Statistical Genomics

Some very basic considerations on the design of experiments and the interpretation of results

Why do we look at genomic data?

- Learn about basic biology.
- Identify drug targets.
- Find biomarkers.
  - Disease risk prediction.
  - Early detection of disease onset.
Bioinformatics and Computational Biology are super-exciting fields to be in! With tons of genomic data being generated, this is a great time to use those skills for clinical and translational research (headed towards personalized medicine).

However, there are some aspects to all of this that are less than super-exciting. Sometimes they get lost in all the hype.

► Even the best device can have poor predictive performance.
► Poor experimental design is common, and can easily do people in. The Hall of Shame is well populated.
► Mistakes are easy to make with these high dimensional data, even with the best of intentions.
► It is not unusual that the technical artifacts in the genomic data are much larger than any biological signal.
► Quite frequently, you do have a “needle in the haystack” problem. The haystack will be hard to move, and your barn might not be large enough for the hay.
► Meet the curse of dimensionality!

Assume you identified a gene signature that predicts the early onset of disease with 99% sensitivity and 99% specificity. What is the probability of a person having the disease given the result is positive, if we randomly select a subject from

► the general population with 0.1% disease prevalence?
► a high risk sub-population with 10% disease prevalence?
Sensitivity  \(\rightarrow\) \(\Pr(\text{positive test} \mid \text{disease})\)
Specificity  \(\rightarrow\) \(\Pr(\text{negative test} \mid \text{no disease})\)
Positive Predictive Value  \(\rightarrow\) \(\Pr(\text{disease} \mid \text{positive test})\)
Negative Predictive Value  \(\rightarrow\) \(\Pr(\text{no disease} \mid \text{negative test})\)
Accuracy  \(\rightarrow\) \(\Pr(\text{correct outcome})\)
Sensitivity → $\frac{99}{(99+1)} = 99\%$
Specificity → $\frac{98901}{(999+98901)} = 99\%$
Positive Predictive Value → $\frac{99}{(99+999)} \approx 9\%$
Negative Predictive Value → $\frac{98901}{(1+98901)} > 99.9\%$
Accuracy → $\frac{(99+98901)}{100000} = 99\%$
Sensitivity  \( \rightarrow \frac{9900}{(9900+100)} = 99\% \)
Specificity  \( \rightarrow \frac{89100}{(900+89100)} = 99\% \)
Positive Predictive Value  \( \rightarrow \frac{9900}{(9900+900)} \approx 92\% \)
Negative Predictive Value  \( \rightarrow \frac{89100}{(100+89100)} \approx 99.9\% \)
Accuracy  \( \rightarrow \frac{(9900+89100)}{100000} = 99\% \)
CD Genomics offers genetic testing panel which is based on a technology that assesses a complex but specific set of sites on the human genome -- Single Nucleotide Polymorphisms (SNPs) -- which determines an individual's likelihood of disease.

What kinds of diseases are you susceptible to?

<table>
<thead>
<tr>
<th>GenSeq™ Disease Susceptibility Panel</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>Cancers (15)</strong></td>
</tr>
<tr>
<td>Breast/Ovarian Cancer, Colorectal Cancer, Pancreatic Cancer, Endometrial Cancer, Esophageal Cancer, Renal Cancer, Bladder Cancer, Prostate Cancer, Hodgkin's Lymphoma, Follicular Lymphoma, Chronic Lymphocytic Leukemia, Meningioma, Abdominal Aortic Aneurysm, Melanoma</td>
</tr>
<tr>
<td><strong>Cardiovascular Diseases (3)</strong></td>
</tr>
<tr>
<td>Hypertension, Coronary Heart Disease, Venous Thromboembolism.</td>
</tr>
<tr>
<td><strong>Neurological Diseases (3)</strong></td>
</tr>
<tr>
<td>Parkinson's disease, Multiple Sclerosis, Alzheimer's Disease</td>
</tr>
<tr>
<td><strong>Metabolic Disease (4)</strong></td>
</tr>
<tr>
<td>Obesity, Gout, Kidney Stones, Gallstones</td>
</tr>
<tr>
<td><strong>Immune System Diseases (3)</strong></td>
</tr>
<tr>
<td>Type I Diabetes, Asthma, Rheumatoid Arthritis</td>
</tr>
<tr>
<td><strong>Endocrine Diseases (3)</strong></td>
</tr>
<tr>
<td>Type II Diabetes, Endometriosis, Hypothyroidism.</td>
</tr>
<tr>
<td><strong>Inflammation (3)</strong></td>
</tr>
<tr>
<td>Chronic Kidney Disease, Ankylosing Spondylitis, Chronic Obstructive Pulmonary Disease</td>
</tr>
</tbody>
</table>

By identifying your carrier status for mutations linked to 34 common diseases' susceptibility, we provide you and your family with the knowledge to help you prepare for the future.

By knowing more about your underlying health risks, you and your doctor can make more informed decisions about your healthcare.
23andMe Bows To FDA's Demands, Drops Health Claims

by SCOTT HENSLEY

December 06, 2013  1:49 PM ET

The maker of a $99 personal genome test blinked.

Silicon Valley's 23andMe said late Thursday that it would comply with the Food and Drug Administration's demand that the company stop marketing health-related genetic tests.

People will still be able to pay 23andMe to have their DNA analyzed to learn about their ancestors. And customers will get a file of their raw genetic info.

But while 23andMe tries to work things out with the FDA, the company won't be telling people they have genetic profiles that predispose them to particular illnesses, or predict their responses to prescription drugs. In other words, no more health claims.

"We remain firmly committed to fulfilling our long-term mission to help people everywhere have access to their own genetic data and have the ability to use that information to improve their lives," Anne Wojcicki, co-founder and CEO of 23andMe, said in a statement. "Our goal is to work cooperatively with the FDA to provide that opportunity in a way that clearly demonstrates the benefit to people and the validity of the science that underlies the test."

The company has a pretty deep regulatory hole to crawl out of.
Association versus Prediction
Week One

<table>
<thead>
<tr>
<th>M</th>
<th>Tu</th>
<th>W</th>
<th>Th</th>
<th>F</th>
</tr>
</thead>
<tbody>
<tr>
<td>C</td>
<td>T</td>
<td>T</td>
<td>T</td>
<td>T</td>
</tr>
<tr>
<td>T</td>
<td>C</td>
<td>C</td>
<td>C</td>
<td>T</td>
</tr>
<tr>
<td>C</td>
<td>C</td>
<td>C</td>
<td>T</td>
<td>C</td>
</tr>
<tr>
<td>T</td>
<td>T</td>
<td>T</td>
<td>C</td>
<td>C</td>
</tr>
</tbody>
</table>

Week Two

<table>
<thead>
<tr>
<th>M</th>
<th>Tu</th>
<th>W</th>
<th>Th</th>
<th>F</th>
</tr>
</thead>
<tbody>
<tr>
<td>T</td>
<td>T</td>
<td>T</td>
<td>C</td>
<td>T</td>
</tr>
<tr>
<td>C</td>
<td>C</td>
<td>C</td>
<td>T</td>
<td>C</td>
</tr>
<tr>
<td>C</td>
<td>C</td>
<td>C</td>
<td>T</td>
<td>T</td>
</tr>
<tr>
<td>T</td>
<td>T</td>
<td>T</td>
<td>C</td>
<td>C</td>
</tr>
</tbody>
</table>

T = treated, C = control, pink = female, blue = male
MECHANISMS OF DISEASE

Mechanisms of disease

Use of proteomic patterns in serum to identify ovarian cancer


Summary

Background New technologies for the detection of early-stage ovarian cancer are urgently needed. Pathological changes within an organ might be reflected in proteomic patterns in serum. We developed a bioinformatics tool and used it to identify proteomic patterns in serum that distinguish neoplastic from non-neoplastic disease within the ovary.

Methods Proteomic spectra were generated by MALDI-TOF mass spectrometry (surface-enhanced laser desorption and ionization). A preliminary "training" set of spectra derived from analysis of serum from 50 unaffected women and 50 patients with ovarian cancer were analyzed by an iterative searching algorithm that identified a proteomic pattern that completely discriminated cancer from non-cancer. The discovered pattern was then used to classify an independent set of 118 masked serum samples: 50 from women with ovarian cancer, and 68 from unaffected women or those with non-malignant disorders.

Findings The algorithm identified a cluster pattern that, in the training set, completely segregated cancer from non-cancer. The discriminatory pattern correctly identified all 50 ovarian cancer cases in the masked set, including all 18 stage I cases. Of the 68 cases of non-malignant disease, 63 were recognized as not cancer. This result yielded a sensitivity of 100% (95% CI 93-100), specificity of 93% (87-99), and positive predictive value of 94% (84-99).

Interpretation These findings justify a prospective population-based assessment of proteomic pattern technology as a screening tool for all stages of ovarian cancer in high-risk and general populations.

Lancet 2002; 359: 572-77

Introduction

Application of new technologies for detection of ovarian cancer could have an important effect on public health, but to achieve this goal, specific and sensitive molecular markers are essential.12 This need is especially urgent in women who have a high risk of ovarian cancer due to family or personal history of cancer, and for women with a genetic predisposition to cancer due to abnormalities in predisposition genes such as BRCA1 and BRCA2. There are no effective screening options for this population.

Ovarian cancer presents at a late clinical stage in more than 80% of patients, and is associated with a 5-year survival of 33% in this population. By contrast, the 5-year survival for patients with stage I ovarian cancer exceeds 90%, and more than 90% of patients are cured of their disease by surgery alone.13 Therefore, increasing the number of women diagnosed with stage I disease should have a direct effect on the mortality and economics of this cancer without the need to change surgical or chemotherapeutic approaches.

Cancer antigen 125 (CA125) is the most widely used biomarker for ovarian cancer.14 Although concentrations of CA125 are abnormal in about 80% of patients with advanced-stage disease, they are increased in only 50-60% of patients with stage I ovarian cancer.15 CA125 has a positive predictive value of less than 10% as a single marker, but the addition of ultrasound screening to CA125 measurement has improved the positive predictive value to about 20%.16 Low-molecular-weight serum protein profiling might reflect the pathological state of organs and aid in the early detection of cancer. Matrix-assisted laser desorption and ionization time-of-flight (MALDI-TOF) and surface-enhanced laser desorption and ionization, time-of-flight (SELDI-TOF) mass spectrometry can profile
Genetic Signatures of Exceptional Longevity in Humans

Paola Sebastiani,1,4 Nadia Solovieff,4 Annibale Puca,2 Stephen W. Hartley,1 Ethlymania Melista,5 Stacy Andersen,2 Daniel A. Dworski,2 Jemima B. Wilk,2 Richard H. Myers,3 Martin H. Steinberg,3 Monty Montano,2 Clinton T. Baldwin,2,7 Thomas T. Perls8

1Department of Biostatistics, Boston University School of Public Health, Boston, MA 02118, USA. 2IRCCS Multimedica, Milano, Italy. Istituto di Tecnologie Biomediche, Consiglio Nazionale delle Ricerche, Segrato, Italy. 3Department of Medicine, Boston University School of Medicine, Boston, MA 02118, USA. 4Section of Geriatrics, Department of Medicine, Boston University School of Medicine and Boston Medical Center, Boston, MA 02118, USA. 5Department of Neurology, Boston University School of Medicine, Boston, MA 02118, USA. 6Departments of Medicine and Pediatrics, Boston University School of Medicine and Boston Medical Center, Boston, MA 02118, USA. 7Center for Human Genetics, Boston University School of Medicine, Boston, MA 02118, USA.

*To whom correspondence should be addressed. E-mail: psebas@bu.edu (P.S.); ttperls@bu.edu (T.T.P.)

Healthy aging is thought to reflect the combined influence of environmental factors and genetic factors. To explore the genetic contribution, we undertook a genome-wide association study of exceptional longevity (EL) in 1655 centenarians and 1267 controls. Using these data, we built a genetic model that includes 150 single nucleotide polymorphisms (SNPs) and found that it could predict EL with 77% accuracy in an independent set of centenarians and controls. Further in silico analysis revealed that 90% of centenarians can be grouped into 19 clusters characterized by different combinations of SNP genotypes—or genetic signatures—of varying predictive value. The different signatures, which attest to the genetic complexity of EL, could be used to develop personalized interventions for individuals at risk of developing age-related diseases such as dementia, hypertension, and cardiovascular disease.

Based upon the hypothesis that exceptionally old individuals are carriers of multiple genetic variants that influence human lifespan, we conducted a genome-wide association study (GWAS) of centenarians. Centenarians are a model of healthy aging, as the onset of disability in these individuals is generally delayed until they are well into their mid-nineties. We studied 801 unrelated subjects enrolled in the New England Centenarian Study (NECS) and found that the age distribution approximately matched controls. NECS subjects were Caucasian who were born between 1890 and 1910 and had an age range of 95 to 119 years (median age 103 years).

Figure S1 in the Supporting Online Material (7) describes the age distribution. Approximately one-third of the NECS sample included centenarians with a first-degree relative also achieving EL, thus enhancing the sample’s power (8). Controls included 243 NECS referent subjects who were spouses of centenarian offspring of children of parents who died at an age above 75 years, and genome-wide SNP data

Articles

Genomic signatures to guide the use of chemotherapeutics

Anil Potti1,2, Holly K Dressman1,3, Andrea Bild1,3, Richard F Riedel1,2, Gina Chan4, Robyn Sayer4, Janiel Cragun4, Hope Cottrill5, Michael J Kelley2, Rebecca Petersen5, David Harpole5, Jeffrey Marks5, Andrew Berchuck1,6, Geoffrey S Ginsburg1,2, Phillip Febo1,3, Johnathan Lancaster4 & Joseph R Nevins1,3

Using in vitro drug sensitivity data coupled with Affymetrix microarray data, we developed gene expression signatures that predict sensitivity to individual chemotherapeutic drugs. Each signature was validated with response data from an independent set of cell line studies. We further showed that many of these signatures can accurately predict clinical response in individuals treated with these drugs. Notably, signatures developed to predict response to individual agents, when combined, could also predict response to multidrug regimens. Finally, we integrated the chemotherapeutic response signatures with signatures of oncogenic pathway deregulation to identify new therapeutic strategies that make use of all available drugs. The development of gene expression profiles that can predict response to commonly used cytotoxic agents provides opportunities to better use these drugs, including using them in combination with existing targeted therapies.
Raw high throughput genomic data always contain artifacts. No exceptions. Really.

Identifying genomic signatures is a super hard problem. Technical artifacts in the data are often much larger than any biological signal.

Not addressing those artifacts can have nasty consequences, in particular when coupled with poor experimental design.

These artifacts include:

- Known systematic biases. For example genomic waves due to GC content.
- Random but possibly reproducible biases. For example laboratory specific artifacts.
- Random non-reproducible biases. For example plate and batch effects.
A comprehensive approach to the analysis of matrix-assisted laser desorption/ionization-time of flight proteomics spectra from serum samples

For our analysis of the data from the First Annual Proteomics Data Mining Conference, we attempted to discriminate between 24 disease spectra (group A) and 17 normal spectra (group B). First, we processed the raw spectra by (i) correcting for additive sinusoidal noise (periodic on the time scale) affecting most spectra, (ii) correcting for the overall baseline level, (iii) normalizing, (iv) recombining fractions, and (v) using variable-width windows for data reduction. Also, we identified a set of polymeric peaks (at multiples of 180.6 Da) that is present in several normal spectra (B1–B6). After data processing, we found the intensities at the following mass to charge (m/z) values to be useful discriminators: 3077, 12 886 and 74 283. Using these values, we were able to achieve an overall classification accuracy of 38/41 (92.6%). Perfect classification could be achieved by adding two additional peaks, at 2476 and 6955. We identified these values by applying a genetic algorithm to a filtered list of m/z values using Mahalanobis distance between the group means as a fitness function.

Keywords: Cross validation / Data cleaning / Discrimination / Genetic algorithm / Mahalanobis distance
Sinusoidal noise removal. Visual inspection of the raw spectra revealed systematic distortions, particularly at the high m/z values: regular sinusoidal noise affected most of the spectra (Fig. 1). This noise was periodic on the time scale, not on the m/z scale. We applied a Fourier transform to several affected spectra, restricting the transform to regions where larger peaks were absent. The period of the noise (roughly 1760 clock ticks) was found to be nearly constant across different fractions and samples, but the phase appeared to be random. We suspect that this phenomenon is linked to the frequency of the alternating current in the power source, but cannot confirm this suspicion without more information. We are certain that it is not due to biology. Sinusoids of the appropriate frequency were fit to the tails of each spectrum, extended to the full spectrum length, and subtracted out. This processing is illustrated in Fig. 2.

The clock is visible in the spectra. Summing the corrected spectra uncovered an unexpected periodic phenomenon – a recurrent dip in intensity every 4096 = 2^{12} clock ticks. Smaller, more complicated periodicities occurred at other powers of 2. These periodicities differed from the sinusoidal noise discussed earlier. The sinusoidal noise was random in phase, and so largely canceled between spectra. Here, we were able to detect the new dip because of reinforcement across spectra. Further, this dip was uniformly present in all 41 averaged spectra. Because this phenomenon occurred at powers of 2, we strongly suspect that it is an artifact related to a computer chip inside the instrument recording the data.
Estimates of the speed of light, with “confidence intervals”.

Tackling the widespread and critical impact of batch effects in high-throughput data

Jeffrey T. Leek, Robert B. Scharpf, Héctor Corrada Bravo, David Simcha, Benjamin Langmead, W. Evan Johnson, Donald Geman, Keith Baggerly and Rafael A. Irizarry

Abstract | High-throughput technologies are widely used, for example to assay genetic variants, gene and protein expression, and epigenetic modifications. One often overlooked complication with such studies is batch effects, which occur because measurements are affected by laboratory conditions, reagent lots and personnel differences. This becomes a major problem when batch effects are correlated with an outcome of interest and lead to incorrect conclusions. Using both published studies and our own analyses, we argue that batch effects (as well as other technical and biological artefacts) are widespread and critical to address.

We review experimental and computational approaches for doing so.

Many technologies used in biology—including high-throughput ones such as microarrays, bead chips, mass spectrometers and second-generation sequencing—depend on a complicated set of reagents and hardware, along with highly trained personnel, to produce accurate measurements. When these conditions vary during the course of an experiment, many of the quantities being measured will be simultaneously affected by both biological and non-biological factors. Here we focus on batch effects, a common and powerful source of variation in high-throughput experiments.

Batch effects are sub-groups of measurements that have qualitatively different behaviour across conditions and are unrelated to the biological or scientific variables in a study. For example, batch effects may occur if a subset of experiments was run on Monday and another set on Tuesday, if two technicians were responsible for different subsets of the experiments or if two different lots of reagents, chips or instruments were used. These effects are not exclusive to high-throughput biology and genomics research, and batch effects also affect low-dimensional molecular measurements, such as northern blots and quantitative PCR. Although batch effects are difficult or impossible to detect in low-dimensional assays, high-throughput technologies provide enough data to detect and even remove them. However, if not properly dealt with, these effects can have a particularly strong and pervasive impact.

Specific examples have been documented in published studies, in which the biological variables were extremely correlated with technical variables, which subsequently led to serious concerns about the validity of the biological conclusions.
December 28, 2007

THE DNA AGE

After DNA Diagnosis: ‘Hello, 16p11.2. Are You Just Like Me?’

By AMY HARMON

The girls had never met, but they looked like sisters.

There was no missing the similarities: the flat bridge of their noses, the thin lips, the fold near the corner of their eyes. And to the families of 14-year-old Samantha Napier and 4-year-old Taygen Lane there was something else, too. In the likeness was lurking an explanation for the learning difficulties, the digestion problems, the head-banging that had troubled each of them, for so long.

Several of the adults wiped tears from their eyes. “It’s like meeting family,” said Jessica Houk, Samantha’s older sister, who accompanied her and their mother to a Kentucky amusement park last July to greet Taygen.

But the two families are not related, and would never have met save for an unusual bond: a few months earlier, a newly available DNA test revealed that Samantha and Taygen share an identical nick in the short arm of their 16th chromosomes.

With technology that can now scan each of an individual’s 46 chromosomes for minute aberrations, doctors are providing thousands of children lumped together as “autistic” or “developmentally delayed” with distinct genetic diagnoses. The symptoms, they are finding, can be traced to one of dozens of deletions or duplications of DNA that were previously hard or impossible to detect.

Some mutations are so rare that they are known only by their chromosomal address: Samantha and Taygen are two of only six children with the diagnosis “16p11.2.”
Rare, high penetrance mutations – use linkage

Common, low penetrance variants – use association

Frequency in population

From Goncalo Abecasis