

Análise de Dados Genéticos

Marcus Nunes

Departamento de Estatística - UFJF

2 e 3 de Outubro de 2014

1 Introdução

2 DNA

3 Métodos

4 Aplicação

Malcolm Gladwell

Especialista - 10.000 horas

Aprendiz - 2.000 horas

Deficiência Crítica

Quem já jogou RPG?

NOME DO PERSONAGEM _____ JOGADOR _____

CLASSE E NÍVEL _____ RAÇA _____ TENDÊNCIA _____ DIVINDADE _____

TAMANHO _____ IDADE _____ SEXO _____ ALTURA _____ PESO _____ OLHOS _____ CABELOS _____ PELE _____



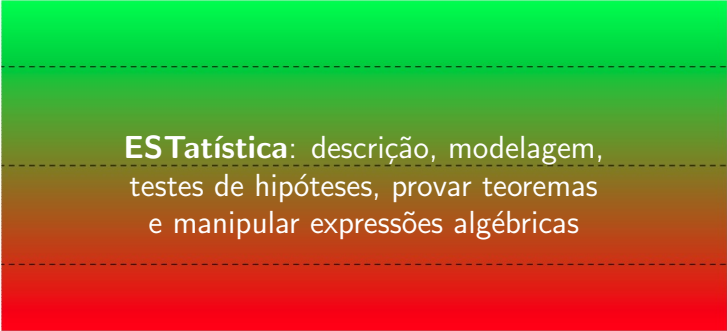
PLANILHA DE PERSONAGEM

HABILIDADE	VALOR	MOD. DE HABILIDADE	VALOR TEMPORÁRIO	MOD. TEMPORÁRIO	TOTAL	FERIMENTOS / PVs ATUAIS	DANO POR CONTUSÃO	DESLOCAMENTO
FOR FORÇA								
DES DESTREZA								
CON CONSTITUIÇÃO								
INT INTELIGÊNCIA								
SAB SABEDORIA								
CAR CARISMA								

PV	CA	TOQUE	SURPRESA	INICIATIVA	TESTE DE RESISTÊNCIA	BÔNUS BASE DE ATAQUE	RESISTÊNCIA À MAGIA	AGARRAR
PV PONTOS DE VIDA	CA CLASSE DE ARMADURA	TOQUE CLASSE DE ARMADURA	SURPRESA CLASSE DE ARMADURA	INICIATIVA MODIFICADOR	FORTITUDE (CONSTITUIÇÃO)			
					REFLEXOS (DESTREZA)			
					VONTADE (SABEDORIA)			

PERÍCIAS	HABILIDADE CHAVE	MOD. DE PERÍCIAS	MOD. DE HABILIDADE	GRADUAÇÃO MÁXIMA (CLASSE / OUTRA CLASSE)	GRADUAÇÃO	OUTROS
<input type="checkbox"/> ABRIR FECHADURAS	DES					
<input type="checkbox"/> ACROBACIA ■	DES*					
<input type="checkbox"/> ADESTRAR ANIMAIS	CAR					
<input type="checkbox"/> ARTE DA FUGA ■	DES*					
<input type="checkbox"/> ATUAÇÃO ()	CAR					
<input type="checkbox"/> ATUAÇÃO ()	CAR					
<input type="checkbox"/> ATUAÇÃO ()	CAR					
<input type="checkbox"/> AVALIAÇÃO ■	INT					
<input type="checkbox"/> BLEFAR ■	CAR					
<input type="checkbox"/> CAVALGAR ■ ()	DES					
<input type="checkbox"/> CONCENTRAÇÃO ■	CON					
<input type="checkbox"/> CONHECIMENTO ()	INT					
<input type="checkbox"/> CONHECIMENTO ()	INT					

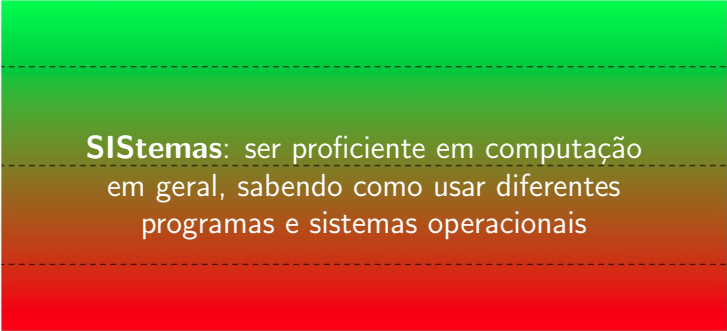
O que um Bioestatístico precisa saber?



ESTatística: descrição, modelagem,
testes de hipóteses, provar teoremas
e manipular expressões algébricas

EST

O que um Bioestatístico precisa saber?

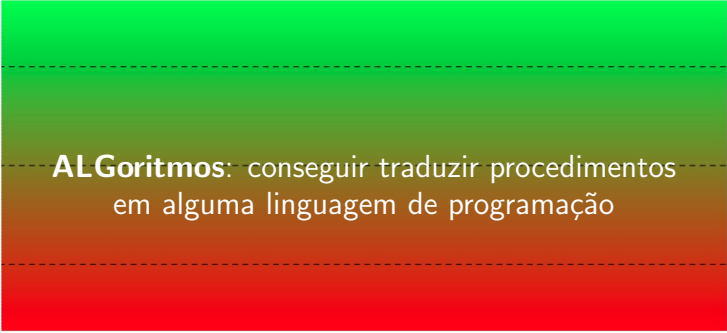


SIStemas: ser proficiente em computação
em geral, sabendo como usar diferentes
programas e sistemas operacionais

EST

SIS

O que um Bioestatístico precisa saber?



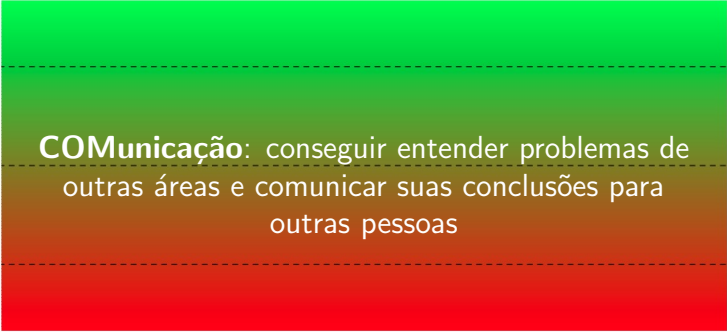
ALGoritmos: conseguir traduzir procedimentos
em alguma linguagem de programação

EST

SIS

ALG

O que um Bioestatístico precisa saber?



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

EST

SIS

ALG

COM

O que um Bioestatístico precisa saber?



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

EST

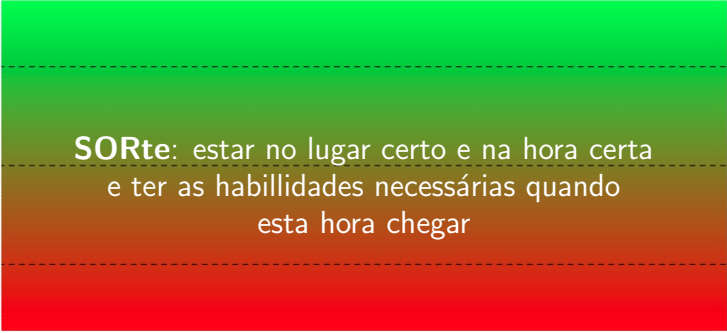
SIS

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

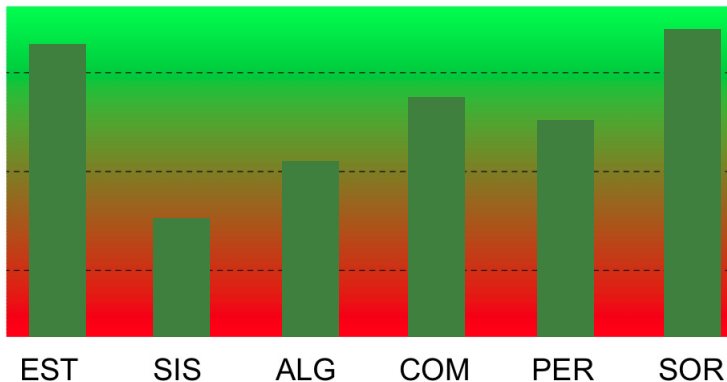
ALG

COM

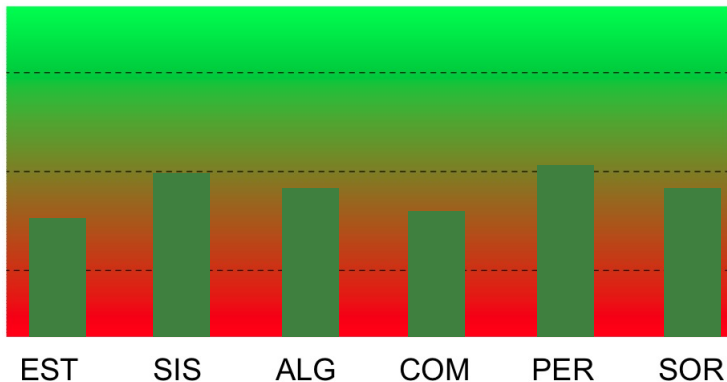
PER

SOR

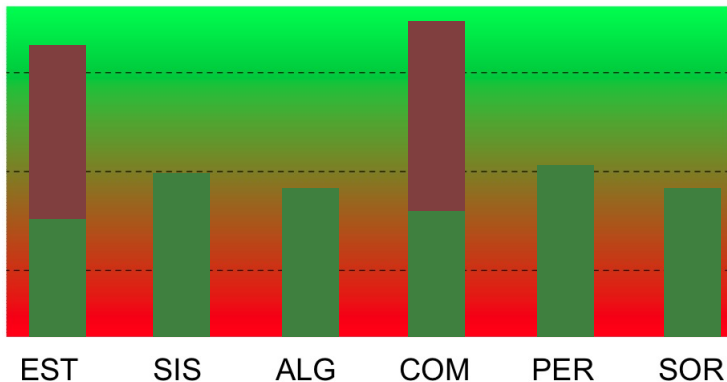
O que eu sei



Estar na média não é bom



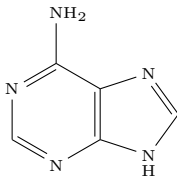
Seja muito bom em algumas áreas



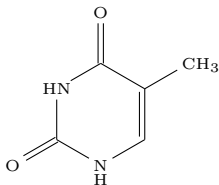
DNA

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

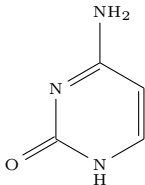
Estrutura Química do DNA



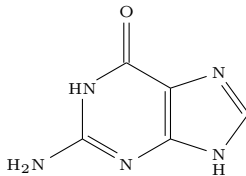
Adenina



Timina



Citosina



Guanina

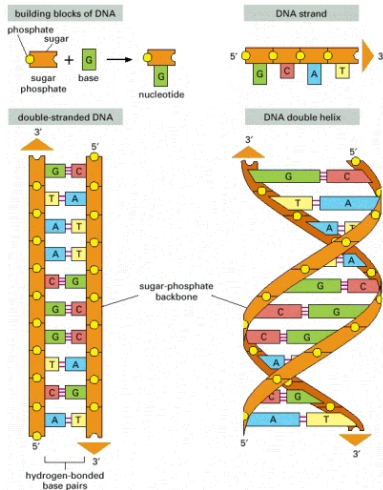
Mas para que serve o DNA?

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

Mais sobre DNA

- 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

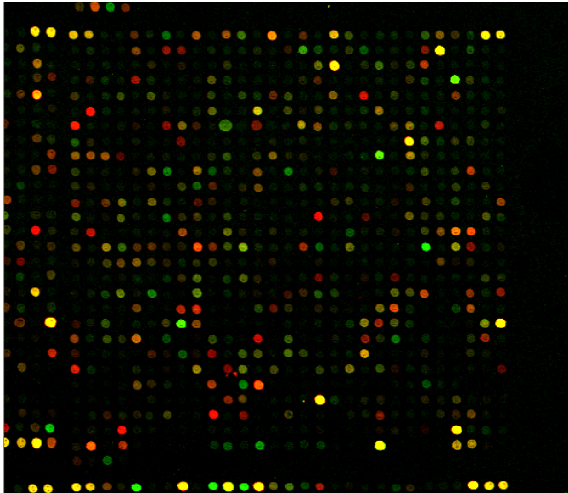
Tecnologias de Sequenciamento

- Sanger sequencing
- Microarrays
- RNA-Seq

Sanger Sequencing

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

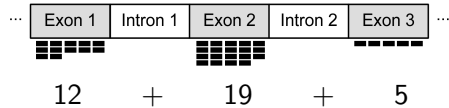
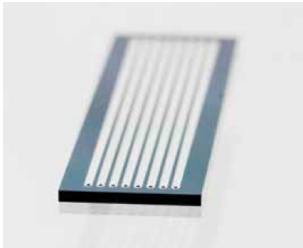
Microarrays



Microarrays

- Estão caindo em desuso
- Preço entre US\$ 190 e US\$ 450 por array
- Maior disponibilidade no mercado

RNA-Seq



RNA-Seq

- Método cuja utilização vem crescendo mais ultimamente
- Cada sequenciamento custa entre US\$ 805 e US\$1.700 (em setembro de 2014)
- Leva 8 horas para ficar pronto

MiniON

- Tecnologia sendo desenvolvida na universidade de Oxford
- Cada chip de sequenciamento custará US\$ 900
- A análise pode ser feita em tempo real

MiniON



Pipeline

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

Preparação da Amostra e Sequenciamento

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

Exemplo de Sequenciamento

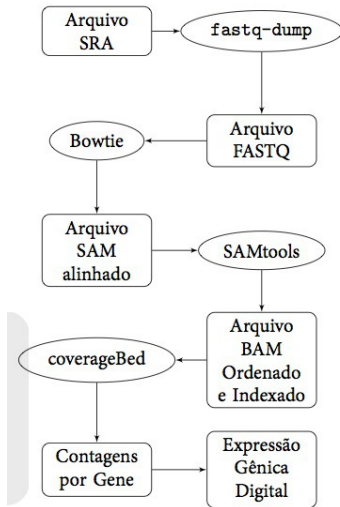
```
@SRR014849.1 EIXKN4201CFU84 length=50  
GGGGGGGGGGGGGGGGGGCTTTTTTTGTTTGGAACCGAAAGGGTTTTGAAT  
+SRR014849.1 EIXKN4201CFU84 length=50  
3+&$#"7F@71,'";C?,B;?6B;:EA1EA 1EA5'9B:
```

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

Alinhamento das Leituras

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

Fluxograma do Pipeline



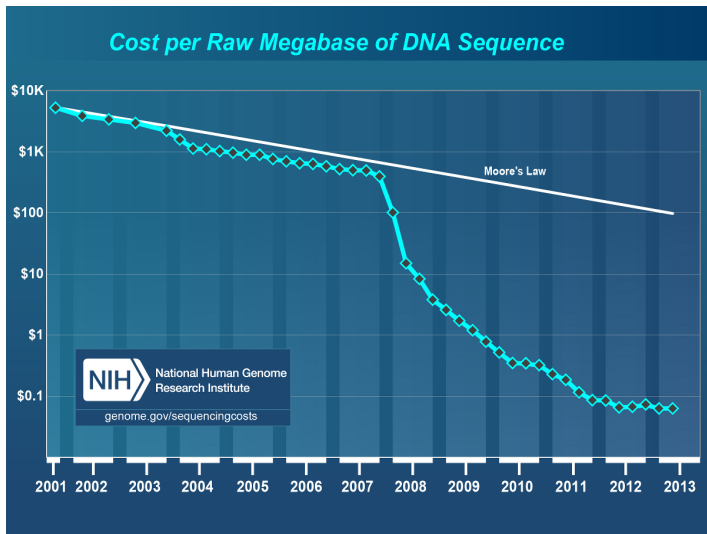
Recursos

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

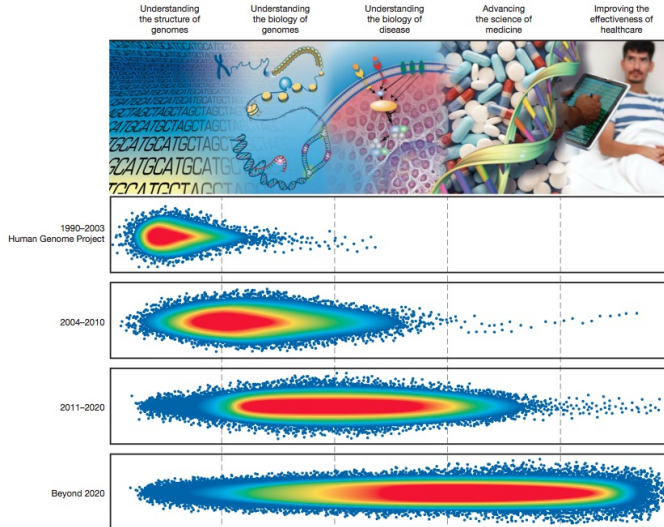
Sequenciar Genomas é Cada Vez Mais Barato

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

Custo de Sequenciamento



Futuro da Genômica



Planejamento de Experimentos

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

Planejamento de Experimentos

- Amostragem
- Replicação
- Agrupamento em blocos
- Aleatorização

Amostragem

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

Aleatorização

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

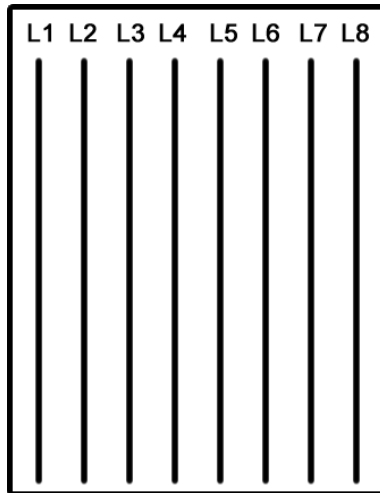
Replicação

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

Agrupamento em blocos

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

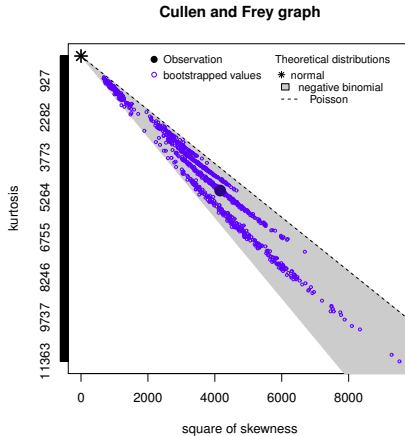
Agrupamento em blocos



Modelagem

- Dados discretos
- Não-normalidade
- Testes múltiplos

Distribuição de Contagens



Distribuições

Poisson

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

Binomial Negativa

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

Modelos Lineares Generalizados

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

Modelos Lineares Generalizados

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

Excesso de Zeros

- Experimentos RNA-Seq geram um grande número de zeros
- Normalmente, estas observações são descartadas na análise
- O excesso de zeros cria sobredispersão
- Propomos um método que é capaz de lidar com o excesso de zeros

Modelos Hurdle e Zero-Inflated

$$f_H(y) = \begin{cases} p, & \text{if } y = 0 \\ (1 - p) \frac{f(y)}{1 - f(0)}, & \text{if } y = 1, 2, 3, \dots \end{cases}$$

$$f_{ZI}(y) = \begin{cases} q + (1 - q)f(0), & \text{if } y = 0 \\ (1 - q)f(y), & \text{if } y = 1, 2, 3, \dots \end{cases}$$

Máxima Verossimilhança - Modelo Hurdle

$$\begin{aligned} L(y|p, \beta, \phi) &= \prod_{i=1}^n (1-p)^{I(y=0)} \\ &\quad \times \left[p \frac{\Gamma(y + \phi^{-1})}{\Gamma(\phi^{-1})\Gamma(y+1)} \left(\frac{\mu}{\phi^{-1} + \mu} \right)^y \left(\frac{1}{(1 + \phi\mu)^{\phi^{-1}} - 1} \right)^{-\phi^{-1}} \right]^{1-I(y=0)} \\ &= \prod_{i=1}^n (1-p)^{I(y=0)} p^{1-I(y=0)} \\ &\quad \times \left[\frac{\Gamma(y + \phi^{-1})}{\Gamma(\phi^{-1})\Gamma(y+1)} \left(\frac{\mu}{\phi^{-1} + \mu} \right)^y \left(\frac{1}{(1 + \phi\mu)^{\phi^{-1}} - 1} \right)^{-\phi^{-1}} \right]^{1-I(y=0)} \\ \mathcal{L}(y|p, \beta, \phi) &= \log(L(y|p, \beta, \phi)) \\ &= \log(L_1(y|p)L_2(y|\beta, \phi)) \\ &= \log(L_1(y|p)) + \log(L_2(y|\beta, \phi)) \\ &= \mathcal{L}_1(y|p) + \mathcal{L}_2(y|\beta, \phi) \end{aligned}$$

Máxima Verossimilhança - Modelo Hurdle

$$\begin{aligned}\mathcal{L}(y_i|\mu, \phi) &= \log\left(\prod_{i=1}^n f(y_i|\mu, \phi)\right) \\&= \log\left\{\prod_{i=1}^n \frac{\Gamma(\phi^{-1})}{\Gamma(\phi^{-1})\Gamma(y_i + 1)} \left(\frac{1}{1 + \phi\mu}\right)^{\phi^{-1}} \left(\frac{\phi\mu}{1 + \phi\mu}\right)^{y_i}\right\} \\&= \sum_{i=1}^n \left\{y_i \log\left(\frac{1}{1 + \mu\phi}\right) - \phi^{-1} \log(1 + \phi\mu) + \log \Gamma(y_i + \phi^{-1})\right. \\&\quad \left.- \log \Gamma(y_i + 1) + \log \Gamma(\phi^{-1})\right\} \\&= \sum_{i=1}^n \left\{y_i \log(\mu\phi) - (y_i + \phi^{-1}) \log(1 + \phi\mu) + \log \Gamma(y_i + \phi^{-1})\right. \\&\quad \left.- \log \Gamma(y_i + 1) - \log \Gamma(\phi^{-1})\right\}\end{aligned}$$

Máxima Verossimilhança - Modelo Hurdle

$$\text{logit}(p) = x_i'(1 - p)$$

$$\log(\mu_i) = x_i'\beta$$

Algoritmo IRLS

- 1 Inicializar β e o preditor linear η
- 2 Calcular os pesos $W^{-1} = \mathbf{V}g'(\beta)$, onde \mathbf{V} é a variância dada por $\kappa''(\theta)$, onde $\theta = [\beta, \phi]$
- 3 Calcular $\mathbf{Z} = \eta + (\mathbf{Y} - \beta)g'(\beta)$
- 4 Regressar \mathbf{Z} nos preditores x_1, x_2, \dots, x_n com pesos W para obter atualizações para β
- 5 Calcular η baseado nas estimações da regressão
- 6 Calcular β como $g^{-1}(\eta)$
- 7 Calcular a função de log-verossimilhança
- 8 Iterar até a convergência

Algoritmo IRLS

$$x_1 = x_0 - \frac{f(x_0)}{f'(x_0)} \Rightarrow x_{n+1} = x_n - \frac{f(x_n)}{f'(x_n)}$$

$$H = \left[\frac{\partial^2 \mathcal{L}}{\partial \beta_j \partial \beta_k} \right]$$

$$H = \begin{pmatrix} \frac{\partial^2 \mathcal{L}}{\partial \beta_1^2} & \frac{\partial^2 \mathcal{L}}{\partial \beta_1 \beta_2} & \cdots & \frac{\partial^2 \mathcal{L}}{\partial \beta_1 \beta_n} \\ \frac{\partial^2 \mathcal{L}}{\partial \beta_2 \beta_2} & \frac{\partial^2 \mathcal{L}}{\partial \beta_2^2} & \cdots & \frac{\partial^2 \mathcal{L}}{\partial \beta_2 \beta_n} \\ \vdots & \vdots & \ddots & \vdots \\ \frac{\partial^2 \mathcal{L}}{\partial \beta_n \beta_1} & \frac{\partial^2 \mathcal{L}}{\partial \beta_n \beta_2} & \cdots & \frac{\partial^2 \mathcal{L}}{\partial \beta_n^2} \end{pmatrix}.$$

Teste de Razão de Verossimilhança

$$H_0 : \beta_q = \beta_{q+1} = \cdots = \beta_{p-1}$$

$$H_1 : \text{nem todos } \beta_k \text{ em } H_0 \text{ são iguais}$$

$$\Lambda(\mathbf{X}) = \frac{\sup_{\theta \in \Theta_0} L(\theta|\mathbf{X})}{\sup_{\theta \in \Theta_1} L(\theta|\mathbf{X})}$$

$$\Lambda(\mathbf{X}) = -2 \log \left(\frac{L_0(\theta|\mathbf{X})}{L_1(\theta|\mathbf{X})} \right),$$

$$\Lambda(\mathbf{X}) \xrightarrow{D} \chi_p^2$$

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

edgeR

- 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

edgeR

- 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 $\text{NB}(m^* \lambda, \phi)$

edgeR

- 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

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

- 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

edgeR

- 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,

$$Z_{tI} \sim \text{NB}(n_I m^* \lambda_t, \phi n_I^{-1}), \quad I \in \{A, B\}$$

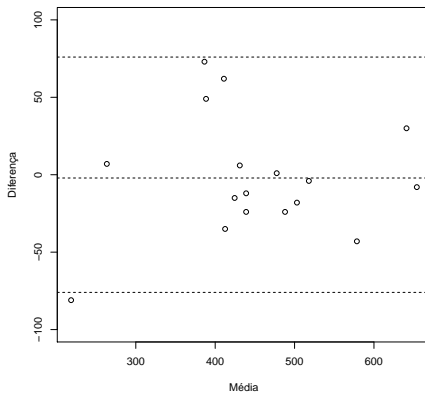
- Condicionando na soma das pseudo-contagens totais, $Z_{tA} + Z_{tB}$ também é uma variável aleatória Binomial Negativa

MA Plot

- 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

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

MA Plot



Comparações Múltiplas

- 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

Comparações Múltiplas

- 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

Correção de Bonferroni

- 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:

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

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

Correção de Bonferroni

- 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

FDR

- 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

FDR

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

FDR

- 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 Blekhman *et al.* (2010)
- As amostras foram obtidas a partir dos fígados de machos e fêmeas de três espécies de primatas: humanos (*Homo sapiens*), chimpanzés (*Pan troglodytes*) e macacos-rhesus (*Macaca mulatta*)

Análise Exploratória dos Dados

- 3 replicações biológicas de cada tratamento
- Cada replicação biológica foi dividida em 2 faixas
- Foram coletadas 20689 características de cada uma das 36 faixas
- No total, são 71 milhões de leituras de 35 pares de base mapeáveis

Colunas da Matriz

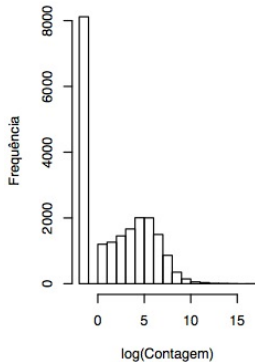
[1]	"EnsemblGeneID"	"R1L1.HSM1"	"R1L2.PTF1"
[4]	"R1L3.RMM1"	"R1L4.HSF1"	"R1L6.PTM1"
[7]	"R1L7.RMF1"	"R2L2.RMF2"	"R2L3.HSM2"
[10]	"R2L4.PTF2"	"R2L6.RMM2"	"R2L7.HSF2"
[13]	"R2L8.PTM2"	"R3L1.RMM3"	"R3L2.HSF2"
[16]	"R3L3.PTM1"	"R3L4.RMF3"	"R3L6.HSM3"
[19]	"R3L7.PTF3"	"R3L8.RMM1"	"R4L1.HSM3"
[22]	"R4L2.HSF1"	"R4L3.RMM3"	"R4L4.PTF1"
[25]	"R4L6.PTM2"	"R4L7.RMF3"	"R4L8.HSM2"
[28]	"R5L1.RMF1"	"R5L2.HSM1"	"R5L3.PTF3"
[31]	"R5L4.RMM2"	"R5L8.RMF2"	"R6L2.PTM3"
[34]	"R6L4.PTM3"	"R6L6.PTF2"	"R8L1.HSF3"
[37]	"R8L2.HSF3"		

Contagens dos Genes

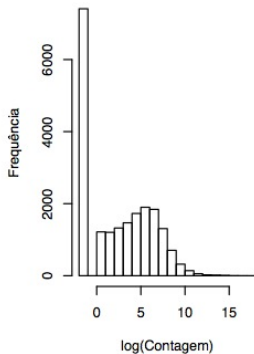
	EnsemblGeneID	R1L1.HSM1	R1L2.PTF1	R1L3.RMM1
1	ENSG000000000003	60	285	207
2	ENSG000000000005	0	1	1
3	ENSG000000000419	17	54	20
4	ENSG000000000457	50	61	68
5	ENSG000000000460	9	6	2
6	ENSG000000000938	32	50	44
7	ENSG000000000971	1726	2617	7207
8	ENSG000000001036	99	135	109
9	ENSG000000001084	155	211	455
10	ENSG000000001167	13	35	52

Histogramas

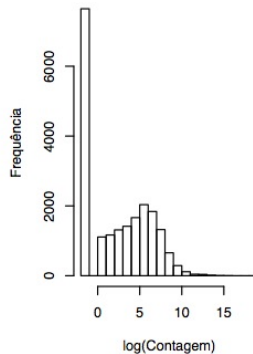
R1L1.HSM1



R1L2.PTF1

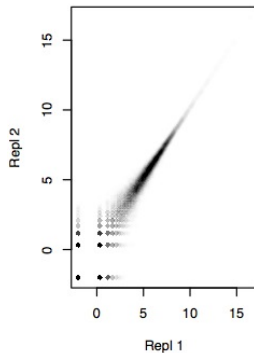


R1L3.RMM1

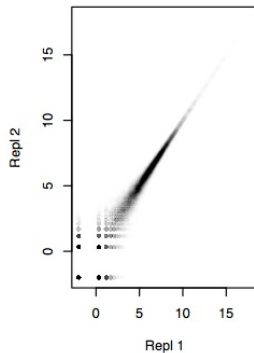


Gráficos de Dispersão

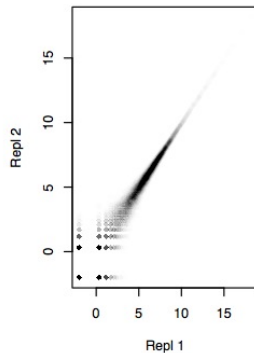
HSM1



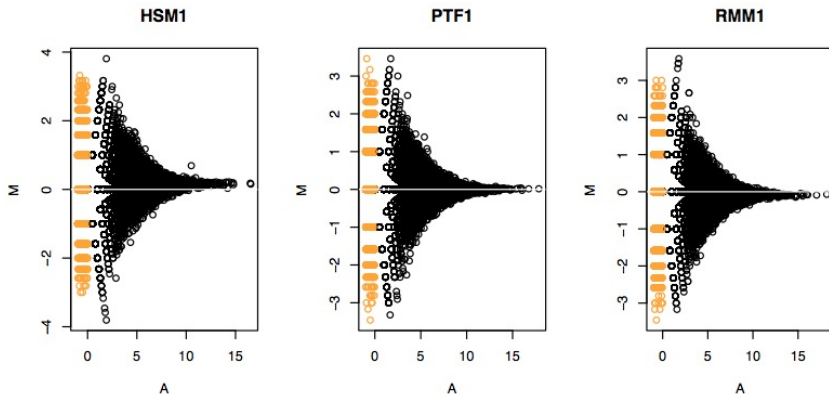
PTF1



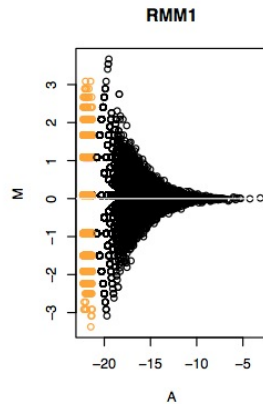
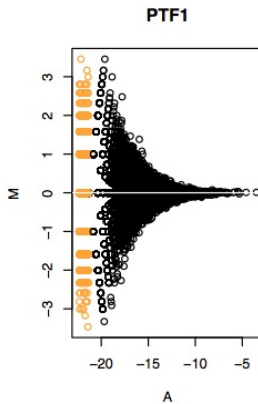
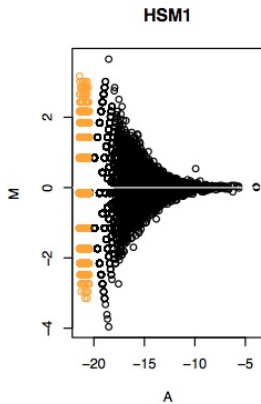
RMM1



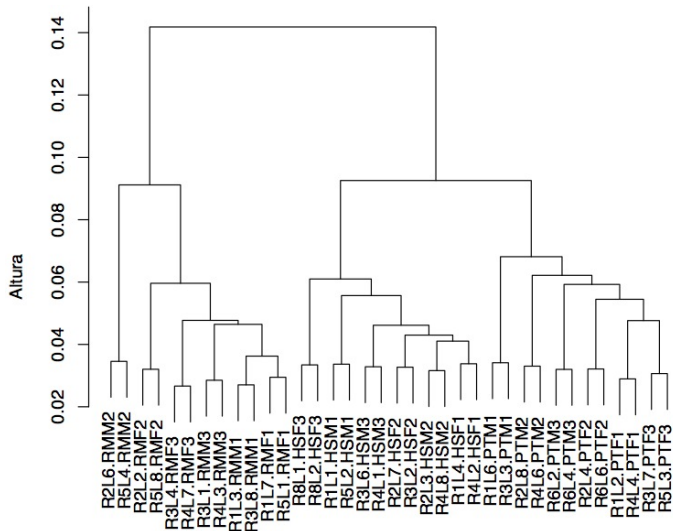
MA Plot



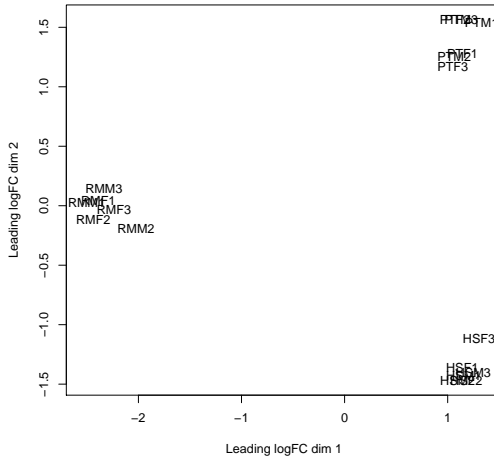
MA Plot Normalizado



Dendrograma



MDS Plot



Teste Exato

```
> HSM.vs.HSF <- exactTest(d, pair = c(1, 2))
> HSM.vs.HSF

## An object of class "DGEEexact"
## $table
##           logFC  logCPM PValue
## ENSG000000000003 -0.06484  6.9082 0.8985
## ENSG000000000005  1.79476 -0.5125 1.0000
## ENSG000000000419 -0.35879  4.1865 0.4994
## ENSG000000000457  0.18363  5.0917 0.7215
## ENSG000000000460  0.02462  1.4939 1.0000
## 14976 more rows ...
##
## $comparison
## [1] "HSF" "HSM"
##
## $genes
## [1] "ENSG000000000003" "ENSG000000000005"
## [3] "ENSG000000000419" "ENSG000000000457"
## [5] "ENSG000000000460"
## 14976 more rows ...
```

Estimação Tagwise - HSM versus MSF

```
> d10 <- estimateTagwiseDisp(d, prior.df = 10)
> TW.HSM.vs.HSF <- exactTest(d10, pair = c(1,
+      2))
> head(TW.HSM.vs.HSF$table)
```

##		logFC	logCPM	PValue
##	ENSG000000000003	-0.06455	6.9082	0.8861
##	ENSG000000000005	1.78073	-0.5125	1.0000
##	ENSG000000000419	-0.35694	4.1865	0.3604
##	ENSG000000000457	0.18184	5.0917	0.6731
##	ENSG000000000460	0.02430	1.4939	1.0000

Estimação Tagwise - HSM versus PTM

```
> TW.HSM.vs.PTM <- exactTest(d10, pair = c(2,  
+      4))  
> head(TW.HSM.vs.PTM$table)
```

##		logFC	logCPM	PValue
##	ENSG000000000003	0.9215	6.9082	0.039618
##	ENSG000000000005	0.8848	-0.5125	1.000000
##	ENSG000000000419	0.3491	4.1865	0.362794
##	ENSG000000000457	1.2639	5.0917	0.003564
##	ENSG000000000460	-0.3789	1.4939	0.536722
##	ENSG000000000938	0.7353	4.4394	0.258405

Estimação Tagwise - HSM versus HSF - FDR

```
> topTags(TW.HSM.vs.HSF, adjust.method = "BH")
```

```
## Comparison of groups: HSM-HSF
```

##	genes	logFC	logCPM	PValue	FDR
## 6515	ENSG00000138131	5.089	5.3944	2.756e-07	0.004129
## 9245	ENSG00000163017	-3.058	3.8267	3.206e-05	0.214623
## 11722	ENSG00000178297	2.189	3.1773	4.298e-05	0.214623
## 5185	ENSG00000128285	-2.304	1.9543	7.879e-05	0.245900
## 11344	ENSG00000175084	-2.163	4.9850	9.342e-05	0.245900
## 6819	ENSG00000140403	-3.190	4.1706	9.848e-05	0.245900
## 4457	ENSG00000120694	-2.669	6.3453	1.234e-04	0.264174
## 12473	ENSG00000185031	-4.979	0.9932	2.070e-04	0.387569
## 750	ENSG00000060566	1.866	5.7418	4.106e-04	0.683441
## 8849	ENSG00000160181	-4.296	0.5679	6.616e-04	0.991113

Estimação Tagwise - HSM versus PTM - FDR

```
> topTags(TW.HSM.vs.PTM, adjust.method = "BH")
```

```
## Comparison of groups: PTM-HSM
```

##	genes	logFC	logCPM	PValue	FDR
## 14119	ENSG00000208587	14.776	7.931	9.663e-64	1.448e-59
## 14118	ENSG00000208570	14.097	8.107	2.310e-54	1.731e-50
## 8570	ENSG00000157399	-6.777	5.252	1.262e-32	6.303e-29
## 5911	ENSG00000134391	-9.228	6.204	2.453e-29	9.186e-26
## 399	ENSG00000025423	-6.002	8.026	4.218e-28	1.264e-24
## 9325	ENSG00000163444	5.757	4.229	6.610e-26	1.650e-22
## 6712	ENSG00000139540	-8.208	6.585	1.885e-24	4.034e-21
## 1901	ENSG00000099834	-6.730	6.475	2.650e-23	4.963e-20
## 14854	ENSG00000219802	9.927	3.164	1.353e-22	2.252e-19
## 14936	ENSG00000220688	4.526	5.146	5.968e-20	8.940e-17

Site do Minicurso

`http://marcusnunes.me/mgest-2014-minicurso/`

Análise de Dados Genéticos

Marcus Nunes

Departamento de Estatística - UFJF

2 e 3 de Outubro de 2014