Dataset Construction Workflow

Process:

Sup 0

- 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)



Sup 1,2,3 + Bulk

Bulk References MDA Data

MDA Data **Bulk Data 1 Bulk Data 2 Bulk Data 3 Contains** No Chimeric No Chimeric No Chimeric Chimeric **Artifacts Artifacts Artifacts Artifacts** Reference Reference Reference Nanopore Bulk 2 **MDA** Bulk 1 Bulk 3 Chimeric Chimeric Chimeric Chimeric Reads Reads Reads Reads Support 0 Support 1 Support 2 Support 3 **Artificial Biological Biological Biological** Chimeric Chimeric Chimeric Chimeric 0/3 Bulk Match 1/3 Bulk Match 2/3 Bulk Match 3/3 Bulk Match **Artificial Biological** Class 0 Class 1