Supplementary Information

MATERIALS AND METHODS

Subjects

Subjects were adolescent/young adult Hodgkin lymphoma (AYAHL)-discordant twin pairs participating in the International Twin Study (Mack *et al*, 1995). Living pairs, resident in the United States or Canada, were eligible if one member had been diagnosed with AYAHL (histologically validated by BNN) before age 50 years, if both members of the pair were alive according to a death index linkage, and if the pair was discordant for reported early childhood fecal-oral exposures as assessed by a questionnaire completed and returned by both members of the pair (n=33 pairs) (Cozen *et al*, 2009). Pairs were excluded if either or both members reported inflammatory bowel disease, ulcers, gastroesophageal reflux or diarrheal diseases, or recent febrile illness of any kind were excluded, in addition to twins who recently changed their diet, or used antibiotics, glucocorticoids, proton pump inhibitors, or probiotics within the last 6 months. Participating twins ranged in age from 35 to 63 years at the time of specimen collection. Body mass index was calculated as self-reported weight in kilograms divided by the square of self-reported height in meters. Hodgkin lymphoma histological subtypes included seven nodular sclerosis, two mixed cellularity, and four not otherwise specified. Detailed treatment data were not available for all case-twins, but the majority received therapeutic radiation plus combination chemotherapy with either mechlorethamine, vincristine, procarbazine and prednisone (MOPP) or adriamycin, bleomycin, vinblastine and dacarbazine (ABVD).

Specimen collection and handling

Participants were provided with a kit that fits over the toilet seat for fecal specimen collection. Specimens were immediately frozen at -20^{0}C, shipped overnight to University of Southern California, and frozen at -80^{0}C after replacement of identifying information with a code number.

DNA preparation and sequencing

Ten to twenty grams of each frozen fecal specimen were pulverized in liquid nitrogen with a mortar and pestle. A 500mg aliquot was suspended in a solution containing 500µl of extraction buffer (200mM Tris (pH8.0), 200mM NaCl, 20 mM EDTA), 210µl of 20% SDS, 500µl of a mixture of phenol:chloroform:isoamyl alcohol (25:24:1, pH7.9), and 500µl of 0.1mm diameter zirconia/silica beads (BioSpec Products, Bartlesville OK). Microbial cells were lysed by mechanical disruption with a bead beater (BioSpec Products) set on high for 2 min at room temperature, followed by extraction with phenol:chloroform:isoamyl alchohol, and precipitation with isopropanol. DNA obtained from three separate aliquots of each fecal specimen were pooled, and the mixture used as a template for PCR of variable region 2 (V2) of bacterial 16S rRNA genes using the primers FWD:AGAGTTTGATCCTGGCTCAG and REV:TGCTGCCTCCCGTAGGAGT (Turnbaugh *et al*, 2009). The resulting amplicons were purified, pooled and subjected to multiplex pyrosequencing using the 454 Life Sciences® FLX platform with standard chemistry.

Data quality assessment and editing

The QIIME pipeline version 1.0 (Caporaso *et al*, 2010b; Edgar *et al*, 2011) was used to remove pyrosequencing reads with low-quality scores (<25), short reads (<200 nt) and reads judged to have sequencing artifacts, using a de-noising algorithm (Turnbaugh *et al*, 2010). Chimeric 16S rRNA sequences, which are PCR artifacts composed of two or more phylogenetically distinct parental sequences, were removed using UCHIME (Wang & Wang, 1996). The analyses reported below were performed on sequences that were denoised and cleared of chimeras.

Sequence analyses and measures of diversity

Species-level operational taxonomical units (OTUs) were obtained de novo by a clustering algorithm (Uclust) that grouped sequences with at least 97% nucleotide identity (97% ID)

(Edgar, 2010). Alpha diversity measures the diversity of OTUs within a sample. The total number of unique OTUs is a measure of alpha diversity that does not consider the relative abundance of OTUs. Most other measures of alpha diversity do take relative abundance of OTUs into account. Because alpha diversity estimates may be particularly sensitive to DNA sequencing errors, we took an additional "conservative" approach: (i) filtered to remove OTUs that had a relative abundance less than 0.1% in each sample; (ii) if a filtered OTU was present in another sample at higher abundance, it was retained; (iii) removed all OTUs that were present in less than 2 samples across the entire dataset; (iv) rarefied the resulting OTU table to 5258 reads per sample.

In addition to defining the number of unique OTUs in each specimen, we used three other tools for estimating alpha diversity. Chao1 is a presence/absence indicator that is bias-corrected for rare taxa (Chao, 1987). Shannon index, which adjusts for relative abundance of an OTU, is defined as (negative) the sum over OTUs of the product of the relative abundance of the OTU times the natural logarithm of the relative abundance (Shannon, 1948). Phylogenetic distance (PD)_whole tree reflects phylogenetic divergence among OTUs within an individual (Faith & Baker, 2006). To estimate PD_whole tree, the phylogenetic tree was estimated by the fasttree method (Price *et al*, 2010). To construct the tree, OTU sequences were aligned with the PyNAST algorithm, and gaps and hypervariable regions were masked using http:/greengenes.lbl.gov/Download/Sequence_Data/lanemask_in_1s_and_0s.txt (Caporaso *et al*, 2010a; Lane, 1991).

Beta diversity measures similarities between microbial communities of two individuals. We estimated beta diversity with UniFrac, which measures the phylogenetic similarity of any two communities based on the degree to which they share branch length on a bacterial tree of life.(Lozupone & Knight, 2005) The phylogenetic tree used for these measurements is the same

as the tree used for PD_whole tree. Weighted UniFrac value considers the abundance of taxa, whereas unweighted UniFrac value does not (Lozupone & Knight, 2005).

For taxonomic analysis, OTUs were assigned to phylum, class, order, family, and genus levels with the untrained Ribosomal Data Project classifier (Wang *et al*, 2007). Relative abundance was estimated as the proportion (%) of OTUs assigned to a taxon.

Statistical analyses

The number of sequences obtained after denoising ranged from 5520 to 23,755 per specimen in the initial analysis (5258 to 22,199 in the conservative OTU-restricted case-control analysis). To compare microbiota diversity within individuals at the same sequence depth, the data were rarified by randomly sampling 1000 times, without replacement, 5520 sequences from each participant. In the conservative analysis, we likewise rarified by performing random sampling 20 times for 5258 sequences per participant. Each participant's alpha diversity measures and other statistics were based on the mean values from 1000 such random samplings of 5520 sequences in the initial analysis and 20 random samplings of 5258 sequences in the conservative analysis. For taxonomic and beta diversity comparisons, we used the mean of 20 such random samples.

Supplementary Table 1. Numbers of filtered, denoised 16S rRNA sequence reads amplified from fecal DNA of 26 individual participants, before and after restriction to operational taxonomic units (OTUs) with relative abundance >0.001 (0.1%).

Sample ID	Number of reads before OTU restriction	Number of reads after OTU restriction
1	5520	5258
2	5665	5388
3	5665	5447
4	5898	5604
5	6365	6221
6	6365	6074
7	6464	6222
8	6817	6370
9	7507	7096
10	7918	7705
11	7935	7492
12	8453	7814
13	9139	8782
14	9222	8714
15	10145	8796
16	10410	9930
17	10644	9684
18	10843	10318
19	11029	10481
20	11345	10806
21	11852	9471
22	11874	11409
23	12005	11368
24	13669	12505
25	16439	15816
26	23755	22199
Total	252,943	236,970

SupplementaryTable 2. Hodgkin lymphoma case-control comparisons of relative abundance of bacterial taxa.

		mean abundance	
Phylum; Class; Order; Family; Genus	***P****	**cases**	**Controls**
Actino;Actino;Coriobacteriales;Coriobacteriaceae;Collinsella	0.03*	0.002	0.004
Bacteroidetes;Bacteroidia;Bacteroidales;Bacteroidaceae;Bacteroides	0.98	0.229	0.230
Bacteroidetes;Bacteroidia;Bacteroidales;Porphyromonadaceae;Barnesiella	0.27	0.005	0.002
Bacteroidetes;Bacteroidia;Bacteroidales;Porphyromonadaceae;Odoribacter	0.88	0.003	0.003
Bacteroidetes;Bacteroidia;Bacteroidales;Porphyromonadaceae;Other	0.37	0.001	0.003
Bacteroidetes;Bacteroidia;Bacteroidales;Porphyromonadaceae;Parabacteroides	0.73	0.018	0.015
Bacteroidetes;Bacteroidia;Bacteroidales;Prevotellaceae;Other	0.65	0.014	0.008
Bacteroidetes;Bacteroidia;Bacteroidales;Prevotellaceae;Paraprevotella	0.77	0.003	0.003
Bacteroidetes;Bacteroidia;Bacteroidales;Prevotellaceae;Prevotella	0.53	0.044	0.014
Bacteroidetes;Bacteroidia;Bacteroidales;Rikenellaceae;Alistipes	0.43	0.050	0.037
Bacteroidetes;Other;Other;Other;Other	0.06	0.001	0.007
Firmicutes;Bacilli;Lactobacillales;Streptococcaceae;Streptococcus	0.33	0.002	0.009
Firmicutes;Clostridia;Clostridiales;Clostridiaceae;Clostridium	0.51	0.002	0.003
Firmicutes;Clostridia;Clostridiales;Eubacteriaceae;Eubacterium	0.20	0.002	0.008
Firmicutes;Clostridia;Clostridiales;Incertae Sedis XIV;Blautia	0.56	0.047	0.053
Firmicutes;Clostridia;Clostridiales;Lachnospiraceae;Coprococcus	0.99	0.006	0.006
Firmicutes;Clostridia;Clostridiales;Lachnospiraceae;Dorea	0.29	0.008	0.011
Firmicutes;Clostridia;Clostridiales;Lachnospiraceae;Other	0.18	0.066	0.077
Firmicutes;Clostridia;Clostridiales;Lachnospiraceae;Roseburia	0.47	0.021	0.026
Firmicutes;Clostridia;Clostridiales;Other;Other	0.27	0.122	0.132
Firmicutes;Clostridia;Clostridiales;Ruminococcaceae;Butyricicoccus	0.87	0.002	0.002
Firmicutes;Clostridia;Clostridiales;Ruminococcaceae;Faecalibacterium	1.00	0.106	0.106
Firmicutes;Clostridia;Clostridiales;Ruminococcaceae;Oscillibacter	0.54	0.022	0.027
Firmicutes;Clostridia;Clostridiales;Ruminococcaceae;Other	0.27	0.082	0.066
Firmicutes;Clostridia;Clostridiales;Ruminococcaceae;Ruminococcus	0.83	0.018	0.020
Firmicutes;Clostridia;Clostridiales;Ruminococcaceae;Subdoligranulum	0.35	0.046	0.032
Firmicutes;Clostridia;Clostridiales;Veillonellaceae;Dialister	0.44	0.005	0.003
Firmicutes;Clostridia;Clostridiales;Veillonellaceae;Megasphaera	0.34	0.000	0.004
Firmicutes;Clostridia;Clostridiales;Veillonellaceae;Phascolarctobacterium	0.82	0.009	0.009
Firmicutes;Clostridia;Other;Other;Other	0.45	0.005	0.004
Firmicutes;Erysipelotrichi;Erysipelotrichales;Erysipelotrichaceae;Catenibacterium	0.33	0.000	0.003
Firmicutes;Other;Other;Other;Other	0.85	0.014	0.014
Other;Other;Other;Other;Other	0.24	0.017	0.026
ProteoBetaproteoBurkholderiales;Alcaligenaceae;Parasutterella	0.25	0.007	0.004
ProteoBetaproteoBurkholderiales;Alcaligenaceae;Sutterella	0.99	0.002	0.002
ProteoOther;Other;Other;Other	0.14	0.001	0.006
Tenericutes;Mollicutes;Anaeroplasmatales;Anaeroplasmataceae;Anaeroplasma	0.34	0.000	0.005

* Paired t-test *P*-value unadjusted for multiple comparisons. Collinsella association by Wilcoxon signed rank test *P*=0.05. Genera with mean relative abundance in the population <0.001 were not compared.

References

Caporaso JG, Bittinger K, Bushman FD, DeSantis TZ, Andersen GL, Knight R (2010a) PyNAST: a flexible tool for aligning sequences to a template alignment. *Bioinformatics* **26**(2)**:** 266-7

Caporaso JG, Kuczynski J, Stombaugh J, Bittinger K, Bushman FD, Costello EK, Fierer N, Pena AG, Goodrich JK, Gordon JI, Huttley GA, Kelley ST, Knights D, Koenig JE, Ley RE, Lozupone CA, McDonald D, Muegge BD, Pirrung M, Reeder J, Sevinsky JR, Turnbaugh PJ, Walters WA, Widmann J, Yatsunenko T, Zaneveld J, Knight R (2010b) QIIME allows analysis of high-throughput community sequencing data. *Nature methods* **7**(5)**:** 335-6

Chao A (1987) Estimating the population size for capture-recapture data with unequal catchability. *Biometrics* **43**(4)**:** 783-91

Cozen W, Hamilton AS, Zhao P, Salam MT, Deapen DM, Nathwani BN, Weiss LM, Mack TM (2009) A protective role for early oral exposures in the etiology of young adult Hodgkin lymphoma. *Blood* **114**(19)**:** 4014-20

Edgar RC (2010) Search and clustering orders of magnitude faster than BLAST. *Bioinformatics* **26**(19)**:** 2460-1

Edgar RC, Haas BJ, Clemente JC, Quince C, Knight R (2011) UCHIME improves sensitivity and speed of chimera detection. *Bioinformatics* **27**(16)**:** 2194-200

Faith DP, Baker AM (2006) Phylogenetic diversity (PD) and biodiversity conservation: some bioinformatics challenges. *Evolutionary bioinformatics online* **2:** 121-8

Lane DJ (1991) *Nucleic Acid Techniques in Bacterial Systematics*. New York: Wiley

Lozupone C, Knight R (2005) UniFrac: a new phylogenetic method for comparing microbial communities. *Applied and environmental microbiology* **71**(12)**:** 8228-35

Mack TM, Cozen W, Shibata DK, Weiss LM, Nathwani BN, Hernandez AM, Taylor CR, Hamilton AS, Deapen DM, Rappaport EB (1995) Concordance for Hodgkin's disease in identical twins suggesting genetic susceptibility to the young-adult form of the disease. *The New England journal of medicine* **332**(7)**:** 413-8

Price MN, Dehal PS, Arkin AP (2010) FastTree 2--approximately maximum-likelihood trees for large alignments. *PloS one* **5**(3)**:** e9490

Shannon CE (1948) A mathematical theory of communication. *The Bell System Techincal Journal*(27)**:** 379-423 and 623-656

Turnbaugh PJ, Hamady M, Yatsunenko T, Cantarel BL, Duncan A, Ley RE, Sogin ML, Jones WJ, Roe BA, Affourtit JP, Egholm M, Henrissat B, Heath AC, Knight R, Gordon JI (2009) A core gut microbiome in obese and lean twins. *Nature* **457**(7228)**:** 480-4

Turnbaugh PJ, Quince C, Faith JJ, McHardy AC, Yatsunenko T, Niazi F, Affourtit J, Egholm M, Henrissat B, Knight R, Gordon JI (2010) Organismal, genetic, and transcriptional variation in the deeply sequenced gut microbiomes of identical twins. *Proceedings of the National Academy of Sciences of the United States of America* **107**(16)**:** 7503-8

Wang GC, Wang Y (1996) The frequency of chimeric molecules as a consequence of PCR co-amplification of 16S rRNA genes from different bacterial species. *Microbiology* **142 (Pt 5):** 1107-14

Wang Q, Garrity GM, Tiedje JM, Cole JR (2007) Naive Bayesian classifier for rapid assignment of rRNA sequences into the new bacterial taxonomy. *Applied and environmental microbiology* **73**(16)**:** 5261-7