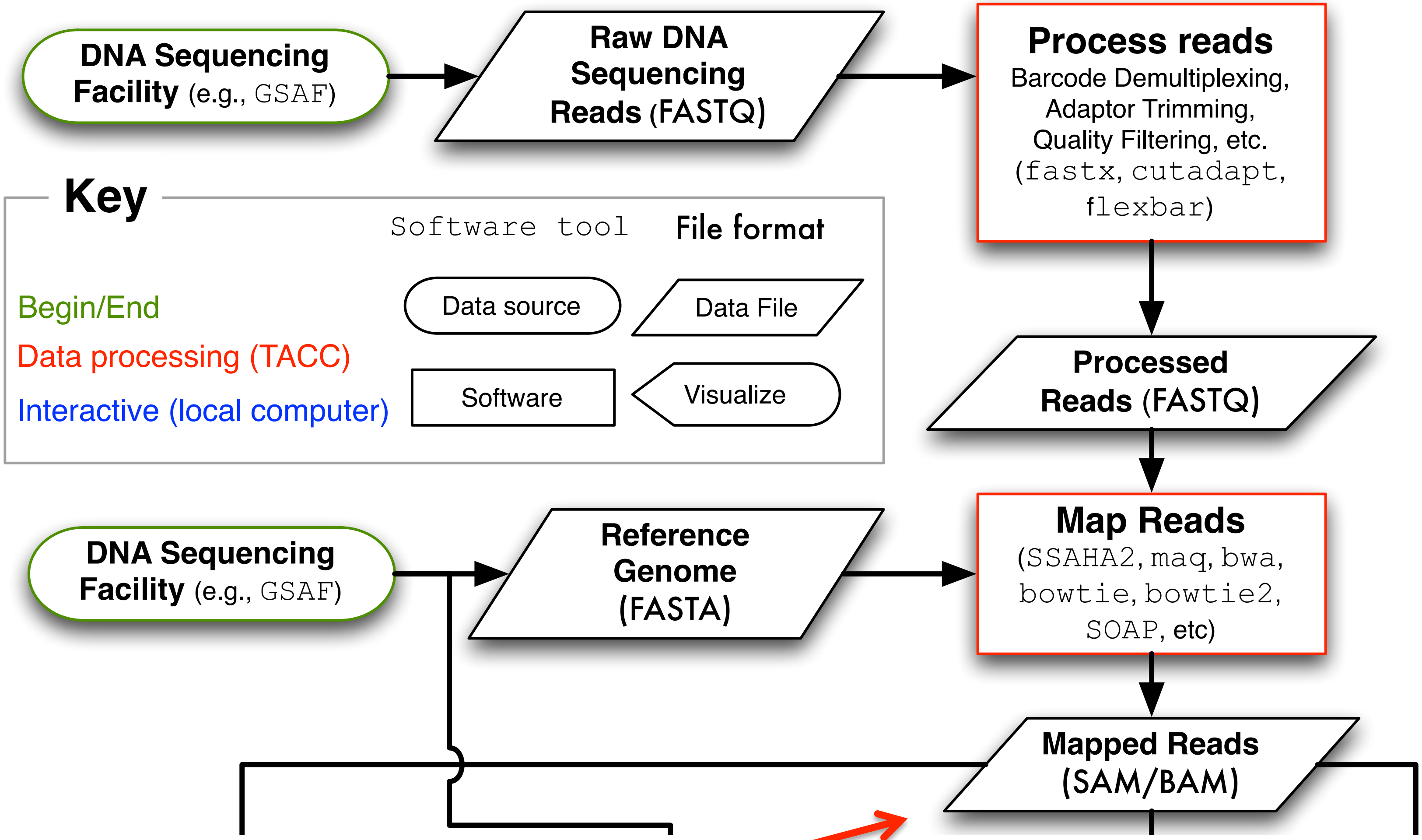
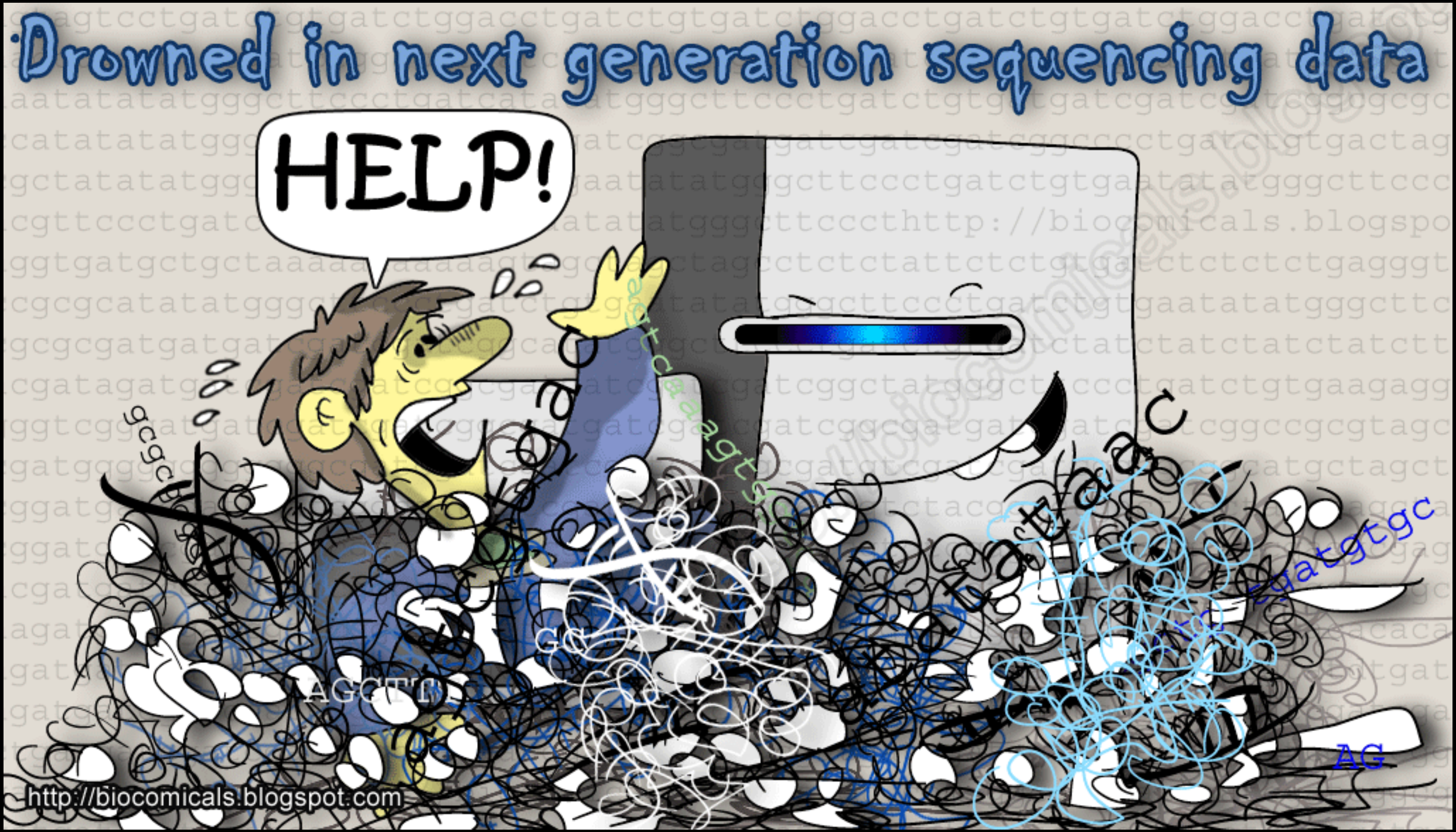
Introduction to Read Mapping


DNA Sequencing Facility (e.g., GSAF)
Raw DNA Sequencing Reads (FASTQ)
Process reads
Barcode Demultiplexing, Adaptor Trimming, Quality Filtering, etc. (fastx, cutadapt, flexbar)
Key
Software tool
File format
Begin/End
Data processing (TACC)
Interactive (local computer)
Data source
Data File
Software
Visualize
Processed Reads (FASTQ)
DNA Sequencing Facility (e.g., GSAF)
Reference Genome (FASTA)
Map Reads
(SSAHA2, maq, bwa, bowtie, bowtie2, SOAP, etc)
Mapped Reads (SAM/BAM)

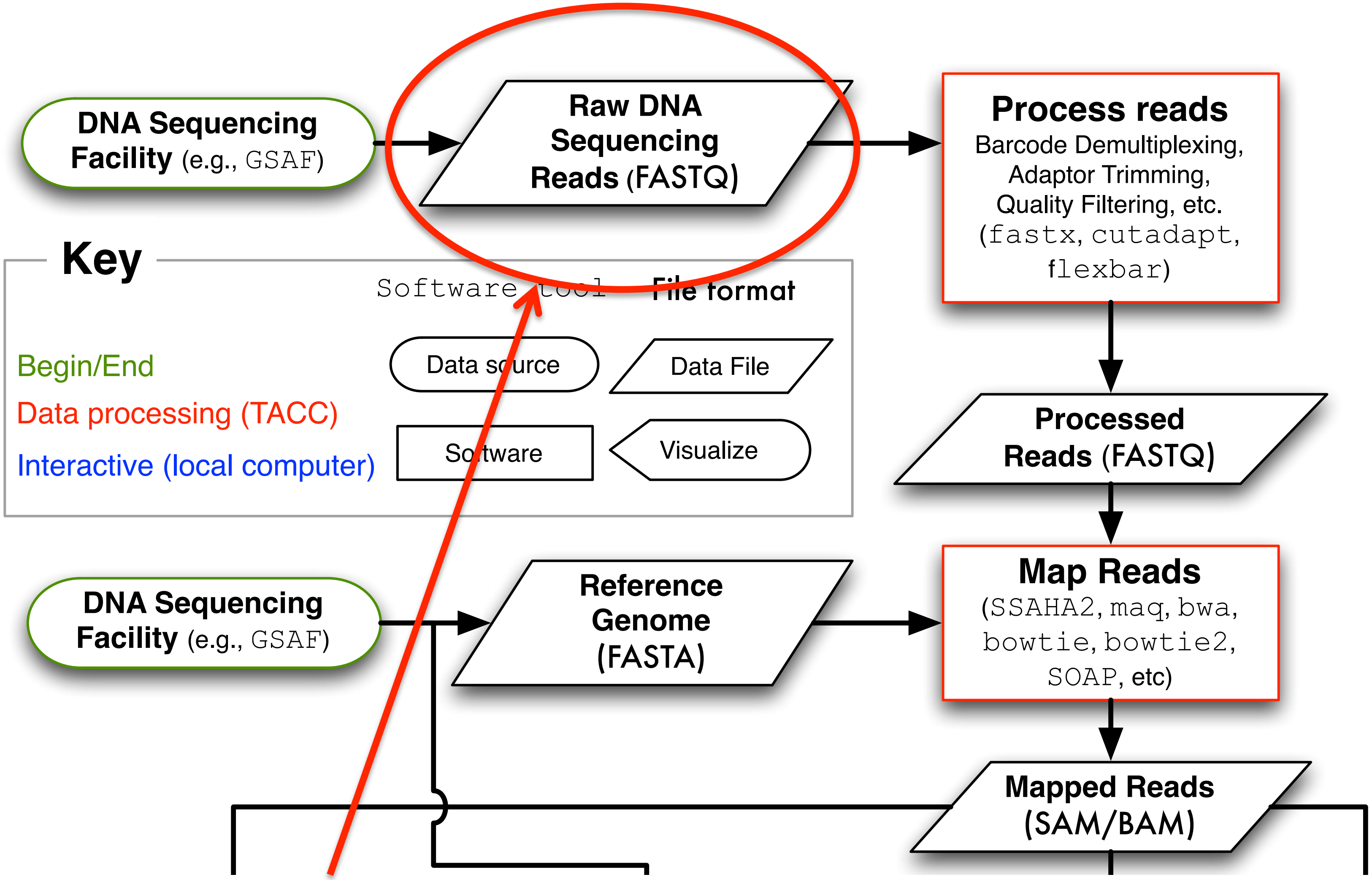
Steps to get you here!


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

Basic steps of mapping reads

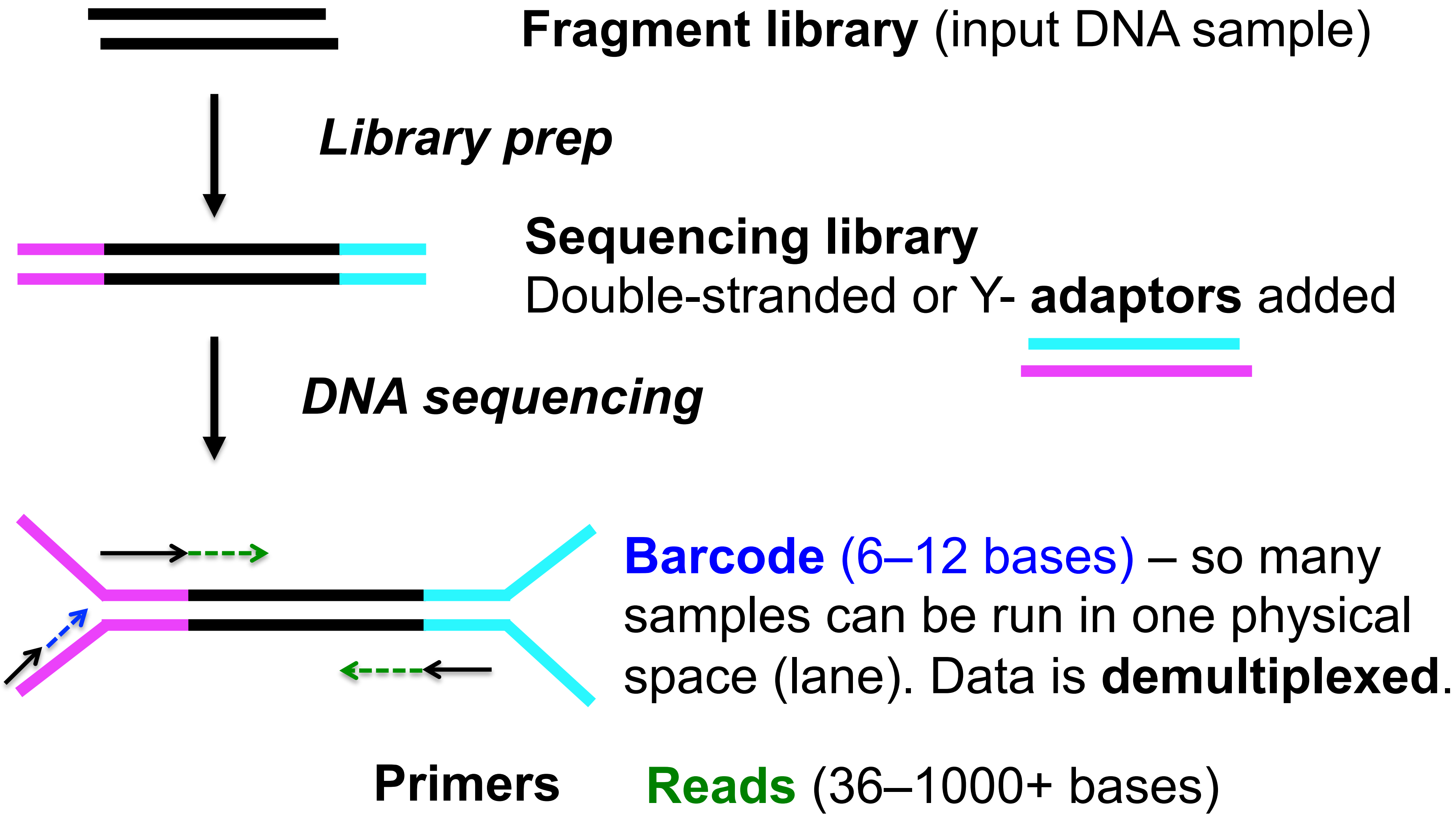
1. Read file quality control and processing
2. Build reference sequence index
3. Map DNA sequencing reads
   - Exact tool/approach depends on sequencing technology and DNA fragment library type
4. Convert result to SAM/BAM database
5. Application specific analysis...
   - These steps are common to any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

Input: Raw DNA Sequencing Reads


DNA Sequencing Facility (e.g., GSAF)
Raw DNA Sequencing Reads (FASTQ)
Process reads
Barcode Demultiplexing, Adaptor Trimming, Quality Filtering, etc. (fastx, cutadapt, flexbar)
Key
Software tool
File format
Begin/End
Data processing (TACC)
Interactive (local computer)
Data source
Data File
Software
Visualize
Processed Reads (FASTQ)
DNA Sequencing Facility (e.g., GSAF)
Reference Genome (FASTA)
Map Reads
(SSAHA2, maq, bwa, bowtie, bowtie2, SOAP, etc)
Mapped Reads (SAM/BAM)

You are here!

Read terminology



Fragment library (input DNA sample)

Library prep

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

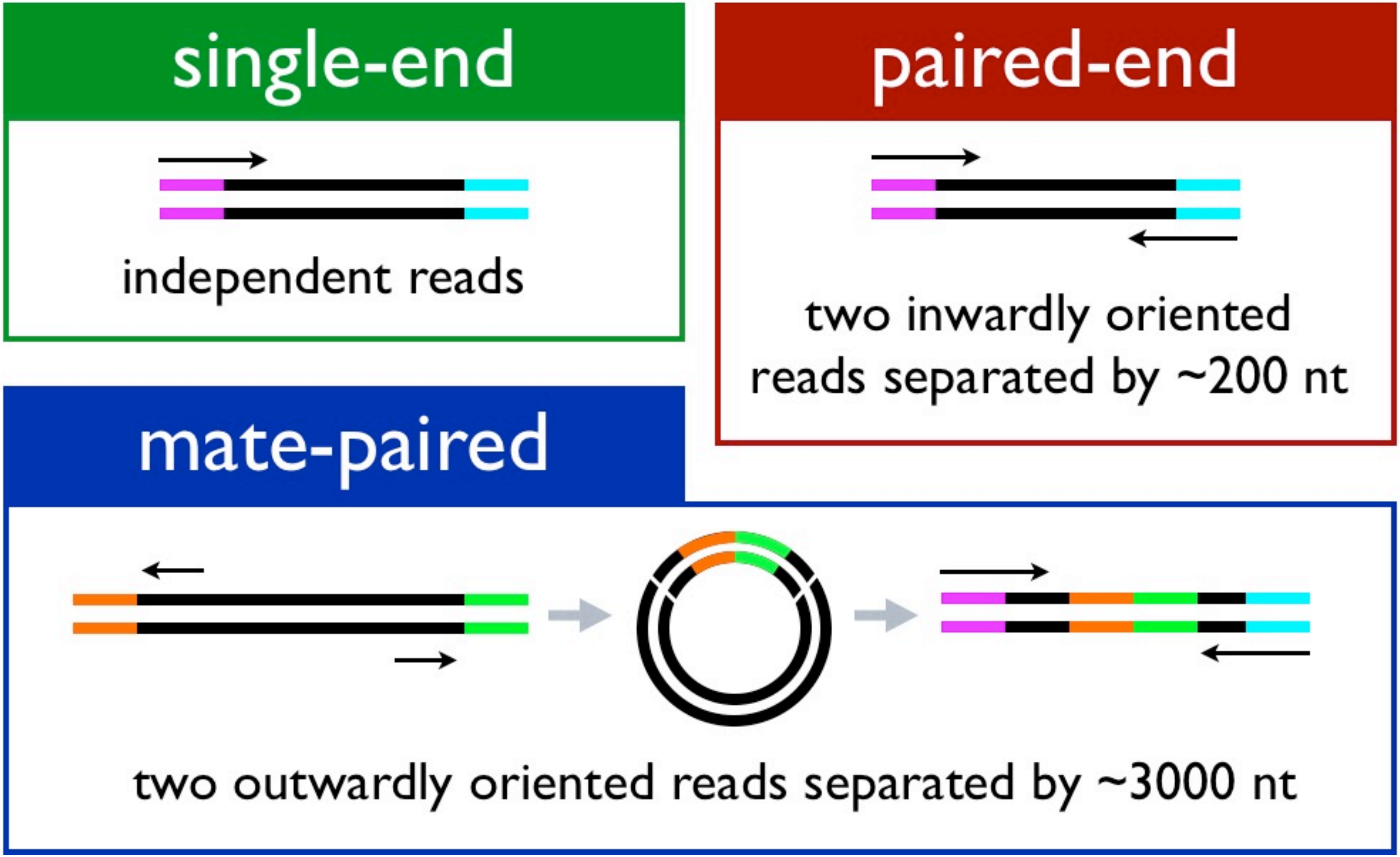
DNA sequencing

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

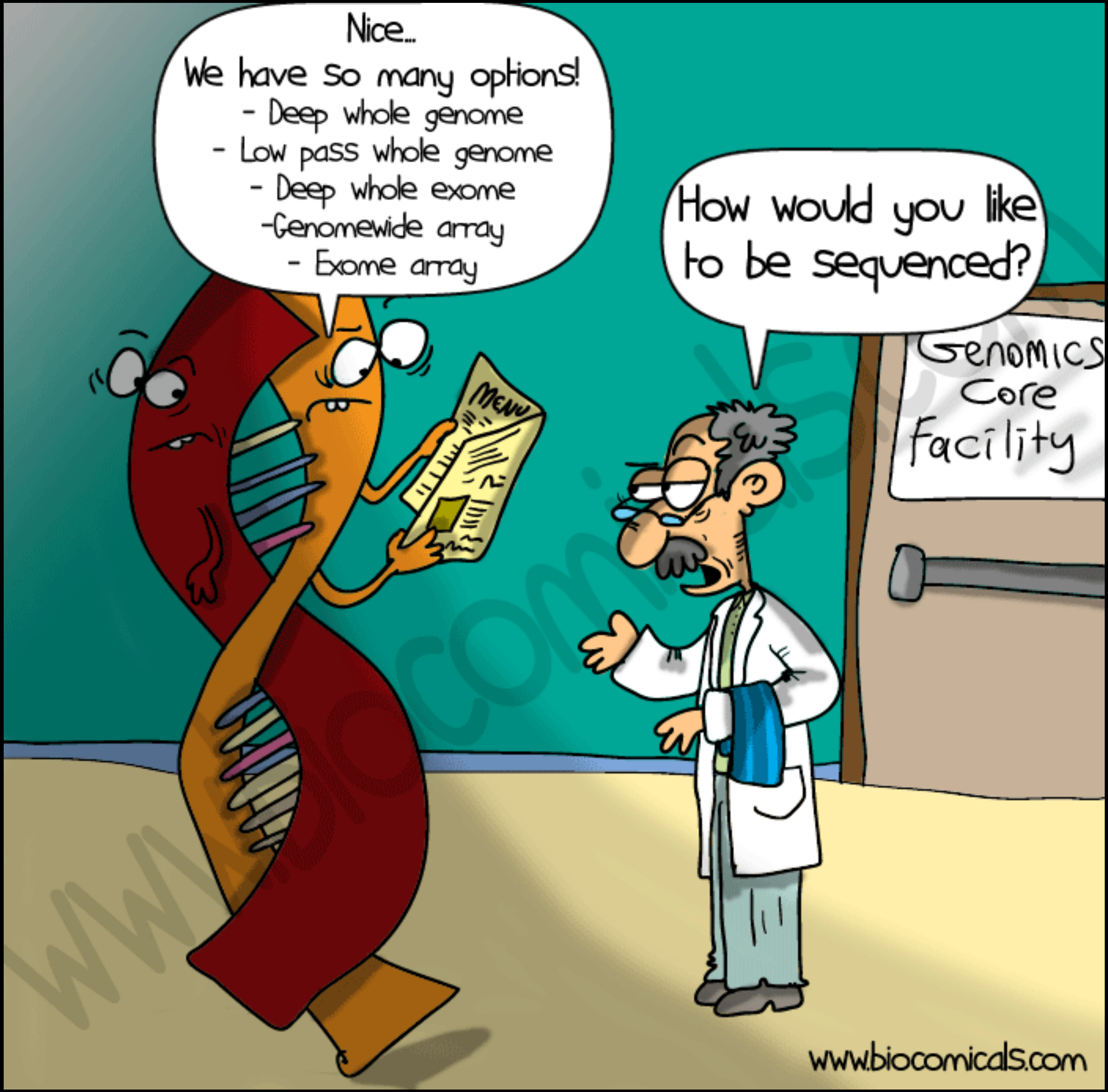
independent reads

paired-end

two inwardly oriented reads separated by ~200 nt

mate-paired

two outwardly oriented reads separated by ~3000 nt



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
```

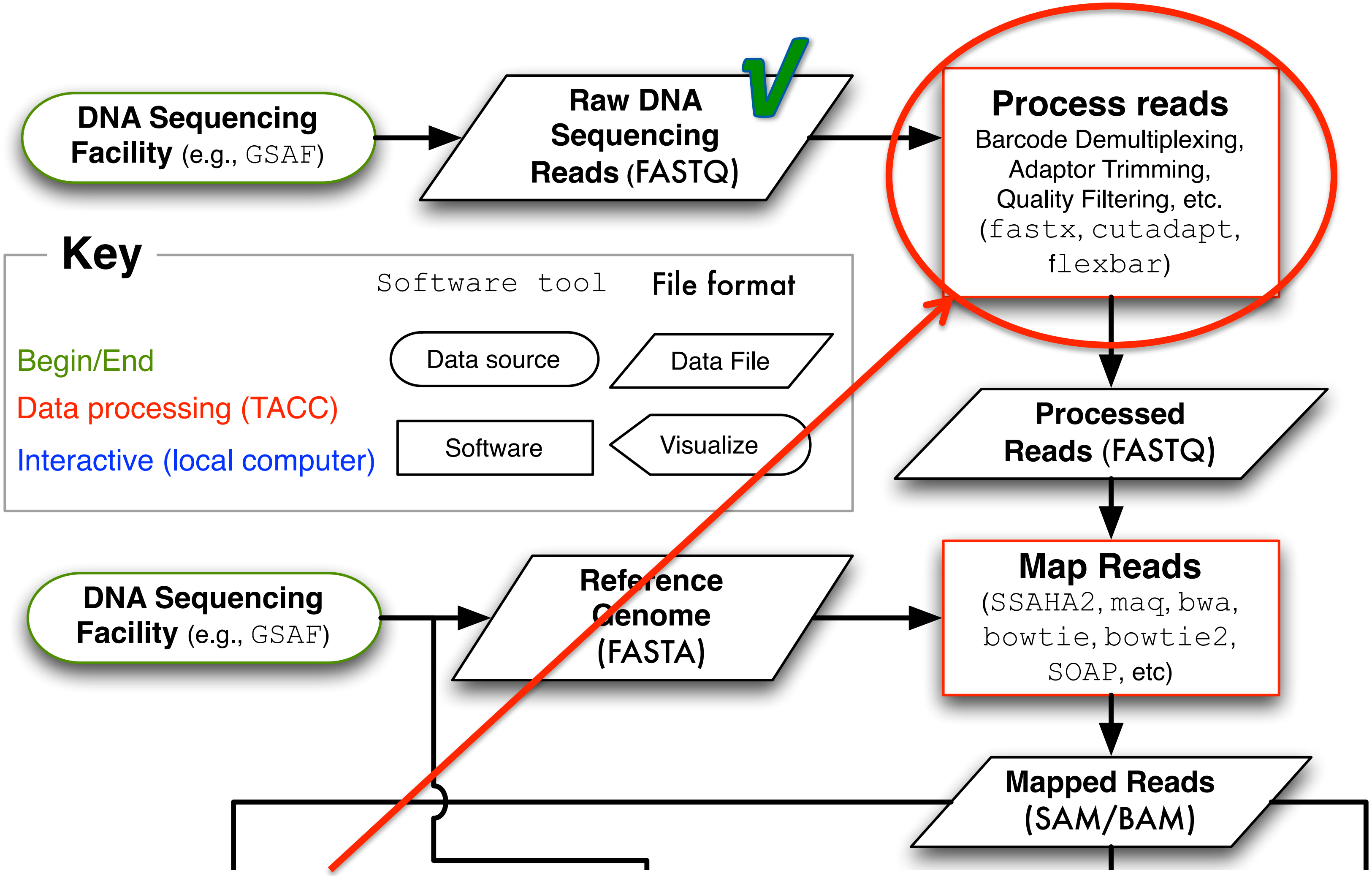
$$\text{Probability of Error} = 10^{-Q/10}$$

(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

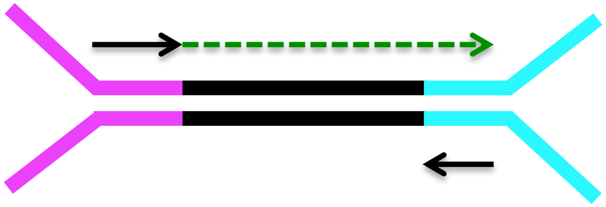

DNA Sequencing Facility (e.g., GSAF)
Raw DNA Sequencing Reads (FASTQ)
√
Process reads
Barcode Demultiplexing, Adaptor Trimming, Quality Filtering, etc.
(fastx, cutadapt, flexbar)
Key
Software tool
File format
Begin/End
Data processing (TACC)
Interactive (local computer)
Data source
Data File
Software
Visualize
Processed Reads (FASTQ)
DNA Sequencing Facility (e.g., GSAF)
Reference Genome (FASTA)
Map Reads
(SSAHA2, maq, bwa, bowtie, bowtie2, SOAP, etc)
Mapped Reads (SAM/BAM)

You are here!

Read sequence quality control

Garbage in = garbage out

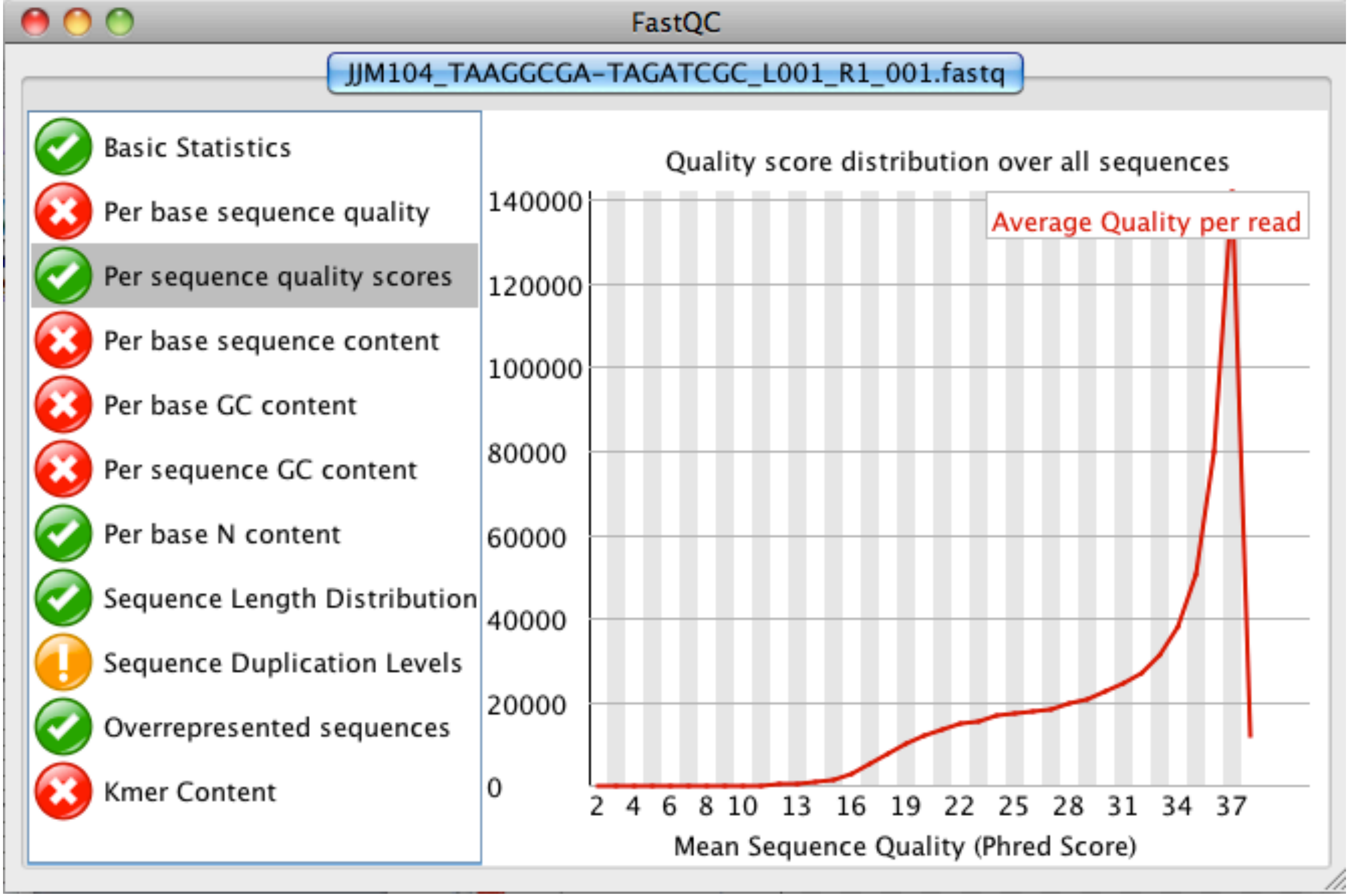
- Contaminated with other samples?
- Sample barcodes removed?
- Adaptor/bar codes trimmed?
  - – Esp. important for MiSeq data
- Trim ends of reads with poor quality?
  - – Less data but higher quality data
- Know your data
  - – Paired reads? Relative orientations?
  - – Technology specific concerns?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



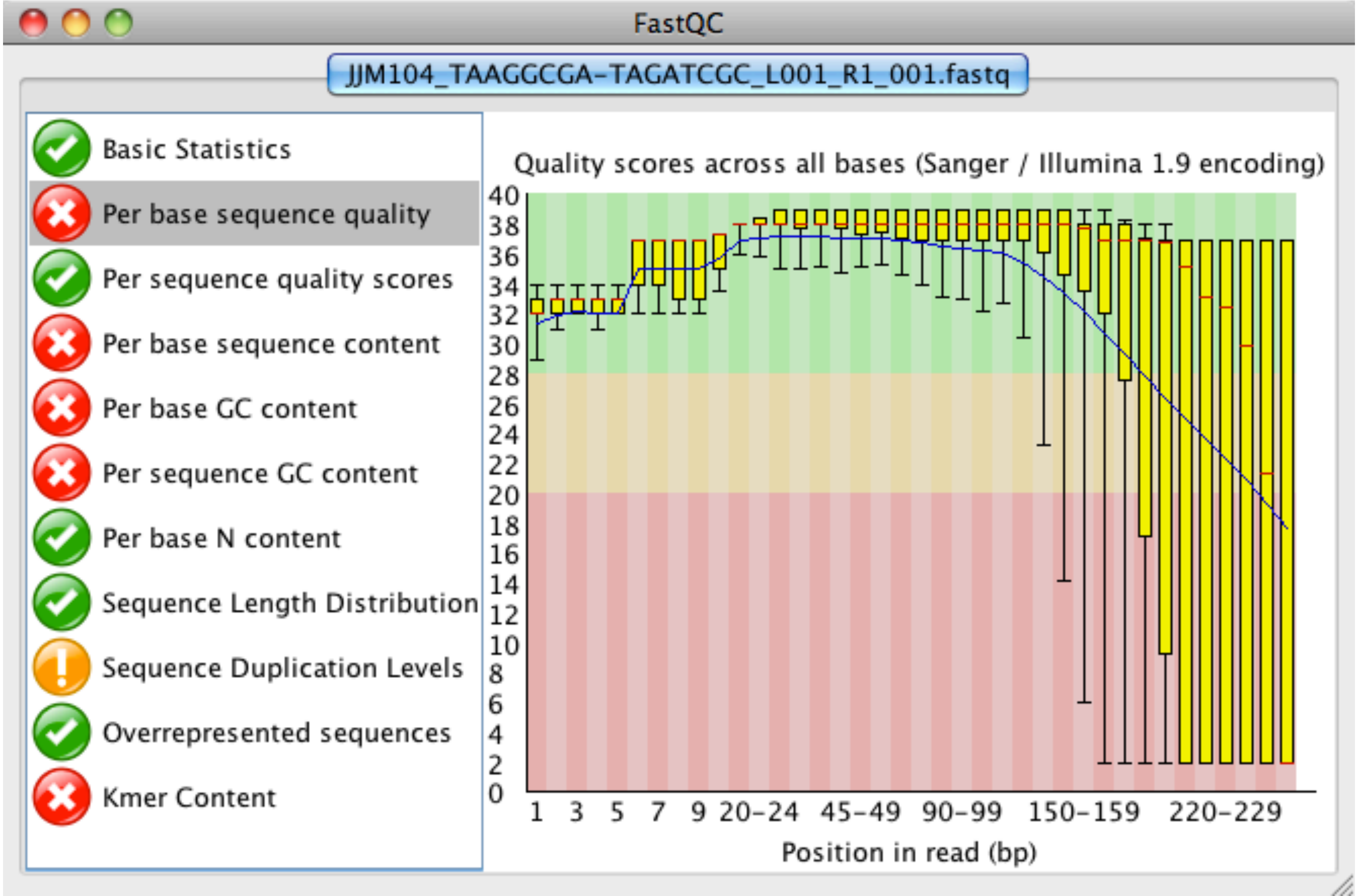
Read quality control software

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

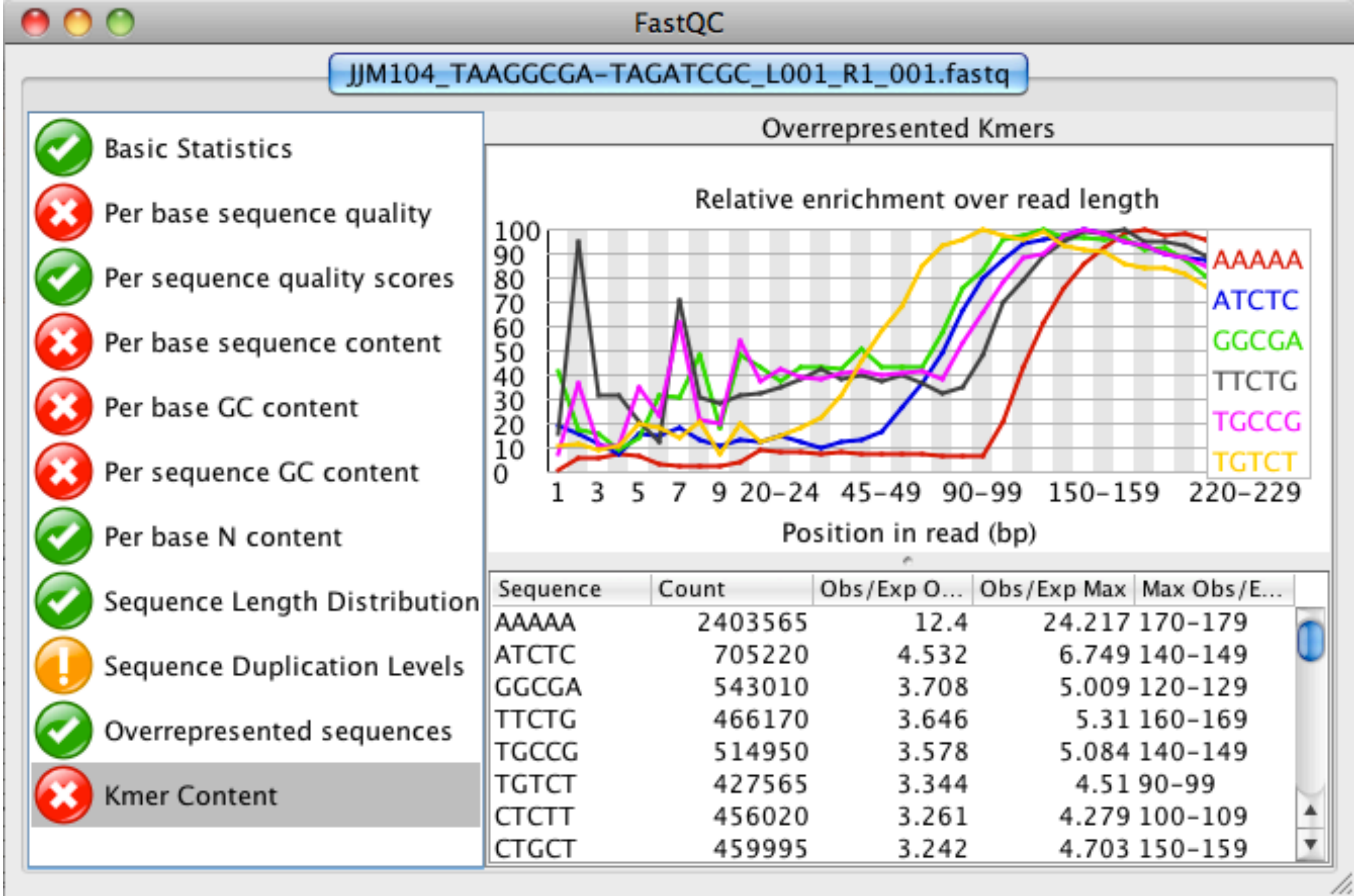

FastQC
JJM104_TAAGGCGA-TAGATCGC_L001_R1_001.fastq
✓ Basic Statistics
✗ Per base sequence quality
✓ Per sequence quality scores
✗ Per base sequence content
✗ Per base GC content
✗ Per sequence GC content
✓ Per base N content
✓ Sequence Length Distribution
! Sequence Duplication Levels
✓ Overrepresented sequences
✗ Kmer Content
Quality score distribution over all sequences
Average Quality per read
140000
120000
100000
80000
60000
40000
20000
0
2 4 6 8 10 13 16 19 22 25 28 31 34 37
Mean Sequence Quality (Phred Score)

Illumina data example


FastQC
JJM104_TAAGGCGA-TAGATCGC_L001_R1_001.fastq
Quality scores across all bases (Sanger / Illumina 1.9 encoding)
40
38
36
34
32
30
28
26
24
22
20
18
16
14
12
10
8
6
4
2
0
1
3
5
7
9
20-24
45-49
90-99
150-159
220-229
Position in read (bp)

- ✓ Basic Statistics
- ✗ Per base sequence quality
- ✓ Per sequence quality scores
- ✗ Per base sequence content
- ✗ Per base GC content
- ✗ Per sequence GC content
- ✓ Per base N content
- ✓ Sequence Length Distribution
- ! Sequence Duplication Levels
- ✓ Overrepresented sequences
- ✗ Kmer Content

Illumina data example



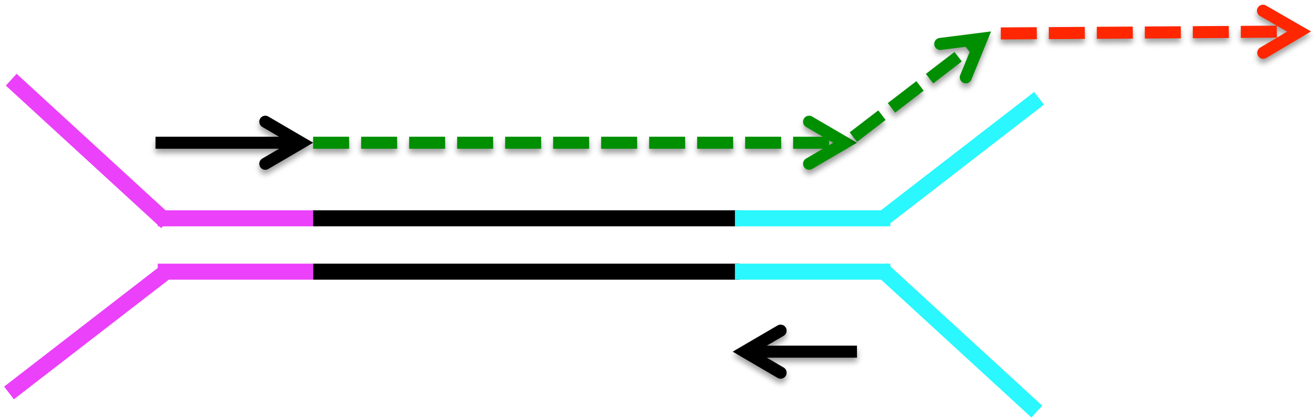
- ✓ Basic Statistics
- ✗ Per base sequence quality
- ✓ Per sequence quality scores
- ✗ Per base sequence content
- ✗ Per base GC content
- ✗ Per sequence GC content
- ✓ Per base N content
- ✓ Sequence Length Distribution
- ! Sequence Duplication Levels
- ✓ Overrepresented sequences
- ✗ Kmer Content

Overrepresented Kmers

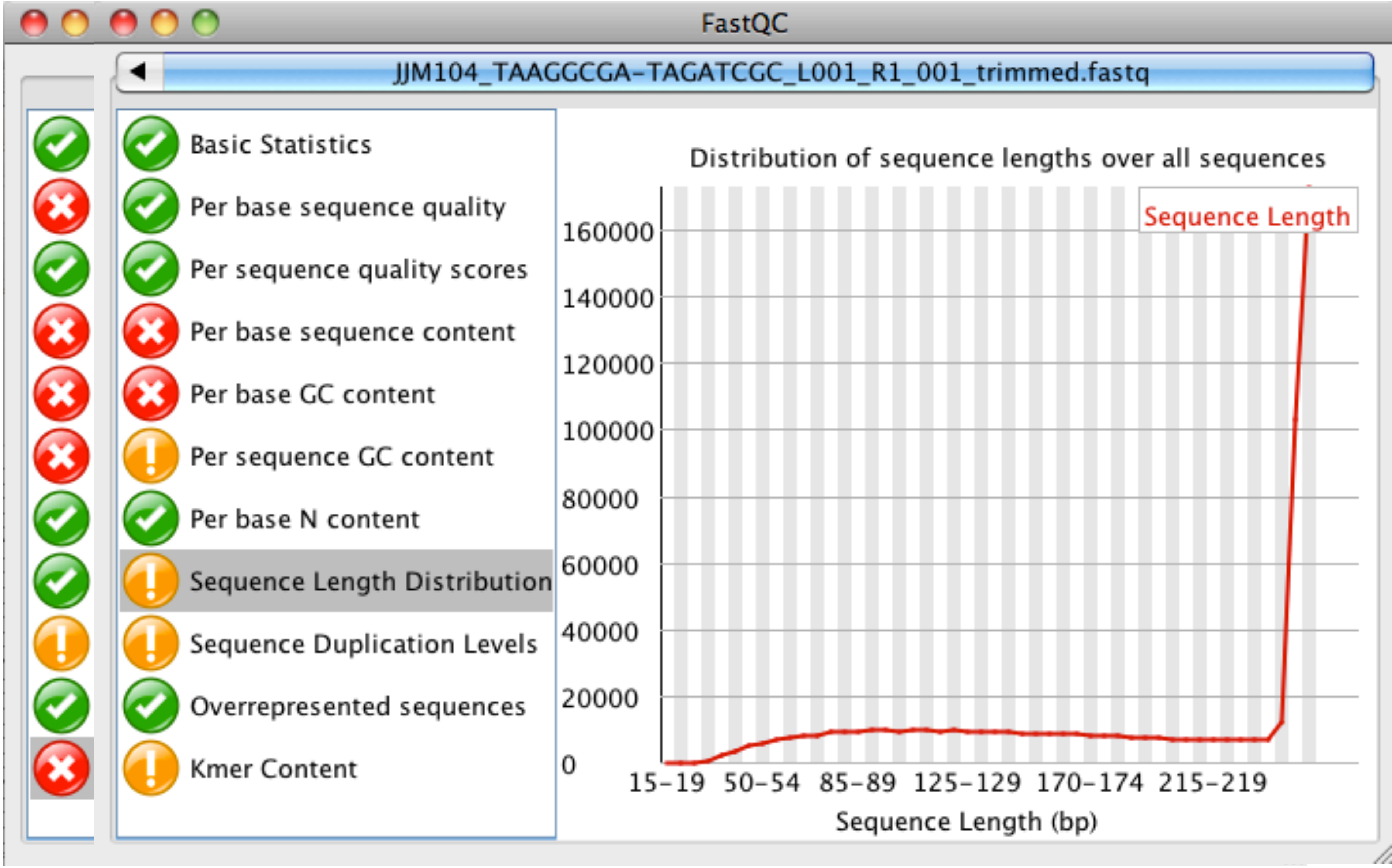
Sequence	Count	Obs/Exp O...	Obs/Exp Max	Max Obs/E...
AAAAA	2403565	12.4	24.217	170-179
ATCTC	705220	4.532	6.749	140-149
GGCGA	543010	3.708	5.009	120-129
TTCTG	466170	3.646	5.31	160-169
TGCCG	514950	3.578	5.084	140-149
TGTCT	427565	3.344	4.51	90-99
CTCTT	456020	3.261	4.279	100-109
CTGCT	459995	3.242	4.703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



Processed Illumina data example

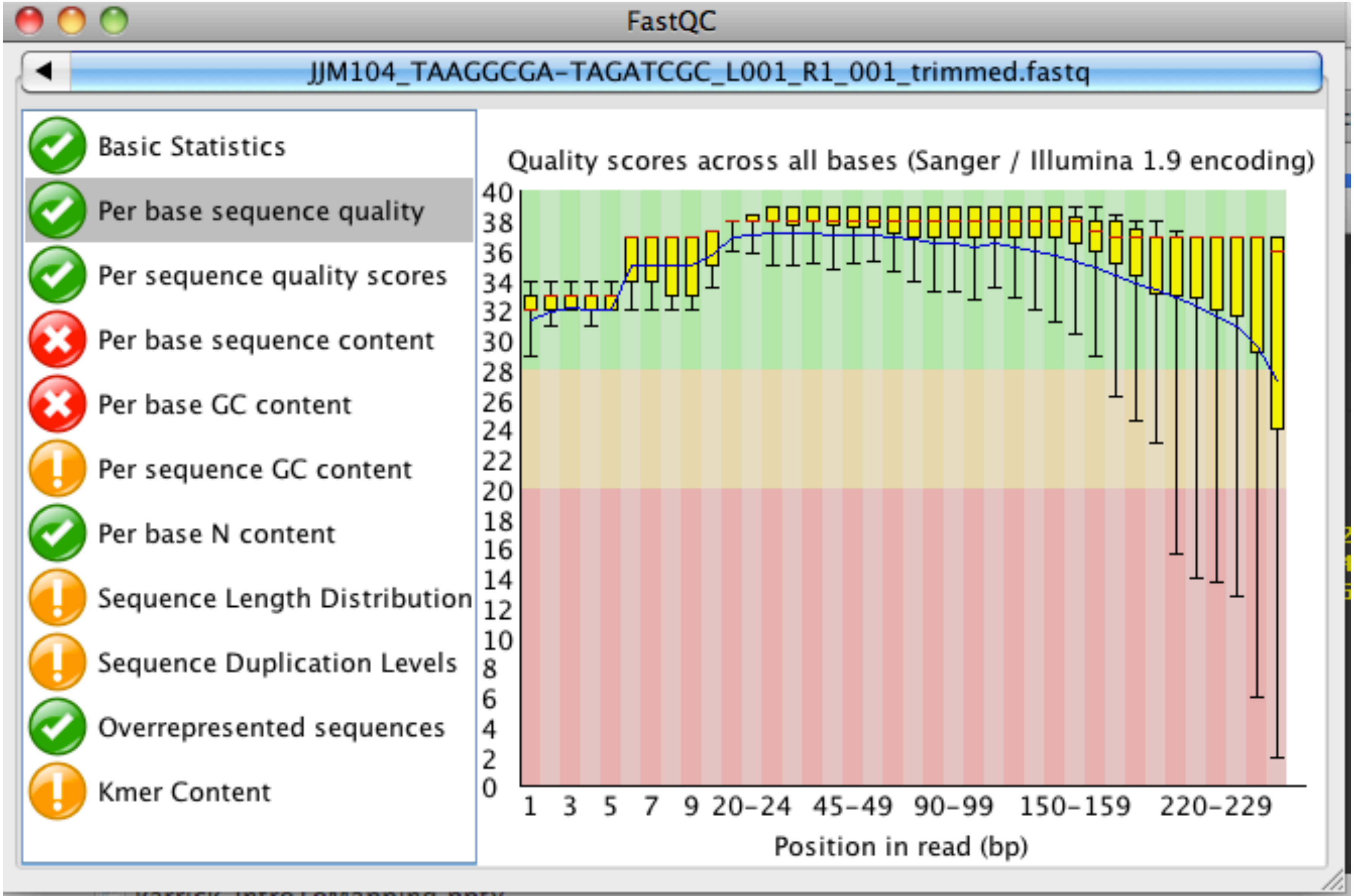


FastQC

JJM104_TAAGGCGA-TAGATCGC_L001_R1_001_trimmed.fastq

- ✓ Basic Statistics
- ✓ Per base sequence quality
- ✓ Per sequence quality scores
- ✗ Per base sequence content
- ✗ Per base GC content
- ! Per sequence GC content
- ✓ Per base N content
- ! Sequence Length Distribution
- ! Sequence Duplication Levels
- ✓ Overrepresented sequences
- ! Kmer Content

Processed Illumina data example



FastQC

JJM104_TAAGGCGA-TAGATCGC_L001_R1_001_trimmed.fastq

- ✓ Basic Statistics
- ✓ Per base sequence quality
- ✓ Per sequence quality scores
- ✗ Per base sequence content
- ✗ Per base GC content
- ! Per sequence GC content
- ✓ Per base N content
- ! Sequence Length Distribution
- ! Sequence Duplication Levels
- ✓ Overrepresented sequences
- ! Kmer Content