# 平成27年度　NGSハンズオン講習会　NGS解析：ゲノムReseq、変異解析（8月4日、27日）講義資料

| 資料名 | ファイル名 |
|---|---|
| 講義資料 | ゲノムReseq、変異解析(PDF:2.30MB) |
| FASTX-Toolkit | http://hannonlab.cshl.edu/fastx_toolkit/ |
| Trimmomatic | http://www.usadellab.org/cms/index.php?page=trimmomatic |
| Bolger et al., Bioinformatics, 2014 | Trimmomatic |
| snpEff | http://snpeff.sourceforge.net/ |
| Cingolani et al., Fly (Austin), 2012 | SnpEff |
| igenome | http://support.illumina.com/sequencing/sequencing_software/igenome.html |
| Zhao and Zhang, BMC Genomics, 2015 | A comprehensive evaluation of ensembl, RefSeq, and UCSC annotations in the context of RNA-seq read mapping and gene quantification |
| GATK Support Forum | http://gatkforums.broadinstitute.org/ |
| vcftools | http://vcftools.sourceforge.net/ |

| ○講義メモ | ハンズオンメモ |
|---|---|
| QC用プログラムは、FASTX-toolkit(原著論文はない)やTrimmomatic (Bolger et al., Bioinformatics, 2014)などが最近よく使われる。 | |
| Reseqはステップごとにbamファイルがどんどん作成されていくので、FASTQファイルサイズの5倍程度のHDD容量を確保しておいたほうが無難 | |
| GATKの後のアノテーションは、フリーソフトのsnpEff (Cingolani et al., Fly (Austin), 2012)が最近よく使われる。 | |
| ゲノムファイルの取得は、Illuminaのigenomeも結構使われるので便利<br>ヒトデータに限って言えば、NCBIとUCSCはかなり似ている。Ensemblはちょっと毛色が違う。hg38はNCBIとUCSCは統一された。RNAはEnsemblが力を入れている印象 | |
| 解凍は、tar zxvf *.tar.gzのオプションが基本。vオプションはあってもなくてもよい | |
| Ensembl, RefSeq, UCSCでどのアノテーションを使うのがいいかに関する評価研究論文。Zhao and Zhang, BMC Genomics, 2015 | |
| GRCh38はデコイ配列を含む。但し、ゲノム解析系はアノテーション情報がついてこれてないので移行しづらい。<br>GRCh38は、1000人ゲノムプロジェクトのコンセンサスを元に作成 | |
| スライド15はリンクのlなどと書いてあるが、実体のファイルがある。 | |
| スライド17で「ln -s ../../WholeGenomeFasta/genome.fa」と書いているが、これは省略形<br>「ln -s ../../WholeGenomeFasta/genome.fa ./genome.fa」が正式<br>同じファイル名の場合は省略してよくて、カレントディレクトリ上に作成する場合も省略してよい。 | |
| headとtail両方で確認。tailで確認するのは、コピー時におしりの部分が欠けることがあるので、それが分かる。 | |
| fastq_quality_filter (ver. 0.0.14)コマンドで「-Q 33」はつけなくても大丈夫そう。 | |
| スライド32で、「bwa mem 2> hoge」とすることでマニュアルをhogeに書きだしてくれる。 | |
| samtools viewのデフォルトは、bamをsamにするが、-Sを指定するとinputがsamであることを意味し、 -bを指定するとoutputがbamであることを意味する。 | |
| 実用上は、パイプを駆使して無駄に多くのファイルを作成しない。<br>「-」はパイプでわたってきた入力ファイルの意味。 「samtools sort」はメモリを食うのでsortだけやらないこともある。<br>ちなみにこの次のsamtools indexは独立してやる。<br>理由はこれまでパイプでつなぐと、bamファイルがない状態でbamのindexファイル(*.bam.bai)だけができてしまうことになるから。 | bwa mem -R "@RG¥tID:1K_ERR038793_1¥tSM:ERR038793¥tPL:Illumina"<br>/home/iu/Desktop/amelieff/Scerevisiae/BWAIndex/genome.fa<br>1K_ERR038793_1_qual.fastq \| samtools view -Sb - \| samtools sort - hoge_sorted |
| スライド37で3列目のマップされたリードをカウントする場合は$3、4列目のマップされなかったリード数は$4にすればよい。 | samtools idxstats 1K_ERR038793_1_qual_sorted.bam > tmp<br>awk '{a += $3} END {print a}' tmp |
| おまけ?!tviewオプションでIGVのような簡単なマッピング結果を眺めることもできる。 | samtools tview 1K_ERR038793_1_qual_sorted.bam |
| bwaのマッピング結果は必ずしも信頼できるわけではないので、 GATKを使って怪しい領域を抜き出して丁寧にアラインメントをやり直す必要がる。<br>これをre-alignmentという。 | |
| スライド38。「-glm」がgenotypingのアルゴリズム<br>-glmをいれなくてもデフォルトのSNPはやってくれる。<br>-glm BOTHでSNPとINDELの両方を同時にやってくれる。<br>別々にやって別ファイルに書きだすのもアリ。 | |
| GATK実行時にエラーが見える(Error Response)が、こういう場合はGATK Support Forum。 | |
| dbSNPに登録されている場合はそれを利用してよりよい結果を返すこともできる。<br>また複数サンプルを同時に入力として与えて検出の信頼度を上げることもできるらしいが、アメリエフさんはまだやったことがないらしい。 | |
| スライド41は94個の変異となったが、実際にやると100個となった。 | |
| IGVで読み込んで、vcfファイルの水色がhomozygoteで青がheterozygote | |
| スライド47。「GATK VariantFiltration」で検索。 | |