The Abdus Salam International Centre for Theoretical Physics

2361-15

School on Large Scale Problems in Machine Learning and Workshop on Common Concepts in Machine Learning and Statistical Physics

20 - 31 August 2012

MACHINE LEARNING IN SYSTEMS BIOLOGY: Bioinformatics for Genomic Medicine

Ole WINTHER

Technical University of Denmark DTU and University of Copenhagen KU Denmark

Bioinformatics for genomic medicine

Ole Winther

Technical University of Denmark (DTU) & University of Copenhagen (KU)

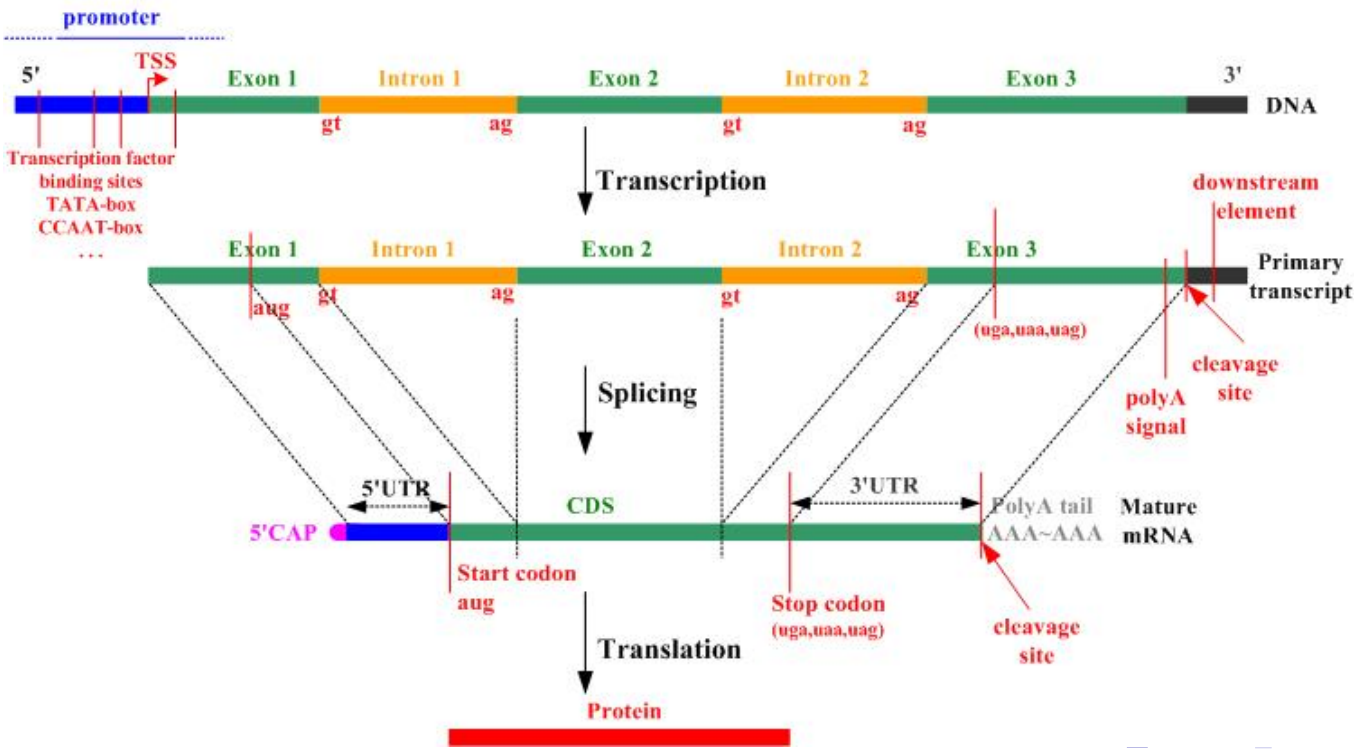
March 28, 2012

- Overall motivation – use genomic data to improve cancer diagnosis and treatment.
- Multidisciplinary collaboration between Bioinformatics Centre, KU and Genomic Medicine and Oncology, Copenhagen University Hospital (Riget).
- Classification and survival analysis for cancer from gene expression data
- Next steps:
  - Include data from more experimental platforms and
  - modeling more aspects of the diagnostic process - therapy selection (personalized medicine)

- Experimental platforms
- Case 1 - Cancer of unknown primary origin (CUP)
  - Classification of cancer and outlier detection
  - Large gene expression profiling dataset
- Case 2 - Prognosis in gastric cancer
  - Random survival forests
  - using gene sets scores as covariates
- Outlook

Experimental platform

- mRNA gene expression profiling
  - Microarray - 50k genes $660
- Exome sequencing - next generation sequencing
  - `23andme.com/exome` 80x coverage of 50M bases $999
- Additional platforms: Non-coding RNAs and proteomics



- Cancer are classified according to their origin.
- CUP - a metastasis is located, but not the primary tumor.
- $2-5\%$ of cancer patients get the CUP diagnosis ~ 20 annually at Riget.
- No primary tumor located in two-third of cases.
- Knowing the origin typically determines treatment.
- Cancer is a very heterogeneous disease - improved molecular characterization will lead to identification of more clinical fitting subtypes.

- Aim: Build classifier for major cancer types
- Phase one - data collection and normalization:
  - Careful curation of 2400+ expression profiles (samples) downloaded from Gene Expression Omnibus (GEO) `http://www.ncbi.nlm.nih.gov/geo/`
  - Training data - 1466 samples: 1299 primary tumors and 167 normal tissue (various organs)
  - Test set - 641 tumor samples: 391 primary tumors and 250 metastases.
  - 57 CUP samples of which 29 remain unknown after work-up.
- 15 cancer types: thyroid, lung, stomach, colon/rectum, pancreas, bile duct/gallbladder, liver, kidney, urinary tract, prostate, breast, ovary, endometrium, cervix uteri, testis cancer, a group of malignant melanomas
- Normalization with Robust multi-chip average (RMA) in R/Bioconductor package.

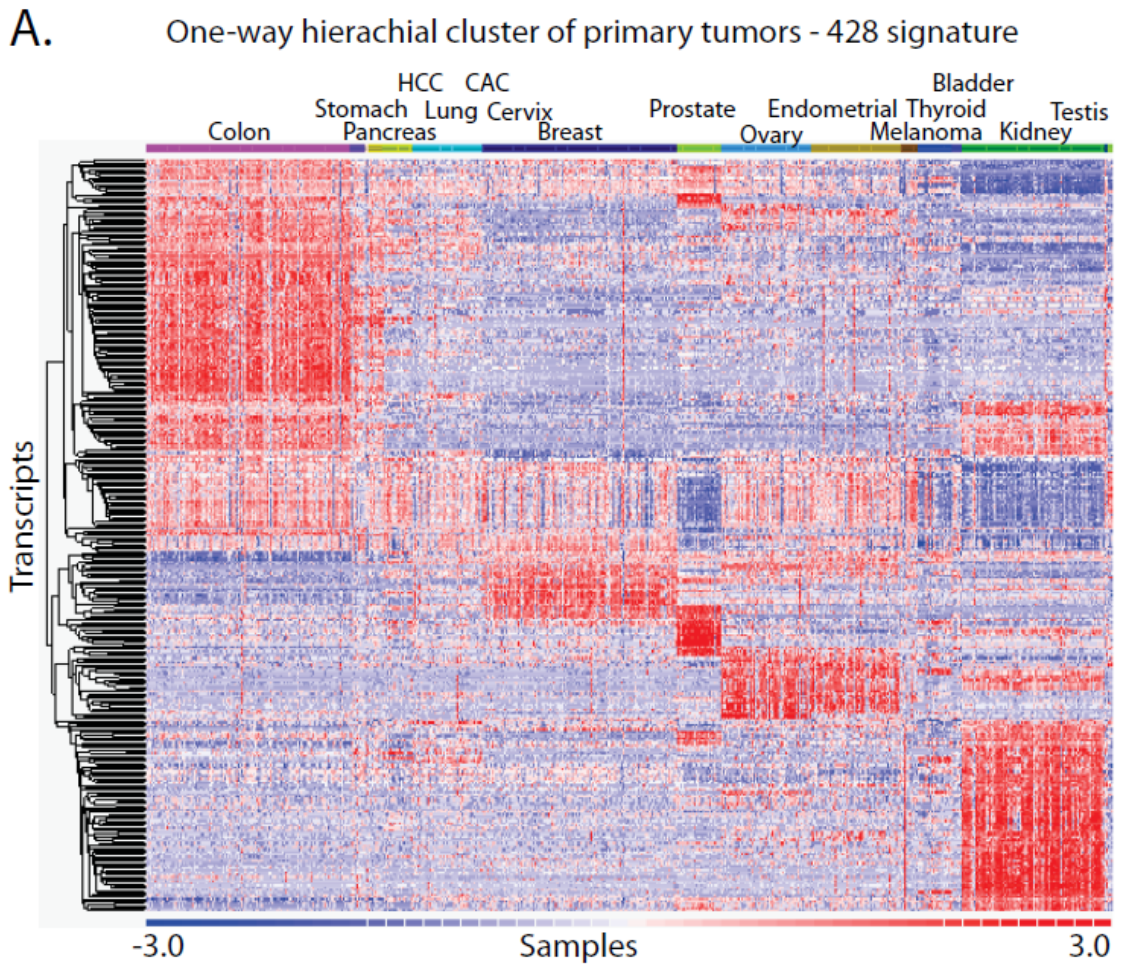
- Phase two - filtering and training of classifier:
  - 47k+ transcript expression values (Affymetrix U133 Plus 2.0)
  - 20k left after variance filtering
- Two-step training of classifier:
  1. Univariate test (F) identification of discriminative probes
  2. Train classifier on selected probes
- Optimal p-value cut-off for F-test ($= 10^{-180}$) found by nested cross-validation

- Initial trials identified linear discriminant analysis (LDA) as the method with lowest error rate
- c is the class label and x the covariate vector (log fold change in expression values).

$$
\begin{aligned}
p(c|x) &= \frac{p(x|c)p(c)}{p(x)} \\
p(x|c) &= \mathcal{N}(x; \mu_c, \Sigma) \\
p(x) &= \sum_c p(x|c)p(c)
\end{aligned}
$$

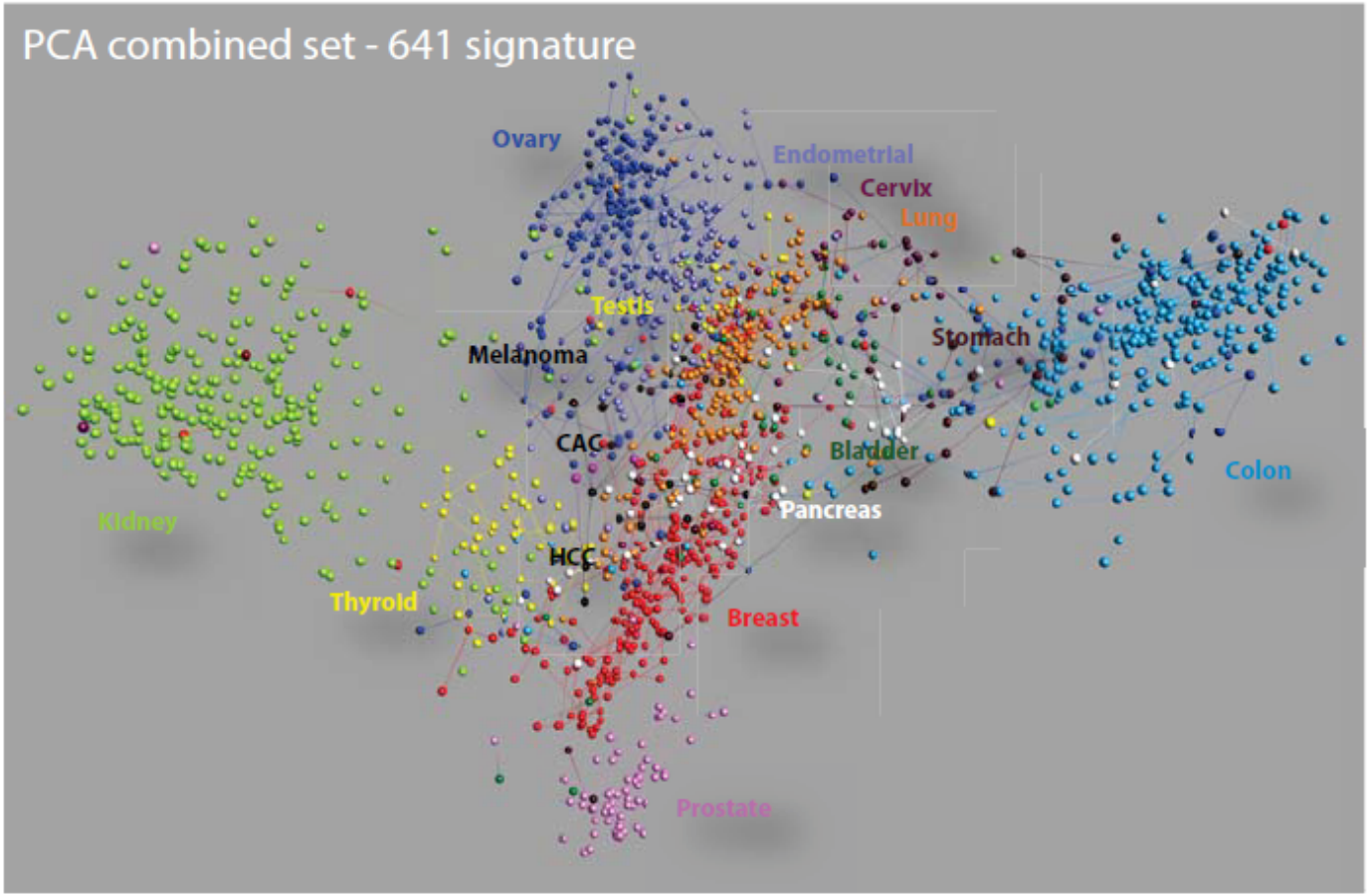
- Test performance 428 probe classifier: 90% and 83% for primary tumors and known metastases, respectively.
- CUP classifier (merge training and test set, 641 probes) had a LOOCV accuracy of 92% and 87%, respectively.

Cancer of unknown primary origin (CUP) classification



Cancer of unknown primary origin (CUP) classification

B.



Classification of CUP patients

ID	Sex/age	Biopsy site	Histology	Path Diag.	Stand of Ref	LDA Pred	Outlier score
14.	F/56	LN neck	PDC	Lung	Lung (CD)	Lung	975
17.	F/57	LN neck	Adenoc.	Lower GI	Colon (RD)	Colon	746
22.	M/55	LN neck	Adenoc.	CUP	Stomach (RD)	Normal	934
23.	M/39	LN retro	PDC	CUP	Kidney (RD)	Kidney	1085
28.	F/58	Peritoneum	PDA	Ovary	Ovary (RD)	Ovary	810
31.	F/40	LN neck	PDA	CUP	Lung (CD)	Stomach	985
34	M/74	Skin	PDA	Lung	Lung (RD)	Lung	898
39.	M/71	Liver	Adenoc.	CUP	Pancreas (CD)	Pancreas	1097
40.	F/44	Liver	Adenoc.	Colon	Colon (RD)	Colon	729
44.	F/43	Kidney	Carc.	CUP	Bladder (RD)	Bladder	1286
49.	M/60	LN neck	PDA	Kidney	Kidney (RD)	Kidney	1223
51.	F/42	LN pelvis	SCC	CUP-SCC	Cervical (RD)	Cervix	828
52.	M/53	Liver	PDA	CUP	CCC (RD)	CCC	923
53.	M/70	Liver	Adenoc.	Lung	Lung (RD)	Lung	1047
57.	M/67	Liver	Adenoc.	CCC	CCC (RD)	HCC	965
66.	F/68	Liver	PDA	CUP	CCC (RD)	Cervix	1100
70.	M/38	Peritoneum	Adenoc.	Stomach	Stomach (CD)	Colon	842
74.	M/62	Leg	Carc.	Adnex tumor	Adnex tumor (RD)	Normal	1010
76.	M/64	Liver	Adenoc.	Lower GI	Small intestine (RD)	Colon	912
77.	M/59	LN axilla	PDC	CUP	Lung (CD)	Breast	978
86.	F/61	LN axilla	Adenoc.	CUP	Lung (RD)	Stomach	1108
88.	F/36	Peritoneum	Adenoc.	Ovary	Ovary (RD)	Cervix	1033
89.	F/57	Liver	PDA	CCC	CCC (RD)	CCC	916
90.	F/71	Peritoneum	Adenoc	Ovary	Ovary (RD)	Ovary	781
92.	M/62	Liver	Malignant tumor	Angiosarcoma	Angiosarcoma (RD)	Normal	1097
95.	M/45	Peritoneum	PDC	DSRCT	DSRCT (RD)	Breast	1098
71+72	M/61	Bone + Kidney	PDC	Kidney	Kidney (RD)	Kidney	1096
							1277
75+87	F/43	Liver	PDA	CCC	CCC (RD)	CCC	925
							1030

Classification of CUP patients

ID	Sex/age	Biopsy site	Histology	Path Diag.	Stand of Ref	LDA Pred	Outlier score
11.	F/58	LN neck	PDA	CUP	CUP (SD)	Ovary	756
13.	F/72	Peritoneum	PDA	CUP	CUP (NSD)	Pancreas	1193
21.	M/63	LN neck	PDC	CUP	CUP (NSD)	Breast	1108
26.	F/67	Skin	PDA	CUP	CUP (NSD)	Breast	971
32.	M/53	LN neck	PDSCC	CUP-SCC	CUP (NSD)	Normal	926
33.	M/58	Skin	PDA	CUP	CUP (NSD)	Colon	1098
41.	M/74	Liver	PDA	Pancreas	CUP (NSD)	Stomach	1040
42.	M/56	Liver	Adenoc.	CUP	CUP (NSD)	Pancreas	994
43.	F/50	LN retro	PDA	CUP	CUP (NSD)	Stomach	797
45.	M/44	Liver	PDC	CUP	CUP (NSD)	Colon	1245
46.	F/76	Liver	Adenoc.	CUP	CUP (NSD)	Normal	1027
47.	F/59	Liver	Adenoc.	CUP	CUP (SD)	CCC	932
48.	F/59	LN neck	PDC	CUP	CUP (NSD)	Ovary	1032
54.	F/67	Liver	Adenoc.	CUP	CUP (NSD)	Normal	1068
55	F/55	Liver	Adenoc.	CUP	CUP (NSD)	Normal	962
58.	F/67	Liver	PDC	CUP	CUP (SD)	CCC	995
61.	M/72	Liver	Carc.	HCC	CUP (NSD)	CCC	1102
64.	F/65	LN inguien	PDA	CUP	CUP (SD)	Lung	1168
65.	M/62	LN neck	PDSCC	CUP-SCC	CUP (NSD)	Breast	929
73.	M/43	LN retro	PDC	CUP	CUP (NSD)	Normal	1020
78.	F/59	Lung	Adenoc.	Lower GI	CUP (NSD)	Lung	1062
80.	F/58	Liver	Adenoc.	CUP	CUP (SD)	CCC	1111
81.	F/71	Liver	PDA	CUP	CUP (NSD)	Breast	1212
82.	F/56	Bone	Adenoc.	CUP	CUP (NSD)	CCC	1209
83.	F/59	Liver	PDA	CUP	CUP (SD)	CCC	1061
91.	F/65	LN axilla	Adenoc.	CUP	CUP (SD)	Lung	939
93.	M/58	Bone	PDSCC	CUP-SCC	CUP (NSD)	Breast	940
94.	F/55	Liver	PDA	CUP	CUP (NSD)	Normal	984
50. + 68	M/41	Adr gl	PDC	CUP	CUP (NSD)	Stomach	978
						Pancreas	1079

- CUP classifier not sex-specific, but some cancers are: ovary, cervical and prostate.
- Unlikely events occur - like men classified as breast - may reflect real biology and limited validity of classification categories.
- Renormalizing class priors $p(c)$:

$$p(c|x) = \frac{p(x|c)p(c)}{p(x)}$$

$$p_{\mathrm{rm}}(c|x) = \frac{p(x|c)p_{\mathrm{rm}}(c)}{p_{\mathrm{rm}}(x)} = p(c|x)\frac{p(x)}{p_{\mathrm{rm}}(x)}\frac{p_{\mathrm{rm}}(c)}{p(c)}$$

- $p_{\mathrm{rm}}(c) = 0$ for cancers not appearing (like ovary in men) and renormalize others appropriately.

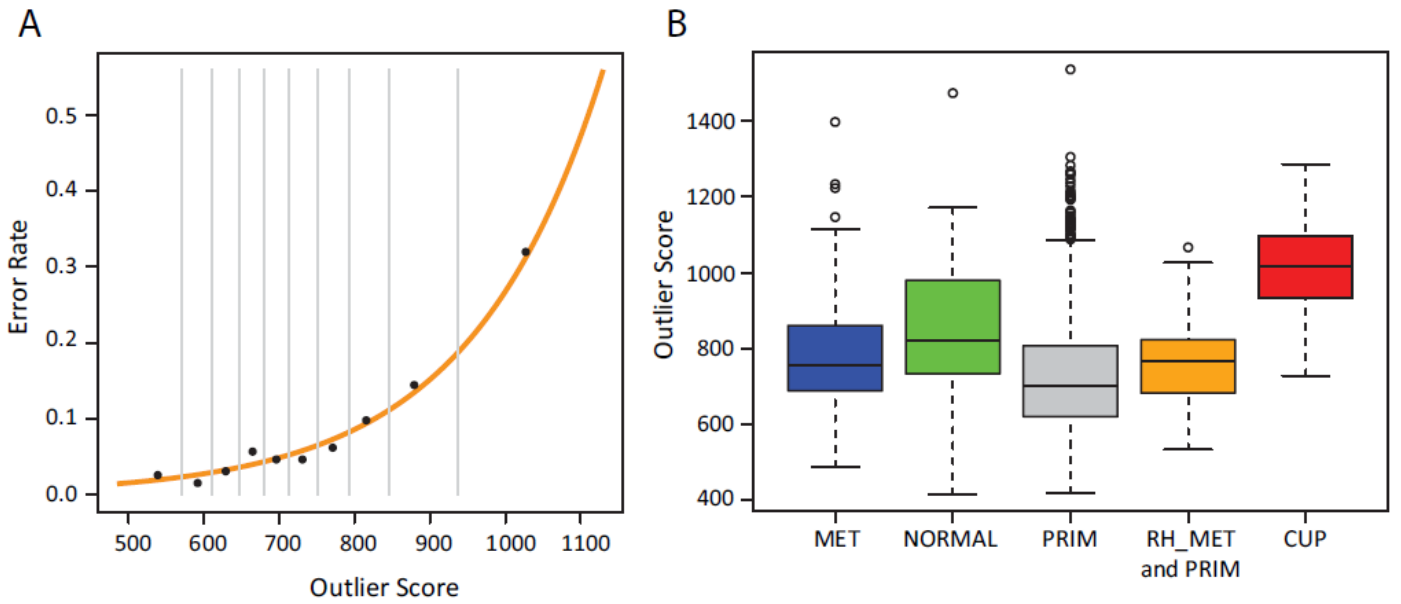
Outlier scores

- $p(x)$ measure how probable x is according to the model - outlier score

$$\mathrm{OS} = -\log p(x)$$

- Use quadratic discriminant analysis (QDA)

$$p(x) = \sum_c p(x|c)p(c) \qquad \text{with} \qquad p(x|c) = \mathrm{N}(x; \mu_c, \Sigma_c)$$



Outlier scores

- Clinical classifier
- Made with R/Sweave
- Input: expression profile
- Output: class probabilities and outlier score
- Notice huge differences on log scale
- Suggests that LDA is not optimal with respect to e.g. Brier score.

Created: Thu Nov 17 11:14:48 2011

R version: R version 2.13.0 (2011-04-13)

File name: `JVik-RH-UPT-104-HG-U133_Plus_2-a1.CEL`

Model file: `./upt_obj.RData`

Output data: `./JVik-RH-UPT-104-HG-U133_Plus_2-a1.CEL`

Sample gender: male

Outlier score: -1196.695

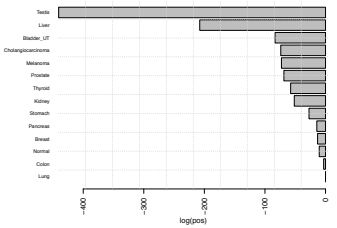


Figure 1: Predictions with scaled posterior in log-scale.

	log(pos)
Lung	-2031.39
Colon	-2034.52
Normal	-2041.78
Breast	-2044.51
Pancreas	-2045.78
Stomach	-2058.68
Kidney	-2082.98
Thyroid	-2088.99
Prostate	-2100.13
Melanoma	-2104.17
Cholangiocarcinoma	-2105.28
Bladder_UT	-2114.75
Liver	-2239.06
Testis	-2471.94

Table 1: Predictions with posterior in log-scale.

Probabilistic PCA

- Probabilistic PCA (Tipping and Bishop, 1999):

$$\begin{aligned} x &= Wz + \epsilon \\ z &\sim \mathcal{N}(z; 0, I) \\ \epsilon &\sim \mathcal{N}(\epsilon; 0, \sigma^2 I) \end{aligned}$$

- Marginalizing z and ϵ

$$p(x|W, \sigma^2) = \mathcal{N}(x; 0, WW^T + \sigma^2 I)$$

- Structured covariance model.

- Log likelihood for **W** and σ^2:

$$\begin{aligned}\log L(\theta; X) &= \sum_n \log p(x_n|W, \sigma^2) \\ &= -\frac{N}{2}\left\{\log\det 2\pi\Sigma + \mathrm{Tr}\left[\Sigma^{-1}S\right]\right\}\end{aligned}$$

- Model covariance: $\Sigma = WW^T + \sigma^2 I$
- Empirical covariance: $S = \frac{1}{n}X^T X$
- Spectral decomposition: $S = U\Lambda U^T$, $\Lambda_{ii} \geq \Lambda_{jj}$ for $i < j$.
- Maximum likelihood solution: $i = 1, \ldots, m$

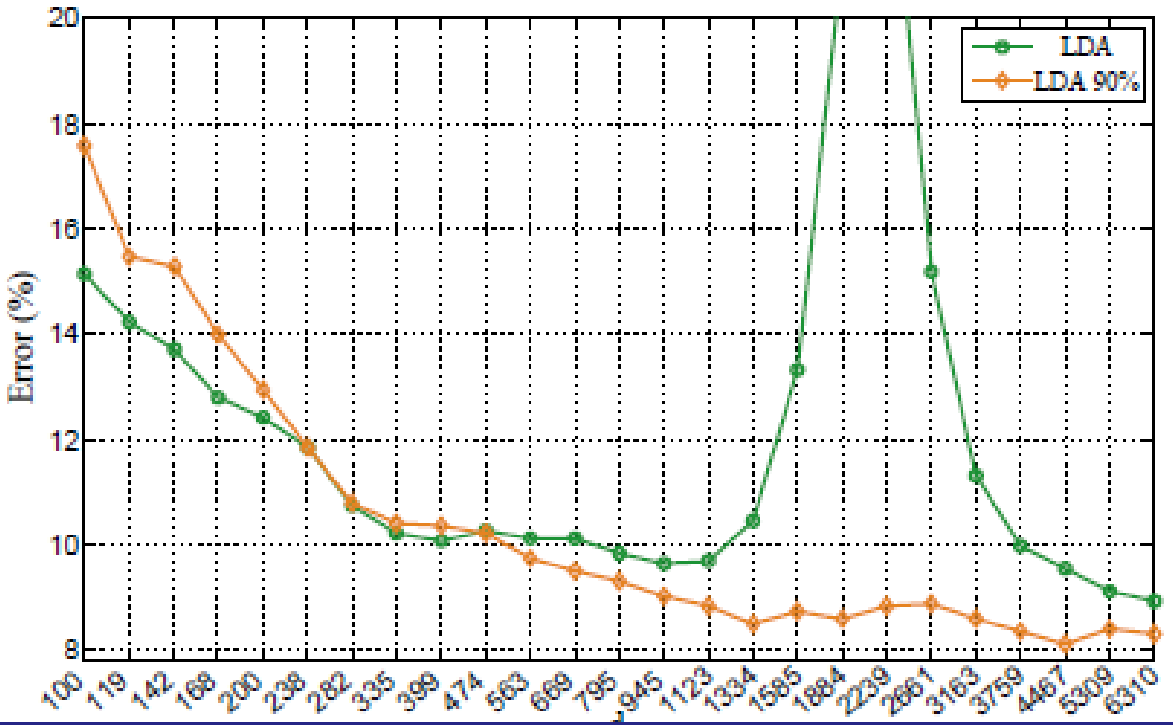
$$w_{i,\mathrm{ml}} = u_i\sqrt{\Lambda_{ii} - \sigma^2_{\mathrm{ml}}}R \qquad \text{and} \qquad \sigma^2_{\mathrm{ml}} = \frac{1}{p-m}\sum_{i=m+1}^{p}\Lambda_{ii}$$

- R arbitrary rotation

- Error versus p for fixed "variance explained"

$$1 - \frac{1}{\mathrm{Tr}\,\Lambda} \sum_{i=m+1}^{p} \Lambda_{ii}$$

- Covariates sorted according to variance.

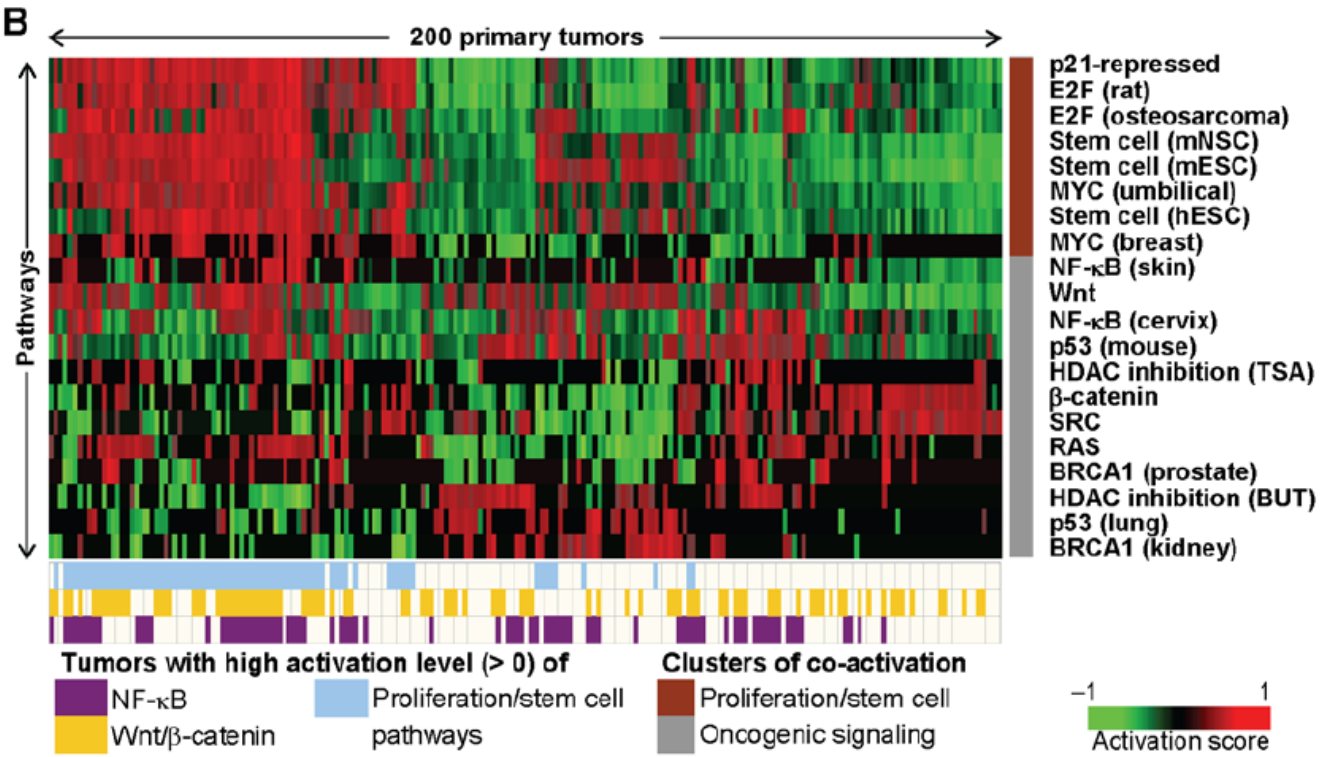


OPEN ACCESS Freely available online PLoS GENETICS

Oncogenic Pathway Combinations Predict Clinical Prognosis in Gastric Cancer

Chia Huey Ooi[1], Tatiana Ivanova[2], Jeanie Wu[2], Minghui Lee[2], Iain Beehuat Tan[3], Jiong Tao[2,4], Lindsay Ward[5], Jun Hao Koo[2], Veena Gopalakrishnan[2], Yansong Zhu[2], Lai Ling Cheng[6], Julian Lee[2], Sun Young Rha[7], Hyun Cheol Chung[7], Kumaresan Ganesan[2], Jimmy So[8], Khee Chee Soo[9], Dennis Lim[10], Weng Hoong Chan[10], Wai Keong Wong[10], David Bowtell[11], Khay Guan Yeoh[12], Heike Grabsch[5], Alex Boussioutas[11,13], Patrick Tan[1,2,14,15]*

1 Duke-NUS Graduate Medical School, Singapore, 2 Cellular and Molecular Research, National Cancer Centre, Singapore, 3 Division of Medical Oncology, National Cancer



- This study inspired us to consider survival analysis and gene expression data.
- Use gene set activation scores instead of gene expression.
- Curated gene sets from Msig database `www.broadinstitute.org/gsea/msigdb/` and scoring according to GAGE (Luo et al, 2009):

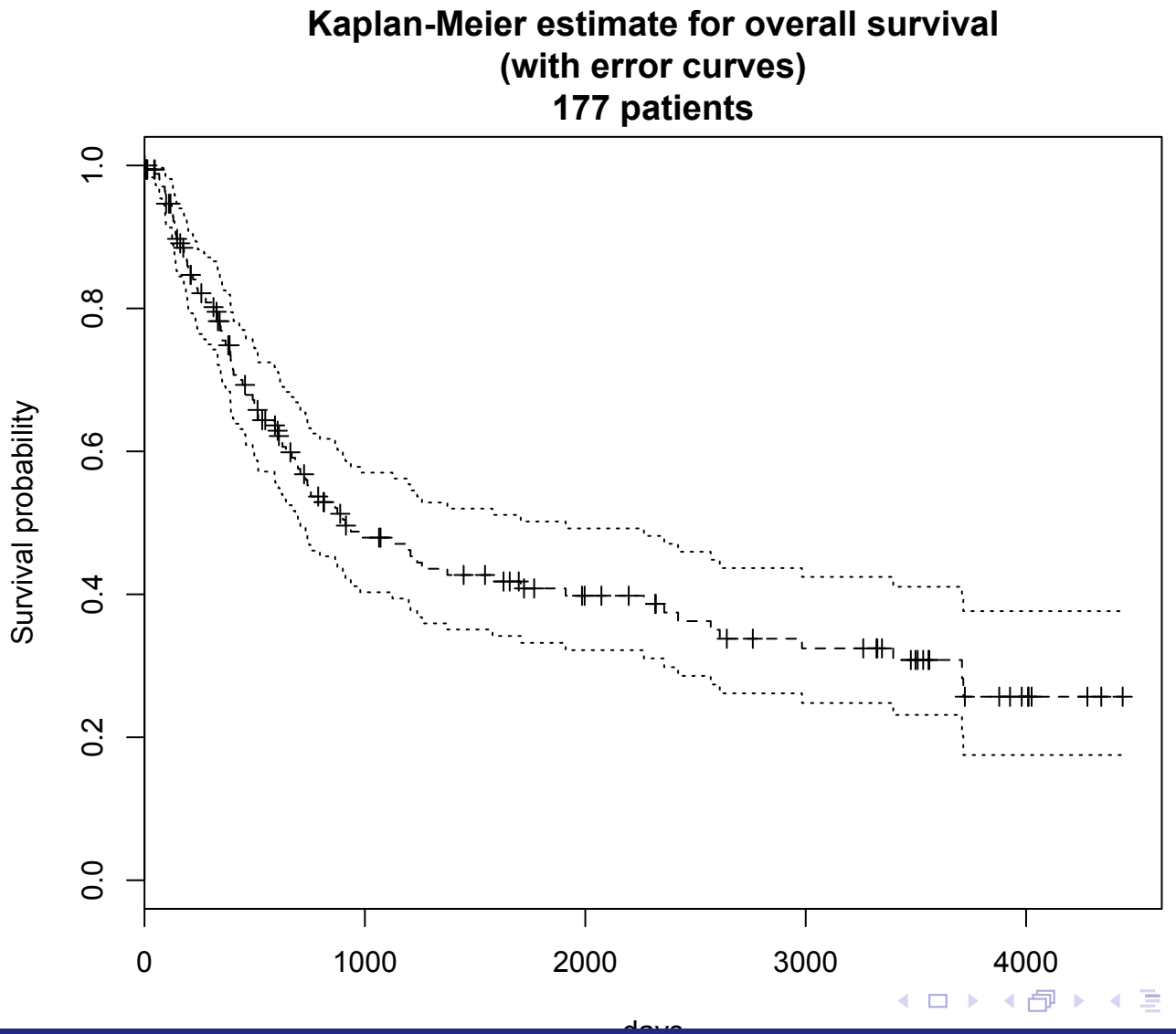
$$t = \frac{m - M}{\sqrt{s^2/n + S^2/n}}$$

- Instead of using a preselected list of gene sets
- we use unbiased search with random survival forest among all gene sets.

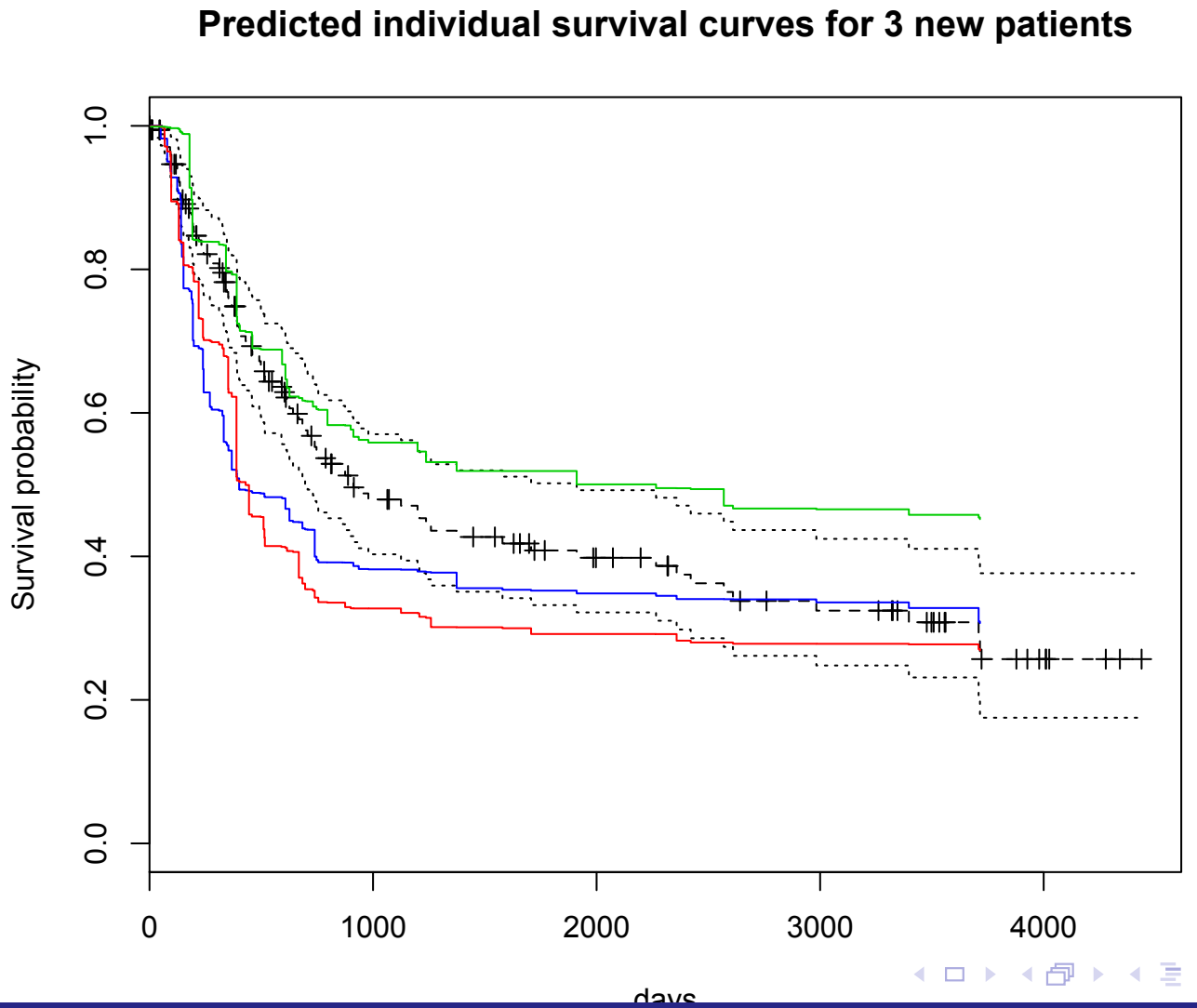
- Random survival forest identifies gene sets which makes biological sense
- according to literature and may be validated by clinicians:

```
[1] "OHM_EMBRYONIC_CARCINOMA_UP_DN_ud"
[2] "ST_FAS_SIGNALING_PATHWAY_b"
[3] "TONKS_TARGETS_OF_RUNX1_RUNX1T1_FUSION_SUSTAINED_IN_MONOCYTE_UP_DN_ud"
[4] "WEIGEL_OXIDATIVE_STRESS_BY_HNE_AND_H2O2_u"
[5] "BIOCARTA_IL1R_PATHWAY_b"
[6] "SNIJDERS_AMPLIFIED_IN_HEAD_AND_NECK_TUMORS_u"
[7] "GARGALOVIC_RESPONSE_TO_OXIDIZED_PHOSPHOLIPIDS_LIGHTYELLOW_UP_DN_ud"
```

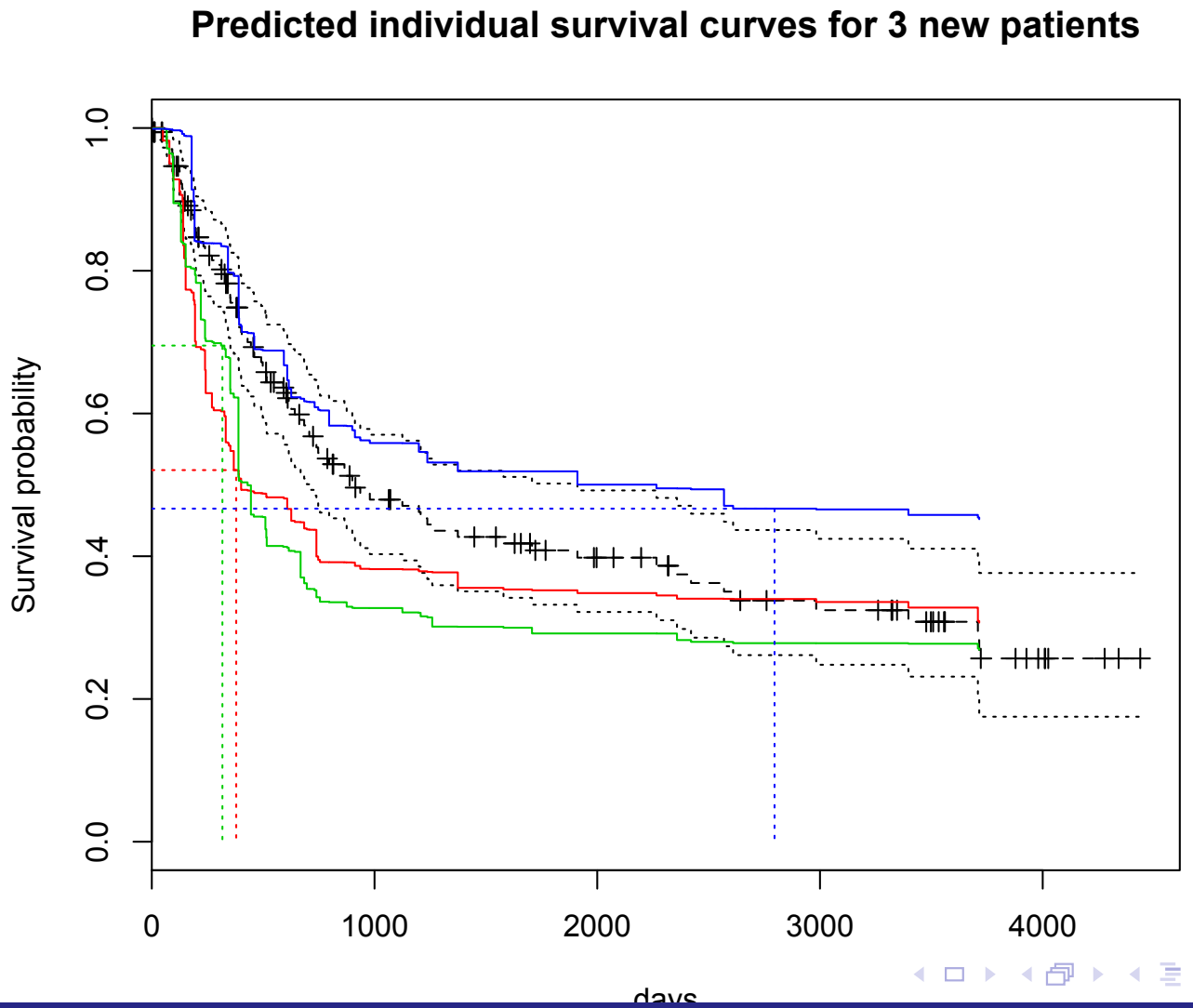
Individualized patient survival curves



Individualized patient survival curves

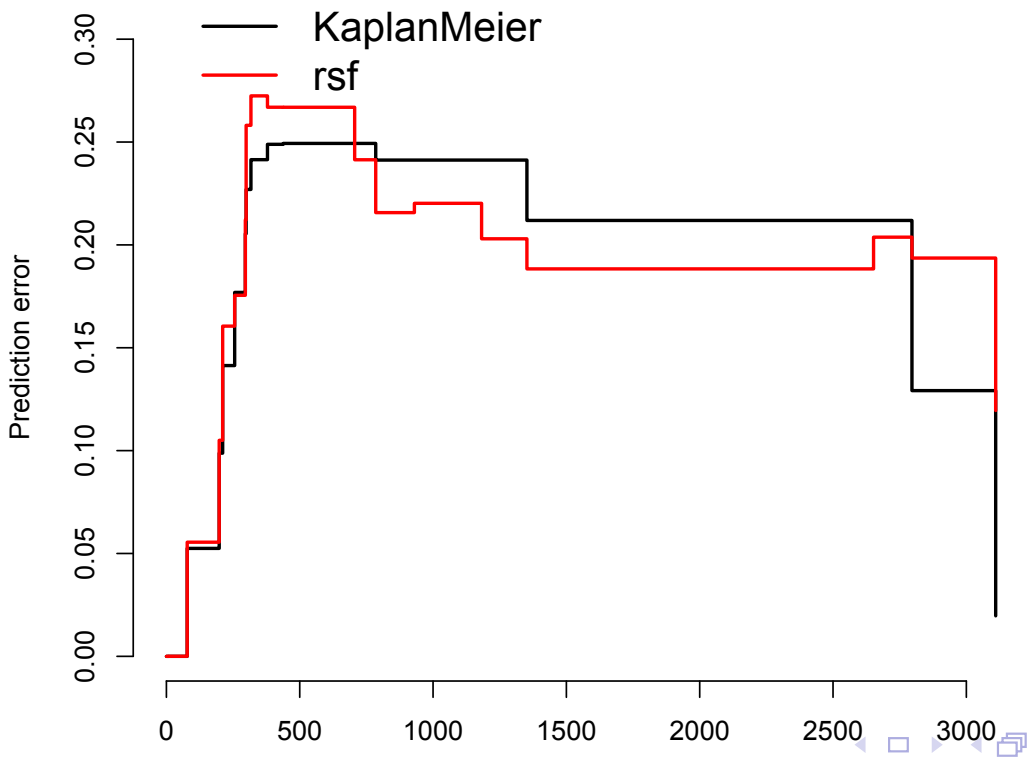


Individualized patient survival curves



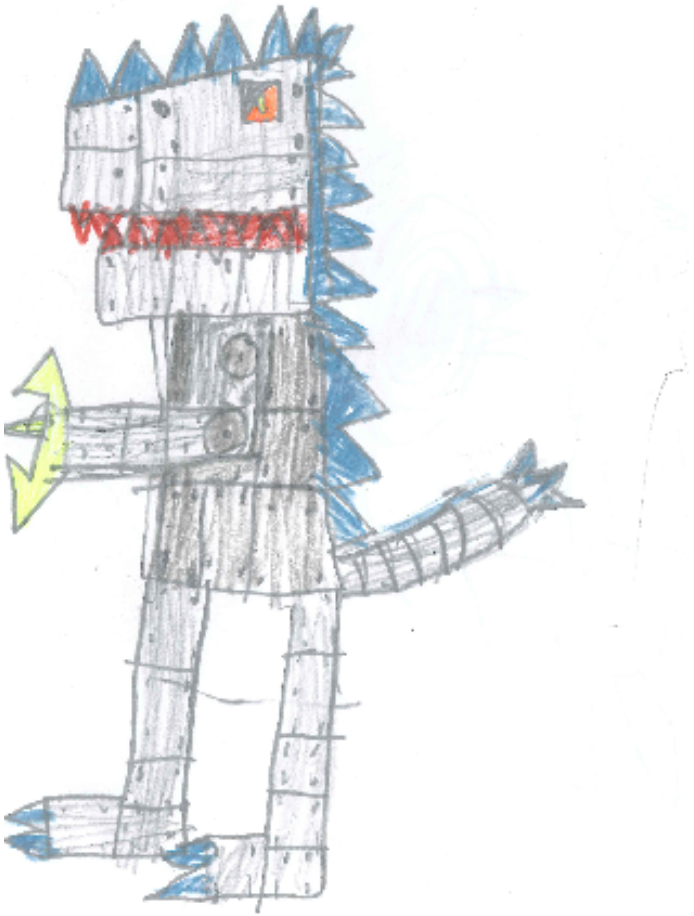
Individualized patient survival curves

- Hold-out (19 of 198) Brier scores
- Kaplan Meier 0.201
- RSF 0.195



Outlook

- Proof of concept of close collaboration with clinicians
- Next steps towards individualized treatment
  - better models
  - richer genomic data
  - predict response to treatment
- International Genomics Consortium (ICG) and The Cancer Genome Atlas (TCGA) provide unprecedented amount of genomic data, but close collaboration with clinicians necessary to get detailed clinical data.



Bioinformatics

- Bogumil Kaczkowski
- Ricardo Henao
- Tomas Martin-Bertelsen
- Anders Krogh

Genomic Medicine, Riget

- Finn Cilius Nielsen
- Lennart Friis-Hansen
- Rehannah Borup

Oncology, Riget

- Gedske Daugaard
- Anne Kirstine Moller