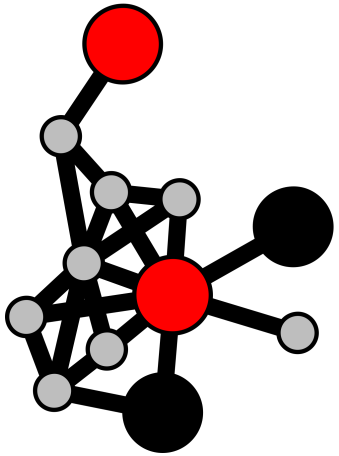
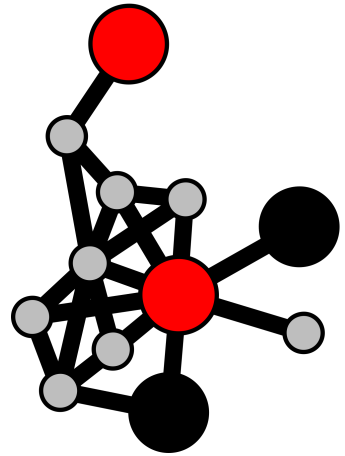
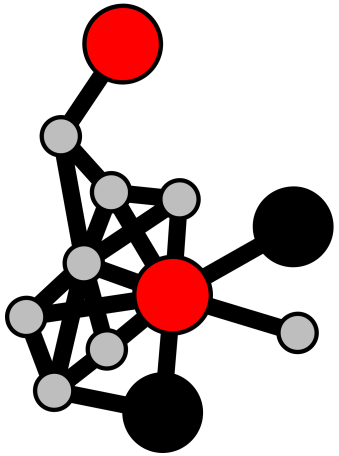
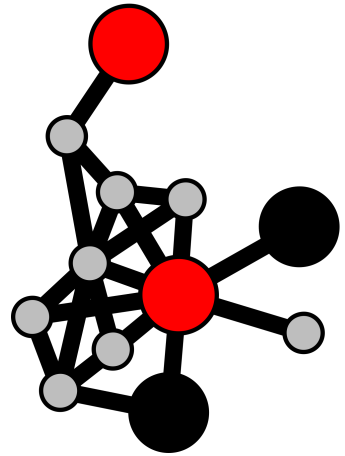




Functional annotation lesson





Targeted HMM-based annotation:

FeGenie introduction and tutorial

Artwork: Nancy Merino

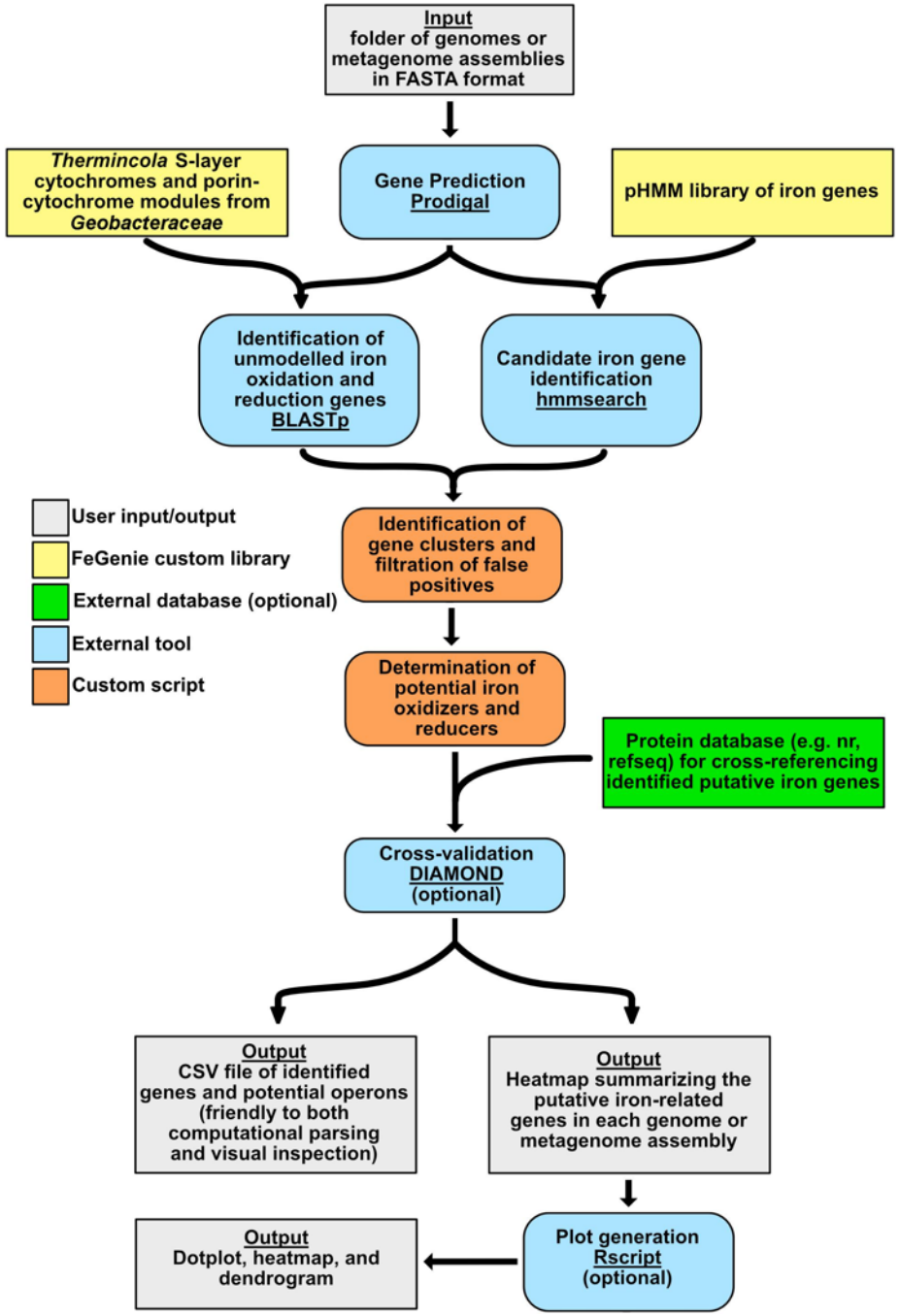
What is FeGenie

1. A collection of HMMs, based on a comprensive set of genetic markers related to iron acquisition/scavenging, transport, efflux, storage, as well as iron redox (dissimilatory reduction and oxidation)
2. A bioinformatics software that uses these HMMs to profile genomic datasets (in FASTA format)
3. Uses rules for filtering out potential false positives (operon structure, genetic co-occurance, bit scores)

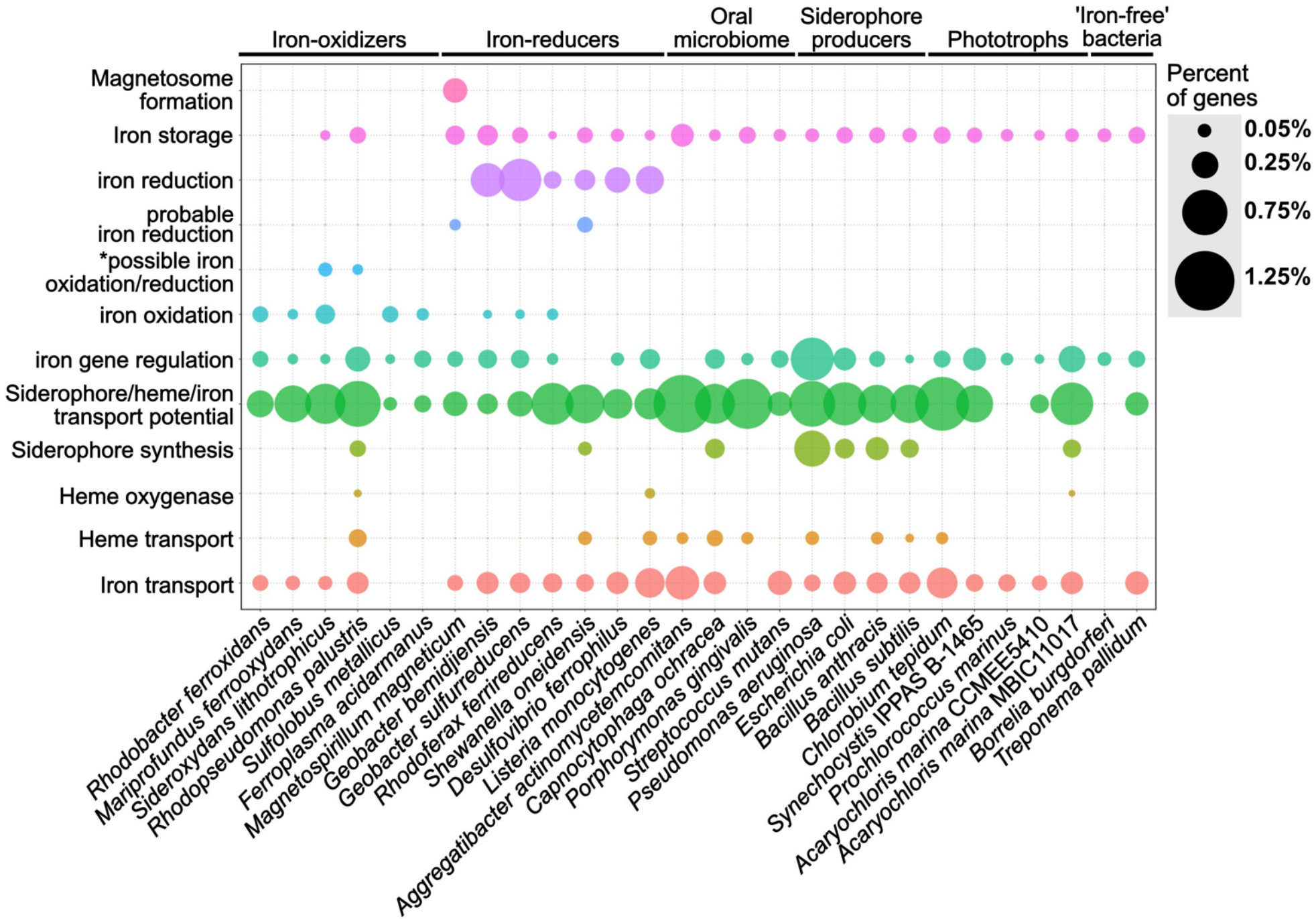
Category	Function	Protein Families
Iron acquisition	Iron(II)/(III) transport	Efe**U**OB[1], FbpABC[2], SfuABC[3], YfuABC[4], **FeoAB(C)**[5], FutA1[6], FutA2[6], FutB[6], FutC[6], YfeABCD[7]
	Heme oxygenase	ChuS[8], ChuZ[9], HemO[10,11], PigA[10,11], Hem**RS**T**U**V[12], HmoB[13], HmuO[14], HugZ[15], HupZ[16], Isd-LmHde[17], IsdG[18], IsdI[19], MhuD[20], **PhuS**[21] (in PhuRSTUVW)
	Heme transport	Has**RA**DE(B)F[22], HmuR**S**T**U**V[22], **HmuY**[23], **HmuY'**[23], HutZ[24], Hxu**C**BA[25], **IsdX1**[26], **IsdX2**[26], Phu**RS**T**U**V**W**[21], **Rv0203**[27]
	Transferrin/Lactoferrin	Tbp**A**B (Lbp**A**B)[28], Sst**AB**CD[29]
	Siderophore synthesis	Acs**A**B**CD**EF[30], AmoA[31], AngR[32], Asb**AB**CDEF[33], DhbACEBF[34], entD-**fepA**-fes-entF-fepEC**GDB**-entCEBA-ybdA[35], IroD in Iro**N**BCDE[36], IucABCD[37,38], **IutA**[37,38], MbtIJABCDEFGH[35], LbtA[39] (in LbtUABC), PchABCDREFGHI[35], PvdQAPMNOFEDJIHLGS[40], PvsABCDE[41], VenB[42], Vab genes in VabR-fur-vabGA-fur-VabCEBSFH-fur-fvtA-vabD[43], Vib genes in VibB-vibEC-vibA-vibH-viuP**DG**C-vibD and viuAB-vibF[44–46], RhbABCDEF-rhrA-rhtA[47]
	Siderophore transport	BesA[48], CbrA**BC**D[49], **TonB**-**ExbB**-**ExbD**[50], Fat**A**B**CD**[51], FecI**RA**B**CD**E[52], FeuABC-yusV[53], Fhu**A**CDB[54–56], **FhuF**[54–56], Fpt**A**BCX[57], FpuA**B**[58], FpuC[58], FpuD[58], FpvI**R**-**FpvA**-FpvGHJKCDEF[59], FvtA in VabR-fur-vabGA-fur-VabCEBSFH-fur-fvtA-vabD[43], HatCD**B**[37], IroNBCDE[36], LbtUABC[39], PirA[60], PiuA[60], Pvu**A**BCDE[41], Viu genes in VibB-vibEC-vibA-vibH-viuP**DG**C-vibD and viuAB-vibF[44–46], **YfiZ**-yfhA[61], YfiY[61], YqjH[62], ybdA and Fep genes in entD-**fepA**-fes-entF-fepEC**GDB**-entCEBA-ybdA[35]
Iron Gene regulation	Transcriptional regulation	**DtxR**[63], FecR (in FecIRABCDE)[52], **FeoC** in FeoAB(C)[5], **Fur**[64], **IdeR**[65], YqjI[62], RhrA in RhbABCDEF-rhrA-rhtA[47]
Iron oxidation and reduction	Iron oxidation	Cyc1[66,67], Cyc2[66,67,68], FoxABC[69], FoxEYZ[70], Sulfocyanin[71], PioA**B**C[72]
	Probable iron oxidation and possible iron reduction	MtoA**B**[73], Cyc2 (cluster 3)
	Dissimilatory iron reduction	CymA[74], Mtr**C**A**B**[75], OmcF[76], OmcS[76], OmcZ[76], FmnA-dmkA-fmnB-pplA-ndh2-eetAB-dmkB[77], DFE_0448-0451, DFE_0461-0465[78]
	Probable iron reduction	MtrCB, MtrAB, MtoAB-MtrC
Iron storage	Iron storage	**Bfr**[79], **DpsA**[80], **Ftn**[81]
Magnetosome-related	Magnetosome formation	MamABEKLMOPQI[82,83] (Note: These genes are found in all known magnetotactic microorganisms, except for *mamL* which is found in magnetite-producing magnetotactic microorganisms[81])

Bolded and underlined HMMs are derived from Pfam or TIGRFAMs databases. Other HMMs were created by using select sequences. See ***Supplementary Table S1*** *for more information, including the corresponding Pfam or TIGRFAMs families and the sequences used to create the HMMs.* [1]*Miethke et al., 2013*, [2]*Adhikari et al., 1996*, [3]*Angerer et al., 1990*, [4]*Gong et al., 2001*, [5]*Lau et al., 2016*, [6]*Katoh et al., 2001*, [7]*Bearden et al., 1998*, [8]*Suits et al., 2006*, [9]*Zhang et al., 2011*, [10]*Friedman et al., 2003*, [11]*Friedman et al., 2004*, [12]*Schneider et al., 2006*, [13]*Park et al., 2012*, [14]*Matsui et al., 2005*, [15]*Hu et al., 2011*, [16]*Sachla et al., 2016*, [17]*Duong et al., 2014*, [18]*Reniere et al., 2010*, [19]*Skaar et al., 2004*, [20]*Graves et al., 2014*, [21]*Ochsner et al., 2000*, [22]*Tong and Guo, 2009*, [23]*Wójtowicz et al., 2009*, [24]*Liu X. et al., 2012*, [25]*Morton et al., 2007*, [26]*Honsa et al., 2014*, [27]*Tullius et al., 2011*, [28]*Gray-Owen et al., 1995*, [29]*Morrissey et al., 2000*, [30]*Carroll and Moore, 2018*, [31]*Barghouthi et al., 1991*, [32]*Wertheimer et al., 1999*, [33]*Oves-Costales et al., 2007*, [34]*May et al., 2001*, [35]*Crosa and Walsh, 2002*, [36]*Hantke et al., 2003*, [37]*Suzuki et al., 2006*, [38]*Martínez et al., 1994*, [39]*Cianciotto, 2015*, [40]*Lamont and Martin, 2003*, [41]*Tanabe et al., 2003*, [42]*Tan et al., 2014*, [43]*Balado et al., 2008*, [44]*Wyckoff et al., 2001*, [45]*Keating et al., 2000*, [46]*Wyckoff et al., 1999*, [47]*Lynch et al., 2001*, [48]*Miethke et al., 2006*, [49]*Mahé et al., 1995*, [50]*Garcia-Herrero et al., 2007*, [51]*Lemos et al., 2010*, [52]*Braun, 2003*, [53]*Peuckert et al., 2011*, [54]*Köster and Braun, 1989*, [55]*Coulton et al., 1987*, [56]*Braun et al., 2002*, [57]*Youard et al., 2011*, [58]*Dixon et al., 2012*, [59]*Brillet et al., 2012*, [60]*Moynie et al., 2017*, [61]*Ollinger et al., 2006*, [62]*Wang et al., 2011*, [63]*Guedon and Helmann, 2003*, [64]*Escolar et al., 1998*, [65]*Rodriguez et al., 2002*, [66]*Castelle et al., 2008*, [67]*Barco et al., 2015*, [68]*McAllister et al., 2019*, [69]*Bathe and Norris, 2007*, [70]*Croal et al., 2007*, [71]*Ilbert and Bonnefoy, 2013*, [72]*Liu J. et al., 2012*, [73]*Jiao and Newman, 2007*, [74]*Castelle et al., 2015*, [75]*Pitts et al., 2003*, [76]*Santos et al., 2015*, [77]*Light et al., 2018*, [78]*Deng et al., 2018*, [79]*Grossman et al., 1992*, [80]*Grant et al., 1998*, [81]*Andrews, 1998*, [82]*Uebe and Schuler, 2016*, [83]*Kolinko et al., 2016*.

Bolded and underlined genes represent HMMs taken from Pfam and TIGRFAMs



General workflow involves a combination of custom scripts and other bioinformatics software for ORF-prediction and orthology searching


Iron-oxidizers
Iron-reducers
Oral microbiome
Siderophore producers
Phototrophs
'Iron-free' bacteria
Magnetosome formation
Iron storage
iron reduction
probable iron reduction
*possible iron oxidation/reduction
iron oxidation
iron gene regulation
Siderophore/heme/iron transport potential
Siderophore synthesis
Heme oxygenase
Heme transport
Iron transport
Percent of genes
0.05%
0.25%
0.75%
1.25%
Rhodobacter ferroxidans
Mariprofundus ferrooxydans
Sideroxydans lithotrophicus
Rhodopseudomonas palustris
Sulfolobus metallicus
Ferroplasma acidarmanus
Magnetospirillum magneticum
Geobacter bemidjiensis
Geobacter sulfurreducens
Rhodoferax ferrireducens
Shewanella oneidensis
Desulfovibrio ferrophilus
Listeria monocytogenes
Aggregatibacter actinomycetemcomitans
Capnocytophaga ochracea
Porphyromonas gingivalis
Streptococcus mutans
Pseudomonas aeruginosa
Escherichia coli
Bacillus anthracis
Bacillus subtilis
Chlorobium tepidum
Synechocystis IPPAS B-1465
Prochlorococcus marinus
Acaryochloris marina CCMEE5410
Acaryochloris marina MBIC11017
Borrelia burgdorferi
Treponema pallidum

Tutorial

Dependencies

- Python (version 3.6 or higher)

Absolutely required dependencies

- Diamond (version 0.9.22.123) (only necessary if you are doing the cross-validation against a reference database)
- BLAST (version 2.7.1+)
- HMMER (version 3.2.1)

Absolutely required dependencies

- Prodigal (version 2.6.3)
- R (version 3.5.1)
- Rscript

Installation

```
git clone https://github.com/Arkadiy-Garber/FeGenie.git
cd FeGenie
bash setup_noconda.sh
```

Put $PATH to FeGenie into your bash profile (.bash_profile or .profile in your home directory):

PATH="/Users/arkadiygarber/topic-functional-annotation/FeGenie-tutorial/FeGenie:$PATH"

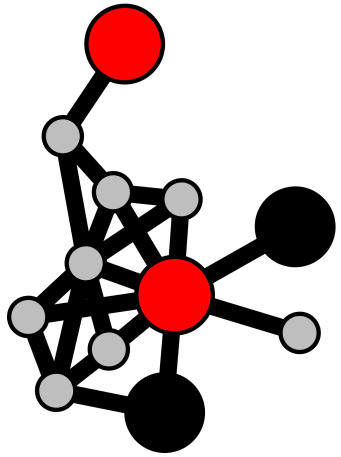
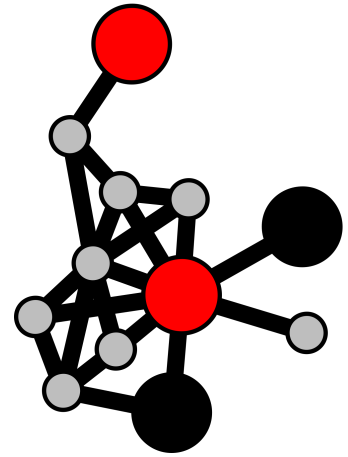
Easy Installation (if you have Conda installed)

```
git clone https://github.com/Arkadiy-Garber/FeGenie.git
cd FeGenie
bash setup.sh
conda activate fegenie
FeGenie.py -h
```

Usage

```
-h, --help              show this help message and exit
-bin_dir BIN_DIR        directory of bins
-bin_ext BIN_EXT        extension for bins (do not include the period)
-d D                    maximum distance between genes to be considered in a
                        genomic 'cluster'.This number should be an integer and
                        should reflect the maximum number of genes in between
                        putative iron-related genes identified by the HMM
                        database (default=5)
-ref REF                path to a reference protein database, which must be in
                        FASTA format
-out OUT                name output directory (default=fegenie_out)
-inflation INFLATION    inflation factor for final gene category counts
                        (default=1000)
-t T                    number of threads to use for DIAMOND BLAST and
                        HMMSEARCH (default=1, max=16)
-bams BAMS              a tab-delimited file with two columns: first column
                        has the genome or metagenome file names; second column
                        has the corresponding BAM file (provide full path to
                        the BAM file). Use this option if you have genomes
                        that each have different BAM files associated with
                        them. If you have a set of bins from a single
                        metagenome sample and, thus, have only one BAM file,
                        then use the '-bam' option. BAM files are only
                        required if you would like to create a heatmap that
                        summarizes the abundance of a certain gene that is
                        based on read coverage, rather than gene counts.
-bam BAM                BAM file. This option is only required if you would
                        like to create a heatmap that summarizes the abundance
                        of a certain gene that is based on read coverage,
                        rather than gene counts. If you have more than one BAM
                        filecorresponding to different genomes that you are
                        providing, please use the '-bams' argument to provide
                        a tab-delimited file that denotes which BAM file (or
                        files) belongs with which genome
--gbk [GBK]             include this flag if your bins are in Genbank format
--orfs [ORFS]           include this flag if you are providing bins as open-
                        reading frames or genes in FASTA amino-acid format
--meta [META]           include this flag if the provided contigs are from
                        metagenomic/metatranscriptomic assemblies
--norm [NORM]           include this flag if you would like the gene counts
                        for each iron gene category to be normalized to the
                        number of predicted ORFs in each genome or metagenome.
                        Without normalization, FeGenie will create a heatmap-
                        compatible CSV output with raw gene counts. With
                        normalization, FeGenie will create a heatmap-
                        compatible with 'normalized gene abundances'
--makeplots [MAKEPLOTS]
                        include this flag if you would like FeGenie to make
                        some figures from your data?. To take advantage of
                        this part of the pipeline, you will need to have
                        Rscipt installed. It is a way for R to be called
                        directly from the command line. Please be sure to
                        install all the required R packages as instrcuted in
                        the FeGenie Wiki: https://github.com/Arkadiy-
                        Garber/FeGenie/wiki/Installation. If you see error or
                        warning messages associated with Rscript, you can
                        still expect to see the main output (CSV files) from
                        FeGenie.
```

Customize your iron-gene annotation pipeline

Jupyter Binder tutorial