The Biostar Handbook: A Beginner's Guide to Bioinformatics[3]

#course

Sequence duplication
k-mers

Sequence duplication

> Duplicates hence fall into two classes:
>
> Natural duplicates - these were identical fragments present in the sample
> Artificial duplicates - these are produced artificially during the sequencing process, PCR amplification, detection errors

序列重复分为两类：天然重复（片段相同），人为重复（PCR扩增，检测）

> There are two main approaches:
>
> Sequence identity - where we find and remove sequences that are have the exact sequence.
> Alignment identity - where we find and remove sequences that align the same way.

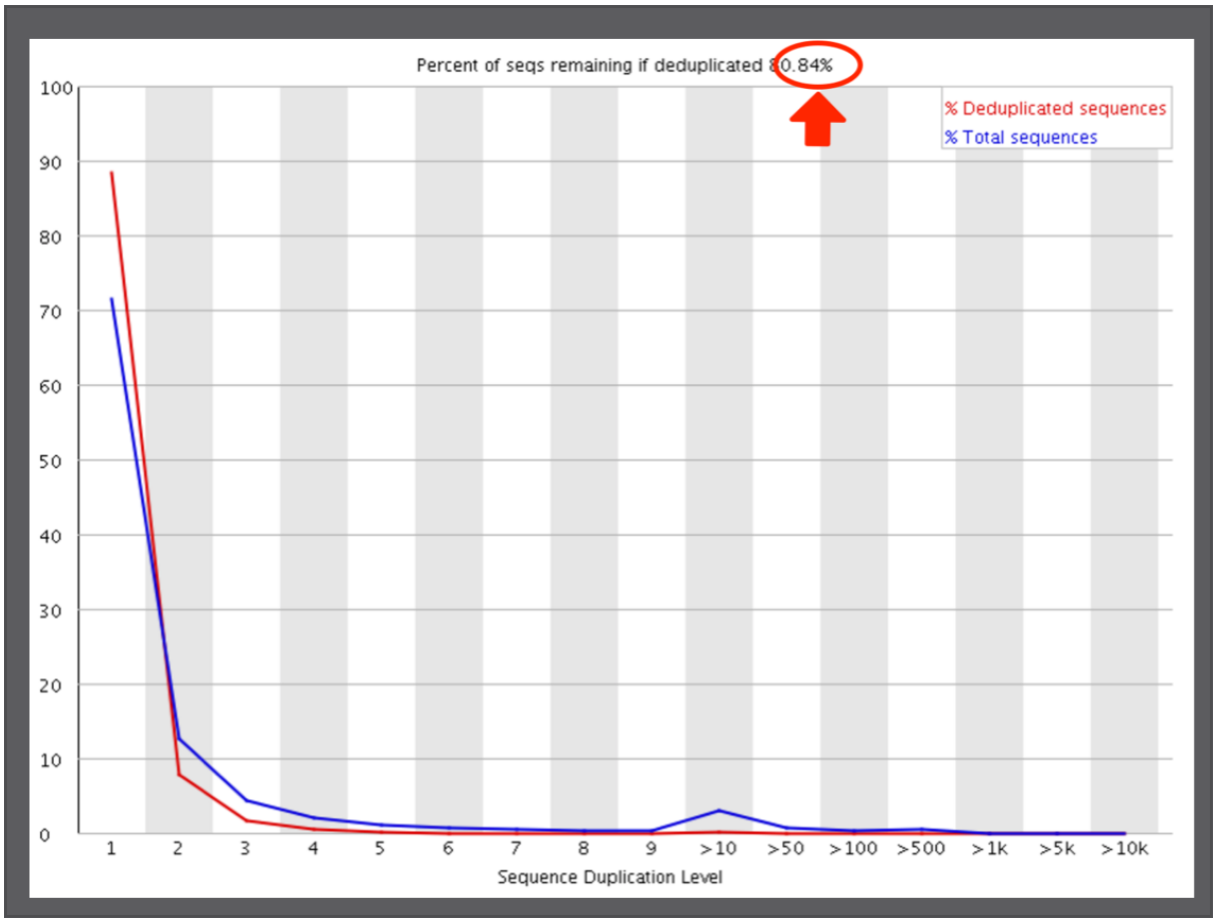
检测重复有两种方法：序列相同，比对相同。

> The adverse effect of read duplication manifests itself mostly during variation detection

读段重复最大的问题是在检测变异上，如果一个变异点重复两次，会产生与实际不符的效果。

> For SNP calling and genome variation detection, the answer is usually yes. For other processes, the answer is usually no.

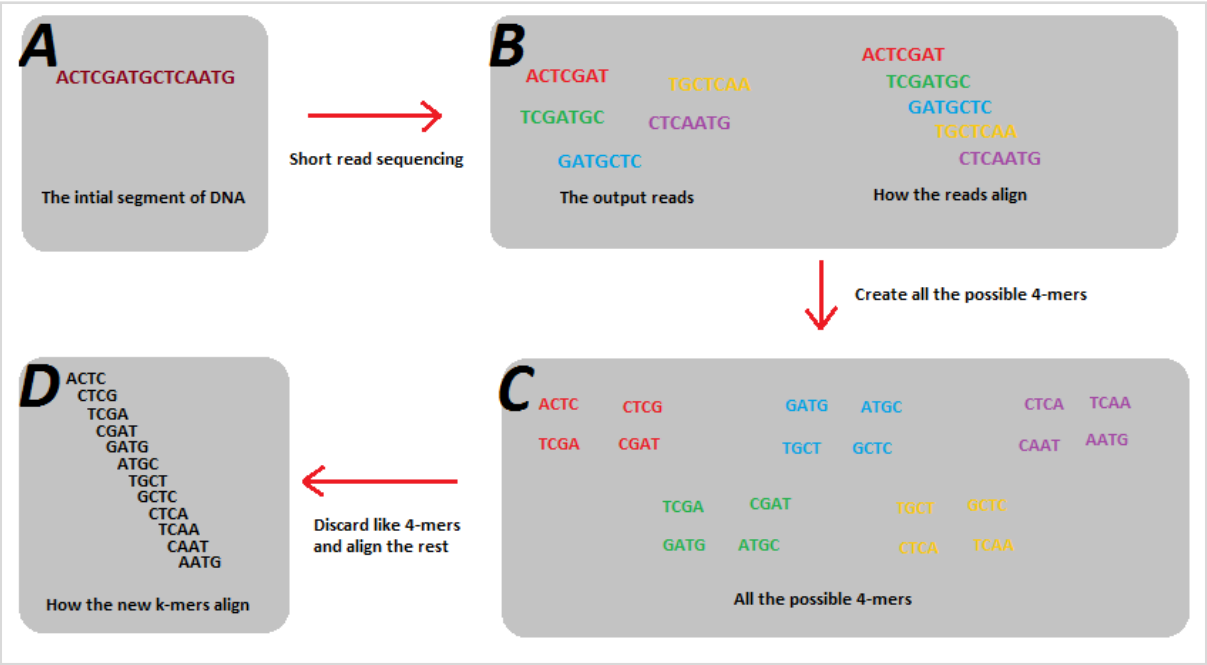
SNP calling和基因组变异检测需要移除重复，其他就不需要。



> FastQC duplicating plot means what percent of the data is distinct. “There are two kinds of duplicate counts: one for all sequences, and another for distinct sequences.

To remove duplicate: picard Tool documentation

k-mers



> There is a new class of methods that detect subsequence composition (called k-mers) that can be used to correct errors.

K-mers

> A k-mer typically refers to all the possible substrings of length k that are contained in a string. For example if the sequence is ATGCA then

K-mers for ATCGA:
*** 2-mers:AT, TG, GC and CA**
*** 3-mers:ATG, TGC and GCA**
*** 4-mers:ATGC, TGCA**
*** 5-mers:ATGCA**

> Error correction: rare k-mers are more likely to be caused by sequence errors.
> Classification: certain k-mers may uniquely identify genomes.
> Pseudo-alignment: new pseudo-aligners can match reads to locations based solely on the presence of common k-mers.

纠错：那些稀有不常见的k-mers，可能仅仅是测序错误.
分类：基因组中特异性的k-mers可以用来区分不同物种。 Classification: certain k-mers may uniquely identify genomes.
Pseudo-alignment(伪比对）：目前RNA-Seq定量分析中出现了一类alignment-free工具，其原理就是先准备不同基因的k-mers的索引，通过将read的k-mers和k-mers索引比较，从而

对基因进行计数。

K-mers for Genome Size Estimation Genome Size Estimation Tutorial | Computational Biology Core

In fasta **lastz** 默认过滤小写字母（重复区）

lastz

一些更加方便的脚本工具：

- estimate*genome*size.pl：https://github.com/josephryan/estimate_genome_size.pl
- KmerGenie：http://kmergenie.bx.psu.edu/
- 华大的GCE： ftp://ftp.genomics.org.cn/pub/gce
- ALLPATHS-LG的 findErrors模块。