

Análisis a gran escala del repertorio expresado de la cadena ligera k del receptor de linfocitos B de sangre periférica en 4 adultos sanos.

Martínez Barnetche J¹, Téllez Sosa JM¹, Gómez Barreto RE¹, Vergara Castañeda E¹, Gutiérrez Peña E², Romero Martínez M³.

1 CISEI, INSP; 2 IMAAS, UNAM; 3 CIEE, INSP

Palabras Clave: Repertorio, diversidad, Kappa, Linfocitos B.

Keywords: Repertory, diversity, Kappa, B Lymphocytes.

Introducción

El sistema inmune adaptativo presente en los seres vivos vertebrados posee la capacidad de generar una extraordinaria diversidad en el repertorio de receptores de antígenos específicos que se expresan de manera clonal en linfocitos B y T (BCR y TCR, respectivamente), se estima que la diversidad potencial de estos repertorios está entre 1×10^{11} y 1×10^{18} . Estos repertorios son entes dinámicos y los procesos que los regulan son conocidos como la teoría de la selección clonal de Burnet. El principal problema para el estudio integral y comprensión del repertorio es su gran diversidad, ya que las técnicas usadas tradicionalmente como la espectratipificación, PCR Múltiplex, FACS tienen sesgos metodológicos inherentes a las mismas técnicas que impiden un análisis integral del repertorio inmune. El desarrollo de técnicas de secuenciación masiva de ADN como la pirosecuenciación permite el análisis del repertorio inmune de manera global, a costos y tiempo razonables, por lo que en este trabajo se diseña una nueva estrategia para investigar la diversidad del repertorio expresado para IgK basada en esta técnica. Al expandir nuestro entendimiento sobre el sistema inmune y sus repertorios se podrá analizar de mejor forma la interacción entre el huésped y agentes patógenos, identificar biomarcadores con fines pronósticos o diagnósticos, detallar la respuesta a vacunas, examinar el comportamiento del repertorio inmunológico en neoplasias y en enfermedades autoinmunes.

Hipótesis: Es posible el estudio de la diversidad expresada del repertorio del receptor de Linfocitos B, desde un punto de vista cualitativo y semicuantitativo, mediante el desarrollo de una nueva estrategia metodológica.

Objetivo: Desarrollar una nueva estrategia metodológica basada en secuenciación masiva en paralelo de ADN, para el estudio de la diversidad expresada de la cadena ligera κ de inmunoglobulina (IgK) humana en sangre periférica de individuo adulto sanos. Describir de manera cualitativa y semicuantitativa el repertorio expresado para IgK en linfocitos B de sangre periférica. Comparar los repertorios expresados para IgK en los sujetos sanos voluntarios.

Resumen

Los organismos vertebrados tienen un sistema inmune adaptativo capaz de generar receptores de antígeno con el potencial de reconocer a casi cualquier antígeno. Cada linfocito T y B expresa un receptor único, esta población de receptores es conocida como repertorio inmunológico. Hasta ahora, el conocimiento de la diversidad y la estructura del repertorio no ha sido reportada con precisión. Con la finalidad de explorar esta diversidad con una profundidad nunca antes lograda, exploramos el uso de nuevas tecnologías de secuenciación masiva de ADN. En este trabajo reportamos un análisis sin precedentes del repertorio expresado para la cadena ligera Kappa en individuos sanos: el número de secuencias obtenidas superó a las registradas en GenBank, se descubrió la expresión de rearreglos no reportados previamente y se reveló convergencia en el repertorio expresado en individuos sanos. Este nuevo abordaje para el estudio del repertorio se presenta como una opción práctica para enriquecer el conocimiento sobre los repertorios de linfocitos, su diversidad y estructura, así como para una mejor comprensión de la respuesta inmune, proyectando la identificación de biomarcadores con fines diagnósticos o terapéuticos.

Palabras clave: Repertorio, diversidad, Kappa, Linfocitos B.

Introducción

El sistema inmune adaptativo presente en los seres vivos vertebrados posee la capacidad de generar una extraordinaria diversidad en el repertorio de receptores de antígenos que se expresan de manera clonal en linfocitos B y T (BCR y TCR, respectivamente). Cada linfocito expresa un receptor único, estimándose que la diversidad potencial del repertorio de linfocitos está entre 1×10^{11} (Davis y Bjorkman 1988) y 1×10^{18} (Sanz 1991). Dicha diversidad le permite identificar y reaccionar potencialmente ante casi cualquier antígeno (Silverstein 1999). El repertorio de linfocitos T y/o B es definido como una colección de linfocitos, caracterizado por su receptor específico para antígeno. Este repertorio es dinámico debido a inactivación funcional, selección negativa de aquellas clonas potencialmente auto-reactivas y selección positiva de clonas reactivas a antígenos a los que el sujeto se ve expuesto a lo largo de su vida, tales procesos son conocidos como la teoría de la selección clonal (Burnet 1959). La gran diversidad en el repertorio de receptores de antígeno para linfocitos representa el problema central para su estudio y comprensión. Las técnicas usadas de manera convencional como la espectratipificación, PCR Múltiplex, FACS tienen como principal desventaja que no realizan un análisis global del repertorio inmunológico por limitaciones inherentes a las técnicas mismas, sesgando así los resultados obtenidos mediante ellas. La pirosecuenciación masiva en paralelo ha sido usada para el estudio del repertorio inmunológico de peces Zebra, permitiendo un análisis cualitativo y semi-cuantitativo sin sesgos metodológicos y ampliando el conocimiento previo sobre el tema (Weinstein, 2009). Por lo que se plantea sea parte de una nueva estrategia metodológica para el análisis de la diversidad de dichos repertorios y las modificaciones que sufren ante un reto antigénico podría brindar la oportunidad para ampliar el conocimiento sobre la dinámica de la selección clonal, además de ayudar a la comprensión de los procesos fisiopatológicos asociados a enfermedades infecciosas y autoinmunes, además de la búsqueda de biomarcadores basados en la diversidad del repertorio expresado (Boudinot, 2008).

Estructura del receptor de antígeno para linfocitos B

El receptor de antígeno de linfocitos B (BCR del inglés B cell receptor) es una glucoproteína heterodimérica tipo I de la familia de las inmunoglobulinas, que está anclada a la superficie celular y se compone de dos cadenas pesadas idénticas (IgH) y dos cadenas ligeras idénticas (IgL) que pueden ser del isotipo κ o λ . La región del BCR que reconoce al antígeno y que confiere especificidad clonal es la región amino-terminal, conocida también como región variable (V), que es diferente para cada clona. La región carboxilo-terminal es conocida como región constante (C).

La especificidad del receptor de antígeno es dada por la combinación de la región variable de IgH e IgL (VH y VL respectivamente) (Davis 2004). La región VH está constituida por recombinación de segmentos V, D y J, en tanto que la región VL está formada por segmentos V y J.

Mecanismos para la generación de diversidad en BCR

La diversidad del repertorio de receptor de linfocitos se genera de forma aleatoria mediante recombinación somática de segmentos génicos codificados en la línea germinal. Este proceso de recombinación se lleva a cabo de manera clonal solamente en linfocitos en etapa de maduración en órganos linfoides primarios. La región VH está compuesta por segmentos V (variable), D (diversity) y J (junction), en tanto que la región VL se compone de segmentos V y J. La dimensión del repertorio está parcialmente fijada por la diversidad combinatoria de los segmentos génicos codificados en la línea germinal, no obstante, debido a la ambigüedad en los sitios de corte y reparación de los extremos involucrados durante la recombinación y a la adición de nucleótidos por la desoxynucleotidil-transferasa terminal (TdT) en las uniones V-D, D-J y V-J la diversidad del repertorio es mucho mayor a la simple combinatoria de elementos de línea germinal (Max, 1999).

La unión del anticuerpo con el antígeno se determina por la estructura de los CDR1, CDR2 en diferentes segmentos del segmento V y sobre todo en la CDR3 de la región VH y VL, cuya variabilidad deriva de la yuxtaposición de diferentes segmentos V, D, J y la adición aleatoria de nucleótidos. Aunque la

contribución del CDR3 de VH es crucial en determinar la especificidad, el apareamiento preferencial con ciertas cadenas de IgL también contribuye a determinar la especificidad del anticuerpo (Davis, 2004).

El estudio de la diversidad del repertorio se ha realizado mediante técnicas convencional como FACS (Kedzierska 2008), espectrotipificación (Yao 2005), PCR múltiplex (Mariani 2009). El problema con estas técnicas es que son poco útiles para el estudio integral del repertorio, poco prácticas, además de ser caras y contener sesgos inherentes a las mismas técnicas impidiendo el análisis cualitativo y semi-cuantitativo del repertorio. Lo anterior conduce al poco conocimiento del repertorio expresado de linfocitos con la necesidad imperante de desarrollar nuevas estrategias para el análisis del mismo.

Mediante el uso de tecnologías de secuenciación masiva (Margulies, et al. 2005) , el empleo de herramientas bioinformáticas y bajo el enfoque de la biología de sistemas, nuestro grupo de trabajo ha desarrollado una nueva estrategia para estudiar la diversidad del repertorio de linfocitos de manera cualitativa y cuantitativa que puede proporcionar información fundamental para comprender mecanismos fisiológicos del sistema inmune adaptativo ante procesos infecciosos, enfermedades autoinmunes y eventos como la vacunación o exposición a ambientes específicos.

El objetivo de este estudio es conocer por primera vez la diversidad del repertorio expresado de la cadena ligera kappa en individuos sanos, sin sesgos, para así obtener nueva información que permita el análisis integral de este sistema biológico permitiendo así el poder diseñar modelos predictivos sobre la respuesta inmune.

Material y Métodos

Obtención de RNA: Mediante venopunción de la parte anterior del codo de 4 individuos adultos sanos con edades entre 29 y 40 años de edad, se colectaron 10 ml. de sangre periférica en un tubo con EDTA, dicha intervención contó con la aprobación del comité de Bioética y de Investigación del INSP. A partir de esta muestra biológica se aislaron células mononucleares en un gradiente de Ficoll. A continuación se purificó el RNA total de estas células por el método de TRIzol. La integridad del RNA fue analizada por electroforesis en gel de agarosa al 1%, su concentración fue determinada por espectrofotometría.

Generación de genotecas de cDNA: Se usó 1µg de RNA total para la síntesis de cDNA, usando el kit comercial SMART RACE cDNA Amplification Kit de Clontech. Se siguió el protocolo descrito por el proveedor para la síntesis de dichas genotecas (Clontech SMART RACE cDNA Amplification Kit User Manual).

Generación de amplicones IgKV: Las genotecas de cDNA fueron usadas como molde para generar genotecas de amplicones de la región variable del gen VLK mediante el método RACE 5' modificado del kit de Clontech. La modificación consiste en sustituir el oligonucleótido UPM del kit por el oligonucleótido FP_Amp_Nup_Piro. Dicho oligo es similar al UPM con la salvedad de que en el extremo 5' contiene una secuencia adaptadora "A" necesaria para el protocolo de Pirosecuenciación masiva en paralelo en la plataforma FLX 454 de Roche.

Los amplicones de IgKV fueron generados mediante PCR RACE 5' usando los oligonucleótidos FP_Amp_Nup_Piro y IGKC_Piro que prima específicamente en el gen que codifica la región constante de las cadenas ligeras Kappa (CLK), este oligo también contiene la secuencia adaptadora "B" para la secuenciación masiva en la plataforma FLX. Ambos oligos están diseñados de acuerdo a los estándares descritos en el protocolo de Pirosecuenciación de amplicones GS del manual GS FLX Amplicon DNA Library Preparation Method. Los productos generados por PCR RACE 5' fueron analizados mediante electroforesis en geles de agarosa al 1%, se

espera un producto con tamaño aproximado de 500 pb, posteriormente fueron purificados usando columnas de Promega.

Secuenciación masiva en paralelo de amplicones IGKV: Los amplicones generados fueron analizados por fluorescencia en el equipo BioAnalyzer 2100 LabChip de Agilent para determinar su concentración y calidad. Se utilizaron aproximadamente 100 ng. de cada librería de amplicones para su amplificación clonal por PCR, usando el estuche GS emPCR III de Roche de acuerdo al protocolo del proveedor. La Pirosecuenciación de las genotecas se realizó en el equipo Genome Sequencer FLX System, desarrollado por 454 Life Sciences Corporation de Roche usando el estuche GS LR70 de acuerdo a las instrucciones del proveedor. Cuatro genotecas fueron secuenciadas, cada una en 2 de ocho carriles disponibles en una placa GS Pico titer LR70, una librería más fue secuenciada en 2 de 16 carriles disponibles de una placa similar. Las secuencias generadas con esta plataforma tienen una longitud promedio de 250 pb. La dirección de la secuenciación fue en antisentido con la finalidad de obtener información del segmento JK, el CDR3 y parte de la región VK.

Procesamiento de datos obtenidos: De acuerdo al software y protocolo de la plataforma GS FLX 454 Life Sciences-Roche se obtiene una base de datos de las secuencias en un archivo formato *fasta*, que fue depurado de secuencias menores a 150 pb, de aquellas que carecían de la secuencia clave o que no correspondían al extremo 5' de CK. Se realizaron análisis con longitudes de 130, 150, 180 y 230 pb a partir del segmento JK hacia el extremo 5', encontrando que con secuencias de 150 pb se obtiene una capacidad de asignación aceptable ($R^2=0.95$).

Asignación de uso de segmento VK para las secuencias obtenidas: Usando BLASTn como herramienta local se asignó el segmento génico usado en cada una de las secuencias (Altschul, 1990). Para validar a BLASTn como método de asignación de segmento VK, la base de datos fue examinada con las secuencias de IgK obtenidas de GenBank y comparadas con los datos reportados en una caracterización reciente (Collins, 2008).

Como complemento a este análisis se usó la herramienta IMGT/V-Quest, que es un sistema disponible en la red que permite identificar los segmentos V y J en cada secuencia, además de brindar información sobre la diversidad de unión, mutaciones, inserciones y supresiones de bases, así como datos sobre variabilidad alélica (Brochet 2008). La limitación de esta plataforma es que el análisis está limitado a 50 secuencias de manera simultánea, por lo que no fue usado con todas las secuencias obtenidas en la secuenciación masiva en paralelo. De cualquier forma sirvió para verificar los resultados de BLASTn en la asignación de segmento para algunas secuencias.

Estimación de la frecuencia clonal y agrupamiento de las secuencias:

Una vez que las frecuencias obtenidas fueron depuradas, se realizó agrupamiento clonal de las mismas usando el software TGICL con los siguientes parámetros $-p$ 98, $-l$ 145 y $-v$ 1. Este software realiza la función de agrupar secuencias idénticas o casi-idénticas, cuyo origen muy probablemente es clonal, mediante este agrupamiento se reduce la complejidad de la base de datos, se define la diversidad clonal en el repertorio expresado y se conoce la frecuencia de cada grupo de secuencias.

Resultados

Secuenciación masiva de genotecas Vk: Una vez depuradas las secuencias obtenidas en las corridas de acuerdo a los criterios descritos en la metodología, se obtuvieron 15,107 secuencias para la primer librería secuenciada (kE), 22,601 para L5, 27,641 para L6, 16,161 para L7, 18,963 para L8 (Tabla I). El primer dato llamativo es que una vez que se realizó la asignación de segmento para estas secuencias, todas las genotecas tuvieron al menos 78% de identidad con algún rearreglo KV-KJ, llegando a ser de 89% para una de ellas, lo que representa un número muy superior al de las secuencias IgK reportadas en GenBank. Para las secuencias restantes se realizó un BLASTn con la intención de encontrar identidad con el genoma humano, se encontró que una proporción importante (5-14%) de estas secuencias tuvieron identidad con el intron 5' del segmento Ck. Para el resto de las secuencias no fue posible encontrar identidad alguna (Tabla I).

Tabla I. Resultados de secuenciación masiva							
Genoteca	secuencias >150 pb	KV-KJ (%)		Intron KJ-KC (%)		Nohits (%)	
kE	15,107	12,910	85.5	1,491	9.9	706	4.7
L5	22,601	19,361	85.7	1,840	8.1	1,400	6.2
L6	27,641	24,751	89.5	1,616	5.8	1,274	4.6
L7	16,161	12,682	78.5	2,270	14.0	1,209	7.5
L8	18,963	16,558	87.3	1,307	6.9	1,098	5.8
GenBank	2,003	2,003	100.0	0	0.0	0	0.0

Tabla I. Resultado del proceso de secuenciación masiva de las 5 genotecas derivadas de sangre periférica de 4 individuos adultos sanos del sexo masculino y las obtenidas del Genbank (<http://www.ncbi.nlm.nih.gov/genbank/index.html>). La segunda columna se refiere al número total de secuencias incluidas para análisis subsecuente. La tercera columna se refiere al número absoluto y % de secuencias catalogadas como re-arreglos de IGK. La cuarta columna describe el total y % de secuencias que mapearon con el intron IGKJ-IGKC y que representan transcritos potencialmente “estériles”. La última columna describe el número y % de secuencias no mapean en el genoma humano y cuya identidad desconocemos.

Uso de la diversidad combinatoria potencial en sujetos sanos: Para caracterizar la proporción del repertorio de IGK que se expresa en Linfocitos B del repertorio potencial, se mapearon los rearreglos de las 5 genotecas mediante BLASTn. El número total de combinaciones VJ para K es de 200 ($40V \times 5J=200$), de las cuales el 48% se encontraban en todas las genotecas y en GenBank. Así mismo, se encontró un “agujero” en el repertorio, es decir, la ausencia de 14% de los rearreglos posibles en todas las genotecas y el

Genbank. Quizás el dato más significativo fue el encontrar que con esta nueva estrategia se obtuvo evidencia de la expresión de rearrreglos que no habían sido reportados previamente en GenBank, representando un porcentaje de 23% del total de combinaciones posibles, mostrando así que la estrategia puede aportar datos que han pasado por alto al usar metodologías convencionales para el estudio del repertorio.

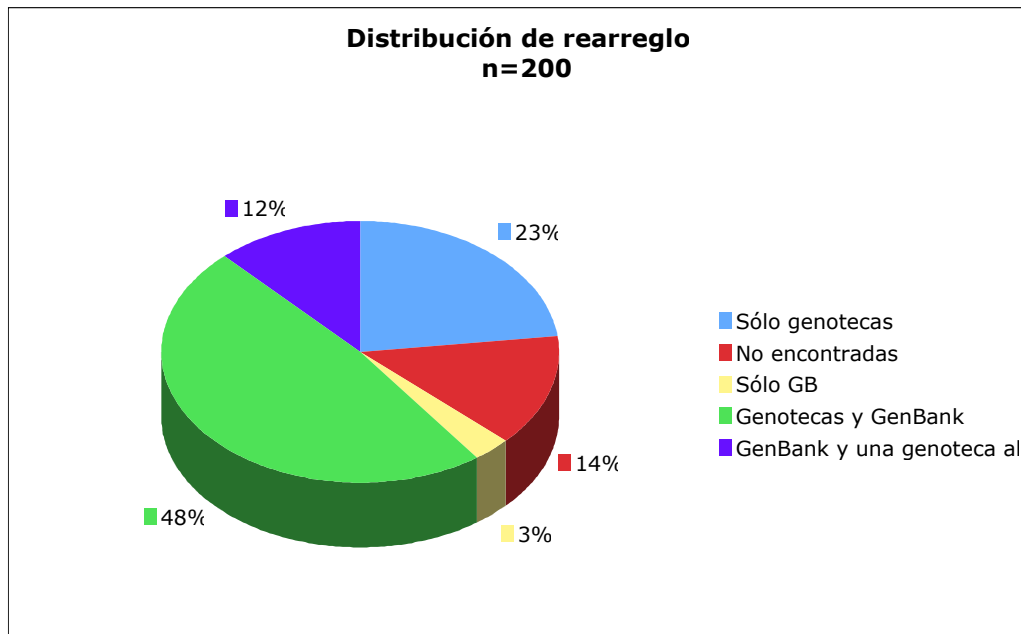


Figura 1. Distribución del rearrreglos posibles (200) generados mediante diversidad combinatoria de segmentos IGKV e IGKJ. En azul se presentan los rearrreglos presentes en las genotecas secuenciadas nunca antes descritos en el GenBank. En verde se representan re-arreglos compartidos en las cinco genotecas y el GenBank.

Comparación de la frecuencia relativa de re-arreglos en sujetos sanos:

Una vez asignados los rearrreglos de cada secuencia, se obtuvieron las frecuencias relativas de cada uno de ellos en cada genoteca para poder compararlas entre sí. Los sujetos de los que se obtuvo la muestra no se encuentran relacionados genéticamente y el repertorio de uno de ellos fue caracterizado en dos experimentos diferentes, al realizar la correlación de las frecuencias relativas en las cinco genotecas se encontró sorprendentemente un alto grado de correlación entre las cinco genotecas. Las cinco genotecas coincidían presentar los mismos rearrreglos altamente representados y un gran número de rearrreglos con frecuencias relativas bajas (Figura 2).

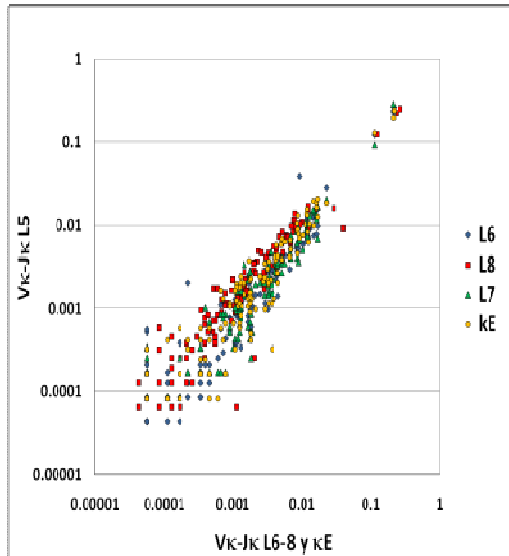


Figura 2. Correlación de la frecuencia relativa de rearrreglos en cinco genotecas de IGK de linfocitos B periféricos provenientes de cuatro individuos no relacionados genéticamente. El Eje x muestra la frecuencia (Log_{10}) de los re-arreglos de las genotecas L6, L7, L8 y kE. El eje y muestra la frecuencia (Log_{10}) de re-arreglos de la genoteca L5. En la esquina superior derecha de la gráfica se aprecian los re-arreglos altamente frecuentes KV1-39/1D-39-J1, KV2-28/D2-28-J5 y KV3-20/D3-20-J2.

Una vez que se estableció la similitud entre las cinco genotecas observadas, se realizó la descripción de las frecuencias de cada rearrreglo, que evidenció la expresión de transcritos para las 5 familias de genes V κ funcionales, se encontró que casi las dos terceras partes del total de las secuencias obtenidas (58%) corresponden a los rearrreglos IGKV1-39/1D-39/J1 (13%), IGKV2-28/2D-28/J5 (19.8%) y IGKV3D-15/J2 (24%). Resalta también que el hecho de que el único gen de la familia IV, IGKV4-1 obtiene casi el 5% de las secuencias totales, de igual manera los genes de la familia III, V3-11/3D-11 y V3-20/3D-20 representan casi el 10% de las frecuencias. De forma tal que tres rearrreglos VJ y 3 segmentos V más, acaparan casi el 75% de las secuencias totales, dejando al resto de los segmentos con frecuencias muy bajas, la mayoría con una frecuencia relativa de alrededor de 0.1% (Figura 3).

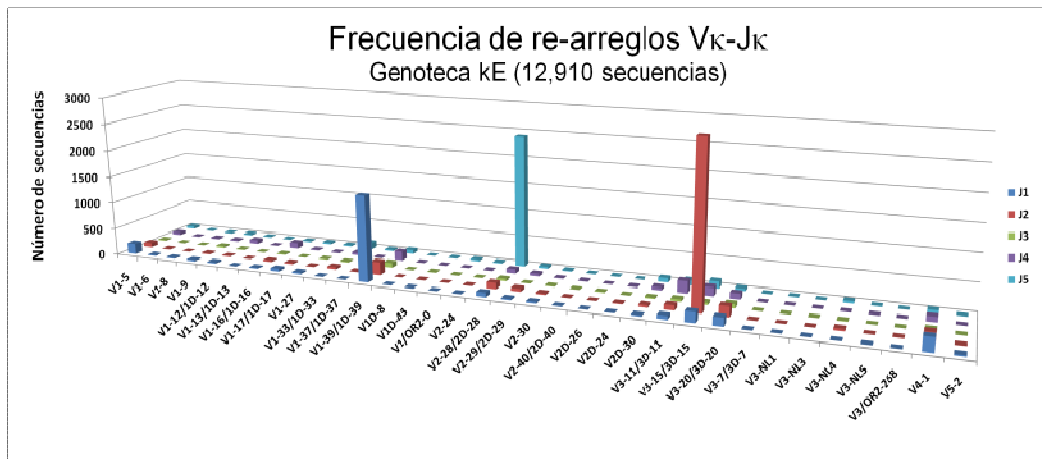


Figura 3. Caracterización de la genoteca kE de acuerdo con la diversidad de re-arreglos IGKV (eje x), IGKJ (eje z) y el número de secuencias no agrupadas (eje y). La caracterización se realizó mediante BLASTn. Nótese que los re-arreglos IGKV1-39/1D-39-IGKJ1; KV2-28/D2-28-J5 y KV3-20/VD3-20-J2 acaparan la mayor cantidad de secuencias.

Cuantificación de la diversidad en grupos de probable origen clonal común:

Una vez conocidas las distribuciones de las frecuencias de secuencias de acuerdo al re-arreglo usado en ellas, se realizó agrupamiento de las mismas con el software TGICL. Al hacer esto, minimizamos la redundancia de la base de datos a secuencias con gran identidad entre sí, que muy probablemente sean del mismo origen clonal, brindando así una nueva perspectiva para el análisis del repertorio expresado en los Linfocitos B. Bajo esta perspectiva sobresale la diferencia que existe entre la proporción de transcritos respecto al uso de segmentos V_k y la proporción de la diversidad clonal del repertorio, en específico de los segmentos con mayor proporción de transcritos: IGKV1-39/1D-39, IGKV2-28/2D-28 y IGKV3D-15, que al ser agrupados presentan una diversidad clonal menor a la obtenida antes del agrupamiento. Caso contrario a los segmentos IGKV1-33/1D-33, IGKV1-9, IGKV2-30, IGKV3-11/3D-11, IGKV3-20/3D-20 e IGKV4-1, cuya proporción de transcritos es baja, pero la proporción de la diversidad clonal llega a ser casi del 10% (Figura 4). De acuerdo a estos datos podemos inferir que ciertos segmentos génicos son transcripcionalmente muy activos, pero su diversidad clonal no es tan extensa. De igual forma se presenta de nueva cuenta el hecho de que pocos segmentos acumulan el mayor porcentaje de la diversidad.

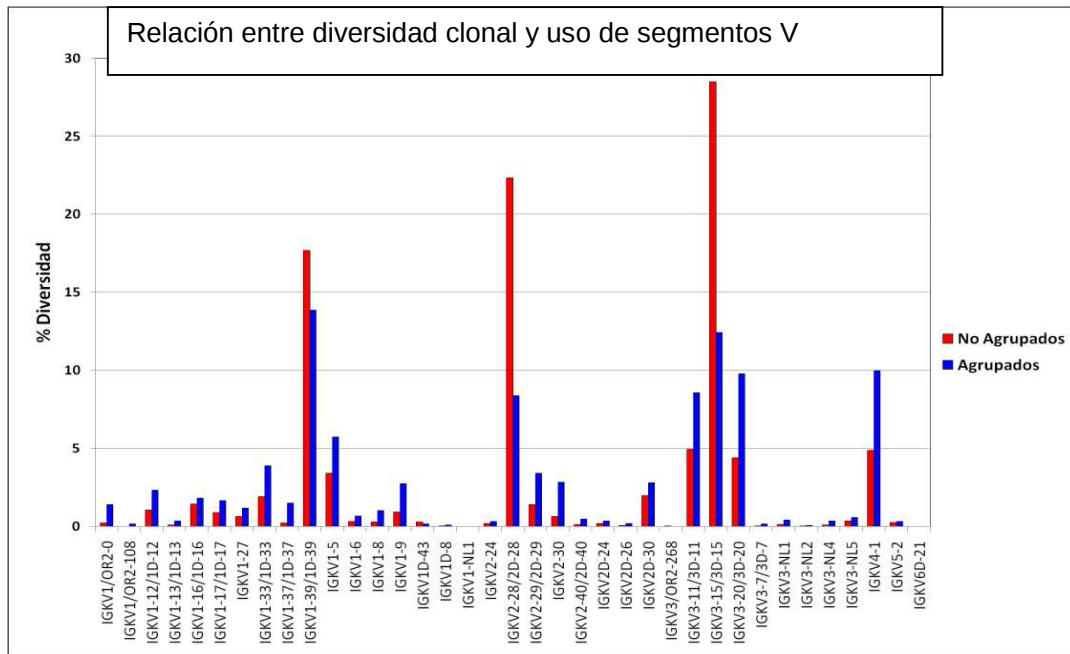


Figura 4. Diversidad del repertorio de acuerdo con el uso de genes IGKV en secuencias no agrupadas y en secuencias agrupadas con TGICL. El agrupamiento reduce la redundancia y permite la definición de grupos clonales. El análisis permite la identificación de diversidad clonal en grupos de linfocitos con el mismo segmento V. Estos datos sugieren que la diversidad clonal en el grupo IGKV2-28/2D-28, IGKV3-15/3D-15 e IGKV4-1 es baja, pero que las clonas son transcripcionalmente muy activas.

Los datos obtenidos en los análisis previos muestran que el repertorio expresado de IgK en individuos sanos puede ser de hasta 87% de la diversidad combinatoria total, en esa diversidad expresada la mayoría de los rearrreglos están representados a bajas frecuencias. El siguiente paso en el estudio del repertorio expresado fue el correlacionar la proporción de la diversidad total con la diversidad de segmento IGKV, al hacerlo para la genoteca kE se manifiesta que los rearrreglos dominantes son altamente redundantes, evidenciando que el incremento en la abundancia de transcritos no está asociado con un incremento proporcional de la diversidad, esto pudiera ser explicado por la activación de ciertas clonas de linfocitos, en particular de células plasmáticas que tienen una gran actividad transcripcional, o bien por la expansión de alguna clona en específico. La correlación de los sujetos sanos contrasta de manera contundente con la correlación obtenida para la base de datos de GenBank, que presenta redundancia limitada.

La distribución de la frecuencia de grupos de probable origen clonal común sigue una ley de potencia: Conforme a la información obtenida en el estudio del repertorio expresado para IGK hemos notado que la mayor parte las secuencias obtenidas (aproximadamente 75%) provienen de un número limitado de rearrreglos VJ y segmentos génicos, en tanto que la mayor parte de la diversidad del repertorio se encuentra representada con una baja frecuencia clonal. Con el objetivo de conocer el tipo de distribución que sigue la clonalidad en el repertorio de IGK, para las 5 genotecas se correlacionó el número de secuencias con probable origen clonal común con el número de clonas diferentes en una escala logarítmica, de esta forma queda de manifiesto que un número limitado de secuencias con probable origen clonal común acaparan la mayoría de los transcritos, en tanto que una gran proporción de la diversidad clonal está representada por un número limitado de transcritos. Esta distribución es característica de los sistemas libres de escala, que describen una proporción 80/20 definida por leyes de potencia (Barabasi, 2003). (Figura 5).

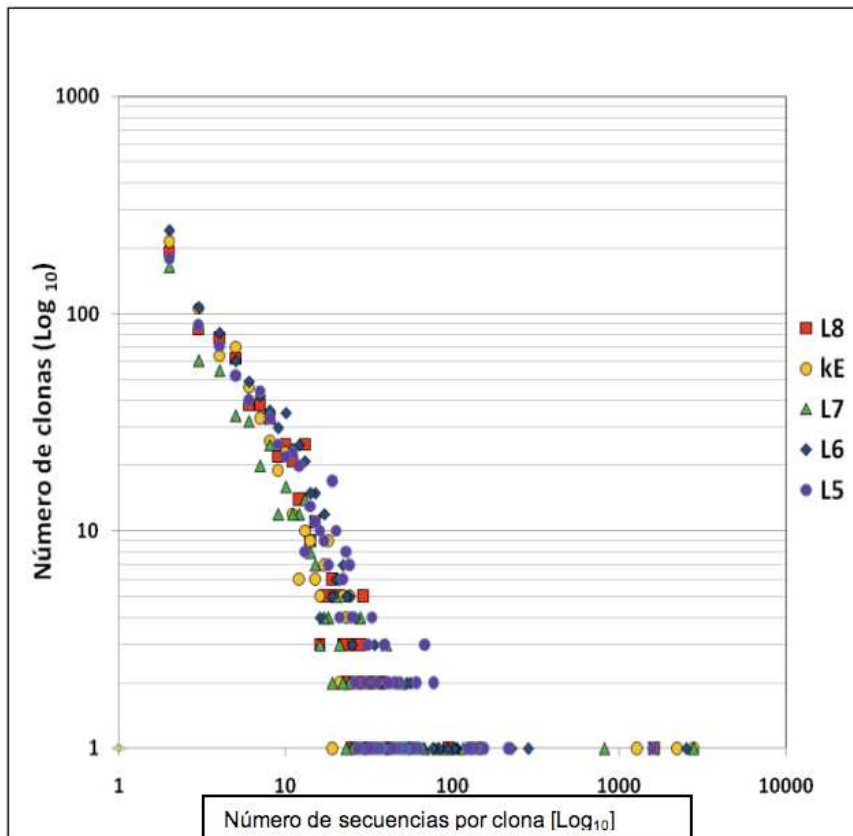


Figura 5. Representación del número de secuencias por clona (eje x) con respecto a el numero de clonas (eje y) en una escala $\log_{10}$ - $\log_{10}$ para las cinco genotecas. Unas cuantas clonas (que contienen el segmento IGKV1-39/1D-39, K2-28/2D-28 e KV3-15/3D-15) acaparan la mayoría de los transcritos (esquina *inferior derecha*), mientras que un alto número de clonas solo tienen pocos transcritos (esquina *superior izquierda*). La distribución de la frecuencia clonal parece ser libre de escala y se determina por una ley de potencia.

Uso de segmentos V en individuos sanos con respecto al Genbank: Las 5 genotecas estudiadas provienen de sujetos sanos, sin antecedentes patológicos de importancia, en tanto que los datos de GenBank provienen de múltiples experimentos e individuos. Para comparar la frecuencia clonal de un sujeto sano contra la frecuencia clonal de GenBank de acuerdo al uso de genes IGKV-IGKJ se usó un cociente normalizado de frecuencia para cada rearreglo, encontrando que los segmentos 1 y 2 de IGKJ están sobre-representados para GenBank, en tanto que los segmentos 4 y 5 se encuentran sobre-representados para el individuo normal, lo anterior en segmentos V relacionados con enfermedades autoinmunes tales como IGKV3-20 (Figura 6-A).

Al estudiar el uso del segmento IGKV3-20/3D-20 en las secuencias obtenidas de los sujetos sanos se encuentra que su frecuencia es de alrededor de 5% del total de las secuencias y de casi el 10% de la diversidad clonal (figura 4), en tanto que para GenBank es de alrededor del 20% para ambos (dato no mostrados). Siendo conocidos los sesgos inherentes de GenBank podría asociarse el uso del segmento IGKV3-20/3D-20 a un proceso patológico, para obtener mayor información al respecto se buscó de manera intencionada la presencia de términos asociados a enfermedades autoinmunes en los registros de GenBank para las secuencias IGKV3-20/3D-20 y se comparó con la aparición de estos términos en el resto de los registros de GenBank para IGK, se descubrió que la proporción de registros que contienen términos relacionados a enfermedades autoinmunes fue mayor para las secuencias que contienen el segmento IGKV3-20/3D-20 que para el resto de las secuencias de GenBank (Figura 6-B).

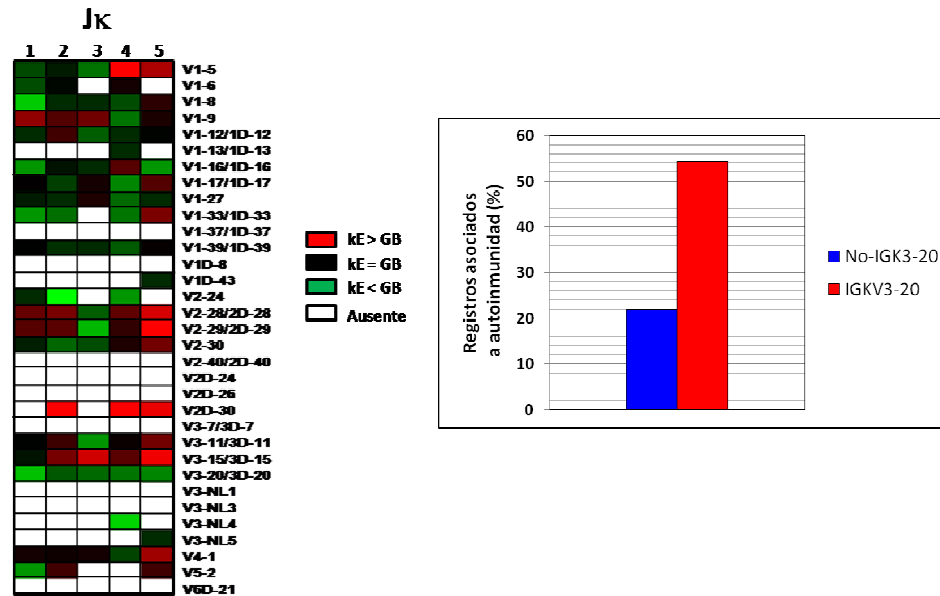


Figura 6: Comparación entre la frecuencia clonal de un individuo sano (kE) con la frecuencia clonal del GenBank según el uso de genes IGKV-IGKJ. **A)** El cociente de la frecuencia clonal relativa (kE/GB) se transformó a $\log_2$, el cual se representa en una escala semi-cuantitativa de color, donde los tonos rojos indican sobre-representación y los verdes sub-representación de kE con respecto al GenBank. Las diferencias entre proporciones fueron evaluadas mediante correlación de dos proporciones. **B)** Proporción de registros del GenBank que utilizan el segmento IGKV3-20/3D-20 que se encuentran enriquecidas de términos asociados con enfermedades autoinmunes “anti-DNA”

Discusión

El análisis de la diversidad del repertorio expresado para IgK en este trabajo brinda información sin precedentes sobre el tema, se encontró la expresión de rearrreglos Vk-Jk que no se habían reportado con técnicas tradicionales, el número de secuencias obtenidas es notablemente superior al que se encuentra reportado en GenBank, se identificaron los rearrreglos más frecuentes en cada individuo y de acuerdo a dichas frecuencias se encontraron similitudes en los repertorios expresados entre diferentes individuos, hallando que la distribución de tales frecuencias sigue una ley de potencia, además de datos que sugieren la edición de receptores de Linfocitos B, tales hallazgos serán descritos de forma amplia a continuación.

En este trabajo se presentan los resultados del uso de una nueva estrategia para el estudio del repertorio expresado de IgK de linfocitos B de sangre periférica de individuos sanos, que a diferencia de las técnicas usadas habitualmente (espectrotipificación, FACS, PCR múltiplex) nos permite abordar su estudio sin presentar sesgos metodológicos que limiten la evaluación de la diversidad real del repertorio. Esto se debe a la modificación de la técnica de RACE 5' con un oligonucleótido complementario a la región constante de la cadena ligera Kappa (CLK) y empleando un oligonucleótido universal con una secuencia conocida para el extremo 5' nos permite generar amplicones para todas los transcritos de IgK contenidos en el RNA, aún cuando no conozcamos el segmento génico que los codifica, logrando así obtener una genoteca que contiene teóricamente toda la diversidad del repertorio expresado para IgK. El locus de Igk involucra una región de aproximadamente 1.8 Mb codificada en el cromosoma 2p11.2 (Lefranc 2005). Se compone de una región proximal de aproximadamente 0.5 Mb que contiene 22-24 genes IGKV, 5 segmentos IGKJ y un segmento IGKC. En dirección centromérica se encuentra una duplicación parcial invertida de aproximadamente 0.45 Mb denominada región distal (D) que codifica para 17-22 genes IGKV y se encuentra separada de la región proximal por un intervalo genómico de 0.8 Mb (Tomlinson, 1995). Se calcula que la

duplicación del locus IGK ocurrió hace aproximadamente 1-2 millones de años y por lo tanto está ausente en primates no humanos (Kawasaki, 2001).

El uso de la pirosecuenciación masiva en paralelo ha demostrado ser útil para el estudio de la diversidad del repertorio, se ha descrito la diversidad del repertorio de la cadena pesada de los anticuerpos de los peces Zebra, encontrando que el uso de los segmentos VDJ no es uniforme, que la distribución de la frecuencia y diversidad del repertorio son similares entre individuos, y que existe convergencia en la diversidad del repertorio para la cadena pesada de anticuerpos entre los individuos estudiados. (Weinstein, 2009). También se ha usado esta nueva tecnología para la caracterización del repertorio de linfocitos T basados en la secuencia del CDR3 y en el subtipo de población efectora al que pertenecen. (Wang, 2009).

Los resultados generados para el estudio del repertorio expresado de la cadena ligera Kappa utilizando secuenciación masiva de nueva generación (Margulies, 2005) arrojaron una cantidad de secuencias casi 50 veces superior a las reportadas previamente en GenBank, además de encontrar expresión de rearrreglos que no habían sido reportados por técnicas previas, 23% de la combinatoria posible de segmentos IGKV/IGKJ, datos que nos sugieren un rendimiento superior de esta metodología, que permitirán una caracterización integral, conociendo así la constitución del mismo, su complejidad estructural y los cambios que experimenta a través del tiempo o por la exposición a diferentes estímulos (Linton, 2004).

El describir el uso de los genes IGKV expresados en cada individuo nos permite identificar los rearrreglos más abundantes, el agruparlos para definir grupos que probablemente tengan un origen clonal común nos permite teorizar sobre la actividad clonal y los procesos selectivos que han ocurrido sobre las mismas (Meffre, 2000), sugiriendo que las clonas más activas probablemente provengan de células plasmáticas con una gran actividad transcripcional, a diferencia de los rearrreglos que ven disminuida su frecuencia relativa al agruparlos clonalmente, hecho que podría sugerir el que se encuentran en estado de reposo.

Al realizar el análisis global del repertorio encontramos para todos los individuos que un grupo pequeño de secuencias con un probable origen clonal común (seis) acaparan un gran porcentaje del total de las secuencias expresadas, lo cual sugiere una distribución libre de escala, obedeciendo una ley de potencia (Barabasi, 2004), datos que coinciden con lo reportado por estudios con una metodología similar (Weinstein, 2009), sorprendentemente los rearrreglos dominantes son los mismos en todas las genotecas y cuentan con una distribución de frecuencias similar, al analizar dichas frecuencias relativas se descubrió una gran correlación entre ellas, hecho que podría sugerir la existencia de un perfil público para IgK, lo cual no necesariamente limitaría el reconocimiento a diversos antígenos, ya que se ha sugerido que la cadena pesada del anticuerpo es la que tiene juega un rol primordial en la especificidad de unión al antígeno (Xu, 2000). Además es posible que el origen de esta distribución de frecuencias en las secuencias para IgK pueda ser explicada por evolución convergente, que los sujetos al estar expuestos a un ambiente similar hayan seleccionado ciertos segmento génicos que brindan estabilidad a los anticuerpos expresados.

Mediante el agrupamiento de secuencias con un probable origen clonal común identificamos que algunos segmentos génicos se encuentran altamente expresados, pero que su diversidad clonal está limitada, tal es el caso de los segmentos IGKV1-39/1D-39, IGKV2-28/2D-28 y IGKV3D-15. Caso contrario a segmentos como IGKV1-33/1D-33, IGKV1-9, IGKV2-30, IGKV3-11/3D-11, IGKV3-20/3D-20 e IGKV4-1 que su frecuencia es relativamente baja, pero su diversidad clonal es abundante. Estos hallazgos sugieren la existencia de clonas transcripcionalmente muy activas y que una gran parte de la diversidad del repertorio expresado para IgK está expresado con frecuencias bajas, que podrían incrementarse ante ciertos estímulos antigénicos.

El identificar que la distribución del repertorio para la cadena ligera Kappa sigue una ley de potencia sugiere que está organizado de manera similar a las redes celulares del metabolismo (Jeong, 2000), redes de regulación transcripcional (Vogelstein, 2000) y redes de interacción de proteínas

(Almaas, 2007). Esto quiere decir que un gran número de clonas se encuentran están poco representadas, en tanto que unas cuantas acumulan la mayor frecuencia. En este tipo de redes su estabilidad depende de la presencia de nodos altamente conectados que interactúan con casi todos los demás nodos de la red, resulta llamativo el hecho de que en un sistema que depende de la diversidad de sus componentes dependa en gran manera de unos cuantos. Generalmente los sistemas que tienen distribución libre de escala se caracterizan porque son robustos ante las fallas, sólo resultan vulnerables cuando un nodo altamente conectado se ve afectado (Jeong, 2001), se han descrito fallas en la respuesta inmune adaptativa que pudieran explicarse por este mecanismo, al eliminarse ciertos segmentos génicos (Silverman, 2006), aunque no se conoce con precisión de que manera se reconfigura la red biológica para mantenerse viable después de la supresión de uno de sus módulos. Es necesario el continuar el estudio del repertorio de las cadenas pesadas de anticuerpos y la cadena ligera lambda para poder tener una imagen completa del repertorio inmune de anticuerpos y conocer su organización.

Durante el análisis de las secuencias obtenidas también se identificaron transcritos del locus IGK que contienen la secuencia del intron localizado en el extremo 5' del segmento IGKC, este dato resulta interesante ya que nos puede referir a la presencia de transcritos estériles, o bien edición de receptores potencialmente auto-reactivos. (Nemazee, 2006). La edición de receptores de linfocitos B es un proceso que ocurre durante su maduración, tiene como objetivo cambiar la especificidad del receptor en caso de auto-reactividad y es el principal mecanismo de tolerancia inmunológica para células B, dicho mecanismo acontece principalmente en el locus Kappa mediante rearreglos secundarios VJ, aunque estos rearreglos secundarios pueden no ser productivos (Halverson, 2004).

Al observar la expresión de los segmentos IGKV3-20/3D-20 en individuos sanos y compararlo con los datos reportados en GenBank se encontró que en los individuos sanos existe un mayor uso de los segmentos distales de IGKJ, dato que junto con la presencia de los transcritos estériles advierte

edición de receptor (Meffre, 2008), ya que este segmento génico ha sido asociado previamente con enfermedades autoinmunes.

La relevancia de la tipificación de un perfil público radica en la posible identificación de nuevos biomarcadores con un origen clonal común, por ejemplo en enfermedades como cáncer, infecciones o padecimientos autoinmunes, el conocer estos marcadores puede facilitar el desarrollo de inmunoterapia para estos padecimientos o servir como marcadores pronósticos de alguna enfermedad específica.

Conclusiones

Se ha realizado un análisis de la diversidad del repertorio expresado para IGK en individuos sanos, se demuestra que el análisis del repertorio a gran escala es viable y accesible. Se descubrió que la distribución de frecuencias en las secuencias y la diversidad clonal entre individuos es inesperadamente parecida. Se encontraron datos que sugieren edición de receptores para IGK. La distribución de las frecuencias de secuencias con un probable origen clonal sigue una distribución libre de escala, al igual que muchos otros sistemas biológicos. Se requiere el establecer un control experimental para verificar la certeza de estos datos, así como el ampliar su aplicación a VH y correlacionar ambos repertorios, además del repertorio de linfocitos T. El uso de esta estrategia para el estudio del repertorio del receptor de linfocitos T y B puede conducir a la identificación de biomarcadores basados en clonalidad, dado que un antígeno específico puede inducir tanto expansión o delección de algunas clonas específicas y al conocer las características del repertorio inmune podemos reconocer estos cambios puntuales.

Limitaciones del estudio

El estudio cuenta con varias limitaciones, el número de sujetos estudiados, es pequeño para poder sugerir la existencia de un perfil público expresado para el repertorio de IgK, debiera ampliarse la muestra y la diversidad de la misma para verificar si los hallazgos encontrados son consistentes. Se ha descrito que el reconocimiento a antígenos depende en mayor parte de la cadena pesada del receptor de inmunoglobulina, por lo que el estudio de la diversidad del repertorio de esta cadena es necesario para poder comprender el repertorio inmune y los cambios que experimenta ante diferentes circunstancias.

De igual forma dado que la estrategia experimental es nueva, es necesario el desarrollar controles experimentales adecuados para la técnica y así asegurar que no exista algún sesgo metodológico que no haya sido identificado aún.

Referencias:

Altschul, S. F., Warren, G., Millar, W., Myers E. W., and D. J. Lipman. 1990. Basic Local Alignment Search Tool. *J Mol Biol.* 215:403-410

Barabasi. AL. *Linked*. Plume Book. Londres. 2003.

Boudinot P., Marriotti-Ferrandiz M. E., Du Pasquier L., Bemansour A., Cazenave P. A., and A. Six. New perspectives for large-scale repertory analysis of immune receptors. *Molecular Immunology*, 2008; 45:2437-2445.

Brochet X, Lefranc MP, Giudicelli V. IMGT/V-QUEST: the highly customized and integrated system for IG and TR standardized V-J and V-D-J sequence analysis. *Nucleic Acids Res.* 2008;36:W503-8.

Burnet FM. *The clonal selection theory of aquired immunity*. Vanderbilt University. Press. Nashville, Tennessee. 1959

Clontech SMART RACE cDNA Amplification Kit User Manual. Protocol No. PT 3269-1. 2008 Clontech Laboratories, Inc.

Collins AM, Wang Y, Singh V, Yu P, Jackson KJ, Sewell WA. The reported germline repertoire of human immunoglobulin kappa chain genes is relatively complete and accurate. *Immunogenetics*. 2008;60(11):669-76.

Davis MM, Bjorkman PJ. T-cell antigen receptor genes and T-cell recognition. *Nature*. 1988;334:395-402.

Davis MM. The evolutionary and structural 'logic' of antigen receptor diversity. *Semin Immunol.* 2004;16(4):239-43.

Jeong H, Mason SP, Barabási AL, Oltvai ZN. Lethality and centrality in protein networks. *Nature*. 2001;411(6833):41-2.

Jeong H, Tombor B, Albert R, Oltvai ZN, Barabási AL. The large-scale organization of metabolic networks. *Nature*. 2000;407(6804):651-4.

Halverson R, Torres R, Pelanda R. Receptor editing is the main mechanism of B cell tolerance toward membrane antigens. *Nature Immunology* 2004;(5): 645-650

Kawasaki K, Minoshima S, Nakato E, Shibuya K, Shintani A, et al. Evolutionary dynamics of the human immunoglobulin kappa locus and the germline repertoire of the V kappa genes. *Eur J Immunol.* 2001;31(4):1017-28

Kedzierka K, La Gruta N, Stambas J, Turner S, Doherty P. Tracking phenotypically and functionally distinct T cell subsets via T cell repertoire diversity. *Mol Immunol.* 2008 February ; 45(3): 607–618.

Lefranc, M. P., and G. Lefranc. *the immunoglobulin facts book*. Academic press, 2001.

Lefranc, M. P., Giudicelli, V., Ginestoux, C., Jabado-Michaloud, J., Folch, G., Bellahcene, F., Wu, Y., Gemrot, E., Brochet, X., Lane, J., Regnier, L., Ehrenmann, F., Lefranc, G., and P. Duroux. 2009. IMGT, the international ImMunoGeneTics information system. *Nucleic Acids Research*. Vol 37.

Linton PJ, Dorshkind K. Age related changes in lymphocytes development and function, *Nature Immunology*, 2004;5(2):133-139

Margulies M, Egholm M, Altman WE, Attiya S, Bader JS, et al. Genome sequencing in microfabricated high-density picolitre reactors. *Nature*. 2005;437(7057):376-80.

Mariani S, Hwang SY, Foglietta M, Bonello L, Vitale C, Coscia M, Fiore F, Massaia B, Bruno B. Comprehensive assessment of the TCRBV repertoire in small T-cell samples by means of an improved and convenient multiplex PCR method. *Exp Hematol*, 2009;37(6):728-738

Meffre E, Casellas R, Nussenzweig MC. Antibody regulation of B cell development. *Nat Immunol*. 2000;1(5):379-85

Meffre E, Wardemann H. B-cell tolerance checkpoints in health and autoimmunity. *Curr Opin Immunol*. 2008;20(6):632-8.

Nemazee D. Receptor editing in B cells. *Adv Immunol*. 2000;74:89-126.

Nemazee D. Receptor editing in lymphocyte development and central tolerance. *Nat Rev Immunol*. 2006;6(10):728-40.

Pertea, G., Huang, X., Liang, F., Antonescu, V., Sultana, R., Karamycheva, S., Lee, Y., White, J., Cheung, F., Parvizi, B., Tsai, J., and J. Quackenbush. 2003. TIGR gene indices clustering tools (TGICL): a software system for fast clustering of large EST datasets. *Bioinformatics*. 22: 651-652.

Roche Applied Science. GS FLX Amplicon DNA Library Preparation Method Manual. 2007 December. Roche Diagnostics GmbH and Roche Applied Science Mannheim. Germany.

Roche Applied Science. GS FLX emPCR Method Manual. 2007 December. Roche Diagnostics GmbH and Roche Applied Science Mannheim. Germany.

Roche Applied Science. GS FLX Sequencing Method Manual. 2007 December. Roche Diagnostics GmbH and Roche Applied Science Mannheim. Germany.

Sanz. Multiple mechanisms participate in the generation of diversity of human H chain CDR3 regions. *J Immunol*. 1991 Sep 1;147(5):1720-9.

Silverman, G., and C. S. Goodyear. 2006. Confounding B-cell defenses: Lessons from a staphylococcal superantigen. *Nature Reviews Immunology*. 6:465-475.

Tomlinson IM, Cox JP, Gherardi E, Lesk AM, Chothia C. The structural repertoire of the human V kappa domain. *EMBO J.* 1995;14(18):4628-38.

Tonegawa S. Somatic generation of antibody diversity. *Nature.* 1983;302(5909):575-81.

Vogelstein B, Lane D, Levine AJ. Surfing the p53 network. *Nature.* 2000;408(6810):307-10.

Von Boehmer H, Melchers F. Checkpoints in lymphocyte development and autoimmune disease. *Nat Immunol.* 2010;11(1):14-20.

Weinstein JA, Jiang N, White RA 3rd, Fisher DS, Quake SR. High-throughput sequencing of the zebrafish antibody repertoire. *Science.* 2009;324(5928):807-10.

Xu JL, Davis MM. Diversity in the CDR3 region of V(H) is sufficient for most antibody specificities. *Immunity.* 2000;13(1):37-45.

Yao XS, Diao Y, Sun WB, Qin M, Tang XY. Analysis of the CDR3 length repertoire and the diversity of TCR alpha chain in human peripheral blood T lymphocytes. *Cell Mol Immunol.* 2007;4(3):215-220