a

Survey of published studies on human stool metagenomes

## Study exclusion criteria:

- · Medication or dietary intervention study
- · Less than 40 total samples
- Used a non-Illumina sequencer
- · Virome or VLP study
- · Early stage disease, rare genetic disorders, non-colon cancers

73 studies (12,957 samples)

Quality control of sequenced reads •

Taxonomic profiling using MetaPhlAn3 •

## Subject exclusion criteria:

- · Unclear mentioning of health status
- Abnormal body weight condition (i.e., underweight, overweight, obese)
- Heavy drug use or substance addiction (e.g., alcohol)
- > 100 years of age
- · Developed disease at an endpoint of a longitudinal study

## Metagenome sample exclusion criteria:

- Low read counts (< 1 million PE reads)
- High proportion of unmapped reads (> 90%)
- High proportion of unknown taxa (> 25%)
- Lacking sufficient taxonomic diversity (< 100 identified taxa)

57 studies (8,104 samples)

Study exclusion criteria:

Less than 20 total samples remaining in the study

## Pooled dataset of 8,069 stool shotgun metagenomes across 12 phenotypes and 54 studies

. colour databot of o,coo ottos onorgan motagonomos doloco la priorio, peo dila o i otadiso			
#	Phenotype	# of studies	# of samples
1	Healthy	46	5547
2	Ankylosing Spondylitis (AS)	1	95
3	Atherosclerotic Cardiovascular Disease (ACVD)	1	214
4	Colorectal Cancer (CC)	9	789
5	Crohn's Disease (CD)	7	284
6	Graves' Disease (GD)	1	100
7	Liver Cirrhosis (LC)	1	152
8	Multiple Sclerosis (MS)	1	224
9	Nonalcoholic Fatty Liver Disease (NAFLD)	1	86
10	Rheumatoid Arthritis (RA)	2	151
11	Type 2 Diabetes (T2D)	4	377
12	Ulcerative Colitis (UC)	6	250



