# 显著检验与多重检验校正

说到显著检验，还可以从东西方文化的差异谈起。东方文明，讲究的是直觉，所以才会有道家的“道”，禅宗的“不立文字，直指人心”。总之，高深的知识只可意会，不可言传。而西方文化，先天讲究逻辑性。两千多年前，苏格拉底式的提问，就是从反复地对话推敲中，找到观念中的逻辑矛盾，从而认识真理。前者凭直觉从正面体验真理，后者更喜欢凭逻辑从对立面证明真理。所以严谨的现代科技很难从中国产生。这种差异也渗透到生活里，中国菜是“盐少许”。德国人就晕了，“少许？少许是什么意思？到底是 0.1g 还是 0.5g？”。德国人更喜欢把厨房改造成实验室，量筒小秤一应俱全。

说到显著检验，举个例子。如果两个人，有一个胖子一个瘦子，哪个更重？如果让中国人张三看见了，“这不是废话吗，当然是胖子重了。”但是如果让一个严谨的日耳曼人彼得看见了，他会说，“这必须要有证据才可以”（必须是日耳曼系才靠谱，拉丁国家也严谨不到哪里去，虽然他们足球踢得好）。于是，彼得拿来一个电子称，把胖瘦两人各称了一遍。结果是：50kg vs 90kg。彼得还是不放心：“虽然从检测结果来看两者有差异，但这个可能是真实差异，但也能我看走眼、电子称不稳定……”。总之，必须要把误差因素考虑上才可以。于是，接下来就是多次测量求平均值、t 检验，非把犯错的概率 P value 算出来才放心。“90.3 > 50.0，P<1.0E-10，嗯，看来结论是胖子重，而且我看走眼的概率是十亿份之一”。

也就是说，在任何一个严谨的科学测量中，我们判断两个数值是否有差异，必须要考虑这个差异可能来源两个方面：可能是真实的差异，也可能来自检测误差。而一般的显著检验的目的，就是计算出观测到的差异来源随机误差的概率，这样才能评判我们的结论是否可靠。例如，通常说的 P value（E value 是 blast 中一种特殊的 p value）小于 1%，就是说我们做出了一个判断（胖子比瘦子重），但这个判断犯错的概率是 1%（这里就是假阳性率，False positive rate）。虽然可能犯错，因为是这个属于小概率事件，我们就忍了吧，于是接受了这个判断。（类似，上街都可能遭遇车祸，因为是小概率事件，所以我们也就忍了……）。

但是，在很多科学实验中，在某些情况下，我们要做多次判断。例如，我们要判断两组样本对应的 10000 个基因的表达量是否在组间存在差异：基因 A 是否有差异？基因 B 是否有差异？基因 C 是否有差异？…..，如此下去，我们要进行 10000 次比较。如果我们以 p value 1% (假阳性的概率是 1%)来作为阈值，并假设每次判断都是彼此独立的，那么即使这 10000 个基因实际上都没有差异，我们也可能会得出有 100 个差异基因的结论（阳性结果的错误率为 100%，也就是下文要提到的 FDR (False Discovery Rate ) 值为 100%）。也就是说，一个小效率事件就在多次反复尝试后，变成了一个多次出现的事件（也就是俗话说的，“常在河边走，怎能不湿鞋”）。如果这 10000 个基因中有 100 个基因真实存在差异的，在 p vlaue 为 1%的阈值标准下，我们可能会得出 199 个基因有差异的结论（阳性结果的错误率，即 FDR 值约为 50%）。从这里，我们可以看到，在进行多次检验后（也就是所说的多重检验，multiple test)，那么基于单次比较的检验标准将变得过于宽松，使得阳性结果中的错误率（FDR 值）已经大到令人不可忍受的地步。

那么怎么办？最好的办法就提高判断的标准（p value)，单次判断的犯错概率就会下降，那么总体犯错的概率也将下降（类似，在多次相亲中，你可以通过提高标准来减少看走眼的概率）。在多重检验中提高判断标准的方法，我们就称之为“多重检验校正”。

最简单严厉的方法要属于 Bonferroni 校正。Bonferroni 法如何校正呢？很简单，还是上面的例子。原来我使用 p value1 %的标准判断是否差异表达，结果对于 10000 个没有差异的基因也会错误地得出“100 个基因差异表达的结论”。那么，我就将 p value 阈值直接提高到 $1\times10^{-6}$（也就是 1% 除以 10000），同样的 10000 次比较之后，平均假阳性次数也依然被

控制在 0.01 次。Perfect，滴水不漏，管控够严了。但这依然有一个问题，标准定太高了，如果一些基因真的存在表达差异，也很有可能达不到我们的阈值标准，被误判为没有差异，这就是假阴性率提高了 （类似如果相亲标准定太高了， 也可能会导致我们错失本来真的合适的另一半）。

于是，各路统计学的大侠设计了各种折中的方案。目前在 RNA-seq 中，使用最普遍的是 Benjamini and Hochberg 在 1995 年第一次提出了 FDR(False Discovery Rate) 的概念以及相应的多重检验校正方法（这个非参数的方法简单、 粗暴、 实用，谷歌显示这篇文章目前被引用了 21670 次，神一般的文章）。其出发点就是基于 Bonferroni 的保守性， 并给出了控制 FDR 的方法（这算是 FDR 控制方法的祖师爷了） 。FDR 就是一种控制阳性结果中的假阳性率的思路。在前面的例子的 10000 次基因差异比较中， 如果我们使用 FDR 为 1%的标准进行检验。如果最后检测出显著差异 （阳性结果） 的基因数是 100 个， 那么其中假阳性的个数就可以被控制在 1 个, 剩下的 99 个则是真实的差异 （阳性结果中的假阳性率被控制在 1%， 而 p value 1%是指单次检验的假阳性率为 1%，两者概念不同） 。FDR 的控制方法，延伸出了一个被校正后的 p value 的概念（比 P value 更严格），称之为 Q value，这个概念是最早是 John Storey （2002） 提出的。 在一般情况下， 大家可以简单一些理解， FDR、Q value、Adjusted p-value 指的是一个东西。

在国内目前大部分公司提供的 RNA-seq 类的生物信息分析中（包括基迪奥公司提供的分析服务），都会使用多重检验校正，而且基本都会使用 Benjamini1995 年报道的方法。在不同公司的结题报告中， 会将多重检验校正后使用的检验标准称为 FDR、Adjusted p-value 或 Q value，大家也不用感到迷惑，是一个东西的不同称呼而已。

不过 Benjamini 的方法依然过于保守。例如，如果将这个方法用于多基因的表达差异分析， Q value 检验的前提假设是所有多重比较的基因都是阴性的 （也就是没有差异） ，而且基因间的关系是独立的。 这显然还是过于严格， 因为样本内本身有些基因就是真实存在差异的，而且基因间往往存在大量互作。所以随后人们开始研究更加 powerful 的方法，现有的方法有 Storey 的，Broberg 的， Dalmasso 的， Guan 的， Strimmer 的等等。 Benjamini 的方法是将 FDR 控制在一个 level 以下， 而之后所有的方法都在试图精确地估计 FDR 。所以后来的这些方法都要 powerful 一些。不过他们所付出的代价就是加入了参数控制， 导致 robustness 的下降。据说 Storey 方法是最流行的 FDR control procedure(For details see Storey's paper published ON PNAS ,2003) 。这是一种利用多重比较结果中的 p value 分布，来预估真实的阳性率的方法，从而提高了 FDR 值预估的准确性。

现有 FDR 控制方法最大的弊端在于，他们假设多重检验中的真阴性的 p-value's 应该符合（ 1）相互间独立（ 2）符合分布于（ 0,1）的均匀分布。这两点假设从实际观察到的数据来看经常是不合理的。 尤其是第二点， 由于在实验中由各种处理因素外的其他因素的干扰 （背景噪音）， p value 值往往并不符合均匀分布。顺便提一句， Storey 和 Leek 在 07 年的 PLOS Genetics 发表了一篇文章专门解决第二个假设的合理性问题。文章中，作者通过对背景噪音的校正，从而使阴性的 p value 接近均匀分布。文章很牛，有兴趣可以看一下。