gmds · LEIPZIG 2006

GMDS Jahrestagung 2006, Leipzig

Data-related Challenges of Genotype - Phenotype studies

JProf Dr. Ulrich Sax,
M.Sc. Yassene Mohammed

Abteilung Medizinische Informatik
CIOffice Medizinische Forschungsnetze
Bereich Humanmedizin
Universität Göttingen
usax@med.uni-goettingen.de

Introduction and Question

- Genome wide association studies
  - ✓ genotyping constantly gets cheaper
  - ✓ many formerly phenome-related projects around complex diseases consider genotyping within the next couple of years.
- challenges
  - A) sufficient case/control numbers
  - B) high quality patient material to be genotyped
  - C) high quality phenotype data
  - D) how can we homogenize the heterogeneous data sources?
  - E) privacy problems!

C) Phenotype data?

- many disease related data collections in the form of registries (phenotype) or biomaterial collections (poss. genotyping)
- Biomaterial banks is likewise scattered due to ethical and legal reasons
- genotyping prices constantly go down, (good) phenotypic annotation is expensive
- The phenotype data could either be captured via
  - ✓ clinical trials
  - ✓ phenotype data from hospital data bases - taking into account the quality uncertainty
  - ✓ phenotype data from disease-related registries.

D) Homogenization

- The challenge starts as soon as genotype and related phenotype data from different data sources are available for further analysis.
  - ✓ Genomic data tends to be more structured than phenotype data, Bioinformatics community is open source and XML based
  - ✓ Phenotype data is kept mostly in traditionally “hand carved”, non-compatible Information systems

- For association studies not only the **data formats** have to be homogenized, more importantly the **content** has to be homogenized. Ontology-approaches using UMLS had some success recently

Butte, A.J. and I.S. Kohane, *Creation and implications of a phenome-genome network.* Nat Biotechnol, 2006. 24(1): p. 55-62.

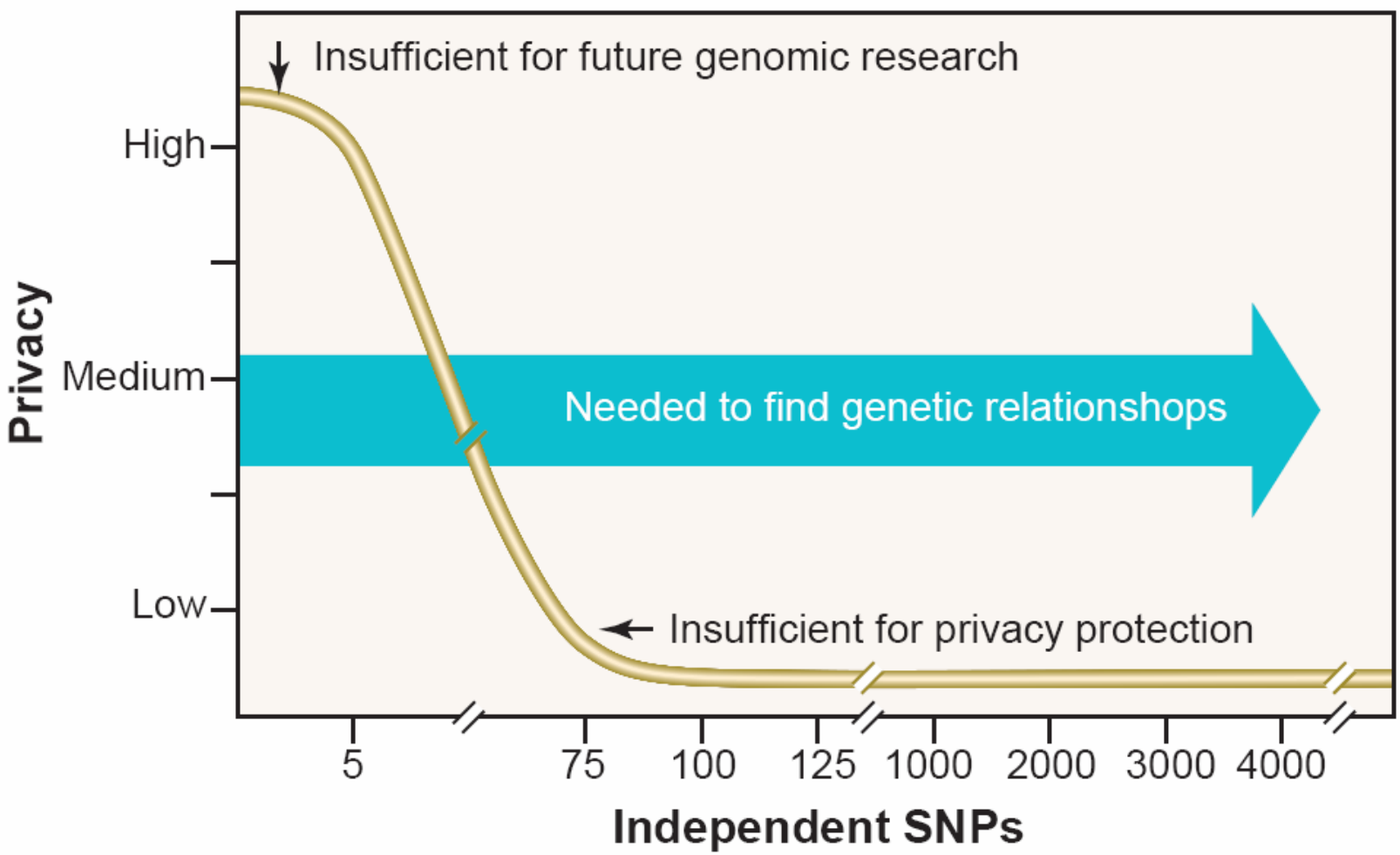
Microarray Data Transfer

Workflow | Array-Data | Database

Removed due to copyright reasons

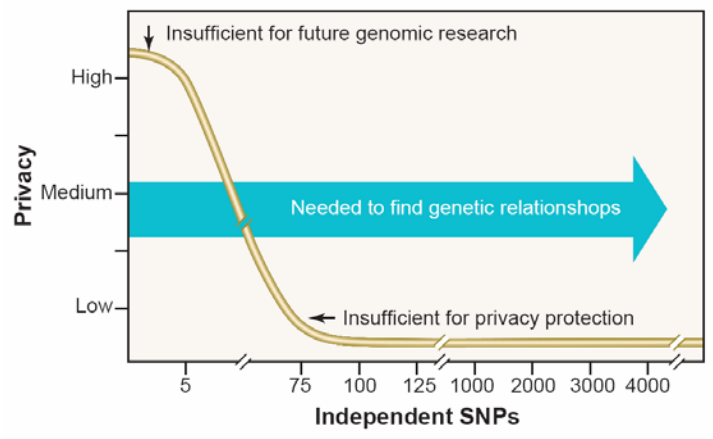
Courtesy: C. Lawerenz, NGFN

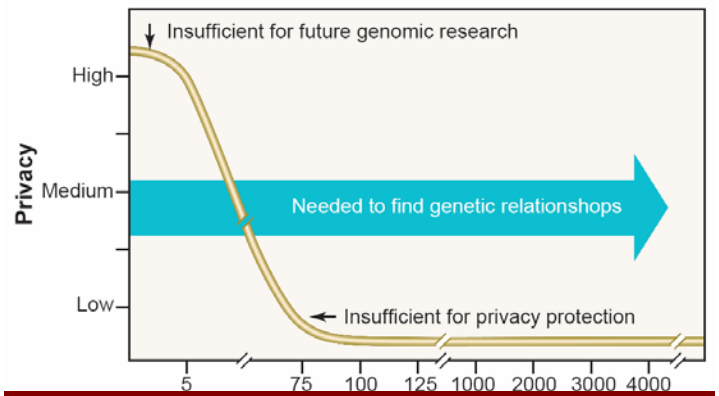
Example: Trade-offs between Single Nucleotide Polymorphisms (SNPs) and Privacy



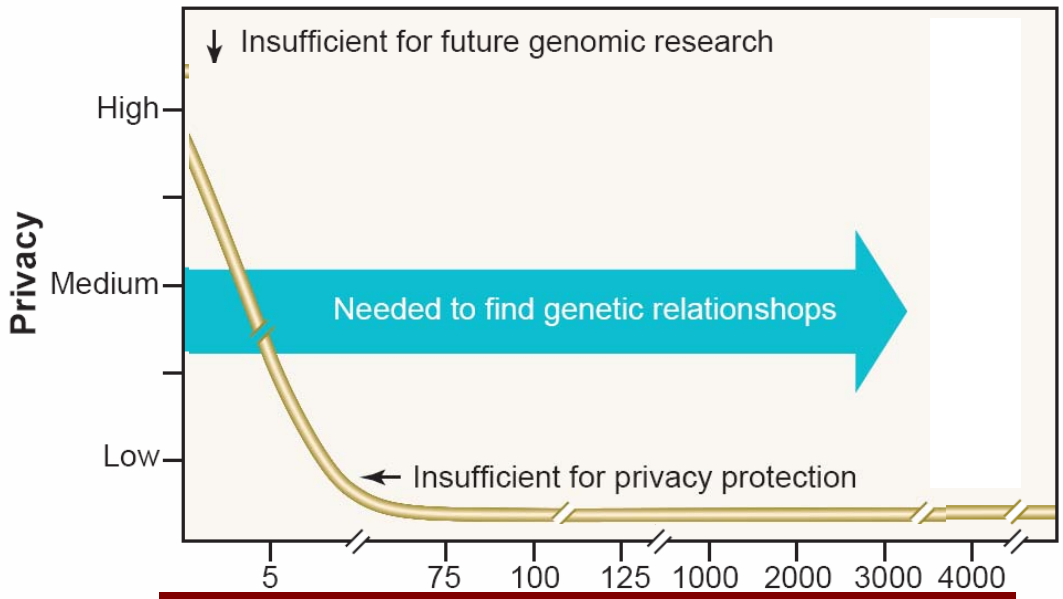
Lin Z, Owen AB, Altman RB. Science. Jul 9 2004;305(5681):183

Privacy concerns: Combination of data





Phenotype data



Combination of Genotype and Phenotype data

E) Privacy?!

Isaac Kohane (i2B2, Boston):

- **The cat is already out of the bag!**
- Most people are not fully aware of the degree to which their blood samples **can be and are used by the pharmaceutical industry**
- (When they learn, some wonder if they should participate in the profits that result from their samples)
- Insurance companies, hospitals **routinely share data** for reimbursement and research.

Genomic Privacy – possible solutions

- Competence networks have to be validated by TMF privacy WG and the federal Privacy Officers
- Privacy in Grid-Computing is a challenge

- Pseudonymization naïve de-identification do not suffice dealing with sequencing and SNP-data!
- K-Anonymity: each data set does not differ from k-1 other data sets (information loss!)
- Methods for giving **patient control of health data disclosure** could provide a mechanism for gathering research data (PHR)

Statistical Disclosure Control (SDC)

- **Microdata**: files with individual observations;
- **Tables**: total values carried by individuals
- **Other statistics** : summaries of different, indices, correlation coefficients, etc.
- → trade-off between the level of protection achieved for the data and the quality of the information released.
- **Information loss**:

- **How (not) to protect genomic data privacy in a distributed network: using trail re-identification to evaluate and design anonymity protection systems** *Journal of Biomedical Informatics, Volume 37, Issue 3, June 2004, Pages 179-192* Bradley Malin and Latanya Sweeney

Elements of Statistical Disclosure Control
Series: Lecture Notes in Statistics , Vol. 155
Willenborg, Leon, **Waal**, Ton de

Discussion

- Given the necessity to capture both environment and genomic state of a patient and their interaction, clinical information systems have to be redesigned.
- More integration work on terminologies and ontologies is to be done.
- Researchers from **medical informatics, bioinformatics and epidemiology** will have to **collaborate** much more intensively than they formerly did.
- One of the main problems may be the different vocabulary and the different background of these researchers.
- Sustainable collaborations would give German Biomedical Informatics a competitive edge in the community.

Take Home Message

- Genome wide association studies
  - ✓ genotyping constantly gets cheaper, phenotypic anotation drives the price!
- challenges
  - A) sufficient case numbers
  - B) high quality patient material to be genotype
  - C) high quality phenotype data
  - D) how can we homogenize the heterogeneous data sources
  - E) privacy problems!
- Researchers from **medical informatics, bioinformatics and epidemiology** will have to **collaborate** much more intensively

Sponsor

- This work was supported by the D-Grid Project **MediGRID**, funded by the Federal Ministry of Education and Research (BMBF), FKZ 01AK803H,
- and by the **Competence Network for Congenital Heart Disease (AHF),** funded by the Federal Ministry of Education and Research (BMBF), FKZ 01G10210.

Bundesministerium
für Bildung
und Forschung

gmds · LEIPZIG 2006

GMDS Jahrestagung 2006, Leipzig

Data-related Challenges of Genotype - Phenotype studies

JProf Dr. Ulrich Sax,
M.Sc. Yassene Mohammed

Abteilung Medizinische Informatik
CIOffice Medizinische Forschungsnetze
Bereich Humanmedizin
Universität Göttingen
usax@med.uni-goettingen.de

References

- [1] Westphal, S.P. *Race for the $1000 genome is o*n. 2002 [cited; Available from: http://www.newscientist.com/article.ns?id=dn2900.
- [2] Greeley, M. *Nothing ventured ... Two Cents on the '$1,000 Genome*'. 2006 [cited 09.04.2006]; Available from: http://www.bio-itworld.
- com/archive/081303/ventured.html.
- [3] Powell, J. and I. Buchan, *Electronic health records should support clinical research.* J Med Internet Res, 2005. 7(1): p. e4.
- [4] Dumitru, R.C. and O. Rienhoff, *Challenges to Patients Medical Records Supporting Clinical Research – Data Quality*, in *1st International Conference on*
- *Information Communication Technologies in Healt*h. 2003: Samos island, Greece.
- [5] MediGRID. *Medical Grid Computing*. 2005 [cited; Available from: www.medigrid.de.
- [6] Lin, Z., A.B. Owen, and R.B. Altman, *Genetics. Genomic research and human subject privacy.* Science, 2004. **30**5(5681): p. 183.
- [7] Malin, B.A., *An Evaluation of the Current State of Genomic Data Privacy Protection Technology and a Roadmap for the Future.* J Am Med Inform Assoc, 2004.
- [8] KN-AHF. *Kompetenznetz Angeborene Herzfehle*r. 2005 [cited 21.09.2005]; Available from: www.kompetenznetz-ahf.de.
- [9] TMF-AG_Biomaterial. 2006 [cited 09.04.2006]; Available from: http://www.tmf-ev.de/site/DE/int/AG/BMB/container_bmb.php.
- [10] HL7, *HL7 Receives ANSI Approval of Three Version 3 Specifications Including CDA, Release 2.* 2005.
- [11] Dolin, R.H., et al. *HL7 Clinical Document Architecture (Release 2.0*). 2004 [cited February 9, 2005]; Committee Ballot #3; Aug 03,2004:[Available from:
- http://hl7.org/library/Committees/structure/CDA.ReleaseTwo.CommitteeBallot03.Aug.2004.zip.
- [12] Bodenreider, O., *The Unified Medical Language System (UMLS): integrating biomedical terminology.* Nucleic Acids Res, 2004. **32 Database issu**e: p. D267-70.
- [13] Butte, A.J. and I.S. Kohane, *Creation and implications of a phenome-genome network.* Nat Biotechnol, 2006. **2**4(1): p. 55-62.
- [14] TMF. *AG Datenschutz*. 2006 [cited 16.03.2006]; Available from: http://www.tmf-ev.de/site/DE/int/AG/DS/container_ag_ds.php.