

ANÁLISE DE DADOS GENÉTICOS: UM PROBLEMA DE BIG DATA A CADA NOVO PACIENTE

RBRAS 2016 - SALVADOR, BA

Marcus Nunes

24 e 25 de maio de 2016

Universidade Federal do Rio Grande do Norte

QUEM SOU EU?

QUEM SOU EU?

- Sou Marcus Nunes, Ph.D. em Estatística pela Penn State University
- Professor na UFRN
- Meus interesses principais são as aplicações da Estatística em grandes conjuntos de dados, como genética, climatologia e saúde
- `marcus.nunes@ccet.ufrn.br`
- `http://marcusnunes.me/rbras-2016/`

SOBRE O QUE É ESTE MINICURSO?

SOBRE O QUE É ESTE MINICURSO?

- Uma **introdução** à análise de dados genéticos
- Vamos entender de onde estes dados vem
- Como eles podem ser preparados para a análise
- E realizaremos testes estatísticos nestes dados

SOBRE O QUE É ESTE MINICURSO?

- O que é Big Data?
- Quem trabalha com Big Data?
- Uma ideia geral sobre DNA
- Fundamentos estatísticos
- Aplicação em um conjunto real de dados

O QUE É BIG DATA?

O QUE É BIG DATA?

- Não existe consenso a respeito de uma definição sobre o que realmente é big data
- A área ainda é nova; não houve tempo para o conhecimento sedimentar
- Em geral, diz respeito a áreas do conhecimento onde as ferramentas de análise de dados tradicionais não são a melhor escolha possível

O QUE É BIG DATA?

- Big Data são os dados que possuem 3 V:
- Volume
- Velocidade
- Variedade

O QUE É BIG DATA?

- Uma outra definição de Big Data se vale da Estatística para ser formulada
- Podemos considerar um conjunto de dados como Big Data se o tempo que levamos para ajustar um modelo aos dados é maior do que o tempo utilizado para a escolha deste modelo

O QUE É BIG DATA?

- Mike Franklin, da Universidade de Berkeley, diz o seguinte:
- “Big Data é todo conjunto de dados caro para manter e manipular e de onde é difícil extrair informações”
- Esta definição é relativa: para alguns, dados na casa dos terabytes podem ser caros para manter; para outros, dados na casa dos petabytes podem ser baratos para manter

QUEM TRABALHA COM BIG DATA?

- Competências de um profissional 100% capacitado para trabalhar com Big Data:
 - Estatística
 - Programação
 - Negócios
 - Conhecer bem a área de atuação (internet, marketing, área financeira, biologia etc)

- Que tipo de profissionais temos no momento?
 - Bons estatísticos e matemáticos que escrevem códigos sem otimização
 - Bons cientistas da computação que entendem um pouco de estatística e matemática
 - Bons cientistas da computação que entendem um pouco de negócios, depois de muita experiência na área
 - Doutores em biologia ou genética
 - Gerentes que sabem fazer estas pessoas trabalharem juntas

QUEM TRABALHA COM BIG DATA?

- Estatísticos
- Programadores
- Físicos
- Cientistas de Dados

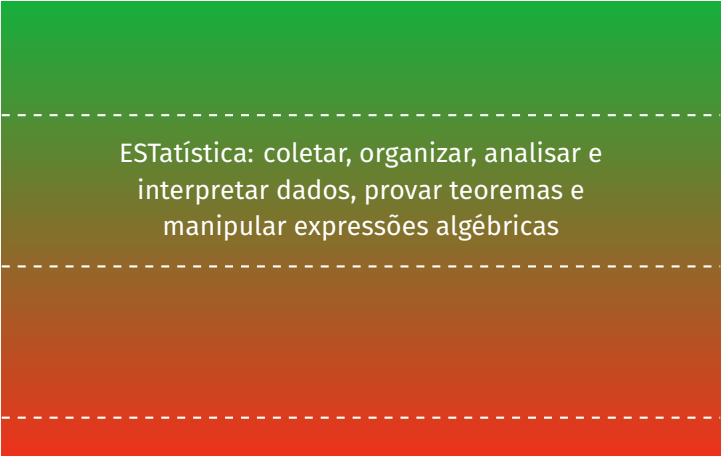
- Cientista de Dados (*Data Scientist*) é um novo nome para Estatístico
- Alguns dizem que o Cientista de Dados é um Estatístico que mora em São Francisco e usa um Mac
- No fundo, ambos são a mesma coisa, embora uma destas profissões trabalhe melhor seu marketing pessoal

QUEM JÁ JOGOU RPG?

QUEM JÁ JOGOU RPG?

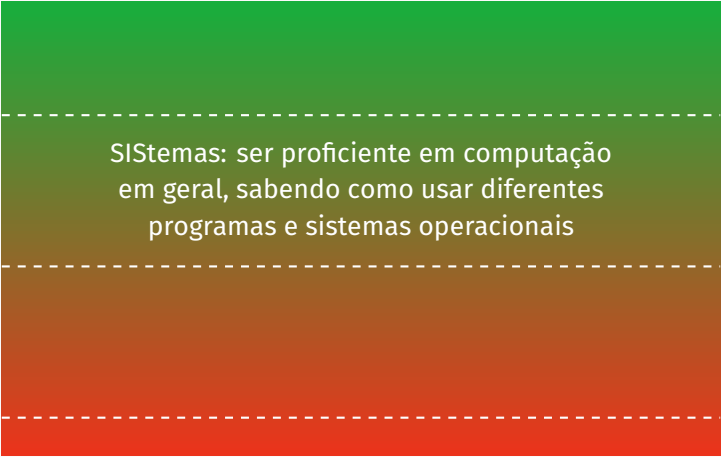
[illegible]





ESTatística: coletar, organizar, analisar e
interpretar dados, provar teoremas e
manipular expressões algébricas

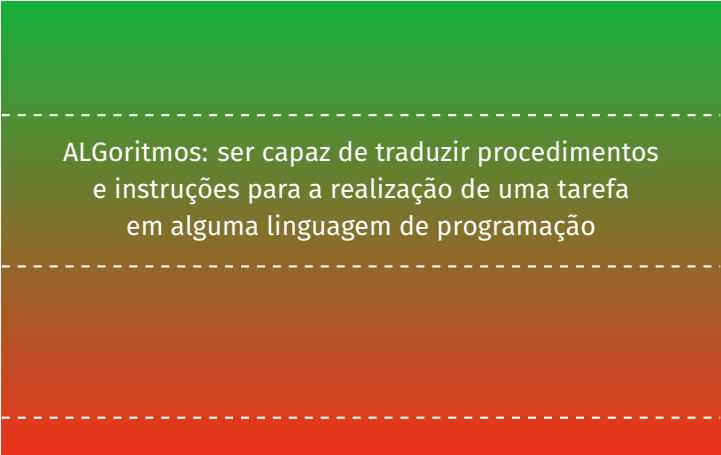
EST



SIStemas: ser proficiente em computação
em geral, sabendo como usar diferentes
programas e sistemas operacionais

EST

SIS

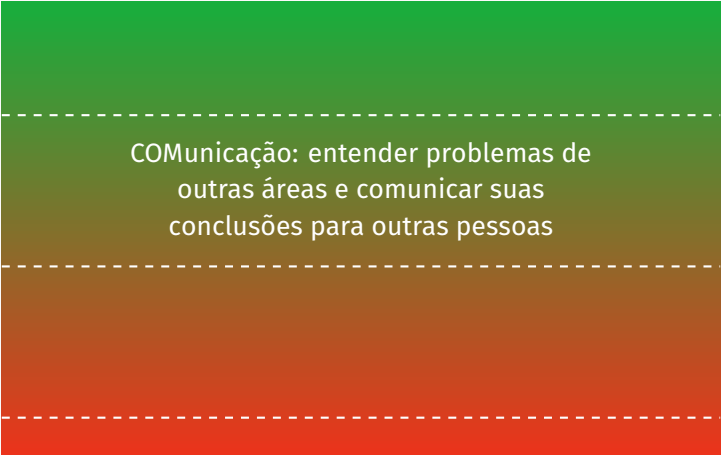


ALGoritmos: ser capaz de traduzir procedimentos
e instruções para a realização de uma tarefa
em alguma linguagem de programação

EST

SIS

ALG



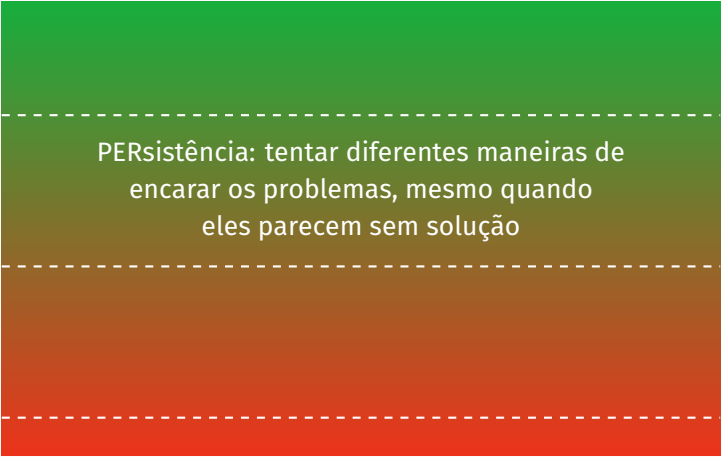
COMunicação: entender problemas de
outras áreas e comunicar suas
conclusões para outras pessoas

EST

SIS

ALG

COM



PERsistência: tentar diferentes maneiras de encarar os problemas, mesmo quando eles parecem sem solução

EST

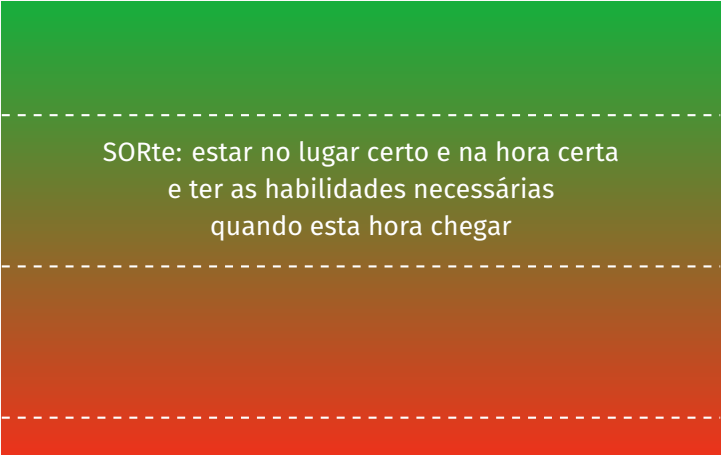
SIS

ALG

COM

PER

O QUE UM BIOESTATÍSTICO PRECISA SABER



SORte: estar no lugar certo e na hora certa
e ter as habilidades necessárias
quando esta hora chegar

EST

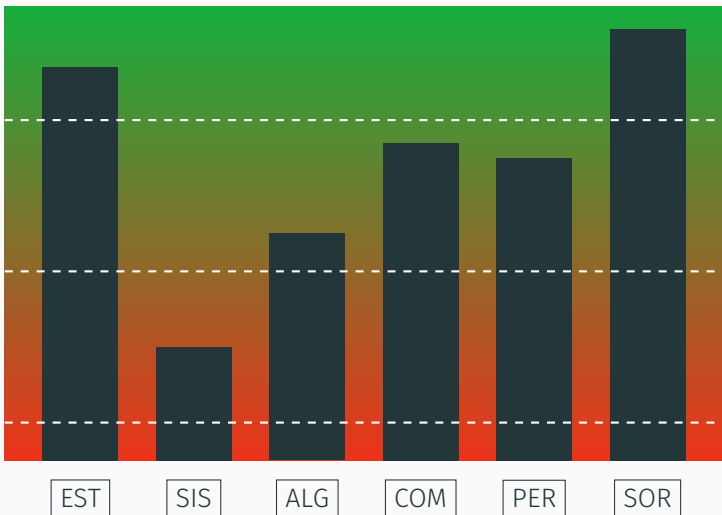
SIS

ALG

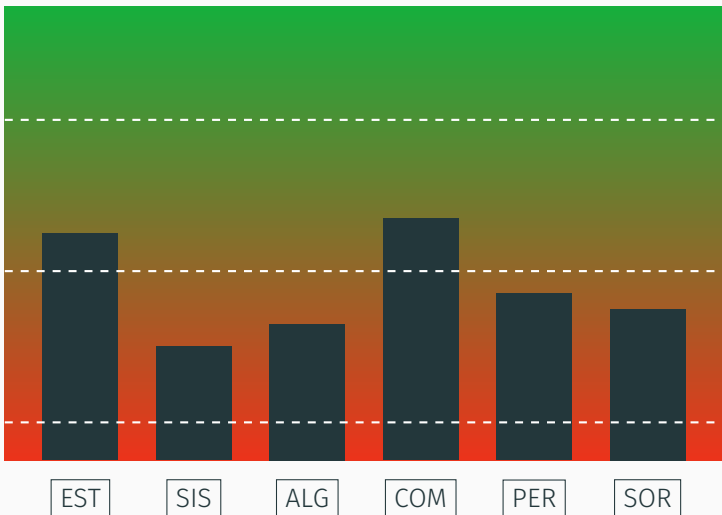
COM

PER

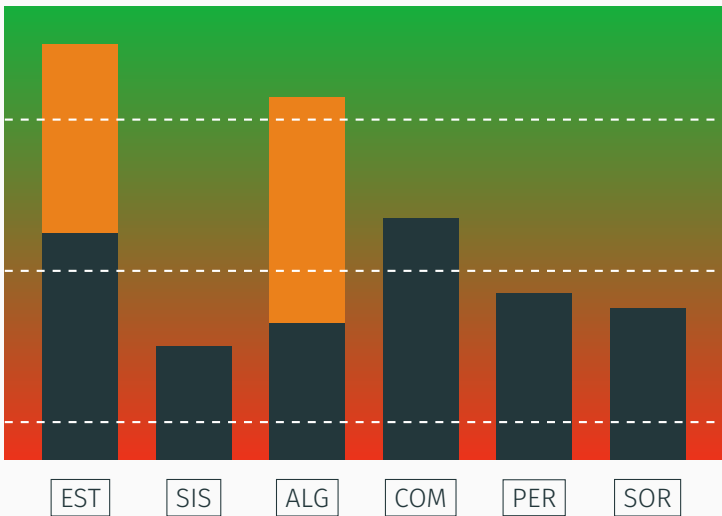
SOR



NÃO É BOM ESTAR NA MÉDIA



SEJA MUITO BOM EM ALGUMAS ÁREAS

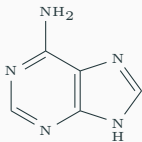


DNA

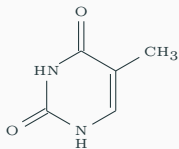


- Descrito pela primeira vez em 1948 (Watson e Crick)
- A genética já era conhecida anteriormente
- Mendel e suas ervilhas
- Francis Galton e a eugenia

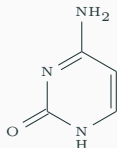
ESTRUTURA QUÍMICA DO DNA



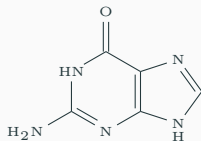
Adenina



Timina



Citosina



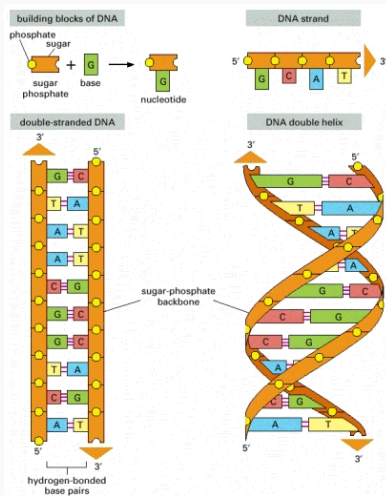
Guanina

MAS PARA QUE SERVE O DNA?

- Tudo
- Cor dos olhos, altura, propensão a sofrer de doenças, testes de paternidade no Programa do Ratinho
- Codifica aminoácidos em proteínas

- Cada nucleotídeo é uma base
- A adenina liga-se apenas com a timina, enquanto a citosina liga-se apenas com a guanina
- O genoma humano possui mais de 3 bilhões de pares de base

ESTRUTURA DO DNA

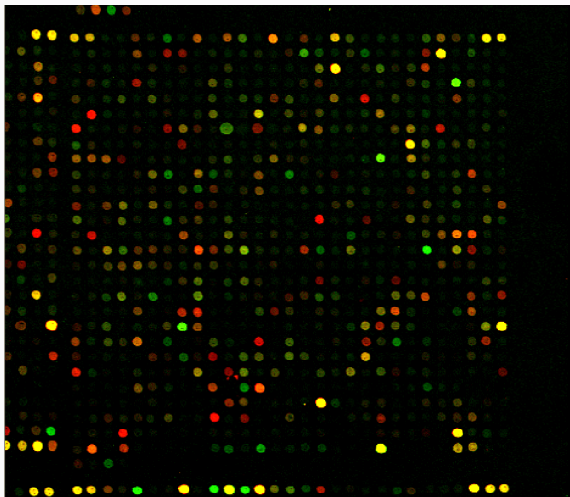


O QUE DESEJAMOS SABER SOBRE O DNA?

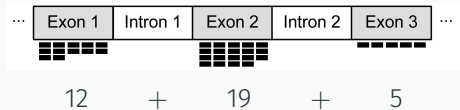
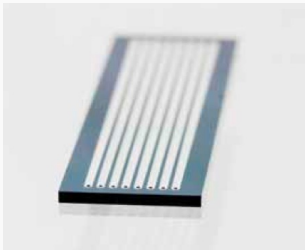
- Expressão gênica
- Processo em que a informação de um gene é utilizada na síntese de um produto gênico
- Em geral, transformar um ou mais aminoácidos em proteínas

- Sanger sequencing
- Microarrays
- RNA-Seq

- Usado no Projeto Genoma Humano
- Custou US\$ 2,7 bilhões
- 13 anos para ficar pronto



- Estão caindo em desuso
- Preço entre US\$ 200 e US\$ 650 por array
- Maior disponibilidade no mercado



- Método cuja utilização vem crescendo mais ultimamente
- Cada sequenciamento custa entre US\$ 40 e US\$ 2.000 (em maio de 2016)
- Leva entre 2 horas e 11 dias para ficar pronta, variando de acordo com a tecnologia utilizada

- Tecnologia desenvolvida na universidade de Oxford e lançada em maio de 2015
- Cada chip de sequenciamento custa US\$ 1000
- Em breve, a análise poderá ser feita em tempo real



1. Preparação da amostra
2. Sequenciamento
3. Alinhamento das leituras
4. Controle de qualidade
5. Análise e descrição dos resultados

- Não nos interessa aqui
- Função de um biólogo ou bioinformata
- Depende da tecnologia utilizada

```
@SRR014849.1 EIXKN4201CFU84 length=50
GGGGGGGGGGGGGGGGGGCTTTTTTTGTTTGGGAACCGAAAGGGTTTTGAAT
+SRR014849.1 EIXKN4201CFU84 length=50
3+&$#"#####7F@71,'";C?,B;?6B;:EA1EA 1EA5'9B:
```

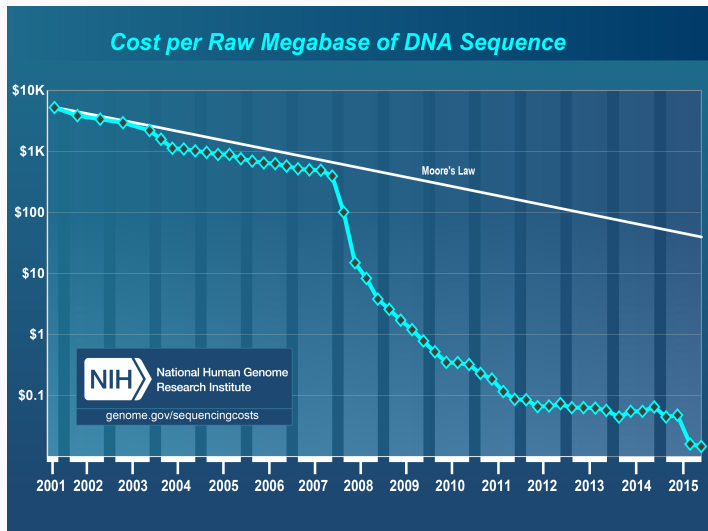
@título e descrição opcional
linha com o que foi sequenciado
+repetição opcional do título
linha com as qualidades da sequência

- Genoma de referência
- bowtie, SAMtools, bedtools
- Análise e descrição dos resultados

- Bioconductor - <http://bioconductor.org/>
- Gene Expression Omnibus (GEO)
<http://www.ncbi.nlm.nih.gov/geo/>
- BioStars - <http://www.biostars.org/>

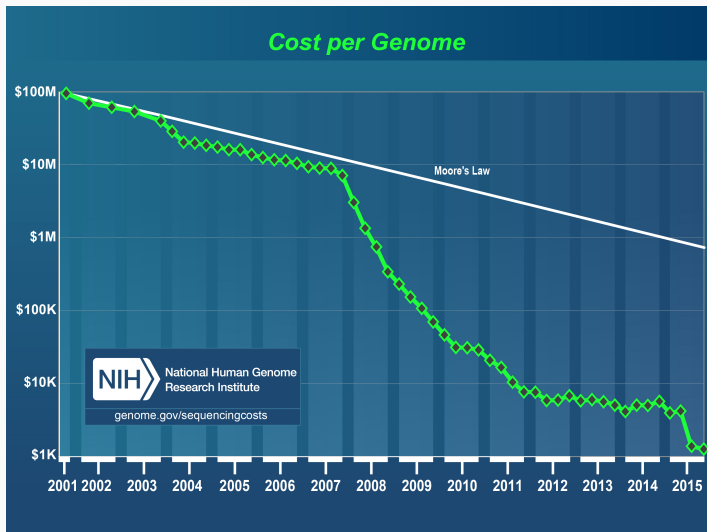
SEQUENCIAR GENOMAS É CADA VEZ MAIS BARATO

- Projeto Genoma Humano: 13 anos, US\$ 2,7 bilhões
- RNA-Seq: 8 horas, entre US\$ 805 e \$1.700
- MiniON: tempo real, US\$ 900



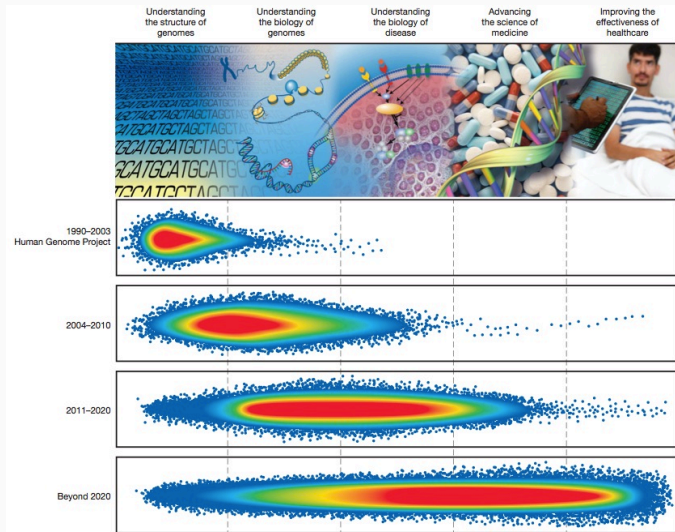
Fonte: <http://genome.gov>

CUSTO DE SEQUENCIAMENTO



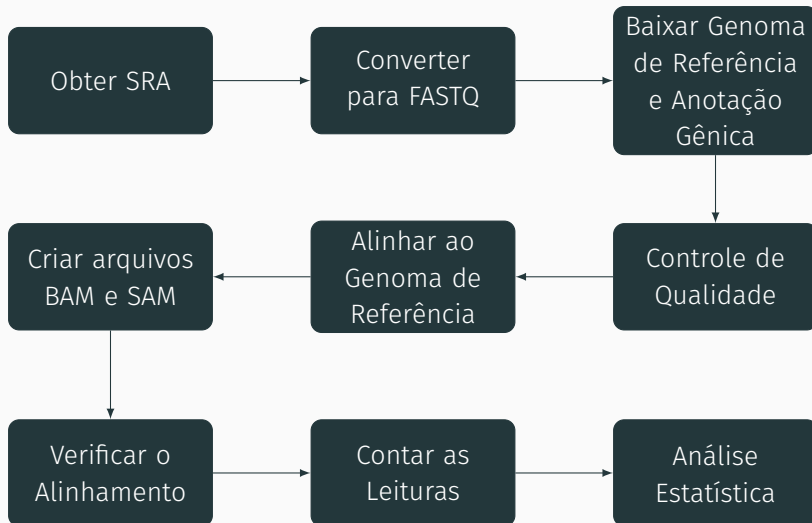
Fonte: <http://genome.gov>

FUTURO DA GENÔMICA



Fonte: Green and Guyer (2011)

MÉTODOS



- Experimentos de RNA-Seq devem ser planejados corretamente
- Máximo de informação
- Mínimo de custo

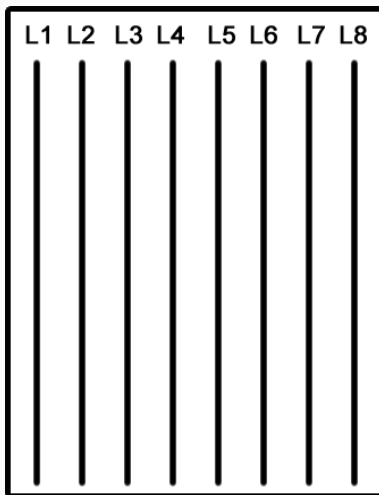
- Amostragem
- Replicação
- Agrupamento em blocos
- Aleatorização

- Ideias similares às de outros tipos de experimentos
- Definir claramente a nossa população de interesse
- Obter amostras representativas

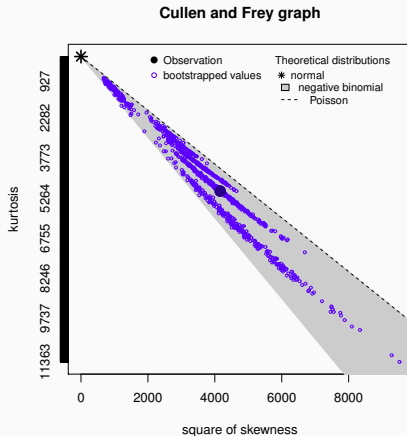
- Fazer comparações entre tratamentos
- Sujeitos distribuídos de maneira aleatória
- Evitar vícios

- Número suficiente de sujeitos no estudo
- Replicação biológica
- Replicação técnica

- Reduzir a variabilidade na análise
- Agrupando sujeitos similares
- Bloco incompleto equilibrado



- Dados discretos
- Não-normalidade
- Testes múltiplos



Poisson

- $f(x|\lambda) = \frac{e^{-\lambda} \lambda^x}{x!}$
- $E(Y) = \lambda$
- $\text{Var}(Y) = \lambda$

Binomial Negativa

- $f(y|r,p) = \binom{r+y-1}{y} p^r (1-p)^y$
- $E(Y) = \frac{pr}{1-p}$
- $\text{Var}(Y) = \frac{pr}{(1-p)^2}$
- $f(y|\mu,\phi) = \frac{\Gamma(y+\phi^{-1})}{\Gamma(\phi^{-1})\Gamma(y+1)} \left(\frac{1}{1+\phi\mu} \right)^{\phi^{-1}} \left(\frac{\phi\mu}{1+\phi\mu} \right)^y$
- $E(Y) = \mu$
- $\text{Var}(Y) = \mu + \phi\mu^2$

- Uma distribuição de probabilidade, da família exponencial, para o vetor resposta Y
- Um preditor linear para a esperança $\eta = X\beta$, que especifica as variáveis explicativas do modelo
- Uma função de ligação $g(\cdot)$ que relaciona η e μ tal que $\eta = g(\mu)$

- $f(y|\boldsymbol{\theta}, \phi) = \exp \left\{ \frac{y_i \theta_i - \kappa(\theta_i)}{\alpha_i(\phi)} + c(y_i|\phi) \right\}$
- $\kappa'(\theta_i) = E(Y)$
- $\kappa''(\theta_i) = \text{Var}(Y)$

EDGER

- Estimador de máxima verossimilhança condicional ajustada pelos quantis (Robinson e Smyth, 2010)
- Todas as amostras i no experimento possuem o mesmo tamanho (*i.e.*, $m_i = m$)
- A soma $Z = Y_1 + Y_2 + \dots + Y_k \sim \text{NB}(km\lambda, \phi k^{-1})$ é verdadeira

- Condicionando a verossimilhança em Z e tomando seu logaritmo natural, temos

$$\mathcal{L}(z|\phi) = \left[\sum_{i=1}^k \log \Gamma(y_i + \phi^{-1}) \right] + \log \Gamma(n\phi^{-1}) \\ - \log \Gamma(z + k\phi^{-1}) - k \log \Gamma(\phi^{-1})$$

- Com a equação acima é possível construir um método de estimação para o parâmetro ϕ

- Seja $m^* = \left(\prod_{i=1}^k m_i\right)^{\frac{1}{k}}$ a média geométrica dos tamanhos das bibliotecas
- Os dados observados são ajustados como se eles tivessem sido amostrados a partir de uma distribuição $NB(m^*\lambda, \phi)$

1. Encontre ϕ , o estimador CML que maximiza a verossimilhança condicional
2. Dada a estimativa de ϕ , estime λ
3. Assumindo que $y_i \sim \text{NB}(m_i\lambda, \phi)$, calcule os percentis observados

$$p_i = P(Y < y_i | m_i\lambda, \phi) + \frac{1}{2}P(Y = y_i | m_i\lambda, \phi), \quad (1)$$

$$i = 1, 2, \dots, k \quad (2)$$

4. Utilizando a interpolação linear das funções dos quantis, gere pseudo-dados de uma distribuição $\text{NB}(m^*\lambda, \phi)$, com quantis p_i
5. Calcule ϕ utilizando a CML nos pseudo-dados
6. Repita os passos 2 a 5 até ϕ convergir

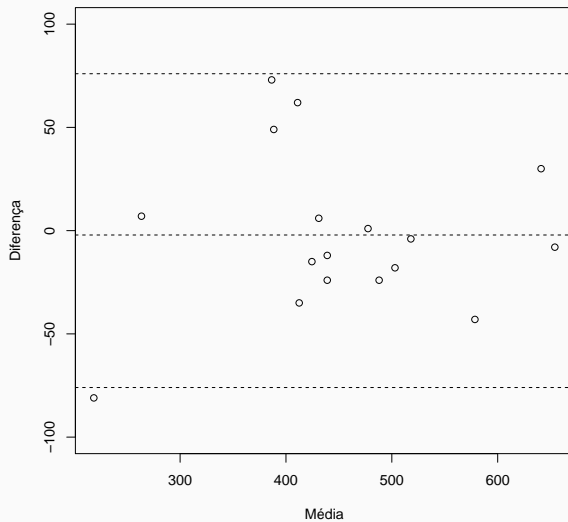
- É possível definir um teste exato
- Para dois grupos A e B, definimos Z_{tA} e Z_{tB} como as somas das pseudo-contagens destes grupos, sobre o número de amostras k_A e k_B . Sob a hipótese nula,

$$Z_{tl} \sim \text{NB}(n_l m^* \lambda_t, \phi n_l^{-1}), \quad l \in \{A, B\}$$

- Condicionando na soma das pseudo-contagens totais, $Z_{tA} + Z_{tB}$ também é uma variável aleatória Binomial Negativa

- O MA Plot é uma aplicação do gráfico de Bland-Altman em estudos genéticos
- Visa detectar diferenças sistemáticas entre duas replicações de um mesmo experimento
- Se estamos interessados na certa característica R de um experimento com duas replicações R_1 e R_2 , então as coordenadas cartesianas (x, y) do MA Plot são dadas por

$$R(x, y) = \left(\frac{R_1 + R_2}{2}, R_1 - R_2 \right)$$



- É como chamamos o fato de realizarmos duas ou mais inferências simultâneas
- No caso de testarmos apenas uma hipótese, definimos uma região de rejeição para controlar a taxa de falsos positivos, conhecidos como Erros do Tipo I, enquanto atingimos o mínimo possível para a taxa de falsos negativos, chamados de Erros do Tipo II
- Conforme o número de testes aumenta, torna-se cada vez mais provável que os grupos controle e tratamento diferenciem-se em pelo menos uma característica apenas devido à chance

- Quando determinamos um nível α para o Erro Tipo I de um teste estatístico, estamos na verdade dizendo que “ $\alpha \times 100\%$ das vezes em que deveríamos rejeitar a hipótese alternativa, nós estamos aceitando-a”
- Ou seja, se testamos a mesma hipótese nula 100 vezes, com um nível $\alpha = 0,05$, rejeitaremos H_0 em 5 destes testes, mesmo H_0 sendo verdade
- Existem diversas maneiras deste problema ser corrigido

- Se o nível desejado para erros do Tipo I em m testes realizados é (no máximo) α , então α/m é o valor da correção de Bonferroni para estes testes
- Justificativa:

$$P(\text{pelo menos um res. sig.}) = 1 - P(\text{nenhum res. sig.})$$

$$P(\text{pelo menos um res. sig.}) = 1 - (1 - \alpha)^m$$

- Se $\alpha = 0,05$ e $m = 100$,

$$P(\text{pelo menos um res. sig.}) = 1 - P(\text{nenhum res. sig.})$$

$$P(\text{pelo menos um res. sig.}) = 1 - (1 - 0,05)^{100}$$

$$P(\text{pelo menos um res. sig.}) = 0,9941$$

- Método conservador

- False Discovery Rate
- Um conjunto de predições possui um percentual esperando de falsas predições
- Para uma série de testes de hipóteses independentes, a FDR é dada por

$$\text{FDR} = E \left(\frac{V}{V + S} \right)$$

onde V é o número de falsos positivos e S é o número de verdadeiros positivos

Verdade	Decisão		Total
	Não-significativo	Significativo	
Hipótese nula	U	V	m_0
Hipótese alternativa	T	S	$m - m_0$
Total	$m - r$	r	m

- Combinamos os p-valores de cada teste num único vetor de p-valores. Após este vetor ser compilado, duas etapas são realizadas:
 1. Ordenar os m p-valores calculados do menor para o maior, denominando-os como $p_{(1)}, p_{(2)}, \dots, p_{(m)}$
 2. Encontrar o maior k tal que $p_{(k)} \leq \frac{k}{m}\alpha$
- Assumindo que os testes de hipóteses são independentes, este método controla a FDR desejada

APLICAÇÃO

- O conjunto de dados analisado aqui foi disponibilizado por Brooks *et al.* (2011)
- Sete amostras de *Drosophila melanogaster*, conhecida popularmente como mosca das frutas
- 3 amostras tratadas com siRNA (short interfering RNA - tratamento) e 4 amostras sem tratamento (controle)

HARDWARE E SOFTWARE NECESSÁRIO

- O ideal é utilizar um máquina com um sistema operacional *nix, seja Linux, Unix ou OS X
- Se possível, com vários processadores
- Um computador com 8 processadores, 8GB de RAM e um HD espaçoso já é um bom começo
- Entretanto, o seu computador pessoal pode dar conta do recado, embora seja um pouco lento
- Se o seu local de trabalho possui um cluster, aproveite-o!

- Alinhador de sequências: tophat2 -
<https://ccb.jhu.edu/software/tophat/index.shtml>
- Visualizador de arquivos: IGV -
<https://www.broadinstitute.org/igv/>

- Programa estatístico: R - <http://r-project.org/>
- Pacotes do R: `ShortRead`, `DESeq`, `edgeR`, `GenomicRanges`, `GenomicFeatures`, `org.Dm.eg.dm` e suas dependências

- Ferramenta para trabalhar com arquivos BAM e SAM: `samtools` - <http://samtools.sourceforge.net/>
- Ferramenta para contagem de leituras mapeadas: `HTSeq` - <http://www-huber.embl.de/HTSeq/doc/overview.html>
- Conversor de arquivos SRA: `SRA Toolkit` - <http://www.ncbi.nlm.nih.gov/Traces/sra/?view=software>

- Gerenciadores de pacotes podem facilitar a instalação destes programas
- Homebrew: <http://brew.sh/> (OS X)
- Linuxbrew: <http://linuxbrew.sh/> (testei no Ubuntu e CentOS)

- Todos os programas citados aqui são gratuitos
- Muitos deles são de código aberto, permitindo que sejam alterados e personalizados de acordo com seu uso
- Além disso, estão em constante atualização

PREPARAÇÃO DOS DADOS

```
> sri <- read.csv("SraRunInfo.csv", stringsAsFactors=FALSE)
> keep <- grep("CG8144|Untreated-", sri$LibraryName)
> sri <- sri[keep, ]
>
> fs <- basename(sri$download_path)
```

```
> fs
```

```
## [1] "SRR031714.sra" "SRR031715.sra" "SRR031716.sra"  
## [4] "SRR031717.sra" "SRR031724.sra" "SRR031725.sra"  
## [7] "SRR031726.sra" "SRR031727.sra" "SRR031708.sra"  
## [10] "SRR031709.sra" "SRR031710.sra" "SRR031711.sra"  
## [13] "SRR031712.sra" "SRR031713.sra" "SRR031718.sra"  
## [16] "SRR031719.sra" "SRR031720.sra" "SRR031721.sra"  
## [19] "SRR031722.sra" "SRR031723.sra" "SRR031728.sra"  
## [22] "SRR031729.sra"
```

```
> for(i in 1:nrow(sri)){  
+   download.file(sri$download_path[i], fs[i])  
+ }
```

```
> stopifnot(all(file.exists(fs)))  
> for(f in fs) {  
+   cmd <- paste("fastq-dump --split-3", f)  
+   cat(cmd, "\n")  
+   system(cmd)  
+ }
```

BAIXAR O GENOMA DE REFERÊNCIA E ANOTAÇÕES GÊNICAS

```
> # baixar o genoma de referencia
>
> system("wget ftp://ftp.ensembl.org/pub/release-70/
fasta/drosophila_melanogaster/
dna/Drosophila_melanogaster.BDGP5.70.dna.toplevel.fa.gz")
> system("gunzip Drosophila_melanogaster.BDGP5.70.dna.toplevel.fa.gz")
>
> # baixar as anotacoes dos genes
>
> system("wget ftp://ftp.ensembl.org/pub/release-70/
gtf/drosophila_melanogaster/
Drosophila_melanogaster.BDGP5.70.gtf.gz")
> system("gunzip Drosophila_melanogaster.BDGP5.70.gtf.gz")
```



```
> system("bowtie2-build -f  
Drosophila_melanogaster.BDGP5.70.dna.toplevel.fa  
Dme_BDGP5_70")
```

```
> library("ShortRead")  
> fqQC <- qa(dirPath=".", pattern=".fastq$", type="fastq")  
> report(fqQC, type="html", dest="fastqQAreport")
```

COLOCAR A TABELA INICIAL NUM FORMATO COM UMA AMOSTRA POR LINHA

```
> sri$LibraryName <- gsub("S2_DRSC_", "", sri$LibraryName)
> samples          <- unique(sri[, c("LibraryName", "LibraryLayout")])
> for(i in seq_len(nrow(samples))) {
+   rw <- (sri$LibraryName==samples$LibraryName[i])
+   if(samples$LibraryLayout[i]=="PAIRED") {
+     samples$fastq1[i] <- paste0(sri$Run[rw], "_1.fastq", collapse=",")
+     samples$fastq2[i] <- paste0(sri$Run[rw], "_2.fastq", collapse=",")
+   } else {
+     samples$fastq1[i] <- paste0(sri$Run[rw], ".fastq", collapse=",")
+     samples$fastq2[i] <- ""
+   }
+ }
```

ADICIONAR DESCRIÇÕES À TABELA DE METADADOS

```
> samples$condition = "CTL"  
> samples$condition[grep("RNAi",samples$LibraryName)] = "KD"  
> samples$shortname = paste(substr(samples$condition,1,2),  
+ substr(samples$LibraryLayout,1,2), seq_len(nrow(samples)), sep=".")
```

```
> samples[, c(1, 2, 5, 6)]
```

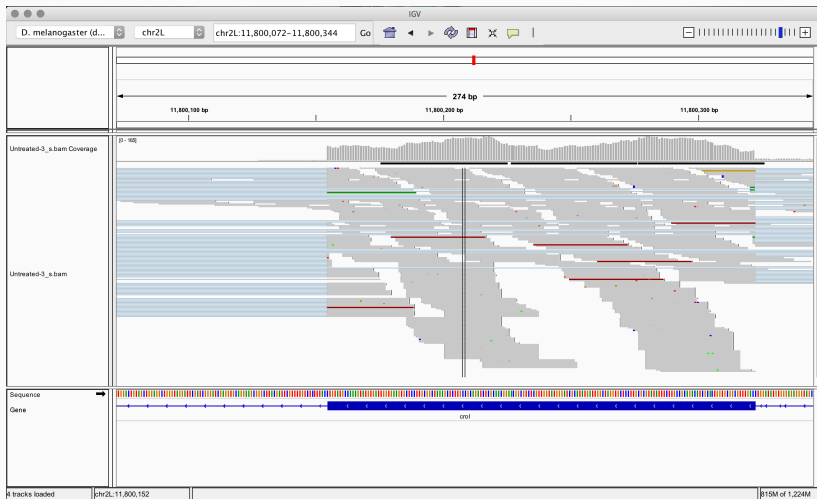
##	LibraryName	LibraryLayout	condition	shortname
## 1	Untreated-3	PAIRED	CTL	CT.PA.1
## 3	Untreated-4	PAIRED	CTL	CT.PA.2
## 5	CG8144_RNAi-3	PAIRED	KD	KD.PA.3
## 7	CG8144_RNAi-4	PAIRED	KD	KD.PA.4
## 144	Untreated-1	SINGLE	CTL	CT.SI.5
## 150	CG8144_RNAi-1	SINGLE	KD	KD.SI.6
## 156	Untreated-6	SINGLE	CTL	CT.SI.7

ALINHAR AS AMOSTRAS COM O GENOMA DE REFERÊNCIA

```
> gf      <- "Drosophila_melanogaster.BDGP5.70.gtf"
> bowind  <- "Dme_BDGP5_70"
> cmd     <- with(samples,
+   paste("tophat -G", gf, "-p 5 -o", LibraryName,
+   bowind, fastq1, fastq2))
> system(cmd)
```

```
> for(i in seq_len(nrow(samples))) {  
+   lib = samples$LibraryName[i]  
+   ob = file.path(lib, "accepted_hits.bam")  
  
+   # classificar por nome, converter para SAM para htseq-count  
+   cat(paste0("samtools sort -n ",ob," ",lib,"_sn"),"\n")  
+   cat(paste0("samtools view -o ",lib,"_sn.sam ",lib,"_sn.bam"),"\n")  
  
+   # classificar por posicao e indice para IGV  
+   cat(paste0("samtools sort ",ob," ",lib,"_s"),"\n")  
+   cat(paste0("samtools index ",lib,"_s.bam"),"\n\n")  
+ }
```

INSPECIONAR OS ALINHAMENTOS UTILIZANDO O IGV




```
> samples$countf <- paste(samples$LibraryName, "count", sep=".")
> gf              <- "Drosophila_melanogaster.BDGP5.70.gtf"
> cmd             <- paste0("htseq-count -s no -a 10 ",
+ samples$LibraryName, "_sn.sam ", gf, " > ", samples$countf)
> cmd
```

Tarefa	Tempo (horas)
Checar a qualidade	2
Organizar os metadados	1
Alinhamento das leituras	6
Contagem das leituras	3
Análise estatística	0,3
Total	12,3

- Note que o tempo de obtenção dos dados, seja através de um experimento ou de download via internet, não está sendo considerado

ANÁLISE ESTATÍSTICA

```
> library("edgeR")  
  
## Loading required package: limma  
  
> counts <- readDGE(samples$countf)$counts
```

FILTRAR OS GENES POUCO EXPRESSADOS E NÃO-INFORMATIVOS

```
> noint <- rownames(counts) %in% c("no_feature", "ambiguous",  
+ "too_low_aQual", "not_aligned", "alignment_not_unique")  
> cpms <- cpm(counts)  
> keep <- rowSums(cpms>1)>=3 & !noint  
> counts <- counts[keep,]
```

```
> head(counts)
```

##	Untreated-3	Untreated-4	CG8144_RNAi-3
## FBgn00000008	76	71	87
## FBgn00000017	3498	3087	3029
## FBgn00000018	240	306	288
## FBgn00000032	611	672	694
## FBgn00000042	40048	49144	70574
## FBgn00000043	15910	18194	31086
##	CG8144_RNAi-4	Untreated-1	CG8144_RNAi-1
## FBgn00000008	68	137	115
## FBgn00000017	3264	7014	4322
## FBgn00000018	307	613	528
## FBgn00000032	757	1479	1361
## FBgn00000042	72850	97565	95760

VISUALIZAR E INSPECIONAR A TABELA DE CONTAGENS

```
> colnames(counts) <- samples$shortname  
> counts <- counts[, order(samples$condition)]  
> head(counts)
```

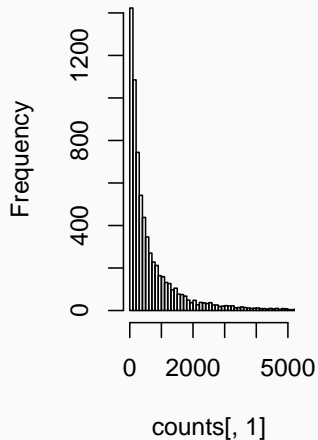
##	CT.PA.1	CT.PA.2	CT.SI.5	CT.SI.7	KD.PA.3	KD.PA.4
## FBgn00000008	76	71	137	82	87	68
## FBgn00000017	3498	3087	7014	3926	3029	3264
## FBgn00000018	240	306	613	485	288	307
## FBgn00000032	611	672	1479	1351	694	757
## FBgn00000042	40048	49144	97565	99372	70574	72850
## FBgn00000043	15910	18194	34171	29671	31086	34085
##	KD.SI.6					
## FBgn00000008	115					
## FBgn00000017	4322					
## FBgn00000018	528					
## FBgn00000032	1361					
## FBgn00000042	95760					
## FBgn00000043	42389					

VERIFICAR AS ESTATÍSTICAS DAS CONTAGENS

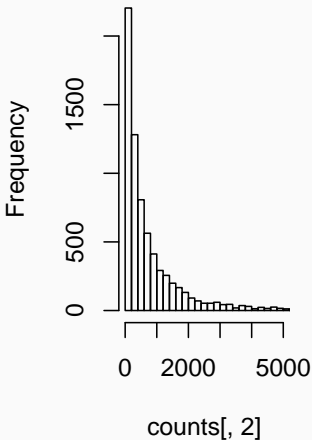
```
> summary(counts)
```

```
##      CT.PA.1      CT.PA.2      CT.SI.5
## Min.   :    4.0   Min.   :    7   Min.   :    7.0
## 1st Qu.:   130.0   1st Qu.:   159   1st Qu.:   326.0
## Median :   359.0   Median :   426   Median :   858.5
## Mean   :   1166.9   Mean   :  1377   Mean   :  2652.6
## 3rd Qu.:   967.2   3rd Qu.:  1085   3rd Qu.:  2178.0
## Max.   :130453.0   Max.   :165299   Max.   :293366.0
##      CT.SI.7      KD.PA.3      KD.PA.4
## Min.   :    0   Min.   :    3   Min.   :    1
## 1st Qu.:   219   1st Qu.:   159   1st Qu.:   174
## Median :   584   Median :   426   Median :   465
## Mean   :   1781   Mean   :   1343   Mean   :   1435
## 3rd Qu.:   1460   3rd Qu.:   1094   3rd Qu.:   1164
## Max.   :206540   Max.   :144953   Max.   :162846
##      KD.SI.6
## Min.   :    3.0
## 1st Qu.:   281.0
## Median :   721.5
## Mean   :   2129.6
```

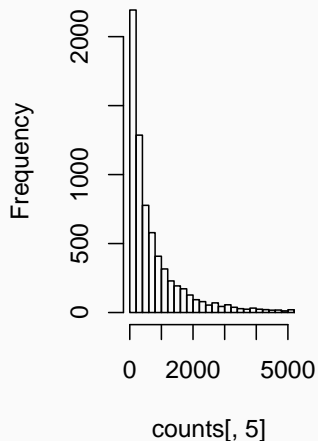
CT.PA.1



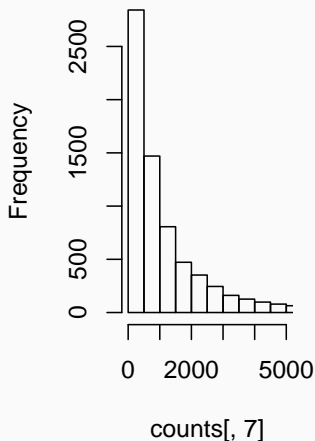
CT.PA.2



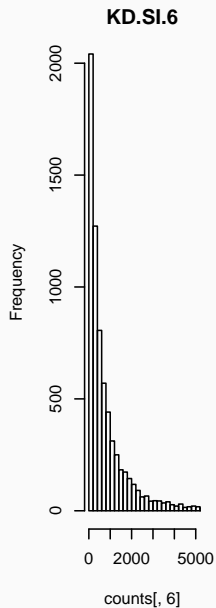
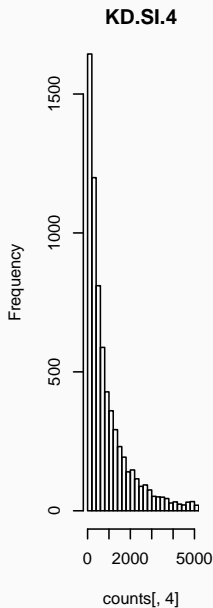
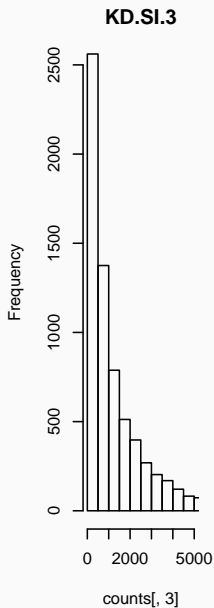
CT.PA.5



CT.PA.7



CONSTRUIR OS HISTOGRAMAS DAS CONTAGENS



```
> d <- DGEList(counts=counts, group=samples$condition)
> names(d)

> counts <- read.csv(file="counts.csv", header=T)
> d      <- DGEList(counts=counts, group=samples$condition)
> names(d)

## [1] "counts" "samples"
```

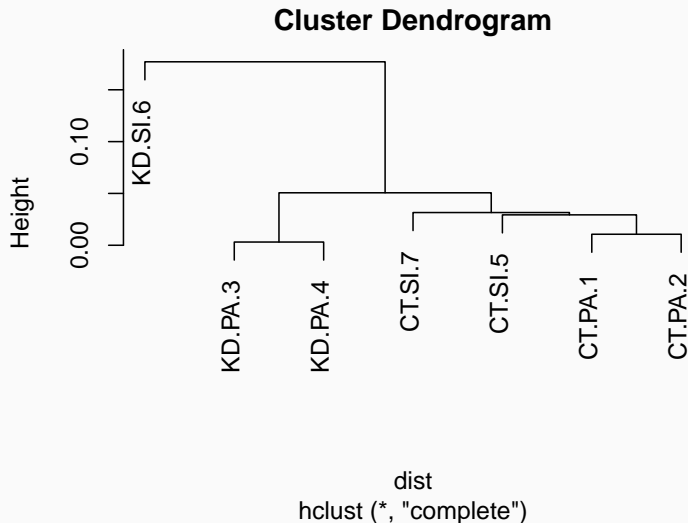
ESTIMAR OS FATORES DE NORMALIZAÇÃO

```
> d <- calcNormFactors(d)
> d$samples
```

##	group	lib.size	norm.factors
## CT.PA.1	CTL	8397136	0.9702373
## CT.PA.2	CTL	9909691	0.9652457
## KD.PA.3	KD	9664838	0.9973330
## KD.PA.4	KD	10325828	1.0146062
## CT.SI.5	CTL	19087995	1.0009795
## KD.SI.6	KD	15324886	1.0391230
## CT.SI.7	CTL	12812818	1.0145053

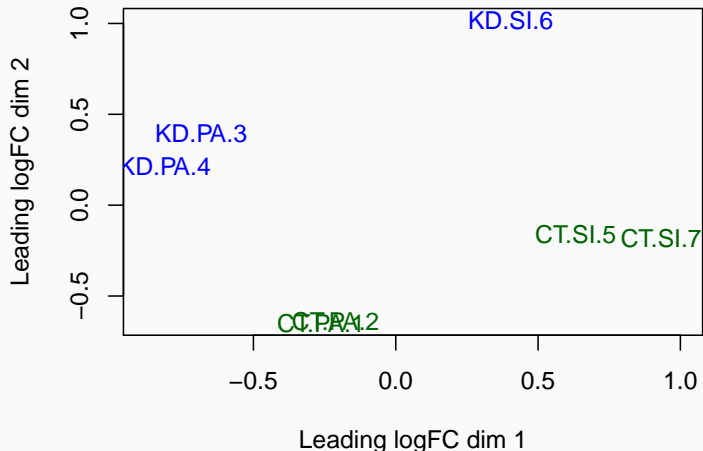
DENDROGRAMA

```
> dist <- as.dist(1 - cor(counts))  
> plot(hclust(dist))
```



INSPECIONAR AS RELAÇÕES ENTRE AS AMOSTRAS

```
> plotMDS(d, labels=samples$shortname,  
+ col=c("darkgreen","blue")[factor(samples$condition)])
```



```
> d <- estimateCommonDisp(d)
> d <- estimateTagwiseDisp(d)
```

GRÁFICO ENTRE A MÉDIA E A VARIÂNCIA

```
> plotMeanVar(d, show.tagwise.vars=TRUE, NBline=TRUE)
```

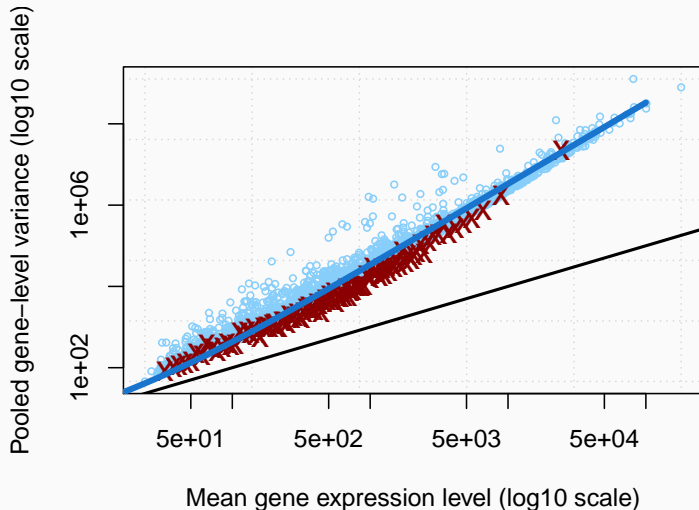
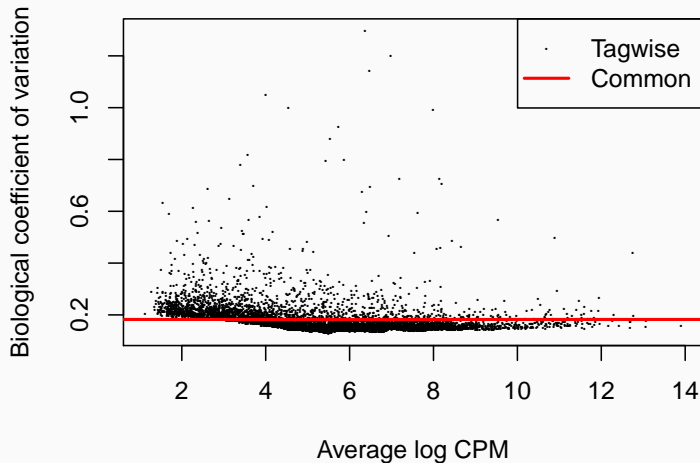


GRÁFICO ENTRE A MÉDIA E A VARIÂNCIA

```
> plotBCV(d)
```



```
> de <- exactTest(d, pair=c("CTL", "KD"))
```

CRIAR AS ESTATÍSTICAS DE DIFERENCIAÇÃO

```
> tt <- topTags(de, n=nrow(d))  
> head(tt$table)
```

##		logFC	logCPM	PValue	FDR
##	FBgn0039155	-4.614626	5.872116	1.053589e-96	7.581626e-93
##	FBgn0025111	2.931199	6.857715	7.621880e-58	2.742352e-54
##	FBgn0039827	-4.027050	4.398979	9.162157e-56	2.197696e-52
##	FBgn0003360	-3.181349	8.421436	1.192060e-54	2.144515e-51
##	FBgn0000071	2.708106	4.733580	1.795362e-40	2.558706e-37
##	FBgn0034736	-3.519673	4.130238	2.133440e-40	2.558706e-37

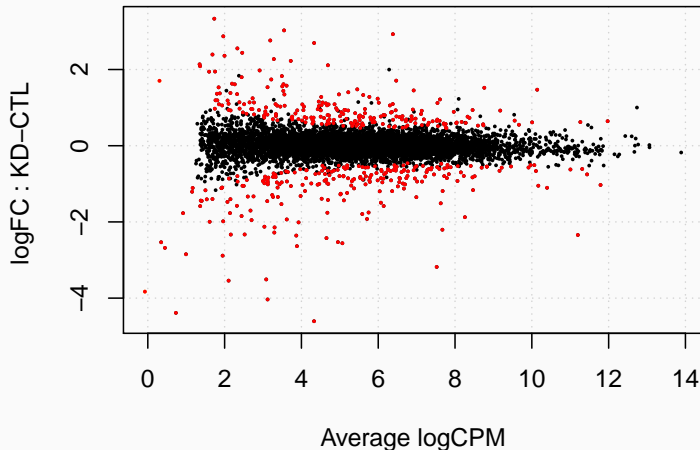
INSPECIONAR AS LEITURAS AJUSTADAS POR TAMANHO DE BIBLIOTECA PARA OS GENES COM MAIOR DIFERENCIAÇÃO

```
> nc <- cpm(d, normalized.lib.sizes=TRUE)
> rn <- rownames(tt$table)
> head(nc[rn, order(samples$condition)], 5)
```

##	CT.PA.1	CT.PA.2	CT.SI.5	CT.SI.7
## FBgn0039155	91.074075	97.958381	100.750047	106.780137
## FBgn0025111	34.244834	31.572498	26.639882	28.464446
## FBgn0039827	39.399970	36.695189	30.094170	34.465059
## FBgn0003360	448.619600	494.600960	589.636377	682.300456
## FBgn0000071	9.082859	9.199933	7.484289	5.846751
##	KD.PA.3	KD.PA.4	KD.SI.6	
## FBgn0039155	3.734803	4.963419	3.516607	
## FBgn0025111	247.430725	254.279776	188.389644	
## FBgn0039827	1.659913	2.768061	2.009490	
## FBgn0003360	62.557957	58.797426	61.791803	
## FBgn0000071	52.079758	55.933915	45.653090	

PLOT MA COM OS GENES COM DIFERENCIAÇÃO SIGNIFICATIVA

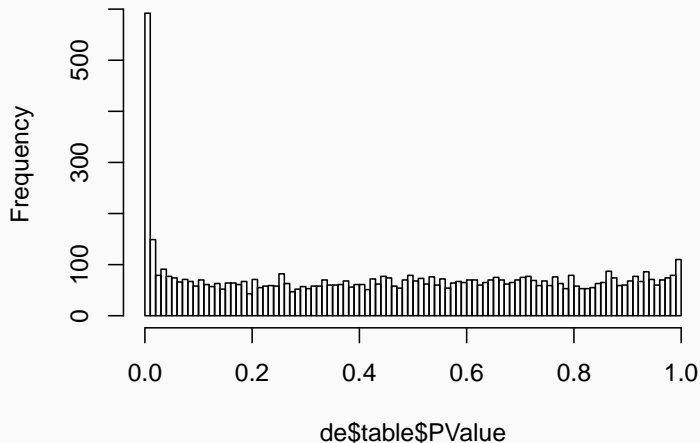
```
> deg <- rn[tt$table$FDR < 0.05]  
> plotSmea(d, de.tags=deg)
```



HISTOGRAMA DOS P-VALORES

```
> hist(de$table$PValue, breaks=100)
```

Histogram of de\$table\$PValue



CONCLUSÃO

- Big Data são 3 V: Volume, Velocidade, Variedade
- Análise de dados genéticos envolve diversas áreas do conhecimento
- É possível trabalhar em vários projetos da área, em equipes multidisciplinares

REFERÊNCIAS

REFERÊNCIAS

- [1] S Anders, D J McCarthy, Y Chen, M Okoniewski, G K Smyth, W Huber, and M D Robinson. Count-based differential expression analysis of RNA sequencing data using R and Bioconductor. *Nature protocols*, 8(9):1765–1786, 2013.
- [2] Paul Livermore Auer. *Statistical design and analysis of next-generation sequencing data*. PhD thesis, 2010.
- [3] Angela N. Brooks, Li Yang, Michael O. Duff, Kasper D. Hansen, Jung W. Park, Sandrine Dudoit, Steven E. Brenner, and Brenton R. Graveley. Conservation of an RNA regulatory map between *Drosophila* and mammals. *Genome Research*, 21(2):193–202, 2011.
- [4] Mark D Robinson, Davis J McCarthy, and Gordon K Smyth. edgeR: a Bioconductor package for differential expression analysis of digital gene expression data. *Bioinformatics*, 26(1):139–40, jan 2010.

ANÁLISE DE DADOS GENÉTICOS: UM PROBLEMA DE BIG DATA A CADA NOVO PACIENTE

RBRAS 2016 - SALVADOR, BA

Marcus Nunes

24 e 25 de maio de 2016

Universidade Federal do Rio Grande do Norte