Integración bioinformática de metodologías ómicas para el estudio de comunidades microbianas en suelos

Lucía Ortiz Rocca[1*], Marcela S. Montecchia [1,2*], Jorge Chalco Vera[3], Martín M. Acreche[3], Olga S. Correa[1], Marcelo A. Soria[1,2]

[1] Universidad de Buenos Aires. Facultad de Agronomía. Cátedra de Microbiología Agrícola. Av. San Martín 4453 (CP1417)
{ortizrocca, mmontecc, correa, soria}@agro.uba.ar
[2] Instituto de Biociencias Agrícolas y Ambientales. INBA (UBA-CONICET) Av. San Martín 4453 (CP1417). C.A.B.A. Argentina
[3] INTA Estación Experimental Agropecuaria Salta-CONICET. Ruta Nac. 68, Km. 172, CP 4403, Cerrillos, Salta, Argentina.
chalco006@gmail.com, acreche.martin@inta.gob.ar

* Ambas autoras contribuyeron igualmente a este trabajo

Resumen. Las comunidades microbianas son esenciales en la dinámica y la sostenibilidad de los suelos. Presentamos un estudio bioinformático integrado de las respuestas de las comunidades bacterianas del suelo a diferentes manejos agronómicos en caña de azúcar en un experimento de larga duración en la EEA INTA Famaillá. Se integraron tres fuentes de datos: abundancia de genes mayormente relacionados con el flujo de gases de efecto invernadero, secuenciación masiva del gen 16S rRNA y secuenciación masiva de ADN total. Detectamos diferencias en las abundancias de algunos genes funcionales. Luego usando qiime2 y otras herramientas analizamos el comportamiento de 12,000 variantes de secuencias derivados de 2.1 millones de secuencias *paired-end* del gen 16S, y encontramos algunas diferencias cuantitativas en la composición de las comunidades. Con los datos de secuenciación masiva de ADN total (32 Giga bases) ensamblamos *contigs* (Megahit) y los agrupamos en *bins* (MaxBin2), obteniendo 374 genomas ensamblados de metagenomas (MAGs). Anotamos los MAGs de mejor calidad con EggNog. El estudio integrado y secuencial nos permitió mantener bajos los costos de la etapa más cara (secuenciación de ADN total) y nos permitió acelerar el descubrimiento de los genomas de interés para los objetivos del estudio.

Palabras clave: metagenómica, secuenciación 16S rRNA, secuenciación ADN total, microbiota del suelo

1. Introducción

El suelo es uno de los hábitats de mayor diversidad biológica de la Tierra y un proveedor natural de bienes y servicios para los ecosistemas [1] y existe abundante evidencia de que la multifuncionalidad de los ecosistemas terrestres depende de la diversidad microbiana de los suelos [2]. Las comunidades microbianas del suelo son muy diversas y muchos de sus integrantes no son cultivables, y sólo se los puede evidenciar a través de la extracción y secuenciación de su ADN. El uso agrícola modifica

las características físicas y químicas del suelo, afectando negativamente a la biota del suelo [2,3]. Las variantes más conservacionistas de manejo agrícola buscan reducir los impactos negativos de la agricultura convencional sobre la biodiversidad y la biota del suelo. El tipo de manejo puede afectar también la dinámica de emisión/mitigación de los gases de efecto invernadero en el suelo. Sin embargo, los efectos observados aún no son claros: de 100 estudios comparativos entre agricultura conservacionista y convencional solamente la mitad mostraron mayores secuestros de carbono en sistemas con mínima labranza comparada con la labranza convencional. Algunos estudios informan mayores emisiones de los gases de efecto invernadero (GEI), óxido nitroso y metano, con manejo conservacionista, mientras que otros encuentran menores emisiones [4].

Las técnicas de secuenciación masiva de ADN son un conjunto de metodologías de biología molecular con las que es posible extraer ADN directamente de una muestra compleja, como es el suelo, y secuenciar partes específicas (marcadores) o la totalidad del ADN obtenido. La metagenómica es la combinación de estas técnicas de biología molecular con técnicas bioinformáticas y estadísticas que permiten estudiar las comunidades microbianas con un enfoque ecológico similar al que se ha usado exitosamente para comunidades de plantas y animales [5]. En este trabajo proponemos un abordaje del análisis de la biodiversidad microbiana integrando tres tipos de fuente de datos metagenómicos: abundancia de genes mayormente relacionados con el flujo de gases de efecto invernadero (GEI), secuenciación masiva de una región hipervariable del gen que codifica para el 16S rRNA y secuenciación masiva de ADN total [6]. Por un lado, el dosaje mediante PCR en tiempo real de la abundancia de genes clave en la dinámica de los GEI y la secuenciación masiva del 16S rRNA permiten comprender mejor las características de aquellas bacterias involucradas con los cambios funcionales y de composición taxonómica de la comunidad. Por el otro, mediante la secuenciación masiva de ADN es posible obtener ensamblados aproximados de genomas de los miembros más abundantes de la comunidad, también conocidos como genomas ensamblados de metagenomas (MAGs de sus siglas en inglés Metagenome assembled genomes) [7].

Los objetivos de este trabajo son: 1) describir la composición de las comunidades microbianas del suelo presentes en lotes con cultivo de caña de azúcar bajo diferentes métodos de implantación y cosecha, con diferentes impactos sobre el suelo, a través del análisis bioinformático y estadístico derivados de la secuenciación de regiones hipervariables del gen 16s rRNA. 2) Analizar los datos de abundancia de varios genes clave en la dinámica de los gases de efecto invernadero (*nirK, nirS, nosZ, amoA* y *pmoA*, también se incluyeron los genes *metT1* y *metT2*, que no son estrictamente genes funcionales, pero son buenos indicadores de la presencia de arqueobacterias metanogénicas). 3) Usar datos de secuenciación de ADN total para ensamblar, agrupar los *contigs* resultantes (*binning*) y anotar física y funcionalmente MAGs. 4) Integrar los datos de las diferentes metodologías para obtener una visión más profunda de las comunidades microbianas, sobre todo bacterianas.

2. Cuerpo del trabajo

2.1 Materiales y métodos

Sitio de estudio, diseño experimental y muestreo de suelos. El campo experimental se localiza en la EEA Famaillá, INTA, provincia de Tucumán (27°00'52.6"S, 65°22'46.5"O). La temperatura media es 13-19ºC, con una gran amplitud térmica anual, y las precipitaciones anuales superan los 900 mm, con humedad relativa media 85% [8]. Se probaron dos sistemas de cultivo: manejo convencional y labranza en franjas sin remoción profunda del suelo en los sitios de tránsito. En el manejo convencional, la implantación del cultivo se realizó en dos pasadas con rastra excéntrica y dos labranzas profundas, removiendo toda la superficie del lote al inicio del ciclo del cultivo. En la labranza en franjas, al implantar el cultivo se efectuó una labranza profunda en los sitios donde luego se colocó la caña semilla. Esta técnica requirió un prototipo de máquina diseñada y desarrollada en el Laboratorio de Terramecánica e Implantación de Cultivos (IIR-CIA-CNIA-INTA ex Castelar). Las parcelas fueron cosechadas en parte de forma convencional y en parte con una cosecha liviana. Los tratamientos combinados de siembra y cosecha se repitieron tres veces en un diseño en bloques completamente aleatorizado con parcelas divididas (Tabla 1).

Tabla 1 . Disposición de los tratamientos de siembra y cosecha. LC: labranza convencional, LF: labranza en franjas, CC: cosecha convencional, CL: cosecha liviana.

Disposición de los tratamientos			
Tratamiento	**Siembra**	**Cosecha**	**Tratamiento**
1	Labranza convencional	Cosecha convencional	LC + CC
2	Labranza convencional	Cosecha liviana	LC + CL
3	Labranza en franjas	Cosecha convencional	LF + CC
4	Labranza en franjas	Cosecha liviana	LF + CL

*Información preliminar del sitio. E*n la región donde se ubica el experimento el tamaño de las comunidades bacterianas varía drásticamente a lo largo del año debido al período de sequía invernal. Los datos analizados son muestras tomadas en marzo, cuando la comunidad bacteriana alcanza su tamaño máximo. Se contaba con alguna información previa con respecto a la dinámica de los gases de efecto invernadero en marzo. No se observaron variaciones en las emisiones de CO_2 asociadas a los manejos agronómicos; no hubo producción detectable de metano en marzo, posiblemente debido a una mayor actividad metanotrófica; mientras que se detectó una mayor emisión de NO_2, en las parcelas con cosecha convencional, aunque poco significativa. También se contaba con información relevante con respecto al nitrógeno en el suelo: no existieron cambios en el amonio del suelo asociados a ninguno de los manejos; sin embargo, se observaron niveles mayores y significativos de nitrato en las parcelas con labranza convencional.

Secuenciación de la región V3-V4 del gen 16S rRNA. Se tomaron muestras de suelo de las diferentes parcelas a una profundidad de 0-10 cm y se extrajo el ADN con un kit comercial (MOBIO PowerSoil DNA Isolation Kit) siguiendo las instruc-

ciones del fabricante. La región hipervariable V3-V4 del gen que codifica para el 16S rRNA se amplificó por PCR con los primers 341F (5'-CCTAYGGGRBGCASCAG-3') y 806R (5'-GGACTACNNGGGTATCTAAT-3'). Esta elección de primers nos permitió detectar y cuantificar un amplio rango de Eubacterias y también una proporción grande de los miembros del filo Euryarchaeota, que concentra las Arqueobacterias metanogénicas. Los productos de amplificación se secuenciaron con un protocolo *paired-end* con fragmentos de 150-pb en Novogene (California, Estados Unidos) en un secuenciador illumina HiSeq.

La mayor parte del procesamiento de los datos se realizó con qiime2 [9]. Brevemente, se revisó la calidad de las secuencias y se decidió remover siete nucleótidos del extremo 5' de ambas secuencias del par *paired-end*, luego se fusionaron y se usó el algoritmo DADA2 para eliminar ruido y extraer variantes de secuencia amplificadas únicas (ASVs por sus siglas en inglés) Luego todas las muestras se rarefaccionaron a 41,000 individuos. Para la clasificación taxonómica de los ASVs se usó la base de datos SILVA [10] y un clasificador Naïve-Baye. Las tablas de abundancias rarefaccionadas se usaron para análisis de diversidades *alfa* y *beta*.

Determinación de la abundancia de genes funcionales por PCR en tiempo real. Se utilizó el método descrito por Lammel et al. [11] con algunas modificaciones para determinar la abundancia de los genes que codifican para las enzimas nitrito reductasa K y S (*nirK*, *nirS*), reductasa del óxido nitroso (*nosZ*), amonio oxidasa A (*amoA*), reductasa de la metil coenzima M (*mcrA*), metano oxygensa (*pmoA*), y metano mono oxigenasa T1 y T2 (*metT1* y *metT2*), y la del gen que codifica para el rRNA 16S.

Secuenciación masiva de ADN total.

Secuenciación y muestreo. A partir del ADN extraído para los pasos anteriores se construyeron muestras compuestas, una para cada uno de los cuatro tratamientos (dos modalidades siembra x dos modalidades de cosecha). La secuenciación se realizó en un equipo de secuenciación Illumina HiSeq (Novogene), con la modalidad "paired-end" de 150-pb de longitud. Una vez obtenidos los datos de secuenciación, se extrajo una muestra de 500.000 reads de cada archivo fastq con el método sample del programa seqtk para realizar un control de calidad con FASTQC. Observado los resultados del control de calidad, llegamos a la conclusión de que era conveniente cortar los primeros cuatro nucleótidos del extremo 5´, debido a que tenían una calidad muy baja, quedando los reads con un longitud de 146-pb. Una vez realizado esto para la totalidad de los reads de cada muestra, la composición de la muestra quedó compuesta de la siguiente manera:

Tabla 2. Detalle de las muestras secuenciadas

Detalle de las muestras secuenciadas				
Tratamiento	**Cantidad de *reads paired- end***	**Longitud de los *reads* (pb)**	**Posición**	**Tamaño del archivo (GB)**
1 → LC-CC	123.556.226	146	*forward*	4.1
			reverse	4.2
2 → LC-CL	138.316.218	146	*forward*	4.6
			reverse	4.7
3 → LF-CC	126.866.762	146	*forward*	4.3
			reverse	4.3
4 → LF-CL	124.243.274	146	*forward*	4.2
			reverse	4.2

Ensamblado. Una vez establecida la muestra, se procedió a realizar un ensamblado con el programa MEGAHIT [12] Debido a los requerimientos de hardware de este software combinado con el tamaño de nuestro conjunto de secuencias, el procesamiento se realizó en uno de los servidores del Centro de Computación de Alto Rendimiento (CECAR) de la Facultad de Ciencias Exactas y Naturales de la Universidad de Buenos Aires (FCEyN, UBA) [13] donde utilizamos un servidor con una CPU AMD Opteron 6320 x4, con 512 GB de memoria RAM (ECC DDR3 1600 MHz) y 2TB de almacenamiento en disco. Para el ensamblado, todos los reads fueron concatenados en dos archivos: uno forward y otro reverse y como longitud mínima por *contig*, se estableció 1000-pb. Además, optamos el preset “meta-large”, que está recomendado por la documentación del programa para comunidades de suelo y la opción “no mercy” para que no restrinja el uso de la RAM. La corrida requirió de ocho días. El control de calidad del ensamblado se realizó con MetaQUAST [14].

Binning. Una vez obtenido el conjunto de *contigs*, se agruparon en *bins* con el programa MaxBin2 [15]. Este paso se realizó en una máquina local equipada con procesador Intel Core i7 (i7-4770, 3.40GHz, 8 threads). El control de calidad del *binning* fue realizado con el programa CheckM[16]. Este paso es fundamental como antesala de la anotación, ya que de aquí se obtienen los datos de calidad y contaminación del *binning*. Como punto de corte, se seleccionó un 70% de completitud de los *bins* para realizar la anotación

Anotación. La anotación de los *contigs* se realizó con eggNOG-mapper v2 [17] sobre los *bins* con menos del 70% de completitud. Este programa produce anotaciones físicas (localización de posibles genes) y funcionales para los genes encontrados de diferente tipo (similitud a genes conocidos, presencia de dominios, anotaciones KEGG, GO, etc.). Como salida, ofrece una tabla de anotaciones en un archivo de texto con tabulaciones como delimitador de campos.

2.2 Resultados

Abundancia de genes relacionados con el flujo de gases de efecto invernadero. El gen *nirK* (nitrito reductasa K) mostró una abundancia significativamente menor en las parcelas con siembra en franjas (P = 0.0255). El gen *nosZ* mostró abundancias mayores asociado a la cosecha convencional (P = 0.0428). Los otros genes funcionales (*amoA, mcrA, nirS, pmoA*) no mostraron efectos significativos para ninguno de las dos variables de manejo. En cuanto a los genes usados para cuantificar abundancias de grupos taxonómicos, sólo para *met T1* se consiguió un modelo con buen ajuste y que mostró mayores abundancias significativas en la cosecha convencional.

Secuenciación masiva del gen 16S rRNA. El conjunto de datos inicial contenía 2.053.203 de pares de secuencias *paired-end*. Después del filtrado, eliminación de ruido y consolidación en secuencias únicas con DADA2 se generó una tabla de abundancia por muestra para 12,010 *features*. La clasificación taxonómica de estos *features* mostró que a nivel taxonómico los filos más abundantes fueron *Proteobacteria, Acidobacteria, Firmicutes, Bacteroidetes, Actinobacteria, Gemmatimonadetes* y *Verrucomicrobia.*

Para llevar a cabo análisis de diversidad *alfa* y *beta* sobre comunidades uniformes, las abundancias de todas las muestras se rarefaccionaron a 41.000 secuencias. Para el análisis de diversidad *alfa* se compararon las riquezas observadas, la uniformidad (*eveness*) y la medida de diversidad de Shannon y se evaluó su significancia estadística con el método robusto de Kruskall-Wallis. Solo se encontraron diferencias significativas para la riqueza observada, que resultó mayor en las parcelas con labranza en franjas (P = 0.037). Para el estudio de la diversidad *beta* se construyeron matrices de distancia para los datos de composición de las comunidades a nivel de *features* con las métricas Bray-Curtis y Unifrac. En este caso también se observaron diferencias significativas, usando el test de permutaciones Adonis, para el tipo de labranza (P = 0.013), pero solo para la matriz de distancia Unifrac (Fig. 1).

Como una aproximación extra al análisis funcional con los datos de secuenciación del rRNA 16S se utilizó Picrust2 con su implementación en qiime2 para inferir anotaciones funcionales. Se derivaron datos de abundancia para 423 vías metabólicas, pero ninguna mostró diferencias asociadas a los manejos agronómicos (datos no mostrados).

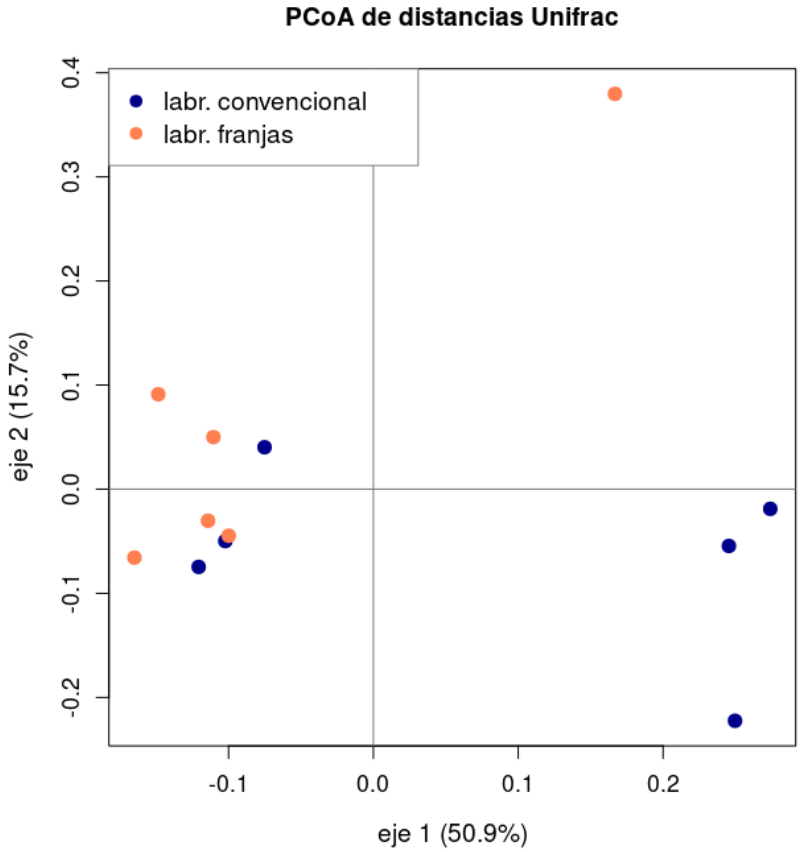

Fig. 1. Análisis de componentes principales derivado de la matriz de distancias Unifrac entre comunidades. Las diferencias entre tipos de labranza son significativas con un P = 0.037.

Debido a que el factor experimental más importante en este análisis resultó ser el tipo de labranza, se buscaron géneros que respondieran en forma diferencial a este factor. Se agruparon los *features* a nivel taxonómico de género y se analizaron cambios en su abundancia con tres métodos diferentes - limma-voom, DESeq2 y corncob- usando los paquetes correspondientes desarrollados para R-Bioconductor. Se encontraron doce géneros diferenciales para el tipo de labranza, pero solo dos de ellos, identificados como *Flavobacterium* y *Chthoniobacter*, fueron detectados por más de un método. En promedio, Flavobacterium es 4.1 veces más abundante en las parcelas bajo labranza convencional, mientras que por el contrario, Chthoniobacter es 2.4 veces más abundante en las parcelas con labranza en franjas.

Secuenciación masiva de ADN total. El análisis de diversidad *beta* basado en la distancia Unifrac de los géneros derivados de los *features* basados en las secuencias del rRNA 16S indicó que las comunidades se estructuraban diferencialmente según el tipo de labranza. Sin embargo, se encontraron pocas diferencias a nivel de abundancia de géneros y de actividades funcionales que puedan considerarse claves para explicar las variaciones en la estructura de las comunidades. Es posible que las diferencias entre comunidades se expliquen, por la sumatoria de pequeñas variaciones en un grupo de taxones presentes en todas las parcelas, entre los que se podrían ubicar aquellos más abundantes. Para conocer mejor las características de los miembros más abundantes de las comunidades, reconstruir sus genomas, anotarlos y obtener información preliminar sobre posibles marcadores moleculares para realizar seguimientos más precisos a campo, realizamos una secuenciación de ADN total a una gran profundidad.

Ensamblado. El ensamblado se realizó con Megahit y produjo un archivo FASTA con 1.041.898 *contigs* que fueron sometidos a una prueba de calidad con MetaQUAST (Tabla 3). A continuación se muestra una tabla con los valores arrojados más relevantes.

Tabla 3 . Resultados del análisis de calidad de los *contigs* en MetaQUAST. N50 es la longitud de *contig* que marca la mitad de la longitud de la muestra, es decir que la mitad de la muestra tiene una longitud de al menos 1697pb. El N90 nos indica entonces, que el 10% de los *contigs* tienen una longitud de 1081pb o más. Un L50 indica la cantidad de *contigs* que hay de igual o mayor tamaño que el N50, es decir que 299.913 *contigs*, tienen 1697pb de longitud, o más. Análogamente, el L90 nos indica que 863.886 *contigs* tienen una longitud de 1081pb o más.

Resultados del análisis de calidad de los *contigs* en MetaQUAST	
Cantidad total de *contigs*	1.041.898 (100%)
Cantidad de *contigs* de longitud <5.000pb	1.013.368 (97.26%)
Cantidad de *contigs* de longitud >=5.000pb y <10.000	21.430 (2.06%)
Cantidad de *contigs* de longitud >=10.000pb	7.100 (0.68%)
Longitud del *contig* más largo	120.370pb
N50	1.697pb
N90	1.081pb
L50	299.913 *contigs*
L90	863.886 *contigs*

Binning. El *binning* fue realizado con MaxBin2 y después de realizar el control de calidad con CheckMA se retuvieron aquellos *bins* con más del 70% de completitud (Tabla 4).

Tabla 4. Resultados del *binning*

Resultados del *binning*	
Cantidad de *bins* obtenidos	374 (100%)
Cantidad de *bins* con más del 70% de completitud	30 (8.06%)

Anotación. Los *contigs* de cada uno de los *bins* se anotaron con EggNog para obtener en primer lugar una anotación física, es decir, las coordenadas de inicio y fin de los marcos de lectura abierto que pueden corresponder con genes, y luego se anotaron funcionalmente. En promedio el 71.7% de los marcos de lectura abierto descubiertos recibieron anotaciones funcionales. A continuación se presenta una figura con un histograma por *bin* de la cantidad de genes encontrados y anotados.

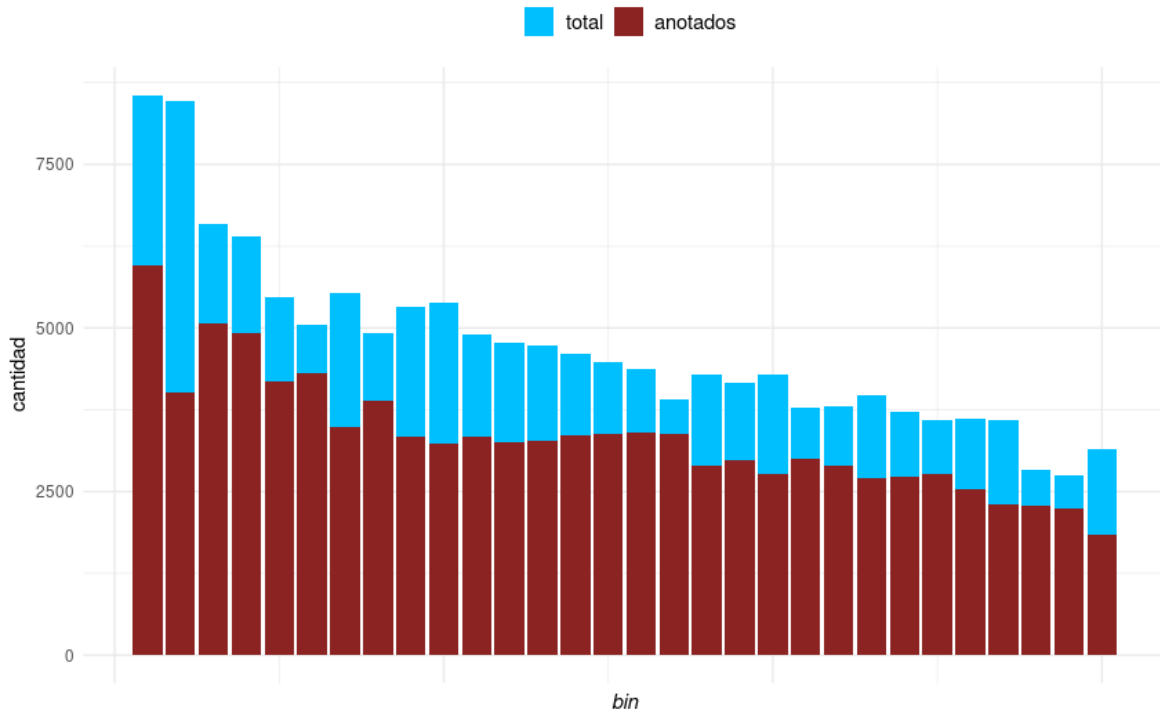

Fig. 2. Cantidad de genes totales encontrados y con coordenadas asignadas por *bin*, y cantidad de genes con al menos una anotación funcional.

2.3 Discusión

Los mayores cambios en la comunidad revelados por el estudio de los amplicones del gen 16S rRNA y por los análisis funcionales, sugieren que la comunidad de procariotas responde a los diferentes manejos agronómicos. Aquí debe destacarse que la secuenciación del gen rRNA 16S fue de una profundidad suficientemente alta como para descubrir cerca de 12,000 *features*, que se agruparon en 477 géneros, 276 de los

cuales tuvieron una abundancia suficientemente buena para realizar una búsqueda de taxones diferenciales. Los costos de los estudios de amplicones bajó en forma marcada. Consideramos que en un estudio como el que presentamos esto debe aprovecharse para producir muestras secuenciadas a gran profundidad, en lugar de muchas muestras con baja. En línea con esto, la profundidad de secuenciación empleada también permitió hacer un estudio de funcionalidades inferidas con Picrust2 con altas abundancias para la mayoría de las vías metabólicas inferidas. Complementando estos hallazgos, el estudio de genes individuales con funciones clave en la producción o mitigación de gases de efecto invernadero por PCR cuantitativa produjo datos de alta precisión mostró, en coincidencia con el análisis de amplicones, que existen pocos de esos genes que muestren abundancias que varien con los manejos agronómicos.

Si bien existen recomendaciones en cuanto a darle menos importancia a las técnicas basadas en secuenciación de amplicones, es importante resaltar que con las profundidades que se alcanzan actualmente es posible analizar un gran número de entidades taxonómicas y que para alcanzar una extensión similar sólo con secuenciación de ADN total implicaría elevados costos tanto de secuenciación como de potencia computacional. En nuestro caso, la integración de métodos nos permitió dejar como objetivo principal de la secuenciación masiva de ADN total el ensamblado de MAGs y también nos sirve de guía para orientar la búsqueda de los MAGs y *contigs* de interés entre la gran cantidad de datos generada. Caso contrario, y para contar con información que se pueda analizar con técnicas estadísticas apropiadas, deberíamos haber secuenciado por separado cada una de las parcelas triplicadas para las cuatro combinaciones de modalidad de siembra y cosecha.

El análisis de los resultados del ensamblado revela que la gran mayoría de los *contigs* son de menos de 5000-pb de longitud, nosotros habíamos configurado MEGAHIT para retener *contigs* con un largo mínimo de 1000-pb. Estos resultados no son los óptimos, pero al considerar el grado de compactación de los genomas procarióticos, con espacios intergénicos cortos y una fracción importante de ADN codificante, es evidente que se puede extraer algo de información útil aún de *contigs* de alrededor de 5000-pb. El uso creciente de metodologías de secuenciación de cadena larga seguramente va a significar una mejora de este problema. Algo parecido sucede con el producto final del *binning:* sólo el 8% de todos los *bins* son los que nos sirvieron porque pasaron un tamaño umbral y tenían valores de completitud altos y de contaminación baja.

Como mencionamos en la descripción inicial del sitio, al momento de tomar las muestras de suelo para este estudio las emisiones de N_2O no mostraban cambios significativos asociados al tipo de labranza; sin embargo, la mayor abundancia de genes nirk en la labranza en franjas podría permitir, a mediano plazo, mitigar las emisiones de N_2O. Además, en esta descripción inicial, las emisiones de metano no eran detectables. Si bien mencionamos como posible causa, el balanceo entre el proceso de metagenogénesis y metanotrofía pero también es necesario destacar que en el análisis de MAGs no encontramos genes de metanogénesis en aquellos *bins* que se corresponden a arqueobacterias. Sugiriendo una explicación alternativa que es que las arqueobacterias metanogénicas reducen su abundancia durante el otoño, porque en otros momentos del año sí se detectó emisión de metano.

Entre los MAGs seleccionados por calidad no encontramos ninguno que se puedan asignar a los géneros *Flavobacterium* y *Chthoniobacter*, debido posiblemente a que no se encuentran entre los géneros más abundantes. Aquí, nuevamente, se destaca la importancia de utilizar más de una técnica. El secuenciado de amplicones del rRNA 16S tiene sesgos y se trabaja con menores profundidades de secuenciación; pero al mismo tiempo, se puede hacer al hacer foco sobre una región particular que resulta especialmente apta para los estudios de diversidad, se alcanza una capacidad de relevamiento de las comunidades que no es tan directa con otras técnicas.

Las *pipelines* utilizadas en este trabajo cuentan con numerosos pasos, en general, en cada uno se pueden tomar varias decisiones en cuanto a su configuración de corrida. Estas decisiones pueden impactar con mayor o menor intensidad sobre el producto final. Una recomendación obvia en este punto es que con frecuencia las opciones por *default* no son las más apropiadas, por lo que es esencial el estudio atento de la documentación. Además, cada paso de la *pipeline* se debe validar con las herramientas adecuadas para el control de la calidad de los resultados de ese paso.

3. Conclusiones

Al ser albergada una enorme cantidad de biomasa microbiana en el suelo, el uso de técnicas metagenómicas para su abordaje y estudio es de mucha utilidad. Pues hacen posible la extracción del ADN de manera directa de una muestra compleja, como en el caso de este trabajo, del suelo. En este trabajo resaltamos la importancia de realizar un análisis integrado de las tres metodologías ómicas ya que realizar primero un análisis de la secuenciación de la región V3-V4 del gen 16S rRNA, nos permitió bajar los costos de la secuenciación de ADN total.

Es posible que con las metodologías de secuenciación de cadena larga nos permitan obtener resultados más significativos en el ensamblado y, consecuentemente, en los MAGs. De todos modos, en este trabajo pudimos obtener 30 MAGs con muy baja contaminación y alta completitud, lo que nos indica que pudiendo acceder a una secuenciación por Illumina también obtenemos resultados concluyentes. Para ello, hacemos hincapié en la importancia de realizar un análisis de calidad en todos los pasos para así poder tomar las decisiones más atinadas para el análisis metodológico.

4. Bibliografía

1. Powlson DS, Gregory PJ, Whalley WR, Quinton JN, Hopkins DW, Whitmore AP, Hirsch PR & Goulding, K W (2011). Soil management in relation to sustainable agriculture and ecosystem services. Food policy, 36, S72-S87.

2. Bender SF, Wagg C, van der Heijden MG (2016). An underground revolution: biodiversity and soil ecological engineering for agricultural sustainability. Trends Ecol Evol 31: 440-452.

3. Ding GC, Piceno YM, Heuer H, Weinert N, Dohrmann AB, et al (2013) Changes of soil bacterial diversity as a consequence of agricultural land use in a semi-arid ecosystem. PLoS ONE 8(3): e59497.

4. Palm C, Blanco-Canqui H, DeClerck F, Gatere L, Grace P (2014). Conservation agriculture and ecosystem services: An overview. Agric Ecosyst Environ 187: 87-105.

5. Vestergaard G., Schulz S., Schöler A., & Schloter M. (2017). Making big data smart—how to use metagenomics to understand soil quality. Biology and Fertility of Soils, 53, 479-484.

6. Semenov, M. V. (2021). Metabarcoding and metagenomics in soil ecology research: achievements, challenges, and prospects. Biology Bulletin Reviews, 11, 40-53.

7. Setubal, J.C. (2021). Metagenome-assembled genomes: concepts, analogies, and challenges. Biophysical Reviews, 13, 905 – 909.

8. Zeman, EA (2019). Resumen agrometeorológico, Abril 2019. Observatorio agrometeorológico INTA EEA Famaillá.

9. Bolyen E, Rideout JR, Dillon MR, Bokulich NA, Abnet CC, Al-Ghalith GA, Alexander H, Alm EJ, Arumugam M, Asnicar F, Bai Y, Bisanz JE, Bittinger K, Brejnrod A, Brislawn CJ, Brown CT, Callahan BJ, Caraballo-Rodríguez AM, Chase J, Cope EK, Da Silva R, Diener C, Dorrestein PC, Douglas GM, Durall DM, Duvallet C, Edwardson CF, Ernst M, Estaki M, Fouquier J, Gauglitz JM, Gibbons SM, Gibson DL, Gonzalez A, Gorlick K, Guo J, Hillmann B, Holmes S, Holste H, Huttenhower C, Huttley GA, Janssen S, Jarmusch AK, Jiang L, Kaehler BD, Kang KB, Keefe CR, Keim P, Kelley ST, Knights D, Koester I, Kosciolek T, Kreps J, Langille MGI, Lee J, Ley R, Liu YX, Loftfield E, Lozupone C, Maher M, Marotz C, Martin BD, McDonald D, McIver LJ, Melnik AV, Metcalf JL, Morgan SC, Morton JT, Naimey AT, Navas-Molina JA, Nothias LF, Orchanian SB, Pearson T, Peoples SL, Petras D, Preuss ML, Pruesse E, Rasmussen LB, Rivers A, Robeson MS, Rosenthal P, Segata N, Shaffer M, Shiffer A, Sinha R, Song SJ, Spear JR, Swafford AD, Thompson LR, Torres PJ, Trinh P, Tripathi A, Turnbaugh PJ, Ul-Hasan S, van der Hooft JJJ, Vargas F, Vázquez-Baeza Y, Vogtmann E, von Hippel M, Walters W, Wan Y, Wang M, Warren J, Weber KC, Williamson CHD, Willis AD, Xu ZZ, Zaneveld JR, Zhang Y, Zhu Q, Knight R, and Caporaso JG. 2019. Reproducible, interactive, scalable and extensible microbiome data science using QIIME 2. Nature Biotechnology 37: 852–857.

10. Quast C, Pruesse E, Yilmaz P, Gerken J, Schweer T, Yarza P, Peplies J, Glöckner FO (2013) The SILVA ribosomal RNA gene database project: improved data processing and web-based tools. Nucl. Acids Res. 41 (D1): D590-D596.

11. Lammel, D. R., Feigl, B. J., Cerri, C. C., & Nüsslein, K. (2015). Specific microbial gene abundances and soil parameters contribute to C, N, and greenhouse gas process rates after land use change in Southern Amazonian Soils. Frontiers in microbiology, 6, 1057.

12. Li, D., Liu, C-M., Luo, R., Sadakane, K., and Lam, T-W., (2015) MEGAHIT: An ultra-fast single-node solution for large and complex metagenomics assembly via succinct de Bruijn graph. Bioinformatics.

13. Centro de Estudios Científicos y Aplicados de la Realidad (CECAR). (n.d.). CECAR. Facultad de Ciencias Exactas y Naturales. Universidad de Buenos Aires. https://cecar.fcen.uba.ar/

14. Mikheenko, A., Saveliev, V., & Gurevich, A. (2016). MetaQUAST: evaluation of metagenome assemblies. Bioinformatics, 32(7), 1088-1090.

15. Wu, Y. W., Tang, Y. H., Tringe, S. G., Simmons, B. A., & Singer, S. W. (2014). MaxBin: an automated binning method to recover individual genomes from metagenomes using an expectation-maximization algorithm. Microbiome, 2, 1-18.

16. Parks, D. H., Imelfort, M., Skennerton, C. T., Hugenholtz, P., & Tyson, G. W. (2015). CheckM: assessing the quality of microbial genomes recovered from isolates, single cells, and metagenomes. Genome research, 25(7), 1043-1055.

17. Cantalapiedra, C. P., Hernández-Plaza, A., Letunic, I., Bork, P., & Huerta-Cepas, J. (2021). eggNOG-mapper v2: functional annotation, orthology assignments, and domain prediction at the metagenomic scale. Molecular biology and evolution, 38(12), 5825-5829.