平成27年度　NGSハンズオン講習会　NGS解析基礎（8月3日、26日）講義資料

資料名	ファイル名
講義資料	NGS解析基礎(PDF:3.07MB)
command.zip	command.zip (zip：2KB)
samtools	http://samtools.sourceforge.net/
DDBJ SRA (DRA)	http://trace.ddbj.nig.ac.jp/dra/index.html
IGV	https://www.broadinstitute.org/igv/
Youtube：Illumina Inc	https://www.youtube.com/user/IlluminaInc
Youtube：PacBio	https://www.youtube.com/user/PacificBiosciences
Youtube：Ion Torrent	https://www.youtube.com/user/iontorrent
Nagasaki et al., Nat Commun., 2015	Rare variant discovery by deep whole-genome sequencing of 1,070 Japanese individuals
bwa	http://bio-bwa.sourceforge.net/
tophat-fusion	http://ccb.jhu.edu/software/tophat/fusion_index.html
SEQanswers	http://seqanswers.com/
Velvet	http://www.ebi.ac.uk/~zerbino/velvet/
Vague	http://www.vicbioinformatics.com/
Powell and Seemann , Bioinformatics, 2013	VAGUE: a graphical user interface for the Velvet assembler

○講義メモ	ハンズオンメモ
samtools viewだけで打つと、ヘルプがでる。[options]の[]は、あってもなくてもいい。 <in.bam>|<in.sam>で、パイプ(|)は、どちらが入力でもいいという意味。 <>は入力として絶対必要なもの。	
samファイルはマップされなかったリードも出力する	
「samtools ngs」などで検索してsamtoolsのサイトにいく。 右上のほうにマニュアルがある。1列目の情報などの詳細を見ることができる。ときどき改訂されているようだ。	
CIGAR記号の説明。12MのMは一致、4IのIはinsertion、など。	
vcfファイルの説明	
FASTA形式は、NCBIは70文字で改行を入れる。lessとgrepを利用	
DDBJ SRA (DRA)でERR038793を検索	
午後は、igvから。less igv.shでファイルの中身を表示。 デフォルトは-Xmx2000mになっている。デカいファイルの場合は、この2000を4000などに変えて利用する。	
Windowsのヒトは、igv.batをダブルクリックで起動すればよい。これは32 bit版のJavaが1.2GBが上限だったという理由がある。	
「samtools view 1K_ERR038793.bam | less」とすればsamに変換した結果をそのままlessで見ることに相当。	
「File - Load from File」でbamファイルを選ぶ。*_sort.bamを選ぶ。ソートされてないものを選ぶと怒られる。	
GUI左下の「Gene」のところで右クリックすると、collapsedやSquishedなどいろいろ選べる。	
bamがあって、baiファイル作成を忘れても、「Tools」メニューの「Run igvtools」で作れるのでべ便利	
「Genomes」メニューの「Create.genome File」でゲノムファイル以外にGene fileなどでおそらくGFF3などアノテーションファイルも読み込ませることができる。	
「firefox fastac_report.html」でFireFoxが起動してhtmlファイルを見ることができる。	
NGS機器の原理の理解はyoutubeなどがいいかも。	
velvetのインストール。普通は、/usr/local/srcなどにソースファイルを置く。	cd ~/Downloads wget -c http://www.ebi.ac.uk/~zerbino/velvet/velvet_1.2.10.tgz cd velvet_1.2.10 make 'MAXKMERLENGTH=79'
マッピングで、segmental duplication（せぐでゅぷ、とも略すらしい)やリピート領域を除外する戦略もあり。 しかしsegmental duplication領域に重要な遺伝子が見つかったりすることもあるので結構ビミョー	
「ngs best practice」が出始めている。	
一つの研究プロジェクトなど、再現性が重要な場合があるときは、ず～っと同じプログラムのバージョンにする。 致命的なバグの場合は、もちろんバージョンを上げる。	
topコマンドでメモリ使用量を見る。	
velvetをアドリブで実行。 以下を実行するとtmpディレクトリが作成されて、その(tmpディレクトリ)中にcontigs.faというアセンブル結果ファイルができる。	cd ~/Desktop/amelieff # single-endでやる場合 velveth tmp 31 -fastq 1K_ERR038793_1.fastq velvetg tmp grep -c ">" tmp/contigs.fa # paired-endでやる場合 velveth tmp2 31 -shortPaired -separate -fastq 1K_ERR038793_1.fastq 1K_ERR038793_2.fastq velvetg tmp2 grep -c ">" tmp2/contigs.fa
velvetをGUIベースで簡単にやるソフトウェア Vague(Powell and Seemann , Bioinformatics, 2013) をインストールして実行。	cd ~/Downloads wget -c http://www.vicbioinformatics.com/vague-1.0.5.tar.gz tar zxf vague-1.0.5.tar.gz cd vague-1.0.5 ./vague
./vagueで起動するGUI上で、「Estimate best k-mer size...」をクリックして出現する 「K-mer estimator」のポップアップ中の「Target genome size」のところの単位は、k, m, gなどと指定するとエラーは吐かない。 しかし、推定されたK-mer sizeが5となってうれしくないので手作業で31などと変えることになる。 Runボタンを押すと(paired-endで1K_ERR038793_1.fastqと1K_ERR038793_2.fastqを指定した場合)コンティグ数が4と出る。 自分で指定したoutput directory中にcontigs.faが作成される。	