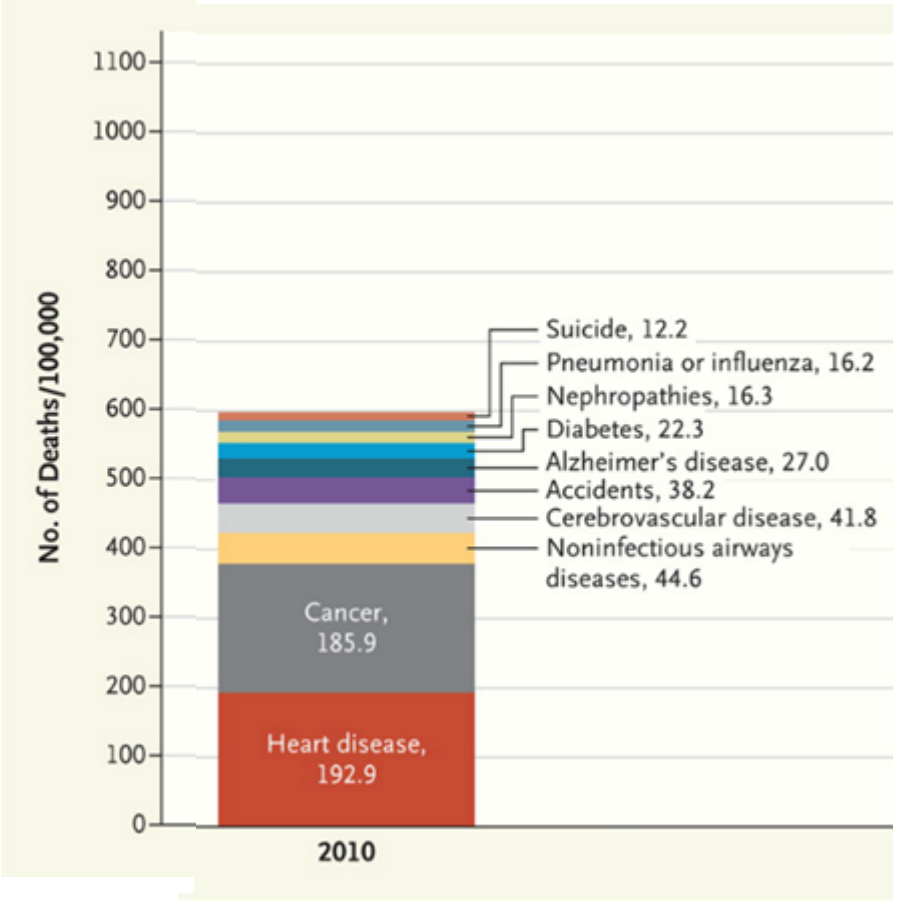
# Cancer bioinformatics: identification of diagnostic and prognostic biomarkers from gene expression data

Garrett Dancik
University of Colorado Denver
December 17, 2012

# Leading causes of death in the U.S. (2010)




**Lifetime probability of developing cancer**

**Males: 50%**
**Females: 33%**

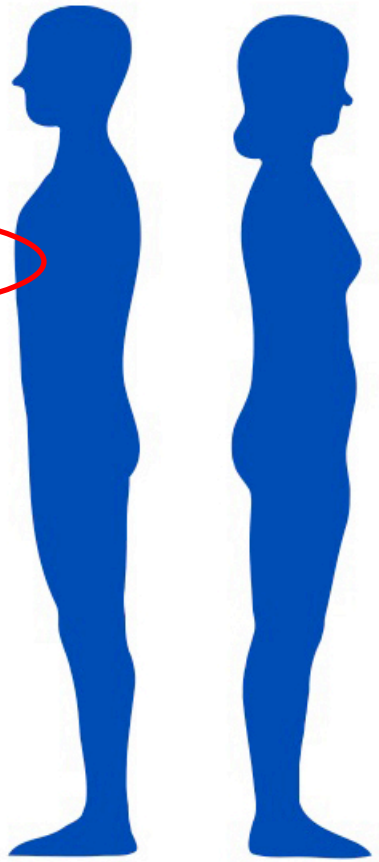
Jones et al. N Engl J Med 2012; 366:2333-2338

# Leading cause of death in the U.S. (2020)

- Coronavirus?
  - https://public.flourish.studio/visualisation/1727839/
  - For about a week, there have been more deaths per day from coronavirus in the U.S. than the average number of daily deaths from any other condition

# 2010 Estimated US Cancer Cases*

| Men 789,620 | | Women 739,940 | |
|---|---|---|---|
| Prostate | 28% | 28% | Breast |
| Lung & bronchus | 15% | 14% | Lung & bronchus |
| Colon & rectum | 9% | 10% | Colon & rectum |
| Urinary bladder | 7% | 6% | Uterine corpus |
| Melanoma of skin | 5% | 5% | Thyroid |
| Non-Hodgkin lymphoma | 4% | 4% | Non-Hodgkin lymphoma |
| Kidney & renal pelvis | 4% | 4% | Melanoma of skin |
| Oral cavity | 3% | 3% | Kidney & renal pelvis |
| Leukemia | 3% | 3% | Ovary |
| Pancreas | 3% | 3% | Pancreas |
| All Other Sites | 19% | 20% | All Other Sites |



Source: American Cancer Society, 2010.

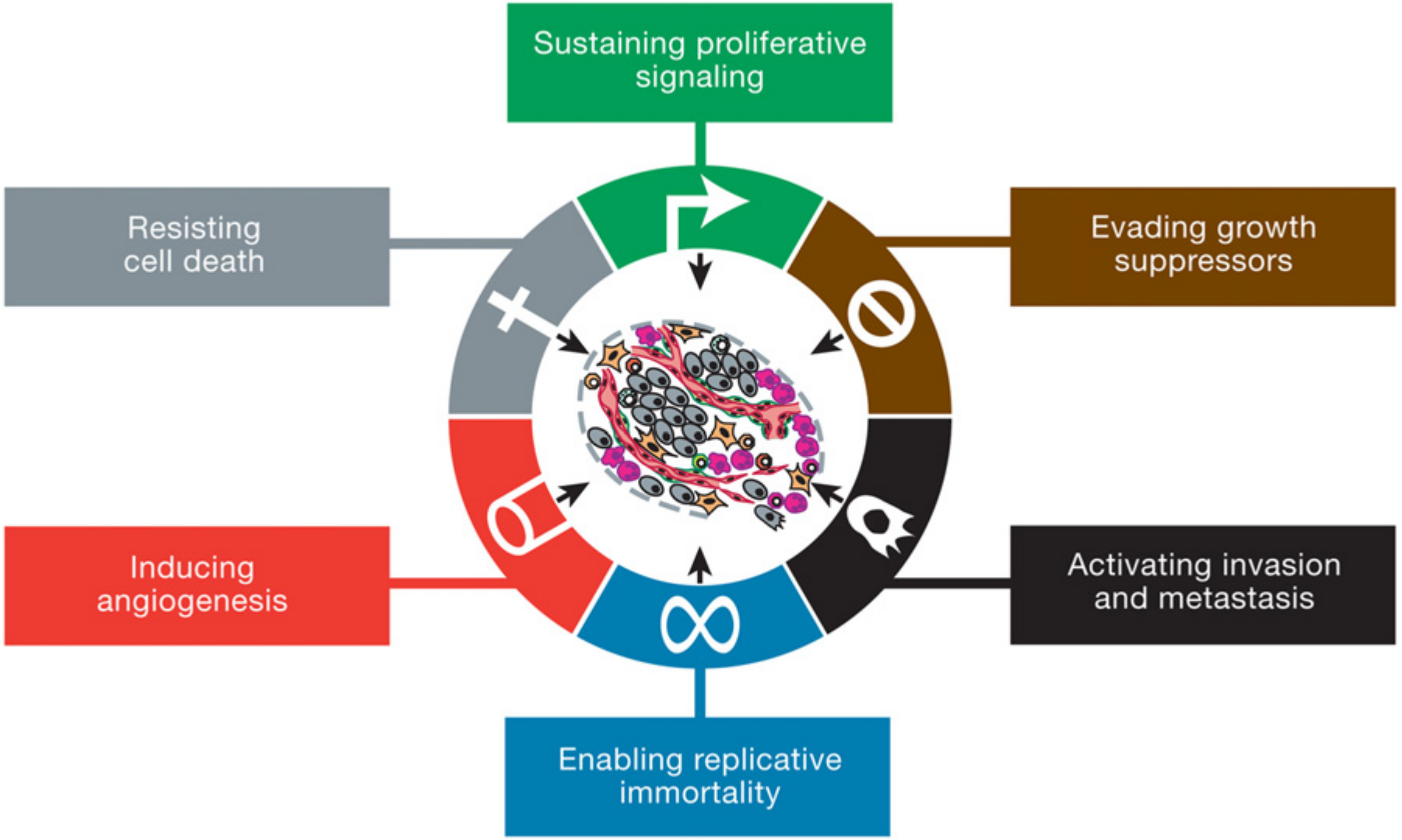
Source: American Cancer Society, 2010.

*Excludes basal and squamous cell skin cancers and in situ carcinomas except urinary bladder.

http://www.cancer.org

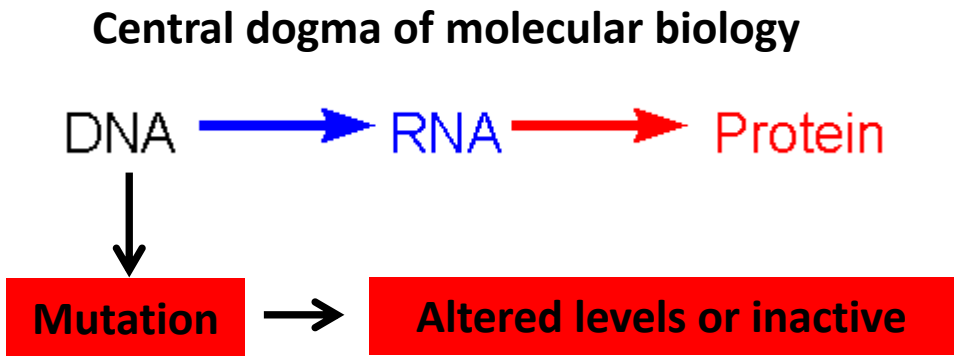
# What is cancer?

# Hallmarks of cancer

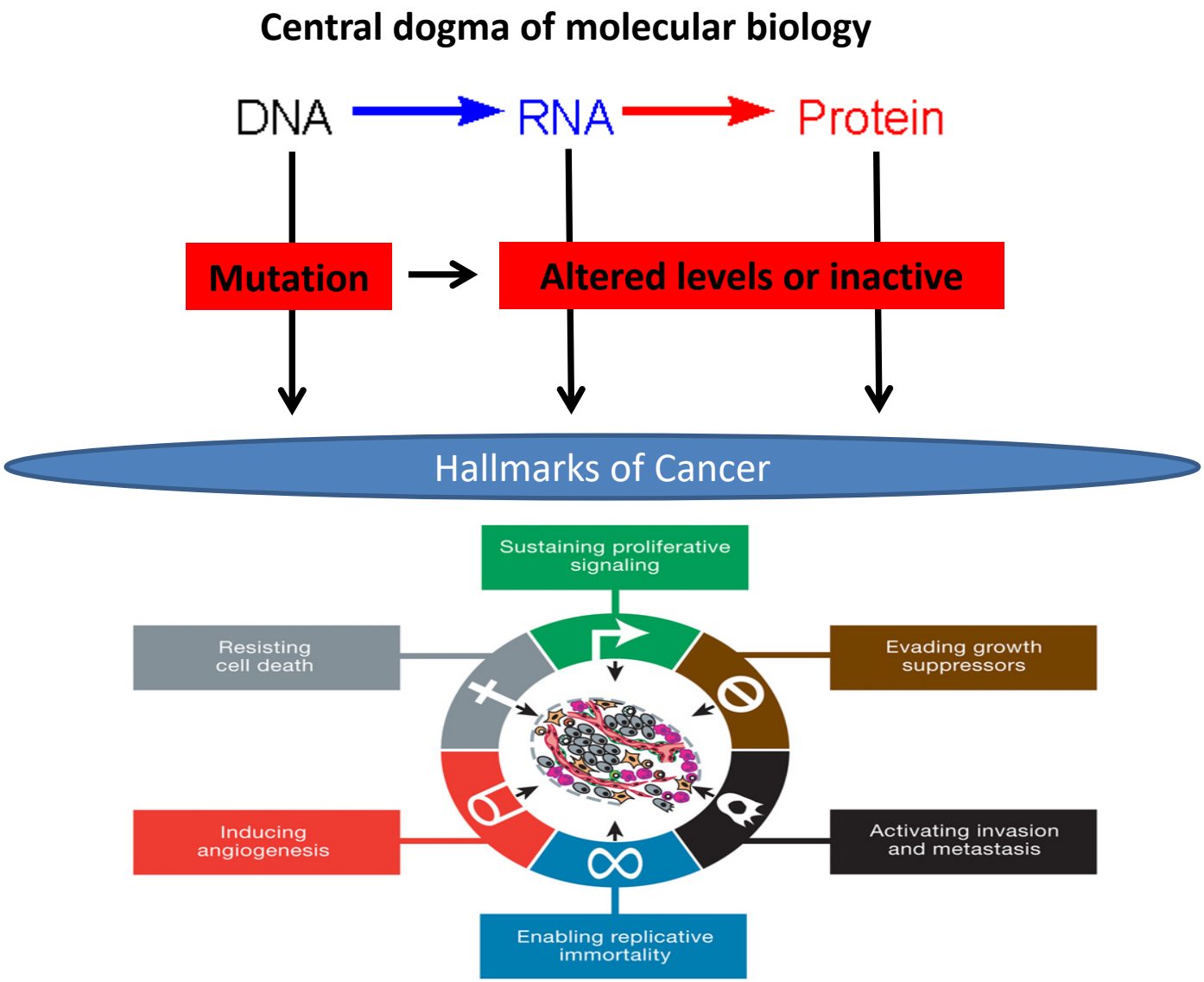



Hanahan and Weinberg. Cell 2011; 144(5):646–674.

# Cancer is a genetic disease

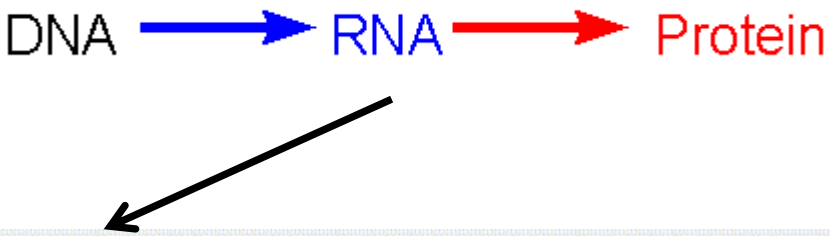


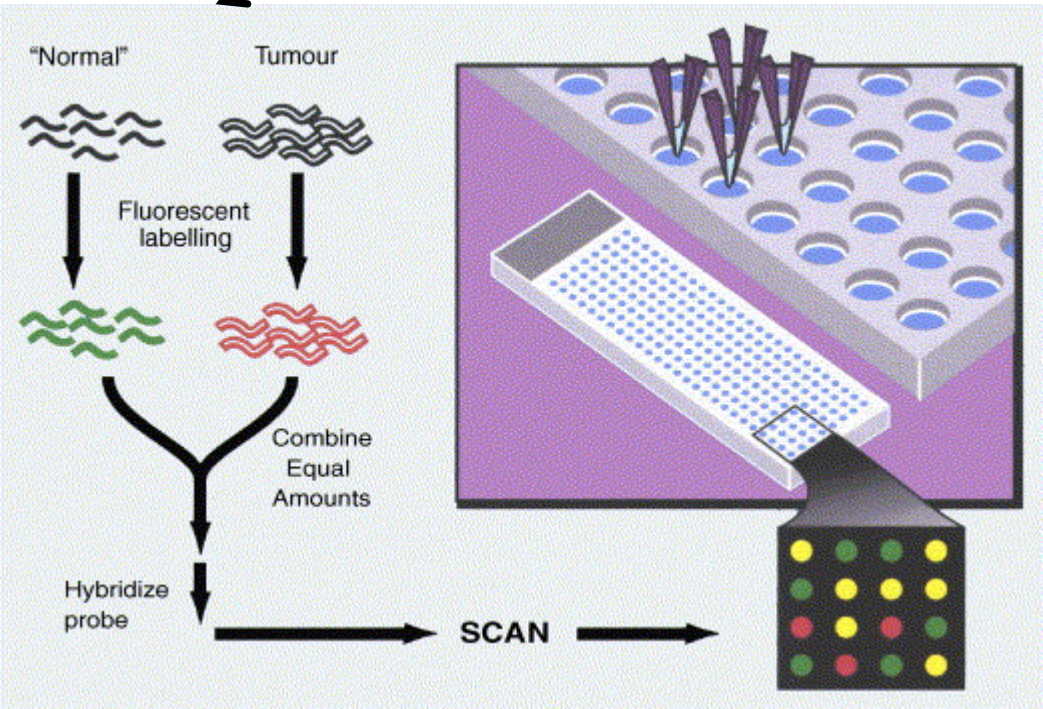
# Cancer is a genetic disease



# Gene expression profiling by microarray

**Central dogma of molecular biology**

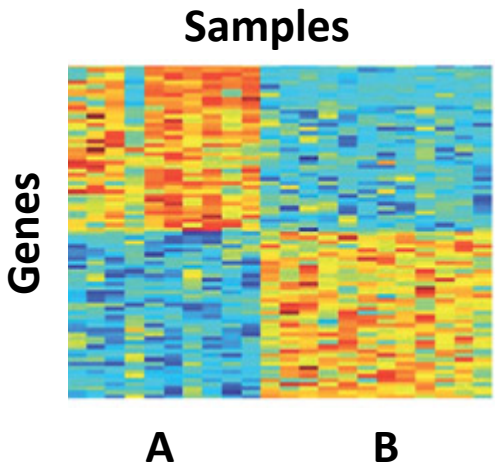



Bowden and Brennan. Br J Oral Maxillofac Surg 2004; 42(6): 42-545

# Biomarkers and personalized medicine

**Gene expression profiles**




Class comparison
- A,B: clinical variable or outcome
  - Tumor type
  - High risk vs. low risk (survival)
  - Responders vs. non-responders
- Classification of new samples:
  - Gene signature
  - Classification method:
    - KNN, SVM, PCA, NCC, etc.

- **Diagnostic biomarker:** a gene or gene signature that is predictive of a clinical variable (e.g., tumor grade)
- **Prognostic biomarker:** a gene or gene signature that is predictive of disease outcome (e.g., survival)

Nature Reviews | Genetics

# A framework to select clinically relevant cell lines by establishing their molecular similarity with patient tumors

Dancik *et a*l. Can. Res. 2011; 71:7398

# Background and motivation

- Cell lines as model systems in cancer
  - Characterization of molecular mechanisms of disease
  - Characterization of activity of therapeutic agents
  - High throughput drug discovery programs
- But….cell lines do not always represent patient tumors
  - Adaptation in culture
  - Cross-contamination
- *In vitro* (cell line) drug sensitivity often does not correlate with drug efficacy in patients

# Motivation and approach

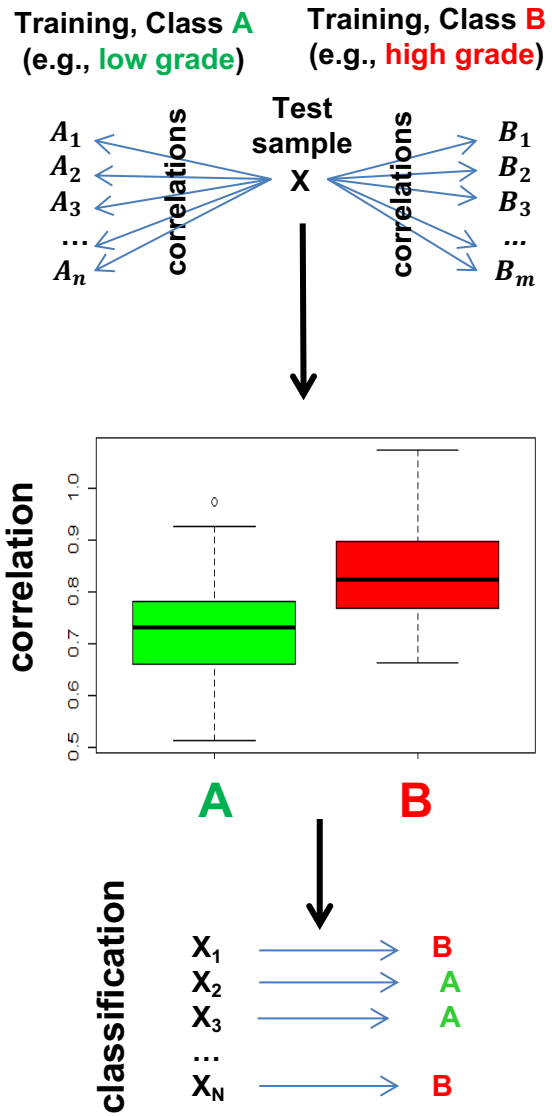
- Objective: identify and **select clinically relevant cell lines** based on their **gene expression profiles**
  - Classify a panel of 36 bladder (BLA-36) cell lines
- Classification objectives
  - **Tissue of origin** (from 10 epithelial tumors)
  - **Stage** (NMI vs. MI)
  - **Grade** (high grade vs. low grade)
  - **Disease specific survival** (high vs. low risk)

# Spearman rank correlation classification method (SRCCM)

**Step 1:** Calculate Spearman correlation between the gene expression profile of test sample (*X*) and gene expression profiles of samples from known classes (classes *A* and *B* here) using a relevant gene signature

**Step 2:** Classification/prediction: the test sample is assigned the class with the highest mean correlation (class B here).

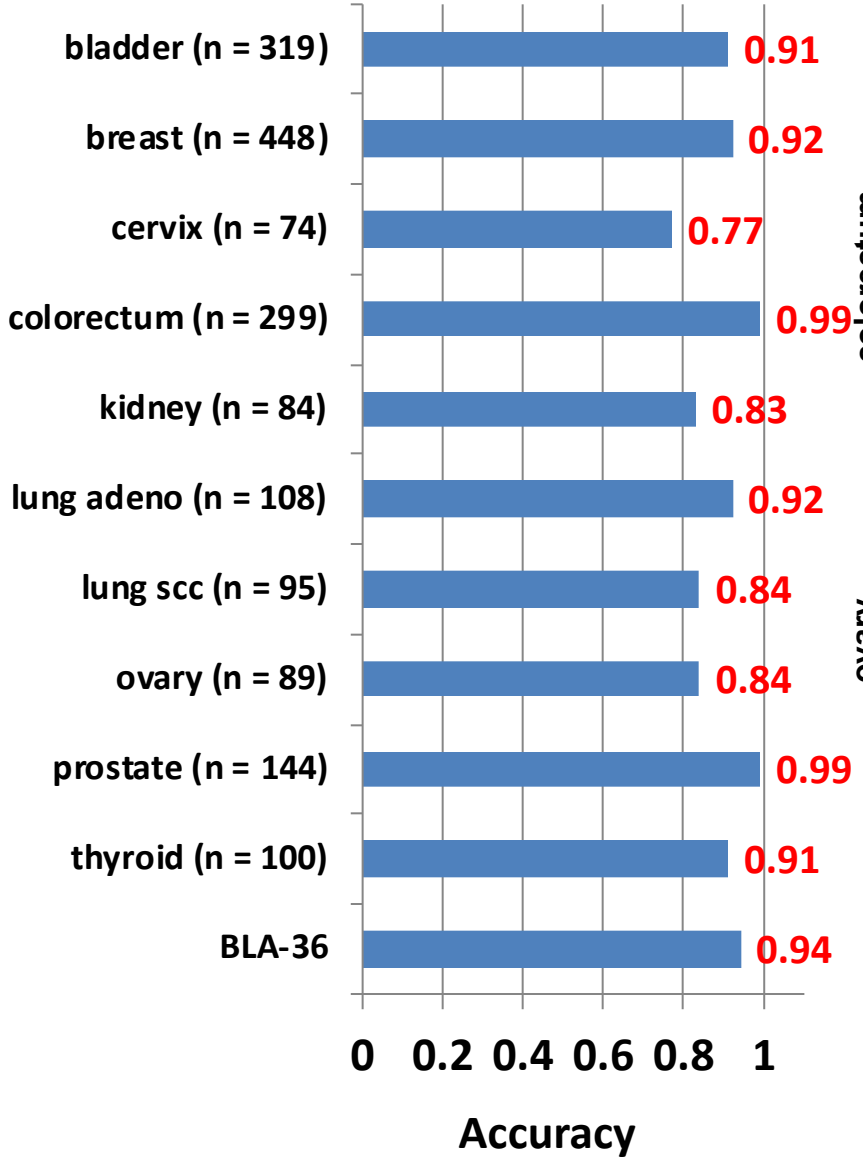
**Step 3:** Repeat steps 1-2 to classify all test samples

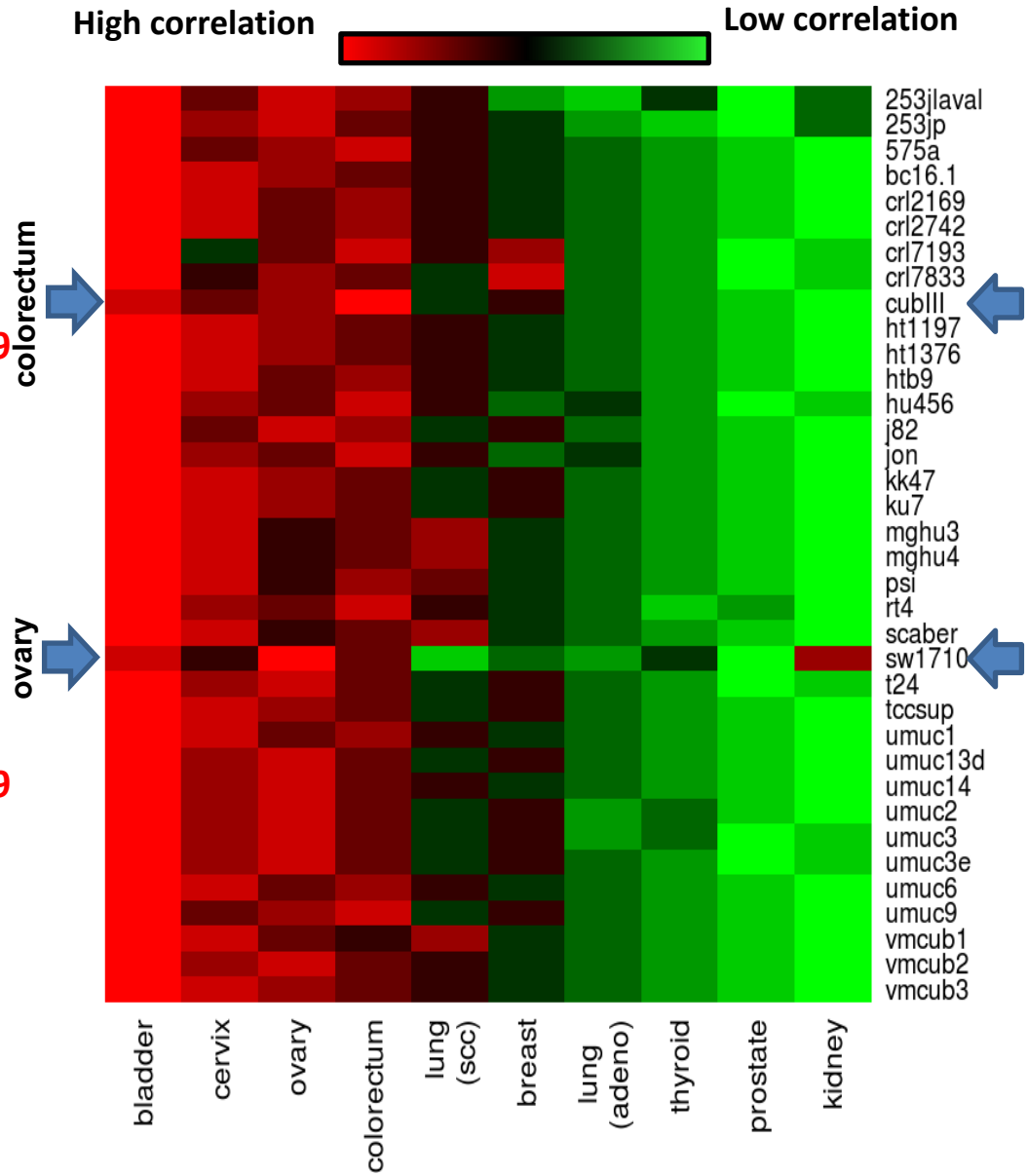


# Tissue of origin classification

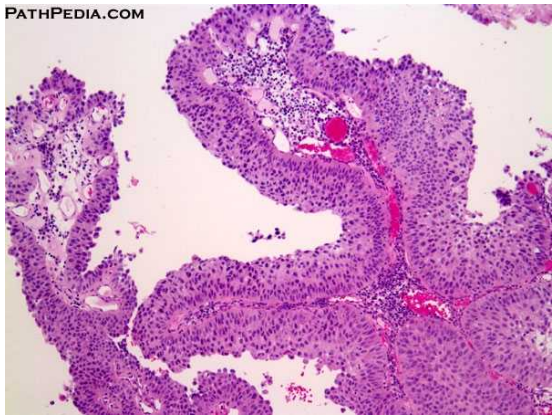
- Clinical relevance of tissue of origin
  - Chemotherapy and radiation therapy efficacy depends on tumor type (Kemp CJ, et al. *Cancer Res* 2001;61(1):327-332)
  - Metastatic site preference is tissue specific
- Do cell lines resemble their derived tissues
  - Previous studies: Only 57% of NCI-60 cell lines resemble presumed tissue of origin (Sandberg R, Ernberg I. *PNAS* 2005;102(6):2052-2057).
  - Survey of 500 leukemia-lymphoma cell lines finds 15% mislabeled. (Drexler HG et al. *Leukemia*. 2003;17(2):416-426)

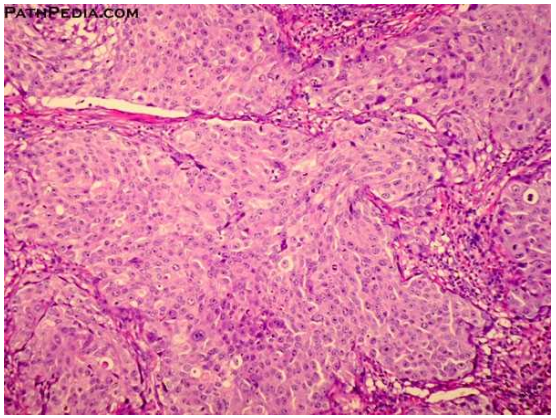
# Tissue of origin classification



# Grade classification



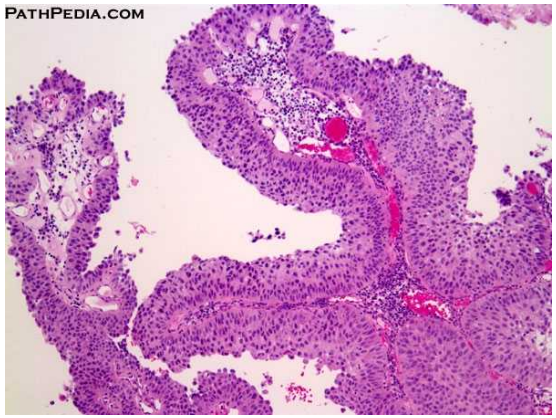



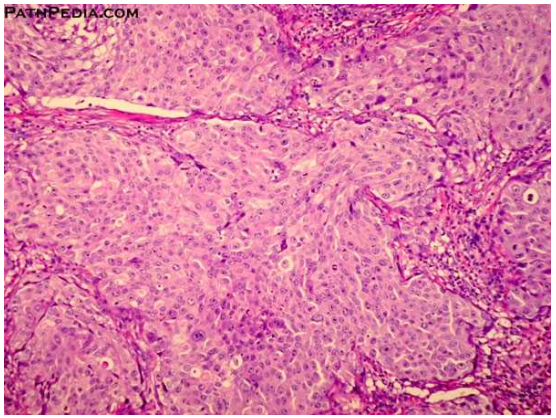

**Low grade** **vs.** **High grade**

Well differentiated vs. poorly differentiated

| Dataset | SRCCM accuracy |
|---|---|
| Lindgren (LOOCV) | 0.875 |
| SC | 0.813 |

# Grade classification



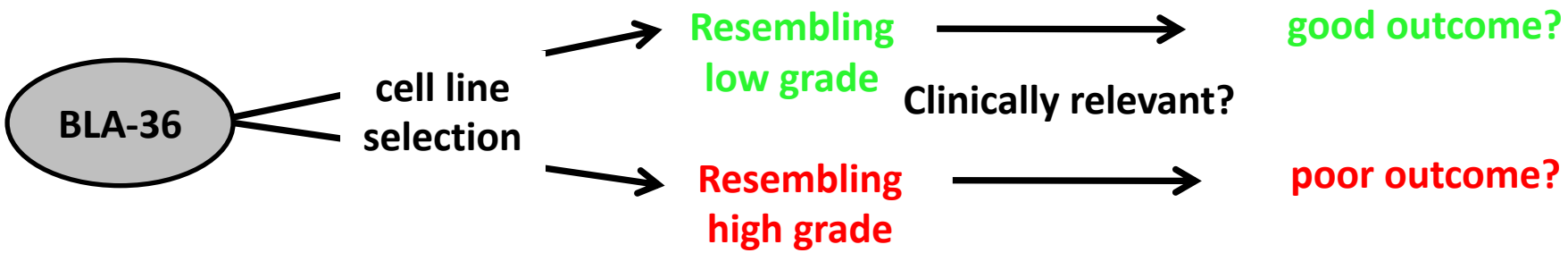



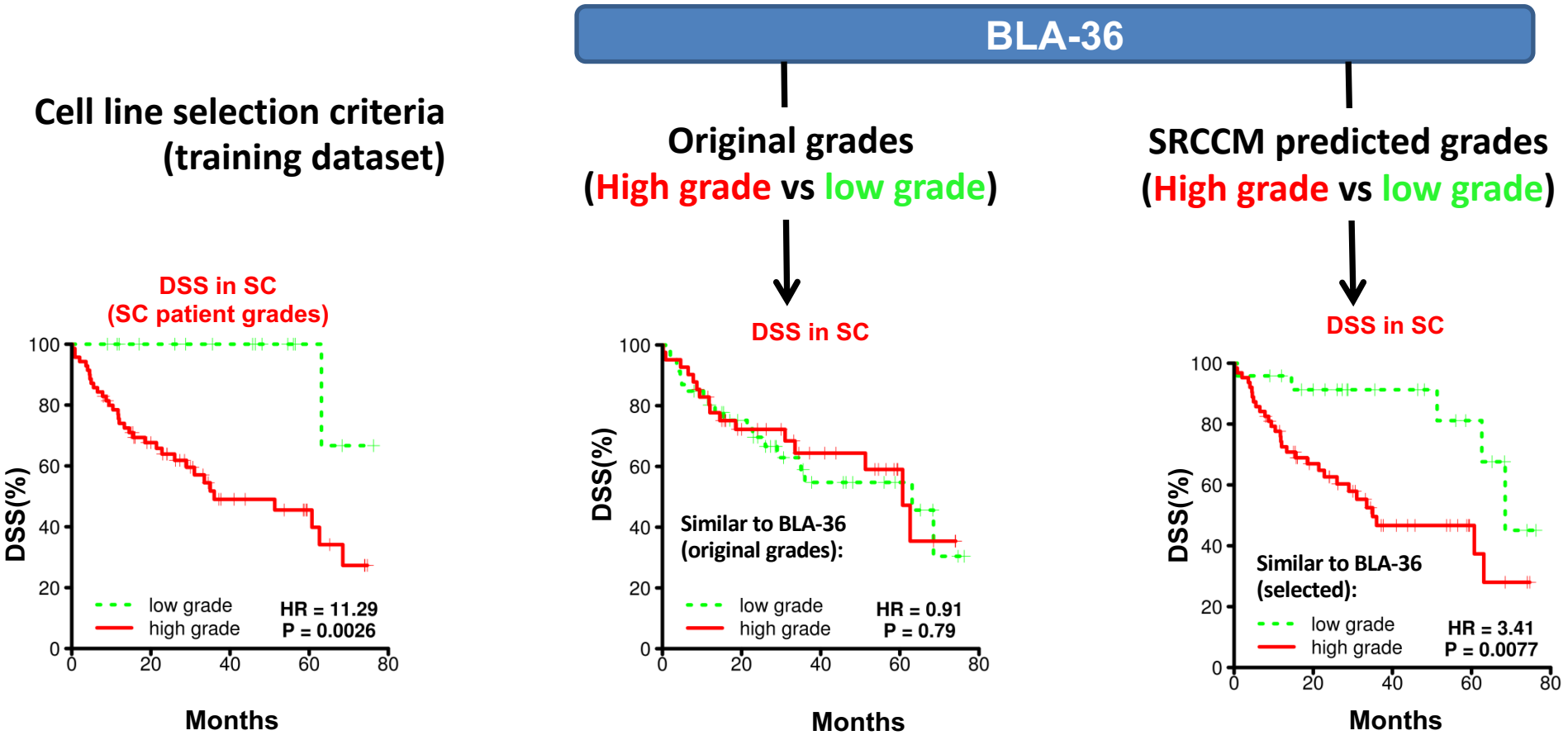

**Low grade** **vs.** **High grade**

Well differentiated vs. poorly differentiated

| Dataset | SRCCM accuracy |
|---|---|
| Lindgren (LOOCV) | 0.875 |
| SC | 0.813 |
| BLA-36 | .571 |

Original tumor grades no longer correlate with survival; correlation is restored through cell line selection via SRCCM

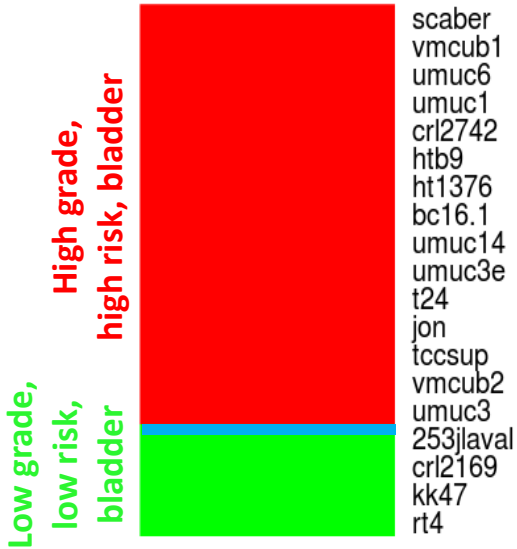


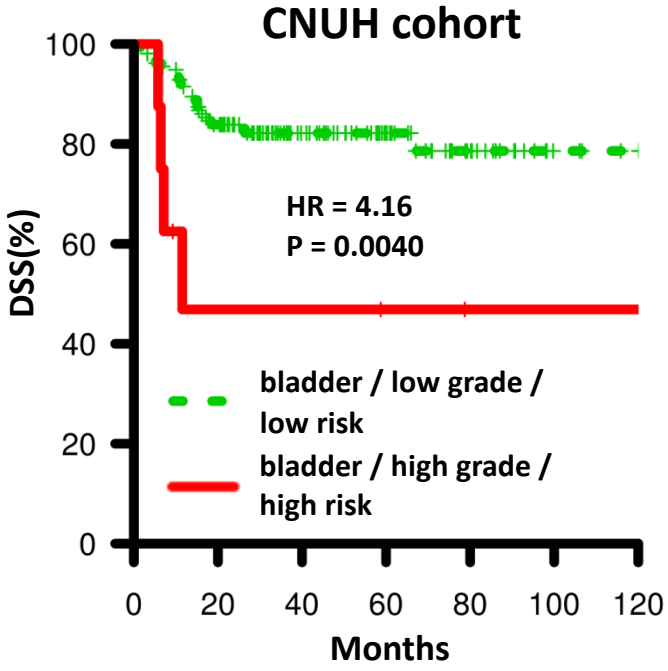
# Selection of the most clinically relevant cell lines by survival risk, grade, and tissue type

## CELL LINE PANEL

Selection of cell lines based on characteristics of interest using SRCCM




Training: cell lines
Testing: new patient cohort



# Summary

- SRCCM algorithm for classification and cell line model selection
- BLA-36
  - Grade: accuracy < 60%, suggesting that many cell lines no longer resemble original tumors with respect to grade
  - Original tumor grade no longer correlates with survival; correlation is restored through SRCCM selection
- Software: Correlation classification method (CCM) http://cran.r-project.org/web/packages/CCM/index.html

# Acknowledgements



- Theodorescu Lab
  - Dan Theodorescu, MD, PhD (PI)
  - Yuanbin Ru, PhD
  - Chuck Owens (lab technician)
- Funding: NIH CA075115.

Thank You!

# Cancer grade and staging

- Tumor grade
  - Normal vs. abnormal
  - Low vs. high grade
- Tumor stage
  - How far has the cancer spread
- Bladder cancer stages
  - Non-muscle invasive (NMI): Ta, T1
    - 5 year survival rate of ~ 90%
    - Progression rate of ~ 20%
  - Muscle invasive (MI): T2-T4
    - 5 year survival rate ~ 50%

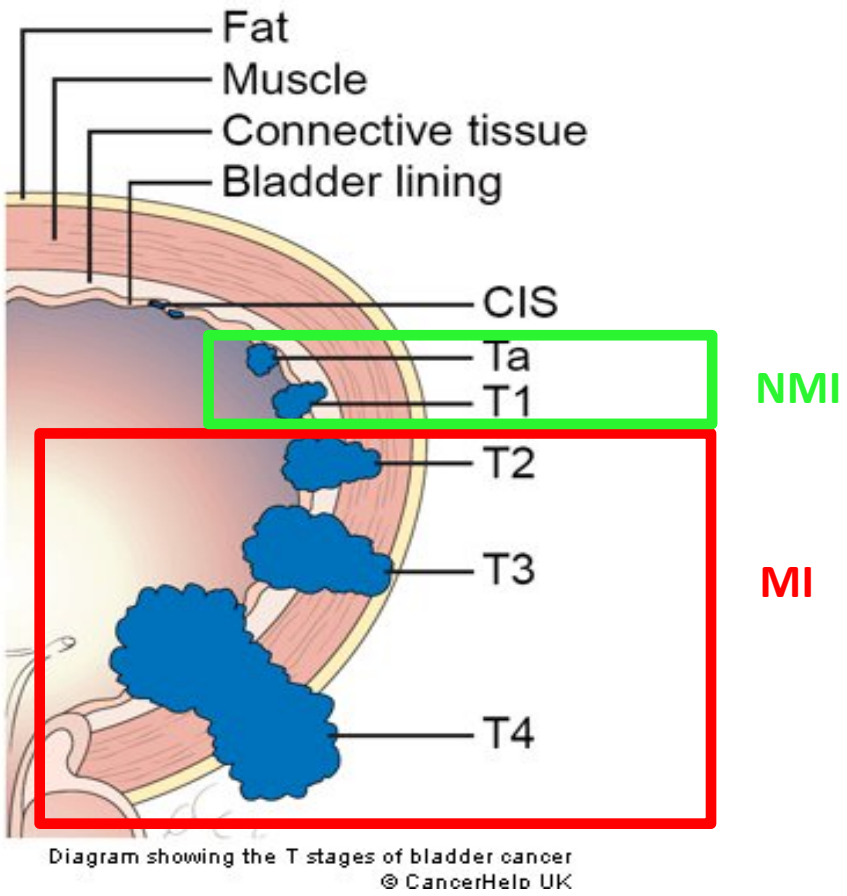


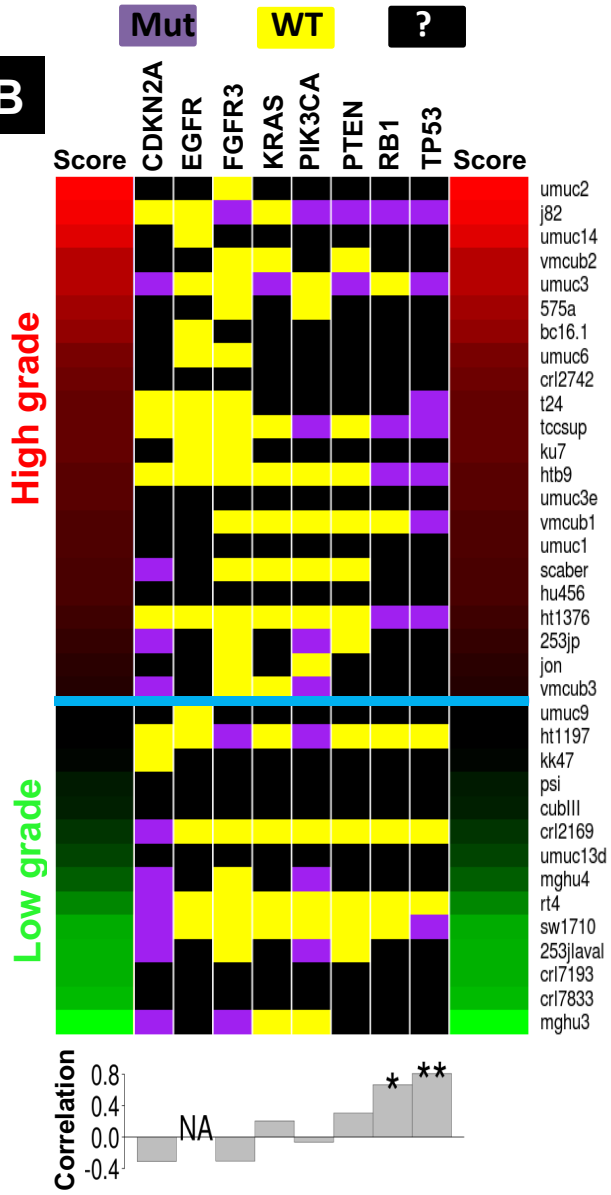

Image: http://www.cancerhelp.org.uk/type/bladder-cancer/treatment/bladder-cancer-stage-and-grade

# Bladder cancer grade classification

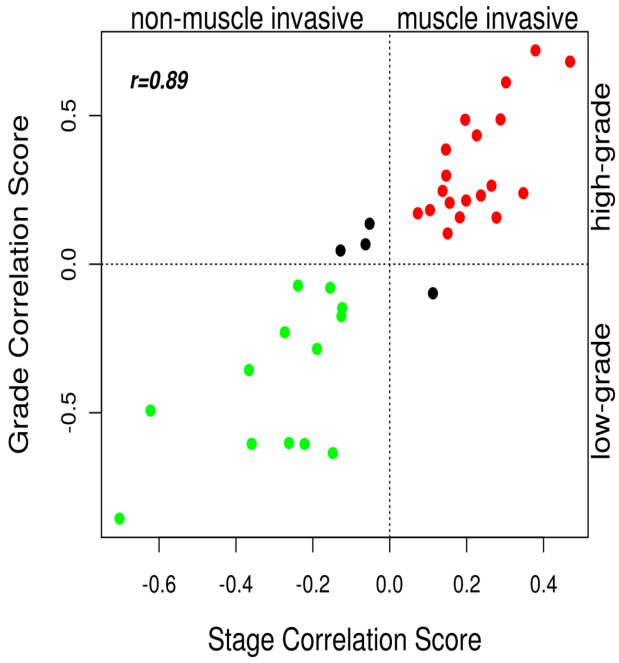
**A**

| | Accuracy |
|---|---|
| LOOCV in Lindgren | 0.875 |
| Independent Validation in SC | 0.813 |
| BLA-36 | .571 |

**B**




**C**



# Presentation Tips

- You are presenting your paper:
  - background, significance, objective, methods, results
- Almost every slide is a picture (or table)
  - From the internet (with reference)
  - From another publication (with reference)
  - From original research

# Presentation Tips

- Presentation is written out and practiced ahead of time
- You do NOT read off of the page
- Additional slides are included at the end
  - For results or background not presented do to time
  - To answer possible questions