it is important to document the parameters you are using to run the analysis, in addition it is very important to document the results at every step, or some metric that speaks to the results. Evaluating these is a form of QC and it will enable you to identify any issues with the data and/or the parameters you are using, as well as alert you to the presence of contamination or systematic biases, etc.

* **number of raw reads**
* **percentage of reads aligned to genome**
* **percentage of reads associated with genes**

MultiQC

General Statistics

Raw reads: **M Seqs**

uniquely mapping reads: **%Aligned**, A good quality sample will have at least 75% of the reads uniquely mapped. The lower the number of uniquely mapping reads means the higher the number of reads that are mapping to multiple locations. ‘STAR: Alignment Scores’ plot visually represents much of this information.

NOTE: The thresholds suggested above will vary depending on the organism that you are working with. Much of what is discussed here is in the context of working with human or mouse data. For example, 75% of mapped reads holds true only if the genome is good or mature. For badly assembled genomes we may not observe a high mapping rate, even if the actual sequence sample is good.

### Complexity

The complexity of the RNA-seq library can be explored a bit with the %Dups column. If a large percentage of the library is duplicated, then this could indicate either a library of low complexity or over-amplification. If there are differences between libraries in the complexity or amplification, then this can lead to biases in the data, such as differing %GC content.

在FastQC报告中，"READ\_PAIR\_OPTICAL\_DUPLICATES" 指的是光学重复读数，这些是由于测序仪器的光学传感器错误地将单个扩增簇识别为多个簇而产生的重复读数.

在测序数据分析中，"READ\_PAIR\_DUPLICATES" 指的是在双端测序（Paired-end sequencing）中，来自同一DNA片段的两个配对读段（read pairs）被认为是重复的。这种现象通常是由于PCR扩增过程中的重复序列产生，即PCR duplicates。这些重复读段在基因组分析中可能不是必需的，因为它们不代表额外的遗传信息，而且在某些分析中，如变异检测或基因表达定量，重复读段可能导致分析结果的偏差。

1. **不同测序类型的正常重复率**：
   * 全外显子测序（WES）和全基因组测序（WGS）的重复率通常在10%左右。
   * 全基因组DNA甲基化测序（WGBS）的重复率可能高于10%。
   * 转录组测序（RNA-seq）的重复率通常在30%到40%左右。
   * 多重PCR测序和捕获Panel测序的重复率与测序的区域以及测序量有关。

### Exploring biases

Within this report we can also explore the bias metrics output by **Qualimap** and FastQC. The 5'-3' bias column denotes whether our data has any 5’ or 3’ biases. These biases could be due to RNA degradation or sample preparation techniques. Generally, we should explore our data more if we have biases approaching 0.5 or 2.

The transcript position plot can also help identify 5’ or 3’ bias in addition to any other coverage issues. We generally expect roughly even coverage.

In addition, we can see whether our different samples have differences in %GC. A GC bias in our data can present as differences in composition of %GC. These biases could be caused by low-complexity libraries, differences in amplification, or library-specific causes.

### Contamination

We can also identify problems with our library or contamination of our samples by looking at the percent of reads that are exonic, intronic or intergenic. High levels of **intergenic reads** is indicative of DNA contamination (**>30%**). Also, if a polyA selection of messenger RNAs was performed, then high percentages of intronic reads would be concerning.

Generally in a good library, we expect **over 60% of reads to map to exons** for mouse and human organisms. For other organisms the percentage depends on how well annotated is the genome.

**Picard**

* 主要功能包括去除 PCR 重复（MarkDuplicates）、数据格式转换、插入片段大小分布分析（CollectInsertSizeMetrics）、文库复杂性评估（EstimateLibraryComplexity）等。

**Qualimap**

* Qualimap 是一个用于评估高通量测序数据质量的开源软件，支持多种实验类型，如全基因组测序、RNA-seq、ChIP-seq 等。
* 它提供了 BAM QC 模块、RNA-seq 模块和功能富集分析等，能够生成包含丰富图表的 PDF 报告，直观展示数据质量。

**RSeQC**

* RSeQC 是一个用于评估 RNA-Seq 数据质量的开源 Python 工具集，提供了多种分析模块，如序列质量、GC 偏差、PCR 偏差、核苷酸组成偏差等。
* 它还可以评估测序饱和度、映射读数分布、覆盖均匀性、链特异性和转录水平 RNA 完整性等。

**Samtools**

* Samtools 是一个用于操作 SAM/BAM 格式文件的工具集，提供了多种命令来查看、排序、索引、合并和转换 SAM/BAM 文件。
* 它还包括一些统计工具，如计算比对深度、覆盖率和标记重复读段等。