HAL open science

The curious case of a hapless mathematical contribution to biology

Eric Tannier

To cite this version:

Eric Tannier. The curious case of a hapless mathematical contribution to biology. 2022. hal-03153696v2

HAL Id: hal-03153696

https://hal.science/hal-03153696v2

Preprint submitted on 21 Mar 2022 (v2), last revised 20 Sep 2022 (v3)

HAL is a multi-disciplinary open access archive for the deposit and dissemination of scientific research documents, whether they are published or not. The documents may come from teaching and research institutions in France or abroad, or from public or private research centers.

L'archive ouverte pluridisciplinaire **HAL**, est destinée au dépôt et à la diffusion de documents scientifiques de niveau recherche, publiés ou non, émanant des établissements d'enseignement et de recherche français ou étrangers, des laboratoires publics ou privés.

Manuscript submitted.

The curious case of a hapless mathematical contribution to biology

Chromosome inversions in *Drosophila*, 1937-1941

Eric Tannier

Abstract This is the story, told in the light of a new analysis of historical data, of a mathematical biology problem that was explored in the 1930s in Thomas Morgan's laboratory at the California Institute of Technology. It is one of the early developments of evolutionary genetics and quantitative phylogeny, and deals with the identification and counting of chromosomal rearrangements in *Drosophila* species from comparisons of genetic maps. A re-analysis of the data produced in the 1930s using current mathematics and computational technologies reveals how a team of biologists, with the help of a renowned mathematician and against their first intuition, came to an erroneous conclusion regarding the presence of phylogenetic signals in gene arrangements. This example illustrates two different aspects of a same piece: 1) the appearance of a mathematical in biology problem solved with the development of a combinatorial algorithm, which was unusual at the time, and 2) the role of errors in scientific activity. Also underlying is the possible influence of computational complexity in understanding the directions of research in biology.

Keywords history of biology · evolutionary genetics · chromosomal inversion · genetic maps · statistics · computational complexity · scientific errors · history of interdisciplinary studies · *Drosophila*

Eric Tannier
Centre de Recherche Inria de l'Université de Lyon
Laboratoire de Biométrie et Biologie Évolutive UMR5558, F-69622 Villeurbanne, France
Tel.: +33-(0)426234474
Fax: +123-45-678910
E-mail: eric.tannier@inria.fr

This is the first time in my life I believe in constructing phylogenies, and I have to eat some of my previous words in this connection. But the thing is so interesting that both Sttt [Sturtevant] and myself are in a state of continuous excitement equal to which we did not experience for a long time.

Theodosius Dobzhansky, letter to Milislav Demerec 1936

I am rather surprised to find myself figuring out hypothetical phylogenies for the Drosophila species, and taking them more or less seriously — after all the uncomplimentary remarks I've published about such procedures.

Alfred Sturtevant, letter to Otto Mohr 1939

These two quotes attest to the renewed interest in phylogeny during the first half of the twentieth century. Marks of enthusiasm such as these, associated with the revival of this old discipline, were common. Among other possible reasons, they are due, on the one hand, to the use of cytological and genetic comparisons, offering direct access to hereditary material, and on the other, to the use of quantified methods, often associated with objectivity. According to Anderson (1937), cytology was like "looking at the cellar window", and is "evidence as to the germplasm itself and is, therefore, of more fundamental importance than the mere architecture erected by the germplasm itself." For Turrill (1938), chromosomes provided "high-powered morphology". For McClung (1908), "The chromosomes are the determinants of characters", and "one cell is sufficient for the identification of the species". "Were our knowledge of cell structure in the grasshopper complete enough we might erect a system of classification based upon cytological characters, just as reasonably as we have designated one using external anatomical structures" (McClung, 1908). As for quantification, the comparisons allowed by precipitin reactions (Strasser, 2010b) made Boyden (1934) write that "The fact that naturalists of recent times have so often forsaken the study of phylogeny is due more to the feeling that such a study is likely to yield little certain progress than to the belief that the problems of phylogeny are unimportant or sufficiently well analyzed."

Of course, the use of both "semantic"[1] characters and quantification, driven by the development of sequencing techniques and computers, was only fully realized in the 1960s by the founders of Molecular Evolution (Suárez-Díaz, 2009; Dietrich, 2016). However the evolutionary genetics program that began in Thomas Morgan's laboratory in 1914, and was subsequently continued by the patners turned rivals Alfred Sturtevant and Theodosius Dobzhansky, had similar epistemological characteristics[2].

The aim of this article is to give an account of a particular moment of this research, focusing on Sturtevant's attempts, over several years and with several successive Ph.D. students, to quantify the number of inversions between homologous genetic linkage groups in two *Drosophila* species. Some aspects

[1] According to the vocabulary of Zuckerkandl and Pauling (1965), this is the directly transmitted hereditary material, and not one of its products, see also Dietrich (1998).

[2] Despite crucial differences in the biological objects have also been discussed (Darden, 2005).

of this research, in particular the attempts to quantify evolutionary divergence, the involvement of the mathematician Morgan Ward, and the errors that resulted, have been overlooked in historical accounts of the study of chromosome evolution (Hagen, 1982, 1984; Kohler, 1994; Gannett and Greisemer, 2004; Smocovitis, 2006) and of the use of quantification in biology (Hagen, 2003; Suárez-Díaz and Anaya-Muñoz, 2008; Suárez-Díaz, 2010; Hagen, 2010).

In the 1930s the use of mathematics, and collaborations with mathematicians was commonplace in biology, and particularly in evolutionary biology. It was even central part to the construction of the modern synthesis (Bowler, 2003). However the type of mathematics in this example is unusual in that it differs from that available to evolutionists, as developed for instance by Fisher, Wright or Haldane for statistics and population genetics. Retrospectively combinatorial and computational aspects are visible, which were handled at the time with underlying[3] systematically applied algorithms on permutations. Some of the questions addressed at the time were only solved 50 years later, and some even remain unsolved today. The difficulties that mathematicians encounter today with these problems were already visible in Sturtevant's attempts. Nevertheless, after trying to solve the same questions myself with today's mathematics and technology, I found three computational and numerical approximations, initially acknowledged as such by the authors, which, after consulting Ward, strangely turned into errors and led to a wrong biological interpretation. This curious case of an unfortunate use of mathematics to solve an evolutionary question illustrates the presence and importance of errors in the practice of science. We could also see it as an example of the often overlooked impact of computational intractability (Papadimitriou, 1993) on biological research.

In the first part of this article, I provide some contextual elements concerning the use of chromosomes in evolutionary studies, both worldwide and in Thomas Morgan's laboratory, in order to highlight the originality of Sturtevant's research. In the second part I describe how Sturtevant progressed from making the first genetic map to the challenge of counting inversions. Alongside historical descriptions, I give my solutions to the described problems using current scientific knowledge. In the third part, I discuss what this exercise can teach us about the unexpected presence of combinatorial algorithmic considerations in 1930s biology, and about the influence of errors and complexity in both past and present research programs.

1 Chromosomes as documents of evolutionary history

In the first half of the twentieth century, the development of genetics and cytology led several researchers and research teams to compare chromosomes and/or linkage groups in order to establish evolutionary relationships and reconstruct evolutionary histories (Hagen, 1982). New markers emerged to delimit and

[3] *I.e* not explicit

classify species or construct phylogenies, such as: the number, shape and size of chromosomes, their behavior during the cell cycle, the position of the centromere or the arrangement of genes. To cite only a few landmarks of this development: at the International Zoological Congress in Boston in 1907, the cytologist Clarence Erwin McClung stated that a character measured within the cell, such as the number of chromosomes, could be considered as informative for phylogenetic classification as any morphological character (McClung, 1908). In 1915 in Berkeley, California, the plant geneticist Ernest Brown Babcock gathered a team to work on the evolution of the flowering plant *Crepis* and contributed to the foundation of the "Bay Area Biosystematists" (Hagen, 1984; Smocovitis, 2009), an influential multidisciplinary group working on plant systematics. In 1926, the International Congress of Plant Science held a joint session involving taxonomists, cytologists and geneticists (Hagen, 1984). In 1937, the field was sufficiently established for Edgar Anderson, from the Missouri Botanical Garden, to write an extensive review on the contribution of cytology to taxonomy in botany (Anderson, 1937). In 1938, Babcock and his collaborator George Ledyard Stebbins Jr, who would become a leading evolutionary biologist (Smocovitis, 2006), published the influential book *The American Species of Crepis*, in which all the genetic and cytological knowledge available at the time was harnessed to decipher the complex evolutionary relationships between members of the the *Crepis* genus (Babcock and Stebbins, 1938; Smocovitis, 2009).

A comparable research program on the fruit fly *Drosophila*, the traditional model organism from which genetics was first developed (Kohler, 1994), was carried out in Thomas Hunt Morgan's genetics laboratory, first at Columbia from 1914 to 1928 and then at Caltech. It was initiated by Charles Metz, born in 1889, who joined in 1912 Morgan's laboratory at Columbia where he became interested in cytology. Metz soon realized that his observation of *Drosophila* chromosomes in anaphase possibly carried phylogenetic information because different species had different chromosomal conformations. Combining data for the presence or absence of microchromosomes and the state of two autosomes (fissioned vs. fused) in 12 *Drosophila* species, Metz managed to classify chromosome organizations into five types. These types were then organized into a tree, where the branches could be interpreted as evolutionary events (Figure 1).

In article published in 1914, from which Figure 1 is reproduced, Metz speculated that differences in chromosome types "may indicate an evolution of chromosomes in the genus" (Metz, 1914). However, in his subsequent articles on the description of chromosome types, Metz became more and more cautious regarding any possible evolutionary interpretation (Kohler, 1994), mainly because of the difficulty in assessing the homology[4] between chromosomes via the technique of independent observation in different species. As

[4] The term homology, in the sense of "common evolutionary origin", was not commonly used at the beginning of the 20th century. The terminology was discussed and ranged from "allelomorph" to "corresponding". I use the current terminology for the sake of consistency and clarity.

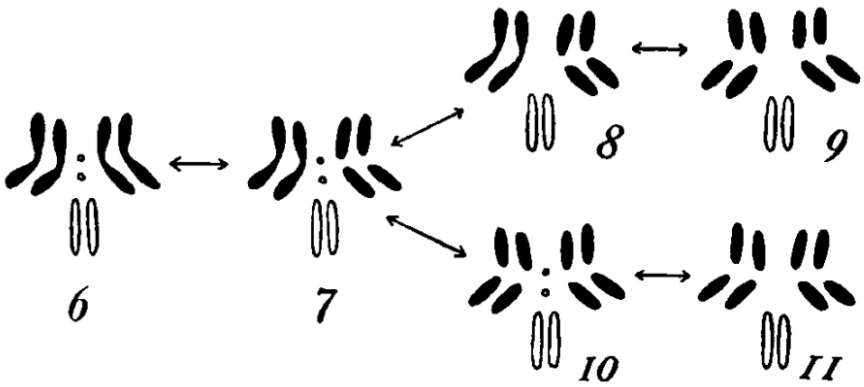

Fig. 1 Reproduction of Figure 1 from Metz (1914). The five different karyotypes from 12 *Drosophila* species, are organized into a tree with a wishful evolutionary interpretation. Nodes 9 and 11 represent the same type of chromosome organization, meaning that the two phylogenetic positions are equally possible. Reproduced with the kind permission of Wiley and the *Journal of Experimental Zoology, Part A: Ecological Genetics and Physiology.*

a result, his subsequent publications (Metz, 1916, 1918) seem more like an organized catalog of chromosome types, with less evolutionary implications.

Then began the search for a technique to assess homology. One method was to produce interspecific hybrids and observe coupled chromosomes during segregation, but this showed little success with *Drosophila* species (Kohler, 1994). Hybrids could be produced but were almost always sterile. Two subsequent techniques would prove more successful for assessing homology and were explored in Morgan's laboratory: gene mapping on chromosomes (from 1917) and hybridization of polytene chromosomes (from 1936).

Charles Metz himself left Columbia University for Washington in 1914 and did not participate further in the activities at Columbia, even though he became an eminent *Drosophila* geneticist. However, while at Columbia he did not work alone and his research program was continued by others. As mentioned in the acknowledgments in his 1914 article (Metz, 1914), he benefited from the help of a young student from Columbia, Alfred Sturtevant.

2 Alfred Sturtevant and Comparative Genetic Mapping, 1921 to 1941

2.1 Genetic maps and predicting inversions

Sturtevant, born in 1891, completed his doctorate in 1914 with Thomas Morgan at Columbia University. One of his legendary achievements was to respond to Morgan's remark, according to which the strength of the genetic linkage between genes, measured from the observation of phenotypic characters, could be related to the physical distance between the genes on a chromosome. From

this idea, Sturtevant defined genetic distance as the percentage of crossing-over between two genes, which he observed from the frequency of associated phenotypes in *Drosophila ampelophila*[5]. As this distance was close to a linear function, it was possible to position genes on a line. This led to the first genetic map, which placed six genes on the "sex-linked" linkage group[6] (Gannett and Greisemer, 2004).

Following Sturtevant's, the same research group produced several other genetic maps. in particular, Morgan and Bridges' 36-marker map of the X-chromosome of *Drosophila melanogaster* (Morgan and Bridges, 1916) was disputed by William Castle (Castle, 1918, 1919) and, by association, several other researchers, who questioned the relevance of the linear model for depicting chromosomes, with responses by Sturtevant et al. (1919); Morgan et al. (1920). Even though each protagonist gave the impression of standing firm on his respective position, the controversy helpd clarify much of the theory, as well as its underlying and *ad hoc* hypotheses.

The real starting point for evolutionary genetic studies was the discovery of mutations in that linear structure. Inversions, *i.e.* evolutionary events reversing the orientation of chromosome segments, were hypothesized by Sturtevant (1921) based on the observation of differences in the arrangement of five "corresponding"[7] genes on chromosome 3 between *Drosophila simulans* and *Drosophila melanogaster*. The inversion hypothesis was confirmed by adding genes, while the comparative mapping carried out by Sturtevant and Plunkett (1926), illustrated in Figure 2, presents a striking visual argument in support of it[8].

Inversions themselves had the same status as linkage groups, that is, they were theoretical objects independent of any direct cytological observation. A cytological demonstration of their existence would be made later with the techniques of Painter (1933).

From this possibility of detecting mutations by comparing genetic maps, Sturtevant developed a comprehensive research project in continuity with his work with Metz. The aim was to map the genes of different *Drosophila* species, asses the homology between these genes and, from the chromosome structure, reconstruct the evolution of these species (Kohler, 1994). This research project was not fully realized, although several publications and many unpublished

[5] Renamed *melanogaster* shorlty after.

[6] Later named the X-chromosome in order to emphasize its peculiarity. The link between chromosomes and linkage groups was already well established, as can be seen by the natural use of "chromosome" in genetic studies from the 1910s onward.

[7] *I.e* homologous, see footnote 3. Homology was deduced from the similarity of phenotype variations during crossing experiments.

[8] Several types of translocations, *i.e.* other mutations of the linear organization of genes along chromosomes, were predicted at the same time (Bridges, 1917; Mohr, 1919; Morgan et al., 1925) and later demonstrated using cytology (Muller, 1929; Dobzhansky, 1930). They were generally considered to be "deficiencies", or abnormalities of karyotypes, possibly resulting from mutagenic conditions. By contrast, inversions were immediately seen as evolutionary patterns susceptible to being used in differentiating species, and thus be a character for taxonomy. Translocations were later used in plant taxonomy by Babcock and Stebbins (1938).

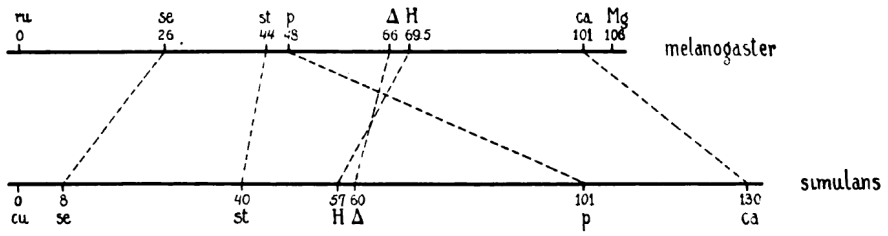

Fig. 2 Reproduction of Figure 1 from Sturtevant and Plunkett (1926): a graphic representation of gene arrangements supporting the existence of inversions and their utility for taxonomy. Linkage group 3 is compared in *Drosophila simulans* (below) and *Drosophila melanogaster* (above). Genes (points) are placed on the line (representing the chromosome or linkage group). Homologous genes are represented by dashed lines. Reproduced with the kind permission of *The Biological Bulletin*.

results[9] confirm they made decisive advances ad well as reveal some challenges. A close look at the mathematical techniques they use helps us understand the progressive introduction of quantification, and how, if it gives the impression of objectivity and fights a "methodological anxiety" (Suárez-Díaz and Anaya-Muñoz, 2008), it is not necessarily a guarantee of greater veracity or accuracy.

2.2 The first mathematical problem: the observed inversion distance

One of the challenges of studying chromosomal arrangements involves the detection of several successive overlapping inversions. Comparing two arrangements that differ by one inversion was easy. However if several overlapping inversions have occurred, which is likely if more distant species were compared, an additional difficulty arose. In 1937, Sturtevant, published with C. C. Tan, a Ph.D. student supervised by himself and Dobzhansky, a comparison of the arrangements of 38 genes along all the chromosomes of *Drosophila melanogaster* and *Drosophila pseudoobscura* (Sturtevant and Tan, 1937). The comparative maps, inferred from the homology of genes deduced from similar phenotypic effects, are reproduced in Figure 3. Inversions are not as visible as in Figure 2 because the species are more distant and thus the accumulation of inversions has blurred the signal.

It is useful to carefully examine both the data and the discussion shown in Figure 3 from the 1937 article by Sturtevant and Tan. A first mathematical problem is stated: given a permutation of letters (the gene order in *melanogaster*), find a sequence of successive inversions transforming it into the alphabetical order (the gene order in *pseudoobscura*). This sequence should have the smallest possible number of inversions, as implied by the term "necessary" in the text. This is the parsimony argument, which was also proposed for comparing DNA or protein sequences by Camin and Sokal (1965). This minimum number has been subsequently named the *inversion distance* of a

[9] Examined by Kohler (1994), who writes that the unpublished part is of wider significance.

legitimate. If the *pseudoobscura* sequence in each arm is arbitrarily taken as an alphabetical one (A B C. . .), then the *melanogaster* sequences become:

X	L	H	F	E	B	A	D	C	K	I	J	G	M	(7)
II *L*	D	E	F	A	C	B	(2)							
II *R*	A	C	E	B	F	D	(4)							
III *L*	C	F	E	B	A	D	(3)							
III *R*	A	E	B	C	F	D	G	(3)						

The numbers in parentheses represent the numbers of successive inversions necessary to turn these sequences into alphabetical ones (in the case of *X* we are not yet certain that six inversions may not be sufficient). The mathematical properties of series of letters subjected to the operation of successive inversions do not appear to have been worked out, so that we are so far unable to present a detailed analysis. It does appear, however, that the five arms (taken together) are definitely more alike in the two species than could result from chance alone.

Fig. 3 Excerpt from Sturtevant and Tan (1937). Chromosome names are given in the column on the left; gene names range from A to M. Numbers in parentheses are the minimum number of inversions that are necessary to transform the arrangement of letters on a line (*melanogaster* arrangement) into the alphabetical order (*pseudoobscura* arrangement). In the paragraph below the letter arrangements, a working program for mathematicians and (not yet existing) computer scientists. Reproduced with permission from *Springer*.

permutation (Fertin et al., 2009). For example, the sequence on chromosome IIL can be transformed into the alphabetical order by two inversions as follows:

$$\underline{DEFA}CB \rightarrow A\underline{FEDCB} \rightarrow ABCDEF.$$

The first inversion concerns the underlined segment $DEFA$, and the second inversion the segment $FEDCB$. It is easy, by enumerating all possible inversions, to see that for this example, one inversion alone cannot transform the initial permutation into the alphabetical order. So the minimum number, *i.e.* the inversion distance of permutation $DEFACB$, is 2. Computing this number becomes tricky when genes and inversions are more numerous.

2.3 Resolution with modern mathematics

Note that in the paragraph in Figure 3, Sturtevant and Tan recognized that for the X-chromosome their best scenario had seven inversions, but they were not certain six was impossible. No detail is given regarding their method for finding the scenario with seven inversions or the reasons why they doubted that seven was minimum number. They probably enumerated many scenarios and could not find one with less than seven, but enumerating all scenarios was considered too long or tedious a task. The cautiousness of their statement was retrospectively a good intuition, since

$$
\begin{aligned}
&\underline{LHFEBA}DCKIJGM\\
\rightarrow\ &AB\underline{EFHLDC}KIJGM\\
\rightarrow\ &ABCD\underline{LHFE}KIJGM\\
\rightarrow\ &ABCDEFH\underline{LKIJG}M\\
\rightarrow\ &ABCDEF\underline{HG}JIKLM\\
\rightarrow\ &ABCDEFGH\underline{JI}KLM\\
\rightarrow\ &ABCDEFGHIJKLM
\end{aligned}
$$

is one of several possible *bona fide* sequences of six successive inversions. It is possible to prove that six inversions are necessary, *i.e.* five is not possible, using the lower bound found by Kececioglu and Sankoff (1995). They define *breakpoints* as pairs of letters that occupy two consecutive places in the initial arrangement, but are not consecutive in the alphabetical order. Thus a pair of breakpoints comprises four letters. If, among those four letters, there are two pairs of consecutive letters in the alphabetical order, the pair of breakpoints is called an *edge*. With b the number of breakpoints, and m the maximum number of edges that do not share breakpoints, Kececioglu and Sankoff (1995) prove that the inversion distance is at least $\frac{2b-m}{3}$. In our case, $b = 9$ and $m = 2$, which makes the lower bound strictly greater than 5.

The "detailed analysis" called for by Sturtevant and Tan (see Figure 3) would have to wait several decades before it became possible with the help of new mathematical and computational techniques (Fertin et al., 2009). However even today, no closed formula or "good" algorithm, *i.e.*, an algorithm that would not require the enumeration of the combinatorial structure, are known to solve the inversion distance problem for any arrangement. Here I did not use any canonical method to find the scenario with six inversions, such a method does not exist. I found this solution while trying to prove that the 6-inversion scenario did not exist, in order to confirm the statement of Sturtevant and Tan (1937). To do so, I assumed its existence, derived some of the properties it should have with the goal of arriving at a contradiction; instead this scenario arose.

The fact that Sturtevant and Tan stated that the result was uncertain is not anecdotal, it is actually important because later on their result was turned into an error. While their passing statement was forgotten, the number seven was taken at face value. Together with two other subsequent approximations this would lead to an erroneous biological conclusion.

2.4 The second mathematical problem: the expected inversion distance

This brings us to the second mathematical problem stated by Sturtevant and Tan, of a statistical nature. The last sentence in Figure 3 states that the arrangements of genes in *pseudoobscura* and *melanogaster* "are definitely more alike in the two species than could result from chance alone." The statistical problem then asks whether the observed gene arrangement has a significantly

lower inversion distance than a random arrangement. The answer requires the computation of an expectation and a variance of the inversion distance for a random permutation. An observation that cannot be attributed to chance (if the observed value falls outside the standard error interval around the expectation) can be considered as the sign of the common origin of the two arrangements.

The word "definitely" in this sentence is interesting for our purpose because it illustrates the progressive extension of the quantification. It means that intuitively, the inversion distances found between *melanogaster* and *pseudoobscura* do not appear to be attributable to chance. This intuition was then turned into a statistical hypothesis in the follow-up paper by Sturtevant and another student, Edward Novitski (Sturtevant and Novitski, 1941). Novitski, like Tan before him, was first a student of Dobzhansky and continued with Sturtevant after Dobzhansky's and Sturtevant's dispute (Novitski, 2005). In each lab, he worked on chromosomes and evolution using different approaches. While working with Sturtevant, he generated a large catalog of homologies, some from the literature and some newly obtained via classical genetic techniques, and carried out a more in-depth mathematical analysis of the 1937 data.

2.5 The call for a professional mathematician

After going over the statements of Sturtevant and Tan (those reproduced in Figure 3), Sturtevant and Novitski announced that they had solicited the help of Morgan Ward, a renowned mathematician from Caltech. Sturtevant himself had a reasonable understanding of mathematics, and Novitski (2005) retrospectively praised his "mathematical mind", compared with Dobzhansky's. However Sturtevant and Novitski probably felt that no easy technique could solve this question and logically solicited the help of an expert.

Morgan Ward (1901-1963) entered Caltech in 1924 as a student, and became a research fellow in 1928. Appreciated by many for his qualities as a teacher, he apparently showed no particular interest in biology, though an acknowledgement can also be found in an article by Dobzhansky and Wright (1941), the only other biology paper, alongside the one studied here, featuring his name. He was more interested in the contribution of his field to physics. His expertise in number theory and Diophantine equations (Lehmer, 1993), which involved the design of calculation methods on integer numbers, might have convinced Sturtevant and Novitski to request his help. The exact mode of collaboration is not known: it is just mentioned in the middle of the article that Ward provided some help. We can suppose one or a few work sessions, where the two mathematical problems, that of the inversion distance and its statistical significance, were exposed and ways to compute the solutions were discussed.

The solution they found to test whether the difference in arrangement was indeed "more alike [...] than could result from chance alone", was for

permutations of 6 genes or less, to enumerate all permutations and for each one, to compute the inversion distance. Then they calculated the mean and standard deviation of all inversion distances. For permutations of eight and nine genes, 60 and 40 permutations were sampled at random instead of the complete enumeration. For higher numbers, permutations were not sampled and the expected inversion distance was obtained by a linear interpolation from smaller number. Indeed, as admitted in the article, "For numbers of loci above nine the determination of [the inversion distance] proved too laborious, and too uncertain, to be carried out" (Sturtevant and Novitski, 1941).

They obtained a mean of 7.6 inversions for 13 genes (see Figure 4), leading them to conclude that, in contrast to their initial intuition, "Evidently the two species are not more alike than could easily result from chance alone". The use of the terms "definitely" in the sentence from 1937 quoted above and "evidently" here suggests several remarks. First, the later statement states that the earlier was evidently a wrong intuition, which tells us something about the scientific personality of Sturtevant: he did not hesitate to admit to himself his supposed error in strong terms. Second, if the latter statement corrects the former by a quantitative assessment of the initial idea, we can note that the intuitive aspect has not been fully eliminated. The authors, after having considered that the differences were "definitely" significant without having calculated them, considered that seven was "evidently" not significantly different from 7.6. However this argument depends on their estimation of the upper bound of the standard deviation ("less than 1", according to the authors). A final remark is that, if we carefully check the calculations, we unfortunately come to the conclusion that the first intuitive argument is correct, and that the revised argument is not. It is sad to note that the willingness of Sturtevant to contradict his own result was itself a scientific error, because in fact the first better reflected the data, according to his own criteria.

It is striking that to this day, we know of no better technique to calculate these numbers. Only the improved performance of computers allows the present day researchers to enumerate all permutations and their inversion distances (for up to 13 genes in 1995 (Galvão and Dias, 2015)[10] instead of up to six in 1941). An asymptotic bound for the mean has been proposed (Bafna and Pevzner, 1996) but it is not applicable to such small values. Consequently I have used the enumeration method to compute, with modern techniques and knowledge, the values for the numbers considered by Novitski, Sturtevant and Ward[11]. I consider these values more accurate than theirs, because I used a complete enumeration of the space instead of an extrapolation. Of course these values are the result of my own understanding of the problem and I cannot

[10] It is a coincidence that the maximum number found in 1995 is precisely the one that biologists struggled with in 1937. That we have not been able to greatly improve our handling of the data is indicative of the inherent computational complexity of the problem.

[11] Note that the corrected values given here were obtained only with the published data and the statistical test proposed by the original authors. However this analysis requires computational tools that were not available at the time. There would probably be a lot more to discover if we were to redo this analysis with new data.

Evidently the two species are not more alike than could easily result from chance alone.

TABLE 4

Comparison of the required and calculated numbers of inversions to change the melanogaster into the pseudoobscura sequences.

ELEMENT	A	B	C	D	E	TOTAL
Loci	13	6	6	6	7	
Inversions required	7	2	4	3	3	19
Inversions calculated	7.6	3.0	3.0	3.0	3.7	20.3

Fig. 4 Excerpt from Sturtevant and Novitski (1941). The letters A, B, C, D and E in the table columns represent the chromosomes and correspond to X, IIL, IIR, IIIL, IIIR, respectively, in Figure 3. The row "Loci" shows the number of genes on each chromosome (corresponding to the number of letters in Figure 3). The row "Inversions required" shows the calculated inversion distances (on the observed arrangements). The row "Inversions calculated" shows the mean inversion distance computed from complete enumeration of permutations, or from samples of permutations, or from interpolation (this is the expected value on random arrangements). A modern calculation finds that all numbers are correct except 7 and 7.6 in column A (and their associated totals), which should be 6 and 7.9. The (wrong) conclusion, which could have been different with the correct numbers, is reprinted above the table. Reproduced with the kind permission of the *Genetics Society of America* and the journal *Genetics*.

discard the hypothesis that a future work will refute them. However I believe this is the best that can be achieved with our current knowledge and technology. This analysis gives an expected inversion distance of 7.9 for 13 genes instead of the interpolated value of 7.6 from 1941 (see Figure 4). The standard deviation is 0.85 instead of the "less than one" estimation from 1941. This is not a big difference, but put end to end, all inaccuracies eventually change the conclusion.

2.6 When the progression of quantification leads to a succession of errors

To recap, there are three small errors or approximations in the 1937 and 1941 articles: the minimum number of inversions (seven required instead of six), the expected number (7.6 instead of 7.9) and the standard deviation ("less than 1" instead of 0.85). Taken together, these change the conclusion. With the corrected calculations, six inversions would have been considered significantly different from 7.9, falling outside the 0.85 standard deviation interval[12].

[12] A *bona fide* statistical test in this case would require a p-value rather than standard deviations. This was not considered in the 1937 and 1941 articles but it is possible to compute an empirical p-value from a sample of 1,000 uniformly sampled random permutations. This gives a probability of achieving six or less inversions for 13 genes of 0.06, a probability of achieving two or less inversions for six genes of 0.2, and a probability of achieving three or less inversions for seven genes of 0.35. Considering each chromosome independently is hardly conclusive. When all chromosomes are taken into account, gene inversions can be

It is retrospectively mind-bogging that Sturtevant and Novitski (1941), assisted by a mathematician, claimed in 1941 to correct the statement of Sturtevant and Tan (1937), while in actual fact they were confirming the only wrong statement of the earlier article, and introduced another error. In 1937 the authors were cautious about the inversion distance number they found, but in 1941 they noted that "this revision does not change the number of inversions required to transform one sequence into the other", thus retaining the erroneous number and ignoring their initial reservations. The help of a mathematician, which ordinarily would have been considered a good idea for such a problem, has perhaps been disastrous in this case, as it undermined, for the wrong reasons, the sound intuitions that the biologists initially had.

Detecting these errors is not just a mathematical exercise or driven by exaggerated attention to detail. It can have historical significance. Success stories are more frequently reported than errors, but sometimes the path taken by scientific research can be influenced by mistakes of different kinds (Firestein, 2015; Livio, 2014). Here, the fact that closely related species, such as *melanogaster* and *pseudoobscura* seemed to have no detectable similarity in gene order may have contributed to orienting genetic research in other directions. Indeed, this conclusion meant that a *Drosophila* phylogeny based on chromosomes was hardly conceivable.

Not much changed after 1941. In a 50-page landmark article on the phylogeny of the *Drosophila* genus, Sturtevant (1942) devoted only two pages to chromosomes and derived no decisive phylogenetic information from them. The article mainly describes comparisons of morphological characters. By contrast, in their book, Babcock and Stebbins (1938) recognized that chromosomes could be used for reconstructing the phylogeny and evolutionary history of the *Crepis* genus, even though *Crepis* is biologically more complex because of the prevalence of hybridization and the diversity of reproductive modes in plants. Babcock began his research on *Crepis* hoping it would be the plant equivalent *Drosophila*, and to explore to what extent the results from Morgan's fly laboratory were generalizable (Smocovitis, 2009). He did not fully succeed in this precise goal but in some aspects went beyond the research in evolutionary genetics and cytology that was carried out on *Drosophila* species.

2.7 Epilogue

Of course the evolutionary genetics project started by Metz and Sturtevant in 1914 did not stop because of a few mathematical errors that were made in the 1930s and 1940s. One important challenge of their project was that it necessitated a prohibitive amount of work to assess the homology of genes and chromosome segments. In the articles analysed here (Sturtevant and Tan, 1937; Sturtevant and Novitski, 1941), a catalog of homologous genes was compiled based on similarities in phenotypic variation. This tedious method, which is

considered significantly different from what would be expected at random based on the usual significance thresholds.

difficult to automate, could not be envisaged beyond a certain evolutionary depth.

The technique invented by Theophilus Painter (Painter, 1933, 1934) to detect homologies between polytene chromosomes was exploited by geneticists and cytologists within a comparative and evolutionary framework (Gannett and Greisemer, 2004). At Caltech, it was used by Theodosius Dobzhansky, first in association with Sturtevant and then independently after their partnership ended. Dobzhansky collected many *pseudoobscura* strains from all over the United States, while Sturtevant collected what he felt was interesting for genetics and, in particular, for his long-standing project of comparing chromosomes from an evolutionary perspective. In 1936, they published together the first phylogenetic tree based on chromosomal inversions (Sturtevant and Dobzhansky, 1936) (see Figure 5).

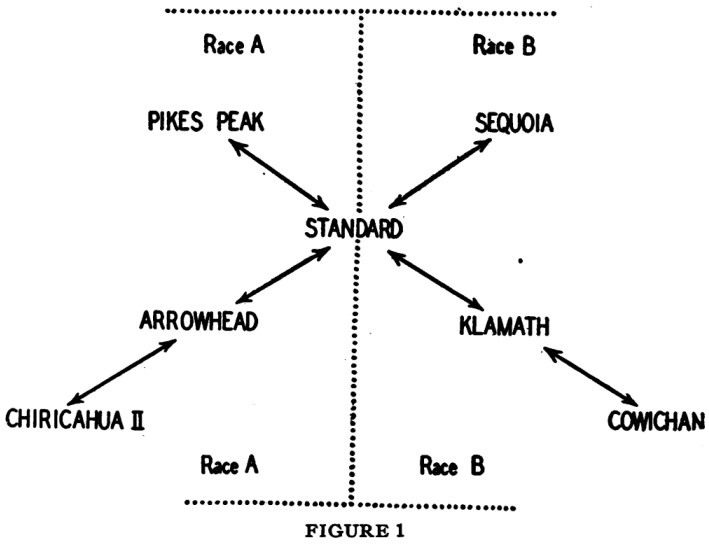

Fig. 5 A phylogeny of seven *Drosophila pseudoobscura* strains, from Sturtevant and Dobzhansky (1936). Reproduced with the tacit permission of the *National Academy of Science* of the United States of America.

Compared to building genetic maps, assessing homology using the cytology of polytene chromosomes was fast and much less costly, which partly explains its immediate and long-lasting success. However, for studying evolution over longer timescales, it was also somewhat limited. If there were more than three overlapping inversions on the same chromosome, the technique yielded almost no interpretable observations. Knowledge of all intermediary steps was required. Neverthelessr Dobzhansky and Powell (1975) followed by others (Carson and Kaneshiro, 1976) finally reconstructed a phylogeny of *Drosophila* species with more than one hundred arrangements and several hundred inver-

sions. Polytene chromosomes are still used to compare insects, and cytology has morphed into cytogenetics, with extremely productive results Carson and Kaneshiro (1976); Brehm (1990); Dutrillaux and Dutrillaux (2012). The problem of computing the inversion distance re-appears in some of these works, alongside with *ad-hoc* solutions Brehm (1990).

From the 1960s onward, it became possible to identify homologies between more distantly related organisms from sequence data. It was only in 1982, probably driven by the availability of genomic sequences, that the inversion problem was redefined in mathematical terms Watterson et al. (1982), without any reference to Sturtevant's papers. This time, this work inspired a long series of mathematical and computational studies Fertin et al. (2009).

3 Discussion

In this section three different aspects of the mathematical component of this historical work are discussed more in depth. First, the type of mathematics required to count inversions is discussed, in the context of a general mathematization of science. It involved the design of algorithms, in particular combinatorial algorithms, which were seldom used by evolutionary biologists, and even by mathematicians. Second, the role of errors in this history and in science in general is discussed. It is striking that the progressive quantification of the question, which aimed to reduce the part left to intuition, and consequently reduce the chance of errors, has in this case been the engine of errors. Third, in searching for the causes of these errors, a special mention needs to me made regarding the computational complexity of these mathematical problems. This illustrates the underestimated influence of intractability in some biology research programs.

3.1 Counting mutations as a computational biology problem

The introduction of measures, quantification methods, statistics and mathematics into evolutionary biology and phylogeny traversed the twentieth century (Hagen, 2003; Sommer, 2008; Suárez-Díaz and Anaya-Muñoz, 2008; Suárez-Díaz, 2010). This tendency is visible in biology, science and society in general (Kay, 1993; de Chadarevian and Kamminga, 1998; Porter, 1996). Several researchers saw this as a possibility to turn phylogeny into a *bona fide* science.

Computing evolutionary distances has been an important activity for establishing phylogenetic relationships. In the first half of the twentieth century this was done for example with serological and immunological reactions (de Chadarevian, 1996; Strasser, 2010b; Hagen, 2010), and later on with DNA hybridization (Suárez-Díaz, 2014). Biologists hoped that results would reflect the amount of divergence between proteins or chromosomes.

With the advent of molecular biology in the 1960s, discrete DNA mutations could be directly quantified (Hagen, 1999, 2003, 2010; Strasser, 2010b; Suárez-Díaz, 2014; Dietrich, 1994, 1998; Morgan, 1998; Sommer, 2008), with the use

of a particular type of mathematics, often aided by computers (Hagen, 2000, 2001; Strasser, 2010a).

The kind of mathematics used by Sturtevant for counting inversions is unusual in this regard. On the one hand, successive mutations in semantic characters were counted, in the same way that substitutions in protein or gene sequences are counted. In that sense, it is closer to the mathematics developed in the 1960s than to the quantifications performed in the 1930s, which was developed as a proxy, *i.e.* "waiting for sequences" (Hagen, 2010). On the other hand, there is a crucial difference between counting inversions and counting point mutations in sequences: as an approximation, the different sites of a sequence subject to point mutations can be considered independent from each other, while with overlapping inversions, gene arrangements are inaccessible to this simplifying hypothesis. These two aspects give a special status to this mathematical problem, and explain why counting inversions, although it precedes counting point mutations by 30 years, is still much less developed.

The technique for counting inversions involves the design of an algorithm. There is no known mathematical formula for estimating the inversion distance. The only way to calculate the inversion distance is to apply successive inversions to the permutation in order to come closer, one step at a time, to the alphabetical order. Sturtevant, his students and perhaps Ward, even if it is not specified in the publications, must have applied this type of method. As they proceeded to calculate the inversion distance for hundreds of permutations, they must have formalized a method. Indeed, not only did they perform the calculations for the permutations stemming from their biological data, but also from the complete set of permutations for up to 6 genes (720 permutations), plus a sample of dozens of permutations of seven to nine genes. They do not describe how they carried this out but admit their method had limits "For numbers of loci above nine the determination of this minimum number proved too laborious, and too uncertain, to be carried out" (Sturtevant and Novitski, 1941). This means that they were certain for permutations with up to nine loci, which is already, for some of them, a difficult exercise. We do not know how they came up with this confidence but we can only imagine they used an automatic method, *i.e.* an algorithm.

Algorithms have been used by mathematicians for a long time and were known to biologists. However their use as mathematical objects was not formalized and few mathematicians were specialized in designing algorithms. Turing's famous articles were published at the time when Sturtevant was carrying out his research (Turing, 1936). This absence of a constituted field with its own practices and applications explains why, despite having constructed an algorithm to solve the inversion problem on dozens of permutations, Sturtevant, Tan and Novitski did not even bother to describe it, even if it must have been a considerable endeavor.

Moreover, almost all algorithms known at the time were algorithms on continuous algebraic structures, allowing for example independence between dimensions and working with one dimension at a time (think of Gaussian elim-

ination for inverting a matrix, Euclid's algorithm for computing the greatest common divisor, or Fisher's Anova). The design of algorithms on combinatorial structures like permutations or graphs was developed in the second part of the 20th century (the description of finding the shortest path in a graph dates back to 1956).

Modern data has not changed this problem much. Even if the possibility of analyzing DNA sequences at the level of the entire chromosome has brought more data, more precision and more evolutionary depth, the principle behind chromosome comparison, unlike the detection of point mutations in genes, has not changed with the availability of sequences and still consists in counting inversions (Pevzner and Tesler, 2003; Murphy et al., 2005). Nevertheless recent developments have been numerous and gave rise to many variants of this problem. For example the possibility of knowing the reading direction of genes has unexpectedly decreased, to a small extent, algorithmic difficulties (Fertin et al., 2009).

To conclude this part by an anecdote, it is ironic that what we today consider a computational biology problem originates from the laboratory of Thomas Hunt Morgan, who allegedly had an aversion to computers. It is said that he banned Friden calculators from the biology department at Caltech, because he mistrusted all quantitative and automatic results[13]. If this attitude seems to run counter to history, the present narrative, made up of errors introduced at the same rhythm as the quantification, does not entirely prove him wrong.

3.2 The importance of errors

It is almost epistemologically impossible to retrace the history of an error. To present the genesis of knowledge while specifying that it is erroneous is already seeing it through the eyes of a subsequent event, that of its falsification. Writing about an error is in itself an anachronism.

On the other hand, placing errors on an equal footing as currently accepted knowledge, without specifying that they have been refuted, also poses an epistemological difficulty. The possibility of studying the history of science without disentangling what is the true from what is the false, according to a current view, is subject to debate Chabot (1999).

This could explain why histories of scientific errors are scarce (Firestein, 2015; Livio, 2014). Errors are often used for educational purposes (Bosch, 2018), to explain how not to make them. Or they can be a way to celebrate the discovery of the truth, by contrast. At best, scientific activity can be seen as a constant effort to track errors Popper (1959).

Nevertheless errors might also be a part, perhaps a major part, of scientific activity. The production of errors, and not their falsification, can be an interesting process. It is all the more interesting when considering the example

[13] This story is attributed to Charlie Munger in Belevin (2007).

described in this article, because the errors appeared and accumulated precisely at a time when quantification progressed, and probably were the result of quantification. Because we usually consider quantification as a process that reduces the possibility of errors resulting from subjectivity and intuition, it is remarkable that in this case the result was the opposite.

Let me remind the reader how, in this example, the accumulation of small errors have been led to a wrong conclusion. First in 1937 Sturtevant and Tan stated that the observed inversion distance of the gene arrangement on the X chromosome of *Drosophila* species was 7. At the time this was not an error because the authors were aware that this number could be 6, even if they did not find a scenario with 6 inversions. In 1941 Sturtevant and Novitski, with the help of Ward, confirmed the number 7 (error number one) and compared it with the expected inversion distance from random arrangements, calculated as 7.6, when the correct value is closer to 7.9. This value should not be considered as an error because it is the result of an interpolation and was not claimed to be the real value. However this value was compared to the observed inversion distance of 7, which falls into a standard deviation interval of "less than one" if centered on the expected value of 7.6. When considering the correct values, we come to the opposite conclusion: 6 does not fall within the standard deviation interval (0.85) centered on 7.9.

Therefore the error number two is to use approximate quantification, knowing but forgetting that they are approximations, to draw a conclusion from the statistical test. The robustness of the biological conclusion is not tested for the three approximations (approximation of the inversion distance, linear interpolation of the expected value, and upper bound of the standard deviation). In this case, the result with less quantification (the result from 1937) is closer to what can be concluded from the data than the result with more quantification (the result from 1941).

We might wonder how enlisting the help of a professional mathematician has had such a disastrous impact on the computations. In all probability Ward concentrated on what he knew best, *i.e.* statistics (computing an expectation and a standard deviation from a sample), and focused less on the problem that he —like everyone else— had no clue about, namely the computation of the inversion distance.

The addition of errors in the second publication is explained by the type of mathematics that we now know to be useful to handle the problem, which was unknown at the time. However the mathematics of counting inversions has hardly improved, because of the intrinsic difficulty of the problem, that is, its computational complexity. This intrinsic difficulty could account in part for this accumulation of errors, and might explain, more generally, the trajectory of some biological research programs.

3.3 The importance of computational complexity

The errors I have reported were obviously not the result of incompetence or poor intuition on the part of scientists involved. They could be due to a lack of real interest in the problem from their part. Indeed, assessing gene homology using genetic techniques was time-consuming, costly, and could not be automated or generalized to more distant species. This meant that large-scale research programs based on this technique had little chance of success. This may explain why the results of the comparison between *melanogaster* and *pseudoobscura* have not been reproduced for other species, and why the mathematical techniques have not been refined and the errors not corrected by additional work.

However several facts tend to contradict the idea that there was a lack of interest from the part of Sturtevant. Sturtevant requested the help of a professional mathematician despite being himself a decent amateur mathematician. Two publications, with two different Ph.D. students, published four years apart, mention the mathematical problem. In the latter, intuitive statements are abandoned for quantified statements. A supposed error in the first publication is corrected in the second. These facts suggest that Sturtevant was reasonably interested in obtaining the right answer to the problems he raised.

One of the reasons why he did not achieve this right answer at the time could be that the mathematical problem raised by successive overlapping inversions is intractable[14]. These problems contain an inherent provably difficulty which prevents the mathematical construction of any tractable solution[15]. Although biologists are often not aware of computational complexity, or do not consider it important, it is a constraint that can influence the direction of biological researches. The example given here illustrates the influence of such a constraint. Today computational sequence alignment tools detect point mutations but not inversions. This is due to the computational complexity of detecting inversions and not to an absence of inversions. In that case computational complexity could explain why some biological processes are extensively studied while others are much less quantified.

Acknowledgements

Thanks to Istvan Miklos for showing me the 1937 article by Sturtevant, and to Vincent Daubin and Bastien Boussau for giving me the opportunity to present part of this work at the Jacques Monod conference in 2016, “Molecules as documents of evolutionary history : 50 years after”. Thanks also to several

[14] Note that this contrasts with the history of protein sequence alignment, where it became possible to compare two related sequences without excessive mathematical involvement (see, for example, (Margoliash, 1963)). I am not saying that sequence alignment did not pose an interesting mathematical but it was inherently easier to solve with the intuitive ideas of biologists than computing an inversion distance.

[15] Finding the minimum number of inversions to transform a sequence of letters into alphabetical order is provably intractable (Caprara and Lancia, 2000).

anonymous historians who have kindly helped me improve the historical aspects of this article, find the relevant secondary literature and get rid of most teleological and anachronistic arguments.

References

Anderson, E. 1937, Jul. Cytology in its relation to taxonomy. *The Botanical Review 3*(7), 335–350.

Babcock, E. B. and G. L. Stebbins 1938. *The American species of Crepis: Their interrelationships and distribution as affected by polyploidy and apomixis.* Washington, D.C: Carnegie Institution of Washington.

Bafna, V. and P. A. Pevzner 1996, February. Genome rearrangements and sorting by reversals. *SIAM J. Comput. 25*(2), 272–289.

Belevin, P. 2007. *Seeking Wisdom: From Darwin to Munger.* PCA Publications.

Bosch, G. 2018, February. Train PhD students to be thinkers not just specialists. *Nature 554*(7692), 277–277.

Bowler, P. J. 2003, 07. *Evolution.* University of California Press.

Boyden, A. 1934. Precipitins and phylogeny in animals. *The American Naturalist 68*(719), 516–536.

Brehm, A. 1990. *Phylogénie de neuf espèces de Drosophila du groupe obscura d'après les homologies de segments des chromosomes polytènes.* Ph. D. thesis, Université de Lyon 1.

Bridges, C. B. 1917. Deficiency. *Genetics 2*, 445–465.

Camin, J. H. and R. R. Sokal 1965, sep. A method for deducing branching sequences in phylogeny. *Evolution 19*(3), 311–326.

Caprara, A. and G. Lancia 2000. Experimental and statistical analysis of sorting by reversals. In D. Sankoff and J. H. Nadeau (Eds.), *Comparative Genomics*, pp. 171–183. Springer.

Carson, H. L. and K. Y. Kaneshiro 1976. Drosophila of hawaii: systematics and ecological genetics. *Annual Review of Ecology and Systematics 7*(1), 311–345.

Castle, W. E. 1918, Feb. Is the arrangement of the genes in the chromosome linear? *Proc Natl Acad Sci U S A 5*(2), 25–32.

—— 1919, Nov. Are genes linear or non-linear in arrangement? *Proc Natl Acad Sci U S A 5*(11), 500–506.

Chabot, H. 1999. *Enquête historique sur les savoirs scientifiques rejetés a l'aube du positivisme (1750-1835).* Ph. D. thesis, Université de Nantes.

Darden, L. 2005, jun. Relations among fields: Mendelian, cytological and molecular mechanisms. *Studies in History and Philosophy of Science Part C: Studies in History and Philosophy of Biological and Biomedical Sciences 36*(2), 349–371.

de Chadarevian, S. 1996, Sep. Sequences, conformation, information: Biochemists and molecular biologists in the 1950s. *Journal of the History of Biology 29*(3), 361–386.

de Chadarevian, S. and H. Kamminga (Eds.) 1998. *Molecularizing Biology and Medicine New Practices and Alliances, 1920s to 1970s.* Taylor and Francis.

Dietrich, M. R. 1994, Mar. The origins of the neutral theory of molecular evolution. *Journal of the History of Biology 27*(1), 21–59.

—— 1998. Paradox and persuasion: negotiating the place of molecular evolution within evolutionary biology. *J Hist Biol 31*(1), 85–111.

Dietrich, M. R. 2016. History of molecular evolution. In *Encyclopedia of Evolutionary Biology.* Elsevier.

Dobzhansky, T. 1930. Translocations involving the third and the fourth chromosomes of drosophila melanogaster. *Genetics 15*(4), 347–399.

Dobzhansky, T. and J. R. Powell 1975. Drosophila pseudoobscura and its american relatives, drosophila persimilis and drosophila miranda. In R. King (Ed.), *Invertebrates of Genetic Interest*, pp. 537–587. Plenum Press.

Dobzhansky, T. and S. Wright 1941. Genetics of natural populations. v. relations between mutation rate and accumulation of lethals in populations of drosophila pseudoobscura. *Genetics 26*, 23–51.

Dutrillaux, A.-M. and B. Dutrillaux 2012. Chromosome analysis of 82 species of scarabaeoidea (coleoptera), with special focus on nor localization. *Cytogenetic and genome research 136*, 208–219.

Fertin, G., A. Labarre, I. Rusu, E. Tannier, and S. Vialette 2009. *Combinatorics of Genome Rearrangements.* London: MIT press.

Firestein, S. 2015. *Failure.* oxford university press.

Galvão, G. R. and Z. Dias 2015, January. An audit tool for genome rearrangement algorithms. *J. Exp. Algorithmics 19*, 1.7:1.1–1.7:1.34.

Gannett, L. and J. R. Greisemer 2004. Classical genetics and the geography of genes. In Rheinberger and Gaudilliere (Eds.), *Classical Genetic Research and Its Legacy*, pp. 57–88. London and New York: Routledge.

Hagen, J. B. 1982. *Experimental Taxonomy, 1930-1950: The Impact of Cytology, Ecology, and Genetics on Ideas of Biological Classification.* Ph. D. thesis, Oregon State University.

Hagen, J. B. 1984, Jun. Experimentalists and naturalists in twentieth-century botany: Experimental taxonomy, 1920–1950. *Journal of the History of Biology 17*(2), 249–270.

—— 1999. Naturalists, molecular biologists, and the challenges of molecular evolution. *Journal of the History of Biology 32*(2), 321–341.

—— 2000. The origins of bioinformatics. *Nature Reviews Genetics 1*(3), 231.

—— 2001. The introduction of computers into systematic research in the united states during the 1960s. *Stud Hist Phil Biol and Biomed Sci. 32*, 291–314.

—— 2003. The statistical frame of mind in systematic biology from quantitative zoology to biometry. *Journal of the History of Biology 36*(2), 353–384.

—— 2010. Waiting for sequences: Morris goodman, immunodiffusion experiments, and the origins of molecular anthropology. *Journal of the History of Biology 43*(4), 697–725.

Kay, L. E. 1993. *The Molecular Vision of Life: Caltech, The Rockefeller Foundation, and the Rise of the New Biology.* Oxford University press.

Kececioglu, J. and D. Sankoff 1995. Exact and approximation algorithms for sorting by reversals, with application to genome rearrangement. *Algorithmica 13*, 180–210.

Kohler, R. E. 1994. *Lords of the fly: Drosophila genetics and the experimental life*. University of Chicago Press.

Lehmer, D. H. 1993. The mathematical work of morgan ward. *Math. Comp. 61*, 307–311.

Livio, M. 2014. *Brilliant Blunders: From Darwin to Einstein - Colossal Mistakes by Great Scientists That Changed Our Understanding of Life and the Universe*. Brilliance Audio.

Margoliash, E. 1963. Primary structure and evolution of cytochrome c. *Proceedings of the National Academy of Sciences 50*(4), 672–679.

McClung, C. E. 1908. Cytology and taxonomy. *Kansas University Science Bulletin 4*(7), 199–215.

Metz, C. W. 1914. Chromosome studies in the diptera. i. a preliminary survey of five different types of chromosome groups in the genus drosophila. *Journal of Experimental Zoology Part A: Ecological Genetics and Physiology 17*(1), 45–59.

—— 1916. Chromosome studies on the Diptera. III. additional types of chromosome groups in the Drosophilidae. *The American Naturalist 50*(598), 587–599.

—— 1918. Chromosome studies on the Diptera. *Zeitschrift für induktive Abstammungs-und Vererbungslehre 19*(3), 211–213.

Mohr, O. L. 1919. Character changes caused by mutation of an entire region of a chromosome in drosophila. *Genetics 4*, 275–282.

Morgan, G. J. 1998. Emile Zuckerkandl, Linus Pauling, and the molecular evolutionary clock, 1959-1965. *J Hist Biol 31*(2), 155–178.

Morgan, T. H. and C. B. Bridges 1916. *Sex-linked inheritance in Drosophila*. Carnegie Inst. Washington, Publ.

Morgan, T. H., C. B. Bridges, and A. H. Sturtevant 1925. *The Genetics of Drosophila*. Bibliographia Genetica.

Morgan, T. H., A. H. Sturtevant, and C. B. Bridges 1920. The evidence for the linear order of the genes. *Proc Natl Acad Sci U S A 6*(4), 162–164.

Muller, H. J. 1929. The first cytological demonstration of a translocation in drosophila. *The American Naturalist 63*(689), 481–486.

Murphy, W. J., D. M. Larkin, A. Everts-van der Wind, G. Bourque, G. Tesler, L. Auvil, J. E. Beever, B. P. Chowdhary, F. Galibert, L. Gatzke, C. Hitte, S. N. Meyers, D. Milan, E. A. Ostrander, G. Pape, H. G. Parker, T. Raudsepp, M. B. Rogatcheva, L. B. Schook, L. C. Skow, M. Welge, J. E. Womack, S. J. O'brien, P. A. Pevzner, and H. A. Lewin 2005, Jul. Dynamics of mammalian chromosome evolution inferred from multispecies comparative maps. *Science 309*(5734), 613–617.

Novitski, E. 2005. *Sturtevant and Dobzhansky: Two Scientists at Odds, With a Student's Recollections*. Bloomington: Xlibris Corporation.

Painter, T. S. 1933. A new method for the study of chromosome rearrangements and the plotting of chromosome maps. *Science 78*, 585–586.

—— 1934. Salivary chromosomes and the attack on the gene. *Journal of Heredity 25*(12), 465–476.

Papadimitriou, C. H. 1993. *Computational Complexity.* Pearson.

Pevzner, P. and G. Tesler 2003, Jan. Genome rearrangements in mammalian evolution: lessons from human and mouse genomes. *Genome Res 13*(1), 37–45.

Popper, K. 1959. *The Logic of Scientific Discovery.* Abingdon-on-Thames: Routledge.

Porter, T. M. 1996. *Trust in numbers.* Princeton University Press.

Smocovitis, V. B. 2006. Keeping up with dobzhansky: G. ledyard stebbins, jr., plant evolution, and the evolutionary synthesis. *Hist. Phil. Life Sci., 28*, 9–48.

—— 2009, aug. The "Plant Drosophila": E. b. babcock, the GenusCrepis, and the evolution of a genetics research program at berkeley, 1915–1947. *Historical Studies in the Natural Sciences 39*(3), 300–355.

Sommer, M. 2008, Sep. History in the gene: Negotiations between molecular and organismal anthropology. *Journal of the History of Biology 41*(3), 473–528.

Strasser, B. J. 2010a. Collecting, comparing, and computing sequences: the making of Margaret O. Dayhoff's atlas of protein sequence and structure, 1954–1965. *Journal of the History of Biology 43*(4), 623–660.

—— 2010b. Laboratories, museums, and the comparative perspective: Alan A. Boyden's quest for objectivity in serological taxonomy, 1924-1962. *Hist Stud Nat Sci 40*(2), 149–182.

Sturtevant, A. H. 1921, Aug. A case of rearrangement of genes in drosophila. *Proc Natl Acad Sci U S A 7*(8), 235–237.

—— 1942. The classification of the genus drosophila, with descriptions of nine new species. *Austin: The University of Texas Publication 4213*, 5–51.

Sturtevant, A. H., C. B. Bridges, and T. H. Morgan 1919, May. The spatial relations of genes. *Proc Natl Acad Sci U S A 5*(5), 168–173.

Sturtevant, A. H. and T. Dobzhansky 1936, Jul. Inversions in the third chromosome of wild races of drosophila pseudoobscura, and their use in the study of the history of the species. *Proc Natl Acad Sci U S A 22*(7), 448–450.

Sturtevant, A. H. and E. Novitski 1941. The homologies of chromosome elements in the genus drosophila. *Genetics 26*, 517–541.

Sturtevant, A. H. and C. R. Plunkett 1926. Sequence of corresponding third-chromosome genes in drosophila melanogaster and d. simulans. *Biol Bull 50*, 56–60.

Sturtevant, A. H. and C. C. Tan 1937. The comparative genetics of Drosophila Pseudoobscura and D. Melanogaster. *Journal of Genetics 34*, 415–432.

Suárez-Díaz, E. 2009, Mar. Molecular evolution: concepts and the origin of disciplines. *Stud Hist Philos Biol Biomed Sci 40*(1), 43–53.

—— 2010. Making room for new faces: evolution, genomics and the growth of bioinformatics. *History and Philosophy of the Life Sciences 32*(1), 65–89.

—— 2014, Aug. The long and winding road of molecular data in phylogenetic analysis. *Journal of the History of Biology 47*(3), 443–478.

Suárez-Díaz, E. and V. H. Anaya-Muñoz 2008, Dec. History, objectivity, and the construction of molecular phylogenies. *Stud Hist Philos Biol Biomed Sci 39*(4), 451–468.

Turing, A. M. 1936. On computable numbers, with an application to the entscheidungsproblem. *Proceedings of the London mathematical society 2*(1), 230–265.

Turrill, W. B. 1938. The expansion of taxonomy with special reference to spermatophyta. *BioL Rev. 13*, 342–373.

Watterson, G., W. Ewens, T. Hall, and A. Morgan 1982, nov. The chromosome inversion problem. *Journal of Theoretical Biology 99*(1), 1–7.

Zuckerkandl, E. and L. Pauling 1965, Mar. Molecules as documents of evolutionary history. *J Theor Biol 8*(2), 357–366.