nature research

Corresponding author(s): Aleksandar D. Kostic and George M. Church

Last updated by author(s): Apr 25, 2019

Reporting Summary

Nature Research wishes to improve the reproducibility of the work that we publish. This form provides structure for consistency and transparency in reporting. For further information on Nature Research policies, see Authors & Referees and the Editorial Policy Checklist.

Statistics

For all statistical analyses, confirm that the following items are present in the figure legend, table legend, main text, or Methods section.

n/a	Confirmed	
☐	☒	The exact sample size (*n*) for each experimental group/condition, given as a discrete number and unit of measurement
☐	☒	A statement on whether measurements were taken from distinct samples or whether the same sample was measured repeatedly
☐	☒	The statistical test(s) used AND whether they are one- or two-sided *Only common tests should be described solely by name; describe more complex techniques in the Methods section.*
☐	☒	A description of all covariates tested
☐	☒	A description of any assumptions or corrections, such as tests of normality and adjustment for multiple comparisons
☐	☒	A full description of the statistical parameters including central tendency (e.g. means) or other basic estimates (e.g. regression coefficient) AND variation (e.g. standard deviation) or associated estimates of uncertainty (e.g. confidence intervals)
☐	☒	For null hypothesis testing, the test statistic (e.g. *F*, *t*, *r*) with confidence intervals, effect sizes, degrees of freedom and *P* value noted *Give P values as exact values whenever suitable.*
☐	☒	For Bayesian analysis, information on the choice of priors and Markov chain Monte Carlo settings
☐	☒	For hierarchical and complex designs, identification of the appropriate level for tests and full reporting of outcomes
☐	☒	Estimates of effect sizes (e.g. Cohen's *d*, Pearson's *r*), indicating how they were calculated

Our web collection on statistics for biologists contains articles on many of the points above.

Software and code

Policy information about availability of computer code

Data collection	no software was used
Data analysis	Unless otherwise noted, all plots were generated in R version 3.4.1 with the ggplot2, dplyr, scales, grid, and reshape2 packages. Large scale data analysis was done on AWS utilizing machines running Ubuntu 16.04. Data curation methods were coded in python version 2.7.12. Unless otherwise noted specifically in the rest of the methods section, code utilized is available to access at https://github.com/kosticlab/athlete and https://github.com/kosticlab/aether. Putative taxonomic abundances were calculated with Metaphlan2. 16S reads were processed with the dada2 pipeline and phyloseq. 16S Veillonella relative abundance modeling for athletes participating in the marathon was done with the R nlme package. Coefficients were were created with the coefplot2 package. Phylogenetic trees were generated from NCBI taxonomy and visualized with phylo.io. All steps in the the processing of raw metagenomic data were done utilizing the Aether package. Raw reads were de novo assembled using megahit. Open reading frames and annotations were generated using prokka. A gene family catalog was generated from the called open reading frames at 95% identity utilizing the CD-HIT software package. A raw abundance count matrix was generated utilizing the gene family catalog, bowtie2, and samtools. The raw abundance count matrix was normalized both by sample and by gene length. Metabolic pathways were queried using MetaCyc and EC IDs were pulled from prokka annotations. Heatmaps were generated with the pheatmap package in R.R was utilized to perform the majority of statistical tests with the exception of pairwise ANOVA tests, for which the SciPy library in python was used. Root mean square error calculations were performed using the plotrix package. The canopy algorithm utilized is available here: https://bitbucket.org/HeyHo/mgs-canopy-algorithm/wiki/Home. Reactions involved with the breakdown of lactate to both propionate and acetate were manually associated with EC IDs using MetaCyc.

For manuscripts utilizing custom algorithms or software that are central to the research but not yet described in published literature, software must be made available to editors/reviewers. We strongly encourage code deposition in a community repository (e.g. GitHub). See the Nature Research guidelines for submitting code & software for further information.

Data

Policy information about availability of data

All manuscripts must include a data availability statement. This statement should provide the following information, where applicable:

- Accession codes, unique identifiers, or web links for publicly available datasets
- A list of figures that have associated raw data
- A description of any restrictions on data availability

All raw sequencing data has been uploaded to NCBI and SRA in the form of the BioProjects PRJNA472785 (16S) and PRJNA472768 (MGX) which are linked to associated BioSamples which are in turn linked to the paired end read files on SRA and correspond to the metadata in the supplement. Additionally, all sequencing files are also available to download from a web browser at https://s3.amazonaws.com/athlete-sequencing-data/index.html.

Field-specific reporting

Please select the one below that is the best fit for your research. If you are not sure, read the appropriate sections before making your selection.

☒ Life sciences ☐ Behavioural & social sciences ☐ Ecological, evolutionary & environmental sciences

For a reference copy of the document with all sections, see nature.com/documents/nr-reporting-summary-flat.pdf

Life sciences study design

All studies must disclose on these points even when the disclosure is negative.

Sample size	No sample size calculation was performed. Leading up to the 2015 Boston Marathon, we attempted to recruit as many participating athletes as possible and collect as many samples as possible from each participant one week before and one week after the race. An equal amount of sedentary controls were then recruited for participation. These sample sizes were sufficient to observe a statistically robust association between Veillonella abundance and time after marathon.
Data exclusions	No data were excluded from the analysis.
Replication	Animal research studies were performed a minimum of three times. All attempts at replication were successful.
Randomization	All animals were randomly allocated into experimental groups.
Blinding	Investigators were blinded to group allocation during data collection and analysis.

Reporting for specific materials, systems and methods

We require information from authors about some types of materials, experimental systems and methods used in many studies. Here, indicate whether each material, system or method listed is relevant to your study. If you are not sure if a list item applies to your research, read the appropriate section before selecting a response.

Materials & experimental systems

n/a	Involved in the study
☒	☐ Antibodies
☒	☐ Eukaryotic cell lines
☒	☐ Palaeontology
☐	☒ Animals and other organisms
☐	☒ Human research participants
☒	☐ Clinical data

Methods

n/a	Involved in the study
☒	☐ ChIP-seq
☒	☐ Flow cytometry
☒	☐ MRI-based neuroimaging

Animals and other organisms

Policy information about studies involving animals; ARRIVE guidelines recommended for reporting animal research

Laboratory animals	All laboratory animals were male Mus musculus strain C57BL/6J aged 8-12 weeks.
Wild animals	The study did not involve wild animals.
Field-collected samples	The study did not involve samples collected from the field.
Ethics oversight	Animal research was approved by the Joslin Diabetes Center IACUC. We have complied with all relevant ethical regulations.

Note that full information on the approval of the study protocol must also be provided in the manuscript.

Human research participants

Policy information about studies involving human research participants

Population characteristics	Athlete 26 total (15 marathon) 65% female (73%) Average age is 27.4 (27.1) Average BMI = 22.4 (22.4) 85% white, 11.5% asian, 4% mixed (73%, 20%, 7%, respectively) 100% from US (96%) Controls 12 total (11 marathon study) 50% female (55%) Average age = 29.5 (29.2) Average BMI = 22.7 (22.9) 50% white, 33% asian, 8% black (45%, 36%, 9% respectively) 58% from US (55%)
Recruitment	Participants were recruited via IRB approved flyers and word of mouth. Our study design relied on prospective selection of the participants, therefore the concern of self-selection bias with regard to outcomes has been eliminated. Each potential participant went though a screening process prior to enrollment to assess study exclusion criteria, which included recent use of antibiotics, pregnancy, travel abroad, or recent inflammatory bowel states. Additionally, participants were determined to be athletes or controls based on screening questions that determined frequency of exercise and training for upcoming athletic competitions.
Ethics oversight	All study participants were recruited following an IRB approved Sports Genomics protocol (#IRB15-0869), conducted at the Wyss Institute for Biologically Inspired Engineering. Each participant read and signed a consent form prior to study enrollment. We have complied with all relevant ethical regulations.

Note that full information on the approval of the study protocol must also be provided in the manuscript.