

1 **Title**

2 **Impact of Salivary and Pancreatic Amylase Gene Copy**
3 **Numbers on Diabetes, Obesity, and Functional Profiles of**
4 **Microbiome in Northern Japanese Population**

5
6 **Authors and Affiliations**

7 Takanori Hasegawa^{1*}, Masanori Kakuta², Rui Yamaguchi², Noriaki Sato³, Tatsuya
8 Mikami⁴, Koichi Murashita⁵, Shigeyuki Nakaji⁶, Ken Itoh⁷, Seiya Imoto^{1*}.

9
10 ¹Health Intelligence Center, The Institute of Medical Science, The University of Tokyo,
11 4-6-1 Shirokanedai, Minato-ku, Tokyo, 108-8639, Japan

12 ²Human Genome Center, The Institute of Medical Science, The University of Tokyo,
13 4-6-1 Shirokanedai, Minato-ku, Tokyo, 108-8639, Japan

14 ³Department of Biomedical Data Intelligence, Graduate School of Medicine, Kyoto
15 University, 54 Shogoin Kawahara-cho, Sakyo-ku, Kyoto, 606-8507, Japan

16 ⁴Innovation Center for Health Promotion, Hirosaki University Graduate School of
17 Medicine, 5 Zaifu-cho, Hirosaki, Aomori, 036-8562, Japan

18 ⁵COI Research Initiatives Organization, Hirosaki University, 5 Zaifu-cho, Hirosaki,
19 Aomori, Japan

20 ⁶Department of Social Medicine, Hirosaki University Graduate School of Medicine, 5
21 Zaifu-cho, Hirosaki, Aomori, 036-8562, Japan

22 ⁷Department of Stress Response Science, Hirosaki University Graduate School of
23 Medicine, 5 Zaifu-cho, Hirosaki, Aomori, 036-8562, Japan

24
25 ***Corresponding author**

26 E-mail addresses: t-hasegw@ims.u-tokyo.ac.jp (T. Hasegawa), mkakuta@ims.u-tokyo.ac.jp (M. Kakuta),
27 rui@ims.u-tokyo.ac.jp (R. Yamaguchi), nori5030@kuhp.kyoto-u.ac.jp (N. Sato),
28 tmika@hirosaki-u.ac.jp (T. Mikami), murasita@hirosaki-u.ac.jp (K. Murashita), nakaji@hirosaki-u.ac.jp
29 (S. Nakaji), itohk@hirosaki-u.ac.jp (K. Itoh), imoto@ims.u-tokyo.ac.jp (S. Imoto).

30
31 **Abstract**

32 Amylase genes reside in a structurally complex locus, and their copy numbers vary
33 greatly, especially among agricultural races. Amylase genes seem to shape the
34 metabolic response to dietary starch, and several studies have reported their association
35 with obesity. Besides, the effect of amylase copy numbers seems to depend on lifestyle,
36 and the mechanism of this effect was partially explained by changes in the oral and gut

37 microbiome compositions; however, a detailed mechanism has been unclarified. In this
38 study, we showed their association with diabetes in addition to obesity, and further
39 discovered a plausible mechanism of this association based on the function of
40 commensal bacterial in a northern Japanese population. First, we confirmed that the
41 amylase copy number in the population tends to be larger than that reported in other
42 studies and that there is a positive association between obesity and diabetes ($p = 1.95E-2$
43 and $3.28E-2$). Second, we identified that relative abundance of some genus level
44 microbiome, *Capnocytophaga*, *Dialister*, and previously reported bacteria, were
45 significantly associated with amylase copy numbers. Finally, through functional
46 gene-set analysis using shotgun sequencing, we observed that the abundance of genes in
47 the Acarbose pathway in the gut microbiome was significantly decreased with an
48 increase in the amylase copy number (p -value = $5.80E-4$), which can partly explain the
49 mechanism underlying obesity and diabetes in populations with high amylase copy
50 numbers.

51

52 **Keywords**

53 amylase copy number, whole-genome sequencing, shotgun metagenomic sequencing,
54 obesity, diabetes, acarbose pathway

55

56

57 **Introduction**

58 Similar to hundreds of human genes, the amylase genes reside in a structurally complex
59 locus, with inversions, deletions, and many duplications. Copy number variation (CNV)
60 of AMY genes, which encode amylase enzymes that digest starch into maltose and
61 dextrin, is considered one of the strongest signals of recent natural selection in the
62 human population. There are three amylase genes; the copy number of AMY1
63 (AMY1-CN), which encodes salivary amylase, varies from 1 to more than 20 copies,
64 whereas the copy numbers of AMY2A (AMY2A-CN) and AMY2B (AMY2B-CN),
65 which encode pancreatic amylase, vary from 1 to more than 6 copies. AMY1-CN is
66 known to be positively correlated with salivary amylase activity and hydrolyzes the
67 alpha bonds of starch and glycogen, initiating the process of starch degradation in the
68 mouth. As the variety in the number of these genes is specific to agricultural races
69 among humans, the CNVs of these three genes are hypothesized to shape the metabolic
70 response to dietary starch (1-3).

71

72 In recent decades, several studies have shown that AMY1-CN and salivary amylase
73 concentrations are associated with obesity and diabetes. Further, an increase in
74 AMY1-CN levels is reported to decrease the risk of obesity (4-9). However, some other
75 studies have reported conflicting results and showed no or contrasting associations of
76 AMY1-CN with obesity (10-12). These conflicting results may be due to the
77 heterogeneity in patient samples in terms of genetic background and lifestyle, as well as
78 the use of different study methods, such as the assessment of AMY1-CN. Interestingly,
79 Rukh et al. reported that the degree of dietary starch intake modifies the relationship
80 between copy number variation in the salivary amylase gene and BMI; thus, an increase
81 in AMY1-CN decreases the risk of obesity in the low starch intake group and increases
82 the risk in the high starch intake group (12). Marcovecchio et al. also showed that an
83 increase in AMY1-CN decreases the risk of obesity in boys, but can increase its risk in
84 girls (9). Some other studies have shown that even though AMY1-CN is associated with
85 salivary amylase levels, the salivary amylase level is more associated with obesity and
86 BMI compared to AMY1-CN (13-15).

87

88 Humans with high AMY1-CN (AMY1H), who can produce high levels of salivary
89 amylase, might thus derive more energy from the same amount of diet compared to
90 those with low AMY1-CN (AMY1L). Compared to AMY1H, AMY1L might be
91 expected to harbor gut microbiomes with a greater capacity for the decomposition of
92 carbohydrates. Thus, AMY1 copy number has also recently emerged as a significant
93 factor in shaping the relative abundance of the specific members in the human oral and
94 gut microbiota (16-18). Further, an experimental study elucidated that the types of
95 microbes in AMY1H drive higher adiposity when transferred to germ-free mice (18),
96 but the detailed mechanism has not been clarified.

97

98 In this study, we aimed to investigate the fundamental relationships of AMY1 and
99 AMY2A with diabetes and obesity in a homogenous northern Japanese population
100 considered to have a high starch intake. We first identified AMY1-CN, AMY2A-CN,
101 and AMY2B-CN, and their mutation distributions using whole-genome sequencing data.
102 In many studies, CNV arrays have been used to detect CNVs; however, recent studies
103 suggest that whole-genome sequencing (WGS) analysis of CNVs is likely to be more
104 efficient and outperforms array-based CNV analysis (19). We utilized a methodology
105 that has been developed to accurately estimate CNVs in AMY regions from sequencing
106 data (20, 21). Secondly, we analyzed the association of the copy numbers of AMY
107 genes (AMY-CNs) with obesity and diabetes. We then analyzed the oral and gut

108 microbiome at the genus level to discriminate high and low AMY-CN individuals using
109 16S rRNA gene sequence analysis. Finally, we obtained shotgun whole-genome
110 sequencing data of the oral and gut microbiome to investigate the functional gene sets
111 of starch degradation pathways that can be related to AMY-CN.

112

113 **Results**

114 ***Distribution of AMY1, AMY2A, and AMY2B copy number variation***

115 We first analyzed the WGS data to detect AMY-CN from 1,479 individuals in the 2015
116 and 2016 participants. For estimating AMY-CN, we applied TIDDIT (20) and
117 AMYCNE (21), and then obtained AMY1A, AMY2A, and AMY2B CNs and their
118 relationships, as illustrated in Fig. 1. In contrast to the average AMY1-CN 10.3 in this
119 population, as described in Fig. 1(a), previous studies have reported much smaller
120 average AMY1-CN; for example, the average CN was 7 in a study from Qatar (13), 8.3
121 in a study from Japan (14), 7.9 in a study from the US (18), and 7.0 in a study from
122 Mexico (16). As illustrated in Fig. 1(b), AMY1 and AMY2A CNs were strongly
123 correlated ($r = 0.60$, $p = 2.82E-144$) because their genetic regions are near and on the
124 same linkage disequilibrium (LD) block (11). On the contrary, AMY2B-CN was not
125 associated, and almost all participants had two copies. Thus, we focused on AMY1 and
126 AMY2A CNs in the following experiments.

127

128 ***SNP patterns on AMY1, AMY2A and AMY2B regions***

129 Next, we summarized the SNP distributions in the AMY gene regions as shown in
130 Table 1 and illustrated in Fig. 2. Contrary to the results of AMY CNV, the number of
131 SNPs was clearly larger in the AMY2A and AMY2B regions than in the AMY1 region.
132 However, a previous study reported that several SNPs with a minor allele frequency of
133 more than 0.3, for example, rs42443712, were detected in the AMY1A region (11). In
134 addition, several SNPs discovered in our analysis did not have a reference SNP ID
135 (rsID) assigned to an identical location. It could be that the genetic variety of AMY1A
136 might not be observed because this cohort is from a region of Japan where genetic
137 diversity appears to be low. Notably, even though the effects of these SNPs on BMI and
138 HbA1c were investigated, no significant effect was detected, similar to a previous study
139 (11). Thus, we focused on the effects of AMY-CN in the following sections.

140

141 ***Effect of AMY CN on HbA1c and BMI***

142 Then, we performed a linear regression analysis of AMY1 and AMY2A CNs on HbA1c
143 and BMI with the explained covariates. Therefore, AMY1-CN and AMY2-CN were

144 positively associated with HbA1c and BMI, and their p-values for the first and second
145 criteria of AMY1-CN and AMY2A-CN were 1.34E-2, 1.95E-2, and 2.76E-2 for HbA1c
146 and 4.71E-1, 3.28E-2, and 6.25E-2 for BMI, respectively, as illustrated in Fig3. Using
147 the predefined threshold ($p = 0.05$), we confirmed the statistically significant association
148 of AMY1 and AMY2A CNs with HbA1c. However, only AMY1A in the second
149 criterion tended to be associated with BMI. These results are partially consistent with
150 those of a previous study (12); further, the association with HbA1c in the Japanese
151 population seems to be a novel observation.

152

153 ***Association analysis between genus level microbiome composition and AMY CNs***

154 In this study, we obtained the oral and gut microbiome composition of the participants
155 in 2015, 2016, 2017, and 2018 through 16S rRNA sequencing. Among these, we
156 extracted the data from participants whose WGS was also obtained to calculate
157 AMY-CNs in 2014 or 2015. Considering the exclusion criteria, we selected 451, 487,
158 426, and 483 participants for the oral microbiome and 435, 475, 412, and 465
159 participants for gut microbiome analysis in 2015, 2016, 2017, and 2018, respectively. In
160 the microbiome analysis, we excluded bacteria for which the sum of compositions of all
161 participants was less than 1.0E-5 and (ii) those that were included in less than 3% of the
162 participants each year. We thus retained 90 and 135 genus level bacteria for the oral and
163 gut microbiome analyses, respectively. Thus, their predefined significant level with
164 Bonferroni correction were $0.05 / 90 = 5.56E-4$ and $0.05 / 135 = 3.70E-4$, respectively.
165 Because it was difficult to integrate microbiome composition data over the entire study
166 period, we calculated integrated p-values using fixed-effect meta-analysis based on the
167 estimates and standard errors. The obtained results are listed in Table 2; we extracted
168 two prominent bacteria associated with AMY1-CNs, as illustrated in Fig.4.

169 *Capnocytophaga* in the oral microbiome tended to be negatively associated with
170 AMY1-CN ($p = 1.38E-6$). In addition, the *Dialister* composition in the gut microbiome
171 increased and decreased with the increase in AMY1 CNs for males and females,
172 respectively, and these differences were statistically significant ($p = 1.02E-6$ for males
173 and $6.00E-6$ for females). In particular, *Dialister* has been reported to be associated with
174 carbohydrate metabolism and host diet; we could reproduce these results (18, 22). The
175 top 10 bacteria and their p-values are presented in the supplementary material.

176

177 ***Functional gene abundance analysis***

178 Through the above experiments, AMY-CNs can be associated with HbA1c levels, BMI,
179 and the oral and gut microbiomes. To further investigate the underlying mechanisms, we
180 analyzed the abundance of functional gene sets in the oral and gut microbiome in

181 association with AMY-CNs using shotgun microbiome sequencing. As we obtained data
182 for 2015 and 2016, we considered participants whose WGS was obtained in 2014 and
183 2015. In this experiment, we generated the data by selecting the participants in 2015 and
184 those participants in 2016 who did not participate in 2015, and integrating them into a
185 single table. A total of 617 participants were thus included in this experiment. Because
186 AMY-CNs determine the capacity of amylase secretion, which decomposes starch into
187 dextrin and maltose, we focused on the starch degradation pathway. In the KEGG
188 pathway database (23), we selected starch and sucrose metabolism (KEGG Identifier:
189 map00500), amino sugar and nucleotide sugar metabolism (map00520), and Acarbose
190 and validamycin biosynthesis (map00525) as potentially associated pathways. For this
191 analysis, the total abundance of the gene set in each pathway was calculated for each
192 participant, and a linear regression analysis with predefined covariates was also
193 performed to analyze the association with AMY-CNs. Here, the predefined p-value with
194 Bonferroni correction was $0.05 / 3 = 1.67E-2$. As a result, the total abundance of genes in
195 map0525 was negatively associated with AMY1-CN under the first and second criteria (p
196 = $5.80E-4$ and $3.97E-2$, respectively) and with AMY2A-CN (p -value = $1.15E-3$), as
197 illustrated in Fig.5. Acarbose is an oral hypoglycemic drug used for treating type 2
198 diabetes, and this decrease could increase blood glucose levels.

199

200 **Discussion and Conclusion**

201 Several studies published in the last decade have shown that AMY1-CN is associated
202 with obesity and diabetes. In contrast, several studies have reported that no significant
203 association was observed among them. These conflicting results may have been
204 influenced by the selection of populations and their lifestyles. For example, in people
205 who eat a low starch diet, a high percentage of salivary amylase digests starch quickly
206 and serum glucose levels become quickly and temporarily high at an early stage,
207 whereas a low percentage of amylase can maintain a high serum glucose level. On the
208 contrary, for people who eat large amounts of starch, a high percentage of salivary
209 amylase may result in high glucose levels, whereas a low salivary amylase level may
210 maintain moderate serum glucose levels. For most human beings, grains and rice are a
211 major source of energy. Thus, it is important to elucidate the impact and mechanism of
212 AMY-CNV on obesity and diabetes.

213

214 First, we assumed that the residents of Hirosaki City consume a relatively large amount
215 of starch, even in Japan. According to a survey of rice consumption in two or more
216 households by the Statistics Bureau, Ministry of Internal Affairs and Communications
217 in Japan (<https://www.stat.go.jp/data/kakei/5.html>), the national average is 64.16
218 [kg/year], the consumption in Morioka city, which is close to Hirosaki city, is 71.61
219 [kg/year], whereas that in Tokyo is 53.16 [kg/year]; thus, the high starch consumption
220 in this area might be related to this observation. Further, the AMY1A region has very
221 few SNPs compared to the AMY2A and AMY2B regions. Thus, it seems that AMY1A

222 exhibits diversity by copy number, and the AMY2A and AMY2B express diversity by
223 mutation. However, in previous studies, more frequent SNPs in the AMY1 region have
224 been reported; these results might thus be region-specific.

225

226 Similar to previous studies, we examined the association between AMY1-CN and
227 HbA1c and BMI. The results showed that both HbA1c levels and BMI increased
228 significantly with the increase in AMY1-CN. This is consistent with the finding of a
229 positive correlation between the increase in AMY1-CN and BMI in the high starch
230 intake group (12), but the association with HbA1c levels is a new finding in the
231 Japanese population. Further, because the average AMY1-CN in this study was clearly
232 greater than that in other studies, there could be a positive selection that favored more
233 AMY1-CN in the past.

234

235 AMY-CN is reported to have a strong correlation with amylase secretion, and has a
236 stronger effect on BMI and the oral and gut microbiome. As the amount of amylase
237 secretion was not measured in this study, we examined the association between
238 AMY1-CN and the oral and gut microbiome. Previously, the association of the human
239 microbiome and AMY-CNs and the effect of microbiome translocation with obesity
240 have been reported (16-18). Those studies have reported that *Prevotella* in the oral
241 cavity and *Ruminococcus* in the gut increased with an increase in AMY1-CN. In our
242 experiment, although not statistically significant, similar tendencies were observed. In
243 particular, in this study, we found that *Capnocytophaga* decreased significantly in the
244 oral cavity as AMY1-CN increased. *Capnocytophaga* is known to be associated with
245 periodontal disease and could be affected by the amount of starch decomposition;
246 however, to the best of our knowledge, no study has been published regarding its
247 associations. In addition, *Dialister* increased significantly in men and decreased
248 significantly in women with an increase in AMY1-CN. *Dialister* has been reported to
249 increase in obese women, but we show a strong correlation with AMY1-CN rather than
250 with BMI and HbA1c levels. The amount of *Dialister* may be thus directly affected by
251 the amount of amylase secreted, which could result in obesity and other consequences.

252

253 Finally, we tried to elucidate the mechanism underlying the effect of the oral and gut
254 microbiome on obesity and diabetes from the viewpoint of changes in the abundance of
255 pathway-based functional genesets. The results showed that the total amount of
256 acarbose and validamycin biosynthesis (map00525) decreased significantly with an
257 increase in both AMY1-CN and AMY2A-CN. Acarbose is an oral hypoglycemic drug

258 used for the treatment of type 2 diabetes, which delays starch breakdown to glucose by
259 inhibiting the action of α -glucosidase. Thus, glucose is absorbed more slowly into the
260 body, and postprandial hyperglycemia can be suppressed. In other words, in the group
261 with a small number of AMY-CN, the expression of acarbose in the gut can be
262 increased and the increase in blood glucose is suppressed; in contrast, with a large
263 number of AMY-CN, the expression of acarbose is decreased and the increase in blood
264 glucose is promoted. Through this mechanism, we could partially explain the causes of
265 an increase in HbA1c levels and BMI in the group with a large AMY-CN.

266

267 Within this study, we considered a particularly limited population in Japan, and so,
268 these results may be specific to lifestyle, genome, and the oral and gut microbiome, and
269 may not be applicable to other populations. However, even in such a specific population,
270 our findings are partially consistent with previous research, and novel findings on
271 AMY-CN have been further obtained.

272

273 **Materials and Methods**

274 *Ethics statement*

275 This study design was approved by the Hirosaki University Hospital Institutional
276 Review Board (Protocol Number Num. 2017-026). Written informed consent was
277 obtained from all the patients in this study. No animal experiments were conducted in
278 this study.

279

280 *Sample collection and preparation*

281 We analyzed the data obtained from the Hirosaki Center of Innovation (COI) health
282 promotion project, which has collected cohort data from the year 2005 in Hirosaki (the
283 population is approximately 10, 000) in northern Japan. Each year, approximately 1,000
284 people have their health status checked through a free participation program. The
285 dataset actually has numerous missing data because the participants do not necessarily
286 undergo a checkup each year. Thus, the number of analyzed participants was displayed
287 in each experiment. Questionnaires and blood tests were collected over the entire study
288 period, and single nucleotide polymorphisms (SNPs) were identified for participants in
289 2014, 2015, 2016, and 2017; WGS analysis was performed for participants in 2014 and
290 2015; 16S rRNA sequencing of the oral and gut microbiomes was performed for
291 participants in 2015, 2016, 2017, and 2018, and whole-genome shotgun sequencing of
292 the oral and gut microbiomes was performed for participants in 2015 and 2016.

293

294 *Participant selection and covariate assessment*

295 The main exclusion criteria were as follows: (1) participants were younger than 18 years
296 of age or were 55 older; (2) those with an estimated glomerular filtration rate (eGFR)

297 below 60 [mL/min/1.73 m²] as calculated from serum creatinine levels and age at
298 admission; (3) those with serum total bilirubin levels above 2 [mg/dl], (4) those with
299 serum albumin levels above 35 [g/dl], (5) those with a BMI above 35 or below 18, (6)
300 those who were prescribed medicine for diabetes mellitus as per their questionnaires, (7)
301 those who were prescribed antidiabetic or antibiotic medication as per their
302 questionnaires, for the analysis of oral and gut microbiome; and (8) those with missing
303 information regarding any of their covariates.

304

305 The participant covariate information was obtained from the blood value tests and
306 questionnaires. The questionnaires were administered upon induction into the study. (a)
307 Age and (b) sex information were obtained from questionnaires (sex was also checked
308 using genetic data), (c) smoking status was classified as never (0) and current and former
309 (1); (d) drinking status was classified as non-drinker (0), and current and former drinker
310 (1). Participants who used medication for (e) high blood pressure, (f) dyslipidemia, (g)
311 steroids, and (h) antibiotics were flagged as non-use (0) and use (1). Finally, we
312 performed principal component analysis (PCA) for genotyping using the algorithm
313 implemented in smartpca (24), and used the top five principal components (PCs) as
314 covariates to represent the population structure based on the genetic correlations among
315 individuals.

316

317 *SNPs and whole genome sequencing*

318 DNA was purified from peripheral whole blood using a DNA extraction kit (QIAamp
319 96 DNA Blood Kit (QIAGEN, Hilden, Germany)) and was extracted from plasma
320 pellets for SNP arrays and whole genome sequencing, respectively. SNP information
321 was obtained from Japonica Array (Toshiba, Tokyo, Japan) (25) and comprised
322 population-specific SNP markers designed from the 1070 whole genome reference
323 panel (26), and whole genome sequencing was performed by Takara Bio Corporation
324 (Shiga, Japan).

325

326 *Analysis of AMY1, AMY2A, and AMY2B copy number variations*

327 To detect AMY1, AMY2A, and AMY2B CNVs from the WGS data, we used
328 AMYCNE (21), which was developed to estimate the AMY CNV from sequencing data.
329 The input data for this software was prepared using TIDIT (20) which was also
330 developed to clarify the structural variants.

331

332 *SNP analysis for AMY1, AMY2A, and AMY2B regions*

333 For SNP detection, we applied the Genomon pipeline
334 (<https://github.com/Genomon-Project>), which has been developed to investigate
335 mutations from exome/whole genome sequencing data. The filter conditions were as
336 follows: minor allele frequency of 0.02 or higher, minimum read depth higher than 8,
337 and 10% posterior percentile of the minor allele count based on a binomial distribution
338 greater than 0.02, and a minimum base quality of 15 or higher.

339

340 ***Microbiome assay***

341 Microbiome samples were obtained from participants' tongue plaques and stool samples
342 for oral and gut microbiomes, respectively, and were stored at -80°C until use. The
343 detailed library preparation method, including PCR conditions, has been described in a
344 previous paper (27). Briefly, the samples were mixed with zirconia beads and lysed using
345 a FastPrep FP100A instrument (MP Biomedicals, Santa Ana, California, USA). DNA
346 was extracted from the bead-treated suspensions using a Magtration System 12GC and
347 GC series MagDEA DNA 200 (Precision System Science, Chiba, Japan) and amplified.
348 The 16S rRNA gene amplicons covering the V3–V4 region were sequenced using
349 Pro341F and Pro805R primers. Sequencing was performed using a paired-end,
350 2×250-base pair cycle run on an Illumina MiSeq sequencing system. Shotgun whole
351 genome sequencing was performed on the HiSeq2500 instrument (Illumina, San Diego,
352 California, USA) with 101 bp paired-end reads.

353

354 ***Calculation of genus level microbiome abundance profiles from 16S rRNA sequences***

355 Adapter sequences and low-quality bases at the 3'-end (threshold = 30) were trimmed
356 using Cutadapt (version: 1.13) (28) from paired-end reads obtained using the Illumina
357 MiSeq sequencing system. In addition, the reads containing N bases and reads shorter
358 than 150 bases were removed. The remaining reads were merged into single reads, and
359 those with a length of 370 to 470 bases were removed using the VSEARCH (29)
360 fastq_mergepairs subcommand (version: version 2.4.3). After merging, we removed
361 reads that were expected to have one or more base errors from the base quality. After
362 removing chimeric reads using the VSEARCH (29) uchime_denovo subcommand, the
363 reads were clustered with 97% sequence identity and the taxa of each cluster were
364 identified using the RDP Classifier (git commit hash: (commit hash:
365 701e229dde7cbe53d4261301e23459d91615999d)) (30). At this time, the result with a
366 confidence value < 0.8, was treated as unclassified. The relative abundance of each
367 taxon was then calculated by dividing the number of reads belonging to the cluster to
368 which the taxon was assigned by the total number of reads.

369

370 ***Calculation of functional gene abundance from shotgun metagenomic sequencing***

371 Adapter sequence and low-quality bases at the 3'-end (threshold = 30) were trimmed
372 using Cutadapt (version: 1.13) from 2×100-base paired-end reads obtained using the
373 Illumina HiSeq 2500 system. Reads containing N bases and reads shorter than 80 bp
374 were removed. Reads from the host genome were also removed using BWA-MEM
375 (version 0.7.15) (31) with GRCh38. From the remaining reads (referred to as non-host
376 reads), contigs were generated by metagenomic assembly using MEGAHIT (version
377 1.1.1) (32). The coding regions of genes on these contigs were predicted using Prodigal
378 (version 2.6.3) (33). After mapping the non-host reads to contigs using BWA-MEM, the
379 relative abundances of the genes were calculated from the average depth of the gene
380 region. KEGG pathway annotations were mapped to genes using a sequence similarity
381 search with MMseqs2 (version: 1d2579627f43662ecaaa0778bd348fc35048976a) (34).
382 We used 1E-10 as the E-value threshold for the similarity search.

383

384 ***Statistical Analysis***

385 In the first criterion for AMY1-CN, participants were divided into low (AMY1-CN
386 range 3–10) and high (AMY1-CN range 11–24) groups, which were separated at the
387 center line. In the second criterion for AMY1-CN, participants were divided into the
388 low (AMY1-CN range, 3–7), medium (range 8-13), and high (range, 14–24) groups.
389 The minimum copy number above 10 percentile and the maximum copy number below
390 90 percentile of AMY1-CN distribution was 7 and 14, respectively. Similarly, for
391 analyzing AMY2A-CN, participants were divided into low (AMY2A-CN range 1–2)
392 and high (AMY2A-CN range 3–5) groups at the center line. We could not analyze the
393 effect of AMY2B-CN because almost all patients had two AMY2B-CNs. For these
394 groups, linear regression analysis using the above-explained covariates was applied to
395 detect statistical significance. A p-value less than 0.05 was considered to indicate
396 statistical significance when considering the corrections for multiple tests.

397

398 ***Acknowledgments***

399 This work was supported by JST COI Grant Number JPMJCE1302. This research used
400 the supercomputing resources provided by the Human Genome Center, Institute of
401 Medical Science, University of Tokyo (<http://sc.hgc.jp/shirokane.html>).

402

403 ***Declaration of interests***

404 The authors declare no competing financial interests.

405

406 **Authors' contributions**

407 T.H. designed the study, analyzed the data, interpreted the data, and wrote the
408 manuscript. M. K., R.Y., and N.S. analyzed the data, interpreted the data, helped write
409 the manuscript, and commented on the drafts. M.A., J.S., I.T., and K.M. performed the
410 medical checkups and gathered and summarized the data. S.I. interpreted the data,
411 helped write the manuscript, commented on the drafts, and supervised the study. All
412 authors have read and approved the final manuscript.

413

414 **Data Availability**

415 The datasets supporting the current study have not been deposited in a public repository
416 [because the data are not public] but are available from the corresponding author, S.I.,
417 on request.

418

419 **References**

420

- 421 1. Pajic P, Pavlidis P, Dean K, Neznanova L, Romano RA, Garneau D, et al. Independent amylase
422 gene copy number bursts correlate with dietary preferences in mammals. *Elife*. 2019;8.
- 423 2. Perry GH, Dominy NJ, Claw KG, Lee AS, Fiegler H, Redon R, et al. Diet and the evolution of
424 human amylase gene copy number variation. *Nat Genet*. 2007;39(10):1256-60.
- 425 3. Inchley CE, Larbey CD, Shwan NA, Pagani L, Saag L, Antao T, et al. Selective sweep on
426 human amylase genes postdates the split with Neanderthals. *Sci Rep*. 2016;6:37198.
- 427 4. Falchi M, El-Sayed Moustafa JS, Takousis P, Pesce F, Bonnefond A, Andersson-Assarsson JC,
428 et al. Low copy number of the salivary amylase gene predisposes to obesity. *Nat Genet*.
429 2014;46(5):492-7.
- 430 5. Mejia-Benitez MA, Bonnefond A, Yengo L, Huyvaert M, Dechaume A, Peralta-Romero J, et al.
431 Beneficial effect of a high number of copies of salivary amylase AMY1 gene on obesity risk in Mexican
432 children. *Diabetologia*. 2015;58(2):290-4.
- 433 6. Bonnefond A, Yengo L, Dechaume A, Canouil M, Castelain M, Roger E, et al. Relationship
434 between salivary/pancreatic amylase and body mass index: a systems biology approach. *BMC Med*.
435 2017;15(1):37.
- 436 7. Elder PJD, Ramsden DB, Burnett D, Weickert MO, Barber TM. Human amylase gene copy
437 number variation as a determinant of metabolic state. *Expert Rev Endocrinol Metab*. 2018;13(4):193-205.
- 438 8. Viljakainen H, Andersson-Assarsson JC, Armenio M, Pekkinen M, Pettersson M, Valta H, et al.
439 Low Copy Number of the AMY1 Locus Is Associated with Early-Onset Female Obesity in Finland. *PLoS*
440 *One*. 2015;10(7):e0131883.

- 441 9. Marcovecchio ML, Florio R, Verginelli F, De Lellis L, Capelli C, Verzilli D, et al. Low AMY1
442 Gene Copy Number Is Associated with Increased Body Mass Index in Prepubertal Boys. *PLoS One*.
443 2016;11(5):e0154961.
- 444 10. Yong RY, Mustaffa SB, Wasan PS, Sheng L, Marshall CR, Scherer SW, et al. Complex Copy
445 Number Variation of AMY1 does not Associate with Obesity in two East Asian Cohorts. *Hum Mutat*.
446 2016;37(7):669-78.
- 447 11. Usher CL, Handsaker RE, Esko T, Tuke MA, Weedon MN, Hastie AR, et al. Structural forms
448 of the human amylase locus and their relationships to SNPs, haplotypes and obesity. *Nat Genet*.
449 2015;47(8):921-5.
- 450 12. Rukh G, Ericson U, Andersson-Assarsson J, Orho-Melander M, Sonestedt E. Dietary starch
451 intake modifies the relation between copy number variation in the salivary amylase gene and BMI. *Am J*
452 *Clin Nutr*. 2017;106(1):256-62.
- 453 13. Al-Akl N, Thompson RI, Arredouani A. High plasma salivary alpha-amylase, but not high
454 AMY1 copy number, associated with low obesity rate in Qatari adults: cross-sectional study. *Sci Rep*.
455 2020;10(1):17918.
- 456 14. Higuchi R, Iwane T, Iida A, Nakajima K. Copy Number Variation of the Salivary Amylase
457 Gene and Glucose Metabolism in Healthy Young Japanese Women. *J Clin Med Res*. 2020;12(3):184-9.
- 458 15. Nakajima K, Higuchi R, Iwane T, Iida A. The association of low serum salivary and pancreatic
459 amylases with the increased use of lipids as an energy source in non-obese healthy women. *BMC Res*
460 *Notes*. 2020;13(1):237.
- 461 16. Leon-Mimila P, Villamil-Ramirez H, Lopez-Contreras BE, Moran-Ramos S, Macias-Kauffer
462 LR, Acuna-Alonzo V, et al. Low Salivary Amylase Gene (AMY1) Copy Number Is Associated with
463 Obesity and Gut Prevotella Abundance in Mexican Children and Adults. *Nutrients*. 2018;10(11).
- 464 17. Hjorth MF, Christensen L, Larsen TM, Roager HM, Krych L, Kot W, et al. Pretreatment
465 Prevotella-to-Bacteroides ratio and salivary amylase gene copy number as prognostic markers for dietary
466 weight loss. *Am J Clin Nutr*. 2020;111(5):1079-86.
- 467 18. Poole AC, Goodrich JK, Youngblut ND, Luque GG, Ruaud A, Sutter JL, et al. Human Salivary
468 Amylase Gene Copy Number Impacts Oral and Gut Microbiomes. *Cell Host Microbe*. 2019;25(4):553-64
469 e7.
- 470 19. Zhou B, Ho SS, Zhang X, Pattni R, Haraksingh RR, Urban AE. Whole-genome sequencing
471 analysis of CNV using low-coverage and paired-end strategies is efficient and outperforms array-based
472 CNV analysis. *J Med Genet*. 2018;55(11):735-43.
- 473 20. Eisfeldt J, Vezzi F, Olason P, Nilsson D, Lindstrand A. TIDDIT, an efficient and comprehensive
474 structural variant caller for massive parallel sequencing data. *F1000Res*. 2017;6:664.
- 475 21. Eisfeldt J, Nilsson D, Andersson-Assarsson JC, Lindstrand A. AMYCNE: Confident copy
476 number assessment using whole genome sequencing data. *PLoS One*. 2018;13(3):e0189710.

- 477 22. Muniz Pedrogo DA, Jensen MD, Van Dyke CT, Murray JA, Woods JA, Chen J, et al. Gut
478 Microbial Carbohydrate Metabolism Hinders Weight Loss in Overweight Adults Undergoing Lifestyle
479 Intervention With a Volumetric Diet. *Mayo Clin Proc.* 2018;93(8):1104-10.
- 480 23. Kanehisa M, Goto S. KEGG: kyoto encyclopedia of genes and genomes. *Nucleic Acids Res.*
481 2000;28(1):27-30.
- 482 24. Price AL, Patterson NJ, Plenge RM, Weinblatt ME, Shadick NA, Reich D. Principal
483 components analysis corrects for stratification in genome-wide association studies. *Nat Genet.*
484 2006;38(8):904-9.
- 485 25. Kawai Y, Mimori T, Kojima K, Nariai N, Danjoh I, Saito R, et al. Japonica array: improved
486 genotype imputation by designing a population-specific SNP array with 1070 Japanese individuals. *J*
487 *Hum Genet.* 2015;60(10):581-7.
- 488 26. Nagasaki M, Yasuda J, Katsuoka F, Nariai N, Kojima K, Kawai Y, et al. Rare variant discovery
489 by deep whole-genome sequencing of 1,070 Japanese individuals. *Nat Commun.* 2015;6:8018.
- 490 27. Takahashi S, Tomita J, Nishioka K, Hisada T, Nishijima M. Development of a prokaryotic
491 universal primer for simultaneous analysis of Bacteria and Archaea using next-generation sequencing.
492 *PLoS One.* 2014;9(8):e105592.
- 493 28. MARTIN M. Cutadapt removes adapter sequences from high-throughput sequencing reads.
494 *EMBnetjournal.* 2011;17(1):10-2.
- 495 29. Rognes T, Flouri T, Nichols B, Quince C, Mahe F. VSEARCH: a versatile open source tool for
496 metagenomics. *Peerj.* 2016;4.
- 497 30. Lan YM, Wang Q, Cole JR, Rosen GL. Using the RDP Classifier to Predict Taxonomic Novelty
498 and Reduce the Search Space for Finding Novel Organisms. *Plos One.* 2012;7(3).
- 499 31. Li H. Aligning sequence reads, clone sequences and assembly contigs with BWA-MEM.
500 *arXiv:13033997v2* 2013.
- 501 32. Li DH, Liu CM, Luo RB, Sadakane K, Lam TW. MEGAHIT: an ultra-fast single-node solution
502 for large and complex metagenomics assembly via succinct de Bruijn graph. *Bioinformatics.*
503 2015;31(10):1674-6.
- 504 33. Hyatt D, Chen GL, LoCascio PF, Land ML, Larimer FW, Hauser LJ. Prodigal: prokaryotic
505 gene recognition and translation initiation site identification. *Bmc Bioinformatics.* 2010;11.
- 506 34. Steinegger M, Soding J. MMseqs2 enables sensitive protein sequence searching for the
507 analysis of massive data sets. *Nat Biotechnol.* 2017;35(11):1026-8.
- 508
509
510
511
512

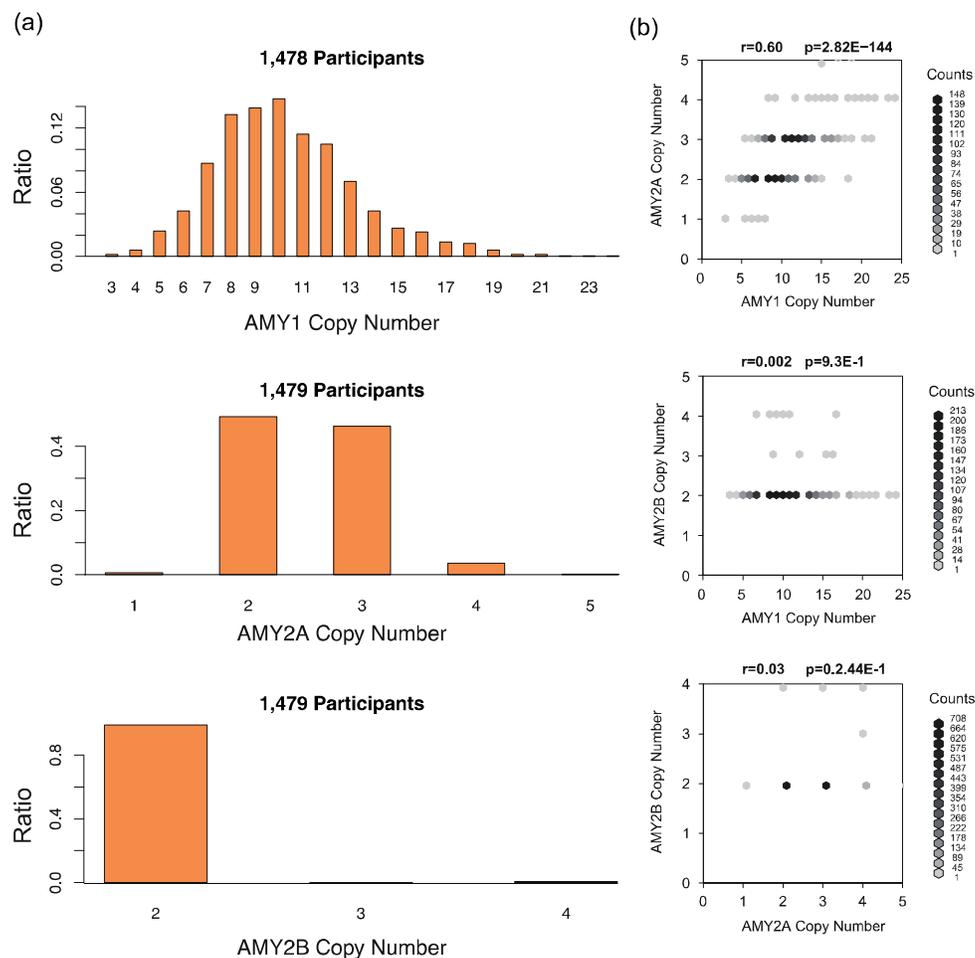
513

514

515

516 **Figure Legends**

517



518

519 **Figure 1. Histogram of AMY1, AMY2A, and AMY2B copy number variations and**

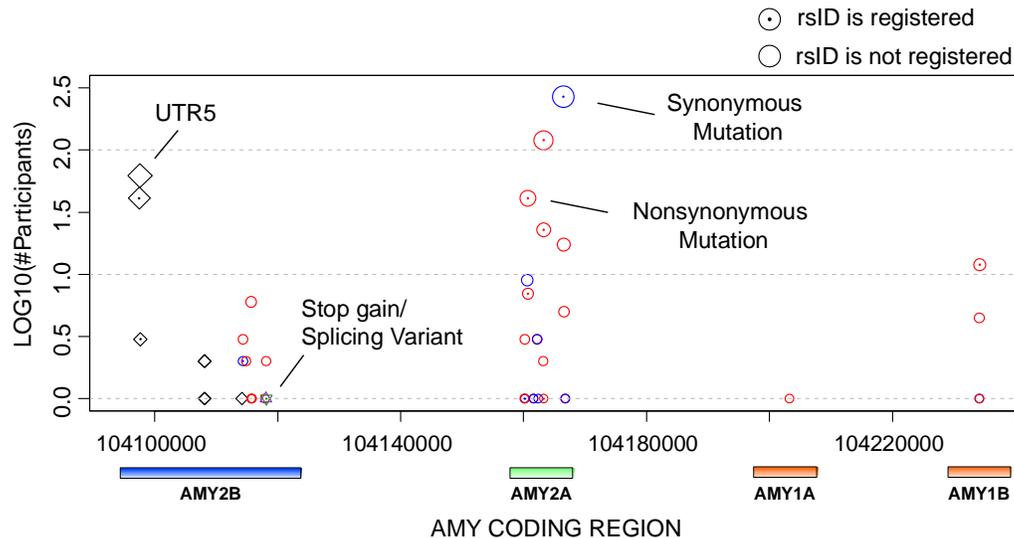
520 **their associations.** In the left of this figure, the histograms of AMY1A, AMY2A, and

521 AMY2B-CN are displayed. The number of analyzed participants are also indicated at

522 the top of each histogram. In the right of this figure, associations among AMY1A,

523 AMY2A, and AMY2B-CN are illustrated as a scatter plot. At the top of each scatter

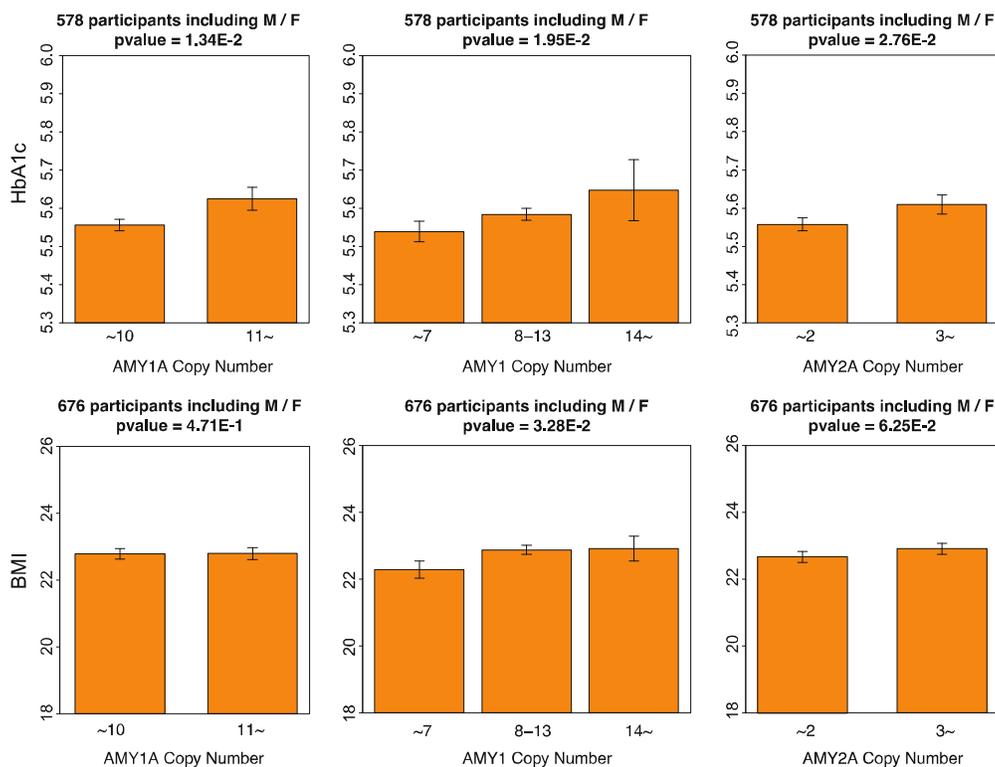
524 plot, r and p indicate Pearson's correlation coefficient and p-value, respectively.



525

526

527 **Figure2. Distributions of SNPs on the exonic regions of AMY1, AMY2A, and**
528 **AMY2B.** In this figure, the vertical and horizontal axes indicate the number of
529 participants with the corresponding SNPs and genomic positions including AMY1A,
530 AMY2A, and AMY2B. Red, blue, green, black, and orange circles indicate
531 nonsynonymous, synonymous, stop-gain (nonsense), splicing site, and 5'
532 prime untranslated region mutations, respectively. The dot at the center of the circles
533 indicates that the SNPs have reference SNP IDs. Genes corresponding to the positions
534 are illustrated at the bottom of the figure.



535

536

537

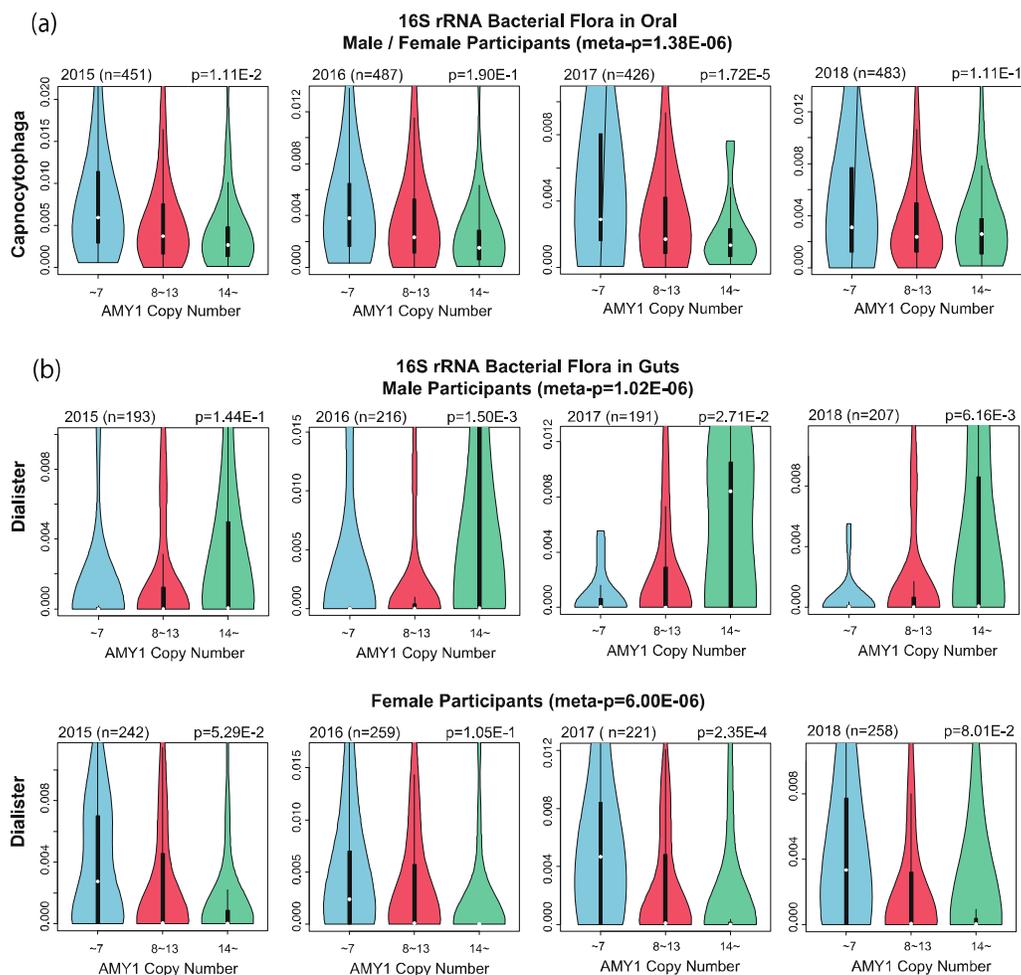
538

539

540

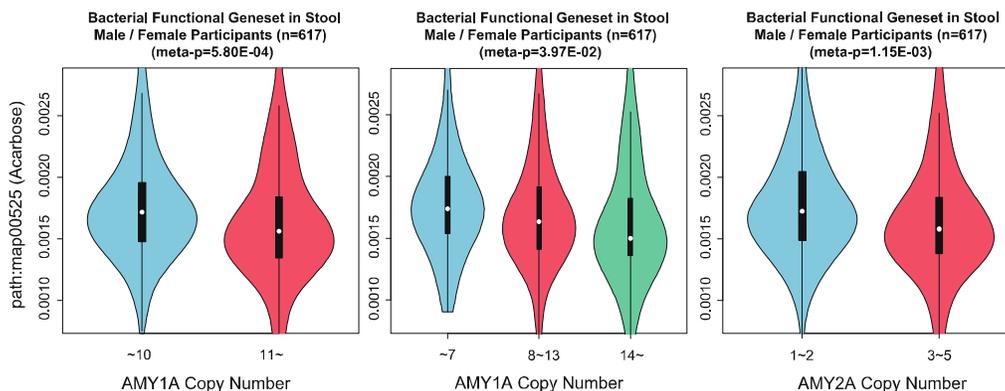
541

Figure3. Association of HbA1c and BMI with AMY1A and AMY2A-CN. In the upper and bottom parts of this figure, the associations of HbA1c and BMI with the criterion 1 and 2 of AMY1-CN and AMY2-CN are illustrated as barplots, respectively. The number of analyzed participants and p-values calculated using linear regression models with predefined covariates are displayed on the barplots.



542

543 **Figure 4. Association of the oral and gut microbiome with AMY1-CN.** In the upper,
544 middle, and bottom panels of this figure, associations of the composition of
545 *Capnocytophaga* for males and females, *Dialister* for males, and *Dialister* for females
546 are illustrated by violin-plot. The number of analyzed participants for each year and the
547 p-values calculated using linear regression models with predefined covariates are
548 displayed on the plots.



549

550 **Figure 5. Associations of abundance of the Acarbose pathway gene set.** The
 551 associations of the total abundance of the gene-set in the Acarbose pathway (KEGG:
 552 map00525) to criterion 1 and 2 of AMY1-CN and AMY2-CN are illustrated as
 553 violin-plots. The number of analyzed participants and p-values calculated by linear
 554 regression models with predefined covariates are displayed on the plots.

555

556 **Tables**

557 **Table 1. Summary of SNPs detected in the AMY1, AMY2A, and AMY2B gene**
 558 **regions**

	exonic (nonsynonymous, synonymous, top)	intronic	UTR5	splicing	intergenic
AMY1	20 (19, 1, 0)	1,131	0	0	583,736
AMY2	564 (251, 292, 1)	13,253	113	1	13,456
AMY1-AMY 2	-	-	-	-	1,229
Total	564	14,384	113	1	609,874

559

560

561

562

563

564

565

566

567

568

569

570 **Supplementary Material**

	AMY1A	AMY2A	AMY1A	AMY2A	AMY1A	AMY2A					
Capnocytophag	1.38492E-06	Alloscardovia	0.00252246	Bacteroides	0.00884817	Mycoplasma	0.00537596	Capnocytoph	9.7757E-07	Howardella	0.00122281
Lautropia	0.002875591	Desulfobulbu	0.00767486	Lautropia	0.01163532	Schwartzia	0.01497963	Fusobacteri	0.00057786	Shuttleworth	0.00617352
Romboutsia	0.004980457	Treponema	0.0258245	Capnocytoph	0.01863692	Tannerella	0.01531352	Parvimonas	0.00645228	Peptoniphilu	0.01131323
Catonella	0.009130838	Achromobact	0.04165186	Catonella	0.01973323	Treponema	0.01586972	Campylobact	0.0128947	Alloscardovia	0.01399269
Saccharibact	0.013648724	Raoultella	0.05813924	Filifactor	0.03337993	Neisseria	0.01607126	Romboutsia	0.01376227	Cryptobacter	0.02490859
Fusobacteriu	0.017850754	Schwartzia	0.06380201	Enterococcus	0.034455	Peptoniphilu	0.02106509	Solobacteriu	0.01485281	Granulicatell	0.0378569
Peptostrepto	0.03286158	Pyramidobac	0.07264064	Escherichia/	0.03862957	Scardovia	0.03884765	Rothia	0.02025476	Lautropia	0.04181737
Enterococcus	0.043417378	Escherichia/	0.08075315	Deinococcus	0.04597273	Raoultella	0.03991252	Blautia	0.02662666	Roseburia	0.04285195
Campylobact	0.043509591	Propionibact	0.08388968	Slackia	0.05350477	Comamonas	0.04269764	Cryptobacter	0.02893532	Niveispirillum	0.04341788
Solobacteriu	0.048392175	Phascolarcta	0.08444056	Peptostrepto	0.05612149	Filifactor	0.04531431	Mogibacteriu	0.03141558	Catonella	0.04462881

Gut Microbiome (genus, p-value)											
Male and Female				Male				Female			
AMY1A	AMY2A	AMY1A	AMY2A	AMY1A	AMY2A	AMY1A	AMY2A	AMY1A	AMY2A	AMY1A	AMY2A
Megasphaera	0.001487438	Anaerostipes	0.0011348	Dialister	1.022E-06	Bifidobacteri	0.00167491	Dialister	6.0041E-06	Flavonifracto	3.4493E-05
Faecalibacte	0.001739769	Flavonifracto	0.00541643	Megasphaera	2.9204E-05	Gemella	0.0027667	Faecalibacte	0.00010316	Dialister	0.00575892
Flavonifracto	0.007713141	Gemella	0.00802741	Bifidobacteri	7.5115E-05	Methanobrev	0.00330896	Megamonas	0.00018452	Barnesiella	0.00604753
Coprococcus	0.010056927	Coprobacter	0.00857085	Veillonella	0.00034992	Stomatobacu	0.00924707	Flavonifracto	0.00036173	Lachnospira	0.00629187
Pediococcus	0.010693062	Stomatobacu	0.00860853	Holdemanell	0.00125158	Anaerostipes	0.01421471	Alloprevotell	0.00283018	Enterococcus	0.00659771
Christensen	0.014092088	Enterococcus	0.00953113	Acidaminoco	0.00259936	Coprobacter	0.01748766	Stomatobacu	0.005165	Bilophila	0.00714666
Ruminococcu	0.014284955	Morganella	0.02101648	Prevotella	0.0107398	Dorea	0.01955454	Veillonella	0.00579802	Succinivibro	0.0082315
Turicobacter	0.025007697	Dialister	0.02249705	Coprococcus	0.01876237	Buttiauxella	0.02320624	Eggerthella	0.01130972	Acidaminoco	0.01515036
Eggerthella	0.029521963	Bilophila	0.02732775	Olsenella	0.02155837	Megamonas	0.02544681	Clostridium X	0.01850992	Anaerofustis	0.02157995
Prevotella	0.042400996	Ruminococcu	0.0294968	Pediococcus	0.02222906	Alloscardovia	0.02775844	Romboutsia	0.02252999	Anaerostipes	0.02568245

571