Theor Appl Genet (1994) 88:289–292
© Springer-Verlag 1994

S. Gerber · F. Rodolphe

An estimation of the genome length of maritime pine (*Pinus pinaster* Ait.)

Received: 15 August 1993 / Accepted: 9 September 1993

Abstract The genome length, in units of Morgans or centimorgans, is a fundamental feature of a species. It can be calculated from a complete linkage map. However, the genome size can be estimated with partial linkage data. Using linkage data obtained by the analysis of a two-dimensional electrophoresis of the proteins contained in an haploid organ, the megagametophyte, we suggest an estimation and a confidence interval of the genome length of a gymnosperm, the maritime pine (*Pinus pinaster* Ait.). The results indicate an important gap between the physical and the genetic maps.

Key words Genome · Linkage · Pine · Gymnosperm · Two-dimensional electrophoresis

Introduction

Thanks to DNA markers, saturated linkage maps covering almost entire genomes are now available in different plant species, including barley (Heun et al. 1991), *Arabidopsis* (Reiter et al. 1992) and soybean (Diers et al. 1992). When the size of such maps is not modified by the addition of new markers, the total length of the genome under study can be considered as having been reached. This length is then obtained as the sum of all mapped intervals. However, even before the achievement of a complete map, genome length can be estimated with partial linkage data. Such an estimate can be used to predict the number of markers necessary to cover the genome and is useful for evaluating the overall relationship between physical and genetic distance, as measured by the number of kilobases of DNA per centimorgan. Hulbert et al. (1988) gave an estimation method based on maximum likelihood techniques. We suggest a justification for this method, which was not given by the authors, and calculate a confidence interval for the estimated value. We then propose an estimation of the genome size of the maritime pine (*Pinus pinaster* Ait.) using this method and a modification given by Chakravarti et al. (1991). The linkage data of Bahrman and Damerval (1989) and of Gerber et al. (1993), obtained with two-dimensional electrophoresis of the proteins contained in the haploid megagametophyte of the seeds of this gymnosperm, are both used. Due to the limited development of extensive mapping in gymnosperm species, such an estimation was not previously available.

Communicated by P. M. A. Tigerstedt

S. Gerber (✉)
ESV, bât 362, F-91405 Orsay Cedex, France

F. Rodolphe
INRA, Laboratoire de Biométrie, F-78370 Jouy en Josas, France

Materials and methods

The seed of pines contains a storage tissue surrounding the embryo, the megagametophyte, which results from an important development of a cell produced from the female gamete and which remains haploid. The tissue can be analysed by two-dimensional electrophoresis which separates proteins according to their isoelectric point and their molecular mass. The comparison of gels obtained by the individual analysis of different megagametophytes from the same tree reveals several kinds of protein variation, including presence/absence, position or quantity modifications. The genetic determinism of these variations can be inferred from the segregations observed among megagametophytes. Using this approach, Bahrman and Damerval (1989) compared 56 megagametophytes of a single maritime pine. The polymorphism of 37 loci accounted for the variations detected. The linkage relationships between these loci were investigated. A similar study was conducted on a sample of 18 pines of the same species (Gerber et al. 1993). An average of 12 megagametophytes per tree were analysed by two-dimensional electrophoresis. Genetic interpretations allowed 84 loci to be described. The 18 trees and each of their megagametophyte progeny were converted into 18 fictitious human families allowing human genetic techniques to be used for the linkage analysis.

The estimation of genome length suggested by Hulbert et al. (1988) can be justified as follows. Let M denote the number of informative pairs of loci i.e., both loci polymorphic in the segregating population studied. The possible linkage of these loci is tested with the lod score method. The lod score of a pair of loci is the decimal logarithm of the ratio of the likelihood calculated under the hypothesis that the loci are linked and the likelihood calculated with independent loci. When the lod score exceeds a certain threshold Z,

the loci are assumed to be linked. Let K denote the number of such pairs. The ratio K/M is then the probability that a pair of loci chosen at random will be declared linked. This probability can also be expressed as a function of G, the genome length, and X, the map distance between two loci for which a lod score of Z is expected. It is equal to $2X/G$ (see Appendix 1). The genome length is then:

$$G = \frac{2MX}{K}. \quad (1)$$

The genetic distance X is related to the recombination rate θ (Haldane's formula):

$$X = -\frac{1}{2}\ln(1-2\theta) \quad (2)$$

and θ is the solution to the equation (Lander and Botstein 1986, cited by Hulbert et al. 1988):

$$Z = n(\theta \log_{10} 2\theta + (1-\theta)\log_{10} 2(1-\theta) \quad (3)$$

where n is the number of gametes studied for the locus pair considered.

An α% confidence interval I can be calculated for the G values (Appendix 2):

$$I_\alpha(G) = \hat{G}(1 \pm n_\alpha K^{-1/2})^{-1}$$

where n_α stands for the value of a centered unit Normal law at an α probability.

The linkage data of Bahrman and Damerval (1989) were reanalysed with the MAPMAKER computer package (Lander et al. 1987), as a backcross, to calculate the lod scores. With this kind of data the number n of gametes studied is constant for each pair of loci, and the equations (1), (2) and (3) can be directly used. When the linkage data are obtained from different progenies, the number n will vary according to the pair of loci considered. In this case, Hulbert et al. (1988) calculate the value of θ and X for each informative pair of loci, with a given Z. The sum of the X values replace (MX) in equation (1). This technique was used with the data of Gerber et al. (1993).

Chakravarti et al. (1991) suggested a simple variation of this method. For a given Z, they chose, among pairs of loci with lod scores greater or equal to Z, the pair with the largest estimated θ value. This θ value is used instead of the one obtained from equation (3).

The method of Hulbert et al. (1988) (method 1) and the method of Chakravarti et al. (1991) (method 2) were both used to estimate the genome length with two values of Z, 2 and 3, for both sets of data.

Results and discussion

In the data obtained from 56 megagametophytes of one pine (Bahrman and Damerval 1989), all the possible pairs associating the 37 loci are informative, that is 666 pairs (37*36/2). Among them, 42 pairs have a lod score of 2 or greater with a maximal distance of 63.6 centimorgans, and 33 pairs have a lod score of 3 or greater with a maximal distance of 51.1 centimorgans.

The data obtained on 18 pines, that is 18 different progenies (Gerber et al. 1993), correspond to 84 loci. Among the 3486 (84*83/2) possible pairs of loci only 2711 were found informative in the different progenies. Of these, 89 pairs have a lod score equal or greater than 2, for a maximal distance of 34.7 centimorgans, while 69 pairs correspond to a lod score of 3 or greater with a maximal distance of 29.0 centimorgans. The results of the estimations of the genome length are given in Table 1.

Chakravarti et al. (1991), using computer simulations and experimental data, compared their method (method 2) with that of Hulbert et al. (1988) (method 1). They observed overestimates with method 1 and less inflated values with method 2. We found the opposite: method 1 always provides the smallest estimates for both sets of data. The genome length estimated with the linkage data of a single pine is less affected by the threshold chosen for the lod score than that obtained on 18 pines. In the experiment of Chakravarti et al. (1991), method 1 appeared to be affected by the sample size whereas method 2 was not. In our case, the estimates obtained with the progeny of one pine or of 18 pines are indeed closer with method 2 than with method 1. However, the estimations are always smaller with linkage data obtained on one progeny. According to these results, the genome size of maritime pine would be about 2000 centimorgans.

Among plant species, pines possess one of the largest amounts of DNA per diploid cell, about 40 picograms (Ohri and Khoshoo 1986). As a comparison, angiosperm trees like ash have 3 pg of DNA per diploid cell, elm and oak have about 2 pg, willow and birch about 1 pg (Bennet and Smith 1991). The DNA content of pines is more than nine times that of maize, 12 times that of tomato and a 100 times that of *Arabidopsis* [according to the data given by Bennet and Smith (1976, 1991)]. The genome of pines is composed of $10–20 \times 10^6$ kilobases (Rake et al. 1980). According to our estimations, one centimorgan would then represent between 5000 and 10000 kilobases. In maize, one centimorgan is supposed to represent between 2000 and 4000 kilobases (Brown and Sundaresan 1991), 500 kilobases in tomato (Messeguer et al. 1991), and about 140 kilobases in *Arabidopsis* (Chang et al. 1988). Comparing different species, there is no linear relationship between recombination and DNA quantity. However, when the DNA amount increases, the chiasma frequency per length of DNA decreases (Rees and Durrant 1986). This value,

Table 1 Estimations of the genome length according to the lod score threshold Z, with two methods and two sets of data

Linkage data		Genome length (5% confidence interval)	
		Z value	
		2	3
One progeny	method 1	1453 cM (1116–2083)	1455 cM (1085–2208)
	method 2	2019 cM (1550–2894)	2061 cM (1537–3128)
Eighteen progenies	method 1	1930 cM (1598–2436)	1787 cM (1446–2339)
	method 2	2114 cM (1750–2668)	2279 cM (1844–2983)

which can be expressed in centimorgans per kilobase, appears to be very low in pines.

The meaning of this observation is unclear. It could be due to a difference of DNA organisation between small and large genomes. Heterochromatin and zones of repetitive DNA sequences are supposed to be associated with a lower recombination rate (Flavell et al. 1985; Chandley and Mitchel 1988). Thuriaux (1977) suggests that recombination is concentrated on structural gene zones. This hypothesis is supported by an observation of Oliver et al. (1992) who found an approximate correlation between the pattern of genetic recombination and that of transcription in an yeast chromosome. Brown and Sundaresan (1991) suggest a model where structural genes and regions accessible to recombination would be interspersed with large, less-recombinogenic regions in the genome. The proportion and the organisation of these regions would explain the differences in rates of recombination per unit length of DNA between species with contrasting genome sizes. In conifers, two examples suggest an unusual organization of DNA sequences in the genome. The ADH genes seem to be larger or else more numerous in pines than in angiosperms (Harry et al. 1989; Kinlaw et al. 1990). Similarly, the sequence of the repeated units of the 18*s*–25*s* genes of pines is more than two times larger (27 kb) than the largest found in angiosperms (12 kb in wheat) (Cullis et al. 1987). It is even larger in spruce (35–43 kb, Bobola et al. 1992). Moreover, these genes are located in a larger number of sites than in angiosperms (Cullis et al. 1987).

The apparent limitation of genome length (in centimorgans) when DNA amount increases could have another explanation. Chromosome size could directly affect recombination. In yeast the DNA sequences of the smallest chromosome undergo recombination at a two-fold higher average than DNA sequences on the largest chromosome (Kaback et al. 1992). Kaback et al. (1992) suggest that this observation could be due to an increased amount of chiasma interference with chromosome size. This physical limitation would be responsible for an upper limit of genetic map sizes when the amount of DNA increases.

The gap between the physical and the genetic maps must be particularly important in pine species. The great DNA amount per cell seems to correspond to a limited genetic map size. This size will be better approached with the development of mapping in these species. However, the meaning of the large number of kilobases per centimorgan will still have to be solved. Since gymnosperm species appeared 140 millions years before the first angiosperms the peculiarities of their genome could be interesting from an evolutionary point of view.

Appendix 1

It is assumed that the loci are randomly distributed on the genome. On a chromosome of length L_i, the position of a locus has a uniform density equal to $1/L_i$. If x stands for the distance between two loci, its density can be written as:

$$f(x) = \int_0^{L_i - x} \frac{1}{L_i^2} dx + \int_x^{L_i} \frac{1}{L_i^2} dx = \frac{2(L_i - x)}{L_i^2}.$$

The probability that a pair of loci on the same chromosome are declared linked is:

$$P(Z, L_i) = \int_0^{L_i} P(\text{lod} - \text{score} > Z) f(x)\, dx$$

$$\text{lod} - \text{score} = \log_{10} \frac{\hat{p}\hat{\theta}^n (1-\hat{\theta})^{N-n} + (1-\hat{p})\hat{\theta}^{N-n}(1-\hat{\theta})^n}{2^{-N}}$$

where $\hat{p}$ and $\hat{\theta}$ stand for the maximum likelihood estimators of p and θ respectively.

$$P[\text{lod} - \text{score} > Z] \Leftrightarrow P[n > z_r \text{ or } n < z_1]$$

for some values z_r and z_1 functions of Z and N_i.

The integral $P(Z, L_i)$ could be calculated, but we use here the following rough approximation:

if θ is close to 1/2 (x large): $P[\text{lod} - \text{score} > Z] \approx 0$

if θ is very small (x very small): $P[\text{lod} - \text{score} > Z] \approx 1$

in-between $P[\text{lod} - \text{score} > Z]$ can be approximated by a symmetrical function in x, hence:

$$P(Z, L_i) \approx \int_0^x f(x)\, dx = \int_0^x \frac{2(L_i - x)}{L_i^2} = \frac{2X}{L_i} - \frac{X^2}{L_i^2}.$$

If X is substantially smaller than L: $P(Z, L_i) \approx \frac{2X}{L_i}$.

Considering m different chromosomes, the probability that a locus is present on the ith chromosome is L_i/G, where $G = \sum_{i=1}^m L_i$ is the genome length. The probability $P(Z, G)$ for two loci to be linked is then:

$$P(Z, G) = \sum_{i=1}^m \left(\frac{L_i}{G}\right)^2 P(Z, L_i) = \sum_{i=1}^m \frac{L_i}{G^2} 2X = \frac{2X}{G}.$$

Appendix 2

Let S stand for the set of all informative pairs of loci.

$$X_t = \sum_{h \in S} X_h.$$

The value of X_h depends on the number of meioses (or gametes) observed for the informative pair h. If these numbers are equal, $X_t = MX$.

$$\hat{G} = 2\frac{X_t}{K}.$$

We define the stochastic variables Y_h:

$Y_h = 1$ if the loci of the informative pair h are declared linked.

$Y_h = 0$ if the loci are declared independant

$$P[Y_h = 1] = \mu_h \approx \frac{2X_h}{G}$$

$$E[Y_h] = \mu_h$$

$$V[Y_h] = \mu_h(1 - \mu_h)$$

and also:

$$K = \sum_{h \in S} Y_h$$

$$E[K] = \sum_{h \in S} E[Y_h]$$

$$V[K] = E[K^2] - E[K]^2$$
$$= E\left[\sum_{h,h' \in S} Y_h Y_{h'}\right] - \left(E\left[\sum Y_h\right]\right)^2$$
$$= \sum_{h \in S} E[Y_h^2] - \sum_{h \in S} (E[Y_n])^2 + \sum_{\substack{h,h' \in S \\ h \neq h'}} (E[Y_h Y_{h'}] - E[Y_h]E[Y_{h'}]).$$

The two first terms are $\sum_{h \in S} V[Y_h]$. The third term is $\sum_{\substack{h,h' \in S \\ h \neq h'}} C[Y_h Y_{h'}] = 0$, since h and h' can belong to two different types:

$h = \{i, j\}$ $h' = \{k, l\}$ then Y_h and $Y_{h'}$ are obviously independent and their covariance is zero.

$h = \{i, j\}$ $h' = \{i, k\}$ then

$$E[Y_{ij} Y_{kl}] = P[Y_{ij} = 1] P[Y_{ik} = 1 / Y_{ij} = 1]$$
$$\approx P[Y_{ij} = 1] P[Y_{ik} = 1] \quad \text{(neglecting side effects)}$$
$$= E[Y_{ij}] E[Y_{ik}].$$

K is the sum of the approximately independent Bernouilli variables Y_h. We can build an approximate confidence interval for K based on the Gaussian approximation:

$$\frac{K}{2X_t} = \hat{G}^{-1} \rightarrow N\left(G^{-1}, \frac{1}{2GX_t}\right)$$

since:

$$V[K] = \sum_{h \in S} \mu_h (1 - \mu_h) \approx \sum_{h \in S} \frac{2X_h}{G}\left(1 - \frac{2X_h}{G}\right)$$
$$= \frac{2X_t}{G} - \frac{4\sum_{h \in S} X_h^2}{G} = \frac{2X_t}{G}.$$

An α confidence interval is:

$$I_\alpha(\hat{G}^{-1}) = \hat{G}^{-1} \pm n_\alpha \left(\frac{1}{2X_t G}\right)^{1/2} \approx \hat{G}^{-1}\left(1 \pm n_\alpha \left(\frac{\hat{G}}{2X_t}\right)^{1/2}\right)$$
$$= \hat{G}^{-1}(1 \pm n_\alpha K^{-1/2})$$
$$I_\alpha(G) = \hat{G}(1 \pm n_\alpha K^{-1/2})^{-1}.$$

References

Bahrman N, Damerval C (1989) Linkage relationships of loci controlling protein amounts in maritime pine (*Pinus pinaster* Ait.). Heredity 63:267–274

Bennett MD, Smith JB (1976) Nuclear DNA amount in angiosperms. Phil Trans R Soc Lond B 274:227–274

Bennett MD, Smith JB (1991) Nuclear DNA amount in angiosperms. Proc R Soc Lond B 334:309–345

Bobola MS, Eckert RT, Klein AS (1992) Restriction fragment variation in the nuclear ribosomal DNA repeat unit within and between *Picea rubens* and *Picea mariana*. Can J For Res 22: 255–263

Brown J, Sundaresan V (1991) A recombination hotspot in the maize A1 intragenic region. Theor Appl Genet 81:185–188

Chakravarti A, Lasher LA, Reefer JE (1991) A maximum likelihood method for estimating genome length using genetic linkage data. Genetics 128:175–182

Chandley AC, Mitchell AR (1988) Hypervariable minisatellite regions are sites for crossing-over at meiosis in man. Cytogenet Cell Genet 48:152–155

Chang C, Bowman JL, Dejohn AW, Lander ES, Meyerowitz EM (1988) Restriction fragment length polymorphism linkage map for *Arabidopsis thaliana*. Proc Natl Acad Sci USA 85: 6856–6860

Cullis CA, Creissen GP, Gorman SW, Teasdale RD (1987) The 25s, 18s, and 5s ribosomal RNA genes from *Pinus radiata* D. Don. Molecular genetics of forest trees. In: Cheliak WM, Yapa AC (eds). Proc second IUFRO working parties in molecular genetics, Canada: 34–40

Diers BW, Keim P, Fehr WR, Shoemaker RC (1992) RFLP analysis of soybean seed protein and oil content. Theor Appl Genet 83:608–612

Flavell RB, O'Dell M, Smith DB, Thompson WF (1985) Chromosome architecture: the distribution of recombination sites, the structure of ribosomal DNA loci and the multiplicity of sequences containing inverted repeats. Molecular form and function of the plant genome. In: van Volten-Doting L, Groot, GSSP, Hall TC (eds). NATO ASI, 83, Plenum Press, New York: 1–14

Gerber S, Rodolphe F, Bahrman N, Baradat Ph (1993) Seed-protein variation in maritime pine (*Pinus pinaster* Ait.) revealed by two-dimensional electrophoresis: genetic determinism and construction of a linkage map. Theor Appl Genet 85:521–528

Harry DE, Mordecai KS, Kinlaw CS, Loopstra CA, Sederoff RR (1989) DNA sequence diversity in alcohol dehydrogenase genes from pines. Proc 20th southern forest tree improvement conference, Charleston, South Carolina: 373–380

Heun M, Kennedy AE, Anderson JA, Lapitan NLV, Sorrells ME, Tanksley SD (1991) Construction of a restriction fragment length polymorphism map for barley (*Hordeum vulgare*) Genome 34:437–447

Hulbert SH, Ilott TW, Legg EJ, Lincoln SE, Lander ES, Michelmore RW (1988) Genetic analysis of the fungus, *Bremia lactucae*, using restriction fragment length polymorphism. Genetics 120: 947–958

Kaback DB, Guacci V, Barber D, Mahon JW (1992) Chromosome size-dependent control of meiotic recombination. Science 256: 228–232

Kinlaw CS, Harry DE, Sederoff RR (1990) Isolation and Characterization of alcohol dehydrogenase cDNAs from *Pinus radiata*. Can J For Res 20:1343–1350

Lander ES, Green P, Abrahamson J, Barlow A, Daly MJ, Lincoln SE, Newburg L (1987) Mapmaker: an interactive computer package for constructing primary genetic linkage maps for experimental and natural populations. Genomics I, 174–181

Messeguer R, Ganal M, de Vicente MC, Young ND, Bolkan H, Tanksley SD (1991) High-resolution RFLP map around the root knot nematode resistance gene (Mi) in tomato. Theor Appl Genet 82:529–536

Ohri D, Khoshoo TN (1986) Genome size in gymnosperms. Pl Syst Evol 153(1–2): 119–131

Oliver SG et al. (1992) The complete DNA sequence of yeast chromosome III. Nature 357:38–46

Rake AV, Miksche JP, Hall RB, Hansen KM (1980) DNA reassociation kinetics of four conifers. Can J Genet Cytol 22:69–79

Rees H, Durrant A (1986) Recombination and genome size. Theor Appl Genet 73:72–76

Reiter RS, Williams JGK, Feldmann KA, Rafalski JA, Tingey SV, Scolnik PA (1992) Global and local genome mapping in *Arabidopsis thaliana* by using recombinant inbred lines and random amplified polymorphic DNAs. Proc Natl Acad Sci USA 89:1477–1481

Thuriaux P (1977) Is recombination confined to structural genes on the eucaryotic genome? Nature 268:460–462