UNIVERSIDAD SANTO TOMAS
PRIMER CLAUSTRO UNIVERSITARIO DE COLOMBIA

Comunicaciones en Estadística
Junio 2013, Vol. 6, No. 1

Cuantificación de variantes genéticas utilizando modelos jerárquicos bayesianos

Quantification of Genetic Variants using Bayesian Hierarchical Models

Jorge Iván Vélez[a]
jorgeivanvelez@gmail.com

Jairo Arturo Ángel[b]
jairoarturoangel@gmail.com

Juan Carlos Correa[c]
jccorrea@unal.edu.co

Resumen

En biología molecular los estudios funcionales son útiles para la caracterización de variantes o mutaciones en el genoma humano vía experimentación con modelos animales (embriones de peces, por ejemplo). Estos experimentos consisten en modificar genéticamente dichos embriones inyectándolos con moléculas de ácido ribonucleico mensajero (mRNA, en inglés) y se caracterizan por ser destructivos, tomar mucho tiempo y generar poca información. En este trabajo se propone e ilustra, con datos reales, una metodología para el análisis estadístico de este tipo de experimentos utilizando un enfoque bayesiano. Los resultados obtenidos con esta metodología concuerdan con lo observado a nivel molecular.

***Palabras clave*:** genética, estudios funcionales, estadística bayesiana, modelos jerárquicos, muestreador de Gibbs.

Abstract

In molecular biology, functional studies play an important role in the characterization of variants or mutations in the human genome by experimenting with animal models (e.g., fish embryos). These experiments, which consist in genetically modifying the embryos by injecting mRNA, are characterized by being destructive, time consuming and generate few information. We propose and illustrate, with

[a]Translational Genomics Group, Genome Biology Department, John Curtin School of Medical Research, The Australian National University, Canberra, ACT, Australia. Grupo de Neurociencias de Antioquia, Universidad de Antioquia, Colombia. Grupo de Investigación en Estadística, Universidad Nacional de Colombia, sede Medellín.

[b]Profesor, Centro de Ciencias Básicas, Universidad Pontificia Bolivariana, Montería, Colombia.

[c]Grupo de Investigación en Estadística, Universidad Nacional de Colombia, sede Medellín. Profesor Asociado, Escuela de Estadística, Universidad Nacional de Colombia, sede Medellín.

real data, a bayesian methodology for the statistical analysis of such experiments. This methodology provides comparable results to those observed at the molecular level.

***Key words*:** Genetics, Functional Studies, Bayesian Statistics, Hierarchical Models, Gibbs Sampler.

1. Introducción

Una mutación se define, en términos generales, como un cambio repentino y espontáneo en la secuencia del genoma de un organismo; en términos estadísticos, esto es equivalente a eventos raros (Roessler et al. 2012). Por esta razón, la determinación de si una mutación específica tiene o no efecto en la secuencia del genoma, altera el producto de un gen o interfiere en el funcionamiento de dicho gen, es de gran interés. Aquellas mutaciones por las cuales el producto del gen se hace menor o este tiene poca o ninguna función, son denominadas mutaciones de *pérdida de funcionalidad* (LOF, por sus siglas en inglés); aquellas en las que el producto del gen adquiere una nueva (pero anormal) función se denominan mutaciones de *ganancia de funcionalidad* (GOF, por sus siglas en inglés).

Con el propósito de entender las bases genéticas y moleculares de las enfermedades, y a su vez poder cuantificar la actividad funcional de las mutaciones de interés, una de las aproximaciones más comunes en este tipo de experimentos es utilizar modelos animales, por ejemplo, peces zebra (o *zebrafish* en inglés). Entre las ventajas que se obtienen al utilizar este modelo *zebrafish* como modelo animal se encuentran, entre otras, su equivalencia taxonómica (Chakraborty et al. 2009) y la homología genética con los humanos (Kari et al. 2007). En este tipo de experimentos, la determinación de si una mutación es LOF o GOF consiste, fundamentalmente, en construir una curva dosis-respuesta y cuantificar la actividad funcional de las variantes o mutaciones de interés (ampliado en las secciones 2.1 y 2.2).

Dentro del interés de los investigadores se encuentra determinar el número de embriones que tendrán determinada característica cuando se inyecta una dosis d^* a un grupo de n^* embriones, y estimar la dosis $\tilde{d}$ a la que los embriones presentan dicha característica en mayor proporción, por lo que, en ambos casos, la construcción de un modelo estadístico apropiado es fundamental. Sin embargo, por tratarse de experimentos destructivos[1], la cantidad de información que se genera es poca y como consecuencia los métodos estadísticos tradicionales (revisados en Ritz 2010) pueden difícilmente ser aplicados. Adicionalmente, puesto que la característica de interés puede variar de un embrión a otro o entre grupos de embriones (variabilidad intra e inter embrión), la incorporación de estas fuentes de variación en el modelo final también es deseable. Infortunadamente, en este tipo de experimentos

[1]Aunque no todos los n embriones inyectados con una dosis d mueren, biológicamente este experimento se considera destructivo, puesto que las inyecciones deben realizarse en un período de tiempo t específico durante la etapa de desarrollo de los embrión y cualquier otra dosis $\tilde{x}$ adicional cambiaría sus características genéticas.

la posibilidad de incluir estas fuentes de variación, utilizando métodos estadísticos tradicionales, es remota y la necesidad de nuevas estrategias de análisis es evidente.

En este documento presentamos una metodología bayesiana para el análisis estadístico de este tipo de experimentos, que constituye una alternativa viable y fácil de implementar en cualquier programa de análisis estadístico. Aunque la aplicación de métodos bayesianos en genética no es nueva (Shoemaker & Painter 1999, Blangero et al. 2005, Ding 2006, Stephens & Balding 2009, Yi et al. 2011, Innocenti et al. 2011, Gompert & Buerkle 2011), la cuantificación de la actividad funcional de variantes genéticas sí lo es. Como ilustración de esta metodología, consideramos la construcción de una curva dosis-respuesta y la cuantificación de una de las variantes genéticas presentadas en Domené et al. (2008), correspondientes a experimentos de rescate (*rescue*, por sus siglas en inglés) (Epstein & Shakes 1995, pp. 468-471) con grupos de embriones de *zebrafish*. La ventaja de esta metodología, como se mostrará más adelante, es que una vez se tienen las distribuciones conjunta y marginales *a posteriori* la inferencia es directa (Gelman et al. 2004, Kerman & Gelman 2006, Barrera & Correa 2008). Adicionalmente, los modelos bayesianos permiten incluir información experimental previa y/o el conocimiento de un grupo de expertos acerca de un parámetro de interés, utilizando técnicas de elicitación (Garthwaite et al. 2005). En experimentos biológicos, este tipo de información es muy valiosa.

2. Metodología

2.1. Curva dosis-respuesta

Considere un estudio funcional cuyo objetivo es cuantificar la actividad de K mutaciones genéticas (por simplicidad, asumiremos $K = 1$). Adicionalmente, durante la ejecución del estudio se realizan m ($m > 1$) experimentos independientes de tipo binomial en los que la característica de interés (también llamada fenotipo) es claramente diferenciable entre unidades muestrales (por ejemplo, embriones de *zebrafish*). Si Y_i es una variable aleatoria que representa el número de embriones con el fenotipo cuando n_i de ellos son inyectados con una dosis d_i de un compuesto particular, entonces $Y_i \sim \text{Binomial}(n_i, \theta_i)$, con $\theta_i \in (0, 1)$ el parámetro de la distribución ($i = 1, 2, \ldots, m$). La función de masa de probabilidad de Y_i es (Casella & Berger 2001):

$$p(Y_i = y_i \,|\, \theta_i) = \binom{n_i}{y_i} \theta_i^{y_i} (1 - \theta_i)^{n_i - y_i} \qquad y_i = 1, 2, \ldots, n_i, \;\; i = 1, 2, \ldots, m \quad (1)$$

Puesto que en muchos casos dosis d altas producen una mayor cantidad de embriones con el fenotipo de interés, es natural pensar que existe una función g tal que $\theta = g(d)$. Teniendo en cuenta la restricción de θ en el espacio parametral, la función $\text{logit}(\theta_i)$ definida como:

$$\log\left(\frac{\theta_i}{1-\theta_i}\right) = \alpha + \beta d_i \qquad i = 1, 2, \ldots, m \tag{2}$$

constituye una buena elección.

La implementación de una estrategia bayesiana requiere la existencia de una función de verosimilitud L y distribuciones *a priori* para los parámetros (Gelman et al. 2004, Berger 2010). Al despejar θ_i de (2) y reemplazarlo en (1), el logarítmo de L para m experimentos está dado por:

$$l = \log L(y_1, y_2, \ldots, y_m \,|\, \alpha, \beta, \mathbf{d}) \propto \sum_{i=1}^{m} \left\{ y_i(\alpha + \beta\, d_i) - n_i \log(1 + e^{\alpha+\beta\, d_i}) \right\} \tag{3}$$

Observe que en esta expresión todas las cantidades son conocidas excepto los parámeros α y β que relacionan linealmente a θ con d.

En el caso de dos parámetros es posible utilizar una rejilla de búsqueda en el rectángulo $[a_1, b_1] \times [a_2, b_2]$ sobre la cual se evalúa (3); este rectángulo constituye las distribuciones *a priori* para α y β (Gelman et al. 2004, Kerman & Gelman 2006). La selección de los valores (a_k, b_k), $k = 1, 2$, puede llevarse a cabo a partir de la información muestral, vía experimentación o basado en información previa (Gelman et al. 2004). Finalmente, la distribución conjunta *a posteriori* para (α, β) se obtiene como:

$$p(\alpha, \beta \,|\, \text{Datos}) \approx \frac{c_1}{\sum\limits_{[a_1,b_1]\times[a_2,b_2]} \sum c_1} \tag{4}$$

con $c_1 = e^{l - \text{máx}\{l\}}$. Similarmente, las distribuciones marginales *a posteriori* son

$$p(\alpha \,|\, \text{Datos}) \approx \frac{p(\alpha, \beta \,|\, \text{Datos})}{\sum\limits_{[a_2,b_2]} p(\alpha, \beta \,|\, \text{Datos})} \tag{5}$$

y

$$p(\beta \,|\, \text{Datos}) \approx \frac{p(\alpha, \beta \,|\, \text{Datos})}{\sum\limits_{[a_1,b_1]} p(\alpha, \beta \,|\, \text{Datos})} \tag{6}$$

Una vez se tienen las distribuciones conjuntas y marginales *a posteriori*, la inferencia es directa (Gelman et al. 2004, Kerman & Gelman 2006, Barrera & Correa 2008) y preguntas tales como (i) cuál es el número de embriones que tendrán el fenotipo cuando un grupo de n^* embriones se inyecta una dosis d^* y (ii) cuál es la dosis $\tilde{d}$ a la que los embriones presentan el fenotipo en mayor proporción, pueden responderse fácilmente (ver Sección 3.1).

2.2. Actividad funcional de una variante genética

Como se mencionó en la sección anterior, uno de los objetivos del análisis de curvas dosis-respuesta es la determinación de la dosis $\tilde{d}$ que genera el fenotipo de interés en mayor proporción. Consideremos ahora un conjunto de m ($m > 1$) experimentos similares a los de la sección 2.1, pero en los que grupos de n_i embriones ($i = 1, 2, \ldots, m$) son ahora inyectados con un compuesto *diferente* y se utiliza una dosis $\tilde{d}$ fija.

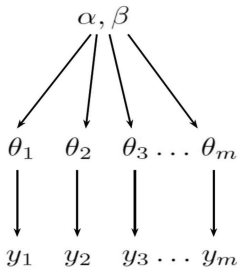

Figura 1: *Estructura del modelo jerárquico propuesto para el análisis bayesiano de la actividad funcional de una mutación. En el primer nivel, combinaciones de los hiperparámetros* (α, β) *generan el parámetro* θ_i*, responsable de la aparición del fenotipo de interés en* y_j *embriones de un grupo de* n_i *embriones que fueron inyectados* ($i = 1, 2, \ldots, m$)*. Fuente: elaboracion propia.*

Si Y_i es el número de embriones inyectados que presentan el *nuevo* fenotipo y n_i (fijo) es el número total de embriones inyectados, entonces $Y_i \,|\, \theta_i \sim \text{Binomial}(n_i, \theta_i)$, con $0 < \theta_i < 1$, $i = 1, 2, \ldots, m$. Por tratarse de una proporción, es razonable pensar que $\theta_i \,|\, \alpha, \beta \sim \text{Beta}(\alpha, \beta)$, con $\alpha, \beta > 0$ los hiperparámetros del modelo; la distribución conjunta de estos es $p(\alpha, \beta)$. De esta forma se tiene entonces que el vector de observaciones $\mathbf{y} = (y_1, y_2, \ldots, y_m)$ puede ser visto como una realización de una estructura jerárquica (vea Figura 1). Comparado con los métodos tradicionales (vea Ritz (2010)), este modelo jerárquico permite considerar la variación que existe de un embrión a otro y/o entre grupos de embriones (variabilidad intra e inter embrión) y no requiere grandes tamaños de muestra.

La distribución *a posteriori* del modelo completo está dada por:

$$\begin{aligned} p(\phi, \boldsymbol{\theta} \,|\, \mathbf{y}) &\propto p(\phi)\, p(\boldsymbol{\theta}|\phi)\, p(\mathbf{y}|\boldsymbol{\theta}) \\ &= p(\phi) \prod_{i=1}^{m} \frac{\Gamma(\alpha+\beta)}{\Gamma(\alpha)\Gamma(\beta)} \theta_i^{\alpha-1}(1-\theta_j)^{\beta-1} \prod_{i=1}^{m} \binom{n_i}{y_i} \theta_i^{y_i}(1-\theta_i)^{n_i-y_i} \\ &= p(\phi) \prod_{i=1}^{m} \text{Beta}(\theta_i|\phi)\ \text{Binomial}(y_i|n_i, \theta_i) \end{aligned} \tag{7}$$

donde $\phi = (\alpha, \beta)$ es el vector de hiperparámetros y $\boldsymbol{\theta} = (\theta_1, \theta_2, \ldots, \theta_m)$ es el vector de parámetros.

El muestreador de Gibbs (ver Casella & George (1992) para una introducción) hace parte de los algorítmos iterativos basados en cadenas de Markov (también denominadas *Markov chain Monte Carlo*, o MCMC) que permiten obtener muestras de la distribución *a posteriori* de un vector de parámetros $\boldsymbol{\Theta}$ de dimensión p cuando esta no tiene forma conocida o $p \to \infty$ (Gelman et al. 2004, Barrera & Correa 2008). Específicamente, el muestreador de Gibbs se utiliza cuando la distribución conjunta de $\boldsymbol{\Theta}$ es desconocida pero la distribución *condicional* de Θ_j es conocida ($j = 1, 2, \ldots, p$). Si definimos $\Theta_{-j}^{t-1} = (\Theta_1^{t-1}, \Theta_2^{t-1}, \ldots, \Theta_{j+1}^{t-1}, \ldots, \Theta_p^{t-1})$ entonces, en la iteración t del algorítmo, $\Theta_j^t \sim p(\Theta_j \,|\, \Theta_{-j}^{t-1}\,, \text{Datos})$, de tal forma que el muestreo de la distribución *a posteriori* es inmediato toda vez que se tengan valores iniciales para $\boldsymbol{\Theta}$ (Gelman et al. 2004, Sección 11.2). Dadas las características de (7), utilizaremos el muestreador de Gibbs para obtener muestras aleatorias de esta distribución.

Las distribuciones *a posteriori* condicionales de (7) son

$$\begin{aligned} p(\theta_i|\phi, \mathbf{y}) &\propto \theta_i^{\alpha-1}(1-\theta_i)^{\beta-1}\theta_i^{y_i}(1-\theta_i)^{n_i-y_i} \\ &= \text{Beta}(\alpha + y_i, n_i + \beta - y_i) \end{aligned} \tag{8}$$

y

$$p(\phi \,|\, \boldsymbol{\theta}, \mathbf{y}) \propto p(\phi) \prod_{i=1}^{m} \text{Beta}(\theta_i|\alpha, \beta) \tag{9}$$

Observe que las distribuciones (7) y (9) están en función de $p(\phi) = p(\alpha, \beta)$ y que a pesar de la estructura jerárquica del modelo, la distribución $p(\theta_i|\phi, \mathbf{y})$ tiene una forma (cerrada) conocida. En la práctica, esto último facilita enormemente la generación de muestras para θ_i ($i = 1, 2, \ldots, m$).

Teóricamente es posible utilizar diferentes distribuciones *a priori* para $p(\phi)$ que reflejen nuestro conocimiento del experimento (Berger 2010, Capítulo 3); la selección de cuál de estas distribuciones es la más adecuada es motivo de extensa investigación (Kass & Wasserman 1996, Casella & Berger 2001, Gelman et al. 2004, Berger 2010). En nuestro caso, las distribuciones *a priori* utilizadas para $p(\phi)$ fueron $p(\phi) \propto (\alpha + \beta)^{-5/2}$ y $p(\alpha, \beta) \propto 1$, esta última también conocida como la distribución *a priori* no informativa de Laplace. Sin embargo, la metodología aquí presentada no se limita a la escogencia de estas distribuciones *a priori*. Observe que $\alpha > 0$ y $\beta > 0$ puesto que corresponden a los parámetros de una distribución Beta.

Para la generación de muestras de la distribución *a posteriori* de $p(\phi)$ se definieron tres muestreadores de Gibbs con $B = 10000$ iteraciones[2]. En el primero, para la iteración t, $\phi^{(t)} \sim N_2\left(\phi^{(t-1)}, \Sigma_\phi\right)$; en el segundo, $\log \phi^{(t)} \sim N_2\left(\phi^{(t-1)}, \Sigma_\phi\right)$ y en el tercero, $\log \phi^{(t)} \sim N_2\left(\phi^{(t-1)}, \Sigma_\phi\right)$ ($t = 1, 2, \ldots, B$). La matriz de varianzas-covarianzas utilizada fue:

$$\Sigma_\phi = \begin{pmatrix} \sigma_\alpha & \sigma_{\alpha,\beta} \\ \sigma_{\alpha,\beta} & \sigma_\beta \end{pmatrix}$$

[2]Dada la complejidad de las distribuciones *a posteriori* conjuntas y marginales, el muestreador de Gibbs se utilizó en conjunto con *rejection sampling*.

con $(\sigma_\alpha, \sigma_{\alpha,\beta}, \sigma_\beta) = (2, 0, 2)$ para el primer muestreador, $(3, 0, 3)$ para el segundo y $(2, 1/3, 3)$ para el tercero. El vector de valores iniciales fue, en todos los casos, $\phi^{(0)} = (1, 1)$. La implementación de estos muestreadores en R (R. Core Team 2013) se encuentra disponible a petición del lector.

3. Aplicación

3.1. Curva dosis-respuesta

Como ilustración, se seleccionaron 5 experimentos tipo *rescue* en *zebrafish* en el que se inyectaron un total de 114 embriones con 1 ng de Hdl MO[3] y diferentes dosis de mRNA del alelo de referencia (*wild-type* o WT, en inglés) del gen *Sine oculis homeobox homolog 3* (*SIX3*). Este gen está ubicado en el cromosoma 2p21 y tiene una longitud de 999 pares de bases (Roessler et al. 2012); mutaciones en este gen han sido asociadas holoprosencefalia (HPE en inglés) (Wallis et al. 1999, Dubourg et al. 2004)[4]. Las duplas dosis/tamaño de grupo consideradas por experimento fueron 2.5/38, 5/36, 10/33, 25/5 y 50/2 y el fenotipo de interés correspondió a la presencia de ambos ojos en el embrión (obtener un embrión *normal*). En la Figura 2a se presenta la curva dosis-respuesta obtenida. Note que mientras en el grupo de 38 embriones la mitad de ellos son normales, la proporción aumenta considerablemente cuando se inyecta una dosis de 50 pg.

La región de evaluación de (4) se definió como $[-2, 2] \times [0, 1]$ gracias a experimentos previos que indicaban que a mayores dosis la proporción de embriones normales era mayor. En las figuras 2b y 2c presentamos la distribución conjunta *a posteriori* para α y β y sus contornos de probabilidad constante, respectivamente. Llama la atención que sin haber definido una estructura de correlación para α y β, los contornos de la distribución conjunta *a posteriori* indiquen que esta existe ($\hat{\rho} = -0.884$, valor-$p < 10^{-10}$).

Tabla 1: *Medidas para las distribuciones a posteriori de α y β en (2). Fuente: elaboracion propia.*

Parámetro	Moda	Media	Mediana	Desviación Estándar	95 %CI
α	-0.636	-0.655	-0.636	0.543	(-1.773, 0.227)
β	0.393	0.414	0.404	0.124	(0.192, 0.636)

[3]Hdl se refiere al gen *headless* (o simplemente *hdl*) en *zebrafish*. MO corresponde a *morpholino oligonucleotide*, moléculas utilizadas para modificar la expresión de un gen. Embriones inyectados solo con Hdl MO sufren alteraciones que dificultan el desarollo de los ojos y otras estructuras a nivel anterior. La coinyección de Hdl MO y WT *SIX3* en *zebrafish* permite *rescatar* los embriones, es decir, obtener embriones *normales* (Domené et al. 2008).

[4]Para mayor información se sugiere consultar el número especial del *American Journal of Medical Genetics* en `http://bit.ly/zVjwQL`.

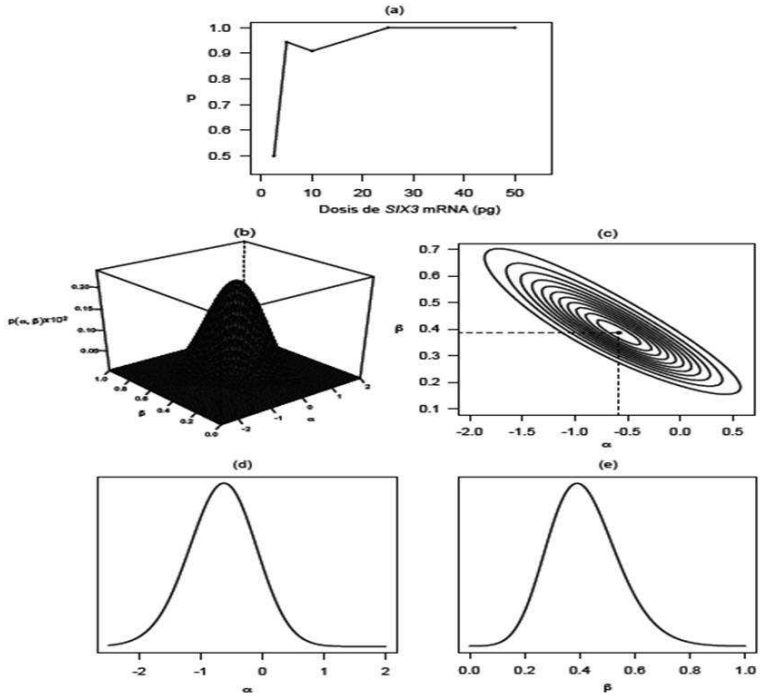

Figura 2: *(a) Curva dosis-respuesta en Domené et al. 2008; (b) distribución a posteriori para* (α, β)*; (c) contornos de (b). Distribuciones a posteriori de (d)* α *y (e)* β*. En (a) el eje y representa la proporción de embriones normales; en (c) la intersección de las líneas punteadas corresponden* $\hat{\alpha}_{MLE}$ *y* $\hat{\beta}_{MLE}$*. La medida pg corresponde a* 10^{-12} *g. Fuente: elaboracion propia.*

Las medidas de resumen para las distribuciones *a posteriori* de α y β se presentan en la Tabla 1 (vea figuras 2d y 2e). Nuestros resultados indican que por cada unidad que se incremente la dosis de WT *SIX3* mRNA, el *odds* de que el embrión sea normal incrementa de 1 a $e^{0.414} = 1.513$ (95 %CI = 1.211–1.967). Comparativamente, los estimadores de máxima verosimilitud (MLE) para α y β utilizando un modelo lineal generalizado (MLG, datos no presentados) son, respectivamente, $\hat{\alpha}_{\text{MLE}} = -0.582$ y $\hat{\beta}_{\text{MLE}} = 0.386$ (ver Figura 2c); el *odds* incrementa de 1 a $e^{0.386} = 1.471$ (95 %CI = 1.161–1.866).

A partir de las distribuciones *a posteriori* conjunta y marginales, se realizaron algunos análisis adicionales que incluyeron (*i*) la comparación del número de embriones normales reales con los *predichos* al utilizar el modelo bayesiano (Figura 3a), (*ii*) el cálculo de la distribución *predicitva* (Barrera & Correa 2008) del número de embriones normales obtenidos cuando n embriones son inyectados con una dosis d fija (Figura 3b), (*iii*) el cálculo del número promedio de embriones normales cuando n es fijo y se varía d (Figura 3c) y (*iv*) el número de embriones normales cuando se varían n y d simultáneamente (Figura 3d).

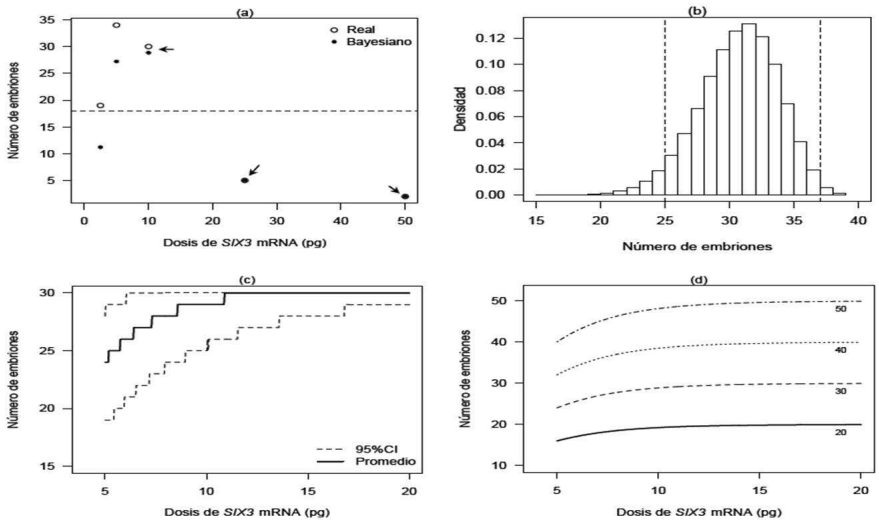

Figura 3: *(a) Análisis de shrinkage para el número de embriones como función de la dosis de SIX3 mRNA. (b) distribución a posteriori del número de embriones con el fenotipo cuando se inyectan 5 pg de SIX3 mRNA en un grupo de 39 embriones. En (c) y (d) se presentan el número promedio de embriones con el fenotipo cuando se inyectan grupos de 30 embriones con diferentes dosis, y se varía el tamaño del grupo y la dosis inyectada, respectivamente. En (c) las líneas punteadas representan el intervalo de confianza (CI) del 95 %. Las flechas en (a) muestran los experimentos en los que la distribución a posteriori proporciona resultados satisfactorios. Fuente: elaboracion propia.*

En el primer caso, el modelo bayesiano produce resultados satisfactorios para dosis superiores 5 pg (Figura 3a) y relativamente buenos para dosis más pequeñas. Una posible explicación de este comportamiento está relacionado con la alta variabilidad en la proporción de embriones normales (Figura 3a) y los tamaños de grupo. El ajuste con el MLG (datos no presentados) es más pobre, independiente de la dosis ($\text{MSE}_{\text{Bayesiano}} = 6.329 \times 10^{-3}$ vs. $\text{MSE}_{\text{MLG}} = 6.691 \times 10^{-3}$). Por otro lado, si se llevara a cabo un *nuevo* experimento con 39 embriones, una dosis $d = 5$ pg produciría $\hat{y}^{\text{pred}} = 31$ (95 %CI = 25–37) embriones normales (Figura 3b), es decir, el 79.5 % de estos. Observe que a pesar de que el tamaño de grupo es similar al utilizado en el experimento real para la misma dosis (36 vs. 39 embriones), las distribuciones *a priori* de α y β, en combinación con la verosimilitud de los datos, producen una proporción de embriones normales *a posteriori* ajustada (94.4 % vs. 79.5 %). Similarmente, nuevos experimentos con grupos de tamaño $n = 30$ y dosis d entre 5 pg y 20 pg (Figura 3b) sugieren que una dosis entre 12 pg y 15 pg sería suficiente para obtener embriones normales. Este resultado es consistente con el obtenido cuando el tamaño de los grupos y la dosis varían simultáneamente (Figura 3d).

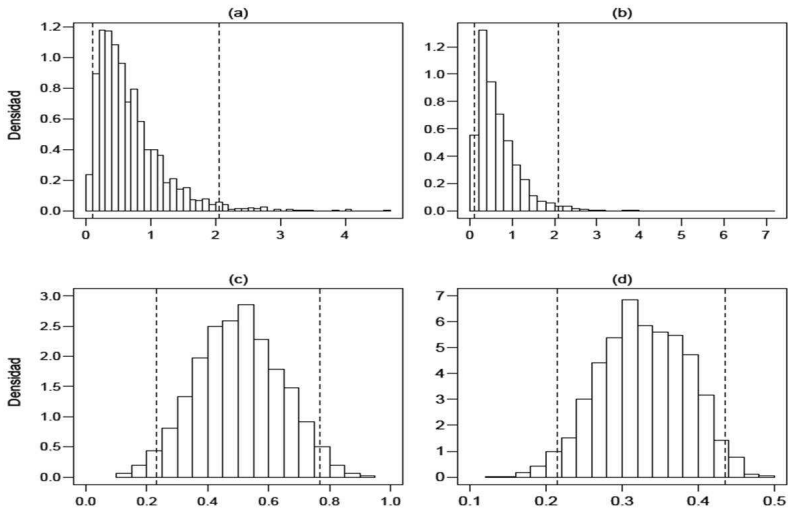

Figura 4: *Distribución a posteriori para (a) α, (b) β, (c) μ y (d) σ cuando la distribución a priori es $p(\alpha, \beta) = (\alpha, \beta)^{-5/2}$. Los parámetros μ y σ corresponden, respectivamente, a la media y desviación estándar de θ en (8). Las líneas verticales corresponden al intervalo de confianza del 95 %. Fuente: elaboracion propia.*

3.2. Análisis de la mutación 605C>T en *SIX3*

Dubourg et al. (2004) analizaron una cohorte de 200 pacientes con HPE y encontraron 7 nuevas mutaciones en el gen *SIX3*. Una de estas mutaciones es 605C>T, ubicada en el dominio Six de la proteina SIX3, y que seleccionaremos para ilustrar nuestro modelo jerárquico Bayesiano (ver Figura 1). Después de inyectar 114 embriones con una dosis de 50 pg de WT *SIX3* mRNA en 4 experimentos independientes, esta mutación, también conocida como T202I (cambio de **T**reonina a **I**soleucina en la posición 220 de la proteína SIX3) fue catalogada como LOF por Domené et al. (2008). Los pares de embriones normales/inyectados fueron 39/40, 3/23, 10/28 y 6/23, respectivamente.

El análisis de convergencia para los muestreadores de Gibbs (Sección 2.2) se llevó a cabo utilizando la prueba de Kwiatkowski-Phillips-Schmidt-Shin (KPSS) (ver Kwiatkowski et al. 1992 y Barrera & Correa 2008 para más información) implementada en la librería `tseries` (Trapletti & Hornik 2011) de R. Si la hipótesis nula es rechazada al utilizar la prueba KPSS, decimos que la cadena de Markov no ha alcanzado la distribución estacionaria (o simplemente *no converge*). Para ambas distribuciones *a priori* solo el muestreador 2 convergió, es decir, el valor-p de la prueba KPSS fue superior a un nivel de significancia $\alpha = 0.05$. En la Figura 4 se presentan, para $p(\alpha, \beta) = (\alpha, \beta)^{-5/2}$, las distribuciones *a posteriori* de α y β, así como de μ y σ, la media y desviación estándar de θ en (8), respectivamente[5].

[5]Los resultados *a posteriori* para $p(\alpha, \beta) \propto 1$ se encuentran disponibles a petición del lector.

Dado que $\theta \sim \text{Beta}(\alpha, \beta)$, $\mu = \alpha/(\alpha+\beta)$ y $\sigma = \sqrt{\alpha\beta/\{(\alpha+\beta)^2(\alpha+\beta+1)\}}$.

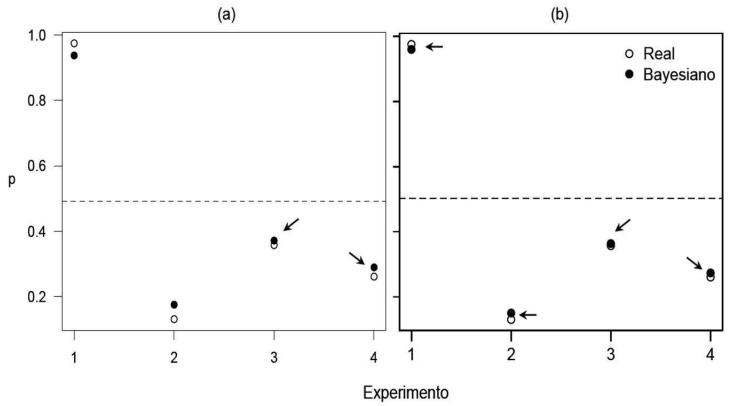

Figura 5: *Análisis de shrinkage para la proporción de embriones con el fenotipo cuando (a)* $p(\alpha) \propto 1$ *y (b)* $p(\alpha, \beta) = (\alpha+\beta)^{-5/2}$. *Las líneas horizontales corresponden a* $\hat{p}_{posterior} = 0.503$. *Las flechas resaltan las diferencias entre los valores reales de p y el valor a posteriori. Note que solo en (b) estas diferencias son despreciables. Fuente: elaboración propia.*

Tabla 2: *Medidas para las distribuciones a posteriori de* α, β, μ *y* σ.

Parámetro	Media	Mediana	95 %CI
α	0.676	0.542	(0.103, 2.053)
β	0.671	0.527	(0.108, 2.090)
μ	0.503	0.503	(0.233, 0.769)
σ	0.328	0.327	(0.215, 0.435)

Las medidas de resumen para las distribuciones *a posteriori* de α y β se presentan en la Tabla 2 (ver figuras 4a y 4b). De acuerdo con estas estimaciones, la proporción *a posteriori* de embriones con el fenotipo es $\hat{p} = 0.503$ (95 %CI = 0.233–0.769); esta estimación es $\hat{p} = 0.508$ (95 %CI = 0.414–0.603) utilizando la aproximación basada en la distribución normal (Casella & Berger 2001). Estos resultados confirman que esta mutación es LOF, puesto que el número de embriones normales que se obtienen corresponden a $\approx 50\,\%$ del que se obtendría con el alelo de referencia. El análisis de *shrinkage* (ver Figura 5) indica que con $p(\alpha, \beta) = (\alpha+\beta)^{-5/2}$ la inferencia *a posteriori* es mejor ($\text{MSE}_{\text{Bayesiano}} = 2.089 \times 10^{-5}$ vs. $\text{MSE}_{\text{MLG}} = 0.111$).

4. Conclusiones

Se ha propuesto una metodología bayesiana para la cuantificación de la actividad funcional de variantes genéticas en la que se maximiza la función de verosimilitud,

utilizando una rejilla de búsqueda y se utilizan distribuciones no conjugadas como distribuciones *a priori*. Como ilustración, se construyó la curva dosis-respuesta para el alelo de referencia del gen *SIX3* y se cuantificó la actividad funcional de la mutación 605C>T en el mismo gen a partir de información proveniente de experimentos con *zebrafish*, caracterizados por ser destructivos, costosos y generar poca información.

Como se mostró durante la construcción de la curva dosis-respuesta y la cuantificación de una de las mutaciones reportadas en Domené et al. (2008), el modelo bayesiano produce mejores resultados que el MLG clásido en términos del MSE, especialmente cuando se tienen pocos experimentos (muy común en biología experimental y molecular). Otros aspectos importantes de esta metodología que son de gran utilidad y aplicación en biología molecular son: (*i*) la posibilidad que existe de incluir información de expertos para determinar la distribución *a priori* de los parámetros de interés (e.g., utilización de técnicas de elicitación)(Garthwaite et al. 2005) y (*ii*) que puedan realizarse inferencias probabilísticas acerca del *verdadero* valor de los parámetros poblacionales o una función de estos (ver Garthwaite et al. (2013) para una amplia discusión) como los presentados en las figuras 3, 4 y 5. En experimentos de este tipo, y en especial en campos de investigación donde generar datos es tan costoso, es fundamental disponer de metodologías de análisis con estas características.

Posibles trabajos futuros incluyen la integración del modelo bayesiano utilizado para la construcción de la curva dosis-respuesta y el Modelo Jerárquico Bayesiano implementado para la cuantificación de experimentos independientes. De esta forma, se incorporaría información acerca del parámetro de interés, e.g., proporción de embriones con el fenotipo, en la cuantificación de la actividad funcional de una variante genética o mutación. La evaluación de otras distribuciones *a priori* para los parámetros constituye un área de trabajo adicional.

Agradecimientos

Los autores agradecen al Dr. Benjamin Feldman del Zebrafish Core, National Institute of Child Health and Human Development, Bethesda, MD, USA, por facilitar los datos presentados en la aplicación, así como los comentarios y sugerencias de un revisor anónimo que ayudó a mejorar sustancialmente este documento. JIV fue financiado parcialmente por The Eccles Scholarship in Medical Sciences, The Fenner Merit Scholarship y The Australian National University High Degree Research Scholarship. JIV agradece el apoyo del Dr. Mauricio Arcos-Burgos.

Recibido: 6 de enero de 2013
Aceptado: 28 de marzo de 2013

Referencias

Barrera, C. J. & Correa, J. C. (2008), 'Distribución predictiva bayesiana para modelos de pruebas de vida vía MCMC', *Revista Colombiana de Estadística* **31**(2), 145–155.

Berger, J. O. (2010), *Statistical Decision Theory and Bayesian Analysis*, 2 edn, Springer-Verlag.

Blangero, J., Goring, H. H. H., Kent, J. W. & Williams, J. T. (2005), 'Quantitative trait nucleotide analysis using bayesian model selection', *Human Biology* **77**(5), 541–559.

Casella, G. & Berger, R. (2001), *Statistical Inference*, Duxbury Press.

Casella, G. & George, E. I. (1992), 'Explaining the Gibbs Sampler', *The American Statistician* **46**(3), 167–174.

Chakraborty, C., Hsu, C., Wen, Z., Lin, C. S. & Agoramoorthy, G. (2009), 'Zebrafish: a complete animal model for in vivo drug discovery and development.', *Curr. Drug. Metab.* **10**(2), 116–24.

Ding, Y. (2006), 'Statistical and Bayesian approaches to RNA secondary structure prediction', *RNA* **12**(3), 323–31.

Domené, S., Roessler, E., El-Jaick, K. B., Snir, M., Brown, J. L., Vélez, J. I., Bale, S., Lacbawan, F., Muenke, M. & Feldman, B. (2008), 'Mutations in the human SIX3 gene in holoprosencephaly are loss of function', *Human Molecular Genetics* **17**(24), 3919–3929.

Dubourg, C., Lazaro, L., Pasquier, L., Bendavid, C., Blayau, M., Duff, F. L., Durou, M.-R., Odent, S. & David, V. (2004), 'Molecular screening of SHH, ZIC2, SIX3, and TGIF genes in patients with features of holoprosencephaly spectrum: Mutation review and genotype−phenotype correlations', *Human Mutation* **24**(1), 43–51.

Epstein, H. F. & Shakes, D. C. (1995), *Methods in Cell Biology*, Vol. 48, Academic Press, Inc.

Garthwaite, P. H., Kadane, J. B. & O'Hagan, A. (2005), 'Statistical methods for eliciting probability distributions', *Journal of the American Statistical Association* **100**(470), 680–701.

Garthwaite, P. H., Kadane, J. B. & O'Hagan, A. (2013), 'Experimental Biology: Sometimes Bayesian statistics are better', *Nature* **494**(35), 35.

Gelman, A., Carlin, J. B., Stern, H. S. & Rubin, D. B. (2004), *Bayesian Data Analysis*, 2 edn, Chapman & Hall/CRC, Washington, D.C.

Gompert, Z. & Buerkle, C. (2011), 'A hierarchical bayesian model for next-generation population genomics', *Genetics* **3**(187).

Innocenti, F., Cooper, G., Stanaway, I., Gamazon, E., Smith, J., Mirkov, S., Ramírez, J., Liu, W., Lin, Y., Moloney, C., Aldred, S., Trinklein, N., Schuetz, E., Nickerson, D., Thummel, K., Rieder, M., Rettie, A., Ratain, M., Cox, N. & Brown, C. (2011), 'Identification, replication, and functional fine-mapping of expression quantitative trait loci in primary human liver tissue', *PloS Genetics* **7**(5), 1–16.

Kari, G., Rodeck, U. & Dicker, A. (2007), 'Zebrafish: an emerging model system for human disease and drug discovery', *Clin. Pharmacol. Ther.* **82**(1), 70–80.

Kass, R. E. & Wasserman, L. (1996), 'The selection of prior distributions by formal rules', *Journal of the American Statistical Association* **91**(435), 1343–1370.

Kerman, J. & Gelman, A. (2006), 'Bayesian Data Analysis using R', *Rnews* **6**(1), 21–24.

Kwiatkowski, D., Phillips, P. C., Schmidt, P. & Shin, Y. (1992), 'Testing the null hypothesis of stationarity against the alternative of a unit root: How sure are we that economic time series have a unit root?', *Journal of Econometrics* **54**(1-3), 159–178.

R. Core Team (2013), *R: A Language and Environment for Statistical Computing*, R Foundation for Statistical Computing, Vienna, Austria. ISBN 3-900051-07-0.
*http://www.R-project.org/

Ritz, C. (2010), 'Toward a unified approach to dose-response modeling in ecotoxicology', *Environmental Toxicology and Chemistry* **29**(1), 220–229.

Roessler, E., Vélez, J. I., Zhou, N. & Muenke, M. (2012), 'Utilizing prospective sequence analysis of SHH, ZIC2, SIX3 and TGIF in holoprosencephaly probands to describe the parameters limiting the observed frequency of mutant gene×gene interactions.', *Mol. Genet. Metab.* **105**(4), 658–64.

Shoemaker, J. & Painter, I. (1999), 'Bayesian statistics in genetics: a guide for the uninitiated', *Bayesian Statistical Methods* **15**(9).

Stephens, M. & Balding, D. (2009), 'Bayesian statistical methods for genetic association studies', *Nat Rev Genet* **10**(10).

Trapletti, A. & Hornik, K. (2011), *tseries: Time Series Analysis and Computational Finance*. R package version 0.10-26.
*http://CRAN.R-project.org/package=tseries

Wallis, D. E., Roessler, E., Hehr, U., Nanni, L., Wiltshire, T., Richieri-Costa, A., Gillesse-Kaesbach, G., Zackai, E. H., Rommens, J. & Muenke, M. (1999), 'Mutations in the homeodomain of the human six3 gene cause holoprosencephaly', *Nat. Genet.* **22**(2), 196–198.

Yi, N., Liu, N., Zhi, D. & Li, J. (2011), 'Hierarchical generalized linear models for multiple groups of rare and common variants: jointly estimating group and individual–variant effects', *PloS Genetics* **7**(12).