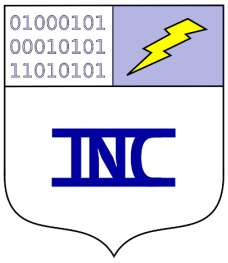

www.intrepidnetcomputing.com

B. Kirkpatrick, Ph.D.
Security Consultant

Phone: 1-406-988-0179
Cell: 1-406-660-7100
bbkirk@intrepidnetcomputing.com

January 5, 2016

Geneticist

RE: Open Letter to Genetics Researchers

Dear Geneticist,

This is a letter to inform the computational genetics communities of changes in their responsibilities to research subjects due to technological advances. Due to the increasing availability of dense genetic data, improvements in statistical data analysis, and changes in computer security, I believe that all human genome data needs to be physically secured in off-line computers. I recommend that all analysis be carried out off-line, and all communication of genetic data to colleagues be done either via wired transmission or off-line with physical media.

Before I discuss these issues and outline my recommendations, I will introduce myself and my qualifications. I am a researcher who specializes in computational family genetics. I have three degrees in computer science: B.S. from Montana State University, M.S. from the University of California Berkeley, and a Ph.D. from the University of California Berkeley under the supervision of Richard M. Karp and Eran Halperin. I did post-doctoral work with Anne Condon in the Department of Computer Science at the University of British Columbia, Canada. I am formerly an Assistant Professor in Computer Science at the University of Miami. I am now the CEO of Intrepid Net Computing where I have the freedom to influence the necessary security precautions for genetic data, the freedom to openly document relevant procedures, the freedom to disclose security breaches, the freedom to consult other computer scientists, and the freedom to delay analysis or methods work should security circumstances become unfavorable.

As you know, modern assays allow us to collect genetic data from loci that are dense, meaning physically located closely together, in the human genome. I am referring to data collection methods beginning with the SNP genotyping microarrays of the early 2000's and including the modern sequencing methods. All of these assays are sufficiently dense to be problematic in the privacy sense, as we will discuss next.

Research results reveal that human genetic data alone, without names, are *identifiable* in the sense that the data distinctly identifies the person almost surely from all other people on the planet. I use 'almost surely' in the statistical sense, meaning that as we collect more data from a human, we are able to distinguish that individual from an increasing number of other humans. This means that the names of the individuals can be recovered using a statistical analysis of genetic data. Indeed this statement is so blindingly obvious, that I have to ask myself why we even started collecting this data. Statistically speaking, the only answer is that the genome is finite and might be finite enough that the probability of identification remains small. However, for decades, studies repeatedly demonstrate that the human genome is sufficiently long to distinctly identify

B. Kirkpatrick, Ph.D.
Security Consultant

Phone: 1-406-988-0179
Cell: 1-406-660-7100
bbkirk@intrepidnetcomputing.com

individuals. Furthermore, even possible existence of genetic doppelgangers does not relieve us of the moral responsibility of being extremely careful.

Bickeboller and Thompson, 1999 [2] statistically demonstrates that it is sufficient to have the genome of a child, the child's name, and the parent's genome in order to identify the name of the parent. Almudevar and Field, 1999 [1] uses statistics to reveal that a sibling's genome and name and the individual's name is all that is necessary to identify the individual. Sankararaman, et al. 2009 [6] reveals something even more disturbing. Given the pooled allele frequencies of a number of unrelated individuals' genotypes and the genotype of a single individual, we can detect whether that individual volunteered for that study. Gitschier, 2009 [4] studied the individuals in the CEPH pedigrees and was able to use the Y-chromosome genotypes and publicly available genealogical data to determine the names of the men in the pedigree. Cowell, 2009 [3] demonstrated that statistics can be used to reconstruct family relationships from genotype data. Furthermore, he demonstrated that if one has access to genealogical data and genotype data, one could identify the names of the individuals. He did this by identifying the tomb near Ekaterinburg, Russia which contains the remains of the family or Russian Zsar Nicholas II. Gymrek, et al. 2013 [5] replicated the above results.

Genetic data is typically stored on the computers that scientists use to analyze the data. This means the computers connected to the sequencing machines store the data, data centers the run the statistical analysis store the data, and often developers of the statistical methods have a subset of the data on their computers. In the past, most of these machines have been networked to the Internet, presumably for ease of communication.

Why do we need to reconsider this approach? Because computer security is a dynamic environment and the computers with the greatest risk of attack are computers that are networked to the Internet. Of the computers networked to the Internet, the risks seem to ranked as follows: wireless computers that travel internationally, wireless computers that travel domestically, and targeted computers with a fixed IP address. Furthermore, computer security can improve or worsen in various localities, just like the weather, and one could reasonably adopt different security measures during a hacking storm than during calm cyber weather. (Disclaimer: the assessment of risk may vary based on the circumstances.)

Worse still is the use of the cloud for genetic data. The cloud technologies are new, untested, and take away from the user the ability to influence data storage. Cloud technologies rely on some form of operating system virtualization. In computer science virtualization is known to be unstable in many aspects, including privacy and security. This is because the virtual operating system often runs inside the native operating system with guaranteed control over access permissions. Furthermore, the user does not know where their data is stored. This last means that the users cannot delete the data. With computing, permanent file deletion is quite technical and depends on storage medium and access permissions. Operating systems have been designed to easily recover accidentally deleted files, since users have clamored for this feature for all of computing history. Unfortunately, this feature makes file deletion, especially in the cloud, very

B. Kirkpatrick, Ph.D.
Security Consultant

Phone: 1-406-988-0179
Cell: 1-406-660-7100
bbkirk@intrepidnetcomputing.com

difficult.

Current computer security measures require knowing how to respond when intrusions occur. It is no longer a question of 'if' they will occur, as it was in the 1980's. Response in the United States now involves a technical assessment of the risk, technical assessment of the affected electronics, reports to law enforcement, and notifications to individuals whose personal data was affected. Indeed, permanent deletion of identifying data essentially requires reporting, as file deletion is so technical. Research data are not exempt from the dizzying and confusing array of state and federal laws that govern Internet crime and reporting of data breaches. The laws are so complicated and untested that there are likely conflicts between the IRB procedures that require confidentiality and the breach reporting laws that require notification.

Furthermore, computational people that handle the data are several degrees of separation away from the IRB confidentially, they are unable to verify that breaches are properly reported to the people who donated their tissue samples. In particular, in cases where the computational people are convinced there was a breach and are convinced of the possibility of malicious intent, it is quite possible that a disagreement with the PI responsible for privacy will prevent reports of breaches reaching the people they are intended to reach.

Of even more concern are the terrifying possibilities that an adversarial and malicious attacker might have in mind, particularly in the twenty-first century. These possibilities cannot be mitigated by breach reports. Medical research subjects who consent to genetic analysis are at extreme risk. As medical research is translated into the clinic where electronic medical records are becoming common, the same risks will apply to everyone. I will outline just a few disaster scenarios, and I am sure more thought by other scientist would result in a much longer list.

Suppose that an adversary has either the genotype or the sequence of several individuals in the same family.

1. An adversary could target a whole family's health with the genomes of several individuals. The adversary would run a genetic risk analysis and target that risk in the opposite manner that a doctor would.
2. When DNA synthesis becomes feasible for the adversary, they could frame a person for a crime by leaving that person's synthesized DNA at a crime scene.
3. Worse yet, the adversary only needs the genome of a close relative of the individual being framed in order to synthesize the 13 CODIS loci.
4. If cloning is advanced enough, an adversary could clone portions of the target's anatomy from their genome. When science is advanced enough, this may compromise typical methods of identification such as fingerprints, iris scanning, and voiceprints.
5. If DNA is ever used for identification (and it already is used this way for crime scene forensics), then an adversary with a copy of an individual's genome sequence would be able to impersonate that person. All the adversary has to do is steal the sequence today and wait 50 years or so until DNA is used for identification.

B. Kirkpatrick, Ph.D.
Security Consultant

Phone: 1-406-988-0179
Cell: 1-406-660-7100
bbkirk@intrepidnetcomputing.com

Certainly it is well-established that primitive versions of these risks exist. However, our risk analysis needs to be reassessed for the modern era. The remaining questions are: whether the reassessed risk is great enough to change IRB procedures and whether there are effective and inexpensive ways to mitigate the risk. I believe that the answers to both questions are a resounding yes.

As for IRB procedures, my personal experience well-argues for revisions. As a graduate student, I was given few directions on data storage or data security. This approach is insufficient, because computer security, electronic privacy, and genetics need to be discussed more frequently together. I now find myself having to remind my faculty colleagues in genetics of the privacy risks inherent in modern computing.

There are very simple and very effective ways to mitigate these risks. These are many well-known security procedures used routinely by the computing industry that would easily and cost-effectively improve security with little sacrifice in convenience.

1. All health data should be stored and analyzed with standards that meet or exceed HIPAA standards.
2. Breaches to health data on a US computer or to a data set collected in the US need to be reported to the Department of Health and Human Services.
3. Data should not be emailed.
4. Data should not be stored in the cloud.
5. Data should be stored on a separate drive or computer from the development code.
6. Moving labs and moving data should be done with extreme care.
7. We should modernize IRB procedures, modernize training, improve security of the data storage, and improve security of the data during analysis.
8. Restrict the data to wired networks, not mobile computers, and restrict the data to standard computer clusters that do not rely on cloud technologies.
9. Research labs should have an off-line intra-network devoted to data analysis and restricted to physical access only.
10. We should develop data analysis methods on simulated data, and have the researchers who collected the data run the actual analysis off-line.

This last measure is justified both by security and by the scientific desire to asses accuracy.

I now consider human genome data to be identifiable despite the fact the NIH has yet to update their definition of identifiable data. I hope that the NIH updates their definition and requires that all scientists, PI's and students, who are involved in analyzing genetic data be trained in IRB procedures and the latest intrusion detection procedures. This means that genetics researchers should be required to learn something about electronic privacy. I also hope that methods researchers who handle genetic data would be required to take a course in the ethics of medical genetics. This seems necessary to understand the many conflicts between various laws and ethical concerns, as well as to more clearly understand the arguments for doing such dangerous

B. Kirkpatrick, Ph.D.
Security Consultant

Phone: 1-406-988-0179
Cell: 1-406-660-7100
bbkirk@intrepidnetcomputing.com

research at all. I hope that IRB procedures are updated to reflect the latest changes in the federal computing laws that require reporting security breaches to the people affected by them.

Sincerely,

B. Kirkpatrick, Ph.D.
Security Consultant

B. Kirkpatrick, Ph.D.
Security Consultant

Phone: 1-406-988-0179
Cell: 1-406-660-7100
bbkirk@intrepidnetcomputing.com

References

[1] Anthony Almudevar and Chris Field. Estimation of single-generation sibling relationships based on dna markers. *Journal of Agricultural, Biological, and Environmental Statistics*, 4(2):pp. 136–165, 1999.

[2] H. Bickeböller and E. A. Thompson. The probability distribution of the amount of an individual's genome surviving to the following generation. *Genetics*, 143:1043–1049, 1996.

[3] R.G. Cowell. Efficient maximum likelihood pedigree reconstruction. *Theoretical Population Biology*, 2009.

[4] Jane Gitschier. Inferential genotyping of y chromosomes in latter-day saints founders and comparison to utah samples in the hapmap project. *Am J Hum Genet*, 84(2):251–258, February 2009.

[5] M. Gymrek, A.L. McGuire, D. Golan, E. Halperin, and Y. Erlich. Identifying personal genomes by surname inference. *Science*, 2013.

[6] S. Sankararaman, G. Obozinski, M.I. Jordan, and E. Halperin. Genomic privacy and limits of individual detection in a pool. *Nature Genetics*, 41(9):965–967, 2009.