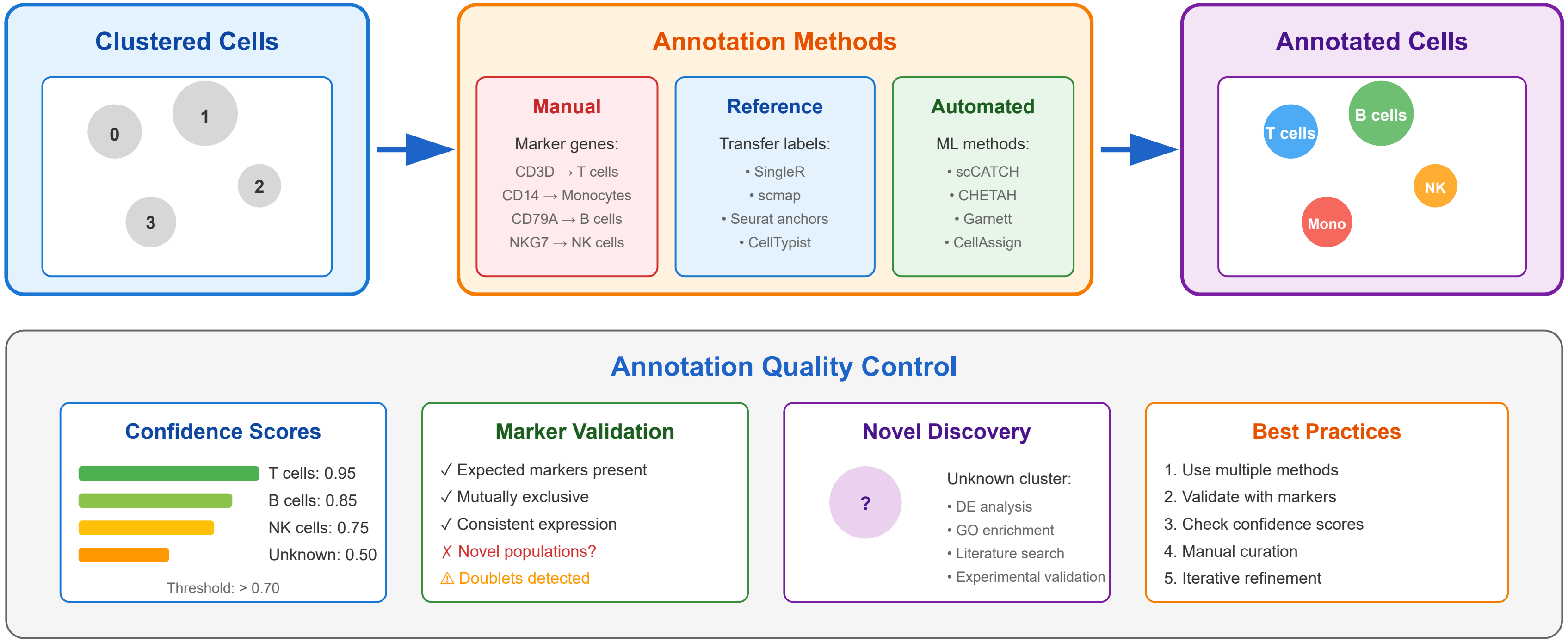
Cell Type Annotation

Cell Type Annotation Pipeline



Clustered Cells

Annotation Methods

Manual

Marker genes:
CD3D → T cells
CD14 → Monocytes
CD79A → B cells
NKG7 → NK cells

Reference

Transfer labels:
- SingleR
- scmap
- Seurat anchors
- CellTypist

Automated

ML methods:
- scCATCH
- CHETAH
- Garnett
- CellAssign

Annotated Cells

Annotation Quality Control

Confidence Scores

T cells: 0.95
B cells: 0.85
NK cells: 0.75
Unknown: 0.50
Threshold: > 0.70

Marker Validation

✓ Expected markers present
✓ Mutually exclusive
✓ Consistent expression
✗ Novel populations?
⚠ Doublets detected

Novel Discovery

Unknown cluster:
- DE analysis
- GO enrichment
- Literature search
- Experimental validation

Best Practices

1. Use multiple methods
2. Validate with markers
3. Check confidence scores
4. Manual curation
5. Iterative refinement

💡 Combine automated tools with manual curation for optimal results

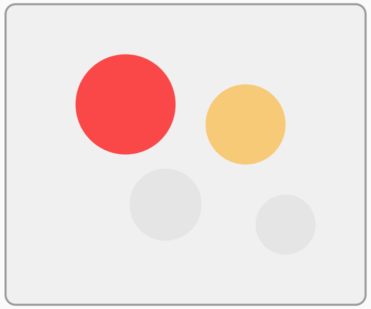
1. Manual Annotation Methods

📊 Marker Gene-Based Identification

Manual annotation relies on expert knowledge to identify cell types based on the expression patterns of known marker genes. This approach requires deep biological understanding and literature review but provides high-quality, interpretable results.

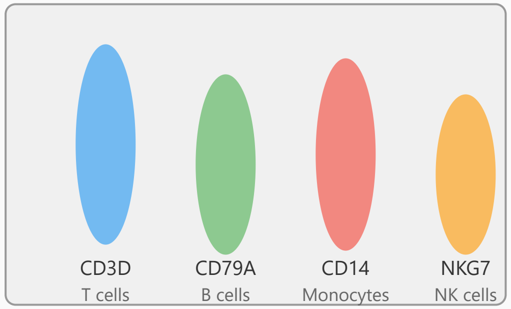
1. **Identify Differential Markers**
   Run differential expression analysis to find genes enriched in each cluster
2. **Visualize Gene Expression**
   Generate feature plots, violin plots, and heatmaps for known markers
3. **Literature Comparison**
   Cross-reference expression patterns with published cell type signatures
4. **Assign Cell Type Labels**
   Manually label clusters based on marker combinations and biological knowledge

Feature Plot: CD3D Expression

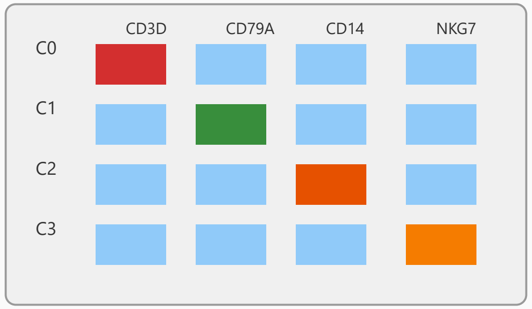


High CD3D = T cells

Violin Plot: Marker Genes



Heatmap: Top Markers



✅ Advantages

- Highly interpretable and biologically meaningful
- Leverages existing biological knowledge
- Flexible and adaptable to novel cell types
- No need for reference datasets

⚠ Limitations

- Time-consuming and labor-intensive
- Requires extensive domain expertise
- Subjective and prone to bias
- Not scalable for large datasets

- Direct visual inspection possible

- Difficult to maintain consistency across studies

💻 Example: Manual Annotation in Seurat (R)

```
# Find cluster markers cluster_markers <- FindAllMarkers(seurat_obj, only.pos = TRUE) # Visualize key markers
FeaturePlot(seurat_obj, features = c("CD3D", "CD79A", "CD14", "NKG7")) # Assign cell type labels based on
markers new.cluster.ids <- c("T cells", "B cells", "Monocytes", "NK cells") names(new.cluster.ids) <-
levels(seurat_obj) seurat_obj <- RenameIdents(seurat_obj, new.cluster.ids)
```

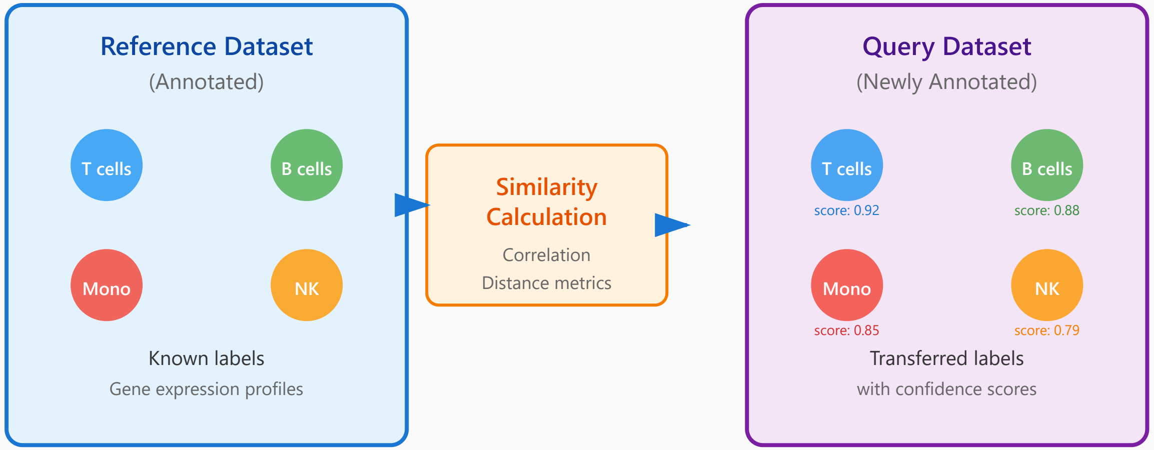
2. Reference-Based Annotation Methods

🔗 Label Transfer from Annotated References

Reference-based methods leverage well-annotated datasets to automatically transfer cell type labels to new query datasets. These approaches compare gene expression profiles between query and reference cells to predict cell identities based on similarity.

1. **Select Reference Dataset**
   Choose a high-quality, well-annotated reference from similar tissue/conditions
2. **Compute Similarity Scores**
   Calculate correlation or distance between query and reference cells
3. **Transfer Labels**
   Assign labels based on nearest neighbors or ensemble voting
4. **Quality Assessment**
   Evaluate confidence scores and validate assignments

Reference-Based Label Transfer Workflow


Reference Dataset

(Annotated)

Known labels

Gene expression profiles

Similarity Calculation

Correlation

Distance metrics

Query Dataset

(Newly Annotated)

Transferred labels

with confidence scores

🔧 Popular Tools & Methods

SingleR

Correlation-based method using curated reference datasets

Seurat Anchors

Integration-based label transfer with CCA/RPCA

scmap

Fast nearest neighbor search for cell type projection

CellTypist

Machine learning classifier with pre-trained models

✅ Advantages

- Fast and scalable for large datasets
- Consistent annotations across studies
- Leverages community-curated references
- Provides confidence scores
- Minimal manual intervention required

⚠ Limitations

- Limited by reference dataset quality
- Cannot identify novel cell types
- Batch effects can reduce accuracy
- May fail for rare cell populations
- Requires appropriate reference selection

💻 Example: SingleR Annotation (R)

```
library(SingleR) library(celldex) # Load reference dataset ref <- HumanPrimaryCellAtlasData() # Run SingleR
annotation predictions <- SingleR(test = query_data, ref = ref, labels = ref$label.main) # Add annotations to
```

```
Seurat object seurat_obj$celltype <- predictions$labels seurat_obj$annotation_score <- predictions$scores
```

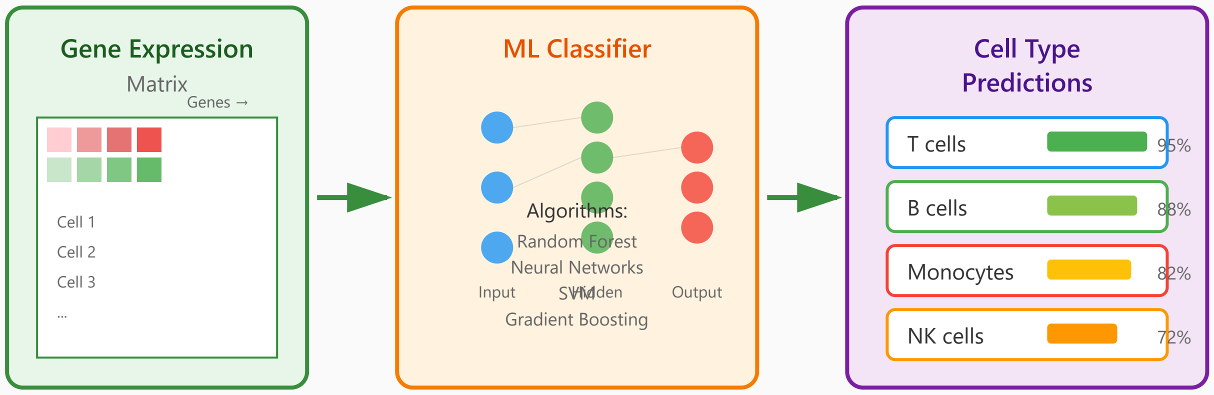
3. Automated Machine Learning Methods

🤖 AI-Powered Cell Type Prediction

Automated methods use machine learning algorithms to classify cell types based on gene expression patterns. These tools can be trained on existing data or use pre-defined marker databases to make predictions without requiring manual inspection.

1. **Data Preprocessing**
   Normalize and prepare expression matrix for ML algorithms
2. **Model Training/Selection**
   Train classifier or use pre-trained model on marker databases
3. **Prediction & Scoring**
   Classify cells and generate probability scores for each type
4. **Validation & Refinement**
   Assess predictions quality and refine low-confidence calls

Machine Learning Classification Pipeline



🛠 Automated Annotation Tools

scCATCH

Evidence-based scoring from marker databases

Garnett

Supervised classifier with marker file specification

CHETAH

Hierarchical classification with confidence scoring

CellAssign

Probabilistic model for marker-based assignment

✅ Advantages

- Highly scalable and reproducible
- Minimal manual curation needed
- Handles complex datasets efficiently
- Can detect subtle expression patterns
- Built-in confidence metrics

⚠ Limitations

- Black-box nature reduces interpretability
- Requires high-quality training data
- May overfit to training dataset biases
- Limited by predefined marker databases
- Difficult to validate novel predictions

💻 Example: CellTypist Annotation (Python)

```
import celltypist from celltypist import models # Load pre-trained model model =
models.Model.load(model='Immune_All_Low.pkl') # Predict cell types predictions = celltypist.annotate(adata,
model=model, majority_voting=True) # Add predictions to AnnData adata.obs['predicted_labels'] =
predictions.predicted_labels adata.obs['conf_score'] = predictions.probability
```

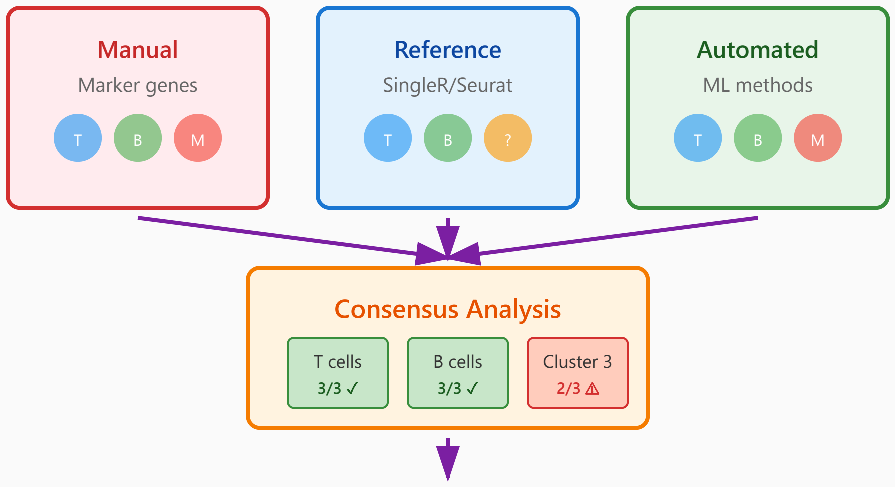
4. Hybrid & Multi-Method Approaches

🔄 Combining Multiple Annotation Strategies

The most robust approach combines manual curation, reference-based methods, and automated tools to leverage the strengths of each strategy. This iterative workflow produces high-confidence annotations while maintaining biological interpretability and scalability.

1. **Initial Automated Annotation**
   Apply multiple automated/reference methods for rapid first-pass labeling
2. **Consensus Analysis**
   Compare predictions across methods, identify agreements and conflicts
3. **Manual Validation**
   Expert review of low-confidence or conflicting predictions using markers
4. **Iterative Refinement**
   Update annotations, re-cluster if needed, and validate final assignments

Integrated Multi-Method Annotation Workflow



Final Curated Annotations

Quality Metrics

✓ High confidence (>90%):	85%
⚠ Medium confidence:	12%
✗ Low confidence (<50%):	3%
Method agreement:	92%
Manual validation:	100%
Novel cell types:	2
Cells annotated:	15,847

📋 Best Practice Checklist

✓ Run 2-3 different annotation methods

✓ Validate with canonical markers

✓ Manually inspect low-confidence cells

✓ Check for doublets/multiplets

✓ Compare results for consistency

✓ Review confidence scores

✓ Document annotation decisions

✓ Investigate novel populations

✅ Advantages

- Maximizes accuracy through consensus
- Balances speed with quality
- Identifies method-specific biases
- Enables discovery of novel cell types
- Provides comprehensive confidence metrics
- Maintains biological interpretability

⚠ Considerations

- Requires more computational resources
- Longer analysis time investment
- Needs expertise across multiple tools
- Resolving conflicts can be subjective
- More complex workflow management

💻 Example: Multi-Method Consensus Workflow

```
# Step 1: Run multiple methods manual_labels <- ManualAnnotation(seurat_obj, marker_genes) singler_labels <-
SingleR(seurat_obj, ref_data) automated_labels <- CellTypist(seurat_obj, model) # Step 2: Create consensus
matrix consensus <- CompareAnnotations( list(manual = manual_labels, singler = singler_labels, automated =
automated_labels) ) # Step 3: Identify high-confidence consensus final_labels <- consensus %>%
filter(agreement >= 2/3) %>% select(consensus_label, confidence_score) # Step 4: Manual review of conflicts
conflicts <- consensus %>% filter(agreement < 2/3) reviewed_labels <- ManualReview(seurat_obj, conflicts)
```

Key Takeaways

Successful cell type annotation requires a strategic combination of methods. Start with automated tools for efficiency, validate with reference datasets for consistency, and refine with manual curation for accuracy. Always assess confidence scores, validate with known markers, and document your annotation decisions for reproducibility.