

1 **The microbiota of breast tissue and its association with tumours**

2 Camilla Urbaniak^{1,2}, Gregory B. Gloor³, Muriel Brackstone⁴, Leslie Scott⁴, Mark Tangney⁵ and

3 Gregor Reid^{1,2,#}

4 **Affiliations:**

5 ¹Lawson Health Research Institute, London, ON, N6A 4V2, Canada

6 ²Department of Microbiology & Immunology, Western University, London, ON, N6A 5C1,

7 Canada

8 ³Department of Biochemistry, Western University, London, ON, N6A 5C1, Canada

9 ⁴London Regional Cancer Program, London, Ontario, N6A 4L6, Canada

10 ⁵Cork Cancer Research Centre, University College Cork, Cork, Ireland

11 # Corresponding author: Gregor Reid, Lawson Health Research Institute, 268 Grosvenor Street,
12 London, Ontario, Canada, N6A 4V2. e-mail: gregor@uwo.ca. Tel: 519-646-6100 x65256.
13 Fax: 519-646-6031

14 **Key words:** breast tissue microbiota, 16S rRNA amplicon sequencing, breast cancer, DNA
15 damage
16

17

18

19

20 **Conflict of interest:** The authors declare no competing financial interests

21

22

23

24

25

26

27

28

29 **Abstract:** In the United States, 1 in 8 women will be diagnosed with breast cancer in her
30 lifetime. Along with genetics, the environment also contributes to disease development but what
31 these exact environmental factors are remain unknown. We have previously shown that breast
32 tissue is not sterile but contains a diverse population of bacteria. We thus believe that the host's
33 local microbiome could be modulating the risk of breast cancer development. Using 16S rRNA
34 amplicon sequencing we show that bacterial profiles differ between normal adjacent tissue from
35 women with breast cancer and tissue from healthy controls. Women with breast cancer had
36 higher relative abundances of *Bacillus*, *Enterobacteriaceae* and *Staphylococcus*. *Escherichia coli*
37 (member of the *Enterobacteriaceae* family) and *Staphylococcus epidermidis*, isolated from breast
38 cancer patients, were shown to induce DNA double stranded breaks in HeLa cells using the
39 γ H2AX assay. We also found that microbial profiles are similar between normal adjacent tissue
40 and tissue sampled directly from the tumour. This novel study raises important questions as to
41 what role the breast microbiome plays in disease development or progression and how we can
42 manipulate this microbiome for possible therapeutics or prevention.

43 **Statement of significance:** This study shows that different bacterial profiles in breast tissue
44 exist between healthy women and those with breast cancer. Higher relative abundances of
45 bacteria, that had the ability to cause DNA damage *in vitro*, were detected in breast cancer
46 patients, as well as a decrease in some lactic acid bacteria, known for their beneficial health
47 effects, including anti-carcinogenic properties. This study raises the important question as to the
48 role of the mammary microbiome in modulating the risk of breast cancer development.

49
50
51

52 **Introduction**

53 Bacteria inhabit numerous body sites and this collection of bacteria, termed the human
54 microbiota, plays an integral role in human development. Changes in the composition of one's
55 microbiota, at various body sites, could promote disease progression, as individuals with
56 periodontitis (1)(2), inflammatory bowel disease (3), psoriasis (4), asthma (5), diabetes (6),
57 bacterial vaginosis (7) and colorectal cancer (8) have different bacterial communities compared
58 with healthy individuals. While it is still unclear whether these microbial differences are a
59 consequence or a cause of the disease, there is evidence in favor of the latter, as healthy animals
60 transplanted with feces from those with obesity, colitis or colorectal cancer then go on to develop
61 disease (9–11).

62 In the United States, 1 in 8 women will be diagnosed with breast cancer in her lifetime.
63 While the etiology of breast cancer is still unknown, it is believed to be due to a combination of
64 both genetic and environmental factors. Support for environmental factors comes from migration
65 studies showing an increased incidence of breast cancer amongst migrants and their descendants
66 after they move from a region of low breast cancer risk to a region of high risk (12, 13). Bacterial
67 communities within the host could be one such environmental factor which has not been
68 considered to date.

69 We have previously shown that a breast tissue microbiome exists in a cohort of Canadian
70 and Irish women (14). To determine whether this local microbiome could have a role in
71 modulating the risk of breast cancer development we examined the breast microbiota of 70
72 women who had either breast cancer (normal adjacent tissue collected), benign tumours (normal
73 adjacent tissue collected) or were disease free. Bacteria isolated from cancer patients were
74 characterized and examined for their ability to induce DNA damage.

75

76 **Methods**77 *Microbiome analysis*78 Tissue collection and processing

79 Fresh breast tissue was collected from 71 women (aged 19 to 90) undergoing breast
80 surgery at St. Joseph's Hospital in London, Ontario, Canada. Ethical approval was obtained from
81 Western Research Ethics Board and Lawson Health Research Institute, London, Ontario, Canada.
82 Subjects provided written consent for sample collection and subsequent analyses. Fifty-eight
83 women underwent lumpectomies or mastectomies for either benign (n=13) or cancerous (n=45)
84 tumours, and 23 were free of disease and underwent either breast reductions or enhancements.
85 For those women with tumours, the tissue obtained for analysis was collected outside the
86 marginal zone, approximately 5 cm away from the tumour. None of the subjects had been on
87 antibiotics for at least 3 months prior to collection.

88 After excision, fresh tissue was immediately placed in a sterile vial on ice and
89 homogenized within 30 min of collection. As an environmental control, a tube filled with 1ml of
90 sterile phosphate-buffered saline (PBS) was left open for the duration of the surgical procedure
91 and then processed in parallel with the tissue samples. As an added control, a skin swab was
92 collected of the disinfected breast area prior to surgery. The swab was placed in 1ml of sterile
93 PBS and then vortexed at full speed for 5min to pellet the contents of the swab. The swab was
94 then removed and the liquid stored at -80°C until DNA extraction.

95 Tissue samples were homogenized in sterile PBS using a PolyTron 2100 homogenizer at
96 28,000 rpm. The amount of PBS added was based on the weight of the tissue in order to obtain a
97 final concentration of 0.4 g/ml. The homogenate was then stored at -80°C until DNA extraction.

98

99

100 DNA isolation

101 After tissue homogenates, in sealed containers, were thawed on ice, 400µl (equivalent to
102 160 mg of tissue) was added to tubes containing 1.2 ml of ASL buffer (QIAamp DNA stool kit;
103 Qiagen) and 400 mg of 0.1-mm-diameter zirconium-glass beads (BioSpec Products). 800µl of the
104 PBS control and 800µl of the skin swab control was also added to tubes containing ASL buffer
105 and beads. Mechanical and chemical lyses were performed on all samples by bead beating at
106 4,800 rpm for 60s at room temperature and then 60s on ice (repeated twice) (Mini-Beadbeater-1;
107 BioSpec Products), after which the suspension was incubated at 95°C for 5 min. Subsequent
108 procedures were performed using the Qiagen QIAamp DNA stool kit according to the
109 manufacturer's protocol, with the exception of the last step, in which the column was eluted with
110 120µl of elution buffer. DNA was stored at -20°C until further use.

111 V6 16S rRNA gene sequencing

112 *PCR amplification*

113 The genomic DNA isolated from the clinical samples was amplified using barcoded
114 primers that amplify the V6 hypervariable region of the 16S rRNA gene (70 base pairs long):

115 V6-Forward:

116 5'ACACTCTTTCCTACACGACGCTCTTCCGATCTnnnn(8)CWACGCGARGAACCTTACC 3'

117 V6-Reverse:

118 5'CGGTCTCGGCATTCCTGCTGAACCGCTCTTCCGATCTnnnn(8)ACRACACGAGCTGACGA
119 C 3'

120 nnnn indicates 4 randomly incorporated nucleotides and "8" represents a specific sample barcode
121 sequence. The PCR was carried out in a 42µl reaction containing 2µl of DNA template (or
122 nuclease-free water as a negative control), 0.15 µg/µl bovine serum albumin, 20µl of 2x GoTaq
123 hot start colorless master mix (Promega) and 10µl of each primer (initial concentration:

124 3.2pmol/μl). Thermal cycling was carried out in an Eppendorf Mastercycler under the following
125 conditions: initial denaturation at 95°C for 2 minutes followed by 25 cycles of 95°C for 1 minute,
126 55°C for 1 minute and 72°C for 1 minute. After amplification, the DNA concentration was
127 measured with the Qubit® 2.0 Fluorometer (Invitrogen) using the broad range assay. Equimolar
128 amounts of each PCR product were then pooled together and purified using the QIAquick PCR
129 purification kit (QIAGEN). The pooled PCR purified sample was then paired-end sequenced on
130 the Illumina Mi-Seq platform using a 150 cycle kit with a 2x80 run at the London Regional
131 Genomics Center, London, Ontario, Canada following standard operating procedures.

132 Sequence processing and taxonomic assignment

133 Custom Perl and Bash scripts were used to de-multiplex the reads and assign barcoded
134 reads to individual samples. Multiple layers of filtering were employed: (i) Paired end sequences
135 were overlapped with Pandaseq, allowing 0 mismatches in the overlapped reads; (ii) Reads were
136 kept if the sequence included a perfect match to the V6 16S rRNA gene primers; (iii) Barcodes
137 were 8mers with an edit distance of >4 and reads were kept if the sequence were a perfect match
138 to the barcode; (iv) Reads were clustered by 97% identity into operational taxonomic units
139 (OTUs) using the Uclust algorithm of USEARCH v7 (15) which has a *de novo* chimera filter
140 built into it; (v) All singleton OTUs were discarded and those that represented $\geq 2\%$ of the reads
141 in at least one sample were kept (a filter for PCR and environmental controls and the skin swabs).
142 Taxonomic assignments for each OTU were made by extracting the best hits from the SILVA
143 database (16) and then manually verified using the Ribosomal Database Project (RDP) SeqMatch
144 tool (rdp.cme.msu.edu) and by BLAST against the Green genes database (greengenes.lbl.gov)
145 Taxonomy was assigned based on hits with the highest percentage identities and coverage. If
146 multiple hits fulfilled this criterion, classification was re-assigned to a higher common taxonomy.

147 Data analysis

148 PCoA plots of weighted UniFrac distances (17) were generated in QIIME (18) by using a
149 tree of OTU sequences built with FASTTREE (19) based on an OTU sequence alignment made
150 with MUSCLE (20). PERMANOVA was used to test for statistical significance between groups
151 using 10000 permutations (QIIME package).

152 Microbiome data are compositional in nature (i.e. proportional distributions that are not
153 independent of each other) and thus has several limitations (21). A simple example is as follows:
154 If a sample has two organisms A (50%) and B (50%) and after antibiotic treatment organism A is
155 completely killed, the proportion of B in that sample will now be 100% even if its actual
156 abundance has not changed. Transforming the data, using centered log-ratios (CLR) alleviates
157 the constraints inherent with compositional data (22) by allowing for subcomposition coherence,
158 linear sample independence and normalization of read counts. CLR transformed data with a
159 uniform prior of 0.5 applied was used when generating the K-means clusterplot and the
160 dendrogram of Euclidian distances.

161 The ALDEx R package version 2 (21) was used to compare the relative abundance of
162 genera. Values reported in this manuscript represent the expected values of 128 Dirichlet Monte-
163 Carlo instances of CLR transformed data. A value of zero indicates that organism abundance is
164 equal to the geometric mean abundance. Thus, organisms more abundant than the mean will have
165 positive values and those less abundant than the mean will have negative values. Base 2 was used
166 for the logarithm so differences between values represent fold changes. Significance was based
167 on the Benjamini-Hochberg corrected p-value of the Wilcoxon rank test (significance threshold
168 p-val < 0.1)

169 The Microbiome Regression-based Kernel Association Test (MiRKAT) (23) was
170 performed in R using the MiRKAT package. Differences in microbiota profiles were tested using

171 a kernel metric constructed from weighted UniFrac, unweighted UniFrac and GUniFrac(24)
172 distances and the Bray-Curtis dissimilarity metric. “Optimal” MiRKAT allows for the
173 simultaneous examination of multiple distance/dissimilarity metrics alleviating the problem of
174 choosing the best one and was performed on the aforementioned metrics. The p-values generated
175 were the mean of 128 Dirichlet Monte-Carlo instances.

176 The R script of “SourceTracker “(version 0.9.1) was used to assess contamination of the
177 tissue microbiota. Tissue samples were designated as “sink” and PBS controls as “source.”

178 Barplots, boxplots, K-means clusterplots and dendograms were all generated in R
179 (<http://www.R-project.org/>).

180 Full details regarding Irish tissue sample collection, patient demographics, DNA
181 extraction protocols and the steps followed to generate the OTU table used for the analysis in
182 Supplementary Figure S4 can be found in our previous publications (14, 25)

183 *DNA damage assay*

184 Bacterial strains

185 Isolates were obtained by plating 100µl of tissue homogenate (normal adjacent tissue and
186 healthy tissue from Canadian subjects) on Columbia blood (CBA), MacConkey and Beef Heart
187 Infusion (BHI) agar plates and incubating both aerobically or anaerobically at 37°C. DNA from
188 single colonies was extracted using the InstaGene Matrix (Bio-Rad) and then amplified using the
189 eubacterial primers pA/pH, which amplifies the complete 16S rRNA gene: pA 5'
190 AGAGTTTGATCCTGGCTCAG 3' pH 5' AAGGAGGTGATCCAGCCGCA 3'

191 The PCR reaction was carried out in 50µl reaction containing 10µl of DNA template (or nuclease
192 free water as a negative control), 1.5mM MgCl₂, 1.0µM of each primer, 0.2mM dNTP, 5µl 10X
193 PCR buffer (Invitrogen), and 0.05 Taq Polymerase (Invitrogen). Thermal cycling was carried out
194 in an Eppendorf Mastercycler under the following conditions: Initial denaturation step at 95°C for

195 2min, followed by 30 cycles of 94°C for 30s, 55°C for 30s and 72°C for 1min. A final elongation
196 step was performed at 72°C for 10min. 40µl of the PCR mixture was then purified using the
197 QIAquick PCR purification kit (Qiagen) and the purified products sent for Sanger sequencing to
198 the London Regional Genomics Centre, London, Ontario, Canada. Sequences were analyzed
199 using the GenBank 16S ribosomal RNA sequences database and the Greengenes database.
200 Taxonomy was assigned based on the highest Max score. Because the 16S rRNA gene does not
201 differentiate members of the *Enterobacteriaceae* family very well, to confirm that our isolates
202 were indeed *E.coli*, we utilized the API[®] 20E strip to differentiate species that are part of the this
203 family. *E.coli* strain IHE3034 was kindly provided by Jean Philippe Nougayrède (INRA,
204 Toulouse, France).

205 Infection assay

206 HeLa cells were maintained and passaged in DMEM/glutamax media (Invitrogen)
207 supplemented with 10% FBS (Invitrogen). On the day of the experiment a 24 well plate
208 containing sterile cover slips was seeded with 0.5ml of 1x10⁵ cells/ml, resulting in 5x10⁴ HeLa
209 cells/well. The plates were then incubated at 37°C with 5% CO₂ for 24 hours after which times
210 the media was removed and the wells washed with sterile PBS. HeLa cells (2 wells for each
211 organism) were then infected at a MOI 100 for 4 hours with either *Staphylococcus epidermidis*
212 (subject 31), *Micrococcus luteus* (subject 8), *Micrococcus sp* (subject 8), *E.coli* (subject 41
213 (isolates H and E), subject 34 and strain IHE3034), *Propionibacterium acnes* (subject 20) and *P.*
214 *granulosum* (subject 20) or at a MOI 1 for 2 hours with *Bacillus cereus* (subject 34 & subject 6).
215 A MOI of 1 was used for *B. cereus* (for 2 hours) instead of 100 (for 4 hours), as used for the other
216 strains, because this was the highest MOI and longest incubation that the HeLa cells could
217 tolerate without dying. The bacterial cultures for infection were prepared by inoculating 5ml of

218 BHI with 1 colony and incubating aerobically at 37°C for 15 hours, with the exception of
219 *Propionibacterium*, which was incubated anaerobically for 72hr. Bacterial cultures were then
220 spun down at 3500g for 10min, washed and resuspended in PBS. Bacterial cells were then
221 diluted to the appropriate concentration in DMEM media containing 10% FBS and 25mM
222 HEPES. 40µM etoposide (Sigma) was used as a technical positive control. The pH was checked
223 at the end of the experiment to ensure consistency between wells.

224 Immunofluorescence

225 After infection, media was removed and HeLa cells were washed 3x with sterile PBS.
226 Cells were then fixed and permeabilized for 12min at room temperature (RT) with a -20°C
227 solution of 95% methanol and 5% acetic acid. Cells were then blocked for an hour with 0.3%
228 Triton-100/5% goat serum. After blocking, a 1/200 dilution of the primary antibody (rabbit anti
229 phospho-H2AX mAb; Cell Signaling technologies) was added and incubated over night at 4°C.
230 After washing a 1/1000 dilution of the secondary antibody (goat anti-rabbit IgG, Alexa Fluor 647
231 conjugate; Cell Signaling technologies) was added and incubated at RT for 30min. Cells were
232 then counter stained with 1µg/ml of DAPI (Life Technologies) for 1min. Cover slips were
233 mounted on microscope slides containing a drop of ProLong Gold antifade mountant (Life
234 Technologies). The experiments were performed three times.

235 Images were captured using the NIKON eclipse TE2000-S digital microscope. Eight
236 fields of view for each replicate were recorded, for a total of 16 fields of view for each condition.
237 Using ImageJ software (version 1.48a), the mean fluorescent intensity of each γH2AX stained
238 cell was measured from the digital images. The digital images were also used to determine the
239 percent of total cells stained positive for γH2AX. This was calculated by dividing the number of

240 red cells (i.e. γ H2AX positive) by the number of blue cells (i.e. DAPI stained) and multiplying by
241 one hundred.

242 Statistics for DNA damage assay

243 Bar graphs of the mean and standard deviation from the 3 experiments were plotted using Prism
244 (version 5.0a). Significance ($p < 0.05$) was tested by a 1 way ANOVA followed by the Dunnet's
245 post hoc test using Prism (version 5.0a)

246 **Results**

247 **Microbiota analysis**

248 16S rRNA amplicon sequencing of the V6 hypervariable region was performed on 70 tissue
249 samples and 38 environmental controls. A full summary of patient demographics can be found in

250 **Supplementary Table S1**. To assess the contribution of environmental contamination towards
251 the overall tissue microbiota, we utilized the contamination predictor tool, "SourceTracker",
252 which compared the microbial population in the tissue samples to that of the phosphate buffered
253 saline (PBS) environmental controls that were processed alongside the tissue samples.

254 **Supplementary Fig. S1** shows that while there is contamination present, it makes up only a
255 small proportion (average 10%) of the overall microbial community in breast tissue. A
256 dendrogram of Euclidian distances of the centered log-ratio (clr) transformed dataset (22) was
257 then constructed to visualize which tissue samples were similar to the PBS controls and to skin
258 swabs collected from the disinfected breast area prior to surgery. As seen in **Supplementary Fig.**
259 **S2**, skin swabs, PBS controls and the no template PCR control (NTC) formed a single cluster,
260 which was separate from most of the tissue samples, indicating distinct microbial profiles. To
261 ensure stringent quality control, we removed those tissue samples (27 of them) that were part of
262 the PBS/skin/NTC group from further analysis (**Supplementary Table S2**). In addition, OTUs

263 present in over 2% abundance in the NTC and PBS controls (11 of them) were also removed
264 from further analysis (**Supplementary Table S2**). 16S rRNA gene sequencing data of the
265 remaining samples and OTUs, showed a diverse population of bacteria consisting of 61 OTUs
266 and 28 genera (**Fig. 1A**) dominated by the phyla *Proteobacteria* and *Firmicutes* (**Fig. 1B**).

267 A comparison of normal adjacent tissue from women with breast cancer with that of tissue
268 from healthy women showed distinctly different bacterial profiles on weighted UniFrac PCoA
269 plots (**Fig. 2A**). The PERMANOVA test performed on the dataset showed that the observed
270 differences were statistically significant (10000 permutations; pseudo F-statistic=14.4; p-value
271 <0.01). Unsupervised K-means clustering of the clr transformed dataset indicated two clusters
272 and the PCA plot in **Figure 2B** shows clear separation between the healthy and cancer groups.
273 Differences between the groups were further confirmed using the Microbiome Regression-based
274 Kernel Association Test (MiRKAT) (**Table 1**).

275 ALDEx2, which allows for the direct comparison of bacterial taxa between groups showed
276 significantly higher compositional abundances of *Prevotella*, *Lactococcus*, *Streptococcus*
277 *Corynebacterium* and *Micrococcus* in healthy patients and *Bacillus*, *Staphylococcus*,
278 *Enterobacteriaceae* (unclassified), *Comamonadaceae* (unclassified) and *Bacteroidetes*
279 (unclassified) in cancer patients (**Fig. 3**) (**Supplementary Table S3**)

280 To assess whether bacteria surrounding the tumour microenvironment might be associated
281 with the severity of cancer, we compared bacterial profiles in normal adjacent tissue from women
282 with various stages of breast cancer. No differences were found based on invasiveness or stage
283 (**Supplementary Fig. S3**). However normal adjacent tissue from women with benign tumours
284 had profiles that were more similar to normal adjacent tissue of women with cancerous tumours
285 rather than tissue from healthy subjects (**Supplementary Table S4**). It is important to note that

286 no differences were observed between tissue samples collected by different surgeons and/or from
287 different surgical rooms

288 We have previously published two reports showing which bacteria are present in tumour
289 tissue and normal adjacent tissue of women from Ireland (14, 25). In this report, we now show,
290 using weighted UniFrac distances, that bacterial communities do not differ between tumour tissue
291 and normal adjacent tissue, both at the population level (**Supplementary Fig. S4A**) and within an
292 individual (**Supplementary Fig. S4B**). Thus, when suitably-collected tumour tissue for
293 microbiome analysis is not available, normal adjacent tissue may be a practical alternative.

294 **Assessment of DNA damage ability of breast tissue isolates**

295 *E.coli* strains belonging to the B2 phylotype harbour the *pks* pathogenicity island, which
296 encodes for machinery for the production of the genotoxin, colibactin. These *pks* + strains have
297 been implicated in colon cancer (26, 27) via its ability to induce DNA double stranded breaks and
298 chromosomal instability (28, 29). As shown in Figure 3, the family *Enterobacteriaceae*, of
299 which *E.coli* is a member, was relatively more abundant in cancer patients compared to healthy
300 controls. For this reason, we wanted to examine whether *E. coli*, cultured from normal adjacent
301 tissue of breast cancer patients, had the ability to induce DNA double stranded breaks (DSB).
302 Cellular levels of γ H2AX, a surrogate marker of DSB, were measured in HeLa cells after
303 incubation with various *E.coli* tissue isolates. *E.coli* IHE3034, which contains the *pks*
304 pathogenicity island and induces DSB (28) was used for comparison.

305 HeLa cells exposed to *E.coli* tissue isolates had significantly higher levels of γ H2AX
306 compared with untreated cells, as measured by mean fluorescent intensity (MFI) and % of cells
307 that stained positive for γ H2AX, with levels equivalent to that induced by *E.coli* IHE3034
308 (**Figure 4**). Additional isolates from breast cancer patients were also examined for the ability to

309 induce DNA damage; (i) *Bacillus* and *Staphylococcus* were tested as these genera were more
310 abundant in cancer patients; (ii) *Micrococcus*, as this genus was higher in healthy individuals and
311 (iii) *Propionibacterium* as there were no differences in relative abundances between cancer
312 patients and healthy controls. Neither *Bacillus*, *Micrococcus* nor *Propionibacterium* isolates
313 induced DSB, whereas *Staphylococcus* did (**Supplementary Fig. S5**).

314 γ H2AX foci can occur and be resolved very quickly in response to DNA damage, thus a time
315 course was performed, with *Bacillus* treated cells analyzed every 15min over a 2 hour period. No
316 statistically significant differences at any time point were observed between treated and untreated
317 cells (data not shown).

318 **Discussion**

319 This study has shown that different bacterial profiles exist in “normal adjacent” breast
320 tissue from women with breast cancer compared with “normal” tissue from healthy controls. In
321 colorectal cancer (CRC) and oral squamous cell carcinoma (OSCC) bacterial profiles in the stool
322 and saliva respectively, also differ between healthy and diseased patients (30–32) with evidence
323 suggesting that changes in this community composition and function may be driving cancer
324 progression at these sites (33, 34). This raises the possibility that the differences observed in the
325 breast could also play a role in breast cancer progression. We acknowledge that the average age
326 differed between the two groups with the cancer cohort having a mean and median age of 62 and
327 the healthy cohort having a mean age of 49 and a median age of 53. Considering that the mean
328 and median age of the benign group was 38 and 36 respectively and the microbial profiles did not
329 differ between the benign and cancer groups, we don't believe that the difference observed
330 between the healthy and cancer group were due to difference in ages. Menopausal status does not
331 appear to be a factor either since no differences in microbial profiles were observed between pre

332 and post menopausal women in the healthy cohort and pre and post menopausal women with
333 either benign or cancerous tumours.

334 *Enterobacteriaceae* and *Staphylococcus* are two taxa found in higher abundance in breast
335 cancer patients compared with healthy controls. Examination of three *E.coli* isolates (a member
336 of the *Enterobacteriaceae* family) and one *Staphylococcus epidermidis* isolate, cultured from
337 normal adjacent tissue of breast cancer patients, all displayed the ability to induce DNA double
338 stranded breaks (DSB). DSB are the most detrimental type of DNA damage and are caused by
339 genotoxins, reactive oxygen species, and ionizing radiation(35). Non-homologous end joining
340 (NHEJ), the mechanism by which DSB are repaired, is extremely error- prone often resulting in
341 missing bases at the site of damage (35). Accumulation of these mis-repairs within the cell over
342 time leads to genomic instability and eventually cancer (36). DSB caused by bacteria such as
343 *Helicobacter pylori* and certain strains of *E.coli* have been shown to induce chromosomal
344 instability with prolonged exposure (29, 37). While the same mechanisms may be involved in
345 the *in vitro* assay described here (or indeed breast tissue transformation), further tests would need
346 to be done to verify whether chromosomal abnormalities do occur subsequent to the DNA
347 damage induced by these breast isolates. In support of this hypothesis, total cell numbers were
348 consistent between all treated and untreated groups, suggesting no induction of apoptosis. It is
349 important to note that bacterial induced DNA damage may not be sufficient in itself to promote
350 breast cancer development unless it occurs in a genetically susceptible host. All genetic and 3-
351 30% of sporadic cancer cases have mutations in DNA repair or DNA checkpoint machinery (38).
352 Thus women who have impaired DNA repair/checkpoints may be more susceptible to bacterial
353 induced DNA damage and may be at a higher risk of developing breast cancer than women
354 without these mutations, even if they have the same “detrimental” microbes in their mammary
355 glands.

356 *Bacillus* was also elevated in breast cancer patients compared with healthy controls,
357 confirming our previous findings (14). While *Bacillus* did not induce DSB like *E.coli* and
358 *S.epidermidis* it could have other pro-carcinogenic effects. One study has shown that a *Bacillus*
359 *cereus* strain, isolated from gingival plaque, metabolizes the hormone progesterone into 5 alpha-
360 pregnane-3,20-dione (5 α P) (39). 5 α P is higher in breast tumours compared with healthy breast
361 tissue (40) and is believed to promote tumour development by stimulating cell proliferation (40,
362 41). While our molecular analysis did not permit species level identification, all *Bacillus* strains
363 cultured from our breast cancer patients were of the species *B.cereus*.

364 An epidemiological study has shown that women who drink fermented milk products
365 have a reduced risk of breast cancer development, irrespective of multivariable risk factors (42).
366 This protection could be attributed to the health promoting properties of the various lactic acid
367 bacteria (LAB) present in fermented products. *Lactococcus* and *Streptococcus*, two such
368 bacteria that were higher in healthy women compared with breast cancer patients, exhibit anti-
369 carcinogenic properties and could play a role in prevention. Natural killer (NK) cells are vital in
370 controlling tumour growth with epidemiological studies showing that low NK cell activity (from
371 peripheral blood mononuclear cells (PBMC)) is associated with an increased incidence of breast
372 cancer (43, 44). *Lactococcus lactis* has been shown to activate murine splenic NK cells,
373 enhancing cellular immunity (45). While no studies have yet been published comparing NK cell
374 functionality in the breast between “normal” (i.e. healthy patients) and “normal adjacent” (breast
375 cancer patients) tissue, it could be assumed, based on the PBMC data, that NK functionality is
376 also impaired in the breast of those with cancer. *Lactococcus sp* present in the mammary glands,
377 could be modulating cellular immunity by maintaining the cytotoxic activity of resident NK cells
378 (46) thus helping to prevent cancer development. *Streptococcus thermophilus* on the other hand,

379 better than any other LAB tested, protects against DNA damage caused by reactive oxygen
380 species by producing antioxidant metabolites that neutralize peroxide and superoxide radicals
381 (47).

382 Orally administered *Lactobacillus sp*, has shown to be protective in animal models of
383 breast cancer (48). While total numbers did not differ between healthy and diseased patients,
384 those with breast cancer may not experience the full anti-carcinogenic benefits afforded by
385 *Lactobacillus* due to the decrease in *Lactococcus* and *Streptococcus*, as LAB have been shown to
386 act in synergy with each other (49).

387 *Prevotella*, which was more abundant in healthy women compared with breast cancer
388 patients, produces the short chain fatty acid (SCFA), propionate. Propionate, like other SCFA,
389 has many beneficial health effects in the gut, one of them being the ability to regulate colorectal
390 tumour growth (50). In both animal and human studies, higher levels of *Prevotella* were observed
391 in the stool of healthy subjects compared to those with CRC (10, 30). However in the oral cavity,
392 patients with OSCC have higher levels of *Prevotella* compared with healthy controls and when
393 *Prevotella* presence was used as a diagnostic tool, the authors could predict 80% of the cancer
394 cases (32). The conflicting association of *Prevotella* between CRC and OSCC could be due to the
395 fact that metabolites function differently at different body sites. While SCFA are anti-
396 inflammatory in the colon and associated with health (51), in the vagina, they are pro-
397 inflammatory and associated with bacterial vaginosis (52). What role *Prevotella* and/or
398 propionate may be playing in breast health (or disease) remains to be determined.

399 It is interesting that the microbiome profile of normal adjacent tissue from women with
400 benign tumours was similar to that of normal adjacent tissue from cancer patients, rather than
401 normal tissue from healthy women and raises the question as to why these women with benign
402 tumours do not have cancer, if we believe there could be a link between bacteria and breast

403 cancer. In women with benign disease, DNA damage caused by bacteria could be responsible for
404 enhanced cellular proliferation leading to tumour formation, similar to what may be occurring in
405 cancer patients, however, other factors that could promote transformation and malignancy of this
406 tumour is reduced in these women compared to those with cancer. One of these factors could be
407 the increased secretion of pro-angiogenic and/or inflammatory molecules from immune and
408 epithelial cells in women who have cancer. Another possibility is that women with benign
409 tumours have lower levels of DNA damaging bacteria than those with cancerous tumours,
410 lowering the probability of multiple onco-genic genes becoming mutated. Further studies
411 following healthy women and those with benign tumours for development of breast cancer could
412 shed more light on which bacterial strains could be driving cancer development.

413 While we have reported differential abundances of certain organisms between health and
414 diseased states, in reality, it is probably not a single organism driving disease progression or
415 protection but rather an interplay of poly-microbial interactions. To get a better understanding of
416 the microbial influence on breast cancer, the functionality of these microbes should be
417 investigated. Further studies examining bacterial metabolites and bacterial-induced host
418 metabolites would provide vital information on the role of bacteria in breast health.

419 **Conclusion**

420 This study has shown that bacterial profiles differ in breast tissue between healthy
421 subjects and normal adjacent tissue of breast cancer patients. Some of the bacteria that were
422 relatively more abundant in breast cancer patients had the ability to induce DNA double stranded
423 breaks. Further studies need to be done to examine whether this DNA damage can lead to
424 chromosomal aberrations and whether the differences in the bacterial profiles are a cause or a
425 consequence of the disease. This study raises important questions as to the role of the breast

426 microbiota in breast cancer development or prevention and whether bacteria could be harnessed
427 for interventions to help prevent disease onset.

428 **Acknowledgements**

429 We would like to thank all the women who provided tissue samples for the study. In
430 addition, members of the Gloor lab for fruitful discussions on 16S rRNA gene sequencing
431 analysis; Dr. Robert Richards and all the clinical and administrative staff at St Joseph's Hospital
432 for their support throughout the study. We are also very grateful to Jean Philippe Nougayrède
433 (INRA, Toulouse, France) for providing us with *E.coli* strain IHE3034 and for his insight and
434 help with our DNA damage assay.

435 **Author contributions**

436 CU designed the study, recruited subjects, collected and prepared the samples for microbial
437 analysis, analyzed 16S rRNA sequencing data, performed and analyzed the DNA damage assay
438 and wrote the manuscript. GG provided co-supervision and input into microbiome acquisition
439 and analysis and manuscript review. MB helped with study design, provided input into data
440 collection and analysis, collected tissue from tumour patients and reviewed the manuscript. LS
441 helped with study design and collected tissue from tumour patients. MT supervised the collection
442 of tissue samples from women in Ireland, provided input into data interpretation and reviewed the
443 manuscript. GR conceptualized the study, helped with study design and manuscript writing,
444 supervised data collection and analysis and provided financial support.

445 **Financial support:** CU was supported by (i) CIHR Strategic Training Program in Cancer
446 Research and Technology Transfer scholarship and (ii) Translational Breast Cancer Research
447 Studentship.

448
449

450 **References**

- 451 1. **Darveau RP**. 2010. Periodontitis: a polymicrobial disruption of host homeostasis. *Nat.*
452 *Rev. Microbiol.* **8**:481–90.
- 453 2. **Ximénez-Fyvie LA, Haffajee AD, Socransky SS**. 2000. Comparison of the microbiota of
454 supra- and subgingival plaque in health and periodontitis. *J. Clin. Periodontol.* **27**:648–57.
- 455 3. **Frank DN, St Amand AL, Feldman RA, Boedeker EC, Harpaz N, Pace NR**. 2007.
456 Molecular-phylogenetic characterization of microbial community imbalances in human
457 inflammatory bowel diseases. *Proc. Natl. Acad. Sci. U. S. A.* **104**:13780–5.
- 458 4. **Gao Z, Tseng C, Strober BE, Pei Z, Blaser MJ**. 2008. Substantial alterations of the
459 cutaneous bacterial biota in psoriatic lesions. *PLoS One* **3**:e2719.
- 460 5. **Hilty M, Burke C, Pedro H, Cardenas P, Bush A, Bossley C, Davies J, Ervine A,**
461 **Poulter L, Pachter L, Moffatt MF, Cookson WOC**. 2010. Disordered microbial
462 communities in asthmatic airways. *PLoS One* **5**:e8578.
- 463 6. **Larsen N, Vogensen FK, van den Berg FWJ, Nielsen DS, Andreasen AS, Pedersen**
464 **BK, Al-Soud WA, Sørensen SJ, Hansen LH, Jakobsen M**. 2010. Gut microbiota in
465 human adults with type 2 diabetes differs from non-diabetic adults. *PLoS One* **5**:e9085.
- 466 7. **Hummelen R, Fernandes AD, Macklaim JM, Dickson RJ, Chagalucha J, Gloor GB,**
467 **Reid G**. 2010. Deep sequencing of the vaginal microbiota of women with HIV. *PLoS One*
468 **5**:e12078.
- 469 8. **Mira-Pascual L, Cabrera-Rubio R, Ocon S, Costales P, Parra A, Suarez A, Moris F,**
470 **Rodrigo L, Mira A, Collado MC**. 2014. Microbial mucosal colonic shifts associated with
471 the development of colorectal cancer reveal the presence of different bacterial and archaeal
472 biomarkers. *J. Gastroenterol.* **50**:167–179.
- 473 9. **Garrett WS, Lord GM, Punit S, Lugo-Villarino G, Mazmanian SK, Ito S, Glickman**
474 **JN, Glimcher LH**. 2007. Communicable ulcerative colitis induced by T-bet deficiency in
475 the innate immune system. *Cell* **131**:33–45.
- 476 10. **Zackular JP, Baxter NT, Iverson KD, Sadler WD, Petrosino JF, Chen GY, Schloss**
477 **PD**. 2013. The gut microbiome modulates colon tumorigenesis. *MBio* **4**:e00692–13.
- 478 11. **Turnbaugh PJ, Hamady M, Yatsunenko T, Cantarel BL, Duncan A, Ley RE, Sogin**
479 **ML, Jones WJ, Roe BA, Affourtit JP, Egholm M, Henrissat B, Heath AC, Knight R,**
480 **Gordon JI**. 2009. A core gut microbiome in obese and lean twins. *Nature* **457**:480–4.
- 481 12. **Le GM, Gomez SL, Clarke CA, Glaser SL, West DW**. 2002. Cancer incidence patterns
482 among Vietnamese in the United States and Ha Noi, Vietnam. *Int. J. Cancer* **102**:412–7.

- 483 13. **Shimizu H, Ross RK, Bernstein L, Yatani R, Henderson BE, Mack TM.** 1991. Cancers
484 of the prostate and breast among Japanese and white immigrants in Los Angeles County.
485 *Br. J. Cancer* **63**:963–6.
- 486 14. **Urbaniak C, Cummins J, Brackstone M, Macklaim JM, Gloor GB, Baban CK, Scott
487 L, O’Hanlon DM, Burton JP, Francis KP, Tangney M, Reid G.** 2014. Microbiota of
488 human breast tissue. *Appl. Environ. Microbiol.* **80**:3007–14.
- 489 15. **Edgar RC.** 2010. Search and clustering orders of magnitude faster than BLAST.
490 *Bioinformatics* **26**:2460–1.
- 491 16. **Pruesse E, Quast C, Knittel K, Fuchs BM, Ludwig W, Peplies J, Glöckner FO.** 2007.
492 SILVA: a comprehensive online resource for quality checked and aligned ribosomal RNA
493 sequence data compatible with ARB. *Nucleic Acids Res.* **35**:7188–96.
- 494 17. **Lozupone C, Knight R.** 2005. UniFrac: a new phylogenetic method for comparing
495 microbial communities. *Appl. Environ. Microbiol.* **71**:8228–35.
- 496 18. **Caporaso JG, Kuczynski J, Stombaugh J, Bittinger K, Bushman FD, Costello EK,
497 Fierer N, Peña AG, Goodrich JK, Gordon JI, Huttley GA, Kelley ST, Knights D,
498 Koenig JE, Ley RE, Lozupone CA, McDonald D, Muegge BD, Pirrung M, Reeder J,
499 Sevinsky JR, Turnbaugh PJ, Walters WA, Widmann J, Yatsunenko T, Zaneveld J,
500 Knight R.** 2010. QIIME allows analysis of high-throughput community sequencing data.
501 *Nat. Methods* **7**:335–6.
- 502 19. **Price MN, Dehal PS, Arkin AP.** 2009. FastTree: computing large minimum evolution
503 trees with profiles instead of a distance matrix. *Mol. Biol. Evol.* **26**:1641–50.
- 504 20. **Edgar RC.** 2004. MUSCLE: a multiple sequence alignment method with reduced time
505 and space complexity. *BMC Bioinformatics* **5**:113.
- 506 21. **Fernandes AD, Reid JN, Macklaim JM, McMurrough TA, Edgell DR, Gloor GB.**
507 2014. Unifying the analysis of high-throughput sequencing datasets: characterizing RNA-
508 seq, 16S rRNA gene sequencing and selective growth experiments by compositional data
509 analysis. *Microbiome* **2**:15.
- 510 22. **Aitchison J.** 1986. *The Statistical Analysis of Compositional Data.* Chapman & Hall,
511 London, England.
- 512 23. **Zhao N, Chen J, Carroll IM, Ringel-Kulka T, Epstein MP, Zhou H, Zhou JJ, Ringel
513 Y, Li H, Wu MC.** 2015. Testing in Microbiome-Profiling Studies with MiRKAT, the
514 Microbiome Regression-Based Kernel Association Test. *Am. J. Hum. Genet.* **96**:797–807.
- 515 24. **Chen J, Bittinger K, Charlson ES, Hoffmann C, Lewis J, Wu GD, Collman RG,
516 Bushman FD, Li H.** 2012. Associating microbiome composition with environmental
517 covariates using generalized UniFrac distances. *Bioinformatics* **28**:2106–13.

- 518 25. **Lehouritis P, Cummins J, Stanton M, Murphy CT, McCarthy FO, Reid G, Urbaniak**
519 **C, Byrne WL, Tangney M.** 2015. Local bacteria affect the efficacy of chemotherapeutic
520 drugs. *Sci. Rep.* **5**:14554.
- 521 26. **Arthur JC, Perez-Chanona E, Mühlbauer M, Tomkovich S, Uronis JM, Fan T-J,**
522 **Campbell BJ, Abujamel T, Dogan B, Rogers AB, Rhodes JM, Stintzi A, Simpson**
523 **KW, Hansen JJ, Keku TO, Fodor AA, Jobin C.** 2012. Intestinal inflammation targets
524 cancer-inducing activity of the microbiota. *Science* **338**:120–3.
- 525 27. **Buc E, Dubois D, Sauvanet P, Raisch J, Delmas J, Darfeuille-Michaud A, Pezet D,**
526 **Bonnet R.** 2013. High prevalence of mucosa-associated *E. coli* producing cyclomodulin
527 and genotoxin in colon cancer. *PLoS One* **8**:e56964.
- 528 28. **Nougayrède J-P, Homburg S, Taieb F, Boury M, Brzuszkiewicz E, Gottschalk G,**
529 **Buchrieser C, Hacker J, Dobrindt U, Oswald E.** 2006. *Escherichia coli* induces DNA
530 double-strand breaks in eukaryotic cells. *Science* **313**:848–51.
- 531 29. **Cuevas-Ramos G, Petit CR, Marcq I, Boury M, Oswald E, Nougayrède J-P.** 2010.
532 *Escherichia coli* induces DNA damage in vivo and triggers genomic instability in
533 mammalian cells. *Proc. Natl. Acad. Sci. U. S. A.* **107**:11537–42.
- 534 30. **Weir TL, Manter DK, Sheflin AM, Barnett BA, Heuberger AL, Ryan EP.** 2013. Stool
535 microbiome and metabolome differences between colorectal cancer patients and healthy
536 adults. *PLoS One* **8**:e70803.
- 537 31. **Wang T, Cai G, Qiu Y, Fei N, Zhang M, Pang X, Jia W, Cai S, Zhao L.** 2012.
538 Structural segregation of gut microbiota between colorectal cancer patients and healthy
539 volunteers. *ISME J.* **6**:320–9.
- 540 32. **Mager DL, Haffajee AD, Devlin PM, Norris CM, Posner MR, Goodson JM.** 2005. The
541 salivary microbiota as a diagnostic indicator of oral cancer: a descriptive, non-randomized
542 study of cancer-free and oral squamous cell carcinoma subjects. *J. Transl. Med.* **3**:27.
- 543 33. **Schwabe RF, Jobin C.** 2013. The microbiome and cancer. *Nat. Rev. Cancer* **13**:800–12.
- 544 34. **Ahn J, Chen CY, Hayes RB.** 2012. Oral microbiome and oral and gastrointestinal cancer
545 risk. *Cancer Causes Control* **23**:399–404.
- 546 35. **Lees-Miller S.** 2003. Repair of DNA double strand breaks by non-homologous end
547 joining. *Biochimie* **85**:1161–1173.
- 548 36. **Khanna KK, Jackson SP.** 2001. DNA double-strand breaks: signaling, repair and the
549 cancer connection. *Nat. Genet.* **27**:247–54.
- 550 37. **Toller IM, Neelsen KJ, Steger M, Hartung ML, Hottiger MO, Stucki M, Kalali B,**
551 **Gerhard M, Sartori AA, Lopes M, Müller A.** 2011. Carcinogenic bacterial pathogen

- 552 Helicobacter pylori triggers DNA double-strand breaks and a DNA damage response in its
553 host cells. *Proc. Natl. Acad. Sci. U. S. A.* **108**:14944–9.
- 554 38. **Negrini S, Gorgoulis VG, Halazonetis TD.** 2010. Genomic instability--an evolving
555 hallmark of cancer. *Nat. Rev. Mol. Cell Biol.* **11**:220–8.
- 556 39. **Ojanotko-Harri A, Nikkari T, Harri MP, Paunio KU.** 1990. Metabolism of
557 progesterone and testosterone by *Bacillus cereus* strain Socransky 67 and *Streptococcus*
558 *mutans* strain Ingbritt. *Oral Microbiol. Immunol.* **5**:237–9.
- 559 40. **Wiebe JP, Muzia D, Hu J, Szwajcer D, Hill SA, Seachrist JL.** 2000. The 4-pregnene
560 and 5alpha-pregnane progesterone metabolites formed in nontumorous and tumorous
561 breast tissue have opposite effects on breast cell proliferation and adhesion. *Cancer Res.*
562 **60**:936–43.
- 563 41. **Wiebe JP.** 2006. Progesterone metabolites in breast cancer. *Endocr. Relat. Cancer*
564 **13**:717–38.
- 565 42. **van't Veer P, Dekker JM, Lamers JW, Kok FJ, Schouten EG, Brants HA, Sturmans**
566 **F, Hermus RJ.** 1989. Consumption of fermented milk products and breast cancer: a case-
567 control study in The Netherlands. *Cancer Res.* **49**:4020–3.
- 568 43. **Imai K, Matsuyama S, Miyake S, Suga K, Nakachi K.** 2000. Natural cytotoxic activity
569 of peripheral-blood lymphocytes and cancer incidence: an 11-year follow-up study of a
570 general population. *Lancet* **356**:1795–9.
- 571 44. **Strayer DR, Carter WA, Mayberry SD, Pequignot E, Brodsky I.** 1984. Low natural
572 cytotoxicity of peripheral blood mononuclear cells in individuals with high familial
573 incidences of cancer. *Cancer Res.* **44**:370–4.
- 574 45. **Kosaka A, Yan H, Ohashi S, Gotoh Y, Sato A, Tsutsui H, Kaisho T, Toda T, Tsuji**
575 **NM.** 2012. *Lactococcus lactis* subsp. *cremoris* FC triggers IFN- γ production from NK and
576 T cells via IL-12 and IL-18. *Int. Immunopharmacol.* **14**:729–33.
- 577 46. **Carrega P, Bonaccorsi I, Di Carlo E, Morandi B, Paul P, Rizzello V, Cipollone G,**
578 **Navarra G, Mingari MC, Moretta L, Ferlazzo G.** 2014. CD56(bright)perforin(low)
579 noncytotoxic human NK cells are abundant in both healthy and neoplastic solid tissues and
580 recirculate to secondary lymphoid organs via afferent lymph. *J. Immunol.* **192**:3805–15.
- 581 47. **Koller VJ, Marian B, Stidl R, Nersesyan A, Winter H, Simić T, Sontag G,**
582 **Knasmüller S.** 2008. Impact of lactic acid bacteria on oxidative DNA damage in human
583 derived colon cells. *Food Chem. Toxicol.* **46**:1221–1229.
- 584 48. **De Moreno de LeBlanc A, Matar C, Thériault C, Perdigón G.** 2005. Effects of milk
585 fermented by *Lactobacillus helveticus* R389 on immune cells associated to mammary
586 glands in normal and a breast cancer model. *Immunobiology* **210**:349–58.

- 587 49. **Shiou S-R, Yu Y, Guo Y, He S-M, Mziray-Andrew CH, Hoenig J, Sun J, Petrof EO,**
588 **Claud EC.** 2013. Synergistic protection of combined probiotic conditioned media against
589 neonatal necrotizing enterocolitis-like intestinal injury. *PLoS One* **8**:e65108.
- 590 50. **Hosseini E, Grootaert C, Verstraete W, Van de Wiele T.** 2011. Propionate as a health-
591 promoting microbial metabolite in the human gut. *Nutr. Rev.* **69**:245–58.
- 592 51. **Louis P, Hold GL, Flint HJ.** 2014. The gut microbiota, bacterial metabolites and
593 colorectal cancer. *Nat. Rev. Microbiol.* **12**:661–672.
- 594 52. **Aldunate M, Srbinovski D, Hearps AC, Latham CF, Ramsland PA, Gugasyan R,**
595 **Cone RA, Tachedjian G.** 2015. Antimicrobial and immune modulatory effects of lactic
596 acid and short chain fatty acids produced by vaginal microbiota associated with eubiosis
597 and bacterial vaginosis. *Front. Physiol.* **6**:164.

598

599

600 **Figure captions**

601 **Figure 1: Breast tissue microbiota in 43 Canadian women identified by 16S rRNA amplicon**
602 **sequencing.** (A) The relative abundances of bacterial genera in different breast tissue samples
603 were visualized by bar plots. Each bar represents a subject and each colored box a bacterial
604 taxon. The height of a coloured box represents the relative abundance of that organism within the
605 sample. Taxa present in less than 2% abundance in a given sample are displayed in the
606 “Remaining fraction” at the top of the graph (gray boxes). As shown by the bar plots, a variety of
607 bacteria were detected in breast tissue. The legend is read from bottom to top, with the bottom
608 organism on the legend corresponding to the bottom colored box on the bar plot. (B) Box plots of
609 the six phyla identified in breast tissue. The box signifies the 75% (upper) and 25% (lower)
610 quartiles and thus shows where 50% of the samples lie. The black line inside the box represents
611 the median. The whiskers represent the lowest datum still within 1.5 interquartile range (IQR) of
612 the lower quartile and the highest datum still within 1.5 IQR of the upper quartile. Outliers are
613 shown with open circles.

614 **Figure 2: Comparison of bacterial profiles between breast cancer patients and healthy**
615 **controls.** (A) Weighted UniFrac principal coordinate (PCoA) plot and (B) K-means clusterplot of
616 centered log-ratio transformed data. Each breast tissue sample, represented by a coloured point is
617 plotted on a three-dimensional, 3-axis plane representing 79% of the variation observed between
618 all samples (A) or 44.85% of the variation on a 2-axis plane (B). Samples (points) that cluster
619 together are similar in biota composition and abundance. The distinct separation between the two
620 groups indicates that bacterial profiles differ between women with and without cancer. The
621 PERMANOVA test performed on the weighted UniFrac distances showed that the observed
622 differences were statistically significant (10000 permutations; pseudo F-statistic=14.4; p-value
623 <0.01)

624
625
626 **Figure 3: Differences in relative abundances of taxa exist between healthy and cancer**
627 **patients.** The top panel shows the bacteria that had statistically significant higher relative
628 abundances in healthy patients compared to those with cancer (i.e. normal adjacent tissue) and
629 the bottom panel shows the bacteria that had statistically significant higher relative abundances in
630 cancer patients compared to healthy controls. The box signifies the 75% (upper) and 25% (lower)
631 quartiles and thus shows where 50% of the samples lie. The black line inside the box represents
632 the median. The whiskers represent the lowest datum still within 1.5 interquartile range (IQR) of
633 the lower quartile and the highest datum still within 1.5 IQR of the upper quartile. Outliers are
634 shown with open circles. Significance was based on the Benjamini-Hochberg corrected p-value
635 of the Wilcoxon rank test (significance threshold p-val < 0.1).

636

637 **Figure 4: DNA damage ability of *E.coli* isolated from breast cancer patients.** *E.coli* was
638 isolated from normal adjacent tissue of 2 patients with breast cancer and tested for its ability to
639 induce DNA double stranded breaks. *E.coli* (isolates H and E) from subject 41, isolate L from
640 subject 34 and strain IHE3034 were incubated with HeLa cells at MOI 100 for 4hr and then
641 stained for γ H2AX and DAPI. Etoposide, a chemical that induces DNA double stranded breaks in
642 eukaryotic cells, was used as a technical positive control. (A) Representative immunofluorescent
643 images of HeLa cells at 1000x magnification. (B) Image J was used to measure the mean
644 fluorescent intensity (MFI) of γ H2AX positive cells from the digitally acquired images. (C)
645 Percent of total cells stained for γ H2AX calculated from the immunofluorescent images. Data
646 displayed in the bar graphs represent the mean +/- SD of 3 experiments representing a total of 48
647 fields of view and approximately 300 cells for each treatment group. ** denotes p-value <0.01
648
649
650

Fig. 1A

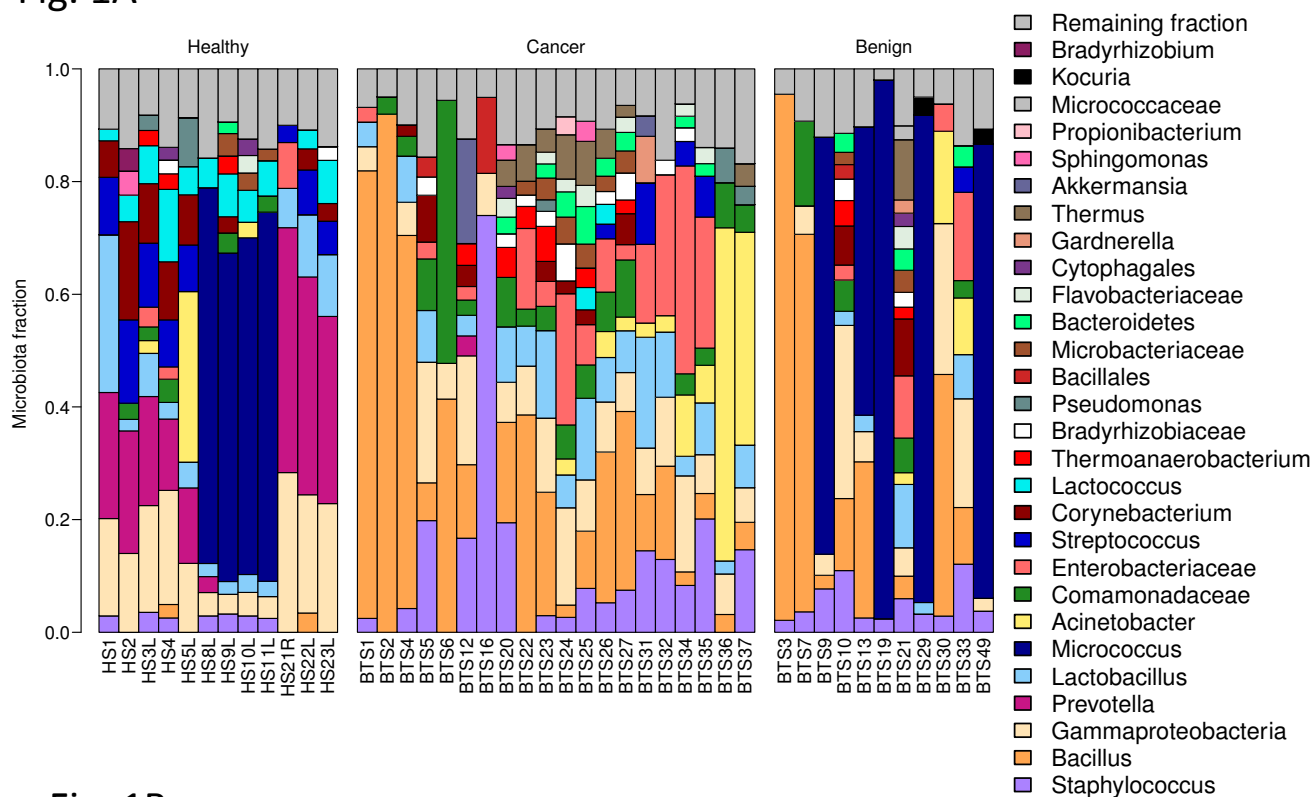


Fig. 1B

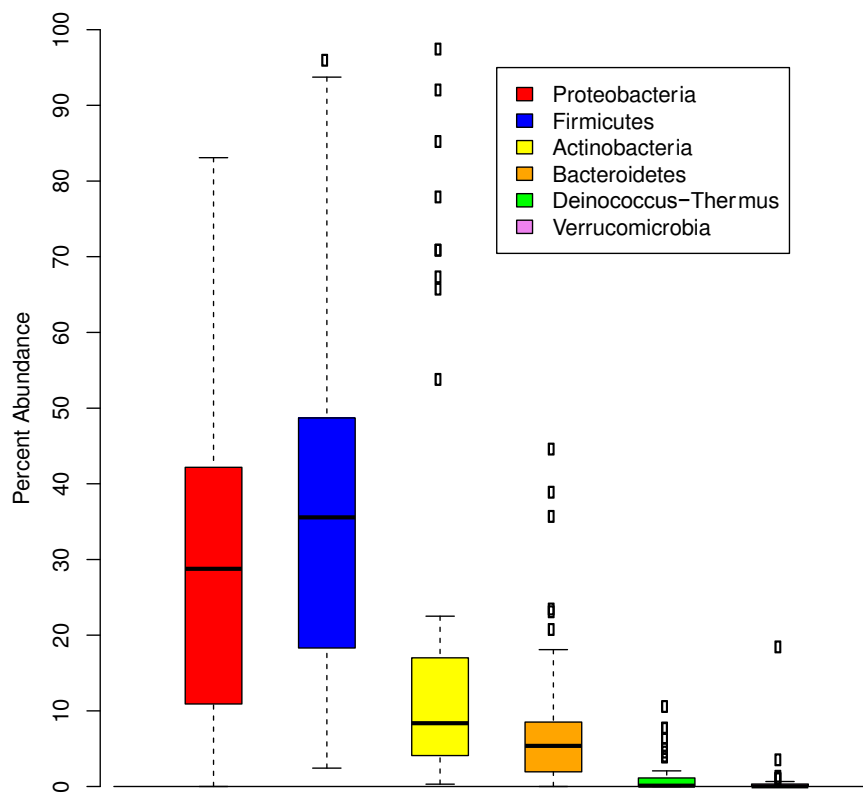


Fig. 2A

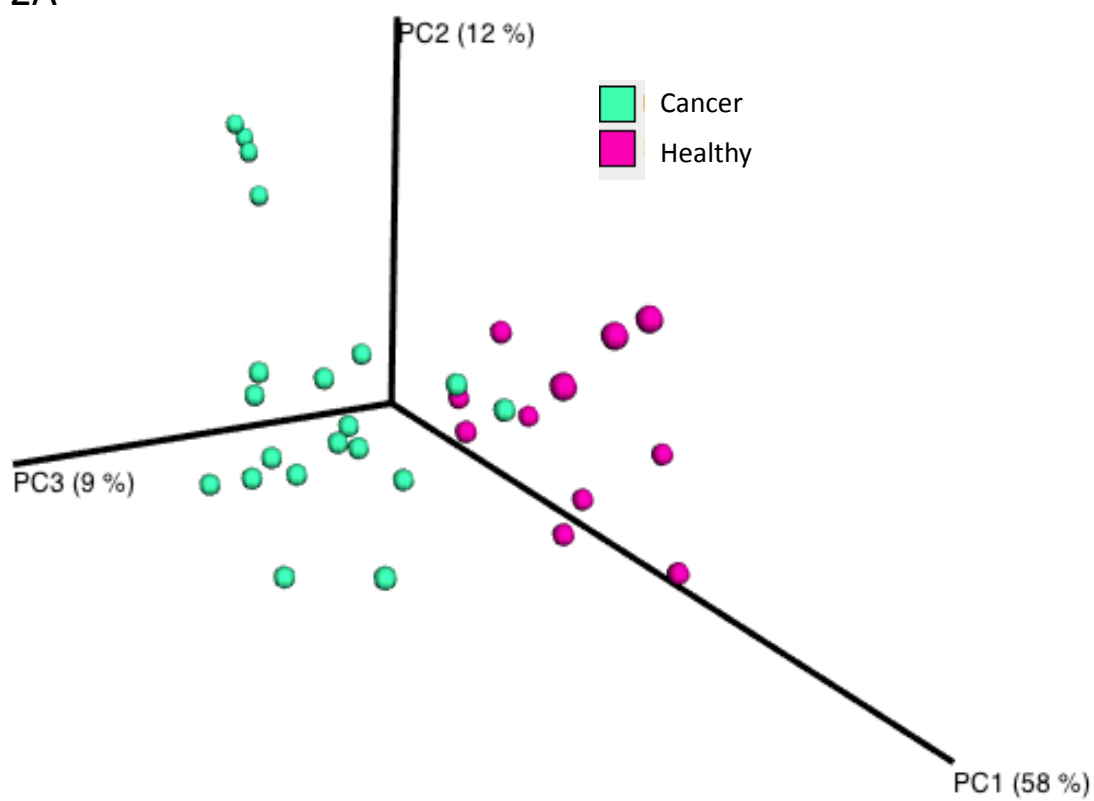
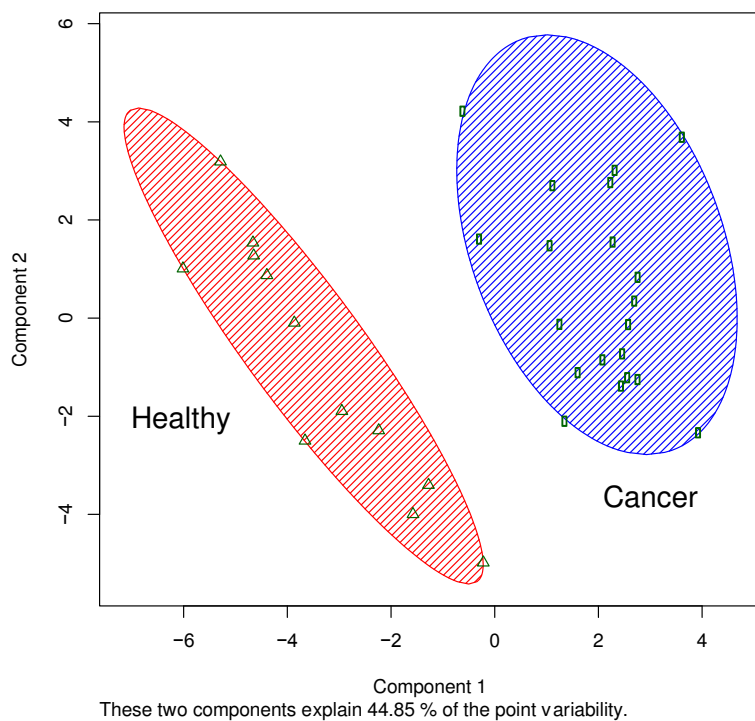


Fig. 2B



These two components explain 44.85 % of the point variability.

Fig. 3

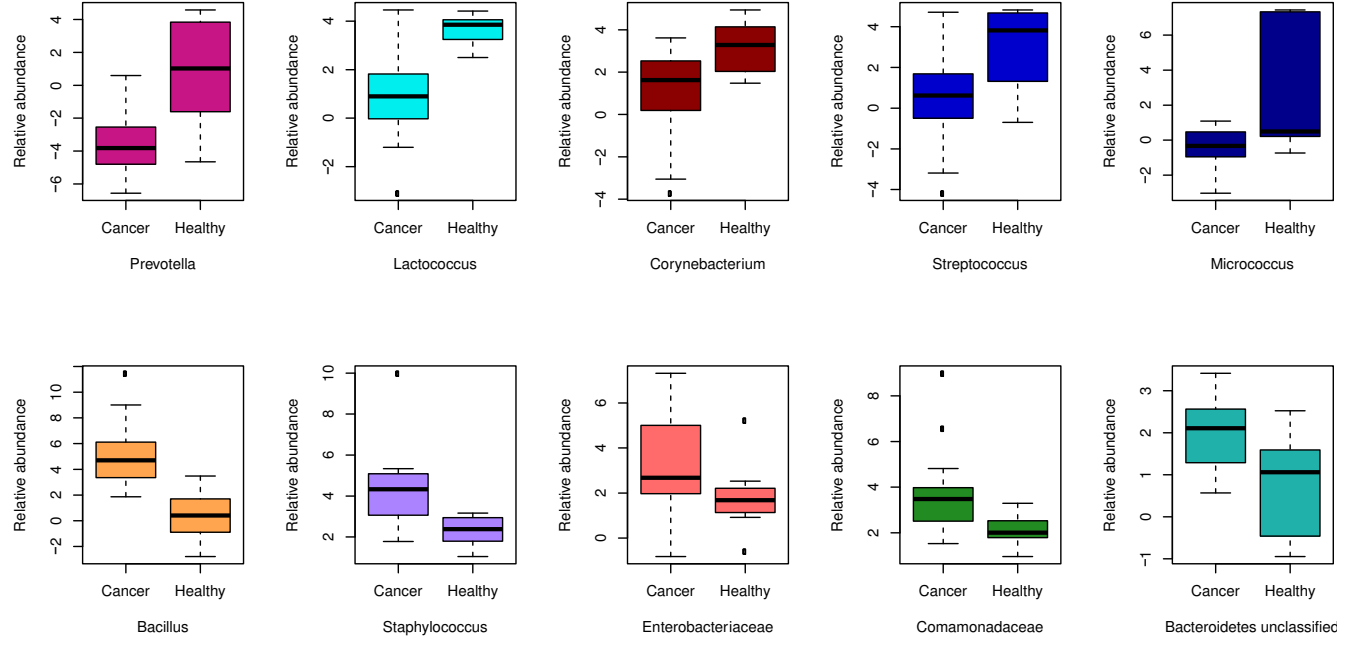


Fig. 4A

Etoposide

E. coli
IHE3034

E. coli
S41H

Untreated

γ H2AX

DAPI

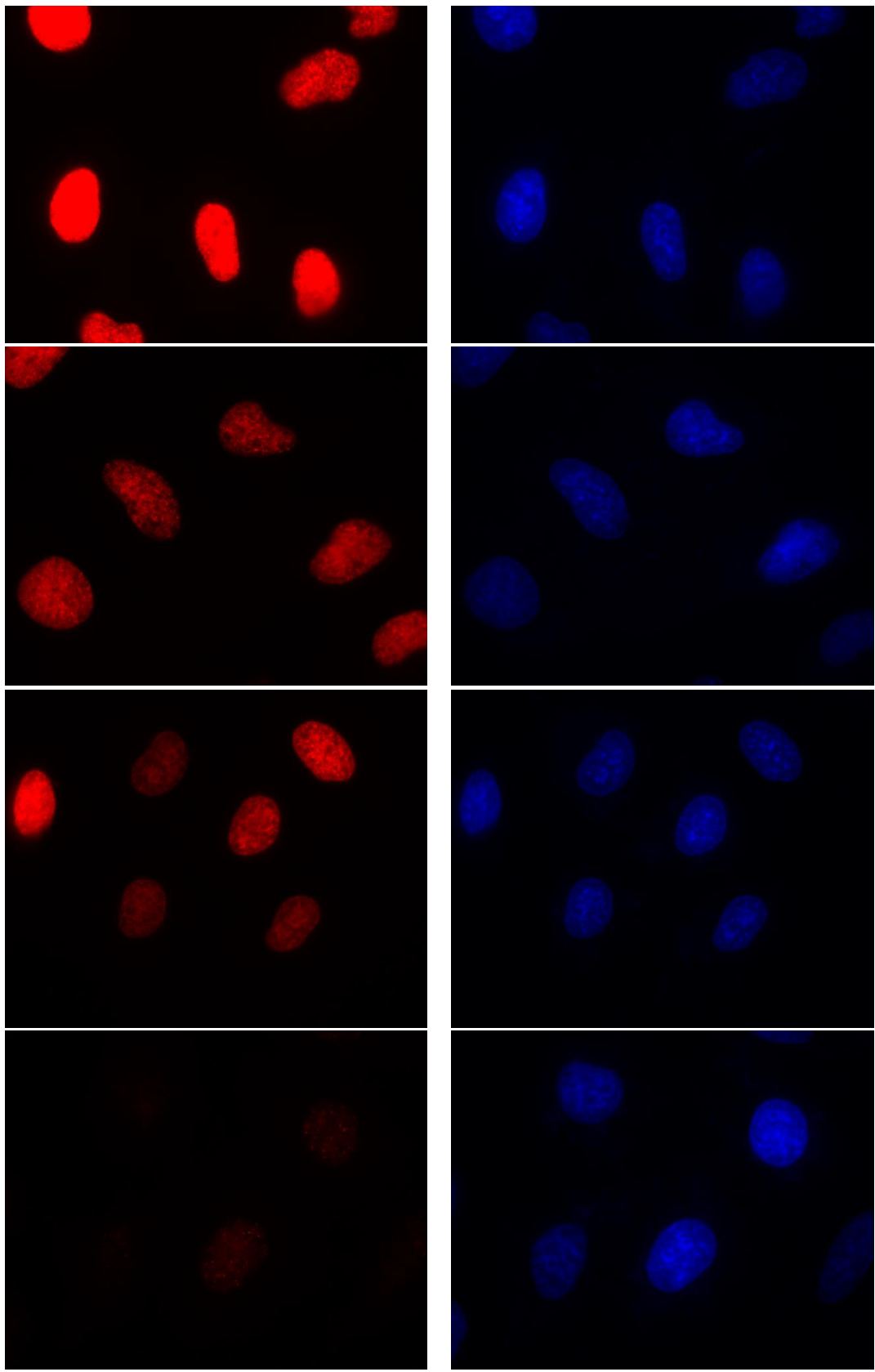


Fig 4B

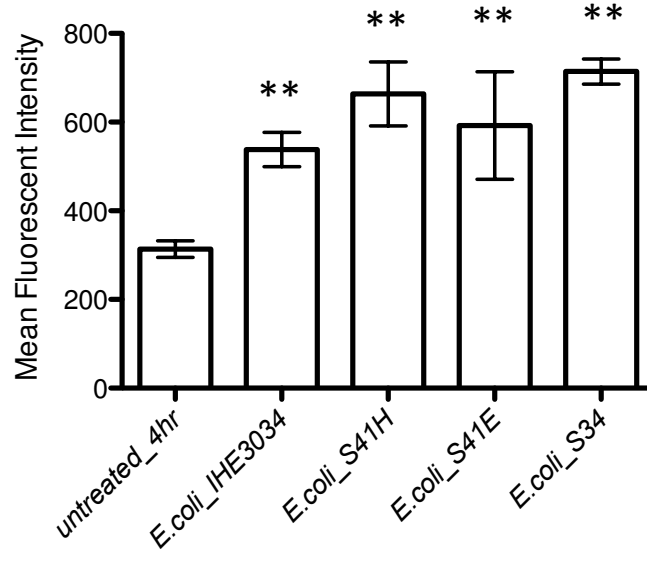


Fig 4C

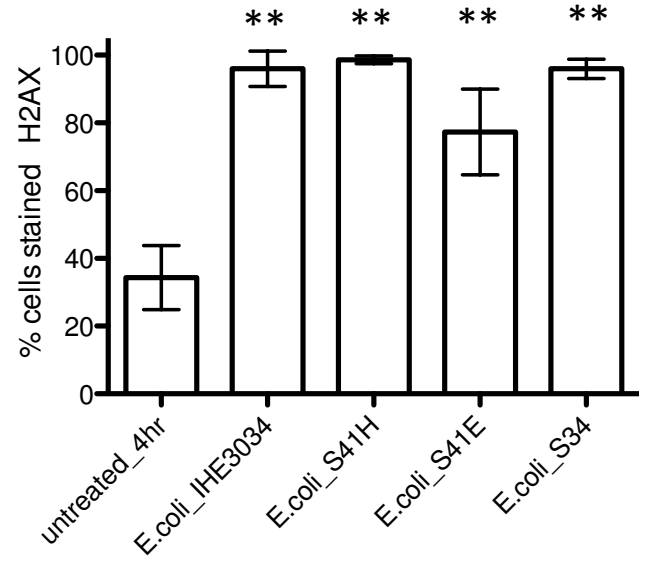


Table 1: Summary of p-values generated by MIRKAT. p-values displayed represent the min, max, average and median values generated from 128 Dirichlet Monte-Carlo instances for each of the 4 distance-based metrics shown (columns 2-5). “Optimal” refers to the p-values obtained when the 4 distance-based metrics are analyzed simultaneously.

	Bray-Curtis	Weighted UniFrac	Unweighted UniFrac	GUniFrac $\alpha=0.5$	Optimal
Min	4.58E-05	2.64E-06	4.12E-06	2.64E-06	0
Max	0.000129826	1.04E-05	0.004191322	1.15E-05	0
Average	7.87E-05	4.52E-06	0.000395004	5.33E-06	0
Median	7.70E-05	4.24E-06	2.05E-04	4.87E-06	0