Comparing the utility of *in vivo* transposon mutagenesis approaches in yeast species to infer gene essentiality

Anton Levitan[1,2], Andrew N. Gale[3], Emma K. Dallon[3], Darby W. Kozan[3], Kyle W. Cunningham[3], Roded Sharan[4] and Judith Berman[1]

[1]School of Molecular Microbiology and Biotechnology, Faculty of Life Sciences, Tel Aviv University, Tel Aviv, Israel

[2]Edmond J. Safra Center for Bioinformatics, Tel Aviv University, Tel Aviv, Israel

[3]Department of Biology, Johns Hopkins University, Baltimore, MD, USA

[4]Blavatnik School of Computer Science, Tel Aviv University, Tel Aviv, Israel

ABSTRACT

Background:

In vivo transposon mutagenesis coupled with deep sequencing enables large-scale genome-wide mutant screens for genes essential in different growth conditions. Six large scale studies have now been performed with three yeast species (*S. cerevisiae, S. pombe* and *C. albicans*), each mutagenized with two of three different heterologous transposons (*AcDs*, *Hermes,* and *PiggyBac*).

Results: We analyzed predictions of gene essentiality for each of the six studies and evaluated the ability of the data to predict gene essentiality using a machine-learning approach. Important data features included a sufficient number of independent insertions and the degree of random insertion distribution. All transposons showed some bias in insertion site preference, both because of jackpot events, specific insertion sequence preferences and preferences for short-range vs long range insertions. For *PiggyBac*, a stringent target sequence limited the ability to predict essentiality in genes with few or no target sequences. Furthermore, the machine learning approach is robust for predicting gene function in less well-studied species by leveraging cross-species orthologs. Finally, comparisons of isogenic diploid vs haploid *S. cerevisiae* isolates identified several genes that are haplo-insufficient, while most essential genes, as expected, were recessive.

Conclusions: We provide recommendations for the choice of transposons and the inference of gene essentiality in genome-wide studies of eukaryotic microbes such as yeasts, including species thathat have been less amenable to classical genetic studies. These include maximizing the number of unique insertions, avoiding transposons with stringent target sequences and a method for cross-species transfer learning.

INTRODUCTION

Work with model yeasts such as *Saccharomyces cerevisiae* and *S. pombe* has pioneered the combination of genetype/phenotype comparisons at a genomic scale. For these yeasts, with genome sequences available for over 20 years[1, 2] and facile gene replacement protocols, have relied heavily on comprehensive collections of deletion mutants[3, 4] for high throughput dissection of specific genotypes as well as for genetic interactions with drugs (reviewed in Lehár et al.[5]) and for gene-gene interactions through systematic analysis of double and triple mutant analysis (e.g., Reguly et al.[6], Kuzmin et al.[7]). In animals and plants that are less amenable to such directed molecular manipulations, the use of heterologous transposons *in vivo* has facilitated genetic analysis, within the limitations imposed by the transposon excision/insertion process[8, 9]. With the advent of deep sequencing, such studies have also become more facile and have been performed in the two model yeasts as well[10, 11].

In vivo transposon mutagenesis generally involves the introduction of a heterologous DNA transposon, along with the genes (e.g., the relevant transposase) required to induce its active transposition into a clonal isolate of a species of interest. Upon induction, the transposase excises the transposon from its original location (excision site) and inserts it into a single new position in the genome. Each cell harbors, at most, a single transposition mutation because the frequency of transposon excision and reinsertion is quite low. The transposon is usually engineered for facile selection of excision and/or reinsertion events, allowing detection and enrichment of these rare events.

In vivo transposon insertion provides several advantages, as it rapidly yields large numbers of mutants in a single step and easily can be performed in parallel strains with different mutations or genetic backgrounds. Because it does not require much prior knowledge, it can also be performed in non-model species, where each experiment is likely to be highly informative. The only transformation steps required are those used to engineer the starting strain. This bypassing the problem of low transformation efficiency in many species. It also avoids the unintended genome alterations (e.g., aneuploidies) that often accompany DNA transformation[12]. The sites of transposon insertion throughout the genome can be identified *en masse* using very large collections of independent insertion event clones, coupled with deep sequencing of the DNA immediately adjacent to the new transposon locus. For example, a high throughput genotype/phenotype analysis of 30 bacterial species grown under >170 different nutrient and stress conditions recently assigned functions to

98thousands of genes including ~300-600 genes per bacterium that are essential
99for viability[13].
100
101Three different transposon systems have been used for *in vivo* mutagenesis in
102yeasts: *AcDs from Zea mays, Hermes from Musca domestica and PiggyBac,*
103*from Trichoplusia ni*. *AcDs* has been used primarily in plant species but was
104also engineered for increased efficiency in the model yeast *S. cerevisiae*[14]
105and, later, in *C. albicans*[15]. *AcDs* does not display any insertion sequence
106preference, although it has a higher frequency of insertions into intergenic
107regions than coding regions and has a bias for reinsertion near the initial site of
108excision. *Hermes* has been used for mutagenesis in *S. pombe* and *S.*
109*cerevisiae*[17, 18]; it prefers to insert at genomic positions with the target
110sequence TnnnnA. *PiggyBac* (PB) has been used in mammalian systems such
111as rat, mouse[19] and also in *S. pombe*[17]. *PB* has a strong preference for
112insertion at TTAA sequences, which are generally more frequent in A-T-rich
113intergenic regions than within coding sequences.
114
115Transposon insertion within an ORF is generally assumed to cause a loss-of-
116function mutation. Identifying the phenotypes associated with loss-of-function
117mutations in specific genes allows the prediction of genetic functions. Cells in
118which the transposon inserted into a gene essential for viability will fail to grow
119and thus be lost from the population. By contrast, cells with mutations in non-
120essential genes are expected to be well-represented in the cell population.
121Insertion of a transposon carrying a strong promoter into an ORF could activate

expression inappropriately; can be useful for the study of gain-of-function mutations.

In vivo transposon mutagenesis studies of yeasts include analysis of *S. cerevisiae* with *Hermes* (this study) or with the mini-Ds derivative of the *AcDs* system[10], in *S. pombe* with *Hermes*[11] and *PB*[20] and in *C. albicans* with *AcDs*[21] and with *PB*[22] (Fig. 1A). In earlier work, we applied a machine learning (ML) approach to infer gene essentiality from the *C. albicans AcDs* data. Here, we modified the ML approach to predict the likelihood of essentiality for the complete set of predicted open reading frames these sixth *in vivo* transposon datasets. We compared the strengths and challenges of the different transposons in each species, with the goal of reaching insights concerning the number of insertion events required for accurate predictions, the distribution of mutations, and the degree to which different transposons, with different sequence dependencies, provided similar or different conclusions. The goal was to provide metrics that assist in determining the advantages and disadvantages of different transposon systems so as to optimize the data produced in a given *in vivo* transposon system and to suggest approaches for generating whole genome data in understudied yeast species.

RESULTS AND DISCUSSION

A comparative analysis of the transposon mutagenesis studies

147We compared six *in vivo* transposon insertion mutagenesis experiments,
148produced using three different heterologous transposons (*AcDs, Hermes* and
149*PiggyBac*) in three different species (*S. cerevisiae, S. pombe and C. albicans*).
150Details of the datasets are provided in the methods section and relevant
151parameters are highlighted in Table 1. The number of transposition events
152detected in the different studies varied considerably, from over 500,000 unique
153insertion sites (hits) and 84 M total reads for the *C. albicans AcDs* (*CaAcDs),* to
154as few as 37,500 unique hits and 6.1 M reads in the *S. pombe PiggyBac* (*SpPB*)
155data set (Table 1). The number of reads per hit also varied considerably, from
156 41 to 170.
157
158**Overview of ML approach for gene essentiality prediction**
159
160We first mapped the hit and read frequency of the transposons in each of the
161three reference genomes (Fig. 1C). Sites of transposon insertion were identified
162based upon targeted sequencing of regions adjacent to the inserted
163transposon. Slightly different sequencing protocols were used in the different
164studies, but all six essentially amplified Tn-adjacent sequences and mapped
165them.
166Theoretically, essential genes have no hits and non-essential genes have many
167hits; however, distinguishing essential and non-essential genes from the data is
168not entirely straight-forward (e.g., Fig. 1C, Gene X_n). To address this
169ambiguity, we extended a previous approach for gene essentiality
170prediction[21]. We first chose input features from transposon data that were
171likely to be informative in the essential/nonessential decision: the number of

172unique insertion sites (hits) per ORF, the degree to which those insertion sites
173were enriched in the population (reads), as well as normalization factors that
174consider the insertion frequency as a function of chromosome position.
175
176Training sets were built using information from the two model species with
177gene essentiality data available from classical genetic approaches (e.g., com-
178prehensive ORF deletion analysis) (reviewed in Giaever and Nislow[23] and
179Spirek et al.[24]) (Table S1 (training sets)). The specificity and sensitivity of the
180approach was analyzed using the AUC (area under the receiver operating char-
181acteristic curve (Fig. 1D). For *C. albicans,* which did not have extensive prior
182knowledge of gene essentiality, we constructed a training set from a core set of
183genes whose orthologs were known to be essential in both model yeasts. This
184approach is likely to be useful for other species that lack sufficient prior knowl-
185edge of gene essentiality to construct a within-species training set.
186
187We assessed the performance of each classifier by producing training sets
188using genes known or inferred to be essential and non-essential, as described
189in the Methods. The classifiers were trained and their performance, assessed
190using the AUC measure, was high across most examined studies (>0.94). The
191one exception was the *SpPB* study, which had far fewer unique insertion sites
192(Table 1) and had an AUC of 0.785. The highest AUC levels were seen with the
193*AcDs* in both *C. albicans* and *S. cerevisiae*. Of note, these two studies also had
194the largest number of total hits and reads. All the considered ML features for
195each ORF in every study and the predicted verdicts of essentiality are provided

in Tables S2-S7. Below, we describe the main insights gained from this comparison.

Insight #1: Optimize the number of independent insertion sites (hits) for highest quality predictions of gene essentiality

The total number of unique insertion sites (hits) and the performance (AUC values) were high highly correlated, and this correlation was statistically significant (Fig. 2a; Pearson's r = 0.892; p-value = 0.0169). By contrast, the total number of sequencing reads showed a weaker correlation with the AUC that was not statistically significant (Fig, 2b; Pearson's r = 0.636; p-value = 0.1741). If we disregard the worst performing *SpPB*, the correlation of the AUCs with the total number of hits rises dramatically to Pearson's r = 0.995; p-value = 0.0003, and the correlation with the total number of sequencing reads remains similarly weak: Pearson's r = 0.652; p-value = 0.2327. Thus, it a library with many independent hits will improve performance and simply increasing the number of sequencing reads is not likely to be sufficient to obtain optimal results. Increasing the number of independent hits requires collection of sufficient numbers of independent colonies soon after transposase induction and the resulting transposon excision and reinsertion. An advantage of *Hermes* is that most insertions occur during stationary phase, so transposase-inducing conditions can be tolerated throughout the growth period. Experimental designs that optimize isolation of independent events are critical.

Insight #2: Avoid libraries with high levels of jackpot events

Jackpot events are the appearance of extraordinarily high numbers of reads in a very small number of insertion sites. When the number of reads greatly exceeds the theoretical number of cell divisions in the experiment, this is likely due to a transposition event happened prior to the induction of transposon excision in the experiment. Jackpot events are a major pitfall in that much sequencing capacity is wasted on detection of a single insertion site. Jackpot events with >1M sequencing reads were present in 4 of the 6 data sets; *SpPB* and *ScHermes* had no major jackpot events (no hits with ≥1000-fold more reads than the average read/hit) (Table 1). Both of these libraries also had far fewer total sequencing reads than the other studies (6M and 18M vs 24-84M for the other libraries).

Of note, within a data set, some individual experiments had jackpot events that were far more than others (Table 1), which would be expected if jackpot events occur stochastically. Importantly, jackpot events were not clearly associated with one of the three species or with the transposon type. This suggests that jackpots arise from technical, rather than biological issues.

Avoiding jackpot events is important because they reduce data quality considerably: the higher the number of reads at a few jackpot sites, the lower the number of informative hits and reads. Avoiding the selection of cells in which a transposon was already mobilized is key to ensuring that the number of hits and reads provide good genome coverage. Dividing the cultures into

dozens of small cultures and then re-pooling these sub-cultures after transposase induction can effectively dilute out most jackpot events. Preparing several independent libraries and sampling a few sequences in each may also be worthwhile. For example, if a tested library shows one sequence twice in one hundred colonies, it is likely to be a >1M jackpot event.

Insight #3. Consider which features are most important in the analysis of a given transposon

In decision tree based algorithms, such as Random Forest[25], every node is a condition of a split of the data by a single feature. The splitting process continues until it reaches a stop condition such as: all the features have been used, the obtained subset is very small or the training labels are the same for the obtained subset. The goal is to reduce entropy (uncertainty) in the data. Entropy is zero when all the labels in the obtained subset are the same; and is maximum when half of the labels are the same in the obtained subset (in a binary classification). Each split of the data by a given feature (node in the tree) reduces the entropy. The importance of a given feature in the Random Forrest classifier is the calculated decrease in entropy contributed by that feature. Here we describe the features of the classifiers, and discuss their relative importance.

For each ORF, we calculated predictive features including the number of hits, number of reads and the length of each ORF, a neighborhood index, which

271normalizes for insertion bias due to genomic position (e.g., proximity to the
272initial excision site in the genome) and a freedom index, which reports the
273proportion of an ORF that is hit-free. The freedom index is especially useful for
274identifying genes with essential domains, that are able to tolerate insertions
275outside of the essential domain. The number of the transposon hits per
276transposon target sequences in an ORF was an additional feature used in the
277analysis (Figure 3), where applicable (in *PB* and *Hermes* studies). Furthermore,
278we calculated the number of hits and the number of reads normalized by the
279length of each ORF. We compared the 'feature importance' for each library to
280ask whether specific features were more important for the different classifiers
281and whether feature importance was characteristic for a given transposon or
282yeast species.
283
284The number of hits per ORF played an important role in determining
285essentiality, with essential genes having far fewer hits than non-essential ORFs
286(~7 times less, on average, across the 6 datasets), consistent with the strong
287correlation between number of hits and the AUC (Fig. 2A). The number of reads
288per ORF played a lesser role in these classifications, also consistent with the
289correlation above (Fig. 2B). Gene length also affected the probability of
290transposon insertion in a gene, and thus was a crucial normalization parameter
291for the numbers of hits and reads for every ORF.
292
293The Neighborhood Index (NI) feature made important contributions in all of the
294classifications (except *SpPB*, which had far less data). Importantly, the NI did
295not differ considerably between the different transposons, consistent with the

idea that chromatin accessibility, 3D chromosome organization and other factors that may bias the insertion site frequency in a given organism affect the frequency of insertion of different transposons in a similar manner.

The Freedom Index (FI) was a major contributor to both *ScAcDs* and *CaAcDs* predictions while results with the *PB* and *Hermes* datasets were mixed (Fig. 3). This is consistent with the idea that *AcDs* does not have a specific target sequence and thus inserts throughout ORFs, while *PB* and *Hermes* have fewer target sequences within ORFs. Thus, the FI more important in *AcDs* experiments because hits occur more randomly throughout an ORF.

The importance of the number of hits in the proximal regulatory sequences (100 bp upstream to the start codon) to the essential/non-essential predictions was only minor, but is highly variable. For example, the impact of *ScHermes* was nearly twice that of *ScAcDs* for this feature. As described further below, we posit that this difference is due to cryptic enhancer/promoter activity in the miniDs transposon in *S. cerevisiae* that not seen with the *Hermes* transposon.

Insight #4: Consider the effect of transposon-specific target sequence specificity

Some transposons have preferred sites of insertion: *Hermes* prefers TnnnnA and *PiggyBac* inserts primarily at TTAA sequences; *AcDs* does not have an insertion site preference. Theoretically, the length of the insertion site sequence necessarily scales inversely with the number of potential unique hit

321sites. However, it was not clear at what insertion sequence length the
322resolution of studies of gene essentiality becomes limiting.
323
324The feature importance of the number of hits per transposon target sequence
325in an ORF, which should be a measure of library saturation, showed a varying
326degree of importance in the *PB* and *Hermes* studies. Curiously, its importance
327wasn't dependent on the type of the transposon or the target sequence
328prevalence in the genome. This likely because target sequences are preferred
329sites of insertion, yet are not exclusive or absolute. For example, *PiggyBac* in
330*C. albicans* had 1.6-fold more unique insertion sites than the theoretical
331number of target sequences in the *C. albicans* genome. By contrast, for both
332*ScHermes* and *SpHermes*, the number of target sequences available far
333outnumbered the number of unique hits. The proportion of target sequences
334not hit ranged from 8.9% for *CaPB* to 85% for *SpHermes* (Table 1) and the
335proportion of hits not in target sequences ranged from 14% in *SpPB* to ~50% in
336*CaPB* as well as both *Hermes* data sets. Thus, we surmise that the preference
337for target sequences is only a minor limitation for both of these transposons,
338except when the total number of hits is very low as in *SpPB*.
339
340Another critical issue is the number of genes that lack any preferred target
341sequences within the ORF; there are 228 and 185 ORFs without a single TTAA
342sequence in *C. albicans* and *S. pombe,* respectively. These ORFS have a lower
343probability of acquiring insertions and, if the genes are non-essential, they are
344much more likely to give false positive information (be predicted essential for
345lack of insertions). Indeed, 155 ORFs without TTAA sequences were predicted

346essential in the *CaPB* data and yet are predicted non-essential in the *CaAcDs*
347study. Similarly, 118 of the 185 ORFs lacking TTAA sequences were predicted
348essential from the *SpPB*study, but were non-essential in the *SpHermes* study.
349We assume that many of these genes are false positives, especially given that
350127 of the 185 ORFs lacking TTAA, including 95 of the 118 aforementioned
351ORFs, were non-essential in classical genetics studies of *S. pombe.*
352
353
354Next, we asked if the number of target sequences within an ORF affected the
355*CaPB* classification performance for that ORF. To address this, we compared
356the performance (AUC) to sets of ORFs filtered to exclude ORFs with different
357numbers of target sequences (from 0 to 10) from the training set used to train
358the classifier (Fig. 4). The AUC increased from ~0.94 for the entire training set
359to >0.98 for the training set containing only genes with 10 or more target sites
360(~50% of the genes in the training set). This suggests that studies using the
361*PiggyBac* transposon may struggle to correctly infer gene essentiality for ORFs
362with low numbers of target sites.
363
364**Insight #5. Consider whether the transposon can activate as well as**
365**disrupt gene expression.**
366
367The prediction of essentiality was based upon the assumption that transposon
368insertion into an ORF disrupted gene expression and produced loss-of-function
369allele. However, this is not necessarily the case for all genes. For example, if
370an insertion allele removes a regulatory domain from a protein, the protein

371may become hyperactive, and thus the result would be a gain-of-function
372allele. Additionally, some transposons may introduce enhancer and promoter
373activities that could increase gene expression in some species. The miniDs
374transposon used in the *ScAcDs* data is likely to contain such activities.[10]
375Consistent with this idea, the *ScAcDs* dataset contains an average of 1.89
376insertions within the first 100 bp of essential genes whereas the other datasets
377including *CaAcDs,* which has a transposon modified from the *miniDs,* contain
378significantly fewer (0.82 insertions in the first 100 bp of Ess genes).
379Additionally, many essential genes of *S. cerevisiae* appeared to tolerate miniDs
380insertions, but not *Hermes* insertions, at sites in the 5’ UTR that are very close
381to the start codon. Thus, the *miniDs* transposon in *S. cerevisiae* may facilitate
382inappropriate activation of gene expression when inserted upstream or within
383certain genes.
384
385**Cross-study analysis**
386
387Knowing the full set of essential and non-essential genes in eukaryotic
388microbes, including pathogens of humans, animals and plants, will improve our
389understanding of common and species-specific properties of these
390understudied organisms. Furthermore, once a transposon library has been
391collected, it can be screened under many other growth conditions to reveal
392genotype/phenotype relationships. *In vivo* transposon analysis of gene
393essentiality is a practical and feasible approach, because the cost in time and
394resources for obtaining libraries is far lower than that for producing engineered
395deletion mutants, especially given that the amount of baseline information

396(other than the genome sequence) about the organisms may be minimal. The
397only technical hurdle is to introduce the heterologous transposon of interest,
398either on a plasmid (where feasible) or into a useful locus within the genome of
399the relevant organism.
400
401An additional challenge is that ML approaches require a high-quality training
402dataset (of gold-standard essential and non-essential genes). For many non-
403model organisms, such training data is too sparse to build a robust training set.
404For *C. albicans,* we circumvented the low numbers of genes already known to
405be essential by relying upon genes that had been determined to be essential
406from comprehensive classical genetic deletion studies in both model yeasts (*S.*
407*cerevisiae* and *S. pombe)* and that had orthologs in *C. albicans.* Training on *S.*
408*cerevisiae* or *S. pombe* orthologs with consistently essential orthologs yielded
409good performance predictions for *C. albicans* (AUC: 0.940 to 0.993, Table 1).
410*CaAcDs* performance was lower when training only on the 66 genes known to
411be essential plus the set of presumed non-essential genes (those that had been
412successfully deleted in *C. albicans* studies, AUC of ~0.92)[21].
413
414Next, we considered the quality of the learning performance for each dataset, if
415we trained on orthologs from one species and predicted essentiality of genes in
416a different organism (Figure 5). The transfer learning performance of the
417classifications was of a comparable quality to the single study classifiers for
418most *AcDs* and *Hermes* cases (Figure 5a and Figure 2). Furthermore, it
419displayed a somewhat symmetrical property: in most cases, there were minor
420differences in performance between train/test and test/train pairs (reducing the

421quality by ~ 0.5% to 5.7%) when the tests were between or among *AcDs* and
422*Hermes* experiments. Conversely, when testing for predictions from *PB* data
423that were trained on either *AcDs* or *Hermes,* the AUCs dropped more
424dramatically (up to ~21.9%). Thus, *PB* data was less transferable than the
425*Hermes* and *AcDs* data.
426
427The low *PB* transferability between *SpPB* and *CaPB* is likely due to the sparser
428target sequence distribution relative to either *Hermes* or *AcDs*, which causes
429*PB* studies to produce false positives as noted above, and thus might
430contribute to reduced performance in cross-study analyses. The lower
431performance of the classifiers in the original *PB* single studies (Table 1), also
432may have contributed to the reduced ability to predict essentiality in pools of
433*PB* mutants using cross-species models.
434
435Reduced differences cross-study performance could also be due to differences
436between the importance of different features in the classifiers for the different
437datasets. To test this possibility, we correlated the vector of the relative
438feature importance for each study with the feature importance in all the other
439studies (Figure 5b). The analysis distinguished 3 groups within the 6 studies,
440based on the correlation coefficient values for feature importance between
441members of the group: *CaAcDs* and *ScAcDs* (Pearson's r = 0.902); *Sp Hermes*
442and *CaPB* (Pearson's r = 0.898); and *SpPB* and *ScHermes* (Pearson's r =
443 0.929). Notably, the quality of the transfer learning predictions appears to be
444independent of both the transposon type and the organism studied, with the

445exception of the *AcDs* studies. We presume that this is due to the lack of a
446specific target sequence for the *AcDs* transposon system.
447
448**Insight #6: As necessary, construct training sets using genes with**
449**orthologs in models where essentiality is known and then validate the**
450**training set manually.**
451
452We suggest that an initial training set of orthologous genes known to be
453essential and non-essential in related model organisms can be used to
454facilitate analysis of a transposon insertion study in a non-model organism with
455sparse essentiality information. An important caveat is that differences
456between gene function in different species can alter gene essentiality of a
457small number of these orthologs; thus, it is important to visually inspect this
458orthologous training set before applying it. The goal is to remove any genes
459with insertion patterns that are highly contradictory to the 'essentiality label'
460that the orthologs provided. For example, for *C. albicans*, the entire
461orthologous training set was reviewed in an unprejudiced fashion by three
462independent inspectors, who visually reviewed the insertion patterns in the
463*CaAcDs* data and manually labeled each gene as essential, non-essential or
464ambiguous. When all three inspectors classified a gene as non-essential (e.g.,
465many insertions throughout an ORF within a genome region that had many
466insertions outside of that ORF) and the orthologs were labeled 'essential' in the
467two model yeasts, we removed that gene from the training set.
468

469Once a training set has been established, and the features for the ORFs have
470been calculated, the Random Forest classifier can be run in a cross-validation
471scheme and the AUC can be calculated using the essentiality labels. This
472provides an efficient approach to obtain information about all of the genes in a
473species that has been sequenced but not subjected to much molecular
474manipulation. Clearly, the same approach can be used to compare the
475essentiality of the same sets of genes grown in different conditions as well,
476potentially providing large amounts of phenotypic data across an entire set of
477ORFs. If applied to a species that had not been the subject of many genetic
478studies, such data would represent a treasure-trove of information about genes
479that had not been previously studied and the phenotypes associated with loss-
480of-function of those genes.
481
482**A Comparative Analysis of Gene Essentiality Predictions**
483
484An important issue is whether different transposon insertion studies in the
485same organism had similar or different predictions from one another and from
486the known essentiality status of deletion mutants, which are by definition 'loss-
487of-function' null alleles. For *S. cerevisiae,* the classifiers displayed a high
488degree of agreement on the final verdicts of gene essentiality (Figure 6a),
489while more discrepancies were evident for the *C. albicans* and *S. pombe*
490studies. Both *PiggyBac* studies predicted a much higher number of essential
491genes than the *AcDs* or *Hermes* studies (Figure 6b and 6c) as expected from
492the paucity of target sequences that are likely to give false positive predictions
493discussed above. For example, the *CaPB* study had an average 5.84 target

494sites per kb in genes likely to be false positives vs. 10.62 target sites per kb in
495all the genes (Mann Whitney U: p-value < $2.38*e^{-78}$). Importantly, when
496compared to the set of essential genes for each species determined by deletion
497analysis, the transposon studies also did quite well, with only 20 to 35% of the
498genes in disagreement. In some cases, such discrepancies were found to be
499due to issues with the deletion collection isolates. For example, ~8% of the
500original *S. cerevisiae* deletion collection carried aneuploidies or gene
501amplifications,[26] and ~10% of *S. pombe* deletion strains retained a wild-type
502copy of the ORF that had been targeted for deletion. Extra copies of the
503'deleted' gene reduces the apparent number of essential genes.
504
505**Gene essentiality in haploid versus diploid strains of *S. cerevisiae***
506
507*S. cerevisiae* is readily grown in both the diploid and haploid states, which
508allows identification of the haplo-insufficient subset of genes among the set of
509essential genes. Based on gene knockout studies, only 2 genes (*NDC1*, *MLC1*)
510were classified as haplo-insufficient,[27, 28] while all other essential genes
511were haplo-proficient (i.e. heterozygous knockouts in diploids were viable). To
512determine whether additional haplo-insufficient genes exist in *S. cerevisiae*, we
513collected *ScHermes* insertions in diploid strain BY4743 and compared them to
514haploid strains BY4741 and BY4742. *Sc Hermes* transposon mutagenesis
515libraries were used with the classifier that had been trained on the *SpHermes*
516haploid training set data, applying the same threshold for classification (Figure
5178). The classifier identified 155 genes as "essential in both haploid and
518diploid", a number far higher than expected. Upon closer analysis, 98

contained regions of poor mapping due to duplications elsewhere in the genome, 50 were categorized as dubious in the Saccharomyces Genome Database (yeastgenome.org), one (*LEU2*) had been deleted in the strains studied, and the two known haploinsufficient genes (*NDC1, MLC1*) had been identified, providing support for this approach. Upon visual inspection of data of the remaining four genes, one essential gene (*BCY1*) appeared haploinsufficient, whereas another (*RPC10*) contained numerous insertions in its 5' UTR in diploids but not haploids, suggesting that it might not be essential (Fig. 9). The other two ORFs predicted to be haploinsufficient are very small (165-225 bp) and also are within regions of sparse insertion density. Thus, we have lower confidence in the data for these two genes. Thus, transposon mutagenesis of a diploid strain successfully revealed the two known haploinsufficient genes and one new one (*BCY1*), which is known to be essential in the conditions employed in the screen, but not essential in other culture conditions[29].

The classifier also identified 29 genes as "haploinsufficient" in diploids and not essential in haploids. Of these, 20 could be dismissed based on their annotation as dubious or the presence of duplicated (unmappable) segments. All of the remaining 9 genes were small (87-528 bp) and were found in regions of sparse insertion density. These genes are annotated in SGD as non-essential and are probably false positives.

These observations raise an important issue about data quality control. It is important to filter dubious and uninformative ORFs from the data set (as was

544done in the analysis of *CaAcDs*[21]. This includes genes with repeated domains
545or duplicate copies in the genome that prevent unambiguous mapping of short
546Illumina reads. Furthermore, predicting the essential/non-essential status for
547short ORFs and especially those located in regions with sparse intergenic
548insertions is more likely to be problematic.
549
550**Insight #7: Prediction quality increases considerably when**
551**uninformative data such as mitochondrial genome sequences and**
552**duplicated genes that are difficult to map.**
553
554**In summary,** we suggest a number of metrics for the inference of gene
555essentiality using *in vivo* transposon mutagenesis studies in yeasts, including
556those with little available genetic data. Maximizing the total number of unique
557transposon insertions is the most critical factor in achieving optimal
558performance of the classification. It can be attained by collecting many
559independent insertion clones, striving to reduce the possible jackpot events in
560the study, maximizing the depth of the coverage and by utilizing a transposon
561with a fairly permissive target sequence or no preferred target sequence.
562Furthermore, while transposons with relatively stringent target sequences have
563some advantages for screens that identify individual mutants, for determining
564gene essentiality they are less robust, as the low number of potential target
565sequences, and especially the lack of any target sequences, in certain ORFs
566will increase the likelihood of falsely classifying non-essential genes as
567essential. Additionally, we think that transposon mutagenesis is an ideal
568approach to gain large amounts of useful genotype/phenotype data

understudied organisms: the cross-species learning methodology allows inference of gene essentiality based on conserved orthologs, especially when coupled with visual screening of the data. Finally, *in vivo* transposon mutagenesis is an incredibly useful tool for high throughout genomic studies, not only of gene essentiality *per se,* but also of genes required under specific selective conditions. We hope that the recommendations provided here will facilitate future work to understand gene in a wide range of yeast species.

576**MATERIALS AND METHODS**

577**Data Acquisition**

578**Experimental**

579*Sc Hermes* data was obtained as follows: The haploid and diploid strains of *S.*
580*cerevisiae* were transformed with plasmid pSG36.[30] A single colony was
581suspended in 100 mL synthetic complete (SC) medium lacking uracil and
582containing 2% galactose, divided into twenty 16 x 150 mm glass culture tubes,
583and shaken for 3 days at 30°C. This protocol yielded ~5 x 10^6 cells bearing
584transposon insertions per mL (~3% of all cells). To enrich for cells bearing
585transposon insertions, the twenty cultures were pooled, centrifuged, and the
586cell pellet was resuspended in 600 mL SC medium containing 2% glucose, 0.1
587mg/mL nourseothricin, and 1 mg/mL 5-fluoroorotic acid, and then shaken
588overnight at 30°C. The cells were pelleted, resuspended in 600 mL of the same
589medium, and cultured as before. Finally, 60 mL of these enriched cells were
590pelleted, resuspended in 600 mL of the same medium, and cultured as before.
591These highly enriched cells were pelleted, resuspended in 15% glycerol, and
592frozen in aliquots at -80°C. To extract genomic DNA, 100 mg of thawed cell
593pellets were washed three times in 1 mL deionized water and extracted using
594Quick-DNA Fungal/Bacterial Miniprep kit (Zymo Research). A total of 2.4 µg of
595purified gDNA was fragmented by sonication in four separate tubes using a
596Diagenode Picoruptor. The fragmented DNA was then end repaired, ligated to
597splinkerette adapters, size selected with AMPure xp beads, and PCR amplified
598in separate reactions using transposon-specific and adapter-specific primers as
599detailed previously.[31] Samples were then PCR amplified to attach Illumina P5
600and P7 (indexed) adapters, purified with AMPure xp beads, mixed with phiX-

601174, loaded into MiSeq instrument (Illumina) and 75 bp of each end was
602sequenced using primers specific for Hermes right inverted repeat and P7.
603Detailed protocols and primer sequences are available upon request. De-
604multiplexed reads were mapped to the *S. cerevisiae* S288C reference genome
605using Bowtie2, and any mapped reads with a quality score < 20 or a mismatch
606at nucleotide +1 were removed. This process was repeated a total of 3 times in
607diploid strain BY4743, 2 times in haploid strain BY4741, and 1 time in haploid
608strain BY4742. The diploid and haploid datasets were combined prior to
609analyses. The *S.cerevisiae Hermes* data (mapped reads and counts) are
610available at http://genome-euro.ucsc.edu/s/CunninghamLab/Hermes%20Vs
611%20AcDs . FastQ files are available from Sequence Read Archive (SRA) and
612the ArrayExpress Experiment Archive (ArrayExpress), which are core
613repositories of the European Nucleotide Archive (ENA) at accession number
614[XXXX] https://www.ebi.ac.uk/ena/about/data-repositories (to be updated when
615accession number is issued).
616
617
618
619**Publicly available databases**
620The rest of the datasets analyzed here were obtained from previously
621published studies. *ScAcDs* was published by Michel et al, 2017[10], from which
622both WT1 & WT2 were combined for the analysis. The data was downloaded:
623https://www.ebi.ac.uk/arrayexpress/experiments/E-MTAB-4885/samples/. *SpPB*
624was published by Li et al, 2011[20] and the data was obtained from the SRA
625database: SRR089408. *Sp Hermes* was published by Guo et al, 2013[22] and

the data was obtained from the SRA database: SRR327340. *CaAcDs* was published by Segal et al, 2018[21] and the data was obtained from the SRA database, where SRR7824843, SRR7824841 and SRR7824838 files, were combined for the analysis. *CaPB* was published by Gao et al, 2018[22] and the data was obtained from the SRA database, where all the following files were for the analysis: DMSO (untreatment): SRR7704188, SRR7704193, SRR7704196; 5-FOA (untreatment: SRR7704189, SRR7704194, SRR7704200; No drug screen: SRR7704195. All the SRR files were obtained using fastq-dump, with the following bash command: fastq-dump --gzip --skip-technical --readids --dumpbase --split-files --clip <SRR*******>

Data Processing

The .fastq files downloaded with fastq-dump, were processed using cutadapt to filter out reads not containing partial transposon sequences. Reads with transposon sequences were trimmed to remove the transposon sequences for alignment purposes, as follows: cutadapt --cores=8 -m 2 -g <primer sequence> <input fastq filename> -o <output fastq filename> --discard-untrimmed --overlap <overlap length>. In the analysis of the *Sc Hermes* study all the sequencing reads contained the transposon and the reads start at the first genomic base, thus required no filtering. In the analysis of the *Sp PiggyBac* we filtered the reads containing the transposons from the rest, by identifying the ACGCAGACTATCTTTCTAGGG sequence, cutting it out and aligning only the remaining part of the relevant reads. In the analysis of the *Ca AcDs* we filtered the reads containing the transposons from the rest, by identifying the GTATTTTACCGACCGTTACCGACCGTTTTCATCCCTA sequence, cutting it out and

aligning only the remaining part of the relevant reads, starting 37bp downstream (Segal et al, 2018[21]). In the analysis of the *Ca PiggyBac,* we filtered the reads containing the transposons from the rest, by identifying the TGCATGCGTCAATTTTACGCAGACTATCTTTCTA sequence, cutting it out and aligning only the remaining part of the relevant reads, starting 3bp downstream. In the analysis of the *Sc AcDs* study we used the published transposon hitmaps of WT1 and WT2. In the analysis of the *Sp Hermes* study we used the published transposon hitmaps (Segal et al, 2018[21]).

Alignment of reads and mapping the transposon hits

bowtie2 indices were created for each organism and gffutils databases were created for each organism's genetic features, using the latest versions of the reference genomes (fasta) and the genomic feature files (gff), which were downloaded from the respective official sources for the three organisms: *S. cerevisiae* from https://downloads.yeastgenome.org, *S. pombe* from ftp://ftp.pombase.org/pombe/ and *C. albicans* http://www.candidagenome.org/download/. Sequencing reads were aligned using bowtie2 with the default settings. The resulting sam files were converted to bam using samtools. bam files were sorted using samtools and indexed via pysam. Transposon hits and their corresponding reads were mapped to the respective genomes and counted in each genomic feature. Transposon target sites were found in every genome using Biopython and counted in each genetic feature.

Gene essentiality classification

Table 2 summarizes the features for machine learning classification that were engineered from the mapped transposon hits, reads and the transposon target sequences in the genomes. Random Forrest classification was performed, using Python's scikit-learn library with the default parameters, except the n_estimators parameter that was increased to 200, and the random_state parameter was fixed at 0, for reproducibility purposes. The results were validated using a 5-fold cross-validation. Essentiality labels for the training set of each organism were obtained previously (Shiftman et al, 2018) and are provided in Table S1.

Thresholds for the essentiality predictions in each classification were chosen as follows: Two metrics were evaluated (Figure 8): 1) Minimum of the Euclidean distance between (0, 1) and the receiver operating characteristic (ROC) curve. 2) Maximum of the vertical distance between the line describing a random choice (a straight line from (0, 0) to (1, 1)) and the ROC curve. The first method was chosen, and we verified that the second metric is reasonably close, to eliminate any possible artifacts. We predicted the essentiality of all the available genes for each organism based on their respective features, and using the aforementioned method to choose the threshold for each binary classification.

Figures were generated using Python's matplotlib and seaborn libraries. The schematics were drawn using Inkscape. Mann Whitney U p-values and Pearson's correlation coefficients and p-values were calculated using Python's Scipy.

FUNDING ACKNOWLEDGEMENTS

European Research Council Advanced Award 340087 (RAPLODAPT) to J.B., the Israel Science Foundation (grants no. 715/18, 757/12) to R.S. by NIH R21-AI130722 and T32-GM007231 to KWC.

A.L. is supported by a fellowship from the Edmond J. Safra Center for Bioinformatics.

References

1. Goffeau A, Barrell BG, Bussey H, Davis RW, Dujon B, Feldmann H, Galibert F, Hoheisel JD, Jacq C, Johnston M, et al: **Life with 6000 genes.** *Science* 1996, **274:**546, 563-547.
2. Wood V, Gwilliam R, Rajandream MA, Lyne M, Lyne R, Stewart A, Sgouros J, Peat N, Hayles J, Baker S, et al: **The genome sequence of *Schizosaccharomyces pombe*.** *Nature* 2002, **415:**871-880.
3. Giaever G, Chu AM, Ni L, Connelly C, Riles L, Veronneau S, Dow S, Lucau-Danila A, Anderson K, Andre B, et al: **Functional profiling of the *Saccharomyces cerevisiae* genome.** *Nature* 2002, **418:**387-391.
4. Kim DU, Hayles J, Kim D, Wood V, Park HO, Won M, Yoo HS, Duhig T, Nam M, Palmer G, et al: **Analysis of a genome-wide set of gene deletions in the fission yeast *Schizosaccharomyces pombe*.** *Nat Biotechnol* 2010, **28:**617-623.
5. Lehar J, Stockwell BR, Giaever G, Nislow C: **Combination chemical genetics.** *Nat Chem Biol* 2008, **4:**674-681.
6. Reguly T, Breitkreutz A, Boucher L, Breitkreutz BJ, Hon GC, Myers CL, Parsons A, Friesen H, Oughtred R, Tong A, et al: **Comprehensive curation and analysis of global interaction networks in *Saccharomyces cerevisiae*.** *J Biol* 2006, **5:**11.
7. Kuzmin E, VanderSluis B, Wang W, Tan G, Deshpande R, Chen Y, Usaj M, Balint A, Mattiazzi Usaj M, van Leeuwen J, et al: **Systematic analysis of complex genetic interactions.** *Science* 2018, **360**.
8. Munoz-Lopez M, Garcia-Perez JL: **DNA transposons: nature and applications in genomics.** *Curr Genomics* 2010, **11:**115-128.

9. Kawakami K, Largaespada DA, Ivics Z: **Transposons as tools for functional genomics in vertebrate models.** *Trends Genet* 2017, **33:**784-801.

10. Michel AH, Hatakeyama R, Kimmig P, Arter M, Peter M, Matos J, De Virgilio C, Kornmann B: **Functional mapping of yeast genomes by saturated transposition.** *Elife* 2017, **6**.

11. Guo Y, Park JM, Cui B, Humes E, Gangadharan S, Hung S, FitzGerald PC, Hoe KL, Grewal SI, Craig NL, Levin HL: **Integration profiling of gene function with dense maps of transposon integration.** *Genetics* 2013, **195:**599-609.

12. Bouchonville K, Forche A, Tang KE, Selmecki A, Berman J: **Aneuploid chromosomes are highly unstable during DNA transformation of *Candida albicans*.** *Eukaryot Cell* 2009, **8:**1554-1566.

13. Price MN, Wetmore KM, Waters RJ, Callaghan M, Ray J, Liu H, Kuehl JV, Melnyk RA, Lamson JS, Suh Y, et al: **Mutant phenotypes for thousands of bacterial genes of unknown function.** *Nature* 2018, **557:**503-509.

14. Lazarow K, Du ML, Weimer R, Kunze R: **A hyperactive transposase of the maize transposable element activator (Ac).** *Genetics* 2012, **191:**747-756.

15. Mielich K, Shtifman-Segal E, Golz JC, Zeng G, Wang Y, Berman J, Kunze R: **Maize transposable elements Ac/Ds as Insertion mutagenesis tools in *Candida albicans*.** *G3 (Bethesda)* 2018, **8:**1139-1145.

754 16. Betrán E, Long M: **Dntf-2r, a young Drosophila retroposed gene**
755 **with specific male expression under positive Darwinian selection.**
756 *Genetics* 2003, **164:**977-988.
757 17. Park JM, Evertts AG, Levin HL: **The Hermes transposon of *Musca***
758 ***domestica* and its use as a mutagen of *Schizosaccharomyces***
759 ***pombe*.** *Methods* 2009, **49:**243-247.
760 18. Edskes HK, Mukhamedova M, Edskes BK, Wickner RB: **Hermes**
761 **transposon mutagenesis shows [URE3] prion pathology**
762 **prevented by a Ubiquitin-targeting protein: Evidence for carbon/**
763 **nitrogen assimilation cross talk and a second function for Ure2p**
764 **in *Saccharomyces cerevisiae*.** *Genetics* 2018, **209:**789-800.
765 19. Zhao S, Jiang E, Chen S, Gu Y, Shangguan AJ, Lv T, Luo L, Yu Z: **PiggyBac**
766 **transposon vectors: the tools of the human gene encoding.** *Transl*
767 *Lung Cancer Res* 2016, **5:**120-125.
768 20. Li J, Zhang JM, Li X, Suo F, Zhang MJ, Hou W, Han J, Du LL: **A piggyBac**
769 **transposon-based mutagenesis system for the fission yeast**
770 ***Schizosaccharomyces pombe.*** *Nucleic Acids Res* 2011, **39:**e40.
771 21. Segal ES, Gritsenko V, Levitan A, Yadav B, Dror N, Steenwyk JL,
772 Silberberg Y, Mielich K, Rokas A, Gow NAR, et al: **Gene essentiality**
773 **analyzed by in vivo transposon mutagenesis and machine**
774 **learning in a stable haploid isolate of *Candida albicans*.** *MBio*
775 2018, **9**.
776 22. Gao J, Wang H, Li Z, Wong AH, Wang YZ, Guo Y, Lin X, Zeng G, Liu H,
777 Wang Y, Wang J: ***Candida albicans* gains azole resistance by**
778 **altering sphingolipid composition.** *Nat Commun* 2018, **9:**4495.

23. Giaever G, Nislow C: **The yeast deletion collection: a decade of functional genomics.** *Genetics* 2014, **197:**451-465.

24. Spirek M, Benko Z, Carnecka M, Rumpf C, Cipak L, Batova M, Marova I, Nam M, Kim DU, Park HO, et al: ***S. pombe* genome deletion project: an update.** *Cell Cycle* 2010, **9:**2399-2402.

25. Breiman L: **Random Forests.** *Machine Learning* 2001, **45:**5-32.

26. Hughes TR, Roberts CJ, Dai H, Jones AR, Meyer MR, Slade D, Burchard J, Dow S, Ward TR, Kidd MJ, et al: **Widespread aneuploidy revealed by DNA microarray expression profiling.** *Nat Genet* 2000, **25:**333-337.

27. Chial HJ, Giddings TH, Jr., Siewert EA, Hoyt MA, Winey M: **Altered dosage of the *Saccharomyces cerevisiae* spindle pole body duplication gene, NDC1, leads to aneuploidy and polyploidy.** *Proc Natl Acad Sci U S A* 1999, **96:**10200-10205.

28. Stevens RC, Davis TN: **Mlc1p is a light chain for the unconventional myosin Myo2p in *Saccharomyces cerevisiae*.** *J Cell Biol* 1998, **142:**711-722.

29. Matsumoto K, Uno I, Oshima Y, Ishikawa T: **Isolation and characterization of yeast mutants deficient in adenylate cyclase and cAMP-dependent protein kinase.** *Proc Natl Acad Sci U S A* 1982, **79:**2355-2359.

30. Gangadharan S, Mularoni L, Fain-Thornton J, Wheelan SJ, Craig NL: **DNA transposon Hermes inserts into DNA in nucleosome-free regions in vivo.** *Proc Natl Acad Sci U S A* 2010, **107:**21966-21972.

31. Bronner IF, Otto TD, Zhang M, Udenze K, Wang C, Quail MA, Jiang RH, Adams JH, Rayner JC: **Quantitative insertion-site sequencing**

(QIseq) for high throughput phenotyping of transposon mutants. *Genome Res* 2016, **26:**980-989.

32. **UCSC Genome Browser on *S. cerevisiae* Apr. 2011 (SacCer_Apr2011/sacCer3) Assembly** [http://genome-euro.ucsc.edu/cgi-bin/hgTracks?db=sacCer3&lastVirtModeType=default&lastVirtModeExtraState=&virtModeType=default&virtMode=0&nonVirtPosition=&position=chrIX%3A183118%2D203117&hgsid=232744249_LxCF1Ofg7R9Jcsy2DHsyDa984DWn]

Figure Legends

Figure 1. Overview of data acquisition and analysis: transposition events to gene essentiality.

a. Three yeast species analyzed (Sp, *S. pombe*; Sc, *S. cerevisiae*; and Ca, *C. albicans*) by in vivo transposition in this study and which transposons (PB, PiggyBac; AcDs and Hermes) were used to mutagenize which species. Note that each species was analyzed with two different transposon systems.

b. Comparison of genome insertion sites for transposition events that initiate from an extrachromosomal plasmid (left, red region of plasmid circle) or a specific chromosomal locus (right, red bar on a given chromosome). Horizontal lines represent multiple copies of the same genome, each of which underwent a single insertion event (green arrow) per genome. While transposition is generally random, a bias for loci in

close proximity to the initial transposon insertion site demands normalization of the final data.

c. Mapped Tnseq analysis of the pool of transposition events yields the chromosomal insertion sites (brown vertical lines) in the reference chromosomes relative to the ORFs (purple regions). A close up of a small region of a single chromosome (olive horizontal line) including 5 ORFs is illustrated.

d. A training set is constructed using known or inferred labels (non-essential, blue; essential, red) together with extracted features calculated from the data and its position relative to ORFs. Features are defined in Table 2.

e. The training set features, as well as features for all ORFs are used as input for Random Forest classification (black rectangle); output is a prediction of essentiality (red or blue as in d), for which an optimal threshold is determined and applied to designate all genes in one of the two categories.

Figure 2. Contribution of unique hits and total number of reads to the quality of ML predictions for gene essentiality/non-essentiality.

Performance of the classifier vs (a) the total number of unique insertion sites (hits) and (b) the total number of sequencing reads, in each study (organism abbreviations as in Fig. 1a; Ac, AcDs; H, Hermes; PB, PiggyBac).

Figure 3. Feature importances in the different classiffiers.

Importance of each feature used in the Random Forest classifier of essentiality for each dataset. Features are described in Table 2; Neighborhood index generally normalizes for non-random insertion frequencies across the genome; Freedom index reports on the largest proportion of an ORF that has no hits, which is a measure of domains that may be essential.[21]

Figure 4. Analysis of the ability of the CaPB classifier to infer gene essentiality in genes with increasing number of target sequences.
When only ORFs with a specific number of target sites are considered (>= x-axis), AUC rises accordingly (red), but the number of ORFs that can be analyzed necessarily decreases (numbers above red dots). This demonstrates the importance of the prevalence of the transposon target sequences in ORFs, for the quality of gene essentiality inference, using in-vivo transposon mutagenesis studies. X axis: Minimum number of target sequences per ORF needed for inclusion in the classification. Y-axis (red): CaPB classifier AUC.

Figure 5. Analysis of ROC AUC values for Random Forest classification trained on data from a different organism and/or transposon in all possible combinations.

a. For each ROC AUC value in the table, training was performed on 80% of the original training set used in the training species/transposon described in the rows. This training data was then used to predict the essentiality of the remaining 20% of the training set in the species/transposon described in the columns. The train/test split ratio was similar to the 5-fold cross-validation performed in the single study analyses.

b. For each study, the vector of the relative feature importance was correlated with the feature importance in all the other studies. Pearson r correlation coefficient values are presented.

Figure 6: Comparison with known essentials genes.

a. Comparison of the essentiality verdicts in *S. cerevisiae*, based on the known essential genes from the literature, ScAcDs and ScHermes classifiers.
b. Comparison of the essentiality verdicts in *S. pombe*, based on the known essential genes from the literature, SpPB and SpHermes classifiers.
c. Comparison of the essentiality verdicts in *C. albicans*, based on the known essential Sp and Sc orthologs from the literature, CaAcDs and CaPB classifiers.

Classification thresholds differ slightly from the previously published analyses [21] based on threshold selection applied systematically to all 6 studies (described in detail in Methods).

Figure 7. Gene essentiality in haploid and diploid *S. cerevisiae*.

Comparison of essential genes in haploid and diploid *S. cerevisiae* analyzed with Sc Hermes. RF classifier was trained on the haploid ScHermes study and predicted gene essentiality in a diploid strain, using the same threshold for the final verdict. Mitochondrial genes were not considered.

Figure 8. Threshold optimization.

Two metrics were evaluated: 1) Minimum of the Euclidean distance between (0, 1) and the receiver operating characteristic (ROC) curve. 2) Maximum of the vertical distance between the line describing a random choice (a straight line from (0, 0) to (1, 1)) and the ROC curve.

Figure 9. Suspected haploinsufficient genes in *S. cerevisiae*.

Four genes suspected to be haploinsufficient in *S. cerevisiae: NDC1, MLC1, RPC10 and BCY1*, as they appear in the UCSD genome browser [32]. *NDC1* and *MLC1* were known to be haploinsufficient. *BCY1* appears to be a previously unknown haploinsufficient gene. *RPC10* might not be essential as it sustained hits in the 5’ UTR in diploids but not haploids.

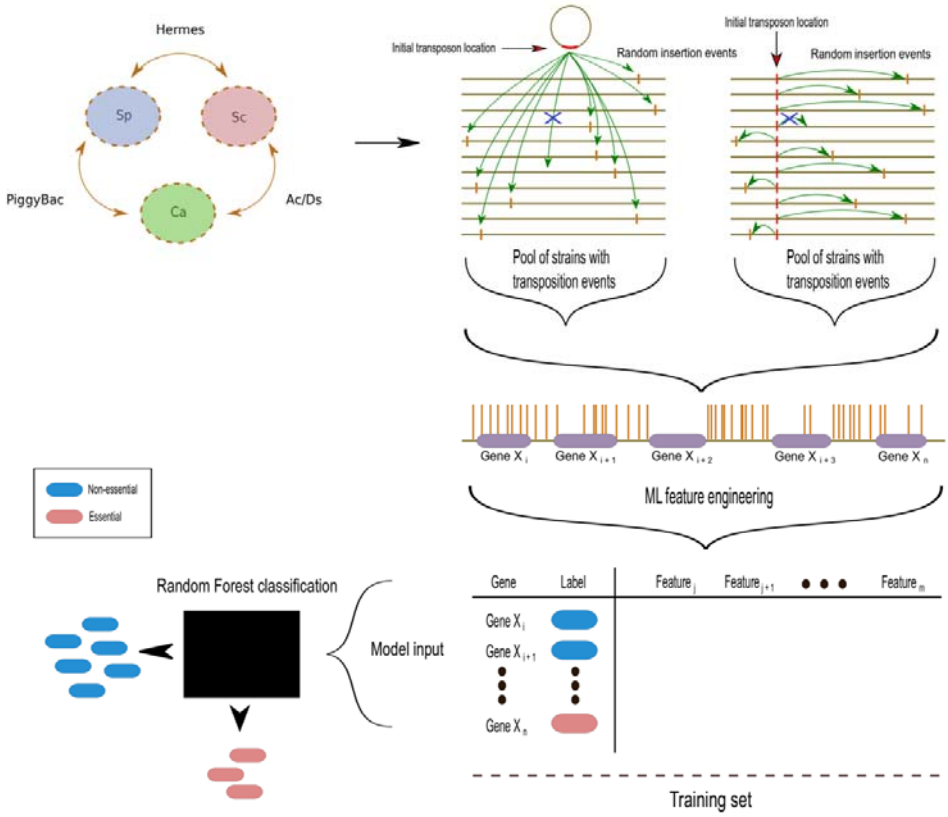


Figure 1. Overview of data acquisition and analysis: transposition events to gene essentiality.

A. Three yeast species analyzed (Sp, *S. pombe*; Sc, *S. cerevisiae*; and Ca, *C. albicans*) by in vivo transposition in this study and which transposons (PB, PiggyBac; AcDs and Hermes) were used to mutagenize which species. Note that each species was analyzed with two different transposon systems.
B. Comparison of genome insertion sites for transposition events that initiate from an extrachromosomal plasmid (left, red region of plasmid circle) or a specific chromosomal locus (right, red bar on a given chromosome). Horizontal lines represent multiple copies of the same genome, each of which underwent a single insertion event (green arrow) per genome. While transposition is generally random, a bias for loci in close proximity to the initial transposon insertion site demands normalization of the final data.
C. Mapped Tnseq analysis of the pool of transposition events yields the chromosomal insertion sites (brown vertical lines) in the reference chromosomes

relative to the ORFs (purple regions). A close up of a small region of a single chromosome (olive horizontal line) including 5 ORFs is illustrated.

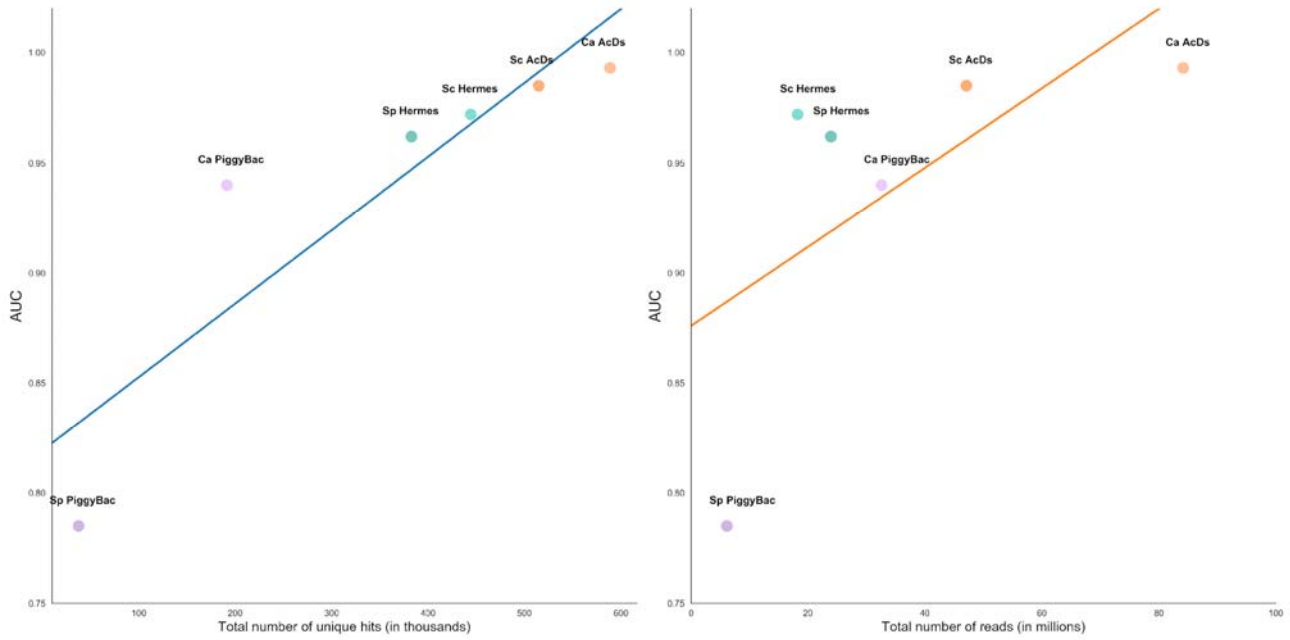
D. A training set is constructed using known or inferred labels (non-essential, blue; essential, red) together with extracted features calculated from the data and its position relative to ORFs. Features are defined in Table 2.

E. The training set features, as well as features for all ORFs are used as input for Random Forest classification (black rectangle); output is a prediction of essentiality (red or blue as in d), for which an optimal threshold is determined and applied to designate all genes in one of the two categories.

F.

Figure 2. Contribution of unique hits and total number of reads to the quality of ML predictions for gene essentiality/non-essentiality.

- Performance of the classifier vs (a) the total number of unique insertion sites (hits) and (b) the total number of sequencing reads, in each study (organism abbreviations as in Fig. 1a; Ac, AcDs; H, Hermes; PB, PiggyBac).



●

Figure 3. Feature importances in the different classifiers.

- Importance of each feature used in the Random Forest classifier of essentiality for each dataset. Features are described in Table XXX; Neighborhood index generally normalizes for non-random insertion frequencies across the genome; Freedom index reports on the largest proportion of an ORF that has no hits, which is a measure of domains that may be essential.[21]

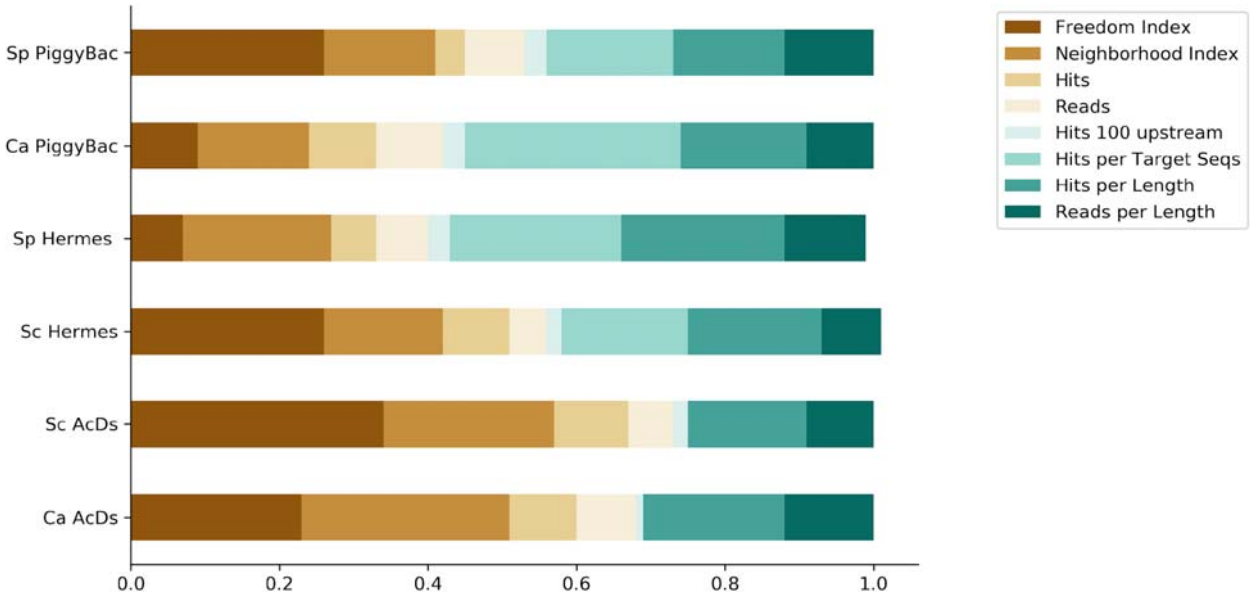


Figure 4. Analysis of the ability of the CaPB classifier to infer gene essentiality in genes with increasing number of target sequences.

- When only ORFs with a specific number of target sites are considered (>= x-axis), AUC rises accordingly (red), but the number of ORFs that can be analyzed necessarily decreases (numbers above red dots). This demonstrates the importance of the prevalence of the transposon target sequences in ORFs, for the quality of gene essentiality inference, using in-vivo transposon mutagenesis studies.
- X axis: Minimum number of target sequences per ORF needed for inclusion in the classification.
- Y-axis (red): CaPB classifier AUC.

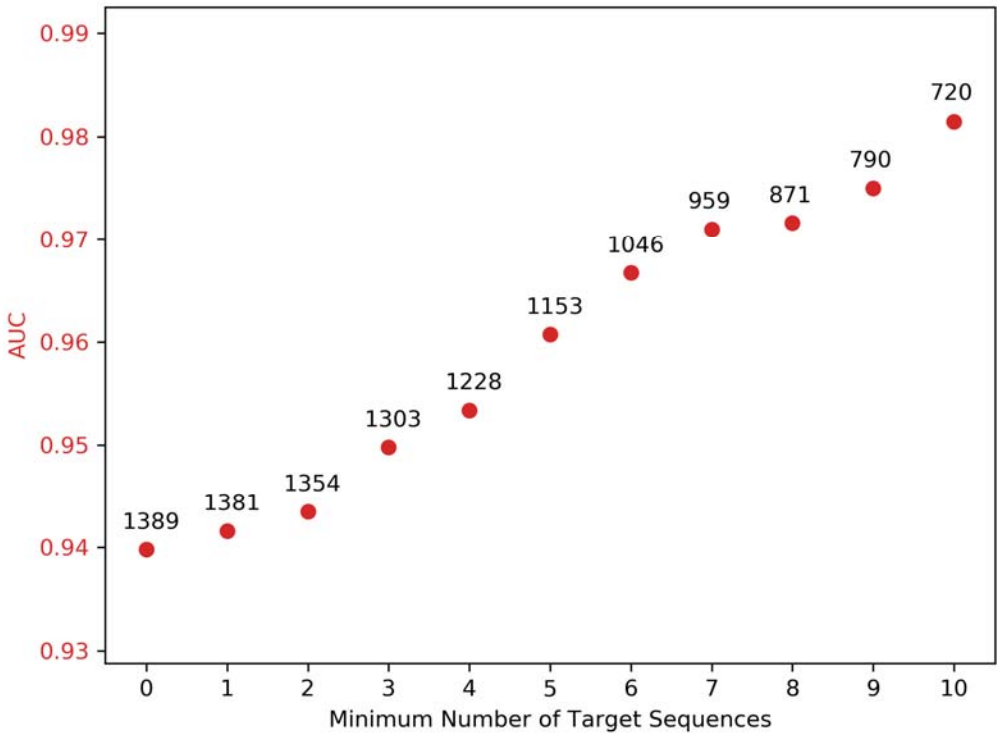
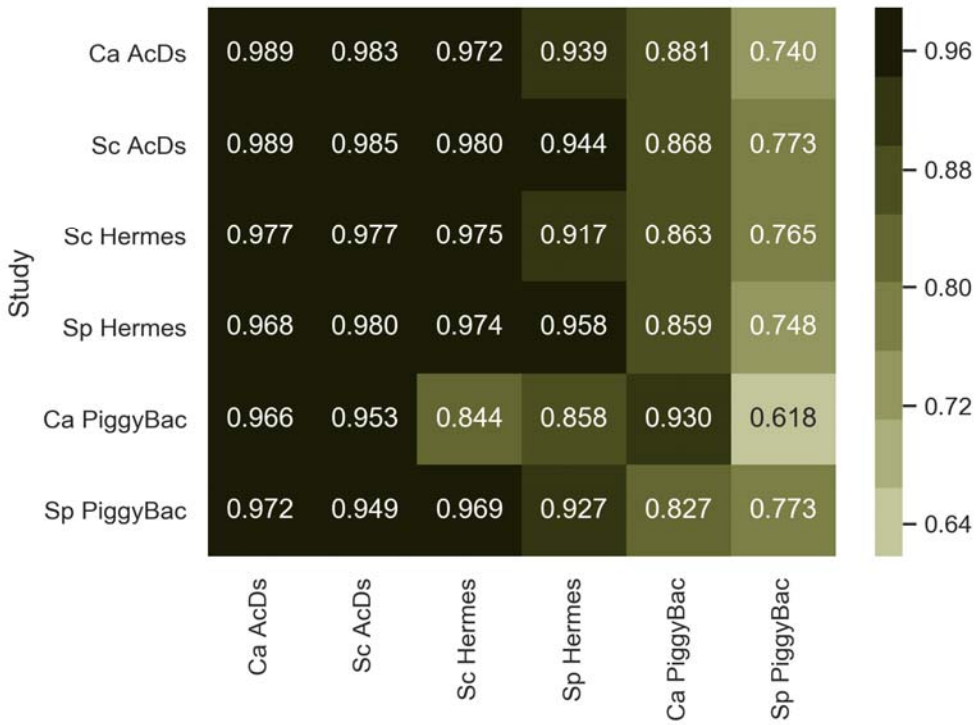


Figure 5. Analysis of ROC AUC values for Random Forest classification trained on data from a different organism and/or transposon in all possible combinations.

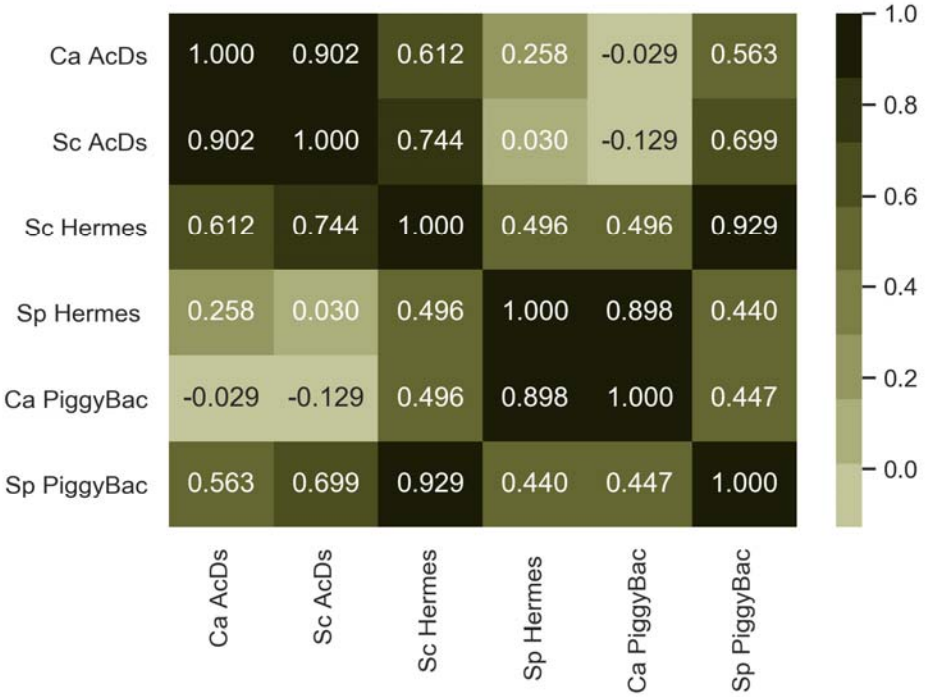
A. For each ROC AUC value in the table, training was performed on 80% of the original training set used in the training species/transposon described in the rows. This training data was then used to predict the essentiality of the remaining 20% of the training set in the species/transposon described in the columns. The train/test split ratio was similar to the 5-fold cross-validation performed in the single study analyses.

B. For each study, the vector of the relative feature importance was correlated with the feature importance in all the other studies. Pearson r correlation coefficient values are presented.

a. Transfer learning AUCs



b. Correlation of feature importance vectors


Ca AcDs
Sc AcDs
Sc Hermes
Sp Hermes
Ca PiggyBac
Sp PiggyBac
1.000 0.902 0.612 0.258 -0.029 0.563
0.902 1.000 0.744 0.030 -0.129 0.699
0.612 0.744 1.000 0.496 0.496 0.929
0.258 0.030 0.496 1.000 0.898 0.440
-0.029 -0.129 0.496 0.898 1.000 0.447
0.563 0.699 0.929 0.440 0.447 1.000
1.0
0.8
0.6
0.4
0.2
0.0

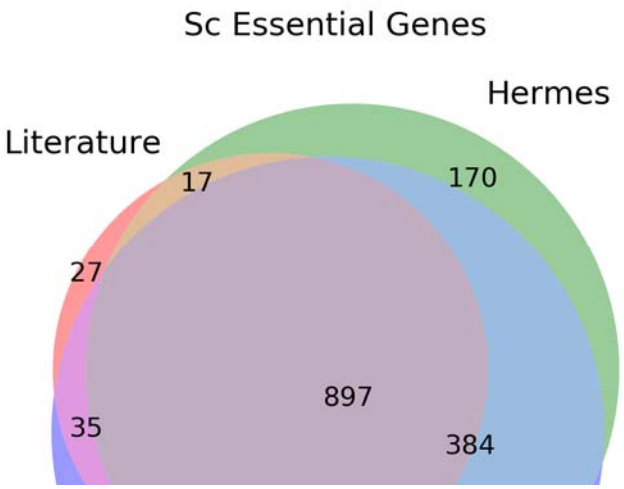
C.

Figure 6: Comparison with known essentials genes.

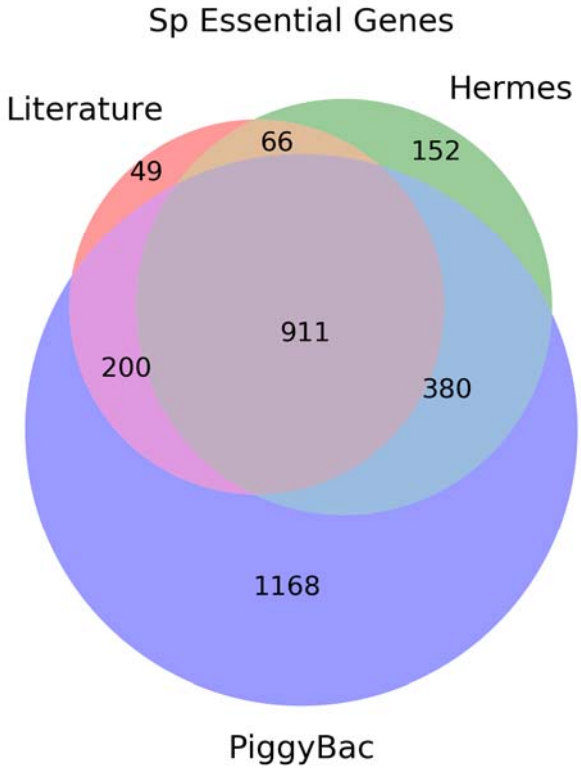
A. Comparison of the essentiality verdicts in *S. cerevisiae*, based on the known essential genes from the literature, ScAcDs and ScHermes classifiers.
B. Comparison of the essentiality verdicts in *S. pombe*, based on the known essential genes from the literature, SpPB and SpHermes classifiers.
C. Comparison of the essentiality verdicts in *C. albicans*, based on the known essential Sp and Sc orthologs from the literature, CaAcDs and CaPB classifiers.

- Classification thresholds differ slightly from the previously published analyses[21]based on threshold selection applied systematically to all 6 studies (described in detail in Methods).

a.



b.



c.

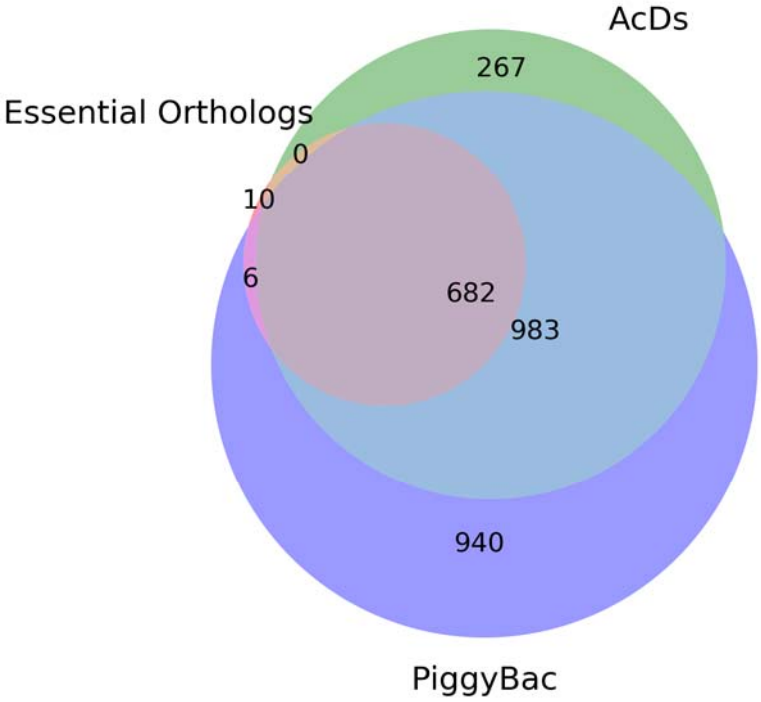


Figure 7. Gene essentiality in haploid and diploid *S. cerevisiae*.

- Comparison of essential genes in haploid and diploid *S. cerevisiae* analyzed with Sc Hermes. RF classifier was trained on the haploid ScHermes study and predicted gene essentiality in a diploid strain, using the same threshold for the final verdict. Mitochondrial genes were not considered.

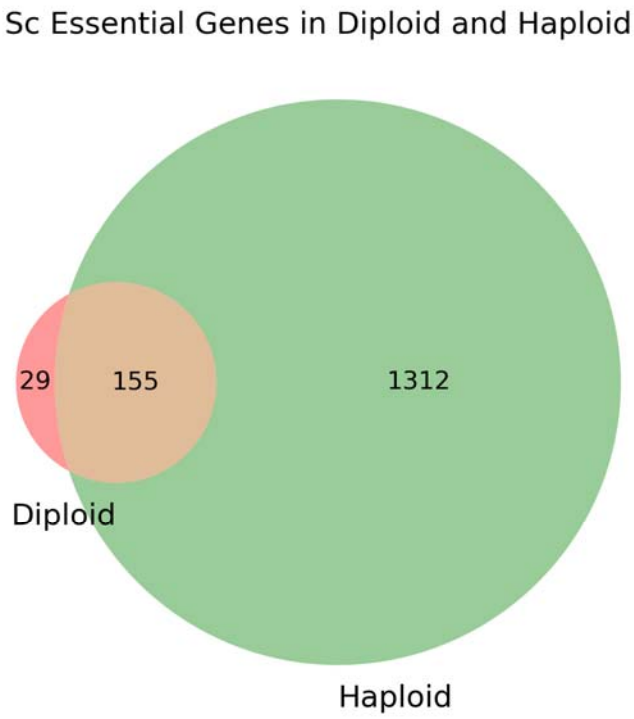
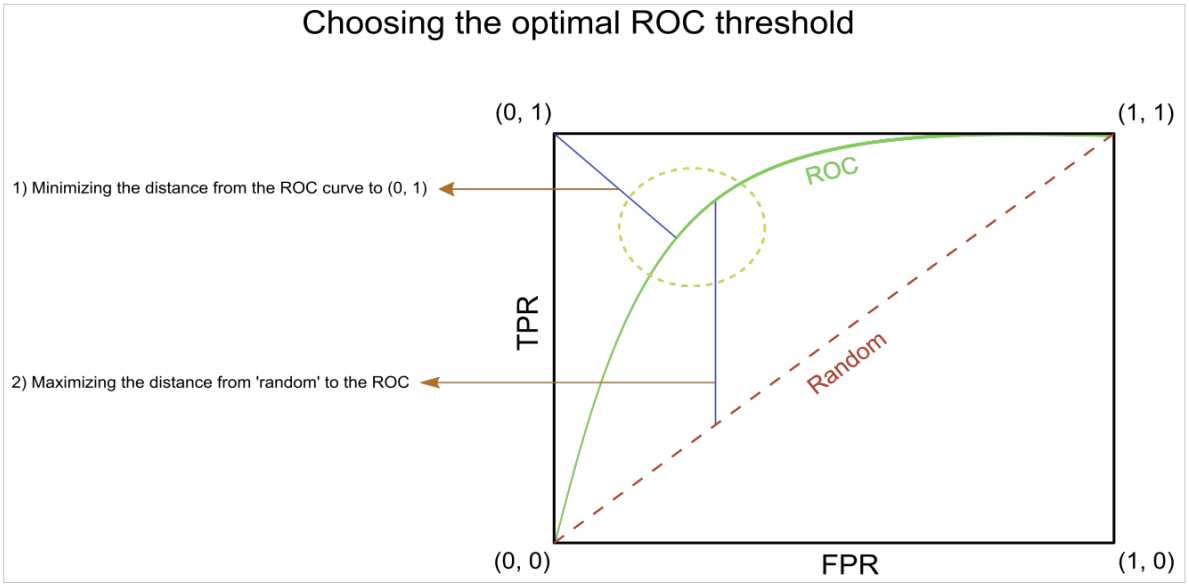


Figure 8. Threshold optimization.

- Two metrics were evaluated: 1) Minimum of the Euclidean distance between (0, 1) and the receiver operating characteristic (ROC) curve. 2) Maximum of the vertical distance between the line describing a random choice (a straight line from (0, 0) to (1, 1)) and the ROC curve.

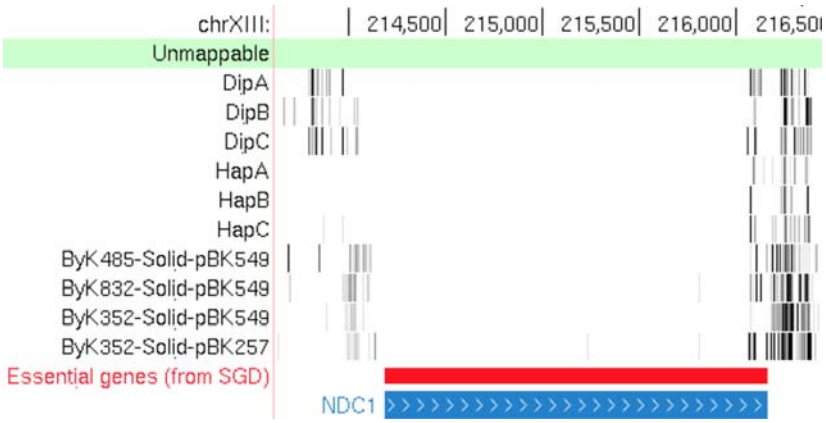


-

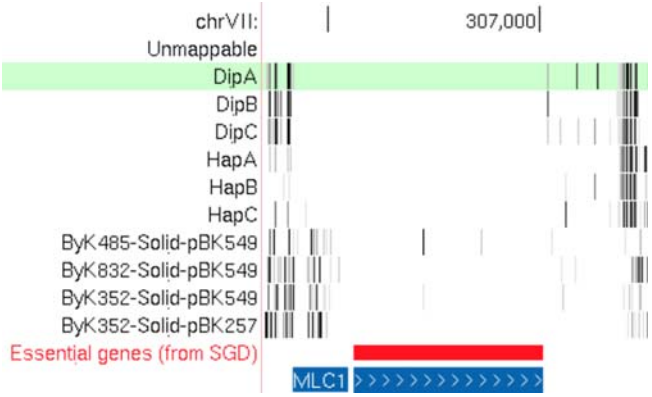
Figure 9. Suspected haploinsufficient genes in *S. cerevisiae*.

- Four genes suspected to be haploinsufficient in*S. cerevisiae*: NDC1, MLC1, RPC10 and BCY1, as they appear in the genome.[32]NDC1 and MLC1 were known to be haploinsufficient. BCY1 appears to be a previously unknown haploinsufficient gene. RPC10 might not be essential as it sustained hits in the 5' UTR in diploids but not haploids.

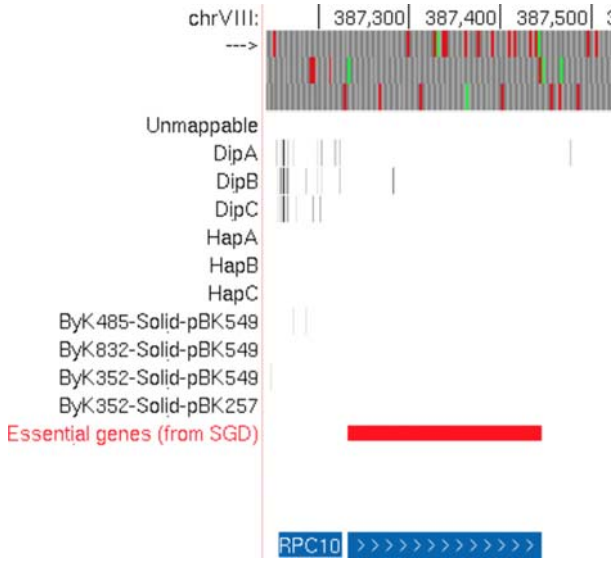
NDC1



MLC1



RPC10



BCY1

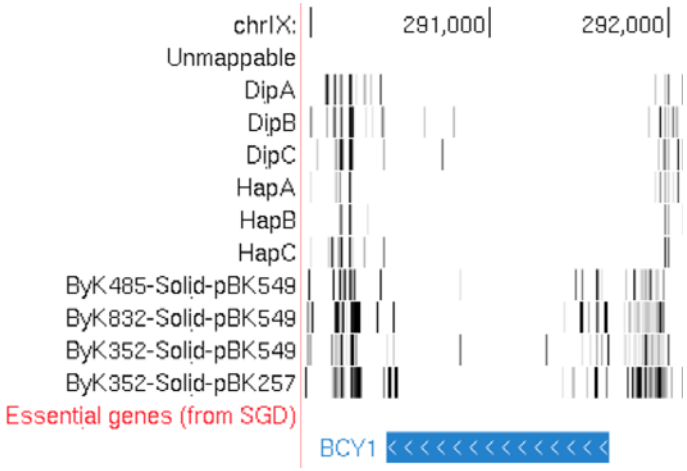

chrIX:
291,000
292,000
Unmappable
DipA
DipB
DipC
HapA
HapB
HapC
ByK485-Solid-pBK549
ByK832-Solid-pBK549
ByK352-Solid-pBK549
ByK352-Solid-pBK257
Essential genes (from SGD)
BCY1

Table 1. Statistics of the transposon data sets

	Ca AcDs	Sc AcDs	Sc Hermes	Sp Hermes	Ca PiggyBac	Sp PiggyBac
Transposon target sequence	-	-	TnnnnA	TnnnnA	TTAA	TTAA
Initial transposon insertion	Genome	Plasmid	Plasmid	Plasmid	Genome	Genome
Number of target sequences (10^3)	-	-	1154.84	1302.41	120.27	111.37
Total number of unique hits (10^3)	588.97	514.89	444.41	382.82	191.49	37.5
Target Sequences without a hit (10^3)	-	-	924.56	1110.2	10.69	79.09
Percent of target sequences without a hit	-	-	80.06%	85.24%	8.89%	71.02%
Percent of hits in target sequences	-	-	51.82%	50.21%	56.74%	86.06%
Total number of reads (10^6)	84.16	47.1	18.22	23.92	32.58	6.14
Average number of reads per hit	143	91	41	62	170	164
Standard deviation reads per hit	9250	4357	137	5069	4584	481
Highest reads per hit (10^3)	3254.83	2210.55	11.72	2355.61	1301.82	48.31
Highest reads per hit / average reads per hit	22761	24292	286	37994	7658	295
Number of over 106 reads per hit	10	2	0	2	1	0
Average number of hits per gene	29.11	44.05	30.36	24.64	11.81	1.32
Number of genes with 0 hits	300	261	332	99	397	2580
Average number of target sequences per gene	-	-	123.64	131.44	10.62	8.39
Number of genes with 0 target sequences	-	-	0	0	228	186
ROC AUC	0.993	0.985	0.972	0.962	0.94	0.785

Table 2. Classification features

Feature	Description
Hits	Number of transposon hits within the ORF
Reads	Number of reads associated with the transposon hits within the ORF
Neighborhood Index (NI)	Number of transposon hits within the ORF, normalized by length of the ORF and the surrounding 10 kbp
Freedom Index (FI)	Length of the largest hit-free region in the ORF, divided by the ORF's length
Hits 100 upstream	Number of transposon hits within the upstream region of the ORF
Hits per Target Seqs	Number of transposon hits divided by the number of transposon target sequences within an ORF
Reads per Length	Number of transposon hits divided by the length of the ORF
Hits per Length	Number of reads associated with the transposon hits divided by the length of the ORF