

平成27年度 NGSハンズオン講習会 NGS解析：RNA-seq（8月5日、28日）講義資料

資料名	ファイル名
講義資料（前半：アメリエフ株式会社）	RNA-seq(PDF:4.16MB)
講義資料（後半：東大）	RNA-seq、カウントデータ取得以降の統計解析(PDF:2.74MB)
deFuse	http://combio.bcrc.ca/software/defuse/
TopHat-Fusion	http://ccb.ihp.edu/software/tophat/fusion_index.html
DDBJ SRA (DRA)	http://trace.ddbj.nig.ac.jp/dra/index.html
TopHat	https://ccb.ihp.edu/software/tophat/index.shtml
UCSC	https://genome.ucsc.edu/

○講義メモ	ハンズオンメモ
画面の拡大または縮小は、「CTRLと+」または「CTRLと-」	
融合遺伝子検出は難しい。deFuseやTopHat-Fusionなどがある。	
DDBJ SRA (DRA)の見方概要	
fastqcコマンド実行時に「-t 2」オプションをつけると、CPU2個を使って処理するので早くなる。	
fastx_trimmerのxは、FASTQ/FASTAどちらでも対応可能という意味	
スライド20は、最初の1塩基と書いている。が、どう考えても最初の7塩基をトリムしたほうがよさそうなのでそうする。	
スライド21で、8塩基目から使いたいの、ととりあえず以下のように実習では打った。「-l 82」は、82塩基目までを残して、それ以外はトリムするというもの。	<pre>cd rnaseq fastx_trimmer -f 8 -l 82 -i ../1K_SRR518891_1.fastq -o 1K_SRR518891_1_s.fastq</pre>
スライド22で、fastx_clipperのオプション。 -lはデフォルトが5だが、マッピングプログラムの性能的に20塩基程度はないとマップできないので、-l 20とする。	
スライド23で、prinseqを推奨。リード中のPolyA/Tの長さに合わせて自動的にトリムしてくれる。	<pre>fastx_clipper -a AAAAA -i 1K_SRR518891_1_s.fastq -o 1K_SRR518891_1_s_notail.fastq</pre>
tagcleanerは捨てすぎる。fastx_clipperはあまり捨ててくれない。cutadaptはちょうどよかった。	
fastq_quality_trimmerで、順番は大事。最初にquality20未満をトリムしてから、80%以上がquality20以上などを残す。	<pre>fastq_quality_trimmer -t 20 -l 30 -i 1K_SRR518891_1_s_notail.fastq fastq_quality_filter -q 20 -p 80 -Q 33 -o 1K_SRR518891_1_clean.fastq</pre>
TopHatのマニュアルを眺める。	
スライド29で、-oで作成するディレクトリ名、-Nが許容するミスマッチ数（デフォルトが2）、--read-edit-distがedit distance。 これはbest practiceによく書かれていたから。 「...Bowtie2Index/genome」までしか記述せず、拡張子は書かないのがポイント。	<pre>tophat -o 1K_SRR518891 -g 3 -N 10 --read-edit-dist 10 --read-gap-length 10 /home/iu/Desktop/amelieff/Scerevisiae/Bowtie2Index/genome 1K_SRR518891_1_clean.fastq ls 1K_SRR518891</pre>
スライド32	<pre>samtools index 1K_SRR518891/accepted_hits.bam igv.sh</pre>
スライド33で、~/Desktop/amelieff/rnaseq/1K_SRR518891中のjunction. bedもIGVで読み込ませた。 「II:45,600-46,000」あたりのジャンクションリードを眺めた。	
スライド36	<pre>cufflinks --min-frags-per-transfrag 2 -o 1K_SRR518891 1K_SRR518891/accepted_hits.bam ll -h 1K_SRR518891</pre>
igenomesに情報がある場合は、それを使うのが王道。 UCSCのTable Browserで各生物種のgenes. gtfまでその場で作成できる。 アノテーションファイル取得に関するTips。 ただし、遺伝子名の列が間違ったりして修正が大変	