Studiare il microbiota: tips and tricks

Eleonora Mastrorilli

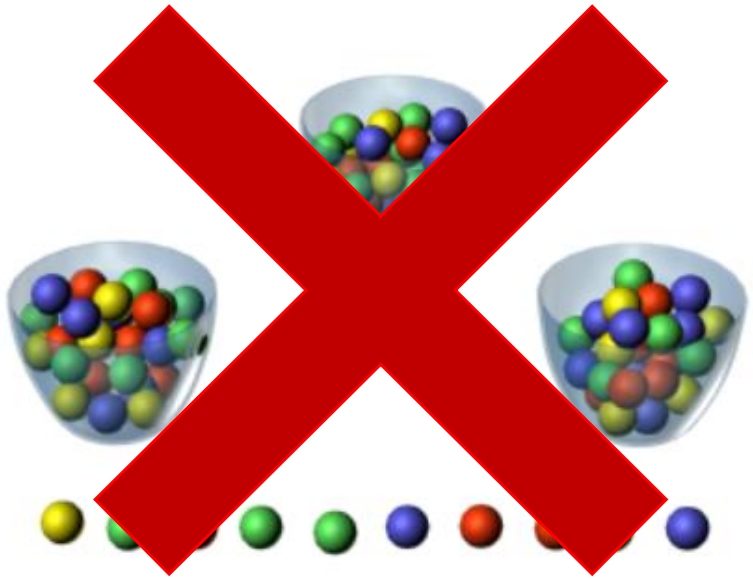
ISZVe

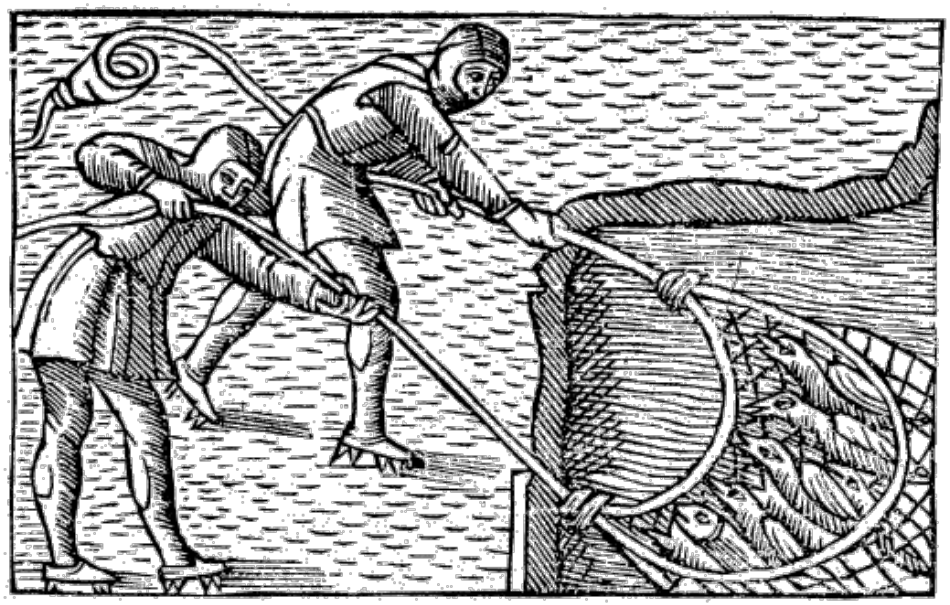
Overview

- Introduzione
- Il tipo di dato
- I dati come CoDa
- La normalizzazione
- La sparsità
- Le fonti di bias
- Cosa si può fare

Introduzione

Una metafora per i dati di sequenziamento 16S

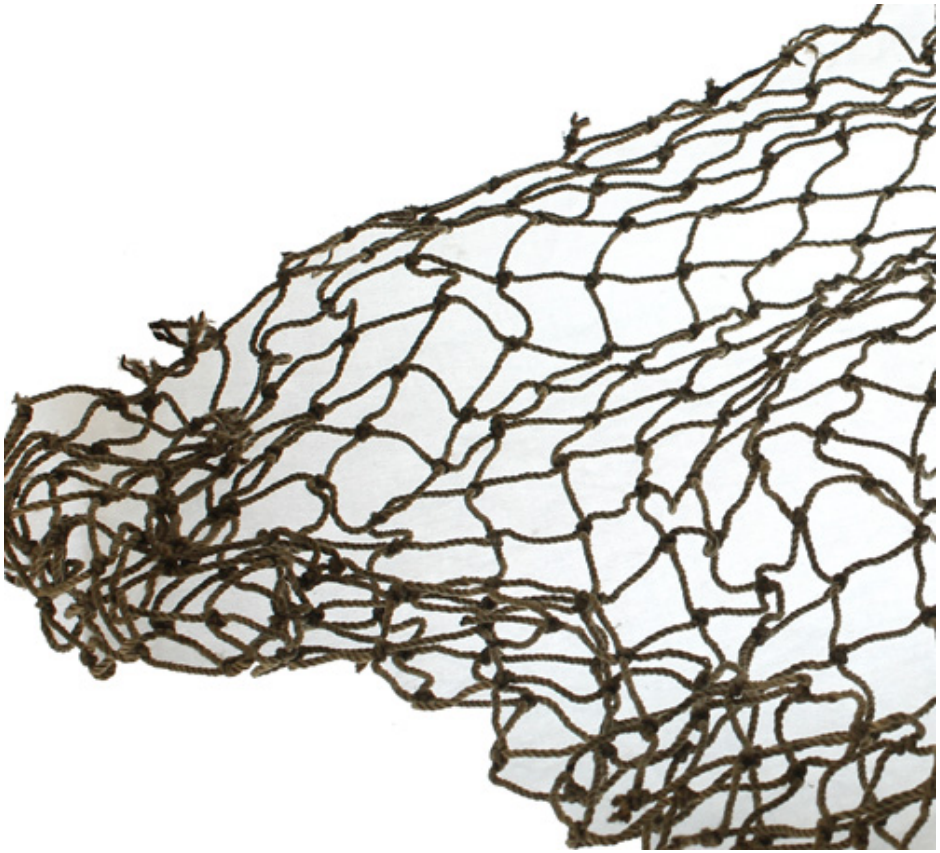




Introduzione

Una metafora ci spiega

- il campione rappresentato
- la «selettività» del processo di campionamento
- la struttura del dato



Il tipo di dato

I dataset sono solitamente caratterizzati da

- *alta dimensionalità* (alto numero di taxa)
- *sottodeterminazione*: n° di taxa >> n° di campioni
- *sparsità*: molti taxa sono presenti solo in un piccolo numero di campioni
- *compositional data*: sono dati vincolati

Il tipo di dato

Inoltre

- la collezione dei campioni
- l'amplificazione via PCR
- il sequenziamento

complicano l'inferenza delle abbondanze reali dai count ottenuti.

I dati come CoDa

- Il numero di count assegnato ai vari organismi in un campione è vincolato da un valore massimo che varia, all'interno della stessa corsa, da campione a campione
  - → *Compositional Data (CoDa):* vettori di elementi non negativi vincolati ad avere somma costante

La sequencing depth

- La somma costante è la *sequencing depth* dedicata al singolo campione
- Ogni campione può avere una seq. depth diversa

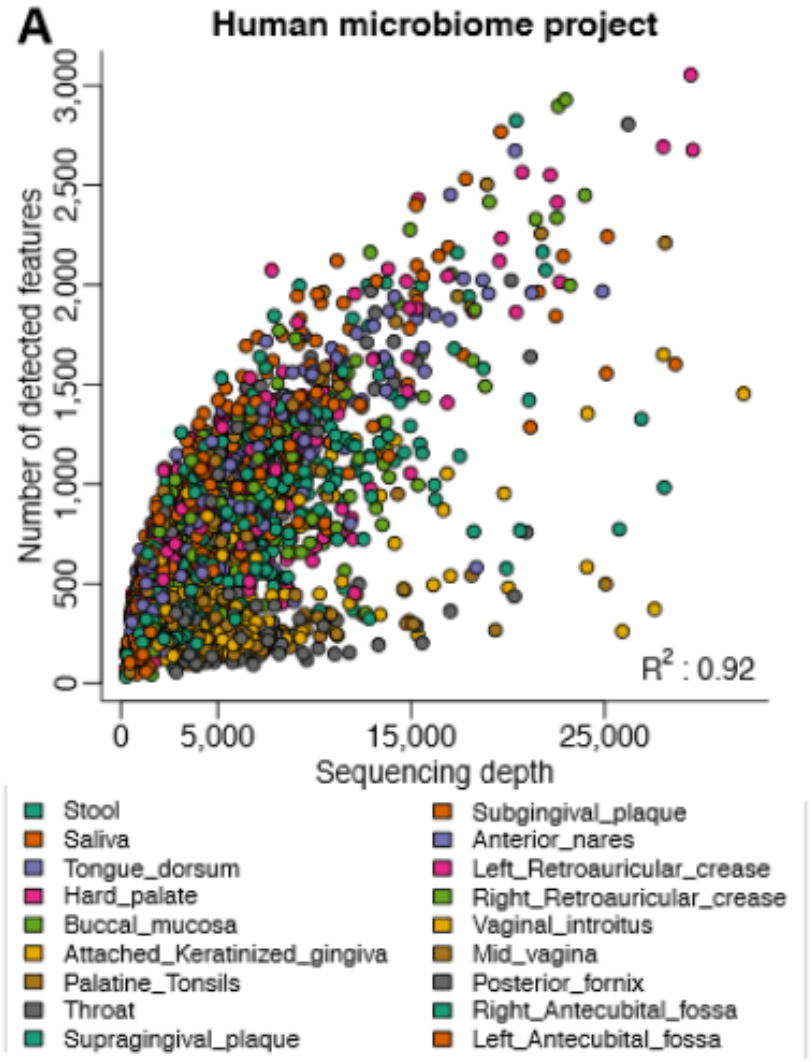


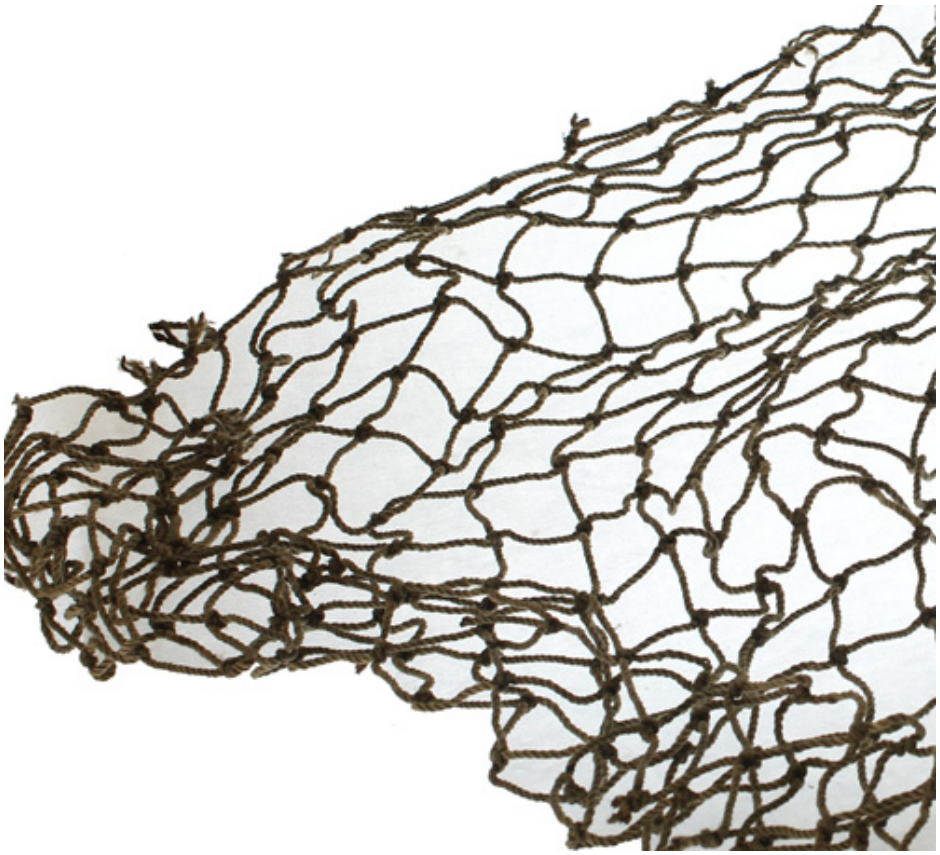
Image from [9]

La sequencing depth

La sequencing depth può variare anche di più ordini di grandezza nella stessa run…

Alcuni campioni hanno poche sequenze

Quando li eliminiamo?



I dati come CoDa

- Pearson (1897) afferma che:

  "spurious correlations" would result, should values constructed as proportions be compared haphazardly

- CoDa sono soggetti al "problema della chiusura":
  - I vari componenti competono per comporre il totale vincolato
  - Il cambiamento nell'abbondanza di uno dei componenti CAUSA cambiamenti apparenti nell'abbondanza degli altri
  - L'indipendenza dei campioni non è più rispettata
  - Si creano errori di COVARIANZE SPURIE

I dati come CoDa

otu 1	10
otu 2	10
otu 3	15
otu 4	40
otu 5	25

N = 100

otu 1	10
otu 2	10
otu 3	10
otu 4	40
otu 5	30

N = 100

otu 1	10
otu 2	10
otu 3	10
otu 4	60
otu 5	10

N = 100

Lovell et al. (2015): "in the absence of any other information or assumptions, correlation of relative abundances is just wrong."

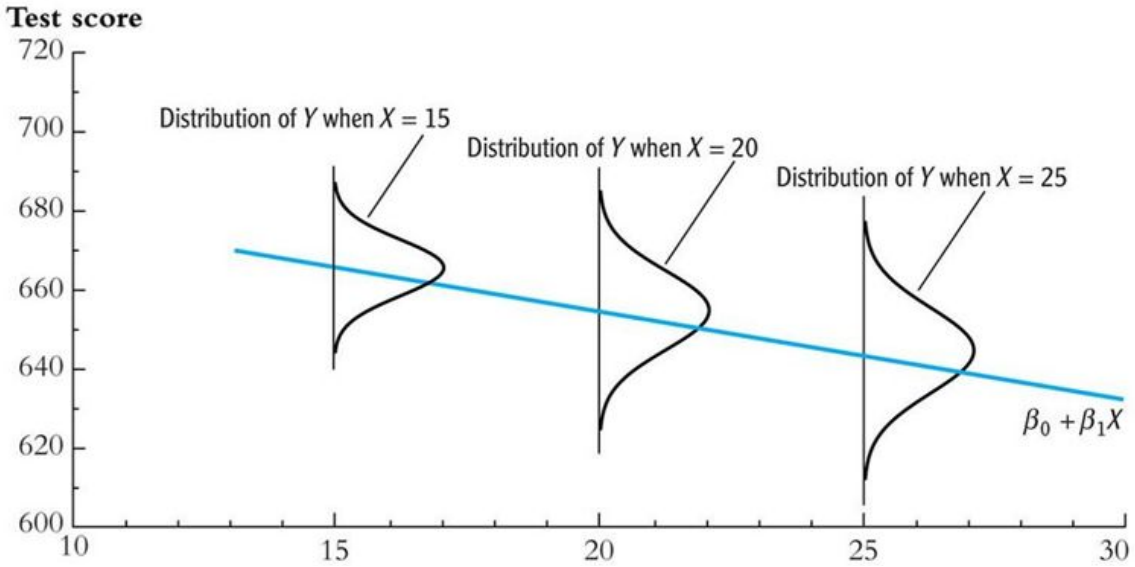
La normalizzazione

- I Compositional Data spesso vengono analizzati dopo una trasformazione
  - divisione per un valore
  - logaritmo

- La scelta del denominatore di solito corrisponde ad uno specifico approccio di *normalizzazione*
  - scegliere una componente come reference
  - centered log-ratio
  - DESeq usa RLE
  - edgeR usa RLE, TMM, quantile
  - metagenomeSeq usa CSS

La normalizzazione

- Perché non basta usare le ***proporzioni***?
  - Proporzioni, percentuali o abbondanze relative NON sono approcci di normalizzazione!
    - Usano un denominatore costante
    - Gli approcci di normalizzazione determinano empiricamente un denominatore unico per ciascun esperimento
  - Perché non tengono conto dell'*eteroschedasticità*



La normalizzazione

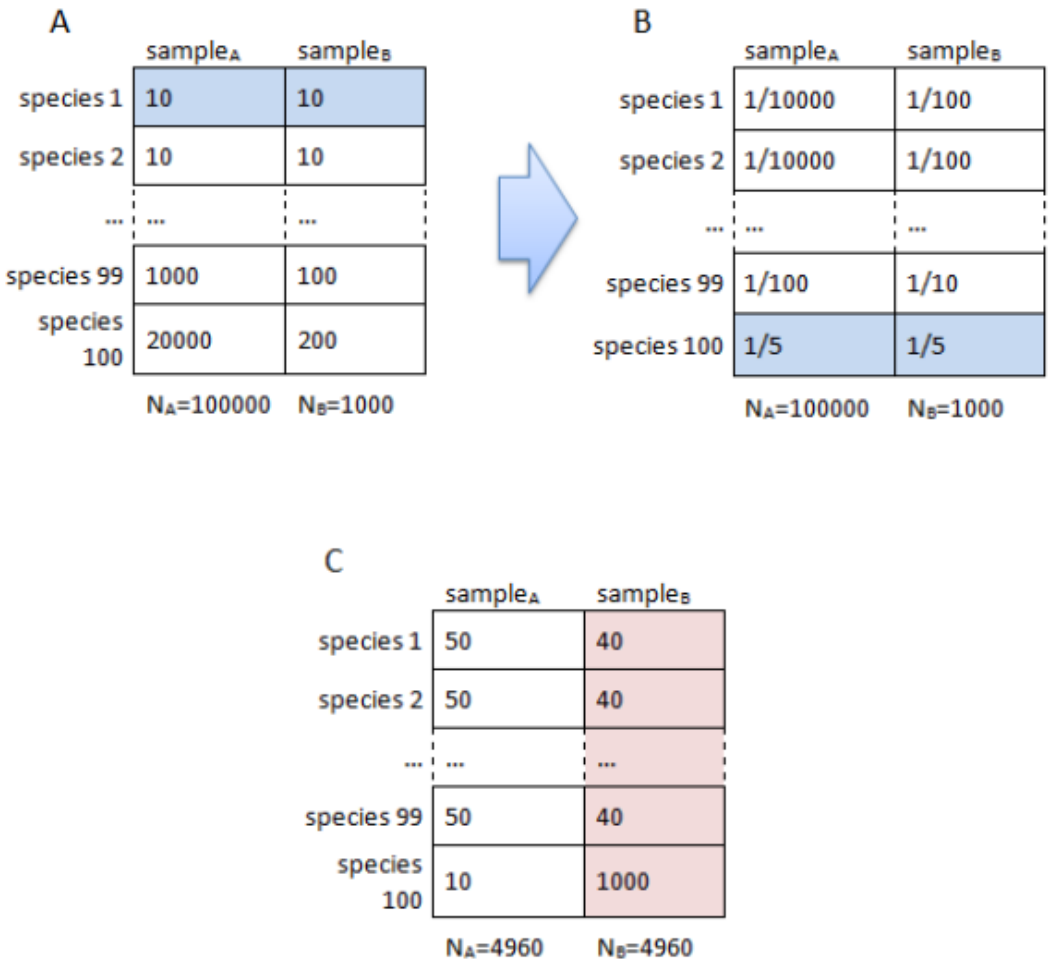
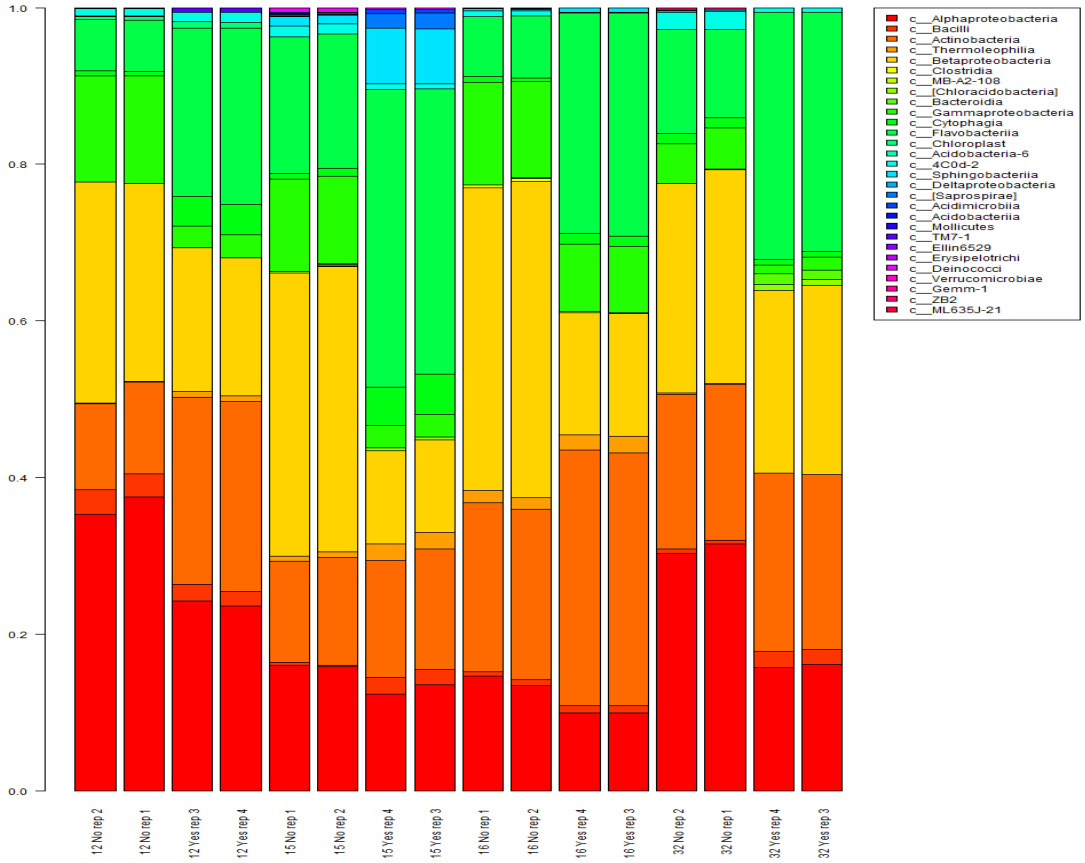


FIGURE 29 DIFFERENT EFFECTS OF SEQUENCING DEPTH ON COUNT DISTRIBUTION: A) SAME ABUNDANCE BUT DIFFERENT RELIABILITY, B) SAME PROPORTION, BUT DIFFERENT VARIABILITY, C) ONE ABUNDANT SPECIES "CONSUMES" MOST OF THE SEQUENCING DEPTH

La normalizzazione

Le proporzioni sono comunque rilevanti…



La normalizzazione

- Perché non basta usare la ***rarefaction***?
  - La rarefaction non è un approccio di normalizzazione!

    Paul J. McMurdie e Susan Holmes nel loro paper[5] la definiscono *inammissibile…*
    - scelgo una sequencing depth minima *N*
    - elimino i campioni con un numero totale di reads inferiore a *N*
    - sottocampiono le librerie rimanenti senza ripetizione in modo che tutte abbiano una sequencing depth totale *N*

Notare che

- la scelta di N è del tutto soggettiva…
- la rarefaction è la scelta di default di QIIME…

La normalizzazione

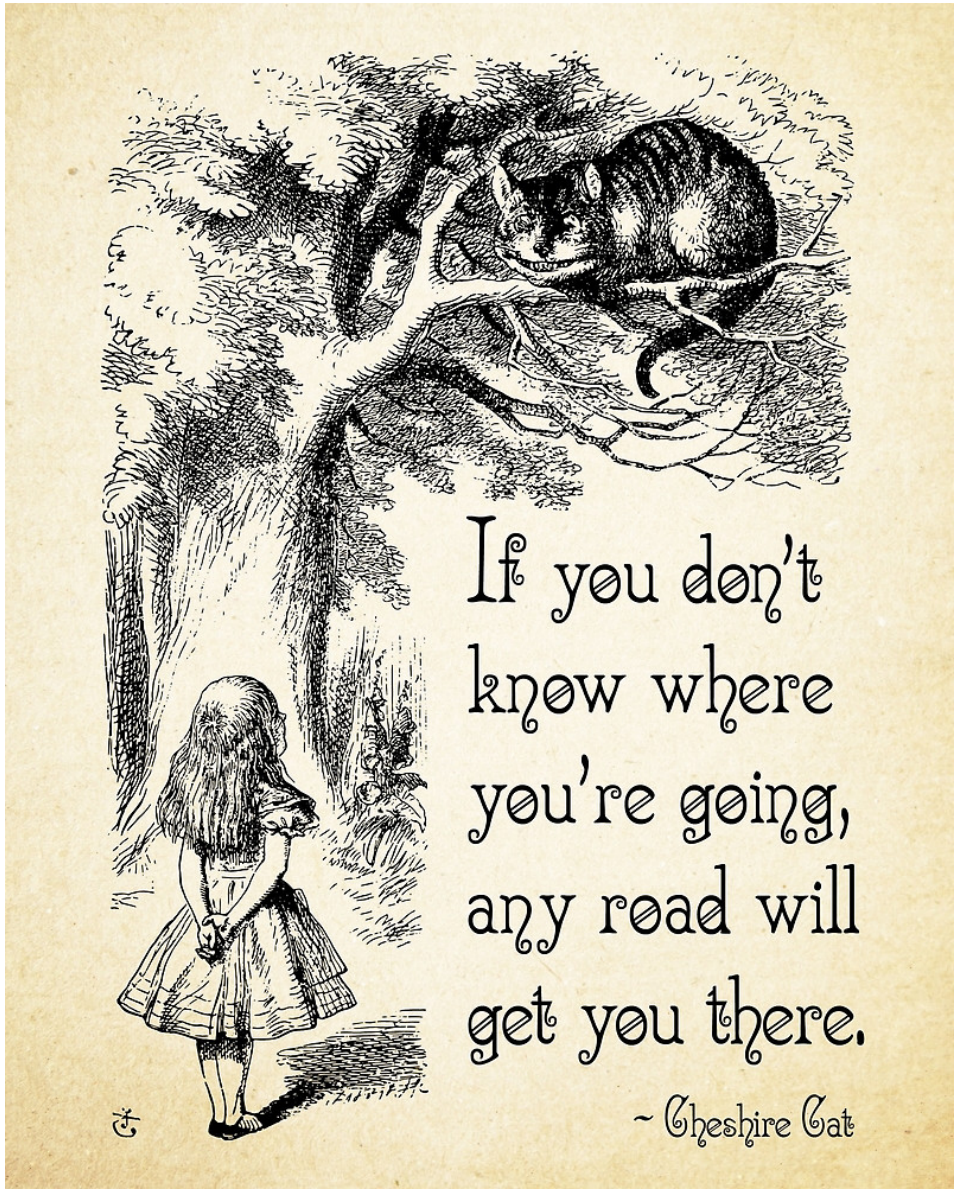
Perché la rarefaction è statisticamente *inammissibile*?

- richiede di omettere dati validi disponibili
- perdo potenza statistica
- aggiungo l'incertezza dovuta alla fase di sotto-campionamento random

La normalizzazione

La risposta è nella domanda stessa!

…tipi di dato diverso forniscono informazioni diverse.



La normalizzazione

Original Abundance

	A	B
OTU1	62	500
OTU2	38	500
Total	**100**	**1000**

Rarefied Abundance

	A	B
OTU1	62	50
OTU2	38	50
	100	**100**

Standard Tests for Difference

P-value	chi-2	Prop	Fisher
Original	0.0290	0.0290	0.0272
Rarefied	0.1171	0.1171	0.1169

La sparsità

C’è un problema che neanche la normalizzazione ci aiuta a mitigare…

$$\begin{pmatrix} 1.0 & 0 & 5.0 & 0 & 0 & 0 & 0 & 0 \\ 0 & 3.0 & 0 & 0 & 0 & 0 & 11.0 & 0 \\ 0 & 0 & 0 & 0 & 9.0 & 0 & 0 & 0 \\ 0 & 0 & 6.0 & 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 7.0 & 0 & 0 & 0 & 0 \\ 2.0 & 0 & 0 & 0 & 0 & 10.0 & 0 & 0 \\ 0 & 0 & 0 & 8.0 & 0 & 0 & 0 & 0 \\ 0 & 4.0 & 0 & 0 & 0 & 0 & 0 & 12.0 \end{pmatrix}$$

La sparsità

- ZERI STRUTTURALI: alcuni zeri sono dovuti alla presenza di taxa poco abbondanti, che troviamo solo in pochi campioni

- ZERI DI «ARROTONDAMENTO»: altri derivano da artefatti di sequenziamento o dalla differenza nella sequencing depth tra campioni diversi

La sparsità

- I modelli parametrici di inferenza statistica fanno fatica a stimare i parametri da dati con molti zeri
- Gli approcci di normalizzazione che sfruttano il passaggio al logaritmo hanno problemi con gli zeri

La sparsità

- Una soluzione per gli zeri di arrotondamento è sostituirli con un valore (*pseudo-count*) → *zero imputation*

- Gli zeri strutturali possono essere modellati usando distribuzioni «zero-inflated»[6]

La sparsità

Una soluzione per gli zeri di arrotondamento è sostituirli con un valore

- non-nullo
- piccolo

Tale valore è detto *pseudo-count* e questa strategia è detta di *zero imputation*

La sparsità

Tuttavia

- la scelta del valore da assegnare
- la difficoltà di rendere le analisi robuste al variare del valore
- la difficoltà di rendere questo approccio non dipendente dal grado di sparsità dei dati

→ non esiste una strategia generalizzata

La sparsità

- Gli zeri strutturali possono essere modellati usando distribuzioni «zero-inflated»[6]
  - i count vengono modellati con una distribuzione
    - gaussiana
    - negativa binomiale
    - ecc.
  - una seconda distribuzione descrive gli zeri presenti nel dataset

La sparsità

Se il mio unico obiettivo è valutare il numero di specie presenti nel campione (*alpha diversity*), ci sono misure in grado di *stimare* il numero di taxa non visti

- Chao index
- first and second order Jackknife estimator
- ACE

usando informazioni su altri taxa «rari»

Il «rare microbiome»

I taxa «rari» sono quelli di cui osservo poche sequenze (o nessuna)

- è un contaminante?
- è una chimera?

ma possono avere funzioni importanti nella loro nicchia ecologica

→ eliminare i 'singletons' è spesso lo standard…

Il «rare microbiome»

Eppure taxa rari possono avere funzioni importanti nella nicchia ecologica investigata

→ ignorare i taxa poco abbondanti (sebbene statisticamente valido) può perdere informazioni importanti!

→ eliminare i ‘singletons’ è spesso lo standard…

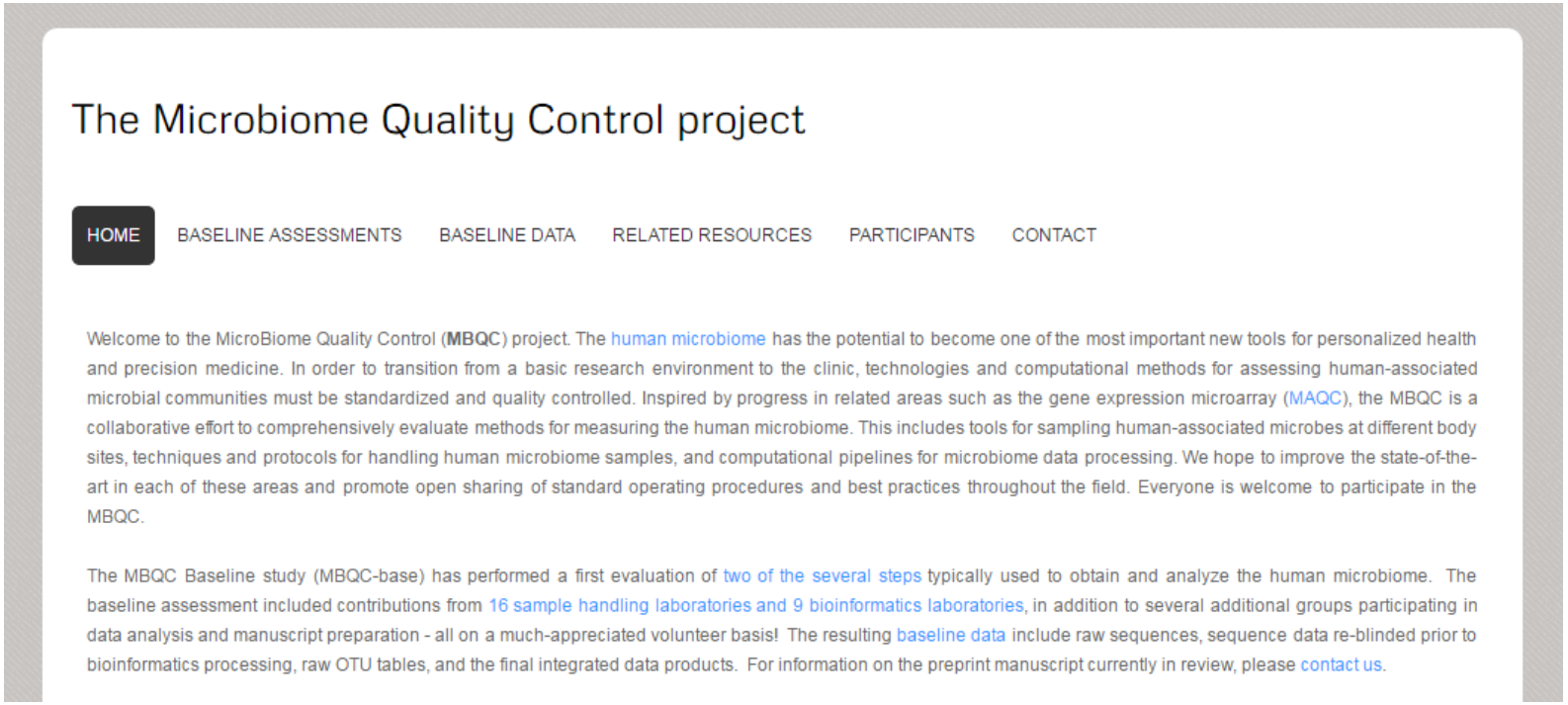
Le fonti di bias

"could anything else explain the results?"

- il 16S copy number
- l'estrazione del DNA
- la PCR
- il sequencing
- …

Le fonti di bias

- Vari studi hanno analizzato quali sono e come si possono mitigare le fonti di bias[3,4]
- Microbiome Quality Control Consortium (http://www.mbqc.org)



Il 16S copy number

Il numero di copie del gene 16S varia molto

- da uno in molte specie
- a 15 in *Photobacterium profundum*

L'abbondanza relativa di un taxa in un campione può attribuirsi

- alla variazione dell'abbondanza dei vari organismi
- alla variazione del numero di copie di 16S tra gli organismi

Il 16S copy number

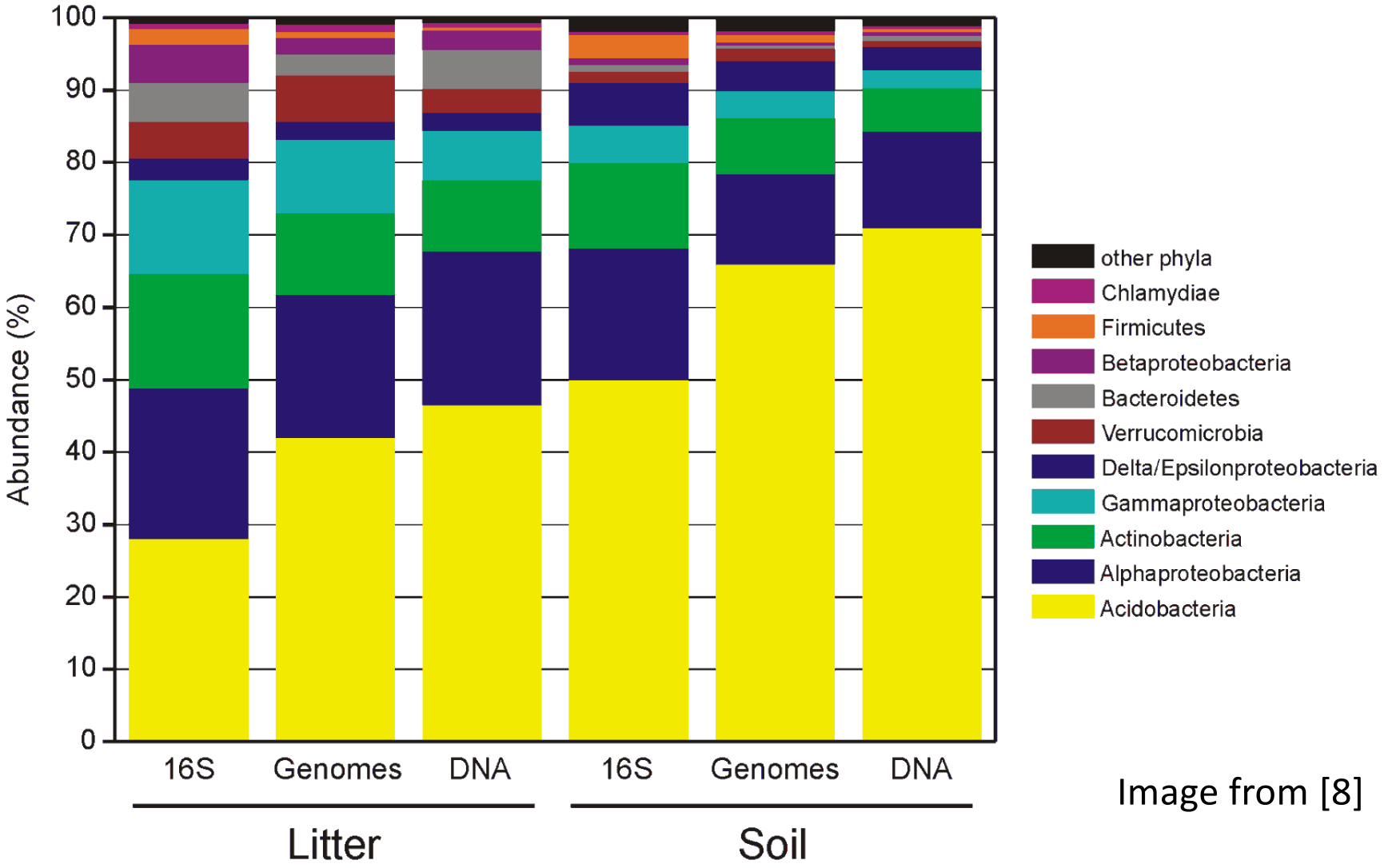


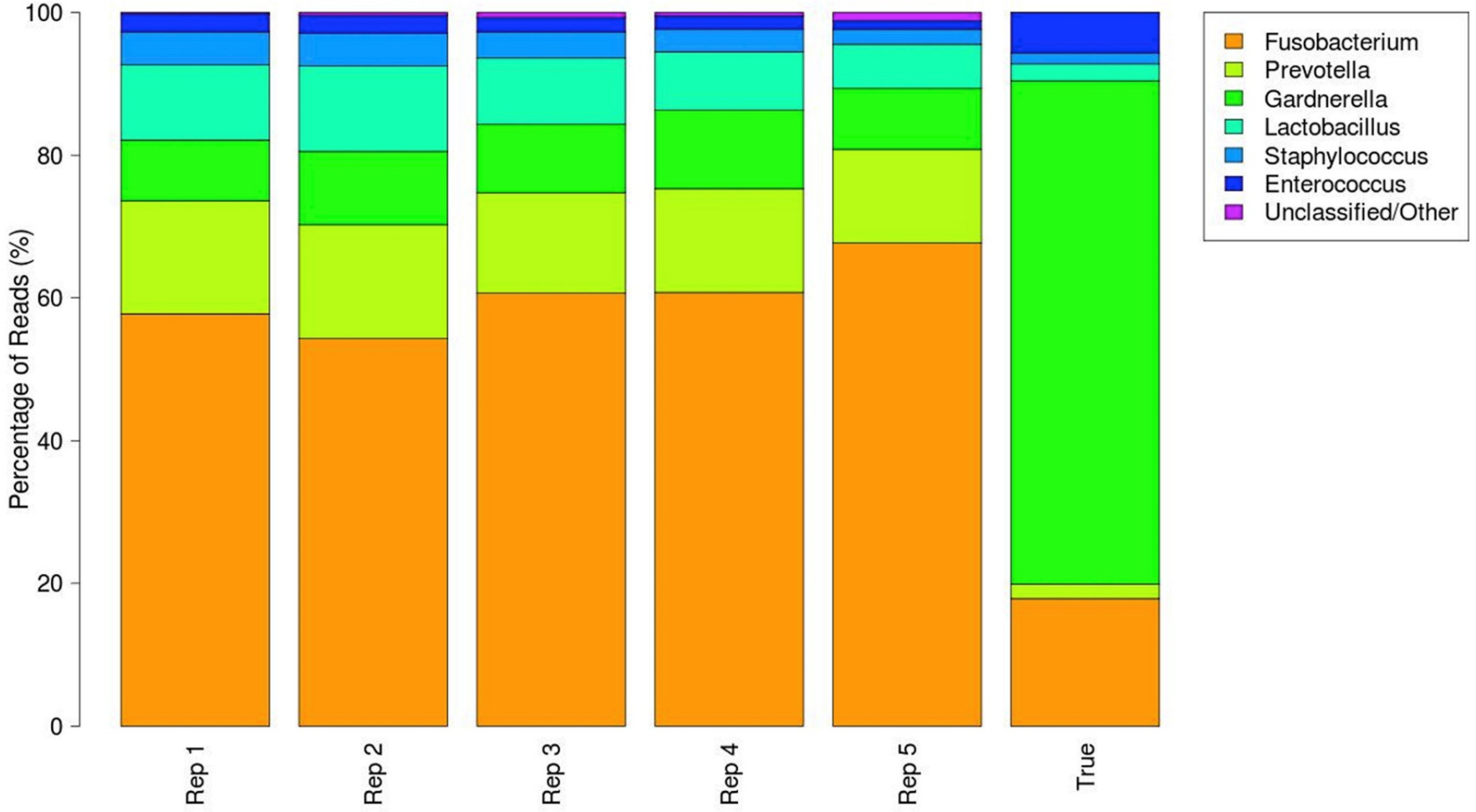
Image from [8]

Altri fattori di bias

1. L'estrazione del DNA
2. La scelta della regione ipervariabile
3. La amplificazione via PCR
4. Il sequenziamento

$$= \sum \; ???$$

Altri fattori di bias



Repliche tecniche VS realtà

Image from [3]

Altri fattori di bias

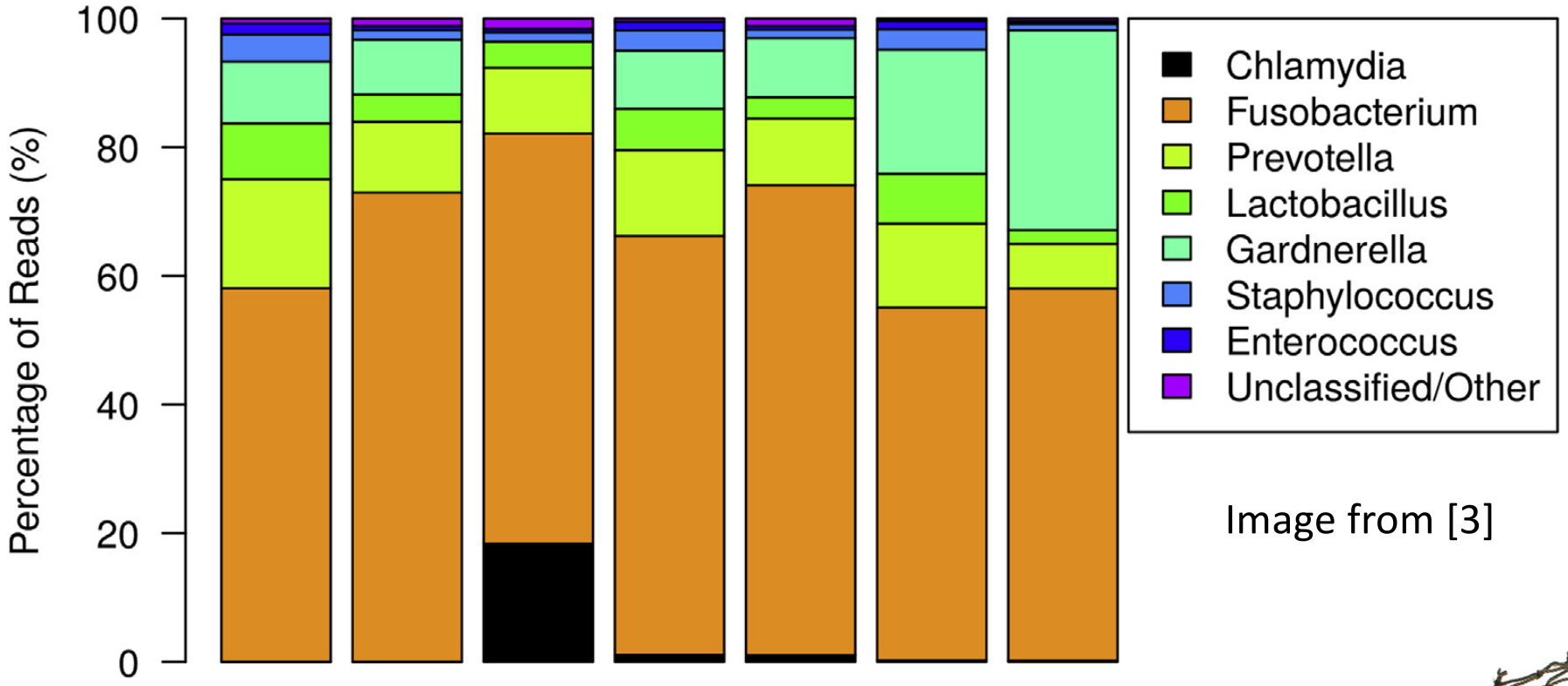
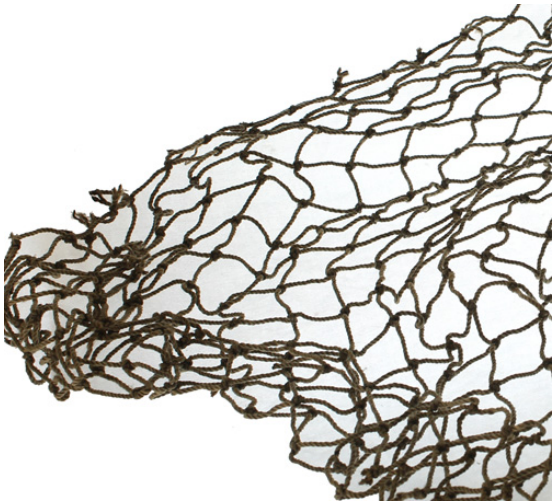


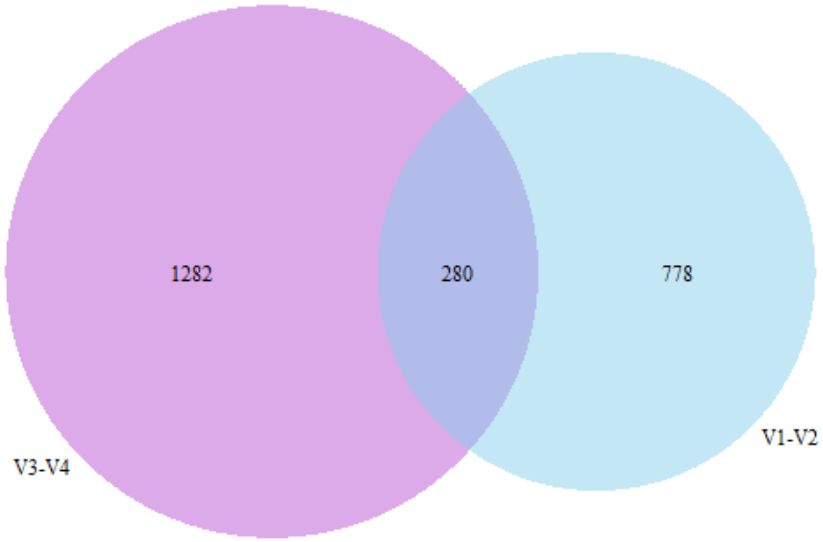
Image from [3]

L'influenza di diversi primer sets

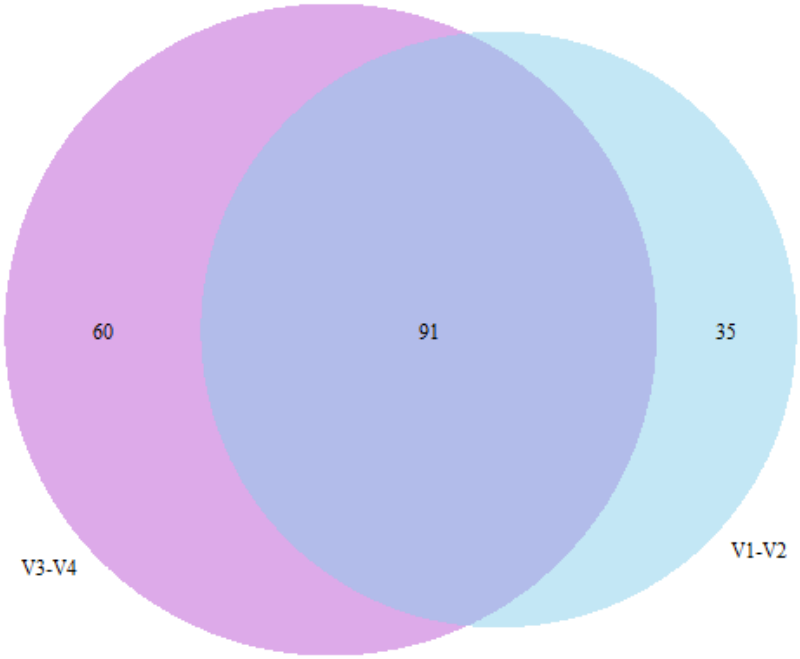


Altri fattori di bias

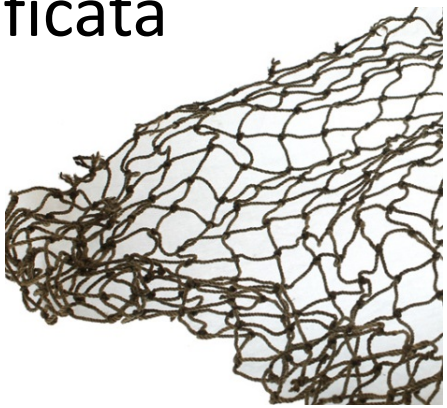
OTU level



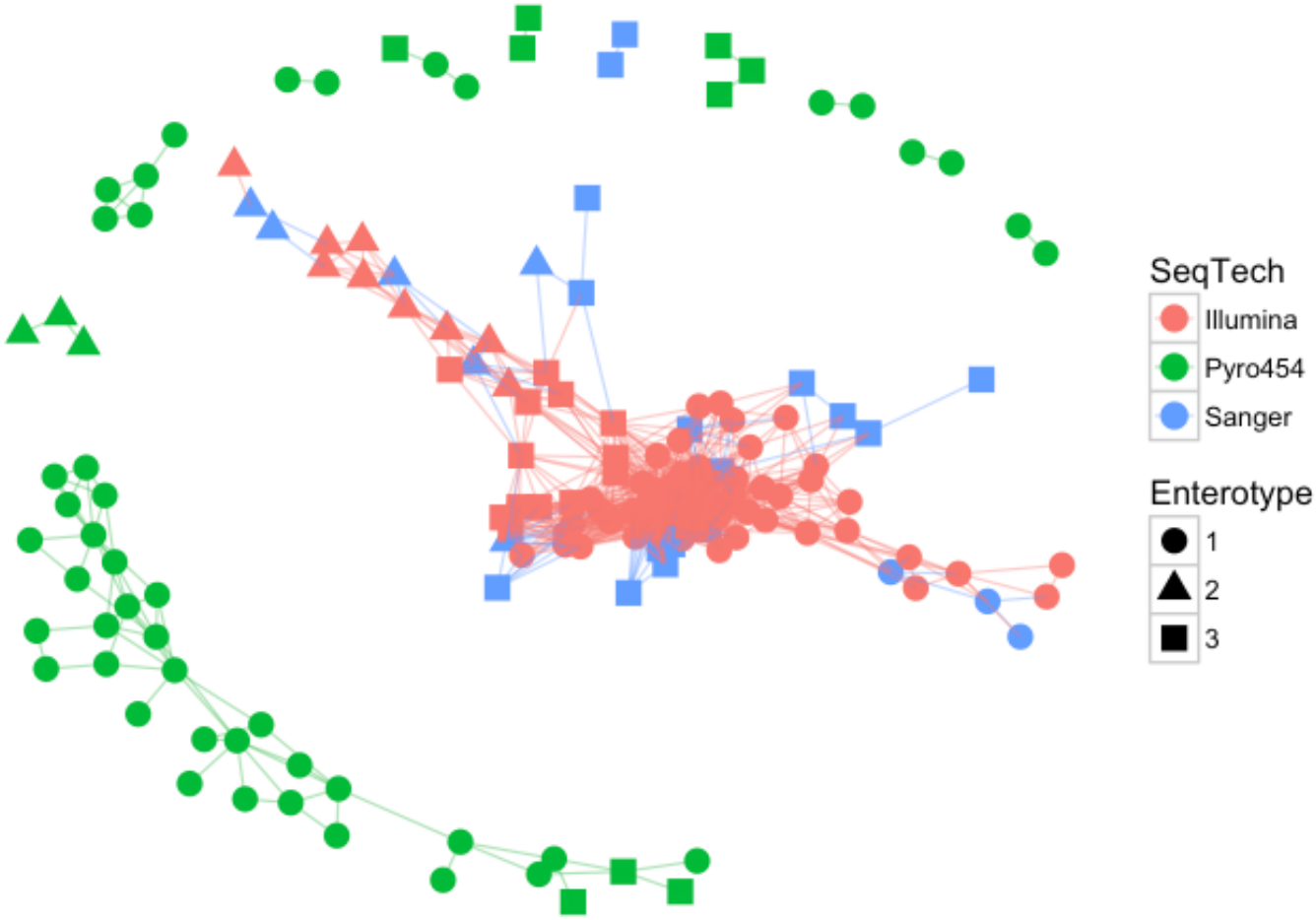
Genera level



L’influenza della regione ipervariabile amplificata



Altri fattori di bias



L’influenza della piattaforma di sequenziamento

Image from [10]

Altri fattori di bias

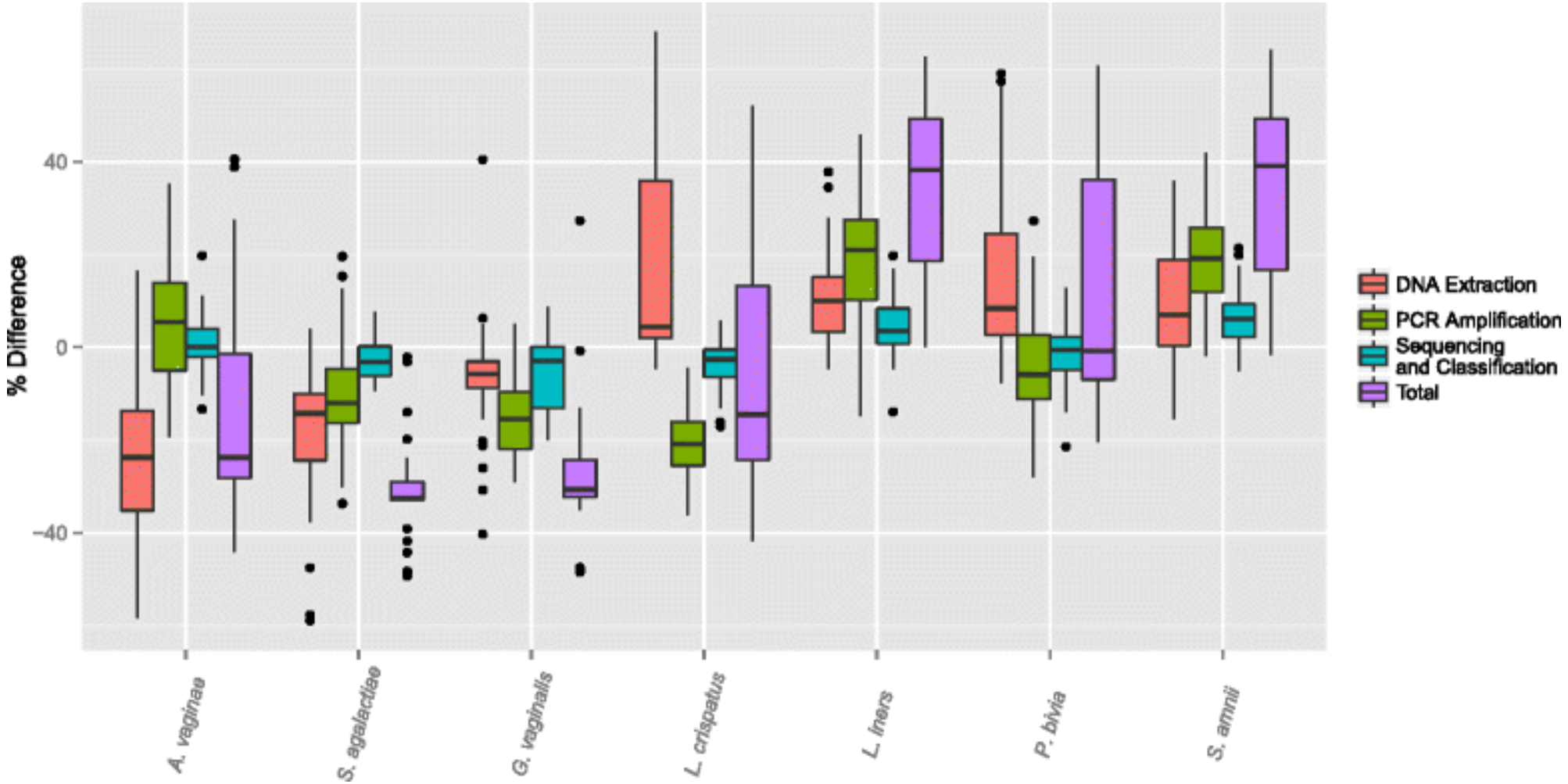


Image from [4]

Cosa fare

Cosa si può fare?

- Effettuare esperimenti di calibrazione dei protocolli usando «mock communities»
- Controllare l'*effetto batch* inserendo
  - *controlli positivi*
  - *controlli negativi*
- Testare e selezionare la regione ipervariabile da targetare in base a
  - esigenze specifiche
  - letteratura
  - ottimizzazioni successive

Wrap up

- I dataset da 16S e le loro caratteristiche
- Strategie diverse per estrarre informazioni
- Le fonti di bias da mitigare

Cosa fare

- Tenere conto, per quanto possibile, dei fattori confondenti (e mitigarli)
- Essere consapevoli dei default delle pipeline utilizzate
- Utilizzare approcci diversi in base al proprio obiettivo

Grazie dell'attenzione!

Calvin and Hobbes by Bill Watterson

Riferimenti utili

1. Tsilimigras, Matthew CB, and Anthony A. Fodor. "Compositional data analysis of the microbiome: fundamentals, tools, and challenges." *Annals of epidemiology* 26.5 (2016): 330-335.
2. Gloor, Gregory B., et al. "It's all relative: analyzing microbiome data as compositions." *Annals of epidemiology* 26.5 (2016): 322-329.
3. Brooks, J. Paul. "Challenges for case-control studies with microbiome data." *Annals of epidemiology* 26.5 (2016): 336-341.
4. Brooks, J. Paul, et al. "The truth about metagenomics: quantifying and counteracting bias in 16S rRNA studies." *BMC microbiology* 15.1 (2015): 1.

Riferimenti utili

5. McMurdie PJ, Holmes S (2014) Waste Not, Want Not: Why Rarefying Microbiome Data Is Inadmissible. PLoS Comput Biol 10(4): e1003531. doi:10.1371/journal.pcbi.1003531
6. Xu L, Paterson AD, Turpin W, Xu W (2015) Assessment and Selection of Competing Models for Zero-Inflated Microbiome Data. PLoS ONE 10(7): e0129606. doi:10.1371/journal.pone.0129606
7. Kembel, Steven W., et al. "Incorporating 16S gene copy number information improves estimates of microbial diversity and abundance." *PLoS Comput Biol*8.10 (2012): e1002743.
8. Větrovský, Tomáš, and Petr Baldrian. "The variability of the 16S rRNA gene in bacterial genomes and its consequences for bacterial community analyses." *PloS one* 8.2 (2013): e57923.

Riferimenti utili

9. Paulson, Joseph N., et al. "Differential abundance analysis for microbial marker-gene surveys." *Nature methods* 10.12 (2013): 1200-1202.
10. McMurdie, Paul J., and Susan Holmes. "phyloseq: an R package for reproducible interactive analysis and graphics of microbiome census data." *PloS one* 8.4 (2013): e61217.