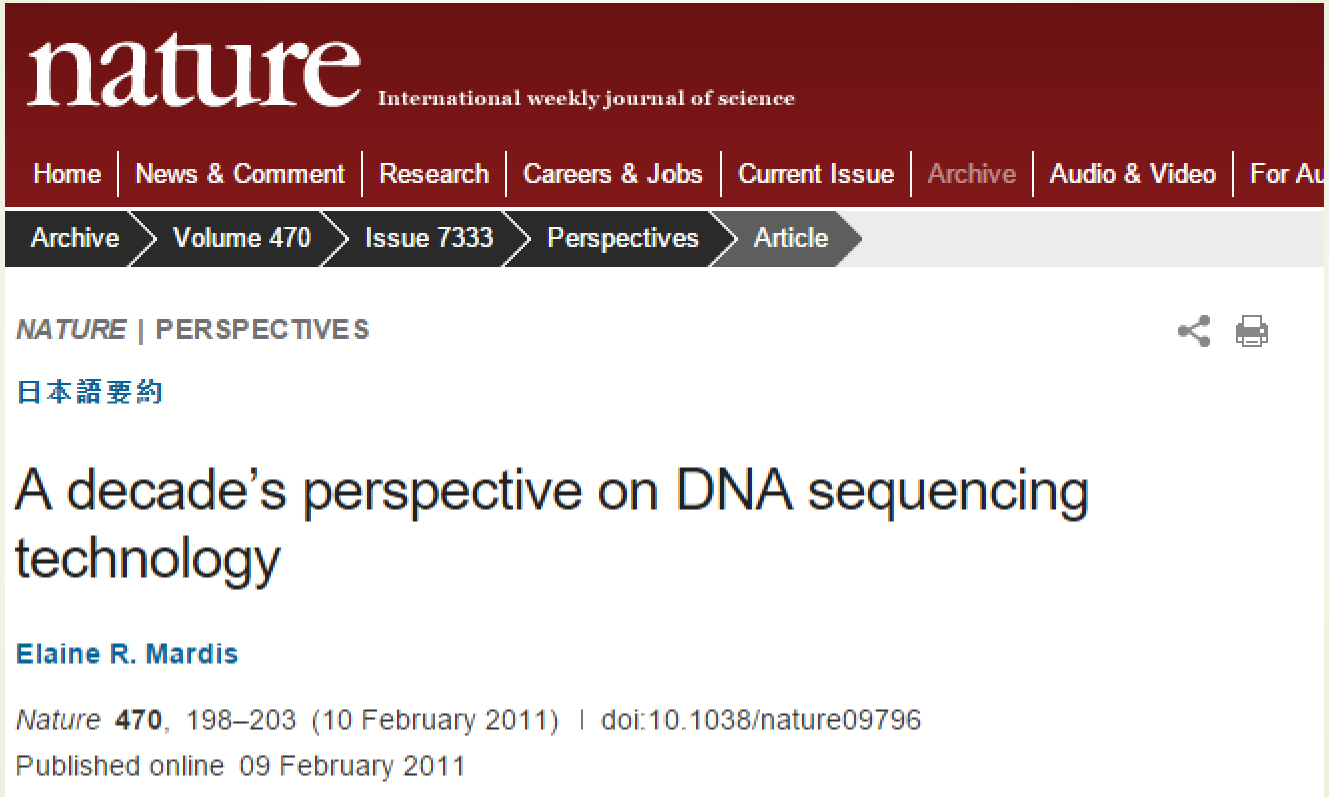
Introduction to NGS

Fotis E. Psomopoulos

CODATA-RDA Advanced Bioinformatics Workshop, 20-24 August 2018, Trieste, Italy

Sequencing Technology



nature International weekly journal of science

Home | News & Comment | Research | Careers & Jobs | Current Issue | Archive | Audio & Video | For Au

Archive > Volume 470 > Issue 7333 > Perspectives > Article

NATURE | PERSPECTIVES

日本語要約

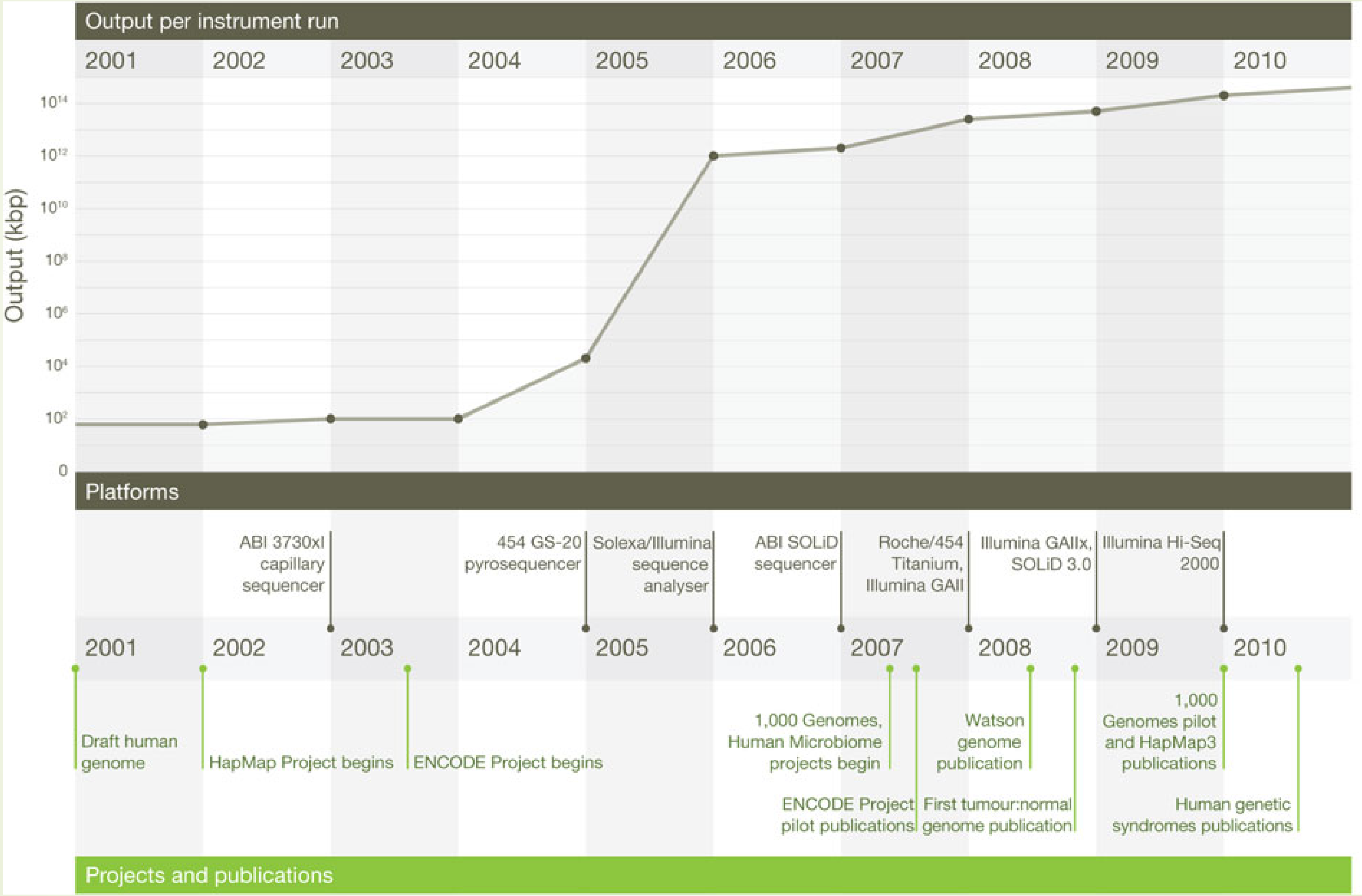
A decade's perspective on DNA sequencing technology

Elaine R. Mardis

Nature **470**, 198–203 (10 February 2011) | doi:10.1038/nature09796

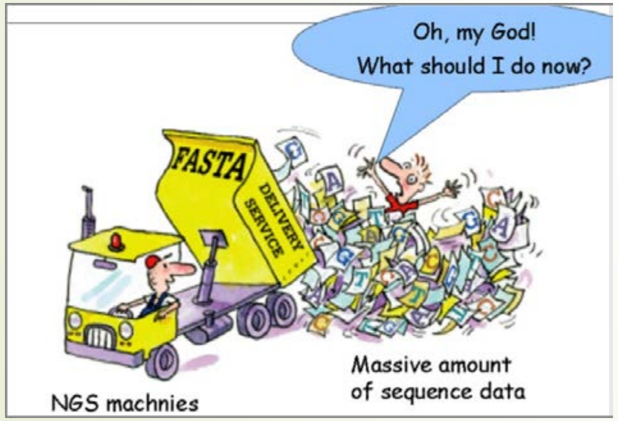
Published online 09 February 2011

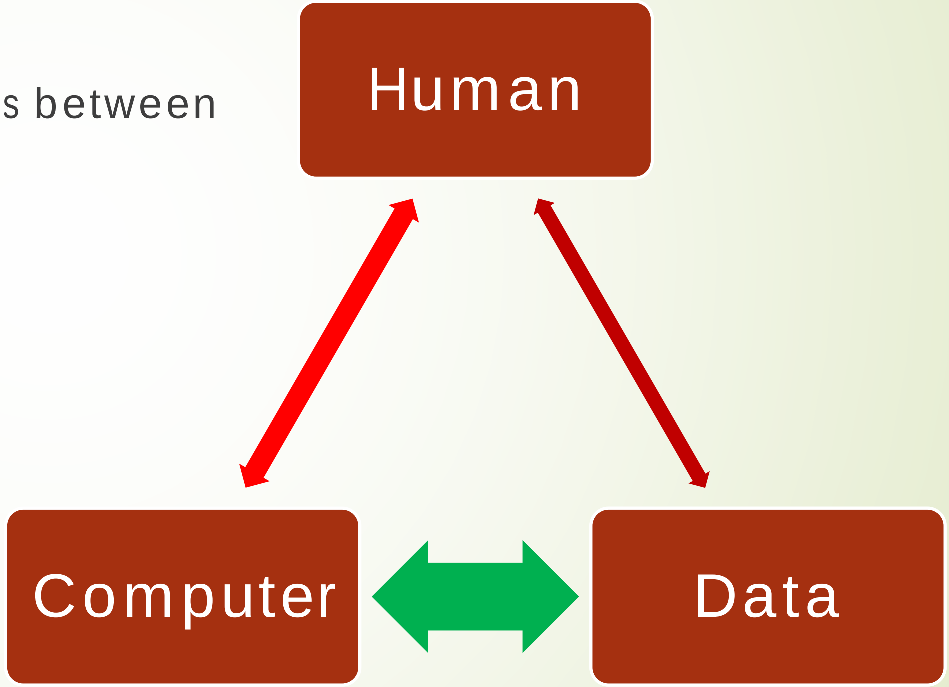
Changes and Timing past decade



The (new) flow of information

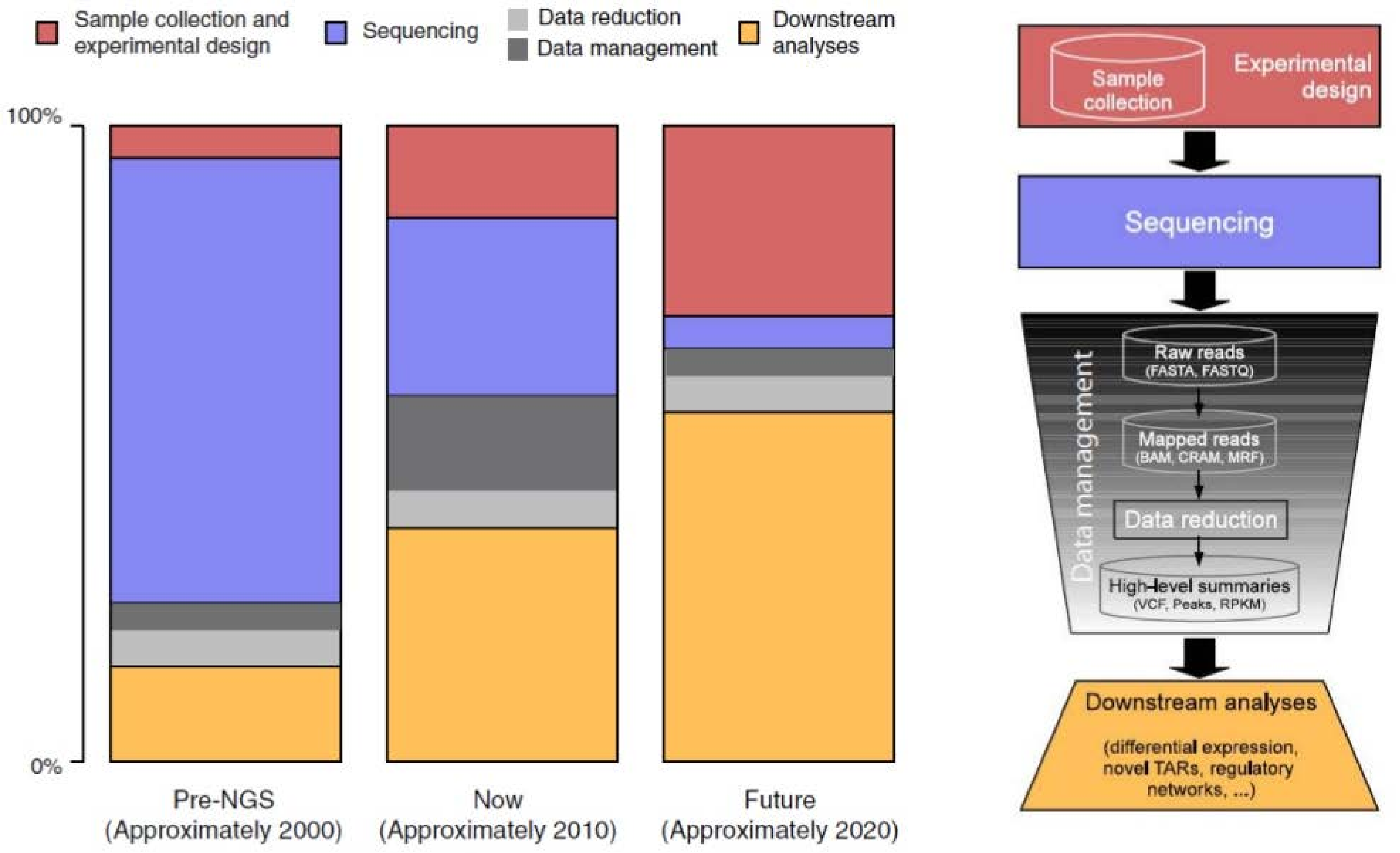
- The trinity of human, data and computer*
  - Extremely high bandwidth between computer and data.
  - Narrow communication channels between human and computer / data.



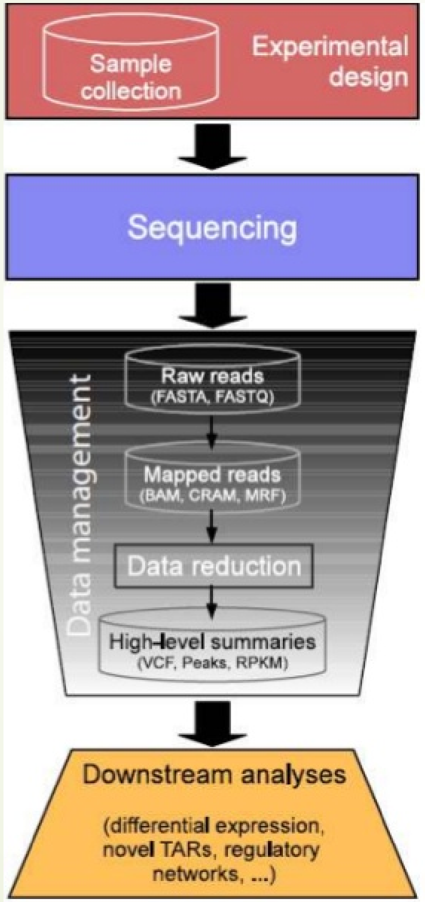


**http://www.kdnuggets.com/2016/08/data-science-challenges.html*

Overview of costs (past, present and near future)



Steps in sequencing experiments


Sample collection
Experimental design
Sequencing
Data management
Raw reads (FASTA, FASTQ)
Mapped reads (BAM, CRAM, MRF)
Data reduction
High-level summaries (VCF, Peaks, RPKM)
Downstream analyses
(differential expression, novel TARs, regulatory networks, ...)

Data analysis

Raw machine reads… What's next?

Preprocessing (machine/technology)

- adaptors, indexes, conversions,…
- machine/technology dependent

Reads with associated qualities (universal)

- FASTQ
- QC check

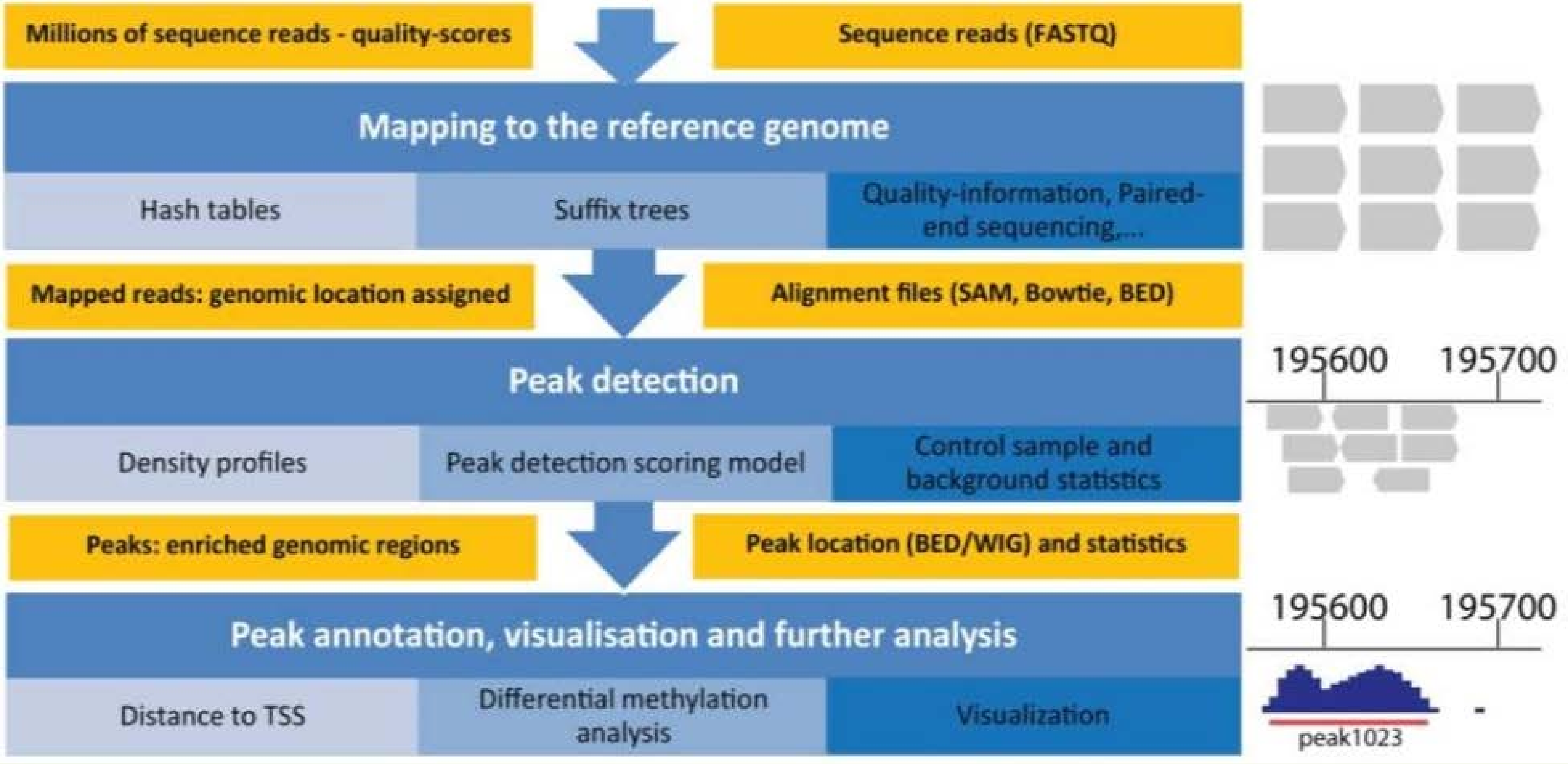
Depending on application (general applicable)

- 'de novo' assembly of genome (bacterial genomes,…)
- Mapping to a reference genome → mapped reads
- SAM/BAM/…

High-level analysis (specific for application)

- SNP calling
- Peak calling

NGS analysis workflow



Millions of sequence reads - quality-scores

Sequence reads (FASTQ)

Mapping to the reference genome

Hash tables

Suffix trees

Quality-information, Paired-end sequencing,...

Mapped reads: genomic location assigned

Alignment files (SAM, Bowtie, BED)

Peak detection

Density profiles

Peak detection scoring model

Control sample and background statistics

Peaks: enriched genomic regions

Peak location (BED/WIG) and statistics

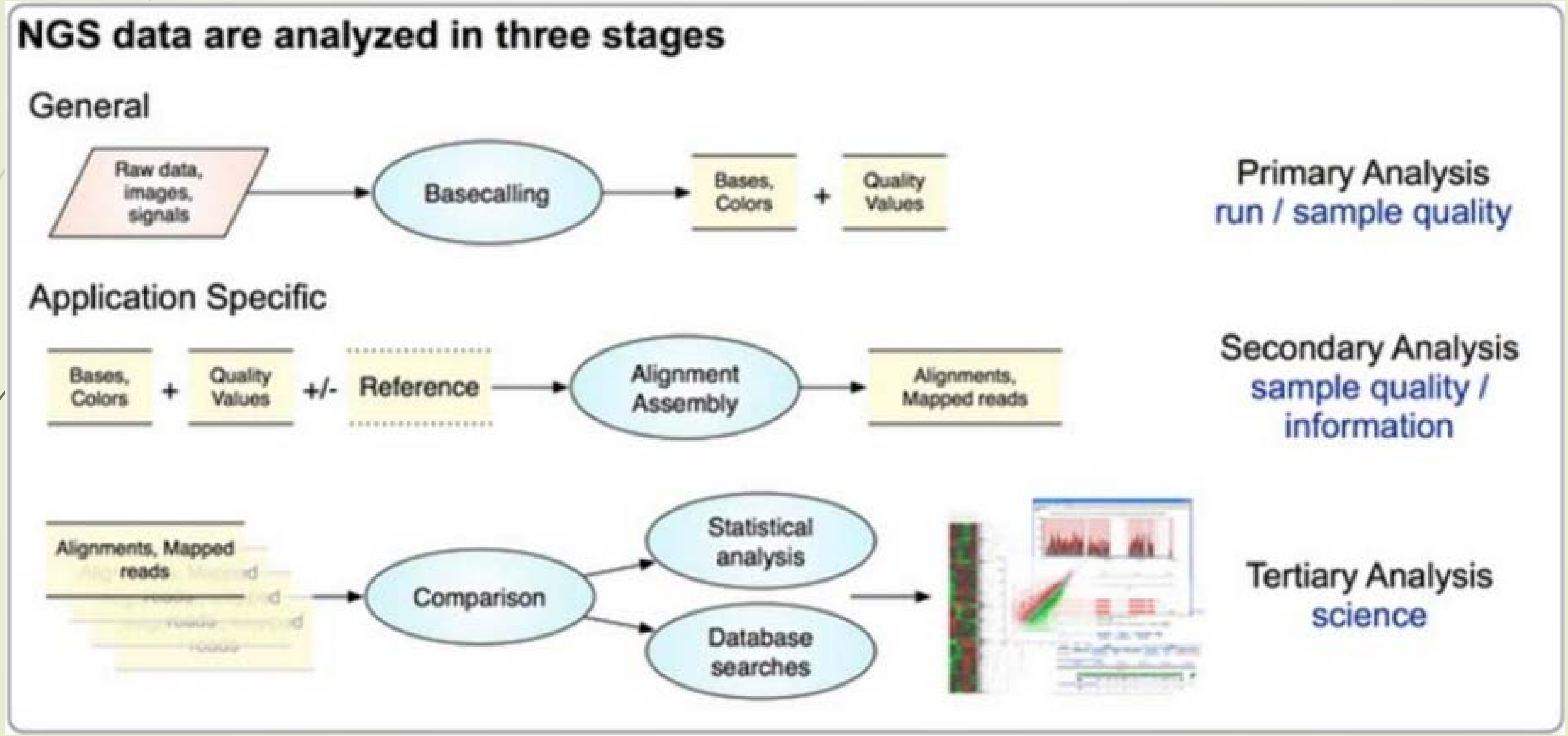
Peak annotation, visualisation and further analysis

Distance to TSS

Differential methylation analysis

Visualization

The three stages of NGS data analysis



NGS data are analyzed in three stages

General

Raw data, images, signals → Basecalling → Bases, Colors + Quality Values

Primary Analysis
run / sample quality

Application Specific

Bases, Colors + Quality Values +/- Reference → Alignment Assembly → Alignments, Mapped reads

Secondary Analysis
sample quality / information

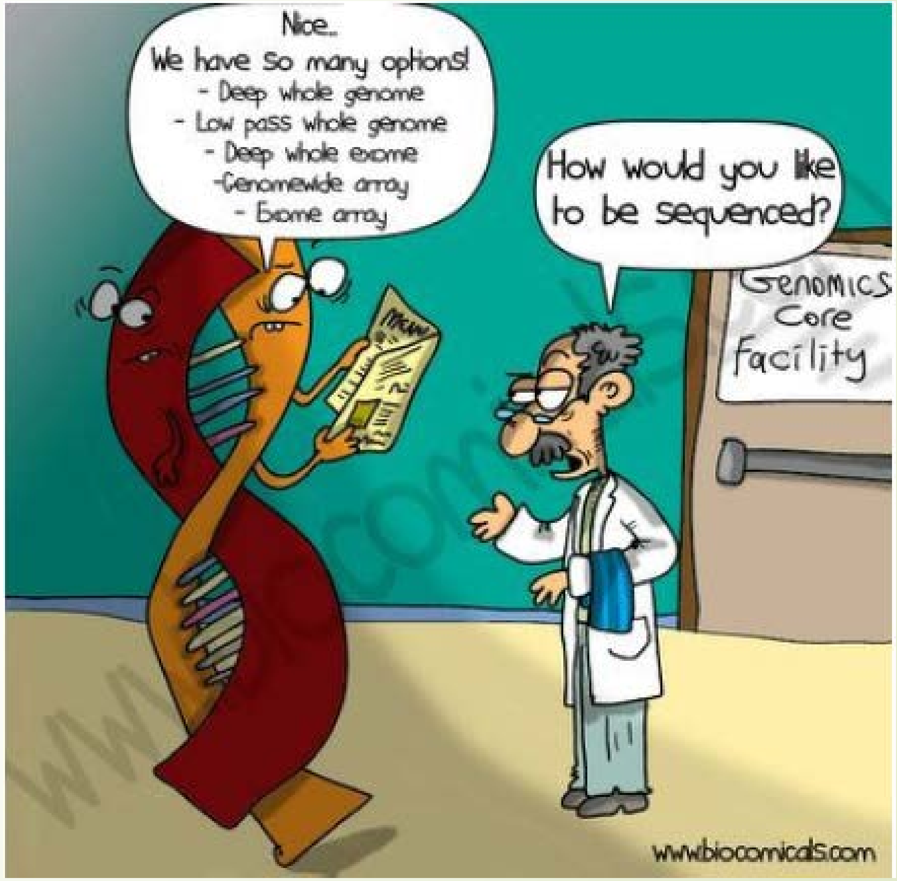
Alignments, Mapped reads → Comparison → Statistical analysis / Database searches →

Tertiary Analysis
science

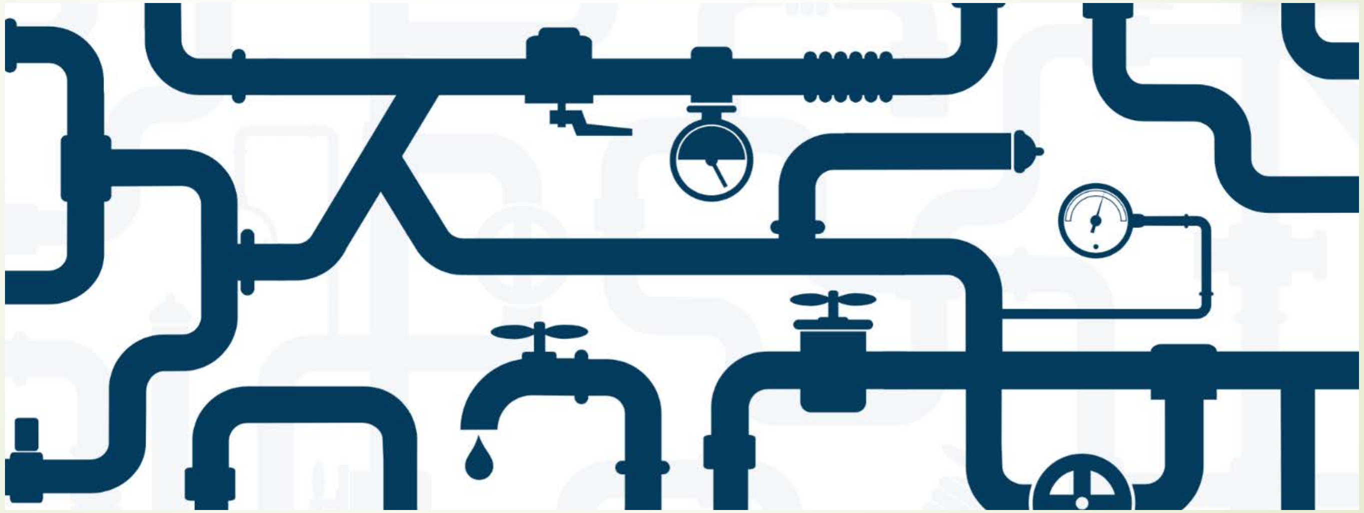
- We will try to provide an overview of all steps in this course

NGS Applications are **sequencing** applications

- Whole Genome Sequencing
- Gene Regulation
- Epigenetic Changes
- Metagenomics
- Paleogenomics
- Transcriptome Analysis
- Resequencing
- ….

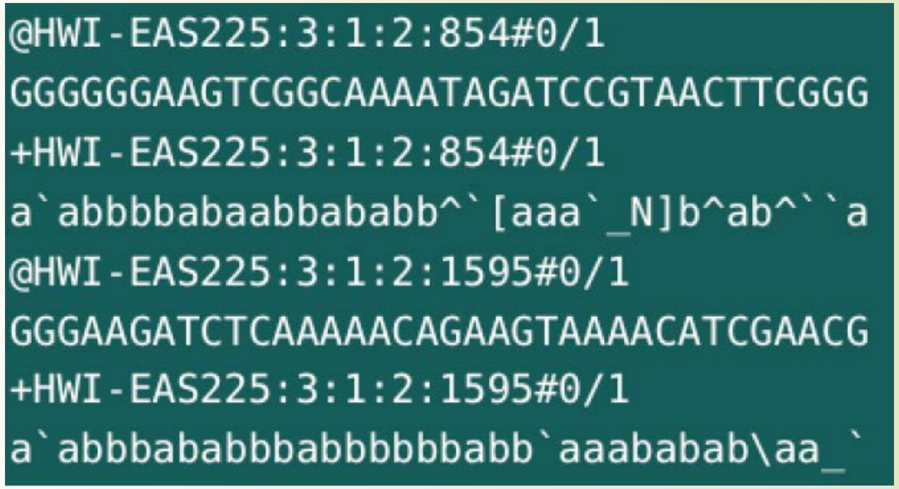


End-to-end computational workflows



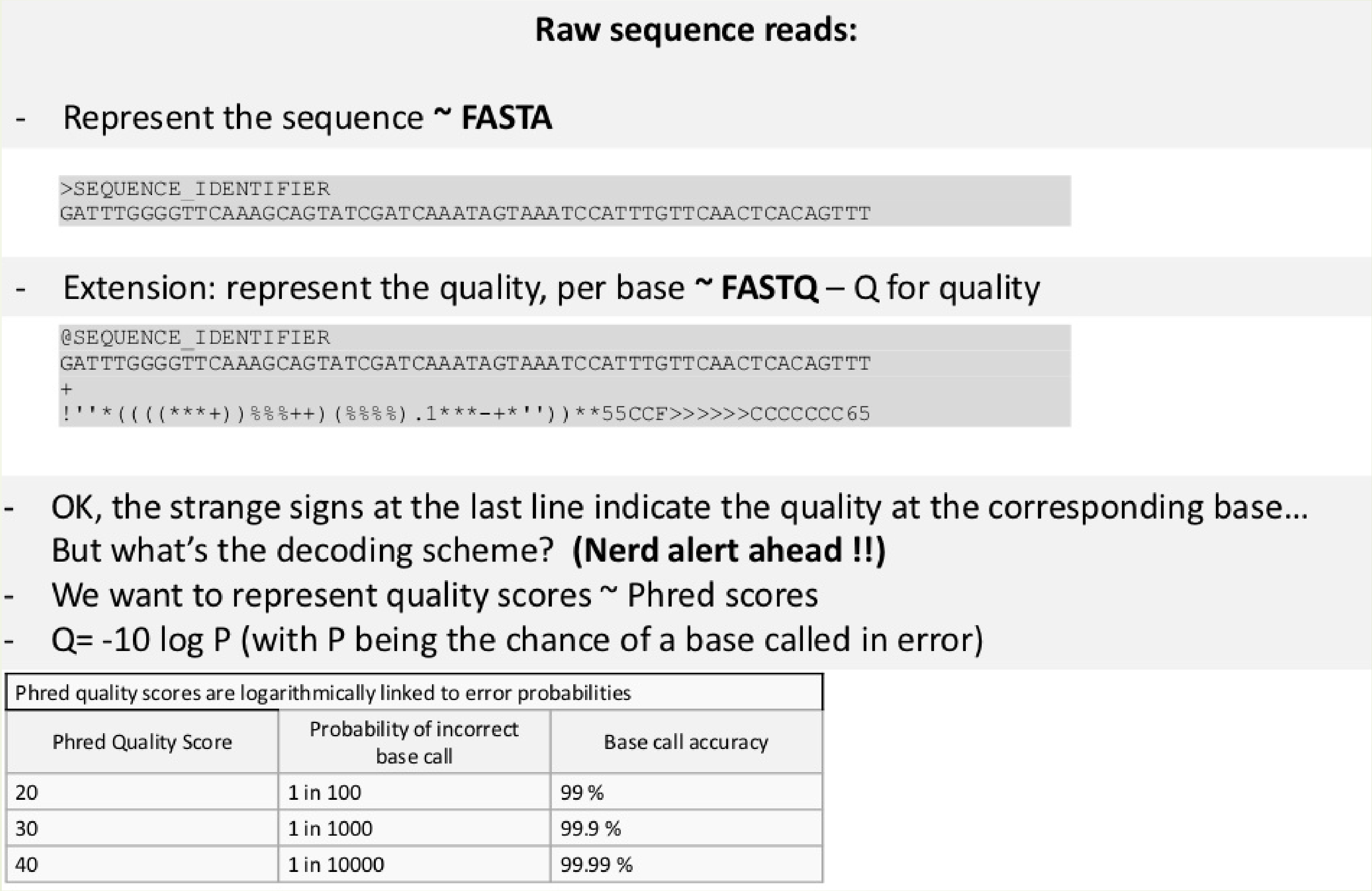
Why QC and preprocessing

- Sequencer output
  - Reads + **quality**
- Natural questions
  - Is the quality of my sequenced data ok?
  - If something is wrong, can I fix it?
- Problem: HUGE files



```
@HWI-EAS225:3:1:2:854#0/1
GGGGGAAGTCGGCAAAATAGATCCGTAACTTCGGG
+HWI-EAS225:3:1:2:854#0/1
a`abbbbabaabbababb^`[aaa`_N]b^ab^``a
@HWI-EAS225:3:1:2:1595#0/1
GGGAAGATCTCAAAAACAGAAGTAAAACATCGAACG
+HWI-EAS225:3:1:2:1595#0/1
a`abbbababbbabbbbbabb`aaababab\aa_`
```

Sequencing Data Formats



Raw sequence reads:

- Represent the sequence **~ FASTA**

```
>SEQUENCE_IDENTIFIER
GATTTGGGGTTCAAAGCAGTATCGATCAAATAGTAAATCCATTTGTTCAACTCACAGTTT
```

- Extension: represent the quality, per base **~ FASTQ** – Q for quality

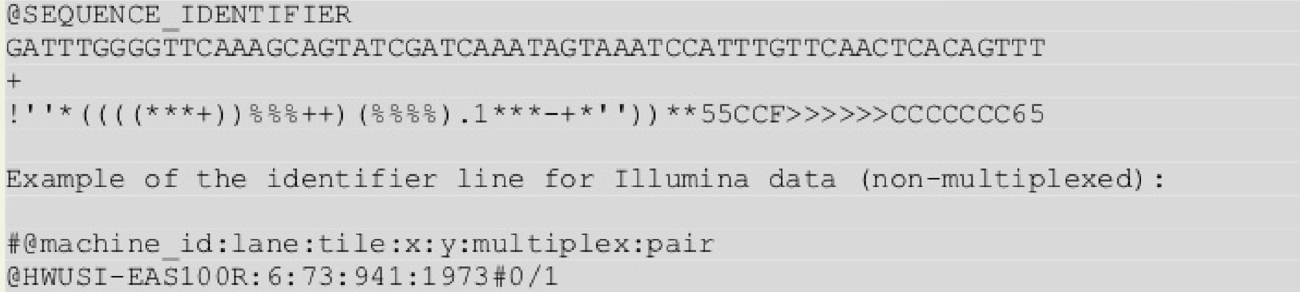
```
@SEQUENCE_IDENTIFIER
GATTTGGGGTTCAAAGCAGTATCGATCAAATAGTAAATCCATTTGTTCAACTCACAGTTT
+
!''*((((***+))%%%++)(%%%%).1***-+*''))**55CCF>>>>>>CCCCCCC65
```

- OK, the strange signs at the last line indicate the quality at the corresponding base… But what's the decoding scheme? **(Nerd alert ahead !!)**
- We want to represent quality scores ~ Phred scores
- Q= -10 log P (with P being the chance of a base called in error)

Phred quality scores are logarithmically linked to error probabilities

Phred Quality Score	Probability of incorrect base call	Base call accuracy
20	1 in 100	99 %
30	1 in 1000	99.9 %
40	1 in 10000	99.99 %

Quality before content



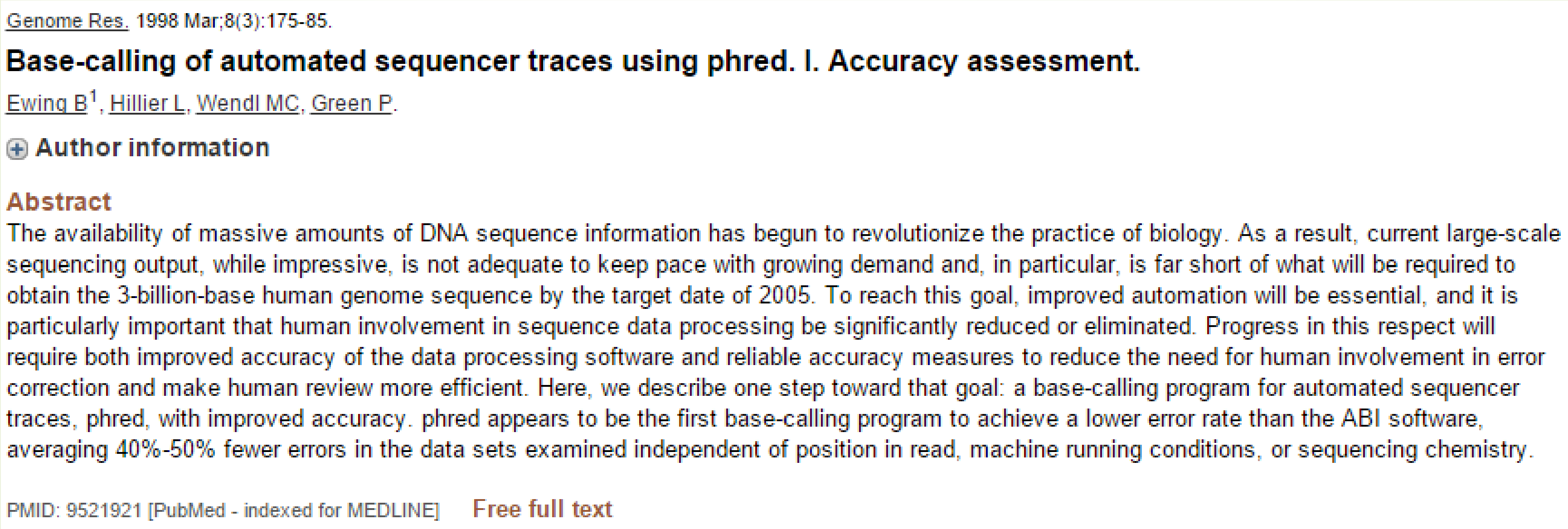
```
@SEQUENCE_IDENTIFIER
GATTTGGGGTTCAAAGCAGTATCGATCAAATAGTAAATCCATTTGTTCAACTCACAGTTT
+
!''*((((***+))%%%++)(%%%%).1***-+*''))**55CCF>>>>>>CCCCCCC65

Example of the identifier line for Illumina data (non-multiplexed):

#@machine_id:lane:tile:x:y:multiplex:pair
@HWUSI-EAS100R:6:73:941:1973#0/1
```

- Phred + 33 → Sanger
- Illumina 1.3 + → Phred +64
- Illumina 1.5 + → Phred +64
- Illumina 1.8 + → Phred +33
- Solid → Sanger

Check your instument + version → FastQC will give you a hint which scoring scheme is probably used

Extensions: FASTQ / FQ

What is quality?



Genome Res. 1998 Mar;8(3):175-85.

Base-calling of automated sequencer traces using phred. I. Accuracy assessment.

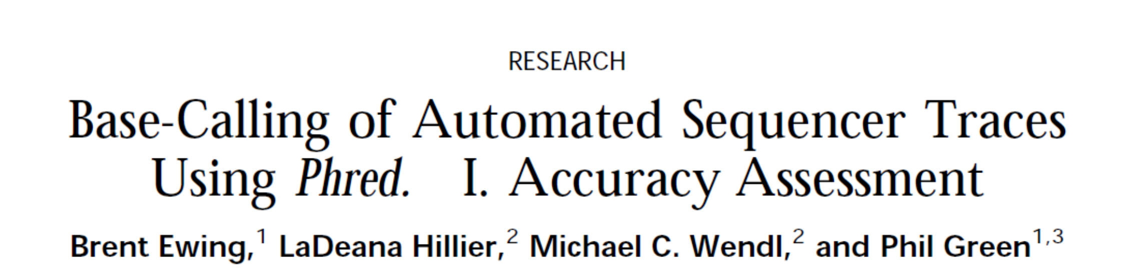
Ewing B[1], Hillier L, Wendl MC, Green P.

⊕ **Author information**

Abstract

The availability of massive amounts of DNA sequence information has begun to revolutionize the practice of biology. As a result, current large-scale sequencing output, while impressive, is not adequate to keep pace with growing demand and, in particular, is far short of what will be required to obtain the 3-billion-base human genome sequence by the target date of 2005. To reach this goal, improved automation will be essential, and it is particularly important that human involvement in sequence data processing be significantly reduced or eliminated. Progress in this respect will require both improved accuracy of the data processing software and reliable accuracy measures to reduce the need for human involvement in error correction and make human review more efficient. Here, we describe one step toward that goal: a base-calling program for automated sequencer traces, phred, with improved accuracy. phred appears to be the first base-calling program to achieve a lower error rate than the ABI software, averaging 40%-50% fewer errors in the data sets examined independent of position in read, machine running conditions, or sequencing chemistry.

PMID: 9521921 [PubMed - indexed for MEDLINE] Free full text

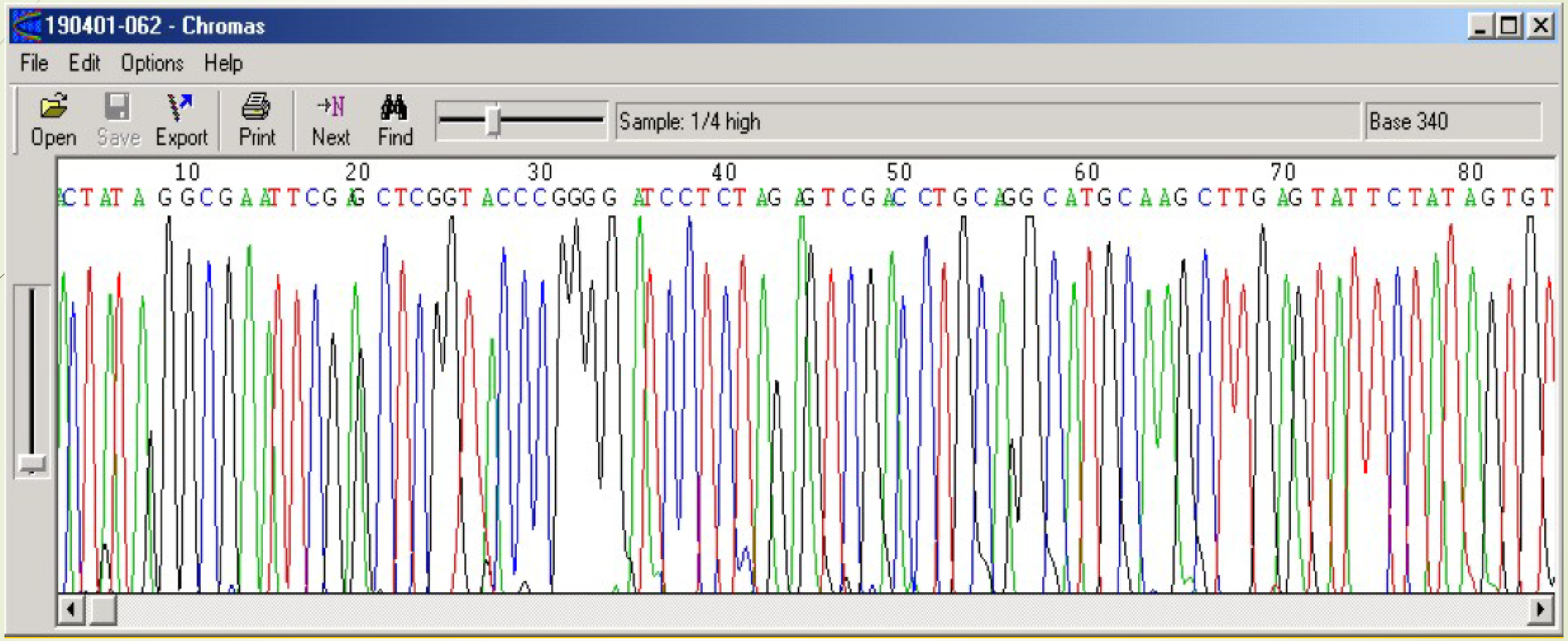


RESEARCH

Base-Calling of Automated Sequencer Traces Using *Phred.* I. Accuracy Assessment

Brent Ewing,[1] LaDeana Hillier,[2] Michael C. Wendl,[2] and Phil Green[1,3]

Trace File (high quality)

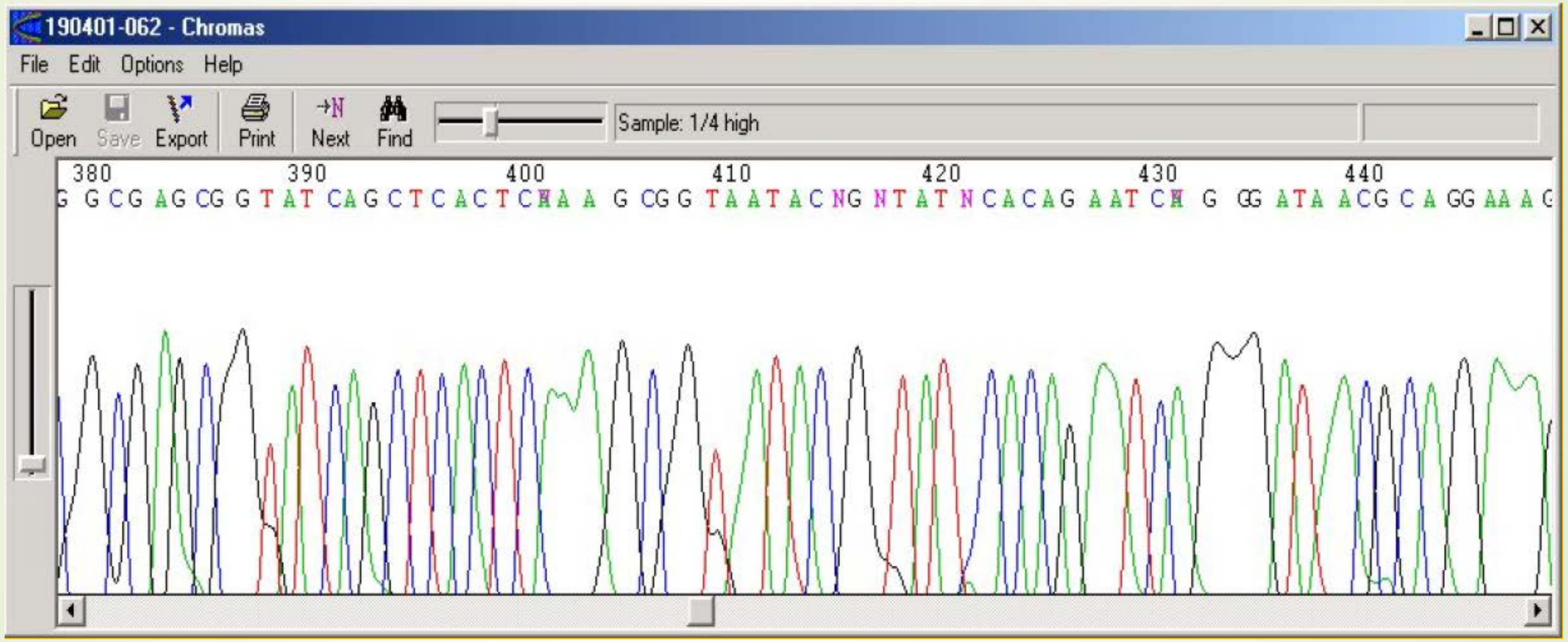


Sample: 1/4 high

Base 340

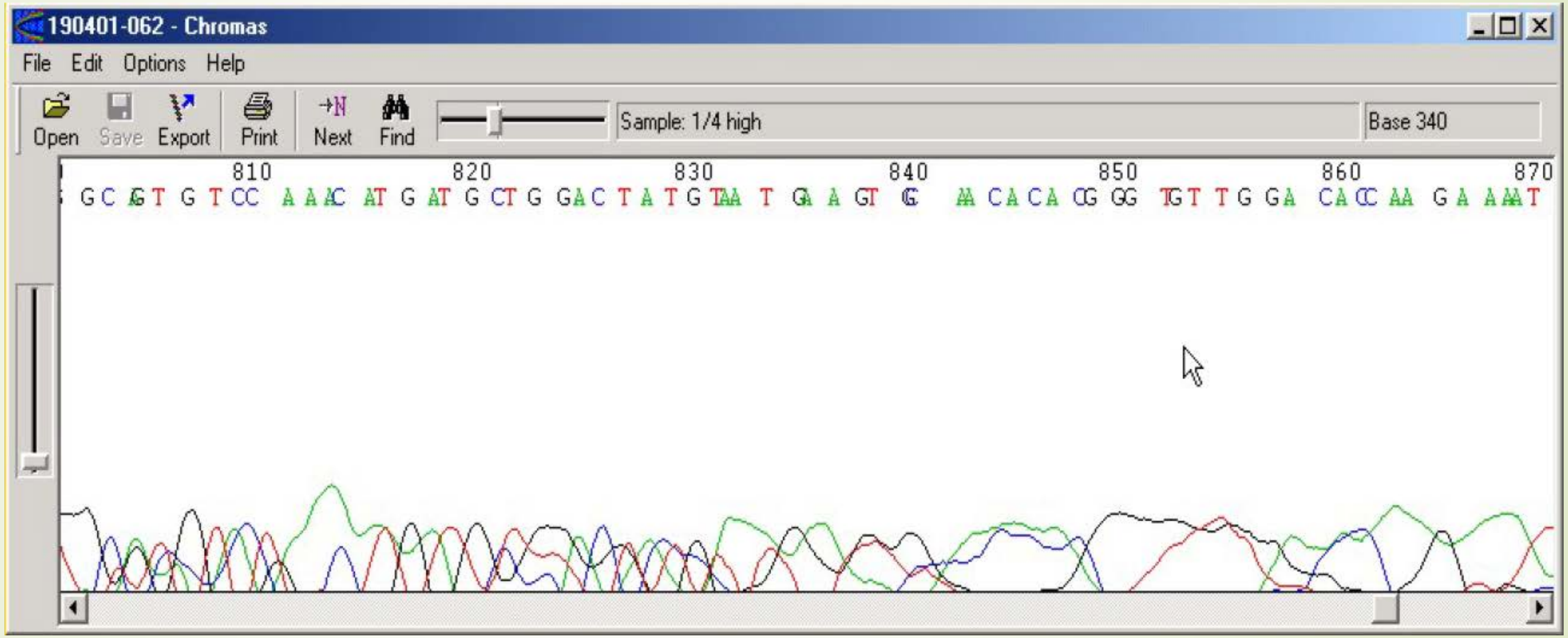
```
ACTATAGGCGAATTCGAGCTCGGTACCCGGGGATCCTCTAGAGTCGACCTGCAGGCATGCAAGCTTGAGTATTCTATAGTGT
```

Trace File (Medium Quality)



```
 380            390           400           410           420           430           440
G GCG AG CG G TAT CAGCTCACTCAA A G CGG TAATACNG NTATNCACAG AAT CA G GG ATA ACG CA GG AA AG
```

Trace File (Low Quality)

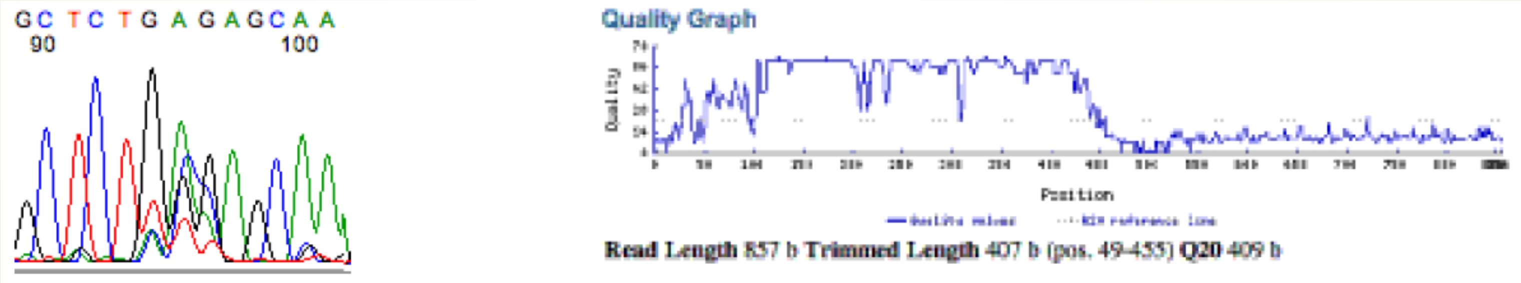

190401-062 - Chromas
File
Edit
Options
Help
Open
Save
Export
Print
Next
Find
810
820
830
840
850
860
870
G C AG T G T CC A A AC AT G AT G CT G GAC T A T G TAA T GA A GT CG AA C A C A CG GG TG T T G GA C A CC AA G A A AAT

Sample: 1/4 high

Base 340

Phred Quality Scores

- Phred is a program that assigns a quality score to each base in a sequence. These scores can then be used to trim bad data from the reads, and to determine how good an overlap actually is
- Phred scores are logarithmically related to the probability of an error:
  - a score of 10 means 10% error probability,
  - 20 means a 1% chance,
  - 30 means a 0.1 chance, etc
- A score of 30 is usually considered the minimum acceptable score.

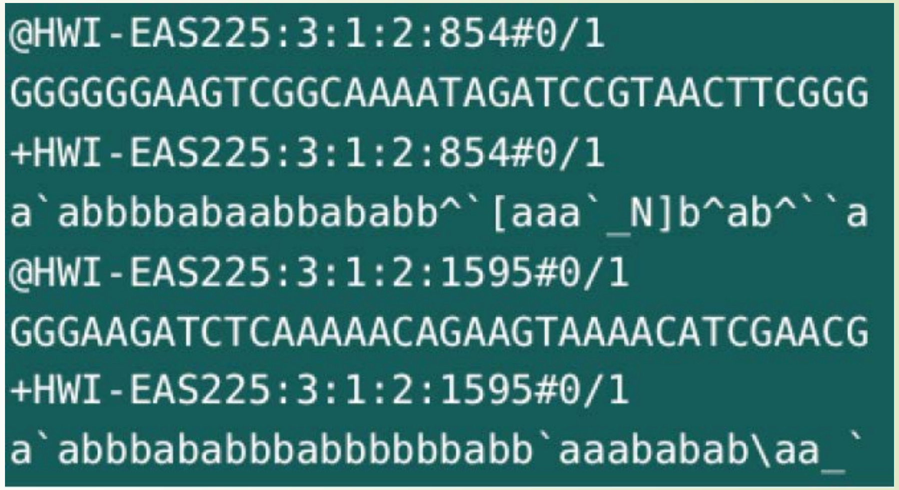


Read Length 857 b Trimmed Length 407 b (pos. 49-455) Q20 409 b

FASTQ File Format

- Each read is represented by four lines:

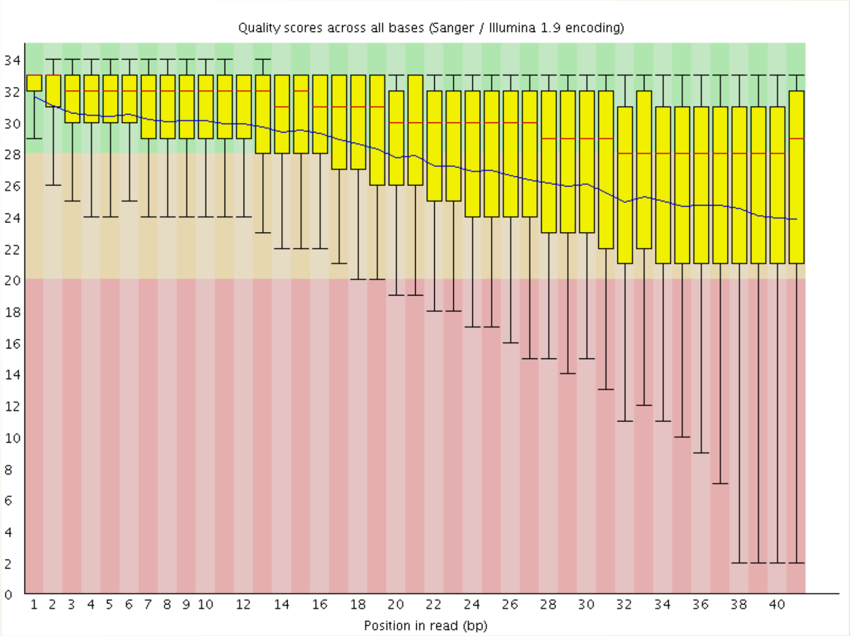
1. @ followed by read ID
2. Sequence
3. + optionally followed by repeated read ID
4. Quality line
   - Same length as sequence
   - Each character encodes the quality of the respective base



```
@HWI-EAS225:3:1:2:854#0/1
GGGGGAAGTCGGCAAAATAGATCCGTAACTTCGGG
+HWI-EAS225:3:1:2:854#0/1
a`abbbbabaabbababb^`[aaa`_N]b^ab^``a
@HWI-EAS225:3:1:2:1595#0/1
GGGAAGATCTCAAAAACAGAAGTAAAACATCGAACG
+HWI-EAS225:3:1:2:1595#0/1
a`abbbababbbabbbbbbabb`aaababab\aa_`
```

FASTQC

- As the name implies, FastQC is way to quickly see some summary statistics to check the quality of your NGS run.
  - It runs both as a GUI (requires Java) and as a command line program.
  - Provides several statistics:
    - Per Sequence Quality
    - Per sequence quality scores
    - Per base sequence and GC content
    - Per Sequence GC Content
    - etc..



Trimming

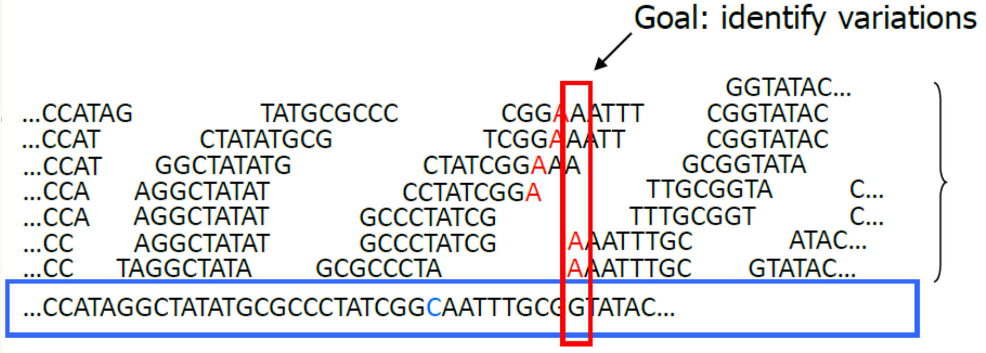
- Knowing quality → Act accordingly
- Adapter trimming
  - May increase mapping rates
  - Absolutely essential for small RNA
    Probably Improves de novo assemblies
- Quality trimming
  - May increase mapping rates
  - May also lead to loss of information
- Lots of software:
  - Cutadapt, Trim Galore!, PRINSEQ, etc.

Mapped Reads

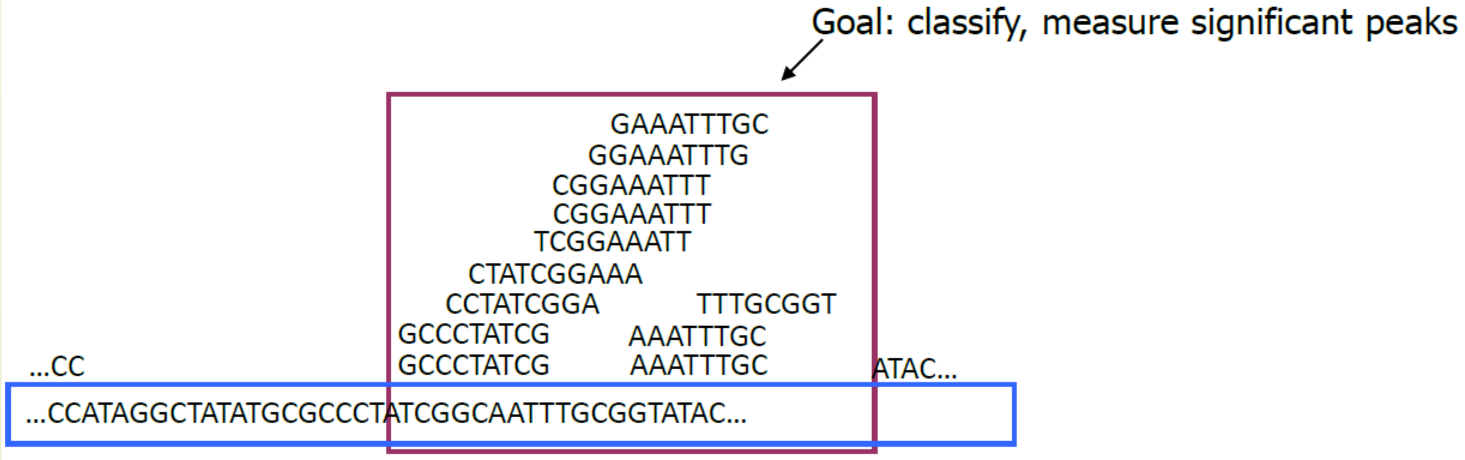
- Mapping: “align” these raw reads to a reference genome
  - Single-end or paired-end data?
  - How would you align a short read to the reference?
- Old-school: Smith-Waterman, BLAST, BLAT,…
- Now: mapping tools for short reads that use intelligent indexing and allow mismatches

Short read applications

- Genotypin

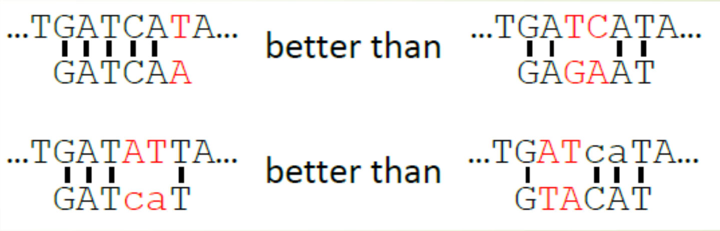

Goal: identify variations
...CCATAG TATGCGCCC CGGAAATTT GGTATAC...
...CCAT CTATATGCG TCGGAAATT CGGTATAC
...CCAT GGCTATATG CTATCGGAAA CGGTATAC
...CCA AGGCTATAT CCTATCGGA GCGGTATA
...CCA AGGCTATAT GCCCTATCG TTGCGGTA C...
...CC AGGCTATAT GCCCTATCG AAATTTGC TTTGCGGT C...
...CC TAGGCTATA GCGCCCTA AAATTTGC ATAC...
GTATAC...
...CCATAGGCTATATGCGCCCTATCGGCAATTTGCGGTATAC...

- RNA-Seq, ChIP-Seq, Methyl-Seq,...


Goal: classify, measure significant peaks

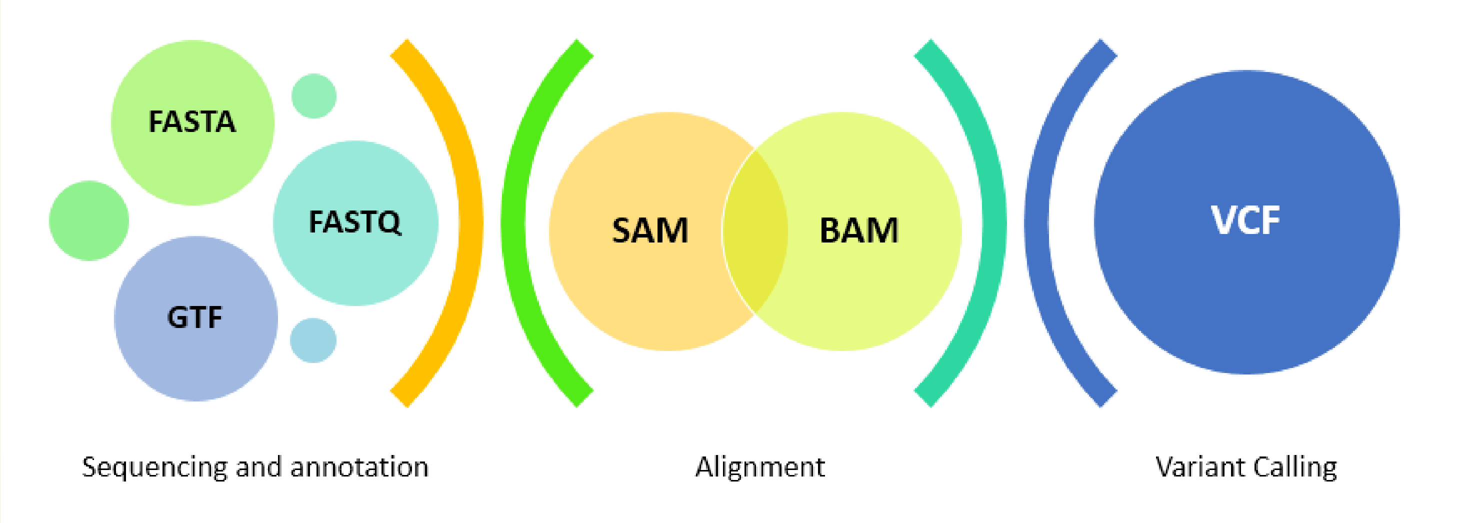
Defining the question

- Given a reference and a set of reads, report at least one “good” local alignment for each read, if one exists
  - Approximate answer to question: **where** in genome did read originate
- What is “good”? For now we concentrate on:
- Fewer mismatches = better
- Failing to align a low-quality base is better than failing to align a high-quality base



Interlude

(not only) NGS File Formats

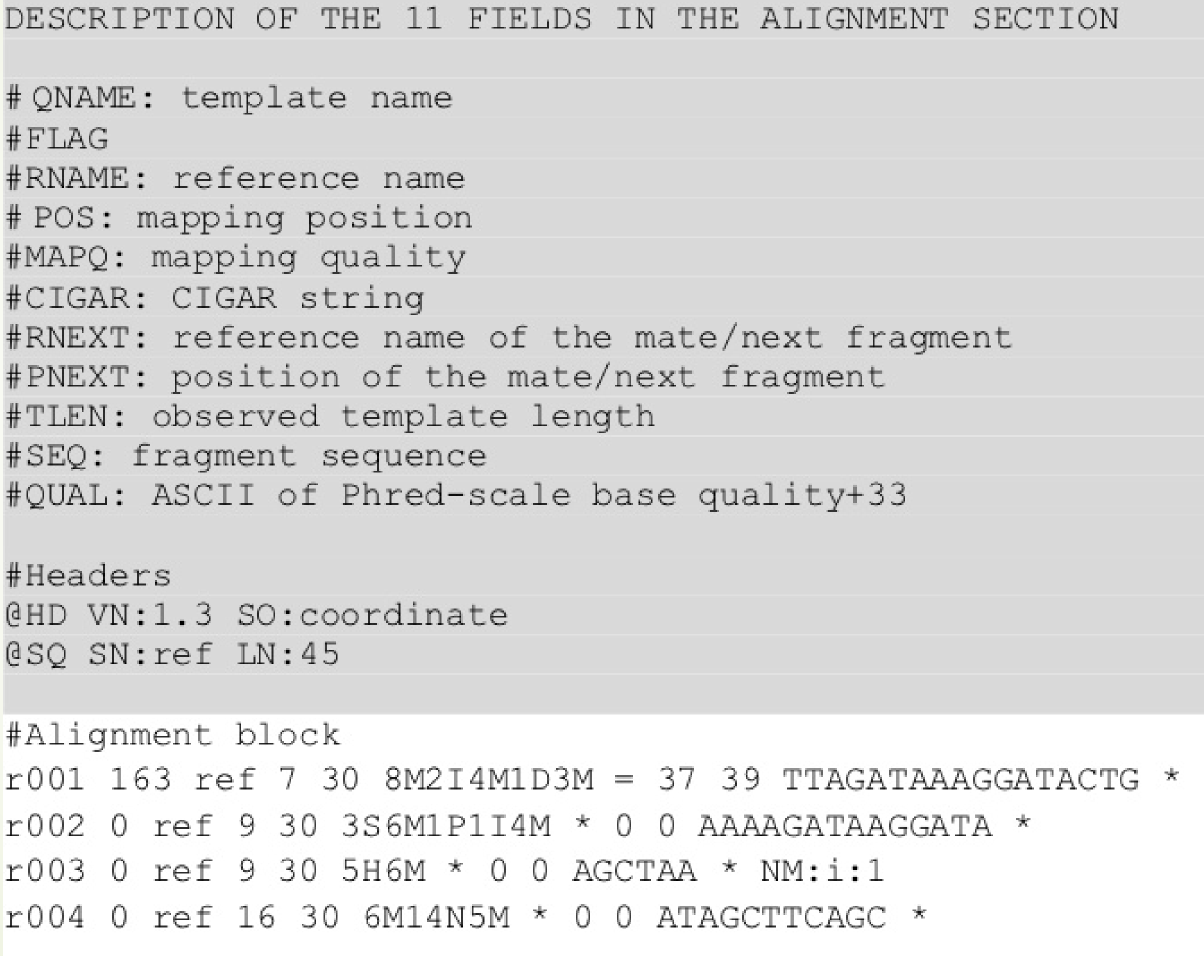


The **S**equence **A**lignment/**M**ap Format

- Generic alignment format
- Supports short and long reads
- Supports different sequencing platforms
- Flexible in style, compact in size, computationally efficient to access

- SAM File Format
  - BAM is the binary version of the SAM file; not human readable but indexed for fast access for other tools / visualization / …

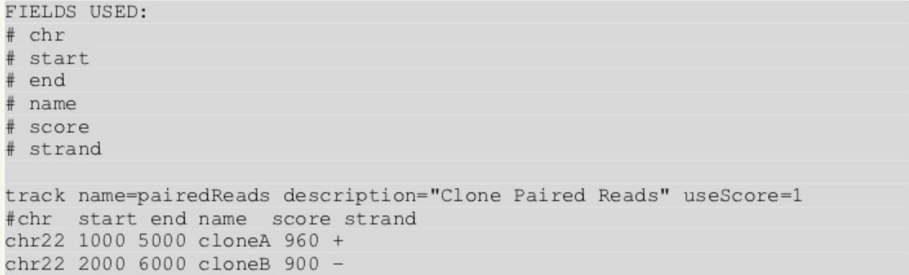
SAM Fields



```
DESCRIPTION OF THE 11 FIELDS IN THE ALIGNMENT SECTION

#QNAME: template name
#FLAG
#RNAME: reference name
#POS: mapping position
#MAPQ: mapping quality
#CIGAR: CIGAR string
#RNEXT: reference name of the mate/next fragment
#PNEXT: position of the mate/next fragment
#TLEN: observed template length
#SEQ: fragment sequence
#QUAL: ASCII of Phred-scale base quality+33

#Headers
@HD VN:1.3 SO:coordinate
@SQ SN:ref LN:45

#Alignment block
r001 163 ref 7 30 8M2I4M1D3M = 37 39 TTAGATAAAGGATACTG *
r002 0 ref 9 30 3S6M1P1I4M * 0 0 AAAAGATAAGGATA *
r003 0 ref 9 30 5H6M * 0 0 AGCTAA * NM:i:1
r004 0 ref 16 30 6M14N5M * 0 0 ATAGCTTCAGC *
```

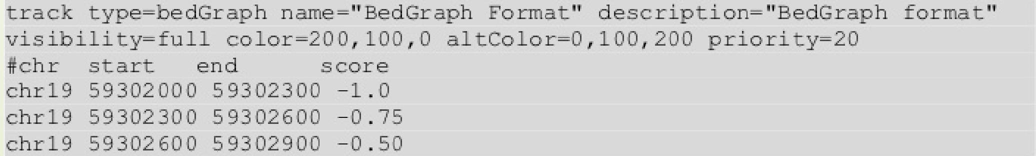
Other useful formats in NGS

- **B**rowser **E**xtensible **D**ata (location / annotation / scores).
  - used for mapping / annotation / peak locations
  - extension: bigBED (binary)



```
FIELDS USED:
# chr
# start
# end
# name
# score
# strand

track name=pairedReads description="Clone Paired Reads" useScore=1
#chr  start end name  score strand
chr22 1000 5000 cloneA 960 +
chr22 2000 6000 cloneB 900 -
```

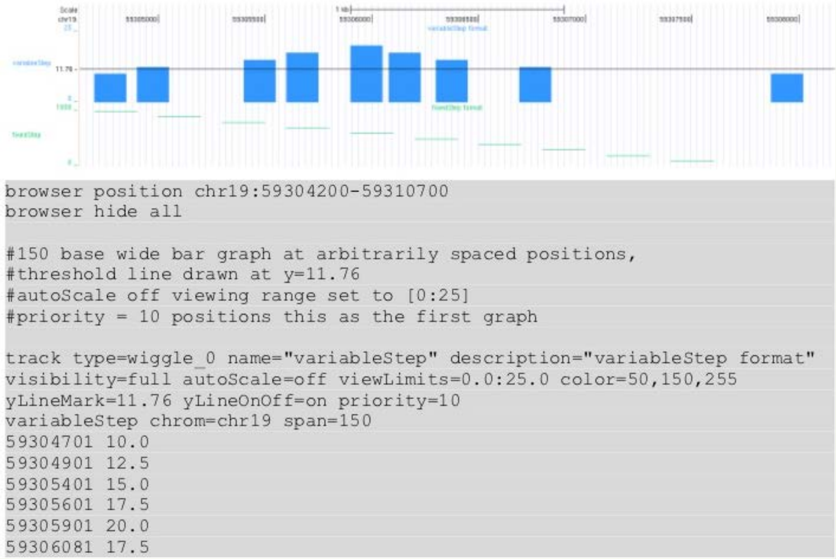
- BEDGraph files (location, combined with score)
  - used to represent peak scores



```
track type=bedGraph name="BedGraph Format" description="BedGraph format"
visibility=full color=200,100,0 altColor=0,100,200 priority=20
#chr  start    end      score
chr19 59302000 59302300 -1.0
chr19 59302300 59302600 -0.75
chr19 59302600 59302900 -0.50
```

Other useful formats in NGS

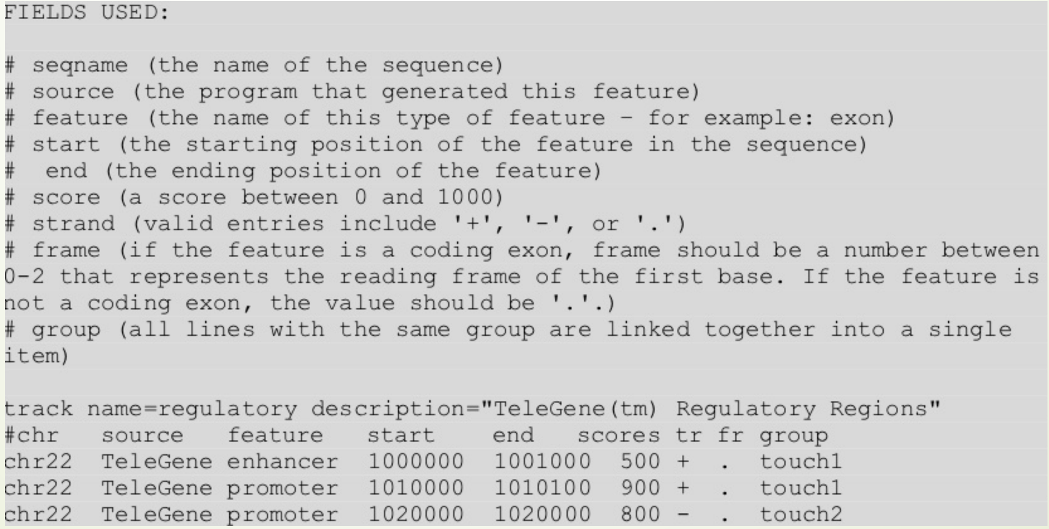
- WIG files (location / annotation / scores): wiggle
  - used for visualization or to summarize data, in most cases count data or normalized count data (RPKM)
  - extension: BigWig – binary versions, often used in GEO for ChIP-seq peaks



```
browser position chr19:59304200-59310700
browser hide all

#150 base wide bar graph at arbitrarily spaced positions,
#threshold line drawn at y=11.76
#autoScale off viewing range set to [0:25]
#priority = 10 positions this as the first graph

track type=wiggle_0 name="variableStep" description="variableStep format"
visibility=full autoScale=off viewLimits=0.0:25.0 color=50,150,255
yLineMark=11.76 yLineOnOff=on priority=10
variableStep chrom=chr19 span=150
59304701 10.0
59304901 12.5
59305401 15.0
59305601 17.5
59305901 20.0
59306081 17.5
```

Other useful formats in NGS

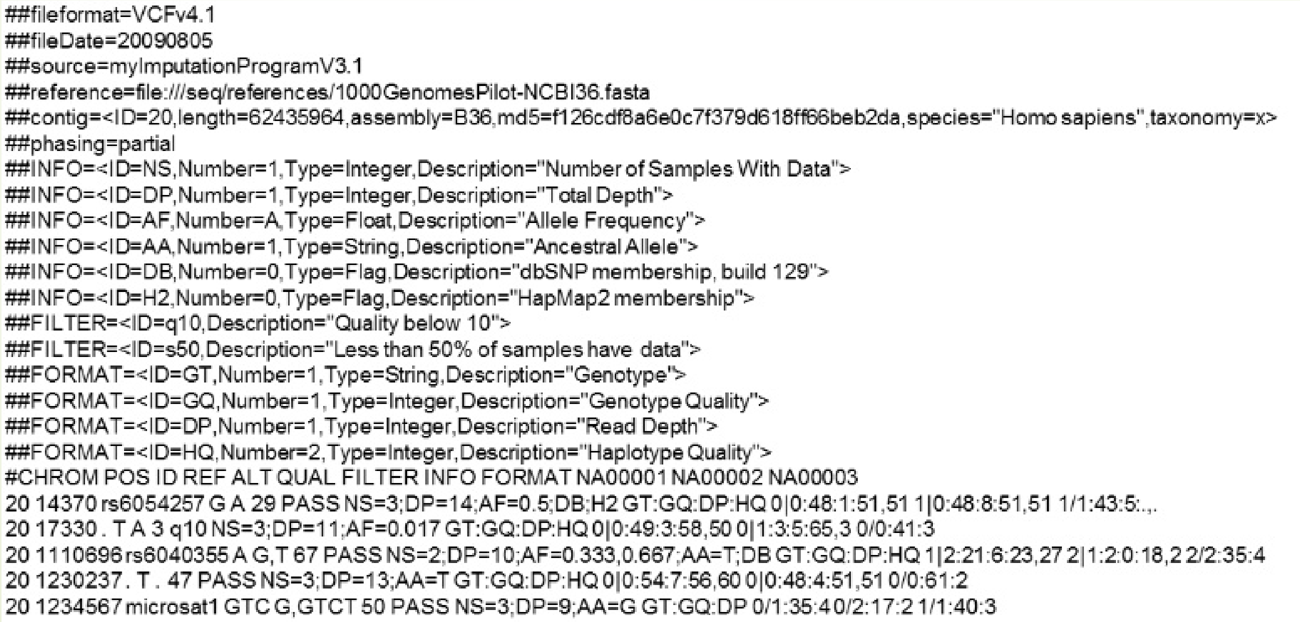
- **G**eneral **F**eature **F**ormat
  - used for annotation of genetic / genomic features, such as all coding genes in Ensembl
  - often used in downstream analysis to assign annotation to regions/peaks/....



```
FIELDS USED:

# seqname (the name of the sequence)
# source (the program that generated this feature)
# feature (the name of this type of feature - for example: exon)
# start (the starting position of the feature in the sequence)
#  end (the ending position of the feature)
# score (a score between 0 and 1000)
# strand (valid entries include '+', '-', or '.')
# frame (if the feature is a coding exon, frame should be a number between
0-2 that represents the reading frame of the first base. If the feature is
not a coding exon, the value should be '.'.)
# group (all lines with the same group are linked together into a single
item)

track name=regulatory description="TeleGene(tm) Regulatory Regions"
#chr   source   feature    start    end     scores tr fr group
chr22  TeleGene enhancer  1000000  1001000  500 +  .  touch1
chr22  TeleGene promoter  1010000  1010100  900 +  .  touch1
chr22  TeleGene promoter  1020000  1020000  800 -  .  touch2
```

Other useful formats in NGS

- **V**ariant **C**all **F**ormat
  - used for SNP representation



```
##fileformat=VCFv4.1
##fileDate=20090805
##source=myImputationProgramV3.1
##reference=file:///seq/references/1000GenomesPilot-NCBI36.fasta
##contig=<ID=20,length=62435964,assembly=B36,md5=f126cdf8a6e0c7f379d618ff66beb2da,species="Homo sapiens",taxonomy=x>
##phasing=partial
##INFO=<ID=NS,Number=1,Type=Integer,Description="Number of Samples With Data">
##INFO=<ID=DP,Number=1,Type=Integer,Description="Total Depth">
##INFO=<ID=AF,Number=A,Type=Float,Description="Allele Frequency">
##INFO=<ID=AA,Number=1,Type=String,Description="Ancestral Allele">
##INFO=<ID=DB,Number=0,Type=Flag,Description="dbSNP membership, build 129">
##INFO=<ID=H2,Number=0,Type=Flag,Description="HapMap2 membership">
##FILTER=<ID=q10,Description="Quality below 10">
##FILTER=<ID=s50,Description="Less than 50% of samples have data">
##FORMAT=<ID=GT,Number=1,Type=String,Description="Genotype">
##FORMAT=<ID=GQ,Number=1,Type=Integer,Description="Genotype Quality">
##FORMAT=<ID=DP,Number=1,Type=Integer,Description="Read Depth">
##FORMAT=<ID=HQ,Number=2,Type=Integer,Description="Haplotype Quality">
#CHROM POS ID REF ALT QUAL FILTER INFO FORMAT NA00001 NA00002 NA00003
20 14370 rs6054257 G A 29 PASS NS=3;DP=14;AF=0.5;DB;H2 GT:GQ:DP:HQ 0|0:48:1:51,51 1|0:48:8:51,51 1/1:43:5:.,.
20 17330 . T A 3 q10 NS=3;DP=11;AF=0.017 GT:GQ:DP:HQ 0|0:49:3:58,50 0|1:3:5:65,3 0/0:41:3
20 1110696 rs6040355 A G,T 67 PASS NS=2;DP=10;AF=0.333,0.667;AA=T;DB GT:GQ:DP:HQ 1|2:21:6:23,27 2|1:2:0:18,2 2/2:35:4
20 1230237 . T . 47 PASS NS=3;DP=13;AA=T GT:GQ:DP:HQ 0|0:54:7:56,60 0|0:48:4:51,51 0/0:61:2
20 1234567 microsat1 GTC G,GTCT 50 PASS NS=3;DP=9;AA=G GT:GQ:DP 0/1:35:4 0/2:17:2 1/1:40:3
```

aaaand back to the story

Mappers

- BowTie2 is the most commonly used aligner
  - Employs an indexing algorithm that can trade flexibility between memory usage and running time
- BWA (mem / aln) is an efficient mapper that is extensively used in RNA-Seq
- STAR aligner, is an general, all-purpose aligner

HiSat2

- Stands for:
  - hierarchical indexing for spliced alignment of transcripts
- **HISAT2** is a fast and sensitive alignment program for mapping next-generation sequencing reads (both DNA and RNA) to a population of human genomes (as well as to a single reference genome).
- HISAT2 searches for up to N distinct, primary alignments for each read
  - Very fast
  - Low memory requirements

We've aligned the data. Then what?

- Depending on the target study.

Gene	Treatment 1			Treatment 2		
1	14	18	10	47	13	24
2	10	3	15	1	11	5
3	1	0	10	80	21	34
4	0	0	0	0	2	0
5	4	3	3	5	33	29
.	.	.	.	.	.	.
.	.	.	.	.	.	.
.	.	.	.	.	.	.
53256	47	29	11	71	278	339
Total	22,910,173	30,701,031	18,897,029	20,546,299	28,491,272	27,082,148

Differential Expression

- To determine if gene 1 is DE, we would like to know whether the proportion of reads aligning to gene 1 tends to be different for experimental units that received treatment 1 than for experimental units that received treatment 2

14 out of 22,910,173		47 out of 20,546,299
18 out of 30,701,031	vs.	13 out of 28,491,272
10 out of 18,897,029		24 out of 27,082,148

How about we try these now?