ANÁLISES ESTATÍSTICAS DE EXPERIMENTOS DE RNASEQ: UMA BREVE ABORDAGEM AOS DESENHOS MULTIFATORIAIS E AOS MODELOS MISTOS

Otávio José Bernardes Brustolini
Laboratório de Bioinformática
Laboratório Nacional de Computação Científica

CIÊNCIA

Curiosidade, dúvida ou observação → Gera perguntas

↓

Sistematização

↓

Modelos

Perguntas:

Como a inteligencia artificial (IA) impacta na ciência?

Os programadores ficarão obsoletos?

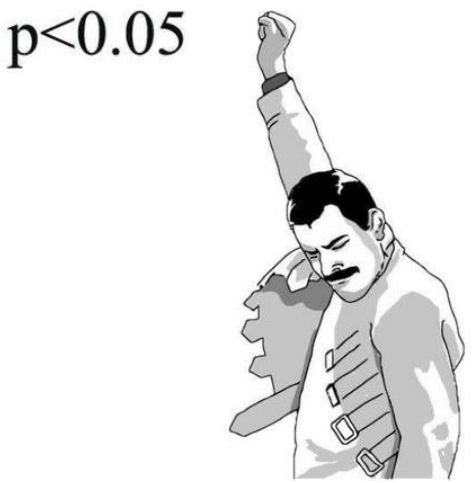
Os bioinformatas virarão biologos teóricos de prompt de IA?

Quais genes respondem a infeção por dengue?

Qual a expressão dos genes nos tecidos cancerigenos comparados com os saudáveis?

ESTATÍSTICA

É a ciência que utiliza as teorias probabilísticas para explicar a frequência da ocorrência de eventos, tanto em estudos observacionais quanto em experimentos para modelar a aleatoriedade e a incerteza de forma a estimar ou possibilitar a previsão de fenômenos futuros, conforme o caso.



* Ferramentas certas para cada tipo de trabalho

"Se duvidarem de sua narrativa, lembre-se das probabilidades"

EXPERIMENTAÇÃO CIENTÍFICA

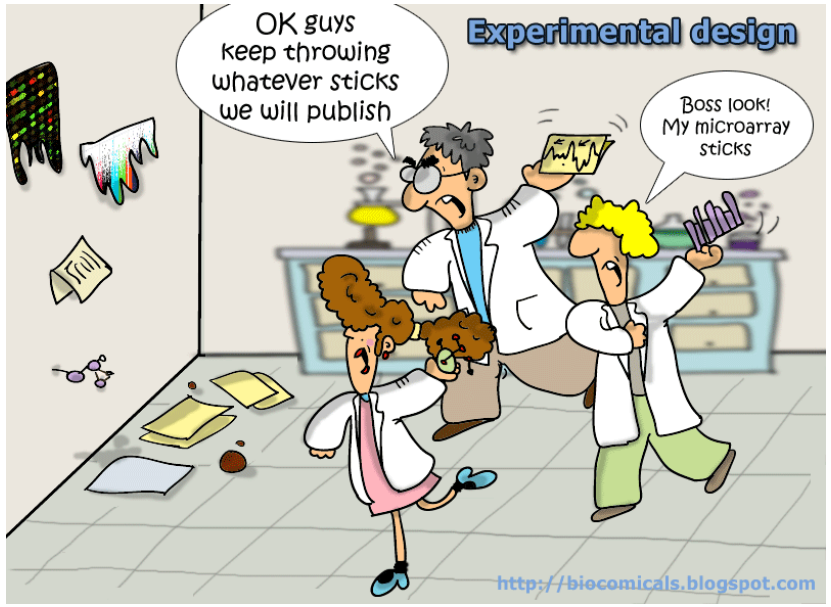
- Processo sistemático que manipula variáveis controladas em um ***ambiente controlado*** para observar e medir os efeitos dessas "manipulações" em outras variáveis.
- Principais elementos da experimentação científica incluem:
  - **Hipótese**: Uma suposição ou previsão baseada em observações iniciais, que será testada.
  - **Variável independente**: A variável que é manipulada ou alterada.
  - **Variável dependente**: A variável que é observada ou medida, afetada pelas alterações na variável independente.
  - **Variáveis controladas**: Fatores que são mantidos constantes durante o experimento para garantir que os resultados sejam atribuídos à variável independente.
- **Reprodutibilidade**: A capacidade de repetir o experimento e obter os mesmos resultados, garantindo a confiabilidade das conclusões.
- **Análise de dados**: A coleta e interpretação dos dados obtidos, usando ferramentas estatísticas para verificar a validade da hipótese.

Fonte: ChatGPT 4o 02/09/24

DESENHO EXPERIMENTAL

Inicia-se com o planejamento e a estrutura ou plano que orienta a condução de um experimento, definindo como os dados serão coletados, analisados e interpretados. Ele envolve a organização de ***fatores controláveis*** e variáveis independentes, a fim de testar hipóteses ou responder perguntas de pesquisa de forma objetiva e reprodutível.

*** Experimentos independentes ou pareados**



Fonte: ChatGPT 4o 02/09/24

DESENHO EXPERIMENTAL

- **Grupo controle e grupo experimental**: Um grupo controle é usado para comparar com o grupo experimental, que recebe a intervenção ou tratamento, permitindo isolar os efeitos específicos da variável testada.
- **Randomização**: Atribuição aleatória de indivíduos aos grupos experimentais ou de controle para evitar vieses e garantir que as diferenças observadas sejam devidas à intervenção.
- **Replicação**: Realização de experimentos em múltiplas amostras ou sujeitos para garantir que os resultados não sejam frutos do acaso.

Fonte: ChatGPT 4o 02/09/24

DESENHO EXPERIMENTAL / VARIÁVEIS

- **Independentes**: São os fatores manipulados pelo pesquisador (como tratamentos ou condições experimentais).
- **Dependentes**: São as respostas ou resultados medidos (como a eficácia de um tratamento).
- **Controle**: São mantidas constantes para evitar que influenciem os resultados.
- **Covariáveis:** variáveis independentes que podem influenciar ou estar associadas ao resultado de um estudo, mas que não são o foco principal da investigação.

TIPOS DE EFEITOS NO MODELO LINEAR

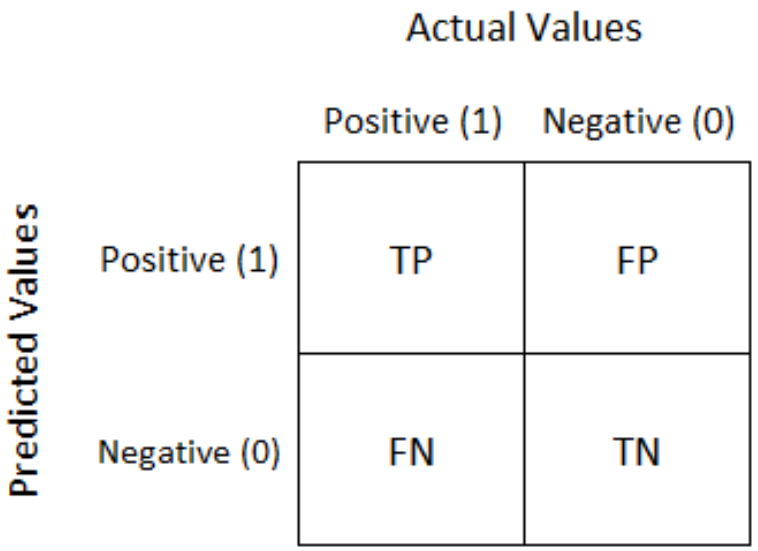
- **Efeito fixo**: constantes ou comuns para toda a população ou amostra.
- **Efeito aleatório**: variação entre diferentes grupos ou unidades de amostra que podem estar relacionadas de maneira hierárquica ou aninhada. São tratados como amostras de uma distribuição.
- Exemplos:
  - **Estudos Longitudinais**: Modelos mistos são frequentemente usados para analisar dados de estudos longitudinais, onde várias observações são feitas no mesmo indivíduo ao longo do tempo.
  - **Dados Hierárquicos**: Em situações onde os dados estão estruturados em diferentes níveis, como estudantes dentro de turmas, turmas dentro de escolas, os modelos mistos permitem capturar a variação entre os diferentes níveis (escola, turma, estudante).
  - **Estudos Multicêntricos**: Em estudos com dados coletados em diferentes locais (por exemplo, hospitais, centros de pesquisa), os modelos mistos podem capturar a variação entre os locais ao incorporar efeitos aleatórios.

FONTES DE VARIABILIDADE

- **Erro**: refere-se a variações imprevisíveis e aleatórias que ocorrem durante a coleta de dados ou medições. Esses erros são causados por fatores imprevisíveis ou desconhecidos que afetam as medições de forma diferente a cada vez.
- **Viés**: distorção consistente e unidirecional nos resultados. É causado por fatores que afetam as medições de maneira constante e previsível. Resulta de problemas no próprio desenho experimental, no método de coleta de dados ou na seleção dos participantes, afetando de maneira previsível os resultados.
- **Confundimento**: relação observada entre uma variável independente (fator que está sendo manipulado ou estudado) e uma variável dependente (resultado ou efeito) é influenciada por uma terceira variável não controlada, conhecida como **variável de confusão** ou **confundidora**. Essa variável confundidora está relacionada tanto com a variável independente quanto com a dependente, o que pode distorcer ou mascarar a verdadeira relação entre elas.

MATRIZ DE CONFUSÃO

- Falso Positivo (erro tipo I) x Falso Negativo (erro tipo II)
- Verdadeiro Positivo (TP) x Verdadeiso Negativo (TN)
- Análise ideal:
  - Maximizar o TP e TN
  - Minimizar o FP e FN
- Métricas: precisão, acurácia, recall, especificidade, etc



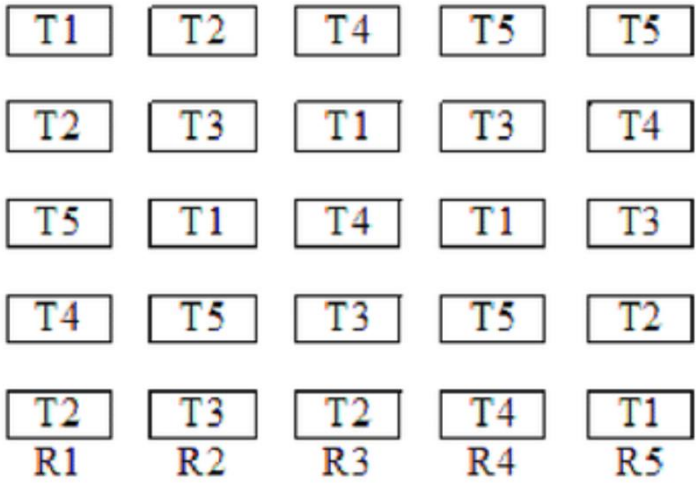
TESTE DE HIPÓTESE

- Inferências sobre uma população com base em uma amostra de dados.
- Verificar a validade de uma suposição (hipótese) sobre um parâmetro populacional, como a média, proporção ou variância, por meio da análise dos dados amostrais.
- Hipótese nula (H_0): A hipótese de que não há efeito, diferença ou relação significativa no experimento ou teste.
- Hipótese alternativa (H_a): A hipótese de que existe um efeito, diferença ou relação significativa.
- Erro tipo I: Rejeitar H0 quando H0 é verdadeira (falso positivo)
- Erro tipo II: Não rejeitar H0 quando H1 é verdadeira (falso negativo)

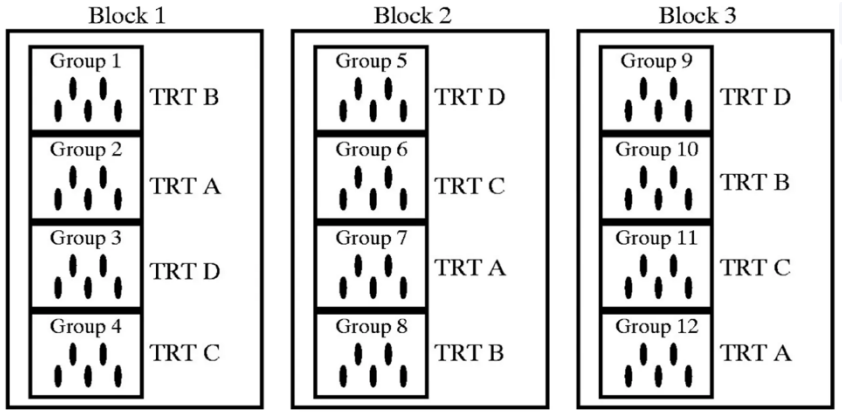
DELINEAMENTO EXPERIMENTAL

- **Inteiramente Casualizados (DIC):** Os tratamentos são atribuídos aleatoriamente às unidades experimentais sem qualquer restrição.
- **Blocos Casualizados (BDC)**: As unidades experimentais são agrupadas em blocos homogêneos, e dentro de cada bloco, os tratamentos são atribuídos aleatoriamente.

DIC



DBC



EXPERIMENTO UNIFATORIAL

- Estuda somente um fator (variável independente) que é manipulado ou investigado para avaliar seu efeito sobre uma variável dependente (resposta).
- Deseja-se entender como diferentes níveis, categorias ou tratamentos de um único fator influenciam o resultado observado.

Biblioteca	Condition
treated1fb	treated
treated2fb	treated
treated3fb	treated
untreated1fb	untreated
untreated2fb	untreated
untreated3fb	untreated
untreated4fb	untreated

1 - Desbalanceado
2 - Um único fator “Condition”

DESeq2 manual: https://bioconductor.org/packages/release/bioc/vignettes/DESeq2/inst/doc/DESeq2.html

EXPERIMENTO UNIFATORIAL / ANOVA

Fonte de Variação	Graus de Liberdade (GL)	Soma de Quadrados (SQ)	Quadrado Médio (QM)	Estatística F
Entre Grupos	$k-1$	SQ_{Entre}	$QM_{Entre} = \frac{SQ_{Entre}}{k-1}$	$F = \frac{QM_{Entre}}{QM_{Dentro}}$
Dentro dos Grupos	$N-k$	SQ_{Dentro}	$QM_{Dentro} = \frac{SQ_{Dentro}}{N-k}$	
Total	$N-1$	SQ_{Total}		

Onde:

- k: Número de grupos ou tratamentos.
- N: Número total de observações (soma de todas as observações em todos os grupos).
- SQ: Soma de Quadrados associada a cada fonte de variação.
- QM: Quadrado Médio, obtido dividindo a Soma de Quadrados pelos respectivos Graus de Liberdade.
- F: Estatística F usada para testar a significância dos efeitos.

Statistical Bioinformatics with R, 2010

EXPERIMENTO MULTIFATORIAL

- Estuda o efeito de dois ou mais fatores (variáveis independentes) simultaneamente sobre a resposta (variável dependente).
- Todos os possíveis níveis dos fatores são combinados, permitindo a análise não apenas dos efeitos individuais (principais) de cada fator, mas também das possíveis **interações** entre eles.

Sample	Condition	Type	SizeFactor
treated1	treated	single-read	1.629707
treated2	treated	paired-end	0.761162
treated3	treated	paired-end	0.830312
untreated1	untreated	single-read	1.143904
untreated2	untreated	single-read	1.791281
untreated3	untreated	paired-end	0.645994
untreated4	untreated	paired-end	0.750728

- Desbalanceado
- Dois fatores: “Condition” e “type”
- Tratamentos combinação

DESeq2 manual: https://bioconductor.org/packages/release/bioc/vignettes/DESeq2/inst/doc/DESeq2.html

DESENHO MULTIFATORIAL / ANOVA

Para um experimento fatorial com dois fatores (por exemplo, Fator A e Fator B), o quadro da ANOVA é estruturado da seguinte forma:

Fonte de Variação	Graus de Liberdade (GL)	Soma de Quadrados (SQ)	Quadrado Médio (QM)	Estatística F
Fator A	$a-1$	SQ_A	$QM_A = \frac{SQ_A}{a-1}$	$F_A = \frac{QM_A}{QM_E}$
Fator B	$b-1$	SQ_B	$QM_B = \frac{SQ_B}{b-1}$	$F_B = \frac{QM_B}{QM_E}$
Interação A x B	$(a-1)(b-1)$	SQ_{AB}	$QM_{AB} = \frac{SQ_{AB}}{(a-1)(b-1)}$	$F_{AB} = \frac{QM_{AB}}{QM_E}$
Erro Experimental	$n-ab$	SQ_E	$QM_E = \frac{SQ_E}{n-ab}$	
Total	$n-1$	SQ_T		

Onde:

- a: Número de níveis do Fator A.
- b: Número de níveis do Fator B.
- n: Número total de observações (repetições x tratamentos).
- ab: Número total de tratamentos (combinações dos níveis dos fatores).
- SQ: Soma de Quadrados associada a cada fonte de variação.
- QM: Quadrado Médio, obtido dividindo a Soma de Quadrados pelos respectivos Graus de Liberdade.
- F: Estatística F usada para testar a significância dos efeitos.

Statistical Bioinformatics with R, 2010

UNIFATORIAL VS MULTIFATORIAL

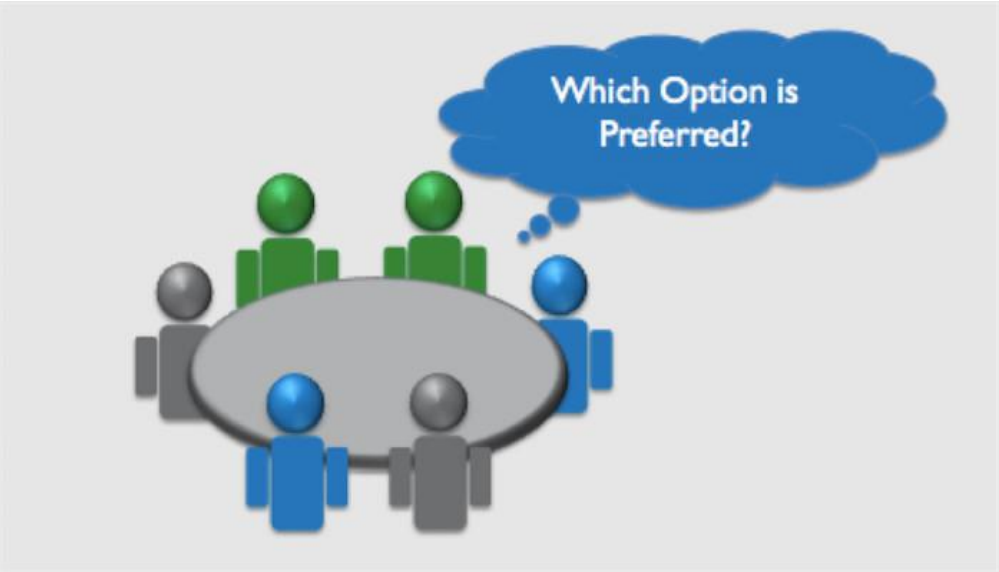
Unifatorial

Biblioteca	Fator1
Lib1	Trat1
Lib2	Trat1
Lib3	Trat1
Lib4	Trat2
Lib5	Trat2
Lib6	Trat2

R fórmula: ~ Fator1

Bifatorial

Biblioteca	Infecção	Celula
Lib1	Mock	Astrocito
Lib2	Mock	Astrocito
Lib3	Mock	Astrocito
Lib4	Zika	Astrocito
Lib5	Zika	Astrocito
Lib6	Zika	Astrocito
Lib7	Mock	Oligodendrocito
Lib8	Mock	Oligodendrocito
Lib9	Mock	Oligodendrocito
Lib10	Zika	Oligodendrocito
Lib11	Zika	Oligodendrocito
Lib12	Zika	Oligodendrocito



R fórmulas:
~ Infeccao + Celula
~ infeccao + Celula + Infeccao:Celula

MODEL MATRIX / INTERCEPT

- O **intercepto** (ou **intercept**) em um modelo de regressão representa o valor esperado da variável dependente (ou resposta) quando todas as variáveis independentes (ou preditoras) são iguais a zero.
- Por padrão as formulas do R usam o intercepto.

Três tratamentos somente um fator (código R):

```
des <- data.frame (virus=c("mock","mock","mock","dengue","dengue","dengue","zika","zika","zika"))
des$virus <- as.factor (des$virus)
des$virus <- relevel (des$virus, ref="mock")
model.matrix (~virus, data=des) # com interceptor
model.matrix (~ -1 + virus, data=des) # sem interceptor
```

	(Intercept)	virusdengue	viruszika
1	1	0	0
2	1	0	0
3	1	0	0
4	1	1	0
5	1	1	0
6	1	1	0
7	1	0	1
8	1	0	1
9	1	0	1

	virusmock	virusdengue	viruszika
1	1	0	0
2	1	0	0
3	1	0	0
4	0	1	0
5	0	1	0
6	0	1	0
7	0	0	1
8	0	0	1
9	0	0	1

MODEL MATRIX / MULTIFATORIAL

- Codigo do R (sem interceptor):
  - *df <- data.frame(biblioteca = c("lib1", "lib2", "lib3", "lib4", "lib5", "lib6", "lib7", "lib8", "lib9", "lib10", "lib11", "lib12", "lib13", "lib14", "lib15", "lib16", "lib17", "lib18"), virus = c("mock", "mock", "mock", "mock", "mock", "mock", "dengue", "dengue", "dengue", "dengue", "dengue", "dengue", "zika", "zika", "zika", "zika", "zika", "zika"), celula = c("macrofago", "macrofago", "macrofago", "neutrofilo", "neutrofilo", "neutrofilo", "macrofago", "macrofago", "macrofago", "neutrofilo", "neutrofilo", "neutrofilo", "macrofago", "macrofago", "macrofago", "neutrofilo", "neutrofilo", "neutrofilo"))*
  - *model.matrix (~0 + virus + celula, data=df, contrasts.arg=list(virus=contrasts(df$virus, contrasts=F),celula=contrasts(df$celula, contrasts=F)))*

	virusmock	virusdengue	viruszika	celulamacrofago	celulaneutrofilo
1	1	0	0	1	0
2	1	0	0	1	0
3	1	0	0	1	0
4	1	0	0	0	1
5	1	0	0	0	1
6	1	0	0	0	1
7	0	1	0	1	0
8	0	1	0	1	0
9	0	1	0	1	0
10	0	1	0	0	1
11	0	1	0	0	1
12	0	1	0	0	1
13	0	0	1	1	0
14	0	0	1	1	0
15	0	0	1	1	0
16	0	0	1	0	1
17	0	0	1	0	1
18	0	0	1	0	1

MODELO LINEAR GENERALIZADO MISTO

1. **Componente Aleatória:**
   - A variável resposta Y_{ij} segue uma distribuição da família exponencial (e.g., binomial, Poisson).
   - A média da distribuição é μ_{ij}.
2. **Componente Sistemática:**
   - O preditor linear é $\eta_{ij} = \mathbf{x}_{ij}^\top \boldsymbol{\beta} + \mathbf{z}_{ij}^\top \mathbf{u}_j$.
   - $\mathbf{x}_{ij}$**:** Vetor de covariáveis para os efeitos fixos.
   - $\boldsymbol{\beta}$**:** Vetor de coeficientes dos efeitos fixos.
   - $\mathbf{z}_{ij}$**:** Vetor de covariáveis para os efeitos aleatórios.
   - $\mathbf{u}_j$**:** Vetor de efeitos aleatórios associados ao nível j.
3. **Função de Ligação:**
   - Relaciona a média μ_{ij} ao preditor linear η_{ij}.
   - $g(\mu_{ij}) = \eta_{ij}$, onde g é a função de ligação apropriada para a distribuição escolhida.

Assunções:

- Os efeitos aleatórios $\mathbf{u}_j$ são normalmente distribuídos com média zero e matriz de covariância $\mathbf{G}$.
- Os efeitos aleatórios e os erros são independentes.

MODELO LINEAR MISTO

O modelo linear misto pode ser representado pela seguinte equação:

$$\mathbf{Y} = \mathbf{X}\boldsymbol{\beta} + \mathbf{Z}\mathbf{u} + \boldsymbol{\epsilon}$$

Onde:

- $\mathbf{Y}$: Vetor de respostas observadas.
- $\mathbf{X}$: Matriz de delineamento dos efeitos fixos.
- $\boldsymbol{\beta}$: Vetor de parâmetros desconhecidos dos efeitos fixos.
- $\mathbf{Z}$: Matriz de delineamento dos efeitos aleatórios.
- $\mathbf{u}$: Vetor de efeitos aleatórios ($\mathbf{u} \sim N(\mathbf{0}, \mathbf{G})$).
- $\boldsymbol{\epsilon}$: Vetor de erros aleatórios ($\boldsymbol{\epsilon} \sim N(\mathbf{0}, \mathbf{R})$).

Assunções:

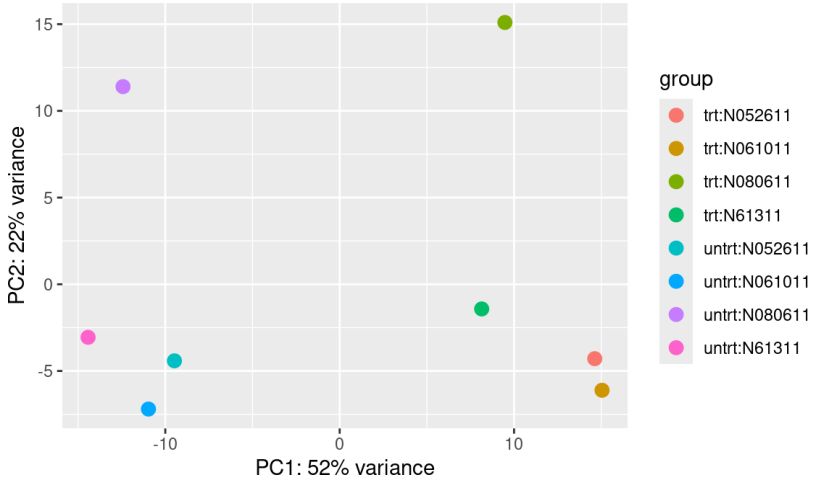
- Os efeitos aleatórios ($\mathbf{u}$) e os erros ($\boldsymbol{\epsilon}$) são normalmente distribuídos com médias zero.
- Os efeitos aleatórios e os erros são independentes entre si.
- As matrizes de covariância $\mathbf{G}$ e $\mathbf{R}$ especificam a estrutura de variância e covariância dos efeitos aleatórios e dos erros, respectivamente.

EXPERIMENTOS EM RNA-SEQ

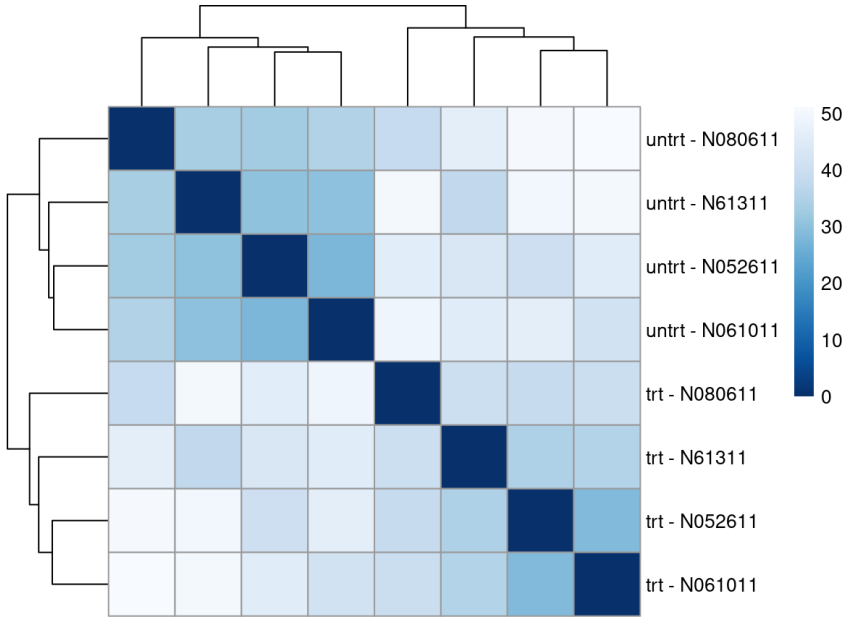
- Bulk RNA-seq (sequenciamento de RNA em massa): técnica de sequenciamento de alto rendimento (NGS) usada para medir a expressão gênica de uma amostra biológica.
- Single-Cell RNA-seq: a técnica de sequenciamento de RNA que permite a análise da expressão gênica em células individuais, em vez de uma população mista de células.
- Sequenciamento ainda caro (mas barateando)
- “Poucos” números de réplicas, em geral 2 ou 3 por tratamento
- Desenho complexos ainda são limitados
- Controle experimental pode ser um desafio em amostras “naturais”.

AVALIAÇÃO DO EXPERIMENTO

Componentes Principais



Mapa de calor e agrupamento com as distancias (euclideana)



NORMALIZAÇÃO

- FPKM/RPKM (Fragments/Reads per Kilobase Milion)
- TMM: presupoe que a maioria dos genes não é diferencialmente expresso
- TMMsp: modificação para alta proporção de zeros
- DESeq2/RLE: expressão relativa por meio da média geométrica por gene
- Upperquartile: quanti 75% das contagem após remoção dos zeros

COEFICIENTE DE VARIAÇÃO BIOLÓGICA (BCV)

- Gene counts should vary according to a Poisson law.
- É o coeficiente de variação com o qual a (desconhecida) verdadeira abundância do gene varia entre as réplicas (amostras)
- Total CV^2 = Tecnica CV^2 + Biológica CV^2
- Variância Biológica >> (muito maior) que a Variância Técnica
- BCV = raiz quadrada (dispersão)
  - Nesse modelos: Variância muito maior do que a média
  - Y ~ NB (média, dispersão)
  - Var(Y) = média + média^2 * dispersão

edgeR manual: https://bioconductor.org/packages/release/bioc/vignettes/edgeR/inst/doc/edgeRUsersGuide.pdf

TESTES ESTATÍSTICOS

- Modelos lineares generalizados (GLM) da família binomial negativa e medelos lineares com transformações
- Testes estatísticos de hipótese por gene segundo o desenho experimental
    - Teste de Wald
    - Teste da razão de verossimilhança
- Estimativa de abundância: baseMean e logCPM
- Comparação das abundâncias em um contraste: log2FC (log2 (fold change)
- Correção do p-valor para testes múltiplos
    - False discover heat (FDR), Bonferroni, Benjamini & Hochberg (BH)
- Escolha da significância estatística (default: 0,5)
- Cutoff para log2foldChange > |1| (?)

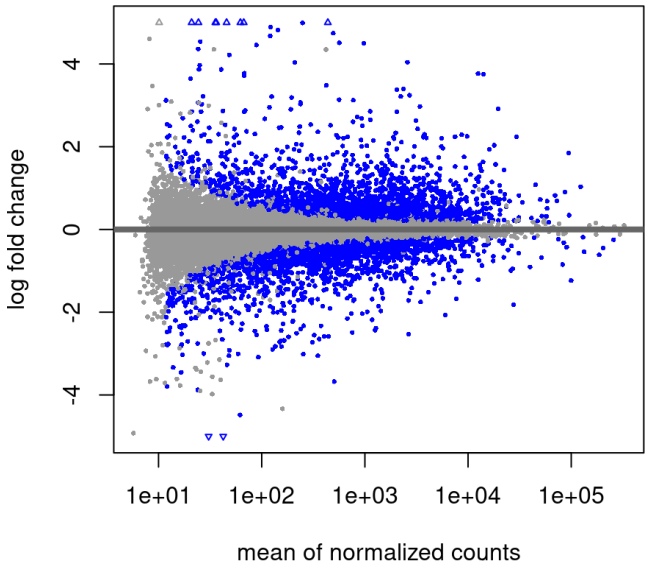
	id character varying (20)	gene_name character varying (40)	basemean double precision	log2fc double precision	padj double precision	gene_type character varying (40)	hgnc_symbol character varying (25)	gene_description text
1	ENSG00000000003	TSPAN6	17.1572017057547	1.97896424765242	0.00190074710325393	protein_coding	TSPAN6	tetraspanin 6 [Source:HGNC Symbol;Acc:HGNC:11858]
2	ENSG00000000419	DPM1	71.5432842100546	-1.02574199140617	3.17731547147863e-06	protein_coding	DPM1	dolichyl-phosphate mannosyltransferase subunit 1, catalytic [Source:HGNC Symbol;Acc:HGNC:3005]
3	ENSG00000000457	SCYL3	27.0983149436443	-0.264888558286573	0.451662536799072	protein_coding	SCYL3	SCY1 like pseudokinase 3 [Source:HGNC Symbol;Acc:HGNC:19285]
4	ENSG00000000460	C1orf112	12.2293586501917	-0.708936120454276	0.133937912592909	protein_coding	C1orf112	chromosome 1 open reading frame 112 [Source:HGNC Symbol;Acc:HGNC:25565]
5	ENSG00000000938	FGR	22.0909642012561	-7.85590275667639	6.87666327967071e-14	protein_coding	FGR	FGR proto-oncogene, Src family tyrosine kinase [Source:HGNC Symbol;Acc:HGNC:3697]
6	ENSG00000000971	CFH	1216.64851481659	0.838680546021523	4.66912853106032e-24	protein_coding	CFH	complement factor H [Source:HGNC Symbol;Acc:HGNC:4883]
7	ENSG00000001036	FUCA2	91.8754675549642	1.84718899620432	1.42579739530742e-11	protein_coding	FUCA2	alpha-L-fucosidase 2 [Source:HGNC Symbol;Acc:HGNC:4008]
8	ENSG00000001084	GCLC	80.1967841270249	0.296192243911436	0.191567362767671	protein_coding	GCLC	glutamate-cysteine ligase catalytic subunit [Source:HGNC Symbol;Acc:HGNC:4311]
9	ENSG00000001167	NFYA	74.261764668447	-0.461385950072636	0.0282275854158563	protein_coding	NFYA	nuclear transcription factor Y subunit alpha [Source:HGNC Symbol;Acc:HGNC:7804]
10	ENSG00000001460	STPG1	23.0984477609741	0.599383595126165	0.158016985693335	protein_coding	STPG1	sperm tail PG-rich repeat containing 1 [Source:HGNC Symbol;Acc:HGNC:28070]
11	ENSG00000001461	NIPAL3	256.914448146056	3.01039215585104	3.33454809481756e-42	protein_coding	NIPAL3	NIPA like domain containing 3 [Source:HGNC Symbol;Acc:HGNC:25233]
12	ENSG00000001497	LAS1L	72.7949852152277	-0.972758750530218	3.52838802866255e-06	protein_coding	LAS1L	LAS1 like, ribosome biogenesis factor [Source:HGNC Symbol;Acc:HGNC:25726]
13	ENSG00000001561	ENPP4	3.21539178178699	1.14005731841628	0.145624474340129	protein_coding	ENPP4	ectonucleotide pyrophosphatase/phosphodiesterase 4 [Source:HGNC Symbol;Acc:HGNC:3359]
14	ENSG00000001617	SEMA3F	46.7950262448008	2.82893696691571	2.71746921457423e-08	protein_coding	SEMA3F	semaphorin 3F [Source:HGNC Symbol;Acc:HGNC:10728]
15	ENSG00000001629	ANKIB1	257.271732104525	1.38791140680547	9.55477978391673e-20	protein_coding	ANKIB1	ankyrin repeat and IBR domain containing 1 [Source:HGNC Symbol;Acc:HGNC:22215]
16	ENSG00000001630	CYP51A1	16.4372520382027	4.03771570887447	0.00029290588306311	protein_coding	CYP51A1	cytochrome P450 family 51 subfamily A member 1 [Source:HGNC Symbol;Acc:HGNC:2649]
17	ENSG00000001631	KRIT1	97.3982875185353	0.0959806152048387	0.669279950761429	protein_coding	KRIT1	KRIT1, ankyrin repeat containing [Source:HGNC Symbol;Acc:HGNC:1573]
18	ENSG00000002016	RAD52	38.3122358431257	-0.899719211844179	0.00182435342905084	protein_coding	RAD52	RAD52 homolog, DNA repair protein [Source:HGNC Symbol;Acc:HGNC:9824]
19	ENSG00000002079	MYH16	22.5042221636887	-4.5748756056223	2.5501170997594e-21	transcribed_unitary_pseudogene	MYH16	myosin heavy chain 16 pseudogene [Source:HGNC Symbol;Acc:HGNC:31038]
20	ENSG00000002330	BAD	46.5295357031922	1.39428791572073	4.71538609998895e-05	protein_coding	BAD	BCL2 associated agonist of cell death [Source:HGNC Symbol;Acc:HGNC:936]
21	ENSG00000002549	LAP3	179.551576557804	-0.406616717654376	0.00461411163770919	protein_coding	LAP3	leucine aminopeptidase 3 [Source:HGNC Symbol;Acc:HGNC:18449]
22	ENSG00000002586	CD99	323.832285696435	2.26447279141293	2.12335966808522e-44	protein_coding	CD99	CD99 molecule (Xg blood group) [Source:HGNC Symbol;Acc:HGNC:7082]
23	ENSG00000002746	HECW1	5.31267621709469	-2.2834374080793	0.00182032410361163	protein_coding	HECW1	HECT, C2 and WW domain containing E3 ubiquitin protein ligase 1 [Source:HGNC Symbol;Acc:HGNC:221
24	ENSG00000002822	MAD1L1	57.5168128055097	0.102602880563044	0.724946739965711	protein_coding	MAD1L1	mitotic arrest deficient 1 like 1 [Source:HGNC Symbol;Acc:HGNC:6762]
25	ENSG00000002834	LASP1	1433.30114850683	1.23392962502247	3.20885551518879e-64	protein_coding	LASP1	LIM and SH3 protein 1 [Source:HGNC Symbol;Acc:HGNC:6513]
26	ENSG00000002919	SNX11	46.3290002380304	0.418577700578125	0.160204816630331	protein_coding	SNX11	sorting nexin 11 [Source:HGNC Symbol;Acc:HGNC:14975]

TESTES ESTATÍSTICOS - PROGRAMAS

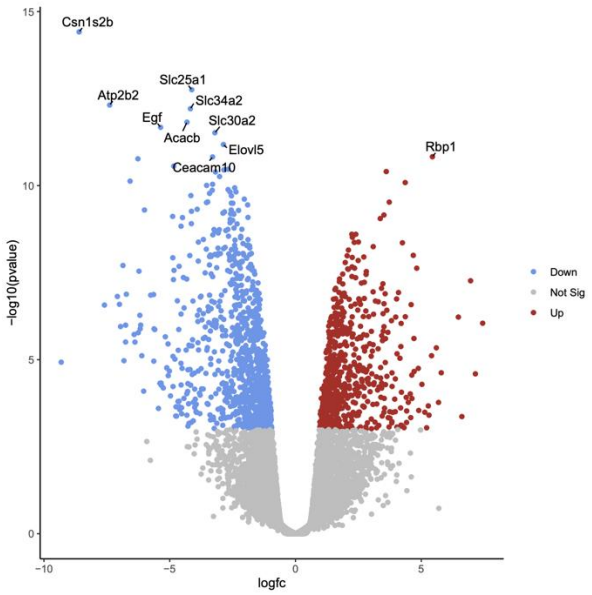
- DESeq2: contrastes, interações simples, séries temporais e correção do erro sistemático (Love et al., 2014)
- edgeR: diversos normalizadores e testes estatisticos (Robinson et al.,2010)
  - Quasi-vossimilhança
  - Teste exato da binomial negativa
  - GLM / teste de Wald
- baySeq: método bayesiano empirico que estima a verosimilhança a posteriori (Hardcastle, 2024)
- limma-voom: flexivel em relação aos modelos, trabalha bem com dados pareados (Law et al., 2014)
- DEGRE: permite inserir efeito aleatório em contrastes pairwise (Machado et al.,2023)
- Cufflinks 2 : calcula (Mortazavi et al., 2008)

GRAFICO DAS EXPRESSÕES

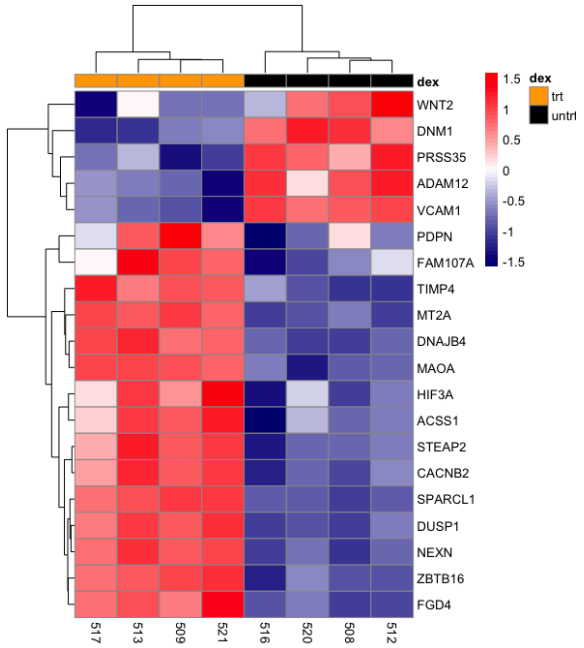
MAplot



Volcano Plot



Heatmap



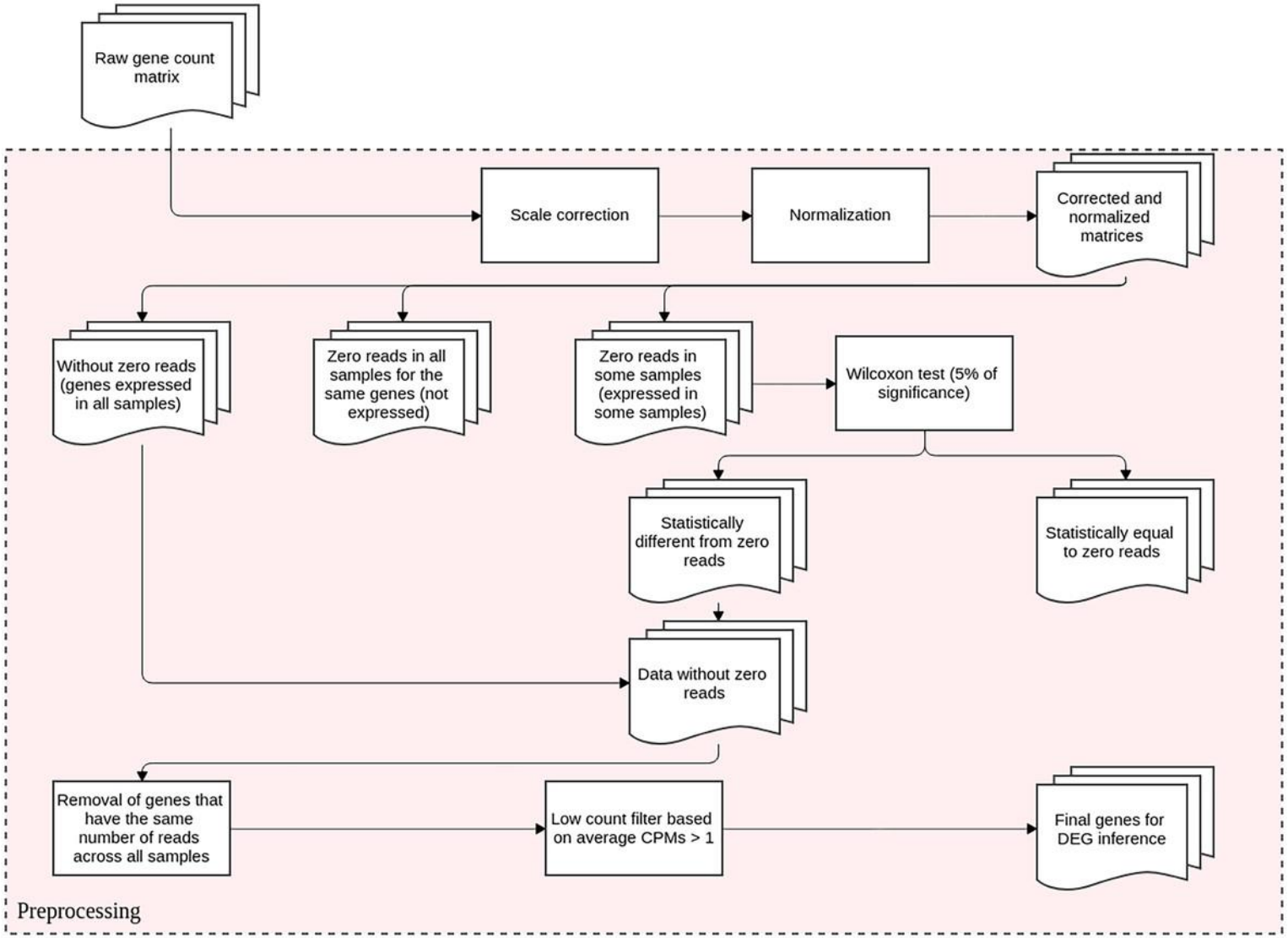
DEGRE

Inferring Differentially Expressed Genes using Generalized Linear Mixed Models

- Aplica “Generalized Linear Mixed Model” com a distribuição binomial negativa
- Teste de Wald para os coeficientes da regressão
- Detecta genes diferencialmente expressos em duas ou mais condições condições
- Alta variabilidade biológica
- Somente os efeitos fixos não são suficientes
- Realiza o pre-processamento de remoção de zeros
- Gera p-valores ajustados para comparaçãp multipla

Machado *et al*, 2022

DEGRE


Raw gene count matrix
Scale correction
Normalization
Corrected and normalized matrices
Without zero reads (genes expressed in all samples)
Zero reads in all samples for the same genes (not expressed)
Zero reads in some samples (expressed in some samples)
Wilcoxon test (5% of significance)
Statistically different from zero reads
Statistically equal to zero reads
Data without zero reads
Removal of genes that have the same number of reads across all samples
Low count filter based on average CPMs > 1
Final genes for DEG inference
Preprocessing

DEGRE: EXEMPLO

- Código R:

```
dir <- system.file("extdata", package = "DEGRE")

tab <- read.csv(file.path(dir,"count_matrix_for_example.csv"))

des <- read.csv(file.path(dir,"design_matrix_for_example.csv"))

results <- DEGRE(count_matrix = tab, p_value_adjustment = "BH", design_matrix
= des, formula = "condition + (1 |sex)")
```

ID	log2FC	P-value	Q-value	averagelogCPM
ENSMUSG00000002104	-3.762	1.749e-10	2.624e-10	8.158
ENSMUSG00000002205	-3.283	3.155e-58	1.056e-57	11.912
ENSMUSG00000002968	-0.635	1.793e-15	3.183e-15	11.860
ENSMUSG00000006335	1.421	3.126e-15	5.333e-15	11.344
ENSMUSG00000018470	-3.294	1.251e-08	1.728e-08	3.283
ENSMUSG00000021254	0.917	8.595e-08	1.150e-07	12.392

DISCUSSÃO

- Como escolher o meu desenho experimental?
- Como saber se o método estatístico é adequado para os meus dados?
- Como saber se tenho “poucos” ou “muitos” genes diferencialmente expressos em um experimento?
- Estamos abusando muito dos testes de hipóteses e p-valores?
- Como saber se tenho relações espúrias nos meus resultados?
- Devo “jogar tudo” para IA?

OBRIGADO!

Github: https://github.com/labinfo-lncc-br
Site: https://www.labinfo.lncc.br
Hugginf Face: https://huggingface.co/Labinfo