

Reproducibility issues while using publicly available metagenomic datasets

Scott Eckert

06/07/2019

Liu Group

Data Reproducibility Bootcamp

Microbiome

- Human microbiome
 - Totality microbial DNA in human body
- 10:1 ratio of bacterial:human cells
- 1.3:1 ratio
 - 0.2kg



Why care about the microbiome?

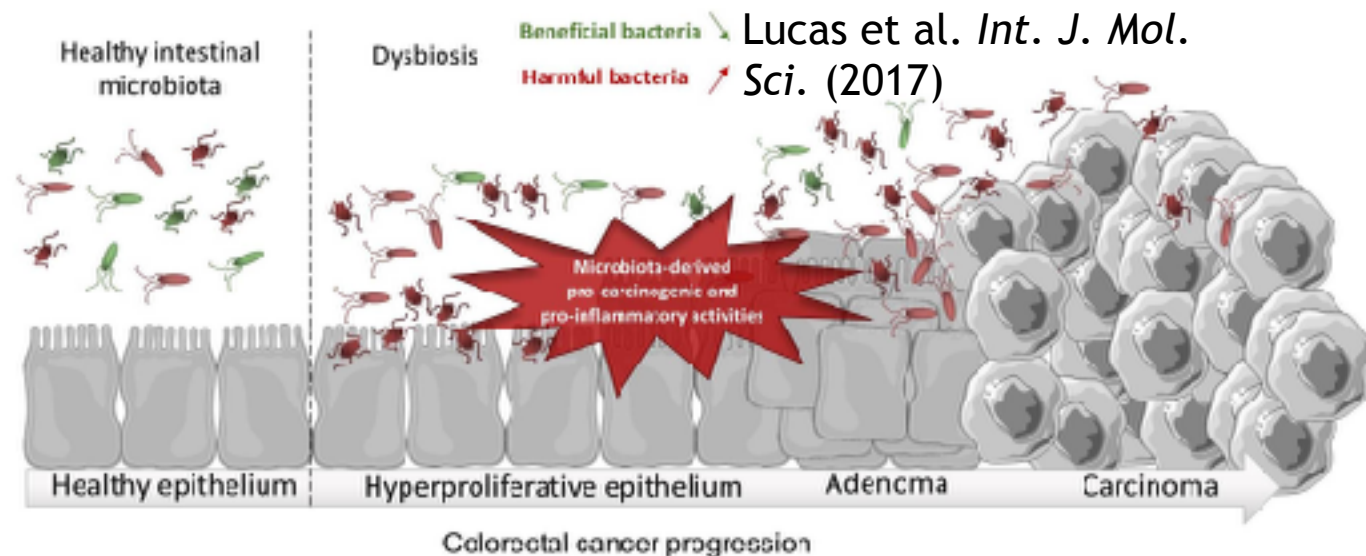
Why care about the microbiome?

ALMOST ALL CASES OF
CERVICAL CANCER
ARE CAUSED BY



HPV

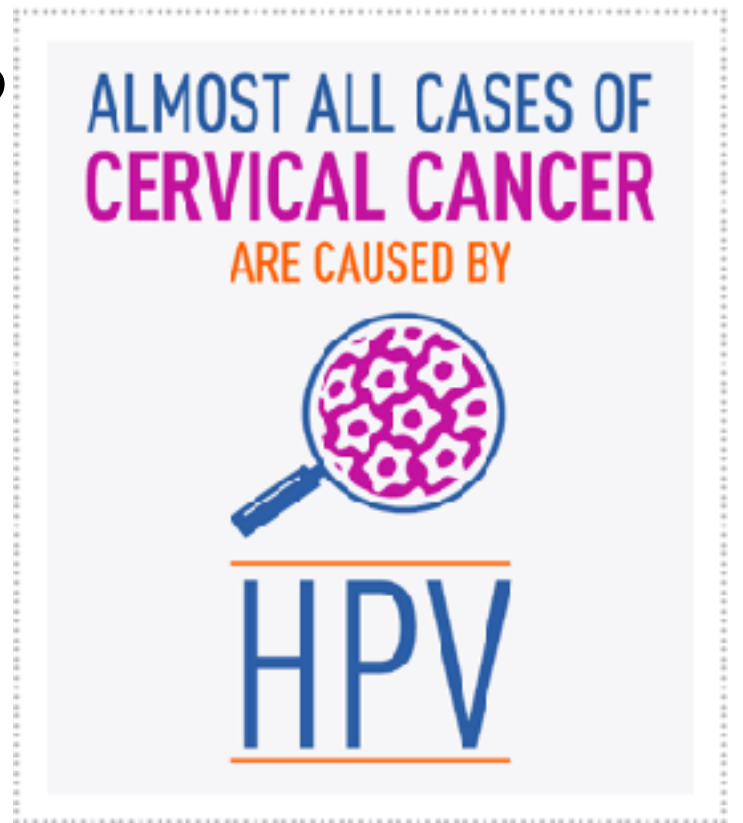
cancer.gov/hpv



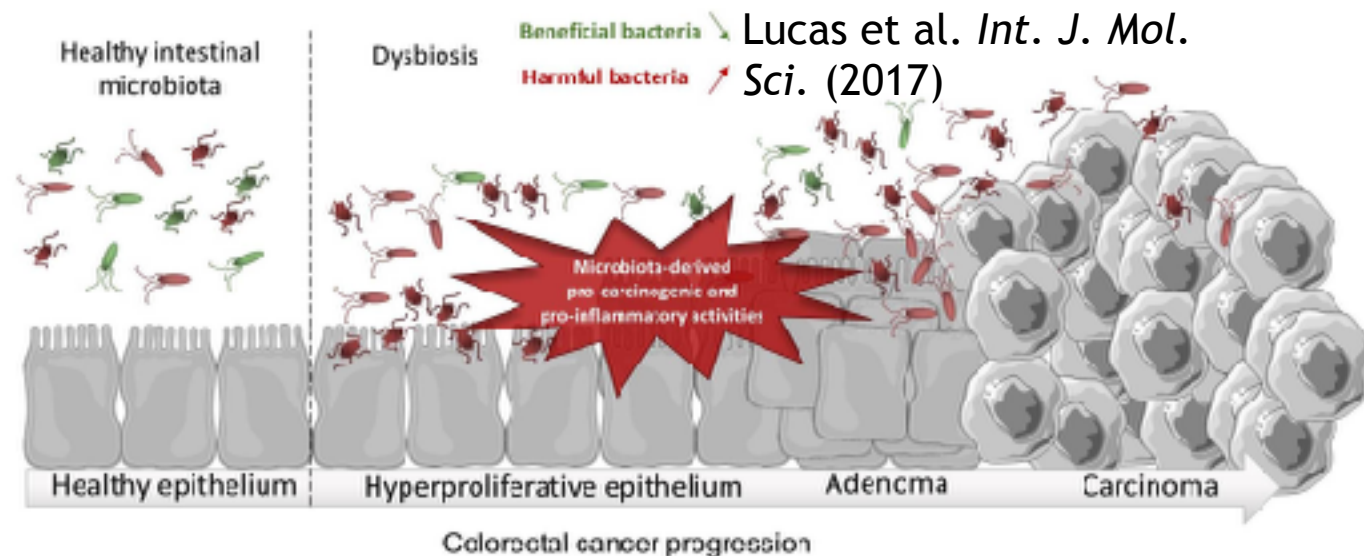
Why care about the microbiome?



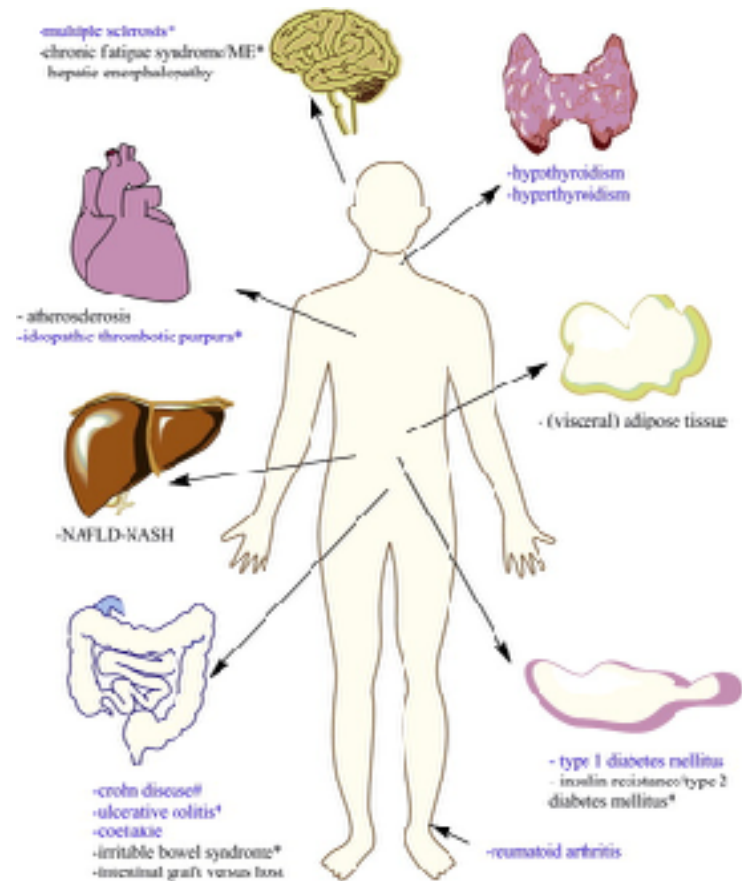
Howard F. Schwartz, Colorado State University,
Bugwood.org



cancer.gov/hpv



Why care about the microbiome?



A. Vrieze et al. *Best Pract Res Clin Gastroenterol* (2013)



Howard F. Schwartz, Colorado State University, Bugwood.org

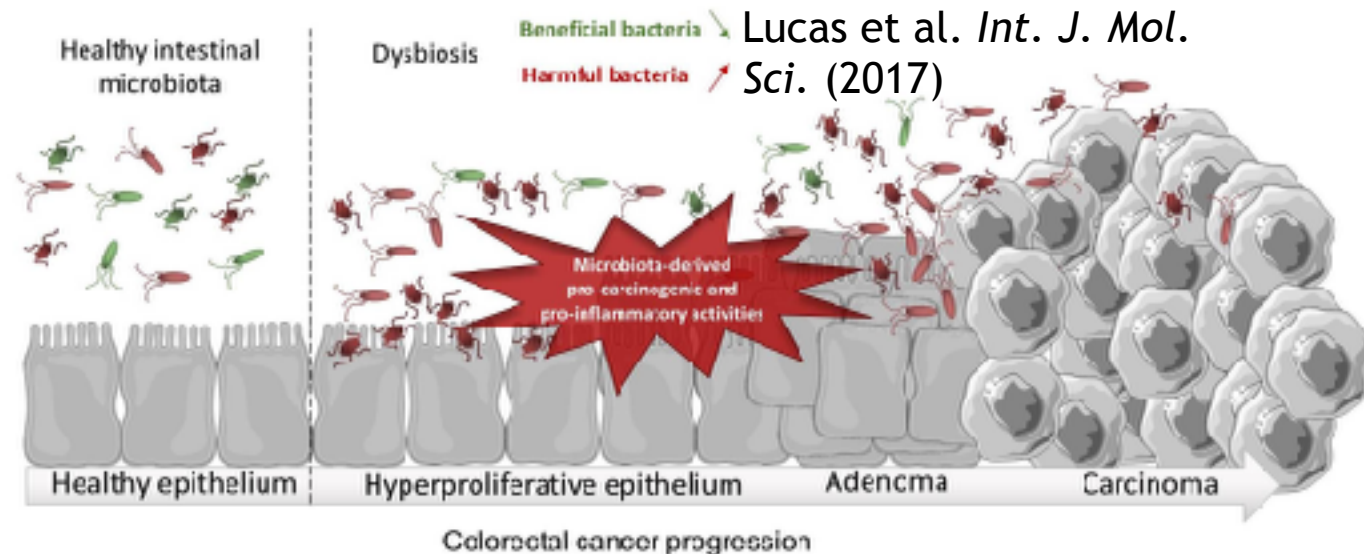
ALMOST ALL CASES OF
CERVICAL CANCER

ARE CAUSED BY



HPV

cancer.gov/hpv



Lucas et al. *Int. J. Mol. Sci.* (2017)

Challenge with microbiome studies

- Critiques:
 - Microbiome is variable, changes with a cup of coffee
 - Not enough phenotypic information



Enriching human microbiome datasets

- Critiques:

- Microbiome is variable, changes with a cup of coffee
- Not enough phenotypic information



- Ideas:

- Call host genotype from off-target WGS reads
- Generate additional related phenotypes/covariates

Published in final edited form as:

Nat Genet. ; 44(6): 631–635. doi:10.1038/ng.2283.

Extremely low-coverage sequencing and imputation increases power for genome-wide association studies

Bogdan Pasaniuc^{1,2,3,*}, Nadin Rohland^{3,4}, Paul J. McLaren^{3,5}, Kiran Garimella³, Noah Zaitlen^{1,2,3}, Heng Li³, Namrata Gupta³, Benjamin Neale³, Mark Daly³, Pamela Sklar⁶, Patrick F. Sullivan⁷, Sarah Bergen³, Jennifer L. Moran³, Christina M. Hultman⁸, Paul Lichtenstein⁸, Patrik Magnusson⁸, Shaun M. Purcell⁹, David W. Haas¹⁰, Liming Liang^{1,2,3}, Shamil Sunyaev^{3,5}, Nick Patterson³, Paul I.W. de Bakker^{3,5,11}, David Reich^{3,4,*}, and Alkes L. Price^{1,2,3,*}

- Random downsampling of 1000 Genomes Project
- Imputation from 0.24x off-target reads from WES
 - Comparable power GWAS from genotype array for common variants

Design



- Call genotypes for individuals in HMP, available studies
- Assess read depth, distribution and G/C content
 - Compare to genotype array data (available for ~300 individuals)
 - Identify other datasets with sufficient read depth



Pasoli et al. *Nature Methods* (2017)

Data acquisition

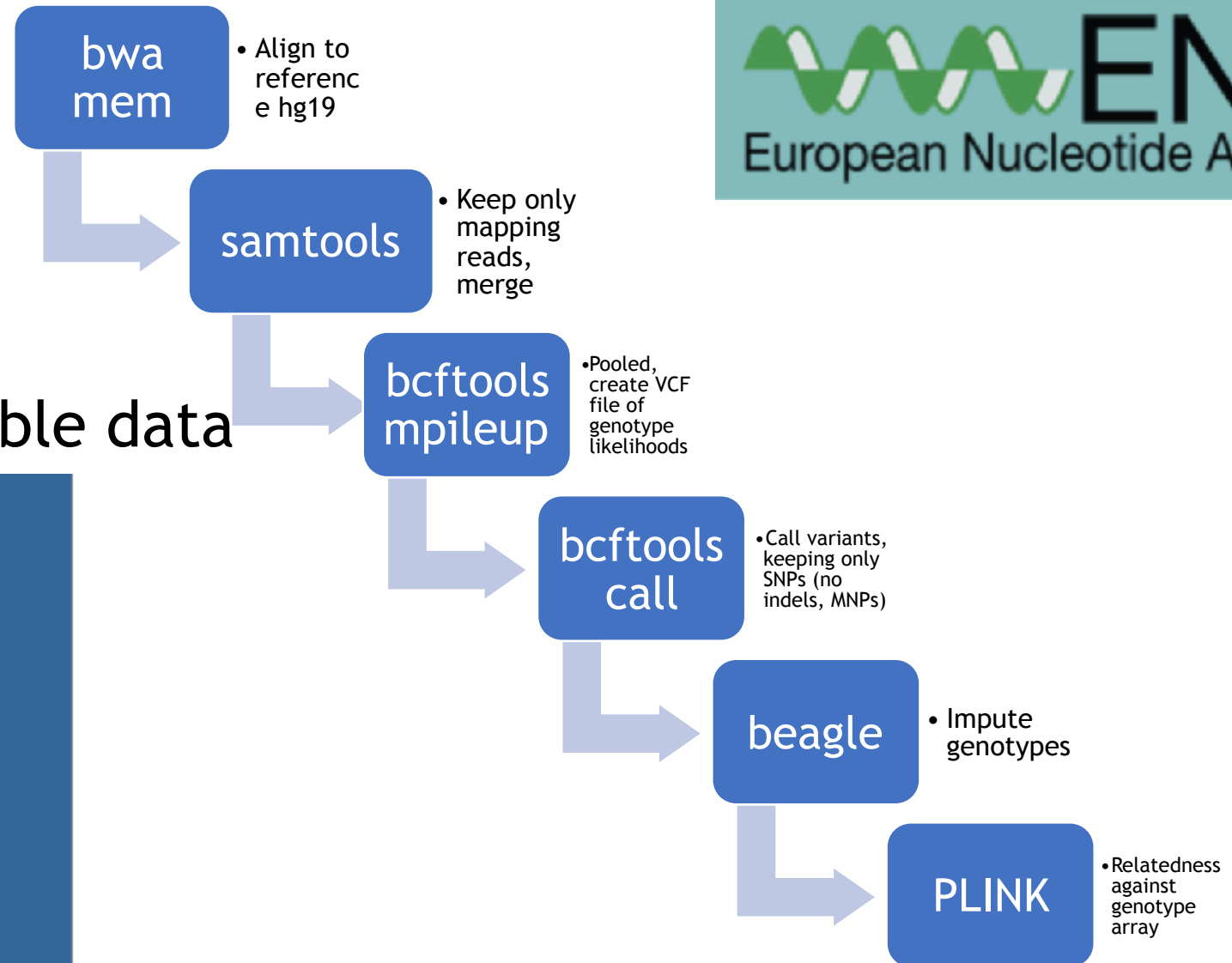
- Sequence Read Archive (SRA) from NCBI
 - SRA-toolkit
 - prefetch and fastq-dump
- European Nucleotide Archive (ENA) from EBI
 - No tool, but standard file paths from 7 or 8 digit



Pipeline



Use publicly available data



Results validation

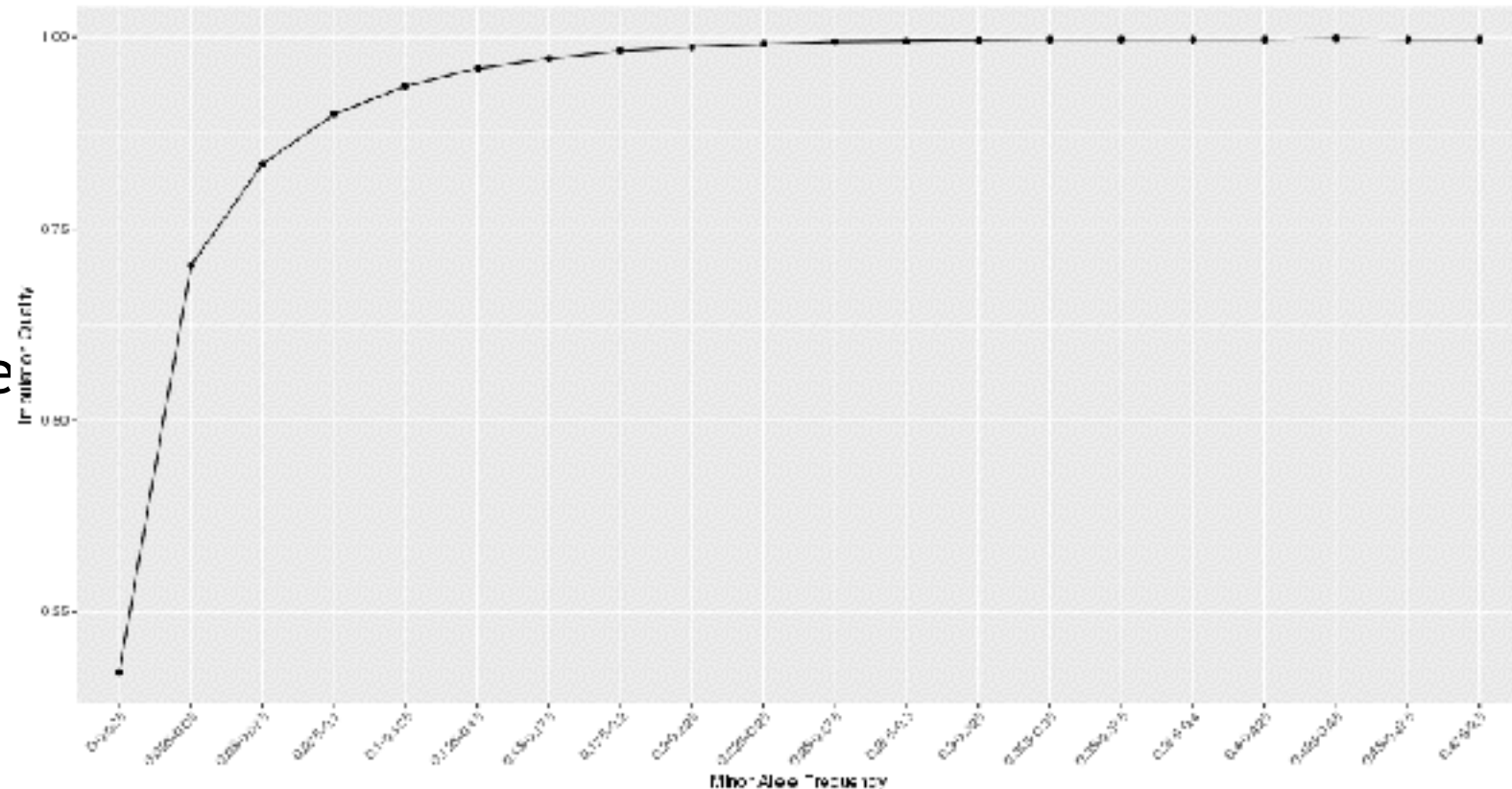


- HMP data structure
 - Many files per individual
 - Variable- different sites and number of replicates
 - Compare to array data
 - Cross validation

```
159814214 SRR060358;SRR060359;SRR059322;SRR059323;SRR062047;SRR062048;SRR062023;SRR062024;SRR059850;SRR059861;SRR060366;SRR060367;SRR061009;SRR062000;SR
763518768 SRR1804164;SRR1804782;SRR1804920;SRR1804951;SRR1804007;SRR1805950;SRR1804426;SRR1804164;SRR1804782;SRR1804320;SRR1804951;SRR1804007;SRR1805950
764325968 SRR061383;SRR061384;SRR046676;SRR061309;SRR061310;SRR061311;SRR061317;SRR061255;SRR062524;SRR061420;SRR061421;SRR061385;SRR061390;SRR063472;SR
909635352 SRR060021;SRR063903;SRR062904;SRR060020;SRR060032;SRR060033;SRR060034;SRR060003;SRR063907;SRR063908;SRR063909;SRR060005;SRR532192;SRR532218;SR
158357646 SRR531987;SRR531997;SRR531998;SRR532401;SRR532404;SRR532406;SRR532502;SRR533032;SRR539795;SRR539796;SRR539826;SRR539828;SRR539867;SRR539868;SR
763577454 SRR061138;SRR062321;SRR061386;SRR061407;SRR062470;SRR062473;SRR062353;SRR062357;SRR062380;SRR062413;SRR062306;SRR062308;SRR061388;SRR061411;SR
763860675 SRR063496;SRR063497;SRR061468;SRR061479;SRR061333;SRR061335;SRR0346684;SRR061370;SRR061391;SRR061326;SRR061368;SRR1804697;SRR061442;SRR061443;S
765216093 SRR1803874;SRR1804325;SRR1803879;SRR1804290;SRR1804435;SRR1804465;SRR1804458;SRR1804870;SRR1804429;SRR1804894;SRR513169;SRR514225;SRR511923;SR
765155902 SRR1804646;SRR1803967;SRR1804068;SRR1803966;SRR1804316;SRR1803965;SRR1804917;SRR1803950;SRR1803997;SRR1803900;SRR1804211;SRR513443;SRR514240;S
160704339 SRR059440;SRR059441;SRR059442;SRR059443;SRR059444;SRR059445;SRR059446;SRR059447;SRR059448;SRR059497;SRR061993;SRR061994;SRR063922;SRR063920;SR
993025008 SRR1031062;SRR1031068;SRR1031158;SRR1031265;SRR1031305;SRR1031421;SRR1031661;SRR1031940;SRR1031383;SRR1031483;SRR1031516;SRR1031653;SRR1031822
150470302 SRR532115;SRR532120;SRR532137;SRR532141;SRR532275;SRR532281;SRR532284;SRR532287;SRR532722;SRR532745;SRR532916;SRR532922;SRR532925;SRR532929;SR
368533040 SRR527884;SRR527916;SRR527952;SRR528002;SRR528034;SRR528087;SRR528156;SRR528212;SRR528230;SRR528337;SRR528367;SRR528418;SRR528516;SRR527884;SR
195044680 SRR1952551;SRR1952554;SRR1952557;SRR1952566;SRR1952567;SRR2241100;SRR2241110;SRR2241111
295137534 SRR532345;SRR532355;SRR532364;SRR532367;SRR532635;SRR532639;SRR532644;SRR532649;SRR532651;SRR533351;SRR533355;SRR533343;SRR533345;SRR533358;SR
868454789 SRR1031061;SRR1031266;SRR1031358;SRR1031359;SRR1031540;SRR1031834;SRR1031840;SRR1031886;SRR1031065;SRR1031149;SRR1031438;SRR1031651;SRR1031772
612472597 SRR1804002;SRR1804048;SRR1804090;SRR1804872;SRR1804308;SRR1804797;SRR1804813;SRR1804876;SRR1804132;SRR1804643;SRR1804049;SRR1804128;SRR1804433
160765029 SRR059500;SRR059501;SRR059538;SRR059539;SRR059542;SRR059543;SRR059548;SRR059549;SRR059556;SRR059557;SRR059518;SRR059519;SRR1564357;SRR059500;S
158328691 SRR1564226;SRR1564226
158418336 SRR2175723;SRR2175760
```


Quality check of imputed genotype

- Data from Castro-Nallar et al. *PeerJ* (2015)
- Oropharyngeal microbiome of SCZ patients compared to controls
- Ti/Tv ratio
 - 2.2:1
- One technical replicate
 - PLINK to check relatedness
 - Label mismatch



Make backups of your work

GitHub



skoote / MiRe Private

Watch 1 Unstar 1 Fork 0

Code Issues 0 Pull requests 0 Projects 0 Wiki Security Insights Settings

Microbiome Refinement [Edit](#)

[Manage topics](#)

● Shell 96.6% ● R 3.4%