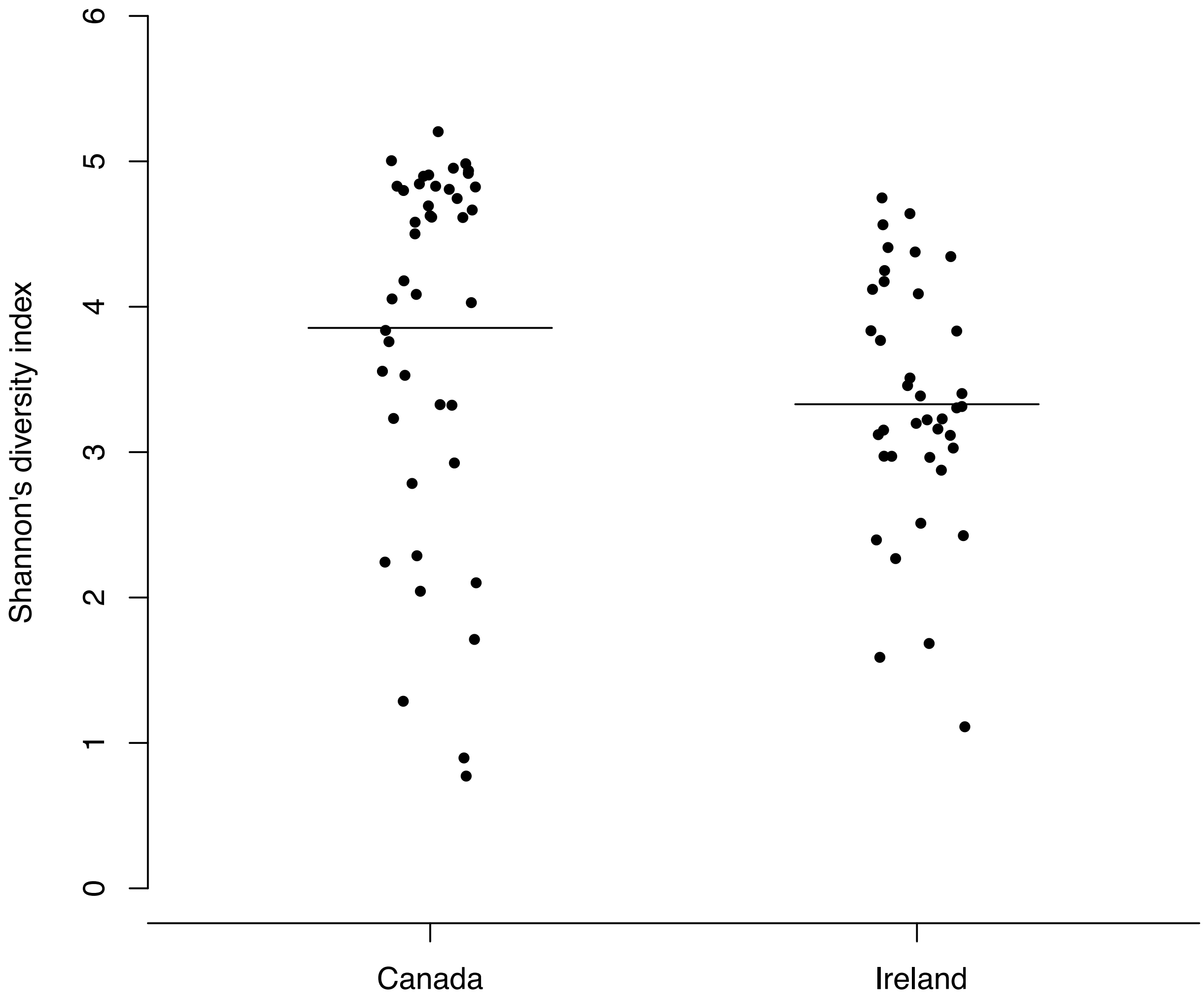

Fig. S1. Bacterial diversity of the breast microbiota. Bacterial diversity within a sample (i.e. alpha diversity) was measured by calculating Shannon's diversity index. Each point on the graph represents a subject with the line representing the mean for all samples. The higher the index the greater the bacterial diversity found within a sample. Since Shannon's diversity index is a logarithmic number with a base of 2, a value of "4" is 2x higher than a value of "3". The mean value for the London samples was 3.9 (n=43) and for the Irish samples, 3.3 (n=33).

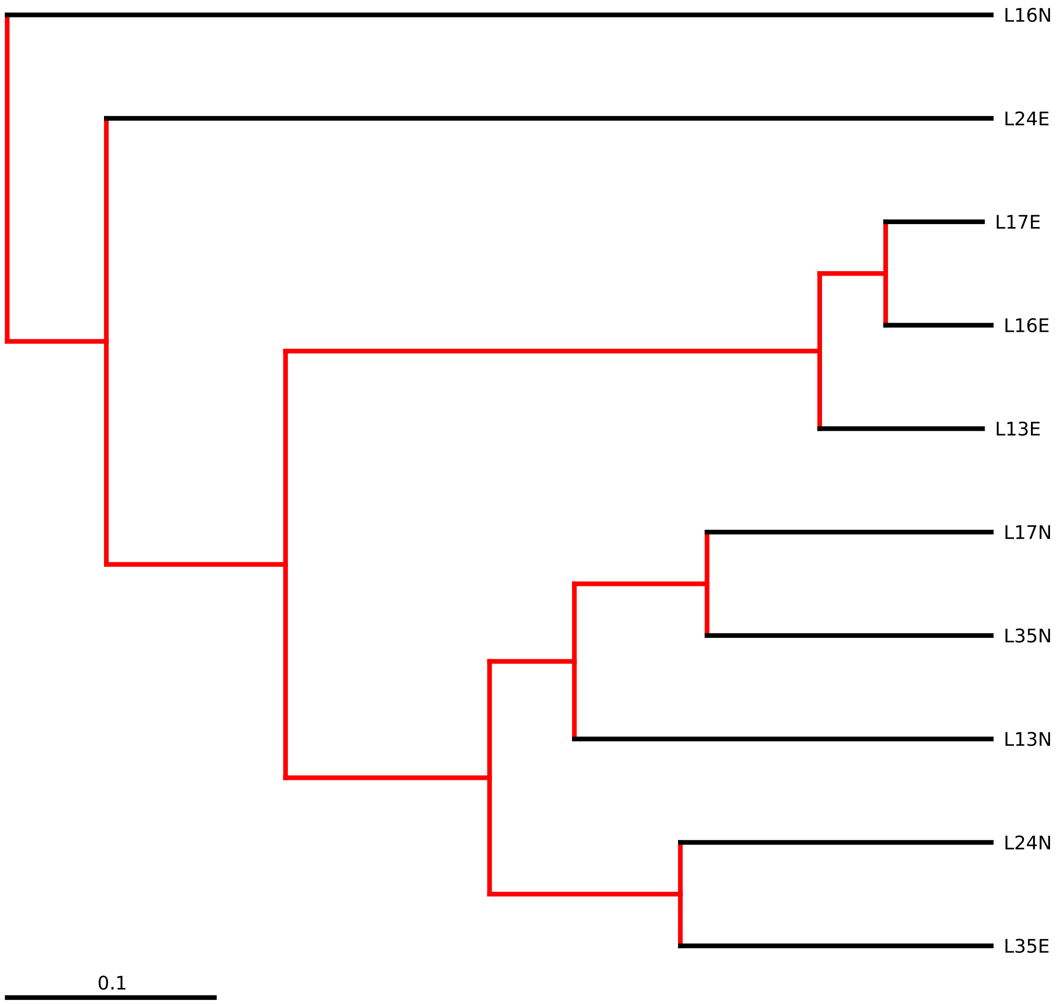

Fig. S2. UPGMA hierarchical clustering comparing Canadian tissue samples with their respective environmental controls. UPGMA hierarchical clustering translates the UniFrac weighted distances into a rooted tree. Samples that cluster together are more similar in bacterial profiles (presence and abundance) than those that are farther apart on the tree. Jackknifing and bootstrapping were performed to measure the robustness of these observations. The red coloured internal nodes provide strong support (75-100%) for similarities or differences between samples. The horizontal scale bar at the bottom indicates 1% sample divergence. As shown above, there is strong support that the bacterial profile between a tissue sample and its respective environmental control is distinct. Reads were rarified to 475 reads/sample. “E”= environmental control and “N”= tissue sample.

Table S1. Primers used in this study. The **N** region in the V6-LT primer refers to the unique barcode sequence. A different barcoded primer was used for each sample. W refers to either A or T and R to either A or G.

Primer Name	Sequence	Ref
V6-LT	5’CCATCTCATCCCTGCGTGTCTCCGACTCAG**NNNNN**CWACGCGARGAACCTTACC 3’	This study
V6-RT	5’CCTCTCTATGGGCAGTCGGTGATACRACACGAGCTGACGAC 3’	
pA	5’ AGAGTTTGATCCTGGCTCAG 3’	(40)
pH	5’ AAGGAGGTGATCCAGCCGCA 3’	

Dataset Legends

Dataset S1. Complete summary of taxonomic results and full length V6 16S rRNA sequence of each OTU. Taxonomy was assigned to operational taxonomic units (OTUs) by using the Seqmatch tool of the Ribosomal database project (RDP). From the top 20 matches to the RDP named isolates database, the full taxonomy was retained for matches with the highest S_ab score. For multiple top matches with equal scores, the lowest common taxonomy was retained (e.g. genus level if multiple species matched equally well). This classification was then verified by BLAST against the Greengenes named isolates database with an output of 100 hits. Taxonomy was assigned based on hits with the highest % identities/coverage. If multiple hits fulfilled this criterion, classification was re-assigned to a higher common taxonomy. In instances where the highest % identity/coverage yielded a single match, if this were less than 90% and the S_ab score from RDP was less than 0.7, taxonomy was assigned at the Family level instead of at the Genus level. The last column shows the sequence of each OTU after clustering at >97% sequence similarity.

Dataset S2. Clinical data of subjects participating in the study. Tissue was collected from 43 women from London, Canada and 38 women from Cork, Ireland. Non-malignant tissue collected adjacent to the tumour is referred to as "normal adjacent" while tissue collected from women without cancer, (i.e those undergoing breast reductions), is referred to as "normal." DCIS= ductal *in situ* carcinoma. Of interest, at least 10 of the patients recruited from Canada had never been pregnant.

Dataset S3. Taxonomic classification of OTU sequences. Taxonomy was assigned to operational taxonomic units (OTUs) by using the Seqmatch tool of the Ribosomal database project (RDP). From the top 20 matches to the RDP named isolates database, the full taxonomy was retained for matches with the highest S_ab score. For multiple top matches with equal scores, the lowest common taxonomy was retained (e.g. genus level if multiple species matched equally well). This classification was then verified by BLAST against the Greengenes named isolates database with an output of 100 hits. Taxonomy was assigned based on hits with the highest % identities/coverage. If multiple hits fulfilled this criterion, classification was re-assigned to a higher common taxonomy.

Supplemental Data References

40. **Edwards U, Rogall T, Blocker H, Emde M, Bottger EC.** 1989. Isolation and direct complete nucleotide determination of entire genes. Characterization of a gene coding for 16S ribosomal RNA. Nucleic Acids Res. **17:**7843-7853.