(19) OFICINA ESPAÑOLA DE PATENTES Y MARCAS

ESPAÑA

(11) Número de publicación: **2 464 817**

(51) Int. CI.:

***C12Q 1/68*** (2006.01)

(12) TRADUCCIÓN DE PATENTE EUROPEA T3

(96) Fecha de presentación y número de la solicitud europea: **31.10.2008 E 08845745 (2)**

(97) Fecha y número de publicación de la concesión europea: **26.02.2014 EP 2222872**

(54) Título: **Método de agrupamiento de muestras para realizar un análisis biológico**

(30) Prioridad:

**31.10.2007 EP 07119761**

(45) Fecha de publicación y mención en BOPI de la traducción de la patente:
**04.06.2014**

(73) Titular/es:

**HENDRIX GENETICS RESEARCH, TECHNOLOGY & SERVICES B.V. (100.0%)**
**Spoorstraat 69**
**5831 CK Boxmeer, NL**

(72) Inventor/es:

**VEREIJKEN, ADRIANUS LAMBERTUS JOHANNUS;**
**JUNGERIUS, ANNEMIEKE PAULA y**
**ALBERS, GERARDUS ANTONIUS ARNOLDUS**

(74) Agente/Representante:

**TOMAS GIL, Tesifonte Enrique**

## DESCRIPCIÓN

Método de agrupamiento de muestras para realizar un análisis biológico

CAMPO DE LA INVENCIÓN

[0001] La invención se refiere al campo de mediciones con resultados categóricos en muestras biológicas, más en particular a métodos para la preparación de la muestra de ensayos biológicos con resultados categóricos. La presente invención proporciona un método de agrupar muestras, el uso de dicho método para genotipificación de una variante alélica. La invención además proporciona un método de realizar un análisis en muestras múltiples, un dispositivo de agrupamiento para agrupar muestras múltiples en una muestra agrupada, un dispositivo de análisis que comprende un procesador que está dispuesto para realizar un análisis en un conjunto de muestra agrupada, un producto de programa informático en un portador que aplica un método de agrupamiento de muestras, y un producto de programa informático en un portador que aplica un método para realizar un análisis en muestras múltiples.

ANTECEDENTES DE LA INVENCIÓN

[0002] Un bioensayo es un procedimiento donde una propiedad, concentración o presencia de un analito biológico se mide en una muestra. Los ensayos biológicos son una parte intrínseca de investigación en todos campos de ciencia, más especialmente en ciencias vitales y especialmente en la biología molecular.

[0003] Un tipo particular de análisis en biología molecular se refiere a genotipificación y secuenciación. Genotipificación y secuenciación se refieren al proceso de determinar el genotipo de un individuo con un ensayo biológico. Métodos actuales incluyen PCR, secuenciación de ADN y ARN, e hibridación para micromatrices de ADN y ARN montadas sobre diferentes portadores tales como placas de vidrio o esferas. La tecnología es intrínseca para pruebas en paternidad/maternidad, en la investigación clínica para la investigación de genes asociados a la enfermedad y en otra investigación dirigida a investigar el control genético de propiedades de cualquier especie por ejemplo escaneo completo del genoma para QTL (Loci de carácter cuantitativo).

[0004] Debido a limitaciones tecnológicas actuales, casi toda la genotipificación es parcial. Es decir, solo una fracción pequeña de un genotipo del individual es determinada. En muchos casos esto es no un problema. Por ejemplo, en la prueba de paternidad/maternidad, solo se investigan 10 a 20 regiones genómicas para determinar la relación o la falta de esta, que es una fracción ínfima de la genoma humano.

[0005] Los polimorfismos de un único nucleótido (SNP) son el tipo más abundante de polimorfismo en el genoma. Con los desarrollos paralelos de densos mapas marcadores SNP y tecnologías para genotipificación SNP de alto rendimiento (p. ej. Kirov G. et al, BMC Genomics (2006), 15(7):1471-2164 and Hoh J et al, BMC Bioinformatics (2003), 22(4): 1471 - 2105), SNP han llegado a ser los marcadores de elección para muchos estudios genéticos. Un número sustancial de muestras se requiere en el mapeo y estudios de asociación o en los experimentos de selección genómica.

[0006] Para proveer para capacidades de genotipificación de alto rendimiento, tecnologías de matriz han sido desarrolladas (p. ej. US 2002/172965 A1). Tales tecnologías están disponibles de proveedores comerciales tal como Affymetrix (microarray-based GeneChip® Mapping arrays), Illumina (BeadArray™), Biotrove (Open Array™) and Sequenom (MassARRAY™). En muchas especies (seres humanos, ganado, plantas, bacterias y virus) un gran número de SNP están disponibles o se harán disponibles pronto. Las nuevas innovaciones han permitido la genotipificación del genoma entero o estudios de asociación y programas de selección de genoma entero asociado para cría de plantas y animales. Aún los costes de tales programas son todavía significativos, requieren presupuestos de hasta varios millones de dólares si las muestras son genotipificadas individualmente. Por lo tanto, los estudios dirigidos a la identificación de SNP en cualquier especie, actualmente implican análisis de solo un número limitado de individuos. La presente invención por lo tanto es de gran importancia ya que permite una reducción muy sustancial del coste de genotipificación.

[0007] Para obtener comprensión completa en la variabilidad genómica es preciso conocer la secuencia completa de (una parte pertinente de) el genoma. No obstante, el coste de determinar la secuencia completa es incluso superior al coste de genotipificación que se ha descrito en el párrafo precedente. A pesar de los costes, se espera que la secuenciación reemplace a la genotipificación para proporcionar genotipos individuales para todo el genoma o regiones específicas de este. La presente invención también proporciona métodos para reducir el coste de secuenciación.

[0008] El agrupamiento de muestras es regularmente usada en estudios en rasgos categóricos como medio para reducir costes de análisis. La presencia de la característica en el agrupamiento, que consiste en una mezcla de diferentes muestras indica la presencia de esta característica en al menos una de las muestras en ese grupo. Depósitos de ADN se usan por ejemplo para:

- estimar frecuencias alélicas en una población.

Tomando una buena muestra de individuos de la población, la frecuencia alélica pura del alelo 1 se calcula como la proporción entre el resultado para el alelo 1 y la suma del resultado para el alelo 1 y el resultado para el alelo 2 en el agrupamiento.

- estudios de asociación caso-control donde casos y controles se dividen en grupos separados, y

- reconstruir haplotipos en un número limitado de individuos y un número limitado de SNP.

En base a las frecuencias alélicas medidas en el agrupamiento, los haplotipos se pueden estimar por diferentes algoritmos tal como probabilidad máxima. El término frecuencia de haplotipo es sinónimo del término distribución conjunta de marcadores.

[0009] Una desventaja importante del agrupamiento de muestras es que la característica medida solo se identifica en el agrupamiento como conjunto, y no en cualquiera de las muestras individuales en el agrupamiento (p. ej. WO 2005/075678 A ; US2003/152942 A1 ; Lindroos K et al, NAR (2002), 30(14):1-9 and Wolford JK et al, Hum Genet (2000), 107(5):483-7). Una excepción es agrupamientos de ADN para tríos de genotipificación (padre, madre y niño) cuando dos agrupamientos en que cada uno consiste en dos individuos son creados (padre + niño y madre + niño). La frecuencia alélica observada en cada grupo es indicativa de los genotipos para los 3 individuos. Este tipo de agrupamiento de muestras proporciona una reducción de coste de 33 % pero sólo es posible con tales tríos. En todos los otros casos, las muestras agrupadas deben ser reanalizadas individualmente para proporcionar resultados para las muestras individuales.

[0010] Así, sería beneficioso proporcionar agrupamientos de muestras para tipos de muestra diferentes de los tríos, a la vez que todavía se provean resultados de la prueba para las muestras individuales dentro de ese agrupamiento.

RESUMEN DE LA INVENCIÓN

[0011] Los presentes inventores han descubierto ahora que individuos aleatorios se pueden agrupar y que genotipos individuales se pueden recuperar de tales agrupamientos cuando la aportación de cada muestra individual en el agrupamiento es una proporción fija de la cada otra muestra, es decir cuando las cantidades de muestra no son equimolares pero proporcionadas en proporciones específicas. Los resultados para muestras individuales se pueden inferir del resultado de prueba agrupado a condición de que la prueba implique una medida cuantitativa de una variable categórica, es decir que la prueba implique una característica discreta o categórica que es medida de forma cuantitativa.

[0012] De hecho, los presentes inventores han encontrado que para el estudio de la presencia de un alelo determinado en un locus determinado en un animal diploide, la mezcla en una proporción de 1:3 de una muestra de ADN de un primer animal diploide con 2 alelos posibles (A o B) en un único locus, con una muestra de ADN de un segundo animal diploide que también tiene 2 alelos posibles (A o B) en el mismo locus, produce la presencia de (2) + (2+2+2) = 8 posibilidades para cualquiera de los alelos en que mezcla, donde la señal del instrumento cuantitativo prevista a partir de un único alelo (p. ej. A) es 12,5 % de la fuerza máxima de señal de muestra. Esto significa que a una intensidad de señal medida de 37.5 % de la máxima fuerza de señal de muestra, la muestra comprende 3 x el alelo A, lo que significa que la señal no se puede derivar del primer animal diploide y pueden solo derivar del segundo animal diploide, indicando que el primer animal diploide tiene genotipo BB y el segundo animal diploide tiene genotipo AB. Asimismo, cuando la intensidad de señal medida es 50 % de la máxima fuerza de señal de muestra, todas las muestras tienen genotipo AB. Cuando la intensidad de señal medida es 0 % de la máxima fuerza de señal de muestra, entonces todas muestras tienen genotipo BB. Los 2 individuos en el agrupamiento tienen en total 3*3 genotipos posibles. Siempre y cuando la exactitud de la medición sea al menos 6.25 %, cada medición se puede asignar a un valor un octavo (1/8) de 100 % o un múltiplo de este. En general, cada resultado de medición posible se puede asignar a un valor $1/(y*((p+1)^0 + (p+1)^1 + (p+1)^2 + (p+1)^{(n-1)})) * 100$ %, donde y=2 (los dos posibles resultados para el alelo A en una posición, alelo está presente o ausente), p es el nivel de ploidía, n es el número de muestras y 100 % es la máxima fuerza de señal de muestra. En total habrá (nivel de ploidía+1)$^n$ genotipos posibles.

[0013] Ahora cuando se agrupan muestras de 3 animales (x, y y z) en una proporción de 1:3:9 (respectivamente, es decir, con un factor de agrupamiento de 3), hay en la teoría un total de 26 posibilidades para cualquiera de los alelos en esa mezcla, donde la señal cuantitativa prevista de un único alelo (p. ej. A) es 3,85 % de la máxima fuerza de señal de muestra. Esto significa que a una intensidad de señal medida de 12 % de la máxima fuerza de señal de muestra, la muestra comprende 3 x el alelo A indicando que el animal x tiene genotipo BB, el animal y tiene genotipo AB, y el animal z tiene genotipo BB. Asimismo, cuando la intensidad de señal medida es 96 % de la máxima fuerza de señal de muestra, la muestra x tiene genotipo AB, mientras que las muestras y y z tienen genotipo AA. Siempre y cuando la exactitud de la medición sea al menos 1.9 %, cada medición se puede asignar a un valor un veintiseisavo (1/26) de 100 % o un múltiplo de

este. (Para una visión de conjunto de resultados posibles para este tipo de experimento agrupado ver los ejemplos a continuación).

[0014] Los inventores han mostrado que este principio se puede usar para un gran número de análisis que implican una medición cuantitativa de un analito en una muestra, donde el resultado del análisis es categórico respecto a una calidad del analito en dicha muestra.

[0015] En un primer aspecto, la presente invención ahora proporciona un método de agrupamiento de muestras que deben ser analizadas en cuanto a una variable categórica, donde el análisis implica una medición cuantitativa de un analito, dicho método de agrupamiento de muestras comprende proporcionar un agrupamiento de n muestras donde la cantidad de muestras individuales en el agrupamiento es tal que los analitos en las muestras estén presentes en una proporción molar de $x^0$: $x^1$: $x^2$: $x^{(n-1)}$, y donde x es un número entero 2 o más alto, tal como 3, 4, 5, 6, 7, o 8, preferiblemente 2 o 3, representando el número de clases de la variable categórica (o el factor de agrupamiento) y n es el número de muestras. La anotación $x^0$: $x^1$: $x^2$: $x^{(n-1)}$ debería ser entendida como en referencia a $x^0$: $x^1$: $x^2$:...: $x^{(n-1)}$, o $x^0$: $x^1$: $x^2$: $x^1$; $x^{(n-1)}$, donde n es el número de muestras e i es un número entero incremental con un valor entre 2 y n.

[0016] Para agrupar individuos poliploides x es igual al (nivel de ploidía+1), así x=2 para un haploide, 3 para un diploide y 5 para un individuo tetraploide con dos alelos posibles en una posición única, x es también igual al número de genotipos posibles.

[0017] Asumiendo que habría tres alelos posibles luego un haploide tendría 3 genotipos posibles (x=3), un diploide tendría 6 genotipos posibles (x=6) y un triploide tendría 10 genotipos posibles (x=10). En un individuo diploide el primer alelo puede existir 0,1 o 2 veces al igual que el segundo y tercer alelo. Esto hace posible agrupar en la misma proporción ($x^0$: $x^1$: $x^2$: $x^{(n-1)}$) como con dos alelos (x es otra vez nivel de poliploidía +1). Intensidades de señal para los 3 alelos se redondean al punto de resultado más cercano $(1/(y*((p+1)^0+ (p+1)^1+ (p+1)^2+ (p+1)^{(n1)}))*100$ %, donde y=2 (alelo 1,2 o 3 está presente o ausente), p= nivel de ploidía y n=número de muestras) para encontrar el número de alelos en la muestra agrupada.

[0018] Así, la proporción entre las dos muestras individuales en el agrupamiento (como un ejemplo) es tal que los analitos en esta están presentes en una proporción molar de 1:x donde x es el número máximo de clases para la característica categórica.

[0019] Métodos donde la cantidad de las muestras individuales en el agrupamiento sirve como secuencia geométrica con proporción común 3 son especialmente adecuados para genotipificación de una variante alélica en individuos diploides, donde cada individuo tiene tres genotipos posibles. El genotipo es la característica categórica que puede tener tres variantes posibles (AA, AB y BB).

[0020] Métodos donde la cantidad de las muestras individuales en el agrupamiento sirve como secuencia geométrica con proporción común 2 son especialmente adecuados para genotipificación de una variante alélica en individuos haploides. Para un ejemplo de esto, se hace referencia a la parte experimental a continuación.

[0021] En otro aspecto, la presente invención se refiere al uso de un método de la invención como se ha descrito anteriormente, para genotipificación de una variante alélica en individuos poliploides o haploides donde el número de clases de la variable categórica (x) es igual a p+1, donde p representa el nivel de ploidía de dicho individuo. Tal uso por ejemplo permite la genotipificación de una variante alélica en un individuo diploide o haploide.

[0022] En otro aspecto, la presente invención se refiere a un método de realizar un análisis en muestras múltiples, que comprende el agrupamiento de dichas muestras según un método de la invención como se ha descrito anteriormente para proporcionar una muestra agrupada y realizar dicho análisis en dicha muestra agrupada. El resultado cuantitativo obtenido es luego redondeado al punto de resultado más próximo (determinado por el número de intervalos teóricos donde la máxima fuerza de señal de muestra se divide para cada resultado posible, ver infra), y la intensidad de señal se asigna al número total de clases de la variable categórica en la muestra agrupada. De esto la variable categórica se determina para cada muestra individual en el agrupamiento teniendo en cuenta la proporción de las diferentes muestras individuales varias en el agrupamiento.

[0023] En otro aspecto, la presente invención proporciona un método de realizar un análisis en muestras múltiples, que comprende realizar un análisis en un conjunto de muestras agrupadas obtenidas por un método de agrupamiento de muestras tal y como se define aquí anteriormente, donde dicha muestra se analiza en cuanto a una variable categórica e implica una medición cuantitativa de un analito en dicha muestra.

[0024] En una forma de realización preferida de este método, un método de realizar un análisis comprende además el paso de deducir a partir de la medición la aportación de las muestras individuales en dicho agrupamiento de muestras.

[0025] En otro aspecto, la presente invención proporciona un dispositivo de agrupamiento para agrupar muestras múltiples en una muestra agrupada que comprende unos aspiradores de muestra para suministrar una muestra agrupada y además comprende un procesador para realizar un método de agrupamiento de muestras tal y como se define aquí anteriormente.

[0026] En otro aspecto, la presente divulgación proporciona un dispositivo de análisis que comprende un procesador que está dispuesto para realizar un análisis en un conjunto de muestra agrupada obtenido por un método de agrupamiento de muestras tal y como se define aquí anteriormente, donde dicho dispositivo está dispuesto para analizar dicha muestra en cuanto a una variable categórica y para realizar una medición cuantitativa de un analito en dicha muestra.

[0027] En una forma de realización preferida de este dispositivo de análisis, el dispositivo comprende además un dispositivo de agrupamiento, de la forma más preferible un dispositivo de agrupamiento como se ha descrito anteriormente.

[0028] En otro aspecto, la presente invención proporciona un producto de programa informático bien en un portador, el cual producto de programa, cuando se carga y ejecuta en un ordenador, una red informática programada u otro equipo programable, aplica un método de agrupamiento muestras tal y como se ha definido aquí anteriormente.

[0029] En otro aspecto, la presente invención proporciona un producto de programa informático bien en un portador, el cual producto de programa, cuando se carga y ejecuta en un ordenador, una red informática programada u otro equipo programable, aplica un método para realizar un análisis en muestras múltiples, dicho método comprende la realización de un análisis en un conjunto de muestras agrupadas obtenido por un método de agrupamiento de muestras tal y como se ha definido aquí anteriormente, donde dicha muestra se analiza en cuanto a una variable categórica e implica una medición cuantitativa de un analito en dicha muestra.

[0030] En una forma de realización preferida de este producto de programa informático, dicho método comprende además el paso de agrupamiento según un método de agrupamiento de muestras tal y como se ha definido aquí anteriormente.

[0031] Usando el método de la presente invención los costes de análisis se pueden reducir inmensamente, es decir típicamente por 50 %, e incluso por 66 % o más.

DESCRIPCIÓN DE LAS FORMAS DE REALIZACIÓN PREFERIDAS

[0032] El término "variable categórica", como se utiliza en este caso, se refiere a una variable discreta tal como una característica o cualidad, por ejemplo la presencia o ausencia de un analito o un característica en esta, o una característica alélica presente o ausente en la forma heterocigótica u homocigótica en un analito. Discreto es sinónimo de categórico y se refiere a no lineal o discontinuo. Una "variable" generalmente se refiere a una característica (categórica) que mide una propiedad de una muestra. Un variable categórica puede ser binaria (que consiste en 2 clases). Un "clase" se refiere a un grupo o categoría a la que una medición puede ser asignada. Así, una variable puramente categórica es una que permitirá que la atribución de categorías y variables categórica tome un valor es decir una de diferentes categorías posibles (clases). En particular, la variable categórica se puede referir a la presencia de un marcador genético tal como un polimorfismo de nucleótido único (SNP) o cualquier otro marcador genético, un alelo, una respuesta inmune, una enfermedad, una capacidad de resistencia, color capilar, género, estado de infección de enfermedad, genotipo o cualquier otra característica o propiedad de una muestra o entidad biológica. Aunque se pueden medir numéricamente, por ejemplo como una señal de analito generada que puede ser recibida, leída y/o registrada por un dispositivo de análisis, las variables categóricas por sí mismas no tienen significado digital y las categorías no tienen ningún límite intrínseco. Por ejemplo, el género es una variable categórica que tiene dos categorías (macho y hembra frecuentemente codificados como 0 y 1) y representan categorías preferiblemente no ordenadas. El genotipo es también un variable categórica con un número de categorías preferiblemente no ordenadas (AA, Aa y aa a veces codificado como 2,1 y 0).

[0033] La muestra en aspectos de la presente invención puede ser cualquier muestra donde una variable categórica debe ser medida. La muestra puede ser una muestra biológica tal como un tejido o muestra de líquido biológico de un animal (incluyendo un humano) o una planta, una muestra medioambiental tal como una muestra de suelo, aire o agua. La muestra puede ser (parcialmente) purificada o puede ser una muestra no tratada (cruda). La muestra es preferiblemente una muestra de ácido nucleico, por ejemplo una muestra de ADN.

[0034] El analito cuya presencia o forma se mide en una prueba cuantitativa puede ser cualquier sustancia química o entidad biológica. En formas de realización preferidas, el analito es una biomolécula y la variable categórica es una variante de dicha biomolécula. Preferiblemente, la biomolécula es un ácido nucleico, en particular un polinucleótido, tal como ARN, ADN, y la variante puede por ejemplo ser un polimorfismo de nucleótido en dicho polinucleótido, por ejemplo una variante alélica, de la forma más preferible un SNP, o la identidad de base de una posición de nucleótidos particular.

[0035] El analito tal y como se define aquí puede así ser una molécula de ADN que exhibe una variable categórica determinada (p. ej. la identidad de base de una posición de nucleótidos particular en esa molécula de ácido nucleico, con un valor categórico de A, T, C o G). La identidad de base de una posición de nucleótidos particular se puede medir usando una prueba cuantitativa, por ejemplo en base a la fluorescencia derivada de una copia de ADNc que incorpora un análogo fluorescente de dicho nucleótido, tal como se conoce en la técnica de secuenciación del ADN. El nivel cuantitativo de la fluorescencia emitida por dicho análogo en una posición particular del ADN y medido por un dispositivo de análisis, es luego asignado a un valor categórico para esta posición de nucleótidos, por ejemplo como una adenina para esta posición.

[0036] En determinar la identidad de base de una posición de nucleótidos particular, la invención concierne a reunir muestras individuales de las cuales la secuencia de nucleótidos de un ácido nucleico particular debe ser determinada. La idoneidad del método de la invención para ensayos de secuenciación (análisis) se puede entender cuando la realización de esos ensayos de secuenciación implica la determinación de una señal de bien una de cuatro bases posibles donde la presencia o ausencia de una señal para cualquier base particular en una posición determinada en por ejemplo un gel de secuenciación corresponde a la presencia o ausencia de esta identidad de base en una posición de nucleótidos particular dentro de dicho ácido nucleico. El agrupamiento de dos muestras antes de realizar el gel de secuencia en la proporción como se describe en este caso permitirá la determinación del origen de cualquier señal particular y así de la secuencia para cada ácido nucleico individual.

[0037] El "analito" puede ser un polipéptido, tal como una proteína, un péptido o un aminoácido. El analito también puede ser un ácido nucleico, una sonda de ácido nucleico, un anticuerpo, un antígeno, un receptor, un hapteno, y un ligando para un receptor o fragmentos de este, un marcador (fluorescente), un cromógeno, radioisótopo. Hecho, el analito se puede formar por cualquier sustancia química o sustancia física que se pueda medir de forma cuantitativa, y que pueda utilizarse para determinar la clase de la variable categórica.

[0038] El término "nucleótido", como se utiliza en este caso, se refiere a un compuesto que comprende una base purina (adenina o guanina) o pirimidina (timina, citosina o uracilo) enlazada al carbono C-1 de un azúcar, típicamente ribosa (ARN) o deoxiribosa (ADN), y además que comprende uno o varios grupos fosfato enlazados al carbono C-5 del azúcar. El término incluye referencia a las unidades estructurales individuales de un ácido nucleico o polinucleótido donde unidades de azúcar de nucleótidos individuales se enlazan a través de un puente de fosfodiéster para formar un esqueleto de fosfato de azúcar con bases de purina o pirimidina pendientes.

[0039] El término "ácido nucleico" como se utiliza en este caso, incluye referencia a un polímero de desoxirribonucleótido o ribonucleótido, es decir un polinucleótido, bien en la forma monocatenaria o bien bicatenaria, y a menos que se limite lo contrario, abarca análogos conocidos con la naturaleza esencial de nucleótidos naturales en cuanto a que ellos hibridan con ácidos nucleicos monocatenarios de modo similar a nucleótidos de origen natural (e. g., ácidos nucleicos peptídicos). Un polinucleótido puede ser de longitud completa o una subsecuencia de un gen nativo o estructural heterólogo o regulador. A menos que se indique lo contrario, el término incluye referencia a la secuencia especificada al igual que la secuencia complementaria de esta. Así, ADNs o ARN con esqueleto modificado por estabilidad o por otras cuestiones son "polinucleótidos" ya que ese término está destinado aquí. Por otra parte, ADNs o ARN que comprenden bases inusuales, tal como inosina, o bases modificadas, tal como bases tritiladas, por nombrar solo dos ejemplos, son polinucleótidos ya que el término se utiliza en este caso.

[0040] El término "medida cuantitativa" se refiere a la determinación de la cantidad de un analito en una muestra. El término "cuantitativo" se refiere al hecho de que la medición se puede expresar en valores numéricos. El valor numérico se puede referir a una dimensión, tamaño, extensión, cantidad, capacidad, concentración, altura, profundidad, anchura, ancho, longitud, peso, volumen o área. La medición cuantitativa puede implicar la intensidad, altura de pico o superficie de pico de una señal de medición, tal como una señal cromogénica o de fluorescencia, o cualquier otra señal cuantitativa. En general, cuando se determina la presencia o forma de un analito, la medición implicará una señal de instrumento. Por ejemplo, cuando se determina la presencia de un SNP, la medición implicará una señal de hibridación, y la medición típicamente suministrará una intensidad de fluorescencia como se ha medido por un fluorímetro. Cuando se determina la presencia de una respuesta inmune, la medición implicará medición de graduación de un anticuerpo y la medición también puede ser típicamente proporcionada como una intensidad de fluorescencia. La necesidad de medición no proporciona un resultado de medición continua, pero se puede referir a intervalos discretos o categorías. La medición también puede ser semicuantitativa. Mientras la medición se pueda determinar en $2^{n-1}$, $3^{n-1}$ o $x^{n-1}$ intervalos parciales y preferiblemente proporcionales de la máxima fuerza de señal de muestra (dependiendo de si el agrupamiento sirve como secuencia geométrica con proporción común 2,3 o x, respectivamente, donde n es el número de muestras en el agrupamiento), la medición es en principio adecuada.

[0041] El término "agrupamiento", como se utiliza en este caso, se refiere al agrupamiento o mezcla de muestras para los fines de maximizar la ventaja a los usuarios. En particular el término "agrupamiento" se refiere a la preparación de una recogida de muestras múltiples para representar una muestra de valor ponderado. La incorporación de muestras múltiples

en una muestra simple es normalmente realizada mediante la mezcla muestras. En la presente invención, la mezcla requiere un pesaje atento de la cantidad de las muestras individuales, donde la cantidad de analito presente en cada muestra es decisiva. Cuando una muestra A tiene una cantidad de analito de 2 g/l y muestra B tiene una cantidad de 1 g/l, estas muestras se deben agrupar en una proporción de volumen de 1:6 para proporcionar la proporción de analito 1:3.

[0042] Cuando dos muestras son por ejemplo agrupadas en una proporción de 1:3 o cuando tres muestras se agrupan en una proporción de 1: 3: 9 como se indica en formas de realización de la presente invención, las frecuencias posibles de las variantes en los depósitos se fija por los objetivos de intervalos de 12,5 % y 3,85 %, respectivamente. Los objetivos de estos intervalos se denominan en este caso como "puntos de resultado" y son equivalentes a los incrementos de paso de la medición cuantitativa hasta alcanzar la máxima fuerza de señal de muestra.

[0043] El término "secuencia geométrica" se refiere a una secuencia de números donde la proporción entre dos términos consecutivos cualesquiera es la misma. En otras palabras, el término siguiente en la secuencia se obtiene multiplicando el término precedente por el mismo número cada vez. Este número fijo se llama proporción común para la secuencia. En una secuencia geométrica de la presente invención, el primer término es 1 y la proporción común es 2 o 3, dependiendo del tipo de muestra.

[0044] El término "máxima fuerza de señal de muestra" se refiere a la señal obtenida del agrupamiento cuando todas muestras en el agrupamiento proporcionan una señal positiva, es decir cuando el 100 % de las muestras individuales son positivas para el analito evaluado. La máxima fuerza de señal de muestra se puede determinar por cualquier método adecuado. Por ejemplo, 50 muestras individuales se pueden medir separadamente para determinar su composición en cuanto al número de eventos discretos presentes entre estas muestras, y posteriormente estas muestras pueden luego ser medidas en un experimento agrupado, donde las fuerzas de señal medidas para la muestra agrupada se muestran en la misma proporción que serían obtenidas sumando todas las fuerzas de señal de todas las muestras individuales.

[0045] Un método de la presente invención se puede realizar con cualquier número de n muestras. No obstante, en la práctica, el número máximo para n se fija por la exactitud del método de medición, es decir la exactitud con la cual una distinción estadísticamente sólida entre dos puntos de resultado consecutivos puede ser determinada. La exactitud (desviación típica) del método debe estar en acuerdo con esto.

[0046] Aplicaciones del método de la presente invención incluyen, pero de forma no limitativa, métodos de genotipificación. La genotipificación basada en agrupamientos de ADN tiene muchas aplicaciones. Los genotipos se pueden usar para mapeo, asociación y diagnóstico en todas especies. Ejemplos de genotipificación específica incluyen a) genotipificación en seres humanos, tales como diagnósticos médicos pero también de seguimiento de tipificiaciones individuales después del agrupamiento del estudio caso - control; b) genotipificación en ganado, tales como tipificaciones individuales en estudios QTL, en los métodos de gen candidato y en las aplicaciones de selección amplia de genoma, y c) genotipificación en plantas por ejemplo para mapeo y estudios de asociación.

[0047] El agrupamiento también puede usarse al secuenciar seres humanos, ganado, plantas, bacterias, virus. Más específicamente el agrupamiento de muestras individuales para secuenciación es pertinente cuando las secuencias de dos o más individuos deben ser comparadas.

[0048] Un método de la presente divulgación para agrupamiento de muestras comprende tomar una submuestra de al menos una primera muestra y una submuestra de al menos una segunda muestra, donde dicha primera y segunda submuestra se incorporan en un único contenedor en cuanto a proporcionar una mezcla de las dos submuestras en forma de una muestra agrupada y donde la proporción de dichas primeras y segundas submuestras en dicha muestra agrupada es 1: 3 o 3: 1 en base la concentración de analito en esta como se describe en este caso. De forma similar, cuando tres muestras están agrupadas (frase que se refiere al hecho de que tres submuestras están mezcladas) la proporción entre la primera, segunda y tercera submuestra (en cualquier orden) que debe ser obtenida en la muestra agrupada es 1: 3: 9 prescrito aquí. Las frecuencias posibles de las variantes en los depósitos se fija por los objetivos de intervalos de 12.5 % y 3.85 %, respectivamente. Las objetivos de estos intervalos se denominan en este caso como "puntos de resultado" y son equivalentes a los incrementos de paso hasta alcanzar la máxima fuerza de señal de muestra.

[0049] Un método de agrupamiento tal y como se define aquí se puede realizar por (el uso de) un dispositivo de agrupamiento. Tal dispositivo adecuadamente comprende un colector de muestra dispuesto para recolectar y entregar una cantidad definida de muestra, por ejemplo en forma de un volumen definido (pero variable). Un colector de muestra adecuada es una pipeta tal como generalmente se aplica en la entrega robótica de muestras y sistemas de tratamiento usado en laboratorios. Tales sistemas de robótica son normalmente equipos de parte superior de banco, que comprenden adecuadamente uno o varios de unos estadios de procesador de microplaca, estaciones reactivas, aspiradores de placa de filtración, y módulos de pipeta robótica basados en neumática y puntas de pipeta desechables. Estos sistemas de robot de muestra son muy adecuados para realizar el método de la presente invención ya que ellos están en última instancia

diseñados para combinar volúmenes de líquido diferentes de distintas muestras en uno o varios tubos de reacción. Por lo tanto, está en el nivel de habilidad del experto en la técnica adaptar tal sistema robótico de pipeta para ejecutar la tarea de combinar diferentes volúmenes de líquidos de distintas muestras en una única muestra agrupada. Tal sistema robótico de pipeta es no obstante solo una forma de realización adecuada de un dispositivo de agrupamiento de muestras para agrupamiento de muestras múltiples en una muestra agrupada, dicho dispositivo comprende un colector de muestra para colectar muestras de múltiples viales de muestra y para entregar muestras en un único vial de agrupamiento para proporcionar una muestra agrupada, y además comprende un procesador que está dispuesto para realizar un método de agrupamiento de muestras tal y como se define aquí. El término "procesador", como se utiliza en este caso, se destina a incluir referencia a cualquier dispositivo de computación donde instrucciones almacenadas y recuperadas de un dispositivo de memoria u otro de almacenamiento son ejecutadas utilizando una o varias unidades de ejecución, tal como una unidad que comprende un dispositivo de pipeta y un brazo de robótica para el movimiento de dicho dispositivo de pipeta entre viales de muestra y agrupamiento de viales de un sistema robótico de pipeta. El término vial debería ser interpretado aproximadamente y puede incluir referencia a un punto de análisis en una colección. Procesadores conforme a la invención pueden por lo tanto incluir, por ejemplo, ordenadores personales, ordenadores centrales, ordenadores de red, estaciones de trabajo, servidores, microprocesadores, DSP, circuitos integrados específicos de solicitud (ASIC), al igual que partes y combinaciones de estos y otros tipos de procesadores de datos. Dicho procesador está dispuesto para recibir instrucciones de un programa de ordenador que aplica un método de agrupamiento de muestras según la presente invención en un dispositivo de agrupamiento tal y como se ha definido aquí anteriormente.

[0050] Un método de agrupamiento de muestras que deben ser analizadas en cuanto a una variable categórica, donde el análisis implica una medición cuantitativa de un analito, dicho método de agrupamiento de muestras comprende proporcionar una agrupación de n muestras donde la cantidad de muestras individuales en el agrupamiento es tal que los analitos en las muestras estén presentes en una proporción molar de $x^0$: $x^1$: $x^2$: $x^{(n-1)}$, y donde x es un número entero de 2 o más alto representando el número de clases de la variable categórica

[0051] Mientras el método de agrupamiento es bastante directo, y puede ser descrito en términos de fórmulas relativamente simples, el método de análisis de muestras agrupadas como se describe en este caso es más intrincado.

[0052] Como se describe en este caso, una variable categórica (p. ej. genotipo) puede tomar un valor es decir una de las diferentes categorías posibles (BB, AB, AA). Estas categorías coinciden con clases de intervalos de resultado. Las categorías se determinan mediante la realización de una medición cuantitativa en un analito (ADN) para un parámetro (p. ej. fluorescencia), y asignando clases para estos valores del parámetro en base a categorización de resultados de análisis, cada uno de las cuales clases representa una variante para dicha variable categórica (ver figura 7). En general, el número total de resultados de análisis posibles (resultados) depende de la naturaleza de la variable categórica. Por ejemplo en el caso de un genotipo de un organismo diploide, el nivel de ploidía determina el número de resultados de análisis posibles. En términos generales, la naturaleza de la variable categórica puede incluir la presencia de diferentes números de variantes o conjuntos del analito (se repite en la Fig. 7) dentro de una muestra. También, el número total de posibles resultados de análisis depende de los valores categóricos diferentes posibles que una repetición puede coger. Un ejemplo del número de resultados de análisis posibles se proporciona en tabla 1.

Tabla 1. Número total de resultados de análisis posibles (resultados) para una medición cuando esta está compuesta por repeticiones del mismo evento.

| Posibles valores para | Número de repeticiones en una muestra | | | |
|---|---|---|---|---|
| 1 repetición | 1 | 2 | 3 | 4 |
| 2 | 2 | 3 | 4 | 5 |
| 3 | 3 | 6 | 10 | 15 |
| 4 | 4 | 10 | .. | .. |
| 5 | 5 | 15 | .. | $(n+_k\ k+1)$ |

N represents the number of possible categorical values or variants for one repeat and k is the number of repeats within the sample. The values provided in the table are calculated based on the formula $(n+_k\ k+1)$.

[0053] Por ejemplo, el genotipo de un diploide individual (2 repeticiones de un alelo dentro de una muestra) es igual a 3 (AA, AB y BB) porque un alelo puede tener solo dos variantes diferentes (A o B). Un triploide (3 repeticiones de un alelo) puede tener 4 genotipos diferentes (AAA, AAB, ABB y BBB),

[0054] Un grupo sanguíneo para un individo es una repetición con cuatro variantes diferentes (A, B, AB o 0).

[0055] La fórmula en la tabla 1 se mantiene para situaciones donde no es importante qué repetición de la variante es medida. Por ejemplo para genotipificación no hay diferencia entre genotipo AB y genotipo BA. No obstante, en el caso de la identidad de las repeticiones es importante luego la fórmula para calcular el número total de posibles resultados de análisis es $n^k$. Esta fórmula luego reemplaza la fórmula ($n +_k K + 1$) en la tabla 1. También todos valores de tabla cambian por consiguiente. Para una situación con 2 repeticiones y 2 resultados posibles para una repetición habrá 4 resultados. Con 3 repeticiones y 3 resultados posibles para una repetición habrá 9 resultados diferentes.

[0056] El número total de resultados de análisis posibles es aplicado aquí como proporción de agrupamiento (p. ej. 1:3:9) y proporciona directamente lo que se llama el "factor de agrupamiento" (3 en el caso de 1:3:9). Por ejemplo cuando se agrupan individuos haploides para genotipificación hay una repetición con 2 variantes posibles por repetición. En tales casos el factor de agrupamiento es igual al 2 (es el número de resultados en la tabla 1).

[0057] Al agrupar 4 individuos luego necesitan ajustarse a la proporción $2^0:2^1:2^2:2^3$.

[0058] Cuando se agrupan individuos diploides el factor de agrupamiento es 3. El agrupamiento de 3 individuos necesita ajustarse a la proporción $3^0:3^1:3^2$.

[0059] El número total de resultados en una agrupación luego es igual a la fórmula siguiente;

$$\textbf{Resultados de agrupamiento totales} = \textbf{factor de agrupamiento}^{\textbf{número de muestras}}$$

[0060] El incremento para las intensidades de señal es luego igual a;

$$\textbf{Incremento} = 1/(\textbf{factor de agrupamiento}^{\textbf{número de muestras}} - 1)*100\ \%$$

O

$$1/(y*((\text{factor de agrupamiento})^0 + (\text{factor de agrupamiento})^1 + (\text{factor de agrupamiento})^2 + \ldots\ldots\ldots..(\text{factor de agrupamiento})^{(n-1))}) *100\ \%$$

Donde n es el número de muestras y y= factor de agrupamiento menos 1.

[0061] Si intensidades de medición están presentes para todas variantes para una repetición (son todos los valores menos uno debido a que el que falta puede luego ser calculado como 1 menos las intensidades de los otros) la fila de parte superior en la tabla 1 es seguida porque esta puede verse como presente o ausente para cada valor de esta repetición que corresponde a 2 resultados posibles para esta repetición. Véase el ejemplo anterior donde 3 alelos posibles se asumen en vez de 2 y donde uno puede medir 3 intensidades de luz diferentes en lugar de 2 (rojo y verde).

[0062] Si hay solo una única medición la tabla 1 puede ser seguida.

[0063] Un método de la presente invención para analizar muestras agrupadas como se contempla aquí comprende el rendimiento de una medición para el analito requerido en dicha muestra agrupada. En el registro de un resultado de medición, por ejemplo una señal de instrumento, el análisis luego implica una serie de pasos que se ejemplifican en gran detalle en los ejemplos proporcionados aquí a continuación.

[0064] Realización de un análisis en un conjunto de muestra agrupada obtenido por un método de la invención donde dicha muestra se analiza en cuanto a una variable categórica, implica una medición cuantitativa de un analito en dicha muestra. El analito es una sustancia química o sustancia física o entidad un parámetro del cual es indicativo para la presencia o ausencia de al menos una variante de dicho variable categórico. Por ejemplo, cuando se determina como una variable categórica el genotipo de un organismo, que tiene alelos variantes A o B, el analito es el ADN del organismo, una sonda de ADN o un marcador genético y el valor absoluto de un parámetro de este analito se puede correlacionar directamente con la presencia (o ausencia) de la variante. La medición cuantitativa para el analito generalmente implicará una intensidad de fluorescencia, una intensidad de radioisótopo, o cualquier medición cuantitativa como un valor para el parámetro de analito. Valores de medición más allá de un umbral determinado o valor categórico generalmente indicarán la presencia de la variante. La medición cuantitativa de un analito en una muestra así se refiere a un analito que señala la presencia o ausencia de una variante de esta variable categórica que se debe analizar en dicha muestra.

[0065] Esencialmente, en un método que analiza una muestra agrupada obtenida por un método de agrupamiento de muestras como se ha descrito aquí, la aportación de las muestras individuales en dicho agrupamiento, es decir, el resultado

para las muestras individuales en el agrupamiento, es determinado de la siguiente manera.

[0066] Primero la máxima fuerza de señal de muestra para un análisis determinado "A" que debe ser realizado en un agrupamiento de n muestras se determina y ajusta a 100 % señal. La fuerza de señal de muestra máxima es la fuerza de señal que se logra cuando el 100 % de las muestras en una agrupación de n muestras es positiva para la variable categórica. La fuerza de señal de muestra máxima se puede determinar mediante una agrupación de prueba de n muestras de referencia positiva y determinando la señal de medición, donde dichas muestras de referencia positiva son positivas con respecto a la variable categórica, y donde n es el número de muestras en los agrupamientos sobre el que el análisis "A" es realizado. La fuerza de señal de muestra máxima para el análisis "A" es registrada o almacenada en la memoria informática para uso posterior. A continuación, el analito de interés se mide en una muestra agrupada obtenida por un método de la presente invención mediante la realización del análisis "A", por el cual la fuerza de señal de la muestra agrupada para el analito es determinada. La fuerza de señal resultante para el analito en la muestra agrupada es registrada, redondeada al punto de resultado más cercano tal como se ha definido anteriormente y opcionalmente almacenada, y luego comparada con la fuerza de señal máxima. Adecuadamente, esta comparación se puede realizar de la siguiente manera. En general, cada resultado de medición posible se puede asignar a un valor $1/(y*(3^0 + 3^1 + 3^2 + 3^{(n-1)}))* 100$ %, donde n es el número de muestras agrupadas, y es un número entero de 2 que representa que "A" está presente o ausente y 100 % es la máxima fuerza de señal de muestra. La anotación $y*(3^0 + 3^1 + 3^2 + 3^{(n-1)})$ debería ser entendida como en referencia a $y*(3^0 + 3^1 + 3^2 + 3^i + 3^{(n-1)})$, donde n es el número de muestras e i es un número entero incremental con un valor entre 2 y n. Por ejemplo para y=2 clases de una variable categórica (marcador ausente y marcador presente), y una agrupación de 4 muestras, con el conjunto de máxima fuerza de señal de muestra a 100 % usando 4 muestras de referencia positiva, hay en total $2*(3^0 + 3^1 + 3^2 + 3^3) = 2 + 6 + 18 + 54 = 80$ puntos de resultado, donde cada resultado de medición posible se puede asignar a un valor 1/80 * 100 % = 1,25 % o un múltiplo de este.

[0067] El resultado para cada muestra en una agrupación de muestras puede ser leído de una simple tabla de resultado, que se puede almacenar en la forma informática legible en una memoria informática, y la cual tabla asigna para cada punto de resultado de pasos incrementales de $1/(y*(3^0 + 3^1 + 3^2 + 3^{(n-1)}))$ * 100 % entre 0 % y 100 % de la fuerza de señal de muestra máxima el valor correspondiente para cada muestra individual en el agrupamiento. Por ejemplo tal tabla de resultados es la tabla como se proporciona en la Tabla 2 a continuación.

[0068] El análisis se completa asignando a cada una de las diferentes submuestras en dicha muestra agrupada la variable categórica.

[0069] Un método de analizar una muestra agrupada tal y como se define aquí se puede realizar por un dispositivo de análisis. Un dispositivo de análisis de la presente invención comprende un procesador que está dispuesto para realizar un análisis en la muestra agrupada obtenido por un método para reunir muestras como se ha descrito anteriormente, donde dicho dispositivo está dispuesto para análisis dicha muestra en cuanto a una variable categórica y para realizar una medición cuantitativa de un analito en dicha muestra. Como se ha indicado anteriormente, la característica única del dispositivo de análisis es que está dispuesto para analizar una muestra agrupada en cuanto a una variable categórica en cada muestra individual en dicha agrupación y para realizar una medición cuantitativa de un analito en dicha muestra. Esencialmente, el dispositivo de análisis está dispuesto para medir y analizar el resultado de medición obtenido para la muestra agrupada e inferir de este resultado la variable categórica en cada muestra individual en una agrupación. Tal dispositivo adecuadamente comprende una unidad de lectura de señal para la medición de la señal de analito en la muestra agrupada. El dispositivo de análisis además comprende adecuadamente una memoria para almacenar el resultado de medición y la tabla de resultado como se ha descrito anteriormente. El dispositivo de análisis además comprende adecuadamente un procesador dispuesto para recuperar datos de memoria y/o de la unidad de lectura, y dispuesto para realizar un cálculo y para realizar un proceso reiterativo donde el resultado de medición para la muestra agrupada se comparan con y se asignan a los resultados correspondientes para las muestras individuales en dicha agrupación que utiliza la tabla de resultado a la que se hace referencia anteriormente; una interfaz de entrada/salida para introducir datos de muestra en la memoria o procesador; y una pantalla conectada a dicho procesador. El procesador está dispuesto para recibir instrucciones de un programa de ordenador que aplica un método de análisis de muestras según la presente invención en un dispositivo de análisis tal y como se define aquí anteriormente. El término "procesador" como se utiliza en este caso se destina a incluir referencia a cualquier dispositivo de computación donde instrucciones recuperadas de un dispositivo de memoria u otro de almacenamiento son ejecutadas utilizando una o varias unidades de ejecución, tal como una unidad de lectura de señal para recibir una muestra agrupada y para realizar la medición de un analito determinando la señal de dicho analito en una muestra o una muestra agrupada.

[0070] Un dispositivo de análisis de la presente invención además incluye el dispositivo de agrupamiento de la invención.

[0071] La invención además proporciona un producto de programa informático en un portador, el cual producto de programa, cuando se carga y ejecuta en un ordenador, una red informática programada u otro equipo programable, aplica un método de agrupamiento de muestras como se ha descrito anteriormente. Esencialmente, el producto de programa informático se

puede almacenar en la memoria del dispositivo de agrupamiento de la invención y se puede ejecutar por un procesador de dicho dispositivo proporcionando dicho procesador con un conjunto de instrucciones que corresponde con los diferentes pasos de proceso del método de agrupamiento.

[0072] La invención además proporciona un producto de programa informático en un portador, el cual producto de programa, cuando se carga y ejecuta en un ordenador, una red informática programada u otro equipo programable, aplica un método para realizar un análisis en muestras múltiples, dicho método comprende la realización de un análisis en un conjunto de muestra agrupada obtenido por un método de agrupamiento de muestras como se ha descrito anteriormente, donde dicha muestra se analiza en cuanto a una variable categórica e implica una medición cuantitativa de un analito en dicha muestra. Esencialmente, el producto de programa informático se puede almacenar en la memoria del dispositivo de análisis de la invención y se puede ejecutar por un procesador de dicho dispositivo proporcionando dicho procesador con un conjunto de instrucciones que se corresponden con los diferentes pasos del proceso del método de análisis. En el producto de programa informático para realizar un análisis, el método introducido en las instrucciones de software comprende además el paso de agrupamiento muestras como se ha descrito anteriormente.

[0073] La presente invención ahora será ilustrada por medio de los siguientes ejemplos no limitativos.

EJEMPLOS

**Ejemplo 1**

*Ejemplo de genotipificación de muestras individuales diploides en cuanto a la presencia de SNP usando 1 agrupamiento de 50 individuos para estandarización*

[0074] Paso 1) 50 individuos fueron evaluados separadamente.

[0075] Para cada SNP y cada individuo nosotros obtuvimos una intensidad para fluorescencia roja (presencia de alelo) y fluorescencia verde (ausencia de alelo) usando dos fluorocromos diferentes en un formato de microarray. La proporción entre las intensidades rojo y verdes no es siempre 1 (o 0) para un animal homocigótico o 0.5 para un animal heterocigótico.

[0076] Los datos en tipificaciones individuales fueron usados para calcular los factores de corrección de las intensidades de señal para todos SNP tipificados.

[0077] Para obtener el factor de corrección más importante (K), un factor de corrección frecuentemente usado para corregir los datos para cualquier rendimiento desigual al representar los alelos, utilizamos señales de genotipos heterocigóticos. Si genotipos heterocigóticos no estaban presentes, asumimos que el SNP estudiado no se está segregando en la población bajo investigación y por lo tanto los resultados para este SNP en los depósitos deberían ser omitidos.

[0078] La omisión de SNP debido a ausencia de heterocigotos en la muestra de 50 individuos puede tener como consecuencia que información en el SNP con baja MAF (frecuencia de alelo menor) podría ser perdida. Para muchas aplicaciones (tal como selección amplia de genoma) esto no es perjudicial porque SNP con frecuencias de alelo menor muy bajas no contribuyen mucho a la exactitud y luego se puede tomar una decisión de no usar datos en estos SNP o no aplicar el factor de corrección.

[0079] El primer factor de corrección (K) que nosotros usamos fue;

$$\mathbf{K = avg\ (Xraw/Yraw)}$$

Donde Xraw es la intensidad medida para rojo, e Yraw es la intensidad medida para verde. Este valor fue determinado de las muestras individualmente genotipadas con genotipo AB.

[0080] En lugar de usar el resultado medio de todas las esferas para un genotipo nosotros también podemos usar los resultados de todas las esferas separadas. Así de una muestra nosotros usamos el resultado medio para Xraw y Yraw o para X e Y o nosotros usamos los resultados de todas esferas separadas de esta muestra.

[0081] Los otros factores de corrección fueron **AAavg** y **BBavg**. AAavg es el promedio de las frecuencias de alelo sin corregir de los genotipos AA. Este valor está previsto que sea próximo a 1. BBavg es el promedio de las frecuencias de alelo sin corregir de los genotipos BB. Este valor está previsto que sea próximo a 0. **AAavg** y **BBavg** fueron calculados utilizando las fórmulas:

$$\mathbf{AAavg = (avg\ (Xrax/Xraw+Yraw)))}$$

Y

**BBavg = (avg (Xraw/(Xraw+Yraw)))**

[0082] Paso 2) un agrupamiento de prueba fue construido incluyendo los 50 individuos del paso 1 anterior. Con este fin la concentración de ADN en ng/µl fue medida en cada muestra individual utilizando un espectrofotómetro de Nanodrop (NanoDrop Technologies, EE. UU.). Todas las muestras de ADN fueron luego diluidas a una concentración estándar de 50 ng/µl antes del agrupamiento en una única muestra. En el agrupamiento de prueba así obtenido nosotros estimamos frecuencias de alelo bien sin corregir o bien en base en a factores de corrección encontrados en el primer paso.

[0083] La frecuencia de alelo sin corregir para el alelo de A se calcula como una proporción entre la intensidad roja dividido por la suma de ambas intensidades de la siguiente manera:

**Frecuencia de alelo sin corregir = Xraw/(Xraw+Yraw)**

[0084] La primera corrección para frecuencia de alelo que aplicamos fue

**Frecuencia de alelo corregida = Xraw/(Xraw+K*Yraw)**

[0085] La segunda corrección que aplicamos fue una normalización.

**Frecuencia de alelo normalizada = (Frecuencia de alelo corregida – BBavg)/AAavg**

[0086] Para ambas corrección y normalización nosotros usamos los 3 genotipos para cada SNP separadamente a partir de las muestras individuales,

[0087] El orden de exactitud de frecuencias de alelo estimado fue: normalizado, (más preciso) corregido (intermedio) y sin corregir (mínimo preciso).

[0088] Esto significa que si no había individuos heterocigóticos en el paso 1 el factor de corrección K se fijó a 0,5, y si no había individuos homocigóticos los factores de corrección AAavg y BBavg se fijaron a 1 y 0, respectivamente.

[0089] Paso 3) nosotros comparamos las frecuencias de alelo calculadas en tipificaciones individuales y en base a los resultados del agrupamiento de prueba. A partir de esto nosotros estimamos un polinomio de cuarto grado donde los resultados reales están en el eje X. Véase figura 1 para un resultado de genotipificación en individuos evaluados separadamente y en el agrupamiento con casi 18000 SNP. La genotipificación fue hecha utilizando el ensayo 8K Chicken SNP iSelect Infinium ensayo (Illumina Inc, EEUU), con SNP uniformemente distribuidos en todo el genoma de pollo (van As et al., 2007). Detalles en el ensayo, flujo de trabajo y chip se pueden encontrar en el sitio web de Illumina (http://www.illumina.com/pages.ilmn?ID=12).
A partir de este polinomio nosotros calculamos la frecuencia de alelo predicha en el agrupamiento de prueba cuando la frecuencia conocida de individuos serían 0, 0.05, 0.1,0.15------0.9,0.95 y 1.

[0090] Poniendo estos resultados en un segundo gráfico con las frecuencias reales en el eje Y, obtuvimos factores de corrección para el tercer paso de corrección, véase figura 2.

[0091] Después de la aplicación de estos factores de corrección, las frecuencias de alelo en el agrupamiento de prueba mostraron una relación lineal con las frecuencias reales, véase figura 3.

[0092] En este experimento con aproximadamente 18,000 SNP sobre el 96 % de las frecuencias de alelo medidas en el agrupamiento de prueba de 50 individuos (y corregido como se describe) estuvieron en el rango de + o - 6,25 % en comparación con los resultados de tipificaciones individuales.

[0093] Para aplicaicón de la invención, los 3 pasos precedentes son preferiblemente realizados antes del análisis real como una "calibración" para mejorar la exactitud del análisis. Estos pasos no obstante no necesitan ser realizados cada vez. La calibración de las mediciones (si se realiza) luego está seguida por: Paso 4) construir agrupamientos de ADN de 2,3 o n individuos en la proporción 1: 3,1: 3: 9 o 1: $3^1$: $3^2$: $3^{(n-1)}$., y someter los depósitos a la medición para genotipificación, donde intensidades de señal se determinan para rojo y verde en un microarray que utiliza el ensayo 18K Chicken SNP iSelect Infinium (véase supra).

[0094] Paso 5) con los factores de corrección encontrados en el paso 1 y paso 3 las frecuencias de alelo se pueden calcular

a partir de las intensidades de señal resultante en el agrupamiento. Con dos individuos en un agrupamiento las frecuencias corregidas predichas dan los puntos de resultado 0 %, 12. 5 %, 25. 0 %, 37. 5 %, 50. 0 %, 62. 5 %, 75. 0 %, 87. 5 % y 100 %.
Redondeo debería ser hecho al punto de resultado más próximo. Los genotipos de los dos individuos se pueden derivar de los resultados como se indica en la tabla 2.

[0095] Con 3 individuos en un agrupamiento el redondeo debería ser hecho al punto de resultado más próximo donde intervalos entre puntos de resultado son 3,85 % (100/($3^3$ -1)) etc.

[0096] Cuanto más cortos sean los intervalos entre los puntos de resultado consecutivos, más precisas se necesita que sean las lecturas de intensidades para permitir la asignación apropiada de un resultado particular para uno de los puntos de resultado. Lecturas más precisas se hará factibles con otro desarrollo de la técnica de genotipificación.

[0097] Para la situación con 2 individuos en una agrupación uno puede decidir usar solo el SNP donde la frecuencia estimada y corregida de alelo en el agrupamiento cae en el ±6,25 % del rango de la frecuencia real en los individuos (véase líneas rojas en la figura 3).

Tabla 2. Puntos de resultado de frecuencias de alelo en muestras agrupadas y genotipos inferidos de los dos individuos en el agrupamiento para un SNP con alelo A y C

| Frecuencia de alelo A en muestra agrupada | Genotipo inferido de individuo 1 (presente en el agrupamiento en 1 parte) | Genotipo inferido de individuo 2 (presente en el agrupamiento en 3 parte) |
|---|---|---|
| 0 | CC | CC |
| 12.5 | AC | CC |
| 25 | AA | CC |
| 37.5 | CC | AC |
| 50 | AC | AC |
| 62.5 | AA | AC |
| 75 | CC | AA |
| 87.5 | AC | AA |
| 100 | AA | AA |

[0098] SNP que muestran una diferencia mayor de un 6,25 % entre resultados agrupados y resultados individuales (en el paso 3) deberían ser omitidos si ninguna otra información está disponible para inferir genotipos individuales.

[0099] Información adicional para inferir genotipos individuales se puede derivar del pedigree de los individuos o de información en los haplotipos que están presentes en la familia o la población a la que el individo pertenece.

[0100] Dependiendo de la repetibilidad de los factores de corrección, los pasos 1, 2 y 3 pueden ser completamente saltados en un análisis nuevo donde condiciones de ensayo se conocen por ser las mismas.

[0101] Cuando se sigue el método de ejemplo 1, ahorros significativos se pueden obtener reduciendo el número total de muestras que necesitan ser analizadas mientras todavía se obtiene resultados fiables en las muestras individuales originales. Reducciones típicas de los números totales de muestras que deben ser analizadas se ejemplifican en la tabla 3.

**Tabla 3**. Ahorros en el número de muestras que deben ser analizadas cuando se agrupan 2 o 3 individuos siguiendo del método de la invención.

| Número de individuos que debe ser genotipificado | Número de muestras cuando se agrupan 2 individuos | | | | Número de muestras cuando se agrupan 3 individuos | | | |
|---|---|---|---|---|---|---|---|---|
| | Número de individuos más agrupamiento | Número de agrupamientos de 2 individuos | Número total de muestras | Reducción del número de muestras que se debe analizar ( %) | Número de individuos más agrupamiento | Número de agrupamientos de 2 individuos | Número total de muestras | Reducción del número de muestras que se debe analizar ( %) |
| 250 | 50+1 | 100 | 151 | 39.6 | 50+1 | 67 | 118 | 52.8 |
| 500 | 50+1 | 225 | 276 | 44.8 | 50+1 | 1.50 | 201 | 59.8 |
| 1000 | 50+1 | 475 | 526 | 47.4 | 50+1 | 317 | 368 | 63.2 |
| 2000 | 50+1 | 975 | 1026 | 48.7 | 50+1 | 650 | 701 | 64.9 |
| 5000 | 50+1 | 2475 | 2526 | 49.5 | 50+1 | 1650 | 1701 | 66.0 |

**Ejemplo 2**

*Ejemplo de genotipificación de m uestras de individuos dipl oides usando 25 agrupam ientos de 2 individuos para estandarización*

[0102] Paso 1) 50 individuos se evalúan separadamente como en el paso 1; ejemplo 1.

[0103] Paso 2) Construir 25 agrupamientos de 2 muestras cada uno en la proporción 1:3 incluyendo los 50 individuos del paso 1 anterior. En estos agrupamientos estimar frecuencias de alelo bien sin corregir o basados en los factores de corrección encontrados en el primer paso.

[0104] Paso 3) comparar la suma de las frecuencias de alelo de las 2 tipificaciones individuales y la frecuencia estimada en los depósitos de 2 muestras individuales. De estos 25 puntos calcular una línea de regresión. El coeficiente de regresión e intersección puede después usarse para corregir las frecuencias estimadas de otros agrupamientos.

[0105] Paso 4) luego construir agrupamientos de ADN de 2,3 o n individuos en la proporción 1: 3,1: 3: 9 o 1:.$3^1$: $3^2$: $3^{(n-1)}$.

[0106] Paso 5) con los factores de corrección encontrados en el paso 1 y paso 3 calcular las frecuencias de alelo de las intensidades de señal resultantes en el agrupamiento.

[0107] Las ahorros en los números de muestra son idénticos a las ahorros en la tabla mencionadas 8 para secuenciación de individuos diploides.

**Ejemplo 3**

*Ejemplo de genotipificación de muestras individuales haploides.*

[0108] Cuando dos muestras haploides son agrupadas y medidas en cuanto a la presencia de alelo A en una posición determinada en el genoma, las proporciones previstas en las mediciones (altura de pico, área bajo el pico, intensidades) son;

**Tabla 4**. Puntos de resultado de frecuencias de alelo en muestras agrupadas y genotipos inferidos de los dos individuos en el agrupamiento para un SNP con alelo A y C

| Frecuencia de alelo A en muestra agrupada | Genotipo inferido de individuo 1 (presente en agrupamiento en 1 parte) | Genotipo inferido de individuo 2 (presente en agrupamiento en 3 partes) |
|---|---|---|
| 0.00 | C | C |
| 0.33 | A | C |
| 0.67 | C | A |
| 1.00 | A | A |

[0109] Si sólo se utilizan agrupamientos de dos muestras puede que no se necesiten factores de corrección. Cuando más muestras son agrupadas factores de corrección probablemente son necesitados. Ellos luego se pueden calcular de agrupamientos de 2 muestras con cantidades iguales del analito para simular individuos diploides heterocigóticos y homocigóticos.

[0110] En el agrupamiento 3 muestras se agrupan en una proporción de 1:2:4, las siguientes proporciones en las mediciones son previstas;

**Tabla 5**. Puntos de resultado de frecuencias de alelo en muestras agrupadas y genotipos inferidos de los tres individuos en el agrupamiento para un SNP con alelo A y C

| Frecuencia de alelo A en muestra agrupada | Genotipo inferido de individuo 1 (presente en agrupamiento en 1 parte) | Genotipo inferido de individuo 2 (presente en agrupamiento en 2 partes) | Genotipo inferido de individuo 2 (presente en agrupamiento en 4 partes) |
|---|---|---|---|
| 0.000 | C | C | C |
| 0.166 | A | C | C |
| 0.333 | C | A | C |
| 0.500 | C | C | A |
| 0.666 | A | C | A |
| 0.833 | C | A | A |
| 1.000 | A | A | A |

**Ejemplo 4**

*Uso de la invención en los protocolos de secuenciación*

[0111] El método de agrupamiento descrito en esta invención se puede aplicar a situaciones donde hay una necesidad de determinar secuencias en 2 o más individuos.

[0112] El agrupamiento de individuos, modelos o productos PCR para secuenciación no es práctica común debido a que el problema esencial cuando se analiza un rastro doble es que dos bases se representan a cada posición y es imposible decir de qué modelo vino cada base ejemplificando solo el rastro.

[0113] Además de modelos deliberadamente agrupados que tienen como resultado trazos dobles, diferentes situaciones biotecnológicas y biológicas se conoce que dan lugar a trazos dobles. Estos se ven en regiones alternativas unidas de un transcrito que son amplificadas por RT-PCR, secuenciadas directas (sin clonación) y experimentos de mutagénesis aleatoria insertional.

[0114] Diferentes métodos han sido descritos para rastrear los haplotipos de secuencias agrupadas o trazas dobles. Flot et al. 2006 describen diferentes métodos moleculares que han sido propuestos para averiguar los haplotipos de un individuo. Por ejemplo secuenciación de productos PCR clonados (p. ej. Muir et al., 2001), SSCP (polimorfismo de conformación monocatenaria).(Sunnucks et al., 2000), electroforesis en gel desnaturalizante con gradiente (DGGE) (Knapp 2005), dilución extrema de ADN a nivel de molécula única (Ding & chantre 2003) y el uso de cebadores de la PCR específicos de alelo (Pettersson et al., 2003). Además diferentes métodos computacionales han sido propuestos para reconstrucción de haplotipo de mezclas de secuencias.

[0115] Todos los métodos descritos, no obstante, pueden ser muy costosos y llevar mucho tiempo y son solo aplicables a fines específicos (p. ej. resecuenciación, unión alternativa, modelos o mezclas PCR amplificadas de dos productos que difieren en la longitud de secuencia, la disponibilidad de una secuencia de genoma de referencia) y no para secuenciación directa estándar de muestras diploides o haploides o de nuevo secuenciación de secuencias completamente desconocidas.

[0116] El agrupamiento de modelos de secuencia siguiendo el agrupamiento descrito en esta invención se puede aplicar a situaciones donde el mismo fragmento de secuencia se puede obtener tanto en individuos como en muestras agrupadas. Esto significa que por ejemplo la secuenciación de escopeta (fragmentos aleatorios cortados) no es adecuada para agrupamiento.

[0117] En todas las aplicaciones mencionadas anteriormente, si el agrupamiento es aplicado ex profeso, cantidades iguales de modelo (muestras, ADN, ARN o producto PCR) son agrupadas.

[0118] Aquí nosotros describimos el agrupamiento de cantidades desiguales de modelo. Para este ejemplo solo la situación para una agrupación que consiste en 2 modelos es descrita, pero la invención puede utilizarse para depósitos de constructos de ADN (o productos de post-PCR) de 2, 3, o n individuos en la proporción $1:3,1:3:9,1:3^1:3^2:3^{(n-1)}$ para organismos diploides y en la proporción de $1:2,1:2:4,1: 2^1:2^2:2^{(n-1)}$ para organismos haploides.

[0119] Condiciones generales que se necesitan cumplir son que el dispositivo de secuenciación sondee modelos (p. ej. en cuanto a fluorescencia) y el cromatograma resultante represente la secuencia del modelo de ADN como una cuerda de picos que están regularmente distanciados y de altura similar.

Paso 1) llevar a cabo reacciones de secuencia para 50 individuos separadamente

[0120] Los datos en las reacciones de secuenciación individuales se utilizan para calcular los factores de corrección de las áreas de pico o alturas de pico para todas las posiciones de base (o nucleótido).

Paso 2) llevar a cabo reacciones de secuencia para 25 agrupamientos de 2 individuos agrupados

[0121] Proporciones de área de pico se utilizan para discriminar entre el primer y segundo pico en la base y picos de ruido. El segundo pico es un porcentaje del primer pico y un valor de umbral se utiliza para discriminar entre picos y picos de ruido. Los datos en las reacciones de secuenciación agrupada se utilizan para calcular los factores de corrección de las áreas de pico o alturas de pico para todas las posiciones de base (o nucleótido).

Paso 3) hacer un gráfico de los resultados de los pasos 1 y 2 y construir la línea de regresión (calcular coeficiente de regresión e intersección).

Paso 4) Construir agrupamientos de ADN (o productos de post-PCR)

[0122] Los agrupamientos se construyen de 2, 3, o n individuos en la proporción $1:3,1:3:9,1:3^1:3^2:3^{(n-1)}$ para organismos diploides y en la proporción de $1:2,1:2:4,1:2^1:2^2:2^{(n-1)}$ para organismos haploides.

Paso 5) con los factores de corrección encontrados en el paso 1,2 y paso 3, el *base-calling* se puede calcular de las intensidades de señal resultantes en el agrupamiento

[0123] En este ejemplo solo 2 nucleótidos potenciales (A y C) en cada posición de base, son mostrados pero el mismo principio funciona para otras combinaciones de 2 de los 4 nucleótidos disponibles que son la base del código genético. La altura de pico media para el nucleótido "A" se fija a 100 y la altura de pico media del nucleótido "C" es 75. Basado en estas alturas de pico, para cada combinación posible de nucleótidos en el agrupamiento de dos muestras haploides las alturas de pico relativas se presentan en la tabla 6. Las alturas de pico relativas para un agrupamiento que consiste en dos modelos diploides se dan en la tabla 7.

**Tabla 6.** Puntos de resultado de frecuencias de alelo en individuos haploides agrupados y no agrupados y genotipo inferido para una posición aleatoria en la secuencia de nucleótidos.

| Genotipo inferido | | Área/altura de pico sin agrupar | | Área/altura de pico agrupado (proporción 1:2) | |
|---|---|---|---|---|---|
| Individuo 1 | Individuo 2 | Primer pico (A) | Segundo pico (C) | Primer pico (A) | Segundo pico (C) |
| A | | 100 | | | |
| C | | | 75 | | |
| | | | | | |
| A | A | | | 100 | |
| A | C | | | 33.3 | 50 |
| C | A | | | 66.6 | 25 |
| C | C | | | | 100 |

**Tabla 7**. Puntos de resultado de frecuencias de alelo en individuos diploides agrupados y no agrupados y genotipo inferido para una posición aleatoria en la secuencia de nucleótidos.

| Genotipo inferido | | Área/altura de pico sin agrupar | | Área/altura de pico agrupado (proporción 1:3) | |
|---|---|---|---|---|---|
| Individuo 1 | Individuo 2 | Primer pico (A) | Segundo pico (C) | Primer pico (A) | Segundo pico (C) |
| AA | | 100 | | | |
| AC | | 50 | 37.5 | | |
| CC | | | 75 | | |
| | | | | | |
| AA | AA | | | 100 | 0 |
| AA | AC | | | 62.5 | 28.125 |
| AA | CC | | | 25 | 56.25 |
| AC | AA | | | 87.5 | 9.375 |
| AC | AC | | | 50 | 37.5 |
| AC | CC | | | 12.5 | 65.625 |
| CC | AA | | | 75 | 18.75 |
| CC | AC | | | 37.5 | 46.875 |
| CC | CC | | | 0 | 100 |

[0124] La tabla 8 indica la reducción del número de reacciones de secuencia comparando la estrategia de agrupamiento de esta invención y la situación sin agrupamiento.

Tabla 8. Ahorro en el número de muestras o reacciones de secuencia cuando se agrupan 2 individuos según el método de la invención.

| Número de invidiuos que se deben secuenciar | Número de agrupamientos o muestras que se deben secuenciar utilizando esta invención | | | Reducción del número de muestras que se deben secuenciar ( %) |
|---|---|---|---|---|
| | Individuos + agrupamientos | Agrupamientos de dos individuos | Número total de muestras | |
| 250 | 50+25 | 100 | 175 | 30 % |
| 500 | 50+25 | 225 | 300 | 40 % |
| 1000 | 50+25 | 475 | 550 | 45 % |
| 2000 | 50+25 | 975 | 1050 | 47,5 % |
| 5000 | 50+25 | 2475 | 2250 | 49 % |

**Ejemplo 5**

*Ejemplo de genotipificación de muestras de individuos diploides utilizando 1 agrupamiento de 50 individuos y 25 agrupamientos de 2 individuos para estandarización utilizando métodos de corrección alternativos. El ejemplo describe diferentes experimentos.*

[0125] Paso 1) 50 individuos fueron evaluados separadamente.

[0126] Igual que en el ejemplo 1, paso 1 pero con método(s) de corrección diferente(s) utilizando las intensidades normalizadas X e Y en lugar de Xraw e Yraw.

[0127] El primer factor de corrección (K) es calculado utilizando X e Y.

$$K = avg (X/Y)$$

donde X es la intensidad normalizada para el alelo A (rojo) e Y es la intensidad normalizada para el alelo B (verde). Este valor fue determinado a partir de las muestras genotipadas individualmente con genotipo AB.

[0128] Los otros factores de corrección **AAavg** y **BBavg** están también basados en X e Y. AAavg es el promedio de las frecuencias de alelo sin corregir de genotipos AA. Este valor está previsto que sea próximo a 1. BBavg es el promedio de las frecuencias de alelo sin corregir de genotipos BB. Este valor está previsto que sea próximo a 0. **AAavg** y **BBavg** fueron calculados utilizando las fórmulas:

$$AAavg= (avg (X/(X+Y)))$$

Y

$$BBavg= (avg (X/(X+Y)))$$

[0129] Todos los factores de corrección K, AAavg y BBavg pueden también ser calculados en base a Xraw e Yraw como en el ejemplo 1, paso 1.

[0130] Si ningún genotipo AA está disponible entre los 50 individuos AAavg se fija a 1.También si ningún genotipo BB está disponible entonces BBavg se fija a 0.

[0131] El paso siguiente es calcular frecuencias de alelo basadas en las tipificaciones individuales para aquellos SNP donde los 50 individuos tienen un resultado.

[0132] Paso 2) un agrupamiento fue construido incluyendo los 50 individuos del paso 1 como en el ejemplo 1, paso 2.

[0133] Frecuencia de alelo sin corregir para el alelo A se calcula como una proporción entre la intensidad roja normalizada (X) dividida por la suma de ambas intensidades normalizadas (X+Y)

**Frecuencia de alelo sin corregir = X/(X+Y) (llamada Raf)**

[0134] La primera corrección para frecuencia de alelo que nosotros aplicamos es

**Frecuencia de alelo corregida = X/(X+K*Y) (llamada Rafk)**

[0135] Si no hubiera ningún genotipo heterocigótico, K puede no ser calculado. En este caso las reglas siguientes pueden ser aplicadas;

[0136] Si **Raf**<0,1 luego **Rafk** se fija a 0.
Si **Raf**>0.9 luego **Rafk** se fija a 1.
En todas las otras situaciones donde K falta **Rafk** se fija igual a **Raf**.

[0137] La corrección de normalización utilizando **AAavg** y **BBavg** no es siempre necesaria cuando se empieza con las intensidades normalizadas X e Y. Si usted empieza con Xraw e Yraw la normalización utilizando AAavg y BBavg se puede aplicar como en el ejemplo 1, paso 2.

[0138] Si se aplica la normalización entonces utilice la siguiente fórmula;

**Frecuencia de alelo normalizada = (Frecuencia de alelo corregida – BBavg)/ AAavg (llamada Rafn)**

[0139] Paso 3) nosotros comparamos las frecuencias de alelo previstas calculadas en tipificaciones individuales en el paso 1 y las frecuencias observadas (corregidas o sin corregir) basadas en los resultados de el agrupamiento de 50 en el paso 2. De esto nosotros calculamos los coeficientes de regresión utilizando el modelo siguiente;

**Frecuencia de alelo prevista=b1*frecuencia observada + b2*frecuencia observada$^2$ + b3*frecuencia observada$^3$ + b4*frecuencia observada$^4$ sin intersección**

[0140] Tanto las frecuencias corregidas (**Rafk** y **Rafn**) como sin corregir (**Raf**) se usan como frecuencia observada en la fórmula anterior.
Por comparación de la frecuencia de alelo esperada con la predicha del modelo el mejor procedimiento de corrección (**Rafk**, **Rafn** o **Raf**) puede ser encontrado.
Los coeficientes de regresión del mejor procedimiento de corrección pueden más tarde usarse para corregir las frecuencias de alelo de los depósitos de 2 individuos en el paso 5a.

[0141] Paso 4) de las 50 muestras individuales se construyen 25 agrupamientos de ADN de 2 individuos en proporción 1: 3. Se anota qué individuo se usa una vez y cual se usa 3 veces en el agrupamiento.

[0142] Paso 5a) Corrección basada en resultados de agrupación de 50 individuos. Con los factores de corrección encontrados en el paso 1 (K, AAavg y BBavg) y paso 3 (factores de regresión b1, b2, b3 y b4) las frecuencias de alelo se pueden calcular a partir de las intensidades de señal resultantes en los agrupamientos, construidos en el paso 4. Primero se calcula Raf o Rafk o Rafn (dependiendo del mejor procedimiento de corrección encontrado en el paso 3) utilizando factores de corrección K, AAavg y BBavg de paso 1.

[0143] Luego Rafc o Rafkc o Rafnc es calculado utilizando los coeficientes de regresión polinomial encontrados en el paso 3 como

**Frecuencia de alelo prevista = b1*frecuencia observada + b2*frecuencia observada$^2$ + b3*frecuencia observada$^3$ + b4*frecuencia observada$^4$ donde frecuencia observada =Raf o Rafk o Rafn**

[0144] Con dos individuos en un agrupamiento las frecuencias corregidas predichas deberían dar los puntos de resultado 0 %, 12. 5 %, 25. 0 %, 37. 5 %, 50. 0 %, 62. 5 %, 75. 0 %, 87. 5 % y 100 %. Se debería redondear al punto de resultado más próximo. Los genotipos de los dos individuos se pueden derivar de los resultados como se indica en la tabla 2 de ejemplo 1.

[0145] Paso 5b) Corrección basada en resultados de agrupaciones de 2 individuos.
**Raf**, **Rafk** y **Rafn** son calculados en base a las intensidades de señal de los agrupamientos construidos en el paso 4 y los

factores de corrección K, AAavg y BBavg encontrados en el paso 1.

[0146] Luego se pueden calcular coeficientes de regresión de polinomiales utilizando el mismo modelo que en el paso 3, ejemplo 5 en base a 20 agrupamientos. Este modelo se puede aplicar en cada SNP separadamente o a través de todos SNP. Las frecuencias de alelo en los otros 5 depósitos se predicen en base a estos factores de regresión como:

**$Rafkc = b1*Rafk + b2*Rafk^2 + b3*Rafk^3 + b4*Rafk^4$ del modelo de regresión con Rafk**

**$Rafn = b1*Rafn + b2*Rafn^2 + b3*Rafn^3 + b4*Rafn^4$ del modelo de regresión con Rafn**

**$Rafc = b1*Raf + b2*Raf^2 + b3*Raf^3 + b4*Raf^4$ del modelo de regresión con Raf**

[0147] Esto se puede repetir 5 veces de manera que todas las muestras se usan para predicción una vez. Las frecuencias de alelo previstas en estos depósitos luego se comparan con las frecuencias de alelo predichas para encontrar el mejor procedimiento de corrección.

[0148] Con dos individuos en un agrupamiento las frecuencias corregidas predichas deberían dar los puntos de resultado 0 %, 12. 5 %, 25. 0 %, 37. 5 %, 50. 0 %, 62. 5 %, 75. 0 %, 87. 5 % y 100 %. Se debería redondear al punto de resultado más próximo. Los genotipos de los dos individuos se pueden derivar de los resultados como se indica en la tabla 2 ejemplo 1.

[0149] Paso 5c) Corrección basada en resultados de agrupamientos de 2 individuos. Otra vía de predicción puede hacerse utilizando coeficientes de regresión lineal múltiple por SNP en las intensidades de luz (X o Xraw e Y e Yraw) en base al modelo siguiente

**Frecuencia de alelo prevista = b1*X + B2*Y**

O

**Frecuencia de alelo prevista = b1*Xraw + b2*Yraw**

[0150] Con estos coeficientes de regresión lineal múltiple las frecuencias de alelo pueden después ser predichas usando

**Frecuencia de alelo predicha = intersección + b1*X + b2*Y**

O

**Frecuencia de alelo predicha = intersección + b1*Xraw + b2*Yraw**

[0151] Los coeficientes de regresión lineal múltiple, como se describen anteriormente, son calculados en base a 20 agrupamientos. Luego las frecuencias de alelo de los otros 5 agrupamientos son predichas en base a estos factores de regresión. Esto se repite 5 veces de manera que todas las muestras se usan para predicción una vez. Las frecuencias de alelo previstas en estos agrupamientos luego se pueden comparar con las frecuencias de alelo predichas para encontrar el mejor procedimiento de corrección.

[0152] Como en el paso 5a y paso 5b los genotipos de los dos individuos se pueden derivar de los resultados como se indican en la tabla 2 de ejemplo 1.

[0153] Paso 6) de otras muestras individuales construir agrupamientos de ADN de 2 individuos en la proporción 1: 3. Se debe anotar qué individuo se usa una vez y cual se usa 3 veces en el agrupamiento como en el paso 4. De estos depósitos nosotros podemos obtener los genotipos utilizando el mejor método de corrección para predicción de la frecuencia de alelo como se describe y usando la tabla 2 de ejemplo 1.

**- Experimento 1**

Aplicación de procedimientos descritos en el ejemplo 5 para análisis SNP de genoma completo usando tecnología Infinium Assay BeadChip (Illumina, Inc. EE. UU.)

[0154] La genotipificación fue hecha en 50 individuos utilizando el ensayo 18K Chicken SNP iSelect Infinium (Illumina Inc, EE.UU.), con SNP uniformemente distribuidos en todo el genoma de pollo (van As et al., 2007). Detalles en el ensayo, flujo de trabajo y chip se pueden encontrar en el sitio web de Illumina (http://www.illumina.com/pages.ilmn?ID=12).

[0155] Para controlar si las frecuencias se pueden estimar con precisión, 8 alelos (de 4 animales diferentes de los 50 individuos genotipados individualmente) fueron combinados en un agrupamiento. Pasos 1 a 3 y paso 5, como se describen en el **ejemplo 5**, fueron tomados excepto la traducción de las frecuencias de alelo predichas en genotipos, utilizando la tabla 2, no fue realizado. En el paso 4 cantidades equimolares de ADN de 4 individuos fueron agrupadas en lugar de ADN de 2 individuos en la proporción 1:3. Si una proporción 1:3 de 2 animales diferentes se usa nosotros podemos considerar que esto es combinar 8 alelos en una agrupación. Usando cantidades equimolares de 4 individuos también 8 alelos son combinados.

[0156] De esta manera 12 agrupamientos fueron compuestos y un agrupamiento de 50 animales como en el paso 1 (se usan las mismas muestras que en los agrupamientos de 4 más las 2 muestras extra). Luego estos 13 depósitos fueron genotipados utilizando un segundo lote de chips Infinium.

[0157] K, AAavg y BBavg por SNP fueron calculados como en el ejemplo 5, paso 1. Entonces, frecuencias de alelo sin corregir y corregidas de el agrupamiento de 50 fueron calculadas como en el ejemplo 5, paso 2. También coeficientes de regresión polinomiales fueron calculados como en el ejemplo 5, paso 3. Además el polinomio y coeficientes de regresión lineal múltiple, como se describe en el paso 5b y 5c, fueron calculados. Esto fue hecho en base a 11 agrupamientos y luego las frecuencias de alelo en el agrupamiento restante fueron predichas utilizando los factores de regresión.

[0158] En este experimento la regresión lineal múltiple en X e Y (intensidades para rojo y verde) dio los mejores resultados. Para resultados finales véase figura 4 y tabla 9.

[0159] En total 4,6 % de las frecuencias de alelo cayeron en la clase incorrecta.
En caso de que estos fueran agrupaciones de 2 individuos en una proporción de 1:3 esto tendría como resultado 3.0 % de errores de genotipificación.

**Tabla 9**. Número de frecuencias de alelo predichas por clase en comparación con las frecuencias de alelo previstas. Los números de la diagonal llevarán a genotipos correctos. Las frecuencias de alelo fuera de la diagonal pero entre cajas conllevarán un error de genotipo. Los otros resultados conllevarán 2 errores de genotipo.

| Frecuencia de alelo prevista | Predicha 0 | 12.5 | 25 | 37.5 | 50 | 62.5 | 75 | 87.5 | 100 | Total |
|---|---|---|---|---|---|---|---|---|---|---|
| 0 | 59489 | 144 | 13 | | | 2 | | 1 | | 59649 |
| 12.5 | 331 | 12888 | 452 | 11 | 3 | 1 | 1 | | | 13687 |
| 25 | 27 | 427 | 12060 | 897 | 10 | 1 | | | | 13422 |
| 37.5 | 2 | | 374 | 11342 | 1026 | 17 | 1 | | | 12762 |
| 50 | | | 4 | 671 | 11590 | 1098 | 27 | | | 13390 |
| 62.5 | 1 | | | 5 | 682 | 11074 | 727 | | 1 | 12490 |
| 75 | | | 1 | | 3 | 779 | 11421 | 494 | 29 | 12727 |
| 87.5 | | | 1 | | 1 | 3 | 528 | 11172 | 416 | 12121 |
| 100 | 10 | | | 3 | 1 | 6 | 5 | 50 | 50896 | 50971 |

**- experimento 2**

**Aplicación de procedim ientos descrita en el ejem plo 5 para análisis SN P utilizando tecnología Veracode Assay (Illumina, Inc. EE.UU.).**

[0160] La genotipificación fue hecha en 50 individuos utilizando el ensayo 96 Chicken SNP Veracode, Golden Gate Assay (Illumina Inc, EEUU), con SNP uniformemente distribuidos en todo el genoma de pollo (paso 1). Detalles en el ensayo, flujo de trabajo y chip se pueden encontrar en el sitio web de Illumina (http://www.illumina.com/pages.ilmn?ID=6)
También 1 agrupamiento de todas las muestras fue construido (como en el paso 2) y 24 depósitos de 2 individuos en la proporción 1:3 (como en el paso 4). Estos 25 depósitos fueron genotipados con un segundo lote de productos químicos. Todas las correcciones fueron hechas como se describe en el paso 1 a 3 de ejemplo 5. La corrección en el paso 5a fue aplicada en los 24 depósitos de 2 usando los factores de regresión polinomiales encontrados en el paso 3.

[0161] Para paso 5b y paso 5c nosotros usamos 23 depósitos cada vez para calcular los factores de regresión (polinomiales en el paso 5b y lineales múltiples en el paso 5c) para ser capaces de predecir las frecuencias de alelo para el agrupamiento restante. En total nosotros hicimos esto 24 veces así todos depósitos fueron usados una vez para predecir las frecuencias de alelo. Los mejores resultados fueron obtenidos utilizando **Rafk (calculado en base a valores normalizados X e Y)** y luego corregidos utilizando los factores de regresión polinomial del paso 5b dando como resultado **Rafkc**.

[0162] En total 84 SNP fueron seleccionados en los individuos. Luego algunos SNP no fueron seleccionados en alguno de los individuos. En total nosotros teníamos 1,906 combinaciones completas de agrupación*SNP.

Tabla 10. Número de frecuencias de alelo predichas por clase en comparación con las frecuencias de alelo previstas. Los números en la diagonal llevarán a genotipos correctos. Las frecuencias de alelos fuera de la diagonal pero dentro de las cajas resultarán en un error de genotipo. Los otros resultados resultarán en 2 errores de genotipo.

| Genotipos Previstos | Predichos CC CC | AC CC | AA C | CC C | AC C | AA AC | CC AA | AC AA | AA AA | Total |
|---|---|---|---|---|---|---|---|---|---|---|
| CC CC | 312 | 9 | | | | | | | | 321 |
| AC CC | 4 | 156 | 4 | 2 | | | | | | 166 |
| AA CC | | 13 | 39 | 7 | 3 | | | | | 62 |
| CC AC | | | 10 | 129 | 7 | 1 | | | | 147 |
| AC AC | | | | 9 | 228 | 12 | 1 | | | 250 |
| AA AC | | | | | 24 | 144 | 5 | | | 173 |
| CC AA | | | | | | 4 | 49 | 9 | | 62 |
| AC AA | | | | | | | 7 | 135 | 1 | 143 |
| AA AA | | | | | | | 1 | 5 | 576 | 582 |
| Total | 316 | 176 | 54 | 147 | 265 | 159 | 64 | 148 | 577 | 1906 |

[0163] En total había 138 (138/1906*100=7,2 %) desapareamientos (tabla 10). Dado que cada observación consiste en 2 muestras individuales esta resultó en 174 errores de genotipo (7,70/1906*2*100=4,46 %), ver tabla 11, La Figura 5 y figura 6.

[0164] El proceso de definir el mejor procedimiento de corrección en este ejemplo (como se hace utilizando el paso 3 (ejemplo 5) y paso 5a, 5b o 5c (ejemplo 5)) también entrega información acerca del número de desapareamientos por SNP. Esto hace posible eliminar un SNP del conjunto para reducir el riesgo de equivocaciones a costa de índices de llamada inferior.

**Tabla 11**. Número de genotipos predichos correctamente

| | Predicho | | | | | | | | | |
|---|---|---|---|---|---|---|---|---|---|---|
| Previsto | CC CC | AC CC | AA CC | CC AC | AC AC | AA AC | CC AA | AC AA | AA AA | total |
| CC CC | 624 | 9 | | | | | | | | 633 |
| AC CC | 4 | 312 | 4 | 0 | | | | | | 320 |
| AA CC | | 13 | 78 | 0 | 0 | | | | | 91 |
| CC AC | | | 0 | 258 | 7 | 1 | | | | 266 |
| AC AC | | | | 8 | 456 | 12 | 0 | | | 477 |
| AA AC | | | | | 24 | 288 | 0 | | | 312 |
| CC AA | | | | | | 0 | 98 | 9 | | 107 |
| AC AA | | | | | | | 7 | 270 | 1 | 278 |
| AA AA | | | | | | | 1 | 5 | 1152 | 1158 |
| Total | 628 | 331 | 83 | 266 | 491 | 297 | 107 | 282 | 1153 | 3642 |

- experimento 3

[0165] Aplicación de procedimientos descrita en el ejemplo 5 para análisis SNP usando otros métodos de genotipificación.

[0166] Los procedimientos descritos en el ejemplo 5 también pueden usarse en cualquier otro método de genotipificación, además de los métodos descritos en el experimento 1 y experimento 2, tal como Affimetrix GeneChip (Affimetrix Inc, EE. UU.) o Tecnologías Agilent.

Ejemplo 6.

*Uso de la invención en los protocolos de secuenciación como en ejemplos 4 pero utilizando otros métodos de corrección*

[0167] Paso 1) Llevar a cabo reacciones de secuencia para 50 individuos separadamente Usar la altura de pico de alelo 1 y altura de pico de alelo 2 como el valor Xraw e Yraw o la altura de pico relativa como X e Y. Altura de pico relativa para alelo 1 es X=X/(X+Y) y altura de pico relativa para alelo 2 es Y=Y(X+Y).
Luego calcular K, AAavg y BBavg del mismo modo que en la genotipificación en el paso 1 del ejemplo 5;

[0168] Paso 2) Llevar a cabo reacciones de secuencia en un agrupamiento de los 50 individuos. Calcular frecuencias de alelo sin corregir y corregidas como en el paso 2 de ejemplo 5;

[0169] Paso 3) calcular frecuencias a partir de secuenciación individual y del agrupamiento. Usar el mismo modelo que en el paso 3 de ejemplo 5 para encontrar los coeficientes de regresión polinomial.

[0170] Paso 4) llevar a cabo reacciones de secuencia para 25 agrupamientos de 2 individuos agrupados

[0171] Paso 5a) Comparar las frecuencias corregidas con frecuencias previstas en base al agrupamiento de los 50 individuos para encontrar el mejor método.

[0172] Paso 5b) calcular Rafnc, Rafkc y Rafc en 5 agrupamientos de 2 individuos utilizando los factores de regresión polinomial encontrados en los otros 20 agrupamientos que utilizan el modelo

**Frecuencia de alelo prevista = b1*frecu encia observada + b2*frecuencia observada $^2$ + b3*frecuencia observada $^3$ + b4*frecuencia observada$^4$ sin intersección**

[0173] Paso 5c) calcular la frecuencia de alelo predicha en 5 agrupamientos de 2 individuos utilizando los coeficientes de regresión lineal múltiple encontrados en los otros 20 depósitos que utilizan el modelo

**Frecuencia de alelo predicha = intersección + b1*X + b2*Y**

O

**Frecuencia de alelo predicha = intersección +b1*Xraw + b2*Yraw**

[0174] De paso 3 y paso 5 determinar el mejor procedimiento de corrección repitiendo el paso 5b y 5c varias veces de manera que todos los agrupamientos se usen para predicción de frecuencias de alelo (validación). Si se necesita otros números para validación pueden ser usados. Por ejemplo uno puede usar 24 agrupamientos para encontrar los factores de regresión y luego predecir uno usando estos factores. En total uno luego necesita repetir esto 25 veces.

[0175] Con el mejor procedimiento de corrección y los factores de corrección y factores de regresión necesitados era posible predecir frecuencias de nuevos agrupamientos y leer los alelos resultantes en la tabla 2.

LEYENDAS DE LAS FIGURAS

[0176]

La Figura 1 muestra en una representación gráfica la correlación entre la frecuencia de alelo en base a datos agrupados (eje y) y la frecuencia de alelo en base a mediciones individuales (eje x).

La Figura 2 muestra en una representación gráfica la relación entre frecuencia de alelo medida en individuos (eje y) y las frecuencias de alelo predichas en el agrupamiento (eje x).

La Figura 3 muestra en una representación gráfica la relación entre la frecuencia de alelo corregida en el agrupamiento (eje y) y las frecuencias de alelo medidas en individuos después de la tipificación individual (eje x).

La Figura 4 muestra en una representación gráfica la diferencia entre las frecuencias de alelo previstas (basadas en tipificaciones individuales) y predichas para el agrupamiento 1 en el experimento 1.

La Figura 5 muestra en una respresentación gráfica la correlación entre las frecuencias de alelo previstas (basadas en tipificaciones individuales) y predichas para todos los agrupamientos en el experimento 2.

La Figura 6 muestra en una representación gráfica la diferencia entre las frecuencias de alelo previstas (basadas en tipificaciones individuales) y predichas para todos los agrupamientos en el experimento 2.

**REIVINDICACIONES**

1. Método de agrupamiento de muestras que deben ser analizadas en cuanto a una variable categórica, donde el análisis implica una medición cuantitativa de un analito, dicho método de agrupamiento de muestras comprende proporcionar un agrupamiento de n muestras donde la cantidad de muestras individuales en el agrupamiento es de manera que los analitos en las muestras están presentes en una proporción molar de $x°: x^1: x^2: x^{(n-1)}$, y donde x es un número entero de 2 o más alto representando el número de clases de la variable categórica.

2. Método según la reivindicación 1, donde el analito es una biomolécula y la variable categórica es una variante de dicha biomolécula.

3. Método según la reivindicación 2, donde la biomolécula es un ácido nucleico.

4. Método según la reivindicación 3, donde la variante es un polimorfismo de nucleótido en dicho ácido nucleico.

5. Método según la reivindicación 4, donde el polimorfismo de nucleótido es un SNP.

6. Método según la reivindicación 3, donde la variante es la identidad de base de una posición particular de nucleótidos.

7. Método según cualquiera de las reivindicaciones precedentes, donde la medición cuantitativa comprende la medición de la intensidad, altura de pico o superficie de pico de una señal de instrumento.

8. Método según la reivindicación 7, donde la señal de instrumento es una señal de fluorescencia.

9. Uso de un método según cualquiera de las reivindicaciones 1-8, para genotipificación de una variante alélica en individuos poliploides o haploides donde el número de clases del variable categórica (x) es igual a p+1, donde p representa el nivel de ploidía.

10. Uso según la reivindicación 9, donde x es 3, para genotipificación de una variante alélica en individuos diploides.

11. Método de realizar un análisis en muestras múltiples, que comprende el agrupamiento de dichas muestras según un método de cualquiera de las reivindicaciones 1-8 para proporcionar una muestra agrupada y la realización de dicho análisis en dicha muestra agrupada.

12. Método de realizar un análisis en muestras múltiples según la reivindicación 11, donde dicha muestra se analiza en cuanto a una variable categórica e implica una medición cuantitativa de un analito en dicha muestra.

13. Método según la reivindicación 12, que comprende además deducir de la medición la aportación de las muestras individuales en dicho agrupamiento de muestras.

14. Dispositivo de agrupamiento para agrupamiento de muestras múltiples en una muestra agrupada que comprende un colector de muestra para suministrar una muestra agrupada y además comprende un procesador que está dispuesto para recibir instrucciones de un programa de ordenador que aplica en el dispositivo de agrupamiento un método de agrupamiento de muestras según cualquiera de las reivindicaciones 1-8.

15. Dispositivo según la reivindicación 14, que incluye además un dispositivo de análisis que comprende un procesador que está dispuesto para realizar un análisis en la muestra agrupada proporcionado por el dispositivo de agrupamiento, donde dicho dispositivo de análisis está dispuesto para analizar dicha muestra en cuanto a una variable categórica y para realizar una medición cuantitativa de un analito en dicha muestra.

16. Producto de programa informático en un portador, el cual producto de programa, cuando se carga y ejecuta en un ordenador, una red informática programada u otro equipo programable, aplica un método de agrupamiento de muestras según cualquiera de las reivindicaciones 1-8.

17. Producto de programa informático según la reivindicación 16, donde el método comprende además el paso de realizar un análisis en muestras múltiples, dicho método comprende la realización de un análisis en la muestra agrupada, donde dicha muestra se analiza en cuanto a una variable categórica e implica una medición cuantitativa de un analito en dicha muestra.

**Figura 1**

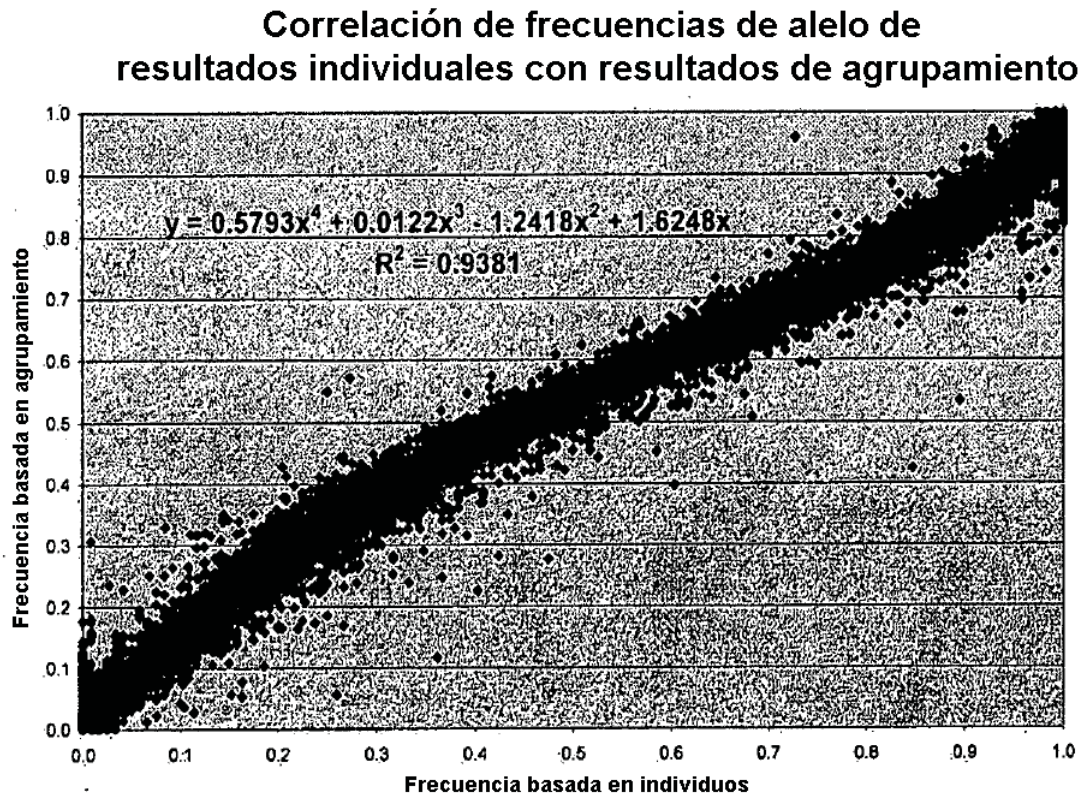

Correlación de frecuencias de alelo de resultados individuales con resultados de agrupamiento
y = 0.5793x^4 + 0.0122x^3 - 1.2418x^2 + 1.6248x
R^2 = 0.9381
Frecuencia basada en agrupamiento
Frecuencia basada en individuos
0.0
0.1
0.2
0.3
0.4
0.5
0.6
0.7
0.8
0.9
1.0

Figura 2

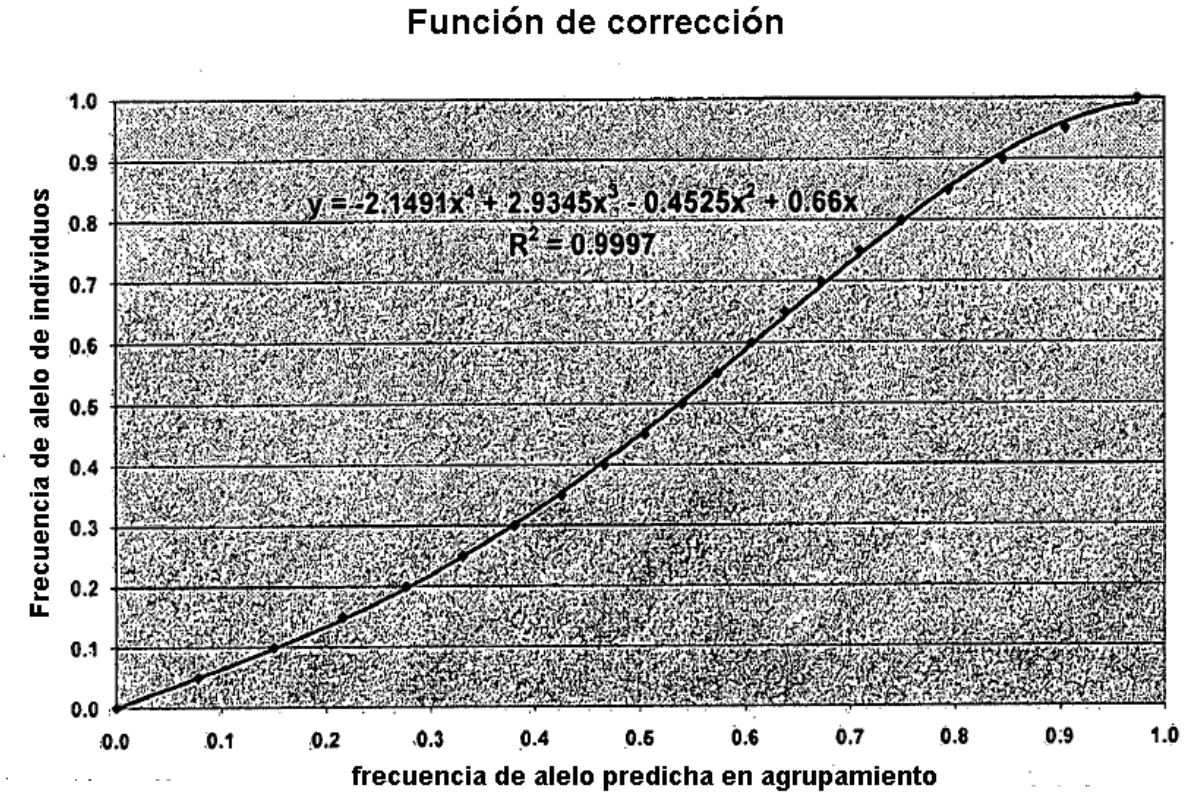

Función de corrección
y = -2.1491x⁴ + 2.9345x³ - 0.4525x² + 0.66x
R² = 0.9997
Frecuencia de alelo de individuos
frecuencia de alelo predicha en agrupamiento
0.0
0.1
0.2
0.3
0.4
0.5
0.6
0.7
0.8
0.9
1.0

**Figura 3**

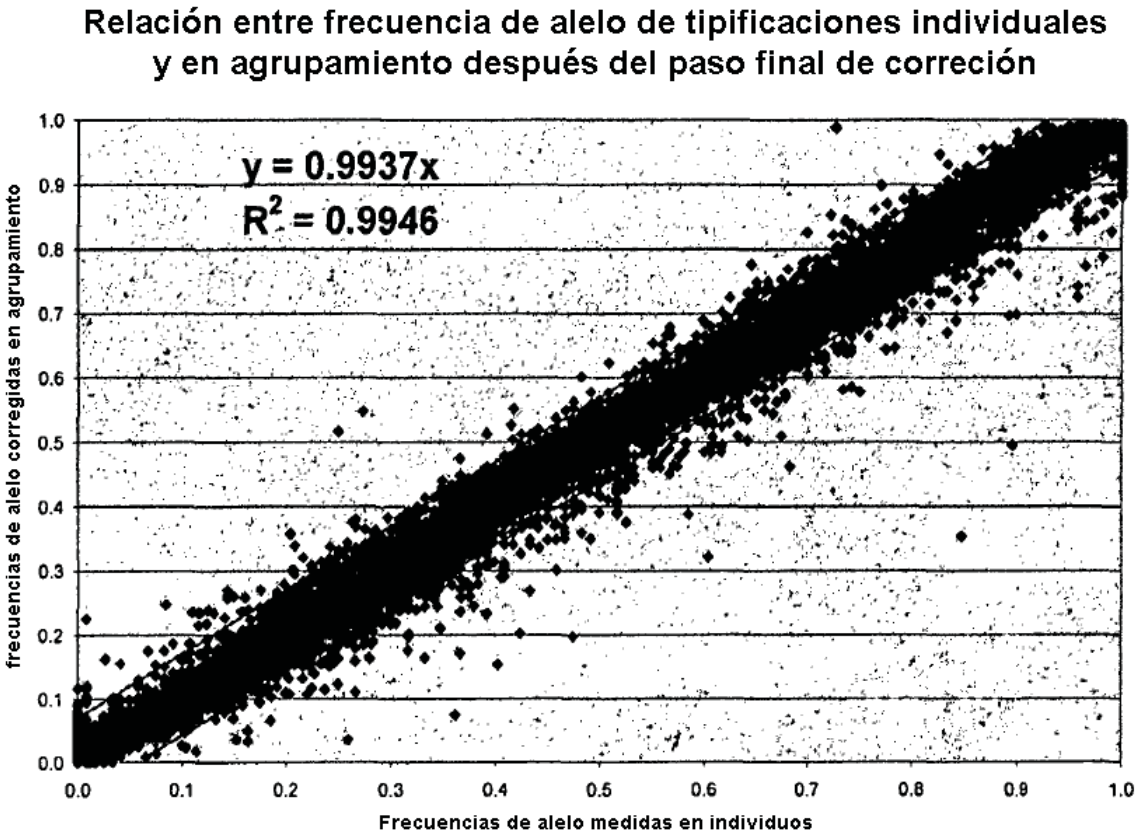

Relación entre frecuencia de alelo de tipificaciones individuales y en agrupamiento después del paso final de correción
y = 0.9937x
R2 = 0.9946
frecuencias de alelo corregidas en agrupamiento
Frecuencias de alelo medidas en individuos
0.0
0.1
0.2
0.3
0.4
0.5
0.6
0.7
0.8
0.9
1.0

Figura 4

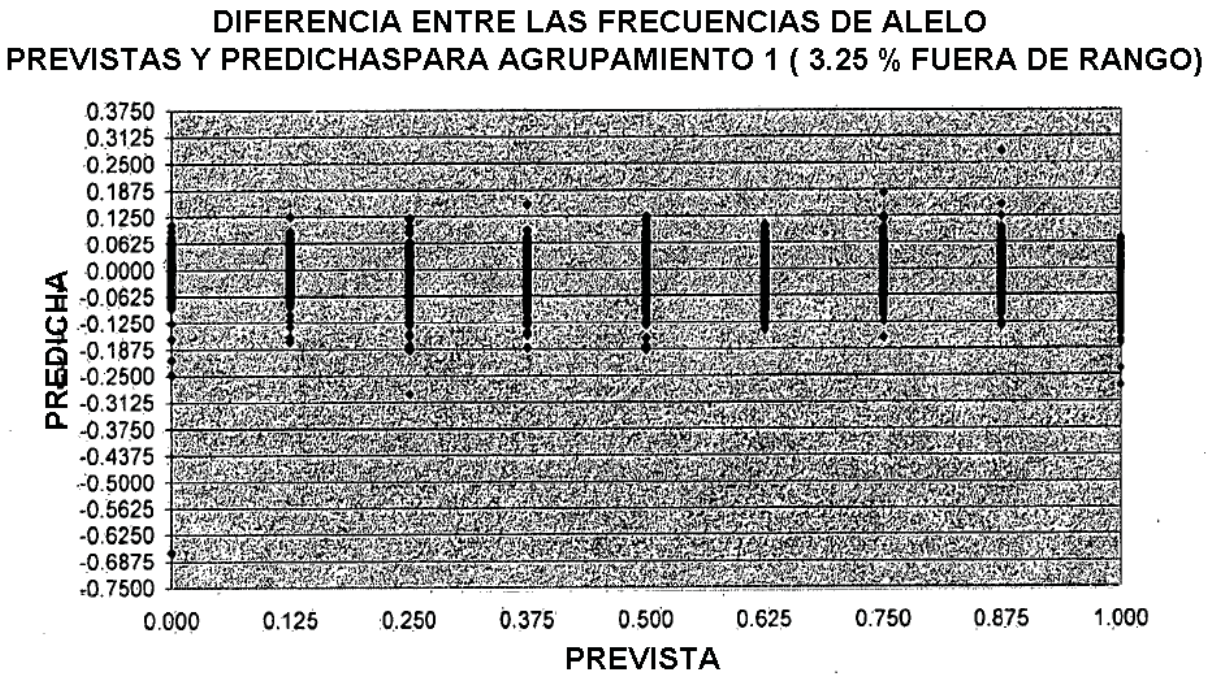



Figura 5

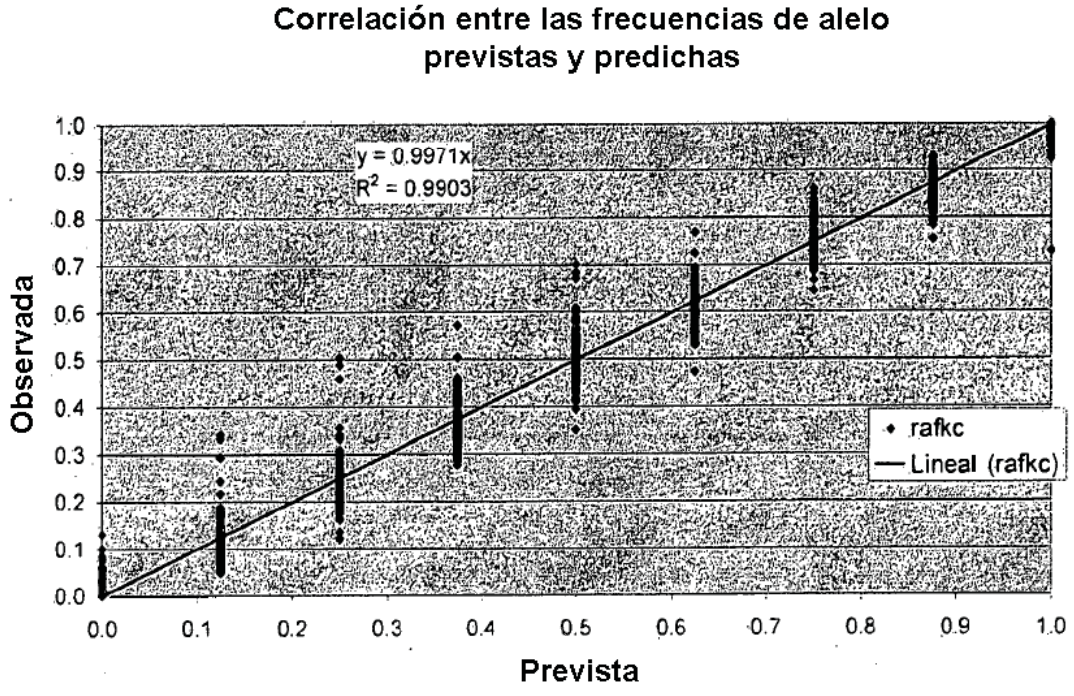


Figure 6

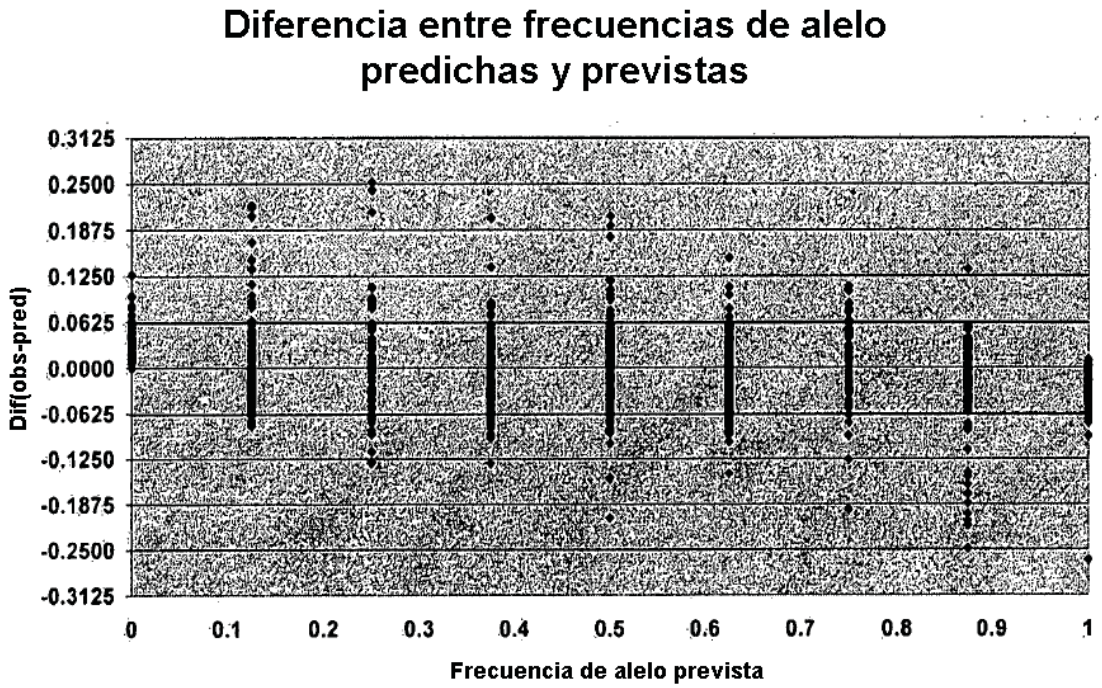
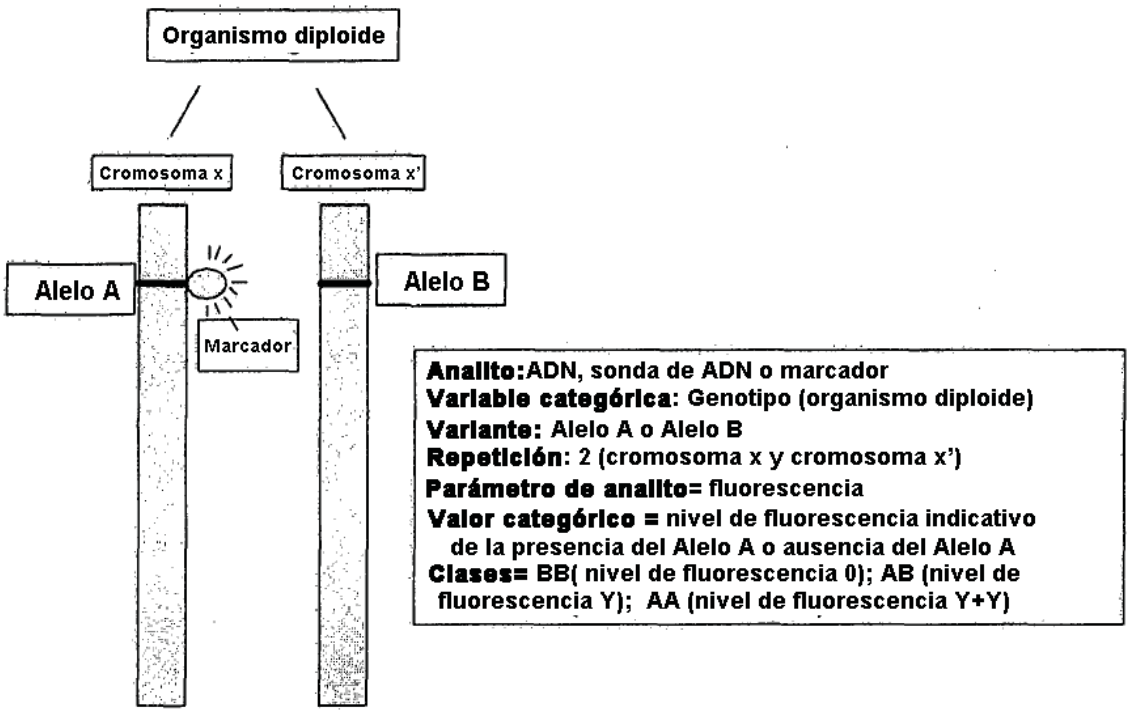

Diferencia entre frecuencias de alelo predichas y previstas
Dif(obs-pred)
0.3125
0.2500
0.1875
0.1250
0.0625
0.0000
-0.0625
-0.1250
-0.1875
-0.2500
-0.3125
0
0.1
0.2
0.3
0.4
0.5
0.6
0.7
0.8
0.9
1
Frecuencia de alelo prevista

Figura 7


Organismo diploide
Cromosoma x
Cromosoma x'
Alelo A
Marcador
Alelo B
**Analito:** ADN, sonda de ADN o marcador
**Variable categórica:** Genotipo (organismo diploide)
**Variante:** Alelo A o Alelo B
**Repetición:** 2 (cromosoma x y cromosoma x')
**Parámetro de analito=** fluorescencia
**Valor categórico =** nivel de fluorescencia indicativo de la presencia del Alelo A o ausencia del Alelo A
**Clases=** BB( nivel de fluorescencia 0); AB (nivel de fluorescencia Y); AA (nivel de fluorescencia Y+Y)