Continuous generation of tandem transposable elements in Drosophila populations provides a substrate for the evolution of satellite DNA

Michael P. McGurk and Daniel A. Barbash

Department of Molecular Biology and Genetics

Cornell University

Ithaca, NY 14853

Abstract:

Eukaryotic genomes are replete with repeated sequences, in the form of transposable elements (TEs) dispersed across the genome or as satellite arrays, large stretched of tandemly repeated sequence. Both types of repeats have attracted much interest due to their considerable variation across species, unusual patterns of evolution, and implications in human disease. The boundary between these categories of repeats is fuzzy, however: Many satellite sequences clearly originated as TEs, though it is unclear how mobile genetic parasites occasionally transform into sometimes megabase-sized arrays. Whatever the generative mechanism, in *Drosophila melanogaster* two TE families have undergone this transition, hinting that its early stages might be observable in a survey of population variation. However, the best available population resources, short-read DNA sequences, are often considered to be of limited utility for the analysis of repetitive DNA due to the challenge of mapping individual repeats to unique genomic locations. Here we develop a new pipeline called ConTExt which demonstrates that reference-free analysis of paired-end Illumina data can identify a wide range of structures of repetitive DNA and reveal a quantitative understanding of repeat sequence variation. Analyzing 85 genomes from five populations of *Drosophila melanogaster* we discover that all three classes of TEs commonly exist as tandem dimers. Our results further suggest that insertion site preference is the major mechanism driving the creation of tandem dimers and that, consequently, dimers form rapidly during periods of active transposition. This abundance of TE dimers has the potential to provide the source material for future expansion into satellite arrays, and we discover one such rare copy number expansion

of the DNA transposon *Hobo* to ~16 tandem copies in a single line. The very process that defines TEs—transposition—, thus, constantly adds to the library of sequences from which new satellite DNAs arise.

Introduction

The evolutionary processes underlying repeat sequence variation in eukaryotes has been debated for decades. Eukaryotic genomes are inundated with two classes of repetitive sequences: transposable elements (TEs), which are dispersed across genomes by a variety of transposition mechanisms, and satellite sequences, which are tandemly repeated sequences that expand, contract, and are homogenized by recombination events. Both classes of repeats are typically enriched near the telomeres and in the heterochromatin surrounding the centromeres, likely because the low frequency of recombination typical of heterochromatin is permissive to their persistence (Charlesworth et al., 1986). However, the essential roles played by telomeres and centromeres in genome integrity suggest that some repetitive sequences are of functional significance. Sequence variation in a human centromeric satellite is associated with mitotic segregation errors resulting in aneuploidy (Aldrup-MacDonald et al., 2016). The structural importance of constitutive heterochromatin means that changes in repeat composition between species can cause reproductive barriers (Ferree and Barbash, 2009). The intimate relationship between repeats and chromatin state further means changes in satellite composition can impact the regulation of genes. Contractions of the human subtelomeric satellite D4Z4 cause facioscapulohumeral muscular dystrophy by altering the chromatin state of nearby genes (Zeng et al., 2009).

Variants of the mostly repetitive *Drosophila melanogaster* Y chromosome have global impacts on gene expression, possibly by titrating chromatin binding factors (Francisco and Lemos, 2014). Satellites may also hijack normal biological processes in ways that provide a reproductive advantage for themselves. There are several examples of satellites engaging in meiotic drive and gamete competition (Fishman and Saunders, 2008) (Hardy et al., 1984), selfish processes whereby alleles bias meiotic segregation or gamete survival to transmit at frequencies greater than Mendelian expectations.

Yet, our understanding of the within-species evolution of satellite sequence remains rudimentary. Many satellite sequences turnover rapidly between closely related species (Lohe and Roberts, 2000). Neutral evolution by unequal exchange provides a partial explanation for this high turnover rate as 1) dramatic changes in copy number can arise from relatively few exchange events and 2) the eventual fate of such an array is its inevitable loss from the genome (Charlesworth et al., 1986). The long-term persistence of some conserved satellites (Strachan et al., 1982) may therefore reflect functional importance, but given their ubiquity, mechanisms to generate new satellites must exist to counter the inevitable loss of neutrally evolving arrays (Charlesworth et al., 1986). The library model of satellite evolution proposes two stages in the emergence of new satellites: 1) amplification processes generate small tandem sequences and 2) some of these sequences expand to large arrays by unequal exchange (Ugarković and Plohl, 2002) . Thus, any process that generates sequence upon which unequal exchange can act is a potential source of new satellites. Polymerase slippage is likely a common path to new simple satellites, where short sequences become tandemly duplicated by successive replication errors until the repetitive stretch is sufficiently large

for unequal exchange to act upon it (Tautz and Schlötterer, 1994) (Charlesworth et al., 1994). Subsequent mutations can then generate more complex monomer types, with the interplay of mutation and unequal exchange then giving rise to higher order repeats (HORs) comprised of highly degenerate versions of the ancestral repeat unit (Prosser et al., 1986).

More enigmatic mechanisms to generate new satellites also exist. TEs are well-known culprits in causing mutations and gene duplication events, but they are also found as tandems arrays in many species, including as centromeric satellites (Meštrović et al., 2015). Perhaps the easiest to understand are those satellites derived from long terminal repeat (LTR) retrotransposons (LTR RTs) (Zhang et al., 2014) (Gong et al., 2012) where the LTRs (typically a few hundred base pairs long) provide substrates for tandem formation by unequal exchange (Ke and Voytas, 1997). Another example is satellites derived from the expansion of tandemly repeated regulatory sequences internal to certain TE families (Macas et al., 2009) (Dias et al., 2014). Yet, even TEs without intrinsic repeats form tandem arrays of complete elements (Caizzi et al., 1993) (Miller et al., 1992), possibly by rolling circle replication (RCR) (Marsano et al., 2003) (Meštrović et al., 2015). A counterpoint to RCR as the generative mechanism is the non-LTR retrotransposon *R1* that has the unusual property of only inserting at a specific site in the multicopy ribosomal RNA genes (rDNA) and also forms a tandem array on the X chromosome in *D. melanogaster* (Kidd and Glover, 1980) (Peacock et al., 1981). The tandem *R1* elements are separated by 33-nt target-site duplications (TSD) of rDNA sequence, consistent with the tandem originating when two *R1* elements inserted in the same rDNA unit and then subsequently expanded by unequal exchange (Roiha et al.,

1981). Tandem dimers of DNA transposons have been found in bacterial genomes, and these also contain TSDs between the tandem elements (Dalrymple, 1987) (Prudhomme et al., 2002), hinting that double insertions may not be limited to elements with insertion site preferences as extreme as in R1.

Whatever the generative process may be, that there are two examples of TEs transitioning to satellite sequence in *D. melanogaster* (Caizzi et al., 1993) (Kidd and Glover, 1980) suggests that it occurs frequently enough that its early stages may be detectable in a survey of population variation. A full assessment of the mechanisms and frequency of TEs to generate tandem arrays, however, remains unexplored, largely due to the wider challenge of applying the most comprehensive population genomic resource available—NGS data—to investigating the evolutionary dynamics of repetitive DNA. One recent analysis of simple satellites (2-10 bp long) across populations of *D. melanogaster* identified population-specific kmers and population differentiation for a number of satellites (Wei et al., 2014), but similar surveys for complex satellites are more challenging. However, the existence of TE-derived satellites provides a potential opportunity: rather than searching for the emergence of satellites from all possible single-copy sequences, tandems arising from repeats that are normally dispersed rather than tandemly arranged, might yield a tractable model for studying the early stages of satellite evolution. To carry out such a survey, we developed a bioinformatic pipeline, ConTExt, that allows us to employ paired-end NGS data to detect structures formed by known repeats and ask how these structures vary across individuals. Applying our pipeline to the Global Diversity Lines, a published panel of 85 *D. melanogaster* strains from five populations (Grenier et al., 2015) , we find that TEs frequently form tandem

dimers by inserting multiple times at the same locus, continuously generating new books in the library of sequences from which satellite arrays can arise.

Methods

Overview

We approach the problem of making inferences about repeats from NGS data using two main steps. First, we employ an alignment strategy that maps repeat-derived reads to the consensus sequences of known repeats. Second, we employ a clustering strategy to infer structures from the distributions of discordant read pairs. Clustering repeat-derived reads requires a more sensitive strategy than reads anchored in unique sequence. Consequently, rather than employ an agglomerative clustering scheme as is common in structural variant discovery, we use mixture modeling to identify patterns in the data. The success of this approach depends upon the choice of an appropriate cluster shape parameter for each sequencing library. We choose this empirically, by constructing a training set of structures for each library and assessing clustering performance over a range of parameter values.

Constructing the repeat index

We downloaded the RepBase repeat annotations (release 19.06) (Bao et al., 2015) for D. melanogaster and supplemented these with a set of simple and complex satellites found at <http://www.fruitfly.org/sequence/sequence_db/na_re.dros>. We also included the rDNA sequence from (Stage and Eickbush, 2007), additional TE sequences from (Zanni et al., 2013) , and extracted subfamilies of the telomeric TE *Het-*

A, the subtelomeric repeat units, the histone cluster, the *stellate* gene, and ubiquitin monomer from the reference genome. Further, the *TART* telomeric TEs contain very long terminal repeats, which we removed from the consensus and placed in separate entries, in accord with how LTRs are handled for LTR retrotransposons. This set of sequences contained redundancies that would have complicated interpreting the alignments; for example, there were over 20 sequences corresponding to the centromeric 359-bp repeat satellite, and two *Protop* variants which simply corresponded to internally deleted elements. To resolve these issues, we manually curated the index by performing all pairwise alignments to identify entries that share considerable homology. To make this tractable, we used a clustering strategy to organize these pairwise alignments into groups of potentially redundant entries. We constructed a graph where nodes represent sequences and any pair of sequences that share a BLASTn alignment of >80% identity were connected with an edge. We used the Louvain algorithm (Blondel et al., 2008) as implemented in the Python package louvain 0.5.3 to partition this graph into communities which we then manually inspected; this is purely an organizational step to simplify manual curation. When two sequences represented the same repeat, we aligned them to the reference, and favored the most complete element that was present at multiple insertions. For the community of 359-bp repeats, we examined the tree of pairwise alignments constructed using the NCBI-BLAST application (Altschul et al., 1990), noted that there were three deep partitions and kept one element in each. We also aligned all remaining elements to themselves to identify any LTRs that had not been removed and gave them their own entries. *Batumi*, *Foldback* and *Vatovio* are not included in subsequent analyses due to errors in

RepBase entries or their propensity for mapping artifacts. While we subsequently refer to the entries in this index as consensus sequences, not all are true consensus sequences; rather, some entries are representative examples.

Read preprocessing

We used Trimmomatic for read quality control (Bolger et al., 2014). We removed all sequence within and subsequent to the 5'-most four nucleotide window where the average PHRED quality score was less than 20. We discard any trimmed reads with lengths less than 40-nt. Because we can only detect a junction if it falls in the gap between reads in a pair, we trimmed any reads longer than 70-bp down to 70-bp from their 3'-ends to increase the size of this gap.

Aligning reads to repeats

The simplest way to determine from which repeat family a read derives is to align reads to a set of repeat consensus sequences. However, the consensus sequences of TE families frequently resemble the most active subfamily and can be quite diverged from older insertions. Reads originating from insertions of old families, like DINE, often do not align even under very permissive alignment parameters. To circumvent this, we employed a two-step alignment procedure, first aligning reads to individual instances of each repeat family, and then collapsing these alignments onto the corresponding consensus sequence.

We first used RepeatMasker (Smit et al., 2015) to both mask release 6.01 of the *D. melanogaster* reference genome (Hoskins et al., 2015) for repeats and to identify the location of all instances of each repeat family, using the most sensitive seed setting. We

extracted all sequences in the reference annotated as a repeat and used these to construct an index of individual insertions. We then used Bowtie2 (version 2.1.0) to align the reads in each read pair as single-end reads to both the repeat-masked reference genome and the index of individual repeats (Langmead and Salzberg, 2012). We used parameters [bowtie2-align -p <# threads> -<phred format of data> --score-min L,0,<-.64 or -1.0> -L 22 -x <Index> -U <infile> -S <outfile>] and use -.64 as the alignment threshold for the reference genome and -1.0 as the alignment threshold for the insertion index.

Alignments were then collapsed onto the corresponding repeat consensus sequences. To accomplish this, we first used BLASTn to align all individual repeats to all repeat consensus sequences and kept those alignments where the percent identity was at least 80%. If an individual repeat aligned multiple times to the set of consensus sequences, we chose the alignment with the lowest e-value. If there were multiple alignments tied for e-values, we chose the longest alignment. We then used the BLAST traceback strings to convert the aligned coordinates of reads from individual insertion to the corresponding coordinates on consensus sequences. If the new aligned length of the read on the consensus was less than 50% or greater than 150% of the original read length, that alignment to the consensus was rejected and the read was marked as unaligned. We updated the CIGAR and MD strings to reflect the new alignment.

Estimating the gap size distribution

The distribution of reads spanning a junction depends upon the size distribution of read pairs in the library. We refer to the distance between the 5’ ends of a concordant

read pair as the insert size, and the interval of sequence between the 3' ends of a read pair reads as the gap. For a junction to be detected, it must be spanned by the gap (junctions interrupting a read will likely prevent its alignment), so for each library we estimated the gap size distribution. Because we knew *a priori* that the average insert size in the GDL generally is around 500 bp, we considered any read pair putatively concordant where the reads 1) map to opposite strands, 2) are oriented toward each other, and 3) have a distance between 3' ends of less than 1kb. On each of the major autosome arms we counted the number of reads with each possible gap size between 0 and 1,000, and normalized this to obtain separate estimates of the gap size distribution (SF 1A) . Rather than fit parametric distributions to the observed gap size frequencies, we employed the nonparametric kernel density estimator with a Gaussian kernel. We chose an optimal bandwidth by 2-fold cross validation (CV). In short, as we constructed a separate gap size distribution for each autosome arm, we have four observations for each read count, which we assign to two complementary subsets at each iteration of CV; within each subset, the frequencies at each gap size are averaged and the distributions are normalized to sum to one. We constructed 100 such subsets, and cross-validated with respect to the mean-squared error (MSE) over a range of bandwidths between 0 and 100, performing an exponential search over four orders of magnitude. We then pooled the reads from across the chromosomes, taking the KDE with the chosen bandwidth. For simulations of read distributions, we use the gap size distribution conditioned on the reads spanning a junction which accounts for small inserts being less likely to span a junction and is approximately given by:

$P(g|J) = \frac{P(g)\times g}{\sum_l P(l)\times l}, g\sim G$, where G is the gap size distribution.

In all subsequent sections, we consider a read pair concordant when its two reads map to opposite strands and are oriented toward each other, and when its gap size falls between the .5% quantile and the 99.5% quantile of the gap size distribution.

Representing paired-end alignments as two-dimensional scatterplots

A read pair can be represented as a point in a two-dimensional space, where the X- axis represents the sequence and strand to which one read maps, and the Y-axis represent the sequence and strand to which the other read maps (SF 1A,B). This is an effective visualization strategy for manually examining the patterns of TE insertions into unique sequence and into repetitive sequence. Organizing reads where both ends map to the same sequence requires additional constraints. For read pairs that map to opposite strands of the same sequence, we assign the reverse strand to the X-axis and the forward strand to the Y-axis. For read pairs that map to the same strands of the same sequence, there is ambiguity as to which axes the reads should be assigned. In the case of forward-forward read pairs, we assign the read with the higher sequence coordinate to the X-axis, and for reverse-reverse, we assign the read with the lower sequence coordinate to the Y-axis.

An Expectation Maximization algorithm for discovering structures

We model the set of all discordant read pairs within each partition with a Gaussian Mixture Model (GMM), where each component distribution represents a pattern of alignments produced by a structure in the sequenced genome. The actual distributions of reads produced by structures are not Gaussian (SF 1B), however GMMs

are often successful in clustering mixtures of non-Gaussian distributions and the approximation makes the problem tractable. We fit the GMM to the data using Expectation Maximization (EM), an iterative procedure that begins with an initial model and iteratively updates the parameters in a manner guaranteed to converge to a local optimum with respect to the model likelihood. This is still a greedy algorithm in that it is not guaranteed to find the global optimum, but as it considers the positions of all read pairs at each iteration it is often able to resolve overlapping distributions of reads. We use EM to update the means and mixing proportions of the model, but hold the covariance constant and tied among all components. The covariance parameter describes the shape of the component distributions; as this relates to the insert size distribution of the sequencing library and does not vary among the individual distributions, it can be learned prior to clustering the data. This constraint coupled with soft-cluster assignments during EM allows the algorithm to automatically identify a reasonable number of clusters given the data, provided that the number of the initial components is greater than the number of real clusters in the data (SF 2A,B). This makes solving the clustering problem tractable and easy to automate because it eliminates the need for model selection with respect to the number of components in the mixture model.

We used the standard update equations for fitting a GMM with EM, and employed the SquareEM acceleration procedure to decrease the number of iterations necessary for convergence (Varadhan and Roland, 2008). In SquareEM, a parameter, α, is computed at each iteration based on how the EM update step behaves over a set of test updates and which controls the rate at which EM updates move through

parameter space; an additional iterative procedure is required to tune α to preserve EM's guarantee that the updated model must be at least as good as the current model. We limited this tuning procedure to 10 iterations, after which if a suitable α is not identified we set α = -1, which corresponds to a single standard EM update (and thus guaranteeing the likelihood increases monotonically). Components are initialized by the Forgy method (i.e. placed on randomly chosen data points), but with the additional constraint that no two components be placed within distance *D* of each other, and are added until there are no eligible data points remaining. *D* is set equal to $\frac{1}{2}\left(\frac{\sigma_1 + \sigma_2}{2}\right)$, where σ_1 and σ_2 are the standard deviations of the component distributions along the first and second eigenvectors of the covariance matrix. This heuristic ensures that the number of components in the model is greater than the number of true clusters in the data and that components are well-distributed across the data, rather than concentrated near high copy structures as would be the case with the standard Forgy method and which would increase the probability that the EM procedure becomes trapped at a poor local optimum. As the EM iterations proceed, components are merged if their means ever attain the same integer values. Very rarely, all components drift sufficiently far from a data point that the computed probabilities for that read pair are incorrectly rounded to zero due to the limits of Python's numerical precision; whenever this happens, the algorithm is simply reinitialized and repeated until it completes without encountering such a rounding error. Each time this happens during initialization, the distance constraint *d* is recursively reduced to $.9d$ to increase the number of components in the model and the algorithm is reinitialized. The EM procedure is terminated when the change in the log-likelihood drops below 10^{-3} or after 200 iterations. In practice, the

algorithm almost always terminates prior to 200 iterations. Hard clusters are subsequently formed by assigning each read pair to the most likely cluster.

When applied to the GDL, the algorithm converged prior to 200 iterations in all but 55 of the 1,137,005 EM runs; 46 of these 55 were cases where a rounding error prevented convergence, and the algorithm was reinitialized. The average time to convergence was .03 seconds (median .004) requiring on average 3.76 SquareEM iterations.

Setting the Model Covariance

An appropriate covariance matrix for the component distributions needs to be identified for each sequencing library before it can be clustered. A bivariate covariance matrix, Σ , oriented at angle θ can be decomposed $\Sigma = Q\Lambda Q^{-1}$, where Q is a matrix with eigenvectors as columns of the form $Q = \begin{bmatrix} \cos(\theta\ ) & -\sin(\theta) \\ \sin(\theta) & \cos(\theta) \end{bmatrix}$ and Λ is a diagonal matrix of eigenvalues $\begin{bmatrix} \sigma_1^2 & 0 \\ 0 & \sigma_2^2 \end{bmatrix}$. Because we know that the distributions should be oriented at an angle of $45°$ relative to the X and Y axes, we set $Q = \begin{bmatrix} \sqrt[-2]{2} & \sqrt[-2]{2} \\ -\sqrt[-2]{2} & \sqrt[-2]{2} \end{bmatrix}$ when the axes represent the same strand and $Q = \begin{bmatrix} \sqrt[-2]{2} & -\sqrt[-2]{2} \\ \sqrt[-2]{2} & \sqrt[-2]{2} \end{bmatrix}$ when the axes represent different strands. Then the matrix Λ of eigenvalues σ_1^2 and σ_2^2 specifies the covariance, and corresponds to the variances of a bivariate Gaussian along the two eigenvectors. Rather than optimize the variances, we optimized with respect to the standard deviations σ_1 and σ_2.

To identify the smallest pair of eigenvalues expected to yield a good clustering of the data, we considered the algorithm's performance on a set of training data. We constructed the training set from insertions of a few high-copy TE families (Roo, Hobo, Baggins, and DMCR1A) into unique sequence on the major chromosome arms (2L, 2R, 3L, 3R, X). To identify insertions, we clustered the set of read pairs where one end is anchored in a repeat sequence and the other on a chromosome arm for each combination of training family, chromosome arm, and strand using single linkage agglomerative clustering with a distance cutoff of δ. We chose δ using the library's insert size distribution to perform 10,000 Monte Carlo simulations of three read pairs spanning a junction and determining the maximum distance on either axis between the adjacent reads in a simulated cluster; δ is set equal to the 99%-quantile of this maximum distance. This clustering strategy is equivalent to that used in PoPoolationTE (Kofler et al., 2015), except that the clustering cutoff is chosen based on the insert size of the library. We kept all such clusters with at least three reads. To exclude aberrant clusters and PCR-duplicated singletons from the training set, we removed any training cluster that did not yield a single cluster upon reclustering with the largest tested covariance matrix $\Lambda = \begin{bmatrix} \sigma_{1,max}^2 & 0 \\ 0 & \sigma_{2,max}^2 \end{bmatrix}$ or which yielded a single cluster upon reclustering with a very small covariance matrix $\Lambda = \begin{bmatrix} 1 & 0 \\ 0 & 1 \end{bmatrix}$.

To choose σ_1 we performed a parameter sweep, assessing clustering performance with $\Lambda = \begin{bmatrix} \sigma_1^2 & 0 \\ 0 & \sigma_{2,max}^2 \end{bmatrix}$ for values of σ_1 between 1 and $\sigma_{1,max}$ with a step size of 1. At each σ_1 we bootstrapped the training set to obtain a sample of 150 training clusters, used SquareEM to recluster each with a 10-component GMM without the

constraint that initial components must be distance *d* apart, and computed average clustering performance, γ. As we expect that each entry in the training set represents a single cluster, we defined clustering performance on each training cluster as the proportion of reads in the largest cluster identified by EM upon reclustering, such that performance equals 1 if the original cluster is recovered and is less than 1 if it is not. To model the expected clustering performance as a function of σ_1 we employed Gaussian Process (GP) Regression as implemented in Scikit-Learn (Pedregosa et al., 2011), an expressive regression model which learns nonlinear relationships from an extensive family of functions; the kernel of the GP is constructed as $Rational\ Quadratic \times (Constant\ +\ WhiteNoise)$ with bounds on the Rational Quadratic kernel's length scales at 1e-20 and 1e20, and the associated hyperparameters chosen by gradient ascent with respect to log-likelihood. We fit with 20 restarts to reduce the odds of becoming trapped at local maxima and use the expectation of the fitted model to set the optimal parameter $\hat{\sigma}_1$ equal to the smallest σ_1 such that the predicted performance is $\gamma \geq \alpha_1$ (SF 1C). We then performed a parameter sweep for σ_2, assessing clustering performance with $\Lambda = \begin{bmatrix} \hat{\sigma}_1^2 & 0 \\ 0 & \sigma_2^2 \end{bmatrix}$ for values of σ_2 between 1 and $\sigma_{2,max}$ with a step size of 1, and use the posterior mode of the fitted GP model to set $\hat{\sigma}_2$ equal to the smallest σ_2 such that the predicted performance is $\gamma \geq \alpha_2$. We specified the covariance matrix of the GMM for the dataset using $\Lambda = \begin{bmatrix} \hat{\sigma}_1^2 & 0 \\ 0 & \hat{\sigma}_2^2 \end{bmatrix}$. We set $\sigma_{1,max} = 333$, $\sigma_{2,max} = 80$, $\alpha_1 = .99$, and $\alpha_2 = .97$ based on manual examination of clustering performance in a few datasets.

Assessing clustering performance

We were primarily concerned with the precision ($\frac{TP}{TP+FP}$) and recall ($\frac{TP}{TP+FN}$) of our algorithm which are dependent on the numbers of True Positives (TP), False Positives (FP), and False Negatives (FN). Here, Positive and Negative respond to the statement "These two reads belong in the same cluster." As this is dependent on knowing the true labels, we assessed performance on simulated data. Ensuring that the algorithm has good recall is important for structure discovery, as a high false negative rate means that the algorithm is likely to identify multiple junctions, where in truth there is only one. Precision, however, is important when studying repeats, as distinct structures often involve similar junctions which yield overlapping read pair distributions that must be resolved; a high false positive rate means that the algorithm will identify a single junction where there are in fact multiple distinct junctions. We employed two tests to each dataset to summarize how the algorithm behaves in each situation. 1) We assessed recall as a function of the number of reads in a cluster and 2) assessed precision as a function of the distance between two junctions along both eigenvectors of the model's covariance.

To assess our ability to recover true clusters as a function of read counts for a particular sample, for each read count from 2 to 20, we clustered twenty sets of ten simulated read pair distributions of that read count and computed the recall as implemented in Scikit-Learn.

To assess our ability to distinguish two nearby junctions in a given dataset, we considered the precision on simulated distributions of 20 reads each spanning junctions spaced some distance apart. We tested precision along both the first and second eigenvectors of the covariance matrix, clustering ten simulated mixture distributions at

each of 150 evenly spaced distances between zero and four times the covariance matrix's standard deviation along the corresponding eigenvector. As the relationship between precision and distance is necessarily sigmoidal, we summarized the trend with the generalized logistic function $f(x) = A + \frac{K-A}{\left(1+Qe^{-B(x-M)}\right)^{1/v}}$ where A and K are the lower and upper asymptotes respectively. We used nonlinear least-squares regression to estimate the parameters Q, B, M, v, fitting with scipy.optimize.curvefit []. We set A=min(Precision) and K=1. The asymptotes complicate fitting, particularly when the initial parameters specified during optimization place the exponential phase of the initial guess outside of the exponential phase of the data. To avoid this, we chose as our initial guess Q=1, B=1, v=1 , such that $f(x) = A + \frac{K-A}{1+e^{M-x}}$ when x=M, $f(x) = A + \frac{1}{2}(K - A)$ and will be halfway between the two asymptotes, so we set M halfway between the largest X such that *y* is at the lower asymptote and the smallest X such that *y* is at the upper asymptote to ensure that the exponential phase of the initial guess falls near the exponential phase of the data. For all but two samples this initialization method found a good solution. In these two samples, we repeated the regression, setting M with the more robust initial guess $M = median(\{x: A < y(x) < K\})$, and obtained reasonable fits both times.

Identifying putative technical artifacts

Chimeric inserts are a common source of false positives when inferring structures from paired-end data. Because they arise when two fragments of the genome are joined by chance during library preparation, they will not produce a diagonal cluster of reads, like a true junction, but rather one or more identical read pairs. The presence

of PCR duplicates means that not all chimeric inserts will be singletons, but rather there may be multiple reads with identical 5'-ends. Thus, we required that inferred structures be supported by two or more reads with distinct 5' coordinates. Clusters that failed to meet this criterion were flagged as putative technical artifacts.

Estimating the location of junctions

Each cluster of read pairs should represent the junction of a structure in the sequenced genome, and the distance between each read pair and the true junction location is the gap size of that read pair. So, we estimated the location of the underlying junction as the average position of all read pairs in the cluster, offset by one-half the average gap size along both the X- and Y-axes. The offset is negative on reverse strands and positive on the forward strands (SF 1B).

Estimating the read depth distributions

Because the GC-content of read pairs can bias estimates of read depth (Dohm et al., 2008), we constructed the joint distribution of the read depth of a position and the expected GC-content of the PCR amplicon for a read pair spanning that position. We computed the expected GC-content of a read pair spanning each position in the reference genome using the read length and insert size distribution. We then combined the coverage and %GC signals into a separate two-dimensional histogram for each autosome arm, where %GC is divided into 1% bins. Rather than fit parametric distributions to the observed read depth frequencies, we employed the nonparametric kernel density estimator with a Gaussian kernel to account for the wide range of shapes

we observed. We chose an optimal bandwidth for each %GC bin separately by 2-fold cross validation (CV). In short, as we constructed a separate conditional read depth distribution for each autosome arm, we have four observations for each read count, which we assign to two complementary subsets at each iteration of CV; within each subset, the frequencies at each read count are averaged and the distributions are normalized to sum to one. Because the frequency of positions where read count = 0 is still partially contaminated by deletions and masked positions, we excluded this bin from CV to avoid oversmoothing. We constructed 100 such subsets, and cross-validated with respect to the mean-squared error (MSE) over a range of bandwidths between 0 and 100, performing an exponential search.

Estimating junction copy number

We estimated the copy number, *N*, of structures by modelling the read count, *X*, of a tandem junction as the sum of the reads originating from each of the *N* copies of the junction in the sequenced genome, where R_{gc} is the read count distribution for a single copy sequence with the same local GC-content as the junction:

$$X = \sum_{i=1}^{N} x_i, \qquad x_i \sim R_{gc}$$

To express beliefs about the copy number of a structure given the number of reads supporting it, we compute the posterior distribution as:

$$P(N|X) \propto P(X|N)$$

To compute the likelihood $P(X|N)$ we make use of the Central Limit Theorem (CLT). As read depth is modelled as the sum of independent and identically distributed

random variables, provided *N* is not very small, $P(X|N)$ should be well approximated by the Normal distribution $Normal(\mu_{gc}N, \sigma_{gc}\sqrt{N})$, where μ_{gc} and $\sigma_{gc}$are the mean and standard deviation of read depth over a single copy junction. In practice, even for small N we found that this provided a good estimate of the posterior distribution, so we employed this approximate likelihood rather than computing the more computationally intensive exact likelihood with convolutions of the read depth distribution. Thus:

$$P(X|N) = P(X), X \sim Normal(\mu_{gc}N, \sigma_{gc}\sqrt{N})$$

We used the posterior distribution to obtain maximum *a posterior* (MAP) estimates of copy number as well as to define 98% credible intervals around the MAP estimates. We computed the normalization factor for the posterior as the sum $\sum_{N=1}^{\lfloor X/\mu_{gc} \rfloor + 2000} P(X|N)$.

Identifying reads containing tandem junctions

Tandem dimers arising from double insertions should contain sequence from the target site within their tandem junctions. The length of intervening sequence may vary depending on the element, and non-LTR RTs are known to have variable TSDs (Linheiro and Bergman, 2012) and consequently, the amount of sequence derived from the tandem elements themselves may be small. We therefore used permissive alignment criteria to identify reads spanning tandem junctions, and then restricted ourselves to tandem junctions whose existence was also supported by read pair alignments. To identify reads that might contain tandem junctions, we went back to the raw sequencing reads for each sample and extracted all reads containing an 11-14 nucleotide seed from the 3’-end of each DNA and non-LTR TE. For DNA transposons,

the seed is the exact 3'-end, but as many non-LTR contain polyA tracts at the 3'-ends of their consensus sequences, we chose a 12-nt sequence within 20 nt of the polyA tract. Because these initial seeds were short we then BLASTed these reads against the repeat consensus sequences and assigned each read to the element with the best HSP as determined by e-value and then realigned each read to that element's consensus. If a read contains a tandem junction there should be two non-overlapping intervals homologous to the repeat, so we first selected the hit with the highest e-value, excluded any other alignments overlapping this hit and among the remaining hits chose the highest e-value alignment. If there were two equally good alignments, we chose the alignment which matched the strand of the first hit. We required that both alignments have e-values less than .05 and that they extend to within 3 nt of the read's end.

To identify intervening sequence, we extracted all sequence between the ends of the tandem repeats. To ensure a sequence and its reverse complement were treated as equivalent, we choose the lexicographically maximum string of each intervening sequence and its reverse complement. To exclude long polyA tracts from the intervening sequences of non-LTR RTs, we removed polyA and polyT tracts five nucleotides or more in length, along with all sequence 5' or 3', respectively.

Inferring euchromatic TE insertions from junctions

A TE insertion generates two junctions, one with either side of the insertion site. These junctions should fall in opposite quadrants of our scatterplots (e.g. (-,-) and (+,+) or (-,+) and (+,-)). Thus, identifying insertions from the set of junctions requires matching pairs of junctions. Because insertions in along the assembled chromosome arms are likely to be far apart, we favored a simple clustering strategy. Thus, we

employed single linkage agglomerative clustering with a distance cutoff of 300 nt to cluster all junctions of a given element in opposite quadrants of each chromosome arm. To ensure we did not miss insertions present in the reference--which will be masked in our reference genome--we disregarded masked nucleotides when computing distance. We estimated the insertion's location as the average estimated position of its junctions.

Mapping tandem dimers to specific TE insertions

To infer the location of tandem dimers, we needed to determine all instances of its intervening sequence in the reference genome and identify a locus containing both the intervening sequence and an insertion of that element. For intervening sequences 9-bp or longer, we used BLASTn with an evalue cutoff of 10 and accepted the top hit and all other hits whose evalues were within 2-orders of magnitude of the highest evalue. For intervening sequences 7- or 8-bp in length, we required an exact match to either the sequence or its reverse complement, and employed string matching. We did not attempt to map any dimer whose intervening sequence was less than 7-bp.

We matched each intervening sequence to a TE insertion if at least one of its the junction's estimated location in the reference was within 150-nt of the intervening sequence. We only considered dimers that could be uniquely mapped to a single insertion.

Gene annotation

We downloaded the RefSeq gene annotations for D. melanogaster from UCSC's Table Browser and excluded all computed genes and RNAs (entries named CG#### or CR####).

Aligning Pacbio reads

We aligned Pacbio reads to the repeat index using BLASTn. We used a linear gap penalty of 2 to account for the high rate of indels and imposed an evalue cutoff of .01.

RESULTS

ConTExt identifies repeats from paired-end Illumina data

Paired-end reads are powerful for detecting the junctions arising from structural rearrangements such as deletions, inversions, translocations , and tandem gene duplication (Bashir et al., 2008) (Rogers et al., 2014). Conceptually the problem of identifying the junctions of structures formed by known repeats is identical but the complication is that repeat derived reads can rarely be mapped to a unique locus in the reference genome. However, such reads can generally be uniquely mapped to a repeat family, and this has been successfully leveraged to identify TE insertions into unique sequence (Kofler et al., 2015) . We extend this idea to identify all classes of junctions involving repetitive sequence, including insertions into unique and repetitive sequence, deletions/inversion internal to a repeat, and tandems. Aligning to the set of all individual repeats present in a reference genome provides the power to detect reads originating from highly divergent variants, but results in reads being distributed across many different sequences. On the other hand, aligning reads to repeat consensus sequences is less powerful in detecting divergent copies, but not only identifies the family to which a read belongs, but also organizes all reads from that family in the same place, greatly

simplifying visualization and downstream analyses. We do both, first using Bowtie2 to align the reads to the set of all individual repeats extracted from the reference genome (including the unmapped contigs), and then converting these alignments from individual repeats to the corresponding consensus sequences (Fig 1A). In trial runs, we recovered about 20% more repeat derived reads using this two-step procedure than when we aligned only to consensus sequences, mostly originating from ancient repeats such as INE. We also aligned reads to the repeat-masked reference genome, allowing the detection of junctions between repeats and unique sequence, as well as estimation of the insert size and read depth distributions. Once reads are organized relative to consensus sequences, we use the alignment patterns to identify junctions (Fig 1B). While agglomerative clustering strategies are successful at identifying structural variation in unique sequence (Medvedev et al., 2009), the alignment patterns of repeat-derived reads are more challenging to resolve. This is because we are collapsing reads derived from up to megabases of sequence onto consensus sequences less than 10kb in length, and so read pairs representing distinct junctions are often crowded and sometimes interspersed. Mixture modelling, however, provides well-founded tools to resolve nearby and overlapping distributions. Each junction in the sequenced genome produces a distinct distribution of discordant read pairs (Fig 1C), and so the set of all discordant read pairs can be thought of as arising from a mixture of many distributions, each of which corresponds to a junction. We employ an accelerated Expectation-Maximization algorithm (Varadhan and Roland, 2008) to fit a mixture model to the set of discordant read pairs, clustering them and allowing us to resolve tightly grouped distributions (Fig 1C). This clustering strategy had >97% recall for junctions supported

at least three reads (SF 3C) and the ability to resolve nearby junctions was consistent across the GDL, speaking to the uniformity of the data's library preparations (SF 3A,B). Once we identify these clusters, we estimate the underlying junctions (Fig 1C) and visualize their distribution across all samples in the dataset (Fig 1D). We estimate the copy number from the read counts of each structure, in a manner that accounts for GC-bias in read depth.

Our pipeline allows us to detect a variety of structures involving repeats, including tandem junctions (Fig 1B, C), insertions into both unique and repetitive sequence (Fig 1E), and internal deletions (Fig 1B, C). While our focus here is on tandems, we note that we successfully identified known internally deleted elements, such as the Th HOBO variants (Fig 1D) (Periquet et al., 1994), the KP nonautonomous P-element (Black et al., 1987) , and the Roo internal deletion hotspot (Fig 3D) (Kaminker et al., 2002) . We also identified known nested repeats such as R-element insertions into the ribosomal RNA genes, including both full-length and distinct 5'-truncated insertions. The population genetics of these structures are under explored and ConTExt is well-suited to address this in future studies.

Transposable elements of all 3 major classes frequently form tandems

Most transposable elements can be detected in tandem in at least one sample, and the three major classes of TEs show distinct patterns of tandem junctions. LTR retrotransposons (LTR RTs) are flanked by direct repeats which provide a substrate for tandem formation, as recombination between LTRs yields structures where adjacent TEs share an LTR (Ke and Voytas, 1997) . The majority of LTR RTs can be detected in tandem (Fig 2C), though most solely involve internal sequence and are present at low-

copy number (SF 4C). These internal-to-internal tandems are consistent with deletions which span the junctions of head-to-tail tandems. We less frequently observe head-to-tail tandems, likely because we only have power to detect tandem LTR RTs when the LTR is shorter than ~330 nt (Fig 2C). However, all LTR retrotransposons with LTRs shorter than 300 nt are detected as head-to-tail tandems in at least one sample. Given this and the abundance of internal-to-internal tandems we conclude that most LTR retrotransposons frequently form tandems. We also find a few instances where tandem junctions are detected between the LTRs themselves; this might arise if much of the TE's internal sequence is lost in a deletion, or if two LTR RTs insert at the same locus.

Unlike LTR RTs, non-LTR RTs and DNA transposons do not provide their own substrates for unequal exchange, yet most can be detected as tandems, demonstrating that additional mechanisms allow tandem formation (Fig 2A,B). Of these, 6 are non-LTR RT TE subfamilies that form the *Drosophila* telomeres, and are known to form head-to-tail tandems whenever two elements of the same family insert consecutively at the same telomere (George et al., 2006). Non-LTR RTs are prone to 5'-truncation due to incomplete reverse transcription during transposition, and consistent with these tandems arising through consecutive insertion events at the same telomere, many telomeric TE tandem junctions are between the tail of one element and the internal region of another (Fig 2B, Fig 3A)). These include a range of distinct truncations, indicating many independent tandem forming events. Most other non-LTR RTs can be detected as tandems in at least one sample, and among these 5'-truncated insertions are abundant, suggesting that these tandems are also generated by transposition. We find tandems of 5' truncated elements at both low (e.g. 1/85) (Fig 3B) and intermediate

(e.g. 40/85) (Fig 3C) population frequency, suggesting that such structures are continuously generated.

DNA transposons are primarily detected as head-to-tail tandems (Fig 2A, SF 4A), however the distribution of junction estimates suggests that not all tandems involve exact head-to-tail junctions, rather some tandem elements are missing sequence from their ends. If junctions with identical coordinates are present in multiple strains, estimation error yields a tight diagonal distribution of junction estimates. This is what we observe for the recently formed P-element tandem junctions (Fig 3E), but Hobo, which is an older resident of the D. melanogaster genome, displays a more diffuse distribution, consistent with small deletions near or spanning the tandem junction, specific to each genome (Fig 3F). This may be a consequence of DNA transposon biology. The ends of DNA transposons are flanked by terminal inverted repeats (TIRs) which are endonuclease cut sites. Tandems lacking these TIRs may experience a reduced rate of DSB and be more stable over time. Consistent with this, all elements in the fixed Bari tandem array have incomplete TIRs (Marsano et al., 2003)(Hoskins et al., 2015).

TE tandems form rapidly during periods of active transposition

The known timeline of P-element's arrival in *D. melanogaster* makes P-element tandems particularly informative of the dynamics and mechanism of tandem formation. *D. melanogaster* stocks collected at the beginning of the 20th century contain no P-elements, but all stocks collected after the 1950s contain dozens (Bingham et al., 1982; Engels, 1992). Given this recent invasion, that over half of the GDL strains contain tandem P-elements, indicates that tandem TEs form rapidly during periods of high transpositional activity (Fig 4A). Head-to-head P-element tandems have been previously

described as generated during a genetic screen (Tower et al., 1993) , but the majority of tandems we observed are head-to-tail, with head-to-head tandems only being observed in 8 strains. This discrepancy may be due to selection removing head-to-head tandems from natural populations over short timescales, as long inverted repeats are prone to forming cruciform DNA secondary structures (Leach, 1994) . Alternatively, it is conceivable that amplicons containing head-to-head tandem junctions form hairpin secondary structures decreasing their PCR amplification efficiency leading to their underrepresentation in sequencing libraries. We reasoned that if TE tandems are driven by multiple insertions at the same site, the tandem junction should contain a target site duplication (TSD) of the insertion site, as has been observed in bacterial DNA transposon dimers (Dalrymple, 1987). We examined reads containing exact head-to-tail junctions and searched for the presence of intervening motifs. The majority of junction sequences identified contain intervening sequence, indicating that most P-element tandems formed through double insertion at a locus. Most such sequences are eight nucleotides in length as is the known P-element TSD and some are are present in the tandem junctions of more than one strain; given P-element's recent arrival, it seems unlikely that these represent alleles identical by descent, but rather likely speak to the limited set of loci into which P-elements insert. We generated a consensus motif (GTCTAGAG) from the 8-mers we identified (Fig 4C); it is quasi-palindromic and nearly identical to the TSD motif identified by Liao and Rubin from an examination of 1,469 P-element insertions (Fig 4D) (Liao, 2000). However, the motifs found at tandem junctions have a higher information content at each position, particularly at the first two

nucleotides, (sign test, *p=.008*), suggesting that P-element tandems are more likely to form at sites that more closely match its preferred target sequence.

Short TE tandems can expand into larger arrays

The abundance of TE tandems that we discovered potentially provides the substrate for expansion by unequal exchange. We discovered a candidate for such an expansion for the DNA transposon Hobo where most strains contain few or no Hobo tandems but Ithacan line I03 contains a copy number expansion estimated at between 13 and 20 copies (Fig 4B). To rule out that this represents multiple independently formed tandems, we again searched for reads containing exact head-to-tail tandem junctions. Unlike with the P-element, we only identified reads containing exact head-to-tail Hobo tandem junctions in four strains, consistent with the distribution of junction estimates which suggested that many Hobo tandems involve elements with deleted terminal sequence. Uniquely in I03, however, we found many such reads indicating that an identical 8-nt motif (GTGGGGAC) resides between the TIRs of the tandem Hobos, strongly suggesting that all elements of the array descend from the same tandem dimer. We mapped the tandem to chromosome 2L, by identifying all Hobo insertions in I03 and then finding the only locus which contains both a Hobo insertion and the motif. We identified only one locus in the reference genome that contained both the motif (position 17,943,032 of chromosome 2L) and a Hobo insertion (somewhere between positions 17,942,976 and 17,943,104 of 2L), suggesting that the Hobo tandem array is found on 2L at position 17,943,032 of the reference approximately 19.5 kb away from the protein coding gene *Beethoven* and near several weakly supported noncoding RNAs (Fig 4E). I03 is the only strain containing a Hobo insertion at this position, indicating that the

tandem likely formed from a recent Hobo insertion. We suggest that this Hobo array reflects the same expansion process that gave rise to fixed, high-copy TE tandems like that R1 and Bari tandem arrays, and indicates that the abundance of low-copy TE tandems represents source material from which larger tandem arrays can arise.

Copy distributions of long tandem repeats

Such expansion events can eventually give rise to very large arrays and become fixed. There are two previously described TE-derived satellites in *D. melanogaster*. The Bari tandem array is comprised of tandemly arrayed copies of the 1.7 kb DNA transposon Bari and is located in the pericentromeric heterochromatin of the right arm chromosome 2 (Caizzi et al., 1993). Previous analyses identified the array in all strains examined (n=10) and estimated its copy number at ~80 repeat units (Caggese et al., 1995) . We also find the Bari array in all 85 GDL strains, ranging from 32 to 130 copies, corresponding to an array ranging from ~54,000 to 220,000 basepairs in length. (Fig 5A). There is little population differentiation, though the Zimbabwe strains are the most variable in copy number. The other known TE-derived satellite is comprised by R1 elements, which insert only a specific site in the 28S ribosomal RNA gene. This tandem was first identified in a large-insert clone containing five tandemly arrayed copies (Kidd and Glover, 1980) and was mapped to the X heterochromatin (Peacock et al., 1981) . We find that the tandem is fixed in the GDL and identify two different high-copy junctions corresponding to the tandem. Both junctions involve sequence from near the 5’ and 3’ ends of R1. BLASTing ISO-1 PacBio reads to the R1 consensus identifies many reads entirely comprised of tandemly arrayed R1 elements and in which both high

copy junctions are interspersed (SF 5A) (Kim et al., 2014). These two junctions together are present on average at a total of 199 copies, ranging from 91 to 273, with Zimbabwe again being the most variable population (Fig 5B). This range corresponds to an array ranging from ~500,000 to ~1,300,000 basepairs in length. We also identify on average 26 tandem junctions consistent with deletions spanning a tandem junction (SF 5B). The number of such junctions is positively correlated with the copy number of the two major junctions (r=.72, p<1e-3), indicating that these fall within the R1 array and suggest an older tandem array where the repeat units are degenerating.

Genomic distribution of tandem dimers

Given that we were able to map the *Hobo* tandem array in I03 using its TSD, we reasoned we might be able to map other dimers to examine their distribution across the genome (Fig 6 A-F). Two related factors complicated this effort. First, many DNA transposon produce small TSDs, meaning that only limited information about their location is contained within the tandem junction. Second, the tandem junctions of non-LTR RT dimers often contain TSDs as well as retrotranscribed polyA tracts of variable length which are uninformative of their location, meaning that a 100-bp read containing a tandem junction may include only short segments of the TEs themselves and provide little indication that the read truly represents a tandem junction. Nonetheless we successfully mapped a number of dimers.

P-elements comprise the majority of mapped dimers, followed by Jockey elements. Despite the abundance of Protop dimers, the 7-nt nucleotide TSDs were insufficient to uniquely map dimers to particular Protop insertions. P-element dimers are significantly closer to the transcription start sites (TSSs) of genes (median distance =

169 bp) than are single insertions (median distance = 430 bp) (Fig 6G). P-elements preferentially insert near the promoters of genes (Spradling et al., 2011), thus the enrichment of dimers near TSSs supports the idea that dimers form at strong insertion sites. This is consistent with the higher information content observed at TSDs within P-element dimer junctions. Indeed, several of the tandems we mapped were adjacent to genes previously identified as the strongest P-element insertion hotspots (Spradling et al., 2011) . Further highlighting the degree to which insertion site preference drives tandem formation, we observe several instances where multiple dimers formed independently near the same gene. The distance between the 11 Jockey dimers we mapped and the nearest gene was comparable to that of single insertions (Fig 6H). Moreover, these dimers confirmed that tandem non-LTR RTs frequently involve elements with distinct 5'-trunctations, further suggesting that these dimers arise by double insertions (Fig 6 A-F).

Discussion

ConTExt successfully identifies repetitive structures in NGS data

Despite the number of NGS population genomics datasets, leveraging them to learn about highly repeated sequence remains a challenging problem. We employed an alignment strategy that maps repeat-derived reads to repeat consensus sequences and used mixture modelling to interpret the patterns. The main advantage of this approach is that there are efficient algorithms for fitting mixture models that take into account the locations of every data point when making clustering choices. This matters less for variant discovery in single-copy sequence, but is necessary if one wishes the resolve

similar but distinct structures within repetitive DNA. Applying this method to a panel of five populations, we observed distinct processes of satellite evolution on-going within a single species. We detected a number of tandem structures, both previously-described and novel, including tandem junctions amongst all telomeric TE families, and large tandem arrays of the *Bari* and *R1* elements. In addition to detecting tandems, we also were able to identify internally deleted TEs as well as nested insertions, such as R-elements inserted into the rDNA.

The strategies we used have some limitations imposed by our reliance upon sequence alignments. Repeat derived reads can only rarely be aligned uniquely to the reference genome which means that in most cases we cannot determine the location of the structures we identify. Further, reliance on consensus sequences to identify repeat-derived reads limits our survey to known repeat families. Structure inference from paired-end alignments has several additional limitations. First, we do not have the power to detect junctions containing intervening sequence longer than the mate pair distance of the library, such as LTRs. Second, chimeric inserts in paired-end data can produce spurious structure discovery calls. To account for this we only considered structures supported by multiple read pairs with distinct coordinates. However, it is unlikely that the general patterns we observe are false positives, as it is more likely that false positives would be dispersed across the consensus sequences rather than concentrated in biologically plausible patterns specific to TE classes. Further, our inability to detect tandems of elements with LTRs longer than the insert size also suggests that false positives do not contribute substantially to patterns of junctions we observe—if head-to-tail dimers were technical artifacts, the length of LTRs would have

no impact on the frequency with which they are detected. Finally, there will be false negatives whenever a dimer is not spanned by any read pairs in a sample where it is truly present.

Mechanisms of tandem formation and evolution

The processes by which TEs transition to satellites are actively ongoing in *D. melanogaster* and continuously add to the library of sequence from which new satellites may arise. Tandem TE dimers from which large arrays can expand are common, with multiple TE dimers being present in most *D. melanogaster* genomes. Several observations suggest that multiple insertions of a TE family at the same site is the amplification process generating dimers. First, many non-LTR RT tandem junctions are between the intact 3' end of one element and the 5'-truncated end of the adjacent element. 5'-truncations arise during integration when reverse transcription terminates prematurely. While a variety of processes can tandemly duplicate sequence, such amplification events are expected to involve two random sequence coordinates. On the other hand, non-LTR RT tandem junctions resulting from integration events will always have an intact 3'-end but may have 5'-truncations. Thus the pattern of non-LTR tandems, where a variety of tandem elements with different 5'-truncations are all adjacent to complete 3'-ends of the same TE family is most readily explained by double insertions. Further, we were able to map several non-LTR RT dimers, and in many cases these involved elements with distinct 5'-truncations. Second, the majority of P-element tandems contain intervening sequence consistent with P-element's known target motif. Double insertions are possible because many TEs induce a staggered cut

at their insertion site, which when repaired yields a pair of direct repeats flanking the insertion. These target site duplications mean that a new TE insertion does not consume its target site but rather preserves it while generating a new one. If TE dimers form through transposition events means then high periods of TE activity should also give high rates of tandems. Indeed, despite only invading the species in the last century, most strains in the GDL contain tandem P-elements. Thus, for those TEs with insertion site preference, the amplification events which form tandem dimers are common and almost guaranteed by the presence of active copies. In contrast, the products of subsequent copy number expansions are much rarer. We observed one such event, a copy number expansion of a *Hobo* dimer in a single line, I03. This rarity may reflect either a very low per generation rate of copy number expansions or instead their instability.

Tandem persistence is dependent upon recombination rate

We found evidence that differential persistence shapes the set of tandems present in a species. Tandem DNA transposons, in particular, are likely to be unstable as their TIRs are endonuclease cut sites, making them subject to high rates of double strand breaks and unequal exchange. We found that most recently formed *P*-element dimers have intact TIRs, but that most *Hobo* dimers lack some terminal sequence. High-copy DNA transposon arrays also lack intact TIRs. We confirmed that the tandemly arrayed *Bari* elements from the assembled genome are missing four nucleotides of sequence from their TIRs (Hoskins et al., 2015; Marsano et al., 2003) . Likewise, a large *P*-element array in *D. guanche* is comprised of elements missing ~100 nt of sequence

from their ends (Miller et al., 1992). Charlesworth et al proposed that tandem arrays in regions with high recombination rates should be lost more rapidly than those in low recombination regions, shaping the genome-wide distribution of satellite sequence (Charlesworth et al., 1986). That only the very recent P-element tandem dimers have intact TIRs while older tandems lack them suggests that this same principle appears to shape the sequence evolution of tandem DNA transposons by favoring the persistence of tandem arrays without the high rate of DSBs induced by intact TIRs, and perhaps more generally any satellite with sequence characteristics that reduce the rate of unequal exchange.

Tandem TEs may impact gene regulation

In addition to providing material for new satellite sequence, that TEs contain intrinsic regulatory elements and are targets for heterochromatin deposition make tandem transposable elements interesting entities in their own right, as emerging tandems may have functional impacts on gene regulation. TE insertions are commonly heterochromatinized by the piRNA pathway (Brennecke et al., 2007) and these silencing marks can spread into nearby genes, reducing their expression (Lee, 2015) . Further, tandem P-element transgenes induce position effect variegation, the strength of which increases with copy number (Dorer and Henikoff, 1994). It is likely, then, that the epigenetic impacts of tandem dimers on the expression of nearby genes is stronger than those of single insertions and that unequal exchange can further intensify this by copy number expansions. Additionally, tandems residing within piRNA clusters could also have global impacts on the activity of the corresponding TE family as well as genes

adjacent to insertions by massively increasing the dosage of piRNAs targeting that element. Conversely, Loehlin et al recently described nonlinear increases in the expression of recently duplicated genes, and suggested that this may be consequence of concentrating regulatory elements (Loehlin and Carroll, 2016). If true for tandem TEs, this may in some circumstances (e.g. before a TE family is recognized by piRNA pathway) increase the activity of tandemly arranged TEs.

ConTExt makes the population genomics of highly repetitive sequence tractable

While we focused on understanding the origination of TE-derived satellite sequence, the method we developed opens the door to addressing a variety of questions about repeat evolution. First, the copy number variation we observe at established TE tandem arrays provides empirical data to which population genetic models of repeat evolution may be fit to elucidate the forces and processes driving their evolution. Second, internal deletions are common to non-autonomous DNA transposons (e.g. Hobo, P-element) which can rapidly replace autonomous elements (Bingham et al., 1982) (Le Rouzic et al., 2007) and engage in hyperparasitic behavior (Robillard et al., 2016). ConTExt offers the possibility of disentangling the copy number dynamics of internally deleted and full-length TEs. Finally, piRNA clusters are commonly comprised of nested and rearranged repeats, which are readily detected by ConTExt and may allow surveys of piRNA cluster variation.

Conclusion

Many satellite sequences clearly originated as TEs and it is unclear how this transition from mobile genetic parasite to sometimes megabase-sized array occurs. To shed light on this source of new satellite sequence we developed a bioinformatic pipeline to utilize paired-end data in a population genomic survey of tandem repeats in *D. melanogaster*. Through this survey we found that transposable element dimers are common, originate by double insertion into the same locus, and can expand into larger arrays. We propose that this is the common mechanism by which TEs without intrinsic repeats generate satellite arrays.

Literature Cited

Aldrup-MacDonald, M.E., Kuo, M.E., Sullivan, L.L., Chew, K., Sullivan, B.A., 2016. Genomic variation within alpha satellite DNA influences centromere location on human chromosomes with metastable epialleles. Genome Res. 26, 1301–1311. doi:10.1101/gr.206706.116

Altschul, S.F., Gish, W., Miller, W., Myers, E.W., Lipman, D.J., 1990. Basic local alignment search tool. J. Mol. Biol. 215, 403–410. doi:10.1016/S0022-2836(05)80360-2

Bao, W., Kojima, K.K., Kohany, O., 2015. Repbase Update, a database of repetitive elements in eukaryotic genomes. Mob. DNA 6, 11.

Bashir, A., Volik, S., Collins, C., Bafna, V., Raphael, B.J., 2008. Evaluation of Paired-End Sequencing Strategies for Detection of Genome Rearrangements in Cancer. PLoS Comput. Biol. 4, e1000051. doi:10.1371/journal.pcbi.1000051

Bingham, P.M., Kidwell, M.G., Rubin, G.M., 1982. The molecular basis of P-M hybrid dysgenesis: The role of the P element, a P-strain-specific transposon family. Cell 29, 995–1004. doi:10.1016/0092-8674(82)90463-9

Black, D.M., Jackson, M.S., Kidwell, M.G., Dover, G.A., 1987. KP elements repress P-induced hybrid dysgenesis in Drosophila melanogaster. EMBO J. 6, 4125–4135.

Blondel, V.D., Guillaume, J.-L., Lambiotte, R., Lefebvre, E., 2008. Fast unfolding of communities in large networks. J. Stat. Mech. Theory Exp. 2008, P10008.

Bolger, A.M., Lohse, M., Usadel, B., 2014. Trimmomatic: a flexible trimmer for Illumina sequence data. Bioinformatics 30, 2114–2120.

Brennecke, J., Aravin, A.A., Stark, A., Dus, M., Kellis, M., Sachidanandam, R., Hannon, G.J., 2007. Discrete Small RNA-Generating Loci as Master Regulators of Transposon Activity in Drosophila. Cell 128, 1089–1103. doi:10.1016/j.cell.2007.01.043

Caggese, C., Pimpinelli, S., Barsanti, P., Caizzi, R., 1995. The distribution of the transposable elementBari-1 in theDrosophila melanogaster andDrosophila simulans genomes. Genetica 96, 269–283. doi:10.1007/BF01439581

Caizzi, R., Caggese, C., Pimpinelli, S., 1993. Bari-1, a new transposon-like family in Drosophila melanogaster with a unique heterochromatic organization. Genetics 133, 335–345.

Charlesworth, B., Langley, C.H., Stephan, W., 1986. The Evolution of Restricted Recombination and the Accumulation of Repeated Dna Sequences. Genetics 112, 947–962.
Charlesworth, B., Sniegowski, P., Stephan, W., 1994. The evolutionary dynamics of repetitive DNA in eukaryotes. Nature 371, 215–220. doi:10.1038/371215a0
Dalrymple, B., 1987. Novel rearrangements of IS30 carrying plasmids leading to the reactivation of gene expression. Mol. Gen. Genet. MGG 207, 413–420.
Dias, G.B., Svartman, M., Delprat, A., Ruiz, A., Kuhn, G.C., 2014. Tetris is a foldback transposon that provided the building blocks for an emerging satellite DNA of Drosophila virilis. Genome Biol. Evol. 6, 1302–1313.
Dohm, J.C., Lottaz, C., Borodina, T., Himmelbauer, H., 2008. Substantial biases in ultra-short read data sets from high-throughput DNA sequencing. Nucleic Acids Res. 36, e105–e105.
Dorer, D.R., Henikoff, S., 1994. Expansions of transgene repeats cause heterochromatin formation and gene silencing in Drosophila. Cell 77, 993–1002.
Engels, W.R., 1992. The origin of P elements in Drosophila melanogaster. BioEssays 14, 681–686. doi:10.1002/bies.950141007
Ferree, P.M., Barbash, D.A., 2009. Species-specific heterochromatin prevents mitotic chromosome segregation to cause hybrid lethality in Drosophila. PLoS Biol. 7, e1000234.
Fishman, L., Saunders, A., 2008. Centromere-Associated Female Meiotic Drive Entails Male Fitness Costs in Monkeyflowers. Science 322, 1559–1562. doi:10.1126/science.1161406
Francisco, F.O., Lemos, B., 2014. How Do Y-Chromosomes Modulate Genome-Wide Epigenetic States: Genome Folding, Chromatin Sinks, and Gene Expression. J. Genomics 2, 94–103. doi:10.7150/jgen.8043
George, J.A., DeBaryshe, P.G., Traverse, K.L., Celniker, S.E., Pardue, M.-L., 2006. Genomic organization of the Drosophila telomere retrotransposable elements. Genome Res. 16, 1231–1240.
Gong, Z., Wu, Y., Koblížková, A., Torres, G.A., Wang, K., Iovene, M., Neumann, P., Zhang, W., Novák, P., Buell, C.R., others, 2012. Repeatless and repeat-based centromeres in potato: implications for centromere evolution. Plant Cell 24, 3559–3574.
Grenier, J.K., Arguello, J.R., Moreira, M.C., Gottipati, S., Mohammed, J., Hackett, S.R., Boughton, R., Greenberg, A.J., Clark, A.G., 2015. Global diversity lines–a five-continent reference panel of sequenced Drosophila melanogaster strains. G3 Genes Genomes Genet. 5, 593–603.
Hardy, R.W., Lindsley, D.L., Livak, K.J., Lewis, B., Siversten, A.L., Joslyn, G.L., Edwards, J., Bonaccorsi, S., 1984. CYTOGENETIC ANALYSIS OF A SEGMENT OF THE *Y* CHROMOSOME OF *DROSOPHILA MELANOGASTER*. Genetics 107, 591–610.
Hoskins, R.A., Carlson, J.W., Wan, K.H., Park, S., Mendez, I., Galle, S.E., Booth, B.W., Pfeiffer, B.D., George, R.A., Svirskas, R., others, 2015. The Release 6 reference sequence of the Drosophila melanogaster genome. Genome Res. 25, 445–458.
Kaminker, J.S., Bergman, C.M., Kronmiller, B., Carlson, J., Svirskas, R., Patel, S., Frise, E., Wheeler, D.A., Lewis, S.E., Rubin, G.M., others, 2002. The transposable elements of the Drosophila melanogaster euchromatin: a genomics perspective. Genome Biol. 3, research0084–1.
Ke, N., Voytas, D.F., 1997. High Frequency Cdna Recombination of the Saccharomyces Retrotransposon Ty5: The Ltr Mediates Formation of Tandem Elements. Genetics 147, 545–556.
Kidd, S.J., Glover, D.M., 1980. A DNA segment from D. melanogaster which contains five tandemly repeating units homologous to the major rDNA insertion. Cell 19, 103–119.

Kim, K.E., Peluso, P., Baybayan, P., Yeadon, P.J., Yu, C., Fisher, W., Chin, C.-S., Rapicavoli, N.A., Rank, D.R., Li, J., others, 2014. Long-read, whole genome shotgun sequence data fro five model organisms. bioRxiv 008037.
Kofler, R., Nolte, V., Schlötterer, C., 2015. Tempo and mode of transposable element activity in Drosophila. PLoS Genet. 11, e1005406.
Langmead, B., Salzberg, S.L., 2012. Fast gapped-read alignment with Bowtie 2. Nat. Methods 9, 357–359. doi:10.1038/nmeth.1923
Le Rouzic, A., Boutin, T.S., Capy, P., 2007. Long-term evolution of transposable elements. Proc. Natl. Acad. Sci. 104, 19375–19380.
Leach, D.R., 1994. Long DNA palindromes, cruciform structures, genetic instability and secondary structure repair. Bioessays 16, 893–900.
Lee, Y.C.G., 2015. The role of piRNA-mediated epigenetic silencing in the population dynamics of transposable elements in Drosophila melanogaster. PLoS Genet. 11, e1005269.
Liao, G. -c., 2000. Insertion site preferences of the P transposable element in Drosophila melanogaster. Proc. Natl. Acad. Sci. 97, 3347–3351. doi:10.1073/pnas.050017397
Linheiro, R.S., Bergman, C.M., 2012. Whole Genome Resequencing Reveals Natural Target Site Preferences of Transposable Elements in Drosophila melanogaster. PLOS ONE 7, e30008. doi:10.1371/journal.pone.0030008
Loehlin, D.W., Carroll, S.B., 2016. Expression of tandem gene duplicates is often greater than twofold. Proc. Natl. Acad. Sci. 113, 5988–5992. doi:10.1073/pnas.1605886113
Lohe, A.R., Roberts, P.A., 2000. Evolution of DNA in Heterochromatin: the Drosophila Melanogaster Sibling Species Subgroup as a Resource. Genetica 109, 125–130. doi:10.1023/A:1026588217432
Macas, J., Koblížková, A., Navrátilová, A., Neumann, P., 2009. Hypervariable 3′ UTR region of plant LTR-retrotransposons as a source of novel satellite repeats. Gene 448, 198–206.
Marsano, R.M., Milano, R., Minervini, C., Moschetti, R., Caggese, C., Barsanti, P., Caizzi, R., 2003. Organization and Possible Origin of the Bari-1 Cluster in the Heterochromatic h39 Region of Drosophila Melanogaster. Genetica 117, 281–289. doi:10.1023/A:1022916817285
Medvedev, P., Stanciu, M., Brudno, M., 2009. Computational methods for discovering structural variation with next-generation sequencing. Nat. Methods 6, S13–S20. doi:10.1038/nmeth.1374
Meštrović, N., Mravinac, B., Pavlek, M., Vojvoda-Zeljko, T., Šatović, E., Plohl, M., 2015. Structural and functional liaisons between transposable elements and satellite DNAs. Chromosome Res. 23, 583–596. doi:10.1007/s10577-015-9483-7
Miller, W.J., Hagemann, S., Reiter, E., Pinsker, W., 1992. P-Element Homologous Sequences are Tandemly Repeated in the Genome of Drosophila guanche. Proc. Natl. Acad. Sci. U. S. A. 89, 4018–4022.
Peacock, W.J., Appels, R., Endow, S., Glover, D., 1981. Chromosomal distribution of the major insert in Drosophila melanogaster 28S rRNA genes. Genet. Res. 37, 209–214.
Pedregosa, F., Varoquaux, G., Gramfort, A., Michel, V., Thirion, B., Grisel, O., Blondel, M., Prettenhofer, P., Weiss, R., Dubourg, V., others, 2011. Scikit-learn: Machine learning in Python. J. Mach. Learn. Res. 12, 2825–2830.
Periquet, G., Lemeunier, F., Bigot, Y., Hamelin, M.H., Bazin, C., Ladevèze, V., Eeken, J., Galindo, M.I., Pascual, L., Boussy, I., 1994. The evolutionary genetics of thehobo transposable element in theDrosophila melanogaster complex. Genetica 93, 79–90. doi:10.1007/BF01435241
Prosser, J., Frommer, M., Paul, C., Vincent, P.C., 1986. Sequence relationships of three human satellite DNAs. J. Mol. Biol. 187, 145–155. doi:10.1016/0022-2836(86)90224-X

Prudhomme, M., Turlan, C., Claverys, J.-P., Chandler, M., 2002. Diversity of Tn4001 Transposition Products: the Flanking IS256 Elements Can Form Tandem Dimers and IS Circles. J. Bacteriol. 184, 433–443. doi:10.1128/JB.184.2.433-443.2002

Robillard, É., Le Rouzic, A., Zhang, Z., Capy, P., Hua-Van, A., 2016. Experimental evolution reveals hyperparasitic interactions among transposable elements. Proc. Natl. Acad. Sci. 113, 14763–14768. doi:10.1073/pnas.1524143113

Rogers, R.L., Cridland, J.M., Shao, L., Hu, T.T., Andolfatto, P., Thornton, K.R., 2014. Landscape of standing variation for tandem duplications in Drosophila yakuba and Drosophila simulans. Mol. Biol. Evol. 31, 1750–1766.

Roiha, H., Miller, J.R., Woods, L.C., Glover, D.M., 1981. Arrangements and rearrangements of sequences flanking the two types of rDNA insertion in D. melanogaster. Nature 290, 749–754.

Smit, A.F.A., Hubley, R., Green, P., 2015. RepeatMasker Open-4.0. 2013–2015. Inst. Syst. Biol. Httprepeatmasker Org.

Spradling, A.C., Bellen, H.J., Hoskins, R.A., 2011. Drosophila P elements preferentially transpose to replication origins. Proc. Natl. Acad. Sci. U. S. A. 108, 15948–15953. doi:10.1073/pnas.1112960108

Stage, D.E., Eickbush, T.H., 2009. Origin of nascent lineages and the mechanisms used to prime second-strand DNA synthesis in the R1 and R2 retrotransposons of Drosophila. Genome Biol. 10, R49.

Stage, D.E., Eickbush, T.H., 2007. Sequence variation within the rRNA gene loci of 12 Drosophila species. Genome Res. 17, 1888–1897.

Strachan, T., Coen, E., Webb, D., Dover, G., 1982. Modes and rates of change of complex DNA families of Drosophila. J. Mol. Biol. 158, 37–54. doi:10.1016/0022-2836(82)90449-1

Tautz, D., Schlötterer, C., 1994. Simple sequences. Curr. Opin. Genet. Dev. 4, 832–837. doi:10.1016/0959-437X(94)90067-1

Tower, J., Karpen, G.H., Craig, N., Spradling, A.C., 1993. Preferential transposition of Drosophila P elements to nearby chromosomal sites. Genetics 133, 347–359.

Ugarković, Đ., Plohl, M., 2002. Variation in satellite DNA profiles—causes and effects. EMBO J. 21, 5955–5959. doi:10.1093/emboj/cdf612

Varadhan, R., Roland, C., 2008. Simple and Globally Convergent Methods for Accelerating the Convergence of Any EM Algorithm. Scand. J. Stat. 35, 335–353. doi:10.1111/j.1467-9469.2007.00585.x

Wei, K.H.-C., Grenier, J.K., Barbash, D.A., Clark, A.G., 2014. Correlated variation and population differentiation in satellite DNA abundance among lines of Drosophila melanogaster. Proc. Natl. Acad. Sci. 111, 18793–18798.

Zanni, V., Eymery, A., Coiffet, M., Zytnicki, M., Luyten, I., Quesneville, H., Vaury, C., Jensen, S., 2013. Distribution, evolution, and diversity of retrotransposons at the flamenco locus reflect the regulatory properties of piRNA clusters. Proc. Natl. Acad. Sci. 110, 19842–19847.

Zeng, W., de Greef, J.C., Chen, Y.-Y., Chien, R., Kong, X., Gregson, H.C., Winokur, S.T., Pyle, A., Robertson, K.D., Schmiesing, J.A., Kimonis, V.E., Balog, J., Frants, R.R., Ball, A.R., Lock, L.F., Donovan, P.J., van der Maarel, S.M., Yokomori, K., 2009. Specific loss of histone H3 lysine 9 trimethylation and HP1gamma/cohesin binding at D4Z4 repeats is associated with facioscapulohumeral dystrophy (FSHD). PLoS Genet. 5, e1000559. doi:10.1371/journal.pgen.1000559

Zhang, H., Koblížková, A., Wang, K., Gong, Z., Oliveira, L., Torres, G.A., Wu, Y., Zhang, W., Novák, P., Buell, C.R., Macas, J., Jiang, J., 2014. Boom-Bust Turnovers of Megabase-Sized Centromeric DNA in Solanum Species: Rapid Evolution of DNA Sequences Associated with Centromeres. Plant Cell 26, 1436–1447. doi:10.1105/tpc.114.123877

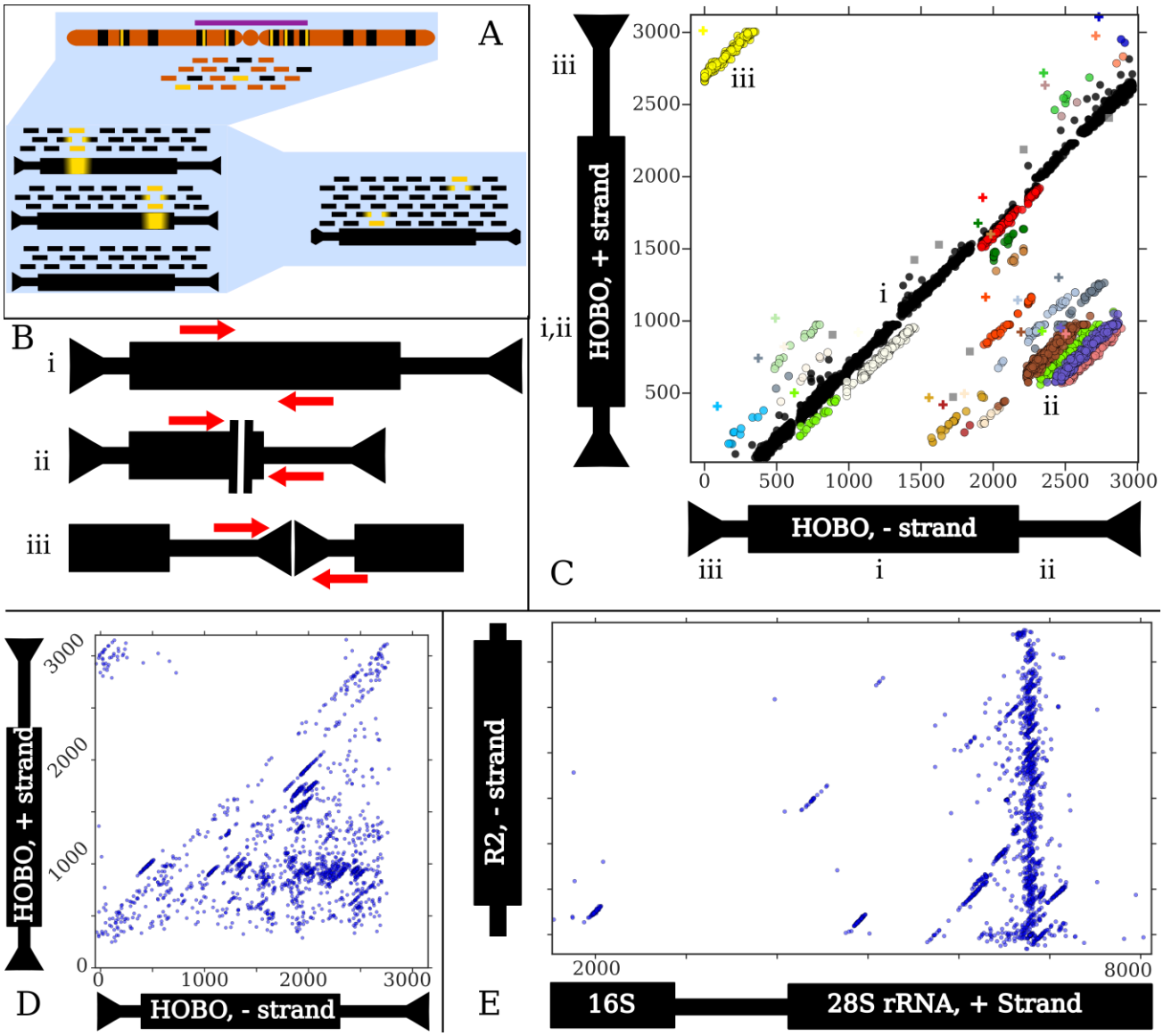


Figure 1 (Main Text Figure, slide 1) : An outline of the ConTExt pipeline and examples of identified structures. A) Sequencing reads are derived from genomic DNA, which has many copies of a particular repeat family (black) dispersed amongst miscellaneous sequence (orange); some instances of the repeat contain divergent regions relative to the consensus (yellow bars), particularly those in the pericentromeric heterochromatin (purple bar). The repeat-derived reads are aligned to a set of individual repeats identified in the reference genome, which includes instances of divergent elements, providing increased power to align reads. Alignments to these individual elements are then collapsed onto a consensus sequence for that repeat family. B) Schematics of paired end reads spanning i) a region of sequence concordant with the consensus, ii) the junction of an internal deletion, and iii) the junction of a head-to-tail tandem. C) Paired-end read alignments from strain I03 to the Hobo element represented as a two-dimensional scatterplot. Each dot represents the 3' end of a read, and its position on X- and Y-axes corresponds to its aligned position on the minus and plus strands of the Hobo consensus, respectively. The Roman numerals indicate how the 3 types of structures shown in B are identified within the scatter plot as well as where the reads map on each of the axes. i) Concordant reads which form the central diagonal, ii) reads spanning internal deletions, iii) reads spanning head-to-tail tandem junctions. Black denotes concordant reads, other colors correspond to clusters identified by the EM algorithm, and grey squares are reads flagged as possible artifacts. The plus symbols indicate the estimated junction location for the cluster with the corresponding color. Note that read pairs where both ends map to the same strand (e.g. head-to-head tandems) require a different scatterplot. D) A scatter plot representing the distribution of junctions involving Hobo across all GDL samples. Each dot represents a junction estimated from an identified cluster. E) A similar plot, depicting all junctions between the minus-strand of R2 and the plus-strand of rDNA.

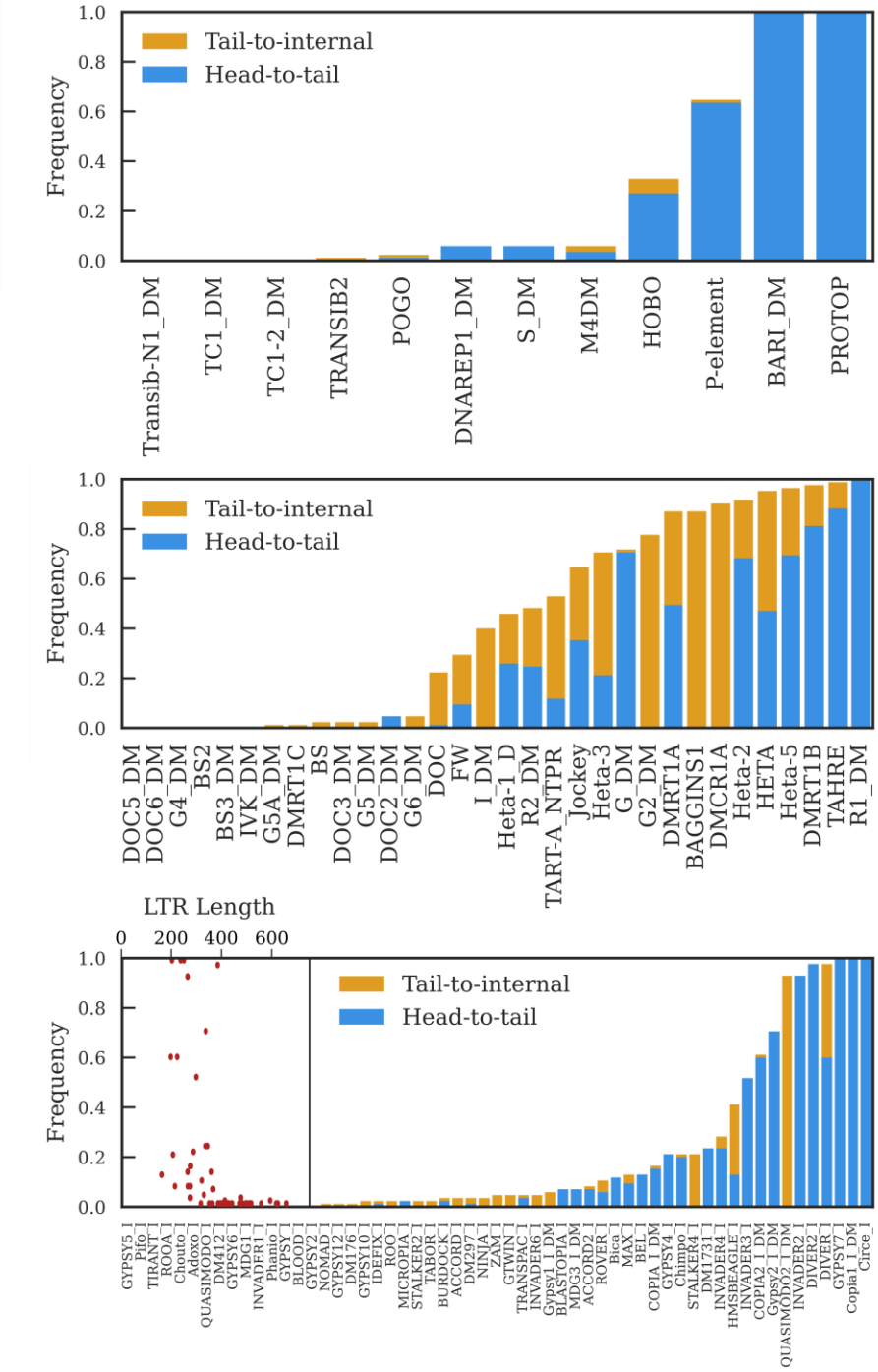


Figure 2: The proportion of samples in which a tandem junction was identified for A) DNA transposon families, B) non-LTR retrotransposon families, and C) LTR retrotransposon families. Head-to-tail tandems have junctions involving the first and last 200-nt of the consensus sequence. Tail-to-internal junctions have junctions between the last 200-nt of the consensus sequence and internal sequence; are consistent with tandems involving 5'-truncated elements, though they can also be formed by nested insertions. The scatter plot inset in C depicts the relation between LTR length and the frequency with we detect an LTR RT family as a head-to-tail tandem.

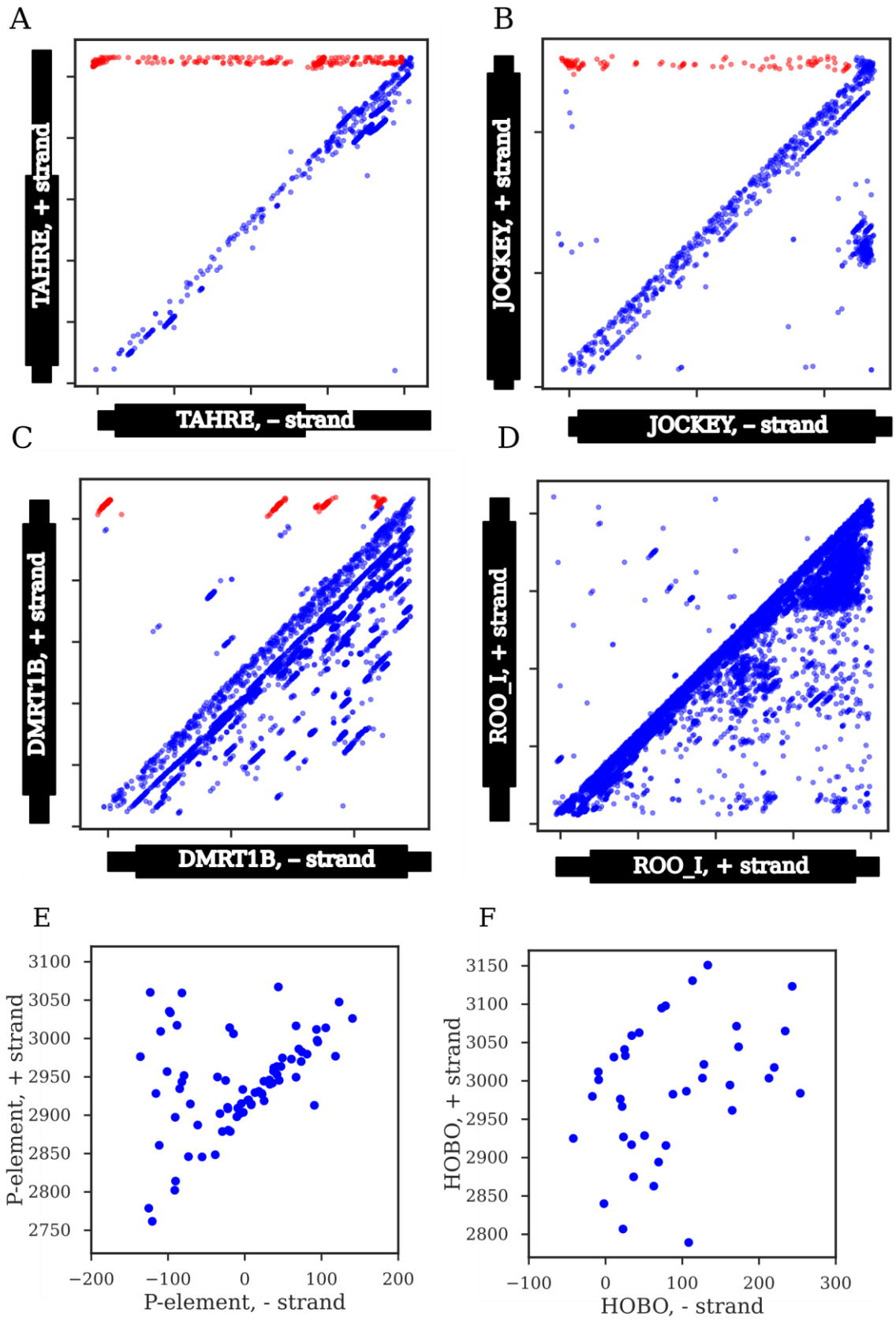

A
TAHRE, + strand
TAHRE, – strand
B
JOCKEY, + strand
JOCKEY, – strand
C
DMRT1B, + strand
DMRT1B, – strand
D
ROO_I, + strand
ROO_I, + strand
E
P-element, + strand
P-element, - strand
3100
3050
3000
2950
2900
2850
2800
2750
−200
−100
0
100
200
F
HOBO, + strand
HOBO, - strand
3150
3100
3050
3000
2950
2900
2850
2800
−100
0
100
200
300

Figure 3: Junction distributions from all samples in the GDL for three non-LTR retrotransposons (A-C), one LTR RT (D), and head-to-tail tandem distributions only for two DNA transposons: P-element (E) and Hobo (F). Each dot represents a junction identified from a single sample. If a junction is present in more than one sample, it will generate a diagonal distribution around the true junction coordinate due to estimation error. Diagonal spread of errors arises from the diagonal read pair distributions from which the junction locations were estimated. Head-to-tail and 5'-truncated tandem junctions are highlighted in red. The distribution of tandem junctions of the telomeric retrotransposon TAHRE (A) and of Jockey (B) are dispersed, with few distinct diagonal clusters, indicating that most individual tandem junctions are low-frequency. By contrast, there are four distinct diagonal clusters of RT1B tandem junctions (C) indicating junctions at moderate to high population frequency and suggesting older tandems. In comparison LTR retrotransposons like Roo (D) do not display this pattern of tail-to-internal junctions. E) The distribution of head-to-tail P-element tandem junctions. Note that most junctions fall within a tight diagonal cluster, consistent with their sharing the same junction coordinate. Several junctions are dispersed above this cluster, consistent with intervening sequence of variable length within the junction. F) The distribution of head-to-tail Hobo tandem junctions. Note that only a few junctions form a tight diagonal cluster, while most are dispersed below the cluster, consistent with small internal deletions spanning the tandem junction in most lines.

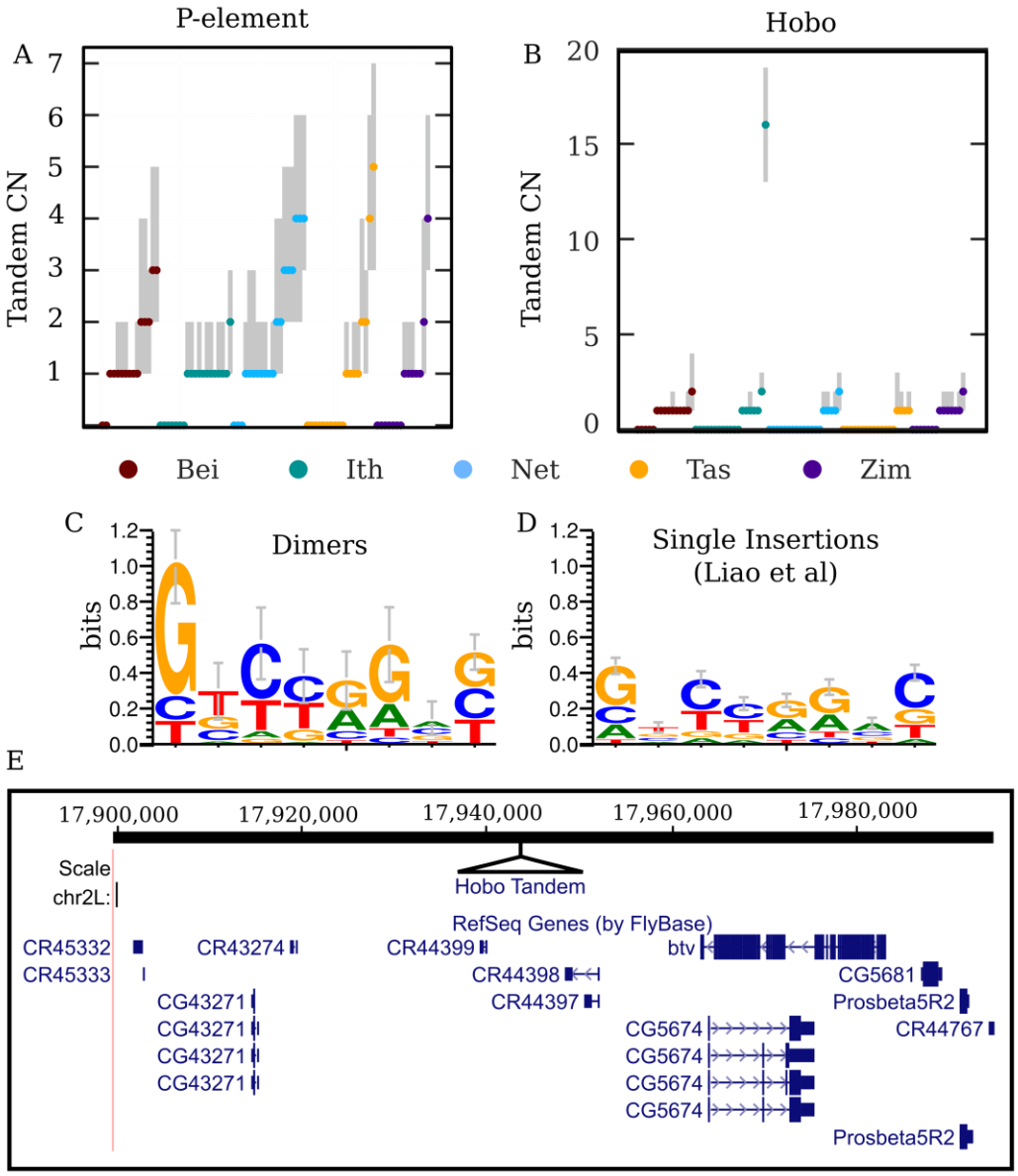


Figure 4: Copy number distributions for P-element (A) and Hobo (B). The dots are maximum a posteriori estimates in a particular strain while the grey lines indicate corresponding 98%-credible intervals; the colors indicate population. C) A Sequence Logo constructed from the 8-nt motifs found within P-element tandem junctions. D) A Sequence Logo constructed from the P-element TSDs reported by Liao et al. E) A UCSC Genome browser view of the region inferred to contain the Hobo tandem array, with the site of the Hobo tandem added in as a black triangle.

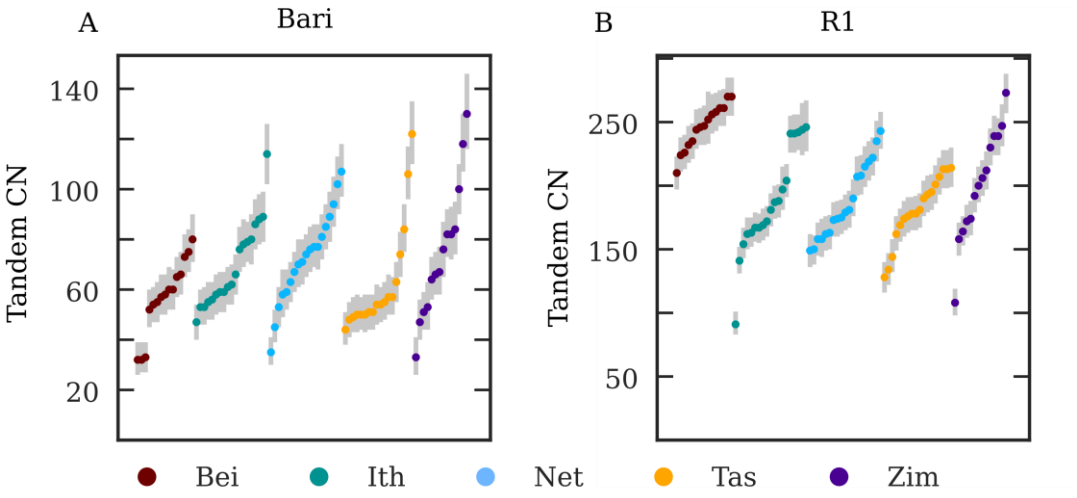


Figure 5: Copy number distributions for the Bari (A) and R1 (B) tandem arrays. The dots are maximum a posteriori estimates in a particular strain while the grey lines indicate corresponding 98%-credible intervals; the colors indicate population.

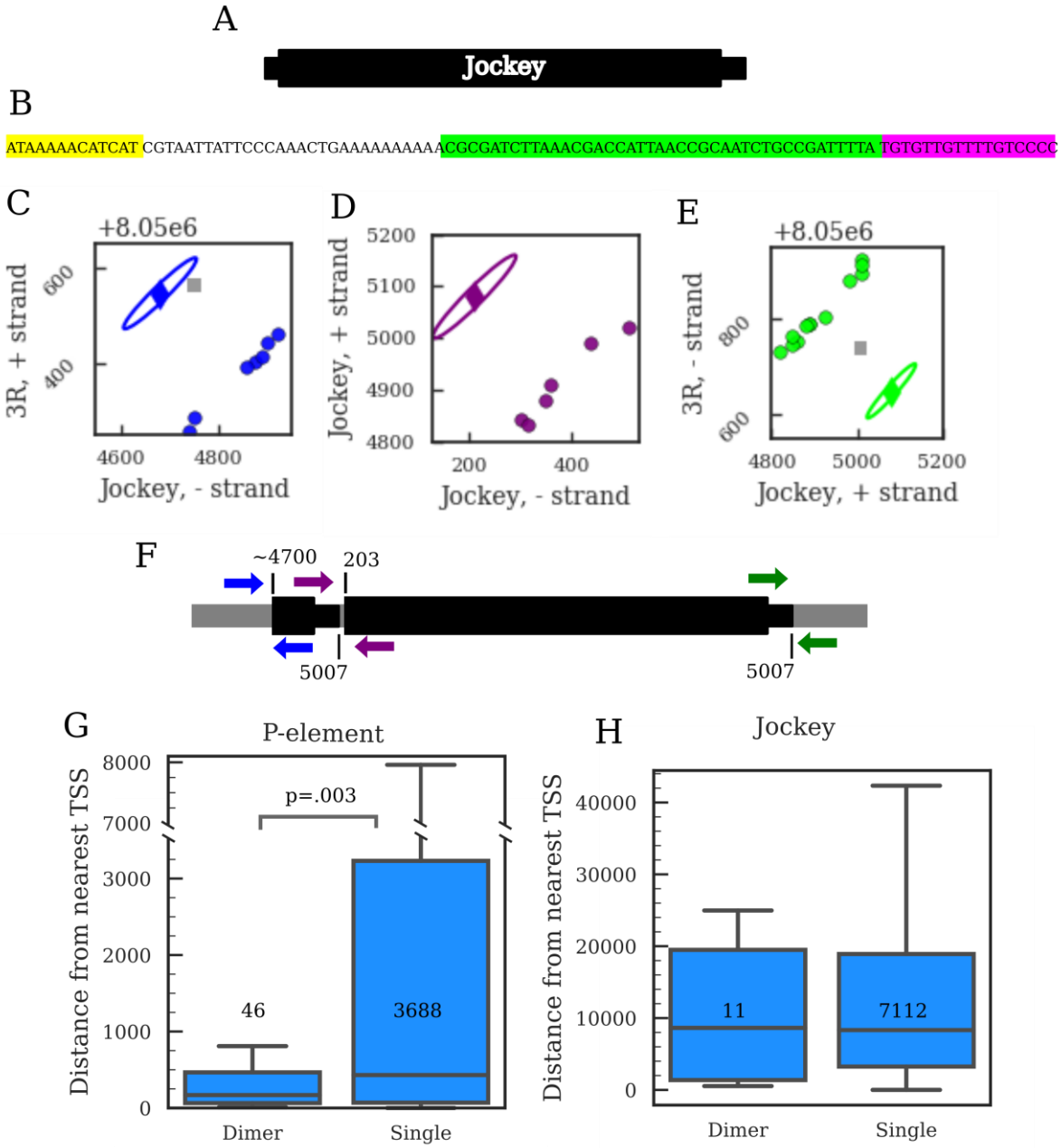
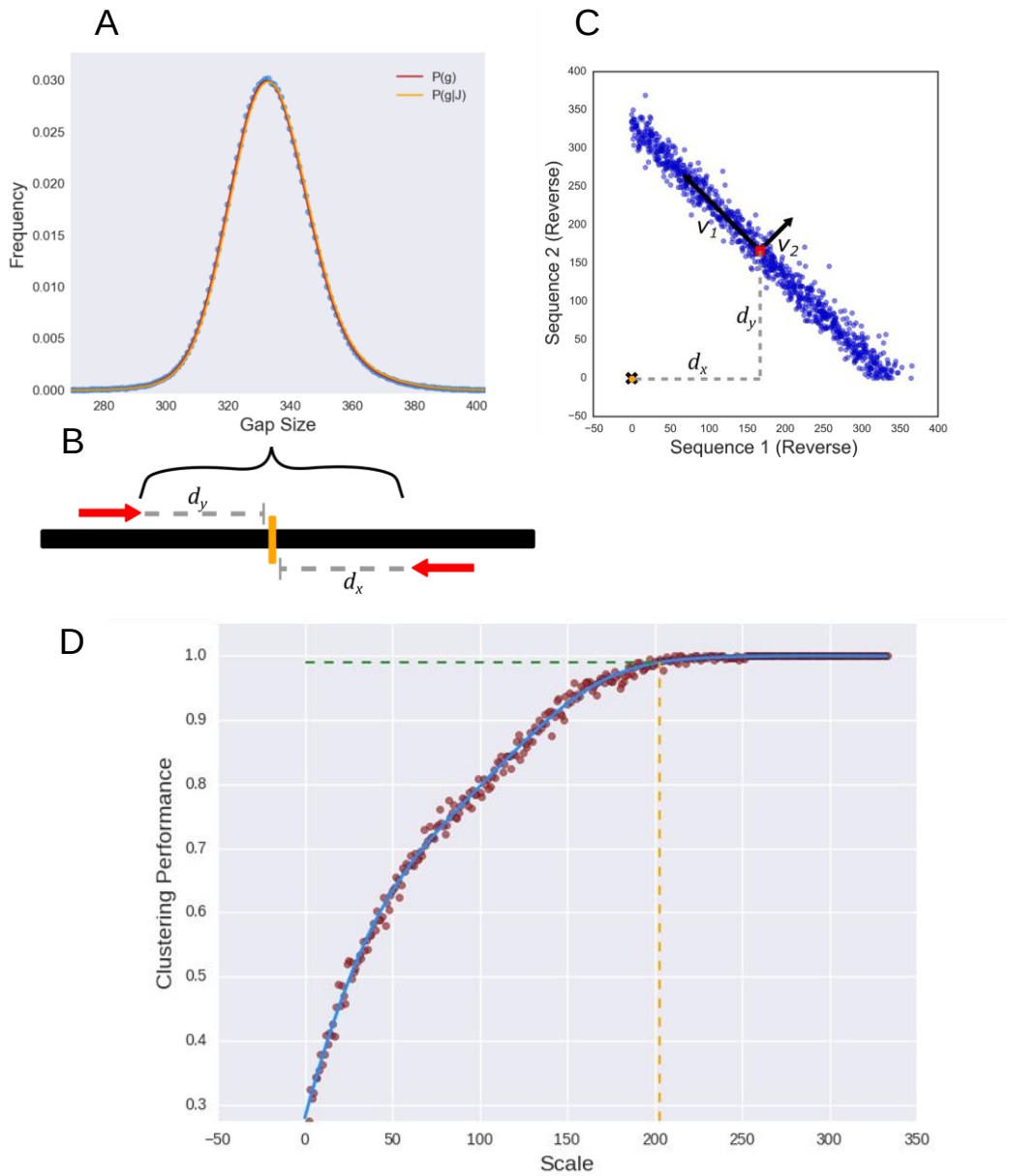
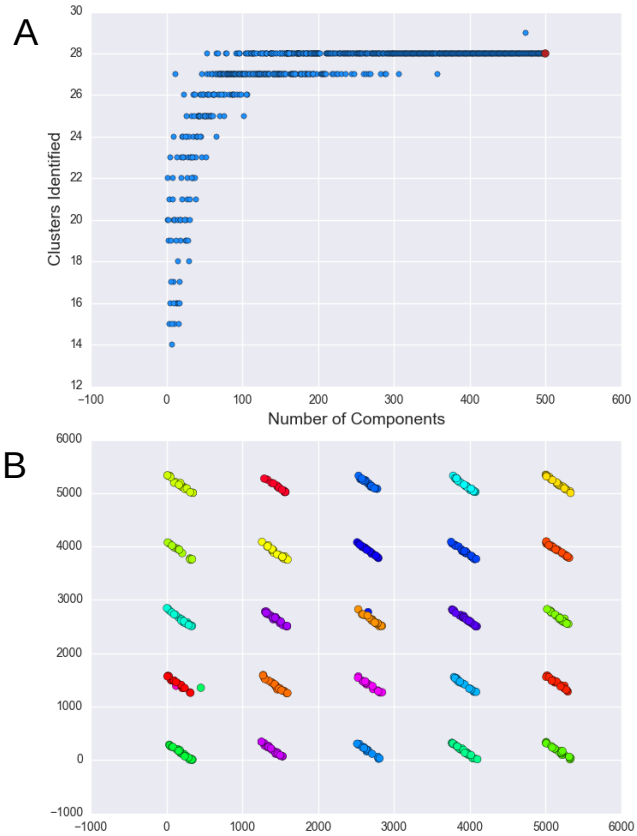


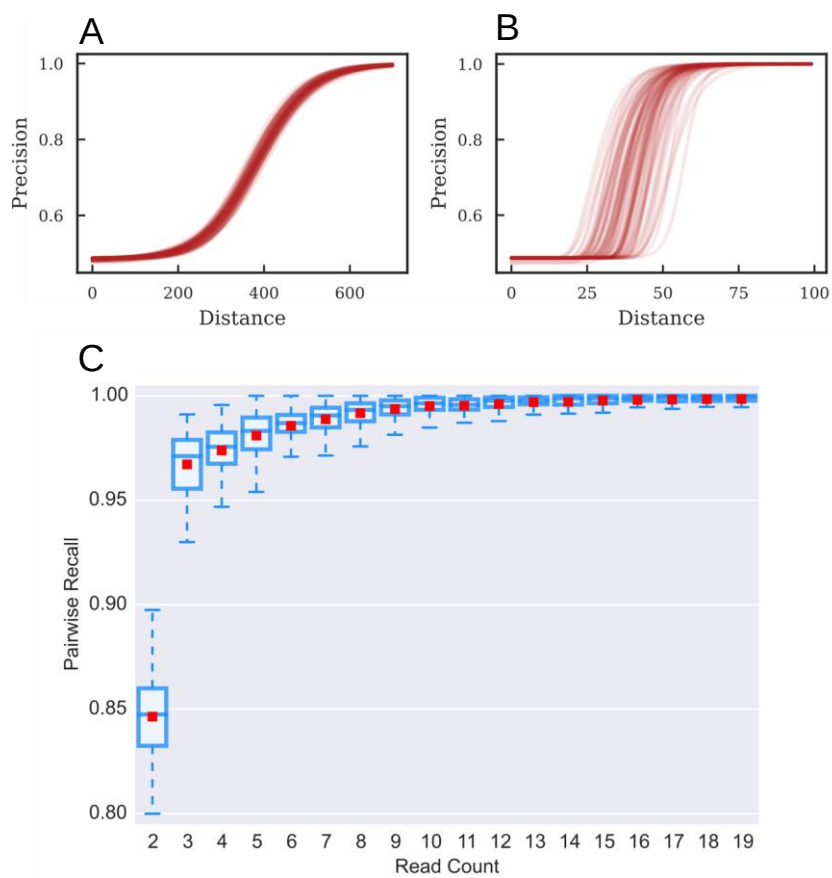
Figure 5: The structure and location of tandem dimers. A) A schematic of a full-length Jockey element; thick box represents CDS. B) An example read containing the 3'-end of one Jockey insertion (yellow) and the truncated 5'-end of a different Jockey insertion (purple) separated by sequence identical to positions 8,050,615 – 8,050,656 of 3R (green). C) Read pairs spanning the junction between the position of ~8,050,600 3R and position ~4,700 of a 5'-truncated Jockey element. D) Read pairs spanning the tandem junction represented in B, between the intact 3'-end of one Jockey element (position 5,007) and a different Jockey element truncated at position 203. E) Read pairs spanning the junction between the intact 3'-end of a Jockey element and position ~8,050,600 of 3R. F) The inferred structure of the tandem dimer based on the identified junctions. The reference sequence and Jockey elements are in 5'->3' orientation. Representative reads are denoted with arrows above the diagram if on the forward strand, and below if on the reverse strand. The three colored read pairs correspond to the three identified junctions: blue (C), purple (D), green (E). G) A boxplots depict the distances to the nearest TSS from P-element dimers and single insertions. The numbers within the plots indicate the number of insertions in each category. The p-value is for a Kolmogorov-Smirnov test. H) A similar plot but for Jockey elements.



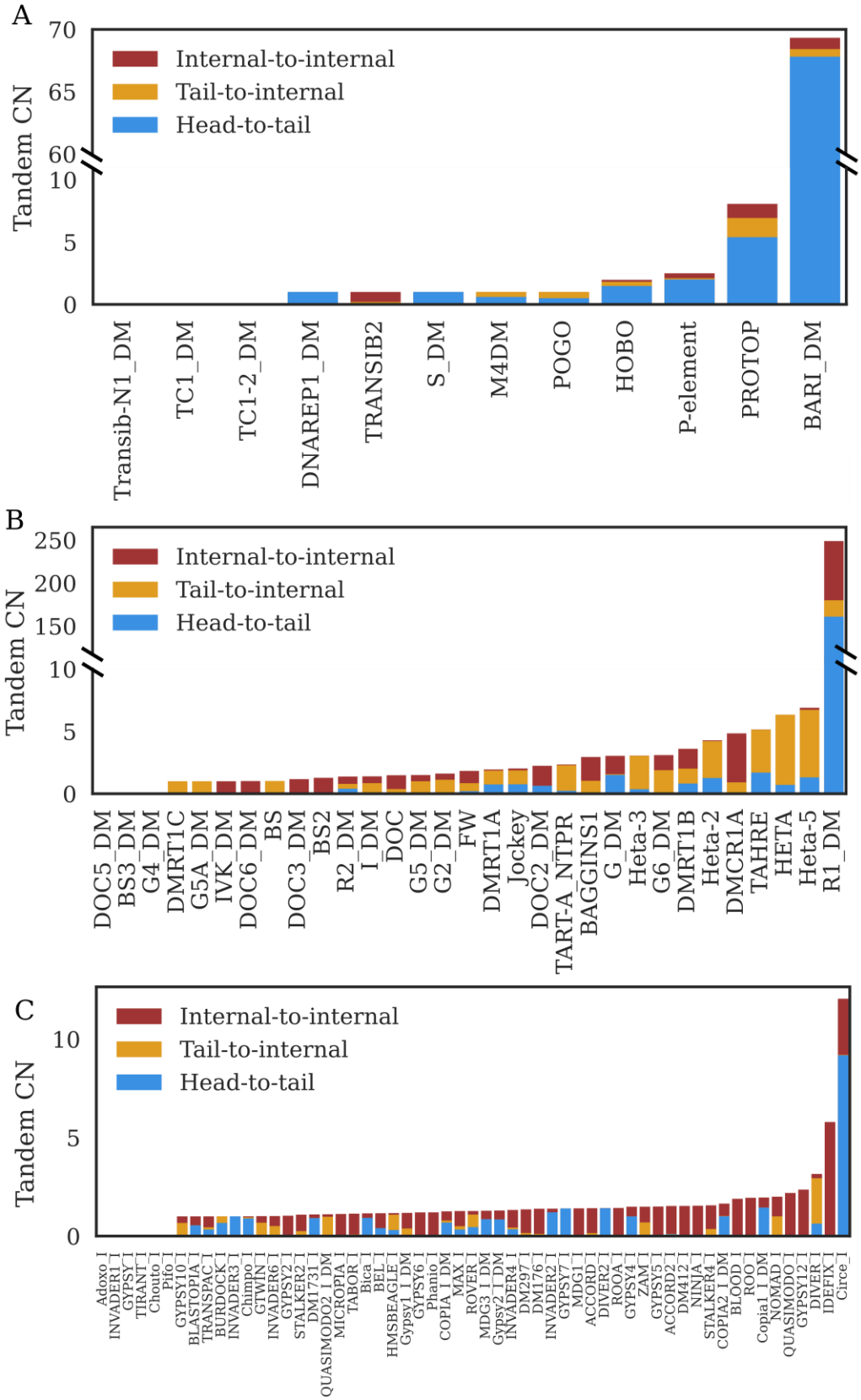
SF 1: Visualizing and approximating read pair distributions. A) A representative gap size distribution from sample N03. Dots represent the observed frequency of specific gap sizes, the red line is the kernel density estimate of the gap size distribution, and the orange line is the gap size distribution conditioned upon the read pairs spanning a junction. B) A schematic depicting a read pair spanning a junction (orange marker). The red arrows depict the forward (above the line) and reverse (below the line) reads. The distance between reads can be partitioned into two components, the distance from the forward read to the junction (d_y) and the distance from the reverse read to the junction (d_x). C) The distribution of 1,000 read pairs spanning the same junction (black x) as a two-dimensional scatterplot, where the Y-axis corresponds to the position of the forward reads and X-axis the position of the reverse reads. The Manhattan distance between read pairs and the junction follows the gap size distribution and points equidistant from the junction fall along 45 degree diagonals yielding diagonal read pair distributions. We estimate the location of the underlying junction as the average read pair coordinate (red square) offset by half the mean gap size along both the X- and Y- axes (orange dot depicts the estimate). There are two perpendicular components of variance in this distribution, one reflecting how the read pair is offset from the junction (V1) and the other reflecting the gap size (V2). D) A depiction of how a scale parameter along an eigenvector (in this case V1) is determined to obtain a suitable Gaussian approximation of the read pair distribution so that the data can be clustered. Each dot is the performance of the EM algorithm given a scale parameter on a bootstrapped sample of read pair distributions (n=150) from the training set. The blue trendline is the fitted Gaussian process regression between scale parameters and clustering performance and represents the expected performance at a given parameter. The green dotted line indicates the performance cutoff and the orange dotted line the parameter at which clustering performance is expected to drop below this cutoff.



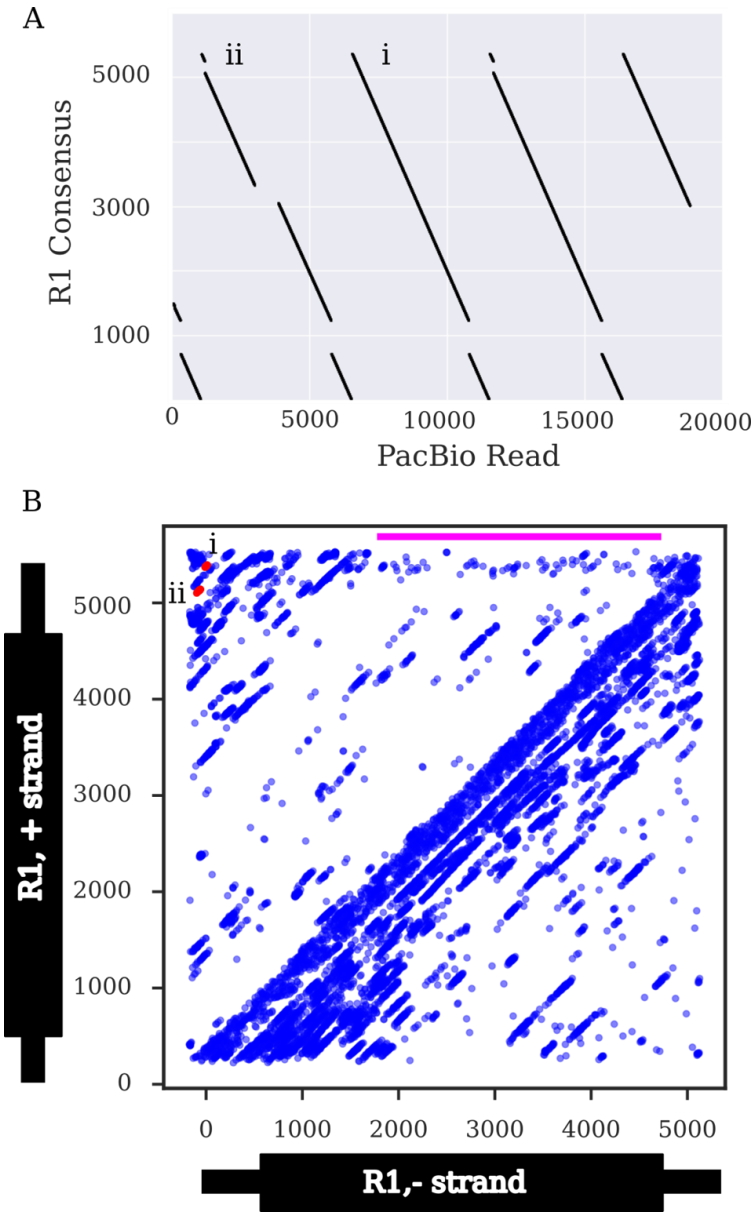
SF 2: A) Number of clusters identified in artificial data by EM as a function of initial components specified when initialized by standard Forgy. We generated 25 artificial distributions from the N14 insert size distribution such that no distributions overlap and clustered using GMMs ranging from 2 to 500 components in size with the appropriate covariance for N14. While this initialization method does not constrain components to be at least distance *d* apart, it does require that each data point be explained by at least one component, and adds additional components if necessary, which is why the smallest cluster count is 14. Even after the number of components exceeds 200, the algorithm tends to recover 28 clusters. Using the distance constrained Forgy initialization scheme identified 28 clusters in 996 of a 1,000 EM runs, and 29 in the remaining four. B) The artificial data clustered with a 500 component GMM (red in A); different clusters are denoted by color.



SF 3: Expected clustering performance for each library in the GDL: A) Each red line depicts the expected precision as a function of the distance between two junctions along the first eigenvector of the model covariance for each sample in the GDL. B) Expected precision as a function of the distance along the second eigenvector. C) The distributions of expected pairwise across the GDL for a given a read count.



SF 4: Average tandem copy number (conditioned on a tandem being present) for A) DNA transposon families, B) non-LTR retrotransposon families, and C) LTR retrotransposon families. Head-to-tail tandems have junctions involving the first and last 200-nt of the consensus sequence. Tail-to-internal junctions have junctions between the last 200-nt of the consensus sequence and internal sequence; are consistent with tandems involving 5'-truncated elements, though they can also be formed by nested insertions. All other tandem junctions are classified as internal-to-internal.



SF 5: A) An alignment between the R1 consensus and a representative PacBio read containing tandem R1 elements. Examples of the two high-copy tandem junctions, one with a small deletion near the tandem junction shifting the read pair distribution, identified by ConTExt are denoted with i and ii. B) The scatterplot of all junctions identified from all lines in the GDL between the forward and reverse strands of R1. In red are the two high copy tandem junctions shown in (A). The purple line indicates a stretch of tandems involving 5'-truncated elements, consistent with more recent double insertions in the rDNA array.