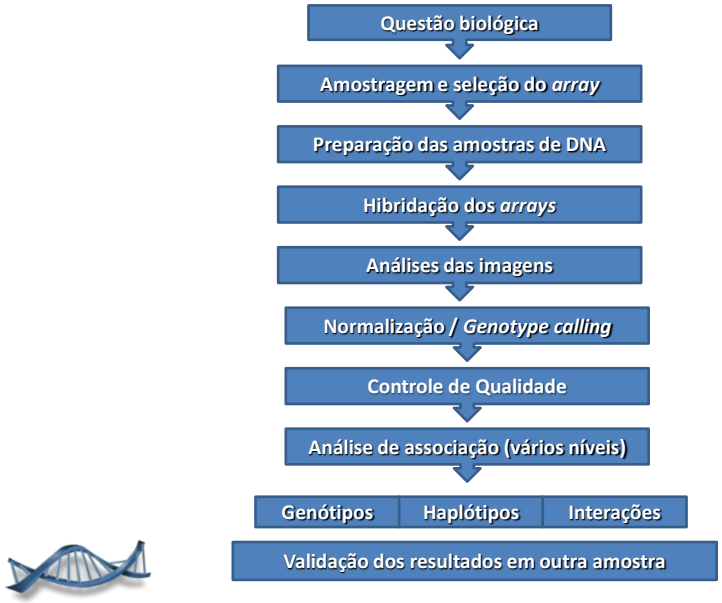
Controle de qualidade de SNPs e de Amostras

Prof. Dr. Fernando Baldi
UNESP-FCAV
Marcos Vinicius Barbosa da Silva
EMBRAPA GADO DE LEITE

Programa do dia

- ✓ **Controle de Qualidade com painéis de alta densidade**
- ✓ **Estudos de associação ampla com um único marcador**
- ✓ **Estudos de associação ampla utilizando haplótipos**
- ✓ **Abordagem de genes idênticos por descendência (IBD)**
- ✓ **Associação com marcadores múltiplos**
- ✓ **Problemas nos estudos de associação ampla do genoma**
- ✓ **Métodos de seleção de SNPs para estudos de associação ampla**
- ✓ **Softwares para estudos de associação ampla**

Etapas em um estudo de associação ampla do genoma

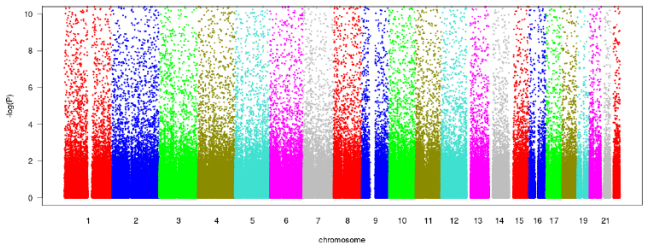


Ziegler et al. (2008)

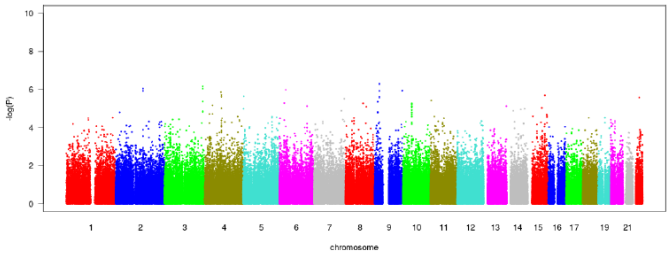
Controle de Qualidade em GWAS

- As tecnologias de genotipagem em larga escala com painéis de alta densidade são muito precisas, mas não são perfeitas .....
  - Alta automatização
- Mesmo com uma precisão de 99,9%, uma análise do genoma de 2.000 indivíduos para 500.000 marcadores da **1.000.000 erros de genotipagem**!
- Erros de genotipagem podem aumentar o número de resultados falso-positivos (falsos SNP significativos)

Por que o controle de qualidade dos dados é importante?



Ger MI FS I, Affymetrix 500k array set, SNPs on chip: 493,840

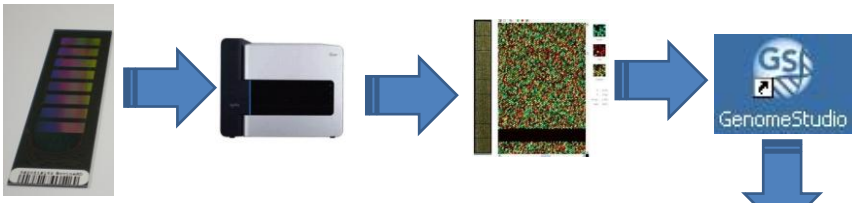


Ger MI FS I, Affymetrix 500k array set, SNPs on chip: 493,840
SNPs passing standard quality control: 270,701

Controle de Qualidade em GWAS

- A qualidade dos dados é importante e os dados genômicos devem ser verificados cuidadosamente:
  - ✓ Padrões incomuns de dados faltantes
  - ✓ Problemas de análises laboratoriais
  - ✓ Problemas nas amostras e chips (efeito de lote)
- Intensidade dos "spots" e atribuição dos genótipos
  - ✓ intensidade de sinal do "spot" é convertido em genótipos
- Depois de atribuídos os genótipos, análises intensivas de controle de qualidade precisam são executadas

Controle de Qualidade em GWAS



Dados Genotípicos consistentes e sem erros

GWAS, GS

Controle de Qualidade em GWAS

Passos no controle de qualidade:

- **Falta de alelos ou perda de Genótipos**
- **Alelos em baixa frequencia (MAF)**
- **Marcadores com alelos fixados ou alta proporção de heterozigotos**
- **Desvios do Equilíbrio de Hardy-Weinberg (H-W)**
- **Controle a nível de indivíduo**
- **Estratificação da população**

Ziegler et al. (2008)

Dados de genótipos perdidos

- Problemas para inferir o genótipo a partir da intensidade de sinal ("*call rate*")
- SNPs são de qualidade questionável se a sua genotipagem falhou em muitos indivíduos
- Problemática em análises múltipla (alguns indivíduos tem, outros não)
- Uma solução prática é a imputação de dados

Menor Frequência de um dos alelos (MAF)

- Menor frequência de um dos alelos (MAF): refere-se a frequência na qual o alelo menos comum ocorre em uma determinada população.
  - muito poucas observações sobre esses alelos
- Por causa do baixo poder para detectar uma associação entre o SNP e a característica de interesse, é razoável excluir esses SNPs.
- O critério de filtro normalmente varia com o tamanho da amostra e os valores variam de 1% a 5%.

Equilíbrio de Hardy-Weinberg (HWE)

- Desvio de HWE podem ser devido à endogamia, mutação, estratificação da população, seleção etc.
- Controle de qualidade: desvios aparentes do HWE por causa de uma tendência de atribuir os homozigotos como heterozigotos (excesso de indivíduos heterozigotos). Critério utilizado p-value<10^{-6}
- Teste de desvios HWE pode ser realizado pelo teste de Pearson ou teste exato de Fischer
  - R "*genetics package*" que implementa ambos os testes de Pearson (*chi-square*) e Fisher

Controle de Qualidade por Amostra

- Excluir indivíduos com menos de 90% dos SNPs genotipados com sucesso (*call rate*)
  - Incluindo os SNPs monomórficos
- **Nível de heterozigose**
  - Um nível muito alto pode ser um indicador de contaminação de DNA. Estimar a média e o desvio padrão da heterozigosidade em todos os indivíduos e excluir todos aqueles fora do intervalo média ± SD*3.
- **Parentesco entre indivíduos** (*Mendelian error:* <1%) e Sexo ID
  - Alelos de um indivíduo diferentes dos alelos dos seus pais biológicos
- **Estratificação da população**
  - Diferenças nas frequências alélicas
  - Pode levar a um viés substancial nas análises estatísticas
  - Análise estratificada por localização geográfica ou grupo (*clusters*)

O Bovine HDBeadchip com Genoma de Nelore

- 777.962 SNP disponíveis
- 48 touros Nelore genotipados
- 332.207 SNP foram eliminados
  - 2.959 sem cromossomo definido
  - 7.198 *Baixo Call rate*
  - 16.222 Problemas de cluster
  - 3.379 Excesso de Heterozigotos
  - 1.353 MAF < 0.05
  - 280.229 monomorficos
- 445.755 SNP para análise

Baldi et al. (2012)

Software

- PLINK

http://zzz.bwh.harvard.edu/plink/

- SVS Golden Helix

https://www.goldenhelix.com/products/SNP_Variation/index.html

- Pacote R snpStats

https://www.bioconductor.org/packages/release/bioc/html/snpStats.html