A COMPARATIVE GENOMICS APPROACH TO THE IDENTIFICATION OF QTL CANDIDATE GENES

Authors: HOWSUN JOW  
– School of Mathematics & Statistics, Newcastle University,  
Newcastle upon Tyne U.K.  
howsun.jow@ncl.ac.uk  

RICHARD J. BOYS  
– School of Mathematics & Statistics, Newcastle University,  
Newcastle upon Tyne U.K.  
richard.boys@ncl.ac.uk  

DARREN J. WILKINSON  
– School of Mathematics & Statistics, Newcastle University,  
Newcastle upon Tyne U.K.  
d.j.wilkinson@ncl.ac.uk  

Abstract:  
• Despite rapid advances in sequencing technology, many commercially relevant species remain unsequenced, and many that are sequenced have very poorly annotated genomes. There is therefore still considerable interest in using comparative approaches to exploit information from well-characterised model organisms in order to better understand related species. This paper develops a statistical method for automating part of a comparative genomics bioinformatic pipeline for the identification of genes and genomic regions in a model organism associated with a QTL region in an unsequenced species. A non-parametric Bayesian statistical model is used for characterising the density of a large number of BLAST hits across a model species genome. The method is illustrated using a test problem demonstrating that markers associated with Bovine hemoglobin can be automatically mapped to a region of the human genome containing human hemoglobin genes. Consequently, by exploiting the (relatively) high quality of genome annotation for model organisms and humans it is possible to quickly identify candidate genes in those well-characterised genomes relevant to the quantitative trait of interest.

Key-Words:  
• Bayesian; non-parametric; density estimation; QTL; BLAST; mapping; comparative genomics.

AMS Subject Classification:  
• 62F15, 62G07, 92D99.
1. INTRODUCTION

The mapping of the genetic component influencing quantitative traits of a species, such as height and weight, can be achieved even in the absence of a complete physical map of a species’ genome. This is called quantitative trait loci (QTL) mapping. One method by which QTLs can be mapped utilizes a map of typed genetic markers in order to establish the statistical correlation between a given quantitative trait and a given point, between two markers, on the genetic map ([12]). This allows for the identification of regions which are highly statistically correlated with the quantitative trait and therefore likely to contain a QTL. These regions can then be sequenced and the genes influencing the quantitative trait can be identified.

This method of finding the genes that influence a particular quantitative trait has its drawbacks. For one thing it is dependent on the quality and resolution of the genetic map used to map the QTLs. A low resolution genetic map would lead to a low resolution QTL map in which relatively large regions are identified as being statistically significant and therefore likely to contain a QTL. This in turn requires the sequencing of large portions of the sequence genome. Alternatively the method can be used on high resolution genetic maps. However, this too has problems: constructing high resolution genetic maps is far from a trivial process and can be expensive and labour intensive, especially for traditional linkage maps.

The method described in this paper uses a comparative genomics approach to locate genes which are correlated with the QTL. It works by first identifying statistically significant QTL regions. Then a high resolution map is constructed by integrating available partial maps of the chromosome in which the QTL regions lie into a single map. There are a number of methods available for integrating partial genetic maps ([14, 18, 16, 19, 6, 13, 11]) and in this paper we use a Bayesian approach to map integration developed by Jow et al. ([11]).

On obtaining a high resolution integrated map, the markers lying between the QTL flanking markers are identified and a BLAST ([1]) search made of their sequences against the genome of a target species. This gives us a series of “hits” on the target genome, that is, locations where the search sequences match. Using these hits it is possible to estimate the probability density of hits across the target genome using, for example, standard kernel density techniques ([17]) or Bayesian alternatives based on Dirichlet processes ([4, 2, 3]). We will use a Bayesian density estimate and then threshold this density to identify regions along the target species which are likely to contain genes performing similar functions to the genes associated with the QTL of the source species.

The rest of this paper is organised as follows. Section 2.1 describes how to construct a Bayesian density estimate from a collection of BLAST hits across
a number of chromosomes. The model is described in detail, with the resulting MCMC algorithm available in the appendix. This section also describes a procedure for determining the location of regions likely to contain genes associated with the QTL. Section 3 validates the MCMC algorithm and implementation on a synthetic example and Section 4 provides a real example of how our method can be used to help identify genes associated with QTLs obtained from the Bovine Hemoglobin genome by using the Human genome.

2. METHODS

In this section we describe how to construct a Bayesian density estimate from a collection of BLAST hits and the procedure for determining the location of intervals likely to contain genes associated with the QTL.

2.1. Bayesian density estimation

Suppose that the target genome consists of $C$ chromosomes with lengths $L_1, ..., L_C$. The data take the form of $n$ BLAST hits describing the location ($y$) and chromosome ($c$) on which each hit was made: $(y_i, c_i), i = 1, ..., n$. Let $n_c$ be the number of observed hits on chromosome $c$. We construct the Bayesian density estimate by modelling these locations as an infinite mixture of normal distributions with unknown means ($\mu$) and variances ($\sigma^2$) and with these parameters $\phi = (\mu, \sigma^2)$ resulting from a Dirichlet process with a particular base distribution. Let $\theta_c$ denote the probability of a hit occurring on chromosome $c$. The formulation of the model is slightly complicated by the need to have a continuous density across the $C$ chromosomes. In summary we have, for $i = 1, ..., n$ and $c_i \in \{1, ..., C\}$,

$$\theta = (\theta_1, ..., \theta_C) | \alpha \sim Dir(\alpha \ell),$$

$$Y_i, c_i | \phi_{ic_i}, \theta_c \sim N(\mu_{ic_i}, \sigma_{ic_i}^2) \times Bern(\theta_{c_i}),$$

$$\phi_{ic_i} | G_{c_i} \sim G_{c_i},$$

$$G_{c_i} | \alpha \sim DP(\alpha, G_{0c_i}),$$

$$G_{0c_i} = U(0, L_{c_i}) \times Inv \Gamma(a, L_{c_i}^2 b),$$

where $DP(\alpha, G_0)$ denotes a Dirichlet process with base measure $G_0$ and concentration parameter $\alpha$, and $\ell$ is the normalized form of $L$, that is, $\ell_c = L_c / \sum_{j=1}^{C} L_j$. The form of the base distribution has been chosen so that it is independent of the scale used to measure the location of the BLAST hits, for example, Mb or b.
All that remains for a full model specification is to choose the prior distribution for $\alpha$. In this paper we take a flexible semi-conjugate form with

$$
\alpha \sim \Gamma(g, h)
$$

Given the data model above and the hit data, it is generally not possible to derive an analytical expression for the probability density at an arbitrary point $y$ on an unknown chromosome $k$. However, numerical sampling methods can be used approximate this (predictive) density as

$$
\pi(y, k|\mathcal{D}) \simeq \frac{1}{Tn} \sum_{t=1}^{T} \sum_{i=1}^{n} \pi(y, k|\phi_{ik}^t, \theta_k^t),
$$

where $\left(\phi_{ik}^t, \theta_k^t\right)$, $t = 1, ..., T$, is a sample from the posterior distribution $\pi(\phi_{ik}, \theta_k|\mathcal{D})$ obtained using an appropriate sampling algorithm. In this paper we have used an MCMC algorithm based on one by Escobar and West ([3]); the algorithm is described in the appendix.

### 2.2. Identification of QTL intervals

On obtaining the probability density of hits across the entire target genome the remaining task is to identify regions with a high probability density. This is done by identifying the highest density regions (HDRs) containing a given percentage of the density; see [10]. For example, a 75% HDR could be found across all the chromosomes. Given that in our model the target genome is one-dimensional, the HDR would be a set of regions across all the chromosomes. These regions can then be searched for genes of interest.

### 3. SIMULATED DATA

To validate our MCMC algorithm and implementation, we simulated a dataset of 200 “hits” spread over three chromosomes with lengths 100 Mb, 200 Mb and 400 Mb. The distribution of the locations of hits on the different chromosomes were taken to be normal distributions on chromosomes 1 and 3 and a mixture of two normal distributions on chromosome 2; see the dashed lines in Figure 1. Also the probability of a hit being located on a particular chromosome was taken as being proportional to the length of the chromosome, that is, with probability $1/7, 2/7$ and $4/7$ for chromosomes 1, 2 and 3 respectively.

We specify the base distribution for the cluster variances ($\sigma^2$) by taking $a = 2.05$ and $b = 0.000105$, so that $E(\sigma^2) = 10^{-4}L^2$ and $SD(\sigma^2) = \sqrt{2}E(\sigma^2)$. 
For example, on chromosome 1 this gives $E(\sigma^2) = (1\text{Mb})^2$, that is, suggests cluster standard deviations are around 1 Mb. We also input fairly weak prior information for $\alpha$ by taking $g = 4$ and $h = 2$, that is, $E(\alpha) = 2$ and $SD(\alpha) = 1$.

Figure 1: The theoretical distribution of hits along the chromosomes (dashed lines) and the Bayesian density estimate obtained from the simulated data (dotted line).

3.1. Results

The MCMC algorithm outlined in the appendix was applied to the simulated dataset. Convergence was assessed by using informal visual methods and the diagnostics suggested by Gelman and Rubin ([5]) and by Heidelberger and Welch ([7]). We found that a burn-in of $100K$ iterations was required to achieve convergence and we then ran the chain for a further $100K$ iterations, thinning the output by taking every $100^{th}$ iterate. This gave a posterior sample of size $1K$ observations from which we could calculate the Bayesian density estimate (2.1) across the (simulated) chromosomes. The results are summarized in in Table 1 and Figure 1, and show that there is a reasonably close match between the theoretical and estimated probabilities of a hit being found on a particular chromosome and between the Bayesian density estimate for the location of hits and their generating distribution.
4. BOVINE HEMOGLOBIN MARKER DATA

To illustrate the power of our method, we now show how the Human genome can be used to help identify genes associated with QTLs obtained from the Bovine Hemoglobin genome. The sequences of molecular markers associated with Bovine Hemoglobin genes were taken from the NCBI “GENE” database ([15]) and the markers we use are given in Table 2. For our analysis, we use the same input parameters \((a, b, g)\) and \((h)\) as in Section 3.

Table 2: Markers associated with Bovine Hemoglobin genes.

<table>
<thead>
<tr>
<th>Marker name</th>
<th>Associated gene</th>
<th>Gene symbol</th>
<th>Sequence length</th>
</tr>
</thead>
<tbody>
<tr>
<td>REN97351</td>
<td>Hemoglobin Beta</td>
<td>HBB</td>
<td>248</td>
</tr>
<tr>
<td>RH69634</td>
<td>Hemoglobin Beta</td>
<td>HBB</td>
<td>141</td>
</tr>
<tr>
<td>PMC115301P1</td>
<td>Hemoglobin Beta</td>
<td>HBB</td>
<td>136</td>
</tr>
<tr>
<td>GDB:178694</td>
<td>Hemoglobin Beta</td>
<td>HBB</td>
<td>300</td>
</tr>
<tr>
<td>HBB</td>
<td>Hemoglobin Gamma</td>
<td>HBG</td>
<td>171</td>
</tr>
<tr>
<td>PMC86017P3</td>
<td>Hemoglobin Gamma</td>
<td>HBG</td>
<td>267</td>
</tr>
<tr>
<td>PMC21968P1</td>
<td>Hemoglobin Epsilon</td>
<td>HBE</td>
<td>989</td>
</tr>
<tr>
<td>Hba-a1</td>
<td>Hemoglobin Alpha</td>
<td>HBA</td>
<td>188</td>
</tr>
<tr>
<td>AW312144</td>
<td>Hemoglobin Alpha</td>
<td>HBA</td>
<td>327</td>
</tr>
<tr>
<td>CB603723</td>
<td>Hemoglobin Zeta</td>
<td>HBZ</td>
<td>312</td>
</tr>
<tr>
<td>BE740596</td>
<td>Hemoglobin Theta 1</td>
<td>HBQ</td>
<td>277</td>
</tr>
<tr>
<td>AW428039.1</td>
<td>Hemoglobin Mu</td>
<td>HBM</td>
<td>193</td>
</tr>
</tbody>
</table>

4.1. Results

A BLAST search of these markers was conducted against the reference Human genome (NCBI 36.3 build) using the parameters listed in Table 3, and gave 188 hits distributed across 15 chromosomes. The MCMC algorithm was
then run on these hit data. As with the simulated data, convergence was assessed by using informal visual methods and standard diagnostics tools. Again, we found that a burn-in of $100K$ iterations was required to achieve convergence.

**Table 3:** BLAST search parameters against the Human genome for the Bovine Hemoglobin markers.

<table>
<thead>
<tr>
<th>BLAST parameter</th>
<th>argument</th>
<th>value</th>
</tr>
</thead>
<tbody>
<tr>
<td>Expectation Value</td>
<td>$-e$</td>
<td>0.1</td>
</tr>
<tr>
<td>Gap Cost</td>
<td>$-G$</td>
<td>5</td>
</tr>
<tr>
<td>Gap Extension Cost</td>
<td>$-E$</td>
<td>2</td>
</tr>
<tr>
<td>Nucleotide Mismatch Cost</td>
<td>$-q$</td>
<td>3</td>
</tr>
<tr>
<td>Nucleotide Match reward</td>
<td>$-r$</td>
<td>2</td>
</tr>
</tbody>
</table>

We then ran the chain for a further $100K$ iterations, thinning the output by taking every $100^{th}$ iterate, to obtain a posterior sample of size $1K$ observations. The results are summarized in Tables 4, 5 and in Figure 2. The posterior probability of a hit being on the target human chromosomes is shown in Table 4.

**Table 4:** Probability of a hit being located on each chromosome of the Human genome (to 3 d.p.).

<table>
<thead>
<tr>
<th>Chromosome</th>
<th>Probability</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>0.032</td>
</tr>
<tr>
<td>3</td>
<td>0.022</td>
</tr>
<tr>
<td>5</td>
<td>0.016</td>
</tr>
<tr>
<td>6</td>
<td>0.005</td>
</tr>
<tr>
<td>7</td>
<td>0.011</td>
</tr>
<tr>
<td>9</td>
<td>0.032</td>
</tr>
<tr>
<td>11</td>
<td>0.620</td>
</tr>
<tr>
<td>12</td>
<td>0.016</td>
</tr>
<tr>
<td>13</td>
<td>0.011</td>
</tr>
<tr>
<td>14</td>
<td>0.005</td>
</tr>
<tr>
<td>15</td>
<td>0.005</td>
</tr>
<tr>
<td>16</td>
<td>0.161</td>
</tr>
<tr>
<td>17</td>
<td>0.027</td>
</tr>
<tr>
<td>19</td>
<td>0.021</td>
</tr>
<tr>
<td>20</td>
<td>0.016</td>
</tr>
<tr>
<td>2, 4, 8, 10, 18, 21, 22, X, Y</td>
<td>$\approx 0$</td>
</tr>
</tbody>
</table>

Table 5 contains the 50%, 60% and 75% highest density regions (HDRs) across all chromosomes, calculated using the method of Hyndman ([10]). Figure 2 gives a graphical view of the HDRs for those chromosomes with a hit probability of at least 5%, that is, for chromosomes 11 and 16. The 50% and 60% HDRs
determined over all chromosomes point to genes of interest only on chromosomes 11 and 16. The aim of our method in this example is to identify regions on the human genome which are associated with the Bovine Hemoglobin genome.

### Table 5: HDR intervals on the Human genome for the Bovine Hemoglobin markers.

<table>
<thead>
<tr>
<th>HDR level</th>
<th>Chromosome</th>
<th>Intervals</th>
<th>Number of candidate genes</th>
</tr>
</thead>
<tbody>
<tr>
<td>50%</td>
<td>11</td>
<td>4.59 Mb– 5.77 Mb</td>
<td>95</td>
</tr>
<tr>
<td></td>
<td>16</td>
<td>0.00 Mb– 0.36 Mb</td>
<td>27</td>
</tr>
<tr>
<td>60%</td>
<td>11</td>
<td>4.37 Mb– 5.98 Mb</td>
<td>104</td>
</tr>
<tr>
<td></td>
<td>16</td>
<td>0.00 Mb– 0.49 Mb</td>
<td>33</td>
</tr>
<tr>
<td>75%</td>
<td>11</td>
<td>3.88 Mb– 6.47 Mb</td>
<td>131</td>
</tr>
<tr>
<td></td>
<td>11</td>
<td>54.52 Mb– 58.11 Mb</td>
<td>121</td>
</tr>
<tr>
<td></td>
<td>16</td>
<td>0.00 Mb– 0.78 Mb</td>
<td>59</td>
</tr>
<tr>
<td></td>
<td>9</td>
<td>124.28 Mb– 124.53 Mb</td>
<td>13</td>
</tr>
<tr>
<td></td>
<td>17</td>
<td>2.51 Mb– 3.72 Mb</td>
<td>35</td>
</tr>
<tr>
<td></td>
<td>17</td>
<td>61.14 Mb– 61.92 Mb</td>
<td>30</td>
</tr>
</tbody>
</table>

**Figure 2:** Figure showing the Bayesian density estimate of BLAST hits across chromosomes 11 and 16 of the Human genome.

If we look in detail at the Human genome, its Hemoglobin genes are located in two clusters on chromosomes 11 and 16, with the \( \beta \)-globin cluster spanning an interval of roughly 5.20–5.25 Mb on chromosome 11 and the \( \alpha \)-globin cluster span-
ning an interval of roughly 0.14–0.17 Mb on chromosome 16. Thus our method has correctly and reasonably accurately identified the appropriate regions on the Human genome.

If we examine the Human genes found within these 60% HDRs, we find the 33 genes located on chromosome 16 listed in Table 6. These include the five known functional genes and two pseudo-genes of the human α-globin locus.

**Table 6:** A list of genes in the 60% HDR for chromosome 16. The genes in bold are the 5 known functional genes present in the Human α-globin locus and those in italics are the two known pseudo-genes ([9]).

<table>
<thead>
<tr>
<th>Ensembl Gene ID</th>
<th>Gene name</th>
</tr>
</thead>
<tbody>
<tr>
<td>ENSG00000220481</td>
<td>Z84812.3</td>
</tr>
<tr>
<td>ENSG00000181104</td>
<td>WASH4P</td>
</tr>
<tr>
<td>ENSG00000219509</td>
<td>Z84723.2</td>
</tr>
<tr>
<td>ENSG00000185203</td>
<td>Z84723.1</td>
</tr>
<tr>
<td>ENSG00000161980</td>
<td>POLR3K</td>
</tr>
<tr>
<td>ENSG00000161981</td>
<td>C16orf33</td>
</tr>
<tr>
<td>ENSG00000007384</td>
<td>RHDF1</td>
</tr>
<tr>
<td>ENSG00000103152</td>
<td>MPG</td>
</tr>
<tr>
<td>ENSG00000103148</td>
<td>C16orf35</td>
</tr>
<tr>
<td>ENSG00000130656</td>
<td><strong>HBZ</strong></td>
</tr>
<tr>
<td>ENSG00000206178</td>
<td><strong>Z84721.1</strong></td>
</tr>
<tr>
<td>ENSG00000206177</td>
<td><strong>HBM</strong></td>
</tr>
<tr>
<td>ENSG00000218072</td>
<td><strong>Z84721.4</strong></td>
</tr>
<tr>
<td>ENSG00000188536</td>
<td><strong>HBA2</strong></td>
</tr>
<tr>
<td>ENSG00000206172</td>
<td><strong>HBA1</strong></td>
</tr>
<tr>
<td>ENSG00000207243</td>
<td>Y_RNA</td>
</tr>
<tr>
<td>ENSG00000086506</td>
<td><strong>HBQ1</strong></td>
</tr>
<tr>
<td>ENSG00000007392</td>
<td>LUC7L</td>
</tr>
<tr>
<td>ENSG00000206168</td>
<td>Z69890.1</td>
</tr>
<tr>
<td>ENSG00000167930</td>
<td>ITFG3</td>
</tr>
<tr>
<td>ENSG00000215289</td>
<td>AC004754.1</td>
</tr>
<tr>
<td>ENSG000000076344</td>
<td>RGS11</td>
</tr>
<tr>
<td>ENSG00000206156</td>
<td>ARHGDIG</td>
</tr>
<tr>
<td>ENSG00000185615</td>
<td>PDI2</td>
</tr>
<tr>
<td>ENSG00000103126</td>
<td>AXIN1</td>
</tr>
<tr>
<td>ENSG000000085504</td>
<td>MRPL28</td>
</tr>
<tr>
<td>ENSG00000129925</td>
<td>TMEM8</td>
</tr>
<tr>
<td>ENSG00000216963</td>
<td>Z97634.3</td>
</tr>
<tr>
<td>ENSG00000103200</td>
<td>NME4</td>
</tr>
<tr>
<td>ENSG00000103202</td>
<td>DECR2</td>
</tr>
<tr>
<td>ENSG000000090565</td>
<td>RAB11FIP3</td>
</tr>
<tr>
<td>ENSG00000201034</td>
<td>Y_RNA</td>
</tr>
<tr>
<td>ENSG00000217816</td>
<td>RP1-196A12.1</td>
</tr>
</tbody>
</table>

Additionally, C16orf35 is known to be involved in the regulation of α-globin. The corresponding list for chromosome 11 contains 104 genes and includes the five
Identification of QTL Candidate Genes

known functional Hemoglobin genes in the human $\beta$-globin locus (HBE1, HBG1, HBG2, HBD and HBB) and one known hemoglobin pseudogene (HBBP1). The annotations available for the remaining genes in these lists show no known direct link to Hemoglobin or its regulation.

5. CONCLUSIONS

In this paper we have developed a method for estimating the density of BLAST hits across chromosomes on a target genome. This estimate can then be used to determine highest density regions (HDRs) on the target genome for genes associated with the QTL of interest.

The method has been shown to work well on both simulated data and real data. In this latter case this involved obtaining BLAST hits for a number of Bovine Hemoglobin markers (given in Table 2) against the Human genome. We were able to construct the density estimate of BLAST hits across the Human genome and thereby determine the highest density regions. The regions obtained were found to contain the Human $\alpha$-globin and $\beta$-globin loci ([8]).

Currently our method uses a fairly superficial treatment of BLAST hits and does not, for example, distinguish between poor BLAST hits and good ones. Future work might involve exploring how to incorporate properly weighted BLAST hits so that the better hits contribute more to the density estimate and this might lead to more accurate HDRs. Also, because the chromosomes have finite length, strictly the density across the chromosomes should have finite support. This could be achieved, for example, by replacing the Gaussian distribution for the location of clusters by (a mixture of) truncated Gaussian distributions. Unfortunately, such a modification does lead to analytical intractability in the calculations underpinning the Bayesian density estimate, though research into using such distributions is also a possible area of future work.

APPENDIX

The MCMC algorithm is a Gibbs sampler for the cluster parameters $\phi_{ic} = (\mu_{ic}, \sigma^2_{ic})$, $i = 1, ..., n$, and the parameters $(\alpha, \beta)$. In the following sections, we derive the posterior conditional distributions for these parameters.
A. The cluster parameters

Here we derive the posterior conditional distributions for the $\phi_{ic_i} = (\mu_{ic_i}, \sigma^2_{ic_i})$, $i = 1, ..., n$. Letting $\phi'_{ic_i} = \{ \phi_{jc_i} : j \neq i \}$, the conditional prior density for $\phi_{ic_i} | \phi'_{ic_i}$ is

$$\pi(\phi_{ic_i} | \phi'_{ic_i}) = \frac{\alpha}{\alpha + n_{ci} - 1} g_{0c_i}(\phi_{ic_i}) + \sum_{j \neq i} \frac{1}{\alpha + n_{ci} - 1} \delta_{\phi_{jc_i}}(\phi_{ic_i})$$

where $g_{0c_i}$ is the probability density corresponding to the distribution $G_{0c_i}$, $n_{ci}$ is the number of observed hits on chromosome $c_i$ and $\delta_y(x)$ is Dirac’s delta function ($\delta_y(x) = 0$ if $x \neq y$ and $\int \delta_y(x) \, dx = 1$).

Multiplying this by the likelihood $\pi(y_i, c_i | \phi_{ic_i}, \theta_{c_i})$, we get the conditional posterior density

$$\pi(\phi_{ic_i} | \phi'_{ic_i}, y_i, \theta_{c_i}) = q_{i0} g_{ic_i}(\phi_{ic_i}) + \sum_{j \neq i} q_{ij} \delta_{\phi_{jc_i}}(\phi_{ic_i})$$

where

$$q_{ij} = \kappa \pi(y_i, c_i | \phi_{jc_i}, \theta_{c_i}) ,$$

$$q_{i0} = \kappa \alpha \int \pi(y_i, c_i | \phi_{ic_i}, \theta_{c_i}) g_{0c_i}(\phi_{ic_i}) \, d\phi_{ic_i} ,$$

$$g_{ic_i}(\phi_{ic_i}) = \frac{\pi(y_i, c_i | \phi_{ic_i}) g_{0c_i}(\phi_{ic_i})}{\int \pi(y_i, c_i | \phi_{ic_i}) g_{0c_i}(\phi_{ic_i}) \, d\phi_{ic_i}}$$

and $\kappa$ is a normalizing constant such that

$$q_{i0} + \sum_{j \neq i} q_{ij} = 1 .$$

We can derive closed form expressions for the densities $g_{ic_i}$ and the $q_{ij}$ by using the base distribution for the Dirichlet process $G_{c_i}$, $G_{0c_i} = U(0, L_{c_i}) \times \text{Inv} \Gamma(a, L_{c_i}^2 b)$, as follows. Let $\phi(\cdot | a, b^2)$ denote the $N(a, b^2)$ density, $\Psi_a(\cdot | b, c)$ the $St(a, b, c)$ density and $\Psi_{\alpha}(\cdot)$ the $t_{\alpha}$ distribution function. Note that if $X \sim t_{\alpha}$ then $b + \sqrt{\tau} X \sim St(a, b, c)$. Also, to simplify notation, we write $\tau = \sigma^2$. Then

$$q_{ij} = \kappa \pi(y_i, c_i | \phi_{ic_i}, \theta_{c_i}) = \kappa \theta_{c_i} \phi(y_i | \mu_{ic_i}, \tau_{ic_i})$$
Identification of QTL Candidate Genes

13

and

\[ q_{i0} = \kappa \alpha \int_0^L \pi(y_i, c_i | \phi_{ic_i}, \theta_{ic_i}) g_{0c_i}(\phi_i) \, d\phi_i \]

\[ = \kappa \alpha \int_0^{L_{ci}} \phi_{ic_i} \phi(y_i | \mu_{ic_i}, \tau_{ic_i}) \times \frac{1}{L_{ci}} \times \left( \frac{L_{ci}^2 b}{a} \right)^a \frac{\tau_{ic_i}^{-a-1}}{\Gamma(a)} \frac{e^{-L_{ci}^2 b/\tau_{ic_i}}}{\tau_{ic_i}} \, d\tau_{ic_i} \, d\mu_{ic_i} \]

\[ = \kappa \frac{\theta_{ic_i}}{L_{ci}} \int_0^{L_{ci}} \psi_{2a}(\mu | y_i, L_{ci}^2 b/a) \, d\mu \]

\[ = \kappa \frac{\theta_{ic_i}}{L_{ci}} \left\{ \Psi_{2a} \left( \frac{1 - y_i/L_{ci}}{\sqrt{b/a}} \right) - \Psi_{2a} \left( \frac{-y_i/L_{ci}}{\sqrt{b/a}} \right) \right\} . \]

Also, for \( 0 \leq \mu_{ic_i} \leq L_{ci} \), \( \tau_{ic_i} > 0 \)

\[ g_{ic_i}(\phi_{ic_i}) = \frac{\pi(y_i, c_i | \phi_{ic_i}, \theta_{ic_i}) g_0(\phi_{ic_i})}{\pi(y_i, c_i | \phi_{ic_i}, \theta_{ic_i}) g_0(\phi_{ic_i})} \, d\phi_{ic_i} \]

\[ = \frac{\phi(y_i | \mu_{ic_i}, \tau_{ic_i}) \times (L_{ci}^2 b)^a \tau_{ic_i}^{-a-1} e^{-L_{ci}^2 b/\tau_{ic_i}} / \Gamma(a)}{\Psi_{2a} \left( \frac{1 - y_i/L_{ci}}{\sqrt{b/a}} \right) - \Psi_{2a} \left( \frac{-y_i/L_{ci}}{\sqrt{b/a}} \right)} \]

\[ \times \sqrt{2\pi} \frac{1}{\Gamma(a)} \left\{ \Psi_{2a} \left( \frac{1 - y_i/L_{ci}}{\sqrt{b/a}} \right) - \Psi_{2a} \left( \frac{-y_i/L_{ci}}{\sqrt{b/a}} \right) \right\} \]

\[ \times \exp \left\{ - \left( \frac{L_{ci}^2 b + (y_i - \mu_{ic_i})^2}{2} / \tau_{ic_i} \right) \right\} . \]

For simulation purposes, it is useful to note that

\[ g_{ic_i}(\phi_{ic_i}) = \pi(\mu_{ic_i}) \pi(\sigma_{ic_i}^2 | \mu_{ic_i}) \]

where

\[ \mu_{ic_i} \sim St(2a, y_i, L_{ci}^2 b/a), \quad 0 \leq \mu_{ic_i} \leq L_{ci} \]

and

\[ \sigma_{ic_i}^2 | \mu_{ic_i} \sim Inv \Gamma \left( a + \frac{1}{2}, L_{ci}^2 b + \frac{(y_i - \mu_{ic_i})^2}{2} \right) . \]

B. The remaining parameters

Here we derive the posterior conditional distributions for \( \alpha \) and \( \theta \). The procedure is a generalisation of that used by Escobar and West ([3]).
Suppose that chromosome \( c \) has \( n_c \) hits arranged in \( k_c \) clusters (\( c = 1, 2, \ldots, C \)). Then the probability function for the number of clusters on chromosome \( c \) is
\[
\pi(k_c|n_c, \alpha, \theta) \propto \begin{cases} 
\alpha^{k_c} \Gamma(\alpha) \Gamma(\alpha + n_c), & \text{if } n_c > 0, \\
1, & \text{if } n_c = 0.
\end{cases}
\]

Let \( \bar{k} = (k_1, k_2, \ldots, k_C) \). As we have independent Dirichlet processes for each chromosome, \( k_c|n_c, \alpha \) are independent for \( c = 1, 2, \ldots, C \) and so
\[
\pi(k|n, \alpha, \theta) \propto \prod_{c=1}^C \alpha^{k_c} \frac{\Gamma(\alpha) \Gamma(\alpha + n_c)}{\Gamma(\alpha + n_c)},
\]
for \( k_c = 1, 2, \ldots, n_c \) if \( n_c > 0 \) or \( k_c = 0 \) if \( n_c = 0 \) (\( c = 1, 2, \ldots, C \)). This can be simplified slightly by letting \( A = \{ c : n_c > 0 \} \) with size \( |A| \), and renumbering the chromosomes so that \( n_c > 0 \) for \( c = 1, 2, \ldots, |A| \) and \( n_c = 0 \) for \( c = |A| + 1, |A| + 2, \ldots, C \), giving
\[
\pi(k|n, \alpha, \theta) \propto |A| \prod_{c=1}^{|A|} \alpha^{k_c} \frac{\Gamma(\alpha) \Gamma(\alpha + n_c)}{\Gamma(\alpha + n_c)}.
\]

The probability function for the number of hits on each chromosome has a multinomial distribution, with
\[
\pi(n|\alpha, \theta) \propto \prod_{c=1}^C \theta_n^{n_c},
\]
and so the likelihood function for \( (\alpha, \theta) \) is
\[
\pi(k, n|\alpha, \theta) = \pi(k|n, \alpha, \theta) \pi(n|\alpha, \theta) \propto \prod_{c=1}^{|A|} \alpha^{k_c} \frac{\Gamma(\alpha) \Gamma(\alpha + n_c)}{\Gamma(\alpha + n_c)} \prod_{c=1}^C \theta_n^{n_c}.
\]

Thus, if we take a gamma \( \Gamma(g, h) \) prior distribution for \( \alpha \), the joint posterior density is
\[
\pi(\alpha, \theta|k, n) \propto \pi(k, n|\alpha, \theta) \pi(\theta|\alpha) \pi(\alpha) 
\propto \prod_{c=1}^{|A|} \alpha^{k_c} \frac{\Gamma(\alpha) \Gamma(\alpha + n_c)}{\Gamma(\alpha + n_c)} \prod_{c=1}^C \theta_n^{n_c + \alpha \ell_c - 1} \alpha^{g-1} e^{-h\alpha}
\propto \prod_{c=1}^{|A|} \Gamma(\alpha) \Gamma(\alpha + n_c) \prod_{c=1}^C \theta_n^{n_c + \alpha \ell_c - 1} \alpha^{g+(\bar{k}-1)|A|-1} e^{-h\alpha},
\]
where \( \bar{k} = \sum_{c=1}^{|A|} k_c/|A| \) be the mean cluster size over chromosomes with hits. Therefore the (conditional) posterior density for \( \theta \) is
\[
\pi(\theta|\alpha, k, n) \prod_{c=1}^C \theta_n^{n_c + \alpha \ell_c - 1},
\]
Identification of QTL Candidate Genes

that is, a \( \text{Dir}(n + \alpha \ell) \) distribution. Also the (conditional) posterior density for \( \alpha \) is
\[
\pi(\alpha | \beta, k, n) \propto \alpha^{G-1} e^{-H \alpha} \prod_{c=1}^{|A|} \frac{\Gamma(\alpha)}{\Gamma(\alpha + n_c)} ,
\]
where \( G = g + (k - 1)|A| \) and \( H = h - \sum_{c=1}^{C} \ell_c \log \theta_c \). Using the identity
\[
\frac{\Gamma(\alpha)}{\Gamma(\alpha + n_c)} = \frac{(\alpha + n_c) B(\alpha + 1, n_c)}{\alpha \Gamma(n_c)} ,
\]
where \( B(\cdot, \cdot) \) is the Beta function, we obtain
\[
(\text{B.1}) \quad \pi(\alpha | \beta, k, n) \propto \alpha^{G-1} e^{-H \alpha} \prod_{c=1}^{|A|} \alpha \Gamma(n_c) .
\]
As the Beta function has integral representation
\[
B(\alpha + 1, n_c) = \int_0^1 x^{\alpha} (1 - x)^{n_c - 1} dx_c
\]
it is clear that
\[
\pi(\alpha, \eta | \beta, k, n) \propto \alpha^{G-1} e^{-H \alpha} \prod_{c=1}^{|A|} \alpha \Gamma(n_c) \eta_c^\alpha (1 - \eta_c)^{n_c - 1} ,
\]
where \( \eta = (\eta_1, \eta_2, ..., \eta_{|A|})' \) are beta distributed auxiliary variables, has distribution (B.1) when marginalised over \( \eta \). Therefore, letting \( \bar{\eta}_A = (\prod_{c=1}^{|A|} \eta_c)^{1/|A|} \) be the geometric mean of the components of \( \eta \), we have
\[
(\text{B.2}) \quad \pi(\alpha | \eta, \beta, k, n) \propto \alpha^{G-1} \exp\{-(H - |A| \log \bar{\eta}_A) \alpha \} \prod_{c=1}^{|A|} (\alpha + n_c) .
\]
Now
\[
\prod_{c=1}^{|A|} (\alpha + n_c) = e_0(n) \alpha^{|A|} + e_1(n) \alpha^{|A|-1} + e_2(n) \alpha^{|A|-2} + \cdots + e_{|A|}(n)
\]
where
\[
e_0(n) = 1, \quad e_1(n) = \sum_{i=1}^{|A|} n_i, \quad e_2(n) = \sum_{1=i<j}^{|A|} n_i n_j, \quad \ldots, \quad e_{|A|}(n) = \prod_{i=1}^{|A|} n_i .
\]
Here the \( e_k(n) \) are elementary symmetric polynomials which may be calculated efficiently by using the Newton–Girard formula
\[
k e_k(n) = \sum_{i=1}^{k} (-1)^{i-1} e_{k-i}(n) S_k(n) \quad \text{where} \quad S_k(n) = \sum_{i=1}^{|A|} n_i^k .
\]
Substituting this power series expansion into (B.2) gives

$$\pi(\alpha|\eta, \theta, k, n) \propto \sum_{i=0}^{\vert A \vert} e_i(n) \alpha^{G + \vert A \vert - i - 1} \exp\{-(H - \vert A \vert \log \bar{\eta}_{g}) \alpha\} ,$$

which is a mixture of Gamma distributions, that is,

$$\alpha|\eta, \theta, k, n \sim \sum_{i=0}^{\vert A \vert} p_i \Gamma(\alpha; G + \vert A \vert - i, H - \vert A \vert \log \bar{\eta}_{g})$$

with mixture proportions

$$p_i = \frac{e_i(n) \Gamma(G + \vert A \vert - i)}{\sum_{j=0}^{\vert A \vert} e_j(n) \Gamma(G + \vert A \vert - j) (H - \vert A \vert \log \bar{\eta}_{g})^{j-i}} , \quad i = 0, 1, ..., \vert A \vert .$$

Finally, for $c = 1, 2, ..., \vert A \vert$, 

$$\eta_c|\alpha, k, n \sim Beta(\alpha + 1, n_c) , \quad \text{independently} .$$

ACKNOWLEDGMENTS

This work was conducted as part of the ComparaGRID project and funded by the UK Biotechnology and Biological Sciences Research Council grant number BBS/B/17158.

REFERENCES


