

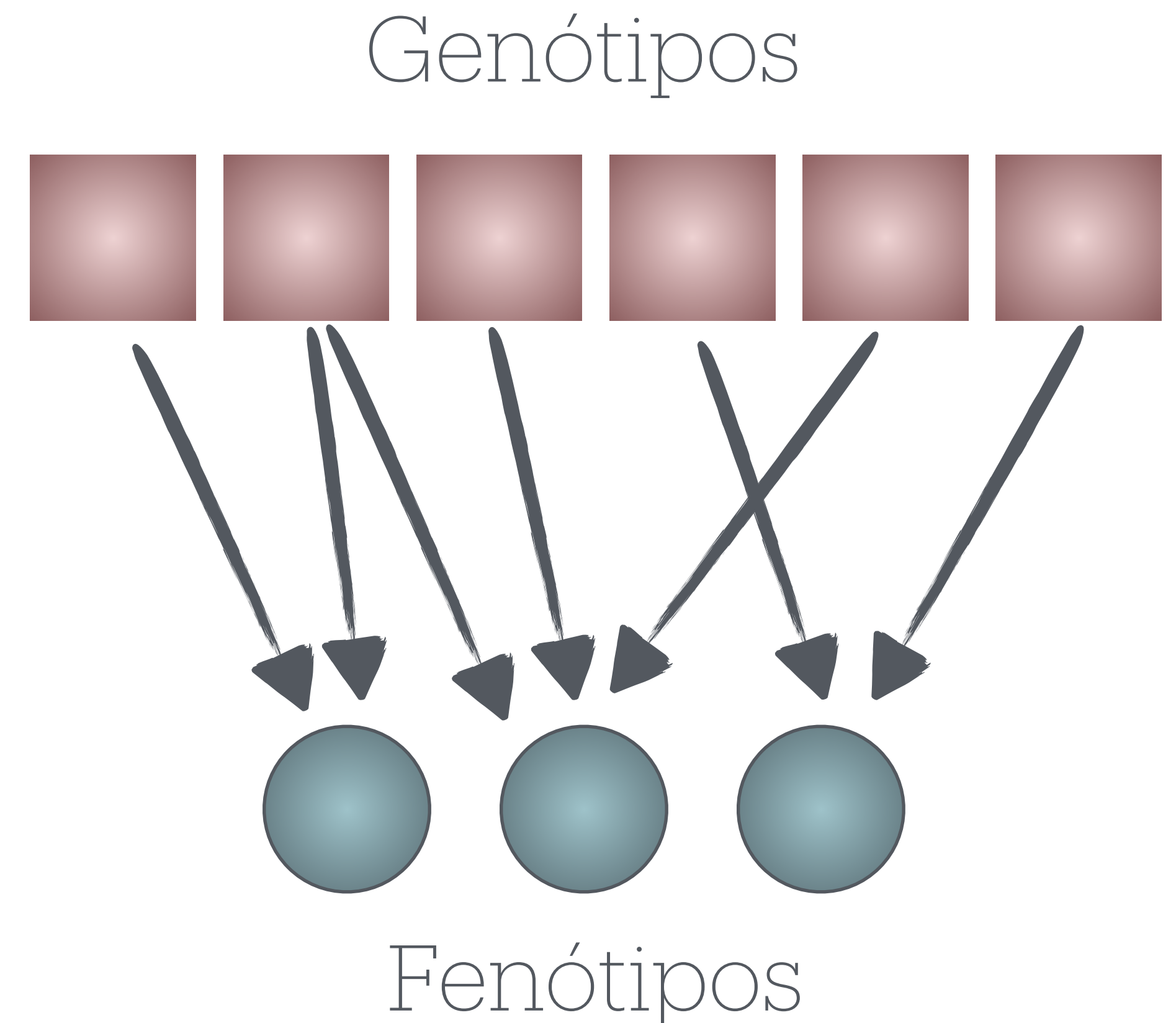
Fenótipos Moleculares e Mapeamento Genético

Desvendando a arquitetura genética de caracteres complexos

Arquitetura genética

O mapa genótipo-fenótipo

- Como os organismos expressam a informação codificada no genoma?
- Como os fenótipos são formados a partir dessa informação?
- Quais variantes genéticas são responsáveis pela variação fenotípica que observamos em populações naturais?
- Quanto da variação é genética e quanto da variação é ambiental?



- Peso, altura, comprimentos, ...
- Probabilidade de desenvolver doença, severidade de doença, ...
-

O que é uma variante genética?

Alelos, SNPs, e outros

- Locus = Uma posição no genoma
- Alelo = diferenças na sequência genética em determinada posição
 - Exemplos: Diferenças de uma única base (SNPs), Indels (inserções e deleções), variação no número de genes (CNVs), ...
- Como essas variantes levam a diferenças fenotípicas?

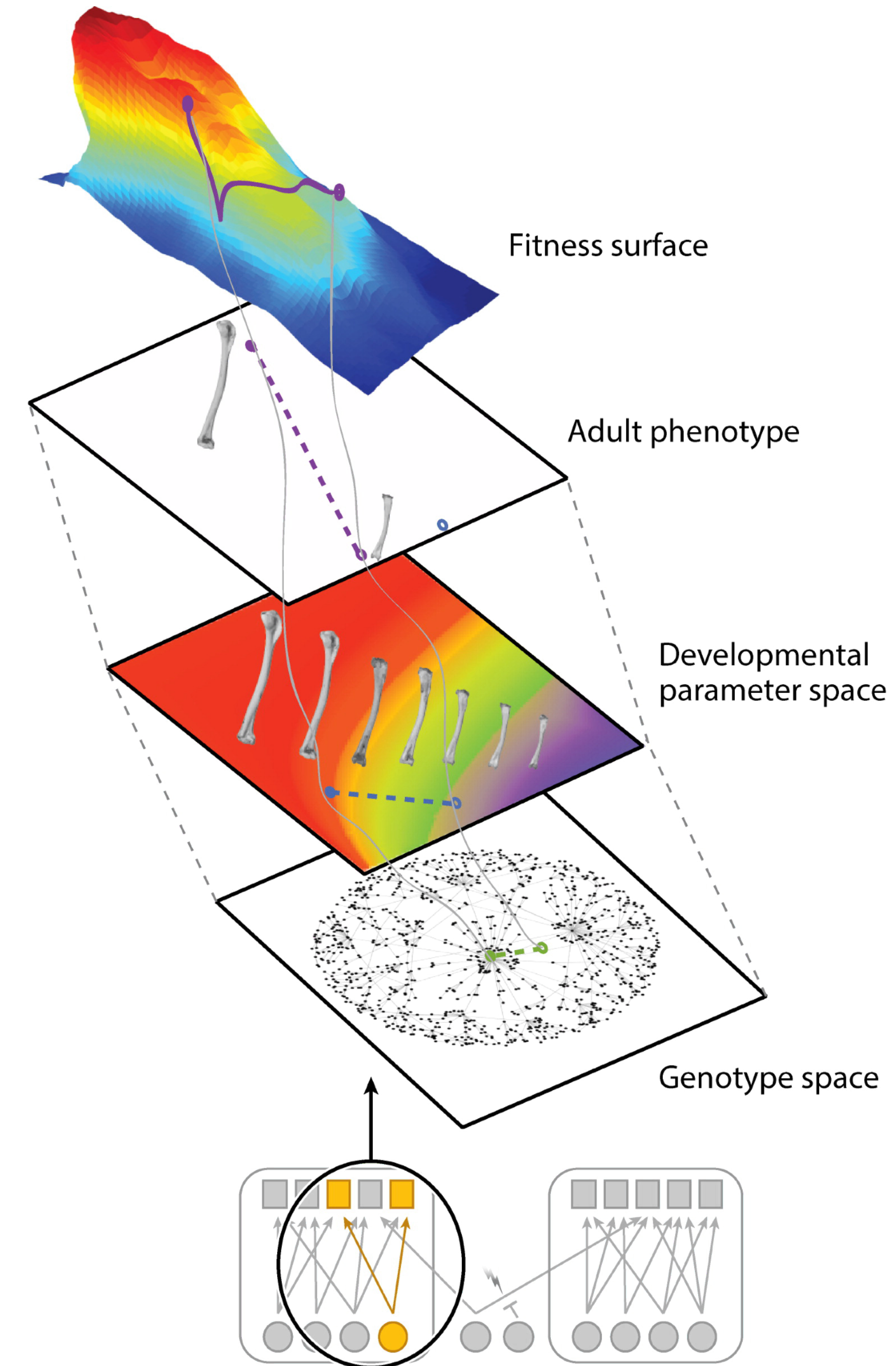
ATGGTAA^GCCTGAG^CTGACTTAGCGT-AT
ATGGTAA^ACCTGAG^TTGACTTAGCGTCAT

↑ ↑ ↑
SNP SNP indel

Desenvolvimento e ontogenia

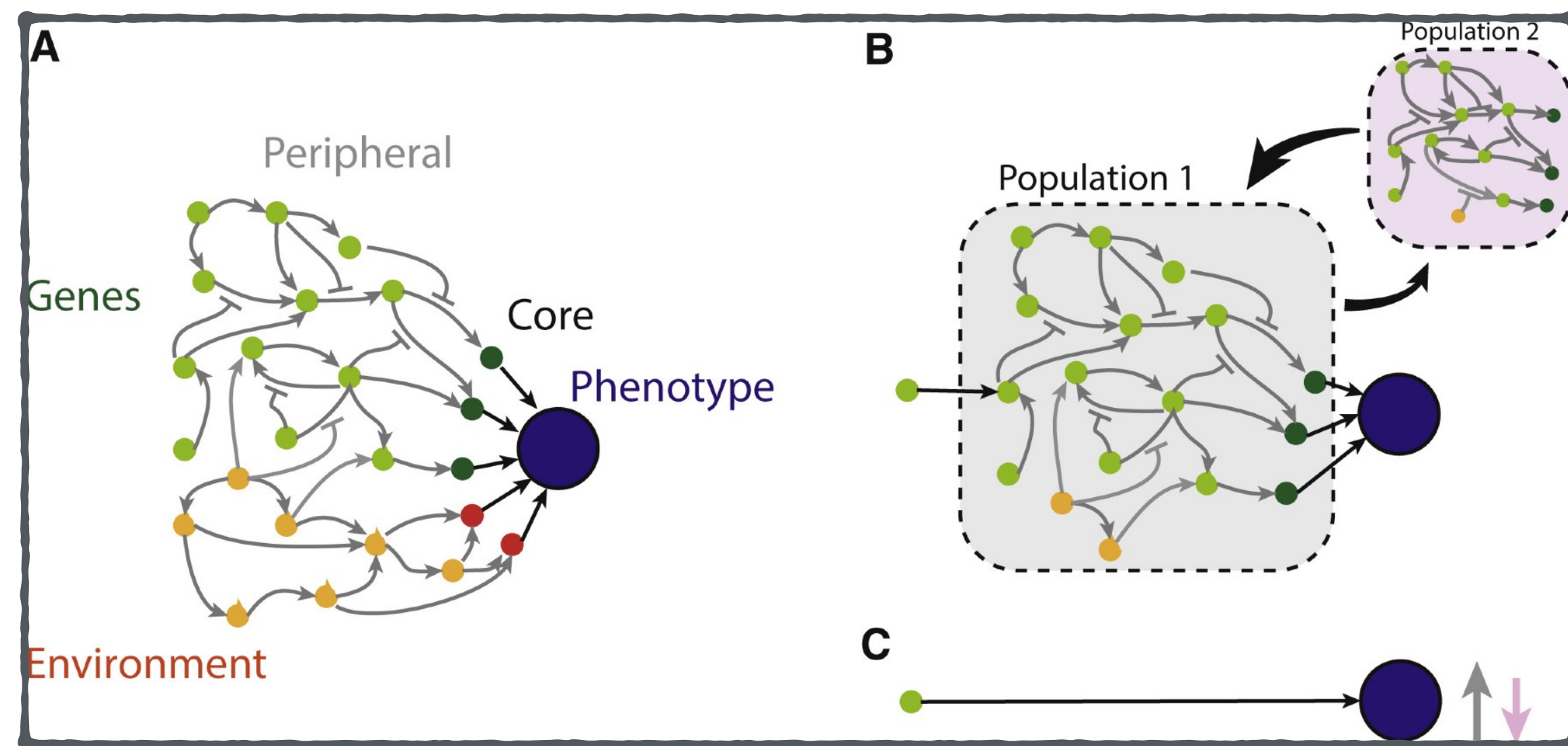
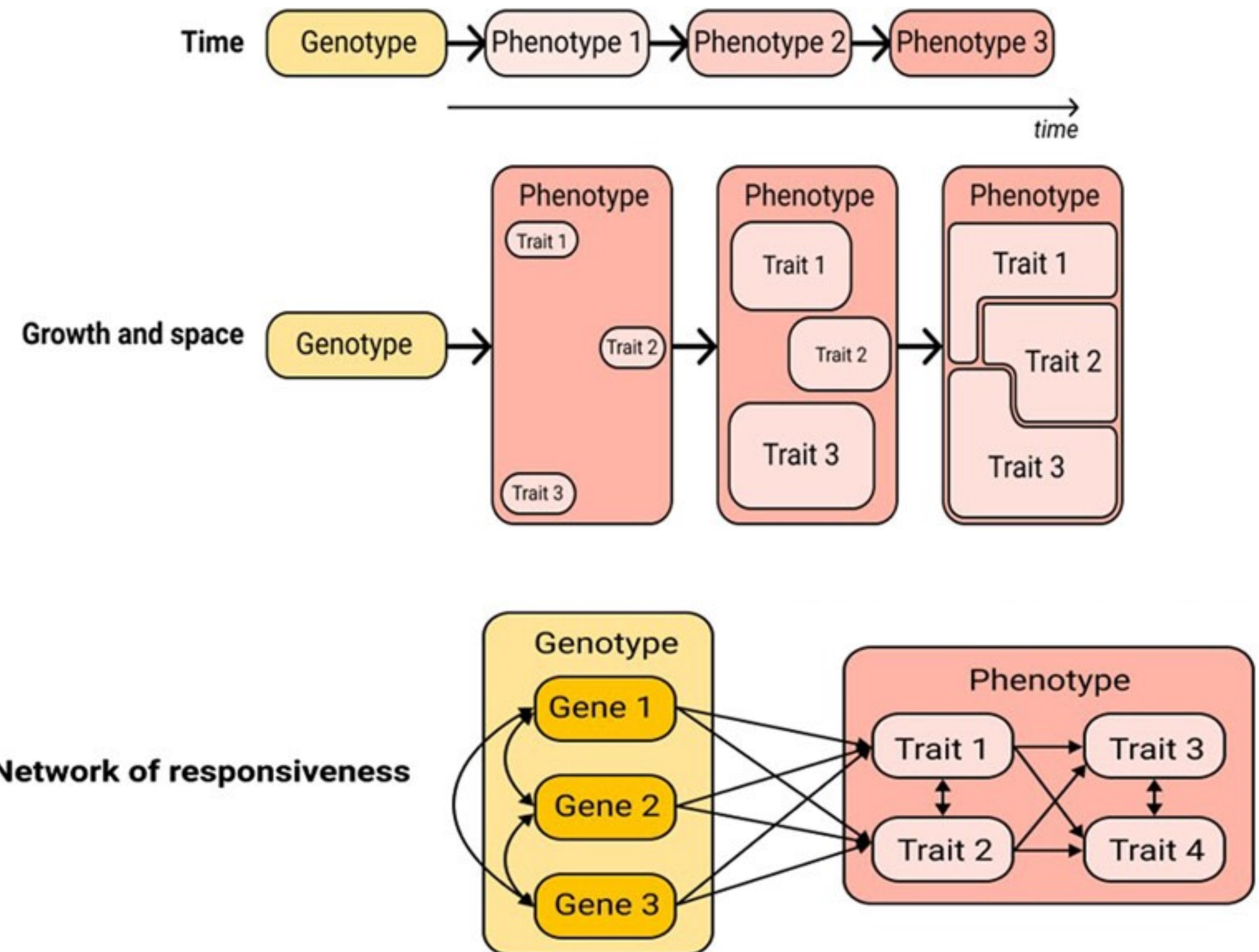
O processo que leva a manifestação do fenótipo

- Todos os fenótipos dos organismos são formados por algum processo de desenvolvimento
- Esse processo leva do genótipo, da sequencia genética, passando por intermediários e processos moleculares, até a manifestação do fenótipo que interage com o ambiente e as forças evolutivas



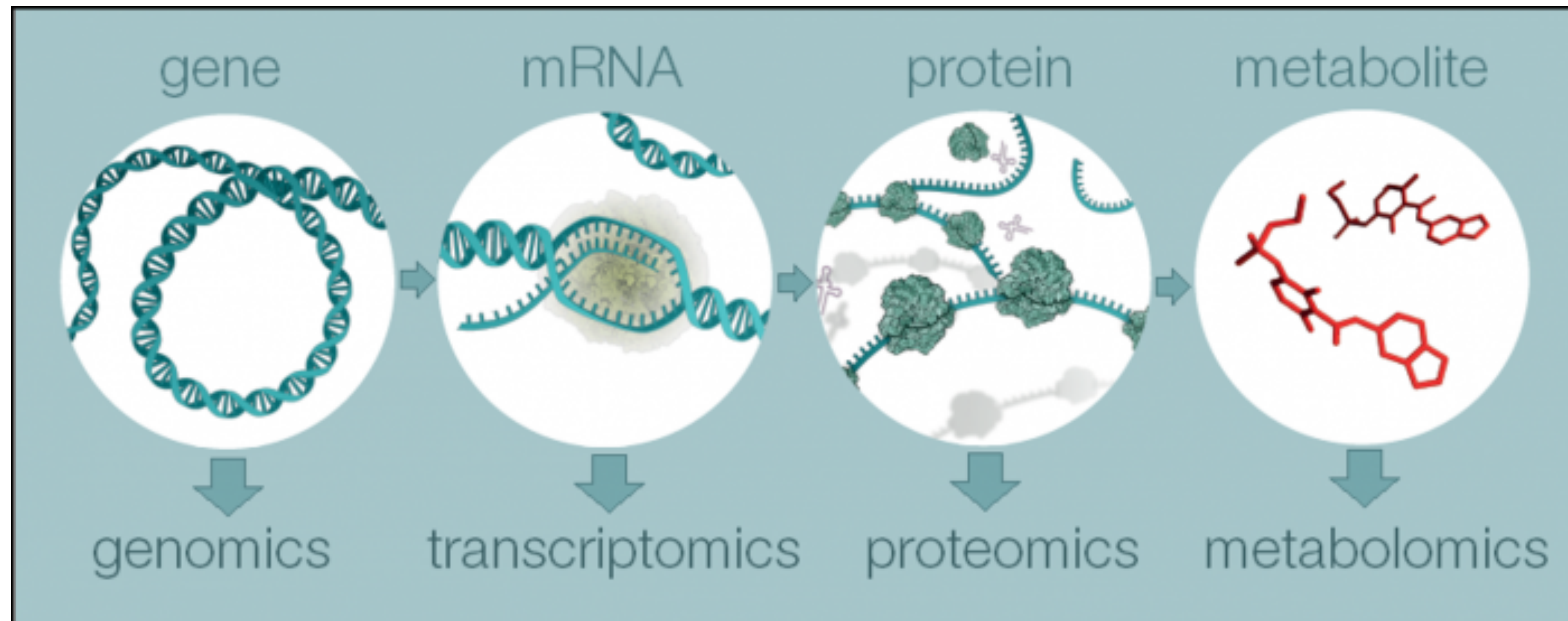
Redes complexas de interações

Redes de interação genéticas e fenotípicas



Níveis de organização

Do genótipo ao fenótipo



Sequenciamento genético

Genomas completos, genótipos e variantes estruturais

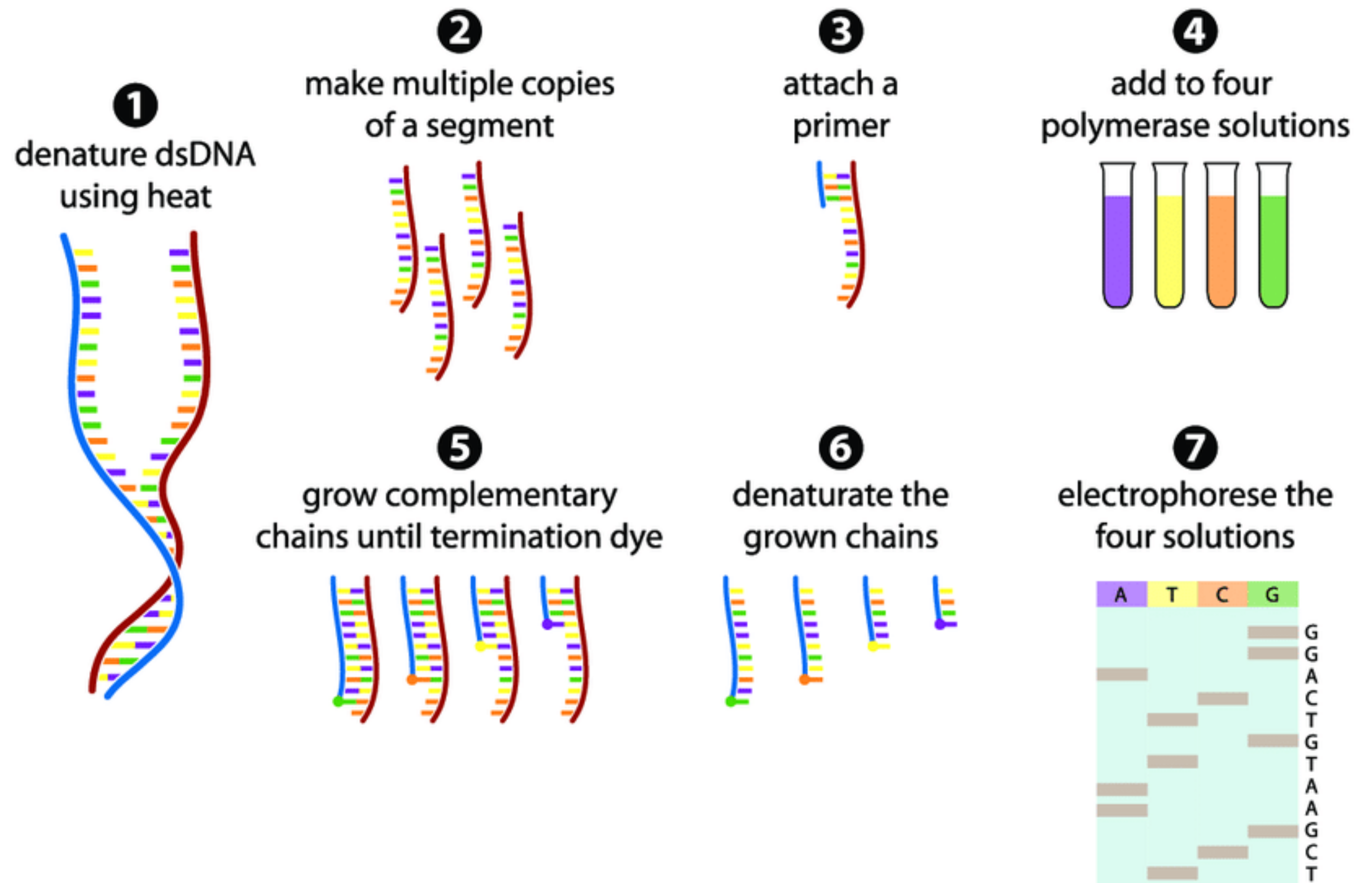
Tipos de variantes genéticas

- Diferenças de sequência
- Variantes estruturais:
 - Inserção-deleção (indel)
 - Inversões
 - Duplicações
 - Translocações

Single nucleotide variant	ATTGGCCTTAACCCCGGATTATCAGGAT ATTGGCCTTAACCCCGGATTATCAGGAT	Structural variants
Insertion–deletion variant	ATTGGCCTTAACCCGATCCGATTATCAGGAT ATTGGCCTTAACCC---CCGATTATCAGGAT	
Block substitution	ATTGGCCTTAACCCCCGATTATCAGGAT ATTGGCCTTAACAGTGGATTATCAGGAT	
Inversion variant	ATTGGCCTTAACCCCCGATTATCAGGAT ATTGGCCTTCGGGGGTTATTATCAGGAT	
Copy number variant	ATTGGCCTTAGGCCTTAACCCCCGATTATCAGGAT ATTGGCCTTA-----ACCTCCGATTATCAGGAT	

Sequenciamento Sanger

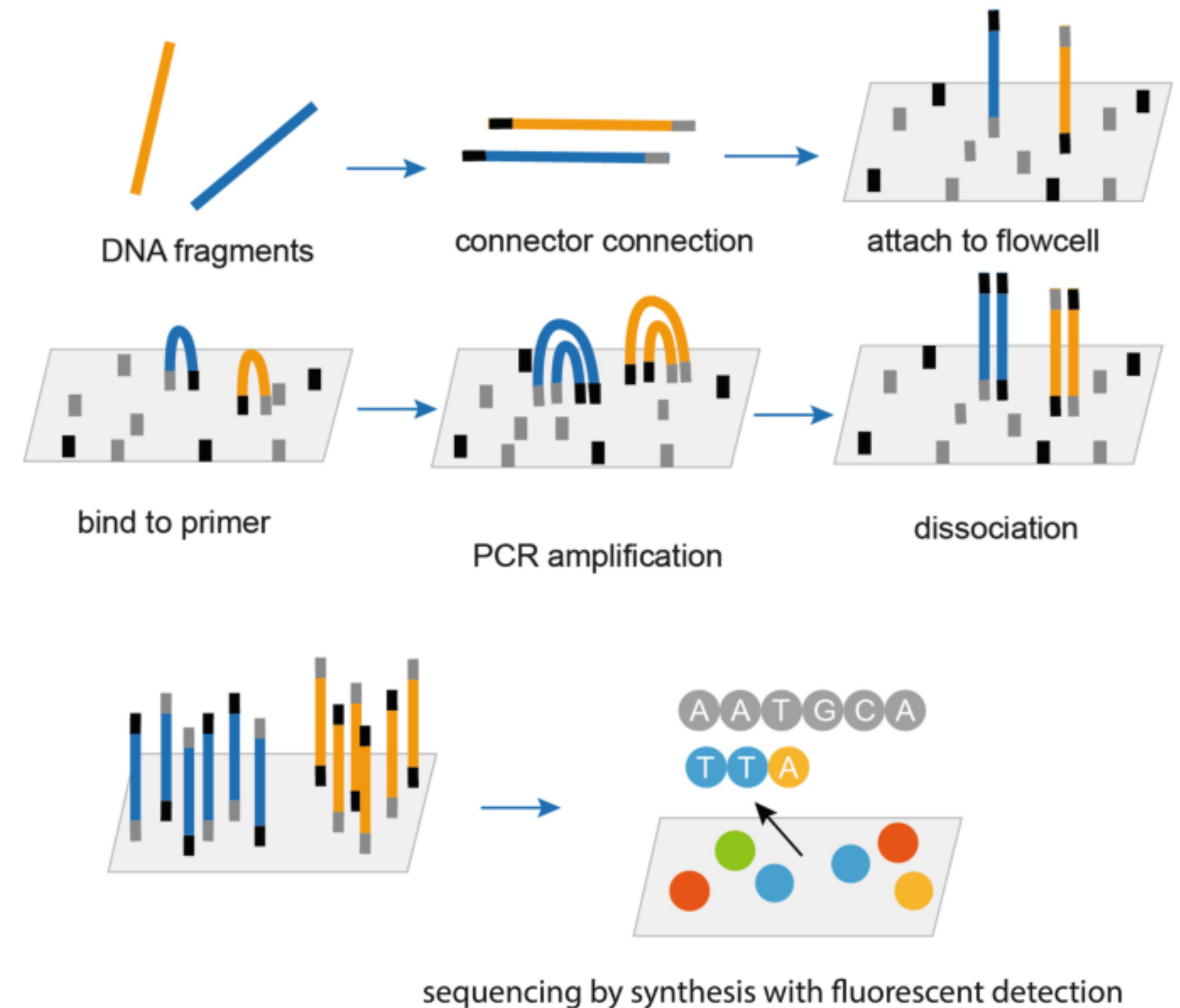
- Sequencias amplificadas são duplicadas com bases especiais que ocasionalmente interrompem a duplicação
- Essas bases especiais podem ser lidas por fluorescência
- Utilizando uma coluna que segrega as sequencias por tamanho, podemos ler a sequencia da molécula original



Sequenciamento de “nova geração”

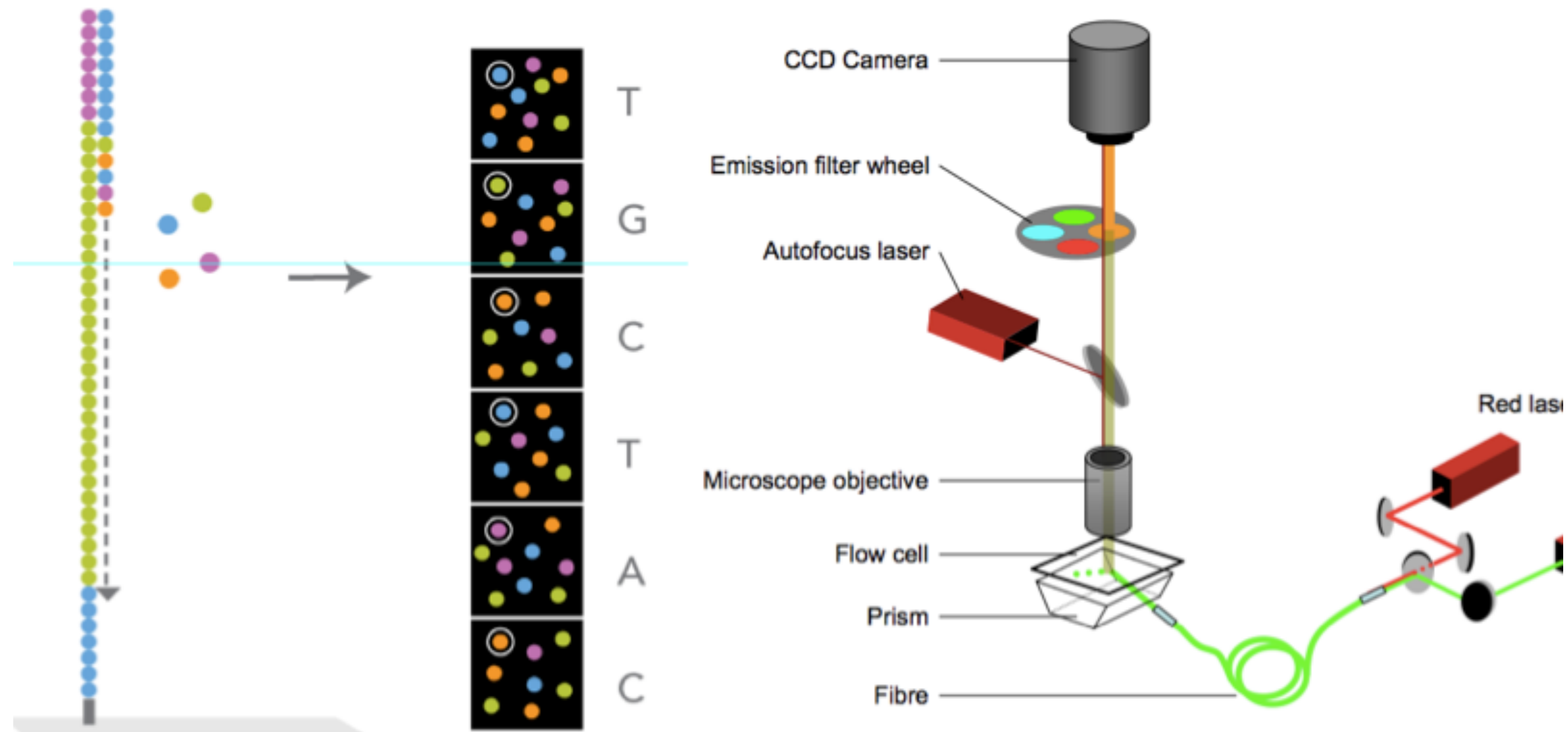
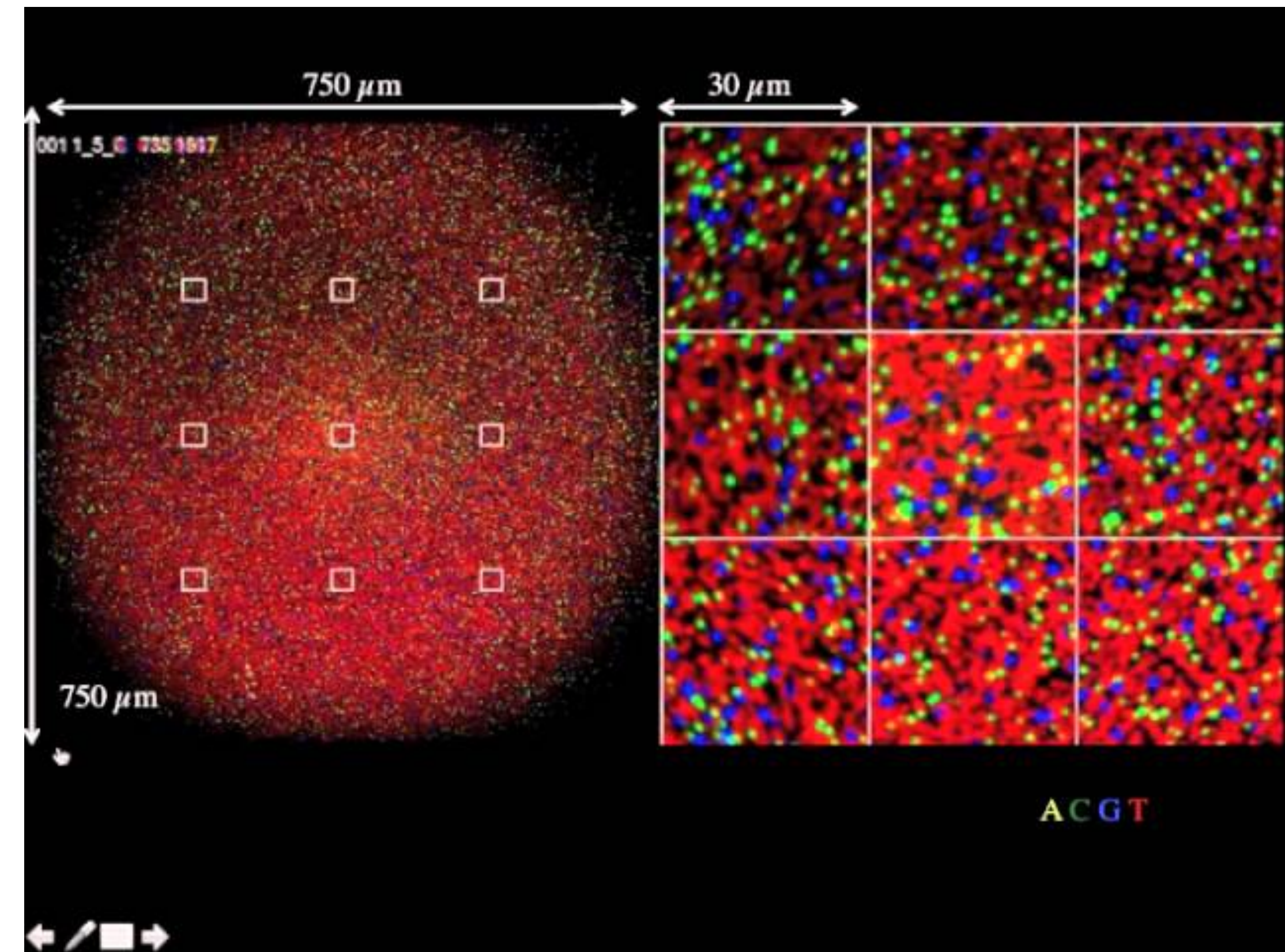
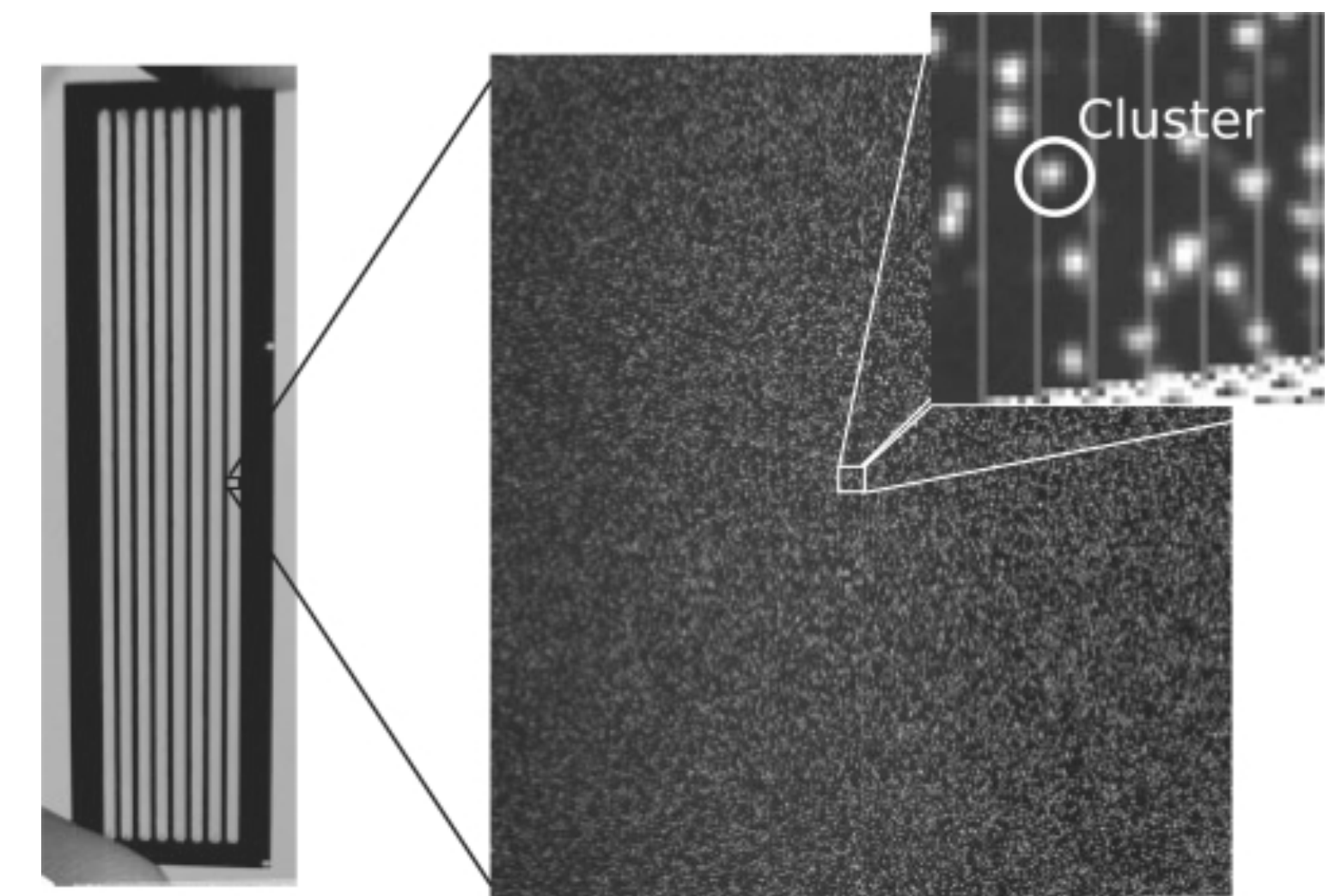
NGS sequencing

- Mesmo princípio do método Sanger, mas com alguma forma de paralelismo
- A maioria dos métodos usa uma estratégia de cortar o genoma em várias moléculas pequenas e sequenciar todas ao mesmo tempo
- Essas sequências pequenas são “montadas” usando um genoma de referência



Illumina flowcell

NGS sequencing

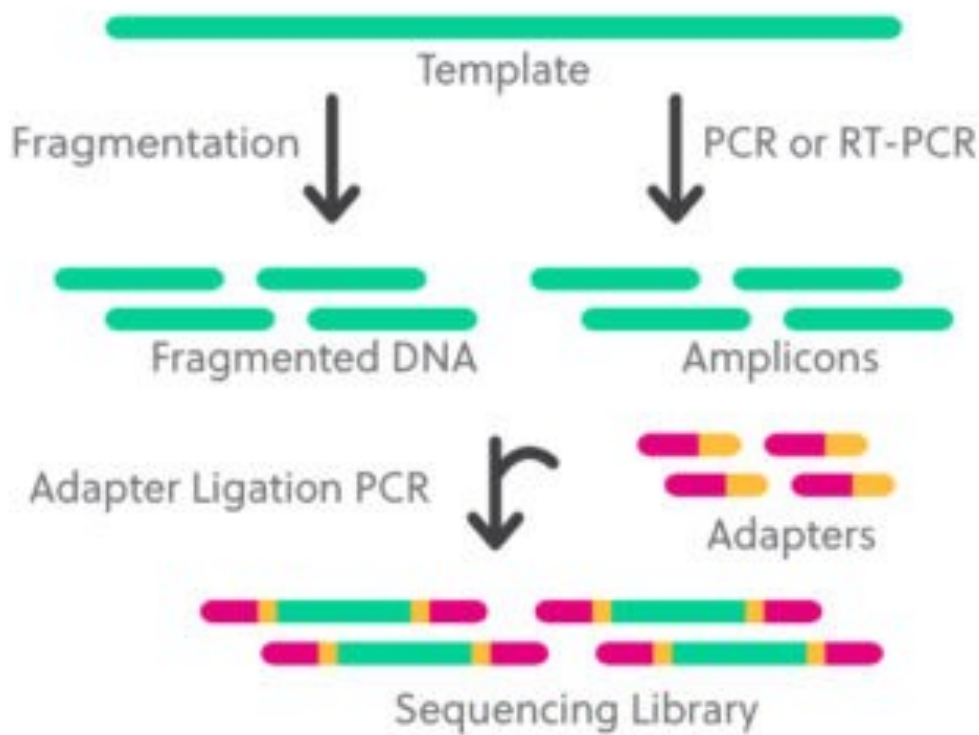


Resultado do sequenciamento

STEP 1: Extraction



STEP 2: Library Prep



STEP 3: Sequencing



STEP 4: Analysis



Arquivos fastq

```
@SIM:1:FCX:1:15:6329:1045 1:N:0:2
TCGCACTCAACGCCCTGCATATGACAAGACAGAATC
+
<>;##=><9=AAAAAAAAAAAA9#:<#<;<<<????#<#<=
```

Alinhamento de sequencias

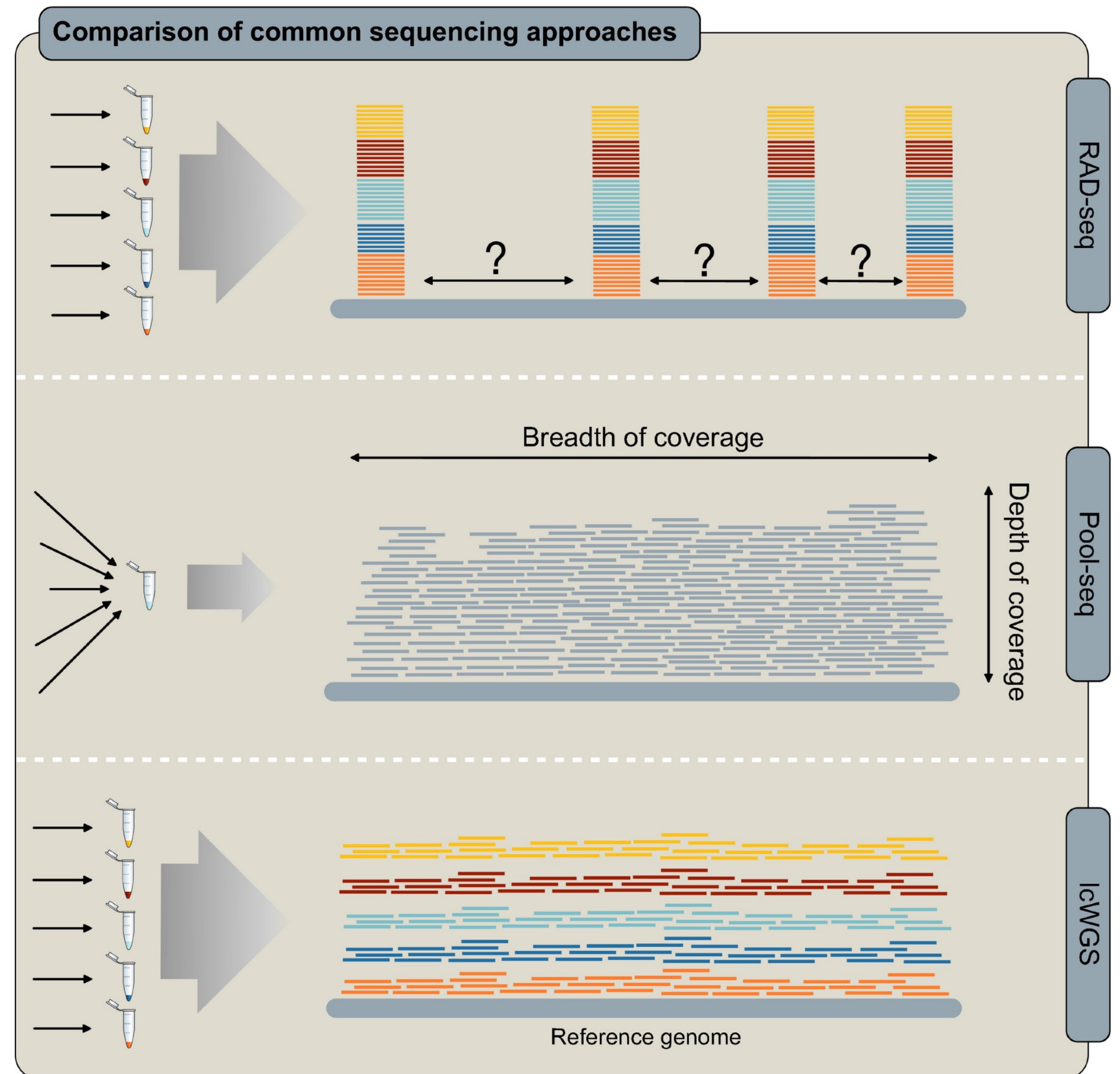
Com genoma de referência



Alguns métodos diferentes de sequenciamento

Nem sempre sobra dinheiro

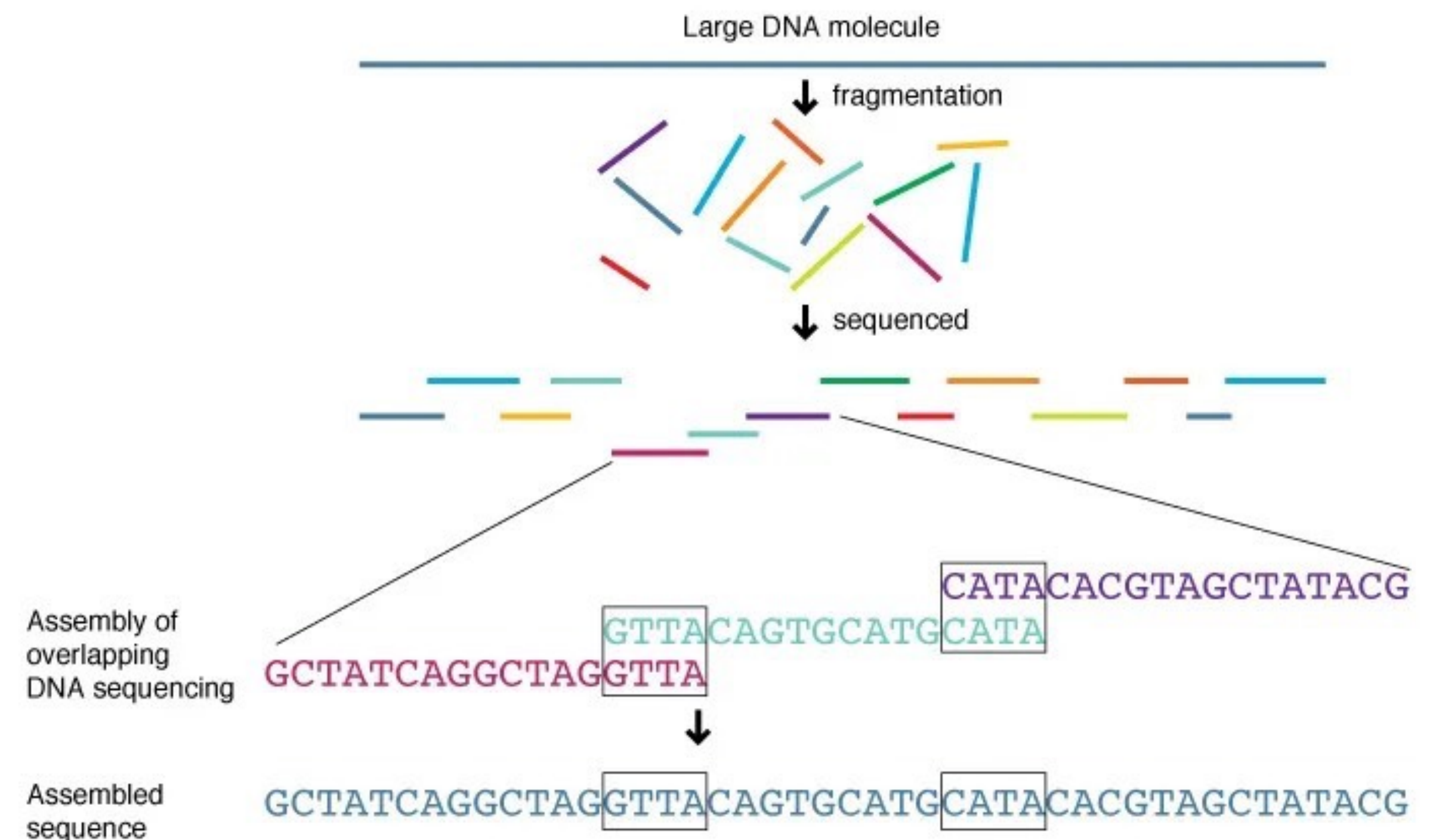
- Nem sempre é factível sequenciar um genoma completo
- As vezes, estamos interessados em outros aspectos da variação genética, como variantes específicas ou só nas frequências alélicas
 - Métodos como Pool-seq ou Rad-seq podem ter um custo/benefício melhor nesses casos
- Outra opção é sequenciar com cobertura baixa



Sequenciamento *de-novo*

Gerando novos genomas

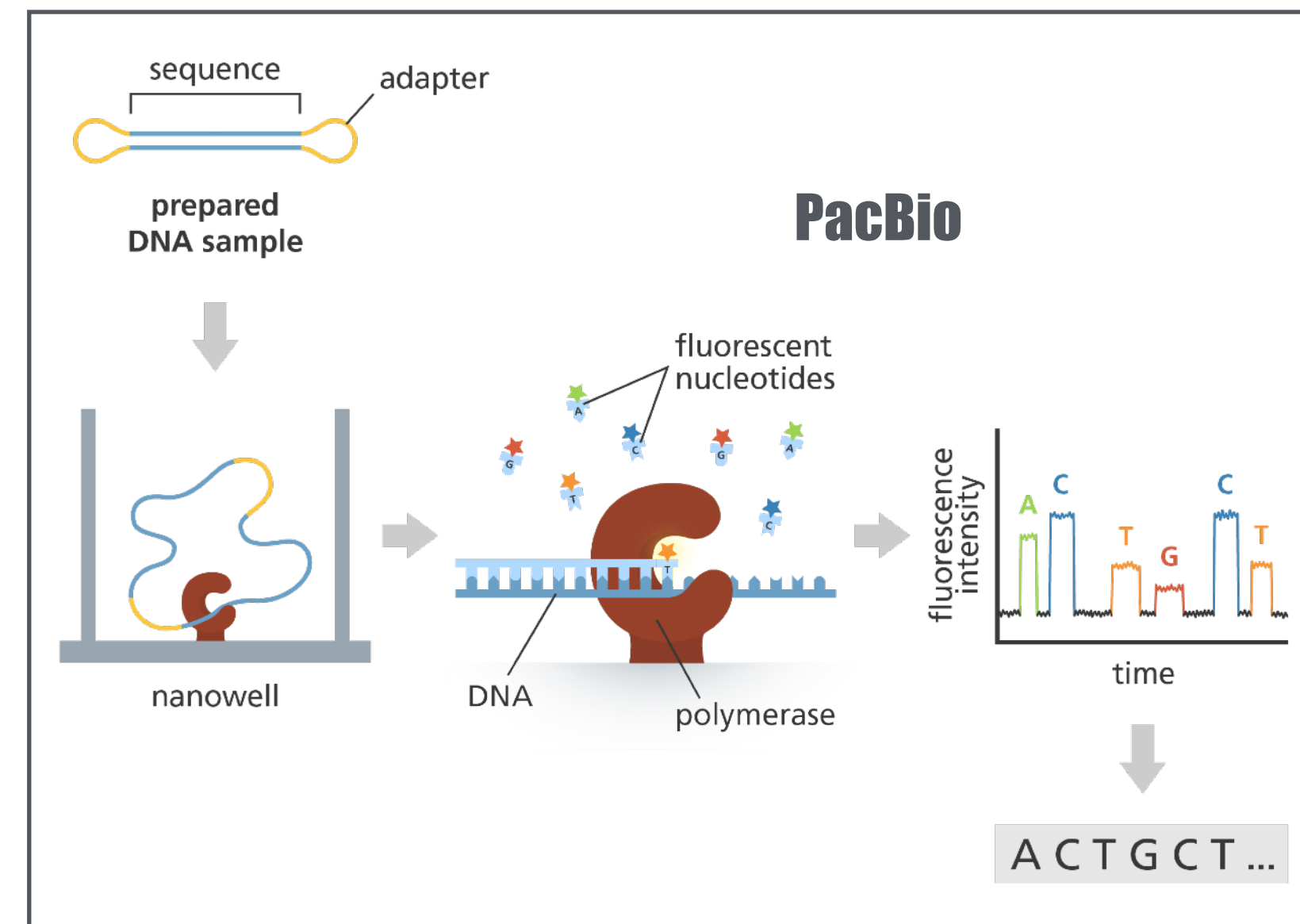
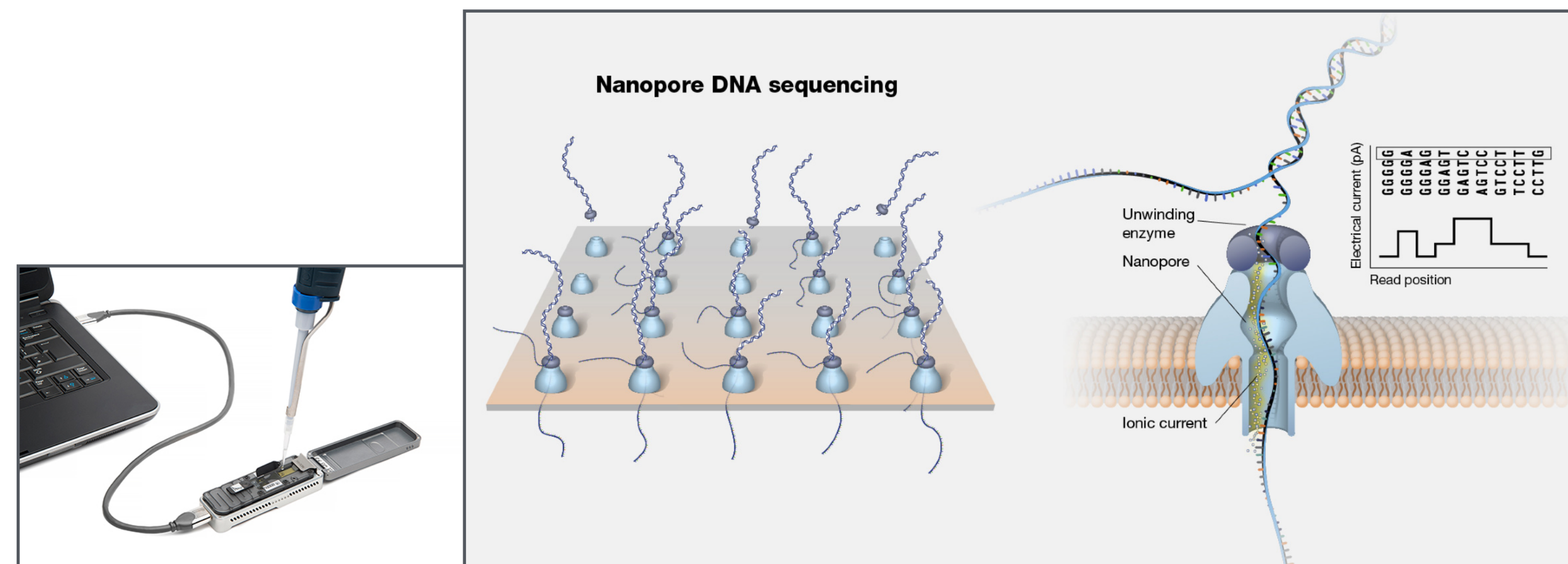
- Sem um genoma de referência, podemos utilizar somente os reads para montar o genoma *de-novo*
- Os reads são alinhados entre si, e montados em sequências contínuas
- Nem sempre é possível sequenciar o genoma completo com essa técnica, principalmente por conta de regiões repetitivas
- Pode ser combinado com sequenciamento “long-reads”, como Nanopore



Sequenciamento Long-read

Oxford Nanopore e PacBio

- Permitem a obtenção de sequencias longas mas com erro alto (10-15%)
- Qual é mais fácil de montar?
 - Quebra-cabeça novinho com peças pequenas?
 - Quebra-cabeça meio detonado com poucas peças?



Nanopore + Illumina

Genomas completos em dias

- Long reads para auxiliar a montagem e short reads para melhorar a qualidade das sequências
- Permite obtenção rápida e eficiente de genomas completos a partir de um único indivíduo
- Cerca de U\$150 por genoma (de drosophila)
- De 1 a 2 dias por genoma
- Possibilita amostragens populacionais de diversas espécies!

PLOS BIOLOGY

OPEN ACCESS PEER-REVIEWED

METHODS AND RESOURCES

Single-fly genome assemblies fill major phylogenomic gaps across the Drosophilidae Tree of Life

Bernard Y. Kim , Hannah R. Gellert, Samuel H. Church, Anton Suvorov, Sean S. Anderson, Olga Barmina, Sofia G. Beskid, Aaron A. Comeault, K. Nicole Crown, Sarah E. Diamond, Steve Dorus, Takako Fujichika, James A. Hemker, [...], Dmitri A. Petrov 

[view all]

Published: July 18, 2024 • <https://doi.org/10.1371/journal.pbio.3002697>

Article	Authors	Metrics	Comments	Media Coverage	Peer Review
					

Abstract

Introduction

Results and discussion

Materials and methods

Supporting information

Acknowledgments

References

Reader Comments

Figures

 **Accessible Data**

See the data 

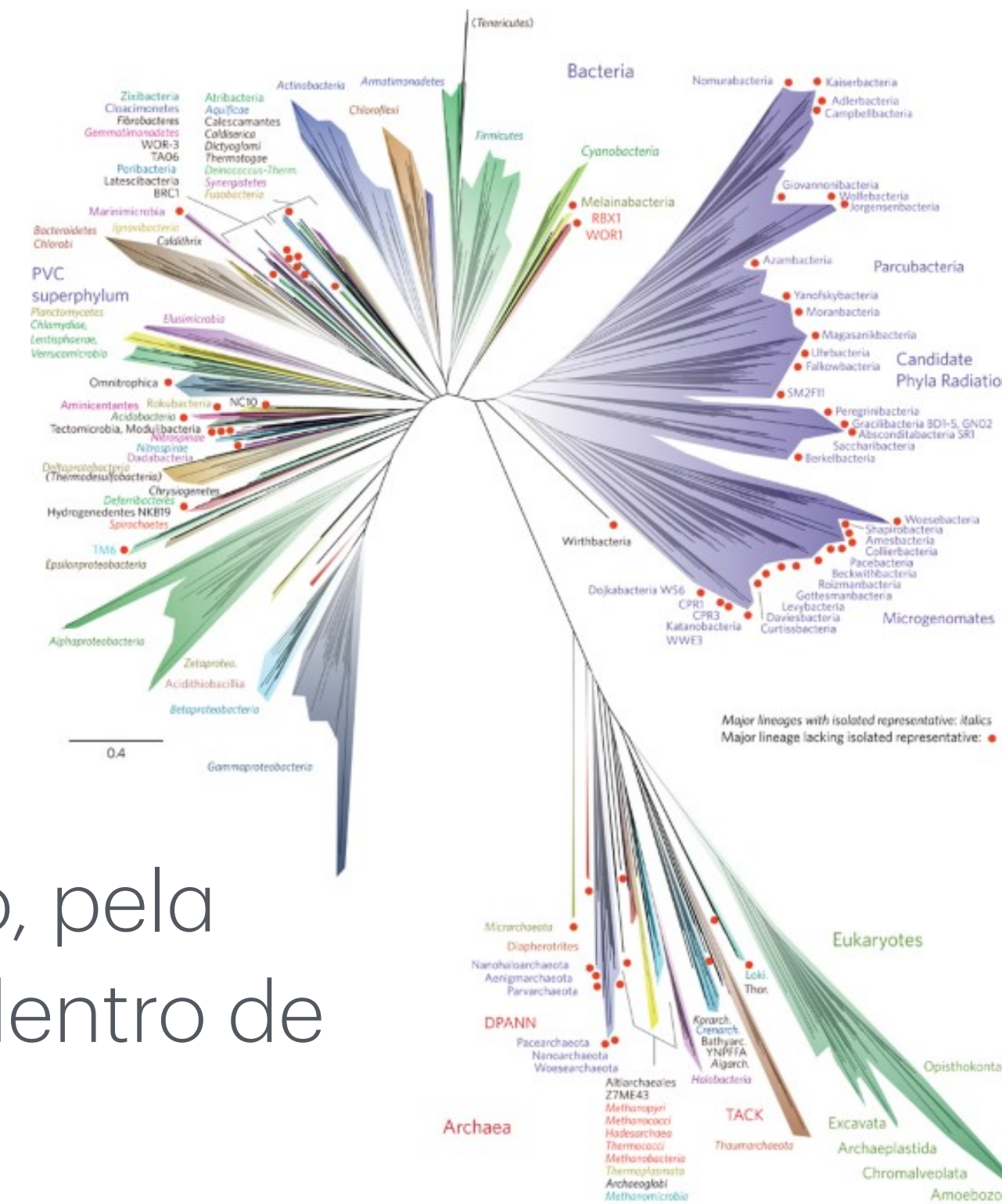
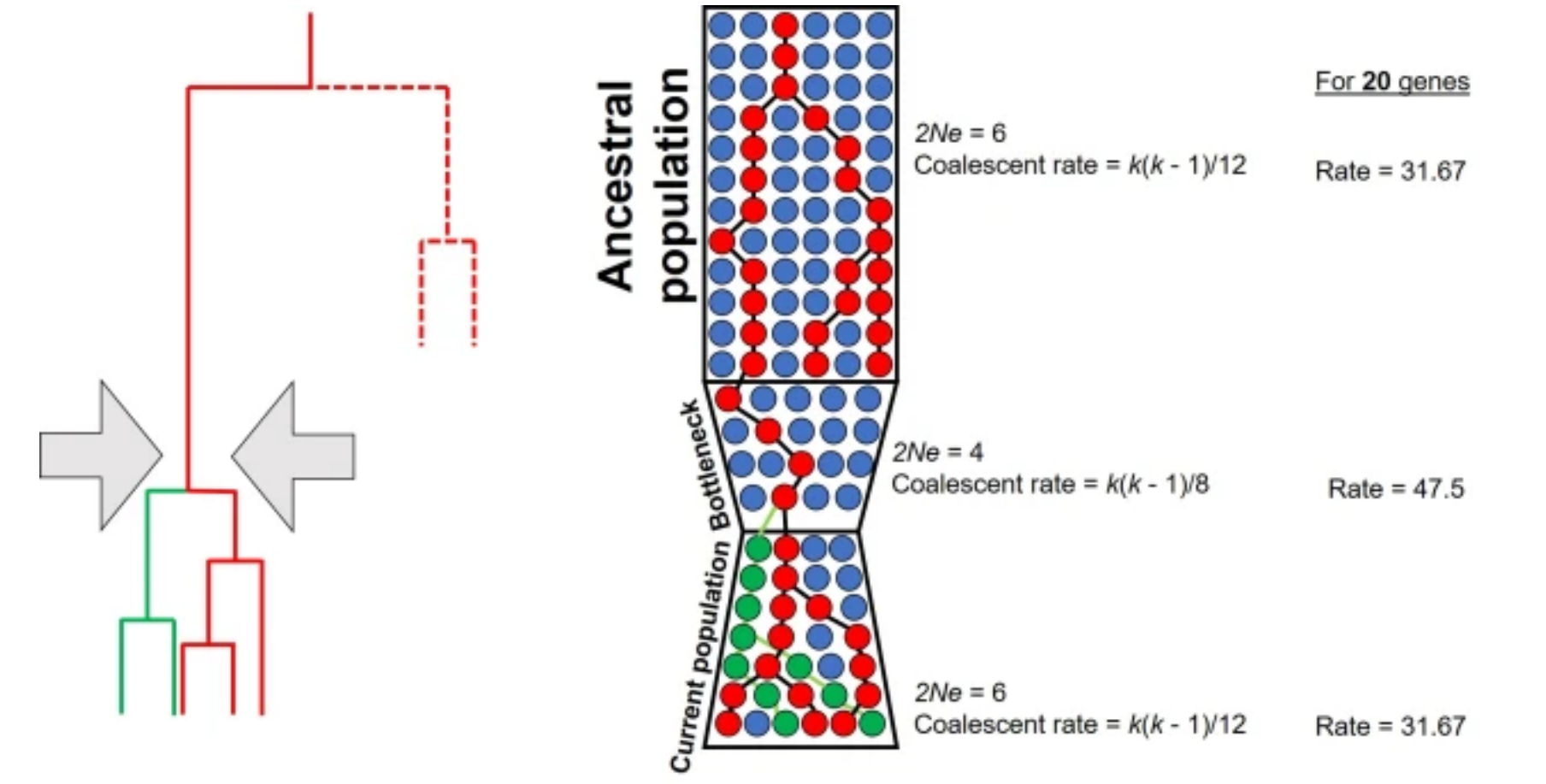
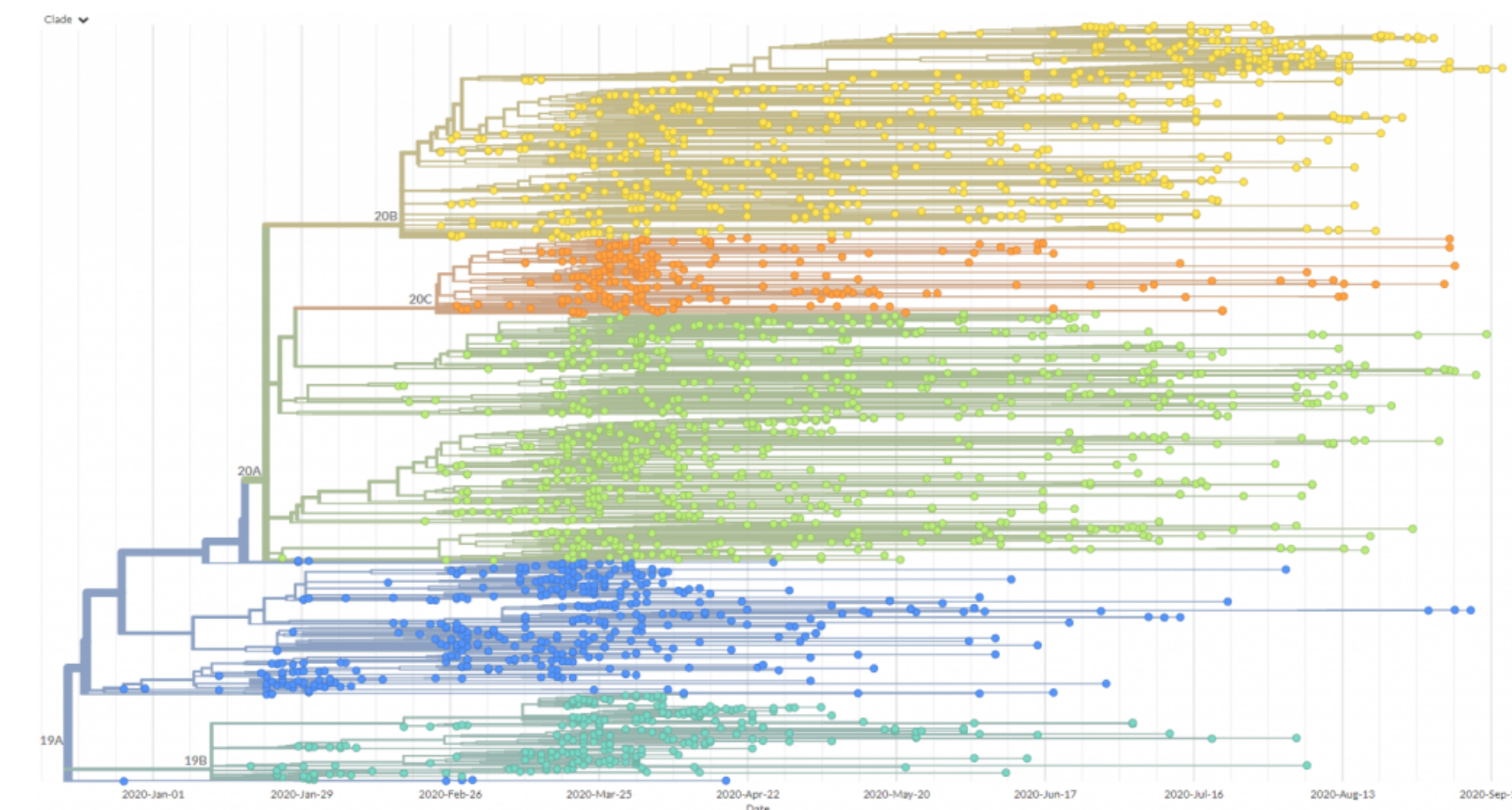
This article includes the Accessible Data icon, an experimental feature to encourage data sharing and reuse. [Find out how](#)

Abstract

Long-read sequencing is driving rapid progress in genome assembly across all major groups of life, including species of the family Drosophilidae, a longtime model system for genetics, genomics, and evolution. We previously developed a cost-effective hybrid Oxford Nanopore (ONT) long-read and Illumina short-read sequencing approach and used it to assemble 101 drosophilid genomes from laboratory cultures, greatly increasing the number of genome assemblies for this taxonomic group. The next major challenge is to address the laboratory culture bias in taxon sampling by sequencing genomes of species that cannot easily be reared in the lab. Here, we build upon our previous methods to perform amplification-free ONT sequencing of single wild flies obtained either directly from the field or from ethanol-preserved specimens in museum collections, greatly improving the representation of lesser studied drosophilid taxa in whole-genome data. Using Illumina Novaseq X Plus and ONT P2 sequencers with R10.4.1 chemistry, we set a new benchmark for inexpensive hybrid genome assembly at US \$150 per genome while assembling genomes from as little as 35 ng of genomic DNA from a single fly. We present 183 new genome assemblies for 179 species as a resource for drosophilid systematics, phylogenetics, and comparative genomics. Of these genomes, 62 are from pooled lab strains and 121 from single adult flies. Despite the sample limitations of working with small insects, most single-fly diploid assemblies are comparable in contiguity (>1 Mb contig N50), completeness (>98% complete dipteran BUSCOs), and accuracy (>QV40 genome-wide with ONT R10.4.1) to assemblies from inbred lines. We present a well-resolved multi-locus phylogeny for 360 drosophilid and 4 outgroup species encompassing all publicly available (as of August 2023) genomes for this group. Finally, we present a Progressive Cactus whole-genome, reference-free alignment built from a subset of 298 suitably high-quality drosophilid genomes. The new assemblies and alignment, along with updated laboratory protocols and computational pipelines, are released as an open resource and as a tool for studying evolution at the scale of an entire insect family.

Utilizando dados de sequencia

- Encontrar relações evolutivas
- Mapeamento genético
- Anotação de genomas
- Entender o processo evolutivo, pela comparação de sequencias dentro de populações e entre espécies

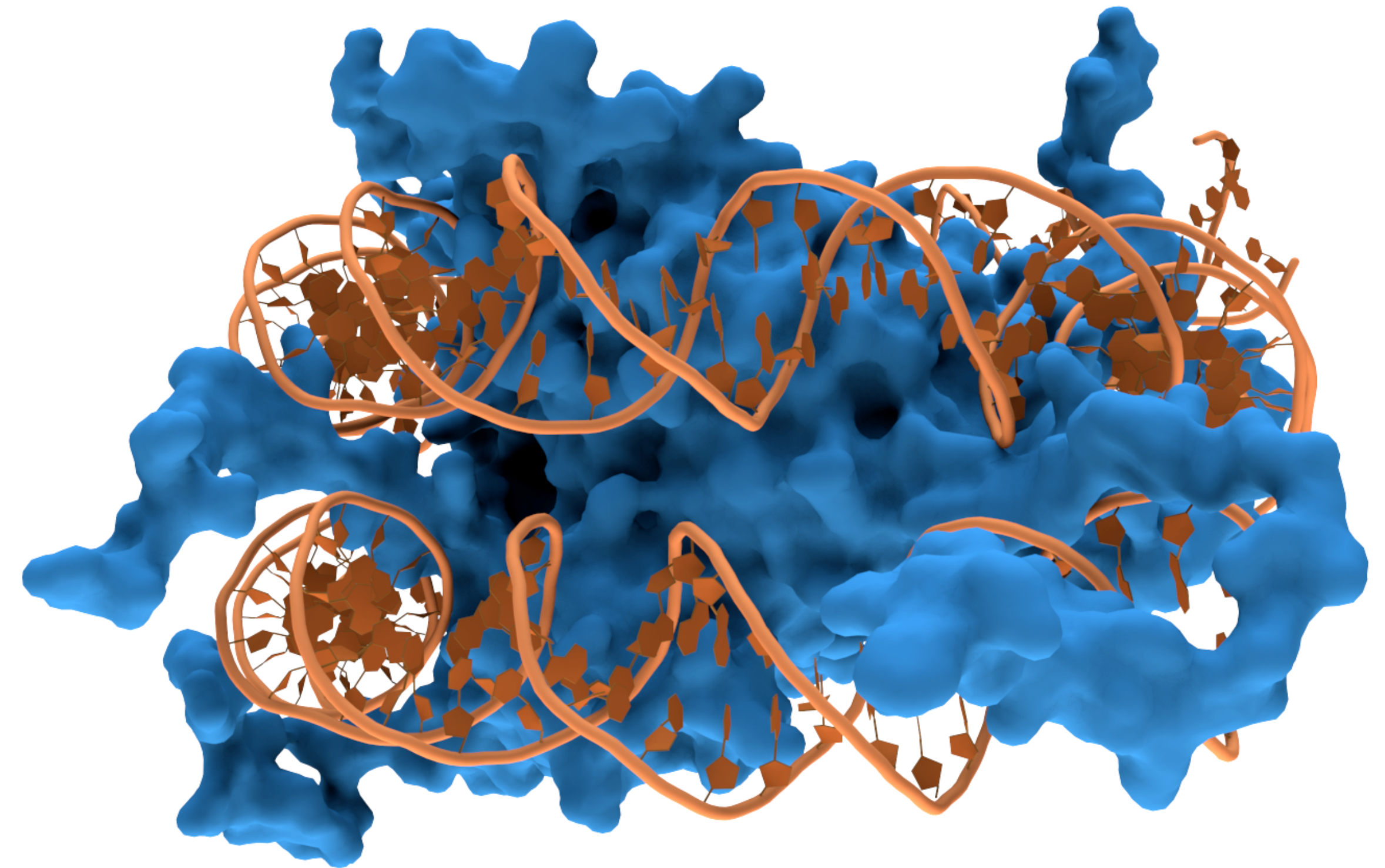


DNA para além da sequência

Outros aspectos da molécula de DNA

Para além das sequências

- A molécula de DNA interage de forma complexa com si mesma e com outras proteínas
- Ensaios moleculares podem ajudar a entender esses processos físicos e como eles se relacionam com a expressão dos fenótipos de interesse
 - Hi-C: interações DNA-DNA
 - ChIP-seq: interações DNA-proteína

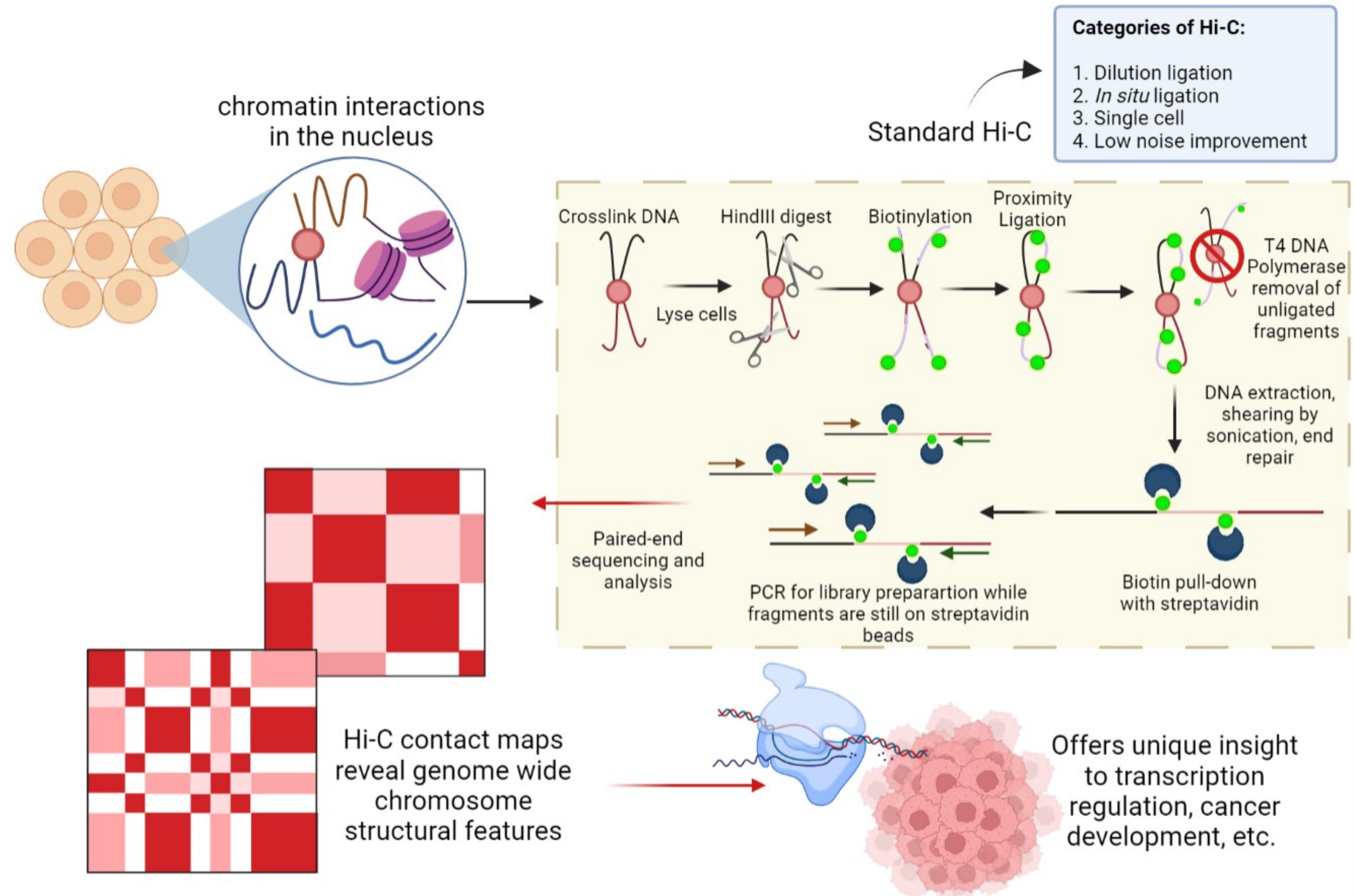


Histona interagindo com a molécula de DNA

Encontrando interações DNA-DNA

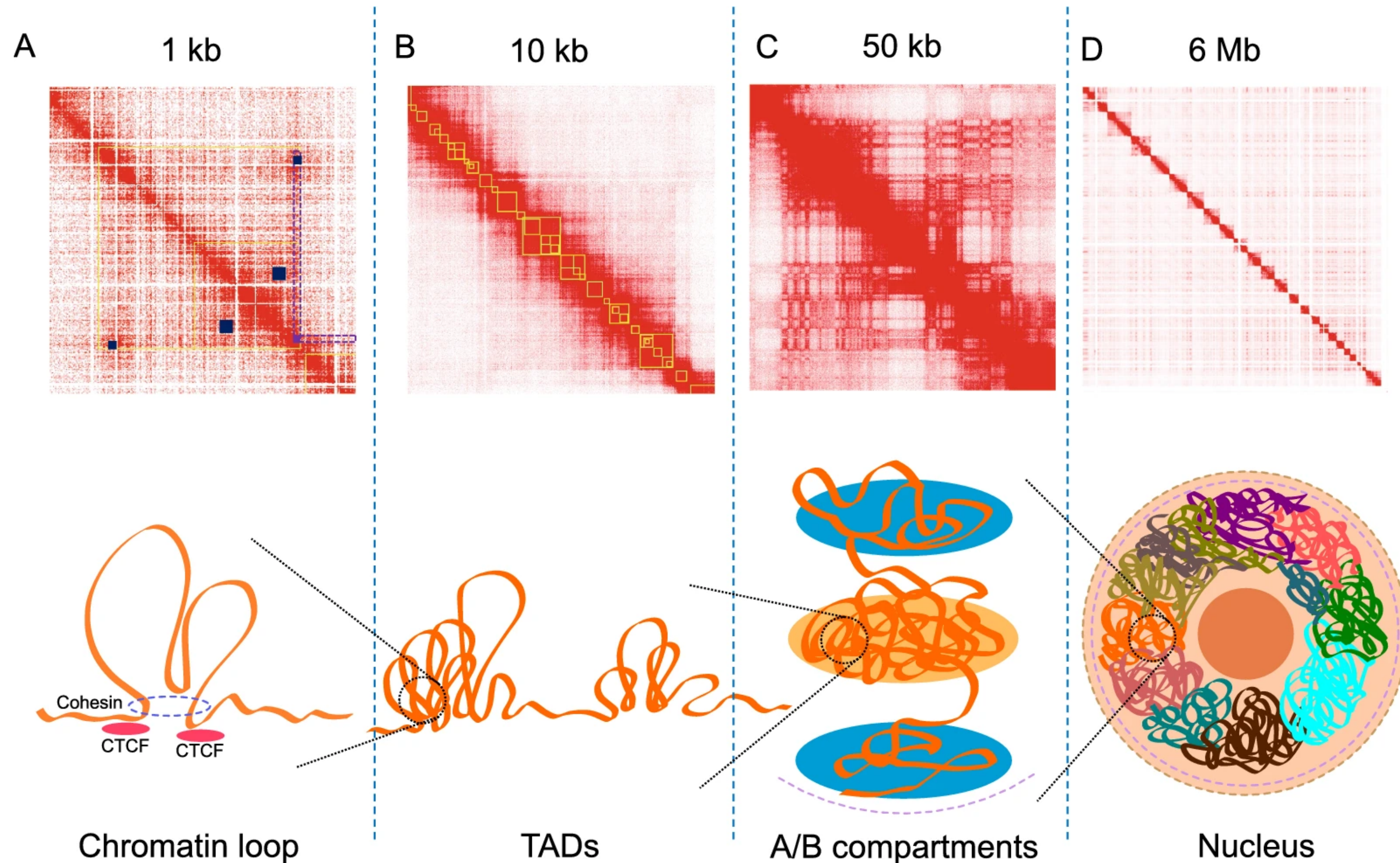
Hi-C e micro-C

- Quimicamente junta regiões de DNA próximas
- Cliva o DNA mantendo os complexos
- Junta as sequências que interagem
- Identifica interações por sequenciamento



Encontrando interações DNA-DNA

Diferentes escalas de interação



Promotores compartilhados

micro-C revela interações regulatórias de longo alcance

Article | Published: 04 May 2022

Transcriptional coupling of distant regulatory genes in living embryos

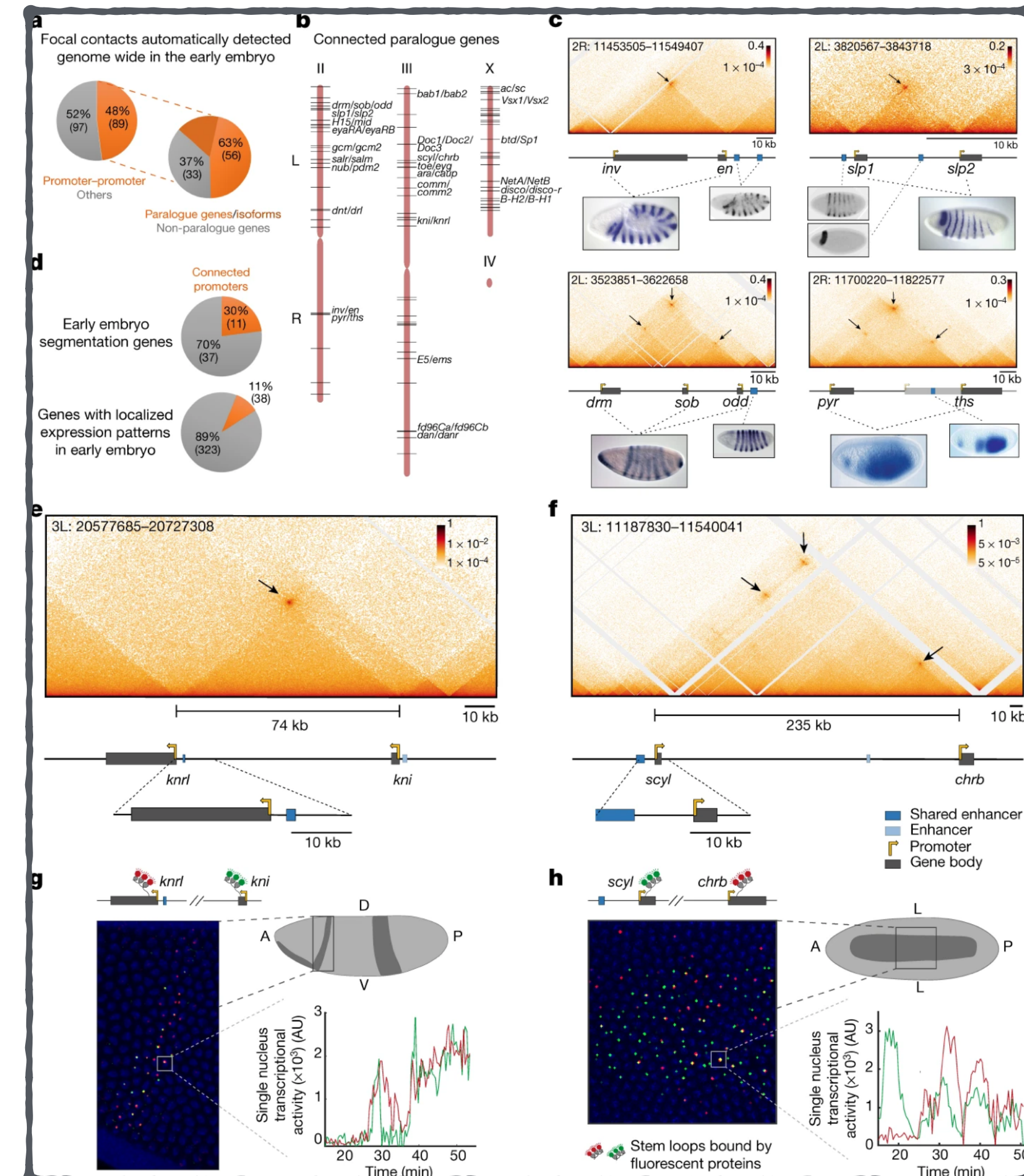
[Michal Levo](#), [João Raimundo](#), [Xin Yang Bing](#), [Zachary Sisco](#), [Philippe J. Batut](#), [Sergey Ryabichko](#),
[Thomas Gregor](#) ✉ & [Michael S. Levine](#) ✉

[Nature](#) **605**, 754–760 (2022) | [Cite this article](#)

22k Accesses | 67 Citations | 159 Altmetric | [Metrics](#)

Abstract

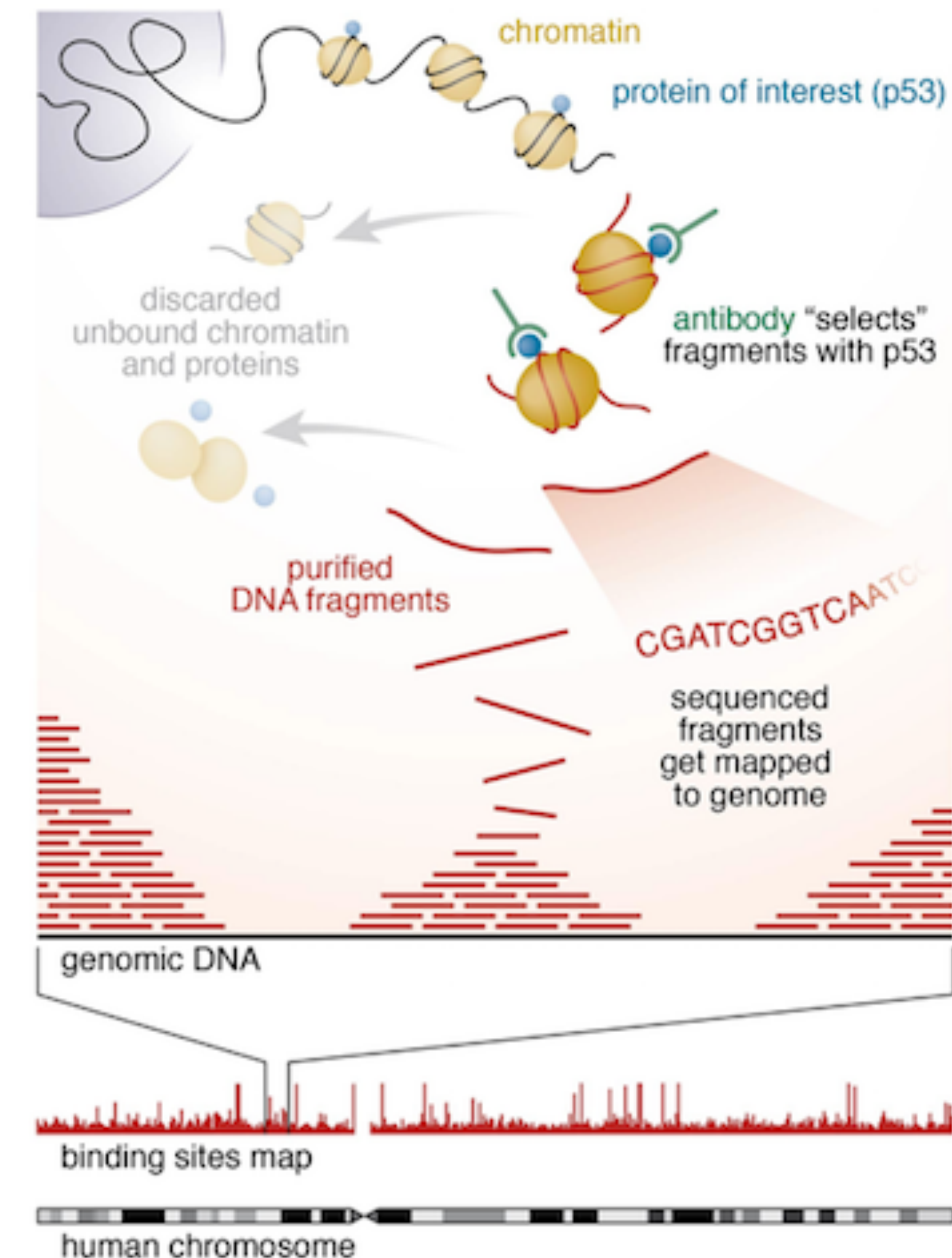
The prevailing view of metazoan gene regulation is that individual genes are independently regulated by their own dedicated sets of transcriptional enhancers. Past studies have reported long-range gene–gene associations^{1,2,3}, but their functional importance in regulating transcription remains unclear. Here we used quantitative single-cell live imaging methods to provide a demonstration of co-dependent transcriptional dynamics of genes separated by large genomic distances in living *Drosophila* embryos. We find extensive physical and functional associations of distant paralogous genes, including co-regulation by shared enhancers and co-transcriptional initiation over distances of nearly 250 kilobases. Regulatory interconnectivity depends on promoter-proximal tethering elements, and perturbations in these elements uncouple transcription and alter the bursting dynamics of distant genes, suggesting a role of genome topology in the formation and stability of co-transcriptional hubs. Transcriptional coupling is detected throughout the fly genome and encompasses a broad spectrum of conserved developmental processes, suggesting a general strategy for long-range integration of gene activity.



Interações DNA proteína

Capturando estados a cromatina

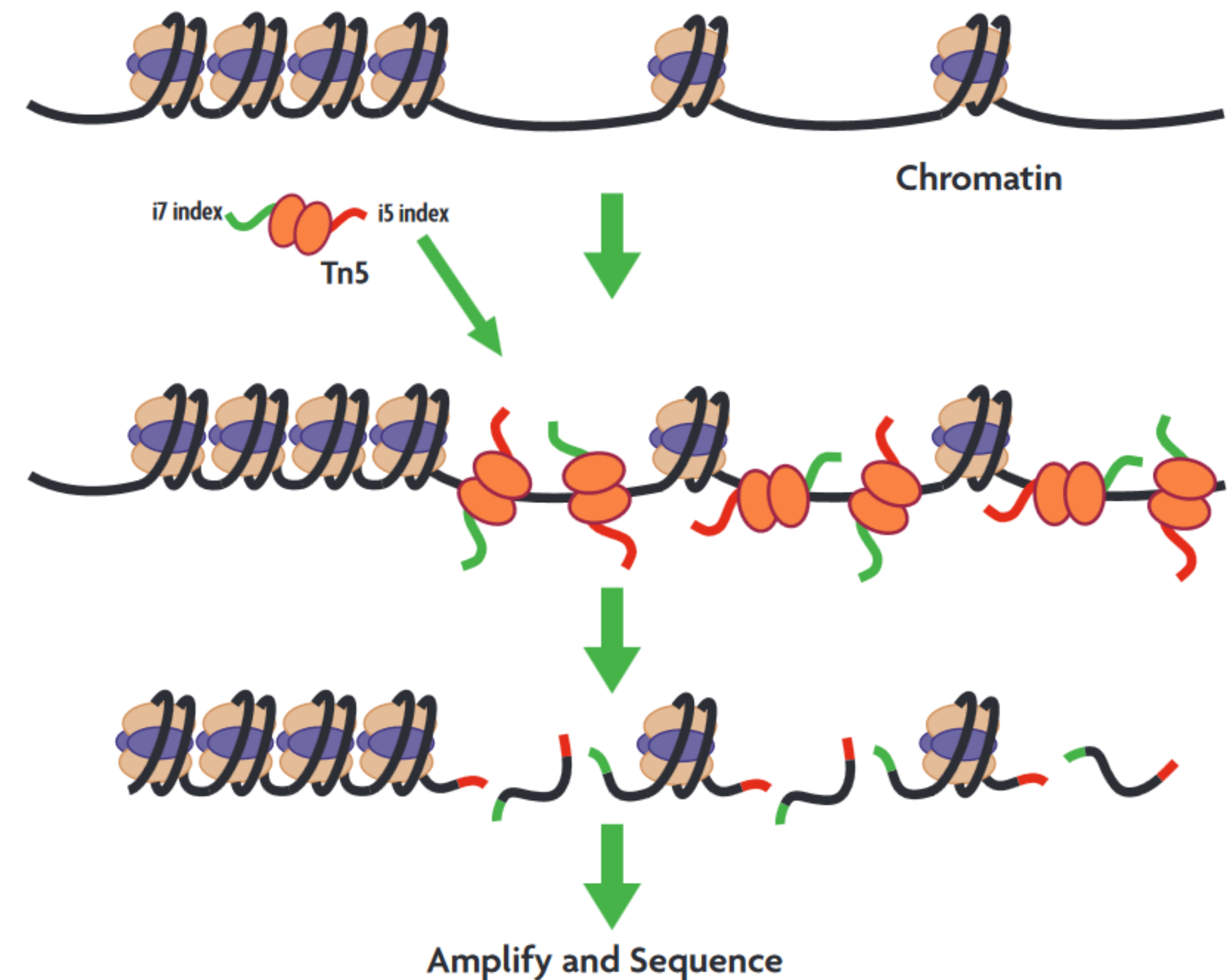
- Imuno-precipitação da cromatina (ChIP-seq) usa anticorpos específicos para determinadas proteínas e sequenciamento
- Complexos DNA-proteína são isolados e o DNA sequenciado



Identificando regiões ativas do genoma

Assay for Transposase-Accessible Chromatin using sequencing
ATAC-seq

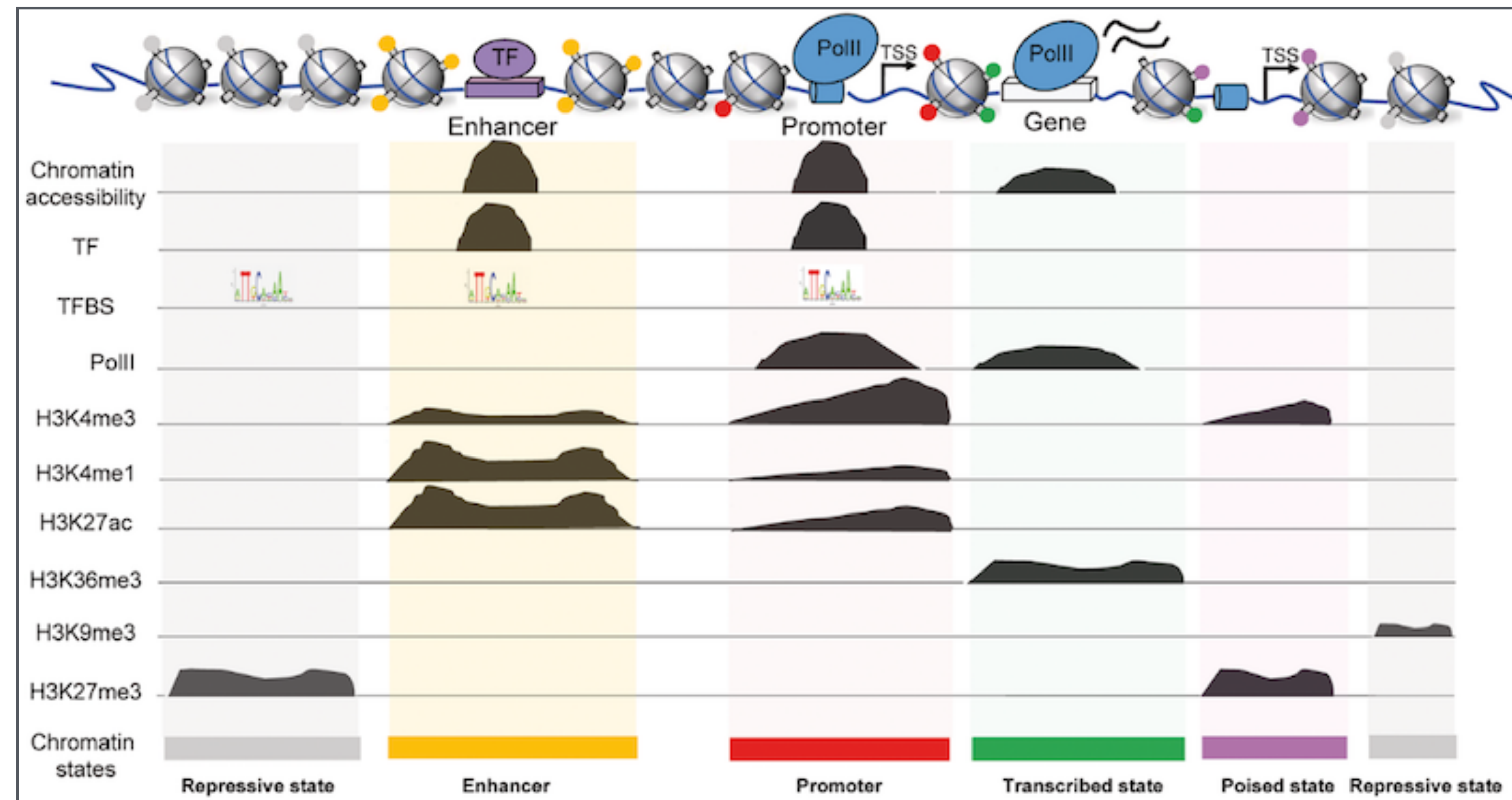
- ATAC-seq procura por regiões “abertas” do genoma, que podem ser ativas
- Novamente, o genoma é clivado e sequenciado, mas somente regiões acessíveis são usadas



Combinando tudo

chromHMM e estados da cromatina

Juntando informações de sequências, Hi-C, ATAC-seq e ChIP-seq, podemos criar mapas detalhados das diferentes regiões ativas e inativas do genoma

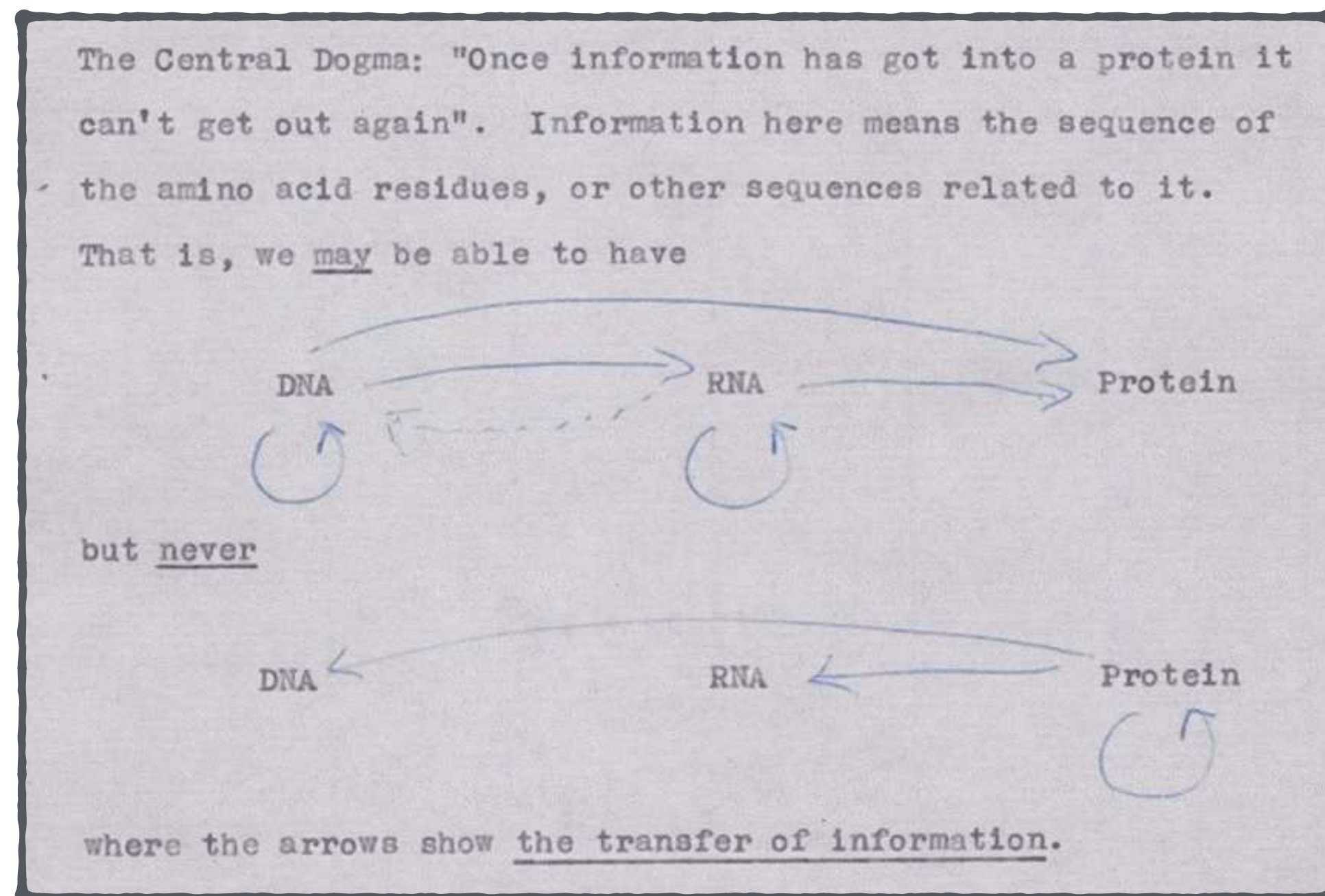
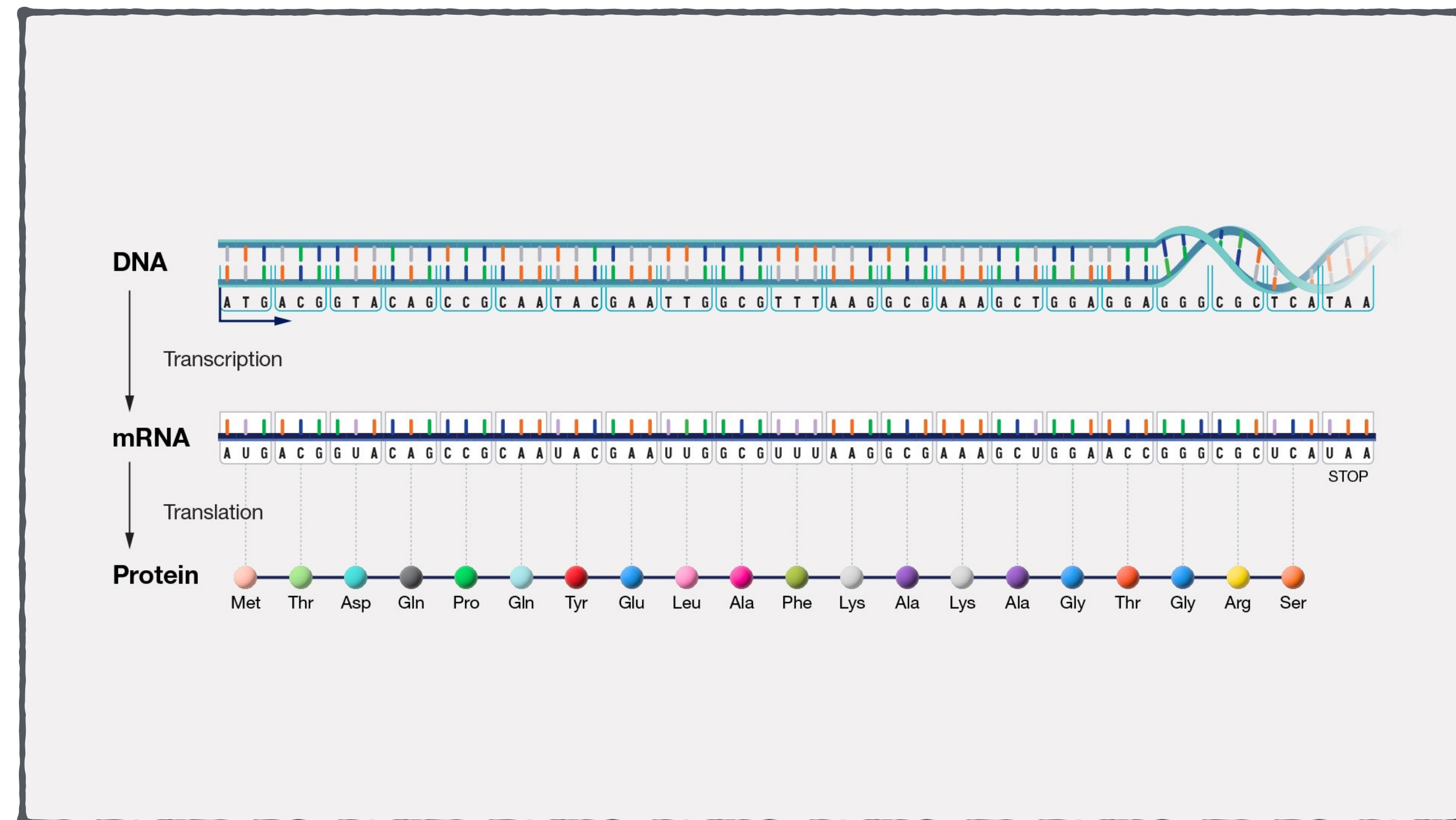


RNA e transcriçãõ

O processo transcricional

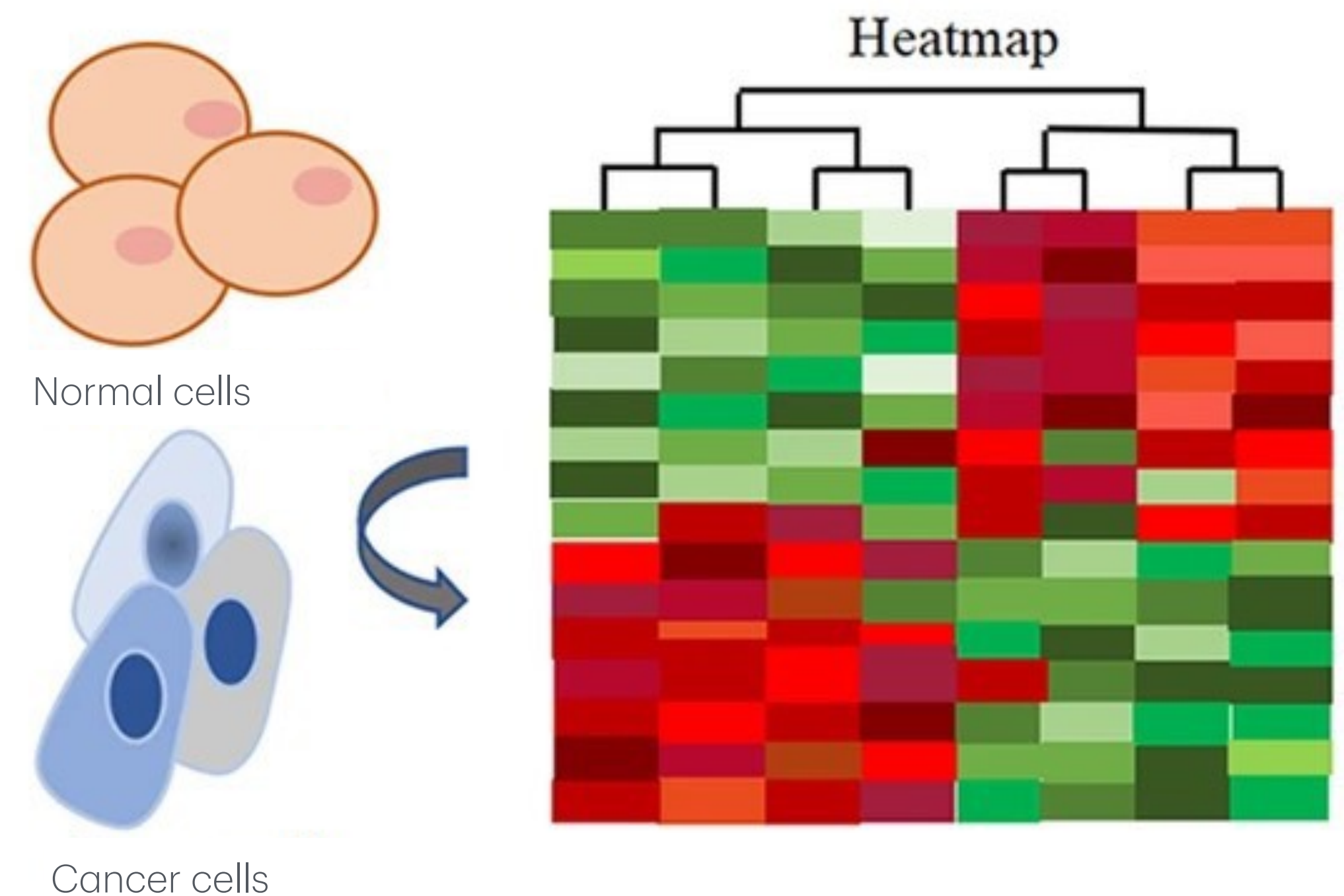
Formando RNA a partir de DNA

- A expressão das proteínas a partir genes codificados no genoma passa pela produção do intermediário de mRNA



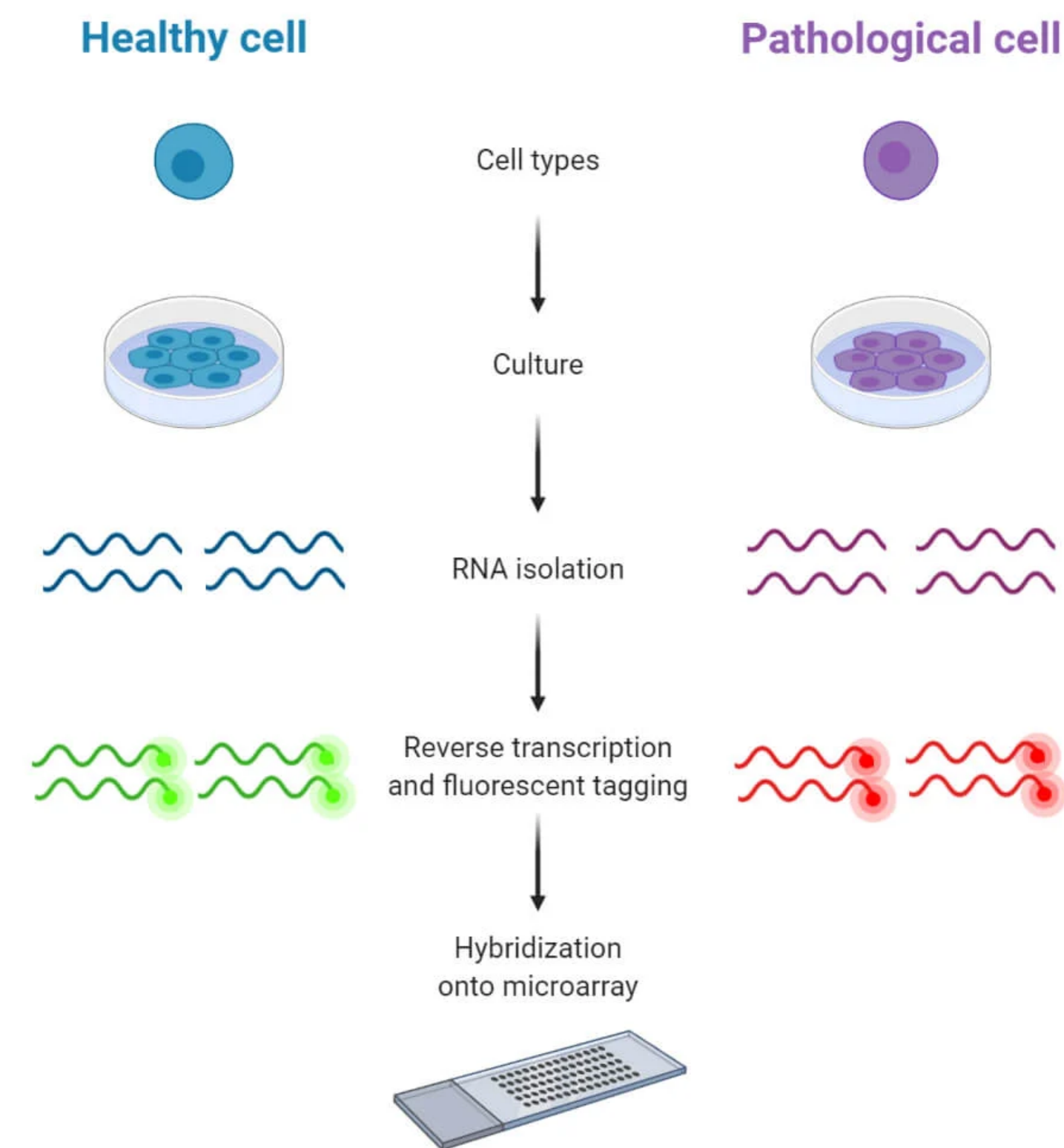
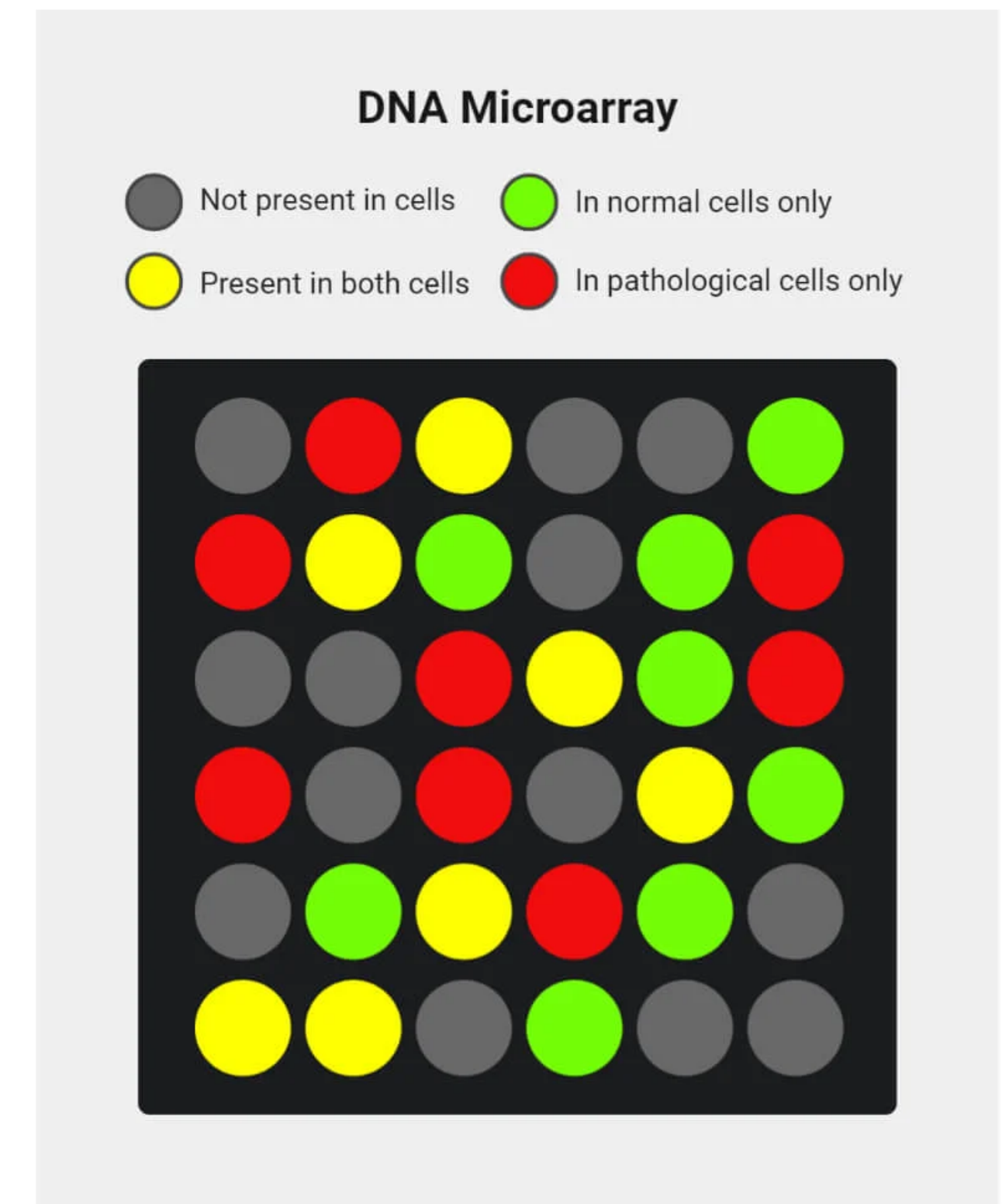
Aferindo níveis de RNA mensageiro

- Uma estratégia para entender o processo de produção do fenótipo é aferir os níveis de diferentes mRNA,
- Identificando quais genes estão sendo transcritos em uma situação fisiológica
- Identificando como a expressão gênica muda em diferentes condições



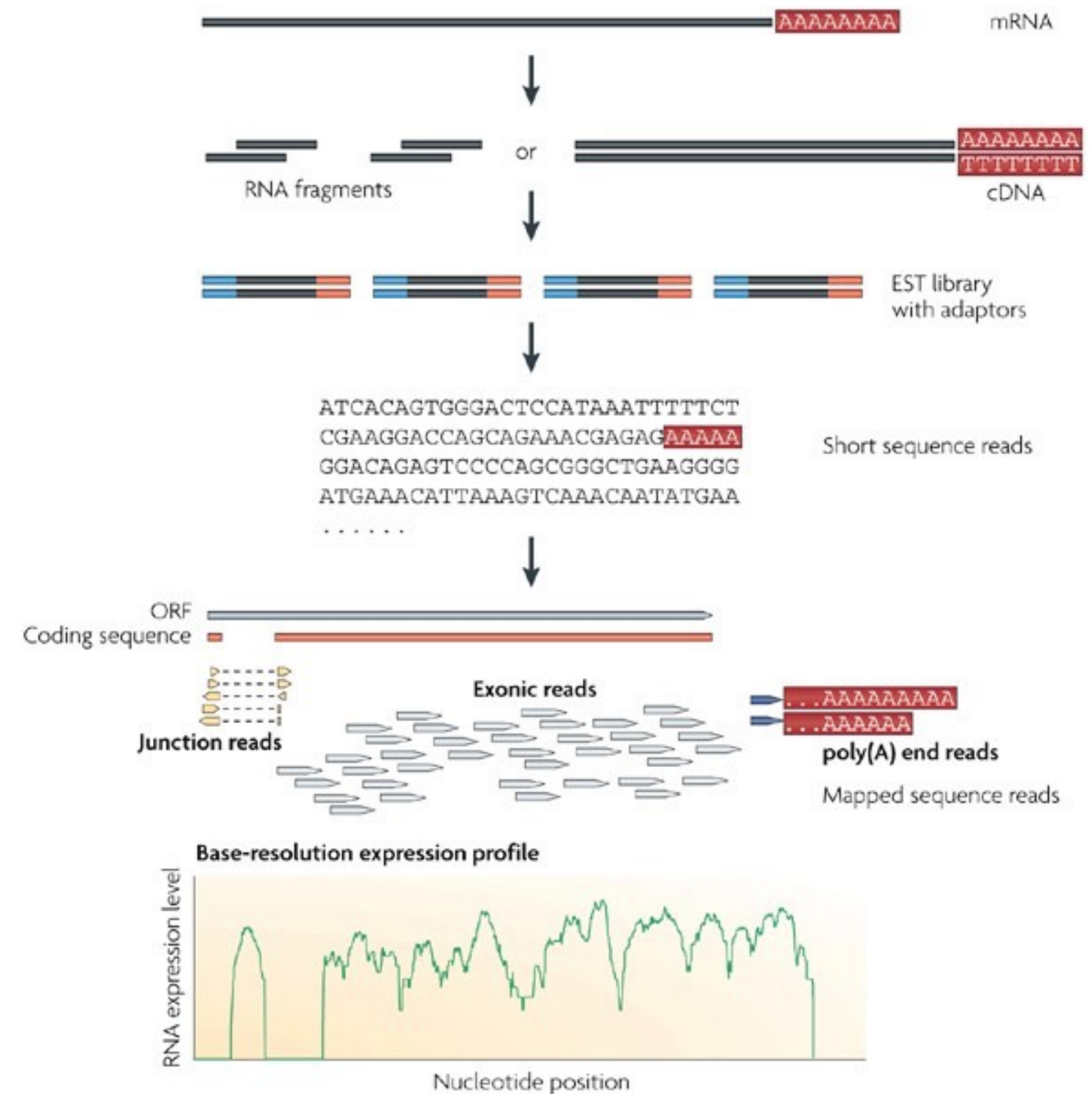
Microarrays

- Microarrays utilizam moléculas de cDNA produzidas a partir do mRNA
- Um conjunto **fixo** de genes tem sua expressão aferida, utilizando a complementariedade e técnicas de fluorescência para identificar uma sequência específica
- Não envolve sequenciamento



RNA-seq

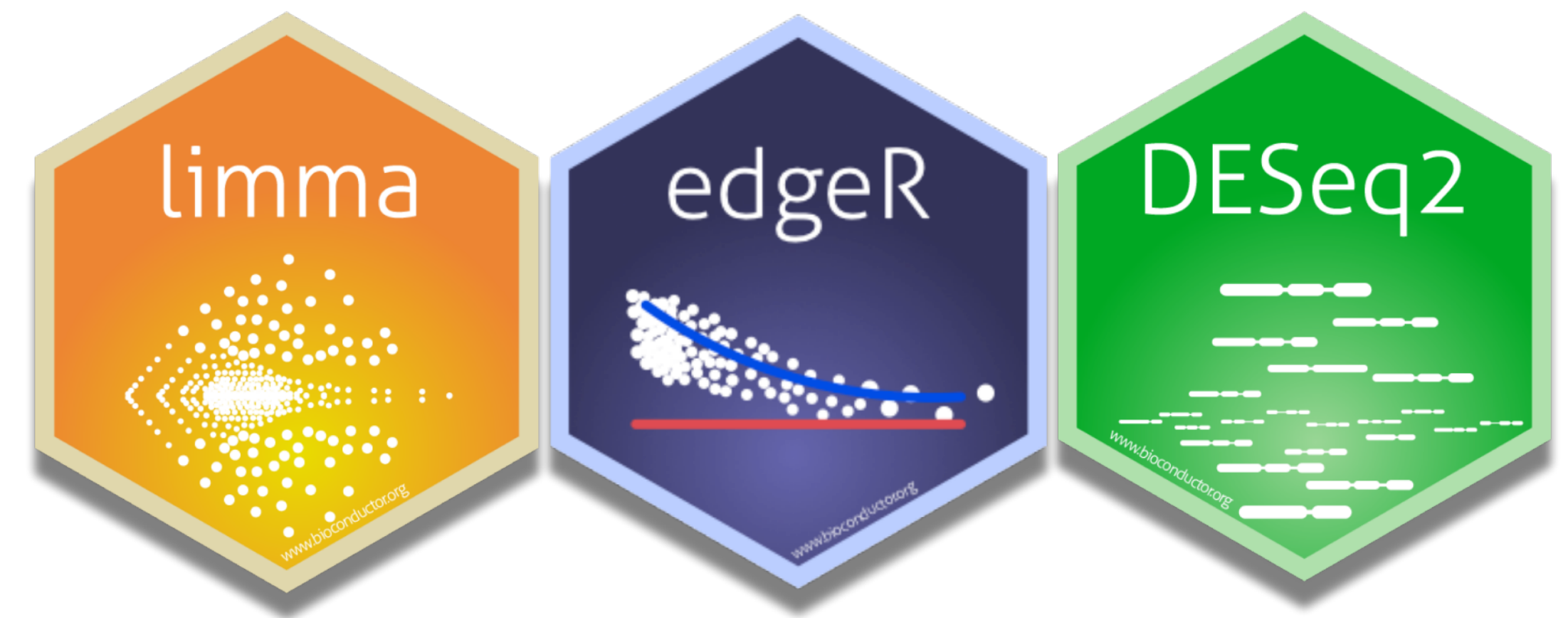
- O RNA-seq é a técnica mais comum para medir níveis de mRNA atualmente
- Consiste em criar bibliotecas de cDNA e sequenciar essas bibliotecas usando short-reads
- Esses reads são então mapeados ao genoma de referências, e a cobertura é usada como proxy do nível de expressão
- Gera “gene counts”, quantos reads são associados a um determinado gene



Analise de RNA-seq

Contagens são difíceis

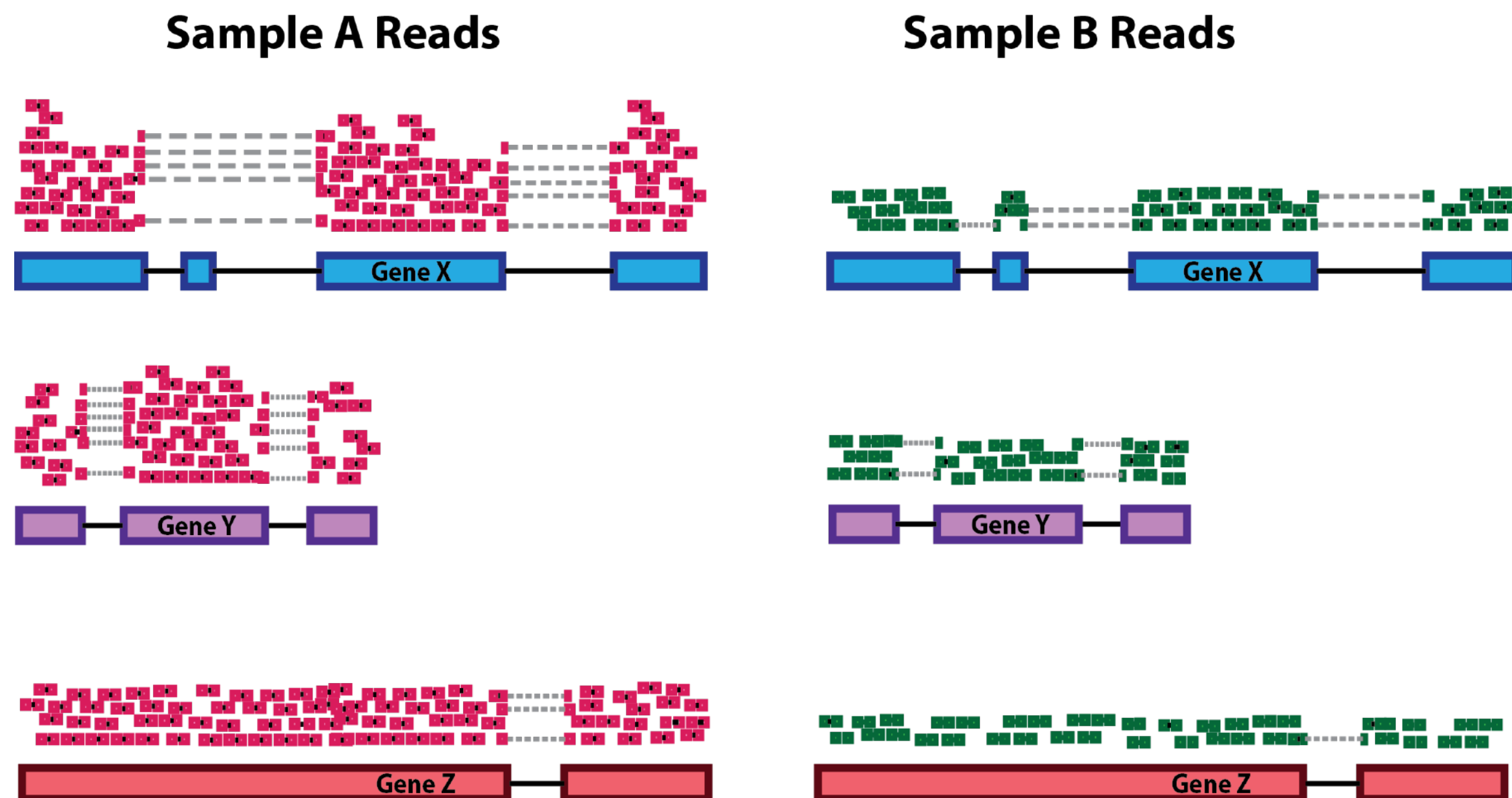
- O uso de dados de contagem em RNA-seq exige métodos estatísticos mais complicados
- Pacotes como limma, edgeR ou DESeq2 tratam os dados de forma diferente para obter uma lista de genes diferencialmente expressos



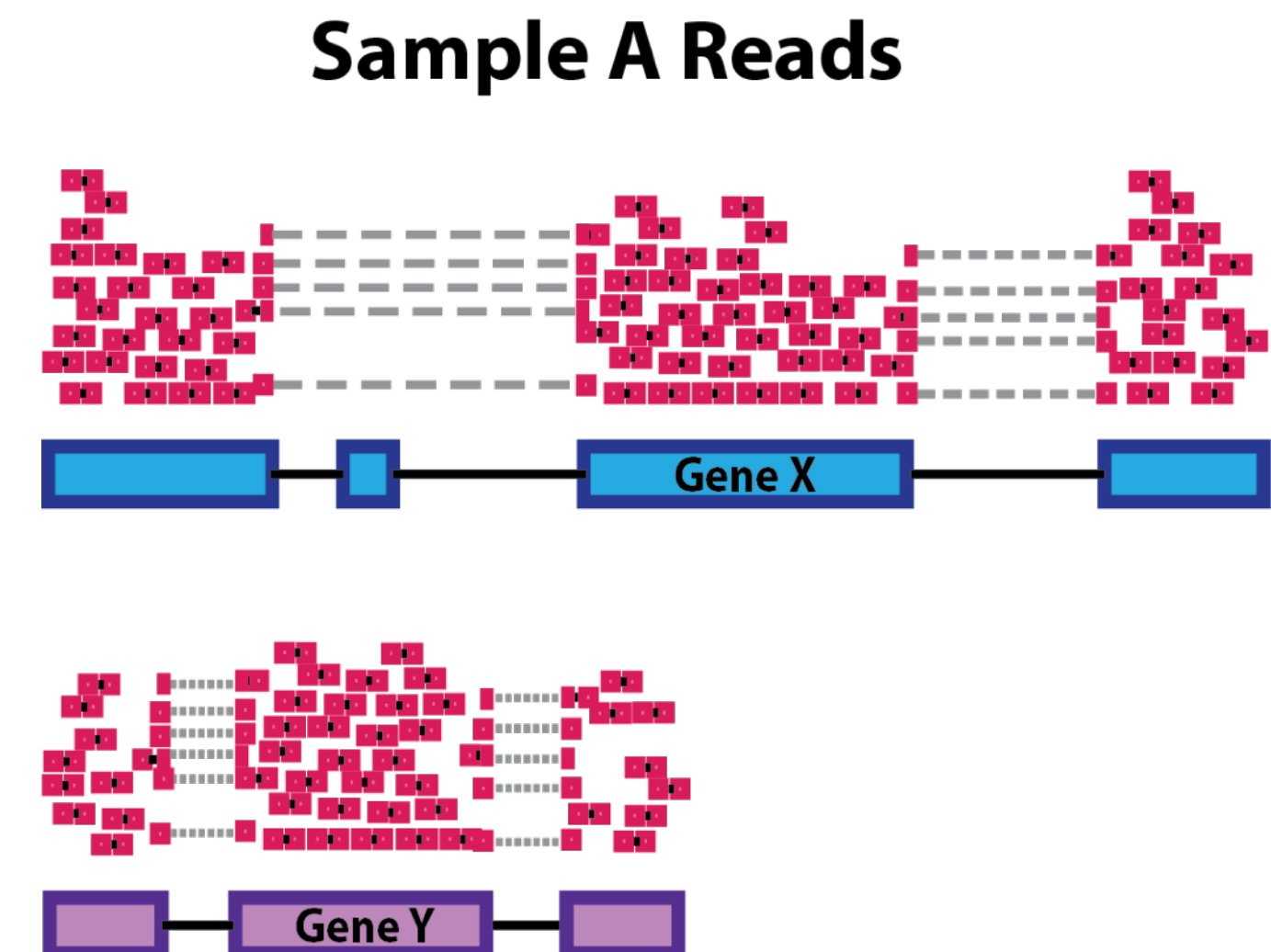
Analise de RNA-seq

Normalizações

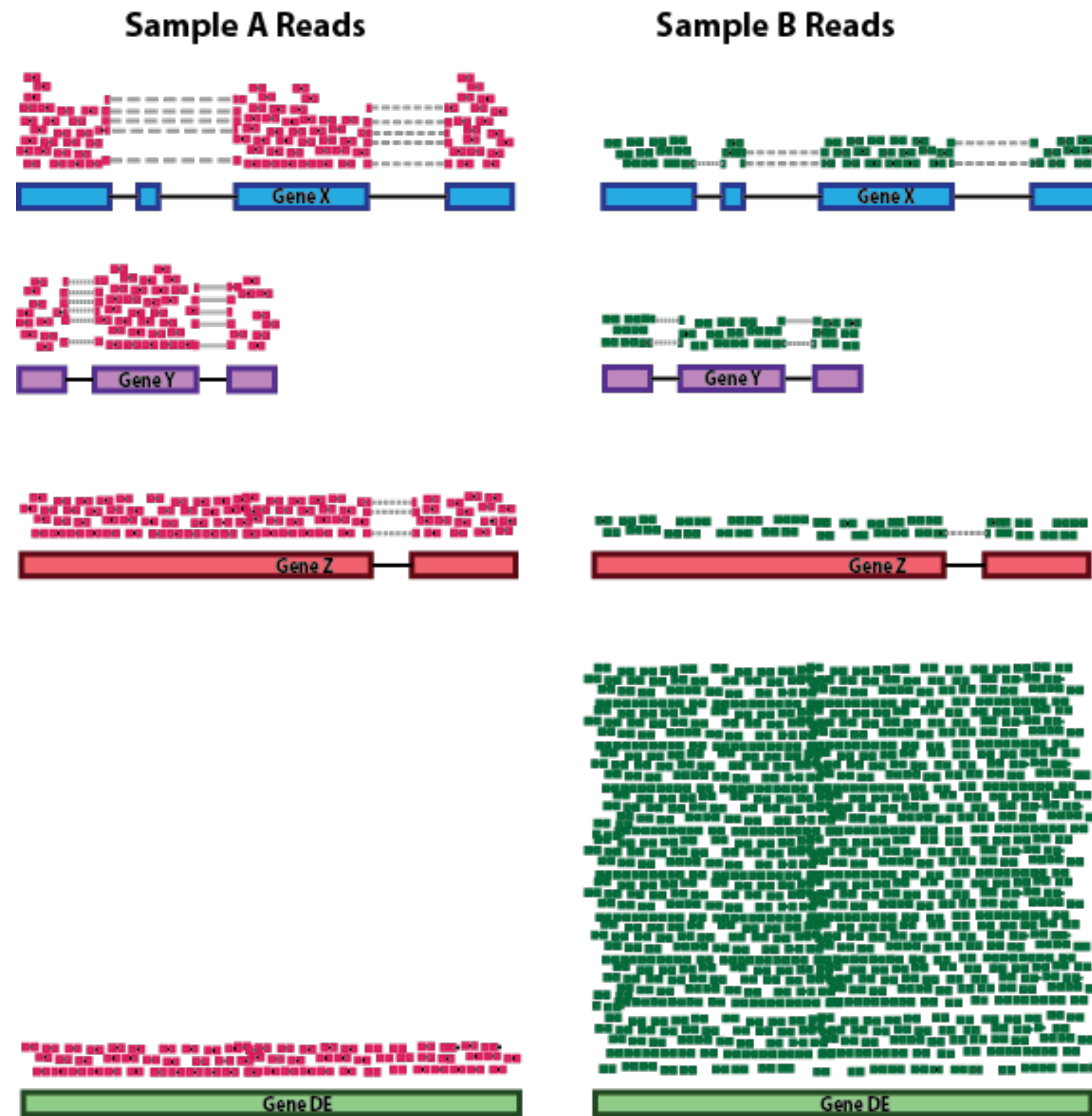
Diferenças de cobertura



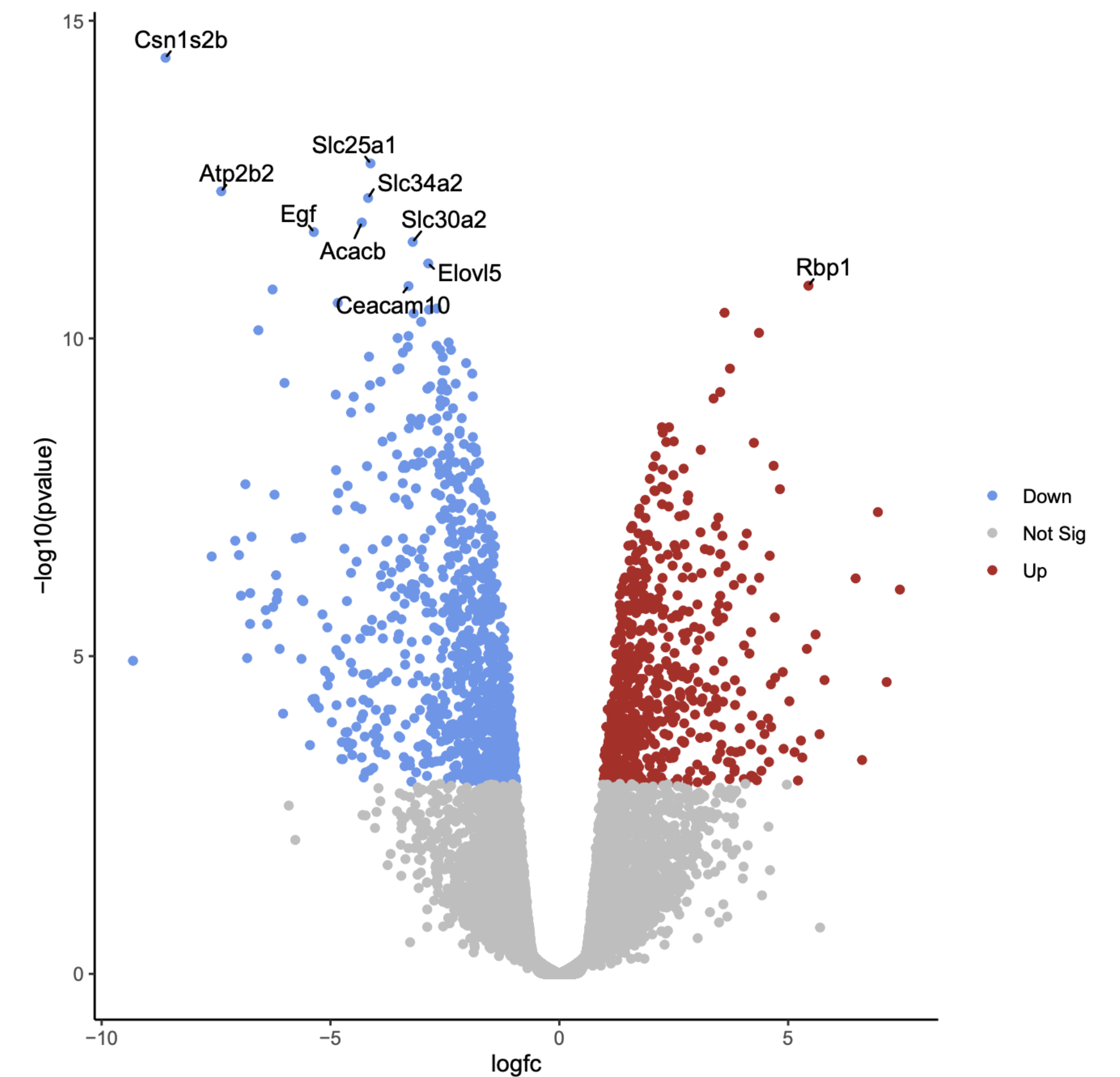
Diferenças de comprimento



Expressão diferencial



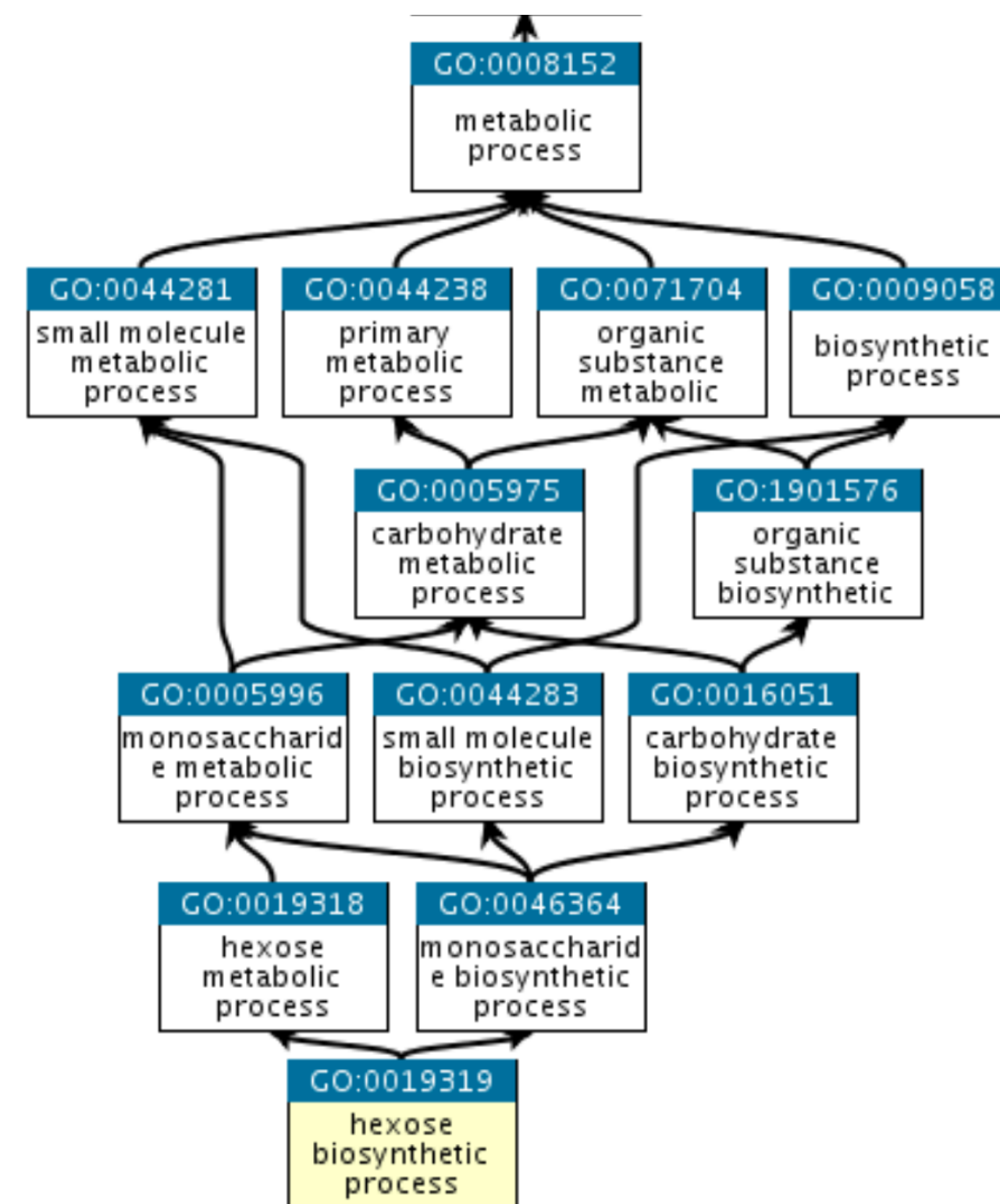
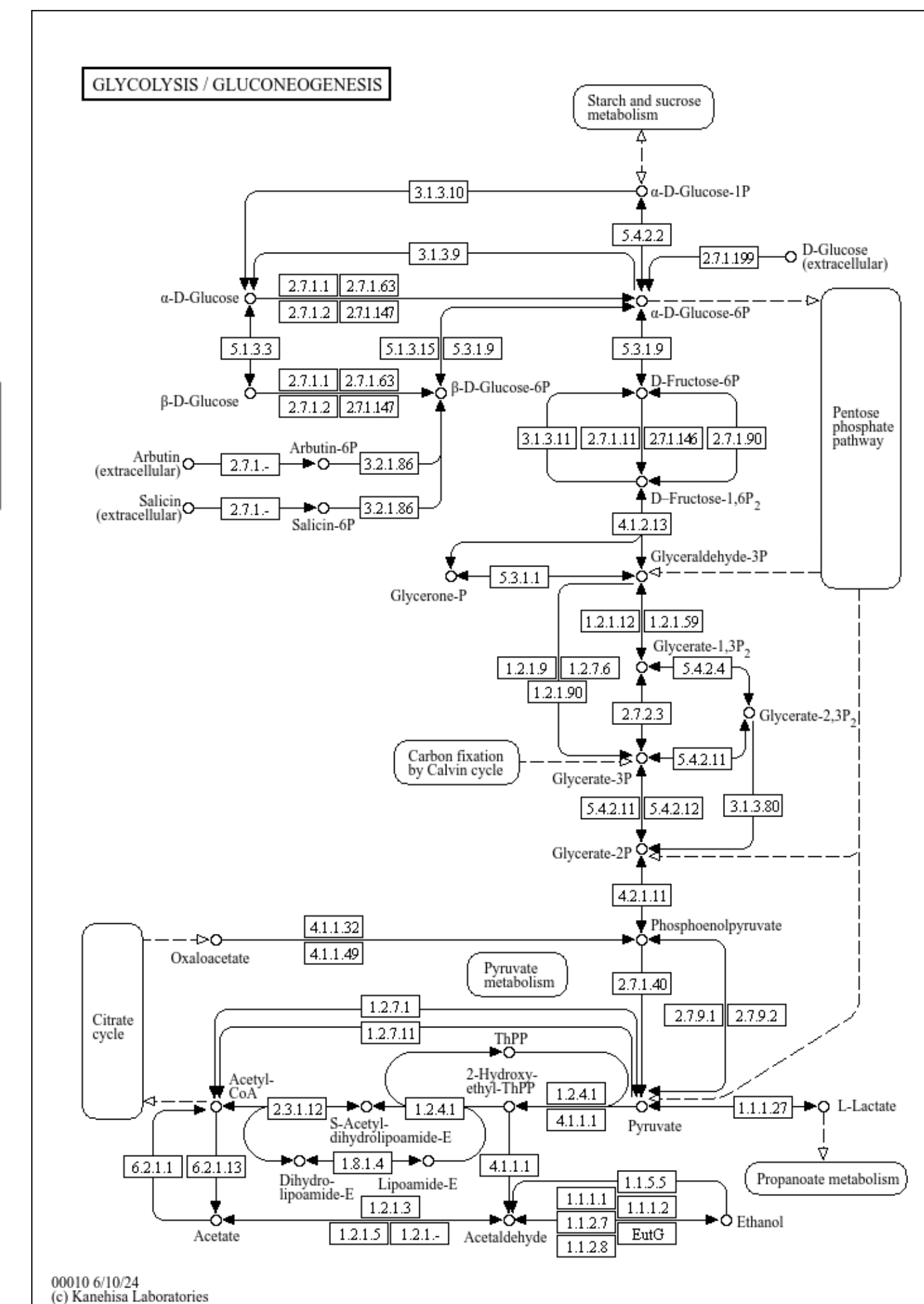
Volcano plot



Análise de enriquecimento funcional

Interpretando dados de expressão diferencial

- A partir de uma lista de genes, podemos utilizar bancos de dados funcionais para entender e quais processos aqueles genes podem estar envolvidos
- **Gene Ontology (GO)**: Sistema hierárquico de anotação de genes e processos biológicos
- **KEGG**: Banco de dados de anotações funcionais, que liga enzimas e proteínas a processos metabólicos



<https://www.genome.jp/kegg/>
<https://geneontology.org/>

Entendendo a regulação gênica

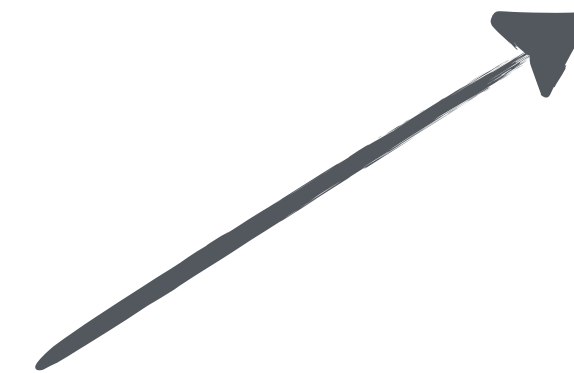
Allele specific expression e eQTLs

- Podemos explorar como o organismo decide expressar um gene ou outro
- Algumas técnicas:
 - Tentar ligar a variação genética com a variação de expressão, usando modelos de mapeamento e encontrando eQTLs (loci associados a traços de expressão gênica)
 - Encontrar relações variacionais entre a expressão de genes diferentes

Entendendo a regulação gênica

Allele specific expression e eQTLs

- Podemos explorar como o organismo decide expressar um gene ou outro
- Algumas técnicas:
 - Tentar ligar a variação genética com a variação de expressão, usando modelos de mapeamento e encontrando eQTLs (loci associados a traços de expressão gênica)
 - Encontrar relações variacionais entre a expressão de genes diferentes



Difícil, pois são muitos caracteres possíveis!
Problem de múltiplos testes se intensifica

Entendendo a regulação gênica

Allele specific expression e eQTLs

- Podemos explorar como o organismo decide expressar um gene ou outro
- Algumas técnicas:
 - Tentar ligar a variação genética com a variação de expressão, usando modelos de mapeamento e encontrando eQTLs (loci associados a traços de expressão gênica)
 - Encontrar relações variacionais entre a expressão de genes diferentes

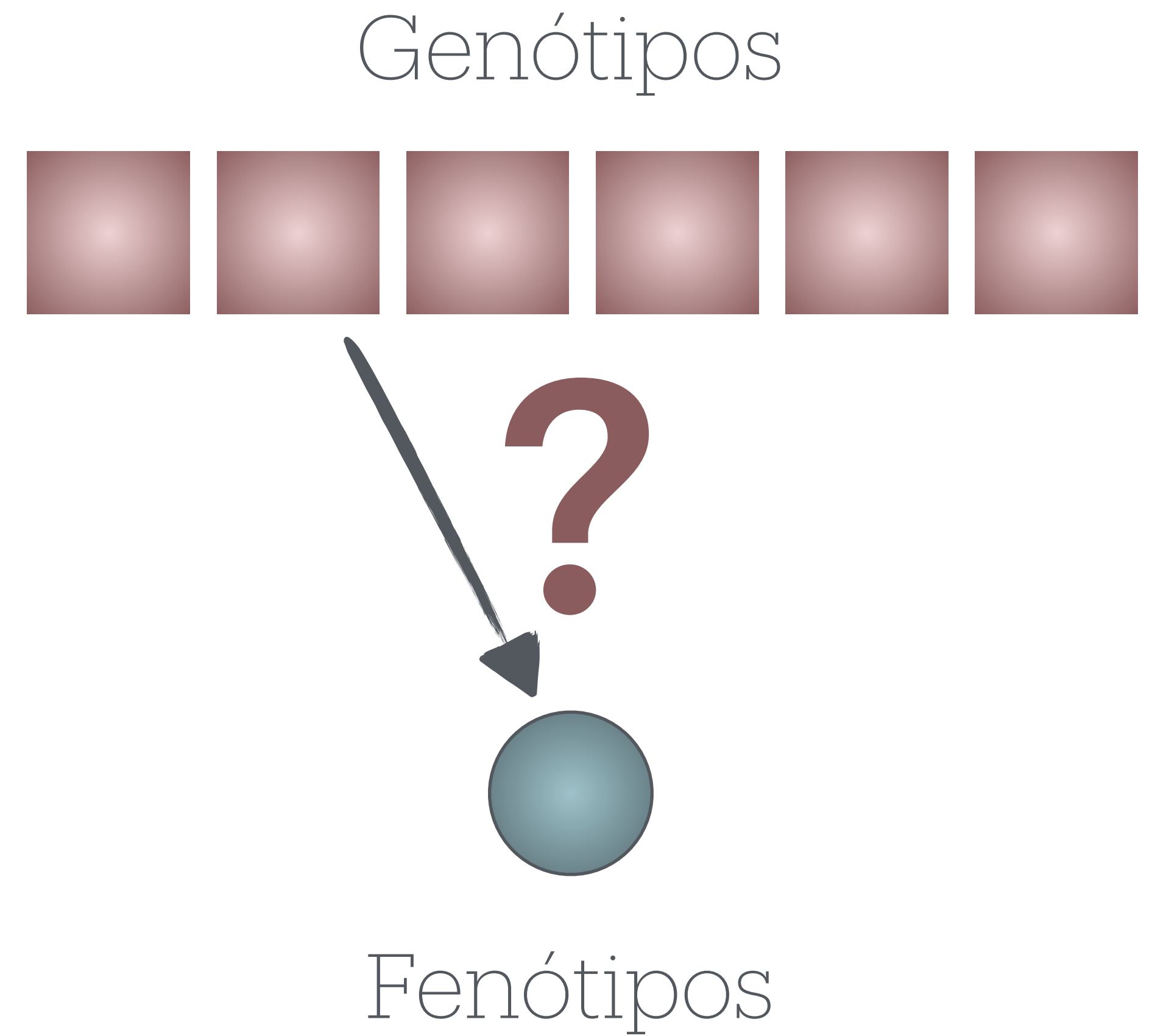


Difícil, pois são muitos caracteres possíveis!
Problem de múltiplos testes se intensifica

Depende de calcular muito, muitos mesmo, pares de correlação (CUIDADO!)

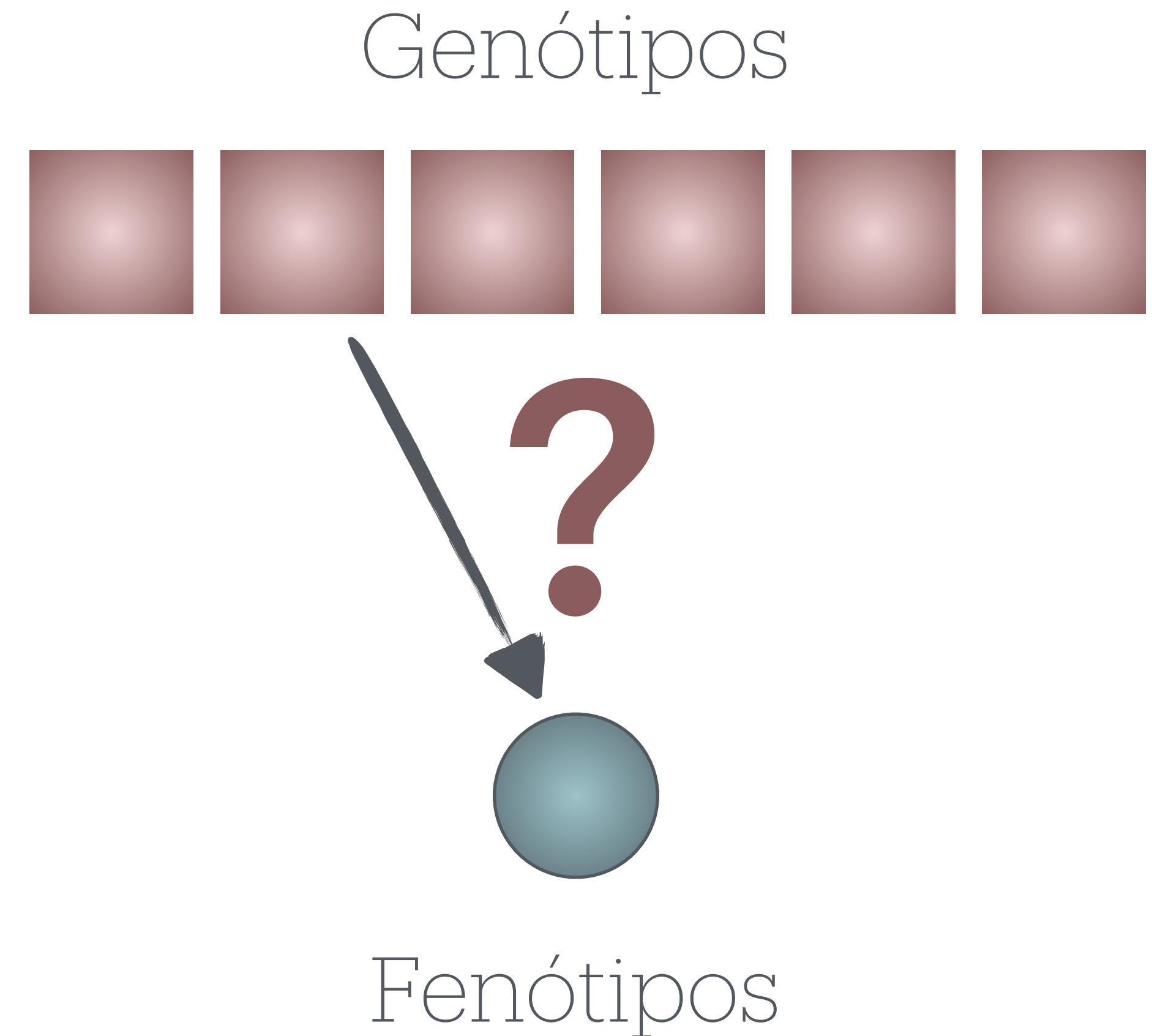
Mapeamento Genético

Como medir o efeito da variação em um locus no fenótipo?



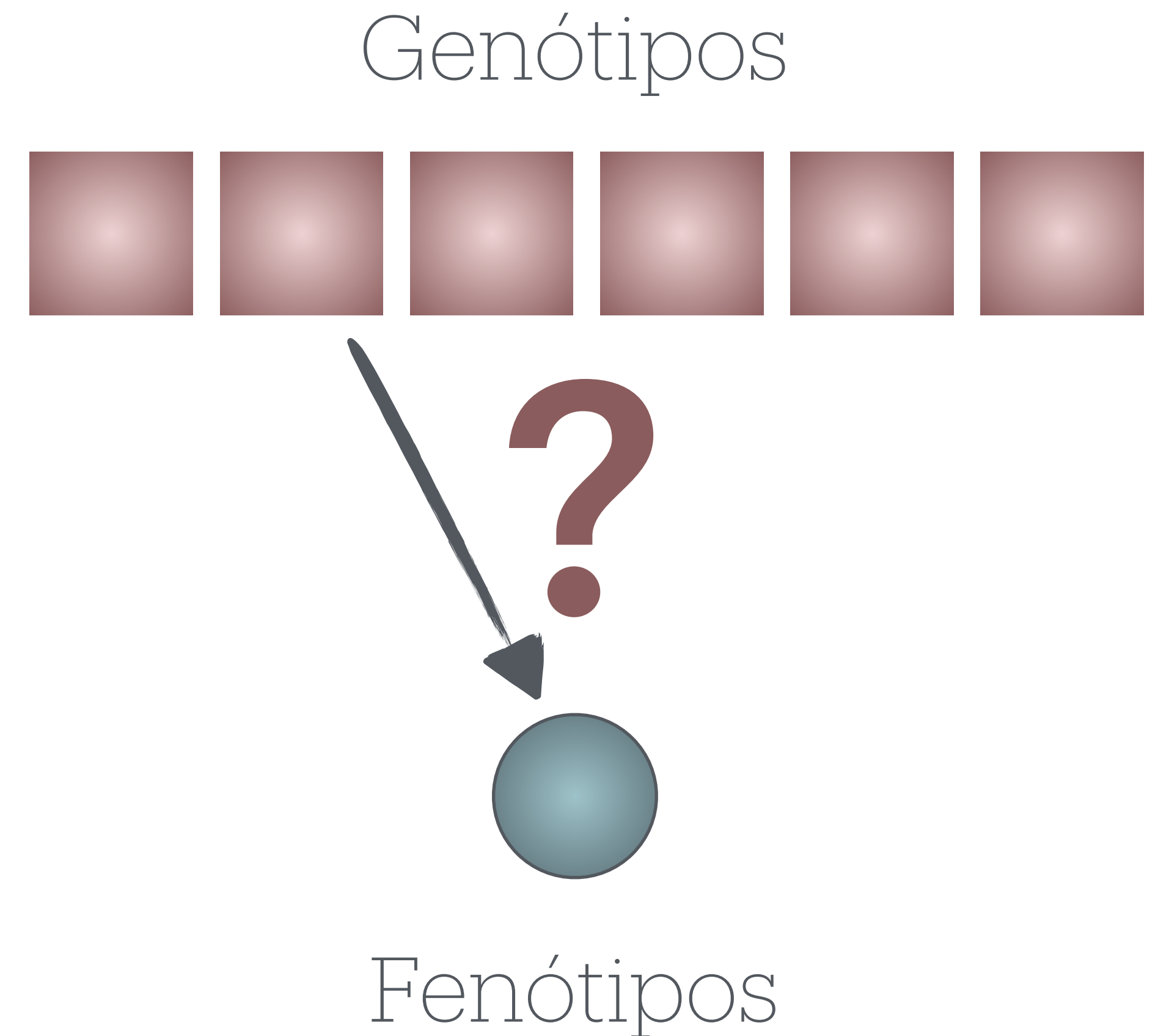
Como medir o efeito da variação em um locus no fenótipo?

- Dado uma população com fenótipos e genótipos conhecidos, como podemos identificar variantes que são responsáveis pela variação observada?



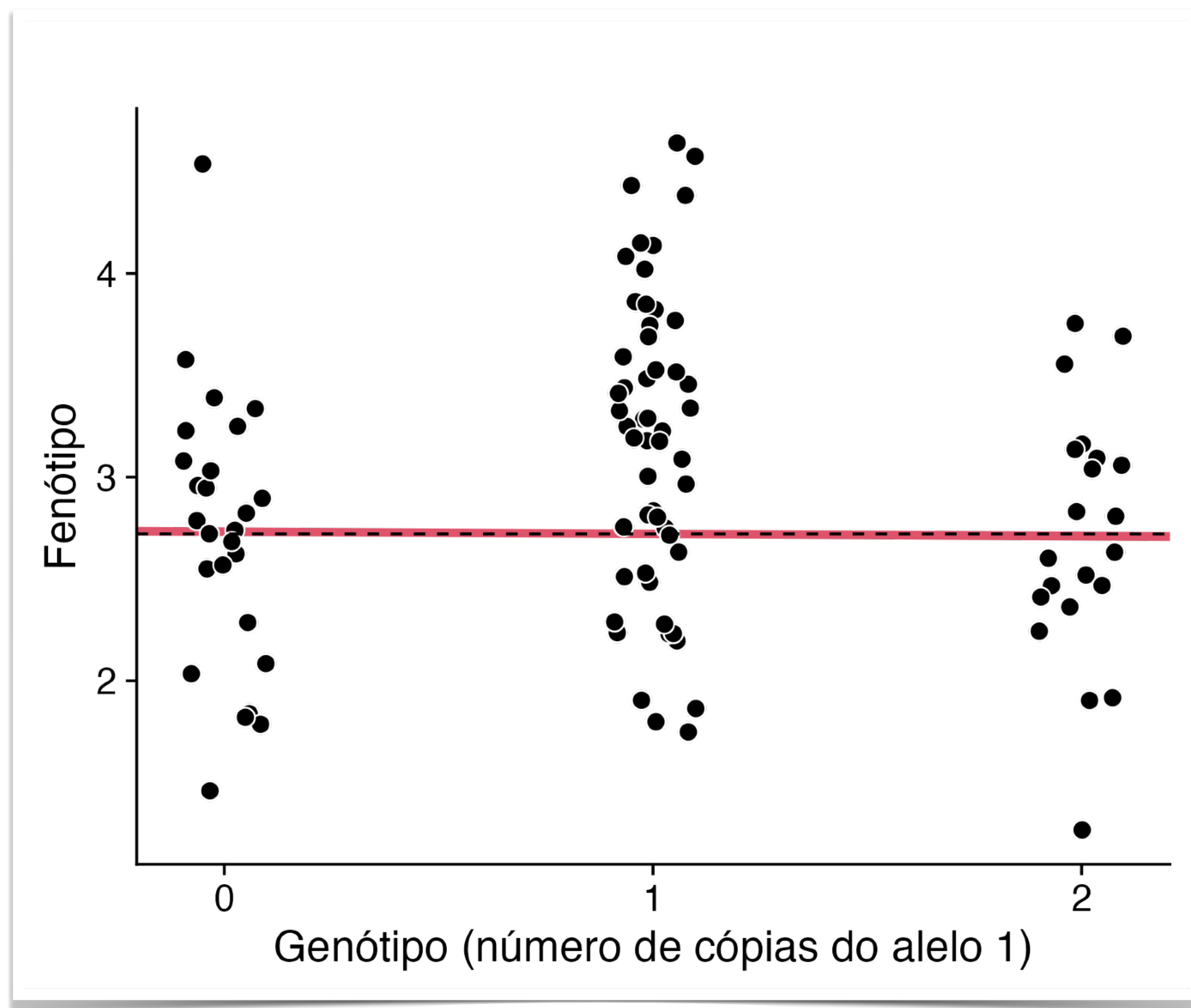
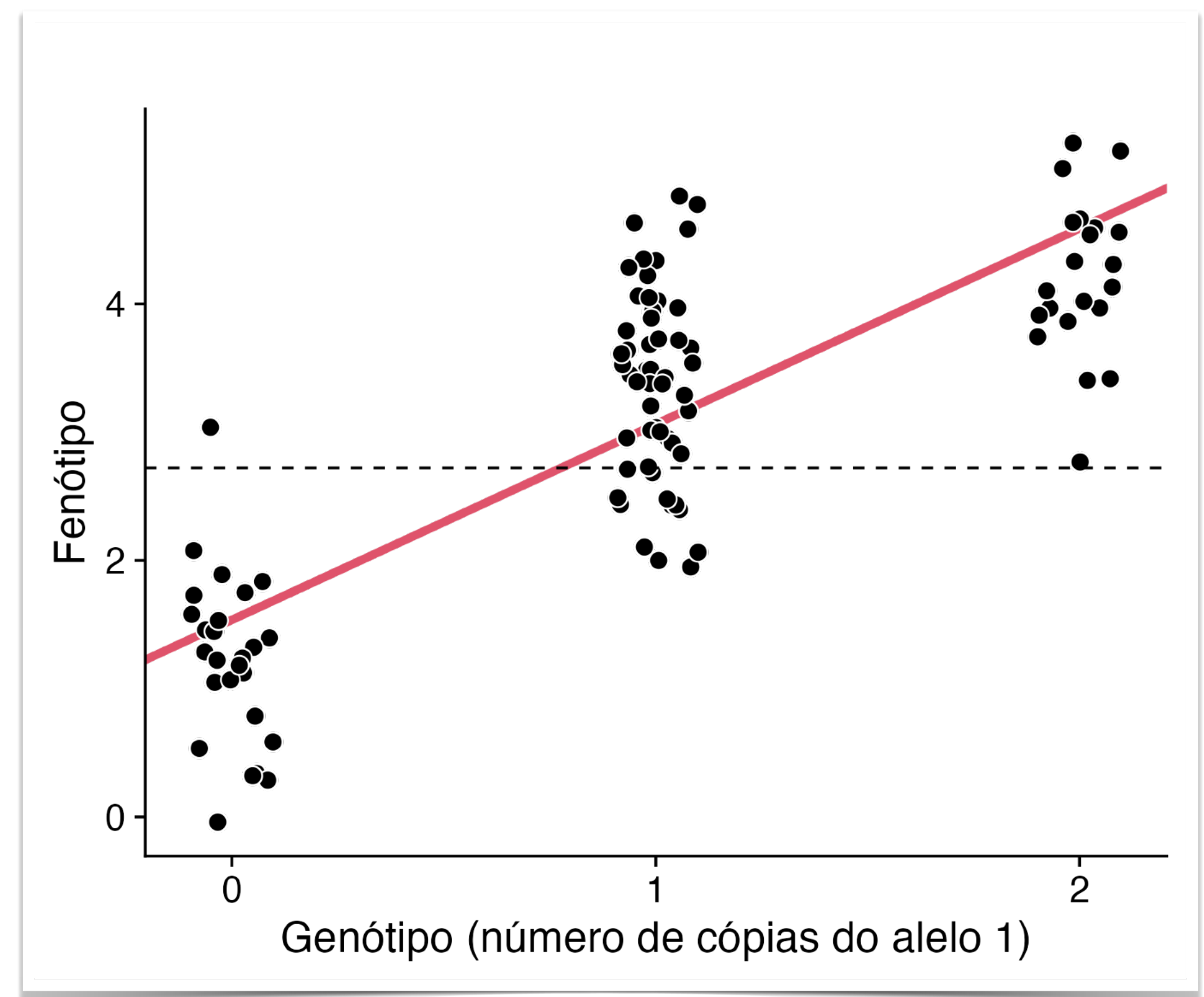
Como medir o efeito da variação em um locus no fenótipo?

- Dado uma população com fenótipos e genótipos conhecidos, como podemos identificar variantes que são responsáveis pela variação observada?
- **Resposta: Procurando associações estatísticas entre variantes genéticas e a distribuição dos fenótipos na população**



O modelo básico de um locus

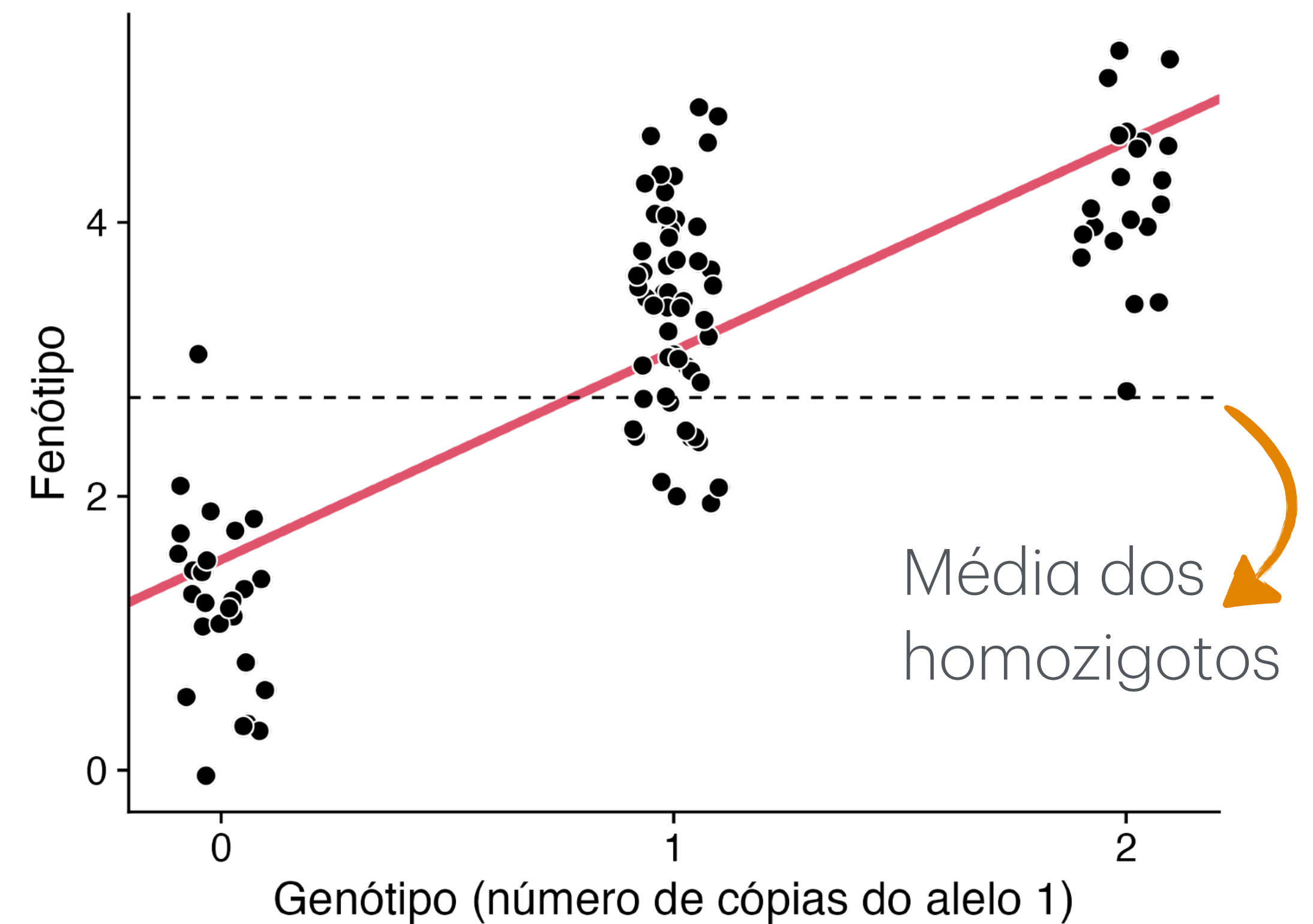
- Objetivo: quantificar a associação entre os genótipos em um locus e o valor de um fenótipo
- Único locus
- Dois alelos: A_1 e A_2



Efeitos genéticos em um locus

Efeitos aditivos e de dominância

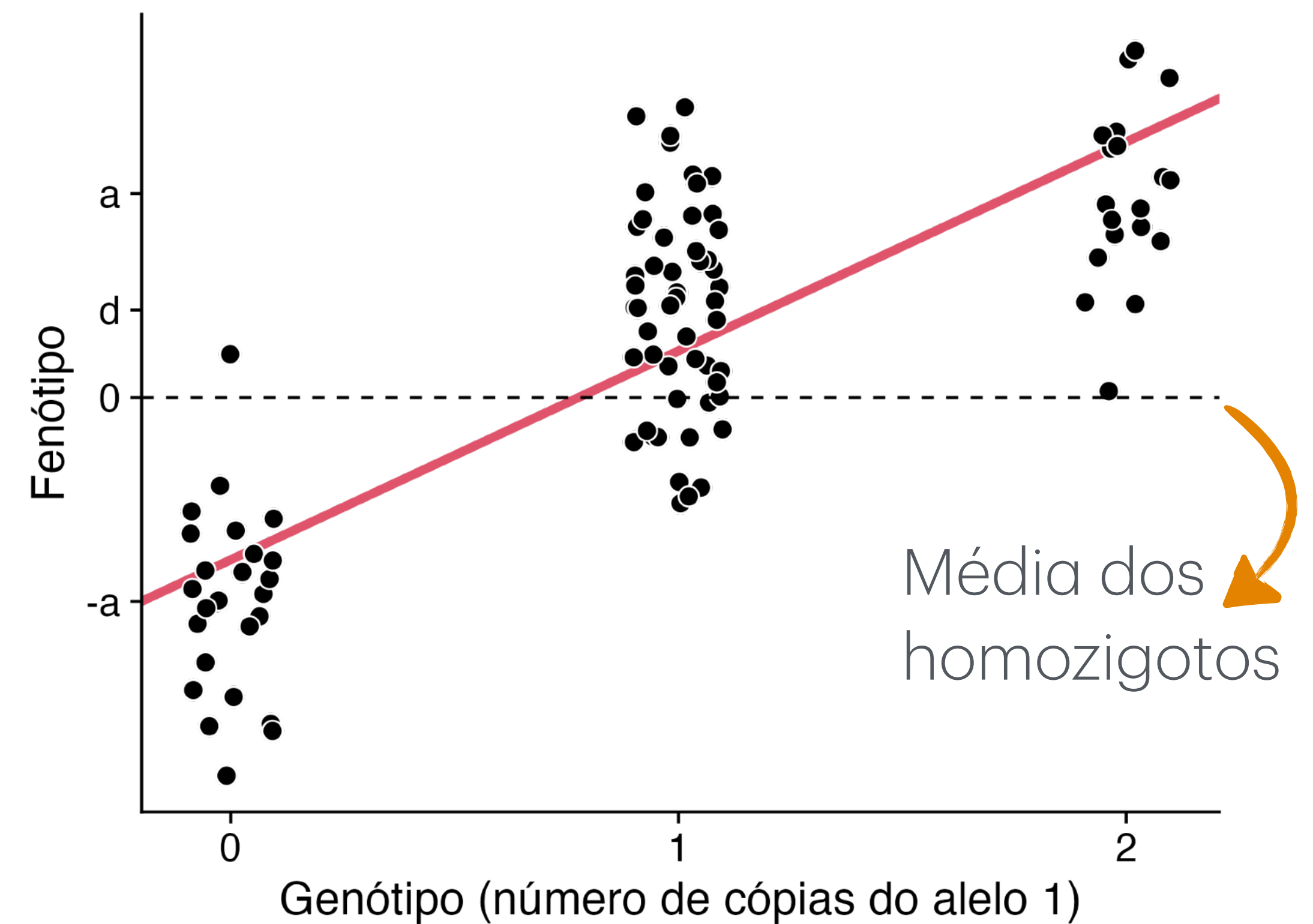
- Uma forma conveniente de quantificar um efeito genético é tomar a média dos homozigotos como um valor de referência



Efeitos genéticos em um locus

Efeitos aditivos e de dominância

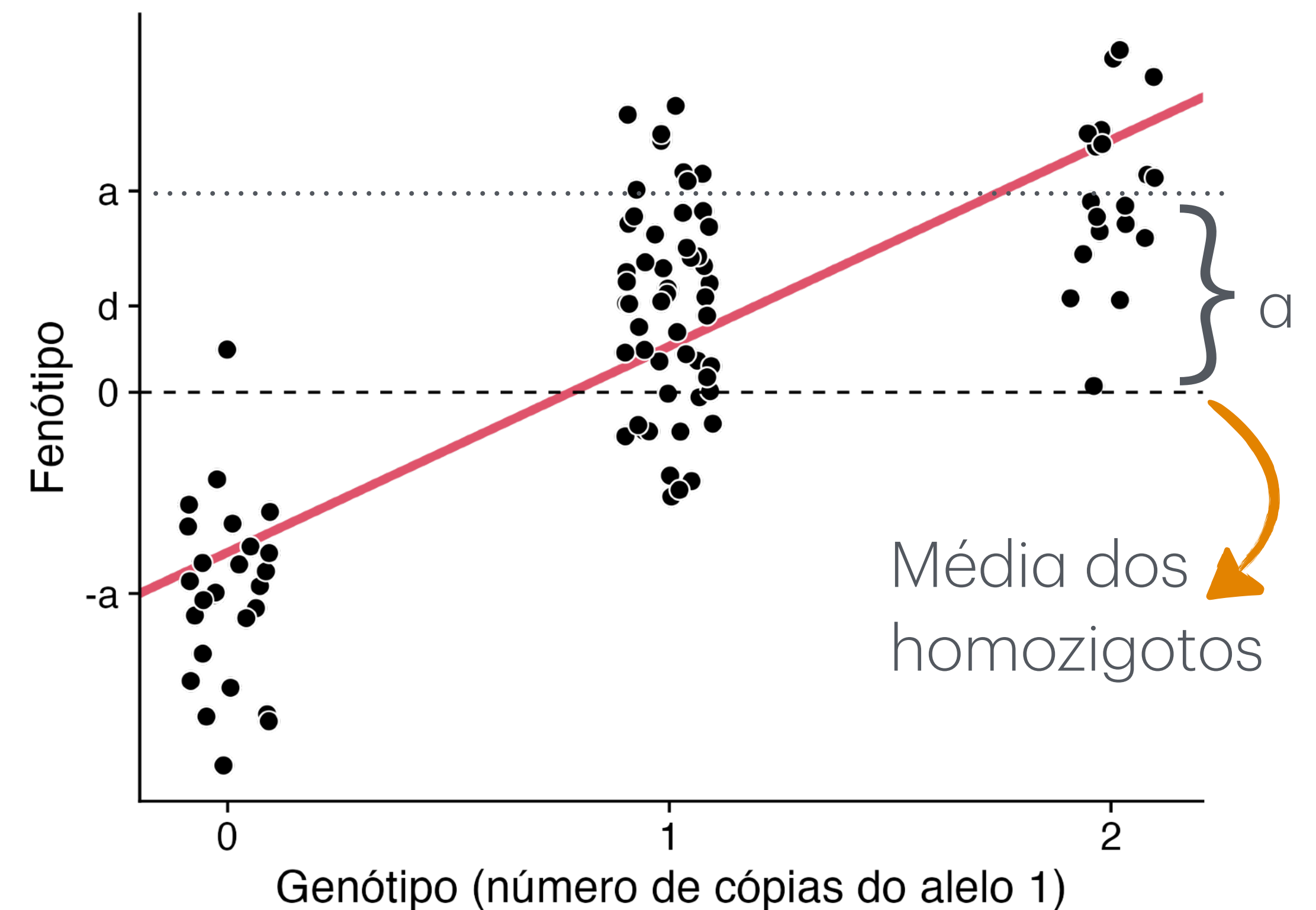
- Uma forma conveniente de quantificar um efeito genético é tomar a média dos homozigotos como um valor de referência



Efeitos genéticos em um locus

Efeitos aditivos e de dominância

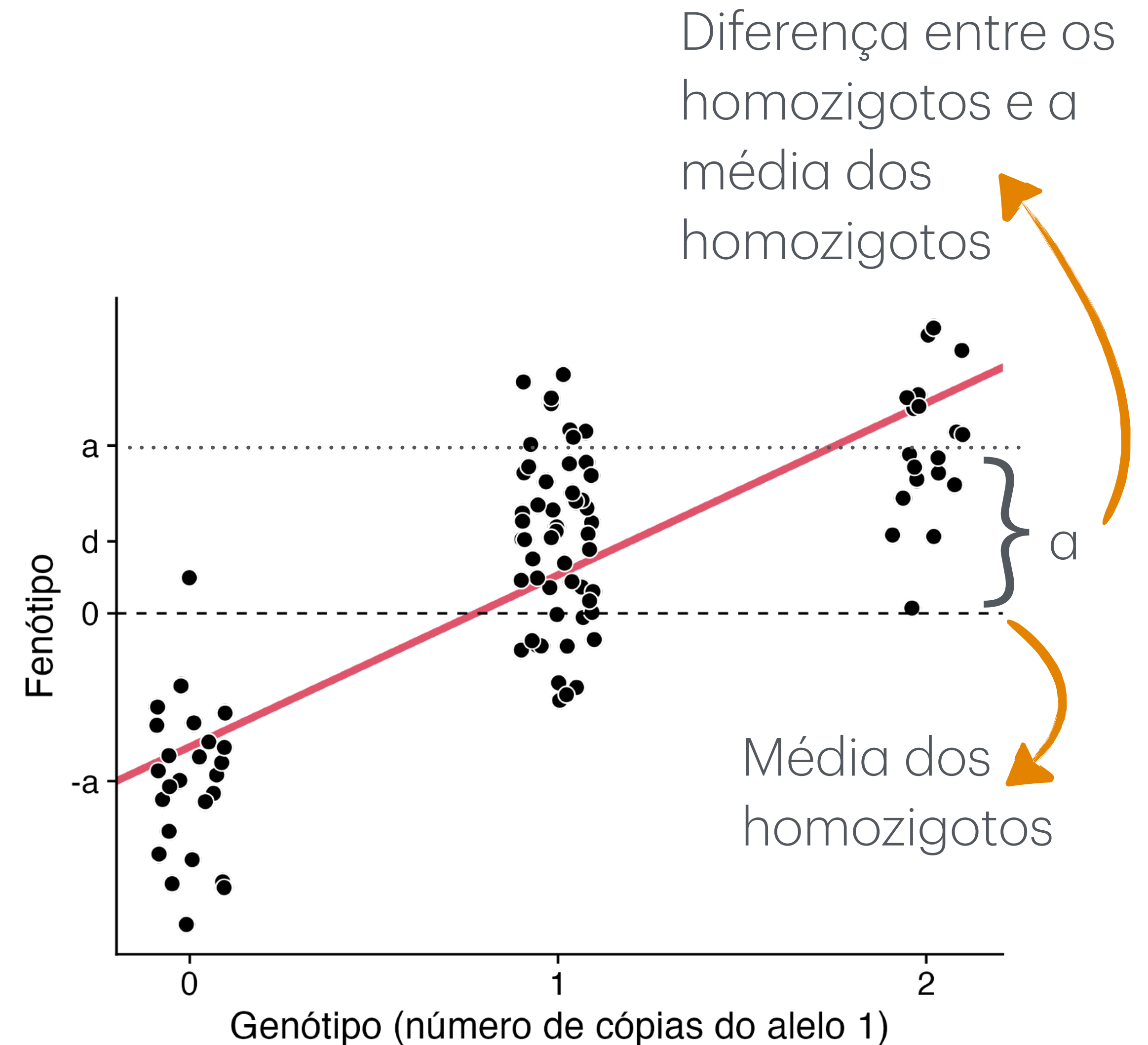
- Uma forma conveniente de quantificar um efeito genético é tomar a média dos homozigotos como um valor de referência
- O efeito de um locus no fenótipo pode ser decomposto em dois componentes
 - a : o efeito aditivo, ou valor genotípico aditivo



Efeitos genéticos em um locus

Efeitos aditivos e de dominância

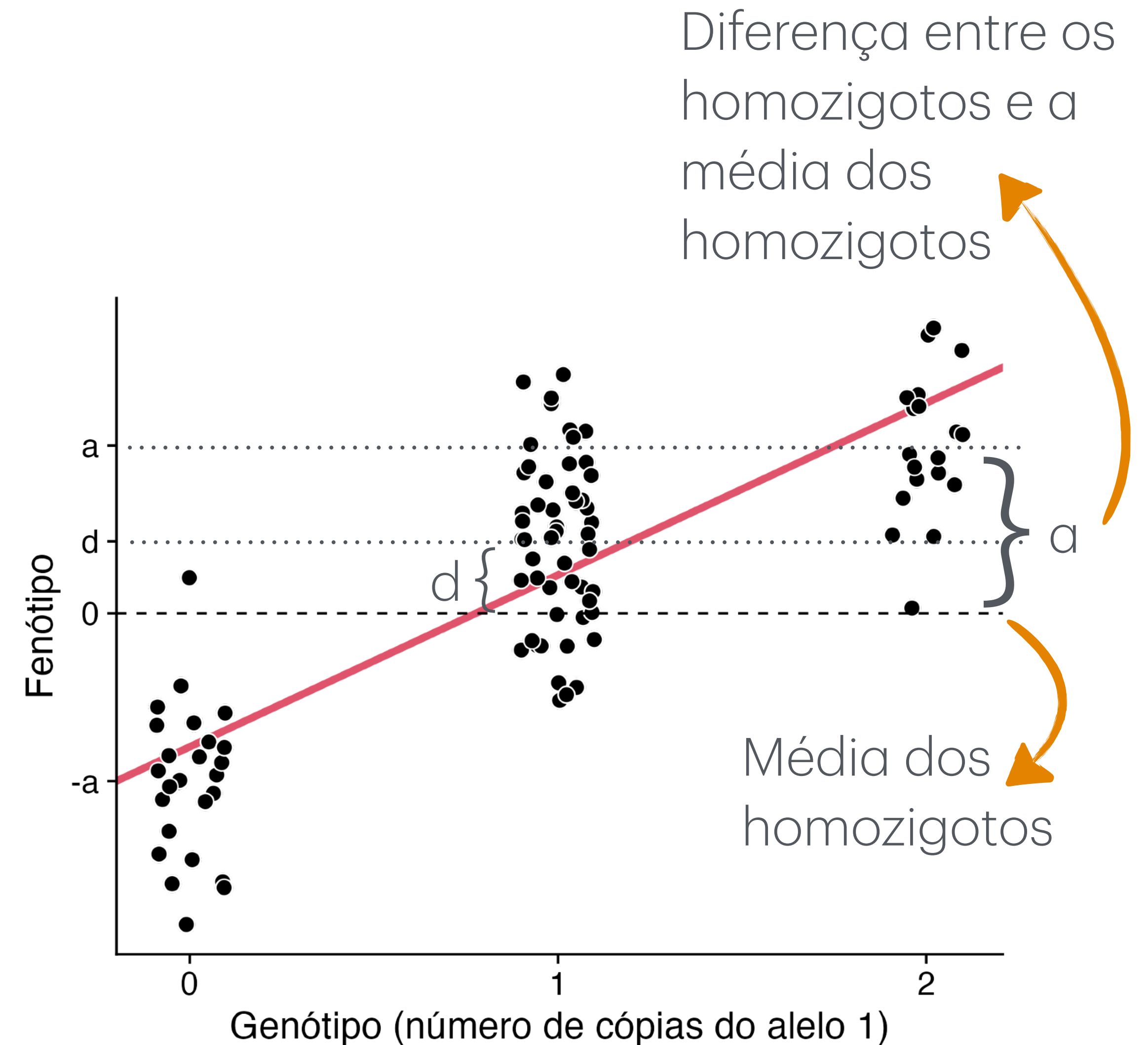
- Uma forma conveniente de quantificar um efeito genético é tomar a média dos homozigotos como um valor de referência
- O efeito de um locus no fenótipo pode ser decomposto em dois componentes
 - a : o efeito aditivo, ou valor genotípico aditivo



Efeitos genéticos em um locus

Efeitos aditivos e de dominância

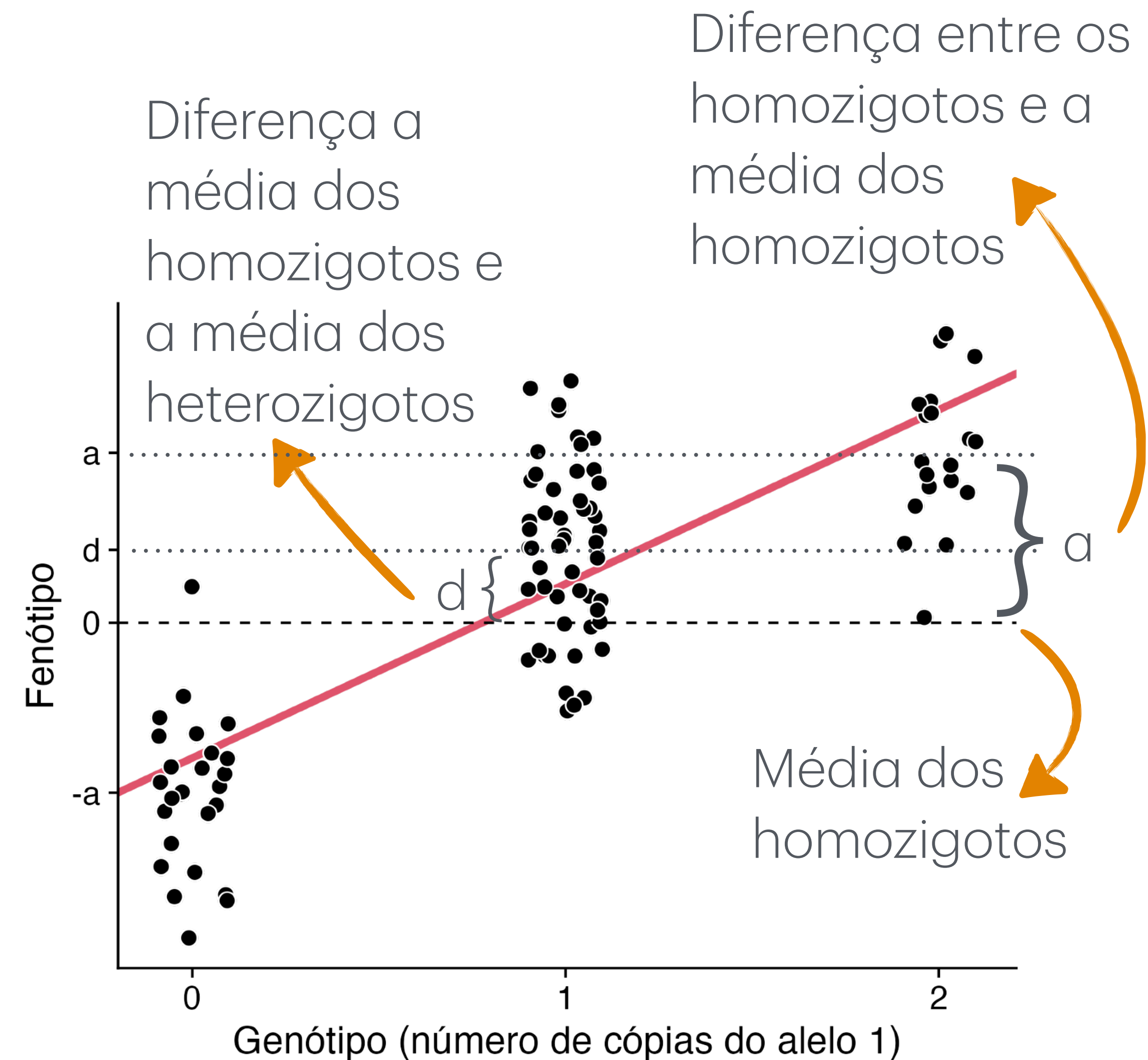
- Uma forma conveniente de quantificar um efeito genético é tomar a média dos homozigotos como um valor de referência
- O efeito de um locus no fenótipo pode ser decomposto em dois componentes
 - **a**: o efeito aditivo, ou valor genotípico aditivo
 - **d**: efeito de dominância, ou valor genotípico de dominância



Efeitos genéticos em um locus

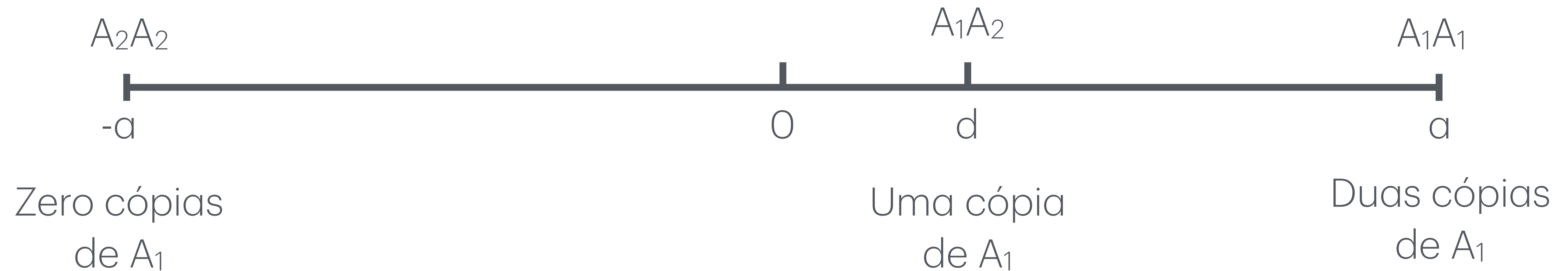
Efeitos aditivos e de dominância

- Uma forma conveniente de quantificar um efeito genético é tomar a média dos homozigotos como um valor de referência
- O efeito de um locus no fenótipo pode ser decomposto em dois componentes
 - **a**: o efeito aditivo, ou valor genotípico aditivo
 - **d**: efeito de dominância, ou valor genotípico de dominância



Efeitos genéticos em um locus

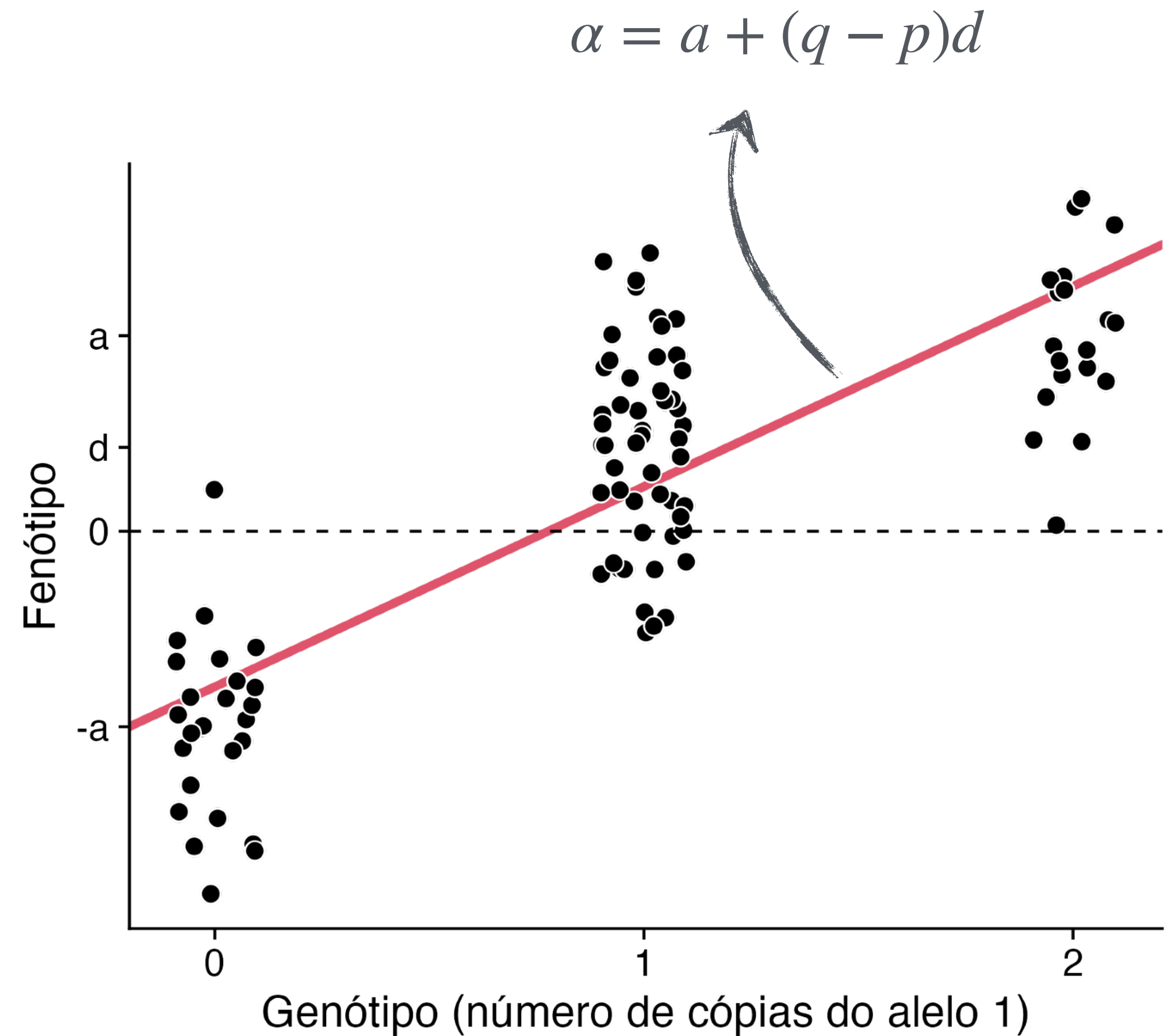
Todos os efeitos são definidos em relação à média dos homozigotos



O efeito genético aditivo

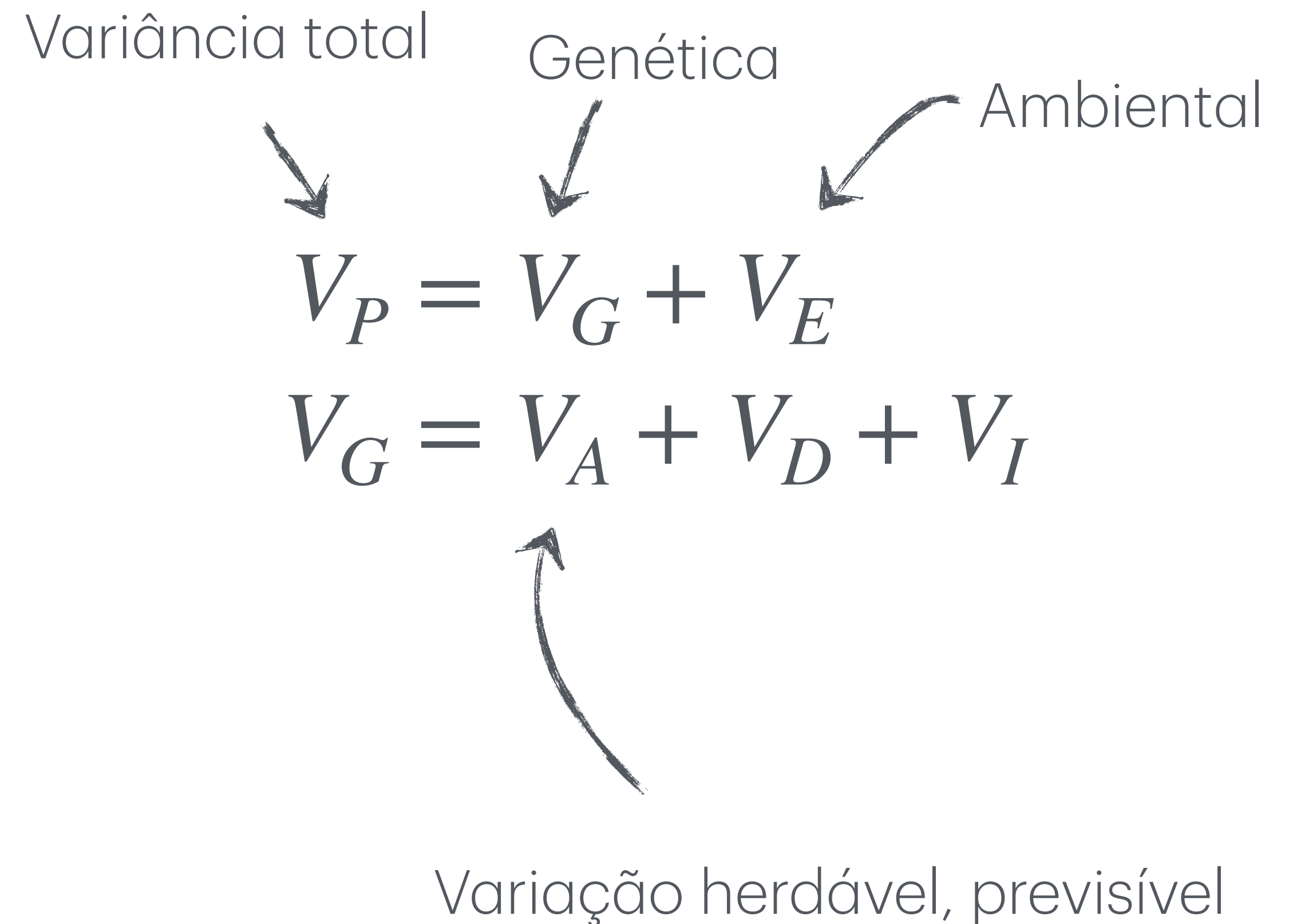
Origem da variação herdável

- A inclinação da reta de regressão entre genótipo e fenótipo é chamada efeito médio de substituição
- Efeito de substituição está relacionado a quão previsível é uma substituição alélica
 - Dois alelos: A_1 e A_2
 - Frequências alélicas: $f(A_1) = p$
 $f(A_2) = q$



Variância genética aditiva

- O efeito de substituição está relacionado com a variação genética aditiva
- A variância genética aditiva permite prever o efeito de seleção direcional, o efeito de deriva genética, e traz informação sobre quanta variação genética está presente na população



Modelo linear de mapeamento

Estimando efeitos genéticos aditivos e de dominância

- Dados:
 - y_i : fenótipo
 - A_i e D_i : componentes aditivos e de dominancia do genótipo
- Parâmetros:
 - a : efeito aditivo
 - d : efeito de dominância
 - μ : média global
 - ϵ_i : erro residual

Exemplo de genótipos:

ID	A1	D1	A2	D2	A3	D3
55	1	0	1	0	1	0
179	0	1	0	1	1	0
640	1	0	1	0	1	0
641	1	0	1	0	1	0
642	1	0	1	0	1	0
647	1	0	1	0	1	0
649	0	1	0	1	1	0
650	0	1	0	1	1	0
652	1	0	1	0	1	0
653	1	0	1	0	1	0
655	1	0	1	0	1	0
660	1	0	1	0	1	0
56	0	1	0	1	0	1
182	-1	0	-1	0	-1	0

Modelo linear de mapeamento

Estimando efeitos genéticos aditivos e de dominância

- Dados:

- y_i : fenótipo
- A_i e D_i : componentes aditivos e de dominancia do genótipo

- Parâmetros:

- a : efeito aditivo
- d : efeito de dominância
- μ : média global
- ϵ_i : erro residual

Modelo de mapeamento:

$$y_i = \mu + aA_i + dD_i + \epsilon_i$$

Exemplo de genótipos:

ID	A1	D1	A2	D2	A3	D3
55	1	0	1	0	1	0
179	0	1	0	1	1	0
640	1	0	1	0	1	0
641	1	0	1	0	1	0
642	1	0	1	0	1	0
647	1	0	1	0	1	0
649	0	1	0	1	1	0
650	0	1	0	1	1	0
652	1	0	1	0	1	0
653	1	0	1	0	1	0
655	1	0	1	0	1	0
660	1	0	1	0	1	0
56	0	1	0	1	0	1
182	-1	0	-1	0	-1	0

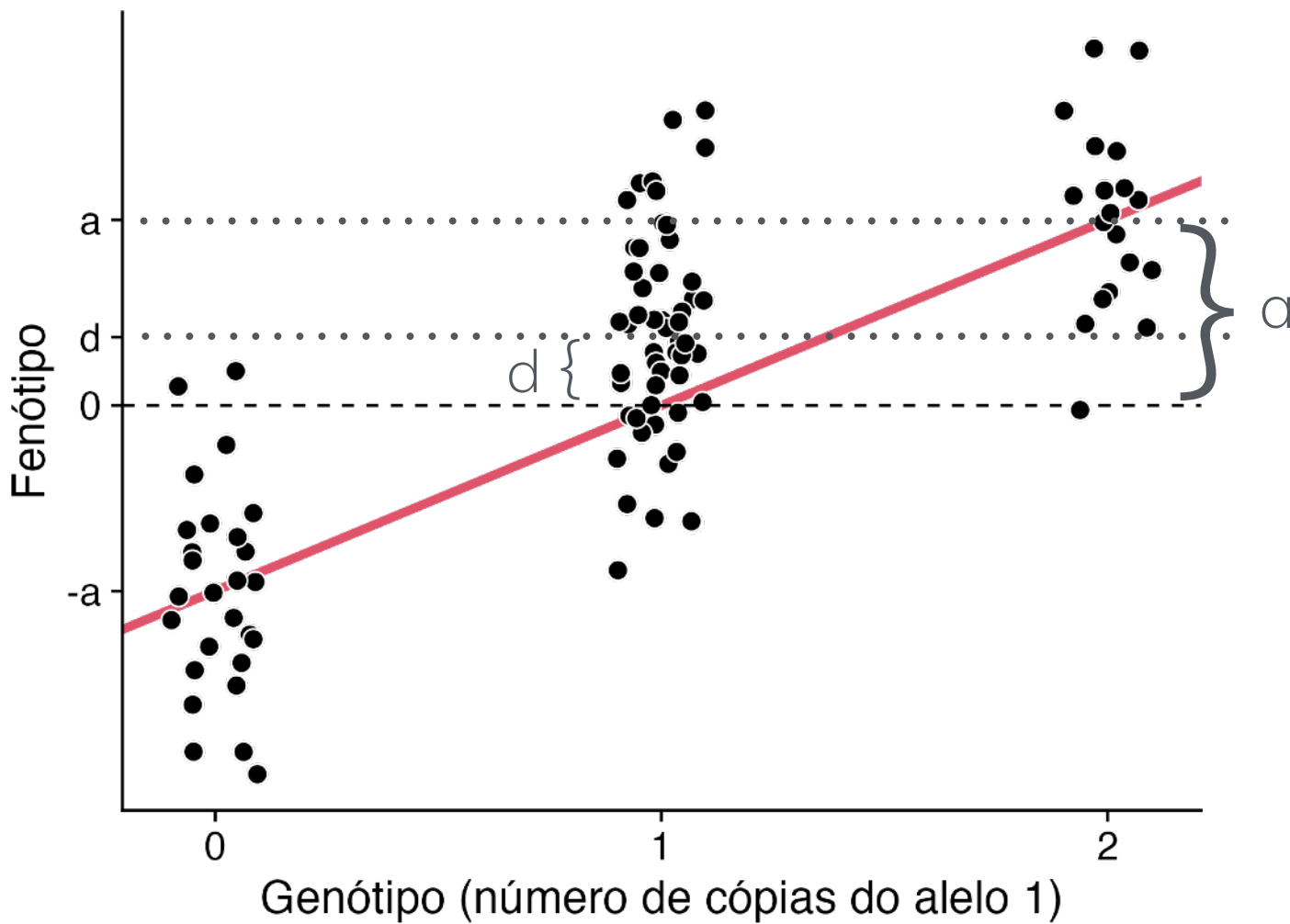
Modelo linear de mapeamento

Estimando efeitos genéticos aditivos e de dominância

- Dados:
 - y_i : fenótipo
 - A_i e D_i : componentes aditivos e de dominancia do genótipo
- Parâmetros:
 - a : efeito aditivo
 - d : efeito de dominância
 - μ : média global
 - ϵ_i : erro residual

Modelo de mapeamento:

$$y_i = \mu + aA_i + dD_i + \epsilon_i$$



Exemplo de genótipos:

ID	A1	D1	A2	D2	A3	D3
55	1	0	1	0	1	0
179	0	1	0	1	1	0
640	1	0	1	0	1	0
641	1	0	1	0	1	0
642	1	0	1	0	1	0
647	1	0	1	0	1	0
649	0	1	0	1	1	0
650	0	1	0	1	1	0
652	1	0	1	0	1	0
653	1	0	1	0	1	0
655	1	0	1	0	1	0
660	1	0	1	0	1	0
56	0	1	0	1	0	1
182	-1	0	-1	0	-1	0

Localizando os locus de interesse

Coeficientes e significância

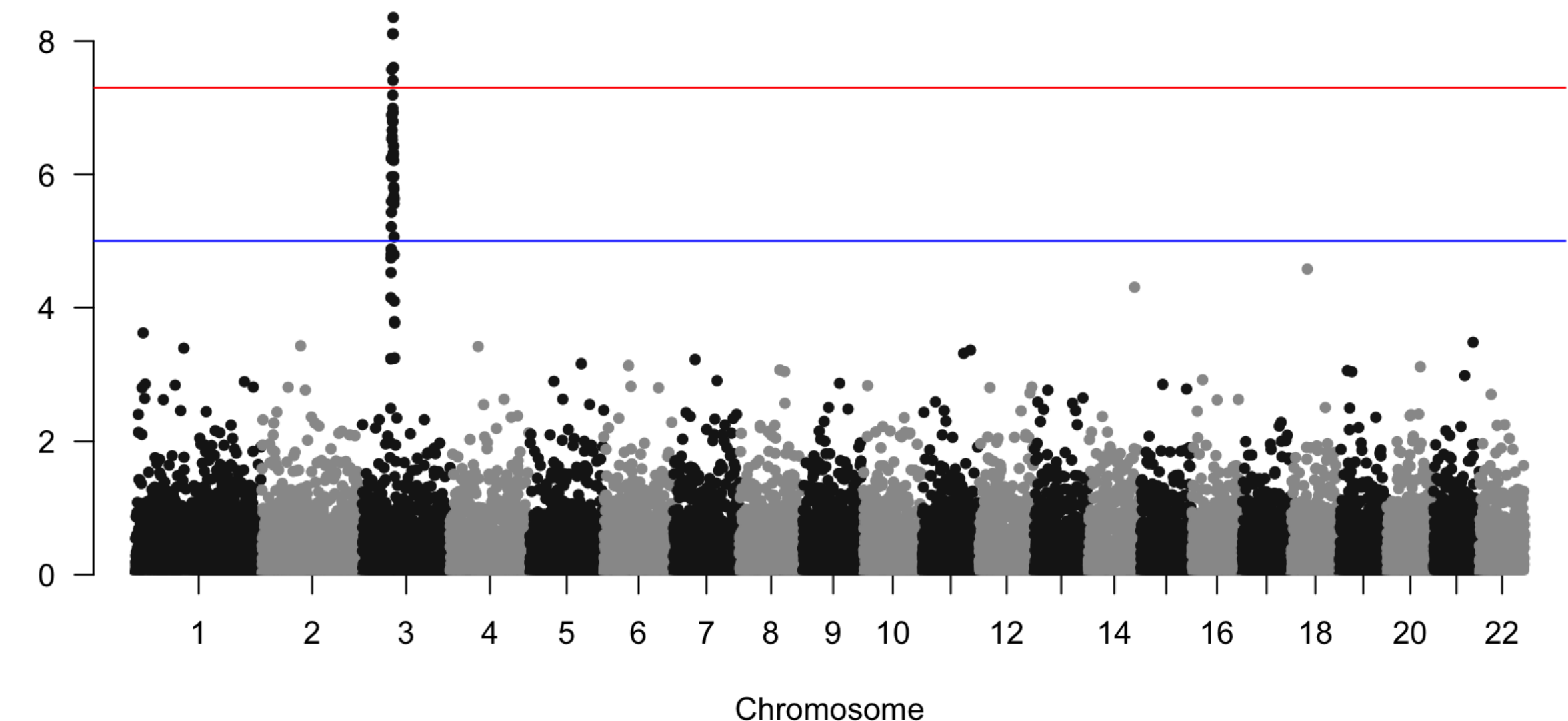
Modelo de mapeamento aditivo:

$$y_i = \mu + aA_i + \epsilon_i$$

Coeficiente ligado ao genótipo

O modelo é aplicado a todos os loci variáveis independentemente, e registramos o p-valor associado a cada locus, testando a hipótese nula que o coeficiente é zero.

Manhattan plot



Localização no genoma

Localizando os locus de interesse

Coeficientes e significância

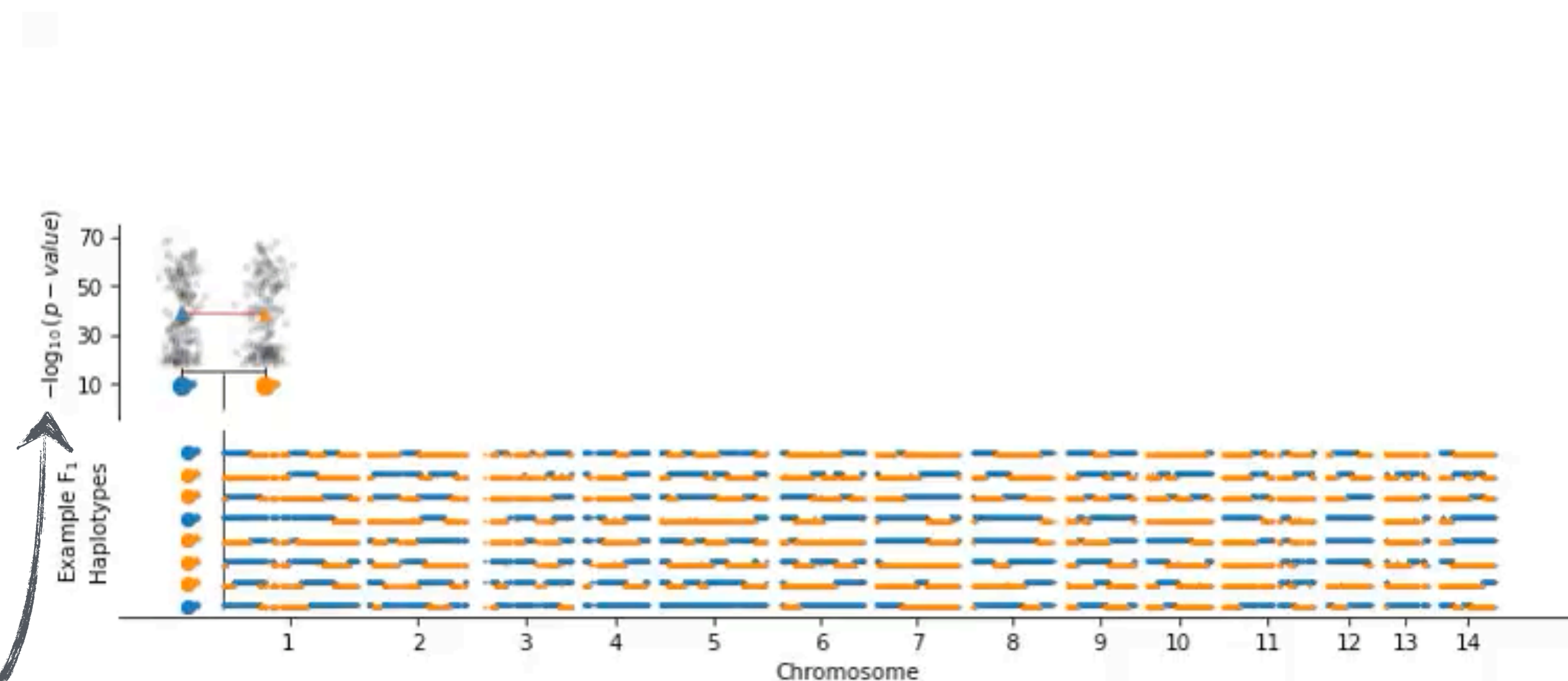
Modelo de mapeamento aditivo:

$$y_i = \mu + aA_i + \epsilon_i$$

Coeficiente ligado ao genótipo

O modelo é aplicado a todos os loci variáveis independentemente, e registramos o p-valor associado a cada locus, testando a hipótese nula que o coeficiente é zero.

Manhattan plot



Localização no genoma

Localizando os locus de interesse

Coeficientes e significância

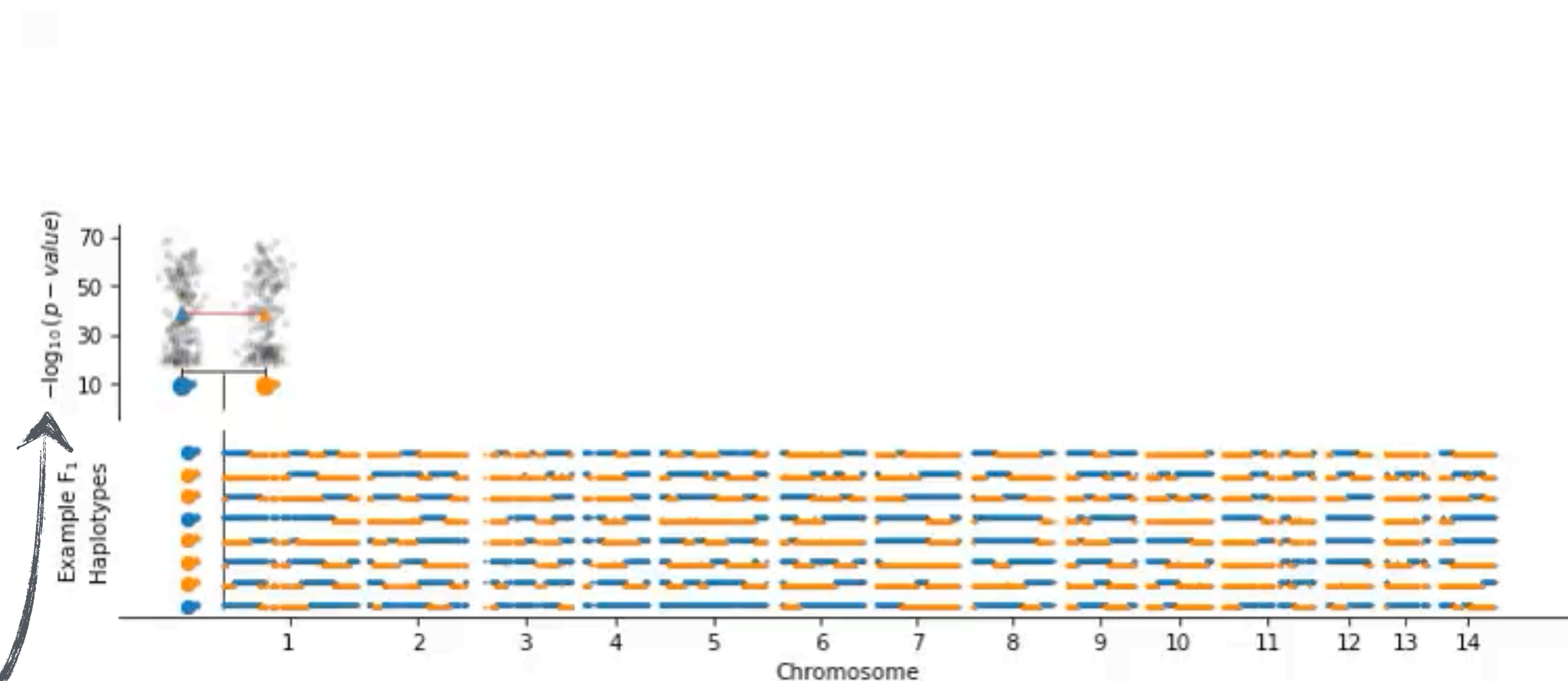
Modelo de mapeamento aditivo:

$$y_i = \mu + aA_i + \epsilon_i$$

Coeficiente ligado ao genótipo

O modelo é aplicado a todos os loci variáveis independentemente, e registramos o p-valor associado a cada locus, testando a hipótese nula que o coeficiente é zero.

Manhattan plot

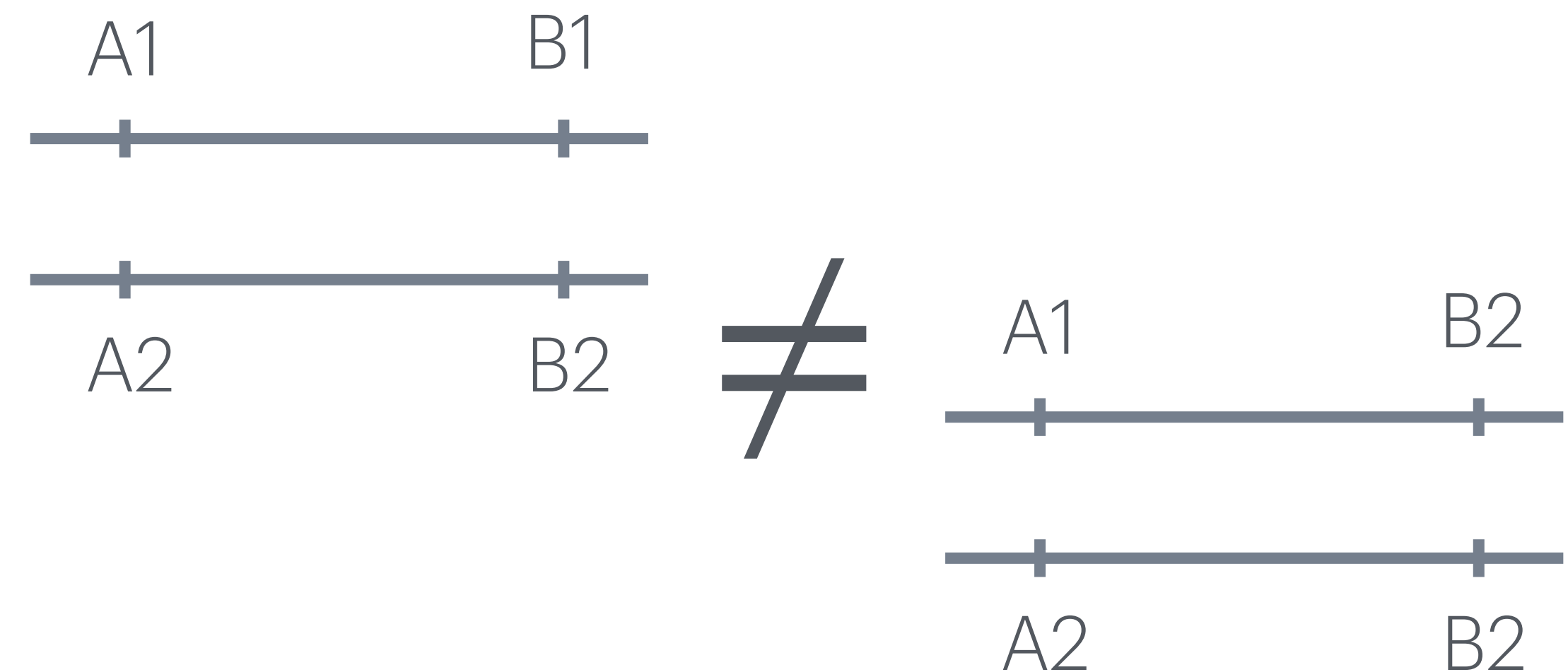
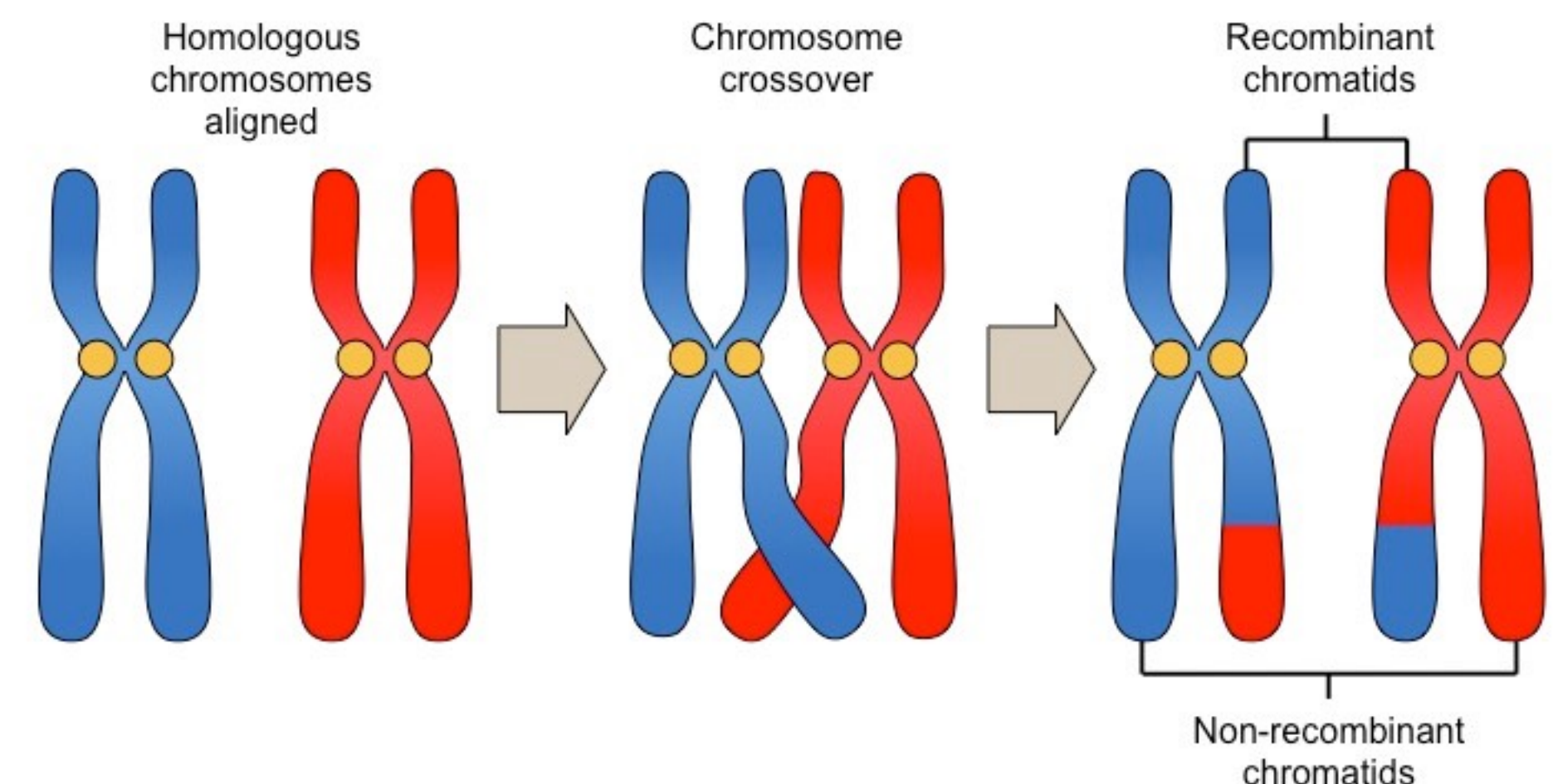


Localização no genoma

Resolução de mapeamento e LD

Desequilíbrio de ligação (LD) é fundamental para o mapeamento genético

- Desequilíbrio de ligação se refere a uma associação estatística entre as frequências alélicas de variantes em loci diferentes no genoma
- O LD é removido por recombinação, e introduzido por migração e pelo processo de mutação

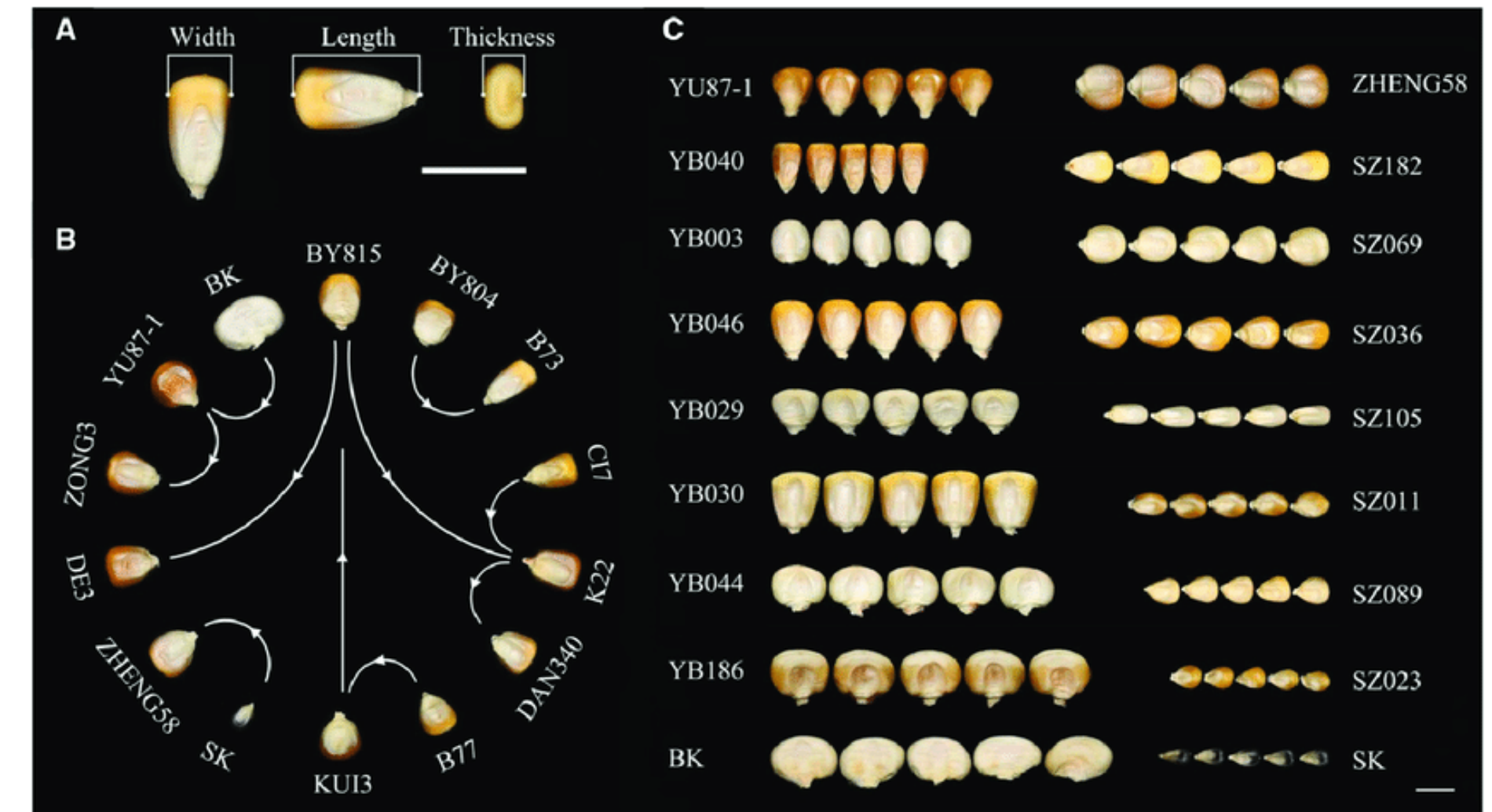


Mapeamento genético em laboratório

Cruzamentos experimentais

QTLs, F2 cross, RIL, AIL, NIL, ...

- Cruzamentos permitem gerar populações experimentais com variação genética ligada ao fenótipo de interesse
- Tipo diferentes de cruzamentos podem ser usado dependendo da pergunta e do sistema biológico
- A maioria dos cruzamento usa a recombinação para gerar variação genética num padrão que é útil para o experimento



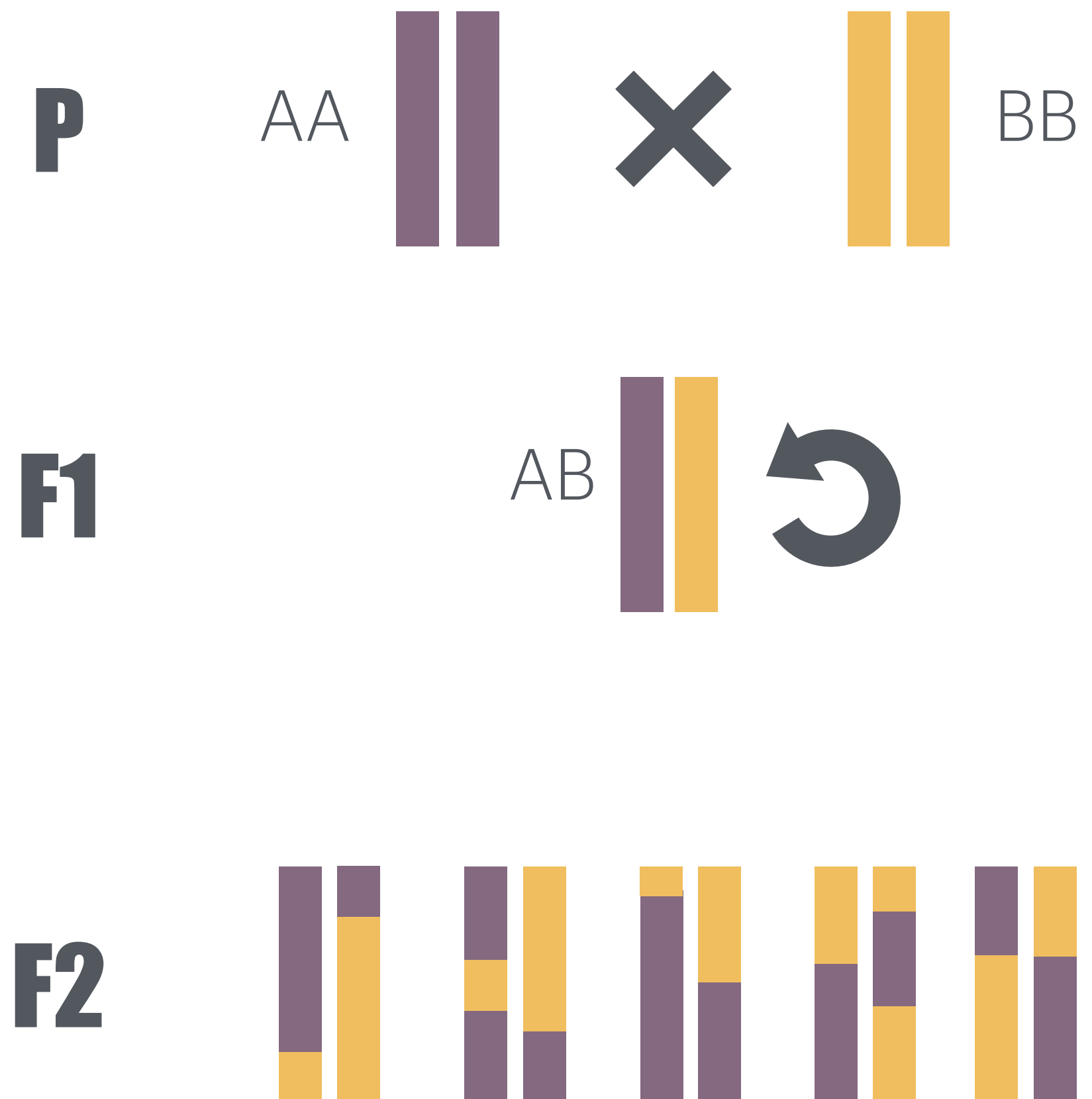
Liu, J. *et al.* The conserved and unique genetic architecture of kernel size and weight in maize and rice. *Plant Physiol.* **175**, 774–785 (2017).



Cruzamento F2

Mapeamento de QTL

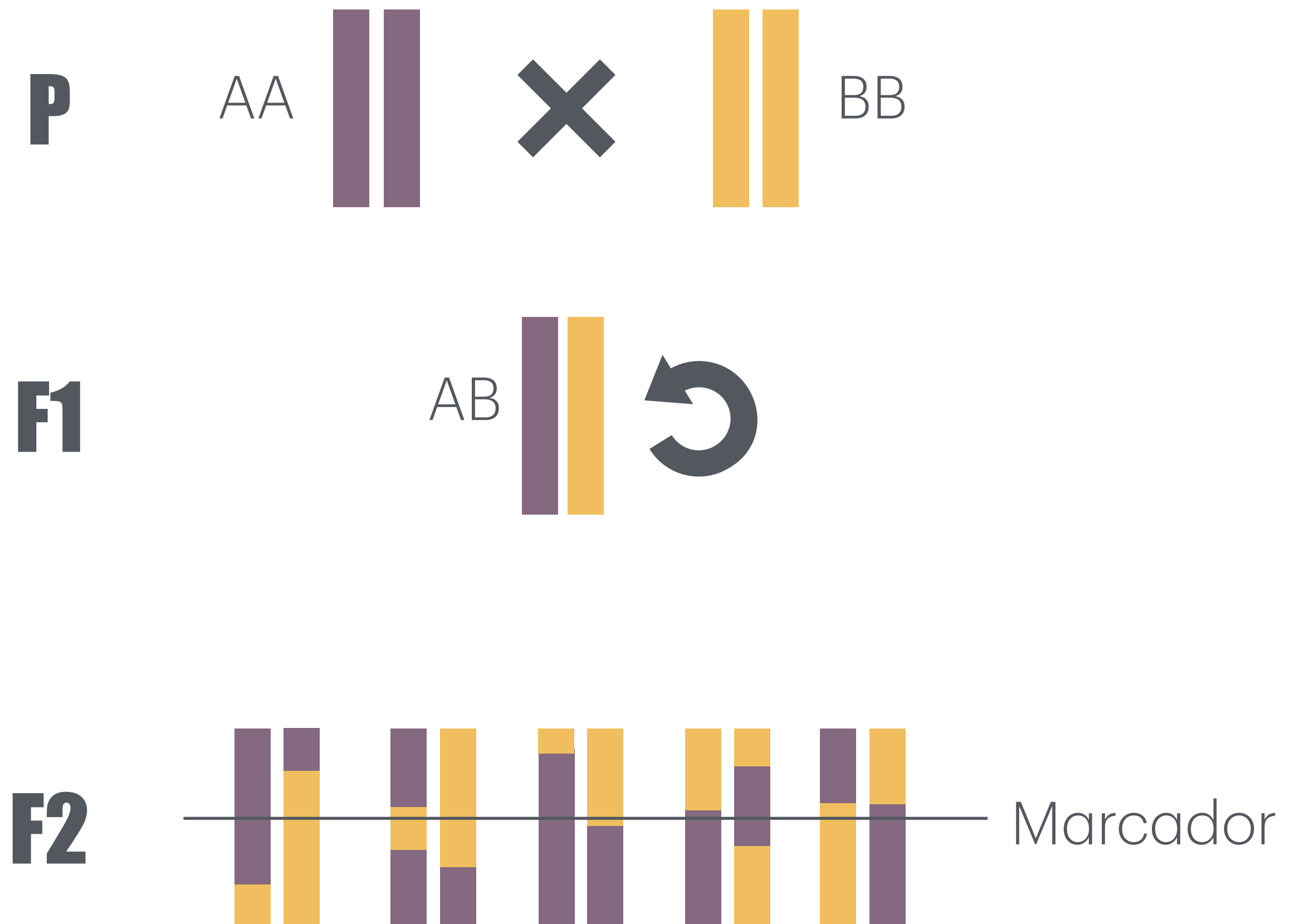
- Começa com linhagens parentais que diferem em algum fenótipo de interesse
- Usa linhagens Parentais para criar uma F1 heterozigota
- Usa a recombinação na F1 para gerar uma F2 variável, que permite localizar regiões associadas ao fenótipo de interesse



Cruzamento F2

Mapeamento de QTL

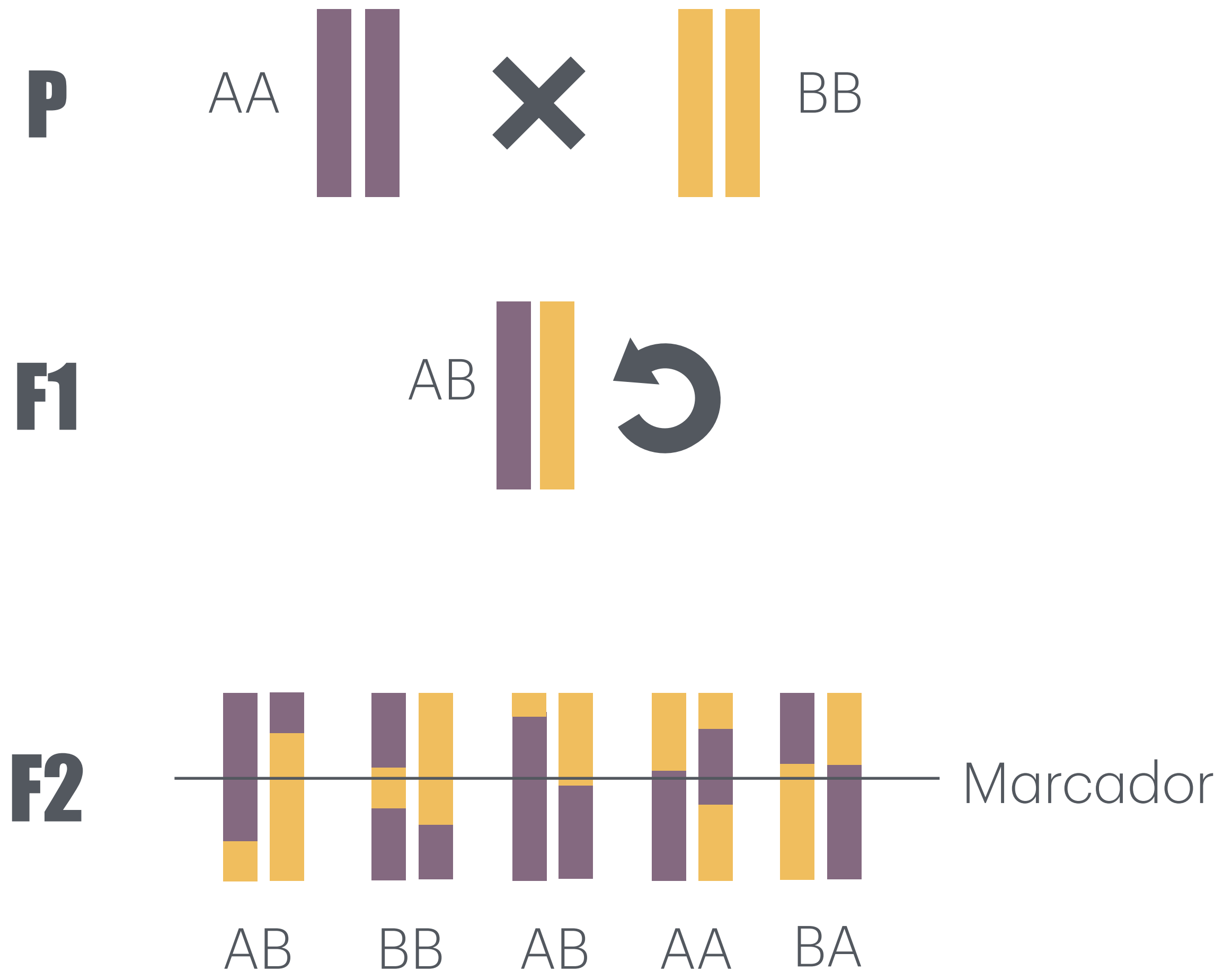
- Começa com linhagens parentais que diferem em algum fenótipo de interesse
- Usa linhagens Parentais para criar uma F1 heterozigota
- Usa a recombinação na F1 para gerar uma F2 variável, que permite localizar regiões associadas ao fenótipo de interesse



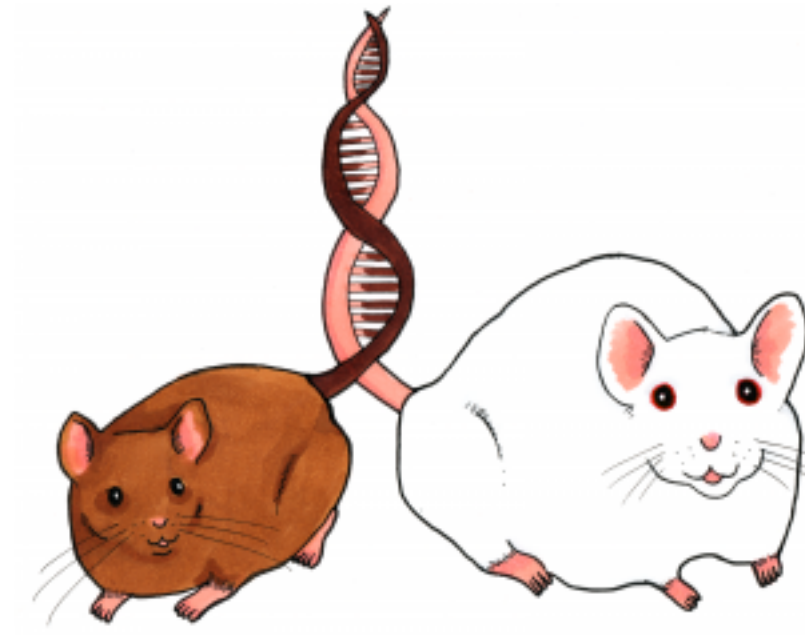
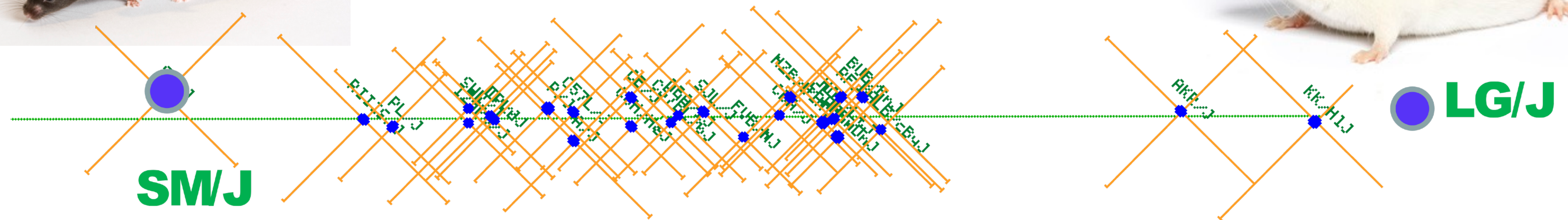
Cruzamento F2

Mapeamento de QTL

- Começa com linhagens parentais que diferem em algum fenótipo de interesse
- Usa linhagens Parentais para criar uma F1 heterozigota
- Usa a recombinação na F1 para gerar uma F2 variável, que permite localizar regiões associadas ao fenótipo de interesse



Cruzamento LG/J - SM/J



Tamanho

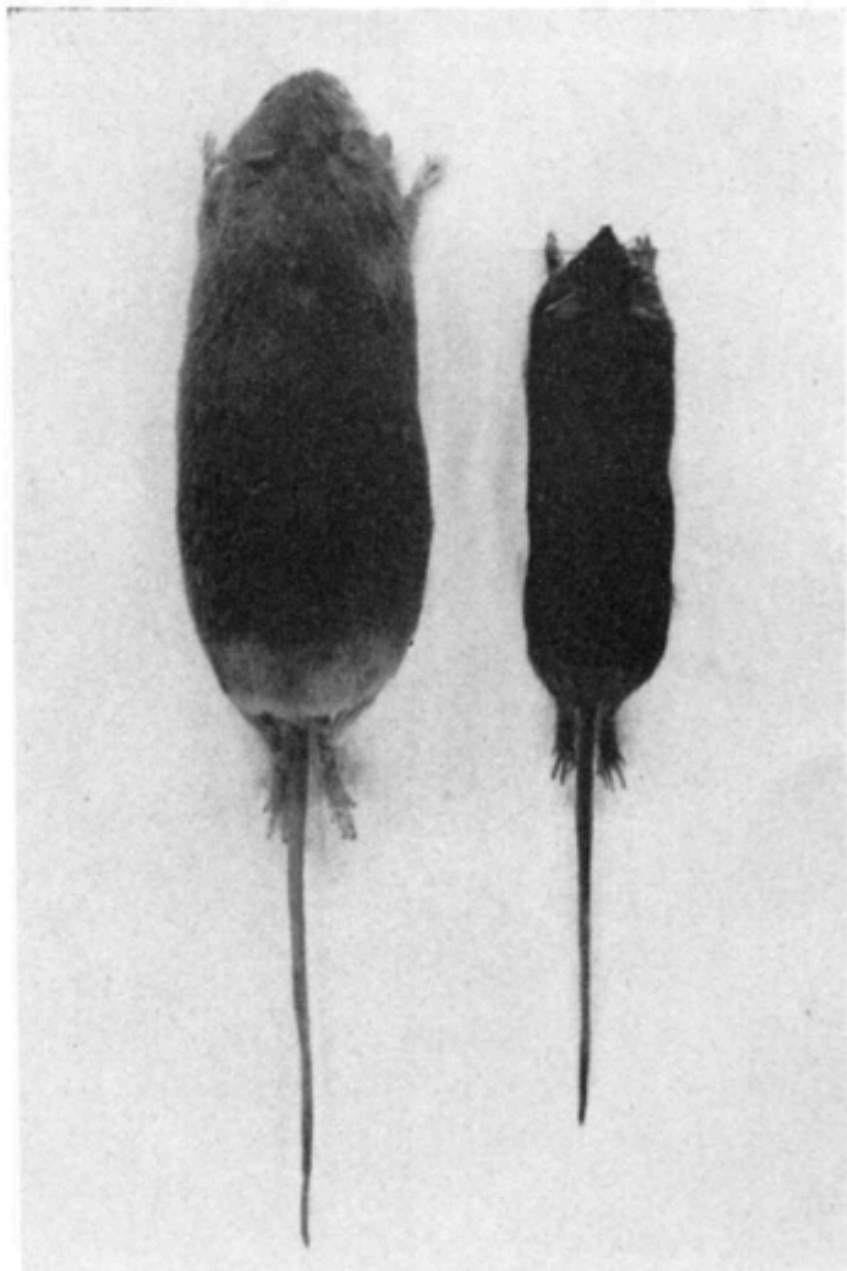
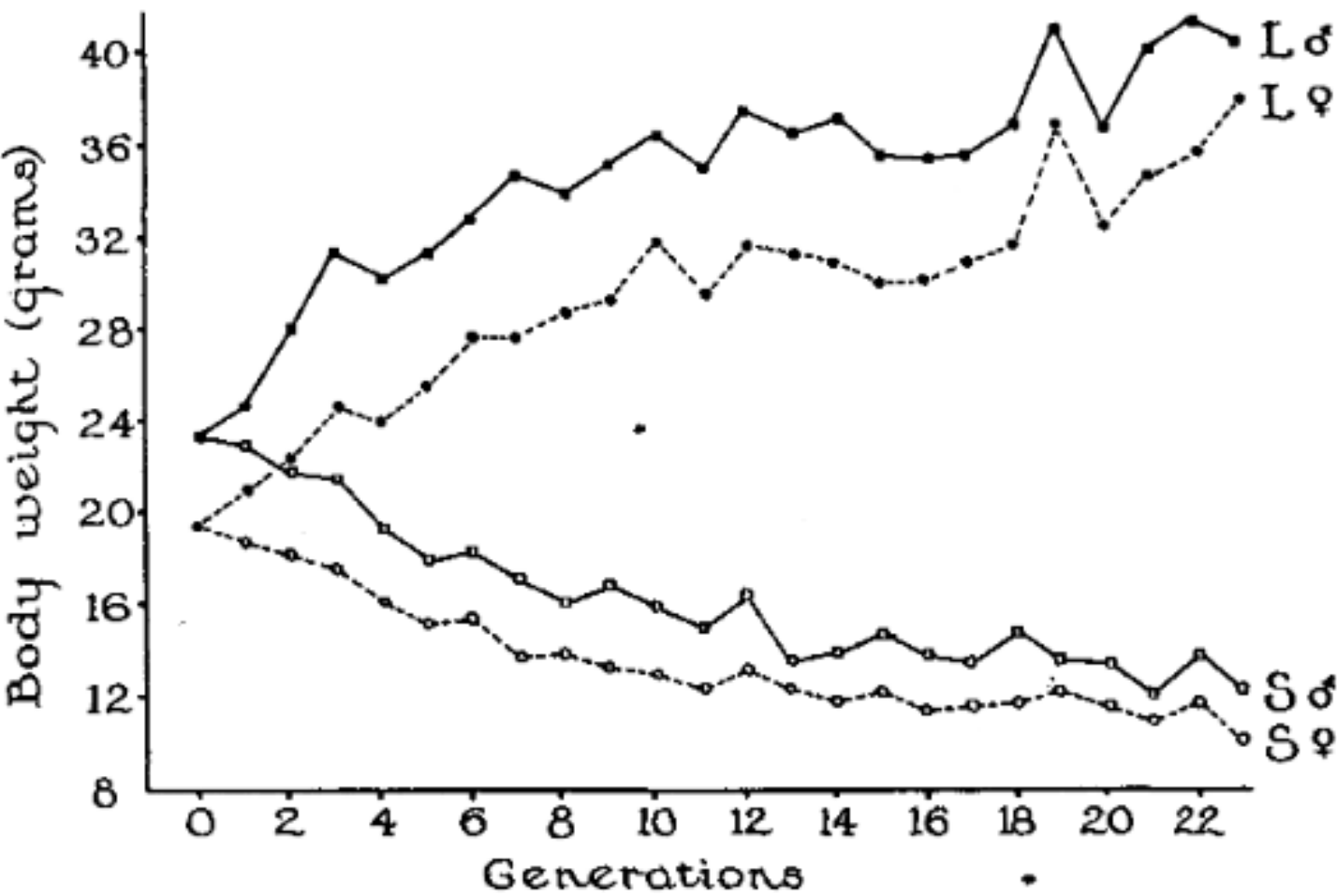
Produtos de seleção artificial para tamanho

SELECTION FOR SMALL AND LARGE BODY SIZE IN THE HOUSE MOUSE*1

JOHN W. MACARTHUR

Department of Biology, Marlboro College, Marlboro, Vermont²

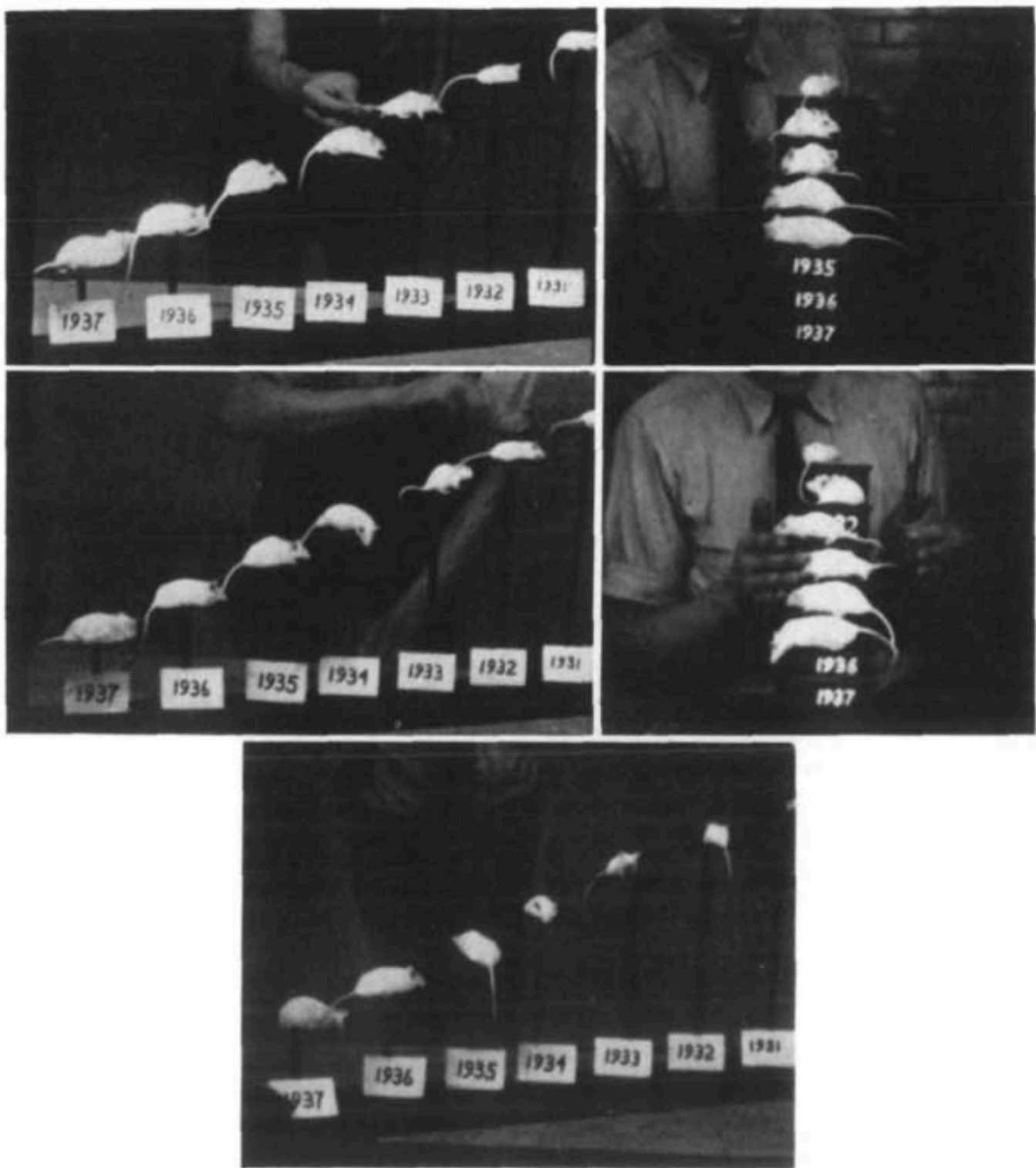
Received July 7, 1948



A STUDY OF THE INHERITANCE OF BODY WEIGHT IN THE ALBINO MOUSE BY SELECTION

H. D. GOODALE

Mount Hope Farm
Williamstown, Massachusetts



SIX YEARS' PROGRESS IN PROGENY SELECTION

FIGURE 3. Three month old males of the 23rd generation; weight of the small 16.2 gms, of the large 58.0 gms.

Fenótipos e loci identificados

JOURNAL ARTICLE

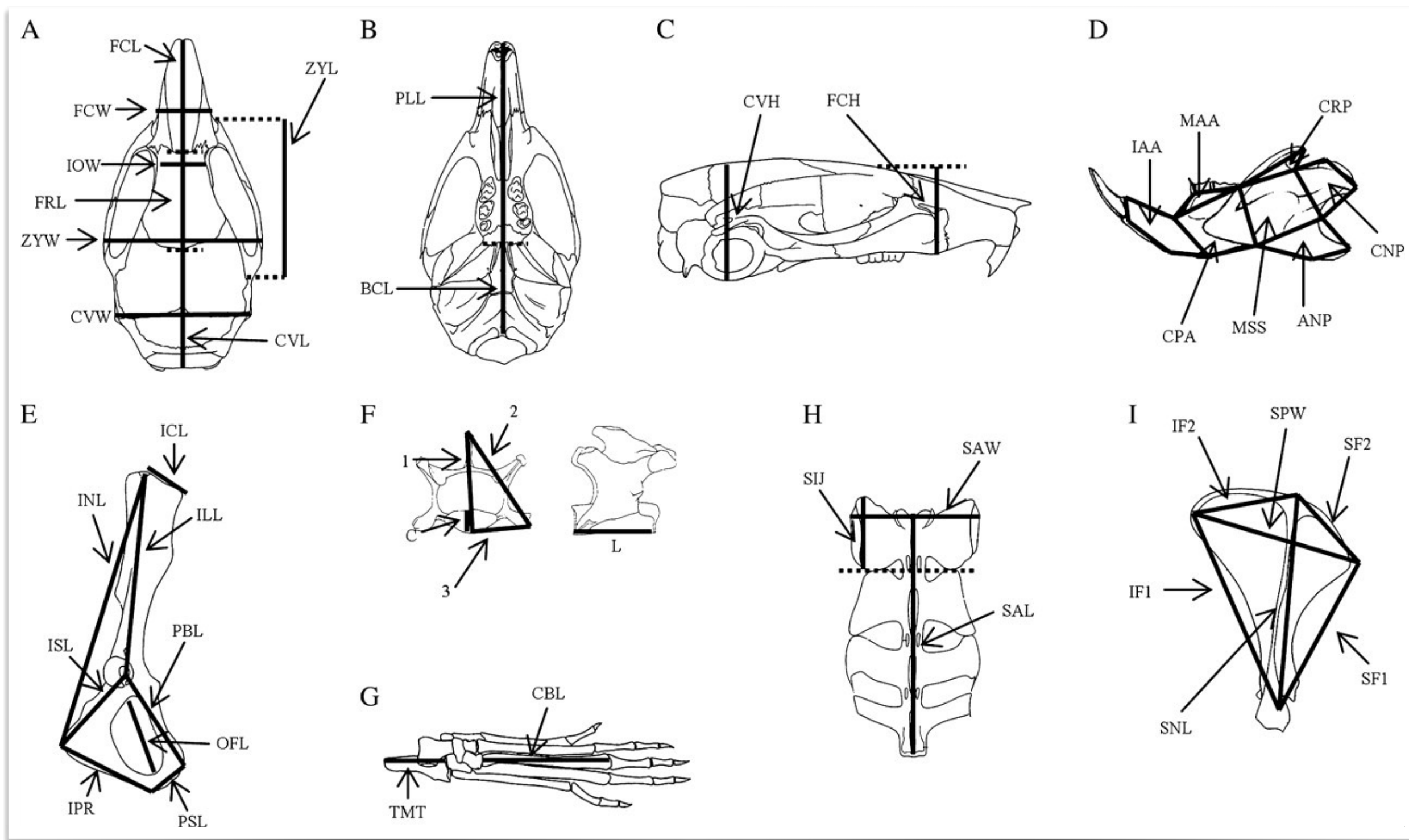
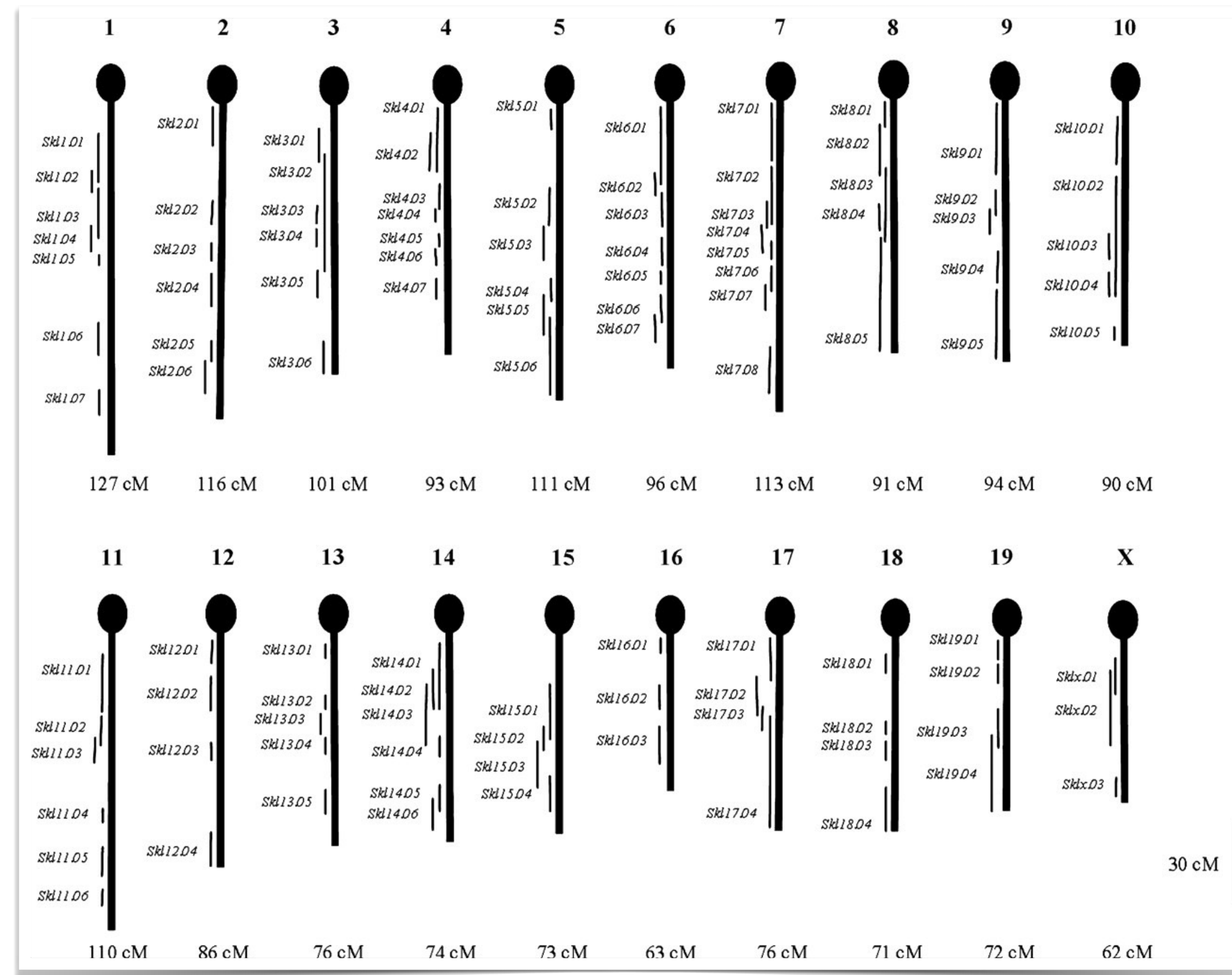
Pleiotropic Patterns of Quantitative Trait Loci for 70 Murine Skeletal Traits

Jane P Kenney-Hunt ✉, Bing Wang, Elizabeth A Norgard, Gloria Fawcett, Doug Falk,
L Susan Pletscher, Joseph P Jarvis, Charles Roseman, Jason Wolf, James M Cheverud

Genetics, Volume 178, Issue 4, 1 April 2008, Pages 2275–2288,

<https://doi.org/10.1534/genetics.107.084434>

Published: 01 April 2008 Article history ▼



Vantagens e desvantagens do desenho F2

Vantagens

- Desenho experimental simples e amplamente aplicável
- Experimento relativamente rápido, só precisa de 2 gerações de cruzamentos
- Computacionalmente rápido de analisar

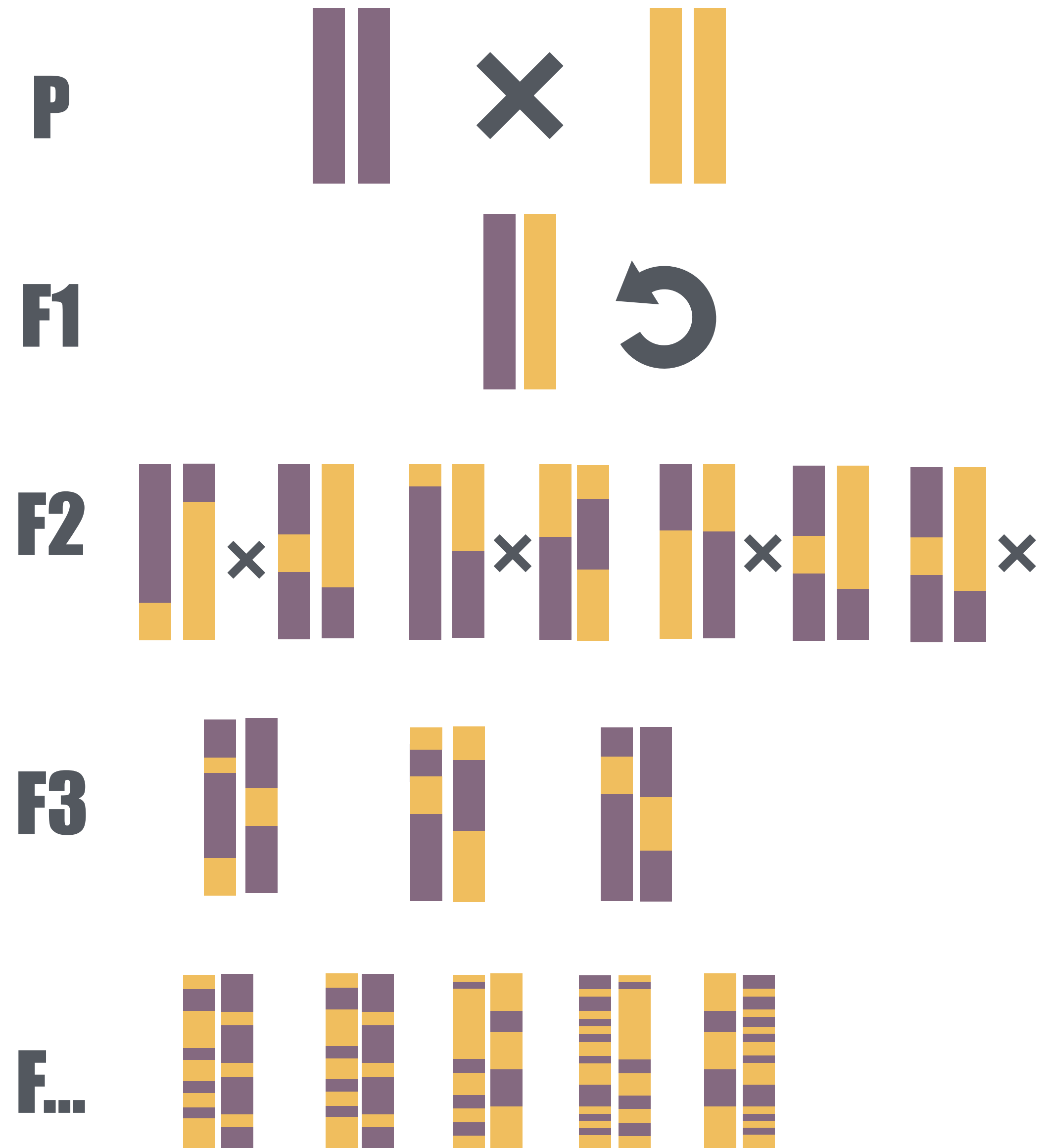
Desvantagens

- Blocos de recombinação podem ser muito grandes, porque só temos uma geração de recombinação
 - Pode ser sanado com uma amostra grande de F2
- É preciso sequenciar ou genotipar todos os indivíduos da F2
 - Potencialmente custoso

Linhas intercruzadas avançadas

Advanced intercross lines (AIL)

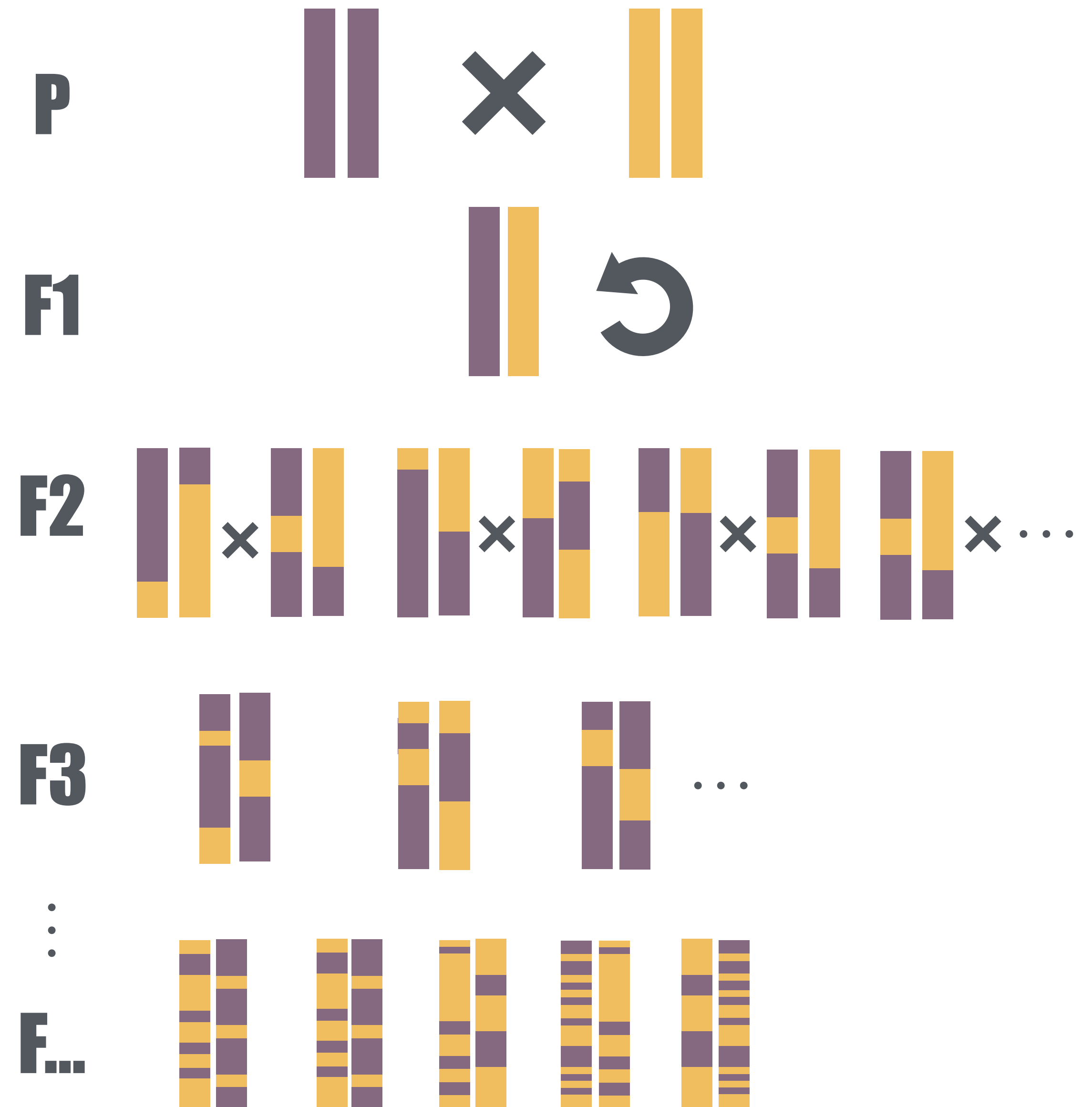
- Em vez de parar na F2, podemos continuar os cruzamentos aleatórios
- Aqui, os modelos estatísticos começam a ficar mais complicados porque temos uma estrutura populacional:
- Indivíduos da mesma família compartilham fenótipos e genótipos, e portanto os modelos tem que levar o parentesco em consideração



Linhas intercruzadas avançadas

Advanced intercross lines (AIL)

- Em vez de parar na F2, podemos continuar os cruzamentos aleatórios
- Aqui, os modelos estatísticos começam a ficar mais complicados porque temos uma estrutura populacional:
- Indivíduos da mesma família compartilham fenótipos e genótipos, e portanto os modelos tem que levar o parentesco em consideração



Modelos de mapeamento em AIL

Adição de termos associados a parentesco

- Dados:
 - y_i : fenótipo
 - A_i : componente aditivo do genótipo
 - GRM : matriz de estrutura familiar
- Parâmetros:
 - a : efeito aditivo
 - β_i : termo de parentesco
 - σ_β : variância aditiva residual
 - μ : média global
 - ϵ_i : erro residual

Modelo de mapeamento aditivo:

$$y_i = \mu + aA_i + \beta_i + \epsilon_i$$

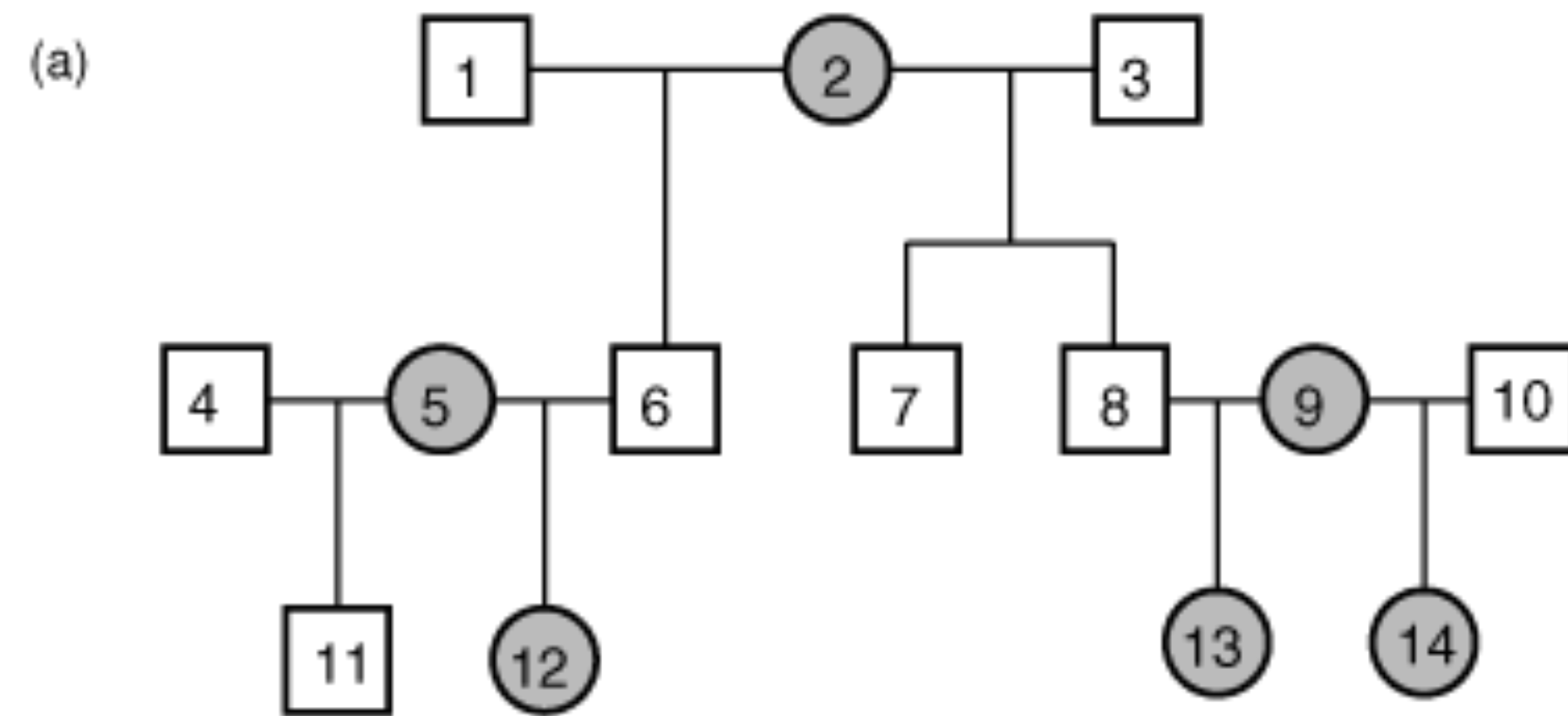
$$\beta_i \sim Normal(0, \sigma_\beta GRM)$$

Termo que permite que indivíduos mais aparentados sejam mais similares

Estrutura populacional é codificada no termo de variância

A matriz de relacionamento genético

GRM - genetic relatedness matrix



Modelo de mapeamento aditivo:

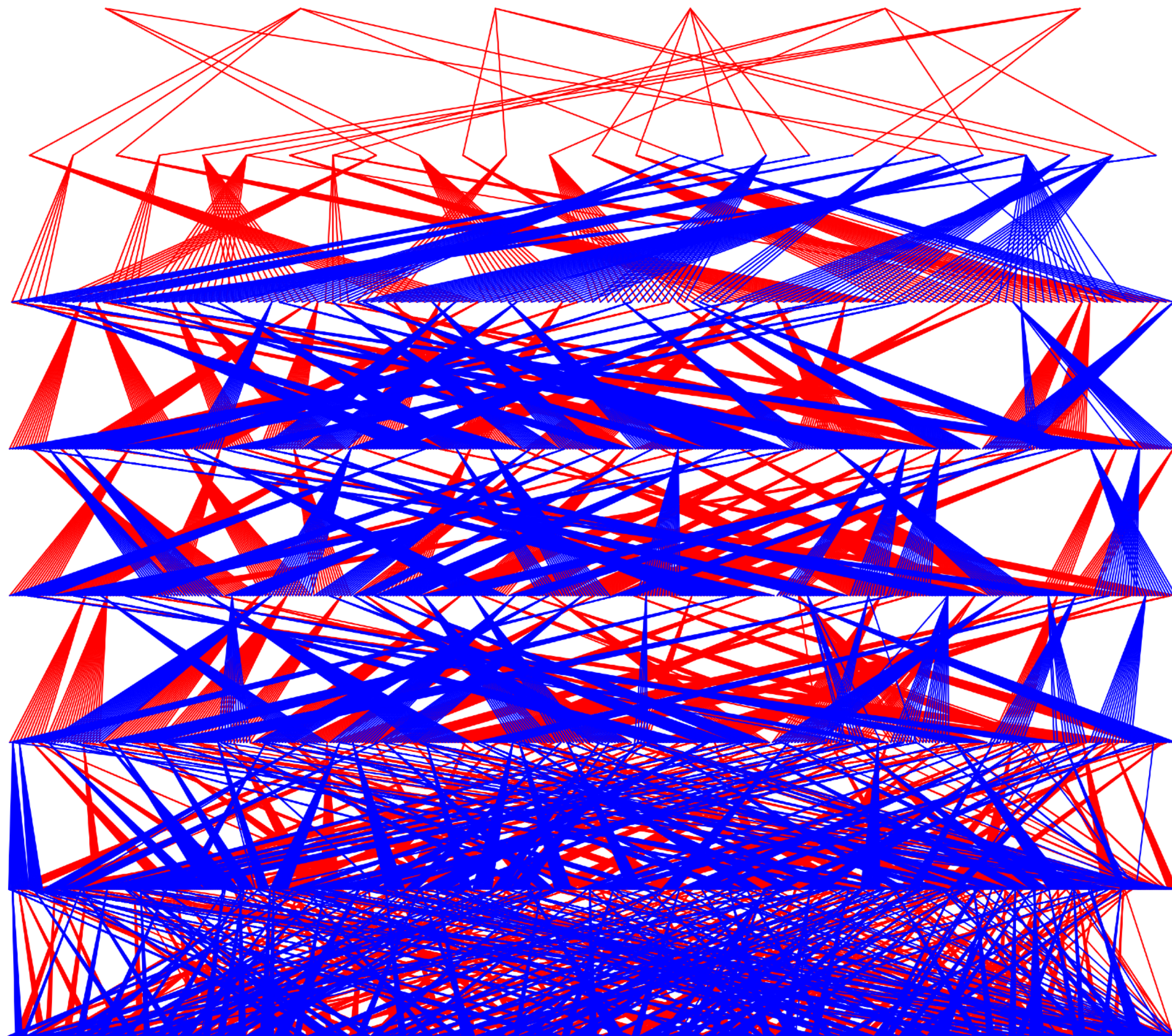
$$y_i = \mu + aA_i + \beta_i + \epsilon_i$$

$$\beta_i \sim Normal(0, \sigma_\beta GRM)$$

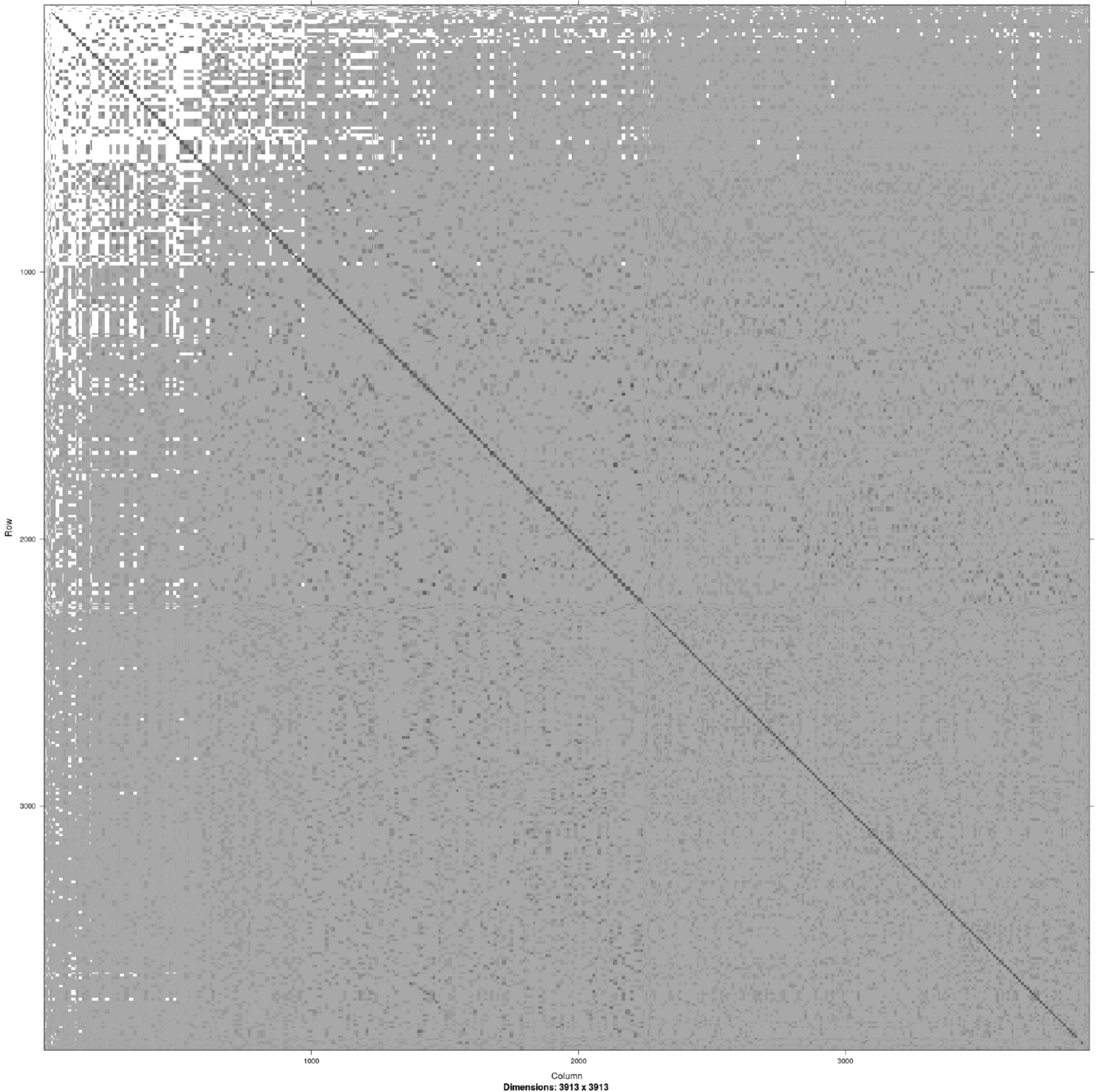
(b)

$$GRM = \begin{pmatrix} 1 & 0 & 0 & 0 & 0 & 0.5 & 0 & 0 & 0 & 0 & 0 & 0.25 & 0 & 0 \\ 0 & 1 & 0 & 0 & 0 & 0.5 & 0.5 & 0.5 & 0 & 0 & 0 & 0.25 & 0.25 & 0 \\ 0 & 0 & 1 & 0 & 0 & 0 & 0.5 & 0.5 & 0 & 0 & 0 & 0 & 0.25 & 0 \\ 0 & 0 & 0 & 1 & 0 & 0 & 0 & 0 & 0 & 0 & 0.5 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 1 & 0 & 0 & 0 & 0 & 0 & 0.5 & 0.5 & 0 & 0 \\ 0.5 & 0.5 & 0 & 0 & 0 & 1 & 0.25 & 0.25 & 0 & 0 & 0 & 0.5 & 0.125 & 0 \\ 0 & 0.5 & 0.5 & 0 & 0 & 0.25 & 1 & 0.5 & 0 & 0 & 0 & 0.125 & 0.25 & 0 \\ 0 & 0.5 & 0.5 & 0 & 0 & 0.25 & 0.5 & 1 & 0 & 0 & 0 & 0.125 & 0.5 & 0 \\ 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 1 & 0 & 0 & 0 & 0.5 & 0.5 \\ 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 1 & 0 & 0 & 0 & 0.5 \\ 0 & 0 & 0 & 0.5 & 0.5 & 0 & 0 & 0 & 0 & 0 & 1 & 0.25 & 0 & 0 \\ 0.25 & 0.25 & 0 & 0 & 0.5 & 0.5 & 0.125 & 0.125 & 0 & 0 & 0.25 & 1 & 0.0625 & 0 \\ 0 & 0.25 & 0.25 & 0 & 0 & 0.125 & 0.25 & 0.5 & 0.5 & 0 & 0 & 0.0625 & 1 & 0.25 \\ 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0.5 & 0.5 & 0 & 0 & 0.25 & 1 \end{pmatrix}$$

Pedigree complexo = GRM complexa

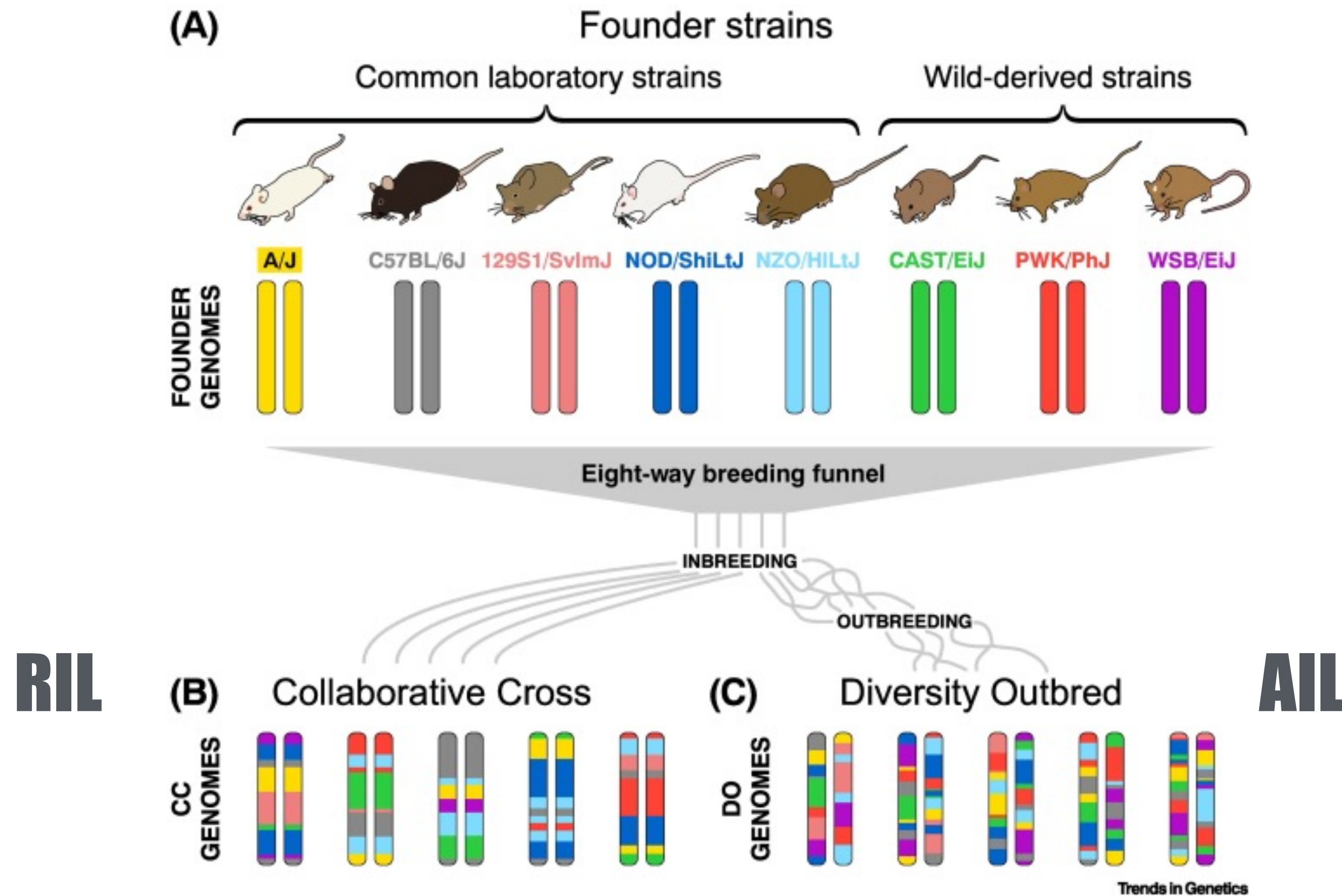


$GRM =$



Populações experimentais com "cara" de naturais

Diversity outbred e outras



Vantagens e desvantagens de AIL

Vantagens

- Permite uma resolução alta na localização dos locus de interesse
- Aplicável para o estudo de interações genéticas (dominância, epistasia, ...)
- Aplicável para fenótipos complexos
- Populações com variação genética com propriedades parecidas com populações naturais

Desvantagens

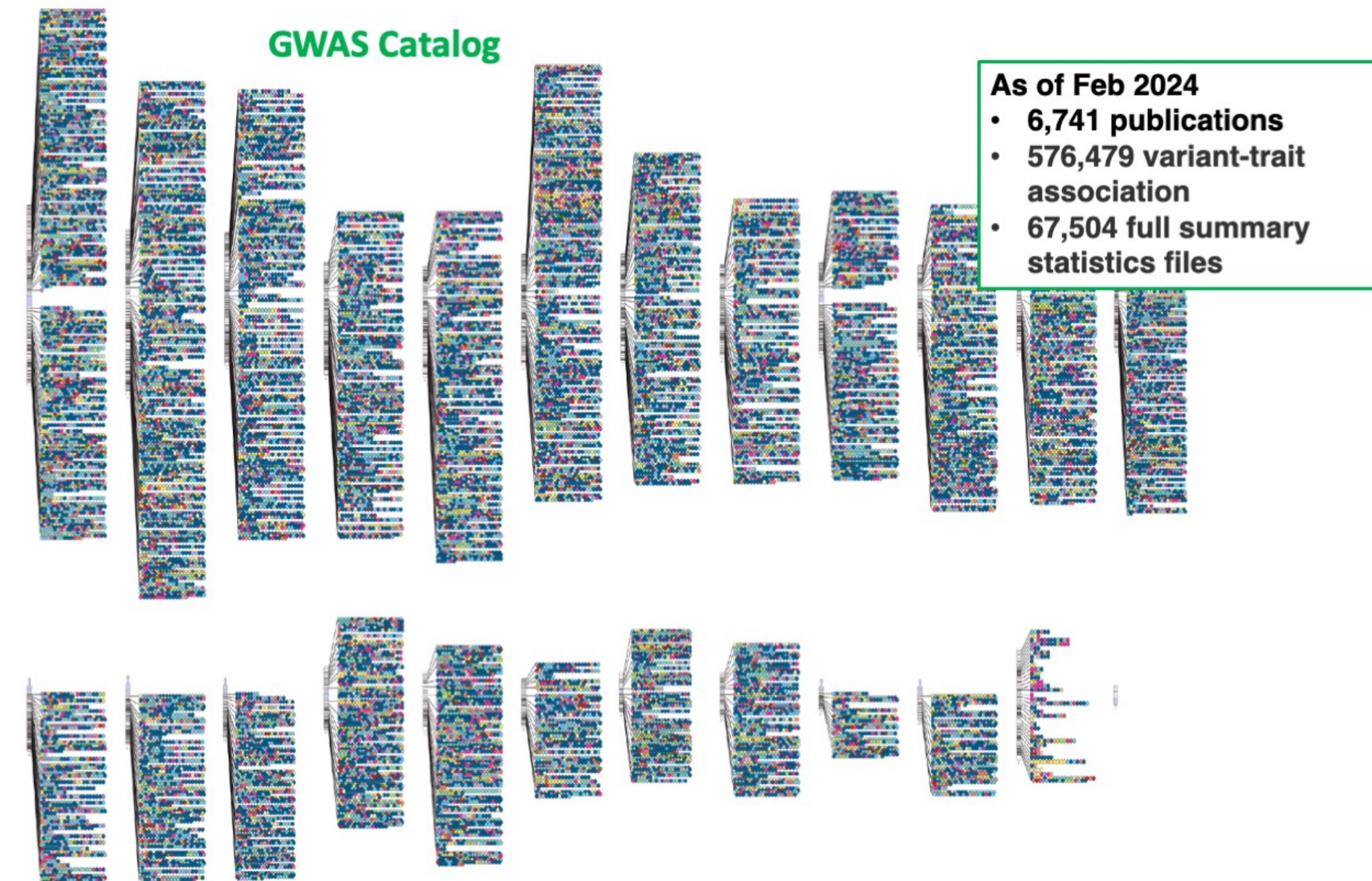
- Exige muitas gerações de cruzamentos
- Todos os indivíduos devem ser genotipados
- Modelos mais complexos de mapeamento

Mapeamento em populações
naturais

Estudo de Associação em Todo o Genoma

Genome Wide Association Studies (GWAS)

- Modelos usados em GWAS são muito parecidos com os usados em mapeamentos em laboratório, mas são aplicados em populações naturais
- São usados para estudar a arquitetura genética de doenças e caracteres complexos em populações humanas e não-humanos



Biobanks e GWAS de larga escala

- Bancos de dados públicos e privados permitem o mapeamento de caracteres complexos
- GIANT consortium - <https://research.23andme.com/dataset-access/>
- UK biobank - <https://www.ukbiobank.ac.uk/>
- Japan biobank - <https://biobankjp.org/en/>

Article | [Open access](#) | Published: 12 October 2022

A saturated map of common genetic variants associated with human height

[Loïc Yengo](#) , [Sailaja Vedantam](#), [Eirini Marouli](#), [Julia Sidorenko](#), [Eric Bartell](#), [Saori Sakaue](#), [Marielisa Graff](#), [Anders U. Eliassen](#), [Yunxuan Jiang](#), [Sridharan Raghavan](#), [Jenkai Miao](#), [Joshua D. Arias](#), [Sarah E. Graham](#), [Ronen E. Mukamel](#), [Cassandra N. Spracklen](#), [Xianyong Yin](#), [Shyh-Huei Chen](#), [Teresa Ferreira](#), [Heather H. Highland](#), [Yingjie Ji](#), [Tugce Karaderi](#), [Kuang Lin](#), [Kreete Lüll](#), [Deborah E. Malden](#), [23andMe Research Team](#), [VA Million Veteran Program](#), [DiscovEHR \(DiscovEHR and MyCode Community Health Initiative\)](#), [eMERGE \(Electronic Medical Records and Genomics Network\)](#), [Lifelines Cohort Study](#), [The PRACTICAL Consortium](#), [Understanding Society Scientific Group](#), ... [Joel N. Hirschhorn](#) 

[+ Show authors](#)

[Nature](#) **610**, 704–712 (2022) | [Cite this article](#)

137k Accesses | **201** Citations | **1418** Altmetric | [Metrics](#)

12 mil variantes associadas a altura em humanos

Biobanks e GWAS de larga escala

- Bancos de dados de mapeamento
- GIANT colaboração de pesquisa
- UK biobank
- Japan biobank

JOURNAL ARTICLE

The genetic architecture of dog ownership: large-scale genome-wide association study in 97,552 European-ancestry individuals

Tong Gong , Robert Karlsson, Shuyang Yao, Patrik K E Magnusson, Olesya Ajnakina, Andrew Steptoe, Laxmi Bhatta, Ben Brumpton, Ashish Kumar, Erik Mélen ... [Show more](#)

[Author Notes](#)

G3 Genes|Genomes|Genetics, Volume 14, Issue 8, August 2024, jkae116,

<https://doi.org/10.1093/g3journal/jkae116>

Published: 31 May 2024 **Article history** ▼

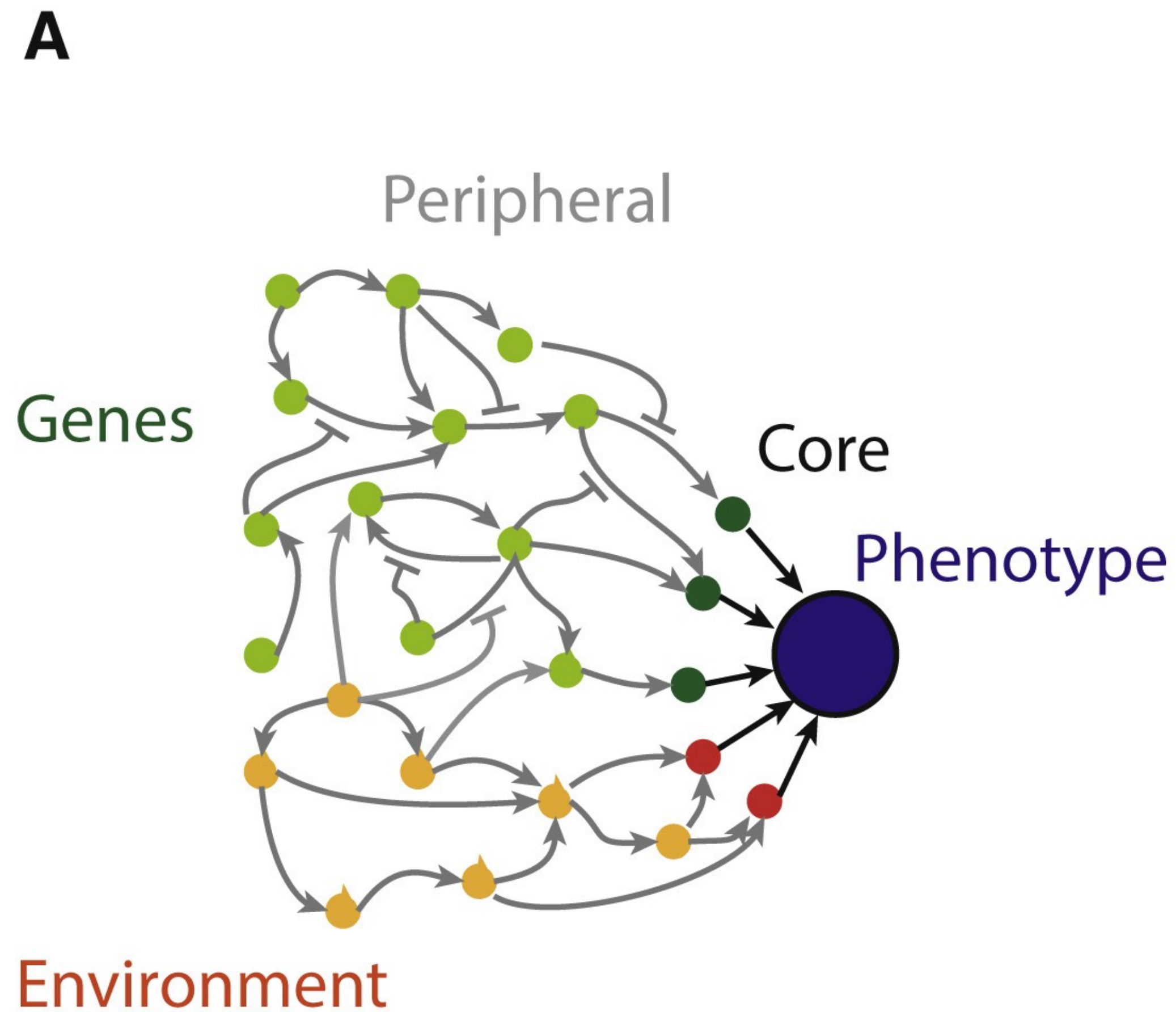
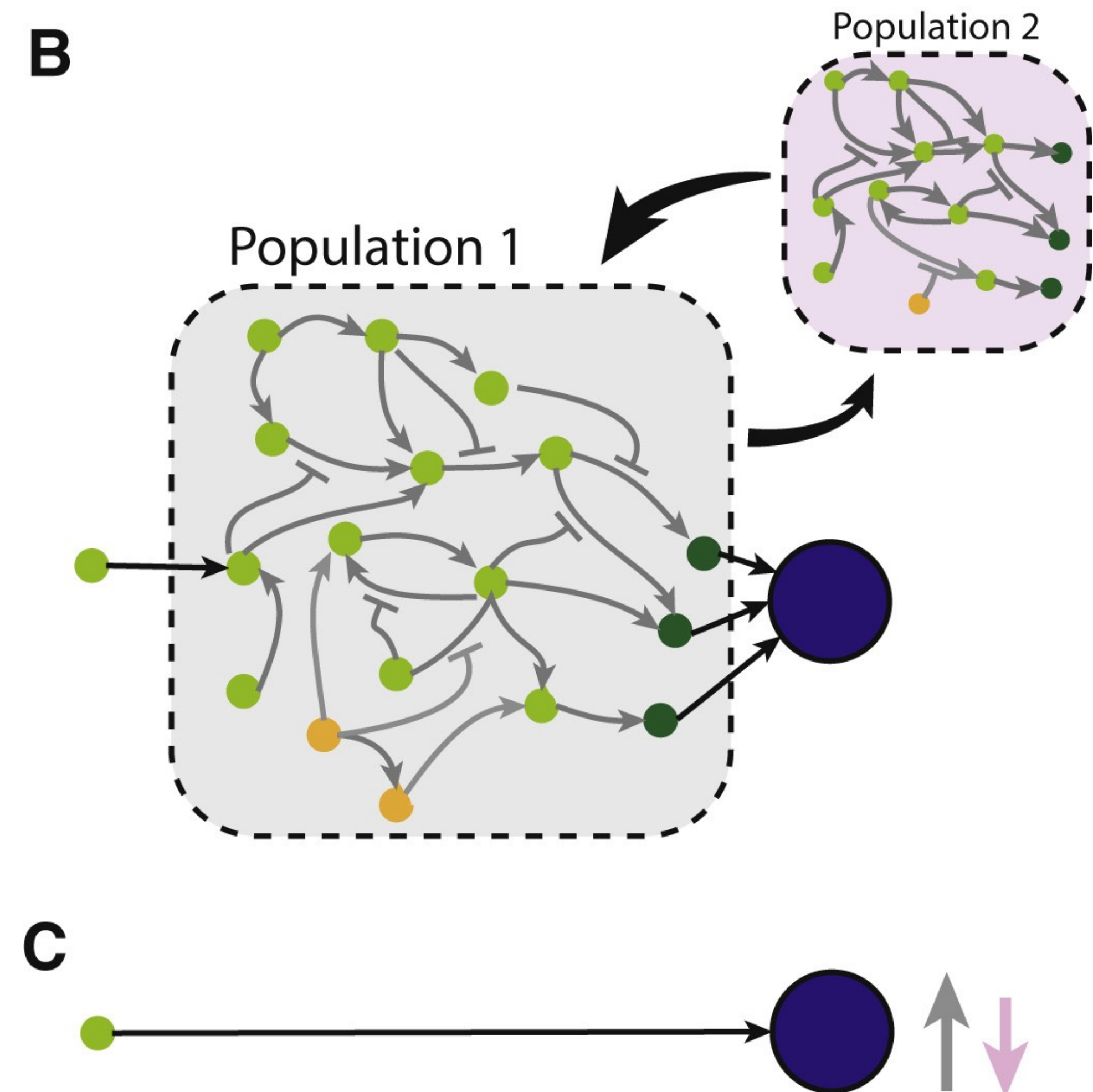
c variants

[Bartell](#), [Saori Sakaue](#), [Marielisa](#), [Miao](#), [Joshua D. Arias](#), [Sarah E.](#), [Shyh-Huei Chen](#), [Teresa Ferreira](#), [Lüll](#), [Deborah E. Malden](#), [23andMe](#), [UK Biobank](#), [MyCode Community Health](#) (work), [Lifelines Cohort Study](#), [The](#), [Joel N. Hirschhorn](#) 

adas a

altura em humanos

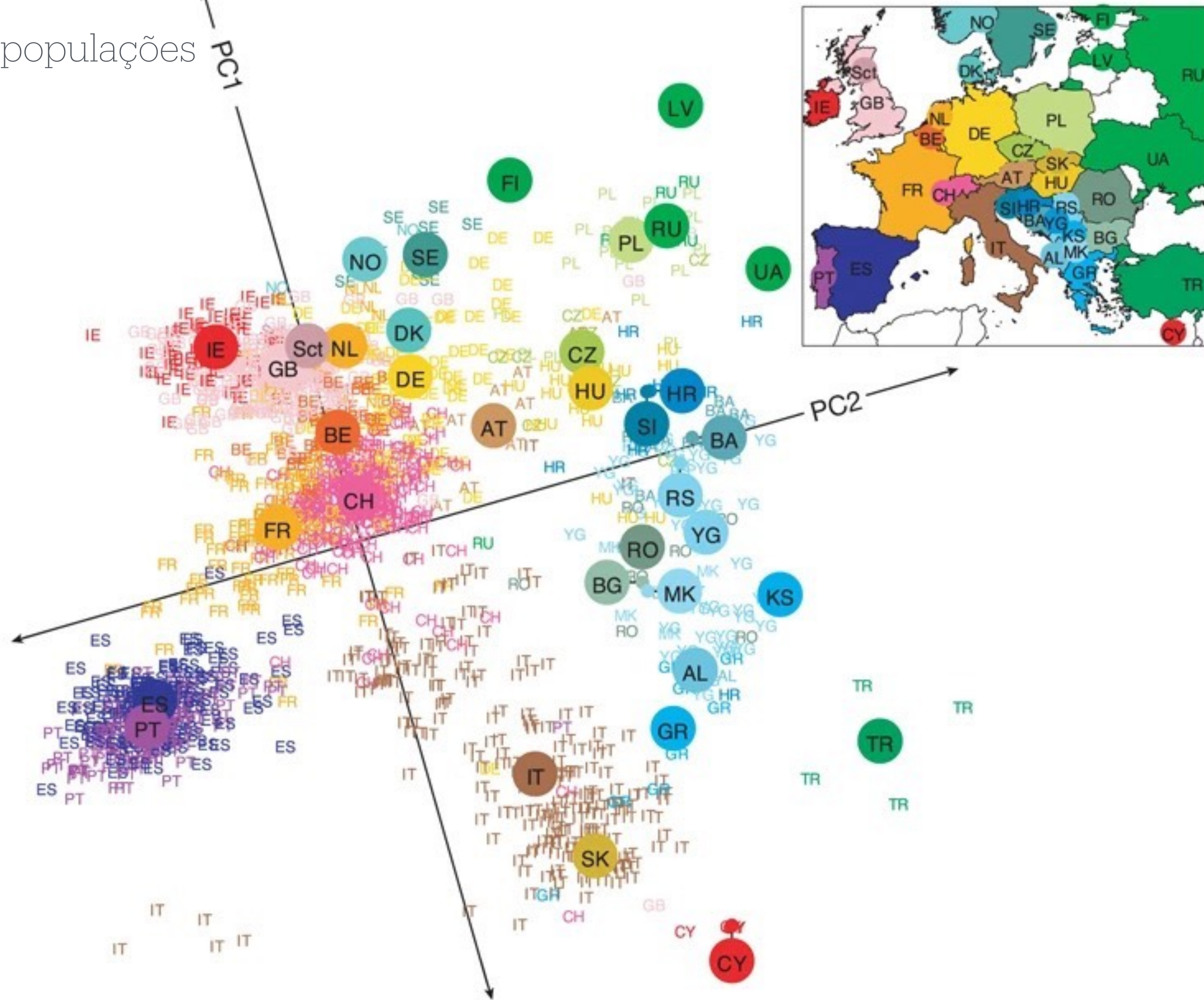
Processo de desenvolvimento do fenótipo pode variar entre populações



Estrutura populacional

História evolutiva e demográfica alteram a composição de populações

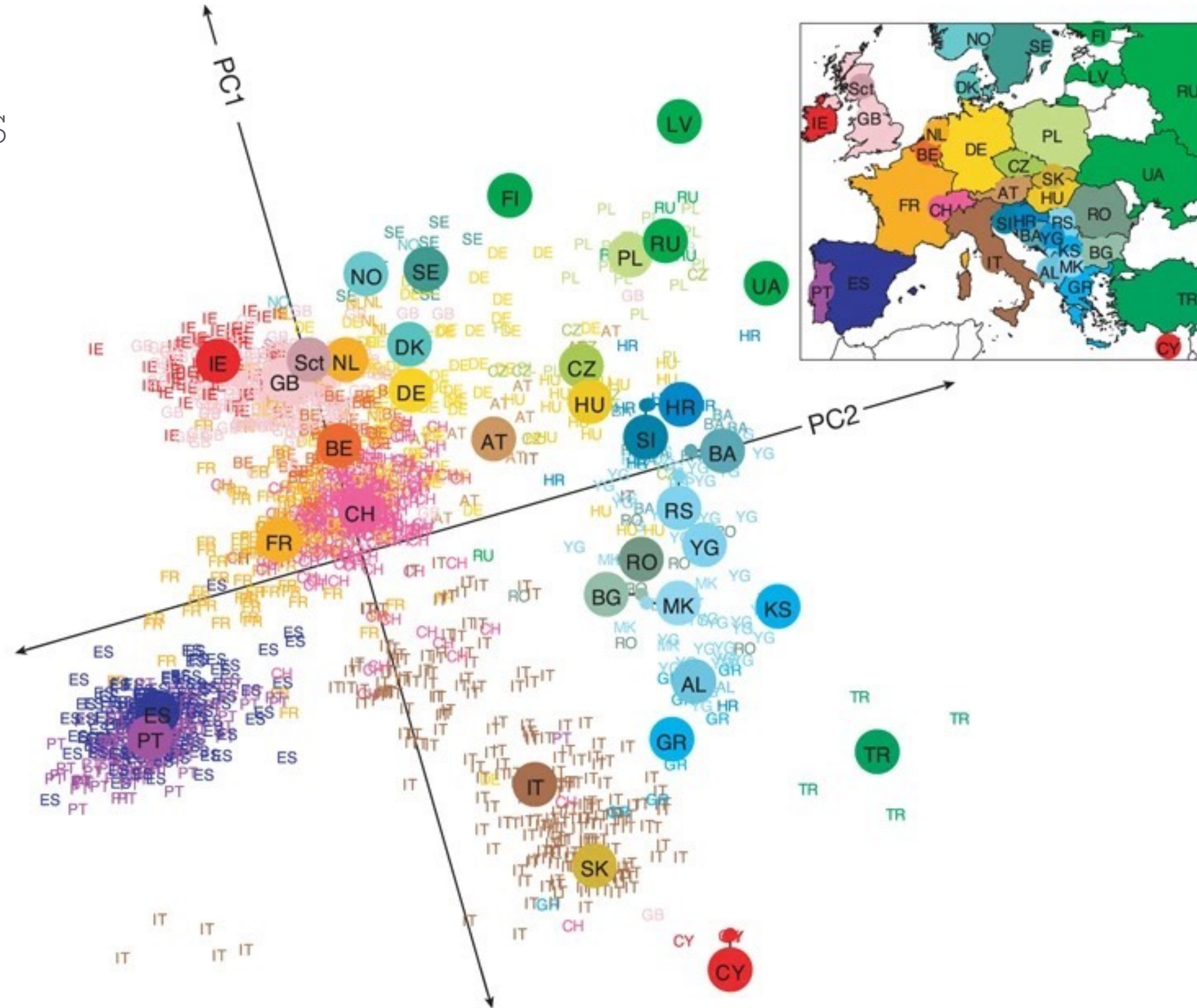
- Tal qual as AIL, indivíduos em populações naturais possuem uma estrutura de parentesco complexa
- Essa estrutura de manifesta dentro e entre populações
- Modelos de mapeamento tem que incluir essa estrutura!
- Usar a matriz de relacionamento genético (GRM) funciona, mas é computacionalmente impossível em bancos de dados grandes



PCA

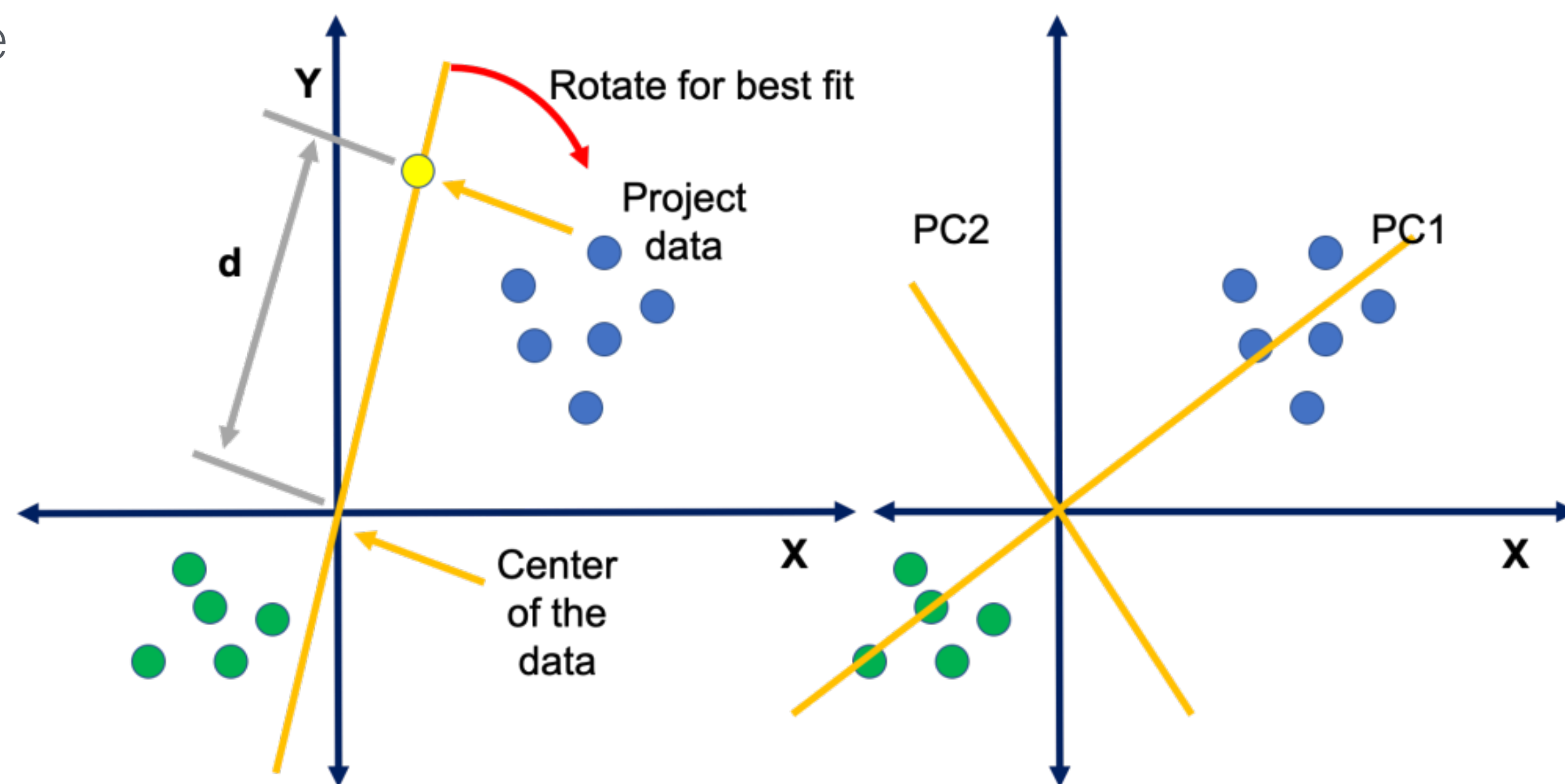
Analise de componentes principais

- Uma ferramenta poderosa para a caracterização da estrutura populacional
- O PCA identifica eixos de diferenciação genética que são frequentemente associados a genealogia
- A análise é feita na matriz de marcadores genéticos (SNPs)



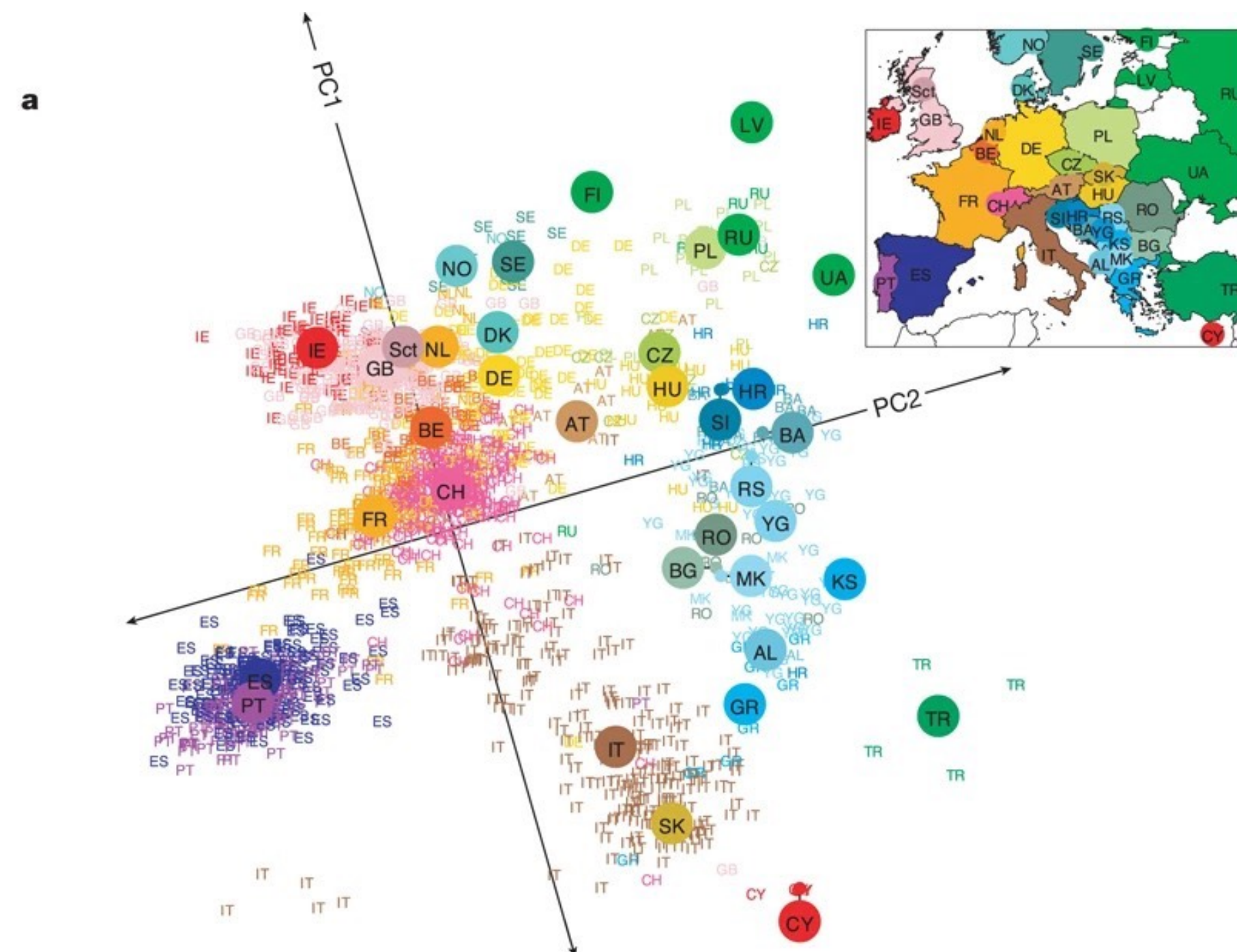
O que é um PCA?

- PCA é uma técnica de representação de dados multivariados muito usada em diferentes áreas
- Muitos usos:
 - Redução dimensional
 - Identificação de grupos
 - Representação da variação em eixos independentes



PCA em genômica

- O PCA, quando aplicado em dados de variantes genéticas, recupera algo da história demográfica das populações



PLOS GENETICS

OPEN ACCESS PEER-REVIEWED

RESEARCH ARTICLE

A Genealogical Interpretation of Principal Components Analysis

Gil McVean

Published: October 16, 2009 • <https://doi.org/10.1371/journal.pgen.1000686>

Article

Authors

Metrics

Comments

Media Coverage

Abstract

Author Summary

Introduction

Results

Discussion

Methods

Acknowledgments

Author Contributions

References

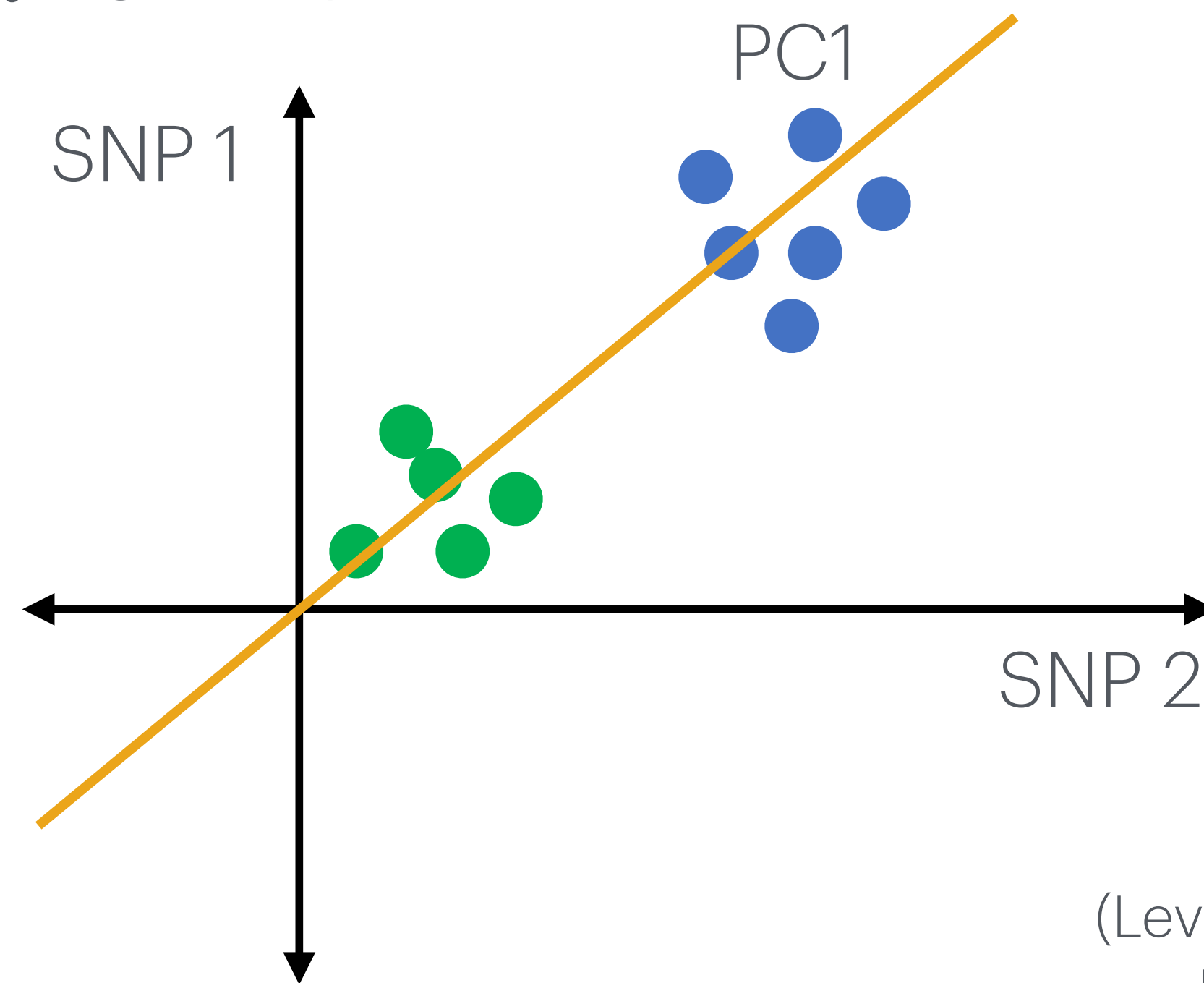
Abstract

Principal components analysis, PCA, is a statistical method commonly used in population genetics to identify structure in the distribution of genetic variation across geographical location and ethnic background. However, while the method is often used to inform about historical demographic processes, little is known about the relationship between fundamental demographic parameters and the projection of samples onto the primary axes. Here I show that for SNP data the projection of samples onto the principal components can be obtained directly from considering the average coalescent times between pairs of haploid genomes. The result provides a framework for interpreting PCA projections in terms of underlying processes, including migration, geographical isolation, and admixture. I also demonstrate a link between PCA and Wright's F_{ST} and show that SNP ascertainment has a largely simple and predictable effect on the projection of samples. Using examples from human genetics, I discuss the application of these results to empirical data and the implications for inference.

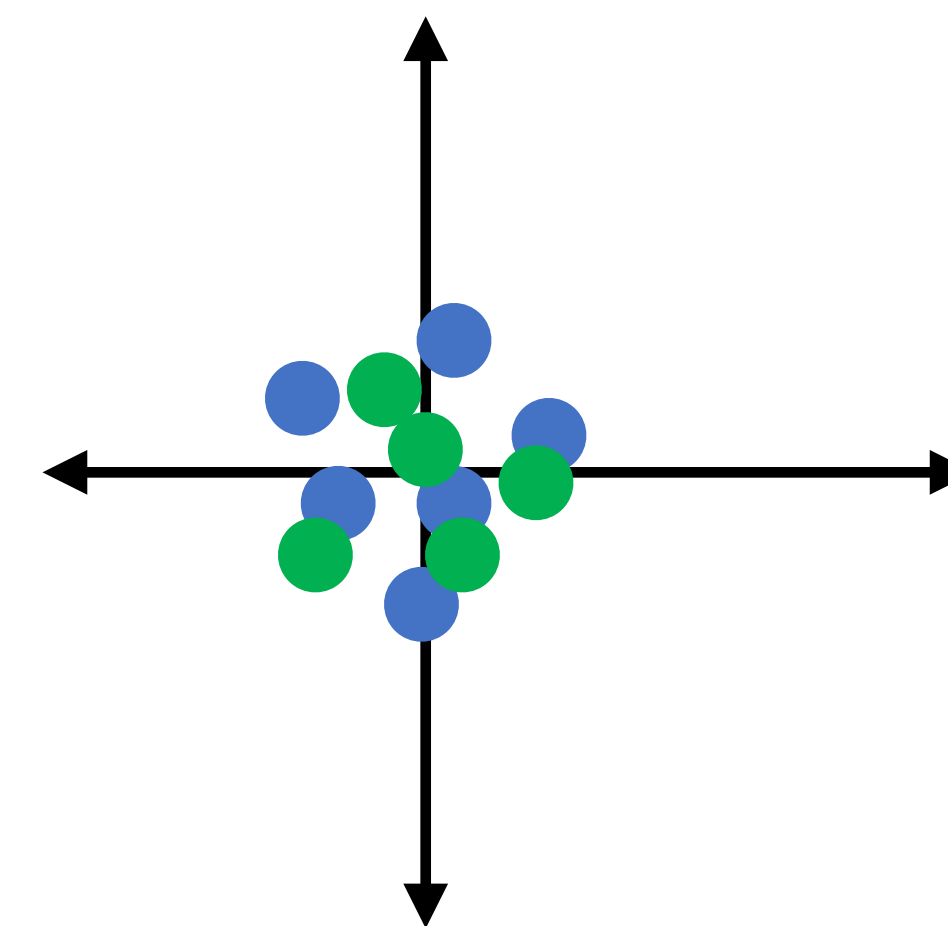
Modelos de mapeamento e PCA

Aproximação do termo de genealogia

Espaço genotípico



Remoção do efeito
do PC1



(Levemente esquemático, na
verdade não haveria variação
nenhuma na direção do PC1)

Modelos de mapeamento e PCA

Aproximação do termo de genealogia

- Na prática, as projeções são incluídas no modelo com co-variáveis

Modelo de mapeamento aditivo com PCs:

$$y_i = \mu + aA_i + \beta_1 PC1_i + \beta_2 PC2_i + \dots + \epsilon_i$$

