AMERICAN SOCIETY FOR MICROBIOLOGY | mSystems®

OBSERVATION

Zebra: Static and Dynamic Genome Cover Thresholds with Overlapping References

Daniel Hakim,[a,b] Stephen Wandro,[e] Karsten Zengler,[a,d,e] Livia S. Zaramela,[a] Brent Nowinski,[e] Austin Swafford,[e] Qiyun Zhu,[f,g] Se Jin Song,[e] Antonio Gonzalez,[a] Daniel McDonald,[a] Rob Knight[a,c,d,e]

[a]Department of Pediatrics, School of Medicine, University of California, San Diego, La Jolla, California, USA
[b]Bioinformatics and Systems Biology Program, University of California, San Diego, La Jolla, California, USA
[c]Department of Computer Science and Engineering, University of California, San Diego, La Jolla, California, USA
[d]Department of Bioengineering, University of California, San Diego, La Jolla, California, USA
[e]Center for Microbiome Innovation, Jacobs School of Engineering, University of California, San Diego, La Jolla, California, USA
[f]School of Life Sciences, Arizona State University, Tempe, Arizona, USA
[g]Biodesign Center for Fundamental and Applied Microbiomics, Arizona State University, Tempe, Arizona, USA

ABSTRACT Assigning taxonomy remains a challenging topic in microbiome studies, due largely to ambiguity of reads which overlap multiple reference genomes. With the Web of Life (WoL) reference database hosting 10,575 reference genomes and growing, the percentage of ambiguous reads will only increase. The resulting artifacts create both the illusion of co-occurrence and a long tail end of extraneous reference hits that confound interpretation. We introduce genome cover, the fraction of reference genome overlapped by reads, to distinguish these artifacts. We show how to dynamically predict genome cover by read count and examine our model in *Staphylococcus aureus* monoculture. Our modeling cleanly separates both *S. aureus* and true contaminants from the false artifacts of reference overlap. We next introduce saturated genome cover, the true fraction of a reference genome overlapped by sample contents. Genome cover may not saturate for low abundance or low prevalence bacteria. We assuage this worry with examination of a large human fecal data set. By compositing the metric across like samples, genome cover saturates even for rare species. We note that it is a threshold on saturated genome cover, not genome cover itself, which indicates a spurious reference hit or distant relative. We present Zebra, a method to compute and threshold the genome cover metric across like samples, a recurrence to estimate genome cover and confirm saturation, and provide guidance for choosing cover thresholds in real world scenarios. Standalone genome cover and integration into Woltka are available: https://github.com/biocore/zebra_filter, https://github.com/qiyunzhu/woltka.

IMPORTANCE Taxonomic assignment, assigning sequences to specific taxonomic units, is a crucial processing step in microbiome analyses. Issues in taxonomic assignment affect interpretation of what microbes are present in each sample and may be associated with specific environmental or clinical conditions. Assigning importance to a particular taxon relies strongly on independence of assigned counts. The false inclusion of thousands of correlated taxa makes interpretation ambiguous, leading to underconstrained results which cannot be reproduced. The importance sometimes attached to implausible artifacts such as anthrax or bubonic plague is especially problematic. We show that the Zebra filter retrieves only the nearest relatives of sample contents enabling more reproducible and biologically plausible interpretation of metagenomic data.

KEYWORDS metagenomics, microbiome, read filtering

The reference overlap problem in taxonomic assignment leads to ambiguously aligned reads. These ambiguities result in either assignment to a higher taxonomic rank leading to a loss in specificity, or equal distribution across all alignments resulting in

Editor Nicholas Chia, Mayo Clinic

Address correspondence to Rob Knight, robknight@eng.ucsd.edu.

The authors declare no conflict of interest.

Received 11 August 2022
Accepted 15 August 2022
Published 8 September 2022

TABLE 1 Ten highest genome cover reference genomes identified in *S. aureus* monocultures

OGU	Covered length	Genome length	Genome cover	Predicted genome cover	Strain	Mean depth (whole-genome)	Mean depth (covered regions)	Reads	Prevalence
G001456215	2,538,281	2,709,797	93.7%	100%	*Staphylococcus aureus* MS4	16,194.45	17,288.73	2.9E + 08	98%
G000072485	437,397	4,851,126	9.0%	10.9%	*Stenotrophomonas maltophilia* K279a	0.12	1.28	3.7E + 03	4%
G000020205	404,791	5,325,729	7.6%	9.5%	*Ralstonia pickettii* 12J	0.10	1.33	3.6E + 03	92%
G000009865	117,070	2,697,861	4.3%	100%	*Staphylococcus haemolyticus* JCSC1435	373.00	8,595.77	6.7E + 06	93%
G000007645	86,798	2,564,615	3.4%	100%	*Staphylococcus epidermidis* ATCC 12228 ASM764v1	271.25	8,014.52	4.6E + 06	93%
G000972575	75,837	2,826,849	2.7%	100%	*Staphylococcus cohnii* subsp. cohnii 532	118.91	4,432.58	2.2E + 06	91%
G000332735	65,933	2,560,716	2.6%	100%	*Staphylococcus warneri* SG1	353.99	13,748.45	6.0E + 06	94%
G001188915	60,333	2,582,931	2.3%	100%	*Staphylococcus schleiferi* 2317-03	134.30	5,749.64	2.3E + 06	91%
G001471555	59,139	2,602,401	2.3%	100%	*Staphylococcus capitis* FDAARGOS_173	179.92	7,917.26	3.1E + 06	93%
G001068545	58,172	2,531,263	2.3%	100%	*Staphylococcus epidermidis* 1056_SEPI	138.31	6,018.25	2.3E + 06	93%

assignment to extraneous species and an illusion of co-occurrence between related species (1).

KrakenUniq (2) introduces techniques to estimate and filter by unique *k*-mer count while SLIMM (3) uses preprocessing to filter by read distribution across multiple bins. However, neither tool makes use of information across samples to inform selection. Because ambiguous reads are necessarily restricted to regions where references overlap, we may use the uniformity of assigned reads to filter artifacts within and across samples overcoming the limitations of these tools to retain rare microbes that may be of phenotypic importance. We report a metric that can be composited across like samples to enable use for rare microbes and a corresponding threshold that scales dynamically by read count and reference length.

We introduce the genome cover metric, the fraction of a reference genome covered by one or more reads, as a measure of read uniformity. With sufficient reads, this metric saturates as the fraction of reference genome overlapped by sample contents. The consequence of the reference overlap problem is that an abundant species may assign read counts to both near and distant relatives within the reference, but only those closest relatives will show a high percentage of saturated genome cover. We propose a genome cover filter to remove extraneous assignments in the Web of Life (WoL) (4) from human samples.

We begin with an evaluation of monocultures where we expect taxonomic assignment to corroborate a single organism. Reads are sourced from 192 *Staphylococcus aureus* monoculture samples selected from lesion- and nonlesion tissues of the skin of human subjects suffering from atopic dermatitis. These data are available through the European Bioinformatics Institute (EBI) (https://www.ebi.ac.uk/ena) under the study identifier PRJEB52498 (ERP137223) and on Qiita (5): https://qiita.ucsd.edu/study/description/11919.

The Web of Life Toolkit App (Woltka [6]) accumulates read counts by splitting evenly across up to 16 matched references. Processing *S. aureus* monoculture in Woltka results in sporadic assignment to over 1,700 reference genomes. Table 1 displays the top 10 assignments ordered by genome cover. As expected, *S. aureus* MS4 dominates these samples by cover and assigned read count. Reference hits are generally filtered via relative abundance thresholds, the fraction of per-sample reads, e.g., 0.01%, and/or prevalence thresholds, the fraction of samples where the organism is detected, e.g., 10% as benchmarked for standard pipelines (7–9). Table 1 and Fig. 1 show these thresholds are insufficient to filter relatives of *S. aureus*.

Nearly 7 million reads were assigned to *Staphylococcus haemolyticus* JCSC1435. These *S. haemolyticus* matches were identified in >90% of samples and far exceed typical filtering thresholds. These assignments are the direct result of the reference overlap problem generating the illusion of co-occurrence of multiple *Staphylococcus* spp. within these samples. Whereas the ~300 million reads of *S. aureus* overlap 94% of its genome, those assigned to

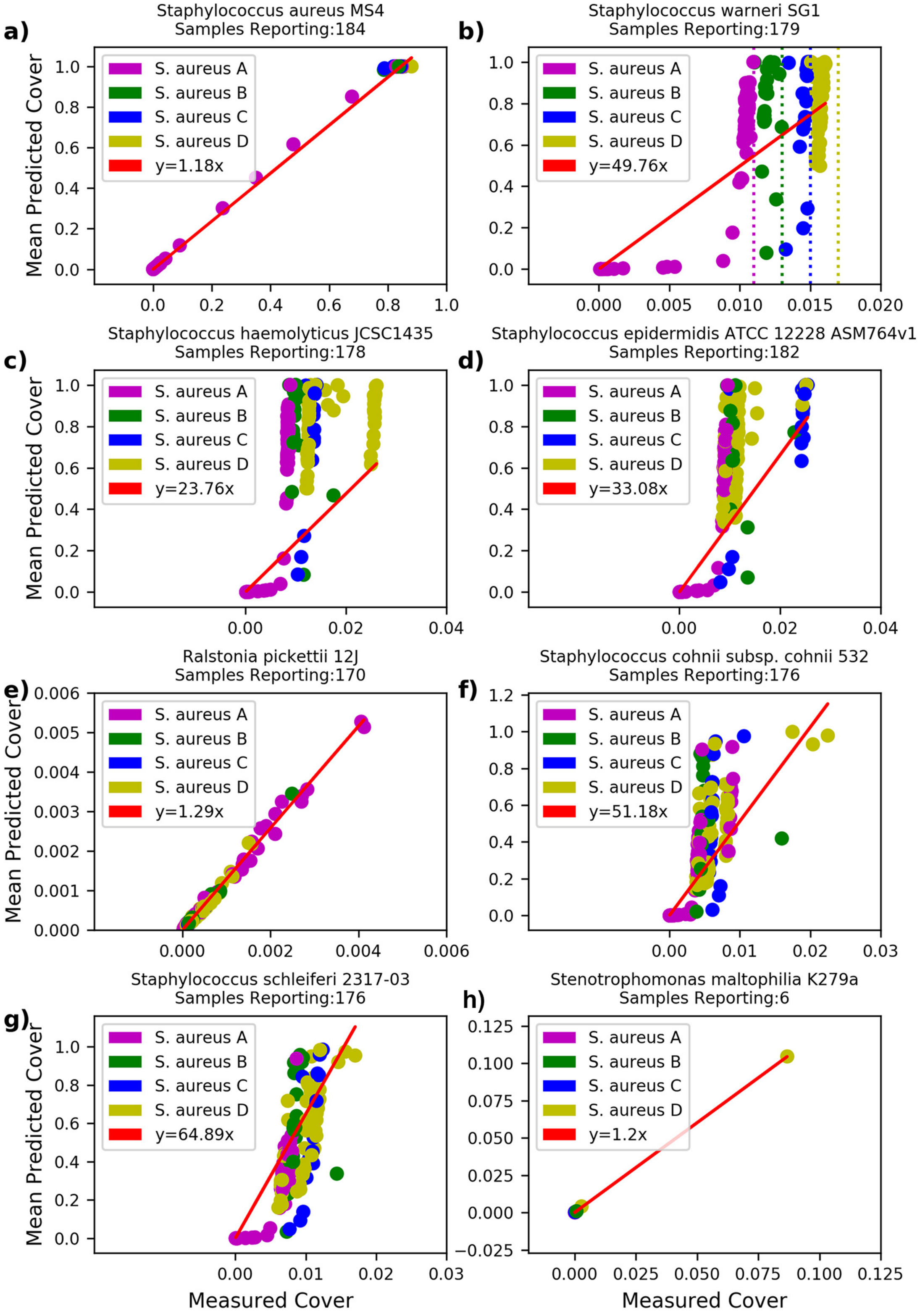

FIG 1 Modeling genome cover by read count differentiates low abundance contaminants from overlapping references in *S. aureus* monocultures. Clusters (A) to (D) determined by thresholding of *Staphylococcus warneri* SG1. Red indicates the line of best fit. A reported slope of 1 with no residual would indicate a perfect model fit. Mean predicted cover calculated using assigned read count and genome length assuming fixed 150 bp read length. As the number of reads increases, measured cover asymptotically approaches the overlap between true sample content and the assigned reference genome.

S. haemolyticus resolve to only 4% of the *S. haemolyticus* genome with high (>8000) mean depth. The high depth regions represent the overlap between the *S. aureus* and *S. haemolyticus* genomes. Analogous arguments may be made for the other *Staphylococcus* spp. in Table 1. The reference overlap problem generates artifacts restricted to overlapping regions. We exploit this to formulate a novel filtering approach.

Fig. S1 models the genome cover metric for a given number of reads of a reference species by assuming reads are uniformly distributed and discretized to nonoverlapping buckets. Closed form mean and standard deviation for the number of covered buckets are given in Equations 3 and 4 (10). In monoculture, this model allows us to filter references whose measured cover lies outside the predicted cover range.

There is a striking difference between the model performance on *S. aureus* versus other *Staphylococcus* species (Fig. 1). For more distant relatives, predicted genome cover is 20 to 60× higher than is measured, strongly indicating that reads are not uniformly distributed across these reference genomes. In contrast, comparing predicted genome cover against measured genome cover on a sample by sample basis based on this metric shows that for *S. aureus*, *Stenotrophomonas maltophilia*, and *Ralstonia pickettii* the predicted and measured cover are linearly related to within a constant factor of ~1.2 (Fig. 1). This constant factor may result from PCR-derived duplicate reads, deviation between the reference genome and the contents of the sample, and/or nonuniformity of read sampling. Thus, *Ralstonia pickettii*, whose cover is low but in agreement with the number of assigned reads, should be considered contamination rather than reference overlap.

Interpreting the x-intercepts as the saturating genome cover between *S. aureus* strains and these references, we observed vertical clusters within the *Staphylococcus* relatives (Fig. 1b) that indicates saturation with even a single high abundance sample. We use this fact to bound thresholds applicable to samples of unknown composition where, due to the possibility of co-occurrence, the assumptions of the dynamic model may not hold.

Ninety percent of reference genomes in WoL are less than 6 million bp in length. By our model, it would take roughly 12,000 reads of length 150 bp to achieve 25% genome cover for a reference of this length. Compositing 100 like samples with roughly 600,000 reads apiece, reaching this threshold requires mean relative abundance 0.02%, in line with standard relative abundance thresholds. This shows that 25% cover is a reasonable threshold for the average microbe to pass in a shallow sequenced 100-sample data set. Table S1 estimates the required number of reads across composited samples to reach a target genome cover threshold within the range of reference lengths in the WoL.

Fig. 2a shows that even the weakest cover thresholds filter 80+% of extraneous reference hits in iMSMS. We do not recommend cover thresholds below 10% however, due to the existence of reference species whose genomes highly overlap common members of the gut microbiome. Fig. 2b and c shows that *Yersinia pestis* (bubonic plague), a false hit that frequently bypasses abundance and prevalence filters (11, 12) has saturated cover around 4.5% in the International Multiple Sclerosis Microbiome Study (iMSMS [13, 14]). This 4.5% cover is due to overlap with *Escherichia coli* and *Klebsiella pneumoniae*. Similarly iMSMS covers 9.5% of *Bacillus anthracis* (*anthrax*) as the result of a few samples containing a close relative of *Bacillus thuringiensis* (Fig. 2d). As these samples are outliers, it is unclear whether 9.5% is the saturated cover.

We further caution against cover thresholds above 90%, even with deep sequencing, due to the imperfect nature of a reference. Table 1 shows a 95% cover threshold would remove even *S. aureus* from the analysis. Table S2 reports genome cover for the WoL relatives of eight ground truth species. *Salmonella enterica* B4212 only appears to overlap 83% of its nearest relative in WoL. If the reference does not contain close relatives, weaker cover thresholds should be employed.

Zebra discards the artifacts of reference overlap, catches problematic species that bypass standard abundance and prevalence filtering, and leads to improved interpretation.

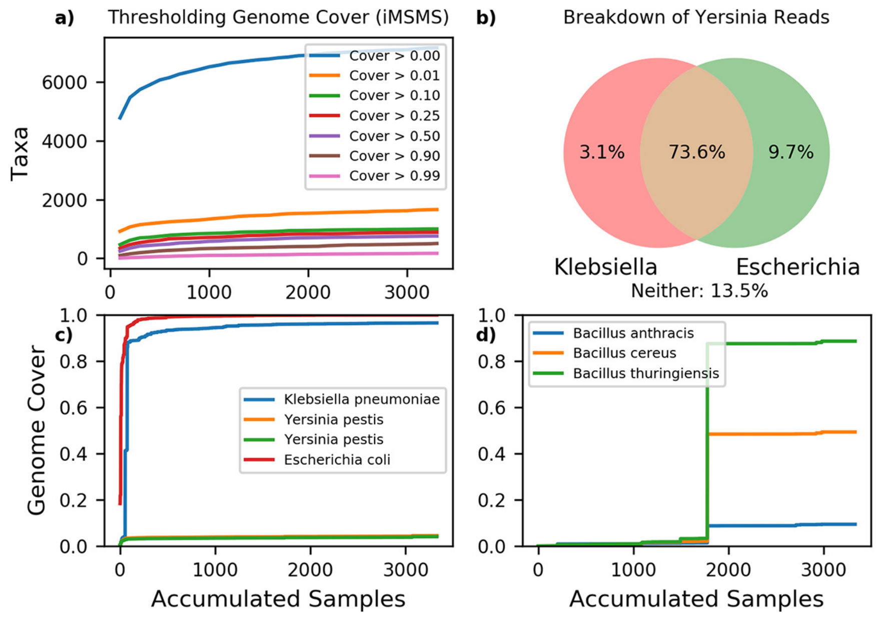

FIG 2 Cumulative genome cover in iMSMS. Genome cover for each reference taxon is accumulated across shuffled iMSMS samples. Sample depth varies from 10^5 to 10^7, median 10^6 reads. (a) Number of taxa passing cover threshold as samples are accumulated. (b) Tracing of reads assigned to *Yersinia pestis*. (c) Accumulated cover of *Yersinia pestis* and related microbes by sample. (d) Accumulated cover of *Bacillus anthracis* and related microbes by sample.

SUPPLEMENTAL MATERIAL

Supplemental material is available online only.

FIG S1, TIF file, 3.5 MB.

TABLE S1, DOCX file, 0.01 MB.

TABLE S2, DOCX file, 0.02 MB.

ACKNOWLEDGMENTS

We acknowledge Gail Ackermann for her support and expertise. This work was supported in part by NIH grants U24CA248454 and 1DP1AT010885 and was supported by a Seed Grant from the Center for Microbiome Innovation at the University of California San Diego to L.Z. and K.Z. We declare that we have no competing interests.

REFERENCES

1. Evans JT, Denef VJ. 2020. To dereplicate or not to dereplicate? mSphere 5. https://doi.org/10.1128/mSphere.00971-19.
2. Breitwieser FP, Baker DN, Salzberg SL. 2018. KrakenUniq: confident and fast metagenomics classification using unique k-mer counts. Genome Biol 19:1–10. https://doi.org/10.1186/s13059-018-1568-0.
3. Dadi TH, Renard BY, Wieler LH, Semmler T, Reinert K. 2017. SLIMM: species level identification of microorganisms from metagenomes. PeerJ 5:e3138. https://doi.org/10.7717/peerj.3138.
4. Zhu Q, Mai U, Pfeiffer W, Janssen S, Asnicar F, Sanders JG, Belda-Ferre P, Al-Ghalith GA, Kopylova E, McDonald D, Kosciolek T, Yin JB, Huang S, Salam N, Jiao J-Y, Wu Z, Xu ZZ, Cantrell K, Yang Y, Sayyari E, Rabiee M, Morton JT, Podell S, Knights D, Li W-J, Huttenhower C, Segata N, Smarr L, Mirarab S, Knight R. 2019. Phylogenomics of 10,575 genomes reveals evolutionary proximity between domains bacteria and archaea. Nat Commun 10:5477. https://doi.org/10.1038/s41467-019-13443-4.
5. Gonzalez A, Navas-Molina JA, Kosciolek T, McDonald D, Vázquez-Baeza Y, Ackermann G, DeReus J, Janssen S, Swafford AD, Orchanian SB, Sanders JG, Shorenstein J, Holste H, Petrus S, Robbins-Pianka A, Brislawn CJ, Wang M, Rideout JR, Bolyen E, Dillon M, Caporaso JG, Dorrestein PC, Knight R. 2018. Qiita: rapid, web-enabled microbiome meta-analysis. Nat Methods 15:796–798. https://doi.org/10.1038/s41592-018-0141-9.
6. Zhu Q, Huang S, Gonzalez A, McGrath I, McDonald D, Haiminen N, Armstrong G, Vázquez-Baeza Y, Yu J, Kuczynski J, Sepich-Poore GD, Swafford AD, Das P, Shaffer JP, Lejzerowicz F, Belda-Ferre P, Havulinna AS, Méric G, Niiranen T, Lahti L, Salomaa V, Kim H-C, Jain M, Inouye M, Gilbert JA, Knight R. 2022. Phylogeny-aware analysis of metagenome community ecology based on matched reference genomes while bypassing taxonomy. mSystems 7:e0016722. https://doi.org/10.1128/msystems.00167-22.
7. Ye SH, Siddle KJ, Park DJ, Sabeti PC. 2019. Benchmarking metagenomics tools for taxonomic classification. Cell 178:779–794. https://doi.org/10.1016/j.cell.2019.07.010.
8. Mallick H, Rahnavard A, McIver LJ, Ma S, Zhang Y, Nguyen LH, Tickle TL, Weingart G, Ren B, Schwager EH, Chatterjee S, Thompson KN, Wilkinson JE, Subramanian A, Lu Y, Waldron L, Paulson JN, Franzosa EA, Bravo HC, Huttenhower C. 2021. Multivariable association discovery in population-scale meta-omics studies. PLoS Comput Biol 17:e1009442. https://doi.org/10.1371/journal.pcbi.1009442.
9. Huttenhower C. 2021. maaslin2 – The Huttenhower Lab. https://huttenhower.sph.harvard.edu/maaslin/.
10. Henry. 2022. Probability distribution of coverage of a set after X independently, randomly selected members of the set. Mathematics Stack Exchange [Internet]. https://math.stackexchange.com/questions/32800/probability-distribution-of-coverage-of-a-set-after-x-independently-randomly.

11. Hsu T, Joice R, Vallarino J, Abu-Ali G, Hartmann EM, Shafquat A, DuLong C, Baranowski C, Gevers D, Green JL, Morgan XC, Spengler JD, Huttenhower C. 2016. Urban transit system microbial communities differ by surface type and interaction with humans and the environment. mSystems 1. https://doi.org/10.1128/mSystems.00018-16.

12. Gonzalez A, Vázquez-Baeza Y, Pettengill JB, Ottesen A, McDonald D, Knight R. 2016. Avoiding pandemic fears in the subway and conquering the platypus. mSystems 1. https://doi.org/10.1128/mSystems.00050-16.

13. The iMSMS Consortium. 2020. Household paired design reduces variance and increases power in multi-city gut microbiome study in multiple sclerosis. Mult Scler 27.

14. iMSMS. IMSMS [Internet]. http://imsms.org/?page_id=21(accessed April 14, 2022).