Microbial Variome Database: Point Mutations, Adaptive or Not, in Bacterial Core Genomes

Sujay Chattopadhyay,*[1] Fred Taub,[1] Sandip Paul,[1] Scott J. Weissman,[2] and Evgeni V. Sokurenko*[1]

[1]Department of Microbiology, University of Washington, Seattle

[2]Center for Childhood Infections and Prematurity Research, Seattle Children's Research Institute, Seattle, Washington

***Corresponding authors:** E-mail: sujayc@u.washington.edu; evs@u.washington.edu.

Associate editor: Howard Ochman

Abstract

Analysis of genetic differences (gene presence/absence and nucleotide polymorphisms) among strains of a bacterial species is crucial to understanding molecular mechanisms of bacterial pathogenesis and selecting targets for novel antibacterial therapeutics. However, lack of genome-wide association studies on large and epidemiologically well-defined strain collections from the same species makes it difficult to identify the genes under positive selection and define adaptive polymorphisms in those genes. To address this need and to overcome existing limitations, we propose to create a "microbial variome"—a species-specific resource database of genomic variations based on molecular evolutionary analysis. Here, we present prototype variome databases of *Escherichia coli* and *Salmonella enterica* subspecies *enterica* (http://depts.washington.edu/sokurel/variome, last accessed March 26, 2013). The prototypes currently include the point mutations data of core protein-coding genes from completely sequenced genomes of 22 *E. coli* and 17 *S. enterica* strains. These publicly available databases allow for single- and multiple-field sorting, filtering, and searching of the gene variability data and the potential adaptive significance. Such resource databases would immensely help experimental research, clinical diagnostics, epidemiology, and environmental control of human pathogens.

Key words: **microbial variome, adaptive evolution, nucleotide polymorphisms, database.**

Letter

In the microbial world, organisms belonging to same species exhibit a diverse set of phenotypic traits, both quantitative and qualitative, resulting in extensive heterogeneity in appearance, metabolism, ability to occupy specific habitats, cause disease, or to resist antimicrobial agents (Orr and Smith 1998; Anisimova and Liberles 2007). Such diversity, in many cases, is defined not only by the presence/absence of certain genes or large gene clusters (Brookfield 2005; Pallen and Wren 2007; Keeling and Palmer 2008) but also by mutational variations in the shared genes, often as single-nucleotide polymorphisms (SNPs) (Feil 2004; Knight et al. 2006; Henriques-Normark et al. 2008). The gene pool of a species is represented by the combination of all genes and their variants that occur in individuals belonging to the given species. The pool is continuously enriched by mutation, sifting and shifting of representative genes, either by random genetic drift or by various selective pressures (Morjan and Rieseberg 2004; Sokurenko et al. 2006; Perron et al. 2008; Rocha 2008). The field of population genetics allows assessment of structure, organization, and dynamics of these gene pools by comparative sequence analysis of several genetic loci across different populations (Kassen and Rainey 2004; Mes 2008). Previously, however, population genetics has been traditionally restricted by the use of limited numbers of loci, due to sequencing and labor expenses (Butlin 2010). This, in turn, has led to major limitations in our ability to detect genome-wide, positively selected variations and to associate them with specific phenotypic diversity that defines fitness in specific environments (Conrad and Hurles 2007). Moreover, lack of experimental knowledge of the genetic basis for most phenotypic traits is another key problem, even for the best studied organisms.

From this perspective, the benefits of whole-genome sequencing and genome-wide search for functional sequence variations are now well recognized. With the flood of new genomic data arising from affordable, rapid, high-quality sequencing technologies, a new field of population genomics has now emerged. It is focused on population-level genetic analysis of variations of individual organisms across the gene pool for a given species (DeLong 2004; Whitaker and Banfield 2006; Brinkman and Parkhill 2008; Stinchcombe and Hoekstra 2008) with the ultimate goal of linking gene function to specific physiological traits (Fournier et al. 2007). We have recently reported a novel analytical approach, Zonal Phylogeny, that detects phylogenetically unlinked mutations at the same amino acid positions, termed hotspot mutations (Sokurenko et al. 2004; Chattopadhyay et al. 2007). These mutations represent a strong indicator of adaptive evolution via molecular convergence of protein variants. Our recent work employing this analytical tool (Chattopadhyay et al. 2009, 2012), along with other studies (Lefebure and Stanhope 2007; Petersen et al. 2007; Soyer et al. 2009), represented initial attempts to provide the research community with a set of positively selected loci that presumably are important in defining the pathogenic potential of organisms, even though the functions of those loci are not yet fully understood.

Here, we present a unique approach to incorporate all results of the molecular evolutionary, selection-based analysis for a microbial species under the umbrella of a free publicly available database. We propose to designate such species-specific resources of genome-wide variations, and their potential adaptive significance, as a "microbial variome," by analogy with the human variome database (http://www.humanvariomeproject.org/, last accessed March 26, 2013). In this article, we announce variome databases for *Escherichia coli* and *Salmonella enterica* subspecies *enterica*, based on the core protein-coding genes from 22 and 17 genomes (supplementary tables S1 and S2, Supplementary material online), respectively. Both for *E. coli* and *Salmonella*, we primarily selected the strains that were clonally distinct, that is, having different alleles of housekeeping genes used for multilocus sequence typing analysis. The only exceptions were two pairs of clonally identical strains—extraintestinal pathogenic strains S88 and UTI89 for *E. coli*, and serovar Paratyphi A strains ATCC 9150 and AKU_12601 for *Salmonella*. As a next step, we will expand the list of strains for both species, along with incorporating all protein-coding genes—the core (included in the present version of the database) and the noncore ones.

Database Description

Microbial variome databases are built with relational database software, with a platform-agnostic, browser-based public front end, to curate species-specific project information and analyzed data. The benefits of a relational database are many and well documented (Codd 1970). The basic idea is to present the genomic data in a way that makes it easy to start from a bird's-eye view of the genomic diversity of an entire species or of specific groups/clones and to drill down, in intuitive ways, to individual genomes, genes, and polymorphisms (fig. 1).

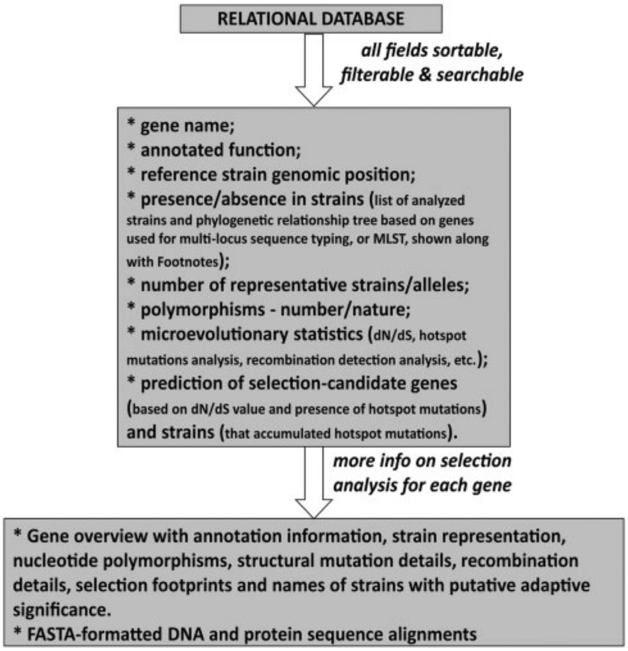

RELATIONAL DATABASE
all fields sortable, filterable & searchable
* gene name;
* annotated function;
* reference strain genomic position;
* presence/absence in strains (list of analyzed strains and phylogenetic relationship tree based on genes used for multi-locus sequence typing, or MLST, shown along with Footnotes);
* number of representative strains/alleles;
* polymorphisms - number/nature;
* microevolutionary statistics (dN/dS, hotspot mutations analysis, recombination detection analysis, etc.);
* prediction of selection-candidate genes (based on dN/dS value and presence of hotspot mutations) and strains (that accumulated hotspot mutations).
more info on selection analysis for each gene
* Gene overview with annotation information, strain representation, nucleotide polymorphisms, structural mutation details, recombination details, selection footprints and names of strains with putative adaptive significance.
* FASTA-formatted DNA and protein sequence alignments

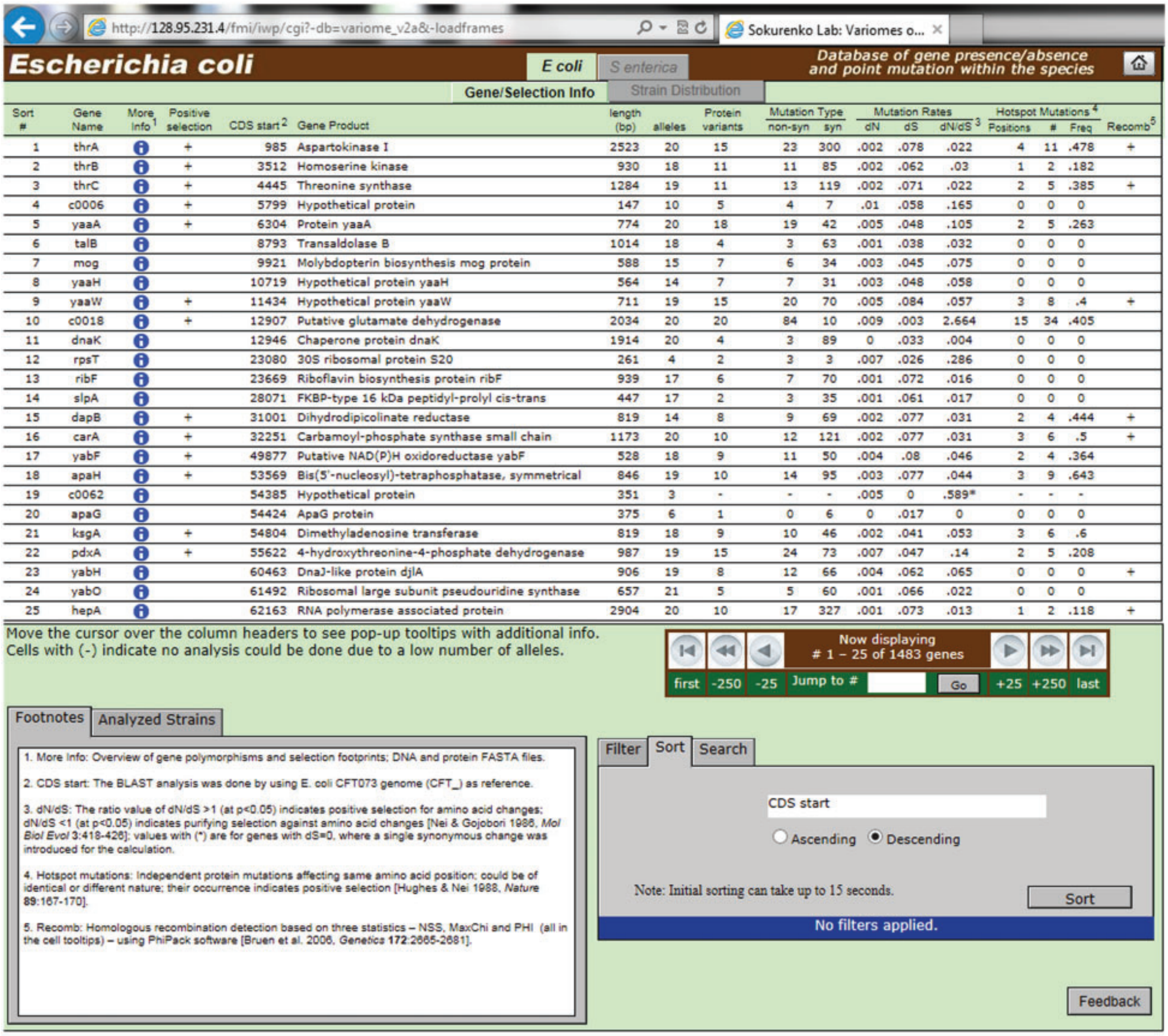
Fig. 1. Schematic architecture of microbial variome database.

The prototype core variome databases of *E. coli* and *Salmonella* are focused on innovative, clear visual presentations of data depicting the core genes' polymorphism diversity. We present the information for both species within a single, unified webpage structure. This makes the relevant data of interest (such as synonymous and nonsynonymous variability, recombination signals, DNA and protein sequence alignments, list of amino acid mutations, and footprints of positive selection based on hotspot mutations and dN/dS statistics) easily accessible at every stage. Although the present version does not offer user any option to enter sequences for comparing with the ones in the database, the user can easily track the gene(s) of interest via sort/filter/search facility and can get access to both DNA and protein sequence data for the corresponding gene(s) as explained below.

For each species database, the website has a pair of overview layouts:

a) "Gene/Selection Info," which includes reference-genome annotation of the gene, number of representative strains and alleles, number and nature (synonymous/nonsynonymous) of polymorphisms, analysis results of dN/dS statistics, hotspot mutations, and recombination detection, along with prediction of positive selection effects based on dN/dS values and presence/absence of hotspot mutations (fig. 2).
b) "Strain Distribution," which indicates strains that accumulate hotspot mutations in positively selected genes (fig. 3).

Each layout offers optional single- and multiple-field sorting, filtering, and keyword-based searching. These layouts enable the user to rapidly search or scan the entire set, or a filtered subset of genes, for data of particular interest. The "More Info" button links to an analysis overview for each gene, along with the corresponding FASTA-formatted DNA/protein sequence alignment data. As an example, figure 4 represents a snapshot of a portion of "Gene Overview" for *Salmonella* gene *fimH*, encoding the type 1 fimbrial adhesin protein. The results suggest that the gene evolves under positive selection via accumulation of hotspot mutations, in the absence of any intragenic recombination (as detected by three recombination detection statistics—pairwise homoplasy index, maximum χ^2, and neighbor similarity score—incorporated in PhiPack software package [Bruen et al. 2006]). This prediction is validated by our recent work on the *fimH* gene with a larger set of *Salmonella* strains, demonstrating adaptive convergent evolution of the adhesive protein (Kisiela et al. 2012). It is worth mentioning here that we assess for only intragenic recombination (i.e., exchange of small internal regions in genes) to detect the presence of hotspot mutations of non-recombinant origin. In contrast, if the polymorphisms result from intergenic recombination (i.e., exchange of larger segments of the chromosome and spanning multiple genes), the parent and recombinant alleles of


http://128.95.231.4/fmi/iwp/cgi?-db=variome_v2a&-loadframes
Sokurenko Lab: Variomes o...

Escherichia coli

Database of gene presence/absence and point mutation within the species

Gene/Selection Info
Strain Distribution

Sort	Gene	More	Positive			length		Protein	Mutation Type		Mutation Rates			Hotspot Mutations[4]			
#	Name	Info[1]	selection	CDS start[2]	Gene Product	(bp)	alleles	variants	non-syn	syn	dN	dS	dN/dS[3]	Positions	#	Freq	Recomb[5]
1	thrA	ⓘ	+	985	Aspartokinase I	2523	20	15	23	300	.002	.078	.022	4	11	.478	+
2	thrB	ⓘ	+	3512	Homoserine kinase	930	18	11	11	85	.002	.062	.03	1	2	.182	
3	thrC	ⓘ	+	4445	Threonine synthase	1284	19	11	13	119	.002	.071	.022	2	5	.385	+
4	c0006	ⓘ	+	5799	Hypothetical protein	147	10	5	4	7	.01	.058	.165	0	0	0	
5	yaaA	ⓘ	+	6304	Protein yaaA	774	20	18	19	42	.005	.048	.105	2	5	.263	
6	talB	ⓘ		8793	Transaldolase B	1014	18	4	3	63	.001	.038	.032	0	0	0	
7	mog	ⓘ		9921	Molybdopterin biosynthesis mog protein	588	15	7	6	34	.003	.045	.075	0	0	0	
8	yaaH	ⓘ		10719	Hypothetical protein yaaH	564	14	7	7	31	.003	.048	.058	0	0	0	
9	yaaW	ⓘ	+	11434	Hypothetical protein yaaW	711	19	15	20	70	.005	.084	.057	3	8	.4	+
10	c0018	ⓘ	+	12907	Putative glutamate dehydrogenase	2034	20	20	84	10	.009	.003	2.664	15	34	.405	
11	dnaK	ⓘ		12946	Chaperone protein dnaK	1914	20	4	3	89	0	.033	.004	0	0	0	
12	rpsT	ⓘ		23080	30S ribosomal protein S20	261	4	2	3	3	.007	.026	.286	0	0	0	
13	ribF	ⓘ		23669	Riboflavin biosynthesis protein ribF	939	17	6	7	70	.001	.072	.016	0	0	0	
14	slpA	ⓘ		28071	FKBP-type 16 kDa peptidyl-prolyl cis-trans	447	17	2	3	35	.001	.061	.017	0	0	0	
15	dapB	ⓘ	+	31001	Dihydrodipicolinate reductase	819	14	8	9	69	.002	.077	.031	2	4	.444	+
16	carA	ⓘ	+	32251	Carbamoyl-phosphate synthase small chain	1173	20	10	12	121	.002	.077	.031	3	6	.5	+
17	yabF	ⓘ	+	49877	Putative NAD(P)H oxidoreductase yabF	528	18	9	11	50	.004	.08	.046	2	4	.364	
18	apaH	ⓘ	+	53569	Bis(5'-nucleosyl)-tetraphosphatase, symmetrical	846	19	10	14	95	.003	.077	.044	3	9	.643	
19	c0062	ⓘ		54385	Hypothetical protein	351	3	-	-	-	.005	0	.589*	-	-	-	
20	apaG	ⓘ		54424	ApaG protein	375	6	1	0	6	0	.017	0	0	0	0	
21	ksgA	ⓘ	+	54804	Dimethyladenosine transferase	819	18	9	10	46	.002	.041	.053	3	6	.6	
22	pdxA	ⓘ	+	55622	4-hydroxythreonine-4-phosphate dehydrogenase	987	19	15	24	73	.007	.047	.14	2	5	.208	
23	yabH	ⓘ		60463	DnaJ-like protein djlA	906	19	8	12	66	.004	.062	.065	0	0	0	+
24	yabO	ⓘ		61492	Ribosomal large subunit pseudouridine synthase	657	21	5	5	60	.001	.066	.022	0	0	0	
25	hepA	ⓘ		62163	RNA polymerase associated protein	2904	20	10	17	327	.001	.073	.013	1	2	.118	+

Move the cursor over the column headers to see pop-up tooltips with additional info.
Cells with (-) indicate no analysis could be done due to a low number of alleles.

Now displaying
1 – 25 of 1483 genes

1. More Info: Overview of gene polymorphisms and selection footprints; DNA and protein FASTA files.

2. CDS start: The BLAST analysis was done by using E. coli CFT073 genome (CFT_) as reference.

3. dN/dS: The ratio value of dN/dS >1 (at p<0.05) indicates positive selection for amino acid changes; dN/dS <1 (at p<0.05) indicates purifying selection against amino acid changes [Nei & Gojobori 1986, *Mol Biol Evol* 3:418-426]; values with (*) are for genes with dS=0, where a single synonymous change was introduced for the calculation.

4. Hotspot mutations: Independent protein mutations affecting same amino acid position; could be of identical or different nature; their occurrence indicates positive selection [Hughes & Nei 1988, *Nature* 89:167-170].

5. Recomb: Homologous recombination detection based on three statistics – NSS, MaxChi and PHI (all in the cell tooltips) – using PhiPack software [Bruen et al. 2006, *Genetics* **172**:2665-2681].

CDS start

○ Ascending ● Descending

Note: Initial sorting can take up to 15 seconds.

No filters applied.

Fig. 2. Snapshot of *Escherichia coli* variome "Gene/Selection Info" layout, showing default sorted list of genes in ascending order of the genomic location in reference genome.

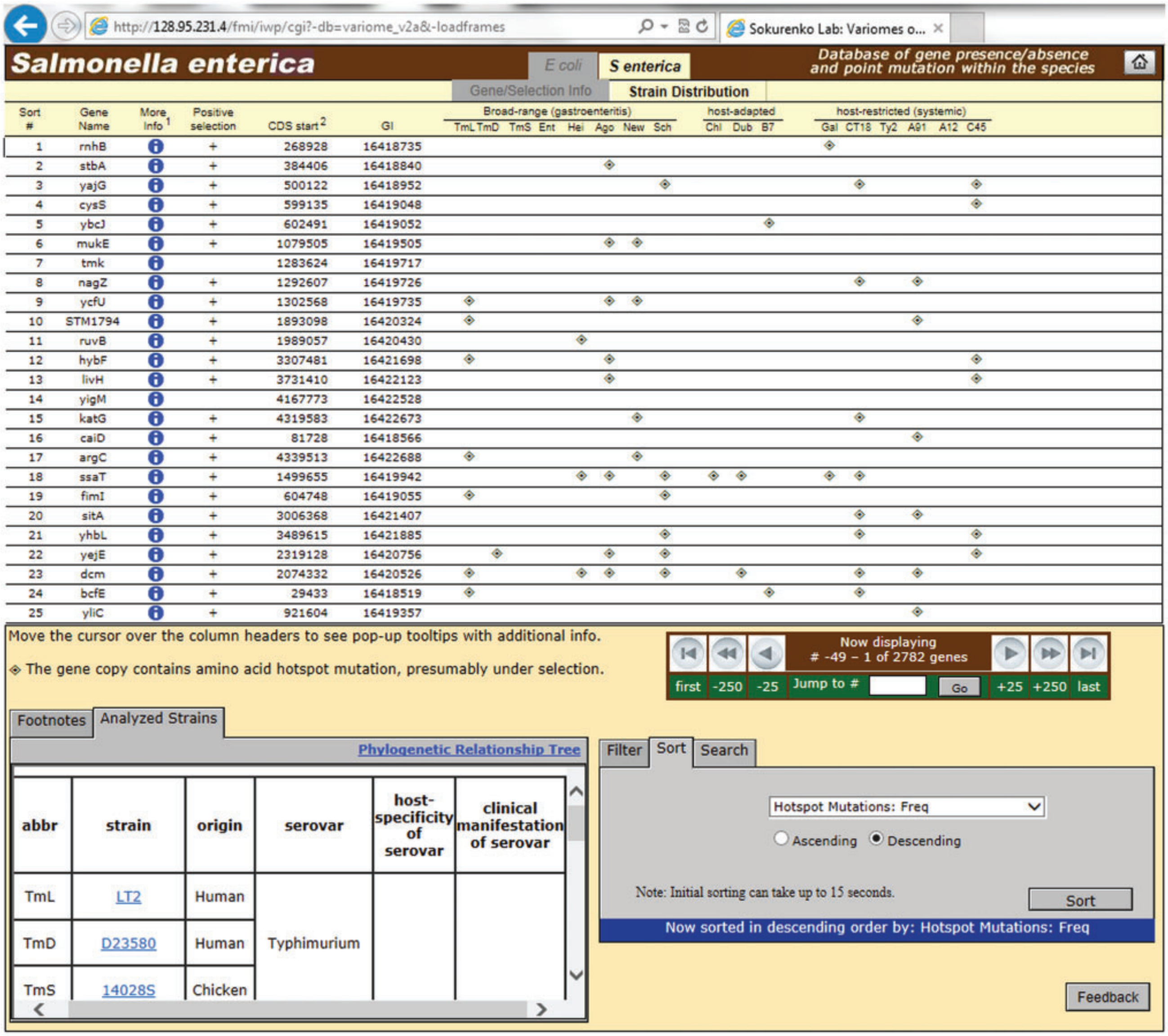
a single gene would be phylogenetically linked and thus not appear as hotspot change in our analysis. Therefore, the identification of positively selected genes via hotspot mutations will not be affected by the presence of intergenic recombination. However, in such cases, the list of SNPs might be mischaracterized, along with the calculations of dN and dS (rate of nonsynonymous and synonymous mutations). Also, we applied high threshold of 95% for both nucleotide sequence identity and length coverage to find an ortholog of each reference gene across the strains. We believe that this selection of high cutoff value minimizes the incorporation of (highly diverse) intergenic recombinants as orthologs.

As proximate goals, the presently available core variome databases enable the *E. coli* and *Salmonella* research communities to easily obtain 1) basic population genetics data on core protein-coding genes across the genomes (based on the reference genome annotation) and 2) information on naturally occurring mutations that are potentially (patho)adaptive for the microorganisms of interest.

Availability

The core *E. coli* and *Salmonella* variome databases are hosted by the Sokurenko's laboratory via the Department of Microbiology, University of Washington, and can be accessed via web browser at http://depts.washington.edu/sokurel/variome, last accessed March 26, 2013. The databases are developed with FileMaker Pro Advanced. They are hosted by FileMaker Server Advanced, which enables access by both FileMaker Pro clients and via any modern web browsers.

The home page of microbial variome databases features a "Downloads" section where Excel files containing the core variome data for *E. coli* and *S. enterica* are available for downloading. Also, each page of the variome databases includes a "Feedback" button. By clicking the button from any page, users are welcome to enter their questions, comments, concerns, and recommendations, along with their email addresses if a response is desired, or anonymously if they so choose. As part of maintenance of the databases, we will regularly respond to the feedbacks from the users.


http://128.95.231.4/fmi/iwp/cgi?-db=variome_v2a&-loadframes
Sokurenko Lab: Variomes o...
Salmonella enterica
E coli
S enterica
Database of gene presence/absence and point mutation within the species
Gene/Selection Info
Strain Distribution

Sort #	Gene Name	More Info[1]	Positive selection	CDS start[2]	GI	Broad-range (gastroenteritis) TmL	TmD	TmS	Ent	Hei	Ago	New	Sch	host-adapted Chl	Dub	B7	host-restricted (systemic) Gal	CT18	Ty2	A91	A12	C45
1	rnhB	ⓘ	+	268928	16418735												◈					
2	stbA	ⓘ	+	384406	16418840						◈											
3	yajG	ⓘ	+	500122	16418952								◈					◈				◈
4	cysS	ⓘ	+	599135	16419048																	◈
5	ybcJ	ⓘ	+	602491	16419052											◈						
6	mukE	ⓘ	+	1079505	16419505						◈	◈										
7	tmk	ⓘ		1283624	16419717																	
8	nagZ	ⓘ	+	1292607	16419726													◈		◈		
9	ycfU	ⓘ	+	1302568	16419735	◈					◈	◈										
10	STM1794	ⓘ	+	1893098	16420324	◈														◈		
11	ruvB	ⓘ	+	1989057	16420430					◈												
12	hybF	ⓘ	+	3307481	16421698	◈					◈											◈
13	livH	ⓘ	+	3731410	16422123						◈											◈
14	yigM	ⓘ		4167773	16422528																	
15	katG	ⓘ	+	4319583	16422673							◈						◈				
16	caiD	ⓘ	+	81728	16418566															◈		
17	argC	ⓘ	+	4339513	16422688	◈						◈										
18	ssaT	ⓘ	+	1499655	16419942					◈	◈		◈	◈	◈		◈	◈				
19	fimI	ⓘ	+	604748	16419055	◈							◈									
20	sitA	ⓘ	+	3006368	16421407													◈		◈		
21	yhbL	ⓘ	+	3489615	16421885								◈					◈				◈
22	yejE	ⓘ	+	2319128	16420756		◈				◈		◈									◈
23	dcm	ⓘ	+	2074332	16420526	◈				◈	◈		◈		◈			◈		◈		
24	bcfE	ⓘ	+	29433	16418519	◈										◈		◈				
25	yliC	ⓘ	+	921604	16419357															◈		

Move the cursor over the column headers to see pop-up tooltips with additional info.

◈ The gene copy contains amino acid hotspot mutation, presumably under selection.

Now displaying # -49 – 1 of 2782 genes

abbr	strain	origin	serovar	host-specificity of serovar	clinical manifestation of serovar
TmL	LT2	Human	Typhimurium		
TmD	D23580	Human			
TmS	14028S	Chicken			

Hotspot Mutations: Freq

○ Ascending ● Descending

Note: Initial sorting can take up to 15 seconds.

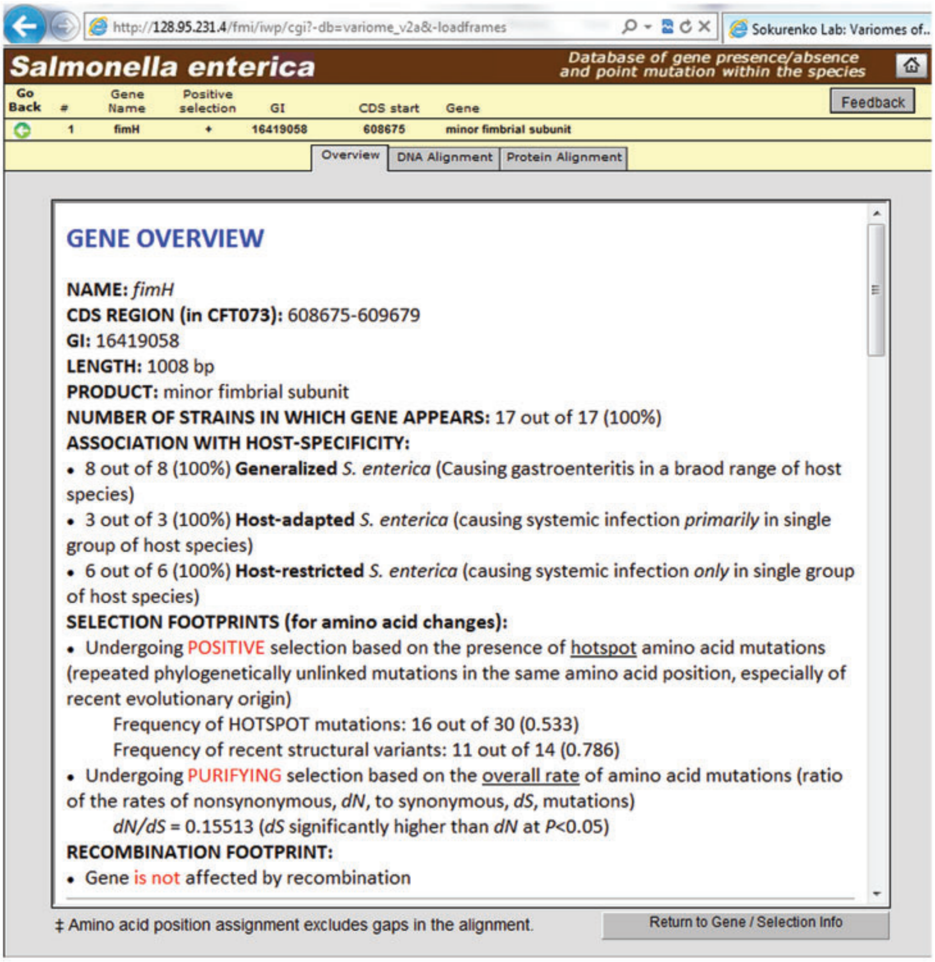
Now sorted in descending order by: Hotspot Mutations: Freq

Fig. 3. Snapshot of *Salmonella* variome "Strain Distribution" layout, showing sorted list of genes in descending order of the frequency of hotspot mutations.

Future Directions

This work presents the first attempt to create microbial variome databases. Besides maintaining the existing database, we plan to periodically update the databases to make them more user-friendly and more informative, based on both user feedback and availability of newly sequenced genomes for these species. With each inclusion of additional data, we need to perform all phylogenetic analyses anew via our newly developed software package TimeZone (Chattopadhyay et al. 2013), followed by the incorporation of new results in the database. As a next step, we also conceptualize the development of a cross-species microbial variome database of closely related species (e.g., *E. coli* and *S. enterica* subspecies *enterica*, or *E. coli* and *Klebsiella pneumoniae*) commonly inhabiting similar host compartments as human microbiota. This potentially could enlighten us about the variety of selection pressures in response to host–pathogen, pathogen–pathogen, and pathogen–commensal interactions.

We believe that the establishment of microbial variome databases signals a paradigm shift in microbial genomics, both quantitatively and qualitatively. These tools can be used to retrieve and analyze detailed information of all genetic variations in a given sequence, their population dynamics, and, most importantly, the action of various types of selection pressures. Along with information on strains' source of isolation, clonal identity, and pathotype, as well as the position of potentially adaptive changes along the affected proteins, this database will offer broad applicability of population genomics tools to experimental research, clinical diagnostics, epidemiology, and environmental control of pathogens. Overall, this knowledge will aid the detection of genotype/phenotype associations and deepen our understanding of bacterial evolution. Our planned layout of microbial variome database (fig. 5,


http://128.95.231.4/fmi/iwp/cgi?-db=variome_v2a&-loadframes
Sokurenko Lab: Variomes of..
Feedback
Overview
DNA Alignment
Protein Alignment
Return to Gene / Selection Info

Salmonella enterica

Database of gene presence/absence and point mutation within the species

Go Back	#	Gene Name	Positive selection	GI	CDS start	Gene
⟲	1	fimH	+	16419058	608675	minor fimbrial subunit

GENE OVERVIEW

NAME: *fimH*
CDS REGION (in CFT073): 608675-609679
GI: 16419058
LENGTH: 1008 bp
PRODUCT: minor fimbrial subunit
NUMBER OF STRAINS IN WHICH GENE APPEARS: 17 out of 17 (100%)
ASSOCIATION WITH HOST-SPECIFICITY:

- 8 out of 8 (100%) **Generalized** *S. enterica* (Causing gastroenteritis in a braod range of host species)
- 3 out of 3 (100%) **Host-adapted** *S. enterica* (causing systemic infection *primarily* in single group of host species)
- 6 out of 6 (100%) **Host-restricted** *S. enterica* (causing systemic infection *only* in single group of host species)

SELECTION FOOTPRINTS (for amino acid changes):

- Undergoing POSITIVE selection based on the presence of hotspot amino acid mutations (repeated phylogenetically unlinked mutations in the same amino acid position, especially of recent evolutionary origin)
  Frequency of HOTSPOT mutations: 16 out of 30 (0.533)
  Frequency of recent structural variants: 11 out of 14 (0.786)
- Undergoing PURIFYING selection based on the overall rate of amino acid mutations (ratio of the rates of nonsynonymous, *dN*, to synonymous, *dS*, mutations)
  dN/*dS* = 0.15513 (*dS* significantly higher than *dN* at P<0.05)

RECOMBINATION FOOTPRINT:

- Gene is not affected by recombination

‡ Amino acid position assignment excludes gaps in the alignment.

Fig. 4. Snapshot of a portion of "Gene Overview" showing the analysis results of *Salmonella fimH* gene.

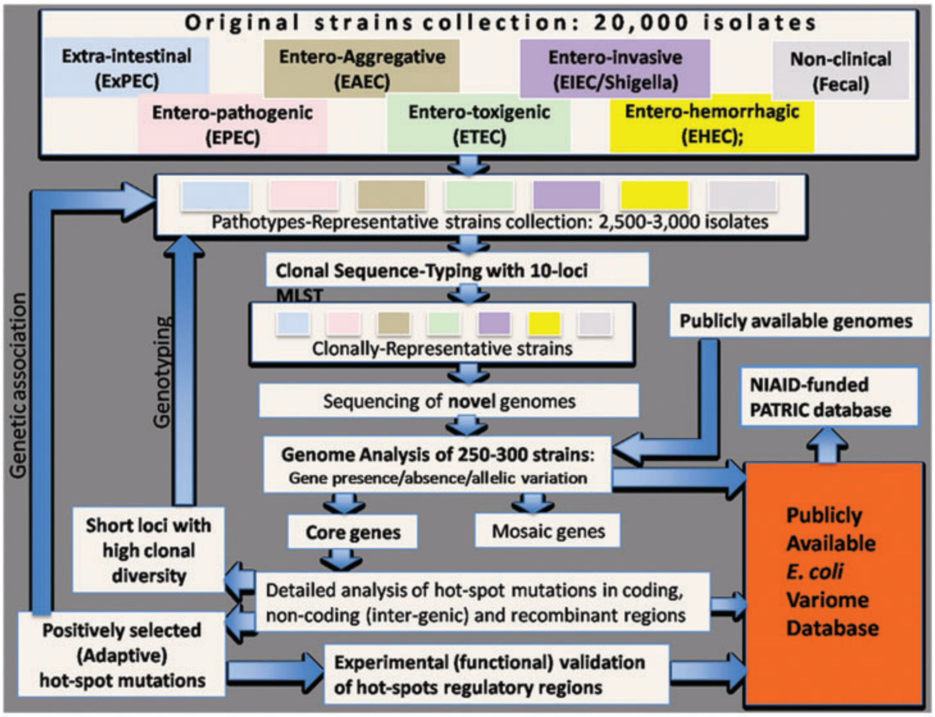

Original strains collection: 20,000 isolates
Extra-intestinal (ExPEC)
Entero-Aggregative (EAEC)
Entero-invasive (EIEC/Shigella)
Non-clinical (Fecal)
Entero-pathogenic (EPEC)
Entero-toxigenic (ETEC)
Entero-hemorrhagic (EHEC);
Pathotypes-Representative strains collection: 2,500-3,000 isolates
Clonal Sequence-Typing with 10-loci MLST
Clonally-Representative strains
Publicly available genomes
NIAID-funded PATRIC database
Sequencing of novel genomes
Genome Analysis of 250-300 strains: Gene presence/absence/allelic variation
Genetic association
Genotyping
Short loci with high clonal diversity
Core genes
Mosaic genes
Detailed analysis of hot-spot mutations in coding, non-coding (inter-genic) and recombinant regions
Positively selected (Adaptive) hot-spot mutations
Experimental (functional) validation of hot-spots regulatory regions
Publicly Available E. coli Variome Database

Fig. 5. Layout plan for *Escherichia coli* variome database.

using *E. coli* as an example) will on the one hand provide information on potential targets for vaccines, antibiotics, and other therapeutic development, while on the other hand, will enable a global surveillance system that can identify newly emerging or re-emerging pathogenic clones and the genetic mechanisms behind such emergence events.

Acknowledgments

This work was supported by the National Institute of Health (NIH) grants R01 GM084318 and RC4 AI092828.

Supplementary Material

Supplementary tables S1 and S2 are available at *Molecular Biology and Evolution* online (http://www.mbe.oxfordjournals.org/).

References

Anisimova M, Liberles DA. 2007. The quest for natural selection in the age of comparative genomics. *Heredity (Edinb)* 99:567–579.

Brinkman FS, Parkhill J. 2008. Population genomics: modeling the new and a renaissance of the old. *Curr Opin Microbiol.* 11:439–441.

Brookfield JF. 2005. The ecology of the genome—mobile DNA elements and their hosts. *Nat Rev Genet.* 6:128–136.

Bruen TC, Philippe H, Bryant D. 2006. A simple and robust statistical test for detecting the presence of recombination. *Genetics* 172: 2665–2681.

Butlin RK. 2010. Population genomics and speciation. *Genetica* 138: 409–418.

Chattopadhyay S, Dykhuizen DE, Sokurenko EV. 2007. ZPS: visualization of recent adaptive evolution of proteins. *BMC Bioinformatics* 8: 187.

Chattopadhyay S, Paul S, Dykhuizen DE, Sokurenko EV. 2013. Tracking recent adaptive evolution in microbial species using TimeZone. *Nat Protoc.* 8:652–665.

Chattopadhyay S, Paul S, Kisiela DI, Linardopoulou EV, Sokurenko EV. 2012. Convergent molecular evolution of genomic cores in *Salmonella enterica* and *Escherichia coli*. *J Bacteriol.* 194:5002–5011.

Chattopadhyay S, Weissman SJ, Minin VN, Russo TA, Dykhuizen DE, Sokurenko EV. 2009. High frequency of hotspot mutations in core genes of *Escherichia coli* due to short-term positive selection. *Proc Natl Acad Sci U S A.* 106:12412–12417.

Codd EF. 1970. A relational model of data for large shared data banks. *Commun ACM.* 13:377–387.

Conrad DF, Hurles ME. 2007. The population genetics of structural variation. *Nat Genet.* 39:S30–S36.

DeLong EF. 2004. Microbial population genomics and ecology: the road ahead. *Environ Microbiol.* 6:875–878.

Feil EJ. 2004. Small change: keeping pace with microevolution. *Nat Rev Microbiol.* 2:483–495.

Fournier PE, Drancourt M, Raoult D. 2007. Bacterial genome sequencing and its use in infectious diseases. *Lancet Infect Dis.* 7: 711–723.

Henriques-Normark B, Blomberg C, Dagerhamn J, Battig P, Normark S. 2008. The rise and fall of bacterial clones: *Streptococcus pneumoniae*. *Nat Rev Microbiol.* 6:827–837.

Kassen R, Rainey PB. 2004. The ecology and genetics of microbial diversity. *Annu Rev Microbiol.* 58:207–231.

Keeling PJ, Palmer JD. 2008. Horizontal gene transfer in eukaryotic evolution. *Nat Rev Genet.* 9:605–618.

Kisiela DI, Chattopadhyay S, Libby SJ, et al. (14 co-authors). 2012. Evolution of *Salmonella enterica* virulence via point mutations in the fimbrial adhesin. *PLoS Pathog.* 8:e1002733.

Knight CG, Zitzmann N, Prabhakar S, Antrobus R, Dwek R, Hebestreit H, Rainey PB. 2006. Unraveling adaptive evolution: how a single point mutation affects the protein coregulation network. *Nat Genet.* 38: 1015–1022.

Lefebure T, Stanhope MJ. 2007. Evolution of the core and pan-genome of Streptococcus: positive selection, recombination, and genome composition. *Genome Biol.* 8:R71.

Mes TH. 2008. Microbial diversity—insights from population genetics. *Environ Microbiol.* 10:251–264.

Morjan CL, Rieseberg LH. 2004. How species evolve collectively: implications of gene flow and selection for the spread of advantageous alleles. *Mol Ecol.* 13:1341–1356.

Orr MR, Smith TB. 1998. Ecology and speciation. *Trends Ecol Evol.* 13: 502–506.

Pallen MJ, Wren BW. 2007. Bacterial pathogenomics. *Nature* 449: 835–842.

Perron GG, Gonzalez A, Buckling A. 2008. The rate of environmental change drives adaptation to an antibiotic sink. *J Evol Biol.* 21: 1724–1731.

Petersen L, Bollback JP, Dimmic M, Hubisz M, Nielsen R. 2007. Genes under positive selection in *Escherichia coli*. *Genome Res.* 17: 1336–1343.

Rocha EP. 2008. Evolutionary patterns in prokaryotic genomes. *Curr Opin Microbiol.* 11:454–460.

Sokurenko EV, Feldgarden M, Trintchina E, Weissman SJ, Avagyan S, Chattopadhyay S, Johnson JR, Dykhuizen DE. 2004. Selection footprint in the FimH adhesin shows pathoadaptive niche differentiation in *Escherichia coli*. *Mol Biol Evol.* 21:1373–1383.

Sokurenko EV, Gomulkiewicz R, Dykhuizen DE. 2006. Source-sink dynamics of virulence evolution. *Nat Rev Microbiol.* 4:548–555.

Soyer Y, Orsi RH, Rodriguez-Rivera LD, Sun Q, Wiedmann M. 2009. Genome wide evolutionary analyses reveal serotype specific patterns of positive selection in selected *Salmonella* serotypes. *BMC Evol Biol.* 9:264.

Stinchcombe JR, Hoekstra HE. 2008. Combining population genomics and quantitative genetics: finding the genes underlying ecologically important traits. *Heredity (Edinb)* 100:158–170.

Whitaker RJ, Banfield JF. 2006. Population genomics in natural microbial communities. *Trends Ecol Evol.* 21:508–516.