The Biostar Handbook: A Beginner's Guide to Bioinformatics[2]

#course

1. **Data format(2)**

GenBank, FASTA represent **curated sequence information**. FASTQ represents **obtained data** via sequencing instrumentation.

- **Genbank:**

```
LOCUS       SCU49845     5028 bp    DNA             PLN       21-JUN-1999
DEFINITION  Saccharomyces cerevisiae TCP1-beta gene, partial cds, and Axl2p
            (AXL2) and Rev7p (REV7) genes, complete cds.
ACCESSION   U49845
VERSION     U49845.1  GI:1293613
KEYWORDS    .
SOURCE      Saccharomyces cerevisiae (baker's yeast)
  ORGANISM  Saccharomyces cerevisiae
            Eukaryota; Fungi; Ascomycota; Saccharomycotina; Saccharomycetes;
            Saccharomycetales; Saccharomycetaceae; Saccharomyces.
REFERENCE   1  (bases 1 to 5028)
  AUTHORS   Torpey,L.E., Gibbs,P.E., Nelson,J. and Lawrence,C.W.
  TITLE     Cloning and sequence of REV7, a gene whose function is required for
            DNA damage-induced mutagenesis in Saccharomyces cerevisiae
  JOURNAL   Yeast 10 (11), 1503-1509 (1994)
  PUBMED    7871890
REFERENCE   2  (bases 1 to 5028)
  AUTHORS   Roemer,T., Madden,K., Chang,J. and Snyder,M.
  TITLE     Selection of axial growth sites in yeast requires Axl2p, a novel
            plasma membrane glycoprotein
  JOURNAL   Genes Dev. 10 (7), 777-793 (1996)
  PUBMED    8846915
REFERENCE   3  (bases 1 to 5028)
  AUTHORS   Roemer,T.
  TITLE     Direct Submission
  JOURNAL   Submitted (22-FEB-1996) Terry Roemer, Biology, Yale University, New
            Haven, CT, USA
FEATURES             Location/Qualifiers
     source          1..5028
                     /organism="Saccharomyces cerevisiae"
                     /db_xref="taxon:4932"
                     /chromosome="IX"
                     /map="9"
     CDS             <1..206
```

" The format has a so-called fixed-width format, where the first 10 characters form a column that serves as an identifier and the rest of the lines are information corresponding to that identifier."

the 1st column is formed by 10 characters
GenBank格式的优势在于可读，但是对于数据分析而言不太合适

"We typically convert a GenBank file to some other, simpler format to work with it.

ReadSeq is a useful conversion tool to use for the support."

ReadSeq 转换格式

"The NCBI Reference Sequence (RefSeq) project provides sequence records and related information for numerous organisms, and provides a baseline for medical, functional, and comparative studies."

Accession prefix	Molecule type	Comment
AC_	Genomic	Complete genomic molecule, usually alternate assembly
NC_	Genomic	Complete genomic molecule, usually reference assembly
NG_	Genomic	Incomplete genomic region
NT_	Genomic	Contig or scaffold, clone-based or WGS[a]
NW_	Genomic	Contig or scaffold, primarily WGS[a]
NS_	Genomic	Environmental sequence
NZ_[b]	Genomic	Unfinished WGS
NM_	mRNA	
NR_	RNA	
XM_[c]	mRNA	Predicted model
XR_[c]	RNA	Predicted model
AP_	Protein	Annotated on AC_ alternate assembly
NP_	Protein	Associated with an NM_ or NC_ accession
YP_[c]	Protein	
XP_[c]	Protein	Predicted model, associated with an XM_ accession
ZP_[c]	Protein	Predicted model, annotated on NZ_ genomic records

how are RefSeq sequences named

RefSeq: NCBI Reference Sequence 提供大多数生物的序列记录和相应的信息，为医学，功能和比较分析提供了基准线

- **FASTA:**

Type	Format(s) [1]	Example(s)
local	lcl|integer lcl|string	lcl|123 lcl|hmm271
GenInfo backbone seqid	bbs|integer	bbs|123
GenInfo backbone moltype	bbm|integer	bbm|123
GenInfo import ID	gim|integer	gim|123
GenBank	gb|accession|locus	gb|M73307|AGMA13GT
EMBL	emb|accession|locus	emb|CAM43271.1|
PIR	pir|accession|name	pir||G36364
SWISS-PROT	sp|accession|name	sp|P01013|OVAX_CHICK
patent	pat|country|patent|sequence	pat|US|RE33188|1
pre-grant patent	pgp|country|application-number|seq-number	pgp|EP|0238993|7
RefSeq [2]	ref|accession|name	ref|NM_010450.1|
general database reference	gnl|database|integer gnl|database|string	gnl|taxon|9606 gnl|PID|e1632
GenInfo integrated database	gi|integer	gi|21434723
DDBJ	dbj|accession|locus	dbj|BAC85684.1|
PRF	prf|accession|name	prf||0806162C
PDB	pdb|entry|chain	pdb|1I4L|D
third-party GenBank	tpg|accession|name	tpg|BK003456|
third-party EMBL	tpe|accession|name	tpe|BN000123|
third-party DDBJ	tpd|accession|name	tpd|FAA00017|

"Lower-case letters may be used to indicate repetitive regions for the genome"

小写字母可能表示重复区，但是目前还是很难判断哪些是重复区。

注lastz默认过滤小写字母（重复区）

注:

1. 序列行不应太长
2. 不同工具对序列行中的超出字符集（核酸ATCG或蛋白质20种)处理不同
3. 序列行如果有多行，除最后一行，前几行应该是等宽的
4. 使用**大写字母**。不同工具会有大小写敏感。如有些工具会认为小写字母是非重复，大写字母象征着重复区域。
5. 不同机构对 > 后面的结构有各自的定义

- **FASTQ:**

"It may be thought of as a variant of the FASTA format that allows it to associate a quality measure to each sequence base, FASTA with QUALITIES"

FASTQ=FASTA+QUALITIES

"the format is very similar to FASTA but suffers from even more flaws than the FASTA

format."

"the @ sign is both a FASTQ record separator and a valid value of the quality string. For that reason it is a little more difficult to design a correct FASTQ parsing program."

"FASTQ format is not required to be line-oriented!"

"each character!''*((((represents a numerical value: a so-called Phred scorePhred quality score - Wikipedia, encoded via a single letter encoding."

Q	Error	Accuracy
0	1 in 1	0%
10	1 in 10	90%
20	1 in 100	99%
30	1 in 1000	99.9%
40	1 in 10,000	99.99%

"When our FASTQ quality contains lower case letters abcdefghi it means that your data is in this older format. Some tools will allow you to set parameters to account for this different encoding."

注:

1. 目前有许多中FASTQ质量编码版本：有+33, +64两类。以!为0的是，+33，以@开头的是+64
2. FASTQ开头的定义也有多个版本

proj 2019-3-29 SeqKit

seqkit

2. **Downloading SRA**

"Entrez is NCBI's primary text search and retrieval system that integrates the PubMed database of biomedical literature with 39 other literature and molecular databases including DNA and protein sequence, structure, gene, genome, genetic variation, and gene expression."

Entrez Direct: A system of NCBI integrates databases

Entrez Direct: E-utilities on the UNIX Command Line - Entrez Programming Utilities Help -

NCBI Bookshelf

entrez

“An accession number”

1. accession number is unique;
2. version number is also unique;
3. when searching accession number, we get the page of the latest version

“NCBI BioProject: PRJN” 研究项目
“NCBI BioSample: SAMN” 材料来源
“SRA Experiment: SRX” 特定样本的单独的测序文库
“SRA Run: SRR or ERR” 存放数据

“paired end reads the data needs to be separated into different files”

```
fastq-dump --split-files SRR1553607
```

3. **Sequencing instrument**

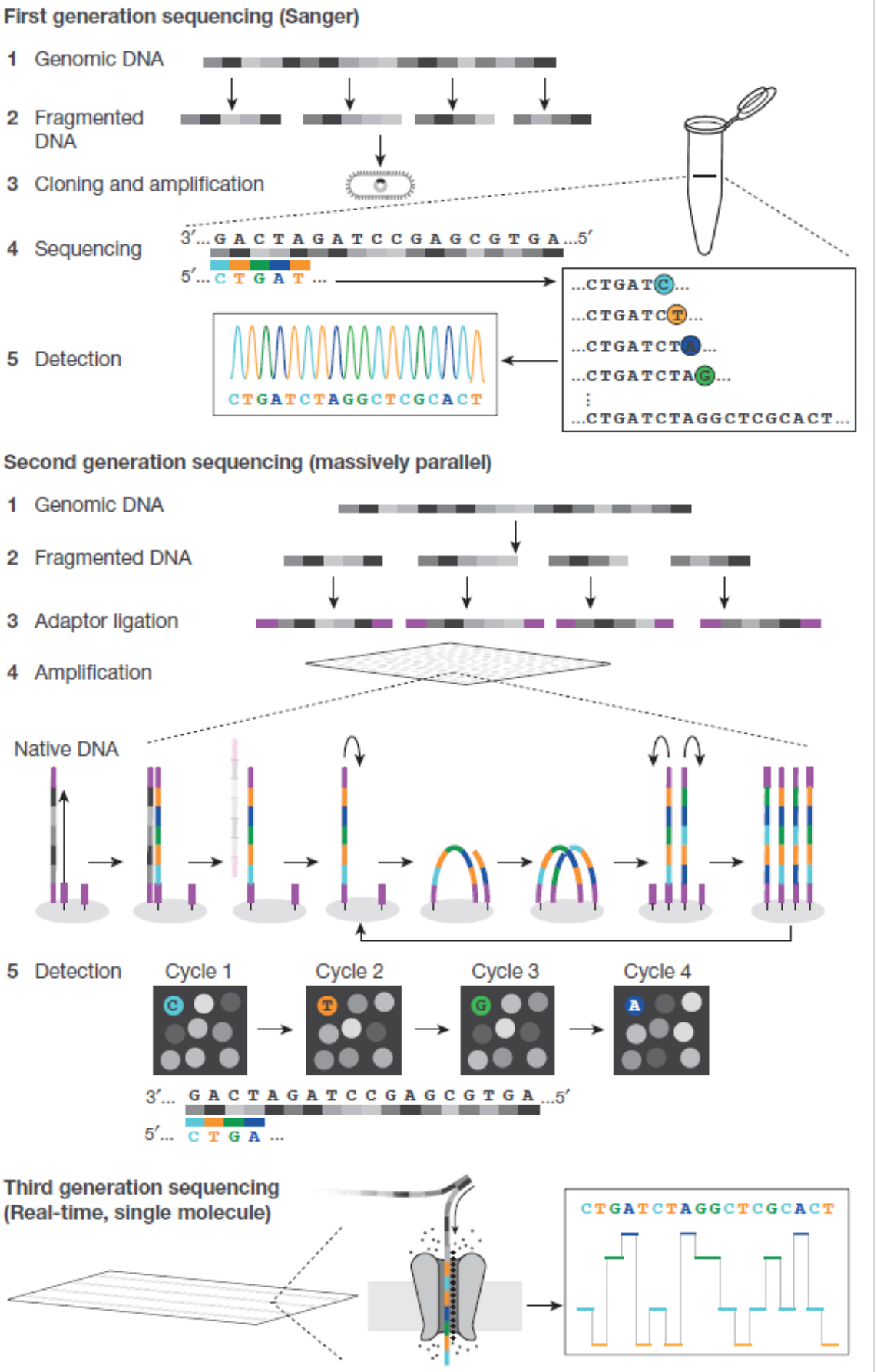

Figure 1 | DNA sequencing technologies. Schematic examples of first, second and third generation sequencing are shown. Second generation sequencing is also referred to as next-generation sequencing (NGS) in the text.

Field guide to next-generation DNA sequencers - GLENN - 2011 - Molecular Ecology Resources - Wiley Online Library
2016 NGS Field Guide: Overview |

“SOLiD and 454”

2005年之后出现了第二代测序，SOLiD 和 454 已经discontinued，只有Illumina一家独大

“Typically the longer read instruments operate at substantially higher error rates than those producing short reads.”

在测序过程中，机器会对每次读取的结果赋予一个值，用于表明它有多大把握结果是对的。从理论上都是前面质量好，后面质量差。并且在某些GC比例高的区域，测序质量会大幅度降低。

“Errors in long reads are easier to correct and account for than errors in short reads.”

“The read may be shorter or even longer than the original fragment. This because the original DNA fragment (some may also call this “template”) has been altered by adding sequencing adapters to it.”

因为测序过程是边合成边测序（SBS），所以在建库的时候，短序列两端会加一些固定的碱基用于桥式PCR扩增，这些固定的碱基就是adapter。一般而言，还可以在接头加一些tag（index），用于标识这个read来自于哪个物种

“The sequencer will typically recognize the XXXX at the beginning and will not report it.”

“the fragment may occasionally be shorter than the measurement.” **“read-through”**

In the situation that adapter is included in the read, which is longer than the fragment

“A simple post processing operation can be used to reverse complement the reads into the most commonly used ––> <–– orientation.”

因此请确保你的双端数据的方向是––> <––

“**Paired-end (PE) sequencing**”

“a method to sequence both ends of a fragment and to make the pairing information available in the data.”

“The two reads are typically stored in separate FASTQ files and are synchronized by name and order”

“Each read in file 1 has a corresponding entry in file 2”

“**Mate-pair sequencing** typically has the same goals as the paired-end approach – it attempts to measure two ends of a fragment.”

“mate-pair DNA fragments are much much longer than PE methods and the read orientations are usually different”

"Mate-paired data is only supported by specialized software."

4. **Coverage**

"C = number of sequenced bases / total genome size"

"10x indicates that on average, each base of the genome will be covered by 10 measurements."

C: genome coverage

each base of the genome will be covered by C measurements

覆盖度不是意味着所有基因组都被覆盖了，而是覆盖率越高，基因组未被检测到的基因越少

"For a sequencing instrument that produces variable reads lengths we have:

C = sum(Li) / G "

"For an instrument with fixed read length L, the formula simplifies to:

C = N * L / G
"

"P = exp(-C)"

P * genome size = number of missing bases

P 碱基丢失率

"Increase your coverages well over the theoretical limits.
Sometime some parts of the genome do not show up in the data. These are called hard to sequence regions, but it is not always clear why they are hard to sequence.
What part of the genome is uniquely coverable with a given read size? Repeat-rich regions of the genome require longer reads to resolve.
Is a given read from one region or is it potentially sourced from multiple regions? We may resolve the source with a longer read."

当然理论覆盖度并不代表现实情况，由于基因组的复杂性，DNA可能也不是真的随机打断,甚至实验protocol还有一定的偏向性。

1. 尽可能增加测序深度
2. 尽管有一些基因组部分很难被测序，但是我们其实清楚这些区域难以测序的原因
3. 基因组的高度重复区域需要更长的读长才能被发现
4.基因组不同区域可能会产生相同的read，你需要更长的读长

“Scientists often use terms such as: **“accessible”, “mappable”, “effective”** genome to indicate that only some parts of the genome can be easily studied.”

5. **FastQC**

“FastQC generates its reports by evaluating a small subset of the data and extrapolating those findings to the entirety of the dataset.”

FastQC的工作原理是通过对总体数据的抽样来评估总体效果，这就是它快(fast)的愿意，毕竟其他一些质量展示软件是老老实实把所有数据都用于作图。

“Most of the time, these symbols are not meaningful.”

没必要太过在意“stoplinght”，但是如果全部红灯，那么数据就要小心了。

“10 corresponds to 10% error (1/10),
20 corresponds to 1% error (1/100),
30 corresponds to 0.1% error (1/1,000) and
40 corresponds to one error every 10,000 measurements (1/10,000) that is an error rate of 0.01%.”

“The yellow boxes contain 50% of the data, the whiskers indicate the 75% outliers.”

“For fixed read length instruments, like the Illumina sequencer, all read lengths are the same. For long read technologies like the PacBio and MinION, the distribution can be a lot more varied.”

“Evaluate (visualize) data quality.
Stop QC if the quality appears to be satisfactory.
If the quality is still inadequate, execute one or more data altering steps then go to step 1”

1.数据可视化评估
2.质量不错就停止QC
3.否则对数据进行修改，返回步骤1

“When the sequencing data looks really bad from the beginning, it is best to move on and collect new data.”

apply QC only when you have problems that need fixing.

“When assembling a de-novo genome, errors can derail the process; hence, it is more important to apply a higher stringency for filtering.”

De novo is more sensitive to data quality than genome sequencing analysis

“Data improvement via quality control is an incremental process with ever-diminishing returns. ”

通过质量控制改进数据是一个渐进过程，收益递减

“the truth of the matter is that objective assessment of changes in data quality is difficult. What you will often see and hear are subjective,”

1.首先QC工具本身质量就不是很好，QC工具之间可能也不一致，不同工具使用相同的参数可能也会有不同的结果。
2.QC的确可能会引入错误，所以尽量避免修改数据

Trimmomatic, BBDuk ,flexbar and cutadapt

- BBDuk part of the BBMap package
- BioPieces a suite of programs for sequence preprocessing
- CutAdapt application note in Embnet Journal, 2011
- fastq-mcf published in The Open Bioinformatics Journal, 2013
- Fastx Toolkit: collection of command line tools for Short-Reads FASTA/FASTQ files preprocessing - one of the first tools
- FlexBar, Flexible barcode and adapter removal published in Biology, 2012
- NGS Toolkit published in Plos One, 2012
- PrinSeq application note in Bioinformatics, 2011
- Scythe a bayesian adaptor trimmer

- SeqPrep - a tool for stripping adaptors and/or merging paired reads with overlap into single reads.
- Skewer: a fast and accurate adapter trimmer for next-generation sequencing paired-end reads.
- TagCleaner published in BMC Bioinformatics, 2010
- TagDust published in Bioinformatics, 2009
- Trim Galore - a wrapper tool around Cutadapt and FastQC to consistently apply quality and adapter trimming to FastQ files, with some extra functionality for MspI-digested RRBS-type (Reduced Representation Bisufite-Seq) libraries
- Trimmomatic application note in Nucleic Acid Research, 2012, web server issue

There also exist libraries via R (Bioconductor) for QC: PIQA , ShortRead