Supplementary Discussion of data presented in Figure 1

Each pyrosequencer run of cecal DNA from lean1 and ob1 littermates yielded 387±95 new groups (s.e.m.) (**Fig. 1A**). The observed relative abundance of COG categories was markedly influenced by sequencing platform (**Fig. 1B**).

As expected, the Firmicutes-enriched, Bacteroidetes-depleted *ob/ob* microbiome is depleted for genes involved in the biosynthesis of lipopolysaccharide (a major component of the outer membrane of Gram-negative bacteria), and enriched for genes involved in cell motility and sporulation (many Firmicutes are motile and form endospores) (**Fig. 1C and 1D; Supplementary Fig. 6**).

EGT assignment notes (**Fig. 1D**): (i) 'Type III secretion systems' are represented by EGTs involved in flagellar assembly (very few EGTs were assigned specifically to the secretion apparatus); (ii) 'Galactose metabolism' includes glycoside hydrolases [alpha-glucosidase (KO1187), beta-galactosidase (KO1190), and alpha-galactosidase (KO7406/7)], and 6-phosphofructokinase (KO0850, catalyzes the rate limiting step in glycolysis); (iii) 'Glycerolipid metabolism' includes glycoside hydrolases [beta-galactosidase (KO1190) and alpha-galactosidase (KO7406/7)] plus glycerol kinase (KO0864, involved in degradation of triglycerides and phospholipids); (iv) 'Glycosphingolipid metabolism' also includes glycoside hydrolases [beta-galactosidase (KO1190) and alpha-galactosidase (KO7406/7)]; (v) 'Reductive carboxylate cycle' and 'Pyruvate/oxoglutarate oxidoreductases' both include genes involved in the citrate cycle [2-oxoglutarate ferrodoxin oxidoreductase (KO0174/5) and succinate dehydrogenase

(KO0238/39/40)], (vi) 'C5-branched dibasic acid metabolism' includes valine and isoleucine biosynthesis from pyruvate [i.e. acetolactate synthase (KO1651/2)[26,27]]. Metabolic capacity is defined based on microbial community gene content. Transcriptomic, proteomic and/or metabolomic data are necessary to confirm predicted activites of genes and their products (e.g. see **Fig. 3**).

Materials and Methods

Animals – All experiments involving mice were performed using protocols approved by the Washington University Animal Studies Committee. Once C57BL/6J *ob/ob*, *ob*/+, and +/+ littermates were weaned, they were housed individually in microisolator cages where they were maintained in a specified pathogen-free state, under a 12-h light cycle, and fed a standard polysaccharide-rich chow diet (PicoLab, Purina) *ad libitum*. Germ-free and colonized animals were maintained in gnotobiotic isolators[28], under a strict 12-h light cycle and fed an autoclaved chow diet (B&K Universal, East Yorkshire, U.K.) *ad libitum*. Fecal samples for bomb calorimetry were collected from mice at 8 or 14 weeks of age, after which time animals were sacrificed.

Community DNA Preparation – The cecal contents used for community DNA sequencing and gas chromatography-mass spectrometry (GC-MS) were obtained, at eight weeks of age, from the same animals used for our previous PCR-based 16S rRNA survey of the gut microbiota[6]: samples had been stored at -80^{o}C (**Supplementary Table 1)**. An aliquot (~10mg) of each sample was suspended while frozen in a solution containing 500 μL of extraction buffer [200 mM Tris (pH 8.0), 200 mM NaCl, 20 mM EDTA], 210 μL of 20% SDS, 500 μL of a mixture of of phenol:chloroform:isoamyl alcohol (25:24:1)], and 500 μL of a slurry of 0.1-mm-diameter zirconia/silica beads (BioSpec Products, Bartlesville, OK). Microbial cells were then lysed by mechanical disruption with a bead beater (BioSpec Products) set on high for 2 min (23^{o}C), followed by extraction with phenol:chloroform:isoamyl alcohol, and precipitation with isopropanol. In order to perform pyrosequencing, DNA was purified further using the Qiaquick gel extraction kit (Qiagen).

Shotgun sequencing and assembly of cecal microbiomes – DNA samples were used to construct plasmid libraries for 3730xl capillary-based sequencing. Pyrosequencing was performed as previously described[11]. Briefly, samples were nebulized to 200 nucleotide fragments, ligated to adaptors, fixed to beads, suspended in a PCR reaction mixture-in-oil emulsion, amplified, and sequenced using a GS20 pyrosequencer (454 Life Sciences, Branford, CT). The Newbler *de novo* shotgun sequence assembler (454 Life Sciences) was used to assemble sequences based on flowgram signal space. This process includes overlap generation, contig layout, and consensus generation. The resulting GS20 contigs were then broken into linked sequences to generate pseudo paired-end reads, and aligned with 3730xl reads using PCAP[29].

Sequences were aligned to reference genomes using the PROmer script in MUMmer[12] (version 3.18). Capillary sequencer reads from each microbiome, the finished genome of the human gut-derived *Bacteroides thetaiotaomicron* type strain ATCC29148[1], and a deep draft genome of the human gut-derived *Eubacterium rectale* type strain ATCC33656 (http://gordonlab.wustl.edu/supplemental/Turnbaugh/obob/) were used as a reference for the pyrosequencer datasets. Coverage was calculated by dividing the sum of all alignment lengths by the length of the reference genome.

Whole genome sequencing and annotation – A draft assembly of *Eubacterium rectale* ATCC33656 was generated from AB36731xl paired end-reads of inserts in whole genome shotgun plasmid and fosmid libraries, as well as from reads produced by the GS20 pyrosequencer. Sequences were assembled using Newbler and PCAP (see above) and ORFs predicted with Glimmer3.01[30] (maximum overlap of 100, minimum length of

110 and a threshold of 30). Each predicted gene sequence was translated, and the resulting protein sequence assigned to InterPro numbers using InterProScan[31] (Release 12.0).

Database search parameters – NCBI BLAST was used to query the non-redundant database (NR), the STRING-extended COG database (179 microbial genomes, version 6.3)[13], a database constructed from 334 genomes available through KEGG (version 37)[14], and the Ribosomal Database Project database (RDP, version 9.33)[32]. Reads with multiple COG/KO hits were counted once for each classification scheme. KO hits were also categorized into CAZy families (http://afmb.cnrs-mrs.fr/CAZY/). KEGG pathway maps are available on-line (http://gordonlab.wustl.edu/supplemental/Turnbaugh/obob/).

NR, COG, and KEGG comparisons were performed using NCBI BLASTX. RDP comparisons were performed using NCBI BLASTN, and microbiomes were directly compared using TBLASTX. A cutoff of e-value < 10^{-5} was used for EGT assignments and sequence comparisons[16] (corresponds to a p-value cutoff of 10^{-12} against the NR and KEGG databases, and 10^{-11} against the COG database). Given this cutoff, we would only expect three false EGT assignments in our combined analyses due to random chance. We also re-analyzed the data using a more stringent cutoff[33] (e-value < 10^{-8}).

Taxonomic assignments of shotgun 16S rRNA gene fragments – Shotgun reads containing a 16S rRNA fragment were identified by BLASTX comparison of each microbiome to the RDP database. 16S rRNA gene fragments were then aligned using the NASTA multi-aligner[34] with a minimum template length of 20 bases and a minimum

percent identity of 75%. The resulting alignment was then imported into an ARB neighbor-joining tree and hypervariable regions were masked using the lanemaskPH filter[35]. Direct BLAST taxonomic assignments were performed through BLASTX comparisons of each microbiome and the NR database. Best-BLAST-hits with an e-value $< 10^{-5}$ were used to assign each read to a given species.

Estimating the total number of orthologous groups – The total estimated number of COGs and NOGs (Non-supervised Orthologous Groups) in each sample was calculated using the lower-limit of the Chao1 95% confidence interval in EstimateS (Version 7.5, R. K. Colwell, http://purl.oclc.org/estimates), based on the number of EGTs assigned to each orthologous group. The number of missed groups was calculated by subtracting the estimated total (Chao1 lower-limit) from the observed number of groups.

Direct comparisons of microbiome sequences – Microbiomes sequenced using the 3730xl instrument were evaluated by reciprocal pairwise TBLASTX comparisons[16]. 8,832 reads were used from each microbiome to limit artifacts that arise from different sized datasets. Each possible pairwise comparison was made by using a BLAST database constructed from each microbiome. Samples were clustered based on the cumulative pairwise BLAST score. An estimate of distance was constructed using the D2 normalization and genome conservation approach previously used for genome clustering[36]. This method calculates a distance score based on the minimum cumulative BLAST score (sum of all best-BLAST-hit scores) between two microbiomes and the weighted average of both self-self comparisons ($D2 = -\ln(\min S_{1v2}, S_{2v1}/\text{average})$. The weighted average is calculated using $\text{average} = \text{squareroot}(2) * S_{1v1} * S_{2v2} / \text{squareroot}(S_{1v1}^2 + S_{2v2}^2)$. The resulting distances were used to create a distance matrix. A tree was

constructed using NEIGHBOR (PHYLIP version 3.64; kindly provided by J. Felsenstein, Department of Genome Sciences, University of Washington, Seattle), and was viewed using Treeview X[37].

Clustering of microbiomes based on predicted metabolic function – Microbiomes were clustered based on the percent representation of EGTs assigned to each COG, KEGG pathway, and phylotype (genome in NR) using Cluster3.0[25]. Percent representation was calculated as the number of EGTs assigned to a given group divided by the number of EGTs assigned to all groups. Single linkage hiearchical clustering via Pearson's correlation was performed on each dataset, and the results were visualized by using the Treeview Java applet[38]. Principal Component Analysis was also performed based on the percent representation of EGTs assigned to KEGG pathways (Cluster3.0[25]), and the data were graphed according to the first two coordinates.

Identification of statistically enriched and depleted metabolic groups – Two methods were used to determine statistically enriched or depleted metabolic groups: the cumulative binomial distribution[3] and a bootstrap analysis[16,17]. The cumulative binomial distribution was used for pairwise comparisons of microbiome COG, KEGG, and taxonomic assignments. The calculation uses the following inputs: number of successes for microbiome 1 (number of EGTs assigned to a given group), number of trials for microbiome 1 (total number of EGTs assigned to all groups), and the expected frequency (number of successes/number of trials for microbiome 2). The probability of having less than or equal to the number of observed EGTs in a given group was then calculated using the cumulative binomial distribution. Depletion was defined as having a probability less than 0.05, 0.01, or 0.001 assuming p equals the expected frequency and that the expected

frequency is normally distributed. Enrichment was defined as having a probability of greater than 0.95, 0.99, or 0.999 given the same assumptions. To minimize false negatives, no corrections for multiple sampling were made. To limit false positives resulting from low sampling, only groups with at least one hit in each microbiome were evaluated.

Xipe[17] (Rodriguez-Brito, version 0.2) was employed for bootstrap analyses of KEGG pathway enrichment and depletion, using the following parameters: 10,000 samples, 10,000 repeats, and three confidence levels (95%, 99%, and 99.9%). Briefly, a dataset composed of the number of EGTs assigned to each KEGG pathway was sampled with replacement from each microbiome 10,000 times. The difference between the number of EGTs per pathway in the first microbiome, and the number of EGTs per pathway in the second microbiome, was calculated for each group. This process was repeated 10,000 times and the median difference calculated for each pathway. A confidence interval was determined by pooling both datasets and comparing 10,000 random samples to 10,000 other random samples. Groups with a larger median difference between microbiomes than the confidence interval were considered significantly different.

Biochemical analyses – Short-chain fatty acids (SCFAs) were measured in nine cecal samples (4 lean, 5 obese) obtained from nine mice that had been used for our previous 16S rRNA gene sequence-based survey [animals C1, C3, C4, C9, C10, C13, C15 (lean2), C17, and C22 in ref. 6]. Two aliquots of each sample were evaluated. SCFA levels were quantified according to previously published protocols[15]: i.e., double diethyl ether extraction of deproteinized cecal contents spiked with isotope-labeled internal

SCFA standards; derivatization of SCFAs with N-tert-butyldimethylsilyl-N-methyltrifluoracetamide (MTBSTFA); and GC-MS analysis of the resulting TBDMS-derivatives.

Bomb calorimetry was performed on 44 fecal samples collected from 22 mice (9 lean, 13 obese). Each mouse was transferred to a clean cage for 24 hours, at which point fecal samples were collected and oven dried at 60^{o}C for 48 hours. Gross energy content was measured using a semimicro oxygen bomb calorimeter, calorimetric thermometer, and semimicro oxygen bomb (Models 6725, 6772 and 1109, respectively, from Parr Instrument Co.). The calorimeter energy equivalent factor was determined using benzoic acid standards. The mean of each distribution was compared using a two-tailed Student's t-Test.

Microbiota transplantation experiments – Germ-free C57BL/6J mice (8-9 weeks old) were colonized with a cecal microbiota obtained from either a lean (+/+) or an obese (*ob/ob*) C57BL/6J donor (n=1 donor and 4-5 recipients/treatment group/experiment; 2 independent experiments). Recipient mice were anesthetized at 0 and 14 days post colonization with an i.p. injection of ketamine (10 mg/kg body weight) and xylazine (10mg/kg) and total body fat content was measured by dual-energy x-ray absorptiometry (Lunar PIXImus Mouse, GE Medical Systems) using previously described protocols[39]. Donor mice were sacrificed at day 0 and recipient mice after the final DEXA on day 14.

16S rRNA sequence-based surveys of the cecal microbiotas of conventionalized mice – Cecal contents were recovered at the time of sacrifice by manual extrusion and frozen immediately at -80°C. DNA was prepared by bead beating,

phenol/chloroform extraction, and gel purification (see above). Five replicate PCRs were performed for each mouse. Each 25 μl reaction contained 50-100 ng of purified DNA from cecal contents, 10 mM Tris (pH 8.3), 50 mM KCl, 2 mM $MgSO_4$, 0.16 μM dNTPs, 0.4 μM of the bacteria-specific primer 8F (5'-AGAGTTTGATCCTGGCTCAG-3'), 0.4 μM of the universal primer 1391R (5'-GACGGGCGGTGWGTRCA-3'), 0.4 M betaine, and 3 units of Taq polymerase (Invitrogen). Cycling conditions were 94°C for 2 min, followed by 35 cycles of 94°C for 1 min, 55°C for 45 sec, and 72°C for 2 min, with a final extension period of 20 min at 72°C. Replicate PCRs were pooled, concentrated with Millipore columns (Montage), gel-purified with the Qiaquick kit (Qiagen), cloned into TOPO TA pCR4.0 (Invitrogen), and transformed into *E. coli* TOP10 (Invitrogen). For each mouse, 384 colonies containing cloned amplicons were processed for sequencing. Plasmid inserts were sequenced bidirectionally using vector-specific primers and the internal primer 907R (5'-CCGTCAATTCCTTTRAGTTT-3').

16S rRNA gene sequences were edited and assembled into consensus sequences using the PHRED and PHRAP software packages within the Xplorseq program[40]. Sequences that did not assemble were discarded and bases with PHRED quality scores <20 were trimmed. Sequences were checked for chimeras using Bellerophon[41] and sequences with greater than 95% identity to both parents were removed (n=535; 13% of aligned sequences). The final dataset (n=4,157 sequences; for ARB alignment and tree see http://gordonlab.wustl.edu/supplemental/Turnbaugh/obob/; for sequence designations see **Supplementary Table 7**) was aligned using the on-line version of the NAST multi-aligner[34] (minimum alignment length=1250; percent identity >75), hypervariable regions were masked using the lanemaskPH filter provided with the ARB database[35], and the

aligned sequences were added to the ARB neighbor-joining tree (based on pairwise distances with the Olsen correction) with the parsimony insertion tool. A phylogenetic tree containing all 16S rRNA gene sequences was exported from ARB and clustered using online UniFrac[21] without abundance weighting.

Supplementary Notes

26. Dailey, F. E. & Cronan, J. E., Jr. Acetohydroxy acid synthase I, a required enzyme for isoleucine and valine biosynthesis in *Escherichia coli K-12* during growth on acetate as the sole carbon source. *J. Bacteriol* **165**, 453-460 (1986).

27. Dailey, F. E., Cronan, J. E., Jr. & Maloy, S. R. Acetohydroxy acid synthase I is required for isoleucine and valine biosynthesis by *Salmonella typhimurium LT2* during growth on acetate or long-chain fatty acids. *J. Bacteriol* **169**, 917-919 (1987).

28. Hooper, L. V. *et al.* (2002) in Molecular Cellular Microbiology, eds. Sansonetti, P. & Zychlinsky, A. (Academic San Diego), Vol. **31**, pp. 559-589.

29. Huang, X., Wang, J., Aluru, S., Yang, S. P. & Hillier, L. PCAP: a whole-genome assembly program. *Genome Res.* **13**, 2164-2170 (2003).

30. Delcher, A. L., Harmon, D., Kasif, S., White, O. & Salzberg, S. L. Improved microbial gene identification with GLIMMER. *Nucleic Acids Res.* **27**, 4636-4641 (1999).

31. Mulder, N. J. *et al.* InterPro, progress and status in 2005. *Nucleic Acids Res.* **33**, D201-205 (2005).

32. Cole, J. R. *et al.* The Ribosomal Database Project (RDP-II): sequences and tools for high-throughput rRNA analysis. *Nucleic Acids Res.* **33**, D294-296 (2005).

33 . Tringe, S. G. *et al.* Comparative metagenomics of microbial communities. *Science* **308**, 554-557 (2005).

34. DeSantis, T. Z. *et al.* NAST: a multiple sequence alignment server for comparative analysis of 16S rRNA genes. *Nucleic Acids Res.* **34**, W394-399 (2006).

35. Ludwig, W. *et al.* ARB: a software environment for sequence data. *Nucleic Acids Res.* **32**, 1363-1371 (2004).

36. Kunin, V., Ahren, D., Goldovsky, L., Janssen, P. & Ouzounis, C. A. Measuring genome conservation across taxa: divided strains and united kingdoms. *Nucleic Acids Res.* **33**, 616-621 (2005).

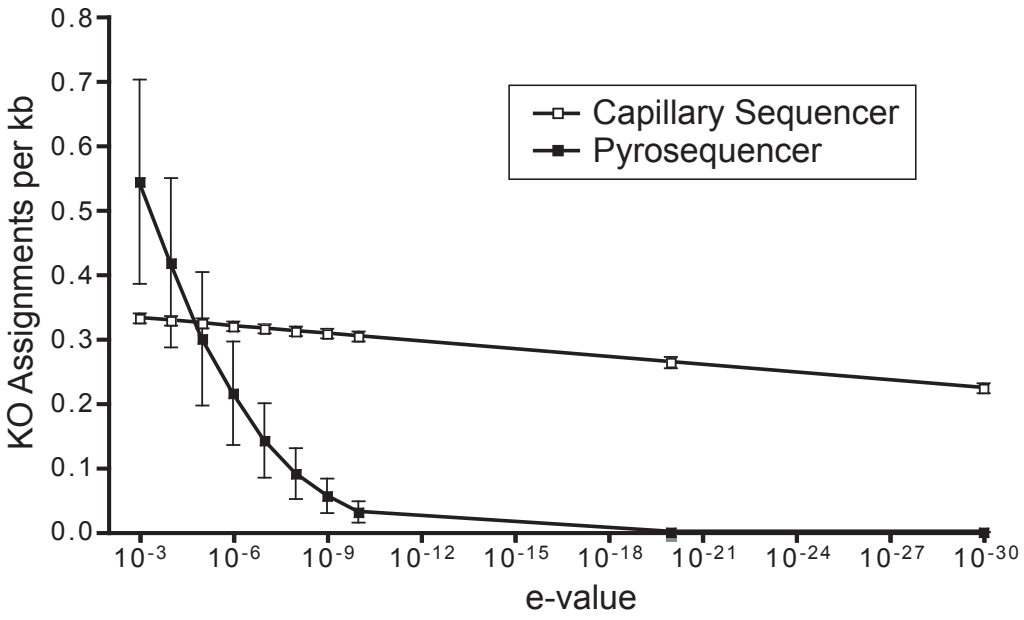
37. Page, R. D. TreeView: an application to display phylogenetic trees on personal computers. *Comput. Appl. Biosci.* **12**, 357-358 (1996).

38. Saldanha, A. J. Java Treeview--extensible visualization of microarray data. *Bioinformatics* **20**, 3246-3248 (2004).

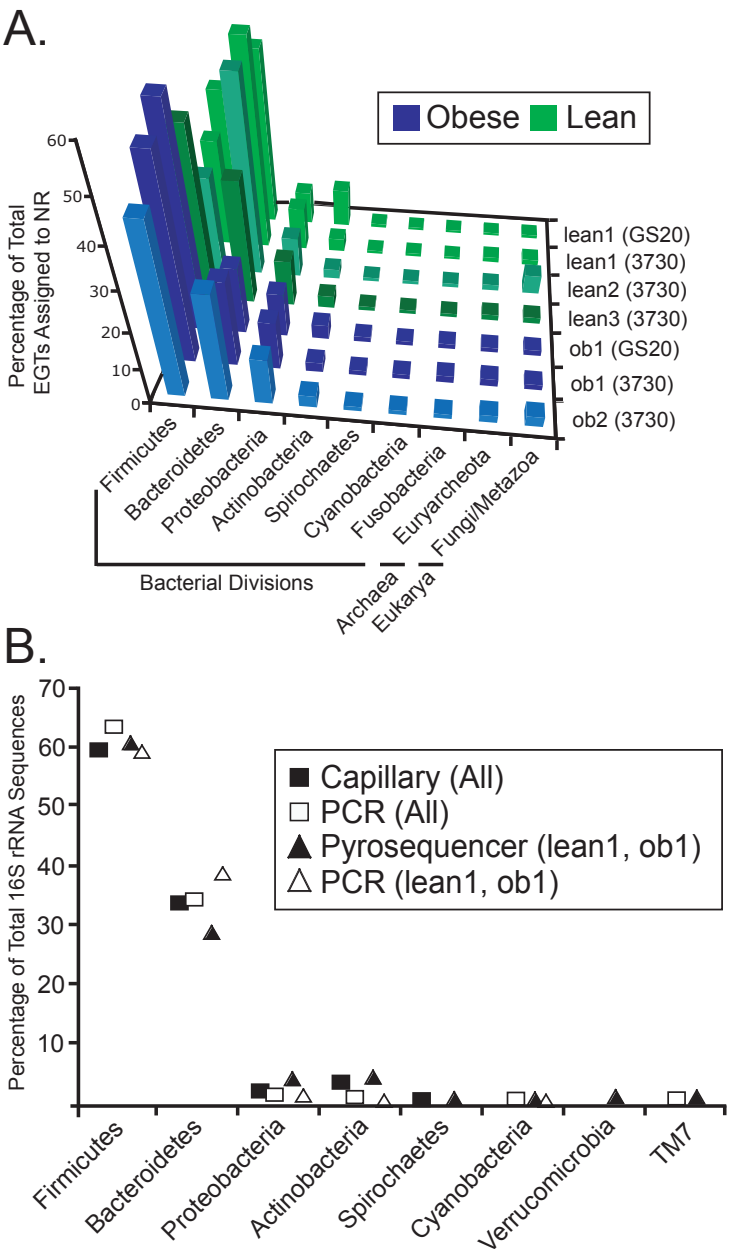
39. Bernal-Mizrachi, C. *et al.* Respiratory uncoupling lowers blood pressure through a leptin-dependent mechanism in genetically obese mice. *Arterioscler Thromb Vasc Biol.* **22**, 961-968 (2002).

40. Papineau, D., Walker, J. J., Mojzsis, S. J. & Pace, N. R. *Appl. Environ. Microbiol.* **71**, 4822-4832 (2006).

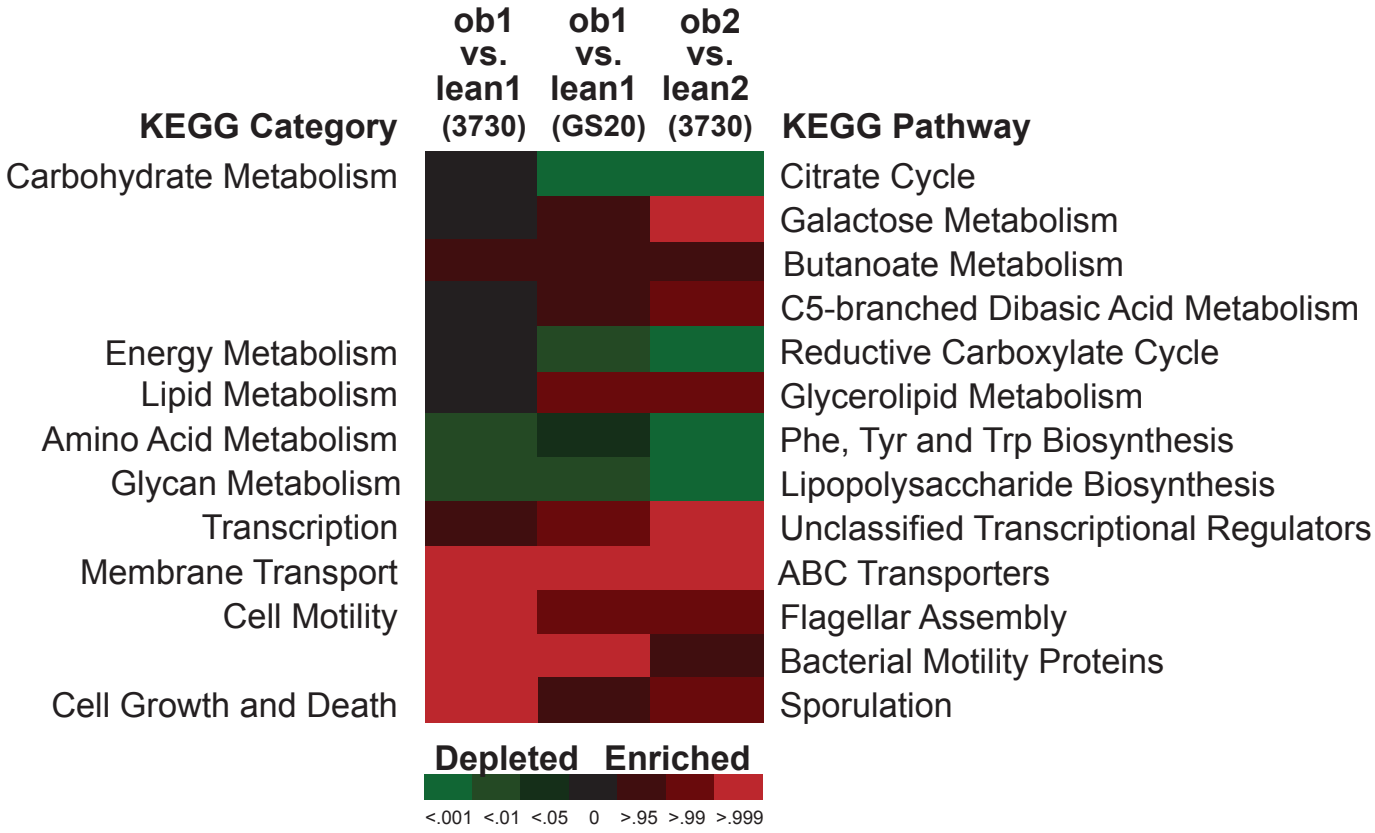
41. Huber, T., Faulkner, G. & Hugenholtz, P. Bellerophon: a program to detect chimeric sequences in multiple sequence alignments. *Bioinformatics* **20**, 2317-2319 (2004).



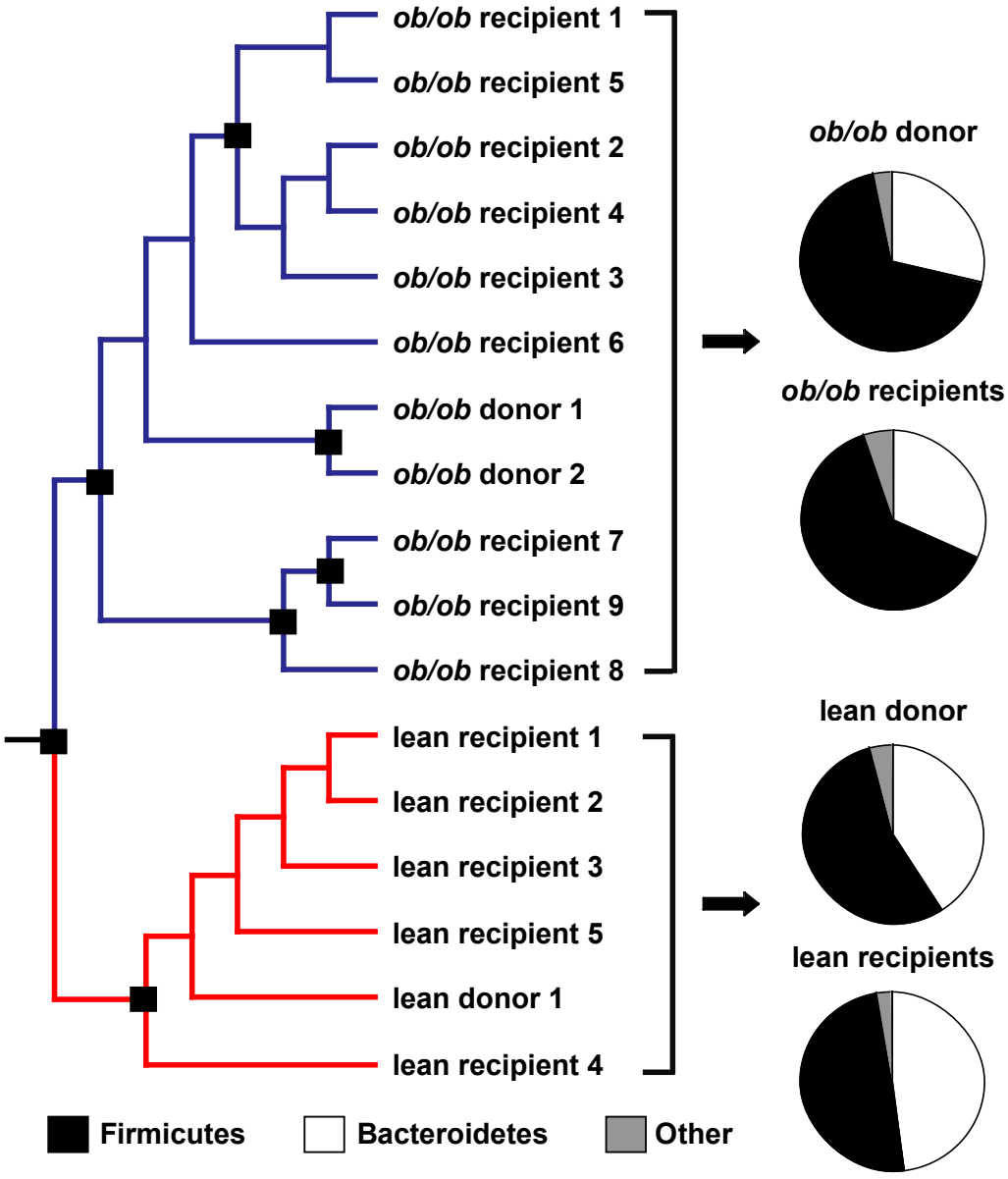
Supplementary Figure 1: The effect of decreasing e-value cut-offs on EGT assignments to the KEGG database from pyrosequencer and capillary sequencer datasets. Points indicate the average number of KO assignments per kb of microbiome sequence. Mean values ± s.e.m. are plotted. The GS20 pyrosequencer and the 3730xl capillary sequencer both resulted in an average 0.3 KO (KEGG orthology) assignments per kb of sequence at an e-value cutoff $<10^{-5}$. However, the number of EGTs present in the pyrosequencer-derived datasets rapidly decays as the e-value cutoff is decreased, whereas the number of EGTs present in the capillary sequencer datasets is relatively stable to $<10^{-30}$.



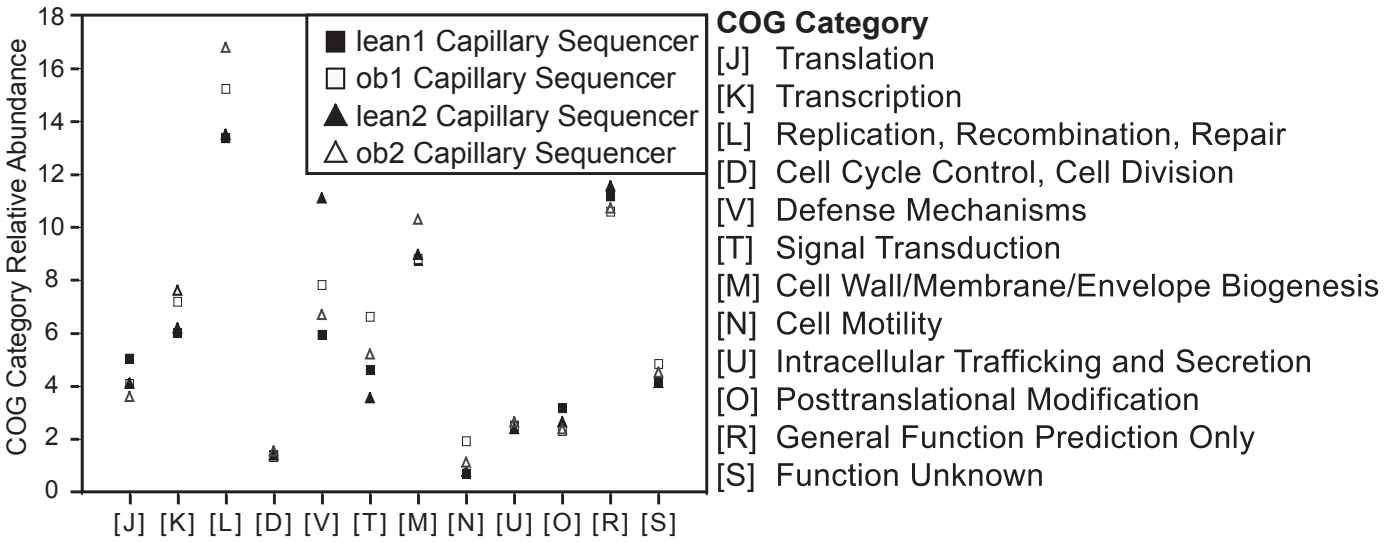
Supplementary Figure 2: Taxonomic assignments of EGTs and 16S rRNA gene fragments. (A) Relative abundance of EGTs (reads assigned to NR, BLASTX with an e-value<10^{-5}) in each cecal microbiome confirms the presence of the indicated bacterial divisions in addition to Euryarcheota. Metazoan sequences (including *Mus musculus* and fungi) are also present at low abundance. Bacterial divisions with greater than 1% representation in at least three microbiomes are shown. (B) Alignment of 16S rRNA gene fragments (black) confirms our previous PCR-derived 16S rRNA gene sequence-based survey[6] (white). Comparisons include all microbiomes sampled with the capillary sequencer (square) and the two microbiomes sampled with the pyro-sequencer (triangle).



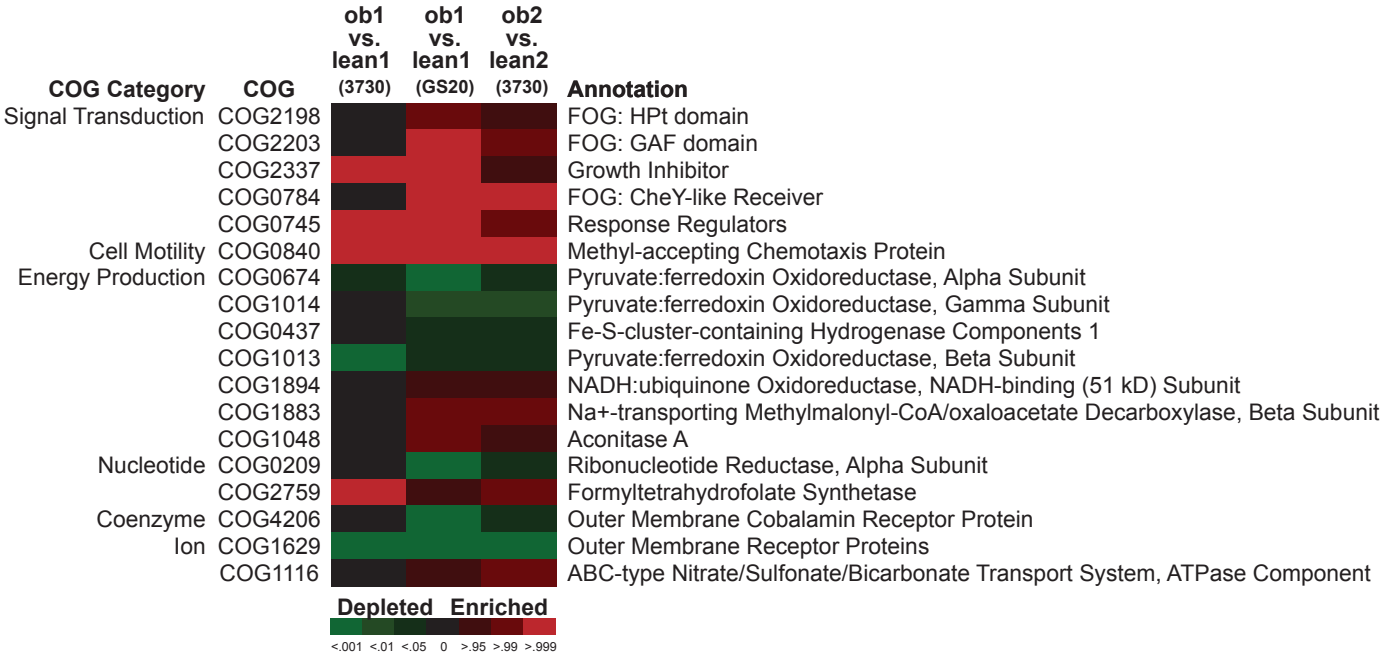
Supplementary Figure 3: KEGG pathways that are enriched or depleted in the cecal microbiomes of both obese versus lean sibling pairs, as indicated by bootstrap analysis of relative gene content. Pathways that are consistently enriched or depleted in the pyrosequencer-based comparison of ob1 versus lean1 littermates, and the capillary sequencer-based comparison of ob2 versus lean2 littermates are shown. Red indicates enrichment and green indicates depletion (brightness denotes level of significance). Black indicates groups that are not significantly changed.



Supplementary Figure 4: Analyses of microbial communities harvested from obese (*ob/ob*) and lean (+/+) C57BL/6J donor mice and colonized gnotobiotic recipients. Online Unifrac clustering[21] of microbial community structure, based on 4,157 16S rRNA gene sequences (see Supplementary Table 7 for number of sequences per sample; ARB tree available at http://gordonlab.wustl.edu/supplemental/Turnbaugh/obob/). Nodes denoted by a black square are robust to sequence number (jackknife values > 0.70, representing the number of times the node was present when 166 sequences were randomly chosen for each mouse for n=100 replicates). Pie charts indicate the average relative abundance of Firmicutes (black), Bacteroidetes (white), and other (grey; includes Verrucomicrobia, Proteobacteria, Actinobacteria, TM7, and Cyanobacteria) in the donor and recipient microbial communities.



Supplementary Figure 5: Relative abundance of COG categories (percentage of total EGTs assigned to COG using BLASTX and e-value<10^{-5}) in the lean1 (black square), ob1 (white square), lean2 (black triangle), and ob2 (white triangle) cecal microbiomes. Microbiomes were characterized by capillary sequencing.



Supplementary Figure 6: COGs that are enriched or depleted in the cecal microbiomes of both obese versus lean sibling pairs, as indicated by binomial comparisons of relative gene content. The COGs shown are enriched or depleted in the pyrosequencer-based comparison of ob1 versus lean1 littermates and the capillary sequencer-based comparison of ob2 versus lean2 littermates. Red indicates enrichment and green indicates depletion (brightness denotes level of significance). Black indicates groups that are not significantly changed.

Supplementary Table 1 – Nomenclature used to designate metagenomic datasets obtained from the cecal microbiota of C57BL/6J *ob/ob*, *ob*/+, and +/+ littermates.

Figure label	Metagenome label	Litter	16S rRNA survey label[1]	Tree label[1]	Host genotype
ob1	PT6	1	C23	M2B-4	*ob/ob*
ob2	PT4	2	C18	M1-2	*ob/ob*
lean1	PT3	1	C21	M2B-1	+/+
lean2	PT8	2	C15	M1-3	*ob*/+
lean3	PT2	2	C16	M1-4	+/+

[1]Samples obtained from our previous 16S rRNA survey (ref. 6).

Supplementary Table 2 – Sequencing results for each cecal microbiome.

Microbiome	Average read length	Number of reads	Sequence
lean1 (GS20)	90.9	1,046,611	94,913,476
ob1 (GS20)	96.4	677,384	65,370,448
lean1 (3730xl)	765	10,752	8,227,047
lean2 (3730xl)	782	11,136	8,705,876
lean3 (3730xl)	706	10,752	7,590,528
ob1 (3730xl)	735	11,136	8,185,880
ob2 (3730xl)	771	8,832	6,811,035
TOTAL	**-**	**1,776,603**	**199,804,290**

Abbrevations: GS20, pyrosequencer; 3730xl, capillary sequencer

Supplementary Table 3 – Assembly of reads from capillary sequencer and pyrosequencer datasets.

Sample	Contigs	Average contig length	Contiged bases[1]	Largest Assembly	N50 contig length (kb)[2]
lean1 (GS20)	102,299	117	11,966,580	2,793	0.109
ob1 (GS20)	56,425	116	6,518,469	2,174	0.109
lean1 (3730xl)	167	1527	254,985	5,500	1.62
lean2 (3730xl)	407	1598	650,499	5,522	1.71
lean3 (3730xl)	224	1528	342,172	3,281	1.59
ob1 (3730xl)	320	1393	445,814	3,225	1.49
ob2 (3730xl)	269	1644	442,210	4,186	1.70
All (3730xl)	2,575	1734	4,465,685	11,213	1.78
All (GS20)	159,245	118	18,809,438	2,708	0.110
All (GS20 and 3730xl)	13,667	898	12,275,469	14,755	0.903

[1]Contiged bases refers to the combined length of all contigs.

[2]N50 contig length refers to the length of the contig, such that 50% of the total contiged bases are present in contigs of greater or equal size.

Assembly of the GS20 pyrosequencer datasets from lean1 (+/+) or ob1 (*ob/ob*) produced very modest contiguity. Note that assembly of all GS20 pyrosequencer data from both lean1 and ob1 did not improve contiguity. However, including the five 3730xl datasets increased the average contig length to 1kb, and the largest contig to >14 kb.

Supplementary Table 4 – Number of EGTs assigned to the NR, COG, and/or KEGG databases.

Microbiome	Total NR EGTs	Total COG EGTs	Total KO EGTs	Total EGTs	Percent unassigned
lean1 (GS20)	48,625	51,481	28,359	56,599	94.6
ob1 (GS20)	33,360	32,819	18,308	39,058	94.2
lean1 (3730xl)	7,973	7,970	2,810	8,462	21.3
lean2 (3730xl)	7,309	7,687	2,723	8,170	26.6
lean3 (3730xl)	7,042	7,119	2,562	7,616	29.2
ob1 (3730xl)	7,331	7,299	2,639	7,859	29.4
ob2 (3730xl)	6,008	6,016	2,053	6,425	27.3

Supplementary Table 5 – Percentage of total assigned reads among each taxonomic domain based on BLASTX searches of the NR database with an e-value cutoff<10^{-5}.

Domain	lean1 3730xl	lean1 GS20	lean2 3730xl	lean3 3730xl	ob1 3730xl	ob1 GS20	ob2 3730xl
Archaea	1.28	0.658	1.55	1.59	2.07	1.23	2.08
Bacteria	95.8	97.9	90.7	95.1	94.4	93.4	92.9
Eukarya	2.36	1.39	7.36	2.74	2.77	4.15	4.19
(Viruses)	0.527	0.065	0.383	0.611	0.709	1.21	0.782

Supplementary Table 6 – KEGG pathways enriched in the pooled *ob/ob* cecal microbiome relative to the pooled lean cecal microbiome (capillary sequencing datasets, ob1+ob2 vs. lean1+lean2+lean3, binomial test, P<0.05).

KEGG Category	KEGG Pathway[1]
Carbohydrate Metabolism	Starch and sucrose metabolism
	Aminosugars metabolism
	Nucleotide sugars metabolism
Amino Acid Metabolism	Lysine biosynthesis
Metabolism of Other Amino Acids	D-Alanine metabolism
Glycan Biosynthesis and Metabolism	N-Glycan degradation
	Glycosaminoglycan degradation
	Glycosphingolipid metabolism
Biosynthesis of Polyketides and Nonribosomal Peptides	Polyketide sugar unit biosynthesis
	Biosynthesis of vancomycin group antibiotics
Transcription	Other and unclassified family transcriptional regulators
Folding, Sorting and Degradation	Type III secretion system
	Membrane Transport
	ABC transporters
Folding, Sorting and Degradation	Phosphotransferase system (PTS)
Signal Transduction	Two-component system
Cell Motility	Bacterial chemotaxis
	Flagellar assembly
	Bacterial motility proteins
Cell Growth and Death	Sporulation

[1]Only pathways with greater than ten hits in both pooled datasets are shown.

Supplementary Table 7 – 16S rRNA gene-sequence libraries from microbiota transplant experiments.

Label in Fig. S4	ARB label	Host Genotype	16S gene sequences
lean donor 1	lean2	+/+	166
ob/ob donor 1	obob1	*ob/ob*	199
ob/ob donor 2	obob2	*ob/ob*	229
lean recipient 1	SWPT11	+/+	248
lean recipient 2	SWPT13	+/+	265
lean recipient 3	SWPT18	+/+	247
lean recipient 4	SWPT19	+/+	278
lean recipient 5	SWPT20	+/+	271
ob/ob recipient 1	SWPT1	+/+	219
ob/ob recipient 2	SWPT2	+/+	268
ob/ob recipient 3	SWPT3	+/+	280
ob/ob recipient 4	SWPT4	+/+	272
ob/ob recipient 5	SWPT5	+/+	290
ob/ob recipient 6	SWPT12	+/+	197
ob/ob recipient 7	SWPT14	+/+	272
ob/ob recipient 8	SWPT15	+/+	198
ob/ob recipient 9	SWPT16	+/+	258
TOTAL	**-**	**-**	**4,157**

Supplementary Table 8 – KEGG pathways depleted in the pooled *ob/ ob* cecal microbiome relative to the pooled lean cecal microbiome (capillary sequencing datasets, ob1+ob2 vs. lean1+lean2+lean3, binomial test, P<0.05).

KEGG Category	KEGG Pathway[1]
Carbohydrate Metabolism	Glycolysis / Gluconeogenesis
	Citrate cycle (TCA cycle)
	Pentose phosphate pathway
	Pentose and glucuronate interconversions
	Fructose and mannose metabolism
Energy Metabolism	Carbon fixation
	Reductive carboxylate cycle (CO2 fixation)
	Pyruvate/Oxoglutarate oxidoreductases
Lipid Metabolism	Fatty acid metabolism
Nucleotide Metabolism	Pyrimidine metabolism
Amino Acid Metabolism	Glutamate metabolism
	Glycine, serine and threonine metabolism
	Cysteine metabolism
	Arginine and proline metabolism
	Phenylalanine, tyrosine and tryptophan biosynthesis
Glycan Biosynthesis and Metabolism	Lipopolysaccharide biosynthesis
Metabolism of Cofactors and Vitamins	Riboflavin metabolism
	Folate biosynthesis
Translation	Ribosome
Folding, Sorting and Degradation	Other ion-coupled transporters

[1]Only pathways with greater than ten hits in both pooled datasets are shown.

Supplementary Table 9 – COG categories involved in information storage and cellular processes that are enriched or depleted in the pooled *ob/ob* cecal microbiome relative to the pooled lean cecal microbiome (capillary sequencing datasets, ob1+ob2 vs. lean1+lean2+lean3, binomial test, P<0.05).

ENRICHED
[K] Transcription
[L] Replication, recombination, repair
[Y] Nuclear structure
[T] Signal transduction
[M] Cell wall/membrane/envelope biogenesis
[N] Cell motility

DEPLETED
[J] Translation
[V] Defense mechanisms
[O] Posttranslational modification, protein turnover, chaperones