华大基因 BGI | HASHED OMICS 华大区块链

华大区块链白皮书

V1.0

数字化生命的价值交互

时间：二零一八年五月

前言

大千世界的万物生长和生老病死都受控于生命的遗传密码——基因组。基因，记录了地球生命几十亿年的演化历史，基因资源的多样性是生命进化和物种分化的物质基础。亿万物种的多样化基因信息就是最天然的、高度压缩的，包含了时空维度的分布式数据库，海量生命数据的相互融合、变异与传承，共同造就了纷繁复杂的生物多样性。科学家们正是通过基因测序技术将物种数字化，然后进行数据分析和价值挖掘，用硅基比特记录和分析碳基生物沧海桑田的变迁，从而认知人类自身与地球万物。

1999 年 9 月 9 日，华大基因伴随着国际人类基因组计划“中国部分”的正式启动而诞生。2001 年，由美、英、法、德、日和中国 6 个国家超过 3000 名科学家共同参与的第一个人类基因组草图绘制完成，耗时 13 年，所用总资金超过 30 亿美元。十几年来，基因测序的成本以“超摩尔定律”的速度下降。2015 年以来，华大基因实现了核心测序工具的突破，相继推出多款国产自主测序平台，以 600 美元的低价引领个人全基因组测序进入百元美金时代。目前全球已测序的高等动植物中（含未公布）约 70%由华大基因和合作者共同完成。2018 年 4 月，被誉为生物“登月计划”的地球生物基因组计划（Earth BioGenome Project，EBP）正式对外发布，华大基因也是该项目联合发起单位，计划在 10 年内对 1000 万至 1500 万种的已知真核生物基因组进行破译，总经费预计需要 47 亿美元。通过数字化地球，全面破译解读亿万物种的遗传密码，推动人类在认知、理解、利用和保护生物多样性等方面迈上全新起点。

随着测序技术及多种生命数字化工具的突破，以大数据驱动的“精准医疗”和“精准健康”时代正在来临。生命是一个多层次的复杂系统，在时空中动态演变。华大基因创造性地提出了以“大人群生命组学大数据（2B4D）”的方法论来认知生命，即从 DNA 开始，遵循生命中心法则，从基因组到蛋白组到跨组学贯穿，从微观到宏观、从生到死的跨尺度、多维度、多模态、全方位、全周期的海量全景式生命大数据解读。从国家层面看，2016 年 6 月，国务院办公厅发布《关于促进和规范健康医疗大数据应用发展的指导意见》，首次将健康医疗大数据提升到国家战略层面。同年 10 月发布的《“健康中国 2030”规划纲要》，也特别强调发展健康产业和医疗大数据、培育健康医疗大数据应用新业态。然而，该领域一直面临协同开放程度不高、权属不明确等问题，数据的采集、生产、存储、传输及计算分析等各流程都涉及到公民的隐私保护和数据安全问题，提升跨机构数据共享的互操作性，最大化发挥数据价值，都成为了刚性迫切的需求。

区块链作为一种集合分布式存储、点对点传输、共识机制、加密算法等技术的组合式创新应用，为个人数据自治和跨机构共享交换，提供了崭新的解决方案。华大基因自 2018 年起开始布局区块链技术应用并加入了超级账本（Hyperledger），由内外部实际场景驱动，逐步推动技术落地，最终目标是与行业伙伴共同搭建基于区块链及密码学技术的数据流通生产级基础设施，以隐私保护为前提，以数据共享为目的，确保全流程可控制、可审计、可监管，从而支撑民生普惠、科学探索和产业应用，使相关主体（个人、政府、医疗机构、科研机构、国家基因库、企业等）共有、共为、共享，在生命时代将个人生命数据确权并将数据资源资产化，形成生命价值可定价、可流通、可交换的全新生态体系。

人类社会正由工业、信息时代加速迈向生命时代。“生优病少、健康长寿、温饱不愁、环境友好”，是美好生活与命运共同体的基础，是人类发展的最大刚需和终极目标。华大追求终极、挑战极限，依托生命“读写存”高效低成本工具，聚焦生命的解读、编写与合成，把“基因科技造福人类”的伟大目标转化为从我做起的内生动力。万物相形以生,众生互惠而成。基因是人人与生俱来的数据资源，存在即合理，人与人之间 0.1%的基因差别定义了人类的多样性，核心正是万物共生与协同，这与区块链所倡导的分布式价值交换不谋而合。与工业、信息时代单纯追求效率、成本的极限不同，生命时代的价值评估更加多维度与多样化。

为推动区块链在生命大数据行业的应用，加速行业伙伴对其应用理解，华大区块链项目组编写了《华大区块链白皮书 V1.0》。白皮书总结了华大对于区块链的理解及相关核心技术积累，分享了区块链及密码学在生命大数据共享、生物智能计算、个人健康激励等实际应用场景的实践案例，并提出了相关建议。白皮书内容详尽、分析透彻、落地场景扎实，具有较好的参考价值。我们认为，从基础设施建设到最终大规模应用，没有捷径可走，在经过基础研发、概念验证、节点部署、开发测试等一系列流程后，华大逐步将生产级的联盟链 BaaS 平台开放给内部各体系及外部合作机构，共同推动行业发展，打造基于区块链的共赢生态。

如果说生物技术（BT）的突破实现了生命大数据的精准量化，以区块链等为代表的新一代的信息技术（IT）则为生命价值的定价与交换提供了必须的基础设施。我们坚信，“BT+IT”，两者深度耦合与互相反馈，必将共同构建生命时代全新的多方协作体系，进而推动人类“健康长寿”终极追求的实现。

华大区块链团队　　2018 年 5 月 27 日

编委会成员

策划顾问：

陈芳、单日强、侯勇、蒋慧、金鑫、李波、刘健、刘靓、刘娜、苗继业、

宋浩、伍利、徐军民、张勇、张玉良、宗洋

研究撰写：

杨梦、潘光明、赵宏德、李艳、潘远航、周童、唐强、李华平、李航、谢青、

李鹏程、曹威、吴家胜、肖鹏、田巍、李亚星

视觉设计：

刘斌、邹康

目录

咨询与合作 ： CHAIN.GENOMICS.CN

导论

近年来，人类基因组计划催生的高通量测序仪，结合质谱仪、高分辨率影像系统等生命组学工具的日益成熟，人们不仅可以将对生命的解读从微米、纳克的尺度推进到纳米、道尔顿的分子和原子级别的微观观察极限，更可以遵从生命中心法则，从基因组到蛋白组到多组学贯穿，实现从微观到宏观、从生到死的跨尺度、多维度、多模态、全方位全周期的海量全景式生命大数据解读和研究，因此以基因组学为基础的个人生命跨组学大数据将迎来爆发式增长。以全基因组测序为例，一个人的全基因组包含 30 多亿个碱基对，折算成测序数据量至少需要 100GB，如果考虑跨组学数据，预估未来每人每年的数据量将达到 1TB，终生数据量将超过 10TB。生命大数据的复杂性，需要百万大人群的数据积累和比较才能总结规律和得出可靠认知，70 亿人的跨组学数据将形成最庞大的数据资源。从健康医疗大数据角度看，IDC 预测到 2020 年，数据量将达到 40 万亿 GB，是 2010 年的 30 倍。生命大数据的发展不仅会颠覆原有医疗保健模式，更将推动循证医学向精准医学转变，为政府公共卫生决策提供支持，促进个人精准健康管理。

海量生命大数据的安全共享和高效应用是精准医疗的基础和生命时代的刚需。但数据的不合理使用通常会导致个人隐私泄露风险，也涉及商业秘密及国家安全问题，使数据共享成为一把双刃剑。目前，个人的医疗健康数据往往由具备数据采集和处理能力的医疗或科研机构进行集中管理。这种管理方式造成了两类问题：其一，个人对于自己的数据没有实际控制权，即便机构有将数据用于科研或交予药厂、保险公司进行商业变现等行为，数据所有者也无从知晓；其二，由于信息系统标准差异与利益格局的障碍，不同的医疗、科研单位、政府、商业机

构之间无法实现数据的安全交换与高效共享。不仅导致了数据的孤立分散，也无法充分挖掘数据价值，尽管每天都有大量的数据产生，但通常无法被整合形成完整的跨组学数据集合并加以分析，进而无法实现精准的健康干预。

深圳国家基因库于 2011 年由国家发改委等四部委批复，并于 2016 年正式对外运营。基因库依托华大基因研究院组建、运营，目前已初步建成“三库两平台”的结构和功能，包括生物样本资源库、生物信息数据库、生物活体库以及数字化平台、基因合成与编辑平台，实现对生物资源和信息的“读写存”，其规模、结构、布局和内容不仅具有世界领先性，更具有唯一性。国家基因库聚全球之力，以“共有、共为、共享”的宗旨和目标实现基因及数据资源的共享利用，为物种多样性提供保障，为精准医学提供大数据支持，在生命时代引领健康人生。

华大基因作为全球最大的基因组学研发机构，肩负着促进生命数据价值流动的责任，华大基因重视个人数据的安全共享与隐私保护，前瞻性地将新兴的区块链及现代密码学等技术架构应用于生命大数据的流通中。从技术角度来讲，区块链可以保障数据信息不可被任何中心化平台非法使用、篡改和删除，使得数据交互方可以不依赖第三方机构进行价值传递，并保证交易记录公开透明、不可篡改，极大地降低信任成本，提高交易效率，形成高效的多方利益分配体系，并为数据共享进行安全、透明的追溯审计。而以安全多方计算为代表的现代密码学技术可确保在保护数据所有者利益和隐私的前提下，实现多方数据可信交换和协同计算，联合挖掘数据价值。华大区块链的目标是搭建 2B4D 数据的共享开放及价值实现的 IT 基础设施，支撑科学探索和产业应用，最终形成一个信任、高效、安全、多方协作的大数据应用生态体系。

1 区块链简介

区块链是利用块链式数据结构来验证与存储数据、利用分布式节点共识算法来生成和更新数据、利用密码学的技术保证数据传输和访问控制的安全、利用由自动化脚本代码组成的智能合约来编程和操作数据的一种全新的分布式基础架构与计算范式。目前，区块链被很多大型机构称为彻底改变业务乃至机构运作模式的重大突破性技术。在金融、物联网、公益慈善、医疗健康、供应链等领域，越来越多的企业机构开始探索区块链在行业中的应用前景，规划基于区块链技术的数据流通路线图。

1.1 区块链兴起与演变之路

区块链技术起源于化名为“中本聪（Satoshi Nakamoto）的学者在 2008 年发表的奠基性论文《比特币：一种点对点电子现金系统》。文章提出，希望可以创建一套“基于密码学原理而非基于信用“的电子支付系统，任何人可以在不知道对方背景信息的情况下进行交易，且不需要第三方的介入。

这篇文章催生了比特币，标志着人类社会的货币体系的全新实验。众所周知，比特币在没有任何中心化机构运营和管理的情况下，多年来运行非常稳定。其原因就在于比特币的发行方式都是由程序和加密算法预先设定后，在全世界的多个节点上运行，没有任何人和机构可以篡改，不受任何单一用户控制。后来，人们把这种基于密码学与分布式存储的底层技术抽象提取出来，称之为区块链技术。

2013 年，19 岁的 Vitalik Buterin 发布了题为“以太坊白皮书：下一代智能合约与去中心化应用平台”的白皮书，提出基于通用的编程语言来创建各种各样的分布式应用，被称为“世界计算机”。2015 年，Linux 基金会发起 Hyperledger（“超级账本”）开源项目，众多金融机构及 IBM、英特尔等巨头加入合作。2016 年起，区块链技术开始从加密数字货币向更多应用场景扩展，引发了全球区块链应用浪潮。2016 年底，区块链技术首次被列入国务院《“十三五”国家信息化规划》，2017 年工信部发布中国首个区块链标准《区块链参考架构》。

区块链根据应用场景和设计不同，主要分为公有链、联盟链和私有链：

(1) **公有链**：以比特币、以太坊和所有数字货币为代表，各个节点可以自由进入或退出区块链网络；

(2) **联盟链**：各个节点通常代表实体组织机构或个人，通常需要经过授权后加入或退出网路。由于各机构间通常存在相关利益，因此需要各方共同参与和维护；

(3) **私有链**：各个节点的准入和退出权限均由内部控制，通常是在特定机构内用于内部数据管理与审计。

1.2 区块链主要优势特点

现有的区块链技术主要包含以下四个特点：

(1) **去中心化：**无需第三方介入，实现点对点的交易、协调和协作。在区块链系统中，没有任何一个机构或个人可以实现对全局数据的控制，而任一节点停止工作都不会影响系统整体运作，这种去中心化的网络将极大地提升数据安全性。

(2) **不可篡改性：**区块链利用加密技术来验证与存储数据、利用分布式共识算法来新增和更新数据，区块链需要各节点参与验证交易和出块；修改任一数据需要变更所有后续记录，修改单节点数据难度极大。

(3) **公开透明与可溯源性：**写入的区块内容将备份复制到各节点中，各节点都拥有最新的完整数据库拷贝且所有的记录信息都是公开的，任何人通过公开的接口都可查询区块数据。区块链中的每一笔交易通过链式存储固化到区块数据中，同时通过密码学算法对所有区块的所有交易记录进行叠加式 HASH 摘要处理，因此可追溯到任何一笔交易历史。

(4) **集体维护性：**区块链去中心化的特征决定了它的集体维护性。传统中心化机构通常要身兼三职：数据存储者、数据管理者和数据分析者，区块链则以对等的方式由各参与方共同维护，各方权责明确，无需向第三方机构让渡权利，实现共同协作。

1.3 区块链核心关键技术

从技术角度来讲，区块链并不是一个全新的技术，而是集成了多种现有技术进行的组合式创新，涉及到以下几个方面：

（1） **共识机制：**常用的共识机制主要有 PoW、PoS、DPoS、PBFT、PAXOS 等（图 1-3-1）。由于区块链系统中没有一个中心，因此需要有一个预设的规则来指导各方节点在数据处理上达成一致，所有的数据交互都要按照严格的规则和共识进行；

共识算法	POW	POS	DPOS	PBFT	PAXOS
应用场景	公有链	公有链	公有链	联盟链	私有链
错误容忍度	<50%节点数	<50%权益	<50%权益	33%节点数	<50%节点数 (Acceptor)
共识效率	低	中	中	高	高
典型应用	比特币	以太坊	BTS	超级账本	传统分布式产品

图 1-3-1

（2） **密码学技术：**密码学技术是区块链的核心技术之一，目前的区块链应用中采用了很多现代密码学的经典算法，主要包括：哈希算法、对称加密、非对称加密、数字签名等。

● **HASH 摘要算法：**HASH 算法的目的是针对不同输入，产生一个唯一的固定长度的输出。HASH 算法有 3 个特点：一是不同的输入数据产生的输出数据必定不同；二是输入数据的微小变动会导致输出的较大不同；三是给定已知输出数据，无法还原出原始的输入数据。常用的 SHA-256 算法就是针对任意长的数据数列输出 256 位数据，实际使用中 SHA256 用于对区块链的每个区块数据进行 HASH 摘要后防止篡

改，同时结合 Merkle Tree 数据结构实现部分区块数据的 HASH 值验证。

- **对称加密算法 :**对称加密算法利用加密密钥对原始数据进行加密处理，然后将加密后的密文发送给接收者，接收者利用同一密钥及相同算法的逆算法对密文进行解密，才能使其恢复成原始数据。在对称加密算法中，使用的密钥只有一个，发收信双方都使用这个密钥对数据进行加密和解密，这就要求解密方事先必须知道加密密钥。区块链技术中常用的对称加密算法有 AES。

- **非对称加密算法：**非对称加密算法需要两个密钥 :公开密钥(Public Key）和私有密钥（Private Key）。公开密钥与私有密钥是一对，如果用公开密钥对数据进行加密，只有用对应的私有密钥才能解密；如果用私有密钥对数据进行加密，那么只有用对应的公开密钥才能解密。其实现机密信息交换的基本过程是：甲方生成一对密钥并将其中的一把作为公用密钥向其它方公开；得到该公用密钥的乙方使用该密钥对机密信息进行加密后再发送给甲方；甲方再用自己保存的另一把专用密钥对加密后的信息进行解密。

- **数字签名算法：** 区块链技术中使用到的数字签名技术用于验证信息的完整性和真实性，基本流程如下：发送者将需要签名的原始数据进行 HASH 摘要，然后对摘要信息用私钥加密后与原始数据一起传送给接收者。接收者只有用发送者的公钥才能解密被加密的摘要信息，然后用同样 HASH 函数对收到的原文产生一个摘要信息，如果与解密

的摘要信息对比相同则说明收到的信息是完整的，在传输过程中没有被修改，否则说明信息被修改过，因此数字签名能够验证信息的完整性。此外，信息发送者拥有私钥且不公开，因此只有发送者本人才能构造基于其私钥的签名信息，可以确保签名真实性。ECDSA 是区块链技术中常用的数字签名技术。

(3) **分布式存储：**区块链是一种点对点网络上的分布账本，每个参与的节点都将独立完整地存储写入区块数据信息。分布式存储区别于传统中心化存储的优势主要体现在两个方面：

一、每个节点上备份数据信息，避免了由于单点故障导致的数据丢失。

二、 每个节点上的数据都独立存储，有效规避了恶意篡改历史数据。

(4) **智能合约：**智能合约允许在没有第三方的情况下进行可信交易，只要一方达成了协议预先设定的目标，合约将会自动执行交易，这些交易可追踪且不可逆转。具有透明可信、自动执行、强制履约的优点。

1.4 区块链未来发展趋势

面对区块链技术带来的机遇与挑战，全球各行各业都在进行积极布局，试图通过这一“组合式创新”技术改变原有的业务与管理模式，构建一个多方参与、安全信任的新型生态体系。区块链的未来发展趋势主要体现在以下几个方面：

(1) **产业渗透：**虽然区块链的底层架构源于比特币，但作为一种通用技术，区块链正加速从数字货币向其他领域渗透，和各行各业创新融合。目前，金融服务、数字资产、慈善公益等行业纷纷投入到区块链应用的探索中，利用日志存证、信息追溯等特点，改变行业内原有的交易不公开透明等问题。相信在未来，区块链将在更多的领域发挥作用。诸如医疗健康等涉及到大规模数据交互的行业，必将通过区块链技术实现数据的可信交易，破除现有的利益壁垒，打造一个全新的数据行业内外安全共享生态体系；

(2) **多中心化：**区块链的核心并不是"为了去中心化而抛弃中心化管理"，而是构建多方信任机制。在未来，随着跨链技术的不断发展，区块链的架构将演变为多方共同参与的可信任体系。即在多方信息不对称、背景不清晰的情况下，构建多方赖以信任与合作的新生态。未来在多中心化和去中心化之间，将会存在一个中间区域，而不同区块链系统根据特定场景需求，将呈现不同的非中心化程度。

(3) **技术融合：**以云计算、大数据、物联网为代表的新一代信息技术正渗透进各行各业。未来区块链的发展必将以技术融合为切入点，共同解决单一技术的不足与难点，扩大应用场景，降低应用成本。以区块链与物联网结合为例，物联网是互联网在实体经济中的延伸，通过计算机技术实现物品与物品之间的信息交换与通信。区块链系统是典型的点对点网络，具有分布式异构特征，天然适合于在物联网中建立各主体的共识机制，制定交互规则，构建去中心化控制的交易网络。因此，

如何通过区块链与其他技术的融合，实现产业创新，将成为区块链未来发展的重要课题。

（4） **标准规范：**企业应用在未来将是区块链的主战场，联盟链将成为主流方向。与公有链不同，在企业级应用中，人们不仅关注通过软件和算法来构建信任基础，更重要的是如何从用户体验与业务需求出发，构建一套基于共识机制、权限管理、智能合约等多维度的生态规则。面对不断演进的区块链技术，同步考虑相应的技术标准和法律法规，增加区块链的可信程度，建立区块链的应用准则加强监管，防范风险。

2 区块链在健康医疗行业的应用

2016 年，美国国家卫生信息技术协调办公室（ONC）发起了区块链技术应用挑战赛，向全球征集将区块链应用于下一代医疗健康数据 IT 系统的建设方案。奥巴马精准医疗计划向 ONC 资助 5 百万美元用于制定一系列的标准和要求，以保护隐私和跨系统数据交换安全。另外，据 IBM 报告预测,全球 56%的医疗机构将在 2020 年前投资区块链技术。在我国发展健康医疗大数据的背景下，将区块链技术应用于健康医疗产业链中的数据交互，确保个人数据不被滥用，促进实现国家和省级人口健康信息平台的互联互通，形成跨部门健康医疗数据资源共享共用格局。同时，区块链所采用的分布式存储技术可避免单点故障而导致的数据库整体性崩溃，减少以往由于中心化存储所带来的大规模数据泄漏事故，为健康医疗行业带来最高级别的数据安全保障。

2.1 当前健康医疗行业数据问题概述

（1） **数据所有权问题：**个人在接受医疗机构诊疗和接受健康管理服务过程中产生的数据被相关机构收集后，往往会在个人不知情的情况下被用于科研，或被用于与药厂进行联合新药研发，或与保险公司联合变现。大部分机构依靠非完全透明的知情同意获得数据授权，或使用去识别化的脱敏匿名数据进行分析应用，但大数据组合挖掘技术的发展依然可能重新识别出个人信息。2013 年，英国曾启动 Care.data 计划，旨在建立一个全国性医疗健康大数据平台，以促进医疗和研究。但项目

仅仅实施了三年，就被 NHS 叫停，原因在于该项目在共享和使用病人数据时，并未征得病人同意，与保险公司合作数据的商业用途也被质疑。因此，为个人健康医疗数据确权，对数据访问者和使用者进行细粒度权限控制和全流程行为透明可追溯，合理合法地约束个人数据使用至关重要。

（2） **数据安全问题**：现有的数据存储模式大多是中心化的，中心化的健康医疗数据存储模式既面临由于单节点系统的网络、硬件的故障风险和安全漏洞，也可能因个人恶意或误操作导致数据库泄露。鉴于个人健康医疗数据的高度敏感性和高价值性，一旦发生泄漏，后果将难以挽回。2015 年 2 月，美国第二大医疗保险服务商 Anthem 被黑客攻破，近 8000 万员工和客户资料被盗。仅 2015 年一年，在美国遭受到黑客入侵的医疗记录就高达 1.12 亿条。据报告，美国每年因此造成的经济损失超过 62 亿元。而基因数据的全局唯一性和其蕴藏的多维度敏感信息，更加增强了数据安全保护需求。2017 年颁布的《福州市健康医疗大数据资源管理暂行办法》是首个行业数据管理办法，其中着重强调了数据安全和隐私。

（3） **数据共享问题**：在现有健康医疗数据管理模式下，由于信息系统的差异与利益格局的壁垒，不同机构之间难以实现数据的安全共享。这种模式不仅导致了数据的分散，也大大降低了数据的利用效率。以中国为例，三甲医院目前使用的医院管理信息系统（Hospital Management Information System，HIS）多由自主开发，不同 HIS 系统间如要共享数据，往往涉及到数据结构和数据接口标准化、数据安全等问题，导

致数据传输困难、效率低下。因此，如何推动不同机构间数据的安全交互、降低共享成本，最大化地发挥数据的科研与产业价值，将是健康医疗大数据时代亟待解决的关键问题。

2.2 精准医疗 VS 隐私保护

精准医疗强调以个人的组学数据驱动精准的个体化诊疗和健康管理，海量个人组学数据的收集和挖掘将催生不可限量的科研成果和产业发展，形成全新的医疗保健范式。但以基因组数据为基础的生命大数据普及应用也同样带来了巨大的个人隐私保护挑战。

基因数据，特别是全基因组数据，具有高度敏感性，是具有全局唯一标识的生物特征数据，且蕴藏着大量多维度的敏感信息。例如人们可以从基因组数据中推断出关于种族、祖源、药物代谢能力、疾病风险、健康和行为、面部特征等信息，同时还涉及相关亲缘关系个体的遗传数据暴露。因此实现基因数据的完全脱敏匿名化异常困难，美国 NIH 和英国 Wellcome Trust 近年也更新了其数据共享政策，严格限制对单个基因型和聚合基因型频率数据的访问。基因数据的泄露还可能导致基因歧视等社会伦理问题。1993 年的《毕尔巴鄂宣言》首次谴责了与基因或生物学特征相关的社会歧视案例。2008 年美国颁布的《遗传信息反歧视法案》（GINA）禁止雇主或保险公司查询个人基因检测结果，以做出对当事人不利的选择。

全球各国政府积极制定了数据保护条例来保护个人基因数据的隐私。美国的《经济和临床健康信息技术法案》（HITECH）要求数据存储者实施物理、行政与技术解决方案保护生物数据免遭泄漏。欧盟于 2018 年 5 月 25 日起全面实施号称“史上最严数据安全法规”的《通用数据保护条例》（GDPR），旨在对个人数据生产与应用的各环节加以安全规范，最大限度保护个人数据隐私。GDPR 把遗传数据（基因数据）和生物特征数据定义为“敏感个人数据”，将受到最严格的安全监管，该条例强调即使是匿名的基因数据，也只有在数据主体同意的前提下才能被处理。基因数据的永久不可篡改性及亲缘群体暴露风险，应该按照比金融数据更高的安全等级严格保护。

我们需要切实地保护个人隐私，但同时也需要通过大数据的共享挖掘来推动精准医学与整个社会的进步。如何在组学数据应用的合理性与隐私保护的必要性间找到一个平衡点，保证数据共享对科研与社会的贡献超过其应用的风险，需要监管者、科学家、医疗人员和企业共同努力，建立基于伦理规范与技术解决方案的数据共享框架。

2.3 国内外健康医疗行业的区块链应用现状

2.3.1 国外健康医疗行业的区块链应用

2016 年，Linux 基金会主导的超级账本开源项目成立了区块链医疗工作小组。同年爱沙尼亚宣布启动基于区块链的全国医疗健康档案安全共享项目。2017 年，美国卫生与公共事业部（HHS）公布了相关企业和机构提交的 70 多份区块链医

疗应用提案。其中 MIT 的 MedRec 区块链电子病历系统采用基于以太坊的智能合约，为跨机构创建分布式的医疗数据授权系统，以满足患者、社区和科研人员的各方需求。2017 年，Google Deepmind 使用英国 NHS 的病人数据开发监测肾病 APP，被英国数据保护监管机构指控违反数据保护法。紧接着谷歌即宣布其正开发基于区块链的“医疗数据审核系统”，确保将来在把数据用于 AI 应用开发时，数据所有者具有完全的数据知情权和控制权。同年，IBM Watson Health（健康医疗项目）和美国食品药品管理局（FDA）签署了一份合作研究计划，旨在用区块链技术研发一种安全、高效、可扩展的医疗数据交易方式。2013 年以来，欧盟委员会就区块链技术在医疗行业的应用统计已投入了超过 3000 万欧元。可见在发达国家，区块链技术已被视为医疗行业数字化转型的核心元素。此外，海外也出现了应用于医疗健康和基因组领域的公有链项目，如，Nebula Genomics（https://www.nebulagenomics.io），MediBloc（https://medibloc.org/en/），MedicalChain（https://www.medicalchain.com）等。这类公有链旨在实现个人数据的点对点授权交易及各方权益分配，从而形成经济激励的生态圈。

2.2.2 国内健康医疗行业的区块链应用

目前国内的区块链应用主要集中在金融服务、社会公益、供应链协同等领域。在健康医疗领域，各机构主要采取搭建基于本地数据库的私有链或基于“医联体”的联盟链，尚未形成大规模的区块链应用。要构建一个大型的跨机构健康医疗数据共享生态，目前仍面临着需求不明确、前期投入较大、政策法规障碍、技术标准缺失、数据质量较低、安全隐私保护制度不完善、系统性能不足等挑战，政府、医院及企业各方也难以快速推动生产级场景的落地应用。华大区块链希望从自身实践做起，积累核心技术和工程化经验，在实际业务中逐步推动技术落地。

3 华大区块链

本章是白皮书的核心章节，我们将从业务目标、技术架构、优势特色及应用场景四个方面重点介绍华大区块链（图 3-1）

图 3-1

3.1 华大区块链的业务目标

华大区块链的目标是通过融合区块链和密码学等技术，打造具有自主核心技术的组学数据共享基础设施，促进数字化生命的价值流动。在保护个人隐私的同时，最大化数据的应用价值。包括两方面的内容：

（1） 从个人层面而言，实现人人、实时、终身的生命 4D 数据隐私保护，为个人数据确权；通过积分激励将数据价值还于个人，同时促进科研及产业应用；

(2) 从组织层面而言，为行业伙伴提供企业级的区块链基础设施与解决方案，形成组学数据与其他健康医疗大数据的共享交互生态体系，最终实现个人（数据所有者）、机构（科研、医疗等）、政府、企业在生命时代共有、共享、共为的多方协作和互惠共赢体系。

3.1.1 华大区块链技术展望

按照涉及到的技术点与提供的服务方式不同，区块链的技术架构主要分 BaaS、PaaS 和 IaaS 三个层次（图 3-1-1）。

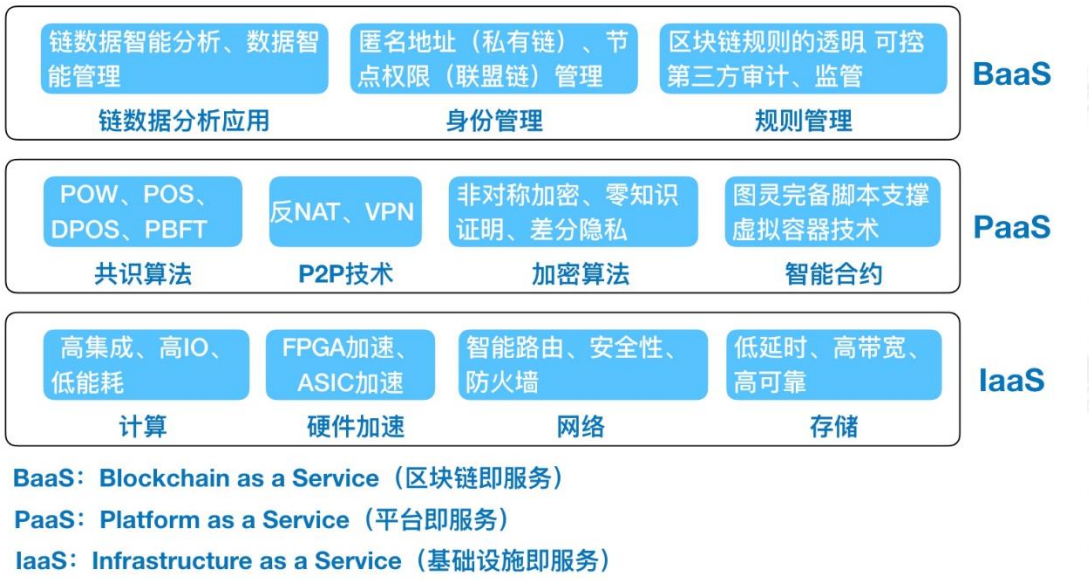

图 3-1-1

在社会资源有限的约束下，传统互联网以信息传递效率优先，中心化架构的 IT 服务大大降低了单用户的信息沟通成本。然而，随着 IT 技术的发展和成本下降，用户的关注点将逐渐从效率转为质量，即关注隐私保护、规则透明平等。由于区块链本身的系统成本高，尤其适合于健康医疗领域这样的高价值行业应用。在华大区块链建设初期，我们以 PaaS 平台为用户提供服务，用户可以自主在平

台上实现数据的安全共享。将来随着云计算、云存储等技术的不断深化成熟，华大区块链将构建自有的开源 BaaS 服务架构，以开源区块链的形式在健康医疗行业形成规模化应用。

区块链技术仍处在早期，目前主要面临以下几点问题：

(1) 多节点共识带来了高信任度，但分布式系统的共识效率性能在企业级的应用有待进一步提升；

(2) 多节点记录存储提高了数据透明度，但对数据安全加密和隐私保护提出更高要求，同时还要保证可用性和监管性；

(3) 智能合约定义了自动执行的逻辑规则，其安全问题有待更加成熟的验证和解决方案；

(4) 区块链技术各成一派，需要跨链技术的进步来增强链与链之间的互操作性。

华大区块链将通过架构优化、技术升级等方式，并和区块链技术生态圈的开发者、标准制定者交流合作，不断提升区块链适用范围与应用价值。同时，华大区块链也将重点关注以下特色技术点的突破：

(1) **基因数字 ID：**通过个人基因 ID 技术，解决现有区块链技术中的数字身份无法安全关联个人实体身份，规避网络应用中的数字权利与现实社会中实体权利难以合法关联的问题；

(2) **后量子加密：**解决现有区块链产品的非对称加密体制无法防止将来量子计算的破解问题；

(3) **安全多方计算：**通过提供安全多方计算解决方案，实现多方数据所有者在不透露数据细节的前提下进行数据协同计算；

(4) **匿名应用：**节点可匿名提供暂时数据给区块链上的第三方应用进行处理，通过瞬时加密机制确保用户隐私；

(5) **乱序存储：**节点可以对数据乱序加扰后分布存储到多个其他节点（包括云平台），数据所有者是恢复原始数据所必须的乱序引索（Index）的唯一拥有者。

3.1.2 华大区块链用来解决什么问题

华大区块链用来解决健康医疗及生命大数据应用的三类矛盾：

(1) 数据应用与隐私保护的矛盾

(2) 数据确权与交互共享的矛盾

(3) 数据安全与加密成本的矛盾

3.1.3 设计原则

华大区块链在设计上遵循以下几个原则：

(1) **2B4D 大数据：**作为全球跨组学数据生产的引领机构，华大创新性地提出了大人群生命组学大数据（2B4D）的概念。2B4D 数据所具有的的人群覆盖度广、数据量大、敏感性高、完整性需求强等特点对区块链的架构与性能提出了极高的要求。因此，从区块链协议、数据结构和功能特性等方面满足 2B4D 大数据的交互共享是华大区块链的第一设计原则。

(2) **自主创新：**华大注重自主创新，目前已在碎片分布式数据存储、基因 ID 等关键领域拥有多项自主知识产权的核心技术与专利，并通过融合差分隐私、非对称数字签名等技术，实现个人生命大数据的可靠存储与安全共享。

(3) **标准化：**华大区块链通过搭建区块链行业应用的标准化体系，实现数据的安全交互与高效共享。包括区块链底层架构标准化与数据共享交互标准化两个层次：**安全高效：**华大区块链在协议设计、架构规划、接口设置、服务部署等方面都遵循这一原则，确保区块链系统运行的可靠与高效。**开放共享：** 华大区块链构建自主可控的 BaaS 区块链平台，发挥运营国家基因库的经验优势，搭建 IT 基础设施，开放区块链服务能力，与行业伙伴共同打造合作共赢生态圈。

3.2 华大区块链技术架构

华大区块链在基于 BaaS、PaaS、IaaS 分层基础上，结合生命健康行业数据应用场景，针对共识算法和密码学算法两个核心技术点进行优化，有效提升系统安全性和业务处理效率。

3.2.1 共识机制

华大区块链使用的共识算法基于 PBFT 基础上做了调整优化，可以称为“PBFT+”共识算法，其核心思想就是针对不同的交易类型实行不同的共识机制。我们把需要上链存储的交易类型分为两大类：事务性交易和非事务性交易。针对不同类型的区块链交易采用不同的共识算法。所谓事务性交易是指需要进行严格

排序和所有节点达成共识的区块链交易，维持同样的状态变更，因此需要使用严格的共识算法确保其执行状态在区块链上保持一致 ;事务性交易之外的其他交易都称为非事务性交易。

针对事务性交易，华大区块链采用 PBFT 算法，具体算法流程图如下所示（图 3-2-1-1）。

其中 C 为发送共识请求的客户端节点，0123 为接受并处理共识请求的服务端节点，3 为故障的服务端节点，共识步骤如下：

(1) **Request：** C 发送请求到任意一个服务端节点，这里是 0；

(2) **Pre-Prepare：**节点 0 收到 C 的请求后广播至节点 123；

(3) **Prepare：**节点 123 收到后记录请求，然后再次广播至所有其他节点，1->023，2->013，3->012；

(4) **Commit：**0123 节点在 Prepare 阶段，若收到超过一定数量的相同请求，则进入 Commit 阶段，广播 Commit 请求；

(5) **Reply：**0123 节点在 Commit 阶段，若收到超过一定数量的相同请求，则对 C 进行反馈。

能够达成共识的计算公式为 $N \geqslant 3F + 1$， N 为参与共识的服务端节点总数，F 为有问题的服务端节点总数。

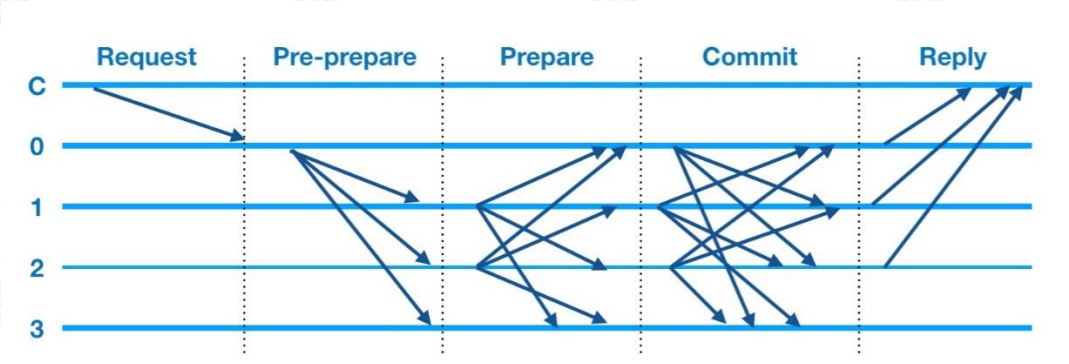

图 3-2-1-1

针对非事务性交易，我们采用简单的排序验证，即发送交易到任一个共识节点上进行交易合法性验证，确认是非事务性交易类型后，通过 Kafka 协议在多个共识节点之间进行统一排序，然后将排序结果广播至链上所有节点（图 3-2-1-2）。

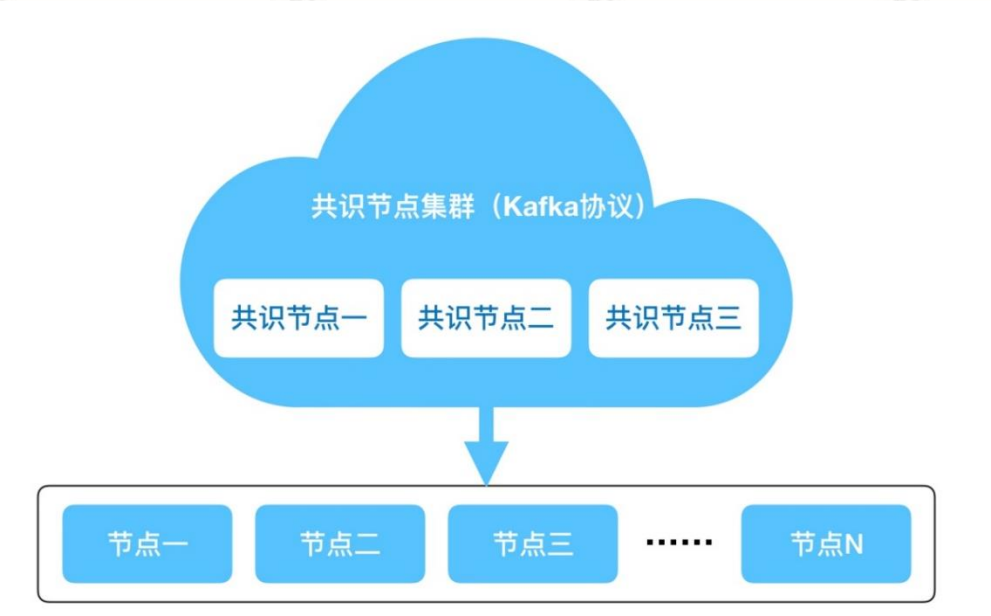

图 3-2-1-2

3.2.2 密码学算法

第一章提到的 HASH 摘要、非对称加密等算法在传统计算机环境中是安全的，在实现 128 位安全级别的情况下可以满足当前区块链应用的加解密速度和存储空间要求。但是考虑到当前量子计算技术的高速发展，目前已知的量子计算 Shor 算法已经能够在较短时间内破解一些常用的传统加解密算法。

传统计算机环境下，常用的大数分解的破解算法包括线性筛选和数域筛选两类，分解模 n 和安全级别位数 b 的关系大致如下：n 的位数约为$\frac{1}{3}b^3 \Big/ (lgb)^2$。在量子计算机环境下，目前已知的 Shor 算法能够实现在$(lgn)^{(1+o(1))}$个量子比特位计算机下破解模 n 的基本运算次数为$(lgn)^{(2+o(1))}$。这意味着如果需要实现同样的安全等级 b，需要的模 n 位数约为$2^{(0.5+o(1))b}$。

由上述模 n 位数和安全等级 b 的关系可知，量子计算环境下，传统的 RSA 加密体制在现有的模 n 位数下已经无法保证安全性，如果要在量子计算环境下实现和传统计算机环境下同等安全性，则需要大为增加 n 值，这意味着公私钥的长度和加解密速度都急剧增长，这就导致加解密算法失去了应用价值。实际上，基于量子计算环境的 Shor 算法，传统的 RSA、ECDSA 算法都不再安全。

另一方面，后量子密码学时代，已经有许多研究证明了量子计算技术的发展并不意味着所有密码学算法都不再安全，目前已知的抗量子算法有基于 HASH 函数的密码算法、基于纠错码的密码算法、多变量二次方程组的密码算法和基于格理论的密码算法。我们在区块链应用中综合考虑算法稳定性、加解密性能和未来兼容性，选择了格密码算法来进行数字签名，确保区块链交易安全性。

同时，华大链支持中国官方制定的 SM2、SM3 算法，SM2 是国家密码管理局于 2010 年 12 月 17 日发布的椭圆曲线公钥密码算法，SM3 密码摘要算法是国家密码管理局 2010 年公布的中国商用密码杂凑算法标准。SM3 算法适用于商用密码应用中的数字签名和验证，是在 SHA-256 基础上改进实现的一种算法。华大链基于 SM2\SM3 密码算法的全面兼容性可以更好的满足国内不同行业的信息安全准入要求。

3.2.3 基因数字 ID

在公有区块链应用中，用户所使用的线上数字身份一般都没有和实体身份对应起来，这样纯粹的虚拟数字身份不具有生物实体的可追溯性，导致线上数字身份的权利义务无法和现实世界中的个人权利义务对应起来，一个典型的问题是比特币用户因为丢失地址私钥后没有任何挽救途径可以找回线上资产。因此，我们认为区块链的应用中需要提供能够同时对应线上数字身份和线下实体身份的技术解决方案。

华大利用自身在基因测序方面的优势，开发出一套基于个人基因序列多态性的新型加密方式。基因序列对应个人的唯一身份，具有最高特异性。同时，基因序列可以产生大量的公私钥对，可实现一次一密的动态加密机制，充分保证了身份认证安全。目前，我们撰写的国家级专利《一种基于个人全基因组数据的数字身份生成方法》正在受理。未来，基因数字 ID 技术将作为华大区块链的核心组件，确保每一次数据交互都安全可靠。

目前线下实体身份通常采用个人脸部识别技术作为身份验证方式（银行系统基于身份证的身份鉴别），但同时基于脸部特征识别技术也存在较多问题，例如

脸部生物特征唯一性无法保证 100%不重复（双胞胎撞脸等）、可复制性较高（整容易容）、人工或机器的脸部识别准确率都无法达到 100%。

我们结合基因技术为区块链的身份鉴别提供高可靠性、高精度的基因 ID 数字身份解决方案(图 3-2-3-1)。在个人基因组中，存在多个短串联重复序列(short tandem repeat, STR），STR 是核心序列为 2-6 个碱基的短串联重复结构，其中每个 STR 中重复序列的重复次数范围在 2-100 之间，对于任何特定个体的全基因组数据，染色体上某个特定位置的 STR 中重复序列的重复次数是固定的，但在不同的个体在相同 STR 的重复序列的重复次数可能不同，这就构成了人群中这些 STR 重复序列的多态性。由于人类基因组中 STR 非常多，通过对这种多态性的检测，就可以明确区分个体与个体的不同。

图 3-2-3-1

个人基因数字 ID 库的应用，可以实现区块链上的每次交易使用不同 ID，这样可以有效保护用户的交易隐私，同时，结合第三方应用，可以对 ID 采取合法性验证，提供匿名数据服务，应用框架如下图所示（图 3-2-3-2）。

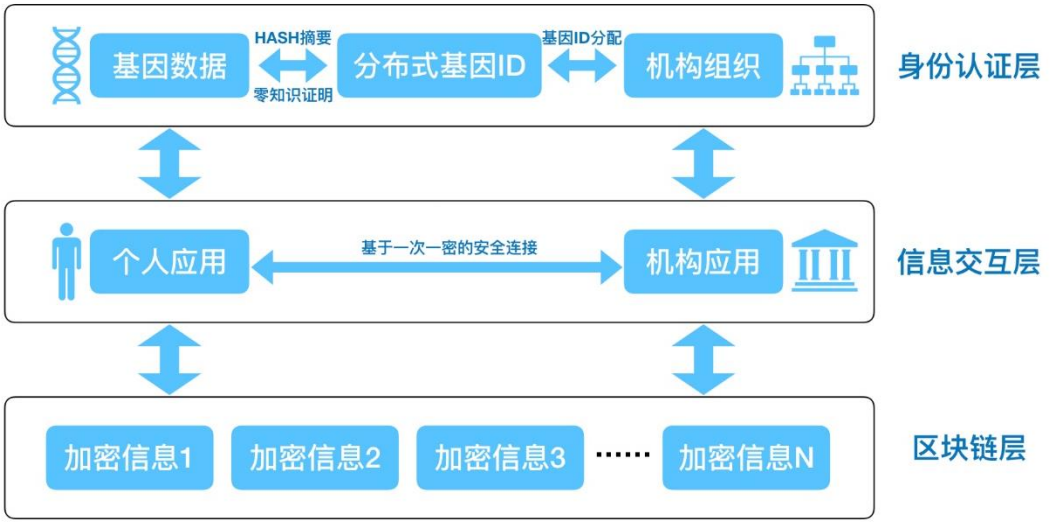

图 3-2-3-2

3.2.4 碎片分布式存储

传统的云存储模式中，用户把完整的数据信息存放到云端数据中心，这种中心化的云存储模式相对高效和低成本而广为流行，但在安全性和隐私泄漏等方面存在风险。我们认为，基于区块链的隐私数据保护需要采取一种全新的分布式云存储方案，这里称为碎片分布式存储，基本思想是将用户的某一完整的隐私数据进行分片操作，然后将不同碎片进行加密扰序后，存储到不同的网络节点上。用户本地保留恢复原始数据所必须的碎片重组索引文件。

碎片化分布式存储可以把数据分布到多个网络节点，各网络节点基于区块链智能合约来提供数据存储服务，在合约有效期内需要定期证明它们能继续提供存储服务的能力。用户需要访问数据或者授权他人获取数据时，需要将访问凭证消息进行数字签名后上链保存，对应网络节点获取到该授权凭证后才提供数据访问

服务。这些凭证在区块链上是公开、透明、可审计的，网络节点自动保证存储合约的一致性。

这种去中心化的碎片化分布式存储方案和区块链技术的结合，可以有效保护用户的数据隐私。数据被分割成小块，经过加密扰序后才会分散存储在众多节点上，能够避免中心化存储的集中式风险，即使某一块数据被泄露，也只是部分而非全部数据。另外，每个数据分片都有多个备份节点，一旦出现某个存储空间提供者长期离线的情况，客户会自动将切片备份到新的提供者中，避免了中心化存储因网络或者物理等原因导致数据丢失的风险。结合上一节提到的基因 ID 技术，我们示意了一种基于个人基因组数据的碎片分布式存储方式，将有效保证个人身份的最大特异性与数据的最高安全性（图 3-2-4）。

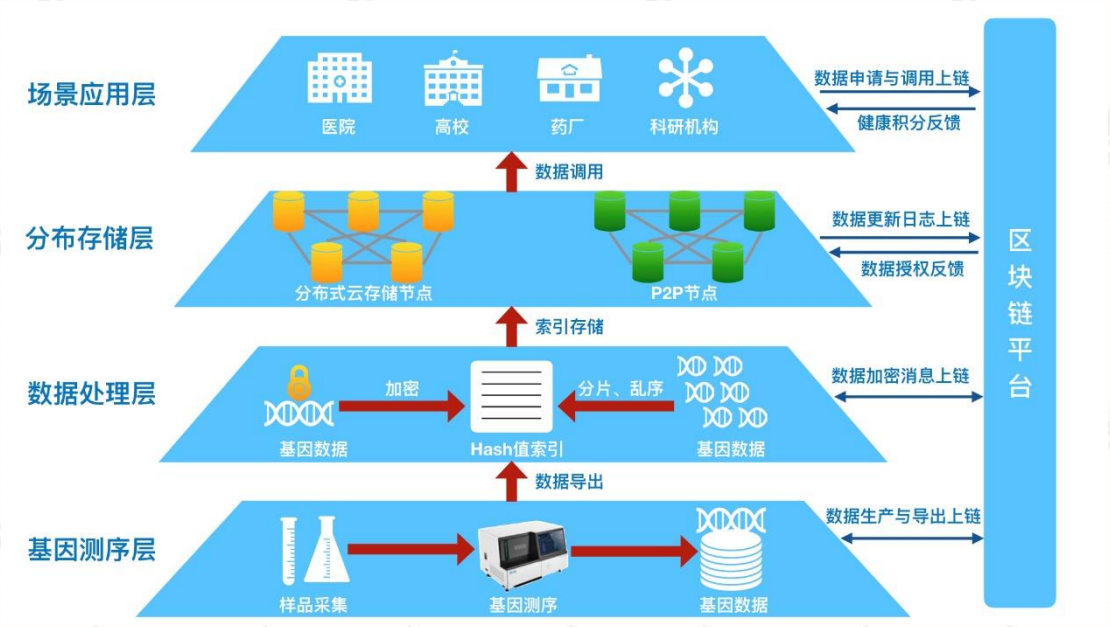

图 3-2-4

3.2.5 安全多方计算

数据流通安全一直是大数据时代难以解决的难题。如何在保护本地数据隐私安全的情况下，促进不同地区、不同机构间的数据共享与协同计算，正引起研究者们的广泛关注。

安全多方计算（SMC）最初由图灵奖得主姚期智院士在 1982 年针对“百万富翁问题”提出，是一种在无可信第三方的条件下，多方之间在互不公开数据的前提下实现协同计算的技术。两方计算框架主要实现原理是基于混淆电路（GC, Garbled Circuit）和不经意传输协议（OT，Oblivious Transfer）的密码学技术，将计算逻辑转化为布尔电路，并加密传输电路及标签数据，最后各方解密获取计算结果。30 年来，也陆续出现了基于秘密分享协议的多方安全计算框架，如 GMW、SPDZ。目前，安全多方计算技术的进步已使其在金融、征信等领域展开应用，但大规模普及商用仍面临特定计算场景的性能瓶颈和可扩展性等问题。

华大通过搭建安全多方计算平台，允许拥有基因数据的各机构在不泄漏原始数据的情况下完成协同计算，将极大促进跨机构开展大人群大队列的组学数据联合研究，也可用于大型基因数据库的安全查询。同时，我们将与区块链技术进行结合，制定一个标准化的组学数据共享协议，将每一次计算与交互的日志进行区块链存证，确保计算过程的公开透明，实现多方共赢。

安全多方计算和区块链的应用框架如下图所示（图 3-2-5），用户数据存储在不同网络节点中，发起计算流程后，通过区块链凭证实现数据的授权。各节点在区块链上接收到授权凭证并确认有效后，由安全多方计算节点进行联合计算。

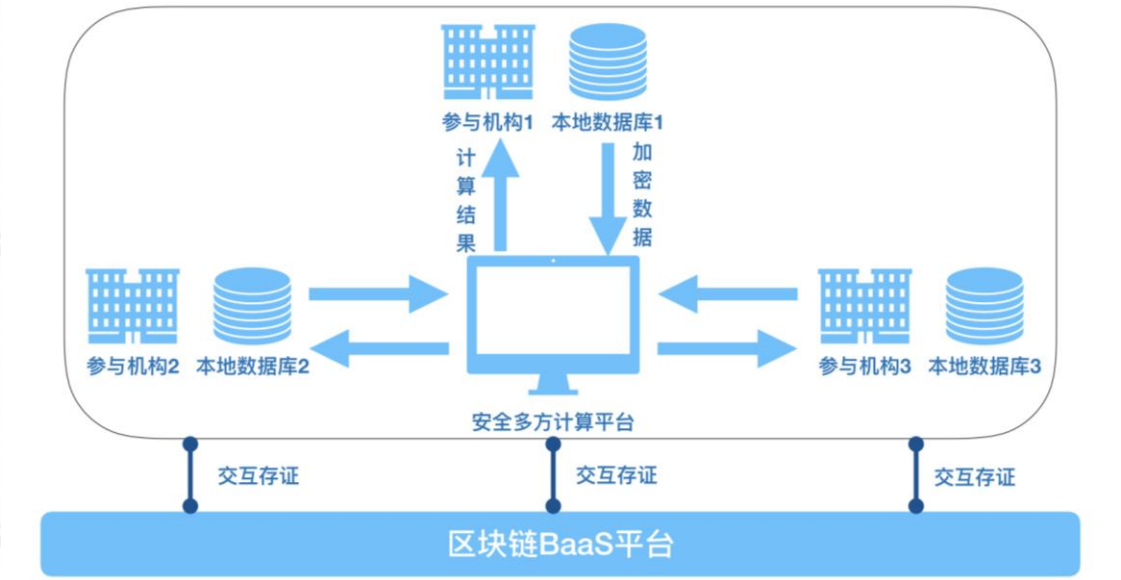

图 3-2-5

3.2.6 BaaS 接口

为了更好的支持上层业务对区块链模块的集成应用，我们在华大区块链设计之初就明确以 BaaS 为建设目标，通过丰富的 API 接口为上层业务及产品提供灵活方便的功能集成、运维部署服务(图 3-2-6)。提供的 BaaS 服务接口包括：

(1) **节点权限认证控制**：基于联盟链的应用场景，各个节点的链上权限及角色配置都可以通过运维软件的 WEB 界面进行配置。每个节点的模块 ID、IP 地址、角色信息（Order 节点、Peer 节点、Endorse 节点等）、组织信息及链通道信息都严格对应，防止越权访问。

(2) **共识算法可插拔**： 默认情况下支持“PBFT+”共识算法，同时支持通过 API 接口调用包括 PBFT、PAXOS 等其他共识算法，实现共识算法的灵活配置。目前华大区块链已实现单节点的平均共识性能在 2000TPS 以上，单节点的平均交易性能在 500TPS 以上。

（3） **加密算法自定义配置**：可以通过 API 接口配置选择不同的数字签名算法（ECDSA\SM2\后量子签名算法）、HASH 摘要算法（SHA256\SHA384\SM3）。

（4） **一键式运维部署**： 我们提供基于 WEB 界面的运维软件实现区块链部署的节点配置，包括节点 IP、节点 ID、节点数据库配置、共识算法及加密算法配置，运维软件根据配置信息自动生成部署脚本，然后执行脚本即可启动区块链服务； 此外，运维软件还提供区块链节点信息查询、区块信息查询功能。

（5） **智能合约动态生成**：基于已开展的应用场景，我们提供在线生成智能合约功能。通过运维软件的 WEB 界面输入交易条件和参数（例如交易价格、交易对象、生效时间、例外条件）后，即可自动生成基于 Go 语言的智能合约代码。

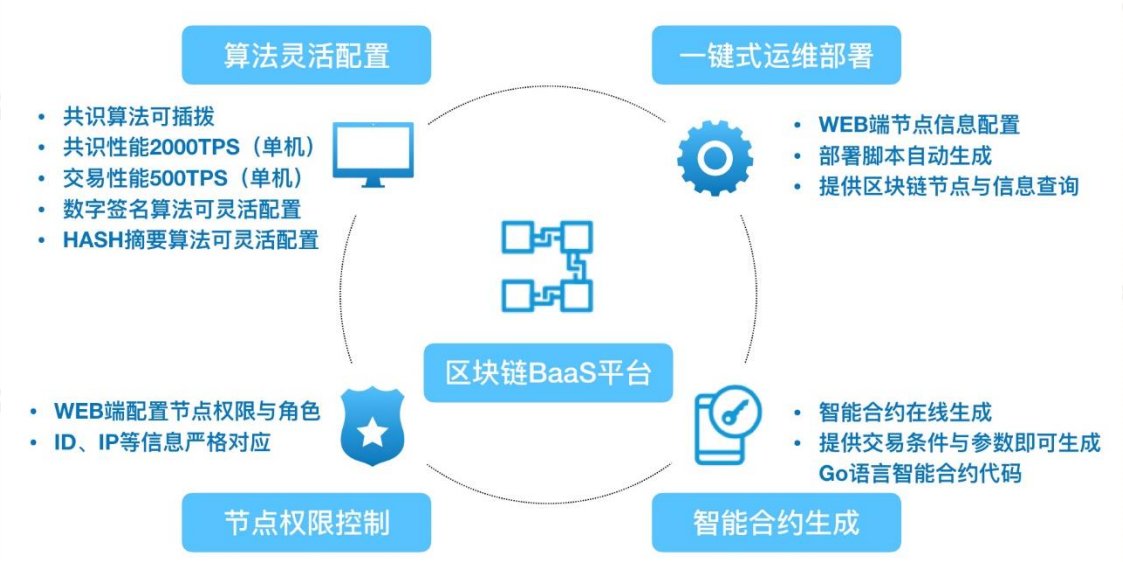

图 3-2-6

3.3 华大区块链优势特色

华大区块链在开发与应用过程中，始终从隐私保护、安全共享、价值交互三方面关注区块链系统架构的搭建与优化，适用于行业内众多应用场景，满足企业级需求，提供安全高效的开放服务平台。

3.3.1 隐私保护

华大区块链参照 GDPR 和国标 GBT 35273-2017 对个人数据的规定，为数据主体确权，实现个人数据的细颗粒度授权控制，确保所有数据交换都如实、不可篡改地记录在区块链上，防止隐私泄露。

(1) 支持对数据信息进行多重加密签名后的链上存储；

(2) 支持凭证撤回功能；

(3) 支持碎片分布式存储与签名验证。

3.3.2 安全共享

华大区块链旨在构建组学数据的共享生态体系，确保数据共享全程可监管，并通过制定数据交互的协议规则，促进数据流通，建立行业标准。

(1) 支持安全多方计算平台；

(2) 支持在云端进行密文分析，实现数据零接触应用；

(3) 支持基于基因数字 ID 的独特生物特征身份认证系统。

3.3.3 价值交互

华大区块链的终极目标是通过个人组学数据的价值交互促进“健康长寿”的人类终极追求实现。基于华大区块链底层架构开发的健康积分激励系统，个人不仅可以授权数据使用挣取积分报酬，也可通过达成健康提升目标（如每日步数达到 10000、体脂率降低至 15%、主动上传每日健康饮食情况等）挣取。所有交互日志都登记在区块链上，积分可用于消费健康促进服务（精准运动、精准营养，健康咨询）等。

（1） 兼容不同机构的数据库与 APP 端接口；

（2） 从数据生产到应用的全贯穿，促进数据价值流动；

（3） 支持积分任务、社群交互等功能，提升用户体验。

3.4 华大区块链应用场景

全球范围内，区块链正加速从数字货币领域演进为与各类实体经济应用场景进行创新融合。自 2018 年开始布局区块链以来，华大积极探索基于“区块链+”的应用场景构建，目前已上线和正在部署包括个人数据确权与价值交互、罕见病公益、HPV 保障计划、区块链与深度学习结合、农产品防伪溯源等方面的应用。未来我们将结合技术进步，不断丰富和完善应用场景。

3.4.1 区块链+跨组学数据：个人生命数据的价值流动

华大从 2015 年推出了国产自主可控的测序仪，核心工具的突破，使得高效低成本的基因技术应用于大人群的疾病防控成为可能。从政府主导的民生实事切

入，将基因技术惠及民生的同时，汇集宝贵的大人群大样本资源，必将形成“大数据驱动”的引领性大科学突破，催生新型的大健康产业。华大已在深圳、长沙、阜阳等城市开展全市范围的孕妇无创产前检测等民生项目，目前华大累计完成超过三百万例的孕妇基因检测。将来将扩展到更多省市，同时也将升级测序数据量，增加疾病防控种类。面对如此海量的基因数据，如何在确保数据安全和隐私保护的前提下进行大数据的挖掘利用，为大人群提供全方位全周期的服务，是华大面临的核心要务。

此外，华大已经用测序、质谱和影像等技术方法对数千名员工进行了连续三年的贯穿组学监测、研究，初步证明了跨组学“4D 大数据”模式是解读、监测健康与疾病状态的最佳途径。华大区块链从自身实践做起，为员工搭建了一套基于区块链的跨组学数据安全共享系统，支撑个人生命数据的价值实现。这是华大区块链的首个应用场景，后续也将推广、升级到华大对外与政府、医院在跨组学数据的科研、临床和产业应用等方面的合作。

以下介绍华大区块链在跨组学数据的应用场景（图 3-4-1）：

（1） 通过生命组学工具（测序仪、质谱仪、影像设备、可穿戴设备等）收集全方位全周期的生命大数据，形成人人、实时、终身的生命健康档案，形成数字化生命；

（2） 所有数据将加密处理，于国家基因库进行统一存储，确保数据硬件安全、物理安全和访问安全；

（3） 用户可通过前端 APP 授权个人数据被内部科研团队、合作医院、健康管理团队等使用，所有的使用日志将以区块链形式记录，用户可实时

查询、授权个人数据使用情况，实现用户对个人数据的控制权，将数据价值还于个人；

(4) 对于授权数据使用或主动提供组学数据的用户，华大以健康积分作为激励；健康积分可用于各类健康促进服务（精准运动、精准营养等），实现个人生命数据价值的正向反馈。

基于区块链技术的数据交互模式现已应用于对内的跨组学数据员工健康计划以及对外的肿瘤关爱计划，已实现近万人的组学数据安全共享与价值实现。随着华大业务覆盖更多人群，该模式将重构个人组学数据的生产交互模式，真正实现“我的数据我掌控、我的健康我做主”。

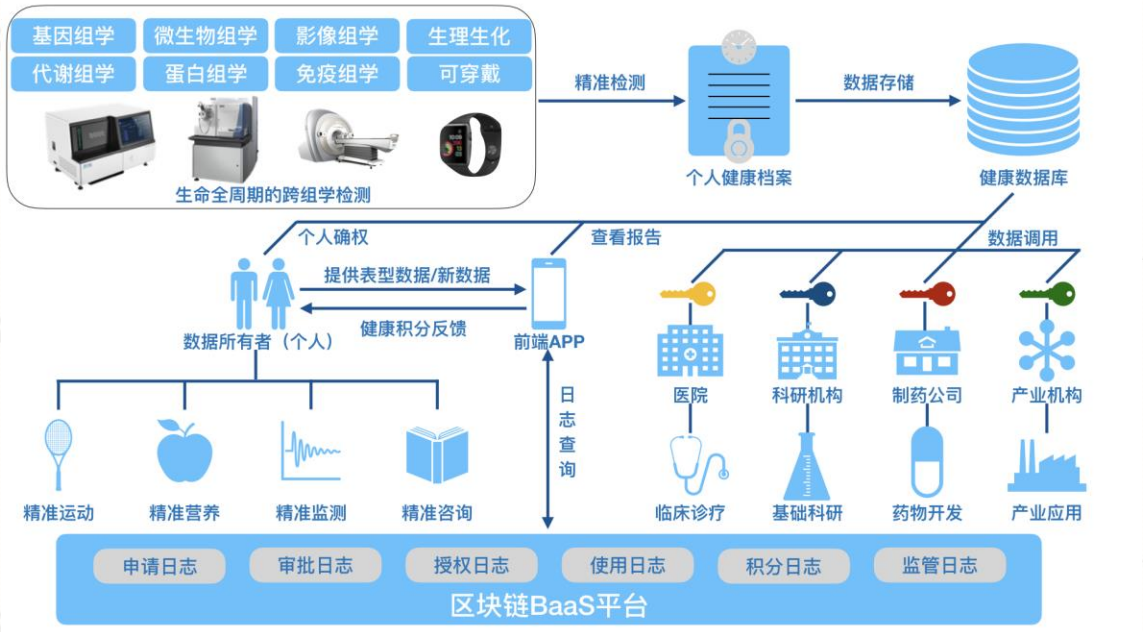

图 3-4-1

3.4.2 区块链+注册申报：医疗器械申报全流程管理

华大智造是华大基因集团旗下定位于生产国产自主可控的生命数字化工具生产商。临床注册审批是国家食品药品监督管理总局（CFDA）准予医疗器械与药品入市许可的必要步骤。在传统申报流程下，由于需要提交的材料较多，易出现报送信息不全、不准、不真。此外，由于审批、检查流程环节众多，信息泄漏的风险点不少，易导致商业核心机密泄漏。华大每年需要报送 CFDA 审批的国产仪器与试剂盒众多，为加强华大内部设备试剂生产、临床试验数据、注册申请材料等全流程追踪管理，实现全程可溯源、信息可追踪、过程可监管，我们正在搭建基于华大区块链的临床注册申报管理平台（图 3-4-2）。未来所有注册申报的数据及信息均将通过此平台登记，防止出现报送信息不全、失真等问题，并减少数据泄漏风险。同时，华大将与 CFDA 探讨建立合作关系，以区块链模式及时共享数据，联合实施审计监管，从而提高申报效率。

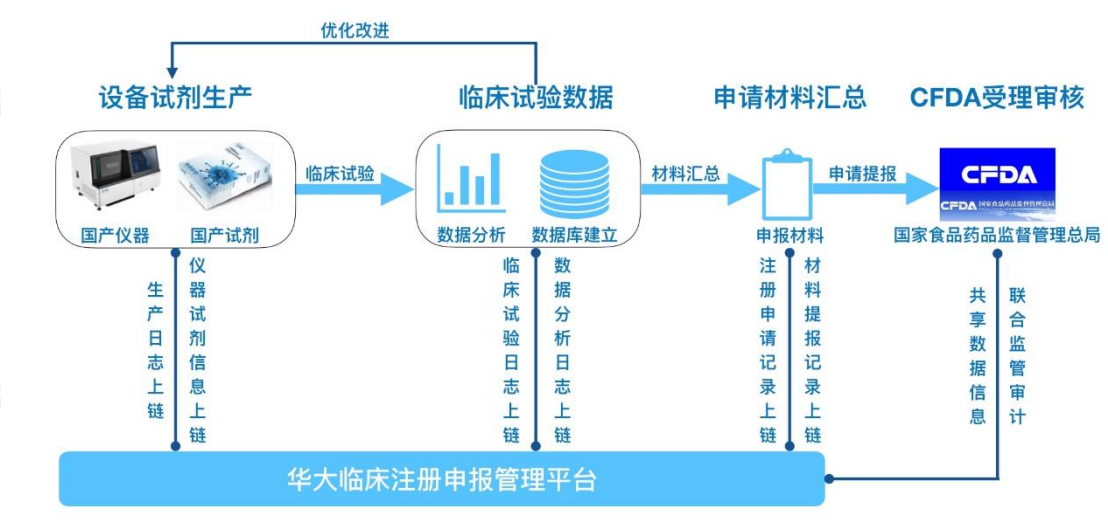

图 3-4-2

3.4.3 区块链+罕见病公益：许一个没有罕见病的未来

罕见病是指发病率极低的疾病，根据世界卫生组织（WHO）的定义，罕见病为患病人数占总人口的 0.65‰～1‰的疾病。华大基因于 2016 年成立了“华大基因罕见病公益基金”，为罕见病患者提供免费基因检测及遗传咨询，帮助罕见病患者查找病因并促进相关科学研究。2017 年又与合作伙伴联合成立了为全球重症地贫血患儿永久免费进行 HLA 配型的“华基金”及为全球 14 岁以下的莱伯氏先天性黑蒙症患者提供基因检测的“光基金”。华大区块链为罕见病相关的公益基金建立基于区块链的管理平台，确保资金用途、捐赠记录、受捐人信息等都将通过区块链进行存证，实现全程公开透明（图 3-4-3）。同时由于罕见病例稀少，相关数据极其分散，对罕见病的诊断、医疗离不开包括患者、病友会、医生、检测机构、科研机构、制药机构、公益基金会、媒体等多方互助才能把资源最优化，最大可能地应对罕见病。华大区块链也将探索通过分布式架构促进罕见病相关方多方协作，同时确保数据隐私、促进数据共享、定义数据价值。

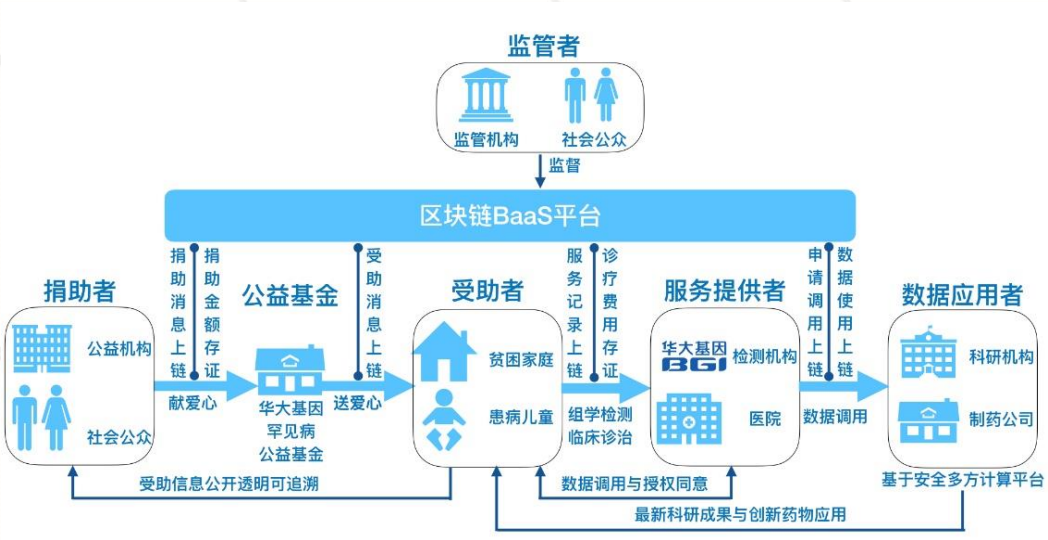

图 3-4-3

3.4.4 区块链+互助保险：HPV 检测保障计划

宫颈癌是女性最常见的恶性肿瘤之一。近年来，我国宫颈癌发病呈现年轻化的趋势，每年因宫颈癌死亡的女性约 3 万人。早在上个世纪的 90 年代，科学家们就已经发现宫颈癌与 HPV（人乳头瘤病毒）的感染密切相关，宫颈癌是目前唯一病因明确、可早发现、早预防的恶性肿瘤。为应对 HPV 感染现状，降低宫颈癌发病率，呵护女性健康，华大基因近期研发出自取样的 HPV 分型基因筛查检测产品，启动大规模的互联网宫颈癌防控计划，同时首期在员工内部试点 HPV 互助保障计划，将基因科技与互联网保险结合，所有购买 HPV 检测试剂盒的费用都将作为互助基金，用于日后互助理赔。购买记录、检测结果、理赔金额等都将通过区块链进行不可篡改地存证，并接受相关机构的监管，保障用户权益。（图 3-4-4）

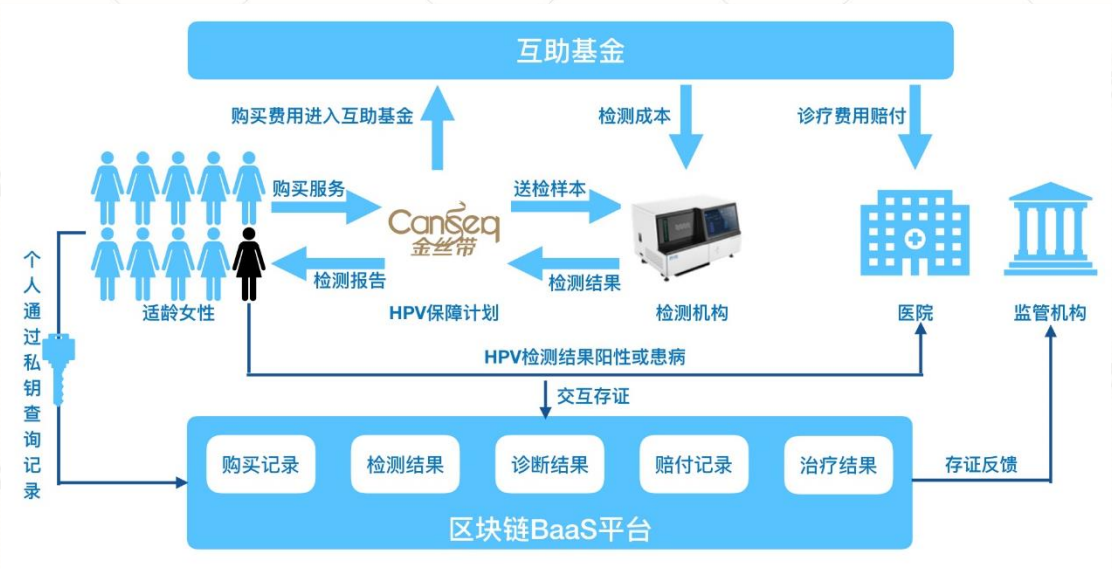

图 3-4-4

3.4.5 区块链+深度学习：从技术融合到生物智能

随着基因数据的爆发性增长，深度学习等算法在基因组分析中的应用逐渐增多。将区块链技术与深度学习等智能算法融合，预先明确算力提供者、算法提供者和数据提供者三方的权责并做好利益分配，才能有效促进基因大数据的挖掘。在区块链上进行待训练数据的身份与权属认证，并通过智能合约发布训练需求，激励算法提供者贡献智慧。算法提供者既可以在本地可信环境中训练模型，还可通过区块链接入第三方算力平台。智能模型训练完成后，其科研与产业应用价值可通过预先定义好的规则回馈给各方（图 3-4-5）。华大区块链创新性将区块链技术用于匹配数据供需方，为数据挖掘引入广泛的市场参与者，从而形成一个多方协作的算法市场与智能计算系统，既可为数据确权，又可最大化发挥数据价值，为最终实现生物智能奠定基础。

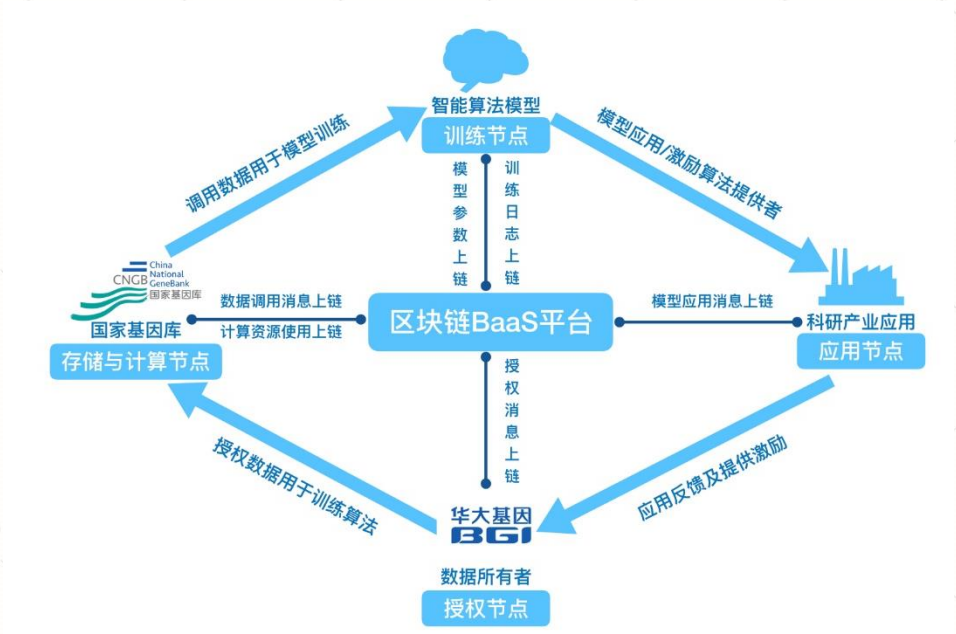

图 3-4-5

3.4.6 区块链+供应链：农产品智慧防伪溯源平台

2017 年，国务院发布《关于积极推进供应链创新与应用的指导意见》，提到建立基于供应链的重要食品质量安全追溯机制。由于我国食品供应链的成熟度较低，物流等基础设施仍薄弱，传统的管理方法和溯源平台无法在短时间内满足民众对食品安全和品质溯源的诉求。

为了解决这一问题，华大区块链携手华大农业，整合物联网和基因检测技术，打造农产品的智慧防伪溯源平台。利用区块链去中心化、数据不可篡改、公开透明、时间戳等特点，将农场、农户、检验检疫、加工贸易、销售、物流仓储等机构加入到联盟链上，形成一个资金流、信息流、产品流的共享链条，做到来源可查，去向可追，责任可究（图 3-4-6）。各个环节参与方以全节点形式参与到流程中，所有数据产生与交互都加密签名后上链存证，充分解决供应链中由于信息不对称导致的交易摩擦，监控管理缺失、数据欺诈导致的质量安全或假冒问题，为消费者提供透明可追溯的全流程信息，形成全新的农业生产管理方式。

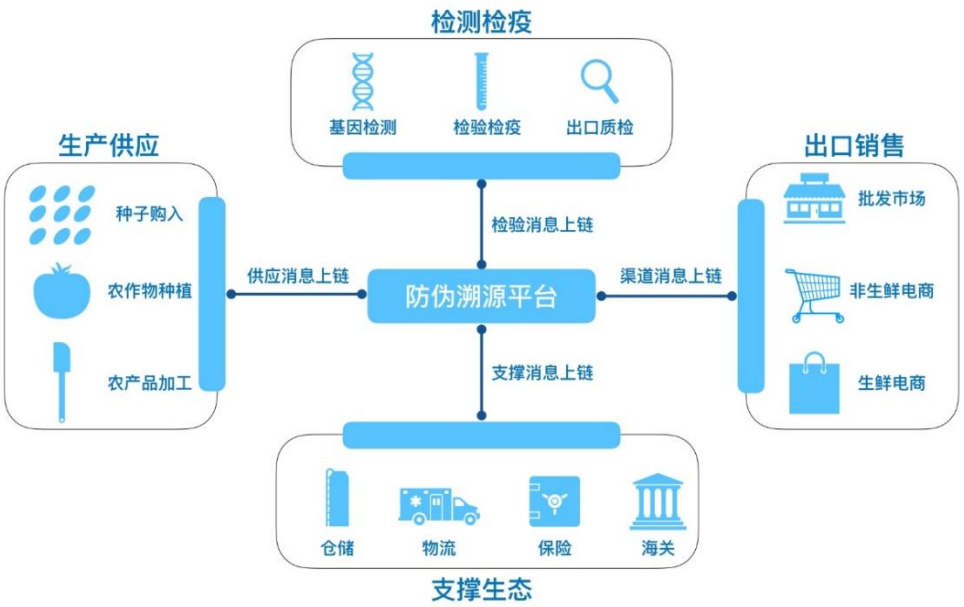

图 3-4-6

4 拥抱生命大数据时代

区块链技术的快速演进，并与实体经济不断融合创新，正在推动可信社会的建立，促进数据价值的流转。人类基因组计划完成已近 20 年，随着组学工具的不断成熟，海量的生命大数据正在不断地被数字化和分析挖掘，也为生命时代的价值度量和交换提供了数据基础。作为全球最大的基因组研发机构，华大基因已与全球 60 多个国家、4000 多家机构、13000 多名合作伙伴建立了广泛的合作关系，共同向着“生优病少、健康长寿、温饱不愁、环境友好“的大目标迈进。华大区块链也紧密围绕集团战略，推动生命大数据共享交换的生态体系建设。

与任何新技术刚兴起时一样，区块链在生命健康行业的落地应用依赖于技术进步、行业认同、政策法规等支持。为推动区块链应用发展，现建议如下：

1. 推动区块链与其他技术的融合

区块链技术不是万能的，也不是孤立存在的。区块链技术广泛应用于生命大数据领域，需要与云计算、数据库、信息安全、密码学等技术在硬件和软件层面上不断融合创新。

2. 依托联盟，行业内广泛合作

以基因数据为基础的生命组学大数据，全球目前仍缺乏统一的共享标准和隐私保护准则。依托国家基因库合作与联盟体系的建设，积极推广区块链技术的应用，在实践中积累经验，形成示范效应。

3. 推进区块链技术标准制定，并与产业标准并进

加快推进区块链关键技术标准的制定，参与国际标准研制工作，并积极与生命健康行业的相关应用标准互动，推动产业应用标准落地。

4. 出台扶持区块链应用发展的政策

借鉴发达国家和地区的先进做法，区块链的规模化应用得益于良好的发展环境与行业政策支持。及时出台相关扶持政策，重点支持平台建设与技术攻关，建立行业监管标准，促进技术长远发展。

数化万物，智在融合。生命因多样而美丽，系统因协作而强大。华大以“造福自己、造福人类”的大目标为指引，用开放的心态拥抱未来，希望与合作伙伴共同探索出一条区块链在生命健康行业的落地应用之路，携手打造生命大数据安全共享和价值交互的全新生态。

术语与缩略语

术语	定义/解释
跨组学	指的是覆盖了从 DNA 到生物体的所有相关生物结构与功能的集合，主要包括基因组学、转录组学、蛋白组学、代谢组学、免疫组学、影像组学等。
区块链	分布式数据存储、共识机制、加密算法等计算机技术的新型应用模式。
去中心化	与“中心化”相对的新型网络内容生产过程，网络各节点高度自治，自由连接，共同创造内容，生产信息。
分布式	与“集中式”相对的数据信息存储模式，是一种不依赖于中心服务器（集群）、利用分布的计算机资源进行计算存储的模式。
哈希	Hash，也称“散列”，一种用于文件压缩映射的计算机算法。
共识机制	区块链系统中实现不同节点间建立信任、获取权益、并就某个提案达成一致的数学算法过程。
非对称加密	一种由公钥和私钥结合的加密算法。
数字签名	基于非对称加密技术实现的电子签名与验证方法。
智能合约	一段部署在区块链上可自动执行的计算机程序，用于交付各参与方预设的承诺。
差分隐私	数学上对通过算法消除个人隐私的统称。
同态加密	一种新型的密码学技术，可以实现对加密数据进行处理后的解密结果与未加密的原始数据进行处理后的结果保持一致。
安全多方计算	在没有第三方介入的情况下，通过计算机协议的方式解决一组互不信任的参与方之间保护隐私的协同计算问题。

基因 ID	利用个人唯一的差异化基因位点序列生成的个人身份认证系统。
健康积分	一种在华大区块链环境下用于数据价值流动的等价交换物，可用于区块链各节点的价值交易

缩略语	原始术语
BT	Biotechnology，即生物科技。
2B4D	大人群生命组学大数据（2B=Big data/Big population 4D 生命数据）
PoW	Proof of Work，即工作量证明
PoS	Proof of Stake，即权益证明
DPoS	Delegate Proof of Stake，即股份授权证明
PBFT	Practical Byzantine Fault Tolerance，即实用拜占庭容错
PAXOS	分布式一致性算法
BaaS	Blockchain as a Service，即区块链即服务
PaaS	Platform as a Service，即平台即服务
IaaS	Infrastructure as a Service，即基础设施即服务
GID	Gene Identification，即基因 ID
BI	Bio-Intelligence，即生物智能

参考文献

1. 《“健康中国 2030”规划纲要》，中共中央、国务院
2. 《国务院办公厅关于促进和规范健康医疗大数据应用发展的指导意见》，国务院办公厅
3. 《中国区块链技术和应用发展白皮书（2016）》，工业和信息部信息化和软件服务业司
4. 《信息安全技术 个人信息安全规范》，中国国家标准化管理委员会
5. EU Commission. 《General Data Protection Regulation》. April 2016
6. National Human Genome Research Institute, United States. 《Genetic Information Nondiscrimination Act of 2008》.May 2008
7. Department of Health and Human Services, United States. 《The Health Information Technology for Economic and Clinical Health Act》, January 2009
8. Nakamoto S. Bitcoin: A peer-to-peer electronic cash system[J]. 2008.
9. Manyika J, Chui M, Brown B, et al. Big data: The next frontier for innovation, competition, and productivity[J]. 2011.
10. Gantz J, Reinsel D. The digital universe in 2020: Big data, bigger digital shadows, and biggest growth in the far east[J]. IDC iView: IDC Analyze the future, 2012, 2007(2012): 1-16.
11. Azencott C A. Machine learning and genomics: precision medicine vs. patient privacy[J]. arXiv preprint arXiv:1802.10568, 2018.
12. Mittos A, Malin B, De Cristofaro E. Systematizing Genomic Privacy Research--A Critical Analysis[J]. arXiv preprint arXiv:1712.02193, 2017.
13. Wang S, Jiang X, Tang H, et al. A community effort to protect genomic data sharing, collaboration and outsourcing[J]. NPJ genomic medicine, 2017, 2(1): 33.
14. Huang Z, Ayday E, Lin H, et al. A privacy-preserving solution for compressed storage and selective retrieval of genomic data[J]. Genome research, 2016, 26(12): 1687-1696.
15. Sousa J S, Lefebvre C, Huang Z, et al. Efficient and secure outsourcing of genomic data storage[J]. BMC medical genomics, 2017, 10(2): 46.
16. Zhang Y, Blanton M, Almashaqbeh G. Secure distributed genome analysis for GWAS and sequence comparison computation[C]//BMC medical informatics and decision making. BioMed Central, 2015, 15(5): S4.

17. Guthrie S, Connelly A, Amstutz P, et al. Tiling the genome into consistently named subsequences enables precision medicine and machine learning with millions of complex individual data-sets[J]. PeerJ Preprints, 2015.
18. Ben-Sasson E, Bentov I, Horesh Y, et al. Scalable, transparent, and post-quantum secure computational integrity[J]. Manuscript.(2017). Slides at https://people. eecs. berkeley. edu/~ alexch/docs/pcpip_bensasson. pdf, 2017.