Data security in genomics: A review of Australian privacy requirements and their relation to cryptography in data storage

Arran Schlosberg¹,²

¹Department of Medical Genomics, Royal Prince Alfred Hospital, Camperdown, NSW 2050, ²Central Clinical School, Sydney Medical School, The University of Sydney, NSW 2006, Australia

E-mail: *Dr. Arran Schlosberg - arran.schlosberg@sydney.edu.au
*Corresponding author

Received: 11 August 2015 Accepted: 06 October 2015 Published: 05 February 2016

Abstract

The advent of next-generation sequencing (NGS) brings with it a need to manage large volumes of patient data in a manner that is compliant with both privacy laws and long-term archival needs. Outside of the realm of genomics there is a need in the broader medical community to store data, and although radiology aside the volume may be less than that of NGS, the concepts discussed herein are similarly relevant. The relation of so-called “privacy principles” to data protection and cryptographic techniques is explored with regards to the archival and backup storage of health data in Australia, and an example implementation of secure management of genomic archives is proposed with regards to this relation. Readers are presented with sufficient detail to have informed discussions – when implementing laboratory data protocols – with experts in the fields.

Key words: Cryptography, genomics, privacy, security, storage

INTRODUCTION

The advent of next-generation sequencing (NGS) brings with it a need to manage large volumes of patient data in a manner that is compliant with both privacy laws and long-term archival needs. Raw sequencing data are processed through an informatics pipeline consisting of multiple algorithms such as alignment and variant calling. A 2011 comparison of common alignment algorithms¹ included six such approaches each of which can be implemented with subtle differences based upon specific software packages and furthermore allow for various configuration directives. These myriad approaches – with the potential for novel future additions – mean that long-term storage of raw instrument data is a prudent approach in order to allow for alternate analyses as guided by changes in best practice. Although National Pathology Accreditation Advisory Council requirements² outline a retention period of 3 years for “calculations and observations from which the result is derived,” jurisdiction-specific legislation³ extends this time frame. Outside of the realm of genomics there is a need in the broader medical community to store data, with radiological domains dealing with magnetic resonance imaging producing volumes comparable or even greater than those of NGS, and the concepts discussed herein are similarly relevant for any volume.
Archival-backup storage of sensitive, large volume data poses a number of technological and legal issues. Data must be maintained in a manner that provides access to those rightfully authorized to have such access while protected against disclosure to and tampering by others. Beyond the potential for malicious acts, there are also technological hurdles posed by data corruption and hardware failures. This need for data integrity may fall under the same legal purview as the need for security.

Although privacy legislation in Australia exists at the Commonwealth, State, and Territory levels, there is a common theme of so-called “privacy principles.” The Australian Privacy Principles (APPs) came into effect in March 2014, thus replacing the National and Information Privacy Principles (IPPs). Nomenclature regarding principles differs between the States with, for example, Victoria’s IPPs and New South Wales’ (NSW) Information Protection Principles which are complemented by the more stringent Health Privacy Principles (HPPs).

A layperson reading of the text of these principles reveals large sections of verbatim reproduction. Of particular note are clauses pertaining to the transfer of information between jurisdictions which prohibit such transfer unless – for example, under NSW’s HPP 14 – “the organisation reasonably believes that the recipient of the information is subject to a law, binding scheme or contract that effectively upholds principles for fair handling of the information that are substantially similar to the HPPs” Victoria’s IPP 9 and the Commonwealth’s APP 8 contain similar allowances which suggest that they may be mutually compatible. Whether or not the privacy requirements of the Health Insurance Portability and Accountability Act (USA) are “substantially similar” is beyond the scope of this article.

Utilizing the NSW HPPs as a benchmark, this article aims to frame the principles in light of practical implications for genomic laboratories. The choice of the NSW State-specific legislation was influenced by the jurisdiction in which I am employed, but, wherever possible, equivalent APPs are referenced.

As the diagnostic-genomics landscape is in its relative infancy regarding such practices, there is limited opportunity for peer benchmarking. Hence I have borrowed from other disciplines in much the same manner as operating theatres’ use of the WHO Surgical Safety Checklist was influenced by the aviation industry. Suggestions for adherence to principles are derived from recommendations by the Australian Signals Directorate (ASD) as they pertain to the protection of sensitive government information.

Regarding terminology, an archive is a moving of data away from a source of regular access, whereas a backup implements protections against the loss of data. Although an archive may be implemented in such a manner that it acts as a backup, it is important to note that a poorly-managed archive does not provide sufficient fault tolerance. However, I shall treat the creation of genomic archives as requiring such characteristics. Thus, for the sake of simplicity, I will use the words archive and backup interchangeably.

PRIVACY PRINCIPLES

The jurisdiction-specific sets of privacy principles vary in their size and scope. However, there are core elements that remain pertinent to NGS data storage, regardless of legislation, as they constitute prudent data management.

Retention and Security

As one would expect, the principles include provisions pertaining to the secure management of health data. There is a requirement (HPP 5, similar to APP 11) to implement “security safeguards as are reasonable [to protect] against loss, unauthorised access, use, modification or disclosure, and against all other misuse.” Interestingly these correlate well with broad domains of cryptography, which are briefly outlined in Table 1.

An additional requirement is that information is retained “no longer than is necessary.” Section 25 of the Health Records and Information Privacy Act (NSW) which defines the NSW HPPs, requires retention “for 7 years from the last occasion on which a health service was provided to the individual” or, in the event that the individual was under the age of 18 years at the time of collection, “until the individual has attained the age of 25 years.” Furthermore, the retention period may be subject to a court or tribunal order which may require that it not be destroyed nor rendered nonidentifiable. Even if this was not the case, given the current cost of procuring

### Table 1: The field of cryptography extends beyond the scope of what many readers may suspect. A selection of cryptographic domains and their respective focuses are outlined

| Domain         | Focus                                                                 |
|----------------|------------------------------------------------------------------------|
| Encryption     | Process of scrambling a message in such a way that only the intended recipient can obtain the original data through decryption |
| Authentication | Means of proving the integrity and authenticity of data; for example, by creating fingerprints for large-volume data |
| Key agreement  | Methods by which parties can decide upon a shared secret, such as an encryption “password” despite non-secure communications channels |
| Signatures     | Proviable attestations to the authorship of data                       |
| Password storage | Allowing for rapid verification, but not decryption nor brute-force inference in the event of exposure of encrypted values |
NGS data, re-sequencing is not economically feasible in the immediate future. With this in mind, we absolutely require a data retention plan rather than simply discarding information, and literature points to similar practice.\[11\]

**Accuracy**
The scope of the principles extends beyond a basic understanding of privacy, to include (HPP 9, similar to APP 10) a requirement that organizations holding health information “ensure that, having regard to the purpose which the information is proposed to be used, the information is relevant, accurate, up to date, complete, and not misleading.”\[12\] The rapidly-changing nature of bioinformatics algorithms is such that the relevancy and completeness of data are variable with time.

The advent of a novel algorithm – and the failure to implement its advances – may render yesterday’s “noise” as tomorrow’s misleading information. It remains to be seen how the true purpose of genomic information is defined; is it a point-in-time test, or does it extend to future reanalysis?

**Transfer of Data**
The existence of provisions, allowing for the transfer of data should specific criteria be met, opens the door to outsourced data storage. The NSW HPPs provide eight circumstances under which transfer is allowed, and their logical grouping by “or” conjunction suggests that only one such criterion need be met. Beyond the provision for transfer to a recipient bound by similar principles, one additional criterion is of note (similar to APP 8):

HPP 14(g): The organization has taken reasonable steps to ensure that the information that it has transferred will not be held, used, or disclosed by the recipient of the information inconsistently with the HPPs.\[13\]

The proper use of encryption, prior to transfer, achieves such a means by rendering information as nonsensical to the recipient – ideally indistinguishable from random noise, as explored in Chapter 3.3 of Ferguson et al.\[12\] According to the ASD, “encryption of data at rest can be used to reduce the physical storage and handling requirements of media or systems containing sensitive or classified information to an unclassified level.”\[13\]

Those managing genomic data are in a position whereby they are required to give proper consideration as to whether or not their practices constitute “reasonable steps.” A loss in confidentiality of genomic data can be considered as a very serious privacy breach, and it is thus prudent to place significant emphasis on their protection. Given that ASD recommendations pertain to information, the breach of which could result in “grave damage to the National Interest,”\[14\] it is left to the reader – and their lawyers – to decide whether they believe that compliance based on the protection of national secrets constitutes sufficient efforts when applied to genomics.

**RISK ANALYSIS**
Loss prevention – be it against technical malfunction or malicious intervention – requires a thorough risk analysis in order to balance the implications of an adverse event against the outlay for protection against it. A simple analytical framework can be borrowed from the financial concept of expected loss. A loss function\[15\] is a statistical function describing the relative probability of losses – for example, the cost to an insurer of a motor-vehicle accident – of varying sizes, and the expected value\[16\] is the mean outcome.

Each potential loss that we face in the storage of genomic data has an associated loss function. The cost may not be directly monetary, but it can be quantified by some means. Issues arise from this analysis: (i) we lack the historical data to make informed decisions as to the definition of the loss function, (ii) such losses are black swan\[17\] events that are improbable yet catastrophic, and (iii) we are undertaking an \(n = 1\) experiment with our data which renders mean values useless as we face all-or-nothing outcomes. Insurers rely on the size of their policy pool to spread financial risk across all policy holders – an approach that I argue is equivalent to outsourcing data storage to highly-redundant cloud vendors.

Given the limiting factors regarding the definition of the loss function, I will only focus on the risks themselves – they are broad in their definition, and readers are encouraged to undertake their own analyses as are relevant to their individual situations. Those, the understanding of which will shed light on the role of cryptography, are included here while additional concepts are in the supplementary material.

**Replication Error**
The process of long-term data handling involves a series of steps with multiple, redundant copies being created. Data transfer mechanisms will, generally, include checking procedures to ensure the integrity of copies, but further checks should be implemented as discussed in sections on data integrity and authenticity.

Given that a change in binary data as small as a single bit may corrupt the underlying meaning this cannot be dismissed as a negligible concern. Often data can be inferred from their context. For example, the binary representations of A, G, C, and T contain sufficient redundant information that the reversion of a single-bit error can be easily inferred. The letter A is represented as 1000001, whereas T is 0101000 – corrupt data of 0010100 are more likely to represent T with only the first bit changed. Encrypted data, however, are such that contextual information is deliberately eroded into random noise – it is computationally infeasible to find the error by brute-force means, and thus a minuscule error may corrupt an entire volume.
False Sense of Security
The science of cryptography is very difficult, and its practical uses – although marginally simpler – remain the domain of experts. Improper use of cryptographic tools amounts to placing a padlock on the gate despite said padlock being made of plastic; we gain the sense of security without any true protection which is an arguably worse scenario as users may behave in a less prudent manner with regards to other security measures.

Another important point to note is that “there is no guarantee or proof of security of an algorithm against presently unknown intrusion methods.”[13] The complex nature of cryptographic algorithms exposes them to weaknesses that are yet to be detected – the academic and security communities undertake rigorous analyses, but they do not know what they do not know. Worse yet is the deliberate inclusion of so-called back-door methods that allow access to data and may in some cases be mandated by law.[14] Such inclusions would entail the inclusion of measures allowing government agencies to decrypt data in a manner akin to tapping a phone line. The belief that a “door” will only allow law enforcement to enter, but will deter malicious adversaries is simply naïve.[19,21] Furthermore, the implications of historical laws limiting the American export of cryptographic tools have resulted in an inadvertent vulnerability that was discovered many years after the laws were no longer relevant.[22,25]

CLOUD STORAGE
Adequate backup procedures rely on the concept of redundancy – the inclusion of multiple levels of protection when perhaps one alone may suffice. The probability of all protections failing simultaneously is less than that of a single mechanism’s deficiency. Means by which such redundancy can be achieved are included in the supplementary material, but I argue that this is a domain that is best outsourced to vendors working at great scale.

Provided that protective layers fail independently of one another, greater redundancy results in greater loss mitigation, but how much is enough? An objective answer requires a level of historical evidence – to define a loss function – that is not available to most laboratories. Even with vendor-supplied failure data there remain site-specific protocols that are subject to failure due to human error.

Infrastructure as a service is the more formal terminology used to describe a subset of “cloud computing” which provides the capability to “provision processing, storage, networks, and other fundamental computing resources.”[24] Infrastructure-as-a-service vendors work at such a scale that they have access to reliable data[25] regarding their hardware architectures and implementation protocols.

Amazon and Google each quote a durability of 99,999,999,999% annually for their S3[26] and Nearline[27] product offerings, respectively. This amounts to the loss, in 1 year, of one data object in every hundred billion – replication across both, or more, platforms can further improve durability. Such objective quantification is beyond the realm of in-house data-recovery protocols. We are thus no longer subjecting our data-protection mechanisms to $n = 1$ experiments regarding loss probabilities. The introduction of scale redefines what was black swan events as being quantifiable and more readily predictable. It is for this reason that cloud storage should be strongly considered as the primary means for achieving quantitatively-assessed risk analyses and mitigation.

Durability may occasionally come at the cost of immediacy and price. Multiple, redundant copies increase the price, but storing data on media that are not actively attached to computers reduces the cost of electricity, as well as the number of required storage interfaces on the computers. This may delay access to data by minutes or hours (as storage media are connected), but given the archival requirements of long-term NGS storage this is not necessarily problematic.

Australian Signals Directorate Certified Cloud Services List
Under the auspices of the ASD, the Information Security Registered Assessors Program[28] undertakes in-depth auditing of cloud providers to “assess the implementation, appropriateness, and effectiveness of [their] system’s security controls.”[29] Successfully audited providers are included on the Certified Cloud Services List and at the time of writing these included specific services from Amazon Web Services, Macquarie Telecom, and Microsoft. Readers are advised to seek the most up to date list.[29]

Outsourcing the management of sensitive health data introduces a new set of concerns, the mitigation of which can be achieved with cryptographic tools.

FUNDAMENTALS OF CRYPTOGRAPHY
The ASD explicitly states that encryption of data at rest – as against during transfer – can be used to reduce the security requirements of storage media for classified information.[10] With this in mind, it is prudent that those making decisions regarding the handling of NGS data have at least a cursory understanding of cryptography, its uses, limitations, and common pitfalls. Cryptography extends beyond the realm of encryption (i.e. encoding data in a means inaccessible to all but the intended recipient); this is by no means intended as a complete treatment of the topic, and interested readers are directed to Ferguson et al.[12]
Although I am repeating an earlier sentiment, it is important to reiterate that improper use of cryptographic tools amounts to installing a plastic padlock on the gate – it looks secure and gives us a sense of protection, but deludes users into a false belief that they can be lax with regards to other protective measures. Even with correct usage it is important to remember that cryptography forms part of a wider framework of data security. There is no point in placing a (titanium) padlock on the gate if the key is left lying around or the windows are left open. General security measures are detailed by Cucoranu et al.,\textsuperscript{[30]} and other resources are included in the supplementary material.

Relation to Privacy Principles
A requirement of the NSW HPPs is protection against “loss, unauthorized access, use, modification, or disclosure” (HPP 5, similar to APP 11) of health information. Each of these is addressed by a particular cryptographic mitigation as described in Table 2.

Threat Analysis: Value and Ability
As with the need to perform a thorough risk analysis regarding data protection, a similar undertaking is relevant to cryptography, but the lens with which the risks are viewed is slightly different. Cryptographer parlance will often refer to an adversary, which is adopted herein.

One must consider both the value of the data being protected, as well as the capabilities (knowledge, resources, etc.) of the adversary. Value is relative, and hence must be considered from the perspectives of adversaries (value gained by access to data), as well as those protecting information (value lost due to a breach in privacy). Furthermore, the value of data compromise may, for an adversary, lie in the tarnishing of reputation rather than in anything intrinsic to the data themselves. With this relative value in mind, we can then consider the extent to which we are prepared to protect our information relative to the combined efforts and capabilities of an adversary.

As an example, financial data hold inherent value that is quantitatively similar for both parties. Genomic data – particularly that without explicit personal identifiers – will likely have a different relative value in that it offers less to an adversary until they can (i) link the data to an individual and (ii) determine a means by which to benefit from the data. This “data reward” will influence the level of resources that an adversary is willing to direct toward unauthorized access to data and thus influence the level of protection that must be instated.

Those protecting data are in a position whereby they must protect all facets of their implementation while an adversary need only find a single vulnerability. Despite all efforts, new security vulnerabilities\textsuperscript{[22,31]} are discovered on a regular basis – an environment which favors the adversary. However, unless an adversary has a reason to target a particular laboratory’s data, it stands to reason that they will preferentially concentrate on a relatively weaker target which offers an equivalent reward for reduced effort. Thus, without any absolute surety regarding security, we can only hope to make access to our protected data relatively more difficult than access to others’.

Kerckhoffs’ Principle
Kerckhoffs’ principle\textsuperscript{[5]} states that: “The security of the encryption scheme must depend only on the secrecy of the [encryption password]…and not on the secrecy of the algorithm”\textsuperscript{[12]} (Paragraph 2.1.1). The interoperability of systems requires common protocols – with every sharing of a protocol with a trusted party there is an increased chance of its being learnt by an adversary. Additionally, publicly-available methods have been heavily scrutinized by experts. Thus, one should not equate the secrecy of a protocol or algorithm with its security.

Basic Terminology
Primitives
A cryptographic primitive is a basic building block of the higher-level cryptographic algorithms, including hash and encryption functions.

Keys
A cryptographic key can be loosely considered as the password provided to a cryptographic primitive in order to perform its task. I say loosely in that the analogy breaks down in certain circumstances, but, for the most part, it is a valuable means by which to understand the concept. Unless specifically stated to the contrary a key should be kept secret, and treated in the same manner as a password; note Box 1.

Randomness
Much of cryptography is focused on the concept of randomness – in contrast to deterministic systems such as the computers that implement cryptographic systems.
The generation of keys is entirely reliant on a source of randomness known as entropy.[34] There is a little point in generating a key in a deterministic manner such that an adversary can repeat the process. A distinction is made between truly random data (e.g., from natural sources such as radioisotope decay) and pseudo-random data which has the statistical appearance of randomness despite its deterministic origins. A (pseudo-) random number generator or (P) RNG is used to provide random input for these needs, and the reader should be aware of the existence of a cryptographically secure PRNG as against its regular counterpart which cannot be used securely due to issues of predictability.[34]

**Data Integrity: Hash Functions as Fingerprints**

With every copy of data that we produce, we introduce a new, potentially weak link in the chain of data protection. A corruption in one copy may propagate through to derivative copies, and we require a means of efficiently checking for data integrity. The simplistic approach of directly comparing two copies has a downfall in that it requires each of them to be present on the same computer. Transferring hundreds of gigabytes of data is both inefficient and itself error-prone.

The concept of digital fingerprints allows for such comparisons and must satisfy certain ideal properties in order to be of use in this scenario:

- **F1** – Fingerprints must be small enough to allow for efficient transfer over a network while ensuring integrity.
- **F2** – The same input data must always result in the output of the same fingerprint.
- **F3** – Different input data must result in the output of different fingerprints.

Close scrutiny of these criteria reveals that they are not mutually consistent. Given that the input data of a fingerprinting mechanism are of unlimited size, for any size fingerprint that is smaller than the input (F1) there must be more than one possible set of original data from which it can be derived (violating F3). This follows from the **pigeon-hole principle**: if we have more pigeons than we do pigeon holes and each pigeon must be placed in a hole then at least one such hole must contain more than one pigeon. Each possible fingerprint can be considered as a pigeon hole, and each possible input a pigeon. A formal treatment of this concept is known as the **Dirichlet box principle**.[35]

F3 is thus relaxed such that the probability of two disparate inputs resulting in the same fingerprints – known as a collision – is minimized. This is achieved through the **avalanche effect**[36] which, in its strictest form, states that “each [fingerprint] output bit should change with a probability of one-half whenever a single input bit is” changed.[37] Thus, the smallest possible discrepancy in copies of data will result in vastly different fingerprints as demonstrated in Table 3.

Cryptographic hash functions (simply hashes from here onward) act as generators of such fingerprints. They should not be confused with their noncryptographic counterparts which lack certain key properties. Beyond the aforementioned properties, cryptographic hashes are such that it is infeasible to:

- **C1** – Determine the input data given the output fingerprint
- **C2** – Determine a different set of input data that will result in the same output fingerprint.

C1 allows for the proof of the contents of data without revealing the contents itself, and C2 protects against the substitution of input data. Note that it is not always sufficient to simply calculate the hash of data in order to achieve authentication as an adversary can easily generate a hash of the data with which they have tampered.

These properties describe the ideal hash function, but their realization is limited by the fact that undiscovered and undisclosed vulnerabilities may exist which compromise (to some extent) the degree with which a particular function meets criteria. It is thus important to have knowledge of which functions are still considered secure. Real-world functions are often named as abbreviations of noninformative names, such as MD (message digest; a digest of data which has the statistical appearance of randomness known as entropy) and secure hashing algorithm (SHA) – each generation of functions is suffixed with a number. The previously-utilized MD5 is now considered “weak,”[38] and so too is SHA-1 with the ASD recommending SHA-2 in its place[13] as a part of a set of approved algorithms called Suite B. There is no need for laboratory data managers to have an intimate understanding of the current state of cryptographic advances suffice to have an appreciation of the ever-changing landscape.

### Table 3: Two very similar genetic regions, with only a single-nucleotide difference, have vastly different fingerprints generated by the SHA512 algorithm—only the first 16 bytes are shown, in hexadecimal notation. This is due to the avalanche effect.[34]. The strict avalanche criterion[37] is met when changing a single bit in input data results in a 50% probability for the change of each output bit, independent of all other changes in output

| Input                                      | Fingerprint    |
|--------------------------------------------|----------------|
| TTTTCTAAATCTACGTATCGCGTAAAACAACG           | d4edcd25bb550d1bdc4c7855a7c4061a |
| TTTTCTAAATCTCCGTATCGCGTAAAACAAACG          | 7aa7b09ddf0d3835393f30c3ca3c43fd |
**Data Authentication**

In certain cases it may be most practical to store the fingerprint alongside the data themselves. For example one may simply wish to compare data to a canonical fingerprint from a specific point in time, such as immediately postsequencing. Although this may alert the user to accidental changes in data, it remains vulnerable to malicious changes in that an adversary – aware, under Kerckhoffs’ principle, of the utilized hash function – can simply replace the fingerprint with that of their altered data.

A specific construct, known as a keyed-hash message authentication code (commonly referred to by its abbreviation, HMAC), combines the data with a (secret) key in order to prevent such malicious changes. For reasons beyond the scope of this article (Paragraphs 5.3.1 and 5.3.2 of Ferguson et al.\[12\]) it does not suffice to hash a simple concatenation of the key and the data, and the HMAC approach is preferred\[10\] as only those with knowledge of the key are able to compute a new (or even check an existing) fingerprint. We have thus achieved compliance with protection against loss and unauthorized modification of health data by verifying data integrity and authenticity, respectively.

**Encryption**

Most people will automatically think of encryption when considering the broader field of cryptography. Encryption (and its counterpart, decryption) are the means by which data (commonly referred to as a message in such circumstances) are “scrambled” (encrypted) in a manner whereby only those with the appropriate key are able to “unscramble” (decrypt) it and access the original message.

The original message is commonly referred to as plaintext while its encrypted counterpart is the ciphertext as the plaintext has been processed by a cipher. One categorization of ciphers is as stream or block based upon whether, respectively, the function processes the plaintext bit-by-bit or in larger blocks. A knowledge of the existence of this separation is all that is required for our purposes as a more informative categorization exists based upon the types of keys in use.

As with hash functions, it is not necessary for the laboratory data manager to have a thorough understanding of encryption beyond an appreciation of its general principles.

**Symmetric**

Ciphers that utilize the same key for both encryption and decryption are known as symmetric. They are computationally efficient\[12\] (Paragraph 2.3) in that they are able to process large volumes of data in a relatively fast manner – this is clearly important for NGS data. There is a drawback in that all parties need to have a prenegotiated key, and keeping said key secret becomes more difficult with each additional party that is privy to its content.

Given a theoretically-ideal cipher, the security is linked to the size of the key. The ASD recommends the use of the Advanced Encryption Standard specification (again as a part of Suite B, and commonly referred to by its abbreviation, AES) in selecting a symmetric cipher and allows for the protection of “TOP SECRET” information with a 256-bit key.\[11\] The requirement to use smaller keys is generally secondary to a constraint in computational resources and is likely a moot point within the laboratory. The need for larger keys is (i) not an option with AES and (ii) unnecessary given the laws of thermodynamics.\[40\]

The use of AES – as a block cipher – requires particular configuration regarding the manner in which each block of data is processed. This is known as a mode of operation – designated as a three-letter suffix – and each mode differs with respect to its provision of both data encryption and even authentication (e.g., AES-GCM\[41\]).

Someone experienced in the use of cryptography should be consulted upon making such a decision. However, it is of note that the electronic codebook mode, with AES, can only be (somewhat) safely used on data smaller than 128 bits (16 bytes) which renders it useless in genomics. Even with such small data it fails to protect the fact that two plaintexts are identical – the ASD forbids its use entirely.\[13\]

**Asymmetric**

Otherwise known as public-key encryption, this involves two distinct (but related) keys – one public, one private, and together known as a key pair. Ignoring their specific mathematical constructs, it suffices to understand their relationship. The private key, kept secret, is used to derive the public key, shared with anyone (including adversaries). Conversely, it is so computationally expensive to determine the private key of a given public key that is considered intractable to the extent that we rely on this difficulty for security. Of the cryptographic keys that I describe, the public key is the only one that does not need to be kept secret.

“Application” of either key to a message is such that it can be reversed only by the key’s counterpart. For the sake of encryption, we thus apply the public key to the plaintext which produces ciphertext that can only be decrypted by the owner of the private key. We are thus able to send a secret message to a specific recipient without – unlike with the symmetric approach – any secret information already shared between the parties involved.

**Data Authentication Revisited: Signatures**

When considering both encryption and authentication we can think of AES and HMAC as counterparts in that they require all parties to have knowledge of a secret key. The reversible means by which asymmetric algorithms...
such as RSA\cite{ref2} (named after its authors Rivest, Shamir, and Adleman) “apply” keys to data make them their own counterparts. By utilizing the private key rather than the public one, we can create a digital signature that allows the owner of the private key to lay claim to a particular message. Everyone else can apply the public key to the signature, compare the outcome to the message, and thus verify the author’s intentions – note that we do not necessarily verify the author as their private key may have fallen into the hands of an adversary. The ASD allows for the use of RSA for both encryption and creation of digital signatures.\cite{ref13}

**Public Key Infrastructure**

Public keys have an inherent problem in that there is no implicit mechanism to verify ownership. An adversary may nefariously publish a public key, claiming that it belongs to someone else, and thus intercept communications. In the scenario whereby they forward the message to the intended recipient – utilizing the actual public key – they have performed what is known as a man-in-the-middle attack\cite{ref12} [Figure 11.2 as applied to a different asymmetric algorithm].

In a scenario involving very few parties who have a secure means of sharing public keys (perhaps in person), there is no issue, but as the number of parties grows (e.g., the Internet) this solution fails. A public key infrastructure utilizes trusted third parties who will independently verify the ownership of public keys and will then attest to that ownership through an asymmetric digital signature of a message akin to “entity A owns public key X” – this attestation is commonly known as a certificate. The public keys of the third parties are then delivered (e.g., built in to browsers) to all parties who can then verify the authenticity of certificates.

This trust model forms the basis of a large proportion of the security of the world wide web despite major shortcomings. Although each individual may choose which third parties to trust, the average user lacks the knowledge to make an informed decision. Any one of the “trusted” third parties may have their private key compromised or, as has already been the case, purposefully misused to create a man-in-the-middle scenario by attesting to false ownership of a public key.\cite{ref41} See Chapters 18-9 of Ferguson et al.\cite{ref12} for a thorough treatment of public key infrastructure.

**Digital Envelopes**

Symmetric algorithms are beneficial in the genomic domain as they are efficient; they can process very large volumes of data in shorter periods of time than can their asymmetric counterparts. Conversely, they lack the benefit of parties not having to preshare a key.

The benefits of both approaches can be combined in what is known as a digital envelope\cite{ref44} whereby a symmetric key is “wrapped” in a public key. No predetermined sharing of secrets is required, and computational efficiency is maintained. Only the intended recipient – the owner of the private key – can unwrap the symmetric key in order to decrypt the message.

**PROPOSED IMPLEMENTATION**

This example allows for a consolidated view, represented schematically in Figure 1, of how a laboratory might implement protections for their data. The risks faced by an individual laboratory should be considered before any such implementation.

While we may think of encrypted communications as occurring between two different, geographically-separated parties, a similar idea is applicable to a single party separated in time. The message sender is analogous to the present time while the future party takes the role of the recipient.

Despite all efforts to protect sensitive data, vulnerabilities remain, and the storage of secret keys is a challenging problem. Specialist hardware security modules exist for this task, but their use is beyond the scope of this document. As genomic archives are rarely accessed we have a great advantage in that encryption keys do not have to remain readily accessible. Cold or off-line storage involves the use of media that are not accessible to a computer, and hence not vulnerable to remote access. Taken to the Nth degree data are stored on hard-copy media, and then protected by physical means. Considering security as a weakest-link problem, any hard copy stored under the same physical-security measures as laboratory instruments provides at least the same level of data protection.

A key pair should be generated for one of the asymmetric algorithms approved by the ASD for use in the agreement of encryption session keys – see Information Security Manual’s\cite{ref15} ASD Approved Cryptographic Algorithms. In the absence of hardware-security-module protection, the private key can then be stored in hard-copy prior to electronic deletion. The use of a QR Code (a 2D barcode with inbuilt error correction\cite{ref45}) allows for transfer back to trusted electronic devices, but a base-64\cite{ref46} human-readable copy should also be printed. As with redundant data backup, the same applies to the paper medium.

Utilizing the public key for digital envelopes, all future archival can be achieved through symmetric encryption of the particular NGS run – each with a newly derived symmetric key that is stored in an envelope. The ephemeral nature of this key (it exists in a plaintext form for only as long as it takes to encrypt the archive) adds a level of security such that an adversary would have to compromise the computer at the exact point of encryption (and hence have access to the raw NGS data anyway).
At this point, we have only implemented the encryption mitigation as outlined in Table 2. The HMAC of the archive’s ciphertext should now be computed, but a different key than that used for encryption should be generated and placed in a digital envelope. This allows us to be in compliance with both the loss- and modification-prevention requirements of the privacy principles. Note that the use of an authenticated mode of operation for our block cipher negates the need for the HMAC, but I am unaware of a simple means by which to achieve this with command-line utilities (see supplementary material notes on implementation).

In the event of a data-recovery scenario, the process is delayed somewhat by the need to convert the decryption key into an electronic format, but this is likely to be considered a worthwhile sacrifice in lieu of the added security. It is my opinion that we have now, in keeping with ASD recommendations, undertaken reasonable steps – if not overly conservative ones – to protect our data for transfer to a third party. In keeping with a defense-in-depth approach, still use a provider from the Certified Cloud Services List.[29] This forms a “trapdoor-like” protocol whereby encryption of data is easy, but decryption requires physical access to a private key which is protected to at least the same extent as laboratory equipment.

### DISCUSSION

The use of cryptography is complex and difficult. Even with a thorough theoretical knowledge, obscure practical
threats known as side-channel attacks exist, and these can be as precise as measuring the timing of a computer’s response to the comparison of unequal hash values. In an ever-changing security environment that is riddled with nuanced problems, it remains a prudent decision to consult with an expert in the field of data security. Beyond this consultation, the use of a third-party auditor/expert should also be considered.

Medical data of all forms need to be kept for extended periods of time, and the future of advances in security threats is unknown. What is considered best practice today may, within the required retention period, become vulnerable to unauthorized access. Much like any quality-assurance efforts undertaken in the laboratory, protective frameworks should be regularly reviewed in light of up-to-date knowledge, and so too should data-recovery processes be routinely checked – any fault in backup mechanisms should be detected as early as possible so as to minimize the time frame during which we are exposed to complete data loss. The practical implications of such an undertaking will likely be beyond the scope of most laboratories, and outsourcing may be a viable alternative. Third parties may be contracted in this regard, but, to the best of my knowledge, no such solution exists within Australia – consideration may need to be given with regards to building capacity in this domain should it not be filled commercially.

Transferring already-encrypted data to third parties negates their ability to perform any meaningful task beyond that of storage. This precludes the use of cloud-based analytical platforms for which high-assurance mitigations against misuse or disclosure are not as easy to implement, and a level of trust in the third party is required. Homomorphic encryption whereby calculations can be performed without decryption of data is, in the field of genomics, very much in its infancy.\[47]\n
In a world of digital mistrust, it is difficult to make confident decisions with regards to information sources. The Suite B recommendations in the ASD’s Information Security Manual\[35] were borne from decisions made by the USA’s National Security Agency, which in light of revelations brought forth by Edward Snowden (archive\[85\]) have a questionable reputation within the broader security community. Leaked internal documents confirmed that they engaged “US and foreign IT industries to covertly influence and/or overtly leverage their commercial product’s designs” with an aim to “insert vulnerabilities into commercial encryption systems, IT systems, networks, and endpoint communications devices” (from an original document as archived).\[49]\n
It is, however, perhaps wise to frame these concerns in light of our objectives in protecting patient data. With both the Australian and USA Governments recommending the use of such algorithms, it is reasonable to believe that any party capable of undermining their security (the National Security Agency included) will have the highest level of resources at their disposal. If considering the relative-value model proposed for determining the extent of security, it is likely that the expenditure of such vast resources would far outweigh the value of our data to such an entity. Furthermore, it is unlikely that encryption will be the weakest link in the chain – an adversary wishing to gain access to our data would face a reduced barrier by instead compromising the source.

From a privacy law perspective, we have sought to take reasonable steps – in following the technological recommendations\[13] of our own government – to adhere to the privacy laws outlined in this paper, and within an ethical framework we have made the decision, to trust these recommendations, in good faith.

**CONCLUSION**

Designations given for secure classification of Australian documents\[14] are such that they represent information for which compromise may result in anything from “damage to… organizations or individuals” (PROTECTED) to “grave damage to the National Interest” (TOP SECRET). Although specific products implementing the ASD recommendations must undergo an evaluation\[19] prior to use in governmental settings, this does not preclude us from utilizing industry-standard implementations in the medical-testing laboratory. In this light, adherence to protections for classified information can hopefully be considered as sufficient for having taken “reasonable steps” in the protection of genomic data.

A completely in-house process for the management of redundant backups cannot be quantified with regards to risks in the same manner as one expects from a cloud vendor. It is thus prudent to consider the outsourcing of such core informatics undertakings. Cloud vendors focus their time on securing their systems, whereas data security is, unfortunately, a secondary endeavor for diagnostic laboratories and hospitals in general. It is my belief that we face greater risks from nonmalicious, accidental losses occurring in-house than from state-sponsored adversaries capable of compromising best-practice cryptographic techniques. However, as with all aspects of this article, the reader is advised to consider their individual situation. The role of computationally-oriented staff in the NGS-focused laboratory can be separated into two distinct categories which are often confused. The bioinformatician deals with the statistical and computational analyses of biological data, whereas the health informatician is tasked with the management – including security – of data
of all types. As Australian genomics laboratories focus more heavily on bioinformatic endeavors, it is important that they too consider these additional roles which fall outside the scope of the bioinformatician but are of key importance in clinical settings.

Acknowledgments
Thank you to Ronald J Trent of the Department of Medical Genomics at Royal Prince Alfred Hospital for his input regarding pertinent content regarding laboratory management. Schematic icons produced by Designer Base, Icomoon, Freepik, SimpleIcon, and Yannick from www.flaticon.com.

Financial Support and Sponsorship
Nil.

Conflicts of Interest
The author is a commercial consultant in the area of data management, including both bioinformatics and health informatics, as well as data security. The author holds no legal qualifications and the contents herein should not be construed as legal advice. The purpose of this document is to provide the reader with an understanding of how technological tools apply to the privacy environment. The proposed implementation acts as an example only, and the specific needs of the individual laboratories should be considered, including seeking legal advice and/or the assistance of experts in the fields of cryptography and data security.

REFERENCES

1. Ruffalo M, LaFramboise T, Koyutürk M. Comparative analysis of algorithms for next-generation sequencing read alignment. Bioinformatics 2011;27:2790-6.
2. National Pathology Accreditation Advisory Council. Requirements for the Retention of Laboratory Records and Diagnostic Material. 6th ed. Canberra, Australia: National Pathology Accreditation Advisory Council; 2005.
3. Health Records and Information Privacy Act (NSW, Australia); 2002.
4. Privacy Amendment (Enhancing Privacy Protection) Act (Commonwealth of Australia); 2012.
5. Information Privacy Act (VIC, Australia) ; 2000.
6. Privacy and Personal Information Protection Act (NSW, Australia); 1998.
7. Health Insurance Portability and Accountability Act (USA); 1996.
8. World Health Organization. WHO | WHO Surgical Safety Checklist and Implementation Manual; 2008. Available from: http://www.who.int/patientsafety/safesurgery/ss_checklist/en/. [Last accessed on 2015 Aug 10].
9. Mahajan RP. The WHO surgical checklist. Best Pract Res Clin Anaesthesiol 2011;25:161-8.
10. Weiser TG, Haynes AB, Lashner A, Dziiek G, Boorman DJ, Berry WR, et al. Perspectives in quality: Designing the WHO surgical safety checklist. Int J Qual Health Care 2010;22:365-70.
11. Gullapalli RR, Desai KV, Santana-Santos L, Kant JA, Bencich MJ. Next generation sequencing in clinical medicine: Challenges and lessons for pathology and biomedical informatics. J Pathol Inform 2012;3:40.
12. Ferguson N, Schneier B, Kohno T. Cryptography Engineering: Design Principles and Practical Applications: Design Principles and Practical Applications. Indianapolis, IN:John Wiley and Sons; 2011.
13. Australian Signals Directorate. Australian Government Information Security Manual Controls; 2013. Available from: http://www.asd.gov.au/publications/Information_Security_Manual_2015_Controls.pdf. [Last accessed on 2015 Aug 08].

http://www.jpathinformatics.org/content/7/1/6
36. Feistel H. Cryptography and computer privacy. Sci Am. 1973;228:15‑23.
37. Webster AF, Tavares SE. On the design of S‑boxes in Advances in Cryptology–CRYPTO’85 Proceedings. Berlin: Springer-Verlag; 1986. p. 523‑34.
38. Wang X, Yu H. In Advances in Cryptology–EUROCRYPT. Berlin: Springer; 2005. p. 19‑35.
39. Bellare M, Canetti R, Krawczyk H. Keying hash functions for message authentication in Advances in Cryptology – CRYPTO’96. Berlin: Springer; 1996. p. 1‑15.
40. Schneider B. Applied Cryptography: Protocols, Algorithms, and Source Code in C. Indianapolis, IN: John Wiley and Sons; 1996. p. 157‑8.
41. McGrew DA, Viega J. In Progress in Cryptology – INDOCRYPT 2004. Berlin: Springer; 2005. p. 343‑55.
42. Rivest RL, Shamir A, Adleman L. A Method for Obtaining Digital Signatures and Public‑key Cryptosystems. Communications of the ACM 21. New York, NY: Association for Computing Machinery; 1978. p. 120‑6.
43. Google Online Security Blog. Maintaining Digital Certificate Security; 2015. Available from: http://www.googleonlinesecurity.blogspot.com.au/2015/03/maintaining-digital-certificate-security. [Last accessed on 2015 Aug 10].
44. EMC Corporation. RSA Laboratories-2.2.4. What is a digital envelope? Available from: http://www.emc.com/emc-plus/rsa-labs/standards-initiatives/what-is-a-digital-envelope.htm. [Last accessed on 2015 Aug 10].
45. ISO, BS. IEC 18004: Information Technology-Automatic Identification and Data Capture Techniques- QR Code Bar Code Symbology Specification 2005.
46. Josefsson S. The Base16, Base32, and Base64 Data Encodings RFC 4648 (Proposed Standard). Internet Engineering Task Force; October, 2006. Available from: http://www.ietf.org/rfc/rfc4648.txt. [Last accessed on 2015 Aug 10].
47. Hayden EC. Extreme cryptography paves way to personalized medicine. Nature 2015;519:400.
48. The NSA Files. The Guardian; 2013. Available from: http://www.theguardian.com/us-news/the-nsa-files. [Last accessed on 2015 Aug 10].
49. Sigint–How the NSA Collaborates with Technology Companies. The Guardian; 2013. Available from: http://www.theguardian.com/world/interactive/2013/sep/05/sigint-nsa-collaborates-technology-companies. [Last accessed on 2015 Aug 10].
50. Australian Signals Directorate. EPL – Evaluated Products List: ASD Australian Signals Directorate. Available from: http://www.asd.gov.au/infsec/epl/. [Last accessed on 2015 Aug 10].