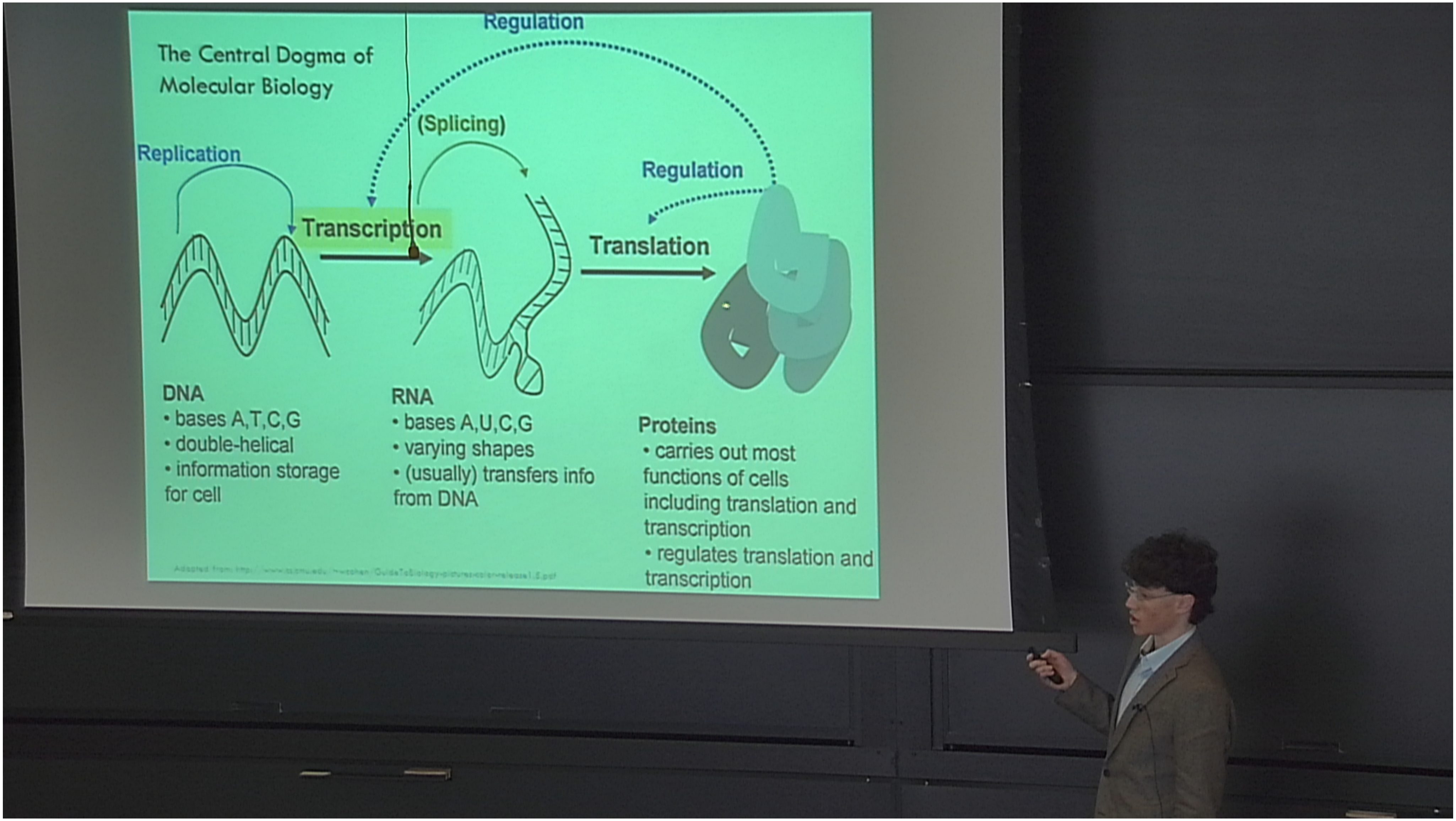
Title: Validation of predicted mRNA splicing mutations using high-throughput transcriptome data

Date: May 07, 2014 04:35 PM

URL: http://pirsa.org/14050059

Abstract: <span>This work has been published:Viner C Dorman SN Shirley BC and Rogan PK (2014)Validation of predicted mRNA splicing mutations using high-throughput transcriptome data [v1; ref status: indexedhttp://f1000r.es/2no]F1000Research20143:8 (doi:10.12688/f1000research.3-8.v1)Additionally this work has been accepted for a highlights presentation at the upcoming Great Lakes Bioinformatics Conference (GLBIO) in Cincinnati Ohio and it was recently presented as a poster at London Health Research Day (LHRD).Abstract:Interpretation of variants present in complete genomes or exomes reveals numerous sequence changes only a fraction of which are likely to be pathogenic. Variants predicted to alter mRNA splicing in particular can be validated by manual inspection of transcriptome sequencing data however this approach is intractable for large datasets. We show that abnormal mRNA splicing patterns are characterized by reads demonstrating either exon skipping cryptic splice site use and high levels of intron inclusion or combinations of these properties. This paper presents Veridical an in silico method for the automatic validation of DNA sequencing variants that alter mRNA splicing. Veridical leverages large numbers of control samples (that lack a putative mutation) applying z-tests to Yeo-Johnson transformed data to normalize read counts of abnormal RNA species in mutant versus non-mutant tissues. With the transformed data the null hypothesis based mainly on either counts of intronic or junctional reads can be rejected for true splicing mutations using conventional parametric statistical methods. </span>



The Central Dogma of Molecular Biology

Regulation

(Splicing)

Replication

Regulation

Transcription

Translation

DNA

- bases A,T,C,G
- double-helical
- information storage for cell

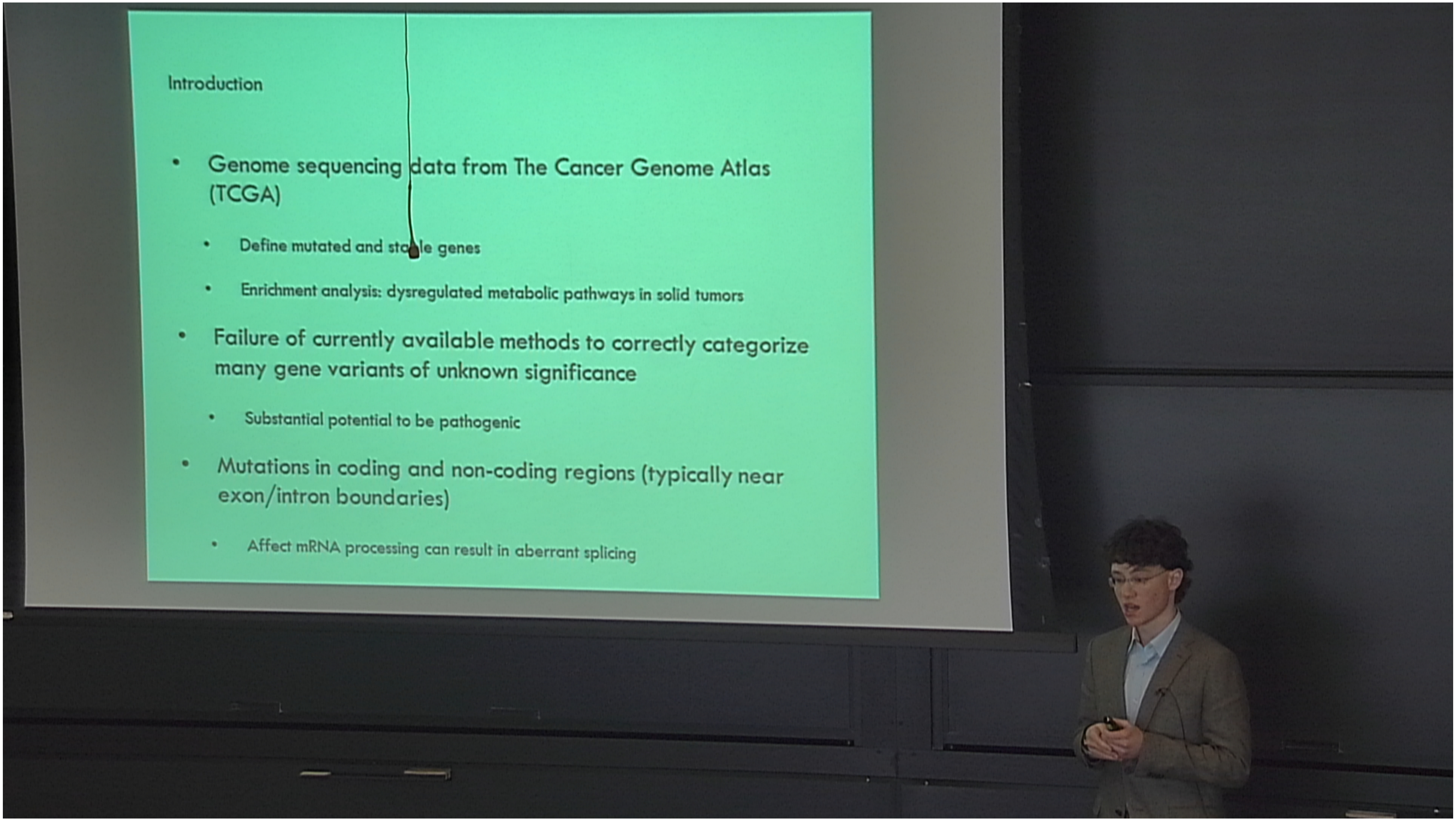
RNA

- bases A,U,C,G
- varying shapes
- (usually) transfers info from DNA

Proteins

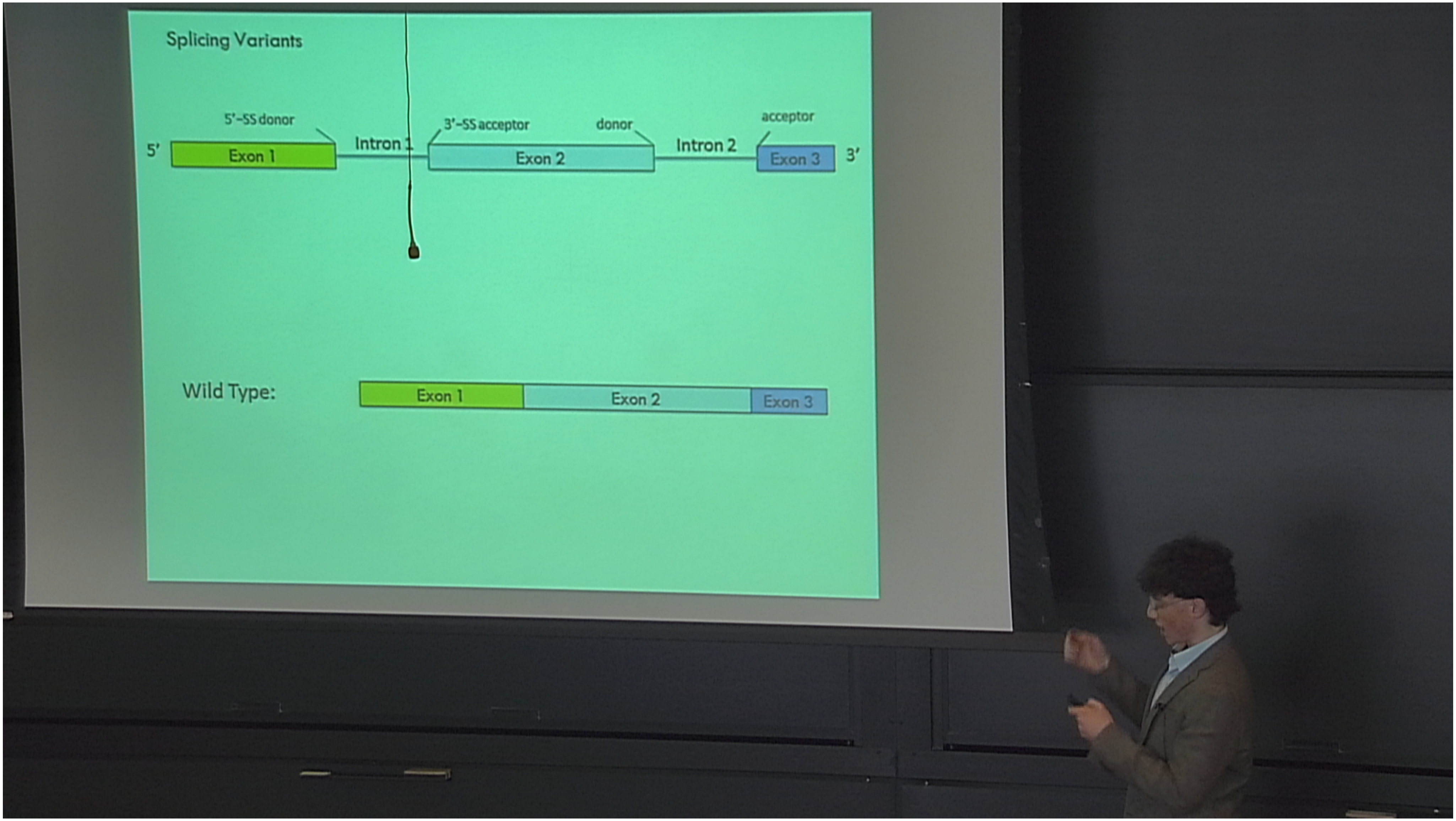
- carries out most functions of cells including translation and transcription
- regulates translation and transcription

Adapted from: http://www.cs.cmu.edu/~wcohen/GuideToBiology-pictures-color-release1.5.pdf



Introduction

- Genome sequencing data from The Cancer Genome Atlas (TCGA)
  - Define mutated and stable genes
  - Enrichment analysis: dysregulated metabolic pathways in solid tumors
- Failure of currently available methods to correctly categorize many gene variants of unknown significance
  - Substantial potential to be pathogenic
- Mutations in coding and non-coding regions (typically near exon/intron boundaries)
  - Affect mRNA processing can result in aberrant splicing



Splicing Variants

5' Exon 1 Intron 1 Exon 2 Intron 2 Exon 3 3'

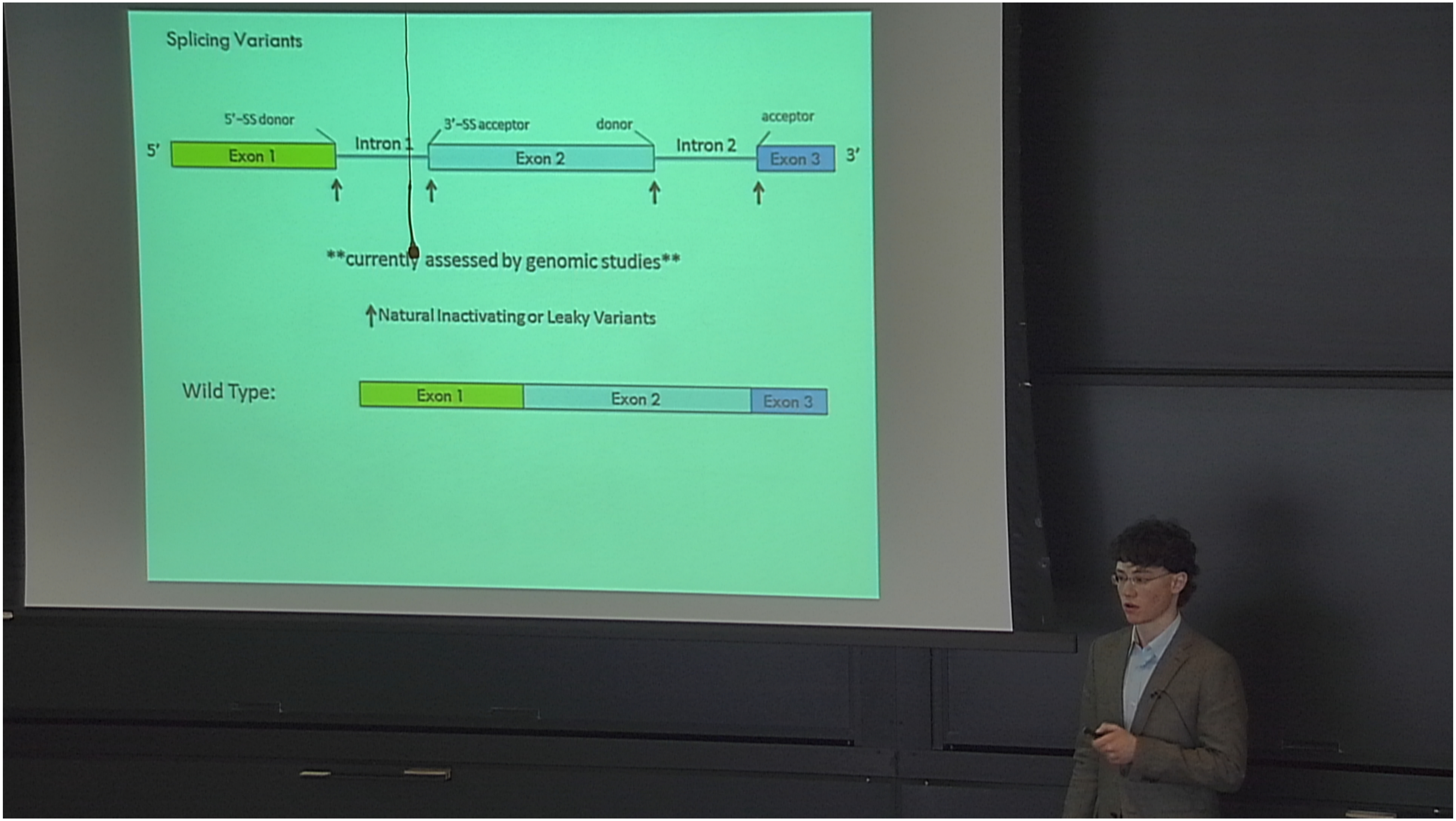
5'–SS donor

3'–SS acceptor

donor

acceptor

Wild Type: Exon 1 Exon 2 Exon 3



Splicing Variants

5'–SS donor

3'–SS acceptor

donor

acceptor

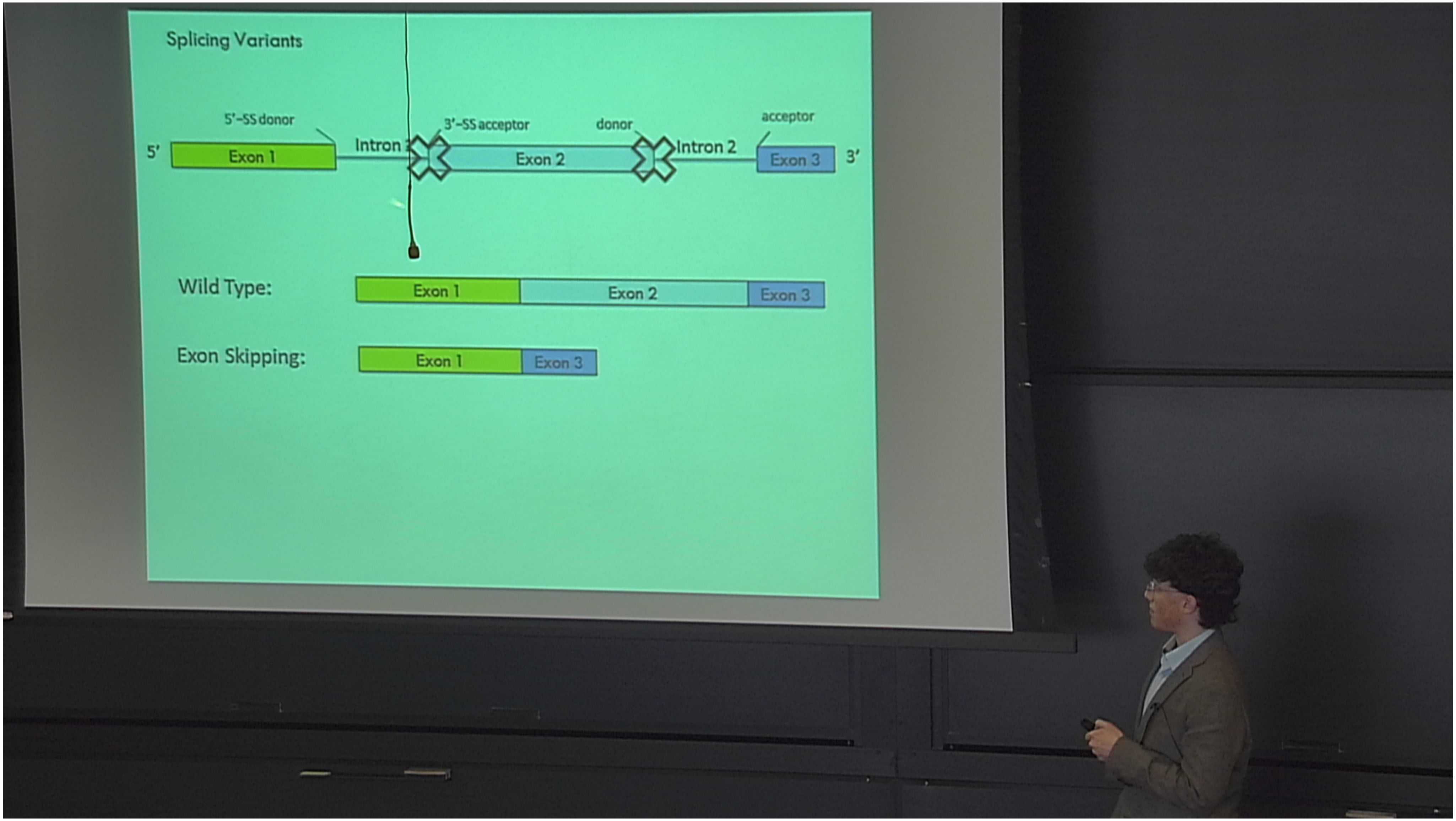
5' Exon 1 Intron 1 Exon 2 Intron 2 Exon 3 3'

↑ ↑ ↑ ↑

currently assessed by genomic studies

↑Natural Inactivating or Leaky Variants

Wild Type: Exon 1 Exon 2 Exon 3

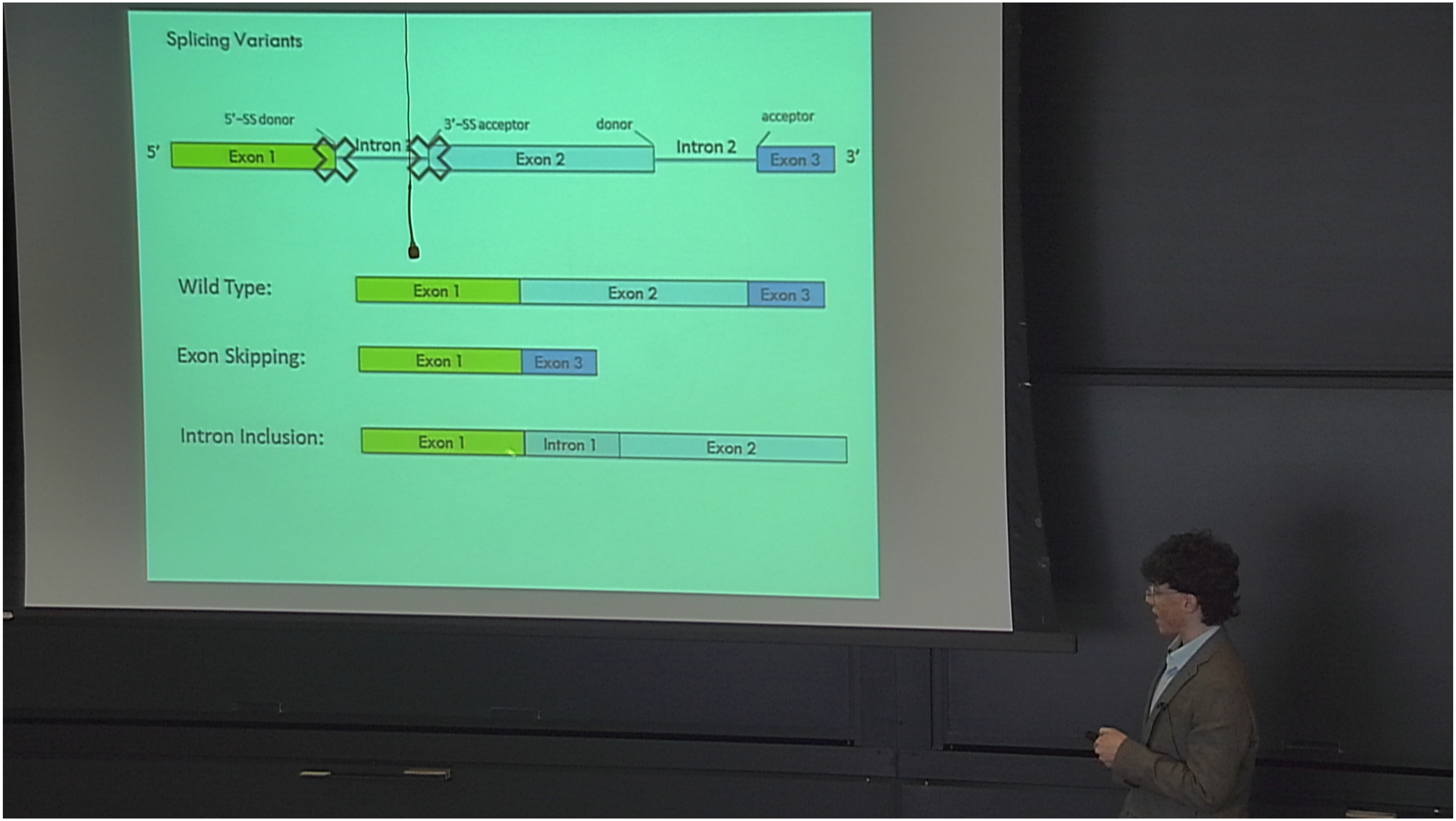


Splicing Variants

5' Exon 1 — 5'–SS donor — Intron 1 — 3'–SS acceptor — Exon 2 — donor — Intron 2 — acceptor — Exon 3 3'

Wild Type: Exon 1 | Exon 2 | Exon 3

Exon Skipping: Exon 1 | Exon 3



Splicing Variants

5' Exon 1 Intron 1 Exon 2 Intron 2 Exon 3 3'

5'–SS donor

3'–SS acceptor

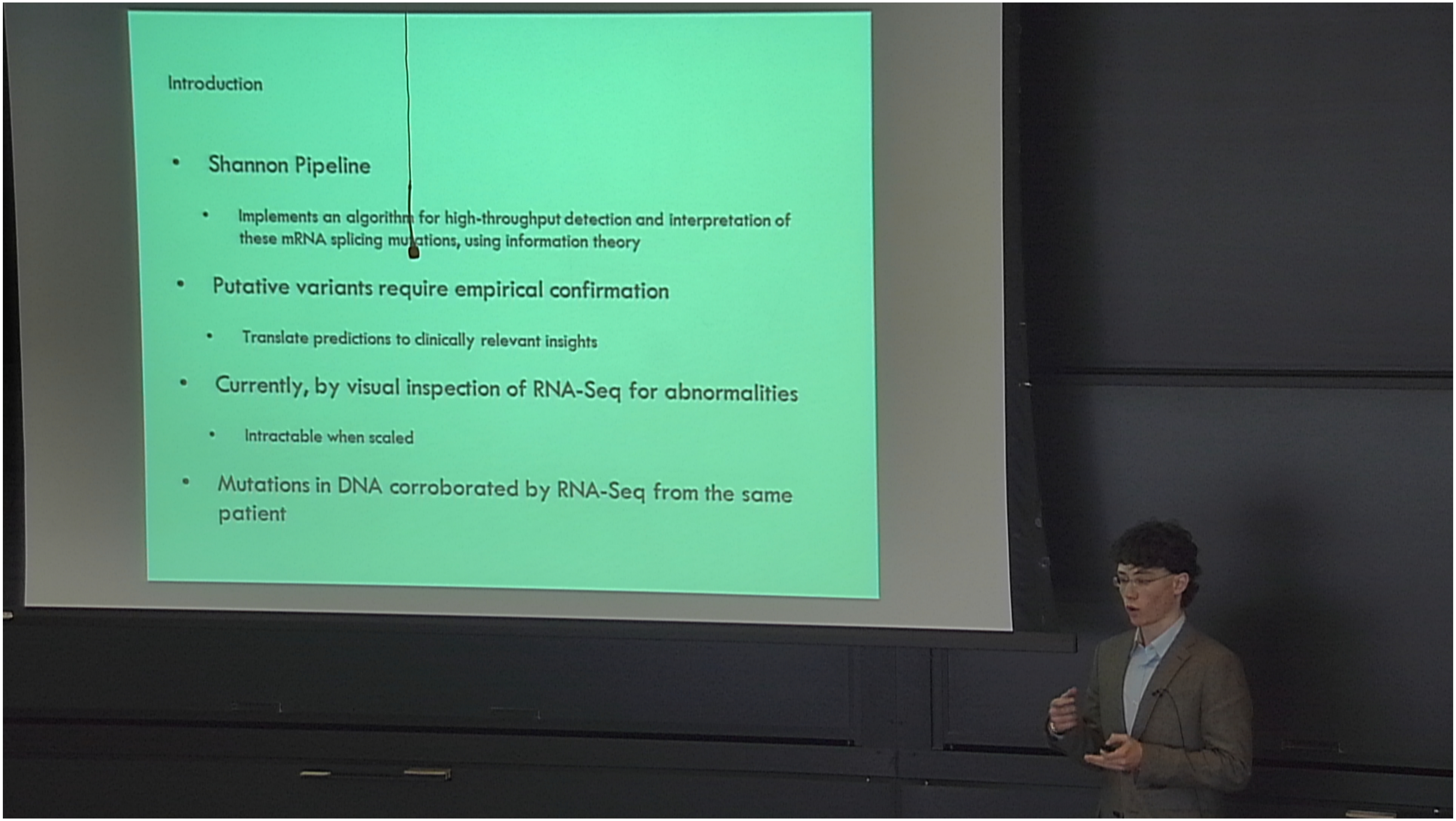
donor

acceptor

Wild Type: Exon 1 | Exon 2 | Exon 3

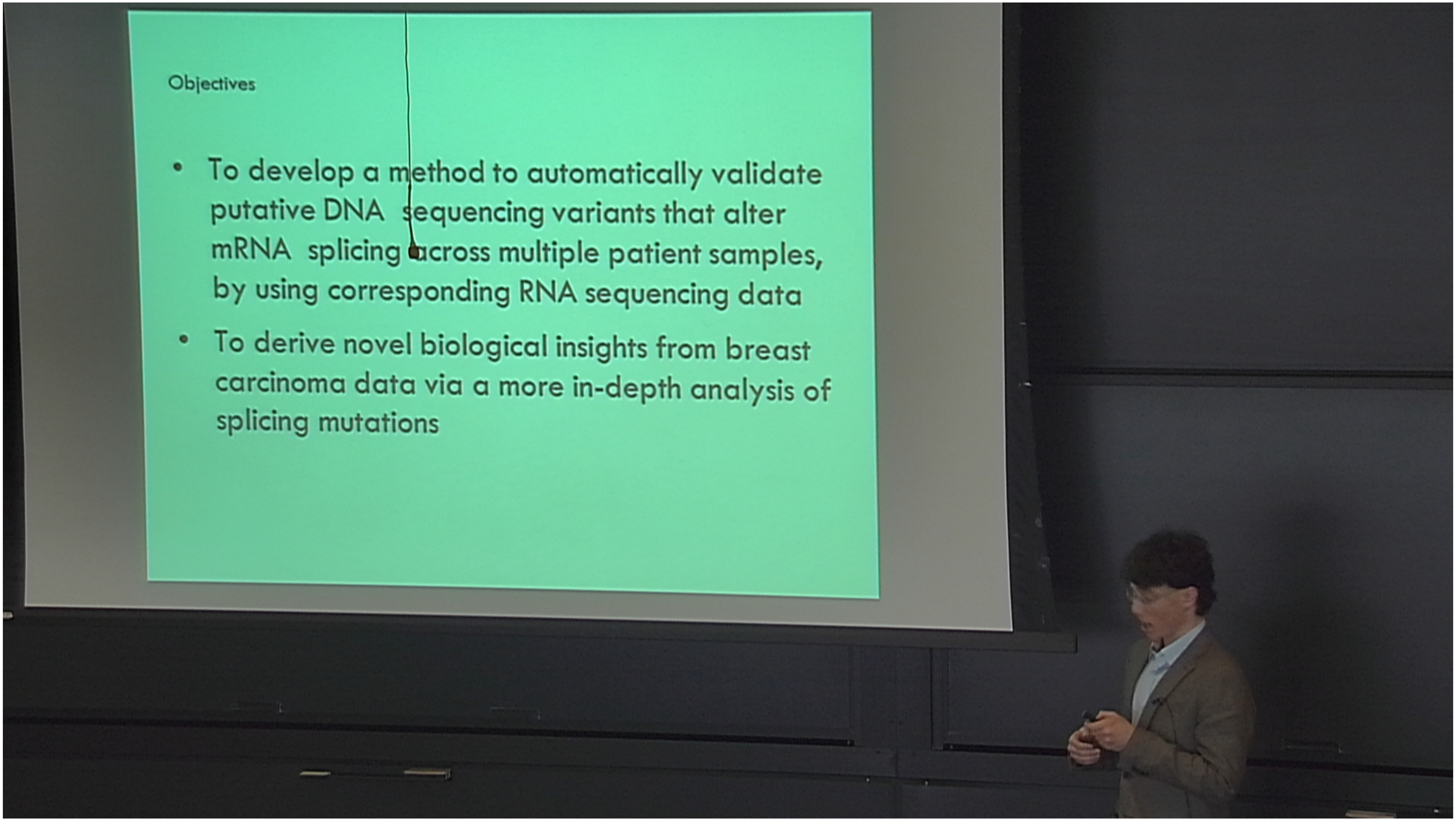
Exon Skipping: Exon 1 | Exon 3

Intron Inclusion: Exon 1 | Intron 1 | Exon 2



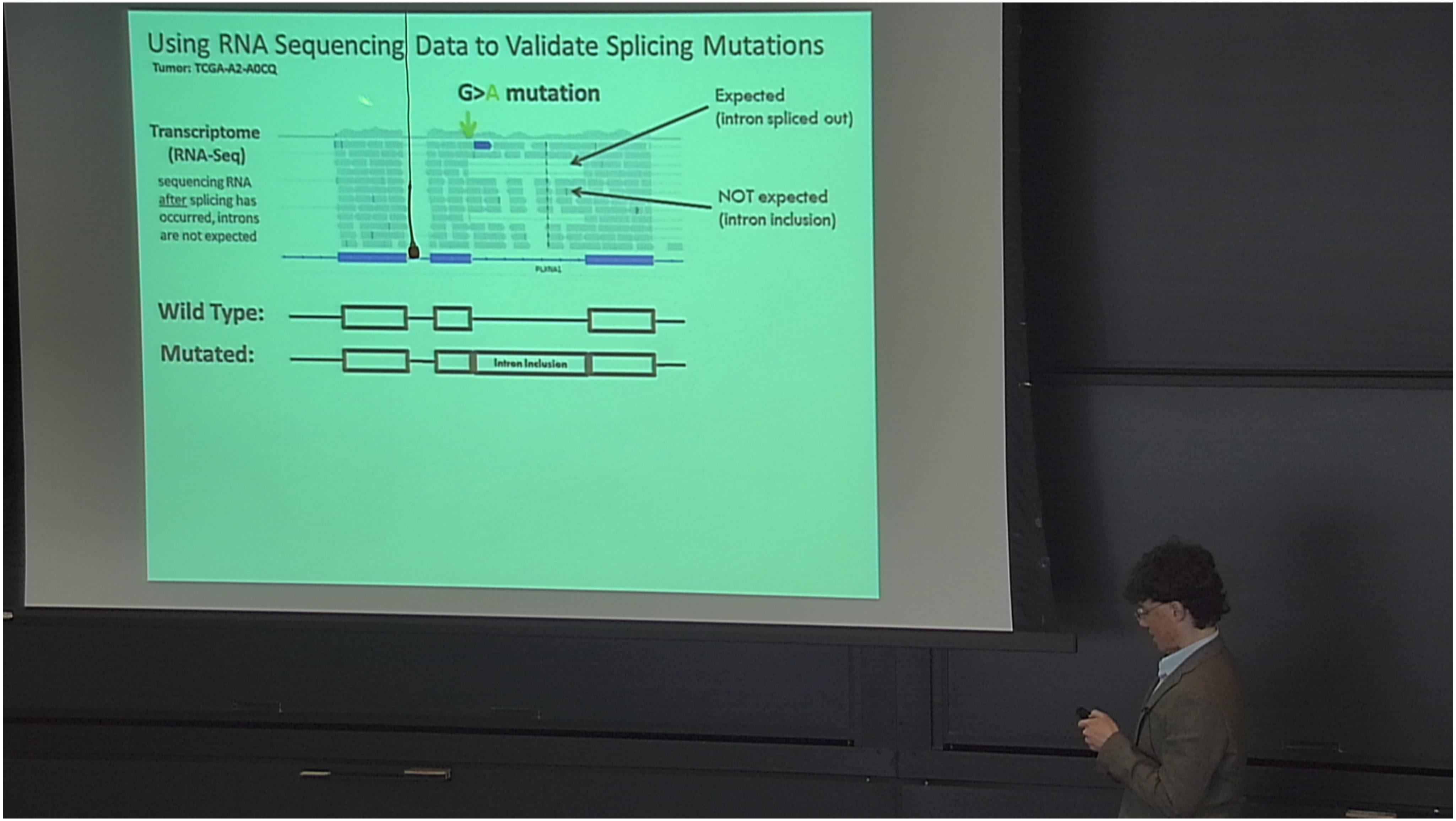
Introduction

- Shannon Pipeline
  - Implements an algorithm for high-throughput detection and interpretation of these mRNA splicing mutations, using information theory
- Putative variants require empirical confirmation
  - Translate predictions to clinically relevant insights
- Currently, by visual inspection of RNA-Seq for abnormalities
  - Intractable when scaled
- Mutations in DNA corroborated by RNA-Seq from the same patient



Objectives

- To develop a method to automatically validate putative DNA sequencing variants that alter mRNA splicing across multiple patient samples, by using corresponding RNA sequencing data
- To derive novel biological insights from breast carcinoma data via a more in-depth analysis of splicing mutations



Using RNA Sequencing Data to Validate Splicing Mutations

Tumor: TCGA-A2-A0CQ

G>A mutation

Transcriptome (RNA-Seq)

sequencing RNA after splicing has occurred, introns are not expected

Expected (intron spliced out)

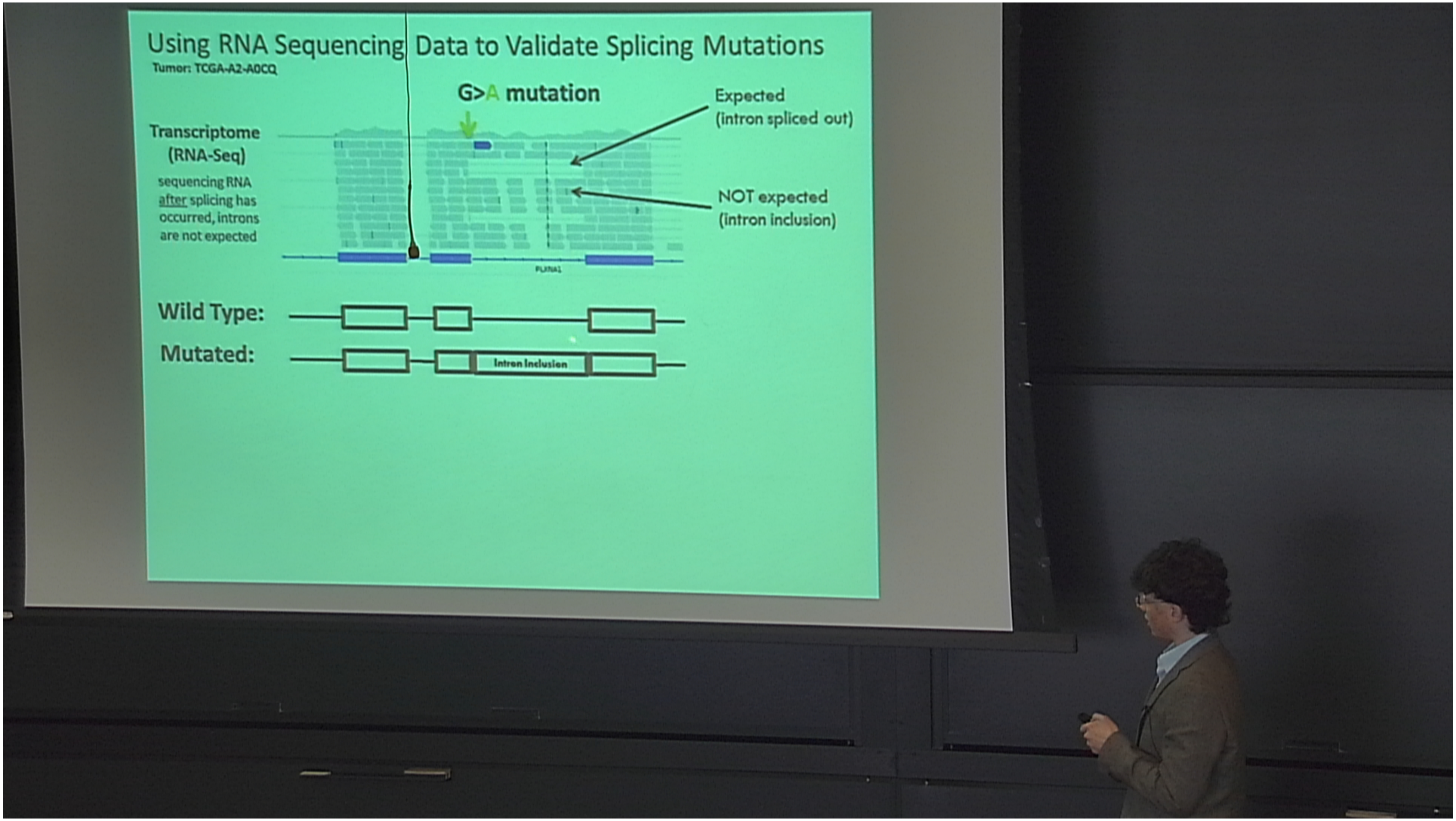
NOT expected (intron inclusion)

PLXNA1

Wild Type:

Mutated:

Intron Inclusion



Using RNA Sequencing Data to Validate Splicing Mutations

Tumor: TCGA-A2-A0CQ

G>A mutation

Transcriptome (RNA-Seq)

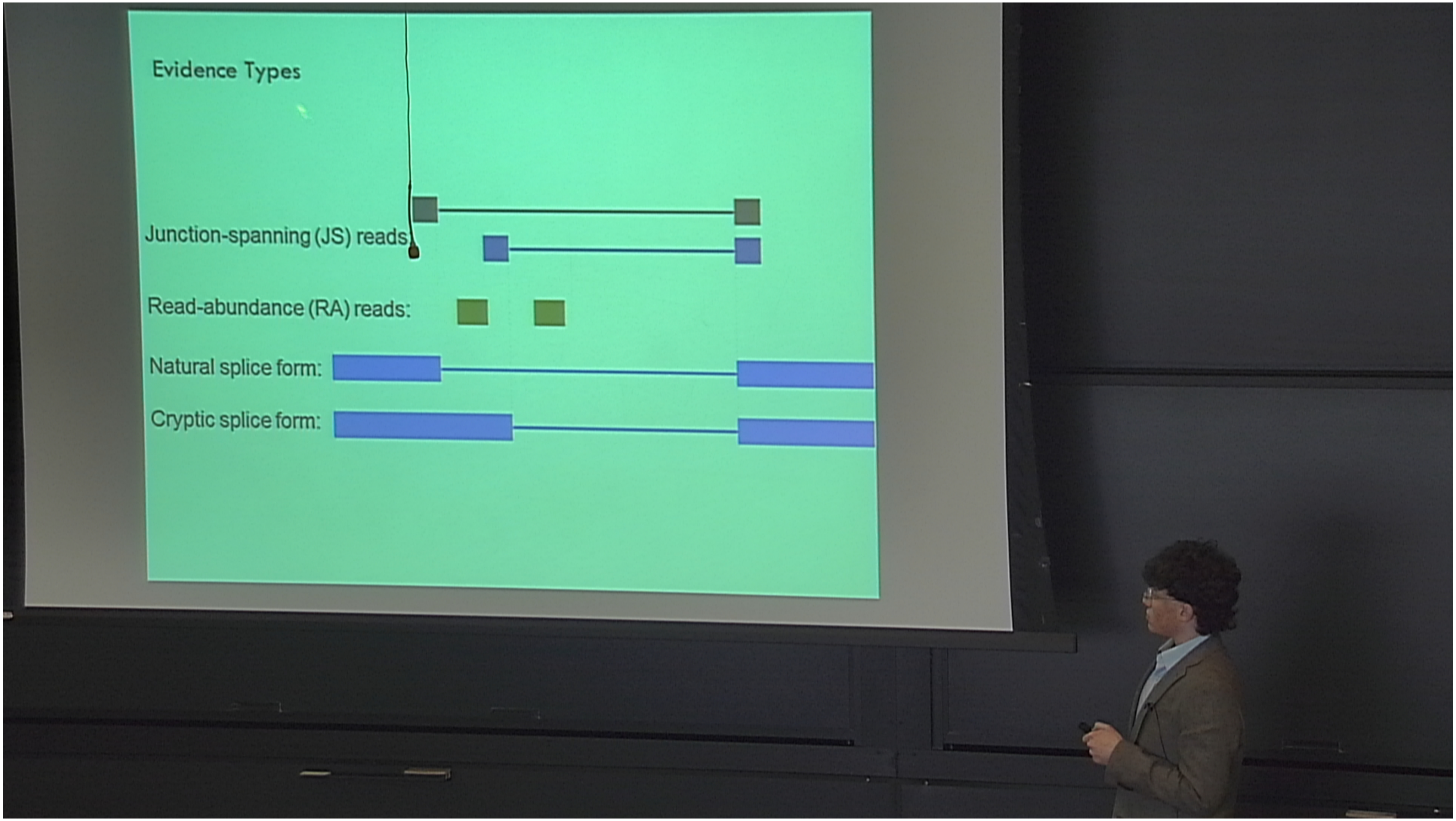
sequencing RNA after splicing has occurred, introns are not expected

Expected (intron spliced out)

NOT expected (intron inclusion)

Wild Type:

Mutated:



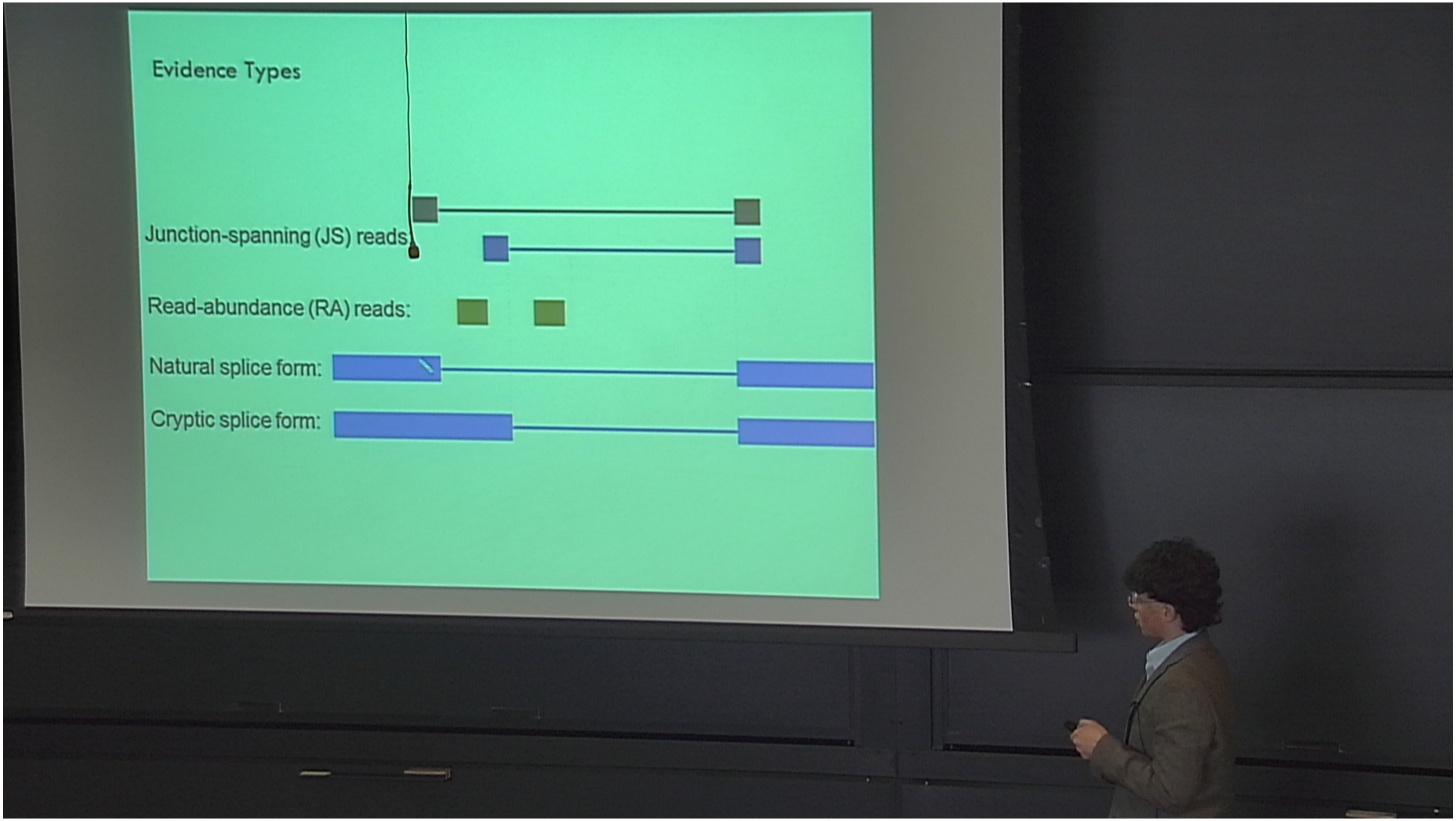
Evidence Types

Junction-spanning (JS) reads:

Read-abundance (RA) reads:

Natural splice form:

Cryptic splice form:



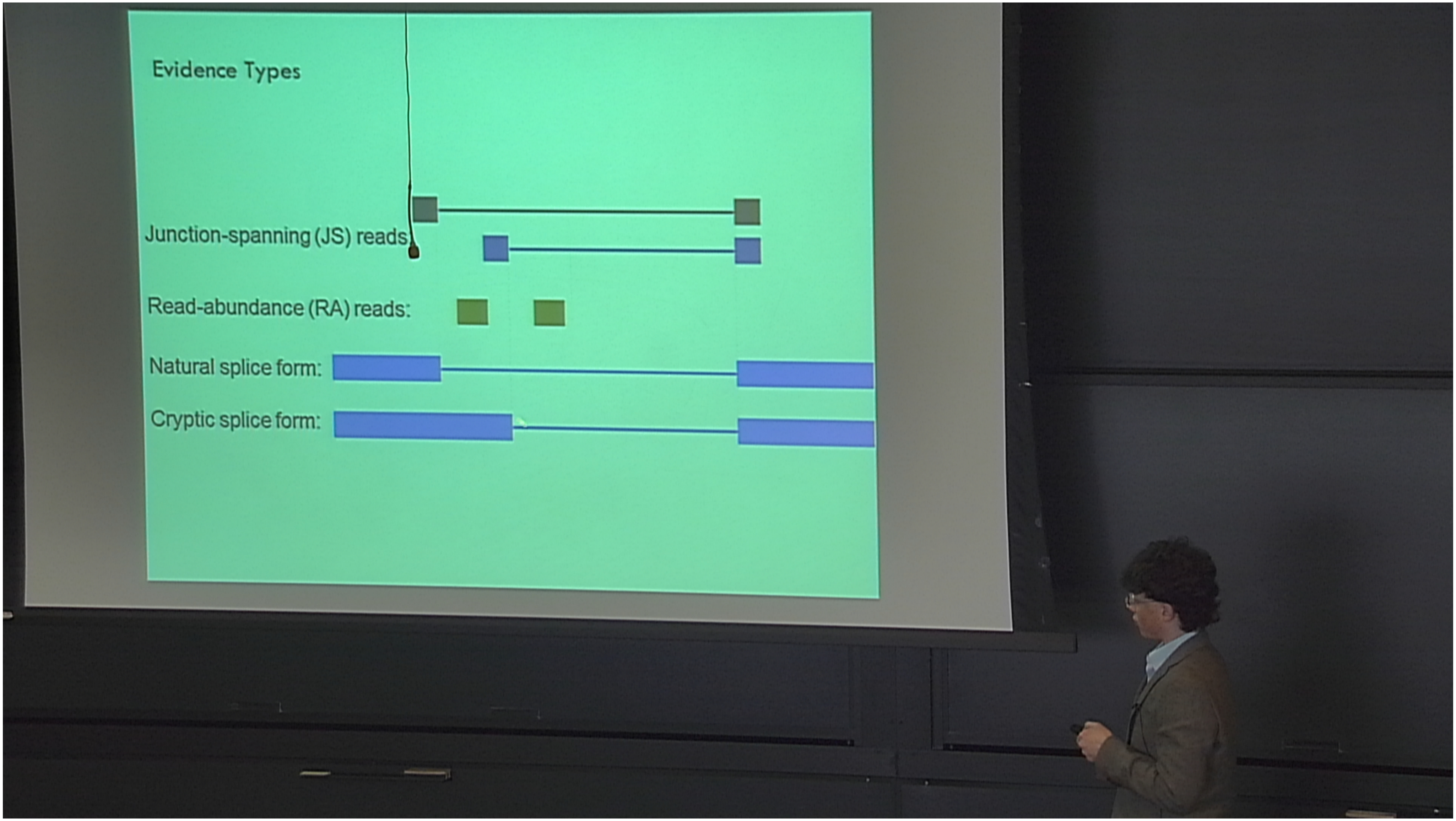
Evidence Types

Junction-spanning (JS) reads:

Read-abundance (RA) reads:

Natural splice form:

Cryptic splice form:



Evidence Types

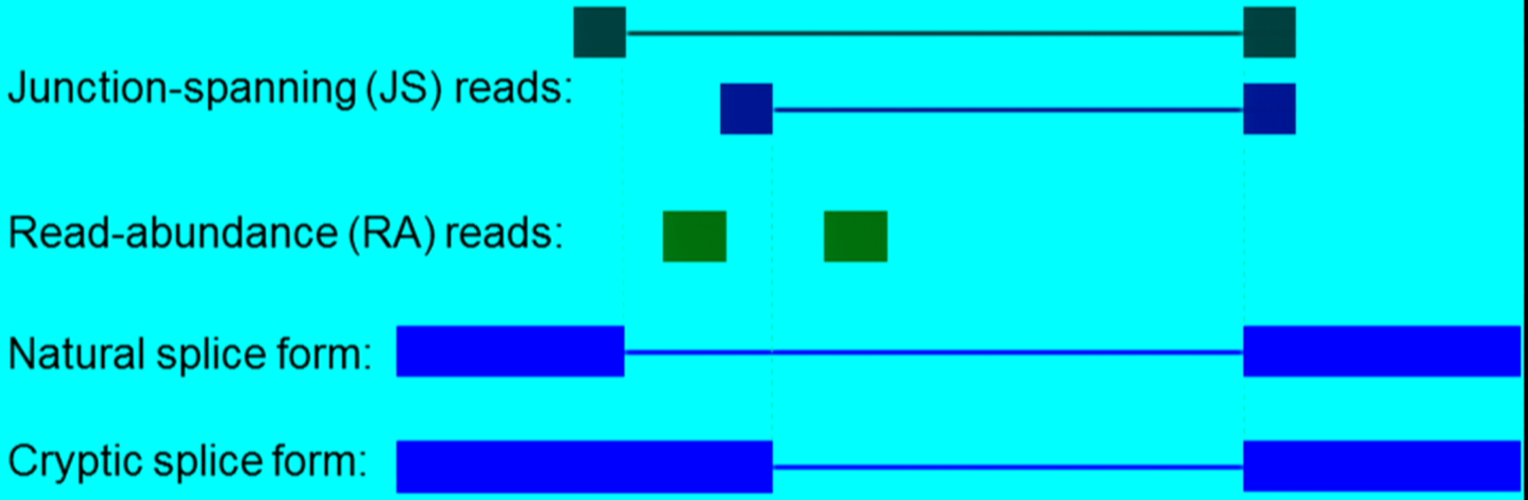
Junction-spanning (JS) reads:

Read-abundance (RA) reads:

Natural splice form:

Cryptic splice form:

Evidence Types

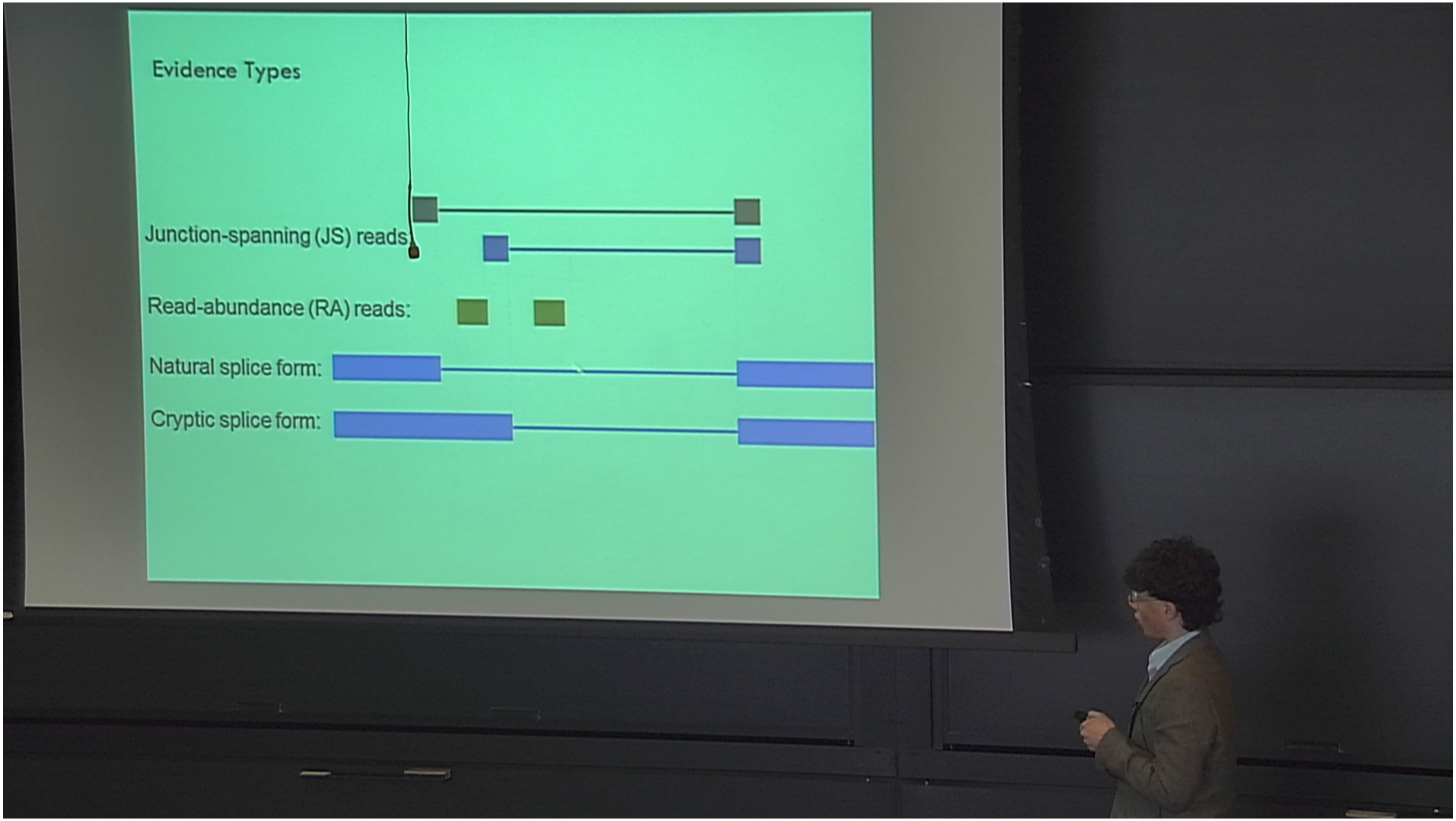


Junction-spanning (JS) reads:

Read-abundance (RA) reads:

Natural splice form:

Cryptic splice form:



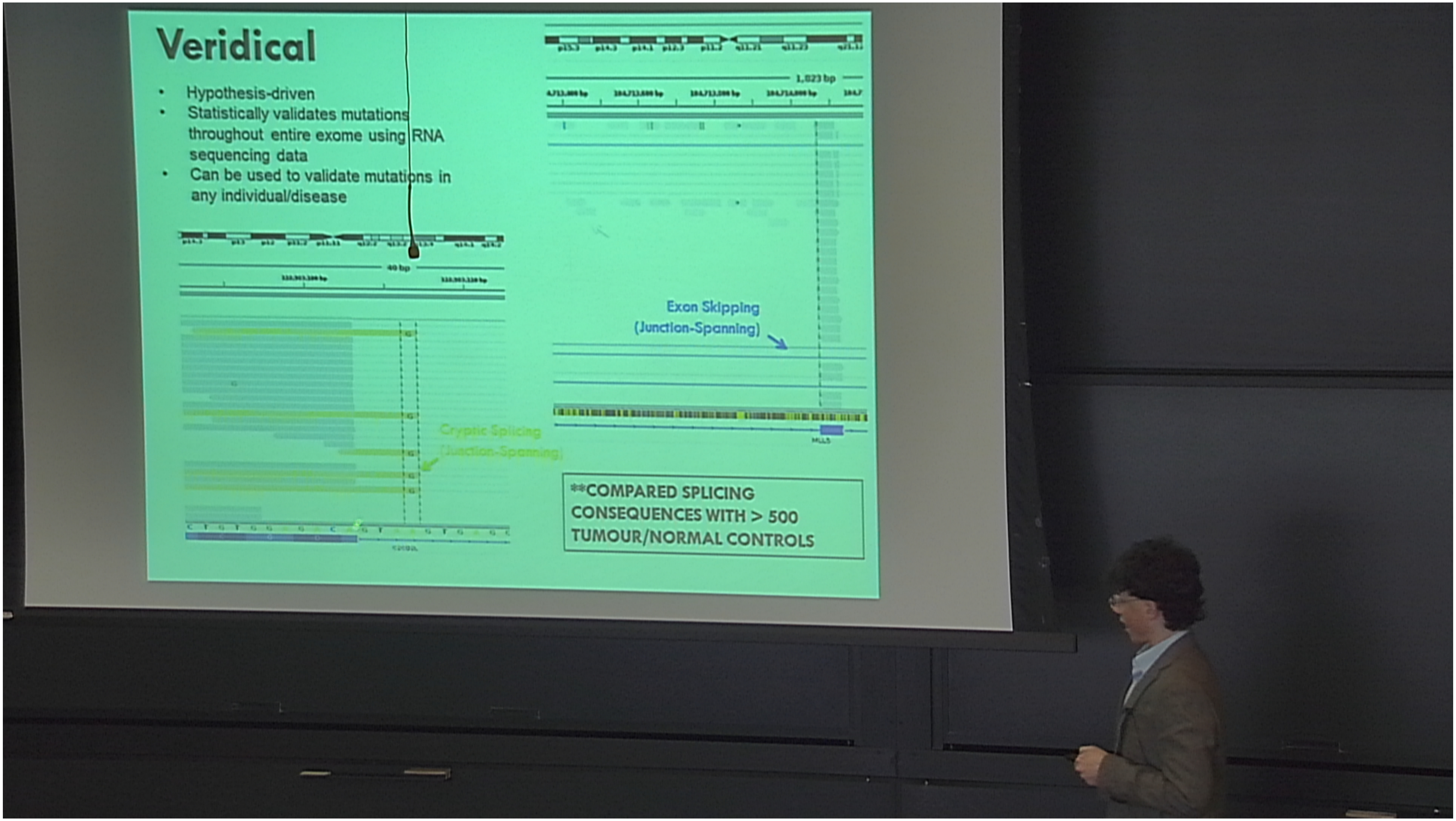
Evidence Types

Junction-spanning (JS) reads:

Read-abundance (RA) reads:

Natural splice form:

Cryptic splice form:



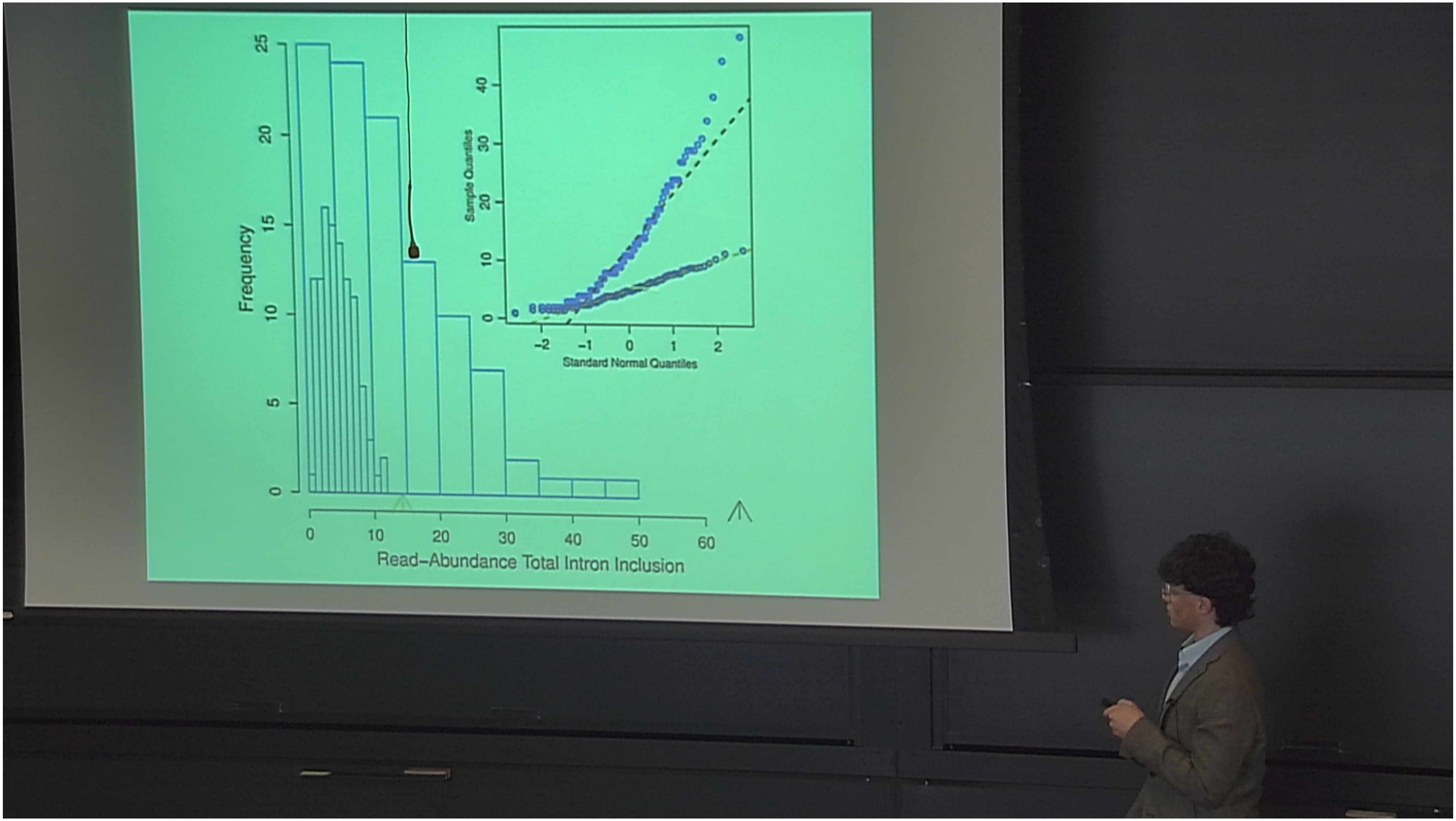
Veridical

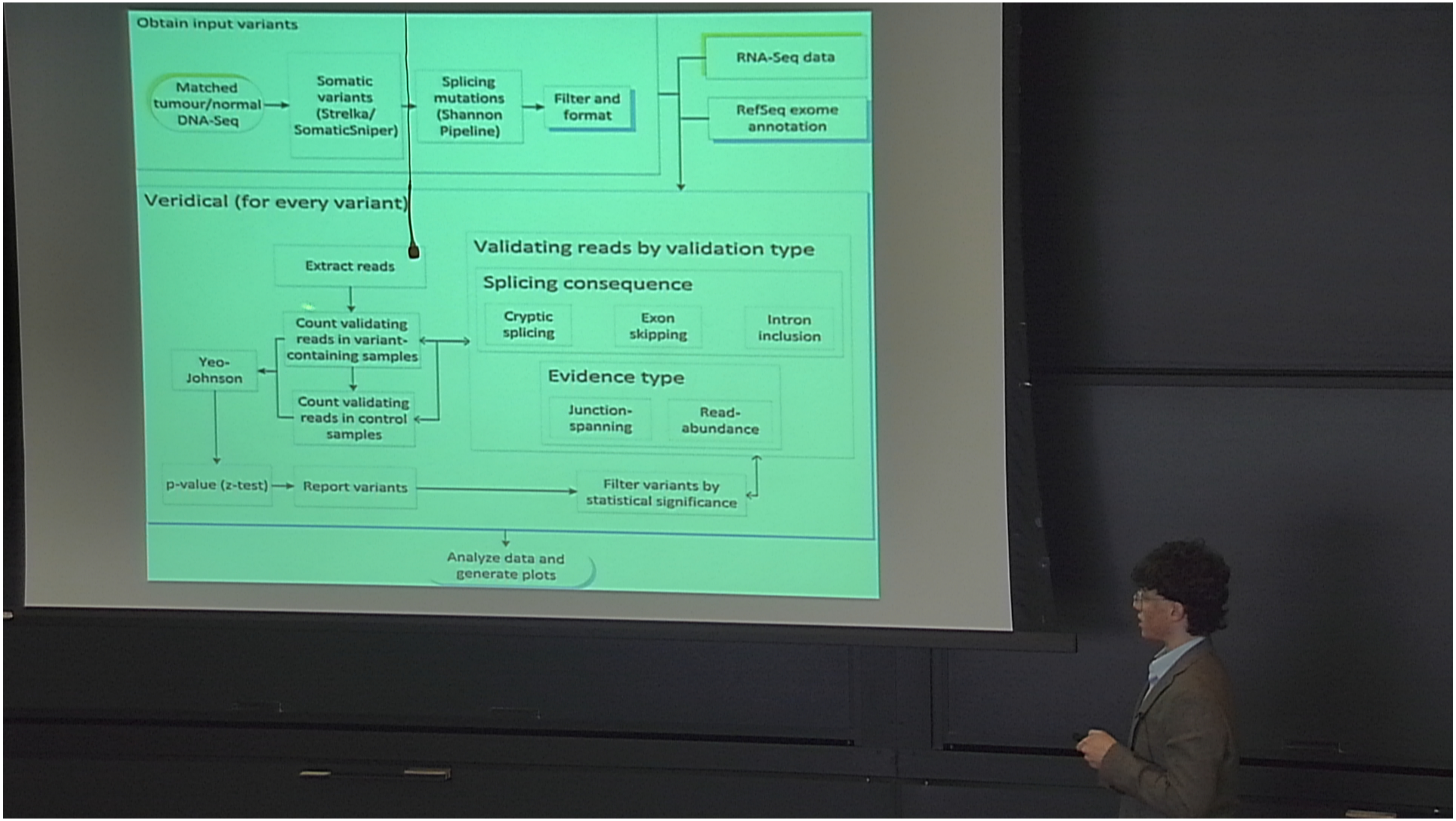
- Hypothesis-driven
- Statistically validates mutations throughout entire exome using RNA sequencing data
- Can be used to validate mutations in any individual/disease

Cryptic Splicing (Junction-Spanning)

Exon Skipping (Junction-Spanning)

**COMPARED SPLICING CONSEQUENCES WITH > 500 TUMOUR/NORMAL CONTROLS


Frequency
0
5
10
15
20
25
0
10
20
30
40
50
60
Read–Abundance Total Intron Inclusion
Sample Quantiles
0
10
20
30
40
−2
−1
0
1
2
Standard Normal Quantiles



Obtain input variants

Matched tumour/normal DNA-Seq

Somatic variants (Strelka/ SomaticSniper)

Splicing mutations (Shannon Pipeline)

Filter and format

RNA-Seq data

RefSeq exome annotation

Veridical (for every variant)

Extract reads

Count validating reads in variant-containing samples

Count validating reads in control samples

Yeo-Johnson

p-value (z-test)

Report variants

Filter variants by statistical significance

Validating reads by validation type

Splicing consequence

Cryptic splicing

Exon skipping

Intron inclusion

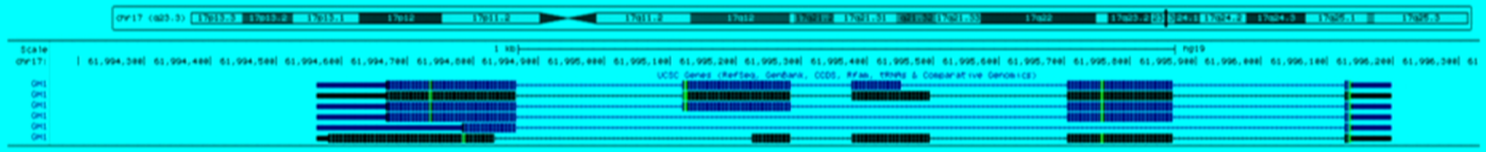
Evidence type

Junction-spanning

Read-abundance

Analyze data and generate plots

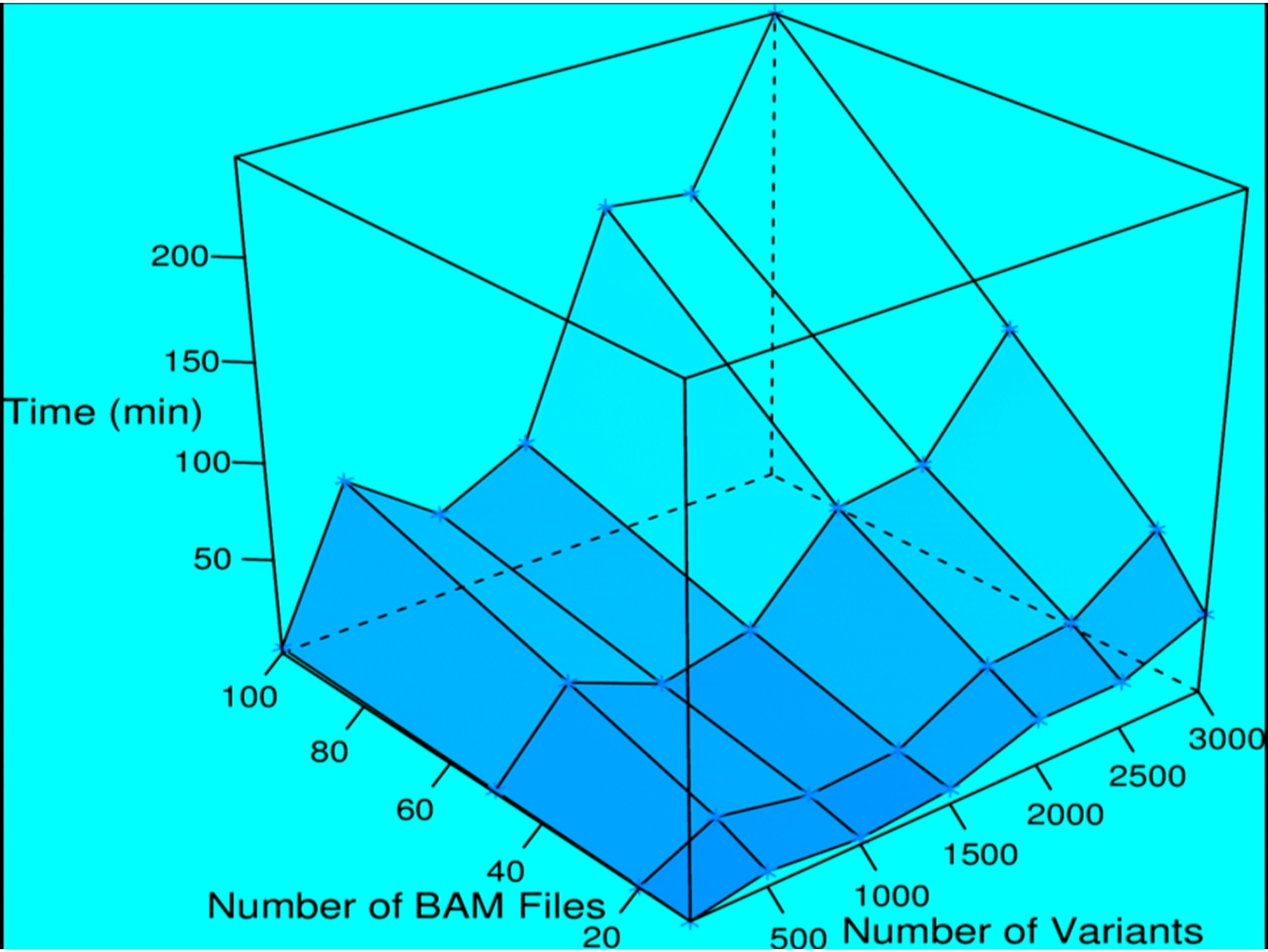
What we are *not* doing:

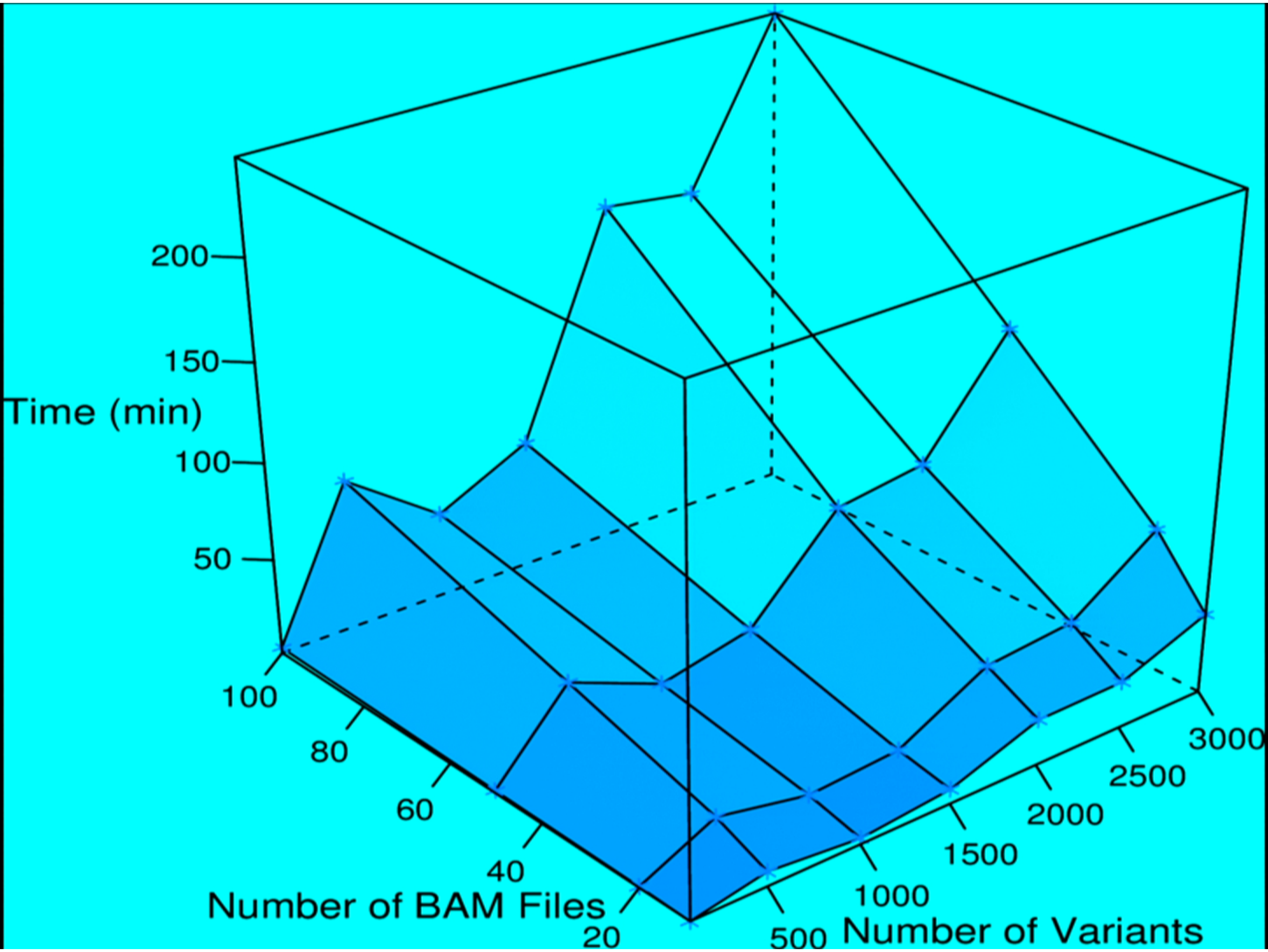


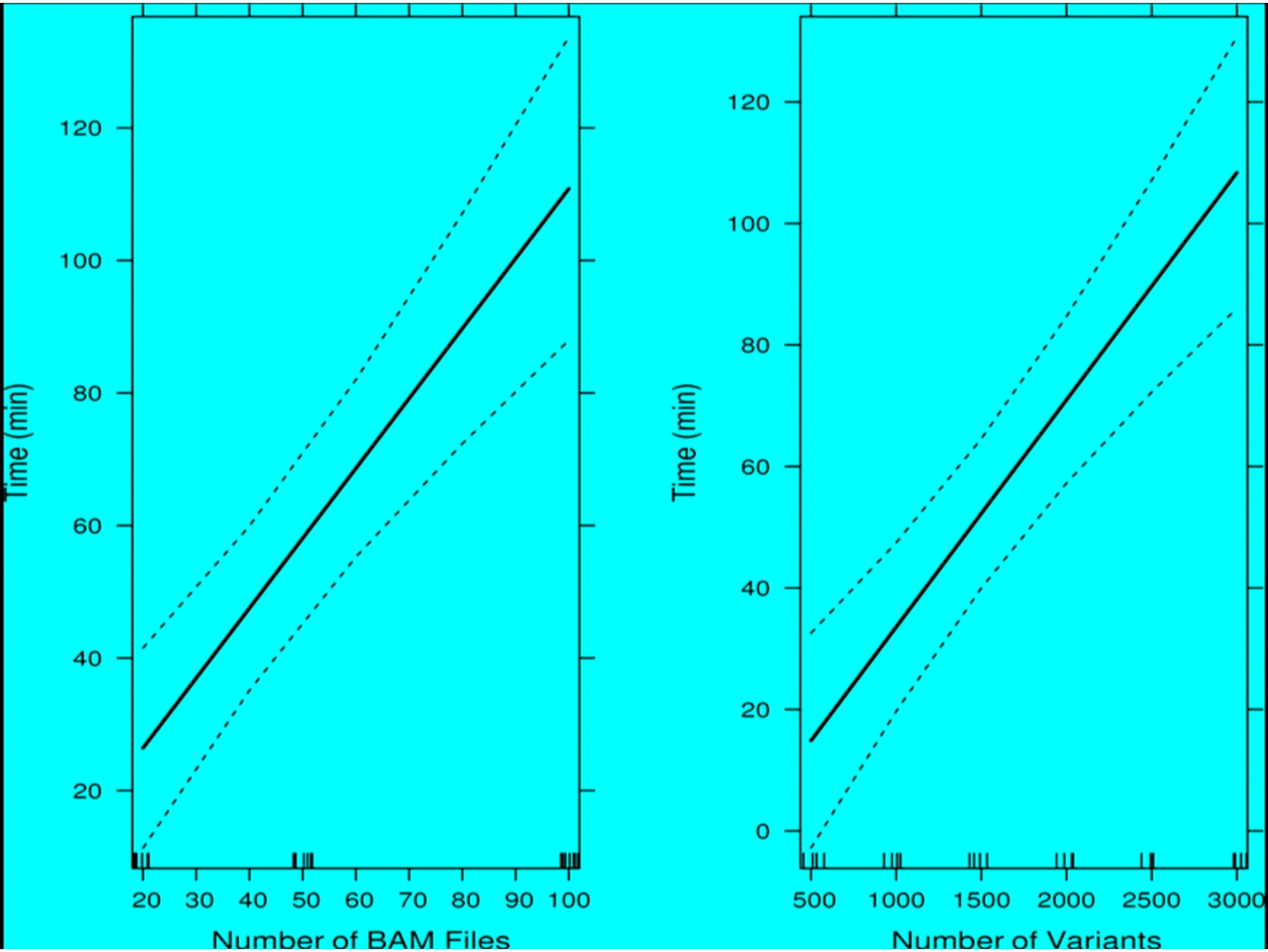
- Alternative splicing is a different problem
  - Many programs address this
    - Multivariate Analysis of Transcript Splicing (MATS)
      - Detects alternative splicing events via MCMC simulation sampling to compute p-values and FDRs
- Deriving a set of putative variants is a different problem
  - Variant Annotation, Analysis and Search Tool (VAAST)
    - Uses a likelihood approach to rank variants by pathogenicity
    - Does not conduct any detailed splicing mutation analyses

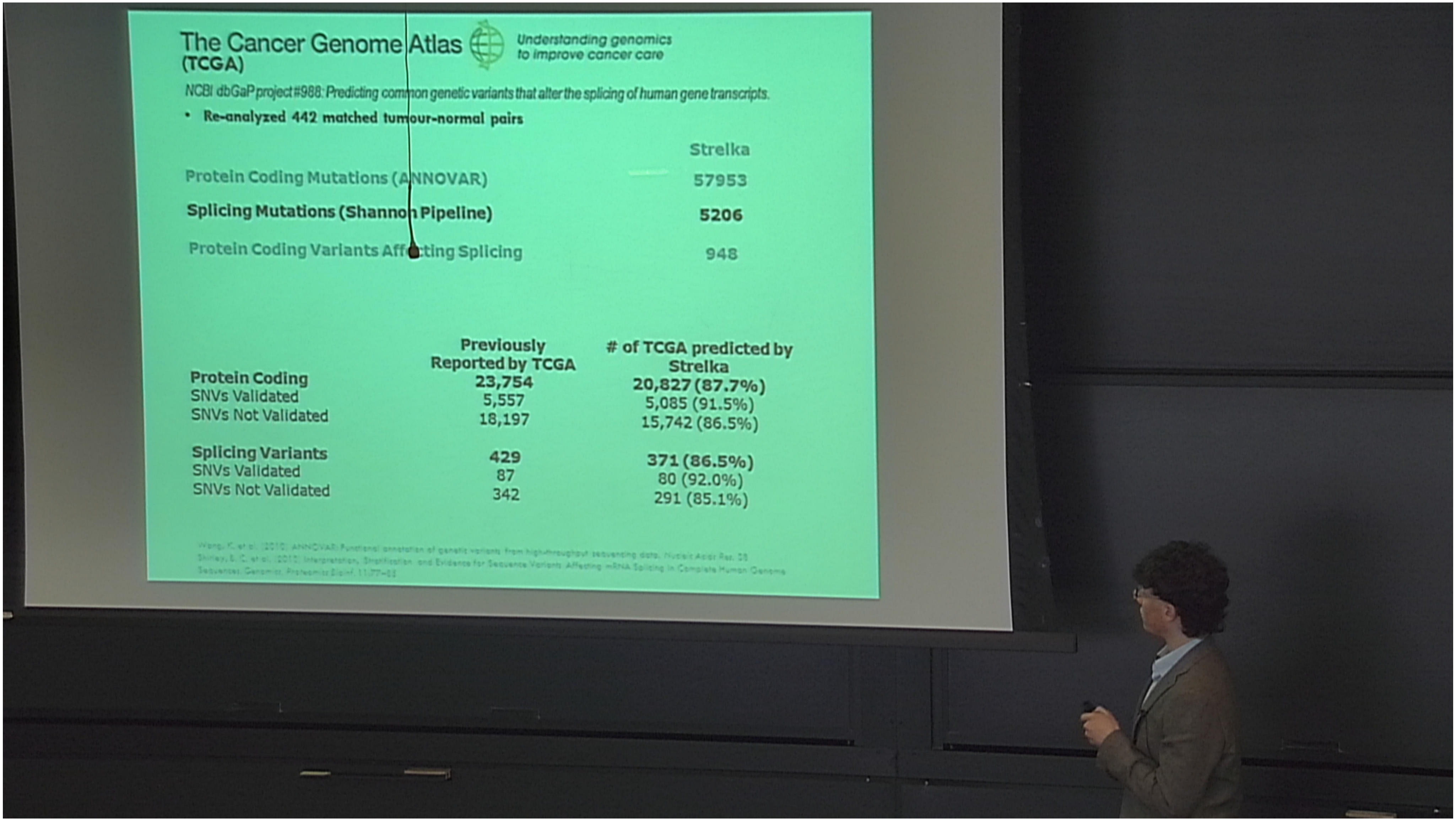
What we *are* doing:

- **Hypothesis testing:** A predicted mutation affects mRNA splicing
  - using variant predictions and an existing exome annotation


200
150
Time (min)
100
50
100
80
60
40
Number of BAM Files
20
500
1000
1500
2000
2500
3000
Number of Variants


Time (min)
200
150
100
50
Number of BAM Files
100
80
60
40
20
Number of Variants
500
1000
1500
2000
2500
3000


Time (min)
120
100
80
60
40
20
20
30
40
50
60
70
80
90
100
Number of BAM Files
Time (min)
120
100
80
60
40
20
0
500
1000
1500
2000
2500
3000
Number of Variants



The Cancer Genome Atlas (TCGA)

Understanding genomics to improve cancer care

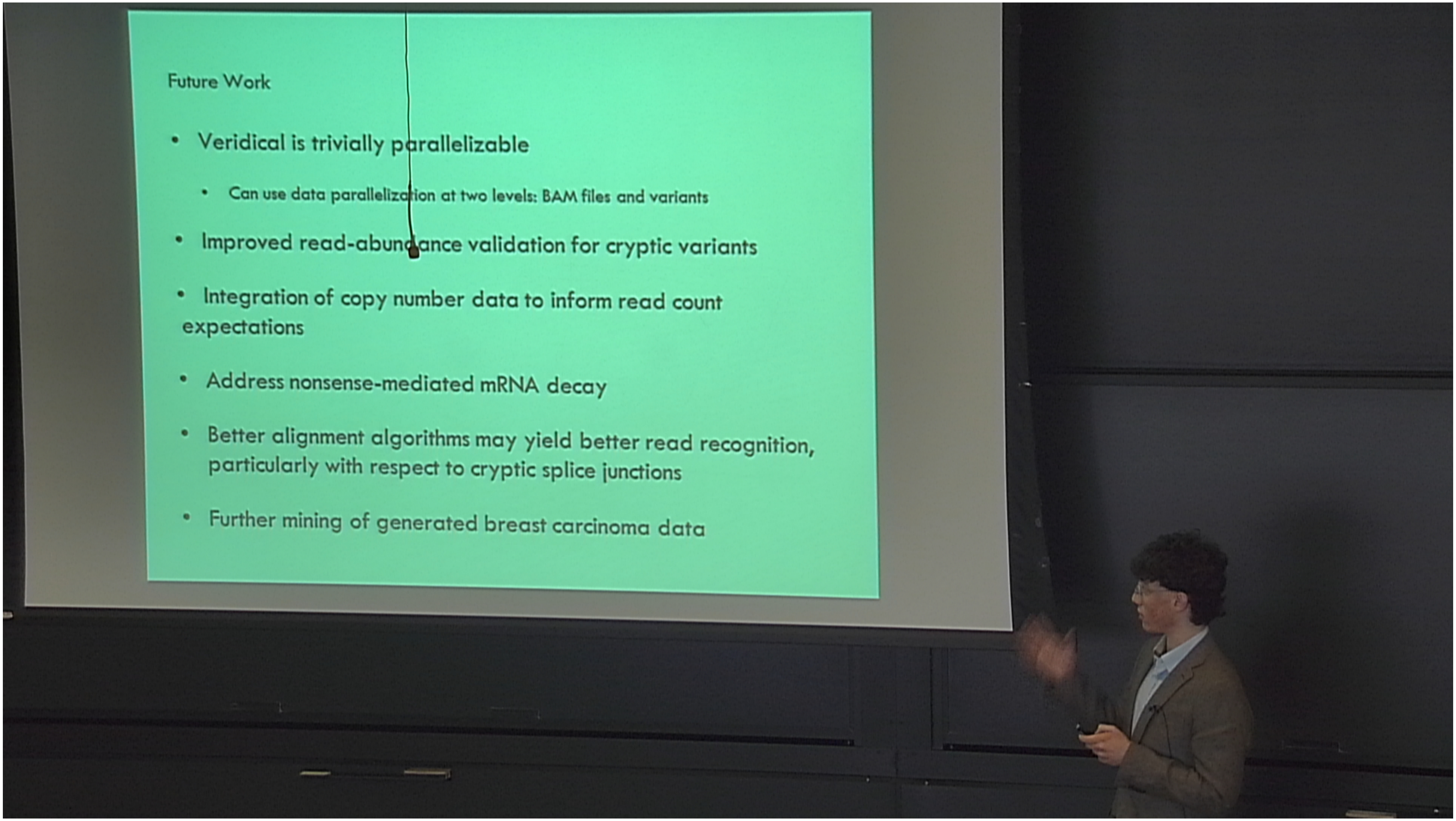
NCBI dbGaP project #988: Predicting common genetic variants that alter the splicing of human gene transcripts.

- **Re-analyzed 442 matched tumour-normal pairs**

	Strelka
Protein Coding Mutations (ANNOVAR)	57953
Splicing Mutations (Shannon Pipeline)	**5206**
Protein Coding Variants Affecting Splicing	948

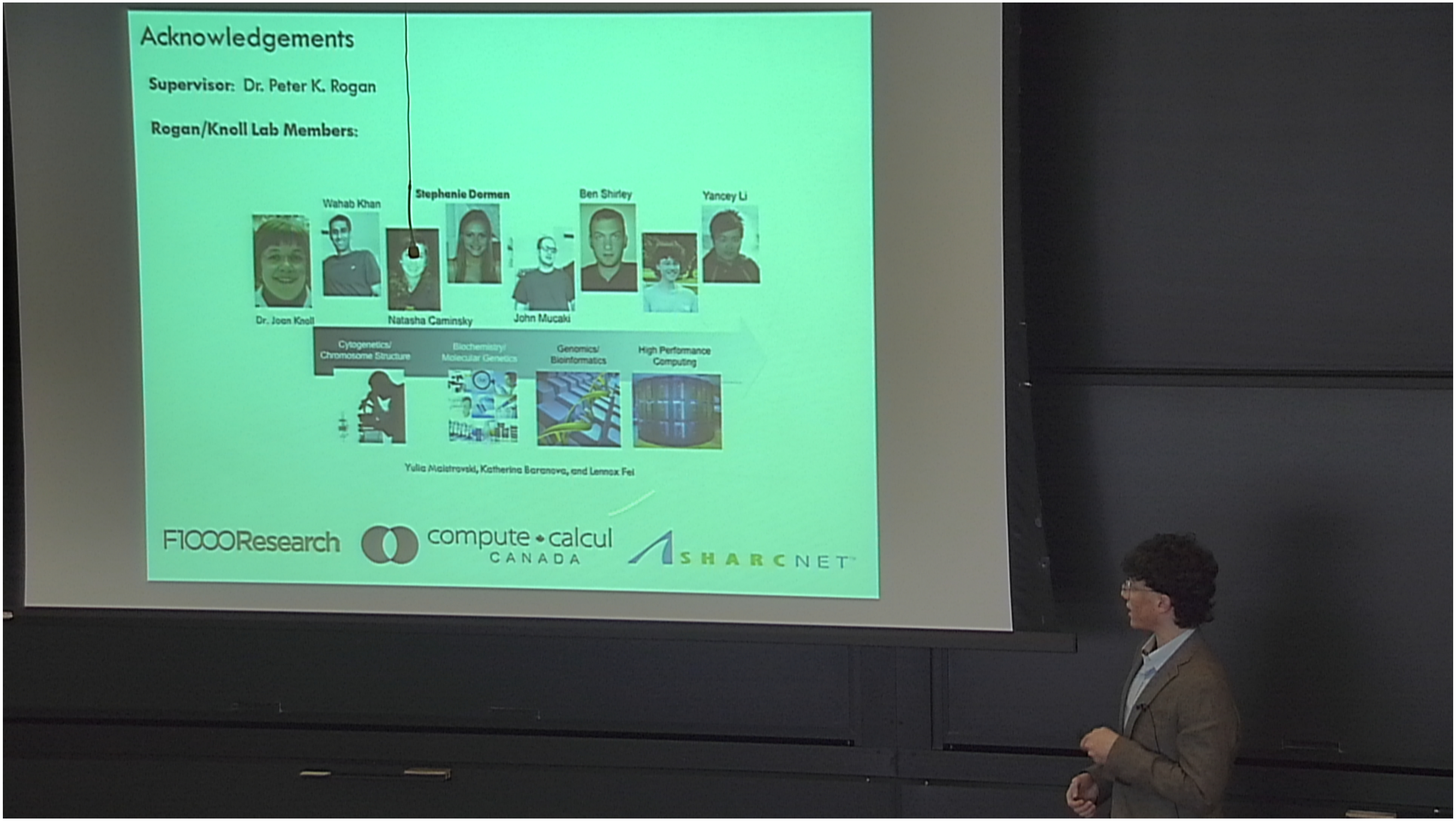
	Previously Reported by TCGA	# of TCGA predicted by Strelka
Protein Coding	**23,754**	**20,827 (87.7%)**
SNVs Validated	5,557	5,085 (91.5%)
SNVs Not Validated	18,197	15,742 (86.5%)
Splicing Variants	**429**	**371 (86.5%)**
SNVs Validated	87	80 (92.0%)
SNVs Not Validated	342	291 (85.1%)

Wang, K. et al. (2010) ANNOVAR: Functional annotation of genetic variants from high-throughput sequencing data. *Nucleic Acids Res.* 38
Shirley, B. C. et al. (2013) Interpretation, Stratification and Evidence for Sequence Variants Affecting mRNA Splicing in Complete Human Genome Sequences. Genomics, Proteomics Bioinf. 11:77–85



Future Work

- Veridical is trivially parallelizable
  - Can use data parallelization at two levels: BAM files and variants
- Improved read-abundance validation for cryptic variants
- Integration of copy number data to inform read count expectations
- Address nonsense-mediated mRNA decay
- Better alignment algorithms may yield better read recognition, particularly with respect to cryptic splice junctions
- Further mining of generated breast carcinoma data



Acknowledgements

Supervisor: Dr. Peter K. Rogan

Rogan/Knoll Lab Members:

Wahab Khan

Stephanie Dermen

Ben Shirley

Yancey Li

Dr. Joan Knoll

Natasha Caminsky

John Mucaki

Cytogenetics/ Chromosome Structure

Biochemistry/ Molecular Genetics

Genomics/ Bioinformatics

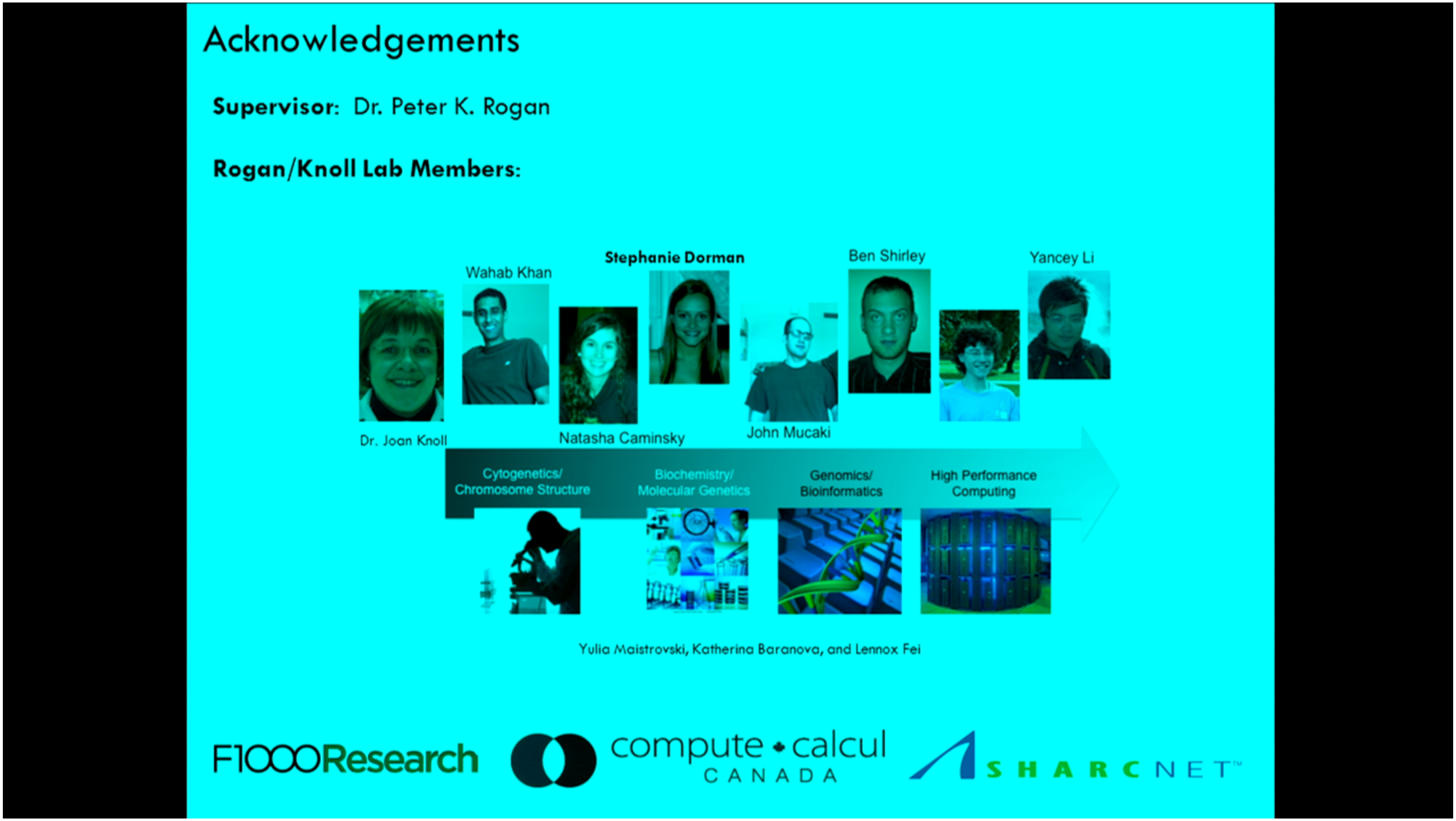
High Performance Computing

Yulia Maistrovski, Katherina Baranova, and Lennox Fei

F1000Research

compute • calcul CANADA

SHARCNET



Acknowledgements

Supervisor: Dr. Peter K. Rogan

Rogan/Knoll Lab Members:

Wahab Khan

Stephanie Dorman

Ben Shirley

Yancey Li

Dr. Joan Knoll

Natasha Caminsky

John Mucaki

Cytogenetics/ Chromosome Structure

Biochemistry/ Molecular Genetics

Genomics/ Bioinformatics

High Performance Computing

Yulia Maistrovski, Katherina Baranova, and Lennox Fei

F1000Research

compute • calcul CANADA

SHARCNET™

1500
2000
2500
3000
Number of Variants