Bayesian Mixture Model Clustering for genotyping of DNA Copy Number Variants

Eleni Giannoulatou

Department of Statistics, Oxford University

August 28, 2009

Human Genome and Genomic Variation

About the human genome:

- It is stored in 23 chromosomes.
- It contains 3 billion chemical nucleotide bases (A, C, T, and G).
- Less than 2% of the genome codes for protein.
- The total number of genes is estimated at around 30,000.
- Genomic variation between individuals can determine phenotypic diversity and disease susceptibility.

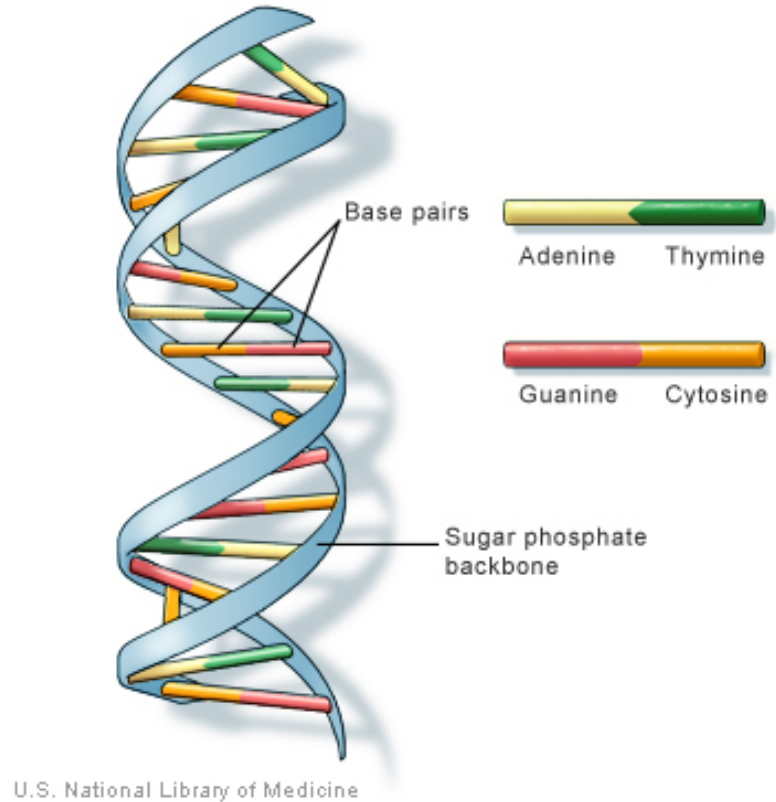
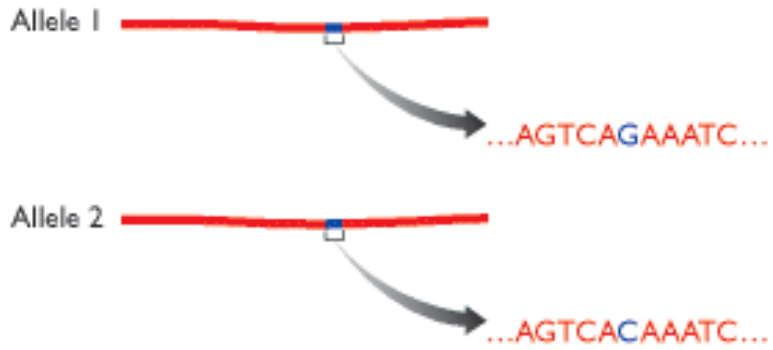


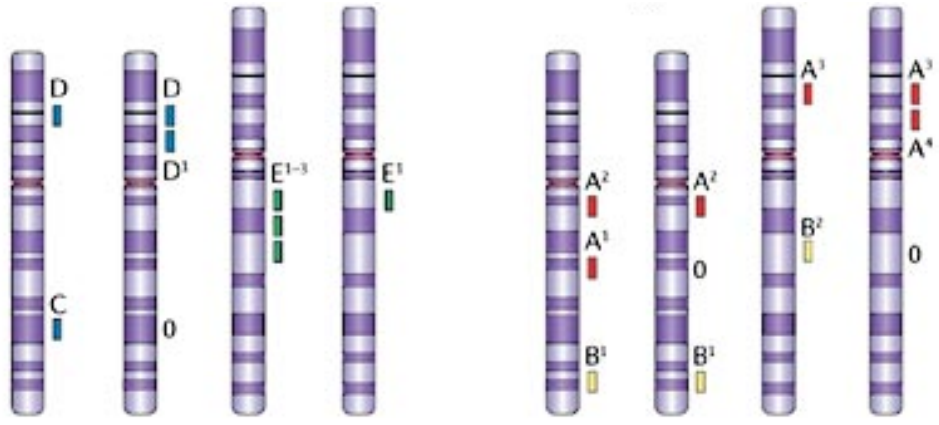
Figure: DNA structure

Variation in the human genome

- Single Nucleotide Polymorphism (SNP): Single base pair position in genomic DNA at which different sequence alternatives (alleles) exist in normal individuals in some population(s).

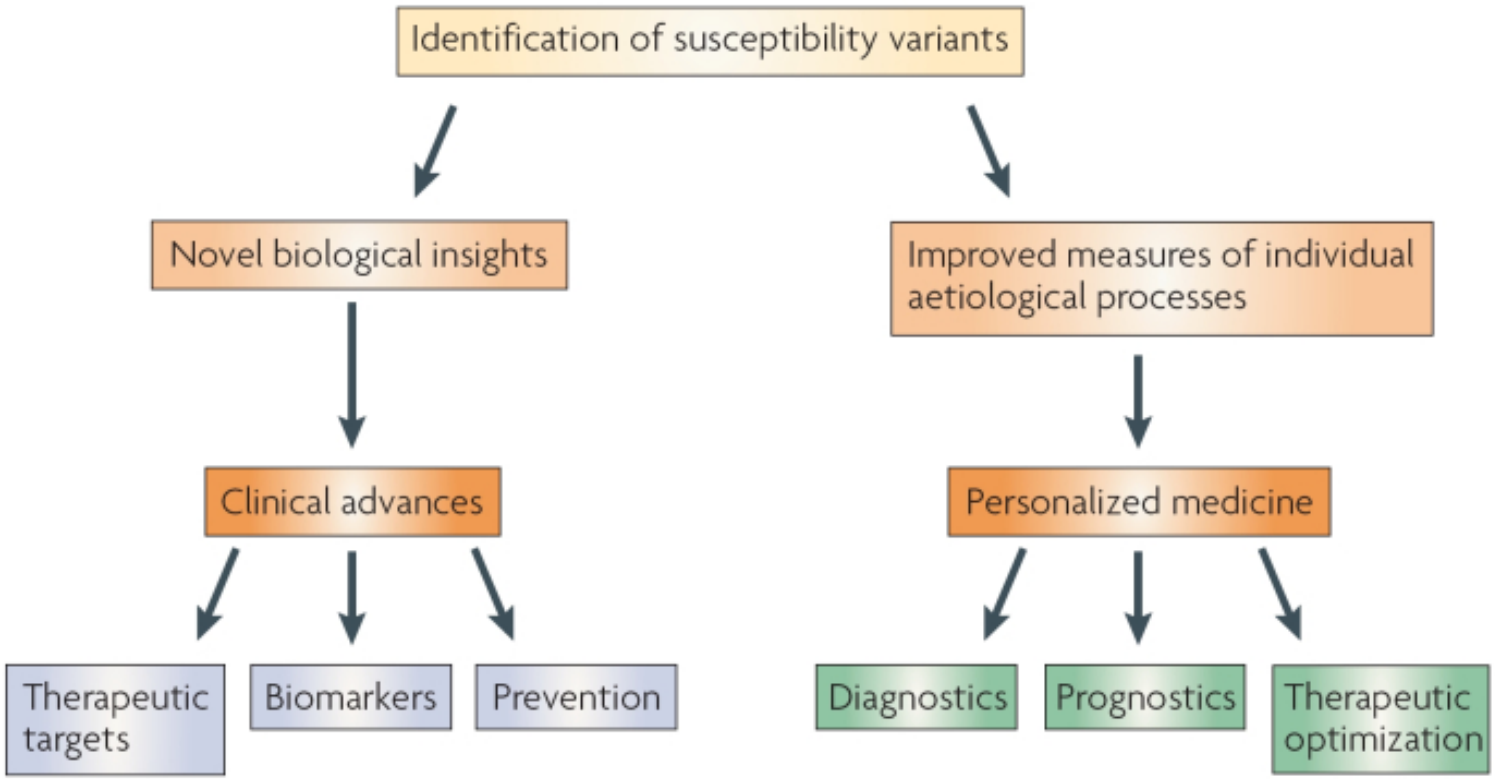


- Copy Number Variation (CNV): DNA fragment that is > 1 kilobases (kb) and is found in variable copy number in comparison with a reference genome



SNPs/CNVs: Why are they important?

- Variation in the DNA sequence can affect disease development as well as response to pathogens, drugs, vaccines.
- Dense maps of SNPs are used in Genome-wide Association Studies: looking for allele-frequency differences between cases (patients with a specific disease) and controls



Nature Reviews Genetics 9:356-369

Genome-Wide Association studies using SNPs

Table 1. Key novel genetics findings using GWA in various complex disease (and related) traits,

Medical field	GWA studies, n	Key novel loci
Ophthalmology		
Age-related macular degeneration	1	*CFH*
Glaucoma	1	*LOXL1*
Endocrinology		
Type 2 diabetes	11	*TCF7L2* and *CDKAL1*
Obesity	5	*INSIG2* and *FTO*
Height	2	*HMGA2* and *GDF5-UQCC*
Cardiology		
Heart disease	4	*CDKN2A / CDKN2B*
Atrial fibrillation	1	*PITX2*
Lipids	3	*MLXIPL*
Oncology		
Prostate cancer	3	8q24
Colorectal cancer	3	8q24 and *SMAD7*
Breast cancer	3	*FGFR2* and *TOX3* (*TNRC9*)
Neuroblastoma	1	*FLJ22536/FLJ44180*
Immunology		
Inflammatory bowel disease	4	*IL23R* and *ATG16L1*
Asthma	1	*ORMDL3*
Type 1 diabetes	2	*CLEC16A* (*KIAA0350*) and 12q13
Ankylosing spondylitis	1	*ERAP1* (*ARTS1*) and *IL23R*
Systemic lupus erythematosus	4	*ITGAM* and *BANK1*
Rheumatoid arthritis	2	*TRAF1/C5*, and *TNFAIP3/OLIG3*
Multiple sclerosis	1	*IL7R* and *IL2RA*
Celiac disease	1	*IL2/IL21*
Host control of HIV-1	1	*HLA-B*5701*
Neurology		
Restless legs syndrome	2	*BTBD9*
Amyotrophic lateral sclerosis	3	*DPP6*
Other		
Gallstones	1	*ABCG8*
Fetal hemoglobin	2	*BCL11A*

Clinical Chemistry 54:7 (2008)

Genome-Wide Association studies using CNVs

Locus	CNV frequency	Clinical phenotype	CNV type	Risk estimate (odds ratio)	Comments
CCL3L1 [9,11]	10–20%	HIV/AIDS susceptibility [9] Rheumatoid arthritis [11]	Deletion Gain: >2 copies	0.67–0.90 1.34	CCL3L1 inhibits HIV cellular entry [50]. Higher CCL3L1 number increases CCL3L1 expression [49]
FCGR3B [10]	Deletion: ~25% Gain: ~15%	Systemic autoimmune disease	Deletion	1.58–2.56[a]	CNV associated with glomerulonephritis in rats and humans [51]
C4 [12]	~40%	Systemic lupus erythematosus	Deletion	Absence: 5.27 Carrier: 1.61 Gains: 0.57	>75% of C4 or C1 deletion carriers have SLE-like disease [12]. Strongest SLE genetic risk factor thus far in blacks [52]
DEFB4 [33,34]	2–12 copies (median 4)	Colonic Crohn disease [33] Psoriasis [34]	Loss: <4 copies Gain: >5 copies	3.06 1.69	↓ number associated with ↓ mucosal gene expression. [33]
GSTM1 [13–16]	Up to 50%	Asthma, lung function, allergic response	Deletion	1.59–1.89	Potent antioxidant. Deletion related to many adverse asthma-related outcomes (see text).

Genomics 93 (2009) 22–26

SNP/CNV genotyping

- Identification of the allelic states of SNPs/CNVs in a large number of individuals.
- The set of alleles that a person has is called a genotype. For this SNP a person could have the genotype AA, AG, or GG.

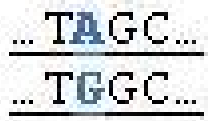


Figure: A part of two chromosomes showing a SNP.

- There are estimated to be 10 million SNPs in the genome - more than 3 million have been charted (International HapMap Project).
- CNV discovery is still in progress!

SNP genotyping – Illumina BeadArray platform

- BeadArray data consist of two channel intensity data that correspond to the two alleles.

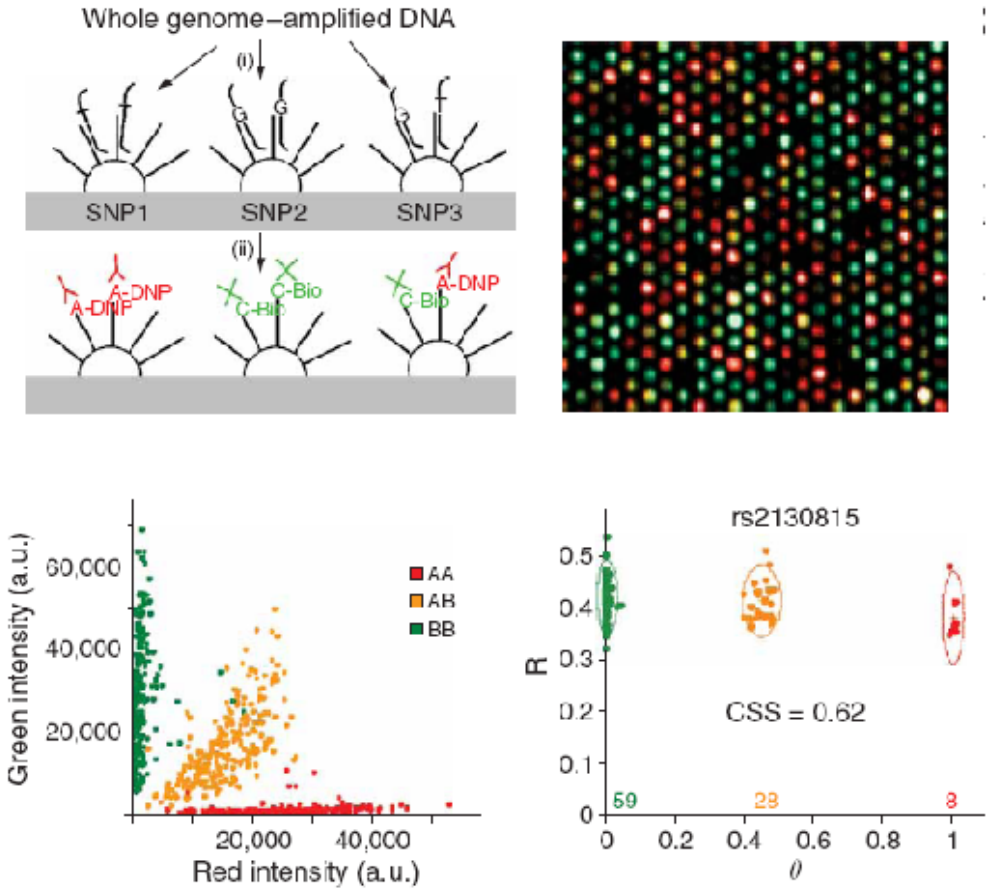


Figure: SNP genotyping using BeadArray Infinium II assay.

Nature Methods 3, 31 - 33 (2006)

Signal Intensity plots

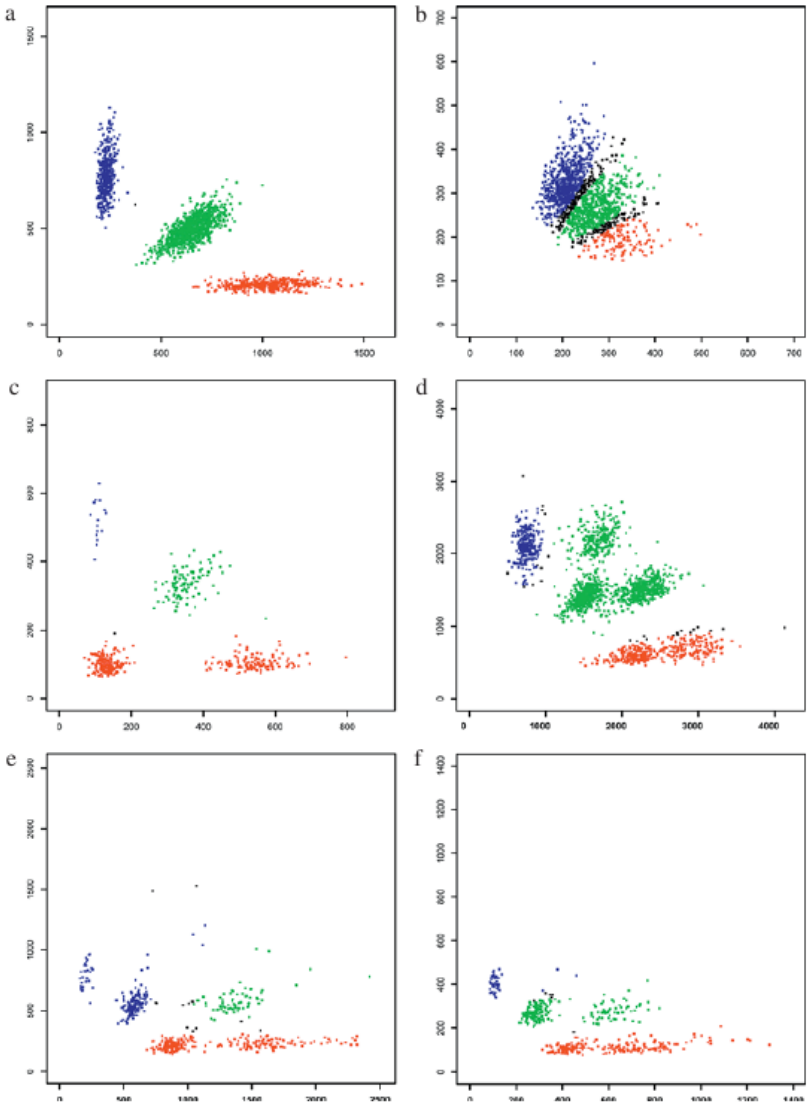


Figure: High and low genotyping quality

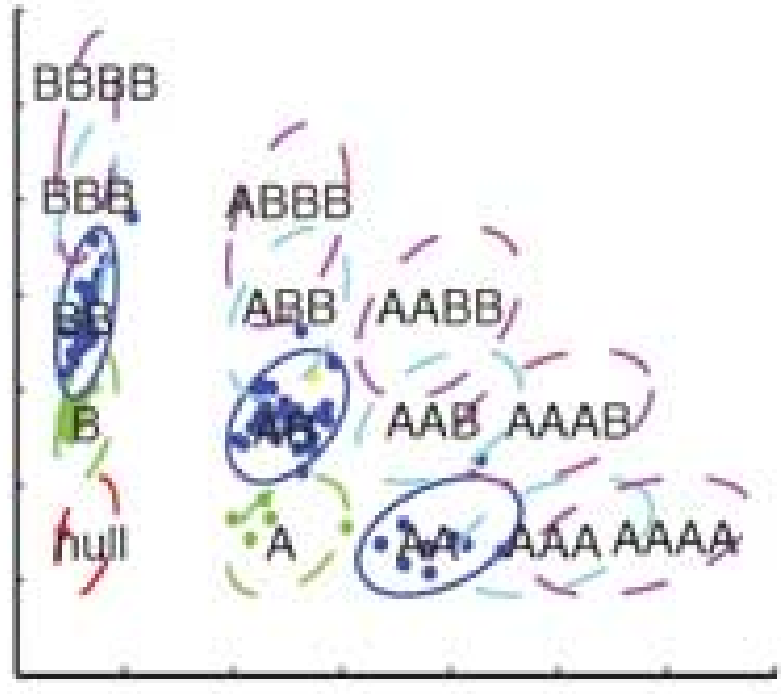
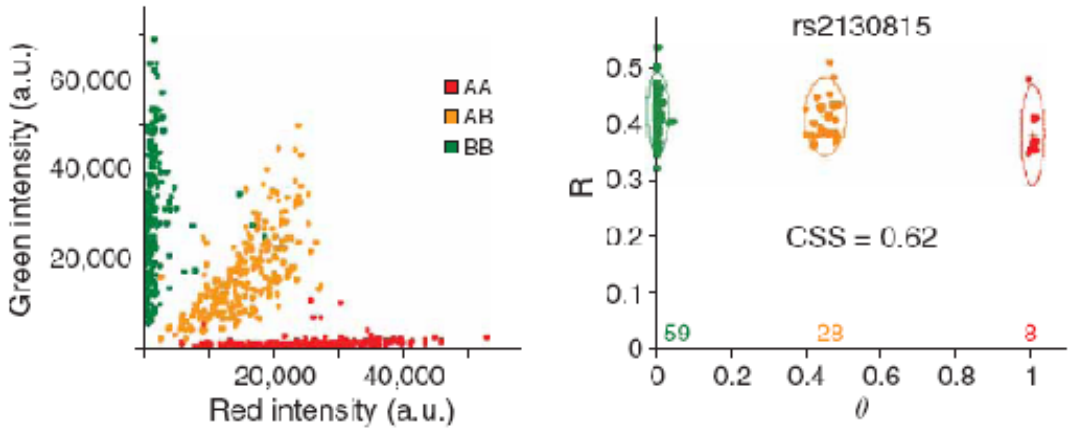


Figure: Combined SNP/CNV information

SNP genotyping algorithms

- Each SNP is interrogated in turn, clustering the allele-specific probe intensities in three classes.
- Reason: Probe intensities vary on a SNP-by-SNP basis.
- Limitations:
  1. Big reference population is needed for SNPs with low MAF. (10,000 samples for a SNP with MAF=1%)
  2. Model parameters must be recalibrated each time the SNP content of an array is modified or a new genotyping array is produced.

GenoSNP: Main principle



- High quality SNP genotyping *within* a sample is enabled without the need for a reference population.
- Inter-class variation is maximized by accounting for dye-specific and bead-specific effects:
  - Clustering is done on the intensities $log_2(x+1), log_2(y+1)$ for each beadpool separately.
- Bayesian Mixture Model using Variational Bayes

Giannoulatou E, Yau C et al, Bioinformatics. 2008 24(19):2209-14

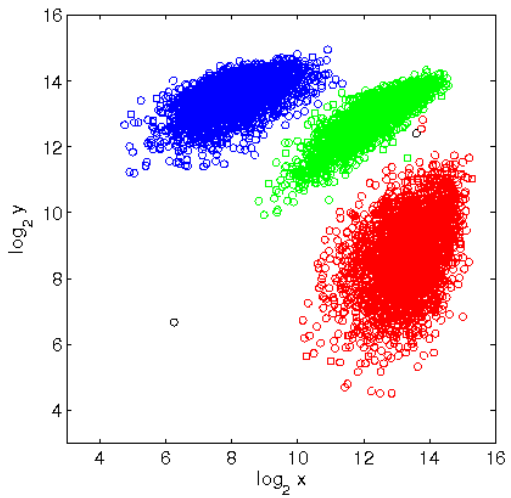
GenoSNP: Statistical Model

$\mathbf{x}_n = \{\log_2(x_n + 1), \log_2(y_n + 1)\}$: the vector of log intensities for the nth SNP. The distribution of the intensities is modelled using a 4-component mixture of Student-t distributions

$$p(\mathbf{x}_n) = \sum_{m=1}^{4} \pi_m \mathcal{S}_m(\mathbf{x}_n; \boldsymbol{\mu}_m, \boldsymbol{\Lambda}_m, \nu)$$

$$\sum_{m=1}^{4} \pi_m = 1$$

Each component corresponds to either one of the three genotype classes AA, AB and BB or a null class to capture outliers.



GenoSNP: Statistical Model

- The SMM can be viewed as a latent variable model as the component label for each data point is unobserved $z_{nm} \in 0, 1$.
- The observed data is still incomplete – the Student-t distribution can be rewritten:

$$\mathcal{S}(\mathbf{x}; \boldsymbol{\mu}, \boldsymbol{\Lambda}, \nu) = \int_0^\infty \mathcal{N}(\mathbf{x}|\boldsymbol{\mu}, u\boldsymbol{\Lambda})\mathcal{G}\left(u\middle|\frac{\nu}{2}, \frac{\nu}{2}\right) du.$$

- The scaling factor is an implicit latent variable on which Gamma prior is imposed.

GenoSNP: Statistical Model

For each data point $\mathbf{x}$ and for each component m, the scale variable u_{nm} given z_{nm} is unobserved. The latent variable model is:

$$p(\mathbf{z}_n|\theta) = \prod_{m=1}^{4} \pi_m^{z_{nm}}$$

$$p(\mathbf{u}_n|\mathbf{z}_n, \theta) = \prod_{m=1}^{4} \mathcal{G}\left(u_{nm}\middle|\frac{\nu_m}{2}, \frac{\nu_m}{2}\right)^{z_{nm}}$$

$$p(\mathbf{x}_n|\mathbf{u}_n, \mathbf{z}_n, \theta) = \prod_{m=1}^{4} \mathcal{N}(\mathbf{x}_n|\boldsymbol{\mu}_m, u_{nm}\boldsymbol{\Lambda}_m)^{z_{nm}}$$

GenoSNP: Statistical Model

The prior for the mixture weight is given by a Dirichlet distribution

$$p(\boldsymbol{\pi}|\boldsymbol{\kappa}) \propto \prod_{m=1}^{4} \pi_m^{(\kappa_m - 1)},$$

and a Normal-Wishart prior used to define the location and scale parameters for each genotype mixture component

$$p(\boldsymbol{\mu}_m, \boldsymbol{\Lambda}_m) = \mathcal{N}(\boldsymbol{\mu}_m|\mathbf{m}_0, \eta_0\boldsymbol{\Lambda}_m)\mathcal{W}(\boldsymbol{\Lambda}_m|\gamma, \mathbf{S}_m)$$

The location and scale parameters of the null class are fixed and set to values to make the distribution relatively flat over the feature space.

GenoSNP: Posterior Inference

Variational Bayes EM algorithm:
Minimisation of the Kullback-Leibler divergence between the true posterior distribution $p(\theta, \mathbf{z}, \mathbf{u}|\mathbf{x})$ and the variational approximation $q(\theta, \mathbf{z}, \mathbf{u})$

$$KL(q, p) \equiv \int q(\theta, \mathbf{z}, \mathbf{u}) \log \frac{p(\theta, \mathbf{z}, \mathbf{u}, \mathbf{x})}{q(\theta, \mathbf{z}, \mathbf{u})} d\theta.$$

Assumption: $q(\theta, \mathbf{z}) = q_\theta(\theta) q_{\mathbf{z}}(\mathbf{z}, \mathbf{u})$. The VB-EM steps are:

$$q_{\mathbf{zu}}^{(t+1)}(\mathbf{z}, \mathbf{u}) \propto \exp\left(E_\theta[\log p(\mathbf{x}, \mathbf{u}, \mathbf{z}|\theta)]\right) \quad (1)$$

$$q_\theta^{(t+1)}(\theta) \propto p(\theta) \exp(E_{\mathbf{z},\mathbf{u}}[\log p(\mathbf{z}, \mathbf{u}, \mathbf{x}|\theta)]) \quad (2)$$

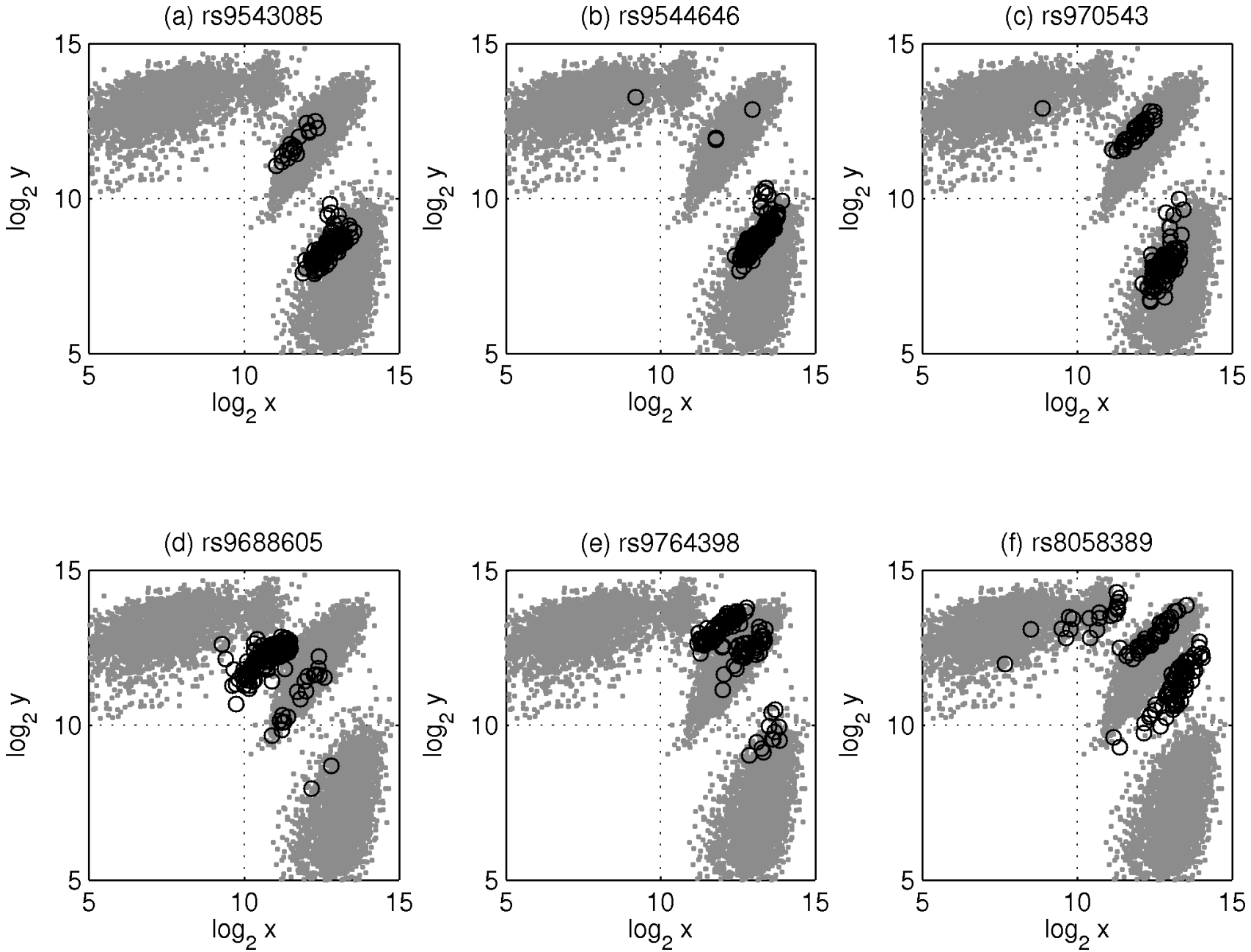
Expressions for the exact parameter updates are given in (Archambeau and Verleysen, 2007)

GenoSNP: Results

Table: Comparison of call rates and accuracy on 120 Hapmap samples genotyped on the HumanHap300Duo BeadChip

Method	Call Rate (%)	False Calls	No Calls	Call Accuracy (%)
GenCall	99.799	38,911	73,295	99.694
Illuminus	99.819	89,025	66,199	99.576
GenoSNP	99.660	88,249	124,613	99.419
GenoSNP-VB	100.000	94,380	143	99.742

GenoSNP: Results



GenoSNP: Results

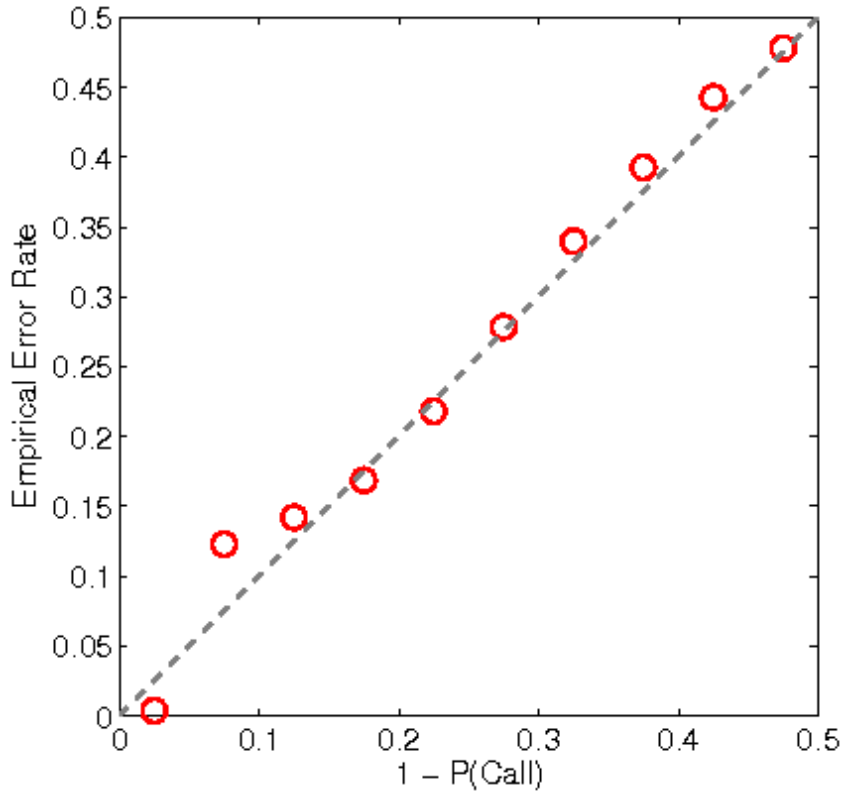


Figure: GenoSNP genotype probabilities are well calibrated with empirical error rates.

CNV calling: Motivation

- Identification of the CNV allelic state of every individual in a cohort.
- The copy number is not assumed to be diploid as in SNP genotyping algorithms but it is inferred. Hence the number of clusters in the data needs to be estimated.
- A mixture model with an excess number of components has the greatest chance of capturing all the true clusters in the data.
- A backward selection procedure can be applied that starting from a excess number of clusters it combines every two clusters and selects for merging the pair with the highest marginal likelihood.

Backward Deletion Procedure and Model Selection

- (a) Initialise the cluster centres uniformly from $\min(x)$ to $\max(x)$.
- (b) For $M = M_{max}$ to 2:
  - 1. Use VB-EM to optimise the hyperparameters.
  - 2. For every pair of clusters (i, j), propose to combine the jth cluster with the ith cluster.:
    - i. Compute a weighted average of the weights, centres and variances for the components i and j.
    - ii. Calculate the approximate log marginal likelihood using the new parameters for the new combined cluster having the i and jth cluster removed from the model.
  - 3. Select the pair (i, j) that has the largest log marginal likelihood and accept this merge.
- (c) Select M that gives the highest log Bayes Factor $BF = \log \frac{p(x|M)}{p(x|M=2)}$

Backward Deletion Procedure and Model Selection

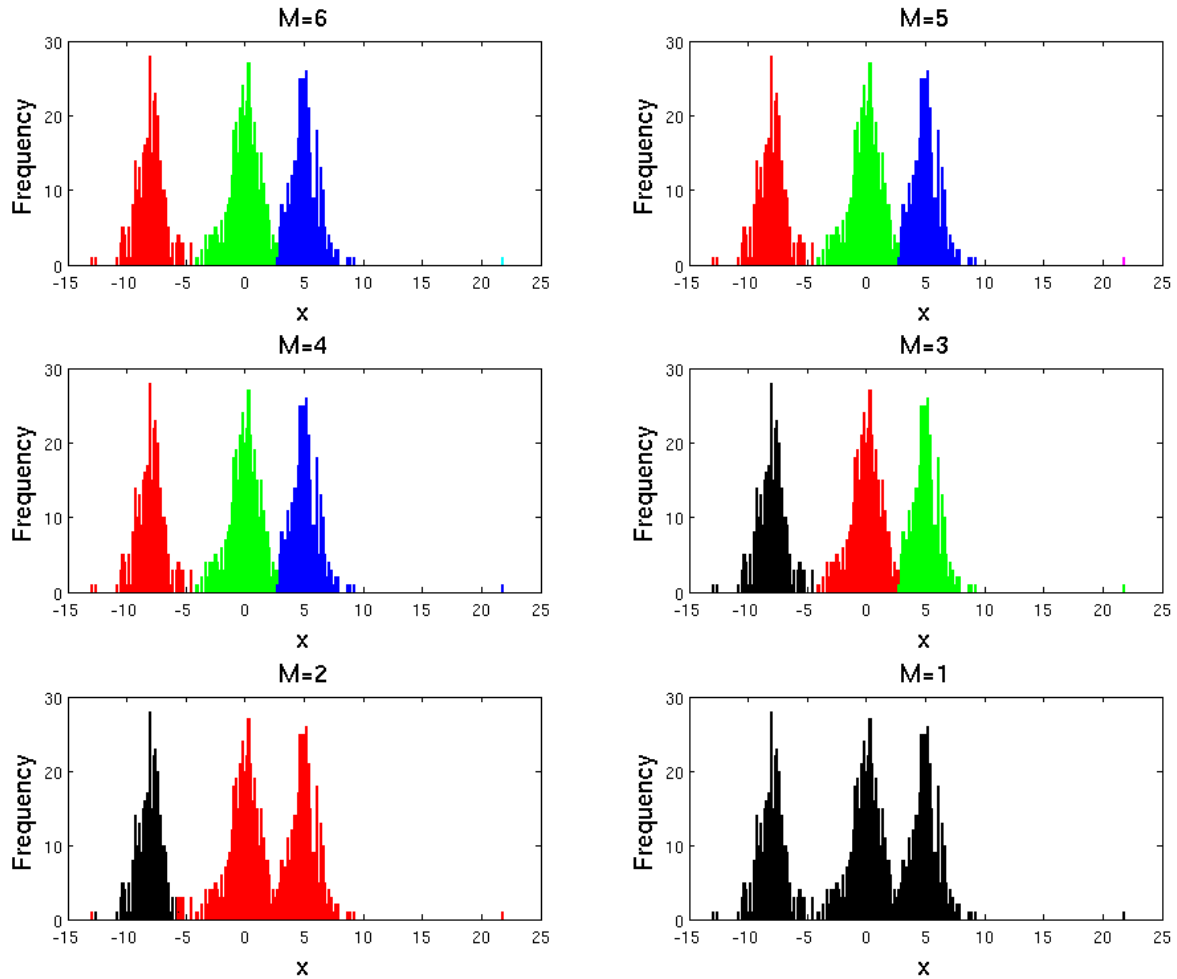


Figure: Histograms of the data showing cluster assignment for number of cluster $M = 1, ..., 6$ (excluding the outlier class).

CNV Clustering results: Model Selection

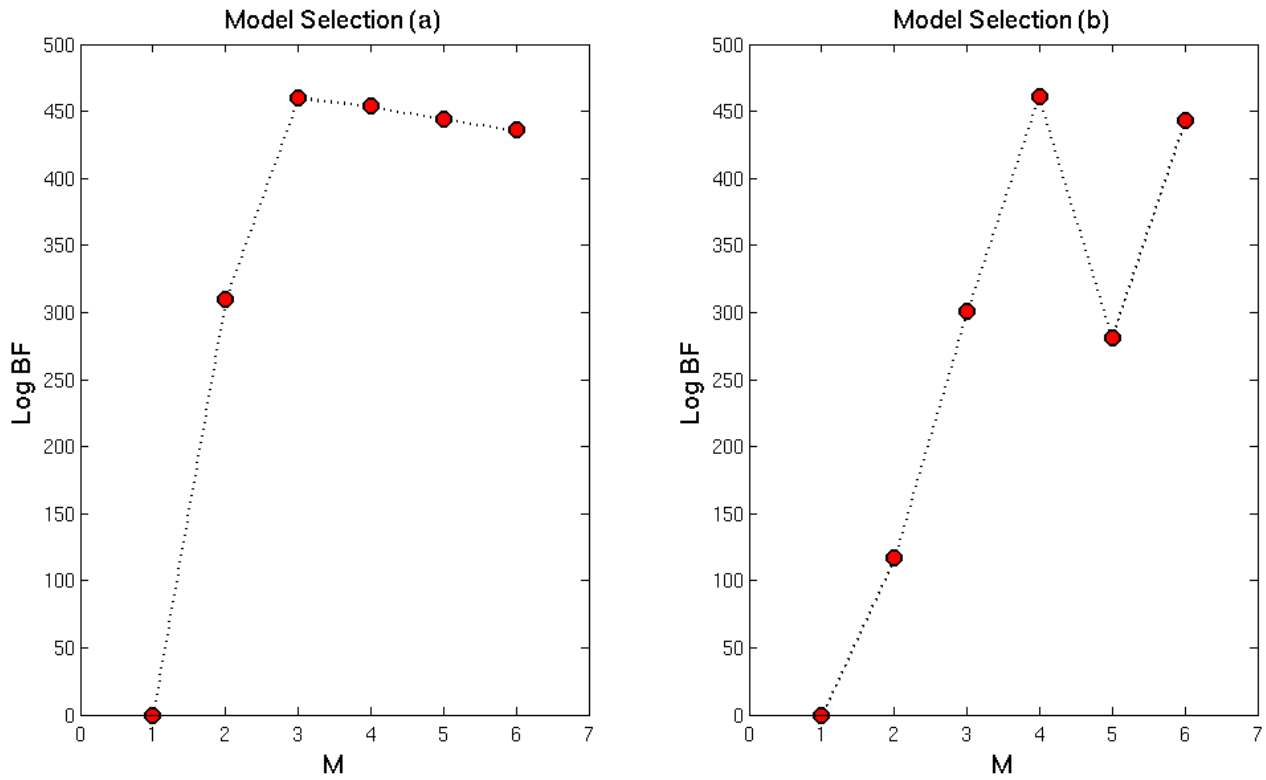


Figure: Log Bayes Factor for each Model (M=1,...6) (a) when we apply Backward Selection of clusters and (b) with no Backward Selection.

Summary

- GenoSNP: a Variational Bayes SNP genotyping algorithm that is able to call genotypes within sample with comparable accuracy to other population-based genotyping algorithms.
- CNV calling method in 1D for targeted studies using robust Bayesian Mixture Model Clustering and Backward Selection of clusters.

Acknowledgements

- Christopher Yau and Chris Holmes (Statistics)
- Chris Holmes/Jotun Hein Groups (Statistics)
- Jiannis Ragoussis and Stefano Collela (Wellcome Trust Centre for Human Genetics)