中国信息化周报/2014 年/8 月/18 日/第 007 版
趋势

解码基因大数据 培育健康产业

华大基因创始人 汪建

基因的大数据到底有多大?一个人的基因在细胞里面的综合是 6×1023，谁还有这样大的数据?每个人身所携带的信息，比现在已知世界上数据的总和还多。一个人的基因组一半基因从父亲来的，一半基因从母亲来的，一个基因组是 3×109，人体身上约有 40 万亿 60 万亿个细胞，A、T、c、G 每三种不同组合变成了基因控制生老病死，所有遗传密码都在这里。蛋白质身上有 20 个不同的氨基酸，在组合上有 2×10。更别提小分子的数量了。一个人的数据绝对不会少于 Google 地图的数据，可惜的是它没有被大家意识到。

高价值的基因大数据

这个大数据的特点是高价值，是人类最贵的价值，因为人的生老病死都在这个数据里面。再看看这种数据怎么获得，天上掉不下来，互联网找不到，需要人开垦，得人去挖掘。基因需要测血，需要测定蛋白质，小分子需要读出来。所有这些数据都要在细胞环境里活动，所以一个细胞就是一个基因生命的基本组成。

我把华大这些年做的基础的研究数据做了一个总结。通过数据可以看出，一晃十几年过去了，每两年我们的数据增长增加一个 0，所以我们将之称为摩尔定律的变种。这个生物学的数据增长远远超过了摩尔定律的增长，它带来的是什么?

“ 人 ”的大数据

大数据的目标都太大了，华大基因把目标定得非常的小，定到个人，以人为本，从我做起。目标很简单，就是建康长寿，不病不傻，具体目标是“ 我的健康我作主，我的生老病死我来掌握 ”。

今天中国工业经济发展使得环境不可承受，人的生命也多少不可承受了。中国每 10 秒钟就因为心脑血管病死亡一个人，高学历、高工资基本上等于高血压、高血脂。

以我自己为例。这些年，我都定期检测激素水平、维生素水平、氨基酸水平，缺什么就补什么，从来不吃任何保健品，下面来看看我长期维持得怎么样。

华大基因曾发表过一篇论文，证明了生物体肠道内的微生物和基因是人的基因的 100 倍，非常非常复杂，华大基因曾经用几十万个 CPU 算了几个月，终于搞明白了肠道微生物是什么。之后，我们创造了一种治疗方法，把肠道不好的微生物改一下，变成好的微生物。不同的饮食习惯、不同的基因和不同的肠道微生物情况导致健康情况不一样。

我把自己肠道微生物换了换，这样就真正实现了“ 我的基因我知道，我的健康我作主，生老病死我来掌控 ”。我的高山速降速度达到每小时 60 公里，还可以玩风帆、登珠峰。

我曾经和中国信息化最好的医院院长聊天，他告诉我他们医院有 60 个 T 的数据，而实际上，我一个人的健康数据量就有 4 个 T。

什么叫大数据?就是人人都要关心、天天都要注意的。我这些大数据怎么来的?从基因到蛋白质，到小分子，全套做下来的。

大数据为健康服务

所以，我们的第一个目标是管控代谢综合症，减少心脑血管病发生和死亡率，如果能够控制到死亡率为每 60 秒死一人，然后能做到 100 万个人的话，我们就能拥有 1018 的数据。这 100 万人的健康数据约等于 2012 年或 2013 年中国全国的数据量，它在科学上的突破在健康产业上可能实现 10 亿到 100 亿的市场。

再说每 10 秒钟死亡一个人的肿瘤。每个人患肿瘤的机会是 22%。2013 年底的统计显示，中

国 5 年的治愈率是 25%，欧美国家是 65%，美国人是 68%。美国肺癌发病率从 90 年代后期就开始下降，中国肺癌发病率持续增长。中国的生产方式到了必须转型提升的阶段，人们的生活方式也是。那么，大数据应该集中在哪里发力?太多的工业化会给国家未来带来什么样的影响?时代正在发生变化，工业进入瓶颈，新型的生物经济正在崛起。一个癌症出现要 15 年的时间，它是从一个基因发生变化到一个细胞发生变化，再到一堆细胞发生变化的过程。怎么打掉它?在它基因发生变化的时候就发现它、就把它抓出来。

以宫颈癌为例，HAV 病毒感染了女性宫颈上皮细胞，时间长了就可以变成癌症。现在如果华大基因在黔西南免费把 HAV 检测了，是不是就能基本控制宫颈癌，是不是可以同时把子宫内膜癌、卵巢癌、乳腺癌都控制了。如果这能够实现，我们对肿瘤早期就可以定性、定量和定位进行分析，这又是一个大数据。依然拿 100 万人做一个基数，依然是 1018 的数据。这个疾病如果在中国 0.1%的人口中做，就是中国最大的数据库，至少可以使癌症早期发现率提早一年以上，如果提早一年以上，5 年存活率至少可以提高 2-3 倍。

最后是 30 秒钟一个出现的出生缺陷，我国拥有 8000 多万残疾人，广东的贫困人口一半因病致贫和因病返贫。先天性盲人疾病都是基因病，通过基因检测是可以做到控制的。怎么从 109 数据中找到一个检基变化来预测疾病，这是一个大数据的过程。华大基因不久前刚与 301 学科启动了百万新生儿天力和联合基因筛选计划，即通过基因分析方法可以比较准确地预测预防新生儿缺陷。

这个计划同样是一个大数据，但很可惜，它依然是一个民间计划。如果对中国每一个新生儿进行这样的筛选，我国的出生缺陷会大大降低，这不光是一个大产业问题，更是一个民生问题和民心问题。

信不信由你，控制出生缺陷是不是千亿万亿的产业?心脑血管病是不是千亿万亿的产业?肿瘤个体化治疗是不是千亿万亿，抗衰老是不是千亿，女同胞的美容是不是更大的产业?我们希望在这些发展过程中，对中国的科学贡献应该有标志性成果，在社会贡献上，我们希望成为未来社会发展的行业标准。

很多人不信华大基因做的事有什么用，我们在过去也很难得到科技部门的支持和理解，但是有一个人很喜欢我们，我们签订了 16 项重要的合同，他就是比尔·盖茨。我说比尔·盖茨，我们不要你的钱，你提出一个项目，我们各拿一半的钱，共同为人类做点事情。

我对黔西南提出一个设想，服务民生，建设医学健康，服务集聚区健康发展配套产业。

贵州黔西南是国家基因宝库，大数据、大科学、大产业最重要的资源所在地，它有生物多样性，民族多样性，是疾病研究的宝贵资源。如果能做出 1018 的大数据来控制遗传性疾病，能够控制黔西南遗传性疾病，我们就能控制中国其他山区的遗传性疾病，相信通过比尔·盖茨也可以推广到全世界去控制这些遗传性疾病。

我们依靠创新激动，依靠服务民生建立一个新的集聚区，来共同减少出生缺陷，减少肿瘤，减少心脑血管病，这三个病一起影响人类健康和生死的 80%。如果在 80%上有所贡献，就不愧对一生。

(本文根据华大基因创始人汪建在云上贵州·大数据国际年会的演讲内容整理而成。)