Prognostic gene signatures: What are they good for?

Rory Stark

17 July 2019

UNIVERSITY OF CAMBRIDGE

CANCER RESEARCH UK | CAMBRIDGE INSTITUTE

Issues:

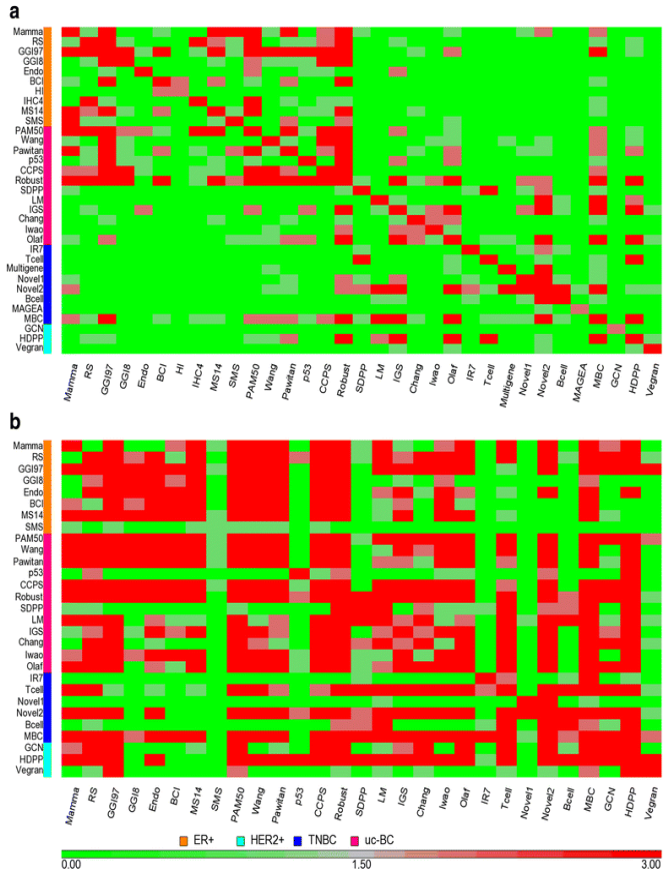
- **What do prognostic signature tell us?**
  - Cancer "landslide"
- **What don't they tell us?**
  - Little overlap between genes in signatures
  - Many possible signatures with similar predictive power
  - Signatures for unrelated phenotypes have similar predictive power

Overlap between signatures

Huang, Murphy, Xu 2018

From 33 breast cancer signatures:

- 2,239 genes present in at least one signature
- 238 overlap in at least two signatures (10.6%)
- 62 overlap at least three signatures
- 1 gene overlaps eight signatures (MKI67/proliferation)
- GO terms a bit better:
  - 988 unique function term significant in at least one signature
  - 195 overlap 2 (20%); 29 overlap 7



CANCER RESEARCH UK | CAMBRIDGE INSTITUTE

Many possible signatures

BIOINFORMATICS ORIGINAL PAPER Vol. 21 no. 2 2005, pages 171–178
doi:10.1093/bioinformatics/bth469

Outcome signature genes in breast cancer: is there a unique set?

Liat Ein-Dor[1,†], Itai Kela[1,3,†], Gad Getz[1,†], David Givol[2] and Eytan Domany[1,]*

van't Veer 70 gene signature derived from 5,852 expressed genes

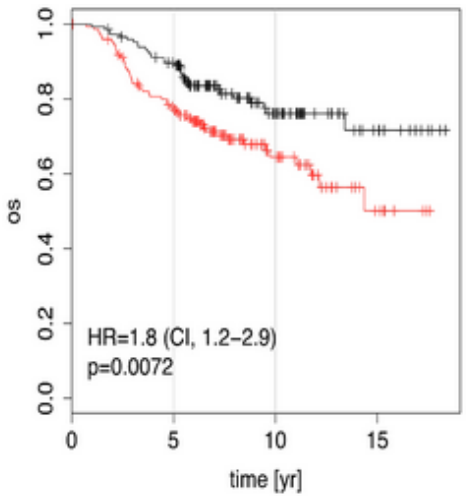
- Published signature:70 genes most correlated with survival
- Randomly sampled 70-gene signatures: of 10,000 random 70-gene signatures, 2,905 (29%) perform better than published signature
- Even signatures composed mostly of lowly correlated genes predict survival (Ein-Dor *et al* 2004)
- Why?
  - Many genes correlated with survival
  - Differences in correlations are small
  - Correlation fluctuate depending which exact samples are used

Signature for unrelated phenotypes are predictive

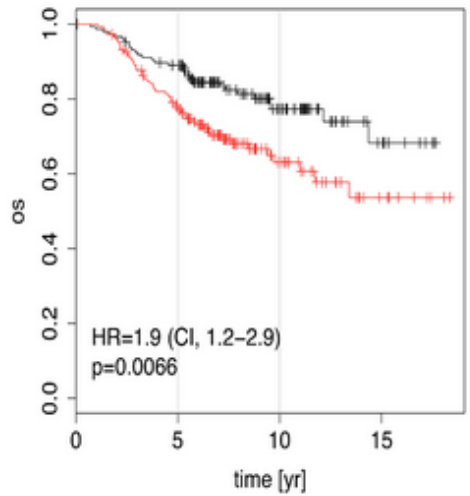
OPEN ACCESS Freely available online PLoS COMPUTATIONAL BIOLOGY

Most Random Gene Expression Signatures Are Significantly Associated with Breast Cancer Outcome

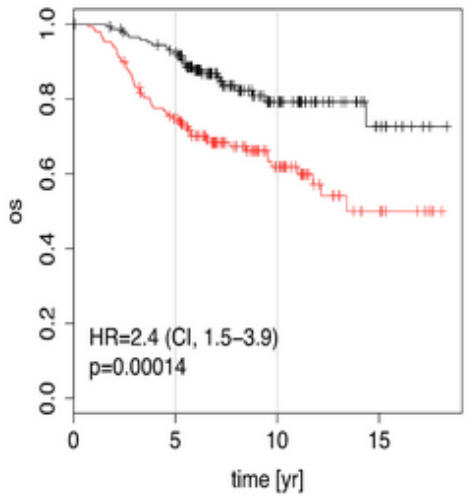
David Venet[1], Jacques E. Dumont[2], Vincent Detours[2,3]*



Post-prandial laughter



Localization of skin fibroblasts



Social defeat in mice

CANCER RESEARCH UK | CAMBRIDGE INSTITUTE

Not secure | software.broadinstitute.org/gsea/msigdb/index.jsp

GSEA
Gene Set Enrichment Analysis

GSEA Home | Downloads | Molecular Signatures Database | Documentation | Contact

- MSigDB Home
- About Collections
- Browse Gene Sets
- Search Gene Sets
- Investigate Gene Sets
- View Gene Families
- Help

MSigDB
Molecular Signatures Database

Molecular Signatures Database v6.2

Overview

The Molecular Signatures Database (MSigDB) is a collection of annotated gene sets for use with GSEA software. From this web site, you can

- **Search** for gene sets by keyword.
- **Browse** gene sets by name or collection.
- **Examine** a gene set and its annotations. See, for example, the GO_NOTCH_SIGNALING_PATHWAY gene set page.
- **Download** gene sets.
- **Investigate** gene sets:
  - **Compute overlaps** between your gene set and gene sets in MSigDB.
  - **Categorize** members of a gene set by gene families.
  - **View the expression profile** of a gene set in a provided public expression compendia.

License Terms

GSEA and MSigDB are available for use under these license terms.

Please register to download the GSEA software, access our web tools, and view the MSigDB gene sets. After registering, you can log in at any time using your email address. Registration is free. Its only purpose is to help us track usage for reports to our funding agencies.

Current Version

MSigDB database v6.2 updated July 2018. Release notes.
GSEA/MSigDB web site v6.3 released January 2018

Collections

The MSigDB gene sets are divided into 8 major collections:

H **hallmark gene sets** are coherently expressed signatures derived by aggregating many MSigDB gene sets to represent well-defined biological states or processes.

C1 **positional gene sets** for each human chromosome and cytogenetic band.

C2 **curated gene sets** from online pathway databases, publications in PubMed, and knowledge of domain experts.

C3 **motif gene sets** based on conserved cis-regulatory motifs from a comparative analysis of the human, mouse, rat, and dog genomes.

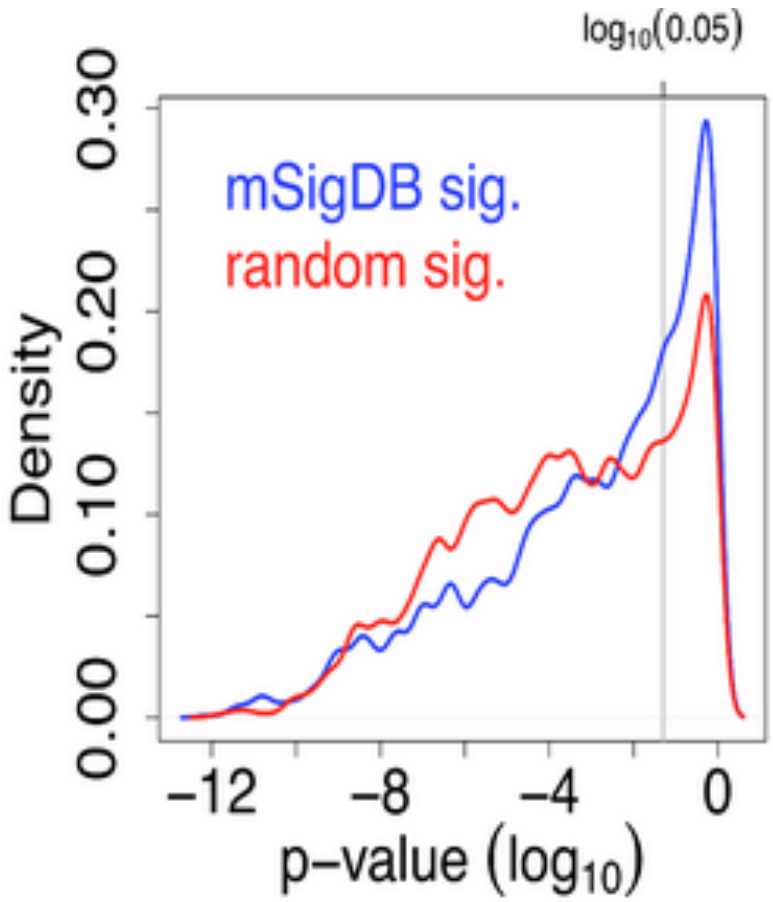
C4 **computational gene sets** defined by mining large collections of cancer-oriented microarray data.

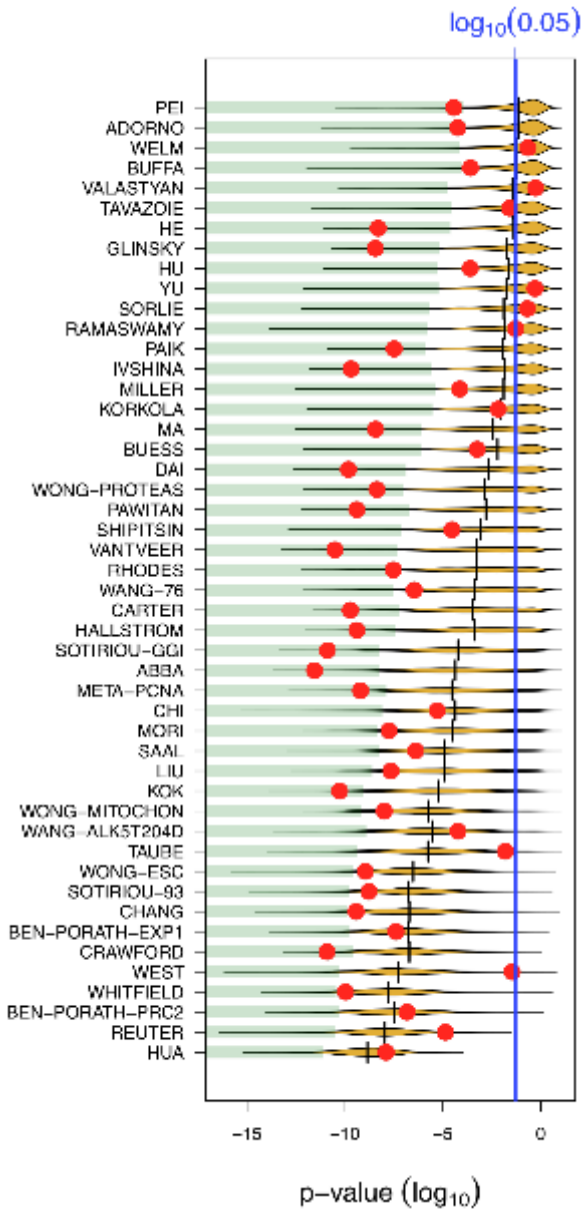
C5 **GO gene sets** consist of genes annotated by the same GO terms.

C6 **oncogenic gene sets** defined directly from microarray gene expression data from cancer gene perturbations.

C7 **immunologic gene sets** defined directly from microarray gene expression data from immunologic studies.

Most random signatures predict survival





CANCER RESEARCH UK | CAMBRIDGE INSTITUTE

SigCheck Bioconductor package

Bioconductor
OPEN SOURCE SOFTWARE FOR BIOINFORMATICS

Home Install Help

Home » Bioconductor 3.9 » Software Packages » SigCheck

SigCheck

platforms all | rank 1243 / 1743 | posts 1 / 1 / 0 / 0 | in Bioc 4.5 years
build ok | updated before release | dependencies 115

DOI: 10.18129/B9.bioc.SigCheck

Check a gene signature's prognostic performance against random signatures, known signatures, and permuted data/metadata

Bioconductor version: Release (3.9)

While gene signatures are frequently used to predict phenotypes (e.g. predict prognosis of cancer patients), it it not always clear how optimal or meaningful they are (cf David Venet, Jacques E. Dumont, and Vincent Detours' paper "Most Random Gene Expression Signatures Are Significantly Associated with Breast Cancer Outcome"). Based on suggestions in that paper, SigCheck accepts a data set (as an ExpressionSet) and a gene signature, and compares its performance on survival and/or classification tasks against a) random gene signatures of the same length; b) known, related and unrelated gene signatures; and c) permuted data and/or metadata.

Author: Rory Stark <rory.stark at cruk.cam.ac.uk> and Justin Norden

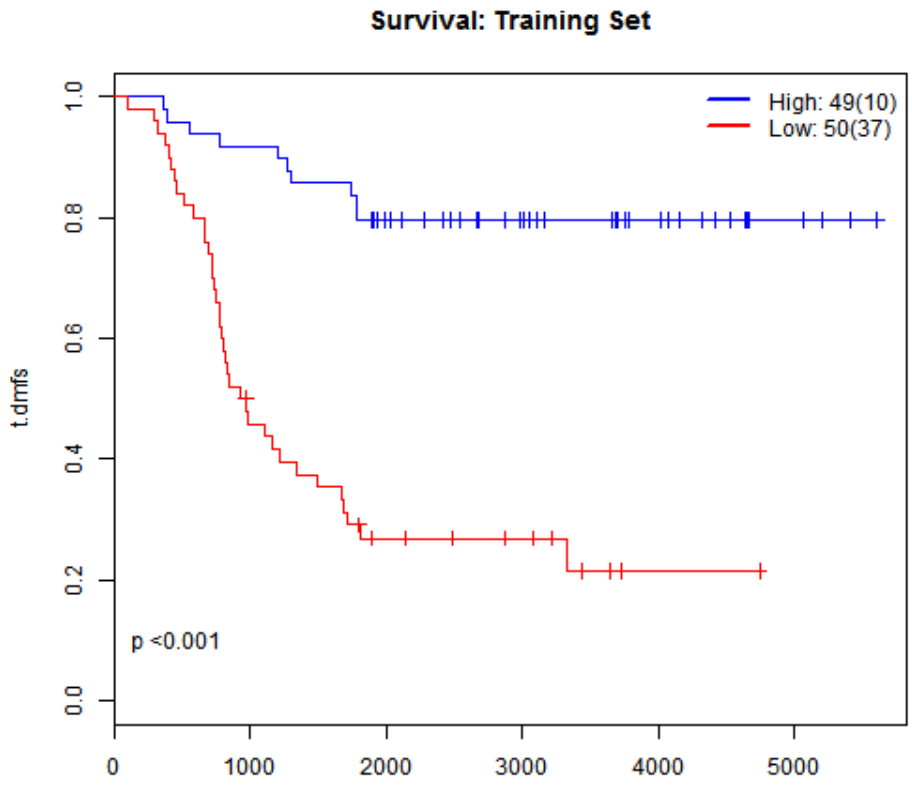
Maintainer: Rory Stark <rory.stark at cruk.cam.ac.uk>

CANCER RESEARCH UK | CAMBRIDGE INSTITUTE

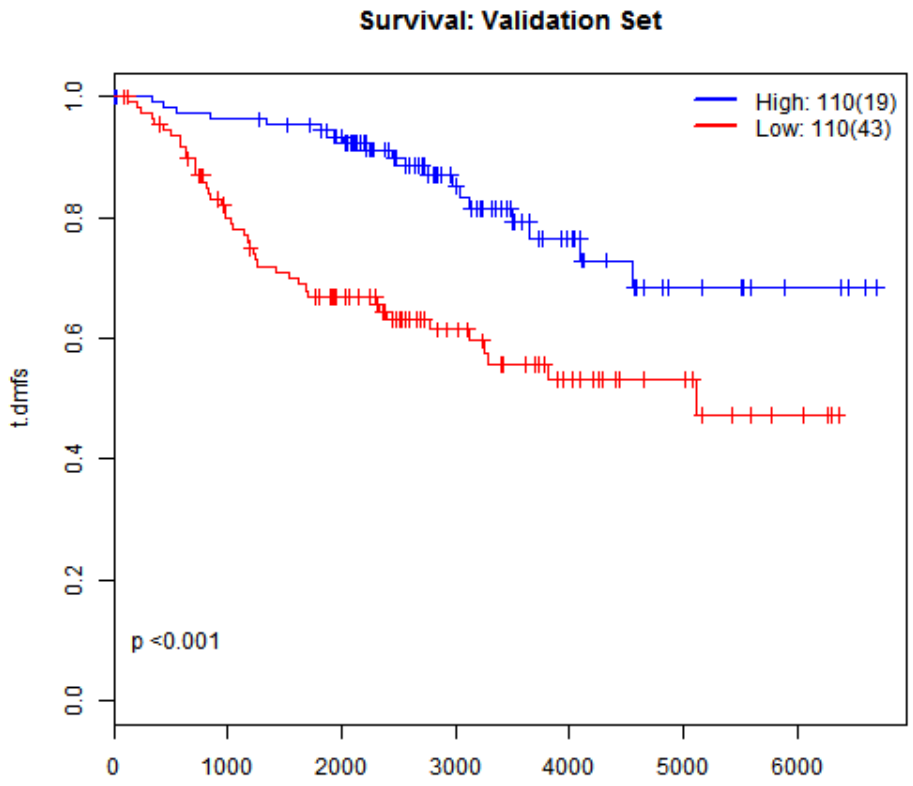
Using SigCheck

- Provide:
  - ExpressionSet (or SummarizedExperiment) object
    - Feature annotation
    - Survival metadata
  - Gene signature
- Analyses
  - Compare to random signature of the same length
    - Useful for exploring inherent prediction power of a dataset
  - Compare to database of existing signatures
  - Check signature against permuted data
- Also incorporates **`MLInterfaces`** package for classification

Example: NKI (van't Veer) Breast Cancer Signature (70 genes)



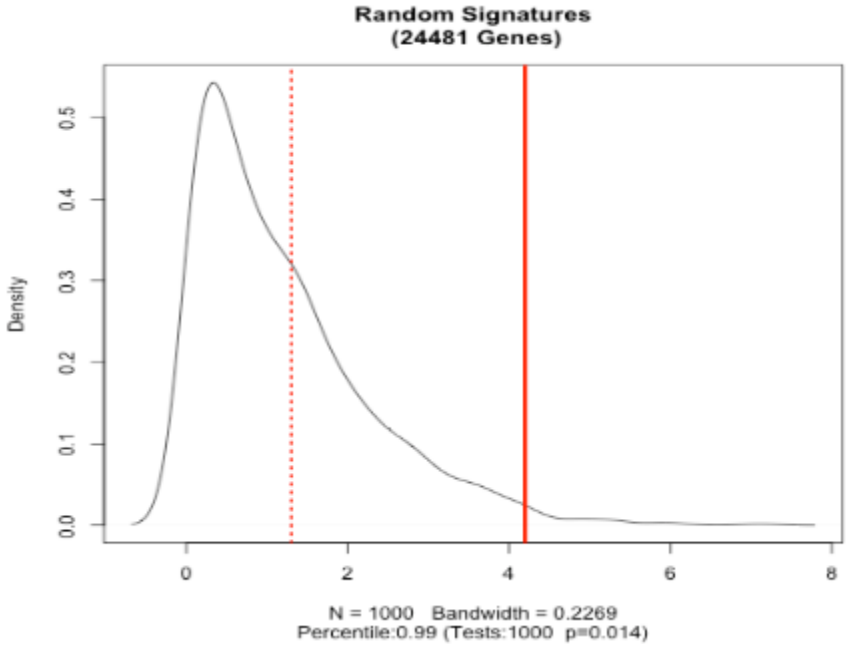
p = 1.09e-08

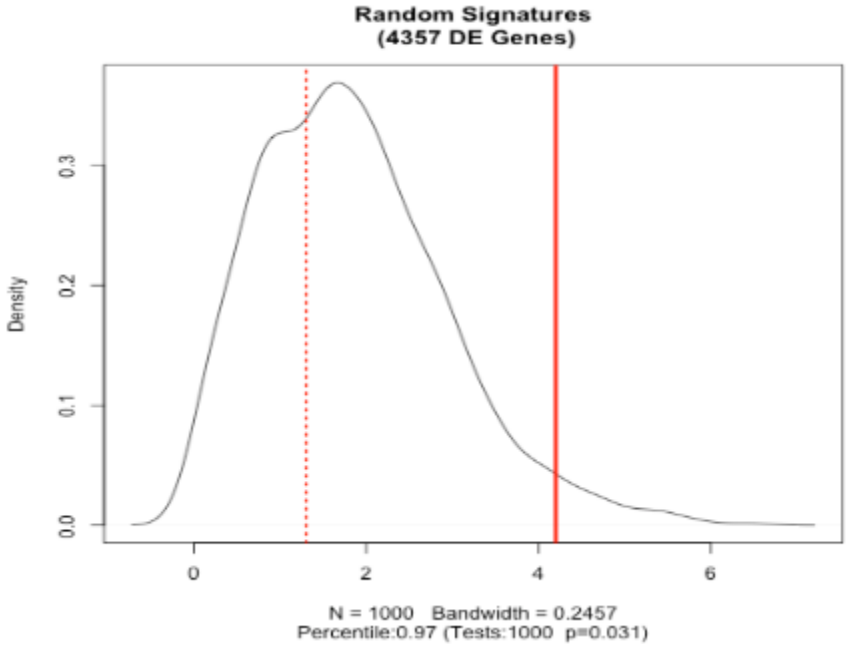


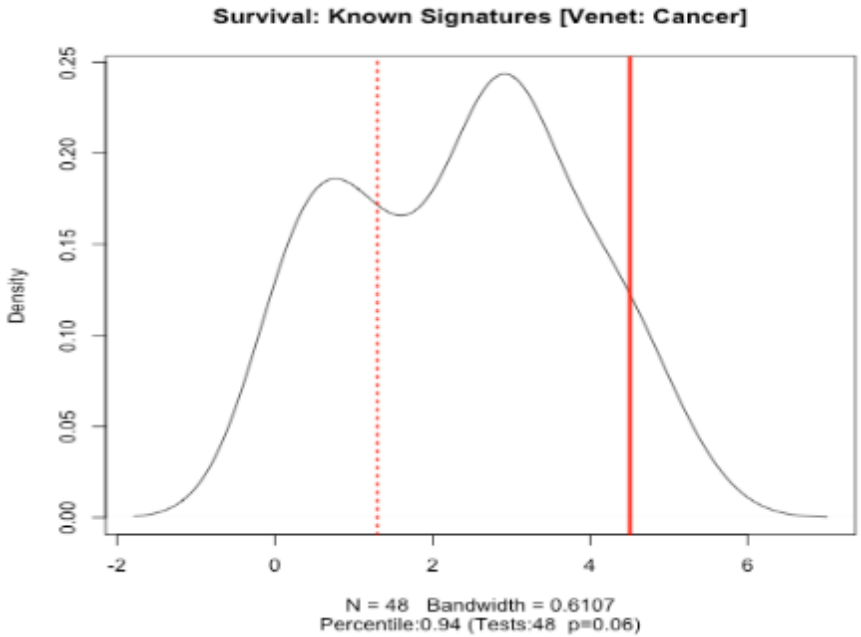
p = 6.0e-05

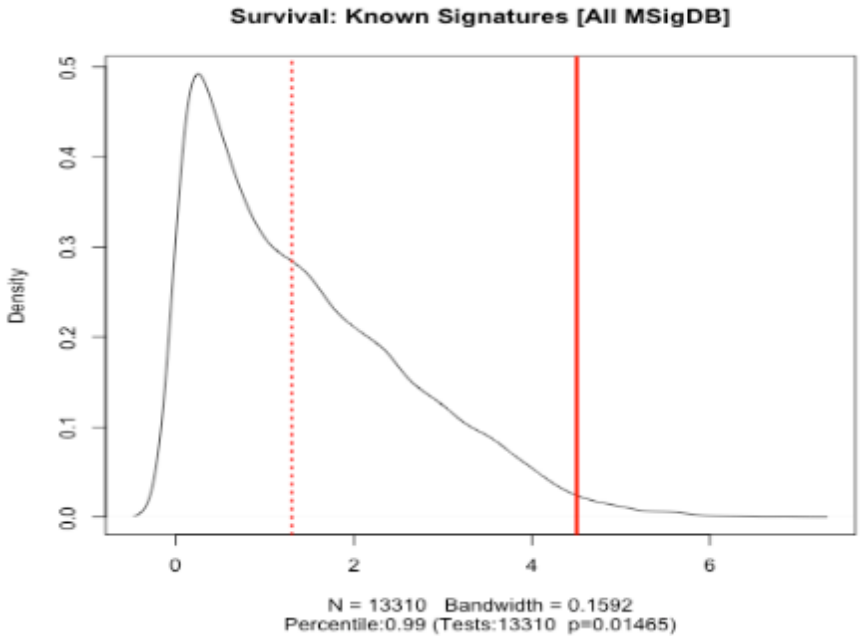
CANCER RESEARCH UK | CAMBRIDGE INSTITUTE

NKI Signature Performance





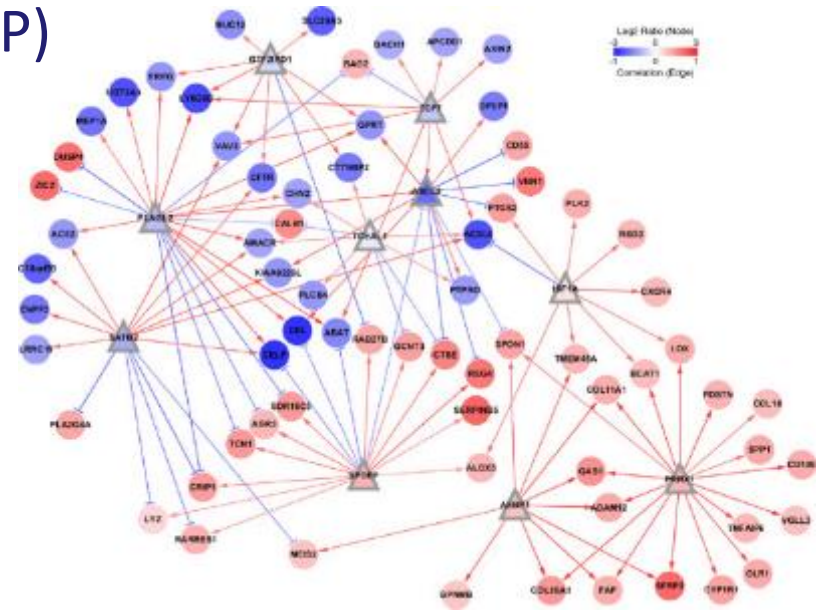




How can signature be enriched for biological meaningful (driver) genes?

Use biological criteria, not just statistical analysis, in selecting genes

- Transcription Factor binding (ChIP)
- Epigenetic marks
- Deeper integration of functional genomics data



Beyond genes

- Pathways/functions
- ARACNe: network analysis of regulons

CANCER RESEARCH UK | CAMBRIDGE INSTITUTE

nature International weekly journal of science

NATURE | LETTER

previous article next article

Differential oestrogen receptor binding is associated with clinical outcome in breast cancer

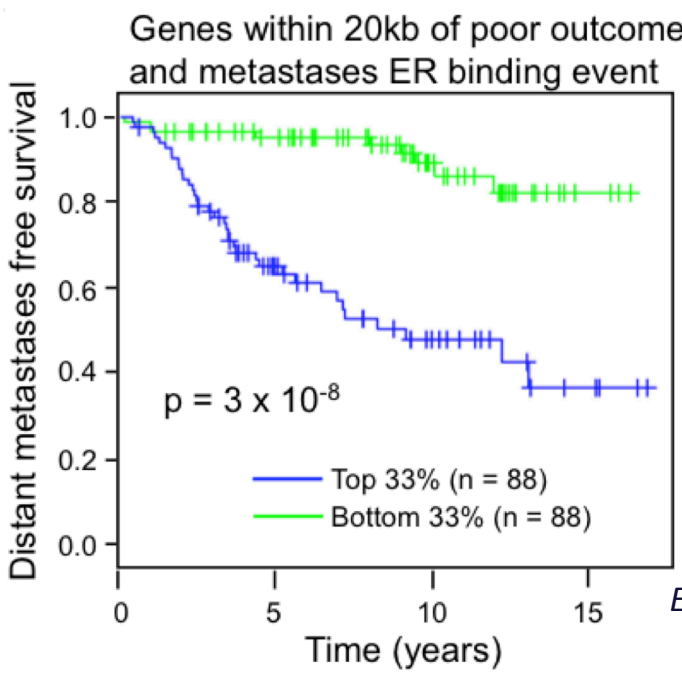
Caryn S. Ross-Innes, Rory Stark, Andrew E. Teschendorff, Kelly A. Holmes, H. Raza Ali, Mark J. Dunning, Gordon D. Brown, Ondrej Gojis, Ian O. Ellis, Andrew R. Green, Simak Ali, Suet-Feung Chin, Carlo Palmieri, Carlos Caldas & Jason S. Carroll

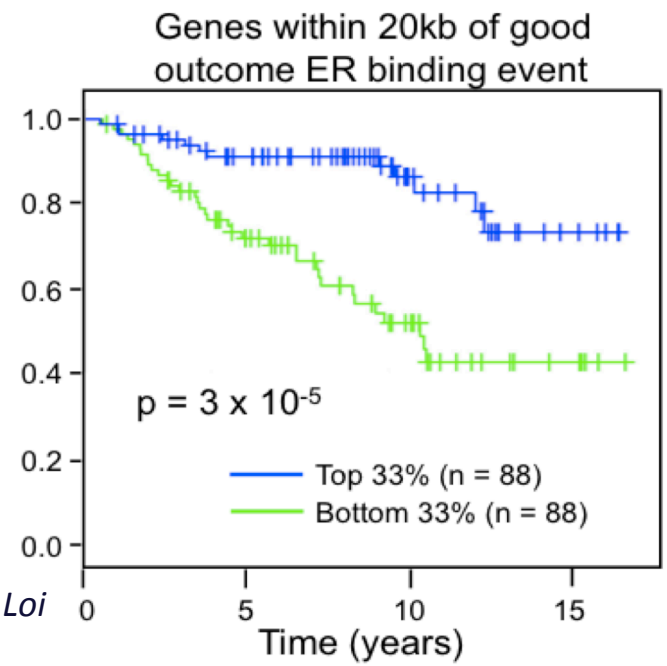
Affiliations | **Contributions** | **Corresponding authors**

Nature **481**, 389–393 (19 January 2012) | doi:10.1038/nature10730

Received 19 May 2011 | Accepted 23 November 2011 | Published online 04 January 2012

Enriching for driver genes





Expression data: Loi et al.

- Signature composed of genes within 20kb of DB sites
  - **265** genes in Poor outcome signature
  - **109** genes in Good outcome signature
- Classifier based on up/down regulation in mRNA expression sets
- Validated in 7 publicly available BC expression datasets

CANCER RESEARCH UK | CAMBRIDGE INSTITUTE

Acknowledgements

- Jason Carroll lab
  - Caryn Ross-Innes
  - Hisham Mohammad
- CRUK-CI Bioinformatics Core
  - Matthew Eldridge
- SigCheck Package Collaborator:
  - Justin Norden

CANCER RESEARCH UK | CAMBRIDGE INSTITUTE