Integrative Analysis of Genomic Properties

Pratyaksha “Asa” Wirapati

Swiss Institute of Bioinformatics
Lausanne, Switzerland

Outline

- The concept of “genomic properties”
- Analysis of genomic properties
- Examples
- Discussions

Background

A flood of disparate genomic data in recent years

Two “axes of integration”:

- “Vertical”

  Various assays (expression, CGH, genetics, clinical, etc.)
  from the same samples (patients, tissues, etc.)

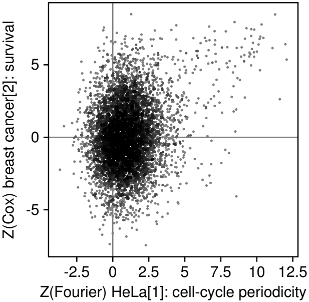
- “Lateral”

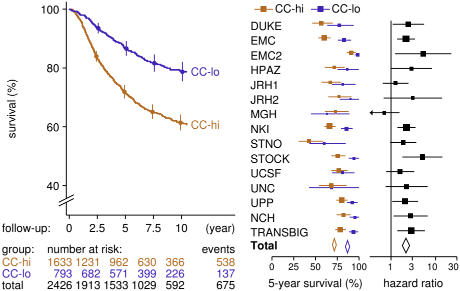
  Relating results of disparate studies (different sample, assays, and even completely different research questions)

  ⇒ connected by “genes”

  Concerted behavior of a group of genes in different contexts may signal a common underlying process

Example 1





group:	number at risk:					events
CC-hi	1633	1231	962	630	366	538
CC-lo	793	682	571	399	226	137
total	2426	1913	1533	1029	592	675

(The association between expression and cell-cycle phase in HeLa cells)
is "associated" with
(The association between expression and survival in breast cancer patients)

[1] Whitfield *et al.* (2002) *Mol Biol Cell* **13**:1977
[2] Wirapati *et al.* (20??) *Submitted, resubmitted, resubmitted, . . .*

The "definition" of genomic properties

Predicates or statements that can be made about each gene in the genome.

Operationally, anything that can be represented as a vector $(T_1, \ldots, T_i, \ldots, T_G)$, where $i = \{1, \ldots, G\}$ are genes in the genome, can be considered a "genomic property vector".

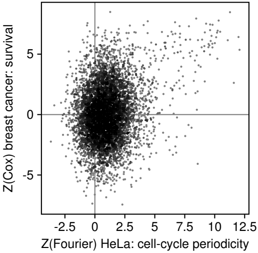
- The notion of "genes" is loosely defined, e.g. gene products, promotor binding sites, intergenic SNPs, etc. can be considered proxies of genes if there is a reasonable mapping scheme
- Context of the properties
  - Broad, e.g. gene ontology annotation
  - Specific population or experimental conditions
  - Individuals (we are not interested in this)

Examples of Genomic Properties

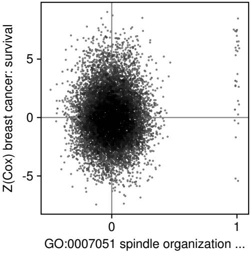
- "Trivial" properties: chromosomal location, etc.
- Gene-by-gene summary results (effect size or test of significance statistics) of genome-wide studies:
  - Expression studies
  - Genetic linkage, e.g. SNP chips
  - ChIP-CHIP binding assays
  - Evolutionary divergence between human and chimp
- Decisions based on the above
  - Prognostic signatures
- Results of annotations or reviews by "experts"
  - Gene Ontology, KEGG, MSigDB, etc.

Example 2

Continuous-Continuous



Continuous-Discrete



Discrete-Discrete

975	14
16192	17

⇒ "Gene sets" are vectors of binary summary statistics

Statistical issue: can the genes be considered "subjects" in sampling experiment?

Dependencies ⇒ p-value is off, but (ab)using the tests of (linear) independence (as ad hoc similarity measures) is found to be useful

Operations on genomic properties

- Construction of genomic property matrices
- Comparison of property vectors (pairwise)
- Aggregation of similar properties
- Visualization of similarity/dependency structure

Construction from primary data

Depends on the nature of each study

For most expression array studies, use gene-wise (generalized) linear models.

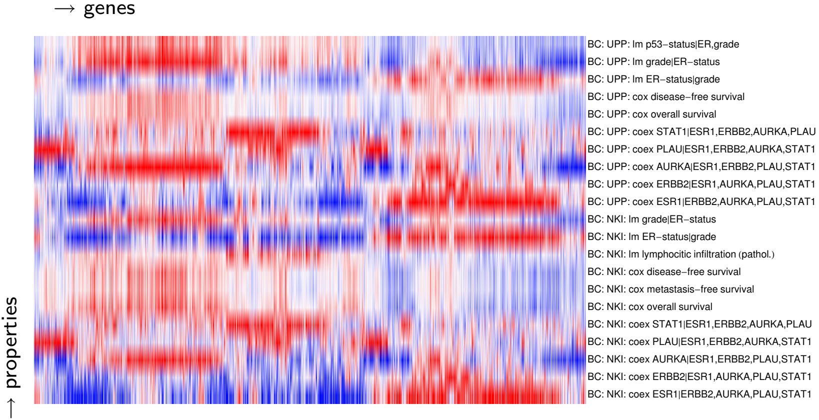
Use Z-scores ($\hat{\beta}/\widehat{\mathsf{SE}}(\beta)$ or $\mathsf{sgn}(\hat{\beta})\sqrt{\mathsf{deviance}}$) of partial tests of coefficients as the "common currency of integration"

It's a function of p-value and still keep the sign of the effect

Under the null, $Z \sim N(0, 1)$.

$Z \approx 4.6 \Leftrightarrow p = 0.05/20000$ (Bonferroni correction)

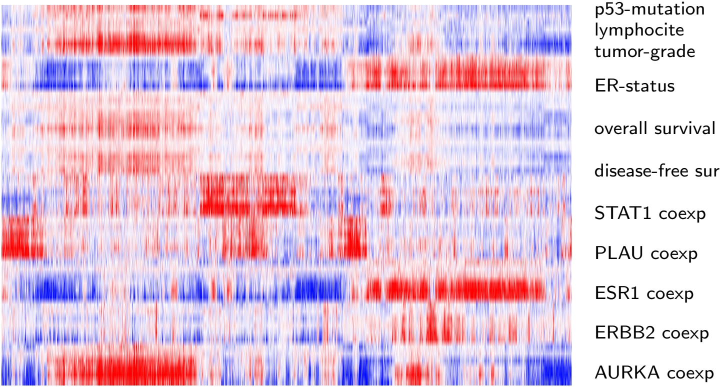
Matrix of Z-scores



Summary profiles of differential expression are identified by the contexts (disease-type, cohort) and regression equations

Multiple questions can be asked on the same dataset

Consistent answers in different cohorts/platforms



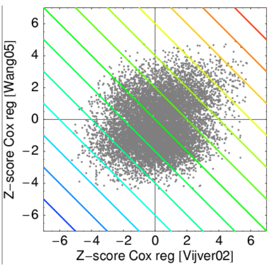
Datasets: NKI (custom Agilent), UPP (Affy U133A,B), STOCK (Affy U133A,B), UNC (Agilent HuA1), NCH (Agilent HuA1), DUKE (95Av2)

Aggregating replicate properties

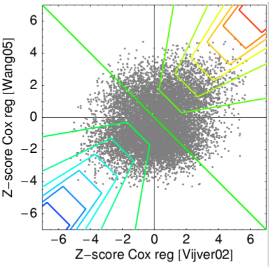
Summary results (of the same question) from multiple cohorts can be combined $\Rightarrow$ stronger significance and economy of thought in understanding many properties

Spectrum of choices for combining:

- "Normalize" and pool (then treat as single cohort)
- Covariate adjust, random effect models
- Combine meta-analytically (i.e. post-hoc)
  - β (only when meaningful)
  - scale-free effect sizes (Pearson's corr., Cohen's d, $Z/\sqrt{n}$)
  - (signed) significance (Z, $-2\log p$)
- Combine decisions (Venn diagram)



fixed-effect
meta analyis



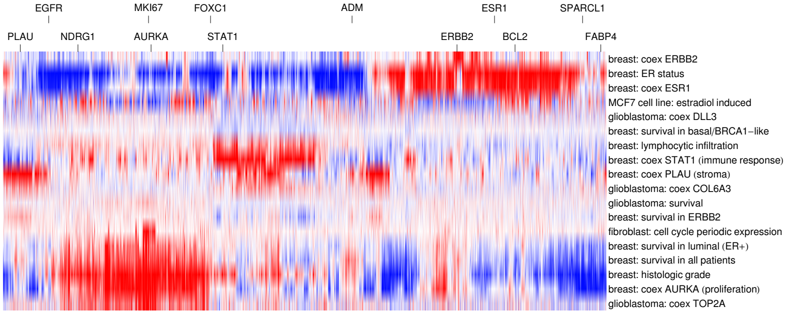
DerSimonian-Laird
random-effect
meta analysis

Appropriate ways to combine summary profiles depend on the data and questions.

For exploration, we just use the inverse-normal method $Z_j = \sum_i Z_{ij}/\sqrt{K_j}$

Broader Scope

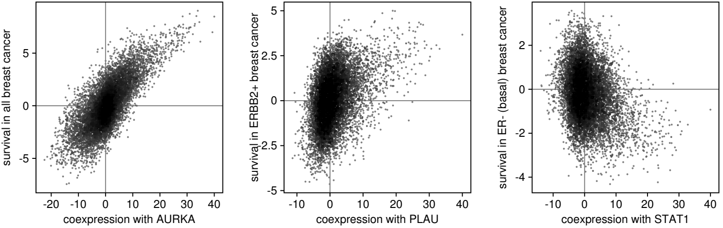
Add more datasets (glioblastoma, MCF7 estradiol-challenge and HeLa cell cycle), and more questions (survival in subtypes)



Similar answers to the similar questions in different diseases

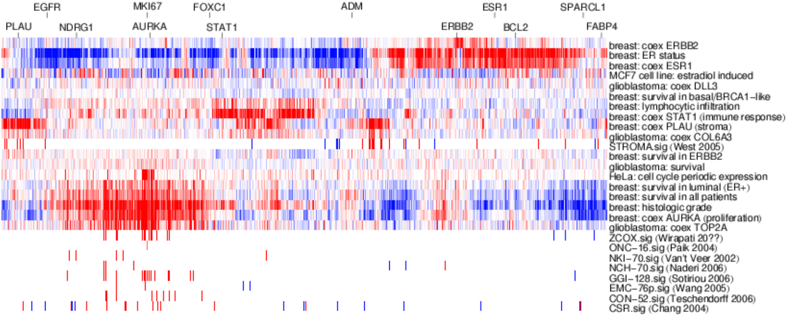
Relationship between tumor-based studies and experimental models

Coexpression modules and survival in breast cancer subtypes



- AURKA (proliferation) module in ER+ ("luminal")
  Sotiriou 2006 *J Natl Cancer Inst* **98***:262*
- PLAU (stroma/invasion) module in ERBB2+ tumors
  Urban 2006 *J Clin Oncol* **24**:4245 (RT-PCR on large independent cohort)
- STAT1 (immune response) might be *protective* in ER- subtype ("basal" or BRCA1-like)
  Ongoing investigation

Reviewing Proposed Prognostic Signatures



- Most breast cancer prognostic signatures are genes “sampled” from the proliferation module ⇒ potentially astronomical number of equivalent signatures

Coanalysis with GO terms and MSigDB

Treat them as binary-value matrices

Huge matrices (thousands of rows, tens of thousands of columns)

However, they are extremely sparse (less that 0.5% of the cells are non-zero) $\Rightarrow$ sparse representations and algorithms

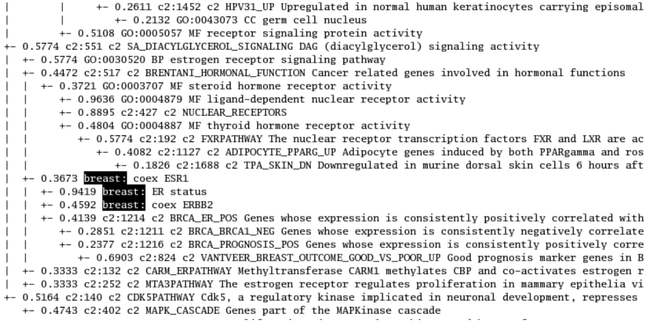
Statistical issues: how to compare?

$\Rightarrow$ Similarity measures for continuous-continous, continuous-discrete, and discrete-discrete should be comparable.

Let's see what happens if we (ab)use linear models (i.e. use correlation).

Organize the properties by finding their minimum (maximum correlation) spanning tree.

“High-tech” graph visualization program



```
|        |     +- 0.2611 c2:1452 c2 HPV31_UP Upregulated in normal human keratinocytes carrying episomal
|        |        +- 0.2132 GO:0043073 CC germ cell nucleus
|        +- 0.5108 GO:0005057 MF receptor signaling protein activity
+- 0.5774 c2:551 c2 SA_DIACYLGLYCEROL_SIGNALING DAG (diacylglycerol) signaling activity
|  +- 0.5774 GO:0030520 BP estrogen receptor signaling pathway
|  +- 0.4472 c2:517 c2 BRENTANI_HORMONAL_FUNCTION Cancer related genes involved in hormonal functions
|  |  +- 0.3721 GO:0003707 MF steroid hormone receptor activity
|  |     +- 0.9636 GO:0004879 MF ligand-dependent nuclear receptor activity
|  |     +- 0.8895 c2:427 c2 NUCLEAR_RECEPTORS
|  |     +- 0.4804 GO:0004887 MF thyroid hormone receptor activity
|  |        +- 0.5774 c2:192 c2 FXRPATHWAY The nuclear receptor transcription factors FXR and LXR are ac
|  |           +- 0.4082 c2:1127 c2 ADIPOCYTE_PPARG_UP Adipocyte genes induced by both PPARgamma and ros
|  |              +- 0.1826 c2:1688 c2 TPA_SKIN_DN Downregulated in murine dorsal skin cells 6 hours aft
|  +- 0.3673 breast: coex ESR1
|  |  +- 0.9419 breast: ER status
|  |  +- 0.4592 breast: coex ERBB2
|  |  +- 0.4139 c2:1214 c2 BRCA_ER_POS Genes whose expression is consistently positively correlated with
|  |     +- 0.2851 c2:1211 c2 BRCA_BRCA1_NEG Genes whose expression is consistently negatively correlate
|  |     +- 0.2377 c2:1216 c2 BRCA_PROGNOSIS_POS Genes whose expression is consistently positively corre
|  |        +- 0.6903 c2:824 c2 VANTVEER_BREAST_OUTCOME_GOOD_VS_POOR_UP Good prognosis marker genes in B
|  +- 0.3333 c2:132 c2 CARM_ERPATHWAY Methyltransferase CARM1 methylates CBP and co-activates estrogen r
|  +- 0.3333 c2:252 c2 MTA3PATHWAY The estrogen receptor regulates proliferation in mammary epithelia vi
+- 0.5164 c2:140 c2 CDK5PATHWAY Cdk5, a regulatory kinase implicated in neuronal development, represses
   +- 0.4743 c2:402 c2 MAPK_CASCADE Genes part of the MAPKinase cascade
```

Discussions

Biology "in-the-large": arrays of genomic studies

“Google Genomics"?

Analysis of many properties (both from experimental results at hand and from annotation databases) should be done simultaneously.

Open statistical issues:

- Similarity measures
- $p \approx n$, but extreme imbalance of signal and noise features
- Graphical models with conditional Gaussian model (mixed discrete and continous variables)?

Acknowledgements

BCF, SIB Lausanne
Mauro Delorenzi
Susanne Kunkel
Eugenia Migliavacca
Thierry Sengstag
Darlene Goldstein

DNA Array Facility
Sylvain Pradervand

Institute Jules Bordet, Brussels
Christos Sotiriou
Benjamin Haibe-Kains
Martine Piccart