# Reproducibility issues while using publicly available metagenomic datasets

Scott Eckert

06/07/2019

Liu Group

Data Reproducibility Bootcamp

# Microbiome

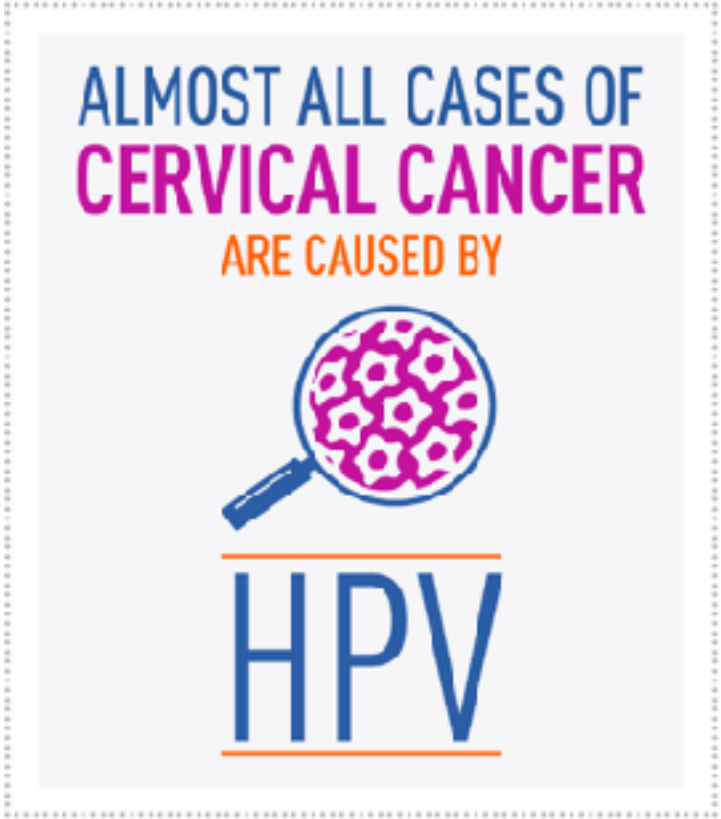
- Human microbiome
  - Totality microbial DNA in human boc
- 10:1 ratio of bacterial:human cell
- 1.3:1 ratio
  - 0.2kg

Savage et al. *Annu Rev Microbiol* (1977)

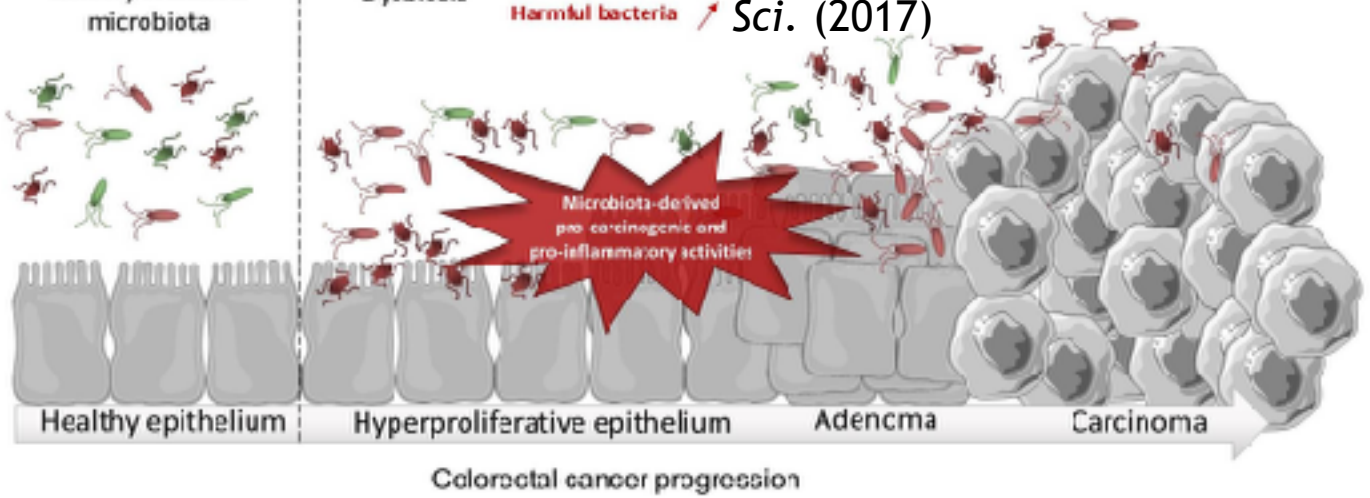
Sender et al. *PLoS Biol* (2016)

# Why care about the microbiome?

# Why care about the microbiome?



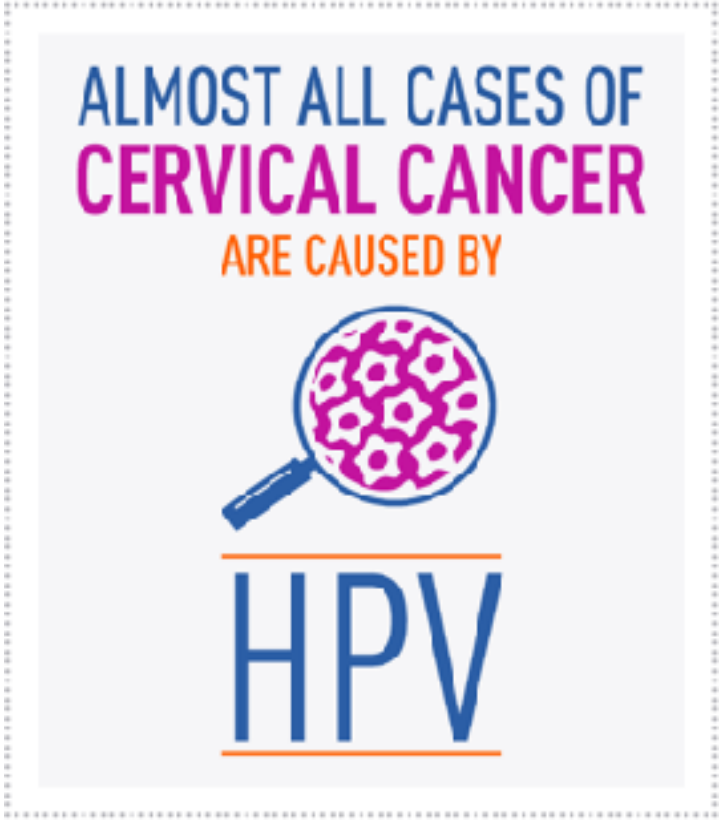

cancer.gov/hpv

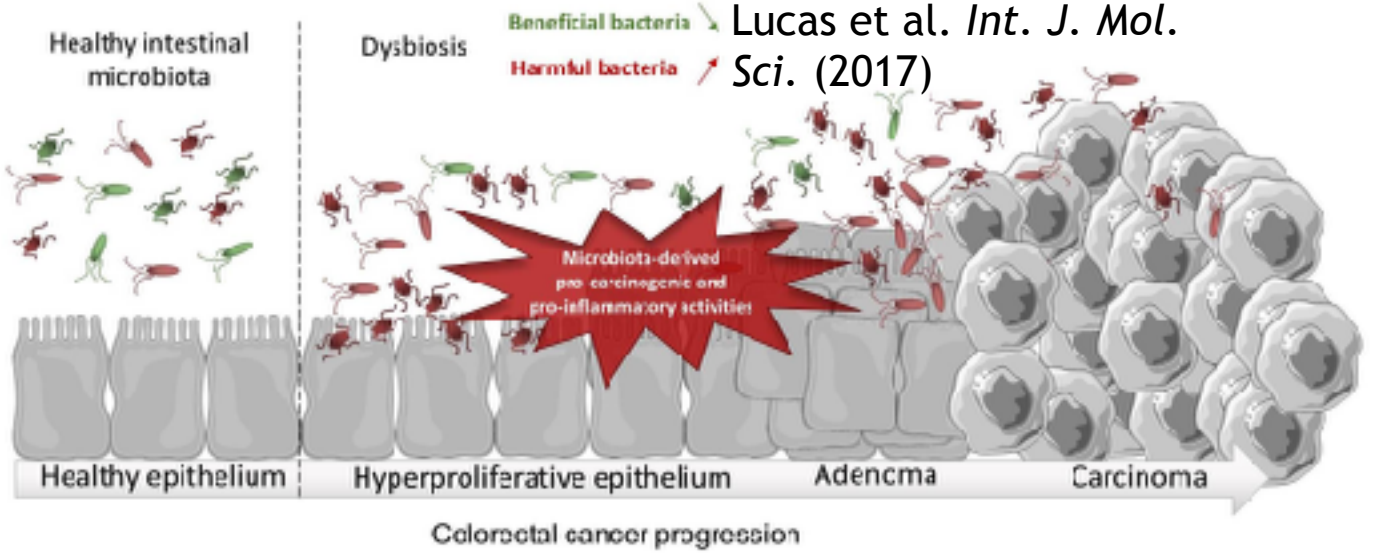



Lucas et al. *Int. J. Mol. Sci.* (2017)

# Why care about the microbiome?

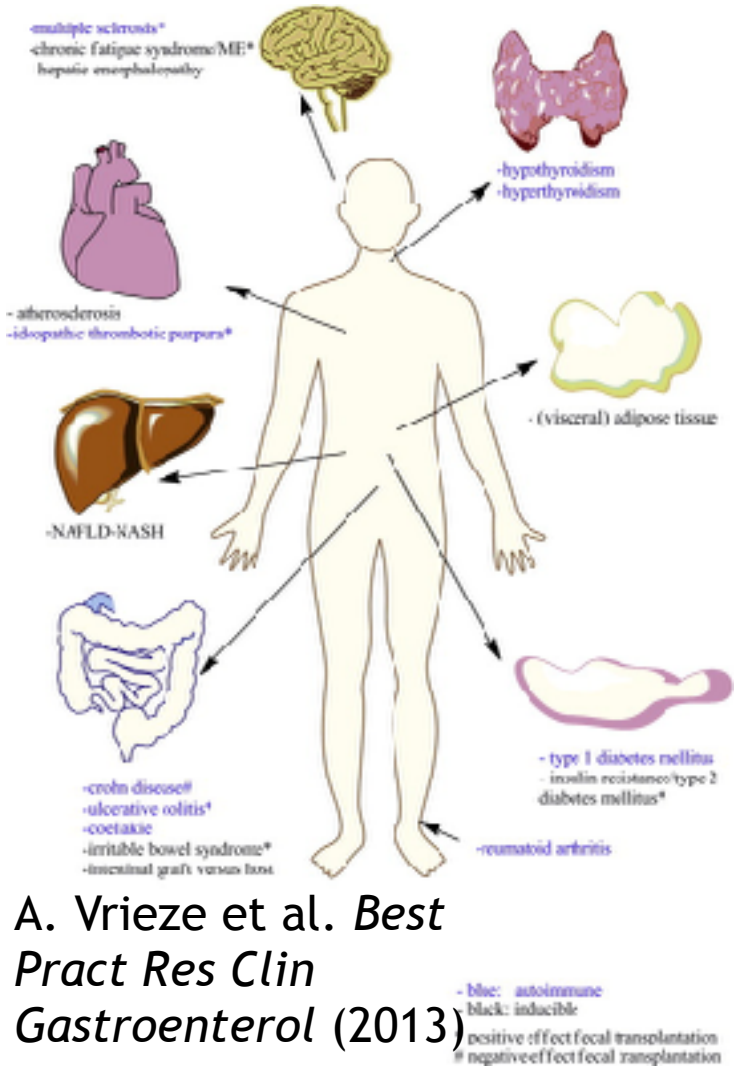

Howard F. Schwartz, Colorado State University, Bugwood.org



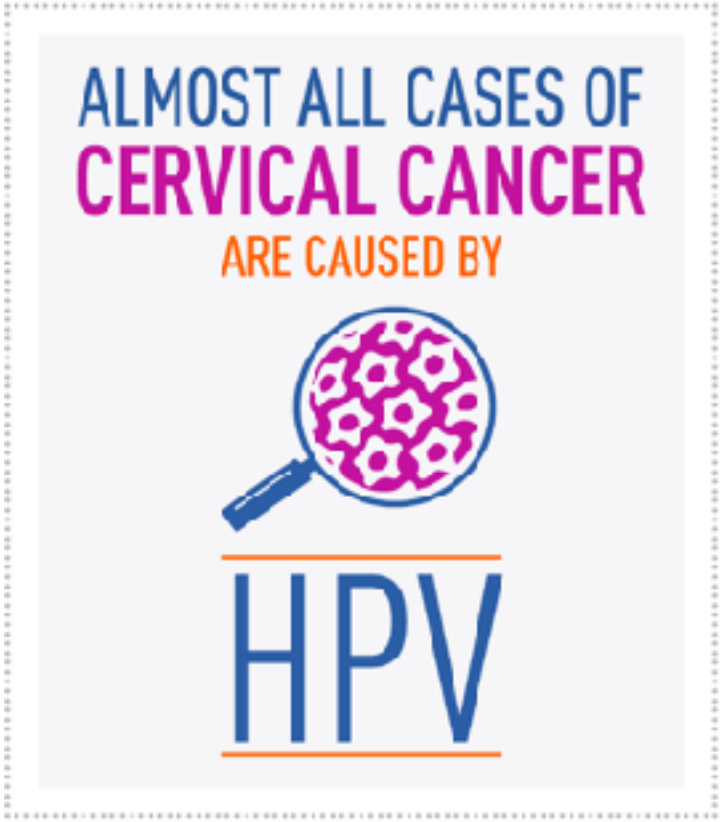

Lucas et al. *Int. J. Mol. Sci.* (2017)

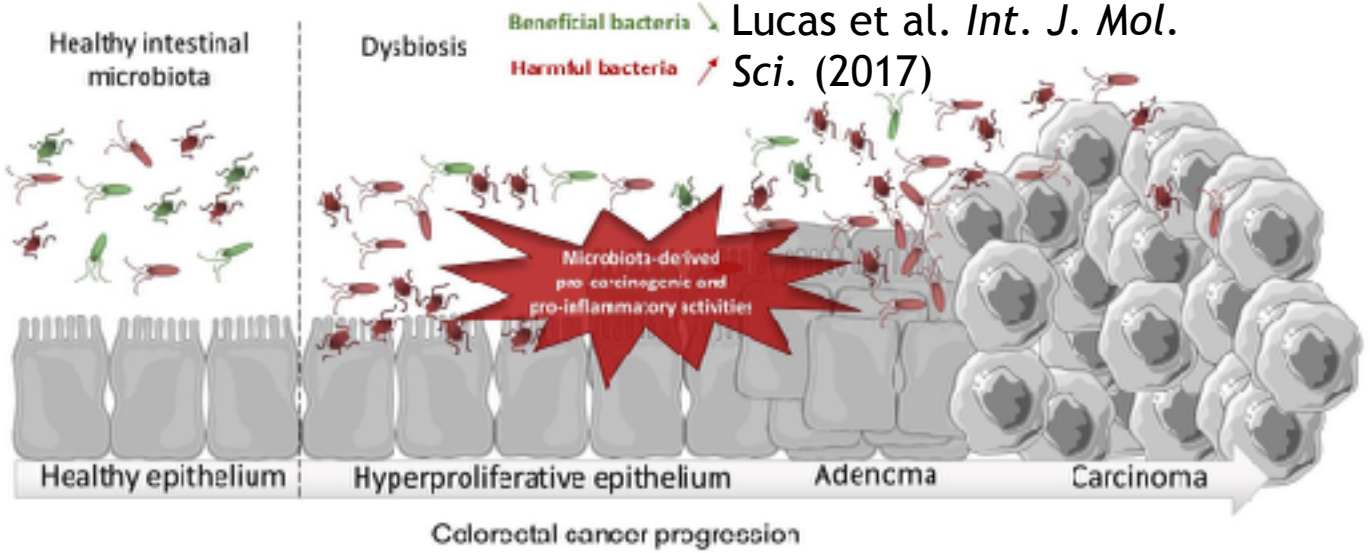
# Why care about the microbiome?




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


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




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 Garimella**[3], **Noah Zaitlen**[1,2,3], **Heng Li**[3], **Namrata Gupta**[3], **Benjamin Neale**[3], **Mark Daly**[3], **Pamela Sklar**[6], **Patrick F. Sullivan**[7], **Sarah Bergen**[3], **Jennifer L. Moran**[3], **Christina M. Hultman**[8], **Paul Lichtenstein**[8], **Patrik Magnusson**[8], **Shaun M. Purcell**[9], **David W. Haas**[10], **Liming Liang**[1,2,3], **Shamil Sunyaev**[3,5], **Nick Patterson**[3], **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

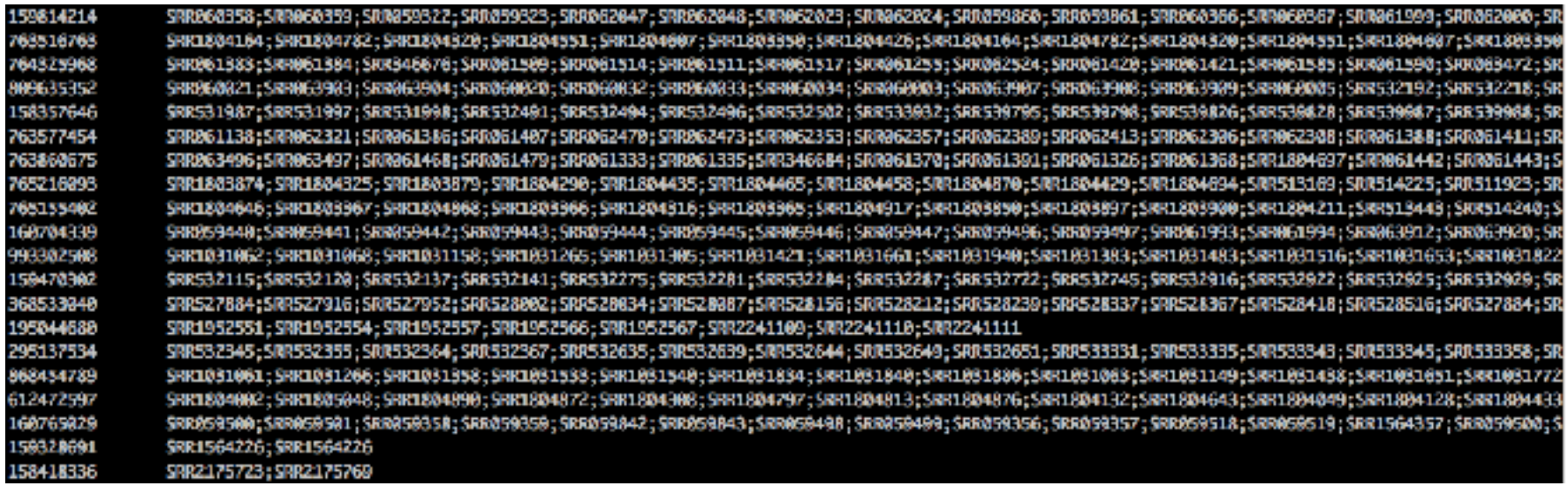
- Sequence Read Archive (SRA) from NCBI
  - SRA-toolkit
    - prefetch and fastq-dump
- European Nucleotide Archive (ENA) from EBI
  - No tool, but standard file paths from 7 or 8 digit


# Pipeline

Use publicly available data

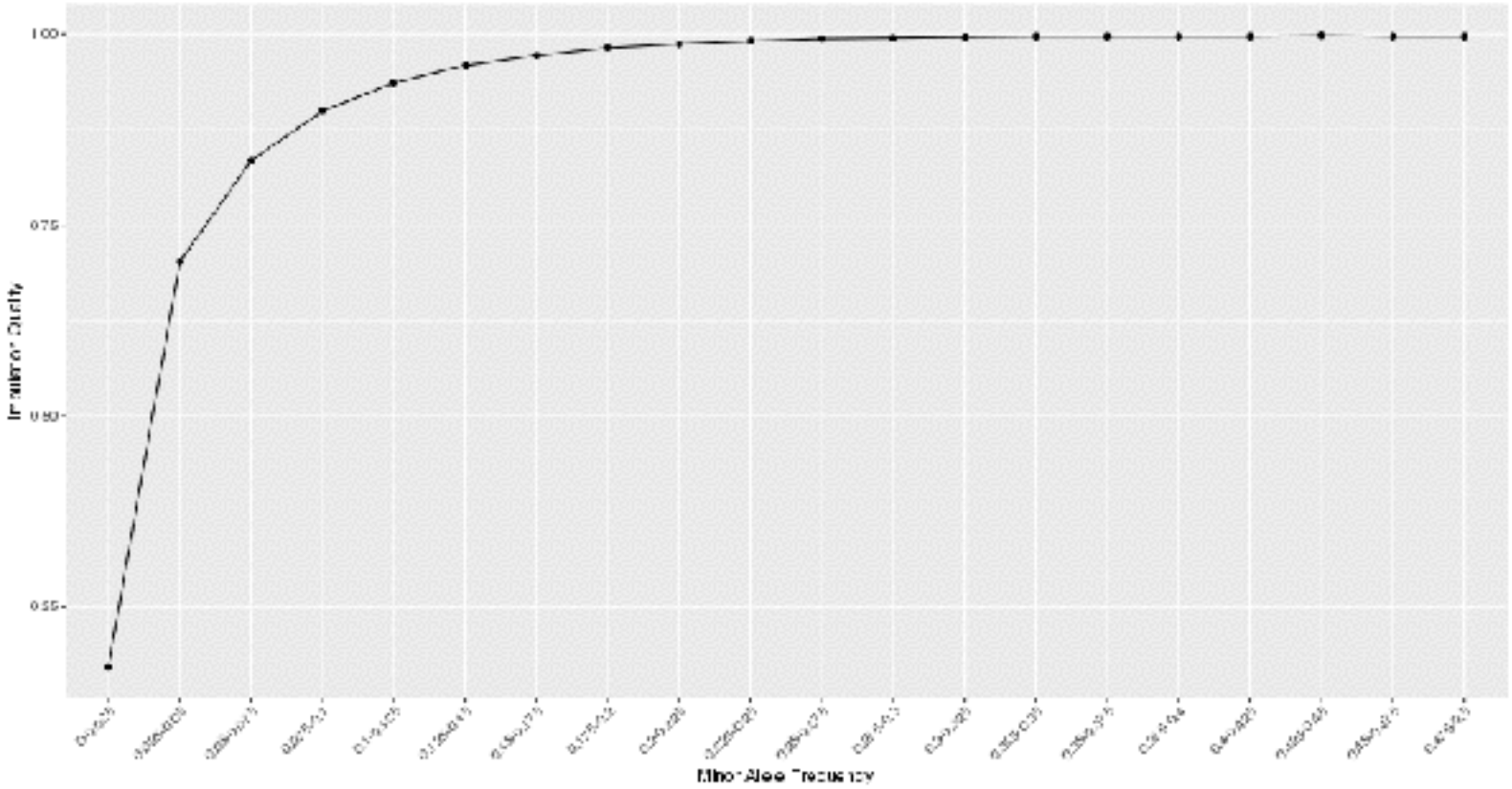

# Results validation

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



# 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 relatedne
  - 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%