SAND2017-9752C

Emulating Genome Security Risks in Realistic Genomics Data Ecosystems

Corey M. Hudson[1,2]

[1]Sandia National Labs, Livermore, CA, [2]Joint BioEnergy Institute, Emeryville, CA

Abstract

Rapid progress in genome sequencing technology and synthetic biology have resulted in significant unquantified risk that must be mitigated, and that the commercial sector is unlikely to address. Between 2006 and 2016 the cost for a genome dropped from ~$20-25 million to ~$1000.[1] This change has been accompanied by an economy of scale, stretching the bounds of consumer data storage and processing, and moving genomics from an exclusively academic venture to a commercial product.[2] Federal departments and agencies have invested heavily in genomics.[3] But, without an understanding of the risk that this move from academic to commercial technology has created, government genomics operations are at risk of the loss of operational security and adversarial surveillance, release of protected data (including intellectual property, personally identifying information and secret information), destruction of data integrity, failure to complete work, and the potential for personalized genomic targeting.[4] The primary algorithms and software that support routine genomics were developed in academic labs and leave open the potential for gaping security holes and often rely heaving on remote resources and databases. Synthetic biology labs contain even more risks, relying on genomics for validation and verification and using cybersecurity resources to automate genetic engineering.[5] A risk inherent to synthetic biology include unintended gene/organism manufacture through hijacked command and control operations. The change, in synthetic biology and genomics from “wet-lab” to computational pursuit has lowered the necessary adversarial sophistication to the level of a cyberthreat and opened a large and under-appreciated risk space.[6]

Modeling Realistic Genomics Facilities

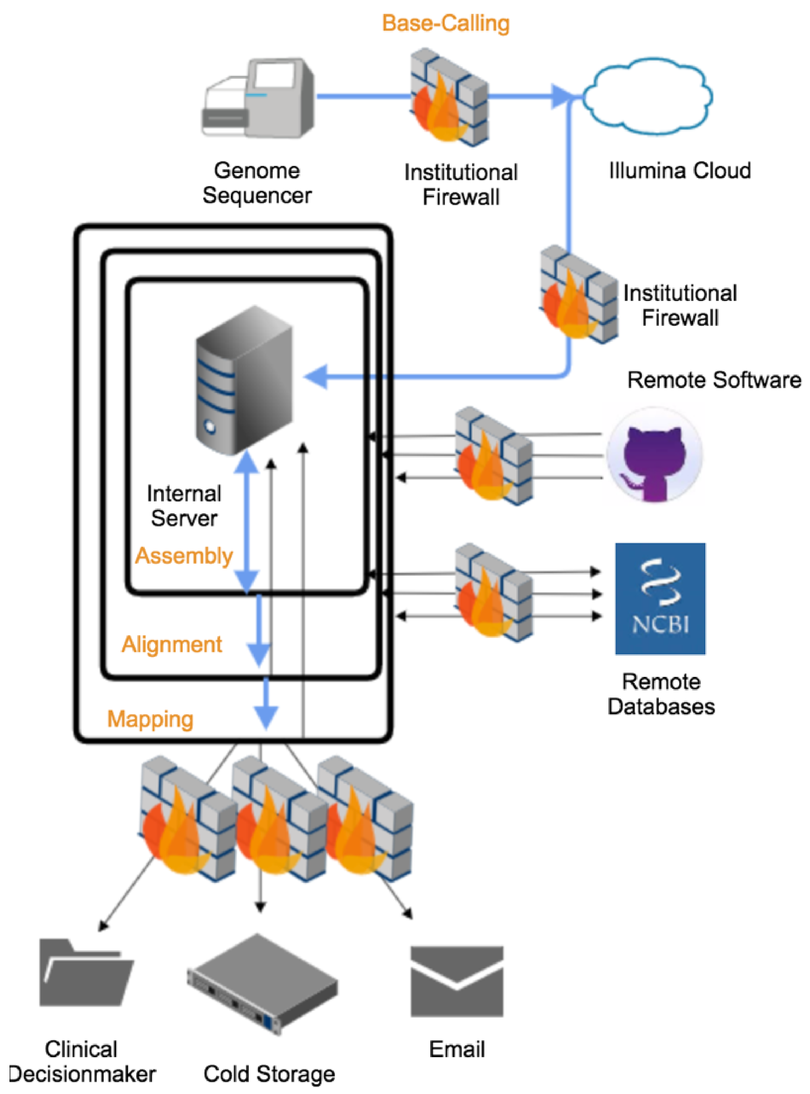


Fig 1 Description of bare-bones bioinformatics pipeline. This figure shows the complex interactions between even simple bioinformatics pipelines between sequencer and the data as it is ultimately used. This includes movement within the network and remote resources. More sophisticated networks contain a higher number of interactions.

What are the Risks?

Risk = Vulnerability x Threat x Impact x Probability

Risks at a facility doing **sequencing genomics**:

- Failure to complete work
- Release of protected data
  - Intellectual property
  - Personally identifying information
  - Secret information
- Destruction of data integrity
- Release of operational security and adversarial surveillance

Consequences around privacy breach

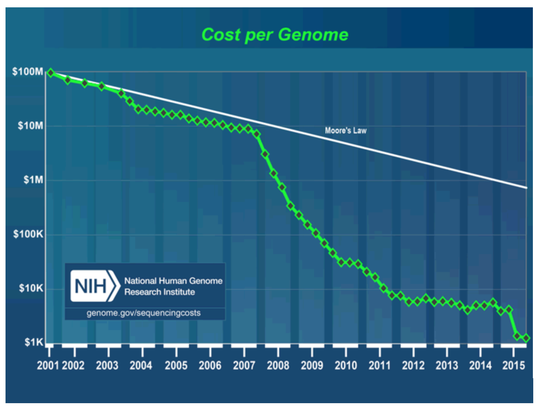
- Loss of faith in genomics studies
- Paternity breach
- Privacy and identification
- Racial or at-risk subgroup identification
- Legal/forensic identification/manipulation
- Phenotype inference
- Genomic access controls
- Genomic targeting

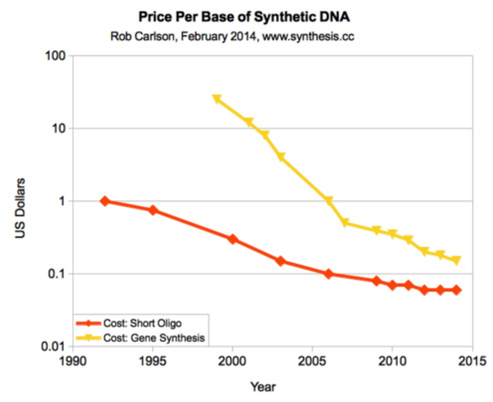
Current

Future

Sandia National Laboratories is a multi-mission laboratory managed and operated by National Technology and Engineering Solutions of Sandia, LLC, a wholly owned subsidiary of Honeywell International, Inc., for the U.S. Department of Energy's National Nuclear Security Administration under contract DE-NA0003525. **SAND Number:**

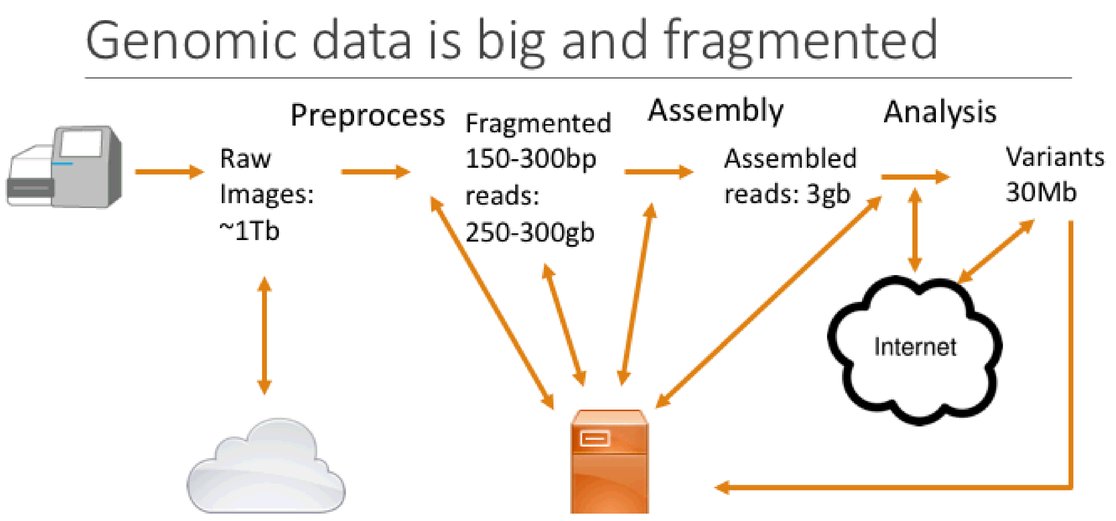
Change in Economy of Scale in Genomics

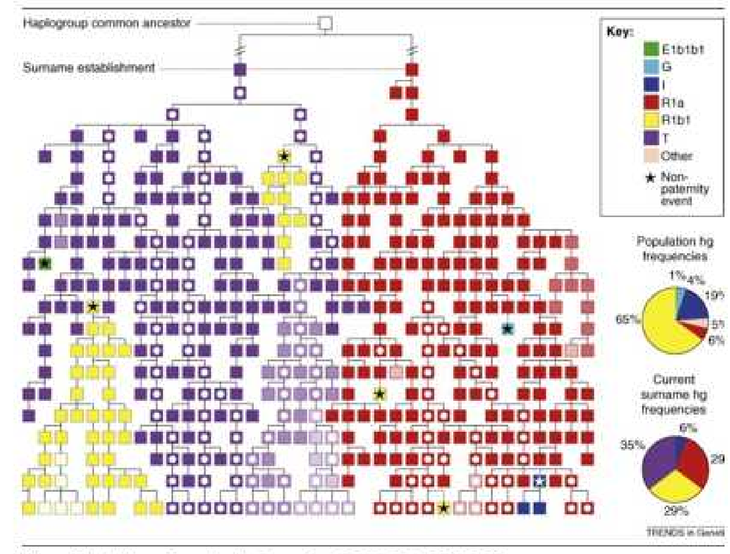




Between 2006 and 2016 the cost for a genome dropped from ~$20-25 million to ~$1000.[1] This change has been accompanied by an economy of scale, stretching the bounds of consumer data storage and processing, and moving genomics from an exclusively academic venture to a commercial product.

How Are Genomics Data Different?





King & Jobling. *Trends in Genetics* 2009 25. 351-360

Genomic data are associational

Every leaked genome leaks data about associated family members.

Asymptotically, this means that genomic data cannot be secured indefinitely.

Funding

Supported by the Laboratory Directed Research and Development program at Sandia National Laboratories, a multi-mission laboratory managed and operated by National Technology and Engineering Solutions of Sandia, LLC, a wholly owned subsidiary of Honeywell International, Inc., for the U.S. Department of Energy's National Nuclear Security Administration under contract DE-NA0003525.