

# **Anotação gênica: princípios gerais e comparação entre metodologias**

Desirrê Petters-Vandresen

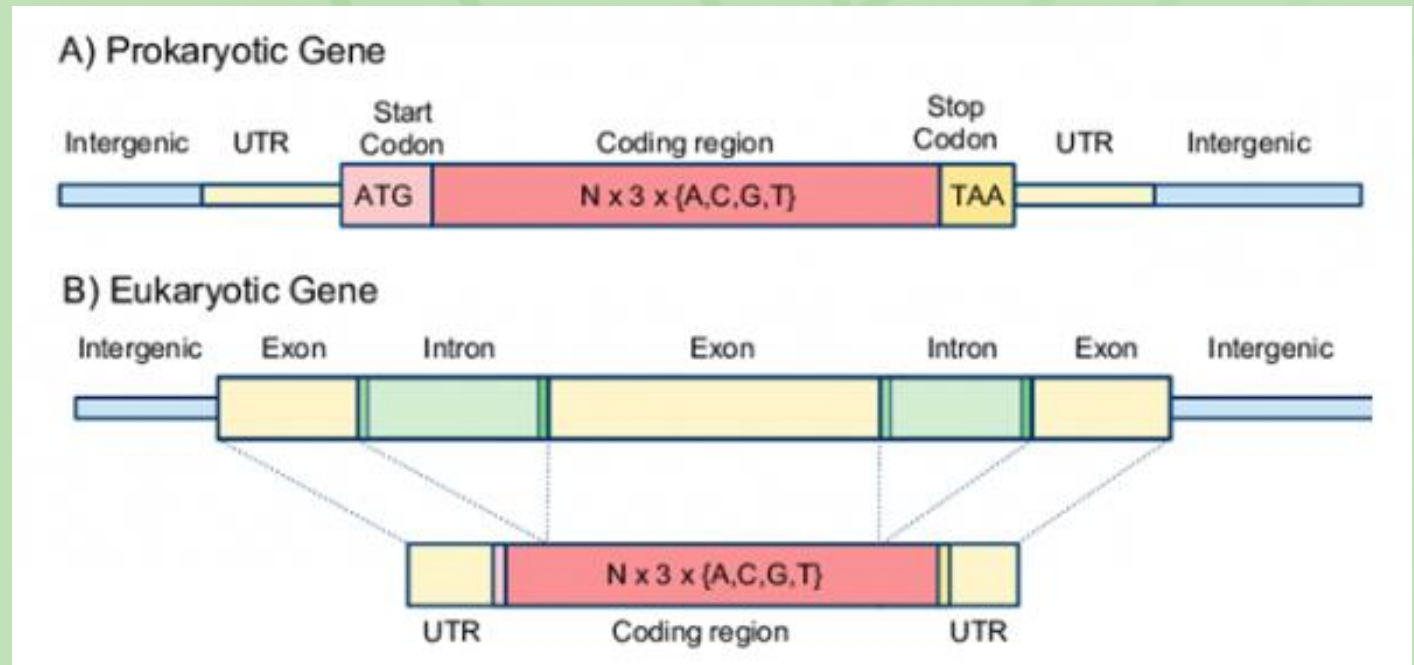
Módulo I – Genômica no Estudo de Microrganismos

# Por que anotar um genoma?

- Sequência do genoma sem anotação: baixa utilidade e aplicabilidade em abordagens funcionais
- Busca por padrões e características que identifiquem genes e sequências relevantes dentro de uma montagem
- Duas categorias principais:
  - *Ab initio*
  - Baseada em homologia

# ***Ab initio* – Visão geral**

- Diretamente baseada na sequência genômica analisada
- Modelos estatísticos treinados para encontrar características presentes em genes:
  - Códon de início e parada
  - Éxons e íntrons
  - Sítios de splicing
  - Sequências intergênicas
  - Sequências transcritas mas não traduzidas



# Ab initio – Vantagem e Desvantagem

- **Vantagem**

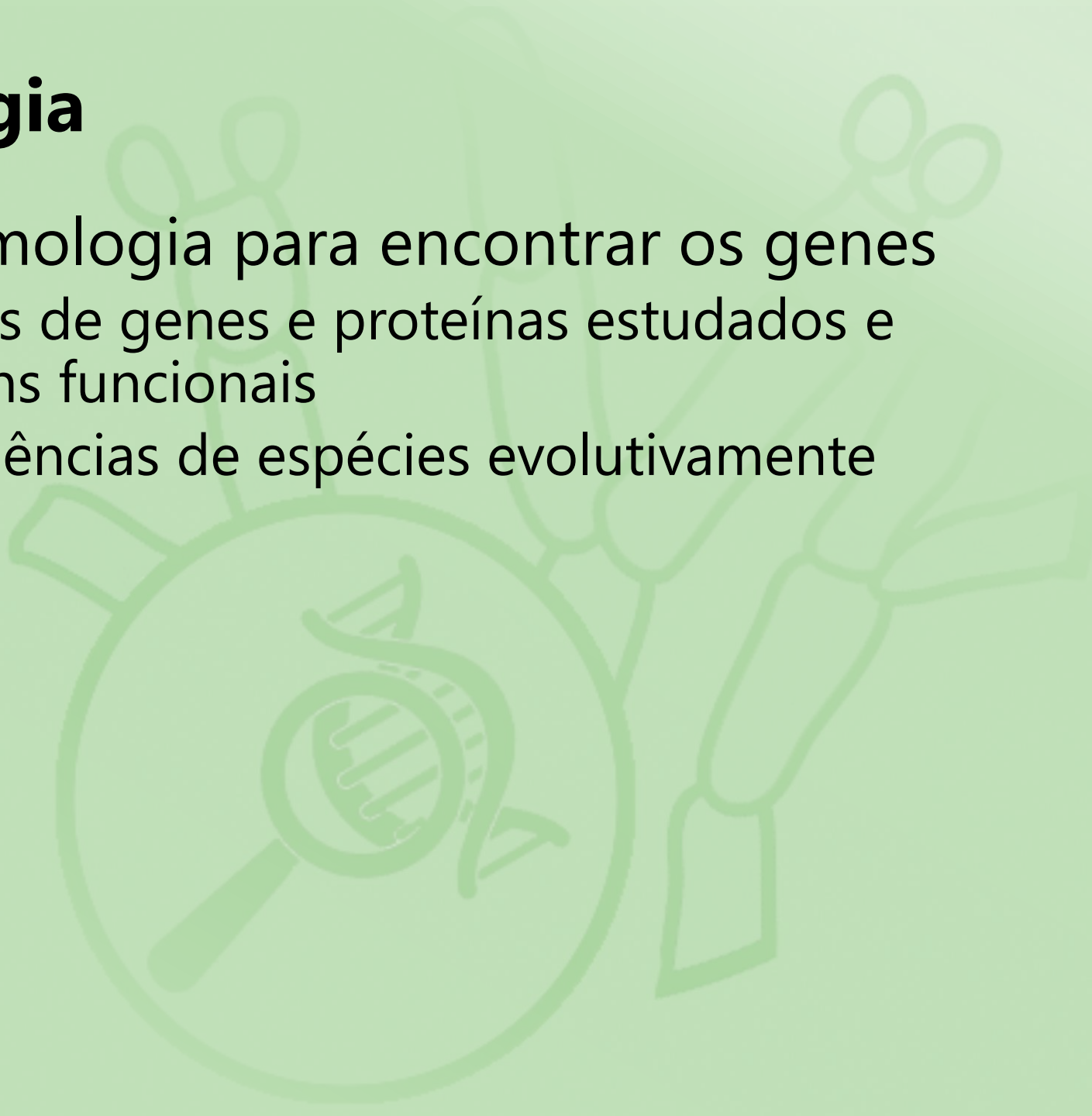
- Pode ser utilizada mesmo quando só sequência do genoma a ser analisado, sem informações adicionais como transcriptoma ou proteoma
- Alguns métodos mais avançados realizam um processo iterativo de anotação e auto-treinamento

- **Desvantagem**

- Exigem um bom “treinamento” a partir de conjuntos de genes previamente anotados com alta qualidade para melhorar a precisão
- Melhores resultados com conjuntos de treinamento que apresentem tamanhos de éxons e íntrons similares e apresentem o mesmo tipo de sítios de splicing
- Os melhores conjuntos de treinamento normalmente são de espécies evolutivamente próximas, e construídos a partir de dados de homologia de sequências

# Baseada em homologia

- Métodos que usam homologia para encontrar os genes
  - Evidências experimentais de genes e proteínas estudados e validados em abordagens funcionais
  - Alinhamentos com sequências de espécies evolutivamente próximas
- Abordagens principais
  - Baseada em RNA
  - Baseada em proteínas



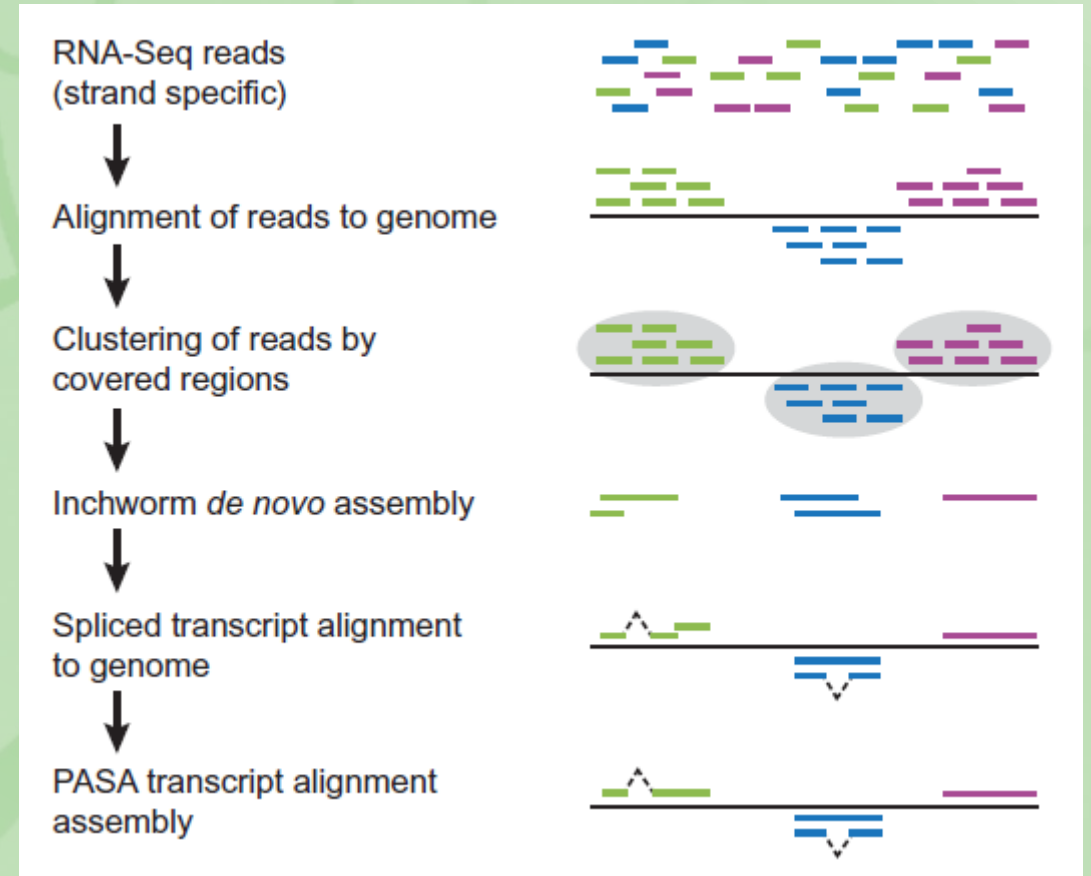


# Baseada em homologia - RNA

- Transcritos provenientes do mesmo organismo do qual a sequência do genoma foi obtido são um tipo de evidência muito preciso:
  - Altamente idênticos ao genoma
  - Determinação precisa dos limites entre éxons-íntrons
  - Determinação de sítio de poli-adenilação
  - Determinação das regiões UTRs
  - Determinação de transcritos alternativos
- Sequências de transcritos podem ser provenientes de:
  - Expressed sequence tags (ESTs)
  - Sequências de cDNA de transcritos completos
  - Sequências de RNA-Seq, no formato de reads ou transcritos montados (transcriptoma)

# Baseada em homologia – RNA-Seq

- Abordagens gerais
  - **Baseada em alinhamento:**  
alinhar os reads contra o genoma e montar os transcritos localmente com base nos alinhamentos
  - **Baseada em transcritos:**  
montar os reads em transcritos, e posteriormente alinhar contra o genoma para determinar as estruturas dos genes
  - Híbrida entre as abordagens anteriores



# Baseada em homologia - Proteínas

- Proteínas são uma fonte de homologia importante para uma anotação, principalmente em termos funcionais
- Particularmente úteis quando não há informação de RNA disponível
- Proteínas conservadas podem auxiliar na anotação de espécies distantes entre si
- Limitação: restrita à proteínas que já tenham sido estudadas e caracterizadas



# Abordagem híbrida: consenso entre *ab initio* e homologia

- Pipelines para execução de softwares de anotação *ab initio* e homologia
- Softwares desenvolvidos para realizar consenso entre todos os resultados, atribuindo pesos distintos para cada um dos métodos
- Em geral:



# Como avaliar a qualidade de uma anotação?

- Avaliação de conteúdo (BUSCO)
- Comparação com anotações prévias confiáveis
  - A quantidade de genes é similar?
  - Anotações funcionais resultam em informações similares?
- Avaliação manual

# Formato GFF3 (General Feature Format)

- Uma feature por linha, 9 colunas delimitadas por tabulações
- **1 (seqid):** nome do cromossomo, contig ou scaffold em que a feature está localizada
- **2 (source):** nome do programa que gerou a anotação, ou da base de dados em que a anotação foi obtida
- **3 (type):** categoria da feature (ex: gene, CDS, mRNA, exon)
- **4 (start):** posição do início da feature no cromossomo, contig ou scaffold
- **5 (end):** posição do final da feature no cromossomo, contig ou scaffold
- **6 (score):** score de confiabilidade, mas muitos softwares não atribuem nenhum valor (.)
- **7 (strand):** indica se a feature está na fita direta/forward (+) ou reversa/reverse (-)
- **8 (phase):** fase de leitura
- **9 (attributes):** informações adicionais sobre a feature, como nome, ou vínculo com outra feature anterior

```
scaffold_10 EVM gene      2697      4438      .      +      .      ID=scaffold_10.1;Name=scaffold_10.1
scaffold_10 EVM mRNA      2697      4438      .      +      .      ID=scaffold_10.1;Parent=scaffold_10.1;Name=scaffold_10.1
scaffold_10 EVM exon      2697      3078      .      +      .      ID=scaffold_10.1.exon1;Parent=scaffold_10.1
scaffold_10 EVM CDS 2697      3078      .      +      0      ID=cds.scaffold_10.1;Parent=scaffold_10.1
scaffold_10 EVM exon      3145      4023      .      +      .      ID=scaffold_10.1.exon2;Parent=scaffold_10.1
scaffold_10 EVM CDS 3145      4023      .      +      2      ID=cds.scaffold_10.1;Parent=scaffold_10.1
scaffold_10 EVM exon      4068      4438      .      +      .      ID=scaffold_10.1.exon3;Parent=scaffold_10.1
scaffold_10 EVM CDS 4068      4438      .      +      2      ID=cds.scaffold_10.1;Parent=scaffold_10.1
```

# Regiões codificantes (CDS) em formato FASTA

- **Linha 1:** identificador da sequência após o sinal de maior (>)
- **Linha 2:** sequência
- Cada sequência corresponde a um gene

```
1 >scaffold_10.1
2 ATGGACGGACCGCTCTCGACGCTGCTCTTACTCCAACGCCATTTTGCTGGTGCTGGGCCTCGCCGGGC
3 TGGCATACATCAGCTTCCGCGCCGCGTACGGCACCGACGTGGGGCGCATCACGGGCATTCTTGAGCCGGG
4 GCACGCGGTGGCGTTCTACGGACACCTCAACTCCAAGGCGCTCGGCAGCGACCAACCCCACTGCGCTGCAG
5 GAGTATTCGGTGAAGAATGGGTGGCCGTTGGTGCAGGTGCGGTTTGGGCAGCGGCGGGTCGTGGTGCTGA
6 ATACGTTTGCGGCGGCGCAGCATTTTCATCATTCGCAATGGGGGGGCGACAATTGACCGCCCGCTGTTTTG
7 GACATTTTCATAAGTTTGTGAGCAATACGCAGGGCGCAACCATTGGCACGTGCGCGTGGGACGCATCGTGC
8 AAGCGCAAGCGCACGGCAATCGGGGCGTACATGACGCGGCCGGCCATCCAGCGCAATGCGCCACTCATCG
9 ACATCGAGGCGCTGGGGCTGGTCGAAGGCATCTTTAACGCCTCGCTGGACGACAACAACAACCCCAAGTGT
10 CGAAGTGGACCCCCGCTCTTTTTCCAGCGGGCTTCGCTCAACTTTGTGCTCATGCTCTGCTACGCGTCG
11 CGGTTCCCGGACATTGACGACCCGCTGCTGCACGAGATTCTGGCCACGGGCAAGACGGTCAGCACGTTTC
12 GCAGCACCAACAACAACATGGCCGACTACGTGCCGCTGCTGCGGTACCTGCCCAACGCGCGGACGGCGAT
13 GGCCAAGCAGGTGACCAAGAAGCGCGACGTGTGGCTCGAGGCGCTGCTGGAGCGCGTGCGCAAAGCCGTG
14 GCGGCCGGCAAGCCCGTGTCTGTCATTGCATCGTCGCTGCTCAAGGAAAAGGGGTCCGAGAAGCTGACAG
15 AGGCCGAGATTCGCTCCATCAACGTGCGGGCTCGTCTCGGGCGGCAGCGACACGATTGCGACGACGGGGCT
16 CGGCGGGCTTGGGTTCTCGCGTCCAAGGAGGGCCAGGCGATTGAGCAAAAGGCGTACGACGAGATTATG
17 AAGGTCTACGCGACGGCCGAGGAGGCGTGGGAGAATTGCGTGCTCGAGGAGAATGTCGAGTACGTGTCG
18 CGCTCGTGCGCGAGATGCTGCGGTACTACTGCGCGATACAGCTGCTGCCACCGCGCAAGACGTGCAAGCC
19 GTTTGAGTGGCATGGCGCACAAATCCCTGCTGGTGTACGGTATACATGAACGCGCAGGCTATCAATCAC
20 GACAAAACCGCATAACGACAGACGCGCACATTTTCCGACCAGAGCGTTGGCTCGATCCCAGCAGTCCGT
21 ACCAGGTGCGGGCTTCCCTACCACTACTCGTATGGCGCGGGCTCGCGAGCATGCACGGCCGTGGCGCTGTC
22 GAACCGGATTCTCTACTGCTACTTTGTGAGGCTGATTGTTTTCGTTCCGCTTCACGGCCAGCGCAGACGCG
23 CCGCCGACGCTGGATTACATTGGATTCAACGAGAACCCGCAGGCGGCGACGGTCGTCCCAAAGACGTTTC
24 GGGTTAACATTGAGGAGAGGCGGCCGAGGGAGGAGCTGGCCAAGAATTTGAGGCGAGTCGAAAGGCCAC
25 TTCTCACCTCGTCTTTACTTAG
```



# Sequências de aminoácidos em formato FASTA

- **Linha 1:**  
identificador  
da sequência  
após o sinal  
de maior (>)
- **Linha 2:**  
sequência
- Cada  
sequência  
corresponde  
a um gene

```
1 >scaffold_10.1
2 MDGPLSTLLSYSNAILLVLGLAGLAYISFRAAYGTDVGRITGIPEPGHAVAFYGHLNSKALGSDHPTALQ
3 EYSVKNGWPLVQVRFGQRRVVVLTFAAAQHFIIRNGGATIDRPLFWTFHKFVSNTQGATIGTSPWDASC
4 KKRRTAIGAYMTRPAIQRNAPLIDIEALGLVEGIFNASLDDNNNPSVEVDPRLFFQRASLNFVLMICYAS
5 RFPDIDDPLLHEILATGKTVSTFRSTNNNMADYVPLLRYLPNARTAMAKQVTKKRDVWLEALLERVRKAV
6 AAGKPVSCIASSLLKEKGSEKLTEAEIRSINVGLVSGGSDTIATTGLGGLGFLASKEGQAIQQKAYDEIM
7 KVYATAEEAWENCVLEENVEYVVALVREMLRYYCAIQLLPPRKTCKPFEWHGAQIPAGVTVYMNAQAINH
8 DKTAYGPDHIFRPERWLDPSSPYQVGLPYHYSYGAGSRCTAVALSNRILYCYFVRLIVSFRFTASADA
9 PPTLDYIGFNENPQAATVVPKTFRVNIEERRPREELAKNFEASRKATSHLVFT
10 >scaffold_10.2
11 MALQTCRRCRKRRIKCDLQLPACTSCQLVDLECLYFDDSLGHDVPRSYLHALSKKVENLESTINAIKSPA
12 AAAPSPTPFSQSDCPTPLQASLDPRGSSASSLGLGTSAGLLENLLKTLVQRSSTQDQSALS RFASRTRDV
13 EDDSALAFPPLKVNFSKLDTQSLQQPHLQRALIEYYAKTVQSSFPLLSKAQIDSLLRYEHPLRQCTAAER
14 LPIYGIFALASNLVSRDLDDKQDQSIASMTWTERFHSYIAGFDSSNAHGAVRMKQNILALCFLALLDLVSPL
15 SPKGGVWEVVGAAASRSYVKVLDLVSSSPEIDDEFERLGHCIIYLLESTLSIHFRIPSLYCNSAPTVIPSG
16 LSEPLVYHTLYTLTQLLNFPKDVSVDMESSIPACLRINLESGPSDVSLGQAQVYLT LHPLFTSPGAGIHC
17 CSPDLLSKIALAAAFITHTHKLNKERRVVSIVVTAENVLQAGAAWAAYLMLHSQRDSPLHDYHVPKPID
18 KLPPMEPIVRCSSLLASFAERWKGGRRFCQAWEAFTELL LADDSLSKMATAPQA
```



# **Anotação de elementos transponíveis: princípios gerais e comparação entre metodologias**

Desirrê Petters-Vandresen

Módulo I – Genômica no Estudo de Microrganismos

# Por que anotar elementos transponíveis (TEs)?

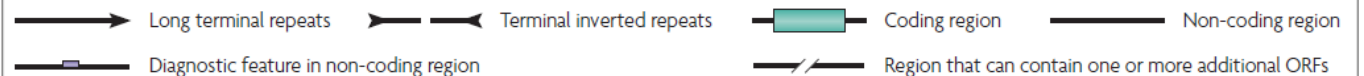
- Porção significativa de muitos genomas eucarióticos e associados com aumento no tamanho de um genoma sem aumento do conteúdo gênico
- Fontes de variabilidade genética e frequentemente associados à genes de patogenicidade, conflitos e interação com hospedeiros

# Classificação de TEs (Classe I)

- Mecanismo de “**cópia e cola**” utilizando transcriptase reversa, com **RNA** como molécula intermediária da transposição
- Autônomos:** transcriptase reversa funcional
- Não-autônomos:** dependem de transcriptases reversas externas
- Capazes de aumentar o número total de cópias dentro do genoma

Classification		Structure	TSD	Code	Occurrence
Order	Superfamily				
<i>Class I (retrotransposons)</i>					
LTR	<i>Copia</i>	→ GAG AP INT RT RH →	4-6	RLC	P, M, F, O
	<i>Gypsy</i>	→ GAG AP RT RH INT →	4-6	RLG	P, M, F, O
	<i>Bel-Pao</i>	→ GAG AP RT RH INT →	4-6	RLB	M
	<i>Retrovirus</i>	→ GAG AP RT RH INT ENV →	4-6	RLR	M
	<i>ERV</i>	→ GAG AP RT RH INT ENV →	4-6	RLE	M
DIRS	<i>DIRS</i>	→ GAG AP RT RH YR →	0	RYD	P, M, F, O
	<i>Ngaro</i>	→ GAG AP RT RH YR → → →	0	RYN	M, F
	<i>VIPER</i>	→ GAG AP RT RH YR → → →	0	RYV	O
PLE	<i>Penelope</i>	← RT EN →	Variable	RPP	P, M, F, O
LINE	<i>R2</i>	RT EN	Variable	RIR	M
	<i>RTE</i>	APE RT	Variable	RIT	M
	<i>Jockey</i>	ORF1 APE RT	Variable	RIJ	M
	<i>L1</i>	ORF1 APE RT	Variable	RIL	P, M, F, O
	<i>I</i>	ORF1 APE RT RH	Variable	RII	P, M, F
SINE	<i>tRNA</i>		Variable	RST	P, M, F
	<i>7SL</i>		Variable	RSL	P, M, F
	<i>5S</i>		Variable	RSS	M, O

## Structural features



## Protein coding domains

AP, Aspartic proteinase	APE, Apurinic endonuclease	ATP, Packaging ATPase	C-INT, C-integrase	CYP, Cysteine protease	EN, Endonuclease
ENV, Envelope protein	GAG, Capsid protein	HEL, Helicase	INT, Integrase	ORF, Open reading frame of unknown function	
POL B, DNA polymerase B	RH, RNase H	RPA, Replication protein A (found only in plants)	RT, Reverse transcriptase	Y2, YR with YY motif	
Tase, Transposase (* with DDE motif)		YR, Tyrosine recombinase			

## Species groups

P, Plants    M, Metazoans    F, Fungi    O, Others

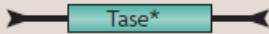
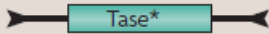
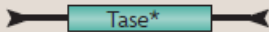
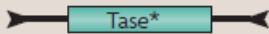
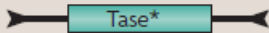
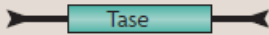
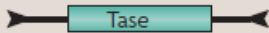


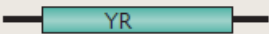


- Mecanismo de "**corta e cola**" utilizando tranposase, com **DNA** como molécula intermediária da transposição

- Autônomos:** tranposase funcional

- Não-autônomos:** dependem de tranposases externas

- Aumentam o número total de cópias dentro do genoma em condições restritas (ex: durante a replicação ou reparo de DNA)

# Classificação de TEs (Classe II)

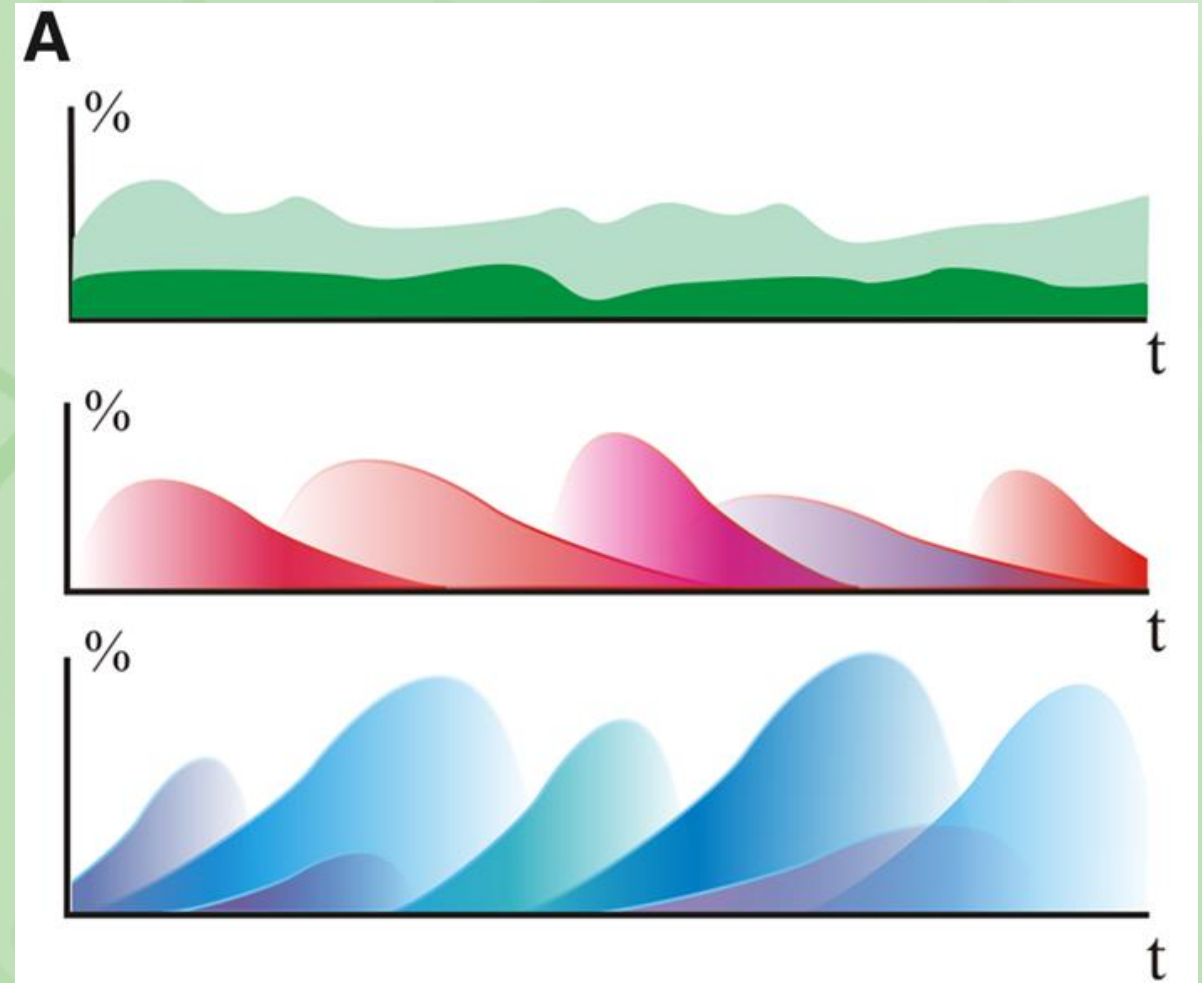
Class II (DNA transposons) - Subclass 1						
TIR	<i>Tc1–Mariner</i>		TA	DTT	P, M, F, O	
	<i>hAT</i>		8	DTA	P, M, F, O	
	<i>Mutator</i>		9–11	DTM	P, M, F, O	
	<i>Merlin</i>		8–9	DTE	M, O	
	<i>Transib</i>		5	DTR	M, F	
	<i>P</i>		8	DTP	P, M	
	<i>PiggyBac</i>		TTAA	DTB	M, O	
	<i>PIF–Harbinger</i>		3	DTH	P, M, F, O	
	<i>CACTA</i>		2–3	DTC	P, M, F	
Crypton	<i>Crypton</i>		0	DYC	F	
Class II (DNA transposons) - Subclass 2						
Helitron	<i>Helitron</i>		0	DHH	P, M, F	
Maverick	<i>Maverick</i>		6	DMM	M, F, O	

Structural features					
	Long terminal repeats		Terminal inverted repeats		Coding region
	Diagnostic feature in non-coding region		Region that can contain one or more additional ORFs		Non-coding region
Protein coding domains					
AP, Aspartic proteinase	APE, Apurinic endonuclease	ATP, Packaging ATPase	C-INT, C-integrase	CYP, Cysteine protease	EN, Endonuclease
ENV, Envelope protein	GAG, Capsid protein	HEL, Helicase	INT, Integrase	ORF, Open reading frame of unknown function	
POL B, DNA polymerase B	RH, RNase H	RPA, Replication protein A (found only in plants)		RT, Reverse transcriptase	
Tase, Transposase (* with DDE motif)		YR, Tyrosine recombinase		Y2, YR with YY motif	
Species groups					
P, Plants	M, Metazoans	F, Fungi	O, Others		

Adaptado de:  
WICKER et al. 2007. **Nature Reviews Genetics**. DOI: [10.1038/nrg2165](https://doi.org/10.1038/nrg2165)

# Dinâmica de TEs dentro de genomas

- **Verde:** TEs “benignos”, neutros ao hospedeiro, com equilíbrio no número de cópias
- **Vermelho:** TEs “agressivos” que invadem o genoma, se expandem e aumentam o número de cópias. O sistema de defesa do hospedeiro não é tão eficiente para reconhecer os elementos invasores e o decaimento no número de cópias é lento
- **Azul:** TEs “dormentes” que possuem picos de expansão dentro do genoma hospedeiro e aumentam o número de cópias. O sistema de defesa é eficiente no reconhecimento das cópias e o decaimento é rápido.





# Reconhecimento e anotação de TEs

- **Desafio**: detectar todos as repetições e TEs, inclusive cópias degeneradas e silenciadas, e que podem não apresentar todas as características de uma cópia ativa
- Estratégias gerais:
  - *De novo*
  - Baseada em homologia
  - Baseada em repetitividade

# De novo – Visão geral

- Diretamente baseada na sequência genômica analisada
- Limitada à elementos que ainda apresentem as características de reconhecimento ao longo da sequência, sem degeneração
- Modelos estatísticos treinados para encontrar características presentes em nos TEs de Classe I e Classe II de diferentes ordens e superfamílias
  - Presença de repetições terminais invertidas
  - Presença de duplicação do sítio alvo
  - Presença de ORFs
  - Presença de sequências associadas a transcriptases reversas, transposases...

Classification		Structure	TSD	Code	Occurrence
Order	Superfamily				
Class I (retrotransposons)					
LTR	<i>Copia</i>	→ GAG AP INT RT RH →	4–6	RLC	P, M, F, O
	<i>Gypsy</i>	→ GAG AP RT RH INT →	4–6	RLG	P, M, F, O
	<i>Bel-Pao</i>	→ GAG AP RT RH INT →	4–6	RLB	M
	<i>Retrovirus</i>	→ GAG AP RT RH INT ENV →	4–6	RLR	M
	<i>ERV</i>	→ GAG AP RT RH INT ENV →	4–6	RLE	M
DIRS	<i>DIRS</i>	→ GAG AP RT RH YR ←	0	RYD	P, M, F, O
	<i>Ngaro</i>	→ GAG AP RT RH YR → → →	0	RYN	M, F
	<i>VIPER</i>	→ GAG AP RT RH YR → → →	0	RYV	O
PLE	<i>Penelope</i>	← RT EN →	Variable	RPP	P, M, F, O

# Baseada em homologia

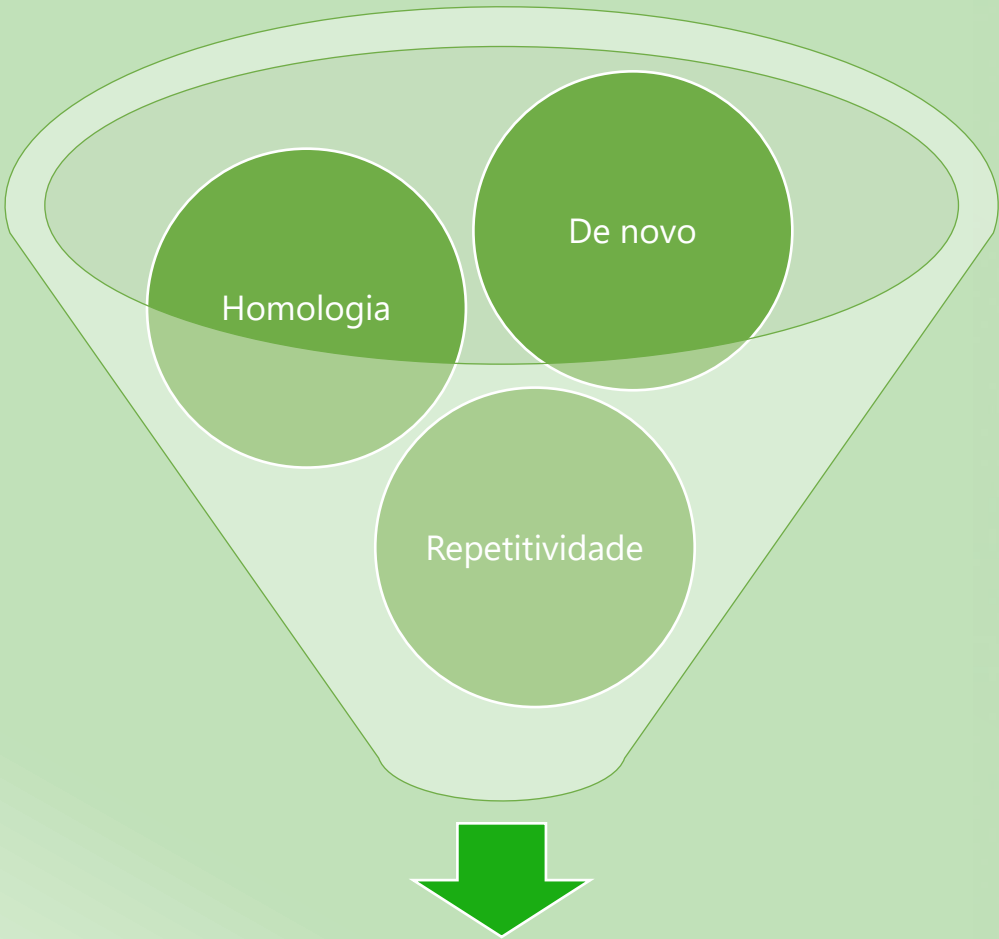
- Métodos que usam homologia para encontrar os TEs
  - Comparação com sequências de TEs previamente caracterizados, disponíveis em bases de dados (ex: Repbase) por meio de alinhamentos
  - Comparação com bibliotecas de elementos de vários grupos de organismos, ou restritas ao grupo do organismo analisado
- **Limitação:** restrita aos TEs já estudados, e TEs muito degenerados podem não ser detectados



# Baseada em repetitividade

- Detecção de sequências ou blocos de sequências repetidas em comparação com o restante do genoma
- Útil para detecção de sequências teloméricas, centroméricas, DNA satélite ou sequências repetitivas que não compõem elementos transponíveis diretamente

# Abordagem híbrida: consenso e re-anotação



- Maior eficiência em comparação ao uso de abordagens isoladas
- Maior sensibilidade em comparação ao uso de bases de dados abrangentes
- Computacionalmente mais exigente: elaboração de pipelines com vários softwares e uso de servidores

Consenso e base inicial  
focada no genoma  
analisado



Nova avaliação baseada em  
homologia



Consenso final de todas as  
rodadas de avaliação,  
classificação e identificação



# Arquivo de saída no formato GFF3

- Estrutura similar ao arquivo de anotação gênica
- Coluna 9 normalmente contém informações sobre o elemento detectado, classificação e identificação
- Exemplos:
  - **Linha 04:** Elemento não categorizado
  - **Linha 05:** LTR-Gypsy (Classe I) incompleto
  - **Linha 06:** LINE (Classe I) incompleto

```
4 scaffold_10 Cap173_annot_REPET_TEs match 396970 397129 0.0 + .
ID=ms387_scaffold_10_noCat_Cap173-B-G1-Map20;Target=noCat_Cap173-B-G1-Map20 1772
1931;TargetLength=2500;TargetDescription=CI:NA struct:(SSR: (TAAA)6_end SSRCoverage:0.22);Identity=65.6
5 scaffold_10 Cap173_annot_REPET_TEs match 387699 388154 0.0 - .
ID=ms388_scaffold_10_RLX-incomp_Cap173-B-G19-Map4;Target=RLX-incomp_Cap173-B-G19-Map4 1
457;TargetLength=5411;TargetDescription=CI:35 coding:(TE_BLRtx: Gypsy-3-I_AF:ClassI:LTR:Gypsy: 5.29% |
Gypsy-34_BG-I:ClassI:LTR:Gypsy: 9.46% TE_BLRx: Gypsy-33_BG-I_1p:ClassI:LTR:Gypsy: 56.79% |
Gypsy-9_BG-I_3p:ClassI:LTR:Gypsy: 6.75% profiles: _RT_pyggy_NA_RT_NA: 99.56%(99.56%) |
_INT_crm_NA_INT_NA: 31.50%(31.50%) | _RNaseH_maggy_NA_RH_NA: 89.26%(89.26%)) struct:(TElength: >4000bps)
other:(SSRCoverage:0.38);Identity=83.4
6 scaffold_10 Cap173_annot_REPET_TEs match 394840 395221 0.0 - .
ID=ms390_scaffold_10_RIX-incomp_Cap173-B-G3-Map20;Target=RIX-incomp_Cap173-B-G3-Map20 1399
1755;TargetLength=2003;TargetDescription=CI:27 coding:(TE_BLRx: Tad1-14_BG_2p:ClassI:LINE:I: 12.36% |
Tad1-31B_BG_2p:ClassI:LINE:I: 8.76%) struct:(TElength: >1000bps) other:(SSRCoverage:0.44);Identity=68.4
```

# **Anotação funcional: efetores, CAZymes, clusters de metabólitos secundários**

Desirrê Petters-Vandresen

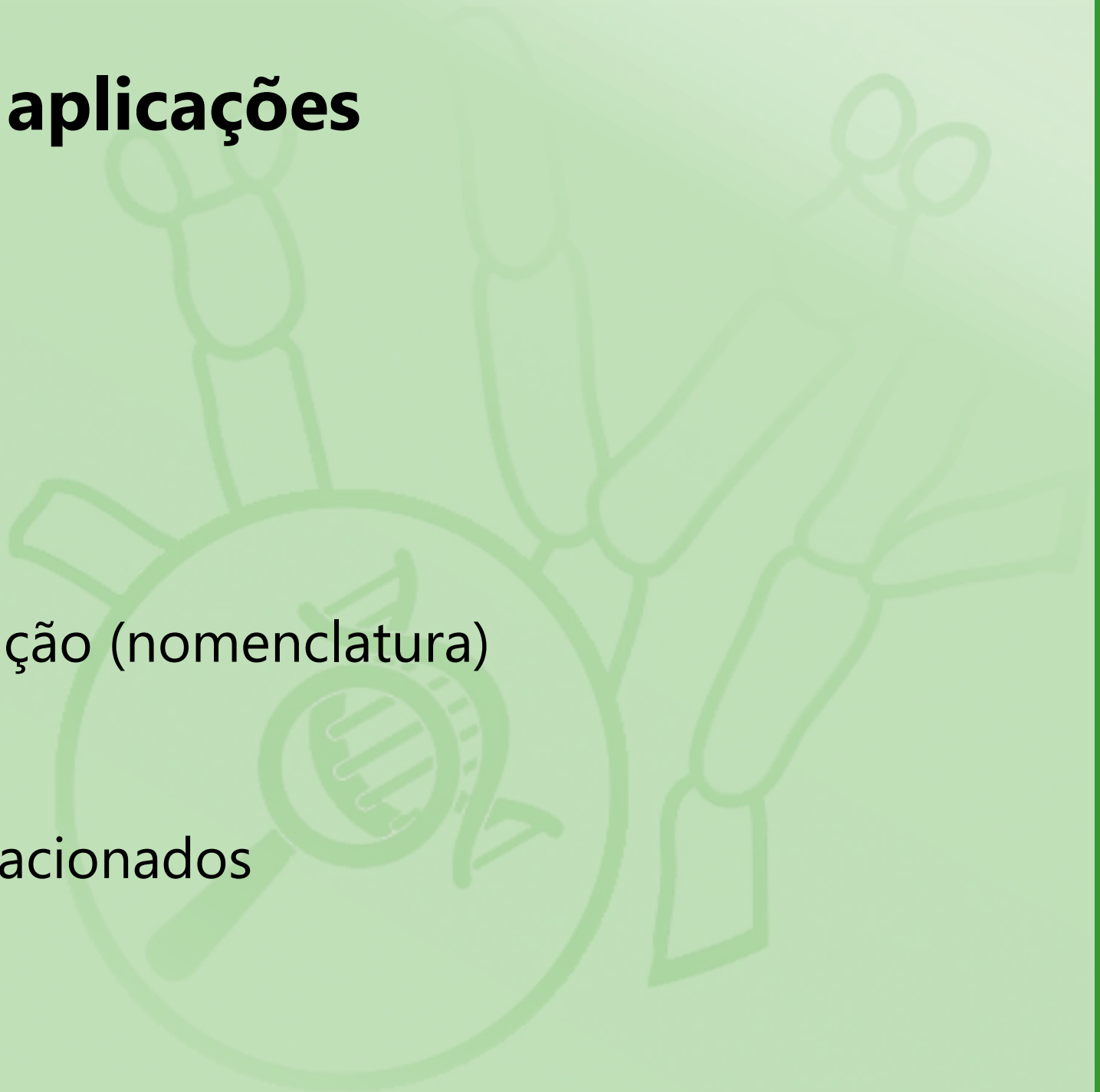
Módulo I – Genômica no Estudo de Microrganismos

# Por que realizar uma anotação funcional?

- Atribuir categorias e funções aos genes anotados e detectados
- Facilidade de estudo dentro das categorias
- Melhor entendimento do modo de vida do organismo analisado
- Possibilidade de comparações funcionais entre organismos de modo de vida similar ou diferente
- Facilidade de seleção de genes candidatos para estudos funcionais com base em categorias de interesse

# Abordagens gerais e aplicações

- Abordagens:
  - *Ab initio*
  - Baseada em homologia
- Aplicações:
  - Classificação e identificação (nomenclatura)
  - Atribuição de função
  - Localização celular
  - Processos biológicos relacionados



# Efetores

- *Ab initio*:
  - Proteínas pequenas ( $< 300$  aa)
  - Ricas em cisteína e pobres em serinas
  - Secretadas (presença de peptídeo sinal, ausência de domínios transmembrana)
  - Menor proporção de aminoácidos alifáticos (apolares e hidrofóbicos)
  - Maior proporção de aminoácidos básicos (polares, positivamente carregados e hidrofílicos)
- Baseada em homologia: bases de dados de efetores ou sequências descritas em estudos prévios
- Alta variabilidade inter e intra-específica: dificuldade de detectar homologia em muitos casos



# Arquivo de saída (lista)

- Em geral, colunas de identificação do gene avaliado, e colunas com a classificação e valor de confiabilidade ou probabilidade

1	# Identifier	Prediction	Probability
2	scaffold_10.6	Non-effector	0.872
3	scaffold_10.13	Non-effector	0.991
4	scaffold_10.23	Non-effector	0.981
5	scaffold_10.40	Non-effector	0.989
6	scaffold_10.41	Effector	0.924
7	scaffold_10.60	Non-effector	0.97
8	scaffold_10.61	Non-effector	0.973
9	scaffold_10.100	Non-effector	0.977
10	scaffold_10.101	Non-effector	0.99
11	scaffold_10.117	Non-effector	0.991
12	scaffold_10.124	Non-effector	0.991
13	scaffold_10.144	Non-effector	0.99
14	scaffold_10.152	Non-effector	0.555
15	scaffold_10.172	Non-effector	0.736
16	scaffold_10.199	Non-effector	0.97
17	scaffold_10.219	Non-effector	0.864
18	scaffold_10.230	Non-effector	0.931
19	scaffold_10.249	Effector	0.873

# CAZymes

- Baseada em homologia:
  - Comparação com CAZymes já caracterizadas em bases de dados de CAZymes (ex: CAZy)
  - Comparação com domínios conservados das classes e famílias de CAZymes:
    - GTs: glycosyltransferases
    - GHs: glycoside hydrolases
    - PL: polysaccharide lyases
    - CE: carbohydrate esterases
    - CBM: carbohydrate-binding modules
    - AA: enzimas auxiliares

# Arquivo de saída (lista)

1	Gene ID	HMMER	Hotpep	DIAMOND	Signalp	#ofTools
2	scaffold_1.1012	GT4(512-682)	N	GT4	N	2
3	scaffold_1.1057	GT90(587-853)	N	N	N	1
4	scaffold_1.1058	GH47(45-507)	GH47	GH47	N	3
5	scaffold_1.1091	AA7(105-310)	N	N	Y(1-23)	1
6	scaffold_1.1104	GH3(90-311)	GH3+CBM1	GH3	Y(1-30)	3
7	scaffold_1.1174	GH5_49(118-421)	N	GH5_49	N	2
8	scaffold_1.1226	GT8(4-224)	GT8	GT8	N	3
9	scaffold_1.1230	GT90(708-991)	GT90	GT90	N	3
10	scaffold_1.1258	AA7(59-483)	N	N	N	1
11	scaffold_1.1296	GH16_18(73-229)	N	N	Y(1-24)	1

- Lista dos genes avaliados, classificação em classe e família de CAZymes e resultados para diferentes ferramentas de identificação utilizadas

# Clusters de metabólitos secundários

- Detecção de clusters
- Comparação com domínios conservados para os diferentes tipos de clusters de metabólitos secundários e diferentes genes presentes no cluster
- Comparação com sequências de genes em clusters já caracterizados (genes diretamente associados à compostos específicos por meio de abordagens funcionais)



# Arquivo de saída (lista simples)

```
6 scaffold_2 scaffold_2 tlpks
scaffold_2.234;scaffold_2.235;scaffold_2.236;scaffold_2.237;scaffold_2.238;scaffold_2.239;scaffold_2.240;
caffold_2.241;scaffold_2.242
scaffold_2.234;scaffold_2.235;scaffold_2.236;scaffold_2.237;scaffold_2.238;scaffold_2.239;scaffold_2.240;
caffold_2.241;scaffold_2.242
7 scaffold_2 scaffold_2 other
scaffold_2.598;scaffold_2.599;scaffold_2.600;scaffold_2.601;scaffold_2.602;scaffold_2.603;scaffold_2.604;
caffold_2.605;scaffold_2.606;scaffold_2.607;scaffold_2.608;scaffold_2.609;scaffold_2.610;scaffold_2.611;s
caffold_2.612;scaffold_2.613;scaffold_2.614;scaffold_2.615;scaffold_2.616
scaffold_2.598;scaffold_2.599;scaffold_2.600;scaffold_2.601;scaffold_2.602;scaffold_2.603;scaffold_2.604;
caffold_2.605;scaffold_2.606;scaffold_2.607;scaffold_2.608;scaffold_2.609;scaffold_2.610;scaffold_2.611;s
caffold_2.612;scaffold_2.613;scaffold_2.614;scaffold_2.615;scaffold_2.616
8 scaffold_3 scaffold_3 nrps
scaffold_3.281;scaffold_3.282;scaffold_3.283;scaffold_3.284;scaffold_3.285;scaffold_3.286;scaffold_3.287;
caffold_3.288;scaffold_3.289;scaffold_3.290;scaffold_3.291;scaffold_3.292;scaffold_3.293;scaffold_3.294
scaffold_3.281;scaffold_3.282;scaffold_3.283;scaffold_3.284;scaffold_3.285;scaffold_3.286;scaffold_3.287;
caffold_3.288;scaffold_3.289;scaffold_3.290;scaffold_3.291;scaffold_3.292;scaffold_3.293;scaffold_3.294
9 scaffold_3 scaffold_3 terpene
scaffold_3.334;scaffold_3.335;scaffold_3.336;scaffold_3.337;scaffold_3.338
scaffold_3.334;scaffold_3.335;scaffold_3.336;scaffold_3.337;scaffold_3.338
```

- Lista de clusters encontrados, classificação e genes pertencentes ao cluster



# Função molecular, localização celular, processo biológico

- Gene Ontology

- **Função molecular (*molecular function*)**: atividades realizadas pelas proteínas de forma mais abrangente, sem especificar a localização celular ou etapa do desenvolvimento em que a atividade ocorre (ex: transporte, atividade catalítica)
- **Localização celular (*cellular component*)**: localização ou compartimento celular em que as proteínas desempenham uma atividade ou atuam de forma estrutura (ex: mitocôndria, ribossomo, citoplasma)
- **Processo biológico (*biological process*)**: processos biológicos abrangentes que são realizados através de atividades específicas (ex: reparo de DNA, transdução de sinal, processos biossintéticos específicos)

# **Função molecular, localização celular, processo biológico**

- Entender o organismo e seus genes em termos mais abrangentes
- Compreender a interação entre diferentes proteínas ou produtos gênicos em uma mesma função molecular ou processo biológico
- Relacionar funções e processos ao modo de vida do organismo e utilizar esta informação em análises comparativas

# **Função molecular, localização celular, processo biológico**

- Comparação com domínios conservados de genes associados à cada categoria
- Comparação com ortólogos caracterizados em cada categoria, e evolutivamente relacionados ao grupo analisado
  - Por exemplo: Ao analisar uma linhagem de uma espécie da classe Dothideomycetes, utilizar os ortólogos desta classe e não ortólogos de Sordariomycetes

Anotação  
gênica

Anotação de  
elementos  
transponíveis

Anotação funcional:  
- Efetores  
- CAZymes  
- Clusters de  
metabólitos  
secundários

Anotação funcional:  
- Função molecular  
- Localização celular  
- Processo biológico

Análises  
comparativas

- Responder à pergunta inicial
  - Testar as hipóteses
- Sugerir novas perspectivas
- Fornecer bases para estudos futuros