Untargeted Metagenomics – Data Processing

Miguel Semedo

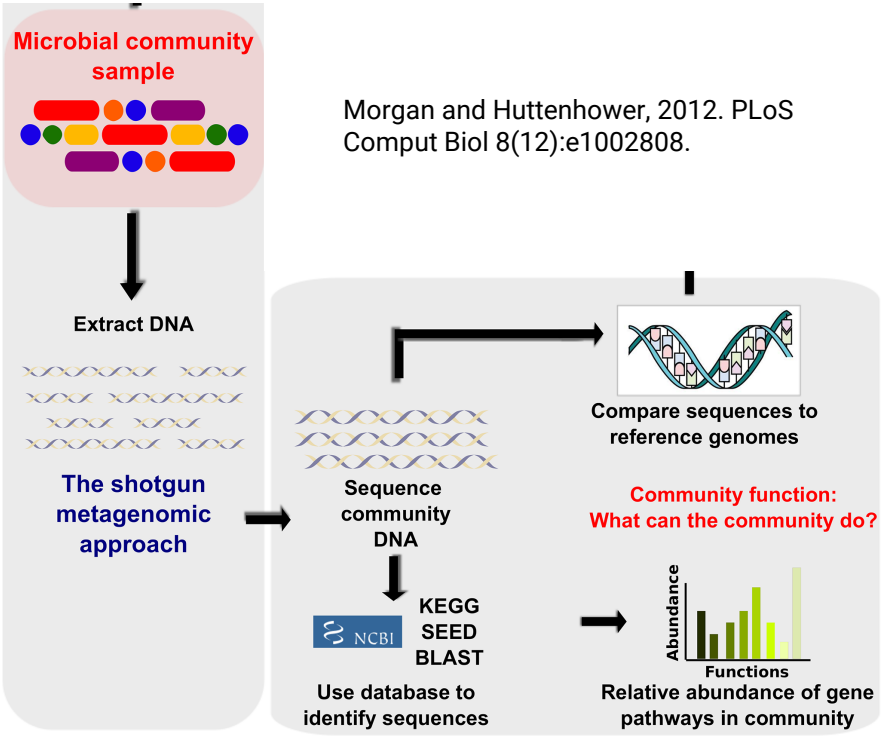
2024/09/05

Funding

Sponsorship

Support

Shotgun sequencing – What for?



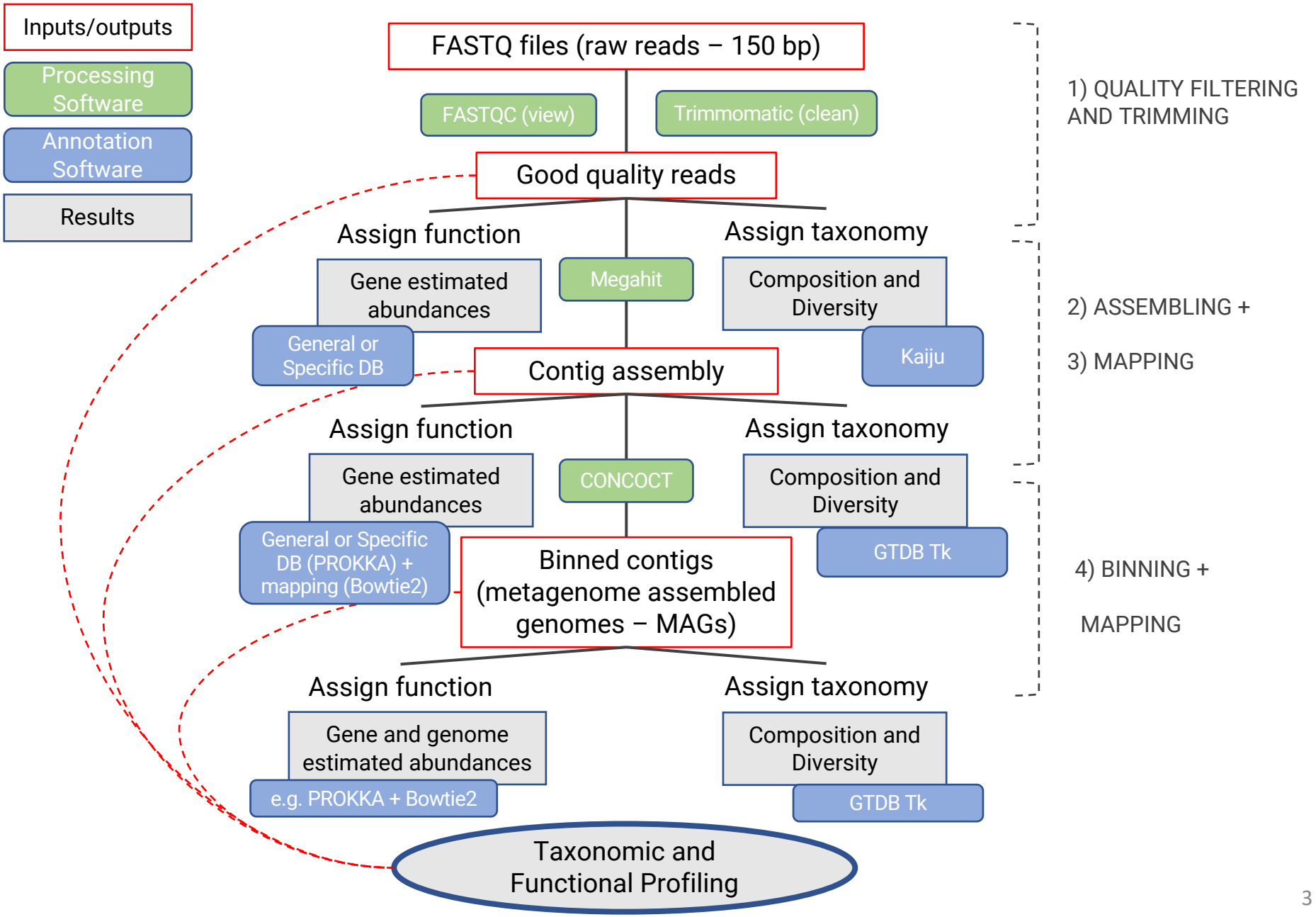
Morgan and Huttenhower, 2012. PLoS Comput Biol 8(12):e1002808.

Untargeted Metagenomics / Shotgun Sequencing / Whole Genome Sequencing

<u>All gene fragments from all organisms (no PCR amplification)</u>

- Who's there?
- What can they do (enzymes, metabolic pathways, biogeochemical cycling, among many others)? - **FUNCTIONS UNCHAINED**!

Shotgun sequencing workflow (from short reads)



Shotgun sequencing workflow (from short reads)

Data Processing

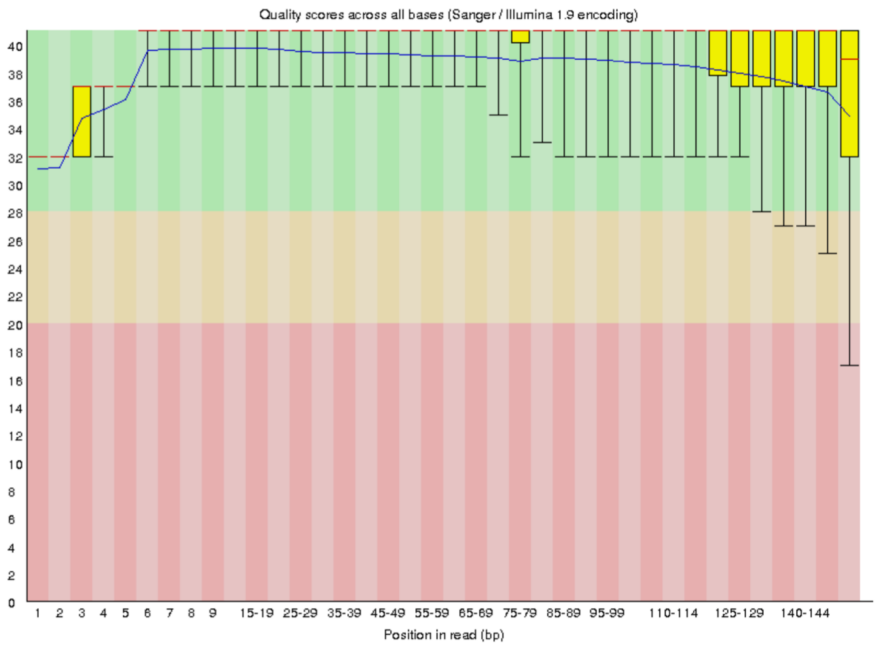
1. Quality filtering and trimming
2. Assembling
3. Mapping
4. Binning

1. Quality filtering and trimming

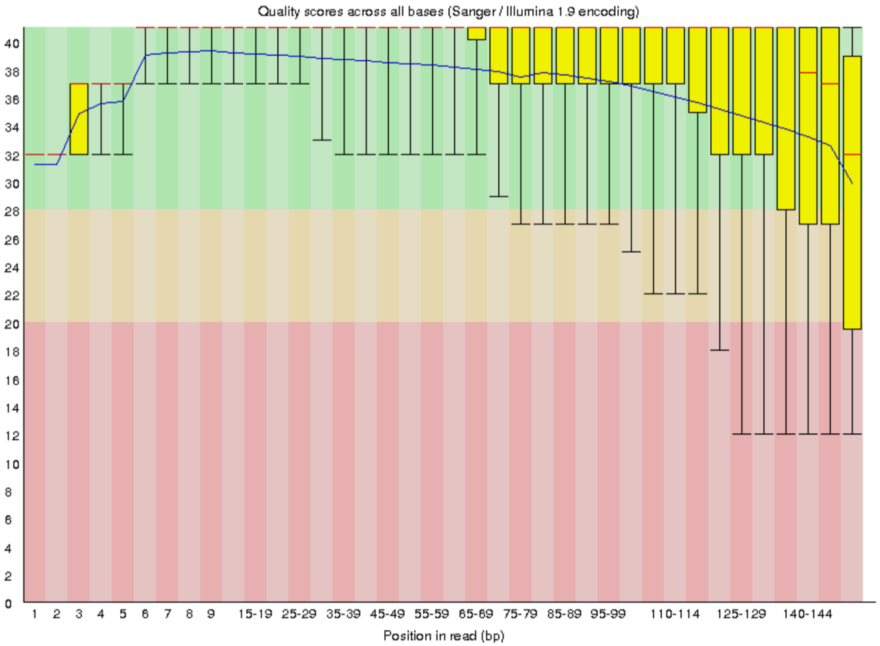
- **FASTQC** to inspect sequences quality (https://www.bioinformatics.babraham.ac.uk/projects/fastqc/). Easy installation.
- Using either own computer or HPC server (easy GUI available)
- `fastqc *fastq -o fastqc_results`
- Output is an html report (open R1 example)

Shotgun sequencing workflow (from short reads)

Per base sequence quality (Fwd)



Per base sequence quality (Rev)



Shotgun sequencing workflow (from short reads)

Data Processing

1. Quality filtering and trimming
2. Assembling
3. Mapping
4. Binning

1. Quality filtering and trimming

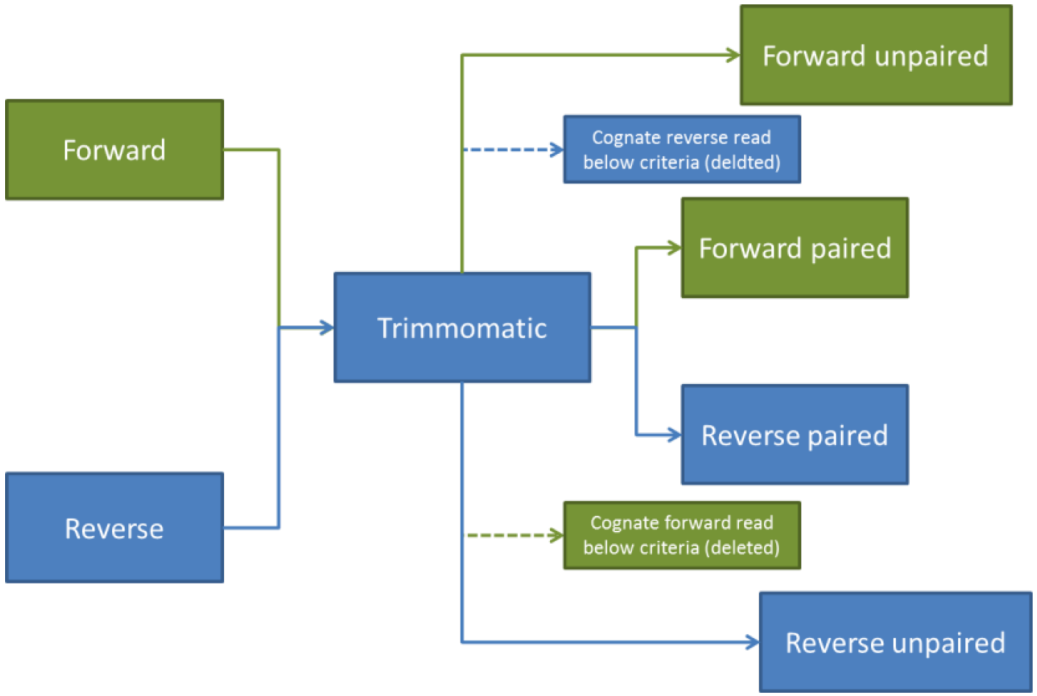
- **FASTQC** to inspect sequences quality (https://www.bioinformatics.babraham.ac.uk/projects/fastqc/). Easy installation.
- Using either own computer or HPC server (easy GUI available)
- `fastqc *fastq -o fastqc_results`
- Output is an html report (open R1 example)
- **TRIMMOMATIC** to remove low quality sequences, adapters, etc.
- `java -jar /usr/local/amd64/abu-dhabi/gcc/bio-programs/trimmomatic-0.33/trimmomatic-0.33.jar PE -threads 4 sample_name_R1_001.fastq sample_name_R2_001.fastq sample_name_1P_001.fastq sample_name_1U_001.fastq sample_name_2P_001.fastq sample_name_2U_001.fastq ILLUMINACLIP:nextera.fa:2:30:10:3:TRUE LEADING:3 TRAILING:3 SLIDINGWINDOW:4:15 MINLEN:36`

Shotgun sequencing workflow (from short reads)

1. Quality filtering and trimming

TRIMMOMATIC

- Removes low quality sequences, adapters, etc.
- Tool for removing low quality base-pairs and adapter sequences
- Java programming language
- Author: Anthony Bolger (USADELLAB)
- Input: FASTQ files



Shotgun sequencing workflow (from short reads)

1. Quality filtering and trimming

TRIMMOMATIC

- `java -jar /usr/local/amd64/abu-dhabi/gcc/bio-programs/trimmomatic-0.33/trimmomatic-0.33.jar PE -threads 4 sample_name_R1_001.fastq sample_name_R2_001.fastq sample_name_1P_001.fastq sample_name_1U_001.fastq sample_name_2P_001.fastq sample_name_2U_001.fastq ILLUMINACLIP:nextera.fa:2:30:10:3:TRUE LEADING:3 TRAILING:3 SLIDINGWINDOW:4:15 MINLEN:36`

- ILLUMINACLIP: removes illumine adapter and other specific sequences
- LEADING: cut bases off the start of a read, if below a threshold quality
- TRAILING: cut bases off the END of a read, if below a threshold quality
- SLIDINGWINDOW: Performs a sliding window trimming approach. Starts scanning at the 5’ end and clips the read once the average quality within the window falls below a threshold.
- MINLEN: remove reads below a specific threshold
- HEADCROP: remove specified number of bases at the start of read

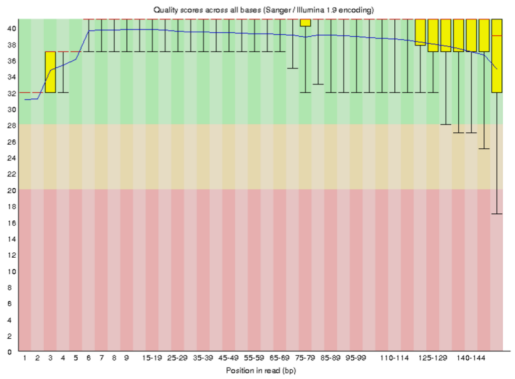
After trimming, always check FASTQC on trimmed samples!

Shotgun sequencing workflow (from short reads)

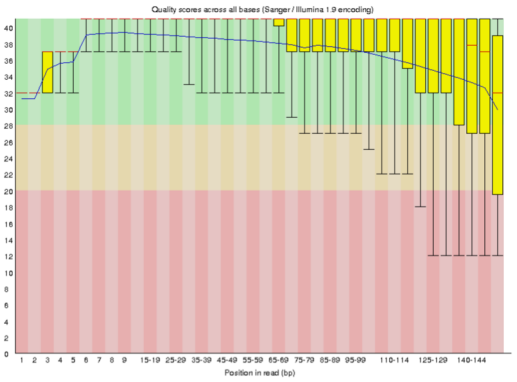
1. Quality filtering and trimming

RAW READS

Fwd (28,485,424)

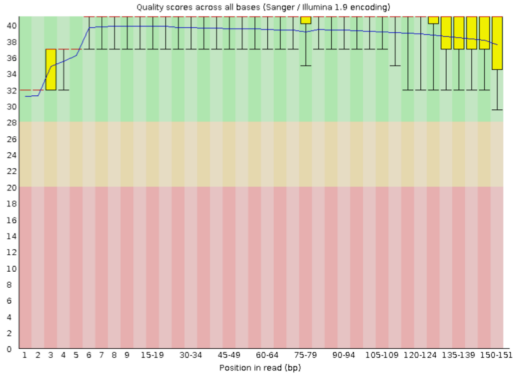


Rev (28,485,424)

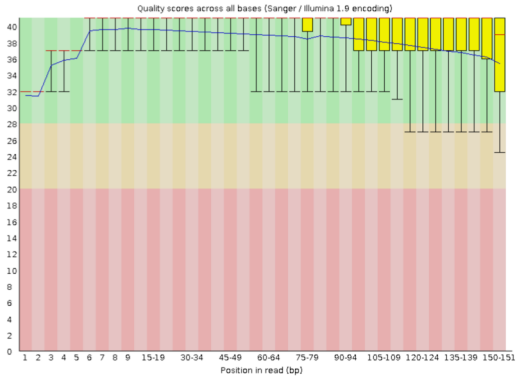


AFTER TRIMMOMATIC

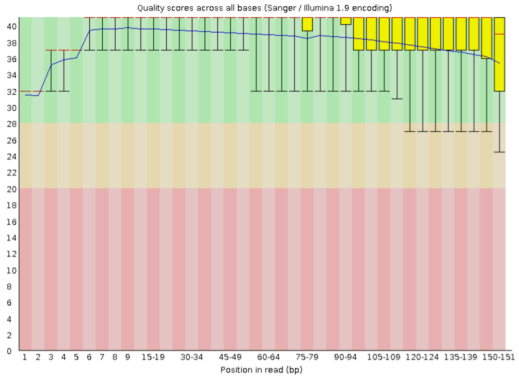
Fwd P (27,559,917)



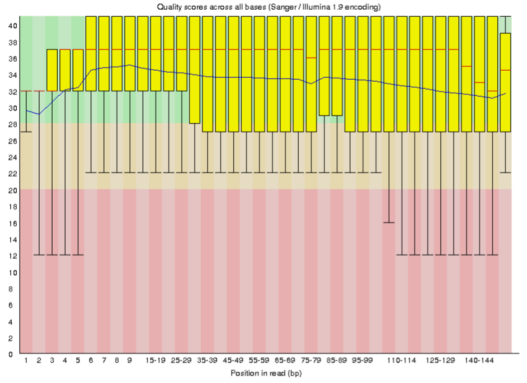
Fwd U (794,850)



Rev P (27,559,917)



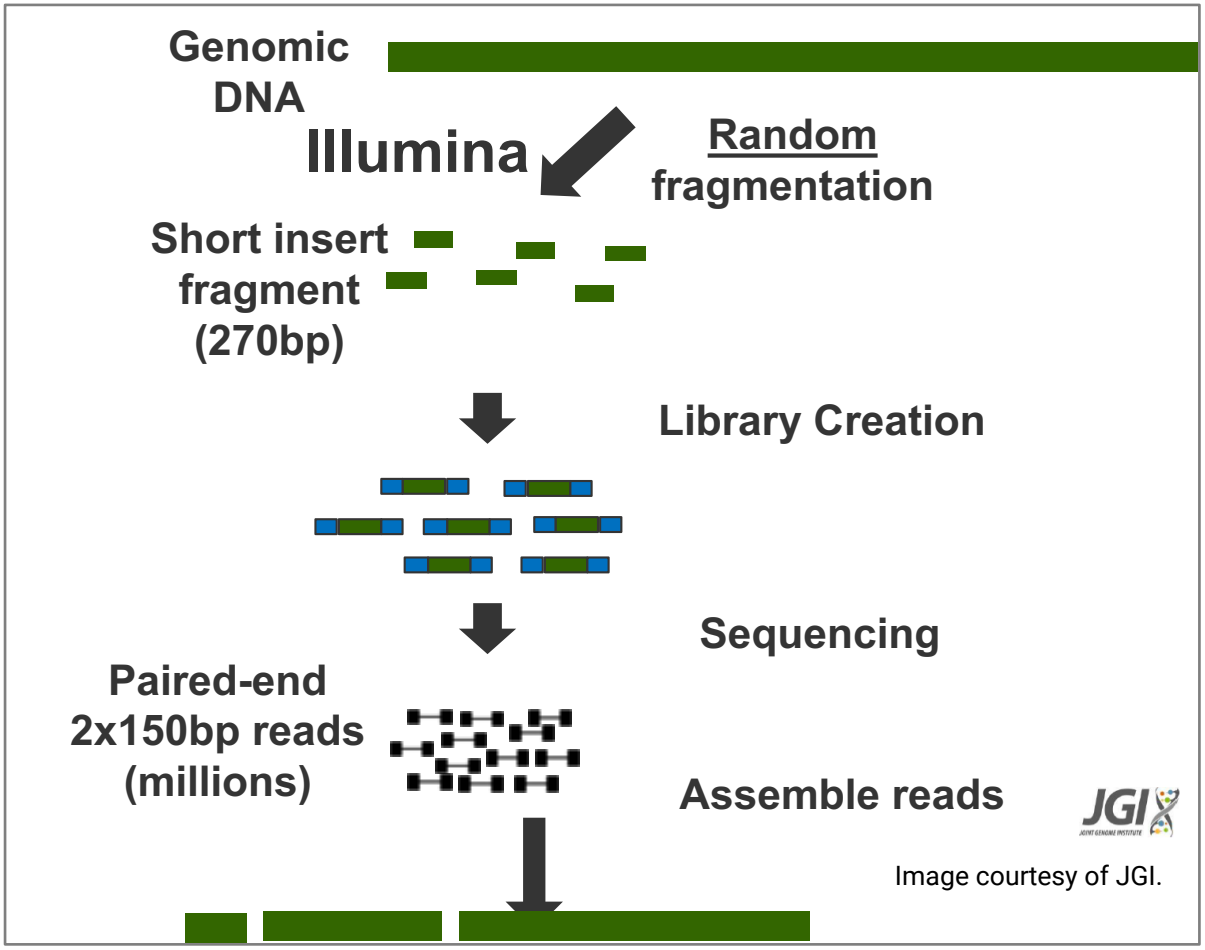
Rev U (80,012)



Shotgun sequencing workflow (from short reads)

2. Assembling

Generating longer reads (contigs), increasing taxonomic and functional resolution, downstream.

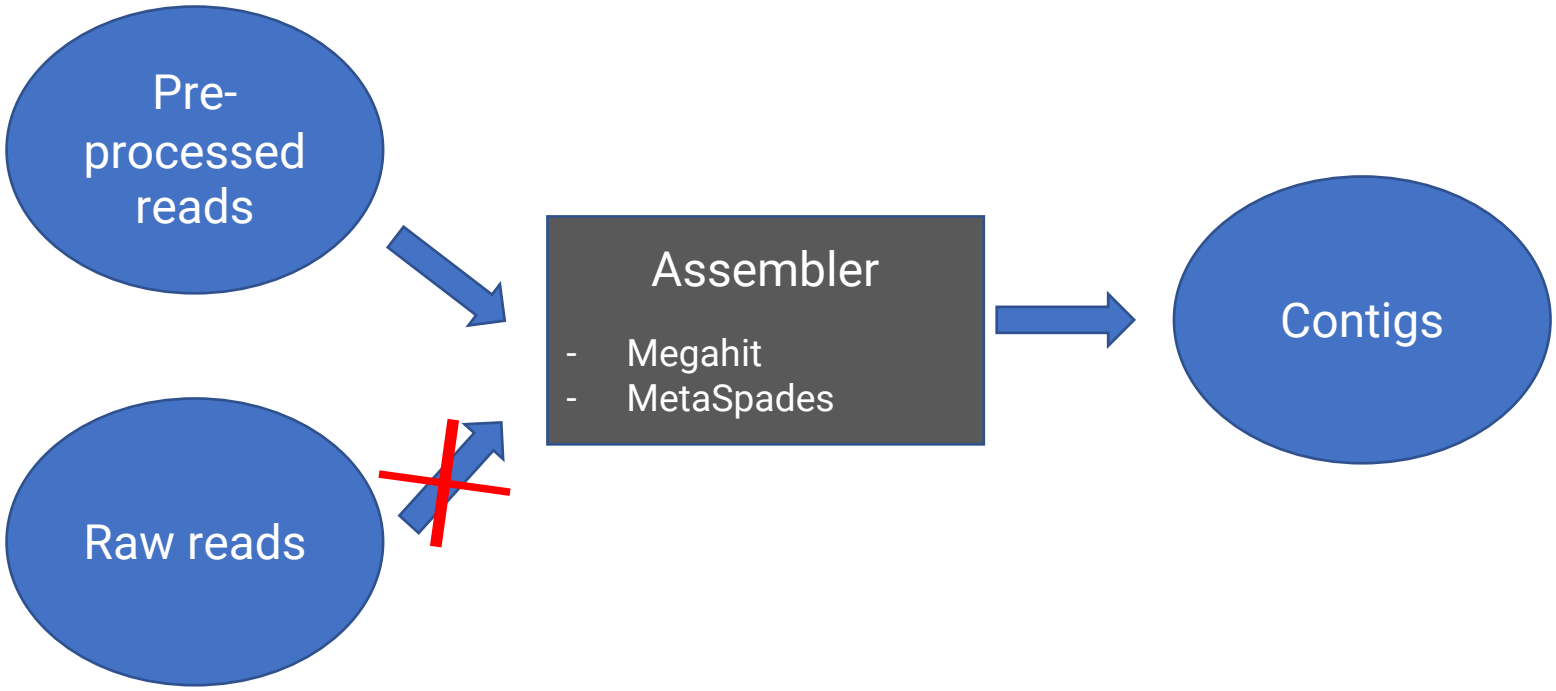


Assembling output: **contigs**

Shotgun sequencing workflow (from short reads)

2. Assembling

Assembling is very computationally intensive. ”Feed it” properly.




```
megahit --presets meta-large --min-contig-len 500 -1
sampleID_1P_001.fastq -2 sampleID_2P_001.fastq -o
megahit_out_sample_500_24 -m 0.9
```

Shotgun sequencing workflow (from short reads)

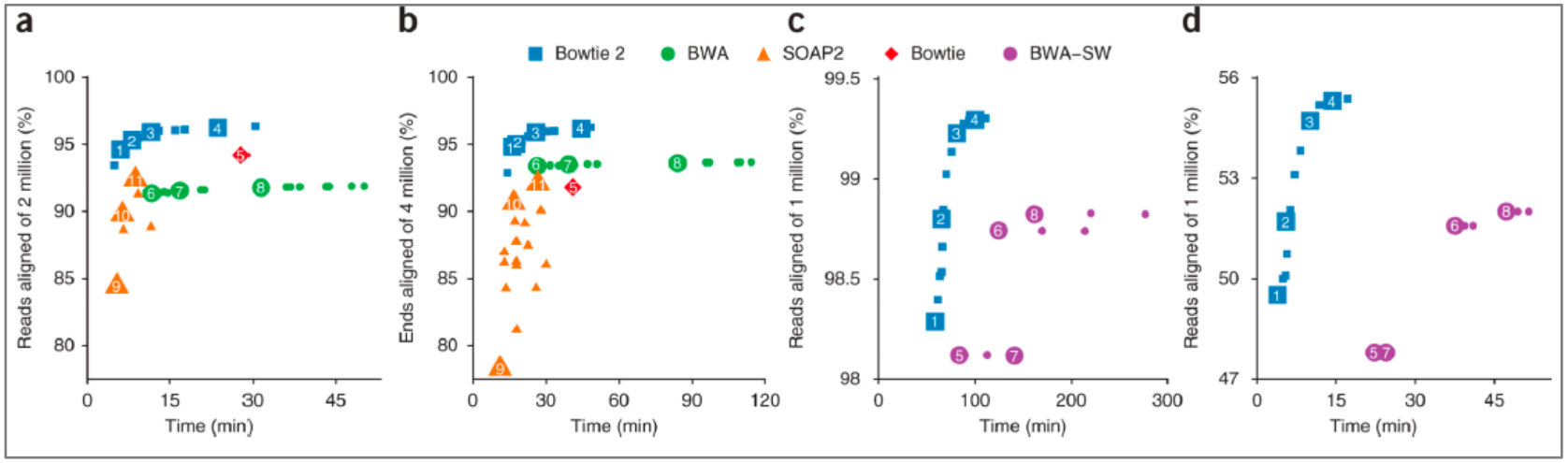
3. Mapping

Mapping/Aligning the clean reads into assembling contigs (or genomes) and estimate contig coverage. Can be used to estimate gene abundances (genes within contigs).

Bowtie2 (others: BWA, SOAP2, etc.)

- Tool for mapping/aligning short reads to longer fragments (contigs or genomes)
- Inputs: Paired reads and reference genome (contigs)
- Outputs: an alignment SAM file

Faster than other commonly used aligners for this read type.

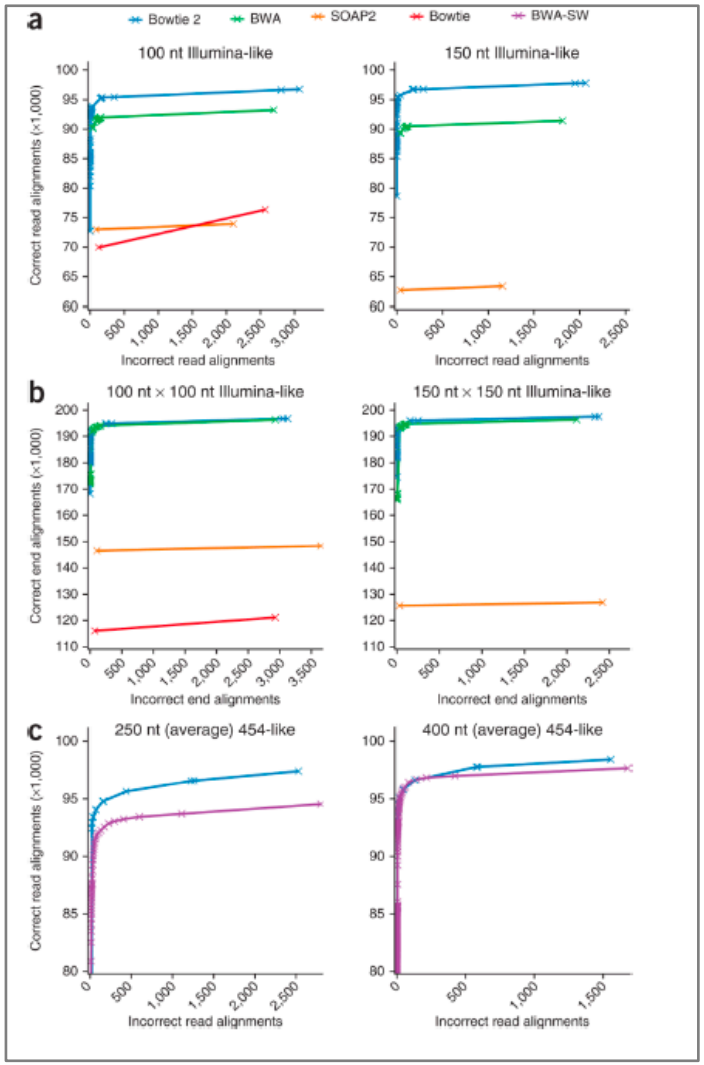


Langmead and Salzberg, 2012, Nat Methods

Shotgun sequencing workflow (from short reads)

3. Mapping

Also more accurate.



Indexing a reference genome (contigs):

```
bowtie2-build final.contigs.24.fa final.contigs.24
```

Aligning the paired reads to the reference genome (contigs). Computationally intensive step

```
bowtie2 --local -N 1 -x final.contigs.24 -1 1735D-62-24_S0_L001_1P_001.fastq -2 1735D-62-24_S0_L001_2P_001.fastq -S contig.local3.alignment.24.sam
```

Converting the sam file to bam so we can count the reads aligned to each contig

```
samtools view -S -b contig.local3.alignment.24.sam > sample24.bam
```

Now the alignment needs to be sorted (30 min/sample with 40 processors).

```
samtools sort sample24.bam -o sample24.sorted.bam
```

Summarizing the indexed bam file (mapped reads per contig)

```
samtools idxstats sample24.sorted.bam | tee sample.24.counts.txt
```

Output: number of mapped reads per contig

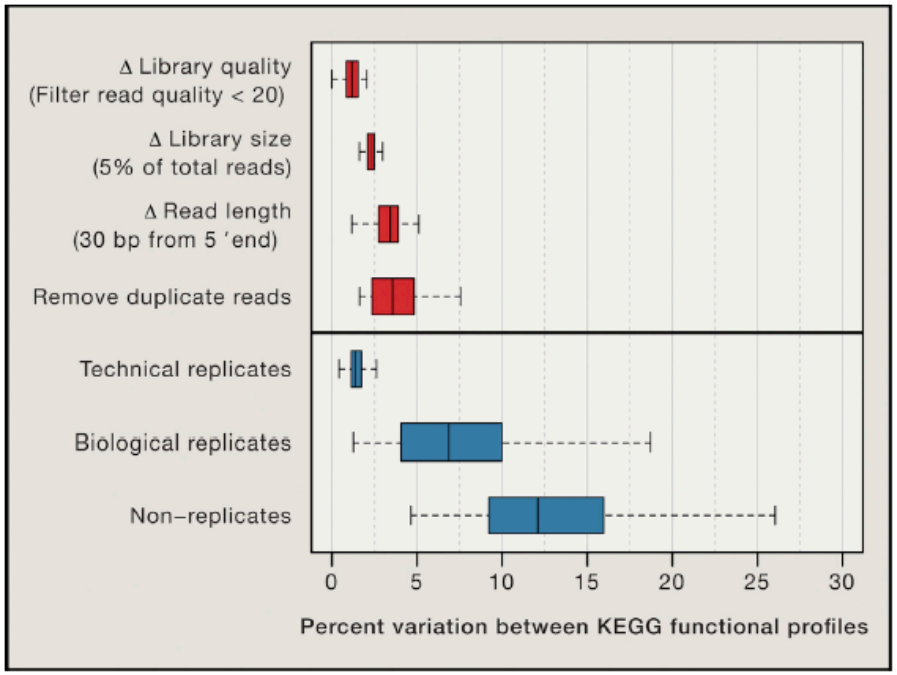
Langmead and Salzberg, 2012, Nat Methods

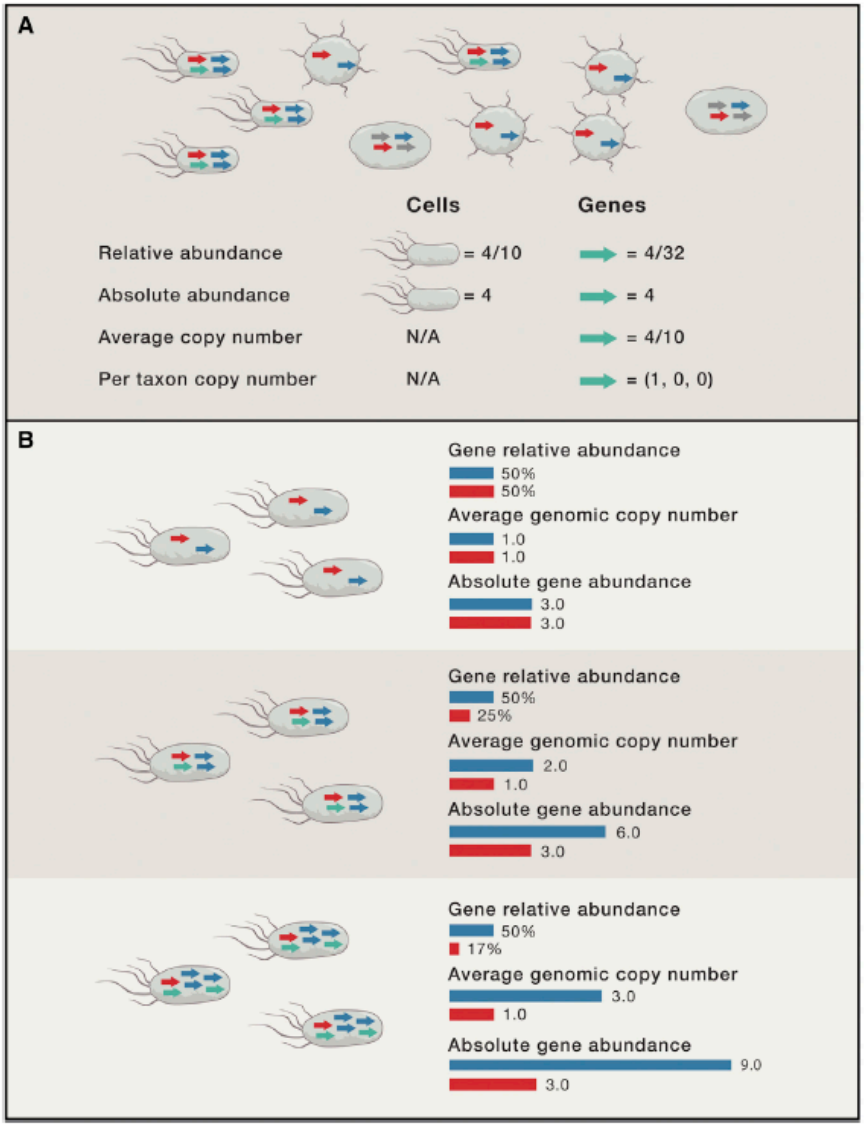
Shotgun sequencing workflow (from short reads)

3. Mapping

Mapping reads to contigs and/or genes can be used to estimate gene abundances

Think carefully about ”abundance” units or normalizing methods you want to use for your research question (**gene vs. gene**, **sample vs sample** comparisons?). Check Nayfach and Pollard, 2016, *Cell*, among others).





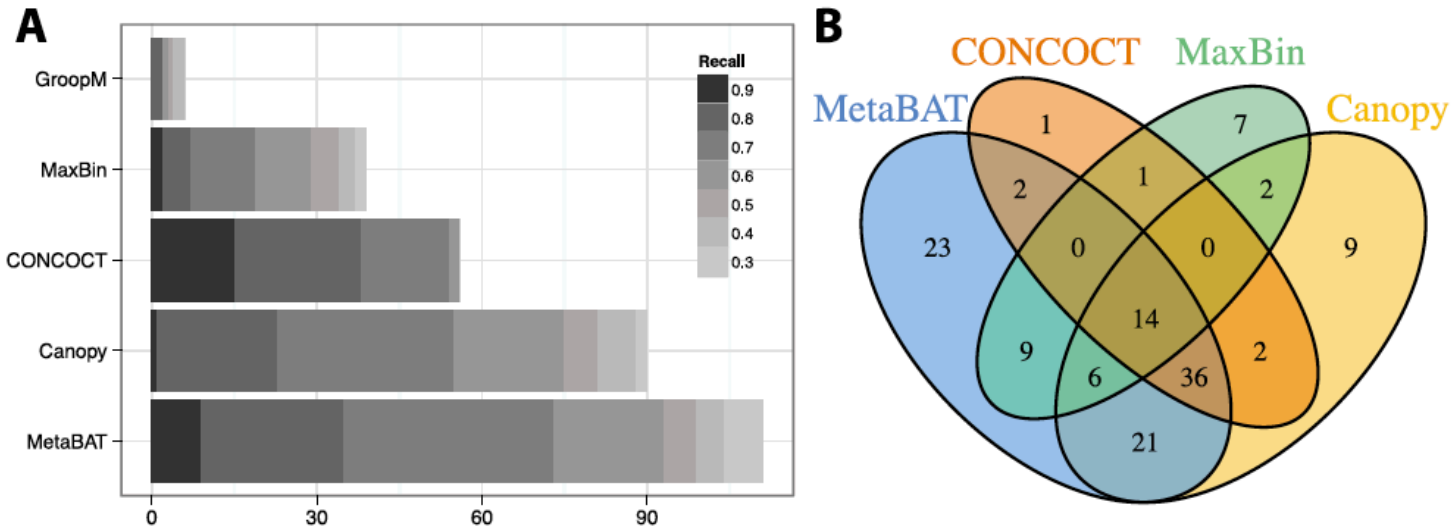
Shotgun sequencing workflow (from short reads)

4. Binning

Binning assembled contigs into putative single genomes (metagenome assembled genomes – MAGs)

MetaBAT2 (others: CONCOCT, MaxBin2, etc.)

- Uses nucleotide composition information and source strain abundance (measured by depth-of-coverage by aligning the reads to the contigs) to perform binning.
- Inputs: Metagenome assemblies (contigs) and the clean reads.
- Outputs: binned contigs



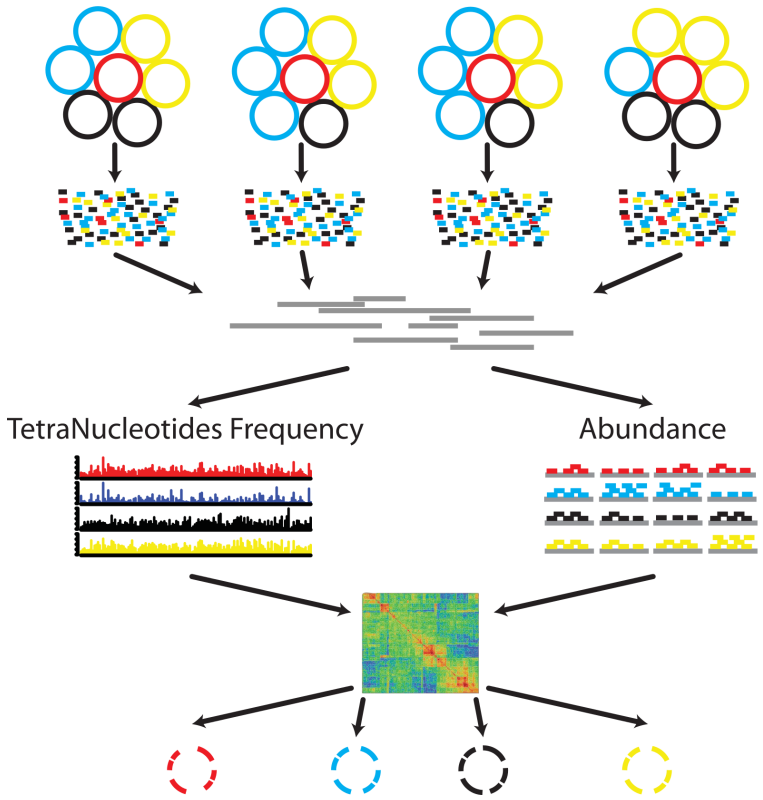
Kang et al., 2015, PeerJ

Shotgun sequencing workflow (from short reads)

4. Binning

Binning assembled contigs into putative single genomes (metagenome assembled genomes – MAGs)

MetaBAT2 (others: CONCOCT, MaxBin2, etc.)



Preprocessing

1. Samples from multiple sites or times
2. Metagenome libraries
3. Initial de-novo assembly using the combined library

MetaBAT

4. Calculate TNF for each contig
5. Calculate Abundance per library for each contig
6. Calculate the pairwise distance matrix using pre-trained probabilistic models
7. Forming genome bins iteratively

Binned contigs can then be used for taxonomic and functional annotation.

MIMAG Standards:

Completeness
> 90% (HQ) or 50%(MQ)
Redundancy/contamination
< 5% (HQ) or 10% (MQ)

Kang et al., 2015, PeerJ

Shotgun sequencing workflow (from short reads)

Processing summary table

	Location	Creek	Core	Raw reads	Trimmed reads	Contigs	Contigs N50 (bp)	Aligned reads	Gene calls	Unique genes (pre-subsample)	Unique genes (post-subsample)
May	Upstream	Reference	b	21438186	20559188	76874	645	1825467	55706	3823	3384
			c	24351101	23422137	62280	607	1655947	39017	3567	3567
		Impacted	b	24415155	23738859	200188	875	9328826	172482	5654	3373
			c	28595973	27823779	253800	858	12102726	214965	5854	3289
	Midstream	Reference	b	37872916	36280805	284820	779	7733672	235440	6142	3844
			c	25621839	24774089	120071	708	2769804	89802	4738	3624
		Impacted	a	25159333	24188071	184185	756	5880198	152017	5427	3573
			b	33367174	32106552	252960	717	8095555	194071	5757	3606
	Downstream	Reference	a	23939789	23374281	134386	694	2952198	96074	4703	3568
			b	26734908	25998243	143503	699	3488016	104070	4762	3476
		Impacted	a	25049011	24024049	175462	686	3302934	125419	5030	3523
			b	22515492	21677934	123524	688	2562007	92503	4509	3449
September	Upstream	Reference	a	28485424	27559917	127098	648	2969027	87632	4627	3540
			b	27246751	26454114	110091	658	2649543	78006	4416	3505
		Impacted	b	24105882	23010409	343472	840	9058675	291269	6733	3867
			c	23999230	23051373	341491	840	9293568	288162	6675	3835
	Midstream	Reference	a	25237220	24557016	115222	672	2456114	75459	4487	3545
			c	32745248	31921279	133811	657	3306746	90019	4727	3547
		Impacted	a	30046745	29020881	226773	729	6629349	173507	5970	3784
			c	28520916	27688113	169577	710	6155569	125476	5351	3669
	Downstream	Reference	a	17666160	16630416	136447	723	2559553	106761	4848	3545
			b	17688053	16972480	125154	720	2591532	96160	4686	3471
		Impacted	a	22073676	21389999	83139	655	1870002	58287	3844	3352
			c	26626579	25869267	117576	676	2847167	86181	4418	3431

Semedo and Song, 2023. STOTEN

Shotgun sequencing workflow (from short reads)

Example results

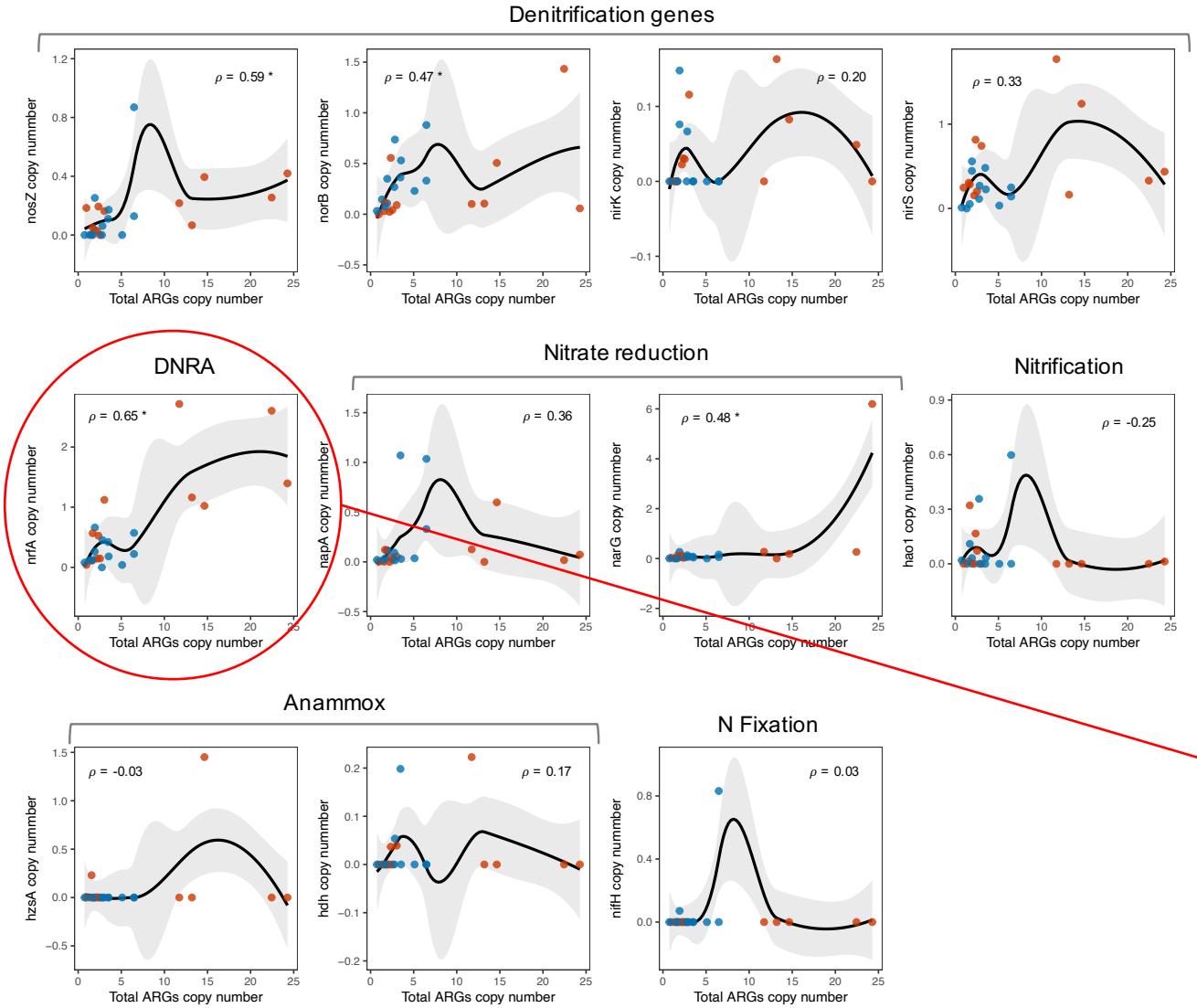
Gene estimated abundances based on reads mapped to contigs

Semedo and Song, 2023. STOTEN

What to do with metagenome sequences

Gene estimated abundances

Gene or metabolic pathway relationships



Driven by co-selection or parallel responses to similar factors?

Semedo and Song, 2023. STOTEN

What to do with metagenome sequences

Pathway co-ocurrence

Taxonomic identification

Functional profiling

Metagenome assembled genomes (MAGs)

Station	Creek	MAG ID	Genus (class/phylum)	# ARG	Nitrate red. narG	Nitrate red. napA	DNRA nrfA	Denitrification nosZ	Denitrification norB	Denitrification nirK	Denitrification nirS	Nitrif. hao1	Anammox hzsA	Anammox hdh	N fix. nifH	% comp.	% cont.
Upstream	Impacted	bin.010	Piscinibacter (Gamma-proteob.)	6		yes	yes				yes					95,6	5,7
		bin.013	UBA3961 (Bacteroidia)	6				yes								87,1	2,0
		bin.057	unclass. (Verrucomicrobiae)	6		yes										99,3	7,4
		bin.017	PHOS-HE28 (Bacteroidia)	5				yes								95,1	3,6
		bin.018	Aestuariivirga (Alpha-proteobacteria)	5	yes	yes	yes			yes						93,2	4,2
		bin.025	unclass. (Acidimicrobiia)	5												95,9	1,8
		bin.037	JAAZBK01 (Acidimicrobiia)	5		yes										94,0	6,3
		bin.042	JJ008 (Bacteroidia)	4				yes								77,6	1,1
		bin.055	JACADZ01 (Alpha-proteobacteria)	4	yes		yes									94,4	2,1
		bin.061	unclass. (Thermoanaerobaculia)	4	yes	yes										66,9	8,4
		bin.016	Methylomirabilis (Methylomirabilia)	3	yes	yes					yes			yes		90,2	3,5
		bin.038	UBA7227 (Anaerolineae)	3	yes	yes	yes				yes					70,4	4,2
		bin.046	JAAUPO01 (Cyanobacteriia)	3												84,1	1,0
		bin.047	unclass. (Kapabacteria)	3		yes		yes								86,9	0,1
		bin.050	Terricaulis (Alpha-proteobacteria)	3	yes											81,6	4,9
		bin.051	unclass. (Ignavibacteria)	3		yes	yes									86,5	1,9
		bin.056	unclass. (Myxococcota)	3	yes	yes					yes					60,7	4,4
		bin.006	unclass. (Gamma-proteob.)	2		yes		yes			yes					95,7	1,9
		bin.019	UBA8403 (Bacteoidia)	2												68,9	1,0
	Ref.	bin.009	Sulfuricella (Gamma-proteob.)	2		yes	yes	yes	yes		yes					85,7	6,9
		bin.004	SM1-50 (Thermoplasmatota)	0												93,2	3,2
		bin.013	PALSA-986 (Thermoproteota)	0												55,5	1,9

Semedo and Song, 2023. STOTEN

What to do with metagenome sequences

Zone	Habitat	Sample ID	Class	Genus	CMPL %	CONT %	Nitrogen Cycle Genes											
							narG	napA	narB	nasA	nrfA	nirK	nirS	norB	nosZ	nifH	pmoA	hao	hzs
Bathypelagic (1000m - 4000m)	Cold Seep	SAMEA5663119	Methanosarcinia	UBA204	72	0.66													
			Spirochaetia	JAHOYX01	88	4.85													
			UBA6098	Unclassified	84	2.20													
	Continental Slope and Abyssal Plain	mgm4510162.3	Alphaproteobacteria	*Amylibacter*	52	0.86													
		mgm4510164.3	Alphaproteobacteria	*Amylibacter*	78	1.68													
				Sneathiella	55	5.33													
			Gammaproteobacteria	50-400-T64	61	8.10													
		mgm4510168.3	Gammaproteobacteria	*Agaribacterium*	93	3.79													
		mgm4510171.3	Gammaproteobacteria	*Colwellia*	56	8.62													
				UBA3067	87	9.99													
		mgm4510172.3	UBA6911	Unclassified	69	3.04													
	Hydrothermal Vent	SAMN03002195	Alphaproteobacteria	*Aurantimonas*	96	1.61													
			Bacteroidia	SM23-62	66	2.15													
			Gammaproteobacteria	GCA-002733105	91	0.42													
				Halomonas	98	2.01													
				Halopseudomonas	99	1.06													
				Marinobacter	98	0.45													
				Methylophaga	79	8.62													
				Pseudohongiella	97	0.06													
		SAMN03002196	Alphaproteobacteria	*Aurantimonas*	100	7.59													
			Aminicenantia	SOIV01	78	4.84													
			Aquicultoria	Unclassified	63	1.81													
			Gammaproteobacteria	*Halomonas*	100	2.44													
				Halopseudomonas	94	0.73													
				Pseudoalteromonas	98	6.55													
				Pseudohongiella	71	0.06													
			Phycisphaerae	HyVt-337	99	5.88													
	Subduction Zone and Trenches	SAMN19461850	Anaerolineae	Unclassified	92	4.55													
		SAMN19461852	Aerophobia	SOJT01	75	1.12													
		SAMN19461860	Actinomycetia	*Mycobacterium*	55	0.63													
			Anaerolineae	E44-bin32	85	5.61													
			Desulfobacteria	Unclassified	82	0.97													
Abyssopelagic (4000m - 6000m)	Hydrothermal Vent	SAMN05571518	Alphaproteobacteria	Unclassified	67	1.03													
			Bipolaricaulia	DRJF01	83	0.00													
			Gammaproteobacteria	HyVt-429	86	0.76													
			SZUA-79	QIJE01	87	0.86													
			Thermodesulfovibrionia	BMS3Bbin07	94	2.73													
			UBA6919	CAJXGR01	66	2.27													
			UBA7883	DRJW01	89	4.44													
	Unclassified	SAMEA5744179	Gammaproteobacteria	GCA-2729495	84	6.65													
				JAACFB01	51	8.55													
Hadopelagic (>6000m)	Subduction Zone and Trenches	SAMN16252685	Gammaproteobacteria	SMWN01	55	1.10													
			Nitrospinia	SZUA-226	86	2.42													
			UBA1144	CR02bin9	53	0.00													
		SAMN16252686	Nitrospinia	SZUA-226	64	0.91													
			Terriglobia	JACZYA01	58	1.78													

Metagenome assembled genomes (MAGs) from public metagenomes

Functional profiling

Associated taxonomic classification

Presence of genes of interest in poorly known habitats

Semedo, Valquaresma, et al (*in prep*)

Questions?