¿QUÉ RESPIRAMOS EN LOS HOSPITALES?

TFM

Graciela de Diego Castilla
Master Data Science Kschool

Graciela De Diego
Graciela.dediego@gmail.com

INTRODUCCIÓN

El trasplante de piel es una de las técnicas aplicadas a los grandes quemados. Estos trasplantes no funcionan igual que los trasplantes de cualquier otro órgano: pueden rechazarse más habitualmente de lo que parece, no solo por incompatibilidades inmunológicas complejas, sino también por mediación de infecciones nosocomiales. [1]

La infección con compromiso sistémico es la principal causa de muerte en el paciente quemado. Este paciente es muy susceptible porque ha perdido su primera línea de defensa contra la invasión de los microorganismos, que es la piel. Además, nos encontramos con un paciente totalmente inmunosuprimido y rodeado de gérmenes, no solamente por sus propios microorganismos, sino también los de su entorno. Por lo tanto, no es una sorpresa que el paciente quemado sea extremadamente susceptible a la infección. [2]

La unidad de quemados encargó al laboratorio de evolución microbiana del Centro de Astrobiología que, con sus técnicas habituales de muestreo aerobiológicas ambientales de campo, organizara un muestreo por la unidad hospitalaria y de UCI de la unidad de quemados, para comprobar si se detecta, con las técnicas de muestreo de campo, una serie de microorganismos que causan estas infecciones. Lógicamente, esta unidad tiene sus propias medidas y controles de calidad que también monitorizan este tipo de microorganismos, pero decidieron realizar un estudio de muestreo desde otro punto de vista como fortalecimiento de su propio control, que ya de por si es más que correcto. El interés por este muestreo es que el laboratorio de evolución microbiana es especialista en aerobiología y la dispersión aérea podría ser una de las causas de estas infecciones.

Los datos que se han obtenido en este muestreo, que se llevó a cabo en 2017, son de muestras tanto superficiales como de muestras aéreas. En estas muestras la metodología que se aplica es el monitoreo de una serie de parámetros mientras se recogen las muestras: temperatura, humedad relativa, partículas aéreas. Posteriormente, en el laboratorio se extrae ADN y se secuencia por Ilumina, revelando qué microorganismos encontramos en las muestras.

El trabajo de Data Science que se aplica en este estudio es muy interesante para el laboratorio de evolución microbiana, ya que muchos de sus muestreos generan este tipo de datos y poder automatizar el análisis de estas muestras supone un ahorro de tiempo muy importante, ampliando la capacidad de respuesta en estudios pequeños y colaboraciones interesantes como la de la unidad de quemados.

Es un trabajo sobre todo descriptivo de la biodiversidad encontrada en ese ambiente en concreto y de búsqueda de patrones que puedan ayudar a entender la distribución de los microorganismos en el ambiente hospitalario.

DESCRIPCIÓN del RAW DATA.

Los datos están contenidos en dos documentos Excel: “Diversidad Hospital.xlsx” y “Metadatos.xlsx”.

“Diversidad Hospital.xlsx” contiene la información taxonómica (los microorganismos) localizados en los muestreos de la unidad de quemados. Tiene dos hojas: una de ellas llamada AIRE que contiene las muestras aéreas, y otra hoja llamada SUP que contiene el muestreo de superficies.

Las columnas de estas tablas son las siguientes:

Taxlevel: indica el nivel taxonómico de cada fila del documento:

1 Dominio
2 Filo
3 Clase
4 Orden
5 Familia
6 Género

rankID: el mismo tipo de información, pero con otro código que aplica jerarquía. Es una información más completa. Por ejemplo, dado que el rankID 0.1 corresponde al Dominio Bacteria, el 0.1.1 (Acidobacteria) y 0.1.2 (Actinobacteria) serán filos que pertenecen al dominio Bacteria. Estos, a su vez, tendrán sus clases y esas clases tendrán sus órdenes, familias y géneros.

Taxon: Nomenclatura, nombre asignado al nivel taxonómico perteneciente a un individuo que aparece en alguna de las muestras. Esos individuos se denominan unidad taxonómica operativa (UTO), también conocida por la sigla OTU. Es una unidad de clasificación seleccionada por el investigador que la utiliza para individualizar a objetos de su estudio, ya sea una especie u otro taxón de cualquier categoría, una morfoespecie, una población, y hasta un individuo, y de este modo poder ordenarlos en una clasificación.

Daughterlevels: no es fiable así que no se ha usado en este análisis

Total: es el número de secuencias que se han encontrado de ese OTU independientemente de en qué muestra sea**.**

El resto de las columnas son muestras, e indican la cantidad de individuos de cada OTU encontrados en cada una de ellas. En la hoja AIRE se describen las muestras aéreas:

- **CO-A** Sala espera interior y pasillo
- **EH-A** Sala espera fuera de la unidad
- **ICU-CO-A** Sala espera interior y pasillos
- **ICU-PR1-A** Hab 408 UCI Con paciente sedado de 4 días
- **ICU-PR2-A** Hab 405 UCI Vacía
- **ICU-TR-A** Sala de curas
- **OR-A** quirófano
- **PR-A** Hab 412-413 zona hospitalaria
- **TR-A** Sala de curas externa

La hoja SUP tiene las mismas columnas, pero contiene el muestreo de superficies. Las muestras han sido recogidas en las mismas salas y se denominan:

- **CO-S** Sala espera interior y pasillo
- **EH-S** Sala espera fuera de la unidad
- **ICU-CO-S** Sala espera interior y pasillos
- **ICU-PR1-S** Hab 408 UCI Con paciente sedado de 4 días
- **ICU-PR2-S** Hab 405 UCI Vacía
- **ICU-TR-S** Sala de curas
- **OR-S** quirófano
- **PR-S** Hab 412-413 zona hospitalaria

- **TR-S** Sala de curas externa

El otro fichero Excel “Metadatos.xlsx” tiene una única hoja llamada METADATOS:

Name: las filas de esta columna son el nombre de las muestras; coinciden con los nombres de las columnas de muestras del Excel “Diversidad Hospital.xlsx”.

Muestra tipo: variable categórica: Aire o Superficie.

Ventilación: variable categórica: ventilación de circuito abierto o circuito cerrado.

Tipo Zona: variable categórica de tipo de zonas que se encuentran en la unidad de quemados: zona de consultas y de hospitalización (visitable) llamada zona Hospitalaria, zona UCI no visitable por personal ajeno salvo permiso, zona Quirófano, y zona externa a la unidad.

Zona: salas donde se ha tomado muestra.

Tª: temperatura a la que se tomó la muestra.

H.realtiva: humedad relativa cuando se tomó la muestra.

Partículas de 0,3 micras: suspendidas en el aire cuando se tomó la muestra.

Initial number of sequences, % of removed chimera, Secuencias antes de restar negativos: controles que se hacen con respecto al dataset original que sale de la unidad de secuenciación por criterios biológicos. No se emplean en el análisis.

Number of OTUs: secuencias únicas en cada muestra sacadas con el Stamp no coincide con el comando unique.

Chao-1: estimador de diversidad basado en la diversidad de la muestra. Cuenta mucho tener individuos poco representados.

Shannon H Index: este índice se representa normalmente como H’ y se expresa con un número positivo, que en la mayoría de los ecosistemas naturales varía entre 0,5 y 5, aunque su valor normal está entre 2 y 3; valores inferiores a 2 se consideran bajos en diversidad y superiores a 3 son altos en diversidad de especies.

Simpson index: es uno de los parámetros que nos permiten medir la riqueza de organismos. En ecología, es también usado para cuantificar la biodiversidad de un hábitat. Toma un determinado número de especies presentes en el hábitat y su abundancia relativa. El índice de Simpson representa la probabilidad de que dos individuos, dentro de un hábitat, seleccionados al azar, pertenezcan a la misma especie. Es decir, cuanto más se acerca el valor de este índice a la unidad, existe una mayor posibilidad de dominancia de una especie y de una población; y cuanto más se acerque el valor de este índice a cero mayor es la biodiversidad de un hábitat.

Preparación de los datos

Adquisición de los datos

Los datos los ha cedido el laboratorio de Evolución Microbiana, del Centro de Astrobiología. Se realizó un muestreo en una unidad de quemados durante un día. Hay que tener en cuenta que fue un gran esfuerzo por parte del hospital, ya que no se suele interrumpir la dinámica hospitalaria. El muestreo se realizó en todo momento bajo su permiso, y siguiendo

estrictamente las indicaciones pertinentes para no romper sus propios métodos de control. Los quirófanos, por ejemplo, fueron muestreados una vez que habían sido usados en sus operaciones, ya que la limpieza de estas instalaciones no debe ponerse en riesgo, y no está permitido ningún tipo de actividad después de una limpieza de quirófano, salvo una cirugía.

Obtención y comprensión de los datos

Los datos son tablas de recuento de secuencias de DNA 16S, obtenidas mediante una extracción de DNA de muestras ambientales obtenidas en diversas salas de la Unidad de quemados de un Hospital.

El proceso resumido es el siguiente:

1. Toma de muestra.
2. Extracción de DNA.
3. PCR para centrar el análisis en el 16S (secuencia del RNA ribosomal) bajo la cual hay gran cantidad de datos de clasificación de microrganismos en bases de datos diferentes. Esta tabla esta creada a través de la base de datos SILVA[3].
4. Secuenciación de todas las secuencias 16 S amplificadas.
5. Análisis de la diversidad de las comunidades encontradas en las muestras. Aquí es donde tiene sentido este trabajo.

Los resultados de la secuenciación tienen una serie de tratamientos previos realizados con mothur[4] y bajo el criterio del grupo de análisis de evolución microbiana como por ejemplo la resta de OTUs consideradas control negativo (OTUs que aparecen debido al uso de determinados kits de extracción de DNA). Los resultados del pre-proceso con mothur se guardan en el Excel “Diversidad Hospital.xlsx” y constituyen el raw data de este trabajo.

Los datos se analizaron con Pandas, cargando las hojas de Excel como DataFrames. Al tener el Excel dos hojas, dos archivos diferentes (uno relativo a muestras de aire, y otro con muestras de superficie) uno de los primeros problemas que debía ser resuelto era unificar esos dos grupos de datos.

Otro problema es que, observando los datos, se puede ver la presencia de filas con nombres (columna “Taxon”) como unclassified o uncultured. Esto corresponde con secuencias que pueden estar taxonómicamente en grupos diferentes, pero son secuencias de especies no estudiadas (no están dentro de las bases de datos de la clasificación, o son secuencias de microorganismos no cultivados y por lo tanto no clasificables). Conociendo su jerarquía (rankID) sí se les puede asignar dentro de algún grupo taxonómico concreto. Lo importante es no perder esa información solo tener un nombre genérico, de modo que todos los OTUs de un mismo nivel (“Taxlevel”) tengan un nombre (“Taxon”) único.

Finalmente, es verdad que el rankID clasifica los OTUs dentro de grupos de forma unívoca. Pero, al unir los dos data frames, esa columna forzosamente pierde todo el sentido ya que se genera independientemente para cada archivo obtenido del Mothur. Por tanto, un mismo OTU puede tener un rankID distinto en cada una de las hojas Excel (AIRE y SUP).

Procesado de OTUs sin clasificar o sin cultivar (Notebook uncultured_unclussified)

Lo primero que se solucionó fue evitar la pérdida de información, completando los nombres de aquellas celdas en las que solo apareciera la palabra uncultured o unclassified, con el nombre del rankID anterior que contuviera información.

Hay que comprender que cada OTU que aparece en el cuadro pertenece a niveles de clasificación diferentes contenidos en la columna de taxlevel y también en el rankID (ver la descripción del raw data). En primer lugar se creó un filtro para localizar las celdas en las que se localizan los contenidos "unclassified", y se comprobó que, en este caso, todos están contenidos en el taxlevel 6.

Para procesar estos datos, se seleccionan todas las filas que tienen unclassified y luego, gracias a la información contenida en la columna rankID, se selecciona el primer nivel que contiene información y se rellena la celda con esa información de tal modo que todos los unclassified sean nombre_taxon_unclussified y se genera una función (*padre_level_unclussified*) que aplica este sistema a todas las filas.

	taxlevel	rankID	taxon	daughterlevels	total	ICU-PR2-S	ICU-CO-S	ICU-TR-S	ICU-PR1-S	OR-S	EH-S	CO-S	TR-S	PR-S
0	0	0	Root	1	171719	23652	15738	15239	18508	19693	40072	15156	11068	12593
1	1	0.1	Bacteria	27	171719	23652	15738	15239	18508	19693	40072	15156	11068	12593
2	2	0.1.1	Acidobacteria	6	780	131	74	148	11	105	0	30	144	137
3	3	0.1.1.1	Acidobacteria_unclassified	1	5	4	0	1	0	0	0	0	0	0
4	4	0.1.1.1.1	Acidobacteria_unclassified	1	5	4	0	1	0	0	0	0	0	0
5	5	0.1.1.1.1.1	Acidobacteria_unclassified	1	5	4	0	1	0	0	0	0	0	0
6	6	0.1.1.1.1.1.1	unclassified	0	5	4	0	1	0	0	0	0	0	0
7	3	0.1.1.2	Acidobacteriia	3	424	48	30	56	11	105	0	30	144	0
8	4	0.1.1.2.1	Acidobacteriales	3	29	10	5	0	7	7	0	0	0	0
9	5	0.1.1.2.1.1	Acidobacteriaceae_(Subgroup_1)	3	23	10	0	0	6	7	0	0	0	0
10	6	0.1.1.2.1.1.1	Acidipila	0	8	8	0	0	0	0	0	0	0	0
11	6	0.1.1.2.1.1.2	Granulicella	0	9	2	0	0	0	7	0	0	0	0

Figura 1 Datos en crudo de las muestras de superficioes.

El resultado es que todos esos nombres que antes eran iguales ahora al ponerle apellidos quedan claramente diferenciados unos de otros. En el ejemplo de la figura, el OTU de la sexta fila ("unclassified", rankID 0.1.1.1.1.1.1) se renombrará como "Acidobacteria_unclassified" para diferenciarlo de otros "unclassified" del taxlevel 6-

El caso de los uncultured es más numeroso y afecta a más taxlevels. Así que se aplica esta función que es muy similar, (*padre_level_uncultured*) pero se le ha añadido una condición para que reconozca el nombre del primer nivel taxonómico que contenga información, de tal manera que si identificasen que nivel con el primer uncultured es, por ejemplo, el nivel 4 , hay que aplicar 3 veces esa función para eliminar todos los uncultured.

Este procesado lo aplicamos sobre las dos hojas del Excel "Diversidad Hospital.xlsx" y generamos otro nuevo Excel "Diversidad Hospital_UN_trated_Hpy.xlsx" también con dos hojas. Y aquí acaba el primer notebook.

Generación de la jerarquía taxonómica (Notebook Columcreator)

Este espacio de trabajo se generó después de varios intentos para realizar la primera unión de archivos. Se conseguía, pero perdiendo la jerarquía que conlleva tener la columna rankID, que contiene información muy útil.

Se parte del Excel generado en el notebook anterior Notebook “Diversidad Hospital_UN_trated_Hpy.xlsx”.

La conclusión, después de darle bastantes vueltas, fue crear columnas con la información de los niveles taxonómicos, es decir optar por la redundancia de información, para poder mantenerla en procesos posteriores. De esta manera se generaron columnas que se llaman: Dominio, Filo, Clase, Orden, Familia y Género, que se rellenaron conn información según la función *(column_creator).* Se vuelve a usar una vez más la columna rankID para poner el nombre del taxón que corresponda.

Esta información se guarda en el Excel 'Diversidad Hospital_columscreator_Hpy.xlsx y en su correspondiente hoja.

Unión de los datos de aire y superficies (Notebook merge_taxones)

Finalmente se lleva a cabo la unión de los dos archivos sin pérdida de información, realizando un ***merge*** considerando la información contenida en las siguientes columnas 'taxon', 'taxlevel', 'dominio', 'phylo', 'clase', 'orden', 'familia', 'genero'.

El resultado es un data fra,e en el que se quedan desdobladas las siguientes columnas:
'rankID_x', 'daughterlevels_x', 'total_x', ’rankID_y', 'daughterlevels_y', 'total_y'.

De ellas solo nos interesa el total: sumamos total_x y total_y lo conservamos como ‘total’.

Las demás columnas desdobladas se eliminan, ya que la información jerárquica la contienen las columnas generadas en el notebook columcreator , que mantienen esa información y son sobre las que se ha generado el merge.

Se guarda todo en un Excel llamado: 'Diversidad Hospital_merge_Hpy.xlsx'. Este documento ya estaría preparado para aplicarle las técnicas de visualización y análisis que se describen a continuación.

Análisis

Descriptiva directa

Es interesante mirar el número de secuencias que contiene cada muestra, y también estudiar cómo de variables son esas secuencias, es decir el número de secuencias únicas que encontramos. Esto se aprecia en la Figura 2: a simple vista se puede ver que hay muestras que contienen más cantidad que otras, siendo la de la mayor cantidad de secuencias en el aire de una habitación de la zona de hospitalización donde se pueden recibir visitas; aunque llama la atención que la variedad de secuencias únicas es de las más bajas.

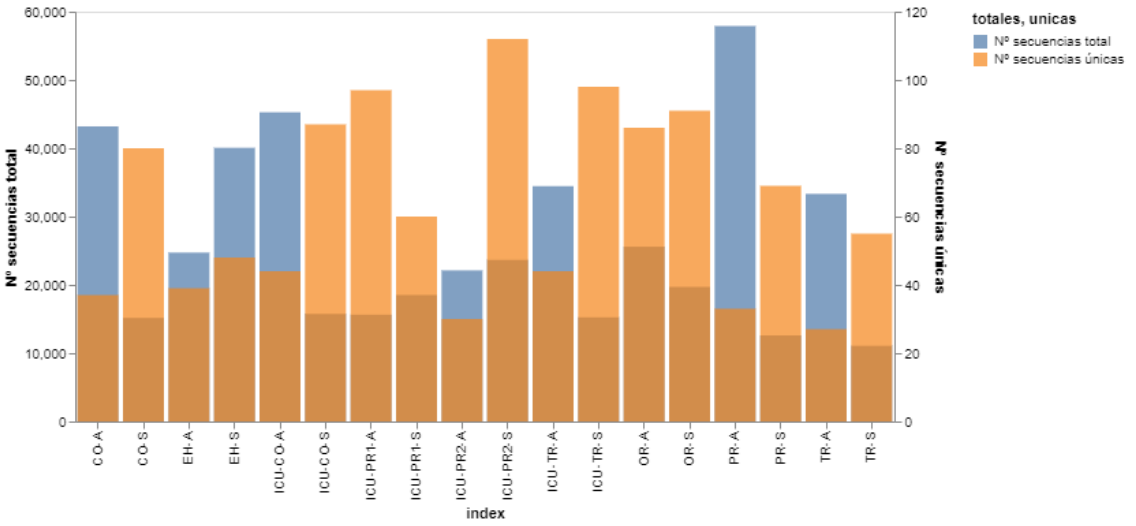

Figura 2 Son dos Gráficos superpuestos realizados con Altair a partir del dataframe df4. Contiene dos ejes independientes, el de la derecha es para las barras naranjas (nº de secuencias genéticas únicas en cada muestra), y el de la izquierda (nº de secuencias totales) para las barras azules.

A su vez también podemos hacer notar que la variedad más alta de secuencias se ha localizado en la superficie de una habitación de UCI vacía. Aunque esto pueda parecer extraño a priori, se puede decir que la unidad de quemados es un ambiente en general bastante limpio, y los dataframes con los que se están trabajando no tienen un gran número de secuencias como en otros estudios medioambientales.

Queda más patente si se hace un describe de df4 (Tabla 1) donde se aprecia que la media de secuencias únicas de las 18 muestras analizadas es de 63 secuencias diferentes eso es un ambiente pobre en diversidad, o extremadamente limpio en el caso de hablar de un hospital.

Tabla 1 Resultados de df4.describe()

df4.describe()	Nº secuencias únicas	Nº secuencias total
count	18.000000	18.000000
mean	63.166667	26328.666667
std	27.173192	13225.047728
min	27.000000	11068.000000
25%	40.250000	15654.000000
50%	57.500000	22885.000000
75%	86.750000	34170.000000
max	112.000000	57900.000000

Otra forma de analizar la diversidad la encontramos bajo los índices de diversidad que contiene la Tabla Excel de Metadatos en la hoja de metadatos. Haciendo una gráfica sencilla de barras y definiendo someramente lo que cada índice explica, podemos afianzar las conclusiones extraídas de Figura 2.

El índice de **Shannon**[5] es una aplicación de la teoría de la información. Se basa en la idea de que la mayor diversidad corresponde a una mayor incertidumbre en elegir de manera aleatoria a una especie en específico en la mayoría de los entornos ecológicos. Varía entre 0,5 y 5: valores

inferiores a 2 se consideran pobres en diversidad y por encima de 3 una diversidad alta, casi todas las muestras están entre 2 y 3.

El índice **Chao-1**[6] está basado en la aparición de especies raras ''singletons'' (que aparecen solo una vez) o "doubletons'' (que aparecen dos veces) en los recuentos de hábitats. Cuanto más aumenta la presencia de "singletons", la estimación de la riqueza aumenta también, por lo que en la población de secuencias, este estimador refleja más al valor de la riqueza observada. En nuestras muestras la riqueza en algunas de ellas se dispara debido a la presencia de esos "singletons". Consideramos que este índice en este tipo de muestra no está siendo representativo ya que no tiene en cuenta demasiado la abundancia de las especies. En nuestras muestras el índice se puede disparar debido a esa aparición de especies raras (en las que solo muestreas una secuencia o dos), parece una muestra muy diversa pero realmente luego en número de secuencias no lo es.

El índice **Simpson**[7] también es conocido cómo índice de la dominancia de especies, y representa la probabilidad de que dos individuos dentro de un hábitat seleccionados al azar pertenezcan a la misma especie, esto implica que los valores son entre 0 y 1 cuanto más cercano a cero mayor diversidad y más cercano a uno mayor dominancia. En la inmensa mayoría de muestras se puede observar que es este índice está muy cercano a uno, lo que quiere decir claramente que en los entornos donde se han tomado las muestras existen especies dominantes.

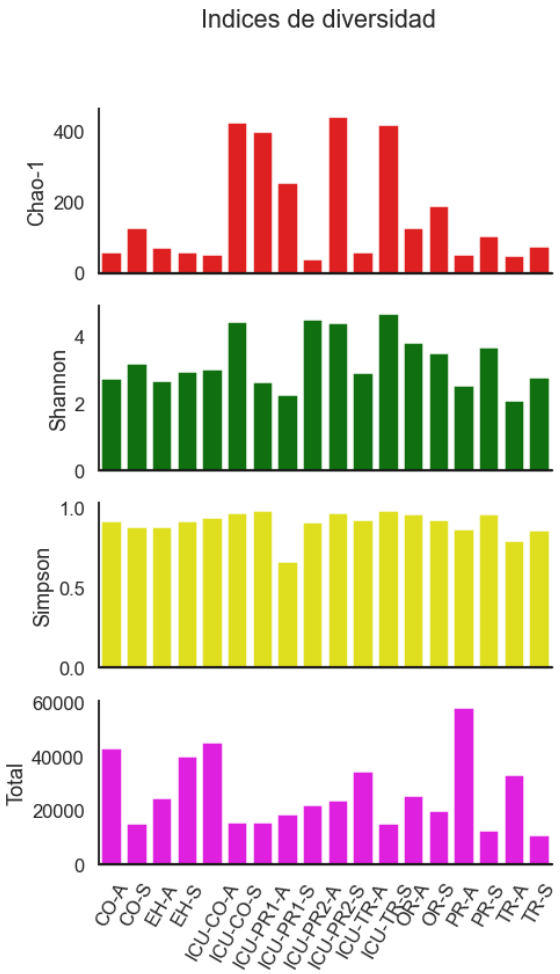

Figura 3 Índices de diversidad

Krona

Otra forma de ver la composición de la muestras es generar un gráfico con la interfaz KRONA (Figura 4), que se encuentra en https://github.com/marbl/Krona/wiki[8] . Es un gráfico interactivo en el que se puede ver la composición taxonómica de todas las muestras seleccionando una a una en la interfaz, así como también realizar búsquedas concretas: se selecciona un taxón de interés y permite la acción de mover esa búsqueda a lo largo de las muestras y dice si está el taxón concreto y en qué proporción.

Para que funcione el programa se han generado mediante un notebook de Python (KRONA.ipynb) archivos de texto, uno por cada muestra, a partir de los datos del Excel 'Diversidad Hospital_merge_Hpy.xlsx', y se han guardado en la carpeta de Krona mostrada en el repo. A partir de ellos, se genera un archivo html usando la utilidad ktImportText de krona, instalada en Linux. El fichero html resultante está incluido en el frontend.

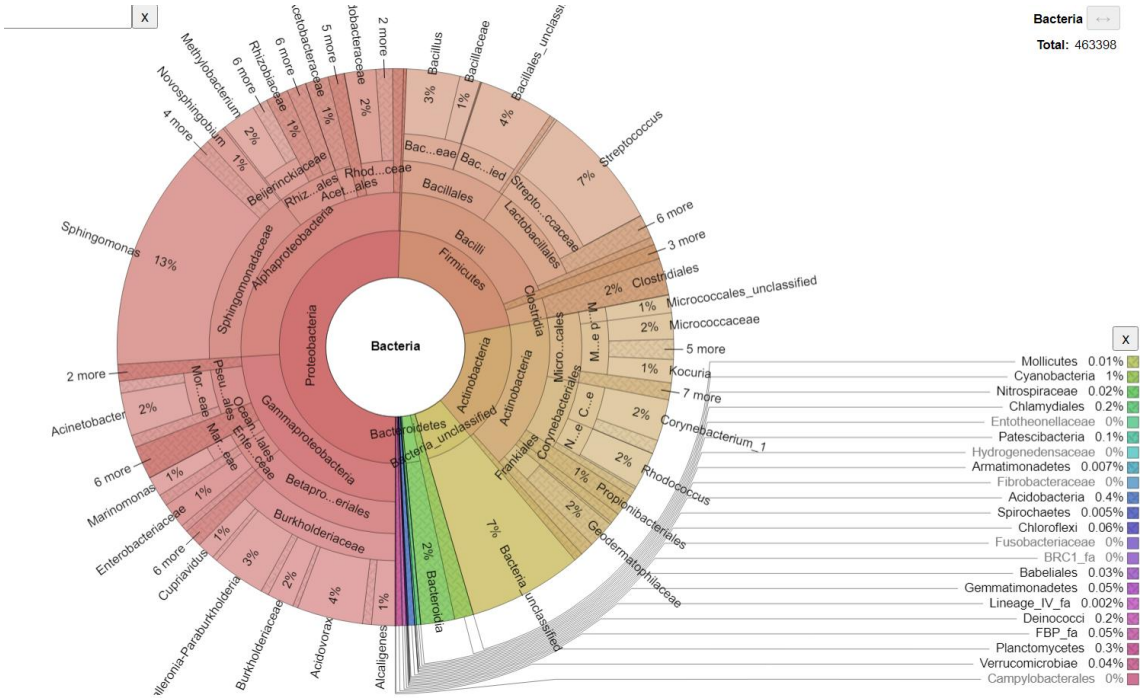

Figura 4 Krona, es un gráfico interactivo aquí solo se muestra una imagen fija, se ve en el archivo HTML de la carpeta krona, ha sido generado a partir de archivos de texto que se han creado mediante código Python partiendo de la tabla Excel 'Diversidad Hospital_merge_Hpy.xlsx'

Barras apiladas tax2 phylos

También es favorable para hablar de diversidad mirar los filos que están más representados en el sistema de estudio, esto es el nivel 2 de la columna taxlevel del archivo 'Diversidad Hospital_merge_Hpy.xlsx'. Esta gráfica, aunque es sencilla, nos da mucha información, y muy visual. Vemos representados muy bien cuatro colores azul, naranja, verde y rojo que corresponden con Proteobacterias, Actinobacteria, Bacteria_unclussified, y Firmicutes, y luego una cantidad de filos muy poco representados en abundancia, y esto es por lo que se puede decir que estas muestras ecológicamente no son muy diversas.

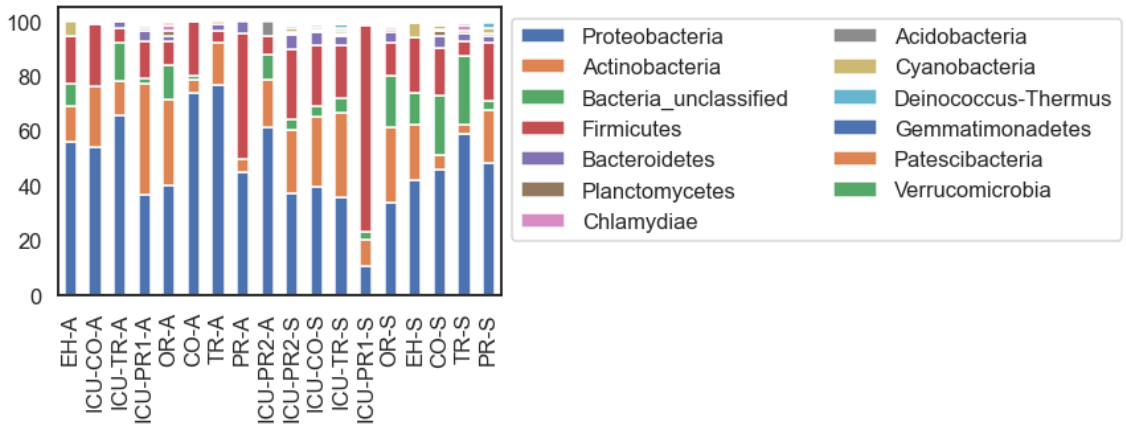

Figura 5 Gráfico de barras apiladas, de nivel filos.

De estos cuatro filos se pueden extraer unas cuantas ideas:

Las **Proteobacterias** es el filo más grande de bacterias gran negativas. Es extenso y hay mucha variedad dentro de este grupo, es común en los muestreos tenerlo bien representado.

Firmicutes es un filo de bacterias Gram positivas cuyos componentes en su gran mayoría son capaces de formar endosporas, estructuras de resistencia que les permiten sobrevivir de manera latente en condiciones extremas de nutrición o de humedad. Es muy común encontrar este filo en muestras aéreas, y tiene sentido encontrarlo en zonas muy limpias del hospital, ya que es de suponer que las bacterias resistentes a los protocolos de limpieza, porque tienen mecanismos para sobrevivir en esas condiciones. En esas condiciones las células no están activas metabólicamente, solo están sobreviviendo hasta encontrar un hábitat con mejores condiciones. Están latentes; por eso no se localizan grandes cantidades de secuencias.

Las **Actinobacterias** es uno de los filos más abundantes, es muy normal tener carga en este filo, es señal de que el muestreo está bien hecho. Como dato curioso se cree que representan el 64% de la biomasa bacteriana y también son productoras de esporas[9]. Muchas de ellas también pueden crear esporas para su dispersión, que también funcionan como formas de resistencia. Algunos de los componentes de este filo han sido descritos como los seres vivos más antiguos de la tierra ya que se han encontrado esporas en el permafrost del ártico y de Siberia, viables[10].

De las **Bacterias_unclussified**[11] son secuencias de DNA de las que solo se pueden decir que vienen del Dominio Bacteria y no se puede decir más. A veces retirarlas del dataframe permite llegar a conclusiones diferentes. En este caso se han incluido en este estudio, aunque se ha comprobado que, si se retira ese grupo de los análisis, no cambian mucho las conclusiones que se tienen en este trabajo. Los archivos generados no se incluyen por no complicar el trabajo pero están analizados. Representan un 7% de las muestras de este estudio y eso no es un valor alto. Para que quede patente el trabajo que aún queda por hacer en ecología y taxonomía microbiana en 2013 se calculaba que solo se conocía el 1% de los microorganismos de suelo[12].

Descriptiva Indirecta

PCA

Para estudiar la diversidad desde otro punto de vista vamos a realizar un PCA, para extraer gradientes de máxima variación, para ver si revelamos algún tipo de agrupamiento. Realizaremos una reducción de dimensionalidad a la matriz generada con el dataframe del taxlevel 6, es decir los géneros, que en principio serán nuestras variables.

La varianza explicada en las cuatro primeras componentes no es excesivamente elevada y las diferentes normalizaciones aunque mejoran los pairplot no mejoran el dato de la varianza, tampoco se pierde en exceso.

Los componentes principales han sido obtenidos tras jugar con varios métodos de normalización eligiendo (np.log1p) que devuelve el logaritmo natural de uno más la matriz de entrada, por elementos. Así se conseguía dispersar alguno de los spots de los pairplots, pero en general no se aprecia ningún tipo de distribución a la que le pueda dar ningún sentido. La siguiente figura muestra el scatter plot de las dos primeras componentes (PC1 y PC2) para las distintas muestras, agrupadas por Filo (taxlevel 2).

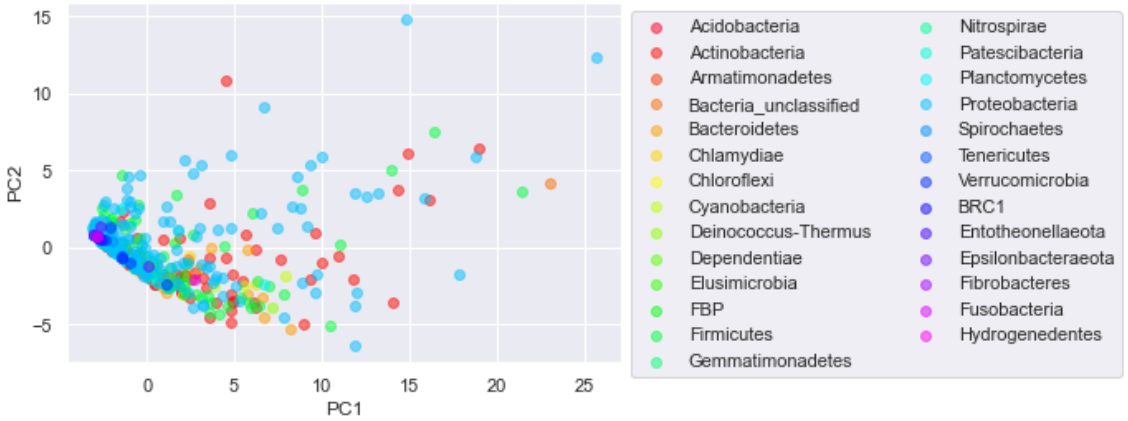

Figura 6 Scatter plot de Matplotlib con los componentes PC1 y PC2 en los ejes, los datos han sido normalizados

Aun así, se miran los loadings, las cargas de los componentes principales en concreto de los cuatro primeros y se analizaron en un scatter generado en altair (Figura 7). Se pueden relacionar, al tener ya en el análisis los nombres de las muestras, con variables categóricas recogidas en la tabla de metadatos, de tal manera que se empezaron a ver clusters de zonas lugares de muestreo: tanto por si eran muestras aéreas o si eran muestras de superficies. Alguna muestra se escapa a esa clusterización, es el caso de la ICU-PR-1-A, pero ya veremos que es una zona de muestreo un poco particular: es una habitación de UCI con un paciente sedado enfermo que llevaba en la habitación 4 días ingresado y puede estar muy influenciada. También, de manera menos clara, se puede ver que existen muestras más vinculadas o parecidas que pueden pertenecer a zonas similares, por ejemplo, quirófanos.

No se puede aventurar a que son debidos estos pesos ya que ni son extremadamente diferentes ni se sabe que variable está actuando más en su valor, pero sí se le puede ir buscando explicaciones testeándolo con las variables categóricas cómo se puede ver en el frontend.

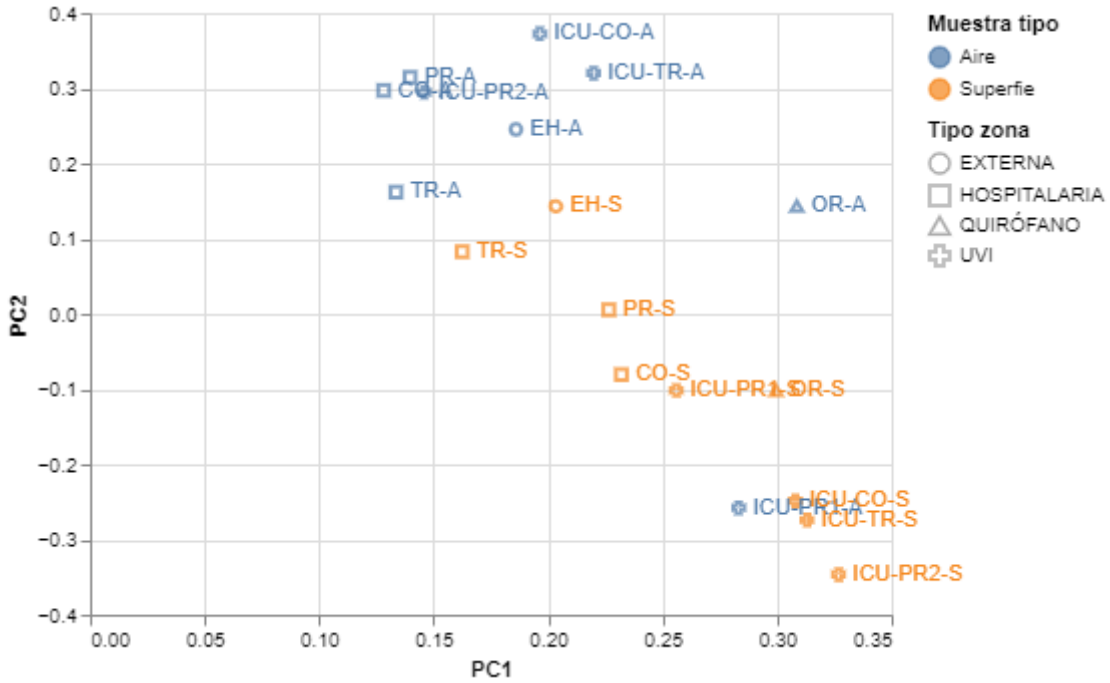

Figura 7 Scatter plot de los loadings de las componentes relacionado con variables categóricas.

RDA

El objetivo del RDA es extraer gradientes de variación en variables dependientes explicadas por variables independientes.

PCA y RDA en apariencia son muy similares. Aunque difieren porque el PCA no está restringido (busca cualquier variable que explique mejor la composición de especies), mientras que el RDA está restringido (busca las mejores variables explicativas). Depende de las longitudes del gradiente.

En este trabajo está implementado en R ya que parecía un paquete más completo para esta técnica y te permitía una mejor visualización, además se ha aplicado una serie de test de discriminación para ver que variables son estadísticamente más significativas.

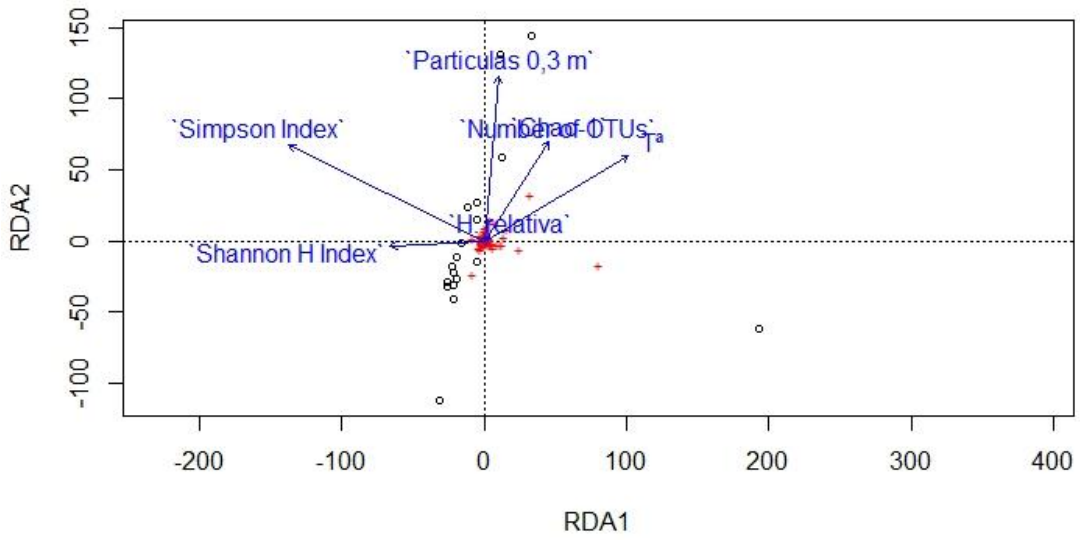

Figura 8 RDA plot donde se ven las variables explicativas, las secuencias en rojo, y las muestras en amarillo

Las únicas variables que han salido significativas en el análisis del RDA han sido el índice de Simpson y el número de individuos de cada muestra, eso, es lo que explicaría parte de la distribución de las muestras en el espacio creado para el RDA.

Network

Por último, se realizó un gráfico network con la librería NetworkX cuyos nodos (amarillos) son las muestras. Las relaciones de las muestras entre sí están basadas en el índice de disimilitud de Bray-Curtis[13]. En grandes rasgos este índice consiste en la composición de las muestras, si dos muestras son iguales en la composición de su población (en este caso secuencias), tiene un valor de 0 y si son distintas totalmente y no comparten ninguna especie tienen un valor de 1. Esto genera una matriz de relaciones, que se puede representar en una network. Se han representado en la red las conexiones entre muestras cuyo índica de Bray-Curtis es inferior a 0.55.

Después se le ha añadido a cada muestra los microorganismos a nivel de género (nodos azules). Como sería un exceso poner todos en una sola imagen, se pone una condición se seleccionan solo aquellos microorganismos que superan el 3% del total de la población, para eliminar secuencias poco representadas. Y las uniones o flechas tienen un grosor y color determinado, tienen peso y color en función de la abundancia.

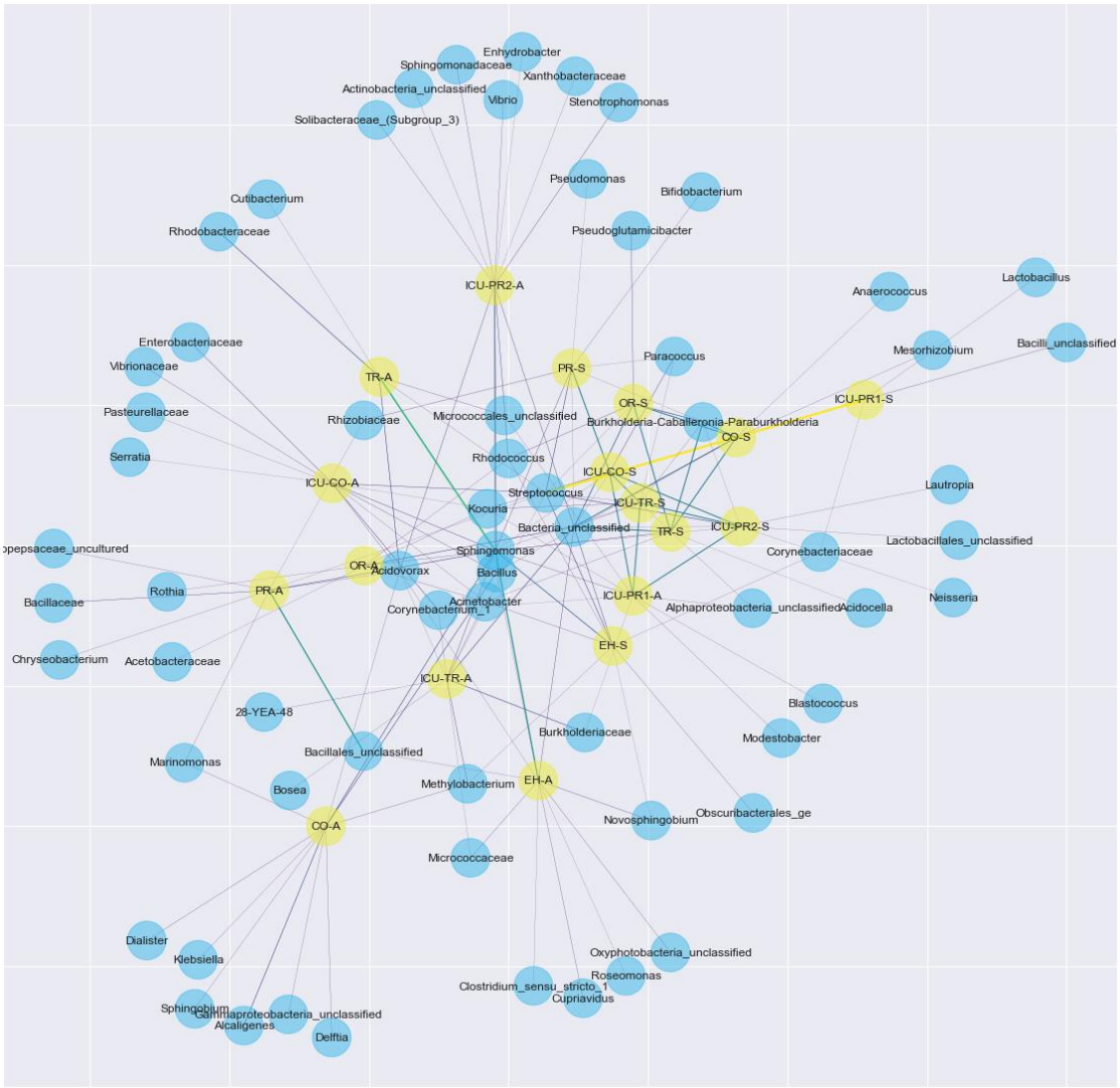

Figura 9 Network con índice de Bray Curtis <0,55 y las secuencias solo aparecen representadas aquellas de nivel 6 género que superen un 3% del total de cada muestra. Las uniones entre los nodos amarillos se seleccionan con el índice de Bray-Curtis, aparecen o no en función del corte que hemos puesto, pero el grosor es fijo. Por el contrario, el grosor de los nodos azules que corresponden a las secuencias pueden tener en común más de un nodo amarillo y además el grosor la línea de unión representa la abundancia.

En base a todo esto llama la atención la flecha que se queda en color amarillo que une una habitación de UCI con enfermo, junto a la secuencia de *streptococcus*. Nadie nos dijo que tipo de infección tenía la persona que estaba en la habitación, pero esto nos hace suponer que probablemente sufriera una infección por *Streptococus*. Esto a su vez afectaba al muestreo de aire y al pasillo que se muestreó en la UCI.

También llama la atención que las muestras de aire de pasillos y sala de espera externa de la unidad CO-A y EH-A no comparten secuencias, al menos no fuertemente con otros puntos de muestreo, seguramente se deba al tipo de ventilación. Las muestras de superficies están más relacionadas entre sí, quedando más en el centro de la red.

Front-end

Se ha realizado un front-end usando streamlit para el análisis interactivo de los datos.

El front-end se lanza desde línea de comandos:

```
streamlit run frontend.py
```

Se incluyen las distintas visualizaciones mencionadas anteriormente, excepto el RDA, esto es:

- Descriptiva directa, incluyendo Krona
- PCA
- Network

Se puede seleccionar de forma interactiva en qué nivel taxonómico (taxlevel) se realiza el análisis. El análisis que hemos descrito (en taxlevel 6, género) se puede hacer por ejemplo por familia o por orden.

También se pueden seleccionar de forma dinámica algunos parámetros de análisis como el tipo de normalización del PCA o los umbrales de la network.

CONCLUSIONES

El muestreo en general demuestra un número bajo de secuencias, y en concreto un número no despreciable son únicas, lo que dispara el índice Chao-1 en muchos puntos del muestreo. Aunque pueden ser algunas de las secuencias muy numerosas, la carga muestral analizada en absoluto lo es (Figura 3). Se puede decir que la limpieza en la Unidad de quemados es muy elevada.

De los nombres de patógenos que nos facilitó el hospital solo aparece el género *Streptococcus* en una habitación en la que estaba ingresado un paciente sedado y con infección. El *Streptococcus* aparece sobre todo en la superficie ICU-PR1-S , en el aire ICU-PR1-A, y en el ICU-COA y, en superficie. Esto se puede comprobar muy bien en el Krona interactivo añadido en el frontend.

Otro género que es patógeno altamente problemático para una unidad de quemados es el género *Staphylococcus,* que aparece en muy baja abundancia en algunas muestras. Sin ser alarmistas, es un microorganismo habitual en la flora de la piel y de la nariz de las personas sanas y por lo tanto es normal su aparición. Aparece al 0,5% en la habitación ocupada de la UCI, pero allí el *Streptococcus* es el predominante representando al 71% de las secuencias de bacterias recolectadas en superficie y el 6% de las secuencias bacterianas del aire de la habitación.

De los filos (Figura 5) representados en el muestreo, aparentemente los mayoritarios tienen sentido que aparezcan, bien por ser bacterias que presentan formas de resistencia ante fenómenos adversos como puedan ser la baja disponibilidad de nutrientes debido a una limpieza exhaustiva, bien porque son grupos muy grandes y su ausencia sería un gran interrogante.

Del análisis con PCA (Figura 7), podemos quedarnos con el agrupamiento que se produce de las muestras aéreas y de las muestras de superficie, aunque hay que decir que la varianza explicada debido a la normalización es del 61% de los cuatro primeros componentes, es algo baja.

Del gráfico Network se puede jugar a agrupar las muestras en función de su similitud, y se puede ver en parte la abundancia de los microorganismos y las diferentes relaciones entre las muestras. aquí se puede ver el peso que tiene el género *Streptococcus* y sin embargo no aparece, aunque existe en casi todas las muestras, el género *Staphylococcus*, ya que su abundancia no es preocupante.

De cara al género *Streptococcus* no sabemos qué especie en concreto es la que está produciendo la infección que tampoco nos la han confirmado ya que es información confidencial. Aunque en las zonas aledañas a la habitación del enfermo continúa siendo elevado la aparición de secuencias, el hospital tiene sus protocolos para evitar este tipo de dispersiones. En el resto de las zonas, si pusiéramos como control la zona externa a la unidad, que contiene un 14% de secuencias de *Streptococcus* del total en superficie y un 0.08 % en aire, este umbral solo se supera en las zonas de la UCI aledañas a la habitación del paciente y que claramente llevarán un protocolo adecuado para evitar la dispersión.

Con toda probabilidad las infecciones se pueden transmitir por el aire, pero en la unidad ya se tiene controlado con sistemas de limpieza (filtrado) el aire. Si bien es verdad que los dos géneros problemáticos aparecen en casi todas las ubicaciones y probablemente forme parte del propio bioma de las personas afectadas con una quemadura, y se den las condiciones para que pasen a ser microorganismos que generen infección, la única forma de no extenderlo es confinándolos en sus habitaciones no compartidas y controlando muy bien la limpieza de salas como quirófanos y salas de curas, que son cosas que el hospital ya hace. Los niveles encontrados no son alarmantes tienen su explicación, y el único nivel preocupante es la habitación del enfermo y ya está siendo tratado.

Bibliografía

1. Rafla K, Tredget EE. Infection control in the burn unit. *Burns*. 2011;37(1):5-15. doi:10.1016/J.BURNS.2009.06.198

2. TCAE en la unidad de quemados - Francisco Lorenzo Tapia - Google Books. Accessed July 17, 2021.

3. Silva. Accessed July 17, 2021. https://www.arb-silva.de/

4. Schloss PD, Westcott SL, Ryabin T, et al. Introducing mothur: Open-Source, Platform-Independent, Community-Supported Software for Describing and Comparing Microbial Communities. *Applied and Environmental Microbiology*. 2009;75(23):7537-7541. doi:10.1128/AEM.01541-09

5. Shannon CE, Weaver W. THE MATHEMATICAL THEORY OF COMMUNICATION. Published online 1949.

6. Statistics AC-SJ of, 1984 undefined. Nonparametric estimation of the number of classes in a population. *JSTOR*. Accessed July 17, 2021.

7. Measurement of Diversity, by E. H. Simpson. Accessed July 17, 2021. http://people.wku.edu/charles.smith/biogeog/SIMP1949.htm

8. Ondov BD, Bergman NH, Phillippy AM. Interactive metagenomic visualization in a Web browser. *BMC Bioinformatics 2011 12:1*. 2011;12(1):1-10. doi:10.1186/1471-2105-12-385

9. Battistuzzi FU, Hedges SB. A Major Clade of Prokaryotes with Ancient Adaptations to Life on Land. *Molecular Biology and Evolution*. 2009;26(2):335-343. doi:10.1093/MOLBEV/MSN247

10. Johnson SS, Hebsgaard MB, Christensen TR, et al. Ancient bacteria show evidence of DNA repair. *Proceedings of the National Academy of Sciences of the United States of America*. 2007;104(36):14401. doi:10.1073/PNAS.0706787104

11. Furtak K, Grządziel J, Gałązka A, Niedźwiecki J. Prevalence of unclassified bacteria in the soil bacterial community from floodplain meadows (fluvisols) under simulated flood conditions revealed by a metataxonomic approachss. *CATENA*. 2020;188:104448. doi:10.1016/J.CATENA.2019.104448

12. Schink B, Stams AJM. Syntrophism Among Prokaryotes. *The Prokaryotes: Prokaryotic Communities and Ecophysiology*. Published online April 1, 2013:471-493. doi:10.1007/978-3-642-30123-0_59

13. Bray JR, Curtis JT. An Ordination of the Upland Forest Communities of Southern Wisconsin. *Ecological Monographs*. 1957;27(4):325-349. doi:10.2307/1942268