思考问题的熊

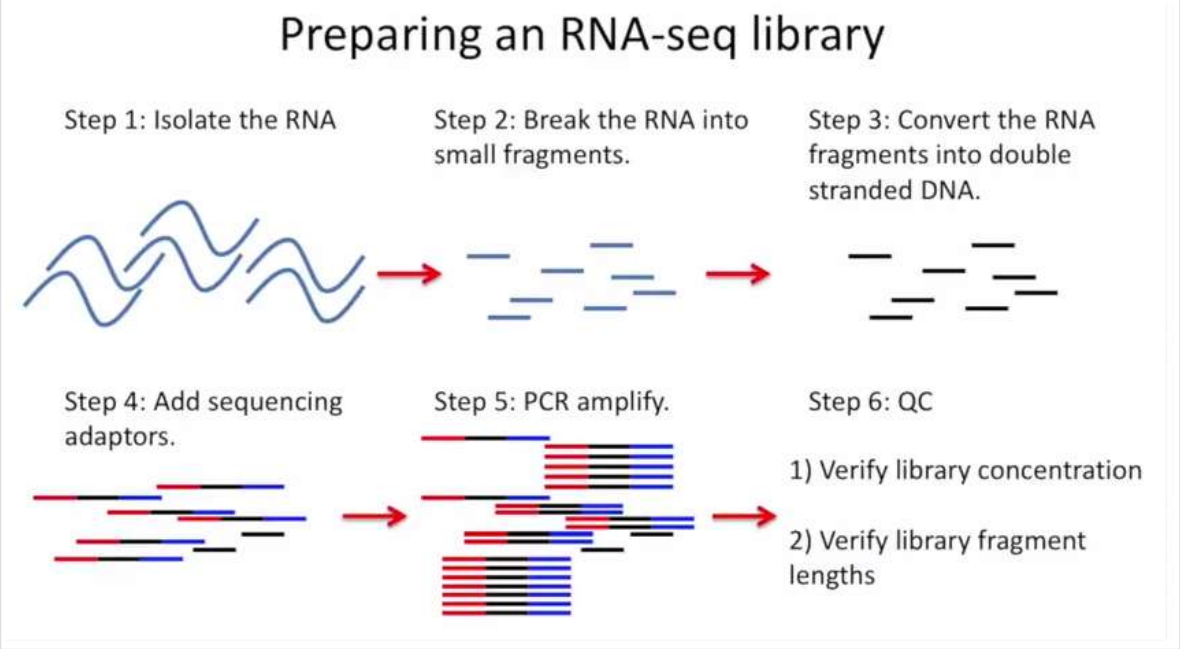
链特异性测序那点事

2017-10-23 | 生物信息 | 19

> 关于链特异性测序的若干问题，很久以前就以为自己想清楚了，但是每次提起它的时候又容易重新产生各种各样的小困惑。于是整理一下，以免以后再时不时犯迷糊。很多东西就是这样，你以为的明白并不是真的明白，一年前的明白和一年后的明白也不是同一个明白。我这么说，不知道你能明白还是不明白。

RNA-seq基本流程

下图是一个大概的RNA-seq基本流程



把RNA破碎成小片段，然后将RNA转变成一条cDNA，这一步需要用到反转录酶 reverse transcriptase (RT) 才能用RNA作为模板合成DNA。

不论是转录还是反转录都需要引物。通常如果我们要mRNA，那就可以用oligo-dT作为RT的引物，但是用它有两个问题，第一个是只能反转录那些有A尾巴的RNA，第二个问题是RT不是一个高度持续性的聚合酶，可能让转录提前发生终止，造成的结果就是3'端要比5'端reads富集，这样就会使得后续定量分析带来bias。

另一种常用的引物称为**随机引物**，随机引物的好处是没有A尾巴的诸如ncRNA也被留下了，而且不会存在明显的3'端偏差。但是很多研究也发现，所谓的随机引物根本就不随机，**这也是测序结果中，通常前6个碱基的GC含量分布特别不均匀的原因**。这几个碱基GC含量均匀很可能不是接头或者barcode那些东西，其实是Illumina 测序RT这一步的random

hexamer priming 造成的bias，很多人在处理数据的时候会把这几个碱基去掉，其实很多时候真多RNA-seq数据去不去掉基本什么影响，不过开头如果有低质量的碱基倒是应该去掉。

随后是第二条链合成，这一步用是DNA聚合酶，以刚才和成的第一条链作为模板。

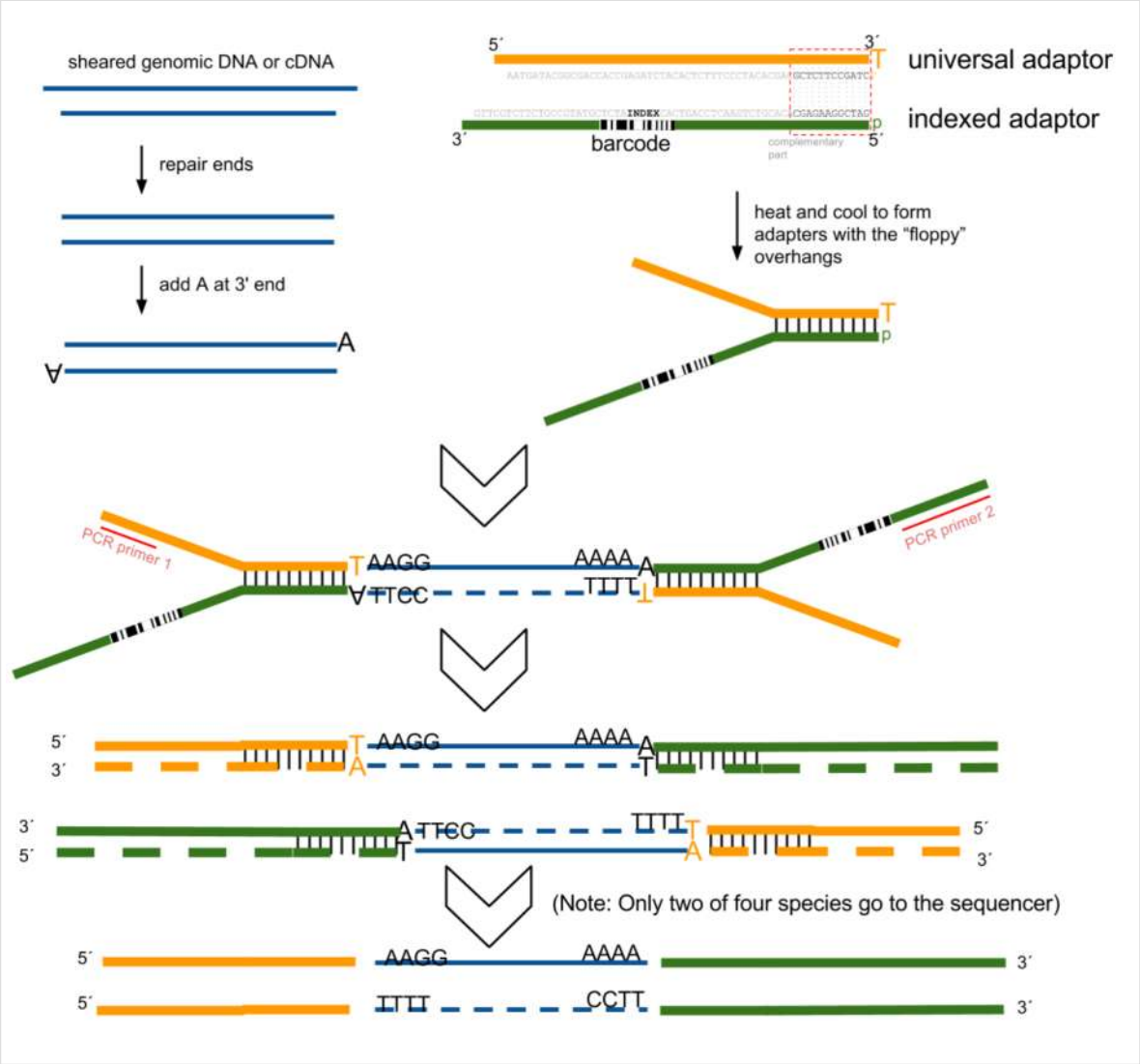
接下来就是在序列两端加上接头，加接头一方面是为了让机器可以识别这些序列，把这些序列固定；二是为了让多个样品可以同时上机，平摊每个样品的测序价格。双端测序为了让read从两边开始延伸，也需要在两端有所需的引物。

Adapter element	Requirement	Location	Function
Amplification element	Required	5′ and 3′ terminus	Clonal amplification of the construct
Primary sequencing priming site	Required	Adjacent to the insert	Initiating the primary sequencing reaction
Barcode/Index	Optional	5′-end of the insert/Between the sequencing priming site and its respective amplification element	Provides a unique label to sequences from different samples. Allows pooling of multiple experiments in a single sequencing reaction.
Paired-end sequencing priming site	Optional	Adjacent to the insert on the side opposite of the primary sequencing priming site	Sequencing into the insert on the end opposite of the primary read
Index sequencing priming site	Optional	Complementary to the 5′-end of the sequencing priming site	Sequencing of the index

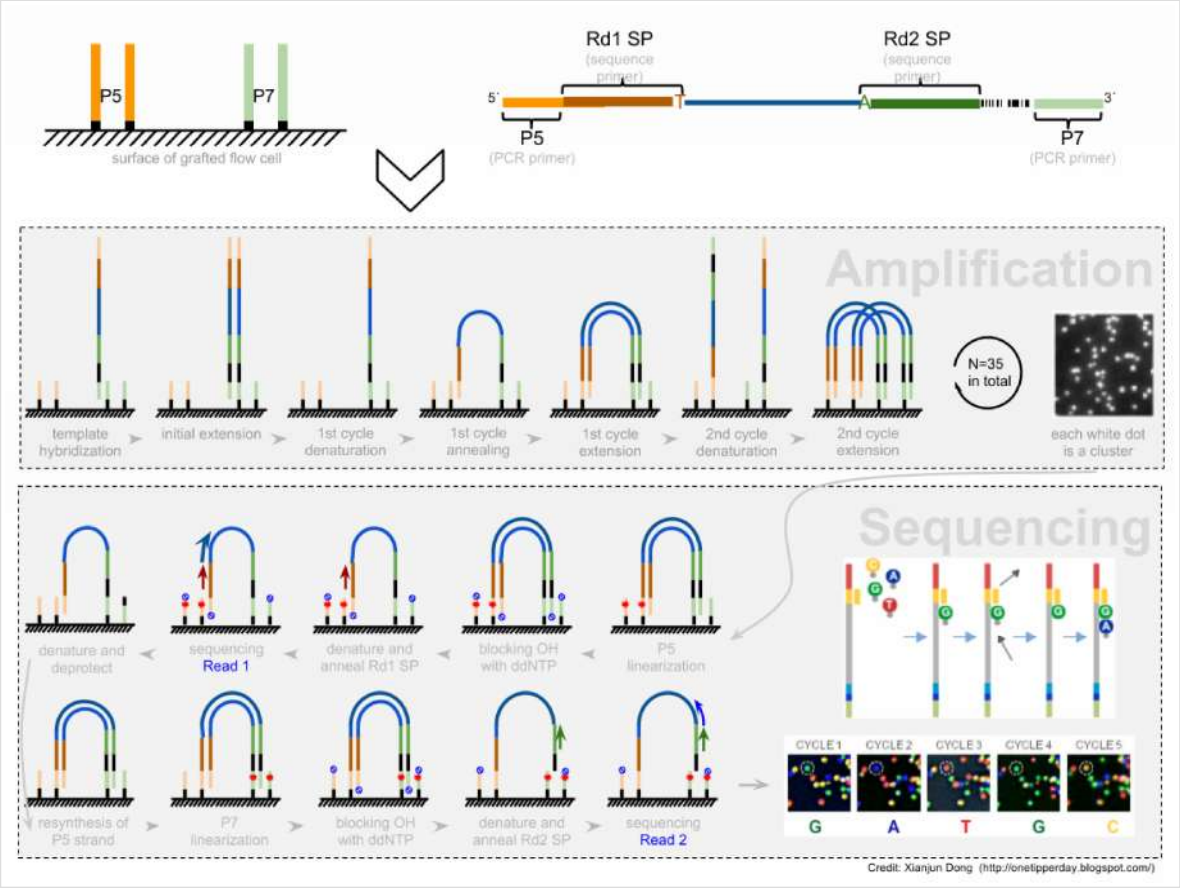
所谓双端测序，因为很多时候read的长度要短于insert，为了增加覆盖度于是就想出了从insert两端同时测序的办法。使得测序深度增加的同时也能够用来判断isoform方向。

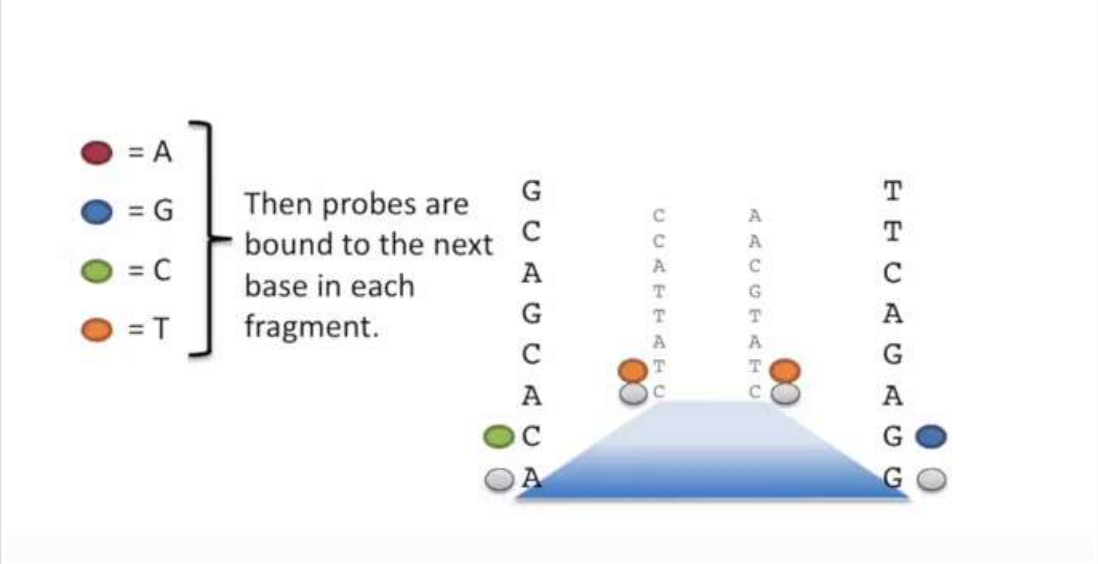
对于illumina数据，有一条5-3的universal adaptor；还有一条是3-5的indexed adatpor，这条引物含有特意的barcode。需要说明的是，在双端测序中，如果insert 不是足够长，那么R1可能就会测到R2的引物，同时R2可能会测到R1引物的反向互补序列。

大概的意思就是下面两张图。



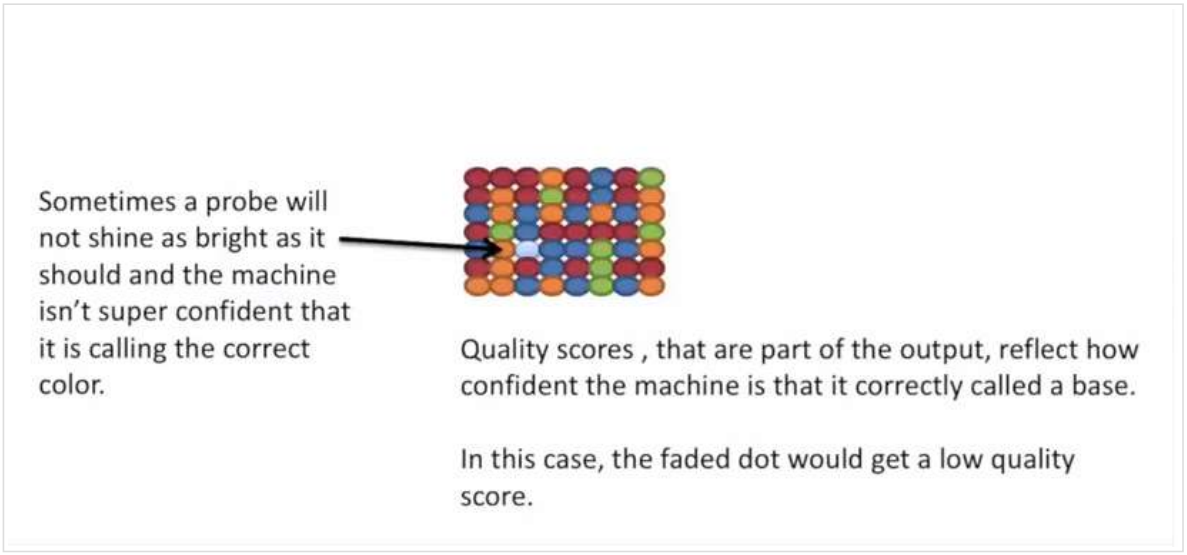
加了接头以后进行PCR的扩增。扩增后就开始测序，测序的过程如下图所示。





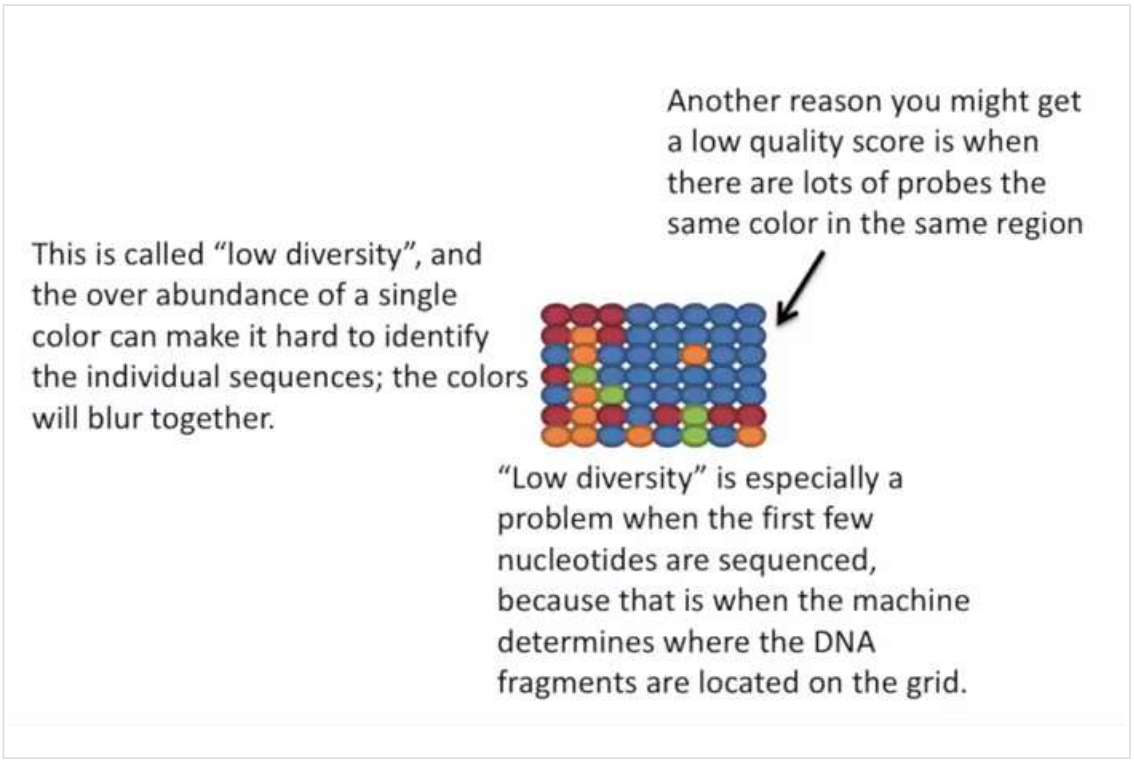
测序的基本思想是机器识别四种碱基发出的不同颜色的荧光，可以理解为一个flow cell 立着非常多序列，机器一层一层扫过去，通过识别荧光而判断这一层每个序列的碱基是什么。

因为一个cell密密麻麻的全是荧光信号，机器并不是总能把每一个判断的非常准确，如果某一个荧光信号没有那么清晰，这个碱基的测序质量就比较低，如下图。



有的时候，如果一大片点都是同一种荧光，机器也可能犯晕，不知道到底哪一个荧光属于哪一个序列。这种情况尤其是在序列的前几个碱基容易发生。

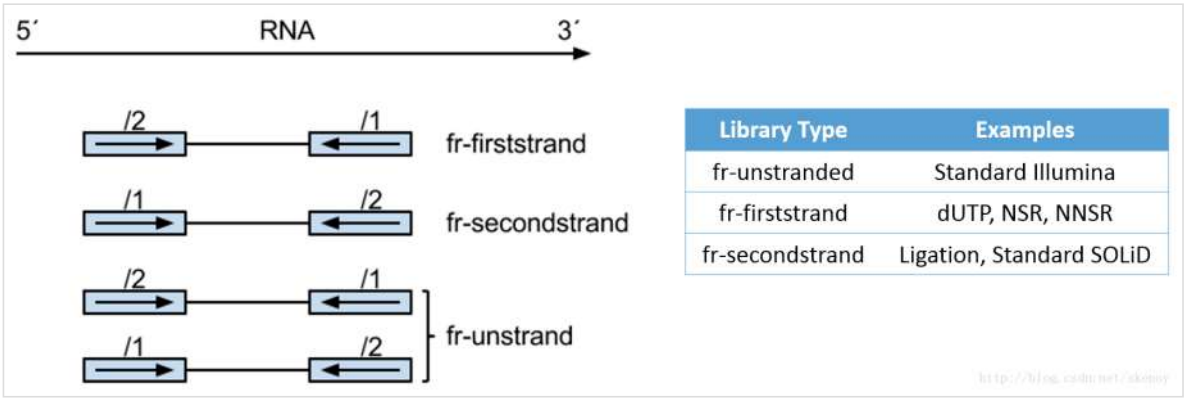
> The sequencing machine uses the first few bases to establish where the cDNA fragments are on the flow cell. If all of the bases in one part of the flow cell are all the same, like 'C', and all show up green in the picture, then the colors will bleed together and it will not be clear where exactly all of the reads are. In contrast, if you have a lot of different colors in a region, it's easier to determine where each one is, even with a little color bleed.



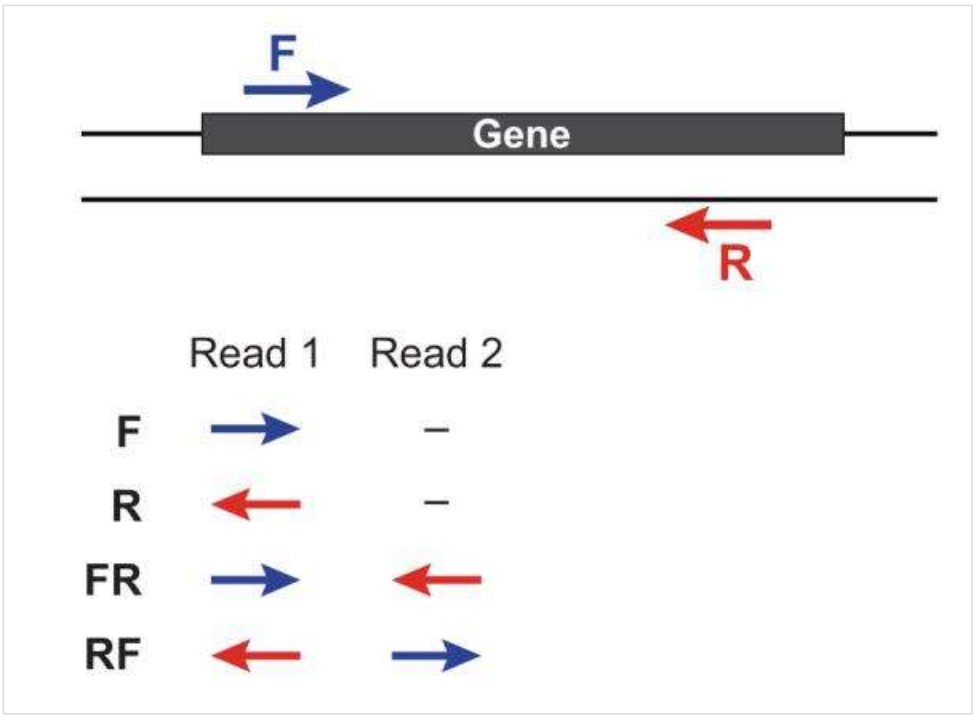
链特异性测序

和普通的RNAseq不同，链特异性测序可以保留最初产生RNA的方向，普通建库方式为什么不行呢？因为传统建库方式通过两个接头的ligation把RNA已经变成了双链DNA，最后的文库中一部被测序的链对应正义链（sense strand），一部分被测序的链测是反义链。

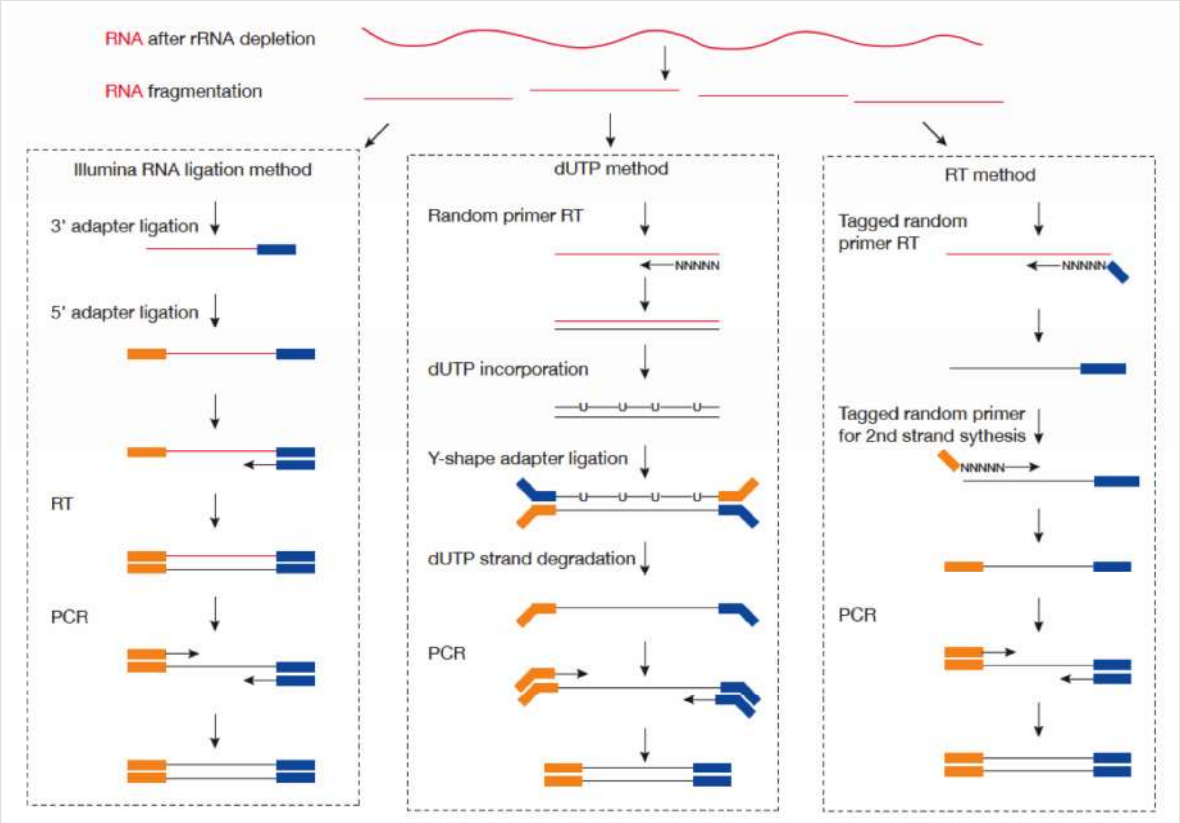
链特异性建库方式有不止一种，对应到不同的软件又有不同的叫法，下面是几种称呼。**要记住的是dUTP 测序方式的名字是fr-firstrand，也是RF。** 至于具体的read方向接下来通过更详细的IGV截图说明问题。



Library Type	Examples
fr-unstranded	Standard Illumina
fr-firststrand	dUTP, NSR, NNSR
fr-secondstrand	Ligation, Standard SOLiD



链特异性建库方式（以目前最常用的dUTP为例，如下图所示）首先利用随机引物合成RNA的一条cDNA链，在合成第二条链的时候用dUTP代替dTTP，加adaptor后用UDGase处理，将有U的第二条cDNA降解掉。



这样**最后的insert DNA fragment都是来自于第一条cDNA，也就是dUTP叫fr-firststrand的原因。**对于dUTP数据，**tophat** 的参数应该为 `–library-type fr-firststrand` 。这里的first-strand cDNA可不是RNA strand，在使用**htseq-count** 时，真正的正义链应该是使用参数 `-s reverse` 得到的结果。

正正反反不清楚

说到链特异性测序，实在让人困惑的是**各种链的概念**，尤其是翻译成中文就更说不清了。

DNA 的正链和负链，就是那两条反向互补的链。参考基因组给出的那个链就是所谓的正链（forword），另一条链是反链（reverse）。但是这正反一定**不能和正义链（sense strand）反义链（antisense strand）混淆**，两条互补的DNA链其中一条携带编码蛋白质信息的链称为正义链，另一条与之互补的称为反义链。但是携带编码信息的正义链不是模板，只是因为它的序列和RNA相同，正义链也是编码链。而反义链虽然和RNA反向互补，但它可是真正给RNA当模板的链，因此反义链也是模板链。

总结两点

1. **正义链（sense strand）= 编码链（coding strand）= 非模板链**

2. **forword strand 上可以同时有sense strand 和 antisense strand。因为这完全是两个不同的概念。**

写这篇文章的原因，就是因为有人问我，链特异性测序数据 htseq-count 的结果是不是应该把正负链的基因分别在-s yes 和-s reverse 两个参数结果中统计出来再做下游分析。这里犯的错误就是我们混淆了基因组正反链和基因正义反义链的概念。

dUTP到底是怎么回事

从前文的一个图我们可以总结出**dUTP方式测序R1文件中read1 的方向和基因的方向（正义链）是相反的，而R2文件中的read2 方向和基因的方向是相同的。**

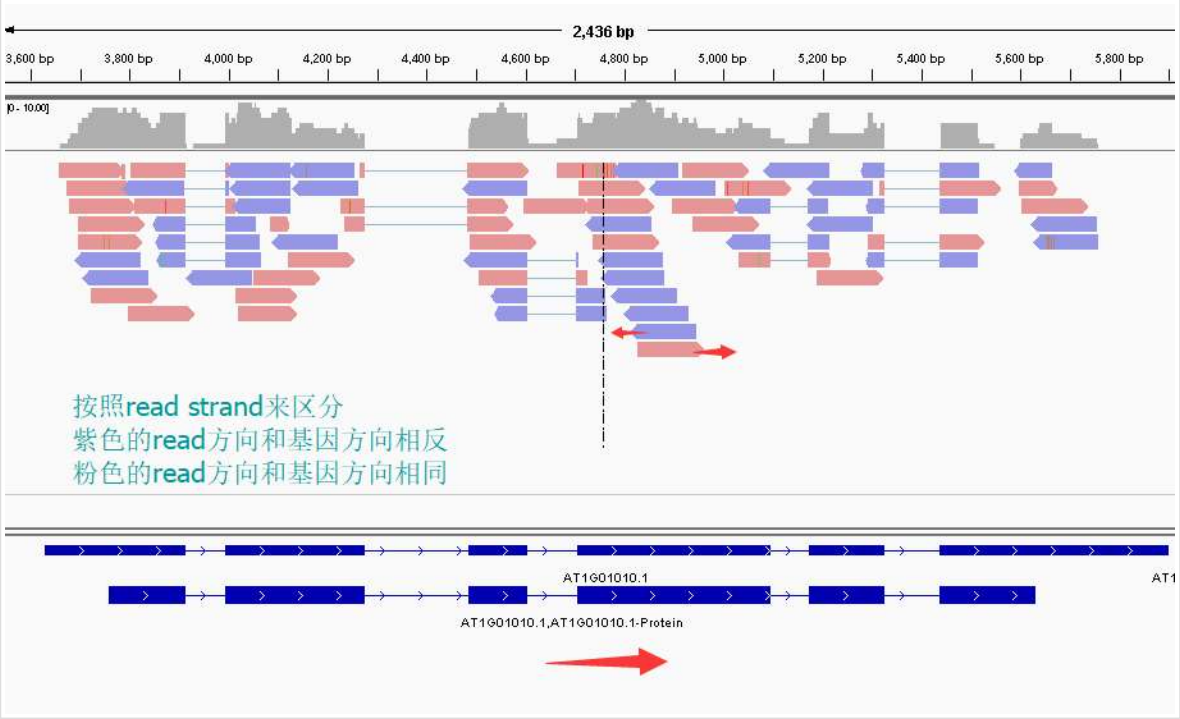
可以参考下面的两个igv文件bam截图。

首先解释一下igv 两个颜色参数的意义

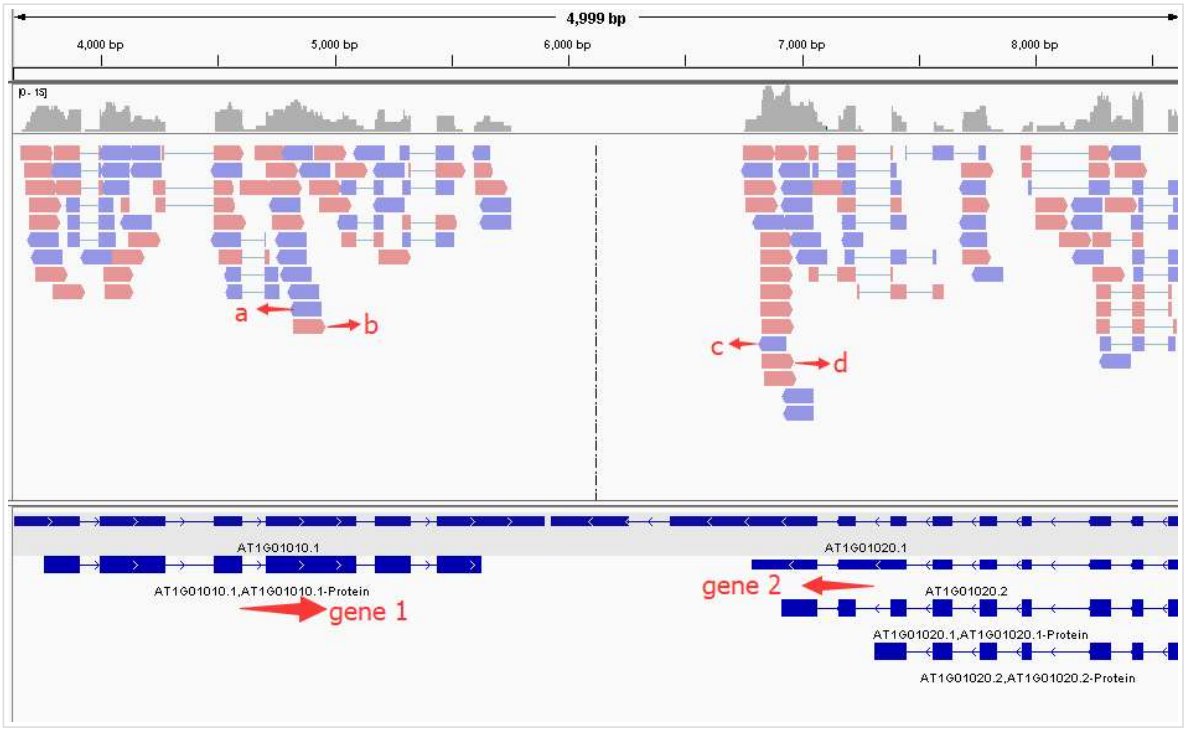
- **Read strand** in pastels, **red for positive rightward (5' to 3') DNA strand, blue for negative leftward (reverse-complement) DNA strand**, and grey for unpaired mate, mate not mapped, or otherwise unknown status.
- **First-of-pair strand** assignment is dependent on RNA transcript directionality and is **useful for directional libraries. Displays reads or read pairs in which the forward read is first (F1 or F1R2) in red and reads or read pairs in which the reverse read is first (R1 or R1F2) in blue.** Unknown status is in gray.
    - For a given transcript, non-directional libraries will show a mix of red and blue reads aligning to the locus.
    - **Directional libraries will show reads of one color in the direction matching the transcript orientation.**

下面这个图示按照igv 颜色选项中的read strand 方向进行区分，可以看到所有**红色read都是在正链方向**（注意正链不是正义链），而所有**蓝色的read都是负链方向。**下面基因的方向是正链方向，也就是和粉色的read同向的，如果你把鼠标放到随意一个粉色的read上，就能看到显示的信息是**second of pair**，也就是**pair中的read2（R2）**；反之如果你在蓝色的read上面，就会显示信息是first of pair，也就是R1。

总结，dUTP测序中pair read 中的read1（R1）和基因方向相反，read2（R2）和基因方向相同



再看下面这张图

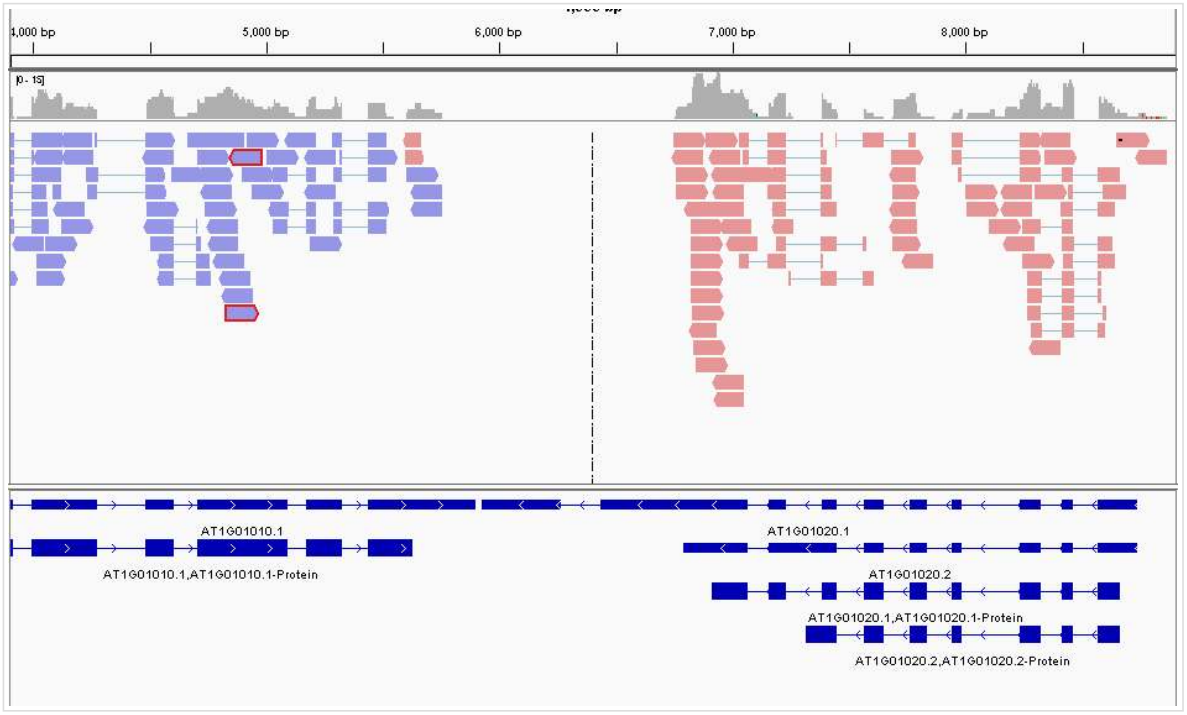


这张图展示了两个基因1和2，我们可以发现**gene1的正义链就在正链上**，而**gene2的正义链其实是在反链上**。看read情况，a，c两个read虽然针对正链负链而言方向一致，都是负链方向，但是如果把**a是pair中的read1（first of pair）**，而**c是pair中的read2（second of pair）**。也就是说，**read方向一致，但一个是read1一个是read2，说明这两个read对应的基因一定是反向的。同样的道理，虽然b，d都是两个方向为负链的read，但是b其实是所在pair的read2（second of pair），而d是所在pair的read1（first of pair）**。

再次强调，dUTP测序中pair read 中的read1（R1）和基因方向相反，read2和基因方向相同

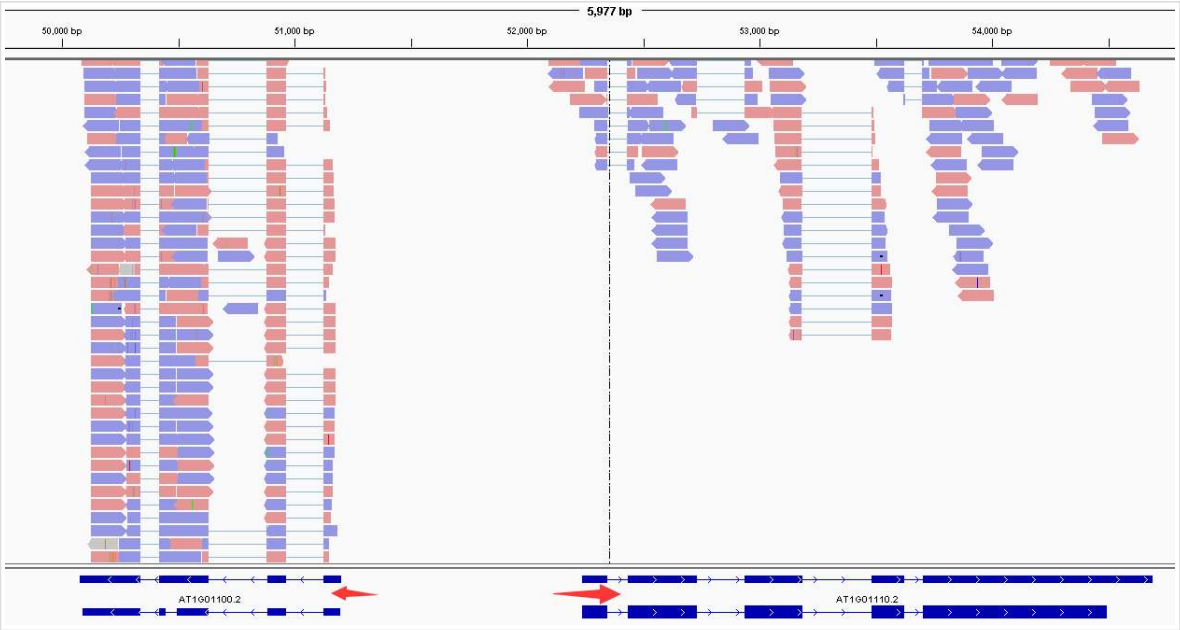
当使用read strand来进行颜色区分的时候，每一个基因上两种颜色的分布应该相对均匀（也就是所谓的pair end）。

如果这个时候把**颜色选项改为按照 `first of pair of strand` 来区分**，会出现下图的变化。



geng1的read全部变成了紫色，而gene2的read全部变成了粉色。

如果是非链特异性测序，在 `first of pair of strand` 模式下，同一个gene相关的read颜色也是明显混杂的。如下图



再一次总结：

- dUTP 链特异性测序中，RNA 方向（gff文件中基因的方向）与read1相反，与read2相同。如果read1比对到基因组正链上，则对应的gene在基因组负链；如果read2比对到基因组正链则对应的gene在基因组正链。
- dTUP 测序方式叫做fr-firstrand（留下的是cDNA第一条链），也是RF。
- 如果dUTP链特异性测序，看基因表达量应该 counts for the 2nd read strand aligned with RNA(htseq-count option -s reverse, STAR ReadsPerGene.out.tab column 3)
- 如果想看反义链是否有转录本（比如NAT）应该用 the 1st read strand aligned with RNA (htseq-count option -s yes，STAR ReadsPerGene.out.tab column 4)

几个常用软件的设置

STAR mpping 时无需特别设置，但如果不是链特异性数据且下游分析要用到cufflinks 则需要增加参数 --outSAMstrandField intronMotif。为的是增加一个XS标签。

> If you have **stranded RNA-seq data, you do not need to use any specific STAR options.** Instead, you need to run Cufflinks with the library option --library-type options. For example, **cufflinks... --library-type fr-firststrand** should be used for the standard dUTP protocol, including Illumina's stranded Tru-Seq.

hisat2 --rna-strandness RF

目的也是给比对序列添加一个XS标签以区分方向，方面cufflinks使用。

> For single-end reads, use F or R. 'F' means a read corresponds to a transcript. 'R' means a read corresponds to the reverse complemented counterpart of a transcript. For paired-end reads, use either FR or RF.
> With this option being used, every read alignment will have an XS attribute tag: '+' means a read belongs to a transcript on '+' strand of genome. '-' means a read belongs to a transcript on '-' strand of genome.

tophat –library-type option fr-firststrand

具体解释参见下表

Library Type	Examples	Description
fr-unstranded	Standard Illumina	Reads from the left-most end of the fragment (in transcript coordinates) map to the transcript strand, and the right-most end maps to the opposite strand.
fr-firststrand	dUTP, NSR, NNSR	Same as above except we enforce the rule that the right-most end of the fragment (in transcript coordinates) is the first sequenced (or only sequenced for single-end reads). Equivalently, it is assumed that only the strand generated during first strand synthesis is sequenced.
fr-secondstrand	Ligation, Standard SOLiD	Same as above except we enforce the rule that the left-most end of the fragment (in transcript coordinates) is the first sequenced (or only sequenced for single-end reads). Equivalently, it is assumed that only the strand generated during second strand synthesis is sequenced.

htseq-count -s reverse/yes(看反义链)

> For `stranded=no`, a read is considered overlapping with a feature regardless of whether it is mapped to the same or the opposite strand as the feature. For `stranded=yes` and single-end reads, the read has to be mapped to the same strand as the feature. For paired-end reads, the first read has to be on the same strand and the second read on the opposite strand. For `stranded=reverse`, these rules are reversed.

RSEM --forward-prob 0（正义链）1（看反义链）

> The RNA-Seq protocol used to generate the reads is strand specific, i.e., **all (upstream) reads are derived from the forward strand. This option is equivalent to --forward-prob=1.0**. With this option set, if RSEM runs the Bowtie/Bowtie 2 aligner, the '–norc' Bowtie/Bowtie 2 option will be used, which disables alignment to the reverse strand of transcripts. (Default: off)
>
> Probability of generating a read from the forward strand of a transcript. Set to 1 for a strand-specific protocol where all (upstream) reads are derived from the forward strand, **0 for a strand-specific protocol where all (upstream) read are derived from the reverse strand**, or 0.5 for a non-strand-specific protocol. (Default: 0.5)

sXpress --rf-stranded / --fr-stranded(看反义链)

> –fr eXpress only accepts alignments (single-end or paired) where the first (or only) read is aligned to the forward target sequence and the second read is aligned to the reverse-complemented target sequence. In directional

> sequencing, this is equivalent to second-strand only.
>
> –rf eXpress only accepts alignments (single-end or paired) where the first (or only) read is aligned to the reverse-completemented target sequence and the second read is aligned to the forward target sequence. In directional sequencing, this is equivalent to first-strand only.

trinity --SS_lib_type RF

> Trinity performs best with strand-specific data, in which case sense and antisense transcripts can be resolved.
>
> *RF*: first read (/1) of fragment pair is sequenced as anti-sense (reverse(*R*)), and second read (/2) is in the sense strand (forward(*F*)); typical of the dUTP/UDG sequencing method.

参数错了又怎样？

到这里，会想问两个问题。有时候我们不知道数据的建库方式是不是链特异性的，如果弄错了结果会怎么样呢？

如果你用STAR mapping 完可以用igv像上文提到的那样，看看是不是链特异性测序。

下面是两个真是数据的count 统计情况。

对于**仅仅进行基因表达定量**来说，把链特异性数据当作普通建库数据来处理，可以观察第2列数据和第4列数据。具体某一个基因而言，影响不会太大，因为绝大多数反义链本身表达量就非常低。
不过可以注意 noFeature 和 ambiguous 这两个值，因为基因组中存在两个基因分别在正链和负链且又重叠的情况，不区分方向会比区分方向的ambiguous数目多一些。因为如果不能通过方向来区分到底属于哪个基因，这样的read就会被认为是ambiguous。

但是因为区分了方向，又会使得noFeature的数目更多一些。不过两者总体影响不会差别非常大。如果不能判断建库方式，在htseq中使用-s no 参数是一个比较保险（虽然不是非常精确）的做法。

	-s no	-s yes	-s reverse
N_noFeature	1290001	16837194	1480658
N_ambiguous	633021	16413	74710
AT1G01010	58	2	56
AT1G01020	65	0	65
AT1G03987	5	5	0
AT1G01030	296	5	291
AT1G01040	901	5	1078
AT1G03993	0	182	0
AT1G01050	428	0	434
AT1G03997	1	7	0

	-s no	-s yes	-s reverse
AT1G01060	85	0	85
AT1G01070	73	0	73
AT1G04003	0	0	0
AT1G01080	1166	15	1151
AT1G01090	2901	0	2901
AT1G01100	1560	0	1560
AT1G01110	82	0	82
AT1G01120	484	0	484
AT1G01130	72	9	63
AT1G01140	518	3	515
AT1G01150	0	1	0
AT1G01160	356	192	551
AT1G04007	4	189	0
AT1G01170	55	11	423

相反，如果把普通建库方式的数据当作链特异性数据来处理。

比如在htseq-count中使用了-s reverse 参数，这个时候**只有R2方向和基因方向相同的pair才用来算作一个count**，所有R2和基因方向不同的pair就被当作no feature了。这样的结果影响可以通过下面的表格观察。

用正常方法数出的noFeature 是6万左右，而用-s yes或者reverse数出来的noFeature 就接近46万了。将近40万的read 被丢掉。

所以，如果把普通建库的数据误当作链特异性数据来处理极有可能会损失大量的数据，如果**弄错了链特异性建库的方式**，那坑你就没几个read剩下了。另外，计算出来的结果自然也会有非常大的差异，是不准确的。

	-s no	-s yes	-s reverse
N_unmapped	729831	729831	729831
N_multimapping	443861	443861	443861
N_noFeature	63787	4591673	4599916
N_ambiguous	594720	27992	27789
AT1G01010	101	54	47

	-s no	-s yes	-s reverse
AT1G01020	84	41	43
AT1G03987	1	0	1
AT1G01030	80	37	43
AT1G01040	499	279	293
AT1G03993	0	33	25
AT1G01050	634	312	337
AT1G03997	0	0	0
AT1G01060	3274	1644	1630
AT1G01070	97	43	54
AT1G04003	0	0	0
AT1G01080	585	303	282
AT1G01090	1768	907	861
AT1G01100	1132	549	583
AT1G01110	60	21	39
AT1G01120	1099	573	526

参考资料1

参考资料2

参考资料3

鼓励创作

赞赏

bioinformatics # RNA-seq

shell脚本应知应会

0