Mud Sticks: On the Alleged Falsification of Mendel's Data

Daniel L. Hartl*,[1] and Daniel J. Fairbanks[†]
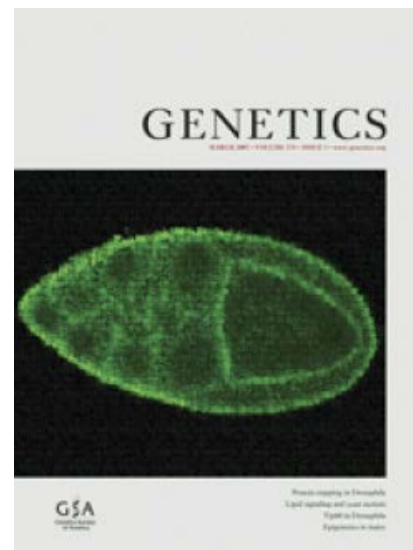
GREGOR Mendel's celebrated paper (MENDEL 1866) is a seemingly inexhaustible source of inspiration and controversy for each succeeding generation of geneticists and historians of genetics (FRANKLIN *et al.* 2007). For the aficionado (or the fanatic) it is studied repeatedly, much as an avid sports fan enjoys each rerun of a classic matchup or a movie buff looks forward to yet another screening of *Casablanca*. Mendel's paper is special for a number of reasons. Its historical importance is beyond dispute, but its layout and style are also alluring. Unassuming and unpretentious, Mendel straightforwardly explains his rationale, his experiments, his results, and his interpretation. For the teacher of genetics, the paper is a cornucopia of raw data obtained from various types of crosses, data of the sort that scarcely exist in today's literature owing to the brevity of communications and the emphasis on summary statistics while the real data, when available at all, are relegated to supplemental online material. Above all Mendel's paper appears to reflect the author's simplicity, modesty, and guilelessness. The statistician R. A. Fisher (FISHER 1936), who devoted what must have been a great deal of his time to reconstructing the timeline and scale of Mendel's experiments, came to the conclusion that "there can, I believe, now be no doubt whatever that [Mendel's] report is to be taken entirely literally, and that his experiments were carried out in just the way and much in the order that they are recounted" (FISHER 1936, p. 132). (An electronic copy of Fisher's article is available online at http://digital.library.adelaide.edu.au/.)

So why the controversies? Many of them relate to what Mendel left unsaid or to events that happened long after his death in 1884. Among the left unsaids was what

Mendel thought about DARWIN's (1859) theory of evolution by means of natural selection (CALLENDER 1988; OREL 1996) or whether Mendel regarded his own experiments as continuing in the long tradition of plant hybridizers (OLBY 1979; MONAGHAN and CORCOS 1990). Among the happened laters was Mendel's lionization by the early 20th century geneticists to such an extent that they overlooked or ignored whatever shortcomings Mendel may have had in his understanding of transmission genetics (OLBY 1979; BOWLER 1989; HARTL and OREL 1992). These are interesting issues and debates with implications for the history of genetics, but none of them impugns Mendel's integrity or undermines the validity of his work (FRANKLIN *et al.* 2007).

The allegation that has impugned Mendel's integrity is FISHER's (1936, p. 132) assertion that "the data of most, if not all, of the experiments have been falsified so as to agree closely with Mendel's expectations," a discovery that Fisher regarded as "abominable" and "shocking" (BOX 1978, pp. 296 and 197, respectively). Much of the discrepancy results from the absence of extreme deviates, and this can largely be explained by unconscious bias in classifying ambiguous phenotypes, stopping the counts when satisfied with the results, recounting when the results seem suspicious, and repeating experiments whose outcome is mistrusted (WRIGHT 1966; EDWARDS 1986; SEIDENFELD 1998). The effects of unconscious bias, arbitrary stopping, or recounting are subtle, and Mendel states that he repeated some experiments. But Fisher's allegation went beyond mere trimming. Fisher alleged that in two particular series of experiments Mendel had the wrong expectation and that Mendel's results fit the erroneous expectation significantly better than Fisher's corrected expectation. Although Fisher somewhat disingenuously attributes the discrepancy to "a possibility among others that Mendel was deceived by some assistant who knew too well what was expected" (FISHER 1936, p. 132), the charge of data falsification has stuck to

[1]*Corresponding author:* Department of Organismic and Evolutionary Biology, Harvard University, The Biological Laboratories, 16 Divinity Ave., Cambridge, MA 02138. E-mail: dhartl@oeb.harvard.edu

*Department of Organismic and Evolutionary Biology, Harvard University, Cambridge, Massachusetts 02138 and
[†]*Department of Plant and Animal Sciences, Brigham Young University, Provo, Utah 84602*

Mendel like dirt sticks to a candidate after a mud-slinging political campaign.

In this article we summarize the nature of Fisher's criticism of the two key series of experiments. We argue that the criticism is not justified by the words that Mendel wrote or by the data presented in his paper. In the second series of experiments, Fisher appears to have misinterpreted the trait that Mendel actually examined (FAIRBANKS and RYTTING 2001). In view of the manner in which Mendel likely carried out the experiment and the trait he probably scored, Mendel's expectation is more nearly correct than Fisher's. In the first series of experiments, the results do not differ significantly from Fisher's expectation, so there is no factual basis to claim a discrepancy. Indeed, in this series of experiments, Mendel repeated an experiment whose outcome he did not like, and the pooled results of the original experiment and the repeat fit Fisher's expectation almost exactly. Our conclusion is that, despite his painstaking reconstruction of Mendel's experiments, Fisher was intemperate if not reckless in alleging deliberate falsification.

In the two series of experiments in question, Mendel performed what we now call progeny tests to ascertain which individuals exhibiting the dominant phenotype were actually heterozygous and which were homozygous. Mendel describes the critical series as follows. "For each separate trial … 100 plants were selected which displayed the dominant character in the first generation, and in order to ascertain the significance of this, ten seeds of each were cultivated [… *von jeder 10 Samen angebaut*]" (MENDEL 1866). (All quotations from Mendel's paper are from an English translation known as the Druery–Bateson translation, which along with the original German text is available online at http://www.mendelweb.org, a wonderful resource created and maintained by Roger B. Blumberg.)

What are the expected values in this experiment? Fisher correctly observed that, given a true 2:1 ratio of heterozygous:homozygous parents, if the diagnosis of parental heterozygosity is based on the presence of one or more homozygous recessive progeny among i offspring, then the ratio of parents classified as heterozygous to those classified as homozygous would be

$$\frac{\text{Het}}{\text{Hom}} = \frac{(2/3)[1-(3/4)^i]}{(1/3)+(2/3)(3/4)^i}. \tag{1}$$

For $i = 30$ this ratio is 0.667:0.333 (or 2:1, which equals Mendel's expectation) but for $i = 10$ the ratio is 0.6291:0.3709 (or 1.7:1, which is Fisher's expectation). In the two series of progeny tests together, Mendel reports a ratio of 720 heterozygous:353 homozygous dominant. This does not differ significantly from Mendel's expectation ($P = 0.76$), but it does differ significantly from Fisher's expectation ($P = 0.0045$). This P-value is the basis of Fisher's statement that "a total devia-

tion of the magnitude observed, and in the right direction, is only to be expected once in 444 trials; there is therefore here a serious discrepancy" (FISHER 1936, p. 129).

Many years later, E. NOVITSKI (2004) suggested an explanation for the discrepancy. Novitski proposed that Mendel required a complete set of 10 progeny with the dominant phenotype to score the parent as homozygous dominant, but that any smaller set of progeny containing one or more homozygous recessives would be sufficient to identify the parent as heterozygous and would have been counted as such. To make this work, Mendel would also have had to eliminate any set of 9 or fewer progeny when all had the dominant phenotype and replace them with an additional set of 10. Such a procedure is biased toward classifying an excess of parents as heterozygous, creating a deviation in the direction opposite that proposed by Fisher. With this as the counting rule, C. E. NOVITSKI (2004) showed that the expected ratio of parents classified as heterozygous to those classified as homozygous is given by

$$\frac{\text{Het}}{\text{Hom}} = \frac{\sum_{i=1}^{10}\binom{10}{i}v^i(1-v)^{10-i}\{(2/3)[1-(3/4)^i]\}}{v^{10}[(1/3)+(2/3)(3/4)^{10}]}, \tag{2}$$

where v is the probability that a given seed gives rise to a progeny plant that can be scored. When v is very close to 1, Equation 2 yields Fisher's expected ratio. But what is the actual value of v? E. NOVITSKI (2004, p. 1134) comments that "no information about the failure rate is available for [the first series of] experiments, but Mendel did give it for a subsequent [dihybrid cross]. … Of 556 seeds, 11 did not yield plants; this is very close to 2% (0.0198)." For $v = 0.98$, Equation 2 yields a ratio of 2.1:1 *vs.* Fisher's expectation of 1.7:1. In other words, for $v = 0.98$ Novitski's counting rule introduces a bias that is comparable in magnitude but opposite in sign to Fisher's correction.

However, a survivorship of 98% results from what we believe is a misreading of Mendel's paper. For the experiment in question Mendel writes that:

> In all, 556 seeds were yielded by 15 plants, and of these there were:
>
> - 315 round and yellow,
> - 101 wrinkled and yellow,
> - 108 round and green,
> - 32 wrinkled and green.
>
> All were sown the following year. 11 of the round yellow seeds did not yield plants, and 3 plants did not form seeds. … From the wrinkled yellow seeds 96 resulting plants bore seed. … From 108 round green seeds 102 resulting plants fruited. … The wrinkled green seeds yielded 30 plants. (MENDEL 1866)

That Novitski misinterpreted Mendel's statement of survivorship is almost certain, since the number 556

appears only once in Mendel's paper, and the number 11 appears only once in connection with failure to grow.

Mendel's results actually imply survivorships of 304/315, 96/101, 102/108, and 30/32, respectively, for individual probabilities of survival of 0.965, 0.950, 0.944, and 0.938, respectively. In other experiments MENDEL (1866) cites survivorships of $405/416 = 0.974$, $90/98 = 0.918$ and $87/94 = 0.926$, $639/687 = 0.930$, and $166/187 = 0.888$. Undoubtedly the survivorship differed from year to year and even from one part of the experimental plot to another because of differences in microenvironment. Taken together, the survivorship data imply a mean $v = 0.94$ with 95% confidence interval 0.93–0.95. The problem is that anywhere in this range of survivorships Equation 2 predicts an expected ratio substantially in excess of 2:1 with a discrepancy that is even greater than Fisher's but in the opposite direction. Novitski's suggestion would therefore seem to be untenable.

Novitski was concerned about progeny tests in which <10 progeny survived, and Fisher speculated that Mendel might actually have cultivated >10. But Mendel is so specific in his statement that he cultivated exactly 10 progeny ("... *von jeder 10 Samen angebaut*") that we should surely take him at his word. If there is anything to which contemporaries who knew Mendel agreed, it was that he was a superb gardener (ILTIS 1932), and any experienced gardener would know exactly how to do this experiment (FAIRBANKS and RYTTING 2001). With more seeds available than progeny plants wanted, the gardener plants 2–3 seeds per hill and thins each hill back to one seedling at some convenient time after germination. Alternatively, Mendel could have germinated the seeds in flats in his greenhouse and transplanted 10 seedlings to his experimental plot. That Mendel did carry out transplants in some experiments is indicated by his statement regarding the monohybrid experiment with long *vs.* short stem, in which he writes, "In this experiment the dwarfed plants were carefully lifted and transferred to a special bed. This precaution was necessary, as otherwise they would have perished through being overgrown by their tall relatives. Even in their quite young state they can be easily picked out by their compact growth and thick dark-green foliage" (MENDEL 1866). Whether multiple seeds were planted per hill in the field or germinated in the greenhouse, either strategy would have allowed Mendel to make optimal use of the ∼30 seeds from each parental plant, cultivate exactly 10 progeny per parental plant when needed, and use no more garden space than actually required for the experiment.

So let us assume that Mendel planted enough seeds to obtain at least 10 seedlings from each parental plant and reexamine Fisher's allegation in this light. We begin with the second series of experiments because it is in these experiments that Fisher found his strongest evidence for data falsification. Fisher refers to these as the "trifactorial data," and they are reported in Section 8 in Mendel's paper. Fisher asserts that his Table II (FISHER 1936, p. 127) shows "Mendel's trifactorial classification of the 639 plants of the second generation ..., which follows Mendel's notation, in which *a* stands for *wrinkled* seeds, *b* for *green* seeds, and *c* for *white flowers*" (FISHER 1936, p. 127). Assuming that white flowers is a completely recessive trait and that all parental genotypes were diagnosed on the basis of the phenotypes of exactly 10 surviving progeny, there is a significant discrepancy. Mendel reports a ratio of 321:152 whereas Fisher's expected values are 297.6:175.4. The chi-square statistic is 4.97, which has an associated *P*-value of 0.026 and is significant.

Two explanations for the discrepancy have been offered; both are plausible and each may have played a role. One is that Fisher misunderstood the trait that Mendel actually scored. The third trait in the trifactorial cross was probably scored not as flower color but rather as axillary pigmentation, a pleiotropic effect of the same mutation that can easily be identified in the seedlings as early as 2–3 weeks after germination and is perfectly correlated with both flower color and seed-coat color (FAIRBANKS and RYTTING 2001). Scoring axillary pigmentation would have allowed Mendel to complete the observations on the progeny seeds and the progeny seedlings within a single growing season (FAIRBANKS and RYTTING 2001). Mendel stated that the trihybrid experiment "was conducted in a manner quite similar to that used in the preceding one" (MENDEL 1866), and the preceding experiment is the dihybrid experiment for seed shape and seed color. Indeed, Mendel could easily study seed shape and seed color in the trihybrid experiment in the same manner as he studied these traits in the dihybrid experiment. However, to determine the genotype of F_2 plants with progeny tests for the third trait (seed-coat, flower, and axillary color), Mendel had to grow F_3 progeny or conduct testcrosses. Fisher presumed that Mendel used the same procedure as in the first series of progeny tests in which "ten seeds of each were cultivated." But in these experiments Mendel did not specify his method, but referred simply to "further investigations [*Weiteren Untersuchungen*]." In any event, if Mendel planted enough seeds to ensure at least 10 seedlings, more than likely he scored them all or, if choosing exactly 10 to preserve after thinning, quite consciously made sure to include any of the seedlings that showed the recessive phenotype. Effectively, the i in Equation 1 becomes >10. With probabilities of germination in the range 0.93–0.95 Mendel may easily have examined ≥20 seedlings, but in fact any significant discrepancy from Fisher's expectation disappears if he had scored as few as 11 seedlings from 70% of the progeny plants and 10 seedlings from the remaining 30%.

Another explanation, offered by WRIGHT (1966), is that some of Mendel's traits could occasionally be recognized

in heterozygous genotypes. Mendel operationally defined a "dominant" trait as one in which the homozygous dominant and heterozygous genotypes could not be distinguished reliably in single individuals. Wright pointed out that what might be true in the case of single individuals need not be true in populations of individuals. A progeny test yields a population of individuals, and Wright suggested that a small percentage of low-level expression in the heterozygous progeny might have allowed Mendel to classify a parent as heterozygous even though none of the offspring of the tested progeny was homozygous recessive.

Wright's suggestion implies that the expected ratio of parental plants classified as heterozygous to those classified as homozygous is given by

$$\frac{\text{Het}}{\text{Hom}} = \frac{(2/3)\{1 - (3/4)^{10} + (3/4)^{10}[1 - (1 - p)^{10}]\}}{(1/3) + (2/3)(3/4)^{10}(1 - p)^{10}},$$
(3)

where p is the penetrance of the recessive trait in heterozygous genotypes. When $p = 0$, Equation 3 yields Fisher's ratio of 1.7:1, and when $p = 1$ it yields Mendel's ratio of 2:1. A value of $p = 0.02$ is sufficient to eliminate any significant discrepancy between the observed data and that expected from Equation 3. Mendel would therefore have needed to identify only a small percentage of heterozygous plants showing signs of the recessive phenotype to make Fisher's expectation spurious. It should, however, be noted that E. Novitski's (2004) assertion that seed-coat color shows incomplete dominance is probably not correct. In varieties of *Pisum sativum* L. with opaque (colored) seed coats, the spotting and coalescence of spots in the seed coat do not reflect incomplete dominance of alleles of the *a* (*anthocyaninless*) locus, which is undoubtedly the locus responsible for anthocyanin pigmentation in the seed coat, flower petals, and leaf axils in Mendel's material (Fairbanks and Rytting 2001). Rather seed-coat spotting is a highly variable quantitative character affected by the alleles of multiple loci as well as by the environment (White 1917; Khvostova 1983).

What about the first series of progeny tests? Mendel described these in the passage quoted below. He was obviously displeased with the result of experiment 5 and admits that he repeated it, giving both the original result and that of the repeated experiment. As Wright (1966) has pointed out in his biologically insightful analysis, this kind of candor would hardly be expected from a person bent on fraud. Mendel writes:

> For each separate trial in the following experiments 100 plants were selected which displayed the dominant character in the first generation, and in order to ascertain the significance of this, ten seeds of each were cultivated.
>
> • Expt. 3. The offspring of 36 plants yielded exclusively gray-brown seed-coats, while of the offspring of 64 plants some had gray-brown and some had white.
>
> • Expt. 4. The offspring of 29 plants had only simply inflated pods; of the offspring of 71, on the other hand, some had inflated and some constricted.
>
> • Expt. 5. The offspring of 40 plants had only green pods; of the offspring of 60 plants some had green, some yellow ones.
>
> • Expt. 6. The offspring of 33 plants had only axial flowers; of the offspring of 67, on the other hand, some had axial and some terminal flowers.
>
> • Expt. 7. The offspring of 28 plants inherited the long axis, of those of 72 plants some the long and some the short axis. . . .
>
> Experiment 5, which shows the greatest departure [from 2:1], was repeated, and then in lieu of the ratio of 60:40, that of 65:35 resulted. (Mendel 1866)

Altogether, the experiments yielded 399 parents classified as heterozygous and 201 parents classified as homozygous. Fisher noted that the expected values from Equation 1 are 377.5 and 222.5, respectively. A chi-square test yields the test statistic 3.31, which has an associated P-value of 0.069. This does not differ significantly from Fisher's expectation; nevertheless, it fanned Fisher's suspicion because he writes, "a deviation as fortunate as Mendel's is to be expected once in twenty-nine trials" (Fisher 1936, pp. 125–126).

Experiments 3 and 7 should probably not be included with the others because these traits can be classified in the seedlings; hence, for the reasons explained above it is likely that effectively >10 seedlings were scored per parental plant. The remaining four experiments include 263 parental plants classified as heterozygous and 137 classified as homozygous as against Fisher's expectation of 251.6 and 148.4 ($P = 0.24$). Among all the progeny tests, the results of experiment 5 are closest to Fisher's expectation ($P = 0.54$). Mendel evidently mistrusted this result, since this is an experiment he chose to repeat. For the two runs of the experiment taken together, the observed ratio is 125:75, which is an almost perfect fit to Fisher's expectation of 125.8:74.2 ($P = 0.90$).

Nevertheless, Fisher was evidently convinced by his analysis, as he referred in private to the discrepancies as "abominable" and a "shocking experience" (Box 1978). The allegation of data falsification is based on his analysis of two series of progeny tests in which he calculated expectations different from Mendel's by correcting for a slight bias in the classification of parental genotypes from only 10 progeny per parent. In the second series of experiments the trait actually scored was probably not flower color, as Fisher inferred, but rather axillary pigmentation. Although these are pleiotropic effects of the same mutation, axillary pigmentation can be scored in the seedlings (Fairbanks and Rytting 2001), and so it is likely that >10 progeny per parental plant were examined. In the first series of experiments, not only axillary pigment but also stem length can be scored in the seedlings (Fairbanks and Rytting 2001). Nevertheless, neither these nor the other progeny tests differ

significantly from Fisher's expected values; hence, there is no factual basis for an allegation of data falsification. Indeed, in the largest number of observations consisting of an original experiment dealing with green *vs.* yellow pods and its repetition, the pooled results agree almost exactly with Fisher's expected values. Let us hope against all experience that Fisher's allegation of deliberate falsification can finally be put to rest, because on closer analysis it has proved to be unsupported by convincing evidence.

LITERATURE CITED

Bowler, P. J., 1989   *The Mendelian Revolution.* Johns Hopkins University Press, Baltimore, MD.

Box, J. F., 1978   *R. A. Fisher: The Life of a Scientist.* Wiley, New York.

Callender, L. A., 1988   Gregor Mendel: an opponent of descent with modification. Hist. Sci. **26:** 41–75.

Darwin, C., 1859   *The Origin of Species.* New American Library, New York.

Edwards, A. W. F., 1986   Are Mendel's results really too close? Biol. Rev. **61:** 295–312.

Fairbanks, D. J., and B. Rytting, 2001   Mendelian controversies: a botanical and historical review. Am. J. Bot. **88:** 737–752.

Fisher, R. A., 1936   Has Mendel's work been rediscovered? Ann. Sci. **1:** 115–137.

Franklin, A., A. W. F. Edwards, D. J. Fairbanks, D. L. Hartl and T. Seidenfeld, 2007   *The Mendel-Fisher Controversy: Can Data be "Too Good to Be True?"* University of Pittsburgh Press, Pittsburgh (in press).

Hartl, D. L., and V. Orel, 1992   What did Gregor Mendel think he discovered? Genetics **131:** 245–253.

Iltis, H., 1932   *Life of Mendel.* Hafner, New York.

Khvostova, V. V., 1983   *Genetics and Breeding of Peas.* Oxonian Press, New Delhi.

Mendel, G., 1866   Versuche über pflanzen-hybriden. Verhandlungen des naturforschenden vereines. Abh. Brünn **4:** 3–47.

Monaghan, F. V., and A. F. Corcos, 1990   The real objective of Mendel's paper. Biol. Philos. **5:** 267–292.

Novitski, C. E., 2004   Revision of Fisher's analysis of Mendel's garden pea experiments. Genetics **166:** 1139–1140.

Novitski, E., 2004   On Fisher's criticism of Mendel's results with the garden pea. Genetics **166:** 1133–1136.

Olby, R., 1979   Mendel, no Mendelian? Hist. Sci. **17:** 53–72.

Orel, V., 1996   *Gregor Mendel: The First Geneticist.* Oxford University Press, Oxford.

Seidenfeld, T., 1998   P's in a pod: some recipes for cooking Mendel's data. Philos. Sci. Arch. (http://philsci-archive.pitt.edu).

White, O. E., 1917   Studies of inheritance in Pisum. II. The present state of knowledge of heredity and variation in peas. Proc. Am. Philos. Soc. **56:** 487–588.

Wright, S., 1966   Mendel's ratios, pp. 173–175 in *The Origin of Genetics*, edited by C. Stern and E. R. Sherwood. W. H. Freeman, San Francisco.