Oh the formats you'll see: from BAM to VCV

Mike Campbell, 10x Genomics

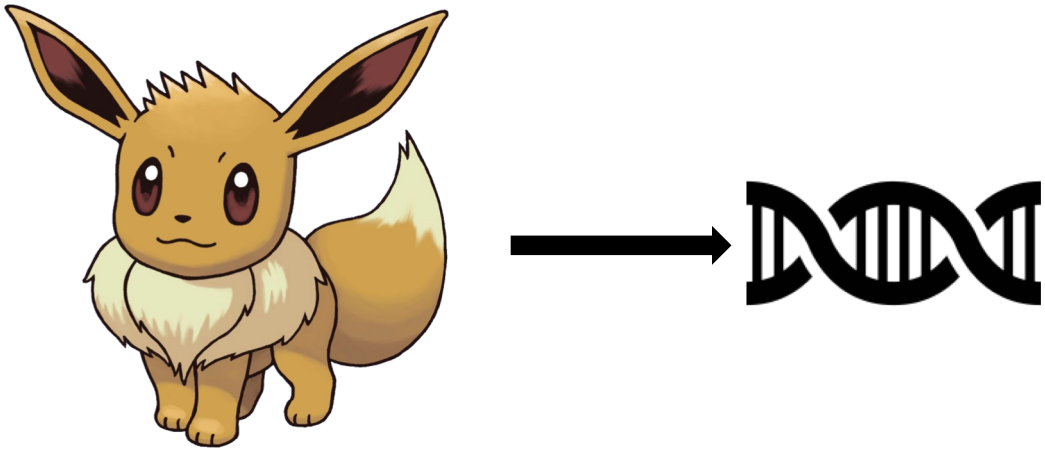
Goals and outline

- Genome project overview
- Understand the importance of standard file formats
- Introduce you to several common formats that you will see in bioinformatics
- Give you some tips on how to parse those files

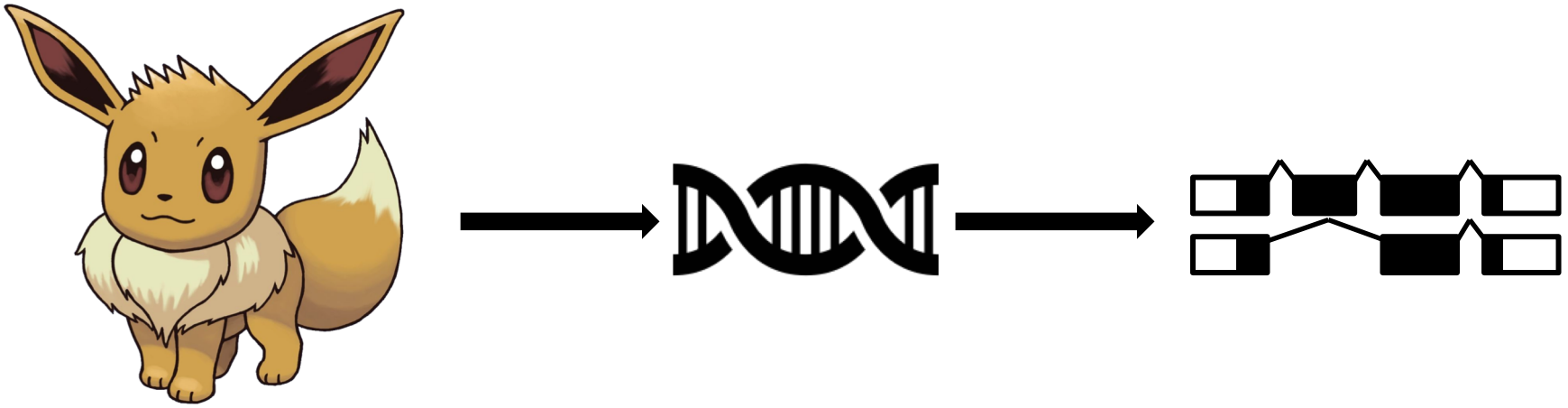
Genome Project Overview

bulbapedia.bulbagarden.net

Genome Project Overview

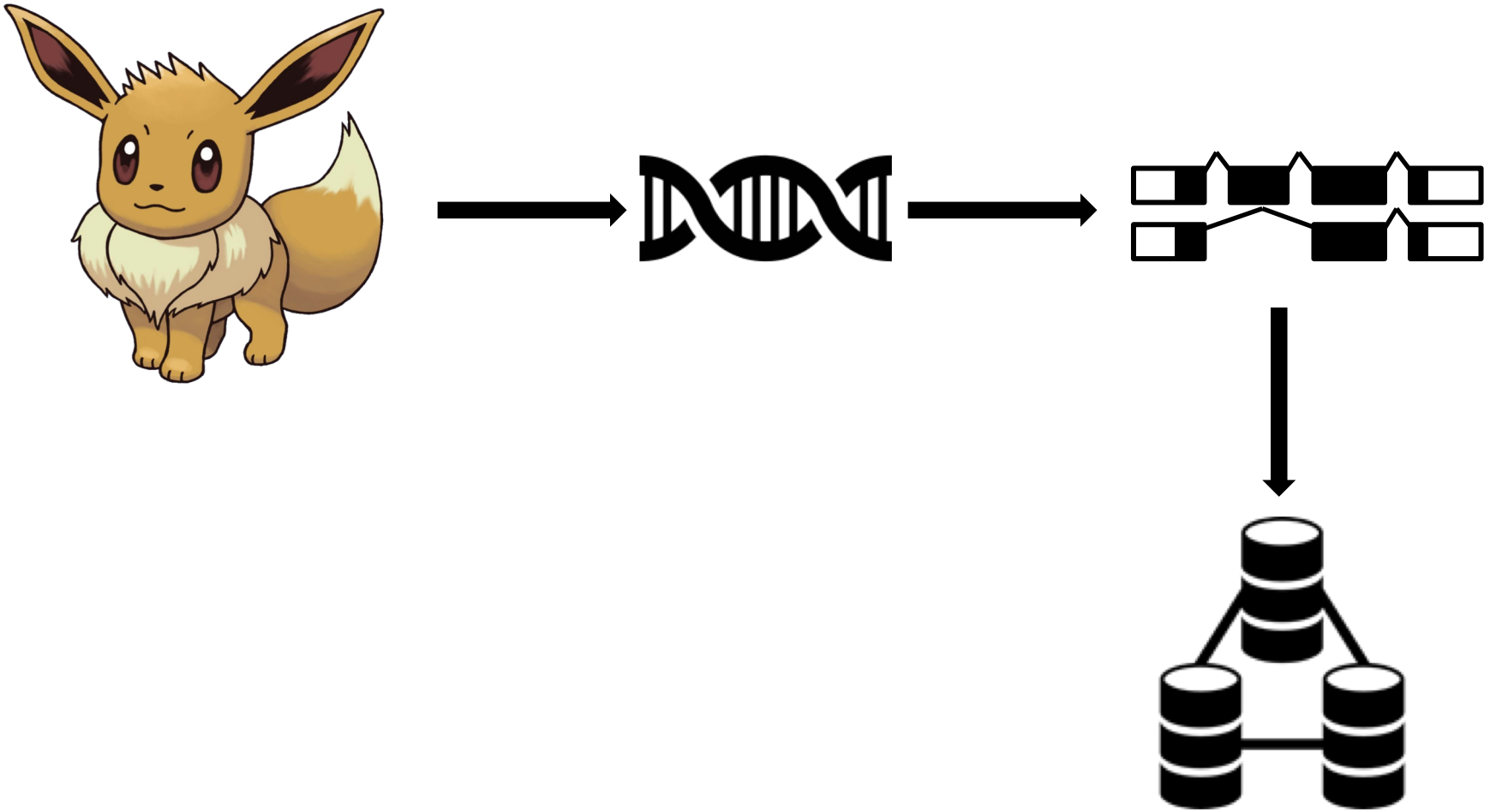


Genome Project Overview



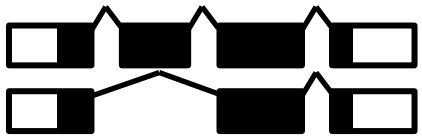
bulbapedia.bulbagarden.net

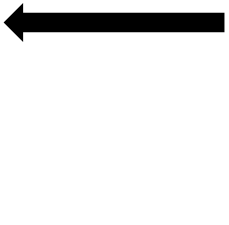
Genome Project Overview



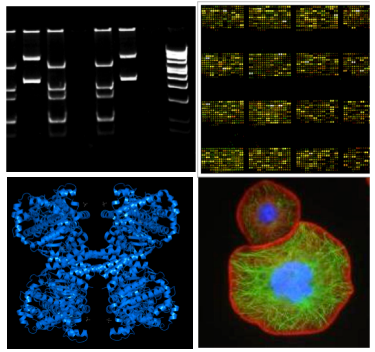
bulbapedia.bulbagarden.net

Genome Project Overview



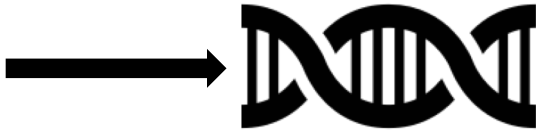


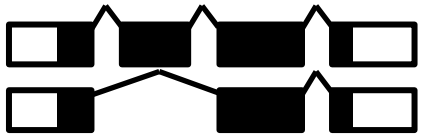
```
>Smg5
MEVTFSSGGSSNASSECAIDGGTNRCRGL
EPNNGTCILSQEVKDLYRSLYTASKQLDD
AKRNVQSVGQLFQHEIEEKRSLLVQLCKQ
IIFKDYQSVGKKVREVMWRRGYYEFIAFV
```

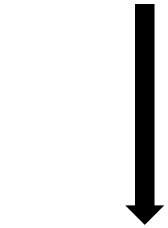


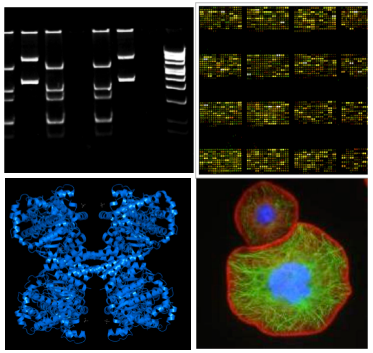
bulbapedia.bulbagarden.net

Genome Project Overview

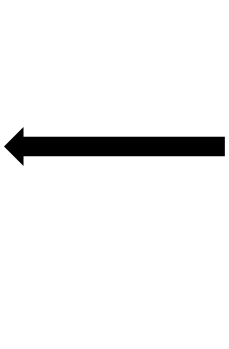








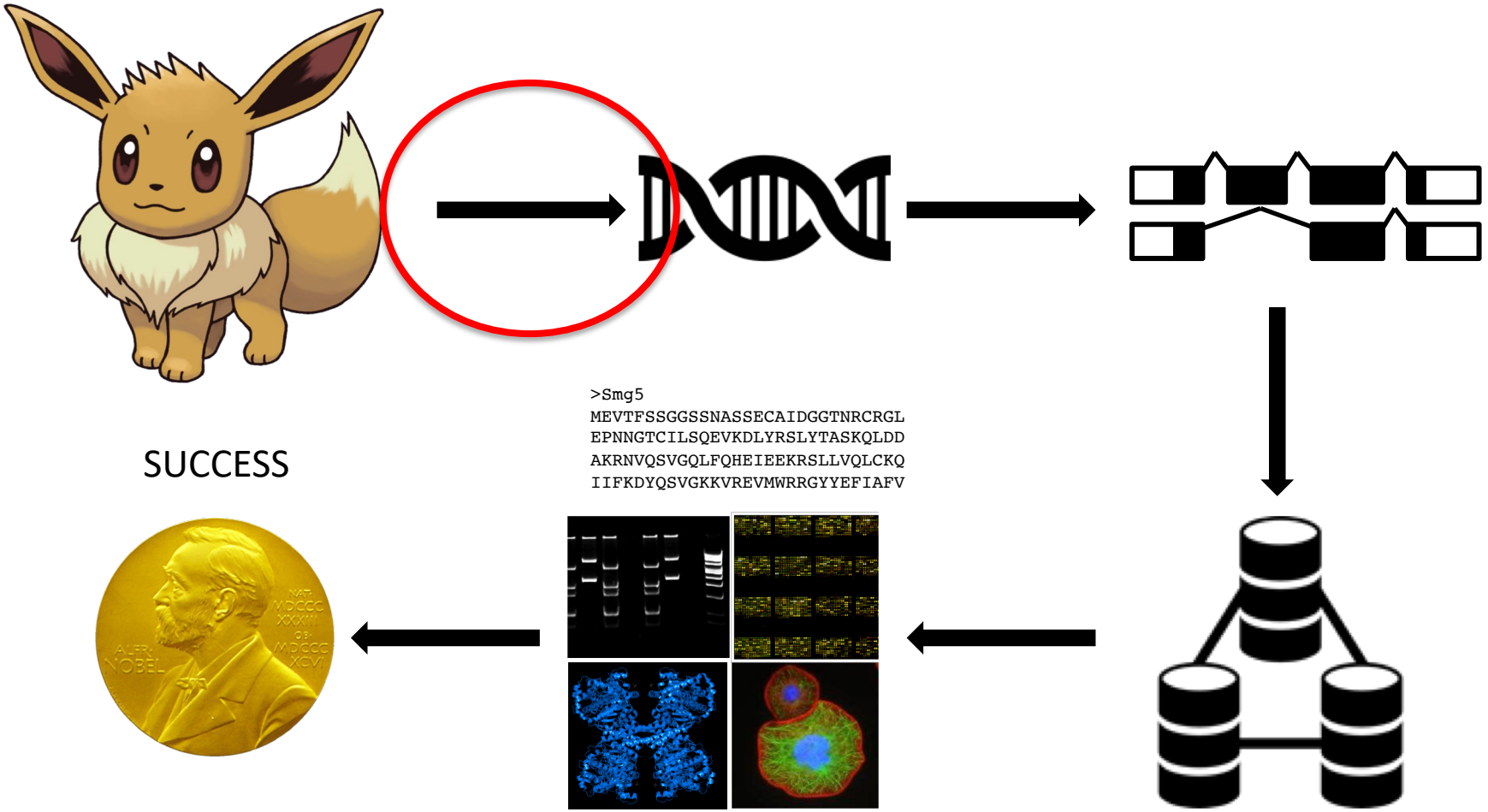
```
>Smg5
MEVTFSSGGSSNASSECAIDGGTNRCRGL
EPNNGTCILSQEVKDLYRSLYTASKQLDD
AKRNVQSVGQLFQHEIEEKRSLLVQLCKQ
IIFKDYQSVGKKVREVMWRRGYYEFIAFV
```



SUCCESS

bulbapedia.bulbagarden.net

Genome Project Overview

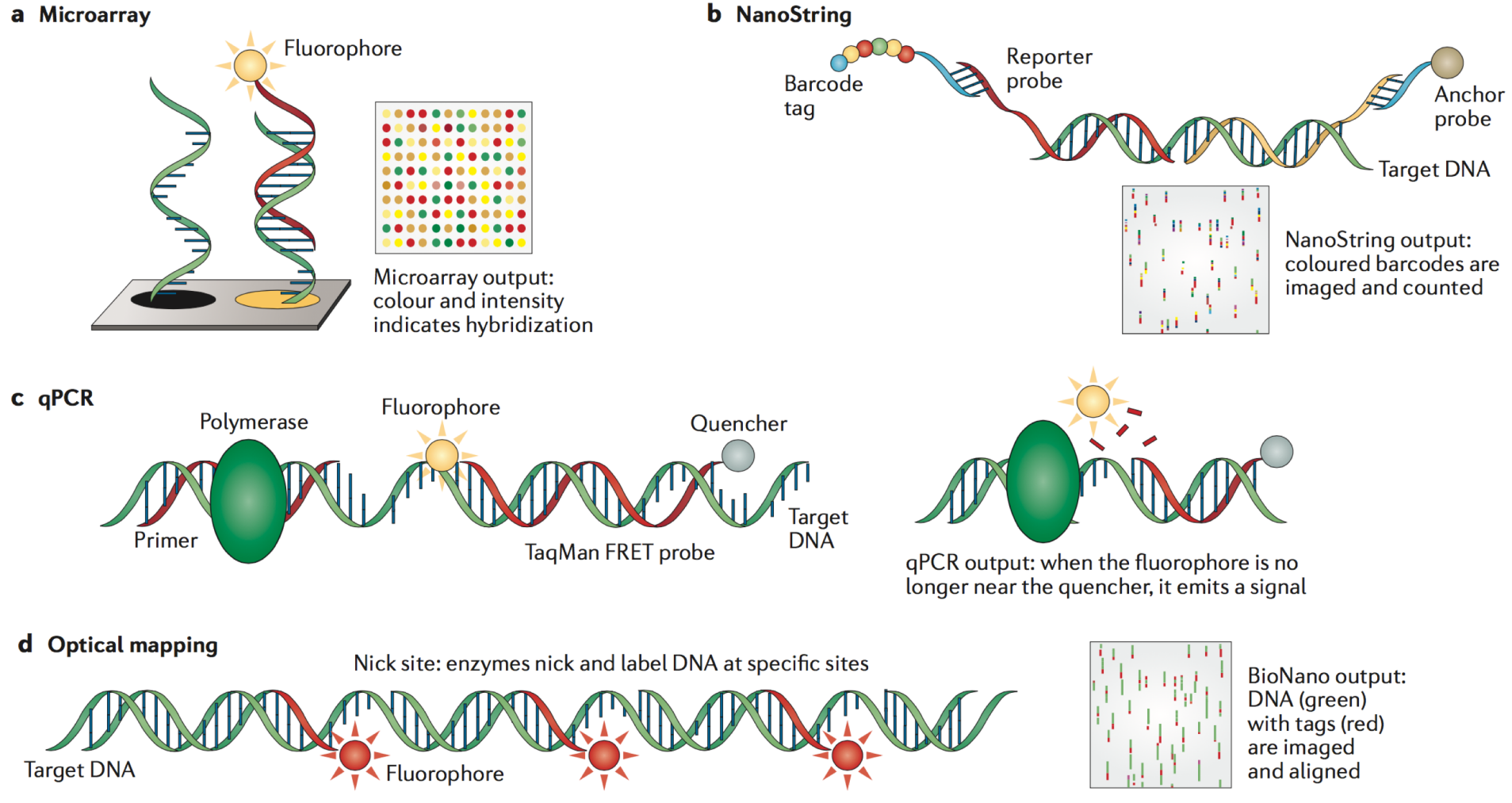


bulbapedia.bulbagarden.net

Coming of age: ten years of next-generation sequencing technologies

Sara Goodwin[1], John D. McPherson[2] and W. Richard McCombie[1]

What did we do before whole genome sequencing?

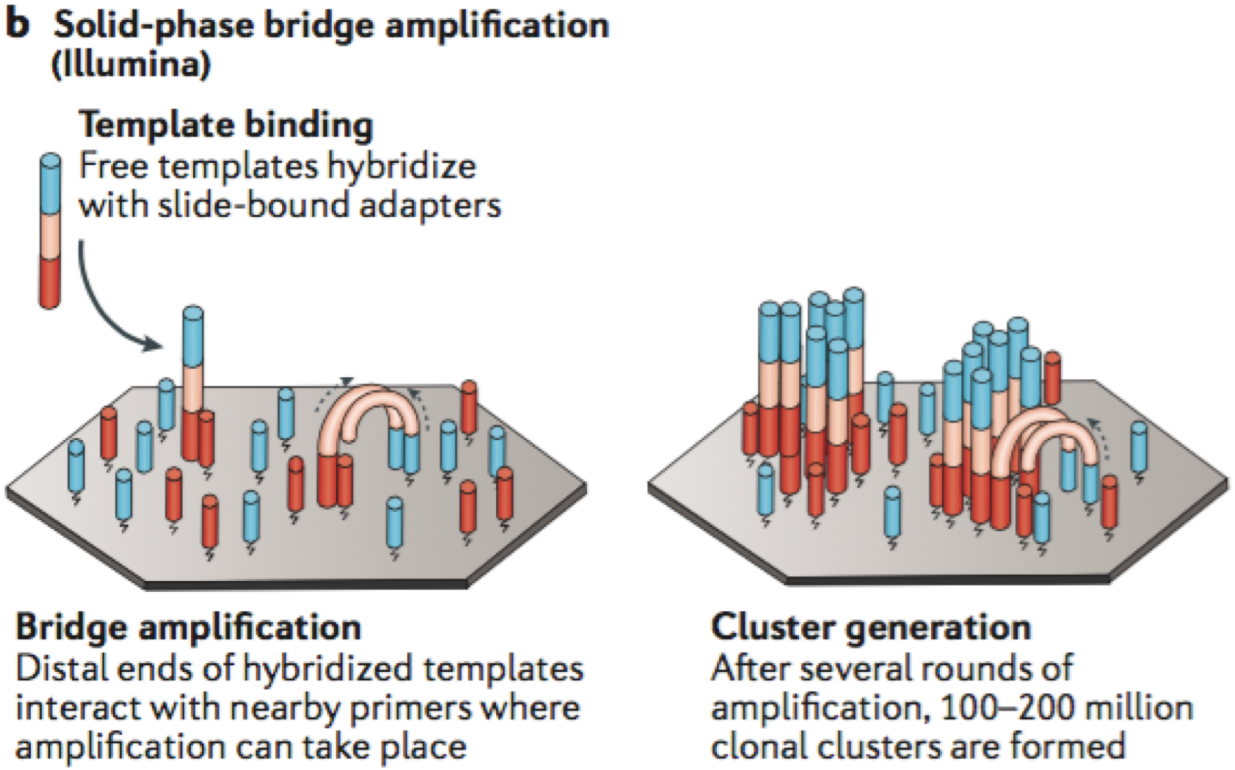


Goodwin et al. 2017

Second Generation Sequencing

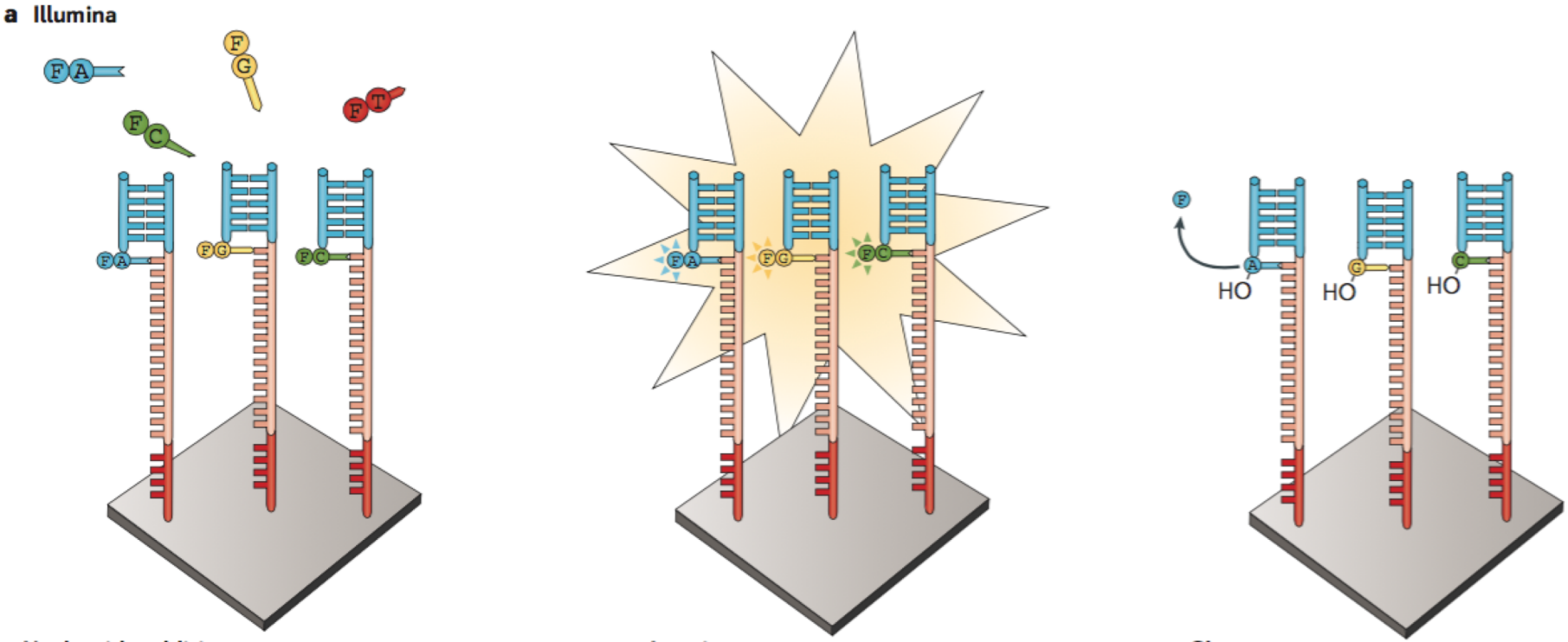
Extract -> amplify -> sequence -> analyze

Illumina



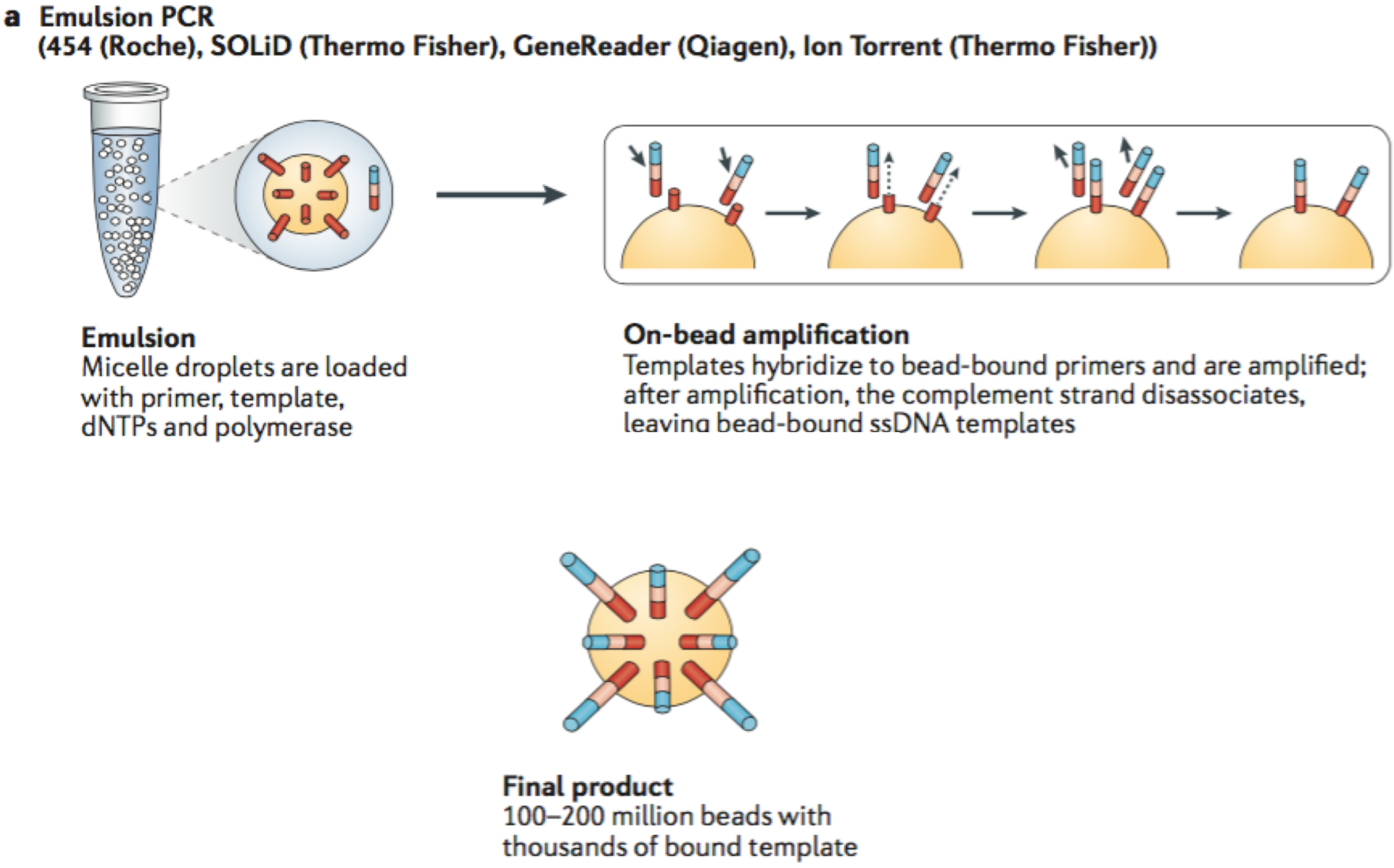
Goodwin et al. 2017

Illumina



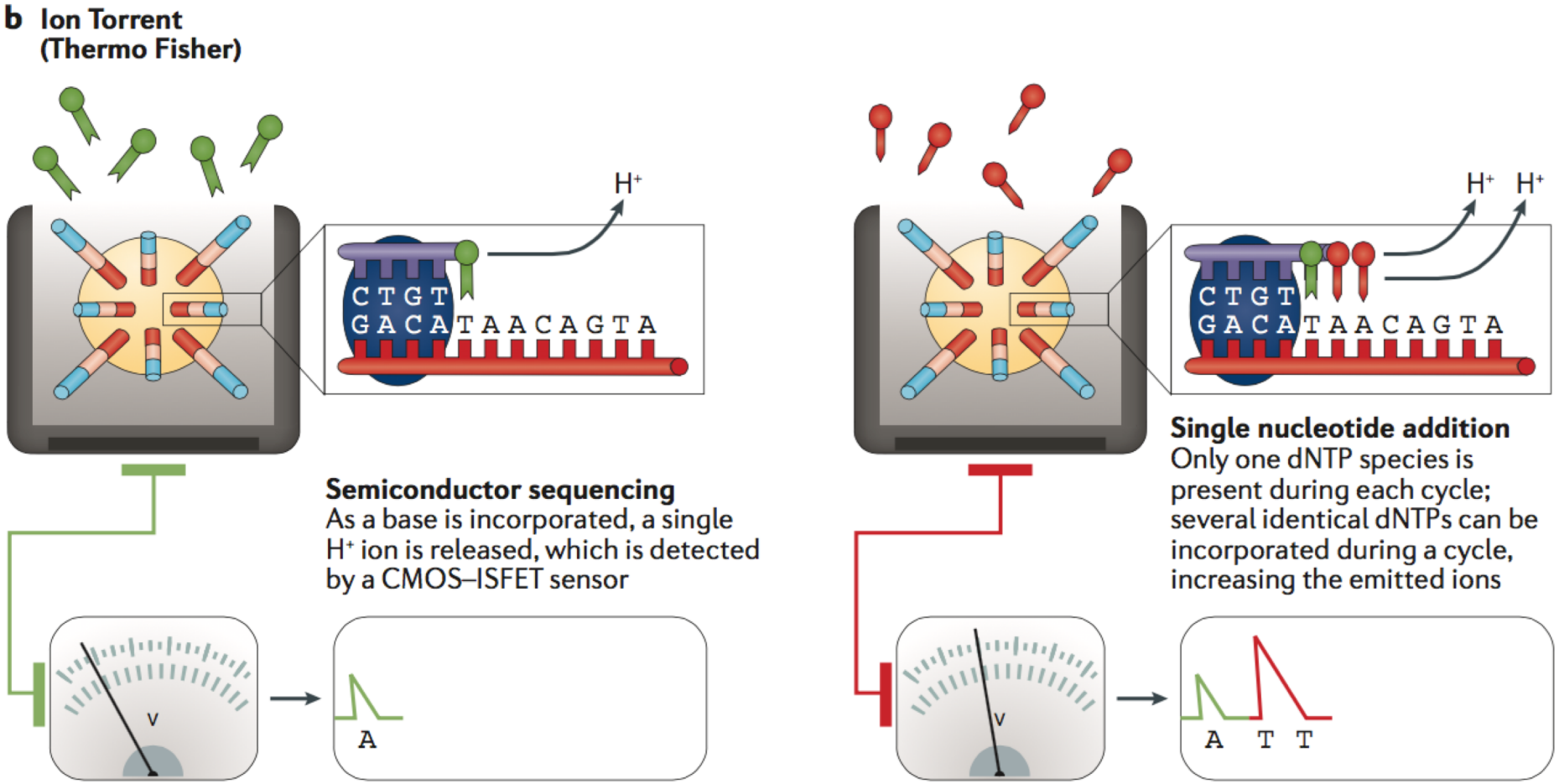
Goodwin et al. 2017

Ion Torent



Goodwin et al. 2017

Ion Torent



Goodwin et al. 2017

Third generation sequencing

Long molecule sequencing

Pac bio

A Real-time long-read sequencing

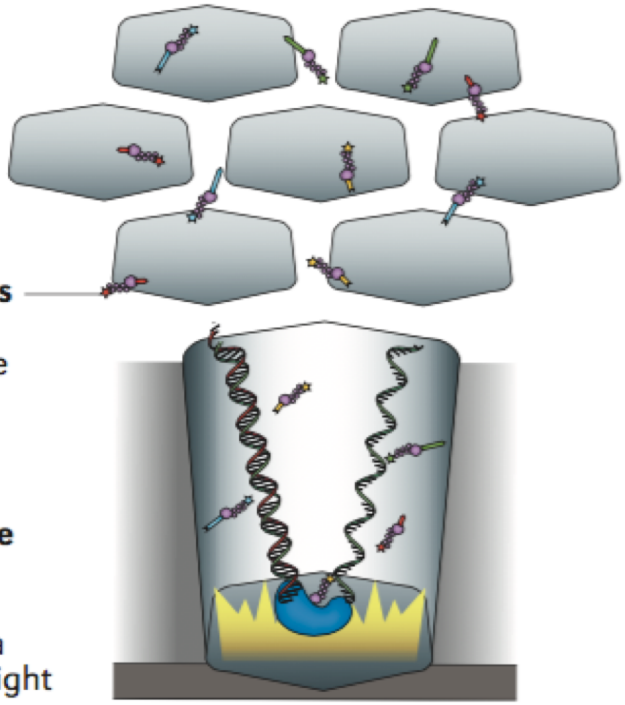
Aa Pacific Biosciences

SMRTbell template
Two hairpin adapters allow continuous circular sequencing

ZMW wells
Sites where sequencing takes place

Labelled nucleotides
All four dNTPs are labelled and available for incorporation

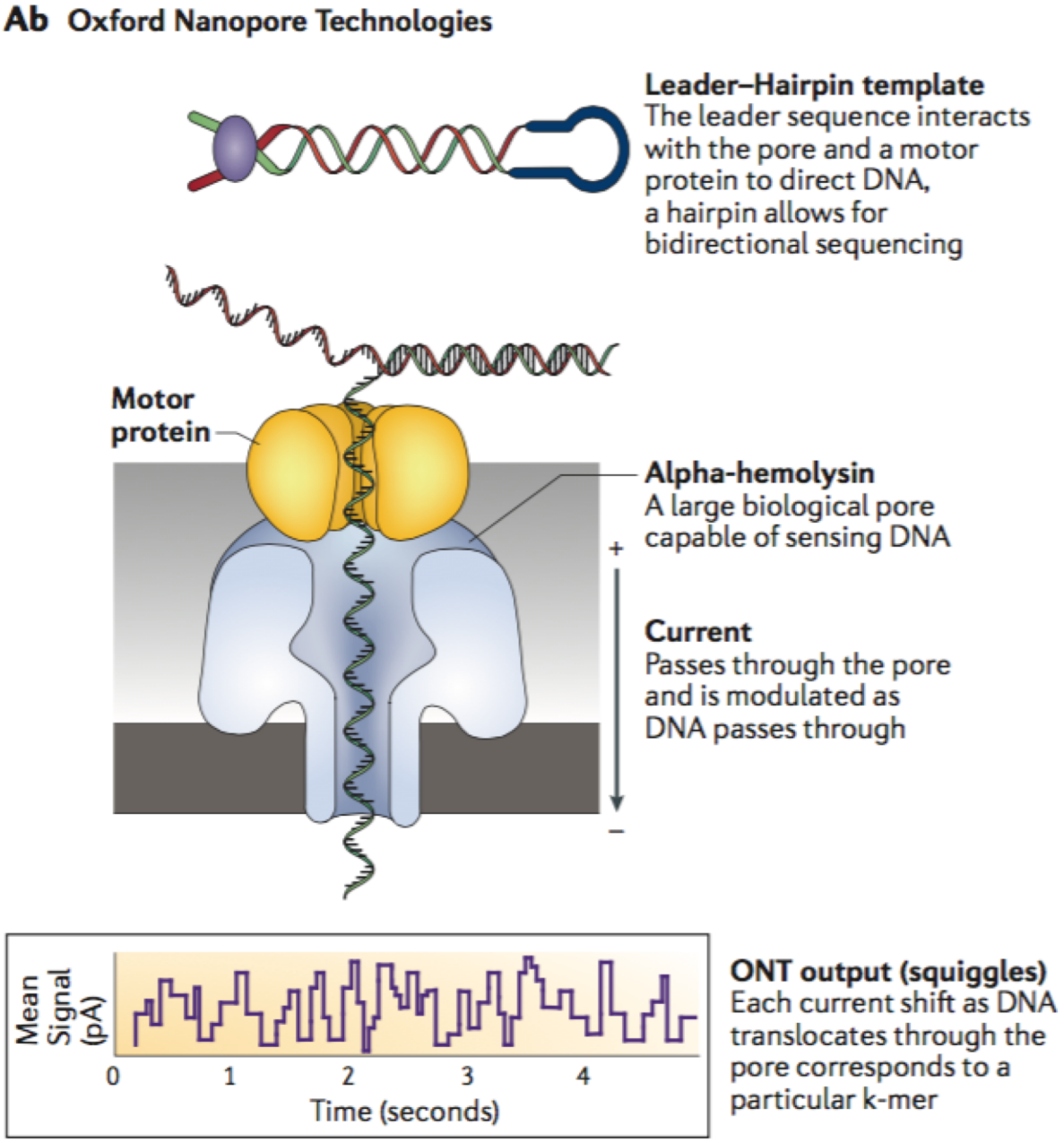
Modified polymerase
As a nucleotide is incorporated by the polymerase, a camera records the emitted light



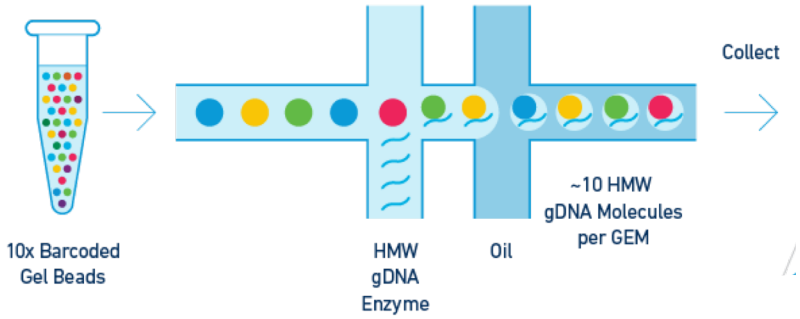
PacBio output
A camera records the changing colours from all ZMWs; each colour change corresponds to one base

Goodwin et al. 2017

Oxford Nanopore

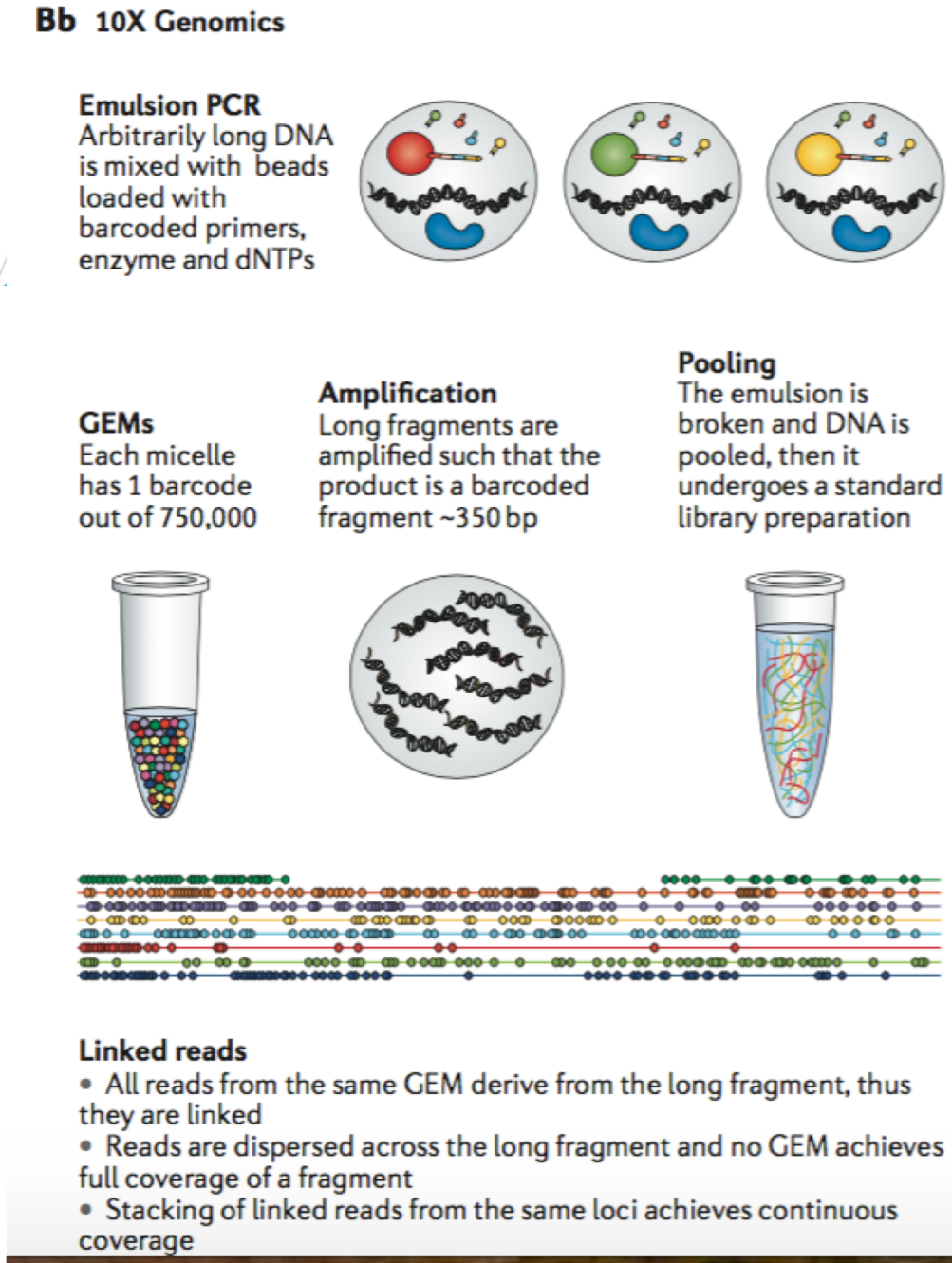


Goodwin et al. 2017



10X genomics linked reads: second generation sequencing technology with a third generation flavor

Goodwin et al. 2017



Whatever technology you use you get standard file formats back

- What is a file format?

Whatever technology you use you get standard file formats back

- What is a file format?
  - A standard way of storing information in a computer file
  - All file formats have specifications some are published others are unpublished

Whatever technology you use you get standard file formats back

- What is a file format?
  - A standard way of storing information in a computer file
  - All file formats have specifications some are published others are unpublished
- Can I tell what kind of file it is by the extension?

Whatever technology you use you get standard file formats back

- What is a file format?
  - A standard way of storing information in a computer file
  - All file formats have specifications some are published others are unpublished
- Can I tell what kind of file it is by the extension?
  - Nope, decent human beings will use an extension that identifies the format, but you can’t count on it
  - Changing a file extension doesn’t change the format of the file

File formats of interest to biologists

- FASTA
- FASTQ
- SAM/BAM
- GFF
- BED
- VCF

FASTA/Pearson format

https://blast.ncbi.nlm.nih.gov/Blast.cgi?CMD=Web&PAGE_TYPE=BlastDocs&DOC_TYPE=BlastHelp

```
>P01013 GENE X PROTEIN (OVALBUMIN-RELATED)
QIKDLLVSSSTDLDTTLVLVNAIYFKGMWKTAFNAEDTREMPFHVTKQESKPVQMMCMNNSFNVATLPAE
KMKILELPFASGDLSMLVLLPDEVSDLERIEKTINFEKLTEWTNPNTMEKRRVKVYLPQMKIEEKYNLTS
VLMALGMTDLFIPSANLTGISSAESLKISQAVHGAFMELSEDGIEMAGSTGVIEDIKHSPESEQFRADHP
FLFLIKHNPTNTIVYFGRYWSP
```

FASTQ (added quality scores)

http://maq.sourceforge.net/fastq.shtml

```
@EAS54_6_R1_2_1_413_324
CCCTTCTTGTCTTCAGCGTTTCTCC
+
;;3;;;;;;;;;;;;7;;;;;;;88
@EAS54_6_R1_2_1_540_792
TTGGCAGGCCAAGGCCGATGGATCA
+
;;;;;;;;;;;7;;;;;-;;;3;83
@EAS54_6_R1_2_1_443_348
GTTGCTTCTGGCGTGGGTGGGGGGG
+EAS54_6_R1_2_1_443_348
;;;;;;;;;;;9;7;;.7;393333
```

SAM/BAM (Sequence alignment map)

https://samtools.github.io/hts-specs/SAMv1.pdf

```
@HD VN:1.5 SO:coordinate
@SQ SN:ref LN:45
r001 99 ref 7 30 8M2I4M1D3M = 37 39 TTAGATAAAGGATACTG *
r002 0 ref 9 30 3S6M1P1I4M * 0 0 AAAAGATAAGGATA *
r003 0 ref 9 30 5S6M * 0 0 GCCTAAGCTAA * SA:Z:ref,29,-,6H5M,17,0;
r004 0 ref 16 30 6M14N5M * 0 0 ATAGCTTCAGC *
r003 2064 ref 29 17 6H5M * 0 0 TAGGC * SA:Z:ref,9,+,5S6M,30,1;
r001 147 ref 37 30 9M = 7 -39 CAGCGGCAT * NM:i:1
```

VCF (Variant call format)

http://samtools.github.io/hts-specs/VCFv4.2.pdf

```
##fileformat=VCFv4.2
##fileDate=20090805
##source=myImputationProgramV3.1
##reference=file:///seq/references/1000GenomesPilot-NCBI36.fasta
##contig=<ID=20,length=62435964,assembly=B36,md5=f126cdf8a6e0c7f379d618ff66beb2da,spe
##phasing=partial
##INFO=<ID=NS,Number=1,Type=Integer,Description="Number of Samples WithData">
##INFO=<ID=DP,Number=1,Type=Integer,Description="Total Depth">
#CHROM POS ID REF ALT QUAL FILTER INFO FORMAT NA00001 NA00002 NA00003
20 14370 rs6054257 G A 29 PASS NS=3;DP=14; GT 0|0 1|0 1/1
20 17330 . T A 3 q10 NS=3;DP=11 GT 0|0 0|1 0/0
20 1110696 rs6040355 A G,T 67 PASS NS=2;DP=10 GT 1|2 2|1 2/2
20 1230237 . T . 47 PASS NS=3 GT 0|0 0|0 0/0
20 1234567 microsat1 GTC G,GTCT 50 PASS NS=3;DP=9 GT 0/1 0/2 1/1
```

BED (Browser Extensible Data)

https://genome.ucsc.edu/FAQ/FAQformat.html#format1

```
browser position chr7:127471196-127495720
browser hide all
track name="ItemRGBDemo" description="Item RGB demonstration"
visibility=2 itemRgb="On"
chr7    127471196  127472363  Pos1  0  +  127471196  127472363  255,0,0
chr7    127472363  127473530  Pos2  0  +  127472363  127473530  255,0,0
chr7    127473530  127474697  Pos3  0  +  127473530  127474697  255,0,0
chr7    127474697  127475864  Pos4  0  +  127474697  127475864  255,0,0
chr7    127475864  127477031  Neg1  0  -  127475864  127477031  0,0,255
chr7    127477031  127478198  Neg2  0  -  127477031  127478198  0,0,255
chr7    127478198  127479365  Neg3  0  -  127478198  127479365  0,0,255
chr7    127479365  127480532  Pos5  0  +  127479365  127480532  255,0,0
chr7    127480532  127481699  Neg4  0  -  127480532  127481699  0,0,255
```

The first three columns are required

Genome browser view

GFF3 format

https://github.com/The-Sequence-Ontology/Specifications/blob/master/gff3.md

```
##gff-version 3
ctg123 . gene            1300  9000  .  +  .  ID=gene0001;Name=sonic;biotype=protein
ctg123 . mRNA            1300  9000  .  +  .  ID=mrna0001;Parent=gene0001;Name=sonic
ctg123 . exon            1300  1500  .  +  .  ID=exon00001;Parent=mrna0001
ctg123 . exon            1050  1500  .  +  .  ID=exon00002;Parent=mrna0001
ctg123 . exon            3000  3902  .  +  .  ID=exon00003;Parent=mrna0001
ctg123 . exon            5000  5500  .  +  .  ID=exon00004;Parent=mrna0001
ctg123 . exon            7000  9000  .  +  .  ID=exon00005;Parent=mrna0001
```

FASTQ (added quality scores)

http://maq.sourceforge.net/fastq.shtml

```
@EAS54_6_R1_2_1_413_324
CCCTTCTTGTCTTCAGCGTTTCTCC
+
;;3;;;;;;;;;;;;7;;;;;;;88
@EAS54_6_R1_2_1_540_792
TTGGCAGGCCAAGGCCGATGGATCA
+
;;;;;;;;;;;7;;;;;-;;;3;83
@EAS54_6_R1_2_1_443_348
GTTGCTTCTGGCGTGGGTGGGGGGG
+EAS54_6_R1_2_1_443_348
;;;;;;;;;;;9;7;;.7;393333
```

Day to day applications: Quality control

$$Q = -10\log_{10}(E)$$

Where E = estimated probability of the base call being wrong

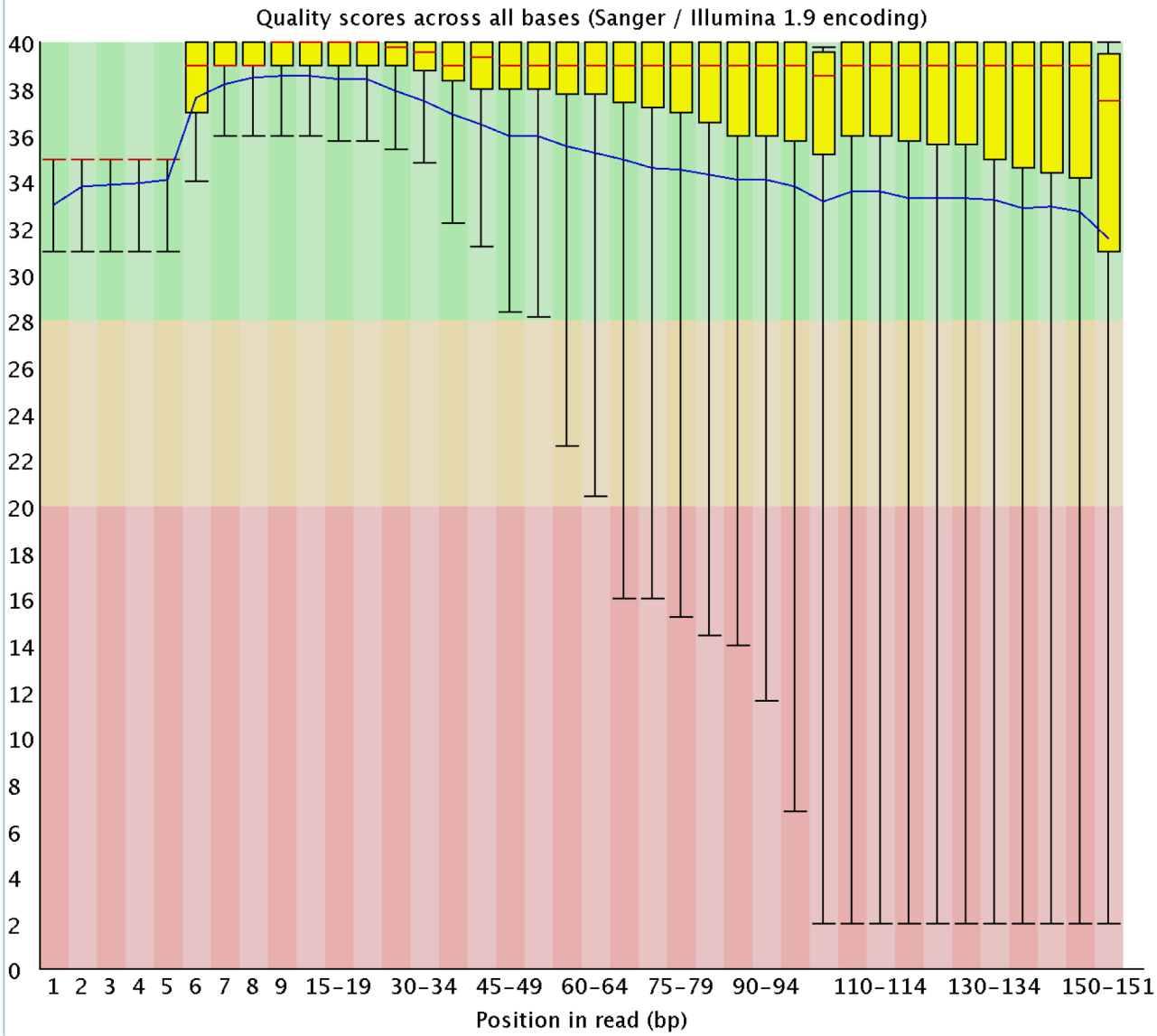
Q	Probability of incorrect base call	Inferred base call accuracy
10	1 in 10	90%
20	1 in 100	99%
30	1 in 1000	99.9%

Quality scores are platform dependent

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)
with 0=unused, 1=unused, 2=Read Segment Quality Control Indicator (bold)
(Note: See discussion above).
L - Illumina 1.8+ Phred+33, raw reads typically (0, 41)

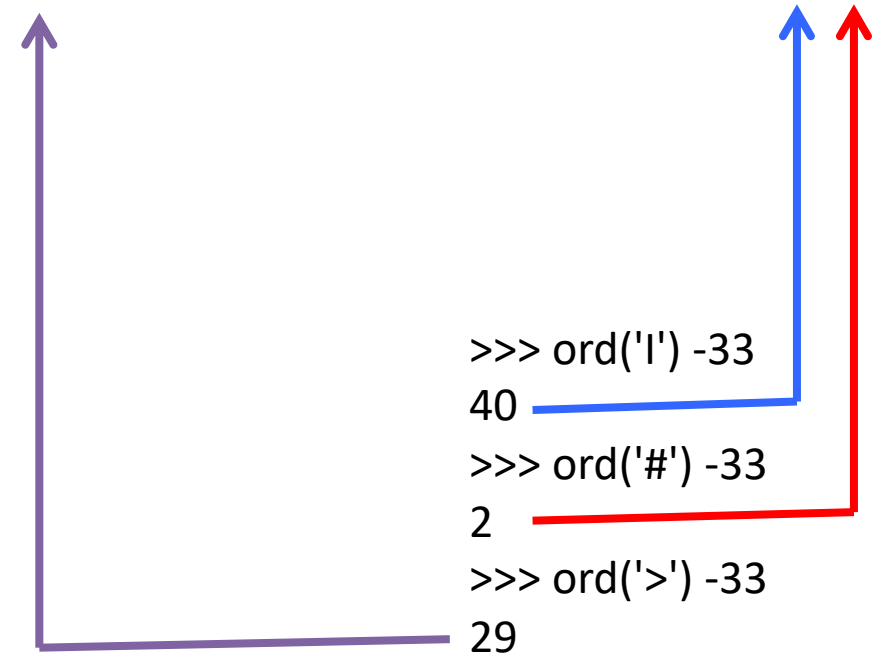
https://en.wikipedia.org/wiki/FASTQ_format

FASTQC for looking at base quality



Quality trimming

```
@D00777:109:HFCWVBCXY:1:1103:1477:2081 1:N:0
GGCTAGCATTTAGGCGATAGAGAATTTTNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNN
+
<<DDHHIIIIIIIIHIIHHIIIIIIIII##############################
```



Day to day applications: Fixing files

FASTA

```
>000263F_quiver
AATGGCAGTTGTGTGCTGTCACATGATGCCCCTTTGCATGGTTGTGTAATGGGTTTTGCAGCCTGCATGATCACGTACAT
GGTGCAAAATCCTCAAAGGTGTAGCCTTTGCAATCCAATGAAGGTTTCCTTTTAAATCACGAACAATGCGCGCTGTAGGC
CACCATGGATGCGTGTGCTATGTTGGCTGTTTGTGCTATTTTTTCGAGAATTTTAATGGCAGAGTGGGGTCCTCATGATG
CAAGTCTGCATAATGGTGTAGTAGTGGGGAATGATAATCTGGTGGTGATAGGGTAGGGTGGCAGTGGTAGCATGGTGGTT
ATGGCTATGAGATTTGCGGTAGTGGCGATCATGGTGATGAAGGGTGGCGTCATGAATGGCTAGTGGAGGTCATGGAGGCG
ACCATGATGGTGCAACAGGGTGGTAGAGACTGGCGGTGGACGACGGTGCTGATAGTAACCACGATTTACGCAGCGGAAAT
AAATCATGATGAAACAAGCATGGAGACAAGCCAAAATCATATGTAAAAGACTGACCTTTTTCTTTGTTGAAAACGCCCAC
AATCTCCTACAATCTTCTGTTGTTTACCTTATCAATTTTCCGGACTGTAATTTCTTAGTGGACAGAAAAAACAATAGCTA
GGGTTAGTCACTCTCTTTTCTGTTTCTAATATATATCTCTTCGATGGGAAGAAGACCACTTTCTTAGAAACTCTTTTTTC
```

GFF3

000263F|quiver maker gene 41412 41819 . + . ID=CASF_028122
000263F|quiver maker mRNA 41412 41819 . + . ID=CASF_028122-RA;Parent=CASF_028122;Name=CASF_028122-RA;Alias=augustus_masked-000263F_quiver-processed-gene-0.9-mRNA-1;_AED=0.20;_QI=0_-1_0_1_-1_1_1_0_135;_eAED=0.20;

Day to day applications: Fixing files

FASTA

```
>000263F_quiver
AATGGCAGTTGTGTGCTGTCACATGATGCCCCTTTGCATGGTTGTGTAATGGGTTTTGCAGCCTGCATGATCACGTACAT
GGTGCAAAATCCTCAAAGGTGTAGCCTTTGCAATCCAATGAAGGTTTCCTTTTAAATCACGAACAATGCGCGCTGTAGGC
CACCATGGATGCGTGTGCTATGTTGGCTGTTTGTGCTATTTTTTCGAGAATTTTAATGGCAGAGTGGGGTCCTCATGATG
CAAGTCTGCATAATGGTGTAGTAGTGGGGAATGATAATCTGGTGGTGATAGGGTAGGGTGGCAGTGGTAGCATGGTGGTT
ATGGCTATGAGATTTGCGGTAGTGGCGATCATGGTGATGAAGGGTGGCGTCATGAATGGCTAGTGGAGGTCATGGAGGCG
ACCATGATGGTGCAACAGGGTGGTAGAGACTGGCGGTGGACGACGGTGCTGATAGTAACCACGATTTACGCAGCGGAAAT
AAATCATGATGAAACAAGCATGGAGACAAGCCAAAATCATATGTAAAAGACTGACCTTTTTCTTTGTTGAAAACGCCCAC
AATCTCCTACAATCTTCTGTTGTTTACCTTATCAATTTTCCGGACTGTAATTTCTTAGTGGACAGAAAAAACAATAGCTA
GGGTTAGTCACTCTCTTTTCTGTTTCTAATATATATCTCTTCGATGGGAAGAAGACCACTTTCTTAGAAACTCTTTTTTC
```

GFF3

000263F|quiver maker gene 41412 41819 . + . ID=CASF_028122
000263F|quiver maker mRNA 41412 41819 . + . ID=CASF_028122-RA;Parent=CASF_028122;Name=CASF_028122-RA;Alias=augustus_masked-000263F_quiver-processed-gene-0.9-mRNA-1;_AED=0.20;_QI=0_-1_0_1_-1_1_1_0_135;_eAED=0.20;