Whole Exon Sequencing simulation and validation pipeline - By David YunTe Lin

Required modules:

seq/samtools/1.1 seq/bowtie/2.2.4 dev/python/2.7.6 seq/bwa/0.7.8 dev/java/jdk1.7

To simulate Whole exome Sequencing:

1. To prepare input files:

Use sametools to index your input file (e.g. samtools faidx ref.fa)

Generate 2bit file (e.g. faToTwoBit ref.fa ref.2bit)

2. Run Wessim\_1: (in ideal target mode a genomic coordinates file is needed)

python Wessim1.py -R ref.fa -B target.bed [options]

Parameters:

-R FILE faidx-indexed (R)eference genome FASTA file

-B FILE Target region .(B)ED file

Parameters for exome capture:

-f INT mean (f)ragment size. this corresponds to insert size when sequencing in paired-end mode. [200]

-d INT standard (d)eviation of fragment size [50]

-m INT (m)inimum fragment length [read\_length + 20]

-x INT slack margin of the given boundaries [0] (only for Wessim1)

-w INT penalty (w)eight for indel in the hybridization [2] (only for Wessim2)

Parameters for sequencing:

-p generate paired-end reads [single]

-n INT total (n)umber of reads

-l INT read (l)ength (bp)

-M FILE GemSim (M)odel file (.gzip)

-t INT number of (t)hreaded subprocesses [1]

Output options:

-o FILE (o)utput file header. ".fastq.gz" or ".fastq" will be attached automatically. Output will be splitted into two files in paired-end mode

-z compress output with g(z)ip [false]

-q INT (q)uality score offset [33]

-v (v)erbose; print out intermediate messages.

To simulate WGS:

To align reads against reference genome:

To align reads using BWA:

1. Index the reference fasta:

bsub -q priority -W 12:00 bwa index -a bwtsw -p chr1 chr1.fa

bsub -q priority -W 12:00 bwa index -a bwtsw -p hg19 chr1.fa

2. Alignment: (Align seperately for paired-ends files)

Example: bwa aln -t 8 -f input.sai hg19 input.fastq

bsub -n 4 -q priority -W 12:00 -e errors\_BWA.%J -o output\_BWA.%J bwa aln -t 4 -f Wes\_1\_cnv\_2.5MX\_A1\_1.sai /groups/nabavi/David/data/chr1\_hg19/Hg19\_chr1 Wes\_1\_cnv\_2.5MX\_A1\_1.fastq

3. Convert to SAM:

Usage :bwa sampe [options] <prefix> <in1.sai> <in2.sai> <in1.fq> <in2.fq>

bwa sampe -f out.sam ref input1.sai input2.sai input1.fq input2.fq

bsub -n 2 -q priority -W 12:00 -e errors\_BWA.%J -o output\_BWA.%J bwa sampe -r "@RG\tID:Con\_Chr1\tSM:Control\_Chr1" -f Sim\_cnv.sam /groups/nabavi/David/data/chr1\_hg19/Hg19\_chr1 sim\_cnv\_R1.sai sim\_cnv\_R2.sai sim\_cnv\_R1.fastq sim\_cnv\_R2.fastq

bwa sampe $REF $output1 $output2 $input1.gz $input2.gz | samtools view -bSu - | samtools sort - $output.bam.prefix

To align using Bowtie2

1. Index fatsa files:

To index:

bsub -q priority -W 12:00 -e errors\_bowite2-build.%J -o output\_bowtie2-build.%J bowtie2-build /groups/nabavi/David/data/chr1\_hg19/chr1.fa hg19

2. Alignment:

bsub -n 8 -q priority -W 12:00 -e errors\_bowite2\_cnv.%J -o output\_bowtie2\_cnv.%J bowtie2 -p 8 -x /groups/nabavi/David/data/chr1\_hg19/hg19 -1 /groups/nabavi/David/outputs/wessim\_output/wessim\_cnv\_allel\_2\_2.5M\_100x\_1.fastq.gz -2 /groups/nabavi/David/outputs/wessim\_output/wessim\_cnv\_allel\_2\_2.5M\_100x\_2.fastq.gz -S cnv\_testAlign.sam

bsub -n 8 -q priority -W 12:00 -e errors\_bowite2\_cnv.%J -o output\_bowtie2\_cnv.%J bowtie2 -p 8 -x /groups/nabavi/David/data/chr1\_hg19/hg19 -S wholeGenome\_ref.sam

3. Convert to BAM, sorting and indexing:

samtools view -b -S <SAM\_file> -o file.bam

bsub -q priority -W 12:00 -e errors\_sam.%J -o output\_sam.%J samtools view -b -S SInC\_sim\_cnv.sam -o SInC\_sim\_cnv.bam

samtools sort <BAM\_file> prefix

bsub -q priority -W 12:00 -e errors\_sam.%J -o output\_sam.%J samtools sort Merged\_Control.bam Merged\_Control\_sorted

samtools index <sorted bam file>

4. Merge BAM files from 2 alleles

bsub -q priority -W 12:00 -e errors\_BWA.%J -o output\_BWA.%J samtools merge Merged\_cnv\_Chr1.bam A1.bam A2.bam