REVSTAT – Statistical Journal
Volume 9, Number 1, March 2011, 1–17

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 (μ) and variances (σ^2) and with these parameters $\phi = (\mu, \sigma^2)$ resulting from a Dirichlet process with a particular base distribution. Let θ_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 $c_i \in \{1, ..., C\}$,

$$\begin{aligned}
\underline{\theta} = (\theta_1, ..., \theta_C)|\alpha &\sim Dir(\alpha \underline{\ell}) \ , \\
Y_i, c_i|\phi_{ic_i}, \theta_{c_i} &\sim N(\mu_{ic_i}, \sigma^2_{ic_i}) \times \text{Bern}(\theta_{c_i}) \ , \\
\phi_{ic_i}|G_{c_i} &\sim G_{c_i} \ , \\
G_{c_i}|\alpha &\sim DP(\alpha, G_{0c_i}) \ , \\
G_{0c_i} &= U(0, L_{c_i}) \times \text{Inv}\,\Gamma(a, L^2_{c_i} b) \ ,
\end{aligned}$$

where $DP(\alpha, G_0)$ denotes a Dirichlet process with base measure G_0 and concentration parameter α, and $\underline{\ell}$ is the normalized form of $\underline{L}$, 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 α. 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

$$\pi(y, k|\mathcal{D}) \simeq \frac{1}{Tn} \sum_{t=1}^{T} \sum_{i=1}^{n} \pi\big(y, k|\phi_{ik}^t, \theta_k^t\big) \ , \tag{2.1}$$

where $(\phi_{ik}^t, \theta_k^t)$, $t = 1, ..., T$, 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 (σ^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 α by taking $g = 4$ and $h = 2$, that is, $E(\alpha) = 2$ and $SD(\alpha) = 1$.

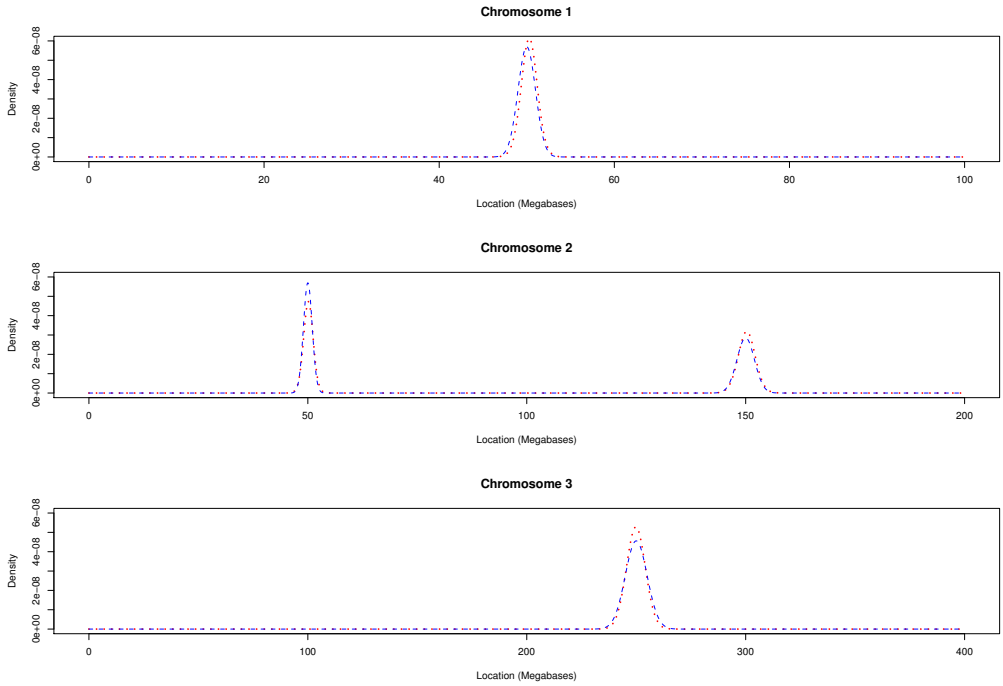

Figure 1: The theoretical distribution of hits along the chromosomes (dashed lines) and the Bayesian density estimate obtained from the simulated data (dotted line).

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 $100\,K$ iterations was required to achieve convergence and we then ran the chain for a further $100\,K$ iterations, thinning the output by taking every 100^{th} iterate. This gave a posterior sample of size $1\,K$ 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.

Table 1: Probability of a hit being located on each (simulated) chromosome (to 3 *d.p.*).

Chromosome	Probability
1	0.142
2	0.285
3	0.573

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.

Marker name	Associated gene	Gene symbol	Sequence length
REN97351	Hemoglobin Beta	HBB	248
RH69634	Hemoglobin Beta	HBB	141
PMC115301P1	Hemoglobin Beta	HBB	136
GDB:178694	Hemoglobin Beta	HBB	300
HBB	Hemoglobin Gamma	HBG	171
PMC86017P3	Hemoglobin Gamma	HBG	267
PMC21968P1	Hemoglobin Epsilon	HBE	989
Hba-a1	Hemoglobin Alpha	HBA	188
AW312144	Hemoglobin Alpha	HBA	327
CB603723	Hemoglobin Zeta	HBZ	312
BE749596	Hemoglobin Theta 1	HBQ	277
AW428039.1	Hemoglobin Mu	HBM	193

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 $100\,K$ iterations was required to achieve convergence.

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

BLAST parameter	argument	value
Expectation Value	−e	0.1
Gap Cost	−G	5
Gap Extension Cost	−E	2
Nucleotide Mismatch Cost	−q	3
Nucleotide Match reward	−r	2

We then ran the chain for a further $100\,K$ iterations, thinning the output by taking every 100^{th} iterate, to obtain a posterior sample of size $1\,K$ 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.*).

Chromosome	Probability
1	0.032
3	0.022
5	0.016
6	0.005
7	0.011
9	0.032
11	0.620
12	0.016
13	0.011
14	0.005
15	0.005
16	0.161
17	0.027
19	0.021
20	0.016
2, 4, 8, 10, 18, 21, 22, X, Y	$\simeq 0$

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.

HDR level	Chromosome	Intervals	Number of candidate genes
50%	11	4.59 Mb – 5.77 Mb	95
	16	0.00 Mb – 0.36 Mb	27
60%	11	4.37 Mb – 5.98 Mb	104
	16	0.00 Mb – 0.49 Mb	33
75%	11	3.88 Mb – 6.47 Mb	131
	11	54.52 Mb – 58.11 Mb	121
	16	0.00 Mb – 0.78 Mb	59
	9	124.28 Mb – 124.53 Mb	13
	17	2.51 Mb – 3.72 Mb	35
	20	61.14 Mb – 61.92 Mb	30

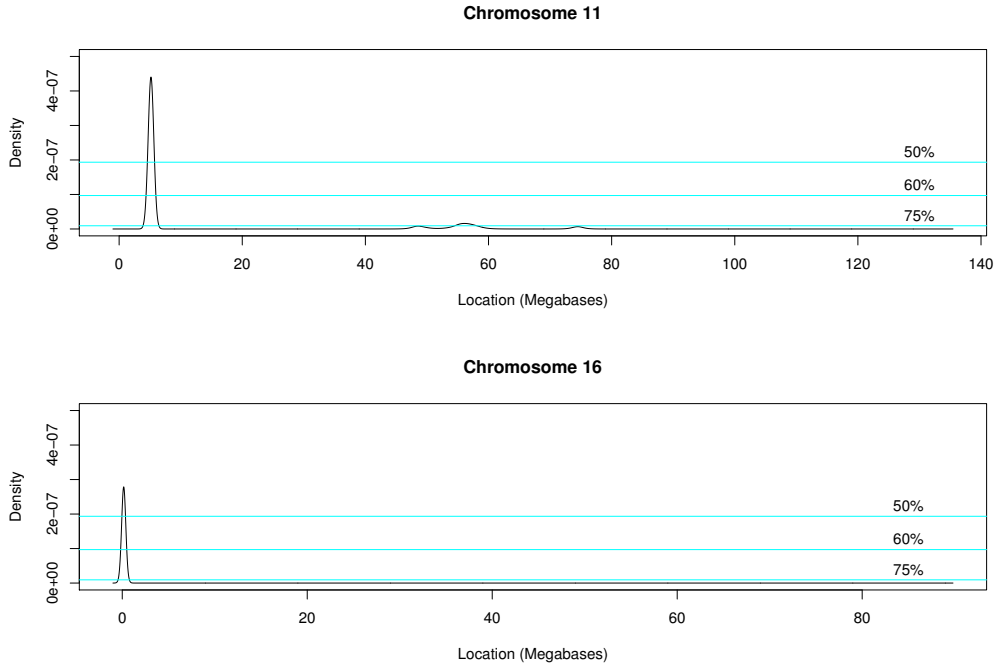

Figure 2: Figure showing the Bayesian density estimate of BLAST hits across chromosomes 11 and 16 of the Human genome.

If we look in detail at the Human genome, its Hemoglobin genes are located in two clusters on chromosomes 11 and 16, with the β-globin cluster spanning an interval of roughly 5.20–5.25 Mb on chromosome 11 and the α-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]).

Ensembl Gene ID	Gene name
ENSG00000220481	Z84812.3
ENSG00000181404	WASH4P
ENSG00000219509	Z84723.2
ENSG00000185203	Z84723.1
ENSG00000161980	POLR3K
ENSG00000161981	C16orf33
ENSG00000007384	RHBDF1
ENSG00000103152	MPG
ENSG00000103148	C16orf35
ENSG00000130656	**HBZ**
ENSG00000206178	*Z84721.1*
ENSG00000206177	**HBM**
ENSG00000218072	*Z84721.4*
ENSG00000188536	**HBA2**
ENSG00000206172	**HBA1**
ENSG00000207243	Y_RNA
ENSG00000086506	**HBQ1**
ENSG00000007392	LUC7L
ENSG00000206168	Z69890.1
ENSG00000167930	ITFG3
ENSG00000215289	AC004754.1
ENSG00000076344	RGS11
ENSG00000206156	ARHGDIG
ENSG00000185615	PDIA2
ENSG00000103126	AXIN1
ENSG00000086504	MRPL28
ENSG00000129925	TMEM8
ENSG00000216963	Z97634.3
ENSG00000103200	NME4
ENSG00000103202	DECR2
ENSG00000090565	RAB11FIP3
ENSG00000201034	Y_RNA
ENSG00000217816	RP1-196A12.1

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

known functional Hemoglobin genes in the human β-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 α-globin and β-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 $\phi_{ic_i} = \left(\mu_{ic_i}, \sigma^2_{ic_i}\right)$, $i = 1, ..., n$, and the parameters $(\alpha, \underline{\theta})$. 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 $\phi'_{ic_i}=\{\phi_{jc_i}\colon j\neq i\}$, the conditional prior density for $\phi_{ic_i}|\phi'_{ic_i}$ is

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

where g_{0c_i} is the probability density corresponding to the distribution G_{0c_i}, n_{c_i} 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

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

where

$$\begin{aligned}
q_{ij} \;&=\; \kappa\,\pi(y_i,c_i|\phi_{jc_i},\theta_{c_i})\;,\\
q_{i0} \;&=\; \kappa\,\alpha\int\pi(y_i,c_i|\phi_{ic_i},\theta_{c_i})\,g_{0c_i}(\phi_{ic_i})\,d\phi_{ic_i}\;,\\
g_{ic_i}(\phi_{ic_i}) \;&=\; \pi(y_i,c_i|\phi_{ic_i})\,g_{0c_i}(\phi_{ic_i})\Big/\int\pi(y_i,c_i|\phi_{ic_i})\,g_{0c_i}(\phi_{ic_i})\,d\phi_{ic_i}
\end{aligned}$$

and κ 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}, $G_{0c_i}=U(0,L_{c_i})\times \mathrm{Inv}\,\Gamma(a,L^2_{c_i}b)$, 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 $\Psi_a(\cdot)$ the t_a distribution function. Note that if $X\sim t_a$ then $b+\sqrt{c}\,X\sim St(a,b,c)$. Also, to simplify notation, we write $\tau=\sigma^2$. Then

$$q_{ij} \;=\; \kappa\,\pi(y_i,c_i|\phi_{ic_i},\theta_{c_i}) \;=\; \kappa\,\theta_{c_i}\phi(y_i|\mu_{ic_i},\tau_{ic_i})$$

and

$$\begin{aligned} q_{i0} &= \kappa\,\alpha \int \pi(y_i, c_i|\phi_{ic_i}, \theta_{c_i})\, g_{0c_i}(\phi_i)\, d\phi_{ic_i} \\ &= \kappa\,\alpha \int_0^{L_{c_i}} \int_0^{\infty} \theta_{c_i}\, \phi(y_i|\mu_{ic_i}, \tau_{ic_i}) \times \frac{1}{L_{c_i}} \times \frac{(L_{c_i}^2 b)^a\, \tau_{ic_i}^{-a-1}\, e^{-L_{c_i}^2 b/\tau_{ic_i}}}{\Gamma(a)}\, d\tau_{ic_i}\, d\mu_{ic_i} \\ &= \frac{\kappa\,\alpha\,\theta_{c_i}}{L_{c_i}} \int_0^{L_{c_i}} \psi_{2a}(\mu|y_i, L_{c_i}^2 b/a)\, d\mu \\ &= \frac{\kappa\,\alpha\,\theta_{c_i}}{L_{c_i}} \left\{ \Psi_{2a}\!\left(\frac{1 - y_i/L_{c_i}}{\sqrt{b/a}}\right) - \Psi_{2a}\!\left(-\frac{y_i/L_{c_i}}{\sqrt{b/a}}\right) \right\}. \end{aligned}$$

Also, for $0 \le \mu_{ic_i} \le L_{c_i},\ \tau_{ic_i} > 0$

$$\begin{aligned} g_{ic_i}(\phi_{ic_i}) &= \frac{\pi(y_i, c_i|\phi_i, \theta_{c_i})\, g_0(\phi_{ic_i})}{\displaystyle\int \pi(y_i, c_i|\phi_{ic_i}, \theta_{c_i})\, g_0(\phi_{ic_i})\, d\phi_{ic_i}} \\ &= \frac{\phi(y_i|\mu_{ic_i}, \tau_{ic_i}) \times (L_{c_i}^2 b)^a\, \tau_{ic_i}^{-a-1}\, e^{-L_{c_i}^2 b/\tau_{ic_i}} \big/ \Gamma(a)}{\Psi_{2a}\!\left(\frac{1-y_i/L_{c_i}}{\sqrt{b/a}}\right) - \Psi_{2a}\!\left(-\frac{y_i/L_{c_i}}{\sqrt{b/a}}\right)} \\ &= \frac{(L_{c_i}^2 b)^a\, \tau_{ic_i}^{-a-3/2}}{\sqrt{2\pi}\ \Gamma(a) \left\{ \Psi_{2a}\!\left(\frac{1-y_i/L_{c_i}}{\sqrt{b/a}}\right) - \Psi_{2a}\!\left(-\frac{y_i/L_{c_i}}{\sqrt{b/a}}\right) \right\}} \\ &\quad \times\ \exp\left\{ -\left(L_{c_i}^2 b + \frac{(y_i - \mu_{ic_i})^2}{2} \right) \Big/ \tau_{ic_i} \right\}. \end{aligned}$$

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

$$\mu_{ic_i} \sim St\big(2a, y_i, L_{c_i}^2 b/a\big)\ , \qquad 0 \le \mu_{ic_i} \le L_{c_i}$$

and

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

B. The remaining parameters

Here we derive the posterior conditional distributions for α and $\underline{\theta}$. 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,...,C). Then the probability function for the number of clusters on chromosome c is

$$\pi(k_c|n_c,\alpha,\underline{\theta}) \propto \begin{cases} \alpha^{k_c}\dfrac{\Gamma(\alpha)}{\Gamma(\alpha+n_c)}, & k_c=1,2,...,n_c\,, \quad \text{if } n_c>0\,, \\ 1, & k_c=0\,, \quad \text{if } n_c=0\,. \end{cases}$$

Let $\underline{k}=(k_1,k_2,...,k_C)$. As we have independent Dirichlet processes for each chromosome, $k_c|n_c,\alpha$ are independent for $c=1,2,...,C$ and so

$$\pi(\underline{k}|\underline{n},\alpha,\underline{\theta}) \propto \prod_{c=1}^{C}\alpha^{k_c}\frac{\Gamma(\alpha)}{\Gamma(\alpha+n_c)}\,,$$

for $k_c=1,2,...,n_c$ if $n_c>0$ or $k_c=0$ if $n_c=0$ ($c=1,2,...,C$). This can be simplified slightly by letting $A=\{c\colon n_c>0\}$ with size $|A|$, and renumbering the chromosomes so that $n_c>0$ for $c=1,2,...,|A|$ and $n_c=0$ for $c=|A|+1,|A|+2,...,C$, giving

$$\pi(\underline{k}|\underline{n},\alpha,\underline{\theta}) \propto \prod_{c=1}^{|A|}\alpha^{k_c}\frac{\Gamma(\alpha)}{\Gamma(\alpha+n_c)}\,.$$

The probability function for the number of hits on each chromosome has a multinomial distribution, with

$$\pi(\underline{n}|\alpha,\underline{\theta}) \propto \prod_{c=1}^{C}\theta_c^{n_c}\,,$$

and so the likelihood function for $(\alpha,\underline{\theta})$ is

$$\pi(\underline{k},\underline{n}|\alpha,\underline{\theta}) = \pi(\underline{k}|\underline{n},\alpha,\underline{\theta})\,\pi(\underline{n}|\alpha,\underline{\theta}) \propto \prod_{c=1}^{|A|}\alpha^{k_c}\frac{\Gamma(\alpha)}{\Gamma(\alpha+n_c)}\times\prod_{c=1}^{C}\theta_c^{n_c}\,.$$

Thus, if we take a gamma $\Gamma(g,h)$ prior distribution for α, the joint posterior density is

$$\begin{aligned}\pi(\alpha,\underline{\theta}|\underline{k},\underline{n}) &\propto \pi(\underline{k},\underline{n}|\alpha,\underline{\theta})\,\pi(\underline{\theta}|\alpha)\,\pi(\alpha) \\ &\propto \prod_{c=1}^{|A|}\alpha^{k_c}\frac{\Gamma(\alpha)}{\Gamma(\alpha+n_c)}\times\prod_{c=1}^{C}\theta_c^{n_c+\alpha\ell_c-1}\times\alpha^{g-1}e^{-h\alpha} \\ &\propto \prod_{c=1}^{|A|}\frac{\Gamma(\alpha)}{\Gamma(\alpha+n_c)}\times\prod_{c=1}^{C}\theta_c^{n_c+\alpha\ell_c-1}\times\alpha^{g+(\bar{k}-1)|A|-1}e^{-h\alpha}\,,\end{aligned}$$

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 $\underline{\theta}$ is

$$\pi(\underline{\theta}|\alpha,\underline{k},\underline{n}) \prod_{c=1}^{C}\theta_c^{n_c+\alpha\ell_c-1}\,,$$

that is, a $Dir(\underline{n}+\alpha\underline{\ell})$ distribution. Also the (conditional) posterior density for α is

$$\pi(\alpha|\underline{\theta},\underline{k},\underline{n}) \;\propto\; \alpha^{G-1}e^{-H\alpha}\prod_{c=1}^{|A|}\frac{\Gamma(\alpha)}{\Gamma(\alpha+n_c)}\;,$$

where $G = g + (\bar{k}-1)\,|A|$ 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

$$\pi(\alpha|\underline{\theta},\underline{k},\underline{n}) \;\propto\; \alpha^{G-1}e^{-H\alpha}\prod_{c=1}^{|A|}(\alpha+n_c)\,B(\alpha+1,n_c)\;. \tag{B.1}$$

As the Beta function has integral representation

$$B(\alpha+1,n_c) \;=\; \int_0^1 x_c^{\alpha}(1-x_c)^{n_c-1}\,dx_c$$

it is clear that

$$\pi(\alpha,\underline{\eta}|\underline{\theta},\underline{k},\underline{n}) \;\propto\; \alpha^{G-1}e^{-H\alpha}\prod_{c=1}^{|A|}(\alpha+n_c)\,\eta_c^{\alpha}(1-\eta_c)^{n_c-1}\;,$$

where $\underline{\eta} = (\eta_1,\eta_2,...,\eta_{|A|})'$ are beta distributed auxiliary variables, has distribution (B.1) when marginalised over $\underline{\eta}$. Therefore, letting $\bar{\eta}_g = \big(\prod_{c=1}^{|A|}\eta_c\big)^{1/|A|}$ be the geometric mean of the components of $\underline{\eta}$, we have

$$\pi(\alpha|\underline{\eta},\underline{\theta},\underline{k},\underline{n}) \;\propto\; \alpha^{G-1}\exp\big\{-(H-|A|\log\bar{\eta}_g)\,\alpha\big\}\prod_{c=1}^{|A|}(\alpha+n_c)\;. \tag{B.2}$$

Now

$$\prod_{c=1}^{|A|}(\alpha+n_c) \;=\; e_0(\underline{n})\,\alpha^{|A|} + e_1(\underline{n})\,\alpha^{|A|-1} + e_2(\underline{n})\,\alpha^{|A|-2} + \cdots + e_{|A|}(\underline{n})$$

where

$$e_0(\underline{n}) = 1\,,\quad e_1(\underline{n}) = \sum_{i=1}^{|A|} n_i\,,\quad e_2(\underline{n}) = \sum_{1=i<j}^{|A|} n_i\,n_j\,,\quad ...,\quad e_{|A|}(\underline{n}) = \prod_{i=1}^{|A|} n_i\;.$$

Here the $e_k(\underline{n})$ are elementary symmetric polynomials which may be calculated efficiently by using the Newton–Girard formula

$$k\,e_k(\underline{n}) = \sum_{i=1}^{k}(-1)^{i-1}e_{k-i}(\underline{n})\,S_k(\underline{n}) \qquad \text{where}\;\; S_k(\underline{n}) = \sum_{i=1}^{|A|} n_i^k\;.$$

Substituting this power series expansion into (B.2) gives

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

which is a mixture of Gamma distributions, that is,

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

with mixture proportions

$$p_i = \frac{e_i(\underline{n})\,\Gamma(G+|A|-i)}{\sum_{j=0}^{|A|} e_j(\underline{n})\,\Gamma(G+|A|-j)\,(H-|A|\log\bar{\eta}_g)^{j-i}}\ , \qquad i=0,1,...,|A|\ .$$

Finally, for $c=1,2,...,|A|$,

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

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

[1] Altschul, S.F.; Gish, W.; Miller, W.; Myers, E.W. and Lipman, D.J. (1990). Basic local alignment search tool, *Journal of Molecular Biology*, **215**(3), 403–410.

[2] Antoniak, C.E. (1974). Mixtures of Dirichlet processes with applications to Bayesian nonparametric problems, *The Annals of Statistics*, **2**(6), 1152–1174.

[3] Escobar, M.D. and West, M. (1995). Bayesian density estimation and inference using mixtures, *Journal of the American Statistical Society*, **90**(430), 577–588.

[4] Ferguson, T. (1973). A Bayesian analysis of some nonparametric problems, *The Annals of Statistics*, **1**(2), 209–230.

[5] Gelman, A. and Rubin, D.B. (1992). Inference from iterative simulation using multiple sequences, *Statistical Science*, **7**(4), 457–511.

[6] Givry, S.; Bouchez, M.; Chabrier, F.; Milan, S. and Schiex, T. (2005). Cartha-GENE: multipopulation integrated genetic and radiation hybrid mapping, *Bioinformatics*, **21**(8), 1703–1704.

[7] Heidelberger, P. and Welch, P. (1982). Simulation run length control in the presence of an initial transient, *Operations Research*, **31**(6), 1109–1144.

[8] Higgs, D.R.; Vickers, M.A. and Wilkie, A.O. (1989). A review of the molecular genetics of the human alpha-globin gene cluster, *Blood*, **73**(5), 1081–1104.

[9] Hubbard, T.J.P.; Aken, B.L.; Beal1, K.; Ballester1, B.; Caccamo, M.; Chen, Y.; Clarke, L.; Coates, G.; Cunningham, F.; Cutts, T.; Down, T.; Dyer, S.C.; Fitzgerald, S.; Fernandez-Banet, J.; Graf, S.; Haider, S.; Hammond, M.; Herrero, J.; Holland, R.; Howe, K.; Howe, K.; Johnson, N.; Kahari, A.; Keefe, D.; Kokocinski, F.; Kulesha, E.; Lawson, D.; Longden, I.; Melsopp, C.; Megy, K.; Meidl, P.; Overduin, B.; Parker, A.; Prlic, A.; Rice, S.; Rios, D.; Schuster, M.; Sealy, I.; Severin, J.; Slater, G.; Smedley, D.; Spudich, G.; Trevanion, S.; Vilella, A.; Vogel, J.; White, S.; Wood, M.; Cox, T.; Curwen, V.; Durbin, R.; Fernandez-Suarez, X.M.; Flicek, P.; Kasprzyk, A.; Proctor, G.; Searle, S.; Smith, J.; Ureta-Vidal, A. and Birney, E. (2007). Ensembl 2007, *Nucleic Acids Res.*, **35** (Database issue), 610–617.

[10] Hyndman, R.J. (1996). Computing and graphing highest density regions, *The American Statistician*, **50**(2), 120–126.

[11] Jow, H.; Bhattacharjee, M.; Boys, R.J. and Wilkinson, D.J. (2010). The integration of genetic maps using Bayesian inference, *Journal of Computational Biology*, **17**, 825–840.

[12] Lander, E.R. and Botstein, D. (1989). Mapping mendialian factors underlying quantitative traits using RFLP linkage maps, *Genetics*, **121**, 185–199.

[13] Liao, W.; Collins, A.; Hobbs, M.; Khatkar, M.S.; Luo, J. and Nicholas, F.W. (2007). A comparative location database (CompLDB): map integration within and between species, *Mammalian Genome*, **18**(5), 287–299.

[14] Morton, N.E.; Collins, A.; Lawrence, S. and Shields, D.C. (1992). Algorithms for a location database, *Annals of Human Genetics*, **56**, 223–232.

[15] Pruitt, K.D.; Tatusova, T. and Maglott, D.R. (2005). Ncbi reference sequence (refseq): a curated non-redundant sequence database of genomes, transcripts and proteins, *Nucleic Acids Research*, **33**, 501–504.

[16] Schiex, T. and Gaspin, C. (1997). *Carthagene: Constructing and joining maximum likelihood genetic maps*. In "Proceedings of the Fifth International Conference on Intelligent Systems for Molecular Biology", pp. 258–267.

[17] Silverman, B.W. (1986). *Density Estimation for Statistics and Data Analysis*, Chapman and Hall.

[18] Stam, P. (1993). Construction of integrated genetic linkage maps by means of a new computer package: Joinmap, *The Plant Journal*, **3**(5), 739–744.

[19] Stassen, H.H. and Scharfetter, C. (2000). Integration of genetic maps by polynomial transformations, *American Journal of Medical Genetics (Neuropsychiatric Genetics)*, **96**(1), 108–113.