Cancer-specific Splicing and Splicing QTLs Revealed By Pan-Cancer Genome Analysis

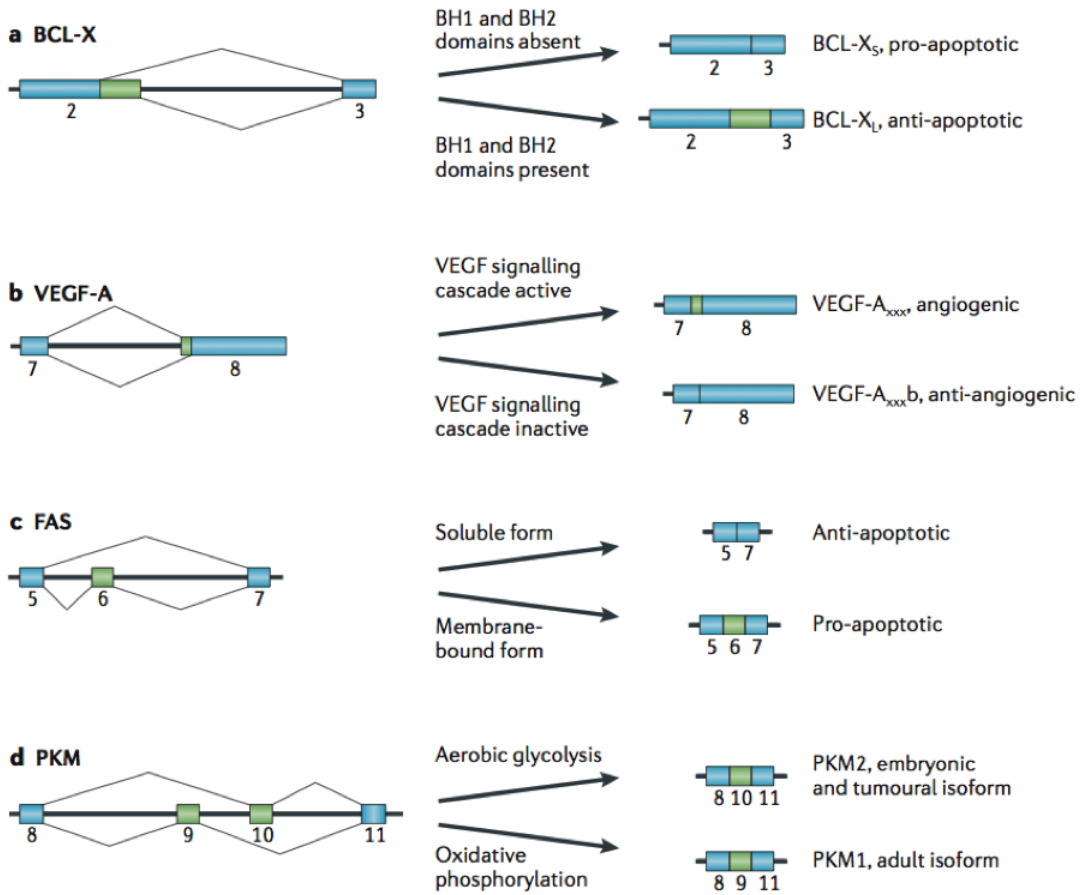
Kjong-Van Lehmann

Memorial Sloan-Kettering Cancer Center

Rätsch Lab, New York City, USA

TCGA Symposium, May 13, 2014

Splicing and Drug sensitivity



S. Bonnal et al. (2012); Nature Reviews Drug Discovery

Analysis Across Multiple Cancer Types

Goals

1. Identify cancer-specific splicing patterns
2. Identify variants regulating splicing in same gene (cis)
3. Identify variants regulating splicing in other cancer genes (trans)

TCGA provides RNA-seq and matching exome data

- RNA-seq $\rightsquigarrow$ Find & quantify splicing events
- Exome $\rightsquigarrow$ Identify variants in exons & flanking intronic regions

Problem: Non-uniform processing (alignments & variant calling)

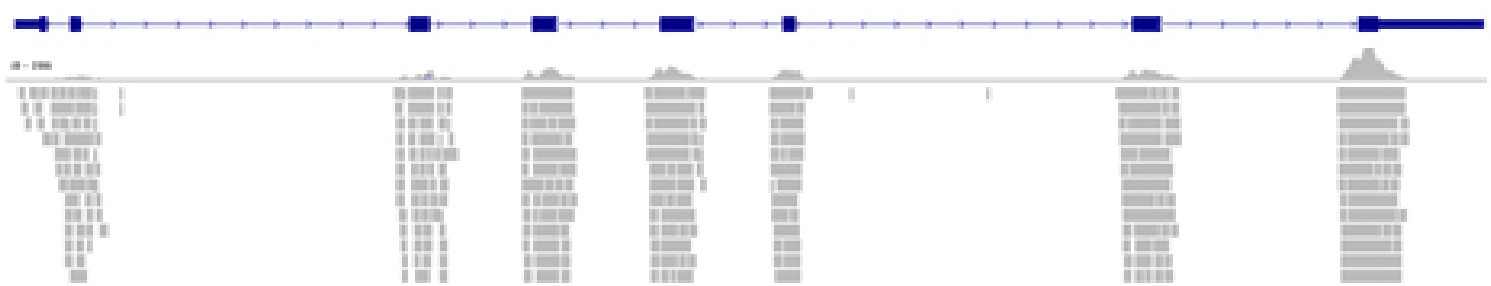
Analysis Across Multiple Cancer Types

Goals

1. Identify cancer-specific splicing patterns
2. Identify variants regulating splicing in same gene (cis)
3. Identify variants regulating splicing in other cancer genes (trans)

TCGA provides RNA-seq and matching exome data

- RNA-seq $\rightsquigarrow$ Find & quantify splicing events
- Exome $\rightsquigarrow$ Identify variants in exons & flanking intronic regions



Problem: Non-uniform processing (alignments & variant calling)

Analysis Across Multiple Cancer Types

Goals

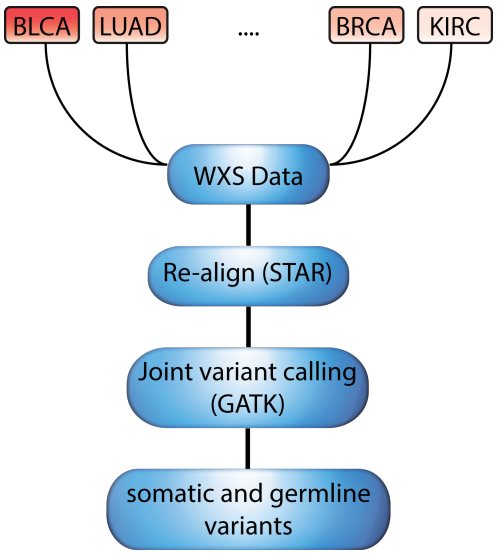
1. Identify cancer-specific splicing patterns
2. Identify variants regulating splicing in same gene (cis)
3. Identify variants regulating splicing in other cancer genes (trans)

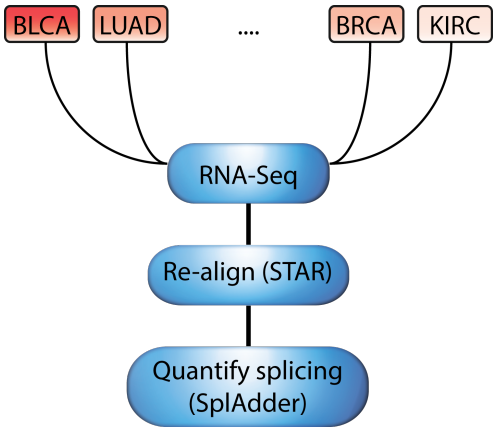
TCGA provides RNA-seq and matching exome data

- RNA-seq $\rightsquigarrow$ Find & quantify splicing events
- Exome $\rightsquigarrow$ Identify variants in exons & flanking intronic regions

Problem: Non-uniform processing (alignments & variant calling)

Re-analysis of Raw Sequencing Data





Computing at cluster colocated with CGHub

Scale: 9,000 exome & 4,500 RNA-seq libraries $\rightsquigarrow$ 400 TB data

$\Rightarrow$ Re-mapping (STAR): $\approx$ 6 CPU years

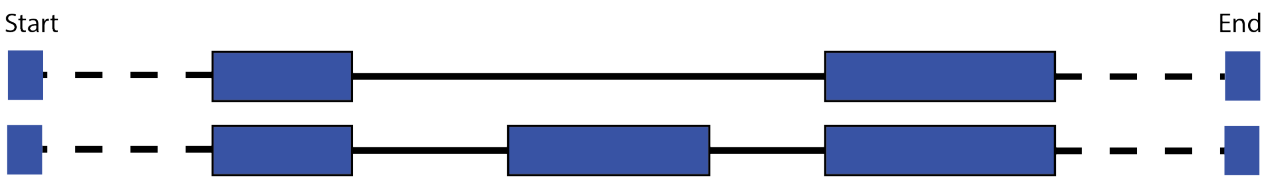
$\Rightarrow$ Variant Calling (GATK U.G. & MuTect): $\approx$ 12 CPU years

$\Rightarrow$ Splice variant quantification (SplAdder): $\approx$ 0.5 CPU years

Detecting Alternative Splice Events with SplAdder

Estimating Splice Index

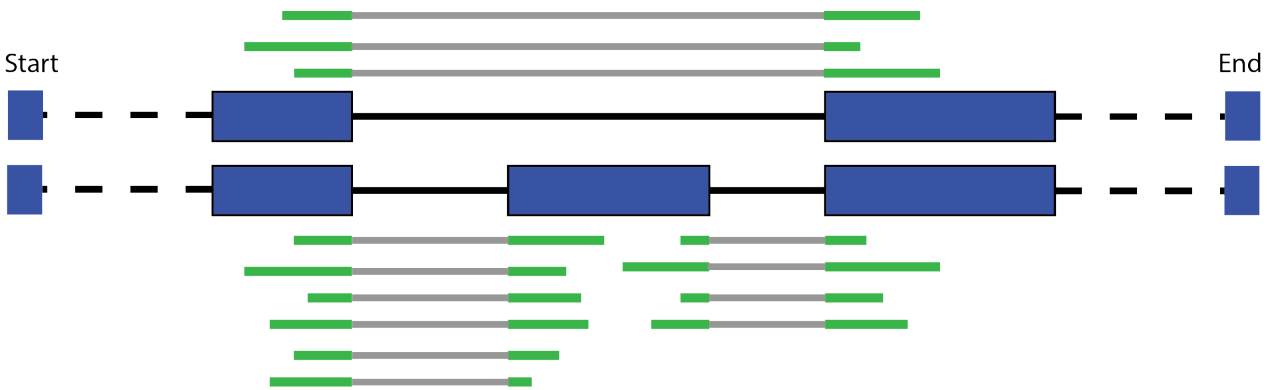
- Build and extend splice-graph using re-aligned reads
- Spliced reads support either Isoform 1 or Isoform 2
- Count reads supporting alternate event



Detecting Alternative Splice Events with SplAdder

Estimating Splice Index

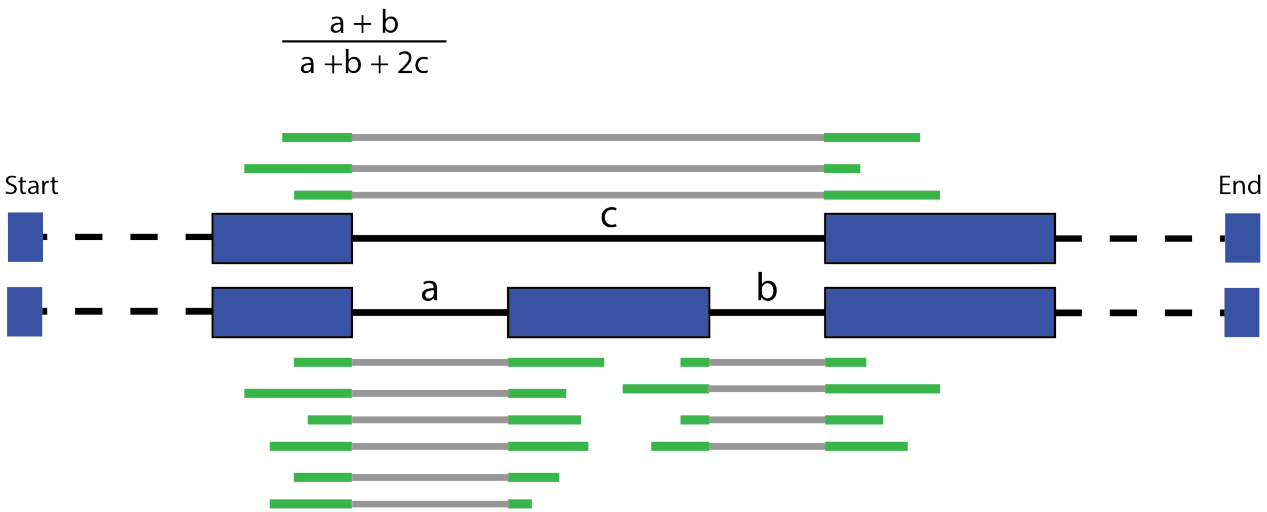
- Build and extend splice-graph using re-aligned reads
- Spliced reads support either Isoform 1 or Isoform 2
- Count reads supporting alternate event



Detecting Alternative Splice Events with SplAdder

Estimating Splice Index

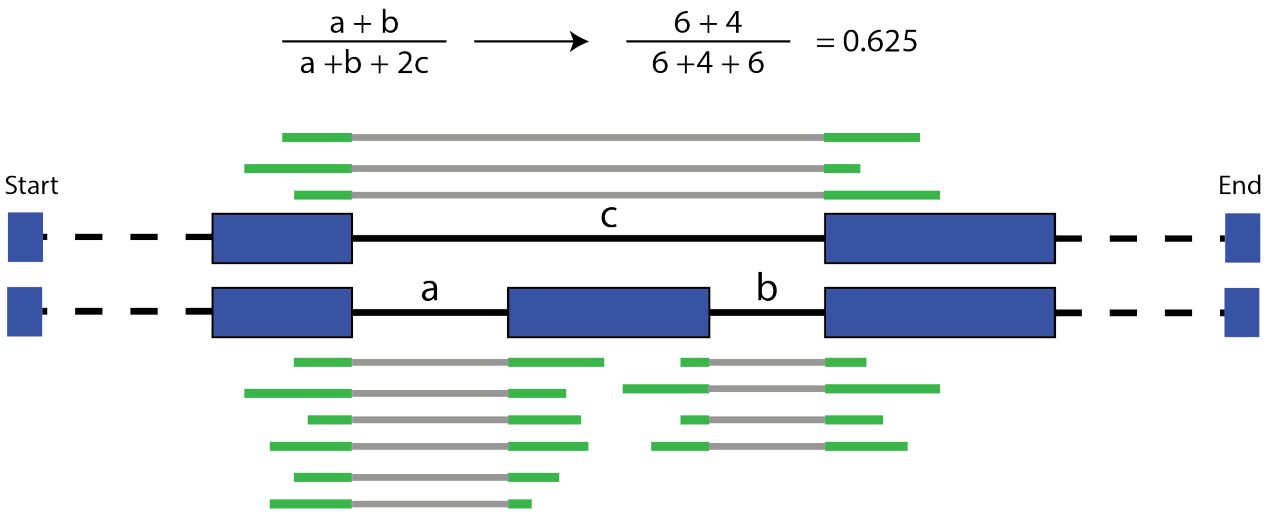
- Build and extend splice-graph using re-aligned reads
- Spliced reads support either Isoform 1 or Isoform 2
- Count reads supporting alternate event



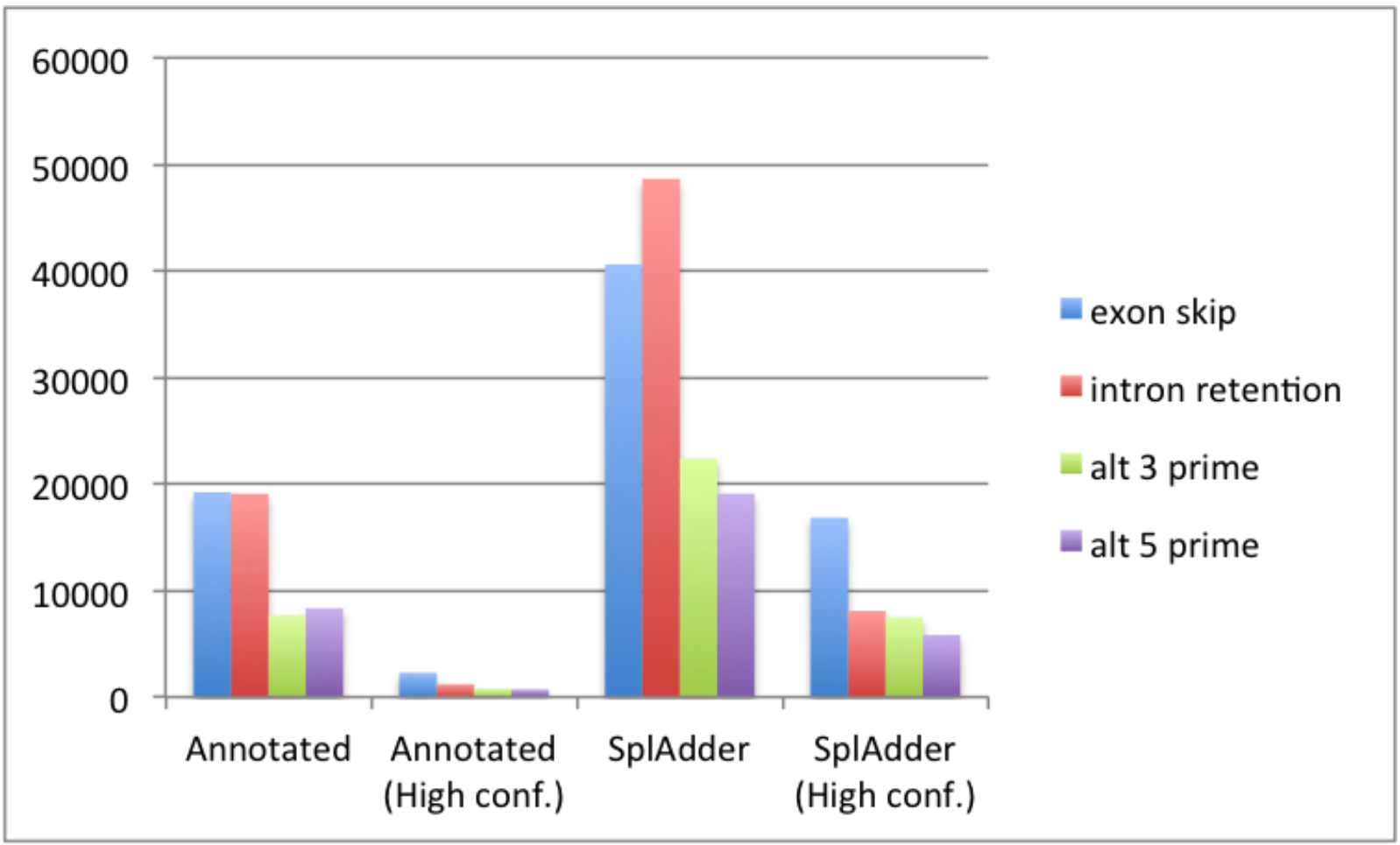
Detecting Alternative Splice Events with SplAdder

Estimating Splice Index

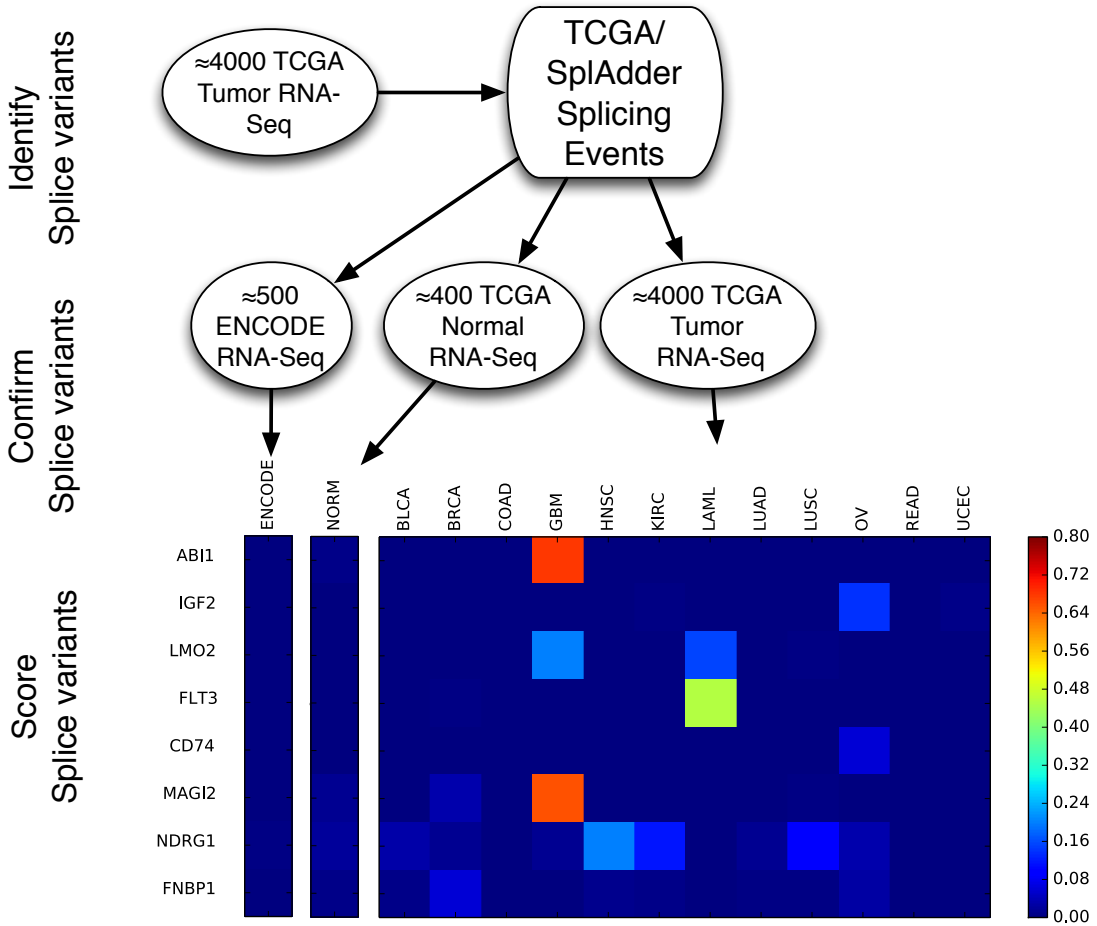
- Build and extend splice-graph using re-aligned reads
- Spliced reads support either Isoform 1 or Isoform 2
- Count reads supporting alternate event



Splicing Variation Across Cancer Types

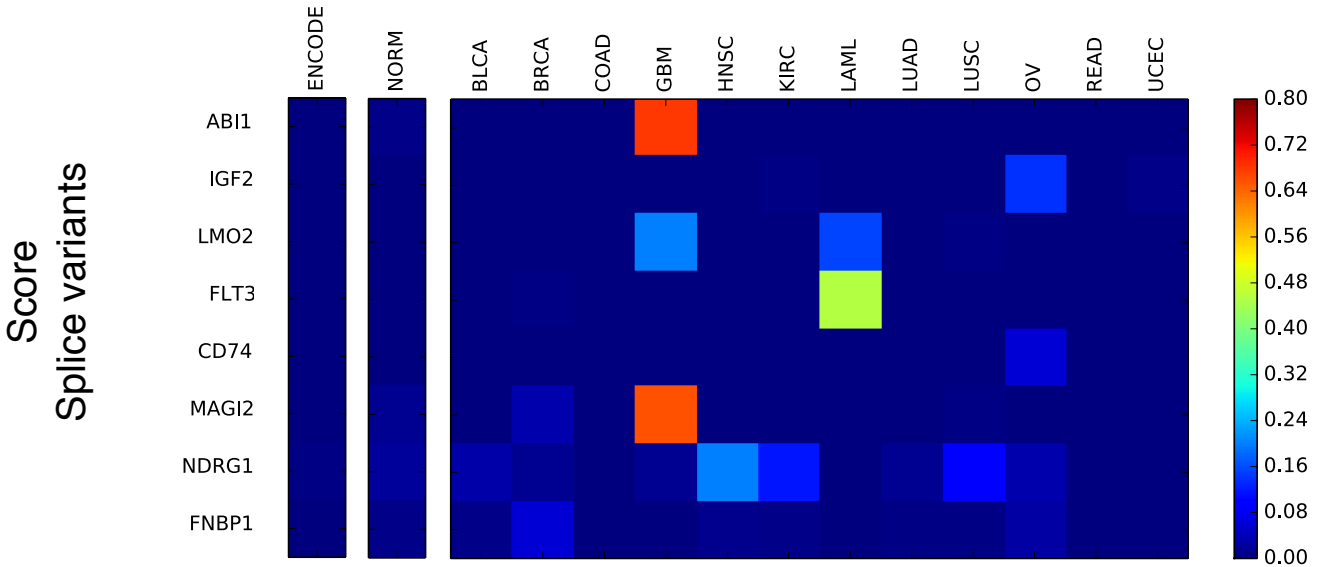


Detection of Cancer-Specific Splicing



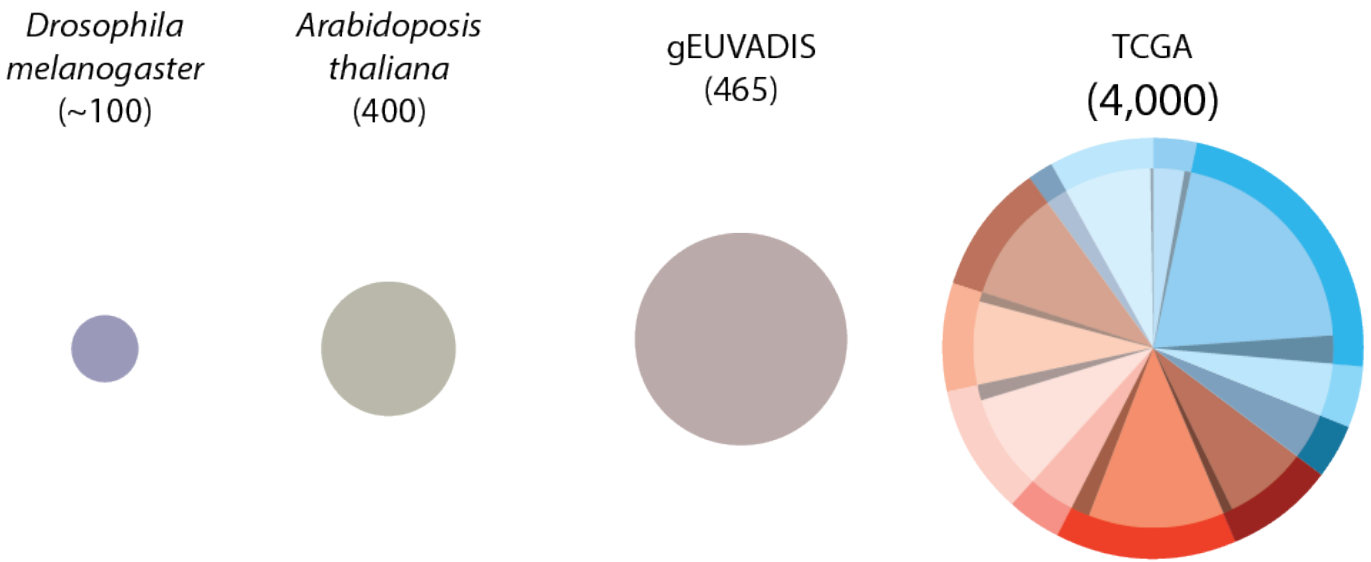
- Detected new splicing events that occur frequently in specific cancer types

Detection of Cancer-Specific Splicing



- Detected new splicing events that occur frequently in specific cancer types
- Needs independent confirmation
- Potential targets for treatment

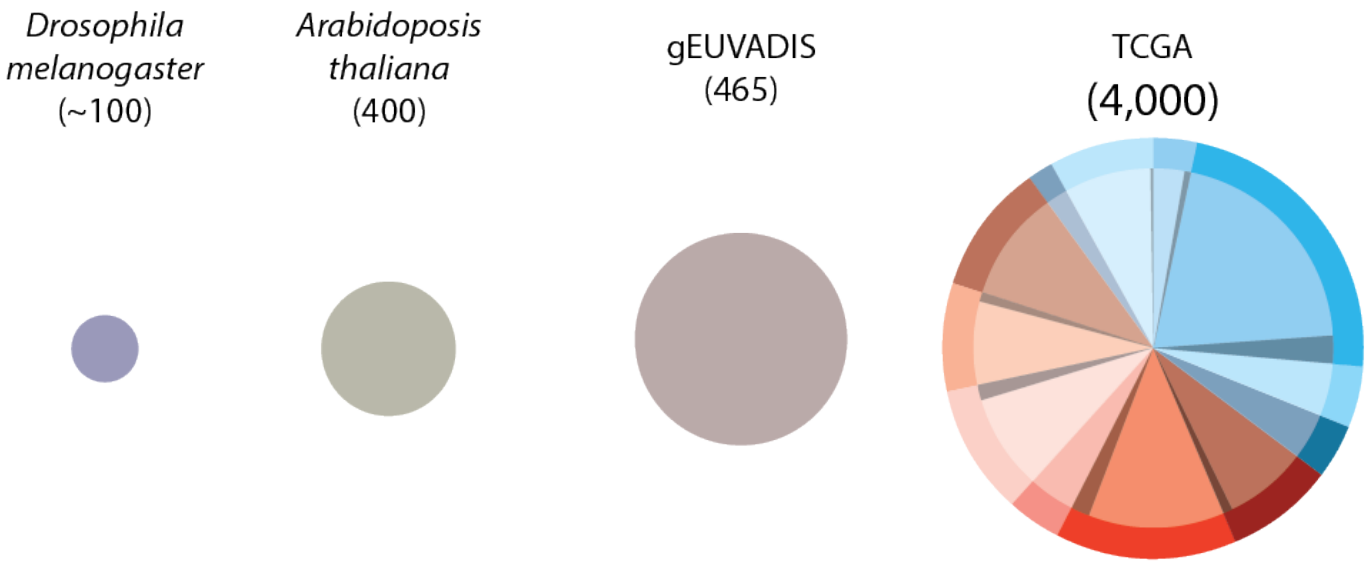
QTL Analyses in Comparison



Challenges in Cancer Genomics

- **Opportunity:** Understand tissue- & cancer specificity of splicing
- **Opportunity:** Large sample size allows to find *trans*-associations
- **Problem:** Heterogeneity and purity of sample
- **Problem:** Germline *vs.* Somatic mutations, many rare variants

QTL Analyses in Comparison



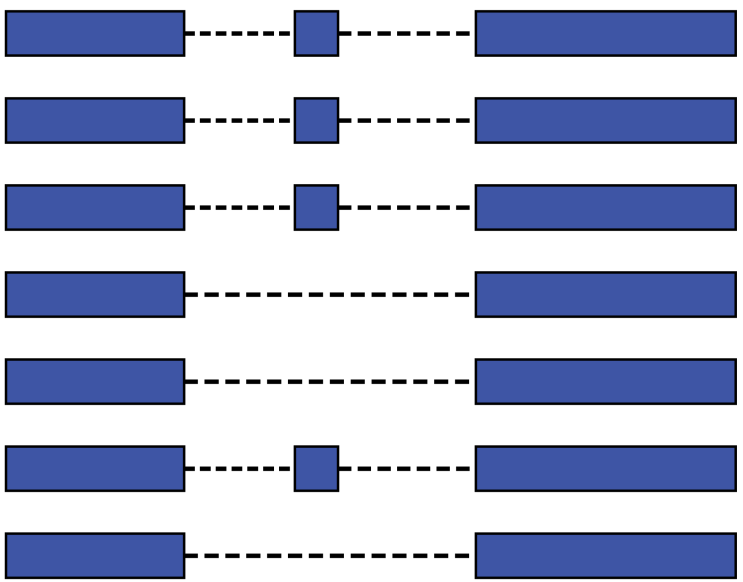
Challenges in Cancer Genomics

- **Opportunity:** Understand tissue- & cancer specificity of splicing
- **Opportunity:** Large sample size allows to find *trans*-associations
- **Problem:** Heterogeneity and purity of sample
- **Problem:** Germline *vs.* Somatic mutations, many rare variants

Common Variant Association Analysis

Modeling cancer and population structure

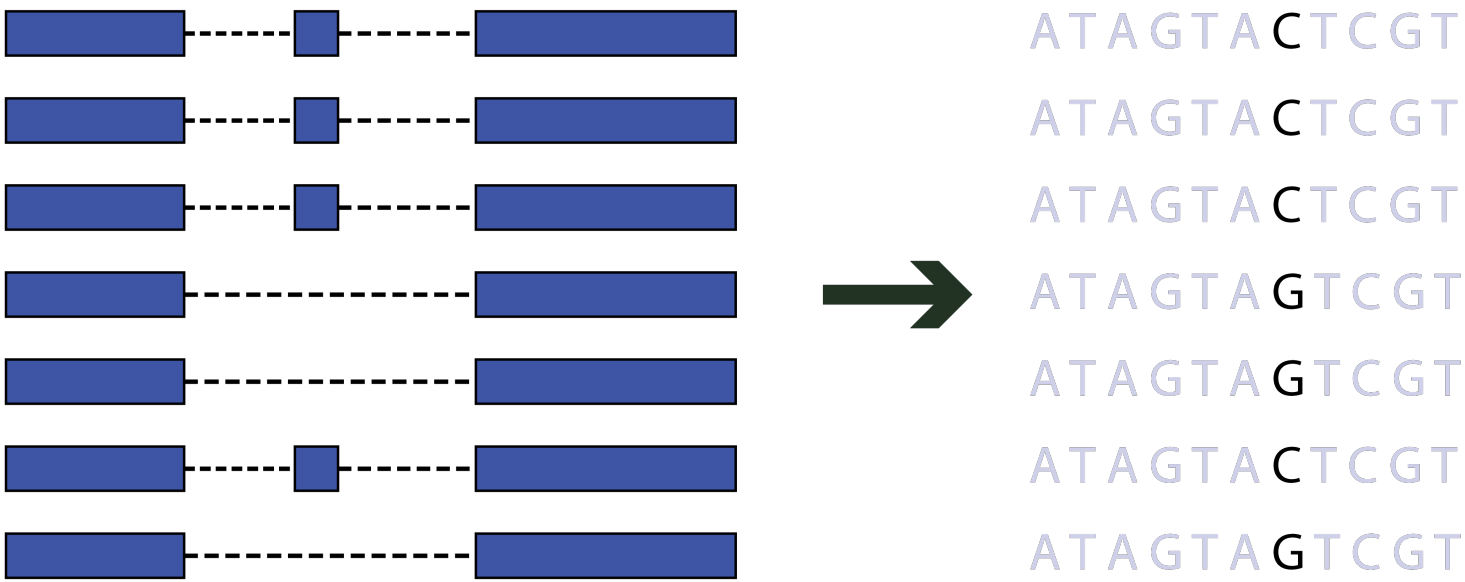
- $Y = X\beta + \text{Pop. Structure} + \text{Cancer Structure} + \epsilon$
- Pop. Structure $\sim N(0, \sigma_p^2 P)$ with $P = X_{\text{germ}} X_{\text{germ}}^T$
- Cancer Structure $\sim N(0, \sigma_c^2 C)$ with $C = X_{\text{soma}} X_{\text{soma}}^T$



Common Variant Association Analysis

Modeling cancer and population structure

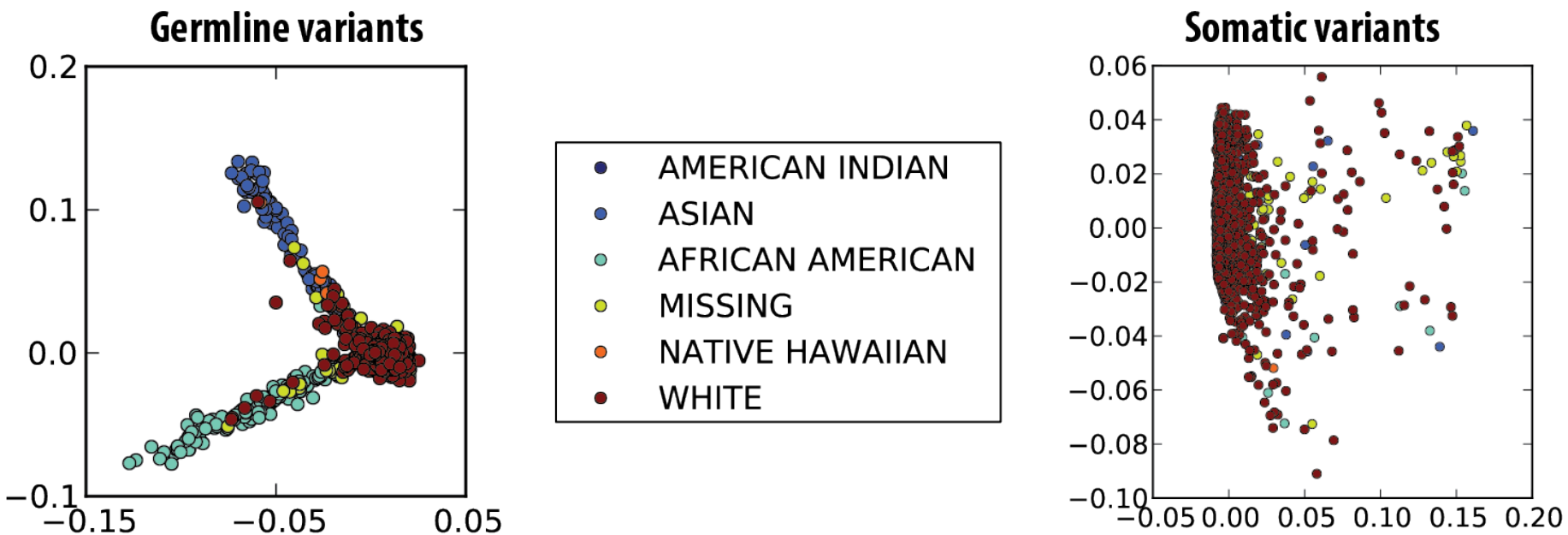
- $Y = X\beta$ + Pop. Structure + Cancer Structure $+\epsilon$
- Pop. Structure $\sim N(0, \sigma_p^2 P)$ with $P = X_{\text{germ}} X_{\text{germ}}^T$
- Cancer Structure $\sim N(0, \sigma_c^2 C)$ with $C = X_{\text{soma}} X_{\text{soma}}^T$



Common Variant Association Analysis

Modeling cancer and population structure

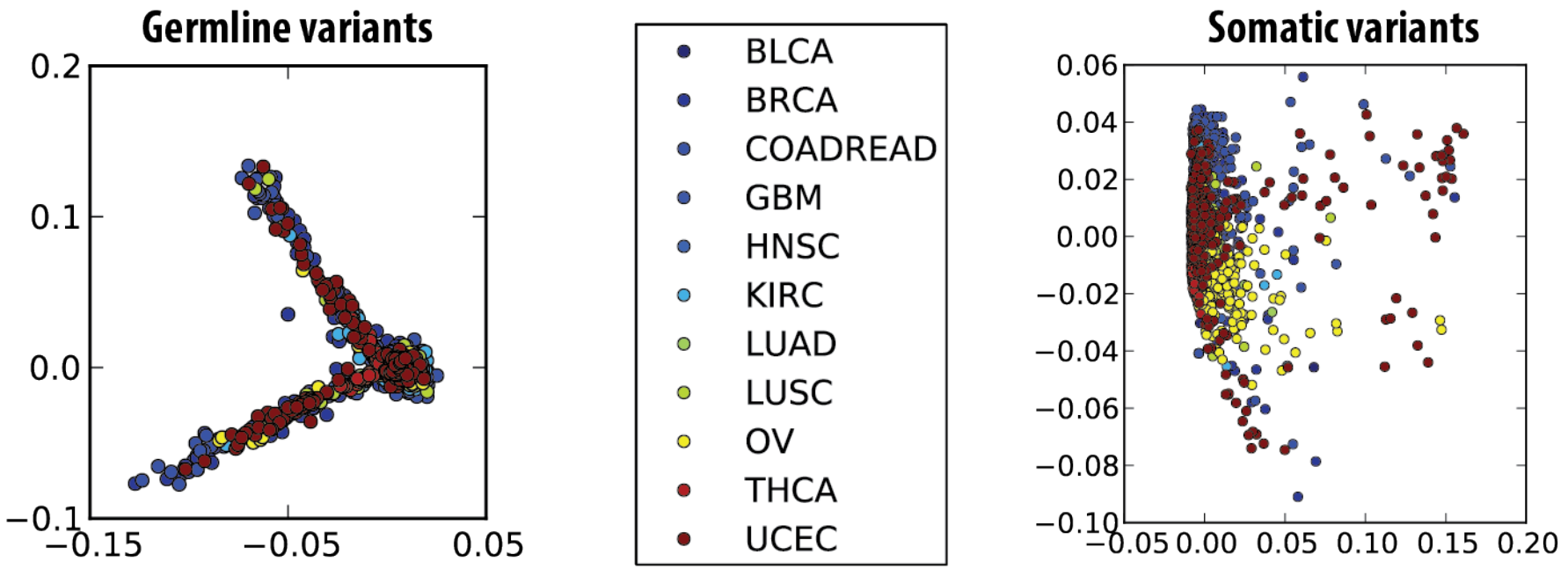
- $Y = X\beta +$ Pop. Structure $+$ Cancer Structure $+\epsilon$
- Pop. Structure $\sim N(0, \sigma_p^2 P)$ with $P = X_{\text{germ}} X_{\text{germ}}^T$
- Cancer Structure $\sim N(0, \sigma_c^2 C)$ with $C = X_{\text{soma}} X_{\text{soma}}^T$



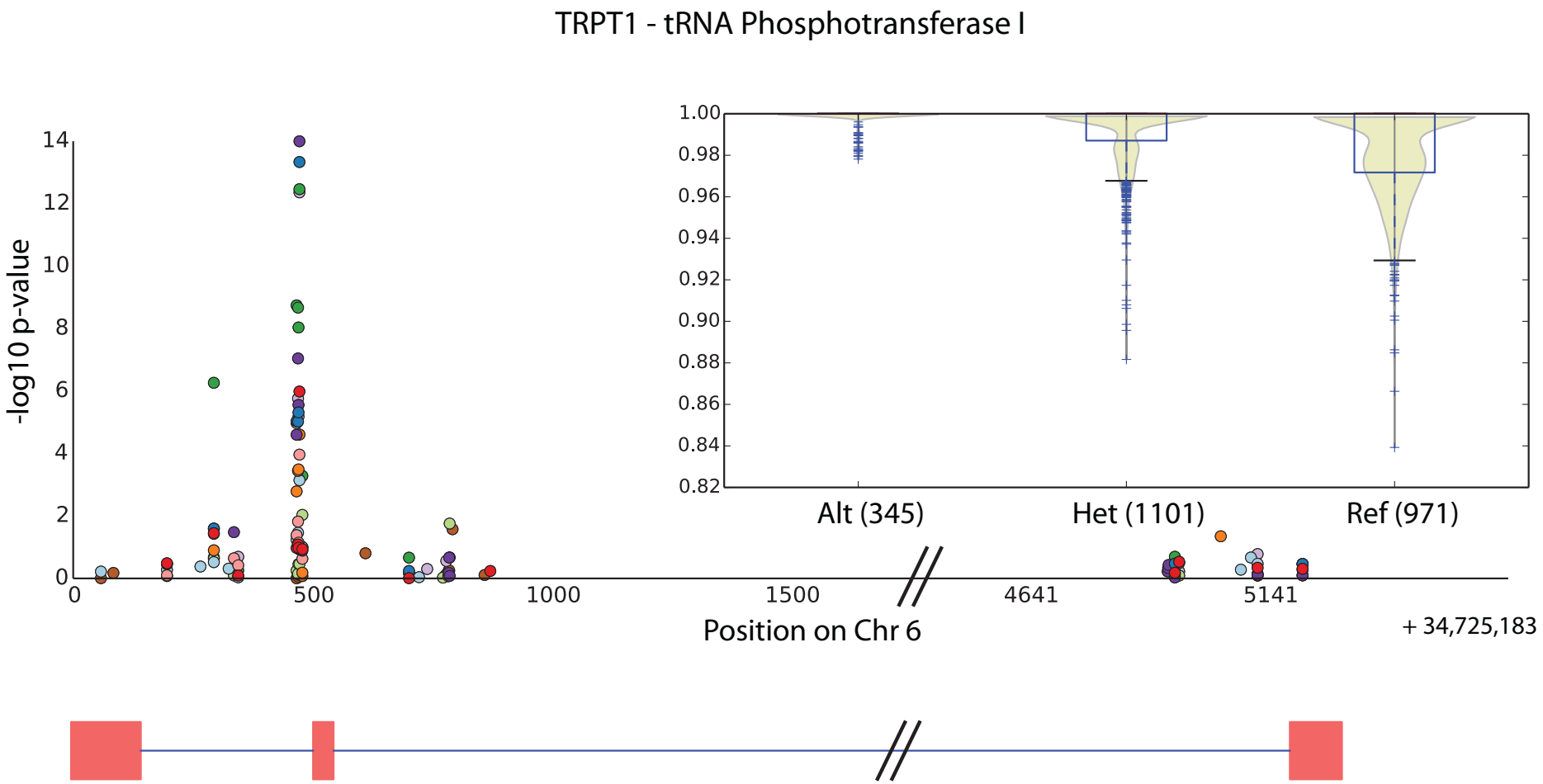
Common Variant Association Analysis

Modeling cancer and population structure

- $Y = X\beta + \text{Pop. Structure} + \text{Cancer Structure} + \epsilon$
- Pop. Structure $\sim N(0, \sigma_p^2 P)$ with $P = X_{\text{germ}} X_{\text{germ}}^T$
- Cancer Structure $\sim N(0, \sigma_c^2 C)$ with $C = X_{\text{soma}} X_{\text{soma}}^T$

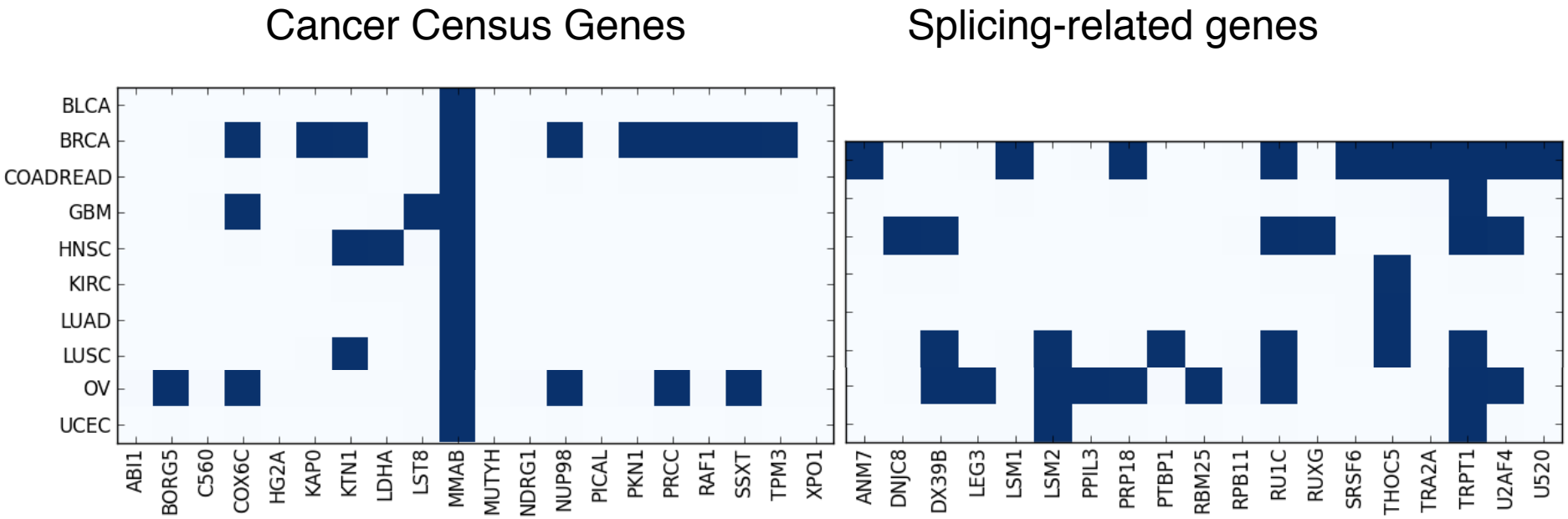


Example: cis-Associations in SNRP-C

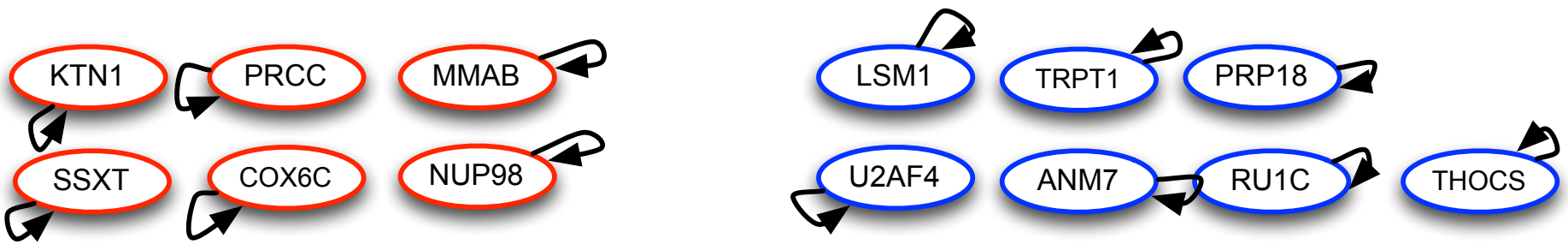


cis-Associations across Multiple Cancer Types

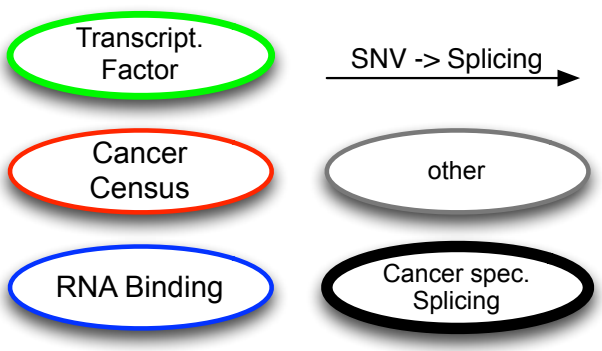
cis-Associations in 45 genes at 5% FDR (in 900 considered genes)

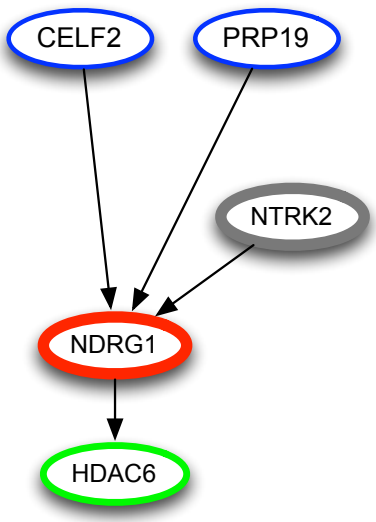


Replicated in multiple cancer types:

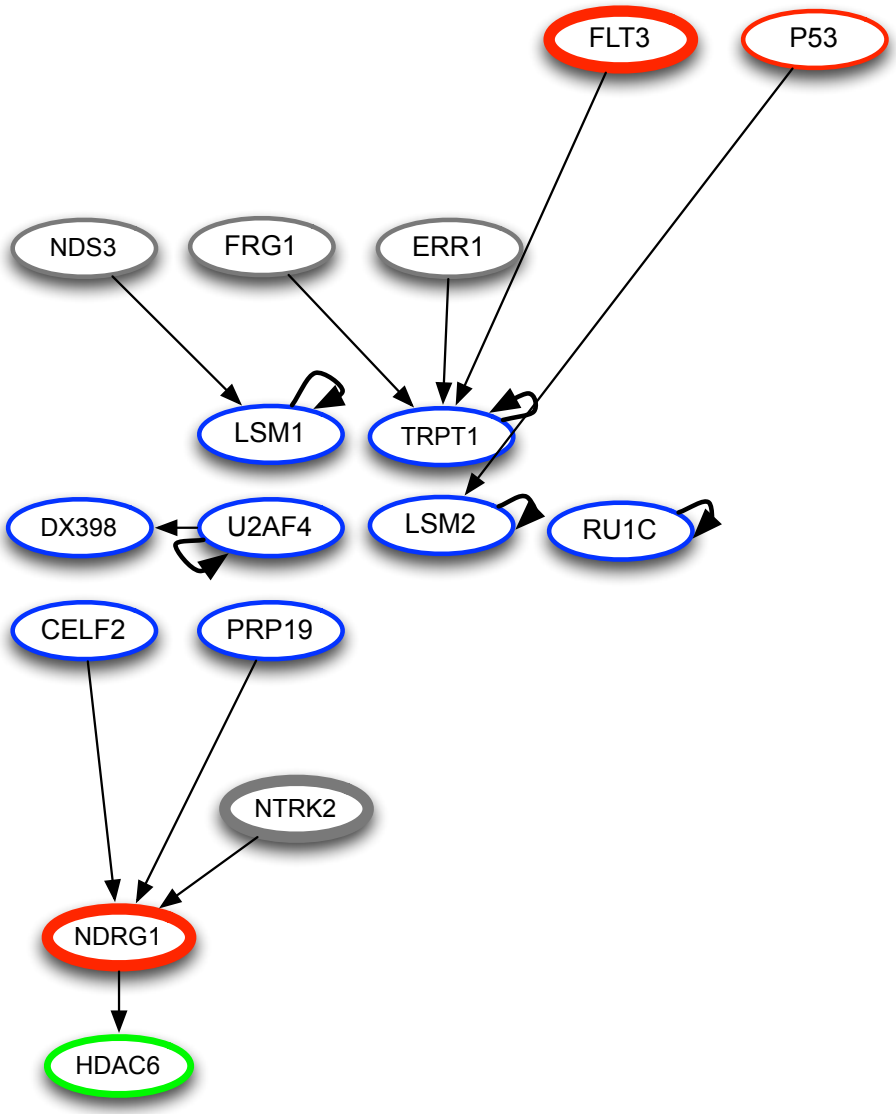


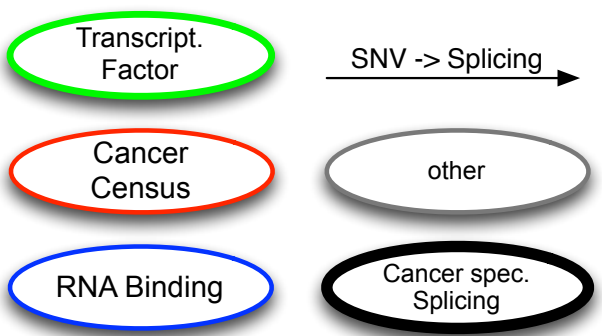
Splicing trans-Associations (FDR 5%)



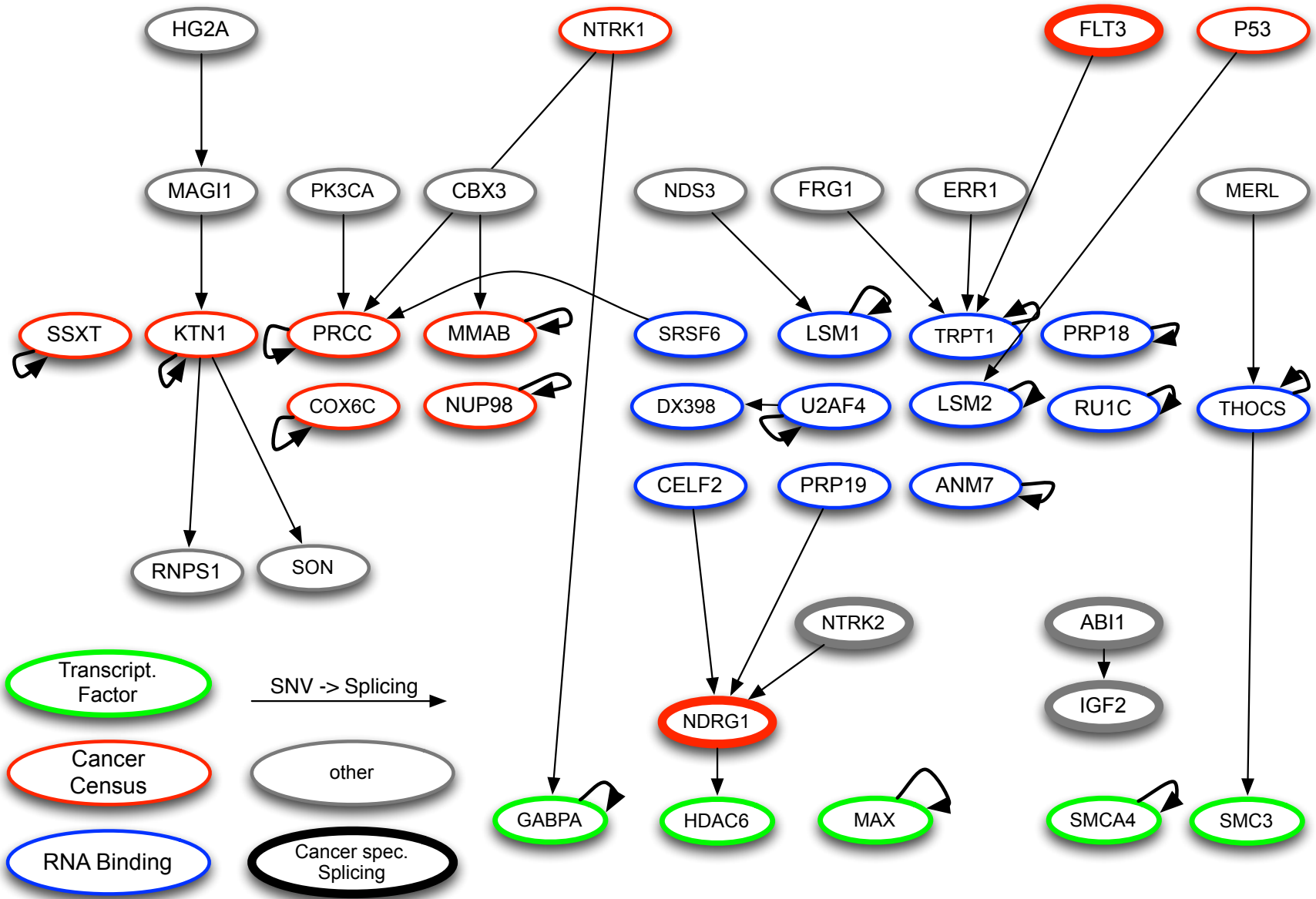


Splicing trans-Associations (FDR 5%)





Splicing trans-Associations (FDR 5%)



Conclusions and future work

- Developed resource of novel & known alternative splice events
- Identified cancer-specific isoforms that appear rarely expressed in normal samples
- Performed common variant association study to map splicing phenotypes
- Sample size in TCGA data enables detection of *trans* associations
- All of these associations still need validation (in particular *trans*)

Acknowledgements

- **André Kahles**
- Cyriac Kandoth
- William Lee
- Cancer Genome Atlas Network
- Nikolaus Schultz
- Robert Klein
- Oliver Stegle
- Gunnar Rätsch

Funded by MSKCC.