The number of genotypic assignments on a genealogy II. Further results for linear systems

N. J. CAMP *
Division of Molecular and Genetic Medicine
University of Sheffield

C. CANNINGS †
School of Mathematics and Statistics and
Division of Molecular and Genetic Medicine
University of Sheffield

N.A. SHEEHAN
Department of Mathematical Sciences
Loughborough University

Abstract

In a previous paper we demonstrated how the number of possible genotypic assignments consistent with the rules of Mendelian genetics and with any known phenotypes could be calculated for an arbitrary genealogy. Here, we present further results for several regular genealogies constructed according to some specified recursive formulae and for which the growth of the genotypic statespace, with increasing genealogical size, can be described by a linear system.

Keywords: Genealogy, State Space, Monte Carlo estimation, Peeling

*Current address; Genetic Epidemiology Unit, University of Utah, Salt Lake City
†Author for correspondence

1 Introduction

The number of possible genotypes which can be assigned to the individuals of a genealogy, or pedigree, subject to the laws of genetics and taking the known phenotypes into account, is of interest for several reasons. In particular, knowledge of the size of the genotypic state space, or some bounds on it, and its relationship to the genealogical structure is relevant to investigations into the effectiveness of Markov chain Monte Carlo methods applied to any analysis of genetic data on groups of related individuals, although the mixing properties of these methods are of paramount concern. A crude upper bound for the number of possibilities, which completely ignores the genealogical structure, is $\prod_i p_i$ where p_i is the number of genotypes which the i^{th} individual can have, given his specified phenotype. In a complex model, where each genotype can give rise to each phenotype, this bound becomes g^n, with g being the number of distinct genotypes for the trait in question and n, the number of individuals. In Camp, Cannings & Sheehan (1994) we showed that a more reasonable estimate of the form γ^n can be given (for n large) where γ depends on the structure of the genealogy and, for example, is often closer to 2 than the crude upper limit of 3 for an autosomal diallelic system. The method used was the "peeling"method of Cannings, Thompson & Skolnick (1978) originally applied to calculate probabilities but shown to be appropriate also for counting states. It is assumed that the reader has some familiarity with the previous paper, although we give a simple example below by way of introduction. We note that the method is suitable for structures other than genealogies e.g. any Markov random field.

The focus in the earlier paper (Camp et al. 1994) was on introducing the methodology and applying it to some simple examples. Here we introduce the notion of a linear system in general and present results for some particular such systems.

2 Review of the Methodology and Notation

2.1 The method

We now review the method by applying it to a genetic system with two alleles, a and b, and Mendelian inheritance for the simple example in Figure 1 (Camp et al. 1994). Taking individual A as the reference individual, the counting is performed via a function $N_A(i)$ where i indexes the genotype of A, coded by $aa = 1$, $ab = 2$ and $bb = 3$, and $N_A(i)$ gives the number of possible states for that part of the genealogy "below" individual A (see Thompson (1986)) i.e. including B, C and D and parts of the genealogy attached to them, when A is of type i. Applying the laws of Mendelian inheritance we have

$$\begin{aligned} N_A(1) &= N_B(1)N_C(1)N_D(1) \\ &\quad +N_B(2)N_C(1,2)N_D(1,2) \\ &\quad +N_B(3)N_C(2)N_D(2) \end{aligned} \tag{2-1}$$

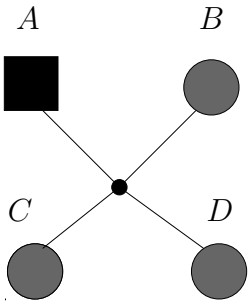

Figure 1: A simple nuclear family depicted as a standard marriage node graph with father A, taken as the reference individual for the following example, mother B and two offspring, C and D. One should imagine this family embedded within a genealogy.

where $N_X(i)$ is the number of states of the genealogy attached at X (and excluding the nuclear family) when X has genotype i, and $N_X(1,2) \equiv N_X(1) + N_X(2)$. By symmetry, $N_A(3) = N_A(1)$ while

$$\begin{aligned} N_A(2) &= N_B(1)N_C(1,2)N_D(1,2) \\ &\quad + N_B(2)N_C(1,2,3)N_D(1,2,3) \\ &\quad + N_B(3)N_C(2,3)N_D(2,3). \end{aligned} \tag{2-2}$$

In the case where there is no genealogy beyond B, C and D, and if all genotypes are possible for B, C and D, then

$$N_B(i) = N_C(i) = N_D(i) = 1, \ \forall\, i,$$

since B is a founder and C and D are both finals, so we conclude that

$$N_A(1) = N_A(3) = 6 \text{ and } N_A(2) = 17$$

from (2-1) and (2-2), respectively. Thus if the genealogy just consists of the nuclear family in Figure 1, the total number of states is 29.

2.2 Linear Systems

In Camp et al. (1994), we defined a *regular* genealogy to be one which can be constructed from simple units according to some specified recursive formula. Here we concentrate on those regular genealogies for which the number of individuals increases linearly with the addition of each new unit. In order to combine these basic building units into genealogies, we define a *linking set* to be a set of individuals common to both the new unit (or units) to be attached and the existing structure. The specification of a linking set induces a natural linking pattern in the basic unit whereby we get *input* individuals who identify with individuals in the previous unit and *output* individuals who will form the connection with the next unit to be added to the structure. A regular *linear* genealogy can be thought

of as one constructed recursively from simple basic units via the same linking set, or combination of linking sets, each time. We are interested in the value of γ, where γ^n describes the number of genotypic states on a genealogy, $\mathcal{G}$, of n individuals for n large, and in the relationship between γ and the structure of $\mathcal{G}$. To this end, we consider the value of γ for different regular linear genealogies for which this value can be calculated formally.

3 Two offspring and two alleles

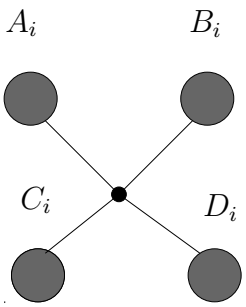

Figure 2: The basic building block consisting of a simple nuclear family with two offspring.

We will begin with a consideration of linear genealogies constructed from the basic unit of a nuclear family with two children depicted in Figure 2 (Camp et al. 1994). Note that this is the simple nuclear family of Figure 1 without the traditional gender distinctions (squares for males, circles for females). At the outset, we will restrict attention to a diallelic genetic system with Mendelian segregation. We commence with the case where a linking set comprises one single individual. Linking sets consisting of more than one individual will be considered in Section 5.

3.1 Input and output individuals are distinct

There are four ways in which the basic building blocks of Figure 2 can be combined to form a linear genealogy with a single linking individual and distinct input and output individuals. These are illustrated in Figure 3. The arrows indicate the direction in which information is passed in the sense that information on the number of states is collated, from the rest of the pedigree via the N-function on the input individual, and from the remaining members of the new block, into the N-function on the output individual. The recursions for the four cases are of the form

$$\mathbf{N_k} = \mathbf{M_J}\mathbf{N_{k-1}}$$

with $\mathbf{N_l} = (N(1), N(2))_l^T$ being a vector of counts for the number of states in the system already included up to stage l for genotypes 1 and 2 ($N(3) = N(1)$, by

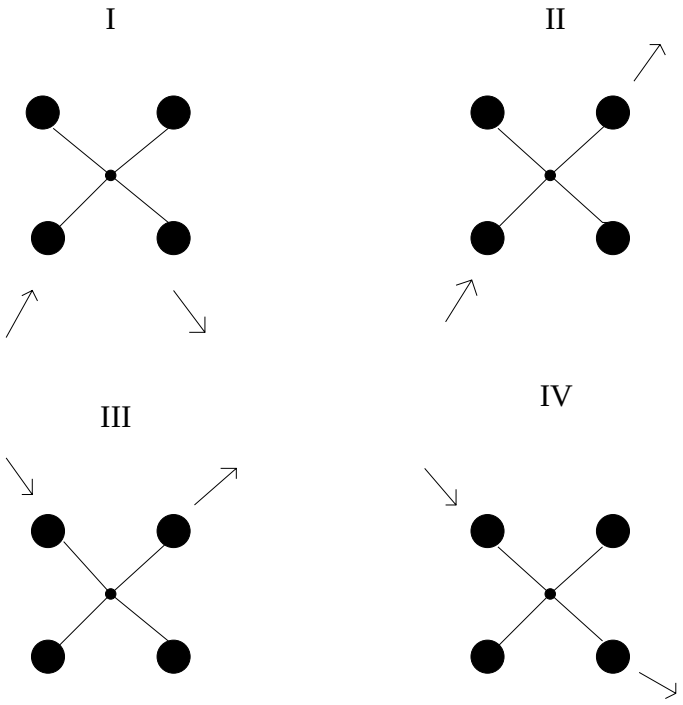

Figure 3: The four basic linking patterns for constructing a regular genealogy which grows linearly from the simple units in Figure 2 where the linking set comprises a single individual and input and output individuals are distinct.

symmetry), and $\mathbf{M_J}$ being the appropriate 2×2 matrix for linking pattern J. The four matrices $\mathbf{M_J}$ are:

$$\mathbf{M_I} = \begin{pmatrix} 5 & 3 \\ 6 & 7 \end{pmatrix}, \mathbf{M_{II}} = \begin{pmatrix} 3 & 3 \\ 10 & 7 \end{pmatrix},$$

and

$$\mathbf{M_{III}} = \begin{pmatrix} 2 & 4 \\ 8 & 9 \end{pmatrix}, \mathbf{M_{IV}} = \begin{pmatrix} 3 & 5 \\ 6 & 7 \end{pmatrix}.$$

As a simple example, we consider a genealogy constructed using the linking pattern of type III in which we take the parents A_i and B_i as the input and output individuals, repectively. By defining $\mathcal{G}_0$ to consist of the single individual, A_0, and $\mathcal{G}_1$ to be the nuclear family of Figure 2, we can construct a genealogy $\mathcal{G}_r$ by identifying individuals B_i and A_{i+1} for $i = 0, 1, 2, \ldots r-1$. The resulting structure is a marriage chain and is shown in Figure 4 for the case when $r = 4$. Thus, as per Camp et al. (1994), we have

$$\mathbf{N_{B_i}} \;=\; \mathbf{M_{III}} \mathbf{N_{B_{i-1}}} \tag{3-1}$$

where

$$\mathbf{N_{B_i}} = (N_{B_i}(1), N_{B_i}(2))^T.$$

The eigenvalues of $\mathbf{M_{III}}$ are $\frac{11 \pm \sqrt{177}}{2}$ so the rate at which the system converges is governed by the larger of these. Hence, if G_r denotes the set of feasible genotypic

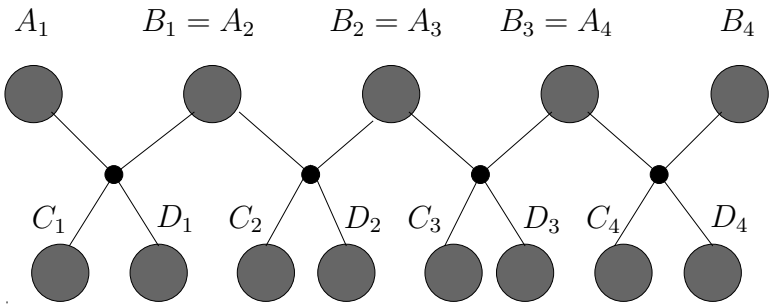

Figure 4: The regular genealogy $\mathcal{G}_4$ which identifies individuals A_i with B_{i+1} for $i < 4$ from the nuclear families of Figure 2.

configurations on a genealogy $\mathcal{G}_r$, then when $\mathcal{G}_r$ is of the type depicted in Figure 4, the addition of another nuclear family to the genealogy will cause the total number of genotypic states

$$|G_r| = 2N_{B_r}(1) + N_{B_r}(2)$$

to increase by a factor

$$\lambda_r = |G_r|/|G_{r-1}|$$

which has $\lambda = \frac{11+\sqrt{177}}{2} \approx 12.15$ as asymptotic value. Since our genealogy grows by 3 individuals with each addition, the average asymptotic increase per individual is given by

$$\gamma = \sqrt[3]{\lambda} \approx 2.30$$

(Camp et al. 1994).

3.2 Maximal Linear Systems

We now prove that the maximal rate of growth of the size of the statespace, using the four patterns of addition defined above (Figure 3), occurs when pattern III is used exclusively. Thus, we prove that $\mathbf{M_{III}}^r$ grows most rapidly amongst products of r matrices selected from the set $\{\mathbf{M_I}, \mathbf{M_{II}}, \mathbf{M_{III}}, \mathbf{M_{IV}}\}$. We shall do this by proving that $\mathbf{M_{III}}^r$ has larger elements than any other sequence.

Definition 1 *A matrix* $\mathbf{A}$ *is said to* dominate *a matrix* $\mathbf{B}$ *if the difference matrix,* $\mathbf{A} - \mathbf{B}$ *is positive*

$$\textit{i.e. } a_{ij} \geq b_{ij} \; \forall \; i, j \textit{ and } a_{kl} > b_{kl} \textit{ for at least some } k, l.$$

Now consider the product of pairs of matrices, $\mathbf{M_i M_j}$. Some of these have no meaning in the present context. For example, for $\mathbf{M_I M_{IV}}$ to occur would require that an individual have two sets of parents. Those pairs which do make sense are displayed in Table 1.

Table 1: Paired matrices $\mathbf{M_i M_j}$ for $i, j \in \{I, II, III, IV\}$ describing the growth of the genotypic statespace for linear genealogies constructed from the basic units in Figure 2, when a type j link is followed by a type i link, for a genetic system with two alleles. The blanks indicate genealogically impossible pairings.

$\mathbf{M_i M_j}$	I	II	III	IV
I		$\begin{pmatrix} 45 & 36 \\ 88 & 67 \end{pmatrix}$	$\begin{pmatrix} 34 & 47 \\ 68 & 87 \end{pmatrix}$	
II		$\begin{pmatrix} 39 & 30 \\ 100 & 79 \end{pmatrix}$	$\begin{pmatrix} 30 & 39 \\ 76 & 103 \end{pmatrix}$	
III	$\begin{pmatrix} 34 & 34 \\ 94 & 87 \end{pmatrix}$	$\begin{pmatrix} 46 & 34 \\ 114 & 87 \end{pmatrix}$	$\begin{pmatrix} 36 & 44 \\ 88 & 113 \end{pmatrix}$	$\begin{pmatrix} 30 & 38 \\ 78 & 103 \end{pmatrix}$
IV	$\begin{pmatrix} 45 & 44 \\ 72 & 67 \end{pmatrix}$	$\begin{pmatrix} 59 & 44 \\ 88 & 67 \end{pmatrix}$	$\begin{pmatrix} 46 & 57 \\ 68 & 87 \end{pmatrix}$	$\begin{pmatrix} 39 & 50 \\ 60 & 79 \end{pmatrix}$

(i = row, j = column)

We immediately have the following dominance patterns amongst the paired matrices of Table 1:

$$\begin{aligned} \mathbf{M_{III}M_{II}} > \mathbf{M_{II}M_{II}} \quad &\text{and} \quad \mathbf{M_{III}M_{II}} > \mathbf{M_{III}M_{I}} \\ \mathbf{M_{IV}M_{II}} > \mathbf{M_{I}M_{II}} \quad &\text{and} \quad \mathbf{M_{IV}M_{II}} > \mathbf{M_{IV}M_{I}} \\ \mathbf{M_{III}M_{III}} > \mathbf{M_{II}M_{III}} \quad &\text{and} \quad \mathbf{M_{III}M_{III}} > \mathbf{M_{III}M_{IV}} \\ \mathbf{M_{IV}M_{III}} > \mathbf{M_{I}M_{III}} \quad &\text{and} \quad \mathbf{M_{IV}M_{III}} > \mathbf{M_{IV}M_{IV}} \end{aligned} \tag{3-2}$$

From these dominance patterns, we can see that there are four potentially maximal such genealogies with corresponding matrices:

$$\mathbf{M} \in \{\mathbf{M_{III}M_{II}}, \mathbf{M_{IV}M_{II}}, \mathbf{M_{III}M_{III}}, \mathbf{M_{IV}M_{III}}\}.$$

In like manner, we can continue to pre and post-multiply these four "maximal" matrix products with $\{\mathbf{M_I}, \mathbf{M_{II}}, \mathbf{M_{III}}, \mathbf{M_{IV}}\}$. From the dominance rules already established for matrix pairings in (3-2), we arrive at the following maximising triples for products of three linking patterns

$$\left.\begin{matrix} \mathbf{M_{III}} \\ \mathbf{M_{IV}} \end{matrix}\right\} \; \mathbf{M_{III}} \; \left\{\begin{matrix} \mathbf{M_{II}} \\ \mathbf{M_{III}} \end{matrix}\right.$$

and, from this, it is straightforward to conclude that the linear genealogy, $\mathcal{G}_r$, constructed from the basic units in Figure 2 with the maximal number of genotypic

states can always be expresssed in the form:

$$\mathbf{N_{X_r}} = \mathbf{A}\mathbf{N_{X_0}}$$

where

$$\mathbf{A} = \left.\begin{array}{c}\mathbf{M_{III}}\\ \mathbf{M_{IV}}\end{array}\right\} \; (\mathbf{M_{III}})^{r-2} \; \left\{\begin{array}{l}\mathbf{M_{II}}\\ \mathbf{M_{III}}\end{array}\right. \quad \text{for } \mathbf{r} \geq \mathbf{2}. \tag{3-3}$$

Note that each of the four possibilities in (3-3) basically describes the same linear genealogy (a marriage chain), the different expressions indicating their different initial input and final output individuals. Thus the maximal genealogy has a state space whose size grows as in the example of Figure 4 (i.e. at the rate λ^r where $\lambda \approx 2.30$).

3.3 Input and output individuals are identical

When we allow the input and output individuals to be one and the same, we can add two more linking patterns as illustrated in Figure 5.

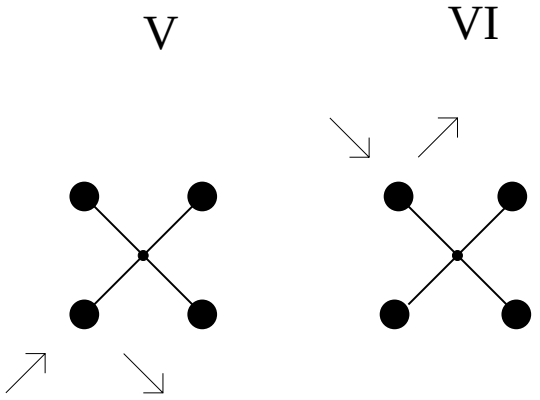

Figure 5: Linking patterns for constructing a regular genealogy where the linking set comprises a single individual who is both input and output.

The addition of a new unit to the existing structure using a linking pattern of type V creates parents and a sibling for some founder individual, whereas a link of type VI causes a spouse and offspring to be assigned to a member of the current genealogy. Clearly, we are restricted in the extent to which we can use a type V link as pedigree members are constrained to have only one set of parents but type VI links can be used repeatedly. Again, the genotypic statespace grows linearly with the addition of each new unit according to the linking patterns in Figure 5 with corresponding matrices

$$\mathbf{M_V} = \begin{pmatrix} 8 & 0 \\ 0 & 13 \end{pmatrix}$$

and

$$\mathbf{M_{VI}} = \begin{pmatrix} 6 & 0 \\ 0 & 17 \end{pmatrix}.$$

3.4 Maximal Statespace with M_J for $J = I, \ldots, VI$

It is clear from the above that a single individual with r spouses (and two offspring in each marriage) will yield a genealogy of the order of 17^r states. The average asymptotic increase per individual for such a structure is given by

$$\gamma = \sqrt[3]{17} \approx 2.57$$

which is larger than the maximising value of 2.30 obtained for the marriage chain structure of Section 3.2. We shall prove that this is in fact the maximal system. The previous approach using matrix domination does not work: it is clear that the diagonal structure of the $\mathbf{M_{VI}}^r$ matrix makes it impossible for this matrix to dominate, or indeed be dominated by, any other matrix product. Our proof depends on exploiting the detailed properties of the relevant $\mathbf{M}$ matrices and is given in Appendix A. Analogously to the earlier result, we can show that if an unrestricted choice of r units is allowed, then the maximising pattern is

$$\mathbf{N_{X_r}} = \mathbf{A}\mathbf{N_{X_0}}$$

where

$$\mathbf{A} = \left.\begin{matrix}\mathbf{M_{III}}\\ \mathbf{M_{IV}}\\ \mathbf{M_{VI}}\end{matrix}\right\} \; (\mathbf{M_{VI}})^{r-2} \; \left\{\begin{matrix}\mathbf{M_{II}}\\ \mathbf{M_{III}}\\ \mathbf{M_{VI}}\end{matrix}\right. \quad \text{for } \mathbf{r \geq 2}. \tag{3-4}$$

Just as before, this represents only one situation: that of one individual with r spouses.

Repeating links of type VI would simulate the growth of some domestic animal pedigrees, for example, where artificial insemination allows a single sire to have thousands of dams. In real life, of course, these animal pedigrees are usually complicated by the presence of inbreeding loops. It would be of interest to consider classes of genealogy with some restrictions imposed. Suppose, for example, that no individual can have more that s spouses. We might reasonably expect that the maximal structure would have u nuclear families connected by type III links with each founder having additional families up to s in total, but we shall not explore this question further here.

4 Multiple offspring and multiple alleles

A similar iterative approach can be taken when the genealogy is constructed from basic units which are not constrained to have two offspring and when the genetic system of interest is not necessarily diallelic. We consider nuclear families with k offspring, as shown in Figure 6 and a genetic system with l alleles, $a_1, a_2, \ldots, a_l$. The four linking patterns of Figure 3 are entirely general in that a link of type I now means that one of the k offspring is chosen as the input individual and another is designated as the output, and so on. Likewise, we can generalise the two linking patterns of Figure 5 relating specifically to the case where the input and output individuals are identical. There are l homozygous genotypes and

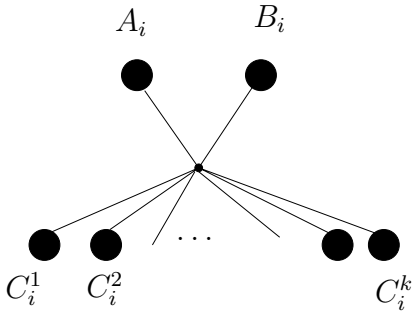

Figure 6: The basic building block consisting of a nuclear family with k offspring.

$\frac{l(l-1)}{2}$ heterozygous genotypes to be accounted for, but the fact that we are not considering any constraints from phenotypic data on the pedigree allows us to simplify the calculations enormously. For any pivot individual X, we have

$$N_X(a_s a_s) = N_X(a_t a_t) \ \forall \ s,t \tag{4-1}$$

and

$$N_X(a_s a_t) = N_X(a_u a_v) \ \forall \ s,t,u,v \text{ where } s \neq t, \ u \neq v$$

by symmetry.

Thus, by labelling a generic homozygous genotype as 1 and a heterozygous genotype as 2, we can express the growth in the genotypic statespace for a pedigree $\mathcal{G}_i$, constructed from $\mathcal{G}_{i-1}$ by the addition of a new unit, as a simple linear system exactly as before:

$$\mathbf{N_{X_i}} = \mathcal{M}\mathbf{N_{X_{i-1}}}$$

where

$$\mathbf{N_{X_i}} = (N_{X_i}(1), N_{X_i}(2))^T$$

and $\mathcal{M}$ is the matrix corresponding to the manner in whuch the new block was linked to $\mathcal{G}_{i-1}$. The six matrices, $\mathcal{M}$, corresponding to the four linking patterns of Figure 3 and the two patterns of Figure 5, are functions of both the number of alleles in the system, l, and the number of offspring, k, in the nuclear family units and are given in these general terms in (4-2) below.

$$\mathcal{M}_{\mathbf{I}} =$$

$$\begin{pmatrix} 1+(l-1)[2^{k-1}+2.3^{k-2}+4^{k-2}(l-2)] & (l-1)[2^{k-1}+3^{k-2}+3.4^{k-2}(l-2)] \\ 2[2^{k-1}+3^{k-2}+3.4^{k-2}(l-2)] & 2+2^k(2l-3)+3^{k-2}+2.4^{k-2}(l-2)(4l-3) \end{pmatrix}$$

$$\begin{aligned} \mathcal{M}_{\mathbf{II}} &= \begin{pmatrix} 2^{k-1}(l-1)+1 & (l-1)[2^{k-1}(l-1)+1] \\ 2[2^{k-1}+3^{k-1}+4^{k-1}(l-2)] & 2^k(l-1)+3^{k-1}+2l4^{k-1}(l-2) \end{pmatrix} \\ \mathcal{M}_{\mathbf{III}} &= \begin{pmatrix} l & l(l-1)2^{k-1} \\ l2^k & 3^k+2.4^{k-1}(l+1)(l-2) \end{pmatrix} \end{aligned} \tag{4-2}$$

$$\mathcal{M}_{\mathbf{IV}} = \begin{pmatrix} 2^{k-1}(l-1)+1 & (l-1)[2^{k-1}+3^{k-1}+4^{k-1}(l-2)] \\ 2[2^{k-1}(l-1)+1] & 2^k(l-1)+3^{k-1}+2l4^{k-1}(l-2) \end{pmatrix}$$

$$\mathcal{M}_{\mathbf{V}} = \begin{pmatrix} 1+(l-1)[2^k+3^{k-1}+(l-2)4^{k-1}] & 0 \\ 0 & 2+(l-1)2^{k+1}+3^{k-1}+2l(l-2)4^{k-1} \end{pmatrix}$$

$$\mathcal{M}_{\mathbf{VI}} = \begin{pmatrix} l+\frac{l(l-1)}{2}2^k & 0 \\ 0 & l2^k+3^k+[\frac{l(l-1)}{2}-1]4^k \end{pmatrix}$$

The total number of states for a linear genealogy $\mathcal{G}_r$ constructed recursively from the units in Figure 6 is then given by

$$|G_r| = lN_{X_r}(1) + \frac{l(l-1)}{2}N_{X_r}(2)$$

for a final output individual X_r.

In order to get some intuition on how the size of the genotypic statespace varies with the number of offspring in the basic building block (Fig 6) and the number of alleles in the genetic system under consideration, we return to a simple structure introduced by Camp et al. (1994). We construct the genealogy $\mathcal{G}_3$ of Figure 7 by linking three units together with a type II linking scheme, taking the parent A_i as the output and the offspring C_i^k say, as the input individuals. Thus, the final output A_3, will be our reference individual and we count the number of states by adding up our R-functions on A_3. Table 2 displays this number for up

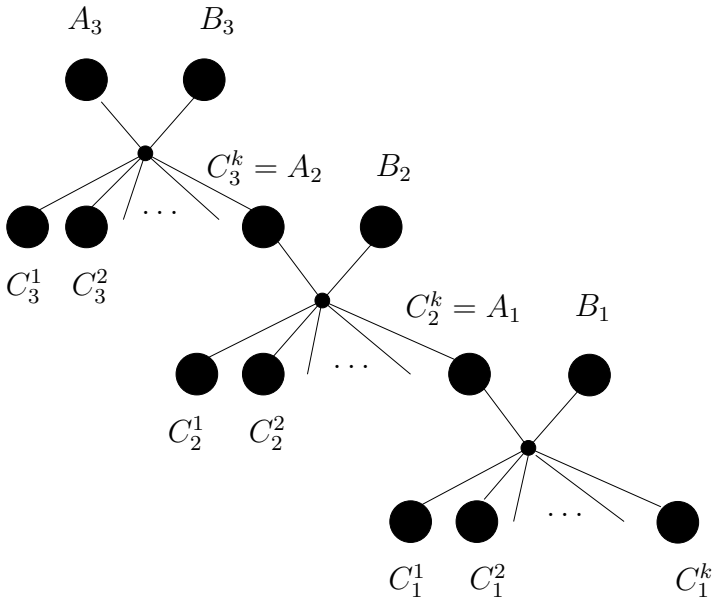

Figure 7: The genealogy $\mathcal{G}_3$, defined by a type II linking scheme, which identifies individuals A_i with C_i^k for $i < 3$ from the basic units of Figure 6.

Table 2: The size of the genotypic state space for the linear genealogy $\mathcal{G}_3$ of Figure 7 for different numbers of offspring, k, in the basic unit and different numbers of alleles, l, in the genetic system. Note that n_3 is the total number of individuals in the genealogy, g is the number of genotypes in the system and $\hat{\gamma} = \sqrt[n_3]{|G_3|}$.

k	n_3	$l =$ $g =$	2 3	3 6	4 10	5 15	6 21
1	7	$N_{A_3}(1)$	168	1833	11836	48745	152706
		$N_{A_3}(2)$	284	3425	22831	95329	300791
		$\|G_3\|$	620	15774	184330	1197015	5428101
		$\hat{\gamma}$	2.501	3.978	5.652	7.384	9.164
		γ	2.318	3.769	5.210	6.642	8.068
2	10	$N_{A_3}(1)$	744	24275	206290	980469	3354956
		$N_{A_3}(2)$	1943	85265	775243	3555203	13099967
		$\|G_3\|$	3431	328620	5476618	40454375	216629241
		$\hat{\gamma}$	2.257	3.562	4.72	5.764	6.82
		γ	2.2125	3.423	4.413	5.287	6.09
3	13	$\|G_3\|$	36297	10381290	225010518	1975765695	10726589271
		$\hat{\gamma}$	2.243	3.465	4.390	5.187	5.91
		γ	2.211	3.374	4.162	4.814	5.378
4	16	$\|G_3\|$	477011	437008161	11246140450	106071690875	595242856581
		$\hat{\gamma}$	2.264	3.468	4.248	4.888	5.444
		γ	2.239	3.406	4.073	4.592	5.029

to four offspring and up to six alleles. The estimated contribution per individual is given by the geometric mean, $\hat{\gamma}$, along with the true value as obtained from the eigenvalues of the appropriate linear system. As can be seen from the small number of worked examples summarised in Table 2, the number of alleles in the genetic system under consideration is, by far, the deciding factor. Note that for $l > 2$, the actual value of γ is *much* lower than the crude upper bound of g, the number of distinct genotypes for the trait. For example, there are 10 genotypes for a system with 4 alleles, yet the value of γ is closer to 4 for nuclear families with more than one offspring. We tend to have higher values of γ in the latter case because of the high ratio of founders to non-founders in the genealogy constructed above.

The matrices in (4-2) have several interrelationships, some of which have been exploited in Appendix A. This is essentially because their elements are obtained by accumulating the number of genotypic states over the possible families in different ways. If we consider an isolated nuclear family of the type depicted in Figure 6 for example, the total number of genotypic possibilities is fixed for this block, however we choose to accumulate the information. Allowing for the multiplicities of homozygotes and heterozygotes noted in Equation (4-1) above, we can write

$$\begin{aligned}(l, \frac{l(l-1)}{2})\mathcal{M}_{\mathbf{X}} \begin{pmatrix} 1 \\ 1 \end{pmatrix} &= l^2 + 2^k[l^2(l-1)] + 3^k[\frac{l(l-1)}{2}] \\ &+ 4^{k-1}[(l+1)l(l-1)(l-2)]\end{aligned}$$

for any matrix $\mathcal{M}_{\mathbf{X}}$ with $X \in \{I, II, III, IV, V, VI\}$. (See Appendix B for the block calculation.) The row sums of the three matrices $\mathcal{M}_{\mathbf{I}}, \mathcal{M}_{\mathbf{IV}}$ and $\mathcal{M}_{\mathbf{V}}$ are identical as are the row sums of $\mathcal{M}_{\mathbf{II}}, \mathcal{M}_{\mathbf{III}}$ and $\mathcal{M}_{\mathbf{VI}}$. In the first case, this is because the output individual (i.e. the pivot onto whom all information is collapsed) is always a parent from the nuclear family most recently added, whereas in the second case, the ouput individual is an offspring from this most recent block. Furthermore, we have relationships between specific elements of these matrices. For any matrix $\mathcal{M}$, the off-diagonal element $(\mathcal{M})_{12}$ represents the number of states on the basic unit corresponding to the case where the output individual is homozygous and the input heterozygous, whereas $(\mathcal{M})_{21}$ represents the reverse situation with heterozygous output and homozygous input. The diagonal elements $(\mathcal{M})_{11}$ and $(\mathcal{M})_{22}$ give the number of states when both pivots are homozygous and heterozygous, respectively. In particular, among the off-diagonal elements, we note that

$$l(\mathcal{M}_{\mathbf{X}})_{12} = \frac{l(l-1)}{2}(\mathcal{M}_{\mathbf{Y}})_{21}$$

for pairings $(X, Y) = (I, I), (III, III), (II, IV)$ and (IV, II). This is due to symmetry between the two offspring in the block in the first instance, symmetry between both parents in the second, and reversabilty of linking directions in each of the last two. In addition, the diagonal elements of $\mathcal{M}_{\mathbf{II}}$ and $\mathcal{M}_{\mathbf{IV}}$ are identical, again because of reversabilty of the linking direction.

We have not attempted to generalise the arguments of Section 3 regarding sequences of units and maximisation for a genetic system with l alleles and nuclear units with k offspring, as this would lead to considerable complications. We note that for the case when input and output individuals are distinct, Camp (1995) has shown numerically that the dominance patterns of (3-2) in Section 3.1 also hold for $\mathcal{M}_{\mathbf{I}}, \ldots \mathcal{M}_{\mathbf{IV}}$ for $l = 2, \ldots, 30$ and $k = 2, \ldots, 12$ and hence the marriage chain structure in Equation (3-3) remains maximal for these values. We hypothesise that if links of types V and VI are also considered, allowing the input and output individuals to be one and the same, the result of Section 3.3 still holds and an individual with multiple marriages will provide the maximising structure. What is certainly true is that $\mathcal{M}_{\mathbf{VI}}$ has the largest eigenvalue, since, for a non-negative matrix, the largest eigenvalue is less than or equal to the largest row sum (Mirsky (1955), Theorem 7.5.4, for example) and, from our earlier comment on row sums, it is clear that the largest eigenvalue is precisely

$$l2^k + 3^k + [\frac{l(l-1)}{2} - 1]4^k,$$

the (2,2) element of $\mathcal{M}_{\mathbf{VI}}$. This suggests that the results of Section 3 might generalise with the number of alleles in the system, l and the number, k, of offspring in the nuclear family block.

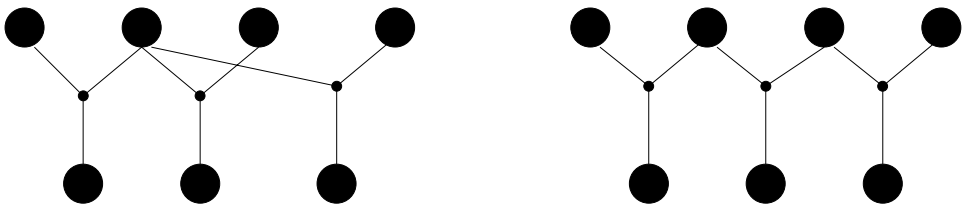

Figure 8: A linear genealogy, (a), constructed from linking nuclear families with a single offspring at a linking set of size 1 which has a larger genotypic statespace than the marriage chain (b). The number of states is γ^7 where for $l = 2$, (a) has $\gamma = 2.41$ and (b) has $\gamma = 2.38$ and for $l = 3$, (a) has $\gamma = 4.02$ and (b) has $\gamma = 3.98$.

As a very simple illustration, Figure 8 depicts the type of structure which we suspect will always have the maximal number of feasible genotypic configurations of all linear genealogies cunstructed from the basic units in Figure 6 via linking sets of order 1, for the case $k = 1$. Both structures displayed in the figure are constructed from the same number of basic nuclear family units but the structure on the left has a type VI link replacing one of the type III links of the straight

marriage chain shown on the right and thus has a larger genotypic statespace. In summary, it would seem that the marraige chain structure can always be improved upon when we allow input and output individuals to be identical.

5 Closed Linear Systems

A linear genealogy, as defined in Section 2.2, is said to be *closed* if there are no new founders added to the genealogy after the initial basic unit or units. Many regular mating systems (Wright 1921) such as repeated parent-offspring, repeated sibling, repeated double first cousins and repeated quadruple half first cousin matings, for example, would be classified as closed linear genealogies.

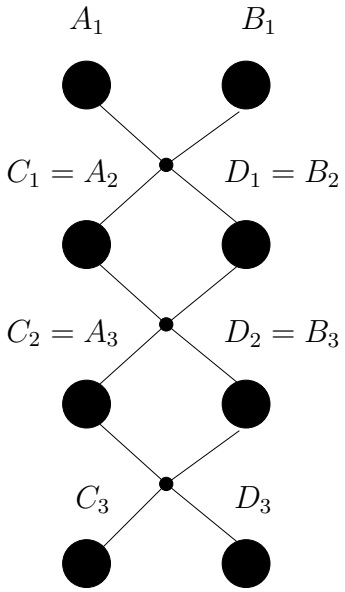

Figure 9: A repeated sib-pair mating depicted, in our usual notation, as the genealogy $\mathcal{G}_3$ which identifies individuals C_i and D_i with A_{i+1} and B_{i+1}, respectively, for $i < 3$.

If we restrict attention to the basic building block of Figure 2, consisting of a nuclear family with two offspring, we can construct the repeated sib-pair mating system of Figure 9 by identifying the offspring of the last unit to be attached with the parents of the next. Now we have a linking set of order 2 and for each nuclear family, the input set which forms the connection with the existing structure comprises the parents, while the output set which will connect with the next block, consists of the two offspring. In the case of systems like repeated double first cousin matings, for example, we have a slightly different situation as is shown in Figure 10. Here we begin with *two* basic units containing all the founders of the resulting genealogy, we add two units at each stage and the required linking set is now of order 4. Furthermore, this is a *disjoint* linking set in the sense that the linking individuals do not all belong to the same basic unit. Nevertheless, all

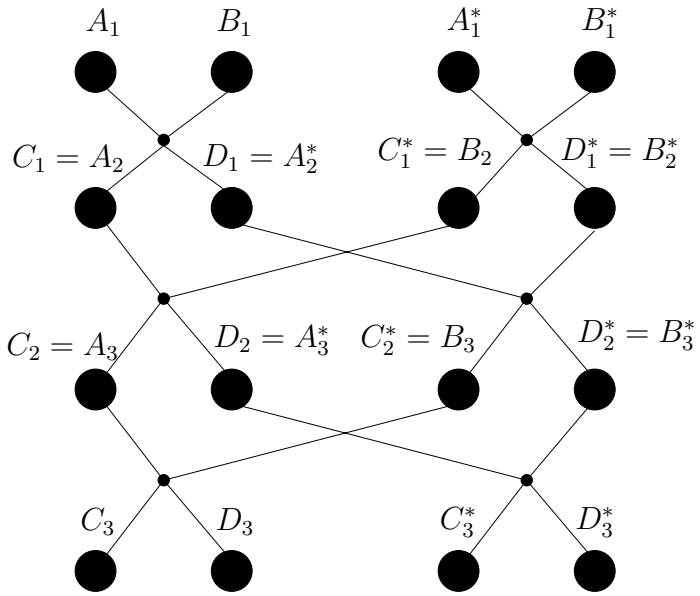

Figure 10: A repeated double first cousin mating system represented as the regular genealogy $\mathcal{G}_3$. The genealogy grows with the addition of two basic units at each stage. At the i^{th} step, we identify one offspring from each block, C_i and C_i^* with the first new set of parents A_{i+1} and B_{i+1} and the remaining pair of offspring D_i and D_i^* with the second parental pairing, A_{i+1}^*, B_{i+1}^*, for $i < 3$. (Note that inbreeding only begins in generation 4.)

these structures have a genotypic statespace which grows as a linear system with the addition of each new unit, or units, just as before. Thus, we can still write

$$\mathbf{N_{S_i}} \quad = \quad \mathbf{M}\mathbf{N_{S_{i-1}}} \tag{5-1}$$

with the distinction now being that the elements of $\mathbf{N_{S_i}}$ relate to the numbers of genotypic configurations corresponding to each of the possible *identity states* (Thompson 1974) on the linking set S after i steps in the construction process. Describing this system exactly, gets complicated when disjoint linking sets are involved, but we can write down the general form of the matrix $\mathbf{M}$ for a closed linear system constructed from nuclear families with two offspring with an arbitrary linking set S and for a genetic system with l alleles.

For any linking set S, suppose we have J_S identity states which are distinct up to permutations among the different alleles of the system. Let $a(j)$ represent the number of distinct alleles in state j, $j \in J_S$. Without loss of generality, we can label the distinct identity states in increasing order with the number of alleles required so that

$$a(j) \geq a(k) \text{ for } j > k,\ j, k \in J_S.$$

The matrix $\mathbf{M} = (m_{rs}),\ r, s \in J_S$ of Equation (5-1), describing the growth of the genotypic statespace for the associated genealogy after a single growing step,

takes the general form:

$$m_{rs} = \begin{cases} 0 & \text{if } a(r) > a(s) \\ \text{independent of } l & \text{if } a(r) = a(s) \\ \text{constant} \times \begin{pmatrix} l - a(r) \\ a(s) - a(r) \end{pmatrix} & \text{if } a(r) < a(s). \end{cases}$$

We recall that an individual element, m_{rs}, of the matrix $\mathbf{M}$ can be interpreted as the contribution to the new count of genotypic configurations on the genealogy corresponding to the output individuals being in identity state r and the input individuals in identity state s. Thus, the reasoning for each of these cases is as follows:

$a(r) > a(s)$: No state of type r on the output set can arise from a state of type s on the input individuals since more alleles are required for state r than are currently available.

$a(r) = a(s)$: The two identity states require the same number of alleles so all alleles in states s are used up in state r.

$a(r) < a(s)$: State r requires less alleles than state s so the remaining $a(s) - a(r)$ alleles can be selected from the unused $l - a(r)$ alleles in the system.

Explicitly, we consider the repeated sib-pair mating structure illustrated in Figure 9 for a genetic system with 4 or more alleles. As the linking set S comprises two individuals here, we have $J_S = 7$ identity states for this example. These identity states are listed in Table 3 in increasing order with increasing number of different alleles involved. The third column of the table shows the number of possible states corresponding to each identity state obtained by permuting amongst the alleles of the genetic system. Thus for instance, state 1 (11,11) simply means that all four alleles of the two linking individuals are the same and hence there are l possible choices for state 1. Similarly, state 2 (11,12) requires two alleles which can be chosen in $\frac{l(l-1)}{2}$ ways. Either of these two alleles can be the allele for the homozygous genotype and either individual can be the homozygote. Hence there are $4\frac{l(l-1)}{2}$ possible genotypic configurations on the pair of individuals in S which are represented by this identity state. The final column in Table 3 gives the number of different offspring genotypes which can result from the mating of the input individuals in any identity state. Classifying the pairwise combinations of these offspring genotypes by the distinct identity states allows us to specify how information is collated from the input to the output individuals with the addition of a new nuclear family to the existing structure. The growth of the genealogy is given by

$$\mathbf{N_{S_i}} = \mathbf{M}\mathbf{N_{S_{i-1}}}$$

Table 3: The seven distinct identity states on the linking set for the repeated sib-pair mating of Figure 9. For each of these, we also give the total number of states for which it codes and the number of possible offspring genotypes. Note that the overall number of states is $\frac{l(l-1)}{2}^2$—the number of genotypes for two individuals and an l-allele system.

$j \in J_S$	State	# Possibilities	# Offspring Genotypes
1	$(11, 11)$	l	1
2	$(11, 12)$	$4\binom{l}{2}$	2
3	$(11, 22)$	$2\binom{l}{2}$	1
4	$(12, 12)$	$\binom{l}{2}$	3
5	$(12, 13)$	$6\binom{l}{3}$	4
6	$(11, 23)$	$6\binom{l}{3}$	2
7	$(12, 34)$	$6\binom{l}{4}$	4

where

$$\mathbf{M} = \begin{pmatrix} 1 & 2(l-1) & 0 & l-1 & 2\binom{l-1}{2} & 0 & 0 \\ 0 & 2 & 0 & 1 & 2(l-2) & 0 & 0 \\ 0 & 0 & 0 & 1 & 0 & 0 & 0 \\ 0 & 4 & 2 & 1 & 6(l-2) & 4(l-2) & 4\binom{l-2}{2} \\ 0 & 0 & 0 & 0 & 6 & 2 & 2(l-3) \\ 0 & 0 & 0 & 0 & 2 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 & 0 & 4 \end{pmatrix}. \qquad (5\text{-}2)$$

For example, if the set S is in state 4 (12,12) at time $i-1$, the only possible identity states at time i are 1 (which can occur in $(l-1)$ ways corresponding to the choices for the unused allele), and $2, 3$ and 4 which can each only happen in one way as they all require both alleles. Taking the appropriate multiplicities from Table 3, we get a total contribution to the overall count of

$$l(l-1) + 7\binom{l}{2} = 9\binom{l}{2}.$$

We can verify this intuitively since each of the possibilities for state 4 on the parent individuals can give rise to $3^2 = 9$ genotypic combinations on the offspring pair. A similar argument for the other states shows that $\mathbf{M}$ should satisfy

$$(l, 4\binom{l}{2}, 2\binom{l}{2}, \binom{l}{2}, 6\binom{l}{3}, 6\binom{l}{3}, 6\binom{l}{4})\,\mathbf{M} =$$
$$(l, 16\binom{l}{2}, 2\binom{l}{2}, 9\binom{l}{2}, 96\binom{l}{3}, 24\binom{l}{3}, 96\binom{l}{4}).$$

Note that for the case $l = 2$, only the first four identity states of Table 3 are required and the appropriate M-matrix is given by the top 4×4 submatrix of $\mathbf{M}$ in (5-2). Similiarly, when the system has 3 alleles, the first six identity states of Table 3 apply and the top left 6×6 submatrix in (5-2) is the relevant M-matrix.

By contrast with the results in Section 4, the eigenvalues of $\mathbf{M}$ in this example are not functionally related to the number of alleles, l, in the system under consideration. It is straightforward to show that the largest eigenvalue is $\lambda \approx 3.7785$ when $l = 2$ and $\lambda \approx 6.60$ for $l \geq 3$. Since two individuals are added to the genealogy at each stage, the average asymptotic increase per individual is given by $\gamma \approx 1.944$ and $\gamma \approx 2.569$, respectively.

6 Discussion

In this paper, we have investigated a class of regularly constructed genealogies whose statespace increases linearly from one growth stage to the next. We began in Section 3 with genealogies which grow with the addition of a single nuclear

family unit of two parents and two offspring. For a two-allele genetic system with Mendelian inheritance, we have shown that when the linking set connecting the new unit to the existing structure is of order 1, a maximal genotypic statespace is obtained when a single individual has multiple marriages. Extending this statement of maximality to the case of multiple offspring in the basic unit and multiple alleles in the genetic system did not prove straightforward, but some numerical experimentation (Camp 1995) has indicated that the result may well generalise. Consequently, we can say that a genealogy of individuals with multiple marriages has a far bigger genotypic statespace than, say, a monogamous structure with the same number of individuals.

Once the linking set comprises two or more individuals, exact treatment of the linear system in question becomes very complicated. In Section 5, however, we have outlined the general form of the linear system in this situation and given the exact asymptotic value of the size of the genotypic statespace as a function of the number of individuals in the genealogy for the repeated sib-pair mating of Wright (1921). A full treatment of closed linear systems, in particular with view to finding a maximal structure analogous to that of Section 3, is at best difficult and beyond the scope of this paper.

7 Acknowledgements

We wish to acknowledge Wellcome Fellowship 036281/Z/92/Z for Nuala Sheehan and Arthritis and Rheumatism Council Grant D0528 for Nicki Camp.

References

Camp, N., Cannings, C. & Sheehan, N. (1994), The number of genotypic assignments on a genealogy I. The method and simple examples, *IMA Journal of Mathematics Applied in Medicine and Biology* **11**, 95–106.

Camp, N. J. (1995), Methods for genealogical analysis, with particular reference to Type II diabetes in the Pima, PhD thesis, University of Sheffield.

Cannings, C., Thompson, E. A. & Skolnick, M. H. (1978), Probability functions on complex pedigrees, *Advances in Applied Probability* **10**, 26–61.

Mirsky, L. (1955), *An Introduction to Linear Algebra*, Clarendon Press.

Thompson, E. A. (1974), Gene identities and multiple relationships., *Biometrics* **30**, 667–680.

Thompson, E. A. (1986), *Pedigree Analysis in Human Genetics*, The Johns Hopkins University Press, Baltimore.

Wright, S. (1921), Systems of mating, *Genetics* **6**, 111–178.

A Maximising the genotypic statespace for a linear genealogy for the case with two offspring and two alleles

To each of the matrices $\mathbf{M} = (m_{ij}) \in \{\mathbf{M_I}, \ldots, \mathbf{M_{VI}}\}$, we associate a 3×3 matrix, $\mathbf{V} = (v_{ij})$ defined by:

$$\begin{aligned} v_{11} = v_{33} &= m_{11} - u, \\ v_{12} = v_{32} &= m_{12}, \\ v_{21} = v_{23} &= m_{21}/2, \\ v_{22} &= m_{22}, \\ v_{13} = v_{31} &= u \end{aligned}$$

where

$$u = 0 \text{ for } \mathbf{M} \in \{\mathbf{M_{II}}, \mathbf{M_{IV}}, \mathbf{M_V}, \mathbf{M_{VI}}\} \text{ and } u = 1 \text{ for } \mathbf{M} \in \{\mathbf{M_I}, \mathbf{M_{III}}\}.$$

Now the $\mathbf{V}$ matrices are the appropriate matrices if the homozygous states *aa* and *bb* (coded as 1 and 3. respectively in Section 3) are treated separately in the recursions. If $\{\mathbf{V_i}\}$ is any sequence of such matrices, $i \in \{I, \ldots, VI\}$, then the total number of states for a linear genealogy constructed from the basic units in Figure 2 via the inferred linking sequence, is given by

$$(1,1,1) \prod_i \mathbf{V_i} \begin{pmatrix} 1 \\ 1 \\ 1 \end{pmatrix} \;=\; (2,1) \prod_i \mathbf{M_i} \begin{pmatrix} 1 \\ 1 \end{pmatrix} \tag{A-1}$$

and we prove that this is maximised when $i = \text{VI}, \ \forall i$.

We begin by defining a set of 3×3 matrices $\mathcal{U} = \{\mathbf{U} = (u_{ij}),\ i, j = 1, 2, 3\}$ where:

1. $u_{ij} \geq 0 \ \forall i, j$
2. $u_{11} = u_{33},\ u_{12} = u_{32},\ u_{21} = u_{23}$ and $u_{13} = u_{31}$
3. $u_{22} > u_{21}$ and $u_{22} > u_{12}$
4. $\sum_i u_{i2} > \sum_i u_{i1}$ and $\sum_i u_{2i} > \sum_i u_{1i}$.

Theorem 1 *$\mathcal{U}$ is closed under multiplication.*

Proof
Suppose $\mathbf{A} \in \mathcal{U}$, $\mathbf{B} \in \mathcal{U}$ and consider $\mathbf{C} = \mathbf{AB}$.

1. $c_{ij} = \sum_k a_{ik} b_{kj} \geq 0$.

2. From the equalities defined on elements of $\mathbf{A}$ and $\mathbf{B}$ ($\in \mathcal{U}$), we have that

$$\begin{aligned} c_{11} &= a_{11}b_{11} + a_{12}b_{21} + a_{13}b_{31} \\ &= a_{33}b_{33} + a_{32}b_{23} + a_{31}b_{13} \\ &= c_{33}. \end{aligned}$$

The other equalites are similarly derived.

3. Using $a_{21} = a_{23}$, and $\sum_i b_{i2} > \sum_i b_{i1}$ we can write

$$\begin{aligned} c_{22} &= a_{21}b_{12} + a_{22}b_{22} + a_{23}b_{32} \\ &= a_{21}(b_{12} + b_{22} + b_{32}) + (a_{22} - a_{21})b_{22} \\ &> a_{21}(b_{11} + b_{21} + b_{31}) + (a_{22} - a_{21})b_{21}, \text{ (since } b_{22} > b_{21}) \\ &= c_{21}. \end{aligned}$$

It can be shown that $c_{22} > c_{21}$ in a similar way.

4.

$$\begin{aligned} \sum_i c_{i2} &= \sum_i \{\sum_k a_{ik}b_{k2}\} \\ &= \sum_k b_{k2}(\sum_i a_{ik}) \\ &= \sum_k b_{k2}(\sum_i a_{i1}) + b_{22}(\sum_i a_{i2} - \sum_i a_{i1}) \end{aligned} \tag{A-2}$$

using $\sum_i a_{i1} = \sum_i a_{i3}$.
Similarly, we have that

$$\sum_i c_{i1} = \textstyle\sum_k b_{k1}(\sum_i a_{i1}) + b_{21}(\sum_i a_{i2} - \sum_i a_{i1}).$$

Now, since $\mathbf{A}, \mathbf{B} \in \mathcal{U}$, we have that

$$\sum_k b_{k2} > \sum_k b_{k1},\ \sum_i a_{i2} > \sum_i a_{i1} \text{ and } b_{22} > b_{21}.$$

Substituting into (A-2) above, we have shown that $\sum c_{12} > \sum c_{i1}$, as required.
The proof of the row sum inequality

$$\sum_i c_{2i} > \sum_i c_{1i}$$

follows by a similar argument.

This concludes the proof that $\mathcal{U}$ is closed under multiplication.

We now consider a set of matrices $\mathcal{V} \subset \mathcal{U}$ which take the form:

$$\mathbf{V}(T,\alpha,\beta,x,y) = \begin{pmatrix} \alpha & y-\alpha-\beta & \beta \\ x-\alpha-\beta & T-2x-2y & x-\alpha-\beta \\ & +2\alpha+2\beta & \\ \beta & y-\alpha-\beta & \alpha \end{pmatrix}$$

where T is fixed and $x, y > u$, for some fixed u. In particular, we note that

$$v_{ij} \geq 0 \Rightarrow x, y \geq (\alpha+\beta),\ \alpha \geq 0,\ \beta \geq 0 \text{ and } T \geq 2(x+y-\alpha-\beta).$$

Theorem 2 *For any sequence of matrices* $\{\mathbf{V_i}_{i=i}^{k}\},\ \mathbf{V_i} \in \mathcal{V}$*, the quantity*

$$(1,1,1)\prod_i \mathbf{V_i} \begin{pmatrix} 1 \\ 1 \\ 1 \end{pmatrix}$$

is maximised by taking

$$\mathbf{V_i} = \mathbf{V}(T, u-w, w, u, u) \text{ where } w \in [0,u],\ i = 2,\ldots,k-1,$$

$$\mathbf{V_1} = \mathbf{V}(T,\alpha,\beta,u,y) \text{ and } \mathbf{V_k} = \mathbf{V}(T,\alpha,\beta,x,u).$$

Proof
Consider

$$\begin{aligned} \prod_{i=1}^{k} \mathbf{V_i} &= (\prod_{i=1}^{r-1} \mathbf{V_i})\mathbf{V_r}(\prod_{i=r+1}^{k} \mathbf{V_i}) \\ &= \mathbf{A}\mathbf{V_r}\mathbf{B},\ r = 2,\ldots,k-1 \\ &= \mathbf{AD} \end{aligned}$$

where $\mathbf{A}, \mathbf{B}$ and $\mathbf{D}$ are the appropriate product matrices and $\mathbf{A}, \mathbf{B}, \mathbf{D} \in \mathcal{U}$. For any matrix, $\mathbf{U} = (u_{ij})$, we will denote the row and column sums by

$$\mathbf{U_{i.}} = \sum_j u_{ij} \text{ and } \mathbf{U_{.j}} = \sum_i u_{ij},$$

respectively. Hence we can write

$$\begin{aligned} (1,1,1)\prod_{i=1}^{k} \mathbf{V_i} \begin{pmatrix} 1 \\ 1 \\ 1 \end{pmatrix} &= (1,1,1)\mathbf{AD} \begin{pmatrix} 1 \\ 1 \\ 1 \end{pmatrix} \qquad \text{(A-3)} \\ &= \sum_i \mathbf{A_{.i}}\mathbf{D_{i.}} \\ &= \mathbf{A_{.2}} \sum_i \mathbf{D_{i.}} - 2\mathbf{D_{1.}}(\mathbf{A_{.2}} - \mathbf{A_{.1}}) \end{aligned}$$

recalling that $\mathbf{D_{1.}} = \mathbf{D_{3.}}$ and $\mathbf{A_{.1}} = \mathbf{A_{.3}}$ since $\mathbf{A}, \mathbf{B} \in \mathcal{U}$.
Now, since $\mathbf{D} = \mathbf{V_r B}$, we have that

$$\mathbf{D_{1.}} = y\mathbf{B_{2.}} - (\alpha + \beta)(\mathbf{B_{2.}} - \mathbf{B_{1.}})$$

and

$$\sum_i \mathbf{D_{i.}} = 2x\mathbf{B_{1.}} + (T - 2x)\mathbf{B_{2.}}.$$

Since $(\mathbf{B_{2.}} - \mathbf{B_{1.}}) \geq 0$ and $(\mathbf{A_{.2}} - \mathbf{A_{.1}}) \geq 0$, the expression in (A-3) is maximised by taking x minimal (and thus maximising $\sum_i \mathbf{D_{i.}}$), y minimal and $(\alpha + \beta)$ maximal (i.e. minimising $\mathbf{D_{1.}}$). This is achieved when $x = y = u = (\alpha + \beta)$ for $r = 2, \ldots, k-1$. For the case $r = 1$, we note that

$$(1, 1, 1) \prod_{i=1}^{k} \mathbf{V_i} \begin{pmatrix} 1 \\ 1 \\ 1 \end{pmatrix} = \sum_i \mathbf{D_{i.}}$$

which is maximised for $x = u$. Similarly, when $r = k$, we only require that $y = u$.

Corollary 1 *If we consider our six* $\mathbf{V}$ *matrices, defined at the outset to correspond to the matrices* $\{\mathbf{M_I}, \ldots, \mathbf{M_{VI}}\}$*, then these all belong to the class* $\mathcal{V}$ *with* $T = 29$ *and* $u = 6$*. For each particular* $\mathbf{V_k}$*, we have that*

$$\alpha = 4, \beta = 1, x = 8 = y \text{ for } k = I,$$

$$\alpha = 3, \beta = 0, x = 8, y = 6 \text{ for } k = II,$$

$$\alpha = 1 = \beta, x = y = 6 \text{ for } k = III,$$

$$\alpha = 3, \beta = 0, x = 6, y = 8 \text{ for } k = IV,$$

$$\alpha = 8, \beta = 0, x = y = 8 \text{ for } k = V, \text{ and}$$

$$\alpha = 6, \beta = 0, x = y = 6 \text{ for } k = VI.$$

Hence, for any product of the $\mathbf{V_i}$*s, the quantity (A-3) is maximised by taking* $\mathbf{V_i} = \mathbf{V_{VI}}$, $\forall \mathbf{i}$.

In particular, we conclude that the total number of states for a linear genealogy constructed from the basic units in Figure 2 via the linking sequence implied by the product $\prod_{i=1} \mathbf{V_i}$, is maximal when links of type VI are used throughout.

B The number of states on the basic building block with k offspring for a genetic system with l alleles

Consider the basic unit of Figure 6 reproduced below with the individuals relabelled. Taking individual A as the reference individual, we will generalise the calculation in Section 2.1 to count the number of possible genotypic configurations consistent with Mendelian inheritance on this unit for a genetic system with l alleles. Denoting the alleles of the system as $a_1, a_2, \ldots, a_l$ and using the result

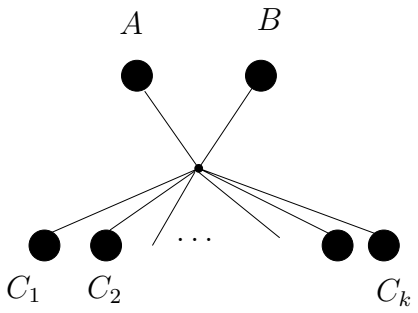

Figure 11: The basic building block consisting of a nuclear family with k offspring.

$$N_X(a_s a_s) = N_X(a_t a_t) \; \forall \; s, t$$

and

$$N_X(a_s a_t) = N_X(a_u a_v) \; \forall \; s, t, u, v \text{ where } s \neq t, \; u \neq v$$

from Equation (4-1) Section 4, it suffices to work out the number of states consistent with one specific homozygous genotype $a_i a_i$ and one specific heterozygous genotype $a_i a_j$ on the reference individual, A.

For the homozygous case, we have

$$\begin{aligned} N_A(a_i a_i) &= 1 \text{ when B is } (a_i a_i) \\ &+ (l-1) \text{ when B is } (a_k a_k) \; , k \neq i \\ &+ (l-1)2^k \text{ when B is } (a_i a_k) \; , k \neq i \\ &+ \frac{(l-1)(l-2)}{2} 2^k \text{ when B is } (a_j a_k) \; , j, k \neq i \\ &= l + (l-1)[(l-2)2^{k-1} + 2^k]. \end{aligned} \tag{B-1}$$

Similarly, for the heterozygous genotype,

$$\begin{aligned} N_A(a_i a_j) &= 2.2^k \text{ when B is } (a_i a_i) \text{ or } (a_j a_j) \\ &+ (l-2)2^k \text{ when B is } (a_k a_k) \; , k \neq i, j \\ &+ 3^k \text{ when B is } (a_i a_j) \\ &+ 2(l-2)4^k \text{ when B is } (a_i a_k) \text{ or } (a_j a_k) \; , k \neq i, j \end{aligned} \tag{B-2}$$

$$+ \quad \frac{(l-2)(l-3)}{2}4^k \text{ when B is } (a_k a_m)\ , k, m \neq i \text{ or } j$$

$$= \quad l.2^k + 3^k + 2.4^{k-1}(l-2)(l+1).$$

The total number of feasible genotypic configurations on the block is then given by:

$$\begin{aligned} N &= lN_A(a_i a_i) + \frac{l(l-1)}{2}N_A(a_i a_j) \qquad \text{(B-3)} \\ &= l^2 + l(l-1)[(l-2)2^{k-1} + 2^k] + \frac{l(l-1)}{2}[l2^k + 3^k + 2.4^{k-1}(l-2)(l+1)] \\ &= l^2 + 2^k l^2(l-1) + 3^k\frac{l(l-1)}{2} + 4^k(l+1)l(l-1)(l-2). \end{aligned}$$