Memorial Sloan Kettering Cancer Center

Genetic risk modeling using machine learning to predict radiotherapy complications and identify key biological correlates

Jung Hun Oh
Assistant Attending

Department of Medical Physics,
Memorial Sloan Kettering Cancer Center

Acknowledgement

GWAS Study
- **Sangkyu Lee**, MSK
- **Joseph Deasy**, MSK
- **Barry Rosenstein**, Mount Sinai School of Medicine
- **Sarah Kerns**, University of Rochester Medical Center
- **Harry Ostrer**, Albert Einstein College of Medicine
- **Jonine Bernstein**, MSK
- **Xiaolin Liang**, MSK
- **Meghan Woods**, MSK
- **Anne Reiner**, MSK

Radiomics
- **Evangelia Katsoulakis**, MSK
- **Yao Yu**, MSK
- **Aditya P. Apte**, MSK
- **Nancy Y. Lee**, MSK
- **Nadeem Riaz**, MSK
- **Vaios Hatzoglou**, MSK

Memorial Sloan Kettering Cancer Center

Part 1. Genome-wide association studies

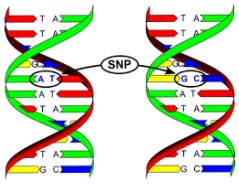
Memorial Sloan Kettering Cancer Center

Background

- Our goal in GWAS is to predict how the risk of radiation toxicity varies between patients, based on germ line genome characteristics
- Previous single-SNP models suffer from multiple-testing correction due to a large number of SNPs being evaluated
  - Important SNPs may fail to achieve genome-wide significance
- Therefore, we have taken a **many-SNP** approach to developing predictive models, using machine learning methods

Memorial Sloan Kettering Cancer Center

Single nucleotide polymorphisms (SNPs)



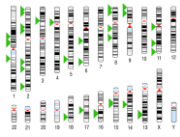
https://www.tubascan.eu

Memorial Sloan Kettering Cancer Center

Genome-wide association studies (GWAS)

Patient 1 …CAAGGTA…
Patient 2 …CAATGTA…
Patient 3 …CAATGTA…
Patient 4 ...CAAGGTA…

Single Nucleotide Polymorphism (SNP) is genetic variation at **one location** in a DNA sequence.



Genome-Wide Association Studies (GWAS) find associations between a disease and such variations across **the whole genome**.

Memorial Sloan Kettering Cancer Center

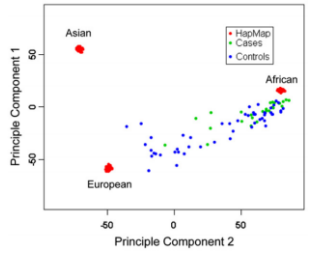
Coding

- Wild type homozygous: 0
- Heterozygous: 1
- Mutant homozygous: 2

- Coded as the number of rare alleles

Memorial Sloan Kettering Cancer Center

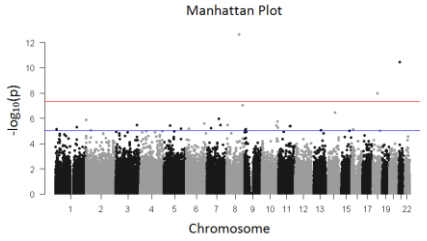
Population structure



Kerns 2010, Int. Rad. Onc. Biol. Phys

Memorial Sloan Kettering Cancer Center

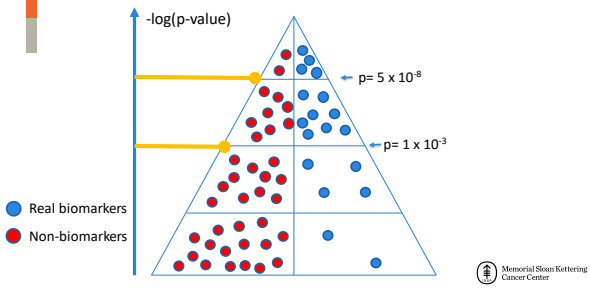
Statistical analysis



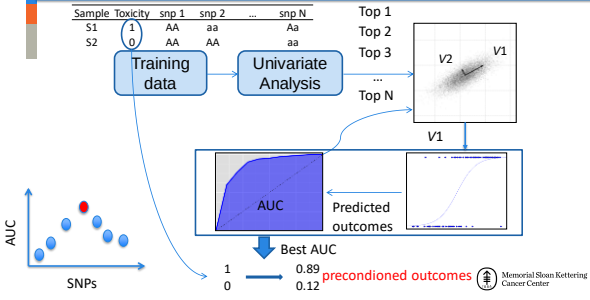
❖ Genome-wide significance level = 5×10^{-8}

Memorial Sloan Kettering Cancer Center

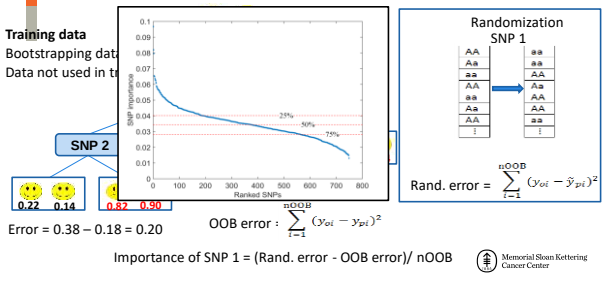
Filtering



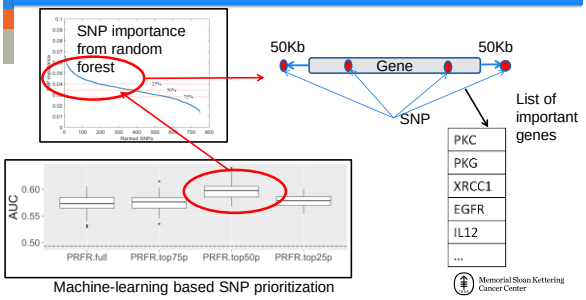
Preconditioning



SNP importance



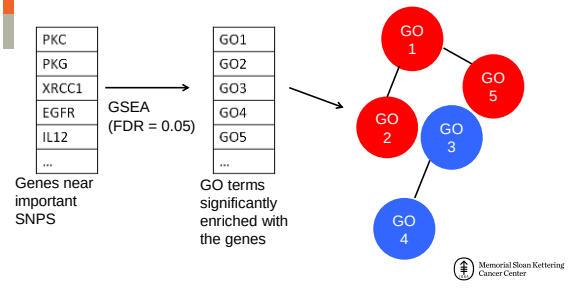
A subset of features



Machine-learning based SNP prioritization

Memorial Sloan Kettering Cancer Center

Gene ontology analysis

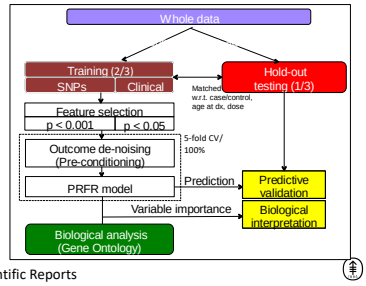


Genes near important SNPS

GO terms significantly enriched with the genes

Memorial Sloan Kettering Cancer Center

Preconditioning Random Forest Regression (PRFR)



Oh et al., 2017. Scientific Reports

Memorial Sloan Kettering Cancer Center

Data

- 368 patients with prostate cancer
  - DNA was genotyped using Affymetrix genome wide array (v6.0)
- Quality control
  - Missing rate > 5% of samples
  - MAF < 5%
  - Hardy-Weinberg equilibrium (p-value < 10^{-5})
  - 613,496 SNPs remained

Memorial Sloan Kettering Cancer Center

Rectal bleeding

Oh et al., 2017. Scientific Reports

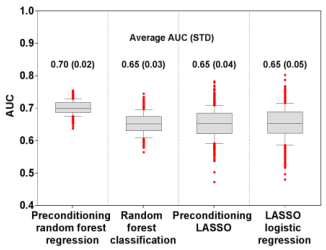
Memorial Sloan Kettering Cancer Center

Data preprocessing

- Outcome: rectal bleeding
  - RTOG ≤ 1 (coded 0) vs RTOG ≥ 2 (coded 1)
- Data split: rectal bleeding
  - Training dataset
    - 243 samples
    - 49 events
    - 749 SNPs (p< 0.001; Chi-square test)
  - Validation dataset
    - 122 samples
    - 25 events
- 5-fold cross validation with 100 iterations

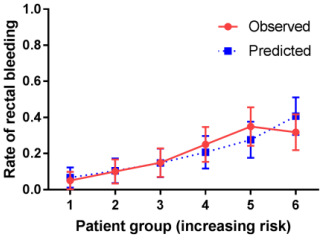
Memorial Sloan Kettering Cancer Center

Performance



Memorial Sloan Kettering Cancer Center

Performance



Memorial Sloan Kettering Cancer Center

Biological analysis

#	GO Processes/Genes
1	Regulation of ion transport CACNA1D,CCL13,DPP6,GCK,GNB4,GPR63,HOMER1,IL1RAPL1,JDP2, KCNIP4,KCNJ6,NLGN1,NOS1AP,PDE4D,PRKCB,PRKG1,VDR
5	Regulation of transmembrane transporter activity CACNA1D,GNB4,HOMER1,NLGN1,NOS1AP,PDE4D,PRKCB,PRKG1

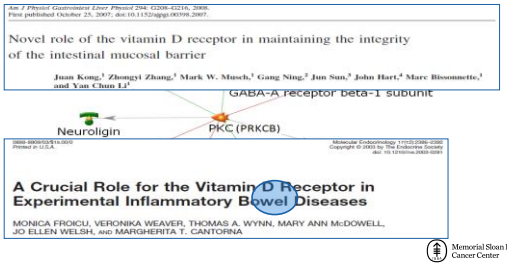
GASTROENTEROLOGY 2005;129:591–608

Epidermal Growth Factor Partially Restores Colonic Ion Transport Responses in Mouse Models of Chronic Colitis

DECLAN F. McCOLE, GERHARD ROGLER, NISSI VARKI, and KIM E. BARRETT
Department of Medicine, School of Medicine, University of California, San Diego, San Diego, California

Memorial Sloan Kettering Cancer Center

Protein-protein network



Memorial Sloan Kettering Cancer Center

Erectile dysfunction toxicity

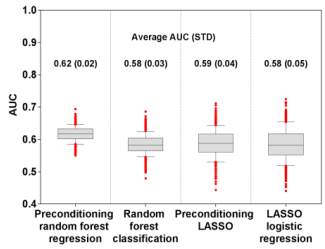
Memorial Sloan Kettering Cancer Center

Data preprocessing

- Outcome: erectile dysfunction
  - SHIM ≤ 7 (coded 1) vs SHIM ≥ 16 (coded 0)
- Data split
  - Training dataset
    - 157samples
    - 88 events
    - 367 SNPs (p< 0.001; Chi-square test)
  - Validation dataset
    - 79 samples
    - 45 events

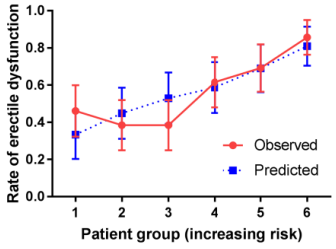
Memorial Sloan Kettering Cancer Center

Performance



Memorial Sloan Kettering Cancer Center

Performance



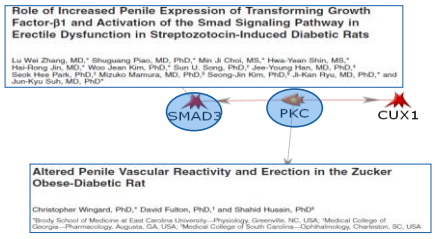
Memorial Sloan Kettering Cancer Center

Biological analysis

Ranking	GO Processes	FDR
1	negative regulation of heart contraction	8.376E-10
2	negative regulation of blood circulation	2.180E-08
3	neutrophil chemotaxis	5.026E-08
4	neutrophil migration	5.883E-08
5	granulocyte chemotaxis	9.684E-08
6	granulocyte migration	1.300E-07
7	positive regulation of locomotion	2.631E-07
8	regulation of muscle system process	5.510E-07
9	regulation of muscle contraction	5.510E-07
10	positive regulation of cell migration	8.960E-07

Memorial Sloan Kettering Cancer Center

Protein-protein network



Memorial Sloan Kettering Cancer Center

Genitourinary Toxicity

Lee et al., 2018. Int J Rad Oncol Biol Phys

Memorial Sloan Kettering Cancer Center

GU symptoms

Symptom Category*	Symptom Name	Training Set Size	Testing Set Size	Event Rate (%)	Modeled?
Irritative (Storage)	Frequency	119	60	23	O
	Urgency	161	81	16	O
	Nocturia	111	56	17	O
Obstructive (Voiding)	Intermittency	164	82	10	X
	Weak Stream	149	75	18	O
	Straining	196	98	5	X
Post-micturition	Incomplete Emptying	168	84	10	X

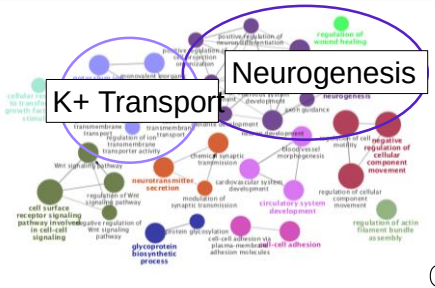
Memorial Sloan Kettering Cancer Center

GU modeling

Symptom Name	Training Set Size	Event rate	# SNPs p<0.001	#clinical p<0.05	PRFR Performance	
					AUC	P-value
Frequency	119	0.23	539	0	0.64	0.06
Urgency	161	0.16	758	0	0.53	0.38
Nocturia	111	0.17	977	1	0.55	0.33
Weak stream	149	0.18	823	0	0.70	0.01

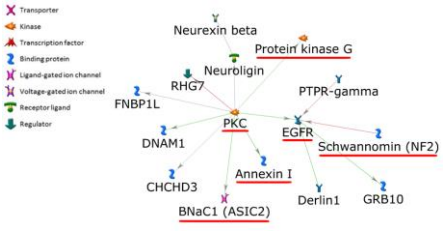
Memorial Sloan Kettering Cancer Center

Gene ontology analysis



Memorial Sloan Kettering Cancer Center

Protein-protein network



Memorial Sloan Kettering Cancer Center

Summary

- We developed a promising method using whole-genome data for deriving predictive risk models for predicting late radiation-induced toxicites
- SNP -> Gene -> Pathway analysis
- Found that biological correlates are tissue specific
- Other studies
  - Secondary contralateral breast cancer
  - Fatigue in breast cancer
  - Weight gain in breast cancer

Memorial Sloan Kettering Cancer Center

Part 2. Radiogenomics

Memorial Sloan Kettering Cancer Center

Data

- **Imaging data:**
  - Pre-treatment CT scans in head and neck cancer were downloaded from the TCIA
  - 77 CT scans were analyzed
  - Using CERR, 104 radiomic features were evaluated
    - Apte, 2018. Medical Physics
  - Feature stability test
  - Volume dependent features were removed
  - 67 features were analyzed

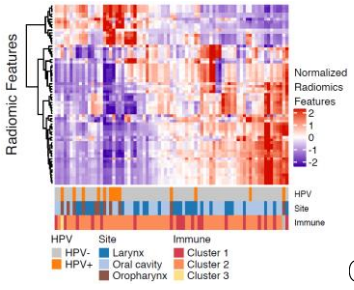
Memorial Sloan Kettering Cancer Center

Data

- **Biological data:**
  - **Recurrent gene mutations**
    - cBioPortal (https://www.cbioportal.org/)
  - **Tumor subtypes**
    - Broad Institute FireBrowse (http://firebrowse.org/)
  - **Immune infiltrates**
    - Thorsson, 2018. Immunity
  - **HPV status**
    - Nulton, 2017. Oncotarget

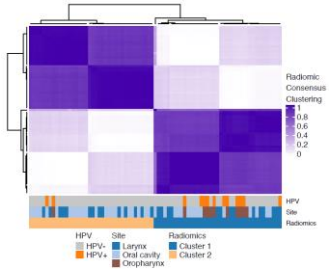
Memorial Sloan Kettering Cancer Center

Clustering



Memorial Sloan Kettering Cancer Center

Consensus clustering

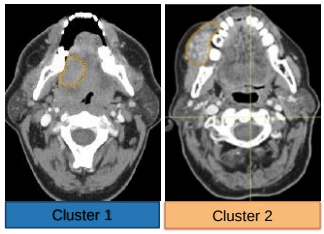


Subsite	Cluster 1	Cluster 2
Oral cavity	15	23
Larynx	14	14
Oropharynx	10	1
P-value	**0.0096**	

HPV	Cluster 1	Cluster 2
Positive	11	2
Negative	28	36
P-value	**0.0127**	

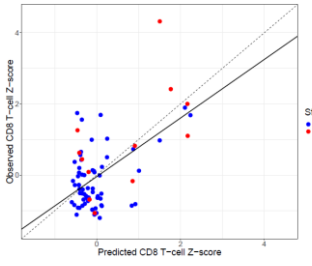
Memorial Sloan Kettering Cancer Center

Representative scans



Memorial Sloan Kettering Cancer Center

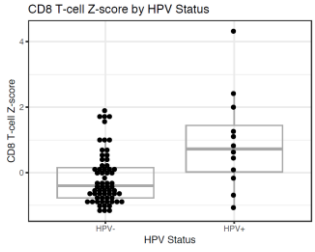
CD8 Prediction using Random Forest



	R^2	P-value
All	0.3	**<0.0001**
HPV+	0.36	**0.0405**
HPV-	0.16	**0.0012**

Memorial Sloan Kettering Cancer Center

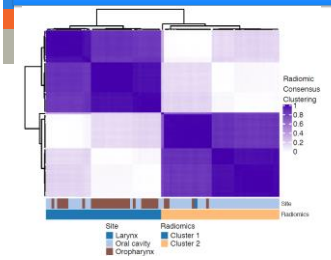
HPV status vs CD8 T-cell



P-value = **0.0061**

Memorial Sloan Kettering Cancer Center

Validation



Subsite	Cluster 1	Cluster 2
Oral cavity	14	37
Larynx	0	1
Oropharynx	27	4
P-value	1.3×10^{-7}	

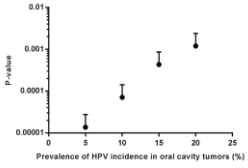
HPV	Cluster 1	Cluster 2
Positive	24	3
Negative	17	39
P-value	4.0×10^{-7}	

- 83 cases (MSKCC)
- Oral cavity: 51, Larynx: 1, Oropharynx: 31

Memorial Sloan Kettering Cancer Center

Difference of HPV status

- **Sensitivity test**
  - Randomly select 5%, 10%, 15%, and 20% of samples in oral cavity tumors
  - Assign them to HPV-positive
  - Iterate 1000 times



- **Prevalence of HPV incidence: 5%**
  P=1.4×10^{-5} (95% CI: 1.3×10^{-5}-1.5×10^{-5})

Memorial Sloan Kettering Cancer Center

Summary

- Found clearly separable radiomic clusters
- The differences in subsite and HPV status between the two radiomic clusters were statistically significant
- Built a machine learning model to predict CD8 T-cell
- Validation using an independent dataset

Memorial Sloan Kettering Cancer Center

Thank you

ohj@mskcc.org

Memorial Sloan Kettering Cancer Center