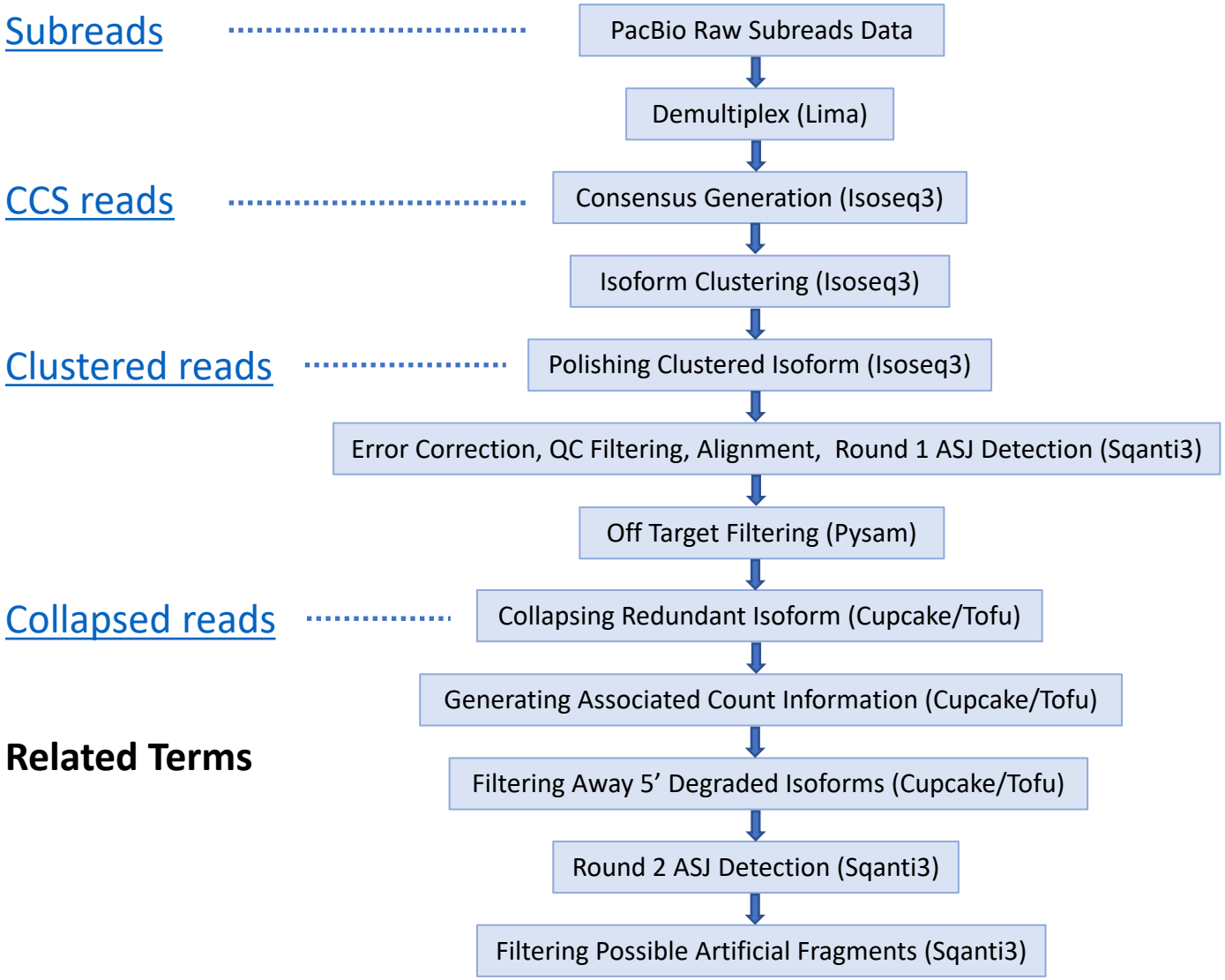
Glossary of Terms in Alternative Splicing Analysis Using PACBIO Targeted Long-Reads Sequencing data

Jieqiong Dai

* Related information and pictures shown in this presentation were collected online from PacBio and associated software packages.

Our Main Pipeline



Related Terms

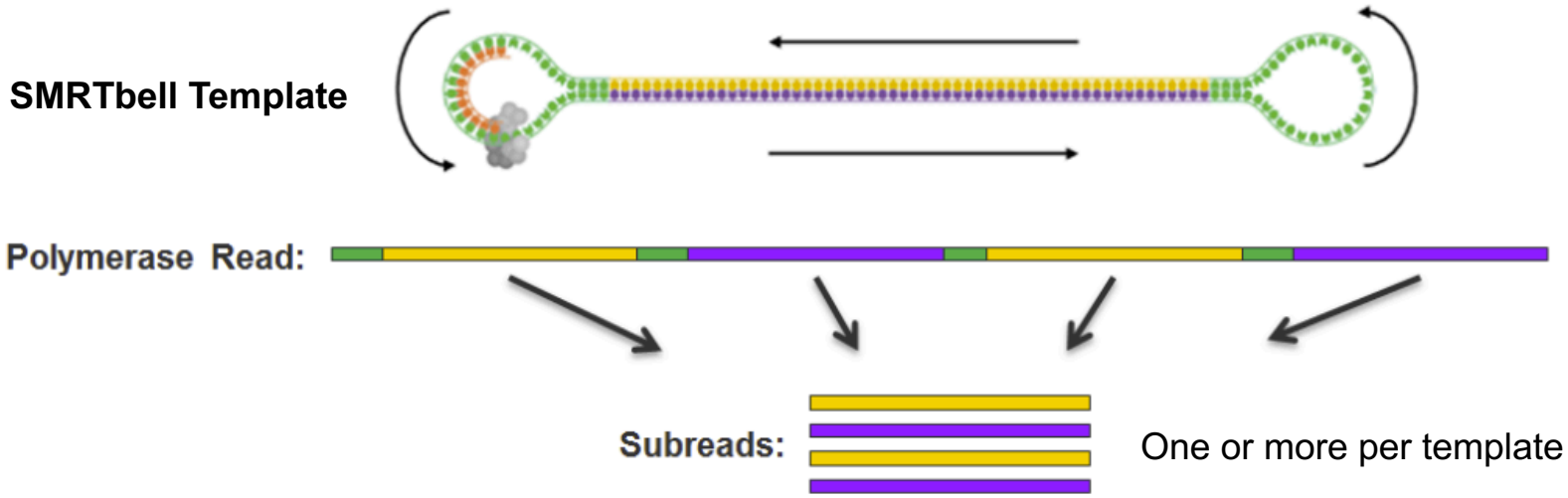
Glossary

- **Subreads**: reads containing sequence from a single pass of a polymerase on a single strand of an insert within a SMRTbell template.
- **CCS reads**: the consensus sequence resulting from alignment between subreads taken from a single ZMW.
- **Clustered reads**: polished ccs reads clustered at the transcript level.
- **Collapsed reads**: clustered reads collapsed at the transcript level; and redundant reads of the same transcript are removed.
- **SQANTI**: structural and quality annotation of novel transcript isoforms, a tool for splicing junction analysis.
- **FSM**: full splice match, a splice-based classification of major PacBio transcripts in SQANTI
- **ISM**: incomplete splice match, a splice-based classification of major PacBio transcripts in SQANTI
- **NIC**: novel in catalog, a splice-based classification of major PacBio transcripts in SQANTI
- **NNC**: novel not in catalog

pipeline

Subread

- Each polymerase read is partitioned to form one or more **subreads**, which contain sequence from a single pass of a polymerase on a single strand of an insert within a SMRTbell template (a double-stranded DNA template capped by hairpin adapters at both ends) and **no** adapter sequences.

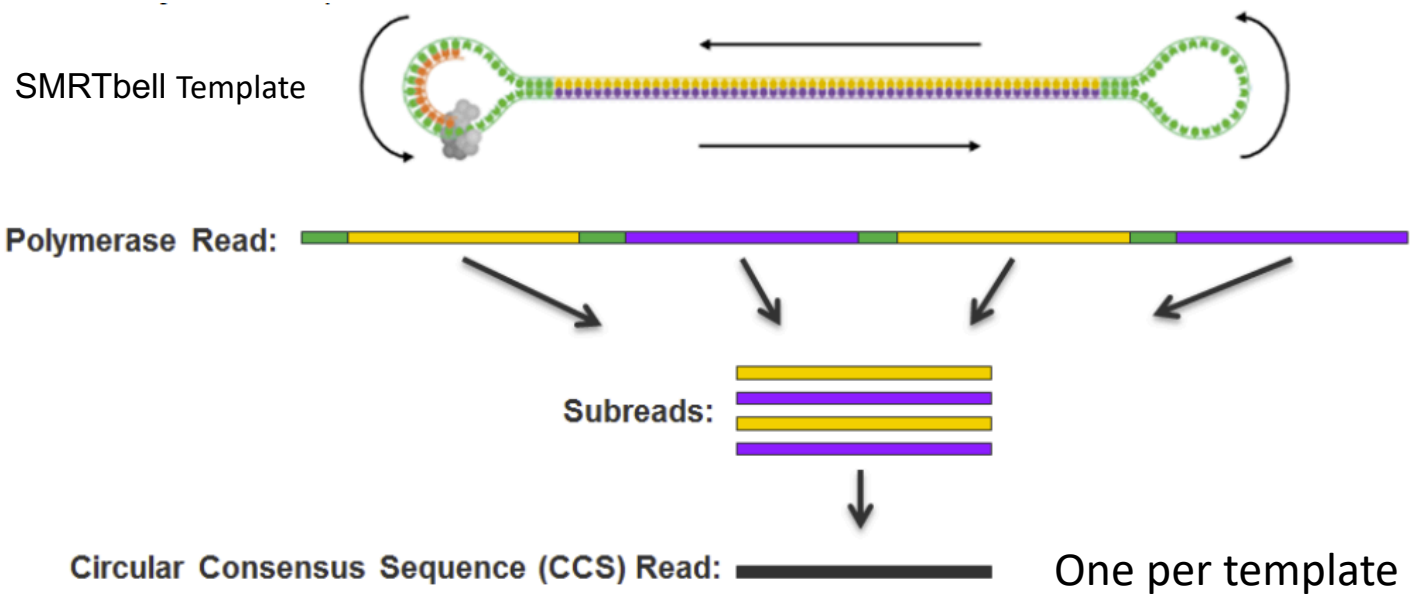


glossary

pipeline

Circular Consensus Sequencing (CCS) Read

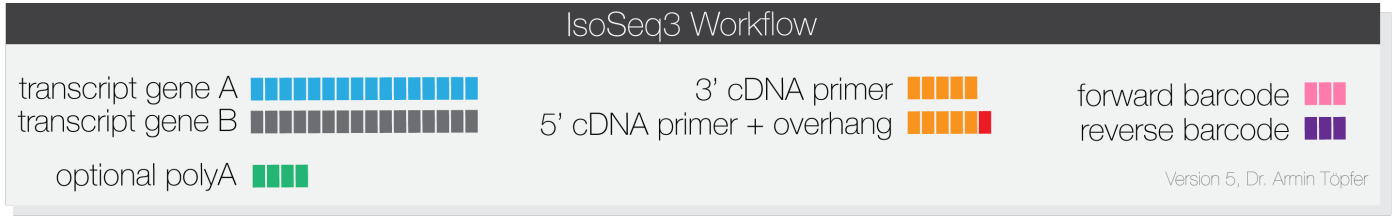
- The consensus sequence results from alignment between subreads taken from a single ZMW (**zero-mode waveguide**: a photonic nanostructure that allows single-molecule-resolved biophysical studies at relatively high concentrations of fluorescent molecules). Generating a CCS read does **not** include or require alignment against a reference sequence but **does** require at least two full-pass subreads from the insert.



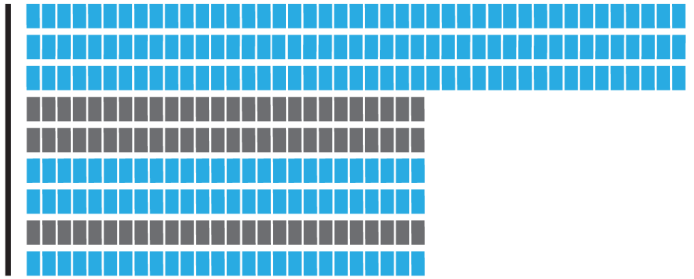
glossary

pipeline

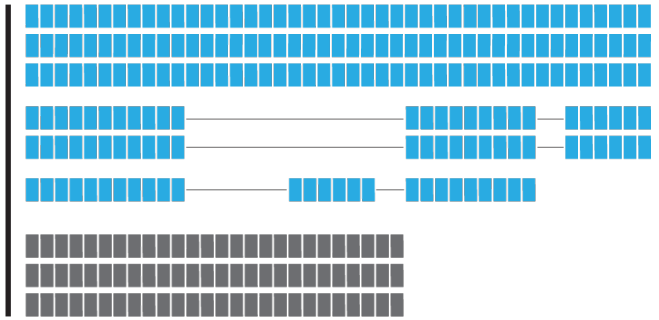
Clustered Reads and Collapsed Reads



CCS reads

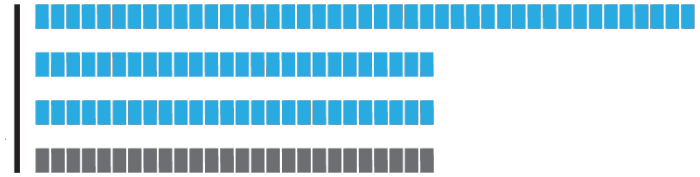


Clustered reads



Isoform level clusters;
Have quantification power

Collapsed reads

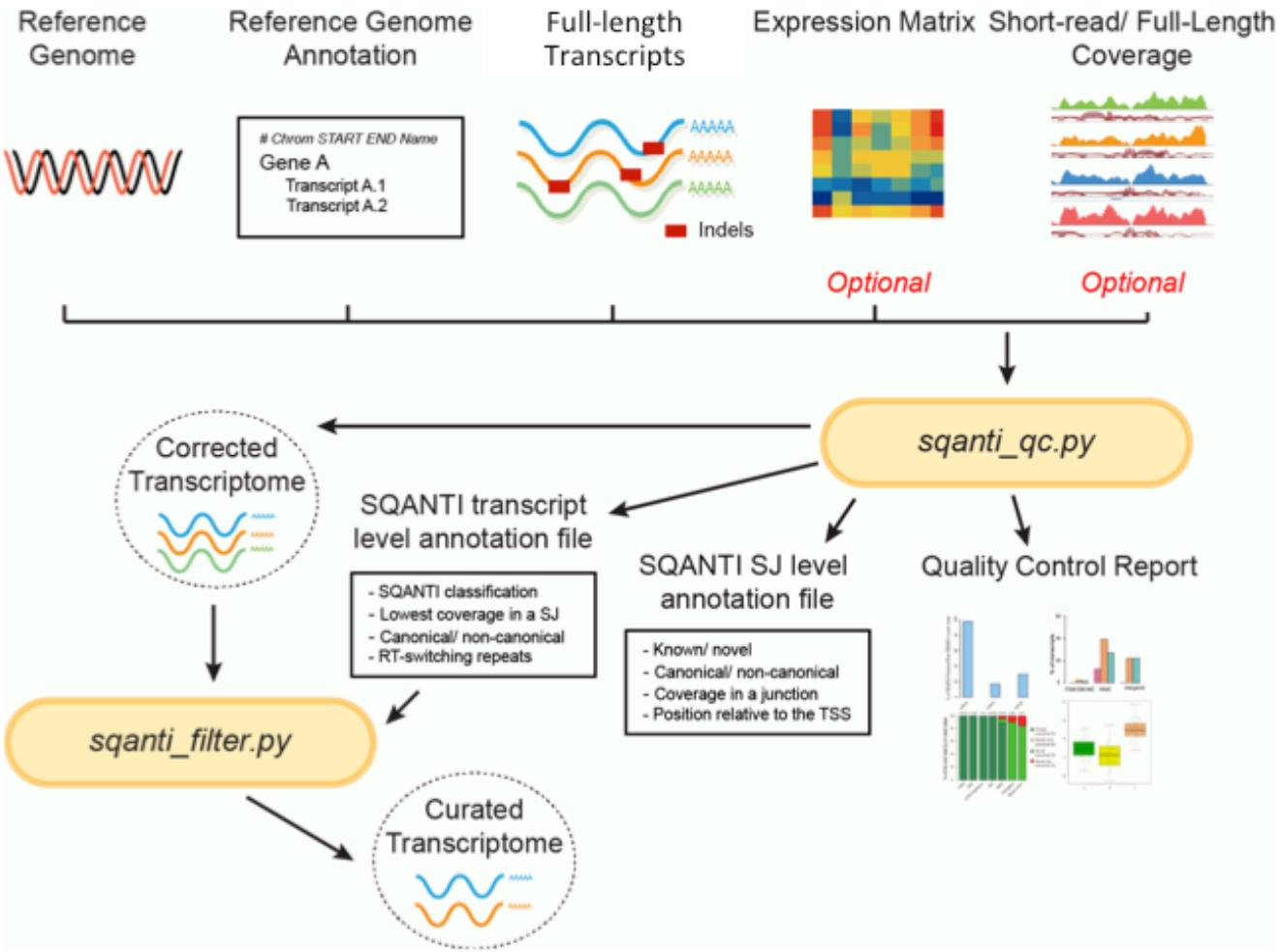


Redundant isoforms removed;
Lose quantification power

glossary

pipeline

SQANTI: Structural and Quality Annotation of Novel Transcript Isoforms



SQANTI pipeline steps:

1. First, as long-read sequencing usually has a high rate of errors along sequences, it performs a **reference-based correction of sequences**.
2. Secondly, it **generates genes models** and **classifies transcripts based on splice junctions**.
3. Third, it **predicts ORFs** for each transcript, obtaining information about the coding potential of each sequence.
4. Finally, it carries out a **deep characterization of isoforms at both transcript and junction level** and **generates a report** with several plots describing in detail the mayor attributes that catalog your set of sequenced isoforms.
5. Together with SQANTI_qc function, the user can use the **SQANTI filter function to remove isoforms potential to be artifacts**.

glossary

pipeline

Splice-based classification of major PACBIO transcripts in SQANTI

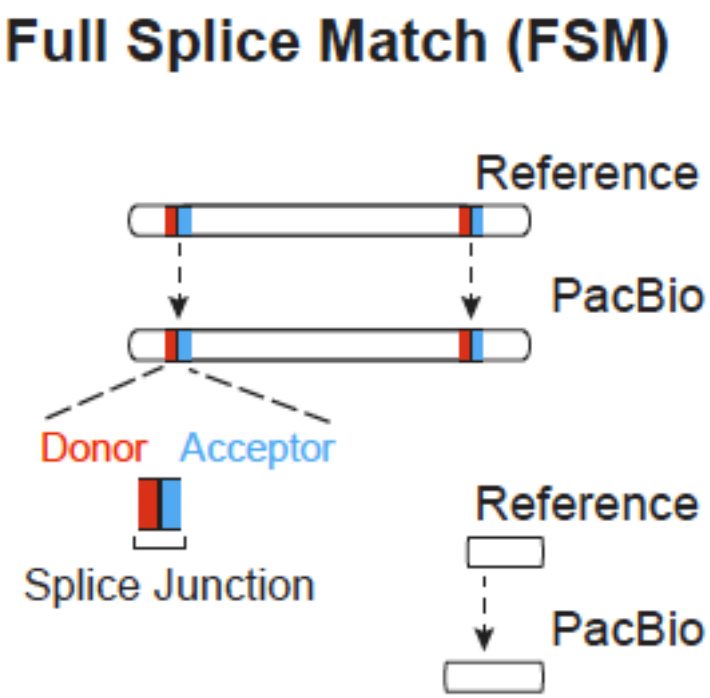
- (I) Known transcripts

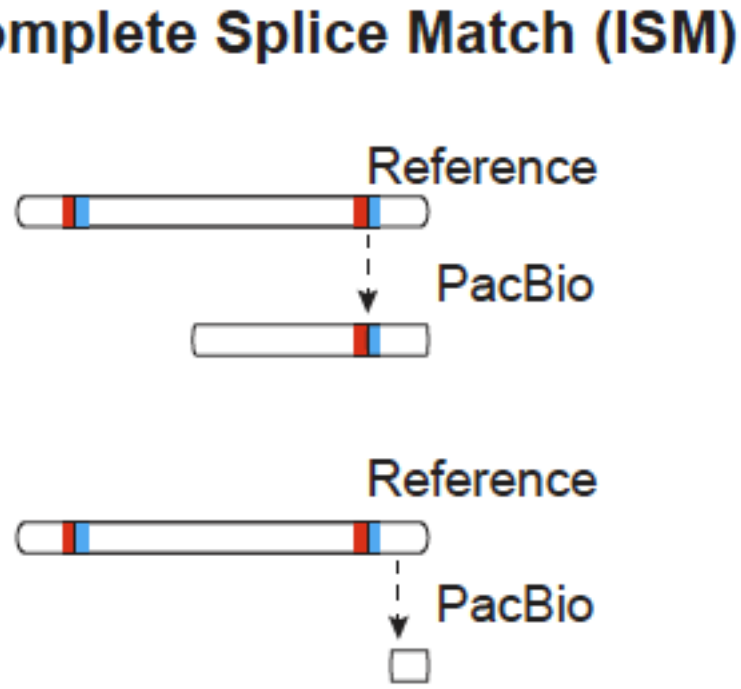
  FSM and ISM
- (II) Novel transcripts from known genes

  NIC and NNC

glossary

Splice-based classification (I): Known Transcripts





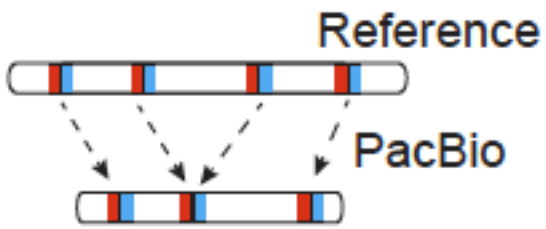
glossary

classification

Splice-based classification (II): Novel Transcripts from known genes

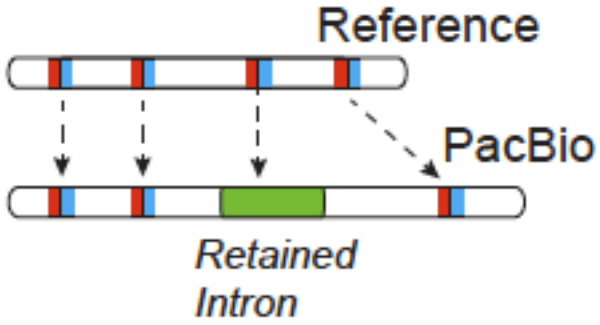
Novel In Catalog (NIC)

Novel SJs with known Donors and acceptors



loss of whole exon

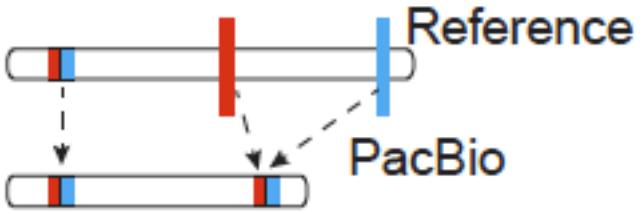
New combination of Known SJs



gain of whole intron

Novel Not in Catalog (NNC)

Novel SJs from novel Donors and/or Acceptors



glossary

classification