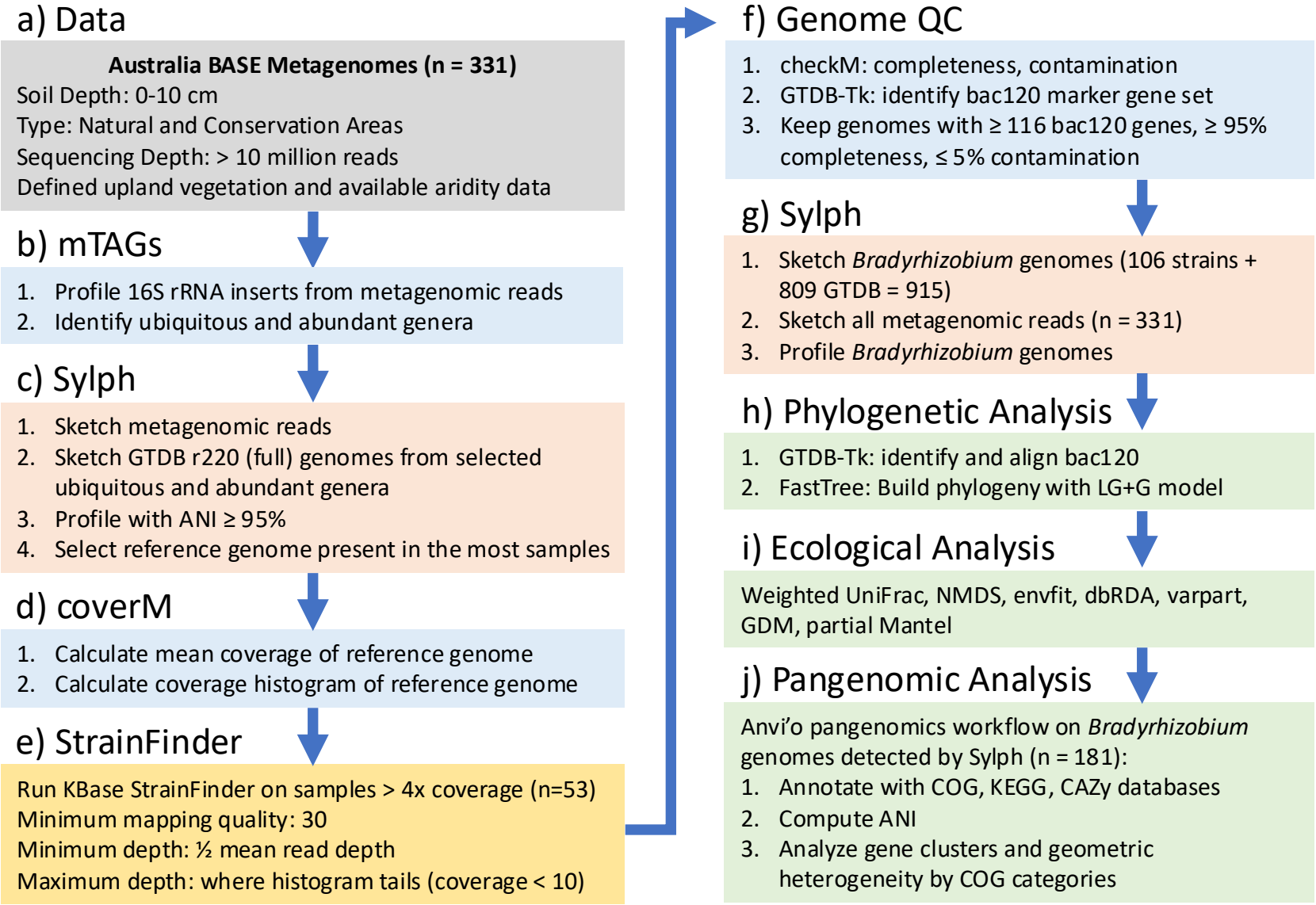
a) Data
Australia BASE Metagenomes (n = 331)
Soil Depth: 0-10 cm
Type: Natural and Conservation Areas
Sequencing Depth: > 10 million reads
Defined upland vegetation and available aridity data
b) mTAGs
1. Profile 16S rRNA inserts from metagenomic reads
2. Identify ubiquitous and abundant genera
c) Sylph
1. Sketch metagenomic reads
2. Sketch GTDB r220 (full) genomes from selected ubiquitous and abundant genera
3. Profile with ANI ≥ 95%
4. Select reference genome present in the most samples
d) coverM
1. Calculate mean coverage of reference genome
2. Calculate coverage histogram of reference genome
e) StrainFinder
Run KBase StrainFinder on samples > 4x coverage (n=53)
Minimum mapping quality: 30
Minimum depth: ½ mean read depth
Maximum depth: where histogram tails (coverage < 10)
f) Genome QC
1. checkM: completeness, contamination
2. GTDB-Tk: identify bac120 marker gene set
3. Keep genomes with ≥ 116 bac120 genes, ≥ 95% completeness, ≤ 5% contamination
g) Sylph
1. Sketch Bradyrhizobium genomes (106 strains + 809 GTDB = 915)
2. Sketch all metagenomic reads (n = 331)
3. Profile Bradyrhizobium genomes
h) Phylogenetic Analysis
1. GTDB-Tk: identify and align bac120
2. FastTree: Build phylogeny with LG+G model
i) Ecological Analysis
Weighted UniFrac, NMDS, envfit, dbRDA, varpart, GDM, partial Mantel
j) Pangenomic Analysis
Anvi'o pangenomics workflow on Bradyrhizobium genomes detected by Sylph (n = 181):
1. Annotate with COG, KEGG, CAZy databases
2. Compute ANI
3. Analyze gene clusters and geometric heterogeneity by COG categories