A novel, privacy-preserving cryptographic approach for sharing sequencing data

Christopher A Cassa,[1,2,3] Rachel A Miller,[3] Kenneth D Mandl[4,5,6]

▶ Additional supplementary files are published online only. To view these files please visit the journal online (http://dx.doi.org/10.1136/amiajnl-2012-001366).

[1]Division of Genetics, Brigham and Women's Hospital, Boston, Massachusetts, USA
[2]Division of Genetics, Harvard Medical School, Boston, Massachusetts, USA
[3]Computer Science and Artificial Intelligence Laboratory (CSAIL), Massachusetts Institute of Technology, Cambridge, Massachusetts, USA
[4]Children's Hospital Informatics Program at Harvard-MIT Health Sciences and Technology, Children's Hospital Boston, Boston, Massachusetts, USA
[5]Division of Pediatrics, Harvard Medical School, Boston, Massachusetts, USA
[6]Manton Center for Orphan Diseases, Children's Hospital Boston, Boston, Massachusetts, USA

Correspondence to
Dr Christopher A Cassa, Brigham and Women's Hospital Division of Genetics, 77 Ave Louis Pasteur, Boston, MA 02215, USA; cassa@alum.mit.edu

Received 25 September 2012
Accepted 4 October 2012
Published Online First 2 November 2012

ABSTRACT

Objective DNA samples are often processed and sequenced in facilities external to the point of collection. These samples are routinely labeled with patient identifiers or pseudonyms, allowing for potential linkage to identity and private clinical information if intercepted during transmission. We present a cryptographic scheme to securely transmit externally generated sequence data which does not require any patient identifiers, public key infrastructure, or the transmission of passwords.

Materials and methods This novel encryption scheme cryptographically protects participant sequence data using a shared secret key that is derived from a unique subset of an individual's genetic sequence. This scheme requires access to a subset of an individual's genetic sequence to acquire full access to the transmitted sequence data, which helps to prevent sample mismatch.

Results We validate that the proposed encryption scheme is robust to sequencing errors, population uniqueness, and sibling disambiguation, and provides sufficient cryptographic key space.

Discussion Access to a set of an individual's genotypes and a mutually agreed cryptographic seed is needed to unlock the full sequence, which provides additional sample authentication and authorization security. We present modest fixed and marginal costs to implement this transmission architecture.

Conclusions It is possible for genomics researchers who sequence participant samples externally to protect the transmission of sequence data using unique features of an individual's genetic sequence.

OBJECTIVE

Hundreds of thousands of individuals have contributed samples to biorepositories for translational research.[1–4] As these specimens are prepared, they are often labeled with patient or participant identifiers when sent to external facilities for sequencing, creating disclosure risk. Even a small amount of disclosed sequence data can be used to match an individual to additional clinical and genomic information from a public biorepository or other DNA database. Furthermore, labeling samples with names or patient identifiers creates the risk that these sequences may be linked to clinical information.

We describe a cryptographic architecture for specimen acquisition and processing that enables the complete removal of patient identifiers (such as medical record numbers (MRNs) and pseudonyms) from samples before external sequencing is conducted. This removes the possibility that sequence data may be linked with patient identifiers outside of the hospital. We seek to define an architecture that can:

1. leverage the use of existing information systems and sample acquisition processes whenever possible, including electronic medical record (EMR) and laboratory processing systems, which generally require the use of patient MRNs or pseudonyms;
2. allow samples to be sequenced externally, without the use of patient identifiers such as MRNs;
3. secure the transmission of genomic data returned from external sequencing laboratories; and
4. robustly re-link those data to the correct individual participant once they are returned.

While pseudonymization is a technique that has gained traction for the sharing of microdata that are not uniquely identifiable, it alone is not sufficient to satisfy these requirements. We outline an approach that does not require any identifiers to be transmitted with samples, which reduces re-identification risk. Furthermore, pseudonyms do not provide assurance that there has been no sample mismatch at the sequencing laboratory, and do not protect the transmission of sequence data (without cryptographic security, such as SSL).

BACKGROUND AND SIGNIFICANCE

Genomic sequencing is often conducted in commercial laboratories external to the location where samples are collected. Often, clinical samples are labeled with a patient's protected health information, including a name, MRN, or other identifiers.[5] While these identifiers enable an uncomplicated chain of custody for laboratory processing and permit rapid reconfirmation, this information poses a quantifiable risk to patient or participant privacy, with the potential to disclose identity, medical conditions, disease risk, and hereditary data. Sequencing laboratories have begun to electronically transmit and store WGS data in the cloud,[6 7] which creates the risk that these data may be viewed by an adversary. We present a novel encryption strategy to protect these transmitted data that does not require public key infrastructure or sample identifiers, and provides other benefits.

We describe three threat models for an adversary who acquires sequence data with an MRN or pseudonym: (1) ability to link to identity or clinical data using an MRN, (2) ability to link to identity using an available genomic dataset (including to pooled genomic data, or family members), and (3) ability to ascertain additional information about phenotype, lineage, and risk for an individual from sequence data.

1. An MRN or pseudonym theoretically enables linkage to clinical data that are recorded and

stored in hospital data systems. These data include information from clinical encounters, demographics, insurance and employment data, and laboratory results. While these data are safe when stored securely and used appropriately by researchers, there is the potential risk that sequence data and patient identifiers could be revealed to a third party while outside of the hospital.

2. Lin et al[8] demonstrated that privacy decreases sharply with the disclosure of a small number of SNP genotypes. In fact, with just 35–70 independent SNP genotypes, it is possible to uniquely identify any individual sequence or confirm that an individual is related with not many more.[9] Additionally, there is the equivalent of a DNA dictionary of individuals, in federal and state criminal DNA databases, which can be used to reliably link individuals and family members.[10 11] Furthermore, it is possible to use a genomic sequence to ascertain the presence of an individual in very large sets of pooled genomic data[12 13] or research articles.[14]
3. The derived genomic data may be used to reveal patient phenotypes, disease propensity, paternity, and lineage. Linkage to additional phenotypic data may be possible from a growing set of samples in biorepository databases. These databases have aggregated both genetic and phenotypic data from participants in large longitudinal clinical studies, including the Framingham and Jackson Heart Studies and the Women's Health Initiative.[15–20] Thousands of research participant records are now available through systems that include the Database of Genotypes and Phenotypes (dbGaP),[21] the NHLBI GO Exome Sequencing Project (ESP),[22] the European Genome-phenome Archive (EGA),[23] and NCBI ClinVar.[24]

We explore the collection, processing, and sequencing of DNA samples in the context of biorepository research, which increasingly includes genomic sequencing. While the sample acquisition process differs across institutions, there are key common steps, including consent, sample acquisition, sample processing, and long-term management (figure 1). When patients share their samples with medical researchers, there is a balance between the expectation of privacy and the need to maintain a robust chain of custody when processing samples. In this article, we define and review a de-identified architecture for DNA sample acquisition and processing which involves the removal of individual identifiers, and protects sequence data using shared secret key cryptography based on the genomic sequence. Additionally, our architecture ensures that only the correct individual's sequence is associated with a participant record, helping to prevent sample mismatch, a silent but serious error.

MATERIALS AND METHODS

We explore a use case of this scheme: a tertiary care setting where sample acquisition orders require the use of MRNs, a common requirement for specimen collection orders in many EMR systems for chain of custody with a research biorepository. The biorepository was required by the institutional review board (IRB) to remove MRNs from genomic samples before shipment to an external sequencing center. This system replaces MRNs with newly generated, randomized identifiers at each step of the sample acquisition and processing pipeline, to maintain a link between the de-identified sample and the patient's MRN. It also secures the electronic transfer of sequence data using a shared secret key that is generated using a subset of the genetic data, which also ensures robust sample re-identification as described in figure 2.

Full details of the de-identification and specimen acquisition processes are provided in the online supplementary material, so here we only summarize the most important steps in this protocol. We first generate a *sample acquisition ID*, which is a randomized identifier that joins the patient MRN during the sample acquisition, stored in a hospital biorepository server. When the sample is processed in a laboratory, we then remove the MRN and add a separate randomized *sample processing ID*. We store the relationship between the sample IDs and the MRN using a secret table in the biorepository server (see online supplementary material). A separate portion of this sample is used to create a local DNA fingerprint, which contains a set of genotypes. We then send the study subject's DNA to an external sequencing center, with only the sample processing ID (this sample processing ID is *not* required for the encryption scheme, however, but may be used if required), and when the results are returned, they arrive using this ID along with an encrypted full genome sequence, and are confirmed to the correct patient identity using the biorepository server.

Encryption of genome sequence data using a subset of genotypes

Genomic sequence data are protected health information that must be adequately protected when transmitted between external facilities and researchers. In this approach, we encrypt the sequence data we intend to transmit using features of each subject's genomic data (which are already naturally shared at both locations) to generate the necessary cryptographic keying material, the equivalent of a password-based shared secret key. We first generate a DNA fingerprint locally, which includes the genotypes of a small set of previously agreed SNP loci, which will serve as the basis of the shared secret key, and adjoin it to the optional randomized identifier. This identifier, alone, is then transmitted with the DNA specimen to an external facility, where the full sequence is generated. The same DNA fingerprint is then derived from the full sequence at the external facility (the same subset of SNP genotypes is selected). These DNA fingerprints are then used at both sites to create the same secret cryptographic key through a process that we describe

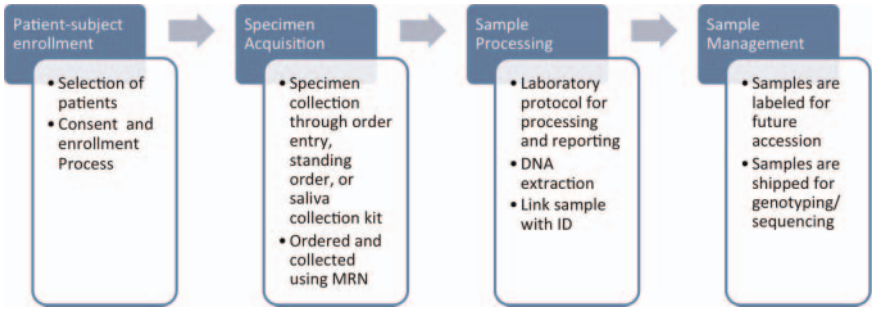

Figure 1 Common steps in the biorepository sample acquisition process. While these steps differ among institutions, many biorepository research projects include enrollment, sample acquisition, sample processing, and long-term management steps. MRN, medical record number.

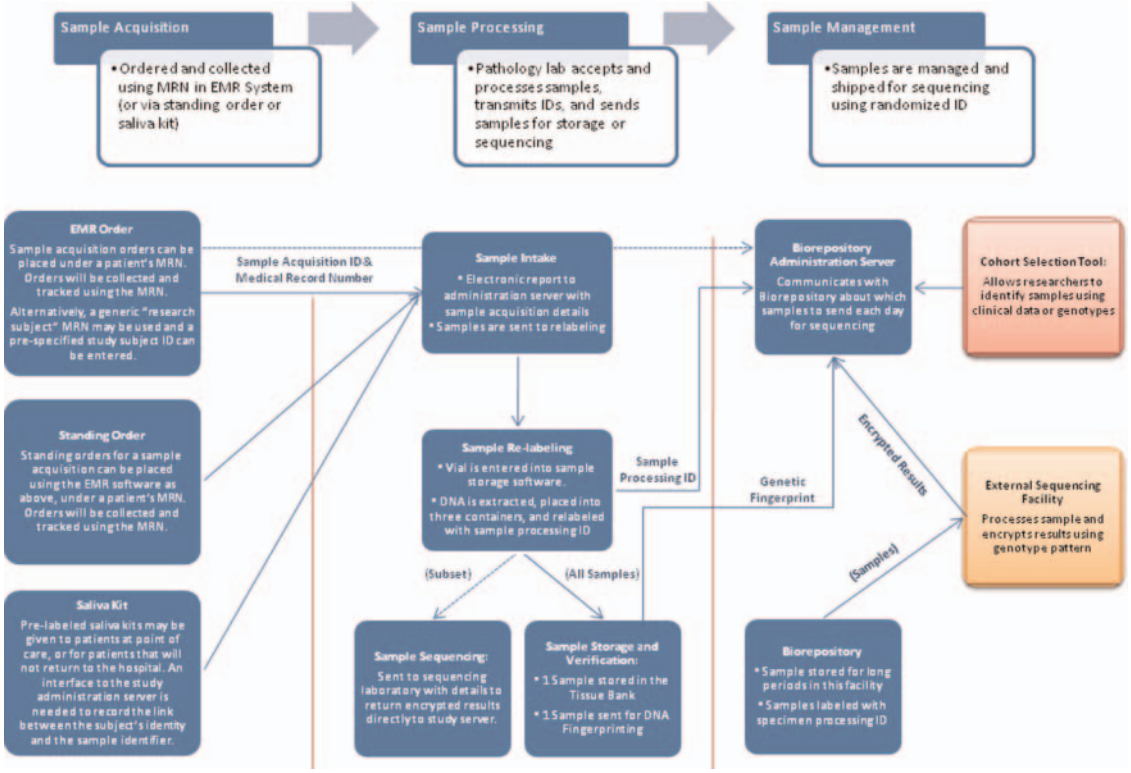

Figure 2 De-identified biorepository sample processing protocol. We describe one possible implementation of a de-identified hospital/biorepository architecture that would allow the use of MRNs to acquire and process samples locally, but would scrub all direct patient identifiers from samples before they are sent to external sequencing laboratories. The full description of this architecture is given in the online supplementary material. EMR, electronic medical record; MRN, medical record number.

separately. The full sequence is encrypted using this shared secret key, and transmitted to the original site with the randomized identifier, where it can be decrypted using the original DNA fingerprint.

We generate a unique, random cryptographic key for each genome so that both the sequencing laboratory and original facility can generate the same key without transmission of sequence data. To do this, we use: (1) a genetic fingerprint, encoded as a bit stream, (2) a randomness extractor to turn the bit stream representing the genetic fingerprint into a smaller truly random bit stream, (3) a (possibly publicly known) truly random data source for the randomness extractor, and (4) a symmetric key encryption algorithm. Optionally, we can include an error correcting code (ECC) to allow a small number of sequencing errors to be made by either party involved in the transmission. The cryptographic scheme is described in figure 3.

We more fully detail each component:

1. Genetic fingerprint. First, we define a standard, genetic fingerprint: a set of highly polymorphic reference SNPs that are commonly included on commercial genome sequencing platforms. Thus, our standard fingerprint will remain extractable from future, larger genomic sequencing and genotyping datasets. Moreover, this requires that for distinct genomes, there is an extremely high probability that this genetic fingerprint will be unique[8] and it is extremely unlikely that any party can guess the genetic fingerprint for any unknown genome, which we can ensure by selecting genotypes that have no known function and vary widely in the population. Based on data from the ESP (N=6500 individuals),[22] 18 468 SNPs in the population have a minor allele frequency of greater than 40%, so there are sufficient highly polymorphic SNPs in the exome such that a subset of these variants can be made to comprise a unique fingerprint. Additionally, this shared symmetric key may be derived from different sets of SNPs if the electronic key is accidentally disclosed. This set of genotypes must be specified between the two parties involved in the data transmission. Each genotype in this fingerprint will be converted into four binary digits, to represent the $2^4=16$ possible combinations of two A/C/G/T alleles at a single position, ordered alphabetically. Using Shannon's entropy, $H(X) = -\sum_{i=1}^{n} p(x_i) \log p(x_i)$, we calculate that each perfectly dimorphic SNP marker in the fingerprint will contribute $-((0.25)\times\log(0.25))-(0.25\times\log(0.25))-(0.5\times\log(0.5))=0.4515$ bit of entropy, where each $p(x_i)$ is the probability of each genotypic combination. While this is the optimal case, the entropy provided by SNPs where the population frequency is approximately split between two allelic choices should be similar.
2. Randomness extractor. While the genetic fingerprint is a bit stream with high entropy, it is not truly random, and truly random secret keys are necessary to obtain the security guarantees promised by encryption functions. In order to use the fingerprint to obtain a truly random bit stream, we employ a randomness extractor.[25 26] Randomness extractors take strings with high entropy, a small number of truly random bits as a seed, and produce a string of bits that appear to be truly random, even when viewed with the seed. Randomness extractors have properties that make them appealing for this activity. First, there is a one-time fixed cost to establish the data used for the randomness

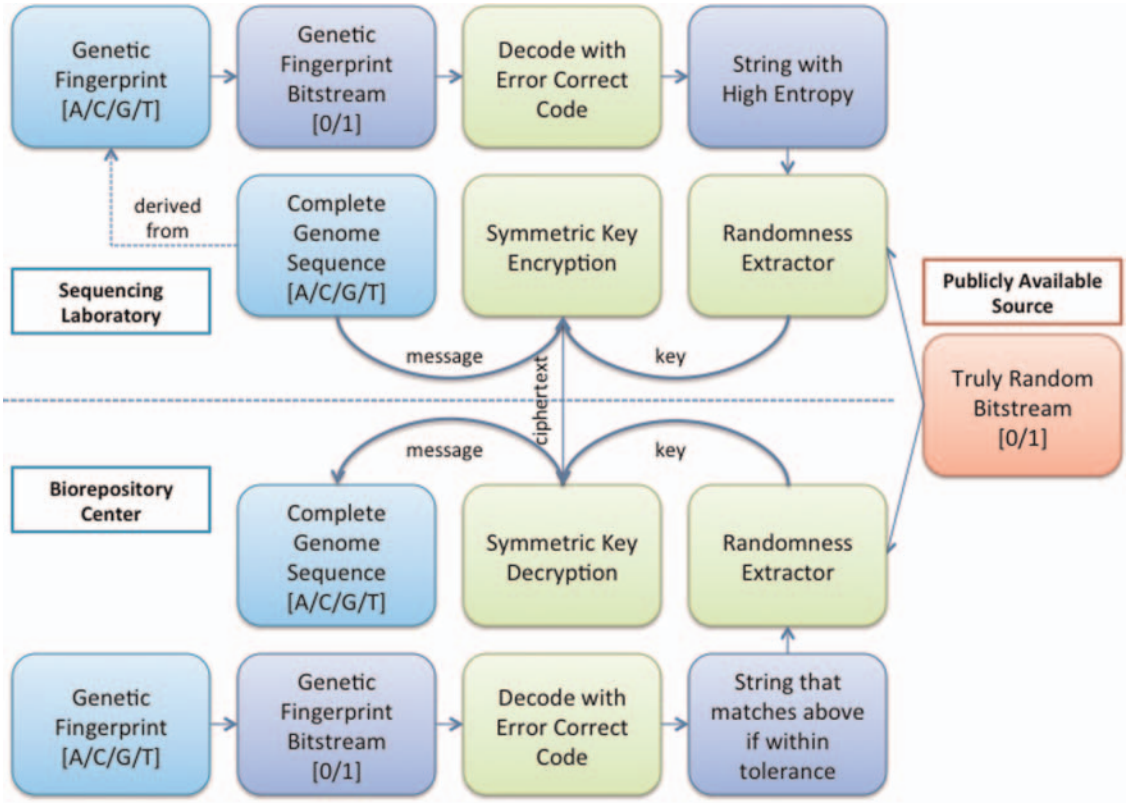

Figure 3 Cryptographic architecture. The shared symmetric key is derived from the genetic fingerprint, which is decoded using error correcting code and then processed through a randomness extractor. This key is used to encrypt the complete genome sequence at the sequencing laboratory, and also to decrypt the complete genome sequence at the biorepository center.

extractor, which can then be used many times on different genomes. Second, the seed used to create the shared secret key does not have to be private, it just must be random, and may be drawn from a mutually agreed set of random data (item 3, below.) Third, each different sequencing provider that processes data for a hospital may use that same, publicly available data, without setting up a new public-key infrastructure (PKI) partnership for each institution. The randomness extractor will act on the DNA fingerprint to generate random bits to be secret keys that are derived from the DNA fingerprint.

3. Random data source. In order for the sequencing laboratory and the biorepository center to generate the same secret key, they will need to use the same seed to extract randomness from the genetic fingerprint. This can be accomplished in two different ways: both sites can use a free public source of random bits for the seed, such as those made available by the NIST Beacon Project[27] or the sequencing laboratory can generate the seed randomly, and send this to the biorepository center in the clear. Due to the properties of randomness extractors, knowledge of this random data will not provide any information about the secret key.
4. Symmetric key encryption algorithm. Because randomness extractors produce a deterministic output for a given input string and seed, both the sequencing laboratory and the biorepository center will generate the same secret key, allowing the use of symmetric key encryption algorithms, such as AES.[27a] Given the genetic fingerprint and the seed, the randomness extractor will output a secret key that appears truly random, and therefore only someone with the genomic fingerprint can reproduce the secret key. The sequencing laboratory will use this key to encrypt the complete genome sequence to secure the transmission of data to the biorepository center; the biorepository center will generate this key and use it to decrypt the full genome sequence. We specify that this key must be at least 128 bits long so that established symmetric key algorithms may be used.
5. Optional use of ECC. Errors in the sequencing process would cause the genetic fingerprints to differ, even with just one error at the sequencing laboratory or the biorepository center. Using an ECC, we can guarantee that both sites will generate the same secret key as long as the number of errors is small. To do this, we view the binary string representing the fingerprint as being encoded in an ECC. As long as the number of errors in this string is small, decoding should lead to the same 'correct' string. Note that the decoded correct string may now be in a different form than the original encodings, and so will not necessarily represent the fingerprint in the same way. Thus decoding with an ECC that allows the number of errors we would like to allow should lead to the same high entropy string; this will in turn be the string used by the randomness extractor to generate the symmetric key.

Because we can select the markers in the genetic fingerprint a priori, it is sensible to select those genotypes that have very high concordance across platforms, with high sequence call accuracies. For this reason, it is highly unlikely that more than one error will be made by the sequencing laboratory and the biorepository center together. This will keep the ECC decoding within the tolerance of errors it can correct.

RESULTS

We define a secure architecture that permits the acquisition and processing of biorepository samples for translational research in a tertiary care setting. This architecture cryptographically protects the data that are transferred, and ensures that only an individual with access to the DNA fingerprint can produce the key needed to reveal the entire genomic sequence. The data transfer uses established symmetric key cryptography, whose keys are partially derived from an extract of the actual sample, leveraging the fact that the specimen transfer includes unique, high-entropy data to obviate the need for additional communication between the sender and receiver. Additionally, by using a portion of the sequence to create the cryptographic key, we provide an assurance that the samples have not been mismatched during processing or sequencing; the conversion will only work if the data match on both sides, representing a patient match.

Balancing privacy: more complex genetic fingerprints versus the introduction of sequencing errors

The privacy assured by this system must be balanced against the risk of introducing sequencing errors. Because sequencing technologies are imperfect, the risk of a sequencing error in a DNA fingerprint increases with each additional genotype that is included. We begin by assessing the cryptographic key space of the genetic fingerprint which will be used by the randomness extractor. The genetic fingerprint should ideally be long enough so that there is sufficient cryptographic key space to avoid a dictionary attack without the use of the randomness extractor, and also sufficient entropy so that it would be difficult to infer it through probabilistic assessment after it has been processed.

As the genetic fingerprint grows in length, even very robust genotyping technologies will produce at least one error. As a function of sequencing accuracy, p, and the number of loci used in the genetic fingerprint, n, the probability of a sequencing error being introduced (at least one but fewer than n errors) is:

$$F(n-1;n,p)=\sum_{i=0}^{\lfloor (n-1)\rfloor}\binom{n}{i}p^i(1-p)^{n-1}$$

The cryptographic key space also grows with the number of loci used in the genetic fingerprint. If dimorphic loci with two alleles are used, there are three possible genotypes for each locus, and the growth of the key space increases exponentially as a function of 3^n.

If high call confidence SNP markers are selected a priori for the genetic fingerprint, it is reasonable to assume a genotyping accuracy of 99.95%, as the overall concordance on a commonly used genotyping platform is around 99.8%, which includes underperforming probes.[28 29] With this specified probability of call accuracy, the value of p(sequencing error(s) present) grows with respect to the cumulative binomial distribution above. This exponential growth of the cryptographic key space and the linear growth of the probability of sequencing errors are shown graphically in figure 4.

This demonstrates that a small, cost-efficient genetic fingerprint, such as one that comprises 20 loci, would have a key space of 3 486 784 401 combinations and would be susceptible to a less than 1% mismatch rate as a result of a sequencing error. While a key space of 3.49 billion is not very large, and would ordinarily be susceptible to a dictionary attack (where all possible key values are generated into a long dictionary list for future look-up, or decryption attempts), we employ a randomness extractor to prevent such an attack by dramatically increasing the entropy of the shared secret key we generate. Other techniques could include the use of salts or additional temporal or institute based identifiers.

To protect against a dictionary attack without the use of a randomness extractor or other advanced cryptographic technique, it is possible to select a DNA fingerprint containing 50 genotypes to achieve a key space of 7.18E23, which at a rate of 1 billion keys per second would take 22.77 million years to exhaust and an expected 11.38 million years to uncover a matching key. With the same baseline sequencing error rate, the probability of a sequencing error in 50 SNPs is below 2.5%. Again, this mismatch rate can be mitigated using other strategies, such as an ECC. Alternatively, the sites can separately encrypt and transmit the complete sequence using another set of distinct genetic fingerprints when the first does not work. If at least one of these sequences matches, then the results can be successfully decrypted. With this strategy, for any reasonable size DNA fingerprint above 10 markers, the probability of re-identification is effectively 1.

Uniqueness of genetic fingerprint for re-identification

Given that there are data from thousands of individuals in research biorepositories, we seek to create unique DNA fingerprints that ensure re-linking of the complete sequence only to the correct individual. We calculate two uniqueness values that quantify this: the probability of the DNA fingerprint matching another random individual and the probability of matching a sibling, because many biorepositories contain family members.

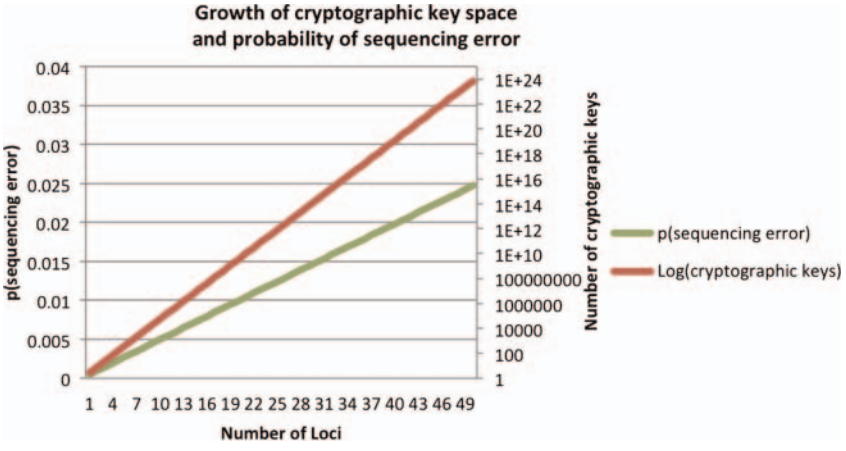

Figure 4 Growth of cryptographic key space and probability of sequencing error, versus number of loci used in the DNA fingerprint. For each additional locus that is included in the DNA fingerprint used to generate the cryptographic material, the probability of at least one sequencing error in that DNA sequence increases linearly (left y-axis) and the total cryptographic key space that is created by that sized DNA fingerprint increases exponentially (right y-axis, on a log scale).

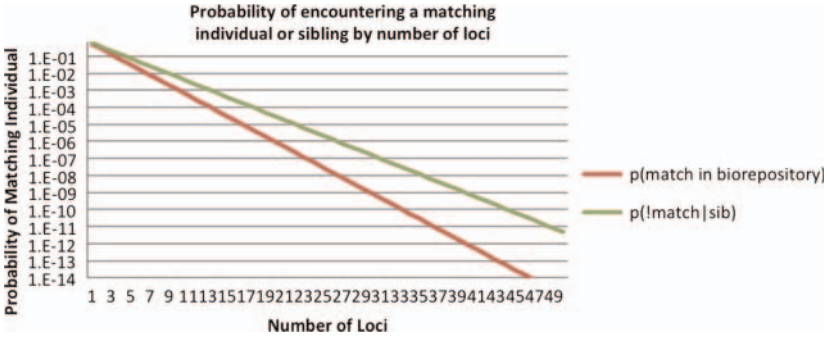

Figure 5 Probability of encountering a matching individual or sibling by number of loci. As the number of dimorphic SNP loci used in the DNA fingerprint increases, the probability of a match in the general population and among family members decreases exponentially. The probabilities of identifying a matching individual are each displayed on a log scale.

For n dimorphic SNP loci, where the population prevalence of both alleles is reliably close to 50% and does not vary appreciably by population group, the probability of identifying a match between two random individuals varies as $(0.5)^n$ and the probability of a match between two siblings is $(0.59375)^n$.[8 9] These exponential growth trends are shown graphically in logarithmic form in figure 5.

Additional constraints and costs

This architecture does incur additional costs for the functionality that must be developed to properly implement it. These costs include the optional generation and transmission of randomized identifiers, the cost of creating a second, local genetic fingerprint (although this is already standard practice at some institutions), and the additional costs associated with cryptographic protection of the genomic data transfer. There are also potential cost savings; specifically, this architecture allows the use of existing hospital infrastructure which depends on the use of MRNs. This potentially reduces the need to hire a separate set of phlebotomy or obviate the need for duplicate laboratory processing, and avoids the need for an additional information infrastructure to manage sample acquisition and processing for a biorepository.

DISCUSSION

At present, it is common for whole genome sequence data to be transferred from sequencing laboratories to researchers on physical drives. This model appears to be changing as sequencing laboratories have begun to transmit, maintain, and analyze these files online and/or in the cloud.[6 7] While differential sequence call files are not very large, it is often preferable for researchers to obtain sequential read data for sequence analysis and realignment, which results in the need to transmit very large files securely. We have developed a cryptographic architecture to support the secure transmission of these large data files by leveraging each individual's unique sequence data.

The mutual benefits of broad public participation in longitudinal biorepository research are enormous, for both patients and researchers, if conducted in a controlled fashion with a keen eye toward participant beneficence.[2] However, new social and ethical risks have been created when biorepository researchers use rich datasets that include large numbers of subjects. Perhaps the most important risk posed to the research subjects is that sharing genetic data might reveal personal or familial propensity to disease. The Genetic Information Non-Discrimination Act (GINA, 2008 H.R. 493) helps protect many individuals and their family members from consequences imposed by employers or health insurers.[30] This law, however, allows for legal consideration of genetic data when setting life, disability, and long-term care insurance premiums.[31] Family members may also be affected, and the risk extends to forensic or criminal investigations for indirect identification of genotype, increasing the number of people who may be identified.[10 11]

Attempts to anonymize genomic data have not been successful as research subjects can often be re-identified uniquely or within a small group of individuals.[32] Through the use of publicly available de-identified data sources, it has been demonstrated that it is often possible to re-identify individuals by linking records to enrich available information about sets of individuals. Malin and Sweeney used a publicly available hospital discharge dataset and combined it with voter records and census data to statistically link individuals within those data sources using zip codes, age, and gender.[32] They were able to uniquely identify patients with rare genetic diseases including a third of all cystic fibrosis patients, half of all patients with Huntington's disease, and even higher numbers of patients with more rare genetic disorders, who were admitted to hospitals in Illinois between 1990 and 1997.

These findings demonstrate that it is possible to directly link publicly available datasets with clinical phenotypes and individual variants. This would certainly alarm some of the patients who had not even personally consented to the release of their healthcare data. Given the characteristics of genomic data, it may not be possible to provide confidentiality or privacy when publishing these data.[8] Further, genomic data and the potential privacy implications of their disclosure are not well understood.[33–37]

Given these potential harms, and that biorepository participants are often engaged long-term at the same clinical settings, there is increased potential to match individual identities from associated phenotypic data and publicly available genotypic data resources. In this architecture, we attempt to mitigate these risks to the participant by protecting against both identifier and sequence data disclosure. This architecture should satisfy IRBs as it allows for removal of MRNs and other protected identifiers from samples before they are processed outside of the hospital, which lowers the risk of harm to human subjects by helping to prevent data disclosure. Additionally, this mode of genome sequence transmission should meet the requirements of the HIPAA Privacy Rule[38] in that we do not simply redact or encode the protected health information that is being transmitted; we individually protect each transmitted sequence using standard cryptographic techniques.

This approach may also help researchers to offer new data while following the NIH Data Sharing Policy, because it provides a novel way to transmit de-identified sequencing data. This could promote the sharing of data with investigators beyond those who initially sequence the biorepository participants. We do not address the important issue of inferential

disclosure of individual biorepository participant identities in this paper,[39] however this approach may be employed to protect against disclosure that could result in inference or prediction of identity.

Alternative cryptographic approaches

We have explored one possible cryptographic approach to protect the transmission of genomic sequence data, focused on the use of an external sequencing laboratory, an increasingly common use case for large scale biorepository research. Other approaches have been explored which provide access to biorepository genome sequence data such as the homomorphic cryptographic approach by Kantarcioglu and Malin[40] and an approach to providing access to genomic data in a personally controlled health record by Adida and Kohane.[41] Future work may explore other cryptographic approaches that do not rely on a shared source of random data, such as the HMAC-based Extract-and-Expand Key Derivation Function (HKDF),[42] as they mature.

There are several distinct advantages to this approach over other cryptographic models, including transmission of data using traditional public-private key cryptography or online data transfer technologies such as SSL. First, this methodology ensures that in order to decrypt and access the full sequence an individual must have access to a copy of the subject's DNA fingerprint. This provides additional protection in that each sample has a unique encryption key set, which would not be the case in traditional public-private key cryptography, where all sequences could be decrypted by a third party who has learned one party's private key. Next, technologies to transmit data securely online, such as SSL, may not scale to the size of data that are being transmitted when processing whole genome sequences. Next, this architecture has the distinct benefit that it provides assurance that there has been no sample mismatch at the sequencing laboratory: in order to decrypt the full sequence, only the correct, matching individual's DNA fingerprint can be used. Finally, while this approach is somewhat more complex than traditional cryptographic approaches, no secret key transmission is required, and thus its implementation in software is actually less complicated and there is no need to transmit or maintain secret cryptographic keys; the random data that are used for the randomness extractor may be in the public domain (such as data available from NIST).

Choice of identifier to transmit to the sequencing laboratory

In the figures and online supplementary material, we included the use of a randomly generated identifier to be transmitted along with the sample to the sequencing laboratory. That identifier would then be transmitted back with the encrypted sequence. As described, it is possible to implement this scheme without sending *any* identifier to the sequencing laboratory. Another possible implementation is that this sample processing identifier—if made sufficiently long and random—could serve as the randomly generated data used as seed data for the randomness extractor. Such an approach would require matching a sample using the local DNA fingerprint to generate shared secret keys for all sequences that have not yet been received, and attempting to use those to decrypt sequence data as they are returned. It is important to note that if an adversary can specify or change the random data that are used by the randomness extractor to non-random data, the generated key will not be random, and the security assumptions using symmetric key cryptography will not necessarily hold. For this reason, it is preferable to use a random data source that is described elsewhere (such as NIST) or transmitted securely.

CONCLUSION

This methodology differs from other standard cryptographic techniques in that it does not require the transmission of any keys, by leveraging the naturally shared genotype data for each individual to encrypt each whole genome sequence. It ensures that the sequence is encrypted throughout the transmission process, and if there is a security failure, such as a disclosure of one genetic fingerprint, the uniqueness of the encryption process ensures that each sample is separate, and will not result in a failure of encryption for a set of samples.

Acknowledgments The authors would like to thank Ben Adida, Sarah Savage, and Tram Tran for helpful advice and discussions.

Contributors CAC conceived and supervised the study, conducted the analysis, and prepared the manuscript. RAM contributed extensively to the cryptographic approach. KDM contributed to the approach and context. All authors revised the manuscript for important intellectual content and approved the version submitted for review.

Funding This research was supported by grant LM010470-01 from the National Library of Medicine and the Manton Center for Orphan Diseases at Children's Hospital Boston (Dr Mandl), and by training grant HD040128 from the National Institute of Child Health and Human Development (Dr Cassa).

Competing interests None.

Provenance and peer review Not commissioned; internally peer reviewed.

REFERENCES

1. **Blow N.** Biobanking: freezer burn. *Nat Methods* 2009;**6**:173–8.
2. **Kohane IS,** Mandl KD, Taylor PL, *et al*. Medicine. Reestablishing the researcher-patient compact. *Science* 2007;**316**:836–7.
3. **Roden DM,** Pulley JM, Basford MA, *et al*. Development of a large-scale de-identified DNA biobank to enable personalized medicine. *Clin Pharmacol Ther* 2008;**84**:362–9.
4. **McGuire AL.** 1000 Genomes on the Road to Personalized Medicine. *Per Med* 2008;**5**:195–7.
5. NCBI. Genetests.org. 2012. http://www.ncbi.nlm.nih.gov/sites/GeneTests/ (accessed 24 Oct 2012).
6. **Illumina I.** BaseSpace: Genomics Cloud Computing. 2012. http://basespace.illumina.com (accessed 24 Oct 2012).
7. **Services AW.** AWS Genomics Event. 2011. http://aws.amazon.com/genomicsevent/ (accessed 24 Oct 2012).
8. **Lin Z,** Owen AB, Altman RB. Genetics. Genomic research and human subject privacy. *Science* 2004;**305**:183.
9. **Cassa CA,** Schmidt B, Kohane IS, *et al*. My sister's keeper?: genomic research and the identifiability of siblings. *BMC Med Genomics* 2008;**1**:32.
10. **Bieber FR,** Lazer D. Guilt by association: should the law be able to use one person's DNA to carry out surveillance on their family? Not without a public debate. *New Sci* 2004;**184**:20.
11. **Bieber FR,** Brenner CH, Lazer D. Human genetics. Finding criminals through DNA of their relatives. *Science* 2006;**312**:1315–6.
12. **Homer N,** Szelinger S, Redman M, *et al*. Resolving individuals contributing trace amounts of DNA to highly complex mixtures using high-density SNP genotyping microarrays. *PLoS Genet* 2008;**4**:e1000167.
13. **Sankararaman S,** Obozinski G, Jordan MI, *et al*. Genomic privacy and limits of individual detection in a pool. *Nat Genet* 2009;**41**:965–7.
14. **Wang RYFL,** Wang X, Tang H, *et al*. Learning your identity and disease from research papers: information leaks in genome wide association study. *CCS '09: Proc of the 15th ACM Conf Comput Commun Secur* 2009;15:534–44.
15. **Benjamin EJ,** Dupuis J, Larson MG, *et al*. Genome-wide association with select biomarker traits in the Framingham Heart Study. *BMC Med Genet* 2007;**8**(Suppl 1): S11.
16. **Genome-Wide Association Studies**. 2008. http://www.genome.gov/20019523 (accessed 24 Oct 2012).
17. **Morton NE.** Into the post-HapMap era. *Adv Genet* 2008;**60**:727–42.
18. **Cappuccio FP,** Oakeshott P, Strazzullo P, *et al*. Application of Framingham risk estimates to ethnic minorities in United Kingdom and implications for primary prevention of heart disease in general practice: cross sectional population based study. *BMJ* 2002;**325**:1271.
19. **Colditz GA,** Coakley E. Weight, weight gain, activity, and major illnesses: the Nurses' Health Study. *Int J Sports Med* 1997;**18**(Suppl 3):S162–70.
20. **Empana JP,** Ducimetiere P, Arveiler D, *et al*. Are the Framingham and PROCAM coronary heart disease risk functions applicable to different European populations? The PRIME Study. *Eur Heart J* 2003;**24**:1903–11.
21. **NCBI.** database of Genotypes and Phenotypes (dbGaP). 2012. http://www.ncbi.nlm.nih.gov/gap (accessed 24 Oct 2012).

22. **NHLBI.** NHLBI GO Exome Sequencing Project. 2012. http://esp.gs.washington.edu/drupal/ (accessed 24 Oct 2012).
23. **EBI.** *The European Genome-phenome Archive*. 2012. http://www.ebi.ac.uk/ega/ (accessed 24 Oct 2012).
24. **NCBI.** ClinVar. 2012. http://www.ncbi.nlm.nih.gov/clinvar/ (accessed 24 Oct 2012).
25. **Vadhan S.** Randomness Extractors. 2012. http://people.seas.harvard.edu/~salil/pseudorandomness/extractors.pdf (accessed 24 Oct 2012).
26. **Venkatesan G,** Umans C, Vadhan S. Unbalanced expanders and randomness extractors from Parvaresh–Vardy codes. *J Assoc Comput Machinery* 2009;**56**:Article 20.
27. **Technology NIoSa.** NIST Randomness Beacon. 2012. http://www.nist.gov/itl/csd/ct/nist_beacon.cfm (accessed 24 Oct 2012).

27a. **EmbeddedSw.net.** Cryptography – 256 bit ciphers: 256bit key – 128bit block – AES. 2012. http://embeddedsw.net/Cipher_Reference_Home.html#AES (accessed 24 Oct 2012).

28. **Nishida N,** Koike A, Tajima A, *et al*. Evaluating the performance of Affymetrix SNP Array 6.0 platform with 400 Japanese individuals. *BMC Genomics* 2008;**9**:431.
29. **Korn JM,** Kuruvilla FG, McCarroll SA, *et al*. Integrated genotype calling and association analysis of SNPs, common copy number polymorphisms and rare CNVs. *Nat Genet* 2008;**40**:1253–60.
30. **Holden C.** Genetic discrimination. Long-awaited genetic nondiscrimination bill headed for easy passage. *Science* 2007;**316**:676.
31. **Hudson KL,** Holohan MK, Collins FS. Keeping pace with the times—the Genetic Information Nondiscrimination Act of 2008. *N Engl J Med* 2008;**358**:2661–3.
32. **Malin BA,** Sweeney LA. Inferring genotype from clinical phenotype through a knowledge based algorithm. *Pac Symp Biocomput* 2002;**41–52**.
33. **Henneman L,** Timmermans DR, van der Wal G. Public experiences, knowledge and expectations about medical genetics and the use of genetic information. *Commun Genet* 2004;**7**:33–43.
34. **Levitt DM.** Let the consumer decide? The regulation of commercial genetic testing. *J Med Ethics* 2001;**27**:398–403.
35. **Miller SM,** Fleisher L, Roussi P, *et al*. Facilitating informed decision making about breast cancer risk and genetic counseling among women calling the NCI's Cancer Information Service. *J Health Commun* 2005;**10**:119–36.
36. **Mouchawar J,** Hensley-Alford S, Laurion S, *et al*. Impact of direct-to-consumer advertising for hereditary breast cancer testing on genetic services at a managed care organization: a naturally-occurring experiment. *Genet Med* 2005;**7**:191–7.
37. **Mouchawar J,** Laurion S, Ritzwoller DP, *et al*. Assessing controversial direct-to-consumer advertising for hereditary breast cancer testing: reactions from women and their physicians in a managed care organization. *Am J Manag Care* 2005;**11**:601–8.
38. **Services USDoHH.** The Privacy Rule. Health Information Privacy. 2002. http://www.hhs.gov/ocr/privacy/hipaa/administrative/privacyrule/index.html (accessed 24 Oct 2012).
39. **Malin B,** Loukides G, Benitez K, *et al*. Identifiability in biobanks: models, measures, and mitigation strategies. *Hum Genet* 2011;**130**:383–92.
40. **Kantarcioglu M,** Jiang W, Liu Y, *et al*. A cryptographic approach to securely share and query genomic sequences. *IEEE Trans Inf Technol Biomed* 2008;**12**:606–17.
41. **Adida B,** Kohane IS. GenePING: secure, scalable management of personal genomic data. *BMC Genomics* 2006;**7**:93.
42. **(IETF) IETF**. HMAC-based Extract-and-Expand Key Derivation Function (HKDF). ISSN:2070-1721 May 2010. http://tools.ietf.org/html/rfc5869 (accessed 24 Oct 2012).