五、研究自然遗传变异，用反向工程学绘制基因型-表型遗传图谱

在人群中发生的自然遗传变异可以在基因型和表型关系的研究中发挥巨大的作用。我们知道，每一对等位基因都可以通过直接或间接的方式对生物体的性状造成影响，但等位基因通过重组和分离又可以在子代中随机分布，所以通过比较个体对同一基因扰动（genetic perturbation）产生的性状之间的异同点，就可以推测出该基因扰动（变异）是造成这一性状表型（例如疾病）的原因还是该表型造成的结果。随后，将这些因果关系构建成一个因果网络，那么“基因型－表型关系图谱（genotype-phenotype map）”的大致框架便搭建起来了。

尽管人类步入基因组时代已有十多个年头了，但是，想要弄明白基因组数据的真正意义还是比获得这些数据要困难得多。全世界的科学家手上已经搜集了大量的基因、转录体、蛋白质以及代谢产物的数据。不过，阐明这些海量数据之间存在的复杂关系却并非一件容易的事情。当代生物学家的中心任务是将这些相互作用联系起来，构建一个可以预测生物体整体运行规律（模式）的模型。

下文中将要介绍如何通过对生物系统进行干扰试验来构建“因果网络”（causal network）模型。有了这个模型，人们就可以很容易地将基因型的改变“翻译”成相应的表型结果。在这个因果网络模型中，有很多因果关系是直接通过对共表达基因的相互关系或细胞组份之间的相互作用得出的，但还有很多因果关系则是对复杂的生理学、行为学、与环境的相互作用等等关系进行综合分析之后得出的（知识框1）。人们将通过观察现象推导出因果关系的方法称作反向工程（reverse engineering），这是因为此类研究的目的不仅仅是鉴定出那些在功能上相关或位点相近的分子，更重要的是希望能够弄清楚整个系统如何以一个有机整体行使功能。

知识框1 因果网络真的存在吗？

实际上，因果网络这个概念一直都存在争议。很多生物学家都对通过统计学方法发现的因果关系表示怀疑，也对因果网络模型到底有多大用处而心存疑虑。因此，很有必要对因果网络在研究基因型与表型关系上的价值进行一番讨论。

大部分人在进行生物学试验的时候都非常倾向于这样一种有关因果关系的日常理论，即一个因素影响另一个因素。而据此得出的结论又通常来自于假设和统计学计算。哪怕在进行单基因敲除这类简单的实验时，也会假定归纳推理法会得出正确的结论——基因型是固定的，它们会引起相应的表型，所有的变量都在试验掌控之中，比如使用野生型对照、单盲（观察者不了解情况blinding observers）检测以及随机重复试验等等。研究人员的这种信心来自于统计学检测，比如t检验中的P值等。他们相信通过统计学检测能发现野生型和突变型之间的差别，但实际上统计学检测本身就是基于一系列假设上的。

在推测概率因果网络（probabilistic causal network）时，使用的假设就更多了，但使用的概念框架（conceptual framework）都是一样的。在一次研究结束之际，会对因果关系得出一个结论，该结论是符合研究者设定的假设和统计学检测手段的。对于遗传学试验来说，大部分人使用的都是单基因干扰试验，而他们进行该试验的目的只是为后续试验找出候选基因。

因果网络是否存在是生物系“快乐时光（happy hour）”里的热门话题。现在更需要回答的问题是，在预测模型的因果网络中究竟哪个部分更重要，各种分子、相互作用、动力学，还是它们都重要或者还有更重要的？一个分子或一个生理过程可能是生物所必需的，但如果它们不发生变化就不会是我们需要的“原因”。从这个意义上说，遗传学中的因果网络实际上和遗传学家对遗传率（heritability，也称遗传力。它不表示一个性状是否会遗传，而只表示能够用遗传学变异来解释的该性状变异的比例）的定义是一样的。因此，从遗传学上分离的群落推导出来的因果网络取决于在该群落中有的遗传变异情况以及该变异在群落中的分布情况。在该框架下，使用因果网络就可以预测出相应的结果。

传统的反向工程方法是先敲掉一个基因，随后观察它对系统造成的影响，或者同时随机敲掉很多基因，然后观察影响。自从R. A. Fisher在1920年提出了有关数量遗传学的观点之后，统计学家就意识到，这种随机的多基因敲除方式实际上是最理想的研究基因型－表型关系的实验方法，而人群中自然发生的遗传变异刚好就是这样一种随机的多基因敲除模型，是最适合于研究的材料。近年，随着众多模式生物、农作物和人类基因组数据的不断积累，已经有人开始使用这些物种的自然变异数据进行基因型－表型关系的研究了。

本文将介绍这方面研究的最新进展。首先，将介绍基本的研究方法和步骤，并会将其与传统的遗传筛选法做一个比较。然后，会就该方法所面临的困难和缺陷做一番讨论。

1.对转录体丰度进行数量遗传学研究

对具有遗传特点的种群进行研究，是发现导致表型改变的基因变异的核心方法。一个常用的方法是将两个近交系品种进行杂交，得到在每个等位基因位点均为杂合型的后代。经过这样一轮典型的杂交试验，就会得到携带有成千乃至百万种遗传位点差异的后代。使用分子标志物可以知道这些子代遗传位点来自亲代基因组中哪一个片段，最终就能发现是哪些遗传位点（即数量性状位点QTL）变异能对表型造成影响（图1 a-d）。

基因型与表型关系研究的最新进展是检测两者之间关联性的“表型状态（phenotypic state）”。其中，最主要的就是对每个待测基因转录体的丰度进行检测。用数量遗传学在全基因组范围内进行转录体丰度的研究有时也被称作遗传基因组学（小词典1），或者叫做数量性状基因表达谱作图（expression QTL mapping）。对基因及其转录体表型之间的关系进行研究，所得到的分析结果可以通过作图的方式——以对应每个转录体的基因在基因组上的物理学图谱位置为纵坐标，以转录体丰度中与遗传变异相关的基因位点为横坐标，直观地表示出来（图1d）。如果杂交体设计得当，那么研究发现的基因型和表型之间的关系就可以说明基因变异是引起表型改变的原因。不过，这种研究对科研人员的实验设计能力、计算机能力和统计分析能力都是一大挑战，尤其要注意避免人为假象的产生。

通常，在研究转录体丰度和基因之间的关系时都会发现很多是与结构基因相关。由于这些基因都有它们自己的QTL，因此，图1d中的点都集中在对角线上。这些相关性大部分都属于顺式作用调控多态性（cis-acting regulatory polymorphisms），只有少部分属于反式作用（in trans）调控。

图1d中呈垂直方向分布的点表示的是“连锁热点（linkage hotspot）”，基因组中该区域发生改变会影响多数转录体的丰度。因此，这些对表型效应（phenotypic effect）具有“强大影响力”的基因位点很有可能是能够在很大程度上影响表型的调控位点。另一种可能是，这些位点的基因发生变异，只会对转录体丰度产生非特异性的影响。这种多效等位基因（pleiotropic

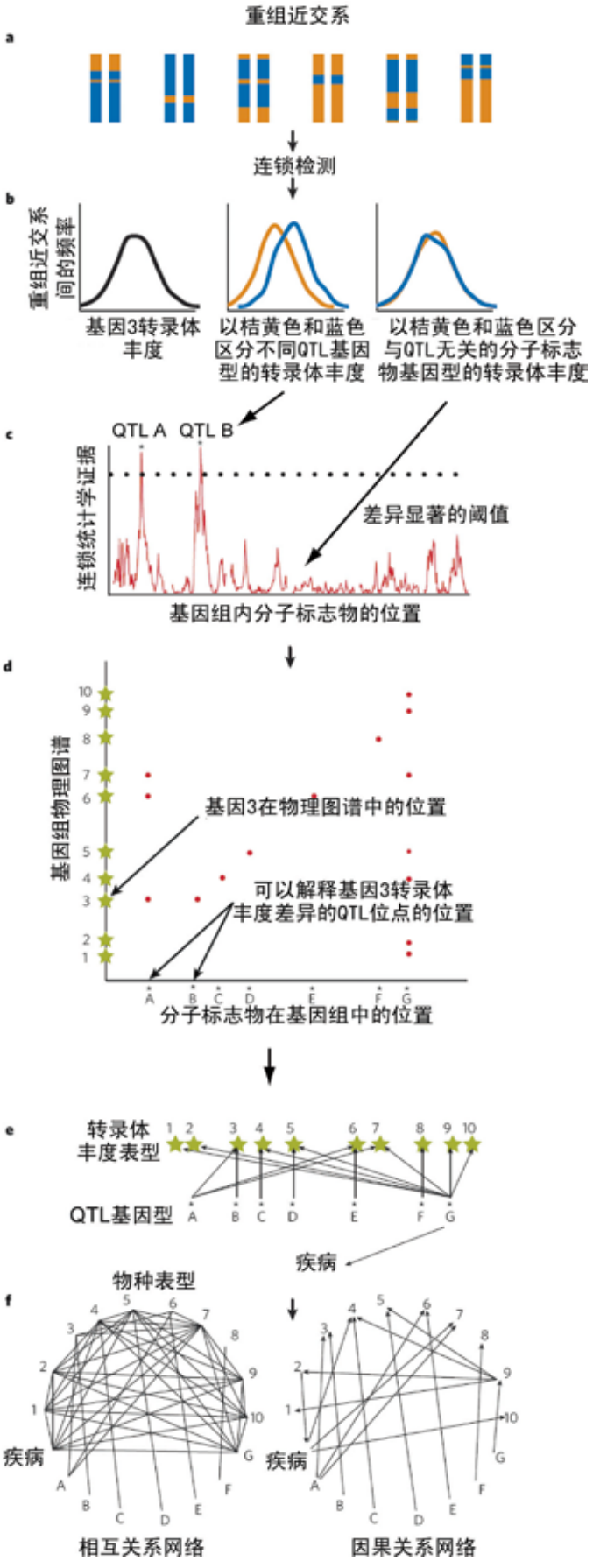

图1 从随机遗传（genetic randomization）到因果网络（causal network）。遗传学上随机分布的种群，例如重组近交系两个亲本基因组的片段在该种群的染色体上是随机分布的，如图中所示的桔黄色和蓝色（a），此图是对表型进行连锁分析的开始。本例中，表型是基因3的转录体的丰度。基因3转录体的丰度在不同的重组近交系中会有所不同（图b左小图）。近交系基因组中的每一个位点都会被检测是否能影响基因3的转录体丰度（图b中间以及右边的小图）。经过统计学分析发现，有两个区域和基因3位点连锁（如图c中星号所示）。这两个区域被称为QTL。如果用同样的试验来检测更多的转录体丰度表型（比如，不止基因3而是基因1~10），就可将获得的能够影响转录体丰度的QTL位点和基因在物理图谱中的位置做成图d那样的图。图中横坐标上的星号表示QTL。纵坐标上的绿色星号代表基因1~10。图d中，沿对角线分布的红点表示局部连锁（local linkage），通常都由顺式作用调控多态性（cis-acting regulatory polymorphism）引起。图中沿纵轴分布的点则表示连锁热点（linkage hotspot）。图e表示的是高水平的因果网络。网络中QTL的变异能引起相应转录体丰度的改变。这些转录体丰度QTL也能与其它（例如疾病）等表型共定位。为了图示方便，只标出QTL G与疾病有关。使用反向工程学构建因果网络就是为了将表型构建成变量，例如判断由QTL G影响的转录体丰度是病因还是结果。虽然性状之间的关联非常复杂（如图f所示），还是能发现QTL G能直接影响基因9。而基因9的转录体丰度又能影响基因1、2、4、5。基因2的转录体丰度是造成疾病的原因，该疾病又能影响基因4、7、10。很多转录体都与该疾病有关，但只有基因2和9的扰动能对该疾病造成影响。

allele）有可能改变细胞内环境的稳态，对多种性状产生影响。

前面提到的使用自然突变资源的数量遗传学方法就比传统的一次改变一个基因的方法要好得多。

首先，使用该方法获得的数据重复性高。初次杂交体中每一个等位基因的作用都可以被反复检测，因为不管这些子代的表型有多么不一样，它们都携带同一个等位基因。以一个只有100个样本参与的研究为例，该研究对每一个等位基因平均会检测50次，而这在传统方法中是无法想象的。

其次，由于这种自然发生的突变是在多个位点同时发生的，因此有助于人们了解它们之间的相互作用关系。通过观察发现，很多转录体丰度相关性状的遗传特征就受基因间相互作用的影响。一个很好的例子就是遗传冗余现象（genetic redundancy）。只有同时将冗余位点也给突变掉，才能观察到该基因的作用。

最后，同时干扰好几个因素相比只改变一个因素能观察到更多的表型改变。复杂的遗传性状都是由多个位点构成的，往往呈现出超亲分离（小词典2）现象。而超亲分离正是导致杂交子代不同转录体丰度性状的原因。种群中大量的表型有助于发现不同性状之间的联系。比如，从这种种群得到的转录体丰度数据通常就能够发现基因间的相互作用关系。

2.因果顺序关系

QTL可以对性状产生直接影响，也可以通过影响一些中间性状对某些性状产生间接影响。于是就产生了一个有趣的问题——表型的改变（例如疾病状态或行为）到底是某个（些）基因转录体丰度改变的原因还是结果呢？不过，只有当转录体丰度改变是造成表型改变的原因时，它才有可能对有机体的性质造成影响，这一点无论在实验室、临床研究还是自然进化的种群中，都是适用的。

在设定了一系列假设的前提下，就会发现这种因果关系了，而且这种关系具有条件独立性（conditional independence）。下面，以A、B、C 这3个相互之间有因果关系的性状为例来具体说明。在标准的马尔科夫假定（standard Markov assumptions）前提条件下，如果性状A的改变造成了性状B的改变，然后又造成了性状C的改变，即A→B→C。这种情况下，如果知道性状B的情况，那么A对C的作用实际上是不需要考虑的，即A和C相对于B来说都是有条件的独立性关系。不过，如果A、B、C之间的因果顺序关系变成B→C→A，那么上述有条件的独立性关系就不存在了。不过，仅靠这种有条件的独立性关系是不能定义一个因果关系的，比如在A←B→C和A→B→C这两种条件下，它们有条件的独立性关系是一样的。不过无论如何，有条件的独立性关系在发现没有直接联系的不同性状之间的直接联系关系的时候还是非常有用的一种工具，哪怕在它们之间的相互因果关系不清楚的情况下（例如不知道是B→C→A还是A→B→C）也能发挥作用。

使用“基因干扰（genetic perturbations）”技术的好处就是因为有这样一条中心法则：基因型的改变会造成表型的改变。但是至少从单个的个体来看，表型的改变却不能反过来影响基因型。因此，如果杂交试验设计得当，通过改变基因型应该是能够改变表型，从而发现它们之间的因果关系的（图2）。

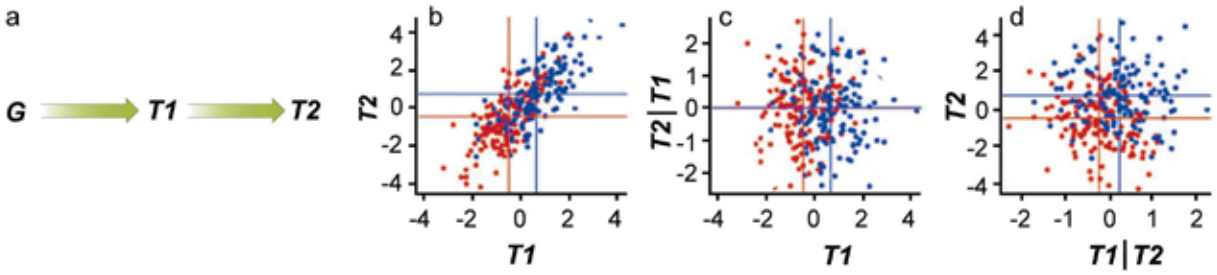

图2 因果的顺序能导致有条件的独立性（conditional independence）。进行因果推导研究的理论基础可以用此图表示。假设一单倍体种群携带有单因果基因位点G，它有两个等位基因。该位点的等位基因状态（allelic state）导致了相应转录体（T1）丰度的不同。同时，其它的一些例如遗传的、环境的和随机的因素也会对T1造成影响。如果使用β_1表示等位基因效应，使用ε表示其它影响因素，例如正态分布噪音，G表示基因型的指示变量，那么$T1=\beta_1 G+\varepsilon$。T1的丰度的改变会引起其下游性状T2的改变，同样T2也会受其它因素影响，因此可以将T2表示成$T2=\beta_2 T1+\gamma$。T1和T2这两个“因果”顺序之间的关系就如a中所示，T1、T2的值则如b中所示。图中的数据代表了300个检测样本。每一数据点都根据基因型被标上了颜色（蓝色表示G的一个等位基因，红色表示G的另一个等位基因）。每一个性状的平均值也都被与相应基因型颜色相同的彩色线所表示出来（上图中基因型是指示变量，红色值为-1，蓝色值为1，$\beta_1=0.5$，$\varepsilon \sim N(0,1)$，T1的均值为0。对于T2来说，$\beta_2=1$，$\gamma \sim N(0,1)$，因此，T2就等于T1加上其它影响因素）。结果发现性状、T1和T2之间互有关联，同时都与基因型有关（b）。T2和G之间的关系受T1调控，T2比T1的值与基因型无关，每一个基因型的平均表型都是一样的（c）。相反，T1比T2的值则与基因型有关（d）。

有好几种方法可以用来分析分离群体（segregating populations）中表型的因果关系。比如，在对QTL进行作图的同时构建因果关系模型；使用正式的统计学方法验证直接的因果关系或者使用包含两种性状和一种QTL组成的数据集构建不同的因果关系模型，然后再借助信息论（information-theory）的标准比较这些模型从而找出最好的那个。

分析与同一QTL有关的不同性状之间的关系也能帮助人们发现该QTL中和这些数量性状相关的基因，而如何找到这些基因是当今数量遗传学领域面临的最大困难。很多这种分析除了要使用到“关系”信息之外还会使用到转录因子结合位点的相关信息、蛋白质之间相互作用的相关信息、该QTL中每一个基因序列多态性的相关信息等等。

在全基因组范围内使用表型相关性（phenotypic correlation）来发现因果转录体（causal transcript）是一个非常有潜力的方法。最近，使用该方法已经发现了很多和人类疾病相关的基因变异。不过，这些变异都位于基因组的非编码区，它们的功能也不太清楚。之所以能发现这些变异就是因为利用了转录体丰度性状和人群中疾病状况的关系结构。不过，在以人群为基础的关联图谱中，可以通过一些外部因素（比如人群的年龄或种族）发现基因型和表型的关系。因此，这些研究中的基因型研究结果并没有使用近交杂交系试验得来的更准确，不过可以使用相应的动物试验予以验证。

3.因果关系网络

要对生物体内的因果关系有一个系统的认识，就需要整合大量的遗传变异信息和表型信息。有好几种方法可以用于这方面的研究，其中使用最广泛的就是贝叶斯网络（小词典3）。所谓贝叶斯网络，就是有关随机变量（在本研究中对应的是表型改变）的一种网络结构图。使用该方法可以发现条件概率（conditional probability）事件。

不过，贝叶斯网络也存在一些问题。

首先，一个条件概率事件可以用好几个贝叶斯网络表示，所以对随机变量（本例中是转录体丰度）的观察不能排它性地确定出一个真正的、直接的关联关系；

其次，由于贝叶斯网络不是一个环状结构，因此缺乏反馈调节模式。不过，这到底会对贝叶斯网络分析方法造成什么不好的影响目前还不甚清楚；

最后，由于贝叶斯网络的可能空间太大，因此给因果关系网络建立运算带来了极大的麻烦。

不过，虽然存在这些问题，但基于以下两点原因，检测遗传学上分离群体的转录体丰度还是目前为止唯一适合用于构建直接贝叶斯网络的方法。原因为：

1. 一个性状和另一个能影响它的性状应该会受同一个遗传扰动（即QTL）影响。这样很容易就筛掉了其它大量的可能的网络结构，极大简化了问题；

2. 虽然构建大范围的因果关系网络还是很困难的一项工作，但如果结合遗传数据等资料，还是可以极大地提高预测质量的。

除了贝叶斯网络之外，结构方程式模型（structural equation model）也是用来分析分离种群间转录体丰度的方法。这些模型包括网络结构中的线性方程组（linear equations organized into a network structure），以及“装配有”同时起到预报器和效应器功能变量的线性模型。虽然结构方程式具有贝叶斯网络所不具备的反馈环状结构，它们还是需要标准的线性模型假设前提。因此，如果使用非线性动态模型来分析转录体丰度就会出现问题。贝叶斯网络在处理非线性问题时会将数据分成简单的几类，例如上调、下调、不变这三种类型，不过这种简单分类法也有其缺陷。

于是出现了另一种方法，使用配对性状关系先构建一个简单的网络，然后再对条件依赖关系进行检测，以对该网络进行修正。使用这种方法有一个明显的优势，就是可以发挥计算机的运算优势。因为基因间的关系都具有模块性（modularity），所以配对研究方法可以将极其复杂的大问题分解为各个模块内的小问题进行解决。

最近，两个有关人类和小鼠疾病的研究项目就使用了上述模块化处理方式，将转录体丰度数据按性状相关模块形式进行了分割。科研人员比较了转录体丰度和疾病表型数据，来判断哪些转录体是导致疾病的原因。在小鼠和人类中都发现了一个模块，不过还需要后续的试验予以证明。虽然这和反向遗传学的要求还差得很远，但现在积累的这些经验（将基因组范围的数据缩小到只有两个性状的数据）还是非常有价值的。

4.数量遗传学是进化遗传学

在个体中，基因型决定表型这一中心法则有助于我们发现直接的因果网络关系。不过，在种群中，表型又会反过来选择基因型。因此，自然变异就形成了一个天然的非致命性遗传变异资源库。在经典的基因型－表型关系图谱无穷小模型（infinitesimal model）下，从自然变异得出的结论有可能只是变异的“副产品（epiphenomenon）”，因为这些自然变异造成的效应并不像人工随机变异那样会带来明显的效果。

在典型研究中，选择效应对QTL的数目和种类的影响是明显的。James Ronald和Joshua Akey发现在酵母杂交系中，在局部QTL中的基因比例要比预计的少。这说明还是

有较低水平的负选择（negative selection）在发挥作用。同样，在众多变异基因中，真正能对重要生理过程起关键性调节作用的基因也为数不多。例如，我们在寻找能改变基因表达水平的基因时，编码转录因子的基因肯定是最有可能符合条件的“候选者”。但实际上，在表达水平QTL（expression QTL）中，这些基因并不多。不过无论如何，哪怕编码转录因子的位点并没有发生基因变异，只要能反映该转录因子活性的指标（例如相应转录体的丰度）的确与该基因型及其相应表型有关，我们还是可以用因果推理的方法来发现转录因子和目的基因之间的联系的。

不过，根据试验研究对象种群的不同，“选择（selection）”方法在筛选遗传变异对表型影响中的作用也不尽相同。近交系杂交（Inbred line cross）种群中往往会有好几种遗传背景不同的群体。因此，我们应该选择能最大化表现表型差异或基因型差异的种群。与种群多样性有关的等位基因的作用也有可能和那些导致种群中最常见遗传多样性位点的作用有所不同。在不同近交系之间的杂交后代中，最有可能发现少见的但作用明显的突变，即连锁热点（linkage hotspot）区域。这种杂交种群有可能会表现出偏向对表型具有较大“影响力”的基因变异，这些基因变异不是单独起作用，而是互相结合着起作用，即是一个遗传变异互相适应（coadaptation）的结果。自然界中存在着广泛的遗传相互作用（genetic interaction），这意味着要建立起基因间的因果关联模型非常困难，虽然最近研究人员在这方面取得了一点进展。

自然变异也是会被种群中与这些变异无关的遗传现象（genetics phenomena）所影响的。而与其它自然选择靶标变异紧密相连的遗传变异在进化过程中是和自然变异相伴发生的，因此，局部重组频率（local recombination rate）和遗传变异水平之间有了联系。同样，人工诱变重组（mutagenic recombination）也能造成同样的结果。位于重组频率较低位点内的基因对遗传的影响比那些位于重组频率较高位点内的基因要小。因此，如果基因位点与其重组频率有关，那么某些功能基因变异对表型造成的影响可能就不会那么明显。

5.遗传系统生物学的前景

构建因果网络还面临许多问题，即使构建成功了，用该网络来预测分离群体中的基因表达情况时还会碰到各种各样的问题。虽然已经有一些论文介绍了几起这方面工作的成功案例，但我们还应该想到那些还没有发表论文的人遇到了多么大的困难。这其中有些问题人们已经有了清晰的认识，不论是从试验方法来说还是从理论分析方面来说，他们都对基因型－表型谱有了更深入的了解。

在使用基因表达数据来绘制因果关系网络图时有一个假设前提，那就是检测各种性状时出现的误差是相同的。如果有一个“因果性状（causal trait）”没有被发现，但受它影响的另一个性状却被检测得很清楚，那么使用这个被检出的“效应性状（effect trait）”来检测“因果性状”比直接对“因果性状”进行检测的效果要更好。在这些情况下，使用条件相关方法（conditional correlation approache）会得到相反的结果（图3）。

通过分析可以得知，出现这种情况，从生物学角度来看是很正常的。比如，细胞中调控因子的数量实际上并不多，但它们的作用效果却可以通过级联方式被放大。因此，如果细胞中编码低丰度转录因子的基因发生变异，就会影响到编码高丰度结构蛋白转录体的数量。问题是，我们要检测这些高丰度转录因子的数量

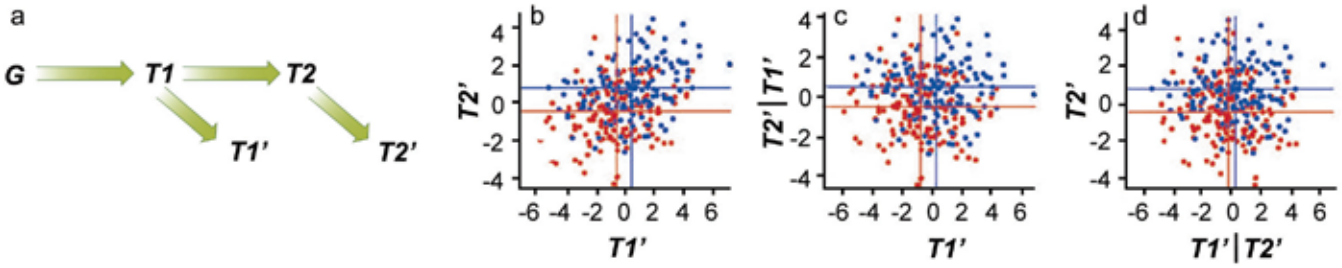

图3 测量误差会混淆因果关系。在图2中，本文已经从检测样本、参数和条件等方面介绍了测量误差对因果关系的影响。简单来说，一单倍体个体有一个因果基因位点G，该位点有两个等位基因，该位点的等位基因状态会造成T1表型的变化，继而会造成T2表型的变化（T1、T2皆为转录体丰度表型）。a：T1、T2的测量值加上正态分布误差得到T1’和T2’。对于T1’来说，正态分布误差是2，而T2’的是0.2。因果顺序如图a所示。b：T1’和T2’互为关联，同时也都与基因型（蓝色表示G的一个等位基因，红色表示G的另一个等位基因）有关。不过条件相关性（conditional correlation）会让我们得出错误的因果网络（如a所示的那样）。c：在考虑T1’的情况下，T2’还是依赖于基因型（如图2中c所示T2比T1的值与基因型无关），由此我们可以得出a中所示的因果顺序。d：在考虑T2’的情况下，T1’几乎不依赖于基因型（如图2中c所示，T1与基因型有关），我们也能得出a中所示的因果顺序。因此，由于测量误差之间的差异，所以用T2’检测T1比用T1’检测T1的效果更好。

很容易，但这些低丰度转录因子的数量往往都会低于目前检测手段的阈值，根本不能被检测到。使用芯片试验有望解决这个问题。对同一个样本的转录体丰度进行多次独立的检测，也可以得出一个稍微准确一点的数据。由于上述种种原因导致我们在构建因果关系网络时有可能会出错。因此，我们应该清楚基于因果关系网络模型做出的预测都是建立在一系列假设前提下的可能发生的情况，最终还是需要用试验来进行验证的。

还有一个需要注意的地方，就是我们用来构建转录体丰度图谱的数据都来自混合细胞样本。因此，测得的数据实际上表示的是混合细胞的情况，而这些情况是由细胞的发育状况和分布状况决定的。比如，在用非同步培养的酵母（即包含有处于细胞周期不同阶段的酵母）做研究对象时就是如此。同样，用动植物的组织或器官作为研究对象时也是如此。我们在研究具有不同遗传背景的小鼠造血干细胞的分裂周期时间不同时，就利用了混合细胞的优势，发现了不同细胞系中哪些基因的表达有差异。

另一个需要注意的问题，就是我们从分离群体（segregating population）得来的数据都只能表示静止的状态（static representation），但缺少时间相关信息（time-series data）就不可能了解动力学状况。我们发现的因果关系也都是与静态表型相关的基因型。在一个反馈系统中，很有可能包含大部分的生物学稳态（steady state），而被我们发现的“因果”往往只是众多能改变表型平衡态（phenotypic balance）原因当中“胜出”的那一个原因。我们在判断一个新的基因变异是否能影响生物体稳态的时候，上述这种“低维度（low-dimension）”的预测模型还可以应付，不过对于真正的反向遗传学来说，还远远不够，还需要更完整的模型。已经发现时间相关数据在解决因果关系网络推理过程中出现的常见问题非常有用，在研究遗传变异时也将时间相关数据考虑进来同样能获得意想不到的结果。

很多构建因果网络模型的方法都会将所有的“因果变量（causal variable）”容纳到模型当中，但这种“完整性”对于大多数生物系统来说其实并没有太大的意义。有很多表型原因都有可能因为检测不到而被忽略，比如未注释的基因、参考基因组中没有的与结构多态性相关的基因以及小RNA的转录体等等。此外，代谢产物的丰度也是经常被忽略的重要因素。还有一些与转录过程无关的调控事件，比如翻

译后调控事件也很容易被忽略。有一些因果关系还可以从器官的发育历程中发现，因为遗传变异对个体的一生都会发挥作用。比如，一个在器官形成早期发挥作用的基因的表达与表型之间的关系，在当时进行检测时无法检测出来，但两者之间的关系却会一直存在。大部分潜在影响基因表达性状的遗传变异都检测不出来，因为它们的“影响力”都不大。要发现它们的最简单的办法就是扩大检测样本量，人们现在也正在做这方面的努力。

生物学中“原因”的种类实在是太多了，大部分试验都只能发现其中很小的一部分（知识框1）。对多种环境因素中通用体系（common lines）的研究给人们提供了一个将遗传变异因素和环境影响因素结合到同一个因果关系网络模型中的可能。分析多个组织中转录体的丰度与研究情景依赖（context-dependent）因果网络同样重要，因为每一个细胞型实际上都是为基因组提供了不同的环境因素。将这些数据收集起来，研究人员就能研究更多条件背景下的因果网络了。

自然变异是进化的基础，是导致遗传病的原因，同时也是各种生物学理论的基石。因此，自然变异是一个非常理想的研究多因素干扰的试验材料，它能帮助科研工作者发现基因型与表型背后的联系。当基因组学发展到系统生物学阶段，自然遗传变异将起到更加重要的作用。

原文检索：Matthew V. Rockman1. (2008) Reverse engineering the genotype–phenotype map with natural genetic variation, *Nature* 456 (7223): 738-744.

YORK/编译

小词典

1. 遗传基因组学（genetical genomics）

遗传基因组学，是指将微阵列技术和数量性状位点（QTL）分析结合起来，在全基因组水平上定位基因表达的QTL（eQTL）。 它为研究复杂性状的分子机理和调控网络提供了全新的手段。
遗传基因组这个概念和研究策略在2001年由Janson和Nap首先提出。到目前为止，遗传基因组学已应用于酵母、老鼠、人以及玉米等植物。研究结果表明：基因表达水平的差异是可遗传的复杂性状；eQTL可以分为顺式作用eQTL和反式作用eQTL。顺式作用eQTL就是某个基因的eQTL定位到该基因所在的基因组区域，表明可能是该基因本身的差别引起mRNA水平的差别。反式作用就是eQTL定位到其它基因组区域，表明其它基因的差别控制该基因mRNA水平的差异。将eQTL结果、基因功能注解以及多种统计分析方法相结合，不仅能更准确地鉴别控制复杂性状及其相关基因表达的候选基因，而且能构建相应的基因调控网络。

2. 超亲分离（transgressive segregation）

超亲分离，指在杂种的分离世代中，出现某种性状超越双亲的个体现象。在遗传学上，这是基于来自双亲的基因累加效应和补充效应。

3. 贝叶斯网络（Bayesian network）

贝叶斯网络是一种概率网络，它是基于概率推理的图形化网络，而贝叶斯公式则是这个概率网络的基础。贝叶斯网络是基于概率推理的数学模型，所谓概率推理就是通过一些变量的信息来获取其它的概率信息的过程。基于概率推理的贝叶斯网络是为了解决不定性和不完整性问题而提出的，它对于解决复杂设备的不确定性和关联性引起的故障有很大的优势，在多个领域中获得广泛应用。
贝叶斯网络又称信度网络，是Bayes方法的扩展。自1988年由Pearl提出后，已经成为近几年来研究的热点。一个贝叶斯网络是一个有向无环图（Directed Acyclic Graph，DAG），由代表变量节点及连接这些节点有向边构成。节点代表随机变量，节点间的有向边代表了节点间的互相关系（由父节点指向其后代节点），用条件概率进行表达关系强度，没有父节点的用先验概率进行信息表达。节点变量可以是任何问题的抽象，如：测试值、观测现象、意见征询等。适用于表达和分析不确定性和概率性的事件，应用于有条件地依赖多种控制因素的决策，可以从不完全、不精确或不确定的知识或信息中做出推理。