

# BITS AND PIECES

## CONTENTS

1. EVOLUTION OF DOMINANCE	1
1.1. Calculations	2
1.2. Fisher v. Wright	2
1.3. Our contribution	3
2. THE FRAGILE-X SYNDROME	4
2.1. The Laird hypothesis	5
2.2. Population genetic calculations	6
2.3. Current status of fragile-X	7
3. EQUILIBRIUM OF MUTATION RATES	7
3.1. Kimura's neutral mutation formula	8
3.2. The CpG dinucleotide	8
3.3. The expected frequency of the CpG dinucleotide	9
4. MITOCHONDRIAL DNA ANALYSIS	10
4.1. Significance test for population differentiation	10
5. EYES ABSENT ( <i>eya</i> )	12
References	13

This section consists of unrelated bits and pieces that I couldn't quite bring myself to leave out of the PIFFLE but probably should have.

### 1. EVOLUTION OF DOMINANCE

Many deleterious genes, in *Drosophila* and in other organisms, are recessive. RA Fisher in 1928 [2] proposed that it is no accident that deleterious genes are recessive. Such recessiveness evolves, he argued,

because it is advantageous for the normal wild type gene to be dominant. Modifier gene(s) would be selected at loci other than the locus in question to produce such dominance, or recessiveness of the mutant gene.

This sounds like a logical and plausible argument. The problem with it is, however, that the selective forces involved are very small. Sewall Wright in 1929 [12] argued against the theory on these grounds.

### 1.1. Calculations.

The genotypes and selective values for a single locus with a deleterious gene  $a$  are as follows:

Genotype	$+/+$	$+/a$	$a/a$
Selective value	1	$1 - hs$	$1 - s$
Frequency (before selection)	$p^2$	$2pq$	$q^2$

$$h = 0.5 \quad \xrightarrow{\text{Evolution of dominance}} \quad h = 0$$

Mutation produces the  $a$  allele at rate  $u$  per generation. An equilibrium is expected when the frequency of the gene  $a$  is

$$\hat{q} = \frac{u}{hs} \tag{1}$$

Selection for modifiers can only occur in heterozygotes, whose frequency is approximately  $2q$ . The reason for the low intensity of selection is the likely small size of  $u$ , presumably no higher than  $10^{-5}$ .

Lowering the value of the selective coefficient  $s$  will raise the equilibrium frequency of  $a$ . However this leads to a proportionate lowering of the intensity of selection, and is unlikely to change the conclusions.

As dominance becomes more complete, lowering the value of  $h$ , the frequency of  $a$  is expected to increase. I don't think that Fisher made much of this in his arguments, although it was certainly mentioned as a factor. Note that  $q$  does not increase without limit as implied by equation (1), since eventually selection against the recessive genotype  $a/a$  becomes the predominant force holding down the  $a$  frequency.

### 1.2. Fisher v. Wright.

The above calculations are presented in the form given by Wright [12], which is closer to current notation than the notation Fisher used. However Fisher's calculations were essentially identical to those presented by Wright. I think that part of the reason that Fisher took offense at

Wright's paper was the suggestion that Wright was adding anything to the theory by the way that he produced his calculations.

The argument between Fisher and Wright was about whether a small constantly acting selective force would have long-term consequences. Fisher said that it would and Wright said that it wouldn't, and that's pretty much the essence of the argument. Fisher believed in the primacy of natural selection as a force in evolution. Wright, on the other hand, was more of a believer in small populations and selective interactions between different genes. I'm not sure that there is yet a decisive answer to the question of whether Fisher's or Wright's world view is more relevant, although it seems clear that Fisher's more extreme statements about the non-existence of neutral gene substitutions have proven to be wrong.

Wright suggested that dominance, rather than evolving, could be a property of the enzymes evolved. To the extent that mutant genes lead to the absence of a functional enzyme, the amount of functional enzyme produced by one copy of the wild-type gene could be enough to do the work normally controlled by two genes.

There is, or evidently was in 1970, still interest and controversy about the topic, otherwise I assume that we wouldn't have been asked to write a review [10]. But it is not clear to me that much has been added to the topic since the original 1928-1929 discussion. Ewens felt sufficiently strongly about it to write a couple of papers on the topic. I recall that we had a somewhat drunken argument in which I put my views. I don't think I have ever apologised for some over the top statements. Sorry Warren.

### 1.3. Our contribution.

Ken Kojima was editing a book on current topics in population genetics, and asked Walter Bodmer to write a review paper on the evolution of dominance. Walter passed it on to me, and I dragged my friend Oliver Mayo into it.

I don't actually currently have a copy of our paper. It is one of the unfortunate side-effects of the current move to online journals that papers published in journals are usually instantly accessible while those published in edited volumes are not. In retrospect I published too much in the latter.

I do feel, in spite of what I have written above, that we did add one facet to the argument. This concerns the necessity to take into account canalisation of development. CH Waddington argued that the

developmental pathways of organisms are organised to oppose the effects of variation in genotype or environment. The end result is thus a normal phenotype under all, or most, circumstances, which Waddington described as canalisation. There are clever experiments in bristle development in *Drosophila* supporting these arguments.

The corollary of this, with respect to the evolution of dominance, is that it is difficult to view the effects of a single gene in isolation. There may be dominance, and this dominance may, indeed, have evolved. But it would be difficult to attribute this evolution to the effects of just one rare gene rather than to a class of genes affecting development of the same character.

I believe that a similar argument may have been put forward by CR Plunkett, a *Drosophila* geneticist of the 1920s and 1930s. This would have been prior to the introduction of the canalisation concept, but I believe that he had the essential idea.

If one accepts this argument, is one taking Fisher's side or Wright's side in the evolution of dominance controversy? I've never been sure about this.

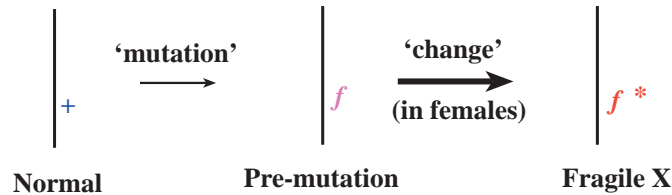
## 2. THE FRAGILE-X SYNDROME

I was introduced to this topic by my friend Charles Laird. Charles was a student with Walter Bodmer in Stanford when I was a postdoc, and subsequently moved to the Zoology Department at University of Washington and then to the Hutchinson Institute in Seattle .

The molecular mechanism of the fragile-X mechanism is now known. At the time it wasn't. It was known that it was an X-linked condition, associated with a syndrome leading to mental retardation and other traits. Grant Sutherland at the Children's Hospital in Adelaide and others had shown that the syndrome was associated with one of a number of 'fragile sites', where chromosomes grown in cell culture tend to break.

Many pedigrees with the syndrome were known, substantially due to the work of Gillian Turner at Prince of Wales Hospital in Sydney. These were analysed by Stephanie Sherman in Newton Morton and Pat Jacobs's group in Hawaii, as well as by other groups. The conclusion was that there must be three states of the gene responsible: a normal state (+), a 'pre-mutation' state ( $f$ ), and a fully mutant state ( $f^*$ ). Furthermore the change from the pre-mutation state to the mutant state

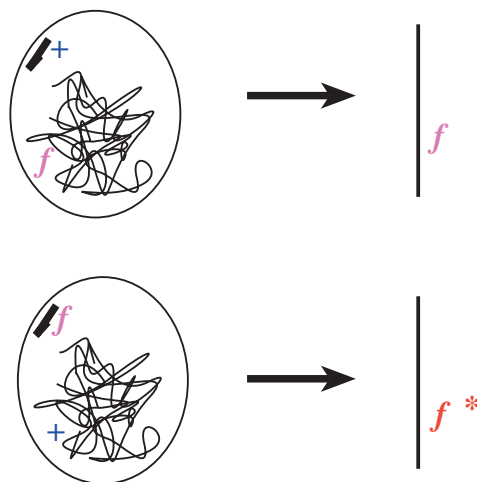
seemed only to happen in females. The three states are shown in the diagram below:



An additional complication was identified by the group in Hawaii, subsequently given the name 'Sherman's paradox'. The 'change' from pre-mutation to mutation seemed to happen at high frequency in some females, and at lower frequency in others. This is a slight simplification of the Sherman paradox, but I believe is its essence. The numbers are complicated by 'ascertainment bias', but are of the order of 70% in the high case and 20% in the low.

### 2.1. The Laird hypothesis.

Under the 'Lyon' hypothesis, only one of the two X chromosomes in a female is active. The other is inactivated and goes into the 'Barr body'. What Charles Laird suggested [6] was that the change from pre-mutation state to mutant state happened when the X-chromosome containing the  $f$  mutation was the 'inactivated' X chromosome in the mother's germline. Charles described this as 'imprinting'. The diagram below shows the two possible situations in a heterozygous female containing the pre-mutation.



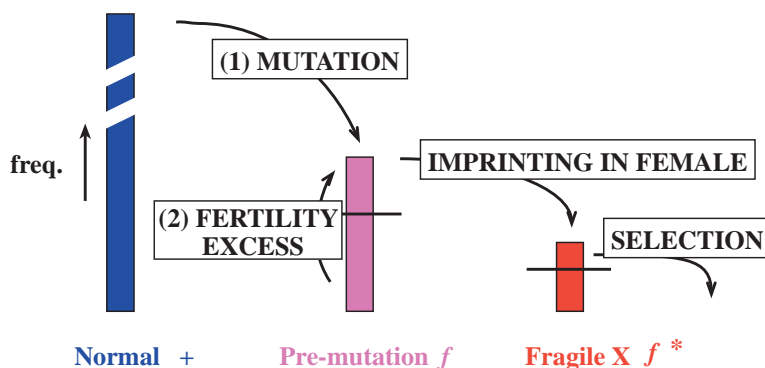
An additional twist to this model neatly explained the Sherman paradox. Charles postulated that there are very few, probably only two,

progenitor cells that give rise to the gonadal cells. Just by chance, therefore, the progenitor cells could all be ones where the pre-mutation is inactivated, giving 100% mutant offspring. On the 2-cell model, 1/4 of  $f/+$  females are expected to give 100% fragile-X offspring, 1/2 are expected to give 50% fragile-X, and 1/4 are expected to give none. The observed segregation patterns fitted well with this model.

## 2.2. Population genetic calculations.

Charles came from Seattle to give a seminar on his hypothesis in Madison, where I was on leave. After a few drinks, I suggested that it should be easy to model the population genetics of this hypothesis, and Charles enthusiastically agreed to a joint project on this.

The results were published a year or two later [8]. We modeled two possibilities: (1) that the primary reason keeping fragile-X in the population, in the face of selection primarily in males, is mutation, (2) that a fertility excess in non-imprinted females could be responsible. Generally we rejected the second hypothesis, and suggested that the analysis of closely linked SNPs should be able to prove the high frequency of the primary mutation event.



Some time after we started the calculations we found out about a paper by Winter [11] that analysed the normal - pre-mutation - mutation scenario, but without giving the specific 50% imprinting probability. I'm not sure that we improved greatly on Winter's calculations. However we did publish a second paper [9] that specifically took into account the probability of ascertainment under Charles's two-cell model. The population implications seemed to fit so well with the observed numbers that at the time I found it difficult to believe that the hypothesis could be wrong.

A brief aside on this second paper: It was published in a journal called 'Brain Dysfunction' with a nominal publication date two years before

our first paper. The story behind this is that I was invited to be on the editorial board of *Brain Dysfunction*, for reasons that are difficult to fathom, but possibly because of a paper that Larry Sandler and I published on some obscure theory on age effects in Down's syndrome. Anyway as an inducement, all people on the editorial board were invited, expenses paid, to a meeting in Troina, a small town in the hills of Sicily. It is one of the few times I have been invited to a meeting expenses paid, and certainly the only time they were paid in small denomination lire notes. It was a most enjoyable meeting, and we got to see the amazing facilities that had been built for treatment of mentally retarded children. Unfortunately the effort to publish a journal from a relatively small Sicilian institution seems not to have been successful, judging by the difficulty of finding the journal now.

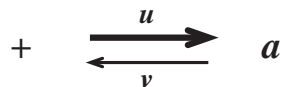
### 2.3. Current status of fragile-X.

Not long after the above events, the site of the fragile-X was cloned and analysed by a couple of groups, including one South Australian one led by Rob Richards and Grant Sutherland. It turned out to be a CGG repeat, that is expanded to a moderate extent in the pre-mutation and then to a much greater extent in the mutation. Furthermore a dozen or so disease states were found to have the same cause, either CGG or CAG repeats amplified.

Most of the triplet repeat genes are not X-linked. The Laird hypothesis deals specifically with X inactivation, and is therefore not relevant to most of these genes. So there is little interest in it any more. Charles was not a 'human geneticist', and I suspect for this reason his hypothesis was never popular in the human genetics community. I don't think that anyone has specifically ruled out the possibility that inactivation has something to do with the triplet expansion. But there is also little reason to suspect that X-inactivation, involving the shutting down of genes, should have anything to do with repeat expansion, which involves some sort of meiotic or pre-meiotic germline abnormality.

## 3. EQUILIBRIUM OF MUTATION RATES

When I learned population genetics, mutation rates were deemed not to be all that important. For instance, a model could be set up in which there is a balance between backward and forward mutation:



The expected equilibrium frequency of the  $a$  gene,  $\hat{q}$  is then

$$\hat{q} = \frac{u}{u + v} \quad (2)$$

But we were taught that such an equilibrium was unlikely, since selection against whatever mutation was involved would over-ride the effects of any back-mutation. So typically the equilibrium would be defined by a balance between mutation and selection, something like equation 1, rather than equation 2. Under this scenario, the actual mutation rate is not a specially important parameter.

### 3.1. Kimura's neutral mutation formula.

Kimura in his 1968 paper [5] turned this theory on its head, arguing that the rate of evolution is primarily determined by the mutation rate. The theory is very simple. If the mutation rate is  $u$  in a population of size  $N$ , the total number of new mutations produced per generation is  $2Nu$ . Looking far into the future, the chance of each of these mutations surviving and taking over the population is  $1/2N$ , providing that the mutation has no effect on fitness. Multiplying the two together, the rate of evolution in the population,  $r$ , is simply equal to the mutation rate per individual gene  $u$ .

Curiously, despite the difficulty some population geneticists such as me initially had in seeing this result, I suspect that it is something that molecular geneticists would have taken for granted. Typically, in those days at least, there would be just one DNA sequence in each of two species. If the two differed in a synonymous substitution, one making no difference to the amino acid involved, what could be more natural than equating the probability of this event to the mutation rate? The complications of whether a gene is substituted into a population or not never enter into the argument.

### 3.2. The CpG dinucleotide.

There is one example in DNA analysis where there is evidence for a systematic bias due to mutation. It has been possible since the earliest days of DNA chemistry to measure the frequencies of the four nucleotides A, C, G and T. Subsequently, but still almost 50 years ago, an ingenious experiment from Arthur Kornberg's group measured the frequencies of the 16 dinucleotide pairs, eg A followed by A (ApA), A followed by C (ApC) etc. [4]. They found that most pairs occurred at frequencies close to expected for the individual nucleotides. However one pair, the dinucleotide CpG, was way down on expectation in mammalian DNA. It occurred at a frequency of only 20% of expectation.

There was no obvious explanation at the time, and I recall that it even led to some discussion about whether DNA really had the properties required of the hereditary material.

Much later it was found that there is high likelihood that C in the CpG dinucleotide will be 'methylated'. This addition of a methyl group to the nucleotide makes C look more like a T. The DNA polymerase controlling DNA replication will sometimes mistakenly put A opposite the methylated C rather than G. The complementary strand thus becomes CpA rather than CpG (reading backwards), and when this in turn is complemented it becomes TpG. The net effect is that the frequency of CpG dinucleotides is reduced, and the frequency of TpG and CpA is increased.

### 3.3. The expected frequency of the CpG dinucleotide.

Adrian Bird is a leading figure in the analysis of vertebrate genomes, particularly the effects of DNA methylation. Marianne Frommer once spent a sabbatical leave with him in Edinburgh. Some years later, when he was working in Vienna, Marianne and I visited him. Over drinks one night, he posed the question of whether continued mutation should lead to the eventual disappearance of the CpG dinucleotide. This led to some discussion of the dynamics of CpG frequencies in populations, and eventually to a joint paper [7].

We put forward the simple formula

$$\hat{q} = \frac{U}{U + V} \tag{3}$$

where  $\hat{q}$  is the equilibrium frequency of the CpG dinucleotide,  $U$  is the mutation rate from the CpG dinucleotide and  $V$  is the back-mutation rate to the dinucleotide. It is a sort of extension of Kimura's formula for a single mutation parameter to the case of a balance between mutation rates.

Equation 3 is exactly analogous to equation 2, the equation that was supposedly not to be trusted in determining the equilibrium gene frequency. There are, however, differences between the two cases. Equation 3 refers to an overall frequency taken over all sequences in the genome. There are presumably many cases where selection would oppose the mutational change, in addition to regions of the genome known as 'CpG islands' where methylation does not occur. But overall, we argued, the currently observed frequencies roughly represent a balance between gain and loss of the CpG dinucleotide.

Equation 3 is evidently a simplification, since there are altogether 16 dinucleotides whose frequency needs to be taken into account in a complete analysis. Differences in frequency between transitions (C-T and A-G) compared to transversions (C-G, C-A, T-A and T-G) also need to be taken into account. We calculated the expected rate of change over time, and applied this to a pseudogene whose time of establishment could be approximately dated.

I worried at the time that back mutation in equation 3 might not be the most important factor in keeping the CpG frequency up, that it might rather be selection or non-methylation. However despite the fact that this paper has been quite widely cited, I haven't had any follow-up from anyone about this.

#### 4. MITOCHONDRIAL DNA ANALYSIS

Sheila van Holst did a PhD in the Fruit Fly Research Centre on mitochondrial DNA of two aboriginal populations. It may seem an odd juxtaposition of themes, but the molecular techniques required to analyse fruit fly mitochondrial DNA are very similar to those required for human DNA. Sheila had a falling out with the laboratory where she initiated this work, and came to us to finish the project.

The two populations that Sheila ended up analysing were a 'riverine' population in western NSW, actually a mixture of two different language groups, and a 'desert' population in the Northern Territory. Sheila was responsible for collecting and obtaining permission for the use of the riverine samples, a tricky prospect given the antagonism of native groups to scientists based on past exploitation. The desert samples were collected by my fellow PhD student from Adelaide University, Barry Boettcher.

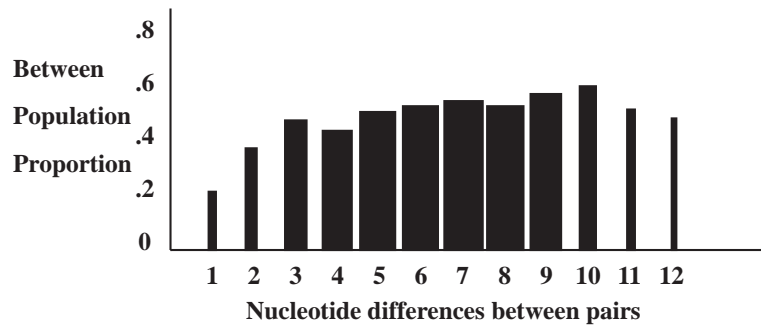
Sheila analysed mainly the control-D region of the mitochondrial genome. A critical advantage of using mitochondrial DNA is that there is a high likelihood that this will be of aboriginal origin despite the high amount of interbreeding that has occurred with europeans. The control loop region is known to be the region that evolves fastest.

##### 4.1. Significance test for population differentiation.

My very minor contribution to this project was to suggest a test of significance to show that there really was significant divergence between the populations. It may seem surprising that it is necessary to do this. But a complicating feature of mitochondrial DNA, in both humans and

fruitflies, is the large amount of polymorphism in present-day populations, that presumably predates any recent separation of populations or even species. It is easy to show that there are frequency differences between populations. It seemed to me that this was a very weak test, although I had considerable difficulty persuading people of this, and that it was necessary to go beyond this to show that there are significant differences in the range of mitochondrial genotypes (haplotypes). The difficulty of doing this is accentuated by the finding that the range of haplotypes found in Sheila's populations actually includes the standard European haplotype.

I suggested two tests. One we called the BEPPI test (BEween Popula-tion Proportion Index - a relic of my inclination to name things). Forty-seven haplotypes were found, differing at 49 positions. All pairs were classified by number of nucleotide differences, and then the proportion of between populations as opposed to within populations was calculated for each difference class. Results are shown in a histogram below, where the width of each class indicates the number of pairs.



As expected, the proportion of 'betweens' increases as the number of differences rises. This was tested by calculating a weighted regression. The significance of this statistic was then tested by simulating an equivalent set of sequences with each sequence assigned to one of the two populations at random. This procedure was repeated many times to obtain a distribution for the expected value of the weighted regression. The observed value turned out to be close to the top 1%.

A second test, the 'connectedness' test was concerned with median networks. Sheila spent a lot of time drawing up such possible networks connecting all sequences. I wrote programs that tried to find the minimum distance from each sequence to a sequence from the opposite population. There were all sorts of rules about what to do when there were two threads of equal length, avoiding circles, joining threads etc.

Then a mean score was calculated for all sequences in the two populations. This was again compared against a distribution of scores when sequences were randomly assigned to the two populations.

This test seemed promising in leading to higher significance levels compared to the BEPPI test. But for reasons that I don't fully recall, perhaps because there were difficulties in implementing the program, in the end the test never saw the light of day.

## 5. EYES ABSENT (EYA)

When I started at Sydney University, there was a very small advanced class, which made it possible to do quite ambitious class practicals compared to what can be done with the class large numbers nowadays. The class did an experiment where we made chromosomes newly collected from wild (winery) populations homozygous, using balancer chromosomes. I've forgotten details, but it's a three generation procedure (see the beginning of the PIFFLE file FITNESS IN DROSOPHILA), which is quite a lot to take on in a class. Maybe we, rather than the students, did some of the earlier crosses.

The main aim of the experiment was to see how many wild chromosomes carry recessive lethal genes. However there was also some chance of finding a visible mutation. One of the students named Philippa Wall reported that she had found a mutation with no eyes. The large red eyes of *Drosophila* are a prominent feature, but nevertheless it would have been easy to miss the mutant flies among all the wild types.

I called the mutation *eyes absent* (*eya*). There were already a couple of such mutations in *Drosophila*, the best known being the 4th chromosome mutation *eyeless* (*ey*). Despite the name, *ey* mutations actually have a sizeable although much reduced number of eye facets. Ours had none, so it looked to be of interest. We knew that the mutation had to be on chromosome 2, and eventually we, probably my colleague Chris Gillies, showed that it was associated with a large chromosome inversion that could be seen in polytene chromosomes under the microscope. Chris used it for many years in cytogenetics lab classes.

Although it looked to be an interesting mutation, I seem to have sat on it for 16 years without sending to anyone. Not sure why! I assume that I was not the only one looking after it at all times since it was used in lab classes, but I was certainly partly responsible, because once I lost it. Most fortunately I had a second stock at the time with the mutation kept over a balancer chromosome, and was able to resurrect

it. Perhaps after that I decided that it was time to send off it to the *Drosophila* stock centre.

I also put a note in *Drosophila* Information Service (DIS 163: 69, 1986). I had intended to add Philippa's name on the note but shamefully forgot when sending it off.

The note was picked up by people in Seymour Benzer's lab, the lab responsible for isolating many of the important mutations affecting development and behaviour. They re-isolated the mutation by mutating laboratory flies and finding mutations that complemented the *eyes absent* phenotype [1] in crosses to our stock. This showed that it was, indeed, a single recessive mutation, and at the same time their procedure meant that their subsequent studies did not need to use our stock that contained the inversion and goodness knows what else. As well as in *Drosophila*, they showed that *eya* also exists in humans, with multiple protein products and a bewildering range of activities [3]. Of course I have had nothing to do with any of this, but it is pleasing that our discovery played a small role in the story.

#### REFERENCES

- [1] N M Bonini, W M Leiserson, and S Benzer. The eyes absent gene: genetic control of cell survival and differentiation in the developing drosophila eye. *Cell*, 72(3):379–95, Feb 1993.
- [2] RA Fisher. The possible modification of the response of the wild type to recurrent mutations. *American Naturalist*, 62:115–126, 1928.
- [3] Rashmi S Hegde, Kaushik Roychoudhury, and Ram Naresh Pandey. The multi-functional eyes absent proteins. *Crit Rev Biochem Mol Biol*, 55(4):372–385, 08 2020.
- [4] J JOSSE, A D KAISER, and A KORNBERG. Enzymatic synthesis of deoxyribonucleic acid. viii. frequencies of nearest neighbor base sequences in deoxyribonucleic acid. *J Biol Chem*, 236:864–875, 1961 Mar.
- [5] M Kimura. Evolutionary rate at the molecular level. *Nature*, 217(5129):624–626, 1968 Feb 17.
- [6] C D Laird. Proposed mechanism of inheritance and expression of the human fragile-x syndrome of mental retardation. *Genetics*, 117(3):587–599, 1987 Nov.

- [7] J Sved and A Bird. The expected equilibrium of the cpg dinucleotide in vertebrate genomes under a mutation model. *Proc Natl Acad Sci U S A*, 87(12):4692–4696, 1990 Jun.
- [8] J. A. Sved and C. D. Laird. Population genetic consequences of the fragile-x syndrome, based on the x-inactivation imprinting model. *Am J Hum Genet*, 46(3), 1990.
- [9] JA Sved and CD Laird. The x-inactivation imprinting model can explain the incidence of the fragile-x syndrom of mental retardation in mother-offspring pairs. *Brain Dysfunction*, 1:245–254, 1988.
- [10] JA Sved and O Mayo. The evolution of dominance. In K Kojima, editor, *Mathematical topics in population genetics*, Berlin, 1970. Springer-Verlag.
- [11] R M Winter. Population genetics implications of the premutation hypothesis for the generation of the fragile x mental retardation gene. *Hum Genet*, 75(3):269–271, 1987 Mar.
- [12] S Wright. Fisher’s theory of dominance. *American Naturalist*, 63:274–279, 1929.