Bio Medical Data Analysis

NGS Data Analysis

1. Data

A short explaination about data:-

The Sample data taken from the prostate cancer tissue and healthy prostate tissue data(2011) Recurrent chimeric RNAs enriched in human prostate cancer identified by deep sequencing using Illumina Sequencing Platform. https://www.ncbi.nlm.nih.gov/sra/?linkname=pubmed_sra&from_uid=21571633

1.1 Sequence Read Archive file

https://trace.ncbi.nlm.nih.gov/Traces/sra/sra.cgi?view=search_seq_name&exp=SRX000001,SRX000002,SRX000003

Comparative transcriptomic analysis of prostate cancer and matched normal tissue using RNA-seq

SRR057629.sra SRR057630.sra SRR057649.sra SRR057650.sra

Prefetch is a part of SRA toolkit this command extract the files from compressed SRA format output stored at /home/mlsi/RNASeq/SRA folder.

In [ ]:
```
prefetch -v "SRA"
```

1.2 Creating .fastq files from .sra files

It is a file which stores a sequance of DNA in letters and quality of it.

In [ ]:
```
fastq-dump --outdir --split-files
```

1.3 Sample Preparation

The test data consists of two commercially available RNA samples: Universal Human Reference(UHR) and Human Brain Reference (HBR).The UHR is total RNA isolated from a diverse set of 10 cancer cell lines. The HBR is total RNA isolated from the brains of 23 Caucasians, male and female, of varying age but mostly 60-80 years old.

> Data File:- UHR + ERCC Spike-In Mix1, Replicate 1 UHR + ERCC Spike-In Mix1, Replicate 2 UHR + ERCC Spike-In Mix1, Replicate 3 HBR + ERCC Spike-In Mix2, Replicate 1 HBR + ERCC Spike-In Mix2, Replicate 2 HBR + ERCC Spike-In Mix2, Replicate 3

In [ ]:
```
wget http://genomedata.org/rnaseq-tutorial/HBR_UHR_ERCC_ds_5pc.tar
```

> To extract the data out of .tar file next step is necessary

In [ ]:
```
tar -xvf HBR_UHR_ERCC_ds_5pc.tar
```

A Term Spike-in is used to bind to a DNA molecule with a matching sequence, known as control checking. Spike-In Control Mixes to each sample. The spike-in consists of 92 transcripts that are present in known concentrations across a wide abundance range (from very few copies to many copies). This range allows us to test the degree to which the RNA-seq assay (including all laboratory and analysis steps) accurately reflects the relative abundance of transcript species within a sample.

2. Creating-Genome-Index

Genome indexing is one type of pre-processing to compress the size of text and to make queries fast.

Here we use chromosome 22 with spike in information.

> In this chr22_with_ERCC92.fa file we have the sequence data of the chromosome 22. It's necessary to merge it with an another file format to create a sort of a mapping guide, inorder to map the sample data with the reference file. This is called indexing.

In [ ]:
```
cd /home/mlsi/GenomeIndices
```

2.1 First 10 lines of file

It can be see that the first 10 lines of the file consists of sequences of 'N's.

In [8]:
```
head chr22_with_ERCC92.fa
```

```
>22 dna_sm:chromosome chromosome:GRCh38:22:1:50818468:1 REF
NNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNN
NNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNN
NNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNN
NNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNN
NNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNN
NNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNN
NNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNN
NNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNN
NNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNN
```

2.2

2.3 10 Lines from approx the middle of this file.

interesting to note the presence of both uppercase and lowercase sequences.

In [11]:
```
head -n 420000 chr22_with_ERCC92.fa | tail
```

```
TGCGTGTCTtctacatgaagaaagtcagtctgtgagttctgtgacggcatcctatggaaa
cggttgggtggatttgcgcccaagttggagagtatggttagcgtgatctgacttaaaact
taggccacatggcacatggaaACTGGAGAGTGTCTGGGGTCCCTCACTTTAAGAACTAGA
ATTTTTCCAAAGGTTGCGTGTGACTCAATAGGAAATGGTGACCCCAAGAAACCAAAGAAA
GGGAGTTGGTGGTTGTGGGGAGACACGGGAGAGCTGCACACCGTGCCCCTTGCCTGGCGC
TGCTGGGCTTCCCACAAAGCCAGTGAGGTTCATCTGTCCCCAGGAAGTCATCAGAGACGG
TGGCTGAGGAGACAGGCACTGGGGGCCTCCGTGGACCCGCTATTGGAGTGAAGTGGCCTG
GATGGAATTGGGTGGACAGGGTGCTGCCTCAGCAGGCTGGGATGTGCTGGGCTGTCAGAA
AGTGCTCATAAATAGCATGCCCCTTGCGAAGACAGGCGAGTCTGCAGACGGCCGTGGCTC
CTCTGTTCTTCCCGAGGCTACAGCAACAGGTTCATTCTGGGATGGGTTGGGGGTGGGGAG
```

2.4 Soft masking

The term Soft masking is a masking of low complexity/repetitive elements, which cause problems for search and clustering algorithm. A lot of the sequence in genomes are repetitive.

> Human genome has (at least) two-third repetitive elements. These repetitive elements are soft-masked by converting the upper case letters to lower case.

3. Get the annotation

Here we use annotations obtained from Ensembl (Homo_sapiens.GRCh38.86.gtf.gz) for chromosome 22 only.

> Annotation files are in GFF or GTF file format. In our protocol .gtf file of chromosome 22 is used.

In [ ]:
```
wget http://genomedata.org/rnaseq-tutorial/annotations/GRCh38/chr22_with_ERCC92.gtf
```

3.1 Show how many unic IDs in the .gtf file

perl command-line helps to display the number of unique gene IDs present in .gtf file

In [12]:
```
perl -ne 'if ($_ =~ /(gene_id\s\"ENSG\w+\")/){print "$1\n"}' chr22_with_ERCC92.gtf | sort | uniq | wc -l
```

```
1318
```

> In above Command perl -ne " will execute the code between single quotes, on the .gtf file, line-by-line And print the output of total unic IDs are in .gtf file in our case it found 1318.

3.2 Show Structure of a single transcript (ENST00000342247) in GTF format

In [ ]:
```
grep ENST00000342247 chr22_with_ERCC92.gtf | less -p "exon\s" -S
```

4. Generate Index

4.1 Why do we index

The reason behind Generating Indexe is used to quickly locate data, that allows multiple libraries to be pooled and sequence together.

Generating index is done by a tool called STAR. STAR is an aligner designed to specifically address many of the challenges of RNA-seq data mapping using a strategy to account for spliced alignments.

The basic options to generate genome indices using STAR are as follows:

In [ ]:
```
STAR --runMode genomeGenerate --runThreadN 2 --genomeDir ~/GenomeIndices
--genomeFastaFiles ~/GenomeIndices/chr22_with_ERCC92.fa --sjdbGTFfile ~/
GenomeIndices/chr22_with_ERCC92.gtf --sjdbOverhang 99
```

> In the next step we can filter out just the information about chromosome sizes from the above created .fa file using the cut function.

In [23]:
```
samtools faidx ~/GenomeIndices/chr22_with_ERCC92.fa
```

4.2 Creating a file with chromosome sizes

In [24]:
```
cut -f1,2 ~/GenomeIndices/chr22_with_ERCC92.fa.fai > ~/GenomeIndices/hg3
8_chr22.chromosome.sizes
```

4.3 Generating a .fa file only containing the information of chr22 without the ERCC2 spike ins!

In [25]:
```
cat chr22_with_ERCC92.fa | perl -ne 'if ($_ =~ /\>22/){$chr22=1}; if ($_
=~ /\>ERCC/){$chr22=0}; if ($chr22){print "$_";}' > chr22_only.fa
```

5. RNA-Seq Data

In [26]:
```
cd /home/mlsi/RNASeq/rawData
```

5.1 Below Command helps to view the first two read records of a file. As we can see above in output.

In [33]:
```
zcat UHR_Rep1_ERCC-Mix1_Build37-ErccTranscripts-chr22.read1.fastq.gz | h
ead -2
```

```
@HWI-ST718_146963544:6:1213:8996:10047/1
CTTTTTTATTTTTGTCTGACTGGGTTGATTCAAAGGTCTGGTCTTTGAGCTCTTAAATTAGTTCTTCTATTT
GGCCTAGTCTGTTGCTAAGGCTGCCAAC

gzip: stdout: Broken pipe
```

5.2 Command below shows total reads are in the first library.

As we can see result below there are 227392 reads in first library.

In [34]:
```
zcat UHR_Rep1_ERCC-Mix1_Build37-ErccTranscripts-chr22.read1.fastq.gz | g
rep -P "^\@HWI" | wc -l
```

```
227392
```

6. FastQC

FastQC provides a simple way to do some quality control checks on raw sequence data coming from high throughput sequencing pipelines. FastQC aims to provide a QC report which can spot

problems which originate either in the sequencer or in the starting library material.

6.1 Performing quality check

The below command shows an output of quality check with parallel command.

In [36]:
```
find . -name "*.fastq.gz" | parallel fastqc -o {//}/ {}
```

```
Analysis complete for HBR_Rep3_ERCC-Mix2_Build37-ErccTranscripts-chr22.r
ead2.fastq.gz
Started analysis of HBR_Rep3_ERCC-Mix2_Build37-ErccTranscripts-chr22.rea
d2.fastq.gz
Approx 5% complete for HBR_Rep3_ERCC-Mix2_Build37-ErccTranscripts-chr22.
read2.fastq.gz
Approx 10% complete for HBR_Rep3_ERCC-Mix2_Build37-ErccTranscripts-chr2
2.read2.fastq.gz
Approx 15% complete for HBR_Rep3_ERCC-Mix2_Build37-ErccTranscripts-chr2
2.read2.fastq.gz
Approx 20% complete for HBR_Rep3_ERCC-Mix2_Build37-ErccTranscripts-chr2
2.read2.fastq.gz
Approx 25% complete for HBR_Rep3_ERCC-Mix2_Build37-ErccTranscripts-chr2
2.read2.fastq.gz
Approx 30% complete for HBR_Rep3_ERCC-Mix2_Build37-ErccTranscripts-chr2
2.read2.fastq.gz
Approx 35% complete for HBR_Rep3_ERCC-Mix2_Build37-ErccTranscripts-chr2
2.read2.fastq.gz
Approx 40% complete for HBR_Rep3_ERCC-Mix2_Build37-ErccTranscripts-chr2
2.read2.fastq.gz
Approx 45% complete for HBR_Rep3_ERCC-Mix2_Build37-ErccTranscripts-chr2
2.read2.fastq.gz
Approx 50% complete for HBR_Rep3_ERCC-Mix2_Build37-ErccTranscripts-chr2
2.read2.fastq.gz
Approx 55% complete for HBR_Rep3_ERCC-Mix2_Build37-ErccTranscripts-chr2
2.read2.fastq.gz
Approx 60% complete for HBR_Rep3_ERCC-Mix2_Build37-ErccTranscripts-chr2
2.read2.fastq.gz
Approx 65% complete for HBR_Rep3_ERCC-Mix2_Build37-ErccTranscripts-chr2
2.read2.fastq.gz
Approx 70% complete for HBR_Rep3_ERCC-Mix2_Build37-ErccTranscripts-chr2
2.read2.fastq.gz
Approx 75% complete for HBR_Rep3_ERCC-Mix2_Build37-ErccTranscripts-chr2
2.read2.fastq.gz
Approx 80% complete for HBR_Rep3_ERCC-Mix2_Build37-ErccTranscripts-chr2
2.read2.fastq.gz
Approx 85% complete for HBR_Rep3_ERCC-Mix2_Build37-ErccTranscripts-chr2
2.read2.fastq.gz
Approx 90% complete for HBR_Rep3_ERCC-Mix2_Build37-ErccTranscripts-chr2
2.read2.fastq.gz
Approx 95% complete for HBR_Rep3_ERCC-Mix2_Build37-ErccTranscripts-chr2
2.read2.fastq.gz
```

6.2 Tool for perform Quality check on multiple fastqc output samples and get one quality report.

A 'multiqc .' can be used to perform Quality check on multiple samples. It parses relevant information from these and generates a single stand-alone HTML report file.

7. Adapter Trimming

Adapter trimmimg can be used for the removal of adapter sequences from the 3’ ends of reads. Adapter sequences should be removed from reads because they interfere with downstream analyses.

In [37]:
```
cd ~/RNASeq/rawDataTrim
```

In [ ]:
```
wget http://genomedata.org/rnaseq-tutorial/illumina_multiplex.fa
```

7.1 Perform Adapter trimming with all *.fastq.gz files

In [39]:
```
flexbar --adapter-min-overlap 9 --adapter-trim-end RIGHT --adapters ~/RN
ASeq/rawDataTrim/illumina_multiplex.fa --pre-trim-left 15 --max-uncalled
325 --min-read-length 27 --threads 4 --zip-output GZ --reads ~/RNASeq/ra
wData/HBR_Rep1_ERCC-Mix2_Build37-ErccTranscripts-chr22.read1.fastq.gz --
reads2 ~/RNASeq/rawData/HBR_Rep1_ERCC-Mix2_Build37-ErccTranscripts-chr2
2.read2.fastq.gz --target ~/RNASeq/rawDataTrim/HBR_1
```

In [41]:
```
flexbar --adapter-min-overlap 9 --adapter-trim-end RIGHT --adapters ~/RN
ASeq/rawDataTrim/illumina_multiplex.fa --pre-trim-left 15 --max-uncalled
325 --min-read-length 27 --threads 4 --zip-output GZ --reads ~/RNASeq/ra
wData/HBR_Rep2_ERCC-Mix2_Build37-ErccTranscripts-chr22.read1.fastq.gz --
reads2 ~/RNASeq/rawData/HBR_Rep2_ERCC-Mix2_Build37-ErccTranscripts-chr2
2.read2.fastq.gz --target ~/RNASeq/rawDataTrim/HBR_2
```

In [42]:
```
flexbar --adapter-min-overlap 9 --adapter-trim-end RIGHT --adapters ~/RN
ASeq/rawDataTrim/illumina_multiplex.fa --pre-trim-left 15 --max-uncalled
325 --min-read-length 27 --threads 4 --zip-output GZ --reads ~/RNASeq/ra
wData/HBR_Rep3_ERCC-Mix2_Build37-ErccTranscripts-chr22.read1.fastq.gz --
reads2 ~/RNASeq/rawData/HBR_Rep3_ERCC-Mix2_Build37-ErccTranscripts-chr2
2.read2.fastq.gz --target ~/RNASeq/rawDataTrim/HBR_3
```

In [43]:
```
flexbar --adapter-min-overlap 7 --adapter-trim-end RIGHT --adapters ~/RN
ASeq/rawDataTrim/illumina_multiplex.fa --pre-trim-left 15 --max-uncalled
325 --min-read-length 27 --threads 4 --zip-output GZ --reads ~/RNASeq/ra
wData/UHR_Rep1_ERCC-Mix1_Build37-ErccTranscripts-chr22.read1.fastq.gz --
reads2 ~/RNASeq/rawData/UHR_Rep1_ERCC-Mix1_Build37-ErccTranscripts-chr2
2.read2.fastq.gz --target ~/RNASeq/rawDataTrim/UHR_1
```

In [44]:
```
flexbar --adapter-min-overlap 9 --adapter-trim-end RIGHT --adapters ~/RN
ASeq/rawDataTrim/illumina_multiplex.fa --pre-trim-left 15 --max-uncalled
325 --min-read-length 27 --threads 4 --zip-output GZ --reads ~/RNASeq/ra
wData/UHR_Rep2_ERCC-Mix1_Build37-ErccTranscripts-chr22.read1.fastq.gz --
reads2 ~/RNASeq/rawData/UHR_Rep2_ERCC-Mix1_Build37-ErccTranscripts-chr2
2.read2.fastq.gz --target ~/RNASeq/rawDataTrim/UHR_2
```

In [45]:
```
flexbar --adapter-min-overlap 9 --adapter-trim-end RIGHT --adapters ~/RN
ASeq/rawDataTrim/illumina_multiplex.fa --pre-trim-left 15 --max-uncalled
325 --min-read-length 27 --threads 4 --zip-output GZ --reads ~/RNASeq/ra
wData/UHR_Rep3_ERCC-Mix1_Build37-ErccTranscripts-chr22.read1.fastq.gz --
reads2 ~/RNASeq/rawData/UHR_Rep3_ERCC-Mix1_Build37-ErccTranscripts-chr2
2.read2.fastq.gz --target ~/RNASeq/rawDataTrim/UHR_3
```

The all above commands are performing adapter Trimming in HBR_1, HBR_2, HBR_3, UHR_1, UHR_2, UHR_3.respectivly. https://support.illumina.com/bulletins/2016/04/adapter-trimming-why-are-adapter-sequencestrimmed-from-only-the–ends-of-reads.html

8. STAR Alignment

8.1 Star Mapping Command

In the next step we conduct mapping , using the SampleNames.text as the output file type.

In [ ]:
```
cat ~/RNASeq/rawDataTrim/SampleNames.txt | parallel -j 1 "mkdir ~/RNASe
q/mapping/{};  cd ~/RNASeq/mapping/{}; STAR --genomeDir ~/GenomeIndices
 --readFilesIn ~/RNASeq/rawDataTrim/{}_1.fastq.gz ~/RNASeq/rawDataTrim/
{}_2.fastq.gz --outSAMattributes All --runThreadN 2 --readFilesCommand z
cat --outStd SAM | samtools sort > ~/RNASeq/mapping/{}/{}.bam"
```

STAR is an aligner designed to specifically address many of the challenges of RNA-seq data mapping using a strategy to account for spliced alignments. To determine where on the human genome our reads originated from, we will align our reads to the reference genome using STAR.

In [48]:
```
parallel -j 1 echo {1}_{2} ::: UHR HBR ::: 1 2 3 > SampleNames.txt
```

> In above command instead of performing mapping of 6 samples in 6 differnt steps, we can use parallel command to parallel the process. For doing so , we create a file named SampleNames.txt containing the names of the samples that needs to be mapped.

9. Count Table

9.1 Estimating the abudance for a single sample

In [6]:
```
featureCounts -a ~/GenomeIndices/chr22_with_ERCC92.gtf -g gene_name -o /
home/mlsi/RNASeq/countTable/featureCounts.txt ~/RNASeq/mapping/*/*.bam
```

```
        ==========      _____ _    _ ____  _____  ______          _____
        =====          / ____| |  | |  _ \|  __ \|  ____|   /\   |  __ \
          =====       | (___ | |  | | |_) | |__) | |__     /  \  | |__) |
            ====       \___ \| |  | |  _ <|  _  /|  __|   / /\ \ |  ___/
              ====     ____) | |__| | |_) | | \ \| |____ / ____ \| |
        ==========    |_____/ \____/|____/|_|  \_\______/_/    \_\_|
        v1.6.4

//========================== featureCounts setting ===========================
======\\
||
||
||             Input files : 6 BAM files
||
||                           P HBR_1.bam
||
||                           P HBR_2.bam
```

```
||                              P HBR_2.bam
||
||                              P HBR_3.bam

||
||                              P UHR_1.bam
||
||                              P UHR_2.bam
||
||                              P UHR_3.bam
||
||
||
||                Output file : featureCounts.txt
||
||                    Summary : featureCounts.txt.summary
||
||                 Annotation : chr22_with_ERCC92.gtf (GTF)
||
||        Dir for temp files : /home/mlsi/RNASeq/countTable
||
||
||
||                    Threads : 1
||
||                      Level : meta-feature level
||
||                Paired-end : no
||
||        Multimapping reads : not counted
||
|| Multi-overlapping reads : not counted
||
||   Min overlapping bases : 1
||
||
||
\\========================================================================
======//

//================================= Running ============================
======\\
||
||
|| Load annotation file chr22_with_ERCC92.gtf ...
||
||    Features : 26155
||
||    Meta-features : 1378
||
||    Chromosomes/contigs : 93
||
||
||
|| Process BAM file HBR_1.bam...
||
||    Paired-end reads are included.
||
||    Assign alignments to features...
||
||    Total alignments : 221532
||
||    Successfully assigned alignments : 174668 (78.8%)
||
||    Running time : 0.00 minutes
||
||
||
|| Process BAM file HBR_2.bam...
||
||    Paired-end reads are included.
||
||    Assign alignments to features...
||
||    Total alignments : 266446
||
||    Successfully assigned alignments : 211695 (79.5%)
||
||    Running time : 0.00 minutes
||
||
||
|| Process BAM file HBR_3.bam...
||
||    Paired-end reads are included.
||
||    Assign alignments to features...
||
||    Total alignments : 241748
||
||    Successfully assigned alignments : 190146 (78.7%)
||
||    Running time : 0.00 minutes
||
```

```
||
||
||
|| Process BAM file UHR_1.bam...
||
||    Paired-end reads are included.
||
||    Assign alignments to features...
||
||    Total alignments : 449976
||
||    Successfully assigned alignments : 351993 (78.2%)
||
||    Running time : 0.01 minutes
||
||
||
|| Process BAM file UHR_2.bam...
||
||    Paired-end reads are included.
||
||    Assign alignments to features...
||
||    Total alignments : 309346
||
||    Successfully assigned alignments : 224926 (72.7%)
||
||    Running time : 0.01 minutes
||
||
||
|| Process BAM file UHR_3.bam...
||
||    Paired-end reads are included.
||
||    Assign alignments to features...
||
||    Total alignments : 352560
||
||    Successfully assigned alignments : 277100 (78.6%)
||
||    Running time : 0.01 minutes
||
||
||
||
||
|| Summary of counting results can be found in file "/home/mlsi/RNASeq/c
ount  ||
|| Table/featureCounts.txt.summary"
||
||
||
\\=================================================================
======//
```

9.2 To check how many columns and their names are in featureCountsTable we run the below command

And we can see there are 13 Columns are in the fetureCountsTable which are Geneid, Chr, Start, End, Strand, Length /home/mlsi/RNASeq/mapping/HBR_1/HBR_1.bam
/home/mlsi/RNASeq/mapping/HBR_2/HBR_2.bam
/home/mlsi/RNASeq/mapping/HBR_3/HBR_3.bam
/home/mlsi/RNASeq/mapping/UHR_1/UHR_1.bam
/home/mlsi/RNASeq/mapping/UHR_2/UHR_2.bam
/home/mlsi/RNASeq/mapping/UHR_3/UHR_3.bam .

In [ ]:
```
cat featureCounts.txt | head
```

9.3 To Find the sequences with most hits in the sample HBR_1.bam!

In [ ]:
```
cat featureCounts.txt | sort -rn -k 7 | head
```

10. Data Visualization

10.1 Createing the bedgraph file out of the .bam file

Data visualization is an essential component of genomic data analysis. However, the size and diversity of the data sets produced by today's sequencing and array-based profiling methods present major challenges to visualization tools. The Integrative Genomics Viewer (IGV) is a high-performance viewer that efficiently handles large heterogeneous data sets, while providing a smooth and intuitive user experience at all levels of genome resolution. A key characteristic of IGV is its focus on the integrative nature of genomic studies, with support for both array-based and next-generation sequencing data, and the integration of clinical and phenotypic data.

In [13]:
```
cat /home/mlsi/RNASeq/rawDataTrim/SampleNames.txt| parallel --eta --verb
ose "cd ~/RNASeq/mapping/{} ; find . -name "{}.bam"; samtools index {}.b
am"
```

```
cd ~/RNASeq/mapping/UHR_1 ; find . -name UHR_1.bam; samtools index UHR_
1.bam

Computers / CPU cores / Max jobs to run
1:local / 1 / 1

Computer:jobs running/jobs completed/%of started jobs/Average seconds to
complete
ETA: 0s Left: 6 AVG: 0.00s  local:1/0/100%/0.0s ./UHR_1.bam
ETA: 0s Left: 5 AVG: 0.00s  local:0/1/100%/0.0s cd ~/RNASeq/mapping/UHR_
2 ; find . -name UHR_2.bam; samtools index UHR_2.bam
ETA: 5s Left: 5 AVG: 1.00s  local:1/1/100%/1.0s ./UHR_2.bam
ETA: 3s Left: 4 AVG: 0.50s  local:0/2/100%/0.5s cd ~/RNASeq/mapping/UHR_
3 ; find . -name UHR_3.bam; samtools index UHR_3.bam
ETA: 2s Left: 4 AVG: 0.50s  local:1/2/100%/0.5s ./UHR_3.bam
ETA: 1s Left: 3 AVG: 0.33s  local:0/3/100%/0.3s cd ~/RNASeq/mapping/HBR_
1 ; find . -name HBR_1.bam; samtools index HBR_1.bam
ETA: 1s Left: 3 AVG: 0.67s  local:1/3/100%/0.7s ./HBR_1.bam
ETA: 1s Left: 2 AVG: 0.50s  local:0/4/100%/0.5s cd ~/RNASeq/mapping/HBR_
2 ; find . -name HBR_2.bam; samtools index HBR_2.bam
ETA: 1s Left: 2 AVG: 0.50s  local:1/4/100%/0.5s ./HBR_2.bam
ETA: 0s Left: 1 AVG: 0.40s  local:0/5/100%/0.4s cd ~/RNASeq/mapping/HBR_
3 ; find . -name HBR_3.bam; samtools index HBR_3.bam
ETA: 0s Left: 1 AVG: 0.40s  local:1/5/100%/0.4s ./HBR_3.bam
ETA: 0s Left: 0 AVG: 0.33s  local:0/6/100%/0.3s
```

https://www.ncbi.nlm.nih.gov/pmc/articles/PMC3603213/ The next step would be to load the .bam files into the igv browser. As we know , the samples would have matched to the chromosome 22 of the human genome. The IGV browser provides a zoom in factor which lets the user to closely compare the sample.

10.2 Create the coverage files (bedgraph) for all BAM files

BigWig is a file format for display of dense, continuous data in a genome browser track, created by conversion from Wiggle (WIG) format. BigWig files are a compressed, indexed, binary format for genome-wide signal data for calculations or experiments. bwtool is a tool designed to read bigWig files rapidly and efficiently. If your data is sparse or contains elements of varying sizes, use the bedGraph format instead of the wiggle format. If you have a very large bedGraph data set, you can convert it to the bigWig format using the bedGraphToBigWig program. https://academic.oup.com/bioinformatics/article/30/11/1618/282756 In the nextstep we convert the .bam files to bedgraph

In [ ]:
```
cat /home/mlsi/RNASeq/rawDataTrim/SampleNames.txt| parallel --eta --verb
ose "cd ~/RNASeq/mapping/{} ; bedtools genomecov -ibam {}.bam -g ~/Genom
eIndices/hg38_chr22.chromosome.sizes -bg | sort -k1,1 -k2,2n > {.}.bedgr
aph"
```

10.3 Here we generate all bigwig coverages from bedgraphs.

In [ ]:
```
cat /home/mlsi/RNASeq/rawDataTrim/SampleNames.txt| parallel --eta --verb
ose "cd ~/RNASeq/mapping/{}; bedGraphToBigWig {}.bedgraph ~/GenomeIndice
s/hg38_chr22.chromosome.sizes /home/mlsi/RNASeq/bigWig/{}.bigwig"`
```