UNIVERSIDADE DE SÃO PAULO
INSTITUTO DO CORAÇÃO – HCFMUSP
LABORATÓRIO DE GENÉTICA E CARDIOLOGIA MOLECULAR

DATA CLEANING

Projeto EPIGEN

Investigador principal: Alexandre C. Pereira
Coordenadora: Andréa R. V. Russo Horimoto
Participantes: José Mauricio Sanches
Enio Akira Oishi
a definir

Belo Horizonte
2012

Preparação do arquivos para análise

- PLINK (Purcell et al., 2007);
- algoritmo Perl para gerar arquivos .map e .ped;
- outputs Affymetrix

BAEPENDI

#CHP File=C:\Pereira\20091120_164832\1186ACP**0007**-20091118-GWS6.birdseed-v2.chp
#Exec GUID=0000044921-1258757365-0000019096-0000025431-0000022945

SNPID	Call	Confidence
SNP_A-2131660	BB	0.001483649
SNP_A-1967418	AA	0.005017896
SNP_A-1969580	BB	0.006377246
SNP_A-4263484	AB	0.004044221

1186-ACP-1	2009-07-13	female	110601-4
1186-ACP-2	2009-07-13	female	19910-1632
1186-ACP-3	2009-07-13	female	105711-611
1186-ACP-4	2009-07-13	male	93817-1089

Data cleaning
Preparação de arquivos para análise

BAEPENDI

"Probe Set ID","dbSNP RS ID","Chromosome","Physical Position","Strand","ChrX pseudo-autosomal region 1","Cytoband","Flank","Allele A","Allele B","Associated Gene","Genetic Map","Microsatellite","Fragment Enzyme Type Length Start Stop","Allele Frequencies","Heterozygous Allele Frequencies","Number of individuals/Number of chromosomes","In Hapmap","Strand Versus dbSNP","Copy Number Variation","Probe Count","ChrX pseudo-autosomal region 2","In Final List","Minor Allele","Minor Allele Frequency","% GC","OMIM"

"**SNP_A-1780419**","rs6576700","1","84647761","-","0","p31.1","GGATACATTTTATTGC[A/G]CTTGCAGAGTATTTTT", "A","G","NM_058248 // intron // 0 // Hs.129142 // DNASE2B // 58511 // deoxyribonuclease II beta /// NM_021233 // intron // 0 // Hs.129142 // DNASE2B // 58511 // deoxyribonuclease II beta /// ENST00000361540 // intron // 0 // Hs.129142 // DNASE2B // 58511 // deoxyribonuclease II beta /// ENST00000370665 // intron // 0 // Hs.129142 // DNASE2B // 58511 // deoxyribonuclease II beta /// ENST00000370662 // intron // 0 // Hs.129142 // DNASE2B // 58511 // deoxyribonuclease II beta","108.402547694204 // D1S2889 // D1S2766 // --- // --- /// 116.765222864103 // D1S2889 // D1S2766 // --- // --- /// 102.95735950781 // --- // --- // TSC586552 // TSC615456","D1S2889 // downstream // 39379 /// D1S1746E // upstream // 140114","StyI // CCATGG_CCTTGG // 1336 // 84647550 // 84648885 /// NspI // GCATGT_ACATGC // 798 // 84647137 // 84647934","0.55 // 0.45 // Caucasian /// 0.6 // 0.4 // Han Chinese /// 0.6222 // 0.3778 // Japanese /// 0.5167 // 0.4833 // Yoruban","0.5333 // Caucasian /// 0.5778 // Han Chinese /// 0.5333 // Japanese /// 0.5 // Yoruban","60.0 // Caucasian /// 45.0 // Han Chinese /// 45.0 // Japanese /// 60.0 // Yoruban","YES","reverse","---","6","0","YES","G // Caucasian /// G // Han Chinese /// G // Japanese /// G // Yoruban","0.45 // Caucasian /// 0.4 // Han Chinese /// 0.3778 // Japanese /// 0.4833 // Yoruban","0.3827352","---"

BAEPENDI

.map

1	rs12565286	0	711153
1	rs12082473	0	730720
1	rs3094315	0	742429
1	rs11240776	0	755132
1	rs2980319	0	766985
1	rs2980300	0	775852

.ped

1 1101 0 0 1 -9 Genótipos....
41 41301 3801 41101 1 -9 Genótipos...

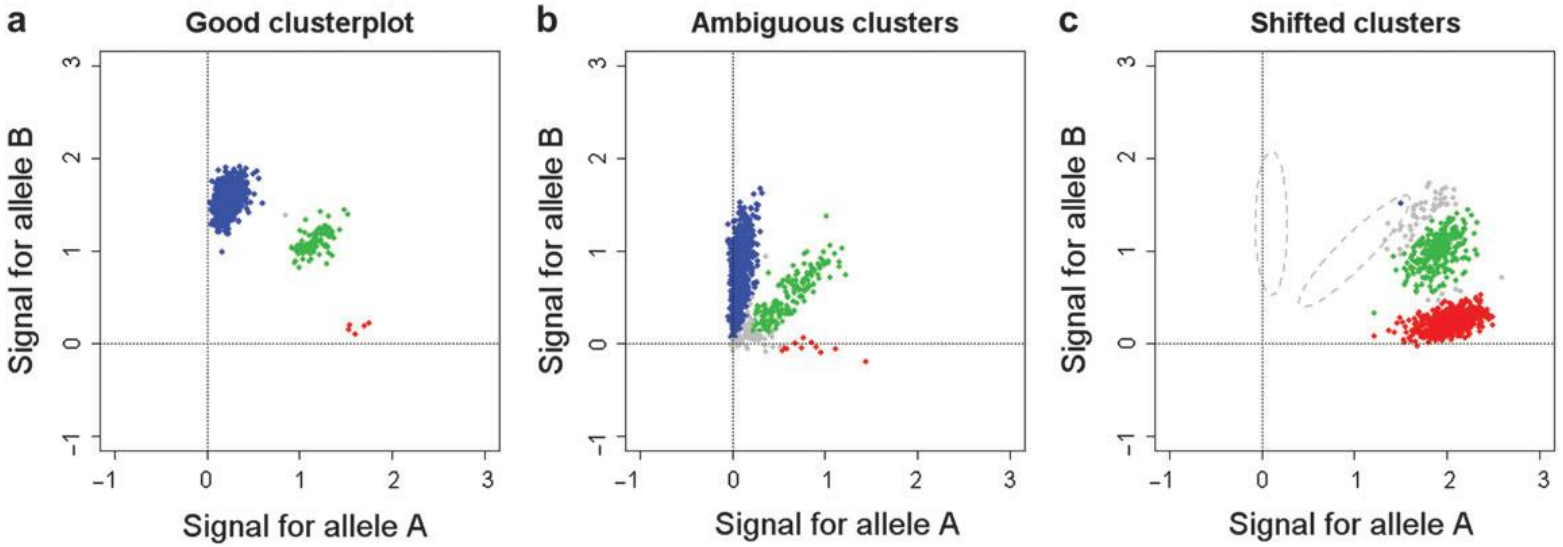
Análise exploratória dos dados

1) Controle de qualidade dos SNPs;

2) Controle de qualidade das amostras.

Data cleaning
Controle de qualidade dos SNPs

Avaliação gráfica dos perfis de hibridização, via clusterplots



Data cleaning
Controle de qualidade dos SNPs

a) avaliação das perdas genotípicas:

a remoção de SNPs, cujo genótipo não pode ser estimado em um grande número de amostras, reduz a probabilidade de manter variantes com genotipagem não acurada na análise.

b) frequência do alelo menor (MAF):

essas variantes estão mais sujeitas a erros de genotipagem e são menos informativas para os estudos de associação

c) análise do equilíbrio de Hardy-Weinberg :

matematicamente, um SNP está em EHW se a probabilidade de observar um dado genótipo for igual à probabilidade de observar os alelos independentemente. Isso pode ser verificado através de um teste de qui-quadrado ou pelo teste exato de Fisher.

Data cleaning
Controle de qualidade dos SNPs

BAEPENDI

Table 1: Genotype Missing

cutoff	ndrop	propdrop	nkeep	propkeep
0.1	**36575**	**3.9**	**893379**	**96.1**
0.05	64123	6.9	865831	93.1
0.01	227752	24.5	702202	75.5

Table 2: Minor Allele Frequency

cutoff	ndrop	propdrop	nkeep	propkeep
0	0	0	929954	100
0.001	34358	3.7	895596	96.3
0.01	**72727**	**7.8**	**857227**	**92.2**
0.05	209531	22.5	720423	77.5
0.1	318504	34.2	611450	65.8

Table 3: Hardy-Weinberg

cutoff	ndrop	propdrop	nkeep	propkeep
1.00e-07	274	0.03	929680	99.97
1.00e-04	**980**	**0.1**	**928974**	**99.9**
0.001	3096	0.3	926858	99.7
0.01	14173	1.5	915781	98.4
0.05	50093	5.4	879861	94.6

Data cleaning
Controle de qualidade das amostras

a) identidade por estado (IBS):

desvios na proporção esperada de alelos compartilhados entre todos os possíveis pares de indivíduos incluídos no estudo são úteis para identificar problemas amostrais, como a ocorrência de amostras duplicadas.

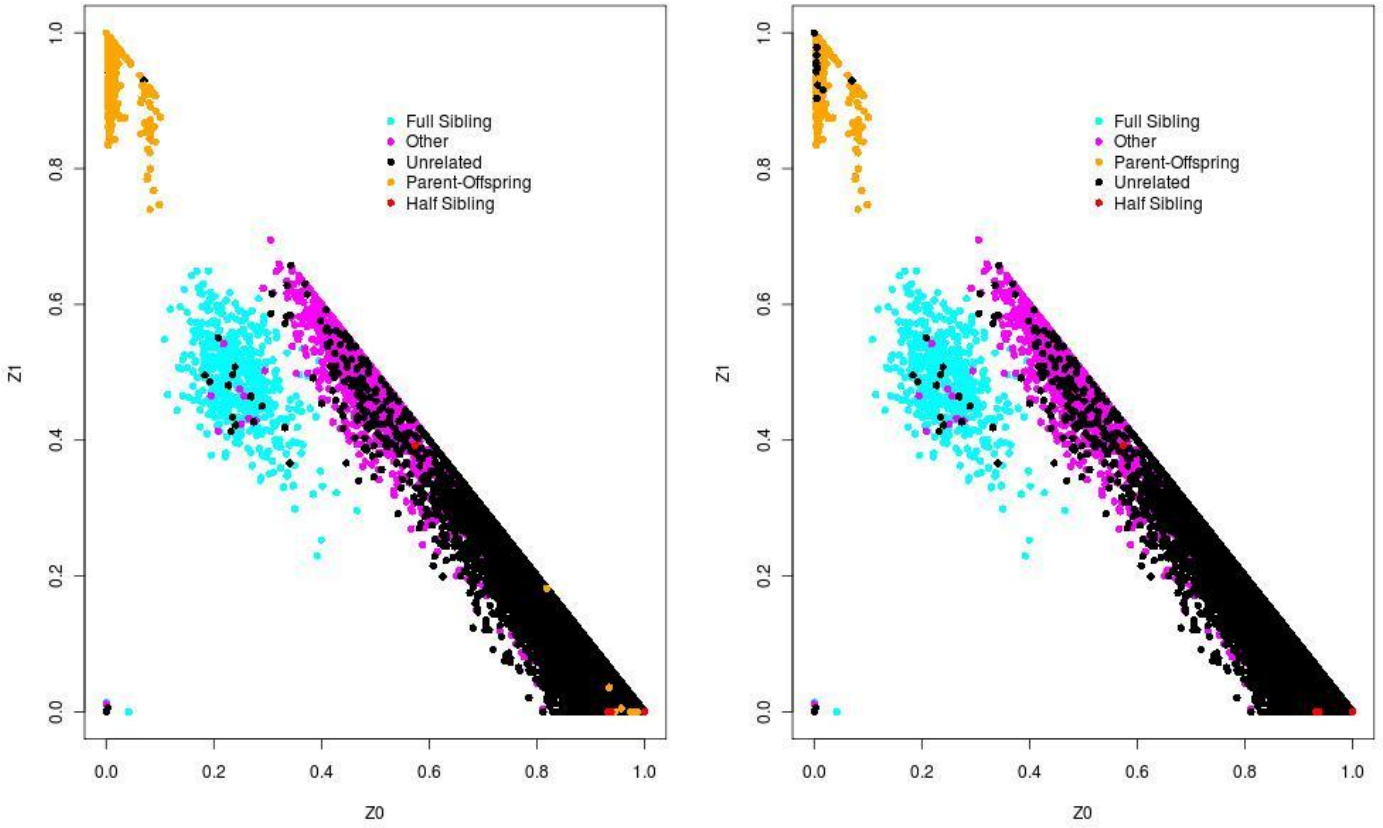
b) checagem de sexo (declarado x computado):

identificação de erros de informação de sexo, trocas de amostra ou ocorrência de problemas cromossômicos.

c) determinação da taxa de heterozigose:

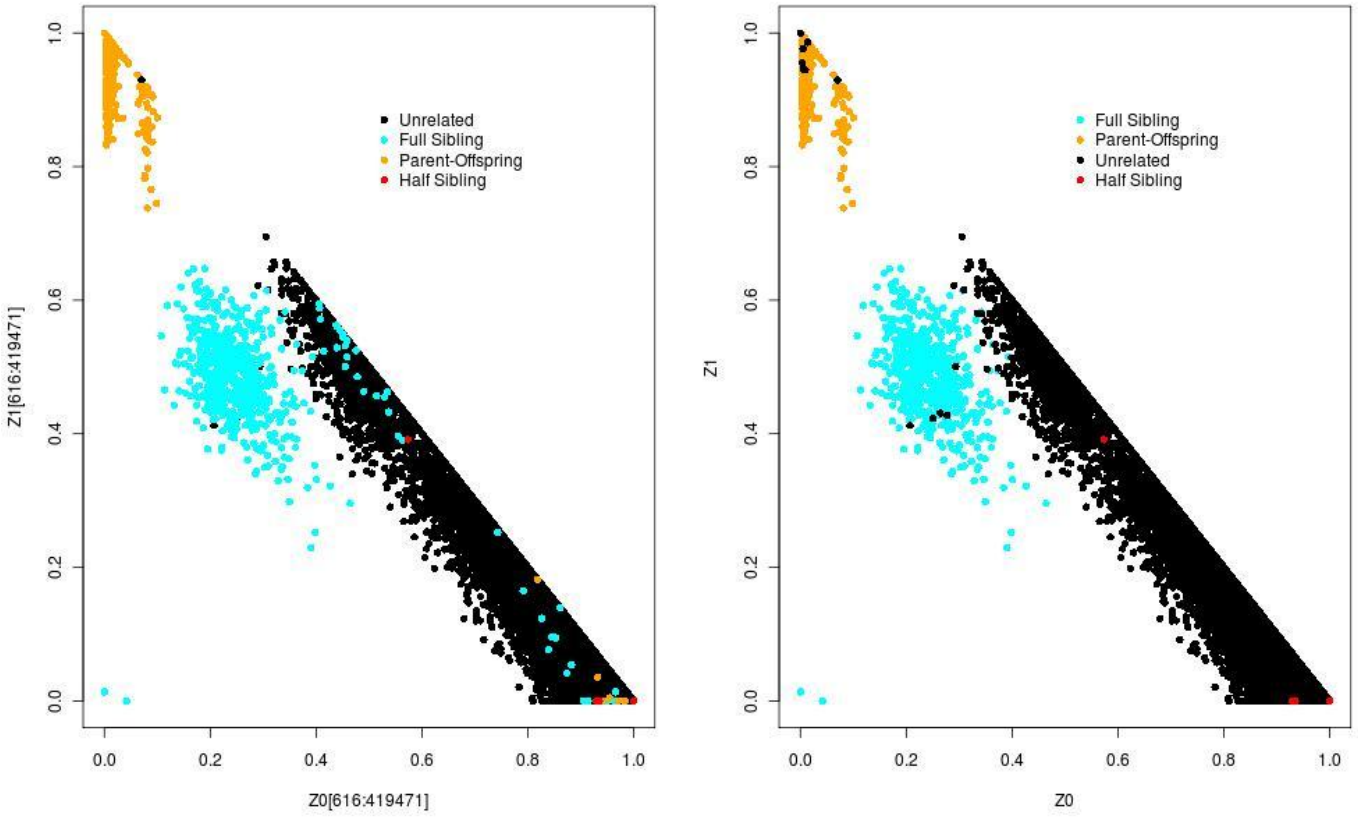
amostras de DNA contaminadas geralmente resultam em taxas de heterozigose mais altas do que o esperado.

BAEPENDI


Full Sibling
Other
Unrelated
Parent-Offspring
Half Sibling
Z1
Z0
0.0
0.2
0.4
0.6
0.8
1.0

Full Sibling
Other
Parent-Offspring
Unrelated
Half Sibling
Z1
Z0
0.0
0.2
0.4
0.6
0.8
1.0

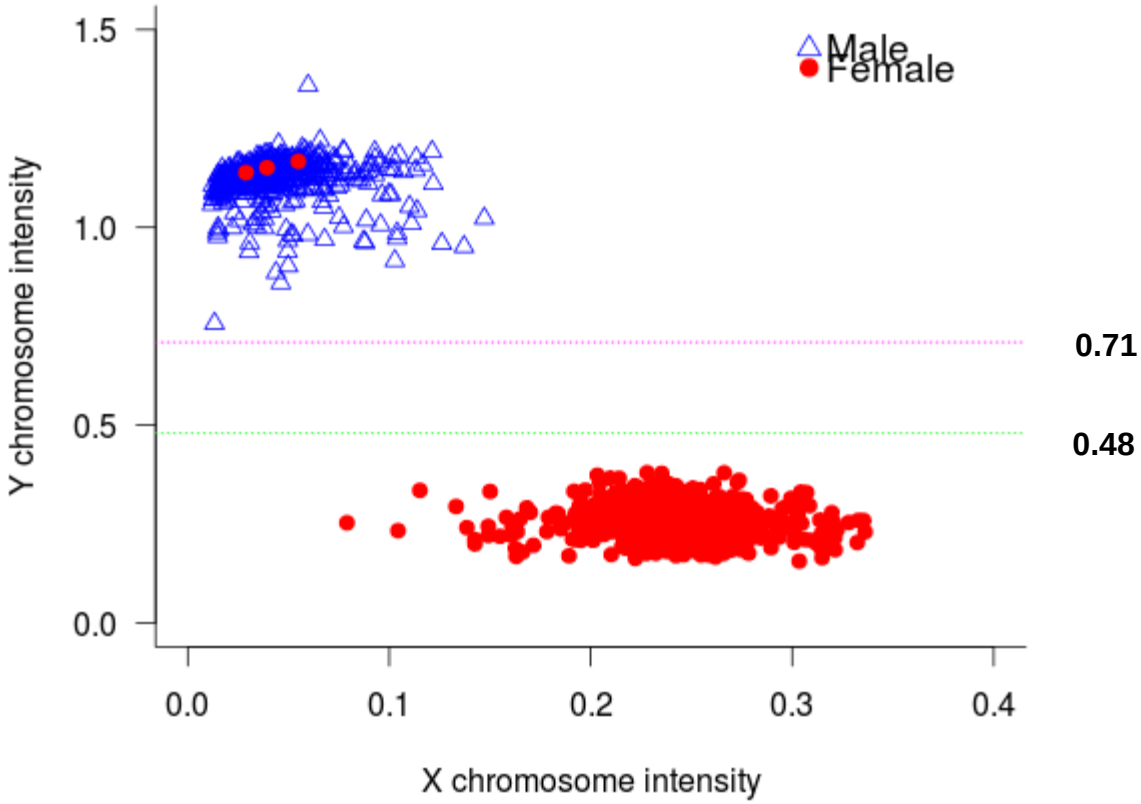
BAEPENDI


Unrelated
Full Sibling
Parent-Offspring
Half Sibling
Z1[616:419471]
Z0[616:419471]
0.0
0.2
0.4
0.6
0.8
1.0

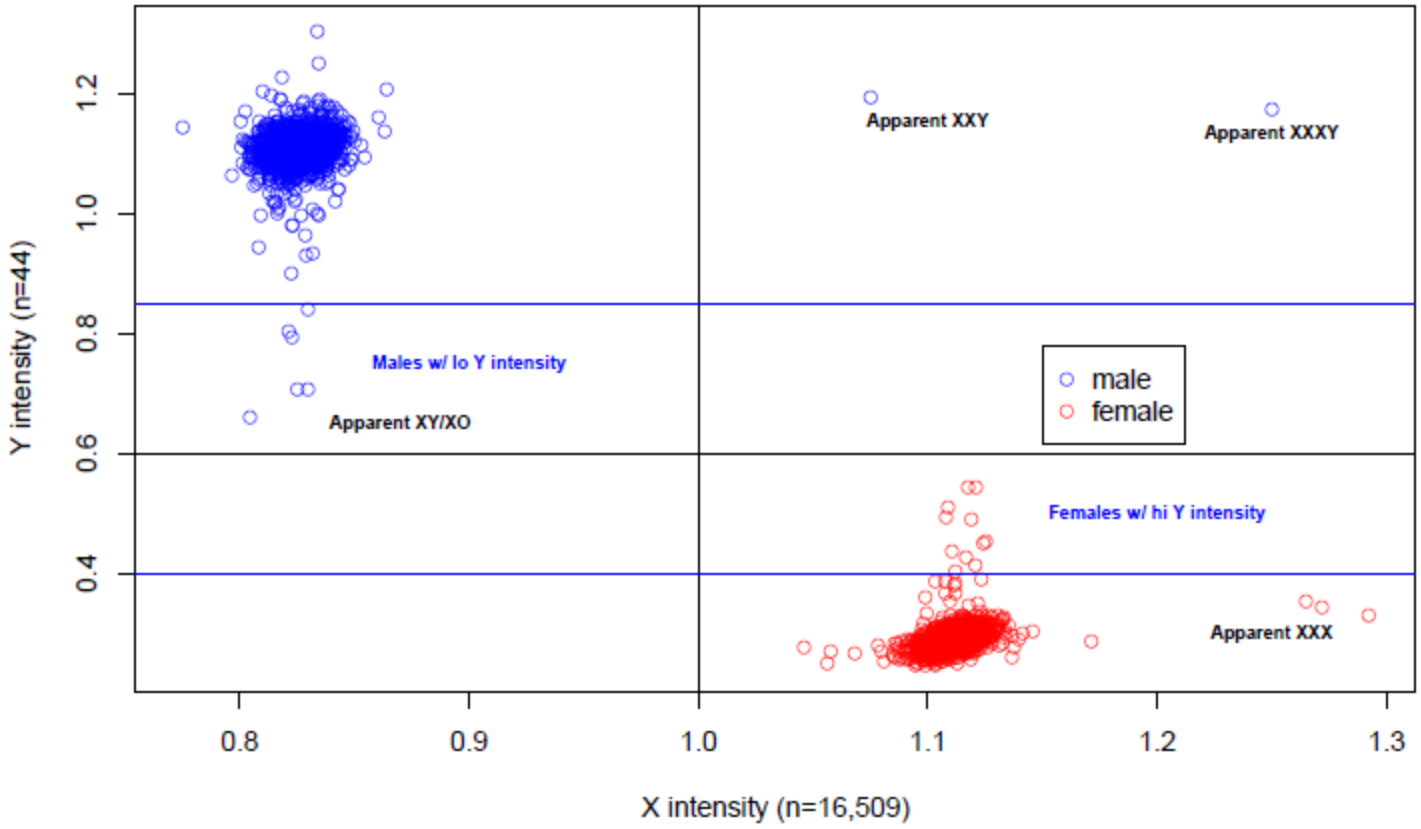
Full Sibling
Parent-Offspring
Unrelated
Half Sibling
Z1
Z0
0.0
0.2
0.4
0.6
0.8
1.0

Data cleaning
Checagem de sexo

BAEPENDI



Data cleaning
Checagem de sexo



DESAFIOS:

1) Armazenamento dos dados;

2) Preparo de arquivos para análise;