



**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.

1 **The curious case of a hapless mathematical**  
2 **contribution to biology**

3 **Chromosome inversions in *Drosophila*, 1937-1941**

4 **Eric Tannier**

5 **Abstract** This is the story, told in the light of a new analysis of historical  
6 data, of a mathematical biology problem that was explored in the 1930s in  
7 Thomas Morgan's laboratory at the California Institute of Technology. It is one  
8 of the early developments of evolutionary genetics and quantitative phylogeny,  
9 and deals with the identification and counting of chromosomal rearrangements  
10 in *Drosophila* species from comparisons of genetic maps. A re-analysis of the  
11 data produced in the 1930s using current mathematics and computational  
12 technologies reveals how a team of biologists, with the help of a renowned  
13 mathematician and against their first intuition, came to an erroneous conclusion  
14 regarding the presence of phylogenetic signals in gene arrangements. This  
15 example illustrates two different aspects of a same piece: 1) the appearance of  
16 a mathematical in biology problem solved with the development of a combi-  
17 natorial algorithm, which was unusual at the time, and 2) the role of errors in  
18 scientific activity. Also underlying is the possible influence of computational  
19 complexity in understanding the directions of research in biology.  
20

21 **Keywords** history of biology · evolutionary genetics · chromosomal inversion ·  
22 genetic maps · statistics · computational complexity · scientific errors · history  
23 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: +33-45-678910  
E-mail: eric.tannier@inria.fr

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

28 Theodosius Dobzhansky, letter to Milislav Demerec 1936

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

32 Alfred Sturtevant, letter to Otto Mohr 1939

33 These two quotes attest to the renewed interest in phylogeny during the  
 34 first half of the twentieth century. Marks of enthusiasm such as these, associ-  
 35 ated with the revival of this old discipline, were common. Among other possible  
 36 reasons, they are due, on the one hand, to the use of cytological and genetic  
 37 comparisons, offering direct access to hereditary material, and on the other,  
 38 to the use of quantified methods, often associated with objectivity. According  
 39 to Anderson (1937), cytology was like “looking at the cellar window”, and is  
 40 “evidence as to the germplasm itself and is, therefore, of more fundamental  
 41 importance than the mere architecture erected by the germplasm itself.” For  
 42 Turrill (1938), chromosomes provided “high-powered morphology”. For Mc-  
 43 Clung (1908), “The chromosomes are the determinants of characters”, and  
 44 “one cell is sufficient for the identification of the species”. “Were our knowl-  
 45 edge of cell structure in the grasshopper complete enough we might erect a  
 46 system of classification based upon cytological characters, just as reasonably  
 47 as we have designated one using external anatomical structures” (McClung,  
 48 1908). As for quantification, the comparisons allowed by precipitin reactions  
 49 (Strasser, 2010b) made Boyden (1934) write that “The fact that naturalists  
 50 of recent times have so often forsaken the study of phylogeny is due more to  
 51 the feeling that such a study is likely to yield little certain progress than to  
 52 the belief that the problems of phylogeny are unimportant or sufficiently well  
 53 analyzed.”

54 Of course, the use of both “semantic”<sup>1</sup> characters and quantification, driven  
 55 by the development of sequencing techniques and computers, was only fully  
 56 realized in the 1960s by the founders of Molecular Evolution (Suárez-Díaz,  
 57 2009; Dietrich, 2016). However the evolutionary genetics program that began  
 58 in Thomas Morgan’s laboratory in 1914, and was subsequently continued by  
 59 the partners turned rivals Alfred Sturtevant and Theodosius Dobzhansky, had  
 60 similar epistemological characteristics<sup>2</sup>.

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

<sup>1</sup> 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).

<sup>2</sup> Despite crucial differences in the biological objects have also been discussed (Darden, 2005).

65 of this research, in particular the attempts to quantify evolutionary diver-  
66 gence, the involvement of the mathematician Morgan Ward, and the errors  
67 that resulted, have been overlooked in historical accounts of the study of chro-  
68 mosome evolution (Hagen, 1982, 1984; Kohler, 1994; Gannett and Greisemer,  
69 2004; Smocovitis, 2006) and of the use of quantification in biology (Hagen,  
70 2003; Suárez-Díaz and Anaya-Muñoz, 2008; Suárez-Díaz, 2010; Hagen, 2010).

71 In the 1930s the use of mathematics, and collaborations with mathemati-  
72 cians was commonplace in biology, and particularly in evolutionary biology.  
73 It was even central part to the construction of the modern synthesis (Bowler,  
74 2003). However the type of mathematics in this example is unusual in that  
75 it differs from that available to evolutionists, as developed for instance by  
76 Fisher, Wright or Haldane for statistics and population genetics. Retrospec-  
77 tively combinatorial and computational aspects are visible, which were handled  
78 at the time with underlying<sup>3</sup> systematically applied algorithms on permuta-  
79 tions. Some of the questions addressed at the time were only solved 50 years  
80 later, and some even remain unsolved today. The difficulties that mathemati-  
81 cians encounter today with these problems were already visible in Sturtevant's  
82 attempts. Nevertheless, after trying to solve the same questions myself with  
83 today's mathematics and technology, I found three computational and nu-  
84 merical approximations, initially acknowledged as such by the authors, which,  
85 after consulting Ward, strangely turned into errors and led to a wrong bio-  
86 logical interpretation. This curious case of an unfortunate use of mathematics  
87 to solve an evolutionary question illustrates the presence and importance of  
88 errors in the practice of science. We could also see it as an example of the of-  
89 ten overlooked impact of computational intractability (Papadimitriou, 1993)  
90 on biological research.

91 In the first part of this article, I provide some contextual elements con-  
92 cerning the use of chromosomes in evolutionary studies, both worldwide and  
93 in Thomas Morgan's laboratory, in order to highlight the originality of Sturte-  
94 vant's research. In the second part I describe how Sturtevant progressed from  
95 making the first genetic map to the challenge of counting inversions. Along-  
96 side historical descriptions, I give my solutions to the described problems using  
97 current scientific knowledge. In the third part, I discuss what this exercise can  
98 teach us about the unexpected presence of combinatorial algorithmic consid-  
99 erations in 1930s biology, and about the influence of errors and complexity in  
100 both past and present research programs.

## 101 1 Chromosomes as documents of evolutionary history

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

---

<sup>3</sup> *I.e* not explicit

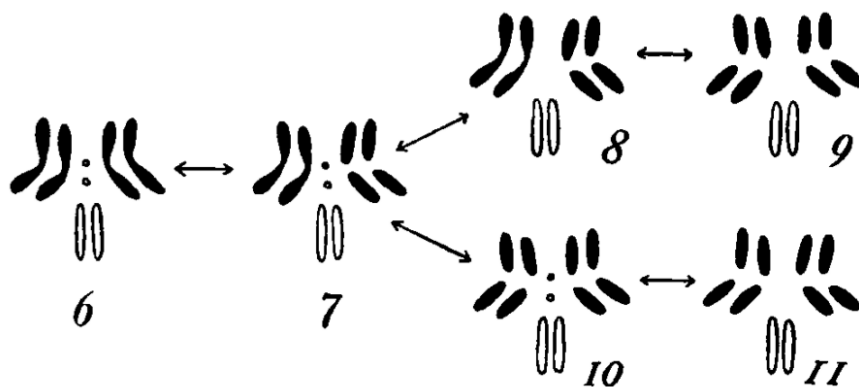
106 classify species or construct phylogenies, such as: the number, shape and size  
107 of chromosomes, their behavior during the cell cycle, the position of the cen-  
108 tromere or the arrangement of genes. To cite only a few landmarks of this  
109 development: at the International Zoological Congress in Boston in 1907, the  
110 cytologist Clarence Erwin McClung stated that a character measured within  
111 the cell, such as the number of chromosomes, could be considered as informa-  
112 tive for phylogenetic classification as any morphological character (McClung,  
113 1908). In 1915 in Berkeley, California, the plant geneticist Ernest Brown Bab-  
114 cock gathered a team to work on the evolution of the flowering plant *Crepis*  
115 and contributed to the foundation of the “Bay Area Biosystematists” (Ha-  
116 gen, 1984; Smocovitis, 2009), an influential multidisciplinary group working  
117 on plant systematics. In 1926, the International Congress of Plant Science  
118 held a joint session involving taxonomists, cytologists and geneticists (Hagen,  
119 1984). In 1937, the field was sufficiently established for Edgar Anderson, from  
120 the Missouri Botanical Garden, to write an extensive review on the contribu-  
121 tion of cytology to taxonomy in botany (Anderson, 1937). In 1938, Babcock  
122 and his collaborator George Ledyard Stebbins Jr, who would become a leading  
123 evolutionary biologist (Smocovitis, 2006), published the influential book *The*  
124 *American Species of Crepis*, in which all the genetic and cytological knowledge  
125 available at the time was harnessed to decipher the complex evolutionary re-  
126 lationships between members of the the *Crepis* genus (Babcock and Stebbins,  
127 1938; Smocovitis, 2009).

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

141 In article published in 1914, from which Figure 1 is reproduced, Metz  
142 speculated that differences in chromosome types “may indicate an evolution  
143 of chromosomes in the genus” (Metz, 1914). However, in his subsequent ar-  
144 ticles on the description of chromosome types, Metz became more and more  
145 cautious regarding any possible evolutionary interpretation (Kohler, 1994),  
146 mainly because of the difficulty in assessing the homology<sup>4</sup> between chromo-  
147 somes via the technique of independent observation in different species. As

---

<sup>4</sup> 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*.

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

150 Then began the search for a technique to assess homology. One method  
151 was to produce interspecific hybrids and observe coupled chromosomes during  
152 segregation, but this showed little success with *Drosophila* species (Kohler,  
153 1994). Hybrids could be produced but were almost always sterile. Two sub-  
154 sequent techniques would prove more successful for assessing homology and  
155 were explored in Morgan's laboratory: gene mapping on chromosomes (from  
156 1917) and hybridization of polytene chromosomes (from 1936).

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

## 163 **2 Alfred Sturtevant and Comparative Genetic Mapping, 1921 to** 164 **1941**

### 165 2.1 Genetic maps and predicting inversions

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

171 this idea, Sturtevant defined genetic distance as the percentage of crossing-  
172 over between two genes, which he observed from the frequency of associated  
173 phenotypes in *Drosophila ampelophila*<sup>5</sup>. As this distance was close to a linear  
174 function, it was possible to position genes on a line. This led to the first genetic  
175 map, which placed six genes on the “sex-linked” linkage group<sup>6</sup> (Gannett and  
176 Greisemer, 2004).

177 Following Sturtevant’s, the same research group produced several other  
178 genetic maps. In particular, Morgan and Bridges’ 36-marker map of the X-  
179 chromosome of *Drosophila melanogaster* (Morgan and Bridges, 1916) was dis-  
180 puted by William Castle (Castle, 1918, 1919) and, by association, several other  
181 researchers, who questioned the relevance of the linear model for depicting  
182 chromosomes, with responses by Sturtevant et al. (1919); Morgan et al. (1920).  
183 Even though each protagonist gave the impression of standing firm on his re-  
184 spective position, the controversy helped clarify much of the theory, as well as  
185 its underlying and *ad hoc* hypotheses.

186 The real starting point for evolutionary genetic studies was the discovery  
187 of mutations in that linear structure. Inversions, *i.e.* evolutionary events re-  
188 versing the orientation of chromosome segments, were hypothesized by Sturte-  
189 vant (1921) based on the observation of differences in the arrangement of five  
190 “corresponding”<sup>7</sup> genes on chromosome 3 between *Drosophila simulans* and  
191 *Drosophila melanogaster*. The inversion hypothesis was confirmed by adding  
192 genes, while the comparative mapping carried out by Sturtevant and Plunkett  
193 (1926), illustrated in Figure 2, presents a striking visual argument in support  
194 of it<sup>8</sup>.

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

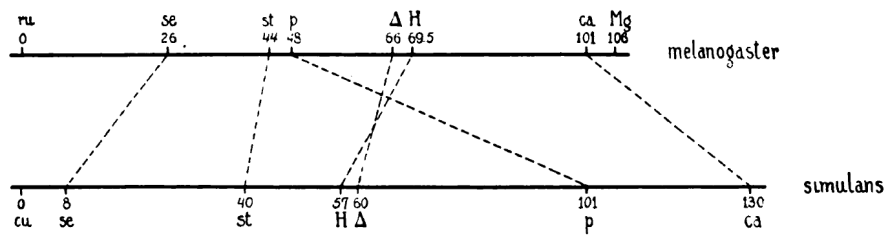
199 From this possibility of detecting mutations by comparing genetic maps,  
200 Sturtevant developed a comprehensive research project in continuity with his  
201 work with Metz. The aim was to map the genes of different *Drosophila* species,  
202 assess the homology between these genes and, from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

<sup>5</sup> Renamed *melanogaster* shortly after.

<sup>6</sup> 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.

<sup>7</sup> *I.e.* homologous, see footnote 3. Homology was deduced from the similarity of phenotype variations during crossing experiments.

<sup>8</sup> 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*.

205 results<sup>9</sup> confirm they made decisive advances as well as reveal some challenges.  
 206 A close look at the mathematical techniques they use helps us understand the  
 207 progressive introduction of quantification, and how, if it gives the impression  
 208 of objectivity and fights a “methodological anxiety” (Suárez-Díaz and Anaya-  
 209 Muñoz, 2008), it is not necessarily a guarantee of greater veracity or accuracy.

## 210 2.2 The first mathematical problem: the observed inversion distance

211 One of the challenges of studying chromosomal arrangements involves the de-  
 212 tection of several successive overlapping inversions. Comparing two arrange-  
 213 ments that differ by one inversion was easy. However if several overlapping  
 214 inversions have occurred, which is likely if more distant species were com-  
 215 pared, an additional difficulty arose. In 1937, Sturtevant, published with C.  
 216 C. Tan, a Ph.D. student supervised by himself and Dobzhansky, a compari-  
 217 son of the arrangements of 38 genes along all the chromosomes of *Drosophila*  
 218 *melanogaster* and *Drosophila pseudoobscura* (Sturtevant and Tan, 1937). The  
 219 comparative maps, inferred from the homology of genes deduced from similar  
 220 phenotypic effects, are reproduced in Figure 3. Inversions are not as visible as  
 221 in Figure 2 because the species are more distant and thus the accumulation of  
 222 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

<sup>9</sup> 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:

$$\underline{DEFACB} \rightarrow \underline{AFEDCB} \rightarrow ABCDEF.$$

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

### 229 2.3 Resolution with modern mathematics

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

$\underline{LHFEBADCKIJGM}$   
 $\rightarrow \underline{ABEFHLDCKIJGM}$   
 $\rightarrow \underline{ABCDLHFEKIJGM}$   
 $\rightarrow \underline{ABCDEFHLKIJGM}$   
 $\rightarrow \underline{ABCDEFHGJIKLM}$   
 $\rightarrow \underline{ABCDEFGHJIKLM}$   
 $\rightarrow \underline{ABCDEFGHIJKLM}$

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

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

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

#### 266 2.4 The second mathematical problem: the expected inversion distance

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

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

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

## 291 2.5 The call for a professional mathematician

292 After going over the statements of Sturtevant and Tan (those reproduced in  
293 Figure 3), Sturtevant and Novitski announced that they had solicited the help  
294 of Morgan Ward, a renowned mathematician from Caltech. Sturtevant himself  
295 had a reasonable understanding of mathematics, and Novitski (2005) retro-  
296 spectively praised his “mathematical mind”, compared with Dobzhansky’s.  
297 However Sturtevant and Novitski probably felt that no easy technique could  
298 solve this question and logically solicited the help of an expert.

299 Morgan Ward (1901-1963) entered Caltech in 1924 as a student, and be-  
300 came a research fellow in 1928. Appreciated by many for his qualities as a  
301 teacher, he apparently showed no particular interest in biology, though an ac-  
302 knowledgement can also be found in an article by Dobzhansky and Wright  
303 (1941), the only other biology paper, alongside the one studied here, featuring  
304 his name. He was more interested in the contribution of his field to physics.  
305 His expertise in number theory and Diophantine equations (Lehmer, 1993),  
306 which involved the design of calculation methods on integer numbers, might  
307 have convinced Sturtevant and Novitski to request his help. The exact mode  
308 of collaboration is not known: it is just mentioned in the middle of the article  
309 that Ward provided some help. We can suppose one or a few work sessions,  
310 where the two mathematical problems, that of the inversion distance and its  
311 statistical significance, were exposed and ways to compute the solutions were  
312 discussed.

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

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

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

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

---

<sup>10</sup> 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.

<sup>11</sup> 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*.

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

## 361 2.6 When the progression of quantification leads to a succession of errors

362 To recap, there are three small errors or approximations in the 1937 and 1941  
 363 articles: the minimum number of inversions (seven required instead of six), the  
 364 expected number (7.6 instead of 7.9) and the standard deviation (“less than  
 365 1” instead of 0.85). Taken together, these change the conclusion. With the  
 366 corrected calculations, six inversions would have been considered significantly  
 367 different from 7.9, falling outside the 0.85 standard deviation interval<sup>12</sup>.

<sup>12</sup> 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

368 It is retrospectively mind-boggling that Sturtevant and Novitski (1941), as-  
369 sisted by a mathematician, claimed in 1941 to correct the statement of Sturte-  
370 vant and Tan (1937), while in actual fact they were confirming the only wrong  
371 statement of the earlier article, and introduced another error. In 1937 the  
372 authors were cautious about the inversion distance number they found, but  
373 in 1941 they noted that “this revision does not change the number of inver-  
374 sions required to transform one sequence into the other”, thus retaining the  
375 erroneous number and ignoring their initial reservations. The help of a math-  
376 ematician, which ordinarily would have been considered a good idea for such  
377 a problem, has perhaps been disastrous in this case, as it undermined, for the  
378 wrong reasons, the sound intuitions that the biologists initially had.

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

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

## 401 2.7 Epilogue

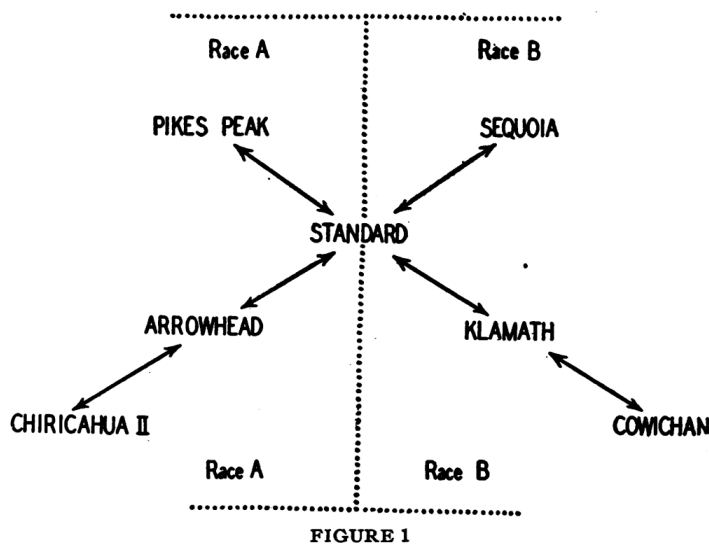
402 Of course the evolutionary genetics project started by Metz and Sturtevant  
403 in 1914 did not stop because of a few mathematical errors that were made  
404 in the 1930s and 1940s. One important challenge of their project was that it  
405 necessitated a prohibitive amount of work to assess the homology of genes and  
406 chromosome segments. In the articles analysed here (Sturtevant and Tan, 1937;  
407 Sturtevant and Novitski, 1941), a catalog of homologous genes was compiled  
408 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.

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

411 The technique invented by Theophilus Painter (Painter, 1933, 1934) to de-  
412 tect homologies between polytene chromosomes was exploited by geneticists  
413 and cytologists within a comparative and evolutionary framework (Gannett  
414 and Greisemer, 2004). At Caltech, it was used by Theodosius Dobzhansky,  
415 first in association with Sturtevant and then independently after their part-  
416 nership ended. Dobzhansky collected many *pseudoobscura* strains from all over  
417 the United States, while Sturtevant collected what he felt was interesting for  
418 genetics and, in particular, for his long-standing project of comparing chro-  
419 mosomes from an evolutionary perspective. In 1936, they published together  
420 the first phylogenetic tree based on chromosomal inversions (Sturtevant and  
421 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.

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

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

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

### 442 3 Discussion

443 In this section three different aspects of the mathematical component of this  
444 historical work are discussed more in depth. First, the type of mathematics  
445 required to count inversions is discussed, in the context of a general mathema-  
446 tization of science. It involved the design of algorithms, in particular combina-  
447 torial algorithms, which were seldom used by evolutionary biologists, and even  
448 by mathematicians. Second, the role of errors in this history and in science  
449 in general is discussed. It is striking that the progressive quantification of the  
450 question, which aimed to reduce the part left to intuition, and consequently  
451 reduce the chance of errors, has in this case been the engine of errors. Third,  
452 in searching for the causes of these errors, a special mention needs to be made  
453 regarding the computational complexity of these mathematical problems. This  
454 illustrates the underestimated influence of intractability in some biology re-  
455 search programs.

#### 456 3.1 Counting mutations as a computational biology problem

457 The introduction of measures, quantification methods, statistics and mathe-  
458 matics into evolutionary biology and phylogeny traversed the twentieth cen-  
459 tury (Hagen, 2003; Sommer, 2008; Suárez-Díaz and Anaya-Muñoz, 2008; Suárez-  
460 Díaz, 2010). This tendency is visible in biology, science and society in general  
461 (Kay, 1993; de Chadarevian and Kamminga, 1998; Porter, 1996). Several re-  
462 searchers saw this as a possibility to turn phylogeny into a *bona fide* science.

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

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



472 of a particular type of mathematics, often aided by computers (Hagen, 2000,  
473 2001; Strasser, 2010a).

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

488 The technique for counting inversions involves the design of an algorithm.  
489 There is no known mathematical formula for estimating the inversion distance.  
490 The only way to calculate the inversion distance is to apply successive inver-  
491 sions to the permutation in order to come closer, one step at a time, to the  
492 alphabetical order. Sturtevant, his students and perhaps Ward, even if it is not  
493 specified in the publications, must have applied this type of method. As they  
494 proceeded to calculate the inversion distance for hundreds of permutations,  
495 they must have formalized a method. Indeed, not only did they perform the  
496 calculations for the permutations stemming from their biological data, but also  
497 from the complete set of permutations for up to 6 genes (720 permutations),  
498 plus a sample of dozens of permutations of seven to nine genes. They do not  
499 describe how they carried this out but admit their method had limits “For  
500 numbers of loci above nine the determination of this minimum number proved  
501 too laborious, and too uncertain, to be carried out” (Sturtevant and Novitski,  
502 1941). This means that they were certain for permutations with up to nine  
503 loci, which is already, for some of them, a difficult exercise. We do not know  
504 how they came up with this confidence but we can only imagine they used an  
505 automatic method, *i.e.* an algorithm.

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

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

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

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

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

### 541 3.2 The importance of errors

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

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

551 This could explain why histories of scientific errors are scarce (Firestein,  
552 2015; Livio, 2014). Errors are often used for educational purposes (Bosch,  
553 2018), to explain how not to make them. Or they can be a way to celebrate  
554 the discovery of the truth, by contrast. At best, scientific activity can be seen  
555 as a constant effort to track errors Popper (1959).

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

---

<sup>13</sup> This story is attributed to Charlie Munger in Belevin (2007).

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

564 Let me remind the reader how, in this example, the accumulation of small  
565 errors have been led to a wrong conclusion. First in 1937 Sturtevant and Tan  
566 stated that the observed inversion distance of the gene arrangement on the X  
567 chromosome of *Drosophila* species was 7. At the time this was not an error  
568 because the authors were aware that this number could be 6, even if they did  
569 not find a scenario with 6 inversions. In 1941 Sturtevant and Novitski, with  
570 the help of Ward, confirmed the number 7 (error number one) and compared it  
571 with the expected inversion distance from random arrangements, calculated as  
572 7.6, when the correct value is closer to 7.9. This value should not be considered  
573 as an error because it is the result of an interpolation and was not claimed to  
574 be the real value. However this value was compared to the observed inversion  
575 distance of 7, which falls into a standard deviation interval of "less than one" if  
576 centered on the expected value of 7.6. When considering the correct values, we  
577 come to the opposite conclusion: 6 does not fall within the standard deviation  
578 interval (0.85) centered on 7.9.

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

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

593 The addition of errors in the second publication is explained by the type  
594 of mathematics that we now know to be useful to handle the problem, which  
595 was unknown at the time. However the mathematics of counting inversions has  
596 hardly improved, because of the intrinsic difficulty of the problem, that is, its  
597 computational complexity. This intrinsic difficulty could account in part for  
598 this accumulation of errors, and might explain, more generally, the trajectory  
599 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<sup>14</sup>. These problems contain an inherent provably difficulty which prevents the mathematical construction of any tractable solution<sup>15</sup>. 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

<sup>14</sup> 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 problem but it was inherently easier to solve with the intuitive ideas of biologists than computing an inversion distance.

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

636 anonymous historians who have kindly helped me improve the historical as-  
637 pects of this article, find the relevant secondary literature and get rid of most  
638 teleological and anachronistic arguments.

## 639 References

- 640 Anderson, E. 1937, Jul. Cytology in its relation to taxonomy. *The Botanical*  
641 *Review* 3(7), 335–350.
- 642 Babcock, E. B. and G. L. Stebbins 1938. *The American species of Crepis: Their*  
643 *interrelationships and distribution as affected by polyploidy and apomixis.*  
644 Washington, D.C: Carnegie Institution of Washington.
- 645 Bafna, V. and P. A. Pevzner 1996, February. Genome rearrangements and  
646 sorting by reversals. *SIAM J. Comput.* 25(2), 272–289.
- 647 Belevin, P. 2007. *Seeking Wisdom: From Darwin to Munger.* PCA Publica-  
648 tions.
- 649 Bosch, G. 2018, February. Train PhD students to be thinkers not just special-  
650 ists. *Nature* 554(7692), 277–277.
- 651 Bowler, P. J. 2003, 07. *Evolution.* University of California Press.
- 652 Boyden, A. 1934. Precipitins and phylogeny in animals. *The American Natu-*  
653 *ralist* 68(719), 516–536.
- 654 Brehm, A. 1990. *Phylogénie de neuf espèces de Drosophila du groupe obscura*  
655 *d’après les homologues de segments des chromosomes polytènes.* Ph. D. the-  
656 sis, Université de Lyon 1.
- 657 Bridges, C. B. 1917. Deficiency. *Genetics* 2, 445–465.
- 658 Camin, J. H. and R. R. Sokal 1965, sep. A method for deducing branching  
659 sequences in phylogeny. *Evolution* 19(3), 311–326.
- 660 Caprara, A. and G. Lancia 2000. Experimental and statistical analysis of  
661 sorting by reversals. In D. Sankoff and J. H. Nadeau (Eds.), *Comparative*  
662 *Genomics*, pp. 171–183. Springer.
- 663 Carson, H. L. and K. Y. Kaneshiro 1976. Drosophila of hawaii: systematics  
664 and ecological genetics. *Annual Review of Ecology and Systematics* 7(1),  
665 311–345.
- 666 Castle, W. E. 1918, Feb. Is the arrangement of the genes in the chromosome  
667 linear? *Proc Natl Acad Sci U S A* 5(2), 25–32.
- 668 ——— 1919, Nov. Are genes linear or non-linear in arrangement? *Proc Natl*  
669 *Acad Sci U S A* 5(11), 500–506.
- 670 Chabot, H. 1999. *Enquête historique sur les savoirs scientifiques rejetés a*  
671 *l’aube du positivisme (1750-1835).* Ph. D. thesis, Université de Nantes.
- 672 Darden, L. 2005, jun. Relations among fields: Mendelian, cytological and  
673 molecular mechanisms. *Studies in History and Philosophy of Science Part*  
674 *C: Studies in History and Philosophy of Biological and Biomedical Sci-*  
675 *ences* 36(2), 349–371.
- 676 de Chadarevian, S. 1996, Sep. Sequences, conformation, information: Bio-  
677 chemists and molecular biologists in the 1950s. *Journal of the History of*  
678 *Biology* 29(3), 361–386.

- 679 de Chadarevian, S. and H. Kammaing (Eds.) 1998. *Molecularizing Biology and*  
680 *Medicine New Practices and Alliances, 1920s to 1970s*. Taylor and Francis.
- 681 Dietrich, M. R. 1994, Mar. The origins of the neutral theory of molecular  
682 evolution. *Journal of the History of Biology* 27(1), 21–59.
- 683 ——— 1998. Paradox and persuasion: negotiating the place of molecular evolu-  
684 tion within evolutionary biology. *J Hist Biol* 31(1), 85–111.
- 685 Dietrich, M. R. 2016. History of molecular evolution. In *Encyclopedia of*  
686 *Evolutionary Biology*. Elsevier.
- 687 Dobzhansky, T. 1930. Translocations involving the third and the fourth chro-  
688 mosomes of drosophila melanogaster. *Genetics* 15(4), 347–399.
- 689 Dobzhansky, T. and J. R. Powell 1975. Drosophila pseudoobscura and its  
690 american relatives, drosophila persimilis and drosophila miranda. In R. King  
691 (Ed.), *Invertebrates of Genetic Interest*, pp. 537–587. Plenum Press.
- 692 Dobzhansky, T. and S. Wright 1941. Genetics of natural populations. v. rela-  
693 tions between mutation rate and accumulation of lethals in populations of  
694 drosophila pseudoobscura. *Genetics* 26, 23–51.
- 695 Dutrillaux, A.-M. and B. Dutrillaux 2012. Chromosome analysis of 82 species  
696 of scarabaeoidea (coleoptera), with special focus on nor localization. *Cyto-*  
697 *genetic and genome research* 136, 208–219.
- 698 Fertin, G., A. Labarre, I. Rusu, E. Tannier, and S. Vialette 2009. *Combina-*  
699 *torics of Genome Rearrangements*. London: MIT press.
- 700 Firestein, S. 2015. *Failure*. oxford university press.
- 701 Galvão, G. R. and Z. Dias 2015, January. An audit tool for genome rearrange-  
702 ment algorithms. *J. Exp. Algorithmics* 19, 1.7:1.1–1.7:1.34.
- 703 Gannett, L. and J. R. Greisemer 2004. Classical genetics and the geography  
704 of genes. In Rheinberger and Gaudilliere (Eds.), *Classical Genetic Research*  
705 *and Its Legacy*, pp. 57–88. London and New York: Routledge.
- 706 Hagen, J. B. 1982. *Experimental Taxonomy, 1930-1950: The Impact of Cy-*  
707 *tology, Ecology, and Genetics on Ideas of Biological Classification*. Ph. D.  
708 thesis, Oregon State University.
- 709 Hagen, J. B. 1984, Jun. Experimentalists and naturalists in twentieth-century  
710 botany: Experimental taxonomy, 1920–1950. *Journal of the History of Bi-*  
711 *ology* 17(2), 249–270.
- 712 ——— 1999. Naturalists, molecular biologists, and the challenges of molecular  
713 evolution. *Journal of the History of Biology* 32(2), 321–341.
- 714 ——— 2000. The origins of bioinformatics. *Nature Reviews Genetics* 1(3), 231.
- 715 ——— 2001. The introduction of computers into systematic research in the  
716 united states during the 1960s. *Stud Hist Phil Biol and Biomed Sci.* 32,  
717 291–314.
- 718 ——— 2003. The statistical frame of mind in systematic biology from quantita-  
719 tive zoology to biometry. *Journal of the History of Biology* 36(2), 353–384.
- 720 ——— 2010. Waiting for sequences: Morris goodman, immunodiffusion experi-  
721 ments, and the origins of molecular anthropology. *Journal of the History of*  
722 *Biology* 43(4), 697–725.
- 723 Kay, L. E. 1993. *The Molecular Vision of Life: Caltech, The Rockefeller Foun-*  
724 *dition, and the Rise of the New Biology*. Oxford University press.

- 725 Kececioglu, J. and D. Sankoff 1995. Exact and approximation algorithms for  
726 sorting by reversals, with application to genome rearrangement. *Algorith-*  
727 *mica* 13, 180–210.
- 728 Kohler, R. E. 1994. *Lords of the fly: Drosophila genetics and the experimental*  
729 *life*. University of Chicago Press.
- 730 Lehmer, D. H. 1993. The mathematical work of morgan ward. *Math. Comp.* 61,  
731 307–311.
- 732 Livio, M. 2014. *Brilliant Blunders: From Darwin to Einstein - Colossal Mis-*  
733 *takes by Great Scientists That Changed Our Understanding of Life and the*  
734 *Universe*. Brilliance Audio.
- 735 Margoliash, E. 1963. Primary structure and evolution of cytochrome c. *Pro-*  
736 *ceedings of the National Academy of Sciences* 50(4), 672–679.
- 737 McClung, C. E. 1908. Cytology and taxonomy. *Kansas University Science*  
738 *Bulletin* 4(7), 199–215.
- 739 Metz, C. W. 1914. Chromosome studies in the diptera. i. a preliminary survey  
740 of five different types of chromosome groups in the genus drosophila. *Journal*  
741 *of Experimental Zoology Part A: Ecological Genetics and Physiology* 17(1),  
742 45–59.
- 743 ——— 1916. Chromosome studies on the Diptera. III. additional types of chro-  
744 mosome groups in the Drosophilidae. *The American Naturalist* 50(598),  
745 587–599.
- 746 ——— 1918. Chromosome studies on the Diptera. *Zeitschrift für induktive*  
747 *Abstammungs-und Vererbungslehre* 19(3), 211–213.
- 748 Mohr, O. L. 1919. Character changes caused by mutation of an entire region  
749 of a chromosome in drosophila. *Genetics* 4, 275–282.
- 750 Morgan, G. J. 1998. Emile Zuckerkandl, Linus Pauling, and the molecular  
751 evolutionary clock, 1959-1965. *J Hist Biol* 31(2), 155–178.
- 752 Morgan, T. H. and C. B. Bridges 1916. *Sex-linked inheritance in Drosophila*.  
753 Carnegie Inst. Washington, Publ.
- 754 Morgan, T. H., C. B. Bridges, and A. H. Sturtevant 1925. *The Genetics of*  
755 *Drosophila*. Bibliographia Genetica.
- 756 Morgan, T. H., A. H. Sturtevant, and C. B. Bridges 1920. The evidence for  
757 the linear order of the genes. *Proc Natl Acad Sci U S A* 6(4), 162–164.
- 758 Muller, H. J. 1929. The first cytological demonstration of a translocation in  
759 drosophila. *The American Naturalist* 63(689), 481–486.
- 760 Murphy, W. J., D. M. Larkin, A. Everts-van der Wind, G. Bourque, G. Tesler,  
761 L. Auvil, J. E. Beever, B. P. Chowdhary, F. Galibert, L. Gatzke, C. Hitte,  
762 S. N. Meyers, D. Milan, E. A. Ostrander, G. Pape, H. G. Parker, T. Raud-  
763 sepp, M. B. Rogatcheva, L. B. Schook, L. C. Skow, M. Welge, J. E. Womack,  
764 S. J. O'brien, P. A. Pevzner, and H. A. Lewin 2005, Jul. Dynamics of mam-  
765 malian chromosome evolution inferred from multispecies comparative maps.  
766 *Science* 309(5734), 613–617.
- 767 Novitski, E. 2005. *Sturtevant and Dobzhansky: Two Scientists at Odds, With*  
768 *a Student's Recollections*. Bloomington: Xlibris Corporation.
- 769 Painter, T. S. 1933. A new method for the study of chromosome rearrange-  
770 ments and the plotting of chromosome maps. *Science* 78, 585–586.

- 771 — 1934. Salivary chromosomes and the attack on the gene. *Journal of*  
772 *Heredity* 25(12), 465–476.
- 773 Papadimitriou, C. H. 1993. *Computational Complexity*. Pearson.
- 774 Pevzner, P. and G. Tesler 2003, Jan. Genome rearrangements in mammalian  
775 evolution: lessons from human and mouse genomes. *Genome Res* 13(1),  
776 37–45.
- 777 Popper, K. 1959. *The Logic of Scientific Discovery*. Abingdon-on-Thames:  
778 Routledge.
- 779 Porter, T. M. 1996. *Trust in numbers*. Princeton University Press.
- 780 Smocovitis, V. B. 2006. Keeping up with dobzhansky: G. ledyard stebbins,  
781 jr., plant evolution, and the evolutionary synthesis. *Hist. Phil. Life Sci.*, 28,  
782 9–48.
- 783 — 2009, aug. The "Plant Drosophila": E. b. babcock, the GenusCrepis,  
784 and the evolution of a genetics research program at berkeley, 1915–1947.  
785 *Historical Studies in the Natural Sciences* 39(3), 300–355.
- 786 Sommer, M. 2008, Sep. History in the gene: Negotiations between molecular  
787 and organismal anthropology. *Journal of the History of Biology* 41(3), 473–  
788 528.
- 789 Strasser, B. J. 2010a. Collecting, comparing, and computing sequences: the  
790 making of Margaret O. Dayhoff's atlas of protein sequence and structure,  
791 1954–1965. *Journal of the History of Biology* 43(4), 623–660.
- 792 — 2010b. Laboratories, museums, and the comparative perspective: Alan  
793 A. Boyden's quest for objectivity in serological taxonomy, 1924-1962. *Hist*  
794 *Stud Nat Sci* 40(2), 149–182.
- 795 Sturtevant, A. H. 1921, Aug. A case of rearrangement of genes in drosophila.  
796 *Proc Natl Acad Sci U S A* 7(8), 235–237.
- 797 — 1942. The classification of the genus drosophila, with descriptions of nine  
798 new species. *Austin: The University of Texas Publication* 4213, 5–51.
- 799 Sturtevant, A. H., C. B. Bridges, and T. H. Morgan 1919, May. The spatial  
800 relations of genes. *Proc Natl Acad Sci U S A* 5(5), 168–173.
- 801 Sturtevant, A. H. and T. Dobzhansky 1936, Jul. Inversions in the third chro-  
802 some of wild races of drosophila pseudoobscura, and their use in the study  
803 of the history of the species. *Proc Natl Acad Sci U S A* 22(7), 448–450.
- 804 Sturtevant, A. H. and E. Novitski 1941. The homologies of chromosome ele-  
805 ments in the genus drosophila. *Genetics* 26, 517–541.
- 806 Sturtevant, A. H. and C. R. Plunkett 1926. Sequence of corresponding third-  
807 chromosome genes in drosophila melanogaster and d. simulans. *Biol Bull* 50,  
808 56–60.
- 809 Sturtevant, A. H. and C. C. Tan 1937. The comparative genetics of Drosophila  
810 Pseudoobscura and D. Melanogaster. *Journal of Genetics* 34, 415–432.
- 811 Suárez-Díaz, E. 2009, Mar. Molecular evolution: concepts and the origin of  
812 disciplines. *Stud Hist Philos Biol Biomed Sci* 40(1), 43–53.
- 813 — 2010. Making room for new faces: evolution, genomics and the growth of  
814 bioinformatics. *History and Philosophy of the Life Sciences* 32(1), 65–89.
- 815 — 2014, Aug. The long and winding road of molecular data in phylogenetic  
816 analysis. *Journal of the History of Biology* 47(3), 443–478.



- 
- 817 Suárez-Díaz, E. and V. H. Anaya-Muñoz 2008, Dec. History, objectivity, and  
818 the construction of molecular phylogenies. *Stud Hist Philos Biol Biomed*  
819 *Sci* 39(4), 451–468.
- 820 Turing, A. M. 1936. On computable numbers, with an application to  
821 the entscheidungsproblem. *Proceedings of the London mathematical soci-*  
822 *ety* 2(1), 230–265.
- 823 Turrill, W. B. 1938. The expansion of taxonomy with special reference to  
824 spermatophyta. *Biol Rev.* 13, 342–373.
- 825 Watterson, G., W. Ewens, T. Hall, and A. Morgan 1982, nov. The chromosome  
826 inversion problem. *Journal of Theoretical Biology* 99(1), 1–7.
- 827 Zuckerkandl, E. and L. Pauling 1965, Mar. Molecules as documents of evolu-  
828 tionary history. *J Theor Biol* 8(2), 357–366.