B2.2—*RecombineSim* (*v2.2*)

Aim: Processing of hybrid spore data and simulation of meiotic event distributions (CO, NCO, Total)

Input(s): Event assignment data, Variant table

Output: Event count tables, experimental IED distributions (individual/aggregated), experimental MLE (γ) fits (individual/aggregated), simulated IED distributions

Req(s): MATLAB (2017a)

RecombineSim constitutes an all inclusive data processing and simulation package specifically designed for hybrid tetrad NGS approaches to recombination mapping (see: Section 2.2). *RecombineSim*, designed within MATLAB (2017a), provides a callable function for automated job queuing:

RecombineSim(*EventAssignmentFile*, *VariantTable*, *Output Folder*, *InputGenotype*, *MergeThreshold*, *SimulatedSampleSize* (*M*), *Mode*, C_{PROB}, *customalpha*, *custombeta*)

Example Queue (M = 1000):

RecombineSim('*EventTable.txt*','*Variants.txt*','*1500_Annotated*','*msh2*',*1500*,*1000*,'*Random*',*0*)

RecombineSim('*EventTable.txt*','*Variants.txt*','*1500_Annotated*','*msh2tel1*',*1500*,*1000*,'*Random*',*0*)

RecombineSim('*EventTable.txt*','*Variants.txt*','*1500_Annotated*','*ndt80AR*',*1500*,*1000*,'*Random*',*0*)

RecombineSim('*EventTable.txt*','*Variants.txt*','*1500_Annotated*','*WT*',*1500*,*1000*,'*Hazard*',*32*)

B2.2.1—Data Processing (Event Counts)

Following event assignment and event merging (see: Figure 2.1), data is primarily specified within event assignment tables, which serve as the primary input for *RecombineSim* (*unused columns omitted):

ID	Meiosis	Threshold	Genotype	GID	Chr	CO_NCO	Midpoint
1	msh2_1	1500	msh2	1	16	NCO	63206
2	msh2_1	1500	msh2	1	10	NCO	360711
3	msh2_1	1500	msh2	1	8	NCO	399094
4	msh2_1	1500	msh2	1	13	NCO	51455
5	msh2_1	1500	msh2	1	8	CO	371642

Individual and averaged event counts are subsequently calculated on a per chromosome and per repeat basis for each event type (CO/NCO). Called midpoint values are utilised to calculate experimental IEDs as the distance between successive events of a given type. Event count and IED information for further analysis is provided to the user within tab delimited .txt files, detailing: (i) IED distributions for individual repeats, per event type (ii) aggregated IED distributions for the genotype, per event type (iii) Event counts for individual repeats, per chromosome and per event type (iv) Averaged event counts per chromosome and per event type with standard deviation values.

B2.2.2—Data Processing (MLE γ fitting)

Maximum likelihood estimation (MLE), via MATLAB's *fitdist* toolbox, is utilised to obtain best fit $\gamma(\alpha,\beta)$ parameters from calculated experimental IED distributions, on a per repeat and per genotype (aggregated IED) basis. $\gamma(\alpha,\beta)$ information is provided to the user within tab delimited .txt files, with 95% confidence interval (CI) values—detailing a range within which the real $\gamma(\alpha,\beta)$ values likely reside.

B2.2.3—Simulation (Virtual Chromosomes)

Virtual chromosomes, upon which simulated event formation occurs, are constructed at a 100bp resolution as binned, numerical arrays proportional in size to *in vivo* (chromosome length*0.01) (*S. cerevisiae*—S288c). Chromosomal lengths are further adjusted to reflect the limit of experimental detection governed by the leftmost and rightmost genetic markers (SNPs/INDELs), creating subtelomeric "dead zones". Any given 100bp bin contains values in the range of [0.0-1.0], denoting the inherent recombination potential (recom(P)) of this loci. Prior to initial event formation, all bins are populated with [1.0]—denoting an equal and full recombination potential. Under conditions of independency (random simulation), recom(P) values remain unaltered.

B2.2.4—Simulation (CO Designation, Site Selection & Event Formation)

Class II CO frequency, a user specified parameter, is set as a decimal fraction in the range of [0.0-1.0] (0-100%) via the C_{PROB} parameter. Subclass designation for any given CO event is determined via a randomly generated number (C) in the range [0.0-1.0]. If $C<C_{PROB}$, the event is designated class II and is randomly assigned a location on the chromosome independently of recom(P) values. If $C>C_{PROB}$, a class I event and subsequent CO interference is generated. Recom(P) values are sensed, in order to determine the position of an interference sensitive class I CO, via a weighted, roulette wheel selection algorithm (RWS). RWS constructs a set of arrays where lengths are proportional in length to Recom(P) values held within each chromosomal bin. These arrays are subsequently concatenated and a position along the joined array (F) is randomly chosen (R). Higher recom(P) values translate into a larger proportion of F, thus a higher probability of the corresponding array segment being selected by R. Bins containing recom(P) values of [0.0] (no recombination potential) are excluded i.e. non-selectable. No such designation check is performed during NCO simulations. During random simulations, the system effectively performs unweighted sampling without replacement—that is, the same 100bp bin cannot be chosen twice. Subsequent to the formation of each event, the potential number of IEDs that would be produced by the current array of events is assessed—taking into account simulated merging at a set threshold (e.g. 1.5kb). Event formation continues until the experimentally observed number of IEDs, as calculated by *RecombineSim*, is obtained—simplifying direct comparisons of model-experimental fit. Additional cells (e.g. N = 1000) are independently simulated and resulting simulated IED distributions, provided to the user in tab delimited .txt files, are averaged to reduce stochastic noise (via MATLAB: *downsample*).

B2.2.5—Simulation (Hazard Functions)

Under *Hazard* or *UniHazard* mode, *RecombineSim* imposes CO interference using data derived or user specified $\gamma(\alpha,\beta)$ values respectively to calculate the corresponding hazard function (h(x)) (EQN 1.1). H(x) is essentially calculated as (PDF/1-CDF)—where PDF is the $\gamma(\alpha,\beta)$ probability distribution function and 1-CDF is the inverse $\gamma(\alpha,\beta)$ cumulative distribution function. The numerator and

denominator of h(x) are differentiable functions asymptotically approaching zero with increasing values of (x). Thus, according to L'Hopital's rule (use of derivatives to evaluate limits involving indeterminate forms) the limiting, upper value of the h(x) ratio (y) can be approximated by calculating $1/\gamma(\beta)$, allowing for more rapid normalisation of any given h(x) to a scale of [0.0-1.0], as opposed to the conditional probability values naturally held by a h(x).

$$h(x) = \frac{f(x)}{1 - F(x)} = \frac{f(x)}{S(x)}$$ **(EQN 1.1)**

Where f(x) = PDF, F(x) = CDF, S(x) = Survival Function

To reduce computational time, resulting h(x)'s are trimmed at 500kb equivalent if applicable and converted into a bidirectional interference function through inversion (1-(hx)) and horizontal concatenation of two oppositely oriented functions (see: Figure 2.2E). Upon generation of a class I CO event, this function is superimposed (through multiplication) onto the virtual chromosome array, centred on the initiating event. Recom(P) values in adjacent bins are therefore altered, reducing them in a distance-dependent manner up to ±500kb away. The bin containing the initiating event and those immediately adjacent are modified to possess values of [0.0] and no further recombination is permitted at this loci.