Editors: Mohamed Kaâniche, mohamed.kaaniche@laas.fr | Aad van Moorsel, aad.vanmoorsel@ncl.ac.uk

E-biobanking: What Have You Done to My Cell Samples?

Paulo Esteves Verissimo and Alysson Bessani | University of Lisbon

Biobanking (collecting and storing human biological material[1]) is becoming extremely important, with deep societal, medical, and scientific implications. A major driving force has been DNA sequencing, which has undergone a revolution in the past six years with the advent of next-generation sequencing (NGS) machines. These machines have lowered sequencing's price and increased its speed, by several orders of magnitude. Figure 1 shows some simple statistics about the recent evolution of genome sequencing's cost.[2] The predictable result is that the number of produced and stored DNA base pairs will skyrocket.

At first sight, the way to deal with all the related data would simply be for the concerned organizations—hospitals, research labs, and so on—to invest more in their IT (for example, see http://bbmri.eu). However, a disturbing trend has recently been unveiled: under fixed cost, the genome-sequencing capacity (DNA base pairs per dollar) is growing more quickly than storage capacity.[3] This trend is bound to continue, with no basic technology solution in sight.

Bottom line: Every day the stored raw data increase by quite a few dozens of terabytes. The burden this imposes on the concerned organizations' IT systems is a major issue because it will become unsustainable in the short term.

Solutions?

An obvious first instance of solutions to this problem would be to improve biological data's storage factor—for example, by aggressively using compression. However, the grand challenges for storage and postprocessing that NGS has introduced have put cloud computing on the agenda.[3] This might well alleviate the pressure on IT and help put the base-pair-versus-megabyte cost gap back on the right side.

This solution might also catalyze a disruptive evolution in biobanking. We believe we're observing the advent of the new era of e-biobanking. That era will witness the evolution toward using public clouds, coexisting and interplaying with dedicated data centers and private clouds.

Although that solution is a road worth traveling, it brings significant security and dependability challenges. As the dependence on computerized representations of physical samples grows, so does the need to guarantee incremental levels of dependability (reliability and availability). Laboratories will assume they can access any cataloged sample at the click of a finger, all day, every day. This implies that challenging standards of availability have been met, by not only the storage but also the computing infrastructure.

But even more worrying is that a perhaps significant part of this material is security and privacy sensitive, such as information on an individual's diseases, or individual genomes. In many countries (especially in Europe), strict regulations for protecting data, including DNA and disease information, prevent using public clouds for those data. However, enormous pressure exists for relaxing those regulations, to the detriment of citizens' privacy.[4] What's often forgotten is that a genome or chronic disease isn't like a password. Once it's compromised, you can't change it; the harm is permanent and can extend to a victim's descendants, ancestors, and other family members.

When trying to meet the economics or usability challenges, such as lowering storage costs or improving data accessibility and sharing, we should always keep in sight nonfunctional aspects such as reliability, availability, integrity, and privacy. Otherwise, the promised shiny e-biobanking future might become very cloudy.

 Copublished by the IEEE Computer and Reliability Societies

Cloudy Weather Ahead

Recent developments show that the movement toward using clouds is irreversible. It's in the road map of many a player, from biobanking consortia, through genomics and bioinformatics communities, to NGS machine vendors (for example, see BaseSpace; www.illumina.com/software/basespace.ilmn). The clouds won't just be private; they'll also be public. The access won't just be restricted to internal users; it will also be public and Web based.

You can't view the move to public clouds in isolation from the greater dependence on IT that e-biobanking is bringing. The challenges we expressed in the previous section must be put in context with a worrying roster of failures that major cloud providers have suffered over the past few years—including data and privacy loss. These failures call for dramatically better cloud resilience, especially when critical applications are at stake.[5]

In addition, a false feeling of security brought by the announced use of standard protection techniques (secure sockets layers, firewalls, anonymization, and so on) might impel people to subject critical data to a high level of threat. However, the risk in protecting digital assets is commensurate with the combined level of threat and degree of vulnerability for those assets.[6] Critical biological data have been reasonably screened from threats until now, even if kept in vulnerable systems, because they were isolated. The pressure to expose data without completely understanding the resulting risk might have serious consequences.

For example, the move to create massive repositories of e-biobanking data includes public but anonymized genome databases and cohort studies of patients. As precious as these data sources are for science, they might be just as damaging for those willing citizens if *reidentification* occurs—that is, public genomics or clinical data are tracked back to individuals. This is easier than it looks: reidentification attacks have already happened in supposedly anonymized real data (including health data) that had been publicly released.[7,8]

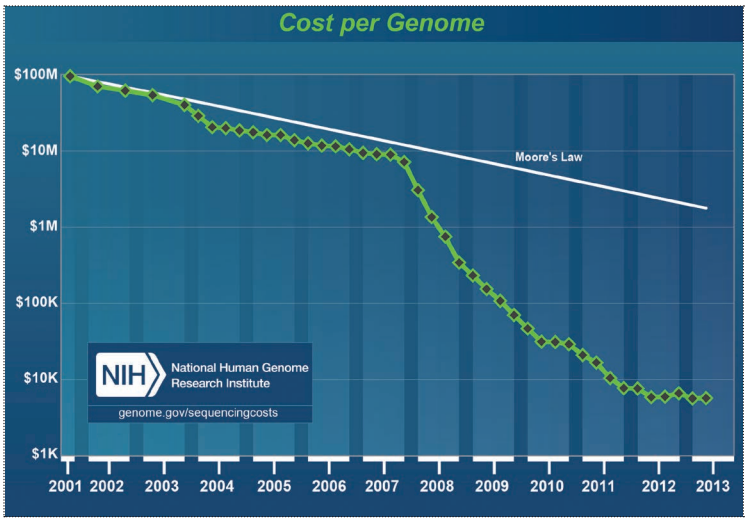

Figure 1. The evolution of genome sequencing's cost.[2] Owing to the decreased cost, the number of produced and stored DNA base pairs will likely skyrocket. (Source: the US National Human Genome Research Institute.)

Going a bit further, in a world of massive but not-so-careful storage and manipulation of genomic data, the prospect of targeted attacks engineered from hacked DNA samples becomes daunting. Threats could get much more worrisome, such as exploring a person's genetic fragilities or massively propagating carefully crafted diseases for terrorism or illegal profit (for example, to increase drug sales). Likewise, cybercriminals might be hired to modify suspects' DNA profiles in police databases. With synthetic DNA technologies, planting false evidence connecting someone to a scene without ever having met that person will no longer be a sci-fi subject: hacking that person's DNA from a database will suffice.[9]

Security and Dependability

The European BiobankCloud project is addressing these major challenges. Here, we focus on how it's handling security and dependability of storage and computing. However, it includes much more interesting work; details are at www.biobankcloud.eu. Meeting the challenges we discussed earlier translates, for cloud ecosystems, into enforcing three objectives, despite threats:

- Promote easy but secure access by both occasional and sophisticated users.
- Enforce dependability against cloud outages and other failures.
- Guarantee data security (integrity and privacy) against unauthorized users and malicious insiders in cloud providers.

Figure 2 depicts a high-level view of the BiobankCloud storage architecture. BiobankCloud is drawing on recent results from yet another European project, TClouds (see www.tclouds-project.eu), to create storage architectures based on a *cloud of clouds*. That is, multiple instances of private and public clouds from several stakeholders and providers will participate in creating the abstraction of a single cloud-based key-value store or file system. E-biobank users and

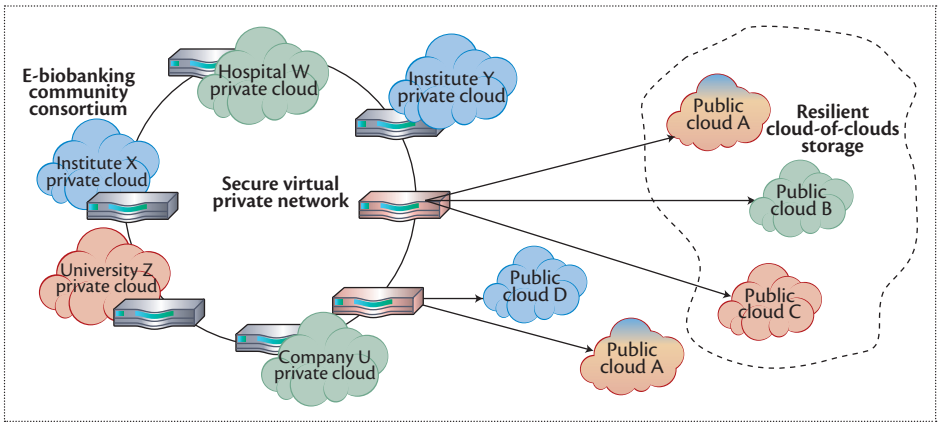

Figure 2. The BiobankCloud architecture. Multiple instances of clouds will participate in creating the abstraction of a single cloud-based key-value store or file system.

administrators will perceive this system as a single seamless cloud. However, underneath, powerful encryption and fault-tolerance mechanisms will ensure that the components interoperate so that no part is a single point of failure due to accidental or malicious threats.

On the other hand, the architecture's modularity will make it easy to set up secure, dependable constellations of private and public clouds belonging to diverse stakeholders, with separation of risk and concerns. Hopefully, this technology will help the e-biobanking vision come true, enabling the creation of ecosystems involving coalitions or consortia of hospitals, biomedical or bioinformatics research institutions, nonprofit organizations for biobanking and genomics research support, and even NGS vendors' own infrastructures. Occasional and power users will be able to easily use the infrastructure in an integrated way, through comfortable platform-as-a-service interfaces. Researchers will be able to set up experiments to automatically share data belonging to several realms, improving the experiments' throughput and turnaround.

However, security and dependability mustn't be compromised. So, data with different criticality levels will reside in different cloud subsets with adequate levels of protection, preserving liability and regulations. For example, privacy-critical data will never leave private clouds yet will be able to be processed by having authorized users ship adequate certified computing functions to those clouds.

Figure 3a suggests a typical example: University Z bioinformaticians performing operations on mixed-criticality data in public and private clouds. Supposing that critical data can't leave some private clouds, BiobankCloud achieves trust by having the job dispatcher ship certified functions to those clouds, running them, and sending the results to the user.

In addition, BiobankCloud technology will significantly leverage the availability of a competitive market of public clouds to optimize e-biobanking costs, initially using them "as is." This will give large communities of users direct access to public data—for example, data anonymized with algorithms the project is developing to overcome vulnerabilities leading to reidentification.

BiobankCloud will also let generic public clouds be part of resilient, highly secure back-end cloud-of-clouds storage. This will address the quickly growing need for storing critical data cost-effectively—that is, complementing private clouds. To achieve this objective, the project will use state-of-the art encryption, coding, and dispersion mechanisms.[10]

Figure 3b illustrates the basic method in a simple way. Data, while still in a trusted environment (the private cloud), is allocated a key and encrypted. Next, *erasure coding* produces m coded chunks from the initial data, with enough redundancy so that the entire data can be recovered from just a few chunks. Then, *secret sharing*, a cryptographic operation, "breaks" the key into several shares, m in this case. This operation ensures that the key can be recovered from just, say, $k < m$ shares, while guaranteeing it can't be recovered from fewer than k shares.

This method resists attacks and failures related to the data's integrity and the key's confidentiality, as long as their severity (the number of chunks or shares affected) isn't higher than a predefined protection threshold. Let's call this threshold f. In Figure 3b, $f = 1$, and recovering $f + 1$ (two chunks and two key shares) is enough to restore the data.

On the other hand, to break privacy, an attacker would have to simultaneously hack into two clouds. Likewise, for the data to become irrecoverable, the attacker would have to compromise all clouds but one and destroy the respective chunks. Besides the redundancy introduced by multiple clouds, their diversity reduces failure probability and further increases the attacker's effort.

Despite this example's simplicity, the scheme is parametric. The threshold f can take any value designers choose, letting them attain arbitrarily high levels of resilience and security, at the cost of increased redundancy.

It will be interesting to see, in the near future, how the several conflicting goals of e-biobanking will be reconciled. On one hand, pressure exists to sequence as many

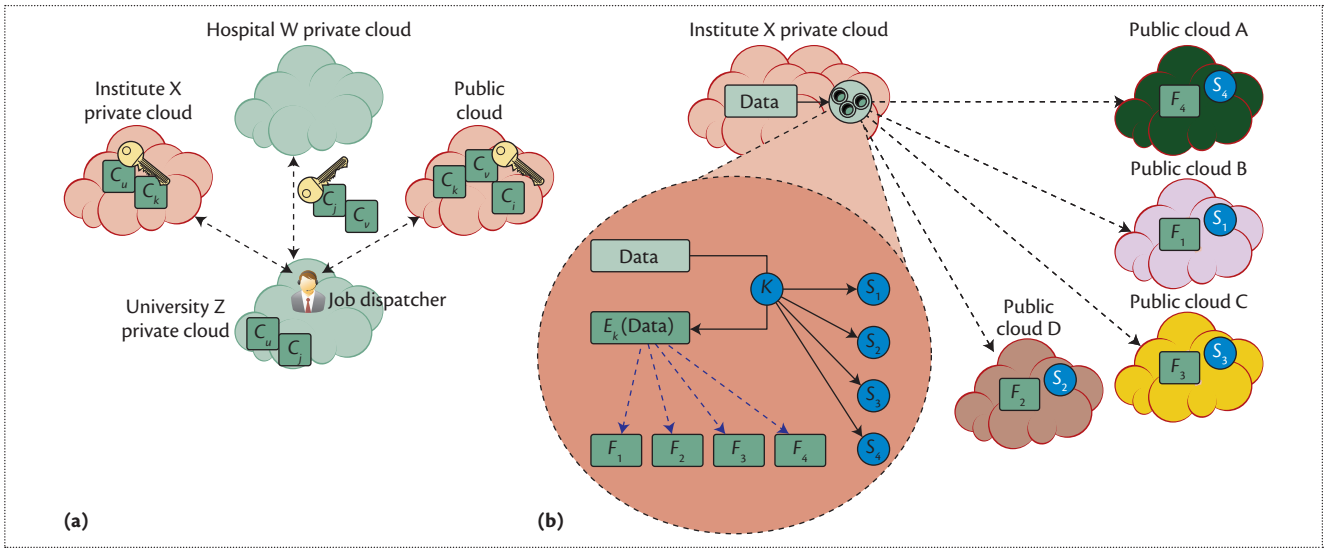

Figure 3. Ensuring security and dependability. (a) Certified computations (for example, MapReduce). Here, bioinformaticians at University Z perform operations on mixed-criticality data in public and private clouds. (b) Resilient cloud-of-clouds storage. This method resists attacks and failures as long as their severity isn't higher than a predefined protection threshold. *C* is a certified function, *K* is the key, E_k(Data) is the encryption of data with *K*, *S* is a share of *K*, and *F* is an erasure-coded data chunk.

genomes as possible and to make a large part of that information available, cost-effectively. On the other hand, the need exists to safeguard both that information and the lives of the people who supplied it. Technologies such as those being developed by BiobankCloud and similar research projects might be helpful, in not only directly securing the latter goals but also facilitating the pursuit of the former ones. This would be achieved by creating a migration path for the secure, dependable use of public clouds to store and process critical data. ■

Acknowledgments

Our BiobankCloud research is funded by the European Commission under grant FP7-ICT-317871 and by the Fundação para a Ciência e a Tecnologia (FCT), through the Multiannual Funding Programme. We warmly acknowledge Francisco Couto and Ulf Leser's comments and input.

References

1. "What Is a Biobank?," Nat'l Biobank Program, 2003; www.biobanks.se/biobank.htm.
2. "DNA Sequencing Costs," US Nat'l Human Genome Research Inst., 2013; www.genome.gov/sequencingcosts.
3. L.D. Stein, "The Case for Cloud Computing in Genome Informatics," *Genome Biology*, vol. 11, no. 5, 2010.
4. C. Bryant, "European Data Protection under a Cloud," *Financial Times*, 28 July 2013; www.ft.com/cms/s/0/dbee868a-f43c-11e2-8459-00144feabdc0.html#axzz2gyIqMlYl.
5. P. Verissimo, A. Bessani, and M. Pasin, "The TClouds Architecture: Open and Resilient Cloud-of-Clouds Computing," *Proc. 2012 IEEE/IFIP 42nd Int'l Conf. Dependable Systems and Networks Workshops*, IEEE, 2012.
6. P. Verissimo and L. Rodrigues, *Distributed Systems for System Architects*, Kluwer Academic, 2001, pp. 377–394.
7. P. Ohm, "Broken Promises of Privacy: Responding to the Surprising Failure of Anonymization," *UCLA Law Rev.*, vol. 57, 2010, p. 1701; http://ssrn.com/abstract=1450006.
8. M. Gymrek et al., "Identifying Personal Genomes by Surname Inference," *Science*, vol. 339, no. 6117, 2013, pp. 321–324.
9. M. Goodman and A. Hessel, "The Bio-Crime Prophecy: DNA Hacking the Biggest Opportunity since Cyber-Attacks," *Wired.co.uk*, 28 May 2013; www.wired.co.uk/magazine/archive/2013/06/feature-bio-crime/the-bio-crime-prophecy.
10. A. Bessani et al., "DepSky: Dependable and Secure Storage in a Cloud-of-Clouds," *Proc. 6th ACM SIGOPS/EuroSys European Systems Conf.* (EuroSys 11), ACM, 2011, pp. 31–46.

Paulo Esteves Verissimo is a professor at the University of Lisbon's Department of Computer Science and Engineering. Contact him at pjv@di.fc.ul.pt.

Alysson Bessani is an assistant professor at the University of Lisbon's Department of Computer Science and Engineering. Contact him at bessani@di.fc.ul.pt.