

Capítulo V

Diversidad genética de poblaciones y detección de selección en la nueva era de la genómica

Andreia de Jesus Amaral Gomes Barbosa Fonseca

PhD. Centro de Investigación Interdisciplinaria en Sanidad Animal,
Facultad de Medicina Veterinaria, de la Universidad de Lisboa.



La selección natural

La selección natural puede darse con o sin presión ambiental; en condiciones ambientales estables, la selección natural mantendrá estable a la población. En un ambiente estable, si ocurre una nueva mutación que sea beneficiosa para la población, esa nueva mutación se mantendrá; con esto aumenta su frecuencia y la población evoluciona. Ante un entorno en constante cambio, la selección natural favorece mutaciones que permiten adquirir un mejor desempeño en el nuevo ambiente, lo que resulta en adaptación y evolución (Darwin, 1859).

Existen tres tipos de selección natural (Stern & Nielsen 2019):

1. Selección direccional: ocurre cuando la selección natural favorece uno de los extremos de la distribución empírica de una característica. Durante muchas generaciones, uno de los fenotipos más extremos se vuelve más común y el otro menos frecuente, o incluso se elimina (figura 6a).
2. Selección disruptiva: ocurre cuando la selección natural favorece los dos extremos de la distribución empírica. A lo largo de varias generaciones, los fenotipos extremos se han vuelto más comunes y este proceso puede eventualmente conducir a la creación de dos nuevas especies (figura 6b).
3. Selección estabilizadora: ocurre cuando la selección natural favorece los fenotipos intermedios de la distribución empírica. Durante muchas generaciones, los valores intermedios se vuelven más comunes y los fenotipos extremos, menos comunes o se eliminan (figura 6c).



Recursos *zoogenéticos*

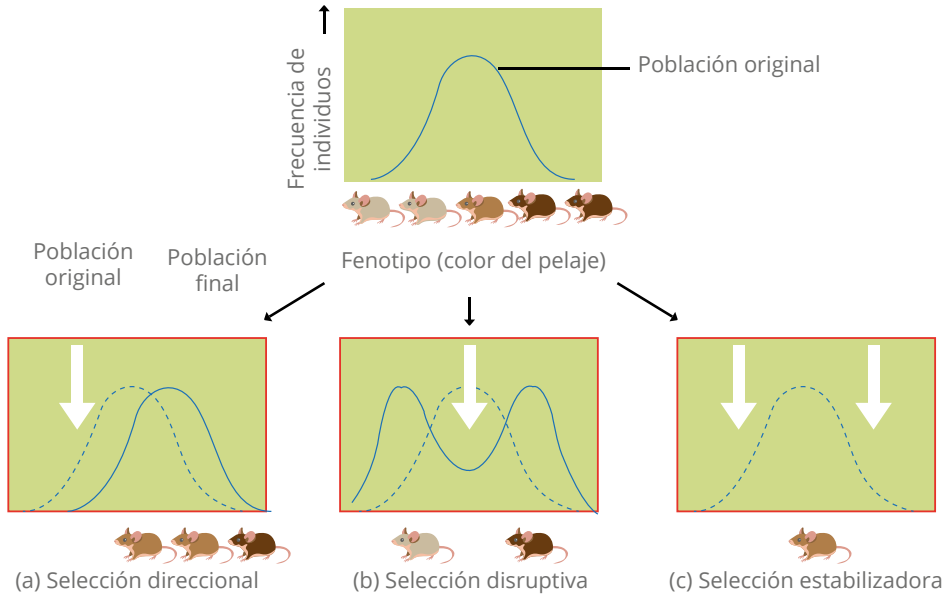


Figura 6. Tipos de selección. a Selección direccional; b. Selección disruptiva; c. Selección estabilizadora. Fuente: Elaboración propia, a partir de “Natural selection” (s. f.)

La selección artificial y la domesticación

La selección artificial es, como su nombre lo indica, no natural y refleja cómo la humanidad comenzó a alterar el patrimonio vivo del planeta Tierra hace aproximadamente 12.000 años. La explotación del patrimonio genético vegetal y animal en beneficio del hombre le permitió asumir por primera vez un papel fundamental en el proceso evolutivo. En un sentido más amplio, esta selección artificial se llama *domesticación* (Roots, 2007). En general, las variantes domesticadas no constituyen especies diferentes de las especies silvestres, de las cuales fueron domesticadas. Cuando las variantes domesticadas tienen una distribución geográfica simpátrica con la especie salvaje de la cual descenden, se produce lo que conocemos como *flujo de genes* (Driscoll et al., 2009).



La domesticación tuvo como consecuencia una disminución de la variabilidad genética, pero se le reconoce la influencia en ecotipos o razas que originó, que son adaptadas a agroecosistemas específicos y que fueron cruciales para las poblaciones humanas. Dichas razas se convirtieron en un patrimonio conjuntamente con sus parientes salvajes (Hoban et al., 2020). Esta enorme diversidad de fenotipos es hoy una oportunidad única para investigar sus genomas y comprender las relaciones entre características y regulación genómica, para que en el futuro se puedan crear normas de manejo de poblaciones que permitan la preservación de características cruciales para la adaptación a los diferentes ambientes.

Detección de huellas de selección en el genoma de la población: búsqueda de valores atípicos

La identificación de huellas de selección asociadas a una característica específica requiere, por un lado, determinar la diversidad genética de una población y, por otro lado, inferir procesos evolutivos no neutros para identificar regiones con un nivel de diversidad de nucleótidos significativamente diferente al esperado (regiones atípicas). Los principales estimadores de diversidad genética de una población se describen a continuación.

Medidores clásicos de diversidad genética dentro de una población

Para reconocer huellas de selección que estén asociadas con características adaptativas o productivas, se utilizan los siguientes estimadores de diversidad genética (Hamilton, 2009):



Recursos *zoogenéticos*

1. Número de posiciones que secretan polimorfismos en una muestra.
2. Promedio de “diferencias por pares”.
3. Tasa de mutación (*scalated mutation rate*).

El primer estimador es el más simple. Una posición de segregación es cualquier posición en un segmento dado de ADN, que es polimórfico (posiciones 2, 6 y 8 en la figura 7). El número de posiciones de segregación (S) se estima mediante la relación entre el número total de posiciones de segregación y el número total de tipos de polimorfismo presentes en las muestras estudiadas. Por lo tanto, es fundamental para una buena estimación de posiciones segregantes realizar una muestra representativa de la población para capturar todas las variantes posibles.

En el segundo estimador, el número medio de diferencias por pares se puede estimar mediante:

$$\Pi = \frac{1}{\binom{n(n-1)}{2}} \sum_{i < j} \Pi_{ij}$$

Donde n es el número de secuencias muestreadas (de modo que $n(n-1)/2$ es el número de posibles comparaciones) y Π_{ij} es la diferencia entre las secuencias i y j (figura 7). Este valor se puede normalizar si se divide por L. La relación Π/L se conoce como diversidad de nucleótidos (π).



Nucleotide diversity (π)										Segregating sites (S and p^s)		
1	A	A	T	G	T	C	A	A	C	G	$d_{12}=0$	Sites 2,6, and 8 have variable base pairs among the four sequences (columns in shade). These are segregating sites. Therefore, for these sequences, $S=3$ and $p^s = 3/10=0.3$ segregating sites per nucleotide site examined.
2	A	A	T	G	T	C	A	A	C	G		
1	A	A	T	G	T	C	A	A	C	G	$d_{13}=0$	
3	A	T	T	G	T	C	A	A	C	G		
1	A	A	T	G	T	C	A	A	C	G	$d_{14}=3$	
4	A	T	T	G	T	G	A	T	C	G		
2	A	A	T	G	T	C	A	A	C	G	$d_{23}=1$	
3	A	T	T	G	T	C	A	A	C	G		
2	A	A	T	G	T	C	A	A	C	G	$d_{24}=3$	
4	A	T	T	G	T	G	A	T	C	G		
3	A	T	T	G	T	C	A	A	C	G	$d_{34}=2$	
4	A	T	T	G	T	G	A	T	C	G		
$\sum d_{ij} = d_{12} + d_{13} + d_{14} + d_{23} + d_{24} + d_{34} = 0 + 1 + 3 + 1 + 3 + 2 = 10$ differences Number of pair of sequences compared = $[n(n-1)]/2 = 6$ $\Pi_{10} = 1.67$ average pairwise differences $\pi = 1.67/10 = 0.167$ pairwise differences per site											Scaled mutation rate (θ) $\theta = 0.3 / ((1/3) + (1/2) + (1/1)) = 0.164$	

Figura 7. Estimadores de diversidad genética.

Nota: Ejemplo práctico de cálculo utilizando secuencias de ADN hipotéticas con diez nucleótidos cada una. Fuente: Elaboración propia, a partir de Hamilton (2009)

El tercer estimador de la diversidad genética dentro de una población viene dado por θ , que es igual a $4Ne\mu$ (Watterson, 1975), donde Ne es el tamaño efectivo de una población y μ es la tasa de mutación. El estimador θ describe la variabilidad esperada en cada nucleótido, y si la evolución fuera completamente



neutra, podría ser utilizado para probar la teoría neutral. Basado en el modelo de *sitio infinito* y apareamiento aleatorio, Watterson (1975) demostró que θ puede estimarse a partir del número de posiciones segregantes (S) en una muestra de secuencias de ADN (n). Si definimos una nueva variable:

$$\sum_{k=1}^{n-1} 1/k$$

Entonces:

1. Utilizando el número absoluto de posiciones segregantes: $\theta=S/a_i$.
2. Usando el valor de posiciones segregantes por nucleótido muestreado: $\theta=p_s/a_i$.

La teoría neutral de la evolución molecular y la detección de valores atípicos en el genoma: la herramienta para la detección de huellas

La teoría neutral de la evolución se basa en la premisa de que la mayoría (si no todas) de las mutaciones que ocurren a bajas frecuencias en la población no afectan la “aptitud” de los individuos (Kimura, 1983). Por lo tanto, las mutaciones que son deletéreas se mantendrán bajas en la población o incluso se eliminarán. Así, los polimorfismos comunes que existen en una población derivan de fenómenos de selección neutral. Otra premisa de esta teoría es que la mayoría de las alteraciones evolutivas provienen de la acción de la deriva genética sobre los alelos neutrales. De acuerdo con la teoría de la evolución molecular neutral, la variabilidad entre poblaciones y dentro de una población ocurre principalmente debido a mutaciones neutrales. Usando el modelo de coalescencia, la diversidad de nucleótidos observada se puede comparar con la diversidad de nucleótidos esperada, en relación con el modelo de evolución neutral para detectar regiones genómicas donde los niveles de diversidad de nucleótidos son significativamente



diferentes de los esperados (Van Tassell et al., 2008). A partir de estos principios se desarrolló la prueba estadística conocida como D de Tajima (1989). Según el modelo de teoría neutral:

$$E(\pi) = \theta = E \left[\frac{S}{\sum_{i=1}^{n-1} \frac{1}{i}} \right] = 4N\mu$$

Dado que S es el número de posiciones segregantes, n es el número de individuos, N es el tamaño efectivo de la población y μ es la tasa de mutación. Si ha tenido lugar la selección, se producirán fluctuaciones demográficas y violaciones de otras premisas del modelo de evolución neutral. De esta forma, el valor de S y, en consecuencia, el valor de $E(\pi)$ serán diferentes a los esperados. La diferencia entre los valores esperados y los valores observados de estos parámetros es el pilar de la prueba D de Tajima. El estimador D es estimado por la diferencia entre $E(\pi)$ y S se llama d y D. Se considera:

$$D = \frac{d}{\sqrt{\hat{V}(d)}}$$

Los valores D fuera del intervalo de confianza corresponden a regiones donde se excluye la hipótesis nula de mutación neutra. Estas regiones que contienen valores de diversidad de nucleótidos significativamente diferentes de los esperados pueden interpretarse como huellas de selección.

Existen varios tipos de huellas de selección (figura 8):

1. Selección positiva: es un tipo de selección direccional en la que se produce una reducción de la variabilidad genética en una determinada región genómica (*selective sweep*). El *sweep* aumenta el alelo mutado en pocas generaciones y preserva la estructura del haplotipo original, ya que el número de eventos de recombinación será limitado, y con esto habrá un aumento en la extensión del



Recursos zoogenéticos

desequilibrio de ligamiento (LD) (figura 8b). Por tanto, además de las estadísticas D de Tajima (1989), deben aplicarse otros métodos; por ejemplo, *extended haplotype homozygosity* (EHH) desarrollado por Sabeti et al. (2002), permite identificar regiones con una gran extensión de LD.

2. Selección purificadora: también es una huella de selección direccional, pero en la que los alelos deletéreos se eliminan de la población, a través de un largo proceso durante el cual la frecuencia del alelo derivado aumenta, lo que resulta en un exceso de alelos de baja frecuencia y una reducción del desequilibrio de ligamiento (figura 8d).
3. Selección de equilibrio: cuando se produce este tipo de selección, hay muchos alelos que se conservan en la población durante varias generaciones; en consecuencia, hay un aumento de la variabilidad (diversidad de nucleótidos) y de alelos con frecuencias intermedias en la población. La extensión del desequilibrio de ligamiento también aumenta (figura 8c).

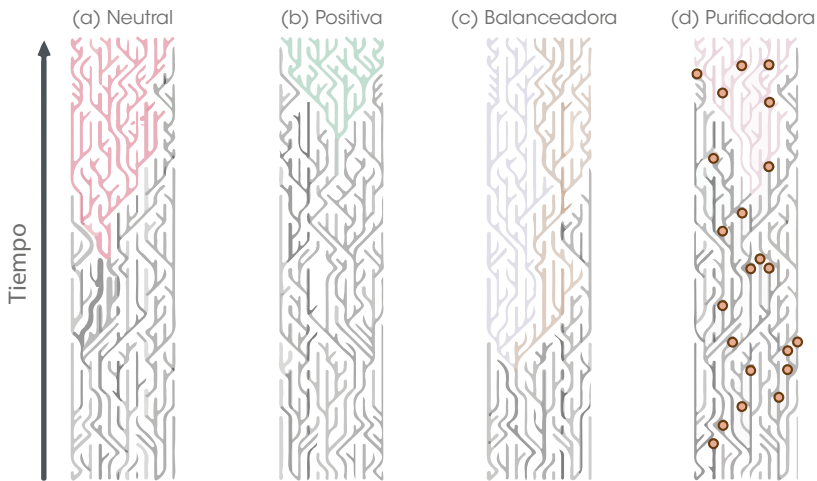


Figura 8. Comparación del modelo neutral con diferentes tipos de selección. a. Demostración de la evolución según el modelo neutral; b. Efecto de selección positiva; c. Efecto de selección balanceadora o de equilibrio; d. Efecto de selección purificante.

Fuente: Elaboración propia, a partir de Awadalla & Hobolth (2009)



Al comparar diferentes poblaciones, se pueden utilizar estimadores de diferenciación genética. Las regiones genómicas que se ubican en las colas de la distribución empírica del índice de fijación (F_{st}) (Wright, 1984), el cual mide el grado de diferenciación que ocurrió entre dos poblaciones, corresponden a huellas de selección que reflejan eventos de divergencia entre razas o poblaciones.

La interpretación de los resultados obtenidos no es trivial, pues existen procesos demográficos que conducen a cambios similares en los genomas; por ejemplo, explosiones demográficas pueden generar una alta ocurrencia de alelos con baja frecuencia en la población. Todos estos métodos son un desafío para distinguir entre señales verdaderamente positivas y otras derivadas del azar o del “ruido” (Kemper et al., 2014). Es posible que la ocurrencia de la subdivisión de la población cree una alta ocurrencia de alelos con frecuencia intermedia; asimismo, el uso del método EHH también provoca un elevado número de falsos positivos o falsos negativos, debido a la heterogeneidad de las tasas de recombinación en todo el genoma, pero también ha permitido muchos avances en la comprensión del proceso de domesticación de varias especies (Amaral et al., 2011; Choi et al., 2015).

Tecnologías genómicas de nueva generación y su aplicación

Los polimorfismos de un solo nucleótido (SNP) son marcadores bialélicos fáciles de evaluar e interpretar, ya que cuentan con amplia distribución en los genomas. Utilizando un conjunto de SNP con cobertura y alta densidad en todo un genoma, los SNP permiten la captura de información de LD incrustada en el genoma, que puede ser utilizada por los estimadores mencionados anteriormente para identificar genes asociados con características y enfermedades. En los animales domésticos, estas herramientas pueden contribuir a: a) mejorar la comprensión del proceso de domesticación y formación de razas; b) desarrollar nuevas teorías de



Recursos *zoogenéticos*

genética de poblaciones; c) identificar los mecanismos genéticos asociados con características de heredabilidad genética compleja, y d) optimizar los métodos de selección en el contexto de la producción animal. De esta manera, se han desarrollado paneles SNP, utilizando diferentes tecnologías (Illumina, Affymetrix, Neogen), que permiten el genotipado simultáneo de unos pocos miles (paneles de baja densidad, 3k, 10k) o muchos miles (500-700k) para una misma muestra con una relación costo-eficiencia factible para diferentes contextos, desde la industria hasta la investigación. En principio, estas nuevas plataformas estuvieron disponibles para animales de granja, para los que ya existía un genoma de referencia.

La disponibilidad de nuevas tecnologías de secuenciación de alta densidad permite hoy en día secuenciar genomas completos de varios animales. Estas tecnologías también se usan de diferentes formas para permitir la identificación rápida de SNP en animales para los cuales el genoma de referencia ya existe o aún no está disponible. Hay diferentes estrategias que se pueden considerar (Davey et al., 2011), como las siguientes:

- Creación de bibliotecas representativas del genoma (RRL): con el uso de enzimas de restricción, se hace un muestreo de regiones del genoma que se va a secuenciar. Se pueden usar una o más enzimas de restricción para producir una o más bibliotecas de fragmentos de ADN distribuidos por todo el genoma. Incluye un paso de selección de los fragmentos de ADN más abundantes en la muestra. Se ha utilizado para la secuenciación conjunta de una muestra de animales representativos de la población, y esto ha permitido la identificación de polimorfismos característicos de una población, pero no a nivel individual; asimismo, permite obtener la secuenciación completa de fragmentos hasta 100 pb. Esta estrategia se utilizó para el desarrollo de los primeros paneles de SNP porcinos (Amaral et al., 2009; Ramos et al., 2009) y bovinos (Van Tassell et al., 2008), en pavos (Kerstens et al., 2009) y en otras especies. En la actualidad, utilizando



códigos de barras, con este método o estrategia es posible la identificación individual antes de agrupar la muestra.

- RAD-seq: emplea enzimas de restricción para realizar muestreos a lo largo del genoma. Como el protocolo implica un paso de *shearing*, permite obtener secuencias entre intervalos más distantes entre sitios de restricción y, de esta manera, secuenciar zonas de la muestra que tienen motivos de restricción en diferentes ubicaciones. Esta solución ha sido muy utilizada para la identificación de polimorfismos en especies sin genoma de referencia y, usando secuenciación en modo *paired-end*, permite con un diseño experimental adecuado reconstruir fragmentos largos.
- GBS-seq: utiliza enzimas de restricción, y es una técnica más reciente que las RRL; no incluye el paso de selección de tamaño (RRL) ni el paso de *shearing* (RAD-seq). Permite obtener fragmentos completos entre 150-350 pb utilizando un diseño experimental en modo *paired-end*. El uso de GBS-seq supone la disponibilidad de un genoma de referencia.

Plataformas de intercambio de datos y su papel en los beneficios socioeconómicos asociados y de conservación de razas locales

La posibilidad de producción masiva de datos genómicos por parte de científicos de todo el mundo genera el potencial para que la comunidad científica contribuya al avance en el conocimiento de la genética y de cómo esta aporta al funcionamiento de la biología de los organismos. El conocimiento generado en razas con una distribución más amplia desde el punto de vista geográfico, o más utilizadas para la producción a mayor escala, ayuda a comprender y a estudiar la genética de razas locales con menores resultados. Para este fin, la comunidad científica ha debatido durante varios años temas asociados al intercambio de



Recursos *zoogenéticos*

datos. Así, junto a la publicación de un estudio, las revistas de mayor reputación requieren el intercambio de datos en repositorios públicos; esto permite a los investigadores explorar un conjunto más amplio de datos, estimula la creatividad y el pensamiento científico, mejora la calidad de la educación académica en todo el mundo y crea sinergias con mayor eficiencia. Por ejemplo, en el caso de estudios de razas locales, estos repositorios son muy útiles, ya que permiten explorar su filogeografía comparando datos de una raza local con otras. Respecto a los estudios de polimorfismos y variabilidad de ADN, en la tabla 5 se muestran los repositorios más utilizados por la comunidad científica.

Tabla 5. Algunos repositorios públicos de datos de secuenciación y genotipado

Nombre	Link	Tipo de datos
Widde	http://widde.toulouse.inra.fr/widde/	Paneles de SNP de bovinos y ovinos.
DRYAD	https://datadryad.org/stash	Todo tipo de datos, incluida información sobre las características de la muestra. Los datos del panel de SNP de diferentes especies están disponibles.
European Nucleotide Archive	https://www.ebi.ac.uk/ena/browser/home	Datos sobre secuenciación de DNA, RNA y epigenética.

Fuente: Elaboración propia

Referencias

Amaral, A. J., Ferretti, L., Megens, H.-J., Crooijmans, R., Nie, H., Ramos-Onsins, S. E., Pérez-Enciso, M., Schook, L. B., & Groenen, M. A. M. (2011). Genome-Wide Footprints of Pig Domestication and Selection Revealed through Massive Parallel Sequencing of Pooled DNA. *PLoS One*, 6(4), e14782. <http://dx.plos.org/10.1371/journal.pone.0014782>



- Amaral, A. J., Megens, H.-J., Kerstens, H. H. D., Heuven, H. C. M., Dibbits, B., Crooijmans, R., Den Dunnen, J., & Groenen, M. A. M. (2009). Application of Massive Parallel Sequencing to Whole Genome SNP Discovery in the Porcine Genome. *BMC Genomics*, *10*, 374. <https://doi.org/10.1186/1471-2164-10-374>
- Awadalla, P., & Hobolth, A. (2009). *Lecture Notes from Module 6: Coalescent Theory. Summer Institute of Statistical Genetics* [Notas de clase]. University of Liège.
- Choi, J.-W., Choi, B.-H., Lee, S.-H., Lee, S.-S., Kim, H.-C., Yu, D., Chung, W.-H., Lee, K.-T., Chai, H.-H., Cho, G.-M., & Lim, D. (2015). Whole-Genome Resequencing Analysis of Hanwoo and Yanbian Cattle to Identify Genome-Wide SNPs and Signatures of Selection. *Molecules and Cells*, *38*(5), 466-473. <https://doi.org/10.14348/molcells.2015.0019>
- Darwin, Ch. (1859). *On the origin of species by means of natural selection, or preservation of favoured races in the struggle for life*. John Murray.
- Davey, J. W., Hohenlohe, P. A., Etter, P. D., Boone, J. Q., Catchen, J. M., & Blaxter, M. L. (2011). Genome-wide genetic marker discovery and genotyping Using next-generation sequencing. *Nature Reviews Genetics*, *12*(7), 499-510. <https://doi.org/10.1038/nrg3012>
- Driscoll, C. A., Macdonald, D. W., & Stephen, J. (2009). From wild animals to domestic pets, an evolutionary view of domestication. *Proceedings of the National Academy of Sciences*, *106*(1), 9971-9978. <https://doi.org/10.1073/pnas.0901586106>.
- Hamilton, M. B. (2009). *Population genetics*. Wiley-Balckwell.
- Hoban, S., Bruford, M., D'Urban Jackson, J., Lopes-Fernandes, M., Heuertz, M., Hohenlohe, P. A., Paz-Vinas, I., Sjögren-Gulve, P., Segelbacher, G., Vernesi, C., Aitken, S., Bertola, L. D., Bloomer, P., Breed, M., Rodríguez-Correa, H., Funk, W. C., Gruber, C. E., Hunter, M. E., Jaffe, R., Liggins, L. ... Laikre, L. (2020). Genetic diversity targets and indicators in the CBD post-2020 Global Biodiversity Framework must be improved. *Biological Conservation*, *248*, 108654. <https://doi.org/https://doi.org/10.1016/j.biocon.2020.108654>
- Kemper, K. E., Saxton, S. J., Bolormaa, S., Hayes, B. J., & Goddard, M. E. (2014). Selection for Complex Traits Leaves Little or no Classic Signatures of Selection. *BMC Genomics*, *15*(1), 246. <https://doi.org/10.1186/1471-2164-15-246>



Recursos zoogenéticos

- Kerstens, H. H. D., Crooijmans, R., Veenendaal, A., Dibbitts, B. W., Chin-A-Woeng, T. F. C., Den Dunnen, J., & Groenen, M. A. M. (2009). Large scale single nucleotide polymorphism discovery in unsequenced genomes using second generation high throughput sequencing technology: Applied to Turkey. *BMC Genomics*, 10(1), 479. <https://doi.org/10.1186/1471-2164-10-479>
- Kimura, M. (1983). *The neutral theory of molecular evolution*. Cambridge University Press. <https://doi.org/10.1017/CBO9780511623486>
- Natural selection. (s. f.). *Biology Dictionary*. <https://biologydictionary.net/natural-selection/>
- Ramos, A. M., Crooijmans, R., Affara, N. A., Amaral, A. J., Archibald, A. L., Beever, J. E., Bendixen, C., Churcher, C., Clark, R., Dehais, P., Hansen, M. S., Hedegaard, J., Hu, Z.-L., Kerstens, H. H., Law, A. S., Megens, H.-J., Milan, D., Nonneman, D. J., Rohrer, G. A., Rothschild, M. F., Smith, T. P. L. ... Groenen, N. A. M. (2009). Design of a High Density SNP Genotyping Assay in the Pig Using SNPs Identified and Characterized by Next Generation Sequencing Technology. *PLoS One*, 4(8), 1-13. <https://doi.org/https://doi.org/10.1371/journal.pone.0006524>
- Roots, C. (2007). *Domestication*. Greenwood Publishing Group.
- Sabeti, P. C., Reich, D. E., Higgins, J. M., Haninah, Z. P. L., Richter, D. J., Schaffner, S. F., Gabriel, S. B., Platko, J. V., Patterson, N. J., McDonald, G. J., Ackerman, H. C., Campbell, S. J., Altshuler, D., Cooper, R., Kwiatkowski, D., Ward, R., & Lander, E. S. (2002). Detecting Recent Positive Selection in the Human Genome from Haplotype Structure. *Nature*, 419(6909), 832-837. <https://doi.org/10.1038/nature01140>
- Stern, A. J., & Nielsen, R. (2019). Detecting natural selection. En D. J. Baldin, I. Moltke, & J. Marioni (Eds.), *Handbook of Statistical Genomics* (pp. 397-40). Wiley. <https://doi.org/10.1002/9781119487845.ch14>.
- Tajima, F. (1989). Statistical Method for Testing the Neutral Mutation Hypothesis by DNA Polymorphism. *Genetics*, 123(3), 585-595. <https://pubmed.ncbi.nlm.nih.gov/2513255>



- Van Tassell, C. P., Smith, T. P. L., Matukumalli, L. K., Taylor, J. F., Schnabel, R. D., Taylor Lawley, C., Haudenschild, C. D., Moore, S. S., Warren, W. C., & Sonstegard, T. S. (2008). SNP discovery and allele frequency estimation by deep sequencing of reduced representation libraries. *Nature Methods*, 5(3), 247-252. <https://doi.org/10.1038/nmeth.1185>
- Watterson, G. A. (1975). On the number of segregating sites in genetical models without recombination. *Theoretical Population Biology*, 7(2), 256-276.
- Wright, S. (1984). *Evolution and the Genetics of Populations*, (Vol. 2). University of Chicago Press.

