METHODS

Study population

Elite professional male athletes (n = 40) and healthy controls (n = 46) matched for age and gender were enrolled for a cross-sectional analysis of the impacts of rigorous physical activity and diet associated with intense athleticism on enteric microbial composition. Recruitment of participants took place in 2011 as previously described in the study.[1] Due to the range of physiques within a rugby team (player position dictates need for a variety of physical constitutions, i.e. forward players tend to have larger BMI values than backs, often in the overweight/obese range) the recruited control cohort was subdivided into two groups. In order to more completely include control participants, the BMI parameter for group inclusion was adjusted to BMI ≤ 25.2 and BMI ≥ 26.5 for the low BMI and high BMI groups respectively. Approval for this study was granted by the Cork Clinical Research Ethics Committee.

Acquisition of clinical and dietary data

Self-reported dietary intake information was accommodated by a research nutritionist within the parameters of a food frequency questionnaire (FFQ) in conjunction with a photographic food atlas as per the initial investigation.[1] Fasting blood samples were collected and analysed at the clinical laboratories of the Mercy University Hospital, Cork. As the athletes were involved in a rigorous training camp we needed to assess the physical activity levels of both control groups. To determine this we used an adapted version of the EPIC-Norfolk questionnaire. Creatine kinase levels were used as a proxy for level of physical activity across all groups.

Collection and processing of biological samples

Upon initial collection, stool and urine samples were stored on ice prior to DNA extraction and purification from the fresh stool using the QIAmp DNA Stool Mini Kit (cat. no. 51504 Qiagen, Crawley, West Sussex, UK),[2] after which samples were stored securely at -80° C. DNA extraction was carried out in accordance with the manufacturer's protocol with the addition of a zirconia bead (11079101z-BSP, 11079110z-BSP, 11079125z-BSP Stratech Scientific) cell disruption bead-beating step (30s X 3). Extracted DNA was stored at -20° C prior to its initial use in 16S rRNA amplicon sequencing, after which DNA samples were stored at -80° Celsius until employment in this current study.

Metagenomic library preparation

Metagenomic library preparation was performed with the Illumina Nextera XT DNA Library Preparation Kit (cat# FC-131-1096, Illumina Inc., USA) in accordance with the manufacturer's protocol (15031942, Illumina). Normalisation of samples to the recommended 0.2 ng/µL per individual library was carried out with the ThermoFisher Qubit 2.0 Flurometric Quantitation system (Q32854, ThermoFisher). Tagmentation and amplification carried out with G-STORM GS1 thermal cycler system. Following the combined enzymatic fragmentation and adapter sequence tagging—tagmentation—and the subsequent amplification of the tagmented DNA, libraries were purified with the AMPure magnetic bead system at a ratio of 1:1.8 (DNA:AMPure) (9A63880, Beckman Coulter). Subsequently, libraries were assessed for appropriate fragment size (~500bp) on the Agilent 2100 Bioanalyzer system. With the libraries passing quality and fragment length requirements, the library preparation was continued on through library normalization, which was met with an additional assessment of suitable molar concentrations (~2 nM) with the KAPA Library Quantification Kit (KK4824, Kapabiosystems) run on a Roche LightCycler 480 (Roche Applied Science). Samples were combined into 8 final pools prior to being shipped on dry ice for sequencing.

Metagenomic Sequencing

Metagenomic libraries from the 86 participants were sequenced on the Illumina Hiseq 2500 (chemistry v4.0) NGS platform by Eurofins Genetic Services Ltd (Ebersberg, Germany). High throughput sequencing was performed on high-output run mode for 2 x 125 bp paired-end reads with the addition of a PhiX library (1%) to estimate sequence quality. Sequencing yielded a total of 344.409 Giga base pairs (Gbp) of raw unfiltered sequences, with an average of 4.15 Gbp (± 1.35 SD) per library and a mean Q30 score of 93.98 (±5.96 SD).

Metagenomic Statistical and Bioinformatics Analysis

Delivered raw FASTQ sequence files were quality checked as follows: contaminating sequences of human origin were first removed through the NCBI Best Match Tagger (BMTagger). Poor quality and duplicate read removal, as well as trimming was implemented using a combination of SAM and Picard tools. Processing of raw sequence data produced a total of 2,803,449,392 filtered reads with a mean read count of 32,598,248.74 (± 10,639,447 SD) per each of the 86 samples. These refined reads were then subjected to functional profiling by the most recent iteration of the Human Microbiome Project (HMP) Unified Metabolic Analysis Network (HUMAnN2 v. 0.5.0) pipeline.[3] The functional profiling performed by HUMAnN2 composed tabulated files of microbial metabolic pathway abundance and coverage derived from the Metacyc database.[4] Microbial pathway data was statistically analysed in the R software environment(v. 3.2.2).[5] PCoA of pathway abundances was compiled with Bray-Curtis index of dissimilarity using the R packages Vegan(v. 2.3-1)[6] & Car. Kruskal-Wallis H test was implemented with compareGroups (v. 2.0) package to appraise pathway variability between athletes and controls.[7] Similarly, Kruskal—Wallis H test derived statistics were produced on PCoA dissimilarity matrices. Semi-supervised PCA-CA-kNN was created using the KODAMA R package (v. 0.0.1).[8]

Pathway correlation plots were compiled with the Corrplot Rpackage (v. 0.73).[9] For participants of which full dietary and clinical data were available (Athletes n = 40, Low BMI control n = 22, and High BMI control n = 20), cor.test of the R package stats was used to perform Pearson product-moment correlation of metagenomics pathways to clinical data with Benjamini-Hochberg False Discovery Rate multiple testing p value adjustment performed with p.adjust, also from the stats package. This process was similarly applied to associations between metabolites and metagenomics pathways. Heatmaps and bargraphs of metagenomics data were generated using ggplot2 (v. 2.1.0).[10] All presented p values were corrected for multiple comparisons using the Benjamini-Hochberg False Discovery Rate (pFDR) method.[11]

Sample preparation for metabonomic analysis

Frozen urine samples (-80°C) were thawed, vortexed and then centrifuged at 1600 × g for 10 minutes to remove particulates and precipitated proteins. Urine samples were prepared for metabolic profiling analysis by reversed phase (RP) and hydrophilic interaction chromatography (HILIC) ultra performance liquid chromatography – mass spectrometry (UPLC-MS) as follows: 200 µl of supernatant was diluted (1:1) with high purity (HPLC grade) water, vortexed, centrifuged at 2700 × g for 20 min and aliquoted for HILIC and RP methods. Quality control (QC) samples were prepared by pooling 50 µl volumes of each sample. During the analysis, the samples were maintained at 4^{o}C in the autosampler. For ^{1}H NMR spectroscopy, 540 µL of urine samples were mixed with 60 µL of phosphate buffer (pH 7.4, 80% D2O) containing 1 mM of the internal standard, 3-(trimethylsilyl)-[2,2,3,3,-2H4]-propionic acid (TSP) and 2mM sodium azide (Na^{3}N), as described previously.[12]

Frozen faecal samples (-80^{o}C) underwent x2 freeze thaw cycles. After thoroughly defrosting, 100mg of homogenised sample was placed in a microtube containing 250 µl of 25% acetonitrile (1 ACN : 3 H_2O), 2mM sodium azide and ~0.05g 1mm Zirconia beads. The microtubes underwent 10 seconds in a Biospec bead beater. Samples were then centrifuged at 16000 x g for 20 mins. Following this the faecal water supernatant was centrifuged through centrifuge tube filters (cellulose acetate membrane, pore size 0.22 µm) to remove any remaining particular matter. The centrifuge tube filters were washed prior to use three times with 25% acetonitrile. The resulting faecal water was prepared for metabolic profiling analysis by HILIC and bile acid profiling UPLC-MS as follows: 150ul of faecal water was diluted 3:1 with acetonitrile. Samples were vortexed and incubated at -20^{o}C for 1 hour. Following this, samples were centrifuged at 4^{o}C at 16000 x g for 1 hour. Quality control (QC) samples were prepared by pooling 20 µl volumes of each faecal water sample and then preparing as above. For ^{1}H nuclear magnetic resonance (NMR) spectroscopy, 50 µl of the filtered faecal water was added to a Pyrex glass tube, which was placed under Nitrogen gas flow for 30 mins or until all the liquid had evaporated. The dried sample was reconstituted with 540 µl of D2O and 60 µl of phosphate buffer solution as described above. The solution was mixed and sonicated for 5 minutes before undergoing further centrifugation at 14000 RPM for 10 mins before 600 µl supernatant was transferred to a NMR tube for ^{1}H-NMR spectral acquisition.

Faecal samples were prepared for targeted analysis of short-chain fatty acids (SCFA) using gas chromatography – mass spectrometry (GC-MS) as previously described.[13] In brief, 100mg of thawed faecal sample was suspended in 1ml of water with 0.5% phosphoric acid. After acidification, samples were vortexed for 2 min and centrifuged for 10 min at 16000 x g. 1ml of the resulting faecal water supernatant was added to 1 ml of ethyl acetate for 2 min and then centrifuged for 10 min at 16000 x g. Prior to analysis, a 600ul volume of the organic phase was transferred into a silanised vial with 4-methyl valeric acid added as the internal standard (IS) at a final concentration of 500uM. Calibration curves of the measured SCFA were derived through analysis of duplicate dilution series of the purchased chemical standards at the beginning and end of the run.

LC-MS Metabolic profiling analysis

Reversed-phased (RP), HILIC and bile acid UPLC-MS metabolic profiling experiments were performed using a Waters Acquity Ultra Performance LC system (Waters, Milford, MA, USA) coupled to Xevo G2 Q-TOF mass spectrometer (Waters, Milford, MA, USA) with an electrospray source. Samples were analysed in a random order, with QCs every ten samples. Urine samples were first analysed using UPLC-MS, with a RP chromatographic method with both positive and negative MS ionisation modes. Secondly, to separate and detect more polar molecules, a HILIC chromatographic stage was used with positive MS ionisation modes. Faecal water samples underwent analysis using HILIC and bile acid profiling chromatographic methods in positive and negative ionisation modes respectively.

HILIC, Reversed-Phase and bile acid profiling liquid chromatographic separation was performed as previously described.[14, 15] Mass spectrometry was performed with the following settings: capillary and cone voltages were set at 1.5 kV and 30 V, respectively. The desolvation gas was set to 1000 L/hr at a temperature of 600°C; the cone gas was set to 50 L/hr and the source temperature was set to 120°C. For mass accuracy a lock-spray interface was used with leucine enkephalin [556.27741 Da ([M+H]+), 554.2615 Da ([M-H]-)] solution used as the lock mass at a concentration of 2000 ng/ml and at a flow rate of 15 µl/min.

^{1}H-NMR Metabolic profiling analysis

^{1}H-NMR spectroscopy was performed on the aqueous phase extracts at 300 K on a Bruker 600 MHz spectrometer (Bruker Biospin, Germany) using the following standard onedimensional pulse sequence: RD – g_{z1} – 90° – t_1 – 90° – t_m – g_{z2} – 90° – ACQ.[12] The relaxation delay (RD) was set at 4 s, 90° represents the applied 90° radio frequency pulse, interpulse delay (t_1) was set to an interval of 4 μs, mixing time (tm) was 10 ms, magnetic field gradients (g_{z1} and g_{z2}) were applied for 1 ms and the acquisition period (AQA) was 2.7 s. Water suppression was achieved through irradiation of the water signal during RD and t_m. For the urine samples, each spectrum was acquired using 4 dummy scans followed by 32 scans while faecal spectrum were acquired using 256 scans and 4 dummy scans and collected into 64K data points. A spectral width of 12 000Hz was used for all the samples. Prior to Fourier transformation, the FIDs were multiplied by an exponential function corresponding to a line broadening of 0.3 Hz.

GC-MS SCFA targeted analysis

The GC-MS targeted SCFA analysis was conducted on an Agilent 7890B GC system, coupled to an Agilent 5977A mass selective detector (Agilent Technologies, USA). The analysis was performed to detect levels of the SCFAs acetate, propionate, butyrate, valerate, isobutyrate, isovalerate, according to a previously described method.[13] The detector was operated in selected ion monitoring (SIM) mode (electron energy 70 eV), scanning the selected characteristic target ion for each measured SCFA (acetate, propionate, butyrate, valerate, isobutyrate, and isovalerate), at the corresponding retention times. Retention times were confirmed prior to analysis through analysis of authentic SCFAs in full scan mode. Samples were analysed in a random order with QCs every ten samples.

LC-MS data treatment

The raw mass spectrometric data acquired were pre-processed using xcms in R and the centwave peak picking method was used to detect chromatographic peaks.[16] The xcms-centwave parameters were dataset specific. Feature grouping across samples was performed using the ‘nearest’ method within xcms. Peak filling, MinFrac (0.5) and QC covariance (0.3) filters were applied to the features. Data was normalised using median fold change normalisation using the median data set as the reference.[17]

^{1}H-NMR data treatment

^{1}H-NMR spectra were automatically corrected for phase and baseline distortions and referenced to the TSP singlet at δ 0.0 using TopSpin 3.1 software. Spectra were then digitized into 20,000 data points at a resolution of 0.0005ppm using an in-house MATLAB R2014a (Mathworks) script. Subsequently, spectral regions corresponding to the internal standard (δ -0.5 to 0.5) and water (δ 4.6 to 5) peaks were removed. In addition, urea (δ 5.4 to 6.3) was removed from the urinary spectra. All spectra were normalised using median fold change normalisation using the median spectrum as the reference.[17]

GC-MS data treatment

GC-MS data was processed using MassHunter Quantitative Analysis (Agilent Technologies) software. Extracted ion chromatograms of the target ion selected for each SCFA were integrated and the peak area was normalised to the internal standard (4-methyl valeric acid) to correct for variability in the instrument response. Calibration curves were constructed by plotting the internal standard normalised area of authentic SCFA standards against the corresponding known SCFA concentrations and used to calculate the measured concentrations of SCFAs in the analysed samples.

Metabonomic Statistical and Bioinformatics Analysis

The resulting ^{1}H-NMR and LC-MS data sets were imported into SIMCA 14.1 (Umetrics) to conduct multivariate statistical analysis. Principal Component Analysis (PCA), followed by Orthogonal Partial Least Squares Discriminant Analysis (OPLS-DA) was performed to examine the data sets and to observe clustering in the results according to the predefined classes. The OPLS-DA models in the current study were established based on one PLS component and one orthogonal component. Unit variance scaling was applied to ^{1}H-NMR data, Pareto scaling was applied to MS data. The fit and predictability of the models obtained was determined by the R^2Y and Q^2Y values, respectively.

Significant metabolites were obtained from LC-MS OPLS-DA models through division of the regression coefficients by the jack-knife interval standard error to give an estimate of the t-statistic. Variables with a t-statistic ≥ 1.96 (z-score, corresponding to the 97.5 percentile) were considered significant. Significant metabolites were obtained from ^{1}H-NMR OPLS-DA models after investigating correlations with correlation coefficients values higher than 0.4.

Univariate statistical analysis was used to examine the SCFA data set. The data was not normally distributed; hence the Mann-Whitney U test was performed to examine differences between classes. P-values were adjusted for multiple testing using the Benjamini-Hochberg False Discovery Rate (pFDR) method.

Metabolite ID

Confirmation of metabolite identities in the NMR data was obtained using 1D ^{1}H NMR sequence with water pre-saturation and 2D NMR experiments such as J-Resolved spectroscopy, ^{1}H-^{1}H TOtal Correlation SpectroscopY (TOCSY), ^{1}H-^{1}H COrrelation SpectroscopY (COSY), ^{1}H-^{13}C Hetero-nuclear Single Quantum Coherence (HSQC) and ^{1}H-^{13}C Hetero-nuclear Multiple-Bond Correlation (HMBC) spectroscopy. In addition, statistical tools such as SubseT Optimization by Reference Matching (STORM) and Statistical TOtal Correlation SpectroscopY (STOCSY) were also applied.[18, 19]

Confirmation of metabolites identities in the LC-MS data was obtained using Tandem MS (MS/MS) on selected target ions with an energy ramp 5-20eV to produce product ions.

Metabolite identification was characterized by a level of assignment (LoA) score that describes how the identification was made.[20] The levels used were as follows: LoA 1: Identified compound, confirmed by comparison to an authentic chemical reference. LoA 2: MS/MS precursor and product ions or 1D+2D NMR chemical shifts and multiplicity match to a reference database or literature to putatively annotate compound. LoA 3: Chemical shift (δ) and multiplicity matches a reference database to tentatively assign the compound.

REFERENCES

1 Clarke SF, Murphy EF, O'Sullivan O, et al. Exercise and associated dietary extremes impact on gut microbial diversity. *Gut* 2014;63(12):1913-20 doi:10.1136/gutjnl-2013-306541

2 Mirsepasi H, Persson S, Struve C, Andersen LO, Petersen AM, Krogfelt KA. Microbial diversity in fecal samples depends on DNA extraction method: easyMag DNA extraction compared to QIAamp DNA stool mini kit extraction. *BMC research notes* 2014;7:50 doi:10.1186/1756-0500-7-50

3 Abubucker S, Segata N, Goll J, et al. Metabolic reconstruction for metagenomic data and its application to the human microbiome. *PLoS computational biology* 2012;8(6):e1002358 doi:10.1371/journal.pcbi.1002358

4 Caspi R, Altman T, Billington R, et al. The MetaCyc database of metabolic pathways and enzymes and the BioCyc collection of Pathway/Genome Databases. *Nucleic acids research* 2014;42(Database issue):D459-71 doi:10.1093/nar/gkt1103

5 R Development Core Team. R: A Language and Environment for Statistical Computing (R Foundation for Statistical Computing, Vienna, 2012). *URL: http:// www. R-project. org* 2015

6 Oksanen J, Blanchet FG, Kindt R, et al. Multivariate analysis of ecological communities in R: vegan tutorial. *R package version 2.3-1.* 2013

7 Subirana I, Sanz H, Vila J. Building bivariate tables: The compareGroups package for R. *Journal of Statistical Software* 2014;57(12):1-16

8 Cacciatore S, Luchinat C, Tenori L. Knowledge discovery by accuracy maximization. *Proceedings of the National Academy of Sciences of the United States of America* 2014;111(14):5117-22 doi:10.1073/pnas.1220873111

9 Wei T. corrplot: visualization of a correlation matrix. R package version 0.60, 2013.

10 Wickham H. *ggplot2: elegant graphics for data analysis*: Springer Science & Business Media, 2009.

11 Benjamini Y, Hochberg Y. Controlling the False Discovery Rate - a Practical and Powerful Approach to Multiple Testing. *J Roy Stat Soc B Met* 1995;57(1):289-300

12 Dona AC, Jimenez B, Schafer H, et al. Precision high-throughput proton NMR spectroscopy of human urine, serum, and plasma for large-scale metabolic phenotyping. *Anal Chem* 2014;86(19):9887-94 doi:10.1021/ac5025039

13 Garcia-Villalba R, Gimenez-Bastida JA, Garcia-Conesa MT, Tomas-Barberan FA, Carlos Espin J, Larrosa M. Alternative method for gas chromatography-mass spectrometry analysis of short-chain fatty acids in faecal samples. *J Sep Sci* 2012;35(15):1906-13 doi:10.1002/jssc.201101121

14 Want EJ, Wilson ID, Gika H, et al. Global metabolic profiling procedures for urine using UPLC-MS. *Nat Protoc* 2010;5(6):1005-18 doi:10.1038/nprot.2010.50

15 Sarafian MH, Lewis MR, Pechlivanis A, et al. Bile acid profiling and quantification in biofluids using ultra-performance liquid chromatography tandem mass spectrometry. *Anal Chem* 2015;87(19):9662-70 doi:10.1021/acs.analchem.5b01556

16 Tautenhahn R, Bottcher C, Neumann S. Highly sensitive feature detection for high resolution LC/MS. *BMC Bioinformatics* 2008;9:504 doi:10.1186/1471-2105-9-504

17 Veselkov KA, Vingara LK, Masson P, et al. Optimized preprocessing of ultra-performance liquid chromatography/mass spectrometry urinary metabolic profiles for improved information recovery. *Anal Chem* 2011;83(15):5864-72 doi:10.1021/ac201065j

18 Posma JM, Garcia-Perez I, De Iorio M, et al. Subset optimization by reference matching (STORM): an optimized statistical approach for recovery of metabolic biomarker structural information from 1H NMR spectra of biofluids. *Analytical chemistry* 2012;84(24):10694-701 doi:10.1021/ac302360v

19 Cloarec O, Dumas ME, Craig A, et al. Statistical total correlation spectroscopy: an exploratory approach for latent biomarker identification from metabolic 1H NMR data sets. *Analytical chemistry* 2005;77(5):1282-9 doi:10.1021/ac048630x

20 Sumner LW, Amberg A, Barrett D, et al. Proposed minimum reporting standards for chemical analysis Chemical Analysis Working Group (CAWG) Metabolomics Standards Initiative (MSI). *Metabolomics* 2007;3(3):211-21 doi:10.1007/s11306-007-0082-2