

PRIVACY HORIZONS: TERRA INCOGNITA

29th International Conference of
Data Protection and Privacy Commissioners

September 25 to 28, 2007
Montreal, Canada



LES HORIZONS DE LA PROTECTION DE LA VIE PRIVÉE : TERRA INCOGNITA

29^e Conférence internationale des commissaires
à la protection des données et de la vie privée

du 25 au 28 septembre 2007
Montréal, Canada

De-identification challenges raised by genetic and genomic data

William W. Lowrance, PhD
(lowrance@iprolink.ch)

September 26, 2007

The physical basis of the challenges

The human genome:

- is extensive and very fine-grained
- influences many personal attributes
- is intrinsic to the body
- doesn't change during the lifetime
- is unique to the individual.

The full genome is carried by the DNA in every cell of the body (except red blood cells).

What genomic data look like

...tttccgtatgcgtagccagactaccctcctagtag...

- through 3,000,000,000 "data-cells," each carrying a/t/g/c.

Altering or inserting just a few a/t/g/c can make a big difference, whether the genome is being considered:

- as a dynamic program-tape, or
- as an intrinsic "barcode."

What genetic data look like

- at sequence scale: |ctag...ctcca|
- at gene scale: "Diabetes-factor gene SLC308A"
- at body scale: "red hair," "heritable renal dysplasia"
- at family scale: pedigree, family health history, other indicators.

The most useful construal of
identifiability for genomic data,
in my view

*"Identifiability" is the potential
associability of data with persons.*

Paths through which genomic data can become identified

- (a) matching genotype to identifiable reference genotype data (such as police, military, or blood-relatives')
- (b) linking genomic+associated data (health, social, etc) with other data
- (c) profiling, i.e. probabilistically describing likely appearance, health factors, or other traits.

Tactics for de-identifying genomic data

- (a) limiting the proportion of genome released
- (b) statistically degrading the data before releasing
- (c) irreversibly de-identifying
- (d) separating the identifiers and key-coding.

Tactic (a): limiting the proportion of genome released

- is done, and can protect
- but often limits usefulness, because often it isn't known in advance which portions of genome are relevant
- difficult to judge how much is "not too much" to release.

Tactic (b): statistically degrading the data before releasing

- can be done, such as by randomly substituting some a/t/g/c
- almost always degrades usefulness, because most analyses depend on precise fine details.

Tactic (c): irreversibly de-identifying

- is occasionally done, such as when the purpose is to survey the background occurrence of some phenomenon, or to provide data for educational use.

Tactic (d): separating the identifiers and key-coding

- works well – if performed carefully, the key is properly safeguarded, and use of the key to reconnect is strictly controlled
- is increasingly being used in activities such as health research.

To de-identify, or not?

Whether and in what ways to de-identify genomic data depends on the:

- character of the data
- consent
- intended uses
- potential for linking to reference genotype or other data
- protections.

Alternatives and complements to de-identification

- Provide access via controlled release (governed by contract, overseen by a stewardship committee, etc)
- Sanction against misuse of the data (such as improper re-identifying) or abuse using the data (such as negative discrimination).

Closing sermon

De-identification is a crucial, practical protection – for both genomic and other kinds of data – and its use must be strongly encouraged!

General ref: Lowrance and Collins,
"Identifiability in genomic research,"
Science 317, 600–602 (August 3, 2007).