Human and bacterial genetic variation shape oral microbiomes and health

Nolan Kamitaki[1,2,3,4,5,6], Robert E. Handsaker[3,4,5], Margaux L.A. Hujoel[1,2,3], Ronen E. Mukamel[1,2,3], Christina L. Usher[7], Steven A. McCarroll[3,4,5,8], Po-Ru Loh[1,2,3]

[1]Division of Genetics, Department of Medicine, Brigham and Women's Hospital and Harvard Medical School, Boston, MA, USA.
[2]Center for Data Sciences, Brigham and Women's Hospital, Boston, MA, USA.
[3]Program in Medical and Population Genetics, Broad Institute of MIT and Harvard, Cambridge, MA, USA.
[4]Stanley Center for Psychiatric Research, Broad Institute of MIT and Harvard, Cambridge, MA, USA.
[5]Department of Genetics, Harvard Medical School, Boston, MA, USA.
[6]Department of Biomedical Informatics, Harvard Medical School, Boston, MA, USA.
[7]Jerome Lipper Multiple Myeloma Center, Dana-Farber Cancer Institute, Boston, MA
[8]Howard Hughes Medical Institute, Harvard Medical School, Boston, MA, USA

Correspondence should be addressed to N.K. (nolan_kamitaki@hms.harvard.edu), S.A.M. (smccarro@broadinstitute.org), or P.-R.L. (poruloh@broadinstitute.org).

Abstract

It is largely unknown which human genetic variants shape a person's oral microbiome and potentially promote its dysbiosis. We characterized the oral microbiomes of 12,519 people by analyzing whole-genome sequencing reads from previously sequenced saliva-derived DNA. Human genetic variation at 11 loci (10 novel) associated with differences in oral microbiome composition. Nearly all of these associations implicated candidate genes with readily interpretable functions, several related to carbohydrate availability. The strongest association (p=3.0x10^{-188}) involved the common *FUT2* W154X loss-of-function variant, which associated with the abundances of 32 bacterial species. Human host genetics also appeared to powerfully shape within-species genetic variation in oral bacteria. Variation at the 11 human loci associated with variation in gene dosages in 68 regions of bacterial genomes. Several such associations implicated interactions of bacterial proteins with histo-blood group antigens presented on host mucosal cell surfaces and salivary proteins. Common, multi-allelic copy-number variation of *AMY1*, which encodes salivary amylase, associated with oral microbiome composition (p=1.5x10^{-53}) and with dentures use in UK Biobank (p=5.9x10^{-35}, n=418k), suggesting that amylase abundance impacts oral health by influencing the oral microbiome. Two other microbiome composition-associated loci, *FUT2* and *PITX1*, also significantly associated with dentures risk, collectively nominating numerous microbial taxa that might contribute to tooth decay.

Introduction

When Antonie van Leeuwenhoek first observed bacteria as “animalcules” in scrapings from his teeth in the seventeenth century, one of his first inquiries was the extent of their variation between people[1]. Oral microbiomes are now known to vary abundantly across people[2–5], and twin studies have shown that some of this variation is heritable[6–8]. However, few human genetic polymorphisms have been associated with the abundances of specific oral microbial species[8–10]; study sizes to date (n<3,000) have provided limited power to detect robust genetic effects. Larger genome-wide association studies (GWAS) of the gut microbiome (n=8-16k) have consistently replicated two genetic effects of variation at the *LCT* and *ABO* loci on gut microbial abundances[11–15], and larger genetic studies of oral microbiomes might be expected to yield similar loci.

Oral pathologies such as dental caries result from dysbiosis of the oral microbiome[16]. If untreated, such pathologies can progress to oral infections which carried high mortality rates prior to modern dentistry and antibiotics[17]. Susceptibility to caries and other oral health outcomes is also strongly influenced by genetics[18–20], and GWAS have identified 47 loci harboring such genetic effects[21]. However, the extent to which some of these genetic effects act by modulating the composition of the oral microbiome is currently unknown. Understanding these interactions could point to microbial drivers of cariogenesis, which are not yet clear[5].

Given these effects of human hosts and resident microbes on each other’s survival and evolutionary trajectory, the oral microbiome is an example of symbiosis[22,23]. The stability of the gut microbiome within individuals[24], its codiversification with humans[25], and abundant structural variation of its microbial genomes[26] all suggest intricate genetic interactions between microbiomes and their human hosts, whereby microbial genomes adapt to genetic variation across people. A recently observed example of such an interaction with the gut microbiome is a structural variant in the *Faecalibacterium prausnitzii* genome that includes genes encoding a GalNAc-metabolizing pathway and interacts with human *ABO* variation[27]. Whether such specific coadaptation commonly occurs in oral microbiomes remains an open question.

Results

The composition of 12,519 oral microbiomes changes with age

To create a data set suitable for exploring variation in the oral microbiome and the way it is shaped by human genetic variation, we analyzed DNA sequencing reads previously generated from whole-genome sequencing (WGS) of saliva samples from 12,519 participants in the SPARK cohort[28] (**Fig. 1a**), building on previous work[29,30]. This sequencing captured a substantial amount of non-human genomic information[31,32], with a median of 8.4% ([4.6%,14.7%], quartiles) of sequencing reads not mapping to the human reference genome (**Extended Data Fig. 1a**). Many of these unmapped reads instead mapped to clade-specific marker genes within microbial genomes (using MetaPhlAn 4; ref. [33]), allowing us to estimate

Figure 1

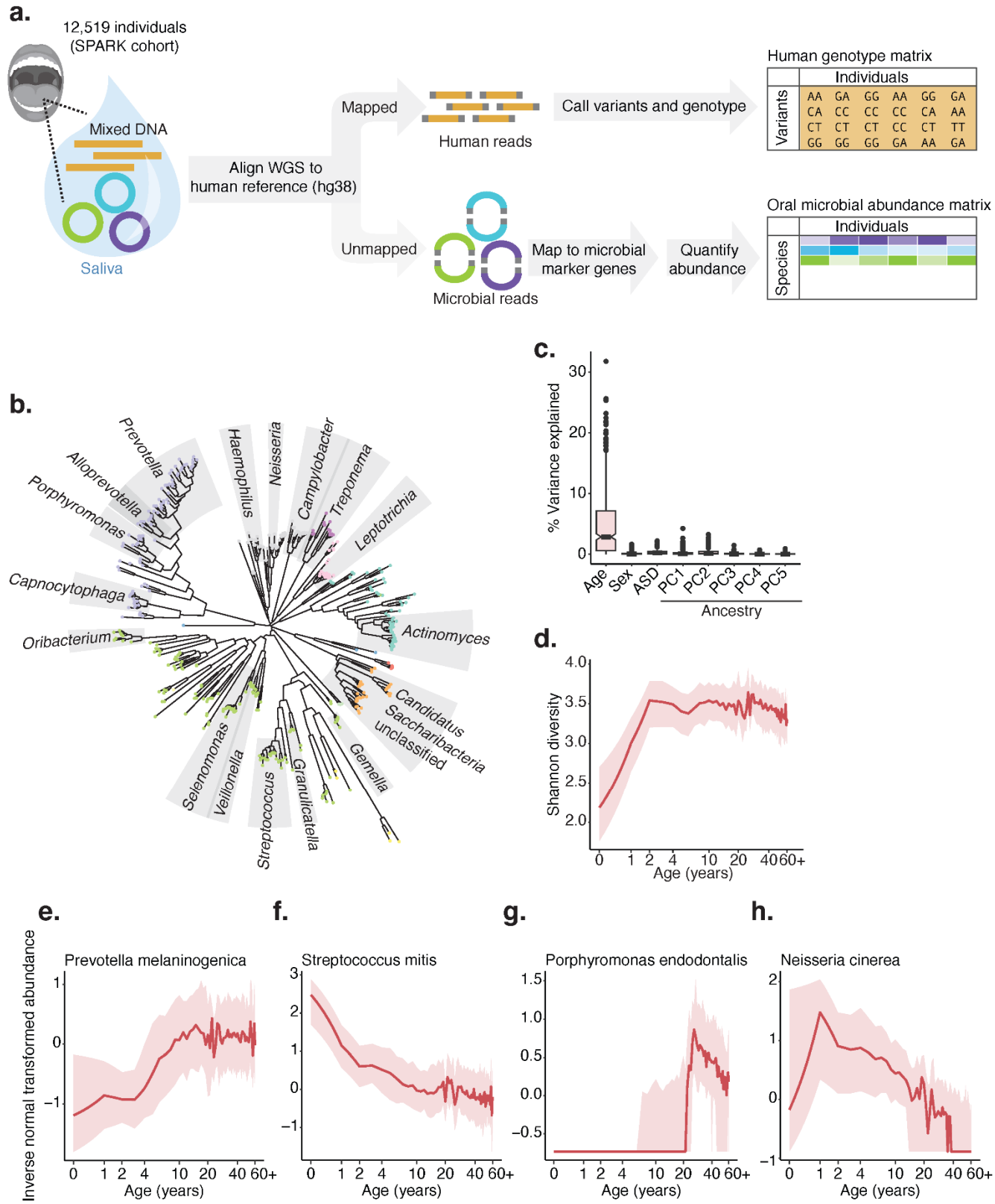

a.
12,519 individuals
(SPARK cohort)
Mixed DNA
Saliva
Align WGS to
human reference (hg38)
Mapped
Human reads
Call variants and genotype
Human genotype matrix
Individuals
Variants
Unmapped
Microbial reads
Map to microbial
marker genes
Quantify
abundance
Oral microbial abundance matrix
Species
b.
Prevotella
Alloprevotella
Porphyromonas
Capnocytophaga
Oribacterium
Haemophilus
Neisseria
Campylobacter
Treponema
Leptotrichia
Actinomyces
Candidatus
Saccharibacteria
unclassified
Gemella
Granulicatella
Streptococcus
Veillonella
Selenomonas
c.
% Variance explained
Age
Sex
ASD
PC1
PC2
PC3
PC4
PC5
Ancestry
d.
Shannon diversity
Age (years)
e.
Prevotella melaninogenica
Inverse normal transformed abundance
Age (years)
f.
Streptococcus mitis
Age (years)
g.
Porphyromonas endodontalis
Age (years)
h.
Neisseria cinerea
Age (years)

Figure 1. Oral microbiomes in 12,519 individuals measured by saliva whole genome sequencing. a, Generation of paired data sets of human genetic variation and oral microbiome composition from whole-genome sequencing of saliva samples from the Simons Foundation Powering Autism Research (SPARK) cohort (n=12,519). Human genetic variants were previously called with DeepVariant and relative abundances of microbial species were estimated with MetaPhlAn 4 from sequencing reads that did not map to the human genome. **b**, Phylogenetic tree based on genomic divergence among 439 microbial species observed in ≥10% of SPARK participants. Phyla are indicated by dot color and genera with more than five species are indicated with labeled gray sectors. **c**, Contributions of age, sex, autism spectrum disorder (ASD) case status, and genetic ancestry principal components (PC1 through PC5) to variation in oral microbial species abundances. For each factor, the fraction of variance in species abundance explained by the factor was computed for each of the 439 species, and the box and whisker plot shows the distribution of this quantity across the 439 species. ASD status explained a median fraction of variance of 0.002. **d**, Species diversity in the oral microbiome as a function of host age. The red line indicates median Shannon entropy, and the shaded region indicates the interquartile range. Oral microbial diversity increases substantially over the first few years of life, plateaus, and then modestly declines in late adulthood. **e-h**, Change in mean abundances of *Prevotella melaninogenica*, *Streptococcus mitis*, *Porphyromonas endodontalis*, and *Neisseria cinerea* over the age range found in the SPARK cohort (truncated at 60 years old due to limited sampling of elderly individuals). Relative abundances were inverse normal rank-transformed (y-axis). Red curves indicate medians; shading indicates interquartile regions.

microbial composition from their counts. This produced the largest collection of oral microbiome profiles (n=12,519) generated to date, measuring the abundances of 439 microbial species (spanning 13 phyla, including one fungal commensal, *Malassezia restricta*) commonly observed in SPARK (≥10% of participants) (**Fig. 1b** and **Extended Data Fig. 1b**). Comparing these profiles across individuals showed that age was a major driver of interindividual variation in oral microbiome composition (**Fig. 1c** and **Extended Data Fig. 1c,d**). Across the lifespan represented in SPARK (age 0–90 years), mean species diversity sharply increased in the first few years of life (representing when the oral cavity is colonized, diet diversifies, and primary teeth are acquired) and then decreased slowly with age[2,34] (**Fig. 1d**). Individual species exhibited vastly different abundance trajectories over the lifespan, with some observed predominantly in adults and others predominantly in children (**Fig. 1e-h**).

In contrast to the strong effects of age on oral microbiome composition, autism spectrum disorder (ASD) case status, sex, and genetic ancestry exhibited minimal effects (**Fig. 1c** and **Extended Data Fig. 1e**). We therefore included all SPARK participants with available WGS data in downstream analyses to maximize statistical power.

Human genetics influences oral microbiome composition through protein abundance, glycosylation, and immunity

To identify human genetic variants that influence interindividual differences in the abundance of one or more microbial taxa, we developed a statistical test to capture pleiotropic effects on many species within an interdependent microbial community[35–37], using principal component analysis to enable efficient genome-wide association testing (**Fig. 2a** and **Methods**). Similar to a recent approach for identifying genetic associations with high-dimensional cell state phenotypes in single-cell RNA-seq data[38], this approach also reduces the multiple-testing burden by testing each genetic variant only once.

Applying this approach to human genetic variants ascertained from the SPARK WGS data identified 11 human loci at which common genetic variation associated with oral microbiome composition ($p<5\times10^{-8}$; **Fig. 2b** and **Table 1**), with only one locus having been previously identified (*SLC2A9*)[9]. Among the 10 new loci, eight loci implicated genes—and in several cases, specific variants—with readily interpretable functions that could explain their associations with microbiome composition.

- Three loci contained genes encoding highly expressed salivary proteins: salivary amylase (encoded by *AMY1*; $p=1.5\times10^{-53}$, top association), submaxillary gland androgen-regulated proteins (encoded by *SMR3A* and *SMR3B*; $p=1.4\times10^{-12}$), and basic salivary proline-rich proteins (encoded by *PRB1–PRB4*; $p=1.1\times10^{-11}$). Likely causal variants at the *PRB* locus are identified in **Extended Data Fig. 2** and **Supplementary Note 1**. These associations appeared to be driven mainly by genetic variants that modify gene expression or copy number (**Table 1**), suggesting that a primary way in which host genetic variation shapes the oral microbiome is by modulating the abundances of salivary proteins. Consistent with these results, heritability-partitioning analysis[39] indicated that genetic effects on oral microbiome

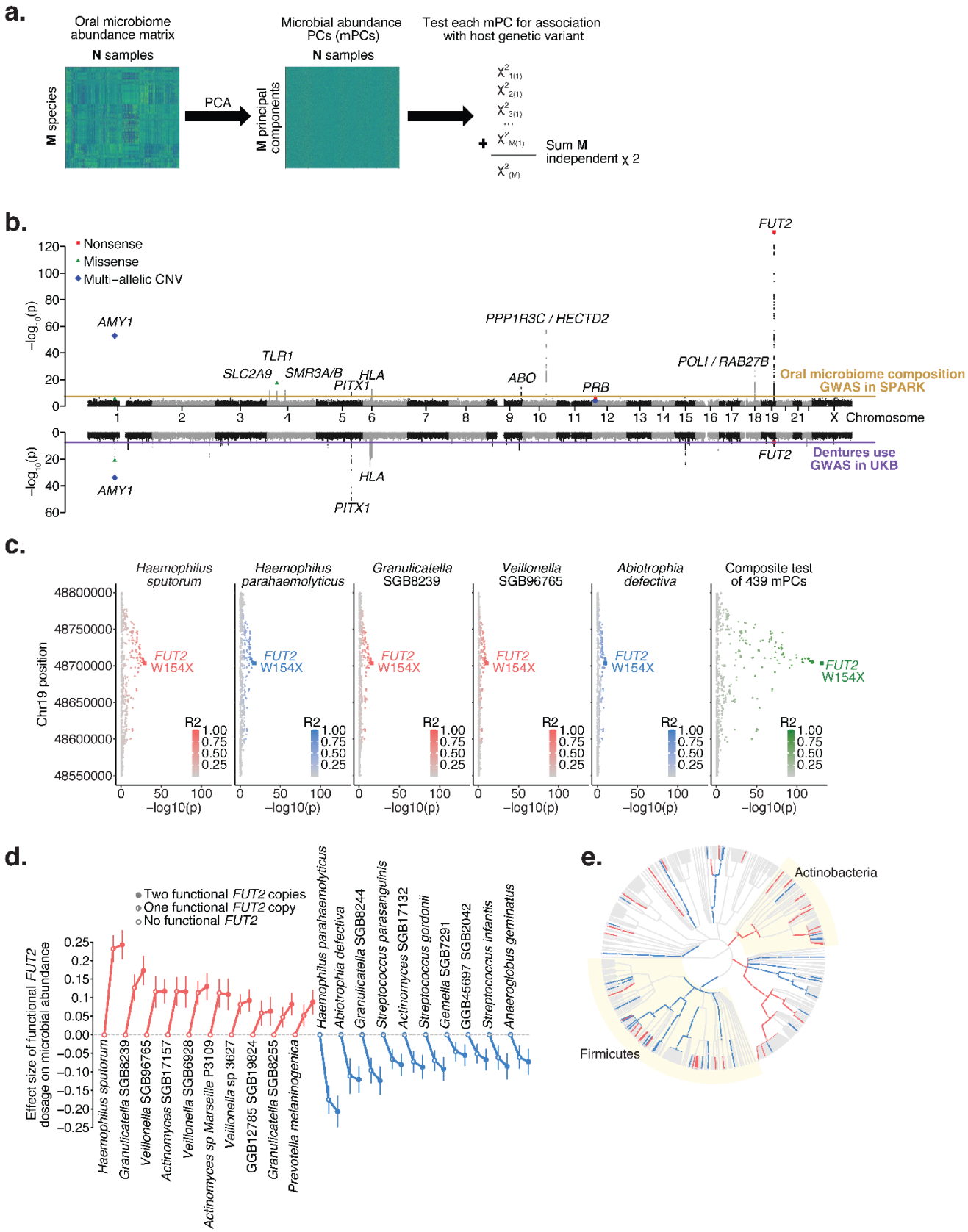

a.
Oral microbiome abundance matrix
N samples
M species
PCA
Microbial abundance PCs (mPCs)
N samples
M principal components
Test each mPC for association with host genetic variant
Sum M independent χ 2
b.
Nonsense
Missense
Multi−allelic CNV
−log10(p)
AMY1
TLR1
SLC2A9
SMR3A/B
HLA
PITX1
PPP1R3C / HECTD2
ABO
PRB
POLI / RAB27B
FUT2
Oral microbiome composition GWAS in SPARK
Chromosome
Dentures use GWAS in UKB
c.
Haemophilus sputorum
Haemophilus parahaemolyticus
Granulicatella SGB8239
Veillonella SGB96765
Abiotrophia defectiva
Composite test of 439 mPCs
Chr19 position
FUT2 W154X
R2
1.00
0.75
0.50
0.25
−log10(p)
d.
Two functional FUT2 copies
One functional FUT2 copy
No functional FUT2
Effect size of functional FUT2 dosage on microbial abundance
Haemophilus sputorum
Granulicatella SGB8239
Veillonella SGB96765
Actinomyces SGB17157
Veillonella SGB6928
Actinomyces sp Marseille P3109
Veillonella sp 3627
GGB12785 SGB19824
Granulicatella SGB8255
Prevotella melaninogenica
Haemophilus parahaemolyticus
Abiotrophia defectiva
Granulicatella SGB8244
Streptococcus parasanguinis
Actinomyces SGB17132
Streptococcus gordonii
Gemella SGB7291
GGB45697 SGB2042
Streptococcus infantis
Anaeroglobus geminatus
e.
Actinobacteria
Firmicutes

Figure 2. Associations of host genetic variants with oral microbiome composition overlap risk loci for dentures use. a, Statistical test for association of a genetic variant with oral microbiome composition. Converting highly-correlated relative abundances of *M* microbial species (left) into *M* orthogonal principal components (middle) allows combining chi-squared statistics (one per PC) into a single chi-squared test statistic with *M* degrees of freedom (right). **b**, Miami plot of genome-wide associations with oral microbiome composition in SPARK (top, n=12,519) and dentures use in UK Biobank (bottom, n=418,039). Protein-altering variants and copy number variants of note are highlighted: nonsense (red squares), missense (green triangles), and multi-allelic CNVs (blue diamonds). **c**, Associations of variants at the *FUT2* locus with relative species abundance for the five microbial species with the strongest associations (left five plots); color indicates effect direction (red, more abundant with functional *FUT2*; blue, less abundant with functional *FUT2*), and color saturation indicates linkage disequilibrium (LD) with rs601338 (*FUT2* W154X). Association strengths from the combined test for association with oral microbiome composition exhibit much greater statistical power (rightmost plot). **d**, Effect sizes (in s.d. units) on relative abundance of microbial species for individuals heterozygous for functional *FUT2* (half-filled circles) and for homozygotes (filled circles) relative to those with no functional *FUT2* (empty circles). For each effect direction, the ten most significantly associated species are shown. Error bars, 95% CIs. **e**, Microbial taxa whose abundance associated with *FUT2* genotype (FDR<0.1) shown on the phylogenetic tree of 439 species (red, taxa whose relative abundances increased with functional *FUT2*; blue, taxa whose relative abundances decreased with functional *FUT2*). Two significantly-associated phyla (Firmicutes and Actinobacteria; $p=1.2\times10^{-4}$ and 4.0×10^{-5}, respectively) are highlighted with yellow sectors. At the species level (outermost circle), dot sizes increase with statistical significance.

Locus	Lead variant	AF	$p_{microbiome}$	$p_{dentures}$	Other information
AMY1	*AMY1* copy number	n/a	1.5×10^{-53}	**5.9×10^{-35}**	Determines *AMY1* gene dosage
SLC2A9	4_9925343_T_G	0.32	2.9×10^{-12}	**0.021**	*SLC2A9* eQTL and sQTL; replicates oral microbiome association[15]; associated with serum urate[40]
TLR1	4_38797027_C_A	0.4	6.2×10^{-18}	0.99	*TLR1* I602S missense; affects immune response[41,42]
SMR3A/B	4_70378787_C_T	0.12	1.4×10^{-12}	**6.0×10^{-3}**	*SMR3A* and *SMR3B* eQTL; *SMR3B* and *MUC7* sQTL
PITX1	5_135173987_C_T	0.59	3.0×10^{-11}	**7.5×10^{-52}**	Variant has unknown function; *Pitx1* knockout causes jaw malformation[43]
HLA	6_32535351_G_A	0.6	2.2×10^{-16}	**7.1×10^{-13} (rs9271236)**	eQTL for six *HLA* class II genes
ABO	9_133266456_C_T	0.18	9.4×10^{-15}	**0.022**	Tags A1 blood group allele[44]
PPP1R3C/ HECTD2	10_91541231_A_G	0.33	8.8×10^{-58}	**0.048**	*HECTD2* and *PPP1R3C* eQTL
PRB1-PRB4	12_11369682_G_A	0.75	1.1×10^{-11}	0.88 (rs7977399)	*PRB2* eQTL; *PRB1* mCNV and *PRB4* R39X also associate with oral microbiome composition (**Supplementary Note 1**)
POLI/ RAB27B	18_54457253_G_A	0.08	4.2×10^{-27}	0.75	*RAB27B*, *POLI*, and *C18orf54* eQTL; *POLI* sQTL
FUT2	19_48703417_G_A	0.45	1.6×10^{-131}	**7.4×10^{-8}**	*FUT2* W154X; primary determinant of secretor status[45]

Table 1. **Eleven loci at which human genetic variants associate with oral microbiome composition**. AF, allele frequency. $p_{microbiome}$, *p*-value for association with oral microbiome composition in SPARK. $p_{dentures}$, *p*-value for association with dentures risk in UK Biobank (with a proxy variant indicated in parentheses for two variants); nominally significant associations ($p_{dentures}<0.05$) are indicated in bold.

composition are enriched at genes specifically expressed in salivary glands (p=0.02, **Extended Data Fig. 3a**).

- Two loci contained genes with established roles in immune function—the HLA class II genes, which encode proteins that present peptides in adaptive immunity, and *TLR1*, encoding Toll-like receptor 1, that binds bacterial lipoproteins in innate immunity. The strongest association at *TLR1* involved a missense variant (rs5743618; p=6.2x10^{-18}) that produces an I602S substitution known to inhibit trafficking of TLR1 to the cell surface, reducing monocyte response to bacterial agonists and reducing neutrophil priming in a recessive manner[41,42]. Consistent with these reports, I602S associated recessively with microbial abundances (p=6.7x10^{-29}; **Extended Data Fig. 3b**).

- Two other loci, *ABO* and *FUT2*, encode glycosyltransferases that together determine expression of histo-blood group antigens on epithelial cells and secreted proteins (in addition to the well-known role of *ABO* in determining blood type). This broader role is important to microbial species that interact with mucosal surfaces, such that both loci are known to influence the gut microbiome[11–14], with some bacterial species utilizing A-antigen saccharides as a carbohydrate source[27]. The variants at *ABO* and *FUT2* that associated most strongly with oral microbiome composition were rs2519093 (p=9.5x10^{-15}), which tags the A1 blood group[44], and rs601338 (p=1.6x10^{-131} additively, p=3.0x10^{-188} recessively), the common *FUT2* W154X nonsense variant that (in homozygotes) produces the non-secretor phenotype lacking histo-blood group antigens in bodily fluids.

- Associations of variants at *PITX1* with oral microbiome composition colocalized with previously-reported associations of these variants with dental caries and dentures use (r^2=0.99 between the top microbiome-associated variant (rs3749751; p=3.0x10^{-11}) and the top dentures-associated SNP at *PITX1*; **Extended Data Fig. 3c**; ref.[21]). *PITX1* is a developmentally-expressed gene that appears to have a role in mandibular tooth morphogenesis (based on a knockout mouse model)[43], suggesting that common genetic variation at *PITX1* might influence tooth morphology, and through it, oral microbiota and dental health.

The shared associations of genetic variants at *PITX1* with both oral microbiome composition and dental health phenotypes led us to wonder whether other genetic influences on the oral microbiome might similarly influence dental health. To explore this question, we performed a genome-wide association analysis of dentures use (a proxy for tooth loss and caries) in the UK Biobank cohort (n=75,156 cases, n=342,883 controls)[46]. Three loci—*AMY1*, *FUT2*, and *PITX1*—contained variants that associated (p<5x10^{-8}) with both oral microbiome composition and dentures use, and at each of these loci, the association patterns colocalized (**Fig. 2b** and **Extended Data Fig. 3c,d**). Moreover, at eight of the 11 loci influencing oral microbiome composition, the most strongly-associated variant also exhibited at least a nominally significant association (p<0.05) with dentures risk (**Table 1**), suggesting that host genetic effects on oral microbiome composition often have downstream effects on oral health. These associations with

dentures risk, together with the functional relevance of genes and variants implicated across the 11 loci, support the robustness of the associations with oral microbiome composition.

Most of these genetic associations appeared to involve effects of human genetic variation on the abundances of several bacterial species, with 141 species-genotype pairs reaching FDR<0.05 across the 11 loci (**Supplementary Tables 1 and 2**). The strong associations at *AMY1* and *FUT2* offered the opportunity for a more detailed investigation of how genetic variation at these loci influences oral microbiomes and oral health. *FUT2* W154X associated with the abundances of 42 of the 439 species (**Fig. 2c-e**). *FUT2* appeared to be nearly but not completely haplosufficient in these associations, with slightly weaker abundance-modifying effects observed among secretor individuals with heterozygous W154X genotype compared to those with two wild-type alleles (**Fig. 2d** and **Extended Data Fig. 3e**). These species only modestly clustered on the bacterial phylogeny (**Fig. 2e**), suggesting widespread interaction of bacteria with human histo-blood group antigens. For several pairs of closely-related species, *FUT2* W154X associated with increased abundance of one species and decreased abundance of the other (**Fig. 2e** and **Extended Data Fig. 4**), possibly reflecting competition between closely related species for ecological niches, with human genetic variation benefiting one species at the expense of the other.

Complex variation at the amylase locus influences microbial abundances and oral health

AMY1 encodes salivary α-amylase, an enzyme that breaks down dietary starches into simple sugars. The dramatic copy number expansion of the amylase locus in humans and other animals[47] has attracted much interest for its theorized role in facilitating recent adaptation to starch-based diets[48–50], but its reported association with human body mass index (BMI) and type 2 diabetes has been controversial[51–55]. *AMY1* copy-number genotypes in SPARK and UK Biobank (which we estimated from WGS depth-of-coverage) exhibited extensive polymorphism[48,51] (2–32 copies per individual; **Fig. 3a** and **Extended Data Fig. 5a**) and high mutability (6.3×10^{-4} [3.7×10^{-4}–11×10^{-4}, 95% CI] mutations per haplotype per generation, similar to an estimate using coalescent modeling[49]; **Extended Data Fig. 5b**).

Copy number variation of *AMY1* generated the strongest association of genetic variation at the amylase locus with oral microbiome composition ($p=1.5\times10^{-53}$; **Fig. 2b**), associating with the abundances of many bacterial species (31 species at FDR<0.05, 16 species at FDR<0.01 with $p=4.7\times10^{-29}$ to $p=0.00013$; **Fig. 3b**). The abundances of these species changed stepwise with *AMY1* copy number, generating a long allelic series with steadily increasing or decreasing abundances (**Fig. 3c**), congruent with the effect of *AMY1* copy number on the abundance of secreted salivary amylase[48,56]. Two of these associations appeared to confirm associations previously observed in smaller candidate-gene studies[53,57] (**Supplementary Note 2**).

AMY1 copy number also associated strongly with dentures use in UK Biobank ($p=5.9\times10^{-35}$, surpassed only by the *PITX1* locus; **Fig. 2b** and **Fig. 3d**). Each additional copy of *AMY1* associated with a 2.1% [1.7%–2.4%] increase in the odds of having dentures, corresponding to

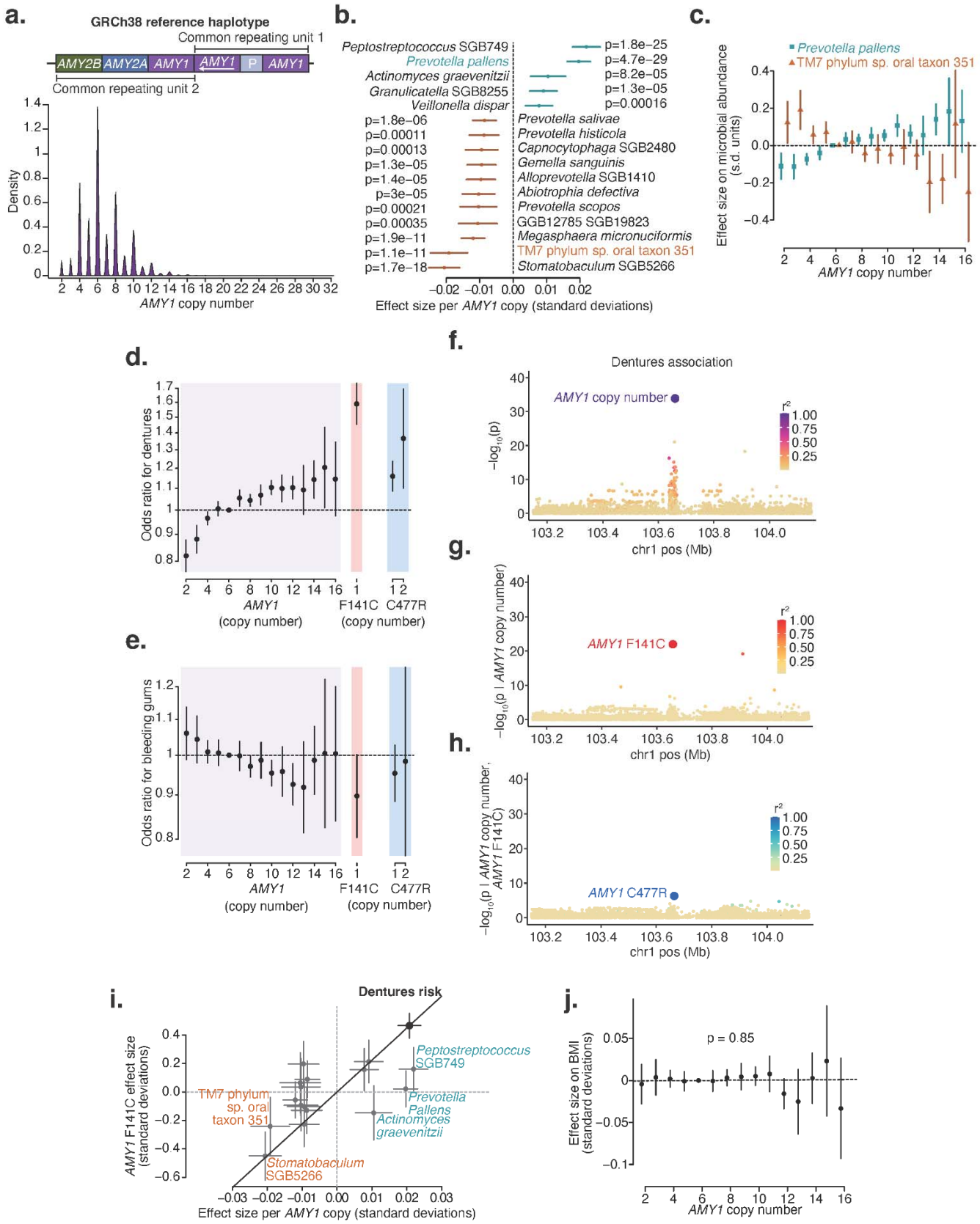


Figure 3. Complex genetic variation of the salivary amylase locus affects the abundance

of several oral microbial species and oral health. a, Distribution of *AMY1* diploid copy number estimates for UK Biobank participants (n=490,415). Inset, diagram of the amylase locus in the human reference genome. The wide range of copy numbers observed across the population is primarily driven by variation in the copy numbers of two cassettes[49,50]. **b**, Effect sizes on relative species abundance for the 16 species most strongly associated with host *AMY1* copy number (FDR<0.01). **c**, Allelic series of effect sizes of *AMY1* copy number genotypes on normalized abundances of *Prevotella pallens* and TM7 phylum sp. oral taxon 351. **d**, Odds ratios for risk of dentures use in UK Biobank (n=418,039) across copy number genotypes of *AMY1* (purple), *AMY1* F141C (red), and *AMY1* C477R (blue). **e**, Odds ratios for risk of bleeding gums in UK Biobank (n=418,039). **f**, Associations of variants at the amylase locus with dentures use. Plotted variants include paralogous sequence variants (PSVs) in the *AMY1* region (for which copy numbers of minor alleles were tested for association). Dot colors indicate LD with *AMY1* copy number. **g**, Associations with dentures use conditioned on *AMY1* copy number. **h**, Associations with dentures use additionally conditioned on *AMY1* F141C copy number. **i**, Comparison of effect sizes for *AMY1* copy number versus *AMY1* F141C copy number on relative abundances of 16 microbial species (from panel b) and on risk of dentures use (blue dot). For some species, the relative effect size of *AMY1* copy number versus *AMY1* F141C copy number on abundance is similar to this ratio for dentures use (blue line), whereas for others, it is not (e.g., *Prevotella pallens*). **j**, Effect sizes of *AMY1* copy number genotypes on BMI in UK Biobank (n=418,150). The line drawn is the best fit across *AMY1* copy numbers. Error bars, 95% CIs in all panels.

a 1.4-fold range in odds across people with 2 to 16 *AMY1* copies (**Fig. 3d**). This association replicated in the *All of Us* cohort[58] (n=230,002; p=3.5x10^{-4} for tooth loss, p=6.1x10^{-3} for caries; **Extended Data Fig. 5c-e**). Surprisingly, *AMY1* copy number associated with decreased risk of bleeding gums in UK Biobank (p=1.5x10^{-6}; **Fig. 3e**), even though gingivitis is considered a risk factor for tooth loss[59,60]. However, the bleeding gums and dentures use phenotypes had little genetic overlap[21] and were slightly negatively correlated (r = –0.07, s.e. 0.0015), suggesting largely independent pathology. These associations were specific to *AMY1*; the copy number of *AMY2A* and *AMY2B* (encoding pancreatic amylase) did not associate with dental phenotypes (**Extended Data Fig. 5f,g**).

Intriguingly, beyond the effect of *AMY1* copy number, two missense variants in *AMY1* carried by 1–2% of UK Biobank participants appeared to confer the largest increases in dentures risk of all common variants in the human genome (OR=1.59 [1.46–1.73] per copy of *AMY1* F141C (p=2.5x10^{-26}); OR=1.16 [1.10–1.23] per copy of *AMY1* C477R (p=8.3x10^{-8}); **Fig. 3d**). These two paralogous sequence variants, which we genotyped by consolidating WGS read alignments to *AMY1A*, *AMY1B*, and *AMY1C*, were typically carried on haplotypes containing three or four copies of *AMY1* and produced the strongest conditional associations with dentures use in two stages of stepwise conditional analysis (**Fig. 3f-h**). The *AMY1* F141C and C477R variants appeared to confer an increase in dentures risk equivalent to increasing *AMY1* copy number by 22.4 [18.3-26.5] and 7.3 [4.6-9.9] copies, respectively (**Fig. 3d**). This apparent gain-of-function effect was surprising given that both variants were predicted to be damaging (PolyPhen-2[61] score of 1.0) and affected amino acid residues far from the active site of amylase.

To investigate the molecular effects of these missense variants, we transfected HEK293T cells with plasmids carrying the reference coding sequence of *AMY1* or sequences encoding F141C and C477R mutant isoforms. Amylase protein expression was analyzed via western blot in both the supernatants and cellular lysates from each transfected line and subsequently purified for downstream enzymatic assays (see **Methods**). Cells with the reference sequence and the F141C sequence secreted amylase protein at similar levels, but cells with the C477R sequence did not; instead, the C477R amylase isoform appeared to be more glycosylated and retained intracellularly (**Extended Data Fig. 6a,b**). The F141C and reference amylase isoforms appeared to have equivalent enzymatic activity (t-test p=0.41, n=32 replicates of a fluorescent starch-degrading assay on purified secreted amylase; **Extended Data Fig. 6b,c**), suggesting that the F141C gain-of-function effect involves another mechanism, perhaps protein binding or activity on another substrate. Whether the retention and glycosylation of the C477R isoform *in vitro* is relevant to oral phenotypes remains to be determined.

The extended allelic series of *AMY1* copy number and missense variants associated with dentures use provided a set of genetic instruments for evaluating which bacterial species might causally contribute to tooth loss. *Prevotella pallens*, whose abundance associated most strongly with *AMY1* copy number (**Fig. 3b,c**), did not appear to be influenced by either *AMY1* F141C or C477R (**Fig. 3i** and **Extended Data Fig. 7a**), suggesting that *P. pallens* either does not causally contribute to dentures risk or does so in combination with other species. In contrast, other species, including a representative of the Saccharibacteria (TM7) phylum, which may reduce

gingival inflammation and bone loss[62], was negatively associated with *AMY1* copy number, F141C, and C477R, consistent with a protective species mediating the associations of these variants with dentures risk (**Fig. 3b,c,i** and **Extended Data Fig. 7a**). Reverse causality (i.e., dentures use causing changes in the oral microbiome that associate with *AMY1* variants) appeared to be unlikely based on the concordance of effect sizes in children and adults (**Extended Data Fig. 7b**).

The UK Biobank and *All of Us* data sets also enabled a rigorous evaluation of whether or not *AMY1* copy number influences body mass index (BMI) among modern humans. *AMY1* copy number did not associate with BMI in UK Biobank (n=418,150, p=0.85; **Fig. 3j**), *All of Us* (n=219,879, p=0.30, **Extended Data Fig. 7c**), or any genetic ancestry within *All of Us* (**Extended Data Fig. 7d-f**).

Genetic associations with bacterial gene dosage identify host-microbe protein interactions

To identify molecular mechanisms by which human genetic variation engages the oral microbiome, we next tested whether the microbial species whose abundances associated with each of the 11 loci might be united by shared biochemical pathway utilization (using HUMAnN[63]; **Extended Data Fig. 8a**, **Supplementary Tables 3 and 4**, and **Supplementary Note 3**). This analysis identified a gene encoding an adhesin in *Haemophilus sputorum* at which sequencing coverage associated much more strongly with *FUT2* W154X than in the rest of its genome, suggesting that the adhesin might interact with FUT2-dependent glycosylation (**Extended Data Fig. 8b** and **Supplementary Note 3**). To search more specifically for similar molecular interactions between human and bacterial proteins, we performed a set of association analyses in which we tested each of the 11 lead variants associated with oral microbiome composition (**Table 1**) for association with gene dosages within microbial genomes[27] (**Fig. 4a**). The key conceptual difference between testing for human genetic effects on microbial abundances (**Fig. 2a**) versus microbial gene dosages (**Fig. 4a**) is that the latter approach searches for effects on relative fitness of bacterial strains that do or do not contain a genomic region (rather than fitness of a bacterial species). Thus, it highlights microbial genomic regions that may contain genes whose products are involved in a host-microbe genetic interaction (**Supplementary Note 4)**.

To implement this approach, we searched specifically for associations of the 11 human genetic variants (**Table 1**) with quantitative measurements of normalized WGS coverage within 500bp regions tiled across 30 bacterial reference genomes (**Fig. 4a**, **Supplementary Table 5**, **Methods**, and **Supplementary Note 4**). This analysis identified 68 regions of 18 bacterial genomes in which normalized read-depth associated with one or more of the 11 human genetic variants (FDR<0.01, **Fig. 4b** and **Supplementary Table 6** and **7**). Normalized read-depth in nearly half of these regions (33/68) associated with *FUT2* secretor status. Eight regions associated with more than one human genetic variant; among them, five regions associated with both *FUT2* secretor status and *ABO*A1* (**Fig. 4b**).

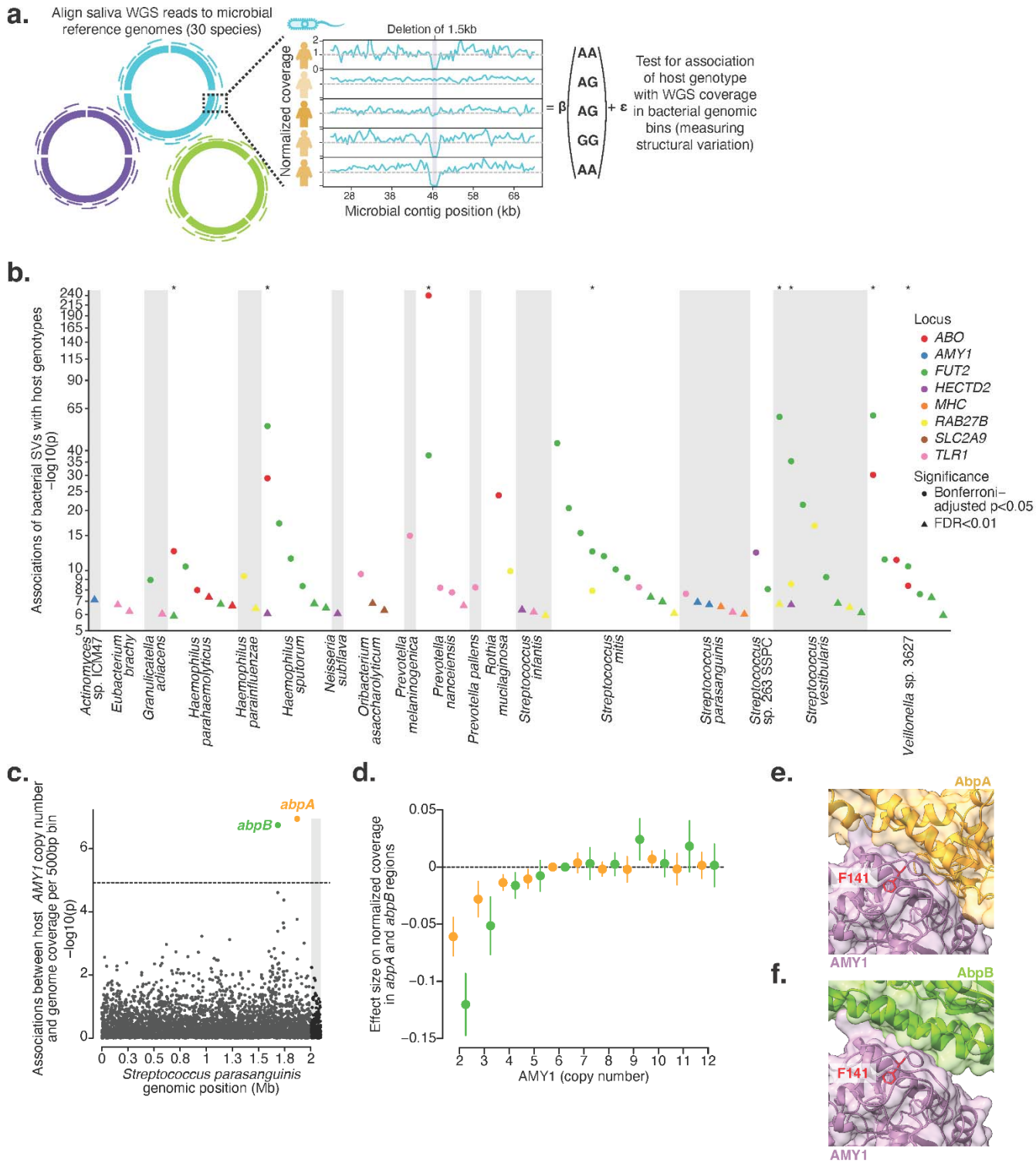


Figure 4. Numerous deletions in oral microbial genomes are selected for by human genetic variation. a, Approach to identify deletions in microbial reference genomes that associate with 11 human genetic variants that influence oral microbiome composition (**Table 1**). Sequencing reads were remapped to 30 microbial reference genomes, after which normalized WGS coverage was computed across 500bp genomic bins and truncated to a maximum of 1. **b**,

Associations between microbial structural variation (based on normalized WGS coverage) and human genetic variants. Each dot indicates a microbial SV region that associated with at least one human genetic variant (circles, Bonferroni-adjusted $p<0.05$; triangles, FDR<0.01). Dot colors indicate human loci, and SV regions that associated with more than one human locus are marked with an asterisk (top). **c**, Associations of deletions in the genome of *Streptococcus parasanguinis* (as estimated by normalized coverage) with *AMY1* copy number (n=12,340). Shading indicates the two assembled contigs of the reference genome. The two significant associations (FDR<0.01) overlap genes encoding amylase-binding proteins, *abpA* and *abpB*. **d**, Allelic series of effect sizes of *AMY1* copy number on normalized coverage in the 500bp bins overlapping *abpA* (orange) and *abpB* (green). Error bars, 95% CIs. **e**, Protein-protein interaction between human AMY1 (purple) and *S. parasanguinis* AbpA (orange) predicted by AlphaFold2. AMY1 residue F141 is highlighted in red. **f**, Analogous to **e**, for *S. parasanguinis* AbpB (green).

Streptococcal amylase-binding protein is selected for in humans with higher *AMY1* copy number

Two genomic regions of *Streptococcus parasanguinis*, a common and generally benign member of dental plaque, associated with host *AMY1* copy number, and these two genomic regions overlapped its *abpA* and *abpB* (amylase-binding protein) orthologs[64,65] (p=1.13x10^{-7} and 1.80x10^{-7}, respectively; **Fig. 4c**). The frequency at which these regions were present in oral microbiomes (based on normalized read-depth) increased with increasing host *AMY1* copy number, forming a non-linear allelic series with larger effects at lower copy numbers (**Fig. 4d**). This relationship mirrors the higher abundance of both *abp* and *AMY1* gene copies in hominins relative to other primates[66]. These associations were specific to *abpA* and *abpB*; abundance of *S. parasanguinis* did not associate with *AMY1* copy number (**Extended Fig. 8c**). Both *S. parasanguinis* amylase binding proteins were computationally predicted by AlphaFold2[67] to interact with the amylase protein face containing the F141C mutation (**Fig. 4e,f**), suggesting that the F141C substitution may modulate interactions with amylase binding proteins to yield its effects on oral health phenotypes.

Genetically encoded A-antigen expression selects for a *Prevotella* glycoside hydrolase

We found that host *ABO*A1* genotype (based on rs2519093) strongly influenced whether *Prevotella* strains – a prevalent oral genus involved in early biofilm formation[68] – carried a gene encoding a glycoside hydrolase likely capable of cleaving fucose groups, given its high amino acid homology (~75%) with α1,2-fucosidases in the glycoside hydrolase 95 (GH95) family (**Fig. 5a-d**). *ABO*A1* genotype associated exceptionally strongly (p=4.8x10^{-19}–8.3x10^{-241}) with normalized WGS coverage across a 3kb segment of the *P. nanceiensis* reference genome (**Fig. 5a**) – a region annotated as a glycoside hydrolase pseudogene due to an N-terminal truncation of the gene in the reference genome. However, assembly of unmapped sequencing reads with mates aligned to the region showed no evidence of such truncation: rather, the reads appeared to originate from a full-length gene, with 95% homology to a glycoside hydrolase found in *P. salivae* (a species not included among the 30 reference genomes analyzed). *ABO*A1* genotype did not associate with the abundance of either *P. nanceiensis* (p=0.39) or *P. salivae* (p=0.85).

*ABO*A1* genotype associated with sequencing coverage in this region only in secretor individuals (p=1.7x10^{-307} in secretors; p=0.44 in non-secretors) with at least one functional copy of *FUT2* that allows expression of histo-blood group antigens on epithelial cells and secreted proteins (**Fig. 5d**). This *FUT2* secretor-dependent effect of host blood group appeared to be driven specifically by A-antigen presentation: among secretors, the fraction of individuals for whom the glycoside hydrolase gene was detectable in saliva-derived DNA increased from 46–48% in non-secretors and individuals with B or O blood type to 71–77% in secretors with A or AB blood type (**Fig. 5b**). We found this association to further reflect the quantity of A-antigen predicted by an individual's diploid genotype, where *ABO*O*, *B*, *A1*, and *A2* alleles exhibited an allelic series of effects on normalized WGS coverage of the glycoside hydrolase gene that was consistent with the abilities of the glycosyltransferases encoded by these alleles to synthesize A-antigen (β=0.29 [0.28, 0.30] for *A1* relative to *O*; β=0.09 [0.07, 0.11] for *A2* relative to *O*; no

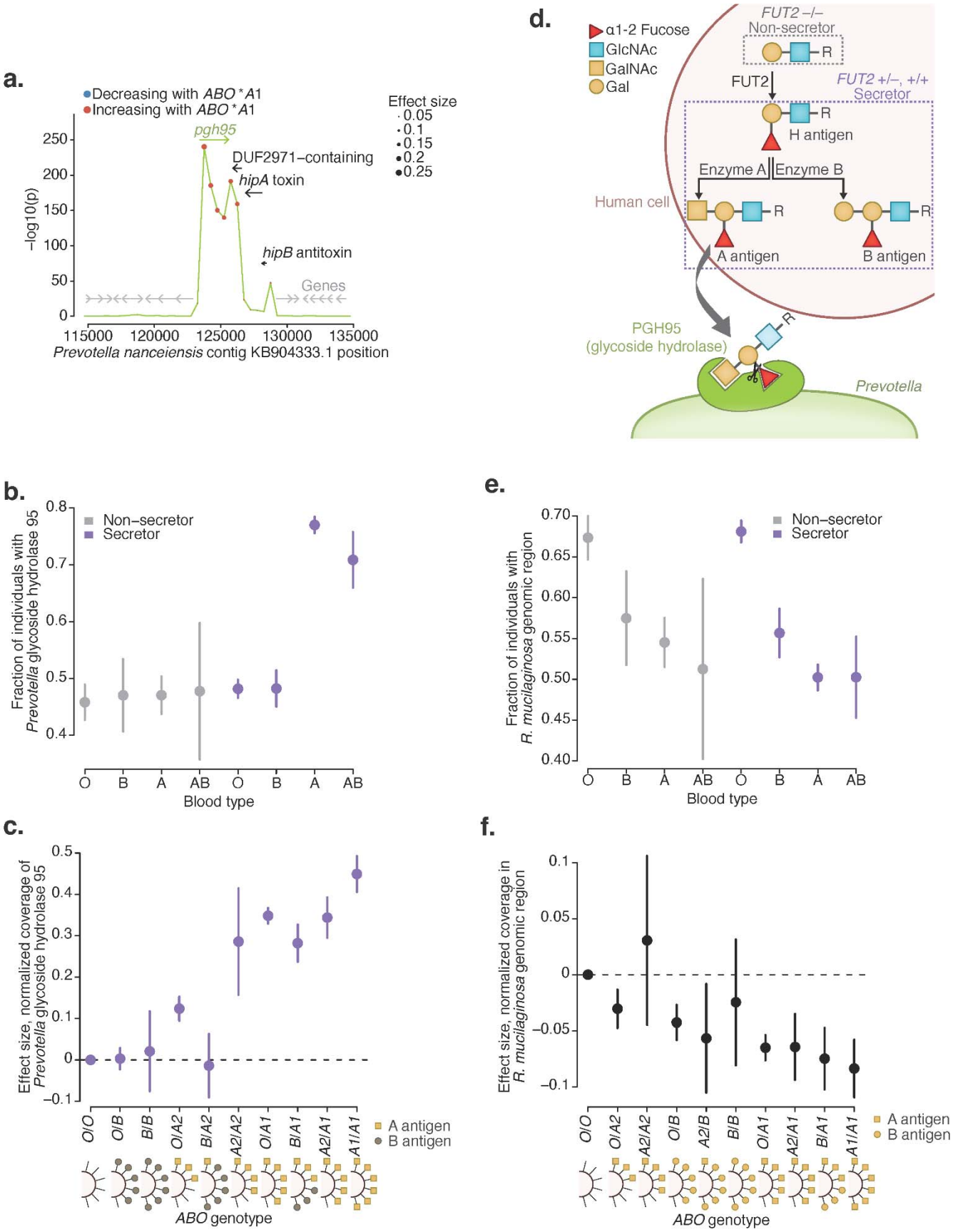

a.
Decreasing with ABO*A1
Increasing with ABO*A1
Effect size
0.05
0.1
0.15
0.2
0.25
pgh95
DUF2971–containing
hipA toxin
hipB antitoxin
Genes
–log10(p)
250
200
150
100
50
0
115000
120000
125000
130000
135000
Prevotella nanceiensis contig KB904333.1 position
d.
α1-2 Fucose
GlcNAc
GalNAc
Gal
FUT2 –/–
Non-secretor
FUT2
FUT2 +/–, +/+
Secretor
H antigen
Enzyme A
Enzyme B
Human cell
A antigen
B antigen
PGH95
(glycoside hydrolase)
Prevotella
b.
Non–secretor
Secretor
Fraction of individuals with Prevotella glycoside hydrolase 95
0.8
0.7
0.6
0.5
0.4
O
B
A
AB
O
B
A
AB
Blood type
e.
Non–secretor
Secretor
Fraction of individuals with R. mucilaginosa genomic region
0.70
0.65
0.60
0.55
0.50
0.45
0.40
O
B
A
AB
O
B
A
AB
Blood type
c.
Effect size, normalized coverage of Prevotella glycoside hydrolase 95
0.5
0.4
0.3
0.2
0.1
0
–0.1
O/O
O/B
B/B
O/A2
B/A2
A2/A2
O/A1
B/A1
A2/A1
A1/A1
A antigen
B antigen
ABO genotype
f.
Effect size, normalized coverage in R. mucilaginosa genomic region
0.1
0.05
0
–0.05
–0.1
O/O
O/A2
A2/A2
O/B
A2/B
B/B
O/A1
A2/A1
B/A1
A1/A1
A antigen
B antigen
ABO genotype

Figure 5. *ABO* variation modulates selection on a glycoside hydrolase gene in *Prevotella* and also engages the oral microbiome in non-secretors. a, Associations of host *ABO*A1* genotype with normalized coverage (truncated at 1) in 500bp bins of the *Prevotella nanceiensis* genome surrounding a glycoside hydrolase gene (n=10,433 in SPARK). Dot color indicates effect direction (red, higher coverage among individuals with A1 blood type) and dot size indicates effect magnitude. Arrows indicate genes: glycoside hydrolase *pgh95* (A3GM_RS0109435, green), other genes overlapping the region associated with *ABO*A1* (black), and nearby genes (gray). **b**, Proportion of individuals whose oral microbiomes carry the *pgh95* gene (n=10,433), stratified by blood type and secretor status (orange, secretor individuals with at least one functional copy of *FUT2*; gray, non-secretors). **c**, Effect sizes on normalized coverage in the *pgh95* region for genotype combinations of common blood type alleles (*O*, *B*, *A2*, and *A1*) relative to *O*/*O* individuals. Analyses were restricted to secretors (n=8,278). Effect sizes appear to reflect expected abundances of A antigens based on the combinations of blood type alleles as depicted in the diagrams below the x-axis (yellow squares, A antigens; gray circles; B antigens). **d**, Inferred interaction between glycosylation of host cell proteins and bacterial glycoside hydrolase. Top (pink background), glycosylation patterns of human mucosal cell surface proteins and secreted proteins depend on an individual's combination of *FUT2* and *ABO* genotypes. In individuals with no functional copies of *FUT2* (non-secretors), Type I H-antigen is not produced, whereas in secretors, Type I H-antigen is produced and can be further glycosylated into A-antigen or B-antigen depending on *ABO* genotype (dashed orange outline). These antigens are then presented on mucosal cell surface and secreted proteins. The associations of *ABO* genotypes with presence of the *pgh95* gene in *Prevotella* strains suggests that the bacterial glycoside hydrolase protein (PGH95, green) is specifically targeting secreted type A antigens and cleaving the α1,2-fucosyl group. **e**, Analogous to **b**, for the *ABO*-associated SV region in *Rothia mucilaginosa* (n=12,475). **f**, Analogous to **c**, for the *ABO*-associated region in *Rothia mucilaginosa*, analyzing secretors and non-secretors together (n=12,475). Error bars, 95% CIs in all panels.

effect of *B* relative to *O*; larger effect of *A*/*A* homozygotes than *A*/*O* heterozygotes relative to *O*). The *B* allele imposed a strong opposing effect when present in an individual heterozygous for an *A1* or *A2* allele (β=-0.074 [-0.12, -0.032] for *B* relative to *O*, p=5.8x10^{-4}, **Fig. 5c**), presumably reflecting competition between A and B transferases for available galactose residues on acceptor H-antigens (**Fig. 5d**).

Taken together, these results suggest that the glycoside hydrolase enables *Prevotella* strains that express it to use type A histo-blood group antigens presented on host mucosal cell surfaces or salivary proteins (in secretors) as a carbohydrate source, similar to a recently observed effect in the gut microbiome[27]. We hypothesize that the glycoside hydrolase binds A-antigens and cleaves the α1,2-fucosyl group synthesized by FUT2 (**Fig. 5d**). Host *ABO* genotypes exhibited an intriguingly different pattern of association with a genomic region of the most abundant species across our cohort, *Rothia mucilaginosa* (p=1.4x10^{-24}; **Fig. 5e,f**). Blood groups A, B, and AB all associated with absence (rather than presence) of this region of the *R. mucilaginosa* genome, and surprisingly, these associations were observed in non-secretors as well as secretors (**Fig. 5e,f**). This region contains genes that encode a protein with no annotated domains and a 3-isopropylmalate dehydrogenase functioning in leucine biosynthesis, leaving the mechanism of association to be determined.

More broadly, the observation of a non-*FUT2*-dependent *ABO* association led us to wonder whether the association of *ABO**A1* genotype with microbiome composition (**Fig. 2b**) might also be partially independent from secretor status. Indeed, within non-secretors the *ABO**A1* association with microbiome composition remained significant (p=0.004). This suggests that some effects of *ABO* variation on the oral microbiome come from cells not dependent on FUT2 for H-antigen production (e.g., blood and endothelial cells, which instead use FUT1). For example, bacteria can produce glycans structurally similar to A- or B-antigens that can then be recognized by anti-A or anti-B antibodies that are made by plasma cells and infiltrate into the mouth[69,70].

Several oral microbial species have host glycosylation-binding adhesins selected for by *FUT2* secretor status

Conversely, many regions in oral microbial genomes associated with *FUT2* secretor status but not *ABO**A1* genotype (**Fig. 4b**), and oral microbiome composition associated much more strongly with *FUT2* secretor status than with any variant at *ABO* (p=3.0x10^{-188} versus p=9.4x10^{-15}; **Fig. 2b**). This pattern contrasted with previous studies of host genetic influences on gut microbiomes (which have generally observed stronger associations at *ABO* than at *FUT2*[12–14]), leading us to wonder whether these regions might point to a molecular mechanism by which *FUT2* secretor status influences oral microbiomes independently of *ABO*.

Examining genes within bacterial genomic regions associated with *FUT2* secretor status identified three classes of bacterial proteins that were each implicated by several genes. Proteins with YadA-like domains were encoded by nine genes in three species: *Veillonella* sp. 3627 (*vadA* through *vadF*), *Haemophilus sputorum* (*hadA* and *hadB*), and *Haemophilus*

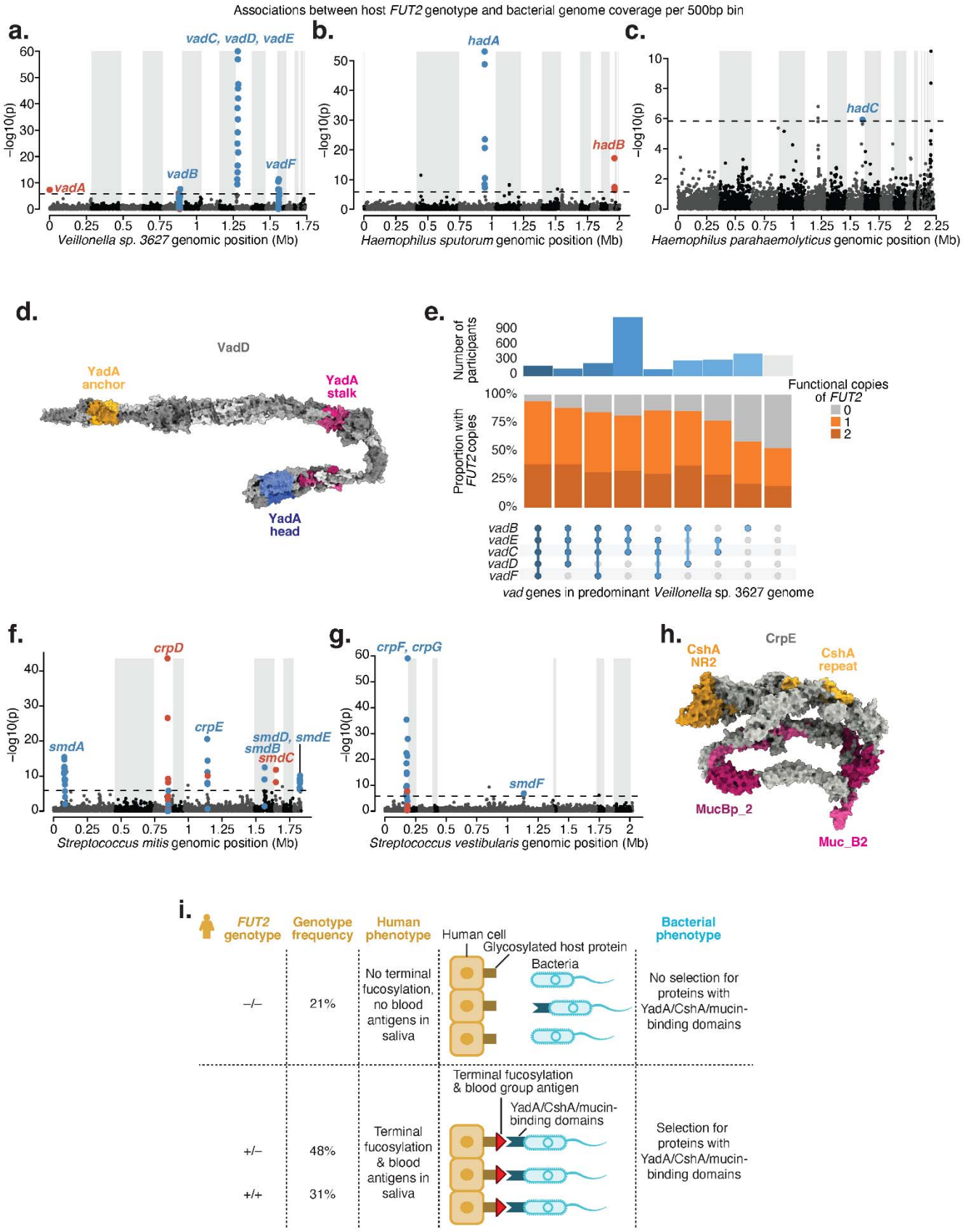

Associations between host FUT2 genotype and bacterial genome coverage per 500bp bin
a.
vadC, vadD, vadE
vadA
vadB
vadF
−log10(p)
Veillonella sp. 3627 genomic position (Mb)
b.
hadA
hadB
Haemophilus sputorum genomic position (Mb)
c.
hadC
Haemophilus parahaemolyticus genomic position (Mb)
d.
VadD
YadA anchor
YadA stalk
YadA head
e.
Number of participants
Proportion with FUT2 copies
Functional copies of FUT2
0
1
2
vadB
vadE
vadC
vadD
vadF
vad genes in predominant Veillonella sp. 3627 genome
f.
crpD
crpE
smdA
smdD, smdE
smdB
smdC
Streptococcus mitis genomic position (Mb)
g.
crpF, crpG
smdF
Streptococcus vestibularis genomic position (Mb)
h.
CrpE
CshA NR2
CshA repeat
MucBp_2
Muc_B2
i.
FUT2 genotype
Genotype frequency
Human phenotype
Human cell
Glycosylated host protein
Bacteria
Bacterial phenotype
–/–
21%
No terminal fucosylation, no blood antigens in saliva
No selection for proteins with YadA/CshA/mucin-binding domains
Terminal fucosylation & blood group antigen
YadA/CshA/mucin-binding domains
+/–
48%
+/+
31%
Terminal fucosylation & blood antigens in saliva
Selection for proteins with YadA/CshA/mucin-binding domains

Figure 6. *FUT2* secretor status selects for three classes of adhesin genes across five microbial species. a, Associations of *FUT2* secretor status with normalized coverage (truncated at 1) in 500bp bins of the *Veillonella* sp. 3627 genome (n=7419). Shading indicates assembled contigs. Significant associations (FDR<0.01) that overlap genes encoding proteins with YadA-like domains are highlighted (blue, genomic region more often present in secretors; red, genomic region more often deleted in secretors). Effect directions are also indicated for bins that did not reach significance but were surrounded by significantly-associated bins. **b**, Analogous to **a**, for *Haemophilus sputorum* (n=8153). **c**, Analogous to **a**, for *Haemophilus parahaemolyticus* (n=7456). **d**, Predicted trimeric structure of VadD (from *V.* sp. 3627), where the head domain (blue) facilitates attachment to host proteins, stalk domains (magenta) flexibility and reach, and anchor domain (yellow) translocation to bacterial surface. **e**, Upset plot of the relative proportions of *FUT2* secretor genotypes (non-secretors in gray, secretors in orange) among individuals with each combination of *vadB*–*vadF* gene deletions in the *V.* sp. 3627 genome. Analysis was restricted to individuals with each gene either primarily present in strains of *V*. sp. 3627 (normalized coverage>0.8) or primarily absent (normalized coverage <0.2). Blue-to-gray shading of sets (bottom) and of bars indicating numbers of individuals per set (top) indicates the number of *vad* genes present. **f**, Analogous to **a**, for *Streptococcus mitis* (n=12,479). Highlighted genes encode proteins that contain either a CshA domain (*crp* genes) or mucin-binding domain (*smd* genes). **g**, Analogous to **f**, for *Streptococcus vestibularis* (n=11,723). **h**, Predicted structure of a portion of CrpE from *S. mitis*. The CshA NR2 (yellow) and mucin-binding domains (magenta) both have lectin activity to their characterized ligands (fibronectin and mucin)[73,74]. **i**, Model of how host *FUT2* genotype selects for bacterial strains expressing proteins with YadA, CshA, or mucin-binding domains that can attach to host cell surface proteins based on the availability of histo-blood type antigens.

parahaemolyticus (*hadC*) (**Fig. 6a-d**). YadA (from *Yersinia pestis*) is a trimeric autotransporter adhesin that aids attachment to host cells by binding components of the extracellular matrix, and some such adhesins are known to recognize host protein glycosylation[71,72] such as the glycosylation added or enabled by FUT2. Five of the seven regions containing these genes were present (i.e., not deleted) more often in the oral microbiomes of secretors than non-secretors, consistent with the hypothesis that the adhesins they encode bind histo-blood group antigens on the host cell surface. This was true of *hadC* in *H. parahaemolyticus* despite this species exhibiting lower abundance in secretors (**Extended Data Fig. 8d**). The genome of *V.* sp. 3627 contained three such regions (containing *vadB* through *vadF*, where *vadC* through *vadE* fall within the same complex region, **Extended Data Fig. 8e**) whose presence or absence was observed largely independently in different microbiomes (**Fig. 6e**). Classifying individuals based on which combination of regions was present within their *V.* sp. 3627 population showed increasing enrichment of secretors among individuals with increasing representation of *vadB*–*vadF* genes in *V.* sp. 3627 (**Fig. 6e**).

FUT2-associated genomic regions additionally implicated two other classes of proteins that appeared to have roles in adhesion to host cells. Four proteins with CshA domains (CrpD and CrpE in *Streptococcus. mitis*, CrpF and CrpG in *S. vestibularis*) and six proteins with mucin-binding domains (MucBP, Muc_B2, MucBP_2) (SmdA through SmdE in *S. mitis*, SmdF in *S. vestibularis*) were encoded by genes in *FUT2*-associated bacterial genomic regions (**Fig. 6f,g**). CshA from *Streptococcus gordonii* binds host fibronectin[73], a heavily glycosylated component of the extracellular matrix. Similarly, mucins have numerous glycosylation sites and are up to 90% carbohydrate by mass[74]. Interestingly, one of these proteins, CrpE in *S. mitis*, appears to contain both a CshA domain and multiple mucin-binding domains (**Fig. 6h**), suggesting that these might function in concert to bind the same host protein or a combination of proximal targets multivalently.

These enrichments of genes encoding proteins with YadA, CshA, and mucin-binding domains were unlikely to occur by chance; the genomes of *V.* sp. 3627, *H. sputorum*, and *H. parahaemolyticus* only contain twelve, four, and eight genes with YadA domains, respectively (Fisher's exact p=7.5x10^{-12}, 3.5x10^{-5}, and 0.021), and the genomes of *S. mitis* and *S. vestibularis* only contain two and three genes with CshA domains (p=7.4x10^{-5} and 2.0x10^{-5}) and ten and three genes with mucin-binding domains (p=3.3x10^{-11} and 0.0087). Most of these genes (15/19) were more commonly present in the oral microbiomes of people who had functional *FUT2*, suggesting that they might encode bacterial lectins that depend on either fucosylation or sugar moieties added by ABO glycotransferase[75] (**Fig. 6i**), whereas the four that are more common in non-secretors might be complementary adhesins that do not depend on glycosylation. This convergence of bacterial genomic adaptations to host *FUT2* genotype broadly suggests that commensal bacteria commonly make use of host histo-blood group antigens not only as a carbohydrate source but also for bacterial attachment to host cell surfaces.

Discussion

This analysis of the largest set of oral microbiome profiles generated to date has identified many specific human genetic variants that contribute to the diversity observed among oral microbiomes[2–4]. The large number of such effects suggests a larger influence of human genetics on the oral microbiome than on the gut microbiome[76,77], perhaps because host cells in the mouth interface more directly with bacteria (in contrast to cells in the gut, which are typically protected by the gut mucosal barrier). Some of these genetic effects on microbial abundances appear likely to mediate associations of the same variants with oral health phenotypes, nominating bacterial species that may contribute to dysbiosis. The salivary amylase gene generates the strongest such shared effect on oral microbiomes and health, driven by both *AMY1* copy number variation and missense mutations in *AMY1*. The expansion of salivary amylase copy number in humans and domesticated animals has been hypothesized to be the result of positive selection driven by the advent of agriculture[47–50]. Our observation here that *AMY1* copy number variation associates with oral phenotypes—but not body mass index—suggests that *AMY1* copy number may instead have been under selection due to effects on oral health, which is conceivable given the high mortality rate of tooth infections before the advent of modern dentistry[17].

The numerous associations we observed between human genetic variants and bacterial gene dosages suggest frequent intergenomic adaptation of microbial species to human hosts and implicate specific molecular interactions likely to drive such adaptation. Most of these associations involved genes within bacterial species whose overall abundances were unaffected by the same human genetic variants, suggesting that genomic adaptations enable many bacterial species to survive equally well across variable host genetic environments. In contrast, a relative abundance association could imply that the microbial genome is unable to flexibly adapt to a particular human variation. The variable gene regions we identified exhibit some breakpoint heterogeneity (**Extended Data Fig. 9a**,**b**) and could either reflect structural variation among circulating strains or recurrent structural mutations, such as in *Helicobacter pylori*[78]. The large number of such effects—observed among just 11 human genetic variants that we tested for association with gene dosages of 30 bacterial species—suggests that this analytical approach may be a powerful way to identify host genetic influences on microbiomes, perhaps because analyzing the balance between members of the same species with and without a variable gene controls for strong environmental influences on species abundance.

We do note a need for care in conducting genetic association studies of microbial abundance phenotypes. We initially observed a strong association ($p=2.5\times10^{-70}$) of oral microbiome composition with variant calls in the ribosomal RNA gene region of the p-arm of chromosome 21; however, these variant calls (which later failed a mappability filter) actually reflected the presence of orthologous bovine rDNA sequences and associated with the abundances of bacterial species used in dairy fermentation, suggesting DNA co-acquisition from recently eaten dairy foods (**Supplementary Note 5**). A recent oral microbiome GWAS[9] reported two associations in centromeric regions that did not replicate here (p=0.68 for rs4911713 and p=0.19 for rs36186689). Additionally, caution is needed in interpreting shared associations of

genetic variants with microbial abundances and oral health phenotypes given the possibilities of pleiotropy and reverse causality. Future data sets will provide increased power to resolve these questions, either through cohorts with human genetic, microbiome, and oral health phenotypes[79] or by Mendelian randomization approaches powered by even larger saliva sequencing data sets—which we have shown here provide rich information about how oral microbiomes are shaped by human genetics.

References

1. Dobell, C. Antony van Leeuwenhoek and His ‘Little Animals’: Being Some Account of the Father of Protozoology and Bacteriology and His Multifarious Discoveries in These Disciplines. (Harcourt, Brace and Company, 1932).
2. Odendaal, M.-L. et al. Host and environmental factors shape upper airway microbiota and respiratory health across the human lifespan. Cell **187**, 4571-4585.e15 (2024).
3. Huttenhower, C. et al. Structure, function and diversity of the healthy human microbiome. Nature **486**, 207–214 (2012).
4. Dewhirst, F. E. et al. The Human Oral Microbiome. J. Bacteriol. **192**, 5002–5017 (2010).
5. Baker, J. L., Mark Welch, J. L., Kauffman, K. M., McLean, J. S. & He, X. The oral microbiome: diversity, biogeography and human health. Nat. Rev. Microbiol. **22**, 89–104 (2024).
6. Gomez, A. et al. Host Genetic control of the oral microbiome in Health and Disease. Cell Host Microbe **22**, 269-278.e3 (2017).
7. Esberg, A., Haworth, S., Kuja-Halkola, R., Magnusson, P. K. E. & Johansson, I. Heritability of Oral Microbiota and Immune Responses to Oral Bacteria. Microorganisms **8**, 1126 (2020).
8. Demmitt, B. A. et al. Genetic influences on the human oral microbiome. BMC Genomics **18**, 659 (2017).
9. Liu, X. et al. Metagenome-genome-wide association studies reveal human genetic impact on the oral microbiome. Cell Discov. **7**, 1–16 (2021).
10. Stankevic, E. et al. Genome-wide association study identifies host genetic variants influencing oral microbiota diversity and metabolic health. Sci. Rep. **14**, 14738 (2024).
11. Kurilshikov, A. et al. Large-scale association analyses identify host factors influencing human gut microbiome composition. Nat. Genet. **53**, 156–165 (2021).

12. Rühlemann, M. C. et al. Genome-wide association study in 8,956 German individuals identifies influence of ABO histo-blood groups on gut microbiome. Nat. Genet. **53**, 147–155 (2021).
13. Lopera-Maya, E. A. et al. Effect of host genetics on the gut microbiome in 7,738 participants of the Dutch Microbiome Project. Nat. Genet. **54**, 143–151 (2022).
14. Qin, Y. et al. Combined effects of host genetics and diet on human gut microbiota and incident disease in a single population cohort. Nat. Genet. **54**, 134–142 (2022).
15. Sanna, S., Kurilshikov, A., van der Graaf, A., Fu, J. & Zhernakova, A. Challenges and future directions for studying effects of host genetics on the gut microbiome. Nat. Genet. **54**, 100–106 (2022).
16. Pitts, N. B. et al. Dental caries. Nat. Rev. Dis. Primer **3**, 1–16 (2017).
17. Thomas, T. T. III. Ludwig's Angina (Part II): An Anatomical, Clinical and Statistical Study. Ann. Surg. **47**, 335–373 (1908).
18. Shaffer, J. R. et al. Clustering Tooth Surfaces into Biologically Informative Caries Outcomes. J. Dent. Res. **92**, 32–37 (2013).
19. Wang, X. et al. Genes and their effects on dental caries may differ between primary and permanent dentitions. Caries Res. **44**, 277–284 (2010).
20. Haworth, S. et al. Heritability of Caries Scores, Trajectories, and Disease Subtypes. J. Dent. Res. **99**, 264–270 (2020).
21. Shungin, D. et al. Genome-wide analysis of dental caries and periodontitis combining clinical and self-reported data. Nat. Commun. **10**, 2773 (2019).
22. Bordenstein, S. R. & Theis, K. R. Host Biology in Light of the Microbiome: Ten Principles of Holobionts and Hologenomes. PLoS Biol. **13**, e1002226 (2015).
23. Kilian, M. et al. The oral microbiome – an update for oral healthcare professionals. Br. Dent. J. **221**, 657–666 (2016).

24. Chen, L. et al. The long-term genetic stability and individual specificity of the human gut microbiome. Cell **184**, 2302-2315.e12 (2021).
25. Suzuki, T. A. et al. Codiversification of gut microbiota with humans. Science **377**, 1328–1332 (2022).
26. Zeevi, D. et al. Structural variation in the gut microbiome associates with host health. Nature **568**, 43–48 (2019).
27. Zhernakova, D. V. et al. Host genetic regulation of human gut microbial structural variation. Nature **625**, 813–821 (2024).
28. SPARK: A US Cohort of 50,000 Families to Accelerate Autism Research. Neuron **97**, 488–493 (2018).
29. Manghi, P. et al. Large-scale metagenomic analysis of oral microbiomes reveals markers for autism spectrum disorders. Nat. Commun. **15**, 9743 (2024).
30. Valles-Colomer, M. et al. The person-to-person transmission landscape of the gut and oral microbiomes. Nature **614**, 125–135 (2023).
31. Kidd, J. M. et al. Exome capture from saliva produces high quality genomic and metagenomic data. BMC Genomics **15**, 262 (2014).
32. Samson, C. A., Whitford, W., Snell, R. G., Jacobsen, J. C. & Lehnert, K. Contaminating DNA in human saliva alters the detection of variants from whole genome sequencing. Sci. Rep. **10**, 19255 (2020).
33. Blanco-Míguez, A. et al. Extending and improving metagenomic taxonomic profiling with uncharacterized species using MetaPhlAn 4. Nat. Biotechnol. **41**, 1633–1644 (2023).
34. Stearns, J. C. et al. Culture and molecular-based profiles show shifts in bacterial communities of the upper respiratory tract that occur with age. ISME J. **9**, 1246–1259 (2015).
35. Zhao, N. et al. Testing in Microbiome-Profiling Studies with MiRKAT, the Microbiome Regression-Based Kernel Association Test. Am. J. Hum. Genet. **96**, 797–807 (2015).

36. Anderson, M. J. A new method for non-parametric multivariate analysis of variance. Austral Ecol. **26**, 32–46 (2001).
37. Calle, M. L. Statistical Analysis of Metagenomics Data. Genomics Inform. **17**, e6 (2019).
38. Rumker, L. et al. Identifying genetic variants that influence the abundance of cell states in single-cell data. Nat. Genet. **56**, 2068–2077 (2024).
39. Finucane, H. K. et al. Heritability enrichment of specifically expressed genes identifies disease-relevant tissues and cell types. Nat. Genet. **50**, 621–629 (2018).
40. Vitart, V. et al. SLC2A9 is a newly identified urate transporter influencing serum urate concentration, urate excretion and gout. Nat. Genet. **40**, 437–442 (2008).
41. Whitmore, L. C. et al. A Common Genetic Variant in TLR1 Enhances Human Neutrophil Priming and Impacts Length of Intensive Care Stay in Pediatric Sepsis. J. Immunol. Baltim. Md 1950 **196**, 1376–1386 (2016).
42. Johnson, C. M. et al. Cutting Edge: A Common Polymorphism Impairs Cell Surface Trafficking and Functional Responses of TLR1 but Protects against Leprosy1. J. Immunol. **178**, 7520–7524 (2007).
43. Mitsiadis, T. A. & Drouin, J. Deletion of the Pitx1 genomic locus affects mandibular tooth morphogenesis and expression of the Barx1 and Tbx1 genes. Dev. Biol. **313**, 887–896 (2008).
44. Goumidi, L. et al. Association between ABO haplotypes and the risk of venous thrombosis: impact on disease risk estimation. Blood **137**, 2394–2402 (2021).
45. Kelly, R. J., Rouquier, S., Giorgi, D., Lennon, G. G. & Lowe, J. B. Sequence and Expression of a Candidate for the Human Secretor Blood Group α(1,2)Fucosyltransferase Gene (FUT2): HOMOZYGOSITY FOR AN ENZYME-INACTIVATING NONSENSE MUTATION COMMONLY CORRELATES WITH THE NON-SECRETOR PHENOTYPE (∗). J. Biol. Chem. **270**, 4640–4649 (1995).

46. Bycroft, C. et al. The UK Biobank resource with deep phenotyping and genomic data. Nature **562**, 203–209 (2018).
47. Pajic, P. et al. Independent amylase gene copy number bursts correlate with dietary preferences in mammals. eLife **8**, e44628 (2019).
48. Perry, G. H. et al. Diet and the evolution of human amylase gene copy number variation. Nat. Genet. **39**, 1256–1260 (2007).
49. Bolognini, D. et al. Recurrent evolution and selection shape structural diversity at the amylase locus. Nature **634**, 617–625 (2024).
50. Yılmaz, F. et al. Reconstruction of the human amylase locus reveals ancient duplications seeding modern-day variation. Science **0**, eadn0609 (2024).
51. Usher, C. L. et al. Structural forms of the human amylase locus and their relationships to SNPs, haplotypes and obesity. Nat. Genet. **47**, 921–925 (2015).
52. Falchi, M. et al. Low copy number of the salivary amylase gene predisposes to obesity. Nat. Genet. **46**, 492–497 (2014).
53. Hasegawa, T. et al. Impact of salivary and pancreatic amylase gene copy numbers on diabetes, obesity, and functional profiles of microbiome in Northern Japanese population. Sci. Rep. **12**, 7628 (2022).
54. Mejía-Benítez, M. A. et al. Beneficial effect of a high number of copies of salivary amylase AMY1 gene on obesity risk in Mexican children. Diabetologia **58**, 290–294 (2015).
55. Shwan, N. A. A. & Armour, J. A. L. No Evidence for Association of BMI with Salivary Amylase Gene Copy Number in the UK 1958 Birth Cohort. Obesity **27**, 1533–1538 (2019).
56. Carpenter, D., Mitchell, L. M. & Armour, J. A. L. Copy number variation of human AMY1 is a minor contributor to variation in salivary amylase expression and activity. Hum. Genomics **11**, 2 (2017).
57. Poole, A. C. et al. Human Salivary Amylase Gene Copy Number Impacts Oral and Gut Microbiomes. Cell Host Microbe **25**, 553-564.e7 (2019).

58. Bick, A. G. et al. Genomic data in the All of Us Research Program. Nature **627**, 340–346 (2024).
59. Chapple, I. L. C. et al. Primary prevention of periodontitis: managing gingivitis. J. Clin. Periodontol. **42**, S71–S76 (2015).
60. Lang, N. P., Schätzle, M. A. & Löe, H. Gingivitis as a risk factor in periodontal disease. J. Clin. Periodontol. **36**, 3–8 (2009).
61. Adzhubei, I. A. et al. A method and server for predicting damaging missense mutations. Nat. Methods **7**, 248–249 (2010).
62. Chipashvili, O. et al. Episymbiotic Saccharibacteria suppresses gingival inflammation and bone loss in mice through host bacterial modulation. Cell Host Microbe **29**, 1649-1662.e7 (2021).
63. Beghini, F. et al. Integrating taxonomic, functional, and strain-level profiling of diverse microbial communities with bioBakery 3. eLife **10**, e65088 (2021).
64. Gwynn, J. P. & Douglas, C. W. I. Comparison of amylase-binding proteins in oral streptococci. FEMS Microbiol. Lett. **124**, 373–379 (1994).
65. Scannapieco, F. A. Saliva-Bacterium Interactions in Oral Microbial Ecology. Crit. Rev. Oral Biol. Med. **5**, 203–248 (1994).
66. Fellows Yates, J. A. et al. The evolution and changing ecology of the African hominid oral microbiome. Proc. Natl. Acad. Sci. U. S. A. **118**, e2021655118 (2021).
67. Jumper, J. et al. Highly accurate protein structure prediction with AlphaFold. Nature **596**, 583–589 (2021).
68. Tett, A., Pasolli, E., Masetti, G., Ercolini, D. & Segata, N. Prevotella diversity, niches and interactions with the human host. Nat. Rev. Microbiol. **19**, 585–599 (2021).
69. Saneshige, Y. & Woodfield, D. G. Detection of ABO Isoagglutinins in Saliva Using Toluene. J. Forensic Sci. Soc. **24**, 489–493 (1984).

70. Bell jr., C. E. & Fortwengler, H. P. Salivary Anti-A and Anti-B Activity of Group O Males. Vox Sang. **21**, 493–508 (1971).
71. Atack, J. M. et al. The Nontypeable Haemophilus influenzae Major Adhesin Hia Is a Dual-Function Lectin That Binds to Human-Specific Respiratory Tract Sialic Acid Glycan Receptors. mBio **11**, e02714-20 (2020).
72. Tram, G. et al. The Acinetobacter baumannii Autotransporter Adhesin Ata Recognizes Host Glycans as High-Affinity Receptors. ACS Infect. Dis. **7**, 2352–2361 (2021).
73. Back, C. R. et al. The Streptococcus gordonii Adhesin CshA Protein Binds Host Fibronectin via a Catch-Clamp Mechanism. J. Biol. Chem. **292**, 1538 (2016).
74. Chatterjee, M. et al. Understanding the adhesion mechanism of a mucin binding domain from Lactobacillus fermentum and its role in enteropathogen exclusion. Int. J. Biol. Macromol. **110**, 598–607 (2018).
75. Poole, J., Day, C. J., von Itzstein, M., Paton, J. C. & Jennings, M. P. Glycointeractions in bacterial pathogenesis. Nat. Rev. Microbiol. **16**, 440–452 (2018).
76. Rothschild, D. et al. Environment dominates over host genetics in shaping human gut microbiota. Nature **555**, 210–215 (2018).
77. Goodrich, J. K. et al. Human genetics shape the gut microbiome. Cell **159**, 789–799 (2014).
78. Suerbaum, S. & Josenhans, C. Helicobacter pylori evolution and phenotypic diversification in a changing host. Nat. Rev. Microbiol. **5**, 441–452 (2007).
79. Divaris, K. et al. Cohort Profile: ZOE 2.0—A Community-Based Genetic Epidemiologic Study of Early Childhood Oral Health. Int. J. Environ. Res. Public. Health **17**, 8056 (2020).
80. Zhou, X. et al. Integrating de novo and inherited variants in 42,607 autism cases identifies mutations in new moderate-risk genes. Nat. Genet. **54**, 1305–1319 (2022).
81. Lloyd-Price, J. et al. Strains, functions and dynamics in the expanded Human Microbiome Project. Nature **550**, 61–66 (2017).

82. Huttenhower, C. et al. Structure, function and diversity of the healthy human microbiome. Nature **486**, 207–214 (2012).
83. Finucane, H. K. et al. Partitioning heritability by functional annotation using genome-wide association summary statistics. Nat. Genet. **47**, 1228–1235 (2015).
84. Hujoel, M. L. A. et al. Insights into the causes and consequences of DNA repeat expansions from 700,000 biobank participants. 2024.11.25.625248 Preprint at https://doi.org/10.1101/2024.11.25.625248 (2024).
85. Loh, P.-R., Kichaev, G., Gazal, S., Schoech, A. P. & Price, A. L. Mixed-model association for biobank-scale datasets. Nat. Genet. **50**, 906–908 (2018).
86. Li, S., Carss, K. J., Halldorsson, B. V., Cortes, A. & Consortium, U. B. W.-G. S. Whole-genome sequencing of half-a-million UK Biobank participants. 2023.12.06.23299426 Preprint at https://doi.org/10.1101/2023.12.06.23299426 (2023).
87. Handsaker, R. E. et al. Large multiallelic copy number variations in humans. Nat. Genet. **47**, 296–303 (2015).
88. Hujoel, M. L. A. et al. Protein-altering variants at copy number-variable regions influence diverse human phenotypes. Nat. Genet. **56**, 569–578 (2024).
89. Matsuda, T. & Cepko, C. L. Electroporation and RNA interference in the rodent retina in vivo and in vitro. Proc. Natl. Acad. Sci. U. S. A. **101**, 16–22 (2004).
90. Peng, Y. et al. Purification and High-Resolution Top-Down Mass Spectrometric Characterization of Human Salivary α-Amylase. Anal. Chem. **84**, 3339–3346 (2012).
91. Schramm, M. & Loyter, A. [90] Purification of α-amylases by precipitation of amylase—glycogen complexes. in Methods in Enzymology vol. 8 533–537 (Academic Press, 1966).
92. Devlin, B. & Roeder, K. Genomic control for association studies. Biometrics **55**, 997–1004 (1999).
93. Meng, E. C. et al. UCSF ChimeraX: Tools for structure building and analysis. Protein Sci. **32**, e4792 (2023).

94. Willems, T., Gymrek, M., Poznik, G. D., Tyler-Smith, C. & Erlich, Y. Population-Scale Sequencing Data Enable Precise Estimates of Y-STR Mutation Rates. Am. J. Hum. Genet. **98**, 919–933 (2016).

95. Frayling, T. M. et al. A Common Variant in the FTO Gene Is Associated with Body Mass Index and Predisposes to Childhood and Adult Obesity. Science **316**, 889–894 (2007).

Methods

Quantification of microbial relative abundances from saliva whole genome sequencing

Saliva-derived whole genome sequencing was previously generated for 12,519 individuals from the Simons Foundation Powering Autism Research (SPARK) cohort of the Simons Foundation Autism Research Initiative (SFARI)[80]. In brief, DNA extracted from saliva was prepared with PCR-free methods for 150bp paired-end sequencing on Illumina NovaSeq 6000 machines. Reads were aligned to human reference build GRCh38 by the New York Genome Center (NYGC) using Centers for Common Disease Genomics project standards.

From the CRAM files previously aligned to GRCh38, we extracted all unmapped reads for subsequent generation and analysis of oral microbiome phenotypes. This retrieved a median of 67.5 million unmapped reads per sample ([35.9m,126.1m], quartiles), comparable with the total number of reads used for prior metagenomic characterization of human microbiomes[81] and consistent with previous analyses of SPARK samples for oral microbiome profiling[29,30]. Unmapped reads were converted to compressed FASTQ with samtools (v1.15.1) and then used as input for microbiome profiling using MetaPhlAn (v4.0.6) with the vOct22 reference database. To evaluate robustness of these oral microbiome profiles, relative abundances of species in SPARK samples were compared with those from samples in the Human Microbiome Project[82] by first subsetting to those profiled in both cohorts and then performing principal coordinate analysis using Bray-Curtis distance (**Extended Data Fig. 1b**).

From the relative abundance phenotype generated for each species by MetaPhlAn, we estimated the fractions of variance explained by covariates (age, sex, autism spectrum disorder, and genetic ancestry principal components; **Fig. 1c**) using ANOVA (finding the sum of squares for each covariate and dividing by the total sum of squares).

Generation of microbial abundance principal components (mPCs)

Relative abundance measures of all microbes were first filtered to 439 entries corresponding to microbial species found in at least 10% of SPARK DNA samples. Rank-based inverse normal transformation across individuals was then performed for each species' abundance. These were then scaled and centered for use as input to principal component analysis to obtain 439 orthogonal microbial abundance principal components (mPCs) representing orthogonal axes of microbial variation.

mPC-based genome-wide association analysis of oral microbiome composition

A straightforward way to search for host genetic effects on microbiomes is to test human genetic variants for association with the abundance of each microbial taxon in turn[11–14]. However, we reasoned that a statistical test designed to aggregate evidence of pleiotropic genetic effects on many species within a microbial community could considerably increase statistical power[35–37]. To perform such a test in a scalable manner (efficient enough to test millions of human genetic

variants), we made use of the decomposition of the microbial abundance matrix into 439 orthogonal mPCs. Specifically, we tested each genetic variant for association with each microbial-abundance PC, after which we summed the 439 test statistics obtained per variant to compute a single, combined association test for each variant (**Fig. 2a**; details in next section). Beyond increasing power to detect pleiotropic effects, the approach reduces multiple-testing burden by testing each genetic variant only once. We evaluated applying this approach to a subset of top axes of microbial variation (rather than all 439 microbial-abundance PCs) but did not observe a further increase in power (data not shown), suggesting that many axes of variation contribute association signal in this data set.

Details of genome-wide association analysis of oral microbiome composition

Variant calling in SPARK was previously performed using DeepVariant (v1.3.0) to produce sample-level VCFs from reads aligned to GRCh38 followed by GLnexus (v1.4.1) to call variants jointly across the cohort. We performed a series of QC steps on the joint call set, starting by converting half-calls to missing and then excluding variants with >10% missingness using plink2 (v2.00a3.6LM). Variants were further excluded if they had a minor allele frequency <1% or if they had a Hardy-Weinberg equilibrium exact test p-value $<1x10^{-6}$ with mid-p adjustment for excessive heterozygosity, leaving 12,525,098 common variants to test for association with oral microbiome phenotypes. Genetic ancestry principal components were generated by LD-pruning variants in 500kb windows with $r^2>0.1$ and then running plink2 --pca approx. No individuals were filtered for outlier heterozygosity after inspection within each genetic ancestry group.

To account for relatedness among SPARK participants, we performed genome-wide association analysis on each mPC phenotype using the linear mixed model implemented in BOLT-LMM (using the --lmmInfOnly flag, as the non-infinitesimal mixed model provided a negligible increase in statistical power) with the following covariates: sequencing batch, age, age squared, square root of age, sex, percent of mapped reads, and the top 10 genetic ancestry principal components. A single father without a recorded age was assigned the average age of other fathers in the dataset. *AMY1* and *PRB1* copy numbers were rescaled to a range of [0,2] and encoded as dosages for association testing.

To test a genetic variant for association with an effect on overall oral microbiome composition, we summed chi-square statistics across the 439 orthogonal mPCs and computed the p-value based on a chi-squared distribution with 439 degrees of freedom. Variants were then filtered to those present in the TOPMed-r3 imputation panel to exclude those in regions of poor mappability, with the MHC region (which was excluded in TOPMed-r3) added back in manually.

Stratified LD score regression for estimating enrichment of heritability at genes with tissue-specific expression

We observed that the same mathematical framework that enables partitioning of heritability via stratified LD score regression on summary statistics from GWAS of a single trait[83] could be extended to analyze test statistics for association with oral microbiome composition (based on

summing chi-squared test statistics across 439 mPCs). Starting from the representation of expected marginal chi-square association statistic for SNP i based on linkage disequilibrium with variants in categories C_k,

$$E\left[\chi_i^2\right] = 1 + Na + N\sum_k \tau_k l(i,k)$$

averaging across the 439 chi-square statistics for each variant gives

$$E\left[\overline{\chi_i^2}\right] = 1 + N\bar{a} + N\sum_k \overline{\tau_k} l(i,k)$$

such that providing $\overline{\chi_i^2}$ as input to S-LDSC generates enrichments corresponding to $\overline{\tau_k}$. Averaged chi-square statistics per variant were used as input to munge_sumstats.py. LDSC was then run with baseline v1.2, weights_hm3_no_hla as weights, and previously described tissue-specific expression bins derived from Genotype-Tissue Expression (GTEx) project samples[39].

Associations of abundances of individual taxa with genetic variants

To avoid test statistic inflation from outlier values, relative abundance measures for each taxon were adjusted for age (using linear regression) and then rank-based inverse normal transformed. Abundances of taxa (of any phylogenetic level) observed in >10% of SPARK samples were then tested for association with host genotypes using BOLT-LMM. The top 20 principal components from PCA on 439 species observed at >10% prevalence (i.e., the top 20 mPCs) were used as covariates to control for the largest axes of variation across samples, along with sequencing batch, age, age squared, square root of age, sex, percent of mapped reads, and the top 10 genetic ancestry principal components. To test loci that might be associated as dominant/recessive rather than additive, BOLT-LMM was rerun using the --domRecHetTest flag.

For estimating effect sizes of specific genotype values such as *FUT2* W154X genotypes (**Fig. 2d**), *AMY1* copy numbers (**Fig. 3c**), or *PRB1* copy numbers (**Extended Data Fig. 2c**), we used linear regression with the same covariates as above, encoding each genotype value (rounded if necessary) as a separate factor. The standard errors estimated by these regressions are slightly underestimated because they do not account for relatedness among SPARK participants, but we determined that this underestimation of standard errors was mild (~7% based on a ~14% inflation of chi-square test statistics computed using linear regression versus a linear mixed model for the 11 genome-wide significant loci (**Extended Data Fig. 10**). To compare effect sizes in adults versus unrelated children, one child was randomly selected from each family in SPARK. BOLT-LMM was used to run linear regression on each of these subsets separately.

Genome-wide association analyses in UK Biobank

Starting from 488,377 individuals in the UK Biobank SNP-array data set[46], individuals were excluded based on the following criteria: 36,008 were removed to drop one relative within pairs of close relatives with kinship coefficient > 0.0884, preferentially keeping individuals if they a) reported having dentures or b) reported not having dentures (i.e., had a non-missing dentures phenotype); 28,701 were removed for not having European genetic ancestry[84]; 1,469 were removed for not having available TOPMed-imputed genotypes (including for chromosome X); 2,601 were removed for not having available WGS data; and 53 were removed for having withdrawn, leaving 419,545 available individuals for genetic association analyses. For the binary oral health phenotypes (dentures use and bleeding gums), 418,039 had non-missing values. For the quantitative BMI z-score phenotype[85], 418,150 had non-missing values.

TOPMed-imputed variants for these individuals were filtered to require minor allele frequency > 0.001 and INFO > 0.3. BOLT-LMM was run in linear regression mode on these samples and variants with the following covariates: age, age squared, sex, genotype array, assessment center, and top 20 genetic ancestry PCs. For estimating effect sizes of specific copy numbers of *AMY1* (**Fig. 3d,e,** and **j**), *AMY2A* (**Extended Data Fig. 5f**), or *AMY2B* (**Extended Data Fig. 5g**), we performed logistic regression (for oral health phenotypes) or linear regression (for BMI) with the same covariates, encoding each copy number (rounded to the nearest integer) as a separate factor.

Phyletic stratification of genetic associations

The phylogenetic tree of all species in the MetaPhlAn 4 database used (vOct22), mpa_vOct22_CHOCOPhlAnSGB_202212.nwk, was first subsetted to the tree spanning nodes with primary label among the 439 species seen at >10% prevalence in the SPARK cohort. This tree was used with graphlan (v1.1.3) for depiction of phylogenetic trees (**Fig. 1b**, **Fig. 2e**, and **Extended Data Fig. 3f**). For comparisons among the effect sizes of a human genetic variant associated with relative abundances of many species, phylogenetic distances between pairs of species were first computed as a cophenetic distance matrix from this tree. For a given index species A, phylogenetic distances between A and other species B were then compared with either a) absolute values of effect sizes for species B (ie. $|\beta_B|$, **Extended Data Fig. 4b,c,g,h**) or b) effect sizes for species B oriented relative to the effect direction for species A (ie. $\text{sign}(\beta_A) \times \beta_B$, **Extended Data Fig. 4d,e,i,j**).

Estimation of *AMY1* copy number in all cohorts

In the SPARK (**Extended Data Fig. 5a**) and *All of Us* cohorts (**Extended Data Fig. 5c**), *AMY1* copy number was estimated by counting whole genome sequencing reads that mapped to the duplicated regions that include *AMY1A* (chr1:103638545-103666411), *AMY1B* (chr1:103685558-103713427), and *AMY1C* (chr1:103732687-103760549) in GRCh38 and normalizing against the total number of reads that aligned in either the 0.5Mb upstream of

AMY2B or the 0.5Mb downstream of *AMY1C*. For the UK Biobank cohort (n=490,415[86], **Fig. 3a**), we applied a more comprehensive read depth normalization pipeline that incorporated sample-specific GC-bias correction inferred from genome-wide alignments (similar to Genome STRiP[87]) before normalizing against read depth in the 0.5Mb regions flanking the amylase locus. We corrected for slight miscalibration of these diploid copy number estimates by fitting a linear model to identify coefficients that centered peaks of copy number estimates at integers.

Among the UK Biobank participants with WGS available, we identified 5,149 siblings that shared both amylase haplotypes IBD2 (based on at most 3 mismatching SNP-array genotypes within a 2Mb window flanking the amylase locus, computed using plink1.9 --genome). Among these IBD2 sibling pairs, 13 pairs were identified as copy number-discordant (and likely to reflect a copy-number mutation in the past generation) based on (i) having *AMY1* copy number estimates that differed by >1.0 and (ii) having *AMY2A* copy number estimates consistent with a duplication or deletion of a commonly variable amylase gene cassette (i.e., +/-1 *AMY2A* copies for *AMY1* copy number discordances of odd parity, and no difference in *AMY2A* copy number for *AMY1* copy number discordances of even parity). We estimated *AMY1* copy number genotyping accuracy by computing the correlation across IBD2 sibling pairs excluding these 13 copy number-discordant pairs.

Testing *AMY1* copy number for association with dental phenotypes and BMI in *All of Us*

We defined the binary tooth loss phenotype as 1 for individuals with at least one recorded instance of “Acquired absence of all teeth” (OMOP concept ID: 40481327, Code: 441935006) and 0 otherwise. Likewise, the caries phenotype was derived from “Dental caries” (OMOP concept ID: 133228, Code: 80967001). For BMI, we generated a normalized z-score phenotype from body mass index (OMOP concept ID: 903124, Code: bmi) by first adjusting for age (in months, determined from time at weight and height measurement) and age squared and then applying inverse rank normal transformation in each sex separately. The BMI z-scores for males and females were then merged together.

Individuals with WGS available for genotyping *AMY1* copy number were first filtered to an unrelated subset of samples (iteratively dropping one individual per related pair with kinship score > 0.1, from relatedness_flagged_samples.tsv). For oral health phenotypes, we performed logistic regression against *AMY1* copy number including age, age squared, sex, and the top 16 genetic ancestry principal components (from ancestry_preds.tsv) as covariates. For BMI, we performed linear regression using only genetic ancestry principal components as covariates as age and sex had already been residualized out. For estimating effect sizes of specific copy numbers of *AMY1* (**Extended Data Fig. 5d,e** and **Extended Data Fig. 7c-f**), we performed logistic regression (for oral health phenotypes) or linear regressions (for BMI) with the same covariates, encoding each copy number (rounded to the nearest integer) as a separate factor.

Paralogous sequence variation in *AMY1*

To identify and genotype paralogous sequence variants (PSVs) from UK Biobank whole genome sequencing data, we used a read-counting approach similar to our previous work[88]. In brief, reads from each sample that had been aligned to any of the three 27.6kb regions in GRCh38 corresponding to *AMY1A* (chr1:103638695-103666261), *AMY1B* (chr1:103685708-103713277), and *AMY1C* (chr1:103732837-103760399) were realigned with bwa (v0.7.17) to the reference sequence of *AMY1A* after filtering out reads with any of the last four SAM flags (-F 0xF00). Read counts supporting each base at each position were tabulated with htsbox (r345) pileup, filtering alignments <50bp and base calls with quality score <20. Individuals were called heterozygous for a PSV allele (having at least one copy of *AMY1* with each of two alleles) if at least 5 reads supported the variant allele in that sample and at least 5 reads did not. PSVs were then filtered to those with heterozygosity >0.002 (resulting in 892 PSVs passing filters). To estimate diploid copy number genotypes for a PSV, we multiplied each individual's diploid *AMY1* copy number by the allelic fraction of the PSV in that individual. In association tests using linear regression with BOLT-LMM, we rounded copy number estimates to integer genotypes.

For follow-up analyses of effect sizes of specific *AMY1* F141C and C477R copy number genotypes, we optimized the assignments of integer copy number genotypes based on manual inspection of histograms of allelic depth-derived PSV copy number estimates (**Extended Data Fig. 11**). Specifically, we assigned copy numbers for each of F141C and C477R using the thresholds [0,0.25)=CN0, [0.25,1.75)=CN1, [1.75,2.7)=CN2, [2.7,3.5)=CN3, and [3.5,5)=CN4. In UK Biobank, F141C had 416,381 (99.2%), 3,124 (0.74%). and 40 (0.0095%) individuals with 0, 1, and 2 copies respectively, and C477R had 412,450 (98.3%), 6527 (1.6%), 547 (0.13%), 19 (0.0045%), and 2 (0.0005%) individuals with 0, 1, 2, 3, and 4 copies respectively. In SPARK, F141C had 12,459 (99.5%) and 60 (0.48%) individuals with 0 and 1 copies respectively, and C477R had 12,343 (98.6%), 172 (1.4%), and 4 (0.032%) individuals with 0, 1, and 2 copies respectively. These threshold-based copy numbers of the alternate alleles were used in logistic regression along with copy numbers of the reference alleles (i.e., F141 and C477) and covariates as above.

Protein expression of AMY1

pCAGEN[89] was a gift from Connie Cepko (Addgene plasmid no. 11160; http://n2t.net/addgene:11160; RRID:Addgene_11160). Codon-optimized sequences encoding reference, F141C, and C477R *AMY1* alleles were synthesized and ordered as gBlocks from Integrated DNA Technologies for cloning into pCAGEN downstream of the CAG promoter. Clones were screened for sequence errors before plasmid preparation using Plasmid Plus Midi Kit (Qiagen, catalog no. 12943). pUC19 (New England Biolabs, catalog no. N3041S) was used as a negative control. 15ug of each plasmid was lipofected into separate 10cm plates of HEK293T (Takara, catalog no. 632180) cells at ~70% confluence using 30uL of Lipofectamine 3000 (Invitrogen, catalog no. L3000015). Medium was switched to serum-free after 24 hours before collection of both supernatant and lysate at 72 hours post-lipofection. 10mL of supernatant was spun at 1000g for 10 minutes to remove cells and debris before the addition of 100uL Halt™ Protease Inhibitor Cocktail (100x, Thermo Scientific, catalog no. 78438) and

100uL EDTA (0.5M). Cells were washed with ice-cold PBS before the addition of cold 1mL RIPA Lysis and Extraction Buffer (Thermo Scientific, catalog no. 89900) with 10uL Halt™ Protease Inhibitor Cocktail (100x) and 10uL EDTA (0.5M). After sufficient solubilization of the cells had occurred, 1uL of Benzonase® Nuclease (250U/uL, Milipore, catalog no. E1014) and 10uL of MgCl2 (1M) were added before incubation at 37C, 500 RPM for 30 minutes. Lysate was then spun at 10,000g for 10 minutes to remove insoluble precipitate. Both supernatant and lysate were stored at -80C until further use.

Western blot of AMY1 in cell culture supernatant and lysate

7.5uL of supernatant or purified lysate was first run denatured and reduced in a 10% Mini-PROTEAN® TGX™ Precast Protein Gel (Bio-Rad, catalog no. 4561036) before wet transfer to nitrocellulose membrane (120V, 2 hours). After evaluation of equal loading by Ponceau S Staining Solution (Thermo Scientific, catalog no. A40000279), the membrane was blocked for 1 hour at room temperature with TBS, 0.1% Tween-20, and 5% w/v nonfat dry milk before washing 3 times for 5 minutes each with TBS-T (TBS, 0.1% Tween-20). Amylase Antibody (G-10, Santa Cruz Biotechnology, catalog no. sc-46657, lot no. G0324) was used as the primary antibody at a 1:200 dilution in TBS-T with 5% w/v milk for incubation overnight at 4C with rotation. Membrane was washed 3 times for 5 minutes each with TBS-T before addition of Anti-mouse IgG, HRP-linked Antibody (Cell Signaling Technology, catalog no. 7076, lot no. 39) as secondary at a 1:2000 dilution in TBS-T with 5% w/v milk for 1 hour incubation at room temperature with rotation. Membrane was washed 3 times for 5 minutes each with TBS-T before detection with Amersham ECL Prime Western Blotting Detection Reagent (Cytiva, catalog no. RPN2236) (**Extended Data Fig. 6a**).

Purification of AMY1 reference and F141C protein

Purification of amylase from supernatant was done using glycogen, adapted from refs[90,91]. All following steps were conducted on ice or at 4C (rotation, centrifugation). In brief, 10mL of supernatant was initially concentrated using prewet 15mL Amicon® Ultra Centrifugal Filter, 30 kDa MWCO (Milipore, catalog no. UFC9030) and put up to a total volume of 900uL with PBS. 600uL of cold ethanol was added slowly to make 40% ethanol v/v. This was then centrifuged at 10,000g for 10 minutes to remove any insoluble precipitate. To the supernatant, 75uL of 0.2M sodium phosphate buffer (pH 8), 75uL glycogen (20mg/mL, Roche, catalog no. 10901393001), and 100uL of ethanol were added in that order. This was then incubated for 5 minutes with end-over-end mixing before centrifugation at 5000g for 6 minutes. The glycogen pellets were then washed twice with 10mM sodium phosphate buffer (pH 8) containing ethanol (40% v/v) before resuspension in 100uL of 50mM MOPS buffer (pH 7, Thermo Scientific, catalog no. J61821-AK) with 5mM CaCl2. These were then incubated at 37C, 500 RPM for 30 minutes to allow for glycogen digestion. The above purification procedure was repeated as necessary to arrive at purified amylase as evaluated on 10% Mini-PROTEAN® TGX™ Precast Protein Gel with InstantBlue® Coomassie Protein Stain (Abcam, catalog no. ab119211) (**Extended Data Fig. 6b**).

Amylase activity assay

Protein was normalized between reference and F141C isoforms by densitometric measurement against linear dilution series (r^2=0.99 with input). The concentration chosen for the assay was determined by having maximal linear activity over the observation window. For each replicate, 10uL of reference or F141C amylase enzyme diluted in 50mM MOPS buffer (pH 7), 5mM CaCl2, and 0.02% w/v BSA (New England Biolabs, catalog no. B9000) was quickly mixed with 10uL of BODIPY™ FL conjugated-starch substrate from EnzChek™ Ultra Amylase Assay Kit (Invitrogen, catalog no. E33651) in 50mM MOPS buffer (pH 7) with 5mM CaCl2. The reaction was then maintained at 20C for 2 hours in a Bio-Rad CFX384 Real-Time PCR Detection System with fluorescence reading taken every minute using FAM fluorophore settings. Fluorescence at 30 minutes relative to initial reading was used as input to a linear model with allele and plate to regress out any run-to-run effects during comparison across replicates (**Extended Data Fig. 6c**).

Bacterial genome reference panel for analyzing bacterial gene dosages

To reduce hypothesis testing burden in association analyses of human genetic variants with bacterial gene dosage phenotypes, we restricted analyses to a set of 30 bacterial genomes representing highly-abundant species and species whose abundances we had found to associate with human genetic variants. Specifically, we selected 30 bacterial species with genomes available in GenBank by including: 1) the five most abundant species in SPARK oral microbiomes, 2) species that associated strongly ($p<4x10^{-11}$) with at least one human genetic variant, and 3) the top two associated species for each locus if not already included. We substituted *Stomatobaculum longum* for the related *Stomatobaculum* SGB5266 which would have been included under 2) but lacked a GenBank assembly. We selected a single genome for each of the 30 species using the following criteria. The SGB centroid was prioritized over other genomes if it corresponded to a GenBank assembly. For cases where the centroid was not a GenBank assembly and multiple genomes corresponded to GenBank assemblies, the reference genome was prioritized if available, or otherwise the highest ranked genome among those listed. A list of these species and GenBank assemblies is included in **Supplementary Table 5**. A bowtie2 index was then built from these merged genomes.

Measuring bacterial gene dosages using read-depth phenotypes

We computed WGS coverage-derived phenotypes informative of structural variation across each of the 30 bacterial reference genomes by first realigning unmapped reads from SPARK saliva WGS to the 30 reference genomes using bowtie2 (v2.5.1) with the --very-sensitive flag. These alignments were then position-sorted within contigs with samtools (v1.15.1). For each WGS sample, we quantified read depth in 500bp bins tiling each of the 30 bacterial reference genomes using mosdepth (v0.3.6), excluding reads with mapping quality <5. For each sample, for each of the 30 bacterial reference genomes, we then median-normalized the bin-level read-depth measurements across the 500bp bins of that reference genome to control for species abundance (such that normalized read-depth measurements had a median of 1 among bins

corresponding to each species). If a sample had <0.5x median coverage across bins corresponding to a given species, we set that sample's normalized read-depth measurements for that species to missing to focus downstream analyses on samples with less-noisy measurements. Finally, we truncated median-normalized read-depth values to the interval [0,1], both to focus on deletions within bacterial genomes and to reduce the influence of outlier measurements that might reflect mismapped reads (potentially derived from either duplicated genomic regions or homologous sequences in microbial species not represented among the 30 reference genomes). We reasoned that these bin-level measurements would capture kilobase-scale deletions of bacterial genomes, circumventing the need to predefine a set of structural variant regions (which was difficult due to the limited sequencing coverage of most species).

Testing bacterial gene dosage phenotypes for association with host genotypes

We used linear regression to test each of the 11 human genetic variants we had found to associate with oral microbiome composition (**Table 1**) for association with normalized read-depth (truncated to [0,1]) in each 500bp bin of each of the 30 bacterial reference genomes. We took two precautions to avoid potential confounders. First, for each of the 30 species, we included as covariates the top 20 principal components of the normalized, truncated read-depth matrix for that species (running PCA after centering and scaling each bin to have a mean of 0 and standard deviation of 1 across samples) to control for linked structural variations (e.g., differences across strains) that could potentially generate non-causal associations in a manner analogous to population structure in GWAS (see **Supplementary Note 4**). Second, we applied a form of genomic control[92] (applied across the 500bp bins of each reference genome) to adjust for remaining test statistic inflation. Specifically, for each pairing of a species and a human genetic variant, we computed the adjustment factor

$$\frac{median(\chi_1^2)}{F^{-1}(0.5)}$$

across the χ_1^2 test statistics for the 500bp bins of that reference genome, where $F^{-1}(x)$ is the inverse cumulative distribution function for a χ_1^2 random variable. The χ_1^2 test statistics were then divided by this factor. This yielded 208 read-depth bins in 18 species that significantly associated with at least one of the 11 human genetic variants (FDR<0.01, **Supplementary Table 6** and **7**) and resolved to 68 unique microbial regions after merging bins within 1.5kb of another significantly associated bin (**Fig. 4b**).

For combinatorial analysis of deletions of the five *vad* genes in *V.* sp. 3627 that associated in the same direction with *FUT2* genotype (*vadB* through *vadF*) (**Fig. 6e**), we first selected individuals whose oral microbiomes had evidence of near-complete presence (>0.8 median normalized coverage) or near-complete absence (<0.2 median normalized coverage) of each *vad* gene (n=3081, representing roughly half of the SPARK samples with coverage of the *V.* sp. 3627 genome reaching the >0.5x threshold for analysis). For each common combination (>2.5% of selected individuals) of presence/absence status of *vadB* through *vadF*, the number of individuals with each *FUT2* W154X genotype were then counted.

AlphaFold2 prediction of protein structures

To predict protein structures for *Streptococcus parasanguinis* AbpA or AbpB (bound to human AMY1), the *Veillonella* sp. 3627 VadD trimer, and *Streptococcus mitis* CrpE, we used AlphaFold (v2.3.1) for multimer prediction with default reference databases, max template date of 2022-01-01, and without Amber relaxation. The TonB domain of AbpB and the signal peptide of VadD were excluded for visualization. To minimize the model size, CrpE was truncated to the region spanning the CshA NR2 domain to the sixth mucin-binding domain (residues 572-2701). Structures were visualized with ChimeraX[93].

Genetically derived blood typing in SPARK

We assigned blood types to SPARK participants based on WGS-derived SNP and indel genotypes using a procedure similar to previous work[13]. The genotype of the rs8176746 missense SNP was first used to determine an individual's dosage (i.e., allele count) of type B alleles (= T allele count) and non-type B dosage (= G allele count).

The rs8176719 indel was next used to determine type O1 dosage (= deletion allele count), which was subtracted from the non-type B dosage to yield non-type B/O1 dosage, as the rs8176719 deletion allele typically occurs on haplotypes that would otherwise be type A alleles. Although this is true in European, East Asian, and American ancestry haplotypes in 1KGP populations, in a small fraction of African (3.2%) and South Asian (0.2%) ancestry haplotypes, the rs8176719 deletion occurs in *cis* with the type B missense allele. As we found 36 SPARK participants who had O1 dosage exceeding non-type B dosage, we subtracted this excess from their type B dosages.

The rs41302905 missense SNP was next used to determine type O2 dosage (= T allele count) and subtracted from the non-type B/O1 dosage to yield type A dosage, as it appears to be in *cis* with type A alleles in all 1KGP populations. O1 and O2 dosages were then merged to compute type O dosage.

The rs56392308 indel was next used to determine the type A2 dosage (= deletion allele count) and subtracted from the type A dosage to yield type A1 dosage. For seven individuals where the type A2 dosage exceeded type A dosage, five appeared to be on type O alleles and two on either type O or type B alleles, so this excess was subtracted from their type A2 dosage.

Enrichment of conserved domains in bacterial genes associated with secretor status

For each of the bacterial species with adhesin genes in dosage variable regions that associated with *FUT2* loss of function (**Fig. 6a-c,f,g**), protein IDs (i.e., WP numbers) for the species were extracted from its RefSeq general feature format (GFF) file. Conserved domains (from NCBI's conserved domain database, CDD) were identified for each protein using a modified version of the provided bwrpsb.pl script (applied to up to 250 proteins at a time). A one-sided Fisher's

exact test was used to identify domains enriched among proteins encoded by genes within read-depth bins that associated with *FUT2* genotype.

Extended Data Figure Legends

Extended Data Figure 1. Oral microbiome composition changes with host age. a, Distribution of the percent of reads not mapping to the human reference (GRCh38) across SPARK participants (n=12,519). The median of 8.39% is indicated by the dotted red line. **b**, Principal coordinate analysis of microbiome profiles from the Human Microbiome Project[82] together with a randomly selected subset of SPARK samples (n=250). PCoA was performed using Bray-Curtis distance on all microbial species profiled in both data sets. The SPARK saliva samples cluster with other samples from oral communities. Points in gray are samples from Human Microbiome Project sites not listed in the legend. **c**, SPARK samples on a UMAP (Uniform Manifold Approximation and Projection) generated from the first 20 principal components of the abundance matrix for the 439 most prevalent species fall on a gradient with respect to host age (color bar on right). **d**, The 439 most prevalent species on a UMAP generated from their loadings onto the first 20 principal components of the abundance matrix fall on a gradient with respect to correlation of relative species abundance with host age (color bar on right). **e**, Among children in SPARK (n=5,760), a UMAP using the same 20 principal components as in **b** shows minimal stratification by autism spectrum disorder case status.

Extended Data Figure 2. **Complex associations of human genetic variation at the *PRB* locus with oral microbiome phenotypes. a**, Diploid copy number of *PRB1* estimated for SPARK participants from WGS read-depth (n=12,519). **b**, Associations of genetic variants at the *PRB* locus with relative abundance of *Stomatobaculum* SGB5266. The association of *PRB1* copy number is highlighted (red point). **c**, Allelic series of effect sizes of *PRB1* diploid copy numbers on relative abundance of *Stomatobaculum* SGB5266 (in s.d. units). Error bars, 95% CIs. **d**, Partial colocalization of associations of genetic variants at the *PRB* locus with *Stomatobaculum* SGB5266 relative abundance (-log10(p), x-axis) and oral microbiome composition (-log10(p), y-axis). *PRB1* copy number (red point) and a common *PRB4* loss-of-function variant (blue point) appear to deviate from a generally concordant pattern of associations, likely reflecting multiple causal effects. **e**, Effect sizes on *Stomatobaculum* SGB5266 relative abundance for each additional copy of *PRB1* compared to each functional copy of *PRB4*.

Extended Data Figure 3. Additional information about associations between human genetic variants and oral microbial abundance phenotypes. a, Stratified LD-score regression coefficients (which quantify evidence of enrichment of heritability in regions surrounding genes with tissue-specific expression) from S-LDSC analysis of chi-squared statistics from the oral microbiome composition GWAS. Heritability enrichment was evaluated for 53 tissues from the Genotype-Tissue Expression Project (GTEx). Error bars, 95% CIs. **b**, Effect sizes (in s.d. units) on relative abundance of SGB9384 for individuals homozygous for either allele of the *TLR1* I602S missense variant relative to heterozygotes. Error bars, 95% CIs. **c**, Colocalization of genetic associations at the *PITX1* locus with dentures use (-log10(p), y-axis) and oral microbiome composition (-log10(p), x-axis). **d**, Analogous to c, for the *FUT2* locus. **e**, Distribution of relative effect sizes on microbial species abundances for individuals homozygous for the *FUT2* secretor allele relative to heterozygotes, across microbial species whose relative

abundance associated with secretor status (FDR<0.05).

Extended Data Figure 4. Phyletic stratification of effects of two human genetic variants on relative abundances of oral microbial species. a, Positive correlation of relative abundance of the Actinobacteria phylum relative to the Firmicute phylum across saliva samples from SPARK participants (n=12,519). **b**, Unsigned effect sizes for associations of FUT2 secretor status with relative abundances of the 439 most prevalent microbial species (y-axis) versus phylogenetic distance from *Haemophilus sputorum* (x-axis). **c**, Analogous to **b**, for *Granulicatella* SGB8239. **d**, Signed effect size of associations of FUT2 secretor status with relative abundances of microbial species (oriented relative to the effect direction for *H. sputorum*) (y-axis) versus phylogenetic distance from *H. sputorum* (x-axis). **e**, Analogous to **d**, for *G.* SGB8239. **f**, Microbial taxa whose abundance associated with the index variant at the *HECTD2*/*PPP1R3C* locus (FDR<0.1) shown on the phylogenetic tree of 439 species (red, taxa whose relative abundances increased with the reference allele; blue, taxa whose relative abundances decreased with the reference allele). Two significantly-associated phyla (Firmicutes and Actinobacteria) are highlighted with yellow sectors. At the species level (outermost circle), dot sizes increase with statistical significance. **g**, Analogous to **b**, for *Granulicatella adiacens* with the *HECTD2*/*PPP1R3C* index variant. **h**, Analogous to **g**, for *Streptococcus vestibularis*. **i**, analogous to **d**, for *G. adiacens* with the *HECTD2*/*PPP1R3C* index variant. **j**, analogous to **i**, for *S. vestibularis*. For **b**-**e**, **g**-**j**, the red line indicates the median effect size, and the shaded region indicates the interquartile range.

Extended Data Figure 5. **Genotyping and mutability of *AMY1* copy number and replication in *All of Us* of associations with oral health. a**, Estimated diploid copy number of *AMY1* for SPARK participants (n=12,519). **b**, Concordance of *AMY1* copy number estimates between 5,149 sibling pairs in the UK Biobank cohort that share both haplotypes identical-by-descent (IBD2) in the region surrounding the amylase locus. Among the 5,149 IBD2 sibling pairs, 13 had copy number-discordant calls (red points) that tended to differ by two copies (11/13), likely corresponding to de novo duplication or deletion of a copy of the common structural cassette containing two *AMY1* genes in an inverted orientation to each other (**Fig. 3a**). Several IBD2 sibling pairs with *AMY1* copy number estimates that differed by close to 1 copy appeared more likely to reflect uncertainty in copy number estimates as they lacked a corresponding *AMY2A* duplication or deletion that would be expected to accompany a duplication or deletion of a single copy of *AMY1*. This gives an estimated germline mutation rate of 6.3×10^{-4} mutations/meiosis ([3.5×10^{-4},11.1×10^{-4}], 95% CIs, similar to recent estimates from haplotype coalescent trees[49]), exceeding the mutation rate of most short tandem repeats[94]. **c**, Analogous to **a**, for the *All of Us* cohort (n=245,377). **d**, Odds ratios for risk of complete tooth loss in *All of Us* (n=230,002) per *AMY1* diploid copy number; error bars, 95% CIs. **e**. Analogous to d, for having caries. **f**, Odds ratios for risk of dentures use in UK Biobank (n=418,039) per *AMY2A* diploid copy number; error bars, 95% CIs. **g**, Analogous to **f**, for *AMY2B* diploid copy number.

Extended Data Figure 6. **AMY1 F141C has enzymatic activity similar to the reference isoform and AMY1 C477R is retained intracellularly. a**, Immunoblotting for amylase transgenically expressed in HEK293T cells. Amylase (~56kD) is found in the supernatant of

cells transfected with the reference sequence of *AMY1* (lane 1) and *AMY1* F141C (lane 2), but not *AMY1* C477R (lane 3) or a controlplasmid (lane 4). The glycogen-purified protein is recognized by anti-amylase (lane 5). Amylase is found in the lysate of cells transfected with the reference sequence of *AMY1* (lane 6), *AMY1* F141C (lane 7), and less abundantly in cells transfected with *AMY1* C477R (lane 8), but not a control plasmid (lane 9). Cross-reactive protein (~50kD) can be seen in the lysate of all samples and smaller AMY1 fragments in lanes containing AMY1. **b**, Glycogen-purified supernatant from cells transfected with the reference sequence of *AMY1* (lane 1) and *AMY1* F141C (lane 2), but not *AMY1* C477R (lane 3) or a control plasmid (lane 4) contains a single protein band at ~56kD. **c**, Starch degradation is similar between equivalent mass dilutions of the reference amylase isoform and AMY1 F141C (n=32 replicates for each allele). Starch degradation was measured as change in FAM relative fluorescence units over 30 minutes (y-axis) after addition of diluted AMY1 to quenched starch substrate.

Extended Data Figure 7. ***AMY1* copy number associates with oral microbiome abundances consistently across age groups and does not associate with BMI in *All of Us*. a**, Comparison of effect sizes for *AMY1* copy number versus *AMY1* C477R copy number on relative abundances of 16 bacterial species (from **Fig. 3b**) and on risk of dentures use (blue dot). For some species (e.g., *Stomatobaculum* SGB5266), the relative effect size of *AMY1* copy number versus *AMY1* C477R copy number on abundance is similar to this ratio for dentures use (blue line), whereas for others, it is not (e.g. *Prevotella pallens*). **b**, Concordance of effect size estimates for *AMY1* copy number on relative abundances of 439 microbial species in adults (x-axis) and unrelated children (y-axis). Species whose abundances associated significantly with *AMY1* copy number are indicated in darker gray (FDR < 0.05) or black (Bonferroni p < 0.05). **c**, Effect sizes on BMI per *AMY1* copy number in *All of Us* (n=219,879). The line drawn is the best fit across *AMY1* copy numbers. As a positive control, we confirmed that the BMI phenotype we tested (see **Methods**) associated strongly with the BMI-associated SNP rs1421085 at *FTO*[95] (p=3.72x10^{-140}). **d**, Analogous to **c**, for the African/African American ancestry subset of *All of Us* participants (n=49,296, p=0.25). **e**, Analogous to **c**, for the American Admixed/Latino ancestry subset of *All of Us* participants (n=38,788, p=0.01). **f**, Analogous to **c**, for the European ancestry subset of *All of Us* participants (n=122,577, p=0.66). Error bars, 95% CIs in all panels.

Extended Data Figure 8. **Pathway abundance associations and details of gene dosage associations. a**, Effect sizes of *AMY1* diploid copy numbers on the relative abundance of reads mapping to genes in the tricarboxylic acid cycle (TCA) pathway across microbial species. **b**, Associations (-log10(p), x-axis) of the relative abundance of reads mapping to each of 2,416 *Haemophilus sputorum* gene families with FUT2 secretor status. Read depth measurements were inverse normal transformed across samples for each gene. Gene families generally fell into two modes, one group not associated with FUT2 secretor status and another that associated at significance similar to that of the relative abundance of *H. sputorum*. One outlier (indicated with the red arrow) associated much more strongly and corresponds to a gene encoding a protein with trimeric autotransporter adhesin domain annotations (UniParc ID J5HXA9). **c**, Effect sizes of *AMY1* diploid copy numbers on relative abundance of *Streptococcus parasanguinis*. **d**, Effect sizes of *FUT2* W154X genotype on normalized coverage at bacterial

genes encoding proteins with YadA-like (adhesin) domain annotations. Colors indicate the effect direction of *FUT2* genotype on the relative abundance of each species (red, increasing with functional copies of *FUT2*; blue, decreasing with functional copies of *FUT2*). **e**, Correlation matrix of normalized coverage in the region of the *V*. sp. 3627 genome surrounding *vadD* and *vadE* revealed a linked deletion to *vadE* that contained *vadC* and nearly passed FDR<0.01 (p=3.32x10^{-6}) in association with *FUT2* W154X genotype, causing samples with *vadE* to tend to also contain *vadC* (**Fig. 6d**) and possibly suggestive of the event introducing *vadD* occurring after the one that produced *vadC* and *vadE*. Arrows indicate *vad* gene locations. Error bars, 95% CIs in all panels.

Extended Data Figure 9. **Left breakpoints of deletions in the genome of *Haemophilus sputorum* that contain the *hadA* gene are bimodal. a**, Coverage of the QEQH01000003.1 contig in the *Haemophilus sputorum* genome for individuals with distinct left breakpoints of deletions containing *hadA*. **b**. Distribution of left breakpoints for the 200 individuals with highest genomic coverage of *H. sputorum* who had strong evidence of the deletion containing *hadA* (<0.1 median normalized coverage at the strongest associated bin). For each individual, mosdepth was run to measure per-base depth, and left and right breakpoints were identified as the first or last base with 10 consecutive zeros, respectively.

Extended Data Figure 10. **Modest inflation of chi-squared test statistics from linear regression in SPARK.** To assess the amount of test statistic inflation incurred by linear regression analyses (which do not account for relatedness among SPARK participants), we computed ratios of chi-squared statistics from linear mixed models (LMM, which account for sample relatedness) to chi-squared statistics from standard linear regression (LR) for the 11 lead variants that associated with oral microbiome composition (in our combined test of 439 mPCs).

Extended Data Figure 11. **Distributions of *AMY1* F141C and *AMY1* C477R copy number estimates. a**, Histograms of copy number estimates for *AMY1* F141C in SPARK (n=12,519). Thresholds used to call integer copy numbers are indicated with dotted red lines. **b**, Analogous to **a**, but for *AMY1* C477R. **c**, Analogous to **a**, but for UK Biobank (n=490,415). **d**, Analogous to **b**, but for UK Biobank.

Data availability

The following data resources are available by application: UKB (http://www.ukbiobank.ac.uk/), All of Us Research Program (https://allofus.nih.gov/), and SFARI SPARK (https://www.sfari.org/resource/spark/). The following data resource is publicly available: Human Microbiome Project (https://hmpdacc.org/). Relative abundances of species in the oral microbiome will be returned to SFARI for release upon request.

Code availability

The following publicly available software resources were used: MetaPhlAn (v.4.0.6, http://segatalab.cibio.unitn.it/tools/metaphlan/index.html), HUMAnN (v.3.8, https://huttenhower.sph.harvard.edu/humann), GraPhlAn (v.1.1.3, http://segatalab.cibio.unitn.it/tools/graphlan/index.html), mosdepth (v.0.3.6, https://github.com/brentp/mosdepth), bcftools (v.1.14, http://www.htslib.org/), samtools (v.1.15.1, http://www.htslib.org/), plink (v.1.90b6.26 and v.2.00a3.7,https://www.cog-genomics.org/plink/), BOLT-LMM (v.2.4.1, https://alkesgroup.broadinstitute.org/BOLT-LMM/), qqman (v.0.1.8, https://cran.r-project.org/web/packages/qqman/index.html), bedtools (v.2.27.1, https://bed-tools.readthedocs.io/en/latest/), AlphaFold2 (v.2.3.1, https://github.com/google-deepmind/alphafold), and ChimeraX (v.1.8, https://www.cgl.ucsf.edu/chimerax/). Custom code used to generate results in this study is available via Zenodo at 10.5281/zenodo.14559458.

Acknowledgements

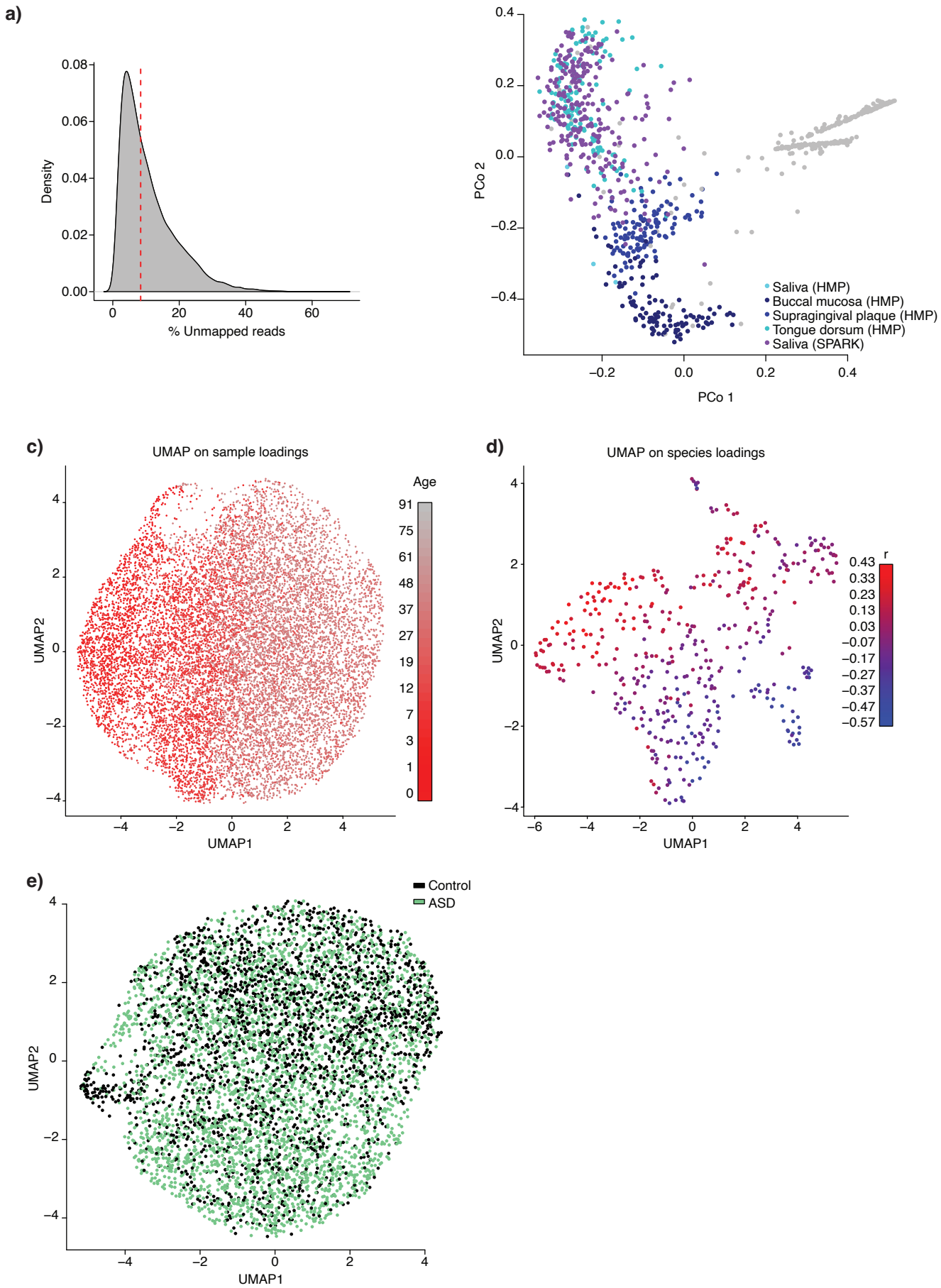
We thank M. Florio and W.-S. Lee for helpful discussions. This research was conducted using the UKB Resource under application number 40709. N.K. was supported by US NIH Fellowship F31 DE034283. R.E.H. and S.A.M. were supported by US NIH grant R01 HG006855. M.L.A.H. was supported by US NIH Fellowship F32 HL160061. R.E.M. was supported by US NIH grant K25 HL150334. S.A.M. was supported by the Howard Hughes Medical Institute. P.-R.L. was supported by US NIH grants R56 HG012698 and R01 HG013110 and a Burroughs Wellcome Fund Career Award at the Scientific Interfaces. The funders had no role in study design, data collection and analysis, the decision to publish or the preparation of the manuscript. The content is solely the responsibility of the authors and does not necessarily represent the official views of the NIH. Computational analyses were performed on the O2 High Performance Compute Cluster supported by the Research Computing Group at Harvard Medical School (http://rc.hms.harvard.edu), the UKB Research Analysis Platform, and the *All of Us* Researcher Workbench. Molecular graphics and analyses performed with UCSF ChimeraX, developed by the Resource for Biocomputing, Visualization, and Informatics at the University of California, San Francisco, with support from National Institutes of Health R01-GM129325 and the Office of Cyber Infrastructure and Computational Biology, National Institute of Allergy and Infectious Diseases. The *All of Us* Research Program is supported by the NIH, Office of the Director: Regional Medical Centers: 1 OT2 OD026549; 1 OT2 OD026554; 1 OT2 OD026557; 1 OT2 OD026556; 1 OT2 OD026550; 1 OT2 OD026552; 1 OT2 OD026553; 1 OT2 OD026548; 1 OT2

OD026551; 1 OT2 OD026555; IAA no. AOD 16037; Federally Qualified Health Centers: HHSN 263201600085U; Data and Research Center:5 U2C OD023196; Biobank: 1 U24 OD023121; The Participant Center: U24 OD023176; Participant Technology Systems Center: 1 U24 OD023163; Communications and Engagement: 3 OT2 OD023205; 3 OT2 OD023206; and Community Partners: 1 OT2 OD025277; 3 OT2 OD025315; 1 OT2 OD025337; and 1 OT2 OD025276. In addition, the *All of Us* Research Program would not be possible without the partnership of its participants.

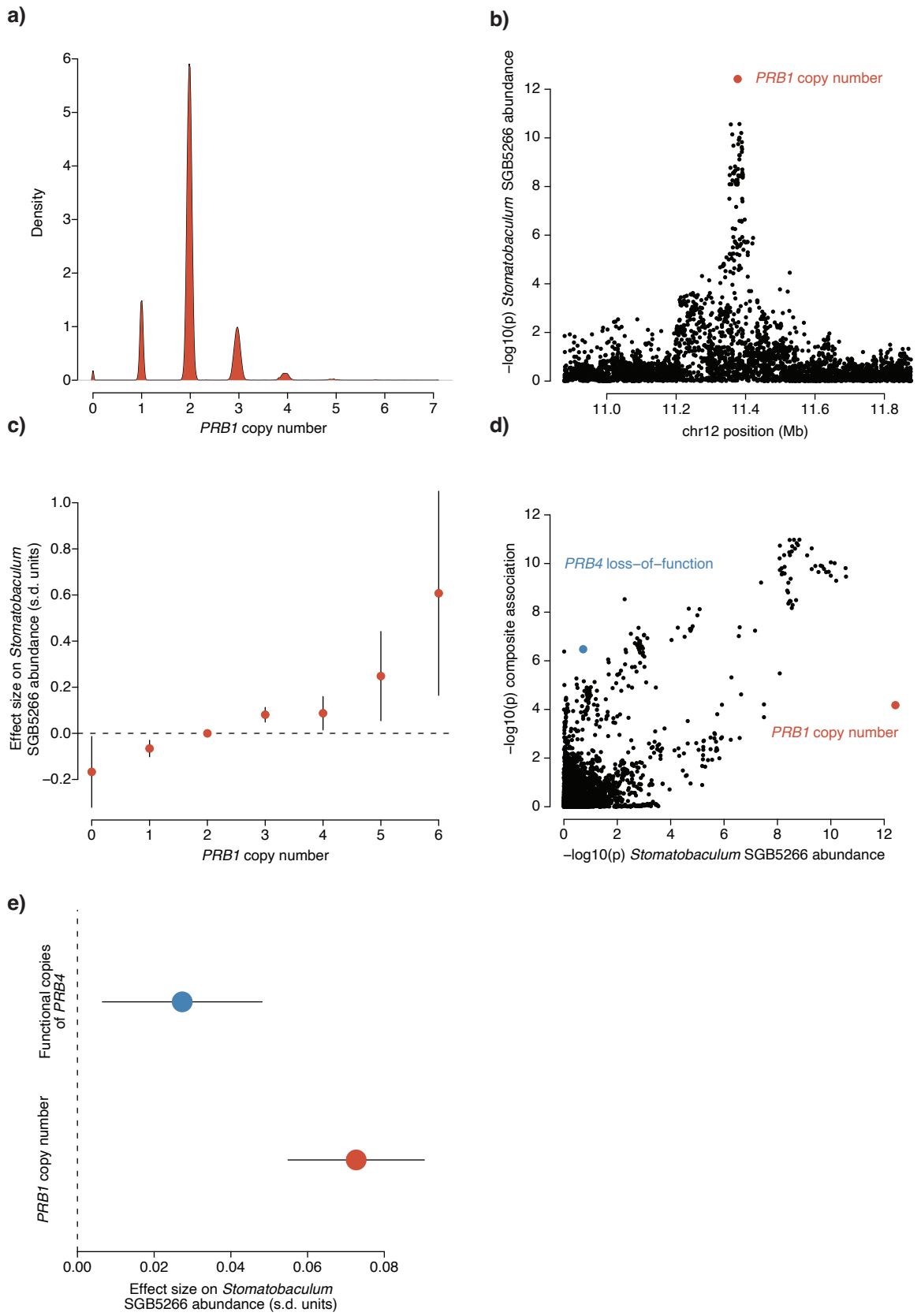
Author contributions

N.K., S.A.M. and P.-R.L. conceived the study design. N.K. and P.-R.L. did the computational and statistical analyses. R.E.H., R.E.M., and P.-R.L. developed the read depth normalization pipeline on UK Biobank. M.L.A.H. helped with performing the genome-wide association for dentures use on UK Biobank. R.E.H. and C.L.U. helped with characterizing variation at the amylase locus. N.K. did the *in vitro* experiments. N.K., S.A.M. and P.-R.L. wrote the manuscript with contributions from all authors.

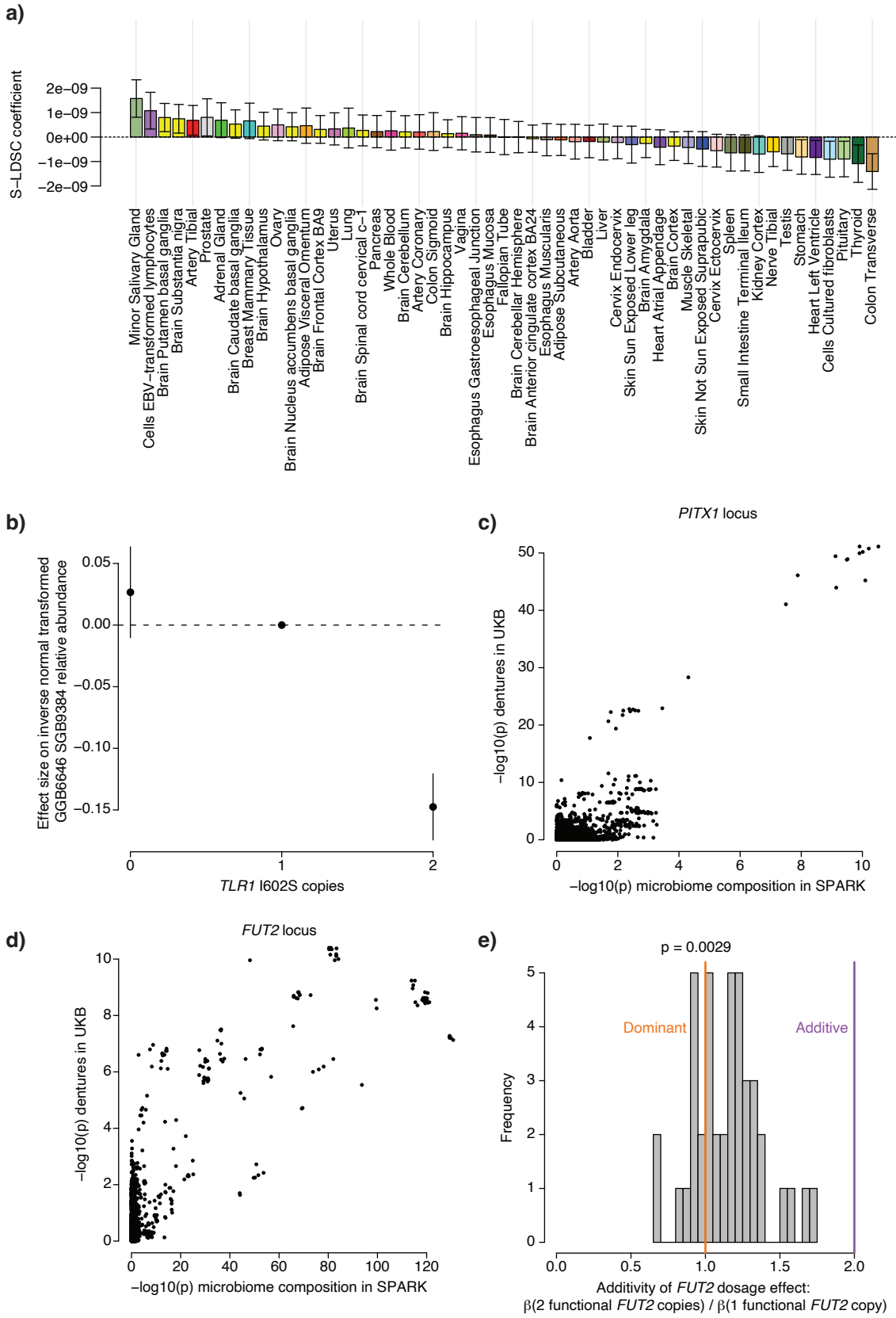
Extended Data Figure 1



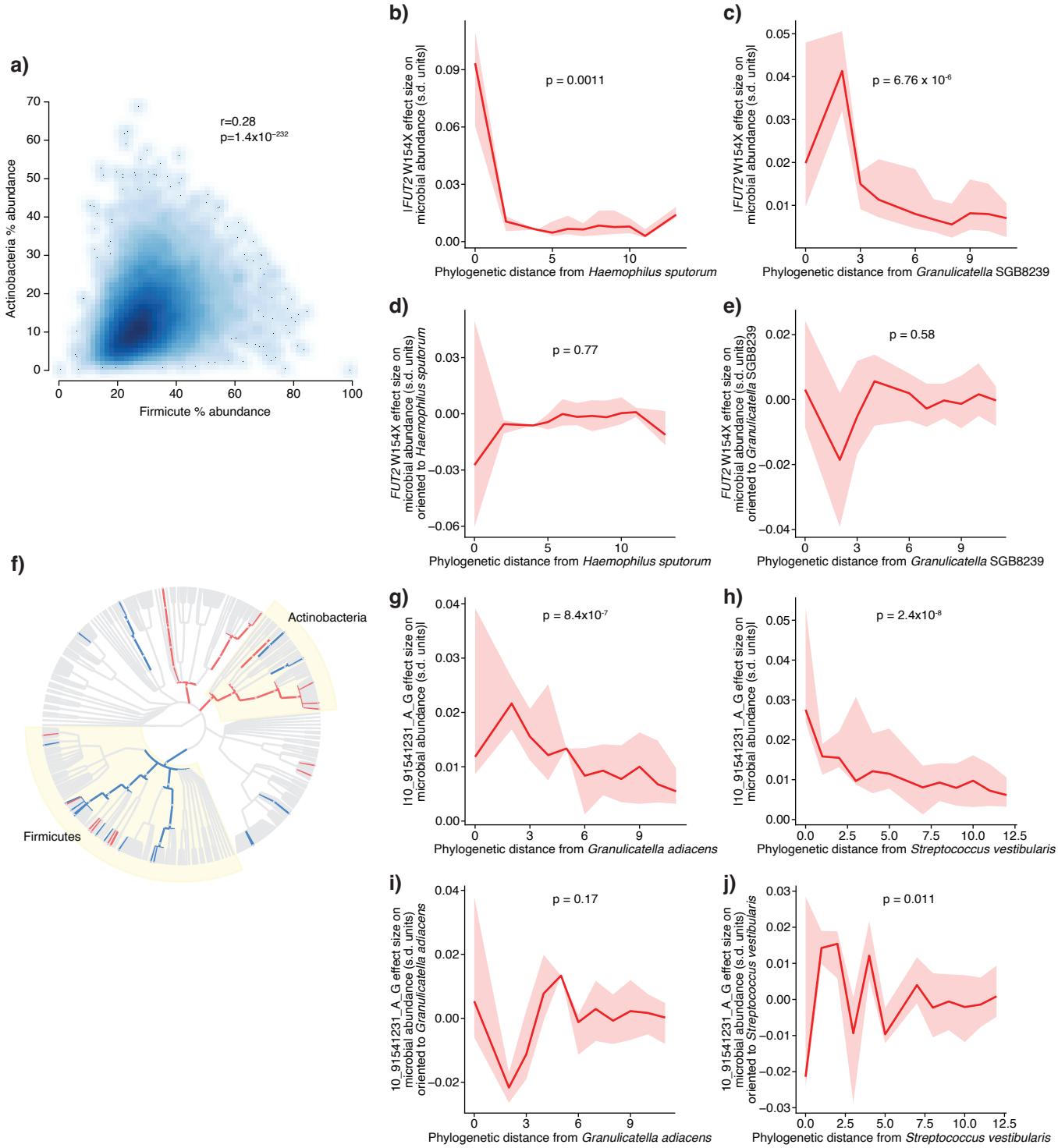
Extended Data Figure 2



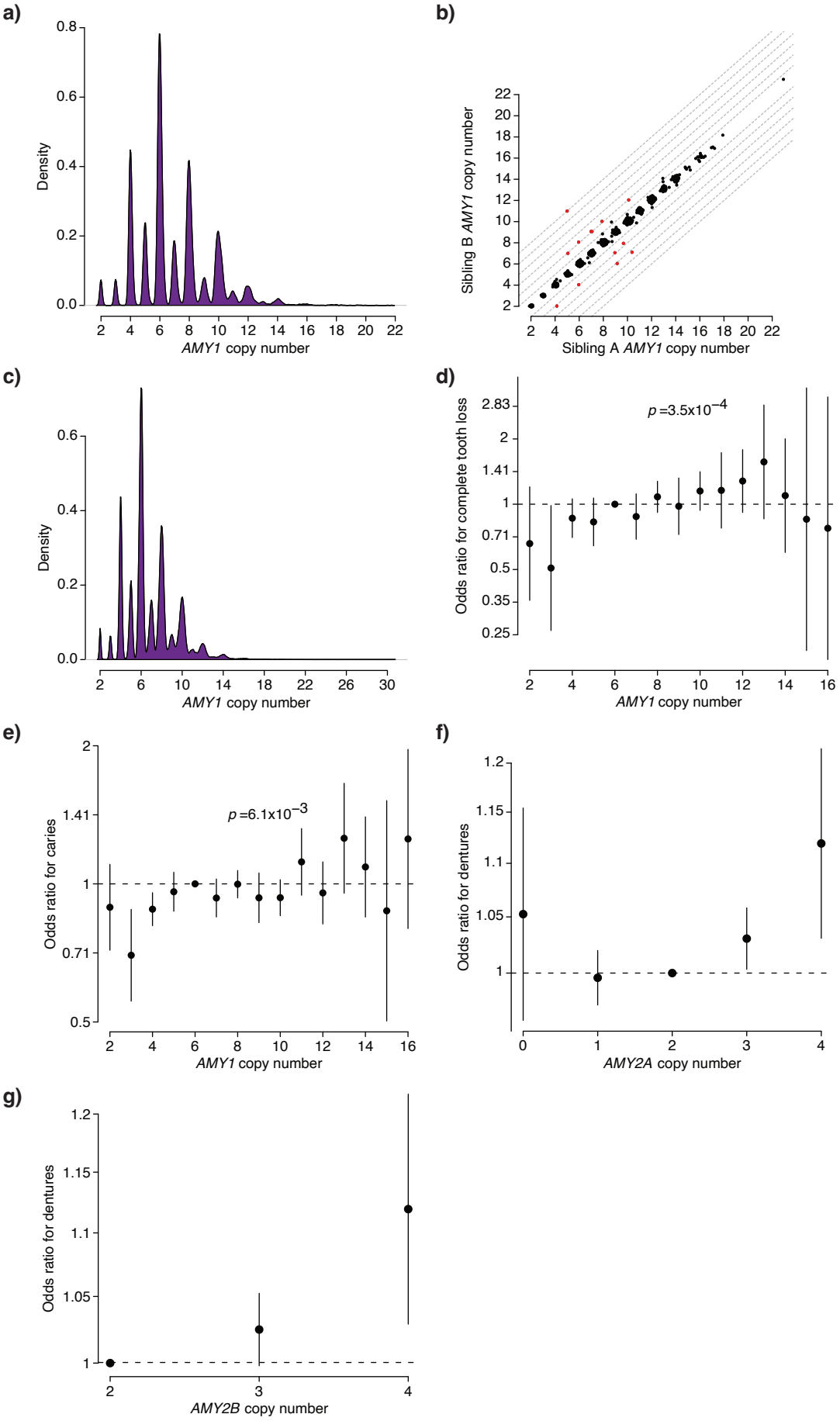
Extended Data Figure 3



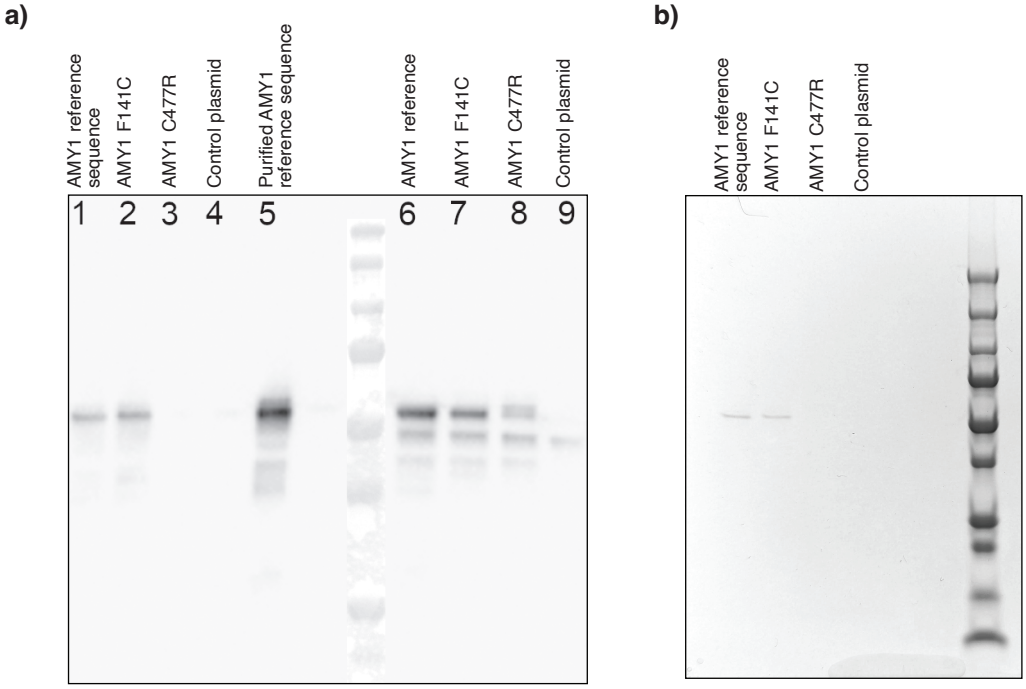
Extended Data Figure 4

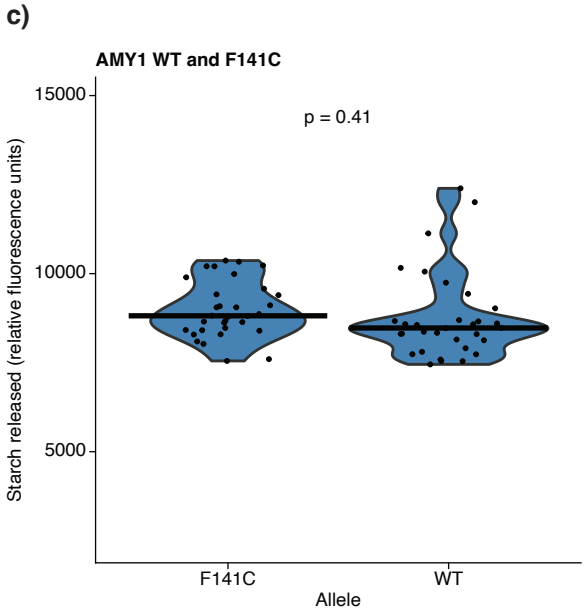


Extended Data Figure 5

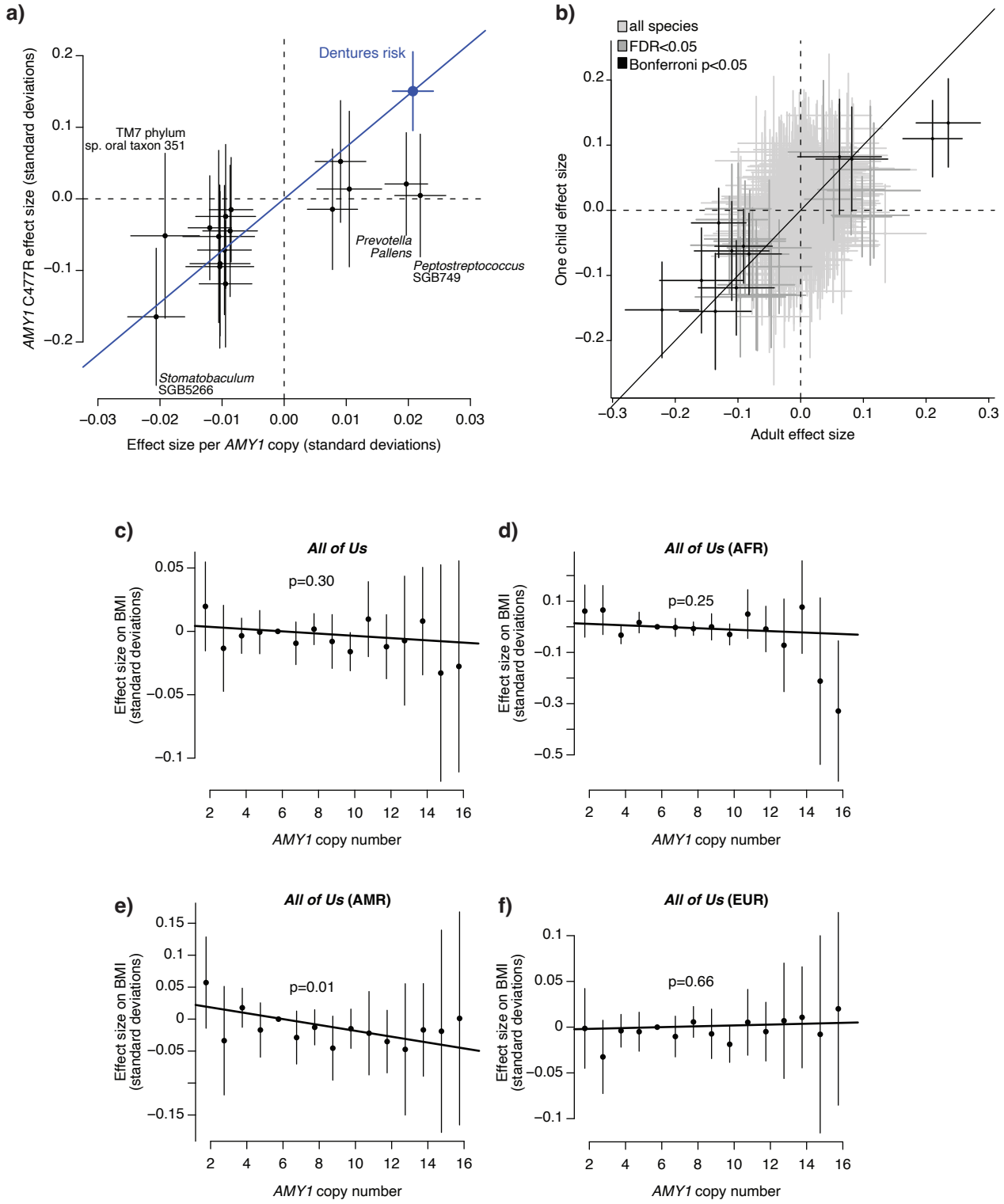


Extended Data Figure 6

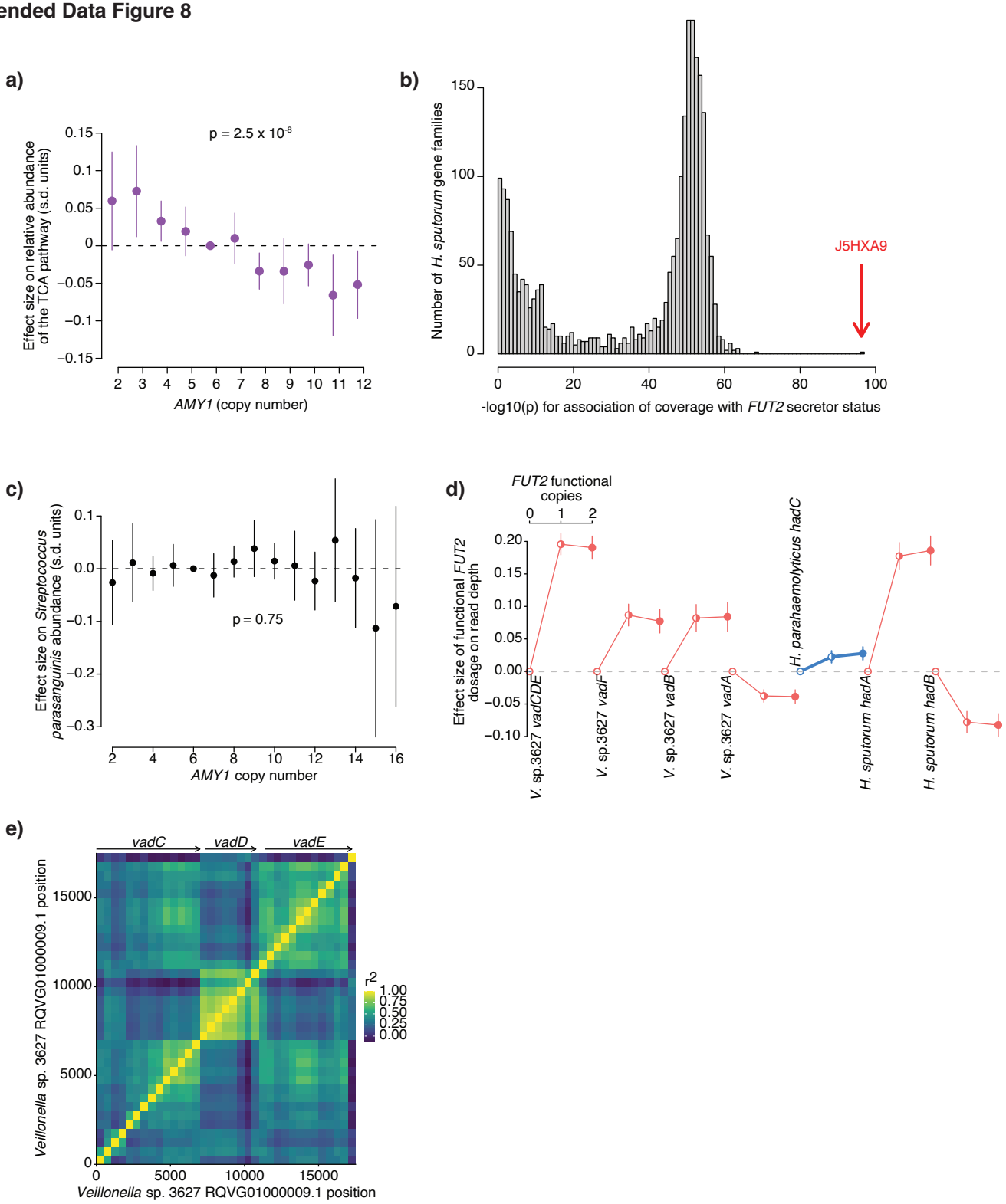




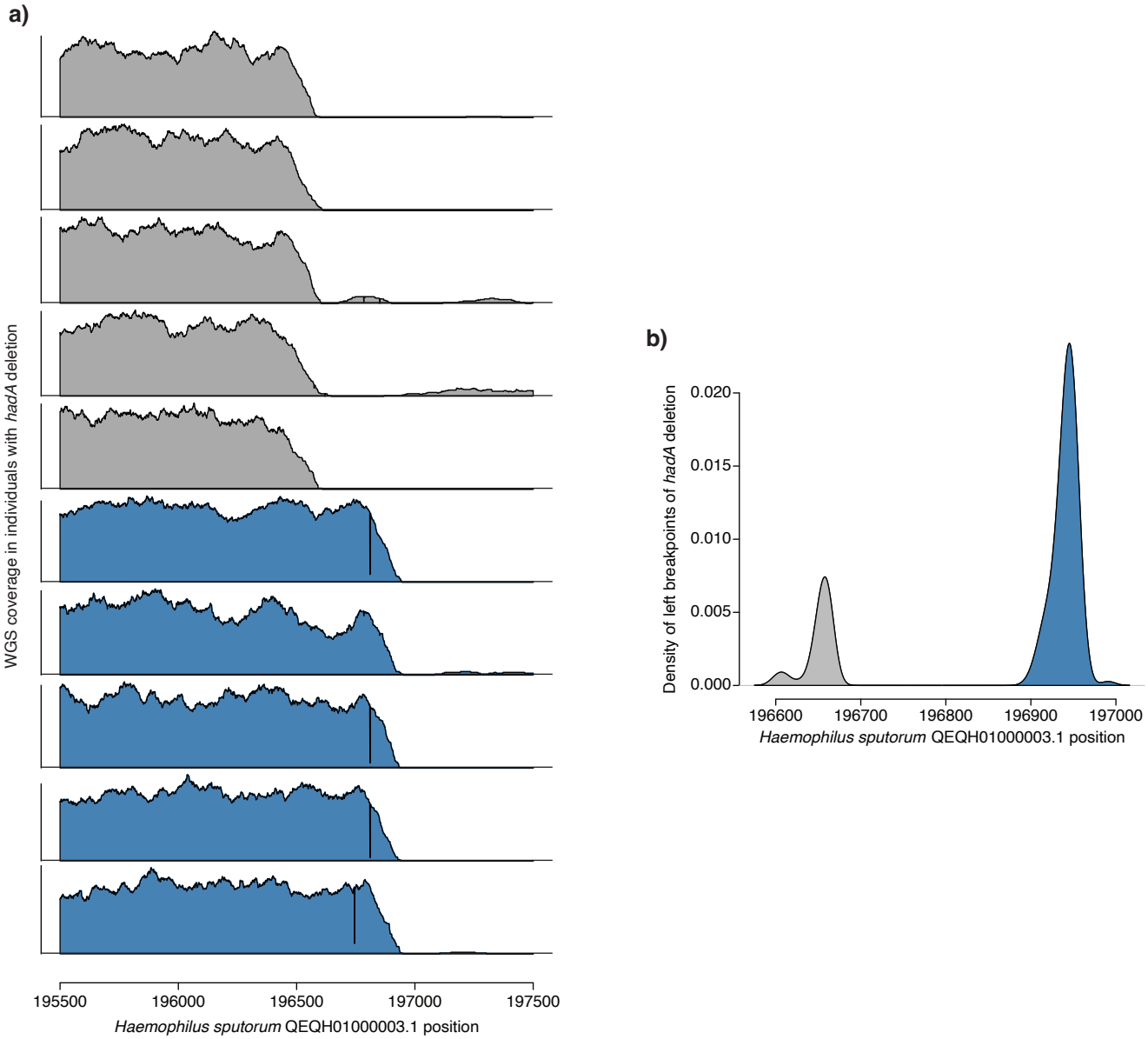
Extended Data Figure 7



Extended Data Figure 8

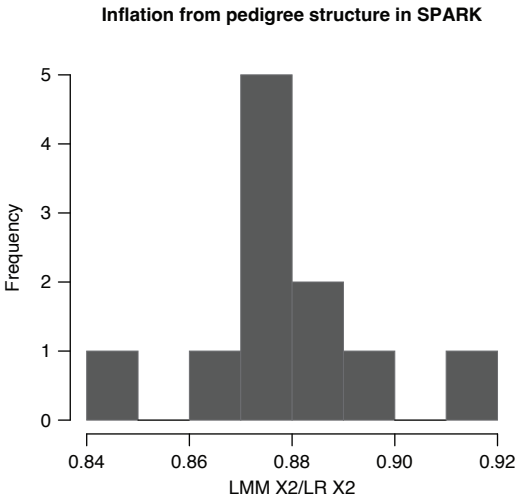
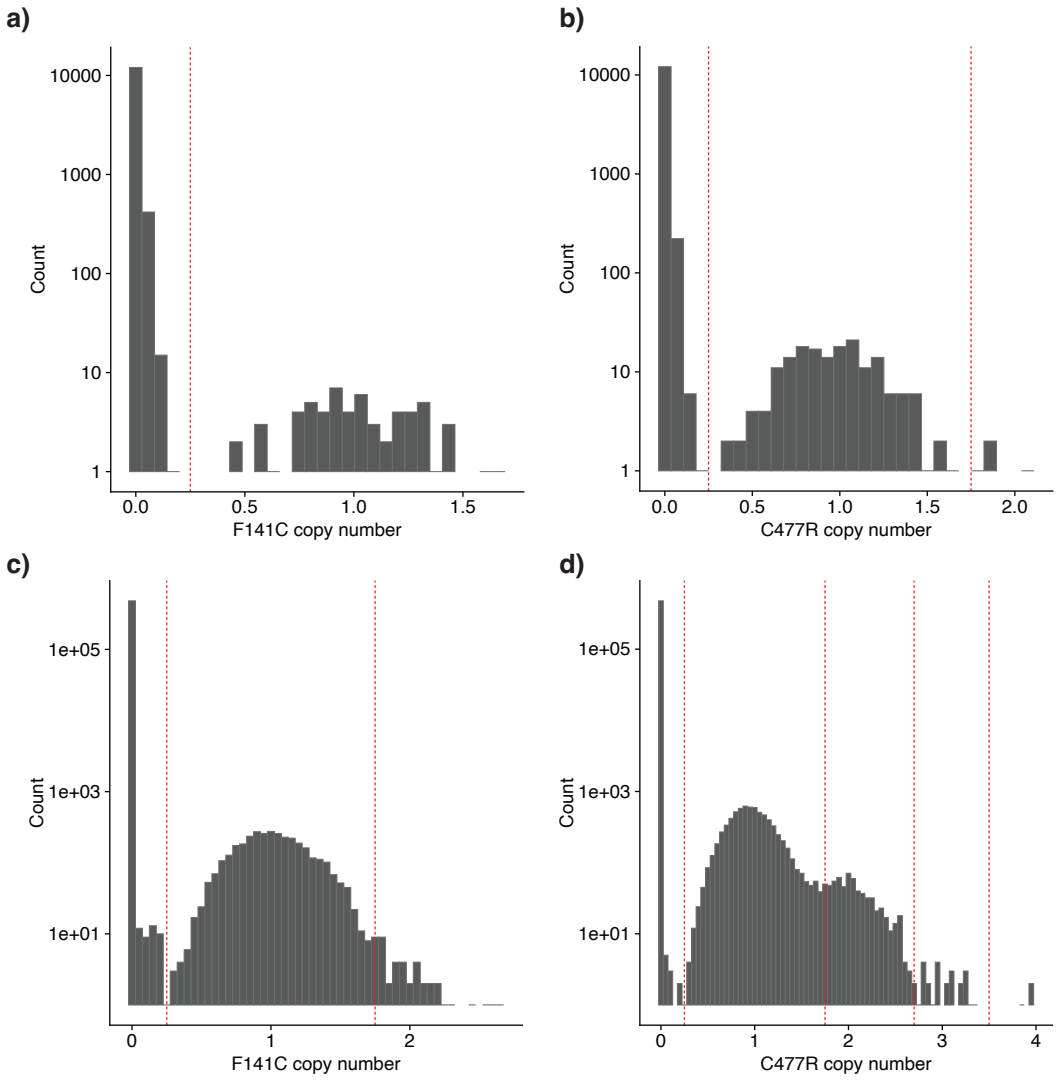


Extended Data Figure 9



Extended Data Figure 10

a)



Extended Data Figure 11