# A Semiparametric Bayesian Model for Comparing DNA Copy Numbers

Luis E. Nieto-Barajas

Department of Statistics, ITAM, Mexico

ISBA 2014 World Meeting

*Cancún, Mexico, July 14-18, 2014*

(joint with Y Ji & V.Baladandayuthapani)

# Contents

- Introduction
- Model
- Inference
- Results

# Objectives

- There has been increasing interest in constructing the genomic architecture of diseases, e.g. breast cancer

# Objectives

- There has been increasing interest in constructing the genomic architecture of diseases, e.g. breast cancer
- Genomic architecture based on DNA copy number alterations

# Objectives

- There has been increasing interest in constructing the genomic architecture of diseases, e.g. breast cancer
- Genomic architecture based on DNA copy number alterations
- CNA = variations (from two) in the copy number of DNA

# Objectives

- There has been increasing interest in constructing the genomic architecture of diseases, e.g. breast cancer
- Genomic architecture based on DNA copy number alterations
- CNA = variations (from two) in the copy number of DNA
- Aim: characterize different subtypes of breast cancer by examining the whole-genome copy number profiles based on multiple samples

# Objectives

- There has been increasing interest in constructing the genomic architecture of diseases, e.g. breast cancer
- Genomic architecture based on DNA copy number alterations
- CNA = variations (from two) in the copy number of DNA
- Aim: characterize different subtypes of breast cancer by examining the whole-genome copy number profiles based on multiple samples
  - Identifying genome aberrations for samples of the same disease subtype

# Objectives

- There has been increasing interest in constructing the genomic architecture of diseases, e.g. breast cancer
- Genomic architecture based on DNA copy number alterations
- CNA = variations (from two) in the copy number of DNA
- Aim: characterize different subtypes of breast cancer by examining the whole-genome copy number profiles based on multiple samples
    - Identifying genome aberrations for samples of the same disease subtype
    - Detecting differences across disease subtypes

# Example

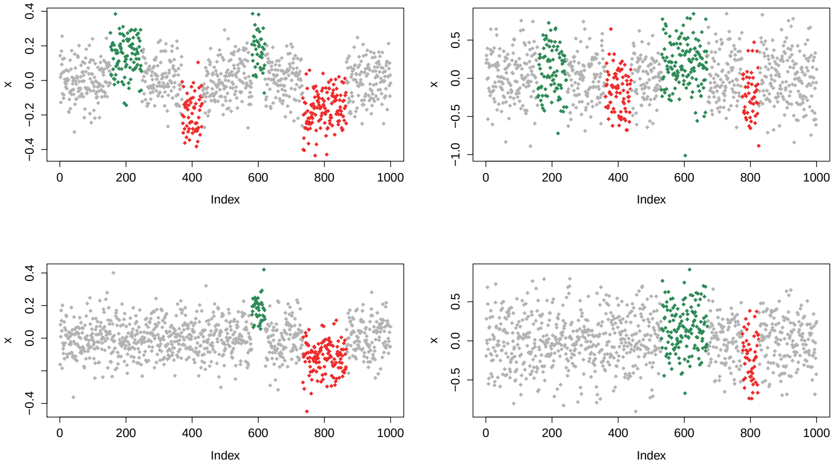


Figure : Simulated genome profile.

## Literature review

Some current approaches to CNA detection are:

- Olshen et al. (2004): Circular binary segmentation (most widely used method)

# Literature review

Some current approaches to CNA detection are:

- Olshen et al. (2004): Circular binary segmentation (most widely used method)
- Guha et al. (2008): Bayesian hidden Markov model

# Literature review

Some current approaches to CNA detection are:

- Olshen et al. (2004): Circular binary segmentation (most widely used method)
- Guha et al. (2008): Bayesian hidden Markov model
- Shah et al. (2007): Hierarchical hidden Markov models for recurrent CNA

# Literature review

Some current approaches to CNA detection are:

- Olshen et al. (2004): Circular binary segmentation (most widely used method)
- Guha et al. (2008): Bayesian hidden Markov model
- Shah et al. (2007): Hierarchical hidden Markov models for recurrent CNA
- Baladandayuthapani et al. (2010): Hierarchical Bayesian random segmentation approach for multiple samples

# Literature review

Some current approaches to CNA detection are:

- Olshen et al. (2004): Circular binary segmentation (most widely used method)
- Guha et al. (2008): Bayesian hidden Markov model
- Shah et al. (2007): Hierarchical hidden Markov models for recurrent CNA
- Baladandayuthapani et al. (2010): Hierarchical Bayesian random segmentation approach for multiple samples
- Yau et al. (2011): mixture model that combines a hidden Markov model for the locations (states), with a Dirichlet process prior for the scales

# Definitions

- Let $\mathcal{A} = \{t_1, t_2, \ldots, t_n\}$ be the index of probes.
  For each array $j$, we assume that there are $n_j$ probes, which are a subset of $\mathcal{A}$.

# Definitions

- Let $\mathcal{A} = \{t_1, t_2, \ldots, t_n\}$ be the index of probes.
  For each array $j$, we assume that there are $n_j$ probes, which are a subset of $\mathcal{A}$.
- For each sample $j = 1, \ldots, J$ we have a partition $\{\Delta_l^j\}_{l=1}^{L_j}$ of $\mathcal{A}$ with $\Delta_l^j = [c_l^j, c_{l+1}^j)$.

# Definitions

- Let $\mathcal{A} = \{t_1, t_2, \ldots, t_n\}$ be the index of probes. For each array $j$, we assume that there are $n_j$ probes, which are a subset of $\mathcal{A}$.
- For each sample $j = 1, \ldots, J$ we have a partition $\{\Delta_l^j\}_{l=1}^{L_j}$ of $\mathcal{A}$ with $\Delta_l^j = [c_l^j, c_{l+1}^j)$.
- We define a common partition $\{\Omega_k\}_{k=1}^{K}$ for all arrays as the union of all partition segments over $j = 1, \ldots, J$. That is, $\Omega_k = [c_k, c_{k+1})$ with $\{t_1 = c_1 < c_2 \cdots < c_{K+1} = t_n\} = \cup_j \{t_1 = c_1^j < c_2^j \cdots < c_{L_j+1}^j = t_n\}$.

# Definitions

- Let $\mathcal{A} = \{t_1, t_2, \ldots, t_n\}$ be the index of probes. For each array $j$, we assume that there are $n_j$ probes, which are a subset of $\mathcal{A}$.
- For each sample $j = 1, \ldots, J$ we have a partition $\{\Delta_l^j\}_{l=1}^{L_j}$ of $\mathcal{A}$ with $\Delta_l^j = [c_l^j, c_{l+1}^j)$.
- We define a common partition $\{\Omega_k\}_{k=1}^{K}$ for all arrays as the union of all partition segments over $j = 1, \ldots, J$. That is, $\Omega_k = [c_k, c_{k+1})$ with $\{t_1 = c_1 < c_2 \cdots < c_{K+1} = t_n\} = \cup_j \{t_1 = c_1^j < c_2^j \cdots < c_{L_j+1}^j = t_n\}$.
- Let $g_j$ indicate the disease subtype for sample $j$. Say $g_j \in \{1, 2\}$.

# Semiparametric model

- Let $Y_{ij}$ be the $\log_2$ ratio of probe $t_i$ at sample $j$.

# Semiparametric model

- Let $Y_{ij}$ be the $\log_2$ ratio of probe $t_i$ at sample $j$.
- Sampling model: For $i = 1, \ldots, n_j$ and $j = 1, \ldots, J$

$$Y_{ij} = \sum_{k=1}^{K} \mu_{k,g_j} I(i \in \Omega_k) + \sum_{l=1}^{L_j} m_{lj} I(i \in \Delta_{lj}) + \epsilon_{ij}, \qquad (1)$$

with $\epsilon_{ij} \overset{\text{iid}}{\sim} \mathrm{N}(0, \sigma_\epsilon^2)$

# Semiparametric model

- Let $Y_{ij}$ be the $\log_2$ ratio of probe $t_i$ at sample $j$.
- Sampling model: For $i = 1, \ldots, n_j$ and $j = 1, \ldots, J$

$$Y_{ij} = \sum_{k=1}^{K} \mu_{k,g_j} I(i \in \Omega_k) + \sum_{l=1}^{L_j} m_{lj} I(i \in \Delta_{lj}) + \epsilon_{ij}, \qquad (1)$$

with $\epsilon_{ij} \overset{\text{iid}}{\sim} \mathrm{N}(0, \sigma_\epsilon^2)$

- That is, $Y_{ij}$ arises from the sum of a population mean $\mu_{k,g_j}$, a sample-specific mean $m_{lj}$, plus a measurement error $\epsilon_{ij}$.

# Semiparametric model

Priors:

- Denote by $\boldsymbol{\mu}_k = (\mu_{k1}, \mu_{k2})$ the vector of population copy number levels for subtypes 1 and 2

$$\boldsymbol{\mu}_k \mid G \overset{\text{ind}}{\sim} G, \quad \text{for } k = 1, \ldots, K$$

$$G = (1-\pi)G_0 + \pi G_1$$

$$G_r | a_r \overset{\text{ind}}{\sim} \mathcal{DP}(a_r, F_r), \; r = 0, 1,$$

# Semiparametric model

Priors:

- Denote by $\boldsymbol{\mu}_k = (\mu_{k1}, \mu_{k2})$ the vector of population copy number levels for subtypes 1 and 2

$$\boldsymbol{\mu}_k \mid G \overset{\text{ind}}{\sim} G, \quad \text{for } k = 1, \dots, K$$

$$G = (1-\pi) G_0 + \pi G_1$$

$$G_r | a_r \overset{\text{ind}}{\sim} \mathcal{DP}(a_r, F_r), \; r = 0, 1,$$

- We define a spike and slab prior in two dimensions $F_0(\boldsymbol{\mu}_k) = N(\mu_{k1} \mid 0, \lambda_0^2) I(\mu_{k1} = \mu_{k2})$ and $F_1(\boldsymbol{\mu}_k) = N_2(\boldsymbol{\mu}_k \mid \mathbf{0}, \boldsymbol{\Lambda}_1)$

# Semiparametric model

Priors:

- Denote by $\boldsymbol{\mu}_k = (\mu_{k1}, \mu_{k2})$ the vector of population copy number levels for subtypes 1 and 2

$$\boldsymbol{\mu}_k \mid G \overset{\text{ind}}{\sim} G, \quad \text{for } k = 1, \ldots, K$$

$$G = (1-\pi)G_0 + \pi G_1$$

$$G_r | a_r \overset{\text{ind}}{\sim} \mathcal{DP}(a_r, F_r), \; r = 0, 1,$$

- We define a spike and slab prior in two dimensions $F_0(\boldsymbol{\mu}_k) = N(\mu_{k1} \mid 0, \lambda_0^2) I(\mu_{k1} = \mu_{k2})$ and $F_1(\boldsymbol{\mu}_k) = N_2(\boldsymbol{\mu}_k \mid \mathbf{0}, \boldsymbol{\Lambda}_1)$
- Introducing a latent indicator $z_k = I(\mu_{k1} \neq \mu_{k2})$

$$\boldsymbol{\mu}_k \mid z_k, G_0, G_1 \overset{\text{ind}}{\sim} G_{z_k}, \;\; z_k \overset{\text{ind}}{\sim} \text{Ber}(\pi), \;\; G_r \overset{\text{ind}}{\sim} \mathcal{DP}(a_r, F_r) \quad (2)$$

# Semiparametric model

Priors:

- For the random effects

$$m_{kj} \overset{\text{ind}}{\sim} \mathrm{N}(0, \tau_j^2), \quad \text{with} \quad \tau_j^2 \overset{\text{iid}}{\sim} \mathrm{IGa}(\alpha_\tau, \beta_\tau).$$

# Semiparametric model

Priors:

- For the random effects

$$m_{kj} \overset{\text{ind}}{\sim} \mathrm{N}(0, \tau_j^2), \quad \text{with} \quad \tau_j^2 \overset{\text{iid}}{\sim} \mathrm{IGa}(\alpha_\tau, \beta_\tau).$$

- For the sample variance:

$$\sigma_\epsilon^2 \sim \mathrm{IGa}(\alpha_\sigma, \beta_\sigma).$$

# Semiparametric model

Priors:

- For the random effects

$$m_{kj} \overset{\text{ind}}{\sim} \mathrm{N}(0, \tau_j^2), \quad \text{with} \quad \tau_j^2 \overset{\text{iid}}{\sim} \mathrm{IGa}(\alpha_\tau, \beta_\tau).$$

- For the sample variance:

$$\sigma_\epsilon^2 \sim \mathrm{IGa}(\alpha_\sigma, \beta_\sigma).$$

- For the precision parameter of the Dirichlet processes:

$$a_r \overset{\text{iid}}{\sim} \mathrm{Ga}(a_\alpha, b_\alpha),$$

for $r = 0, 1$.

# Semiparametric model

Posteriors:

- We update jointly $(\boldsymbol{\mu}_k, z_k)$

# Semiparametric model

Posteriors:

- We update jointly $(\boldsymbol{\mu}_k, z_k)$
- Posterior conditional of $m_{lj}0$, $\sigma_\epsilon^2$ and $\tau_j^2$ are conditionally conjugate

# Semiparametric model

Posteriors:

- We update jointly $(\boldsymbol{\mu}_k, z_k)$
- Posterior conditional of $m_{lj}0$, $\sigma_\epsilon^2$ and $\tau_j^2$ are conditionally conjugate
- Posterior conditional of $a_r$ is not conditionally conjugate and requires a MH step

# Semiparametric model

Posteriors:

- We update jointly $(\boldsymbol{\mu}_k, z_k)$
- Posterior conditional of $m_{lj}0$, $\sigma_\epsilon^2$ and $\tau_j^2$ are conditionally conjugate
- Posterior conditional of $a_r$ is not conditionally conjugate and requires a MH step
- Also implement a re-sampling step for $\boldsymbol{\mu}_k$

# Calling aberrations

- Key parameters of interest are: $\boldsymbol{\mu}_k = (\mu_{k1}, \mu_{k2})$ and $z_k$, and $m_{lj}$

# Calling aberrations

- Key parameters of interest are: $\boldsymbol{\mu}_k = (\mu_{k1}, \mu_{k2})$ and $z_k$, and $m_{lj}$
- Calling CNA across samples: compute

$$\mathsf{P}(|\mu_{k1}| \geq c_1 \mid \text{data}) \;\text{ and }\; \mathsf{P}(|\mu_{\mathrm{k}2}| \geq \mathrm{c}_2 \mid \text{data}),$$

for values of $c_1$ and $c_2$ to achieve a certain FDR

# Calling aberrations

- Key parameters of interest are: $\boldsymbol{\mu}_k = (\mu_{k1}, \mu_{k2})$ and $z_k$, and $m_{lj}$
- Calling CNA across samples: compute

$$\mathrm{P}(|\mu_{k1}| \geq c_1 \mid \mathrm{data}) \ \text{and} \ \mathrm{P}(|\mu_{\mathrm{k2}}| \geq \mathrm{c}_2 \mid \mathrm{data}),$$

  for values of $c_1$ and $c_2$ to achieve a certain FDR
- Calling differential CNA across disease subtypes: compute

$$\mathrm{P}(\,\{|\mu_{k1}| \geq c_1 \text{ or } |\mu_{\mathrm{k2}}| \geq \mathrm{c}_2\}\,\&\,\{\mathrm{z_k} = 1\} \mid \mathrm{data}\,),$$

# Calling aberrations

- Key parameters of interest are: $\boldsymbol{\mu}_k = (\mu_{k1}, \mu_{k2})$ and $z_k$, and $m_{lj}$
- Calling CNA across samples: compute

$$\mathsf{P}(|\mu_{k1}| \geq c_1 \mid \mathrm{data}) \text{ and } \mathsf{P}(|\mu_{\mathrm{k2}}| \geq \mathrm{c}_2 \mid \mathrm{data}),$$

  for values of $c_1$ and $c_2$ to achieve a certain FDR
- Calling differential CNA across disease subtypes: compute

$$\mathsf{P}(\{|\mu_{k1}| \geq c_1 \text{ or } |\mu_{\mathrm{k2}}| \geq \mathrm{c}_2\} \,\&\, \{\mathrm{z_k} = 1\} \mid \mathrm{data}),$$

- Sample specific: segment-specific mean copy number is

$$(\mu_{k,g_j} + m_{l,j})$$

# Simulated Data

- $n = 1,000$ probes, with locations from 1 to $n$

# Simulated Data

- $n = 1,000$ probes, with locations from 1 to $n$
- For group $g = 1$, we took 4 regions of CNA around $\{200, 400, 600, 800\}$, alternating gain and loss

# Simulated Data

- $n = 1,000$ probes, with locations from 1 to $n$
- For group $g = 1$, we took 4 regions of CNA around $\{200, 400, 600, 800\}$, alternating gain and loss
- Group $g = 2$ contains only two regions of CNA at $\{600, 800\}$, (gain and loss)

# Simulated Data

- $n = 1,000$ probes, with locations from 1 to $n$
- For group $g = 1$, we took 4 regions of CNA around $\{200, 400, 600, 800\}$, alternating gain and loss
- Group $g = 2$ contains only two regions of CNA at $\{600, 800\}$, (gain and loss)
- Aberration widths $\sim$ Ga$(2.5, 0.05)$ (accommodates large and short segments)

# Simulated Data

- $n = 1,000$ probes, with locations from 1 to $n$
- For group $g = 1$, we took 4 regions of CNA around $\{200, 400, 600, 800\}$, alternating gain and loss
- Group $g = 2$ contains only two regions of CNA at $\{600, 800\}$, (gain and loss)
- Aberration widths $\sim \text{Ga}(2.5, 0.05)$ (accommodates large and short segments)
- We took level zero for the neutral zones and a positive / negative random value $\text{Un}(0.1, 0.25)$ for the gain/loss zones

# Simulated Data

- $n = 1,000$ probes, with locations from 1 to $n$
- For group $g = 1$, we took 4 regions of CNA around $\{200, 400, 600, 800\}$, alternating gain and loss
- Group $g = 2$ contains only two regions of CNA at $\{600, 800\}$, (gain and loss)
- Aberration widths $\sim$ $\mathrm{Ga}(2.5, 0.05)$ (accommodates large and short segments)
- We took level zero for the neutral zones and a positive / negative random value $\mathrm{Un}(0.1, 0.25)$ for the gain/loss zones
- We added random errors $\mathrm{N}(0, \sigma^2)$ to the mean profiles, with $\sigma^2 \in \{0.1, 0.3\}$ to show low and high levels of noise in the log2 ratios

# Simulated Data

- We generated 100 profiles

# Simulated Data

- We generated 100 profiles
- To test our model under different conditions, only a percentage $\omega 100\%$ of the 100 profiles presented the shared aberrations

# Simulated Data

- We generated 100 profiles
- To test our model under different conditions, only a percentage $\omega 100\%$ of the 100 profiles presented the shared aberrations
- The remainder $(1-\omega)100\%$ were all neutral, showing only white noise around zero.

# Simulated Data

- We generated 100 profiles
- To test our model under different conditions, only a percentage $\omega 100\%$ of the 100 profiles presented the shared aberrations
- The remainder $(1-\omega)100\%$ were all neutral, showing only white noise around zero.
- We took three prevalence levels, $\omega \in \{1, 0.6, 0.3\}$

# Simulated Data

- We generated 100 profiles
- To test our model under different conditions, only a percentage $\omega 100\%$ of the 100 profiles presented the shared aberrations
- The remainder $(1-\omega)100\%$ were all neutral, showing only white noise around zero.
- We took three prevalence levels, $\omega \in \{1, 0.6, 0.3\}$
- Therefore, we had a total of 6 different scenarios: (3 prevalence levels × 2 noise levels).

# Simulated Data

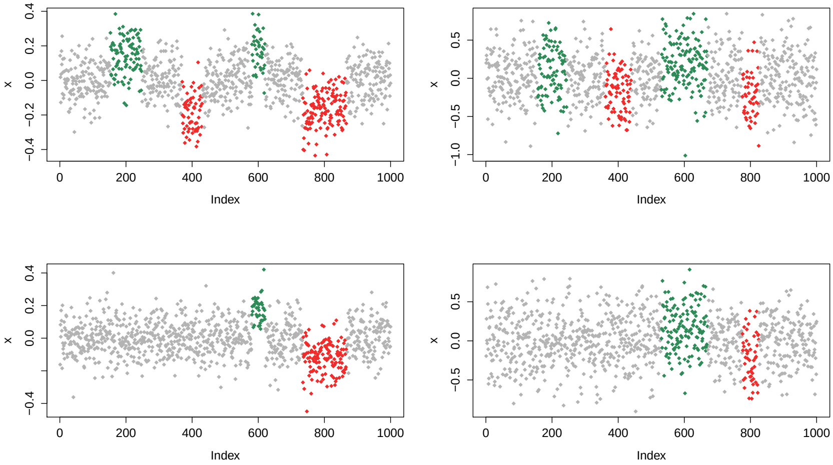


Figure : Simulated genome profile.

# Simulated Data

- S-s partitions $\{\Delta_l^j\}$ were obtained from CBS with $\alpha = 0.01$

# Simulated Data

- S-s partitions $\{\Delta_l^j\}$ were obtained from CBS with $\alpha = 0.01$
- Prior specifications: $\lambda_0^2 = \lambda_1^2 = \lambda_2^2 = 100$, $(\alpha_a, \beta_a) = (1, 1)$, $\sigma_\epsilon^2$, $(\alpha_\sigma, \beta_\sigma) = (2, 1)$

# Simulated Data

- S-s partitions $\{\Delta_l^j\}$ were obtained from CBS with $\alpha = 0.01$
- Prior specifications: $\lambda_0^2 = \lambda_1^2 = \lambda_2^2 = 100$, $(\alpha_a, \beta_a) = (1, 1)$, $\sigma_\epsilon^2$, $(\alpha_\sigma, \beta_\sigma) = (2, 1)$
- The crucial parameter $\tau_j^2$ (variance of the s-s r.e.)

# Simulated Data

- S-s partitions $\{\Delta_l^j\}$ were obtained from CBS with $\alpha = 0.01$
- Prior specifications: $\lambda_0^2 = \lambda_1^2 = \lambda_2^2 = 100$, $(\alpha_a, \beta_a) = (1, 1)$, $\sigma_\epsilon^2$, $(\alpha_\sigma, \beta_\sigma) = (2, 1)$
- The crucial parameter $\tau_j^2$ (variance of the s-s r.e.)
  - Large $\tau_j^2 \Rightarrow$ s-s effects capture most of the variability of the data, leaving little for the population mean

# Simulated Data

- S-s partitions $\{\Delta_l^j\}$ were obtained from CBS with $\alpha = 0.01$
- Prior specifications: $\lambda_0^2 = \lambda_1^2 = \lambda_2^2 = 100$, $(\alpha_a, \beta_a) = (1, 1)$, $\sigma_\epsilon^2$, $(\alpha_\sigma, \beta_\sigma) = (2, 1)$
- The crucial parameter $\tau_j^2$ (variance of the s-s r.e.)
    - Large $\tau_j^2 \Rightarrow$ s-s effects capture most of the variability of the data, leaving little for the population mean
    - Small $\tau_j^2 \Rightarrow$ variability of the data is shared between the population effects and the s-s effects

# Simulated Data

- S-s partitions $\{\Delta_l^j\}$ were obtained from CBS with $\alpha = 0.01$
- Prior specifications: $\lambda_0^2 = \lambda_1^2 = \lambda_2^2 = 100$, $(\alpha_a, \beta_a) = (1, 1)$, $\sigma_\epsilon^2$, $(\alpha_\sigma, \beta_\sigma) = (2, 1)$
- The crucial parameter $\tau_j^2$ (variance of the s-s r.e.)
  - Large $\tau_j^2 \Rightarrow$ s-s effects capture most of the variability of the data, leaving little for the population mean
  - Small $\tau_j^2 \Rightarrow$ variability of the data is shared between the population effects and the s-s effects
- We took $(\alpha_\tau, \beta_\tau) = (3, 0.01)$

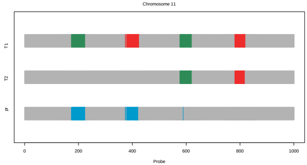
# Simulated Data

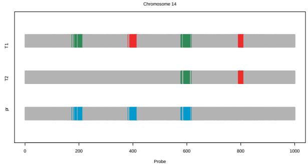
- S-s partitions $\{\Delta_l^j\}$ were obtained from CBS with $\alpha = 0.01$
- Prior specifications: $\lambda_0^2 = \lambda_1^2 = \lambda_2^2 = 100$, $(\alpha_a, \beta_a) = (1, 1)$, $\sigma_\epsilon^2$, $(\alpha_\sigma, \beta_\sigma) = (2, 1)$
- The crucial parameter $\tau_j^2$ (variance of the s-s r.e.)
  - Large $\tau_j^2 \Rightarrow$ s-s effects capture most of the variability of the data, leaving little for the population mean
  - Small $\tau_j^2 \Rightarrow$ variability of the data is shared between the population effects and the s-s effects
- We took $(\alpha_\tau, \beta_\tau) = (3, 0.01)$
- Ran Gibbs sampler for 10,000 iterations with a burn-in of 1,000, keeping every other draw

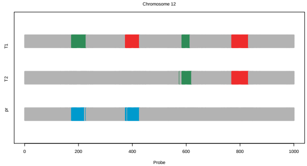
# Simulated Data

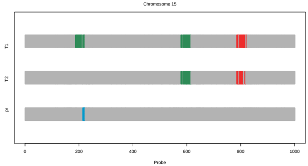
- S-s partitions $\{\Delta_l^j\}$ were obtained from CBS with $\alpha = 0.01$
- Prior specifications: $\lambda_0^2 = \lambda_1^2 = \lambda_2^2 = 100$, $(\alpha_a, \beta_a) = (1, 1)$, $\sigma_\epsilon^2$, $(\alpha_\sigma, \beta_\sigma) = (2, 1)$
- The crucial parameter $\tau_j^2$ (variance of the s-s r.e.)
  - Large $\tau_j^2 \Rightarrow$ s-s effects capture most of the variability of the data, leaving little for the population mean
  - Small $\tau_j^2 \Rightarrow$ variability of the data is shared between the population effects and the s-s effects
- We took $(\alpha_\tau, \beta_\tau) = (3, 0.01)$
- Ran Gibbs sampler for 10,000 iterations with a burn-in of 1,000, keeping every other draw
- We call differential CNAs with a $\mathrm{FDR} = 5\%$ and thresholds $c_1 = c_2 = c$ with $c = 0.10, 0.05, 0.03$ for the 100%, 60% and 30% prevalence levels

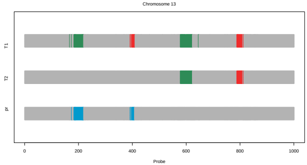
# Simulated Data

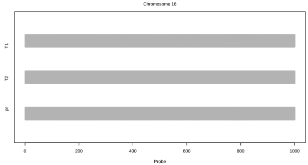












# Breast Cancer Data

- UTMDACC conducted arrayCGH experiments using samples from 122 patients

# Breast Cancer Data

- UTMDACC conducted arrayCGH experiments using samples from 122 patients
- Tumor samples of 122 patientes are: 60 - ER+, 11 - PR+, and 51 - TN

# Breast Cancer Data

- UTMDACC conducted arrayCGH experiments using samples from 122 patients
- Tumor samples of 122 patientes are: 60 - ER+, 11 - PR+, and 51 - TN
- Concentrated on comparing ER+ and TN (111 samples in total)

# Breast Cancer Data

- UTMDACC conducted arrayCGH experiments using samples from 122 patients
- Tumor samples of 122 patientes are: 60 - ER+, 11 - PR+, and 51 - TN
- Concentrated on comparing ER+ and TN (111 samples in total)
- We split the data on chromosomes

# Breast Cancer Data

- UTMDACC conducted arrayCGH experiments using samples from 122 patients
- Tumor samples of 122 patientes are: 60 - ER+, 11 - PR+, and 51 - TN
- Concentrated on comparing ER+ and TN (111 samples in total)
- We split the data on chromosomes
- Sample-specific partitions $\{\Delta_l^j\}$ were obtained from CBS with $\alpha = 0.01$

# Breast Cancer Data

- UTMDACC conducted arrayCGH experiments using samples from 122 patients
- Tumor samples of 122 patientes are: 60 - ER+, 11 - PR+, and 51 - TN
- Concentrated on comparing ER+ and TN (111 samples in total)
- We split the data on chromosomes
- Sample-specific partitions $\{\Delta_l^j\}$ were obtained from CBS with $\alpha = 0.01$
- Same prior specifications as in simulated data

# Breast Cancer Data

- UTMDACC conducted arrayCGH experiments using samples from 122 patients
- Tumor samples of 122 patientes are: 60 - ER+, 11 - PR+, and 51 - TN
- Concentrated on comparing ER+ and TN (111 samples in total)
- We split the data on chromosomes
- Sample-specific partitions $\{\Delta_l^j\}$ were obtained from CBS with $\alpha = 0.01$
- Same prior specifications as in simulated data
- We call differential CNA with a $\mathrm{FDR} = 5\%$ with thresholds $c_1 = c_2 = 0.2$ for all chromosomes

# Breast Cancer Data

- We found CNA differences between the two cancer subtypes in 16 of the 23 chromosomes

# Breast Cancer Data

- We found CNA differences between the two cancer subtypes in 16 of the 23 chromosomes
- Predominantly in chromosomes 3 –7, 9 – 12, 14 – 19, and 23

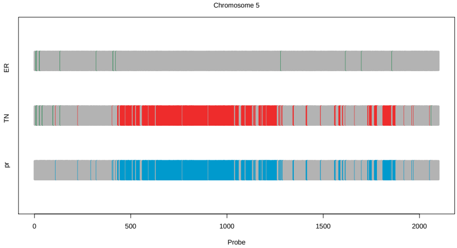
## Breast Cancer Data

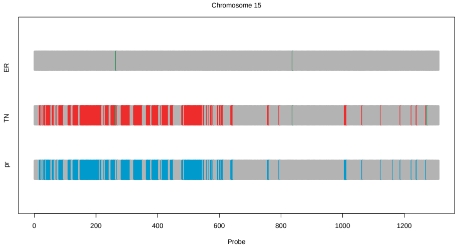
- We found CNA differences between the two cancer subtypes in 16 of the 23 chromosomes
- Predominantly in chromosomes 3 –7, 9 – 12, 14 – 19, and 23
- Chromosome 5 is confirmatory

# Breast Cancer Data

- We found CNA differences between the two cancer subtypes in 16 of the 23 chromosomes
- Predominantly in chromosomes 3 –7, 9 – 12, 14 – 19, and 23
- Chromosome 5 is confirmatory
- Chromosome 15 is a new finding

# Breast Cancer Data





# Breast Cancer Data

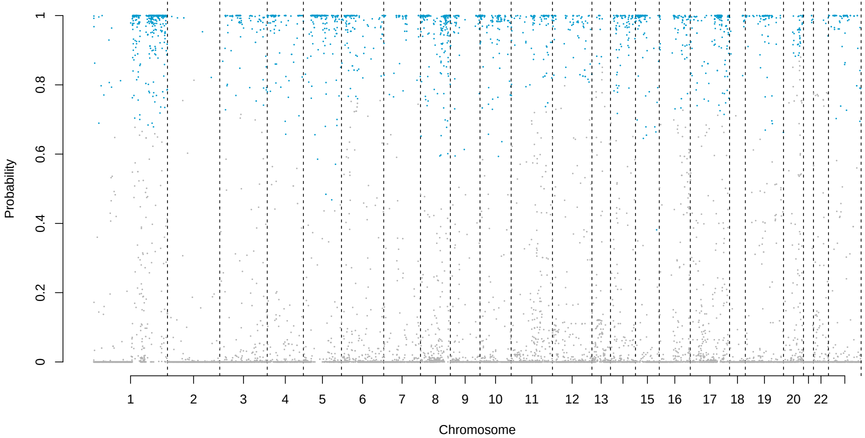


Figure : Differential CNA probabilities for all chromosomes.

# References

Baladandayuthapani, V., Ji, Y., Talluri, R., Nieto-Barajas, L.E., and Morris, J.S. (2010). Bayesian random segmentation models to identify shared copy number aberrations for array CGH data. *Journal of the American Statistical Association* **105**, 1358–1375.

Guha, S., Li, Y., and Neuberg, D. (2008). Bayesian hidden Markov modeling of array CGH data. *Journal of the American Statistical Association* **103**, 485–497.

Olshen, A.B., Venkatraman, E.S., Lucito, R., and Wigler, M. (2004). Circular binary segmentation for the analysis of array-based DNA copy number data. *Biostatistics* **4**, 557–572.

Shah, S.P., Lam, W.L., Ng, R.T., and Murphy, K.P. (2007). Modeling recurrent DNA copy number alterations in array CGH data. *Bioinformatics* **23**, 450–458.

Yau, C., Papaspiliopoulos, O., Roberts, G., and Holmes, C. (2011). Bayesian non-parametric hidden Markov models with applications in genomics. *Journal of the Royal Statistical Society, Series B* **73**, 33–57.