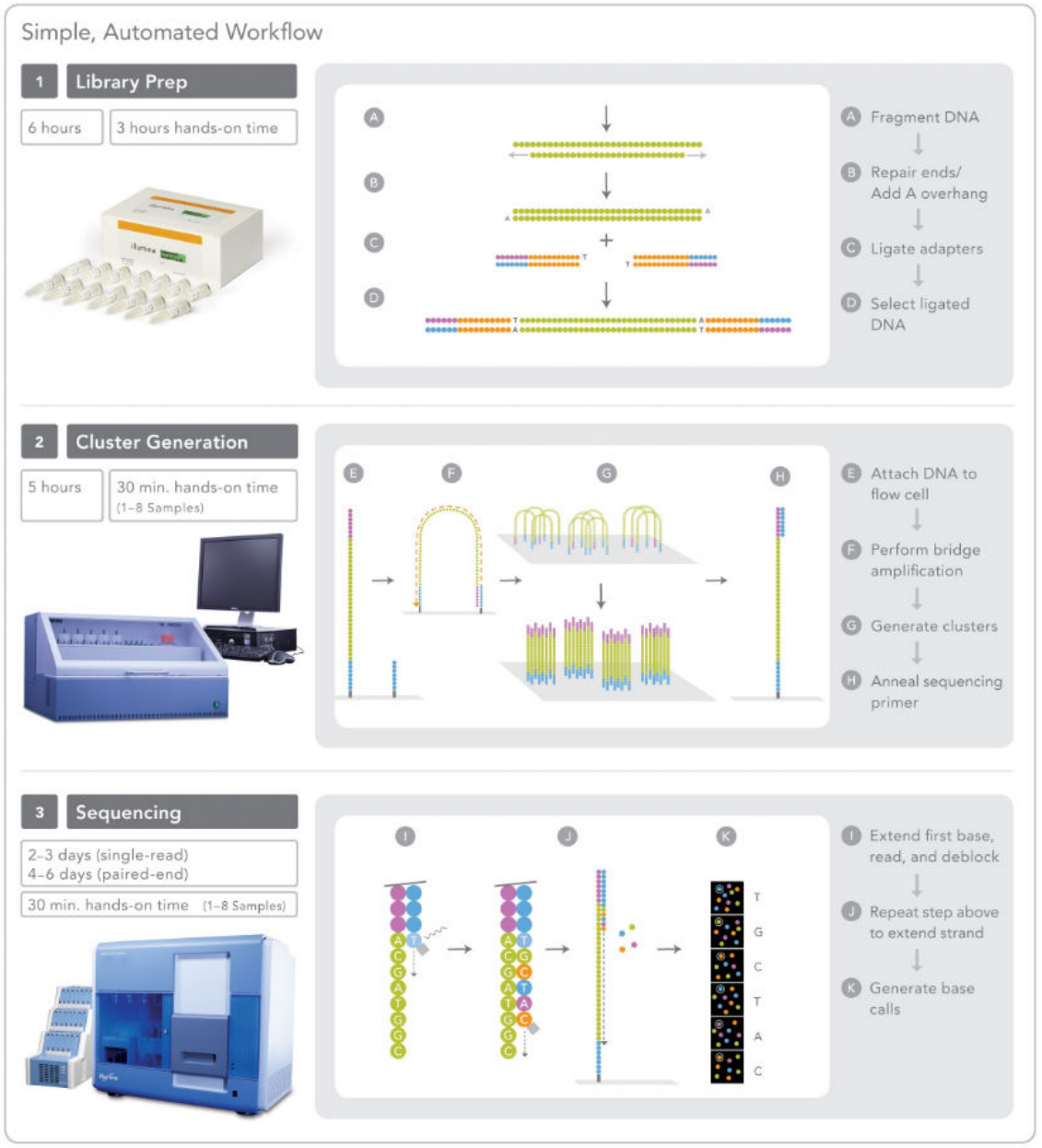
Whole-genome Screening

- Random fragmentation and library construction.
- Attachment and amplification of fragments in flow cell.
- Sequential addition of nucleotides, and detection by chemi-luminescence.
- Build contigs from the shot-gun sequences.
- Can generate billions of bps of sequence in a few days, without biological cloning (single run with two ~150-bp paired ends, 6 billion fragments read).



Some Central Goals of Population Genomics

- Identifying chromosomal regions subject to various forms of selection, and estimating the strength of selection operating on various classes of sites.

- Using neutral sites to estimate the power of the major forces of evolution:

  The key measureable parameters $2N_e u$ and $2N_e r$, respectively, equal the ratios of the power of mutation (u) and the power of recombination (r) to the power of drift ($1/2N_e$); and relative to each other yield an estimate of u/r.

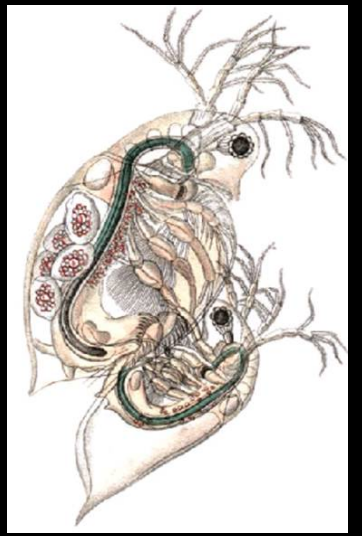
- Estimating aspects of population structure:

  levels of inbreeding; relatedness; population substructure.

- Establishing fine-scale genetic maps, identifying recombination-hotspot motifs, etc.

- Inferring historical changes in population size.

The 5000 *Daphnia pulex* Genomes Project:

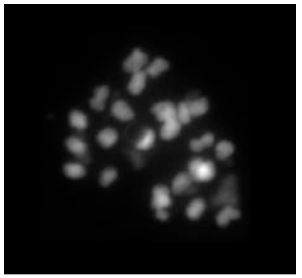


- Obtain the sequences of 96 diploid genomes from ~50 populations over the full geographic range of sexual *D. pulex*.

- Provide an unprecedented resource for the research community:
  - A complete list of functional genes and of their variant nucleotide sites.
  - A complete assessment of molecular variation at all genetic loci across the species.

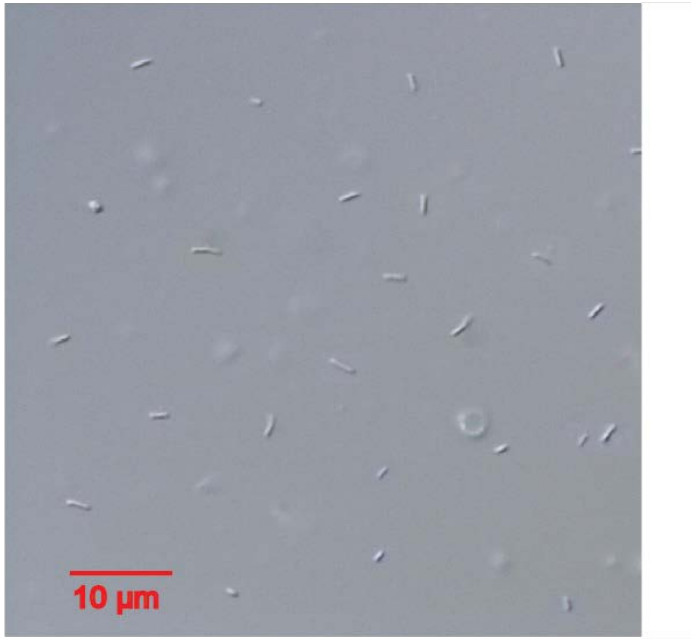
- Will be used to test several hypotheses:
  - Mechanisms by which genome architecture evolves, most notably the origin of introns.
  - The consequences of the loss of recombination.
  - The consequences of long-term population bottlenecks.

General Protocol:

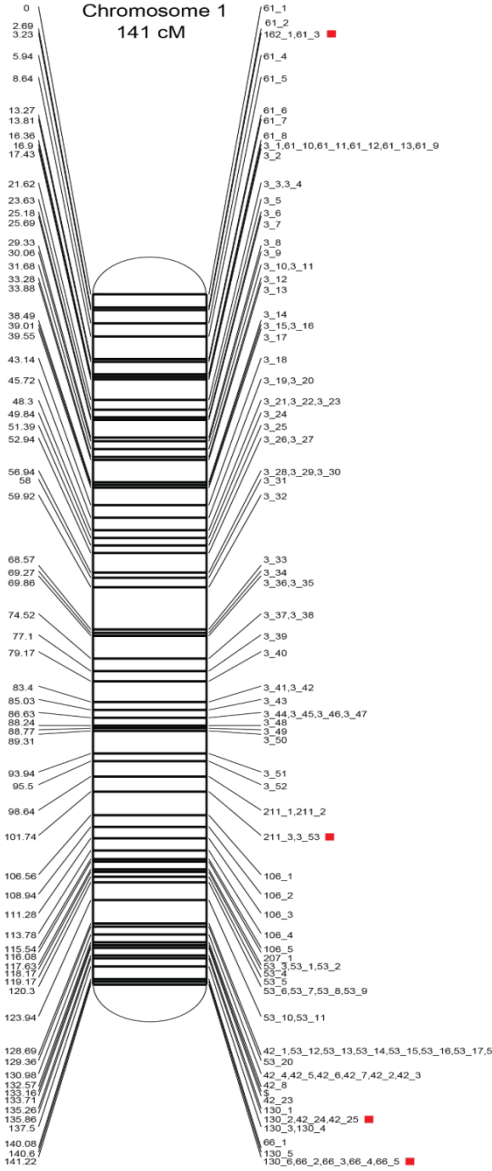


- Collect and isolate animals at hatch out in the spring; verify with diagnostic markers.
- Grow up, isolate and fragment the DNA.
- Bar-code each clone's DNA with a unique 8-mer oligonucleotide.
- Normalize and pool the DNA, and sequence paired ends in a single Illumina run -- yields ~4 billion base pairs of reads / clone (~20x coverage).
- Bin the sequences based on their barcodes and map to the reference strain.
- Use maximum-likelihood methods to estimate allele and genotype frequencies, linkage disequilibrium, individual relatedness, population subdivision, etc.

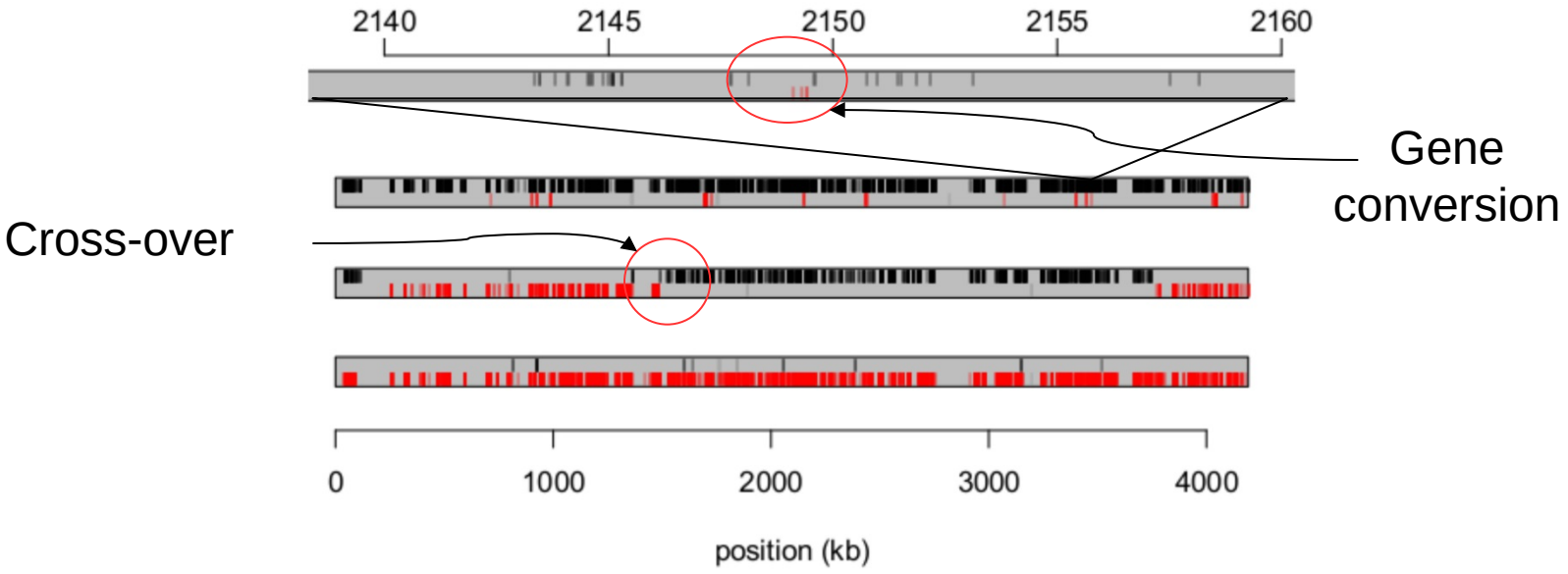
Generation of a High Resolution Genetic Map with Single-sperm Whole-genome Sequencing



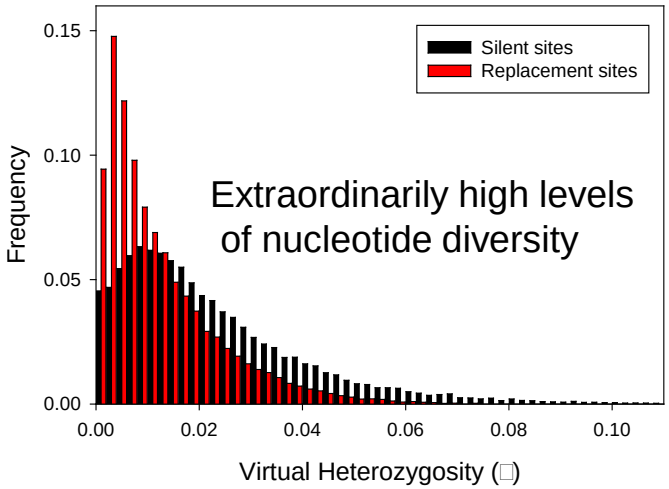
	D. pulex map	Previous map
Total map distance (cM)	1451	1206
No. of markers	1672	185
No. of scaffolds mapped	187	73
Basepairs of genome mapped (Mbp)	131.9	73.9
Average inter-marker distance (cM)	0.87	7

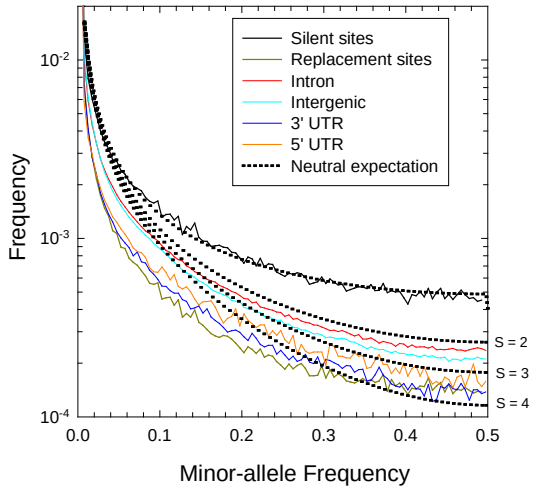


Complete Survey of Recombinational Features by Single-sperm Sequencing



A first glimpse: Kickapoo Pond, Illinois (a natural woodland pond).





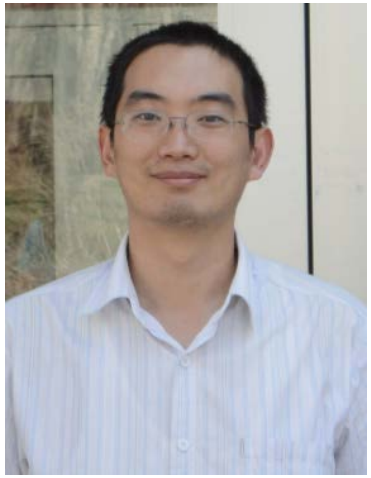
Silent-site diversity behaves neutrally.

Matt Ackerman

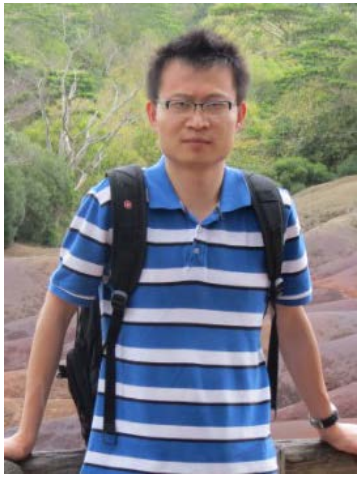
Takahiro Maruki

Ken Spitze

Abe Tucker



Sen Xu



Zhiqiang Ye

Help wanted in developing the community-level resource:

- Population samples of 96 sexual *D. pulex / pulicaria* isolated after resting-egg hatching.
- 96 global isolates of obligately asexual *Daphnia pulex*.
- 96 global isolates of cyclically parthenogenetic *D. pulex / pulicaria* – northern and southern hemispheres; Asia, Europe, South and Central America, South Pacific.
- Gene expression – mRNAs under different conditions, different life stages, different tissues.
- Phenotyping.

Goal and product: the “complete” genome sequences of all of these by 2017.

All data are freely available.

Approaches to Population-Genomic Analysis

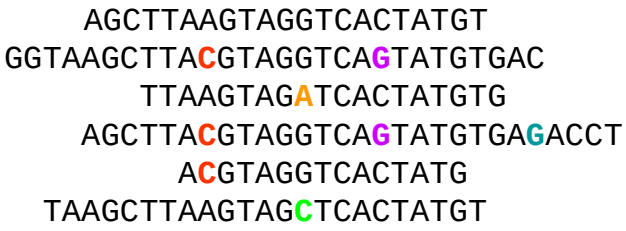
- Pooled population sampling.

- Single-individual analysis.
  - Mean population heterozygosity.
  - Patterns of linkage disequilibrium.

- Allele-frequency estimation from individual sampling.

Central Challenges for Analysis of High-throughput Data From Single Diploid Individuals

1) Error rates as high as 0.01, resulting from a variety of sources generate false heterozygosity and encourage inflated estimates of rare alleles.



- Most methods to estimate error rates are arbitrary, and “off-the shelf” estimates are not reliable – errors in the error rate lead to errors in population parameter estimates.

- Subjective methods for discarding potentially problematical sequences can lead to downward bias in genetic-variation estimates and discard substantial amounts of data.

2) At low to moderate coverages (n), there is a high probability that only one of the two alleles at the site within a diploid individual will have been sequenced, $2(1/2)^n$, creating the false impression of homozygosity; and singleton sampling occurs at rate $2n(1/2)^n$, creating the false impression of errors.

With n = 5, 37.5% of sites have these problems; with n = 10, this becomes 2.1%.

Estimation of Population-genetic Parameters With Single Diploid Individuals

000H00HH0000H0000H0000H0000H00H0H00000000HH0000000000000000000H0000H000H000000000000H000H000000000000000

- Estimates of average site-specific heterozygosity (π) can be achieved genome-wide, within chromosomal regions, or at classes of sites with particular functional significance.

- Estimates of the correlation of "zygosity" across pairs of sites (Δ) separated by various physical distances yields information on the degree of linkage disequilibrium (LD) and insight into recombination rates.

- In randomly mating populations, these single-individual estimates are quite representative of the more conventional population parameter estimates.

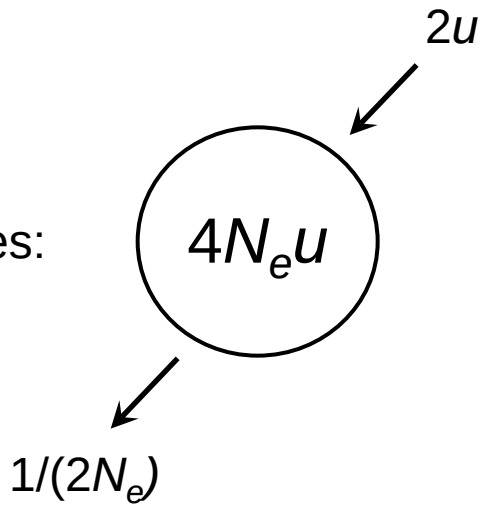
- π provides an estimate of $4N_e u$ for neutral sites (N_e being the effective population size, and u being the per-site mutation rate), assuming drift-mutation equilibrium.

- Δ can be used to acquire an estimate of $4N_e r$, where r is the recombination rate between sites, assuming drift-mutation-recombination equilibrium.

Nucleotide Diversity at Silent Sites in Protein-coding Genes Estimates $4N_eu$

u = base-substitution mutation rate
N_e = genetic effective population size

Expected nucleotide heterozygosity at neutral sites:



ACA
ACC
ACG
ACT

← Silent site

The Four Threonine Codons

A Maximum-Likelihood Approach

- **Data structure (quartet)**: a site that has been sequenced n times within an individual will have a sequence profile (n_A, n_C, n_G, n_T), with the sum of the four elements equaling n.

- Goal is to obtain estimates of the heterozygosity (π) and disequilbrium (Δ) that best explain the data, using the data themselves to estimate and factor out the nuisance error rate (ε).

- Assumes that all read fragments are properly aggregated, either by *de novo* assembly in the case of long reads or via a “reference genome” in the case of short reads, with complicating regions involving paralogs and mobile elements masked out.

- The raw sequence reads may be subject to trimming and quality control prior to analysis.

- The error structure of the data is assumed to be homogeneous, with each nucleotide site having the same probability of misassignment, but more complex scenarios are readily implemented.

Estimation of Average Heterozygosity Within an Individual, π

- The unit of observation at each nucleotide site is the quartet – the set of reads for all four nucleotides.
- For the full range of candidate values of π and ε, the likelihood of the data at each site is obtained by considering the probabilities of the observed data conditional on all possible genotypic states.
- Must condition on whether the site is truly homozygous or heterozygous.

Conditional on the site being homozygous,

the likelihood of the observed data is obtained by summing over the likelihoods conditional on all four possible homozygous types (AA, CC, GG, and TT, with respective relative probabilities p_1, p_2, p_3, and p_4),

$$\ell_1(n_1, n_2, n_3, n_4) = \sum_{i=1}^{4} p_i \cdot \underbrace{p(n - n_i; n, \epsilon)}_{\text{probability of } (n - n_i) \text{ errors}} \cdot \underbrace{\binom{n - n_i}{n_j}\binom{N - n_1 - n_j}{n_k}(1/3)^{n-n_i}}_{\text{trinomial probability of distribution of error types}}$$

Conditional on the site being heterozygous for nucleotides *i* and *j*, must incorporate:

1) the probabilities of the six alternative heterozygous genotypes,
2) the error probability distribution,
3) the binomial sampling distribution of the alternative alleles.

genotype frequencies

probability of ($n - n_i$ - n_j) errors

$$\ell_2(n_1,n_2,n_3,n_4) = \sum_{i=1}^{4}\sum_{j>i}^{4} 2p_ip_j \cdot b(n-n_i-n_j;n,2\epsilon/3)\cdot$$

$$p(n_i;n_i+n_j,0.5)\cdot b(n_k;n-n_i-n_j,0.5)/S$$

allele sampling

binomial probability of distribution of error types *k* and *l*

summed heterozygote frequencies to normalize to a sum of 1.0.

The total likelihood of the observed data at the site, given the assumed values of π and ε:

$$\ell(n_1, n_2, n_3, n_4) = (1 - \pi)\ell_1(n_1, n_2, n_3, n_4) + \pi\ell_2(n_1, n_2, n_3, n_4)$$

The total likelihood of all of the data is the product of the above over all sites, yielding the log likelihood:

$$L = \sum N(n_1, n_2, n_3, n_4) \cdot \ln\left[\ell(n_1, n_2, n_3, n_4)\right]$$

The goal is to obtain the values of π and ε that maximize L.

Two-site population genetics: traditional definition of linkage disequilibrium (LD):

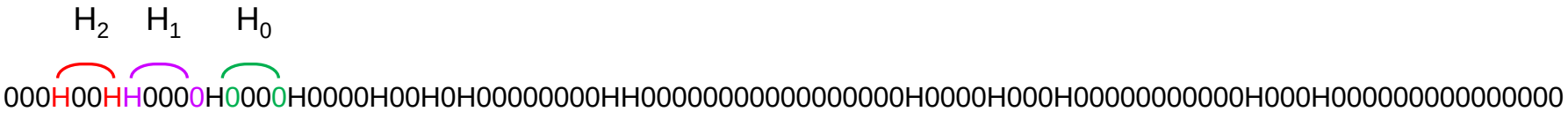
Assuming two alleles segregating at two sites (A and a; and B and b), there are four gamete types: AB, Ab, aB, and ab.

Coefficient of linkage disequilibrium:

D = frequency of the AB gamete – (frequency of allele A)(frequency of allele B)

$= P_{AB} - (p_A \cdot p_B)$

Estimating the Correlation of Heterozygosity for All Pairs of Sites Separated by *d* Nucleotides, Δ_d



Expected Frequencies of Pairs of Sites:

Double homozygotes: $H_0 = (1-\Delta)(1-\pi)^2 + \Delta(1-\pi)$

$= (1-\pi)^2 + \Delta\pi(1-\pi),$

Double heterozygotes: $H_2 = \pi^2 + \Delta\pi(1-\pi),$

Mixed homozygote / heterozygote pairs: $H_1 = 2\pi(1-\pi)(1-\Delta).$

$n_{a1}, n_{a2}, n_{a3}, n_{a4}, n_{b1}, n_{b2}, n_{b3}, n_{b4}$ = octet of observed nucleotide counts at sites *a* and *b*

Estimating the Correlation of Heterozygosity for All Pairs of Sites Separated by *d* Nucleotides, Δ_d.

Likelihood of the data observed at a pair of sites:

$$\ell\left(n_{a1}, n_{a2}, n_{a3}, n_{a4}, n_{b1}, n_{b2}, n_{b3}, n_{b4}\right) = \overbrace{\left[(1-\pi)^2 + \Delta\pi\left(1-\pi\right)\right]}^{\text{joint homozygosity}} \ell_{1a}\ell_{1b} + \overbrace{\left[\pi^2 + \Delta\pi\left(1-\pi\right)\right]}^{\text{joint heterozygosity}} \ell_{2a}\ell_{2b} + \underbrace{\left[\pi\left(1-\pi\right)\left(1-\Delta\right)\right]}_{\text{heterozygosity / homozygosity}} \left(\ell_{1a}\ell_{2b} + \ell_{1b}\ell_{2a}\right),$$

ℓ_{1a}, ℓ_{1b} = likelihoods of read quartets, conditional on homozygosity at loci *a* and *b*, and error rate ε.
ℓ_{2a}, ℓ_{2b} = likelihoods of read quartets, conditional on heterozygosity at loci *a* and *b*, and error rate ε.

Total log likelihood summed over all pairs of sites:

$$L = \sum N\left(n_{a1}, n_{a2}, n_{a3}, n_{a4}, n_{b1}, n_{b2}, n_{b3}, n_{b4}\right) \cdot \ln\left[\ell\left(n_{a1}, n_{a2}, n_{a3}, n_{a4}, n_{b1}, n_{b2}, n_{b3}, n_{b4}\right)\right]$$

Some useful features of the individual-based measure of LD:

1) Δ is a measure of the deficit of the frequency of homozygote-heterozygote pairs from the random expectation.

$$\Delta = 1 - \frac{H_1}{2\pi(1-\pi)}$$

2) For a randomly mating population, the expected value of Δ is equivalent to a scaled measure of the average squared population measure of LD.

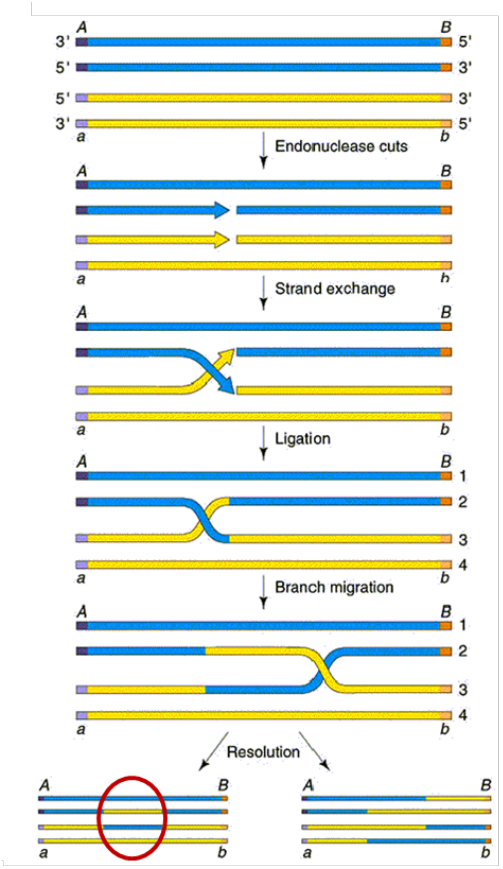
$$E(\Delta) \simeq \frac{4E(D^2)}{\pi(1-\pi)}.$$

3) The expected value of Δ can be expressed in simple terms of the population mutation and recombination rates, $\theta = 4N_e u$, $\rho = 4N_e r$.

$$E(\Delta) \simeq \frac{\theta(18+\rho)}{18+13\rho+\rho^2}$$

How is Δ expected to scale with the distance between sites?

$$E(\Delta) \simeq \frac{\theta(18+\rho)}{18+13\rho+\rho^2}$$ where $\theta = 4N_e u$ and $\rho = 4N_e r$



Gene conversion only

Crossing-over plus gene conversion

Effective recombination rate ≈
cross-over rate + single-site conversion rate

$r = c\,[xd + (1\text{-}x)L(1 - e^{-d/L})]$

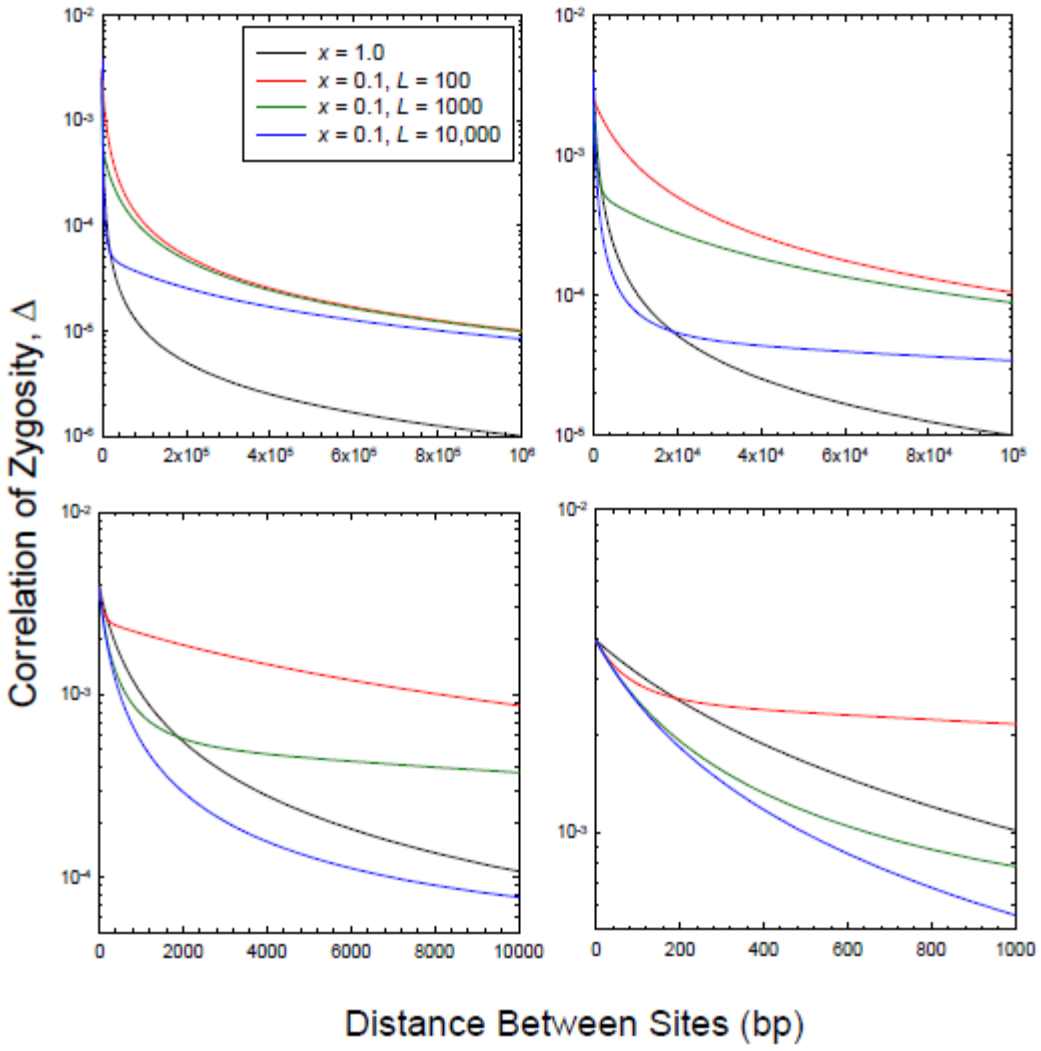
where c = total recombination rate per site,
x = fraction of recombination events resolved as crossovers,
d = distance between sites (bp),
L = average length of conversion tract (bp).

From: Andolfatto and Nordborg (1998); Langley et al. (2000).

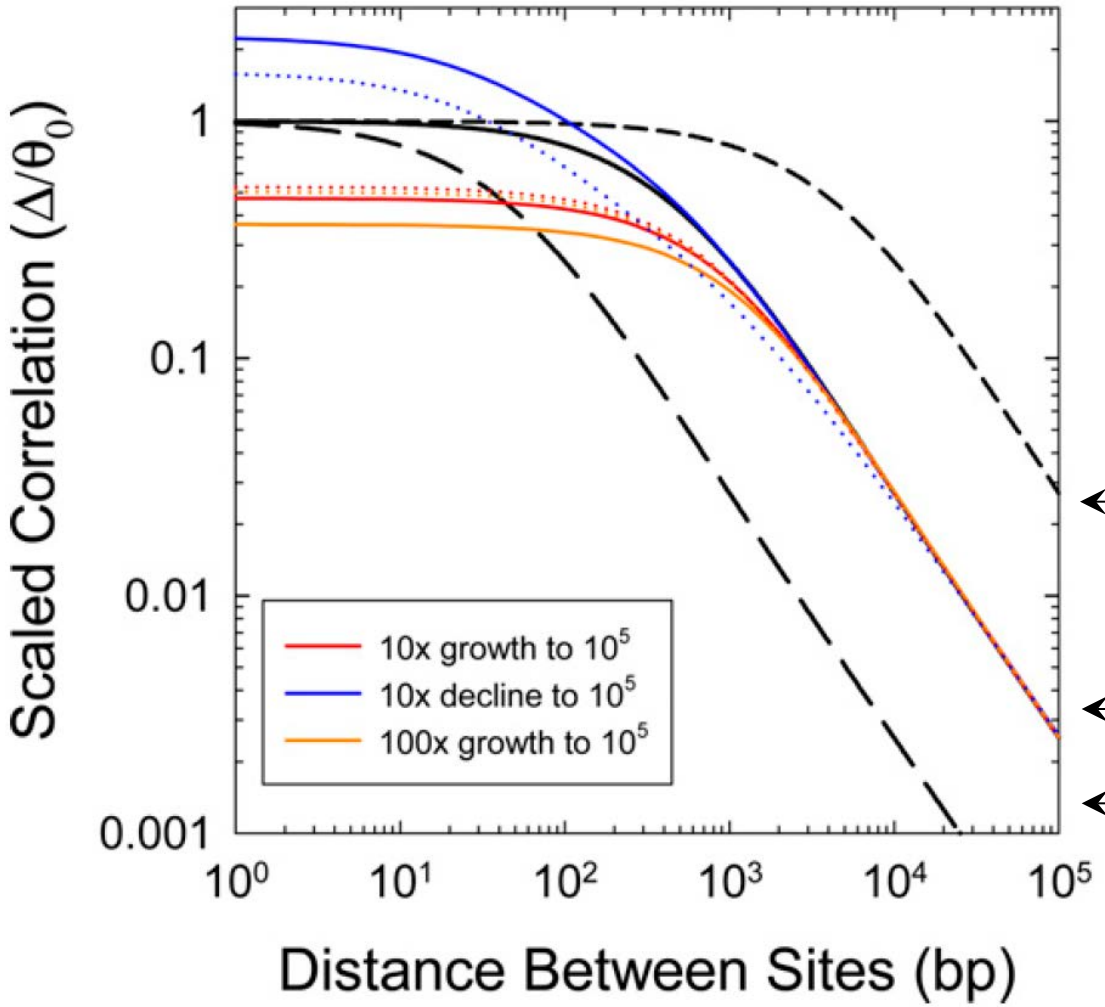
For $d >> L$, $r \approx xcd$

For $d << L$, $r \approx (x + 2)cd$

At Drift-Mutation-Recombination Equilibrium, Δ Declines With Increasing Physical Distance Between Sites, at a Decelerating Rate



Effects of Historical Changes in Population Size on LD Profiles



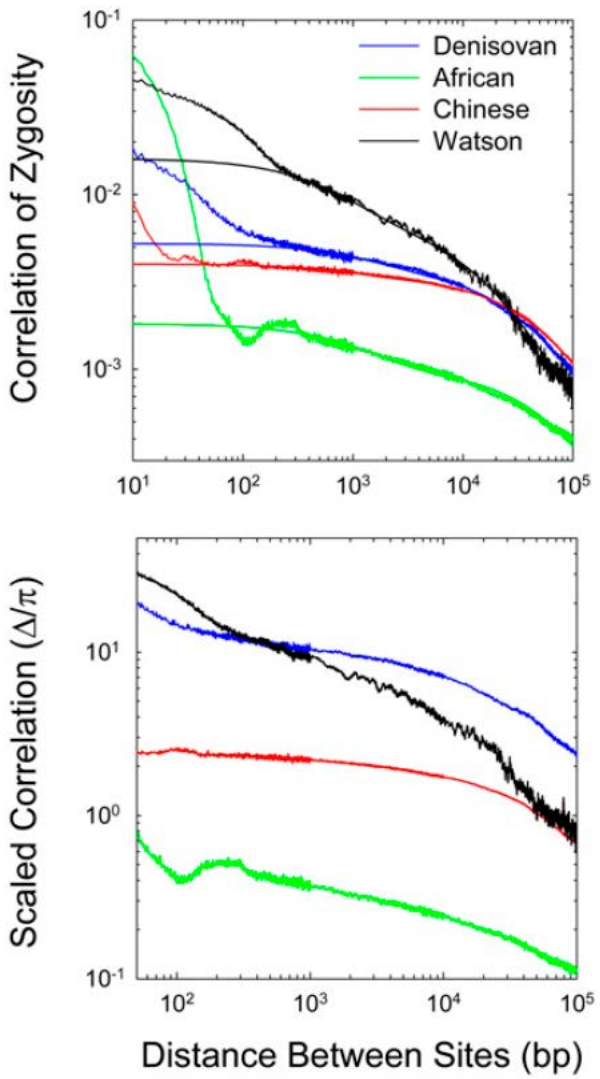
Population sizes:

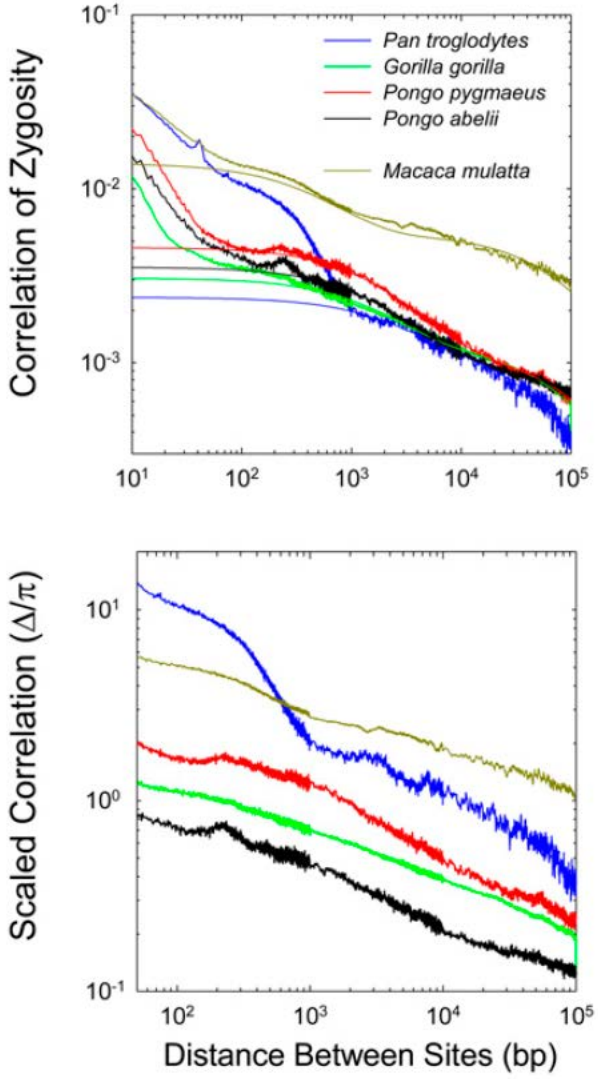
← Small

← Medium, $N_e = 10^5$

← Large

LD-Distance Profiles in Primates





Most Recombination Events Involve Gene Conversion Without Crossing-over

Table 1 Estimates of the features of recombination from the pattern of decline of Δ with physical distance between sites, derived from genomic sequences of single individuals of vertebrates (File S1)

Species	θ	θ'	ρ_1	x	$\bar{L}$
Primates					
Gorilla gorilla	0.0031	0.0031	0.00057	0.067	3792
Macaca mulatta	0.0026	0.0149	0.00218	0.018	1068
Pan troglodytes	0.0010	0.0024	0.00033	0.177	4286
Pongo abelii	0.0053	0.0036	0.00062	0.043	5662
P. pygmaeus	0.0027	0.0047	0.00064	0.069	6852
Homo sapiens (Archaic Denisovan)	0.0004	0.0051	0.00022	0.235	3208
H. sapiens (African)	0.0036	0.0018	0.00064	0.066	1970
H. sapiens (Chinese)	0.0016	0.0039	0.00015	0.216	1833
H. sapiens (Watson)	0.0010	0.0161	0.00146	0.168	1389
Nonprimate mammals:					
Ailuropoda melanoleuca	0.0013	0.0018	0.00023	0.250	16267
Canis familiaris	0.0009	0.0082	0.00503	0.020	1183
Loxodonta africana	0.0013	0.0221	0.00286	0.021	1084
Ornithorhynchus anatinus	0.0013	0.1709	0.28970	0.025	608
Nonmammalian vertebrates:					
Anolis carolinensis	0.0026	0.0094	0.00214	0.058	1927
Fugu rubripes	0.0032	0.0042	0.00128	0.025	7238
Petromyzon marinus	0.0044	0.0061	0.00118	0.052	2863

θ', an inferred estimate of $4N_e u$ from the fit to Equation 8, not necessarily equivalent to the independently derived ML estimate given as θ); $\rho_1 = 4N_e c$, where c is the recombination rate between adjacent sites; x, the fraction of recombination events leading to crossovers; and $\bar{L}$, the mean conversion-tract length (in base pairs).

Acknowledgments:

Bernhard Haubold, Group Leader, Max Planck Institute for Evolutionary Biology

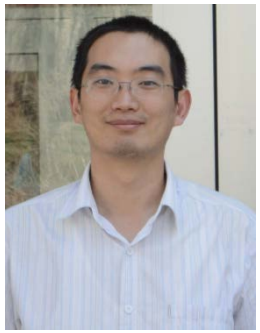
MOLECULAR ECOLOGY

Molecular Ecology (2010), **19** (Suppl. 1), 277–284 doi: 10.1111/j.1365-294X.2009.04482.x

mlRho – a program for estimating the population mutation and recombination rates from shotgun-sequenced diploid genomes

BERNHARD HAUBOLD,* PETER PFAFFELHUBER† and MICHAEL LYNCH‡

http://guanine.evolbio.mpg.de/mlRho/



Sen Xu

Takahiro Maruki

Peter Pfaffelhuber