Machine Learning for Personalised Medicine

James McMurray

PhD Student
Department of Empirical Inference

04/02/2014

Outline of talk

MAX-PLANCK-GESELLSCHAFT

What is "Personalised Medicine"?

- Any application of personalisation in healthcare.
- Become much more popular recently, resulting in a broad term, like "Big Data".
- Main application is the use of increasing amounts of genetic data.
- Recently made headlines with Angelina Jolie's masectomy on account of carrying the BRCA1 gene.

14 May 2013 Last updated at 17:02 GMT

Share

Angelina Jolie has double mastectomy due to cancer gene

COMMENTS (393)

The BBC's Fergus Walsh explains the background to Angelina Jolie's decision

Hollywood actress Angelina Jolie has undergone a double mastectomy to reduce her chances of getting breast cancer.

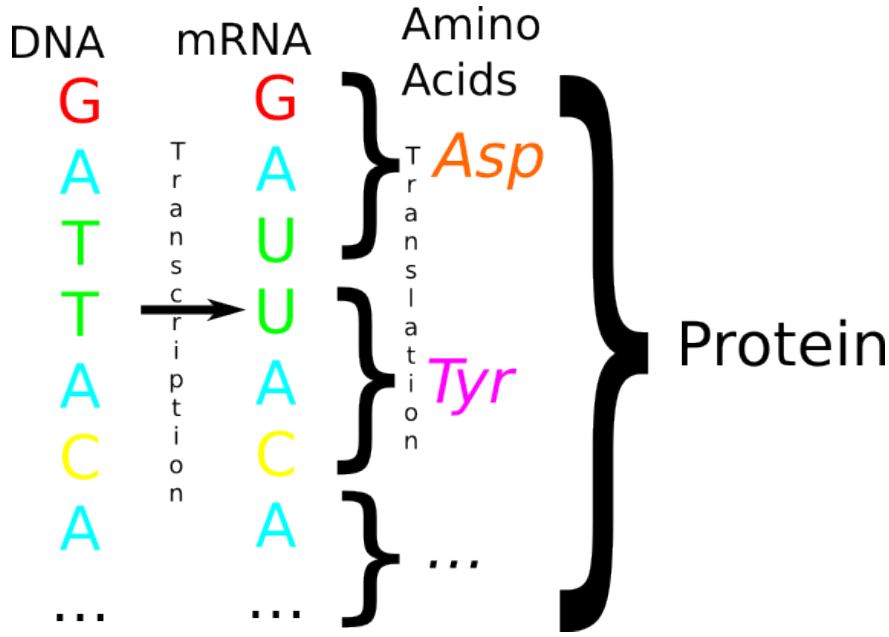
Related Stories

BBC News, 14 May 2013

MAX-PLANCK-GESELLSCHAFT

Genetics Recap

- DNA is made up of 4 bases: Adenine (A), Guanine (G), Thymine (T) and Cytosine (C).
- Triplets of bases (codons) code for amino acids.
- DNA is transcripted to mRNA, which is translated to produce amino acids.
- Genes are sequences of DNA which code for proteins (sequences of amino acids).
- The amount of protein produced from a gene is termed gene expression



MAX-PLANCK-GESELLSCHAFT

Genetics Recap: Important terms

Genotype: The genetic composition of an organism. I.e. the entire sequence of genetic bases which make up the organism's DNA.

Epigenotype: Heritable patterns of DNA methylation, and modifications to chromatin proteins that package DNA - resulting in stable changes in gene expression.

Environment: The physical conditions which the organism is subject to. For example, diet or exposure to allergens.

Phenotype: Observable traits of an organism, resulting from its genotype, epigenotype, environment and interactions thereof. I.e. height, or the presence of a disease.

SNP: A Single Nucleotide Polymorphism - a mutation of a single base in the genome often altering gene function.

MAX-PLANCK-GESELLSCHAFT

What is Personalised Medicine: Main tasks

- Target drugs to specific groups based on genetic compatibility:
  - Drug discovery is very expensive, it would be useful to still be able to market drugs which only work with certain subpopulations.
  - For example, *Elitek*, a drug to treat tumor lysis syndrome, requires a genetic test before prescription to ensure the patient does not have glucose-6-phosphate dehydrogenase deficiency which would result in a dangerous accumulation of peroxides[1].
  - Warfarin is an anticoagulant, also used as rat poison. Excess Warfarin can cause internal bleeding, and the necessary dosage depends upon the levels of the enzyme vitamin K epoxide reductase (VKOR). The VKORC1 gene encodes this enzyme and variations of this gene will affect the optimal warfarin dosage[1].

[1] Gillham, *Genes, Chromosomes, and Disease*, 2011

What is Personalised Medicine: Main tasks

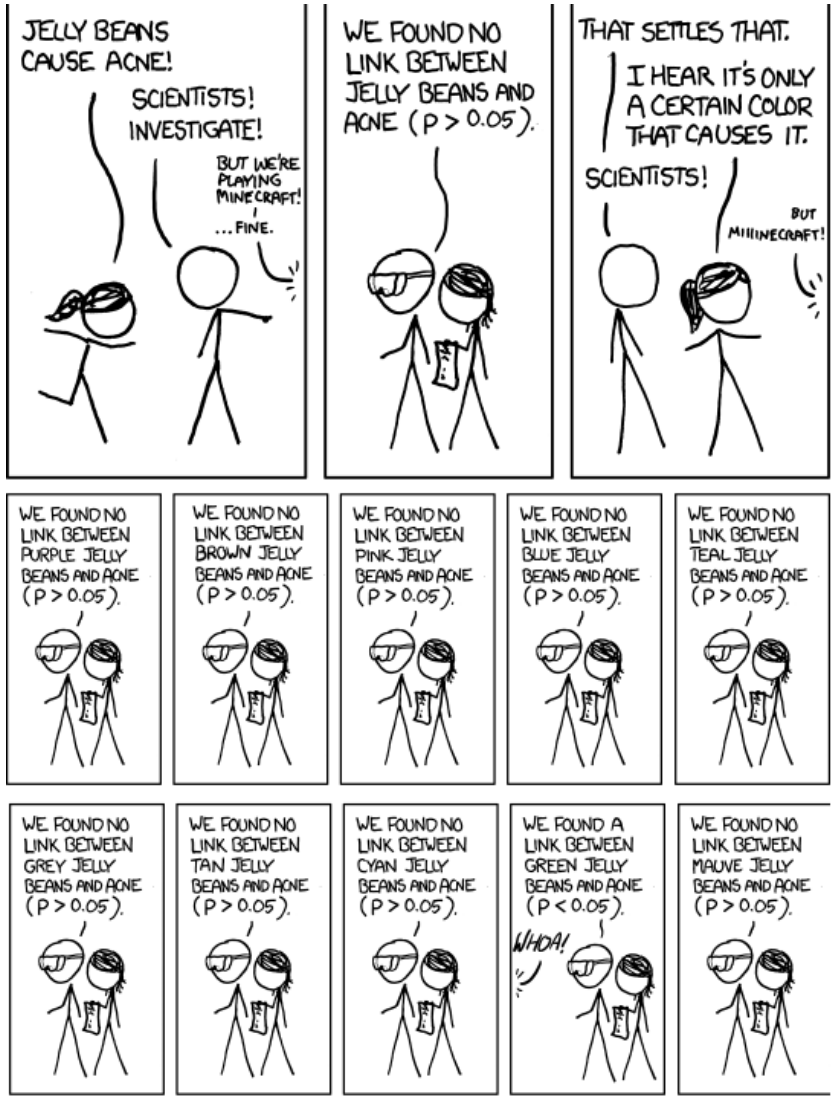
- Provide recommendations/diagnosis based on the combination of environmental and genetic information:
  - Much more data available with new direct-to-consumer genetic sequencing, and the growth of digitised medical and shopping records and smart watches providing greater environmental information.
  - QuantifiedSelf online community tracks and analyses personal data for health and fitness.
  - Fitoop is an example of a new startup attempting to usefully combine the large amounts of different data.
  - IBM's Watson is applying hypothesis generation to diagnosis.
  - Clear applications of Machine Learning.

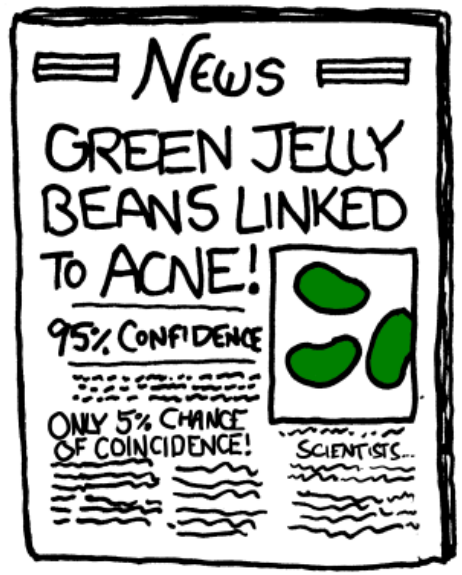
MAX-PLANCK-GESELLSCHAFT

What is Personalised Medicine: Main tasks

- Discover causal genetic, epigenetic and environmental interactions for phenotypic traits:
  - Main aim of scientific research, for example in Genome Wide Association Studies (GWASs).
  - Must beware of confounders and false positives.
  - For example, in GWASs it is standard to use Linear Mixed Models.
  - Involves a large number of statistical tests.
  - Must correct for multiple hypothesis testing.
  - I.e. if you have 1000 fair coins, and flip them each ten times, there is a $\sim 98\%$ chance that at least one coin will land 10 heads and appear significantly biased.
  - Population stratification is a confounding problem, where cases and controls are drawn from different populations. I.e. Chopstick gene.

MAX-PLANCK-GESELLSCHAFT



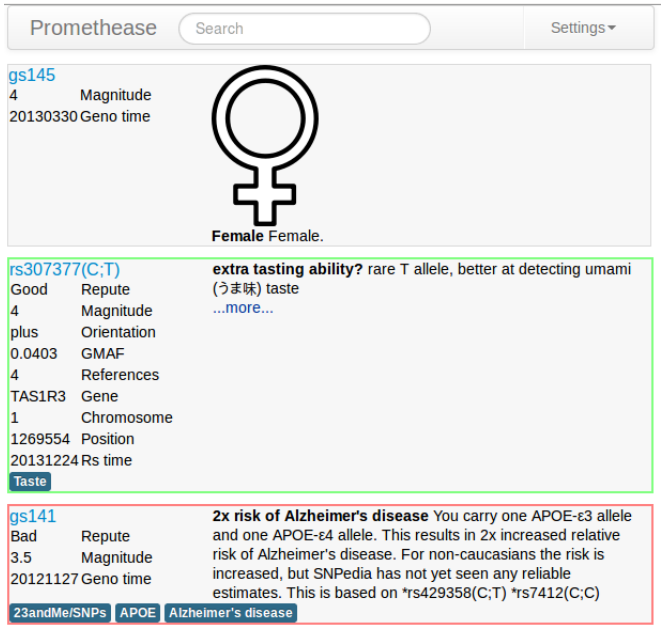


XKCD: Significant
(Edited to fit)

MAX-PLANCK-GESELLSCHAFT

What is Personalised Medicine: Main tasks

- Predicting phenotype from genotype:
  - Not concerned with causality or confounders, just correlation. Like an insurance company for example.
  - Often used in selective breeding of cattle and plants.
  - Promethease is an online tool which uses SNPedia, an online wiki of SNPs and their posited correlations, to produce a profile for a user given their sequence data.

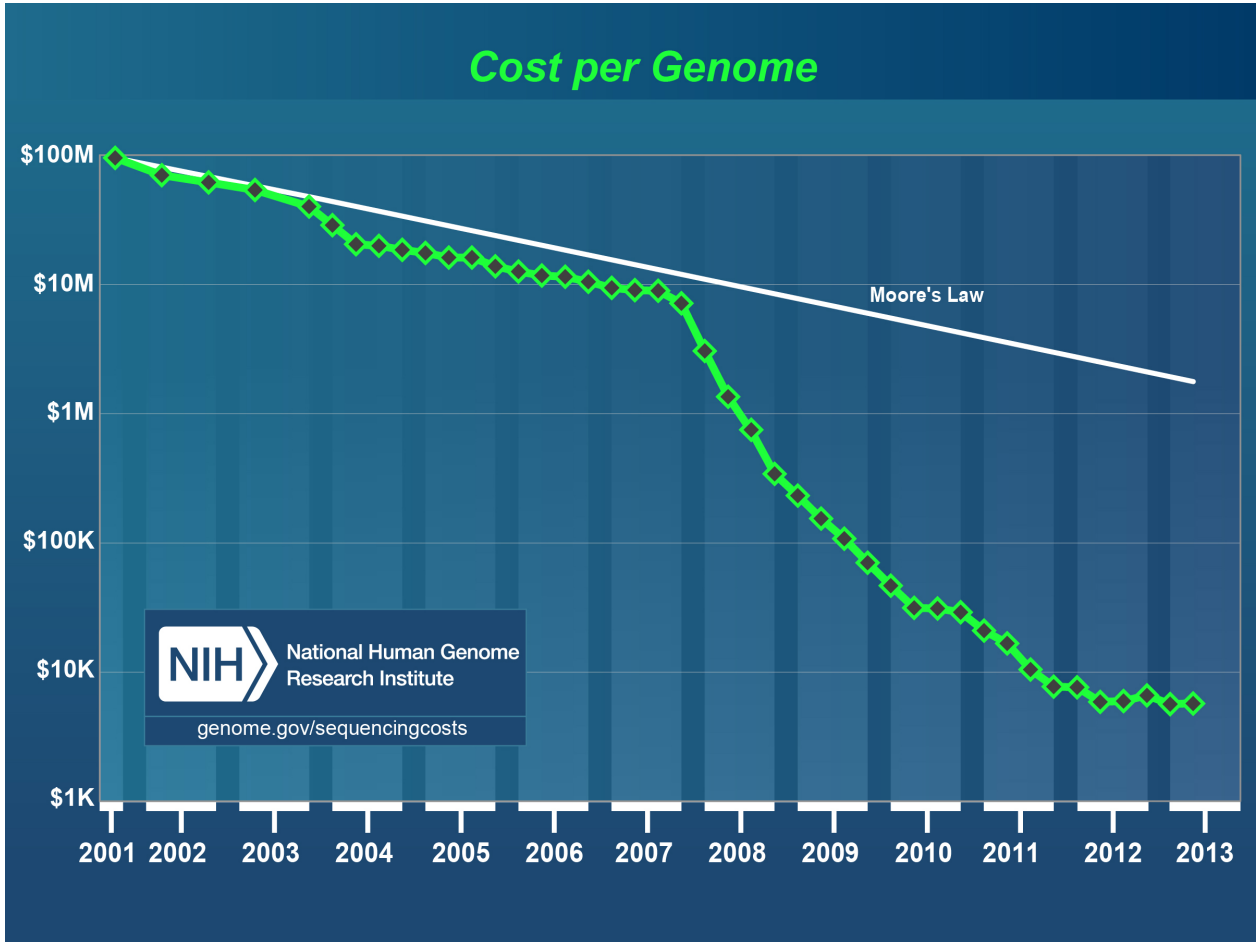


Promethease example

MAX-PLANCK-GESELLSCHAFT

The decreasing cost of sequencing

- Genetic sequencing has become much cheaper:



Data from the NHGRI Genome Sequencing Program (GSP) up to April 2013:
http://www.genome.gov/sequencingcosts/

MAX-PLANCK-GESELLSCHAFT

Applications of Machine Learning in Research

Machine Learning has applications in the two major types of genetic studies:

Genome Wide Association Studies (GWASs):

Look at the association of SNPs (single base mutations) with phenotypes, to produce a "risk factor" if a significant association is found.

Expression Quantitative Trait Loci (eQTL) mapping:

Gene expression can be regulated by genetic loci (bases in the genome) close (cis) or far (trans) from the gene itself. eQTL mapping studies attempt to discover which loci regulate the expression of which genes.

Other possible applications such as protein function prediction, drug discovery, etc.

My current work: Application of Causal Inference

- Fusi, Stegle *et al.* developed PANAMA[2] model (Probabilistic ANAlysis of genoMic dAta) for eQTL studies, which uses the Gaussian Process Latent Variable Model (GPLVM) to model confounding factors.
- Zhang *et al.* developed IGPLVM[3] - permits interpretation of causal relations between observed variables, by allowing arbitrary noise correlations between the latent variables.
- Currently attempting to implement IGPLVM in GPy.

[2] N. Fusi, O. Stegle and N. D. Lawrence, "Joint modelling of confounding factors and prominent genetic regulators provides increased accuracy in genetical genomics studies", PLoS Computational Biology, 2012

[3] Zhang, K., Schölkopf, B., and Janzing, D. (2010). "Invariant Gaussian Process Latent Variable Models and Application in Causal Discovery". UAI 2010.

MAX-PLANCK-GESELLSCHAFT

Interesting problems

- Missing data not at random - Sherlock Holmes' curious incident
- Combining different data sources and propagating uncertainty
- Many more SNPs than sequences: $p >> n$ - can we incorporate prior information from known interaction pathways, etc.?
- Scalable algorithms
- Unobserved confounders - population stratification, environmental effects, etc.
- Is genomics a social science?[4]
  - Large number of potential causes, individually small in their effects.
  - Causes are non-independent and non-additive.
  - Randomized experimentation is not possible.

[4] Eric Turkheimer, "Genome Wide Association Studies of Behavior are Social Science", *Philosophy of Behavioral Biology*

Conclusion

- Growing area of research with many applications for Machine Learning.
- Lots of industrial interest:
  - Pharmatics, IBM Watson, Fitoop, HealthTech challenge
- Need to avoid "Genetic Horoscope": 23AndMe
- Want to take part?
  - Can obtain some data from the Personal Genomes Project, sometimes Kaggle
  - DIYgenomics and Genomera are attempting to "crowdsource" studies.
  - DREAM challenges (Dialogue for Reverse Engineering Assessments and Methods).

Questions?

MAX-PLANCK-GESELLSCHAFT