# VEuPathDB Data Analysis Methods

VEuPathDB draws data from many sources. To facilitate comparisons across data sets, we analyze all data with standardized, data type-specific analyses. All data of one type are analyzed with the same workflow. Although our results may show some differences from an author's publication, our re-analysis of the data makes it feasible to compare data sets from very different sources and to update the data analysis with contemporary methods. For transparency, the methods we use to analyze data are presented here.

### Contents



## Genome analyses

Genome sequence and annotation are analyzed by the EBI Pipeline supplemented with three in-house analyses. In the rare case that the EBI pipeline cannot be applied to a genome, we use a series of in-house analyses in lieu of the EBI Pipelines.

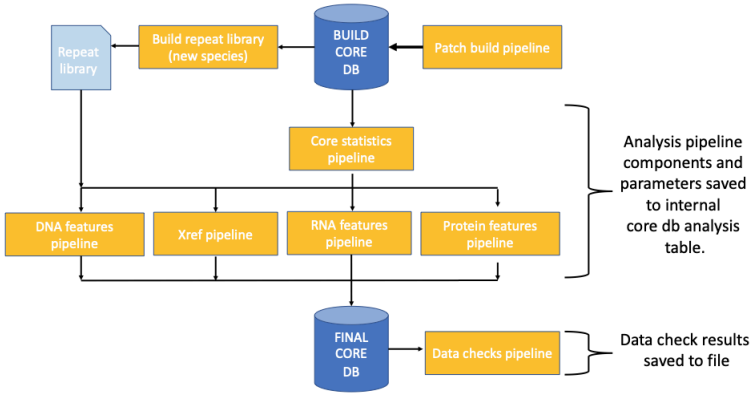
### EBI Pipeline

VEuPathDB employs the Ensembl genome analysis for analyzing genomic sequence to enhance annotations. While most of the genomic sequence (FASTA) are integrated into VEuPathDB from an INSDC repository, genome annotation (GFF3) may come from either the INSDC repository or a community submission.

Core database pipelines (figure 1)- Primary genomic sequence and structural annotation data are loaded into a core database and run through 6 pipelines: core statistics, DNA feature annotation, external cross reference annotation, RNA gene annotation, repeat feature annotation, and protein feature annotation. The main pipelines applied to the core database and their components are listed in table 1.

Based on evolutionarily informed expectations of gene content of near-universal single-copy orthologs, BUSCO metric complements technical metrics like N50. The current version employed by VEuPathDB is v5.4.7. In "genome" mode, BUSCO utilized the MetaEuk gene predictor, while the official protein set of each genome was employed for BUSCO runs in "proteins" mode. BUSCO scores on VEuPathDB are generated with the Ensemble pipeline which computes scores based on the docker provided by BUSCO itself.

Configuration details for each pipeline are determined in Ensembl hive pipeline config files for each pipeline. Since the Ensembl pipelines may change to accommodate bioinformatic advances, pipeline component programs (e.g. Interpro for protein features), versions, and parameters are recorded in the core db analysis table. Final data check results are saved to disk and manually reviewed to determine if the final core db is suitable for release to be loaded into the GUS system.




Core database analysis pipelines and hive components

Example ehive pipelines, modules, programs and parameter data from coredb analysis table

## Supplements to the EBI Pipelines

VEuPath DB supplements the EBI pipeline with workflows that produce data for EST alignments, Open reading frames, and synteny (Table).

EST alignments: BLAT is applied to EST sequences that have been blocked using RepeatMasker.

Open reading frame generation: Open reading frames are generated from genomic DNA or EST sequences. The analysis produces a gff file containing the ORFs (50 or more amino acid translations of the input nucleic acid) for the 6 reading frames. The translations in all 6 reading frames do not necessarily begin with MET but always end at a stop codon. ORF names are in the form template-frame-start-end, e.g. AAEL01000396-5-5847-4366.

Synteny: VEuPathDB uses an in-house script called runMercator to run pair-wise alignments that employs Mercator and MAVID for comparative genome analysis. Mercator generates orthology maps using genomes and exon coordinates to create exon translations for protein BLAT alignments. The orthology maps are used as a guide for MAVID which also uses a phylogenetic newick tree to generate gene alignments.

Product description annotations: In genomes with greater than 90% uninformative gene product names such as 'unspecified product' or 'hypothetical', Pfam domain names are substituted for product descriptions for genes with assigned Pfam domains and uninformative product descriptions. Electronically transferred gene product descriptions are amended with ‘domain containing protein’ and the details of the electronic transfer are chronicled in the gene page Product Descriptions table.

Details for the supplements to the EBI pipelines

## In-house genome analyses in Lieu of the EBI Pipeline

On rare occasions the EBI pipeline cannot be applied to a genome. For example, genomes that are not housed at an INSDC repository cannot be analyzed by the EBI pipeline. VEuPathDB uses the following in-house analyses in lieu of the EBI pipeline.

BLAT against NRDB: For every genome, VEuPathDB runs BLAT alignements of the annotated proteins against the GenBank Non-Redundant Protein Sequence Database (NRDB) to identify possible relationships and alignments outside the scope of VEuPathDB-supported organisms.

Compute open reading frames: Open reading frames are generated from genomic DNA or EST sequences. The analysis produces a gff file containing the ORFs (50 or more amino acid translations of the input nucleic acid) for the 6 reading frames. The translations in all 6 reading frames do not necessarily begin with MET but always end at a stop codon. ORF names are in the form template-frame-start-end, e.g. AAEL01000396-5-5847-4366.

DNA repeats: The Tandem Repeats Finder program locates and displays tandem repeats in genomic sequences.

EST alignments: BLAT is applied to EST sequences that have been blocked using RepeatMasker.

Protein domain annotations: InterProScan scans protein sequences against the protein signatures of the InterPro member databases and generates a file containing the domain matched, description of the InterPro entry, GO descriptions and E-values.

Signal peptide prediction: Signal P is used to identify signal peptides and their likely cleavage sites. A signal peptide is a short peptide present at the N-terminus of most newly synthesized proteins that are destined towards the secretory pathway.

Syntenic sequences: VEuPathDB uses an in-house script called runMercator to run pair-wise alignments that employs Mercator and MAVID for comparative genome analysis. Mercator generates orthology maps using genomes and exon coordinates to create exon translations for protein BLAT alignments. The orthology maps are used as a guide for MAVID which also uses a phylogenetic newick tree to generate gene alignments.

Transmembrane domain prediction: TMHMM is used to predict transmembrane domain presence and topology from protein sequences.

tRNA gene prediction: tRNAScan identifies transfer RNA genes in transcript or genome sequences.

Details for the VEuPathDB in-house pipelines

## Preserving annotations with LiftOff

We have developed a pipeline to run LiftOff to preserve old annotations. First, the distance and flank parameters are incrementally changed to find the combination with the lowest flank number and the fewest missing features. Features are retrieved from the source GFF3 and passed to LiftOff. Then we use agat_sp_fix_cds_phases to calculate phase information and identify any transferred gene models that are incomplete or otherwise invalid. AGAT's agat_sp_extract_sequences is then used to extract CDS sequences for the transferred genes on the new genome. Next, missing CDS regions are identified, and the pipeline produces datasets of what those changes are on a model-by-model basis. It also outputs summary graphics and a corrected GFF3 with metadata regarding model validity (i.e. if it contains stop codons, matches the original sequence, or has any missing regions).

The GFF3 of the lifted annotation is then loaded as a track on the new genome. Each model is colour coded according whether the transfer is valid (it does not have stop codons or missing regions) and if the protein on the new genome matches the protein from the source genome.

Transfers are considered invalid where coding sequence has a missing CDS region or internal stop codon and where ncRNA sequences do not match between the source and transfer. Coding sequences with mismatching protein sequence are not considered immediately invalid, but should be examined to determine if the annotation should be kept, fixed or discarded.

---

# Orthology

The identification of orthologs is an important concept in gene evolution, commonly used to infer likely gene function(s) in newly-sequenced genomes. Orthologs are proteins in different species that have diverged solely through speciation, and are therefore relatively likely to retain similar function (Gabaldón & Koonin 2013). Paralogs are proteins that arose through gene duplication; while initially identical, duplication provides the opportunity for functional divergence and evolutionary selection. "In-paralogs" refer to recent duplications that are likely to retain similar function; "out-paralogs" derive from more ancient duplications predating divergence of the species of interest, and are more likely to have diverged in function. See Chen 2007 and the Orthology Benchmarking Service (Altenhoff 2020) for descriptions and assessments of widely-used ortholog detection algorithms

## OrthoMCL

VEuPathDB websites use the OrthoMCL algorithm (Li 2003) to infer orthology and paralogy, employing protein sequence similarity (BLASTP) to relate proteins across the tree of life, several normalization steps to improve performance on complex eukaryotic genomes, and Markov clustering to group these proteins into ortholog groups. Each protein in every species is assigned to precisely one OrthoMCL ortholog group (e.g. OG6_135465). The OrthoMCL website allows users to explore ortholog groups and the proteins they contain, including sequences, similarity metrics, alignments, domain architecture, and phyletic pattern profiles. Users can also assign putative ortholog groups to proteins from additional genomes of interest (Fisher 2011). Because of the large and rapidly growing number of available genomes, the most recent implementation of OrthoMCL has been modified to improve scalability and usability, as described here.

### Orthology on the Gene Page

All Gene Pages in VEuPathDB databases include an 'Orthology and Synteny' section providing: a link to the ortholog group on the OrthoMCL database, a list of orthologs within the same VEuPathDB database component (e.g. PlasmoDB; see the OrthoMCL database for orthologs in other species), and a JBrowse genome browser display of syntenic chromosomes in other strains and species (shaded to indicate orthology).

### Orthology-based functional prediction

Many genes have yet to be functionally characterized, but function can often be inferred using orthology and protein feature information. VEuPathDB assigns putative Enzyme Commission (EC) numbers (e.g. 2.7.4.3) based on a heuristic scoring algorithm that considers the length, sequence similarity, and domain architecture of other proteins in the same ortholog group that have previously been annotated with this EC number. Scoring details and other statistics are provided in a table in the 'Functional Prediction' section of the Gene Page; further details on EC prediction are available here. The inferred EC number is also represented in metabolic pathways graphs and tables accessible via the 'Pathways and Interactions' section of the Gene Page.

### Searches for genes based on orthology

Several searches employ orthology to identify genes of possible interest. The Orthology Phylogenetic Profile Search, identifies genes based on the taxonomic groups in which orthologs are or are **not** found (i.e. genes displaying a specific pattern of conservation across the tree of life). For example, users seeking a pathogen-specific therapeutic target may wish to search for genes present in the pathogen(s) of interest, but not in non-pathogenic species or host species (e.g. mammals). The Paralog Count Search allows users to identify genes based on the number of paralogs identified, e.g. genes that have undergone a recent expansion through one or more duplications. Finally, any list of genes identified through previous searches or Search Strategies (tutorial) may be converted into a list of orthologs in other species. This is particularly useful when seeking to identify genes in an understudied species based on functional information in more intensively characterized species (e.g. essentiality data derived from mutational screens conducted in model organisms).

---

# Proteomics

VEuPathDB integrates the results of proteomics experiments as peptides aligned to a reference genome or as abundance data assigned to a gene. We do not reanalyze the raw mass spec data but instead use an in-house plugin that loads found peptides or abundance data from tab delimited input files of a specific format.

Details for the VEuPathDB in-house proteomics pipeline

---

## RNA Sequence

VEuPathDB integrates RNA-Seq data from many different experiments and analyzes all data with the same EBI RNA-Seq analysis pipeline. The RNA sequence data that we integrate is processed at EBI.

The following is a general outline of the analysis process.

- Trim poor quality data (Trimmomatic)
- HiSAT2 alignment to a reference genome
- HT-Seq-count to tally aligned reads per gene
- Convert count data to transcripts per kilobase million (TPM) for use in visualisations and fold-change searches
- Use count data as input to DESeq2 to determine differential expression

EBI RNA-Seq pipeline details

---

## Single-cell RNA Sequence

VEuPathDB supports scRNA-Seq data as expression values and cell cluster projections displayed in the CELLXGENE interactive data mining application. VEuPathDB loads the data directly from data providers without reanalyzing raw data. Data is shared with VEuPathDB as:

- A (sparse) matrix of the normalised expression for each gene in each cell.
- Coordinates for the cluster analysis such as UMAP, PCA or PHATE projections.
- Metadata about the cells which can originate from experimental parameters (e.g., which sample a given cell came from), qa output (e.g., number of features and number of RNAs counted for each cell), or derived metadata such as cluster assignments, of which there may be several if the clustering has been done in different ways.

---

## ChIP-Sequence

VEuPathDB integrates ChIP-Seq data from many different experiments and sources. DNA seq data are aligned to the reference genome using Bowtie2. Alignment results are converted to bigwig and displayed in JBrowse.

---

## Copy Number Variation

VEuPathDB uses coverage from whole genome sequencing data to estimate gene and chromosome copy numbers in sequenced strains. The bowtie2 alignments generated during SNP analysis are used as a starting point. HTseq-count is used to count the number of reads that align to each gene and the values are converted to transcripts per million (TPM). Assuming that the median TPM value represents a single copy gene on a chromosome of constitutive ploidy, we can infer gene or chromosome duplications by comparing the TPM values for individual genes or the median TPM for individual chromosomes to the whole genome median using custom scripts based on the method described in PMID: 22038252. Additionally, coverage is calculated in 1 kb bins across the genome, normalised to the constitutive ploidy and converted to bigwig format for visualisation in JBrowse.

Haploid number and gene dose are metrics used to define copy number in VEuPathDB. Haploid number is the number of genes on an individual chromosome. Gene dose is the total number of genes in an organism, accounting for copy number of the chromosome. For example, a single-copy gene in a diploid organism has a haploid number of 1 and a gene dose of 2.

### Searches for genes based on Copy Number Variation

There are two searches that query copy number data in VEuPathDB. The first, Identify Genes based on Copy Number (CNV) returns genes that are present at copy numbers within a range that you specify. The search can be configured to return genes based on the haploid number or gene dose. The second search, Identify Genes based on Copy Number Comparison (CNV), returns genes for which the copy number varies between the reference and your chosen isolates. This search compares the estimated copy number of a gene in the resequenced strain(s) with the copy number in the reference annotation. The copy number in the reference annotation is calculated as the number of genes that are both on the same chromosome and in the same ortholog group as the gene of interest. We infer that these genes have arisen as a result of tandem duplication of a common ancestor. In this search, the metric for copy number is the haploid number, which is the number of copies of a gene on a single chromosome.

---

## Genetic Variation and SNP calling

VEuPathDB analyzes whole genome resequencing data to call single nucleotide polymorphisms of isolates. The method employed by VEuPathDB to call SNPs from short read sequencing like Illumina reads, follows these steps:

- Reads are aligned to the reference genome using bowtie2
- The resulting BAM file from bowtie2 is sorted and a pileup file using samtools is generated
- Reads around indels are realigned using GATK
- SNPs and indels are called consensus sequence using VarScan is generated:
  - P value <= 0.01

  - minimum aligned reads >= 5
  - minimum read frequency >= 0.8
- SNP calls where coverage is >2.5x the median coverage are removed to limit erroneous calls in repeat regions
- At each SNP position “like reference” calls are generated for each strain that is identical to the reference to give the full picture of each SNP

---

## Microarray data

VEuPathDB integrates microarray data from high density oligonucleotide as well as spotted arrays. In general, the data comes to us as intensities associated with probes. VEuPathDB does not reanalyze the original fluorescence data. We process the data we receive according to the following outline:

- Map the array probes to the reference genome's transcriptome
- Filter the data to remove outliers.
- Normalize
  - For one channel data we perform a robust multi-array average (RMA) normalizations.
  - For two channel data we perform a Loess normalization
- Compute the average probe intensity per gene.
- Compute the expression average per gene.
  - First, average the technical replicates.
  - Second, average the biological replicates (if any).
- Optional: perform differential expression analysis if there is a sufficient number of biological replicates.

---

## Protein Array data

VEuPathDB integrates protein array data from serum antibody microarray experiments. In general, the data comes to us as intensities associated with probes. VEuPathDB does not reanalyze the original fluorescence data. Although each experiment and data set can have special considerations, we process the data according to the following outline:

- Map the array probes to the reference genome's transcriptome
- Filter the data to remove outliers.
- Normalize
  - For one channel data we perform a robust multi-array average (RMA) normalizations.
  - For two channel data we perform a Loess normalization
- Compute the average probe intensity per gene.
- Compute the expression average per gene.
  - First, average the technical replicates.
  - Second, average the biological replicates (if any).
- Optional: perform differential expression analysis if there is a sufficient number of biological replicates.

---

## Metabolic Pathways

VEuPathDB integrates metabolic pathways from KEGG and MetaCyc. For TriTrypDB, pathways are also integrated from LeishCyc and TrypanoCyc. Metabolic pathways are associated with genes via Enzyme Commission annotations.

The **VEuPathDB** Bioinformatics Resource Center makes genomic, phenotypic, and population-centric data accessible to the scientific community. **TriTrypDB** provides support for these organisms.

This project is funded in part by the US National Institute of Allergy and Infectious Diseases (Contract HHSN75N93019C00077), with additional support from the Wellcome Trust (Resource Grants 212929 & 218288).