Instituto de Genética
Facultad de Agronomía
Universidad Central de Venezuela

Este material es usado exclusivamente para la docencia en cursos del Postgrado en Agronomía, orientación Mejoramiento Genético de Plantas. Muestra conceptos y modelos básicos sobre el uso de marcadores para caracterizar la diversidad y estructura genética de las poblaciones, obtención de mapas genéticos y detección de loci responsables de la variación de caracteres cuantitativos (QTL). No pretende ser exhaustivo en cada tema sino parcializado para enfocarse en sus fundamentos genéticos con fines docentes.

A.D.-P.

USO DE LOS MARCADORES GENÉTICOS EN LOS ESTUDIOS DE DIVERSIDAD GENÉTICA, MAPEO Y DETECCIÓN DE QTL EN PLANTAS

Antonio Díaz-Pérez
Maracay 2005

1. DIVERSIDAD GENÉTICA

El estudio de la diversidad genética parte de su caracterización. Si se dispone de los marcadores genéticos, esta caracterización se hace en función de variantes alélicas. Dicha variación es procesada mediante índices y métodos genético-poblacionales y evolutivos.

Para medir la variabilidad genética un marcador genético debe cumplir con los siguientes criterios: i) distinguir la expresión alélica a nivel de individuos (reconocimiento de los heterocigotos); ii) determinar el efecto de cada sustitución alélica de manera específica para cada locus y distinguir las sustituciones en otros loci; iii) detectar todas las sustituciones de bases; y, iv) muestrear al azar los loci, independientemente de su función o nivel de polimorfismo.

Dependiendo del tipo de marcador, se cumplirán parcial o totalmente las condiciones previamente señaladas. Por ejemplo, los marcadores microsatélites, de herencia codominante, garantizan el cumplimiento de la primera condición mientras que los AFLP no. La sensibilidad de los marcadores de ADN para detectar todas las sustituciones de bases posibles (condición 3) los hace preferibles a las isoenzimas, estas últimas sensibles a un 30% de las variaciones a nivel del ADN. Los microsatélites y RAPD generalmente se ubican en todo el genoma, en zonas transcritas y en zonas sin función aparente alguna; los RFLP y en mayor grado las isoenzimas están restringidos a zonas transcritas, lo cual introduce un sesgo en el muestreo del genoma (condición 4).

Los alelos son diferentes formas de un locus. Implica variaciones a nivel de los nucleótidos. Cuando se trabaja con marcadores genéticos, un locus no necesariamente coincide con un cistrón, ya que puede ubicarse en regiones no codificadoras.

2.1 Cálculo de las frecuencias alélicas

Considere un locus *A* con n_{uu} homocigotos y n_{uv} heterocigotos (conteo genotípico) en una muestra de tamaño n. El número n_u de alelos A_u sería

$$nu = 2n_{uu} + \Sigma_{v \neq u}\, n_{uv}$$

y la frecuencia alélica muestral

$$\tilde{p}_u = n_u / 2n$$

El hecho de que $\mathcal{E}(\tilde{p}_u) = p_u$, indica que en el caso de que se tomen varias muestras de una misma población (muestreo estadístico), el promedio de la frecuencia del alelo *Au* para todas ellas será igual a la frecuencia del alelo en la población (p_u). Las frecuencias $\tilde{p}_u$ de las muestras varían entre sí, lo cual se puede cuantificar a través de la varianza de la frecuencia alélica muestral

$$Var(\tilde{p}_u) = 1/2n\,(p_u + P_{uu} - 2p_u^2),$$

donde P_{uu} es la frecuencia genotípica del homocigoto A_uA_u. Note que esta expresión involucra las frecuencias de la población p_u y P_{uu}. Si se sustituyen las frecuencias muestrales en dicha ecuación, es decir, $\tilde{p}_u$ y $\tilde{P}_{uu}$, se obtiene un estimado de la varianza de $\tilde{p}_u$, el cual se denomina $\widehat{Var}(\tilde{p}_u)$.

Si el n de la muestra es mayor que 30, asumiendo que $\tilde{p}_u$ se distribuye aproximadamente como una distribución normal, se puede construir un intervalo de confianza para p_u. La expresión general para los límites de confianza vienen dados por la expresión

$$\tilde{p}_u \pm z_{1-\alpha/2} [\widehat{Var}(\tilde{p}_u)]^{1/2}$$

en el que $z_{1-\alpha/2}$ es el valor de la abscisa de la distribución normal tipificada asociada con error tipo I igual a α. Para un α=0,95, el valor de z es de 1,96.

Cuando se asume la existencia del equilibrio de Hardy y Wienberg (EHW), que se alcanza cuando los miembros de la población se aparean al azar, los alelos como las frecuencias genotípicas se distribuyen al azar. En este caso, la frecuencia de cada alelo se distribuye de acuerdo a la distribución binomial. Bajo el EHW

$$Var(\tilde{p}_u) = p_u(1-p_u) / 2n.$$

2.2 Análisis de una población

Según el escenario de una población fija de (Weir 1996), sólo se consideraría el efecto del muestreo estadístico en una población predeterminada. La población no varía genéticamente. La información que se analiza a partir de los marcadores proviene de una muestra de n individuos, tomados de una población de tamaño N. La escogencia de estos individuos se denomina *Muestreo Estadístico*. Los estadísticos que se obtienen de esta muestra pueden ser distintos de otra muestra tomada de la misma población. Aun cuando, no se toman muestras repetidas dentro de una población, la teoría estadística predice cuánta variación es probable entre dichas muestras, variación que sirve para hacer inferencias acerca de la población a partir de la muestra.

2.3 Comparación de frecuencias entre poblaciones

La comparación de poblaciones-fijas distintas se puede hacer a través de las frecuencias genotípicas si las poblaciones no se encuentran en EHW. Si se toma una muestra de una población de tal manera que cada miembro tiene la misma oportunidad de ser considerado, a la vez que los individuos son muestreados independientemente, se alcanza el EHW. Bajo esta condición los conteos genotípicos y alélicos (o frecuencias absolutas) se distribuyen multinomialmente, por lo que se pueden comparar las frecuencias alélicas según esta distribución.

Un procedimiento que facilita la comparación estadística entre poblaciones-fijas es la prueba de χ^2 basado en una tabla de

contingencia. Las poblaciones (factor 1) se arreglan en filas y los genotipos (factor 2) en columnas. En la prueba de χ_C^2, las frecuencias esperadas se obtienen a partir de

$$E_{ij} = \Sigma_i^{c1} \Sigma_j^{c2} (X_{i.} * X_{.j}) / X..$$

donde

E_{ij} = es la frecuencia absoluta esperada para la i-ésima clase del factor 1 y la j-ésima clase del factor 2
$Xi.$ = es el total marginal con la i-ésima clase del factor 1
$J.j$ = es el total marginal con la el j-ésima clase del factor 2
$X..$ = es el total general

Los valores esperados se calculan bajo el supuesto de que la hipótesis nula es verdadera (independencia de los factores), la cual se distribuye aproximadamente como χ_C^2, *con (m-1)*(n-1)* grados de libertad, donde *n* es el número de clases del factor 1 y *m* el número de clases del factor 2. Para una buena aproximación, los valores esperados no deberían ser tan pequeños.

El estadístico χ_C^2 se obtendría como

$$\chi_C^2 = \Sigma_i^{c1} \Sigma_j^{c2} (X_{ij} - E_{ij})^2 / E_{ij}$$

En presencia del EHW, las frecuencias genotípicas se sustituyen por las frecuencias alélicas.

Un segundo procedimiento es el remuestreo numérico o *bootstrapping*. Con él se pueden hacer inferencias sobre p_u a partir de las frecuencia muestral $\tilde{p}_u$. Se construyen las varianzas y los intervalos de confianza a partir de una serie de muestras nuevas, generadas de la muestra original. Cada muestra nueva se forma al tomar n individuos al azar con reemplazo. Algunos individuos pueden salir repetidos en una muestra nueva; otros no salir inclusive. A partir de cada muestra nueva se obtiene $\tilde{p}_u$, lo que permite caracterizar la distribución de los estimados.

Un intervalo de confianza para un porcentaje de confiabilidad (1-α)%, se construye con dos valores estimados de $\tilde{p}_u$ en los cuales se enmarcan el (1-α)% central del total de estimados. Dos poblaciones se consideran con distintas frecuencias alélicas si los intervalos de confianzas respectivos no se solapan.

2.4 Equilibrio de Hardy-Weinberg y coeficiente de endocría

En el Equilibrio de Hardy y Weinberg (EHW) las frecuencias genotípica de la población se presentan según la unión al azar de las gametas de la generación anterior. En una población sin subdivisiones, el EHW se puede perder debido a una mayor proporción de cruzamientos entre parientes que al azar, al cruzamiento entre semejantes fenotípicos y a la selección, entre otros. La falta de ajuste de una población al EHW se puede caracterizar de muchas formas; sólo se tratarán tres de ellas: a) mediante una prueba de χ^2, b) a través de pruebas exactas y c) mediante el coeficiente *f* ó coeficiente de endocría.

En la prueba de $\chi 2$ los valores absolutos observados para cada una de las clases genotípicas se confrontan con las esperadas, que se obtienen de la siguiente manera

$$P_{uu} = p_u^2$$
$$P_{uv} = 2p_u p_v, \text{ para u} \neq \text{v}$$

Los grados de libertad de la prueba serán igual a el total de clases menos uno, menos el número de parámetros estimados en la prueba: para dos alelos sería tres grados en total, menos uno, menos otro grado de libertad al estimarse la frecuencia de un alelo (el otro no es necesario pues es complementario al primero); para tres alelos serían un total de seis, menos uno, menos dos más (ya que se estiman dos frecuencias alélicas; la tercera se obtiene por complemento). El estimado de $\chi 2$ para dos alelos se calcularía de la siguiente manera

$$\chi^2 = (n_{uu} - np^-_u{}^2)^2 / np^-_u{}^2 + (n_{uv} - n2p^-_u p^-_v)^2 / n2p^-_u p^-_v + (n_{vv} - q^-_u{}^2)^2 / nq^-_u{}^2$$

Con la aparición de marcadores hipervariables, como los microsatélites, es común la existencia de múltiples alelos en la caracterización de pocos individuos. Ello trae como consecuencia que el número de posibles genotipos en EHW sea muy grande, lo cual rara vez se satisface en muestras pequeñas. Por otra parte, la prueba χ^2 requiere que en todas las clases genotípicas se presenten al menos 5 individuos. Una forma de tratar esta situación es mediante el uso de pruebas exactas.

Las pruebas exactas buscan comprobar si las frecuencias genotípicas observadas se aproximan a los productos de las frecuencias alélicas observadas. Estas pruebas evalúan todas las posibles combinaciones de frecuencias genotípicas para las frecuencias alélicas observadas, rechazando la hipótesis de EHW si las frecuencias genotípicas observadas tienden a ser inusuales bajo EHW. Para el caso de dos alelos, la probabilidad de las frecuencias genotípicas, asumiendo EHW, condicional a las frecuencias alélicas observadas es

$$Pr = [(n!/n_{uu}!n_{uv}!n_{vv}!)(p^-_u{}^2)^{n_{uu}}(2p^-_u p^-_v)^{n_{uv}}(p^-_v{}^2)^{n_{vv}}] / [(2n)!/n_u!n_v!](p^-_u)^{n_u}(p^-_v)^{n_v}$$

Las probabilidades de todas las combinaciones se ordenan y aquéllas que se encuentran en la proporción menos probable α (0,05) del total se consideran la zona de rechazo.

El coeficiente de endocría *f* es la correlación de los dos alelos que forman un individuo diploide o equivalentemente la probabilidad de que los dos alelos provengan de un ancestro común. El coeficiente *f* puede ser estimado observando registros genealógicos particulares; sin embargo, en poblaciones, una medida promedio de *f* se obtiene como el grado en que la proporción observada de heterocigotas se aleja de la proporción esperada en EHW.

Se pueden definir los parámetros *Ho* y *He* como la proporción observada y esperadas respectivamente de heterocigotos en la población. Entonces *f* se obtendría como

$$f = (He - Ho) / He = 1 - Ho/He$$

A partir de una muestra, dos estimados de f ^ serían

$$f \wedge = 1 - [n_{uv} / 2np^{-}_{u}(1-p^{-}_{u})],$$

obtenido por el método de máxima verosimilitud

$$f \wedge = 1 - [(n-1)n_{uv}/n]/ [2np^{-}_{u}p^{-}_{v} - n_{uv} / 2n],$$

por el método de los momentos.

El último presenta un menor sesgo que el primero, pero puede tener una mayor varianza.

2.5 Índices de diversidad génica

Se pueden identificar dos tipos:

a) Riqueza alélica: número de alelos distintos.
b) Uniformidad: distribución de las frecuencias alélicas.

Las frecuencias alélicas no son índices de diversidad genética, por lo que es necesario hacer una serie de transformaciones. Entre los índices de diversidad más empleados se encuentran:

El porcentaje de loci polimórficos (P): es una guía aproximada del nivel de variación en la muestra. Un locus se define como polimórfico si la frecuencia del alelo mayoritario es menor o igual a 0,99; no obstante, esta definición es arbitraria, lo cual hace a este índice ambiguo. Es dependiente del tamaño de muestra. Hace poco uso estadístico de la información dado que es insensible al nivel de polimorfismo en un locus. Combina elementos de ambos conceptos de diversidad sin enfatizar en ninguno de ellos.

El número promedio de alelos por locus (A): enfatiza en el componente *riqueza alélica* de la diversidad. Parece ser importante en el estudio de las constricciones genéticas. Es dependiente del tamaño de muestra porque a) presenta una gran varianza cuando el tamaño de muestra es pequeño, subestimando en forma general el número real en la población y b) incluye genes deletéreos, la mayoría con baja frecuencia en la población, que contribuyen escasamente a la variación génica de la población. Otra desventaja sería que considera a todos los alelos por igual, sin tomar en cuenta su importancia biológica.

Uniformidad de las frecuencias alélicas: es derivada del Índice de diversidad ecológica de Simpson. Esta medida o sus transformaciones han recibido varios nombres; no obstante, la Heterocigosidad promedio o Diversidad Génica ha sido la más empleada en estudios genéticos.

La diversidad génica (*H*) se interpreta superficialmente como la probabilidad de encontrar dos alelos distintos en una población. Este índice puede utilizarse en cualquier organismo, autógamo o alógamo, haploide o poliploide. Se basa en que la heterocigosidad (*H*) en una población **no panmíctica** no está relacionada con la frecuencia de heterocigotos, sino que es una medida de la variación génica. Originalmente se aplicó al polimorfismo isoenzimático de allí que se interprete también como la diferencia promedio de codones de todos los individuos de una población.

Los dos términos principales relacionados con este índice son:

$$H + J = 1$$

donde

H = diversidad génica
J = identidad génica

H es la heterocigosidad promedio (aritmético) y *J* es la homocigosidad promedio (aritmético) de todos los loci evaluados. La homocigosidad en un locus se obtiene a través de

$$j\hat{} = \Sigma \bar{p}_i^2$$

y la heterocigosidad como

$$h\hat{} = 1 - \Sigma \bar{p}_i^2,$$

donde $\bar{p}_i$ es la frecuencia del i-ésimo alelo.

La variación génica en una población (x) se puede medir a través del número promedio de diferencias entre codones pertenecientes a dos genes escogidos al azar. Dado que pueden existir al menos una diferencia entre cualquier par de alelos distintos, el *Número Mínimo de Diferencias entre Codones por Locus entre dos Genomas al Azar* [$D_{x(m)}$] viene dado por

$$D_{x(m)} = 1 - J$$

lo cual es igual a *H*. Existen otros estimados, derivados del primero:

$$D_x = -\log_e J$$

donde *J* se obtiene por media aritmética. Es un estimado estándar

$$D'_x = -\log_e J'$$

donde *J* se obtiene por media geométrica. Es un estimado máximo.

El concepto de diferencias entre codones es útil para medir las diferencias génicas entre dos poblaciones o en la partición de la diversidad génica en poblaciones subdivididas. En la práctica, todos los estimados *D* se refieren a diferencias entre codones que son detectados a través de una técnica. En el caso de electroforesis de isoenzimas, la técnica detecta solamente un 30% de las diferencias entre codones. Otra desventaja sería que cada cambio mutacional de un gen se cuenta como una diferencia entre codones, aun si involucra varias de ellas.

La varianza de *H* incluye un componente intralocus y otro interlocus. Este último se debe a que diferentes loci están afectados por sustituciones de bases en forma independiente. Además, las poblaciones naturales incluyen una mezcla de loci con diferentes grados de evolución, con tasas de mutación y coeficientes selectivos diferentes.

Una varianza aproximada de *H* se obtiene asumiendo que los *h* de distintos loci son equivalentes a los h d poblaciones repetidas. De esta manera,

$$V(H)= \Sigma_t^r (h_t-H)^2/r(r-t),$$

donde *t* se refiere al t-ésimo locus y *r* es el número de loci estudiados. Esta ecuación asume que cada h_r se distribuye como una variable normalmente distribuida con media *H*.

Índice de información (Shannon y Weaver): mide la cantidad de información. No se tiene claro su significado genéticamente. Se usa a menudo en ecología para cuantificar la diversidad entre especies. Es más sensitivo al concepto de riqueza alélica que *h*, aunque es una medida compuesta de ambos conceptos. Se calcula como

$$-\Sigma p_i \log(p_i),$$

en la que p_i es la frecuencia de la i-ésima categoría.

2.6 Desequilibrio de ligamiento

Las asociaciones entre caracteres son generadas por el desequilibrio de ligamiento (*DL*). El *DL* es la correlación entre alelos de diferentes loci. La magnitud de la correlación depende del germoplasma que se considere. En términos generales, la magnitud y dirección de las correlaciones pueden variar entre un grupo y otro. De esta manera, un marcador pudiera estar asociado, dependiendo del grupo genético, a diferentes marcadores y caracteres morfológicos y agronómicos.

El Desequilibrio de Ligamiento (*DL*) es aquella condición de las frecuencias gaméticas de dos o más genes en la cual existe una desviación con respecto a las frecuencias esperadas suponiendo independencia de los mismos. Si se consideran dos genes con dos alelos: gen A (alelos A_1 con frecuencia p_r y A_2 con frecuencia p_s) y el gen B (alelos B_1con frecuencia q_u y B_2 con frecuencia q_v).

Las frecuencias gaméticas en la población considerando *DL* serían

$$P_{ru}\,(A_1B_1) = p_r\,q_u + DL$$
$$P_{rv}\,(A_1B_2) = p_r\,q_v - DL$$
$$P_{su}\,(A_2B_1) = p_s\,q_u - DL$$
$$P_{sv}\,(A_2B_2) = p_s\,q_v + DL$$

La estimación del *DL* requiere conocer las frecuencias gaméticas (P_{ru}, P_{rv}, P_{su}, P_{sv}). Sin embargo, estas frecuencias no están directamente disponibles en individuos diploides, salvo que se asuma apareamientos al azar en la población. Por otra parte, a veces interesa solamente si dos loci se encuentran en *DL* y no su estimación. Una forma simple de verificar la presencia de *DL* es determinar si las frecuencias genotípicas conjuntas para los dos loci se apartan del producto de las frecuencias genotípicas de cada locus (hipótesis nula de no desequilibrio). Se busca entonces la probabilidad de que los genotipos para dos loci sean condicionales de las frecuencias genotípicas para un locus. Esto se expresa como

$$Pr\ [(nrsuv)/(nrs),(nuv)] = (\Pi_{r,s}\ n_{rs}!\ \Pi_{u,v}\ n_{uv}!) / (n!\ \Pi_{r,s,u,v}\ n_{rsuv}!)$$

Esta probabilidad se compara con todas las demás probabilidades generadas por permutación y si cae en la cola de la distribución se rechaza la hipótesis nula de no DL.

Cuando el *DL* es cero, por ejemplo en poblaciones bajo panmixia por muchas generaciones, las frecuencias P_{ij} se comportan como dos genes independientes, no importando la frecuencia de recombinación *r* entre los loci. Esto imposibilita la predicción de la variabilidad morfoagronómica a partir de la variabilidad de los marcadores. El DL puede ser evidente cuando se comparan poblaciones que han estado aisladas reproductivamente, cada una con alelos divergentes para varios loci.

Dado que el *DL* no implica necesariamente un ligamiento físico entre loci, el *DL* se puede perder si se modifican las condiciones iniciales del germoplasma. La asociación conseguida en un conjunto de accesiones pudiera romperse durante el mejoramiento genético de las mismas, pues se requerirían cruzamientos y selección de materiales que modificarían las frecuencias genotípicas iniciales de la población.

El *DL* se rompe con la recombinación en los dobles heterocigotos. Estos genotipos poseen frecuencias mayores en poblaciones con preferencia de polinización cruzada que en preferentemente autógamas, por lo que se espera, en general, que las poblaciones autógamas presenten mayor *DL* que las alógamas. En otras palabras, se pueden conseguir una mayor proporción de bloques de ligamiento en las autógamas que en las alógamas.

2.7 Estructura genética de las poblaciones

La estructura genética de las poblaciones está relacionada con la presencia de divisiones dentro de las mismas. Estas divisiones pueden originarse por barreras geográficas, temporales y biológicas. A continuación se considera una gran población compuesta de múltiples poblaciones (subpoblaciones o demos) producto de la división, donde existe un cierto grado de interacción entre ellas.

Se han propuestos distintos modelos que tratan de explicar la estructura genética de una población. El más simple es el modelo de islas, en el cual se asume que cada subpoblación recibe inmigrantes del resto, a una tasa constante en el tiempo (*m*), sin importar las distancias geográficas. En este sentido es poco realista. Otro modelo incorpora la distancia geográfica entre subpoblaciones, dando lugar al modelo de piedras de paso. Si se considera que los individuos de una población se encuentran distribuidos en un continuo espacial, sin fraccionamiento evidente en ella, el modelo de aislamiento por distancia es el más adecuado.

La estructura genética de una población puede explicarse en términos generales por la acción conjunta de la migración, selección, mutación y deriva genética. Ya que la mayoría de los marcadores genéticos se consideran neutralmente selectivos, probablemente no detecten el efecto de la selección.

Otro punto que se acostumbra a tratar es la interacción entre los individuos dentro de una subpoblación. Es de interés determinar si existe EHW o en estimar el coeficiente de endocría en función del sistema de apareamiento de esos individuos.

Estadísticos F

Wright propuso una medida de diferenciación genética entre subpoblaciones en función de las varianzas de sus frecuencias alélicas. Esta medida, denominada F_{ST}, está relacionada con la correlación entre las gametas de individuos distintos dentro de una subpoblación. En este sentido, F_{ST} es analógo a una correlación intraclase, por lo que es equivalente a la variabilidad entre las medias de las subpoblaciones, que en este caso son las frecuencia alélicas. En una población, viene dado como

$$F_{ST} = S^2_p / p\,(1\text{-}p).$$

donde S^2_p es la varianza de las frecuencias del alelo en las distintas subdivisiones. El denominador representa a la varianza máxima posible, la cual ocurriría si todas las subpoblaciones tuviesen un alelo fijado. Para cada alelo existirá un estimado de F_{ST}, por ello, más adelante se mostrará una manera de agruparlos.

F_{ST} también se le conoce como *índice de fijación*, pues mide el grado en que las subpoblaciones tienden a fijar alelos particulares. Se considera que dentro de cada subpoblación hay apareamientos al azar y que no existen mutación ni selección, por lo tanto, está afectado exclusivamente por la deriva genética. De allí el papel importante que juega el tamaño de las subpoblaciones en determinar su valor. No obstante, en las poblaciones naturales estos factores están presentes, dificultando la interpretación, aunque no invalida la utilidad de este coeficiente como un índice de diferenciación genética.

Debido a que la división subpoblacional establece un efecto parecido a la endocría, fue necesario diferenciar la endocría producida por apareamientos consanguíneos (f ó F_{IS}) de la generada por deriva genética (F_{ST}). La primera se restablece si los individuos dentro de una subpoblación se dejan aparear al azar, la segunda no, por lo que F_{ST} produce efectos irreversibles en la disponibilidad de variabilidad genética la evolución, como efectos drásticos en el fenotipo medio de la subpoblación por depresión por endocría. La única manera de contrarrestar el efecto de la deriva es el cruzamiento de individuos de subpoblaciones distintas.

Los coeficientes F_{IS} y F_{ST} se pueden agrupar en uno solo, o coeficiente de endocría total (F_{IT}). En una población con tres niveles jerárquicos la relación entre ellos viene dada por

$$(1\text{-}F_{IT}) = (1\text{-}F_{IS})\,(1\text{-}F_{ST})$$

Esta relación se puede extender a cualquier jerarquía, agregando nuevos términos a lado derecho de la ecuación. Suponiendo que se tiene la siguiente jerarquía: Subpoblación (S) dentro de Región (R) dentro de la Población (T), los estadísticos F quedarían

$$(1-F_{IT}) = (1-F_{IS})\,(1-F_{SR})\,(1-F_{RT})$$

Como medida de variabilidad genética, F_{IS} no tiene una utilidad clara, pues indica la variación que hay entre individuos dentro de una Subpoblación. En cambio, F_{SR} y F_{RT} indican el grado de variabilidad genética en la población Total debido a la división entre Subpoblaciones y Regiones.

Algunas características de los estadísticos F para una jerarquía simple son:

Coeficiente de endocría **(F_{IS})**: mide la reducción de heterocigosidad de un individuo debido a apareamientos no aleatorios dentro de su subpoblación. Puede ser negativo, ya que es una correlación, lo cual ocurriría si se evitan los apareamientos consanguíneos dentro de cada subpoblación.

Índice de fijación **(F_{ST})**: mide el efecto de la subdivisión poblacional a través de la reducción de heterocigosidad en una subpoblación debido a la deriva genética. Siempre es positivo, ya que es la relación de dos varianzas. F_{ST} puede variar entre 0 y 1. Wright sugiere que valores entre

0 - 0,05	indican una diferenciación genética (DG) pequeña
0,05 - 0,15	indican una DG moderada
0,15 - 0,25	indican una DG grande
más de 0,25	indican una DG muy grande

Sin embargo, el autor aclara que un F_{ST} menor que 0,05 no es un valor despreciable.

Coeficiente de endocría total de un individuo **(F_{IT})**: incluye una contribución de la ausencia de panmixia dentro de la subpoblación (F_{IS}) y otra de la subdivisión (F_{ST}). En otras palabras, mide la reducción en la heterocigosidad de un individuo con respecto a la población total.

Los estadísticos *F* se pueden estimar de la siguiente manera:

$$F_{IS} = 1 - H_I/H_S$$

$$F_{ST} = 1 - H_S/H_T$$

$$F_{IT} = 1 - H_I/H_T$$

donde

$\boldsymbol{H_I}$ es la heterocigosidad de un individuo en la subpoblación. Se puede interpretar como la heterocigosidad promedio de todos los genes de un individuo o como la probabilidad de heterocigosidad de cualquier gen.

$\boldsymbol{H_S}$ es la heterocigosidad esperada en un individuo en una x-ésima subpoblación equivalente bajo panmixia. Para dos alelos, este parámetro viene dado como $H_S=2p_x\,q_x$,

$\boldsymbol{H_T}$ es la heterocigosidad esperada en un individuo en la población total bajo panmixia. Representa la heterocigosidad si todas las poblaciones se aparean al azar; si la frecuencia alélica promedio entre subpoblaciones es p_0, entonces $H_T=2p_0q_0$.

Nei y Chesser (1983) propusieron un método de estimación de *F* similar al anterior, pero incorpora modificaciones en el cálculo de H_I, H_S y H_T para considerar el error de muestreo.

Los estadísticos *F* también se puede interpretar como la probabilidad de que dos genes tomados al azar provengan de un ancestro común, término conocido como *endocría*. F_{IS} mide la endocría producto de apareamientos consanguíneos dentro de las Subpoblaciones, por ejemplo, cuando ocurre autofecundación o cruzamientos entre hermanos. F_{SR} y F_{RT} miden la endocría producto de la deriva genética, que a su vez puede interpretarse como el producto de apareamientos consanguíneos lejanos que tiende a incrementarse en Subpoblaciones y Regiones pequeñas en tamaño.

Complementariamente, $1\text{-}F=P$, denominado *Índice Panmíctico*, es la probabilidad de que dos genes tomados al azar no provengan de un ancestro común. De esta forma, $1\text{-}F_{IT} = P_{IT}$ es la probabilidad de que los dos genes dentro de un individuo tomado al azar de la Población Total no provengan de un ancestro común. Si se considera que

$$P_{IT} = P_{IS}P_{SR}P_{RT} = (1\text{-}F_{IT}) = (1\text{-}F_{IS})\ (1\text{-}F_{SR})\ (1\text{-}F_{RT})$$

P_{IT} viene siendo el producto de la probabilidad de que este individuo escape a la endocría por apareamientos consanguíneos dentro de su Subpoblación (P_{IS}) y de la probabilidad de que escape simultáneamente a la endocría por efecto de la deriva en las Subpoblaciones (P_{SR}) y en las Regiones (P_{RT}).

Cockerham (1969, 1973) propuso una forma distinta de obtener estimados relacionados con los estadísticos *F*. Este método está explícitamente discutido en Weir y Cockerham (1984). El método se basa en el análisis de la varianza, donde la variable estudiada es la presencia o ausencia de un alelo particular. Si se considera que a_{ij} representa el j-ésimo alelo del i-ésimo individuo, la frecuencia de este alelo x_{ij} es definida como

$$x_{ij} = 1 \quad \text{si} \quad a_{ij} = A$$
$$x_{ij} = 0 \quad \text{si} \quad a_{ij} = no\ A$$

A partir de esta definición se estimaron las correlaciones entre alelos: dentro de individuos, entre individuos dentro de subdivisiones (θ) y entre individuos de toda la población (*F*). Para ello se usan en las covarianzas el coeficiente de endocría y el coeficiente de consanguinidad. Las correlaciones se relacionan con los componentes de la varianza de la siguiente manera

$$(1 - F)\ p\ (1\text{-}p) = \sigma^2_w$$

$$(F - \theta)\ p\ (1\text{-}p) = \sigma^2_b$$

$$\theta\ p\ (1\text{-}p) = \sigma^2_a$$

$$\sigma^2_t = \sigma^2_w + \sigma^2_b + \sigma^2_a = p\ (1\text{-}p)$$

siendo *p* la frecuencia del alelo en la población y σ^2_w, σ^2_b y σ^2_a, los componentes de varianzas asociados con las diferencias entre alelos dentro de individuos, entre individuos

dentro de las subpoblaciones y entre subpoblaciones, respectivamente. Despejando de las expresiones anteriores se obtienen los análogos de los estadísticos *F*

$$F = [\sigma^2_b + \sigma^2_a] / [\sigma^2_w + \sigma^2_b + \sigma^2_a] = F_{IT}$$

$$\theta = [\sigma^2_a] / [\sigma^2_w + \sigma^2_b + \sigma^2_a] = F_{ST}$$

$$(F - \theta) / (1 - \theta) = [\sigma^2_b] / [\sigma^2_w + \sigma^2_b] = F_{IS}$$

Los estimados de estos parámetros están reseñados en Weir y Cockerham (1984) para variaciones en el tamaño de las muestras y el número de subpoblaciones.

A diferencia de la definición dada por Wright para F_{ST}, θ puede medir cualquier diferenciación genética sin importar la causa que la origine. Entre ellas se pueden mencionar no sólo la deriva genética, sino la selección natural dentro y entre subpoblaciones y la migración.

De las tres expresiones anteriores se desprende que θ es la proporción de la variación total debida a diferencias entre subpoblaciones; mientras que el análogo a F_{IT}, la proporción debida a diferencias entre individuos dentro de subpoblaciones. Cuando existen apareamientos al azar dentro de las subpoblaciones, $F = \theta$, lo cual significa que no se debería esperar variabilidad entre individuos dentro de una subpoblación.

Estos parámetros son estimados para cada alelo y para distintos loci. Por lo tanto, el siguiente paso es agruparlos y finalmente realizar las pruebas de hipótesis pertinentes. Un agrupamiento que genera estimados insesgados por locus $\theta\wedge_w$ es

$$\theta\wedge_w = \Sigma_u a_u / \Sigma_u (a_u + b_u + w_u)$$

siendo *a*, *b* y *c* los componentes de varianzas relacionados con el u-ésimo alelo. Cuando se consideran distintos loci, el agrupamiento se haría según

$$\theta\wedge_w = \Sigma_l \Sigma_u a_{lu} / \Sigma_l \Sigma_u (a_{lu} + b_{lu} + w_{lu})$$

con la única diferencia de que ahora se le añade un componente relacionado con el l-ésimo locus. Con el uso de las técnicas de remuestreo (bootstrapping o jackknife) se pueden obtener los intervalos de confianza para los estimados de f (F_{IS}), θ (FST) y F (F_{IT}). En este caso, la unidad de muestreo pueden ser los loci o las subpoblaciones.

Implícito en el método propuesto por Cockerham (1969, 1973) está la idea de muestreo genético, adicional a la de muestreo estadístico. La variación asociada con el muestreo genético se origina del muestreo inherente en la transmisión del material genético de padres a hijos. En replicas distintas, diferentes genes pudieron transmitirse por simple azar. Una subpoblación cualquiera de tamaño N, de donde se obtiene una muestra, se considera como una réplica de infinitas subpoblaciones independientes que se originaron de la misma población original de referencia.

La atención que se le debe dar a cada tipo de variación dependerá de los objetivos del estudio. Como señala Weir (1996), si hay interés en una población, es posible ignorar el muestreo genético que ha hecho que esta población sea diferente a las demás. Por otra parte, si se quiere concluir sobre todas las poblaciones réplicas originadas bajo circunstancias similares, debe considerarse el muestreo genético.

Nei (1986) difiere de Cockerham (1969) ya que considera difícil que en la naturaleza se encuentren poblaciones muy grandes, constantes en tamaño a través del tiempo. El modelo de Cockerham puede incorporar el efecto de migración, mutación y selección, siempre y cuando se mantenga la independencia entre subpoblaciones; algo improbable. En este sentido, el método Nei y Chesser (1983) abandona las suposiciones anteriores y reformula los estadísticos *F* en función de las subpoblaciones existentes. No se requieren por lo tanto, suposición alguna sobre los factores que han causado la diferenciación entre subpoblaciones, ni considerar poblaciones replicadas.

Índices de diversidad génica de Nei

Nei (1973) plantea una forma distinta de evaluar la diferenciación genética entre subpoblaciones. La variación de las frecuencias alélicas puede analizarse en términos de de heterocigosidad o bien como diversidad génica. Este método es independiente de fuerzas evolutivas como mutación, selección y migración. De igual forma, se puede aplicar a cualquier número de alelos, individuos diploides o monoploides, de reproducción sexual o asexual, siempre y cuando se dispongan de las frecuencias alélicas.

El fundamento del método de Nei consiste en que la diversidad génica total (H_T) de la población se puede analizar en componentes de la diversidad génica dentro (H_S) y entre (D_{ST}) subpoblaciones. Conceptos como diversidad génica e identidad génica ya fueron discutidos en el caso de una población.

El cálculo de cada componente se logra de la siguiente manera:

considere

$$J_i = \Sigma_k \, p_{ki}^2$$

como la identidad génica dentro de una subpoblación, donde p_{ki}^2 es la frecuencia del k-ésimo alelo en la i-ésima subpoblación,

$$J_S = \Sigma_i \, J_i / s$$

como la identidad génica subpoblacional promedio (s = nro. de subpoblaciones),

$$J_T = \Sigma_k \, p_{.k}^2$$

como la identidad génica en toda la población, donde

$$p_{.\,k} = \Sigma_i \, w_i \, p_{ki} / s,$$

siendo w_i es un ponderado para la i-ésima subpoblación $(\Sigma_i w_i = 1)$

que también se puede calcular como

$$J_T = \Sigma_k (\Sigma_i w_i p_{ik})^2$$

Por complemento,

$$H_T = 1 - J_T$$

$$H_S = 1 - J_S$$

$$D_{ST} = H_T - H_S$$

mientras que la magnitud de diferenciación genética entre subpoblaciones en relación a la Población Total es

$$G_{ST} = D_{ST}/H_T$$

el cual varía entre 0 y 1 y se le puede llamar Coeficiente de diferenciación genética. G_{ST} es equivalente a F_{ST} en que no son negativos. Ambos son idénticos cuando hay dos alelos por locus; en caso contrario, G_{ST} es una media ponderada de F_{ST} para todos los alelos.

Finalmente se establece la siguiente relación:

$$H_T = H_S + D_{ST}$$

Al igual que los estadísticos *F*, está relación puede extenderse para incluir otros niveles de jerarquía. Cuando se trata de Subpoblaciones (*S*) dentro de Regiones (*R*) dentro de una Población Total (*T*)

$$H_T = H_S + D_{SR} + D_{RT}$$

En donde D_{SR} representa la diversidad génica entre Subpoblaciones y D_{RT} la diversidad entre Regiones. Cada uno de los tres términos se puede dividir entre H_T para obtener la proporción que representan de la diversidad total.

Según el método de Nei, las medias aritméticas de los estimados de H_T y H_S se obtienen para todos los loci disponibles. De estos promedios se obtiene D_{ST} y finalmente G_{ST}. Si en cambio, se obtienen los G_{ST} para cada locus y se promedian, el resultado es diferente.

Para conocer en forma global la diferenciación genética entre subpoblaciones, es conveniente *utilizar un gran número de loci, incluyendo los monomórficos.* Ello genera un muestreo representativo del genoma.

2.8 Otras alternativas para el estudio de la diversidad genética

En el caso de que no se disponga de información alélica, la diversidad genética puede estimarse de patrones electroforéticos. Estos registros consideran la proporción de patrones o de bandas dentro de las subpoblaciones.

Un índice que permite descomponer la diversidad genética dentro y entre poblaciones es el de Índice de Información de Shannon *H'* (Hutcheson, 1970); se obtiene de la siguiente manera

$$H' = \Sigma_{i=1}^{m} = p_i \ln p_i,$$

en el que pi es la frecuencia del I-ésimo patrón o banda de un gel. *H'* se puede agrupar por caracteres, países, etc, y sus componentes pueden ser detectados estadísticamente mediante un análisis de varianza. Esto es posible, gracias a que en muestras relativamente grandes, la distribución de *H'* se aproxima a la normal.

Excoffier *et al.* (1992) propusieron una adaptación del método de Cockerham (1969) para obtener análogos de los estadísticos *F*. En vez de utilizar la presencia o ausencia de alelos, se usa la presencia o no de haplotipos. El método se conoce como Análisis de Varianza Molecular (AMOVA). Otros autores posteriormente han usado este método con patrones (haplotipos) RAPD, ya que la dominancia completa de los mismos impiden utilizar los métodos de Cockerham (1969) y Nei (1973).

2.9 Semejanza genética entre individuos/poblaciones

Para estimar la semejanza entre pares de unidades biológicas (UB) es conveniente adoptar el arreglo de datos en la forma de una matriz $n \times t$, donde *t* representa las diferentes UB cuyas semejanzas se obtienen en función de *n* caracteres

	UB		
Caracteres	**1**	**2 ...**	**... t**
1	X_{11}	X_{12}	X_{1t}
2	X_{21}	X_{22}	X_{2t}
.			.
.			.
n	X_{n1}	X_{n2}	X_{nt}

El uso de los coeficientes de semejanzas está limitado por la escala de la medida de las variantes del carácter evaluado. Por ejemplo, no es conveniente utilizar un coeficiente para datos cuantitativos con registros binarios. En otros casos, algunos coeficientes necesitan información genética (frecuencias alélicas, p.e.); mientras que en otros, se requiere información fenotípica.

Para fines prácticos, los coeficientes de semejanzas se pueden dividir en:
de distancias: miden la distancia entre UB en un espacio definido.

de asociación: miden la correspondencia de las variantes de los caracteres; en casos especiales pueden ser considerados distancias.

de correlación: miden la proporcionalidad e independencia entre pares de UB.

Coeficientes de distancias

Solamente se hará referencia a dos distancias genéticas: la Distancia Genética de Nei (*D*; Nei 1972) y la distancia cuadrada promedio *D1* de Goldstein *et al*. (1995). La distancia *D* ha sido empleada extensivamente; sin embargo, con la disponibilidad de los marcadores microsatélites fue necesario buscar nuevos estimadores que tomaran en cuenta el patrón de mutación 'de paso' (Stepwise Muattion Model: SMM) de los microsatélites. La distancia genética de Nei se adapta a marcadores que mutan según el Modelo de Infinitos Alelos (IAM), en el cual cada mutación genera un alelo completamente distinto al anterior. En cambio, en el SSM cada mutación puede originar alelos previamente existentes, ya que ocurren de acuerdo a la adición o eliminación de secuencias repetitivas. La variación de secuencias repetitivas parece seguir una distribución normal, en el que un mayor número de adiciones o eliminaciones se hacen cada vez menos probables, independientemente del número de repeticiones ya existentes.

La importancia de utilizar correctamente ambas distancias se hace evidente cuando se quiere estimar el tiempo de divergencia entre los miembros de un grupo. En estudios de simulación, Takezaki y Nei (1996) han conseguido que ambas distancias, bajo IAM y SMM, respectivamente, se incrementan linealmente a medida que el tiempo transcurre. Es importante considerar que ambas distancias tienden a presentar coeficientes de variación relativamente grandes, lo cual disminuye la eficiencia de caracterizar la topología verdadera del árbol. Por lo tanto, si ese es el objetivo, otras distancias, como D_A de Nei y la D_C de Cavalli-Sforza y Edwards deberían considerarse.

Distancia genética de Nei (D)

La Distancia genética de Nei *D* (Nei 1972) entre poblaciones está expresada en función de sus frecuencias génicas. Desde el punto de vista genético, la medida más apropiada de distancia es el número de diferencias entre nucleótidos (o codones) por unidad de longitud de ADN. Ya que es difícil disponer de la secuencia de todo el genoma, Nei desarrolló un método estadístico en el cual el Número Promedio de Diferencias entre codones por Locus se puede estimar a partir de datos de frecuencia génica.

Si se supone que:

$$j_i = \Sigma p_{ki}^2,\ j_y = \Sigma p_{ky}^2\ ,\ j_{iy} = \Sigma p_{ki} p_{ky}$$

entonces

$$J_i = \Sigma_r j_{i\,(r)} / r\ ,$$

donde *r* es el número de loci

$$J_y = \Sigma_r j_{y\,(r)} / r$$
$$J_{iy} = \Sigma_r j_{iy\,(r)} / r$$

La proporción de loci (alelos) diferentes escogidos de dos genomas al azar en las poblaciones respectivas sería:

$$D_{i(m)} = 1\text{-}J_i$$
$$D_{y(m)} = 1\text{-}J_y$$

mientras que el estimado mínimo de diferencias entre codones entre dos genomas escogidos al azar, uno de cada población es

$$D_{iy(m)} = 1\text{-}J_{iy}$$

El estimado Mínimo de Diferencias Netas entre Codones por Locus cuando se extrae las diferencias intrapoblacionales se obtiene por

$$D_m = D_{iy(m)} - [D_{i(m)} + D_{y(m)}]/2$$

D_m se conoce como *Distancia Genética Mínima.* Debe considerarse como un estimado aproximado del Número de Diferencias Netas entre Codones cuando $D_{xy(m)}$ es grande.

Si los cambios en los codones son independientes, el número medio de diferencias "**netas**" entre codones se puede obtener por

$$D = -\log_e I \text{ (Distancia Genética Estándar)},$$

donde

$$I = J_{xy}/(J_x J_y)^{1/2}$$

Es la identidad normalizada.

Existe otra medida, la *Distancia genética máxima* (*D´*), la cual considera errores de muestreo y deriva genética. En todo caso, *D* es preferible que *D´*, además de que puede ser utilizada en el estudio de comparaciones de individuos entre especies como dentro de ellas.

Distancia D1

La distancia *D1* (Goldstein *et al.*, 1995) parte de un SMM estricto, donde sólo se permiten ganancias o pérdidas de una sola repetición, con una probabilidad de $\mu/2$, independientes de la cantidad de repeticiones *i* que existan en ese alelo. *D1* se puede definir como la diferencia al cuadrado promedio en números de repeticiones de dos alelos tomados de diferentes poblaciones, aisladas por t generaciones en el pasado.

Un estimador de *D1* en términos del número de repeticiones entre dos alelos i, i' cualesquiera sería

$$\Delta_{ii'} = (i - i')^2.$$

El promedio (Δm) de $\Delta_{ii'}$ entre todos los alelos muestreados, uno de cada población, es un estimador insesgado de *D1*. así

$$\Delta m = \Sigma_r \Sigma_i \Sigma_{i'} (i - i')^2 f_i f_{i'} / r$$

donde, f_i f_i son las frecuencias alélicas de i y i' en las muestras de la primera y segunda población, respectivamente y *r* el número de loci.

Existe otra distancia genética relacionada que tiene una menor varianza, calculada como

$$(\delta\mu)^2 = \Sigma_r [(\Sigma_i\, i\, f_i) - (\Sigma_{i'}\, i'\, f_{i'})]^2 / r$$

Distancia euclideana

Esta distancia no posee una interpretación genética clara. Ha sido empleada para manipular información obtenida de frecuencias alélicas. Para las poblaciones j y k, se obtiene a través de la fórmula

$$D_{jk} = [\Sigma_i^n (X_{ij} - Y_{ik})^2]^{1/2}$$

En la que X_{ij} y X_{ik} son las frecuencias del i-ésimo alelo en la j-ésima y k-ésima población, respectivamente. Dado que D_{jk} aumenta con el número de caracteres utilizados en la comparación, se utiliza una distancia promedio, igual a

$$djk = (D_{jk}^2 / n)^{1/2}$$

2.9 Coeficientes de asociación

Existen muchos coeficientes de asociación. En los modelos más comunes, los coeficientes de asociación se obtienen a partir de caracteres binarios, codificados por conveniencia como ´1´ y ´0´. Esta nomenclatura puede utilizarse para expresar la respectiva presencia versus ausencia de una variante genética en un locus. Las combinaciones de la presencia/ausencia de dicha variante cuando se comparan dos UB se expresan en el siguiente cuadro

		UB J	
		1	**0**
UB K	**1**	*a*	*b*
	0	*c*	*d*

Cada coeficiente hace uso distinto de las combinaciones expresadas en la tabla anterior.

Entre los coeficientes de mayor uso se tienen:

Coeficiente de Jaccard (S_J): varía entre 0 y 1. En su clase, es el más simple de los coeficientes. Omite los apareamientos negativos. Es monotónico con el coeficiente de Dice. Se calcula de la siguiente manera:

$$S_J = a / (a+b+c)$$

Coeficiente de Dice (S_D): está relacionado con el coeficiente de Jaccard. Otorga el doble de importancia a los apareamientos positivos (a) que a los negativos (d). Se obtiene a través de

$$S_D = 2a/(2a+b+c)$$

Coeficiente de apareamiento simple (S_{SM}): cuando $1\text{-}S_{SM}$ tiende a cero o a uno, es igual a la distancia euclideana. Por tal motivo, la raíz del complemento de este coeficiente ($1\text{-}S_{SM}$) es

una función métrica que puede representarse en un espacio euclideano. Se calcula como

$$S_{SM} = a+d / (a+b+c+d)$$

2.10 Procesamiento de las distancias/coeficientes

Los valores de las distancias/coeficientes se agrupan en una matriz cuyas filas y columnas se refieren a las UB. El orden de las UB es el mismo tanto en filas como en columnas, por lo que los valores en la diagonal principal representan la comparación de un UB consigo mismo. La matriz es cuadrada, de orden *t x t*, donde t es el número de UB. La matriz se expresa como

			UB		
UB	**1**	**2**	**3**	**. . .**	***t***
1	S_{11}	S_{12}	S_{13}		S_{1t}
2	S_{21}	S_{22}	S_{23}		S_{2t}
3	S_{31}	S_{32}	S_{33}		S_{3t}
.	.	.	.		.
.	.	.	.		.
.	.	.	.		.
t	S_{t1}	S_{t2}	S_{t3}	. . .	S_{tt}

La matriz es simétrica, ya que la semejanza entre a y b es la misma que entre b y a, así que por convención se utiliza el triángulo inferior izquierdo. Para establecer patrones (cualquier propiedad descriptiva en la distribución de los UB) a partir de la matriz de semejanzas, se utiliza por lo general algoritmos matemáticos. Algunos de estos algoritmos generan "cluster" o grupos.

El procedimiento de agrupamiento consiste en obtener una o más particiones de un conjunto de UB. El algoritmo de Ligamiento Promedio UPGMA (Unweighted Pair Group Method using Arithmetic Average) requiere de una especie de similitud o disimilitud promedio entre una UB (o cluster) candidata y un cluster ya existente. Cada UB en ese cluster es ponderado equitativamente, independientemente de su subdivisión estructural.

La disimilitud U_{jk} entre dos cluster cualesquiera *j* y *k* se define como

$$Ujk = (1 / t_j t_k) \Sigma_{jk} U_{jk} ,$$

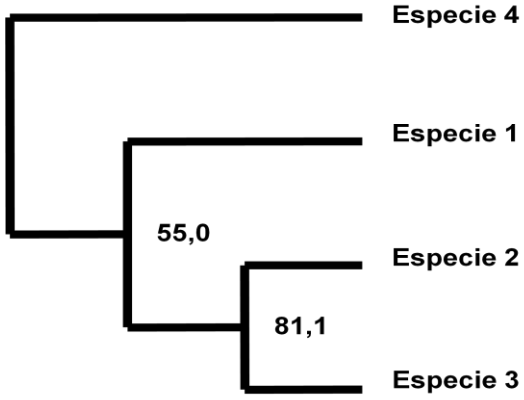
donde *t* representa el número de UB en cada cluster. El algoritmo trabaja de la siguiente manera (ver ejemplo en los anexos):

1. Se buscan los grupos (cluster) menos disímiles, los cuales formarán un nuevo grupo. La disimilitud entre UB que no se unan en nuevos cluster permanecen sin cambio alguno de la matriz original.

2. A partir de la primera matriz agrupada, se buscan nuevos pares menos disímiles de la misma manera que en la matriz original. Aquí entra en juego el criterio *no ponderado*.

3. Se generan nuevas matrices hasta conseguir una matriz con un valor final de disimilitud.

4. Estos agrupamientos jerarquizados se expresan gráficamente en un dendrograma o árbol filogenético. Para evaluar la consistencia del agrupamiento o clado, respectivamente, se recurre a la técnica de re-muestreo (bootstrapping) para generar múltiples árboles a partir de la matriz de datos original. La unidad de muestreo son los loci. En cada nudo se presenta la proporción de árboles que mostraron dicho agrupamiento/clado, valor que se conoce como valor-P. En el dendrograma siguiente se observa que el agrupamiento/clado entre las especies 2 y 3 se presentó en un 81,1 % de los árboles. Si bien el valor-P es alto, es recomendable que esté por encima del 95%. El agrupamiento/clado de la especie 1 con las dos anteriores no fue muy robusto, con un 55%. Ello indica que este grupo de tres especies no es consistente, existiendo la posibilidad de que la especie 1 forme por sí sola un grupo.



2.11 Métodos de ordenamiento

A veces es conveniente complementar la información de clasificación, obtenida a partir del dendrograma, con un método de ordenamiento. Es recomendable hacer un cluster como primer paso, y si el fenograma parece poco satisfactorio, realizar un estudio de ordenamiento. Estos métodos, preferiblemente en tres dimensiones, son importantes para entender la estructura poblacional con mayor detalle. El método de ordenamiento más utilizado es el Análisis de Componentes Principales (ACP), a pesar de que existen otros como el Análisis de Coordenadas Principales que usa la matriz de coeficientes de semejanza.

El ACP un método multivariado utilizado principalmente para reducir la cantidad de información. Requiere de variables con distribución normal (multinormal); no obstante, en la mayoría de los casos se hace con fines descriptivos, sin considerar pruebas de hipótesis, por lo que se omite la multinormalidad de las variables.

Si se considera los registros de una muestra de n individuos para cada una de la p variantes genéticas (para todos los loci)

$$X_j \ (j = 1,...,p)$$

los registros se pueden escribir en notación matricial como

$$X^* = [X_{ij}] \ ,$$

donde X_{ij} denota el registro de la j-ésima variable para el i-ésimo individuo. Si los registros de X^* se expresan como desviaciones de su respectiva media, se origina la matriz

$$X = [x_{ij}].$$

Si $X'X$ es la suma de cuadrados corregida y los productos cruzados, la matriz de varianza-covarianza viene dada por

$$S = (X'X) / (n-1).$$

El objetivo del ACP es transformar la matriz X de p variantes, las cuales pueden estar correlacionadas, en otra matriz Y de p variables hipotéticas no correlacionadas, cuyas varianzas disminuyen de la primera a la última. Esto se logra utilizando una matriz ortogonal U

$$Y = XU$$

donde las columnas de U son los vectores latentes normalizados de la matriz $S(pxp)$, arreglados de tal manera que el primer vector posee la *Raíz Latente* (λ) mayor de S, el segundo la segunda raíz, y así sucesivamente.

Si y_1 es la nueva variable, correspondiente a la primera columna de Y, representa por lo tanto el primer Componente Principal (CP). Éste es una suma ponderada de las variables x_j(j=1,...,p) y se puede escribir como

$$y_1 = x_1u_{11} + x_2u_{21} + ... + x_pu_{p1}$$

donde u_{j1} es el peso dado a la j-ésima variable en este componente. La varianza de y_1 es

$$Var\ (y_1) = \lambda_1 \text{ (o autovalor)}$$

Siendo λ_1 es la raíz latente mayor.

Cuando un conjunto de variables observadas están relacionadas, los primeros CP derivados de ellas explican una parte grande de su varianza; así que, sin una pérdida apreciable de la información original, las variantes pueden remplazarse por un conjunto pequeño de variables derivadas.

El ACP se caracteriza por ser una representación confiable de las distancias entre grupos (cluster) mayores; pero es notorio su sesgo cuando se tratan vecinos cercanos. En este aspecto es contradictorio al UPGMA, el cual generalmente reproduce fielmente las distancias entre elementos cercanos; pero distorsiona las distancias entre miembros de grupos grandes, p.e. a nivel de los troncos del dendrograma.

Referencias bibliográficas

Cockerham C. 1969. **Variance of gene frequencies**. Evolution 23:72-84.

Cockerham C. 1973. **Analyses of gene frequencies**. Genetics 74: 679-700.

Excoffier L., Smouse P. & J.M. Quattro. 1992. **Analysis of Molecular Variance Inferred From Metric Distances**

Among DNA Haplotypes: Application to Human Mitochondrial DNA Restriction Data. Genetics 131: 479-491.

Goldstein D., Ruiz-Linares A., Cavalli-Sforza L.L. & M.W. Feldman. 1995. **An evaluation of genetic distances for use with microsatellite loci.** Genetics 139: 463-471.

Hutcheson K. 1970. **A test for comparing diversities based on the Shannon formula**. J. Theor. Biol. 29:151-154.

Nei M. 1972. **Genetic distance between populations**. Amer. Naturalist 106: 283-292.

Nei M. 1973. **Analysis of gene diversity in subdivided populations**. . Proc. Natl. Acad. Sci. 70: 3321-3323.

Nei M. & R.K. Chesser. 1983. **Estimation of fixation indices and gene diversities**. Ann. Hum. Genet. 47: 253-259.

Nei M. 1986, **Definition and estimation of fixation indices**. Evolution 40: 643-645.

Takezaki N. & M. Nei. 1996. **Genetic distances and reconstruction of phylogenetic trees from microsatellite DNA**. Genetics 144: 389-399.

Weir B. y Cockerham C. 1984. **Estimating F-statistics for the analysis of population structure**. Evolution 38: 1358-1370.

Referencias Generales:

Brown A. & B. Weir. 1983. **Measuring genetic variability in plant populations**. In: Isozymes in plant genetics and breeding. (Tanksley y Orton Eds.). Elsevier. p. 219-255.

Hartl D. 1988. **A primer of population genetics**. Sinauer Associates Inc. cap 1 y 2.

Maxwell A. 1977. **Multivariate analysis in behavioural research**. Chapman and Hall. Cap. 4.

Nei M. 1975. **Molecular population genetics and evolution**. North-Holland Publishing Company. cap 6 y 7.

Sneath P. & R. Sokal. 1973. **Numerical taxonomy**. Freeman and Company. cap 4 y 5.

Weir B. 1996. **Genetic Data Analysis II**. **Methods for discrete population genetic data**. Sinauer Associates, Inc. Publishers Sunderland, Massachusetts.

2. MAPEO CROMOSÓMICO

2.1. Modo de herencia de caracteres monogénicos

Los datos observados se confrontan con una segregación teórica esperada en función del grado de dominancia intra-alélica. En plantas, se usan preferentemente poblaciones segregantes como la F_2, Retrocruzas (Rc), Dobles Haploides (DH) y Líneas Endocriadas Recombinantes (RIL), cuya segregación fenotípica se conoce de antemano. En el caso de una población de individuos diploides, donde los padres poseen los genotipos aa y bb y los fenotipos A y B, respectivamente, se esperarían las siguientes proporciones fenotípicas (genotípicas)

Población	codominancia	dominancia completa
F_2	1A : 2 AB : 1 B (1 aa : 2 ab : 1 bb)	3 A : 1 B (1 aa + 2 ab) : 1 bb
Rc (F1 X P2)	1 AB : 1 B (1 ab : 1 bb)	1 A : 1 B (1ab : 1 bb)
DH RIL	1 A : 1 B (1 aa : 1 bb)	1 A : 1 B (1 aa : 1 bb)

El ajuste de los datos observados a los esperados se establece por el método estadístico de bondad de ajuste χ^2, donde la distribución χ^2 es una razón entre la suma de cuadrados entre de los datos observados y una varianza teórica fijada (Mather 1967). El estadístico χ^2 se calcula de la siguiente (Mather 1957):

$$\chi^2 = \Sigma^k (n_i - e_i)^2 / e_i$$

donde

k el número de genotipos descendientes distinguibles
n_i el número de descendientes con el *i-ésimo* genotipo
e_i, el número esperado correspondiente.

La divergencia χ^2 no debería exceder un umbral preestablecido para no rechazar la hipótesis nula. El umbral se obtiene de la correspondiente probabilidad (valor probabilístico o valor-p) de ocurrencia a partir de una distribución χ^2 con *k-1* grados de libertad. Para un error tipo I del 5 % (valor-p), el umbral de rechazo para tres clases (2 gl) es de 5,99 y de 3,84 para dos clases (1 gl). Un valor-p bajo (correspondiente a un χ^2 por encima del umbral) sugiere que la segregación en ese locus está distorsionada. La desviación pudiera ser causada por el azar y/o por selección en alguno de los alelos.

2.2 Ligamiento genético

El ligamiento genético entre dos loci es una asociación que se establece a nivel de la meiosis, específicamente durante la fase de paquitene. Cuando se consideran poblaciones experimentales como una F_2 o una retrocruza, dos genes presentan asociación si se encuentran en un mismo cromosoma. Esto permite definir un grupo de ligamiento como aquel conjunto de genes en el cual existe al menos una asociación entre un gen y cualquier otro. Teóricamente, un

cromosoma se corresponde con un grupo de ligamiento; no obstante, por causa de artificios experimentales puede haber dos o más grupos de ligamiento asignados a un cromosoma. A medida que se amplia la cobertura del genoma, con un mayor número de marcadores (saturación), se llega a la condición teórica esperada.

En todo estudio de mapeo se buscan dos objetivos:

- detección de ligamiento
- estimación del grado de asociación (ligamiento)

Ambos no son necesariamente secuenciales, pues algunos programas de computación detectan y estiman el grado de ligamiento simultáneamente.

Inferir el comportamiento de una característica de interés a través de los marcadores moleculares es una herramienta que hace óptimo el proceso de diagnóstico y selección de variantes específicas. La Selección Asistida por Marcadores (Marker Assisted Selection) es una metodología que establece, por lo tanto, asociaciones entre dos características: el marcador (variable predictora) y el carácter de interés (variable marcada).

2.3 Detección de ligamiento

Prueba de ligamiento (bondad de ajuste χ^2)

Esta prueba parte del supuesto de que cualquier conjunto de datos que no se ajusten a una segregación de genes independientes implica ligamiento. Las segregaciones esperadas sobre las que se fundamenta la Hipótesis nula (Ho) para dos genes independientes se presentan en el siguiente cuadro:

Población	Dominancia Completa	Codominancia
F2	9:3:3:1	1:2:1:2:4:2:1:2:1
Rc	1:1:1:1	1:1:1:1
RIL	1:1:1:1	1:1:1:1
DH	1:1:1:1	1:1:1:1

La divergencia de los datos observados con respecto a los esperados, partiendo de una segregación de genes independientes, viene dada por

$$\chi^2 Tot = \sum_i (a_i - nm_i) / nm_i$$

donde n es el total de individuos evaluados y a_i y m_i el número de individuos observados y la frecuencia esperada para la i-ésima clase, respectivamente.

Esta variación Total (o Suma de Cuadrados Totales) incluye tres componentes:

- desviación del gen A con respecto a la segregación mendeliana (Ho: 1:2:1 en F_2, 1:1 en cruza de prueba, 1:1 en doble-haploides) (χ^2A)
- desviación del gen B con respecto a la segregación mendeliana (χ^2B)
- desviación de la segregación conjunta (genes A y B) con respecto a la independencia (χ^2L)

Por descarte el χ^2L viene dado por la diferencia entre la variación total y las divergencias de las segregaciones individuales de los dos genes

$$\chi^2 L = \chi^2 Total - \chi^2 A - \chi^2 B$$

Este valor se confronta con χ^2 teórico, que indica la probabilidad de que el valor de χ^2 observado sea debido al azar en presencia de la Ho. Si el valor observado está por encima del teórico se rechaza la Ho de no ligamiento y se infiere que los datos se comportan como si los dos genes estuviesen ligados.

Prueba de máxima verosimilitud

Uno de los criterios estadísticos para aceptar o rechazar el ligamiento entre dos o más genes es el Lod-Score (LS), originalmente propuesto por Morton (1955) aunque adaptado posteriormente para muestras fijas no secuenciales (Chotai 1984, Gerber y Rodolphe 1994). Éste se define como el Log_{10} de la relación que existe entre la probabilidad de que los datos provengan de loci que están ligados y la probabilidad de que los datos se deriven de loci independientes. El término *Lod* son las siglas del "logaritmo de la razón de las probabilidades (odds)".

El LS se puede definir mediante la razón:

$$z(\theta) = log_{10} [P(r;\theta) / P(r;0,5)].$$

Considerado como una función de θ, $P(r;\theta)$ es llamada la función de verosimilitud, donde θ es el parámetro de la frecuencia de recombinación y r el estimado muestral. El numerador de la razón se lee como la "probabilidad de que se hayan obtenido los datos observados asumiendo un valor de r dado"; por otra parte, el denominador se interpretaría como "la probabilidad de que se hayan obtenido los datos observados asumiendo un r de 0,5".

La prueba LS rechaza la hipótesis nula de independencia entre los genes [Ho (θ=0,5)] para valores altos de la razón

$$P(r;\theta max) / P(r;0,5);$$

$P(r;\theta max)$ denota el máximo de $P(r;\theta)$ en el intérvalo $0,5 \geq \theta \geq 0$, es decir el valor de r que más se ajusta a los datos. El valor θmax con el cual se consigue este máximo se denomina el *estimado de máxima verosimilitud* de θ, y su cálculo se detallará más adelante.

La prueba del LS es más potente que la tradicional prueba de la Tabla de Contingencia cuando las segregaciones individuales de cada gen son mendelianas; pero al contrario, no es robusta cuando existen distorsiones. El límite convencional para la aceptación del ligamiento es un LS mayor o igual a 3,0, el cual se corresponde con una relación de 1000:1 a favor del ligamiento. Los valores de probabilidad asociados con cada LS, basados en una distribución χ^2 con 1 grado de libertad se muestran en el siguiente cuadro

Lod-Score y sus valores de probabilidad (error tipo I) asociados

Lod-Score	Nivel de significación*
1,0	0,0160
1,5	0,0043
2,0	0,0012
2,5	0,0003
3,0	0,0001

* Niveles de significación para la prueba de una sola cola

Debido a que el valor umbral de 3,0 se adaptó de estudios de genética humana, a veces se le considera demasiado estricto para aceptar ligamientos en especies con un menor número de cromosomas, en las cuales se supone que existe una mayor probabilidad *a priori* de ligamiento. Por el contrario, un valor umbral bajo incrementa la oportunidad de detectar ligamiento si está presente, pero aumenta el riesgo de aceptar el ligamiento de dos loci cuando en realidad no existe.

Prueba de contingencia

La prueba de detección de ligamiento $\chi^2 L$ asume que cualquier locus se desvía de la segregación mendeliana esperada por efectos del azar. En esta situación, el azar queda minimizado en el estimador χ^2 L. Sin embargo, cuando se sospecha que las desviaciones son sistemáticas, es recomendable utilizar una prueba de contingencia. Por otra parte, la prueba LS también se basa en que cada locus no se desvía de las proporciones mendelianas, por lo que no se recomienda su empleo en caso contrario. Ante ello, se sugiere el empleo de una prueba de contingencia (χ_C^2), la cual no presupone segregación mendeliana alguna.

En la prueba de χ_C^2, las frecuencias esperadas se obtienen a partir de

$$E_{ij} = \Sigma_i^{c1} \Sigma_j^{c2} (X_{i.} * X_{.j}) / X..$$

donde

E_{ij} = es la frecuencia absoluta esperada para la i-ésima clase del factor 1 y la j-ésima clase del factor 2
$Xi.$ = es el total marginal con la i-ésima clase del factor 1
$J.j$ = es el total marginal con la el j-ésima clase del factor 2
$X..$ = es el total general

Los valores esperados se calculan bajo el supuesto de que la hipótesis nula es verdadera (independencia de los factores), la cual se distribuye aproximadamente como χ_C^2 , con (m-1)*(n-1) grados de libertad, donde n es el número de clases del factor 1 y m el número de clases del factor 2. Para una buena aproximación, los valores esperados no deberían ser tan pequeños.

El estadístico χ_C^2 se obtendría como

$$\chi_C^2 = \Sigma_i^{c1} \Sigma_j^{c2} (X_{ij} - E_{ij})^2 / E_{ij}$$

Un punto importante, es que la prueba $\chi^2 L$ es preferible a la de contingencia cuando se presentan desviaciones monofactoriales debidas al azar. Sin embargo, no es fácil determinar si las desviaciones son aleatorias, de allí que se sugiera realizar una prueba de contingencia que complemente a la prueba de ligamiento $\chi^2 L$.

2.4 Cuantificación de la frecuencia de recombinación

La cuantificación de la intensidad de ligamiento se basa en la frecuencia de rompimiento y reunión de los cromosomas homólogos entre loci, por lo que es un estimado de la proporción de cromosomas recombinados y en organismos diploides, la proporción de gametas recombinantes. *La intensidad de ligamiento se mide a través de la frecuencia de recombinación,* designada como ***r***. El siguiente ejemplo muestra gráficamente este concepto:

Asuma el caso hipotético de un organismo F_1 con sólo 4 células germinales de genotipo AB/ab. Suponiendo que el 50% de sus células sufren entrecruzamiento (M = 0,5) durante la meiosis se tendrían los siguientes productos meióticos:

	Parentales	Recombinantes
CÉLULA 1 (sin EC):	2 AB, 2 ab	
CÉLULA 2 (sin EC):	2 AB, 2 ab	
CÉLULA 3 (con EC):	1 AB, 1 ab,	1 Ab, 1 aB
CÉLULA 4 (con EC):	1 AB, 1 ab,	1 Ab, 1 aB

En total:

Gametas paternas
AB = 6/16
ab = 6/16
Gametas recombinantes
Ab = 2/16
aB = 2/16

r = frecuencia gametas recomb. = 2(2/16) = ¼ = 0,25 (25%)

de allí que las frecuencias de cada gameta se pueden expresar como:

Gametas Paternas: AB y ab con frecuencia (1-r)/2 c/u
Gametas Recombinantes: Ab y aB con frecuencia r/2 c/u

Puede observarse que la frecuencia de recombinación es sólo la mitad de la frecuencia de entrecruzamiento (M=2*r*), lo cual se deduce porque de cada célula que sufre entrecruzamiento, se originarán dos cromátidas recombinantes. Si se considerara un porcentaje de entrecruzamiento del 100%, *r* sería igual a 0,5, en cuyo caso los genes se comportarían como si fuesen independientes.

Retrocruza

Permite determinar *r* en una forma sencilla y directa. Con caracteres dominantes, uno de los padres (neutral) sólo

poseerá genes en condición recesiva, de tal forma de inferir la constitución genética de las gametas de la F_1. Observando directamente los fenotipos de los individuos de esta población se obtiene *r*.

Si el individuo F_1 del ejemplo anterior se cruzase con un individuo completamente recesivo (cruza de prueba)

AB/ab x ab/ab

Gametas		**Genotipo**	**Fenotipo**	**Frec. Esperada**	**Valor observado**
F1	**Padre**				
(1-r)/2 AB	1 ab	AB/ab	AB	(1-r)/2	*a1*
r/2 Ab	1 ab	Ab/ab	Ab	r/2	*a2*
r/2 aB	1 ab	aB/ab	aB	r/2	*a3*
(1-r)/2 ab	1 ab	ab/ab	ab	(1-r)/2	*a4*

r = individuos recombinantes / Total

$$r = (a2 + a3) / (a1 + a2 + a3 + a4)$$

Este método, si bien es directo, está limitado a sólo aquellas situaciones en la que se dispone de la cruza de prueba, conseguida en muchos casos a través de la retrocruza de una F_1 con uno de los padres. No obstante, en muchos organismos es más fácil autofecundar y obtener una F_2 que hacer cruzamientos forzados.

Método de máxima verosimilitud

Este método puede aplicarse a cualquier población experimental (F_2, Rc, DH, F_3) y es más eficiente (estadísticamente) que el método anterior. Fue propuesto inicialmente para estimar *r* del cruce de doble heterocigotos (Haldane 1919, Fisher y Balmukand 1928). Para simplificar la comprensión del método, se estimará *r* mediante una cruza de prueba.

En esta población las frecuencias esperadas para dos genes, definidas por $\boldsymbol{m_i}$ (i = 1,2,3,4), se pueden conocer en términos de *r*, parámetro que se quiere determinar. La verosimilitud de observar una familia está dada por un término de la expansión

$$(m1 + m2 + m3 + m4)^n$$

donde *n* es el número total de individuos de la familia, lo que resulta en

$$[n! / a1!\ a2!\ a3!\ a4!] \ * \ (m1)^{a1}\ (m2)^{a2}\ (m3)^{a3}\ (m4)^{a4}.$$

El método de máxima verosimilitud depende de la maximización de esta expresión, con respecto a *r*. Para tal fin, el logaritmo de la misma es más cómodo matemáticamente para su maximización, ya que también tendrá su máximo para el mismo valor de *r* que la expresión no transformada. Así, el logaritmo de la expansión de verosimilitud, definida por *L*, es:

$$L = C + a1 \log m1 + a2 \log m2 + a3 \log m3 + a4 \log m4$$

donde C es una constante que depende del término de verosimilitud, el cual desaparece al momento de la derivación. La derivación maximiza el término:

$$dL/dr = a1(d \log m1)/ dr + a2(d \log m2)/ dr + a3 (d \log m3)/ dr + a4 (d \log m4)/dr = 0$$

Una de las soluciones de la ecuación será el valor de θmax, que daría el valor de r que más se ajusta a los datos observados. Allard (1956) presenta las soluciones para las derivadas dependiendo de la familia evaluada. Para esta cruza de prueba (en acoplamiento):

$$dL/dr = - (a1+a4)/(1-r) + (a2+a3)/(r) = 0$$

despejando r quedaría

$$r = (a2 + a3) / (a1 + a2 + a3 + a4)$$

que corresponde a la cantidad de individuos recombinantes entre total, deducción idéntica a la obtenida por el método de la retrocruza. El error estándar del estimado para el caso de esta cruza de prueba se obtiene a través de

$$Sr = [r(1-r)/n]^{1/2}$$

Métodos de los momentos

Cuando los factores monogénicos no segregan mendelianamente, se prefiere este método sobre el de máxima verosimilitud para estimar la frecuencia de recombinación. En algunas situaciones se obtienen estimados exactos; mientras que en otras se reduce el error originado por la pobre segregación si se compara con el método de máxima verosimilitud (Immer 1929). El método está adaptado para caracteres dominantes, donde sólo se registran 4 clases fenotípicas, a saber:

A>a, B>b

A-B- (a)
A-bb (b)
AaB- (c)
Aabb (d)

Si se emplean caracteres codominantes es necesario transformar los fenotipos codominantes de los marcadores de acuerdo a la nomenclatura anterior. Se puede asignar la dominancia a cualquiera de los dos alelos segregantes en cada locus.

El cociente $K= ad/bc$ es igualado a su esperanza matemática (Fisher y Balmukand 1928, Immer 1929):

$$K = \mathcal{E}(a)\mathcal{E}(d) \,/\, \mathcal{E}(c)\mathcal{E}(c) = P(2 +P) / (1 – P)^2$$

donde P está en función de la proporción de las gametas portadoras de las combinaciones alélicas AB y ab. De acuerdo a la fórmula anterior, P se consigue como una solución de

$$(K-1)P^2 – 2(K+1)P + K = 0$$

Finalmente *r* se calcula como

en fase de acoplamiento: $r = 1 - (P)^{1/2}$
en repulsión: $r = (P)^{1/2}$

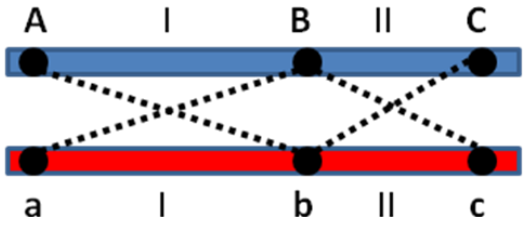
La varianza de *P* se consigue mediante

$$2P(1-P)(2+P) / n(1+2P)$$

donde *n* es el tamaño de la progenie.

2.5 Principios del mapeo cromosómico

Las frecuencias de recombinación entre pares de loci ubicados en un mismo cromosoma deben ser encuadradas en un Mapa Cromosómico. El mapeo cromosómico involucra al menos tres loci ubicados en un cromosoma. Aquellos loci muy próximos físicamente deberían presentar recombinaciones poco frecuentes y viceversa. Si asumimos el siguiente ejemplo en el cual el orden de los genes es A, B y C, con la siguiente fase parental:



Y se realiza el cruce **ABC/abc x abc/abc** con los siguientes resultados:

ABC/abc = 360; abc/abc = 363; ABc/abc = 47
abC/abc = 55; Abc/abc = 112; aBC/abc = 99
AbC/abc = 8; aBc/abc = 10

La frecuencias de recombinación vienen dadas por

$$r_{AB} = fRSi + fDR$$

en la que

fRSi = frecuencia de recombinantes simples de la i-ésima región, siendo i = I, II
fDR = frecuencia de los doble recombinantes

de esta manera

rI_{AB} = (112+99) + (8+10) / 1054 = 0,217 (21,7%)
rII_{BC} = (47+55) + (8+10) / 1054 = 0,114 (11,4%)
r_{AB+BC} = 0,217 + 0,114 = 0,331

Se debe enfatizar que las frecuencias de recombinación no son aditivas. Si se calcula *r* entre A y C directamente de los datos:

$r_{AC} = (Ac + aC) / Total$ = [(47+112) + (55+99)] / 1054
r_{AC} =0,297

se obtiene un valor menor al conseguido por la sumatoria de los *r* individuales, es decir, 0,331.

Esto se entiende porque cada frecuencia de recombinación para un segmento dado incluye a los dobles recombinantes, los cuales, al considerar dos secciones contiguas, se neutralizan dando la apariencia de que se está considerando una gameta paterna. De allí que varios eventos de recombinación son enmascarados y r_{AC} tiende a ser menor que r_{AB+BC}.

Esta relación se expresa como:

$$r_{AB+BC} = fRSI + fRSII + 2fDR$$

lo que sugiere que las sumas individuales poseen un componente adicional no detectado en los datos observados igual a *2DR*, por lo que las *r* no son aditivas. Como los recombinantes observados entre A y C son aquellos que provienen de un solo evento de recombinación, bien sea entre A y B ó entre B y C, entonces

$$r_{AC} = r_{AB}\,(1-r_{BC}) + r_{BC}\,(1-r_{AB})$$

$$r_{AC} = r_{AB} + r_{BC} - 2r_{AB}r_{BC}$$

El factor $r_{AB}r_{BC}$ (*fDR*) se hace mayor a medida que los *r* de cada uno de los segmentos se hacen más grande, lo cual disminuye la aditividad de *r* como medida de mapeo lineal.

Distancias de mapeo

Una distancia de mapeo debe ser *Aditiva*, de tal manera que cuando nuevos loci sean añadidos al mapa, las distancias anteriores no tengan que ser reajustadas. También debe medir el número total de entrecruzamientos (*M*) (tanto los pares como los impares). Por naturaleza, esta medida es aditiva, en la que el número de entrecruzamientos (*M*) entre A y C es igual a los M entre A y B más los M entre B y C.

Algunas funciones de mapeo que tratan de predecir M en función de *r* fueron propuestas por Haldane (1919) y Kosambi (1944). La Distancia de Mapeo de Haldane es la más simple pues no asume interferencia, lo que se puede interpretar como la ocurrencia de entrecruzamientos al azar e independientes a lo largo del todo el genoma:

$$M = ln(1-2r) / 2 \quad \text{(en Morgans) y}$$

$$r = ½\,[1-exp(1-2M)]$$

La distancia de Kosambi incluye interferencia parcial

$$M = ¼\,ln\,[(1+2r)/(1-2r)] \quad \textit{(en Morgans) y}$$

$$r = [1-exp(-4M)] / [2(1+exp(-4M))]$$

No hay una relación universal entre la distancia de mapeo y la distancia física entre dos loci. Un centiMorgan (o unidad de mapeo) se corresponde con una distancia de aprox. 10 Kb a 1000 Kb, dependiendo de la especie. Incluso dentro de un

mismo cromosoma puede haber cambios dramáticos. Los entrecruzamientos tienden a ser suprimidos cerca del centrómero y los telómeros, lo cual aumenta la cantidad de pares de bases por cM. La tasa de recombinación puede variar también con el sexo (*Drosophila*).

Ordenamiento de los loci

Los mapas genéticos se construyen con mayor confianza cuando se incluye información de múltiples loci. En tal sentido, los métodos de mapeo deben adaptarse a esta condición. Esto no es fácil, pues el número de combinaciones posibles es elevado. Para determinar el orden más probable de los marcadores dentro de un grupo de ligamiento, el programa MAPMAKER (Lander *et al*. 1987; Lander y Green 1987) por ejemplo emplea el siguiente procedimiento:

- Para cada posible orden de ese grupo se calcula la máxima verosimilitud según los datos observados, considerando simultáneamente las distancias de mapeo entre todos los marcadores.
- Luego se comparan estas verosimilitudes y se escoge el orden más probable como respuesta.

Este tipo de análisis no es práctico, incluso para grupos de tamaño mediano (6 a 10 marcadores) porque para un grupo de N marcadores existirían N!/2 combinaciones. En estos casos, se obtienen órdenes de subconjuntos del grupo de ligamiento, para después solaparlos y finalmente mapear cualquier marcador sobrante en función de los ya mapeados.

El proceso de mapeo para pequeños grupos de loci se hace a través del Algoritmo *EM*. Este algoritmo ofrece una aproximación general y potente para obtener los estimados de máxima verosimilitud, aun si hay datos faltantes. En plantas, se adaptó a partir de desarrollos teóricos aplicados a la genética humana. El algoritmo *EM* busca en un espacio multidimensional las frecuencias de recombinación para todos los intervalos. El procedimiento general se describe a continuación.

Asuma

$$\theta = \{\theta_1, \ldots, \theta_{m-1}\}$$

Como un vector de un conjunto de frecuencias de recombinación entre loci que maximiza la probabilidad de que los datos observados se hayan derivado de él.

$$M_1, \ldots, M_m$$

Representan los *m* loci genéticos, alineados de acuerdo a un orden cromosómico asumido, y θi a la fracción de recombinación entre los loci adyacentes m_i y m_{i+1}.

Entonces *EM* remplaza mediante tanteo un $\theta^{antiguo}$ en un θ^{nuevo}, donde éste último posee una mayor verosimilitud que el anterior.

EM busca el θ final suponiendo un estimado inicial ($\theta^{antiguo}$) para todas las fracciones de recombinación entre los loci considerados. En el primer paso (E=*expectativa*) usa el $\theta^{antiguo}$

como si fuese el θ real, y calcula el valor esperado para los datos completos, es decir, el número esperado de meiosis recombinantes y no recombinantes en cada intervalo. Este paso se conoce también como Reconstrucción Genética, referida al cálculo del número esperado de meiosis recombinantes. La Reconstrucción Genética más generalizada y empleada para las poblaciones vegetales se basa en las "Cadenas Ocultas de Markov".

El siguiente paso (M=*maximización)* usa los valores esperados de los datos completos como si fuesen los verdaderos, y calcula el estimado de máxima verosimilitud θ^{nuevo} para las fracciones de recombinación.

Finalmente tantea de nuevo los pasos E y M hasta que los estimados $\theta^{antiguo}$ y θ^{nuevo} converjan.

Referencias bibliográficas

Allard R. 1956. **Formulas and tables to facilitate the calculation of recombination values in heredity**. Hilgardia 24: 235-278.

Chotai J. 1984. **On the lod score method in linkage analysis**. Ann. Hum. Genet. 48: 359-378.

Fisher R.A & B. Balmukand. 1928. **The estimation of linkage from the offspring of selfed heterozygotes.** J. Genet. 20: 79-92.

Gerber S. & F. Rodolphe. 1994. **Estimation and test for linkage between markers: a comparison of Iod score and χ^2 test in a linkage study of maritime pine (*Pinus pinaster* Ait.)**. Theor. Appl. Genet. 88: 293-297.

Haldane J. 1919. **The probable errors of calculated linkage values, and the most accurate method of determining gametic from certain zygotic series**. J. Genet. 8: 291-297.

Haldane J. 1919. **The combination of linkage values, and the calculation of distances between the loci of linked factors**. J. Genet. 8: 299-309.

Immer F.R. 1929. **Formulae and tables for calculating linkage intensities**. Genetics 15: 81-98.

Kosambi D. 1944. **The estimation of map distances from recombination values**. Ann. Eugen. 12:172-175.

Lander E. & P. Green. 1987. **Construction of multilocus genetic linkage maps in humans**. Proc. Natl. Acad. Sci. 2363-2367.

Lander E.S., Green P., Abrahamson J., Barlow A., Daly M.J., Lincoln S.E. & L. Newburg. 1987. **MAPMAKER: an interactive computer package for constructing primary genetic linkage maps of experimental populations**. Genomics 1:174-181.

Morton N.E. 1955. **Sequential tests for the detection of linkage**. Am. J. Hum. Genet. 7:277-318.

Referencias Generales:

Mather K. 1957. **The measurement of linkage in heredity**. Methuen & Co. LTD. pag. 13-25.

Mather K. 1967. **The elements of biometry**. Methuen & Co. LTD. pag. 73-86.

3. DETECCIÓN Y MAPEO DE QTL EN PLANTAS

Mediante asociaciones estadísticas entre marcadores genéticos y un carácter cuantitativo se pueden detectar, ubicar y estimar el efecto de los factores genéticos que controlan al mismo o *Quantitative Trait Loci* (QTL). Los QTL se emplazan en regiones particulares del genoma lo que permite caracterizarlos como unidades mendelianas y manipularlos de forma más simple. Como requisito inicial es importante disponer de mapas genéticos densamente saturados para posteriormente localizar los QTL. Los principios básicos del mapeo ya se discutieron en el capítulo anterior.

No se pretende hacer una revisión exhaustiva de los métodos de mapeo; sólo se discutirán los más simples y de mayor uso, en particular los Modelos lineales (un marcador a la vez) y el Mapeo por intérvalo.

3.1 Generalidades

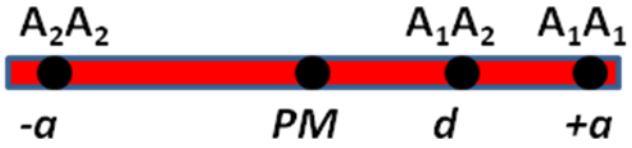
Una característica cuantitativa comúnmente está condicionada por poligenes, cada uno con efecto pequeño sobre el fenotipo, por lo que es difícil discriminar los efectos individuales de cada gen. En general dicha característica está muy afectada por el medio ambiente. Estadísticamente tiende a ser registrada a través de mediciones y explicada por medio de medias y varianzas. Al verificarse que los rasgos cuantitativos eran explicados por la conjunción de genes de herencia simple, se buscó asociarlos con genes marcadores a través del ligamiento genético facilitando su ubicación en un mapa.

Considere *r* como la frecuencia de recombinación entre un marcador genético, llamado M1 y un locus (Q) que codifica para una característica cuantitativa (QTL), y *r´* la frecuencia recombinación entre Q y un segundo marcador M2.

Si se conociera la ubicación en el mapa de *M1 y M2, r y r´*, esto permitiría ubicar al *QTL*. El mapeo genético de los QTL descansa sobre la idea de que los marcadores genéticos que tienden a ser transmitidos en conjunto con valores específicos de la característica cuantitativa tienen una alta probabilidad de estar cerca del gen que afecta a ese carácter.

Valores genotípicos

Las relaciones entre los valores genotípicos se pueden expresar en la escala:



donde

PM (Punto Medio) = punto medio entre los valores genotípicos de los homocigotos
-a = valor genotípico del homocigoto A2A2
+a = valor genotípico del homocigoto A1A1
d = valor genotípico del heterocigoto A1A2 (especifica el grado de dominancia)

El valor de *a* refleja el componente aditivo de la variabilidad genética. Es el que se puede ir acumulando por selección y está directamente relacionado con la cantidad de alelos favorables en un organismo. El término favorable es relativo y dependerá de cuál alelo aporte un fenotipo ventajoso en ciertas condiciones ambientales, naturales o antrópicas.

Por otra parte, *d* es un término asociado con la interacción de dominancia e influye en el componente no aditivo de la varianza genética. Dado a que sólo se expresa en individuos heterocigotos, no puede medirse en homocigotos, por lo que no suele relacionarse con la cantidad de alelos favorables en un individuo.

Primeros estudios

Sax (1923) en *Phaseolus vulgaris*, estudió la asociación entre el peso y el grado de pigmentación del grano. Él observó en una F_2 que la pigmentación del grano se asociaba con el peso del mismo. Los granos pigmentados estaban condicionados por un alelo dominante (P); mientras que los no pigmentados por la alternativa recesiva (p).

Los registros del cruce fueron:

L1	x	L2
(56 g; Pigmentada)		(21 g; No pigmentada)
P		***p***

F_2

P	**X= 29 g**
p	**X= 26,4 g**

Las relaciones entre los distintos genotipos fueron

- *PP / pp = 4,3 (P <0,05)*
- *Pp / pp = 1,9*
- *(PP-pp)/ (Pp-pp) = 2*

lo cual se interpreta como un incremento constante entre pp – Pp – PP de 2, por lo que se concluyó que P estaba ligado de manera aditiva al peso del grano.

3.2. Métodos estadísticos para el análisis de los QTL

Es común en todos estos métodos conocer el efecto de *a*, *d* y *r*. Sin embargo, se pueden clasificar de acuerdo al uso de la información disponible:

- Un marcador a la vez (modelos lineales)
- Mapeo por intervalo
- Mapeo múltiple

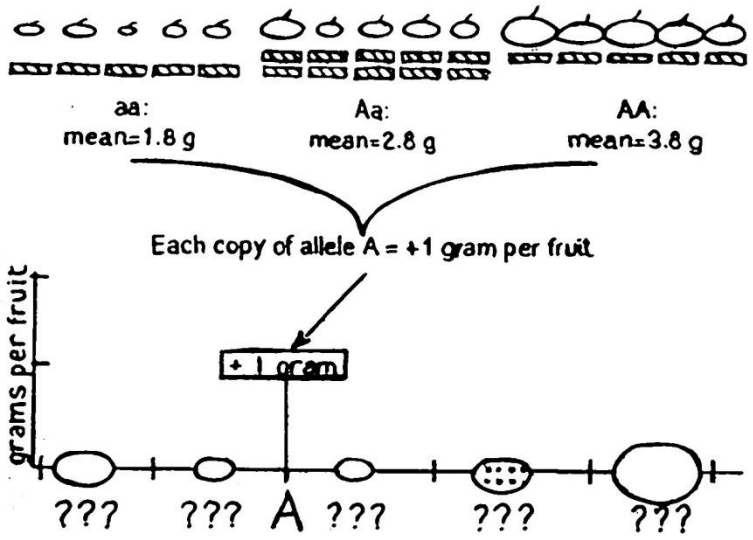
Sólo los dos primeros serán descritos aquí. En todos los métodos se registra el carácter cuantitativo y el genotipo del marcador genético para cada individuo o un grupo de individuos que representan una línea endogámica recombinante (RIL) o doble haploide (DH). Se acostumbra que en la fase de plántula se registre el marcador genético y posteriormente en campo, luego del transplante en algunas especies, la variable cuantitativa.

Salvo algunas excepciones, los métodos se basan en la evaluación de caracteres con distribución normal, por lo que no es raro el uso de los métodos estadísticos paramétricos, como los derivados del modelo lineal y los análisis de regresión

Modelos lineales

Estos métodos usan la información de un marcador a la vez. No requieren de un mapeo cromosómico, pues consideran la asociación de la característica cuantitativa con cada marcador por separado. Los derivados del modelo lineal son los más sencillos en su concepción y en su aplicación. A través de estos métodos se puede determinar superficialmente cuáles marcadores presentan asociación con los QTLs, y hacer un descarte preliminar.

En la siguiente figura se muestra gráficamente el principio que subyace en los modelos lineales. Los diversos genotipos del marcador genético se identifican mediante patrones de bandas en un gel de electroforesis. La característica cuantitativa corresponde al tamaño del fruto.



Tomado de Paterson *et al*. (1991)

Cada genotipo del marcador puede estar asociado con el peso medido de los individuos con ese patrón. En la figura anterior se observa que los individuos con la banda inferior (genotipo aa) presentan una media para el peso de 1,8 g, el patrón con dos bandas (Aa) de 2,8 g y aquellos con la banda superior (AA) de 3,8 g. Por lo tanto, el marcador genético A muestra una asociación con la característica cuantitativa. La asociación se muestra porque existen diferencias de 1 g entre cada homocigoto y el heterocigoto y 2 g entre ambos homocigotos. Ello indica que por cada alelo A que posea un

individuo, éste poseerá 1 g adicional partiendo del nivel basal de 1,8 g.

No obstante, no se tiene la certeza sobre la ubicación de los genes que codifican la característica. La razón es la siguiente: para una asociación entre el marcador y el QTL pueden existir dos alternativas

- Un QTL muy próximo al marcador con efecto pequeño (frutos pequeños cerca de la letra A en la figura)
- Un QTL muy alejado al marcador con efecto grande (frutos grandes alejados de la letra A en la figura)

De allí surge la primera limitación de los métodos de un marcador a la vez: no es posible estimar con certeza el efecto del QTL, denominado *a*, y su distancia del marcador (r). Por lo consecuencia, estos métodos pueden usarse como una primera aproximación al estudio de los QTL. Esta incertidumbre, como se verá posteriormente, se puede definir estadísticamente.

Detección estadística del efecto de un QTL

Detectar estadísticamente las diferencias entre clases para los modelos lineales depende del número de clases consideradas. Cuando se trata de dos, lo ideal sería una prueba de *t*. Para tres clases, una prueba de *F*, comúnmente denominada Análisis de varianza o ANOVA (en inglés), permite detectar diferencias entre las clases; posteriormente una prueba de contraste entre las medias de las clases (por ejemplo la prueba de *t*) complementaría el análisis. Otra vía es a través de un Análisis de Regresión Simple, lo cual es equivalente. En este caso, si el coeficiente de regresión es distinto de cero, indicaría que el marcador está asociado con el carácter cuantitativo.

En un ANOVA y en la prueba de *t* se detecta la diferencia (δ) entre las clases del marcador si ésta es lo suficientemente grande, dada una varianza (σ^2) dentro de las clases, lo cual implica asociación. El nivel de significación está en función del nivel de confianza de la prueba.

Las diferencias entre clases no necesariamente indican el contraste real entre los valores genotípicos. En los modelos lineales, la frecuencia de recombinación (*r*) enmascara las diferencias reales de la siguiente manera

δ en una F_2

$$M1M1 - M2M2 = 2a\,(1\text{-}2r)$$
$$M1M2 - (M1M1 + M2M2)/2 = d\,(1\text{-}2r)^2$$

δ en una Rc

$$(A)\ M1M1 - M1M2 = (a\text{-}d)\,(1\text{-}2r)$$
$$(B)\ M1M2 - M2M2 = (a+d)\,(1\text{-}2r)$$

$$A\text{-}B = 2d\,(1\text{-}2r)$$
$$A+B = 2a\,(1\text{-}2r)$$

δ en una población de doble haploides (DH)

$$M1M1 - M2M2 = 2a\,(1\text{-}2r)$$

En todos los casos, *a* y *d* están sesgados (subestimados) por (*1-2r)*. Mientras mayor es el valor de *r*, mayor será el sesgo que posee la diferencia observada entre las clases del marcador. Ahora se entiende por que un QTL con efecto grande (*a*) y *r* grande puede aparentar un QTL pequeño con *r* pequeño. Por ejemplo, en el caso de los siguientes QTL detectados en una población de DH

QTL A: a = 2; r = 0,3
QTL B: a = 1; r = 0,1

El δ para los dos QTL

$\delta_A = 2(2)\,[1\text{-}2(0{,}3)] = 1{,}6$

$\delta_B = 2(1)\,[1\text{-}2(0{,}1)] = 1{,}6$

Surge por lo tanto una restricción para el uso de estos modelos: el valor real de la diferencia sólo será detectado si el marcador se encuentra exactamente en el QTL, es decir, que el *r* = 0. No hay forma de conocer *a* y *d* salvo que *r* sea cero. Si *r* = 0,5, la diferencia real, a pesar de que pueda ser inmensamente grande, no se detectará con la prueba, pues

$\delta = 2a\,[1 - 2\,(0{,}5)] = 0.$

Otras limitaciones del modelo surgen también cuando *r* es diferente de cero. En el siguiente cuadro se describen las medias y varianzas teóricas para cada clase del marcador en una F_2.

Genotipo	Media	Varianza
M1M1	$a(1\text{-}2r) + 2dr(1\text{-}r)$	$2a^2r(1\text{-}r)+2d^2rs$ $-4adr(1\text{-}3r+2r^2)$
M1M2	$d(1\text{-}2r+2dr^2)$	$2a^2r(1\text{-}r)+2d^2rs$
M2M2	$-a(1\text{-}2r) + 2dr(1\text{-}r)$	$2a^2r(1\text{-}r)+2d^2rs$ $+4adr(1\text{-}3r+2r^2)$

$s= 1\text{-}3r+4r^2\text{-}2r^3$

Cuando *r* = 0, las varianzas de las tres clases son iguales a $2a^2r$; pero para *r* > 0, las varianzas entre las clases son distintas. El ANOVA asume que las varianzas entre clases son iguales, por lo que se estaría violando uno de los supuestos del método.

En conclusión para los modelos lineales. i) Si las varianzas son distintas, se sugiere no utilizar el término del error del ANOVA. ii) No se puede precisar la ubicación del QTL ni su efecto real, ya que no se pueden estimar *a*, *d* y *r*, pues las dos primeros parámetros estarían subestimados por un factor (*1-2r*). iii) Debido al efecto del entrecruzamiento (r > 0) pueden aparecer diferentes genotipos del QTL dentro de una clase del marcador; eso genera distribuciones mixtas dentro de cada clase MiMj debidas a la segregación de Q1Q1 + Q1Q2 + Q2Q2, inflando las varianzas dentro de las clases y disminuyendo la sensibilidad de la prueba.

Ejemplo

El siguiente caso ilustra la aplicación del modelo en tomate para la tolerancia a la salinidad (Bretó *et al*. 1994). El estudio

asoció la expresión de la característica cuantitativa a 2 marcadores isoenzimáticos, 10 RFLP y 2 RAPD. Cada descendencia F_2 del cruce *L. esculentum* cv Madrigal x *L. pimpinellifolium* Línea 1 (donadora de la tolerancia) se evaluó para número de frutos (FN), peso del fruto promedio (FW) y peso de fruto total (FT) en condiciones salinas.

Se declaró que un marcador estaba asociado con la salinidad cuando las medias de cada clase del marcador (para rendimiento) eran distintas significativamente de acuerdo a la prueba de t con estimados no agrupados (pool) de la varianza, la cual se especifica como

$$t = (Xm_1 - Xm_2) / [(s_1^2/n_1) + (s_2^2/n_2)]^{1/2}$$

donde

Xm_i = media de la i-ésima clase marcadora
s_i = desviación estándar de la i-ésima clase
n_i = número de individuos evaluados en la i-ésima clase

El valor de t, que sigue la distribución de Student, se compara con un t tabulado con

$$(s^2_1/n_1 + s^2_2/n_2)^2 / [(s^2_1/n_1)^2 / (n_1-1)] + [s^2_2/n_2)^2 / (n_2-1)]$$

grados de libertad.

Los resultados se muestran a continuación. Cuando son significativos, los genotipos se ordenan según sus medias mostrándose el estadístico *t* para la diferencia entre clases hmocigotas. NS significa que las clases genotípicas no son significativas.

Marcador	Peso total	Número de frutos	Peso del fruto
Aco-1	NS	NS	*PP, EP > EE* −2
Est-4	NS	NS	NS
TG 18	NS	NS	NS
TG 23	NS	NS	NS
TG 24	NS	*PP, EP > EE* −3.4	NS
TG 30	NS	NS	NS
TG 43	NS	*PP* ≥ *EP* ≥ *EE* −2.5	NS
TG 48	NS	*PP, EP > EE* −2.5	*EE > EP, PP* 2.9
TG 63	NS	NS	NS
TG 68	NS	NS	NS
TG 123	*PP, EP > EE* −2.9	*PP, EP > EE* −3.8	NS
TG 134	NS	NS	*EE, EP > PP* 3.4
PD2	NS	NS	NS
PD3	NS	NS	NS

Se detectaron varios QTL relacionados con la tolerancia a la salinidad; 6 marcadores mostraron cierto grado de asociación con los componentes de rendimiento: 3 de ellos con FW, 4 con FN y 1 con TW. El marcador TG 48 estuvo relacionado tanto con FW como con FN, lo que sugiere la existencia de efectos pleiotrópicos en esta región del genoma. Para los marcadores ligados a los QTL que afectan a FN, las plantas F_2 con el alelo P en el locus marcador (derivado de *L. pimpinellifolium*), tanto en dosis simple como en doble, produjeron más frutos en condiciones salinas que las plantas

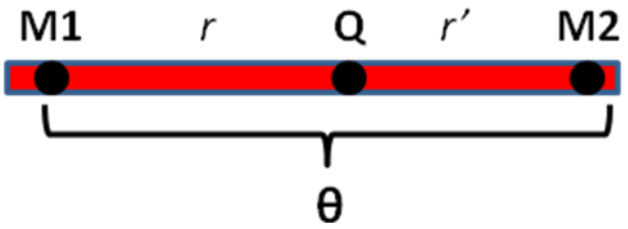
homocigotas para el alelo E. En el caso de FW, las plantas con el alelo P en los loci TG48 y TG134 presentaron frutos más pequeños y por el contrario, frutos más grandes en Aco-1.

Los autores concluyen que en el mejoramiento para rendimiento bajo condiciones salinas, la selección debería estar orientada hacia aquellas plantas con el alelo P en los loci Aco-1, TG24, TG123 y TG43, y con el alelo E en los demás loci; de igual manera sugieren desarrollar dos líneas distintas con los genotipos EE y PP en el locus TG48 para incrementar FW y FN, respectivamente

Mapeo por intervalo

El mapeo por intervalo surge de la necesidad de estimar *a*, *d* y *r*. El método evalúa si un QTL se encuentra entre dos marcadores. Algunas de las ventajas que muestra sobre los modelos lineales son su mayor potencia para detectar QTL, ya que se disminuye el efecto de las distribuciones mixtas; disminución el error Tipo I; menor sensibilidad a la no-normalidad; además de generar intervalos de confianza para estos parámetros.

En el mapeo por intervalo están contenidas dos frecuencias de recombinación (*r* y *r´*) y tres loci distribuidos linealmente de la siguiente manera



Donde

$\theta = r + r'$ sin interferencia
$\theta = r + r' - 2rr'$ con interferencia

Modelo de regresión lineal

El modelo de regresión se basa en que la variable dependiente (en este caso el carácter cuantitativo) se puede explicar a través de una o varias variables (marcadores genéticos).

Considere que Y_{ij} es la observación en la i-ésima planta con la j-ésima clase del marcador, y Γ_j la probabilidad de que un alelo Q esté presente en un individuo dada la j-ésima clase marcadora. Q se considera un alelo favorable en el QTL para la característica cuantitativa. Cada clase marcadora está representada por una combinación alélica de los dos marcadores circundantes al QTL, donde M1M2, M1m2, m1M2, m1m2 pueden tomar los valores respectivo de j = 1, 2, 3, 4. Tanto M1 y m1, como M2 y m2 representan las dos variantes alélicas de cada marcador. Además, asuma que i = 1, 2, ..., n_j es el número de individuos en la j-ésima clase,

Según una hipótesis con sólo efectos aditivos en el locus Q, en la que sólo interesa la presencia de los alelos favorables y no las combinaciones heterocigóticas (dominancia), Y_{ij} viene dada por

$$Y_{ij} = 2q + (Q\text{-}q)\,\Gamma_j$$

En el que

2q es el punto de corte de la regresión; nivel basal o genotipo sin alelos favorables (qq)

(Q-q) es la pendiente de la regresión; representa incremento en el rendimiento de acuerdo al número de alelos favorables Q que presenta el i-ésimo individuo en la j-ésima clase. Señala el valor de "a" del QTL

Γ_j variable que representa la probabilidad de que el i-ésimo individuo posea un alelo favorable; está en función de la clase marcadora

En otras palabras, se busca estimar el rendimiento de un individuo en función de su clase marcadora. En la ecuación general de regresión $2q = \beta_0$ y $(Q\text{-}q) = \beta_1$. Por lo tanto, el modelo se puede traducir en

$$\hat{y}_{ij} = \hat{\beta}_0 + \hat{\beta}_1 \Gamma_j + \varepsilon_{ij}$$

añadiendo el error experimental ε_{ij}.

Faltaría definir los valores que tomaría Γ_j dependiendo de la clase marcadora. Faltaría definir los valores que tomaría Γ_j dependiendo de la clase marcadora. Este modelo, descrito en Martínez y Curnow (1992) se aplica a una retrocruza, aunque puede ser adaptado a cualquier otra población. En tal sentido, los Γ_j estarán referidos a la Rc.

En una Rc, la gameta que define al fenotipo de cada individuo es la aportada por la F_1. Si se obtiene la probabilidad de que se presente un alelo Q dado las combinaciones alélicas en las gametas (M1M2, M1m2, m1M2, m1m2), se obtendrán los Γ_j. Por ejemplo, Γ_1 es la probabilidad de que esté presente un alelo Q dada la gameta del marcador M1M2. Trataremos de desarrollar solamente este valor. Considere dos padres con combinaciones alélicas **M1QM2/M1QM2** y **m1qm2/m1qm2** (representando el padre recesivo qq). Se asume de antemano que Q está entre los loci marcadores.

Así, un individuo de la F1 formaría las siguientes gametas

Combinación	**tipo**	**probabilidad**
M1QM2	paterna	*(1-r)(1-r´)/2*
m1qm2	paterna	*(1-r)(1-r´)/2*
M1Qm2	RSII	*(r´ - rr´) / 2*
m1qM2	RSII	*(r´ - rr´) / 2*
M1qm2	RSI	*(r – rr´) / 2*
m1QM2	RSI	*(r – rr´) / 2*
M1qM2	DR	*rr´/ 2*
m1Qm2	DR	*rr´/ 2*

RS = recombinantes simples; DR = doble recombinantes

La probabilidad de que ocurran gametas M1M2, sumando la tercera columna será

$P\ (M1M2) = (1-r)(1-r´)/2 + rr´/2 = ½\ (1-\theta)$

$\Gamma_1 = P(M1QM2) / P\ (M1M2) = P\ (M1QM2 / M1M2) = [(1-r)(1-r´)/2] / [½\ (1-\theta)]$

$\Gamma_1 = 1 – [r\ (\theta-r) / (1-2r)(1-\theta)]$

En el supuesto de que se desee incorporar interferencia completa en el modelo, los valores de Γ en el cuadro que se muestra a continuación deben cambiarse a: $\Gamma_1 = 1- \Gamma_4 = 1$ y $\Gamma_2 = 1- \Gamma_3 = (\theta-r)/\theta$.

Valores de Γ dado el genotipo del marcador sin interferencia

Γ_1	M1M2	$1 – [r\ (\theta-r) / (1-2r)(1-\theta)]$
Γ_2	M1m2	$(1-r)\ (\theta-r) / \theta\ (1-2r)$
Γ_3	m1M2	$1- \Gamma_2$
Γ_4	m1m2	$1- \Gamma_1$

Para cada valor de r, $r´$ y θ deberá existir un valor de Γ, y por consecuencia, un valor de β_0 y β_1, de allí que podemos expresar el modelo como

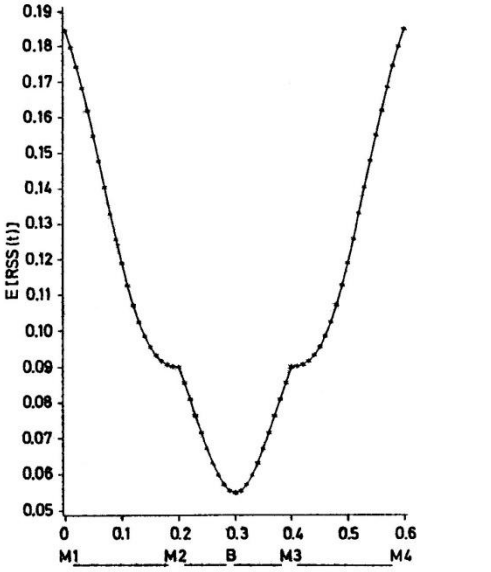
$$\hat{y}_{ij}\ (r) = \hat{\beta}_0\ (r) + \hat{\beta}_1(r)\ \hat{\Gamma}_j\ (r)$$

El fundamento del modelo de regresión se expresa así. Primero se fija un r para obtener los estimados de β_0 y β_1. Luego se repite este paso para todos los r posibles del intervalo definido por θ. Finalmente, se escoge aquel r que genere las menores Sumas de Cuadrados Residuales (SCR; fuente del error en el ANOVA), que son un indicativo del ajuste del modelo a los datos observados. Una vez conocido ese r, inmediatamente se conoce la posición del QTL y el estimado de a. La SCR se obtienen como

$$SCR = \Sigma^{4}_{j=1} \Sigma^{nj}_{i=1}\ (y_{ij} - \hat{y}_{ij}(r)\ )^2$$

Los resultados se pueden expresar como un gráfico de dos dimensiones, en el que el eje de las ordenadas expresa las SCR para cada uno de los r en el eje de las abscisas. Por ejemplo, Martínez y Curnow (1992) pusieron a prueba el modelo utilizando una población obtenida por simulación en el que la característica cuantitativa se explicaba a través de un QTL ubicado en el medio de dos marcadores M2 y M3, con un $a = 1$.

La siguiente figura muestra una sima (menor valor para la SCR o RSS), exactamente en la ubicación del QTL, a un $r = 0,1$ entre M2 y Q (en la figura expresado como B). En este caso el coeficiente de regresión $\beta_1(0,1) = 1,00$, coincide con el valor simulado de a = 1.



E[RSS(t)] for each point between the markers M_1 and M_4. Size of the effect B = 1. No interference

El modelo de regresión lineal puede adaptarse para más de un QTL y se puede considerar tres marcadores a la vez. Ello le confiere ciertas ventajas sobre el método de máxima verosimilitud y el modelo de regresión con dos marcadores a la vez. Hay que incluir en el modelo ya no 4 clases para una retrocruza, sino 8, además de que el número de β aumenta de 2 a 3 y se añaden dos probabilidades condicionales, ψ y △, aparte de Γ. En términos generales, aunque el modelo de regresión se puede aplicar a cualquier número de marcadores, el análisis simultáneo está restringido, pues el tamaño de la muestra se incrementaría exponencialmente de tal forma de obtener una precisión deseada.

Mapeo por intervalo simple

Este método localiza un QTL en cualquier ubicación de un grupo de ligamiento (Lander y Botstein 1989). Es el método incluido en el programa MAPMAKER/QTL (Lander *et al*. 1987). Para una retrocruza, el fenotipo (Φ_i) y el genotipo (g_i) para el i-ésimo individuo están relacionados de acuerdo a la tradicional regresión lineal

$$\Phi_i = a + bg_i + \varepsilon_i$$

donde

- a media de la población
- g_i es una variable indicadora que toma valores de (0,1) según el número de alelos Q
- b indica el efecto fenotípico estimado de una sustitución alélica en un QTL hipotético
- ε_i es una variable normal con media cero y varianza σ^2

Si se registran los valores fenotípicos para un grupo de individuos para cada clase marcadora, interesa obtener de allí los estimados de a, b y σ^2 que maximicen la probabilidad de ocurrencia de los datos observados L(a, b, σ^2), conocidos como estimados de máxima verosimilitud. Como la distribución normal es continua, recurrimos a la función de densidad de probabilidades para obtener la probabilidad de ocurrencia de cada observación.

En la retrocruza se esperan dos distribuciones, cada una correspondiente a las dos clases genotípicas para el locus QTL: Qq y qq; la primera con una copia del alelo Q y la otra sin copia alguna. Cada clase por lo tanto, de acuerdo a la ecuación anterior, poseerá una media distinta pues bg_i en la primera clase será igual a *b* y en la segunda igual a cero:

	media	**varianza**	**alelos Q**
Clase Qq	$a + b$	σ^2	1
Clase qq	a	σ^2	0

La función de densidad para el i-ésimo individuo dentro de cada clase estará definida por

$$F(y) = [1/\sigma(2n)^{1/2}] \ exp\ [-(x-\mu)^2 / 2\sigma^2]$$

Que para la

Clase Qq (1)

$$F(y) = [1/\sigma(2n)^{1/2}] \ exp\ [-(\Phi_i-(a + b))^2 / 2\sigma^2] = L(1);$$

Clase qq (0)

$$F(y) = [1/\sigma(2n)^{1/2}] \ exp\ [-(\Phi_i-a)^2 / 2\sigma^2] = L(0).$$

El problema consiste en determinar a cual de las dos clases pertenece un individuo, conociendo solamente la información del genotipo del marcador. Debido a la incertidumbre no se debería descartar ninguna de las dos, sino integrar las probabilidades de ocurrencia obtenidas de las dos funciones de densidad. Esto se logra al añadir un coeficiente a cada función de densidad, que indique la probabilidad de que cada una de las clases pueda ocurrir en el i-ésimo individuo, dado el genotipo del marcador. Este coeficiente se simboliza como $G_i(1)$ ó $G_i(0)$ según sea la clase, y depende de *r*, *r´* y θ de una forma similar a la segunda tabla de la página 28.

La probabilidad de ocurrencia del i-ésimo individuo queda definida como

$$G_i(0)\, L_i(0) + G_i(1)\, L_i(1)$$

Finalmente, las probabilidades de ocurrencia de todos los individuos evaluados deben ser multiplicadas entre sí para obtener la función de verosimilitud

$$L\ (a, b, \sigma^2) = \Pi_i\ [G_i(0)\, L_i(0) + G_i(1)\, L_i(1)].$$

Para cada valor de *r* se obtienen distintas verosimilitudes para la función $L(a, b, \sigma^2)$. Las frecuencias de recombinación entran en $G_i(1)$ y $G_i(0)$ de una forma similar al procedimiento que concluyó con los estimados del cuadro de la página 28.

Detección del QTL

Los estimados de máxima verosimilitud se comparan con los estimados obtenidos suponiendo que $b=0$, es decir, asumiendo que no hay QTL alguno (μ, 0, σ^2) contenido en el par de marcadores M1 y M2.

La evidencia de que existe un QTL se extrae del LOD-Score, o el logaritmo de la relación entre la probabilidad de que los datos observados se hayan derivado de un QTL ubicado en el intervalo definido por un par de marcadores y la probabilidad de que los datos provengan de un QTL ausente de ese intervalo

$$LOD = log_{10} (L (a, b, \sigma^2) / L (\mu , 0, \sigma^2)).$$

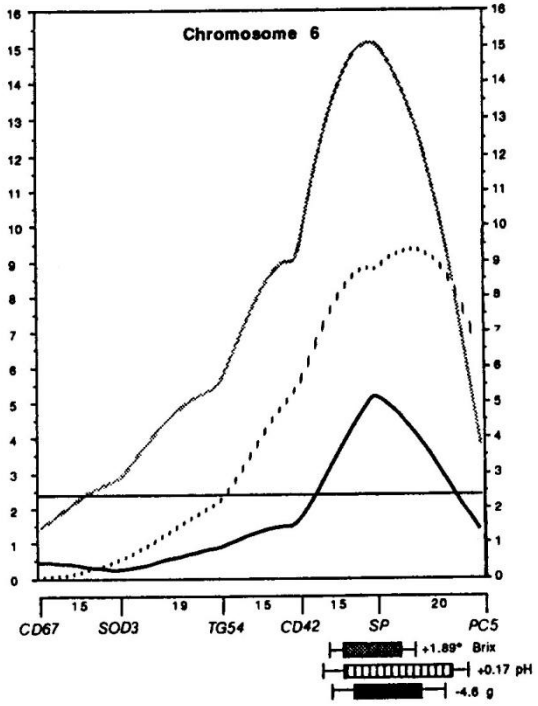
El LOD-Score se puede interpretar como "Cuán probable los datos observados se derivaron de un QTL presente que de uno ausente". Si el LOD sobrepasa un umbral, generalmente entre 2 y 3, se toma como indicio de que el QTL está en el intervalo evaluado.

Al igual que en el método de regresión lineal, para diferentes valores de *r* se obtienen diferentes verosimilitudes, y por consecuencia diferentes valores de LOD-Score. De allí se elabora una gráfica en dos dimensiones que expresa la relación entre los diferentes valores de *r* en el eje de las abscisas y el LOD-Score en el eje de las ordenadas. El umbral de aceptación de un QTL se representa con una recta horizontal para el valor del LOD seleccionado. En el *r* en que la curva sobrepase este umbral, se declara la ubicación del QTL.

Ejemplo

De Paterson *et al*. (1988) se extrajo la figura de la derecha de un mapa de verosimilitud del cromosoma 6 de tomate. Se midió el peso, concentración de sólidos solubles (grados Brix) y pH del fruto en 273 individuos de la retrocruza (*Lycopersicon esculentum* cv. UC82B x *L. chmielewskii* accesión LA1028) x *Lycopersicon esculentum* cv. UC82B. Se seleccionaron un subconjunto de marcadores polimórficos, 63 RFLP y 5 isoenzimas, espaciados en intervalos de aproximadamente 20 cM en los 12 cromosomas del tomate.

Se observa un QTL para cada una de las tres características señaladas en el mapa. Se declaró la presencia del QTL cuando el LOD sobrepasó el umbral de 2,5. Los intervalos de confianza señalan las regiones a ambos extremos del pico donde hay una caída del LOD de 1 (barra) ó 2 (líneas); lo que equivale a una disminución de la verosimilitud de la existencia del QTL en esa región de 10 ó 100 veces, respectivamente. Los tres QTL coinciden en ubicación, en la región cubierta por los marcadores CD42, SP y PC5. Esta conjunción de picos pudiera interpretarse como un efecto pleiotrópico, pues la misma región codifica para las tres características.



Desventajas

La prueba de detección no es realmente una prueba por intervalo, ya que no puede ser independiente de los efectos de otros QTL fuera del intervalo. Si hay más de un QTL en un cromosoma, la prueba estadística estará afectada por esos QTL y los estimados de posición y efecto del QTL probablemente estarán sesgados. Como consecuencia, este método no es eficiente con el uso de los datos, ya que la información de otros marcadores diferente a la pareja M1 y M2 no se utiliza.

Mapeo por intervalo compuesto

El mapeo compuesto se propuso para minimizar los problemas que presenta el mapeo por intervalo (Zeng 1994). Representa una variación de éste que añade a su vez la información de los marcadores externos al intérvalo mediante una regresión múltiple

$$\Phi_i = a + bg_i + \Sigma b_k x_{ik} + \varepsilon_i$$

donde

b_k = es el coeficiente de regresión parcial del k-ésimo marcador sobre el fenotipo Φ

x_{ik} = es un coeficiente conocido del k-ésimo marcador para el i-ésimo individuo.

Los estimados de b, b_k y ε se obtienen a través de las soluciones de matrices, detalladas en la página 1461 de Zeng (1994).

Detección del QTL

Las hipótesis por ser probadas son las siguientes

$H_0 : b = 0$
$H_1 : b \neq 0$

El estadístico de prueba es la Relación de Verosimilitud (LR) que se calcula como

$$LR = -2ln(L_0/L_1)$$

en la que

L_0 = función de verosimilitud bajo la hipótesis nula

El LR se comporta de acuerdo a una distribución χ^2, con grados de libertad que varían en función del número de marcadores o de individuos. Por otra parte, a través de permutaciones se puede obtener la probabilidad de conseguir un LR mayor que el observado y por medio de técnicas de re-muestreo como Bootstrapping y Jacknife, la desviación estándar del estadístico de prueba y la de los efectos de aditividad y dominancia.

El QTLCARTOGRAPHER (QTLCART, Basten *et al.* 1994.) es un algoritmo de computación que se encarga de obtener los estimados de los efectos aditivos, dominancia, epístasis y posición del QTL según este modelo. Presenta a su vez alternativas para efectuar un análisis previo de los datos según regresión simple, además de que realiza el mapeo por intervalo simple. Necesita ser alimentado con un mapa cromosómico, preferiblemente mediante el MAPMAKER, pues posee la opción de formatear la salida de este mapa para utilizarlo posteriormente.

La filosofía del programa es tratar de fijar cualquier variación (o ruido de fondo) del carácter cuantitativo que no se deba a la segregación genética dentro del intervalo bajo estudio. El QTLCART presenta diferentes submodelos que pueden probarse para detectar y estimar los efectos de un QTL. Algunos de los más importantes son:

a) Modelo 1: usa todos los marcadores para controlar el "background" genético
b) Modelo 2: usa todos los marcadores no ligados.
c) Modelo 4: usa un marcador por cromosoma (excepto el que se está probando).
d) Modelo 6: usa un número fijado de marcadores previamente asociados al carácter a través de regresión paso a paso, y una ventana de longitud variable a los lados del intervalo que no se considera para controlar el background genético.

Referencias bibliográficas

Basten C. J., Weir B. S. & Zeng Z.-B. 1994. **Zmap-a QTL cartographer**. In: Proceedings of the 5th World Congress on Genetics Applied to Livestock Production: Computing Strategies and Software, edited by C. Smith, J. S. Gavora, B. Benkel, J. Chesnais, W. Fairfull, J. P. Gibson, B. W. Kennedy and E. B. Burnside. Volume 22, pages 65-66. Published by the Organizing Committee, 5th World Congress on Genetics Applied to Livestock Production, Guelph, Ontario, Canada.

Bretó M.P., Aśins M.J. & E.A. Carbonell. 1994. **Salt tolerance in Lycopersicon species. III. Detection of quantitative trait loci by means of molecular markers**. Theor. Appl. Genet. 88: 395-401.

Lander E.S., Green P., Abrahamson J., Barlow A., Daly M.J., Lincoln S.E. & L.A. Newberg. 1987. **MAPMAKER: an interactive computer package for constructing primary genetic linkage maps of experimental and natural populations**. Genomics 1: 174-181.

Lander E. & D. Botstein. 1989. **Mapping mendelian factors underlying quantitative traits using RFLP linkage maps**. Genetics 121: 185-199.

Martínez O. & R. Curnow. 1992. **Estimating the locations and the sizes of the effects of quantitative trait loci using flanking markers**. Theor. Appl. Genet. 85:480-488.

Paterson A.H., Lander E.S., Hewitt J.D., Peterson S., Lincoln S.E. & S.D. 1988. Tanksley. **Resolution of quantitative traits into Mendelian factors by using a complete linkage map of restriction fragment length polymorphisms**. Nature 335:721-726.

Paterson A., Tanksley S. & M. Sorrels. 1991. **DNA markers in plant improvement**. Adv. Agron. 46:39-90.

Sax K. 1923. **The association of size differences with seed-coat pattern and pigmentation in *Phaseolus vulgaris***. Genetics : 552-560.

Zeng Z. 1994. **Precision mapping of quantitative trait loci**. Genetics 136: 1457-1468.

Referencias generales:

Falconer D. y T. Mackay. 1997. **Introduction to quantitative genetics**. 4ta. Edición. Longman