Dataset Construction Workflow

Process:

- 1. Extract chimeric reads from MDA & Bulk data
- 2. Compare MDA reads with 3 Bulk references
- 3. Classify based on support count (0-3)
- 4. Binary classification: Artificial vs Biological
- 5. Training data for sequence classification

Classification:

Artificial (Sup 0) Biological (Sup 1,2,3)

Bulk References MDA Data

MDA Data

Contains Chimeric

Artifacts Nanopore

Bulk Data 1

No Chimeric **Artifacts**

Compare MDA reads ALL 3 Bulk reference **Bulk Data 2**

No Chimeric **Artifacts**

Reference

Bulk Data 3

No Chimeric **Artifacts**

Reference

MDA Chimeric Reads



Bulk 1 Chimeric

Reads

Bulk 2

Chimeric Reads

Bulk 3 Chimeric Reads

Support 0

Artificial Chimeric

0/3 Bulk Match

Support 1

Biological Chimeric

omparison & Classification

1/3 Bulk Match

Support 2

Biological Chimeric

2/3 Bulk Match

Support 3

Biological Chimeric

3/3 Bulk Match

Artificial

Class 0

Sup 0

Biological Class 1

Sup 1,2,3 + Bulk