孟德尔为大数据时代的科研带来的些许启示

孙丽雅（上海交通大学医学院附属精神卫生中心）
贺　林（上海交通大学 Bio-X 研究院）

在世期间遭受误解和冷落似乎是不少原创性科学先驱的共同遭遇，然而，真正推动时代进步的，恰恰是这些敢于挑战权威、彰显真理的人们。对他们的纪念也总是姗姗来迟但却绵远悠长。奠定现代遗传学基础的 19 世纪奥地利科学家孟德尔（Gregor Johann Mendel, 1822-1884）就是其中的一员。2022 年 7 月 20 日，是他诞辰 200 周年的纪念日。站在 200 年后的今天回顾他的科研故事，仍然能给早已处于大数据时代的我们带来些许启发。

在孟德尔的时代，遗传现象已经得到广泛的研究，杂交试验也在多处开展。两株不同性状的纯种植物杂交后产生的后代呈现 1:3 的表型规律也已有学者发现，但当时遗传界的权威理论是达尔文提出的“混合遗传”和“泛生”论。“混合遗传”认为后代继承了父母的所有性状，表现出的是这些性状的平均值；而“泛生”论则认为父母的所有性状（包括后天获得的性状）是通过名为“微芽”或“泛子”的粒子从全身汇集到生殖细胞，混合后传递给子代的。这两个理论虽均未得到科学上的证明，但当时的研究者多为描述型的博物学家而非统计分析型的数学科学家，在这种情况下，达尔文的理论成了被普遍接受的一套解释，没有人去深究 1:3 背后的隐藏含义。然而孟德尔没有满足于此，他凭借着自己认真严谨的科学思维完成了三步重要的工作：

一是选择了豌豆作为首选的研究材料

豌豆是严格闭花授粉的植物，自然状态下获得的都是严格自交的后代，这使得人工异花杂交的效果能够清晰呈现；而且孟德尔选择进行考察的豌豆性状，如花色、高矮茎等都差别明显，易于区分，遗传稳定，大大方便了后续的统计学分析。因此，考虑到真实世界的多样性和复杂性，如何在研究初期不盲目跟风，而是根据试验目的，选择恰当的研究对象和合理的终点评估指标，也是当今科学研究在设计阶段所要注意的。

二是重新思考和构建了遗传的潜在机制

孟德尔勇敢地跳出当时的权威解释体系，基于自身对遗传机制的长期观察、理解和思考，认为可能存在颗粒性的“遗传因子”，可以将一些性状在代际间稳定地遗传下去。由于那时还未发展出分子生物学，显然这个“遗传因子”是孟德尔推想

出来的，但却相当前瞻性地预测了遗传物质“基因”的存在；而对不同性状之间关系的理解，孟德尔也没有停留在简单“混合”的层面，创造性地定义了显性性状（对应显性基因型）和隐性性状（对应隐性基因型）这一对重要的概念，为后续解释杂交实验的结果提供了有效的理论支持。可见，在尊重事实的基础上，通过构建模型探索现象背后的本质，是永远不会过时的科学方法和态度。后续他采用数理统计，进一步用客观数字证明了自己的模型假设是合理的，便成了如虎添翼般的有效尝试。

三是揭示了遗传过程中的随机性

孟德尔提出的“分离定律”和“自由组合定律”，统称为“孟德尔遗传定律”。分离定律揭示了配子形成时等位基因分离，然后随机进入单个配子的过程，自由组合定律则进一步表明不同基因之间可以随机组合后再传递给下一代。这两个过程都体现了大自然对遗传这一过程所赋予的随机性，至今仍在数理统计领域发挥重要作用。分离定律被应用于统计因果推断中的孟德尔随机化算法，自由组合定律则构成了计算机最优解计算中遗传算法的核心步骤——交叉算子。可见，孟德尔遗传定律不仅揭示了生物遗传的重要规律，展现出的大自然智慧也为现代数理算法提供了不可替代的工具和思路。

如今，在大数据时代，无论是自然科学还是社会科学，各种空间和时间尺度节点上的观测数据正以前所未有的速度越积越多，如同孟德尔时代诸多博物学家周游世界带回来的海量标本、绘画或文字描述一般。那么，如何不被这些信息和数据所淹没，如何合理地辨析以及挖掘这些数据，甚至能够在假说的引导下选择性地只分析其中一部分必要的数据，便能见微知著，着实是非常考验当代研究者的智慧。“一花一世界，一叶一如来”，现代科学对研究者的科学直觉，试验设计能力和数理分析基础都提出了相当高的要求。在一头扎入数据海洋之前，先有从宏观到微观的审慎思考以及预判，合理地选择观测和分析的对象，再根据观测的结果对之前的预判或假设进行调整和校正，如此反复，不失为逐步接近真理的一条行之有效的途径。