**HMP1-II**

Artículos asociados:

* Reference: The Human Microbiome Project Consortium. Structure, function and diversity of the healthy human microbiome. Nature 486, 207–214 (2012). <https://doi.org/10.1038/nature11234>
* Reference: Lloyd-Price, J., Mahurkar, A., Rahnavard, G., Crabtree, J., Orvis, J., Hall, A. B., ... & Huttenhower, C. (2017). Strains, functions and dynamics in the expanded Human Microbiome Project. Nature, 550(7674), 61-66. <https://doi.org/10.1038/nature23889>

Códigos de Proyecto y fuentes de los archivos Fastq:

* PRJNA48479 (Presente en HMP\_2012 y HMP\_2017), este proyecto tiene metadata asociada, pero los archivos asociados no están disponibles en el Proyecto porque hacen referencia a los archivos raw que están protegidos en la dbGap. Sin embargo, los archivos pasados por un control de calidad están disponibles en la web del HMP (<https://portal.hmpdacc.org/>).
* PRJNA275349 (Presente en HMP\_2017).

Metadata:

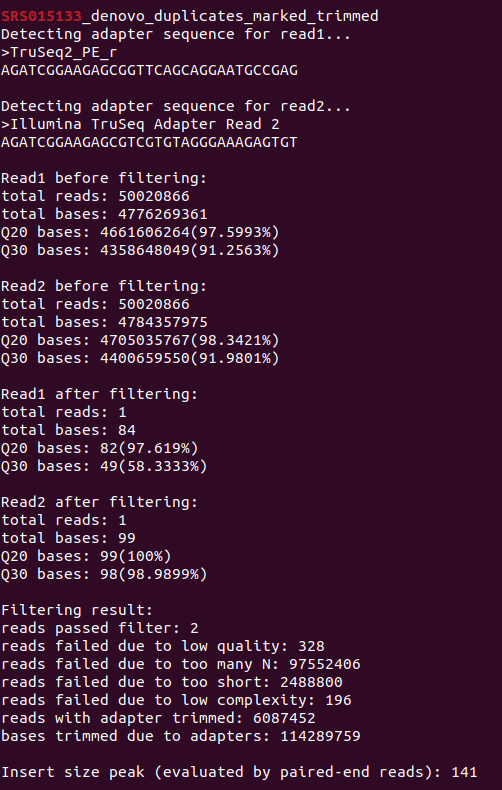
* En el paquete R curatedMetagenomicData hay metadata curada de HMP\_2012.
* En la publicación de Lloyd-Price en “Extended Data Table 1 Metadata summary and data availability”, vemos que todos los individuos son adultos entre 18-40 años. Además, indican que “Metadata availability: public metadata are available with the metagenomic taxon abundances table (<https://www.hmpdacc.org/hmsmcp2/>); private metadata are available through dbGaP with accession number [phs000228.v3.p1](https://www.ncbi.nlm.nih.gov/projects/gap/cgi-bin/study.cgi?study_id=phs000228.v3.p1).” Columnas eran:
  + SN: Sample alias (en base a los proyectos de ENA).
  + RANDSID: Identificador de Individuos (basándonos en el paquete de R y que hicimos un unique y el numero de identificadores coindiciía con el núemro de individuos de la publicación de HMP\_2017).
  + VISNO : Número de visita(hay algunas que aparecen como 02S y 03S). Se comprobó que dentro de un mismo individuo no había VISNO ‘2’ y ‘02S’, o un ‘3’ y ‘03S’, por lo que no hay motivo para preocuparse.
  + SNPRNT : Sample alias (aunque no siempre coincide con la columna SN).
  + Gender : (Male / Female)
  + WMSPhase : Creemos que corresponde a la pase en el proyecto en la que se generó la muestra. Estoy casi seguro de ello.
  + SRS : Corresponde al secondary\_sample\_accession de los proyectos de ENA.
* En la publicación de Lloyd-Price en “Extended Data Figure 1: Extended body-wide metagenomic taxonomic profiles in HMP1-II.” indican que “The combined HMP1-II datasets include a total of 2,355 metagenomes (724 previously published and 1,631 new, including 252 technical replicates). These span the project’s **six targeted body sites** (anterior nares, buccal mucosa, supragingival plaque, tongue dorsum, stool, and posterior fornix) in addition to at least 20 samples each from 3 additional sites, of the 18 total sampled sites: retroauricular crease, palatine tonsils, and subgingival plaque. Metagenomes are now available for at least one body site for a total of 265 individuals.”
* En cuanto al de HMP\_2012 hay unas muestras que había sospecha de contaminación en otros artículos posteriores (SRS014287, SRS050422, SRS062427). Habría que filtrarlas.

IMP. En base a las publicaciones de 2012 y 2017, he llegado a la conclusión de que efectivamente hay varias replicas técnicas, a veces dentro de la misma fase WMSPhase, y entre diferentes fases, es por tanto la combinación RANDSID + VISNO. En el artículo de 2012 decían que “Microbiome samples were collected from up to 18 body sites at one or two time points from 242 individuals clinically screened for absence of disease”, mientras que en la publicación de 2017 en base a la “Supplemental Table S1: Detailed body site sampling statistics. The number of subjects for which a given number of visits (1-3) are available with 16S profiles from HMP1 1 (300 subjects), shotgun metagenomes (expanded in HMP1-II, 265 subjects)”. Es decir, para estas muestras para cada individuo podemos tener como máximo hasta 3 visitas, y esto solo se cumple cuando usamos esa conbinación. Además, los números individuos coinciden casi a la perfección con lo que se dice en esta tabla:

* **Tabla-Gut-Stool(WGS) Publicación 2017**: 1 Visit(89), 2 Visits(98) y 3 Visits(62).
* **Combinación RANDSID + VISNO tabla metadata HMP\_2017 Public**: 1 Visit(89), 2 Visits(99) y 3 Visits(60). Tener en cuenta que en mi caso quite 3 muestras de un mismo individuo por sospecha de contaminación y que he quitado algunos accession extra lo que explica las pequeñas diferencias.

Procesado:

* **Descarga de la tabla de URLs en web HMP1.**
* **HMP\_2017\_Metadata**: Se tomó la tabla de metadata asociada a la publicación de 2017 de la web del HMP y se tomó como referencia. Después se filtró la tabla quedándonos solo con aquellos que la columna “STArea”=Gut y “STSite”=Stool. Dedando un total de 553 muestras de stool. También se filtraron los 3 sample\_accession sospechosos de contaminación (SRS014287, SRS050422, SRS062427) quedando 550 sample\_accessions.
* **HMP ENA Projects**: Se comprobó que de los 550 sample\_accessions de interés había un total de 463 sample\_accessions asociados al proyecto PRJNA275349 y 87 sample\_accessions más asociados al proyecto PRJNA48479. Se hizo un script propio para obtener el “instrument\_model” asociado a cada sample accession. Después de revisar estos resultados la mayoría eran Illumina, pero hay algunas excepciones (14 excepciones), después de lo cual se filtraron otros 5 sample\_accessions, quedando 545 sample\_accessions:
  + 3 sample\_accessions que aparecían unicamente como “454 GS FLX Titanium” se decidió quitarlos por seguridad (SRS022093, SRS016990, SRS013216). Se intento descargarlos por probar y los 3 daban error al intentar descargarlos (ERROR 404: Not Found.). Además no aparecen en la tabla de metadata de HMP\_2012 curatedMetagenomicData.
  + 1 sample\_accession que tiene “'454 GS FLX Titanium' y 'Illumina Genome Analyzer II'”pero una sola URL en la web (SRS019068). Se decidió quitarlo por seguridad. Además, no aparece en la tabla de metadata de HMP\_2012 de curatedMetagenomicData.
  + 9 sample\_accesions tienen '454 GS FLX Titanium' y 'Illumina Genome Analyzer II'” pero si que tienen urls distintas para 454 y Illumina (SRS045645, SRS055982, SRS015794, SRS058723, SRS019030, SRS015369, SRS051882, SRS022524, SRS063040). Estos 9 se mantienen.
  + 1 sample\_accession tiene '454 GS FLX Titanium' y 'Illumina Genome Analyzer II'(SRS064645), pero solo he visto url de descarga que no ponían /454/. Aunque si que aparece en la tabla de metadata de HMP\_2012 curatedMetagenomicData. Por seguridad se filtra también.
* **HMP 2012 curatedMetagenomicData**: se consiguio la metadata del paquete en su versión 3.8.0. Se hizo una prueba para ver cuanta metadata podiamos relacionar directamente hay 142 sample\_accessions comunes del HMP\_2012 de los que se dispone de metadata en el paquete (en especial su edad).
* **Tabla de urls de la web del HMP**: vimos los 545 sample\_accessions de interés estaban disponible para descargar en Web HMP. Asi que se decició simplificar y trabajar directamente con estos archivos. Se encontraron las siguientes peculiaridades:
  + Las propias urls:
    - En la mayoría de casos tenemos multiples maches para cada sample\_accession. Esto se debía a la presencia de ciertas urls que aparecían como “/v1/” y otras “/v2/”, y también otras “/454/”. La interpretación que son subcarpetas dentro de su servidor con diferentes versiones (/v1/ y /v2/) . En el caso de las “/454/” muestras secuenciadas con roche 454, en vez de Illumina (solo 12 y en /v1/). Se decidió filtrar las urls a solo las que apareciera “/v2/”, es decir, trabajar con la que entendemos que es la versión más reciente.
    - Se hizo una prueba con el accession SRS16018, bajando las urls “v1” y “v2”, y en este caso los archivos generados eran identicos. De hecho en la tabla tienen el mismo md5.
    - Además la tabla en la columna url, tiene en realidad dos urls separadas por una coma, se trabajo con la primera parte, ya que se comprobo que era la que tenia los archivos de interes.
  + 6 sample\_accessions extra no presentes en la publicación del HMP\_2017 (SRS016585, SRS023176, SRS045000, SRS045324, SRS054659, SRS142505). Estas aparecen todas en el proyecto privado PRJNA48479. El hecho de que sean muestras del HMP\_2012 y que no aparezcan en la tabla de metadata de la web de la referencia de 2017, me da mala espina asi que estás 6 muestras se han descartado.
* **Combinación Metadata**: (merge\_HMP\_metadata.py)
  + Primer Merge:
    - **HMP\_2017\_Metadata(referencia):** HMP\_2017\_Public\_Metadata\_stool\_without\_suspect\_samples\_without\_454.tsv
    - **Instrument\_Model\_ENA\_Projects:** HMP\_2017\_ENA\_projects\_instrument\_models\_without\_454\_modif.tsv / Este archivo a parte de estar filtradas las 5 excepciones ['SRS022093', 'SRS016990', 'SRS013216', 'SRS019068', 'SRS064645'], en las 9 excepciones (SRS045645, SRS055982, SRS015794, SRS058723, SRS019030, SRS015369, SRS051882, SRS022524, SRS063040), se dejo solo el modelo illumina, ya que nos aseguramos de que no se cogía la URL 454.
  + Segundo Merge: En la publicación también se hablaba de réplicas técnicas. Se comprobó cuantas combinaciones únicas de las columnas [RANDSID, VISNO] se repetían más de una vez (74 combinaciones RANDSID + VISNO). Asi que para tener esto en cuenta en el caso de que tuvieramos metadata para una de las réplicas en el paquete de R, se duplico esta información extra al otro sample\_accession asociado a la misma combinación RANDSID + VISNO.
    - Tabla resultante de Primer Merge.
    - **HMP\_2012\_curatedMetagenomicData** HMP\_2012\_curatedMetagenomicData\_v3.8.0.tsv
  + Arreglo extra, solo hay un caso de más de una coincidencia para réplicas en el HMP\_2012\_curatedMetagenomicData → ['SRS047014', 'SRS049164']. Al final, tenemos 190 sample\_accessions con metadata extra, como había 142 coincidencias directas, eso quiere decir que hay 48 sample\_accessions replicas que ahora tienen metadata extra también.
* Dereplicar réplicas técnicas eligiendo aquella que tuviera un mayor número de lecturas. Después de la derreplicación en base a las columnas (RANDSID + VISNO) y al número de conteos de los raw fastq, nos quedamos con 467 muestras (934 Fastqs, ya que eran PAIRED) de 248 individuos.
* Eliminación de la muestra **SRS015133.** Una vez se procesaron todas las muestras y se obtuvo la track\_table para los archivos procesados, se vio que los ficheros de esta muestra presentaban un recuento total de 1 lectura, tanto para el control de calidad con fastp como para bowtie2. Al observar el .log de fastp se comprobó que el problema estaba en la gran cantidad de secuencias que presentaban Ns y en la pequeña longitud de estas, por lo que se decidió eliminar.



Description Summary:

* **Tipo Dataset**: Poblacional.
* **Estado Individuos**: Healthy
  + **Antes Filtros**:
    - **N.º Muestras=Nº Individuos**: No (2,355 muestras/ 265 individuos de diversos body sites).
  + **Después Filtros (Filtrado de las muestras procedentes de GUT, excepciones y réplicas técnicas):**
    - **N.º Muestras=Nº Individuos**: No (466 muestras/ 248 individuos).