

0.1 Title Page

SuperPhy:

Predictive genomics for the bacterial pathogen *Escherichia coli*

Matthew D Whiteside^{1†}, Chad R Laing^{1*†}, Akiff Manji¹, Peter Kruczkiewicz¹, Eduardo N Taboada¹ and Victor PJ Gannon¹

p MDW: matthew.whiteside@phac-aspc.gc.ca

CRL: chad.r.laing@phac-aspc.gc.ca

AM: akiff.manji@gmail.com

PK: peter.kruczkiewicz@gmail.com

ENT: eduardo.taboada@phac-aspc.gc.ca

VPJG: vic.gannon@phac-aspc.gc.ca

*Correspondence:

chad.r.laing@phac-aspc.gc.ca

¹Laboratory for Foodborne

Zoonoses, Public Health Agency
of Canada, Twp Rd 9-1, T1J 3Z4
Lethbridge, Canada

Full list of author information is
available at the end of the article

[†]Equal contributor

Abstract

Background: Predictive genomics is the translation of raw genome sequence data into an assessment of the phenotypes exhibited by the organism. For bacterial pathogens, these phenotypes can range from environmental survivability, to the severity of human disease associated with them. Significant progress has been made in the development of generic tools for genomic analyses that are broadly applicable to all microorganisms; however, a fundamental missing component is the ability to analyze genomic data in the context of organism-specific phenotypic knowledge, which has been accumulated from decades of research and can provide a meaningful interpretation of genome sequence data.

Results: In this study, we present SuperPhy, an online predictive genomics platform (<http://lfz.corefacility.ca/superphy/>) for *Escherichia coli*. The platform integrates the analyses tools and genome sequence data for all publicly available *E. coli* genomes and facilitates the upload of new genome sequences from users under public or private settings. SuperPhy provides real-time analyses of thousands of genome sequences with results that are understandable and useful to a wide community, including those in the fields of clinical medicine, epidemiology, ecology, and evolution. SuperPhy includes identification of: 1) virulence and antimicrobial resistance determinants 2) statistical associations between genotypes, biomarkers, geospatial distribution, host, source, and phylogenetic clade; 3) the identification of biomarkers for groups of genomes on the basis of presence / absence of specific genomic regions and single-nucleotide polymorphisms and 4) *in silico* Shiga-toxin subtype.

Conclusions: SuperPhy is a predictive genomics platform that attempts to provide an essential link between the vast amounts of genome information currently being generated and phenotypic knowledge in an organism-specific context.

Keywords: genomics; bioinformatics; epidemiology

Background

Whole-genome sequencing (WGS) of bacterial isolates generates the complete DNA sequence of each organism. WGS provides the greatest possible resolution of any typing method, the sequence is easily transferable, and its analyses can reveal important phenotypic insights such as the presence of virulence factors or anti-microbial resistance determinants. Current benchtop sequencers such as the Illumina MiSeq and the newly developed USB-sized Oxford Nanopore sequencer have made it possible for real-time WGS to be performed in the laboratory as well as on the front-line,

¹as was recently seen in the 2014 Ebola outbreak, and in managing a hospital out-¹
²break of *Salmonella* [1, 2, 3, 4].²

³ WGS will likely replace current typing and sub-typing methods due to its low³
⁴cost, high information content, portability, and speed of analyses. It is now being⁴
⁵used in real-time for: the identification of the source of foodborne outbreaks [5],⁵
⁶surveillance [6, 7], epidemiological investigations [7], industrial applications [8, 9],⁶
⁷population studies [10, 11], routine typing [12], regulation [13], providing point-of-⁷
⁸care insight for clinicians [14, 15], informing veterinary practice [16], and helping⁸
⁹inform public-health decisions [17].⁹

¹⁰ WGS is now the *de facto* standard for bacterial strain analyses and the global¹⁰
¹¹community is coming together to help store and best utilize this rapid in-¹¹
¹²flux of information under the Global Microbial Identifier network ([http://www.](http://www.globalmicrobialidentifier.org/)¹²
¹³[globalmicrobialidentifier.org/](http://www.globalmicrobialidentifier.org/)). This international effort currently involves¹³
¹⁴32 countries, many of which have their own national or regional programs to¹⁴
¹⁵best utilize WGS data in public health, epidemiological and research contexts,¹⁵
¹⁶such as the GenomeTrakR initiative of the Food and Drug Administration in the¹⁶
¹⁷United States of America ([http://www.fda.gov/Food/FoodScienceResearch/](http://www.fda.gov/Food/FoodScienceResearch/WholeGenomeSequencingProgramWGS/)¹⁷
¹⁸[WholeGenomeSequencingProgramWGS/](http://www.fda.gov/Food/FoodScienceResearch/WholeGenomeSequencingProgramWGS/)), the Integrated Rapid Infectious Disease¹⁸
¹⁹Analysis (IRIDA) platform in Canada (<http://www.irida.ca/>), and the Patho-¹⁹
²⁰NGen-Trace project within the European Union ([http://patho-ngen-trace.eu/](http://patho-ngen-trace.eu/project/)²⁰
²¹[project/](http://patho-ngen-trace.eu/project/)).²¹

²² Recently, several platforms have emerged that attempt to provide additional con-²²
²³text in addition to the raw WGS data. For instance PATRIC provides pre-computed²³
²⁴analyses for public genomes, including annotation, protein families, antibiotic re-²⁴
²⁵sistance identification and comparative pathway analysis [18]. MicroScope provides²⁵
²⁶an expert-guided annotation pipeline, as well as comparative analyses based on²⁶
²⁷shared gene content [19]. The Integrated Microbial Genomes (IMG) project is also²⁷
²⁸a combined genome annotation and analysis platform, that additionally allows for²⁸
²⁹genomic data submissions by the user [20]. BIGSdb allows local comparisons among²⁹
³⁰genomes using a multi-locus sequence typing approach, and allows phenotypic data³⁰
³¹to be stored along with the genomic information [21]. The Harvest suite of tools³¹
³²allows for fast core-genome alignments and interactive visualizations for thousands³²
³³of genomes [22]. Other platforms focus on a specific organism, such as Sybil, a plat-³³

¹form for the comparative analyses of *Streptococcus pneumoniae* based on BLASTP¹
²searches [23].²

³ The large initiatives that generate and collect the tens- and hundreds-of thousands³
⁴of genome sequences, and the platforms that host and analyze the public data pro-⁴
⁵vide an enormous benefit. Even though WGS and basic comparative analyses is⁵
⁶commonplace, meaningful interpretation of the raw data in a phenotypic context,⁶
⁷also known as predictive genomics, lags considerably behind [24]. Microbiologists of-⁷
⁸ten have organism-specific knowledge that can meaningfully inform the WGS data,⁸
⁹but which is not incorporated into a generic analysis. The ability to interactively⁹
¹⁰explore species-specific data that contains organism-specific knowledge from experts¹⁰
¹¹in the field is of tremendous value. A recent study on outbreak investigations using¹¹
¹²WGS also listed a main obstacle of routine adoption as ‘a paucity of user-friendly¹²
¹³and clinically focused bioinformatics platforms’ [25]. While some components nec-¹³
¹⁴essary for phenotypic prediction based on WGS data have been developed, there is¹⁴
¹⁵currently no single integrated platform built to provide predictive genomic analyses¹⁵
¹⁶for organism-specific end-users.¹⁶

¹⁷ Here we present SuperPhy, a predictive genomics platform that brings organism-¹⁷
¹⁸specific knowledge to comparative genomic analyses. SuperPhy incorporates knowl-¹⁸
¹⁹edge from research on the pathogenesis and epidemiology of *E. coli*, as well as the¹⁹
²⁰tremendous amount of genotypic and phenotypic data that have previously been²⁰
²¹generated. This knowledge is used within SuperPhy to discover relationships among²¹
²²and about sub-groups. It allows non-bioinformaticians to quickly analyze new data²²
²³against the background of other sequenced *E. coli*, facilitating novel insights.²³

²⁴ We have previously developed Panseq, software that performs comparative ge-²⁴
²⁵nomics in a pan-genome context, identifying differences in the accessory genome and²⁵
²⁶single nucleotide variations within the core genome [26]. SuperPhy utilizes the pan-²⁶
²⁷genomic output from Panseq to identify: 1) virulence and antimicrobial resistance²⁷
²⁸determinants 2) epidemiological associations between specific genotypes, biomark-²⁸
²⁹ers, geospatial distribution, host, source, and other metadata in an interactive and²⁹
³⁰explorable setting; 3) statistically significant clade-specific genome markers (pres-³⁰
³¹ence / absence of specific genomic regions, and single-nucleotide polymorphisms)³¹
³²for bacterial populations; and 4) *in silico* Shiga-toxin subtyping for genomes that³²
³³possess *stx* genes.³³

SuperPhy allows the submission of genomes in a private or public context and is continually updated with the influx of public *E. coli* data from GenBank, allowing researchers to quickly analyze and compare new genomes with other publicly available sequenced *E. coli* strains. Predictive genomics provides an essential link between the vast numbers of genomes currently being generated and organism-specific phenotypic knowledge.

1 Platform Features

1.1 Navigation and Overview

The layout of the SuperPhy website (<https://lfz.corefacility.ca/superphy>) provides universal and quick access to the major components of the platform: ‘Group Analyses’ provides an interactive environment for comparing groups of strains based on metadata types or user-created strain-groupings, and determining statistically significant biomarkers (both the presence / absence of genomic regions and SNPs) for these groups; ‘VF and AMR’ provides an ontology of both virulence genes and AMR determinants, and the ability to select groups of genomes and factors based on the provided ontologies. Output includes a summary of the presence / absence of selected VF and AMR factors among the strains of interest; ‘Group Browse’ provides an interface to examine groups of strains, and their distribution in both a geospatial and phylogenetic context simultaneously; ‘My Data’ provides an interface for uploading and modifying user-submitted genomes that are available only to the user; ‘Home’ provides a landing page and an overview of the features of the site. Additionally, an in-depth examination and report on an individual strain, including all known metadata, Shiga-toxin subtype (if applicable), phylogenetic and geospatial information, and a summary of virulence factor and anti-microbial resistance determinants can be accessed by selecting ‘detailed information’ from any genome in the platform.

1.2 Strain Selection

SuperPhy provides three methods of selecting *E. coli* genomes for analyses, that are consistent across the site: list-, tree-, and map-based selections. The platform is based heavily on metadata, and as such provides a unified metadata control panel that displays the metadata fields and their associated values for each genome

¹across each of the three views. The metadata control panel also allows filtering and¹
²selecting genomes that match given metadata criteria. ²

³ 1) List-based selection provides a table-based interface to the genomes and their³
⁴metadata, with private and public genome sets afforded their own sections. ⁴

⁵ 2) Tree-based selection provides an interactive phylogeny that can be manipulated⁵
⁶to expand / contract clades, and from which clade and individual genome selection⁶
⁷can be made. Metadata is appended to each leaf node of the tree, and branches⁷
⁸containing more than one genome have the metadata for the entire branch sum-⁸
⁹marized as an interactive bar-chart that displays the frequency of values within⁹
¹⁰selected metadata categories. This summary is an excellent way to visually discern¹⁰
¹¹clade differences, and allows an effective representation of thousands of genomes in¹¹
¹²tree form that would otherwise be intractable. An example of the phylogenetic tree¹²
¹³with metadata clusters is shown in Figure 1. ¹³

¹⁴ 3) Map-based selection provides a Google Maps interface to geospatial genome¹⁴
¹⁵selection, along with a table-view of the metadata for the genomes in the map.¹⁵
¹⁶Just as in the list-based view, the displayed metadata fields for each genome can¹⁶
¹⁷be changed, and used to filter the displayed genomes. As an example, we show the¹⁷
¹⁸map when a user searches for ‘United Kingdom’ in Figure 2. ¹⁸

¹⁹ ¹⁹

²⁰1.3 Website Usage Tutorials ²⁰

²¹Every page of the SuperPhy platform includes a guided tutorial introduction using²¹
²²the IntroJS plugin (<https://usablica.github.io/intro.js/>). This tutorial pro-²²
²³vides a walk-through of all the major features and how to use them, and is activated²³
²⁴by clicking the large red ‘Introduction’ button located on each page. ²⁴

²⁵2 Implementation ²⁵

²⁶2.1 Webserver Application and Database ²⁶

²⁷Genome data and analyses are administered using a PostgreSQL 9.3 database with²⁷
²⁸a schema adapted from the Generic Model Organism Database (GMOD) Chado²⁸
²⁹schema [27]. The Chado relational database schema uses a flexible, ontology-centric²⁹
³⁰approach to organizing biological entities, relationships, properties and analyses.³⁰
³¹Entries in generic tables are assigned types using a mutable, controlled vocabulary.³¹
³²By not defining entity types directly into the relational layer, the database can be³²
³³highly adaptable and can grow to add new analyses or biological data. ³³

¹ The application layer for the SuperPhy website is build using the Model-View-¹
²Controller (MVC) Perl CGI::Application framework (<http://www.cgi-app.org/>).²
³The phylogenetic tree display and interaction is built on top of the Data Driven Doc-³
⁴uments (D3) JavaScript library (<http://d3js.org/>). Geospatial views are built us-⁴
⁵ing the Google Maps JavaScript API v3 ([https://developers.google.com/maps/](https://developers.google.com/maps/documentation/javascript/)⁵
⁶[documentation/javascript/](https://developers.google.com/maps/documentation/javascript/)). Group comparisons are processed and displayed us-⁶
⁷ing the RStudio Shiny web application framework for R [28].⁷

⁸ The webserver application code base, database schema and public data are hosted⁸
⁹on Github at <https://github.com/superphy/version-1>.⁹

¹²2.1.1 Access to Uploaded Data¹²

¹³Users can upload genomes and metadata and choose between three access levels to¹³
¹⁴govern their use: ‘public’ information is available to all users; ‘private’ information¹⁴
¹⁵is only available for the genome uploader and additional users they select; and¹⁵
¹⁶‘private until a specified date’ data is released to ‘public’ data after a specified¹⁶
¹⁷date. Users may also designate other registered users for whom the data will be¹⁷
¹⁸available. Private data is accessible only to designated users, but can be combined¹⁸
¹⁹with public data for user-specific analyses. Users can create custom genome-groups¹⁹
²⁰that can be saved, and all results may be downloaded for offline analyses.²⁰

²¹ Uploaded data undergo a series of checks to ensure the quality of the data. Data²¹
²²are rejected if any of the following conditions are met: 1) Greater than 1000 con-²²
²³tigs; 2) Genome size less than 3 Mbp or greater than 7.5 Mbp; 3) Invalid nucleotide²³
²⁴characters (all IUPAC characters are valid); 4) The MD5 checksum of the concate-²⁴
²⁵nated contigs already exists in the database; 5) The SNP string for the pan-genome²⁵
²⁶alignment is identical to another strain in the database.²⁶

²⁷ Uploaded genomes undergo two checks to ensure the data are of a minimum²⁷
²⁸quality, and that the genomes being uploaded belong to the species *E. coli*. We²⁸
²⁹initially identified genomic regions present in at least 70% of the genomes, referred²⁹
³⁰to as the ‘conserved core’. All genomes are considered to be *E. coli* if: 1) they³⁰
³¹contain at least 1500 conserved core regions, and 2) The presence of at least three³¹
³²*E. coli* species-specific regions. The derivation of these markers is presented in the³²
³³‘Pan-genome’ subsection of the ‘Example analyses’.³³

2.2 Acquisition of public *Escherichia coli* genomes

SuperPhy is continually and automatically updated with closed and draft genomes of *Escherichia coli* from GenBank using the script https://github.com/superphy/version-1/Sequences/ncbi_downloader.pl. All metadata present in the GenBank submissions are extracted automatically using the script https://github.com/superphy/version-1/Sequences/genbank_to_genodo.pl. For the initial bulk upload, a second phase of manual curation was carried out to ensure all available metadata was included, even if it was stored in a non-standard way during the initial submission. The complete list of 1641 public *E. coli* genomes present in the SuperPhy database at the time of manuscript preparation, along with all extracted metadata is available at (https://github.com/superphy/version-1/Data/metadata_table.csv). A summary of the metadata fields used in SuperPhy, as well as the percentage of the public genomes containing information for a particular metadata category is presented in Table 1.

2.3 Comparative Genomic Analyses

Our pan-genomic analyses tool, Panseq is used for the background comparative analyses [26]. It iteratively adds new genomic sequences, and compares them to those already stored in the platform. This computational approach allows a continuous influx of new sequence data without large time or memory requirements. In this way, the complete pan-genome of all sequences in the database is determined. Annotations for these regions are determined by querying the GenBank NR protein database via BLASTx.

Differences in the accessory genome and the single nucleotide variation in the core genome are obtained and used by SuperPhy in downstream applications including the construction of discriminatory and robust phylogenies, and in the pre-computed data for bio-marker identification among groups of genomes.

2.3.1 Tree Construction

SuperPhy provides a dynamic maximum-likelihood phylogenetic tree that is continuously updated to include all *E. coli* genomes currently in the database, as likelihood approaches to phylogenetic reconstruction have been shown to be superior to distance and parsimony approaches [29]. An initial phylogenetic tree for SuperPhy was constructed using conserved genomic regions from the 1641 *E. coli* genomes

¹obtained from GenBank. The conserved regions were aligned using Muscle [30, 31]¹
²and input into FastTreeMP to build a minimum-evolution tree [32]. To achieve suf-²
³ficient resolution in branch lengths to disambiguate strains, the double-precision³
⁴version of FastTree was used [32]. As new genomes are uploaded to SuperPhy, they⁴
⁵are incorporated into the multiple sequence alignment and a new tree is rebuilt,⁵
⁶which becomes the tree used for all analyses within the SuperPhy platform.⁶

⁷2.3.2 Virulence and Anti-microbial Resistance Markers⁸

⁹The presence / absence of virulence and AMR genes are computed using Panseq,⁹
¹⁰The non-redundant query set of AMR genes from the Comprehensive Antibiotic¹⁰
¹¹Resistance Database (CARD) [33] is used for *in silico* AMR determinant screen-¹¹
¹²ing. All AMR genes are organized and stored in the database according to their¹²
¹³CARD-assigned Antibiotic Resistance Ontology annotation to aid in identifying¹³
¹⁴the presence of different antimicrobial resistance mechanisms . The virulence gene¹⁴
¹⁵database was constructed by obtaining all gene alleles of known virulence factors¹⁵
¹⁶for *E. coli* from the Virulence Factor Database [34], supplemented with additional¹⁶
¹⁷virulence factors from ‘*Escherichia coli*: Pathotypes and Principles of Pathogen-¹⁷
¹⁸esis, 2nd Ed.’ [35], and additional published literature, which effectively doubled¹⁸
¹⁹the number of virulence factors in the database. To avoid duplication of factors,¹⁹
²⁰all AMR and virulence factor sequences were clustered based on similarity using²⁰
²¹BLASTclust with default settings; the longest allele was selected for each gene, ex-²¹
²²cept in cases where sequence similarity was less than 90%, in which case multiple²²
²³alleles were included [36].²³

²⁴In addition to providing the presence / absence of virulence and AMR factors, Su-²⁴
²⁵perPhy stores the sequence of the individual alleles for each genome, and constructs²⁵
²⁶a phylogeny based on each single gene. This allows one to compare the relationships²⁶
²⁷among genomes based on a single virulence or AMR attribute and to examine the²⁷
²⁸sequence variation of the gene at the individual base level, as the multiple sequence²⁸
²⁹alignment (MSA) can also be displayed, as shown in Figure 3²⁹

³⁰2.3.3 Group Comparisons³¹

³¹The statistical identification of markers that differ between groups based on both³¹
³²single nucleotide polymorphisms and the presence / absence of genomic loci is imple-³²
³³mented using a two stage approach: 1) The ‘approximate’ vectorized Fisher’s Exact³³

¹Test (FET) from the R corpora package is calculated ([http://cran.r-project.](http://cran.r-project.org/web/packages/corpora/index.html)¹
²[org/web/packages/corpora/index.html](http://cran.r-project.org/web/packages/corpora/index.html)), and the 100 most-significant results are²
³then subject to the FET from the base R statistical package [37]. All single-³
⁴nucleotide polymorphisms and genomic presence / absence data reside in the⁴
⁵database, and require only the retrieval and P-value computation for the strains⁵
⁶of interest for the real time analysis of genome markers.⁶

⁷ The R Shiny interface is used for group creation and all metadata fields are pre-⁷
⁸populated for all strains in the database. This makes comparing, for example, all⁸
⁹human and non-human strains of a given serotype as simple as selecting groups⁹
¹⁰based on the serotype and host metadata fields, and clicking the compare button.¹⁰
¹¹Additionally, custom groups of any genomes can be created and saved to a user-¹¹
¹²profile so they become available whenever the user is logged in. These custom groups¹²
¹³can include private genomes available only to the logged-in user, in addition to any¹³
¹⁴public genomes.¹⁴

¹⁵ 2.4 Stx Typing ¹⁶

¹⁷Shiga-toxin (Stx) subtype assignment, when a strain possesses a copy of one or¹⁷
¹⁸more of *stx1* or *stx2*, is calculated based on a phylogenetic tree generated from¹⁸
¹⁹concatenated and aligned a and b subunits for each of Stx1 and Stx2. Clades specific¹⁹
²⁰to a Shiga-toxin subtype were identified based on the scheme presented by Scheutz²⁰
²¹et al. (2012) [38]. Membership in these pre-defined clades is used to identify the²¹
²²subtype of the toxin gene; those strains that fall outside of known sub-type clades²²
²³are marked as unknown. Multiple sequence alignments of the Stx genes are stored²³
²⁴in the database for reference and comparison.²⁴

²⁵ 2.5 Geospatial Visualization ²⁶

²⁶The geospatial visualizations provide an interactive map interface for selecting and²⁶
²⁷and searching genomes and groups of genomes. SuperPhy leverages Google Maps²⁷
²⁸along with the companion Javascript library, Google Maps API (V3).²⁸

²⁹Genome location data is geocoded for latitude and longitude during the process²⁹
³⁰of adding a new strain to the platform. To reduce the computational overhead in³⁰
³¹rendering thousands of genome map markers, the marker clustering algorithm Mark-³¹
³²erClusterPlus for Google Maps V3 [http://google-maps-utility-library-v3.](http://google-maps-utility-library-v3.googlecode.com/svn/trunk/markerclustererplus/docs/reference.html)³²
³³[googlecode.com/svn/trunk/markerclustererplus/docs/reference.html](http://google-maps-utility-library-v3.googlecode.com/svn/trunk/markerclustererplus/docs/reference.html) was³³

¹implemented. Locations within a distance of 60 pixels on the map are clustered¹
²into a single marker rendered at the geometric center of the cluster, and a count of²
³the number of genomes is displayed. 3

⁴ All geospatial views are accompanied by a dynamic and sortable table of genome⁴
⁵metadata that is by default sorted by country. Users also have the option of sorting⁵
⁶by province, state and city. The table is dynamic and updates to display informa-⁶
⁷tion for the genomes visible on the map. Locations for each *E. coli* strain can be⁷
⁸downloaded for offline manipulation. 8

⁹ 9

¹⁰2.6 Continuous Integration 10

¹¹The user community is able to provide feedback as the platform evolves in the form¹¹
¹²of feature requests and bug reports using the ‘Issues’ section at [https://github.](https://github.com/superphy/version-1/issues)¹²
¹³[com/superphy/version-1/issues](https://github.com/superphy/version-1/issues). This will ensure the platform evolves in a way¹³
¹⁴that is most beneficial to those who use it. 14

¹⁵3 Results and Discussion 15

¹⁶3.1 Pan-genome 16

¹⁷At the time of writing, 2324 publicly available *E. coli* genomes from GenBank¹⁷
¹⁸had been analyzed for incorporation into the SuperPhy platform [39]. *E. coli* is¹⁸
¹⁹a ubiquitous, gram-negative bacterial species found in the intestines of healthy¹⁹
²⁰mammals, with only a small subset causing disease in humans or animals [40]. The²⁰
²¹population structure of *E. coli* was initially described as being broadly distributed²¹
²²among four large and two smaller phylogenetic groups [41, 42]. Recent studies have²²
²³found that the species has an open pan-genome, meaning that the addition of new²³
²⁴genomes is likely to add additional genes to the pool [43]. The pan-genome of *E. coli*²⁴
²⁵is highly variable, with around 80% of an individual genome comprised of accessory²⁵
²⁶genes and the remainder from the shared core genome [44]; a stable proportion of²⁶
²⁷approximately 4000 genes are present in at least 50% of the genomes [45]. 27

²⁸The pan-genome distribution of these 2324 *E. coli* genomes as 1000 bp genomic²⁸
²⁹segments is presented in Figure 4. As can be seen, the majority (29.7 Mbp) of the²⁹
³⁰37.44 Mbp pan-genome is present in fewer than 100 genomes, with the core genome³⁰
³¹size (present in at least 2300 genomes) observed to be 1.86 Mbp. Only 5.84 Mbp³¹
³²of the pan-genome was found in greater than 100 genomes, but fewer than 2300³²
³³genomes. Based on these results, we selected a ‘conserved core’ of 3598 genomic³³

¹regions, defined as those present in at least 70% of the 2324 genomes. The conserved¹
²core is used within SuperPhy to identify SNPs that are used in phylogenetic tree²
³building, as well as in the quality filtering of uploaded genomes. ³

⁴ Additionally, we endeavored to identify genomic regions that were specific to the⁴
⁵species *E. coli*. To achieve this we screened the ‘conserved core’ against genomes⁵
⁶from a subset of *E. coli* and other bacterial species, the results of which are presented⁶
⁷in Table 2. The *E. coli* genomes contained more of the ‘conserved core’ regions than⁷
⁸any of the other genomes examined, although genomes from *Shigella spp.* contained⁸
⁹nearly as many, which is not surprising given that *Shigella spp.* has long been known⁹
¹⁰to be very similar to *E. coli* [46]. Recent work using the analyses of whole genome¹⁰
¹¹sequence data of both *Shigella spp.* and *E. coli* showed *Shigella spp.* to form three¹¹
¹²separate monophyletic clades within the *E. coli* species [47], and that there was a¹²
¹³mixing of traditional *Shigella spp.* within these clades. The analyses performed in¹³
¹⁴this study to find *E. coli* specific regions treated *Shigella spp.* as distinct from *E.*¹⁴
¹⁵*coli*; had they been considered as sub-groups within *E. coli*, the number of species-¹⁵
¹⁶specific markers would likely have increased. ¹⁶

¹⁷ The results shown in Table 2 were filtered based on the distribution among these¹⁷
¹⁸19 genomes to identify genomic regions present in only the *E. coli* genomes, re-¹⁸
¹⁹sulting in 33 candidates; the raw data table is available at [https://github.](https://github.com/superphy/version-1/Sequences/genome_content_panseq/binary_table.txt)¹⁹
²⁰[com/superphy/version-1/Sequences/genome_content_panseq/binary_table.](https://github.com/superphy/version-1/Sequences/genome_content_panseq/binary_table.txt)²⁰
²¹[txt](https://github.com/superphy/version-1/Sequences/genome_content_panseq/binary_table.txt). These 33 candidates were screened against the GenBank ‘nr’ and ‘WGS’²¹
²²databases using the ‘bacteria’ taxid to limit the search; the raw BLAST re-²²
²³sults are available at [https://github.com/superphy/version-1/Sequences/](https://github.com/superphy/version-1/Sequences/genome_content_panseq/UBOHWGTR015-Alignment.xml)²³
²⁴[genome_content_panseq/UBOHWGTR015-Alignment.xml](https://github.com/superphy/version-1/Sequences/genome_content_panseq/UBOHWGTR015-Alignment.xml) and [https://github.](https://github.com/superphy/version-1/Sequences/genome_content_panseq/UD4GVA26015-Alignment.xml)²⁴
²⁵[com/superphy/version-1/Sequences/genome_content_panseq/UD4GVA26015-Alignment.](https://github.com/superphy/version-1/Sequences/genome_content_panseq/UD4GVA26015-Alignment.xml)²⁵
²⁶[xml](https://github.com/superphy/version-1/Sequences/genome_content_panseq/UD4GVA26015-Alignment.xml). Based on these queries using a 90% total sequence identity threshold, we re-²⁶
²⁷moved all putative species-specific regions that were identified in genomes from²⁷
²⁸bacteria other than *E. coli*, and were left with the ten species-specific regions pre-²⁸
²⁹sented in Table 3. ²⁹

³⁰ The correlation between the species-specific regions and the ‘conserved core’ re-³⁰
³¹gions among the 2324 *E. coli* genomes is presented in Figure 5. As can be seen, not³¹
³²all species-specific markers were found in all strains; however, most *E. coli* genomes³²
³³contained at least 8 of the markers and all contained at least 3 given the quality³³

checks for assembled genomes previously described. A general trend was observed¹ where genomes with higher ratios of ‘Genome size’ / ‘No. contigs’ contained both² more ‘conserved core’ regions and species-specific regions, indicating that the qual-³ity of genome assembly affects the number of genomic regions that can be identified⁴ at a given sequence identity threshold. Based on these results, any genome in the⁵ SuperPhy database is defined as *E. coli* if it possesses at least three of the species⁶ specific markers and at least 1500 of the conserved core genomic regions.⁷

Of the 2324 genomes examined, only 1641 had metadata beyond the name of the⁸ strain. As such, the initial SuperPhy database contained only these 1641 genomes⁹ to facilitate a metadata driven approach to genomic analysis.¹⁰

3.2 Predictive Markers for Sub-groups¹²

A ‘group’ of bacteria can be defined in numerous ways, from spatially or temporally¹³ co-located strains, to those sharing biochemical utilization patterns, or those that¹⁴ occupy a clade of a phylogenetic tree. Regardless of how a group is defined, users¹⁵ are generally interested in defining characteristics that are predictive of the group,¹⁶ and can be used to discriminate its members from those of other related genomes.¹⁷ SuperPhy utilizes both the presence / absence of genomic regions, and SNPs within¹⁸ shared regions to define markers statistically predictive of a group. These identified¹⁹ biomarkers have potential downstream application in *in silico* diagnostics or simple²⁰ wet-lab tests for the identified markers.²¹

As an example, we utilized the ‘Group Analyses’ feature of SuperPhy to identify²² SNPs that were statistically predictive for *E. coli* of serotype O157:H7 with respect²³ to those of all other *E. coli*. This is demonstrated in Figure 6, where the SNPs²⁴ are ranked from most- to least-significant. The marker ID for each SNP, the poly-²⁵morphism being examined, the p-value, the false discovery rate adjusted p-value,²⁶ and the presence / absence of each SNP for the two groups being examined are²⁷ displayed. The marker ID provides a link to a ‘SNP Information’ page (Figure 7),²⁸ which identifies the pan-genome region the SNP is found in, the allele frequency²⁹ of SNPs for all genomes in the database, the putative function of the region given³⁰ by the top BLAST hit, and an option to download detailed SNP information for³¹ each genome. The download includes the genomic location, allele, and upstream /³² downstream sequences for all genomes in the database.³³

¹ In addition to providing groups based on metadata categories such as serotype,¹
²and providing group vs. non-group comparisons, SuperPhy allows multi-way group²
³vs. group comparisons, as shown in the example of Figure 8, where ‘isolation host’³
⁴is selected and the categories ‘Bos taurus (cow)’, ‘Homo sapiens (human)’, and⁴
⁵‘Environmental source’ are used to generate comparisons between all combinations⁵
⁶of the categories. This facilitates more rapid identification of group and sub-group⁶
⁷predictive markers for the genomes being examined. 7

⁸3.3 Distribution of the *eae* gene 8

⁹Within the species *E. coli*, there are a subset of strains that attach to human intesti-¹⁰
¹¹nal epithelial cells via an attaching and effacing mechanism, the requisite apparatus¹¹
¹²for which is encoded in a genomic island known as the locus of enterocyte effacement¹²
¹³(LEE) [48]. As an example of the ‘VF and AMR’ functionality within SuperPhy, we¹³
¹⁴identified the distribution of the LEE gene *eae* among the 1641 public genomes in¹⁴
¹⁵the SuperPhy database. All virulence factors are stored using controlled ontologies,¹⁵
¹⁶which facilitate easy addition and retrieval of related data. The ontological category¹⁶
¹⁷‘LEE-encoded TTSS effector’ provided the *eae* alleles, and they were selected, along¹⁷
¹⁸with all 1641 public genomes. The results are presented in an interactive matrix of¹⁸
¹⁹gene presence / absence, as well as allele copy number (Figure 9). Within the 1641¹⁹
²⁰genomes examined, 662 possessed the *eae* gene. Additionally, SuperPhy provides a²⁰
²¹table of the results for download, where subsequent offline manipulation is possible.²¹

²²3.4 Analyses of Geographical and Phylogenetic Clusters 22

²³The ‘Group Browse’ section of SuperPhy provides a means for selecting, filtering and 23
²⁴exploring groups of genomes utilizing the three modes of genome selection, namely 24
²⁵the tree, map and list views. These allows users to view geographical clusters in 25
²⁶terms of their corresponding position in a phylogenetic tree. For example, using the 26
²⁷map view, and the hierarchical listing of locations, all genomes with the isolation 27
²⁸location of Santa Clara, California, United States were selected and their corre- 28
²⁹sponding positions on the phylogenetic tree automatically highlighted, as shown in 29
³⁰Figure 10. Here it is evident that although all six genomes were isolated from Santa 30
³¹Clara, California on the same day, the genomes do not form their own cluster on the 31
³²phylogenetic tree. On the tree, all nodes that contain a selected genome are shown 32
³³as blue-filled squares, while those that do not are white-filled squares. Similarly, all 33

¹selected genomes appear on the tree as blue-filled circles, and those not selected as ¹
²white-filled circles. All six selected genomes from Santa Clara are not visible on the ²
³tree at once, as they are not all closely related and the tree needed to be zoomed in ³
⁴for readability. Genomes CS02 and CS06 are both visible, on separate branches of ⁴
⁵the tree, indicating they are less related to each other, and the other four genomes ⁵
⁶from Santa Clara, than the genomes with which they group most closely. ⁶

⁷ This ability to quickly examine geographical strain clusters in a phylogenetic ⁷
⁸context would prove extremely useful in determining if a group of genomes from ⁸
⁹the same time and place originated from a single bacterial clone, as in an outbreak ⁹
¹⁰situation or in the routine surveillance of a location such as a food-processing plant, ¹⁰
¹¹to determine whether bacterial isolates were that of a persistent strain. ¹¹

¹² Conversely, within SuperPhy one can also select a phylogenetic clade and have ¹²
¹³the geographical locations of all strains shown. The ability to break apart a clus- ¹³
¹⁴ter of strains that are related at the genome level into geographical and metadata ¹⁴
¹⁵categories has use in source tracking of strains, and in determining the geographi- ¹⁵
¹⁶cal dissemination of bacterial clones over time. As an example, genomes from the ¹⁶
¹⁷serotype O104:H4 outbreak that occurred in Germany in 2011 were chosen. This ¹⁷
¹⁸outbreak was the first caused by strains of O104:H4 that were found to have acquired ¹⁸
¹⁹the *stx2* gene through lateral gene transfer, which is thought to have been the con- ¹⁹
²⁰tributing factor that led to the high rates of acute illness in healthy adults observed ²⁰
²¹throughout the outbreak [49]. As can be seen in Figure 11, the O104:H4 strains ²¹
²²containing the *stx2* gene are nearly identical on the phylogenetic tree; however, the ²²
²³source of isolation of these bacteria, visible on the map, shows the dissemination ²³
²⁴of the bacterial clone from the German epicenter to countries such as Denmark, ²⁴
²⁵the United Kingdom, Canada, and the United states, which were determined to be ²⁵
²⁶travel-acquired infections. ²⁶

²⁷ ²⁸ ²⁹4 Conclusions ²⁹

³⁰ Predictive genomics and platforms that easily facilitate it are poised to become ³⁰
³¹the translation layer between the vast amounts of sequence data and biological ³¹
³²knowledge in a specific domain that is needed to test hypotheses. SuperPhy allows ³²
³³users to make some of these genotype / phenotype correlations, and platforms like ³³

it will become increasingly important in transforming raw genome data into useful knowledge.

Current work involves the addition of previously published *in silico* serotyping schemes to SuperPhy, and the expansion of the platform to include the bacterial pathogens *Salmonella enterica* and *Campylobacter jejuni*. Lastly, a representational state transfer (REST) application programming interface (API) is being designed to allow programmatic interaction with the SuperPhy platform, which will help ensure that SuperPhy does not become a data silo but can instead contribute to a dynamic and growing web of biological knowledge.

Availability and Requirements

Project name: Superphy **Project home page:** <https://lfz.corefacility.ca/superphy> **Operating system(s):** Platform independent (modern web-browser; the most recent Firefox or Chrome for best experience) **Programming languages:** Perl, Coffeescript / Javascript, R **License:** Apache2

List of abbreviations

WGS: whole-genome sequencing **DNA:** deoxyribonucleic acid **GMOD:** generic model organism database **Stx:** Shiga-toxin **AMR:** anti-microbial resistance **CARD:** comprehensive antibiotic resistance database **SNP:** single-nucleotide polymorphism

Availability of supporting data

The project is entirely open source under the Apache 2 license <https://www.apache.org/licenses/LICENSE-2.0>. All code and any additional files referenced in the manuscript are available at the GitHub repository <https://github.com/superphy/version-1>.

Competing interests

The authors declare that they have no competing interests.

Authors' contributions

Designed the project: VPJG, CRL, MDW

Coded the platform: MDW, AM, JM, CRL, PK

Wrote the manuscript: CRL, MDW, AM, VPJG

Contributed ideas; read, edited, and approved the manuscript: MDW, CRL, AM, PK, ENT, VPJG

Acknowledgements

Thanks to Nicolas Tremblay for excellent metadata mining, and Omar Zabaneh, Peter Shen, Michael Benediktson, and Waqar Gill for contributing to early versions of this project. This work is funded in part by the Public Health Agency of Canada and a grant from the Genomics Research and Development Initiative.

Author details

¹Laboratory for Foodborne Zoonoses, Public Health Agency of Canada, Twp Rd 9-1, T1J 3Z4 Lethbridge, Canada.

²Laboratory for Foodborne Zoonoses, Public Health Agency of Canada, Twp Rd 9-1, T1J 3Z4 Lethbridge, Canada.

References

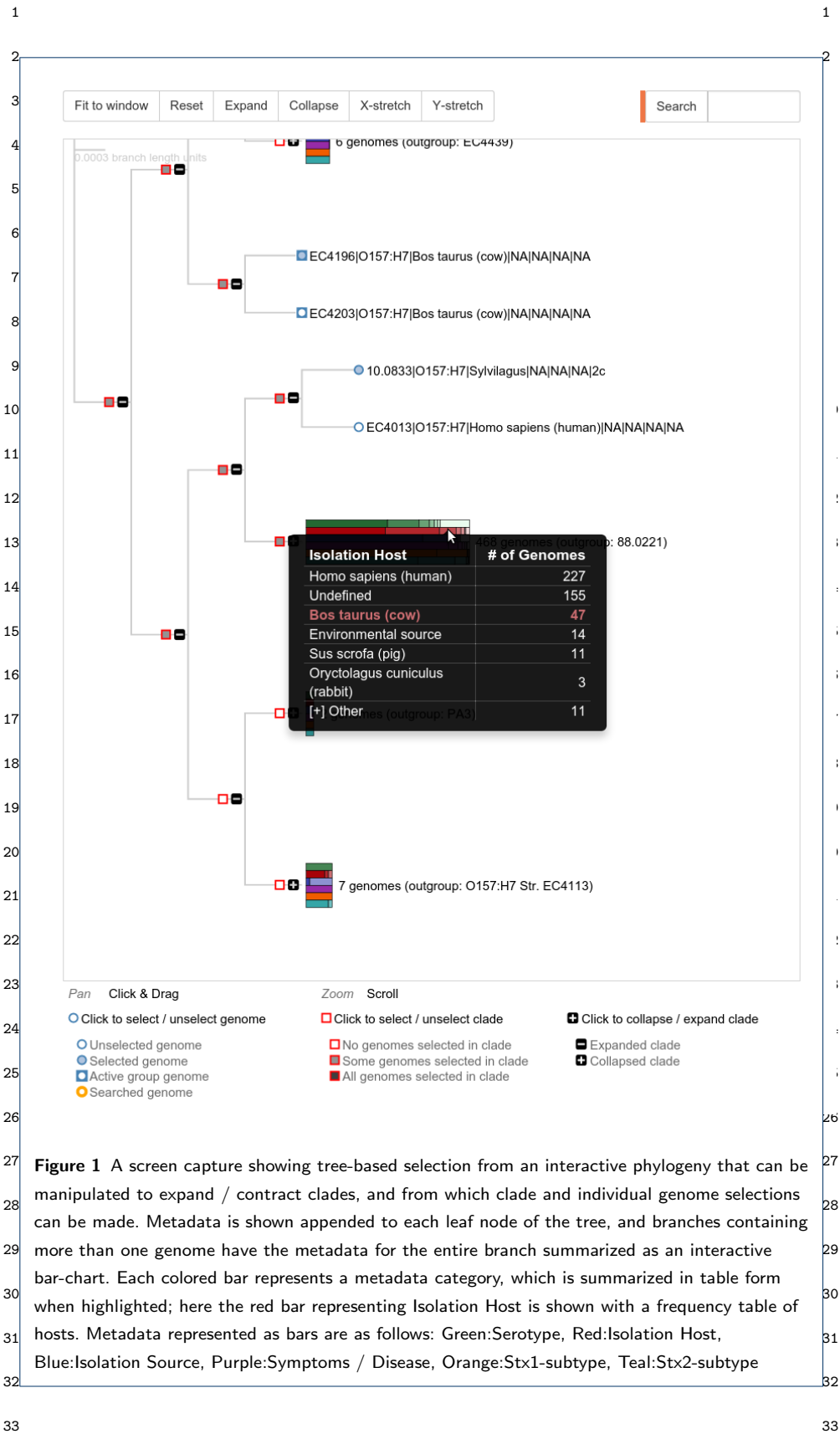
1. Jones, B.: Technology: Nanopore sequencing for clinical diagnostics. *Nature Reviews Genetics* **16**(2), 68–68 (2015). doi:[10.1038/nrg3895](https://doi.org/10.1038/nrg3895). Accessed 2015-05-27
2. Gilchrist, C.A., Turner, S.D., Riley, M.F., Petri, W.A., Hewlett, E.L.: Whole-Genome Sequencing in Outbreak Analysis. *Clinical Microbiology Reviews* **28**(3), 541–563 (2015). doi:[10.1128/CMR.00075-13](https://doi.org/10.1128/CMR.00075-13). Accessed 2015-05-27

- 1 3. in Biomedical Informatics at University of Birmingham, N.L.M.S.T.F.: How a small backpack for fast genomic
2 sequencing is helping combat Ebola. [http://theconversation.com/
3 how-a-small-backpack-for-fast-genomic-sequencing-is-helping-combat-ebola-41863](http://theconversation.com/how-a-small-backpack-for-fast-genomic-sequencing-is-helping-combat-ebola-41863) Accessed
4 2015-05-27
- 4 4. Quick, J., Ashton, P., Calus, S., Chatt, C., Gossain, S., Hawker, J., Nair, S., Neal, K., Nye, K., Peters, T.,
5 Pinna, E.D., Robinson, E., Struthers, K., Webber, M., Catto, A., Dallman, T.J., Hawkey, P., Loman, N.J.:
6 Rapid draft sequencing and real-time nanopore sequencing in a hospital outbreak of Salmonella. *Genome
7 Biology* **16**(1), 114 (2015). doi:[10.1186/s13059-015-0677-2](https://doi.org/10.1186/s13059-015-0677-2). Accessed 2015-06-01
- 6 5. Graham, R.M.A., Doyle, C.J., Jennison, A.V.: Real-time investigation of a *Legionella pneumophila* outbreak
7 using whole genome sequencing. *Epidemiology & Infection* **142**(11), 2347–2351 (2014).
8 doi:[10.1017/S0950268814000375](https://doi.org/10.1017/S0950268814000375). Accessed 2014-11-21
- 8 6. Zankari, E., Hasman, H., Kaas, R.S., Seyfarth, A.M., Agers, Y., Lund, O., Larsen, M.V., Aarestrup, F.M.:
9 Genotyping using whole-genome sequencing is a realistic alternative to surveillance based on phenotypic
10 antimicrobial susceptibility testing. *Journal of Antimicrobial Chemotherapy* **68**(4), 771–777 (2013). Accessed
11 2013-10-28
- 11 7. Cody, A.J., McCarthy, N.D., Rensburg, M.J.v., Isinkaye, T., Bentley, S., Parkhill, J., Dingle, K.E., Bowler,
12 I.C.J.W., Jolley, K.A., Maiden, M.C.J.: Real-time genomic epidemiology of human *Campylobacter* isolates
13 using whole genome multilocus sequence typing. *Journal of Clinical Microbiology*, 00066–13 (2013). Accessed
14 2013-10-28
- 13 8. Andreevskaya, M., Johansson, P., Laine, P., Smolander, O.-P., Sonck, M., Rahkila, R., Jskelinen, E., Paulin, L.,
14 Auvinen, P., Bjrkroth, J.: Genome Sequence and Transcriptome Analysis of Meat-Spoilage-Associated Lactic
15 Acid Bacterium *Lactococcus piscium* MKFS47. *Applied and Environmental Microbiology* **81**(11), 3800–3811
16 (2015). doi:[10.1128/AEM.00320-15](https://doi.org/10.1128/AEM.00320-15). Accessed 2015-05-27
- 16 9. Mazzaglia, A., Studholme, D.J., Taratufolo, M.C., Cai, R., Almeida, N.F., Goodman, T., Guttman, D.S.,
17 Vinatzer, B.A., Balestra, G.M.: *Pseudomonas syringae* pv. *actinidiae* (PSA) Isolates from Recent Bacterial
18 Canker of Kiwifruit Outbreaks Belong to the Same Genetic Lineage. *PLoS ONE* **7**(5), 36518 (2012).
19 doi:[10.1371/journal.pone.0036518](https://doi.org/10.1371/journal.pone.0036518). Accessed 2015-05-27
- 18 10. Nasser, W., Beres, S.B., Olsen, R.J., Dean, M.A., Rice, K.A., Long, S.W., Kristinsson, K.G., Gottfredsson, M.,
19 Vuopio, J., Raisanen, K., Caugant, D.A., Steinbakk, M., Low, D.E., McGeer, A., Darenberg, J.,
20 Henriques-Normark, B., Beneden, C.A.V., Hoffmann, S., Musser, J.M.: Evolutionary pathway to increased
21 virulence and epidemic group A *Streptococcus* disease derived from 3,615 genome sequences. *Proceedings of
22 the National Academy of Sciences* **111**(17), 1768–1776 (2014). doi:[10.1073/pnas.1403138111](https://doi.org/10.1073/pnas.1403138111). Accessed
23 2014-11-21
- 22 11. Kopac, S., Wang, Z., Wiedenbeck, J., Sherry, J., Wu, M., Cohan, F.M.: Genomic heterogeneity and ecological
23 speciation within one subspecies of *Bacillus subtilis*. *Applied and Environmental Microbiology*, 00576–14
24 (2014). doi:[10.1128/AEM.00576-14](https://doi.org/10.1128/AEM.00576-14). Accessed 2014-11-21
- 24 12. Zhang, S., Yin, Y., Jones, M.B., Zhang, Z., Kaiser, B.L.D., Dinsmore, B.A., Fitzgerald, C., Fields, P.I., Deng,
25 X.: Salmonella Serotype Determination Utilizing High-Throughput Genome Sequencing Data. *Journal of
26 Clinical Microbiology* **53**(5), 1685–1692 (2015). doi:[10.1128/JCM.00323-15](https://doi.org/10.1128/JCM.00323-15). Accessed 2015-05-22
- 26 13. Halachev, M.R., Chan, J.Z., Constantinidou, C.I., Cumley, N., Bradley, C., Smith-Banks, M., Oppenheim, B.,
27 Pallen, M.J.: Genomic epidemiology of a protracted hospital outbreak caused by multidrug-resistant
28 *Acinetobacter baumannii* in Birmingham, England. *Genome Medicine* **6**(11), 70 (2014).
29 doi:[10.1186/s13073-014-0070-x](https://doi.org/10.1186/s13073-014-0070-x). Accessed 2014-11-21
- 28 14. Grad, Y.H., Lipsitch, M.: Epidemiologic data and pathogen genome sequences: a powerful synergy for public
29 health. *Genome Biology* **15**(11), 538 (2014). doi:[10.1186/s13059-014-0538-4](https://doi.org/10.1186/s13059-014-0538-4). Accessed 2014-11-21
- 29 15. Jr, W.M.D., Westblade, L.F., Ford, B.: Next-generation and whole-genome sequencing in the diagnostic clinical
30 microbiology laboratory. *European Journal of Clinical Microbiology & Infectious Diseases* **31**(8), 1719–1726
31 (2012). doi:[10.1007/s10096-012-1641-7](https://doi.org/10.1007/s10096-012-1641-7). Accessed 2015-05-27
- 31 16. Biek, R., O'Hare, A., Wright, D., Mallon, T., McCormick, C., Orton, R.J., McDowell, S., Trewby, H., Skuce,
32 R.A., Kao, R.R.: Whole Genome Sequencing Reveals Local Transmission Patterns of *Mycobacterium bovis* in
33 Sympatric Cattle and Badger Populations. *PLoS Pathog* **8**(11), 1003008 (2012).

- doi:[10.1371/journal.ppat.1003008](https://doi.org/10.1371/journal.ppat.1003008). Accessed 2015-05-27
17. Lemke, A.A., Harris-Wai, J.N.: Stakeholder engagement in policy development: challenges and opportunities for human genomics. *Genetics in Medicine* (2015). doi:[10.1038/gim.2015.8](https://doi.org/10.1038/gim.2015.8). Accessed 2015-05-27
18. Wattam, A.R., Abraham, D., Dalay, O., Disz, T.L., Driscoll, T., Gabbard, J.L., Gillespie, J.J., Gough, R., Hix, D., Kenyon, R., Machi, D., Mao, C., Nordberg, E.K., Olson, R., Overbeek, R., Pusch, G.D., Shukla, M., Schulman, J., Stevens, R.L., Sullivan, D.E., Vonstein, V., Warren, A., Will, R., Wilson, M.J.C., Yoo, H.S., Zhang, C., Zhang, Y., Sobral, B.W.: PATRIC, the bacterial bioinformatics database and analysis resource. *Nucleic Acids Research* **42**(D1), 581–591 (2013). Accessed 2014-01-15
19. Vallenet, D., Belda, E., Calteau, A., Cruveiller, S., Engelen, S., Lajus, A., Le Fevre, F., Longin, C., Mornico, D., Roche, D., Rouy, Z., Salvignol, G., Scarpelli, C., Thil Smith, A.A., Weiman, M., Medigue, C.: MicroScope—an integrated microbial resource for the curation and comparative analysis of genomic and metabolic data. *Nucleic Acids Research* **41**(D1), 636–647 (2012). Accessed 2013-09-12
20. Markowitz, V.M., Chen, I.-M.A., Palaniappan, K., Chu, K., Szeto, E., Pillay, M., Ratner, A., Huang, J., Woyke, T., Huntemann, M., Anderson, I., Billis, K., Varghese, N., Mavromatis, K., Pati, A., Ivanova, N.N., Kyrpides, N.C.: IMG 4 version of the integrated microbial genomes comparative analysis system. *Nucleic Acids Research* **42**(D1), 560–567 (2013). Accessed 2014-01-15
21. Jolley, K.A., Maiden, M.C.: BIGSdb: Scalable analysis of bacterial genome variation at the population level. *BMC Bioinformatics* **11**(1), 595 (2010). doi:[10.1186/1471-2105-11-595](https://doi.org/10.1186/1471-2105-11-595). Accessed 2014-12-18
22. Treangen, T.J., Ondov, B.D., Koren, S., Phillippy, A.M.: Rapid core-genome alignment and visualization for thousands of microbial genomes. *bioRxiv*, 007351 (2014). doi:[10.1101/007351](https://doi.org/10.1101/007351). Though many microbial species or clades now have hundreds of sequenced genomes, existing whole-genome alignment methods do not efficiently handle comparisons on this scale. Here we present the Harvest suite of core-genome alignment and visualization tools for quickly analyzing thousands of intraspecific microbial strains. Harvest includes Parsnp, a fast core-genome multi-aligner, and Gingr, a dynamic visual platform. Combined they provide interactive core-genome alignments, variant calls, recombination detection, and phylogenetic trees. Using simulated and real data we demonstrate that our approach exhibits unrivaled speed while maintaining the accuracy of existing methods. The Harvest suite is open-source and freely available from: <http://github.com/marbl/harvest>. Accessed 2014-11-20
23. Riley, D.R., Angiuoli, S.V., Crabtree, J., Hotopp, J.C.D., Tettelin, H.: Using Sybil for interactive comparative genomics of microbes on the web. *Bioinformatics* **28**(2), 160–166 (2012). Accessed 2013-09-12
24. Fricke, W.F., Rasko, D.A.: Bacterial genome sequencing in the clinic: bioinformatic challenges and solutions. *Nature Reviews Genetics* **15**(1), 49–55 (2014). doi:[10.1038/nrg3624](https://doi.org/10.1038/nrg3624). Accessed 2015-05-27
25. Sherry, N.L., Porter, J.L., Seemann, T., Watkins, A., Stinear, T.P., Howden, B.P.: Outbreak investigation using high-throughput genome sequencing within a diagnostic microbiology laboratory. *Journal of Clinical Microbiology* **51**(5), 1396–1401 (2013). Accessed 2013-10-28
26. Laing, C., Buchanan, C., Taboada, E.N., Zhang, Y., Kropinski, A., Villegas, A., Thomas, J.E., Gannon, V.P.J.: Pan-genome sequence analysis using panseq: an online tool for the rapid analysis of core and accessory genomic regions. *BMC Bioinformatics* **11**, 461 (2010). Accessed 2011-01-18
27. Mungall, C.J., Emmert, D.B.: A Chado case study: an ontology-based modular schema for representing genome-associated biological information. *Bioinformatics* **23**(13), 337–346 (2007). Accessed 2014-01-15
28. Racine, J.S.: RStudio: A platform-independent IDE for R and Sweave. *Journal of Applied Econometrics* **27**(1), 167–172 (2012). Accessed 2014-11-14
29. Guindon, S., Dufayard, J.-F., Lefort, V., Anisimova, M., Hordijk, W., Gascuel, O.: New algorithms and methods to estimate maximum-likelihood phylogenies: assessing the performance of PhyML 3.0. *Systematic Biology* **59**(3), 307–321 (2010). doi:[10.1093/sysbio/syq010](https://doi.org/10.1093/sysbio/syq010). Accessed 2011-07-08
30. Edgar, R.C.: MUSCLE: multiple sequence alignment with high accuracy and high throughput. *Nucleic Acids Res.* **32**(5), 1792–1797 (2004)
31. Edgar, R.C.: MUSCLE: a multiple sequence alignment method with reduced time and space complexity. *BMC Bioinformatics* **5**, 113 (2004)
32. Price, M.N., Dehal, P.S., Arkin, A.P.: FastTree 2 approximately maximum-likelihood trees for large alignments. *PLoS ONE* **5**(3), 9490 (2010). Accessed 2013-09-13

33. McArthur, A.G., Wagelchner, N., Nizam, F., Yan, A., Azad, M.A., Baylay, A.J., Bhullar, K., Canova, M.J., Pascale, G.D., Ejim, L., Kalan, L., King, A.M., Koteva, K., Morar, M., Mulvey, M.R., O'Brien, J.S., Pawlowski, A.C., Piddock, L.J.V., Spanogiannopoulos, P., Sutherland, A.D., Tang, I., Taylor, P.L., Thaker, M., Wang, W., Yan, M., Yu, T., Wright, G.D.: The comprehensive antibiotic resistance database. *Antimicrobial Agents and Chemotherapy* **57**(7), 3348–3357 (2013). Accessed 2014-01-17
34. Chen, L., Xiong, Z., Sun, L., Yang, J., Jin, Q.: VFDB 2012 update: toward the genetic diversity and molecular evolution of bacterial virulence factors. *Nucleic Acids Research* **40**(D1), 641–645 (2011). Accessed 2014-01-17
35. Donnenberg, M.: *Escherichia Coli: Pathotypes and Principles of Pathogenesis*. Academic Press, ??? (2013)
36. Altschul, S.F., Madden, T.L., Schffer, A.A., Zhang, J., Zhang, Z., Miller, W., Lipman, D.J.: Gapped BLAST and PSI-BLAST: a new generation of protein database search programs. *Nucleic Acids Research* **25**(17), 3389–402 (1997). doi:[PMC146917](https://doi.org/10.1093/nar/25.17.3389). Accessed 2008-10-29
37. for Statistical Computing, R.F.: R: A Language and Environment for Statistical Computing. R Development Core Team, Vienna, Austria (2005)
38. Scheutz, F., Teel, L.D., Beutin, L., Pirard, D., Buvens, G., Karch, H., Mellmann, A., Caprioli, A., Tozzoli, R., Morabito, S., Strockbine, N.A., Melton-Celsa, A.R., Sanchez, M., Persson, S., O'Brien, A.D.: Multicenter evaluation of a sequence-based protocol for subtyping Shiga toxins and standardizing Stx nomenclature. *Journal of clinical microbiology* **50**(9), 2951–2963 (2012). doi:[10.1128/JCM.00860-12](https://doi.org/10.1128/JCM.00860-12)
39. Benson, D.A., Cavanaugh, M., Clark, K., Karsch-Mizrachi, I., Lipman, D.J., Ostell, J., Sayers, E.W.: GenBank. *Nucleic Acids Research*, 1195 (2012). Accessed 2014-11-14
40. Tenaillon, O., Skurnik, D., Picard, B., Denamur, E.: The population genetics of commensal *Escherichia coli*. *Nature Reviews. Microbiology* **8**(3), 207–217 (2010). doi:[10.1038/nrmicro2298](https://doi.org/10.1038/nrmicro2298). Accessed 2011-07-20
41. Selander, R.K., Caugant, D.A., Ochman, H., Musser, J.M., Gilmour, M.N., Whittam, T.S.: Methods of multilocus enzyme electrophoresis for bacterial population genetics and systematics. *Appl Environ Microbiol* **51**(5), 873–84 (1986)
42. Goulet, P., Picard, B.: Comparative electrophoretic polymorphism of esterases and other enzymes in *Escherichia coli*. *Journal of General Microbiology* **135**(1), 135–143 (1989)
43. Medini, D., Donati, C., Tettelin, H., Maignani, V., Rappuoli, R.: The microbial pan-genome. *Current Opinion in Genetics & Development* **15**(6), 589–594 (2005). doi:[10.1016/j.gde.2005.09.006](https://doi.org/10.1016/j.gde.2005.09.006). Accessed 2009-05-21
44. Lukjancenko, O., Wassenaar, T.M., Ussery, D.W.: Comparison of 61 sequenced *Escherichia coli* genomes. *Microbial Ecology* **60**(4), 708–720 (2010). doi:[10.1007/s00248-010-9717-3](https://doi.org/10.1007/s00248-010-9717-3). Accessed 2011-03-23
45. Gordienko, E.N., Kazanov, M.D., Gelfand, M.S.: Evolution of pan-genomes of *Escherichia coli*, *Shigella spp.*, and *Salmonella enterica*. *Journal of Bacteriology* **195**(12), 2786–2792 (2013). doi:[10.1128/JB.02285-12](https://doi.org/10.1128/JB.02285-12). Accessed 2014-11-18
46. Pupo, G.M., Lan, R., Reeves, P.R.: Multiple independent origins of Shigella clones of *Escherichia coli* and convergent evolution of many of their characteristics. *Proc Natl Acad Sci U S A* **97**(19), 10567–72 (2000)
47. Sahl, J.W., Morris, C.R., Emberger, J., Fraser, C.M., Ochieng, J.B., Juma, J., Fields, B., Breiman, R.F., Gilmour, M., Nataro, J.P., Rasko, D.A.: Defining the Phylogenomics of Shigella Species: a Pathway to Diagnostics. *Journal of Clinical Microbiology* **53**(3), 951–960 (2015). doi:[10.1128/JCM.03527-14](https://doi.org/10.1128/JCM.03527-14). Accessed 2015-07-28
48. Croxen, M.A., Law, R.J., Scholz, R., Keeney, K.M., Wlodarska, M., Finlay, B.B.: Recent advances in understanding enteric pathogenic *Escherichia coli*. *Clinical Microbiology Reviews* **26**(4), 822–880 (2013). doi:[10.1128/CMR.00022-13](https://doi.org/10.1128/CMR.00022-13). Accessed 2014-11-21
49. Mellmann, A., Harmsen, D., Cummings, C.A., Zentz, E.B., Leopold, S.R., Rico, A., Prior, K., Szczepanowski, R., Ji, Y., Zhang, W., McLaughlin, S.F., Henkhaus, J.K., Leopold, B., Bielaszewska, M., Prager, R., Brzoska, P.M., Moore, R.L., Guenther, S., Rothberg, J.M., Karch, H.: Prospective Genomic Characterization of the German Enterohemorrhagic *Escherichia coli* O104:H4 Outbreak by Rapid Next Generation Sequencing Technology. *PLoS ONE* **6**(7), 22751 (2011). doi:[10.1371/journal.pone.0022751](https://doi.org/10.1371/journal.pone.0022751). Accessed 2011-07-22

Figures



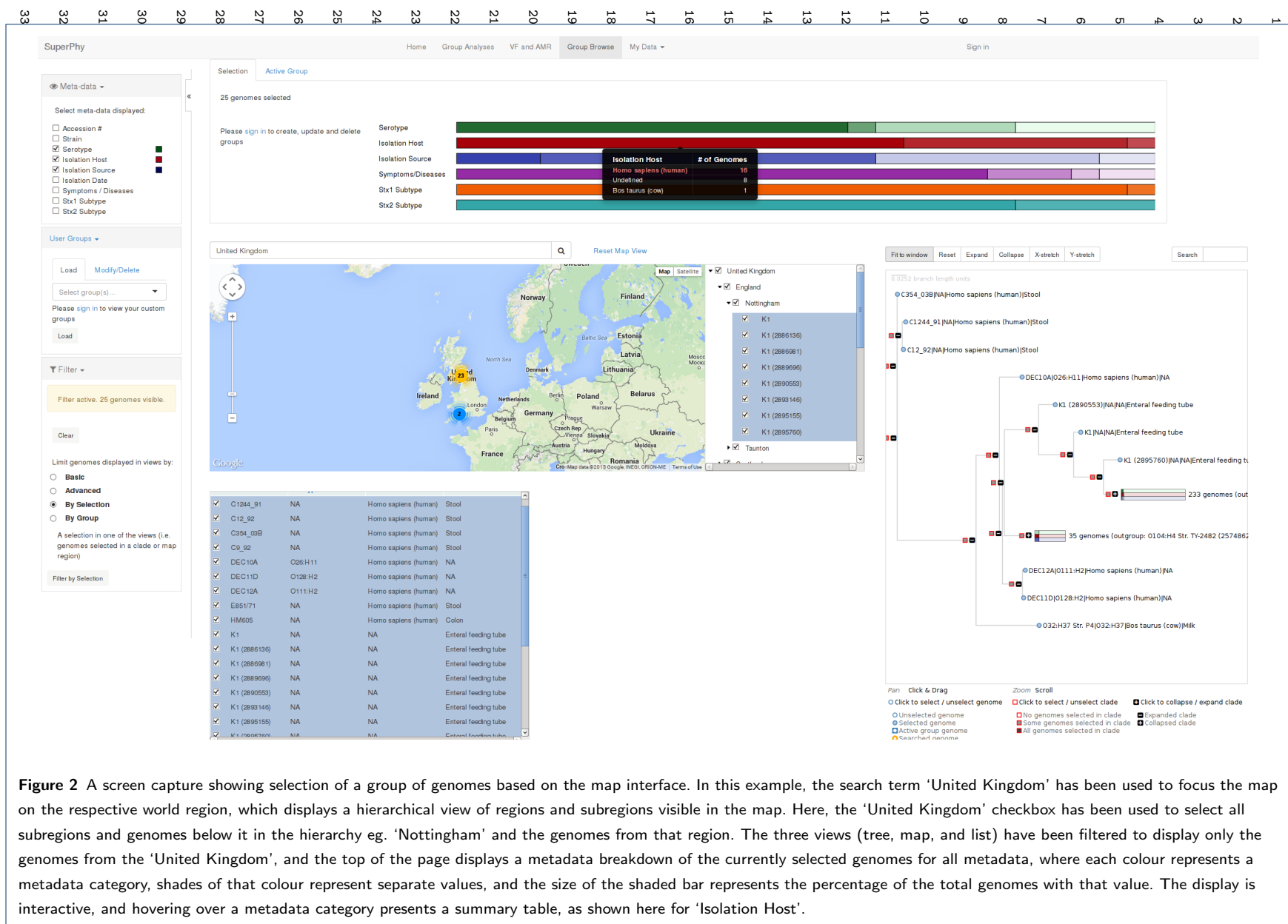
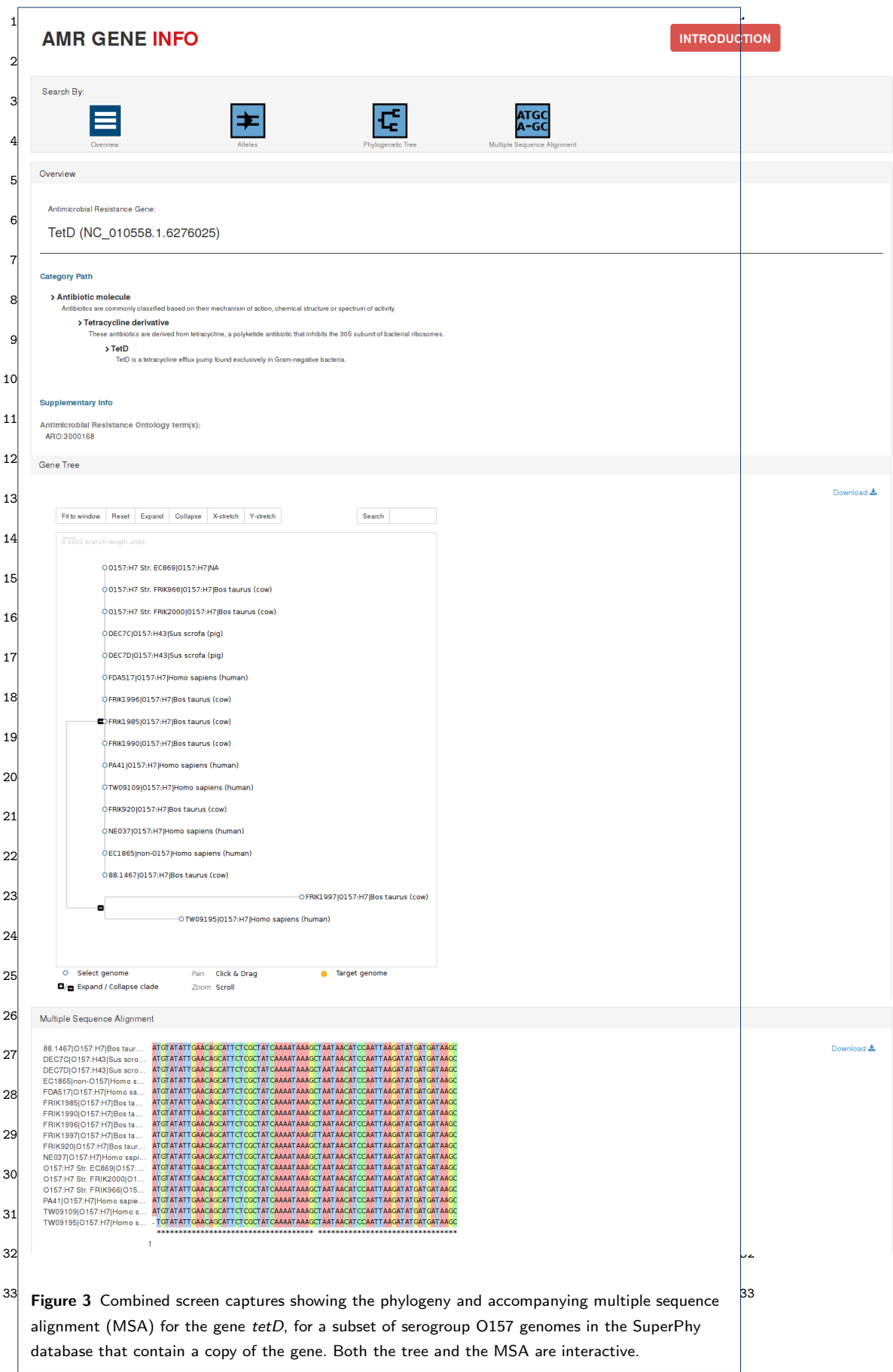
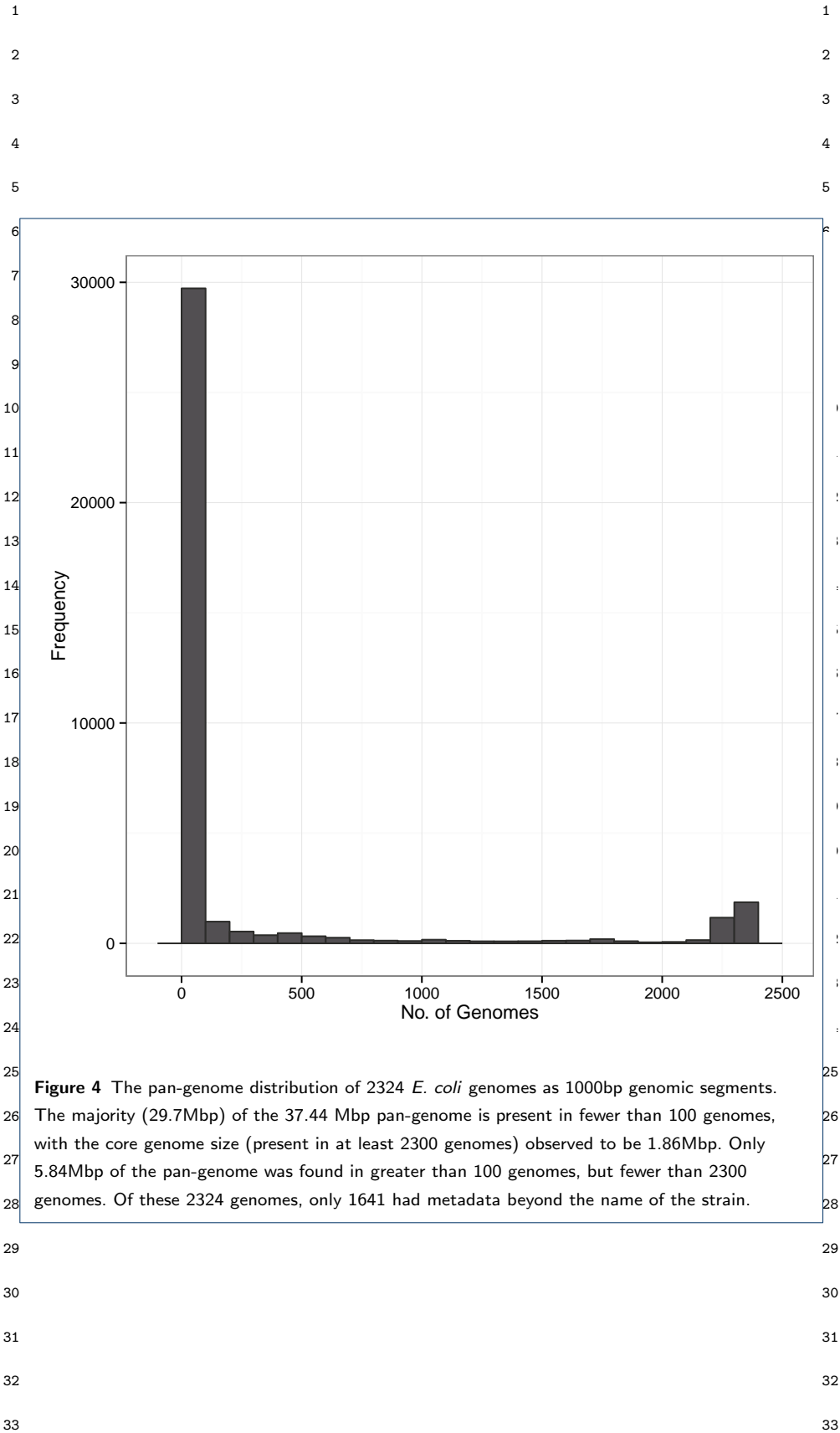
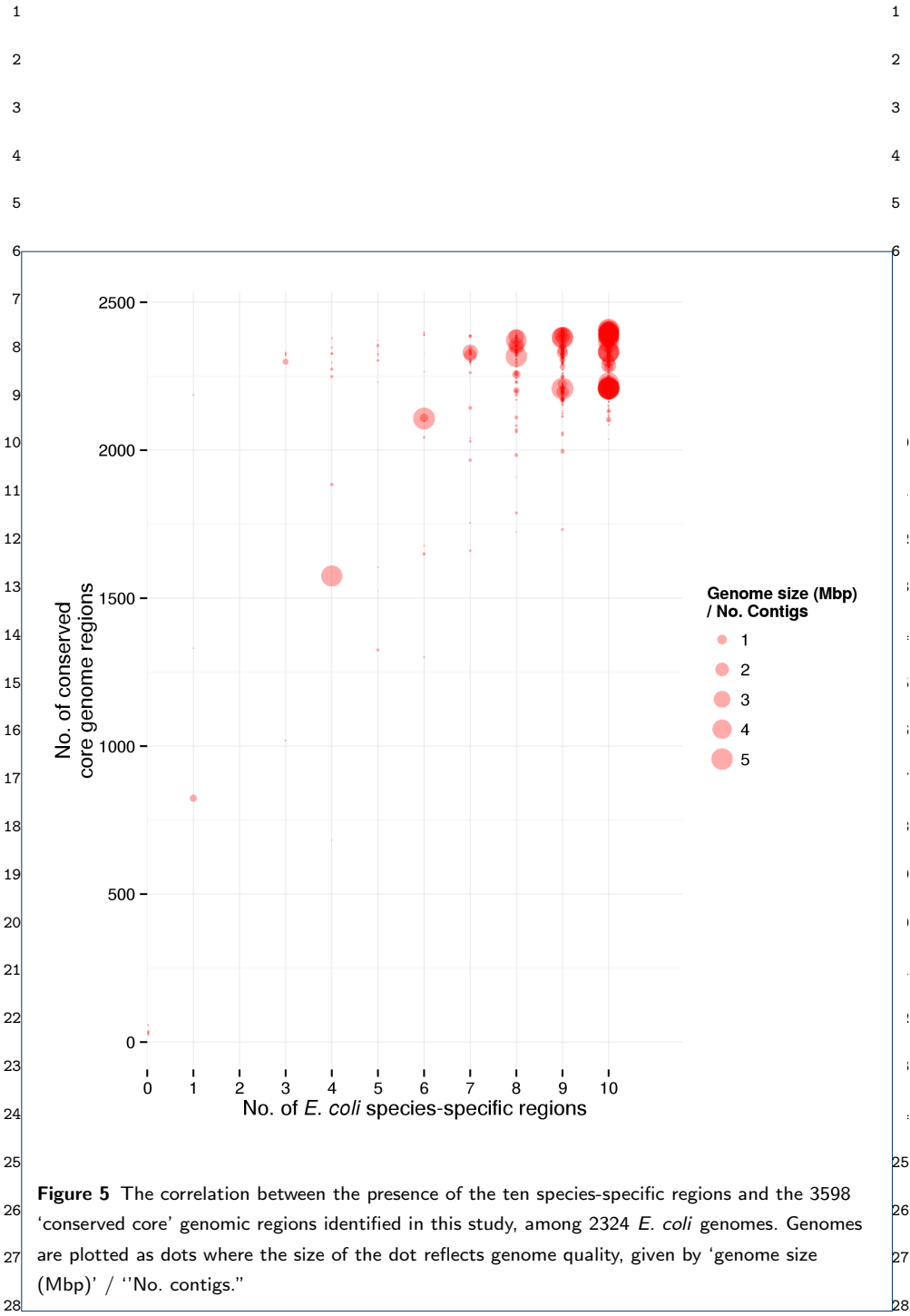


Figure 2 A screen capture showing selection of a group of genomes based on the map interface. In this example, the search term 'United Kingdom' has been used to focus the map on the respective world region, which displays a hierarchical view of regions and subregions visible in the map. Here, the 'United Kingdom' checkbox has been used to select all subregions and genomes below it in the hierarchy eg. 'Nottingham' and the genomes from that region. The three views (tree, map, and list) have been filtered to display only the genomes from the 'United Kingdom', and the top of the page displays a metadata breakdown of the currently selected genomes for all metadata, where each colour represents a metadata category, shades of that colour represent separate values, and the size of the shaded bar represents the percentage of the total genomes with that value. The display is interactive, and hovering over a metadata category presents a summary table, as shown here for 'Isolation Host'.







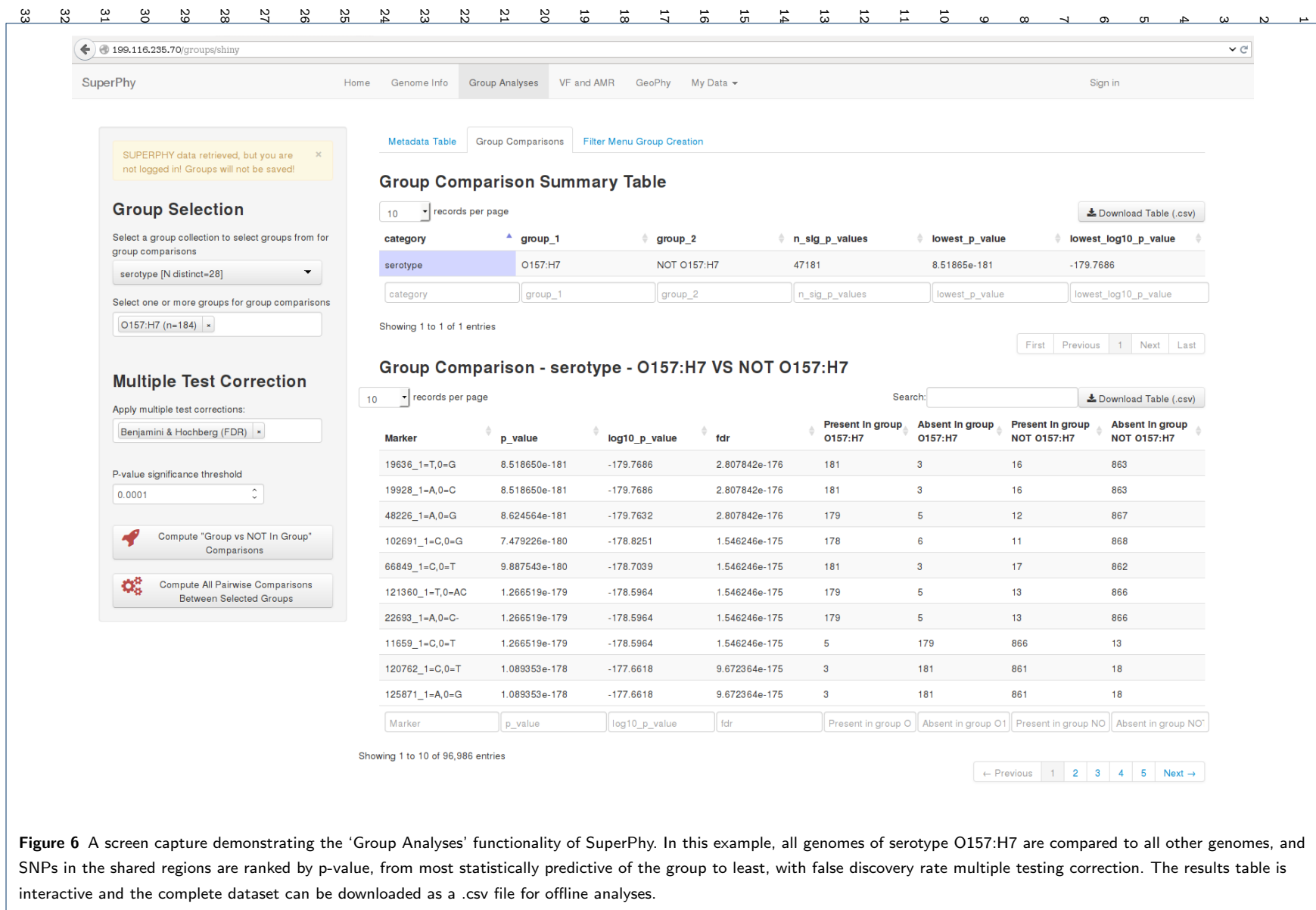


Figure 6 A screen capture demonstrating the ‘Group Analyses’ functionality of SuperPhy. In this example, all genomes of serotype O157:H7 are compared to all other genomes, and SNPs in the shared regions are ranked by p-value, from most statistically predictive of the group to least, with false discovery rate multiple testing correction. The results table is interactive and the complete dataset can be downloaded as a .csv file for offline analyses.

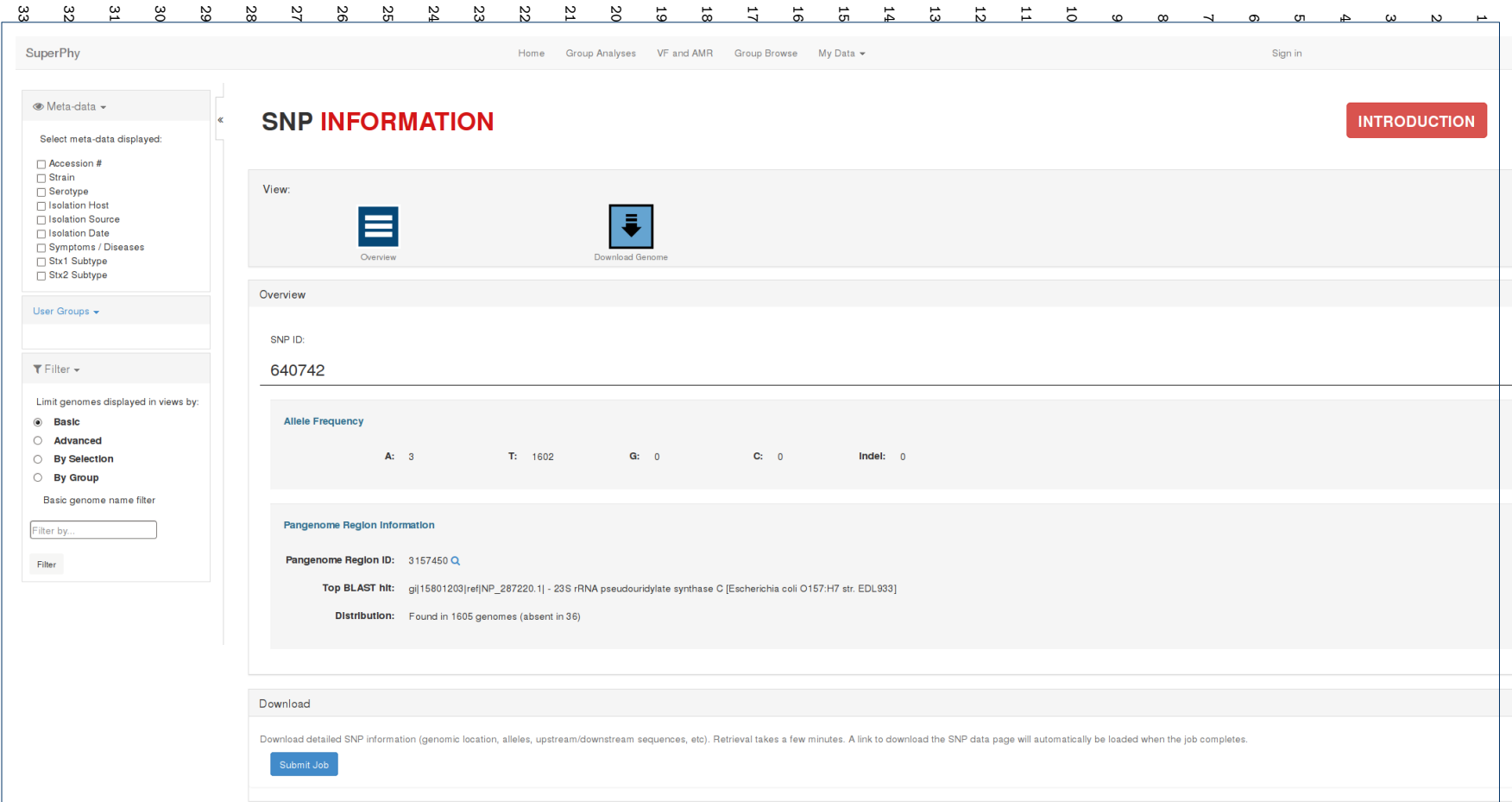


Figure 7 A screen capture demonstrating the ‘SNP Information’ page, where a SNP of interest can be more fully examined. The page identifies the pan-genome region the SNP is found in, the allele frequency of SNPs for all genomes in the database, the putative function of the region given by the top BLAST hit, and an option to download detailed SNP information for each genome. The download includes the genomic location, allele, and upstream / downstream sequences for all genomes in the database.

199.116.235.70/groups/shiny

SuperPhy Home Genome Info Group Analyses VF and AMR GeoPhy My Data Sign in

SUPERPHY data retrieved, but you are not logged in! Groups will not be saved!

Group Selection

Select a group collection to select groups from for group comparisons

isolation_host [N distinct=10]

Select one or more groups for group comparisons

Bos taurus (cow) (n=60) Homo sapiens (human) (n=492) Environmental source (n=18)

Multiple Test Correction

Apply multiple test corrections:

Benjamini & Hochberg (FDR)

P-value significance threshold: 0.0001

Compute "Group vs NOT In Group" Comparisons

Compute All Pairwise Comparisons Between Selected Groups

Group Comparison Summary Table

10 records per page Download Table (.csv)

category	group_1	group_2	n_sig_p_values	lowest_p_value	lowest_log10_p_value
isolation_host	Bos taurus (cow)	Homo sapiens (human)	6276	9.751610e-15	-13.709894
isolation_host	Bos taurus (cow)	Environmental source	100	1.866261e-06	-5.427998
isolation_host	Homo sapiens (human)	Environmental source	911	3.920230e-08	-7.105658

Showing 1 to 3 of 3 entries

First Previous 1 Next Last

Group Comparison - isolation_host - Bos taurus (cow) VS Homo sapiens (human)

10 records per page Search: Download Table (.csv)

Marker	p_value	log10_p_value	fdr	Present in group Bos taurus (cow)	Absent in group Bos taurus (cow)	Present in group Homo sapiens (human)	Absent in group Homo sapiens (human)
36883_1=A,0=T	9.751610e-15	-13.70989	4.762150e-10	9	51	328	164
36885_1=A,0=G	9.751610e-15	-13.70989	4.762150e-10	51	9	164	328
90574_1=A,0=G	1.117583e-13	-12.70872	3.211006e-09	52	8	183	309
90575_1=A,0=CG	1.337729e-13	-12.61137	3.211006e-09	52	8	184	308
36881_1=A,0=G	1.643821e-13	-12.58804	3.211006e-09	7	53	298	194
20033_1=A,0=G	2.108677e-13	-12.37496	3.256781e-09	46	14	136	356
90581_1=A,0=G	4.120251e-13	-12.32260	3.256781e-09	8	52	305	187
90577_1=C,0=T	4.120251e-13	-12.32260	3.256781e-09	52	8	187	305
62894_1=A,0=G	2.599234e-13	-12.29521	3.256781e-09	10	50	323	169
103267_1=A,0=G	5.146470e-13	-12.21509	3.256781e-09	48	12	153	339

Showing 1 to 10 of 67,459 entries

Previous 1 2 3 4 5 Next

Figure 8 A screen capture demonstrating the 'all pairwise comparisons' between selected metadata categories. In this example, those genomes under the metadata category 'Isolation Host' are compared pairwise in all possible combinations for the categories 'Bos taurus (cow)', 'Homo sapiens (human)', and 'Environmental source'. The resulting SNPs in the shared regions for each comparison are ranked by p-value, from most statistically predictive of the group to least, with false discovery rate multiple testing correction. The results table is interactive and the complete dataset can be downloaded as a .csv file for offline analyses.

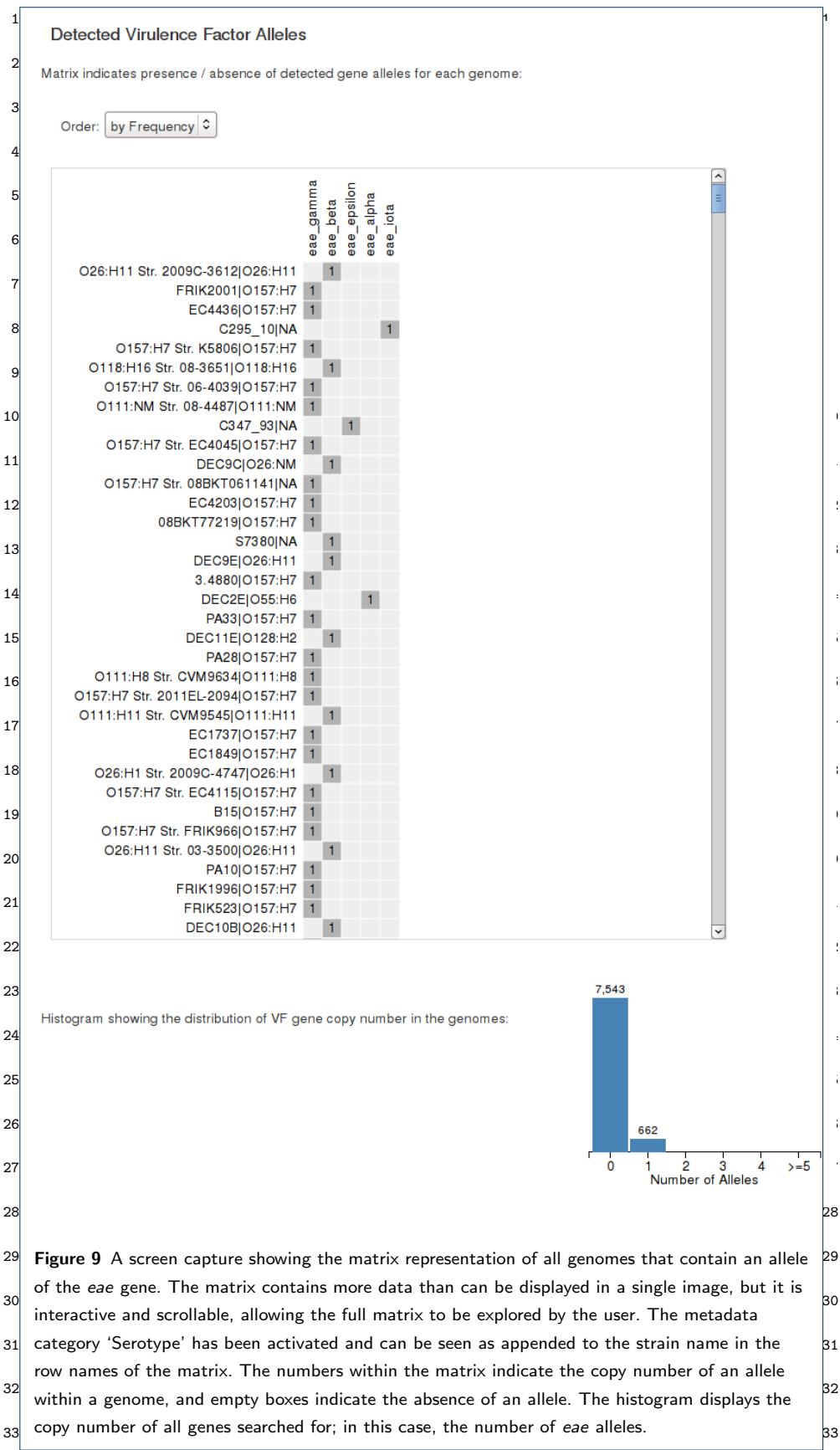


Figure 9 A screen capture showing the matrix representation of all genomes that contain an allele of the *eae* gene. The matrix contains more data than can be displayed in a single image, but it is interactive and scrollable, allowing the full matrix to be explored by the user. The metadata category 'Serotype' has been activated and can be seen as appended to the strain name in the row names of the matrix. The numbers within the matrix indicate the copy number of an allele within a genome, and empty boxes indicate the absence of an allele. The histogram displays the copy number of all genes searched for; in this case, the number of *eae* alleles.

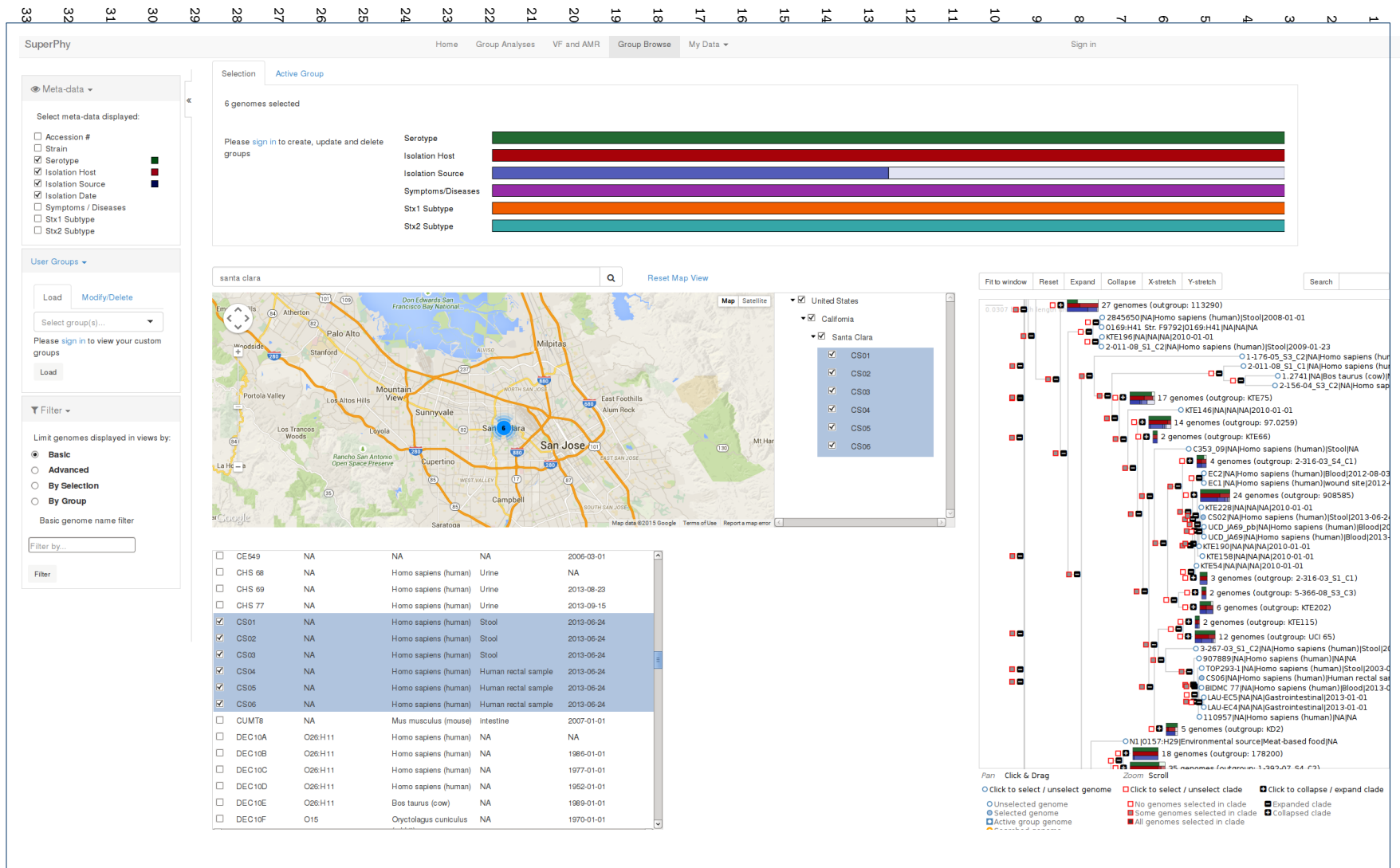


Figure 10 A screen capture showing the ability to group genomes by geographical location and simultaneously examine their phylogenetic position. In this example, all six genomes from Santa Clara, California, United States are selected and highlighted in the map, tree and list views. The available metadata shows that all six genomes were isolated from human sources on the same day; however, their phylogenetic positioning indicates that they are not all from a clonal source. On the tree, all nodes that contain a selected genome are shown as blue-filled squares, while those that do not are white filled squares. Similarly, all selected genomes appear on the tree as blue-filled circles, and those not selected as white-filled circles.

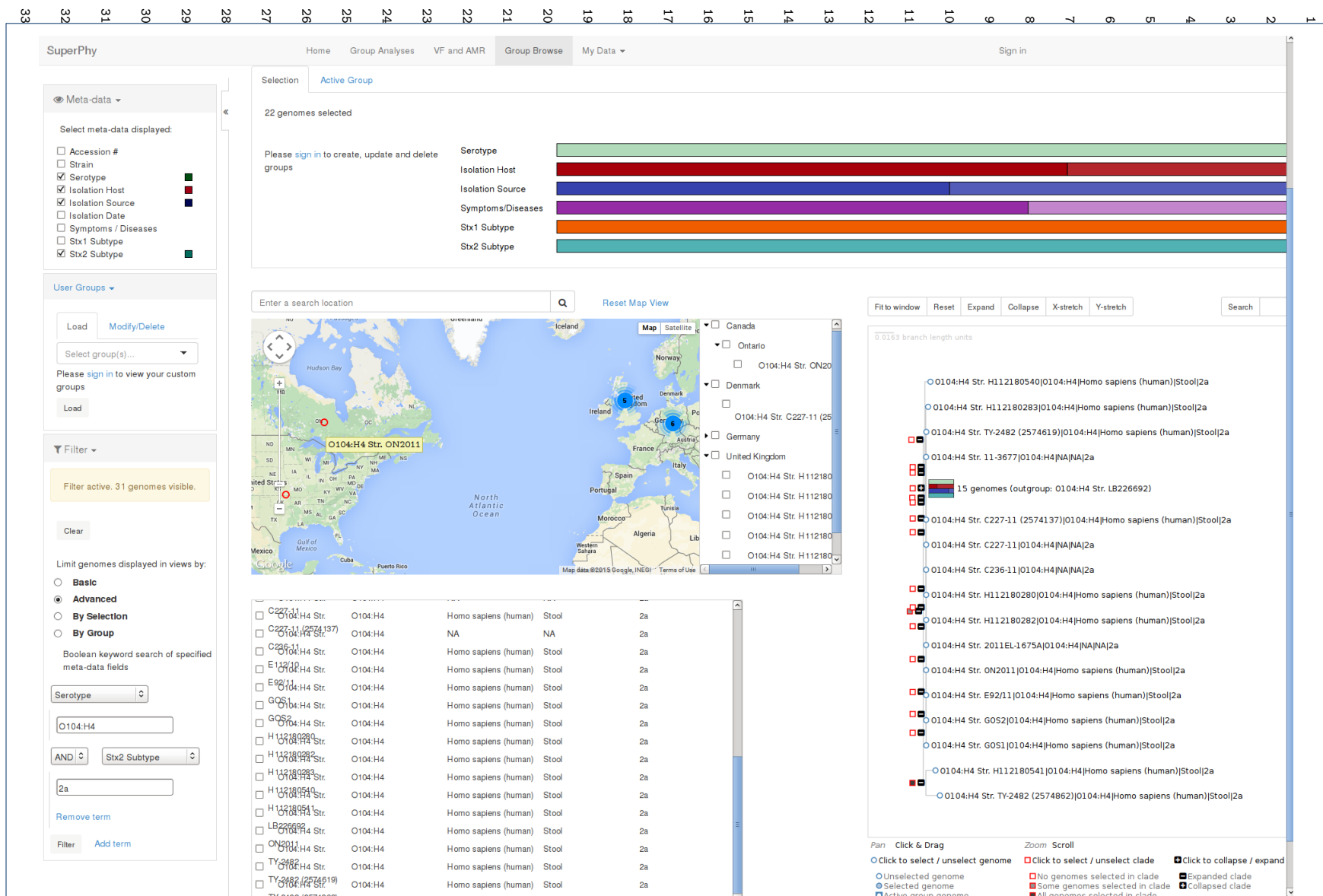


Figure 11 A screen capture showing genomes from the *E. coli* O104:H4 outbreak that occurred in Germany in 2011. The phylogeny of the outbreak strains shows their clonality, and the metadata, visible on the map, shows the dissemination of the bacterial clone from the German epicenter to countries such as Denmark, the United Kingdom, Canada, and the United States, which were determined to be travel-acquired infections.

¹ Tables	1
² Table 1 The percentage of genomes that contain metadata for each of the metadata fields in the	2
³ initial public data set of 1641 <i>E. coli</i> in the SuperPhy database.	3
4	4
5	5
6	6
7	7
8	8
9	9
10	10
11	11
12	12
13	13
14	14
15	15
16	16
17	17
18	18
19	19
20	20
21	21
22	22
23	23
24	24
25	25
26	26
27	27
28	28
29	29
30	30
31	31
32	32
33	33

Metadata field	Percentage
Location	85
Host	79
Date of Isolation	63
Source	52
Serotype	44
Stx2 subtype	23
Stx1 subtype	18
Disease syndrome	6

Table 2 The number of conserved core genomic regions present in 19 selected bacterial genomes, from the total 3598 conserved core genomic regions found in at least 70% of the 2324 *E. coli* genomes examined.

Genome	No. 'conserved core' genes
<i>E. coli</i> O103:H2,12009	3563
<i>E. coli</i> O157:H7, EDL933	3557
<i>E. coli</i> K-12, MG1655	3550
<i>E. coli</i> , UMN026	3483
<i>E. coli</i> O7:K1, CE10	3448
<i>E. coli</i> O83:H1, NRG 857C	3289
<i>Shigella sonnei</i> , 53G	3259
<i>Shigella flexneri</i> 2002017	3148
<i>Shigella boydii</i> , CDC 3083-94	2965
<i>Shigella dysenteriae</i> , 1617	2683
<i>Escherichia fergusonii</i> ATCC 35469	1619
<i>Salmonella enterica</i> subsp. Enterica serovar Typhimurium str. 14028S	95
<i>Citrobacter rodentium</i> , ICC168	77
<i>Klebsiella oxytoca</i> , E718	50
<i>Klebsiella pneumoniae</i> subsp. Pneumoniae, 1084	50
<i>Klebsiella variicola</i> , At-22	46
<i>Escherichia blattae</i> , DSM 4481	27
<i>Staphylococcus aureus</i> , 04-02981	0
<i>Listeria monocytogenes</i> , 07PF0776	0

Table 3 The ten *E. coli* species-specific genomic regions identified in this study based on a total sequence identity of 90%, their location in the K12 reference genome MG1655, the number out of 22324 *E. coli* genomes each region was found in, and their putative function based on the top scoring BLASTx hit.

Region ID	Start bp	Stop bp	No. Genomes	Putative function
3160548	347258	346259	2238	Propionate catabolism operon regulatory protein PrpR
3160296	537566	536567	2256	2-hydroxy-3-oxopropionate reductase
3160113	538566	537567	2248	Allantoin permease
3159571	541565	540567	2275	Purine permease ybbY
3159389	542566	541567	2268	Glycerate kinase
3158844	545665	544666	2261	Allantoate amidohydrolase
3158667	546665	545666	2272	Ureidoglycolate dehydrogenase
3159808	1588200	1587201	2171	FimH protein
3160196	4411062	4410063	2261	Hypothetical protein
3158082	4456632	4457631	2074	Mur ligase family, glutamate ligase domain protein