Functional Principal Components Analysis for RNA-Seq Curve based Data

Abstarct

Transcript block,

Subtype

Now

Challendge

Molecular signature

1, miRNA: 3828(1595 miRNA\_primary\_transcript and 2233 mature miRNA)

2, non-allele specific expression mRNA

3, allele specific expression mRNA

4, traditional mRNA(array technique)

5, DNA methylation array

Serous Cystadenocarcinoma

420 RNA-seq bam file

489 miRNA-seq bam file

inferred platinum resistance: refractory=death in 6mo or less, sensitive=survival for 12mo or more

Two records in Follow up for certain patients

TCGA-42-2582

TCGA-42-2587

TCGA-42-2588

TCGA-42-2590

TCGA-42-2591

另外还有设置重复，TCGA-23-1023病人的样本进行了重复RNA-seq

以下病人收集了两个样本

TCGA-04-1644

TCGA-09-2055

TCGA-10-0925

TCGA-13-0901

TCGA-23-1023

TCGA-24-0975

162人完全使用Carboplatin+Taxol

perl drugpair2.pl | perl -lane 'print $1 if /Carboplatin Taxol/ig'

435人使用Carboplatin+Taxol组合，并额外使用其他一些药物。

perl drugpair2.pl | perl -lane 'print $1 if /Carboplatin(.\*)Taxol/ig'

使用的药物少于2种的情况下的药物组合情况

perl drugpair2.pl | perl -lane 'print $\_ if (! defined @F[2])'

Background

Removal of the tumor usually is the first step of the cancer therapy and then chemotherapy: Combination chemotherapy with a platinum-based drug and a taxane drug delivered intraperitoneally (through the abdominal cavity). Sometimes, clinical trials of biologic drugs (targeted therapy) following combination chemotherapy will be conducted.

Chemotherapy is often given as a combination of drugs. Combinations usually work better than single drugs because different drugs kill cancer cells in different ways. Each of the drugs in this combination is approved by the Food and Drug Administration (FDA) to treat cancer or conditions related to cancer. ***Carboplatin-Paclitaxel Combination*** can be used to treat: Non-small cell lung cancer that has spread and Ovarian cancer. This combination may also be used with other drugs or treatments or to treat other types of cancer.

Data

Processing of the Cancer Genome Atlas (TCGA)'s next-generation sequencing data and their comparisons to our data

The TCGA data on the next-generation sequencing were downloaded from its data portal (https://browser.cghub.ucsc.edu/browser/).

The sequence tags were aligned against cDNA database (GRCh37.55) from the Ensembl (www.ensembl.org/) using the SOAP2 (v2.19) program [15], which is an ultrafast short read alignment tool. Up to 1 mismatch was allowed during the alignment. If a tag was mapped to more than 2 genes, then it was filtered out. If a tag mapped to two genes, its count was assigned to the more abundant gene. The count numbers of all mapped tags to a gene were summed to represent the expression abundance of the gene. The expression levels were normalized to one million to calculate the expression level in TPM (transcript per million) for each gene. The comparisons of the TCGA's data to our data were performed at gene levels.

TCGA Analysis

Clinical information were downloaded from the website1 while all the omics data were download from website2 with customized perl script.

Phenotype selection

1, tumor size ~ (number of mutation + eclipse time from onset)

2, survival duration ~( therapy scheme + mutation + methylation + expression signature )

3, response ~ (mutation, methylation, expression signature)

4,

bcr\_patient\_barcode TCGA-G4-6293

bcr\_drug\_barcode TCGA-G4-6293-D21800

bcr\_drug\_uuid b280e2a2-cbfb-45dd-9277-bb4a0c42f558

clinical\_trail\_drug\_classification [Not Available]

date\_of\_form\_completion ########

days\_to\_drug\_therapy\_end 235

days\_to\_drug\_therapy\_start 25

drug\_category [Not Applicable]

drug\_name Fluorouracil

measure\_of\_response [Not Available]

number\_cycles 4

prescribed\_dose 500

prescribed\_dose\_units mg/m2

regimen\_indication ADJUVANT

regimen\_indication\_notes [Not Applicable]

regimen\_number 1

route\_of\_administration Intravenous (IV)

therapy\_ongoing NO

therapy\_type Chemotherapy

therapy\_type\_notes [Not Available]

total\_dose 2000

total\_dose\_units mg/m2

tx\_on\_clinical\_trial [Not Available]

imputation

as to the patients with missing days to tumor recurrence from surgical operation are imputed with the days days to tumor progression or days to death.

今天用AgilentG4502A\_07\_3表达芯片的data对数据进行了cluster analysis

#ProgressionFreeSurvival is missing if the patient has progressed/recurred and there is no date of progression/recurrence indicated or the patient is dead and no date of progression/recurrence is indicated.

\*The PlatinumFreeInterval is negative if the patient had received treatment after the date of progression/recurrence or if the date of last platinum treatment is after the date of last follow-up.

Webiste1: <https://tcga-data.nci.nih.gov/docs/dictionary/TCGA_BCR_DataDictionary.xml>

Website2:

**FOLFOX** is a [chemotherapy regimen](http://en.wikipedia.org/wiki/Chemotherapy_regimen) for treatment of [colorectal cancer](http://en.wikipedia.org/wiki/Colorectal_cancer), made up of the drugs

FOL– [Folinic acid](http://en.wikipedia.org/wiki/Folinic_acid) ([leucovorin](http://en.wikipedia.org/wiki/Leucovorin" \o "Leucovorin))

F – [Fluorouracil](http://en.wikipedia.org/wiki/Fluorouracil) (5-FU)

OX – [Oxaliplatin](http://en.wikipedia.org/wiki/Oxaliplatin) (Eloxatin)[[1]](http://en.wikipedia.org/wiki/FOLFOX" \l "cite_note-1)

Age: age\_at\_initial\_pathologic\_diagnosis

<https://wiki.nci.nih.gov/display/TCGA/RNASeq+Version+2>

There are two analysis pipelines used to create Level 3 expression data from RNA Sequence data. The first approach used at TCGA relies on the [RPKM](https://wiki.nci.nih.gov/x/VxNCB) method, while the second method uses MapSplice to do the alignment and RSEM to perform the quantitation.

References:

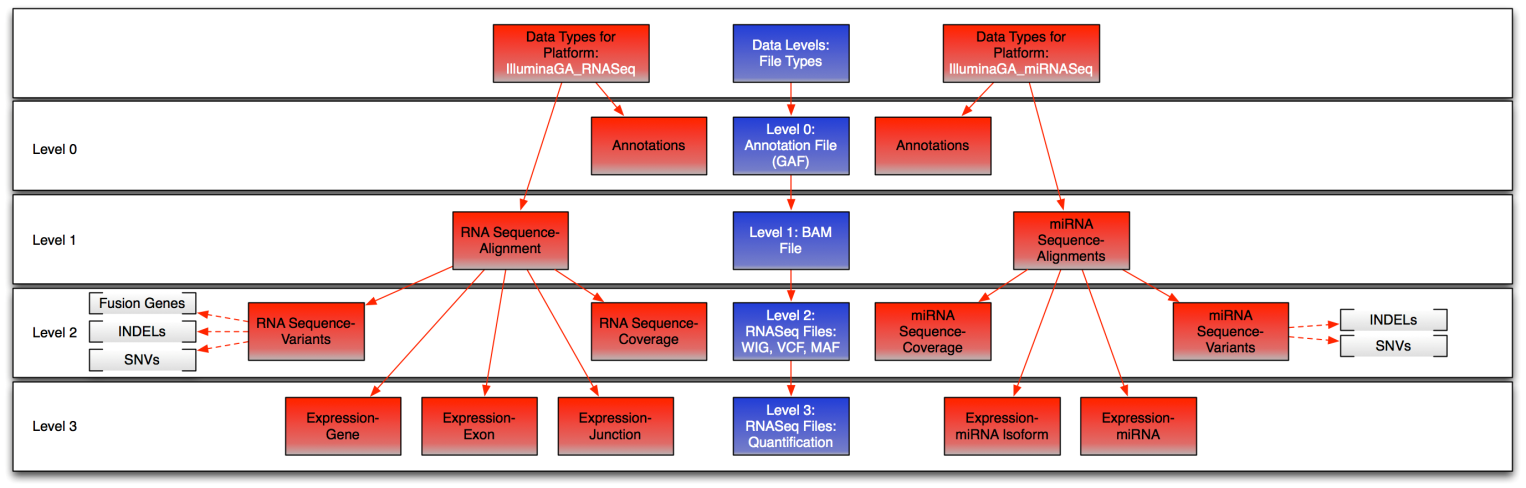
Li B, Ruotti V, Stewart RM, Thomson JA, Dewey CN. (2010)  
**RNA-Seq gene expression estimation with read mapping uncertainty.**   
Bioinformatics. Feb 15;26(4):493-500.

[Pubmed Link](http://www.ncbi.nlm.nih.gov/pubmed/20022975)   
[Bioinformatics link](http://bioinformatics.oxfordjournals.org/content/26/4/493.long) [Exit Disclaimer logo](http://www.cancer.gov/global/web/policies/page8)

Wang K, Singh D, Zeng Z, Coleman SJ, Huang Y, Savich GL, He X, Mieczkowski P, Grimm SA, Perou CM, MacLeod JN, Chiang DY, Prins JF, Liu J. (2010)  
**MapSplice: accurate mapping of RNA-seq reads for splice junction discovery.**   
Nucleic Acids Res. Oct;38(18):e178.

[Pubmed Link](http://www.ncbi.nlm.nih.gov/pubmed/20802226)   
[Nucleic Acids Research link](http://nar.oxfordjournals.org/content/38/18/e178.long) [Exit Disclaimer logo](http://www.cancer.gov/global/web/policies/page8)

数据由两所科研单位产生：北卡罗莱纳大学 (Universities of North Carolina，USA)和加拿大的迈克尔·史密斯基因组科学中心（Michael Smith Genome Sciences Centre，Canada）。



GAF (generic annotation file)文件被称为level 0的数据，GAF信息在SDRF文件的Annotation REF列也有反映。

Bam文件是level 1 的数据，会被提交到dbGap数据库中，分析人员通过数据库探测技术动态监控Bam文件的动态变化。(<https://tcga-data.nci.nih.gov/datareports/bamTelemetryReport.htm>)。

Wig文件(描述genome coverage)的文件，及各种序列变异文件（SNVs，INDELs，Fusion genes）等隶属于level 2的数据。level 3的数据包括量化后的基因表达数据，exon表达数据和Junction区域的表达信息。

目前TCGA暂不公布level 1 和level 2 的数据，只公布level 3的数据，此次现在的level 3 的数据尚不包含Expression-Splice Variants信息，因为对Expression-Splice Variant公认的统计方法还尚未建立。

Mega-TAB（<http://tab2mage.sourceforge.net/docs/magetab_docs.html>）

所有数据结构说明网址（[https://wiki.nci.nih.gov/display/TCGA/RNASeq#RNASeq-IlluminaGAmRNADGEValidationRules](https://wiki.nci.nih.gov/display/TCGA/RNASeq" \l "RNASeq-IlluminaGAmRNADGEValidationRules) ）

RNA-seq比对技术经过第一代技术[Bo Li]和第二代技术[Kai Wang]的不断完善，渐进成熟。

Subtype analysis

Matlab script

<http://www.genome.duke.edu/labs/nevins/ReevesSupplement1/figure1supp.html>

RNA-seq

<http://seqanswers.com/forums/showthread.php?t=19172>

http://seqanswers.com/forums/showthread.php?t=19007

MicroRNA

Human miRNA were download from miRBase release 19[[1](#_ENREF_1)].

MicroRNA存在多种形式，最原始的是pri-miRNA，长度大约为300~1000个碱基；pri-miRNA经过一次加工后，成为pre-miRNA即microRNA前体，长度大约为70~90个碱基；pre-miRNA再经过Dicer酶酶切后，成为长约20~24nt的成熟miRNA。

Method

GENCODE Genes V14 (wgEncodeGencodeV14) Track Description

The GENCODE Genes track (version 14, October 2012) shows high-quality manual annotations merged with evidence-based automated annotations across the entire human genome generated by the GENCODE project. The GENCODE gene set presents a full merge between HAVANA manual annotation process and Ensembl automatic annotation pipeline. Priority is given to the manually curated HAVANA annotation using predicted Ensembl annotations when there are no corresponding manual annotations. The annotation was carried out on genome assembly GRCh37 (hg19).

miRNA-SEQ method

Alignment of reads to genome, annotation of reads using reference databases, and measurement of expression levels by annotation. All post-alignment analysis was performed with MiRNA Profiling v0.2.6 (<http://www.bcgsc.ca/platform/bioinfo/software>)

[**LIBRARY\_CONSTRUCTION\_PROTOCOL**]

Small RNAs, containing microRNA (miRNA), in the flow-through material following mRNA purification on a MultiMACS separator (Miltenyi Biotec, Germany) are recovered by ethanol precipitation. MiRNA-seq libraries are constructed using a 96-well plate-based protocol developed at the BC Cancer Agency, Genome Sciences Centre. Briefly, an adenylated single-stranded DNA 3' adapter is selectively ligated to miRNAs using a truncated T4 RNA ligase2 (NEB Canada, cat. M0242L). An RNA 5' adapter is then added, using a T4 RNA ligase (Ambion USA, cat. AM2141) and ATP. Next, first strand cDNA is synthesized using Superscript II Reverse Transcriptase (Invitrogen, cat.18064 014), and serves as the template for PCR. Index sequences (6 nucleotides) are introduced at this PCR step to enable multiplexed pooling of miRNA libraries. PCR products are pooled, then size-selected on an in-house developed 96-channel robot to enrich the miRNA containing fraction and remove adapter contaminants. Each size-selected indexed pool is ethanol precipitated and quality checked on an Agilent Bioanalyzer DNA 1000 chip and quantified using a Qubit fluorometer (Invitrogen, cat. Q32854). Each pool is then diluted to a target concentration for cluster generation and loaded into a single lane of a GAIIx or HiSeq 2000 flow cell for sequencing with a 31-bp main read (for the insert) and a 7-bp read for the index

RNA-SEQ Method

Illumina paired-end RNA sequencing reads were aligned to GRCh37-lite genome-plus-junctions reference using BWA version 0.5.7. This reference combined genomic sequences in the GRCh37-lite assembly and exon-exon junction sequences whose corresponding coordinates were defined based on annotations of any transcripts in Ensembl (v59), Refseq and known genes from the UCSC genome browser, which was downloaded on August 19 2010, August 8 2010, and August 19 2010, respectively. Reads that mapped to junction regions were then repositioned back to the genome, and were marked with 'ZJ:Z' tags. BWA is run using default parameters, except that the option (-s) is included to disable Smith-Waterman alignment. Finally, reads failing the Illumina chastity filter are flagged with a custom script, and duplicated reads were flagged with Picard's MarkDuplicates

The uuid value was of an illegal format. A uuid is a 16-byte (128-bit) number. In its canonical form, a uuid is represented by 32 hexadecimal digits, displayed in 5 groups separated by hyphens, in the form 8-4-4-4-12 for a total of 36 characters (32 digits and 4 hyphens).  The requested action will not be performed.  Please examine your submitted UUID to make sure it is correctly formatted. If it looks correct please contact the helpdesk.

最新的Gencode V14显示在基因组中仅仅有不到110 M的碱基用于翻译成蛋白质 (107434781, 3.43% of GRCh37[3,137,144,693])

Method,

So suppose we have 400 samples and 3000 sites for each gene. And there are 4 subtype of the ovarian cancer, the proportion is 50:50:100:200.

Plink 在二代测序时代的应用

科研工作的初期更多的是学习和积累，一定程度上很强调系统的锻炼。

最基本的科学规范和默认尝试需要牢记在心。

Affection status, by default, should be coded:

-9 missing

0 unaffected

1 affected

数据必须分开存储，保证安全，设置稳定可靠地验证程序，数据分析时数据来源的正确性。一般而言需要设置dat文件，map文件，cov文件，phe文件等。

Dat文件，row为individual，column为snp，gene or site等信号信息

Map文件，row为每个site的信息（position , gene symbol,此文件可以进行丰富）

Cov文件，和phe可能相关的individual非phe信息都可以放在这里

Phe文件，phenotype信息都放在这里

Ped file需要有统一的格式（）

The PED file is a white-space (space or tab) delimited file: the first six columns are mandatory:

前六列需要固定为一下内容

Family ID

Individual ID

Paternal ID

Maternal ID

Sex (1=male; 2=female; other=unknown)

Phenotype

Quantitative traits with decimal points must be coded with a period/full-stop character and not a comma, i.e. 2.394 not 2,394

Genotypes (column 7 onwards) should also be white-space delimited; they can be any character (e.g. 1,2,3,4 or A,C,G,T or anything else) except 0 which is, by default, the missing genotype character.

**Map file**

Map文件需要在plink的基础上进行较大的变动，我们的目的是把map文件转变为bedtools等软件可以进行处理的格式。By default, each line of the MAP file describes a single marker/region and must contain exactly 4 columns:

chromosome (1-22, X, Y or 0 if unplaced)

Base-pair start position (bp units)

End-pair start position (bp units)

Snp(rs#)/region identifier

Genetic distance (morgans)

SNP的start和end可以采用同一位置的原则，即start=end，注意对于position的0base还是1base的问题，本人建议采用1base，尽管这样增加了计算量，但是更容易和现有科学体系match，避免不必要的错误的产生。

对于map文件格式的思考实际上是个重要的问题，但目前我还没有找到理想的解决方案。显然对于genome而言，position是个unique的标示， 每个position有很多annotation也就是存在冗余性。这些annotation可能不同维度的（一个position同时有甲基化和snp等信息），但也可能是同一维度的（一个position被多个基因占有）。因此map文件可能非常非常大，如果全基因单base分辨率的话，可能会有几十个G的数据量[对于hg19，仅仅exon但碱基分辨率的话，文件大小是1.2G]，并且随着生命科学的发展，此数据量会越来越大。

就我目前的知识容量我认为需要建立的资料有

1. 所在区域/位点的snp，或者snp的个数
2. 所在区域/位点的CpG信息，是否在CpG island上
3. 所在区域/位点的Gene信息，基因的个数

perl -lane 'print "@F[1]\t@F[2]\t@F[4]\t@F[5] \t@F[3] \t@F[12]"' refgene

|  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- |
| chrom | txStart | txEnd | name2 | name | strand |
| chr1 | 66999824 | 67210768 | SGIP1 | NM\_032291 | + |
| chr1 | 33546713 | 33585995 | ADC | NM\_052998 | + |

**Minimum non-overlapping transcript units (MNOTU)**

Minimum non-overlapping transcript units (MNOTU), download exon data from latest UCSC GENECODE V14[[2](#_ENREF_2), [3](#_ENREF_3)], RefGene[[4-6](#_ENREF_4)] and merge overlapping region and then 27,768 units for encode V14 and 282,027 units for refGene. statistic the number of the genes in this region. Note, sort bed file first then merge the bed file with the above principle.

perl -lane 'print " @F[2]\t@F[4]\t@F[5] \t@F[13] \t@F[12] @F[1]\t "' refgene

Table, **The distribution of the number of th gene in the MNOTU(27756) for encode gene**

|  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- |
| **1** | **2** | **3** | **4** | **5** | **6** | **7** | **8** | **9** | **10** | **11** | **12** | **13** | **14** | **15** | **16** |
| **20009** | **4233** | **1744** | **860** | **403** | **223** | **114** | **64** | **43** | **20** | **16** | **12** | **5** | **5** | **2** | **3** |

**The distribution of the number of th gene in the MNOTU(27756) for reference gene**

|  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- |
| 1 | 2 | 3 | 4 | 5 | 6 | 7 | 8 | 9 | 10 | 11 | 12 | 13 | 14 | 15 |
| 17843 | 2257 | 515 | 151 | 73 | 46 | 23 | 17 | 4 | 14 | 9 | 1 | 2 | 2 | 1 |

Drug相关的基因在liu,2012[[7](#_ENREF_7)]中发现了227个药物相关基因。我们提取了这227个基因，

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| 1 | 2 | 3 | 4 | 9 |
| 171 | 30 | 4 | 1 | 1 |

|  |  |  |  |
| --- | --- | --- | --- |
|  | Unique | multi | sum |
| Genome | 17843 | 3119 | 20962 |
| Drug | 171 | 36 | 207 |

Pearson's Chi-squared test with Yates' continuity correction

Pvalue=0.3619

没有显著性差异

我们把liu找到的227个基因中171个**MNOTU作为feature进行cluster analysis，文件为drug\_mnotu\_feature.bed，内容如附件**

第一部分，

Bam数据基本资料

下载412人的143个bam文件，其中TCGA-23-1023-01有两个bam文件【check是什么原因】。

Bash file command as follow,

for i in {1..20}

do

perl unique.tmp.pl | perl -lane 'print $\_ if $\_ eq \$i' | wc -l

echo "$i"

done

**pedcreat.pl [合并一次412人的colon cancer需要33个小时，合并好后的ped文件为154G]**

#!/usr/bin/perl

use strict;

use Cwd;

chdir getcwd;

open OUT, ">tcga\_rnaseq.ped";

my @file=glob("TCGA\*.reads");

my ($fam,$iid,$fai,$mai,$sex,$phe);

$fai=0;$mai=0;$sex=0;$phe=1;

foreach my $file(@file){

#TCGA-04-1361-01A-01R-1565-13\_GRCh37-lite\_rnaseq.bam.reads

my ($sample,undef)=split/\_GRCh37/,$file;

$fam++;

$iid=$sample;

open F,$file;

chomp(my @content=<F>);

my $b=join "\t",@content;

print OUT "$fam\t$sample\t$fai\t$mai\t$sex\t$phe\t$b\n";

#print "$fam\t$sample\t$fai\t$mai\t$sex\t$phe\t$b\n";

}

unique.tmp.pl

#!/usr/bin/perl

use Cwd;

my $dir=getcwd;

sub unique (&@){

my $a\_ref = shift;

my %hash = map {$\_ => 1} @$a\_ref;

my @unique\_array = sort (keys %hash);

return @unique\_array;

}

open F,"mnotu.encode14.bed";

while(<F>){

my @line=split/\t/;

my @gene=split /;/,$line[3];

my $num=unique(\@gene);

print "$num\n";

}

1. Kozomara, A. and S. Griffiths-Jones, *miRBase: integrating microRNA annotation and deep-sequencing data.* Nucleic Acids Res, 2011. **39**(Database issue): p. D152-7.

2. Flicek, P., et al., *Ensembl 2011.* Nucleic Acids Res, 2011. **39**(Database issue): p. D800-6.

3. Harrow, J., et al., *GENCODE: producing a reference annotation for ENCODE.* Genome Biol, 2006. **7 Suppl 1**: p. S4 1-9.

4. Pruitt, K.D., T. Tatusova, and D.R. Maglott, *NCBI reference sequences (RefSeq): a curated non-redundant sequence database of genomes, transcripts and proteins.* Nucleic Acids Res, 2007. **35**(Database issue): p. D61-5.

5. Pruitt, K.D., T. Tatusova, and D.R. Maglott, *NCBI Reference Sequence (RefSeq): a curated non-redundant sequence database of genomes, transcripts and proteins.* Nucleic Acids Res, 2005. **33**(Database issue): p. D501-4.

6. Pruitt, K.D., T. Tatusova, and D.R. Maglott, *NCBI Reference Sequence project: update and current status.* Nucleic Acids Res, 2003. **31**(1): p. 34-7.

7. Liu, Y., et al., *Integrated analysis of gene expression and tumor nuclear image profiles associated with chemotherapy response in serous ovarian carcinoma.* PLoS One, 2012. **7**(5): p. e36383.