Overview of generic* metagenomics workflow When working with your own data you should never follow any pipeline blindly. There can be critical differences based on your data. *This is generic; specific workflows can vary on the order of steps here and how they are done. might be done by sequencing facility sequencing demultiplex (split samples by barodes) quality filter/trim fastq files fasta files facility (remove adapters/primers) fastqc/multiqc @HISEQ2500:282:1:1101:1220:1944 1 Some tools: Some tools: >HISEQ2500:282:1:1101:1220:1944 ATCGGATCG.. sabre ATCGGATCG. trimmomatic <G.<G<AGGII. · fastx demux (usearch/vsearch) · bbduk.sh (bbtools suite of tools) • idemp fastx barcode splitter (fastx-toolkit) read-based no-assembly path analysis consider testing assemblies with and w/o Some tools: Some tools: TIPP/SEPP assembly digital normalization bbnorm metaphlan2 path diainorm humann2 sourmash MetaQUAST is a great Count Table kraken tool for comparing assemblies obj_1 **Analysis** map individual sample reads to (co)-assembly Generate coverage obj_2 306 323 (co)-assembly 217 Some tools: information (mapping) phyloseq SpiecEasi Breakaway MaAsLin Some assemblers and tools: DivNet Megahit (assembler) DESeq2 Some tools: CORNCOB SPAdes (assembler) bowtie2 idba-ud (assembler) bwa MetAMOS (assembler and analysis pipeline) MetaCompass (reference-guided) Gene calling MetagenomeScope (visualize assembly graphs) Recovering genomes ____ A note on MAGs: Functional/taxonomic MAGs (metagenome-assembled genomes) are from metagenomes not the same thing as isolate genomes. They are composite representative genomes of closely profiling related genomic lineages. anvi'o (interactive manual curation of bins; and much more) Some common genomics stuff Some tools: CONCOCT (kmer-based and coverage-based binning; also incorporated in anvi'o) prodigal (identifies open reading frames) COCACOLA (kmer-based, coverage-based, and incorporates paired-read linkage of contigs) prokka (runs prodigal and performs annotations) MetaBAT2 (kmer-based and coverage-based binning tool) **Phylogenomics** GHOSTKOALA (web-hosted KEGG annotations) BinSanity (primarily coverage-based, optional second round kmer-based binning tool) BLAST (protein nr db/refseg/COGs) checkm (genome-level taxonomy; and much more) Comparative DASTool (a tool for evaluating bins recovered by different methods) DESMAN (tool aimed at resolving strains) genomics **Pangenomics** Env. distributions 1,002 1:1 ortholog 354,229 AA Some tools: · anvi'o (integrated HMMs for common single-copy gene sets; integrated pangenomic workflow for identifying orthologs via OrthoMCL) PanOCT (identifies orthologs utilizing synteny information) StrainPhlÀn/PanPhlAn (tools for strain-level analyses) MUSCLE (alignment software) FastTree (very fast, pseudo-maximum likelihood tree builder) RAxML (maximum likelihood tree builder) Mauve (whole-genome alignment)