Estimating Recombination Rates

LRH selection test, and recombination

- Recall that LRH tests for selection by looking at frequencies of specific haplotypes.
- Clearly the test is dependent on the recombination rate.
- Higher recombination rate destroys homozygosity
- It turns out that recombination rates do vary a lot in the genome, and there are many regions with little or no recombination

Daly et al., 2001

- Daly and others were looking at a 500kb region in 5q31 (Crohn disease region)
- 103 SNPs were genotyped in 129 trios.
- The direct approach is to do a case-control analysis using individual SNPs.
- Instead, they decided to focus on haplotypes to corect for local correlation.
- The study finds that large blocks (upto 100kb) show no evidence of recombination, and contain only 2-4 haplotypes
- There is some recombination across blocks

Daly et al, 2001

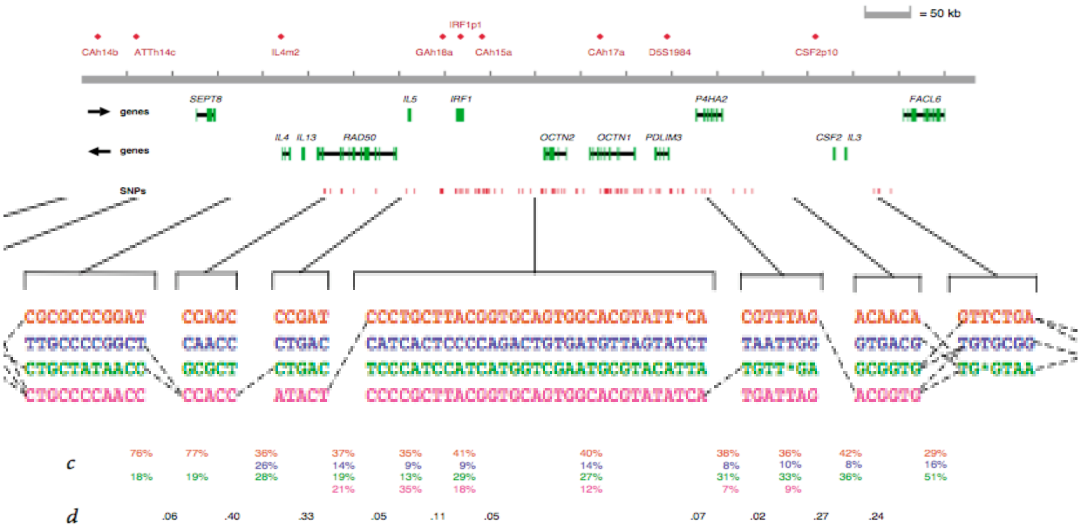
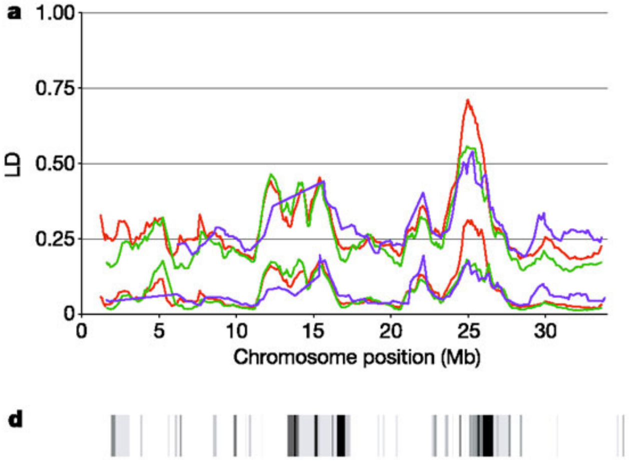


Fig. 2 Block-like haplotype diversity at 5q31. ***a***, Common haplotype patterns in each block of low diversity. Dashed lines indicate locations where more than 2% of all chromosomes are observed to transition from one common haplotype to a different one. ***b***, Percentage of observed chromosomes that match one of the common patterns exactly. ***c***, Percentage of each of the common patterns among untransmitted chromosomes. ***d***, Rate of haplotype exchange between the blocks as estimated by the HMM. We excluded several markers at each end of the map as they provided evidence that the blocks did not continue but were not adequate to build a first or last block. In addition, four markers fell between blocks, which suggests that the recombinational clustering may not take place at a specific base-pair position, but rather in small regions.

230 **nature genetics • volume 29 • october 2001**

Recombination in human chromosome 22 (Mb scale)

Dawson et al. Nature 2002



Q: Can we give a direct count of the number of the recombination events?

Recombination hot-spots (fine scale)

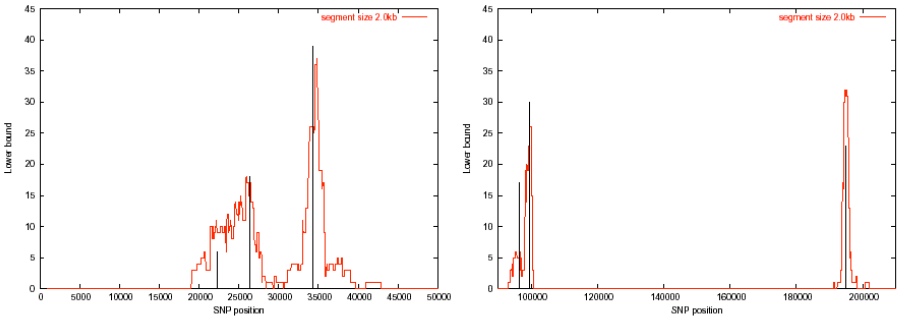
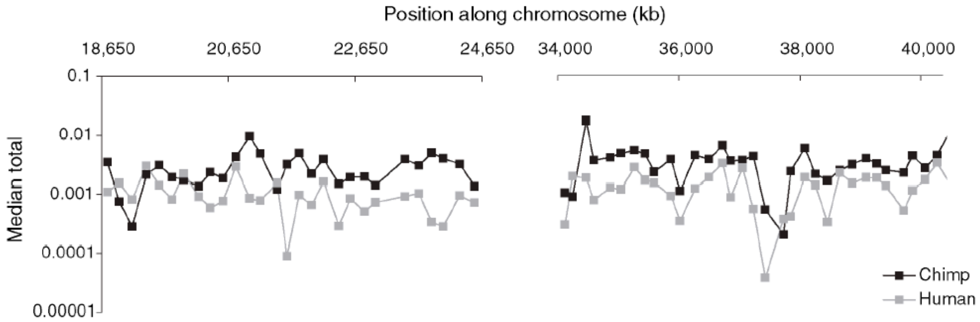


Figure 7: Plot of recombination lower bounds (on a 2-kb scale) for the 216-kb segment of the class II region of the major histocompatibility complex (MHC). The vertical black lines (height scaled by logarithm of the mean recombination rate obtained from sperm typing for that hotspot) show the approximate locations of the center of the six hotspots inferred using sperm crossover analysis by Jeffreys et. al. [22]. The TAP2 hotspot [23] is the last hotspot near the 200-kb region.

Recombination rates (chimp/human)



- Fine scale recombination rates differ between chimp and human
- The six hot-spots seen in human are not seen in chimp

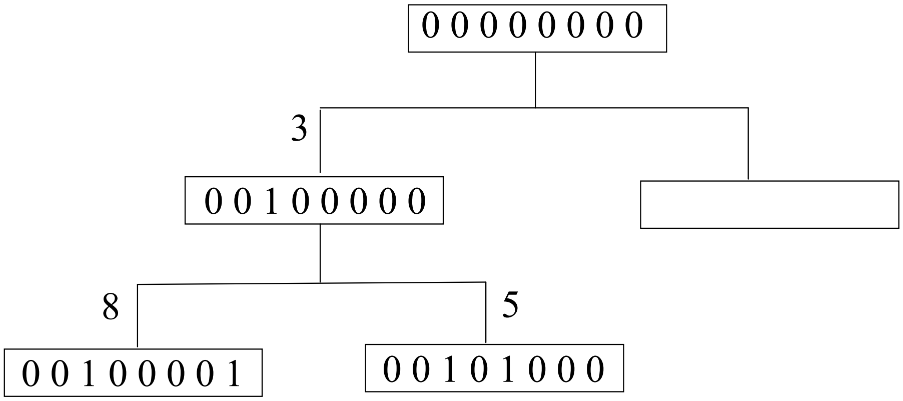
Estimating recombination rate

- Given population data, can you predict the scaled recombination rate ρ in a small region?
- Can you predict fine scale variation in recombination rates (across 2-3kb)?

Combinatorial Bounds for estimating recombination rate

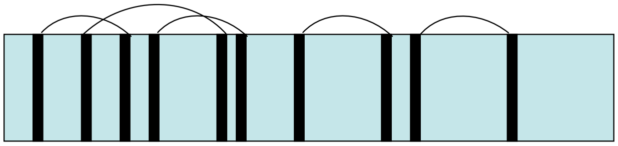
- Recall that expected #recombinations = ρ log n
- Procedure
  - Generate N random ARGs that results in the given sample
  - Compute mean of the number of recombinations
- Alternatively, generate a summary statistic s from the population.
- For each ρ, generate many populations, and compute the mean and variance of s (This only needs to be done once).
- Use this to select the most likely ρ
- What is the correct summary statistic?
- Today, we talk about the min. number of recombination events as a possible summary statistic. It is not the most natural, but it is the most interesting computationally.

The Infinite Sites Assumption & the 4 gamete condition



- Consider a history without recombination. No pair of sites shows all four gametes 00,01,10,11.
- A pair of sites with all 4 gametes implies a recombination event

Hudson & Kaplan

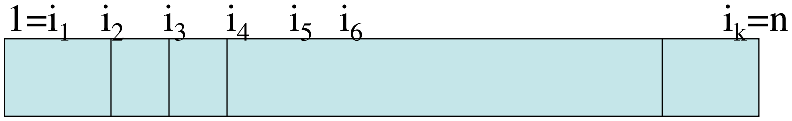


- Any pair of sites (i,j) containing 4 gametes must admit a recombination event.
- Disjoint (non-overlapping) sites must contain distinct recombination events, which can be summed! This gives a lower bound on the number of recombination events.
- Based on simulations, this bound is not tight.

Myers and Griffiths'03: Idea 1

- Let B(i,j) be a lower bound on the number of recombinations between sites i and j.

Define Partition $P = 1 = i_1 < i_2 < \ldots < i_k = n$



$$R(P) = \sum_{j=1}^{k-1} B(i_j, i_{j+1}) \text{ is a lower bound for all P!}$$

- Can we compute $\max_P R(P)$ efficiently?

The R_m bound

Let $R_m(j) = \max_{P_j} R(P_j)$, for all partitions of the first j columns

Computing $R_m(j)$ for all j is sufficient (why?)

$$\text{for } j = 2 \ldots n$$
$$R_m(j) = \max_{1 \le k < j} R_m(k) + B(k, j)$$

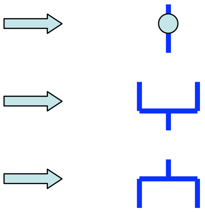
Improved lower bounds

- The R_m bound also gives a general technique for combining local lower bounds into an overall lower bound.
- In the example, R_m=2, but we cannot give any ARG with 2 recombination events.
- Can we improve upon Hudson and Kaplan to get better local lower bounds?

0 0 0
0 0 1
0 1 0
0 1 1
1 0 0
1 0 1
1 1 0
1 1 1

Myers & Griffiths: Idea 2

- Consider the history of individuals. Let H_t denote the number of distinct haplotypes at time t
- One of three things might happen at time t:
  - Mutation: H_t increase by at most 1
  - Recombination: H_t increase by at most 1
  - Coalescence: H_t does not increase



The R_H bound

$H \equiv$ Number of extant & distinct halotypes
$E \equiv$ Number of mutation events
$R \equiv$ Number of Recombination events

$$H \le R + E + 1$$

$$\Rightarrow R \ge H - E - 1$$

$$\text{Infinite sites} \Rightarrow \text{E} \le \text{S}$$

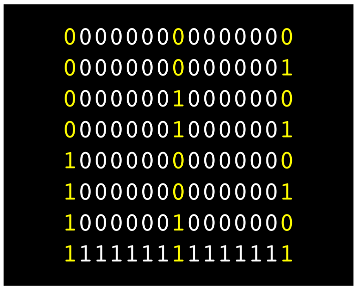
$$R \ge H - S - 1$$

Ex: R>= 8-3-1=4

```
0 0 0
0 0 1
0 1 0
0 1 1
1 0 0
1 0 1
1 1 0
1 1 1
```

R_H bound

- In general, R_H can be quite weak:
  - consider the case when S>H
- However, it can be improved
  - Partitioning idea: sum R_H over disjoint intervals
  - Apply to any subset of columns. Ex: Apply R_H to the yellow columns



```
000000000000000
000000000000001
000000010000000
000000010000001
100000000000000
100000000000001
100000010000000
111111111111111
```

Caveat : Computing $\max_{H' \subseteq H} R(H')$ is NP - complete!

(BB'05)

Computing the R_H bound

- Goal: Compute
  - Max H' R(H')
- It is equivalent to the following:
- Find the smallest subset of columns such that every pair of rows is 'distinguished' by at least one column
- For example, if we choose columns 1, 8, rows 1,2, and rows 5,6 remain identical.
- If choose columns 1,8,15 all rows are distinct.

```
  123456789012345
1:000000000000000
2:000000000000001
3:000000010000000
4:000000010000001
5:100000000000000
6:100000000000001
7:100000010000000
8:111111111111111
```

(BB'05)

Computing R_H

- A greedy heuristic:
  - Remove all redundant rows.
  - Set of columns, C=Ø
  - Set S = {all pairs of rows}
  - Iterate while (S<>Ø):
    - Select a column c that separates maximum number of pairs P in S.
    - C=C+{c}
    - S=S-P
  - Return n-1-|C|

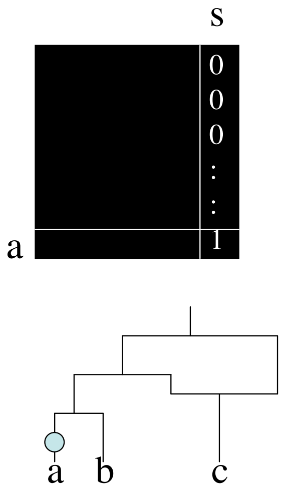
Computing R_H

- How tight is R_H?
- Clearly, by removing a haplotype, R_H decreases.
- However, the number of recombinations needed doesn't really change

0 0 0
0 0 1
0 1 0
▭
1 0 0
1 0 1
▭
1 1 1

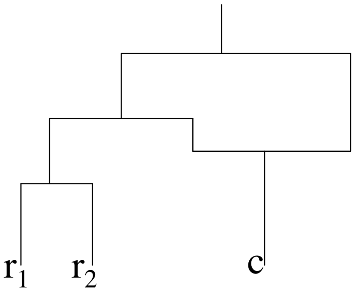
R_s bound: Observation I

- *Non-informative column:* If a site contains at most one 1, or one 0, then in any history, it can be obtained by adding a mutation to a branch.
  - EX: if a is the haplotype containing a 1, It can simply be added to the branch without increasing number of recombination events
  - R(M) = R(M-{s})



R_s bound: Observation 2

- *Redundant rows*: If two rows h_1 and h_2 are identical, then
  - R(M) = R(M-{h_1})



Rs bound: Observation 3

- Suppose M has no non-informative columns, or redundant rows.
  - Then, at least one of the haplotypes is a recombinant.
  - There exists h s.t.

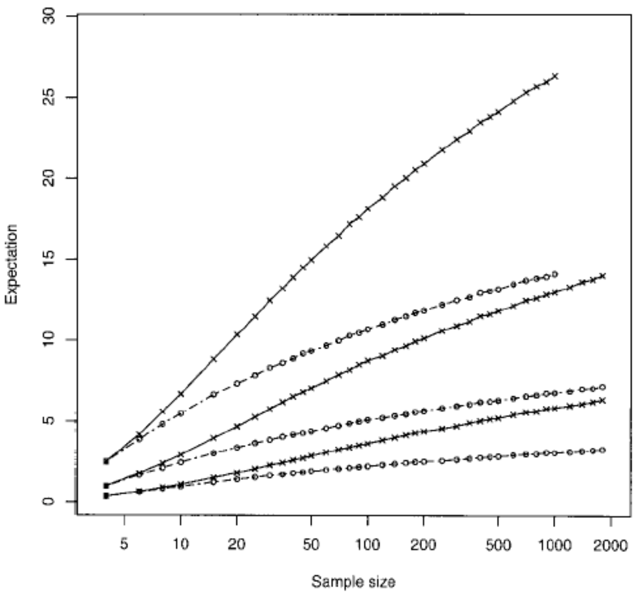
    $R(M) = R(M-\{h\})+1$
  - Which h should you choose?

R_s bound (Procedural)

```
Procedure $Compute\_R_s(M)$
    If ∃ non-informative column s
        return ($Compute\_R_s(M-\{s\})$)
    Else if ∃  redundant row h
        return ($Compute\_R_s(M-\{h\})$)
    Else
        return (1 + $\min_h$($Compute\_R_s(M-\{h\})$))
```

Results

FIGURE 3.—The effect of sample size on the expected minima for three values of $\theta = \rho$. Each point is estimated from 10,000 simulated samples. The dotted lines with circles give values for $\mathbb{E}(R_m)$, and the solid lines with crosses give values for $\mathbb{E}(R_h)$. The three pairs then correspond to $\theta = \rho = 5$ (bottom pair), 10 (middle pair), and 20 (top pair).



Additional results/problems

- Using dynamic programming, R_s can be computed in 2^n poly(mn) time.
- Also, Rs can be augmented to handle intermediates.
- Are there poly. time lower bounds?
  - The number of connected components in the conflict graph is a lower bound (BB'04).
- Fast algorithms for computing ARGs with minimum recombination.
  - Poly. Time to get ARG with 0 recombination
  - Poly. Time to get ARGs that are galled trees (Gusfield'03)

Underperforming lower bounds

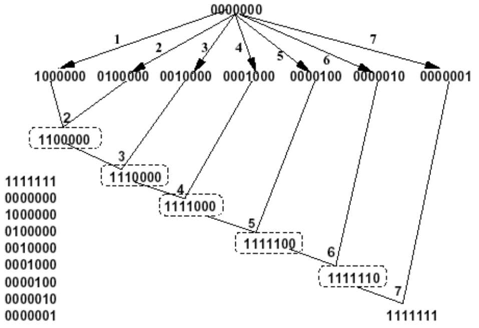
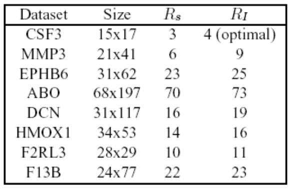


Figure 3: A set of 9 haplotypes for which R_s is 1 and a phylogenetic network for the set of haplotypes with 6 recombination events ($R_I = 6$)



Dataset	Size	R_s	R_I
CSF3	15x17	3	4 (optimal)
MMP3	21x41	6	9
EPHB6	31x62	23	25
ABO	68x197	70	73
DCN	31x117	16	19
HMOX1	34x53	14	16
F2RL3	28x29	10	11
F13B	24x77	22	23

Table 2: Comparison of the number of detected recombination events using R_s and R_I for the phased haplotype datasets for various genes obtained from the SeattleSNP project [31]

- Sometimes, R_s can be quite weak
- An R_I lower bound that uses intermediates can help (BB'05)

LPL data set

- 71 individuals, 9.7Kbp genomic sequence
  - – R_m=22, R_h=70

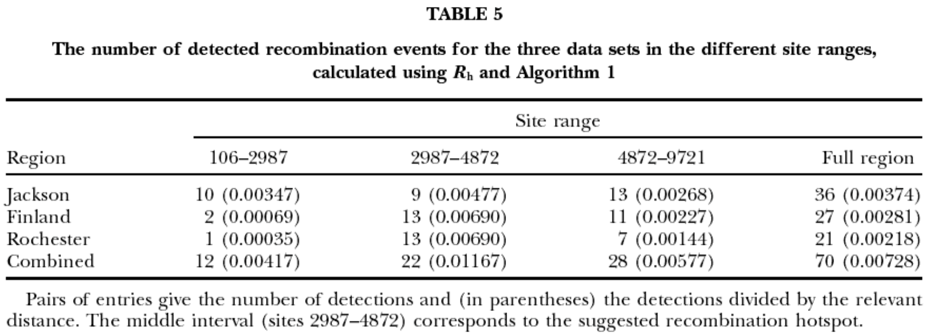


TABLE 5

The number of detected recombination events for the three data sets in the different site ranges, calculated using R_h and Algorithm 1

	Site range			
Region	106–2987	2987–4872	4872–9721	Full region
Jackson	10 (0.00347)	9 (0.00477)	13 (0.00268)	36 (0.00374)
Finland	2 (0.00069)	13 (0.00690)	11 (0.00227)	27 (0.00281)
Rochester	1 (0.00035)	13 (0.00690)	7 (0.00144)	21 (0.00218)
Combined	12 (0.00417)	22 (0.01167)	28 (0.00577)	70 (0.00728)

Pairs of entries give the number of detections and (in parentheses) the detections divided by the relevant distance. The middle interval (sites 2987–4872) corresponds to the suggested recombination hotspot.