Artículo original

Distribuciones y asociaciones entre rasgos estructurales y de localización genómica en microsatélites de bacterias patógenas

Distributions and associations between structural and genomic location traits in microsatellites of pathogenic bacteria

Carlos Miguel Martínez Ortiz[1]* https://orcid.org/0000-0001-8618-7301
Alejandro Rivero Bandinez[1] https://orcid.org/0000-0003-2396-346X
Nibaldo Hernández Mesa[1] https://orcid.org/0009-0006-2208-9116

[1]Instituto de Ciencias Básicas y Preclínicas "Victoria de Girón". Universidad de Ciencias Médicas de La Habana, Cuba.

*Autor para la correspondencia: cmmo@infomed.sld.cu

RESUMEN

Introducción: La presencia de microsatélites en bacterias y organismos eucariontes es mayor que la esperada por azar y se desconocen en detalle las fuerzas evolutivas y los rasgos estructurales que influyen en sus dinámicas de distribución y localización en el genoma. Tampoco se conocen las asociaciones entre dichos rasgos, lo que permitiría elucidar la dinámica molecular y evolutiva de estos marcadores.

Objetivo: Analizar las distribuciones y asociaciones entre los rasgos estructurales de los microsatélites de acuerdo con el tamaño de la unidad repetida, la región que

Esta obra está bajo una licencia: https://creativecomons.org/licenses/by-nc/4.0/deed.es_ES

ocupan en el genoma con relación a los genes, el tipo de genoma (principal o plasmídico), la composición del patrón y el grado de polimorfismo.

Métodos: Se escogieron 59 especies bacterianas, pertenecientes a 6 filos, todas de interés biomédico. Se procesaron 2780 y 966 secuencias genómicas y plasmídicas respectivamente. La detección de los microsatélites polimórficos se realizó con *MIDAS* y *PSSRExtractor*.

Resultados: Se encontraron 33506 microsatélites. La mitad de ellos se encuentran en regiones codificantes y sus flancos. Sus distribuciones por tamaño del patrón y región relativa a los genes mostraron diferencias significativas. Los tamaños del patrón influyen en el polimorfismo de sus loci y en otros rasgos como el número de copias y la entropía de los flancos. La composición nucleotídica también mostró relación con el grado de polimorfismo.

Conclusiones: Los análisis de asociación global, así como las frecuencias y distribuciones particularizadas para las 59 especies patógenas constituyen un valioso registro bioinformático que permite orientar estudios de carácter experimental relacionados a las dinámicas de los microsatélites y sus fuerzas evolutivas, para una mejor compresión de la biología molecular de la patogénesis.

Palabras clave: microsatélites; SSR, repetidos en tándem; patogenicidad; polimorfismo de SSR; número de copias variable.

ABSTRACT

Introduction: The presence of microsatellites in bacteria and eukaryotic organisms is higher than expected by chance, and the evolutionary forces and structural features that influence their distribution dynamics and localization in the genome are not fully understood. Neither are the associations between these traits known,

Esta obra está bajo una licencia: https://creativecomons.org/licenses/by-nc/4.0/deed.es_ES

which would allow us to elucidate the molecular and evolutionary dynamics of these markers.

Objective: To analyze the distributions and associations between the structural features of microsatellites according to the size of the repeat unit, the region they occupy in the genome in relation to genes, the type of genome (main or plasmid), the pattern composition and the degree of polymorphism.

Methods: 59 bacterial species belonging to 6 phyla, all of biomedical interest, were selected. 2780 and 966 genomic and plasmid sequences were processed, respectively. The detection of polymorphic microsatellites was performed with MIDAS and PSSRExtractor.

Results: 33506 microsatellites were found. Half of them are located in coding regions and their flanks. Their distributions by pattern size and region relative to genes showed significant differences. Pattern sizes influence the polymorphism of their loci and other features such as copy number and flank entropy. The nucleotide composition also showed a relationship with the degree of polymorphism.

Conclusions: Global association analyzes as well as the particularized frequencies and distributions for the 59 pathogenic species constitute a valuable bioinformatics record that allows to guide experimental studies related to the dynamics of microsatellites and their evolutionary forces, for a better understanding of the molecular biology of pathogenesis.

Keywords: microsatellites; SSR; tandem repeats; pathogenicity; SSR polymorphism; variable number of copies.

Recibido: 12/08/2023

Aceptado: 23/03/2024

Esta obra está bajo una licencia: https://creativecomons.org/licenses/by-nc/4.0/deed.es_ES

Introducción

Los microsatélites o repetidos de secuencias cortas (SSR, siglas en inglés), son secuencias de ADN constituidas por patrones pequeños de hasta 6 nucleótidos repetidos en tándem. El principal proceso mutacional que les da origen es el deslizamiento de cadena durante la replicación. La clasificación se basa principalmente en el tamaño de la unidad repetida o patrón, aunque existen discrepancias en cuanto a estas categorías.[1]

Su presencia en bacterias y organismos eucariontes es mayor que la esperada por el azar y se desconocen en detalle las fuerzas evolutivas y los rasgos estructurales que influyen en sus dinámicas. La variabilidad de los SSR es uno de los procesos que impulsan la plasticidad genómica. Las regiones genómicas que los contienen son potencialmente hipermutables debido a las expansiones (inserciones) y contracciones (deleciones) de sus patrones repetitivos y en bacterias se han reportado frecuencias de mutación tan elevadas como 10^{-1} por locus por generación.[2] Es por ello que el polimorfismo de los *loci* de los SSR es el fundamento de los métodos de genotipificación entre los que se encuentran la tipificación basada en el número variable de repetidos en tándem o el análisis de repetidos variables multilocus, que se aplican de manera rutinaria para la identificación de cepas patógenas.[3,4, 5]

La variación en la longitud del microsatélite, causada por el llamado deslizamiento de cadena durante la replicación, da lugar a cambios fenotípicos en respuesta a cambios ambientales que permiten una mejor adaptación en las bacterias. Cuando

Esta obra está bajo una licencia: https://creativecomons.org/licenses/by-nc/4.0/deed.es_ES

estos cambios afectan el fenotipo, producen la llamada variación de fase, la cual se ha observado en múltiples especies bacterianas.[6, 7, 8]

Un rasgo distintivo de los SSR es su distribución diversa entre las especies, incluso en especies muy relacionadas, lo que indica que están sujetas a cambios evolutivos rápidos.[9] Un análisis de 300 genomas de procariotas mostró que la distribución de los SSR varía con las especies, el tamaño de los genomas y el contenido de G+C.[10] Más específicamente, los SSRs con patrones pequeños (1 - 4 pb) son más abundantes en patógenos con adaptación al hospedero que tienen genomas reducidos (< 2 Mb) y bajo contenido de G+C (< 40 %), tales como *Mycoplasma y Haemophilus* spp. Por el contrario, los SSRs con patrones mayores (5 - 11 pb) son más frecuentes en gérmenes oportunistas con genomas mayores (> 4 Mb) y alto contenido de G+C (> 60 %) tales como *Burkholderia* y *Anabaena* spp [11, 12]. Basado en estas observaciones, se puede hipotetizar que las distribuciones diferenciales de los SSR en las bacterias se pueden correlacionar con la patogenicidad, aunque se necesitan más evidencias para corroborarlo.

La distribución de los SSR a lo largo del genoma muestra diferencias significativas entre regiones codificadoras y no codificadoras. En general, los SSR de mononucleótidos y dinucleótidos son excluidos de regiones codificantes, probablemente dado a que tienen mayor posibilidad de causar mutaciones de desplazamiento del marco de lectura en los genes. Por el contrario, aquellos SSR cuyos patrones son múltiplos de 3, están sobrerrepresentados en las regiones codificantes ya que sus expansiones o contracciones no afectan el marco de lectura de los genes. [13, 14, 15]

Estas distribuciones contrastantes en los microsatélites se deben analizar con una mayor dimensionalidad, observándolas en el contexto de otros rasgos estructurales que poseen estas secuencias, como pueden ser el tamaño del patrón,

Esta obra está bajo una licencia: https://creativecomons.org/licenses/by-nc/4.0/deed.es_ES

el número de copias, el grado de imperfección o el grado de polimorfismo. Por tales motivos el presente trabajo pretende analizar las distribuciones y asociaciones entre los rasgos estructurales de los SSR de acuerdo con el tamaño de la unidad repetida, su localización en el genoma relativa a los genes, el tipo de genoma (principal o plasmídico).

Métodos

Secuencias genómicas y variables de estudio

Se realizó una investigación no experimental de corte bioinformático en la que se analizaron las secuencias genómicas de 59 especies bacterianas, pertenecientes a 6 filos, escogidas por su interés médico al contener variantes patógenas en humanos. Los genomas bacterianos fueron extraídos del sitio del NCBI (ftp://ftp.ncbi.nlm.nih.gov/refseq/release/bacteria/), repositorio público y de libre acceso. El total de secuencias procesadas (todas RefSeq), genómicas y de plásmidos, fue 2780 y 966 respectivamente.

Las variables escogidas para el análisis descriptivo e inferencial son extraídas de las salidas de estos programas y miden rasgos estructurales de los SSR detectados (Tabla 1).

Tabla 1- Variables empleadas en el análisis

Variable(s)	Descripción
Copias	Cantidad de repeticiones de la unidad repetida del SSR.
Región	Localización en el genoma donde se posiciona el SSR con relación a los genes que codifican para proteínas (CDS: codificante, UTR 5´: flanco 5´ del CDS, UTR 3´: flanco 3´ del CDS, No-CDS: otra región distinta a las anteriores).

Esta obra está bajo una licencia: https://creativecomons.org/licenses/by-nc/4.0/deed.es_ES

Entropía 5´ y Entropía 3´	Entropía informacional de los flancos 5´ y 3´ del SSR. También llamada entropía de Shannon (en honor a Claude E. Shannon [16]). Aplicada a una secuencia *X*, con *N* posibles caracteres, mide su grado de complejidad. $$H(X) = -\sum_{i=1}^{N} p(x_i) \log_2 p(x_i)$$ $p(x_i)$ es la probabilidad del carácter x_i según sus frecuencias en *X*.
Inexactitud	Grado de inexactitud del SSR al compararlo con su versión exacta en el alineamiento
Longitud del Patrón	Cantidad de bases nucleotídicas de la unidad repetida: mono- (1 pb), di- (2 pb), tri- (3 pb), tetra- (4 pb), penta- (5 pb) y hexanucleótido (6 pb)
Patrón	Secuencia nucleotídica de la unidad repetida.
PIC (Siglas en inglés para *Polymorphism Information Content*)	Contenido de información polimórfica. También se conoce como heterocigocidad promedio esperada o diversidad genética de Nei. Da una medida de la probabilidad de que, para un *locus* único, un par de alelos escogidos al azar en la población sean diferentes $$PIC = 1 - \sum_i p_i^2$$ p_i es la probabilidad del alelo *i* según su frecuencias en la población del *locus*
Tipo de Genoma	Origen de la secuencia genómica: Principal o Plásmido.
Tamaño del Genoma	Cantidad de pares de base del genoma

Métodos computacionales y estadísticos

Para la detección de los SSR y la estimación de su grado de polimorfismo se emplearon los programas *MIDAS* y *PSSRExtractor* [17] (versiones 1.12 y 1.0 respectivamente disponibles en https://github.com/cmmo2020/PSSRExtractor).

La aplicación *MIDAS* no extrae información relacionada con las anotaciones estructurales o funcionales hechas a las secuencias genómicas. Por esta razón, se implementó el *script SSRFeatExtractor* en Java SE, empleando librerías de Biojava

Esta obra está bajo una licencia: https://creativecomons.org/licenses/by-nc/4.0/deed.es_ES

(Legacy), para extraer la localización de los SSR con respecto al gen codificador de proteínas más cercano.

Se implementó el script SSRMerge en Java SE para eliminar la redundancia de loci de los SSR debido a la elevada probabilidad de encontrar locus similares conservados que se repiten en distintas secuencias genómicas pertenecientes a una misma especie. La comparación de los flancos de los SSR presentes en todos los ficheros se hizo por alineamiento de secuencia global Nedleman-Wunsch y cuando dos SSR presentan más de un 90 % de identidad se escoge uno y se desecha el otro.

Todo el análisis descriptivo e inferencial se llevó a cabo empleando el lenguaje de programación R 4.0 y el entorno de programación RStudio 1.1.463. Para el análisis inferencial de asociación entre variables categóricas se empleó la prueba no paramétrica X^2 de Pearson; y entre variables categóricas y cuantitativas, es decir comparación entre tres o más grupos cuando la variable cuantitativa no cumple los criterios de normalidad, se empleó la prueba no paramétrica Kruskal-Wallis. Para contrastar la normalidad de las variables cuantitativas se empleó la prueba de Shapiro–Wilk.

Resultados

El tamaño promedio de los genomas principales analizados fue 3,32 Mb (rango: 0,8 Mb *Mycoplasma pneumoniae* - 7 Mb *Mycobacterium smegmatis*) y el de los plásmidos 0,13 Mb (rango: 0,01 - 0,32 Mb). Se detectaron un total de 33 506 loci de microsatélites (promedio por especie: 574 y rango: 43 *Treponema pallidum* - 3619 *Burkholderia pseudomallei*) y el PIC promedio por especie fue 0,15 (rango: 0,03 *Bordetella pertussis* - 0,41 *Haemophilus influenzae*). La distribución en frecuencias

Esta obra está bajo una licencia: https://creativecomons.org/licenses/by-nc/4.0/deed.es_ES

relativas porcentuales para los SSR según el tamaño de la unidad repetida fue: mono- (1pb) 18,3, di- (2pb) 7,1, tri- (3pb) 37,5, tetra- (4pb) 21,2, penta- (5pb) 5,1 y hexanucleótido (6pb) 11,8 (Tabla 2, valores resumidos tomados de la base de datos generada en el estudio y disponible en la plataforma de la revista como documento adjunto *SSR_Bacteria_BD.xlsx*).

Tabla 2- Distribución de los SSR para filos y especies, de acuerdo a la longitud del patrón y valores del Contenido de Información Polimórfica asociados.

Esta obra está bajo una licencia: https://creativecomons.org/licenses/by-nc/4.0/deed.es_ES

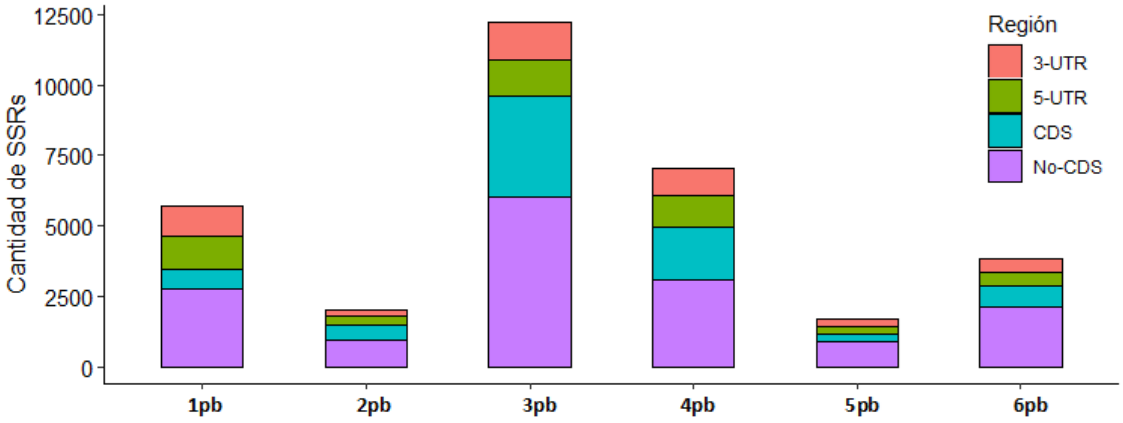
Filo	Especie	Gs	T(Mb)	Ps	T(Mb)	% Mono	Di	Tri	Tetra	Penta	Hexa	Total	PIC
Actinobacterias	Gardnerella vaginalis	7	1.7			56.8	0.0	15.1	20.0	2.7	5.4	185	0.198
Actinobacterias	Mycobacterium abscessus	23	4.9	3	0.15	2.2	0.4	58.5	20.2	3.4	15.2	506	0.136
Actinobacterias	Mycobacterium leprae	4	3.2			13.4	22.4	46.3	10.4	4.5	3.0	67	0.236
Actinobacterias	Mycobacterium smegmatis	4	7			0.4	0.4	66.9	18.7	2.3	11.3	257	0.09
Actinobacterias	Mycobacterium tuberculosis	81	4.4			5.2	0.4	81.3	7.4	1.6	4.1	557	0.076
Bacteroidetes	Bacteroides fragilis	14	5.3	6	0.06	23.4	5.2	29.1	38.0	2.7	1.6	368	0.15
Chlamydiae	Chlamydia pneumoniae	8	1.2			46.1	0.0	38.2	10.1	2.2	3.4	89	0.115
Chlamydiae	Chlamydia trachomatis	25	1.04	5	0.01	52.1	7.3	26.0	9.4	3.1	2.1	96	0.092
Firmicutes (Bacilli)	Bacillus anthracis	32	5.2	20	0.13	18.9	6.6	33.2	30.1	4.0	7.2	376	0.136
Firmicutes (Bacilli)	Bacillus cereus	52	5.3	65	0.3	19.0	6.1	29.1	31.4	5.4	9.0	1639	0.183
Firmicutes (Bacilli)	Bacillus subtilis	81	4.1	12	0.07	26.6	2.1	30.1	34.3	3.7	3.2	755	0.116
Firmicutes (Bacilli)	Bacillus thuringiensis	46	5.5	115	0.27	22.5	6.5	29.6	30.2	4.3	6.9	1738	0.173
Firmicutes (Bacilli)	Enterococcus faecalis	28	2.9	17	0.07	36.0	1.0	26.8	30.7	2.4	3.1	381	0.107
Firmicutes (Bacilli)	Enterococcus faecium	61	2.7	65	0.13	52.6	1.3	17.7	23.9	2.2	2.3	599	0.1
Firmicutes (Bacilli)	Listeria monocytogenes	89	2.9	14	0.15	20.6	2.5	31.1	24.3	3.8	17.8	399	0.189
Firmicutes (Bacilli)	Staphylococcus aureus	190	2.8	27	0.03	16.4	9.0	31.1	33.1	3.8	6.5	676	0.109
Firmicutes (Bacilli)	Staphylococcus epidermidis	17	2.5	16	0.03	5.9	9.2	49.4	29.5	3.7	2.2	271	0.133
Firmicutes (Bacilli)	Streptococcus agalactiae	48	2.1			27.0	7.0	28.8	27.0	5.1	5.1	215	0.122
Firmicutes (Bacilli)	Streptococcus mitis	5	2			15.8	12.9	26.7	32.7	9.9	2.0	101	0.134
Firmicutes (Bacilli)	Streptococcus mutans	9	2			17.6	3.4	44.5	28.6	3.4	2.5	119	0.146
Firmicutes (Bacilli)	Streptococcus pneumoniae	51	2.1			37.9	7.1	23.7	21.8	4.3	5.2	211	0.174
Firmicutes (Bacilli)	Streptococcus pyogenes	96	1.8			33.6	3.1	27.7	19.5	7.5	8.6	292	0.15
Firmicutes (Bacilli)	Streptococcus sanguinis	5	2.4			11.0	7.3	43.9	28.0	6.1	3.7	82	0.164
Firmicutes (Bacilli)	Streptococcus suis	31	2.2			33.6	10.4	23.6	23.9	6.2	2.3	259	0.194
Firmicutes (Clostridia)	Clostridioides difficile	25	4.2	4	0.11	13.4	11.3	34.6	35.0	3.1	2.8	575	0.124
Firmicutes (Clostridia)	Clostridium botulinum	30	3.9	18	0.18	6.7	12.5	38.0	34.1	4.6	4.2	1620	0.156
Firmicutes (Clostridia)	Clostridium tetani	2	2.8	2	0.07	2.7	7.1	34.5	46.0	2.7	7.1	113	0.169
Firmicutes (Mollicutes)	Mycoplasma pneumoniae	17	0.8			37.0	18.5	24.1	13.0	1.9	5.6	54	0.327
Proteobacteria (Alfa)	Rickettsia prowazekii	10	1.1			22.1	2.1	36.8	32.6	4.2	2.1	95	0.076
Proteobacteria (Beta)	Burkholderia mallei	32	2.9			2.2	25.6	39.4	5.9	6.0	20.9	1272	0.157
Proteobacteria (Beta)	Burkholderia pseudomallei	152	3.6	1	0.3	3.2	16.9	26.4	6.3	12.2	35.0	3619	0.363
Proteobacteria (Beta)	Neisseria gonorrhoeae	35	2.2	1	0.03	27.0	4.5	39.7	13.8	6.9	8.7	378	0.239
Proteobacteria (Beta)	Neisseria meningitidis	68	2.2			28.3	2.3	37.1	19.7	4.5	8.0	512	0.224
Proteobacteria (Beta)	Bordetella bronchiseptica	20	5.2			8.9	13.1	50.8	16.5	4.1	6.6	1000	0.136
Proteobacteria (Beta)	Bordetella parapertussis	15	4.8			8.4	11.3	51.5	16.0	5.2	7.6	406	0.073
Proteobacteria (Beta)	Bordetella pertussis	65	4.1			11.6	10.9	50.1	14.7	4.4	8.2	631	0.029
Proteobacteria (Epsilon)	Campylobacter jejuni	88	1.7	11	0.04	51.0	2.2	15.8	22.2	3.4	5.4	537	0.174
Proteobacteria (Epsilon)	Helicobacter pylori	120	1.6	3	0.01	44.4	6.8	19.0	19.0	4.0	6.7	1053	0.334
Proteobacteria (Gamma)	Acinetobacter pittii	11	4	12	0.09	9.0	1.4	40.3	34.1	4.5	10.7	290	0.174
Proteobacteria (Gamma)	Aeromonas hydrophila	21	4.8	5	0.16	8.0	1.0	60.1	18.5	3.6	8.8	842	0.122
Proteobacteria (Gamma)	Coxiella burnetii	10	2	2	0.05	38.3	0.0	21.8	22.6	5.3	12.0	133	0.183
Proteobacteria (Gamma)	Enterobacter cloacae	22	4.9	17	0.14	7.4	4.3	54.7	24.4	2.5	6.7	746	0.123
Proteobacteria (Gamma)	Escherichia coli	345	5	194	0.11	30.3	3.2	31.3	19.0	4.4	11.9	1862	0.133
Proteobacteria (Gamma)	Francisella tularensis	18	1.9	1	0.03	20.5	5.5	29.2	20.5	3.7	20.5	219	0.159
Proteobacteria (Gamma)	Haemophilus influenzae	42	1.8			19.2	1.5	16.3	30.6	8.5	24.0	412	0.407
Proteobacteria (Gamma)	Klebsiella aerogenes	20	5.4	10	0.2	6.2	2.3	73.1	12.5	1.6	4.2	1110	0.114
Proteobacteria (Gamma)	Klebsiella pneumoniae	205	5.3	172	0.16	20.8	3.0	46.4	15.6	3.4	10.9	1728	0.107
Proteobacteria (Gamma)	Legionella pneumophila	50	3.4	5	0.1	30.8	0.5	21.1	39.0	4.6	4.1	413	0.09
Proteobacteria (Gamma)	Pseudomonas aeruginosa	144	6.6	21	0.32	14.8	3.0	66.0	20.2	5.2	22.5	1138	0.112
Proteobacteria (Gamma)	Shigella dysenteriae	18	4.7	16	0.17	20.0	0.0	40.3	21.0	4.1	14.5	290	0.18
Proteobacteria (Gamma)	Shigella flexneri	29	4.6	17	0.17	24.8	1.9	36.7	16.3	4.8	15.6	270	0.144
Proteobacteria (Gamma)	Vibrio cholerae	43	2.5	1	0.2	16.6	4.3	38.4	28.0	2.4	10.4	211	0.12
Proteobacteria (Gamma)	Vibrio parahaemolyticus	48	2.8	8	0.09	15.8	2.0	29.4	29.6	4.5	18.6	398	0.182
Proteobacteria (Gamma)	Yersinia enterocolitica	18	4.7	12	0.08	22.8	3.2	34.3	20.0	7.2	12.6	470	0.188
Proteobacteria (Gamma)	Yersinia pestis	27	4.6	9	0.09	23.1	3.2	31.1	27.9	4.0	10.8	251	0.152
Spirochaetes	Borrelia burgdorferi	8	0.9	49	0.03	33.2	5.5	32.7	20.7	3.7	4.1	217	0.164
Spirochaetes	Leptospira interrogans	12	4.3	10	0.3	36.7	3.9	15.3	28.9	10.7	4.5	308	0.133
Spirochaetes	Treponema denticola	1	2.8			26.9	5.8	28.8	30.8	5.8	1.9	52	0.094
Spirochaetes	Treponema pallidum	2	1.1			60.5	9.3	2.3	23.3	0.0	4.7	43	0.115
		2780	**3.329**	**966**	**0.13**	18.3	7.1	37.5	21.2	5.1	11.8	**33506**	**0.1559**
			*		*								*

Gs:Genomas Principales; **Ps**: Plásmidos; **T(Mb)**: Tamaño en Mb; **PIC**: Contenido de Información Polimórfica; * Valores promedios.

Se obtuvieron las frecuencias de los SSR según el tamaño del patrón para las diferentes regiones relativas a genes codificadores de proteínas (Figura 1). De

(cc) BY-NC Esta obra está bajo una licencia: https://creativecomons.org/licenses/by-nc/4.0/deed.es_ES

acuerdo con esta definición, la etiqueta No-CDS no se refiere solo a regiones intergénicas, ya que en esta región pueden localizarse también todos los genes que codifican para otros ARN no mensajeros. Estas frecuencias pertenecen a ambos tipos de genomas (principal y plasmídico).



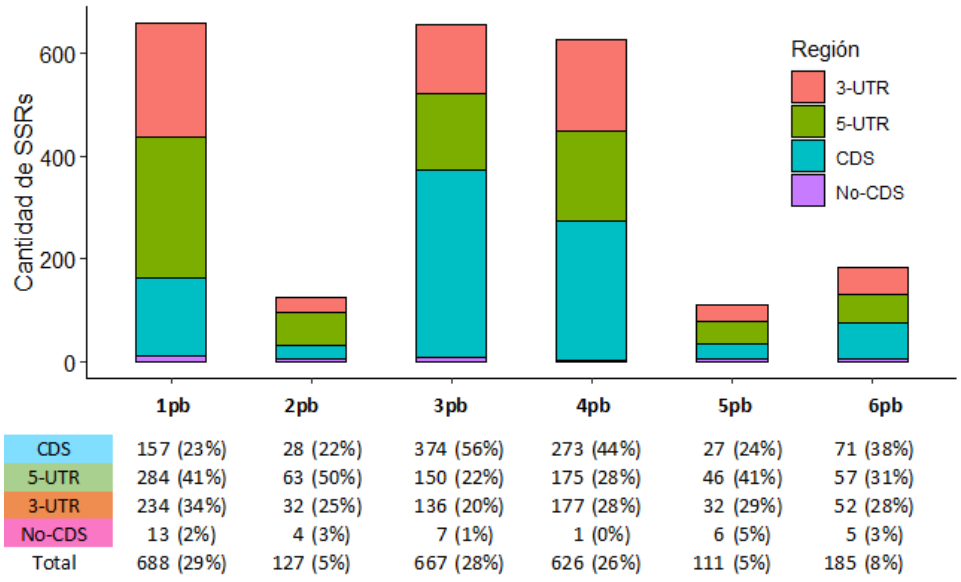
	1pb	2pb	3pb	4pb	5pb	6pb
CDS	728 (12%)	650 (27%)	3664 (30%)	1833 (26%)	266 (16%)	807 (21%)
5-UTR	1277 (21%)	391 (16%)	1253 (10%)	1133 (16%)	285 (17%)	476 (12%)
3-UTR	1146 (19%)	285 (12%)	1364 (11%)	964 (14%)	271 (16%)	481 (12%)
No-CDS	2899 (48%)	1052 (44%)	6133 (49%)	3111 (44%)	875 (52%)	2142 (55%)
Total	6050 (18%)	2378 (7%)	12414 (37%)	7061 (21%)	1697 (5%)	3906 (12%)

Fig. 1- Distribución de las cantidades de SSR por longitud del patrón y la región genómica, en todo el genoma.

Como las regiones definidas son conjuntos disjuntos y a su vez unidas cubren todo el espacio genómico analizable, lo primero que se puede destacar por simple inspección visual es que los SSR que se localizan en el conjunto unión CDS + UTRs fueron aproximadamente la mitad del total detectado, para todos los tamaños del patrón. Es decir, podemos afirmar que la mitad de los SSR detectados en los genomas pertenece a regiones codificantes y a sus flancos UTR. También se observa que los 5-UTR y los 3-UTR están igualmente distribuidos para todos los patrones, con diferencias de muy pocos puntos porcentuales, de modo que no se

Esta obra está bajo una licencia: https://creativecomons.org/licenses/by-nc/4.0/deed.es_ES

aprecia un marcado sesgo por uno u otro en particular. No obstante, teniendo en cuenta todas las regiones, el análisis de proporciones mediante prueba X^2 para las variables cualitativas tamaño del patrón y región, mostró asociación entre ellas ($p < 0,01$).



	1pb	2pb	3pb	4pb	5pb	6pb
CDS	157 (23%)	28 (22%)	374 (56%)	273 (44%)	27 (24%)	71 (38%)
5-UTR	284 (41%)	63 (50%)	150 (22%)	175 (28%)	46 (41%)	57 (31%)
3-UTR	234 (34%)	32 (25%)	136 (20%)	177 (28%)	32 (29%)	52 (28%)
No-CDS	13 (2%)	4 (3%)	7 (1%)	1 (0%)	6 (5%)	5 (3%)
Total	688 (29%)	127 (5%)	667 (28%)	626 (26%)	111 (5%)	185 (8%)

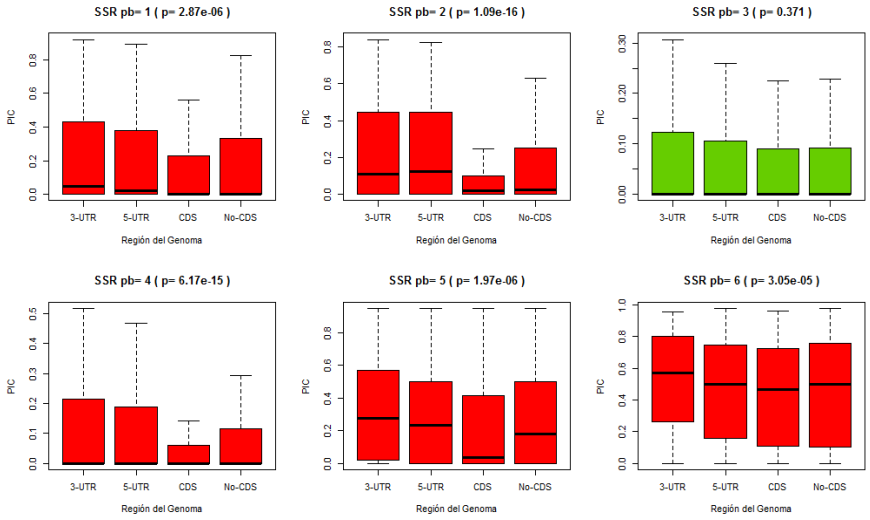
Fig. 2- Distribución de las cantidades de SSR por longitud del patrón y la región genómica en plásmidos.

La Figura 2 representa el análisis para el genoma plasmídico (aunque muy inferior en cuanto al tamaño y la cantidad de secuencias analizadas). Al igual que en el análisis anterior, las variables cualitativas tamaño del patrón y región mostraron asociación entre ellas (prueba X^2, $p < 0,01$). No obstante, se aprecian dos diferencias notables en los plásmidos con respecto a la tendencia general (cuyo peso está marcado por los genomas principales). La primera diferencia es que las frecuencias de los SSR en regiones relacionadas a genes que no codifican para

Esta obra está bajo una licencia: https://creativecomons.org/licenses/by-nc/4.0/deed.es_ES

proteínas (No-CDS) fueron prácticamente nulas y la segunda diferencia es que los SSR de tamaño 1 pb superaron tenuemente a los de 3 pb y 4 pb (29 %, 28 % y 26 % respectivamente).

La Figura 3 muestra las distribuciones y comparaciones de los valores medios de PIC para las diferentes regiones relativas a los genes, para cada uno de los tamaños de los patrones. El resultado es interesante pues existieron asociaciones significativas ($p < 0,05$) para todos los tamaños del patrón excepto para tamaño 3 pb ($p > 0,05$). Es decir, las regiones que ocupan los SSR con respecto a los genes influyen en los valores de PIC excepto para patrones de 3 pb. También se aprecia que los valores de PIC para patrones de 3 pb fueron los menores entre todos los tamaños.



p: Prueba Kruskal-Wallis

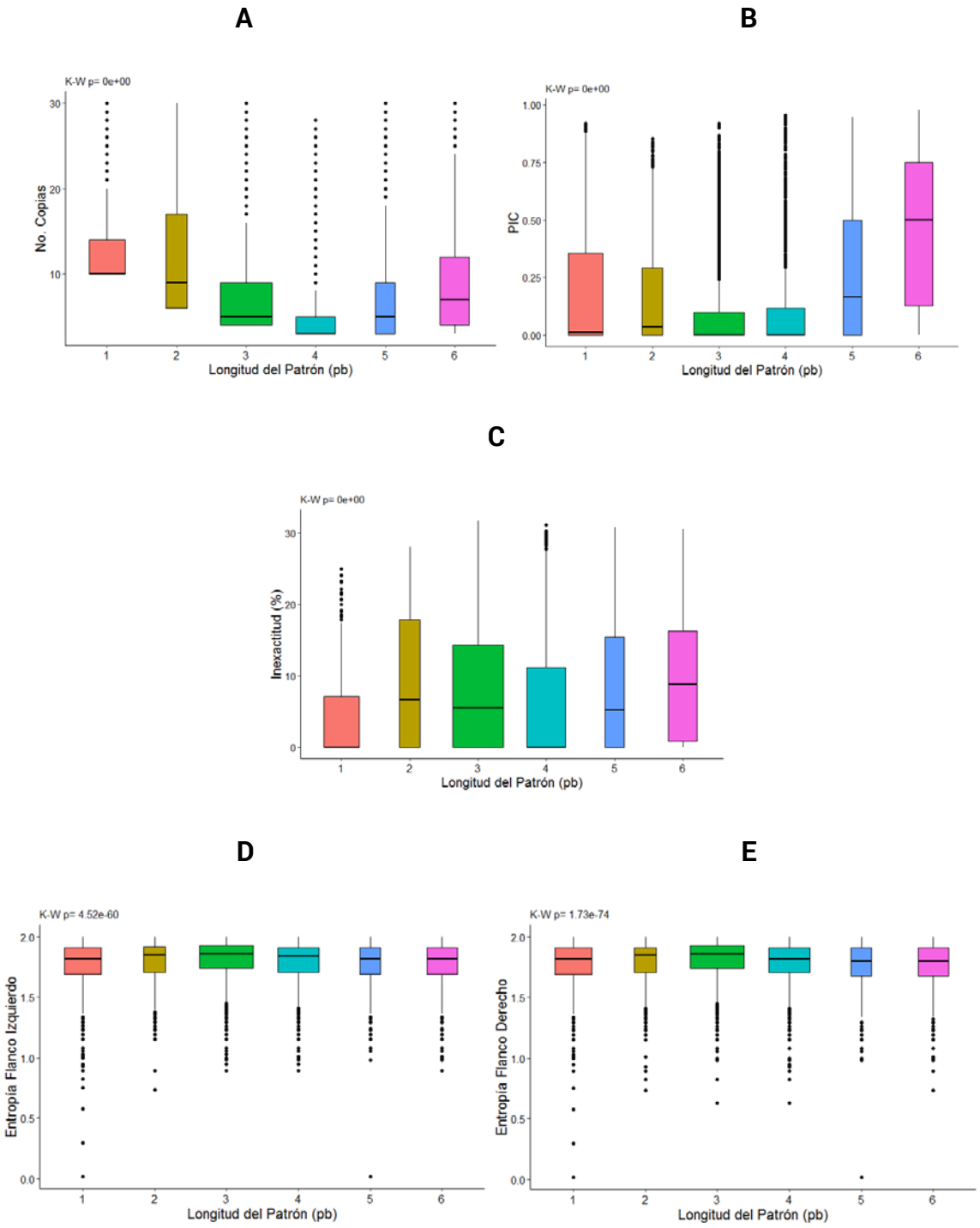
Fig. 3- Comparaciones del Contenido de Información Polimórfica (PIC) en regiones que ocupan en el genoma con relación a los genes, para cada longitud del patrón.

Al analizar las asociaciones entre los tamaños de los patrones y los valores de PIC, así como para otros rasgos propios de la estructura y la composición del

Esta obra está bajo una licencia: https://creativecomons.org/licenses/by-nc/4.0/deed.es_ES

microsatélite como son número de copias, porciento de inexactitud y entropía de los flancos 5´ y 3´ (Figura 4), vemos que las mismas fueron significativas ($p < 0,05$) por la prueba Kruskal-Wallis para comparación de grupos. De modo que, el tamaño del patrón del SSR es un factor que influye en los rasgos relacionados con la estructura y la capacidad mutacional de los SSR, como son el PIC y el número de copias, y en rasgos relacionados directamente con la composición de bases, como son la inexactitud y la entropía de los flancos. En ninguna de las distribuciones se apreció una tendencia en particular (creciente o decreciente).

Esta obra está bajo una licencia: https://creativecomons.org/licenses/by-nc/4.0/deed.es_ES



K-W: Prueba Kruskal-Wallis

Fig. 4- Distribuciones de los microsatélites según los tamaños del patrón para el número de copias (**A**), valores de PIC (**B**), porciento de inexactitud (**C**) y entropía de flancos 5´ y 3´ (**D** y **E**)

Esta obra está bajo una licencia: https://creativecomons.org/licenses/by-nc/4.0/deed.es_ES

El análisis de la composición nucleotídica se hizo necesario para complementar los resultados anteriores. En la Tabla 3 se muestran las frecuencias de cada patrón, de 1 a 3 pb, de acuerdo con su composición nucleotídica y los valores correspondientes de PIC promedio. Cada patrón en la tabla representa un conjunto de patrones que son sus permutaciones cíclicas y secuencias complementarias. La extensión de esta representación a patrones de mayor orden se vuelve prohibitivo, al menos para el análisis visual, por la cantidad de combinaciones que se generan.

Tabla 3- Frecuencias y valores promedio de Contenido de Información Polimórfica para los microsatélites con unidades repetitivas de 1-3 pb (mono-, di- y trinucleótidos)

Mononucleótido			Dinucleótido			Trinucleótido		
UR	Frec	PIC	UR	Frec	PIC	UR	Frec	PIC
A/T	0,67	0,16	AC	0,02	0,24	AAC	0,04	0,15
C/G	0,33	0,21	AG	0,04	0,31	AAG	0,08	0,12
			AT	0,43	0,12	AAT	0,16	0,13
			CG	0,51	0,07	ACC	0,09	0,08
						ACG	0,05	0,07
						ACT	0,02	0,14
						AGC	0,15	0,07
						AGG	0,01	0,10
						ATC	0,08	0,10
						CCG	0,31	0,09

UR: Unidad repetitiva (patrón). Los patrones mostrados representan subconjuntos de patrones que se complementan o constituyen permutaciones cíclicas de los mismos (p. ej. **AC** –> CA/TG/GT).

Discusión

Los organismos procariotas presentan generalmente genomas pequeños donde cerca del 90 % codifica para proteínas o ARN. Este grado de compactación de los genomas bacterianos posee una elevada proporción de regiones codificantes, *i.e.* genes que codifican para proteínas, lo cual explica en parte la preponderancia y el elevado porcentaje de los SSR de trinucleótidos. La expansión o contracción de tripletes de bases, o fragmentos de longitud múltiplos de tres, evaden las

Esta obra está bajo una licencia: https://creativecomons.org/licenses/by-nc/4.0/deed.es_ES

mutaciones de desplazamiento del marco de lectura, lo que permite que estas mutaciones sean selectivamente neutras. A esto debe sumarse también, como posible causa, el sesgo en el uso de codones y las secuencias repetitivas a nivel nucleotídico que poseen las estructuras secundarias de proteínas codificadas por estos genes.[14,18]

Es de destacar la elevada frecuencia que mostraron los SSR de tetranucleóidos (21 %) las cuales, al no expandirse o contraerse en múltiplos de tres, pueden producir mutaciones con desplazamiento del marco de lectura que cambian drásticamente la función del producto de los genes o la invalidan totalmente y, por esta misma razón, están sujetas a la presión selectiva. La variación de fase, que implica activar y desactivar la expresión de ciertos genes, también está mediada por las repeticiones de tetranucleótidos. La dispersión de estas repeticiones dentro de los genes que codifican las moléculas de virulencia permite a las bacterias adaptarse a los cambios del entorno del huésped sin aumentar la tasa general de mutación.[19]

En este estudio se hallaron especies donde los repetidos de tetranucleóidos son preponderantes, p. ej. *Bacillus cereus, Bacillus subtilis, Clostridioides difficile, Clostridium tetani, Enterococcus faecalis, Enterococcus faecium, Staphylococcus aureus, Streptococcus mitis*. Dada esta característica, las especies ejemplificadas pueden constituir modelos que permitan estudiar en particular el fenómeno de las repeticiones de tetranucleótidos en bacterias y su rol en procesos como la variación de fase la cual facilita estrategias de adaptación como la evasión inmunológica y la tolerancia al estrés ambiental.[20]

Con respecto a la localización relativa a los genes (Figura 1), es significativo el hecho de que, para mononucleótidos, los SSR tienen mucha menor presencia en CDS (12 %) que en sus regiones UTR 5′ y 3′ (21 % y 19 % respectivamente). También

Esta obra está bajo una licencia: https://creativecomons.org/licenses/by-nc/4.0/deed.es_ES

se aprecia este comportamiento para patrones de 5 pb, siendo prácticamente iguales las frecuencias. El resto de los patrones tienen similar comportamiento en cuanto a sus proporciones por regiones, con ordenamiento descendente de No-CDS, CDS y UTR. Los SSR de 1 pb, 2 pb, 4 pb y 5 pb, que no producen secuencias de longitud múltiplo de tres al expandirse o contraerse en una unidad, tienen la misma posibilidad de crear desplazamientos en el marco de lectura. No obstante, mientras menor es el patrón menos energía necesita para disociarse de su complementario y esto favorece la formación de suficiente cadena simple para la formación del lazo que produce el deslizamiento de cadena en el ADN. Si la presencia de los SSR de un nucleótido se considera como un mecanismo que aporta variación genética para una mejor adaptación del microorganismo en el hospedero, se puede conjeturar que la selección ha desfavorecido su presencia en regiones codificantes, a cambio de favorecerlas en regiones reguladoras aledañas. Esto se traduce en un cambio cuyo efecto se da a nivel de la expresión de determinadas proteínas, y no en regiones donde pueden causar mayor número de mutaciones de desplazamiento del marco de lectura o anti-sentido.

El análisis en plásmidos (Figura 2) reveló dos diferencias fundamentales con respecto a la tendencia general observada en los genomas principales: la ausencia de los SSR en zonas no codificantes o cercanas a los genes y el predominio de los SSR de 1 pb. La primera diferencia no es sorprendente pues se sabe que los plásmidos contienen en su mayoría genes codificadores de proteínas que, aunque no son esenciales para su metabolismo básico, son necesarios para la supervivencia del microorganismo. La segunda diferencia sí es novedosa, en cuanto a las frecuencias de los SSR, y apunta necesariamente hacia la función de estos genes y su composición de secuencia. Los plásmidos portan genes que en lo fundamental se relacionan con funciones como la resistencia a antibióticos, patogenicidad o la capacidad de adaptarse a nuevas condiciones. Los plásmidos

Esta obra está bajo una licencia: https://creativecomons.org/licenses/by-nc/4.0/deed.es_ES

se replican independientemente del cromosoma principal y pueden ser transferidos horizontal y verticalmente. Como consecuencia, su comportamiento evolutivo debe diferir del que posee el genoma principal y por lo tanto la composición genética de ambos ha de ser diferente.[21, 22]

Es interesante cómo se mantuvo la característica de ser más propensos los SSR en regiones UTR que en la región codificadora (41 % y 34 % versus 23 %) para el tamaño de 1 pb. En general, el comportamiento que mantiene proporciones similares entre CDS y UTR para cada tamaño de patrón, se conserva para ambos tipos de genomas. Al parecer, la característica de estos loci, es decir la ubicación con respecto a los genes, derivó en tiempos evolutivos anteriores a la propia diferenciación de ambos tipos de genomas, sin embargo, no ocurre lo mismo para sus frecuencias totales según el tamaño del patrón. De modo que podemos concluir que los SSR de tamaños de patrones 1 pb y 4 pb están sobrerrepresentados en los genomas plasmídicos y por lo tanto pudieran influir en las funciones específicas de adaptación conferidas a los genes que los componen.

La alta frecuencia de los SSR de trinucleótidos en los genomas bacterianos y a su vez su baja tasa mutacional, sin preferencias por alguna región en particular (Figura 3), son un indicador de que estas secuencias no juegan un papel distintivo como proveedoras de la variación genética. Su elevada presencia se justifica más por su rol funcional en la formación de estructuras secundarias ya sea en proteínas que están en mayor proporción, como hélices alfa que tienen marcada periodicidad de 3 pb, o en moléculas de ARN no mensajero. Para el resto de los tamaños fue significativo el hecho que los SSR tuvieron los valores menores en regiones codificantes y menores en los UTR. Esto se puede interpretar como una tendencia, viendo a estas regiones como protegidas por la selección contra los cambios de inserción y deleción de bases, las cuales afectan drásticamente el producto de sus genes. De modo que, los llamados cambios de fase que se producen en genes

Esta obra está bajo una licencia: https://creativecomons.org/licenses/by-nc/4.0/deed.es_ES

bacterianos debido a estas secuencias, reportados en regiones codificadoras, al parecer ocurren en mayor proporción en regiones reguladoras por mecanismos que no afectan directamente la estructura de las proteínas sino sus niveles de expresión.

En relación con la composición de los patrones en los SSR de 1 pb hasta 3 pb se observó una clara tendencia en la que los patrones más frecuentes no fueron los de mayor PIC (Tabla 3). Este resultado es interesante por dos razones: 1ra existe un marcado sesgo en la literatura en cuanto al análisis composicional de los patrones de los SSR debido al alto costo combinatorio de las variables a analizar y 2da hace evidente que las fuerzas evolutivas que gobiernan la variabilidad de los loci de los SSR no van en la misma dirección que aquellas que favorecen la permanencia de su patrón composicional. Otro resultado sorprendente es que los tractos de A/T que, como sabemos, forman enlaces de menor energía y son más propensos a desestabilizar la doble hélice del ADN, tuvieron menor PIC que los tractos de C/G. De modo que la fuerza de complementariedad de bases no parece ser un factor definitorio en el fenómeno molecular de deslizamiento de cadena del ADN que genera las mutaciones de expansión o contracción características en los SSR. Los patrones con mayor PIC promedio (0,31) fueron los dinucleótidos del grupo **AG**/GA/TC/CT y los de menor PIC promedio fueron los del grupo **CG**/GC. Este resultado es interesante pues los del grupo (**CG**/GC) son repetidos inversos, es decir, sus secuencias complementarias son iguales, y sin embargo se ha reportado que estos repetidos inversos son capaces de formar estructuras cruciformes en el ADN bacteriano.[23]

Esta obra está bajo una licencia: https://creativecomons.org/licenses/by-nc/4.0/deed.es_ES

Conclusiones

De los microsatélites encontrados aproximadamente la mitad de ellos se encuentran en regiones codificantes y sus flancos. Sus distribuciones por tamaño del patrón y región relativa a los genes mostraron diferencias significativas. Los tamaños del patrón influyen en el polimorfismo de sus loci y en otros rasgos como el número de copias y la entropía de los flancos. La composición nucleotídica también mostró relación con el grado de polimorfismo. Los análisis de asociación a nivel global y las frecuencias y distribuciones específicas para las 59 especies patógenas estudiadas, representan un valioso recurso bioinformático. Estos datos permiten guiar investigaciones experimentales sobre las dinámicas de los microsatélites y las fuerzas evolutivas que los moldean. De esta manera, se contribuiría a una mejor comprensión de la biología molecular subyacente a la patogénesis.

Referencias bibliográficas

1. Jonika M, Lo J, Blackmon H. Mode and Tempo of Microsatellite Evolution across 300 Million Years of Insect Evolution. Genes [Internet]. 2020 Aug 16;11(8):945. Available from: http://dx.doi.org/10.3390/genes11080945

2. Rando OJ, Verstrepen KJ. Timescales of genetic and epigenetic inheritance. Cell. 2007 Feb 23;128(4):655-68. https://doi:10.1016/j.cell.2007.01.023.

3. Peters T, Bertrand S, Björkman JT, Brandal LT, Brown DJ, Erdõsi T, *et al.* Multi-laboratory validation study of multilocus variable-number tandem repeat analysis

Esta obra está bajo una licencia: https://creativecomons.org/licenses/by-nc/4.0/deed.es_ES

(MLVA) for Salmonella enterica serovar Enteritidis, 2015. Eurosurveillance. 2017;22(9):30477.

4. Zhou K, Aertsen A, Michiels CW. The role of variable DNA tandem repeats in bacterial adaptation. FEMS Microbiology Reviews. 2014;38(1):119-41.

5. Lindstedt BA. Genotyping of selected bacterial enteropathogens in Norway. Int J Med Microbiol. 2011 Dec;301(8):648-53. https://doi:10.1016/j.ijmm.2011.09.005.

6. Gor V, Ohniwa RL, Morikawa K. No Change, No Life? What We Know about Phase Variation in Staphylococcus aureus. Microorganisms [Internet]. 2021 Jan 25;9(2):244. Available from: http://dx.doi.org/10.3390/microorganisms9020244

7. Harhay GP, Harhay DM, Bono JL, Capik SF, DeDonder KD, Apley MD, *et al.* A Computational Method to Quantify the Effects of Slipped Strand Mispairing on Bacterial Tetranucleotide Repeats. Scientific Reports. 2020;10(1):1633.

8. Phillips ZN, Tram G, Seib KL, Atack JM. Phase-variable bacterial loci: how bacteria gamble to maximise fitness in changing environments. Biochemical Society Transactions. 2019;47(4):1131-41.

9. Kassai-Jáger E, Ortutay C, Tóth G, Vellai T, Gáspári Z. Distribution and evolution of short tandem repeats in closely related bacterial genomes. Gene. 2008 Feb 29;410(1):18-25. https://doi:10.1016/j.gene.2007.11.006.

10. Mrázek J, Guo X, Shah A. Simple sequence repeats in prokaryotic genomes. Proceedings of the National Academy of Sciences. 2007;104(20):8472-7.

11. Moxon R, Bayliss C, Hood D. Bacterial contingency loci: the role of simple sequence DNA repeats in bacterial adaptation. Annu Rev Genet. 2006;40:307-33. https://doi:10.1146/annurev.genet.40.110405.090442. PMID: 17094739.

Esta obra está bajo una licencia: https://creativecomons.org/licenses/by-nc/4.0/deed.es_ES

12. Treangen TJ, Abraham AL, Touchon M, Rocha EP. Genesis, effects and fates of repeats in prokaryotic genomes. FEMS Microbiol Rev. 2009 May;33(3):539-71. https://doi:10.1111/j.1574-6976.2009.00169.x. PMID: 19396957.

13. Orsi RH, Bowen BM, Wiedmann M. Homopolymeric tracts represent a general regulatory mechanism in prokaryotes. BMC Genomics. 2010;11(1):102.

14. Lin W-H, Kussell E. Evolutionary pressures on simple sequence repeats in prokaryotic coding regions. Nucleic Acids Research. 2011;40(6):2399-413.

15. Guard J, Rivers AR, Vaughn JN, Rothrock, Jr. MJ, Oladeinde A, Shah DH. AT Homopolymer Strings in Salmonella enterica Subspecies I Contribute to Speciation and Serovar Diversity. Microorganisms [Internet]. 2021 Oct 1;9(10):2075. Available from: http://dx.doi.org/10.3390/microorganisms9102075

16. Shannon, Claude E. A Mathematical Theory of Communication. Bell System Technical Journal. 1948; 27 (3): 379–423. doi:10.1002/j.1538-7305.1948.tb01338.

17. Martínez CM, Rivero A. Methodology for in silico mining of microsatellite polymorphic loci. Revista Cubana de Informática Médica. 2019; 11:2-17.

18. Zhang H, Li D, Zhao X, Pan S, Wu X, Peng S, et al. Relatively semi-conservative replication and a folded slippage model for short tandem repeats. BMC Genomics. 2020;21(1):563.

19. Christopher D, Bayliss, Tamsin van de Ven E, Richard M. Mutations in polI but not mutSLH destabilize Haemophilus influenzae tetranucleotide repeats. The EMBO Journal, 2002. https://doi:10.1093/EMBOJ/21.6.1465.

20. Václav, Brázda., Miroslav, Fojta., Richard, P., Bowater. Structures and stability of simple DNA repeats from bacteria. Biochemical Journal, (2020). https://doi:10.1042/BCJ20190703.

Esta obra está bajo una licencia: https://creativecomons.org/licenses/by-nc/4.0/deed.es_ES

21. Tazzyman SJ, Bonhoeffer S. Why There Are No Essential Genes on Plasmids. Mol Biol Evol. 2015 Dec;32(12):3079-88. https://doi:10.1093/molbev/msu293.

22. Richard D, Ravigné V, Rieux A, Facon B, Boyer C, Boyer K, *et al.* Adaptation of genetically monomorphic bacteria: evolution of copper resistance through multiple horizontal gene transfers of complex and versatile mobile genetic elements. Mol Ecol. 2017 Apr;26(7):2131-2149. https://doi:10.1111/mec.14007.

23. Brazda V, Fojta M, Bowater RP. Structures and stability of simple DNA repeats from bacteria. Biochem J. 2020 Jan 31;477(2):325-339. https://doi:10.1042/BCJ20190703. PMID: 31967649; PMCID: PMC7015867.

Conflicto de intereses

Los autores declaran no tener ningún conflicto de intereses.

Consideraciones éticas

Los datos de las secuencias genómicas utilizadas son absolutamente públicos y de libre acceso (ftp://ftp.ncbi.nlm.nih.gov/refseq/release/bacteria/). El procesamiento y análisis de los mismos, así como los resultados expuestos, no incurren en conflictos éticos de ninguna índole.

Contribución de autoría

Conceptualización: Carlos Miguel Martínez Ortiz, Alejandro Rivero Bandinez.

Análisis Formal: Carlos Miguel Martínez Ortiz.

Investigación: Carlos Miguel Martínez Ortiz, Alejandro Rivero Bandinez, Nibaldo Hernández Mesa.

Metodología: Carlos Miguel Martínez Ortiz.

Esta obra está bajo una licencia: https://creativecomons.org/licenses/by-nc/4.0/deed.es_ES

Software: Carlos Miguel Martínez Ortiz, Alejandro Rivero Bandinez.

Supervisión: Carlos Miguel Martínez Ortiz, Nibaldo Hernández Mesa.

Redacción – borrador original: Carlos Miguel Martínez Ortiz.

Redacción – revisión y edición: Carlos Miguel Martínez Ortiz, Alejandro Rivero Bandinez, Nibaldo Hernández Mesa.

Esta obra está bajo una licencia: https://creativecomons.org/licenses/by-nc/4.0/deed.es_ES