ICS 07.080
Q 8499

团　　体　　标　　准

T/SZAS 9—2019

基因数据流通区块链存证应用指南

Application guidelines of blockchain certifications for genomic data flow

2019－12－07 发布　　2019－12－23 实施

深圳市标准化协会　　发 布

目　次

前　言

本标准编写格式遵循了 GB/T 1.1-2009 给出的规则编写。

本标准由深圳华大基因科技有限公司提出。

本标准由深圳市标准化协会归口。

本标准主要起草单位：深圳华大基因科技有限公司、深圳华大智造科技有限公司、深圳华大基因股份有限公司、深圳赛西信息技术有限公司、深圳吉因加医学检验实验室。

本标准主要起草人：蒋慧、刘健、伍利、单日强、杨梦、潘光明、丁远彤、李士森、李倩一、杜佳婷、吴昊、姜华艳、王晶晶、吕雪、王溪、陈永胜、黄毅。

本标准为首次发布。

基因数据流通区块链存证应用指南

1 范围

本标准规定了基因数据流通区块链存证应用指南的术语和定义、缩略语、基因数据流通区块链存证应用模型、基因数据流通区块链存证应用原则、基因数据流通区块链存证应用参与方、基因数据流通区块链存证应用关键过程、基因数据流通区块链存证应用系统评估等。

本标准适用于计划使用区块链存证服务的基因数据组织和机构,为建设和实现区块链存证系统提供参考。

2 规范性引用文件

下列文件对于本文件的应用是必不可少的。凡是注日期的引用文件,仅注日期的版本适用于本文件。凡是不注日期的引用文件，其最新版本（包括所有的修改单）适用于本文件。

GB/T 25069-2010 信息安全技术 术语

GB/T 30989-2014 高通量基因测序技术规程

GB/T 35890-2018 高通量测序数据序列格式规范

ISO 20387-2018 生物技术 - 生物建库 - 生物建库一般规则（Biotechnology -- Biobanking -- General requirements for biobanking）

3 术语和定义

下列术语和定义适用于本文件。

3.1

共识consensus

在分布式节点间达成区块数据一致性的认可。

3.2

分布式账本distributed ledger

在分布式节点间共享并使用共识机制实现具备最终一致性的账本。

3.3

加密 encipherment / encryption

对数据进行密码变换以产生密文的过程。一般包含一个变换集合，该变换使用一套算法和一套输入参量。输入参量通常被称为密钥。

[GB/T 25069-2010,定义2.2.2.60]

3.4

区块链blockchain

区块链是使用密码技术链接将共识确认过的区块按顺序追加而形成的分布式账本。

3.5

智能合约smart contract

存储在分布式账本中的计算机程序，其共识执行结果都记录在分布式账本中。

注：本文中，除非特殊说明，合约代指图灵完备的智能合约，即从智能合约代码、智能合约运行时环境均支持图灵完备。

3.6

共识机制consensus mechanism

在分布式节点间达成共识的规则和程序。

3.7

摘要算法digest algorithm

摘要函数；Hash函数，通常通过将任意长度的消息输入变成固定长度的短消息输出来保障数据的完整性。

注：本文中，除非特殊说明，数据摘要及摘要信息均指对输入数据使用摘要算法进行运算处理后得到的输出数据。

3.8

鉴权机制 authentication mechanism

验证用户是否拥有访问系统权利的机制。

注：本文中，鉴权机制也简称为鉴权。

3.9

区块链存证 blockchain proof of existence

为了保证存证信息（电子数据）的完整性和真实性，采用区块链技术实现多节点共识的存证服务。

注：本文中，区块链存证也称为blockchain certifications。

3.10

基因数据 genomic data

与生物先天遗传或后天获得的基因特征相关的数据，一般通过对该生物体样本的基因测序得到，能提供关于该生物体生理或健康方面独特的信息。

3.11

基因测序 sequencing

指分析特定DNA或RNA片段的碱基序列，以获得其碱基排列方式的过程。

[GB/T 30989-2014，定义3.18]

3.12

人类遗传资源 human genetic resources

人类遗传资源包括人类遗传资源材料和人类遗传资源信息，其中人类遗传资源材料是指含有人体基因组、基因等遗传物质的器官、组织、细胞等遗传材料。人类遗传资源信息是指利用人类遗传资源材料产生的数据等信息资料。

3.13

数据获取 acquisition
取得生物样本、基因数据或关联数据的行为。
[ISO 20387-2018，定义3.2]

3.14

数据采集方 data collector
进行数据获取行为的自然人、法人、机构及组织。

3.15

数据主体 data subject
生物样本、基因数据及关联数据源自的个体。

3.16

数据控制方 data controller
可单独或共同决定处理数据目的与方式的自然人、法人、机构及组织。

3.17

数据申请方 data requester
为获得数据查看、处理、使用等权限向数据控制方提出申请的自然人、法人、机构及组织。

3.18

数据持有方 data owner
持有数据的自然人、法人、机构及组织。

3.19

数据监管方 data supervisory authority
对数据全生命周期进行监督管理的独立权力机构。

3.20

数据应用方 data user
使用数据的自然人、法人、机构及组织。

3.21

关联数据 associated data
任何与生物样本及基因数据有关的信息，包括但不限于表型、临床、处理过程等信息。
[ISO 20387-2018，定义3.3]

3.22

FASTA 格式 FASTA format

FASTA是基于文本的、保存生物序列(通常是核酸序列)的、每两行表示一条序列的标准格式。

3.23

FASTQ 格式 FASTQ format

FASTQ是基于文本的、保存生物序列(通常是核酸序列)和其测序质量信息的、每四行表示一条序列的标准格式。

[GB/T 35890-2018，定义3.9]

3.24

SAM/BAM格式 SAM/BAM format

SAM是基于文本的、存储核酸序列和其测序质量信息的、以每一行表示一条序列、每行以制表符分割成11列的标准格式，测序质量信息使用ASCII字符表示，BAM是SAM格式的二进制格式。

[GB/T 35890-2018，定义3.10]

3.25

生物样本建库 biobanking

获取、存储及应用生物样本的过程，亦可包括任何与生物样本收集、准备、测试、分析、分发有关的活动。

[ISO 20387-2018，定义3.6]

3.26

数据销毁 data disposal

移除生物样本、基因数据及其关联数据的行为。

[ISO 20387-2018，定义3.19]

3.27

数据处理 data processing

在数据生命周期对生物样本、基因数据及关联数据进行的任何活动。

[ISO 20387-2018，定义3.36]

3.28

离线传输offline transfer

通过非互联网传输通道，将在线数据复制到另一在线系统的数据复制方式。

4 缩略语

下列缩略语适用于本文件。

API：应用编程接口 （Application Programming Interface）

VCF：变异识别格式（Variant Call Format）

MAF：突变注释格式（Mutation Annotation Format）

5 基因数据流通区块链存证应用模型

5.1 应用模型概述

基因数据流通区块链存证应用模型包含应用原则、相关参与方和关键过程，见图1。

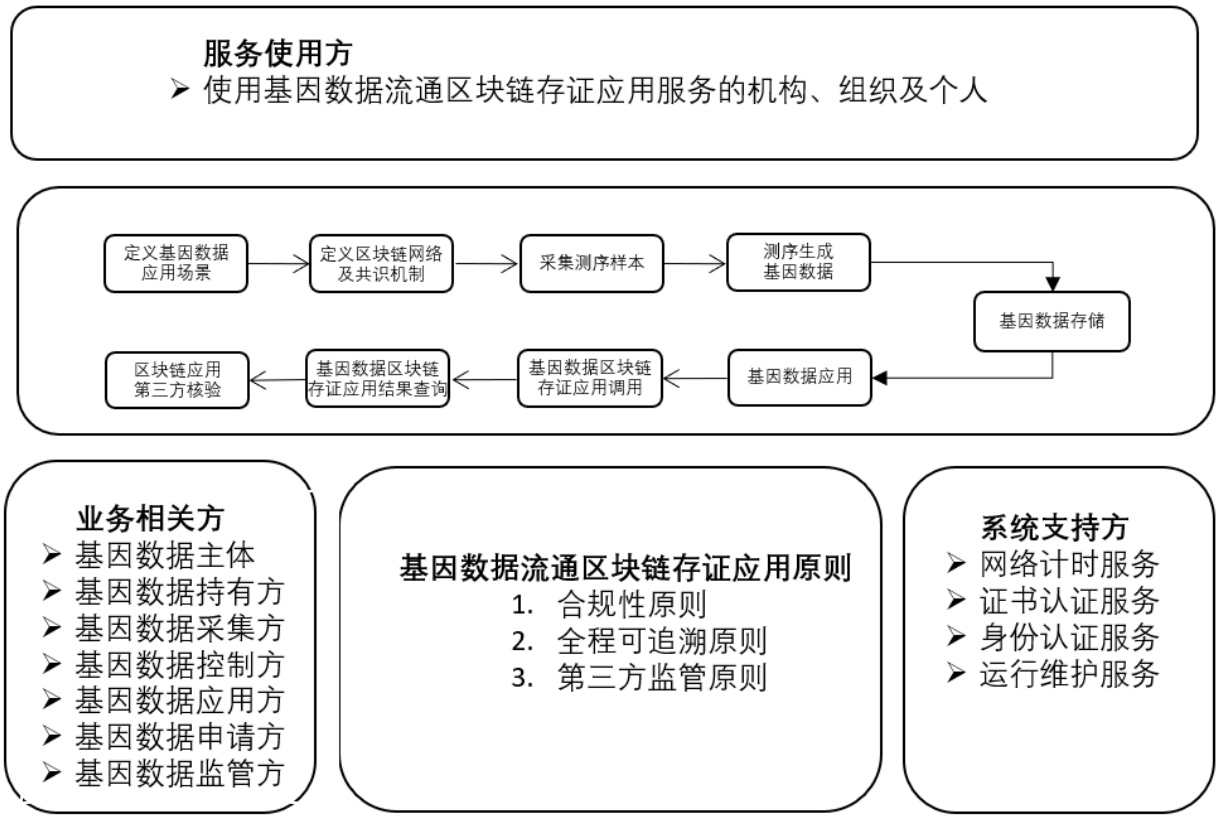


图 1 基因数据流通区块链存证应用模型

5.2 应用原则

基因数据流通区块链存证应用原则包括合规性原则、全程可追溯原则和第三方监管原则。

5.3 参与方

相关参与方分为区块链基因数据存证应用业务相关方、区块链基因数据存证应用服务使用方、区块链基因数据存证应用系统支持方。其中，基因数据存证应用业务相关方包含了基因数据主体、基因数据采集方、基因数据控制方、基因数据持有方、基因数据应用方、基因数据申请方、基因数据监管方。

5.4 关键过程

5.4.1 基因数据流通区块链存证应用关键过程

基因数据流通区块链存证应用关键过程包括：

a) 定义基因数据流通区块链存证应用场景；
b) 定义区块链网络及共识机制；
c) 采集测序样本；
d) 测序生成基因数据；
e) 基因数据存储；
f) 基因数据应用；
g) 基因数据区块链存证应用调用；
h) 基因数据区块链存证应用结果查询；
i) 区块链存证应用第三方核验。

5.4.2 关键过程和业务流程对应关系

基因数据流通区块链存证应用全景图可参考附录A。

基因数据业务流程中各业务环节如图2所示：

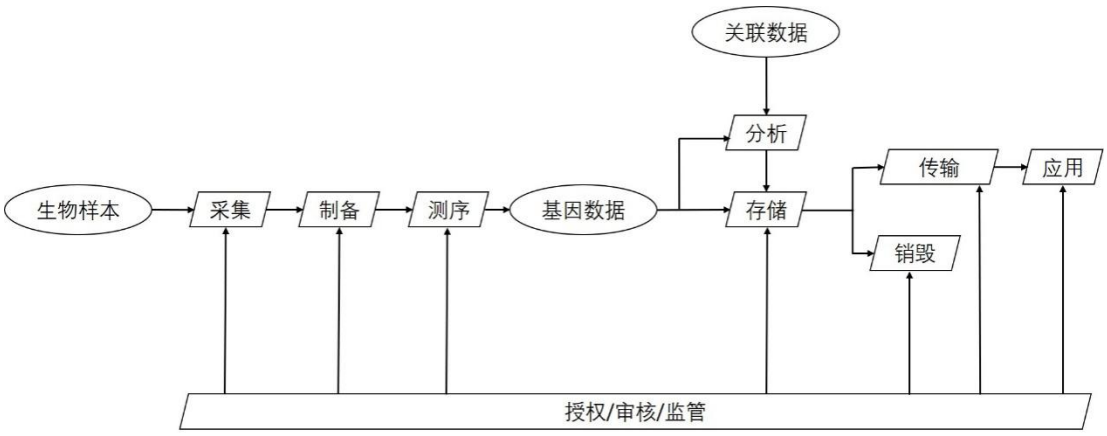


图 2 基因数据存证业务环节

6 基因数据流通区块链存证应用原则

6.1 合规性原则

基因数据业务系统应全面满足法律性要求，在区块链存证系统应用中，业务系统应在多个参与方所属地的法规要求竞合且不一致时遵循最为严格的法规要求， 业务系统宜：

a) 遵循全部参与方所属地的行业规章；
b) 遵循全部参与方的内部规章规范；
c) 设计中考虑未来法规趋严的前瞻性要求。

6.2 全程可追溯原则

基因数据业务系统需实现基因数据流通的全流程可追溯，应包括：

a) 对基因数据流通全流程进行存证。基因数据在不同机构之间的外部流通，应实现存证上链。机构内部的基因数据流通，宜内部存证；

b) 基于基因数据存储内容进行存证追溯。考虑单个基因数据存储文件到多个基因数据存储文件之间和多个基因数据存储文件到单个基因数据文件之间的转换；
c) 针对基因数据的发送方和接收方均进行存证追溯；
d) 支持离线传输的存证追溯；
e) 确保同一个数据文件在不同环节的存证中具有唯一的数据标识，建议使用 SHA512 生成数据摘要作为标识；
f) 针对基因数据业务参与方的厂商机构代码进行预先分配并保证唯一性。

6.3 第三方监管原则

基因数据业务存证系统应支持第三方监管机构接入：
a) 应支持第三方监管的独立区块链节点部署；
b) 机构之间的基因数据流通存证应支持第三方监管的 API 接口调用查询；
c) 宜支持按地域等属性划分的不同层级的第三方监管接入。

7 基因数据流通区块链存证应用参与方

7.1 业务相关方

基因数据流通区块链存证应用业务相关方包括：
a) 基因数据主体；
b) 基因数据持有方；
c) 基因数据采集方；
d) 基因数据控制方；
e) 基因数据应用方；
f) 基因数据申请方；
g) 基因数据监管方。

7.2 服务使用方

基因数据流通区块链存证应用服务使用方指使用区块链存证应用服务的组织、机构以及个人，包括业务相关方的机构、组织及个人和非业务相关的机构、组织及个人。

7.3 系统支持方

基因数据流通区块链存证应用系统根据实际需求，可以引入外部第三方提供支持服务，例如网络计时服务、证书认证服务、身份认证服务、运行维护服务等。

7.4 业务参与机构

各业务流程环节的主要业务参与方和关键过程的对应关系参见表1：

表 1 业务流程环节和关键过程对应关系

关键过程	相关业务参与方	基因数据系统业务目标	参与方机构	可调用的区块链服务
采集测序样本	数据主体、数据申请方	基因数据测序授权	数据主体：自然人用户； 数据申请方：提供基因测序服务的厂商；	存证服务、智能合约授权服务、查询服务
测序生成基因数据	数据控制方、数据应用方	基因数据测序、基因数据存储	数据控制方：提供基因测序服务的厂商； 数据应用方：实际负责基因测序活动的厂商；	存证服务、查询服务
基因数据处理	数据控制方、数据应用方	基因数据测序	数据控制方：提供基因测序服务的厂商； 数据应用方：实际负责基因测序活动的厂商；	存证服务、查询服务
基因数据分析	数据主体、数据申请方、 数据控制方、数据应用方	基因数据应用	数据主体：自然人用户； 数据控制方：提供基因数据存储服务的厂商； 数据申请方：面向系统用户提供基因数据分析服务的厂商； 数据应用方：提供基因数据应用分析服务的厂商；	存证服务、智能合约授权服务、查询服务
基因数据存储/传输	数据主体、数据控制方、数据应用方、数据监管方	基因数据共享、基因数据第三方存储	数据主体：自然人用户； 数据控制方：提供基因数据存储服务的厂商； 数据应用方：提供基因数据应用分析服务的厂商； 数据监管方：具有监管资质的第三方独立机构或组织；	存证服务、查询服务；

注 1：同一厂商在基因数据业务存证活动中可对应多个业务参与方角色；

注 2：同一厂商在不同基因数据业务关键过程中可对应不同业务参与方角色。

8 基因数据流通区块链存证应用关键过程

8.1 定义基因数据存证应用场景

分析基因数据存证应用场景，识别需要满足的法规性要求，针对基因数据进行等级定义，明确基因数据的隐私性和安全性程度，制定清晰的区块链基因数据存证应用目标，包括基因数据业务目标、网络安全目标、合规性目标。应区分人类遗传资源类别的基因数据和非人类遗传资源类别的基因数据。

8.2 定义区块链网络及共识机制

建立区块链基因数据存证应用系统时，应根据存证应用场景及目标定义或选择区块链网络及共识机制,包括区块链网络安全机制、节点共识机制、区块链跨链机制。具体内容如下：

a) 网络安全机制：系统鉴权及准入机制、节点服务器安全加固机制；
b) 节点共识机制：包括基于工作量的共识机制、基于投票的共识机制以及基于第三方公信力的共识机制。
c) 区块链跨链机制：选择不同区块链系统之间实现信息交互的跨链机制；
d) 对于涉及到跨境监管的基因数据，应选用基于许可链的区块链系统并采用严格的节点准入机制确保基因数据流通的合规性。

8.3 采集测序样本

8.3.1 采集生物样本

采集生物样本的过程包括：

a) 数据收集方按照相关法规及伦理要求向数据主体提出数据收集申请，和/或向数据监管方备案；
b) 向数据主体收集生物样本、基因数据和/或关联数据；
c) 进行生物样本建库；
d) 备案信息、数据主体同意记录进行区块链存证；
e) 采集测序样本获得的数据信息进行区块链存证，其中与个人基因组有关的基因数据应进行隐私保护后方能进行区块链存证，可采取的隐私保护方法有脱敏处理、摘要信息上链等。

8.3.2 采集生物样本过程中的存证内容

采集生物样本过程中上链存证的信息内容应包括：

a) 采集厂商的机构代码；
b) 采集样本事件信息：包括但不限于时间、地点、采集方式、样本编码、样本存储方式、样本存储地点；
c) 脱敏后的授权主体信息；
d) 脱敏后的授权记录；
e) 如销毁样本，则应保存样本销毁信息：包括但不限于时间、地点、采集方式、样本编码、样本存储方式、样本存储地点。

8.4 测序生成基因数据

8.4.1 测序生成基因数据

测序生成基因数据的过程包括：

a) 对生物样本进行制备（采集生物样本后处理为可进行基因测序材料的过程，可包括但不限于离心、纯化、固定、扩增、过滤、培养、冷冻与解冻等）；
b) 对制备完成的生物样本进行测序，获得基因数据原始数据，基因数据原始数据一般为 FASTQ 格式。

8.4.2 测序生成基因数据过程中的存证内容

测序生成基因数据过程中上链存证的信息内容应包括:

a) 对基因数据原始数据进行区块链存证，其中与个人基因组有关的基因数据应进行隐私保护后方能进行区块链存证，可采取的隐私保护方法有脱敏处理、摘要信息上链等；
b) 对于大容量的基因原始数据可采用压缩或者数据摘要方式上链存证；
c) 测序样本数据和基因原始数据的对应关系进行区块链存证，对应关系应信息应包括：原始数据存证索引、测序样本数据存证索引、测序仪器设备操作信息。

8.5 基因数据存储与传输

8.5.1 基因数据存储与传输

基因数据存储与传输过程包括：
a) 获得以下格式基因数据并存储：FASTA 格式、FASTQ 格式、BAM 格式、SAM 格式、VCF 格式、MAF 格式等；
b) 不同数据存储或控制机构之间的数据传输。

8.5.2 基因数据存储与传输过程中的存证内容

跨机构流通的基因数据在流出机构和流入机构同时进行区块链存证，存证信息应包含以下内容：
a) 基因数据文件标识：当前传输的基因数据文件的唯一标识；
b) 来源数据的文件标识：当前基因数据文件的上一级源文件的文件标识；
c) 流出厂商机构代码：持有当前传输数据文件并发起本次数据流通的厂商机构；
d) 流入厂商机构代码：接受当前传输数据文件并进行内部存储的厂商机构；
e) 流出厂商机构节点信息：数据传输发起方的节点信息，包括但不限于传输方式、源 IP 地址、域名等；
f) 流入厂商机构节点信息：数据接受方的节点信息，包括但不限于传输方式、源 IP 地址、域名等；
g) 存证上传节点信息：上传本次存证的节点信息，包括但不限于厂商机构名称、源 IP 地址、域名等；
h) 基因数据文件内容信息：对基因数据文件本身的属性信息汇总，应包括：
    1) 基因数据文件格式（数据类别：原始数据、比对数据、突变位点数据等）；
    2) 基因数据文件大小；
    3) 基因数据来源（测序物种、测序平台、时间等）；
    4) 基因数据处理流程（分析流程名称、版本等）；
    5) 基因数据隐私保护等级（是否脱敏、属于限制跨境流通管制范围、是否属于特殊人类遗传资源）；
i) 如销毁基因数据，则应在文件内容信息中备注销毁信息，包括销毁时间、地点、销毁授权记录等。

8.6 基因数据应用

8.6.1 基因数据应用

基因数据应用包括：
a) 基因数据质量检测；
b) 基因数据与生物样本、关联数据进行整合；
c) 对基因数据进行合并、分割、压缩等操作。

8.6.2 基因数据应用过程中的存证内容

基因数据应用过程中的基因数据文件发生更新时（包括原有文件的升级以及产生新的基因数据文件），应对基因数据文件更新记录进行区块链存证，存证的更新信息内容应包括：

a) 新的基因数据文件标识：当前新生成的基因数据文件的唯一标识；
b) 来源数据的文件标识：当前新生成的基因数据文件的上一级源文件的文件标识，即原有基因数据文件的文件标识；
c) 基因数据文件应用的厂商机构代码：持有当前基因数据文件并发起本次基因数据文件更新的厂商机构；
d) 基因数据文件应用的厂商机构节点信息：数据应用方的节点信息，包括但不限于源 IP 地址、域名、机器用户名等；
e) 存证上传节点信息：上传本次存证的节点信息，包括但不限于厂商机构名称、源 IP 地址、域名等；
f) 基因数据文件内容信息：对新生成的基因数据文件本身的属性信息汇总，应包括：
    1) 基因数据文件格式（数据类别：原始数据、比对数据、突变位点数据等）；
    2) 基因数据文件大小；
    3) 基因数据来源（测序物种、测序平台、时间等）；
    4) 基因数据处理流程（分析流程名称、版本等）；
    5) 基因数据隐私保护等级（是否脱敏、属于限制跨境流通管制范围、是否属于特殊人类遗传资源）。

8.7 基因数据流通区块链存证应用调用

基因数据流通区块链存证应用调用中宜：

a) 功能分层及解耦。业务系统服务和区块链系统服务分层实现且解耦；
b) 区块链系统提供存证应用结果查询服务；
c) 基因数据业务系统通过调用区块链存证系统服务提供数据存证服务、数据查询服务；
d) 基因数据业务系统提供异步回调接口给区块链存证系统上传服务调用执行结果。

8.8 基因数据流通区块链存证应用结果查询

基因数据流通区块链存证应用系统宜：

a) 支持基因数据业务层的存证结果查询；
b) 支持区块链系统层的存证应用结果查询；
c) 支持基于区块链节点的存证应用结果查询。

8.9 基因数据流通区块链存证应用第三方核验

基因数据流通区块链存证应用系统宜：

a) 支持法律法规指定的第三方核验，包括数据有效性、身份合规性、时间准确性、算法安全性；
b) 支持业务参与方的核验，包括存证核验；
c) 支持法律法规指定的第三方监管机构针对基因数据业务流程的线上实时合规性审计。

9 基因数据流通区块链存证应用系统评估

基因数据流通区块链存证应用系统评估宜包括：

a） 对系统合规程度的评估；

b） 对系统区块链存证业务服务和基因数据存证业务服务的完备程度评估；

c） 对系统业务可用性的评估；

d） 对系统业务易用性的评估；

e） 对系统业务可扩展性的评估；

f） 对系统业务其他质量属性的评估。

附　录A
（资料性附录）
基因数据流通区块链存证应用全景图

A.1　基因数据流通区块链存证系统应用概述

区块链存证系统提供基本的区块链存证应用服务，基因数据业务系统通过调用区块链存证应用服务实现对基因数据的全流程跟踪及监管，包括：

a）　基因数据存证和溯源：从基因数据采集、基因数据存储到基因数据传输和应用，基因数据在不同机构和节点之间的流通都通过区块链系统实现上链存证；

b）　基因数据审计和监管：第三方监管机构可以通过区块链系统上的基因数据存证对基因数据流通、基因数据应用合规性进行监管审计；

A.2　基因数据流通区块链存证系统应用框架

如图A.1所示，基因数据流通区块链存证应用全景图给出了基因数据流通的区块链存证应用框架，概括描述了基因数据业务系统和区块链存证系统应用服务的逻辑组合关系，表达了如下主要过程：

a）　系统用户使用基因数据存证业务系，触发基因数据存证业务流程；

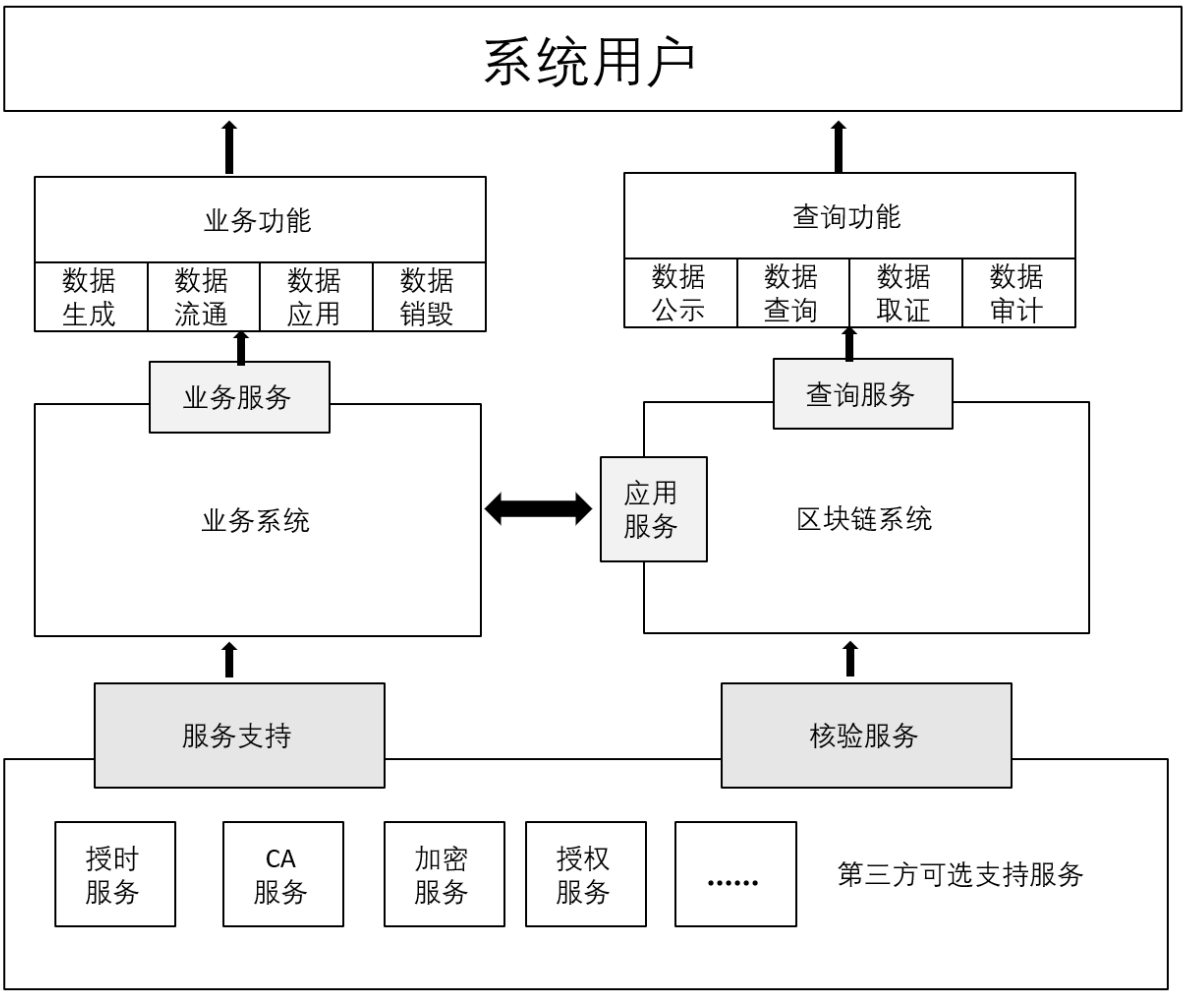
b）　基因数据业务系统根据业务流程进行内部处理，组合调用区块链存证系统服务；

c）　基因数据业务系统根据业务流程进行内部处理，组合调用第三方系统服务；

d）　基因数据业务系统接收并处理区块链存证系统的服务调用执行结果；

e）　基因数据业务系统接收并处理第三方系统的服务调用执行结果；

f）　基因数据业务系统完成基因数据业务流程处理，提供处理结果给系统用户。



图A.1　基因数据存证应用全景图

参 考 文 献

[1] CBD-Forum-001-2017 区块链参考架构
[2] CBD-Forum-003-2018 区块链存证应用指南