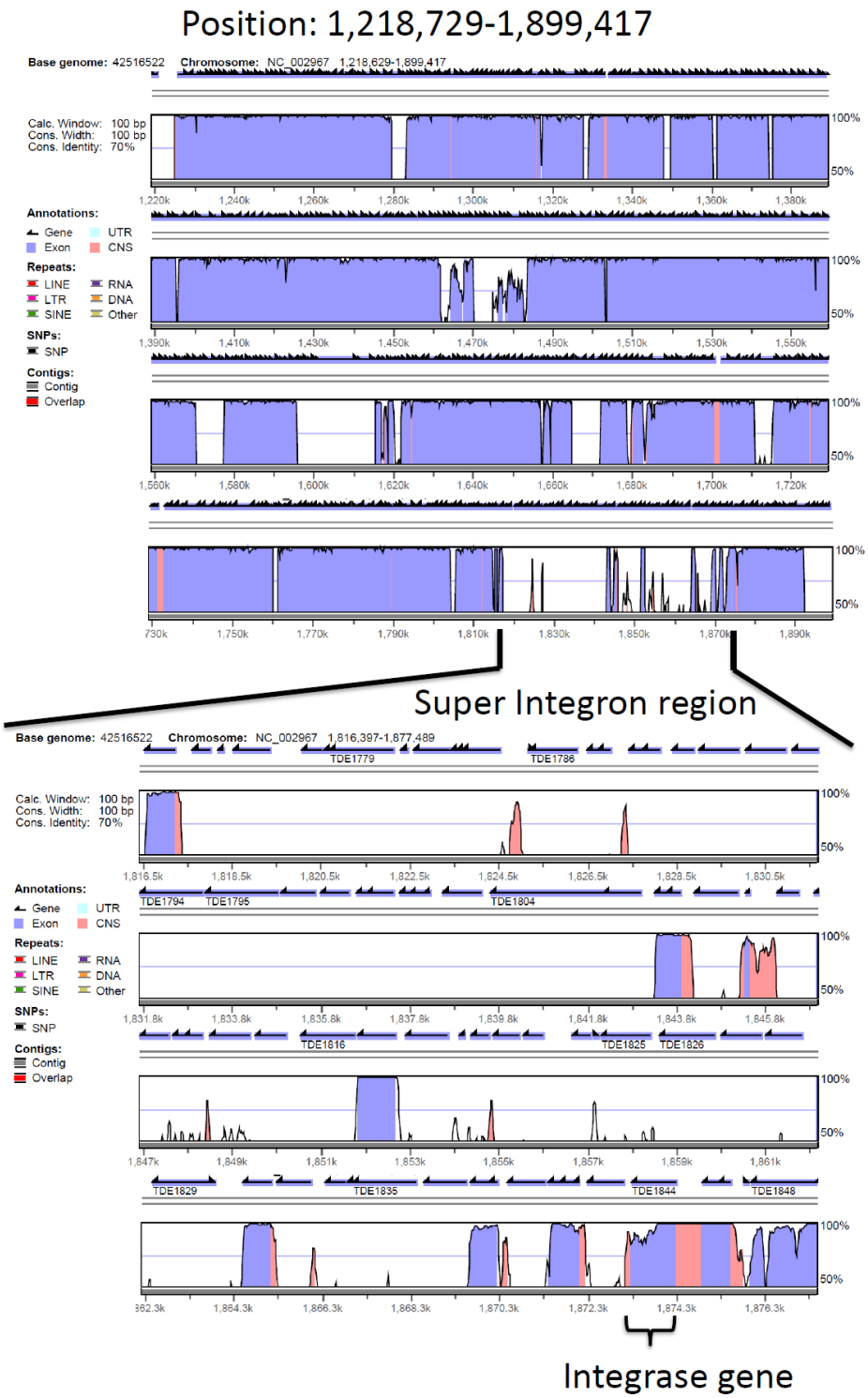

Figure S1. The mapping of *Treponema denticola* F0402 contig ADEC01000014 to the ATCC 35405 genome. The plots are generated using RankVISTA web service.

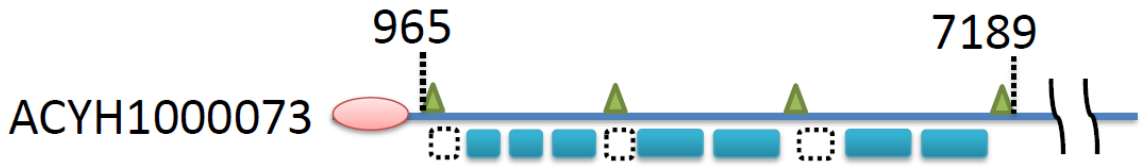

Figure S2. The predicted integron recombination sites and genes in the contig ACYH1000073 of *T. vincentii*. Triangles are recombination sites, rectangles represent the integron genes, and the oval is the *IntI* gene. We use solid rectangles to represent the genes that pass our integron gene discovery threshold, and dashed rectangles are open reading frames that do not meet the criteria.

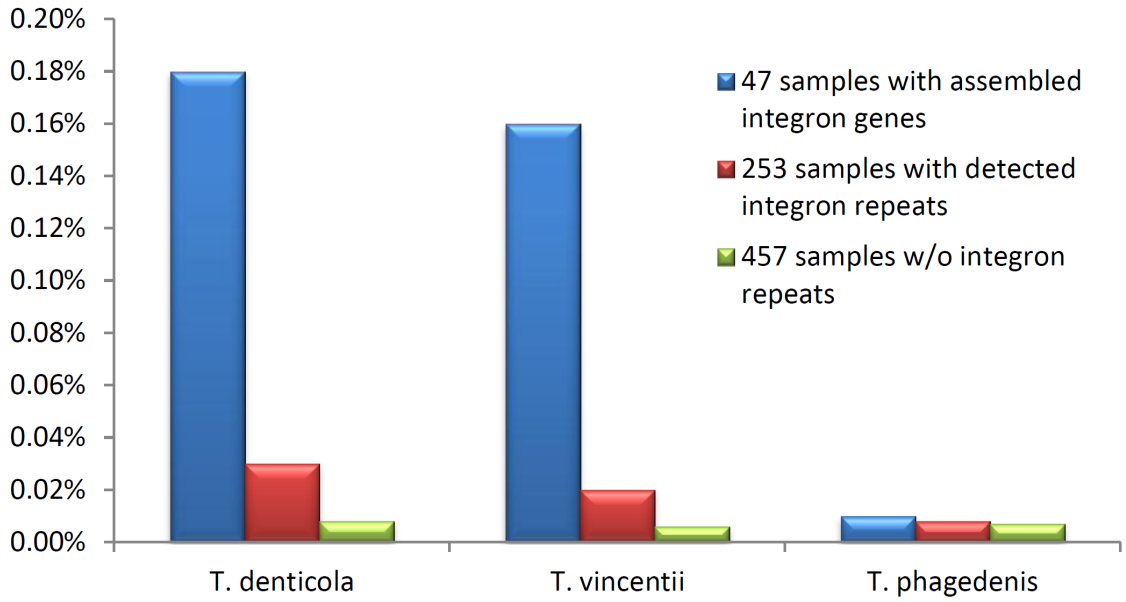

Figure S3. Comparison of the average abundances of the three integron-containing *Treponema* species in the HMP samples of different categories: samples with assembled integron genes (shown in blue), samples with detectable recombination sites (but no integron genes are assembled; shown in red), and samples without recombination sites detected (in green). The abundances in each HMP sample were estimated by mapping paired-end shotgun sequences of the HMP datasets onto the genomes (or genome drafts), by BWA (Li et al. 2009). Both reads in a pair are counted if at least one read maps to the genomes. Reads that map to common regions of genomes from different species are considered for all corresponding species in the estimation of the abundances. This chart confirms the existence of these *Treponema* species in the HMP datasets, with *T. denticola* and *T.vincintii* being more abundant in the samples. The mapping results are consistent with the results of the identification of *attC* sites and the integron gene cassettes: the samples with integron gene cassettes identified have the most *T. denticola* and *T. vincintii*, and the samples without recombination sites identified have the lowest presence of these species. This figure also suggests that the integron genes we identified are more likely to be from *T. denticola* and *T. vincintii*.

Table S1. Identified integron gene numbers for each sample using constrained assembly approach and whole metagenome assembly.

Sample-ID	Constrained assembly approach	Whole metagenome assembly
SRS011115	13	0
SRS011126	12	0
SRS011152	2	1
SRS011255	0	2
SRS013533	28	1
SRS013705	17	2
SRS013836	2	0
SRS013950	36	1
SRS014470	4	0
SRS014476	37	7
SRS014477	42	0
SRS014573	45	0
SRS014578	8	8
SRS014691	5	0
SRS015215	19	11
SRS015434	17	0
SRS016331	64	0
SRS017209	10	1
SRS017227	0	5
SRS017691	2	0
SRS018157	13	0
SRS018739	43	4
SRS019029	3	0
SRS019071	1	0
SRS022143	2	0
SRS022149	35	1
SRS022602	8	1
SRS023595	66	3

SRS024441	32	1
SRS024561	0	1
SRS042643	29	0
SRS045313	3	0
SRS047113	8	1
SRS047634	11	0
SRS049318	42	8
SRS049389	18	0
SRS050244	5	0
SRS050669	1	0
SRS051930	12	1
SRS055378	8	5
SRS055401	2	0
SRS057205	1	0
SRS058053	2	1
SRS058808	18	0
SRS062544	20	1
SRS063215	0	5
SRS063603	51	6
SRS063932	11	9
SRS063999	4	0
SRS064774	2	0
SRS075404	12	0

Table S2. The COG functional category distributions of the integron gene cassettes identified in different human body locations.

COG Functional Categories[1]	Supragingival plaque	Tongue dorsum	Subgingival plaque
[C] Energy production and conversion	1	0	0
[D] Cell cycle control, cell division, chromosome parititioning	8 (11)[2]	3	1
[E] Amino acid transport and metabolism	2 (4)	1	0
[F] Nucleotide transport and metabolism	1	0	0
[G] Carbohydrate transport and metabolism	1	0	4
[H] Coenzyme transport and metabolism	1	1 (8)	0
[I] Lipid transport and metabolism	1 (3)	0	0
[J] Translation, ribosomal structure and biogenesis	3	0	1
[K] Transcription	9	4	1
[L] Replication, recombination and repair	10 (12)	10 (14)	5 (6)
[M] Cell wall/membrane/envelope biogenesis	3 (4)	2 (3)	0
[N] Cell motility	8 (10)	2 (5)	1
[O] Posttranslational modification, protein turnover, chaperones	0	1	0
[P] Inorganic ion transport and metabolism	0	1	0
[R] General function prediction only	*45 (67)*	*13 (14)*	*7 (8)*
[S] Function unknown	*59 (72)*	*18 (21)*	*14 (15)*
[T] Signal transduction mechanisms	6 (10)	3	2 (3)
[U] Intracellular trafficking, secretion, and vesicular transport	2 (3)	1	0
[V] Defense mechanisms	2	3 (4)	1

[1]: the functional categories (including [A] RNA processing and modification, [B] Chromatin structure and dynamics, [Q] Secondary metabolites biosynthesis, transport and catabolism, [W] Extracellular structures, [Y] Nuclear structure, and [Z] Cytoskeleton) that have no gene cassettes are not listed in the table.

[2]: Number of genes is obtained by clustering the genes at a 97% identity threshold for each functional category within each location. Numbers within parentheses indicate the number of genes before clustering.

Table S3. Functions of genes related to species other than *T. denticola* or *T. vincentii*

Species (# of genes)	Gene functions	Number of genes[1]
Bacillales		
	Hydrolase	2 (3)
	Hypothetical	2 (3)
Bacteroidetes		
	Hypothetical	5 (6)
	DNA-cytosine methyltranferase	1 (1)
Clostridiales		
	Hypothetical	3 (4)
	D-alanine--D-alanine ligase	1 (2)
	Type II restriction enzyme HphI	1 (1)
	Acetyltransferase (GNAT) family	1 (1)
	Toxon-antitoxin system, antitoxin component, XRE family	1 (1)
	Hydrolase, NUDIX family	1 (1)
	Toxon-antitoxin system, toxin component, Txe/Yoe family	1 (1)
	ABC transporter, ATP-binding protein	1 (2)
	Toxon-antitoxin system, toxin component, RelE family	1 (1)
Flavobacteriaceae		
	Hypothetical transmembrane protein	1 (3)
	Hypothetical	1 (1)
	FRG domain protein	1 (1)
Gammaproteobacteria		
	Hypothetical membrane protein	1 (3)
	Type II restriction enzyme BanI	1 (1)
	DNA (cytosine-5-)-methyltransferase	1 (1)
	Hypothetical	7 (7)
Kosmotoga olearia		
	Methyltransferase type 11	1 (8)
Ricinus communis		
	Hypothetical protein	1 (10)
Spirochaeta caldaria DSM 7334		

	toxin-antitoxin system, toxin component, PIN family (PilT domain)	1 (4)
	Prevent-host-death family	1 (1)
	Hypothetical	4 (20)
Treponema phagedenis F0421		
	Restriction endonuclease	3 (3)
	Hypothetical	4 (4)
Treponema succinifaciens DSM 2489		
	XRE family transcriptional regulator	1 (1)
	Plasmid maintenance system killer	1 (1)
	hypothetical	8 (13)
	Transcriptional modulator of MazE/toxin, MazF	1 (5)

[1]: The numbers indicate the unique gene numbers by clustering the genes using a 97% identity threshold. Number of genes before clustering is shown within parentheses.

Table S4. The COG function category distributions of the integron gene cassettes shared by two or more samples.

COG Functional Categories[1]	Genes shared between exactly two samples	Gene shared among three or more samples
[D] Cell cycle control, cell division, chromosome partitioning	2	2
[E] Amino acid transport and metabolism	0	1
[G] Carbohydrate transport and metabolism	1	0
[H] Coenzyme transport and metabolism	0	1
[I] Lipid transport and metabolism	0	1
[J] Translation, ribosomal structure and biogenesis	1	0
[K] Transcription	2	1
[L] Replication, recombination and repair	1	4
[M] Cell wall/membrane/envelope biogenesis	3	0
[N] Cell motility	0	3
[P] Inorganic ion transport and metabolism	0	0
[R] General function prediction only	*7*	*8*
[S] Function unknown	*12*	*9*
[T] Signal transduction mechanisms	0	2
[U] Intracellular trafficking, secretion, and vesicular transport	0	1
[V] Defense mechanisms	0	1

[1]: the functional categories (including [A] RNA processing and modification, [B] Chromatin structure and dynamics, [C] Energy production and conversion, [F] Nucleotide transport and metabolism, [O] Posttranslational modification, protein turnover, chaperones, [Q] Secondary metabolites biosynthesis, transport and catabolism, , [W] Extracellular structures, [Y] Nuclear structure, and [Z] Cytoskeleton) that have no gene cassettes are not listed in the table.

References

Li H and Durbin R 2009. Fast and accurate short read alignment with Burrows-Wheeler transform. Bioinformatics **25**(14): 1754-1760.