Dark Bio: An Open Platform for Personal Health Analysis

Your genome and every layer beyond it, on a device only you control

Péter Szilágyi
peter@dark.bio · dark.bio

Dark Bio AG
Gartenstrasse 6, 6300 Zug, Switzerland
CHE-332.917.708

DRAFT · May 2026

Abstract

A person's health data, their genome most of all, is the most sensitive and the most permanent information they will ever possess. Today, almost none of it is under their control. It is scattered across sequencing labs, hospitals, testing services, app vendors, and wearable platforms, each of which becomes a custodian, a point of leverage, and a point of failure. Analysing it at all has meant handing that data to someone else, and once handed over it cannot be recalled. A leaked genome is leaked forever; for the individual, and in part for everyone related to them.

Dark Bio removes the custodian. Instead of moving health data to wherever it needs to be analysed, the analysis is moved to the data. All of a person's health data is held on a small device, an Ark, under their sole physical possession. Analysis programs are brought to this enclave and run locally inside a deterministic sandbox that has no network and no writable storage. Code can compute on the data but has no uncontrolled channel through which to leak it. The only output is a small result that the owner reviews and explicitly releases from an independent device they already trust, their phone. Every access is recorded in a tamper-evident, cryptographically signed journal.

This inversion does more than protect privacy. Because trust rests in the execution environment rather than in the developer, anyone can publish an analysis, and so health analysis becomes an open, composable ecosystem, the way open source made software an open, composable ecosystem. The Ark is the enabler, but the ecosystem is the purpose. This document describes the architecture, the trust model, the current state of the system, and the many independent directions along which it is built to grow.

A note on scope: Dark Bio does not perform genomic sequencing or similar laboratory work. It begins where that work ends, once a person holds their own data, and concerns itself with only what happens afterwards.

Table of Contents

1 Introduction

In March 2025, the consumer genetics company 23andMe filed for Chapter 11 bankruptcy [1]. By then it had built genetic profiles for roughly fifteen million customers, and that database was a central asset in the court-supervised sale that followed [2]. Two years earlier, a security breach had compromised about fourteen thousand customer accounts directly, and through them exposed genetic and ancestry information belonging to roughly seven million others [3].

It is tempting to read this as one company's failure, but it is something more general. The entire field has made the same structural choice: to keep a person's health data somewhere other than with that person. Once that choice is made, the rest follows. There must be an operator. The operator must be trusted. And any operator, given enough time, can be breached, sold, subpoenaed, or can change their terms. Because genomic data is permanent, the consequences of any such event are permanent too.

Unlike a password or a credit-card number, a genome cannot be rotated, reissued, or revoked. It is the same on the day a person is born and the day they die, and it is substantially shared with their parents, their siblings, and their children. It can indicate a predisposition to disease long before any symptom appears, which is one reason genetic information is given special protection in law in many countries. A genome is not a secret that can be re-secured after a disclosure; it is a fact about a person that, once exposed, stays exposed.

None of this is to claim that a genome stays hidden in everyday life: DNA traces are shed in any hair, saliva, or surface a person touches. But recovering it requires physical access, one person at a time. A database can be reached remotely, and exposes everyone in it at once.

A genome is also only the beginning. The same reasoning applies to blood panels, to the continuous streams from wearable sensors, to medical imaging, to the microbiome, and to every other omic layer that modern medicine is learning to read. Each is sensitive on its own. Together, correlated and longitudinal, they form a uniquely complete and lasting record of a person. Assembling that record is genuinely valuable, because much of the insight in modern health analysis comes from reading one layer against another. Under today's model, though, that fuller picture can be assembled only by concentrating everything with a single custodian and trusting that the arrangement holds.

Dark Bio is built on a different premise: that hope is not a security model.

Today, to analyse any health data, that data travels to the analysis: to a laboratory's pipeline, a cloud service, an application vendor's servers. Dark Bio reverses that direction. The data stays where it is, on a device the person physically holds, and the analysis travels to the data instead. Because each analysis runs inside a sandbox that is structurally incapable of reaching the outside world, it does not matter who wrote it. It can come from a university, a company, an independent developer, or an AI agent, and it still has no way to send data out of the device.

That single property, untrusted code held within a trustworthy boundary, is what turns a privacy device into a platform. An analysis reads raw data only while it runs, inside the sandbox on the owner's own device. Publishing an analysis therefore no longer requires a data-sharing agreement, an institutional affiliation, or a gatekeeper's permission. Anyone can write an analysis. Anyone can run anyone else's. Analyses can build on other analyses. This is the same dynamic that turned software from a product shipped by a few firms into an open commons built by many.

2 Problem

Personal health data is hard to protect for four structural reasons. They are not separate grievances; each compounds the others, and together they explain why stronger policy and better intentions have repeatedly fallen short.

The custodial model fails irreversibly. Any custodian can eventually fail, and for health data that failure cannot be undone. Most data is recoverable: a leaked password is changed, a cloned card reissued. A genome cannot be reissued, and the consequences of its disclosure outlast any individual remedy.

For example, in the United States, the Genetic Information Nondiscrimination Act [4] bars use of genetic information in health insurance and employment, but leaves life insurance, disability insurance, and long-term care insurance untouched. Women carrying BRCA1 or BRCA2 variants, including those with no history of cancer who underwent preventive surgery to reduce their risk, have been denied life insurance on the basis of those test results [5].

None of these consequences is reversed by a breach notification or a settlement; the person may be compensated, but never made whole, because the information cannot be made private again. It takes only one failure, at one custodian, at one moment. Given enough time, such a failure becomes likely; and unlike most failures, this one can never be undone. A model that fails this way does not provide security. It postpones the loss.

Fragmentation hides the multi-omic picture. A person's health data is scattered across the parties that produced it: the genome with a sequencing laboratory, blood panels with testing services, imaging with a hospital, the heart-rate and sleep history with a wearable vendor. No single party holds the whole picture, and the person, who alone has a claim to all of it, holds none of it in a usable form.

Yet the most valuable insights come from combining these layers. A person on several medications faces interaction risks that depend on both what they take and how their body handles each drug. The genome shows the metabolism; the prescription list shows the combination; together they flag real interactions [6]. Each half alone is incomplete: the genome does not know what someone is taking, the prescription list does not know how their body handles each drug. The genome lives with a sequencing lab; the prescription list with a doctor or pharmacy; the person who has a claim to both holds neither in a usable form.

Analysis of that kind is effectively impossible today, because assembling the inputs means gathering them with a single company, which is the most dangerous thing a person could do with their data.

Only data holders can put analyses to use. A method can be developed on public or synthetic data, but validating it, and running it for the people it is meant to help, takes their real data, and that access belongs to whoever holds it, a custodian or someone contracting with one. Polygenic risk scores, statistical methods that estimate disease risk from many genetic variants together, are a case in point: the methods themselves are openly published and widely studied, but new score development remains heavily concentrated in the small number of institutions that hold population-scale genetic cohorts [7]. Independent researchers and developers are shut out not for lack of skill but for lack of data. Progress moves at the pace of a few large organisations.

AI cannot be trusted with the data. Language models can now read the medical literature, reason about genetic variants, and write competent analysis code, which could put expert-level interpreta-

tion within everyone's reach. But they are probabilistic systems: an instruction to "never reveal X" is simply more text added to the input, and a model cannot reliably separate a trusted instruction from untrusted content, so a crafted document can quietly redirect what it does. Prompt injection of this kind is well documented and, at present, unsolved [8]. An agent that can read a genome can be steered, by a single planted line in a file, into placing that genome in an outbound request. The most capable tool for interpreting health data is also the one that can least be trusted to hold it.

2.1 Existing approaches

A number of systems have been built to address these problems, each making a different tradeoff. What they share is structural: the data, or live access to it, rests with a party the user must trust.

- **Institutional databases.** Hospitals and sequencing laboratories restrict analysis to in-house staff. But medicine runs on referral, consultation, and collaboration, so the data is shared outward routinely, by design. Each of those transfers is decided by the institution holding the data, not by the patient, who in practice can neither see how far their data has travelled nor limit it.
- **Trusted execution environments.** Services such as GenomesDAO [9] hold genomic data in cloud vaults built on a processor's trusted execution environment: each runs as a virtual machine with hardware-encrypted memory, and the user's data within it can be unlocked only with a key the user holds on their own device. The model is designed so that not even the operator can reach the data. The difficulty is not that the design is unsound; it may well work exactly as described. It is that the data sits in the operator's data centre, on hardware the user does not hold, inside a software stack the operator builds and updates at its own discretion. Attestation can establish what that stack is at a given moment, but the user cannot prevent it from changing afterwards, and the integrity of the audit ledger rests on the same operator.
- **Compute-to-data platforms.** Ocean Protocol [10] and similar systems keep a dataset with its provider and let approved algorithms run against it in an isolated environment, returning only results. But the data owner does not operate that environment. The dataset is processed on infrastructure the provider runs, and the owner has to take it on trust that the isolation held, that nothing copied the data, and that the provider was not compromised, with no way to verify any of it. Compute-to-data assumes an organisation that already holds a dataset and wants to expose it for computation. A person's own genome has no such organisation behind it, so the model would still require a custodian to hold the data. It changes what that custodian can offer, not whether there has to be one.
- **Health-data marketplaces.** Sequencing.com [11] runs a "DNA App Store" for genetic analysis. To publish, a developer registers under an agreement and submits code that Sequencing.com reviews and approves before listing. Because each app runs on real genetic data, that review is the safeguard: a user who runs an app is trusting that the review caught anything harmful, with no way to check the code themselves. It also makes the marketplace necessarily curated rather than open, with the platform deciding what may be built and how quickly the ecosystem can grow.
- **Consumer health apps.** Apps that pull data from many sources into one place, increasingly with AI-driven coaching, give a person an integrated and interpreted view of their health. But each delivers that view as a vendor-run service: obtaining it means entrusting an ever more complete health profile to one company, under terms the user has to take on trust, cannot audit, and cannot carry elsewhere. The aggregation is delivered. The custody problem is not solved, only made more comfortable.

- **On-device machine learning.** Platform vendors run sensitive computation on the user's own hardware. Google's Private Compute Core [12] isolates on-device machine-learning features within Android so that sensitive data is processed without leaving the phone, and Apple keeps its personal-intelligence processing on-device, extending to a verifiable private cloud only for heavier work [13]. But these systems run the vendor's own models, for the vendor's own features. They are not open platforms: an independent developer cannot place an analysis inside this protected processing. Third-party code on a phone reaches health data the usual way instead, through APIs that hand it out to the app that asks.
- **Fully homomorphic encryption.** FHE allows computation directly on encrypted data, and for genomic analysis it is advancing quickly, with working demonstrations at real scale [14]. But it answers a different question than custody: it is a way to compute on data, not a place to keep it. FHE can complement a custody solution; it is not one in itself.

* * *

Each of these approaches leaves the same question to be answered: who must be trusted with the data? Dark Bio's response is that the question itself is the problem. The data should not be anywhere that requires it to be answered.

3 Precedent

Health analysis today is produced by a small number of institutions and consumed by everyone else. Computing was organised the same way until the personal computer. The transition that followed happened through three changes, each building on the one before it.

Ownership of the hardware moved to the individual. The personal computer was a possession of its user rather than an asset of an institution: what ran on it, and when, was the owner's decision alone.

The machine became a platform. Published system interfaces, development tools, and later application stores meant that software for it could be written by anyone, not only the hardware's vendor. A person who needed a capability could build it themselves rather than wait for a company to provide it.

Distribution opened. A program written by one person could be obtained and run by any other, then extended and recombined into further work. This turned personal computing from a set of isolated tools into a compounding ecosystem.

These three changes produced three conditions: hardware its user owns, a platform any developer can build on, and work that moves freely between users.

Health analysis has only ever had the institutional form. Its data sits on hardware its subjects do not own, which forecloses the first condition and, with it, the other two. Dark Bio is aiming to establish all three for health data: hardware the person owns, a platform any developer can target, and analyses that can be shared and composed.

The difference from computing is the cost of error. A fault in early consumer software cost its user time; a fault in systems that access a person's genome can disclose data that can never be made private again. Each of the three conditions must be built so that it cannot become a way for data to leak out.

4 Sovereignty

The first condition is the one the others rest on: the data must be the person's own, held on hardware they possess, and reachable only through them. Health analysis fails this at the outset, because the data lives with a custodian. Putting it on a device the person holds removes that custodian, but possession by itself is not sovereignty: a held device can still be lost, stolen, or seized; the computer it is plugged into can be compromised; the company that built it can be compelled to open it.

4.1 Dedicated device

The Ark is a single device built for one purpose: to hold a person's health data and to run analyses on it. It is deliberately not a phone or a laptop. A general-purpose computer carries too much software, too many ways in, and too wide an attack surface to be trusted with data that cannot be made private again once it has leaked. The Ark runs only its own firmware and does only this.

It reaches the outside world through a single link, a wired connection to a host computer, and uses no wireless networking of its own. It treats that host as untrusted: nothing the Ark does depends on the machine it is plugged into being clean. The host carries the Ark's traffic and interprets the protocol that drives the device, but sensitive traffic is encrypted end-to-end, to the owner's phone or to the cloud. The host holds no key to it, and the actions that could expose data are authorised on a separate device the host cannot speak for. A hostile host can stall the Ark or refuse to relay its messages, but it cannot read what the Ark holds or act in the owner's place.

One case lies outside this guarantee: importing data through the host. A dataset loaded from the host computer passes through it in the clear, so a host compromised at that moment could copy it. The exposure is limited to the act of import, and it disappears when data is encrypted to the Ark at its source, before it ever reaches the host.

4.2 Two-factor encryption

The health data on the Ark is encrypted while it sits in storage, and decrypting it requires two separate keys. One is hardware-bound to the device itself. The other is held only by the owner, on their phone, and is never sent to Dark Bio or to anyone else. Both keys are needed together to access the data.

One key would be enough to lock the data away, but not to keep it safe over time. Encrypted data can be copied, and a copy is patient: it can be carried off and kept for as long as it takes to obtain its key, whether by theft, by coercion, or by a single careless moment. But patience only works against a key that is a secret kept somewhere, and one of the two keys is not: bound into the Ark's own hardware, it is in no copy of the data and beyond the reach of waiting. So the patient copy is defeated, and reaching that key means taking the Ark itself, a physical theft, not a quiet copy. Even then the device is not enough: decryption also requires the owner's key, both present at the same moment, never assembled at leisure. Neither key alone reveals anything, so losing one is not a breach, only a warning that the data now rests on the key that remains.

The data is therefore beyond the reach of everyone but its owner. The Ark by itself holds only the hardware key, and so cannot read what it stores; it decrypts the data only after the owner approves an unlock and their phone supplies the second key, an access lost when power is cut. A device that is

lost, stolen, or seized is inert in another's hands. The manufacturer never receives the owner's key at all, so it cannot decrypt the data, and cannot be compelled to hand over what it has no means to read.

4.3 Control through companion

The Ark holds the data, but it does not decide what is done with it. That decision belongs to a second device: the owner's own phone, running a Companion application. The two are paired once, in a step that can only be done with the Ark physically in hand, from which point the phone is the controller.

The Companion holds the owner's key, without which the stored data stays sealed. That key can be backed up, so that a lost or replaced phone does not cost the owner access to their data; like the key itself, the backup is the owner's to keep and is never given to Dark Bio. The Companion is also where every consequential action is authorised. Unlocking the data, running an analysis, releasing a result: each is presented to the owner on their phone, in plain terms, and proceeds only when they approve it there. The two devices reach each other over a channel encrypted end-to-end.

Holding the data on one device and authorising its use from another is deliberate. An attacker who reaches one has not reached the other: the Ark stores the data but cannot act on it alone, and the phone can authorise but holds no health data to take.

The split has a further consequence. The Ark is needed only while in use, and otherwise sits powered off and connected to nothing, presenting none of the standing target an always-online system does. And because the two devices are seldom in the same place, an attacker set on having both faces two separate physical thefts, not one remote intrusion.

4.4 Cryptographic audit trail

The Ark keeps a journal of its own security-significant events: each time it starts, each pairing, each firmware update, each attempt to unlock the data, whether it succeeded, was refused, or timed out, and each analysis run, recorded with the parts of the data it was allowed to read. Its purpose is to leave a trail: anything done to the device that its owner did not do is recorded, and the owner can review that record afterwards on their Companion.

Each entry is written once and chained to the one before it by a cryptographic hash, with the latest entry signed by the Ark; when the Companion fetches the journal it verifies this chain, and reports any broken links rather than passing over them. An entry that is altered, removed, or inserted breaks the chain. The journal is not what makes the device trustworthy, and is not meant to be; it is the means by which interference with the device leaves a trace its owner can later find.

* * *

These four together remove the custodian. No operator holds the data, so none can be breached, sold, or compelled, and the failure this document opened with has no equivalent here. What remains to be trusted is far narrower, and open to inspection: the firmware is published and built reproducibly, the hardware runs only firmware that carries a valid signature, and the device attests to the version it runs, so anyone can confirm an Ark runs exactly the published firmware and nothing else. Updates take effect only at the owner's choice, never silently, and a release built to weaken the device would not pass that same public check. With the data secured, the next section takes up what can be done with it, given that it never leaves the device.

5 Extensibility

The data is secure, and it does not move. To analyse it, the analysis must therefore come to the data: a program is brought onto the Ark and run against what is stored there. The useful programs, though, will not be ones the owner wrote; they will come from researchers, companies, and developers the owner has never met. Running a stranger's code on the device that holds a person's genome is, on its face, a reckless thing to do. It is safe only if that code, whoever wrote it, has no way to carry the data off; and if it is safe, then it no longer matters who wrote it, and anyone can write one. That is the second condition: a platform open to every builder.

5.1 Trust in the sandbox

An analysis runs inside a sandbox on the Ark, built around one idea: a program may see the data, but must have no way to send it anywhere. So the sandbox gives it nothing to send with. It has no network and thus no route to the outside world.

The analysis can read the data it is given and compute on it. Its only output is a bounded result, far smaller than what it read. Whether that result goes any further is the owner's to decide.

Because the boundary holds whatever the code does, the author does not have to be trusted. Nobody needs to vet the developer, confirm their institution, or read the program first. A malicious analysis is held by exactly the boundary that holds an honest one: it reaches no network, leaves nothing behind, and the most it can do is write what it chooses into the one bounded result the owner reviews before it goes anywhere. Trust no longer rests on the author, so the platform has no reason to be closed: anyone can write an analysis, and anyone can run another's.

The platform is open in one further way. Every analysis runs as wasm, a low-level format that code in almost any language compiles to, so a developer builds for the Ark in a language they already know. And because every analysis reaches the sandbox in that one form, an analysis in one language can be built on as readily as one in another, with none of the usual barriers between them.

5.2 Data indexed for analysis

A sandbox that anyone can build for is only as open as it is easy to build for. If writing an analysis meant first turning the owner's data files into something queryable, parsing each format, building the indexes, cross-referencing one source against another, only specialists would write one, and the platform would be open in name only.

The Ark does this work itself, when data is loaded onto it. The files are transcoded, indexed, and cross-referenced, then presented to each analysis as a read-only filesystem. The indexing is deep enough that an analysis queries the data at a high level: one file gives the genotype at a chromosome position, another a named gene, another a blood marker. The analysis itself parses nothing and builds no index.

The data is offered as files as reading one needs no client library and no protocol; an analysis reads a gene or a marker as any program reads a file, and what its author learns is just the layout, the paths and what lies at them. A programming interface would have each analysis link a client and track that interface as it changed; reading a file needs none of that.

What an author writes, then, is the analysis itself. It names the data it needs and is given only that, then opens files and reads values; the device has already done everything beneath that.

5.3 AI agents as builders

The platform is built for AI agents as much as for people. As an agent can read across the medical literature far more widely than a person can, and turn what it finds into analyses; it can write one for the specific question a person is asking, on demand, rather than waiting for a suitable app.

This has not been safe to do before. A model cannot reliably separate a trusted instruction from the contents of a file, so an agent that can see a genome can be talked into sending it elsewhere. What changes on the Ark is that an agent building an analysis never sees the genome at all.

An agent is a developer like any other. It is given the layout of the data, the files that exist and the form of each, and writes an analysis against it; what it writes runs in the ordinary sandbox, contained exactly as any other code is.

The most capable tool for interpreting a genome can therefore be used on one without being trusted to hold it; to the boundary, an analysis from an agent and one from a researcher are the same thing. What the boundary does not settle is whether an analysis is any good: an agent, like any author, can write code that runs safely and still reaches a wrong answer, and whether its conclusions are sound is a separate question from whether it is safe to run.

* * *

The Ark is now a platform, holding a person's data and running against it, safely, the work of any builder. That work has so far stayed on the one device; how analyses spread between people, and build on one another, is the next section.

6 Shareability

Extensibility established that anyone can write an analysis for the Ark. For those analyses to form an ecosystem, two more things are needed: an analysis must be able to reach everyone who might run it, and a new analysis must be able to build on the ones already written. Distribution and composability are the third condition.

6.1 Sharing analyses

An analysis is a single file, a compiled wasm program. Anyone can publish one the way they would publish any file, on a web server or a code repository, with no registry to enter. This is safe only because of the sandbox. Existing marketplaces have to review and approve every analysis they list, since each runs on real data and a harmful one would otherwise reach users unchecked. An analysis that cannot exfiltrate data carries no such risk, and can be shared as freely as any other file.

Dark Bio runs a Hub, a place to find analyses, and for their authors a place to sell them. It is a catalogue and a payment system, not a gatekeeper: it indexes what authors publish so an owner can find an analysis suited to their question, and takes a listing down only when abuse is reported. But it's not a necessity. An analysis can be published and obtained with no Hub involved, so Dark Bio does not control which analyses exist or who runs them, it's just a convenience feature to Ark owners.

6.2 Composing analyses

An analysis need not be written from scratch. It reads files and may write files, so the output of one analysis can be the input of another, and something complex can be assembled from simpler parts. A program that gathers variants bearing on a condition, a model that turns those into a risk estimate, a step that renders the estimate as a readable report: each can be a separate analysis, by a different author, and a fourth can be little more than the three run in sequence.

Analyses divide into two kinds, by what they produce. Some produce a report, for the owner to read, and that is the end of a chain. Others produce data, a new layer of the filesystem, read by whatever analysis comes next exactly as the original data is read. The platform keeps the two kinds apart, which also settles what can leave: a report can be reviewed and released, but data is data, and stays on the device like the rest of it.

Composed this way, a field stops rebuilding what it has, the way software has always grown. Health analysis never could, because the data an analysis runs on has been locked away with custodians. On the Ark it is in one place, open to every analysis, so analyses can finally build on one another.

* * *

The three conditions are now in place: the data stays with its owner, on hardware they hold; anyone can build for it, because trust rests in the boundary and not in the author; and the analyses they build are shared and composed freely, while the data itself never moves. Each made the next possible. Together they are the basis for a field of health analysis with no custodian in it, one anyone can contribute to and anyone can draw on. What that field looks like is the subject of the next section.

7 Ecosystem

The three conditions are worth having for what they make possible. The same device and the same open platform serve one person analysing their own health, two people with a question that needs both their genomes, and a whole population in a study. None of it requires a custodian.

7.1 Analysis on one device

The commonest case is a person analysing their own data, and the range of it is wide. One analysis reports which drugs a person's genome makes ineffective or dangerous for them, the pharmacogenomic guidance most prescribing still goes without. Another screens for the recessive conditions they carry and could pass on. Another computes a polygenic risk score for a common disease from variants spread across the genome. Each is a single program, written once, run by anyone, on a device that already holds the data it needs.

These read the genome alone. Other analyses read more than one layer at once. A predisposition inferred from the genome means little by itself; set against years of blood panels and the daily record of a wearable, it becomes specific enough to act on. That reading needs the genome, the bloodwork, and the wearable history in one place, which no institution holds and which a person has had no safe way to gather. On the Ark they are already together, and an analysis can move across all of them.

Nor must an analysis exist before the question it answers. A person can ask an AI agent how their own variants bear on a new finding, or what the research implies for someone with their particular markers. The agent reads the literature, works out which of the device's data are relevant, and writes an analysis for that question in particular, without ever seeing the data itself. The result is an analysis fitted to one person, produced on demand.

The device also travels. Handed to a specialist clinic, it lets the clinic run its own diagnostic pipeline against the owner's data, on the clinic's premises, with nothing copied into the clinic's systems. The owner approves each step from their Companion, and the journal records what the clinic's software did and which data it read. A second opinion stops meaning a second copy of one's genome in a second institution's keeping.

7.2 Collaborating devices

Some questions are not about one genome but two. Whether a couple both carry the same recessive condition, and what that would mean for a child, is one of them. Two Arks can settle it between themselves. Each verifies that the other is a genuine device; the carrier screen runs on each genome separately, on its own device; and the devices exchange only what the joint result needs, the carrier flags, never the genomes they came from. Each partner sees the combined finding on their own Companion. Neither has shown the other their genome, and no clinic or service has seen either. The same arrangement answers any question that spans related genomes.

7.3 Trade-secret algorithms

Most analyses are published openly, but the platform does not require it. An author may want to sell one without giving away how it works, releasing neither its source nor its compiled binary. The Hub delivers such an app encrypted to the buyer's Ark, and only after the device's attestation proves it

genuine. The host that carries the bytes sees only ciphertext, and the Ark decrypts the app and runs it in the ordinary sandbox. The Ark runs only signed firmware, and that firmware offers no way to read a loaded app back out, so the owner cannot lift the binary off the device either. The owner loses nothing by this: trust in an analysis never rested on reading its code, only on the sandbox that holds it, and that sandbox holds a closed app no differently from an open one.

Some authors will not ship a binary at all, sealed or not. A drug developer may keep its risk model on its own systems alone, offering the use of it but never the program itself. Such an algorithm cannot come to the Ark, so the direction reverses: rather than the analysis coming to the data, the data goes to the algorithm, encrypted. The Ark encrypts what the algorithm needs with fully homomorphic encryption, which lets a computation run on ciphertext and return an encrypted result. The ciphertext leaves the device, the algorithm runs on it elsewhere, and the encrypted result returns to the Ark, which alone holds the key to open it. The company computes on the genome without ever seeing it, and its algorithm never leaves its own systems; neither side gives up its secret.

7.4 Population studies

Research needs scale, and scale has meant a pool: a biobank, a central database, a population's data gathered in one place and exposed by being gathered. Much of it need not be. A research group can publish an analysis like any other; thousands of people can choose to run it, each on their own device, against their own data; and each returns only the summary the study asked for, a set of variant frequencies, a correlation, a statistic, never the data beneath it. Those summaries combine into a population-scale result, with no raw data pooled to produce it and no genome leaving the device it sits on. Where a study's questions can be answered from aggregates, and a great many can, a population can be studied without a population's data being collected anywhere.

Whether people take part is a separate question from whether they can. Most, asked to contribute to research that means nothing to them, will not, and nothing here changes that. But a study of a particular condition is addressed, above all, to the people who live with it, and they have a reason of their own to take part: research into their condition is what stands to improve its treatment, or one day cure it. The system does not manufacture that motive; the condition supplies it. It just removes the price participation used to carry, a permanent copy of one's genome left in an institution's keeping.

* * *

None of this is a set of separate products. It is one device and one platform, used by a single person or by thousands, for one clinical question or for a study spanning a population. What makes it one thing is what has held throughout: the data does not move. Every use described here leaves each person's data where it has always been, on the device they hold. This is the ecosystem the three conditions produce, an open and compounding body of health analysis with no custodian anywhere in it. The next section sets out what the design does not solve, and what is not yet built.

8 Limitations

Every section so far has described what the system does. This one describes what it does not do. The limits are of two kinds: boundaries of the design, which no amount of building will move, and gaps between the design and what exists today, which building will close.

8.1 Boundaries of the design

The system protects a person's data from everyone except that person; it cannot protect it from the person. An owner compelled, by law or by force, to unlock the device and release a result will do so, and the system carries the instruction out like any other. Two keys, a separate Companion, an audit trail: none of them resist an owner made to cooperate. The design leaves coercion as the only way in: no quiet path around the owner, no operator to be served an order instead, no copy held elsewhere to subpoena. An attacker is left with one option, and it is one the owner would see being used. That is a meaningful narrowing, and it is not the same as protecting a person who can be compelled.

The Ark is built by Dark Bio, and the owner trusts that it was built honestly. Firmware that is published and reproducibly built lets anyone confirm what software a device runs, but it says nothing about the silicon beneath. A manufacturer could, in principle, place something in the hardware that no inspection of the firmware would find, and an owner cannot themselves rule that out. The design moves trust off operators and developers; it does not move it off the manufacturer. It reduces that trust to one party and one moment, the making of the device, rather than a relationship that must hold for as long as the data exists.

A person with the device in hand, and the right equipment, can lift its storage and copy it. The copy is encrypted and still needs both keys to open, so it does not surrender the data. Its real threat is to the record. An older copy of the storage can later be written back, returning the device to a past state and erasing the journal's account of everything since. The journal is tamper-evident against entries altered, removed, or inserted within it; it is not, on its own, evidence against a whole earlier copy being restored in place. Catching that needs a hardware counter the Ark does not yet carry.

An analysis cannot move data out in bulk, but it can write into the result it returns, and the result is what the owner releases. A malicious one could fold a fragment of the data into a result that otherwise reads normally, and an owner reviewing it would not be sure to notice. The channel is narrow, far narrower than the data behind it, and nothing leaves until the owner approves it. But it is not nothing, and a determined analysis can use the width that remains.

The sandbox governs what an analysis may do, not whether it is any good. An analysis that runs safely can still be wrong: a risk miscalculated, a variant misread, a confident conclusion resting on nothing. Code an AI agent wrote can be wrong the same way, under the same guarantees. The boundary makes an analysis safe to run; it does not make it correct, and nothing here should be read as claiming otherwise. Whether an analysis is sound is settled the way a field settles such things, through testing, review, and reputation, and the platform leaves room for those rather than standing in for them.

8.2 First mile problem

The Ark takes no part in producing a person's raw health data. A genome is sequenced by a laboratory, bloodwork runs through clinical instruments, imaging is done by a hospital, even a wearable's readings begin on a device the owner did not build. Whoever does this work handles the sample and sees the data first. The reasonable question is whether anything that follows can matter, given that.

It does, for two reasons. The harm at the first mile is a different shape from the harm afterwards. A laboratory's exposure of one sample is bounded to that one transaction, the one company, the one time. The custodial accumulation of a lifetime of records, by contrast, is bounded only by how long the operator lasts. Addressing the second does not require solving the first.

The second reason is that the first mile is not fixed where it is. The instruments that produce health data could, in principle, encrypt their output to the owner's Ark before it leaves them, so that the laboratory operating the instrument never sees it. Nothing in the design of a sequencer or an analyser prevents this, and such a model would fit naturally with what the Ark provides. Instruments of that kind do not exist today, and building them is not Dark Bio's role: it is for instrument makers, accreditation bodies, and regulators to drive. But it is a direction the field can move in, not a wall.

Sequencing in particular is the layer of the first mile where the technology has already moved closest to the owner. Portable nanopore-based devices are small enough to sit on a desk and are in use today in field and point-of-care settings [15]. Whether such devices become cheap and accessible enough for routine personal use is a market question rather than a settled outcome, but the trajectory of the hardware over the last decade has been steadily toward smaller, more affordable instruments. The first mile in this layer narrows as that continues.

None of this is to claim the first mile is solved. It is to say that the custodial problem after data arrival is distinct from the first-mile problem, and that addressing one should not require solving the other.

8.3 Losing the device

The Ark is one device, and hardware fails. A device that is lost, broken, or destroyed takes its data with it, and the two-key encryption that protects that data is exactly what makes it unrecoverable from the device alone. Against this, an owner can keep a backup, held in Dark Bio's cloud or somewhere of their own choosing.

A backup cannot be as safe as the data on a working device. To be restorable onto a replacement, it cannot be sealed to the hardware key of the device that failed, since that key is gone with it; a backup is therefore a copy that the device-bound half of the two-key scheme no longer guards. It is still encrypted: whatever holds it, Dark Bio's cloud included, keeps only ciphertext, and a breach of that store gives an attacker nothing they can open. But it needs a second factor of its own to replace the missing device key, so that a stolen phone is not by itself enough to open it. A backup trades part of the on-device guarantee for the survival of the data, and that trade is the owner's alone to make.

9 Conclusion

Dark Bio is a system for personal health analysis that requires no custodian. The data stays with the person, on a device they hold, and analysis is brought to the data instead of the data to the analysis. Each analysis runs inside a boundary it cannot reach past, so the person running it need not trust whoever wrote it. From that one property the rest evolves: trust placed in the boundary and not the author, a platform anyone can build on, analyses that circulate and build on one another while the data itself never moves. A device built to protect health data becomes the ground for an open field.

The pattern is not new, computing went through it once. The personal computer made the hardware the individual's own, published interfaces made it a platform anyone could build on, and the software written for it became something shared and combined. Sovereignty, extensibility, and shareability are those same three stages, carried over to a person's health data.

The cost of failure changes. The custodial model fails for everyone held in it at once, as the 23andMe breach did. The Ark fails one device at a time: an owner compelled, a device physically taken apart, a device lost, each reaching a single person and no one else. That is grave for that person, and not to be waved away; but it is not the population-scale event a central breach is. Our design does not make failure impossible. It makes it local.

The custodial model gathers a population's permanent data into one place and asks everyone to trust that it holds, indefinitely. Dark Bio asks for something narrower, that a person trust a device they hold in their hands. That is the exchange, and a far smaller thing to ask.

Appendix A: Reference diagrams

The two figures below summarise the architecture described in the body. The first shows what an Ark and its Companion hold, and how the two keys combine to keep the data encrypted. The second shows how an analysis reaches an Ark, who places it there, and how the Companion sits between the owner and the device.

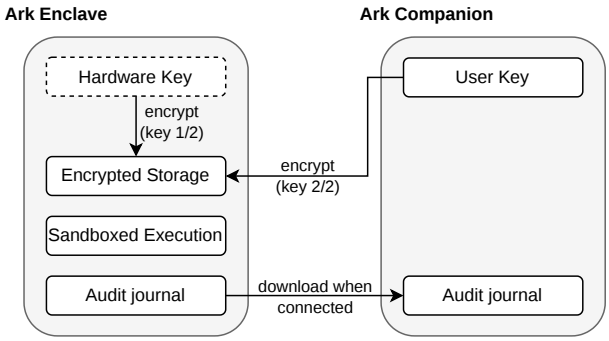

Figure 1: The Ark and the Companion. The Ark holds encrypted storage, a sandboxed execution environment, and an audit journal. Encryption combines two keys: the hardware key, bound to the Ark and never leaving it, and the user key, held on the Companion. Neither alone is sufficient to decrypt the data. The audit journal is mirrored to the Companion whenever the two are connected.

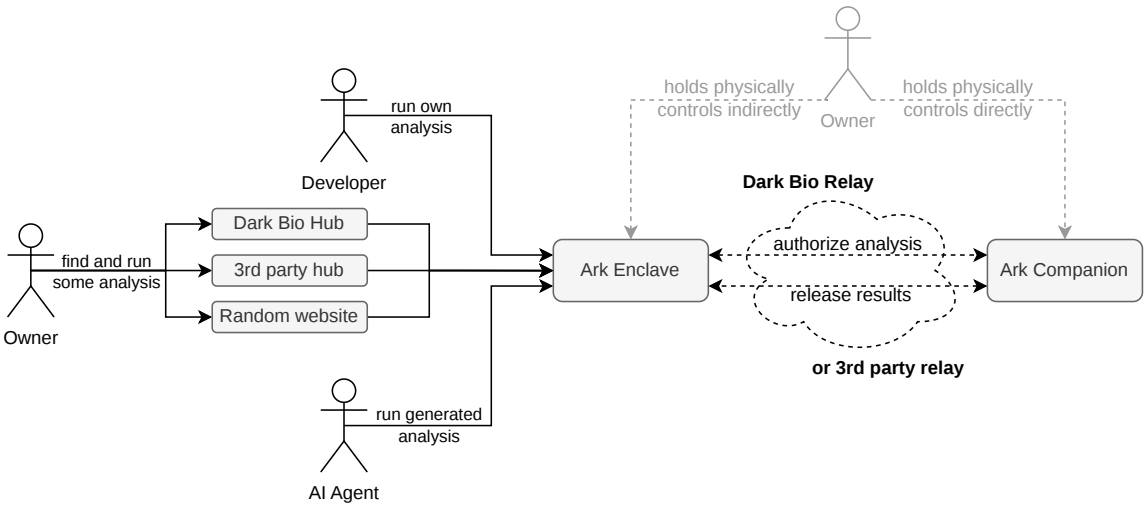

Figure 2: How an analysis runs on an Ark. Analyses can come from the owner themselves, from a developer, or even an AI agent; and can be found through Dark Bio's hub, a third party's, or any other source the owner chooses. The owner authorises an analysis through the Companion they hold; the request and the result pass between the Companion and the Ark through a convenience relay, which can be Dark Bio's or a third party's.

References

[1] 23andMe, "23andMe Initiates Voluntary Chapter 11 Process to Maximize Stakeholder Value Through Court-Supervised Sale Process." [Online]. Available: https://investors.23andme.com/news-releases/news-release-details/23andme-initiates-voluntary-chapter-11-process-maximize

[2] "Judge OKs sale of 23andMe — and its trove of DNA data — to a nonprofit led by its founder," *NPR*, June 2025, [Online]. Available: https://www.npr.org/2025/06/30/nx-s1-5451398/23andme-sale-approved-dna-data

[3] "23andMe confirms hackers stole ancestry data on 6.9 million users," *TechCrunch*, Dec. 2023, [Online]. Available: https://techcrunch.com/2023/12/04/23andme-confirms-hackers-stole-ancestry-data-on-6-9-million-users/

[4] U.S. Congress, "Genetic Information Nondiscrimination Act of 2008 (Public Law 110-233)." [Online]. Available: https://www.congress.gov/bill/110th-congress/house-bill/493

[5] Joly et al., "Time to End the Use of Genetic Test Results in Life Insurance Underwriting," *Journal of Law and the Biosciences*, 2021, [Online]. Available: https://pmc.ncbi.nlm.nih.gov/articles/PMC8607993/

[6] Swen et al., "A 12-gene pharmacogenetic panel to prevent adverse drug reactions: an open-label, multicentre, controlled, cluster-randomised crossover implementation study," *The Lancet*, Feb. 2023, [Online]. Available: https://www.thelancet.com/journals/lancet/article/PIIS0140-6736(22)01841-4/fulltext

[7] "Global trajectories of polygenic risk score research: a systematic bibliometric review of precision medicine, equity, and clinical translation," *Frontiers in Medicine*, 2026, [Online]. Available: https://www.frontiersin.org/journals/medicine/articles/10.3389/fmed.2026.1779659/full

[8] OWASP, "OWASP Top 10 for Large Language Model Applications." [Online]. Available: https://owasp.org/www-project-top-10-for-large-language-model-applications/

[9] GenomesDAO, "Genomes.io." [Online]. Available: https://genomes.io/

[10] "Compute-to-Data." [Online]. Available: https://docs.oceanprotocol.com/developers/compute-to-data

[11] Sequencing.com, "Adding Apps to the DNA App Store: App Market API Guide." [Online]. Available: https://sequencing.com/developer-documentation/api-guides/app-market-api

[12] "Android Private Compute Core Architecture," 2022. [Online]. Available: https://arxiv.org/abs/2209.10317

[13] Apple, "Private Cloud Compute: A new frontier for AI privacy in the cloud." [Online]. Available: https://security.apple.com/blog/private-cloud-compute/

[14] "Secure large-scale genome-wide association studies using homomorphic encryption," *Proceedings of the National Academy of Sciences*, 2020, [Online]. Available: https://www.pnas.org/doi/10.1073/pnas.1918257117

[15] “Portable nanopore-sequencing technology: Trends in development and applications,” *Frontiers in Microbiology*, 2023, [Online]. Available: https://www.frontiersin.org/journals/microbiology/articles/10.3389/fmicb.2023.1043967/full