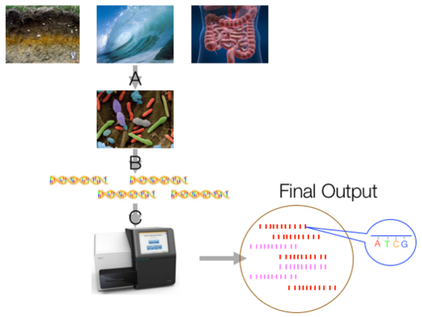
A Metagenomics Method for Simultaneously Identifying Microbial Species and Estimating their Abundance in Multiple Samples

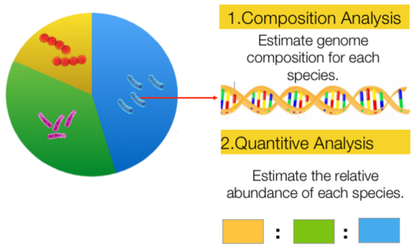
Xin Xing[1], Jun Liu[2], Wenxuan Zhong[1]

1.Department of Statistics, University of Georgia. 2. Department of Statistics, Harvard University.

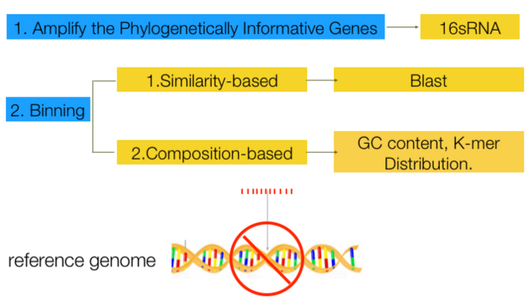
Metagenomics



Two main objectives in metagenomics



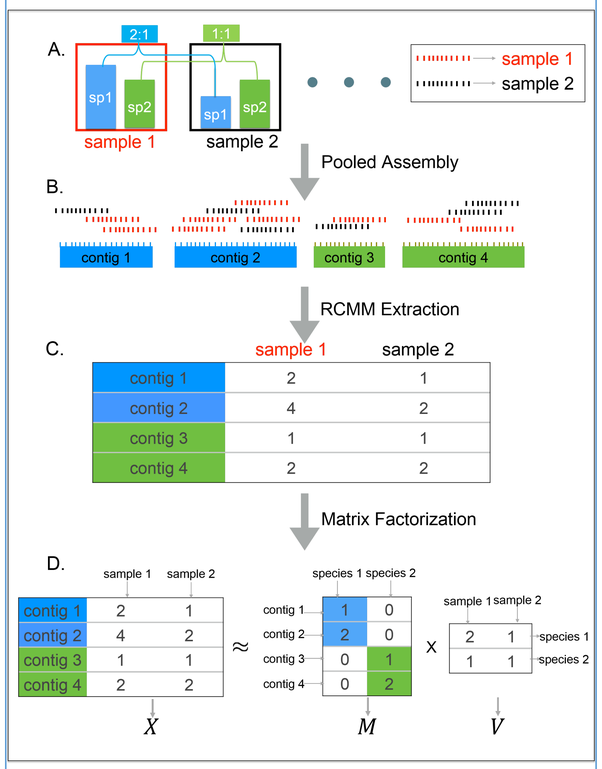
Existing Methods



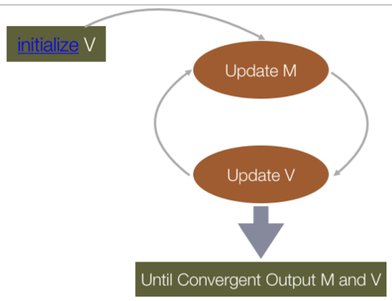
Optimization

$$
\begin{aligned}
\text{Minimize}: \quad & ||X - MV||_F \\
\text{Subject To}: \quad & ||M_i||_{L_0} < c \qquad i = 1, \ldots, n \\
& M_{ik} \geq 0 \qquad i = 1, \ldots, n \quad k = 1, \ldots, K \\
& V_{kj} \geq 0 \qquad j = 1, \ldots, p \quad k = 1, \ldots, K
\end{aligned}
$$

Where M_i is the $i'th$ row vector of matrix M.The $||M_i||_{L_0} = \#\{\text{nonzeoro element in } M_i\}$. $||A||_F = \sqrt{Trace(AA^T)}$ is the Frobenius norm of matrix.



Algorithm



The Uniqueness of the Matrix Factorization

$$
\begin{aligned}
\text{Minimize}: \quad & ||X - MV||_F \qquad (1) \\
\text{Subject To}: \quad & ||M_i||_{L_0} < c \qquad i = 1, \ldots, n \qquad (2) \\
& M_{ik} \geq 0 \qquad i = 1, \ldots, n \quad k = 1, \ldots, K \\
& V_{kj} \geq 0 \qquad j = 1, \ldots, p \quad k = 1, \ldots, K
\end{aligned}
$$

Theorem 1. (Identifiability of the decomposition) *Subject to the constraints in 2, $X = MV$ is a factorization of X and M has full column rank. If $X = ST$ is another factorization and S has full column rank then there exist a monomial matrix B such that $S = MB$ and $T = B^{-1}V$.*

Choose the number of species K

BIC-type criterion

$$f(X_{i1}, \cdots, X_{iP} | I_{ik} = 1) = \prod_{j=1}^{P} \exp^{-M_{ik}V_{kj}} \frac{X_{ij}^{M_{ik}V_{kj}}}{X_{ij}!},$$

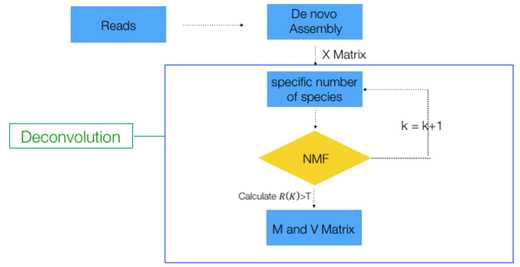
Where I_{ik} is a latent Bernoulli variable that take value one with probability τ_k

$$L(\theta; X) = \prod_{i=1}^{N} \sum_{k=1}^{K} \tau_k \prod_{j=1}^{P} \exp^{-\ell_i \alpha_k V_{kj}} \frac{X_{ij}^{-\ell_i \alpha_k V_{kj}}}{X_{ij}!},$$

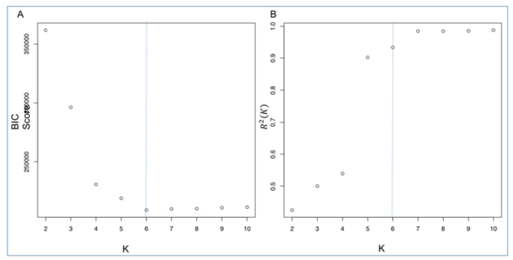
Where $M_{ik} = l_i \alpha_k$, l_i is the number of base pairs of the ith contig and α_k is propositional to the relative abundance of the kth species in the pooled sample.

$$BIC(k) = -2 \ln L(\hat{\theta}; X) - (P * K + 2K) \ln N$$

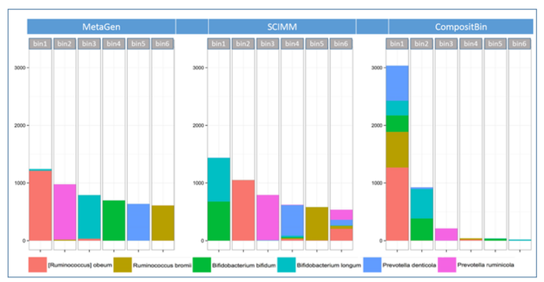
Pipeline



Empirical Performance of two criteria



Classification Accuracy (M matrix)

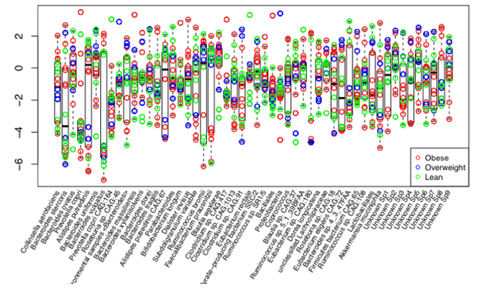


A Case Study in Obesity Development

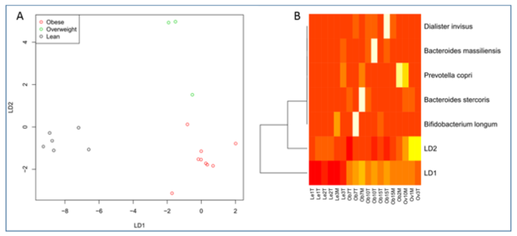
They extracted DNA from faeces of 18 samples that coming from 6 families with pairs of twins and their maternal parents. According to their body mass index (BMI), the 18 subjects were classified into three groups: obese group (BMI ≥ 30), overweight group (25 ≤ BMI ≤ 30) and lean group (BMI≤ 25).

Reference:
Peter J Turnbaugh, Micah Hamady, Tanya Yatsunenko, Brandi L Cantarel, Alexis Duncan,Ruth E Ley, Mitchell L Sogin, William J Jones, Bruce A Roe, Jason P Aourtit, et al. **A core gut microbiome in obese and lean twins**. nature.

Discovered 43 known species and 9 unknown species



Selected Obese Associated Species



Gut microbiota network