

1 **Microbiota long-term dynamics and prediction of acute graft-versus-host-**
2 **disease in pediatric allogeneic stem cell transplantation**

3
4
5 **Anna Cäcilia Ingham^{1#}, Katrine Kielsen^{2, 3}, Hanne Mordhorst¹, Marianne Ifversen³, Frank M.**
6 **Aarestrup¹, Klaus Gottlob Müller^{2, 3, 4}, Sünje Johanna Pamp^{1*}**

7
8
9
10
11
12 ¹ Research Group for Genomic Epidemiology, Technical University of Denmark, Kongens Lyngby,
13 Denmark.

14
15 ² Institute for Inflammation Research, Department of Rheumatology and Spine Disease, Copenhagen
16 University Hospital, Rigshospitalet, Copenhagen, Denmark.

17
18 ³ Department of Pediatrics and Adolescent Medicine, Copenhagen University Hospital Rigshospitalet,
19 Copenhagen, Denmark.

20
21 ⁴ Institute of Clinical Medicine, University of Copenhagen, Copenhagen, Denmark.

22
23
24
25
26 # Present address: Department of Bacteria, Parasites and Fungi, Statens Serum Institut, Copenhagen,
27 Denmark.

28
29 *Correspondence: sjpa@dtu.dk
30 Technical University of Denmark, Research Group for Genomic Epidemiology, 2800 Kongens Lyngby,
31 Denmark.

32
33
34 **Keywords**

35 Gut, oral, and nasal microbiota; HSCT; acute GvHD; immune reconstitution; microbiome; antibiotics;
36 amplicon sequence variants; machine learning; prediction; holobiont.

41 **Abstract**

42

43 **Background**

44 Patients undergoing allogeneic hematopoietic stem cell transplantation (HSCT) exhibit changes in their
45 gut microbiota and are experiencing a range of complications, including acute graft-versus-host disease
46 (aGvHD). It is unknown if, when, and under which conditions a re-establishment of microbial and
47 immunological homeostasis occurs. It is also unclear whether microbiota long-term dynamics occur at
48 other body sites than the gut such as the mouth or nose. Moreover, it is not known whether the patients'
49 microbiota prior to HSCT holds clues to whether the patient would suffer from severe complications
50 subsequent to HSCT. Here, we performed integrated host-microbiota analyses of the gut, oral, and nasal
51 microbiotas in 29 children undergoing allo-HSCT.

52

53 **Results**

54 The bacterial diversity decreased in the gut, nose, and mouth during the first month and reconstituted
55 again 1-3 months after allo-HSCT. The microbial community composition traversed three phases over one
56 year. Distinct taxa discriminated the microbiota temporally at all three body sides, including *Enterococcus*
57 spp., *Lactobacillus* spp., and *Blautia* spp. in the gut. Of note, certain microbial taxa appeared already
58 changed in the patients prior to allo-HSCT as compared to healthy children. Acute GvHD occurring after
59 allo-HSCT could be predicted from the microbiota composition at all three body sites prior to HSCT, in
60 particular from *Parabacteroides distasonis*, *Lachnospiraceae* NK4A136 sp. and *Lactobacillus* sp.
61 abundances in the gut. The reconstitution of CD4+ T cells, T_H17 and B cells was associated with distinct
62 taxa of the gut, oral, and nasal microbiota.

63

64 **Conclusions**

65 This study reveals for the first time bacteria in the mouth and nose that may predict aGvHD. Surveillance
66 of the microbiota at different body sites in HSCT may be of prognostic value and could assist in guiding
67 personalized treatment strategies. The identification of distinct bacteria that have a potential to predict
68 post-transplant aGvHD might provide opportunities for an improved preventive clinical management,
69 including a modulation of microbiomes. The host-microbiota associations shared between several body
70 sites might also support an implementation of more feasible oral and nasal swab sampling-based analyses.
71 Altogether, the findings suggest that both, host factors and the microbiota, could provide actionable
72 information to guiding precision medicine.

73 Background

74 In allogeneic hematopoietic stem cell transplantation (allo-HSCT), the infusion of donor derived stem cells
75 is employed as a curative treatment for various types of hematologic and non-hematologic disorders [1].
76 In allo-HSCT patients, the human gut microbiota changes subsequent to transplantation, which may in
77 part be attributable to antimicrobial treatment and conditioning regimens [2–4]. Butyrate-producing
78 bacteria affiliated with the order *Clostridiales* are depleted in the gut early after transplantation, while
79 *Proteobacteria*, and *Lactobacillales* such as *Enterococcus* spp. expand, possibly due to both increased
80 oxygen levels in the intestinal lumen in the absence of butyrate, and antimicrobial resistance [2–5].
81 However, microbiota dynamics in HSCT patients have so far mainly been monitored in detail during the
82 first month post HSCT and not over longer periods of time. Hence, it is unclear whether and when the
83 microbiota re-establishes to similar microbial community structures as prior to HSCT.

84 Conditioning-induced intestinal epithelial permeability might promote bacterial translocation and
85 bacteremia [6]. This is recognized as the initial step in the pathogenesis of acute graft-versus-host disease
86 (aGvHD) [7]. Acute GvHD is a common side effect of allo-HSCT in which alloreactive donor T cells exhibit
87 cytotoxic activity against healthy tissue in the host, including the gut epithelium [7]. Acute GvHD severity
88 can be distinguished in four grades dependent on the extent of organs affected: Grade 0-I presents as no
89 or mild, and grade II-IV as moderate to severe aGvHD. Recently, studies have suggested that a lower gut
90 microbiota diversity is associated with aGvHD and aGvHD-related mortality and that certain bacterial taxa
91 dominating post HSCT may be involved in promoting aGvHD [3,8–12]. However, it has not been examined
92 whether microbiota composition prior to HSCT has a predictive value in forecasting possible aGvHD
93 severity, and which is addressed in the present study.

94 The microbiota exerts immunomodulatory function on the host's adaptive immune system, for example
95 on T cells [13]. For instance, human commensal gut strains affiliated with *Bacteroides* and *Clostridia* can
96 induce T regulatory (T_{reg}) cells in germ-free mice [14]. Recent findings suggest that functionally different
97 T cell subsets, such as T helper 17 (T_H17) and T_{reg} cells are involved in the pathogenesis of aGVHD [15–
98 17]. The microbiota at body sites other than the gut, such as the oral and nasal cavities, have also been
99 suggested to be involved in immunomodulation [18]. We have previously proposed that the gut
100 microbiota is associated with immune cell reconstitution after allo-HSCT [4]. However, it is unknown if the
101 microbiotas at other mucosal sites are affected by allo-HSCT, whether they are associated with aGvHD,
102 and whether they are associated with recovery of the patients' immune system.

103 Here, we monitored the microbiota dynamics in the gut, oral, and nasal cavities in pediatric allogeneic
104 HSCT patients over a period of one year. At all three body sites, we identify distinct temporal bacterial
105 abundance trajectories. In a machine learning approach, we predict aGvHD severity from pre-transplant
106 microbiotas in the gut, oral, and nasal cavities which may be useful for early preventive managements in
107 the clinical setting. By relating the microbiota composition to immune cell counts, inflammation and
108 infection markers, antibiotic treatment, clinical outcomes, and patients' baseline parameters, we uncover
109 similarities in host-microbial associations at different body sites.

110

111 Results

112 We characterized long-term microbiota dynamics in pediatric allo-HSCT at three body sites: the gut, and
113 oral and nasal cavities (Figure 1). Fecal samples, buccal swabs, and anterior naris swabs were collected

114 from 29 children at 10 time points over a one-year period: Twice prior to HSCT, on the day of HSCT, weekly
115 during the first month after HSCT, and at three follow-up time points up to twelve months post HSCT
116 (Figure 1). Microbial community dynamics in these samples were determined by 16S rRNA gene profiling.
117 A total of 709 patient samples (212 fecal samples, 248 oral swabs, and 249 nasal swabs from 10 time
118 points) were characterized. Upon sequence filtering (see Methods), we retained 2465 ASVs for the fecal,
119 377 ASVs for the oral, and 197 ASVs for the nasal core microbiota sets. We predicted the development of
120 aGvHD severity from pre-transplant gut, oral, and nasal microbial abundances using machine learning. In
121 addition, we assessed multivariate associations between the microbiota at the different body sites and
122 immune reconstitution, immune markers, and clinical outcomes. Immune reconstitution was determined
123 through quantitative measurements of T, B, and NK cells, and other leukocyte subpopulations in
124 peripheral blood (Figure 1). We assessed systemic inflammation through levels of C-reactive protein (CRP),
125 and measured procalcitonin as an approximation of infection (Figure 1, see Methods).

126

127 **Patient cohort and outcomes**

128 The 29 children had a median age of 8.2 years (range: 2.5-16.4) at the time of HSCT. Nine patients (31%)
129 had no or mild aGvHD (grade 0 or I) and 20 patients (69%) developed moderate to severe aGvHD (grade
130 II-IV) at median +14 days following HSCT (range: day +9 to day +61) (Supplementary Table S1, Additional
131 file 1; and <https://doi.org/10.6084/m9.figshare.13567502>). The main organs involved in aGvHD included
132 the skin (all), intestinal tract (n=3), and the liver (n=2). During the follow-up period of 21.4 months on
133 average (range: 10.1 – 32.7 months), two patients (7%) relapsed and one patient underwent a donor
134 lymphocyte infusion. Three patients (10%) died (one relapse-related death at day +91 and two treatment-
135 related deaths at days +111 and +241, respectively). Due to their low incidence, we did not focus our
136 analysis on relapse and mortality. For 25 patients (86.2%) ≥ 1 bacterial infection indicated by positive
137 microbial culture was reported throughout the monitored period. All patients were treated
138 prophylactically with trimethoprim and sulfamethoxazole prior to HSCT. In cases of fever or clinical signs
139 of infections, antibiotic treatment with meropenem (28 patients), vancomycin (24 patients), ciprofloxacin
140 (20 patients), phenoxymethylpenicillin (14 patients), or other antibiotics was commenced according to
141 culture-based results or clinical presentation.

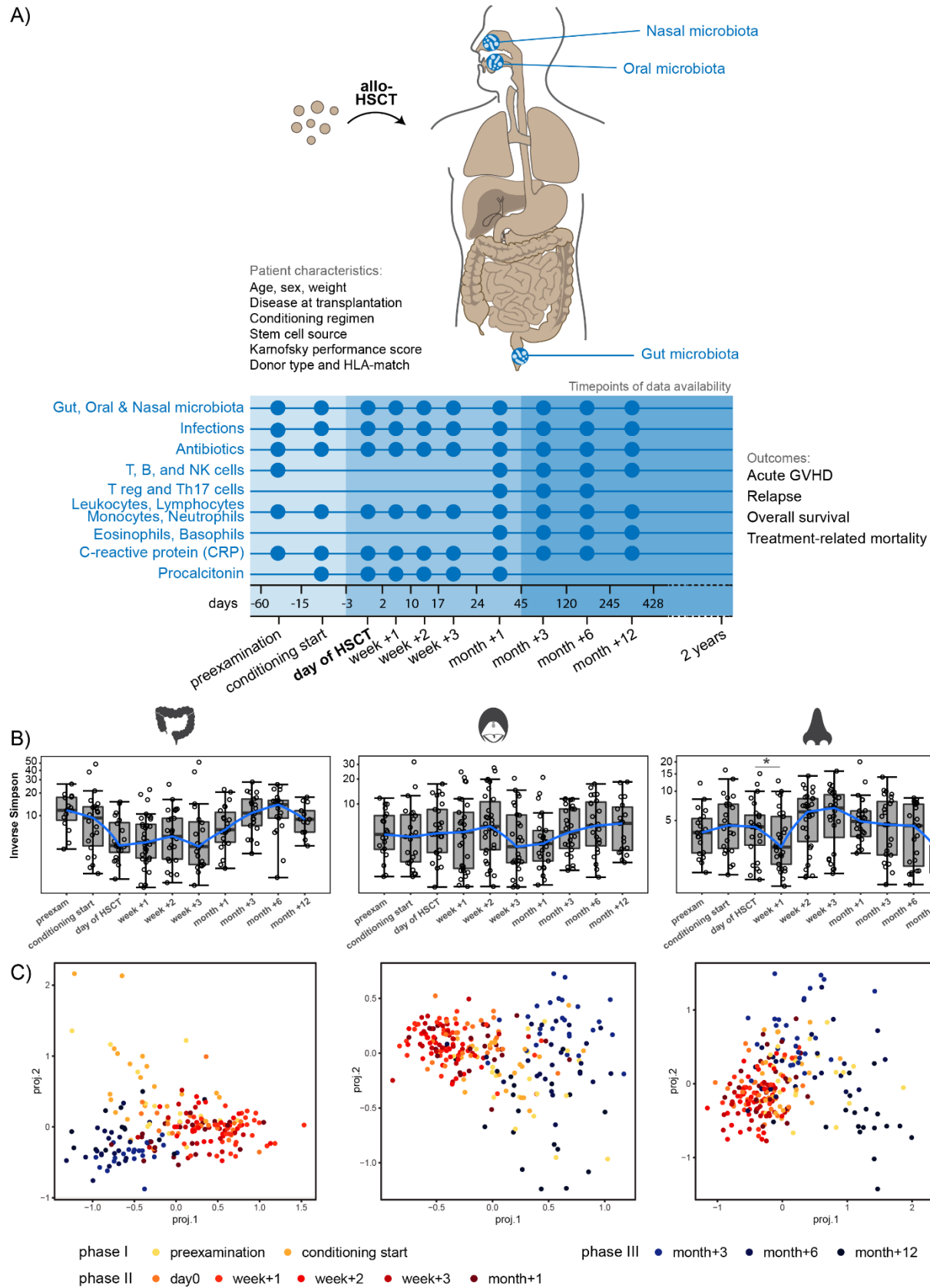
142

143 **Bacterial alpha diversity decreases in relation to allo-HSCT at all three body sites**

144 Alpha diversity (Inverse Simpson) in the gut was overall the highest, followed by the oral cavity, and the
145 nose (Figure 1B). The lowest alpha diversity was observed within the first month post HSCT for all three
146 body sites. However, the exact time points were somewhat different for each body site: the day of HSCT
147 to week +3 for the gut, week +3 for the oral cavity, and week +1 for the nasal cavity. The decrease in
148 microbial diversity was significant for the nasal cavity, where the median alpha diversity decreased from
149 4.43 at the start of conditioning to 2.65 in week +1 ($P = 0.02$) (Figure 1B). Alpha diversity increased again
150 at all body sites thereafter. However, alpha diversity was lower again at month +12 in the nasal cavity.

151

152



154 **Figure 1. Monitoring gut, oral, and nasal microbiota and the host immune system in allogeneic hematopoietic**
155 **stem cell transplantation (HSCT).** A) Twenty-nine children were monitored before, at the time of, and immediately
156 post allogeneic HSCT, as well as at late follow-up time points. Patients' baseline characteristics, clinical outcomes, as
157 well as immune cell counts, and inflammation and infection markers over time were monitored. Patient
158 characteristics are described in detail in Table S1 (Additional File 1). Host immune system parameters were related
159 to longitudinal dynamics of the gut, oral, and nasal microbiota that was assessed at the denoted time points. B)
160 Bacterial alpha diversity before, at the time of, and after HSCT at each body site, displayed on a log₁₀ transformed
161 y-axis for visualization purposes. Asterisks indicate significant differences in median inverse Simpson index between
162 time points * $P < 0.05$. C) Tree-based sparse linear discriminant (LDA) analyses by time point in relation to HSCT. For
163 fecal samples, the positive LDA scores were observed for samples collected immediately post HSCT. For both oral
164 and nasal samples, the positive LDA scores were observed for samples from before HSCT and from late follow up-
165 time points.
166

167 **Microbial community composition in patients prior to HSCT differs from healthy controls**

168 We hypothesized that the bacterial alpha diversity at the first sampling time point (preexamination) might
169 already be lower in these patients as compared to age-matched healthy children due to the treatment
170 given prior to the referral to allo-HSCT and enrolment in this study. To assess this, we compared the gut
171 microbiota at preexamination to that of healthy children (median age 6.8 years) [19]. As expected, the
172 alpha diversity was 2.4-fold lower in the patients at preexamination (median InvSimpson 11.7) as
173 compared to the healthy children (median InvSimpson 28.2) (Supplementary Figure S1A, Additional File
174 2). Bacterial composition differed between the two groups (anosim, $p=0.001$, $R=0.44$, Figure S1B). This
175 difference was to a certain extent due to a larger variation within the HSCT group (betadisper, $p<0.001$)
176 (Supplementary Figure S1 B, Additional File 2). Through linear discriminant analysis (LEfSe) and differential
177 abundance analysis (DeSeq2), we found taxa that were significantly more abundant in the patients already
178 at preexamination as compared to the healthy controls: these included *Bacilli* (e.g. *Lactobacillus*,
179 *Enterococcus*), *Erysipelotrichaceae*, and *Enterobacteriaceae* (e.g. *Klebsiella*). In contrast, certain taxa were
180 more abundant in the healthy children, such as *Prevotella*, *Ruminococcaceae* (e.g. *Ruminococcus*), and
181 *Akkermansia*, as compared to the patients at preexamination (Supplementary Figure S1 C and D,
182 Additional File 2; and <https://doi.org/10.6084/m9.figshare.13614230>).
183

184 **Temporal microbial community dynamics appear in three interlaced phases over one year**

185 For a more detailed assessment of gut, oral, and nasal ASVs that best characterized samples from different
186 time points, we performed tree-based sparse linear discriminant analyses (LDA). We observed at all three
187 body sites that samples divided into three partly interlaced phases; phase I: samples at pre-examination
188 and conditioning start, phase II: day of HSCT to month +1, and phase III: month +3 to month +12 (Figure
189 1C). Interestingly, samples from phase I and III overlapped for the oral and nasal cavities, suggesting a
190 possible return of microbial communities from later time points to a state similar to before HSCT. Of note,
191 the nasal community composition at month +12, that exhibited low alpha diversity, was different from
192 samples of week +1 (phase II) that also exhibited low alpha diversity (Figure 1B and C).

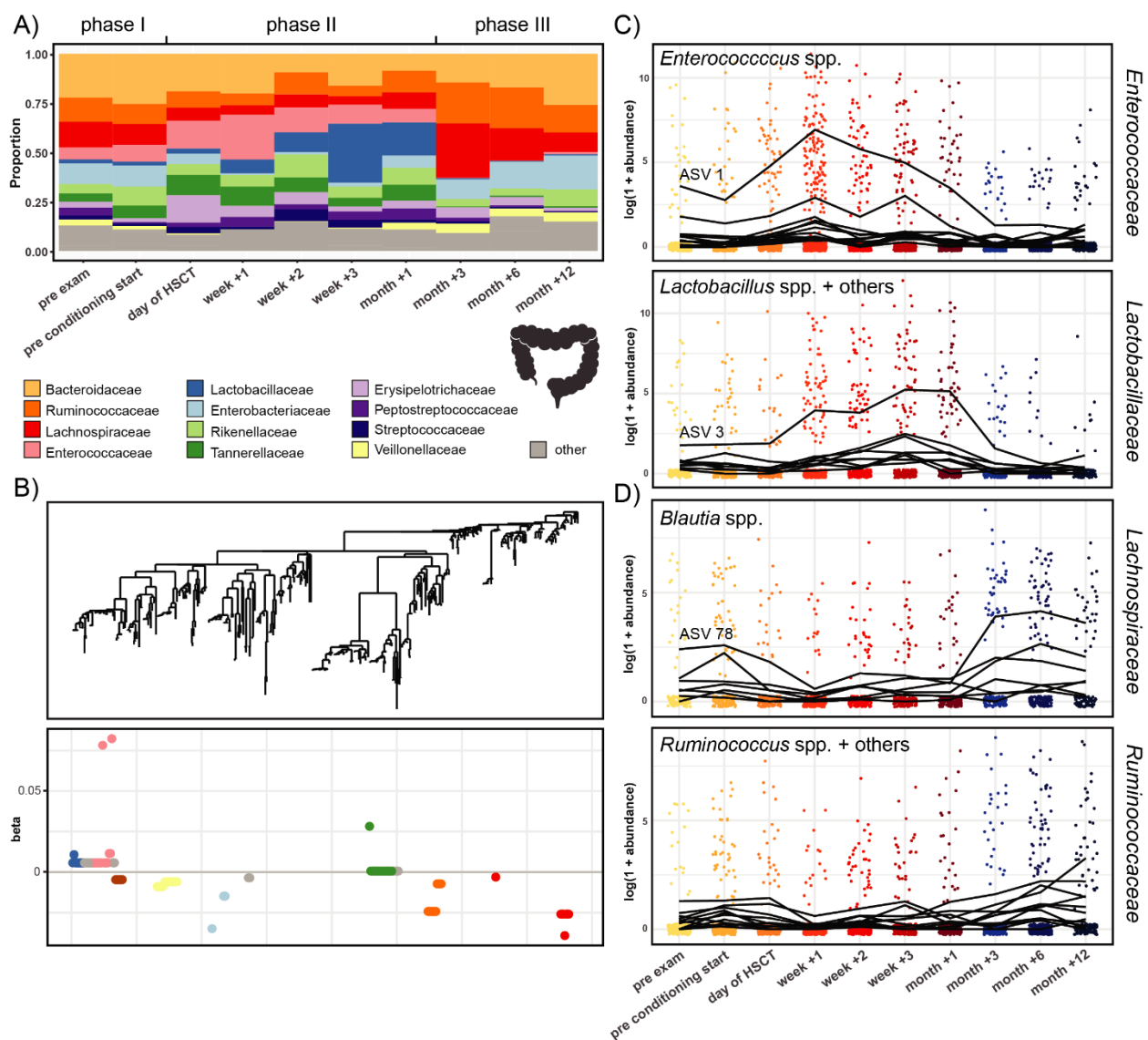
193 To get a more detailed view of the microbial abundance dynamics, we examined the 12 most abundant
194 families at each body site, respectively (Figures 2A, 3A, Additional File 2: Figures S2 and S3A). In the gut,
195 we observed a reduction in *Lachnospiraceae* in phase II, immediately after HSCT, from 13% at pre-
196 examination to 4.7% in week +1, followed by a recovery to 27.5% in month +3 at the start of phase III
197 (Figure 2A). Concurrently, an expansion of *Enterococcaceae* in phase II (pre-examination: 6.1%; week +1:

198 22.8%) and *Lactobacillaceae* in phase II (pre-examination: 2%; week +1: 7%) occurred, followed by a
 199 reduction in phase III from month +3 onwards to 0.2% and 0.6%, respectively (Figure 2A).

200 In the oral cavity, we observed a reduced relative abundance of *Actinomycetaceae* for several time points
 201 in phase II as compared to the time points in phase I (prior to HSCT) and at later follow-up time points.
 202 For example, *Actinomycetaceae* abundances were 9.7% at pre-examination and 2.9% in week +3 (Figure
 203 3A). Furthermore, *Streptococcaceae* abundances were lower from the day of HSCT until week +2
 204 compared with before HSCT and late follow-up time points (pre-examination: 44.6%; week +1: 23.3%;
 205 month +3: 51.3%, Figure 3A).

206 In the nasal cavity, we observed a reduced relative abundance of *Corynebacteriaceae* and *Moraxellaceae*
 207 at most time points in phase II, as compared to samples from phase I and III (Additional File 2: Figure S3).
 208 For example, *Corynebacteriaceae* abundances were 28.7% at pre-examination and 0.7% in week +1
 209 (Additional File 2: Figure S3).

210
 211



212

213 **Figure 2. Temporal microbial community dynamics in the gut.** A) Relative abundances over time of the 12 most
214 abundant families in the gut. B) Tree-based sparse linear discriminant analysis (LDA). Coefficients of discriminating
215 clades of ASVs on the first LDA axis, colored by taxonomic family, and plotted along the phylogenetic tree. C)
216 Trajectories of ASVs affiliated with the families *Enterococcaceae* and *Lactobacillaceae*, with increasing abundances
217 after HSCT. The most abundant discriminating ASV for each family is indicated. D) Trajectories of ASVs affiliated with
218 the families *Lachnospiraceae* and *Ruminococcaceae*, with decreasing abundances after HSCT and recovery at late
219 follow-up time points. The most abundant discriminating ASV for *Blautia* spp. is indicated. Detailed taxonomic
220 information and LDA-coefficients of the displayed ASVs are listed in Additional File 1: Table S2.

221

222 **Distinct *Enterococcus*, *Lactobacillus*, and *Blautia* lineages discriminate the gut microbiota temporally**

223 In order to determine which specific taxa in the gut were driving the differences between samples in the
224 LDA (Figure 1C), we examined the individual discriminating ASVs. In general, in tree-based sparse LDA,
225 ASVs with positive LDA coefficients are overrepresented in samples with positive LDA scores, while ASVs
226 with negative LDA coefficients likewise are associated with samples with negative LDA scores (Figures 1C,
227 2B, 2C, and 2D). The LDA revealed 19 clades (total 102 ASVs) in the gut that best separated samples by
228 time point (Figure 2B). The two most discriminating clades with positive LDA-coefficients comprised ASVs
229 of the family *Enterococcaceae* and *Lactobacillaceae* (Figure 2B). The ASVs of these two clades increased
230 in abundance from the day of HSCT (*Enterococcaceae*) and week +1 (*Lactobacillaceae*), respectively, in
231 support of the family abundances and in line with the positive LDA scores of phase II samples (Figures 2A,
232 2C, and 1C). Of note, the order *Lactobacillales* and genus *Lactobacillus* (family *Lactobacillaceae*) appeared
233 already to be higher at pre-examination as compared to healthy children (Supplementary Figure S1D,
234 Additional File 2). From month +3 onwards, their abundances decreased again to levels comparable to the
235 time of pre-examination (i.e. pre-treatment) (Figure 2C). All members of the *Enterococcaceae* clade, with
236 the exception of one ASV, were *Enterococcus* spp. (Additional File 1: Table S2). The most abundant and
237 most frequently observed *Enterococcus* was ASV 1 (Figure 2C and Additional File 1: Table S2). More
238 detailed sequence analysis of the partial 16S rRNA gene sequence using SINA and BLAST alignments
239 revealed that it belonged to the *Enterococcus faecium* group. The most abundant and most frequently
240 observed *Lactococcus* was ASV 3 (Figure 2C and Additional File 1: Table S2), and its partial 16S rRNA gene
241 sequence exhibited a high sequence similarity to *Lactobacillus rhamnosus*.

242 The two most discriminative clades with negative LDA-coefficients included two individual ASVs and one
243 clade of the *Lachnospiraceae* family, and two *Ruminococcaceae* clades (Figure 2B, Additional File 1: Table
244 S2). The abundances of these ASVs decreased in week +1 and recovered from month +3 onwards,
245 returning to abundances comparable with before HSCT or higher (Figure 2D), in agreement with the
246 abundance patterns for those families (Figure 2A). Of note, the family *Ruminococcaceae* appears already
247 to be lower at pre-examination as compared to healthy children (Supplementary Figure S1 C and D,
248 Additional File 2). All ASVs within the *Lachnospiraceae* group belonged to the genus *Blautia* (Additional
249 File 1: Table S2). The most abundant and most frequently observed *Blautia* was ASV 78 (Figure 2D and
250 Additional File 1: Table S2), and its partial 16S rRNA gene sequence exhibited a high sequence similarity
251 to *Blautia wexlerae*.

252

253 **Distinct *Actinomyces* and *Streptococcus* lineages discriminate the oral microbiota temporally**

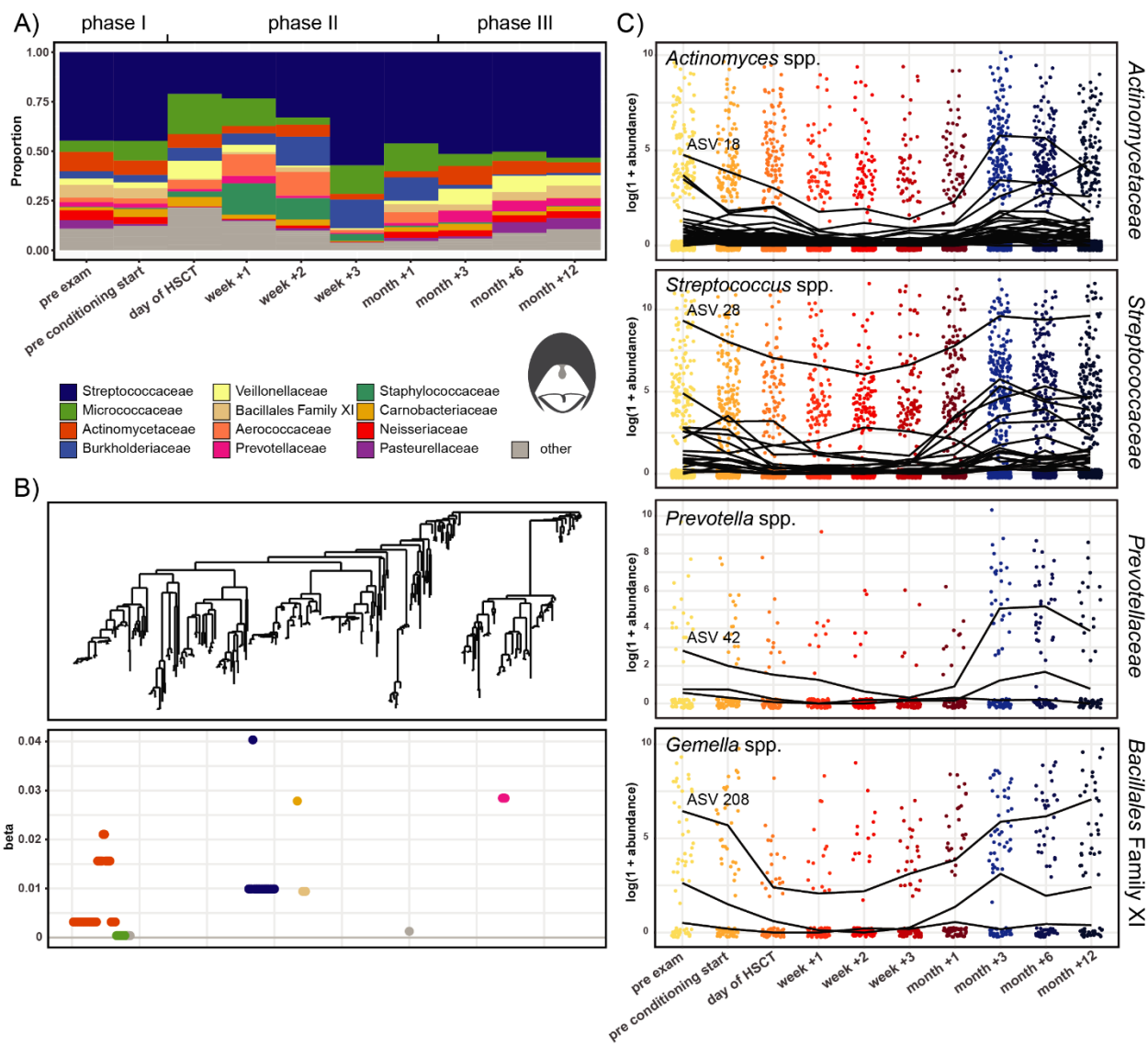
254 The tree-based sparse LDA identified 10 clades of in total 71 ASVs in the oral cavity that best separated
255 samples by time points along the first axis (Figure 3B). The two largest discriminating groups of ASVs were

256 affiliated with *Actinomycetaceae* and *Streptococcaceae* (Figure 3B, Additional File 1: Table S2). The most
257 abundant and among the most frequently observed ASVs were *Actinomyces* ASV 18 and *Streptococcus*
258 ASV 28 (Figure 3C and Additional File 1: Table S2), and their partial 16S rRNA gene sequence exhibited a
259 high sequence similarity to the *Actinomyces viscosus* and *Streptococcus mitis* groups, respectively.
260 Additional discriminating ASVs were affiliated with *Prevotellaceae*, and *Bacillales* Family XI (*Gemella* spp.),
261 respectively. The most abundant and frequently observed ASVs were affiliated with *Prevotella*
262 *melaninogenica* (ASV 42) and *Gemella sanguis* (ASV 208). In agreement with the relative family abundance
263 dynamics, these clades shared a pattern of depletion from the day of HSCT or week +1 onwards (phase
264 II), until their abundances recovered from month +3 onwards (phase III) (Figures 3A and 3C) to an
265 abundance similar to before HSCT, as observed for *Ruminococcaceae* and *Lachnospiraceae* in the gut.

266

267 **Distinct *Corynebacteriaceae* and *Streptococcaceae* lineages discriminate the nasal microbiota** 268 **temporally**

269 The LDA revealed 30 discriminating nasal clades on axis 1 (comprising in total 36 ASVs), many of which
270 consisted of individual ASVs (Additional File 2: Figure S3B). ASVs affiliated with the same family did not
271 always covary in abundance. The *Corynebacteriaceae*, *Streptococcaceae*, and *Moraxellaceae* ASVs all had
272 positive LDA-coefficients, i.e. their abundances decreased after HSCT and increased again from month+3
273 onwards (Additional File 2: Figures S3B and S3C). The most abundant and most frequently observed
274 *Corynebacteriaceae* was ASV 14 (Additional File 2: Figure S3C and Additional File 1: Table S2), and its
275 partial 16S rRNA gene sequence exhibited a high sequence similarity to *Corynebacterium propinquum*.



276
 277 **Figure 3. Temporal microbial community dynamics in the oral cavity.** A) Relative abundances over time of the 12
 278 most abundant families in the oral cavity. B) Tree-based sparse linear discriminant analysis (LDA). Coefficients of
 279 discriminating clades of ASVs on the first LDA axis, colored by taxonomic family, and plotted along the phylogenetic
 280 tree. C) Trajectories of ASVs affiliated with the families *Actinomycetaceae*, *Streptococcaceae*, *Prevotellaceae*, and
 281 Family XI (Class *Bacillales*), with decreasing abundances after HSCT and recovery at late follow-up time points. The
 282 most abundant discriminating ASV for each family is indicated. Detailed taxonomic information and LDA-coefficients
 283 of the displayed ASVs are listed in Additional File 1: Table S2.

284
 285

286 Acute GvHD severity can be predicted from gut microbiota composition prior to HSCT

287 To reveal potential associations between the gut microbiota and the severity of acute GvHD, we examined
 288 the 12 most abundant families at each body site in patients with no or mild (grade 0-I) and moderate to
 289 severe (grade II-IV) aGvHD. In the gut, *Tannerellaceae* were less abundant at time points before HSCT in
 290 patients with grade 0-I compared to grade II-IV, especially at pre-examination and at start of conditioning
 291 (Figure 4A). In order to predict aGvHD (grade 0-I versus grade II-IV) from microbial abundances at time
 292 points up until the time of stem cell infusion, we implemented machine-learning models (see Methods –

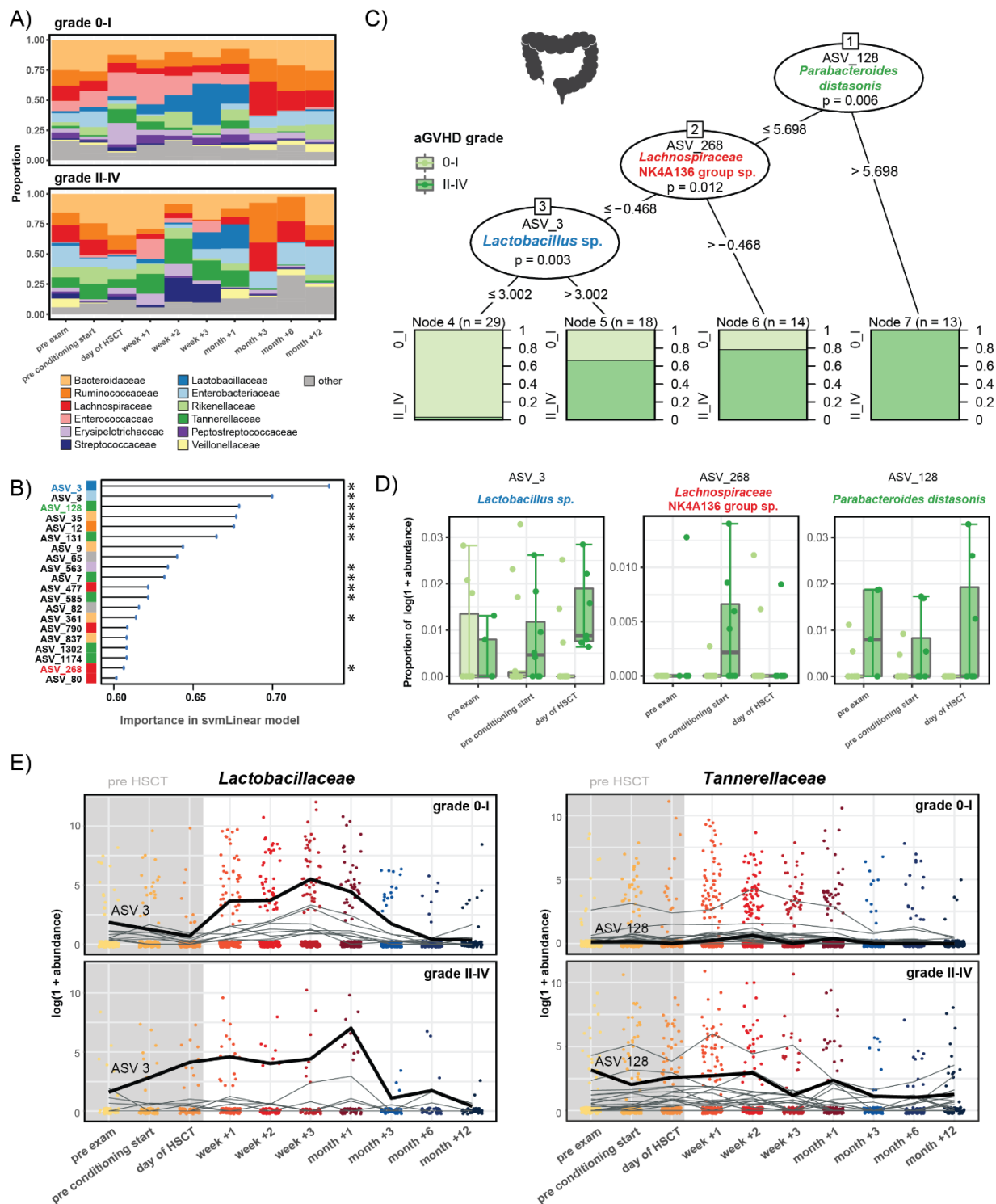
293 Statistical analysis). This analysis revealed 3 significant predictive ASVs in the gut: ASV 128 (*Parabacteroides*
294 *distasonis*, Tannerellaceae, $P < 0.01$), ASV 268 (*Lachnospiraceae* NK4A136 group sp., Lachnospiraceae, P
295 = 0.01) and ASV 3 (*Lactobacillus* sp., Lactobacillaceae, $P < 0.01$) (Figures 4B and 4C, and Additional File 1:
296 Table S3). This means, high abundances of these ASVs before HSCT were associated with the subsequent
297 development of aGvHD grade II-IV post HSCT (Figure 4C). For instance, all pre-transplant samples with a
298 variance stabilized abundance >5.7 of ASV 128 (*Parabacteroides distasonis*) and 67% with a variance
299 stabilized abundance >3 of ASV 3 (*Lactobacillus* sp.) originated from patients who later developed aGvHD
300 grade II-IV (Figure 4C). In agreement, log transformed relative abundances of these ASVs were mostly
301 higher at pre-examination, conditioning start, and the day of HSCT in patients who later developed aGvHD
302 grade II-IV compared with those exhibiting grade 0-I (Figure 4D). For instance, the average abundance of
303 ASV 128 (*Parabacteroides distasonis*) was 5.5 times higher at pre-examination in grade II-IV versus in
304 grade 0-I patients (Figure 4D). The temporal trajectory of ASV 3 (*Lactobacillus* sp.) also revealed a higher
305 abundance at time points up to the transplantation in patients with grade II-IV aGvHD compared to those
306 with grade 0-I (Figure 4E). Within the *Lactobacillaceae* identified by the LDA, this pattern seemed to be
307 restricted to ASV3 (Figure 4E). ASV 128 (*Parabacteroides distasonis*) was part of the discriminating group
308 of *Tannerellaceae* identified in the LDA (Figure 4E, and Additional File 1: Table S3). Its trajectory faceted
309 by aGvHD severity confirmed the observation of increased pre-HSCT abundances in patients with
310 subsequent development of aGvHD grade II-IV (Figure 4E).

311

312 **Acute GvHD severity can be predicted from oral microbiota composition prior to HSCT**

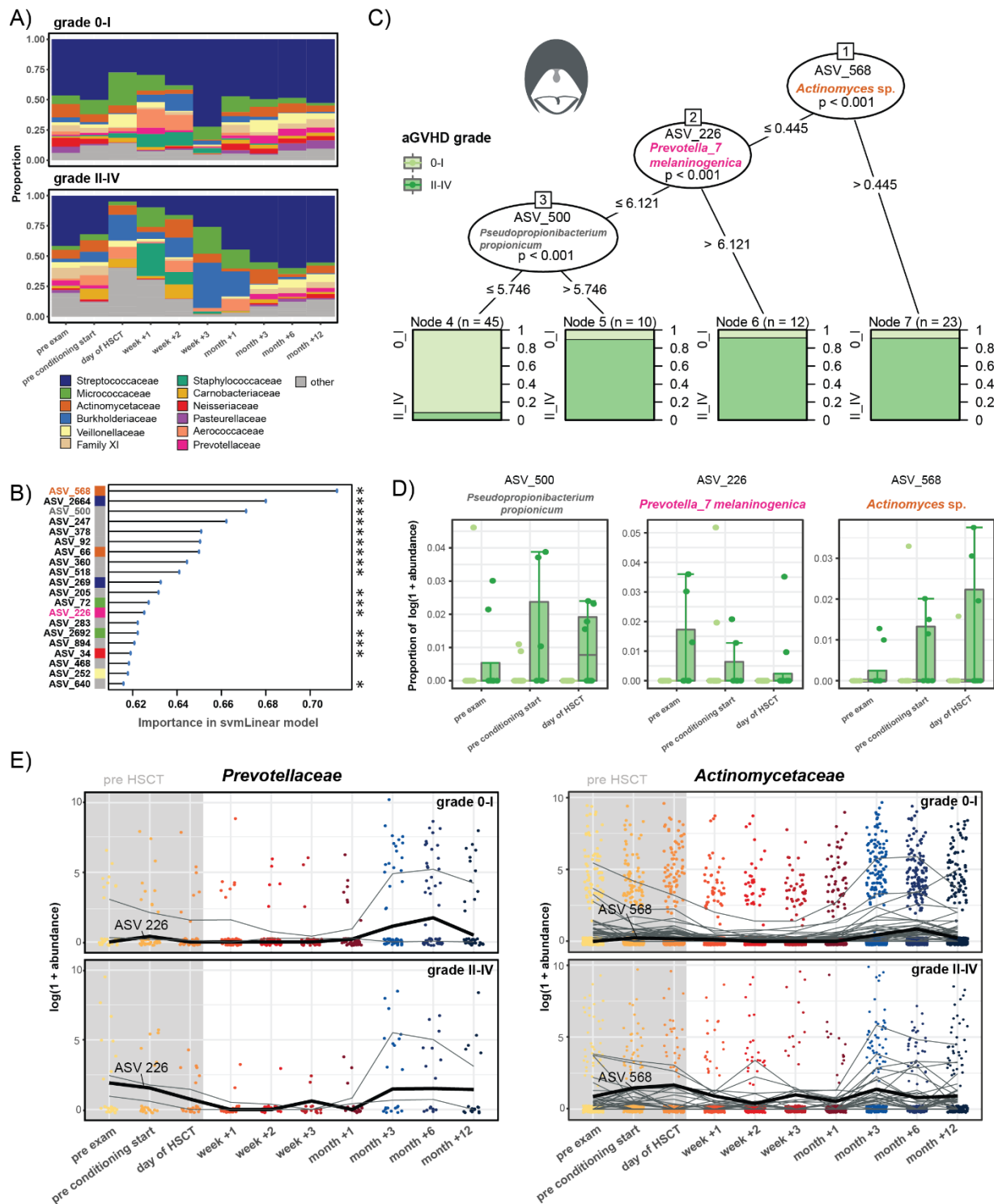
313 In the oral cavity, the bacterial community before HSCT in patients with grade II-IV aGvHD was
314 characterized by a lower relative abundance of *Neisseriaceae*, and higher relative abundances of
315 *Aerococcaceae* and *Prevotellaceae*, compared with grade 0-I aGvHD, especially at pre-examination and
316 conditioning start (Figure 5A). Our machine learning approach predicted aGvHD severity (grade 0-I versus
317 II-IV) from the abundances of 3 significant oral ASVs pre-HSCT: ASV 568 (*Actinomyces* sp.,
318 Actinomycetaceae, $P < 0.001$), ASV 226 (*Prevotella melaninogenica*, Prevotellaceae, $P < 0.001$) and ASV
319 500 (*Pseudopropionibacterium propionicum*, Propionibacteriaceae, $P < 0.001$) (Figures 5B and 5C, and
320 Additional File 1: Table S3). High abundances of these ASVs before transplantation predicted the
321 development of aGvHD grade II-IV after HSCT (Figure 5C). For instance, 91% of samples with a variance
322 stabilized abundance >0.4 of ASV 568 (*Actinomyces* sp.) and 92% of samples with a variance stabilized
323 abundance >6.1 of ASV 226 (*Prevotella melaninogenica*) originated from patients with subsequent
324 development of aGvHD grade II-IV (Figure 5C). In support, pre-HSCT log transformed relative abundances
325 of these ASVs were higher in those patients. For example, the median relative abundance of ASV 500
326 (*Pseudopropionibacterium propionicum*) on the day of HSCT was 10 times higher in grade II-IV versus in
327 grade 0-I patients (Figure 5D). Temporal trajectories of oral *Actinomycetaceae* and *Prevotellaceae*,
328 identified also in the LDA, showed that the abundances of ASV 226 (*Prevotella melaninogenica*) and ASV
329 568 (*Actinomyces* sp.) were higher at time points up to the transplantation in patients with grade II-IV
330 versus those with grade 0-I (Figure 5E).

331



332
 333 **Figure 4. Machine learning-based prediction of aGVHD severity from the pre-HSCT gut microbiota composition.** A)
 334 Relative abundances of the 12 most abundant families over time in the gut in patients with aGVHD grade 0-I versus
 335 II-IV. B) Importance plot of top 20 predictive gut ASVs identified by the svmLinear model with importance scores
 336 indicating the mean decrease in prediction accuracy in case the respective ASV would be excluded from the model.
 337 The final cross-validated svmLinear model predicted aGVHD (0-I versus II-IV) from the abundances of gut ASVs pre-
 338 HSCT with 86% accuracy (95% CI: 65% to 97%). The ASVs that were also confirmed by Boruta feature selection are
 339 indicated with asterisk. C) Conditional inference tree (CTREE) displaying ASVs identified as significant split nodes by

340 nonparametric regression for prediction of aGvHD. Numbers along the branches indicate split values of variance
 341 stabilized bacterial abundances. The terminal nodes show the proportion of samples originating from patients (n =
 342 number of samples) with aGvHD grade 0-I vs II-IV. D) Boxplots depicting the log transformed relative abundances of
 343 the predictive ASVs at time points up to the transplantation in aGvHD grade 0-I compared with grade II-IV patients.
 344 E) Trajectories of *Lactobacillaceae* and *Tannerellaceae* ASVs that were identified by tree-based sparse LDA, including
 345 ASV 3 and ASV 128 that were predictive for aGvHD (bold lines), in patients with aGvHD grade 0-I vs II-IV.
 346



347

348 **Figure 5. Machine learning-based prediction of aGvHD severity from the pre-HSCT oral microbiota composition.**

349 A) Relative abundances the 12 most abundant families over time in the oral cavity in patients with aGvHD grade 0-I
350 versus II-IV. B) Importance plot of top 20 predictive oral ASVs identified by the svmLinear model with importance
351 scores indicating the mean decrease in prediction accuracy in case the respective ASV would be excluded from the
352 model. The final cross-validated svmLinear model predicted aGvHD (0-I versus II-IV) from the abundances of oral
353 ASVs pre-HSCT with 92% accuracy (95% CI: 73% to 99%). The ASVs that were also confirmed by Boruta feature
354 selection are indicated with asterisk. C) Conditional inference tree (CTREE) displaying ASVs identified as significant
355 split nodes by nonparametric regression for prediction of aGvHD. Numbers along the branches indicate split values
356 of variance stabilized bacterial abundances. The terminal nodes show the proportion of samples originating from
357 patients (n = number of represented samples) with aGvHD grade 0-I vs II-IV. D) Boxplots depict the log transformed
358 relative abundances of the predictive ASVs at time points up to the transplantation in aGvHD grade 0-I compared
359 with grade II-IV patients. E) Trajectories of *Prevotellaceae* and *Actinomycetaceae* ASVs that were identified by tree-
360 based sparse LDA, including ASV 226 and ASV 568 that were predictive for aGvHD (bold lines), in patients with aGvHD
361 grade 0-I vs II-IV.

362

363

364 **Acute GvHD severity can be predicted from nasal microbiota composition prior to HSCT**

365 The proportion of nasal *Neisseriaceae* prior to HSCT was higher in patients with aGvHD grade 0-I as
366 compared to grade II-IV (Additional File 2: Figure S4A). In contrast, *Actinomycetaceae* and
367 *Corynebacteriaceae* exhibited a higher abundance in aGvHD grade II-IV patients prior to HSCT compared
368 to those with grade 0-I (Additional File 2: Figure S4A). We found two ASVs significantly predicting aGvHD
369 grade with opposite effects, ASV 66 and ASV 47. A high pre-HSCT abundance of ASV 66 (*Actinomyces* sp.,
370 *Actinomycetaceae*, $P = 0.03$) predicted development of aGvHD grade II-IV. The partial 16S rRNA gene
371 sequence of ASV 66 exhibited a high sequence similarity to *Actinomyces viscosus*. A total of 94% of
372 samples with a variance stabilized abundance >6.4 of ASV 66 originated from patients with subsequent
373 development of aGvHD grade II-IV (Additional File 2: Figures S4B and S4C). In support, pre-HSCT log
374 transformed relative abundances of ASV 66 (*Actinomyces* sp.) were 2.3 times higher in patients with
375 aGvHD grade II-IV compared to those with grade 0-I (Additional File 2: Figure S4C). In contrast, high pre-
376 HSCT abundance of ASV 47 (*Rothia* sp., $P = 0.03$) predicted that patients would be spared from aGvHD.
377 The partial 16S rRNA gene sequence of ASV 47 exhibited a high sequence similarity to *Rothia aeria*. All
378 nasal samples with a variance stabilized pre-HSCT abundance >-3.05 of ASV 47 (*Rothia* sp.) originated from
379 patients who subsequently developed no or mild aGvHD (grade 0-I) (Additional File 2: Figure S4B and S3C).

380

381 **Reconstitution of CD4+ T cells and the T_H17 subpopulation is associated with gut, oral, and nasal
382 microbiota**

383 In order to characterize associations between the microbiota and immune cell counts, immune markers,
384 and clinical outcomes in HSCT that potentially might impact our predictions of aGvHD, we implemented
385 two multivariate multi-table approaches, namely sparse partial least squares (sPLS) regression and
386 canonical correspondence analyses (CCpNA). Using sPLS regression, we identified three clusters of ASVs
387 for each body site, respectively (Figures 6A, and Additional File 2: S5A and S6A), which was supported by
388 the CCpNA (Figure 6B, and Additional File 2: S5B and S6B). Several cell populations of the adaptive immune
389 response were associated with one cluster each at all three body sites according to the sPLS analysis.
390 These included T cell counts at late follow-up time points, particularly CD4+ T cells in months +3 and +6,
391 and the subpopulation of T_H17 cells in months +1 and +3. In the gut, high numbers of these adaptive
392 immune cell populations were associated with high abundances of mainly *Lachnospiraceae*,

393 *Ruminococcaceae*, and *Lactobacillaceae* ASVs (gut cluster 1, Figure 6A). Of note, two of the *Lactobacillus*
394 spp. ASVs in gut cluster 1 (ASV 31 and ASV 586) were also observed as members of the group of
395 *Lactobacillaceae* that discriminated samples from different time points in the LDA (Figure 2C). In the oral
396 cavity, the same lymphocyte subsets were positively correlated with specific *Flavobacteriaceae*,
397 *Prevotellaceae*, *Veillonellaceae*, and *Neisseriaceae* ASVs (oral cluster 3, Additional File 2: Figure S5A). The
398 nasal cluster 1 that was affiliated with high T cell counts comprised predominantly *Veillonellaceae*
399 (Additional File 2: Figure S5A). The nasal cluster 3 was characterized by high T cell counts at pre-
400 examination and exhibited a high abundance of ASV 47 (*Rothia* sp.) and other taxa that were associated
401 with no to mild aGvHD (grade 0-I) (Additional File 2: Figure S4).

402 In the CCpnA, we observed that samples in gut cluster 1 (mainly from months +3 and +6) belonged to
403 patients with benign primary diseases, who received conditioning regimens involving fludarabine (Figure
404 6B). Moreover, these patients had a high number of bacterial and viral infections and were treated often
405 with phenoxymethylpenicillin compared to the overall patient population. In the oral cavity, samples
406 associated with CD4+ T cell reconstitution similarly stemmed from late follow-up time points and from
407 pre-examination. Patients in oral cluster 3 were generally treated with few antibiotics. The CCpnA of the
408 nasal data set indicated that patients with high CD4+ T cell and T_H17 cell counts at late follow-up time
409 points exhibited moderate to severe aGvHD (grade II-IV). Furthermore, these patients were treated with
410 meropenem, ciprofloxacin, and vancomycin more often compared with the remaining patient population
411 (Figure S6B). Most samples in the nasal cluster 1 were collected in weeks +2 and +3.

412

413 **Reconstitution of B cells is associated with gut, oral, and nasal microbiota**

414 At all three body sites, B cell counts at several late follow-up time points exhibited associations with
415 microbial abundances. High B cell counts were positively correlated with high abundances of
416 *Ruminococcaceae*, *Lachnospiraceae*, and *Rikenellaceae*, as well as few *Veillonellaceae* and
417 *Lactobacillaceae* in the gut (cluster 2, Figure 6A). In addition, the gut cluster 2 was associated with high
418 NK cell counts in month +1. In the oral cavity, ASVs within the small cluster 1, particularly ASV 422
419 (*Actinomyces odontolyticus*) and ASV 546 (*Veillonella parvula*), were positively correlated with these cell
420 counts, whereas ASVs affiliated with *Staphylococcaceae* and *Lactobacillaceae* (oral cluster 2) exhibited
421 negative correlations (Additional File 2: Figure S5A). ASV 422 (*Actinomyces odontolyticus*) was also
422 observed within the group of *Actinomycetaceae* ASVs in the LDA of the oral microbiota. In the nasal cavity,
423 abundances of *Streptococcaceae*, *Moraxellaceae*, and *Corynebacteriaceae* within nasal cluster 3 were
424 positively correlated with high B cell counts, particularly in month +3 (Additional File 2: Figure S6A). The
425 CCpnA indicated that samples in gut cluster 2 were taken predominantly in week +2, whereas samples in
426 oral cluster 1 were mainly collected in months +3 and +6 (Figures 6B and Additional File 2: Figure S5B).

427 Both the gut and oral CCpnA indicated that the associations between B cell counts and microbial
428 abundances predominantly occurred in patients who underwent a conditioning regimen without TBI and
429 without fludarabine (in contrast to conditioning regimens involving TBI or fludarabine). Furthermore,
430 these patients were treated with ceftazidime, vancomycin, and ciprofloxacin, but sparsely with other
431 antimicrobial agents (Figure 6B and Additional File 2: Figure S5B). The CCpnA on the gut data set revealed
432 that samples in this cluster (gut cluster 2) originated from both patients diagnosed with malignant diseases
433 and benign diseases (Figure 6B).

434

436 **Figure 6. Multivariate associations of the gut microbiota with immune and clinical parameters in HSCT.** A)
437 Clustered image map (CIM) based on sparse partial least squares (sPLS) regression analysis (dimensions 1, 2, and 3)
438 displaying pairwise correlations >0.3 / <-0.3 between ASVs (bottom) and continuous immune and clinical parameters
439 (right). Red indicates a positive correlation, and blue indicates a negative correlation, respectively. Based on the sPLS
440 regression model, hierarchical clustering (clustering method: complete linkage, distance method: Pearson's
441 correlation) was performed resulting in the three depicted clusters. B) Canonical correspondence analysis (CCpNA)
442 relating gut microbial abundances (circles) to continuous (arrows) and categorical (+) immune and clinical
443 parameters. ASVs and variables with at least one correlation >0.3 / <-0.3 in the sPLS analysis were included in the
444 CCpNA. The triplot shows variables and ASVs with a score >0.3 / <-0.3 on at least one of the first three CCpNA axes,
445 displayed on axis 1 versus 2 with samples depicted as triangles. The colored ellipses (depicted with 80% confidence
446 interval) correspond to the clusters of ASVs identified by the sPLS-based hierarchical clustering. Abbreviations not
447 mentioned in text: ATGmm, anti-thymocyte globulin; B_, blood; BU, busulfan; CY, Cyclophosphamide; DonorMatch6,
448 matched unrelated donor; FLU_other, fludarabine combinations without thiotepa; GvHD.Prophylaxis1, treatment
449 with cyclosporine; GvHD.Prophylaxis7, treatment with cyclosporine and methotrexate; immat_B, immature B cells;
450 K_d100, Karnofsky score on day +100; K_pre, Karnofsky score before HSCT; m1, month+1; m3, month+3; m6,
451 month+6; m12, month+12; mat_B, mature B cells; MEL, melphalan; total_B, total B cells; P_, plasma; parasitic,
452 parasitic infection; pre_cond, before conditioning start; pre_exam, pre-examination; THIO, thiotepa; viral, viral
453 infection; VP16, Etoposide.

454

455 **Body site-specific immune-microbial associations**

456 In addition to immune-microbial associations shared between two or three of the examined body sites,
457 we observed a few patterns that were exclusive to individual sites. In cluster 3 in the gut, we observed
458 ASVs primarily affiliated with *Bacteroidaceae* and *Tannerellaceae* whose abundances showed positive
459 correlations with eosinophil counts in months +3, +6, and +12. In the oral cavity, the sPLS analysis revealed
460 a sub-cluster of oral cluster 3 comprising ASVs affiliated with various families, e.g. ASV 1172 (*Actinomyces*
461 *sp.*), which was also identified as one of the discriminating *Actinomycetaceae* ASVs in the LDA. In the sPLS
462 analysis, this sub-cluster was associated with high counts of T_{reg} and T_H17 cells at late follow-up time points
463 (Additional File 2: Figure S6A).

464

465 **Discussion**

466 Both the microbiota and the immune system are subject to major changes during allogeneic HSCT. Failure
467 to re-establish host-microbial homeostasis might have adverse consequences for the patients, such as
468 prolonged immune deficiency. Long-term surveillance of microbial dynamics is required to understand i)
469 the shifts in the microbial community structure induced by HSCT and its accompanying treatments, and
470 ii) at which time points and under which conditions re-establishment of immunological and microbial
471 homeostasis occurs. Such knowledge may be of great prognostic value and may assist in guiding
472 personalized treatment strategies. Here, we present a comprehensive assessment of temporal microbial
473 abundance trajectories from before, at the time of, and after HSCT, to late follow-up time points up to
474 one year.

475

476 We have identified a group of *Ruminococcaceae*, and a clade of *Blautia* spp. (*Lachnospiraceae*), temporally
477 discriminating microbial community structure in the gut in relation to HSCT. We show a clear pattern of
478 depletion of fecal *Blautia* spp. immediately post HSCT, as well as their recovery from month +3 post HSCT
479 onwards. One could describe the trajectories of these potentially beneficial taxa as a “smile”-shape.
480 Previous studies have associated the taxonomic families of *Ruminococcaceae* and *Lachnospiraceae* (both

481 class *Clostridia*), and especially the genus *Blautia* (family *Lachnospiraceae*), with lower mortality, lower
482 GvHD, and higher bacterial diversity in adult allo-HSCT recipients [4,9,20–22]. In turn, a loss of those taxa
483 after HSCT was associated with subsequent adverse outcomes. Our findings extend the potential of
484 *Blautia* spp. abundances as an indicator of favorable clinical outcomes, as we characterize abundance
485 dynamics in children and provide important insight into the time point for the expected return to
486 abundances comparable to pre-HSCT time points (i.e. between month +1 and +3).

487
488 Adverse effects, like bacteremia and GvHD, have been found to accompany an expansion of the genus
489 *Enterococcus* post transplantation [2,3,6,23]. We have found a characteristic expansion of this genus, as
490 well as of certain *Lactobacillaceae* after HSCT, in agreement with other recent studies [4,6,11]. In addition,
491 we were able to show a decrease of *Enterococcus* spp. and *Lactobacillaceae* from month +3 to abundances
492 comparable to pre-HSCT levels. The abundance of these taxa over the course of one year might be
493 described as a “frown”-shaped trajectory. As for the “smile”-trajectory of potentially beneficial taxa, the
494 “frown”-trajectories of these taxa could be the first step towards a novel basis to evaluate the re-
495 establishment of patients’ microbial homeostasis and associated convalescence. Importantly,
496 *Enterococcus* was already higher abundant in the patient cohort at preexamination prior to HSCT as
497 compared to the healthy age-matched cohort, most likely due to prior chemotherapy and antibiotic
498 treatment given before referral to HSCT. Knowledge about the abundance level of *Enterococcus* before
499 HSCT could therefore provide valuable information about potential high-risk individuals already prior to
500 transplantation. It should be noted, however, that despite the observed different abundance levels in
501 patients and healthy controls, and the further expansion of *Enterococcus* post HSCT being in line with
502 previous studies, our multivariate analyses did not reveal direct detrimental host-microbial associations
503 of *Enterococcus* in the present cohort.

504
505 We have to our knowledge for the first time determined long-term dynamics of the oral and nasal
506 microbiota in allogeneic HSCT patients. Interestingly, we identified abundance trajectories of
507 phylogenetically closely related groups of *Actinomycetaceae*, *Streptococcaceae*, *Prevotellaceae*, and
508 Family XI (*Gemella* spp., Class *Bacillales*) in the oral cavity, resembling the “smile”-shaped trajectories
509 observed in gut. These taxa are part of the normal oral microbiota. Our findings are in agreement with
510 previous studies reporting the detection of fewer *Prevotella* spp. and *Streptococcus* spp. in the oral cavity
511 during the first month post HSCT [24,25]. In addition, our current study provides insight into the time of
512 recovery of these taxa in month +3 after HSCT.

513
514 For the oral cavity, a post-transplant expansion of *Enterococcus* spp. and *Staphylococcus* spp. has been
515 reported previously [25,26]. Consistently, we observed an increased relative abundance of
516 *Staphylococcaceae* during the first month post HSCT, but we did not identify *Enterococcus* spp. or
517 *Staphylococcus* spp. as significant drivers of temporal dynamics in the oral cavity. Previously, increased
518 *Enterococcus* abundances post HSCT were found predominantly in patients who developed oral mucositis,
519 which was not directly assessed in our study [25,27]. Therefore, our findings suggest that further
520 investigation of taxa that exhibit “smile”-like abundance trajectories could be relevant in direct relation
521 to oral mucositis. Especially *Actinomycetaceae*, *Streptococcaceae*, and *Prevotellaceae*, when low-

522 abundant, might be candidates for bacterial predictors of oral mucositis, and furthermore might be
523 employed to facilitate preventive management.

524
525 In the nasal cavity, the microbiota did not exhibit temporal patterns as distinct as the “smile“- and
526 “frown“- shaped trajectories in the gut and the oral cavity. One could speculate that nasal bacterial
527 abundance patterns might be more individualized, which might in turn conceal pronounced patterns
528 when looking at the patient population as a whole. However, certain host-microbial associations observed
529 in the gut were reflected in the nasal cavity. For instance, reconstitution of CD4+ T cells and the T_H17
530 subset were associated with distinct groups of ASVs at all three body sites.

531
532 Together, these findings suggest that the oral and potentially also the nasal cavity might constitute easily
533 accessible microbial niches suitable for investigating host-microbial associations in the context of HSCT,
534 similar to current strategies for the gut. While mucous membranes that are in close association with
535 distinct microbial communities characterize all three niches, it is more feasible to collect buccal and
536 anterior naris swabs during clinical routine as compared to collecting fecal samples. Fecal sample
537 collection is dependent on bowel movements, which often are impaired in this patient group. Therefore,
538 our study provides valuable knowledge for possible future applications that could include the monitoring
539 of oral microbial dynamics in clinical routine, which might be easier to implement than routine fecal
540 sampling.

541
542 We identified AVSs with the potential to predict post-transplant aGvHD, which might open opportunities
543 to improved preventive clinical management, for example by intensified prophylactic immunosuppression
544 for patients at increased risk. Some ASVs were significant for both, discriminating the microbiota in long-
545 term dynamics as well as in the prediction of aGVHD severity from the microbiotas prior to HSCT, such as
546 ASV 3 (*Lactobacillus* sp.) in the gut, as well as ASV 568 (*Actinomyces* sp.) and ASV 226 (*Prevotella*
547 *melaninogenica*) in the oral cavity. While we do not yet understand the biological mechanisms underlying
548 this observation, these taxa could be of particular interest for a long-term monitoring in pediatric HSCT
549 patients, starting prior to HSCT. Like the gut microbiota, the oral and nasal commensal residents might be
550 of systemic relevance, and a more holistic picture of microbial influences might be drawn by examining
551 various niches with bacterial communities potentially interacting across body sites. In light of intimate
552 host-microbiota interactions, the microbial community patterns might also be a marker for underlying
553 changes occurring in the immune system.

554
555 High abundances at late follow-up time points of two fecal *Lactobacillus* spp. that expanded after HSCT
556 showed positive correlations with T cell reconstitution. This is in line with previous studies suggesting that
557 the expansion of *Lactobacillus*, a genus commonly associated with probiotic properties, might promote
558 immune homeostasis and thereby exert a protective effect to limit *Enterococcus* expansion [4,23,28]. A
559 potential explanation indicated by our results might be that high *Lactobacillus* abundances outlasting
560 enterococcal dominance promotes T cell reconstitution. However, the associated cell populations include
561 T_H17 cells which can facilitate inflammation, and therefore it is difficult to determine whether the
562 observed *Lactobacillus* expansion is exclusively beneficial [13]. However, Th17 cells could perhaps add to

563 the host defense in these patients and therefore be beneficial for local homeostasis, although with the
564 unusual cost of harmful inflammation.

565
566 Furthermore, we found associations of high *Lachnospiraceae* and *Ruminococcaceae* in the gut with rapid
567 B and NK cell reconstitution, which is in support of our previous study [4]. These two *Clostridiales* families
568 play an important role in providing the host with short-chain fatty acids (SCFAs), such as butyrate [5,29].
569 A study demonstrated that SCFAs can facilitate the differentiation of human naïve B cells to plasma cells
570 in culture [30]. Whether SCFAs also directly influence B cell proliferation is yet unknown.

571
572 We have made several observations in which infections and/or antibiotic treatments were associated with
573 the abundance of specific bacterial clusters at certain body sites, immune cell counts, and aGvHD. For
574 example, patients whose samples were represented by gut microbiota cluster 1 experienced a high
575 number of infections and were treated often with phenoxymethylpenicillin compared to the overall
576 patient population. In contrast, patients affiliated with gut microbiota cluster 2 experienced treatment
577 with ceftazidime, vancomycin, and ciprofloxacin, but sparsely with other antimicrobial agents.
578 Furthermore, patients affiliated with oral microbiota cluster 3 were generally treated with few antibiotics,
579 and, patients whose sample were represented by the nasal microbiota cluster 1 were treated often with
580 meropenem, ciprofloxacin, and vancomycin compared with the remaining patient population. However,
581 it is challenging to interpret these observations, as these patient samples were also associated with other
582 features, such as an increased or decreased abundance of certain immune cells (see Additional file 3 for
583 further discussion), or the patients were exposed to other treatments as well, such as TBI or fludarabine.
584 Overall, however, our observations are consistent with previous reports that antimicrobial treatment is
585 associated with changes in microbiota composition in patients undergoing allo-HSCT and might impact
586 clinical outcomes [4,11,31–33]. It will be important to gain a more mechanistic understanding of the
587 possible effects of antimicrobial treatment to disentangle the effect of antibiotics from that of other
588 medications and host responses. Such insight could for example allow selecting more suitable
589 antimicrobials for treatment in HSCT patients that spare the elimination of beneficial taxa, whose decline
590 might be associated with more severe clinical outcomes. The choice of antibiotic treatment might also be
591 important to take into consideration in patients that might potentially be referred to HSCT eventually,
592 given that we already observed certain changes in the microbiota in the patients at referral compared to
593 healthy controls. The microbiota at referral already exhibited some features that were associated with
594 more severe side effects.

595
596 Associations between aGvHD severity and the microbiota have to date merely been based on logistic
597 regression and correlation analyses [8,34–36]. In addition, microbial abundances at the time of neutrophil
598 recovery or engraftment were assessed, i.e. at time points shortly before, concurrent to, or potentially
599 after aGvHD onset [8,16,36]. Here, we have implemented machine learning techniques to take the
600 assessment of microbiota-aGvHD relations from correlative to predictive modeling: We presented
601 evidence that aGvHD severity may be predicted from pre-HSCT microbial abundances in the gut, as well
602 as in the oral and nasal cavities. This could open up opportunities for the future where microbial markers
603 guide early interventions to prevent aGvHD. This could include a modulation of the microbiota of patients
604 predicted to be at high risk with synthetic microbiotas containing beneficial bacteria, including probiotics.

605 Notably, we have to our knowledge for the first time revealed microbial taxa in the oral and nasal cavity
606 that may predict aGvHD. A further discussion on possible connections between specific microbial taxa of
607 the gut, oral, and nasal cavity, immune responses, and aGvHD can be found in Additional file 3.

608

609 **Conclusions**

610

611 With the present study we bring forward a comprehensive framework of host-microbial associations in
612 allogeneic HSCT. We focused on long-term microbial dynamics, demonstrating distinct microbial
613 abundance patterns of disturbance and recovery, as well as making predictions about aGvHD from the
614 pre-transplant microbiota. We discovered that the microbial community composition in patients prior to
615 HSCT already differs somewhat from healthy controls in regard to key microbial taxa, opening up
616 opportunities for potential preventive measure in the future. Moreover, we confirmed the depletion of
617 *Blautia* spp. and expansion of *Enterococcus* spp. in the gut after HSCT and expand this knowledge by
618 precisely defining which phylogenetically closely related sequence variants of these genera are
619 characteristic for those patterns, and when they return to pre-HSCT levels. We identified similar patterns
620 for members of the oral and nasal microbiota and propose month +3 post-transplant as a possible
621 universally crucial time point for microbiota reconstitution after HSCT. We demonstrate that high
622 abundances of for example an intestinal *P. distasonis* ASV, and an oral *P. melaninogenica* ASV pre-HSCT
623 predict the development of moderate to severe aGvHD post-transplant. When relating microbial
624 abundances with immune cell counts, we found rapid B and NK cell reconstitution to be associated with
625 high abundances of *Lachnospiraceae* and *Ruminococcaceae*, which also depended on antibiotics treatment.
626 Distinct ASVs at all three body sites were associated with T_H17 cell counts, suggesting future research on
627 a potential immunomodulatory involvement of the microbiota in inflammation regulation, which might
628 play a role for aGvHD development. We have discovered host-microbial associations shared between two
629 or more of the examined body sites. This may open up opportunities for implementing a more feasible
630 oral and nasal swab sampling into research and clinical diagnostic activities to design more precise patient
631 treatment strategies to reduce serious side effects and improve immune and microbiota reconstitution.

632

633

634 **Materials and Methods**

635

636 **Patient recruitment and sample collection**

637 We recruited 29 children (age range: 2.5 - 16.4 years) who underwent their first myeloablative allogeneic
638 hematopoietic stem cell transplantation at Copenhagen University Hospital Rigshospitalet (Denmark)
639 between November 2015 and October 2017. We provide detailed information about the patients' clinical
640 characteristics in Table S1 (Additional File 1). Every patient underwent a myeloablative conditioning
641 regimen starting on day -10 for patients receiving a graft from a haploidentical donor, and on day -7 for
642 patients with sibling or matched unrelated donors (Additional File 1: Table S1). One patient had a donor
643 lymphocyte infusion on day +223 after the first transplantation. Immune cell count date of this patient
644 was excluded from our analysis from the time of donor lymphocyte infusion. We grouped the patients
645 into four categories of conditioning regimens: 1. TBI_CY_or_TBI_VP16 (n=6; TBI + cyclophosphamide or

646 TBI + etoposide), 2. BU_CY_VP16_MEL_combos (n=6; Combinations of busulfan, cyclophosphamide,
647 etoposide and melphalan), 3. FLU_THIO (n=12; subgroups: fludarabine + busulfan + thiotepa (n=6);
648 fludarabine + treosulfan + thiotepa (n=4); fludarabine + thiotepa (n=1); fludarabine + cyclophosphamide
649 + thiotepa (n=1)), and 4. FLU_other (n=5; subgroups: fludarabine + busulfan (n=2); fludarabine +
650 cyclophosphamide (n=2); fludarabine + treosulfan (n=1)) (Additional File 1: Table S1). The following
651 sampling time points were defined: pre-examination (between day -57 and day -15), around the start of
652 conditioning (between day -14 and day -3 and latest 2 days after conditioning start), at time of HSCT
653 (between day -2 and day +2), and weekly during the first 3 weeks after transplantation (week +1: day +3
654 to day +10, week +2: day +11 to day +17, week +3: day +18 to day +24) (Figure 1A). Broader intervals
655 applied to follow-up time points: Month +1 (between days +25 and +45), month +3 (between days +46
656 and +120), month +6 (between days +121 and +245), and month +12 (between day +246 and +428). Acute
657 GvHD was graded by daily clinical assessment of skin, liver and gastro-intestinal manifestations according
658 to the Glucksberg criteria [37]. We group aGvHD severity into grade 0-I and grade II-IV, reflecting clinical
659 practice where grade I represents limited alloreactivity with no (or very limited) impact on the overall
660 clinical outcome of HSCT, and therefore no need for medical treatment of these patients, such as the use
661 of glucocorticoids, which is first-line treatment for grade II-IV aGvHD.
662 To address certain specific questions, we also analyzed the microbiota (from time point 0) of a cohort of
663 18 healthy children that were part of a previous study [19]. The median age of these children was 6.8
664 years (interquartile range 4.6 to 9.6). A total of 30 fecal samples were obtained (11 children provided two
665 samples each within an interval of six months). The children did not receive any antibiotics within the
666 month prior to sample collection. The samples were processed in the same way as the fecal samples of
667 the patients of this study (described below).

668

669 **Infections and antibiotics**

670 Records of bacterial, fungal, viral, and parasitic infections and antibiotic treatment from before HSCT
671 (from day -30 or at the collection time of the first microbiota sample in case this was earlier) until month
672 +12 (day +428) were taken into consideration (or as long as data was available for the most recent
673 patients; data accessed in July 2018). This corresponds to the length of the sampling period of fecal and
674 swab samples.

675

676 **Analysis of immune cell subpopulations**

677 Leukocyte counts were recorded daily during hospitalization starting prior to HSCT, and later weekly in
678 the outpatient clinic by flow cytometry (Sysmex XN) or microscopy (CellaVision DM96 microscope) in case
679 of very low counts. Monitored subpopulations included lymphocytes, monocytes, neutrophils, basophils,
680 and eosinophils.

681

682 **Analysis of T, B and NK cells in peripheral blood**

683 T, B, and NK cell counts in $\times 10^9/L$ were determined at pre-examination, and in month +1, +3, +6, and +12.
684 Trucount Tubes (Becton Dickinson, Albertslund, Denmark) were used to quantify these cell types in
685 peripheral blood on a FC500 flow cytometer (Beckman Coulter, Copenhagen, Denmark). For
686 immunofluorescence staining, the following conjugated monoclonal antibodies were used for CD3+ T
687 cells, CD3+CD4+ T cells and CD3+CD8+ T cell quantification: CD3-PerCP, CD3-FITC, CD4-FITC, CD8-PE

688 (Becton Dickinson). CD45-PerCP, CD16/56-PE antibodies were used to determine NK cells based on their
689 CD45+CD16+CD56+ phenotype. For B cells, total B cells (CD45+CD19+), mature B cells
690 (CD45+CD19+CD20+) and immature B cells (CD45+CD19+CD20-) were differentiated by using CD20-FITC
691 and CD19-PE antibodies.

692

693 **Subtyping of T cells**

694 Peripheral blood samples were collected in month +1, +3 and +6 for isolation of peripheral blood
695 mononuclear cells (PBMCs) by gradient centrifugation of heparinized blood with Lymphoprep™ (Axis-
696 Shield, Oslo, Norway). PBMCs were washed in PBS (Life Technologies, Invitrogen, Paisley, U.K.) three times
697 and then resuspended in RPMI 1640 buffer containing HEPES (Biological Industries Israel Beit-Haemek Ltd,
698 Kibbutz Beit-Haemek, Israel), L-glutamine (GIBCO, Invitrogen, Carlsbad, CA) and Gentamycin (GIBCO), 30%
699 fetal bovine serum (Biological Industries) and 10% Dimethyl Sulfoxide (VWR, Herlev, Denmark) for cryo-
700 preservation in liquid nitrogen.

701 T cell subsets, i.e. T_H17 cells and T_{reg} cells, were quantified from frozen PBMCs by flow cytometry on a
702 FACS Fortessa III flow cytometer (Becton Dickinson, Albertslund, Denmark). PBMCs were thawed and
703 washed before incubation with Fixable viability stain 620 (Becton Dickinson) and a set of conjugated
704 monoclonal antibodies for 30 minutes on ice: CD3-APC-A750 (Beckmann Coulter), CD4-PE-Cy7 (Beckmann
705 Coulter), CD8-A700 (Becton Dickinson), CD25-PE (Becton Dickinson), CD39-PerCP-Cy5.5 (Beckmann
706 Coulter), CD196-BV510 (Biolegend, San Diego, USA), CD127-BV711 (Biolegend), CD161-BV650 (Becton
707 Dickinson) and CD45RA-BV786 (Becton Dickinson). Next, PBMCs were washed and incubated with
708 transcription factor buffer set (BD) for 45 min on ice. Afterwards, PBMCs were washed twice and
709 intracellular monoclonal antibodies were added and incubated for 45 minutes on ice: RORγT-A488
710 (Becton Dickinson), FOXP3-A647 (Becton Dickinson) and Helios-PB (Beckmann Coulter). TH17 cells were
711 determined by the CD4+RORγT+ phenotype, and T_{reg} cells by the CD4+CD25^{high}FOXP3+ phenotype.
712 Absolute cell counts in x10⁹/L were obtained by multiplying the frequency of T_H17 and T_{reg} cells with the
713 CD4+ T cell count from the same time point.

714

715 **Quantification of inflammation and infection markers**

716 Markers were measured at the Department of Clinical Biochemistry, Copenhagen University Hospital
717 Rigshospitalet, Denmark. As a marker of infection, plasma procalcitonin was determined by sandwich
718 electrochemiluminescence immunoassays (ECLIA). As a marker of systemic inflammation, CRP was
719 measured by latex immunoturbidimetric assays (LIA).

720

721 **DNA isolation from fecal, oral, and nasal samples and 16S rRNA gene sequencing**

722 A total of 212 fecal samples for analysis of the intestinal microbiota were collected from 29 patients at
723 the 10 time points described above. The gut microbiota was characterized at ≤6 time points in 9 patients
724 (31%), at 7-8 time points in 13 patients (45%) and at 9-10 time points in 7 patients (24%) (Additional File
725 1: Table S1). DNA from fecal samples, one blank control per extraction round (thereof sequenced: 14),
726 one mock community sample (Biodefense and Emerging Infectious Research (BEI) Resources of the
727 American Type Culture Collection (ATCC) (Manassas, VA, USA), Catalog No. HM-276D) per sequencing run

728 and two collection tube controls was isolated using the QIAamp Fast DNA Stool Mini kit (Qiagen, Venlo,
729 Netherlands), following the manufacturer's instructions with modifications according to [38].
730 We collected 248 buccal swabs (3x at ≤6 time points (10%), 11x at 7-8 time points (38%), 15x at 9-10 time
731 points (52%)) and 249 anterior naris swabs (3x at ≤6 time points (10%), 9x at 7-8 time points (31%), 17x at
732 9-10 time points (59%)). DNA from swab samples, one blank control per extraction round (thereof
733 sequenced: 28), one mock community sample per run, two collection tube controls, and two sampling
734 swab controls was isolated using the QIAamp UCP Pathogen Mini kit (Qiagen, Venlo, Netherlands), with
735 the 'Protocol: Pretreatment of Microbial DNA from Eye, Nasal, Pharyngeal, or other Swabs (Protocol
736 without Pre-lysis)' and subsequently the 'Protocol: Sample Prep (Spin Protocol)', following the
737 manufacturer's instructions with the following modifications: 550µl instead of 500µl Buffer ATL was used
738 during pretreatment; DNA was eluted twice with 20µl Buffer AVE into 1.5 ml DNA LoBind tubes
739 (Eppendorf, Hamburg, Germany) instead of the tubes provided with the kits.
740 Library construction and sequencing on an Illumina MiSeq instrument (Illumina Inc., San Diego, CA, USA)
741 was performed at the Multi Assay Core facility (DMAC), Technical University of Denmark. DNA
742 concentration of each sample was measured using a NanoDrop spectrophotometer (Thermo Scientific,
743 Waltham, MA, USA). Library construction was performed according to the *16S Metagenomic Sequencing*
744 *Library Preparation* protocol by Illumina [39]: The V3-V4 region of the 16S ribosomal RNA gene were
745 amplified in a PCR in each sample and in the controls, using the following previously evaluated primers,
746 preceded by Illumina adapters [40]: 341F (5'-
747 TCGTCGGCAGCGTCAGATGTGTATAAGAGACAGCCTACGGGNGGCWGCAG-3') and 805R (5'-
748 GTCTCGTGGGCTCGGAGATGTGTATAAGAGACAGGACTACHVGGGTATCTAATCC-3'). Amplicons were then
749 analyzed for quantity and quality in an Agilent 2100 Bioanalyzer with the use of an Agilent RNA 1000 Nano
750 Kit (Agilent Technology, Santa Clara, CA, USA). Subsequently, the amplicons were purified on AMPure XP
751 Beads (Beckman Coulter, Copenhagen, Denmark) according to the manufacturer's instructions. Illumina
752 adapters and dual-index barcodes were then added to the amplicon target in a PCR according to Illumina
753 [39] using the 96 sample Nextera XT Index Kit (Illumina, FC-131–1002). A final clean-up of the libraries was
754 performed in another PCR step, using AMPure XP Beads (Beckman Coulter, Copenhagen, Denmark)
755 according to the manufacturer's instructions, followed by a confirmation of the target size in an Agilent
756 2100 Bioanalyzer (Agilent Technologies). Before sequencing, DNA concentration was determined with a
757 Qubit (Life Technologies, Carlsbad, CA, USA) and libraries were pooled. In preparation for sequencing, the
758 pooled libraries were denatured with NaOH, diluted with hybridization buffer, and heat denatured. 5%
759 PhiX was included as an internal control for low-diversity libraries. Paired-end sequencing with 2 × 300bp
760 reads was performed with a MiSeq v3 reagent kit on an Illumina MiSeq instrument (Illumina Inc., San
761 Diego, CA, USA).

762

763 **16S rRNA gene sequence pre-processing**

764 Raw sequence reads were demultiplexed based on sample-specific barcodes and 'read 1' and 'read 2'
765 FASTQ files for each sample were generated on the Illumina MiSeq instrument by the MiSeq reporter
766 software. Primers were removed by using cutadapt (version 1.16) [41] at a tolerated maximum error rate
767 of 15% for matching the primer sequence anchored in the beginning of each read. In the case that at least
768 one read of a pair did not contain the primer, the pair was discarded. Only pairs in which the forward read

769 contained the forward primer (341F) and the reverse read contained the reverse primer (805R) were
770 retained.

771 The resulting reads were further processed using the R package DADA2 (version 1.8) to infer high-
772 resolution amplicon sequence variants (ASV) [42]. Forward and reverse reads were truncated at 280 bp
773 and 200 bp respectively. This way, the majority of reads retained a quality score >25 according to MultiQC
774 analysis [43]. These truncation thresholds also ensured an overlap of 480 bp (expected amplicon length
775 of 460 bp + 20 bp), allowing to merge forward and reverse reads. Samples were pooled for the sample
776 inference step (*dada()* function) to increase the power for detecting rare variants. Default values were
777 used for all other quality filtration parameters in DADA2. DNA from samples with a read count <10,000
778 after preliminary chimera and contaminant removal were re-sequenced. DNA from feces samples with a
779 read count <5,000 were re-extracted. Eventually, chimeras were identified by sample and removed from
780 the whole data set (over all sequencing runs) based on a consensus decision (*removeBimeraDenovo()*
781 function, method “consensus”). Taxonomic assignment on ASVs was done by using the Silva reference
782 data base (version 132), formatted for DADA2 [44]. Additional species assignment by exact reference
783 strain matching was performed using the Silva species-assignment training data base, formatted for
784 DADA2 [44].

785 The resulting ASV and taxonomy tables were integrated with the R package phyloseq and its dependencies
786 (version 1.24.0) [45]. The data was split into two data sets, one containing feces sample data and one
787 containing nasal and oral swab data. Subsequently, contaminant removal was performed with the R
788 package decontam [46]. Potential technical batch effects by sequencing run, 96-well plate, extraction kit,
789 extraction round, experimenter, and extraction date were assessed by ordination (Principal Coordinates
790 Analysis (PCoA)).

791 For both, the fecal sample data set and the swab data set, contaminants were identified by sequencing
792 run as a batch effect and a subsequent calculation of a consensus probability. For the feces sample data
793 set, contaminants were identified by both, increased prevalence in 14 blank extraction controls and by
794 relating ASV frequency to post-PCR sample DNA concentration, assuming inverse correlation (method
795 “both”, frequency threshold: 0.2, prevalence threshold: 0.075) [46]. After manual evaluation of edge
796 cases, 89 ASVs were removed from the fecal sample data set as contaminants. In an additional step, we
797 identified 7 contaminants from 2 sampling tube controls (method and thresholds as stated above). In
798 total, 96 ASVs were removed as contaminants from the fecal sample data set.

799 For the swab sample data set, contaminants were identified by both, increased prevalence in 28 blank
800 extraction controls and by relating ASV frequency to post-PCR sample DNA concentration (method “both”,
801 frequency threshold: 0.1, prevalence threshold: 0.6) [46]. A more stringent threshold for prevalence
802 compared to frequency was chosen here, given the low biomass of the swab samples, accompanied by
803 post-PCR DNA concentrations similar to those in blank controls. After manual evaluation of edge cases,
804 1137 ASVs were removed from the swab sample data set as contaminants. In an additional step, we
805 identified 16 contaminants from 2 sampling tube controls and 2 swab controls (method “both”, frequency
806 threshold: 0.075, prevalence threshold: 0.5). In total, 1153 ASVs were removed as contaminants from the
807 oral and nasal swab sample data set.

808 For each subset, we created a phylogenetic tree by de novo alignment of the inferred ASVs, following a
809 previously described workflow [47]. First, we performed multiple alignment with the package DECIPHER
810 [48]. Subsequently, we built a neighbor-joining tree using the package phangorn [49], based on which we

811 fitted a GTR+G+I (Generalized time-reversible with Gamma rate variation) maximum likelihood tree. The
812 phylogenetic tree for each data set (fecal, oral, and nasal) was then integrated with the respective
813 phyloseq object.

814 Next, we took core subsets of the ASVs remaining after contaminant removal using the function *kOverA()*
815 from R package *genefilter* [50]. In the fecal set, 2465 ASVs with ≥ 5 reads in ≥ 2 samples were retained.
816 With ≥ 5 reads in ≥ 10 samples, 509 ASVs were retained from the oral sample set, and 602 ASVs from the
817 nasal sample set. Additional manual contaminant filtering was applied to the oral and nasal core sets.
818 ASVs affiliated with taxonomic families commonly found in both the oral or nasal cavity and the gut were
819 only retained in the oral sample set in case they had ≥ 10 reads in ≥ 10 samples. ASVs of families only
820 expected in the gut were removed from the oral and nasal sample sets after manually assessing their
821 abundances. Subsequently, we retained 377 ASVs in the oral sample set, and 197 ASVs in the nasal sample
822 set.

823 For the comparison of the fecal microbiota in preexamination samples ($n=15$) of HSCT patients and
824 healthy children ($n=18$), these data were combined in a phyloseq object. The same set of putative
825 contaminants was removed from the healthy data set as were identified within the full fecal data set of
826 HSCT patients. Subsequently, a core subset was taken as described above (retaining ASVs with ≥ 5 reads
827 in ≥ 2 samples).

828

829 **Statistical analysis**

830 Statistical analyses and generation of graphs was performed in R (version 3.5.1, R Foundation for
831 Statistical Computing, Vienna, Austria) [51]. The R scripts documenting the major steps of our statistical
832 analyses are available from figshare (<https://doi.org/10.6084/m9.figshare.12280001>). Sequencing data,
833 and experimental and clinical data (<https://doi.org/10.6084/m9.figshare.12280028>) were integrated for
834 analysis by using the R package *phyloseq* and its dependencies [45]. We also provide the resulting
835 *phyloseq* objects through figshare (<https://doi.org/10.6084/m9.figshare.12280004>). Plots were
836 generated with the packages *ggplot2* [52], *mixOmics* [53], *treeDA* [54], *caret* [55], and *partykit* [56,57].
837 From the core sets of ASV counts for each body site, bacterial alpha diversity (denoted by the inverse
838 Simpson index) was calculated and compared between time points by using a Friedman test with
839 Benjamini-Hochberg correction for multiple testing, and a post-hoc Conover test. To gain insight into
840 changes of microbial abundances over time in relation to HSCT, we agglomerated ASV counts on family
841 levels with the function *tax_glom()* in *phyloseq* [45]. Thereafter, we displayed the relative abundances of
842 the 12 most abundant families at each body site for each time point. We also depicted relative abundances
843 over time on family level in patients with aGvHD grade 0-I versus grade II-IV.

844 In order to determine which particular ASVs are relevant in temporal microbial abundance dynamics at
845 each body site, we implemented tree-based sparse linear discriminant analysis (LDA) with the package
846 *treeDA* [54]. This supervised method implements prior information about phylogenetic relationships
847 between ASVs to perform supervised discrimination of classes, here time points, and induces sparsity
848 constraints to increase interpretability [58]. Leaves and nodes of the phylogenetic tree, representing $\log+1$
849 transformed ASV abundances and the sums thereof respectively, were used as predictive features. The
850 core oral and nasal sets were used as input as described above, while the fecal set was further reduced to
851 389 ASVs with >5 reads in >10 samples for this analysis. Leave-one-out cross validation (LOOCV) was
852 performed to choose the optimal minimum number of predictive features ensuring sparse, interpretable

853 models. The resulting LDA models had 9 components. By default, this number corresponds to the number
854 of predicted classes (here 10 time points) less one. To identify relevant components, we plotted sample
855 scores colored by time points along each component and plotted the components pairwise against each
856 other (Figure 1C). Thereby, we revealed that the first LDA-component for each body site showed the
857 highest sample scores and best separated the samples by time point. Therefore, we proceeded with
858 displaying temporal trajectories of clades of predictive features (ASVs) on the first component. For
859 selected groups of predictive ASVs we displayed trajectories for patients with aGvHD grade 0-I versus with
860 grade II-IV.

861 Next, we implemented machine learning models to predict aGvHD grade post-transplant from preceding
862 ASV abundances. The strategy and R code for the machine learning approach was partially adapted from
863 a previous approach [59,60]. As a preparative step for this analysis, we variance-stabilized the ASV count
864 data. To do so, we first performed size factor estimation for zero-inflated data on the core data sets for
865 each body site with the package GMPR [61]. Subsequently, we transformed the data by using the
866 function *varianceStabilizingTransformation()* in the package DESeq2 [62]. The function implements a
867 Gamma-Poisson mixture model to account for both library size differences and biological variability [63].
868 For the prediction of aGvHD grade, we compared the performances of four different classifiers (random
869 forest (rf), boosted logistic regression (LogitBoost), support vector machines with linear kernel
870 (svmLinear), and support vector machines with radial basis function kernel (svmRadial)) using the package
871 caret [55]. We took subsets of the phyloseq objects comprising only the time points preceding aGvHD
872 onset: pre-examination, conditioning start, and at the time of HSCT. Prior to fitting the models, we
873 excluded ASVs with near zero variance, i.e. those that were not differentially abundant between any
874 samples, by using the function *nearZeroVar()* in package caret [55]. Thereby we obtained sets of 238, 186,
875 and 100 ASVs for the fecal, oral, and nasal data set, respectively, which were then assessed as potential
876 predictors of subsequent aGvHD. All classifiers were trained on a randomly chosen subset of 70% of the
877 data to build a predictive model evaluated on a test set (30% of the data). Splitting was performed in a
878 way that samples from the same patient at different time points were kept together in either the testing
879 or training set to ensure that the outcome of a patient can only appear in either the testing set or the
880 training set, but not both. Thirty iterations of 10-fold cross-validation were performed for each classifier,
881 both with and without up-sampling. Up-sampling refers to the process of replacement-based sampling of
882 the class with fewer samples (here aGvHD grade II-IV) to the same size as the class with more samples
883 (here aGvHD grade 0-I) to achieve a balanced design. SvmLinear on up-sampled data was chosen as the
884 best performing predictive model for all three data sets (gut, oral, and nasal). Subsequently, we performed
885 Boruta feature selection using the package Boruta [64]. The Boruta algorithm is a Random Forest
886 classification based wrapper that compares the importance of real features to that of so called 'shadow
887 attributes' with randomly shuffled values. Features that are less important than the 'shadow attributes'
888 are iteratively removed. Here, we retained those ASVs in each data set that were both, among the 50
889 most important predictors in the svmLinear model and confirmed by the Boruta algorithm (Additional File
890 1: Table S3). Subsequently, we fitted a CTREE on each set of selected predictors (17 gut, 26 oral, and 12
891 nasal ASVs) by using the package partykit (Additional File 1: Table S3) [56,57]. In the CTREE analysis the
892 effect of the predictive ASVs on aGvHD grade is evaluated in a nonparametric regression framework. Using
893 CTREE, we found 3 significant ASVs each in the gut and in the oral data set, and two significant ASVs in the
894 nasal data set. CTREE iteratively tests if the abundance of any ASV has a significant effect on aGvHD grade.

895 In the case that a significant relation is found, the ASVs with the largest effect is picked as a node for the
896 tree. The procedure is then recursively repeated until no further significant effect of any ASV on aGvHD is
897 found. We plotted the result as a tree featuring the significant split nodes, represented by the ASVs and
898 the Bonferroni-corrected p-values indication significant influence of their abundance on aGvHD grade.
899 The terminal nodes of the tree show the proportion of samples stemming from patients with aGvHD grade
900 0-I versus II-IV, under the condition of the abundance split criterion described on each branch. Since we
901 used variance stabilized bacterial abundances as input for the machine learning analyses, abundances can
902 be presented as negative values in some cases and are therefore not easy to interpret intuitively.
903 Therefore, we additionally displayed the log-transformed relative abundances of all ASVs significantly
904 predicting aGvHD in boxplots at the three investigated time points (pre-examination, conditioning start,
905 and at the time of HSCT).

906 Subsequently, we were interested in associations between the fecal, oral, and nasal microbiota and
907 immune cell counts, and clinical outcomes in HSCT. Records of immune markers, and immune cell counts
908 contained left- and right-censored measurements, i.e. observations below or above the detection (or
909 recording) limit, respectively. In order to use these data in analyses that do not tolerate censored records,
910 we needed to impute the censored data. Therefore, we first fitted the non-parametric maximum
911 likelihood estimator (NPMLE, also called Turnbull estimator) for univariate interval censored data on each
912 variable that contained censored records, using the function *ic_np()* in the R package *icenReg* [65].
913 Subsequently, censored records were imputed, informed by the model that was fitted on the entirety of
914 observed and censored data of each variable, using the *imputeCens()* function [65]. Next, we took the
915 median of measurements for the time points defined above for those immune markers, and immune cell
916 counts that have been measured more frequently than that. This way, we obtained comparable data sets.
917 Continuous immune marker and cell count data that was systematically missing for certain sampling time
918 points was split by time points and unavailable time points were excluded. Missing values in continuous
919 immune marker and cell count data were imputed for variables with $\leq 50\%$ missingness. Simultaneous
920 multivariate non-parametric imputation was performed using the R package *missForest* [66]. Variables
921 with more than 50% missing values were excluded from the analysis.

922 Next, we implemented two multivariate multi-table approaches to gain a detailed understanding of how
923 the fecal, oral, and nasal microbiota might be associated with immune cell counts, immune markers, and
924 clinical outcomes in HSCT. Evaluated clinical outcomes comprised acute GvHD (grade 0-I versus II-IV),
925 relapse, overall survival, and treatment-related mortality. Furthermore, we included bacterial alpha
926 diversity (inversed Simpson index), antibiotic treatment, infections, Karnofsky scores before conditioning
927 and at day +100, and patients' baseline parameters (age, weight, sex, primary disease, malignant versus
928 benign primary disease, conditioning regimen (including ATG treatment), chemotherapeutic agents'
929 dosages, TBI treatment and dosage, stem cell source, GvHD prophylactic regimen, donor type
930 (sibling/matched unrelated/haploidentical), donor HLA-match, and donor sex).

931 For each body site, we performed sparse partial least squares (sPLS) regression by using the function *sp/s()*
932 in the package *mixOmics* [53]. In sPLS regression, two matrices are being integrated and both their
933 structures are being modelled. Here, we used variance stabilized ASV abundances as explanatory variables
934 and all continuous clinical and immune parameters as response variables. The method allows multiple
935 response variables. Collinear, and noisy data can be handled by this method as well [67]. We did not limit
936 the number of response variables to be kept for each component (keepY) prior to model calculation. The

937 number of explanatory variables (ASVs) to be kept on each component (keepX) was set to 25 after running
938 the sPLS regression models for each body site with a range of values between 20 and 40 for keepX,
939 showing results robust to keepX. The *perf()* function was used to inform the choice of 3 relevant
940 components. Based on the sPLS regression models for each body site, we then performed hierarchical
941 clustering with the *cim()* function, using the clustering method “complete linkage” and the distance
942 method “Pearson’s correlation”. Thereby, we generated matrices of coefficients indicating correlations
943 between ASV abundances and continuous clinical and immune parameters.
944 Subsequently, we carried out canonical (i.e. bidirectional) correspondence analysis (CCpNA), which is a
945 multivariate constrained ordination method. This method allow us to assess associations of both
946 categorical and continuous clinical and immune parameters to ASV abundances. We included ASVs and
947 variables with a correlation of >0.2 / <-0.2 (oral and nasal data set) or >0.3 / <-0.3 (fecal data set) in the sPLS
948 analysis into the CCpNA, and additionally included categorical variables that could not be included in the
949 sPLS. The method was implemented with the *cca()* function in package *vegan* [68]. It implements a Chi-
950 square transformation of the $\log+1$ transformed ASV count matrix and subsequent weighted linear
951 regression, followed by singular value decomposition. We depicted the CCpNA results as a triplot with plot
952 dimensions corresponding in length to the percentage of variance explained by each axis. At each body
953 site, we identified three clusters of ASVs through hierarchical clustering based on the first three latent
954 dimensions of each sPLS analysis (Figure 6A, and Additional File 2: Figures S5A and S6A). The CCpNA
955 analyses reinforced the cluster separations and additionally provided insight into associations with
956 categorical variables, including patient baseline parameters, the occurrence of infections, antibiotics
957 treatment, and clinical outcomes (Figure 6B, and Additional File 2: Figures S5B and S6B).
958 We compared bacterial alpha diversity and community composition in the gut of HSCT patients at
959 preexamination with that of healthy children. Alpha diversity (inverse Simpson index) between the two
960 groups was compared by a Kruskal-Wallis test. Community composition was visualized in a principal
961 coordinates analysis (PCoA), and analysis of similarities (ANOSIM, package *vegan*) was used to assess
962 significant differences in the means of rank dissimilarities between the two groups. DESeq2 was employed
963 for identification of differentially abundant genera among the top 100 most abundant genera with >10
964 total reads [62]. Differences in relative abundance of genera identified as differentially abundant were
965 visualized in a heat tree (package *metacoder*) [69]. Higher taxonomic level differential abundance was
966 assessed by linear discriminant analysis effect size (LEfSe) on centered-log ratio (CLR) transformed data
967 with an LDA cutoff of 4 (package *microbiomeMarker*) [70]. LefSe accounts for the hierarchical structure
968 of bacterial phylogeny, thereby allowing identification of differentially abundant taxa on several
969 taxonomic levels (here kingdom to genus). For additional information see
970 <https://doi.org/10.6084/m9.figshare.13614230>.

971

972

973 **List of abbreviations**

974 AML: Acute myeloid leukemia

975 ASV: Amplicon sequence variant

976 ATG: Anti-thymocyte globulin

977 CCpNA: Canonical correspondence analysis

978 CML: Chronic myeloid leukemia

979 CRP: C-reactive protein
980 CTREE: Conditional inference tree
981 ECLIA: Electrochemiluminescence immunoassays
982 GvL effect: Graft-versus-leukemia effect
983 (a)GvHD: (Acute) graft-versus-host disease
984 HSCT: Hematopoietic stem cell transplantation
985 IDS: Immunodeficiency syndromes
986 IEA: Inherited abnormalities of erythrocyte differentiation or function
987 IMD: Inherited disorders of metabolism
988 LIA: Latex immunoturbidimetric assay
989 LDA: Linear discriminant analysis
990 LogitBoost: Boosted logistic regression
991 LOOCV: Leave-one-out cross validation
992 MDS: Myelodysplastic or myeloproliferative disorders
993 MM: Multiple myeloma
994 NHL: Non-Hodgkin lymphomas
995 NPMLE: Non-parametric maximum likelihood estimator
996 OL: Other leukemia
997 OTU: Operational taxonomic unit
998 PBMC: Peripheral blood mononuclear cell
999 PCoA: Principal Coordinates Analysis
1000 Rf: Random forest
1001 SAA: Severe aplastic anemia
1002 SCFA: Short-chain fatty acid
1003 sPLS: Sparse partial least squares analysis
1004 svmLinear: Support vector machines with linear kernel
1005 svmRadial: Support vector machines with radial basis function kernel
1006 TBI: Total body irradiation
1007 T_H17 cell: T helper 17 cell
1008 T_{reg} cell: T regulatory cell
1009 UCB: Umbilical cord blood

1010

1011 **Declarations**

1012

1013 **Acknowledgements**

1014 We thank the patients and their families for their participation in this study. We also thank Marlene
1015 Danner Dalgaard and Neslihan Bicen (Multi Assay Core facility (DMAC), Technical University of Denmark)
1016 for library construction and sequencing. Sequence pre-processing described in this paper was performed
1017 using the DeiC National Life Science Supercomputer at DTU. Furthermore, we would like to thank Patrick
1018 Murigu Kamau Njage (Technical University of Denmark) for helpful discussions related to machine
1019 learning models.

1020

1021 **Author's contributions**

1022 A.C.I., K.K., K.G.M., and S.J.P. designed the research; A.C.I., K.K., H.M, M.I., and S.J.P. performed the
1023 research; A.C.I., and S.J.P. contributed analytic tools; A.C.I., and S.J.P. analyzed the data; A.C.I. and S.J.P.
1024 wrote the manuscript; and K.K., M.I., F.M.A., and K.G.M. edited the manuscript.

1025

1026 **Funding**

1027 This work was supported by the European Union's Framework program for Research and Innovation,
1028 Horizon2020 (643476), and by the National Food Institute, Technical University of Denmark.

1029

1030 **Availability of data and materials**

1031 The 16S rRNA gene sequences are available through the European Nucleotide Archive (ENA) at the
1032 European Bioinformatics Institute (EBI) under accession number PRJEB30894. The datasets generated
1033 and/or analysed in this study as well as the R code used to analyze the data are available from the figshare
1034 repository [https://figshare.com/projects/Microbiota_long-](https://figshare.com/projects/Microbiota_long-term_dynamics_and_prediction_of_acute_graft-versus-host-disease_in_allogeneic_stem_cell_transplantation/80366)
1035 [term_dynamics_and_prediction_of_acute_graft-versus-host-](https://figshare.com/projects/Microbiota_long-term_dynamics_and_prediction_of_acute_graft-versus-host-disease_in_allogeneic_stem_cell_transplantation/80366)
1036 [disease_in_allogeneic_stem_cell_transplantation/80366](https://figshare.com/projects/Microbiota_long-term_dynamics_and_prediction_of_acute_graft-versus-host-disease_in_allogeneic_stem_cell_transplantation/80366) (see also individual links in the Methods
1037 section).

1038

1039 **Ethics approval and consent to participate**

1040 Written informed consent was obtained from the patients and/or their legal guardians. The study protocol
1041 was approved by the local ethics committee (H-7-2014-016) and the Danish Data Protection Agency.

1042

1043 **Consent for publication**

1044 Not applicable.

1045

1046 **Competing interests**

1047 The authors declare that they have no competing interests.

1048 **References**

- 1049 1. Chabannon C, Kuball J, Bondanza A, Dazzi F, Pedrazzoli P, Toubert A, et al. Hematopoietic stem cell
1050 transplantation in its 60s: A platform for cellular therapies. *Sci Transl Med* [Internet]. American
1051 Association for the Advancement of Science; 2018 [cited 2018 Aug 2];10:eaap9630. Available from:
1052 <http://www.ncbi.nlm.nih.gov/pubmed/29643233>
- 1053 2. Shono Y, van den Brink MRM. Gut microbiota injury in allogeneic haematopoietic stem cell
1054 transplantation. *Nat Rev Cancer* [Internet]. Nature Publishing Group; 2018 [cited 2018 Feb 21]; Available
1055 from: <http://www.nature.com/doi/10.1038/nrc.2018.10>
- 1056 3. Holler E, Butzhammer P, Schmid K, Hundsrucker C, Koestler J, Peter K, et al. Metagenomic Analysis of
1057 the Stool Microbiome in Patients Receiving Allogeneic Stem Cell Transplantation: Loss of Diversity Is
1058 Associated with Use of Systemic Antibiotics and More Pronounced in Gastrointestinal Graft-versus-Host
1059 Disease. *Biol Blood Marrow Transplant* [Internet]. 2014 [cited 2015 Oct 22];20:640–5. Available from:
1060 <http://www.sciencedirect.com/science/article/pii/S1083879114000755>
- 1061 4. Ingham AC, Kielsen K, Cilieborg MS, Lund O, Holmes S, Aarestrup FM, et al. Specific gut microbiome
1062 members are associated with distinct immune markers in pediatric allogeneic hematopoietic stem cell
1063 transplantation. *Microbiome* [Internet]. BioMed Central; 2019 [cited 2019 Sep 18];7:131. Available from:
1064 <https://microbiomejournal.biomedcentral.com/articles/10.1186/s40168-019-0745-z>
- 1065 5. Rivera-Chávez F, Lopez CA, Bäumlér AJ. Oxygen as a driver of gut dysbiosis. *Free Radic Biol Med*
1066 [Internet]. Pergamon; 2017 [cited 2018 Feb 18];105:93–101. Available from:
1067 <https://www.sciencedirect.com/science/article/pii/S0891584916304361?via%3Dihub>
- 1068 6. Taur Y, Xavier JB, Lipuma L, Ubeda C, Goldberg J, Gobourne a., et al. Intestinal Domination and the
1069 Risk of Bacteremia in Patients Undergoing Allogeneic Hematopoietic Stem Cell Transplantation. *Clin
1070 Infect Dis* [Internet]. 2012;55:905–14. Available from:
1071 <http://cid.oxfordjournals.org/lookup/doi/10.1093/cid/cis580>
- 1072 7. Ghimire S, Weber D, Mavin E, Wang X nong, Dickinson AM, Holler E. Pathophysiology of GvHD and
1073 Other HSCT-Related Major Complications. *Front Immunol* [Internet]. Frontiers; 2017 [cited 2018 Oct
1074 19];8:79. Available from: <http://journal.frontiersin.org/article/10.3389/fimmu.2017.00079/full>
- 1075 8. Golob JL, Pergam SA, Srinivasan S, Fiedler TL, Liu C, Garcia K, et al. Stool Microbiota at Neutrophil
1076 Recovery Is Predictive for Severe Acute Graft vs Host Disease After Hematopoietic Cell Transplantation.
1077 *Clin Infect Dis* [Internet]. Oxford University Press; 2017 [cited 2018 Nov 23];65:1984–91. Available from:
1078 <https://academic.oup.com/cid/article/65/12/1984/4085173>
- 1079 9. Jenq RR, Taur Y, Devlin SM, Ponce DM, Goldberg JD, Ahr KF, et al. Intestinal *Blautia* Is Associated with
1080 Reduced Death from Graft-versus-Host Disease. *Biol Blood Marrow Transplant* [Internet]. 2015 [cited
1081 2016 May 9];21:1373–83. Available from:
1082 <http://www.sciencedirect.com/science/article/pii/S1083879115002931>
- 1083 10. Han L, Zhao K, Li Y, Han H, Zhou L, Ma P, et al. A gut microbiota score predicting acute graft-versus-
1084 host disease following myeloablative allogeneic hematopoietic stem cell transplantation. *Am J
1085 Transplant*. 2020;20:1014–27.

- 1086 11. Peled JU, Gomes ALC, Devlin SM, Littmann ER, Taur Y, Sung AD, et al. Microbiota as predictor of
1087 mortality in allogeneic hematopoietic-cell transplantation. *N Engl J Med.* 2020;382:822–34.
- 1088 12. Stein-Thoeringer CK, Nichols KB, Lazrak A, Docampo MD, Slingerland AE, Slingerland JB, et al. Lactose
1089 drives *Enterococcus* expansion to promote graft-versus-host disease. *Science* (80-). 2019;366:1143–9.
- 1090 13. Honda K, Littman DR. The microbiota in adaptive immune homeostasis and disease. *Nature*
1091 [Internet]. Nature Publishing Group; 2016 [cited 2018 Aug 21];535:75–84. Available from:
1092 <http://www.nature.com/articles/nature18848>
- 1093 14. Atarashi K, Tanoue T, Oshima K, Suda W, Nagano Y, Nishikawa H, et al. Treg induction by a rationally
1094 selected mixture of Clostridia strains from the human microbiota. *Nature* [Internet]. Nature Publishing
1095 Group; 2013 [cited 2018 Sep 6];500:232–6. Available from:
1096 <http://www.nature.com/articles/nature12331>
- 1097 15. Kielsen K, Ryder LP, Lennox-Hvenekilde D, Gad M, Nielsen CH, Heilmann C, et al. Reconstitution of
1098 Th17, Tc17 and Treg cells after paediatric haematopoietic stem cell transplantation: Impact of
1099 interleukin-7. *Immunobiology* [Internet]. 2018 [cited 2018 Feb 7];223:220–6. Available from:
1100 <http://www.ncbi.nlm.nih.gov/pubmed/29033080>
- 1101 16. Han L, Jin H, Zhou L, Zhang X, Fan Z, Dai M, et al. Intestinal Microbiota at Engraftment Influence
1102 Acute Graft-Versus-Host Disease via the Treg/Th17 Balance in Allo-HSCT Recipients. *Front Immunol*
1103 [Internet]. 2018 [cited 2018 May 17];9:669. Available from:
1104 <http://www.ncbi.nlm.nih.gov/pubmed/29740427>
- 1105 17. Ratajczak P, Janin A, Peffault de Latour R, Leboeuf C, Desveaux A, Keyvanfar K, et al. Th17/Treg ratio
1106 in human graft-versus-host disease. *Blood* [Internet]. American Society of Hematology; 2010 [cited 2018
1107 Nov 19];116:1165–71. Available from: <http://www.ncbi.nlm.nih.gov/pubmed/20484086>
- 1108 18. Larsen JM. The immune response to *Prevotella* bacteria in chronic inflammatory disease.
1109 *Immunology* [Internet]. Wiley/Blackwell (10.1111); 2017 [cited 2018 Nov 25];151:363–74. Available
1110 from: <http://doi.wiley.com/10.1111/imm.12760>
- 1111 19. De Pietri S, Ingham AC, Frandsen TL, Rathe M, Krych L, Castro-Mejía JL, et al. Gastrointestinal toxicity
1112 during induction treatment for childhood acute lymphoblastic leukemia: The impact of the gut
1113 microbiota. *Int J Cancer.* Wiley-Liss Inc.; 2020;147:1953–62.
- 1114 20. Weber D, Oefner PJ, Hiergeist A, Koestler J, Gessner A, Weber M, et al. Low urinary indoxyl sulfate
1115 levels early after transplantation reflect a disrupted microbiome and are associated with poor outcome.
1116 *Blood* [Internet]. 2015 [cited 2015 Oct 22];126:1723–8. Available from:
1117 <http://www.bloodjournal.org/content/126/14/1723>
- 1118 21. Taur Y, Jenq RR, Perales M, Littmann ER, Morjaria S, Ling L, et al. The effects of intestinal tract
1119 bacterial diversity on mortality following allogeneic hematopoietic stem cell transplantation.
1120 *Transplantation* [Internet]. 2014;124:1174–82. Available from:
1121 <http://www.bloodjournal.org/content/bloodjournal/124/7/1174.full.pdf?sso-checked=true>
- 1122 22. Andermann TM, Peled JU, Ho C, Reddy P, Riches M, Storb R, et al. The Microbiome and
1123 Hematopoietic Cell Transplantation: Past, Present, and Future. *Biol Blood Marrow Transplant* [Internet].

- 1124 Elsevier; 2018 [cited 2018 May 29]; Available from:
1125 <https://www.sciencedirect.com/science/article/pii/S1083879118300879?via%3Dihub>
- 1126 23. Jenq RR, Ubeda C, Taur Y, Menezes CC, Khanin R, Dudakov J a., et al. Regulation of intestinal
1127 inflammation by microbiota following allogeneic bone marrow transplantation. *J Exp Med* [Internet].
1128 2012;209:903–11. Available from: <http://www.jem.org/cgi/doi/10.1084/jem.20112408>
- 1129 24. Verma D, Garg PK, Dubey AK. Insights into the human oral microbiome. *Arch Microbiol* [Internet].
1130 Springer Berlin Heidelberg; 2018;200:525–40. Available from: [http://dx.doi.org/10.1007/s00203-018-](http://dx.doi.org/10.1007/s00203-018-1505-3)
1131 1505-3
- 1132 25. Osakabe L, Utsumi A, Saito B, Okamatsu Y, Kinouchi H, Nakamaki T, et al. Influence of Oral Anaerobic
1133 Bacteria on Hematopoietic Stem Cell Transplantation Patients: Oral Mucositis and General Condition.
1134 *Transplant Proc* [Internet]. Elsevier; 2017 [cited 2018 Mar 12];49:2176–82. Available from:
1135 <http://www.ncbi.nlm.nih.gov/pubmed/29149979>
- 1136 26. Soga Y, Maeda Y, Ishimaru F, Tanimoto M, Maeda H, Nishimura F, et al. Bacterial substitution of
1137 coagulase-negative staphylococci for streptococci on the oral mucosa after hematopoietic cell
1138 transplantation. *Support Care Cancer*. 2011;19:995–1000.
- 1139 27. Olczak-Kowalczyk D, Daszkiewicz M, Krasuska-Sławińska, Dembowska-Bagińska B, Gozdowski D,
1140 Daszkiewicz P, et al. Bacteria and Candida yeasts in inflammations of the oral mucosa in children with
1141 secondary immunodeficiency. *J Oral Pathol Med*. 2012;41:568–76.
- 1142 28. Chung H, Pamp SJ, Hill JA, Surana NK, Edelman SM, Troy EB, et al. Gut immune maturation depends
1143 on colonization with a host-specific microbiota. *Cell* [Internet]. Elsevier; 2012 [cited 2018 Aug
1144 16];149:1578–93. Available from: <http://www.ncbi.nlm.nih.gov/pubmed/22726443>
- 1145 29. Mathewson ND, Jenq R, Mathew A V, Koenigsknecht M, Hanash A, Toubai T, et al. Gut microbiome-
1146 derived metabolites modulate intestinal epithelial cell damage and mitigate graft-versus-host disease.
1147 *Nat Immunol* [Internet]. NIH Public Access; 2016 [cited 2018 May 15];17:505–13. Available from:
1148 <http://www.ncbi.nlm.nih.gov/pubmed/26998764>
- 1149 30. Kim M, Qie Y, Park J, Kim CH. Gut Microbial Metabolites Fuel Host Antibody Responses. *Cell Host*
1150 *Microbe* [Internet]. Elsevier Inc.; 2016;20:202–14. Available from:
1151 <http://dx.doi.org/10.1016/j.chom.2016.07.001>
- 1152 31. Shono Y, Docampo MD, Peled JU, Perobelli SM, Velardi E, Tsai JJ, et al. Increased GVHD-related
1153 mortality with broad-spectrum antibiotic use after allogeneic hematopoietic stem cell transplantation in
1154 human patients and mice. *Sci Transl Med* [Internet]. 2016 [cited 2016 May 23];8:339ra71-339ra71.
1155 Available from: <http://stm.sciencemag.org/content/8/339/339ra71>
- 1156 32. Weber D, Jenq RR, Peled JU, Taur Y, Hiergeist A, Koestler J, et al. Microbiota Disruption Induced by
1157 Early Use of Broad Spectrum Antibiotics is an Independent Risk Factor of Outcome after Allogeneic Stem
1158 Cell Transplantation [Internet]. *Biol. Blood Marrow Transplant*. Elsevier Inc.; 2017. Available from:
1159 <http://linkinghub.elsevier.com/retrieve/pii/S1083879117302756>
- 1160 33. Weber D, Hiergeist A, Weber M, Dettmer K, Wolff D, Hahn J, et al. Detrimental effect of broad-
1161 spectrum antibiotics on intestinal microbiome diversity in patients after allogeneic stem cell

- 1162 transplantation: Lack of commensal sparing antibiotics. *Clin Infect Dis* [Internet]. 2018 [cited 2018 Sep
1163 20]; Available from: <http://www.ncbi.nlm.nih.gov/pubmed/30124813>
- 1164 34. Liu C, Frank DN, Horch M, Chau S, Ir D, Horch EA, et al. Associations between acute gastrointestinal
1165 GvHD and the baseline gut microbiota of allogeneic hematopoietic stem cell transplant recipients and
1166 donors. *Bone Marrow Transplant Adv online Publ* [Internet]. 2017;doi:1–8. Available from:
1167 <https://www.nature.com/bmt/journal/vaop/ncurrent/pdf/bmt2017200a.pdf>
- 1168 35. Biagi E, Zama D, Nastasi C, Consolandi C, Fiori J, Rampelli S, et al. Gut microbiota trajectory in
1169 pediatric patients undergoing hematopoietic SCT. *Bone Marrow Transplant* [Internet]. Nature Publishing
1170 Group; 2015 [cited 2018 Jul 2];50:992–8. Available from: <http://www.nature.com/articles/bmt201516>
- 1171 36. Mancini N, Greco R, Pasciuta R, Barbanti MC, Pini G, Morrow OB, et al. Enteric Microbiome Markers
1172 as Early Predictors of Clinical Outcome in Allogeneic Hematopoietic Stem Cell Transplant: Results of a
1173 Prospective Study in Adult Patients. *Open Forum Infect Dis* [Internet]. Oxford University Press; 2017
1174 [cited 2018 Dec 8];4. Available from:
1175 <http://academic.oup.com/ofid/article/doi/10.1093/ofid/ofx215/4367678>
- 1176 37. Glucksberg H, Storb R, Fefer A, Buckner CD, Neiman PE, Clift RA, et al. Clinical manifestations of
1177 graft-versus-host disease in human recipients of marrow from HL-A-matched sibling donors.
1178 *Transplantation* [Internet]. 1974;18:295–304. Available from:
1179 <http://www.ncbi.nlm.nih.gov/pubmed/4153799>
- 1180 38. Knudsen BE, Bergmark L, Munk P, Lukjancenko O, Prieme A, Aarestrup FM, et al. Impact of Sample
1181 Type and DNA Isolation Procedure on Genomic Inference of Microbiome Composition. *bioRxiv*
1182 [Internet]. 2016;1:064394. Available from: <http://biorxiv.org/lookup/doi/10.1101/064394>
- 1183 39. 16S Metagenomic Sequencing Library Preparation. [Internet]. [cited 2018 Apr 17]. Available from:
1184 https://support.illumina.com/content/dam/illumina-support/documents/documentation/chemistry_documentation/16s/16s-metagenomic-library-prep-guide-15044223-b.pdf
- 1187 40. Klindworth A, Pruesse E, Schweer T, Peplies J, Quast C, Horn M, et al. Evaluation of general 16S
1188 ribosomal RNA gene PCR primers for classical and next-generation sequencing-based diversity studies.
1189 *Nucleic Acids Res* [Internet]. Oxford University Press; 2013 [cited 2018 Apr 17];41:e1–e1. Available from:
1190 <http://academic.oup.com/nar/article/41/1/e1/1164457/Evaluation-of-general-16S-ribosomal-RNA-gene-PCR>
1191
- 1192 41. Martin M. Cutadapt removes adapter sequences from high-throughput sequencing reads.
1193 *EMBnet.journal* [Internet]. 2011 [cited 2018 Jun 26];17:10. Available from:
1194 <http://journal.embnet.org/index.php/embnetjournal/article/view/200>
- 1195 42. Callahan BJ, McMurdie PJ, Rosen MJ, Han AW, Johnson AJA, Holmes SP. DADA2: High-resolution
1196 sample inference from Illumina amplicon data. *Nat Methods* [Internet]. 2016 [cited 2016 Jul 28];13:581–
1197 3. Available from: <http://www.nature.com/nmeth/journal/v13/n7/full/nmeth.3869.html>
- 1198 43. Ewels P, Magnusson M, Lundin S, Källner M. MultiQC: summarize analysis results for multiple tools
1199 and samples in a single report. *Bioinformatics* [Internet]. Oxford University Press; 2016 [cited 2018 Sep
1200 11];32:3047–8. Available from: <https://academic.oup.com/bioinformatics/article->

- 1201 [lookup/doi/10.1093/bioinformatics/btw354](https://doi.org/10.1093/bioinformatics/btw354)
- 1202 44. Callahan B. Silva taxonomic training data formatted for DADA2 (Silva version 132). 2018 [cited 2018
1203 Jun 26]; Available from: <https://doi.org/10.5281/zenodo.1172783#.WzJRh15uQOA.mendeley>
- 1204 45. McMurdie PJ, Holmes S. phyloseq: An R Package for Reproducible Interactive Analysis and Graphics
1205 of Microbiome Census Data. Watson M, editor. PLoS One [Internet]. Public Library of Science; 2013
1206 [cited 2018 Jan 24];8:e61217. Available from: <http://dx.plos.org/10.1371/journal.pone.0061217>
- 1207 46. Davis NM, Proctor DM, Holmes SP, Relman DA, Callahan BJ. Simple statistical identification and
1208 removal of contaminant sequences in marker-gene and metagenomics data. Microbiome [Internet].
1209 BioMed Central; 2018 [cited 2018 Dec 27];6:226. Available from:
1210 <https://microbiomejournal.biomedcentral.com/articles/10.1186/s40168-018-0605-2>
- 1211 47. Callahan BJ, Sankaran K, Fukuyama JA, McMurdie PJ, Holmes SP. Bioconductor workflow for
1212 microbiome data analysis: from raw reads to community analyses. F1000Research [Internet].
1213 2016;5:1492. Available from:
1214 <http://www.ncbi.nlm.nih.gov/pubmed/27508062>
1215 <http://www.pubmedcentral.nih.gov/articlerender.fcgi?artid=PMC4955027>
- 1216 48. Wright ES. Using DECIPHER v2.0 to Analyze Big Biological Sequence Data in R. R J. 2016;8:352–9.
- 1217 49. Schliep KP. phangorn: phylogenetic analysis in R. Bioinformatics [Internet]. Oxford University Press;
1218 2011 [cited 2018 Nov 27];27:592–3. Available from: <http://www.ncbi.nlm.nih.gov/pubmed/21169378>
- 1219 50. Gentleman R, Carey V, Huber W, Hahne F. genefilter: methods for filtering genes from microarray
1220 experiments. R package version 1.58.1. 2017.
- 1221 51. R Core Team. R: A Language and Environment for Statistical Computing [Internet]. R Foundation for
1222 Statistical Computing, Vienna, Austria; 2018. Available from: <https://www.r-project.org/>
- 1223 52. Wickham H. ggplot2: Elegant Graphics for Data Analysis. Springer Verlag New York; 2016.
- 1224 53. Rohart F, Gautier B, Singh A, Lê Cao K-A. mixOmics: An R package for ‘omics feature selection and
1225 multiple data integration. Schneidman D, editor. PLOS Comput Biol [Internet]. Public Library of Science;
1226 2017 [cited 2017 Dec 11];13:e1005752. Available from: <http://dx.plos.org/10.1371/journal.pcbi.1005752>
- 1227 54. Fukuyama J. treeDA: Tree-Based Discriminant Analysis. [Internet]. 2017. Available from:
1228 <https://github.com/jfukuyama/treedda>
- 1229 55. Kuhn M, Wing J, Weston S, Williams A, Keefer C, Engelhardt A, et al. caret: Classification and
1230 Regression Training. R package version 6.0-80. [Internet]. 2018. Available from: [https://cran.r-
1231 project.org/package=caret](https://cran.r-project.org/package=caret)
- 1232 56. Hothorn T, Zeileis A, Cheng E, Ong S. partykit: A modular toolkit for recursive partytioning in R. J
1233 Mach Learn Res. 2015;16:3905–9.
- 1234 57. Hothorn T, Hornik K, Zeileis A. Unbiased Recursive Partitioning: A Conditional Inference Framework. J
1235 Comput Graph Stat [Internet]. Taylor & Francis; 2006 [cited 2018 Nov 24];15:651–74. Available from:

- 1236 <http://www.tandfonline.com/doi/abs/10.1198/106186006X133933>
- 1237 58. Fukuyama J, Rumker L, Sankaran K, Jeganathan P, Dethlefsen L, Relman DA, et al. Multidomain
1238 analyses of a longitudinal human microbiome intestinal cleanout perturbation experiment. *PLoS Comput*
1239 *Biol* [Internet]. Public Library of Science; 2017 [cited 2018 Mar 2];13:e1005706. Available from:
1240 <http://www.ncbi.nlm.nih.gov/pubmed/28821012>
- 1241 59. Njage PMK, Henri C, Leekitcharoenphon P, Mistou M, Hendriksen RS, Hald T. Machine Learning
1242 Methods as a Tool for Predicting Risk of Illness Applying Next-Generation Sequencing Data. *Risk Anal*
1243 [Internet]. John Wiley & Sons, Ltd (10.1111); 2018 [cited 2018 Dec 24];risa.13239. Available from:
1244 <https://onlinelibrary.wiley.com/doi/abs/10.1111/risa.13239>
- 1245 60. Njage PMK, Leekitcharoenphon P, Hald T. Improving hazard characterization in microbial risk
1246 assessment using next generation sequencing data and machine learning: Predicting clinical outcomes in
1247 shigatoxigenic *Escherichia coli*. *Int J Food Microbiol* [Internet]. Elsevier; 2019 [cited 2018 Dec
1248 24];292:72–82. Available from:
1249 <https://www.sciencedirect.com/science/article/pii/S0168160518308936#f0005>
- 1250 61. Chen J, Zhang L. GMPR: Geometric Mean of Pairwise Ratios. R package version 0.1.3. 2017.
- 1251 62. Love MI, Huber W, Anders S. Moderated estimation of fold change and dispersion for RNA-seq data
1252 with DESeq2. *Genome Biol* [Internet]. BioMed Central; 2014 [cited 2018 Jan 24];15:550. Available from:
1253 <http://genomebiology.biomedcentral.com/articles/10.1186/s13059-014-0550-8>
- 1254 63. McMurdie PJ, Holmes S. Waste Not, Want Not: Why Rarefying Microbiome Data Is Inadmissible.
1255 McHardy AC, editor. *PLoS Comput Biol* [Internet]. 2014 [cited 2016 Apr 13];10:e1003531. Available from:
1256 <http://dx.plos.org/10.1371/journal.pcbi.1003531>
- 1257 64. Kursu MB, Rudnicki WR. Feature Selection with the Boruta Package. *J. Stat. Softw.* 2010. p. 1–13.
- 1258 65. Anderson-Bergman C. **icenReg** : Regression Models for Interval Censored Data in R. *J Stat Softw*
1259 [Internet]. 2017;81. Available from: <http://www.jstatsoft.org/v81/i12/>
- 1260 66. Stekhoven DJ, Buhlmann P. MissForest--non-parametric missing value imputation for mixed-type
1261 data. *Bioinformatics* [Internet]. Oxford University Press; 2012 [cited 2018 Sep 18];28:112–8. Available
1262 from: <https://academic.oup.com/bioinformatics/article-lookup/doi/10.1093/bioinformatics/btr597>
- 1263 67. Lee D, Lee W, Lee Y, Pawitan Y. Sparse partial least-squares regression and its applications to high-
1264 throughput data analysis. *Chemom Intell Lab Syst* [Internet]. Elsevier; 2011 [cited 2018 Jan 24];109:1–8.
1265 Available from: <https://www.sciencedirect.com/science/article/pii/S016974391100150X>
- 1266 68. Oksanen J, Blanchet FG, Friendly M, Kindt R, Legendre P, McGlinn D, et al. **vegan**: Community
1267 Ecology Package. R package version 2.5-2. [Internet]. 2018. Available from: [https://cran.r-](https://cran.r-project.org/package=vegan)
1268 [project.org/package=vegan](https://cran.r-project.org/package=vegan)
- 1269 69. Foster ZSL, Shapton TJ, Grünwald NJ. Metacoder: An R package for visualization and manipulation of
1270 community taxonomic diversity data. *PLoS Comput Biol.* 2017;13:1–15.
- 1271 70. Cao Y. **microbiomeMarker**: microbiome biomarker analysis. R package version 0.0.1.9000.

1272 <https://github.com/yiluheihei/microbiomeMarker>. 2021.

1273 **Figure and Table Legends**

1274

1275 **Figure 1. Monitoring gut, oral, and nasal microbiota and the host immune system in allogeneic**
1276 **hematopoietic stem cell transplantation (HSCT).** A) Twenty-nine children were monitored before, at the
1277 time of, and immediately post allogeneic HSCT, as well as at late follow-up time points. Patients' baseline
1278 characteristics, clinical outcomes, as well as immune cell counts, and inflammation and infection markers
1279 over time were monitored. Patient characteristics are described in detail in Table S1 (Additional File 1).
1280 Host immune system parameters were related to longitudinal dynamics of the gut, oral, and nasal
1281 microbiota that was assessed at the denoted time points. B) Bacterial alpha diversity before, at the time
1282 of, and after HSCT at each body site, displayed on a log₁₀ transformed y-axis for visualization purposes.
1283 Asterisks indicate significant differences in median inverse Simpson index between time points * $P < 0.05$.
1284 C) Tree-based sparse linear discriminant (LDA) analyses by time point in relation to HSCT. For fecal
1285 samples, positive LDA scores were observed for samples collected immediately post HSCT. For both oral
1286 and nasal samples, positive LDA scores were observed for samples from before HSCT and from late follow
1287 up-time points.

1288

1289 **Figure 2. Temporal microbial community dynamics in the gut.** A) Relative abundances over time of the
1290 12 most abundant families in the gut. B) Tree-based sparse linear discriminant analysis (LDA). Coefficients
1291 of discriminating clades of ASVs on the first LDA axis, colored by taxonomic family, and plotted along the
1292 phylogenetic tree. C) Trajectories of ASVs affiliated with the families *Enterococcaceae* and
1293 *Lactobacillaceae*, with increasing abundances after HSCT. The most abundant discriminating ASV for each
1294 family is indicated. D) Trajectories of ASVs affiliated with the families *Lachnospiraceae* and
1295 *Ruminococcaceae*, with decreasing abundances after HSCT and recovery at late follow-up time points. The
1296 most abundant discriminating ASV for *Blautia* spp. is indicated. Detailed taxonomic information and LDA-
1297 coefficients of the displayed ASVs are listed in Additional File 1: Table S2.

1298

1299 **Figure 3. Temporal microbial community dynamics in the oral cavity.** A) Relative abundances over time
1300 of the 12 most abundant families in the oral cavity. B) Tree-based sparse linear discriminant analysis (LDA).
1301 Coefficients of discriminating clades of ASVs on the first LDA axis, colored by taxonomic family, and plotted
1302 along the phylogenetic tree. C) Trajectories of ASVs affiliated with the families *Actinomycetaceae*,
1303 *Streptococcaceae*, *Prevotellaceae*, and Family XI (Class *Bacillales*), with decreasing abundances after HSCT
1304 and recovery at late follow-up time points. The most abundant discriminating ASV for each family is
1305 indicated. Detailed taxonomic information and LDA-coefficients of the displayed ASVs are listed in
1306 Additional File 1: Table S2.

1307

1308 **Figure 4. Machine learning-based prediction of aGvHD severity from the pre-HSCT gut microbiota**
1309 **composition.** A) Relative abundances of the 12 most abundant families over time in the gut in patients
1310 with aGvHD grade 0-I versus II-IV. B) Importance plot of top 20 predictive gut ASVs identified by the
1311 svmLinear model with importance scores indicating the mean decrease in prediction accuracy in case the
1312 respective ASV would be excluded from the model. The final cross-validated svmLinear model predicted
1313 aGvHD (0-I versus II-IV) from the abundances of gut ASVs pre-HSCT with 86% accuracy (95% CI: 65% to

1314 97%). The ASVs that were also confirmed by Boruta feature selection are indicated with asterisk. C)
1315 Conditional inference tree (CTREE) displaying ASVs identified as significant split nodes by nonparametric
1316 regression for prediction of aGvHD. Numbers along the branches indicate split values of variance stabilized
1317 bacterial abundances. The terminal nodes show the proportion of samples originating from patients ($n =$
1318 number of samples) with aGvHD grade 0-I vs II-IV. D) Boxplots depicting the log transformed relative
1319 abundances of the predictive ASVs at time points up to the transplantation in aGvHD grade 0-I compared
1320 with grade II-IV patients. E) Trajectories of *Lactobacillaceae* and *Tannerellaceae* ASVs that were identified
1321 by tree-based sparse LDA, including ASV 3 and ASV 128 that were predictive for aGvHD (bold lines), in
1322 patients with aGvHD grade 0-I vs II-IV.

1323
1324 **Figure 5. Machine learning-based prediction of aGvHD severity from the pre-HSCT oral microbiota**
1325 **composition.** A) Relative abundances the 12 most abundant families over time in the oral cavity in patients
1326 with aGvHD grade 0-I versus II-IV. B) Importance plot of top 20 predictive oral ASVs identified by the
1327 svmLinear model with importance scores indicating the mean decrease in prediction accuracy in case the
1328 respective ASV would be excluded from the model. The final cross-validated svmLinear model predicted
1329 aGvHD (0-I versus II-IV) from the abundances of oral ASVs pre-HSCT with 92% accuracy (95% CI: 73% to
1330 99%). The ASVs that were also confirmed by Boruta feature selection are indicated with asterisk. C)
1331 Conditional inference tree (CTREE) displaying ASVs identified as significant split nodes by nonparametric
1332 regression for prediction of aGvHD. Numbers along the branches indicate split values of variance stabilized
1333 bacterial abundances. The terminal nodes show the proportion of samples originating from patients ($n =$
1334 number of represented samples) with aGvHD grade 0-I vs II-IV. D) Boxplots depict the log transformed
1335 relative abundances of the predictive ASVs at time points up to the transplantation in aGvHD grade 0-I
1336 compared with grade II-IV patients. E) Trajectories of *Prevotellaceae* and *Actinomycetaceae* ASVs that
1337 were identified by tree-based sparse LDA, including ASV 226 and ASV 568 that were predictive for aGvHD
1338 (bold lines), in patients with aGvHD grade 0-I vs II-IV.

1339
1340 **Figure 6. Multivariate associations of the gut microbiota with immune and clinical parameters in HSCT.**
1341 A) Clustered image map (CIM) based on sparse partial least squares (sPLS) regression analysis (dimensions
1342 1, 2, and 3) displaying pairwise correlations $>0.3/<-0.3$ between ASVs (bottom) and continuous immune
1343 and clinical parameters (right). Red indicates a positive correlation, and blue indicates a negative
1344 correlation, respectively. Based on the sPLS regression model, hierarchical clustering (clustering method:
1345 complete linkage, distance method: Pearson's correlation) was performed resulting in the three depicted
1346 clusters. B) Canonical correspondence analysis (CCpNA) relating gut microbial abundances (circles) to
1347 continuous (arrows) and categorical (+) immune and clinical parameters. ASVs and variables with at least
1348 one correlation $>0.3/<-0.3$ in the sPLS analysis were included in the CCpNA. The triplot shows variables
1349 and ASVs with a score $>0.3/<-0.3$ on at least one of the first three CCpNA axes, displayed on axis 1 versus
1350 2 with samples depicted as triangles. The colored ellipses (depicted with 80% confidence interval)
1351 correspond to the clusters of ASVs identified by the sPLS-based hierarchical clustering. Abbreviations not
1352 mentioned in text: ATGmm, anti-thymocyte globulin; B_, blood; BU, busulfan; CY, Cyclophosphamide;
1353 DonorMatch6, matched unrelated donor; FLU_other, fludarabine combinations without thiotepa;
1354 GvHD.Prophylaxis1, treatment with cyclosporine; GvHD.Prophylaxis7, treatment with cyclosporine and
1355 methotrexate; immat_B, immature B cells; K_d100, Karnofsky score on day +100; K_pre, Karnofsky score

1356 before HSCT; m1, month+1; m3, month+3; m6, month+6; m12, month+12; mat_B, mature B cells; MEL,
1357 melphalan; total_B, total B cells; P_, plasma; parasitic, parasitic infection; pre_cond, before conditioning
1358 start; pre_exam, pre-examination; THIO, thiotepa; viral, viral infection; VP16, Etoposide.

1359
1360 **Supplementary Table S1. Patient characteristics.** Abbreviations: HLA, human leukocyte antigen; TBI, total
1361 body irradiation; CY, Cyclophosphamide; VP16, Etoposide; BU, Busulfan; MEL, Melphalan; GvHD, graft-
1362 versus-host disease.

1363
1364 **Supplementary Table S2. Taxonomy of a subset of LDA clade members and corresponding LDA-**
1365 **coefficients in the gut, oral cavity, and nasal cavity.**

1366
1367 **Supplementary Table S3. Taxonomy of aGvHD predictors within the fecal, oral, and nasal microbiota.**
1368 ASVs that were significantly predicting aGvHD severity according to the conditional inference tree
1369 regression model are highlighted in bold. Of the 50 most important gut ASVs identified by the svmLinear
1370 model, 17 were confirmed by Boruta feature selection and are listed here. In the oral and nasal cavities,
1371 26 and 12 ASVs were confirmed by Boruta selection, respectively. Listed in bold are those ASVs with a
1372 significant predictive effect on aGvHD severity, tested in a regression framework with CTREE (see
1373 Methods).

1374
1375 **Figure S1. The gut microbiota in the HSCT patients at pre-exam differs from the gut microbiota of age-**
1376 **matched healthy children.** A) Fecal bacterial alpha diversity (inverse Simpson index) was 2.4-fold higher
1377 in healthy children (n=18) compared to children at pre-examination before HSCT (n=15). B) Fecal bacterial
1378 composition was significantly different between the two groups (anosim, p=0.001, R=0.44), and within-
1379 group variance was significantly greater in the HSCT group (betadisper, p<0.001). C) The taxa which best
1380 explain differences in community structure between HSCT patients at preexamination and healthy
1381 children were identified by analysis of LEfSe (Linear discriminant analysis Effect Size). LEfSe accounts for
1382 the hierarchical structure of bacterial phylogeny, thereby allowing identification of differentially abundant
1383 taxa on several taxonomic levels (here: kingdom to genus). Count data was centered-log ratio (CLR)
1384 transformed within the LEfSe analysis. The higher the LDA score (log10), the higher the effect size of the
1385 respective taxon in explaining group difference. Here, we show taxa with an LDA score >4. D) Differentially
1386 abundant genera between the two groups were additionally identified by DESeq2. Of the top 100 most
1387 abundant genera (of the whole gut microbiota data set), eighteen genera were significantly more
1388 abundant in healthy children (yellow), and 15 genera were significantly more abundant in the patients at
1389 preexam (purple). Differences in median proportions of these genera (and their supertaxa) are displayed
1390 in a heat tree. See also additional information at <https://doi.org/10.6084/m9.figshare.13614230>.

1391
1392
1393 **Figure S2. Most abundant taxonomic families in the gut, oral cavity, and nasal cavity in allo-HSCT**
1394 **patients.** Rank abundance curves displaying the proportions of the 12 most abundant taxonomic families
1395 at each body site (gut, oral cavity, and nasal cavity).

1396

1397 **Figure S3. Tree-based sparse linear discriminant analysis revealing nasal ASVs that distinguish time**
1398 **points from each other in relation to HSCT.** A) Relative abundances over time of the 12 most abundant
1399 families in the nasal cavity. B) Coefficients of discriminating clades of ASVs on the first LDA axis, colored
1400 by taxonomic family, and plotted along the phylogenetic tree. C) Trajectories of ASVs in one discriminating
1401 group, affiliated with the family *Corynebacteriaceae*, with decreasing abundances after HSCT and recovery
1402 at late follow-up time points. The most abundant discriminating ASV is indicated. Detailed taxonomic
1403 information and LDA-coefficients of the displayed ASVs are listed in Table S2.

1404
1405 **Figure S4. Machine learning-based prediction of aGvHD severity from nasal microbial abundances pre-**
1406 **HSCT.** A) Relative abundances of the 12 most abundant families over time in the nasal cavity in patients
1407 with aGvHD grade 0-I versus II-IV. B) Importance plot of top 20 predictive nasal ASVs identified by the
1408 svmLinear model with importance scores indicating the mean decrease in prediction accuracy in case the
1409 respective ASV would be excluded from the model. The final cross-validated svmLinear model predicted
1410 aGvHD (0-I versus II-IV) from the abundances of nasal ASVs pre-HSCT with 76% accuracy (95% CI: 56% to
1411 90%). The ASVs that were also confirmed by Boruta feature selection are indicated with asterisk. C)
1412 Conditional inference tree (CTREE) displaying ASVs identified as significant split nodes by nonparametric
1413 regression for prediction of aGvHD. Numbers along the branches indicate split values of variance stabilized
1414 bacterial abundances. The terminal nodes show the proportion of samples originating from patients with
1415 aGvHD grade 0-I vs II-IV (n = number of samples). D) Boxplots depict the log transformed relative
1416 abundances of the predictive ASVs at time points up to the transplantation in aGvHD grade 0-I compared
1417 with grade II-IV patients.

1418
1419 **Figure S5. Multivariate associations of the oral microbiota with immune and clinical parameters in HSCT.**
1420 A) Clustered image map (CIM) based on sparse partial least squares (sPLS) regression analysis dimensions
1421 1, 2, and 3, displaying pairwise correlations $>0.2/ <-0.2$ between oral ASVs (bottom), and continuous
1422 immune and clinical parameters (right). Red indicated positive correlation, and blue indicates negative
1423 correlation, respectively. Based on the sPLS regression model, hierarchical clustering (clustering method:
1424 complete linkage, distance method: Pearson's correlation) was performed resulting in the three depicted
1425 clusters. B) Canonical correspondence analysis (CCpNA) relating oral microbial abundances (circles) to
1426 continuous (arrows) and categorical (+) immune and clinical parameters. ASVs and variables with at least
1427 one correlation $>0.2/ <-0.2$ in the sPLS analysis were included in the CCpNA. The triplot shows variables
1428 and ASVs with a score $>0.3/ <-0.3$ on at least one the first three CCpNA axes, displayed on axis 1 versus 2
1429 with samples depicted as triangles. The colored ellipses (depicted with 80% confidence interval)
1430 correspond to the clusters of ASVs identified by the sPLS-based hierarchical clustering. For visualization
1431 purposes, a focused section of the CCpNA triplot is shown. Abbreviations are described in Figure 6.
1432 Additional abbreviations: fungal, fungal infection; haploident, haploidentical donor; hemo, hemoglobin;
1433 leuko, leukocytes; lympho, lymphocytes; w1, week+1; w2, week+2; w3, week+3.

1434
1435 **Figure S6. Multivariate associations of the nasal microbiota with immune and clinical parameters in**
1436 **HSCT.** A) Clustered image map (CIM) based on sparse partial least squares (sPLS) regression analysis
1437 dimensions 1, 2, and 3, displaying pairwise correlations $>0.2/ <-0.2$ between nasal ASVs (bottom), and
1438 continuous immune and clinical parameters (right). Red indicated positive correlation, and blue indicates

1439 negative correlation, respectively. Based on the sPLS regression model, hierarchical clustering (clustering
1440 method: complete linkage, distance method: Pearson's correlation) was performed resulting in the three
1441 depicted clusters. B) Canonical correspondence analysis (CCpNA) relating nasal microbial abundances
1442 (circles) to continuous (arrows) and categorical (+) immune and clinical parameters. ASVs and variables
1443 with at least one correlation >0.2 / <-0.2 in the sPLS analysis were included in the CCpNA. The triplot shows
1444 variables and ASVs with a score >0.3 / <-0.3 on at least one the first three CCpNA axes, displayed on axis 1
1445 versus 2 with samples depicted as triangles. The colored ellipses (depicted with 80% confidence interval)
1446 correspond to the clusters of ASVs identified by the sPLS-based hierarchical clustering. For visualization
1447 purposes, a focused section of the CCpNA triplot is shown. Abbreviations are described in Figures 6 and
1448 S5. Additional abbreviations: DonorMatch8, unrelated donor with 1 HLA mismatch; PB, peripheral blood.