**Date: 26 Aug, 2019**

**Datatset Title:** Empirical measures of mutational effects define neutral models of regulatory evolution in Saccharomyces cerevisiae

**Dataset Creators:** Andrea Hodgins-Davis, Fabien Duveau, Elizabeth Walker, and Patricia J Wittkopp

**Dataset Contact:** andreahd@umich.edu

**Funding:** NIH F32 GM115198 to AHD, EMBO ALTF 1114-2012 to FD, and NSF MCB-1021398, NIH 1R35GM118073, and NIH R01 GM108826 to PJW

**Research Abstract:**

Understanding how phenotypes evolve requires disentangling the effects of mutation generating new variation from the effects of selection filtering it. Tests for selection frequently assume that mutation introduces phenotypic variation symmetrically around the population mean, yet few studies have tested this assumption by deeply sampling the distributions of mutational effects for particular traits. Here, we examine distributions of mutational effects for gene expression in the budding yeast Saccharomyces cerevisiae by measuring the effects of thousands of point mutations introduced randomly throughout the genome. We find that the distributions of mutational effects differ for the ten genes surveyed and are inconsistent with normality. For example, all ten distributions of mutational effects included more mutations with large effects than expected for normally distributed phenotypes. In addition, some genes also showed asymmetries in their distribution of mutational effects, with new mutations more likely to increase than decrease the gene’s expression or vice versa. Neutral models of regulatory evolution that take these empirically determined distributions into account suggest that neutral processes may explain more expression variation within natural populations than currently appreciated.

**Methods:**

The data are flow cytometry measurements of the level of expression of a YFP reporter gene driven by one of 10 different promoters for a collection of budding yeast genotypes in which point mutations have been induced by genome-wide mutagenesis.

**Instrument/Software:**

Flow cytometry data were collected on the BD Accuri C6 fitted with a HyperCyt autosampler (IntelliCyt). Data analyses were performed in R v3.3.3 using a variety of custom scripts made available with the manuscript. R packages flowCore v1.5 and flowClust v3.22 were used to process raw fluorescence measurements.

**All Supporting Data and Metadata**

Supplementary File 1: Primers used to generate and sequence confirm reporter constructs. Provided directly in the supplementary information published with the manuscript.

Supplementary File 2: R code for processing raw .fcs files, normalizing phenotypes by plate controls, filtering outliers, and calculating mean phenotypes by promoter. Provided directly in the supplementary information published with the manuscript.

Supplementary File 3: Layout spreadsheet with experimental metadata linking .fcs files to samples. Made available here at Deep Blue Data.

Supplementary File 4: R code used to contrast mean phenotypes among promoters. Provided directly in the supplementary information published with the manuscript.

Supplementary File 5: R code for evolutionary simulations. Provided directly in the supplementary information published with the manuscript.

Supplementary File 6: Processed data file including mean phenotypes. Made available here at Deep Blue Data.

Raw flow cytometry data are made available on FlowRepository (record FR-FCM-ZYUW) at the URL <https://flowrepository.org/id/FR-FCM-ZYUW>.

**Files contained here:**

**Supplementary File 3.zip** contains metadata describing the identity and treatment conditions of each individual sample scored on the BD Accuri in this experiment. Spreadsheets contain metadata for analyses of one promoter construct each. Corresponding raw flow cytometry data are made available on FlowRepository.org Experiment: FR-FCM-ZYUW. Briefly, each row in these files contains identifying metadata describing each sample analyzed by high-throughput flow cytometry, including:

ASSAY: represents in which experiment the sample was collected

STRAIN: represents the genotype of the strain. CTRL indicates wild type unmutagenzied colonies used to adjust for systematic effects across plates. Otherwise, strain names contain the promoter driving the reporter gene, the number of and position in the plate in which they are found, and an indication of whether the genotype represents an unmutagenized (SHAM) control isolated to a single cell in parallel with the mutagenized population.

CONDITION: indicates whether the colony measured was a control for plate effects (“CTRL”), promoter-matched unmutagenized sham-treated control (“SHAM”), P-TDH3 genetic background unmutagenized sham-treated control (“TDH3.SHAM”), mutagenized cell isolated randomly (“EMS”), or non-randomly (“EMS.LOW”, “EMS.HIGH”).

ROUND: internal experimental information

DAY: date of flow cytometry data collection

PLATE: unique identifier for each 96-well plate of cells scored

REP: unique identifier for each flow cytometry replicate in which fluorescence was measured

STACK.ORDER: identifies the depth of the plate in the shaking incubator towers. 1 indicates the top, 4 indicates the bottom.

FLOW.RUN: unique data collection bout on the instrument

BLOCK: indicates whether the FLOW.RUN was the first or second data collection bout from the plate

ROW: indicates position of the sample in rows of the 96-well plate

COLUMN: indicates position of the sample in columns of the 96-well plate

WELL/POSITION: indicate row and column where sample is located in 96-well plate

FILENAME: path to file

SKIP: manual annotation

**Supplementary File 6.zip** contains processed flow cytometry data including new data produced in this work in PROCESSED.DATA.txt and two experiments containing previously collected data for the promoter TDH3 in files TDH3.EXP1.txt and TDH3.EXP.2.txt. Scripts generating this data are made available as Supplementary Material with the bioRxiv manuscript (and as Datasets 2 and 3 in the published work).

Within PROCESSED.DATA.txt file, each row in these files represents the summary of replicate measurements of a single genotype with:

STRAIN: a unique name for the individual line of yeast

CONDITION: a description of the condition used to isolate the line. EMS indicates a mutagenized cell randomly isolated without respect to fluorescence level. SHAM indicates an unmutagenized cell with a reporter construct matching the assay. TDH3.SHAM indicates an unmutagenized cell carrying a P-TDH3-YFP promoter construct for standardizing among assays. EMS.LOW and EMS.HIGH cells were isolated by fluorescence-activated cell sorting because they had a phenotype more extreme than the 2% or 98% of the control unmutagenized (sham) distribution.

ASSAY: the experiment in which data was collected for this genotype

PLATE: indicates the 96-well plate containing the genotype

CTRL.GENO: indicates the genotype of the strain used to infer plate effects for this sample

CONTAM.CHECK: data quality check variable, should always be NONE for inclusion

SORT/SCORE/EXP: variables indicating the experiment, now redundant

N: represents the number of independent replicate measurements contributing to the estimate of expression described by the row

P.VAL.MEDIAN: tests the hypothesis that this strain differs from the sham population

Format of remaining variables:

“Summary statistic applied to replicates” + “Single cell feature quantified within a clonal population of cells measured on the flow cytometer” + “Scaling of metric” (e.g. ABS: on the absolute instrument scale, WT: scaled relative to the median of the promoter-matched unmutagenized phenotype, TDH3: scaled relative to the median of the TDH3 sham controls included on every plate)

For example:

MEAN.YFP.MEDIAN.ABS: represents the mean of replicate measurements of the genotype describing the median adjusted fluorescence on an arbitrary fluorescence scale (after processing of raw data described in SI Appendix)

MEAN.YFP.MEDIAN.WT: represents the mean of replicate measurements of the genotype describing the median adjusted fluorescence score scaled to be relative to the wild type YFP level (defined as the median of the sham genotypes)

MEAN.YFP.MEDIAN.TDH3: represents the mean of replicate measurements of the genotype describing the median adjusted fluorescence score scaled to be relative to the un-mutagenized TDH3 expression level (defined as the median of the TDH3 sham genotypes)

MEAN.FSC.CORRECT: represents the mean of replicate measurements of the genotype describing the median corrected cell size as measured by forward scattered area (logFSC.A)

SD.YFP.MEDIAN.ABS: represents the MAD of replicate measurements of the genotype describing the median adjusted fluorescence on an arbitrary fluorescence scale (after processing of raw data described in SI Appendix)

SD.YFP.MEDIAN.WT: represents the MAD of replicate measurements of the genotype describing the median adjusted fluorescence score scaled to be relative to the wild type YFP level (defined as the median of the sham genotypes)

SD.YFP.MEDIAN.TDH3: represents the MAD of replicate measurements of the genotype describing the median adjusted fluorescence score scaled to be relative to the un-mutagenized TDH3 expression level (defined as the median of the TDH3 sham genotypes)

SD.FSC.CORRECT: represents the MAD of replicate measurements of the genotype describing the median corrected cell size as measured by forward scattered area (logFSC.A).