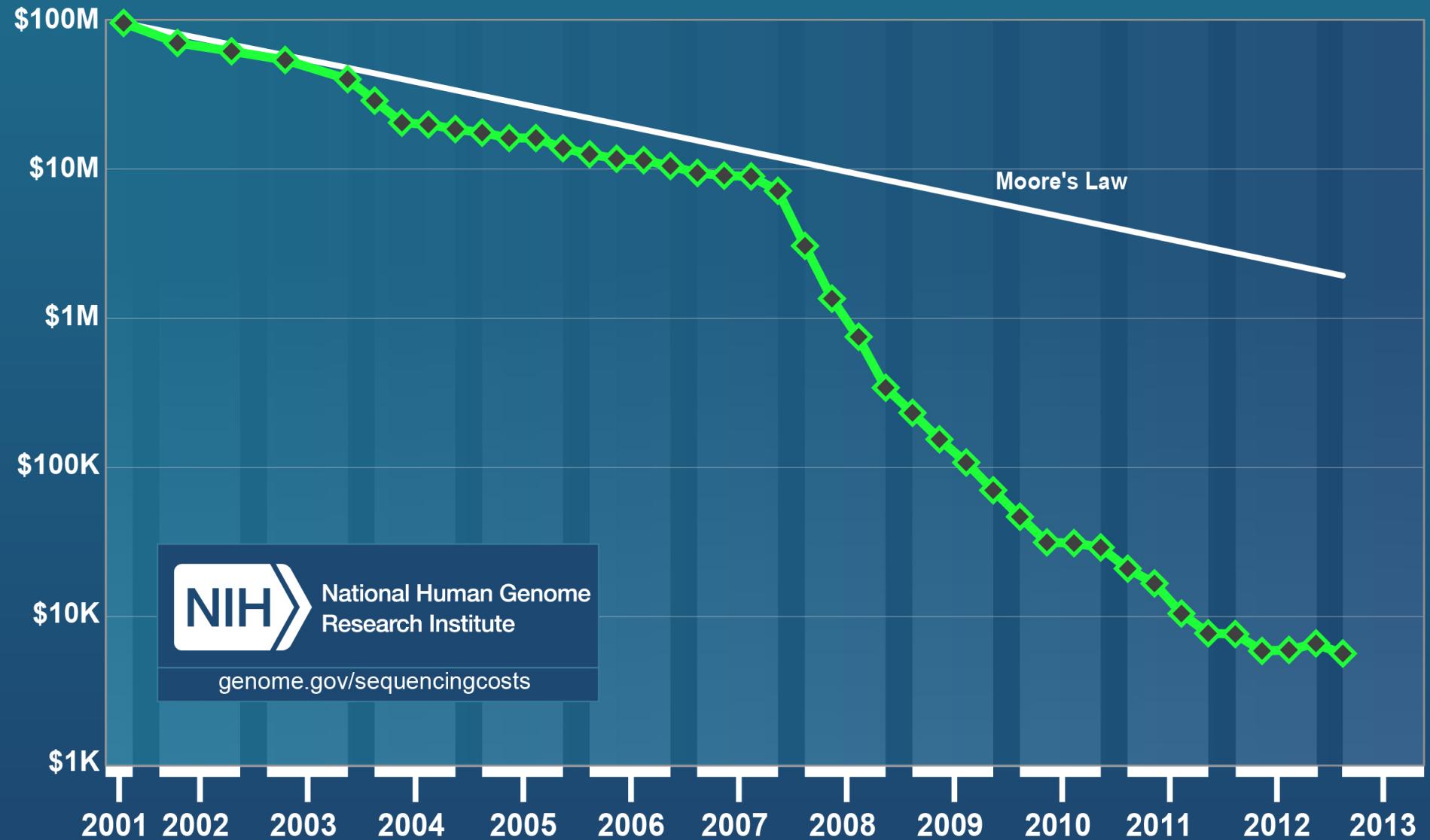


Transformando a medicina com abordagem sistêmática de dados genéticos

Mauricio Carneiro

Broad Institute of MIT and Harvard

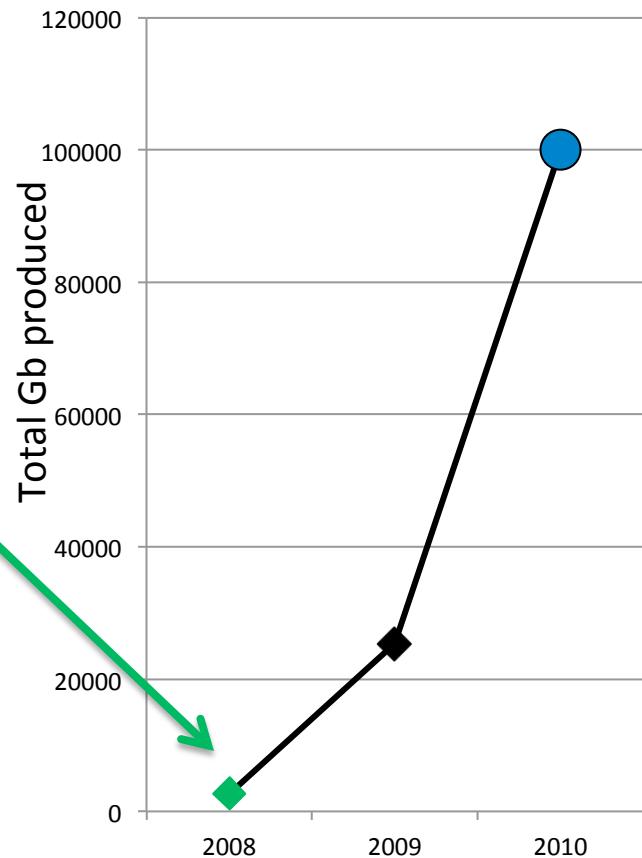
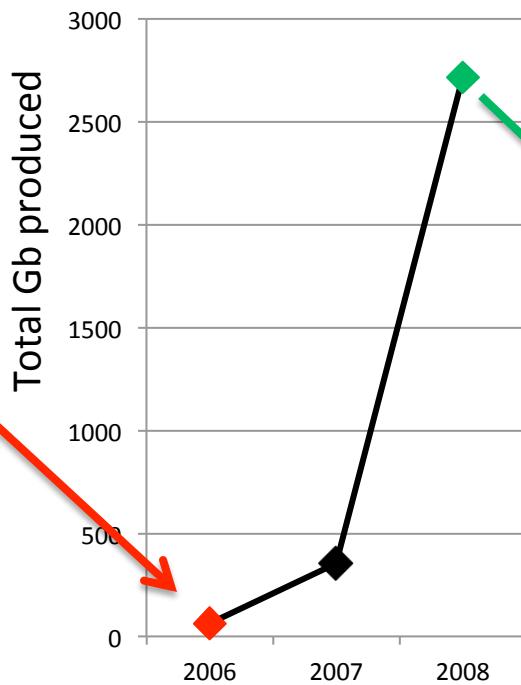
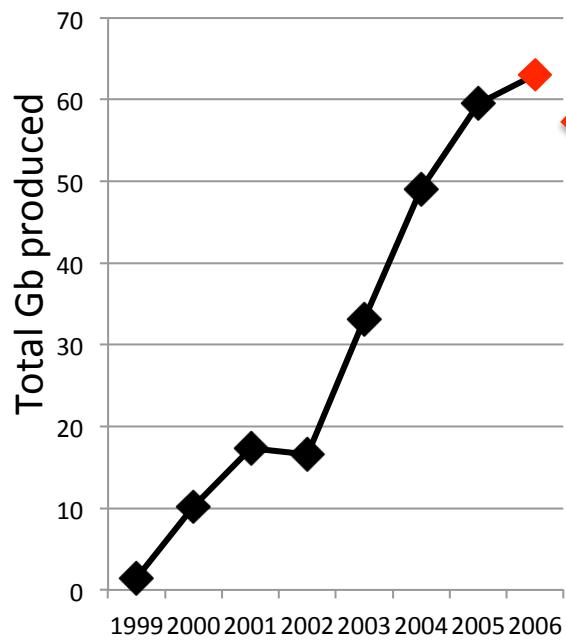
Cost per Genome



Sequenciadores no mercado

- Illumina HiSeq 2500
 - Padrão do mercado, maior volume de dados e qualidade por base. Referencia para sequenciamento humano.
 - (MiSeq) Versão de bancada do HiSeq com menor volume para projetos pequenos (1 exoma por rodada)
- Life Technologies Ion Proton
 - Sequenciamento por semicondutores é o principal competidor no mercado. Tecnologia é acurada para identificação de mutações pontuais mas sofre de erros sistemáticos em inserções e deleções.
 - (Ion) Versão de bancada do Ion Proton.
- Pacific Biosciences RS II
 - Único sequenciador com *reads* de até 20,000 bases e capaz de identificar metilação de bases. Excelente para montagens de pequenos genomas (microorganismos).
- **Futuros sequenciadores com potencial de mudar o mercado**
 - Oxford Nanopores, GNUMBio e QIAGEN.
- **Histórico (em desuso)**
 - Solid, 454 e Sanger.

Sequenciamento no Broad Institute



...2011 = 500,000 Gb

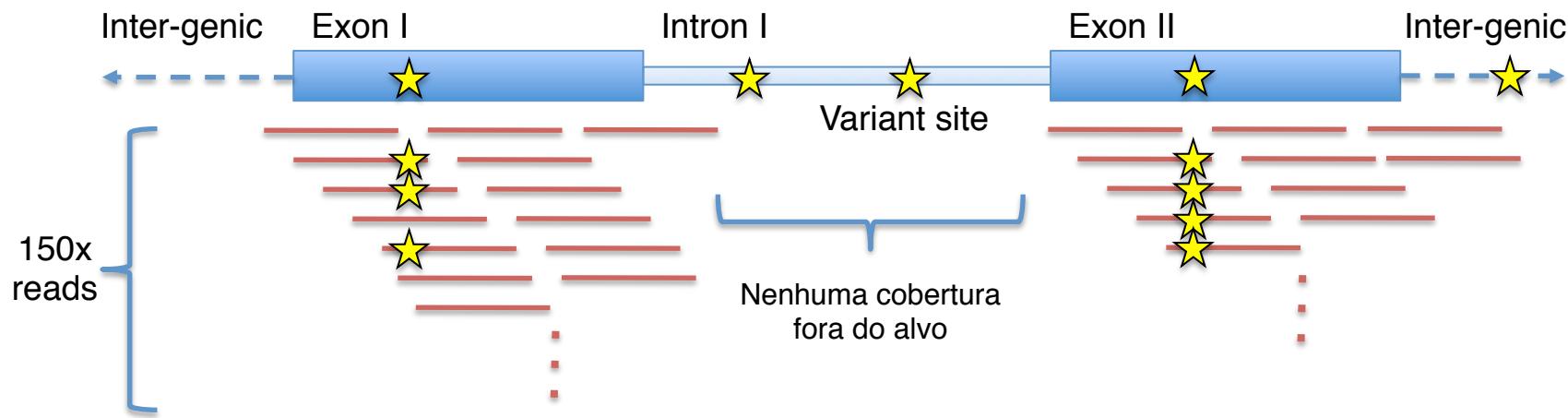
...2013 = 1.5 Pb



Sequenciamento humano para projetos de pesquisa clínica vem em 4 modelos

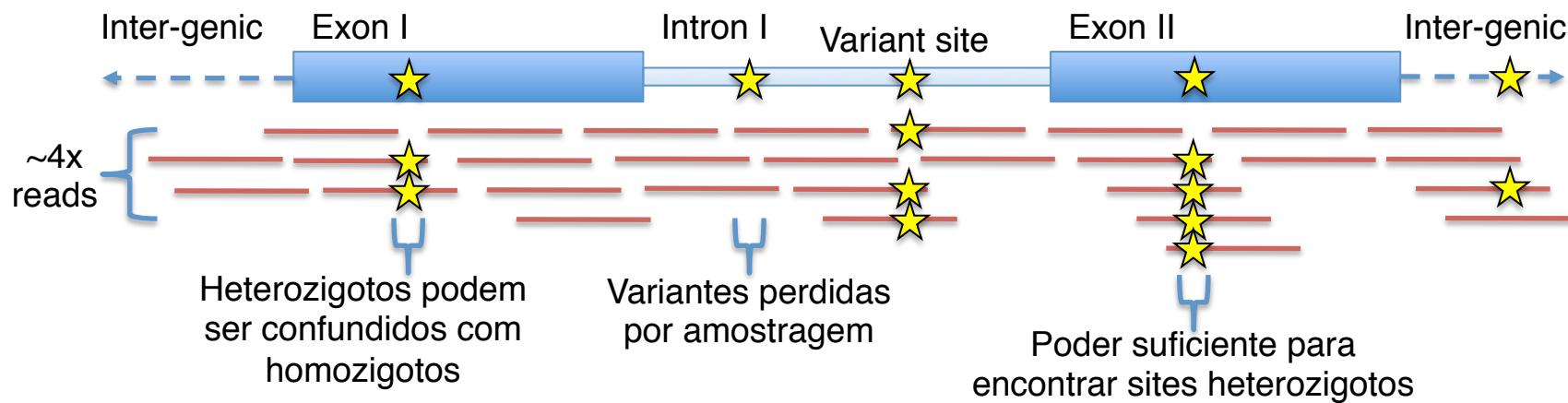
- Sequenciamento de alvos específicos (**targeted**)
 - Geralmente utilizado para identificar alvos de interesse específico (ex: APOE para alzheimer)
- Exoma completo (**WES**)
 - Típico projeto para avaliação correlação genética para doenças e pesquisa. (ex: diabetes, autismo, distrofia muscular...)
- Genoma completo em baixa cobertura (**low pass WGS**)
 - Usado para identificar variação de baixa frequência em população (ex: 1000 genomes project)
- Genoma completo em alta cobertura (**high pass WGS**)
 - Quando é preciso entender *tudo** sobre um paciente (ex: avaliação clínica).

Desenho para captura de exoma



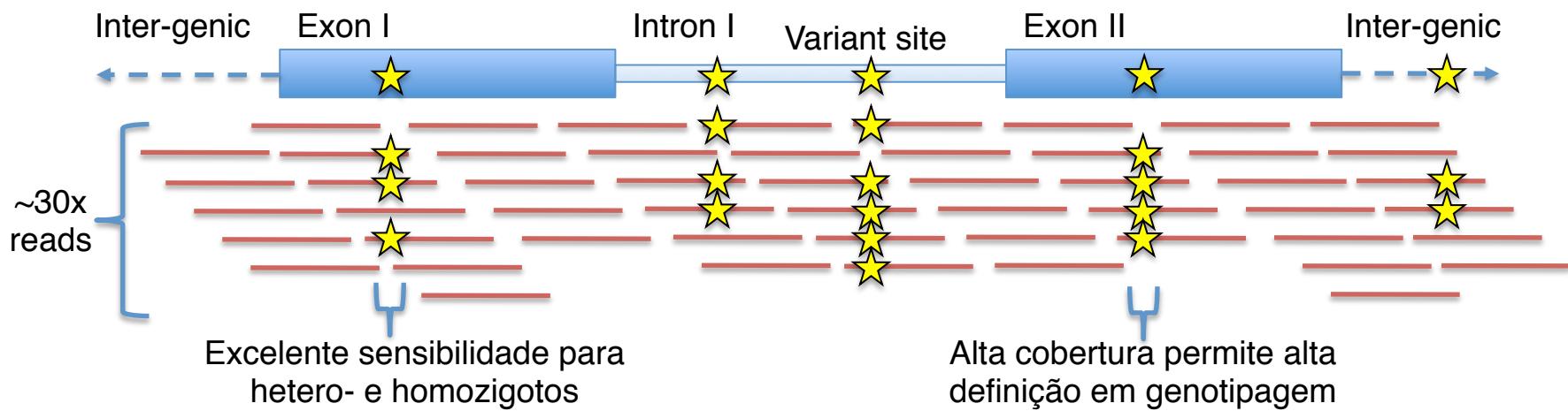
Bases sequenciadas	~32Mb	Variantes por amostra	~20K
Cobertura	80% @ 20x	% variação no genoma	0.5%
Produção	5 Gb	Pr{singleton discovery}	~95%
# lanes de HiSeq 2500	~0.33	Pr{common allele discovery}	~95%

Genoma completo em baixa cobertura



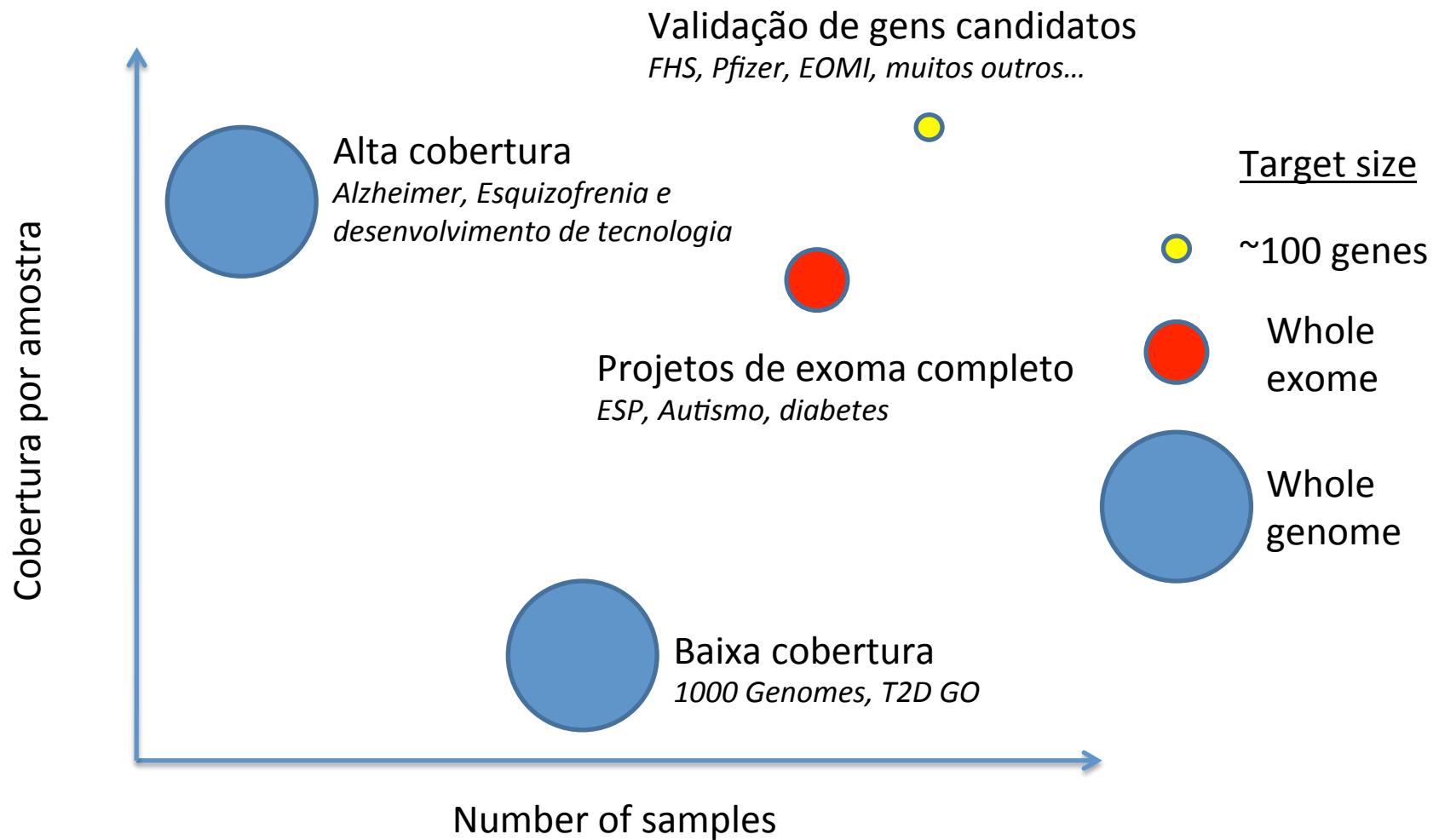
Bases sequenciadas	~3 Gb	Variantes por amostra	~3M
Cobertura	Avg. 4x	% variação no genoma	~90%
Produção	20 Gb	Pr{singleton discovery}	<50%
# lanes de HiSeq 2500	~1.25	Pr{common allele discovery}	~99%

Genoma completo em alta cobertura



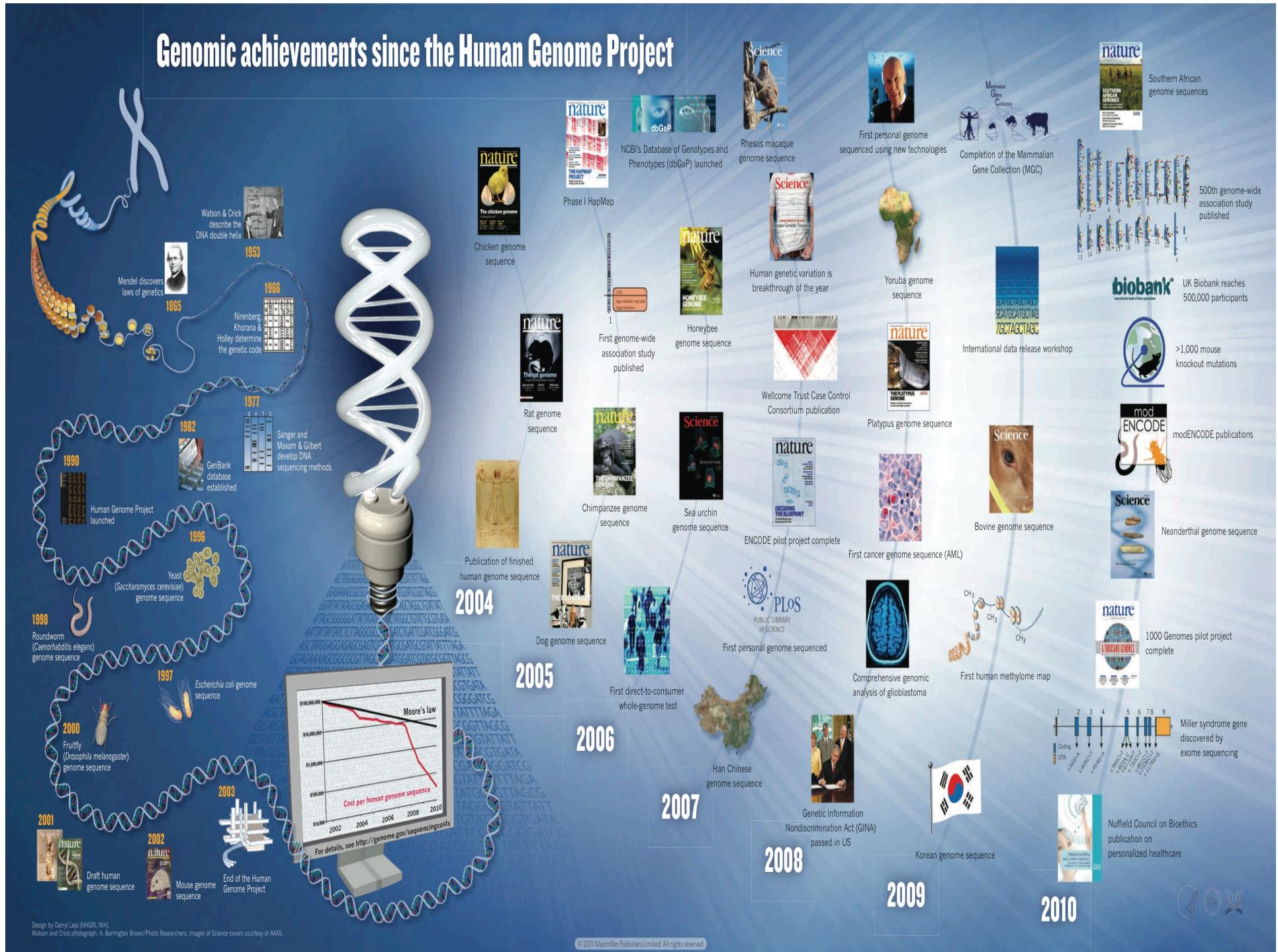
Bases sequenciadas	~3 Gb	Variantes por amostra	~3-5M
Cobertura	~30x	% variação no genoma	>99%
Produção	100 Gb	Pr{singleton discovery}	>99%
# lanes de HiSeq 2500	~8 lanes	Pr{common allele discovery}	>99%

Usando uma variedade de desenhos experimentais,
de acordo com o objetivo do projeto



ENTENDENDO DADOS DE SEQUENCIAMENTO

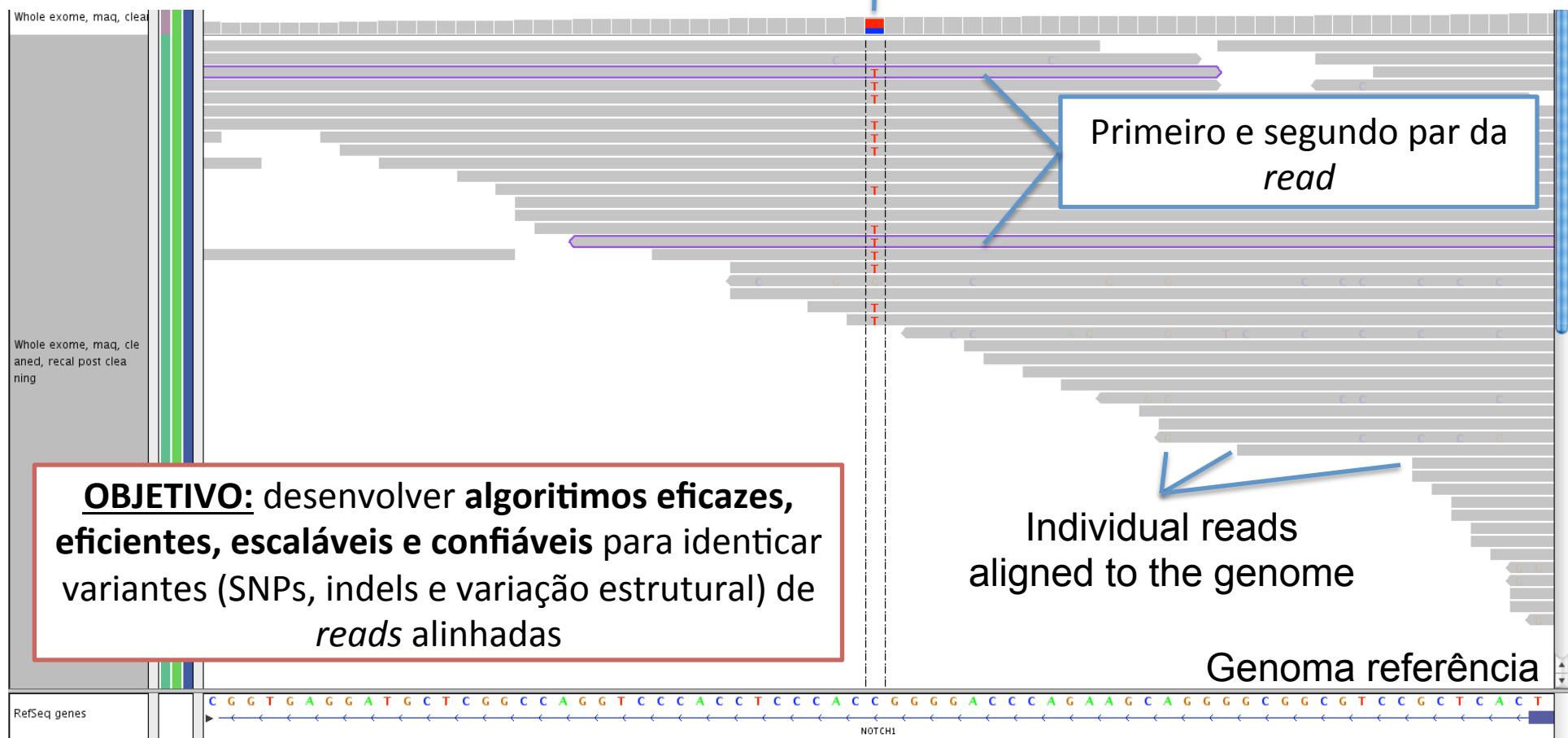
Genomic achievements since the Human Genome Project



NGS cria uma oportunidade sem precedentes de caracterizar variação genética em milhares de pacientes a um custo acessível

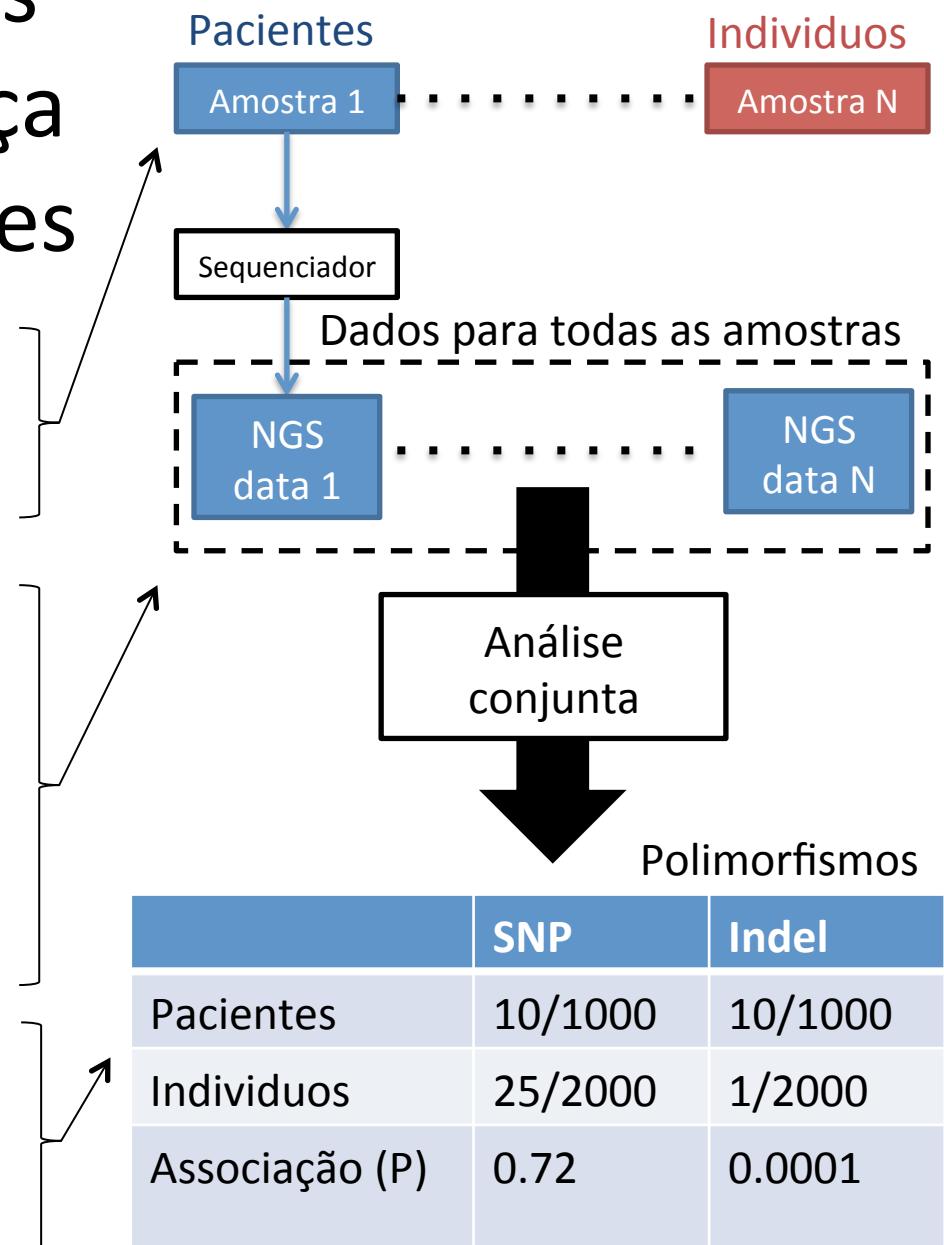
Bases não referência são coloridas;
bases referência são cinza

heterozigoto C/T



Como descobrir sites envolvidos em doença em três passos simples

1. Obter milhares de pacientes afetados e dezenas de milhares de indivíduos não afetados.
2. Sequenciar amostras
 - Descobrir polimorfismos (SNPs, indels, etc) entre as amostras.
 - Determinar o genótipo de todas as amostras em cada site variante.
3. Buscar diferenças sistemáticas em genótipos de pacientes vs indivíduos em todos os sites.



A análise de todos os projetos começa com uma matriz quadrada de variantes x amostras

The diagram illustrates a genotype matrix. On the left, a vertical blue arrow labeled "Todas as amostras (casos e controles)" points right, indicating the scope of the analysis. A horizontal blue arrow at the top indicates the number of samples. To the left of the matrix, a vertical blue arrow labeled "3M variantes" points down, indicating the number of variants. The matrix itself has "Site" and "Variante" as columns and "Amostra 1", "Amostra 2", "...", and "Amostra N" as rows. The matrix is divided into four main sections: SNP, Indel, CNV, and other variants. Each section contains two rows of data. The first row for each section shows a ratio (e.g., 1:10, 1:100, 1:1000), an allele pair (e.g., A/C, T/TC, T/), and a genotype call (e.g., 0/0, 0,10,100). The second row shows a ratio (e.g., 0/1, 0,20,200), an allele pair (e.g., 0/0, 0,20,200), and a genotype call (e.g., 0/0, 0,100,255). The "Genótipos" column on the right lists the three genotypes: 0/0 ref, 0/1 het, and 1/1 hom-alt. The "Verosimilhança" column on the right describes the phred-scaled probability of each genotype.

		Todas as amostras (casos e controles)					
		Site	Variante	Amostra 1	Amostra 2	...	Amostra N
SNP	1:10	A/C	0/0 0,10,100	0/1 20,0,200	...	0/0 0,100,255	
	1:100	T/TC	0/0 0,10,100	0/0 0,20,200	...	1/0 255,0,255	
Indel	1:1000	T/	0/0 0,10,100	0/1 20,0,200	...	0/0 0,100,255	
	
CNV	X:1234	G/T	0/1 10,0,100	0/1 20,0,200	...	1/1 255,100,0	
	

Genótipos:
0/0 ref
0/1 het
1/1 hom-alt

Verosimilhança:
A probabilidade A/B/C (phred-scaled) de a amostra ser hom (A), het (B), hom-alt (C).

...mas esses genomas tem que ser analisados consistentemente

Genomica
Clínica

Um paciente afetado



VS

Comparado com vários controles



Encontrar variantes consistentes com o modelo da doença (dominante, recessivo), normalmente condicionado nas variantes estarem ausentes / raras nos controles

Genomica
Pesquisa

Muitos pacientes afetados



VS

Comparado com vários controles



Encontrar variantes enriquecidos/depletados em individuos afetados, relativo aos controles

Todos individuos devem ser consistentemente analisados, para que as diferenças entre grupos sejam por **diferenças genéticas reais**

Infelizmente a análise não é tão simples

- Processo de amostragem dificulta a análise direta
- Distribuição de leituras (*reads*) não é uniforme ou independente
 - Algumas regiões tem redundância muito grande enquanto outras regiões possuem quase nenhuma.
- Sequenciadores cometem erros e tem limites
 - Erros podem ser sistemáticos ou aleatórios.
 - Sequências com alto conteúdo de G/C são difíceis de manipular/amplificar e portanto, sequenciar.
- Alinhamento não é perfeito
 - O genoma tem muitas regiões copiadas, de rápida mutação e mutações estruturais (recombinações, grandes deleções, etc.)

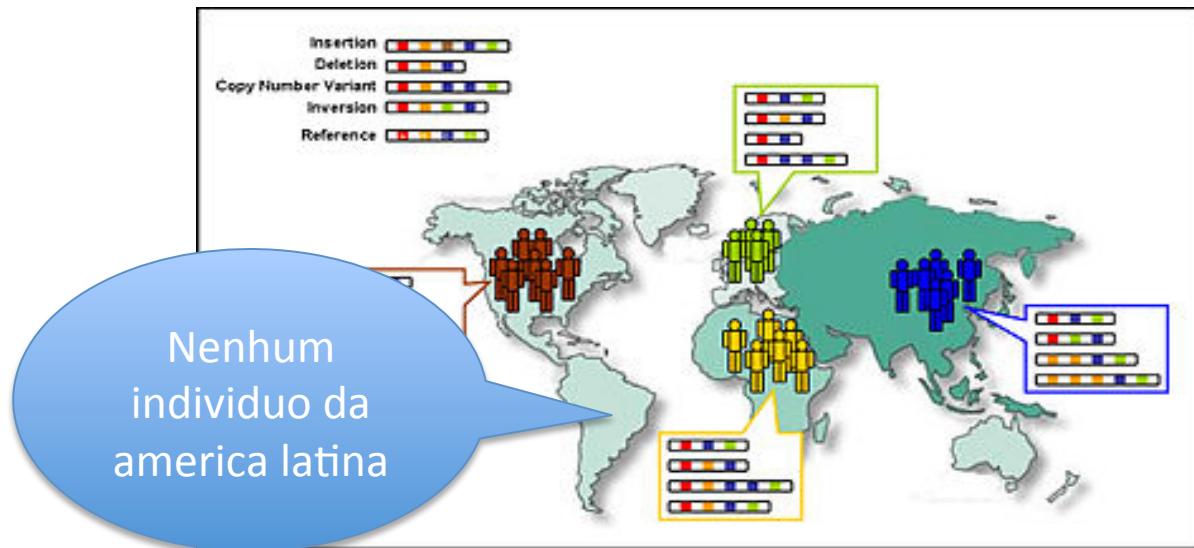
Regiões difíceis do genoma não podem ser trivialmente analisadas manualmente



VIABILIZANDO A ANÁLISE DE DADOS GENÉTICOS

1000 Genomes Project

- Consórcio com a missão de mapear todas as variantes com frequência superior a 1% na população mundial.
- Começou em 2007 e influenciou a maneira de fazer pesquisa na área com 22 instituições internacionais, envolvendo indústria e academia.
- Identificou imediatamente a necessidade de uma conduta mais sistemática da análise e processamento de dados e 11 grupos iniciaram o desenvolvimento colaborativo de algoritmos.
- Nosso grupo (GATK) representou o Broad Institute nesse consórcio. Em dois anos de projeto, se tornou o padrão internacional.



GATK (Genome Analysis Toolkit)

O que é o GATK:

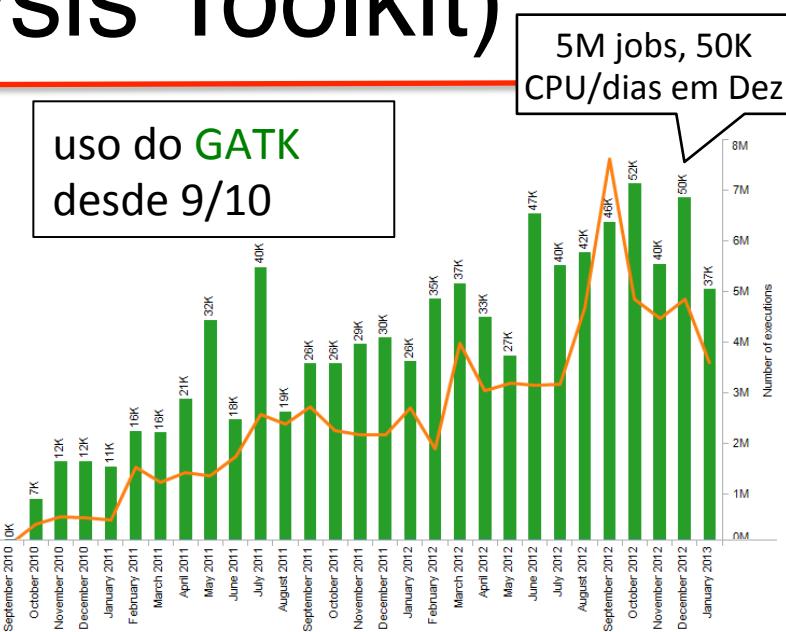
- Um kit de ferramentas para análise de dados genéticos produzido e distribuído livre e aberto pelo Broad Institute.
- O GATK também é uma plataforma de desenvolvimento.

Impacto

- Mais de 700 ferramentas (muitas desenvolvidas fora do Broad Institute).
- Várias ferramentas do GATK são hoje o padrão na indústria para análise de dados genéticos.
- Mais de 3,000 visitas únicas por dia no site do GATK.
- Principal tecnologia em grandes projetos: 1000 Genomes, TCGA, ESP e a maioria dos grandes projetos financiados pelo NHGRI
- Base para o *Archon X Prize*, *Genomes in a Bottle* e outros padrões internacionais de referência

Publicações

- DePristo et al. A framework for variation discovery and genotyping using next-generation DNA sequencing data. (2011) *Nat. Genet.*
- McKenna et al. (2010) The Genome Analysis Toolkit: a MapReduce framework for analyzing next-generation DNA sequencing data. *Genome Res.*



Análise de variação de SNP e Indels é um problema de modelagem bayesiana

Prior do genótipo Verossimilhança do genótipo

$$\Pr\{G|D\} = \frac{\Pr\{G\} \Pr\{D|G\}}{\sum_i \Pr\{G_i\} \Pr\{D|G_i\}}, \text{ [Bayes' rule]}$$
$$\Pr\{D|G\} = \prod_j \left(\frac{\Pr\{D_j|H_1\}}{2} + \frac{\Pr\{D_j|H_2\}}{2} \right) \text{ where } G = H_1 H_2$$

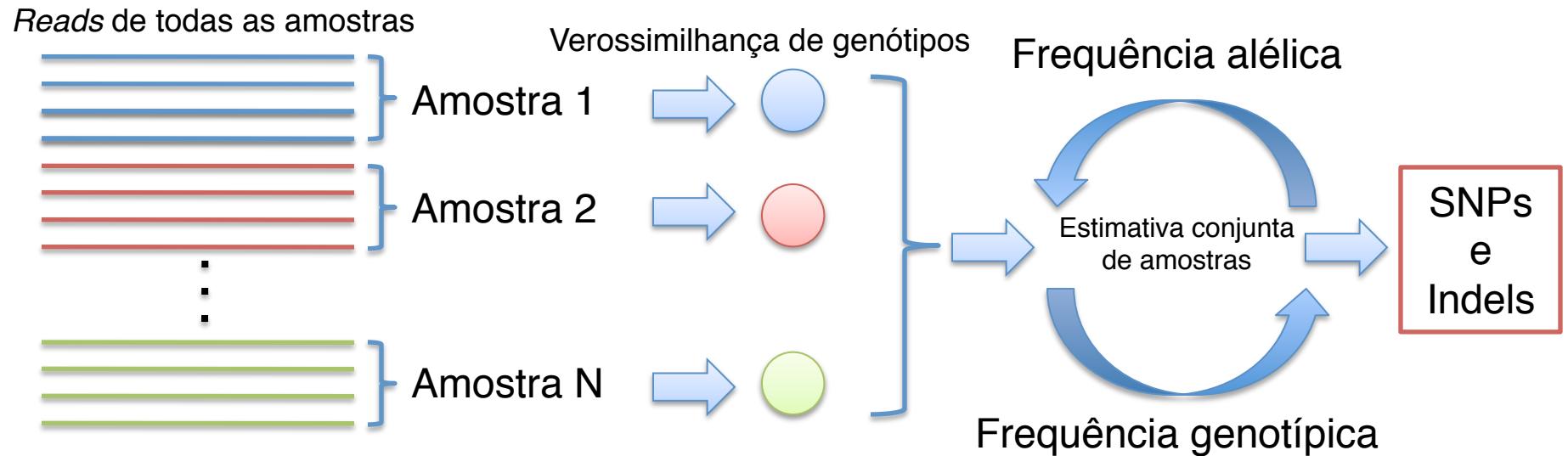
Modelo Bayesiano

Pr{D|H} is the haploid likelihood function

premissa diplóide

- Inferencia: qual é o genótipo G de cada amostra dado a observação das *reads* D para cada amostra?
- Calcula-se via lei de Bayes' a probabilidade de todos os G possíveis
- Expansão produtorial assume que as *reads* são independentes
- Depende de uma função de versossimilhançapara estimar a probabilidade de cada amostra dado um haplotipo candidato

Análise de multiplas amostras integra as verossimilhanças para estimar a frequênciade variação

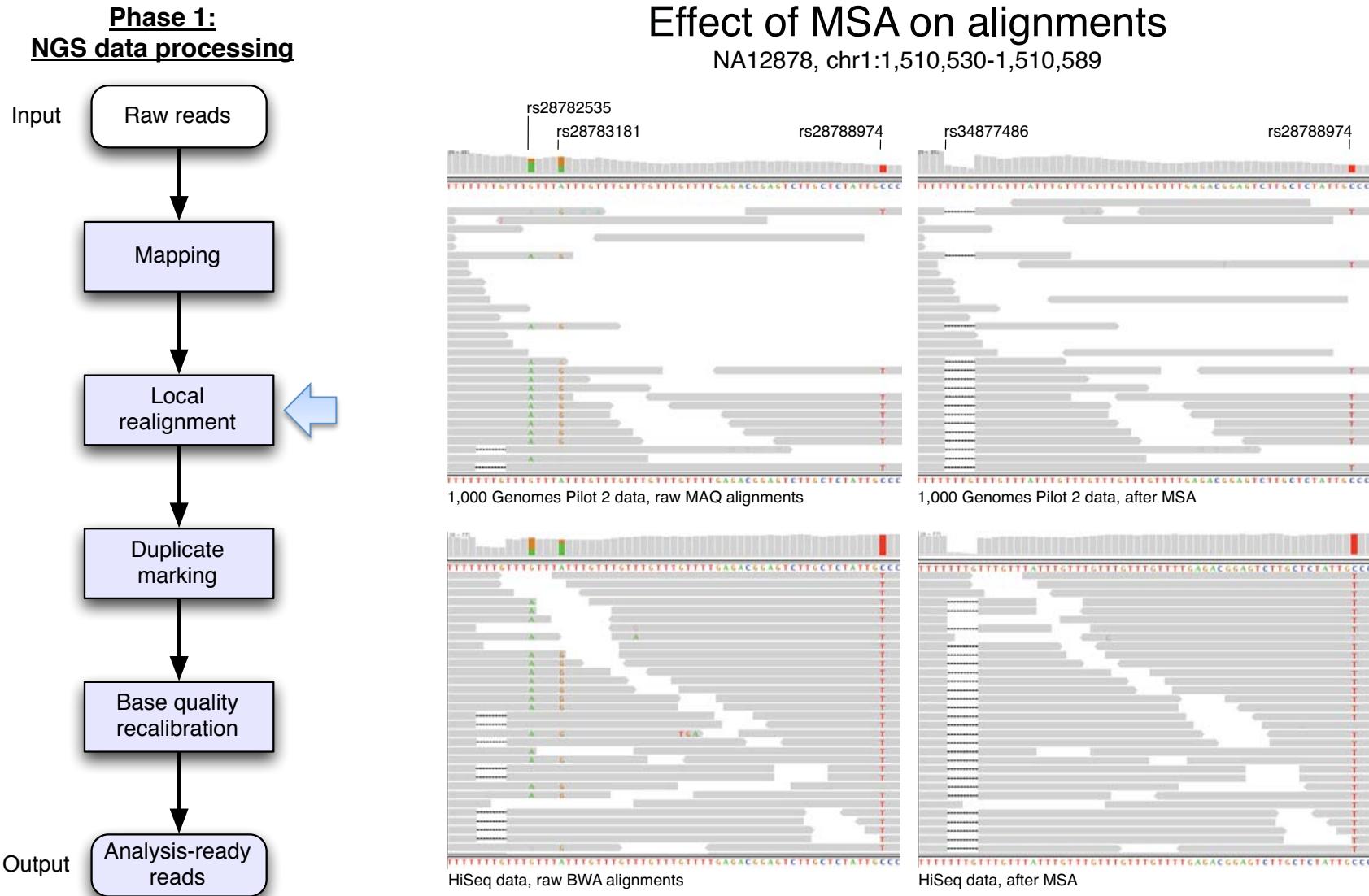


- Estimativa simultânea de:
 - Espectro de frequênciade alélica $\Pr\{AF = i | D\}$
 - A probabilidade que a variante existe $\Pr\{AF > 0 | D\}$
 - Designação de genótipos para cada amostra

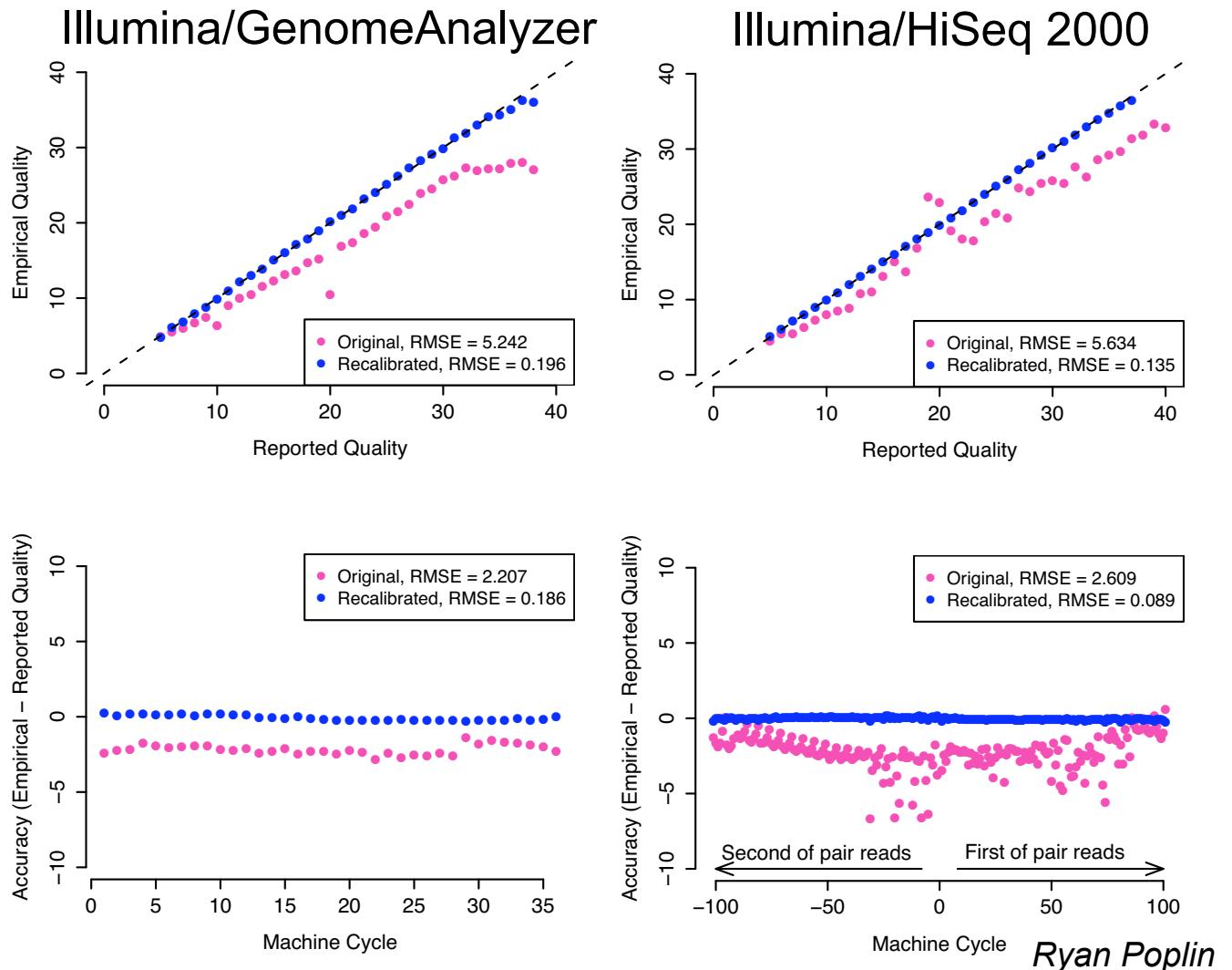
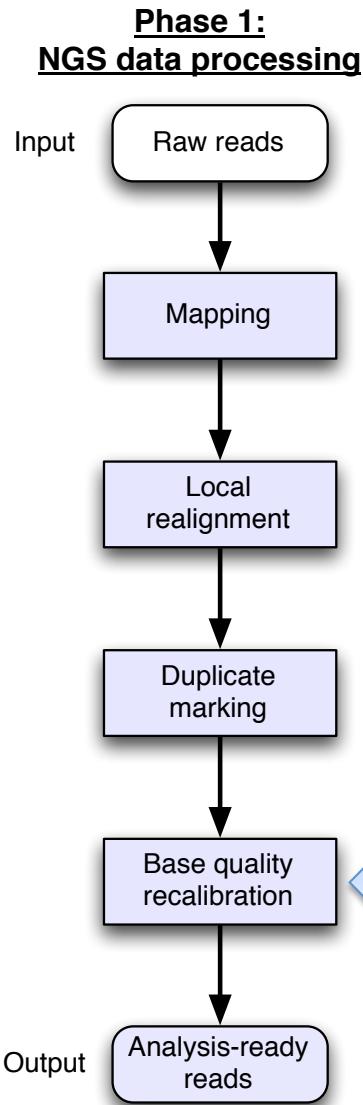
Dificuldades da avaliação dos genótipos usando a modelagem Bayesiana

- O **alinhamento** das *reads* não concordam em eventos de inserção/deleção principalmente em zonas de alta ambiguidade (ex: homo-polímeros)
- A função de verossimilhança (que depende da **estimativa de erro** para cada base nas *reads*) não possuía a acurácia necessária para distinguir erro de variação real
- Apesar do rigor estatístico, **falsas variantes** são impossíveis de ser detectadas com uma visão local dos dados.

Melhor alinhamento através de realinhamento local de inserções e deleções



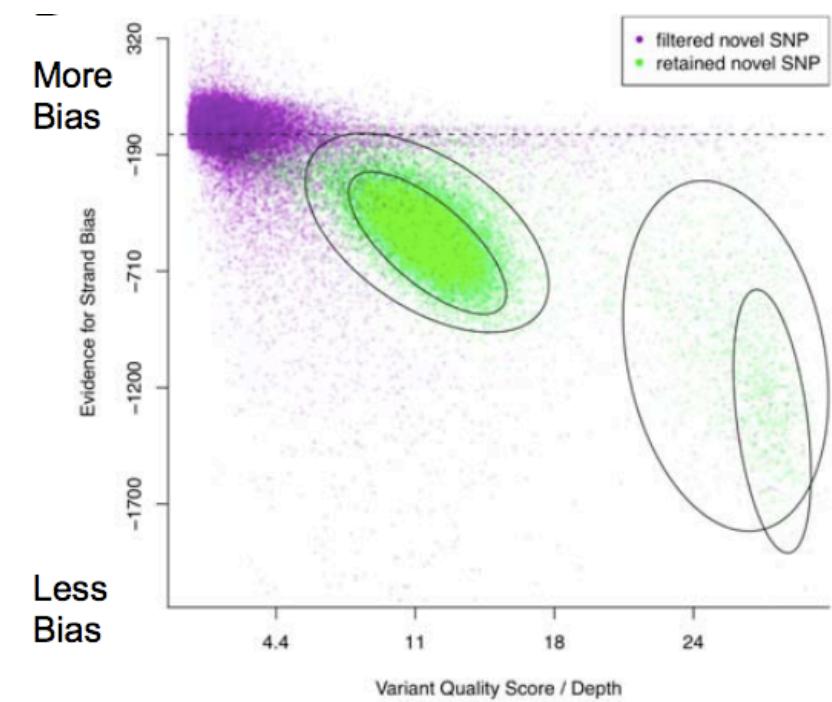
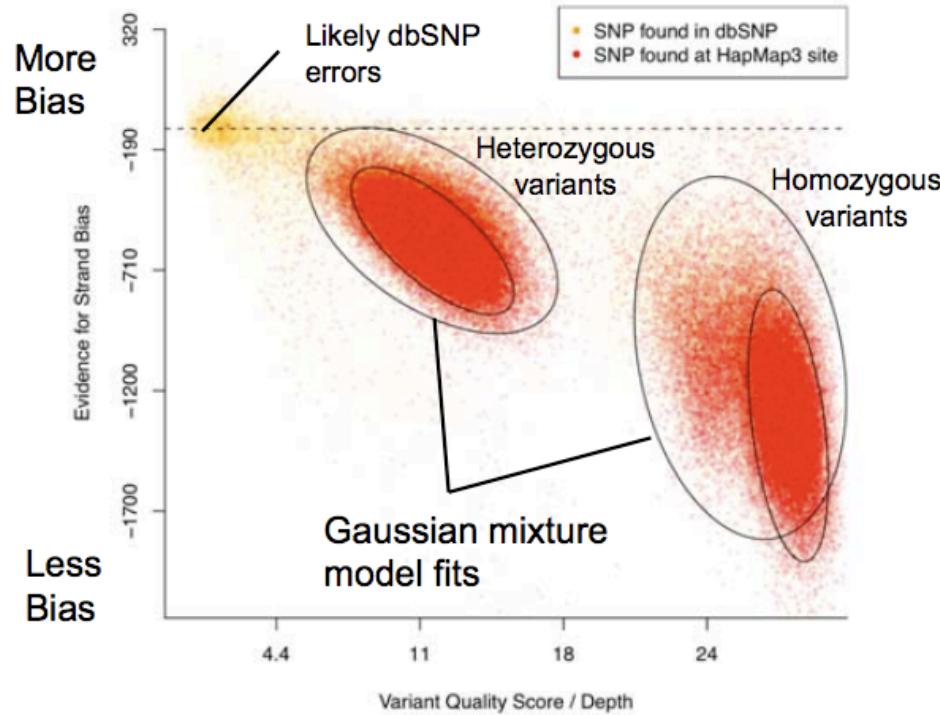
Melhor estimativa de erro através de análise sistemática de probabilidades de cada base



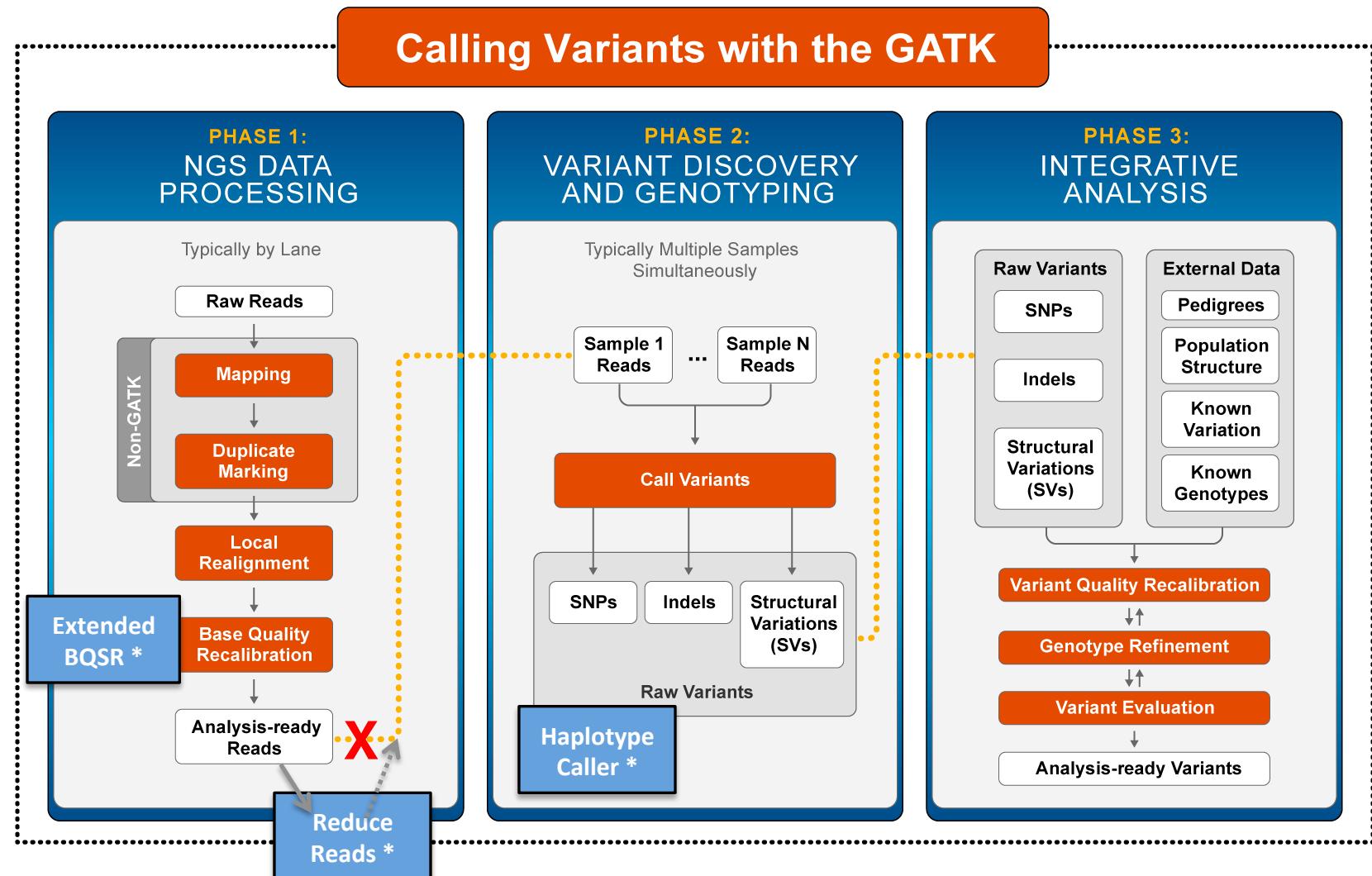
Usando *machine learning* para identificar falsas variantes devido a erro/bias de sequenciamento

Treina-se *Gaussian mixture model* de covariantes de erro usando sites de variantes validadas

Re-avaliação de novos sites através de consistência com GMM clusters



Enfim, um “best practices pipeline” para análise

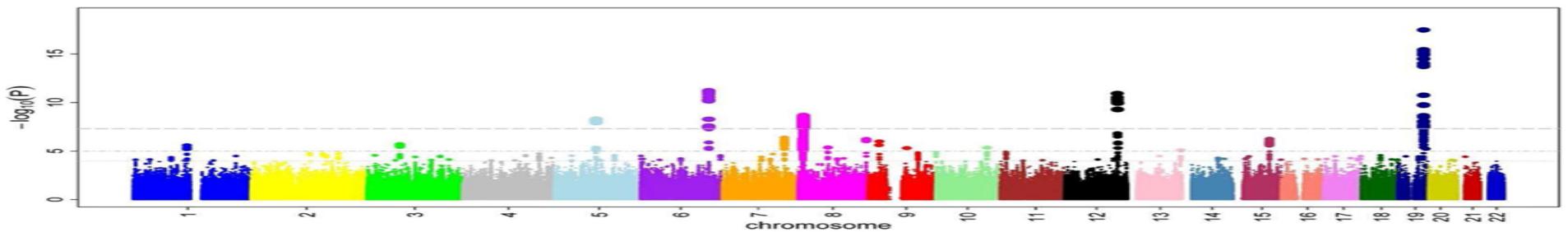


Veja <http://www.broadinstitute.org/gatk> para mais informações

QUEBRANDO AS LIMITAÇÕES DE HOJE

Associação genética de esquizofrenia

- Projeto começou logo após o sequenciamento do genoma humano com ~100 pacientes e não encontrou nenhuma correlação genética significativa.
- Após segunda etapa de financiamento, projeto expandiu para 700 pacientes e 4 gens (figura abaixo) se mostraram potenciais candidatos significativos, outros ainda sem evidencia suficiente
- Limitações computacionais impediram a análise da 3^a fase do projeto com ~18,000 pacientes que confirmaria os gens candidatos.



Entendendo as limitações computacionais

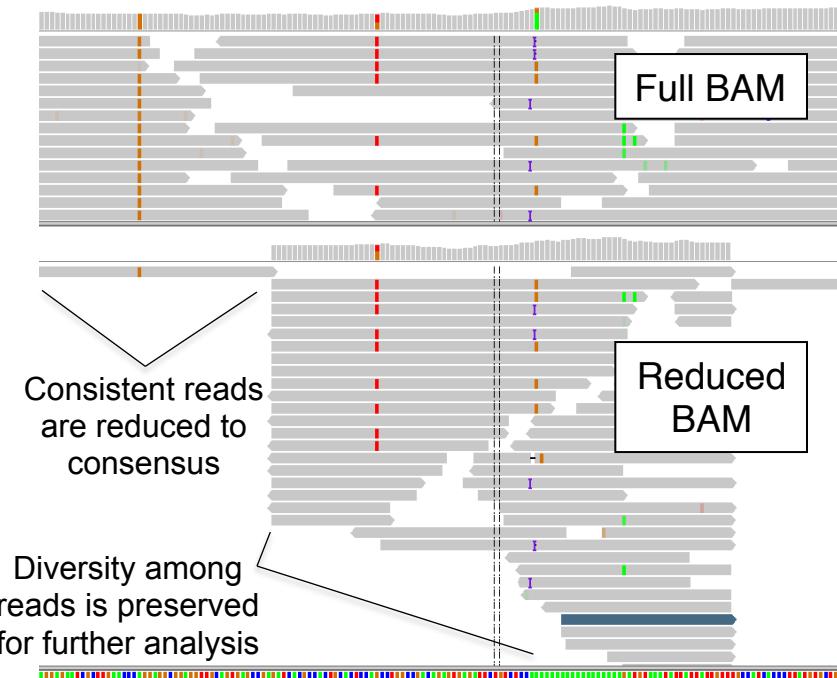
- Processamento de 14,000 exomas precisa manipular 280,000 Gb de dados (**280Tb**)
- Outros projetos no Broad Institute já estão sequenciando mais de 30,000 exomas para estudos de associação (ex: Diabetes tipo 2), isso equivale a **600Tb** de dados. O objetivo de processar 1M de exomas até o final de 2013 traduzem em **20Pb**.
- Cada amostra precisa passar por todas as etapas de processamento (best practices pipeline) que envolve aproximadamente 2-4 cpus/dia de computação por exoma, traduzindo em **28,000-60,000 cpus/dia** só de processamento (fora análise conjunta – que é inviável em termos de memória e processamento atualmente).
- Eses números seriam na ordem de *petabytes* (**18Pb** para diabetes) se os dados fossem do genoma completo.

Perspectiva: A base de dados completa de busca web do Google é de **850Tb**

Tipo de dado	Tamanho em disco
Exoma completo	~20Gb
Genoma completo (30x)	~600Gb

Um algoritimo para reduzir a representação de dados genéticos

Tamanho original	~ 20 GB (exome)
Tamanho reduzido	~ 100-200 MB
Compressão	~ 50x-100x
SNP calling	mesmo
Indel calling	mesmo
Tempo de execução	>50x faster



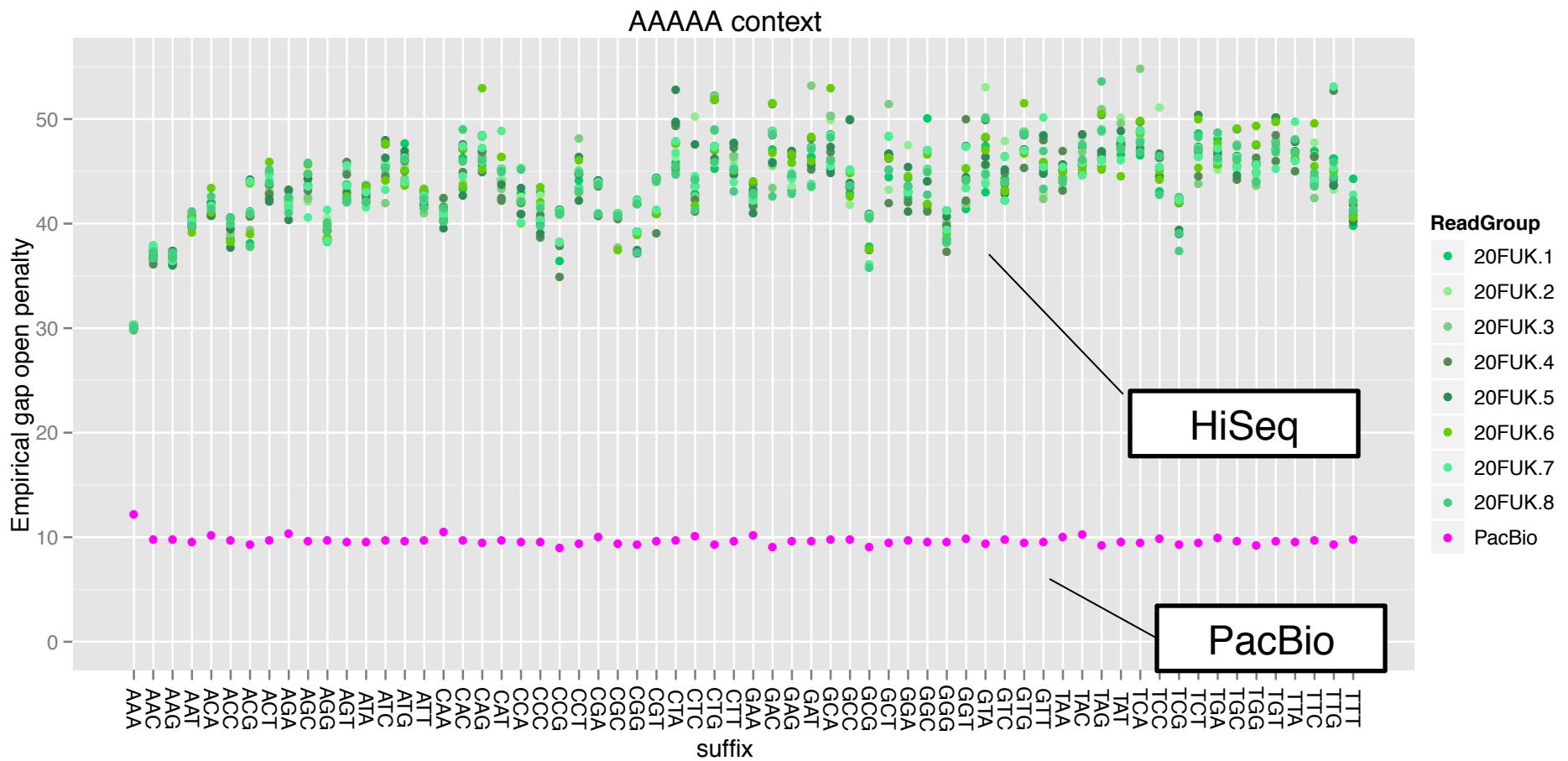
Usando esse algoritimo, o projeto de esquizofrenia identificou 78 gens com alta confiança ($p < 10^{-20}$) e iniciou um projeto de análise de genoma completo.

Quais são as dificuldades que enfrentamos hoje?

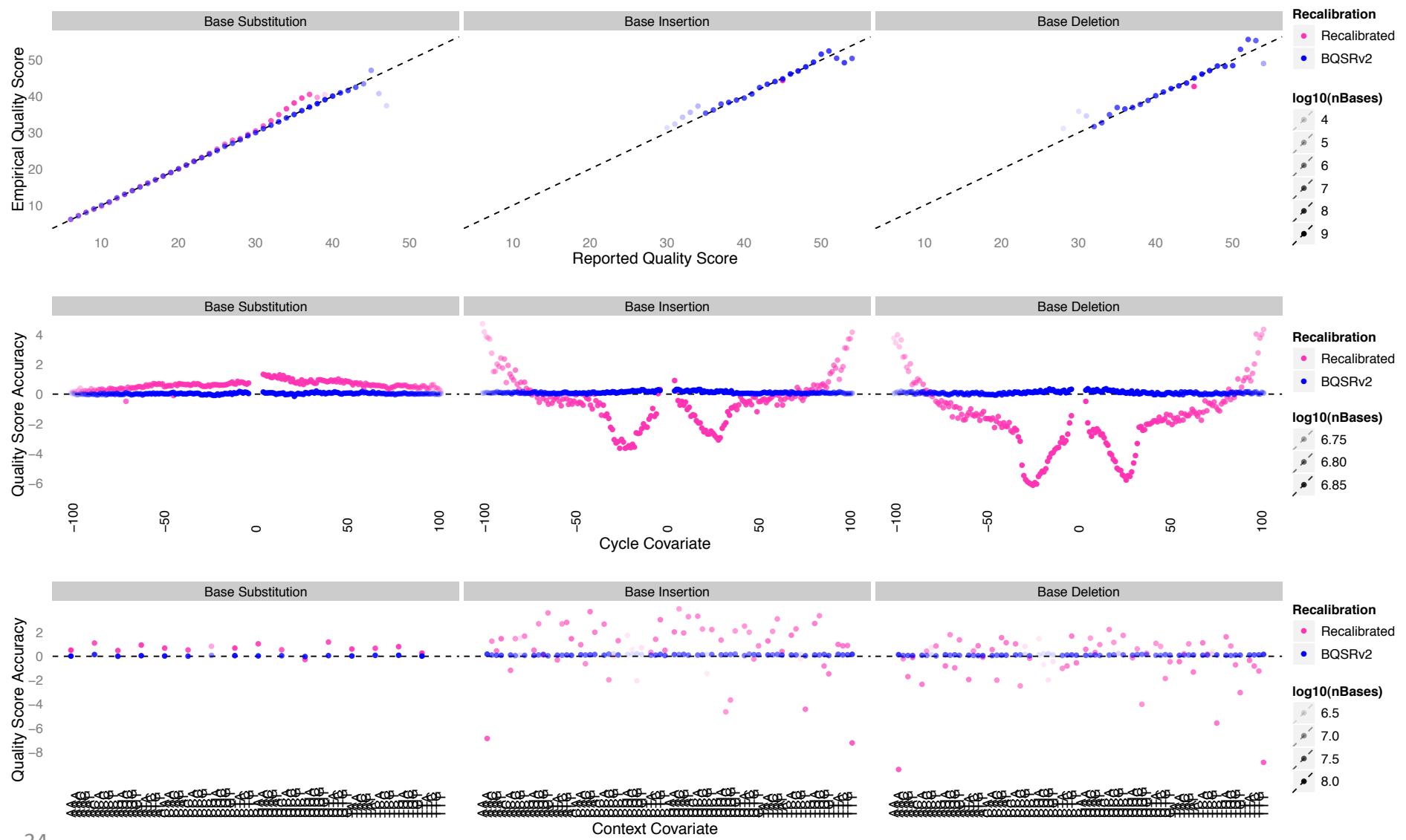
- **Limitações técnicas**
 - Falta de cobertura em regiões importantes do genoma (gens associados a doenças!)
 - Processo de erro complexo
 - Ainda confunde-se erro de sequenciamento com variantes reais. (falso positivos)
 - Procedimento tem que ser muito conservador, e força a perda de variantes reais em regiões difíceis (falso negativos)
- **Limitações analíticas**
 - Mal interpretação de regiões difíceis com os modelos atuais.
 - Inserções, deleções e variantes estruturais ainda são o grande desafio.
- Não há **poder estatístico** para identificar associação em estudos genéticos.
 - Precisamos analizar mais amostras
 - Para isso, é necessário que abixe o custo de sequenciamento e o agregamento de dados já existentes no mundo.

Variante	Taxa de validação média
SNPs	99%
Indels curtos	80%
Indels longos	50%

Taxa de erro de indel varia em diferentes plataformas



Expandindo o conceito de probabilidade de erro empírica em cada base para indels

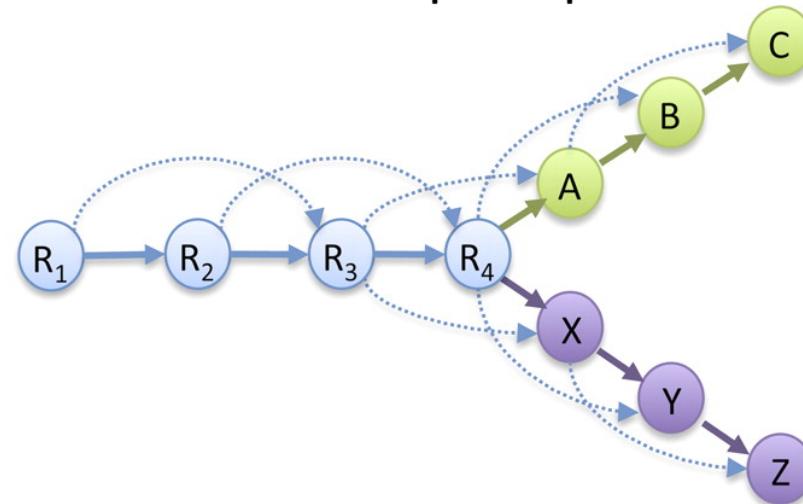


Usando *de novo assembly* local para propor haplotypes para o modelo de identificação de variantes

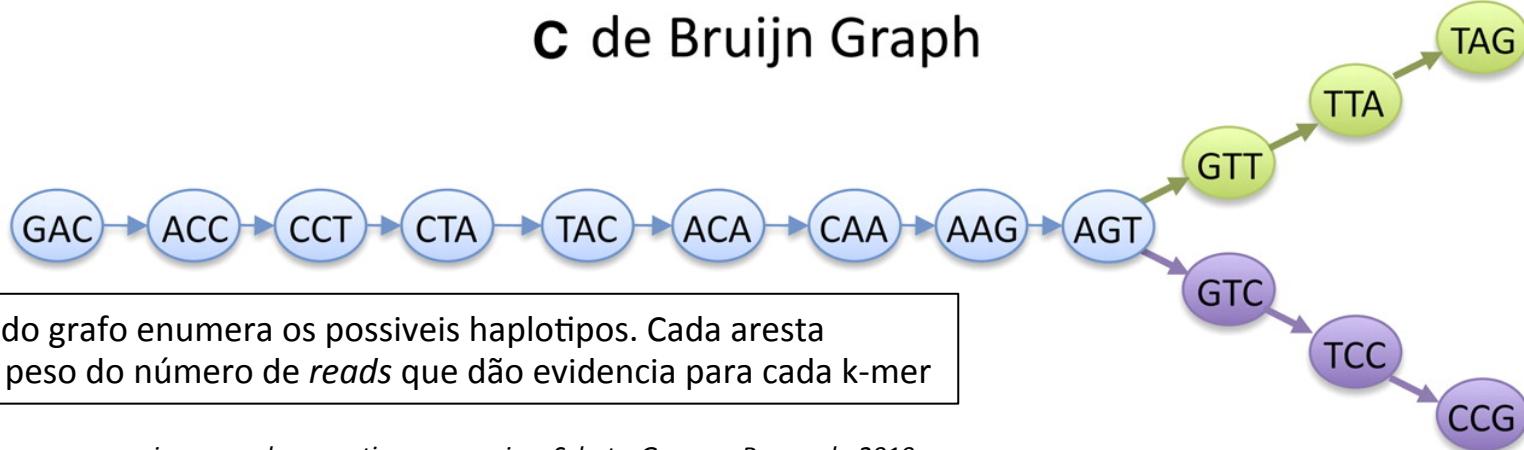
A Read Layout

R ₁ :	GACCTACA
R ₂ :	ACCTACAA
R ₃ :	CCTACAAAG
R ₄ :	CTACAAGT
A:	TACAAGTT
B:	ACAAGTTA
C:	CAAGTTAG
X:	TACAAGTC
Y:	ACAAGTCC
Z:	CAAGTCCG

B Overlap Graph



C de Bruijn Graph



Travessia do grafo enumera os possíveis haplotipos. Cada aresta contém o peso do número de *reads* que dão evidencia para cada k-mer

We often find consistent (artifactual) alleles at the sites of larger events because they cannot be properly modeled by the mappers



Novos algoritmos aumentam significativamente a precisão de regiões difíceis e indels

Tipo de variante	Validação antiga (gold standard)	Novos algoritmos (2013)
SNPs	99%	99%
Indels curtos	80%	97%
Indels longos	50%	93%

Todos esses programas já estão disponíveis na última distribuição do GATK (versão 2.5), e o desenvolvimento tem sido intenso nos últimos meses.

Agregando dados

O FUTURO DO SEQUENCIAMENTO DE DADOS MÉDICOS

O maior limite hoje é o tamanho da amostra

- Suponha o sequenciamento de 500 pacientes com Alzheimer.
- A descoberta de um indel que causa perda de função em um gene ligado ao cérebro está presente em 10 pacientes
- O que eu posso dizer sobre a relação da variante com a doença?

Associação da variante 20:12345 C/CAT *indel* com risco de Alzheimer's (exemplo fictício)

Comparação	Usando só 500 pacientes	Contra 500 controles	Contra 1K controles	Contra 10K controles	Contra 100K controles
Pacientes afetados	10/1000	10/1000	10/1000	10/1000	10/1000
Controles	None	0/1000	1/2000	10/20000	100/200000
Associação (P-value)	None	10^{-2}	10^{-4}	10^{-8}	10^{-10}

O que eu aprendi? Eu deveria ter sequenciado controles Nem remotamente significativo Indistinguível do ruído Praticamente significativo! Descoberta importante!!

Agregação de dados genéticos em massa: o futuro da medicina genética

- Custo de sequenciamento já caiu 1 milhão de vezes, e permitiu a explosão de sequenciamento sobre a base genética de doenças
- Através da agregação de todos os dados clínicos sequenciados do mundo, será possível acelerar drasticamente o processo de descoberta.
 - exemplo, doenças genéticas raras afetam 100 crianças nascidas por ano. Cada hospital vê zero ou um caso. Sozinhos, nada é aprendido. Combinados, o resultado é evidente.
 - exemplo, o poder de associar variantes com doenças aumenta drasticamente com o número de amostras de controle. Amostras afetadas podem ser usadas como controle para *outras doenças*.
- Dados genéticos agregados vão ser necessários para guiar a interpretação da medicina diagnóstica do futuro para toda a prática clínica

O desafio dessa geração

- Comunidades científica, médica e de pacientes ainda não estão organizadas para concretizar essa oportunidade e nem sequer estão seguindo esse caminho.
- Dados vivem em silos: por instituições, doenças, tecnologia utilizada, projetos,... sem visibilidade mútua.
- Não existem procedimentos regulatórios para viabilizar o compartilhamento de informação médica e genética, nem sequer para pesquisa.
- Não existe expertise e capacitação computacional nem padronização mundial para que os resultados tenham relevância em compartilhamento

Como uma comunidade precisamos construir uma plataforma para armazenar, processar, analisar e interpretar uma quantidade extraordinária de dados genéticos para pesquisa e aplicações clínicas