**Developmental regulation of human cortex transcription**

1. STUDY DESIGN

Collecting the data:

Because I want to compare the differences between genes of fetus and adult, I collect randomly 10 samples, 5 of which are fetus and other are adult. In each aged classification, I also choose randomly 2 females and 3 males.

The RNA-seq of these samples are collected in the date of publication 2014-11-13

1. SUMMARY CHOICES:

**Processing raw data of original publication in Galaxy website:** [**Galaxy (usegalaxy.org)**](https://usegalaxy.org/) **to get tidy data (count table)**

1. Get data

* Choose “Download and Extract Reads in FASTA/Q” in GET DATA section. Then type SRR accession and choose gzip compressed before clicking execute.
* Choose “FASTQ splitter” in FASTA/FASTQ in order to split the initial fastq read into 2 reads

1. Quality control

* Choose “FastQC” in FASTQ Quality Control. Then, chose multiple datasets including read 1 and read 2 and execute. This is a list of links to see quality control of each read:
* After checking QC of each read, cut adapt every read and then use the adapted reads to prepare for alignment:

+ Choose “Cutadapt” in FASTA/FASTQ and then choose pair-end reads. Then, put read 1 and read 2 in FASTQ/A file #1 and #2 respectively.

+ Choose “Adapter Options”:

* Maximum error rate: 0.1
* Minimum overlap length: 3

+ Choose Filter Options:

* Minimum length (R1): 20
* Minimum length (R2): 20

+ Choose Read Modification Options:

* Quality cutoff: 20

1. Alignment with HISAT2:

* Choose “HISAT2” in RNA-seq:

+ Source for the reference genome: Use a built-in genome

+ Select a reference genome: Human (Homo sapiens) (b37): hg19

+ Using paired-end and put adapted read 1 and adapted read 2 in FASTQ/A file #1 and #2 respectively.

+ Other parameters are set default and then execute

1. Get feature counts:

* Choose “featureCounts” in RNA-seq:

+ Using bam file, which is an output of HISAT2 and choose unstranded

+ Gene annotation file: featureCounts built-in

+ Built-in genome: hg19

+ Output format: Gene-ID "\t" read-count (MultiQC/DESeq2/edgeR/limma-voom compatible)

+ In Advanced options:

* GFF feature type filter: exon
* GFF gene identifier: gene\_id

1. Get tidy data (count table):

* Merge count data of each sample by using Column Join on Collection
* Using annotateMyIDs to annotate the gene\_id from feature counts:

+ File with IDs: put the file that is merged above including all samples

+ Organism: Human

+ ID Type: Entrez

+ Output columns: ENSEMBL, ENTREZID, SYMBOL

1. THE CODE BOOK
2. **Count table (feature genes)**

|  |  |  |
| --- | --- | --- |
| **Notes** | | |
| Output Created | | 08-AUG-2021 19:40:35 |
| Comments | |  |
| Input | Data | D:\word\bioinformatics\personal project\PRJNA245228\tidy data\count.csv |
| Active Dataset | DataSet1 |
| Filter | <none> |
| Weight | <none> |
| Split File | <none> |
| N of Rows in Working Data File | 28696 |
| Syntax | | CODEBOOK ENTREZID [n] ENSEMBL [n] SYMBOL [n] SRR1554534 [s] SRR1554535 [s] SRR1554568 [s]  SRR1554561 [s] SRR1554567 [s] SRR1554536 [s] SRR1554541 [s] SRR1554539 [s] SRR1554538 [s]  SRR1554537 [s]  /VARINFO POSITION LABEL TYPE MEASURE VALUELABELS MISSING ATTRIBUTES  /OPTIONS VARORDER=VARLIST SORT=ASCENDING MAXCATS=5000  /STATISTICS COUNT PERCENT MEAN STDDEV QUARTILES. |
| Resources | Processor Time | 00:00:00.28 |
| Elapsed Time | 00:00:00.30 |

|  |
| --- |
| **Warnings** |
| Value labels, counts, and/or percents not displayed for the following variables or multilple response sets because the number of unique, valid values exceeds the specified maximum or default maximum of 200:  ENTREZID,ENSEMBL,SYMBOL. |

|  |  |  |
| --- | --- | --- |
| **ENTREZID** | | |
|  | | Value |
| Standard Attributes | Position | 1 |
| Label | Entrez Gene ID |
| Type | String |
| Measurement | Nominal |

|  |  |  |
| --- | --- | --- |
| **ENSEMBL** | | |
|  | | Value |
| Standard Attributes | Position | 2 |
| Label | Ensembl Gene ID |
| Type | String |
| Measurement | Nominal |

|  |  |  |
| --- | --- | --- |
| **SYMBOL** | | |
|  | | Value |
| Standard Attributes | Position | 3 |
| Label | Gene symbol |
| Type | String |
| Measurement | Nominal |

|  |  |  |
| --- | --- | --- |
| **SRR1554534** | | |
|  | | Value |
| Standard Attributes | Position | 4 |
| Label | Sample 1 |
| Type | Numeric |
| Measurement | Scale |
| N | Valid | 28696 |
| Missing | 0 |
| Central Tendency and Dispersion | Mean | 3428.07 |
| Standard Deviation | 25268.711 |
| Percentile 25 | 2.00 |
| Percentile 50 | 223.00 |
| Percentile 75 | 2211.00 |

|  |  |  |
| --- | --- | --- |
| **SRR1554535** | | |
|  | | Value |
| Standard Attributes | Position | 5 |
| Label | Sample 2 |
| Type | Numeric |
| Measurement | Scale |
| N | Valid | 28696 |
| Missing | 0 |
| Central Tendency and Dispersion | Mean | 4677.15 |
| Standard Deviation | 39507.319 |
| Percentile 25 | 4.00 |
| Percentile 50 | 294.00 |
| Percentile 75 | 3279.00 |

|  |  |  |
| --- | --- | --- |
| **SRR1554568** | | |
|  | | Value |
| Standard Attributes | Position | 6 |
| Label | Sample 3 |
| Type | Numeric |
| Measurement | Scale |
| N | Valid | 28696 |
| Missing | 0 |
| Central Tendency and Dispersion | Mean | 5026.94 |
| Standard Deviation | 15012.371 |
| Percentile 25 | 4.00 |
| Percentile 50 | 321.00 |
| Percentile 75 | 4336.50 |

|  |  |  |
| --- | --- | --- |
| **SRR1554561** | | |
|  | | Value |
| Standard Attributes | Position | 7 |
| Label | Sample 4 |
| Type | Numeric |
| Measurement | Scale |
| N | Valid | 28696 |
| Missing | 0 |
| Central Tendency and Dispersion | Mean | 4839.56 |
| Standard Deviation | 28045.323 |
| Percentile 25 | 3.00 |
| Percentile 50 | 289.00 |
| Percentile 75 | 3095.00 |

|  |  |  |
| --- | --- | --- |
| **SRR1554567** | | |
|  | | Value |
| Standard Attributes | Position | 8 |
| Label | Sample 5 |
| Type | Numeric |
| Measurement | Scale |
| N | Valid | 28696 |
| Missing | 0 |
| Central Tendency and Dispersion | Mean | 6447.25 |
| Standard Deviation | 19442.831 |
| Percentile 25 | 6.00 |
| Percentile 50 | 475.00 |
| Percentile 75 | 5667.50 |

|  |  |  |
| --- | --- | --- |
| **SRR1554536** | | |
|  | | Value |
| Standard Attributes | Position | 9 |
| Label | Sample 6 |
| Type | Numeric |
| Measurement | Scale |
| N | Valid | 28696 |
| Missing | 0 |
| Central Tendency and Dispersion | Mean | 2296.34 |
| Standard Deviation | 72557.248 |
| Percentile 25 | .00 |
| Percentile 50 | 89.00 |
| Percentile 75 | 879.00 |

|  |  |  |
| --- | --- | --- |
| **SRR1554541** | | |
|  | | Value |
| Standard Attributes | Position | 10 |
| Label | Sample 7 |
| Type | Numeric |
| Measurement | Scale |
| N | Valid | 28696 |
| Missing | 0 |
| Central Tendency and Dispersion | Mean | 7702.16 |
| Standard Deviation | 25972.923 |
| Percentile 25 | 5.00 |
| Percentile 50 | 443.50 |
| Percentile 75 | 6023.50 |

|  |  |  |
| --- | --- | --- |
| **SRR1554539** | | |
|  | | Value |
| Standard Attributes | Position | 11 |
| Label | Sample 8 |
| Type | Numeric |
| Measurement | Scale |
| N | Valid | 28696 |
| Missing | 0 |
| Central Tendency and Dispersion | Mean | 4137.50 |
| Standard Deviation | 22770.711 |
| Percentile 25 | 4.00 |
| Percentile 50 | 270.00 |
| Percentile 75 | 2887.50 |

|  |  |  |
| --- | --- | --- |
| **SRR1554538** | | |
|  | | Value |
| Standard Attributes | Position | 12 |
| Label | Sample 9 |
| Type | Numeric |
| Measurement | Scale |
| N | Valid | 28696 |
| Missing | 0 |
| Central Tendency and Dispersion | Mean | 6628.34 |
| Standard Deviation | 20076.631 |
| Percentile 25 | 5.00 |
| Percentile 50 | 455.00 |
| Percentile 75 | 5672.00 |

|  |  |  |
| --- | --- | --- |
| **SRR1554537** | | |
|  | | Value |
| Standard Attributes | Position | 13 |
| Label | Sample 10 |
| Type | Numeric |
| Measurement | Scale |
| N | Valid | 28696 |
| Missing | 0 |
| Central Tendency and Dispersion | Mean | 5683.63 |
| Standard Deviation | 17247.397 |
| Percentile 25 | 4.00 |
| Percentile 50 | 330.50 |
| Percentile 75 | 4799.50 |

1. **Sample**

|  |  |  |
| --- | --- | --- |
| **Notes** | | |
| Output Created | | 08-AUG-2021 19:06:18 |
| Comments | |  |
| Input | Data | D:\word\bioinformatics\personal project\PRJNA245228\sample\_data\pheno\_sample.csv |
| Active Dataset | DataSet2 |
| Filter | <none> |
| Weight | <none> |
| Split File | <none> |
| N of Rows in Working Data File | 10 |
| Syntax | | CODEBOOK Run\_sample [n] age.group [n] age [s] sex [n]  /VARINFO POSITION LABEL TYPE MEASURE VALUELABELS MISSING ATTRIBUTES  /OPTIONS VARORDER=VARLIST SORT=ASCENDING MAXCATS=200  /STATISTICS COUNT PERCENT MEAN STDDEV QUARTILES. |
| Resources | Processor Time | 00:00:00.00 |
| Elapsed Time | 00:00:00.01 |

**Variable information of samples**

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| **Run\_sample** | | | | |
|  | | Value | Count | Percent |
| Standard Attributes | Position | 1 |  |  |
| Label | Name of each sample |  |  |
| Type | String |  |  |
| Measurement | Nominal |  |  |
| Valid Values | SRR1554534 |  | 1 | 10.0% |
| SRR1554535 |  | 1 | 10.0% |
| SRR1554536 |  | 1 | 10.0% |
| SRR1554537 |  | 1 | 10.0% |
| SRR1554538 |  | 1 | 10.0% |
| SRR1554539 |  | 1 | 10.0% |
| SRR1554541 |  | 1 | 10.0% |
| SRR1554561 |  | 1 | 10.0% |
| SRR1554567 |  | 1 | 10.0% |
| SRR1554568 |  | 1 | 10.0% |

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| **age.group** | | | | |
|  | | Value | Count | Percent |
| Standard Attributes | Position | 2 |  |  |
| Label | classifying that sample is fetus or adult |  |  |
| Type | String |  |  |
| Measurement | Nominal |  |  |
| Valid Values | adult |  | 5 | 50.0% |
| fetus |  | 5 | 50.0% |

|  |  |  |
| --- | --- | --- |
| **age** | | |
|  | | Value |
| Standard Attributes | Position | 3 |
| Label | the age of each sample |
| Type | Numeric |
| Measurement | Scale |
| N | Valid | 10 |
| Missing | 0 |
| Central Tendency and Dispersion | Mean | 20.447880 |
| Standard Deviation | 22.0883966 |
| Percentile 25 | -.402700 |
| Percentile 50 | 18.058200 |
| Percentile 75 | 41.580000 |

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| **sex** | | | | |
|  | | Value | Count | Percent |
| Standard Attributes | Position | 4 |  |  |
| Label | Gender of each sample |  |  |
| Type | String |  |  |
| Measurement | Nominal |  |  |
| Valid Values | female |  | 4 | 40.0% |
| male |  | 6 | 60.0% |