CEN5789 - Genômica e Bioinformática

Almir Moreira Gonçalves Júnior
Amanda Cristina Fraga de Albuquerque
Ana Carolina Lopes da Silva

Pangenoma e Genômica Comparativa

Piracicaba
Outubro de 2024

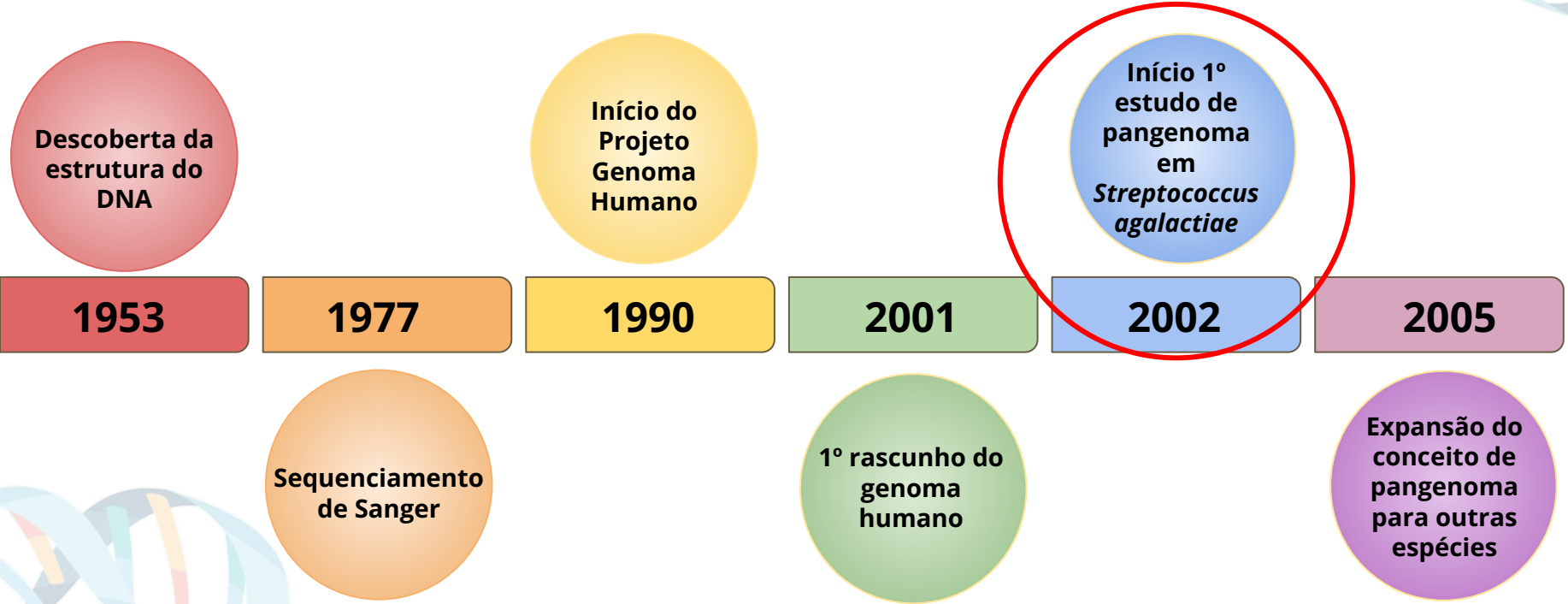
Introdução

Transformação na Biologia: Com o avanço das tecnologias de sequenciamento, a genômica revolucionou nosso entendimento sobre a biodiversidade e a evolução;

Variabilidade Genética: O genoma de cada espécie não é um conjunto fixo e uniforme; existe uma vasta diversidade genética entre diferentes cepas de uma mesma espécie;

Impacto Evolutivo: Esse estudo permite revelar os processos evolutivos, entender adaptações específicas e a especialização de organismos em seus respectivos nichos.

A Jornada do Sequenciamento: Como Chegamos ao Pangenoma



Pangenoma de *Streptococcus agalactidae*

Primeiro Estudo de Pangenoma em Bactérias

Tettelin *et al.* (2005) sequenciaram múltiplas cepas de *Streptococcus agalactiae* e observam que o genoma de cada cepa não é idêntico, mas apresenta uma combinação de genes comuns e exclusivos.

Introduz-se o termo "pangenoma" para descrever o conjunto completo de genes em todas as cepas de uma espécie

RESEARCH ARTICLE | BIOLOGICAL SCIENCES |

Genome analysis of multiple pathogenic isolates of *Streptococcus agalactiae*: Implications for the microbial "pan-genome"

Hervé Tettelin, Vega Masignani, Michael J. Cieslewicz, Claudio Donati, Duccio Medini, Naomi L. Ward, Samuel V. Angiuoli, Jonathan Crabtree, Amanda L. Jones, A. Scott Durkin, Robert T. DeBoy, Tanja M. Davidsen, Marirosa Mora, Maria Scarselli, Immaculada Margarit y Ros, Jeremy D. Peterson, Christopher R. Hauser, Jaideep P. Sundaram, William C. Nelson, Ramana Madupu, Lauren M. Brinkac, Robert J. Dodson, Mary J. Rosovitz, Steven A. Sullivan, Sean C. Daugherty, Daniel H. Haft, Jeremy Selengut, Michelle L. Gwinn, Liwei Zhou, Nikhat Zafar, Hoda Khouri, Diana Radune, George Dimitrov, Kisha Watkins, Kevin J. B. O'Connor, Shannon Smith, Teresa R. Utterback, Owen White, Craig E. Rubens, Guido Grandi, Lawrence C. Madoff, Dennis L. Kasper, John L. Telford, Michael R. Wessels, Rino Rappuoli, and Claire M. Fraser -42 Authors Info & Affiliations

September 19, 2005 | 102 (39) 13950-13955 | https://doi.org/10.1073/pnas.0506758102

O que é um Pangenoma?

Número total de genes não redundantes presentes num conjunto de dados.
Esse conjunto de dados é dividido em três categorias:

- Genoma núcleo;
- Genoma acessório;
- Genes espécie-específicos.

(Muzzi & Donati, 2011)

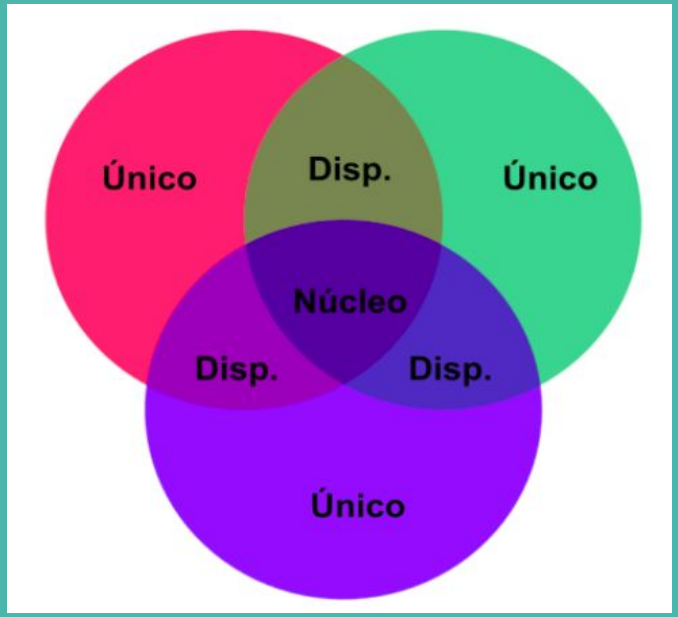


Diagrama de Venn ilustrando as categorias de um pangenoma: Núcleo (genes compartilhados por todos os isolados), Genoma Acessório (genes presentes em alguns, mas não em todos os isolados) e Único (genes exclusivos de um isolado).
Fonte: Guimarães, 2020.

Genoma núcleo

- Presentes em todos os isolados, essenciais para funções básicas (replicação, tradução e homeostase celular);
- Quanto mais relacionados filogeneticamente, maior o número de genes no núcleo.

Genoma acessório

- Presentes em alguns, mas não todos os isolados;
- Relacionados a funções específicas, como sobrevivência, resistência a antibióticos e virulência;
- Origem provável: transferência horizontal e adaptação ao ambiente.

Genes espécie-específicos

- Genes presentes em apenas um isolado;
- Funções adaptativas: virulência e patogenicidade em organismos patogênicos ou rotas metabólicas em não-patogênicos.

Udaondo *et al*. (2016); Machado *et al.* (2019).

Eucariotos

- Conjunto de todos os genes presentes em diferentes cepas;
- Alta diversidade genética entre cepas devido à transferência horizontal de genes;
- Genes exclusivos conferem adaptação e resistência a ambientes específicos.

Procariotos

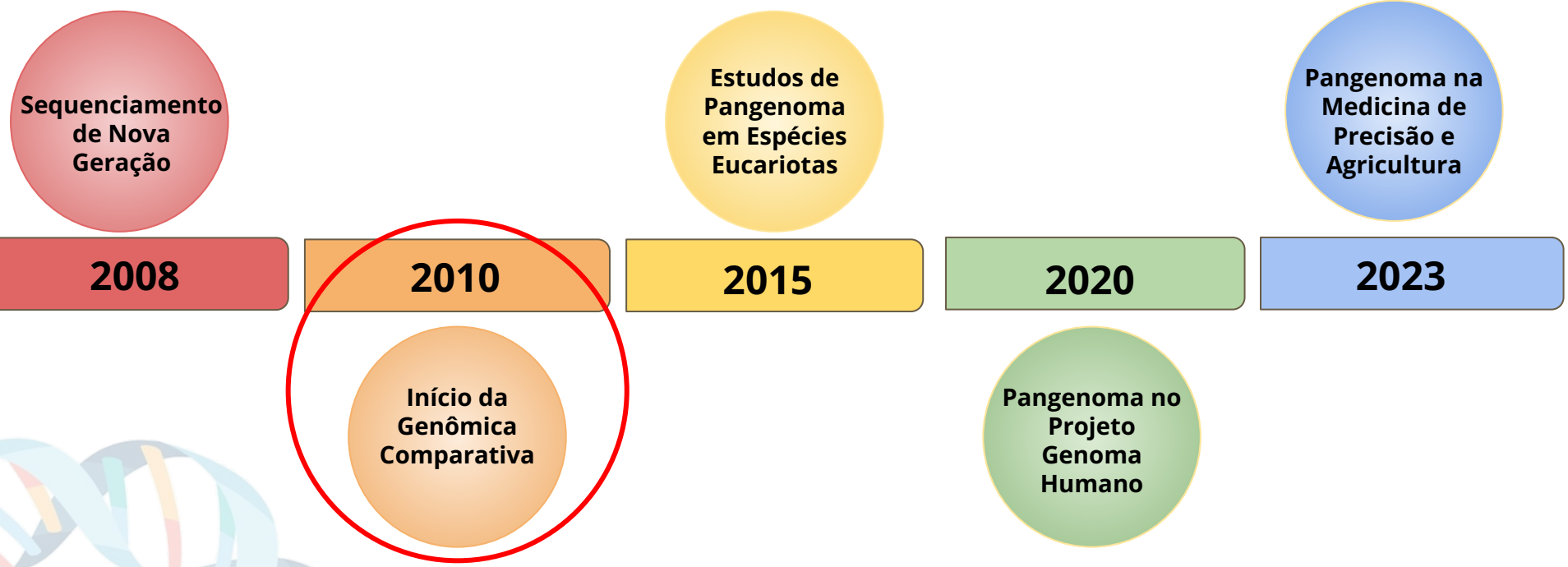
- Coleção de todas as sequências de DNA na espécie;
- Menor variabilidade genética entre indivíduos;
- Inclui genes e sequências regulatórias que influenciam características e adaptações.

Outras classificações: Soft, Shell e Cloud Genome

- Soft genoma representa os genes presentes em 95% dos genomas;
- Shell genoma é a definição para os genes compartilhados entre 10% e 88% dos genomas;
- Cloud genoma que agrupa genes presentes em menos de 10% dos genomas.

Bezuidt *et al.* (2016)

A Jornada do Sequenciamento: Como Chegamos ao Pangenoma



O que é Genômica Comparativa?

- Compara genomas de diferentes organismos;
- Identifica semelhanças e diferenças genéticas;
- Explora variações evolutivas e adaptações moleculares;
- Analisa a presença/ausência e organização de genes;
- Relaciona características genéticas às funções e adaptações em diferentes ambientes.

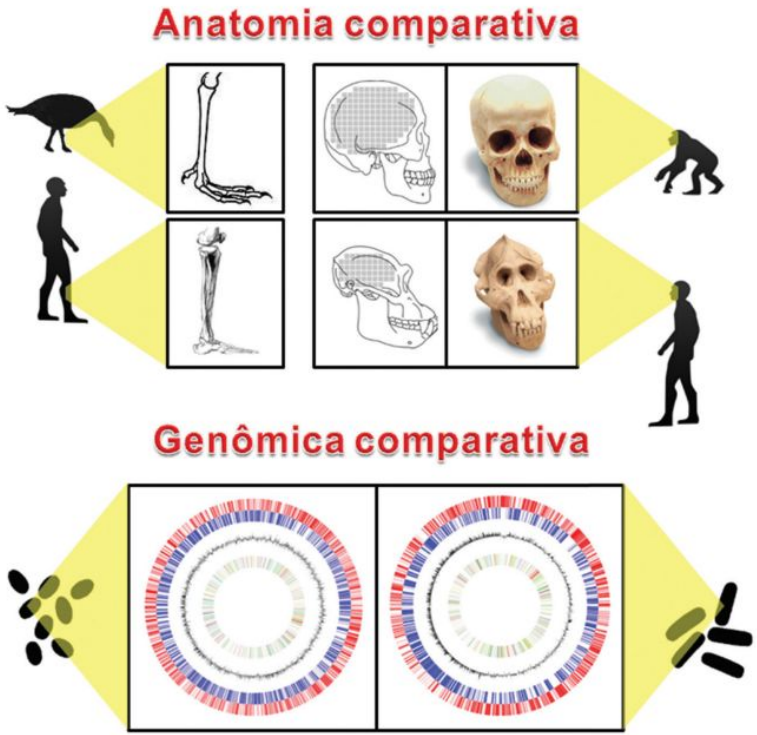
Importância e Aplicações

Estudo da Evolução:

Compara genomas para entender relações filogenéticas e evolução de características.

Identificação de Genes Funcionais:

Revela genes conservados entre espécies, indicando funções essenciais em processos biológicos.



Fonte: Prosdócimi & Moreira (2015)

Importância e Aplicações

Análise de Doenças:

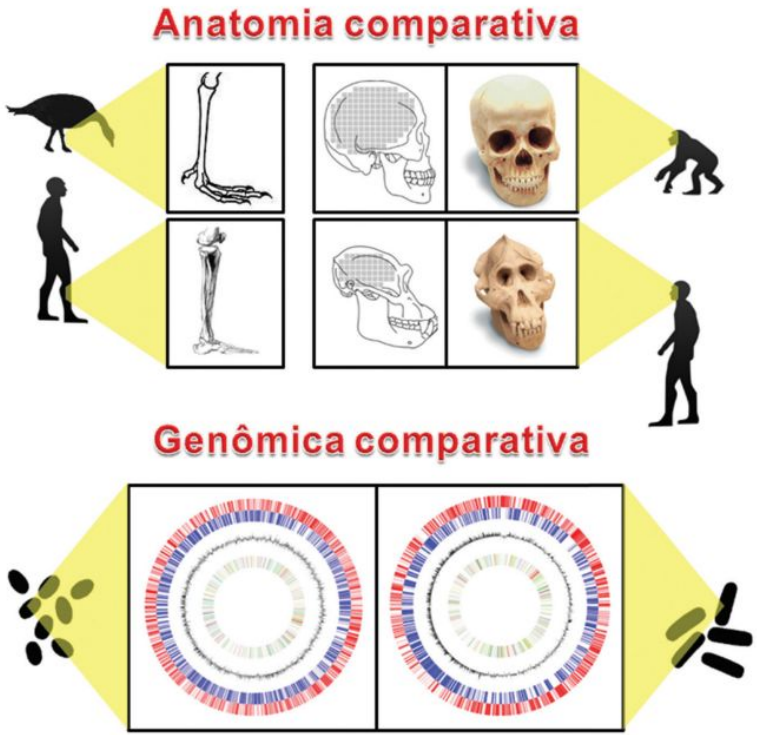
Compara expressão gênica em células normais e tumorais, identificando alvos terapêuticos para o câncer.

Biotecnologia e Melhoramento Genético:

Auxilia na seleção de genes para resistência a doenças e aumento de produtividade em plantas e animais.

Estudos de Sintenia:

Analisa a organização estrutural dos genes, investigando impactos na evolução e adaptação.



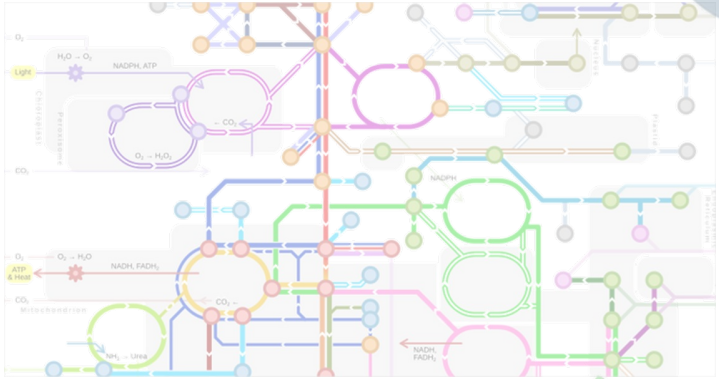
Fonte: Prosdócimi & Moreira (2015)

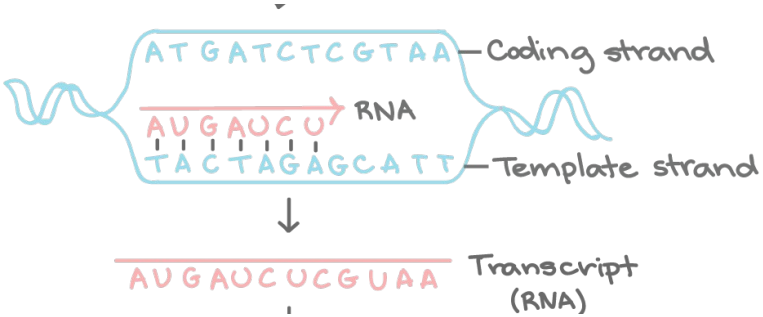
Por que comparar genomas?

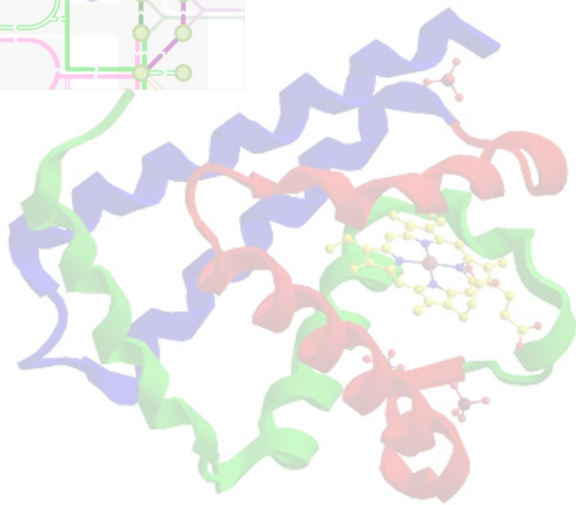
1. Descrever características hereditárias;
2. Organismos relacionados compartilham genes ancestrais;
3. Elementos funcionais codificantes podem ser semelhantes em organismos distintos;
4. Identificação de elementos funcionais no genoma;
5. Compreensão funcional em organismos modelos podem ser transmitidas para organismos não-modelo;
6. Revelam processos e restrições evolutivas.

Genômica comparativa

- Genes
- Transcritos de RNA
- Proteínas
- Vias bioquímicas e metabólicas

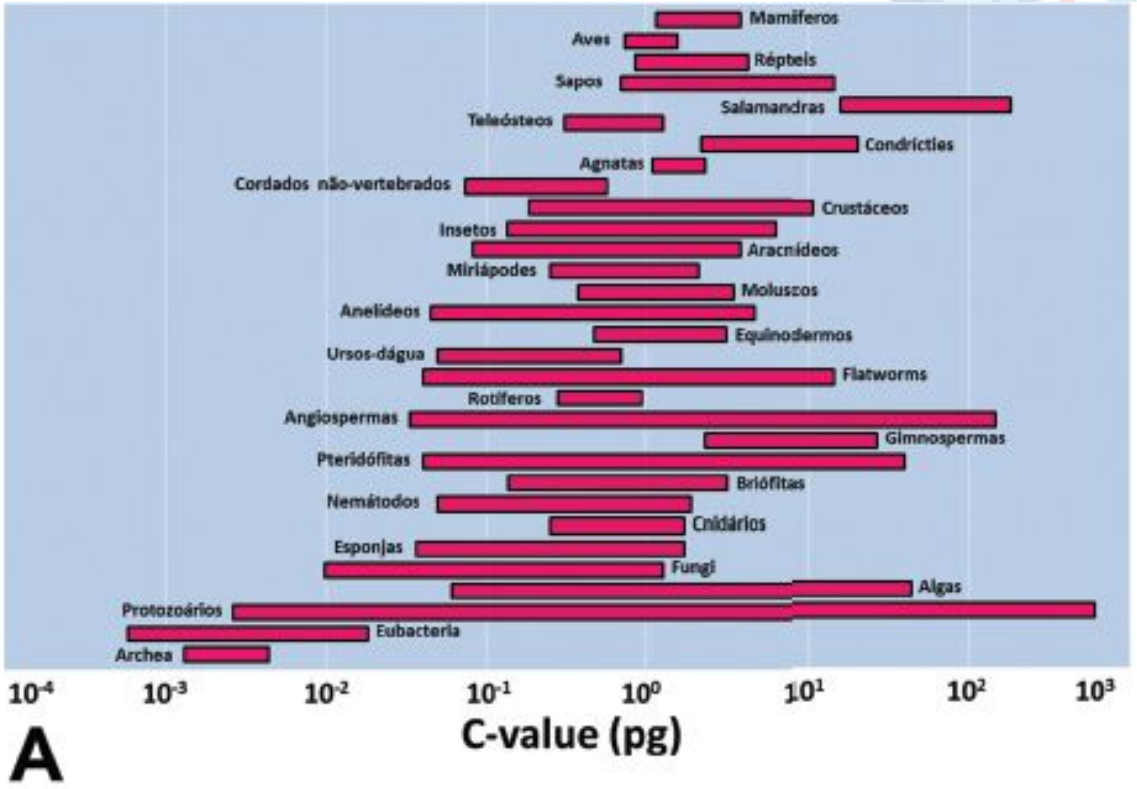




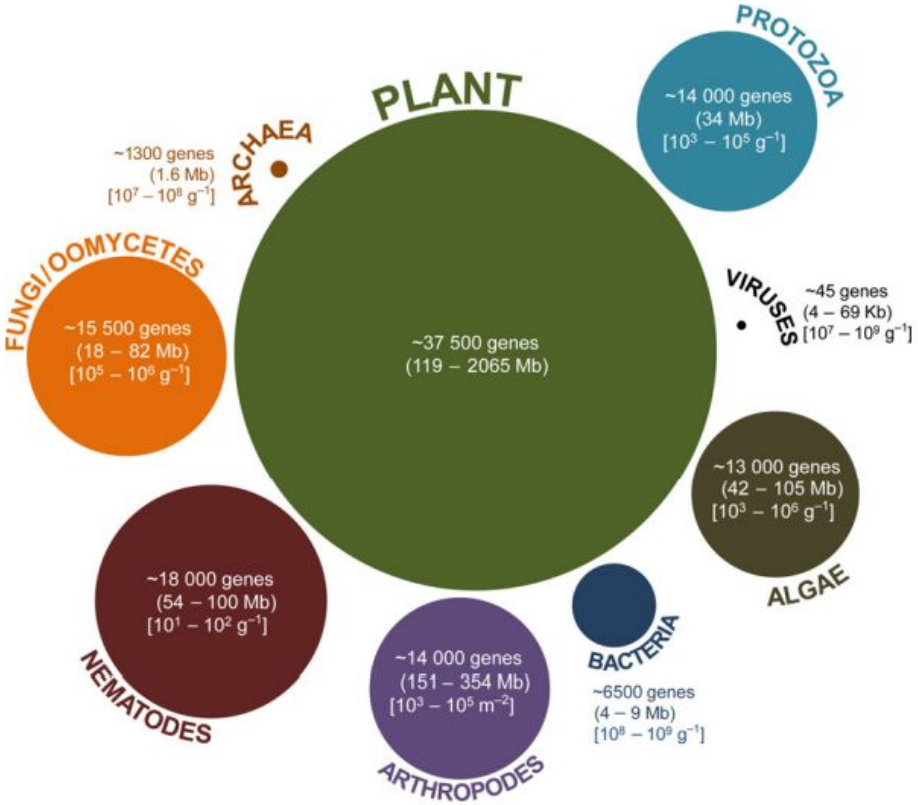


Genômica comparativa - Genes

1. Inicialmente feitas em escala quantitativa



Genômica comparativa - Genes



doil: 10.1111/1574-6976.12028

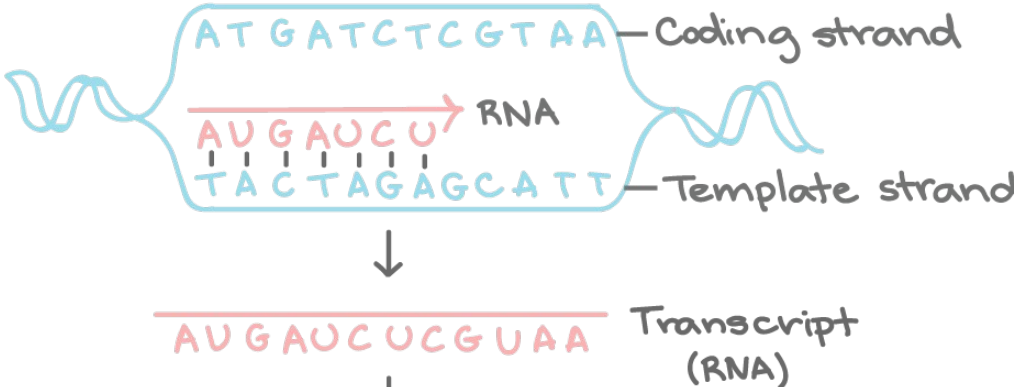
Genômica comparativa - Genes

2. É importante realizar a catalogação e a comparação de vários fatores em projetos de genômica comparativa. Pode-se calcular, por exemplo

a. O número de genes por cromossomo entre diversas espécies;
b. O número médio de éxons ou íntrons de cada gene;
c. A distribuição de nucleotídeos A, C, T ou G pelos genes;
d. A utilização de códons preferenciais nos genes codificadores de proteínas;
e. A utilização de pares de nucleotídeos (ou dinucleotídeos) entre os genes.

Genômica comparativa - Transcritos de RNA

- Transcriptômica comparativa e os genes oriundos de DNA expresso em RNA mensageiro, ou cDNA.
- Ao analisar transcritos totais em condições celulares contrastantes, revela como as células respondem a diferentes estímulos regulando a ativação e desativação de genes.

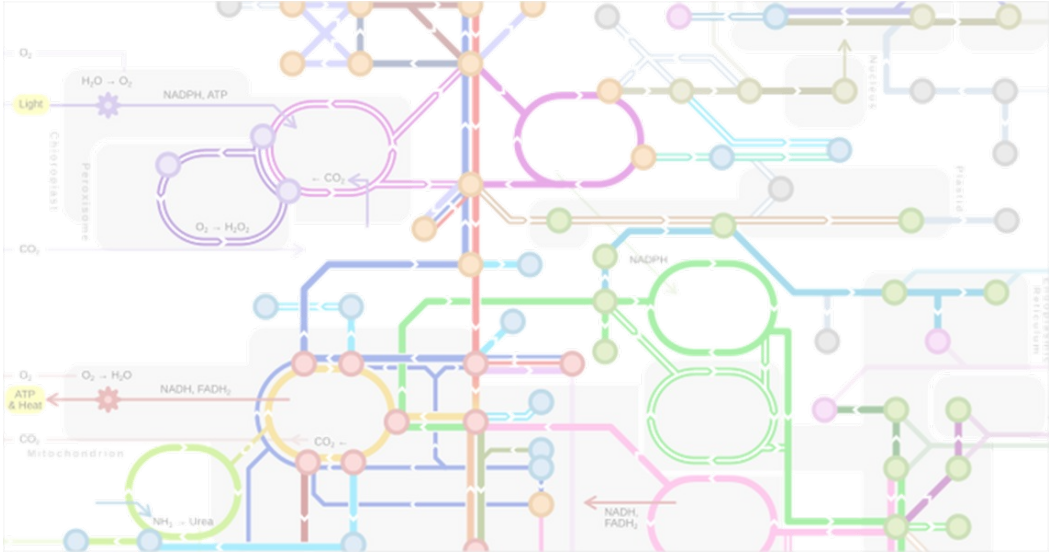


Genômica comparativa - Proteínas

- Quais proteínas estão ou não presentes em um determinado genoma e que, através do agrupamento das proteínas por via bioquímica, possam identificar quais vias um organismo é capaz – ou não – de produzir.

Genômica comparativa - Vias bioquímicas e metabólicas

- Permite compreender quais vias metabólicas ou subprodutos, diferentes organismos conseguem produzir para fins diversificados, quase sempre relacionados à adaptação diante de condições fisiológicas as quais foram expostos.



Recapitulando...

Genes **Conservados**:

Os genes de um genoma que são compartilhados entre outros genomas

Genes **Homólogos**:

São genes conservados e que são derivados de um mesmo gene de um organismo ancestral

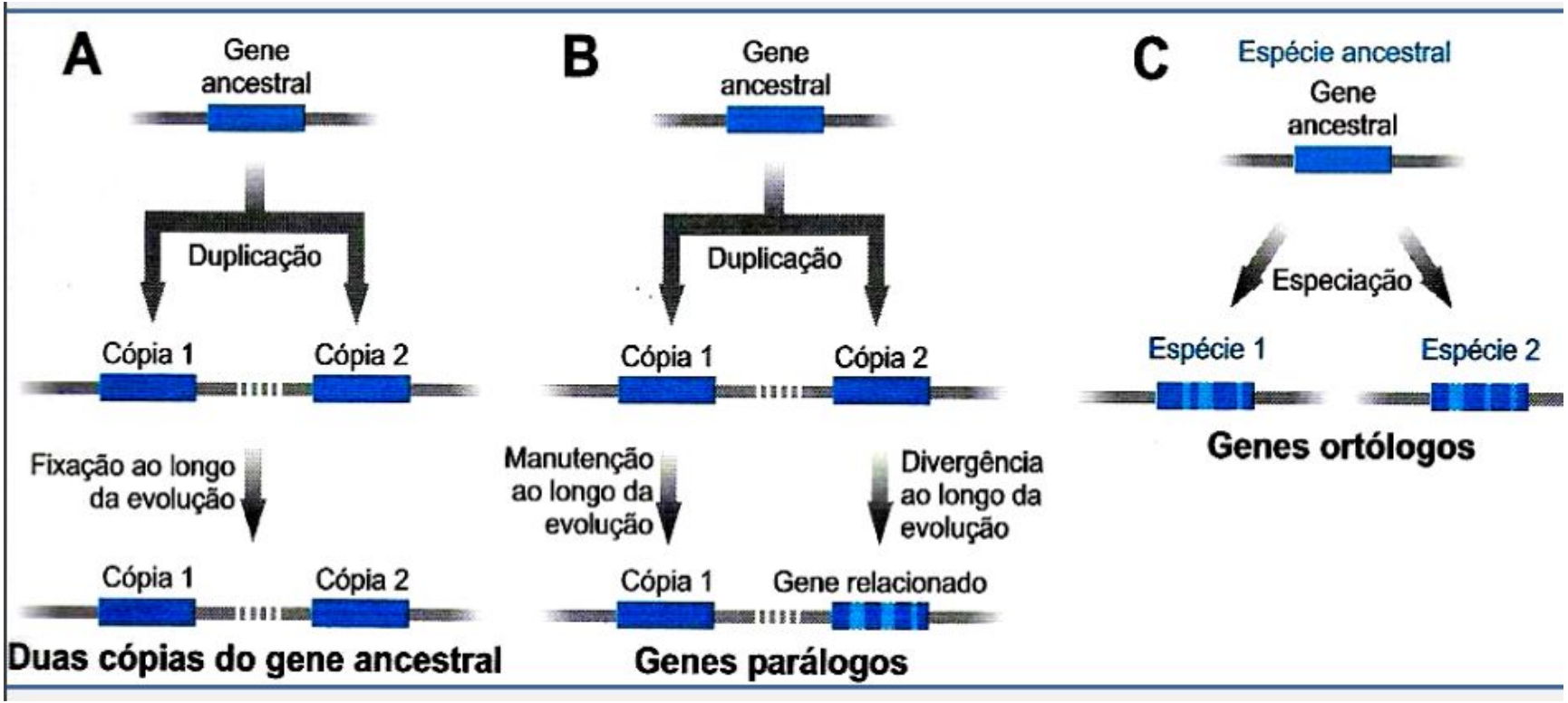
Recapitulando...

Genes **Ortólogos**:

Originados por eventos de especiação

Genes **Parálogos**:

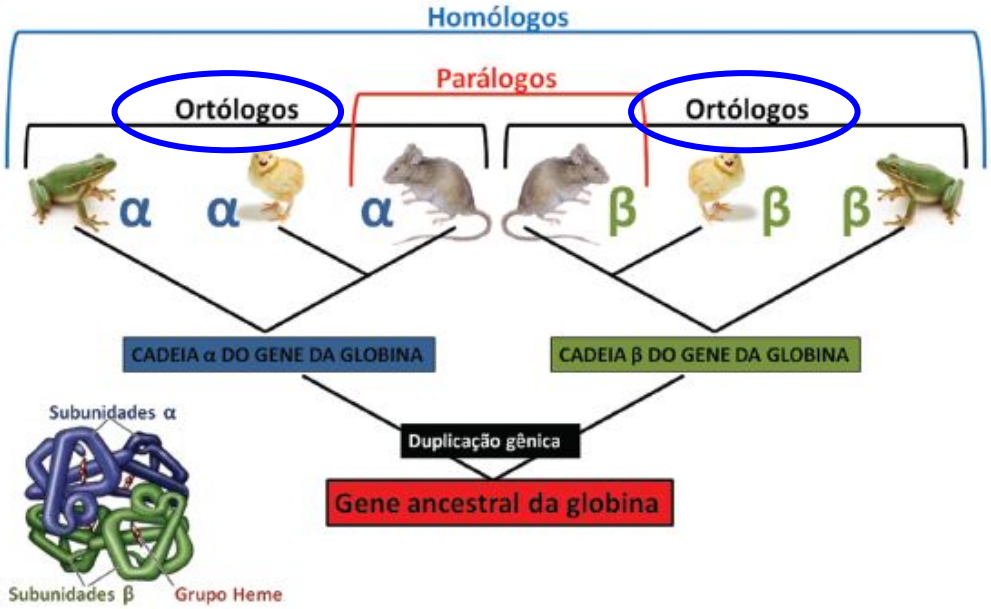
Duplicação dentro de de uma mesma espécie


A
Gene ancestral
Duplicação
Cópia 1
Cópia 2
Fixação ao longo da evolução
Cópia 1
Cópia 2
Duas cópias do gene ancestral
B
Gene ancestral
Duplicação
Cópia 1
Cópia 2
Manutenção ao longo da evolução
Divergência ao longo da evolução
Cópia 1
Gene relacionado
Genes parálogos
C
Espécie ancestral
Gene ancestral
Especiação
Espécie 1
Espécie 2
Genes ortólogos

Inferir Homologia

- Ajuda a compreender a distância evolutiva entre os organismos.

- Inferir funções a genes recém-sequenciados.



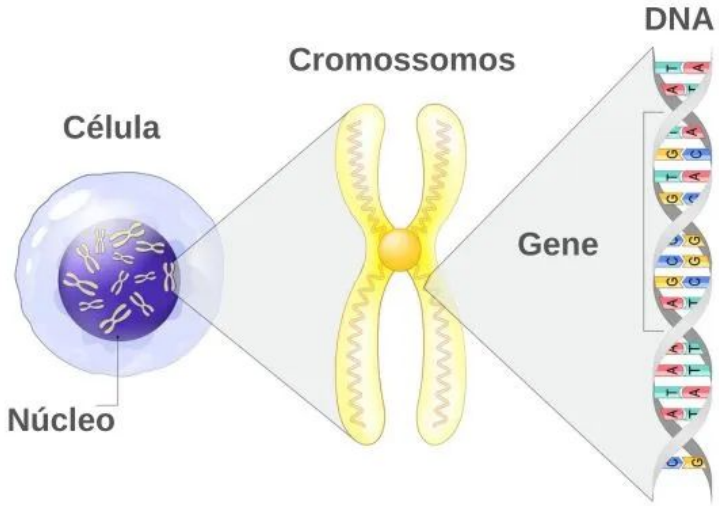
Inferir Homologia

- Anotação genômica, identificação de famílias de proteínas e reconstrução de árvores filogenéticas.

- Além disso, os grupos ortólogos podem auxiliar na análise funcional e evolutiva dos organismos

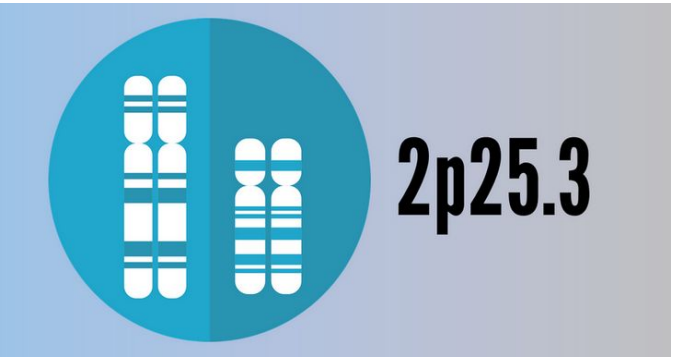
Sintenia

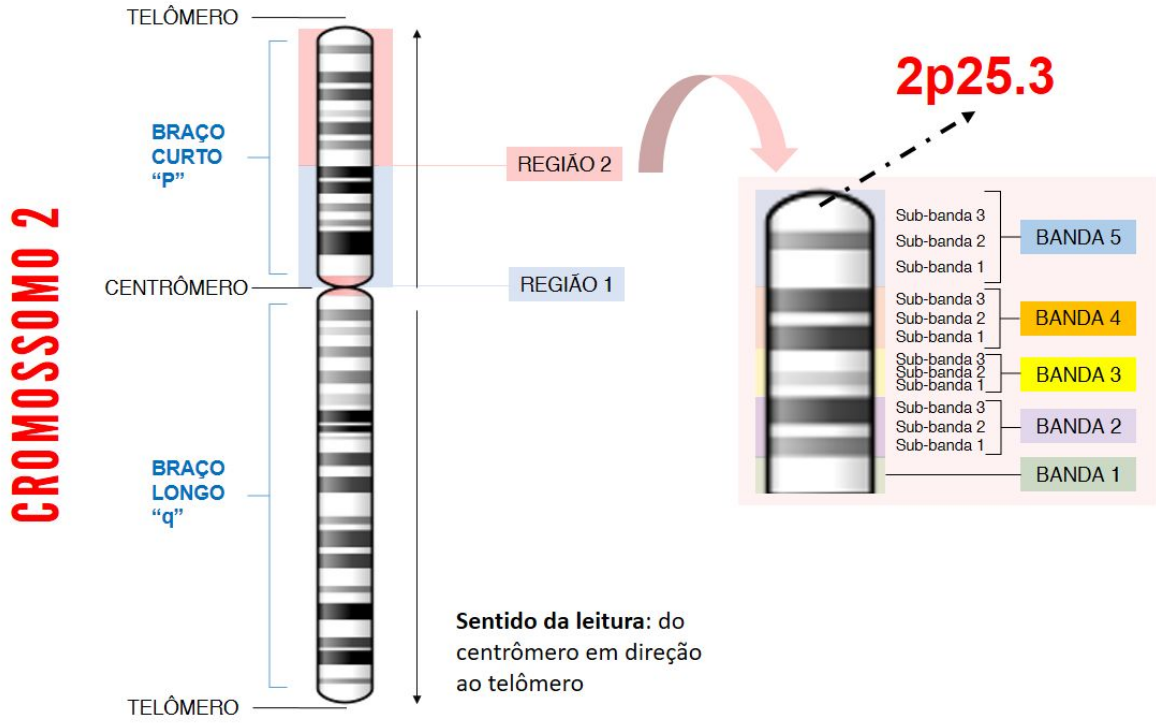
Como a ordem dos genes ao longo dos genomas foi se modificando ao longo do tempo, caso venham a ser alterados.



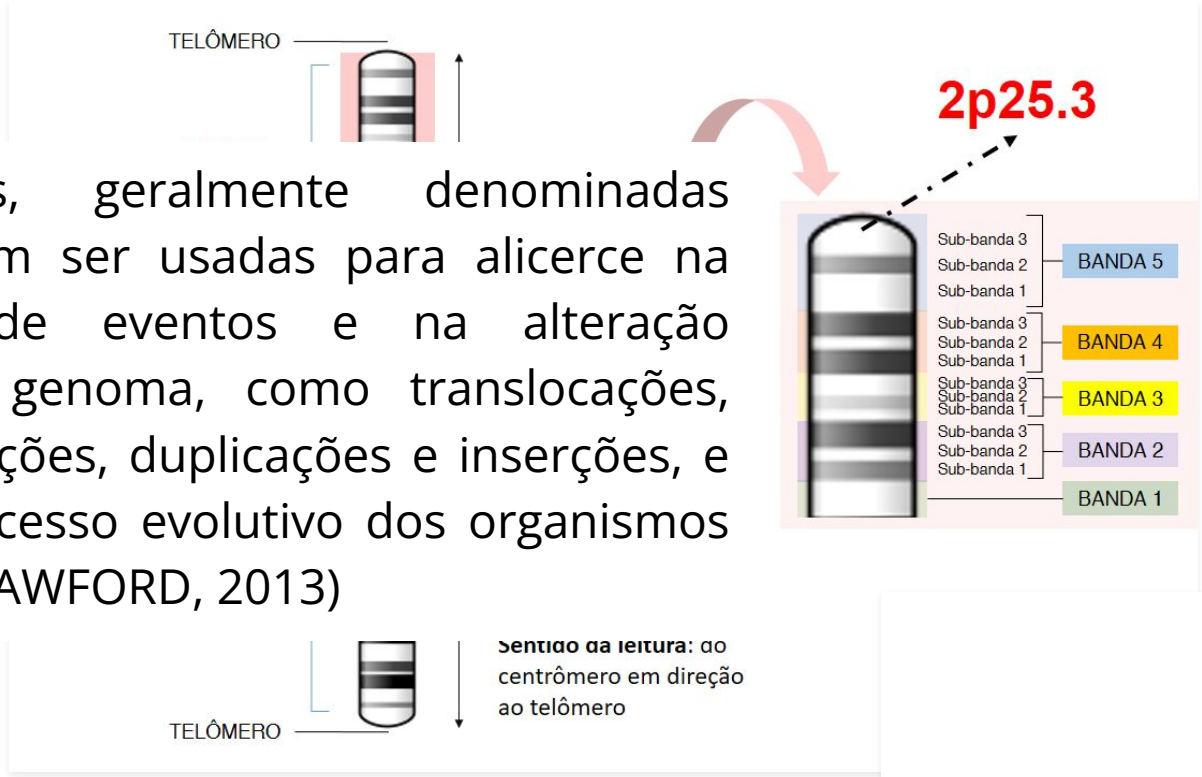
Estudar a comparação entre a ordem dos genes em diferentes organismos.

Sintenia


2p25.3

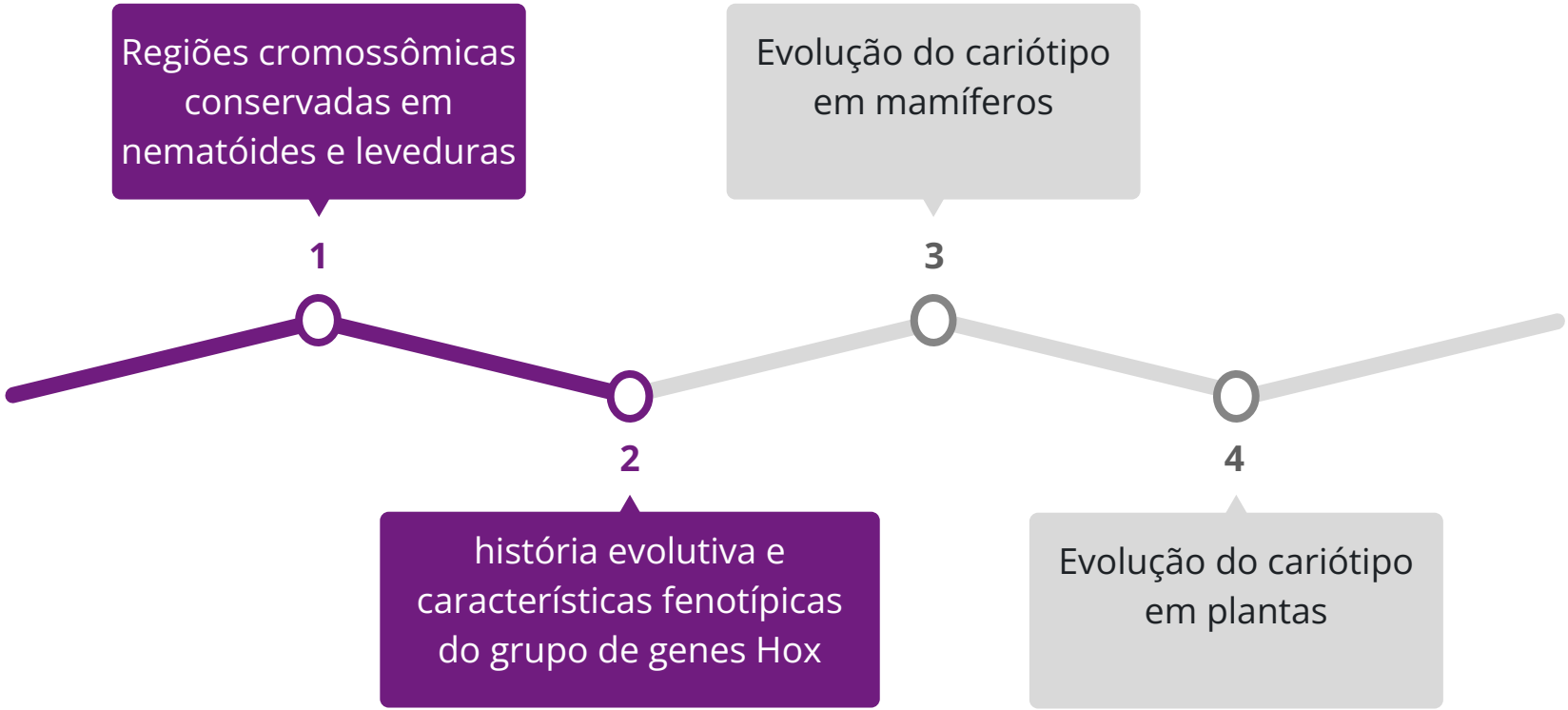

CROMOSSOMO 2
TELÔMERO
BRAÇO CURTO "P"
CENTRÔMERO
BRAÇO LONGO "q"
TELÔMERO
REGIÃO 2
REGIÃO 1
2p25.3
Sub-banda 3
Sub-banda 2
Sub-banda 1
BANDA 5
BANDA 4
BANDA 3
BANDA 2
BANDA 1
Sentido da leitura: do centrômero em direção ao telômero

Sintenia



Estas regiões, geralmente denominadas “blocos”, podem ser usadas para alicerce na identificação de eventos e na alteração estrutural do genoma, como translocações, inversões, deleções, duplicações e inserções, e refletem o processo evolutivo dos organismos analisados. (CRAWFORD, 2013)

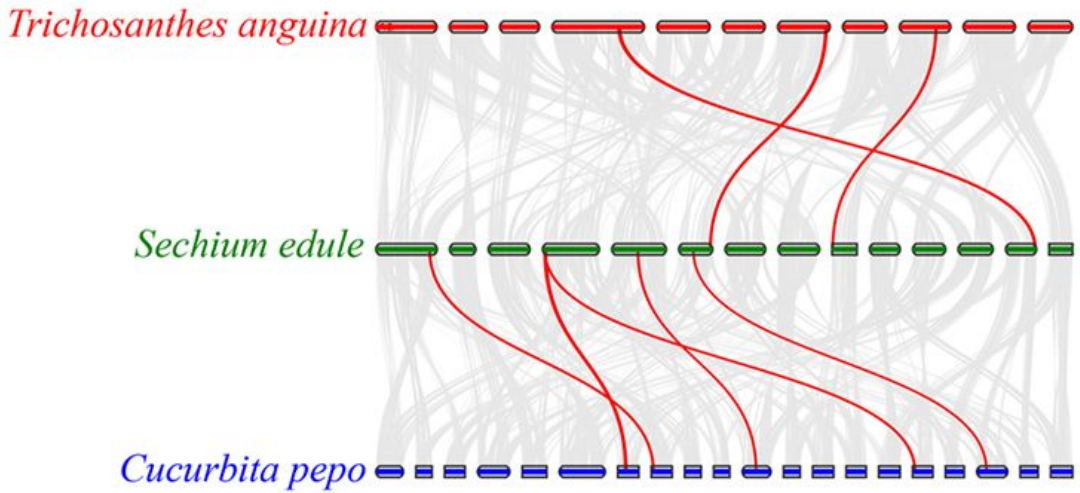
Inferindo Sintenia



Inferindo Sintenia

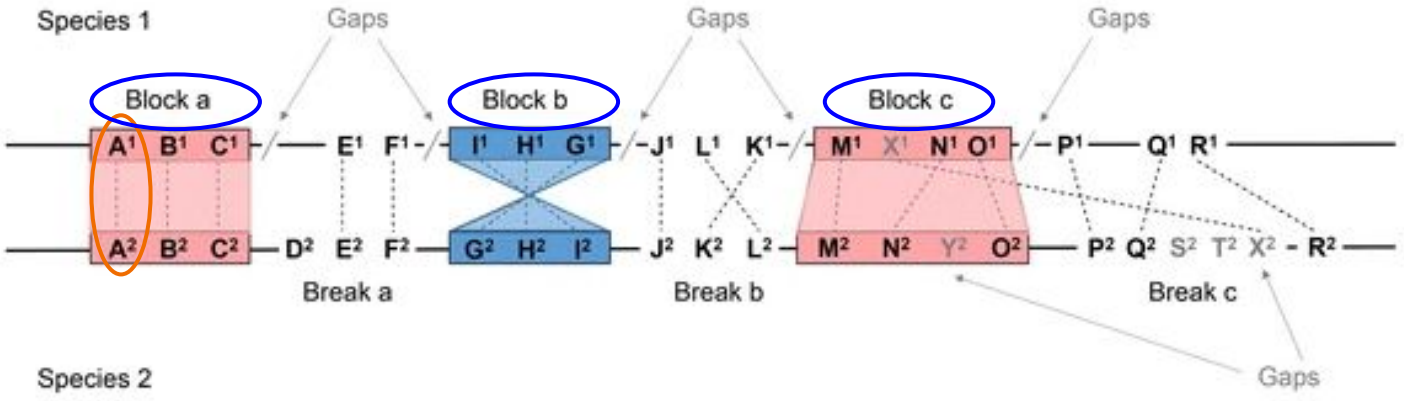
Bloco de sintenia:

- Conjunto de genes ortólogos e que estão organizados na mesma ordem em ambos os genomas comparados.



Inferindo Sintenia

- Para que um bloco de sintenia seja identificado, é necessário que haja um número mínimo de genes ortólogos (chamados de âncoras) que estejam co-arranjados.



https://doi.org/10.1186/s12859-018-2026-4

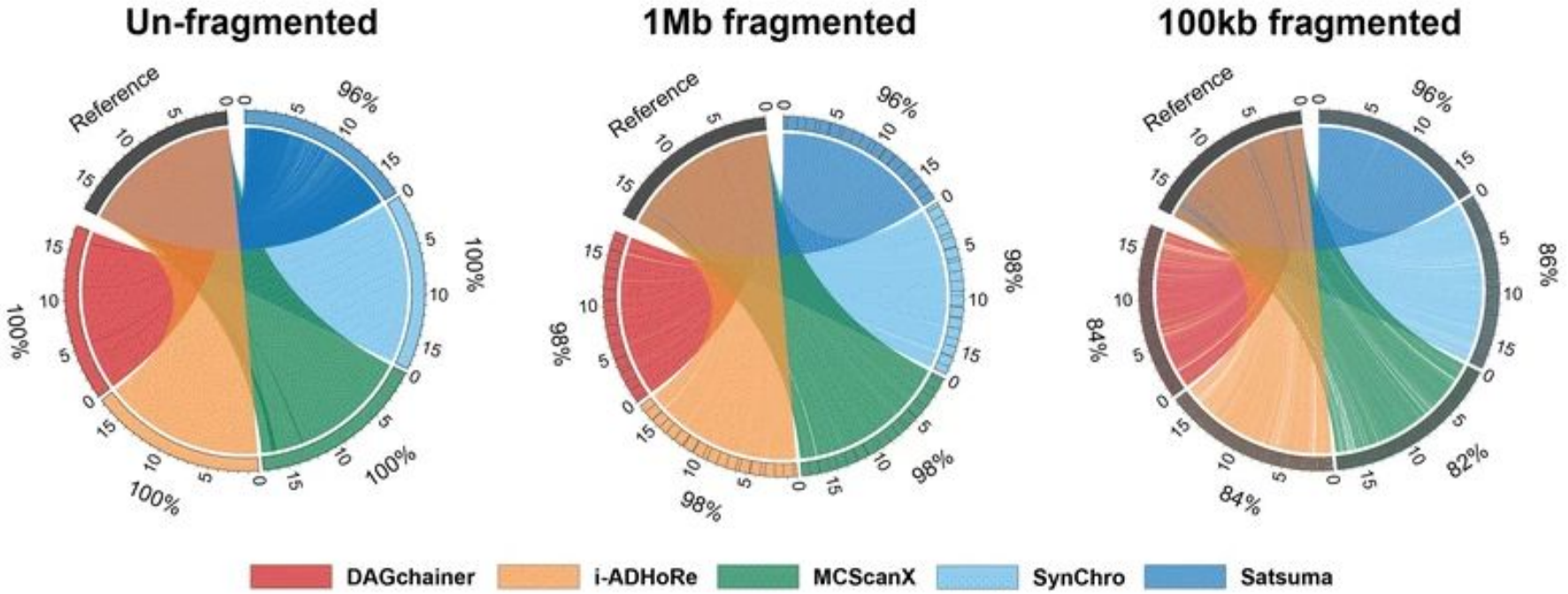
Programas para inferir sintenia em âncoras

Programa	Aplicação	Tipos de genoma
DAGchainer	Encontra blocos de genes conservados com base em alinhamentos de pares.	Comparação de genomas de espécies próximas
i-ADHoRe	Detecta regiões homólogas e sintenia	Genomas poliplóides
MCScanX	Duplicação segmentar, traça blocos de sintenia e calcula a conservação de genes.	Múltiplos genomas ou dentro de genomas duplicados
SynChro	Calcula sintenia global e loval	Genomas próximos ou distantes
Satsuma*	Sintenia global entre genomas	Genomas de espécies distantes

*baseada somente em alinhamentos de nucleotídeos

Inferindo Sintenia

Caenorhabditis elegans

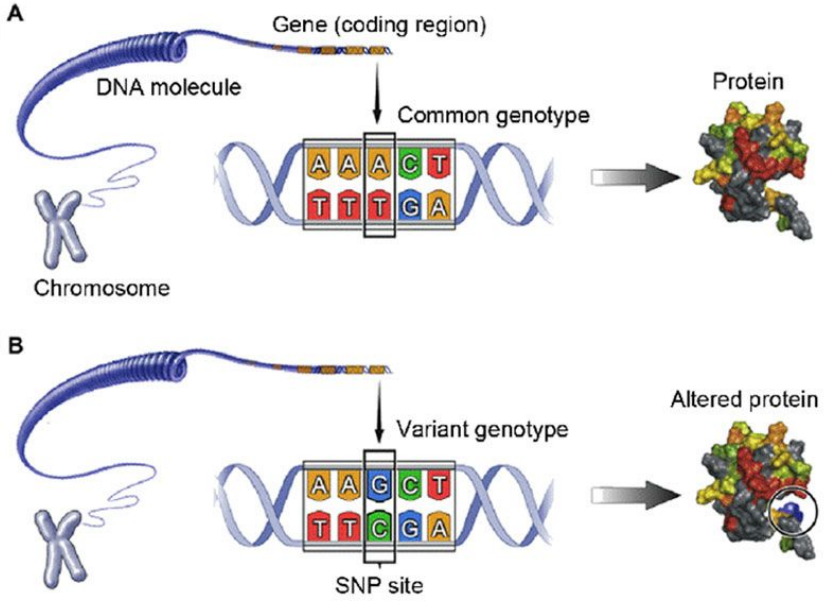


https://doi.org/10.1186/s12859-018-2026-4

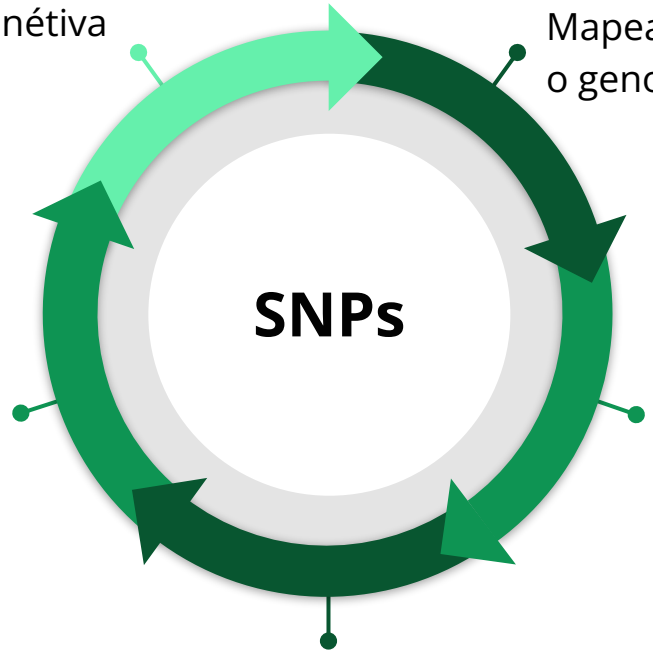
SNPs (*Single Nucleotide Polymorphism*)

- Variação genética que ocorre quando uma única base de nucleotídeo em uma sequência de DNA é alterada;

- Para ser considerada polimorfismo, essa variação deve estar presente em pelo menos 1% dos indivíduos



SNPs (*Single Nucleotide Polymorphism*)



Diversidade Genétiva

Mapeamento de associação em todo o genoma

Estudos de história evolutiva populacional

Seleção genômica

Relações filogenéticas

Os SNPs são os mais abundantes e robustos, viáveis para genotipagem automatizada de alto rendimento, altamente reprodutíveis e podem ser usados para identificar variantes, substituindo os marcadores anteriores devido ao alto rendimento, eficiência e custo-efetividade.

https://doi.org/10.1186/s12863-023-01109-6

Exemplo

RESEARCH Open Access

Single nucleotide polymorphism (SNP) markers for genetic diversity and population structure study in Ethiopian barley (*Hordeum vulgare* L.) germplasm

Mihret Yirgu[1,2*], Mulugeta Kebede[3], Tileye Feyissa[4], Berhane Lakew[5], Aemiro Bezabih Woldeyohannes[6] and Mulusew Fikere[7]

https://doi.org/10.1186/s12863-023-01109-6

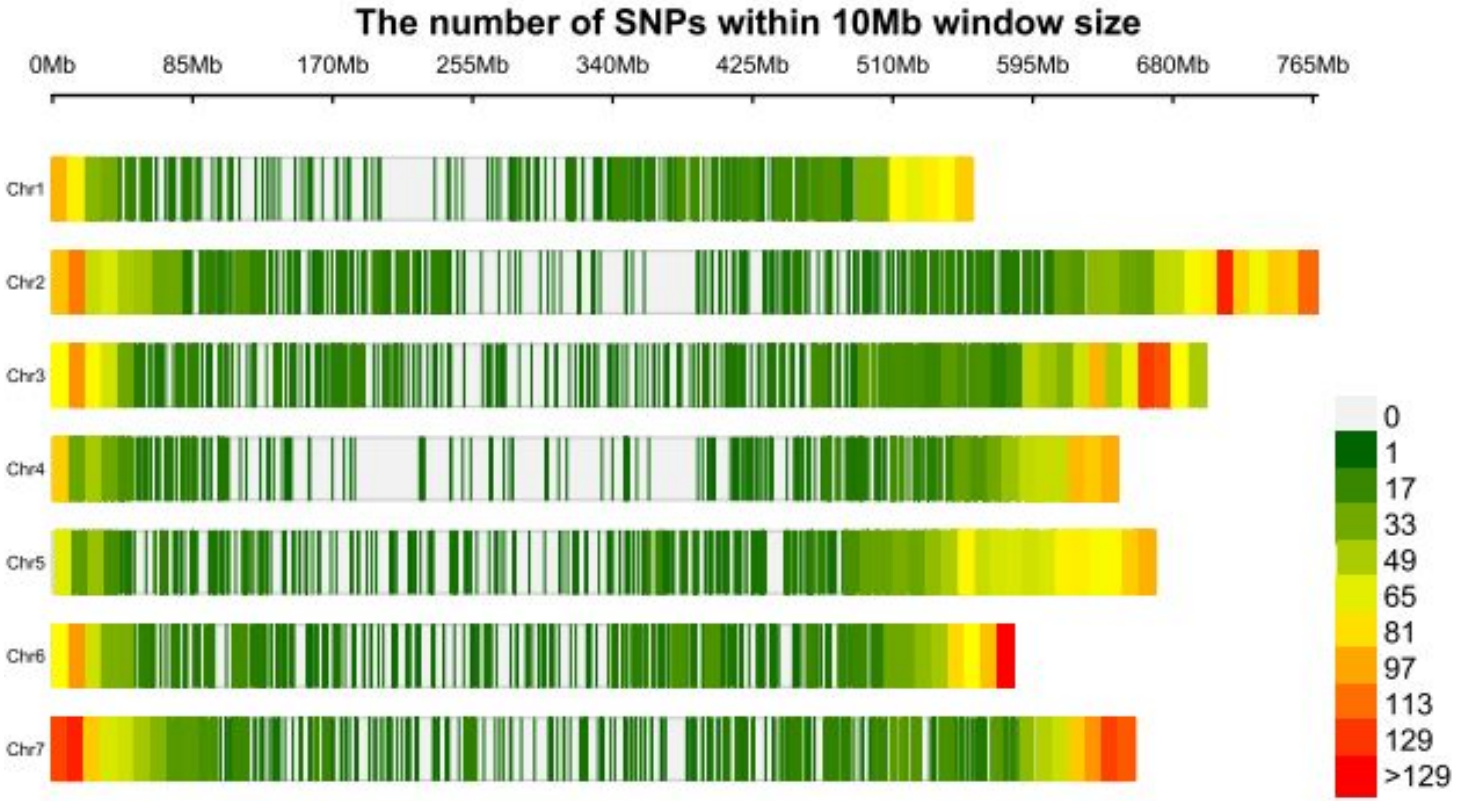


Fig. 1 Distribution of SNP markers within 10 Mb window size across seven chromosomes. Colored bars are SNP counts in 10 Mb interval

https://doi.org/10.1186/s12863-023-01109-6

Como os SNPs são identificados?

- Os SNPs são identificados através de técnicas de sequenciamento de DNA, como o sequenciamento de nova geração (NGS) e o genotipagem em larga escala

Programa	**Característica**
GATK (Genome Analysis Toolkit)	Utilizado para chamadas de variantes, incluindo SNPs e indels
PLINK	Análise em larga escala e voltada para estudos de associação
IGV (Integrative Genomics Viewer)	permite inspecionar alinhamentos e variantes manualmente.
GEMMA	ajusta efeitos genéticos e ambientais.
SnpEff / SnpSift	Ferramentas para anotação e filtro de SNPs com base em bancos de dados genômicos.

Aplicação

1. **Estudos de associação genética (GWAS)**: Identificar variantes ligadas a doenças complexas.
2. **Predição de risco genético**: Avaliar a suscetibilidade a doenças.
3. **Seleção genética em agricultura**: Melhorar características desejáveis em plantas e animais.
4. **Medicina personalizada**: Ajustar tratamentos com base no perfil genético.
5. **Estudo de evolução**: Investigar padrões de ancestralidade e adaptação populacional.

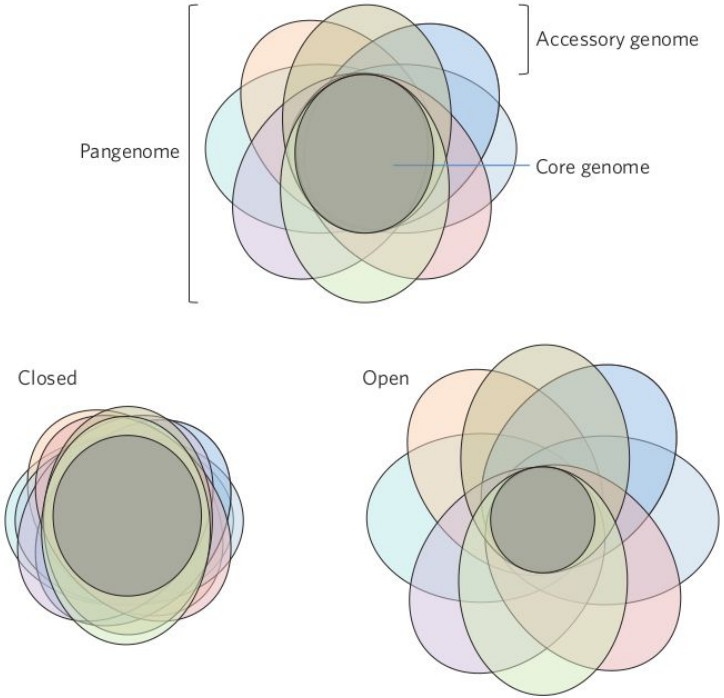
Como construir um Pangenoma?

- Métodos de genômica comparativa são necessários para detectar regiões conservadas e únicas entre um conjunto de genomas.
- Mas quantos genomas são necessários para construir um pangenoma?

Pangenoma Fechado x Aberto

Fechado: tem um tamanho finito, pois após a adição de um certo número de genomas, o pangenoma pode ser totalmente caracterizado.

Aberto: de tamanho ilimitado, pois continua a crescer conforme o número de genomas aumenta.



Estimativa Tamanho Pangenoma

- É um genoma aberto ou fechado?
- Qual o tamanho do pangenoma?
- Quantos genomas preciso considerar?

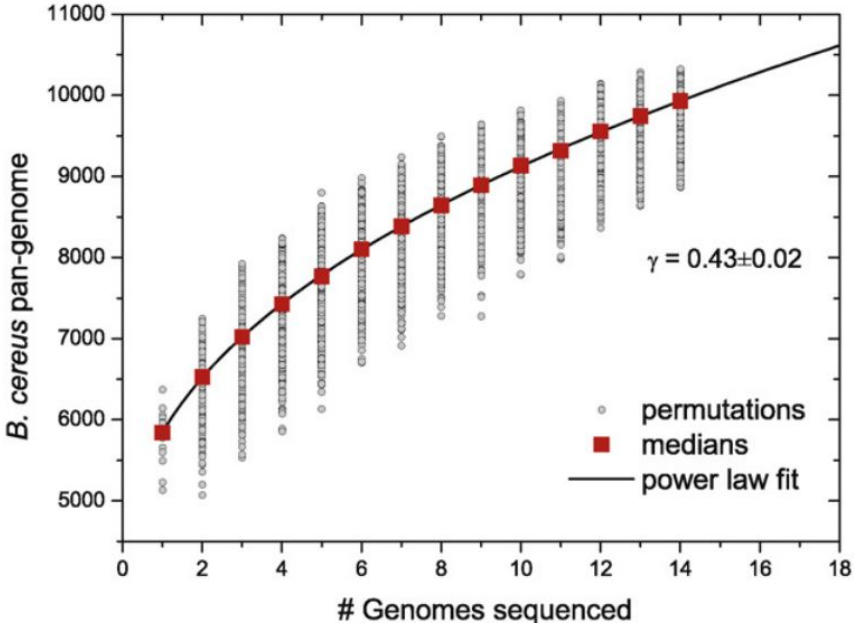
$$n = \kappa N^{\gamma}$$

n → Tamanho Pangenoma

N → Número genomas

Y > 0 indica genoma aberto

Heaps' Law



Tettelin et al (2008)
doi:10.1016/j.mib.2008.09.006

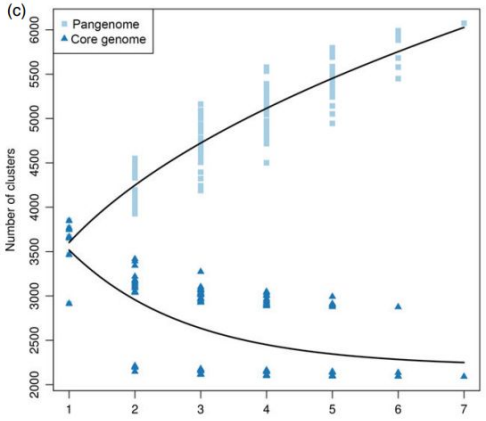
Quais fatores considerar?

- Qualidade dos genomas montados;
- Resolução Filogenética (espécie, gênero,...);
- Tipo e qualidade da anotação (genes, ORFs, CDSs)*;
- Detecção de genes ortólogos e os parâmetros (e-value; % identidade, %cobertura)*;
- Seleção apropriada de amostras.
    - A seleção de um pequeno número de indivíduos intimamente relacionados pode resultar em subestimação do tamanho do pangenoma.

Indivíduos diversos

↓

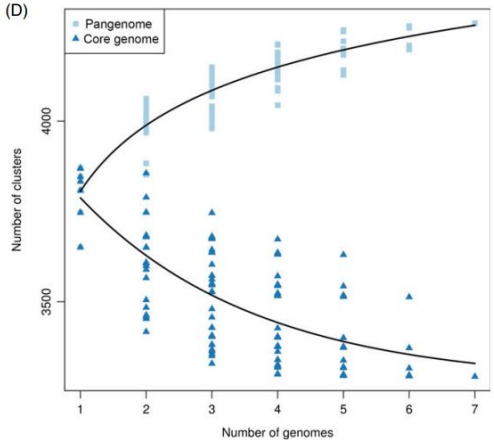
Genoma Aberto



Indivíduos semelhantes

↓

Genoma Fechado

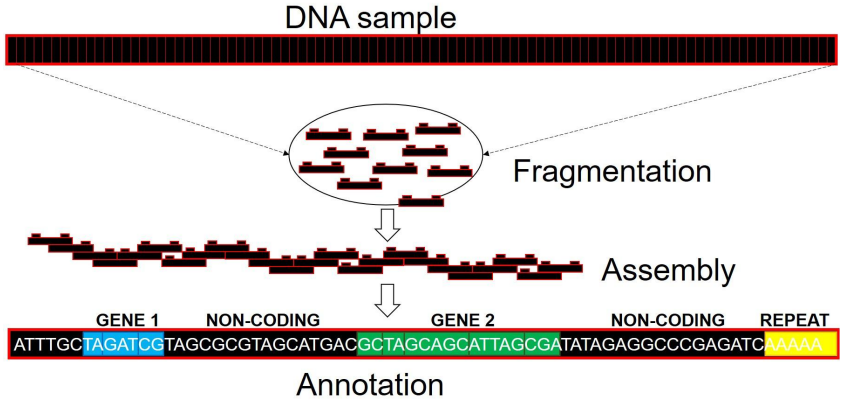


Golicz et al (2016)
doi: 10.1111/pbi.12499

* Se aplica em modelos de pangenoma baseados em genes

Formatos de construção de Pangenoma

- Elemento da montagem
  - Gene: anotação por ortólogos
  - Sequência
    - Alinhamento: Referência; Alinhamento Múltiplo; Grafos
    - Alinhamento Free: Grafo Bruijn

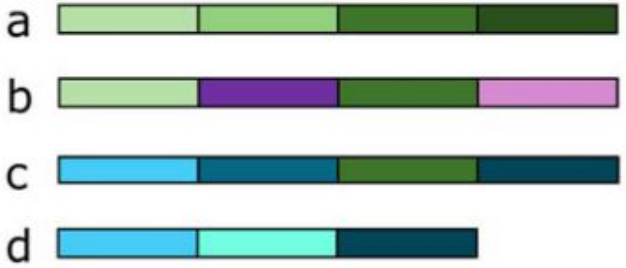


Formatos de construção de Pangenoma

- Processamento
  - Iterativo: um genoma por vez
  - Não iterativo: múltiplos genomas ao mesmo tempo

Iterativo

Alinhamento Múltiplo



Montagem Pangenoma - Gene (Ortólogos)

Componentes	EDGAR*	PGAT*	PGAP	GET-HOMOLOGUS*
Modelo anotação	bidirectional best hits (BBH) utilizando BLAST	BLASTP	InParanoid com BLAST	OrthoMCL, COGtriangles, (BBH) utilizando BLAST
Parâmetro cutoff ortólogos	BLAST Score Ratio (BSR)	%Identidade e %Cobertura	%Identidade e %Cobertura	%Identidade e %Cobertura
Modelo clusterização famílias genes		BLASTP	BLASTALL com Markov Cluster (MCL)	Pfam database para domínios de proteínas

Zekic et al (2016)
https://doi.org/10.1007/978-1-4939-7463-4_2

*Modelos desenvolvidos para Procariotos

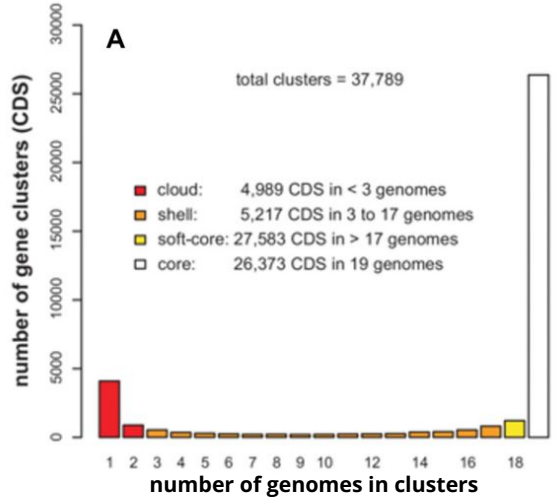
Montagem Pangenoma - Sequência

Tipo	Referência	Alinhamento	Grafo	Grafo Bruijn
Característica	→ Indexa variantes nos genomas em relação referência; → Match exato ou por similaridade → Recomendado genomas parecidos	→ Múltiplas sequências são alinhadas; → Identificação dos segmentos cores e identificação variantes cada input.	→ Alinhamento não por uma referência e sim totalmente indexado com referência multi-genoma.	→ Sem alinhamento, utiliza K-mers, e sem referência.
Exemplos	RCSI; MuGI; Journaled String Tree (JST); BWBBLE; PBWT	Panseq; Harvest	GenomeMapper; GCSA; PanCake	SplitMEM; TwoPaCo;BCALM; cdBGs; Cortex; Vari

Zekic et al (2016)
https://doi.org/10.1007/978-1-4939-7463-4_2

Módulos das ferramentas

- Clusterização genes homólogos
- Estimativa tamanho pangenoma
- Perfil pangenoma: core e genes acessórios
- Identificação SNPs
- Matriz Pangenoma
- ANI e Árvore filogenética
- Anotação e busca de funções



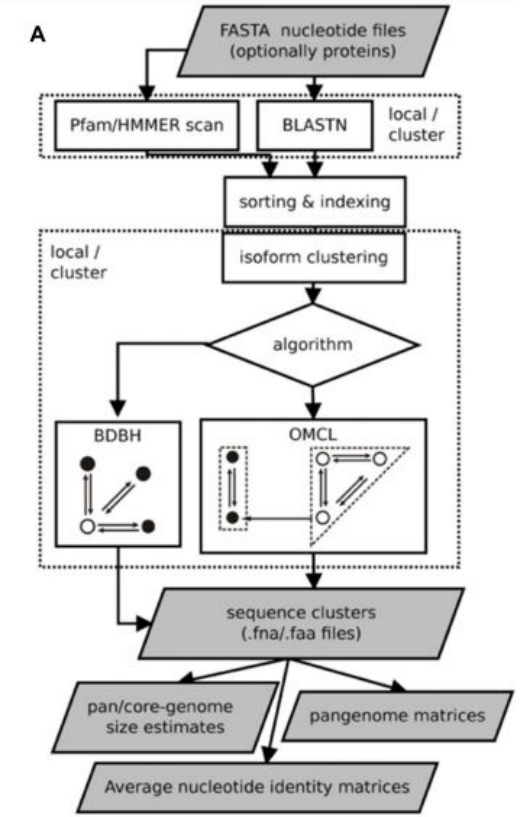
Aplicações Pangenoma Procariontes

- Organismos patogênicos:
  - Em Pseudomonas aeruginosa, a alta resistência aos tratamentos tentados pode estar associada a maior variabilidade genômica entre as espécies.
  - Para Acinetobacter baumannii foram detectadas regiões conservadas em nível de proteína que podem ser estudadas como potenciais alvos de vacinas.

E Pangenoma Eucariontes?

Estudo de caso: **Analysis of Plant Pan-Genomes and Transcriptomes with GET_HOMOLOGUES-EST, a Clustering Solution for Sequences of the Same Species**

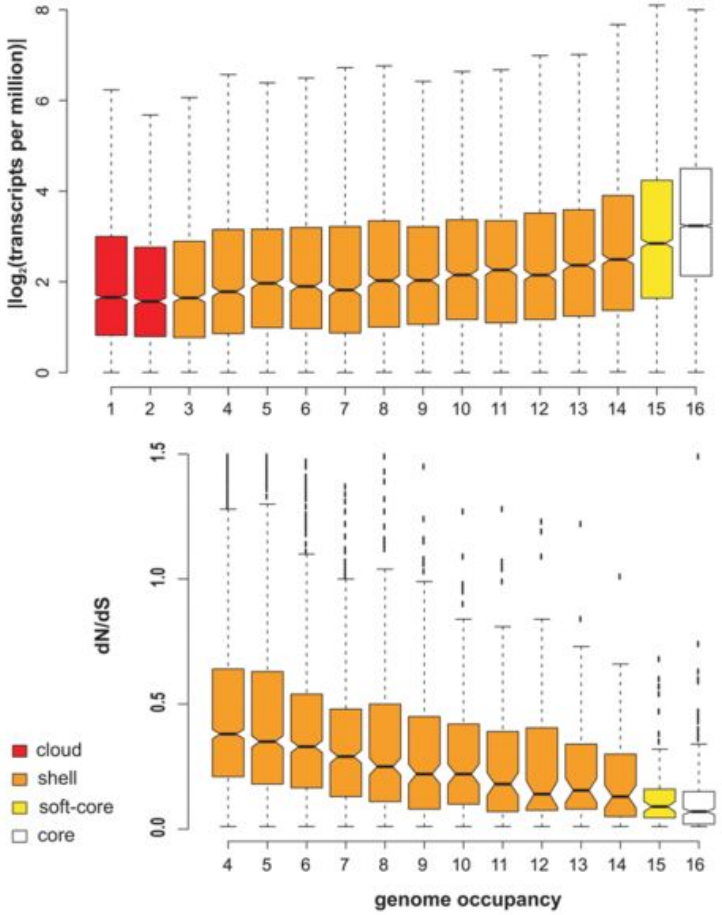
- Objetivo: Comparar a construção do pangenoma com genoma e com transcriptoma (pan-transcriptomas)
- Planta referência: Arabidopsis thaliana
- Resultado: Pangenoma 10% menor utilizando transcriptoma.



E Pangenoma Eucariontes?

Estudo de caso: **Analysis of Plant Pan-Genomes and Transcriptomes with GET_HOMOLOGUES-EST, a Clustering Solution for Sequences of the Same Species**

- Pan-transcriptoma com cevada
  - Genes do core têm maior expressão
  - Genes acessórios têm maiores taxas de substituição não sinônima, menos conservados



Pangenoma além das CDRs

- **Transposons (TEs)**
  - A expressão genética é afetada não apenas por regiões codificadoras, mas também por regiões não codificadoras, como por elementos Cis-regulatory;
  - TEs agem como elementos Cis-regulatory, pois podem influenciar a expressão de genes ao se inserirem em diferentes locais do genoma.

Pangenoma além das CDRs

- **Variações estruturais (SVs)**

Pequenas:

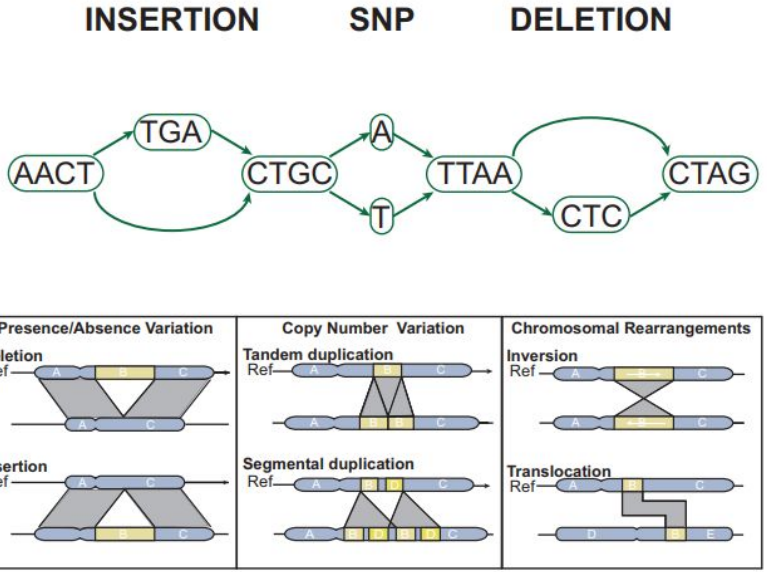
→ Single nucleotide polymorphisms (SNPs)

→ InDel

Grandes:

→ Presence and Absence Variations (PAVs)

→ Copy Number Variants (CNVs)



Pangenoma em Grafos Eucariontes

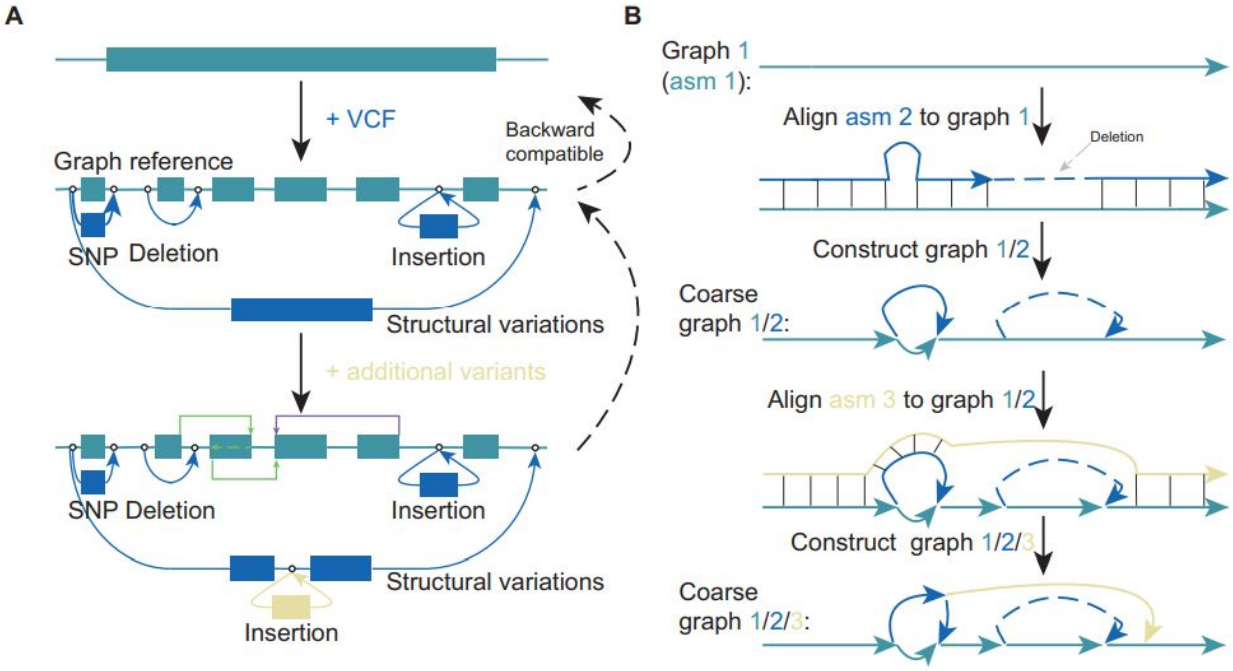
- **Dois tipos de construção:**
  - Genoma de referência e informação de variação: Seven Bridges; vg, ...
  - Alinhamento: Minigraph; MGR;...

Tools	Methods	Variation type	Computational speed
vg	vcf + ref	SNP+InDel+SV	Fast
SevenBridges	vcf + ref	SNP+InDel+SV	Intermediate
Minigraph	Alignment-based	InDel+SV	Fast
MGR	Alignment-based	SNP+InDel+SV	Low
Seqwish	Alignment-based	SNP+InDel+SV	Fast
NovoGraph	Alignment-based	SNP+InDel+SV	Low
PGGB	Alignment-based	SNP+InDel+SV	Low
Cactus	Alignment-based	SNP+InDel+SV	Fast

Pangenoma em Grafos Eucariontes

Seven Bridges and vg

Minigraph



Pangenoma em Grafos Eucariontes

- **Armazenamento:**

Tipo de dado:

S: nó (sequência)

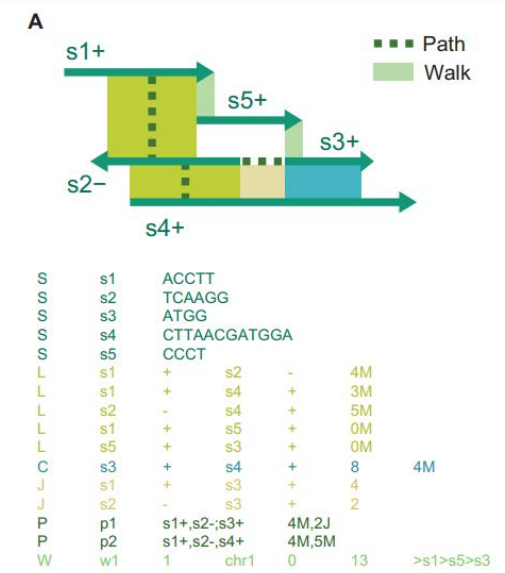
L: aresta que conecta os nós

J: salto, segmento sem associação

C: contenção, um segmento contido em outro

P: caminho do genoma de referência e do haplótipo

W: passeio orientado no grafo



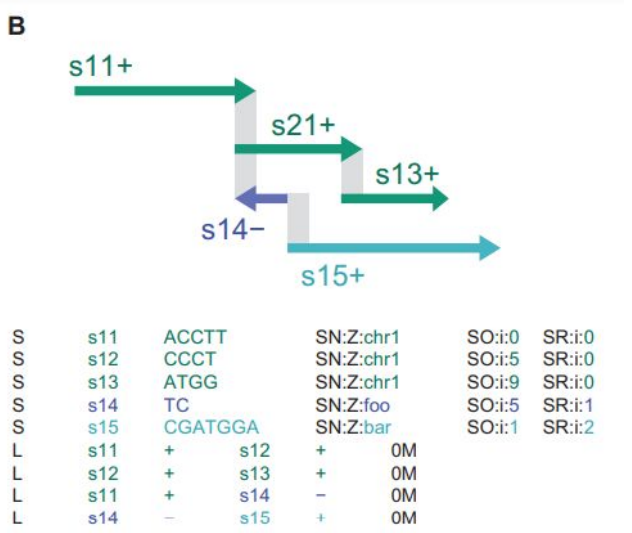
Graphical fragment assembly (GFA)

Coordenada:

SN: Número cromossomo

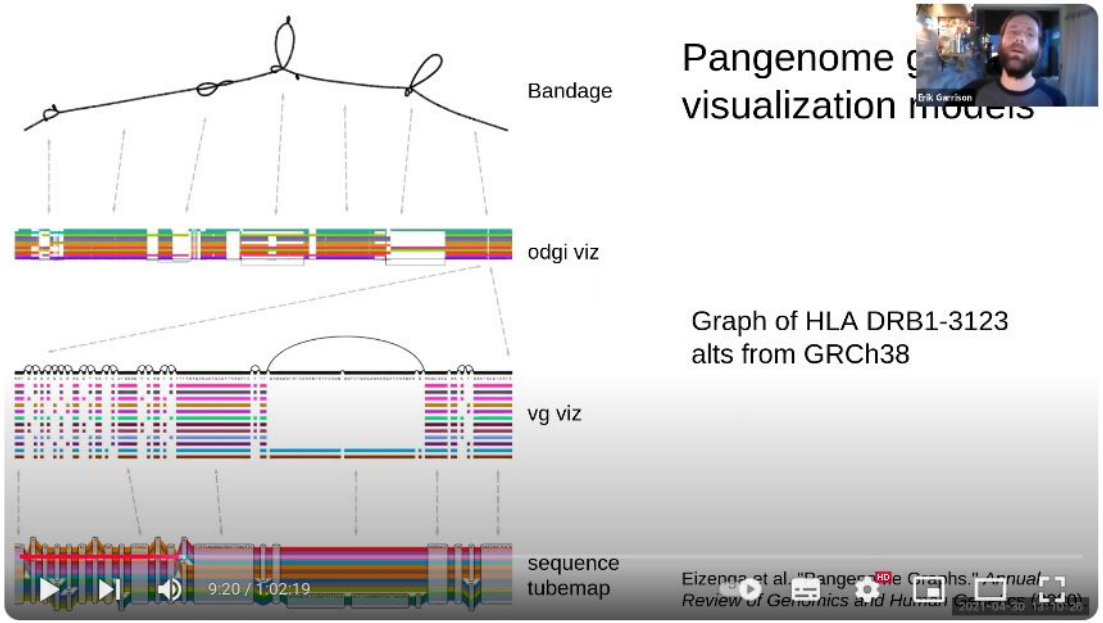
SR: Coordenada do nó

SO: Fonte do nó



Reference GFA (rGFA)

Visualização Pangenoma em Grafos



Building pangenome graphs:
https://www.youtube.com/watch?v=PGLg4n1UhKc

Aplicações Pangenoma Plantas

- Genes “dispensáveis” desempenham papéis importantes na evolução e na interação entre as plantas e o ambiente.
- Associação de dados fenotípicos possibilita identificar variações que controlam o fenótipo de uma espécie, o que pode ser utilizado em culturas agrícolas.
    - Pode ser mais eficiente utilizar SVs do que SNPs para a avaliação de fenótipos.
- Processos biológicos afetados por CNVs/PAVs: produção de metabólitos, tempo de floração, tolerância à submersão, absorção de fósforo, resposta ao estresse biótico
- A identificação de PAVs pode estar associada com regiões HOT, que são mais propícias às mutações.

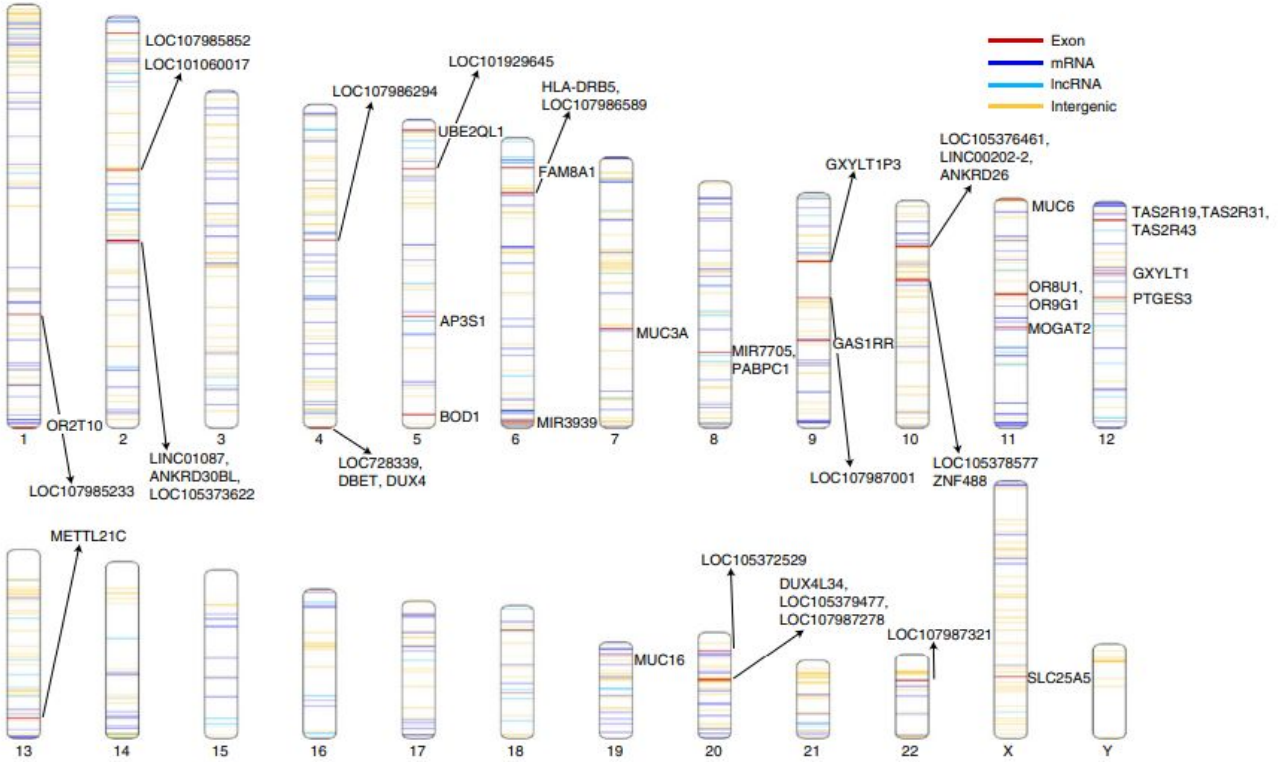
Pangenoma Humano

Estudo de caso: **Assembly of a pan-genome from deep sequencing of 910 humans of African descent**

- Objetivo: Comparar o pangenoma africano com o genoma humano de referência;
- Amostras genomas: 910 indivíduos descendentes africanos;
- Resultados:
  - O pangenoma africano contém ~10% mais DNA do que o genoma de referência;
  - 387 dos novos contigs se enquadram em 315 genes de codificação de proteínas, e o restante parece ser intergênico.

Pangenoma Humano

Mapa do genoma humano mostrando as localizações de todos os contigs do pangenoma africano.



Referências Bibliográficas

GUIMARÃES, Amanda Munari. Abordagens bioinformáticas no estudo pangenômico de *Xanthomonas campestris*. 2020. Dissertação de Mestrado. Universidade Federal de Pelotas.

LIU, Dang; HUNT, Martin; TSAI, Isheng J. Inferring synteny between genome assemblies: a systematic evaluation. BMC bioinformatics, v. 19, p. 1-13, 2018.

MADE, Bantayehu Bekele; BEYENE, Dereje. The Role of Host Genetics in the Immune Response to Sars Cov-2 and covid-19 Susceptibility and Severity. Clinical Research and Studies, v. 1, n. 2, p. 2835-2882, 2022.

MENDES, Rodrigo; GARBEVA, Paolina; RAAIJMAKERS, Jos M. The rhizosphere microbiome: significance of plant beneficial, plant pathogenic, and human pathogenic microorganisms. FEMS microbiology reviews, v. 37, n. 5, p. 634-663, 2013.

MOREIRA, Leandro Márcio; PROSDÓCIMO, Francisco. Genômica comparativa. In: MOREIRA, Leandro Márcio (org.). Ciências genômicas: fundamentos e aplicações. Ribeirão Preto: Sociedade Brasileira de Genética, 2015. cap. 4, p. 81-99.

TETTELIN, Hervé et al. Genome analysis of multiple pathogenic isolates of *Streptococcus agalactiae*: implications for the microbial "pan-genome". Proceedings of the National Academy of Sciences, v. 102, n. 39, p. 13950-13955, 2005.

Obrigado pela atenção!
Dúvidas?