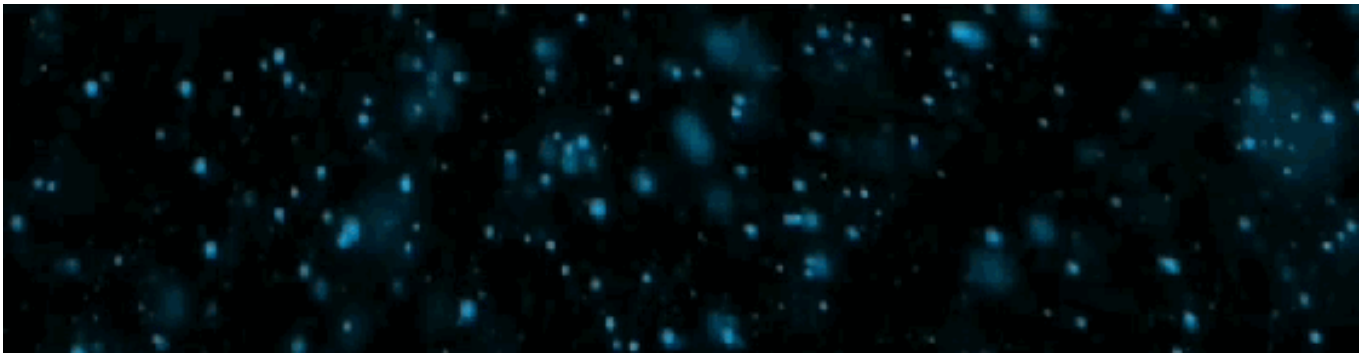
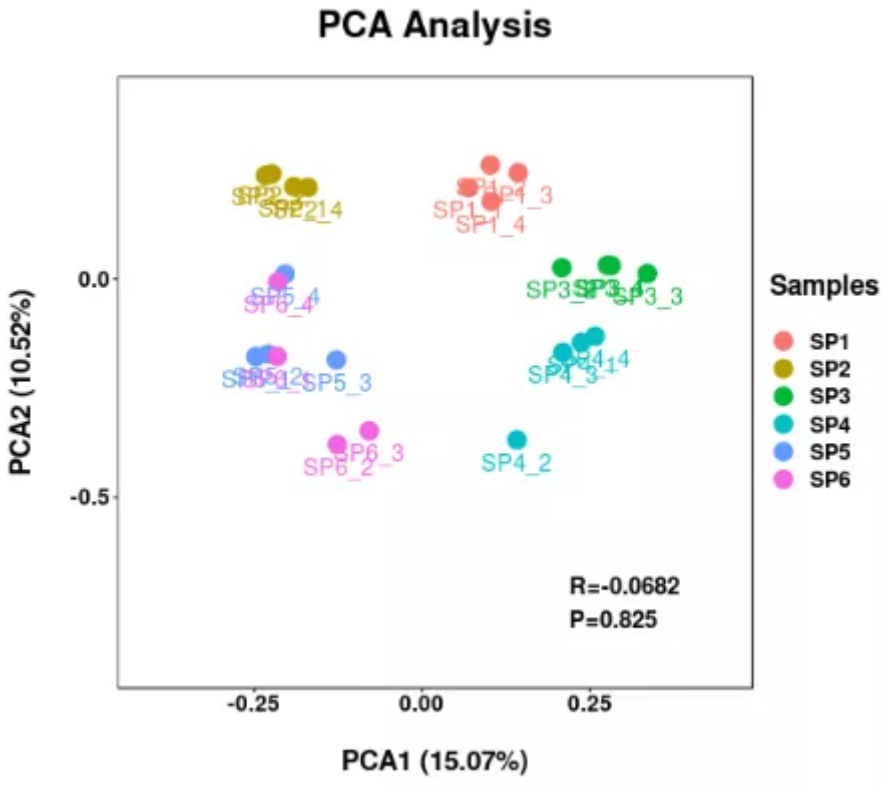
读完就懂主成分分析PCA

原创 运营部-HFR 联川生物 2020-03-05



当我们拿到很多类型的报告的时候，其中可能都会包括一张PCA图，一张二维坐标或者三维坐标散点图，其中的点或聚集或分散，可能还标上了不同的颜色，看起来简直就像是夜空中最闪亮的焰火，初次见到可能还真得费劲琢磨一番。什么是PCA？为什么它又有着如此重要的地位？我们究竟可以从PCA结果中看出哪些信息呢？



一张典型的二维PCA散点图

PCA全名principal component analysis，即主成分分析，听起来倒是非常的简单清爽，但是这主成分三个字里可是大有玄机。简单地说，主成分分析是一组变量通过正交变换转变成另一组变量的分析方法，来实现数据降维的目的。转换后得到的这一组变量，即是我们所说的主成分。

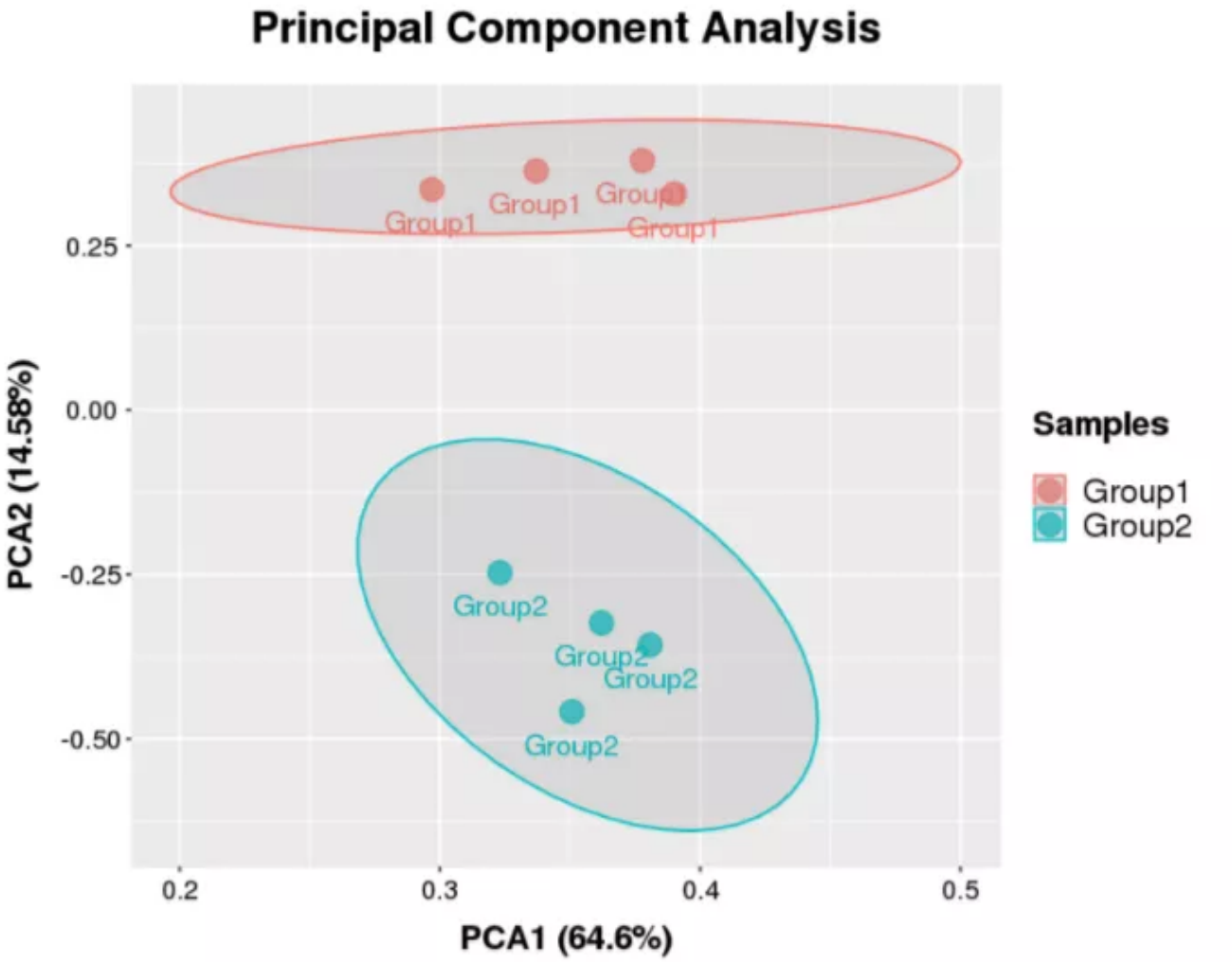
降维？降维又是什么？降维打击？可能直接上概念比较抽象，那么我们先来举个栗子。比方说，我们手里有一组重测序得到的变异数据，有n个突变位点，或者有一组转录组表达量数据，有n个转录本的表达量信息。那么我们就相当于有了一组n个变量，这个n可能非常大，可能随随便便就上万，甚至十万百万。想要直接比较两个或多个数据，显然就十分困难。而经过主成分分析，这样一组包含n个变量的数据经过转换变成了一组包含r个变量的数据，其中r<n，这样的过程即是降维，得到即是r个主成分。这r个主成分会依据方差的大小进行排序，称作主成分（PC）1、主成分2、……主成分r。而每个主成分的方差在这一组变量中的总方差中所占的比例，即是主成分的贡献度。通常来说，我们仅考察贡献度前2或者前3的主成分，经过可视化后，即得到了二维或三维PCA散点图。

可能看到这里你会问，这个过程我明白了，但是为什么较多的变量经过数据变换之后变成了较少的几个变量呢？这不是会有大量的信息丢失掉吗？如果你考虑到这个问题，那么恭喜，这说明你对主成分分析的思考已经很深了。

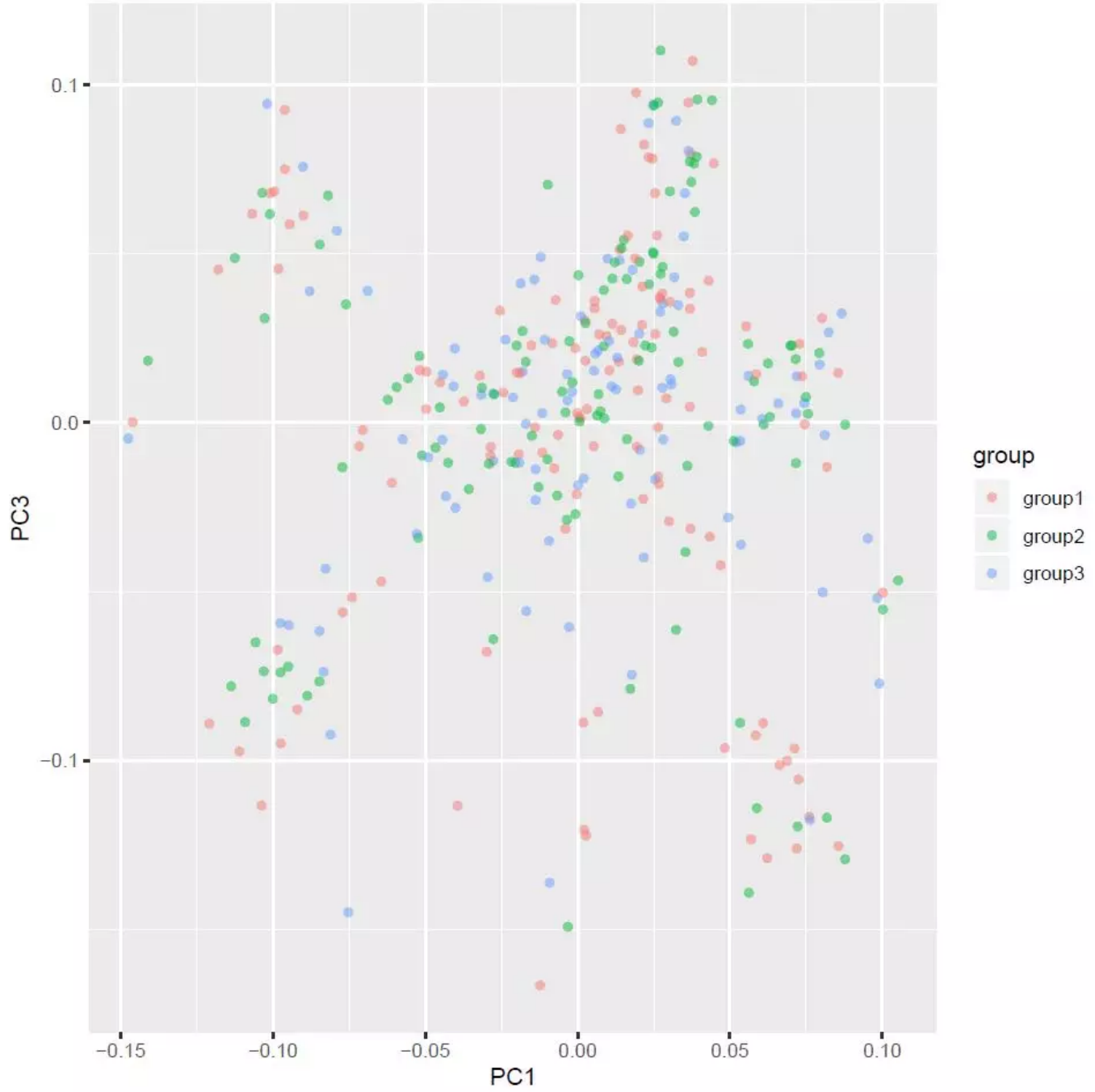
在我们最开始得到的一组变量中，变量之间并不是完全相互独立的。例如我们一个位点发生了变异，那么与之连锁的几个位点也大概率会发生变异；或者一个基因的表达量发生了变化，同一条通路中的其他基因的表达量也大概率会发生变化，即变量之间是存在相关性的。极端一点，假设两个位点完全连锁，那么我们去掉其中一个突变的所有信息，并不会影响总的信息含量。主成分分析也即是基于这样一种思想开展的，将变量之间根据相关性进行分解、合并和降维，类似于从n维空间到r维空间的投影。如果对具体的计算方法感兴趣，有很多相关的资料可供参考，当然也有很多工具可以方便我们直接对数据进行主成分分析。

那么具体到我们的报告当中，PCA图又说明了什么问题呢？

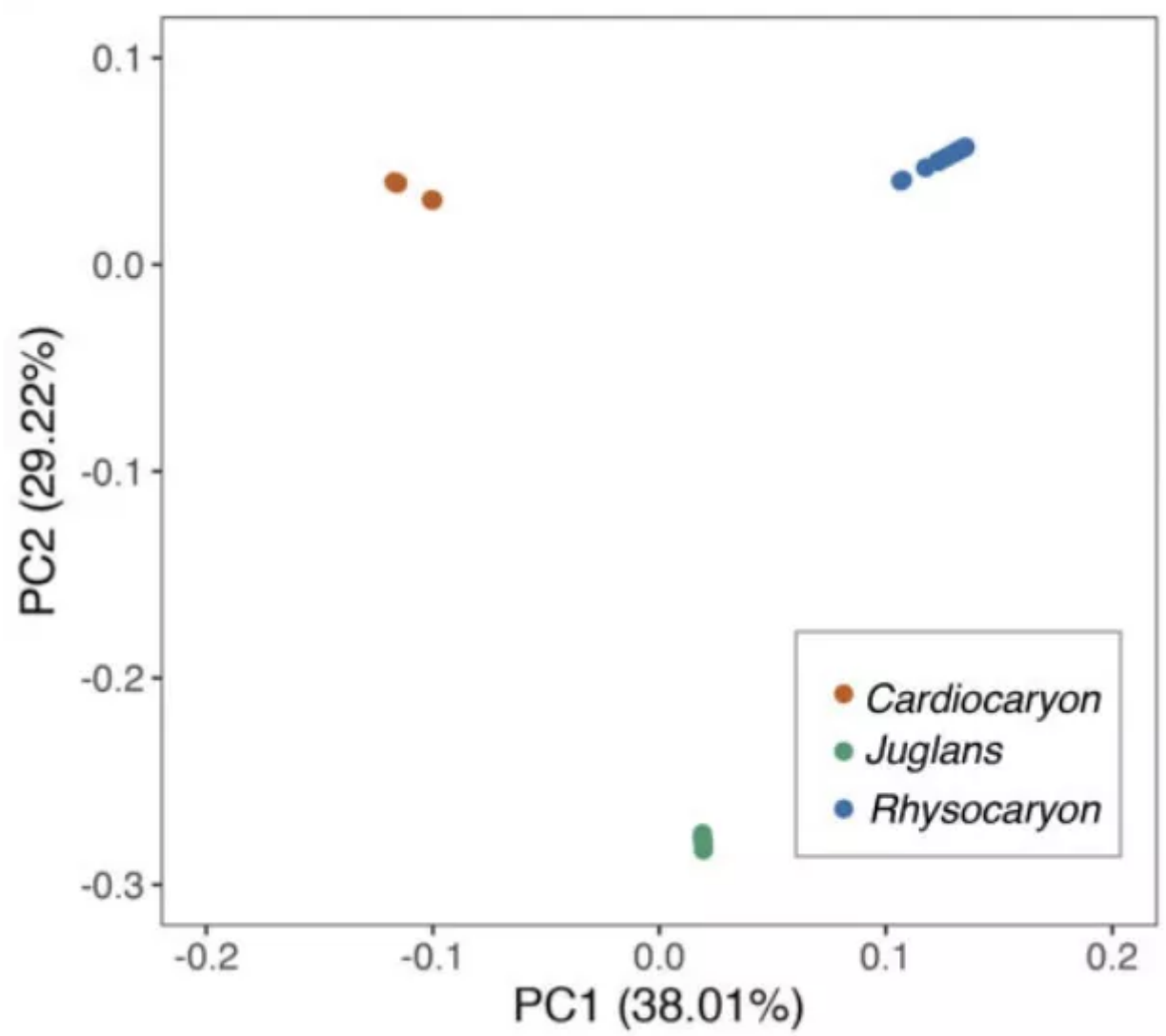
在分析过程中，PCA可以让我们非常直观地看出各个样本之间的相似性。例如在一张PCA散点图中，数个样本的点聚在一起，那么就说明这几个样本之间的相似性非常高；反之，如果几个样本的点非常分散，则说明这几个样本之间的相似性比较低。例如下图，几个组的样本对应的散点在组内呈现相互聚集的情况，说明组内的重复性比较好，样本数据非常相似，而组间则有较好的区分度。有的时候为了说明组内样本的相似程度，还会用一个椭圆将同一组的样本对应的散点全部囊括起来。



不过并不是所有的PCA结果都是这种明显的分群结果会比较好，还是要根据分析的目的而定。例如在GWAS分析当中，这种“天女散花”一般的PCA散点图，正说明了样本之间不具备明显的亚群分化，适宜进行后续的GWAS分析。



PCA能得到的信息不止于此，例如在群体进化研究当中，杂交种与其亲本进行PCA聚类的时候，杂交种会在PC1介于两个亲本之间，而在PC2上与亲本呈现较大的差异。下图即是一个典型的例子，杂交形成的栽培核桃与其两组亲本野生核桃的PCA分析示意。



栽培核桃与野生核桃的PCA聚类(Zhang et al. 2019)

通过以上几个例子，我想你已经清楚地知道了，在应对不同的分析目的的时候，PCA可以从不同的侧面对数据的状况进行整体的反映。当然了，手里有数据的话，也可以自己尝试进行对数据进行PCA分析。如果面对做PCA的诸多工具眼花缭乱不知所措的话，不妨上联川的云平台试试看，也许你能在这里找到新的天地。

联川生物云平台：

https://www.lc-bio.cn/overview

小姐姐告诉你2秒钟的PCA图2分钟就能学会

三分钟绘制一张优美的三维PCA图

参考文献：