BMI 826

Advanced Data Analysis

Karl Broman

Biostatistics & Medical Informatics, UW–Madison

kbroman.org
github.com/kbroman
@kwbroman

Course web: kbroman.org/AdvData

What is data analysis?

What is data analysis?

- Answer questions with data

What is data analysis?

- Answer questions with data
- Identify/develop appropriate methods to do so

What is data analysis?

- Answer questions with data
- Identify/develop appropriate methods to do so
- Quantify uncertainty
- Assess appropriateness of the method
- Identify problems in the data
- Understand where the data came from and possible biases or other limitations
- Manage and organize data
- Manage/organize/develop/test software and analyses so they are reproducible and correct

Important principles

1. You'll never know all the methods
2. Focus on the question and data, not the method
3. "Because you can" is not a good reason to do something

This course

- Data analysis projects
- Tools for organizing analyses so that they are reproducible
- Stories of data analysis projects, with lessons

Lesson 1

Follow up artifacts

They might be the most interesting results

Am. J. Hum. Genet. 63:861–869, 1998

Comprehensive Human Genetic Maps: Individual and Sex-Specific Variation in Recombination

Karl W. Broman,[1] Jeffrey C. Murray,[2,3] Val C. Sheffield,[2,4] Raymond L. White,[5] and James L. Weber[1]

[1]Marshfield Medical Research Foundation, Marshfield, WI; Departments of [2]Pediatrics and [3]Biology, University of Iowa, and [4]Howard Hughes Medical Institute, Iowa City; and [4]Eccles Institute for Human Genetics, University of Utah, Salt Lake City

Summary

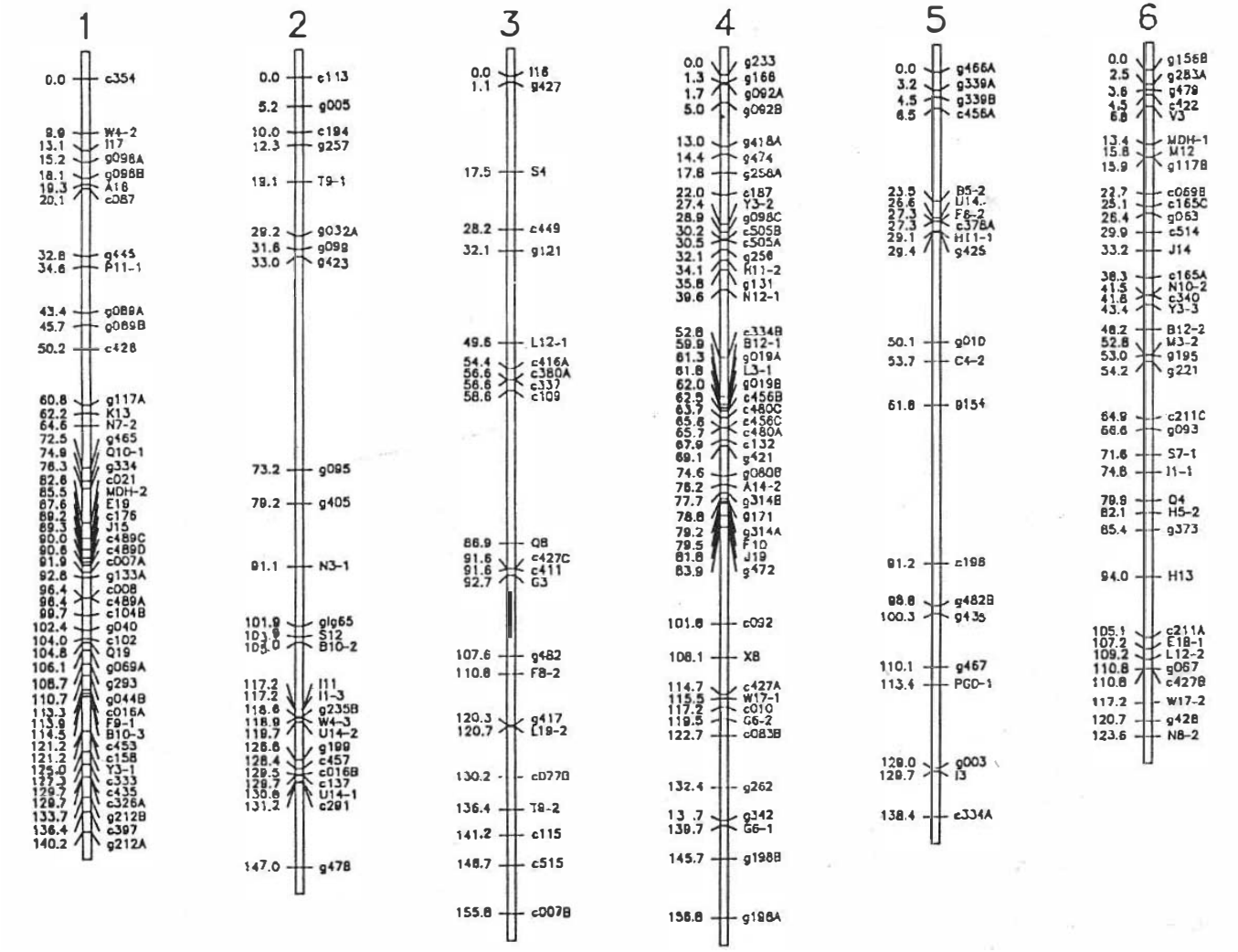
Comprehensive human genetic maps were constructed on the basis of nearly 1 million genotypes from eight CEPH families; they incorporated >8,000 short tandem-repeat polymorphisms (STRPs), primarily from Généthon, the Cooperative Human Linkage Center, the Utah Marker Development Group, and the Marshfield Medical Research Foundation. As part of the map building process, 0.08% of the genotypes that resulted in tight double recombinants and that largely, if not entirely, represent genotyping errors, mutations, or gene-conversion events were removed. The total female, male, and sex-averaged lengths of the final maps were 44, 27, and 35 morgans, respectively. Numerous (267) sets of STRPs

Introduction

Polymorphic DNA markers and their corresponding maps are an essential resource for localization of genes via linkage analysis, for characterization of meiosis, and for providing a foundation for the construction of physical maps. Although physical maps, including genome sequences, can provide the order of tightly linked polymorphisms, the physical maps do not provide genetic distances or other recombination data.

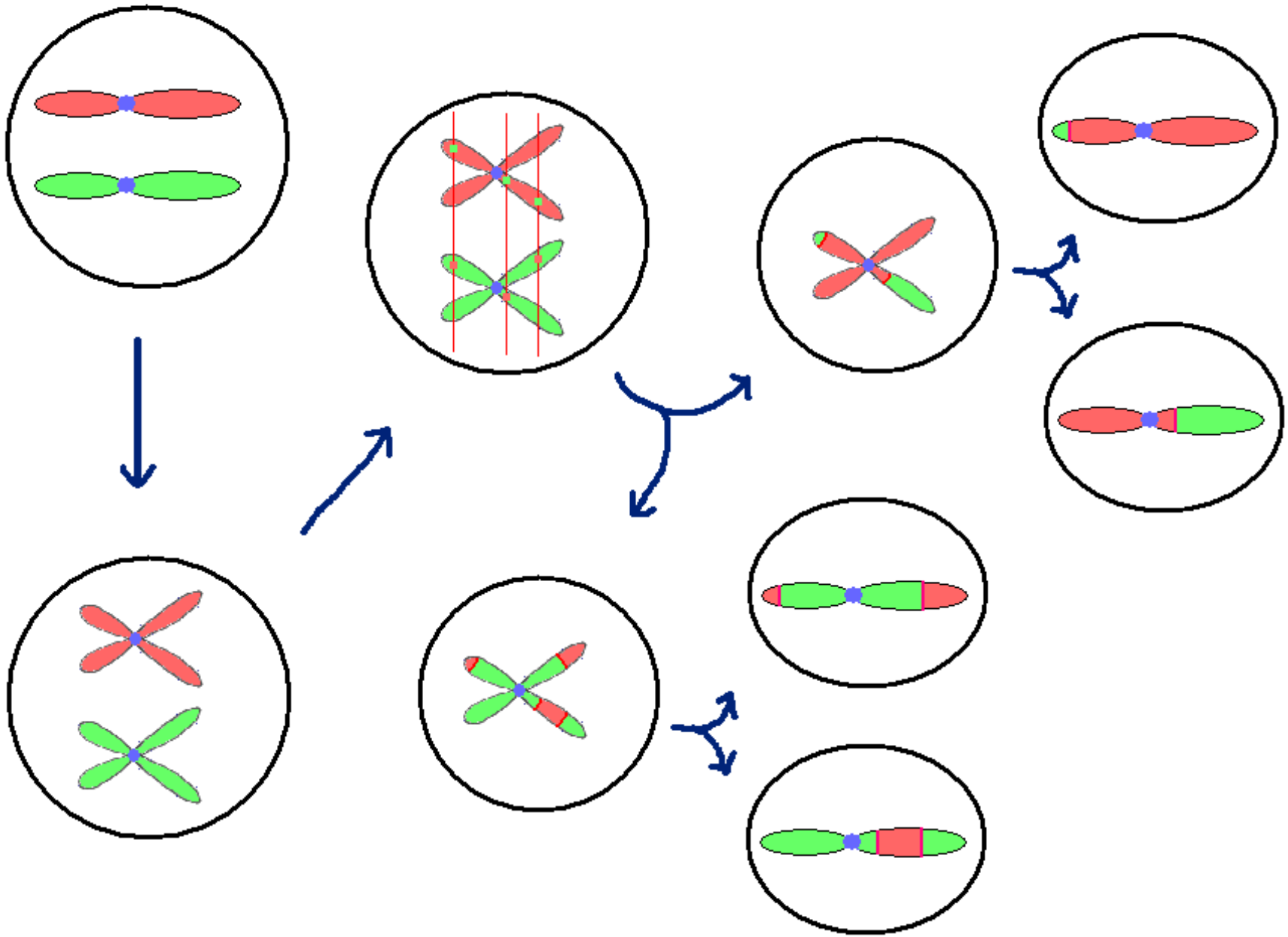
The era of human genome-scale genetic-map construction was heralded by the landmark paper by Botstein et al. (1980), in which both the use of DNA polymorphisms, as opposed to protein polymorphisms or other measurable phenotypes, in linkage mapping and an ef-

Eucalypt genetic map

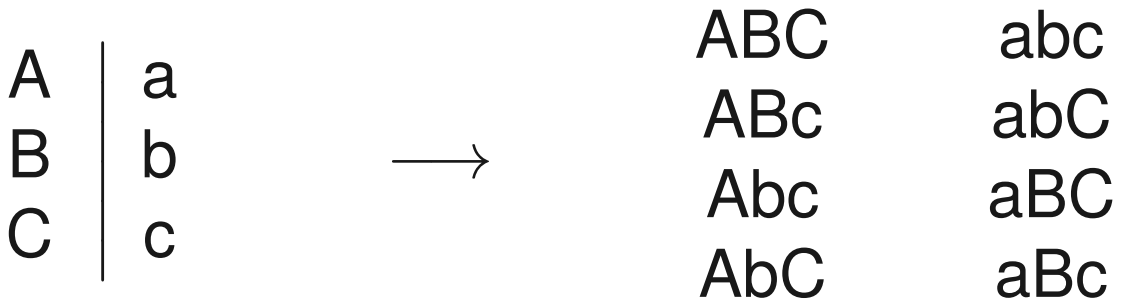


Byrne et al., Theor Appl Genet 91:869–875, 1995

Meiosis



Ordering markers

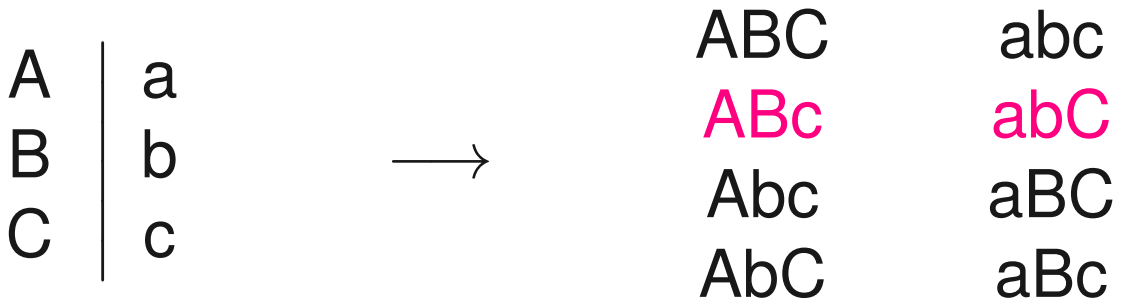


Marker orders: A–B–C A–C–B B–A–C

Ordering markers

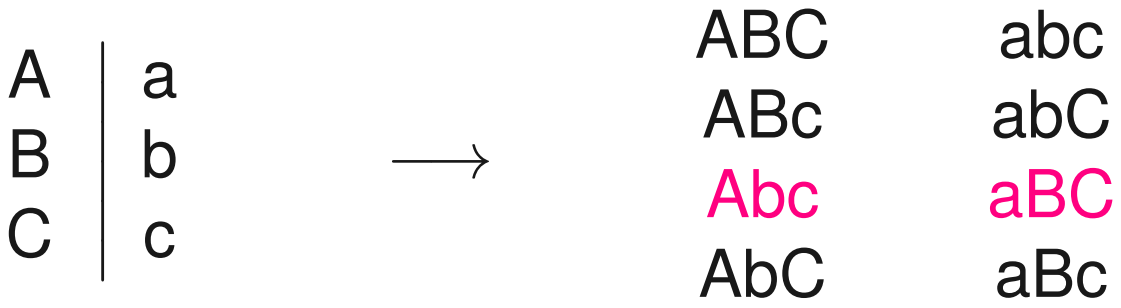
Marker orders: A–B–C A–C–B B–A–C

Ordering markers



Marker orders: A–B–C A–C–B B–A–C

Ordering markers



Marker orders: A–B–C A–C–B B–A–C

Ordering markers

A	a
B	b
C	c

$\longrightarrow$

ABC	abc
ABc	abC
Abc	aBC
AbC	aBc

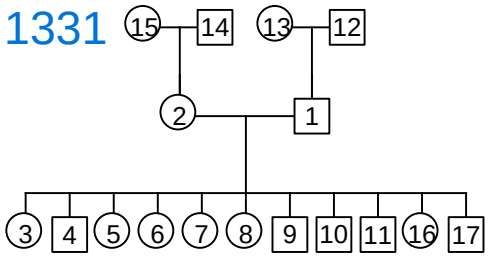
Marker orders: A–B–C A–C–B B–A–C

With M markers, there are M!/2 possible orderings.

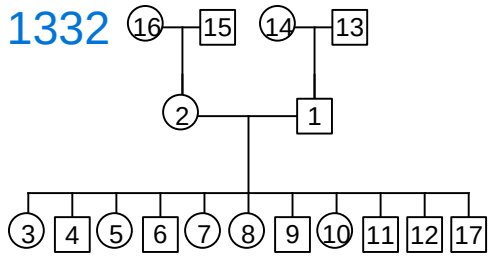
For $M = 100$, $M!/2 \approx 10^{157}$

CEPH pedigrees

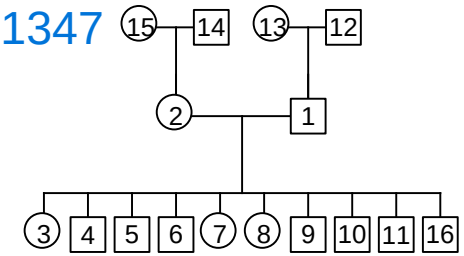
1331



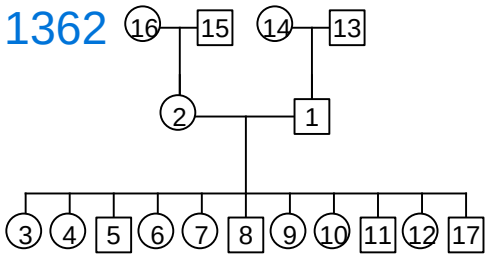
1332



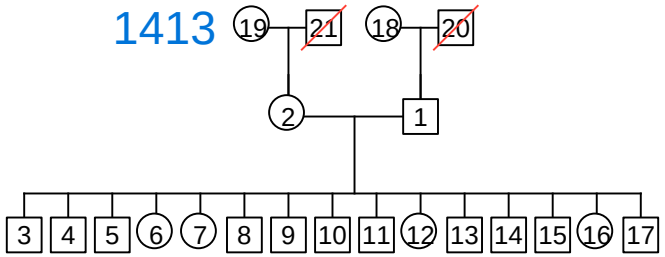
1347



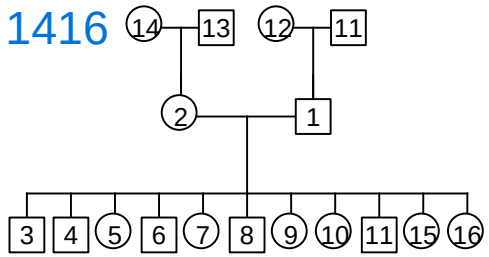
1362



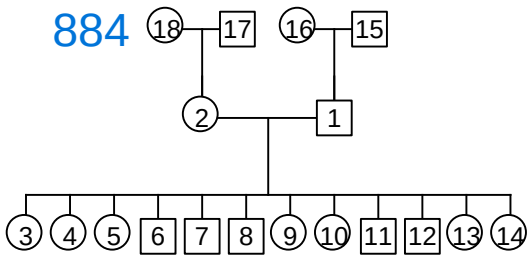
1413



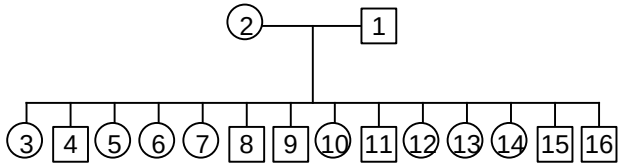
1416



884



102



Marshfield genetic maps: Tasks

- Assemble data
- Understand marker names
  AFM, UT, CHLC (GATA etc.), Mfd, D*S*
- Identify cryptic duplicates
- Order markers and identify genotyping errors
  Removed 764 / 969,425 genotypes

CRIMAP chrompic

```
1332-03 ma -11-i--11--111-i111-11-1111i--1111i-1111-i--11---1--11-1111-1-1i1---1...
1332-03 pa 0000----0000000o00o00-000-000-0000o00-000-00000-00001---000-00-o000-0...

1332-04 ma -11-i--11--111-1111-11-i111i--i1111-1111-i--11---1--11-1111-1-11i--11...
1332-04 pa 1111----1111111111i11-1i1-111-i111i11-111-11111-11111---111-11-1i1111...

1332-05 ma -11-i--11--111-i111-11-1111o--0000o-0000-o--00---0--00-0000-0-0o0--00...
1332-05 pa 0000----0000000o00o00-000-111-1111i11-111-1111--11111---111-11-i11111...

1332-06 ma -00-o--00--000-o000-00-0000o--0000o-0000-o--00---0--00-0000-1-11i--11...
1332-06 pa 1111----1111111i11i11-111-111-1111i11-111-11111-11111---111-11-1i1111...

1332-07 ma -00-o--00--000-o000-00-0000o--0000o-0000-o--00---0--00-0000-0-0o0--00...
1332-07 pa 1111----1111111i11i11-111-111-1111i11-111-1111--11111---111-11-i11111...

1332-08 ma -10-o--00--000-00-0-00-0000o--o0000-0000-o--00---0--11-1111-1-1i1--11...
1332-08 pa 0000----000000000-o00-010-000-o000o00-000-00000-00000---000-00-o00000...

1332-10 ma -11-i--1---111-i111-11-1111i--1111i-1111-i--11---1--11-1111-1-1i1--11...
1332-10 pa 1000-----000000o00o00-000-000-0000o00-000-00000-00000---000-00-o00000...

1332-11 ma -11-o--00--000-o000-00-0000o--0000o-0000-o--00---0--00-0000-0-0o0--00...
1332-11 pa 1111----1111111i11i11-111-111-1111i11-111-11111-11111---111-11-i11111...

1332-12 ma -00-i--11--111-i111-11---11i--1111i-1111-i--11---1--11-1111-1-1i1---1...
1332-12 pa 0000----0000000o00o00-0---000-0000o00-000-00000-00000---000-00-o000-0...

1332-17 ma -11-i--1---11--i111-1--1111i--1111i-1111-i--11---1--11-1100-0-00o--00...
1332-17 pa 0000-----0000--o00o00-000-000-0000o-0-000-0000--00000---000-00-0o0000...
```

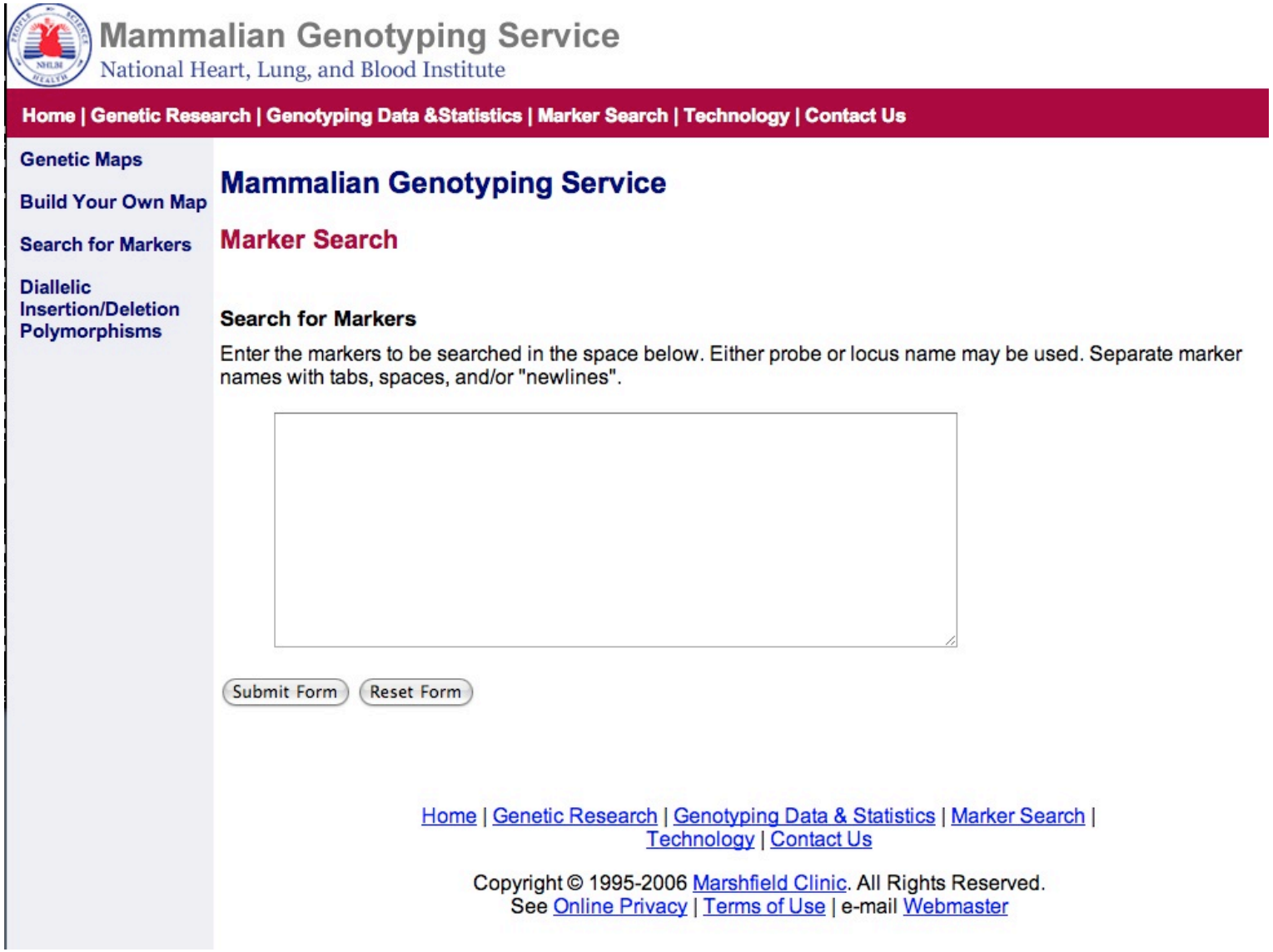
CRIMAP chrompic

```
1332-03 ma -11-i--11--111-i111-11-1111i--1111i-1111-i--11---1--11-1111-1-1i1---1...
1332-03 pa 0000----0000000o00o00-000-000-0000o00-000-00000-00001---000-00-o000-0...

1332-04 ma -11-i--11--111-1111-11-i111i--i1111-1111-i--11---1--11-1111-1-11i--11...
1332-04 pa 1111----1111111111i11-1i1-111-i111i11-111-11111-11111---111-11-1i1111...

1332-05 ma -11-i--11--111-i111-11-1111o--0000o-0000-o--00---0--00-0000-0-0o0--00...
1332-05 pa 0000----0000000o00o00-000-111-1111i11-111-1111--11111---111-11-i11111...

1332-06 ma -00-o--00--000-o000-00-0000o--0000o-0000-o--00---0--00-0000-1-11i--11...
1332-06 pa 1111----111111i11i11-111-111-1111i11-111-11111-11111---111-11-1i1111...

1332-07 ma -00-o--00--000-o000-00-0000o--0000o-0000-o--00---0--00-0000-0-0o0--00...
1332-07 pa 1111----111111i11i11-111-111-1111i11-111-1111--11111---111-11-i11111...

1332-08 ma -10-o--00--000-00-0-00-0000o--o0000-0000-o--00---0--11-1111-1-1i1--11...
1332-08 pa 0000----000000000-o00-010-000-o000o00-000-00000-00000---000-00-o00000...

1332-10 ma -11-i--1---111-i111-11-1111i--1111i-1111-i--11---1--11-1111-1-1i1--11...
1332-10 pa 1000-----000000o00o00-000-000-0000o00-000-00000-00000---000-00-o00000...

1332-11 ma -11-o--00--000-o000-00-0000o--0000o-0000-o--00---0--00-0000-0-0o0--00...
1332-11 pa 1111----111111i11i11-111-111-1111i11-111-11111-11111---111-11-i11111...

1332-12 ma -00-i--11--111-i111-11---11i--1111i-1111-i--11---1--11-1111-1-1i1---1...
1332-12 pa 0000----0000000o00o00-0---000-0000o00-000-00000-00000---000-00-o000-0...

1332-17 ma -11-i--1---11--i111-1--1111i--1111i-1111-i--11---1--11-1100-0-00o--00...
1332-17 pa 0000-----0000--o00o00-000-000-0000o-0-000-0000--00000---000-00-0o0000...
```

Top of chr 22

	Marker	Dnumber	sex-ave(cM)	female(cM)	male(cM)
1	ATA2G02	Unknown	0.00	0.00	0.00
			1.79	0.00	2.60
2	GATA198B05	Unknown	1.79	0.00	2.60
			2.27	3.32	0.00
3	AFM217xf4	D22S420	4.06	3.32	2.60
			4.26	4.51	5.42
4	AFM288we5	D22S427	8.32	7.83	8.02
			5.25	7.52	3.00
5	265yf5	D22S425	13.57	15.35	11.02
			0.03	0.00	0.65
6	GGAA10F06	D22S686	13.60	15.35	11.67
			0.84	0.00	0.82
7	AFMa037zd1	D22S539	14.44	15.35	12.49
			0.00	0.00	0.00
8	AFM292va9	D22S446	14.44	15.35	12.49
			3.27	5.91	0.00
9	Mfd51	D22S257	17.71	21.26	12.49

Marker search



10th worst graph

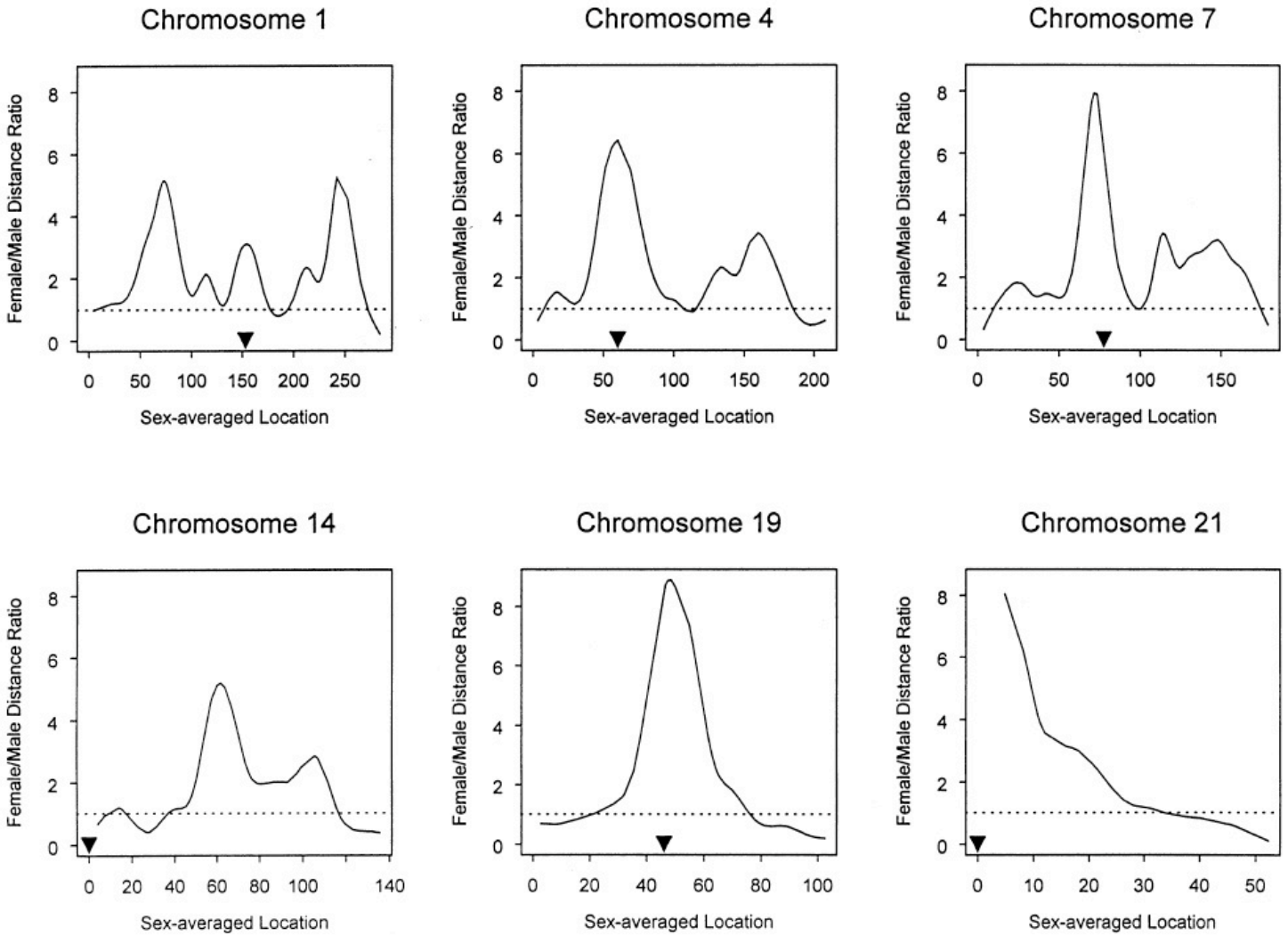
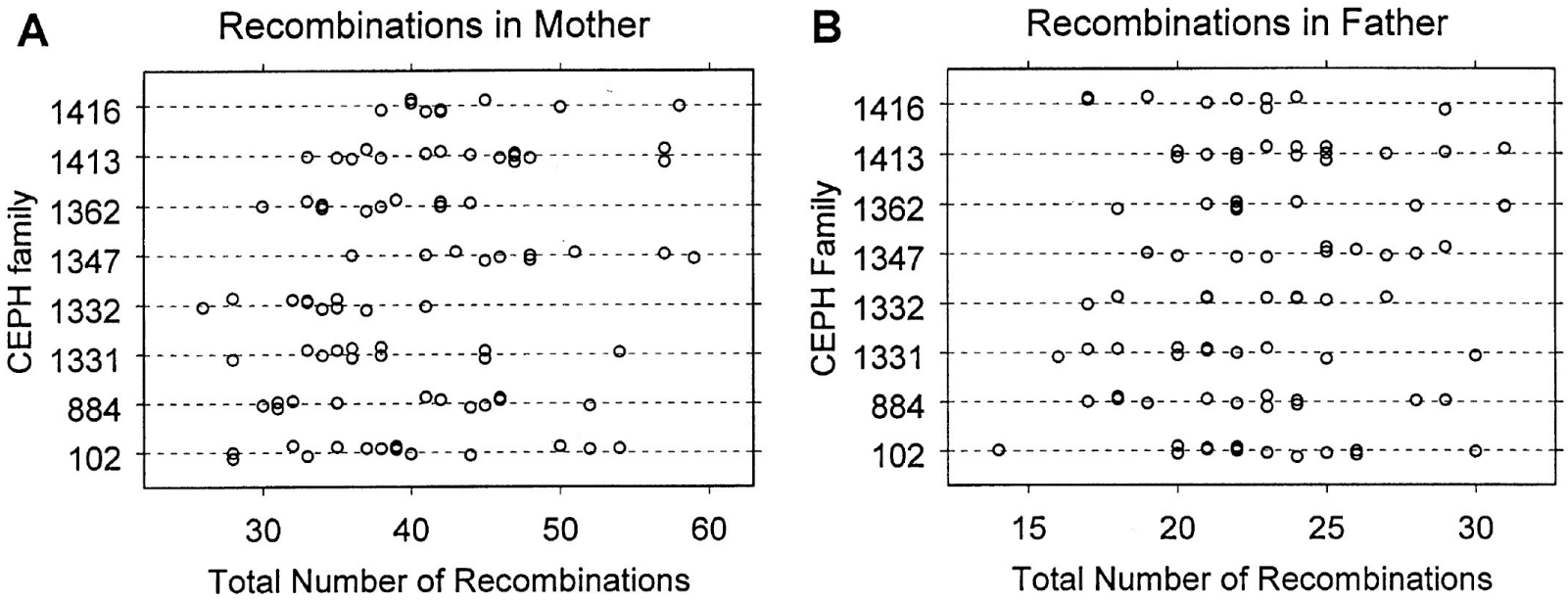


Figure 1 Plots of the female:male genetic-distance ratio against sex-averaged genetic location (in cM) along six selected chromosomes. Approximate locations of the centromeres are indicated by the triangles. The dashed lines correspond to equal female and male distances.

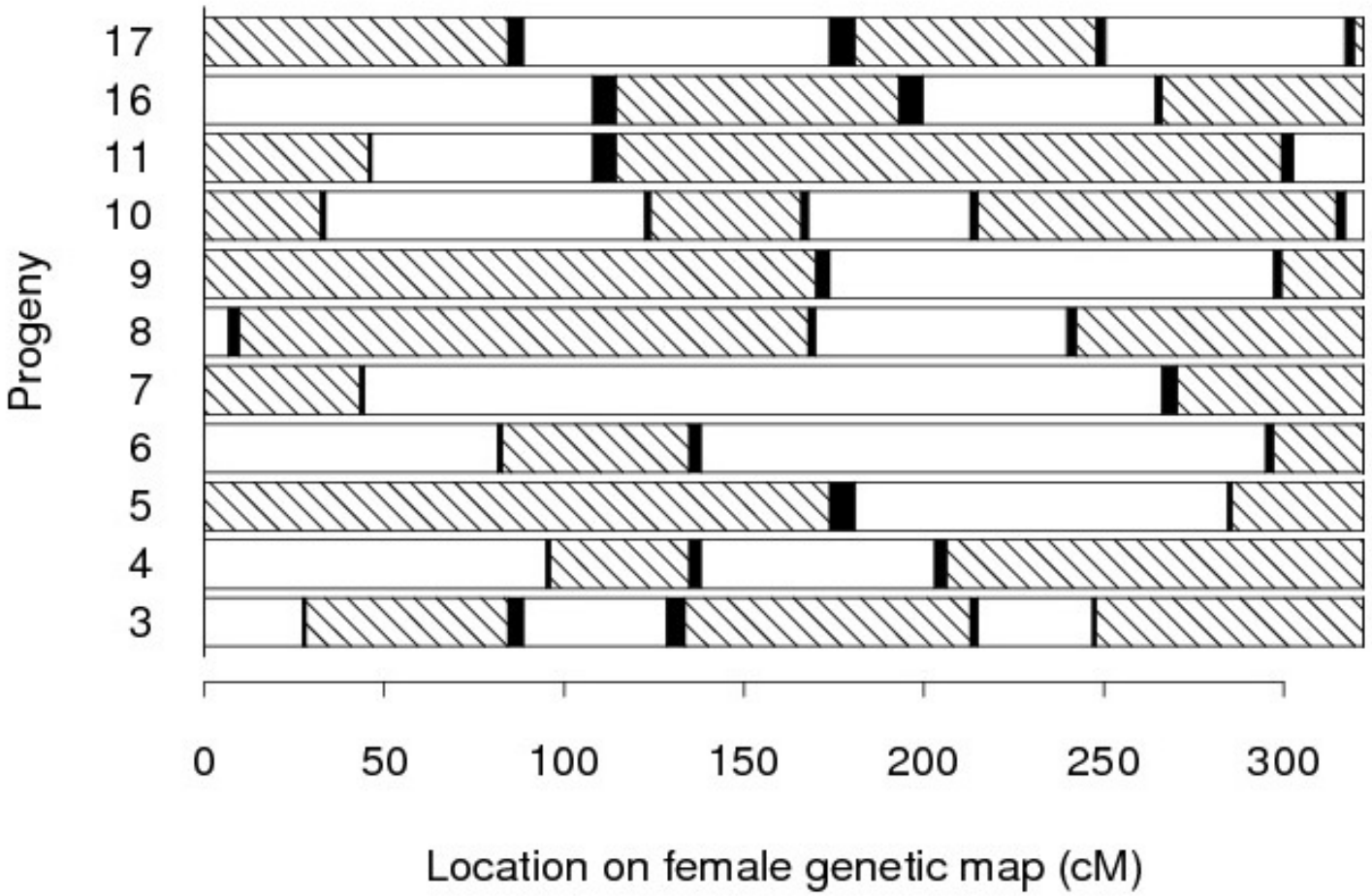
Broman et al., Am J Hum Genet 63:861–869, 1998

Total no. crossovers



Broman et al., Am J Hum Genet 63:861–869, 1998

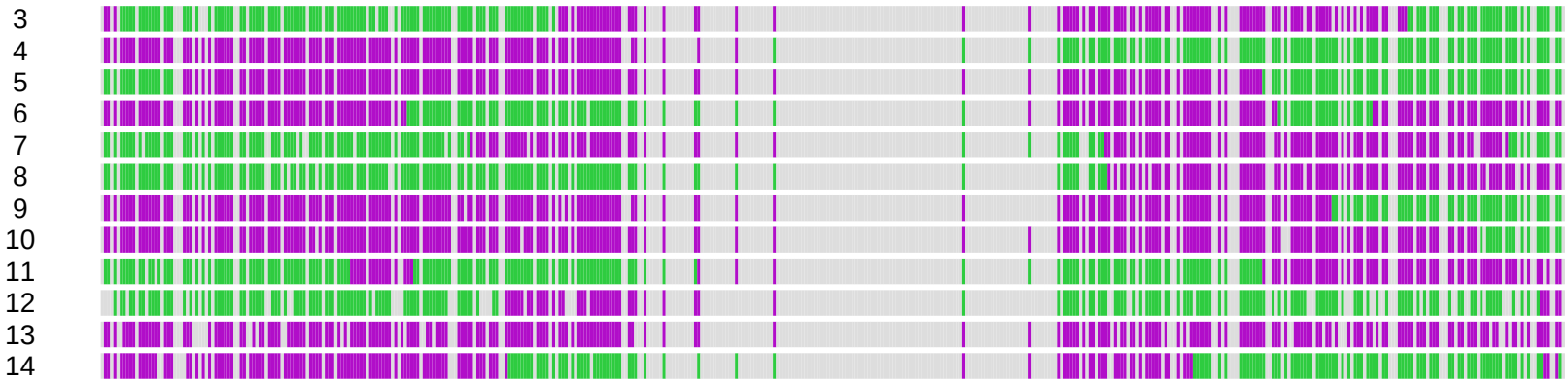
Crossover locations



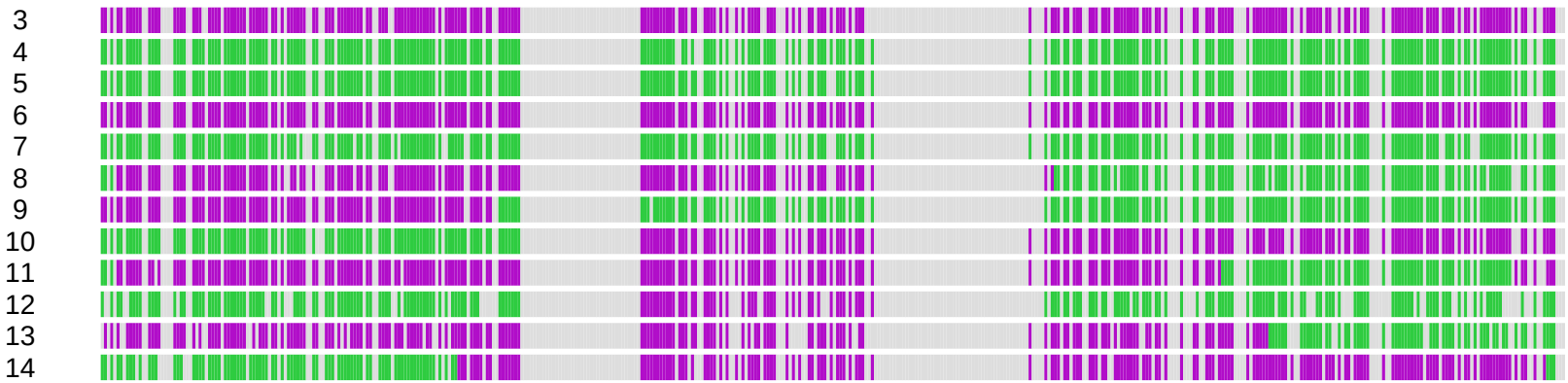
Broman et al., Am J Hum Genet 63:861–869, 1998

Family 884, chr 6

Maternal chromosomes



Paternal chromosomes



Am. J. Hum. Genet. 65:1493–1500, 1999

Long Homozygous Chromosomal Segments in Reference Families from the Centre d'Étude du Polymorphisme Humain

Karl W. Broman and James L. Weber

Marshfield Medical Research Foundation, Marshfield, WI

Summary

Using genotypes from nearly 8,000 short tandem-repeat polymorphisms typed in eight of the reference families from the Centre d'Étude du Polymorphisme Humain (CEPH), we identified numerous long chromosomal segments of marker homozygosity in many CEPH individuals. These segments are likely to represent autozygosity, the result of the mating of related individuals. Confidence that the complete segment is homozygous is gained only with markers of high density. The longest segment in the eight families spanned 77 cM and included 118 homozygous markers. All individuals in family 884 showed at least one segment of homozygosity: the father and mother were homozygous in 8 and 10 segments with an average length of 13 and 16 cM, respectively, and covering a total of 105 and 160 cM, respectively. The

a nearly limitless supply of DNA, making these families available for genotyping by investigators around the world. Many thousands of short tandem-repeat polymorphisms (STRPs) have been genotyped within a subset of eight of the CEPH families. These data provide a uniquely comprehensive view of the genomes of these individuals, which allows analyses that would not be possible on the basis of data from a more typical genome scan of 400 markers.

We recently constructed new genetic maps based on these families (Broman et al. 1998). As part of that work, we screened the data for apparent tight double-recombination events indicative of genotyping errors or mutations. In the process, we identified several long segments of noninformative markers in family 884, caused by long stretches of homozygous markers in the parents of that family.

Autozygosity

Homozygous Segments for Individual 884-02

Chromosome (Markers)	Cytogenetic Band(s)	Length (cM)	Proportion Homozygous	LOD Score
3 (D3S1571–D3S1617)	q28	4.9	9/9	5.53
4 (GATA144E02–D4S189)	p11-q12	11.1	21/21	12.26
5 (D5S398–D5S401)	q11-q14	29.8	77/77	46.21
6 (D6S1711–D6S278)	q11-q22	35.3	109/113	48.12
8 (D8S506–D8S385)	q22-q23	8.0	28/30	12.35
9 (D9S1802–D9S250)	q33	6.5	18/18	9.53
12 (D12S103–D12S1680)	q13-q21	11.3	43/43	21.82
16 (D16S494–D16S3107)	q21-q22	8.8	26/26	17.23
16 (D18S450–GATA51E05)	q21-q22	40.3	84/84	49.79
22 (D22S1156–D22Sl179)	q13	3.9	21/21	15.81

Broman and Weber, Am J Hum Genet 65:1493–1500, 1999

Am. J. Hum. Genet. 66:1911–1926, 2000

Characterization of Human Crossover Interference

Karl W. Broman and James L. Weber

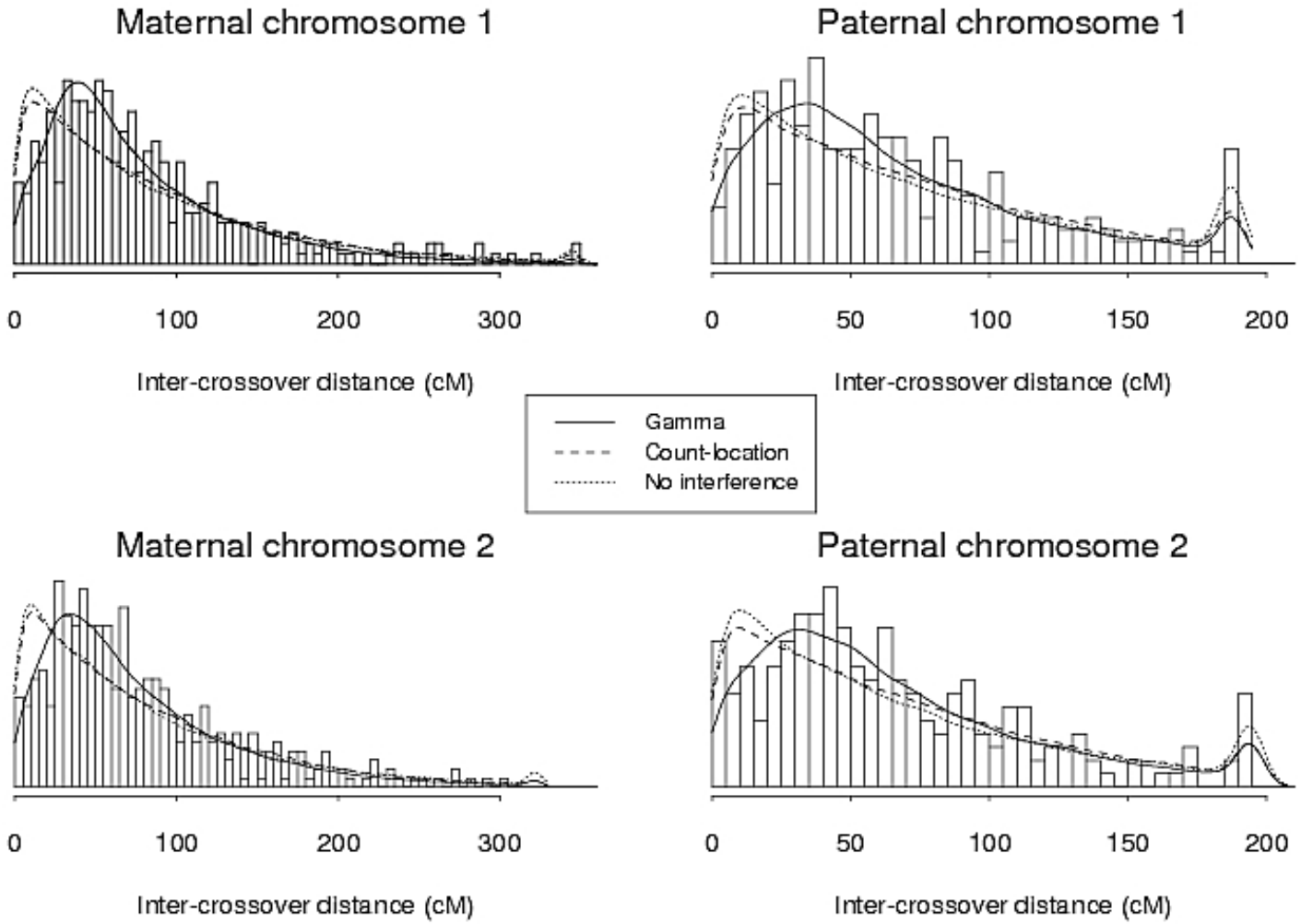
Marshfield Medical Research Foundation, Marshfield, WI

We present an analysis of crossover interference over the entire human genome, on the basis of genotype data from more than 8,000 polymorphisms in eight CEPH families. Overwhelming evidence was found for strong positive crossover interference, with average strength lying between the levels of interference implied by the Kosambi and Carter-Falconer map functions. Five mathematical models of interference were evaluated: the gamma model and four versions of the count-location model. The gamma model fit the data far better than did any of the other four models. Analysis of intercrossover distances was greatly superior to the analysis of crossover counts, in both demonstrating interference and distinguishing between the five models. In contrast to earlier suggestions, interference was found to continue uninterrupted across the centromeres. No convincing differences in the levels of interference were found between the sexes or among chromosomes; however, we did detect possible individual variation in interference among the eight mothers. Finally, we present an equation that provides the probability of the occurrence of a double crossover between two nonrecombinant, informative polymorphisms.

Introduction

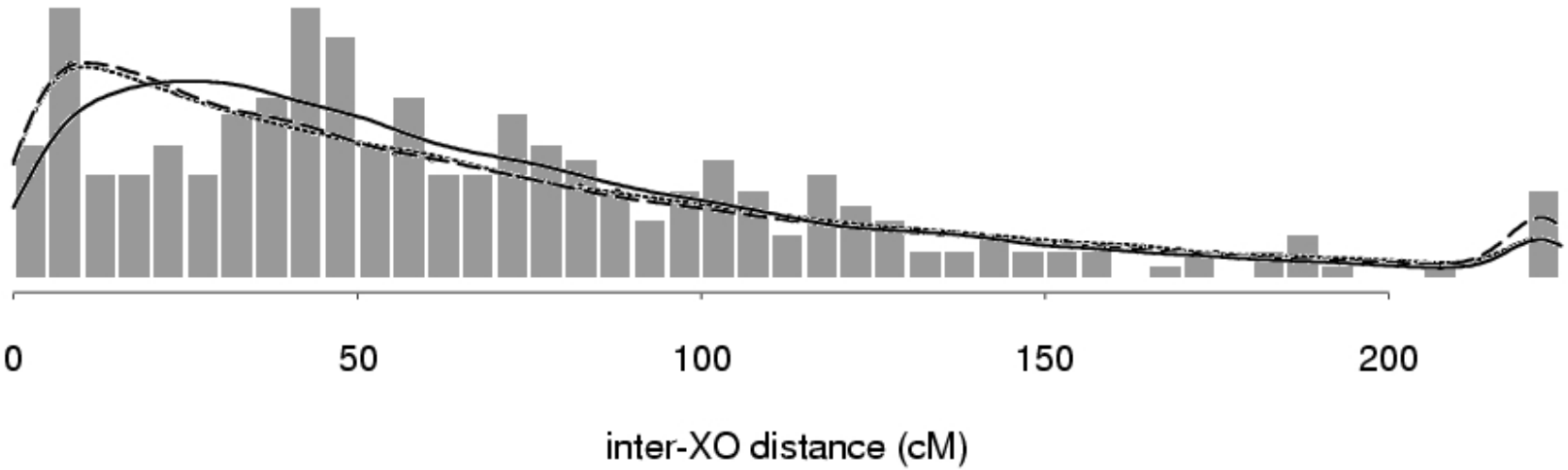
Crossover interference may be defined as the nonrandom placement of crossovers along chromosomes in meiosis. Interference was identified soon after the development of the first working models for the recombination process (Sturtevant 1915; Muller 1916). Strong evidence for matid interference is a dependence in the choice of strands involved in adjacent chiasmata. There is little consistent evidence for the presence of chromatid interference in experimental organisms (Zhao et al. 1995*a*), and any inference with regard to chromatid interference generally requires that data be available for all four products of meiosis (so-called "tetrad data");

Crossover interference

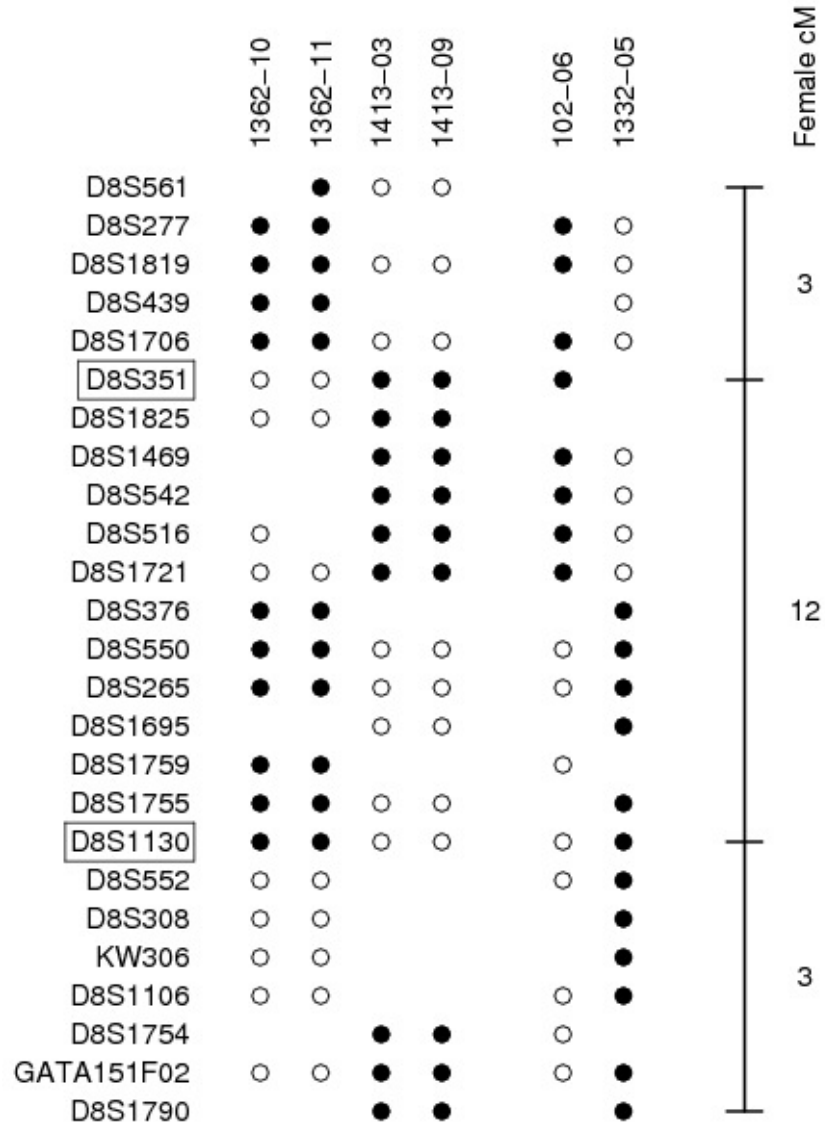


Broman and Weber, Am J Hum Genet 66:1911–1926, 2000

Maternal chr 8

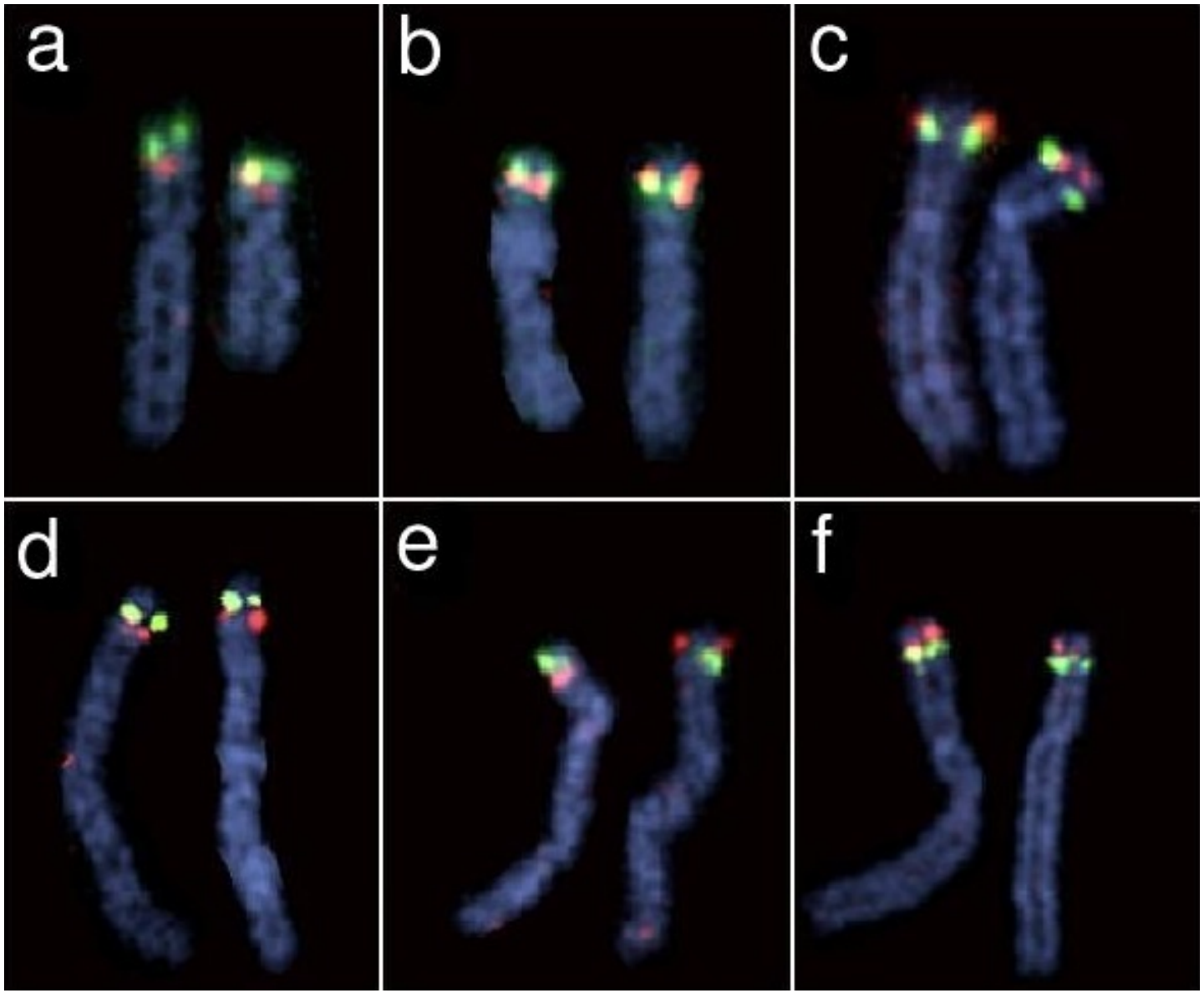


Apparent triple XOs



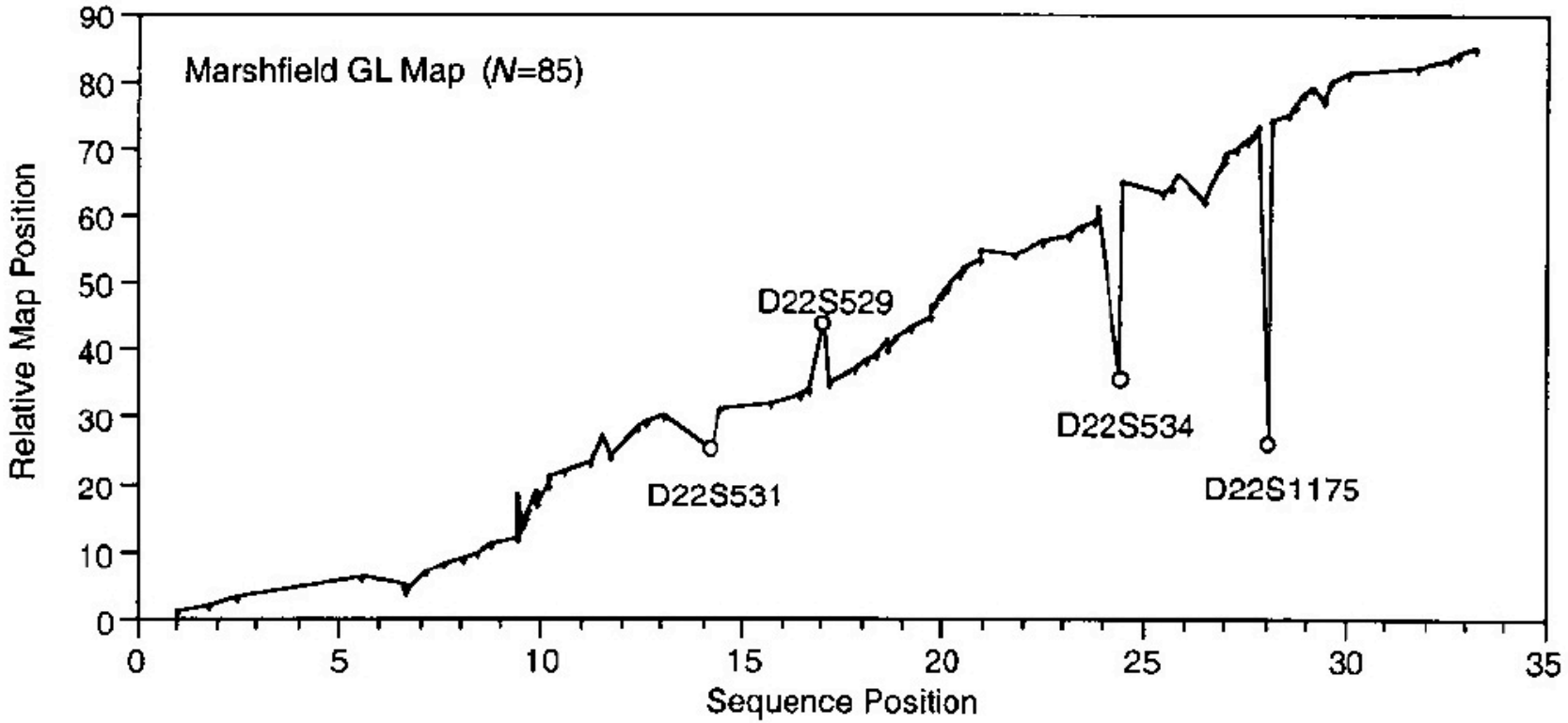
Broman et al., In: *Science and Statistics: A Festschrift for Terry Speed*, 2003

Chr 8p inversion



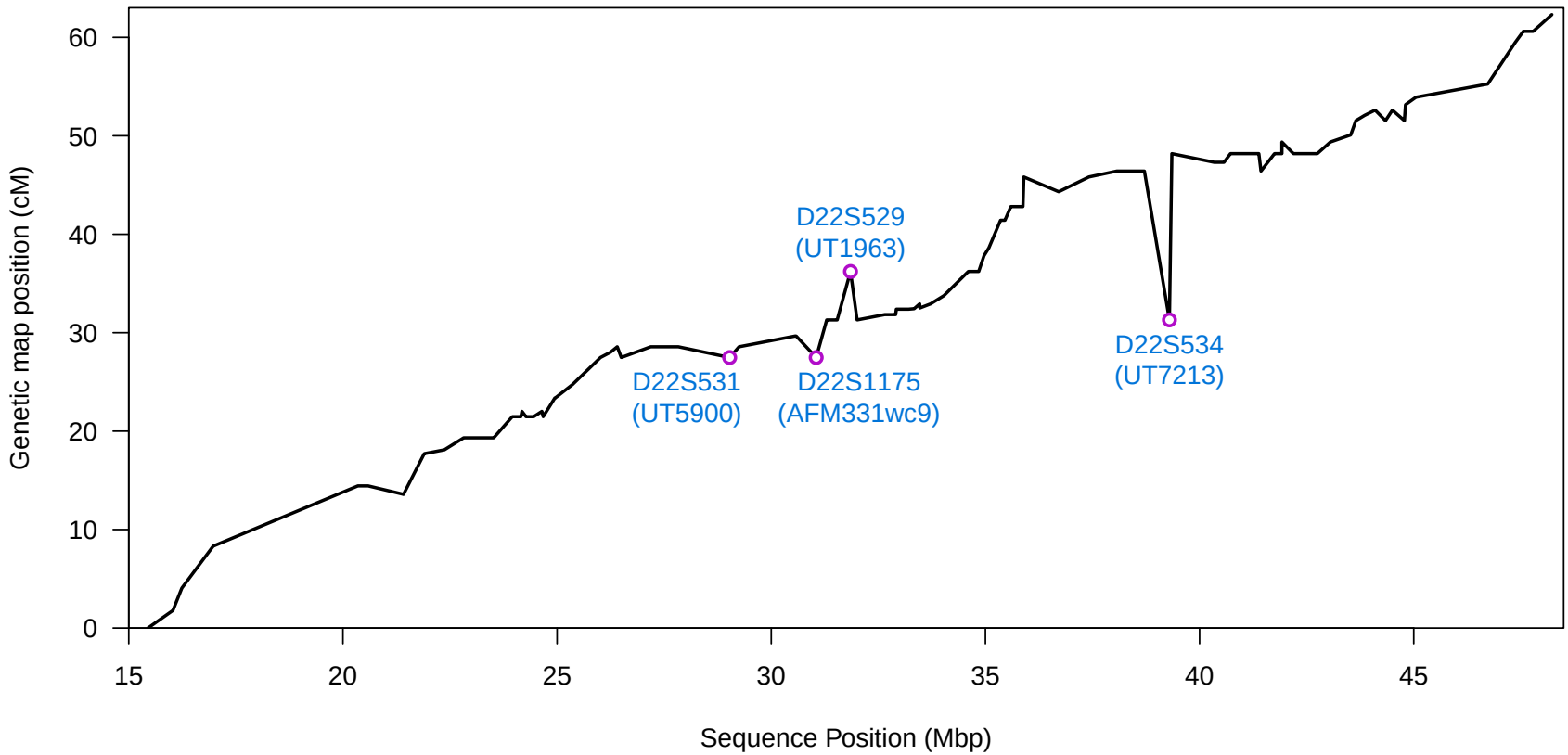
Broman et al., In: *Science and Statistics: A Festschrift for Terry Speed*, 2003

Comparison to sequence



Matise et al., Am J Hum Genet 70:1398–1410, 2002

Comparison to sequence



Lesson 1

Follow up artifacts

They might be the most interesting results

Lesson 2

The simplest things
can be the most important