Open science & genomic privacy

Chloé-Agathe Azencott

CBIO, Mines ParisTech – Institut Curie – INSERM U900, Paris (France)

April 1st, 2016 – DALI

http://cazencott.info chloe-agathe.azencott@mines-paristech.fr @cazencott

Computational biology

- Analyzing **large amounts** of human genetic and clinical data to **generate biological hypotheses.**

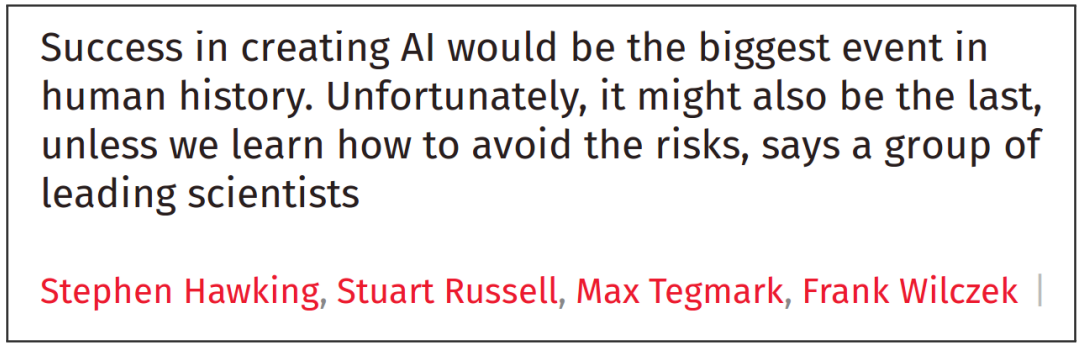
- **Positive impact on society**
  - Biological findings
  - Data-driven medicine
  - Precision medicine
  - Computer-aided diagnosis

What about negative impact?

Should I worry about it?

- I am a member of **society.**
- I am funded by **public money.**
- If I don't, who else will? Isn't it **other people's job?**

  Social scientitsts, ethicists, lawmakers, etc.



Data sharing in computational biology

- **More data** $\Rightarrow$ better algorithms.
- Utilize data maximally.
- Make the most out of public research funding.

Image source: Hyperbole and a half

Big, open data is awesome...

... but so is **privacy**.

Genetic privacy: Why care about it?

- Information about **you**.
- Information about **your family**.
- **Genetic discrimination**.

Genetic discrimination

Being **treated differently** because you have (or are perceived to have) a **genetic mutation** that increases your risk of an inherited disorder.

- Matthewman, W. D. (1984). **Genetic testing: Can your genes screen you out of a job?** Howard LJ, 27, 1185.

Legislation against genetic discrimination

From the **Declaration of Bilbao** (1993) to Article 21 of the **EU Charter of Fundamental Rights** (effective 2009).

- **France (March 2002):** prohibits any discrimination based on genetic characteristics.
- **USA (April 2008), GINA:** restricted to employment and health insurance.
- **Germany (July 2009), Gendiagnostikgesetz.**
- **CalGINA (2012):** housing, mortgage lending, employment, education and public accommodations.

Fear of genetic discrimination

And yet

- **No genetic discrimination law** in e.g. Canada.
- **Fear of genetic discrimination** is still strong [Green et al., 2015].
- Wauters, A. and Van Hoyweghen, I. (2016). **Global trends on fears and concerns of genetic discrimination: a systematic literature review.** Journal of Human Genetics.


SARAH ZHANG SCIENCE 02.01.16 7:00 AM

DNA GOT A KID KICKED OUT OF SCHOOL—AND IT'LL HAPPEN AGAIN

http://www.wired.com/2016/02/schools-kicked-boy-based-dna/

How to **protect** genomic privacy?

Image source: http://www.perspecsys.com/

Anonymization is not enough

Anonymization of records is not enough.

- Your inclusion in the study **will affect the results** of the study;
- The results of the study **will** give (with high probability) **new information about you.**

Anonymization is not enough

2006:

- Identification of individuals in a data base using **genetic markers** corresponding to their phenotye (e.g. skin/hair/eye color) [Malin].

2008:

- **Deanonymization of Netflix data** [Narayanan & Shmatikov].
- Assessing **whether a given genotype is part of a cohort** summed up by **allele frequencies** [Homer et al].

  $\Rightarrow$ NIH and Wellcome Trust **policy update**.

Anonymization is not enough

2009:

- **Quantitative guidelines** for releasing a limited number of SNPs without compromising privacy [Sankararaman et al.].
- Also identify the **phenotype** associated with this genotype [Jacobs et al.].
- Homer et al. extended to only requiring a few hundred SNPs (instead of full genotype) [Wang et al.].

2012:

- Predict **SNPs from gene expression** [Schadt et al.].
- Predict **surnames from Y-STRs** and public genealogical data bases [Gymrek et al.].

Are there **alternative approaches** that provide **appropriate participant privacy** while **maximizing scientific impact?**

http://www.stockmonkeys.com

k-anonymity

- **k-anonymity**: Censor information until it becomes impossible to distinguish one person from $k - 1$ others [Sweeney, 2002].
- **l-diversity**: At least l "well-represented" values for each sensitive attribute [Machanavajjhala et al., 2007].
- **t-closeness**: Bound by t the distance between the distribution of a sensitive attribute within an anonymized group and its distribution within the whole data [Li et al., 2007].

  Not well-suited to **high-dimensional settings.**

Differential privacy

Maximize the potential of a database while **minimizing the chances of identification.**

- Can we guarantee that the **privatized** version of what is released is nearly the same, whether you're included in the study or not?

$$\frac{P(\mathcal{M}(\mathcal{D}) = C)}{P(\mathcal{M}(\mathcal{D} \cup \{x\}) = C)} \leq e^{\epsilon}$$

- **Noise-injection mechanisms**, e.g. Laplace, exponential, or algorithm-specific.
- Price to pay: **accuracy** of the algorithms.

Differential privacy & precision medicine

Differential privacy in personalized warfarin dosing
[Fredrikson et al., 2014]

- Can you predict **genotype** from **black-box model** and marginals, dosage, basic demographics?

  genotype: values of SNPs in two genes of interest (CYP2C9 and VKORC1)

- With current differential privacy mechanisms, model inversion attacks can only be prevented **at the price of exposing patients to increased risk** of stroke, bleeding, and mortality.

Is promising privacy realistic?

- **Trust Not Privacy** [Erlich et al., 2014]

  Transparency, increased control and reciprocity.

- **Secure cloud computing**

  E.g. The Pan-Cancer Analysis of Whole Genomes (PCAWG)

- **Restrictions** on access to data

  A **burden** for (junior) researchers.

Privacy is dead

- Inform participants that their privacy **cannot be guaranteed**, and seek consent nonetheless.
  - The Personal Genome Project
  - OpenSNP
  - 1000 Genomes German cohort.
- **P4 medicine:**
  Preventive, Predictive, Personalized and **Participatory**.

source: http://www.flickr.com/photos/wwworks/

References I

- Misha Angrist.
  Open window: When easily identifiable genomes and traits are in the public domain.
  PLOS ONE, 9(3):e92060, 2014.
- Madeleine P. Ball, Joseph V. Thakuria, Alexander Wait Zaranek, Tom Clegg, Abraham M Rosenbaum, et al.
  A public resource facilitating clinical use of genomes.
  Proceedings of the National Academy of Sciences, 109(30):11920–11927, 2012.
- R. J. Bayardo and Rakesh Agrawal.
  Data privacy through optimal k-anonymization.
  In 21st International Conference on Data Engineering, 2005. ICDE 2005. Proceedings, pages 217–228, 2005.
- Joppe W. Bos, Kristin Lauter, and Michael Naehrig.
  Private predictive analysis on encrypted medical data.
  Journal of Biomedical Informatics, 50:234–243, 2014.
- Paul R. Burton, Madeleine J. Murtagh, Andy Boyd, James B. Williams, Edward S. Dove, et al.
  Data Safe Havens in health research and healthcare.
  Bioinformatics, 31(20):3241–3248, 2015.
- Fida Kamal Dankar and Khaled El Emam.
  The application of differential privacy to health data.
  In Proceedings of the 2012 Joint EDBT/ICDT Workshops, EDBT-ICDT '12, pages 158–166, New York, NY, USA, 2012. ACM.
- Cynthia Dwork.
  Differential Privacy.
  In Automata, Languages and Programming, number 4052 in Lecture Notes in Computer Science, pages 1–12. Springer Berlin Heidelberg, 2006.
- Cynthia Dwork.
  The promise of differential privacy: A tutorial on algorithmic techniques.
  In Proceedings of the 2011 IEEE 52Nd Annual Symposium on Foundations of Computer Science, pages 1–2, Washington, DC, USA, 2011.

References II

- Yaniv Erlich and Arvind Narayanan.
  Routes for breaching and protecting genetic privacy.
  Nature reviews Genetics, 15(6):409–421, 2014.
- Yaniv Erlich, James B. Williams, David Glazer, Kenneth Yocum, Nita Farahany, Maynard Olson, Arvind Narayanan, Lincoln D. Stein, Jan A. Witkowski, and Robert C. Kain.
  Redefining genomic privacy: Trust and empowerment.
  PLOS Biol, 12(11):e1001983, 2014.
- Matthew Fredrikson, Eric Lantz, Somesh Jha, Simon Lin, David Page, and Thomas Ristenpart.
  Privacy in pharmacogenetics: An end-to-end case study of personalized warfarin dosing.
  In 23rd USENIX Security Symposium, pages 17–32, 2014.
- Dov Greenbaum, Andrea Sboner, Xinmeng Jasmine Mu, and Mark Gerstein.
  Genomics and privacy: Implications of the new reality of closed data for the field.
  PLoS Computational Biology, 7(12), 2011.
- Melissa Gymrek, Amy L. McGuire, David Golan, Eran Halperin, and Yaniv Erlich.
  Identifying personal genomes by surname inference.
  Science, 339(6117):321–324, 2013.
- Arif Harmanci and Mark Gerstein.
  Quantification of private information leakage from phenotype-genotype data: linking attacks.
  Nature Methods, 13(3):251–256, 2016.
- Nils Homer, Szabolcs Szelinger, Margot Redman, David Duggan, Waibhav Tembe, et al.
  Resolving individuals contributing trace amounts of DNA to highly complex mixtures using high-density SNP genotyping microarrays.
  PLOS Genet, 4(8):e1000167, 2008.

References III

- Hae Kyung Im, Eric R. Gamazon, Dan L. Nicolae, and Nancy J. Cox.
  On sharing quantitative trait GWAS results in an era of multiple-omics data and the limits of genomic privacy.
  American Journal of Human Genetics, 90(4):591–598, 2012.

- Kevin B. Jacobs, Meredith Yeager, Sholom Wacholder, David Craig, Peter Kraft, et al.
  A new statistic and its power to infer membership and phenotype in a genome-wide association study using genotype frequencies.
  Nature genetics, 41(11):1253–1257, 2009.

- Yann Joly, Edward S. Dove, Bartha M. Knoppers, Martin Bobrow, and Don Chalmers.
  Data sharing in the post-genomic world: The experience of the international cancer genome consortium (ICGC) data access compliance office.
  PLOS Comput Biol, 8(7):e1002549, 2012.

- Tatiana Komarova, Denis Nekipelov, and Evgeny Yakovlev.
  Estimation of treatment effects from combined data: Identification versus data security.
  NBER Chapters, National Bureau of Economic Research, Inc, 2014.

- N. Li, T. Li, and S. Venkatasubramanian.
  t-closeness: Privacy beyond k-anonymity and l-diversity.
  In IEEE 23rd International Conference on Data Engineering, 2007. ICDE 2007, pages 106–115, 2007.

- Grigorios Loukides, Aris Gkoulalas-Divanis, and Bradley Malin.
  Anonymization of electronic medical records for validating genome-wide association studies.
  Proceedings of the National Academy of Sciences, 107(17):7898–7903, 2010.

- L. Low, S. King, and T. Wilkie.
  Genetic discrimination in life insurance: empirical evidence from a cross sectional survey of genetic support groups in the United Kingdom.
  BMJ, 317(7173):1632–1635, 1998.

- Jeantine E. Lunshof, Ruth Chadwick, Daniel B. Vorhaus, and George M. Church.
  From genetic privacy to open consent.
  Nature Reviews Genetics, 9(5):406–411, 2008.

References IV

- Ashwin Machanavajjhala, Daniel Kifer, Johannes Gehrke, and Muthuramakrishnan Venkitasubramaniam.
  L-diversity: Privacy beyond k-anonymity.
  ACM Trans. Knowl. Discov. Data, 1(1), 2007.

- Bradley Malin.
  Re-identification of familial database records.
  AMIA Annual Symposium Proceedings, 2006:524–528, 2006.

- Lukasz Olejnik, Kutrowska Agnieszka, and Claude Castelluccia.
  I'm 2.8% Neanderthal – the beginning of genetic exhibitionism?
  In Workshop on Genomic Privacy, Amsterdam, 2014.

- PGP Consortium, George Church, Catherine Heeney, et al.
  Public access to genome-wide data: Five views on balancing research with privacy and protection.
  PLOS Genet, 5(10):e1000665, 2009.

- Laura L. Rodriguez, Lisa D. Brooks, Judith H. Greenberg, and Eric D. Green.
  The complexities of genomic identifiability.
  Science, 339(6117):275–276, 2013.

- Sriram Sankararaman, Guillaume Obozinski, Michael I. Jordan, and Eran Halperin.
  Genomic privacy and limits of individual detection in a pool.
  Nature Genetics, 41(9):965–967, 2009.

- Eric E. Schadt, Sangsoon Woo, and Ke Hao.
  Bayesian method to predict individual SNP genotypes from gene expression data.
  Nature Genetics, 44(5):603–608, 2012.

- Latanya Sweeney.
  K-anonymity: A model for protecting privacy.
  Int. J. Uncertain. Fuzziness Knowl.-Based Syst., 10(5):557–570, 2002.

References V

- Latanya Sweeney, Akua Abu, and Julia Winn.
  Identifying participants in the personal genome project by name.
  SSRN Scholarly Paper ID 2257732, Social Science Research Network, Rochester, NY, 2013.

- Peter M. Visscher and William G. Hill.
  The Limits of Individual Identification from Sample Allele Frequencies: Theory and Statistical Analysis.
  PLOS Genet, 5(10):e1000628, 2009.

- D. Vu and A. Slavkovic.
  Differential privacy for clinical trial data: Preliminary evaluations.
  In IEEE International Conference on Data Mining Workshops, 2009, pages 138–143, 2009.

- Rui Wang, Yong Fuga Li, XiaoFeng Wang, Haixu Tang, and Xiaoyong Zhou.
  Learning your identity and disease from research papers: Information leaks in genome wide association study.
  In Proceedings of the 16th ACM Conference on Computer and Communications Security, pages 534–544, New York, NY, USA, 2009. ACM.

- Jun Zhang, Zhenjie Zhang, Xiaokui Xiao, Yin Yang, and Marianne Winslett.
  Functional mechanism: Regression analysis under differential privacy.
  Proc. VLDB Endow., 5(11):1364–1375, 2012.