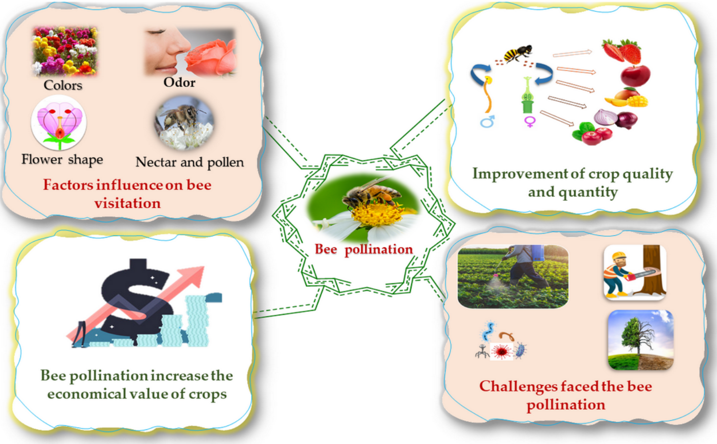
Optimize sequencing depth for shotgun metagenomics of pollination system by rarefaction, using a modular profiling pipeline

Cong Liu

2021.9

Bee health is a crucial issue


Colors
Odor
Flower shape
Nectar and pollen
Factors influence on bee visitation
Improvement of crop quality and quantity
♂
♀
Bee pollination
Bee pollination increase the economical value of crops
Challenges faced the bee pollination

Khalifa *et al.*, 2021

Diverse gut microbiome impacts bee health

Taxonomic diversity

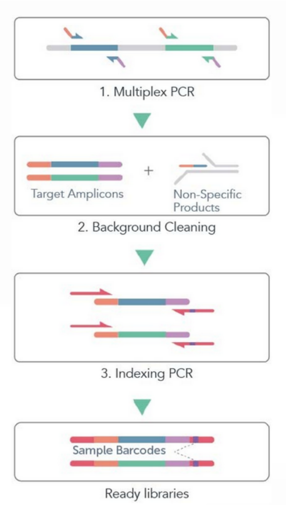
- Bacteria:
  Core bacteria: *Bifidobacterium sp.*, *Frischella sp.*, *Gilliamella sp.*, *Snodgrassella sp.*, *Lactobacillus sp.*
  None-core bacteria: *e.g.* *Bartonella sp.*, *Apibacter sp.*, *Enterobacter sp.* *Klebsiella sp.*
- Fungi:
  yeasts: *e.g.* *Saccharomyces sp.*, *Zygosaccharomyces sp.*, *Wickerhamomyces sp.*
  pathogens: *e.g.* *Nosema sp.*
- Viruses:
  phages: *e.g.* *Badaztecvirus sp.*, *Bigbernvirus sp.*, *Blindbaselvirus sp.*
  host-infecting: *e.g.* deformed wing virus (*Iflavirus sp.*), Lake Sinai virus (*Sinaivirus sp.*)
- Other eDNA signature:
  e.g. plants, arthropods

Diverse gut microbiome impacts bee health

Functional diversity

- Food digestion:
  pectin breakdown,
  sucrose hydrolysis,
  mannose metabolism, *etc*
- Parasite defence:
  Crithidia,
  Paenibacillus larvae,
  Nosema sp., *etc*
- Chemical detoxification:
  cadmium,
  copper,
  selenate, *etc*

Amplicon sequencing for exploring microbiome



Source: https://www.paragongenomics.com/targeted-sequencing/amplicon-sequencing/

Limitations of amplicon sequencing for bee microbiome studies

Loss of taxonomic diversity:

- Different taxonomic clades require different barcode regions,
- Amplicon sequencing only captures taxon diversity within a certain clade.
- Bee microbiome is composed of diverse clades.

Unreliable function inference:

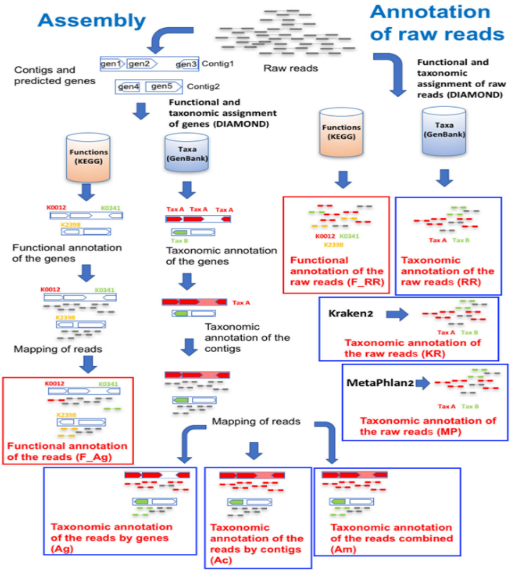
- No information on functional gene clusters is provided.
- Function potentiality is inferred from pre-sequenced genomes.
- Bee bacterial symbionts are diversified at strain level, indicating gene content variation.

Shotgun metagenomics provides a solution to overcome limitations of amplicon sequencing

Advantages of shotgun metagenomics:

- Capturing DNA fragments unselectively
- Illustrating diversity of multiple taxonomic clades
- Providing information on functional gene content

Shotgun metagenomics: challenge of data analysis


Assembly
Annotation of raw reads
Raw reads
Contigs and predicted genes
gen1
gen2
gen3
Contig1
gen4
gen5
Contig2
Functional and taxonomic assignment of genes (DIAMOND)
Functional and taxonomic assignment of raw reads (DIAMOND)
Functions (KEGG)
Taxa (GenBank)
Functions (KEGG)
Taxa (GenBank)
K0012
K0341
K2398
Functional annotation of the genes
Tax A
Tax A
Tax A
Tax B
Taxonomic annotation of the genes
K0012
K0341
K2398
Mapping of reads
Tax A
Taxonomic annotation of the contigs
K0012
K0341
K2398
Functional annotation of the reads (F_Ag)
K0012
K0341
K2398
Functional annotation of the raw reads (F_RR)
Tax A
Tax B
Taxonomic annotation of the raw reads (RR)
Kraken2
Tax A
Tax B
Taxonomic annotation of the raw reads (KR)
MetaPhlan2
Tax A
Tax B
Taxonomic annotation of the raw reads (MP)
Mapping of reads
Taxonomic annotation of the reads by genes (Ag)
Taxonomic annotation of the reads by contigs (Ac)
Taxonomic annotation of the reads combined (Am)

Tamames et al. (2019)

Shotgun metagenomics: challenge of sequencing depth determination

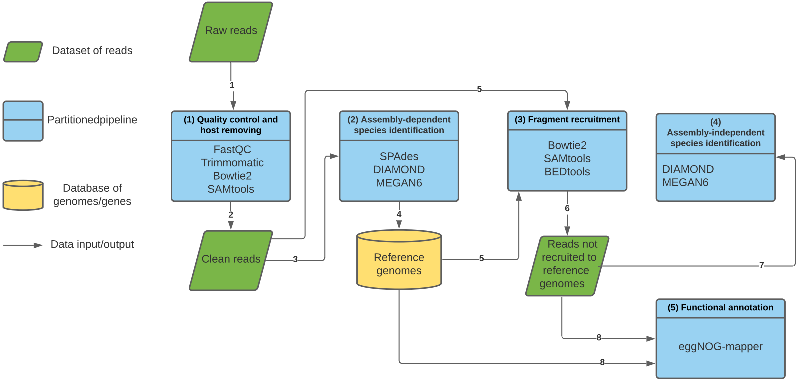
Cost for $>$ 12 Gbp/sample shotgun metagenomics or $>$ 40k 16S reads/sample (100 samples)

Company	Microeco Tech Group Co.,	APExBIO	GeneCloudBio
Technology	NovaSeq	Not mentioned	Illumina
Shotgun metagenomics	Q20 $>$ 85%	Not mentioned	Q20 $>$ 85%
	540 RMB/sample (6 Gbp)+59 RMB/Gbp extra depth	1888 RMB/sample (10 Gbp)	2500 RMB/sample (12 Gbp)
	Total: 101200 RMB (∼11244 pounds)	Total: $>$ 188800 RMB (∼20978 pounds)	Total: $>$ 250000 RMB (∼27778 pounds)
16S amplicon	155 RMB/sample (50 k reads)	188 RMB/sample	340 RMB/sample (50 k reads)
	Total: 15500 RMB (∼1722 pounds)	Total: 18800 RMB (∼2089 pounds)	Total: 34000 RMB (∼3778 pounds)

Aims

- Combining assembly-dependent and -independent methods for metagenomic data analysis
- Optimizing sequencing depth to balance sequencing cost and reliable diversity analysis of microbiomes from different host species

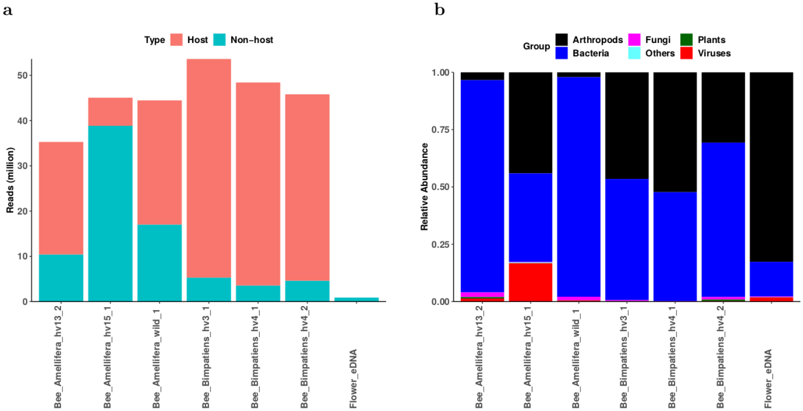
Integrated pipeline for taxonomic/functional profiling of shotgun metagenomic data



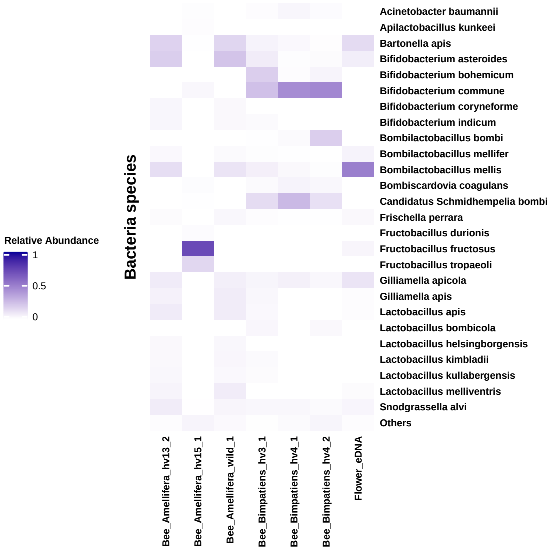
Samples and sequence data

- Three honey bees (*Apis mellifera*)
  Three bumble bees (*Bombus impatiens*)
  One flower washes of *Erigeron annuus*
- 2 × 150bp read pairs
- Deep sequencing
- Analyzed using the integrated pipeline

Diverse communities composed of multiple taxonomic clades were identified



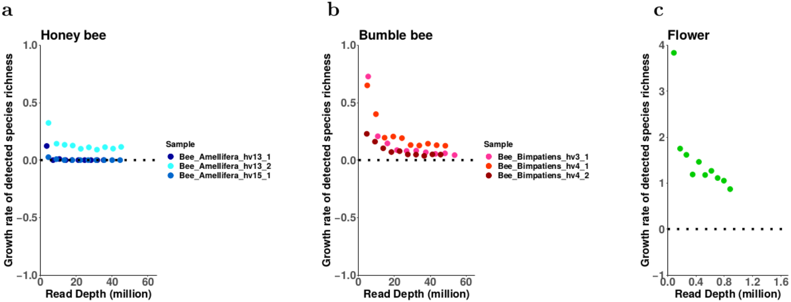
Core bacterial symbionts indicates good quality and representativity of samples



Species common in pollination system indicate sample quality and representativity

- Arthropods:
  Dominated by *Apis* and *Bombus*
- Plants:
  Crops including rape, soybean, sunflower, radish
- Fungi:
  Dominated by *Nosema ceranae* and yeasts
- Viruses:
  Phages and arthropod-infecting viruses

Integrated pipeline provided improvement in species identification



Sequencing depth was simulated by randomly subsampling original datasets

Advantages of integrated pipeline

- Improvement in species identification
  Combination of assembly-free method helps solve species not represented by assembly.
- Flexibility
  Modularity provides capability for incorporating of alternative tools.
- Transparency
  Output files generated by each step are recorded and can be inspected easily for troubleshooting.

Integrated pipeline could be evaluated more comprehensively

- Microbiome of other host species
- Mock metagenomic dataset
- Comparison with other strategies, *e.g.* MG-RAST, SqueezeMeta, Kraken.

Aims

- Combining assembly-dependent and -independent methods for metagenomic data analysis
  Integrated pipeline
- Optimizing sequencing depth to balance sequencing cost and reliable diversity analysis of microbiomes from different host species

Measure diversity by Hill numbers

Hill numbers:

$$D^{(q)} = (\sum_{i}^{S}(p_i)^q)^{\frac{1}{1-q}} \tag{1}$$

p_i: the relative abundance of ith species or functional gene cluster (KO)

q: parameter

S: number of categories

Advantages:

- Replication principle
- Modulate sensitivity to relative abundances via order q
- Related to widely used diversity indexes:

$$D^{(0)} = S \tag{2}$$

$$D^{(1)} = e^{-\sum_i p_i log p_i} \tag{3}$$

$$D^{(2)} = \frac{1}{\sum_i (p_i)^2} \tag{4}$$

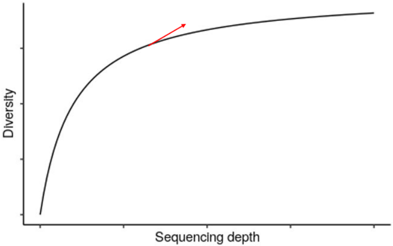
Computing expected diversity of given sequencing depth

- Assume the original datasets is almost complete
  i.e. almost all species/KOs are represented by the original datasets
- Simulate shallow sequencing by random subsampling (10%-90% at interval of 10%)
- Fit to asymptotic accumulation models
- Multimodel inference based on Akaike weight

$$D^{(q)}(x) = \sum_{i} w_i D_i^{(q)}(x) \tag{5}$$

x: sequencing depth
$D_i^{(q)}(x)$: ith fitted model describing relationship between sequencing depth and Hill numbers
w_i: Akaike weight of ith model calculated from small sample unbiased Akaike information criterion (AICc)

Optimizing sequencing depth according to slope of rarefaction curve

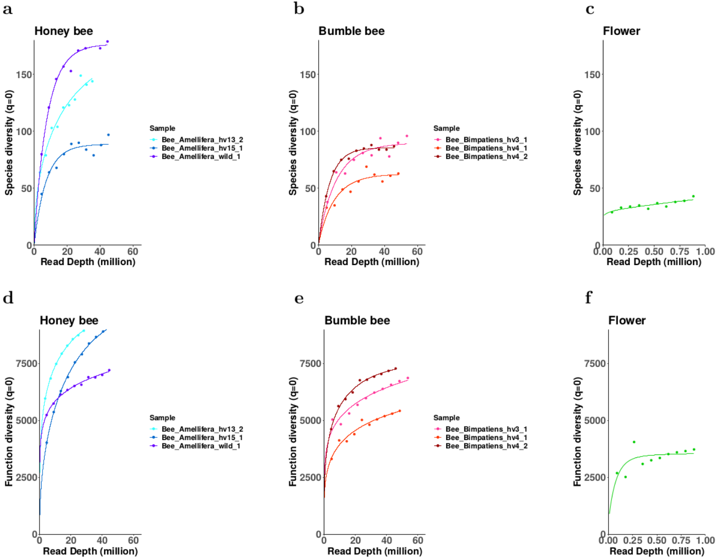


$$\frac{dD^q}{dx} = \sum_i w_i \frac{dD_i^{(q)}}{dx} \tag{6}$$

$$Asymptote = \lim_{x \to +\infty} D^{(q)}(x) \tag{7}$$

$$Completeness = \frac{D^{(q)}(x)}{Asymptote} \tag{8}$$

Verify completeness of original datasets by rarefaction curves for species/KO richness



Optimal sequencing depth for diversity estimation

Species diversity estimation:

- *Bee_Amellifera_hv13_2* and *Flower_eDNA* were dropped for incompleteness (final slope > 1 and completeness $< 80\%$)
- $q = 0$ (species richness):
  Slope < 0.1 provided completeness $> 95\%$
  Honey bees: 40.33 million (12.10 Gbp)
  Bumble bees: 42.49 million (12.75 Gbp)
- $q = 1$ or 2 (reduced emphasis on rare species)
  Slope < 0.01 provided completeness $> 95\%$
  Honey bees: 18.57 million or 5.57 Gbp ($q = 1$) and 17.45 million or 5.24 Gbp ($q = 2$)
  Bumble bees: 40.33 million or 12.10 Gbp ($q = 1$) and 24.77 million or 7.43 Gbp ($q = 2$)

Function diversity estimation:

- All datasets were incomplete (final slope > 15)

Sequencing depth can be optimized for species diversity estimation

Species diversity estimation:

- 12.0 Gbp (honey bees) and 12.9 Gbp (bumble bees) would be sufficient for capturing species richness
- Shallower sequencing can be adopted when little emphasis is given on rare species

Function diversity estimation:

- Deep sequencing is valuable

Limitations:

- Small sample size (3 honey bees, 3 bumble bees and 1 flower eDNA)
- Lack of repeat in sequencing depth subsampling

Summary

- The integrated pipeline provides benefits in terms of results, flexibility and transparency
- For species diversity detection, 12 Gbp for honey bees and 12.9 Gbp for bumble bees would be sufficient
- Shallower sequencing can be adopted with reduced emphasis on rare species
- For function diversity, deep sequencing would be recommended
- Similar pilot studies for large scale metagenomic project of other host species help budget management