RNA README: UPDATED 8/11/2020

RNA derived from Novartis Nigerian patients / TCGA FASTA files from 2017

Input

Labeled, processed raw counts: NigerianTCGA\_quants (ALL)

NigerianTCGA\_quants-lnc (LONG NONCODING RNA)

NigerianTCGA\_quants-proteincoding (PROTEIN CODING)

NigerianTCGA\_quants-proteincoding-IHC (PROTEIN CODING)

*Raw-Files: Nigerian-RNA-RSEM-Raw -> Genes/Isoforms separated (Nigerian-RNA-RSEM)  
TCGA-RNA-RSEM-Raw -> Genes/Isoforms separated (TCGA-RNA-RSEM)*

*RSEM: Nigerian-RNA-RSEM*

*TCGA-RNA-RSEM*

Input / Phenotype-Data

TCGA-JSON-parson-2.ipynb -> Maps between old and new legacy TCGA names

TCGA\_legacy\_RNAseq\_metadata.2018-11-28

TCGAethnicity.txt

IHC-PAM50-Correlation -> Map between IHC and PAM50 labels for Life History and Perou projects

WABCS\_RNAseq-WGS\_overlap\_IDs

*Clinical and Pathologic Phenotypes*

Nature Communications WABCS final

WABCS\_Sample\_Phenotype\_v5\_2018-03-07

WABCS-U01\_Samples\_Phenotype\_master2B

BRCA\_ER.tsv

TCGA\_clinical-11092018

***Life History***

-> Life History WGS <-> RNA mapping and TP53/GATA3/silent genome/14qLOH

*-*GATA3\_gene\_dosage\_allgroups\_NAP  
-QUIET\_NIG\_list

TP53\_gene\_dosage\_allgroups\_NAP

WABCS\_RNAseq\_IDs.txt

WABCS\_RNAseq-WGS\_overlap\_IDs

**DE**

Perou Nigeria\_109\_samples PAM50 Calls 9-12-19 UPDATED 9-21-19-PSR

RNAseq-PAM50-allsamples.txt

Output:

HRD / ESR -> Gene-specific analyses requested by Funmi. Analyses are based on minimal normalization for libraries WITHOUT complete differential expression analysis performed.

Life-History-WGS-Validation -> Results for UK collaborative project

Overall-RNA -> Results for UNC collaborative project

Status -> Project update presentations

TCGA-Nigerian-RNAseq

Analysis **(README LOCATED HERE)**

NigerianTCGArawcountsDeSeq2-pc2.RMD

NigerianTCGArawcountsDeSeq2-proteincoding-IHC.RMD

NigerianTCGA-lnc.RMD

NigerianTCGArawcountslimma-voomDE-PAM50.RMD

Code

\*TCGA-JSON-parser-2.ipynb (redundant backup)

\*Proteincodingparse.RMD (redundant backup)

Docs Figures/HTML

Raw-Files:

*raw\_quants.tar.gz*

*Raw-Files:*

*Nigerian-RNA-RSEM-Raw -> Genes/Isoforms separated (Nigerian-RNA-RSEM)  
TCGA-RNA-RSEM-Raw -> Genes/Isoforms separated (TCGA-RNA-RSEM)*

-Raw counts derived from HTSeq (Dominic, completed before my arrival):

*RSEM:*

*Nigerian-RNA-RSEM*

*TCGA-RNA-RSEM*-RSEM produced by <https://github.com/tbrunetti/psychENCODE_data_analysis> (Dominic, completed before my arrival):

-Confirmed all RNA genes were the same across all files

\*Uses code/ NigerianTCGAgenecheck.Rmd

**STEPWISE CHECK AND CODE SCRATCH (DO NOT USE):**

-DeSEQ2: \*Uses code / NigerianTCGArawcountsDeSeq2.Rmd

-Limma/Voom: \*Uses code / NigerianTCGArawcountsVoom.Rmd

**ANNOTATION AND PHENOTYPE INFORMATION**

-Legacy TCGA identifiers and current TCGA identifiers were linked

\*Uses code / TCGA-JSON-parser-2.ipynb + TCGA\_legacy\_RNAseq\_metadata.2018-11-28.json / TCGA\_clinical-11092018.tsv

-RNA genes separated into protein coding and non-protein coding

\*Uses code/proteincodingparse.Rmd

**DEMONSTRATION OF METHOD OF NORMALIZATION OF RNASEQ DATA**

**Project File: NigerianTCGArawcountsDeSeq2-pc2.Rmd**

*OLOPADE LAB:* NigerianTCGA\_quants *-> 219 files, 133 Nigerian and 86 TCGA* with raw counts

**ANNOTATION:**

TCGA Ethnicity: Nature Communications WABCS final: Supplementary data 1e: RNA-seq samples Race

PAM50: Nature Communications WABCS final: Supplementary data 1e: RNA-seq samples PAM50

Batch: WABCS-U01\_Samples\_Phenotype\_master2B

**NO EXCLUSIONS PERFORMED but**

**Unmatched = No annotation provided**

**NFWT = Not found in master table or Jason’s table**

**DUPWT = Duplicate sample per master table (first batch chosen)**

*CHUCK PEROU PROJECT:* NigerianTCGA\_quants-proteincoding *-> 180 files, 96 Nigerian and   
84 TCGA with raw counts*

**Project File: NigerianTCGArawcountsDeSeq2-proteincoding-PAM50.Rmd**

**Output Files: Output/Overall-RNA**

**EXCLUDED:**

2 TCGA samples: 2 samples Excluded due to PCA with abnormal counts compared to other TCGA samples

4637f6a3-9e07-4ac4-be5a-c893a27c8fe2 RNA -> 8785012f-f73e-4d68-87cf-1d804af32782 (black)

8999059b-74b5-4388-afdb-457c9fd24421 RNA -> 8785012f-f73e-4d68-87cf-1d804af32782 (black)

37 Nigerian samples

16 Nigerian samples: Poor coverage, not found or unmatched

|  |  |
| --- | --- |
| **RNAseq Library ID** | **Issues Notes** |
| LIB-008206 | Mismatching Tumor, Low coverage |
| LIB-008196 | Wrong ID |
| LIB-008188 | Wrong ID |
| LIB-006813 | Low coverage |
| LIB-006809 | Low coverage |
| LIB-006808 | Low coverage |
| LIB-006806 | Low coverage |
| LIB-006802 | Low coverage |
| LIB-006799 | Low coverage |
| LIB-006798 | Low coverage |
| LIB-04674 | K562 control |
| LIB-04656 | Low coverage |
| LIB-04646 | Wrong ID |
| LIB-04608 | Low coverage |
| LIB-04590 | Wrong ID |
| LIB-04588 | Wrong ID |

21 duplicates or not included in Perou samples ->

*LIB-03328WT*

*LIB-04588wt*

*LIB-04604wt*

LIB-04607wt

LIB-04609wt

LIB-04622wt

LIB-04623wt

LIB-04626wt

LIB-04629wt

LIB-04630wt

LIB-04660wt

LIB-04661wt

LIB-04662wt

LIB-02352

LIB-02353

LIB-02354

LIB-02355

LIB-02356

LIB-02357

LIB-02358

LIB-02359

LIB-006801

**ANNOTATION:**

TCGA Ethnicity: Nature Communications WABCS final: Supplementary data 1e: RNA-seq samples Race

Subtype: Perou PAM50 (Perou Nigeria\_109\_samples PAM50 Calls 9-12-19 UPDATED 9-21-19-PSR)

*NASER PROJECT:* NigerianTCGA\_quants-IHC *-> 116 files, 49 Nigerian and 66 TCGA with raw counts*

**Project File: NigerianTCGArawcountsDeSeq2-proteincoding-IHC.Rmd**

**Output Files: Output/Life-History-WGS-Validation**

**EXCLUDED:**

16 Nigerian samples: Poor coverage, not found or unmatched (chart as above)

67 samples = UNSURE = Duplicates + Not included on WABCS\_RNAseq-WGS\_overlap\_IDs.txt (84 samples)

LIB-04662; LIB-03316

LIB-02353; LIB-03325; LIB-04661

LIB-03324; LIB-04660

LIB-03318; LIB-04629  
LIB-03335; LIB-04626

LIB-02359; LIB-03332; LIB-04623

LIB-02355; LIB-03323; LIB-04609

LIB-03317; LIB-04607

LIB-02352; LIB-03320; LIB-04604

LIB-02358; LIB-03322; LIB-03328

LIB-02353; LIB-03325; LIB-04661

|  |  |
| --- | --- |
| LIB-04622 /  LIB-02357 | Duplicate -> Using sampleLIB-03321 (consistent with Chuck project) |
| LIB-02356 | Duplicate -> Using sample LIB-04606 (consistent with Chuck project) |
| LIB-04630 | Duplicate -> Using sample LIB-03319 (consistent with Chuck project) |

20 TCGA samples: 15 Ethnicity not reported or Asian or Hispanic

1 IHC not available  
4 samples Excluded due to PCA with abnormal counts compared to other TCGA samples (preventing visualization of PCA groups)  
9ac00677-9ba1-4ee9-bf02-ecb04533f45d f0787165-6f58-4d67-b510-928eea2c4882 (white)  
67d1ea12-766f-4eb9-af7e-4545975b076c RNA -> 9435447e-d65f-408b-863b-6576b1d652dd (white)

4637f6a3-9e07-4ac4-be5a-c893a27c8fe2 RNA -> 8785012f-f73e-4d68-87cf-1d804af32782 (black)

8999059b-74b5-4388-afdb-457c9fd24421 RNA -> 8785012f-f73e-4d68-87cf-1d804af32782 (black)

**ANNOTATION:**

TCGA Ethnicity: Nature Communications WABCS final: Supplementary data 1e: RNA-seq samples Race

Subtype: Naser IHC status / IHC from Nature Communications DNA for corresponding patient (Nigerian-IHC-sample-list-Naser)

Genetic markers (LOH, GATA3, TP53): Naser annotation

*----*

*Considered analyses not performed:*

*\*Allele-specific expression for specific SNVs, copy number variations or indels identified as different*

*\*ASE Read Counter -> --countOverlapReadsType COUNT\_FRAGMENTS (so that it doesn't count fragment overlap at paired read sites) / known or predicted from CGI only (we can manipulate (keep CGI result in the VCF INFO field) and convert CGI tsv file back to single-subject VCF (two samples actually, a T/N pair).)*

*\*Subclonal expression: Correlation to finalized life history anaylsis with DriverNet or BayCount*

*\*Presence of signature switching (within Nigerian group) // HRD*

*\*Early vs late drivers or signatures (within Nigerian group) -> Temporal evolution of signatures*

*\*Immunogenomics expression? / HLA LOH*