Creating supermatrices and phylogenetic datasets with SuperCRUNCH

Daniel Portik

Postdoctoral Researcher
University of Arizona

Phylogenetic Data Workflows

Handful of tools available for assembling phylogenetic data

Phylogenetic Data Workflows

Handful of tools available for assembling phylogenetic data

Each has great features!

PhyLoTa	(Sanderson et al. 2008)
PHLAWD	(Smith et al. 2009)
phyloGenerator	(Pearse & Purvis 2013)
SUMAC	(Freyman, 2015)
SUPERSMART	(Antonelli et al. 2017)
PhylotaR	(Bennett et al. 2018)
PyPHLAWD	(Smith & Walker 2018)

Phylogenetic Data Workflows

Common limitations for my work needs

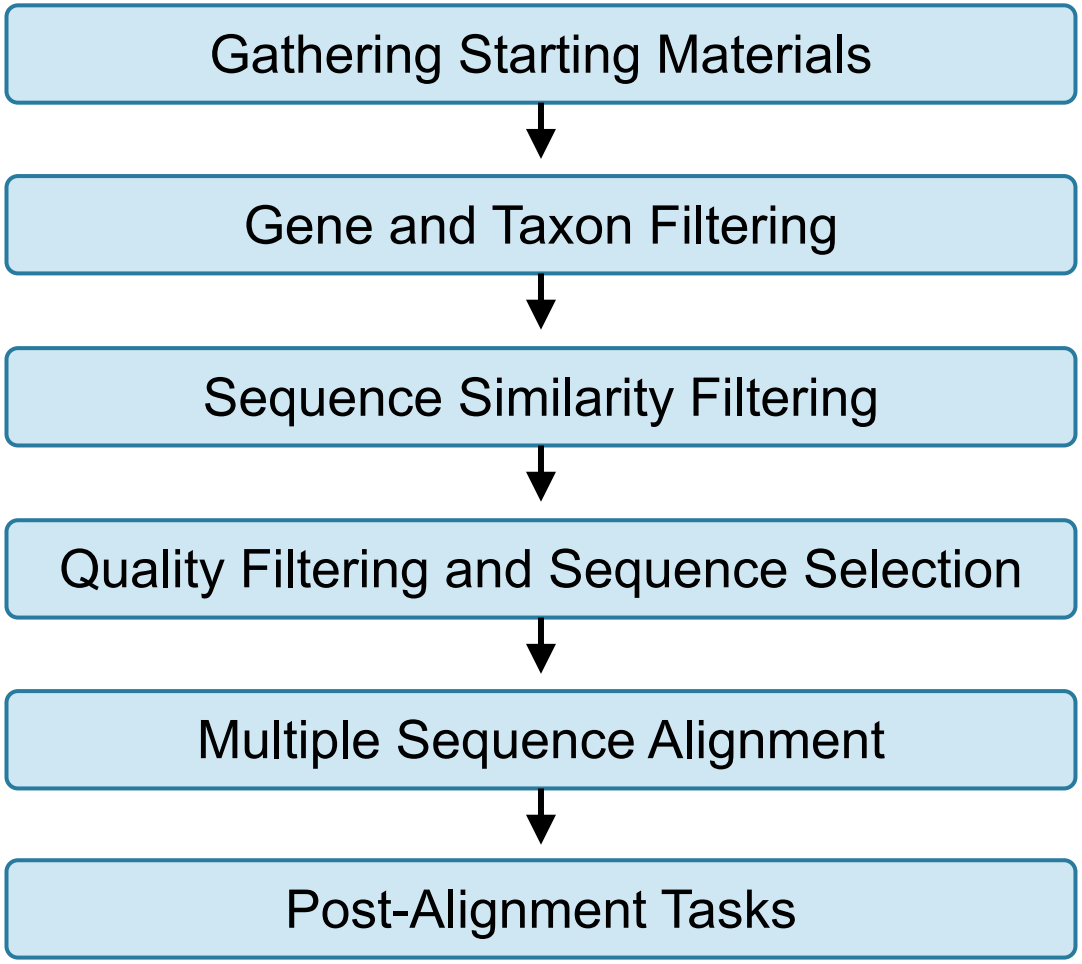
- Sequences fetched using GenBank release database
- Reliance on NCBI Taxonomy for identifying sequences
- Automated clustering of whole sequence sets
- Black-box methods for sequence filtering and selection

SuperCRUNCH

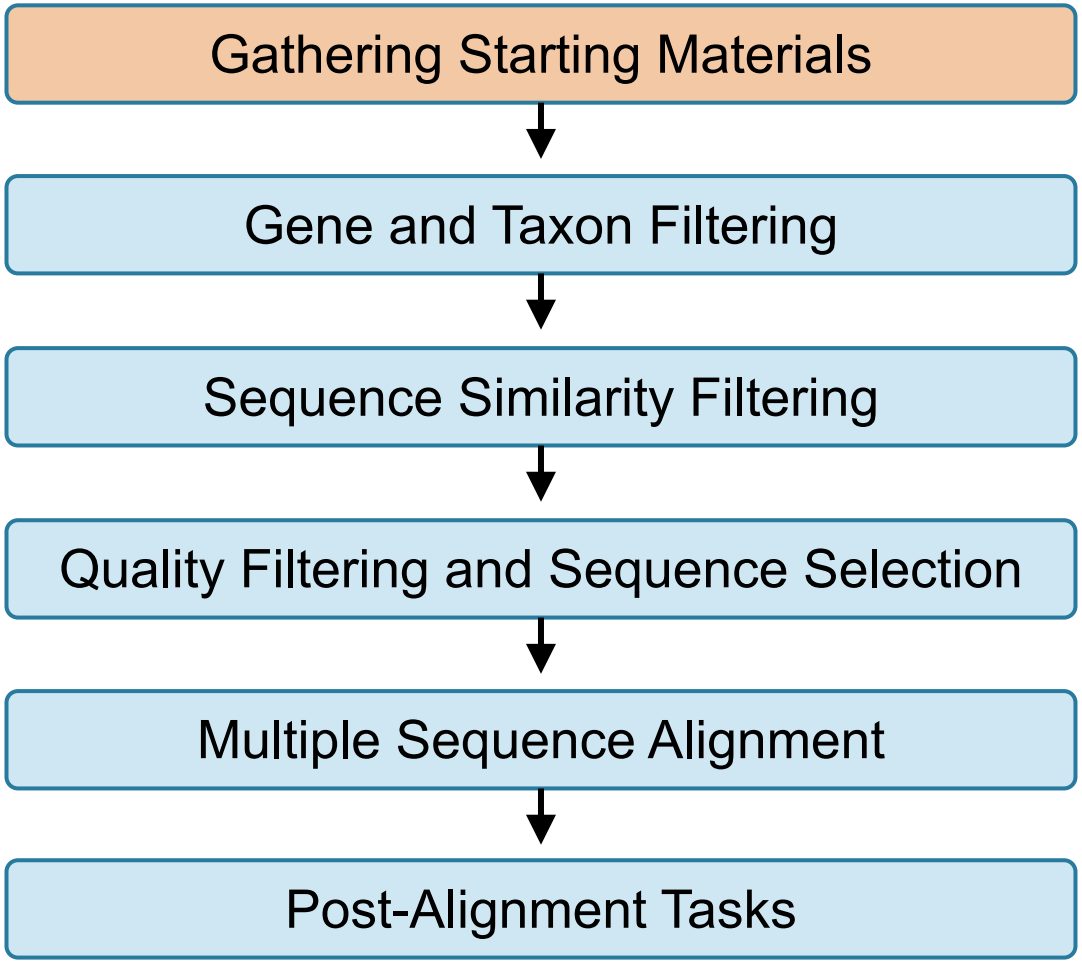
FOR PHYLOGENETIC DATA

- Bioinformatic toolkit for phylogenetic data
- Allows ANY fasta-formatted sequence data to be used
- Targeted searches based on list of taxa and genes
- Transparent methods for orthology-filtering and sequence selection, including detailed output files
- Can be used to assemble traditional supermatrices as well as population-level datasets
- Can process large sequence capture datasets (UCE's)

SuperCRUNCH Workflow



SuperCRUNCH Workflow



Gathering Starting Materials

Sequence Data

Taxon List

Gene List

Gathering Starting Materials

Sequence Data | Taxon List | Gene List

NCBI/GenBank sequences and/or Locally generated sequences

Gathering Starting Materials

Sequence Data | Taxon List | Gene List

NCBI/GenBank sequences and/or Locally generated sequences

Fasta format:

```
>KP820543.1 Callisaurus draconoides voucher R45 aryl hydrocarbon
receptor (anr) gene, partial cds
CACAATGAGAAAGCCTTGATAAACCGTGATCGGACTTTGCCACTCGTTGAAGAAATAGATGAGAGCT…

>KP820544.1 Urosaurus ornatus voucher UWBM7587 aryl hydrocarbon
receptor (anr) gene, partial cds
TAAAATCTCCTTTGAAAGGAACCTTTTTGTGGACACCAGGGATGAATTAGGTAATGTAATGGCCAGA…
```

Gathering Starting Materials

Sequence Data | Taxon List | Gene List

NCBI/GenBank sequences and/or Locally generated sequences

Fasta format: > **Unique Identifier** **Taxon Label** **Description**

```
>KP820543.1 Callisaurus draconoides voucher R45 aryl hydrocarbon
receptor (anr) gene, partial cds
CACAATGAGAAAGCCTTGATAAACCGTGATCGGACTTTGCCACTCGTTGAAGAAATAGATGAGAGCT…

>KP820544.1 Urosaurus ornatus voucher UWBM7587 aryl hydrocarbon
receptor (anr) gene, partial cds
TAAAATCTCCTTTGAAAGGAACCTTTTTGTGGACACCAGGGATGAATTAGGTAATGTAATGGCCAGA…
```

Gathering Starting Materials

Sequence Data | Taxon List | Gene List

Text file with species and subspecies labels

```
Callisaurus draconoides crinitus
Callisaurus draconoides draconoides
Callisaurus draconoides inusitanus
Cophosaurus texanus
Holbrookia maculata
Uma exsul
Uma inornata
Uma notata
Uma paraphygas
Uma rufopunctata
Uma scoparia
```

Gathering Starting Materials

Sequence Data | Taxon List | Gene List

Text file with species and subspecies labels

```
Callisaurus draconoides crinitus
Callisaurus draconoides draconoides
Callisaurus draconoides inusitanus
Cophosaurus texanus
Holbrookia maculata
Uma exsul
Uma inornata
Uma notata
Uma paraphygas
Uma rufopunctata
Uma scoparia
```

List Sources

- Taxonomic Databases
- NCBI Taxonomy Browser
- Directly from sequences (via SuperCRUNCH)

Gathering Starting Materials

Sequence Data | Taxon List | Gene List

Tab-delimited text file with search terms

- Search terms include a gene abbreviation and description
- Can include any number and combination of genes (mtDNA, nuclear, seq-cap)
- Tetrapod 5k UCE set and other example files on github

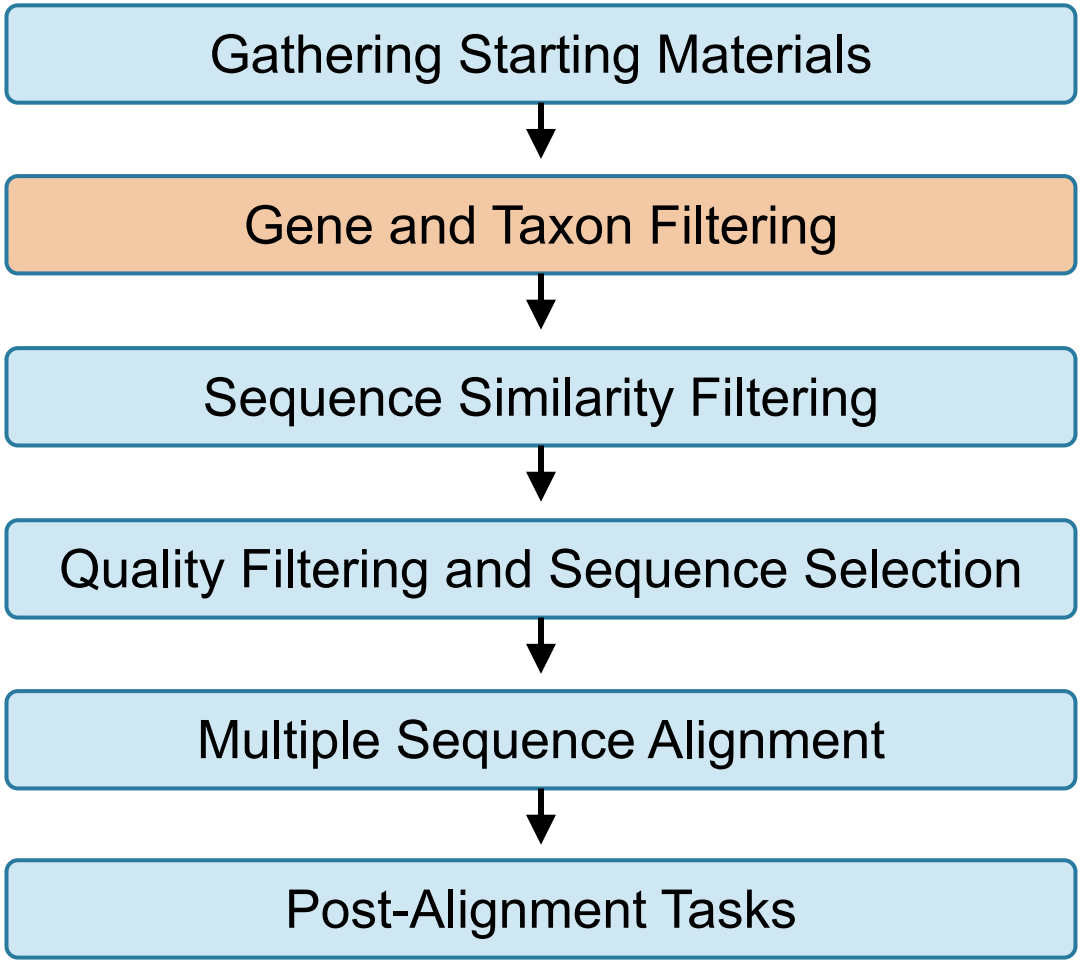
Gathering Starting Materials

Sequence Data | Taxon List | **Gene List**

Tab-delimited text file with search terms

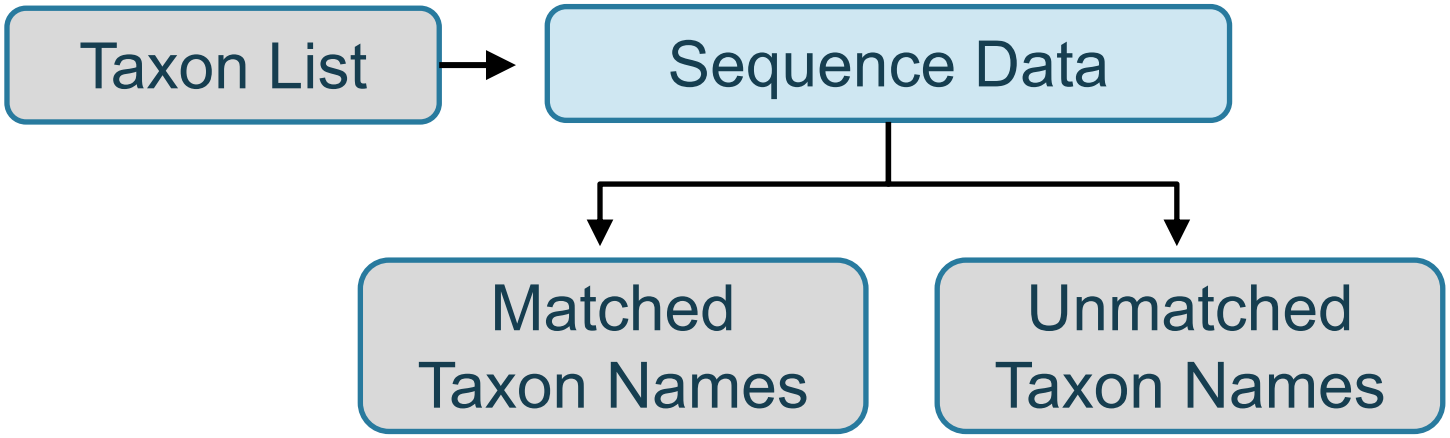
Label	Abbreviation	Description
CMOS	CMOS;C-MOS	oocyte maturation factor
CXCR4	CXCR4	chemokine C-X-C motif receptor 4;C-X-C …
DLL1	DLL1;DLL	distal-less
DNAH3	DNAH3	dynein axonemal heavy chain 3;dynein ax…
ECEL1	ECEL1;ECEL	endothelin converting;endothelin conver…
ENC1	ENC1	ectodermal neural cortex 1
EXPH5	EXPH5	exophilin;exophilin 5;exophilin-5;exoph…
FSHR	FSHR	follicle stimulating hormone;follicle-s…

SuperCRUNCH Workflow



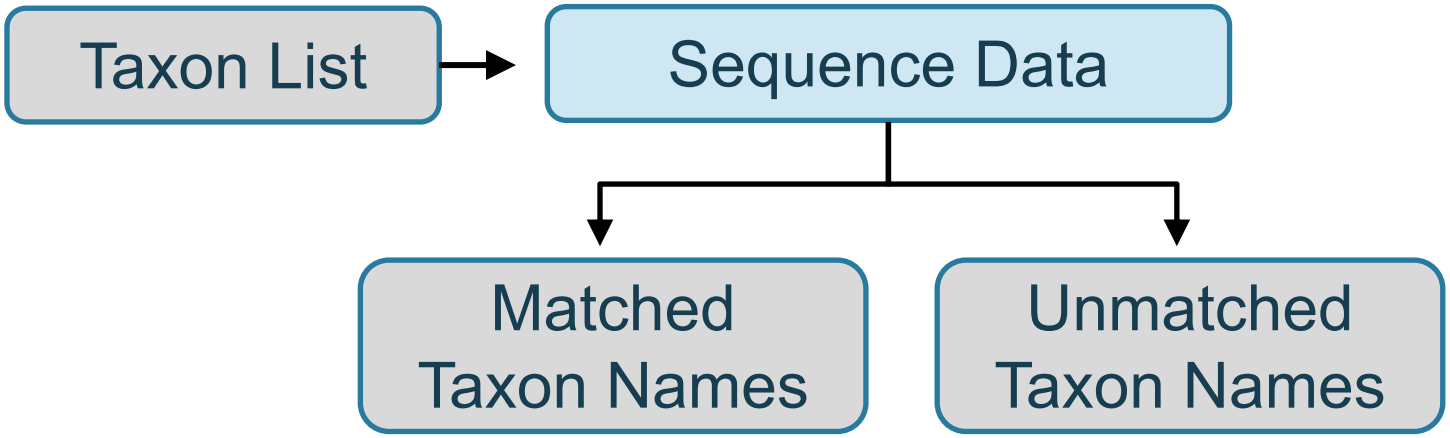
Gene and Taxon Filtering

Optional assessment and cleaning step



Gene and Taxon Filtering

Optional assessment and cleaning step

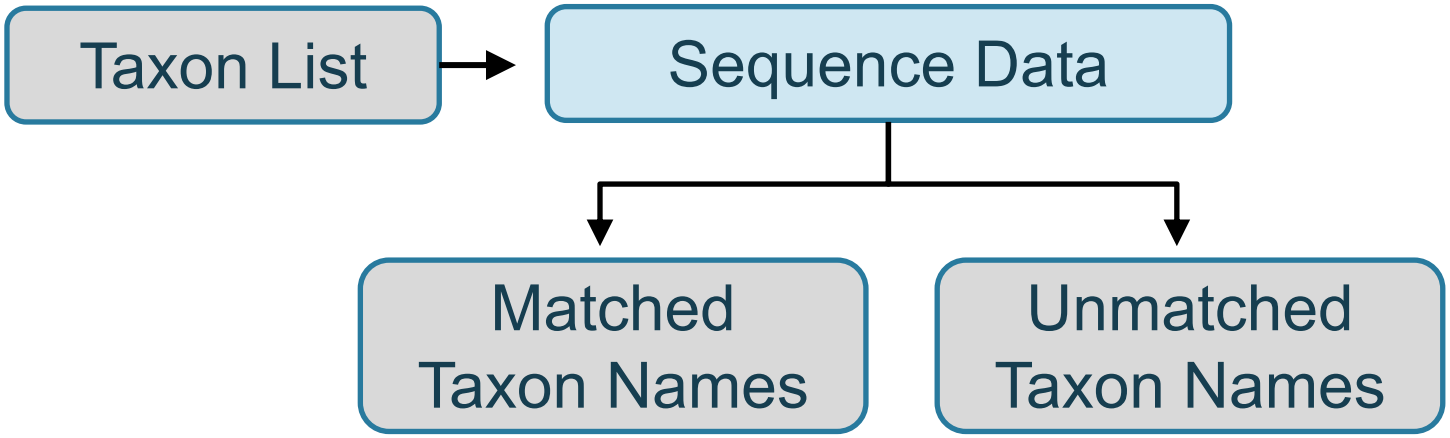


Bad Records

```
A.porcus mitochondrial
Calumma sp.
Jp 2016507218-a/5
Unverified calotes
```

Gene and Taxon Filtering

Optional assessment and cleaning step

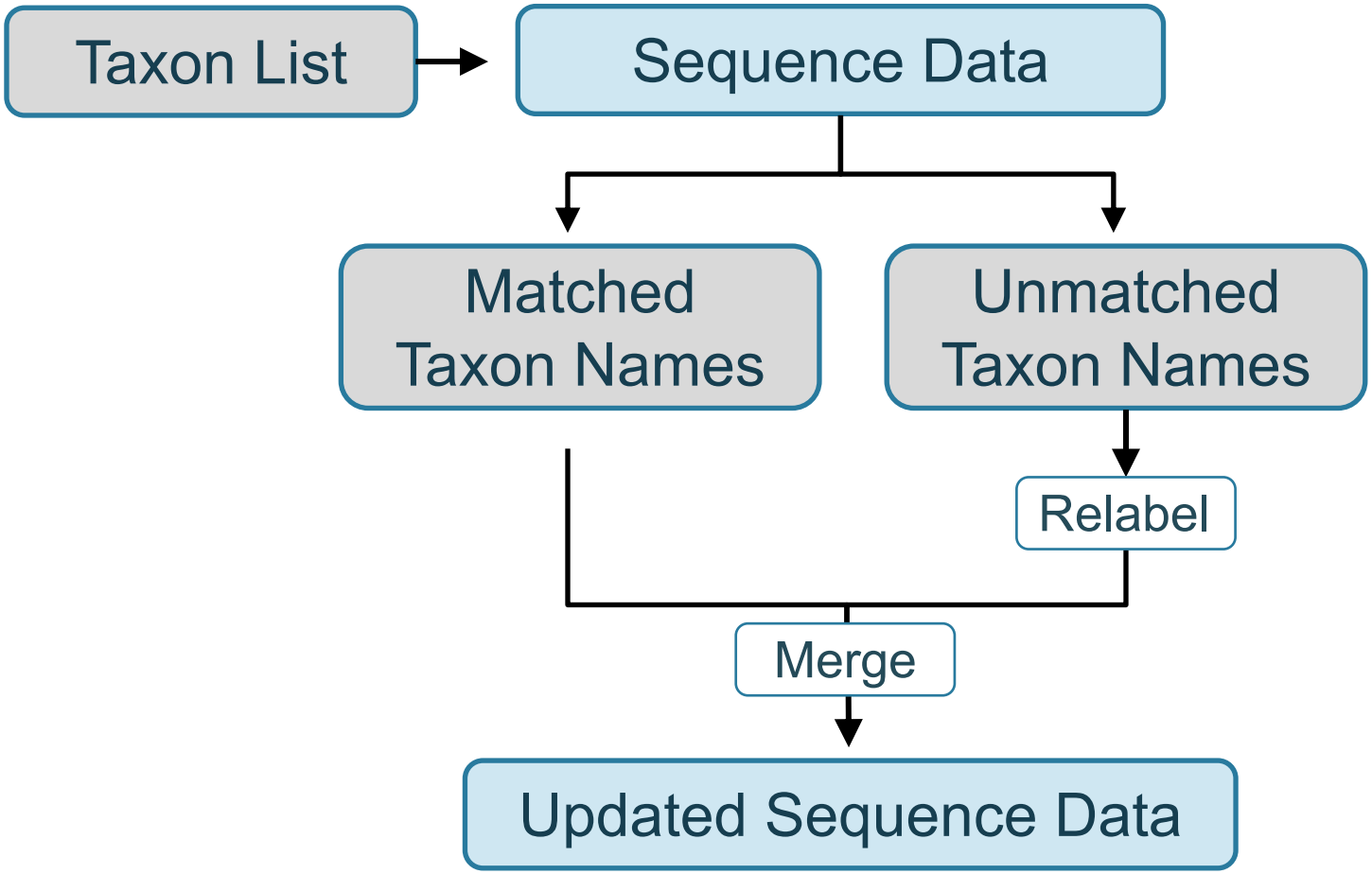


Good records, but need to be corrected

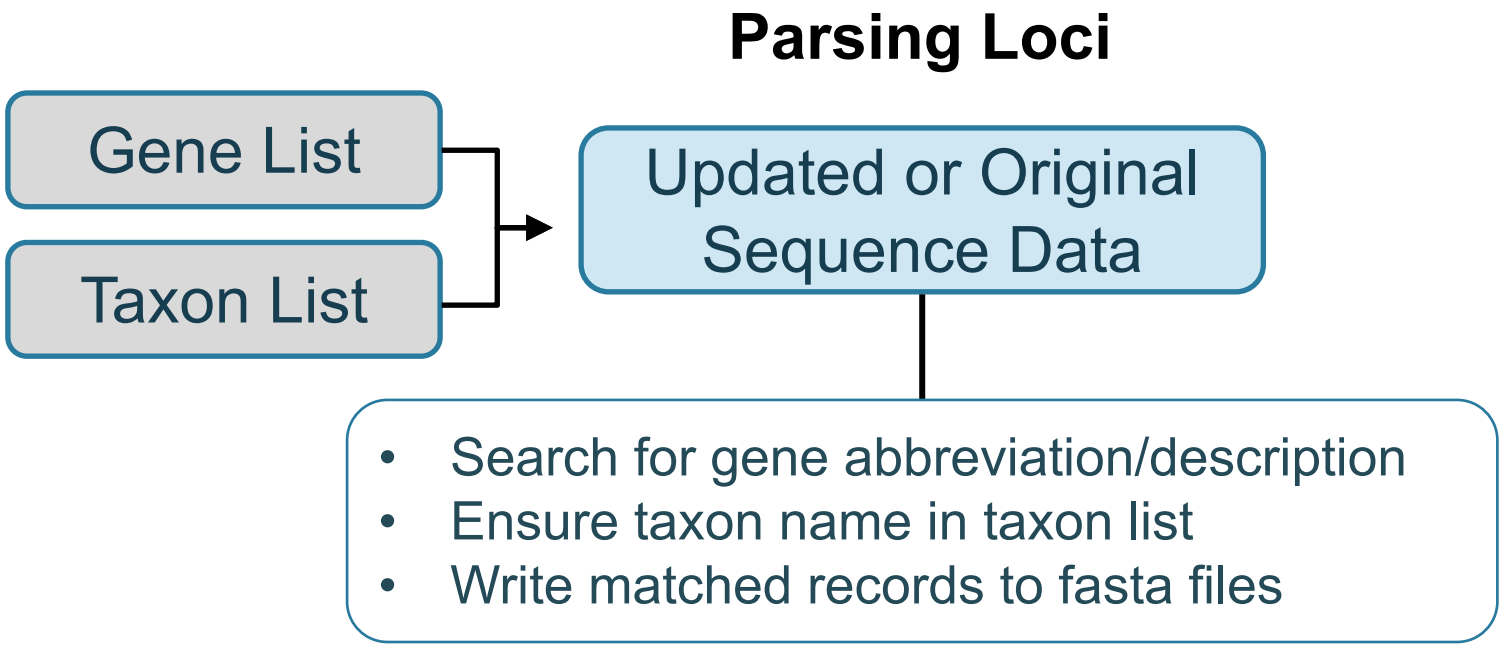
Uromastix acanthinura	→	Uromastyx acanthinura	Spelling errors
Causus defilippi	→	Causus defilippii	
Vipera raddei	→	Montivipera raddei	Change in taxonomy

Gene and Taxon Filtering

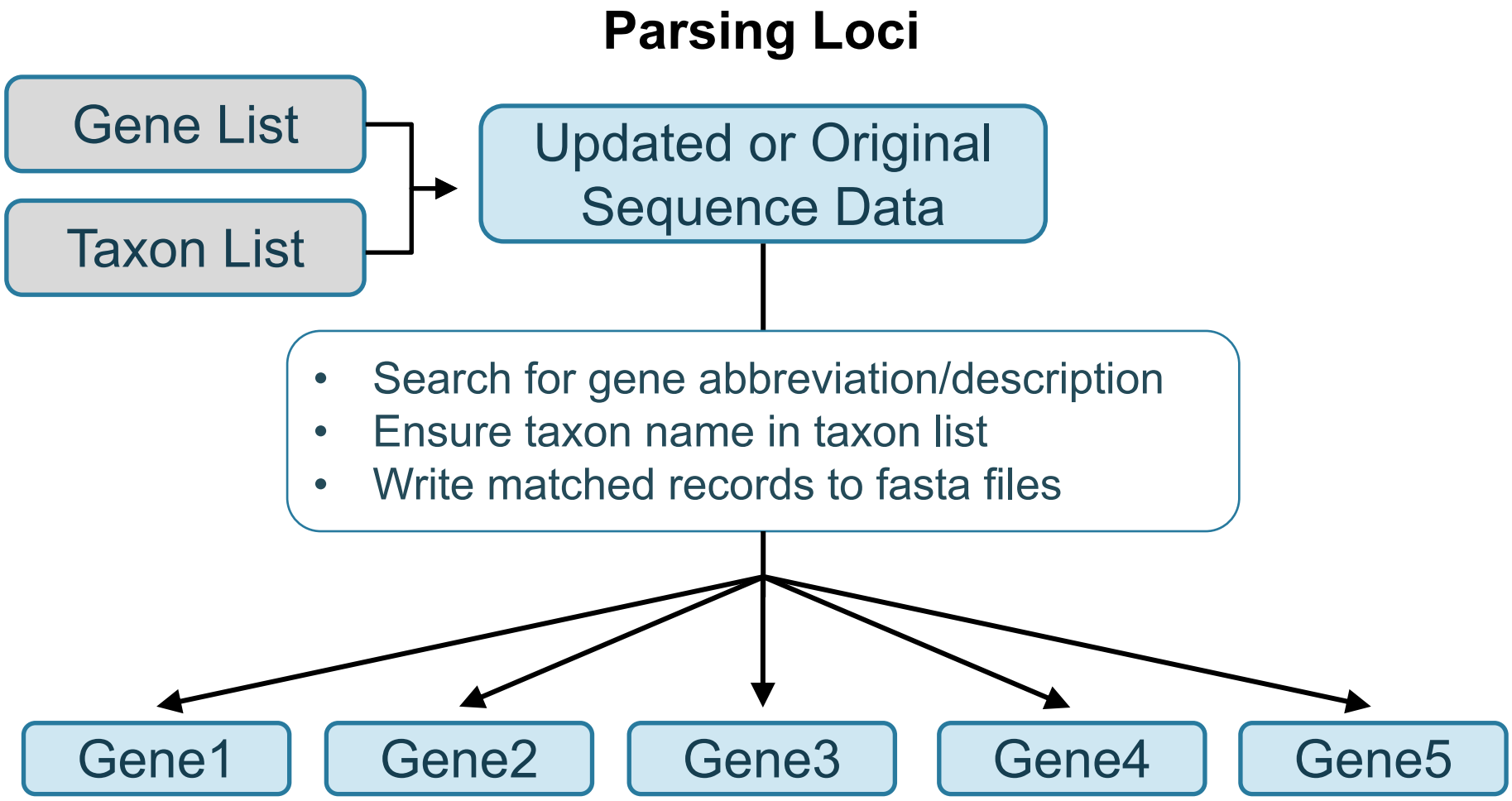
Optional assessment and cleaning step



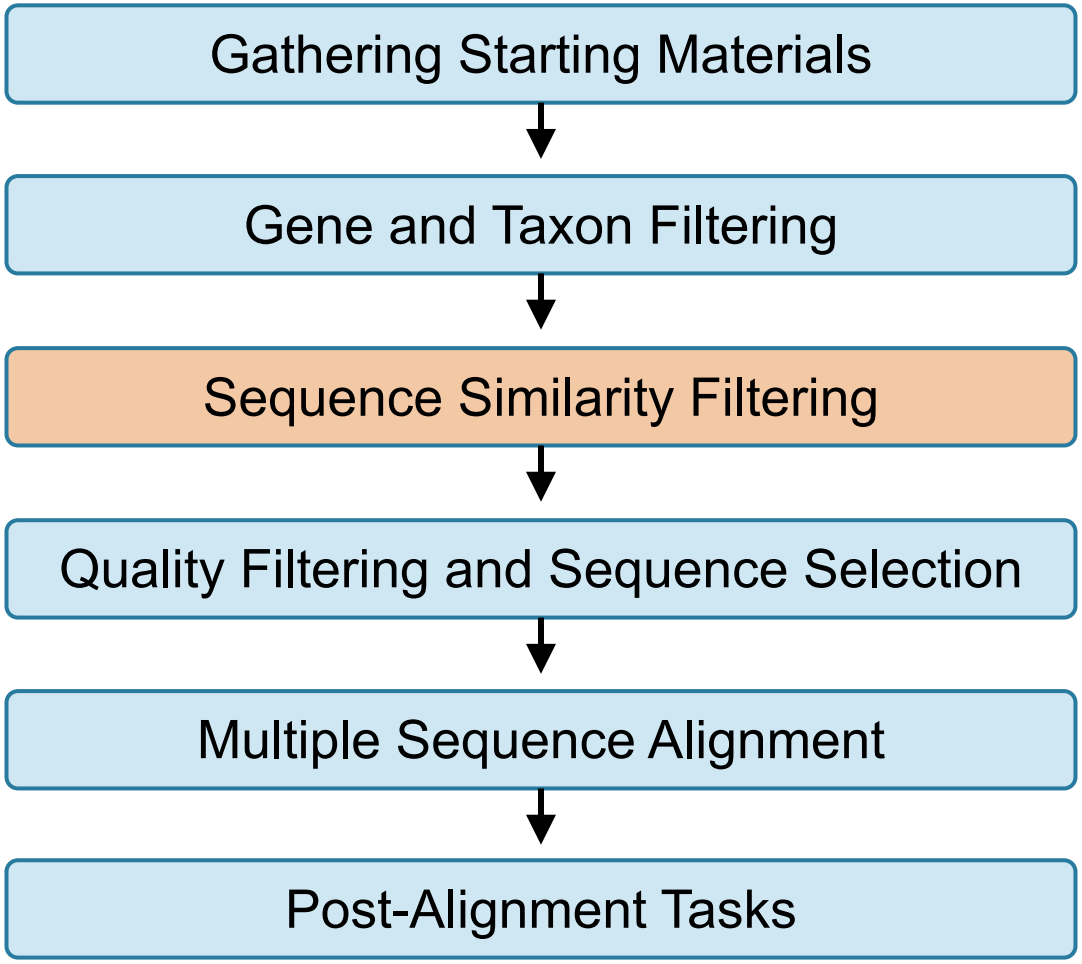
Gene and Taxon Filtering



Gene and Taxon Filtering



SuperCRUNCH Workflow

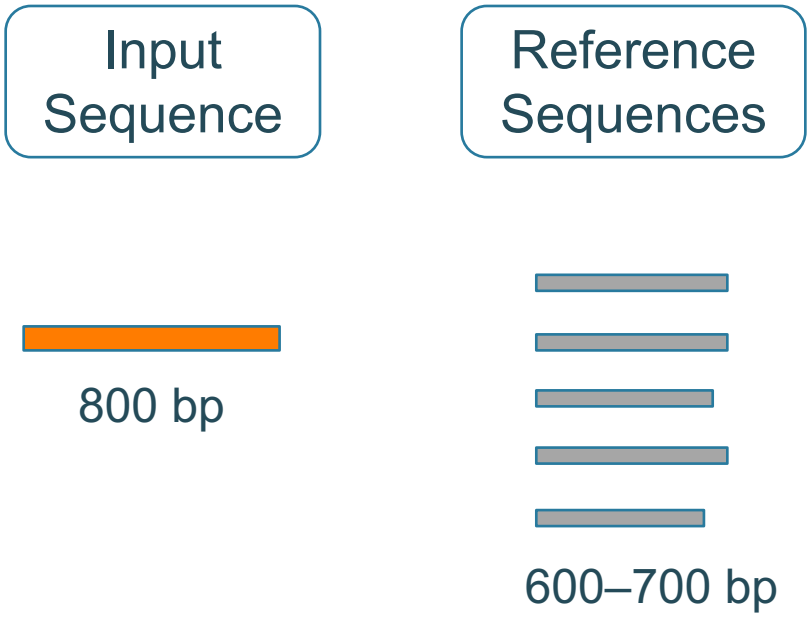


Sequence Similarity Filtering

Use BLASTn to identify and extract target sequence

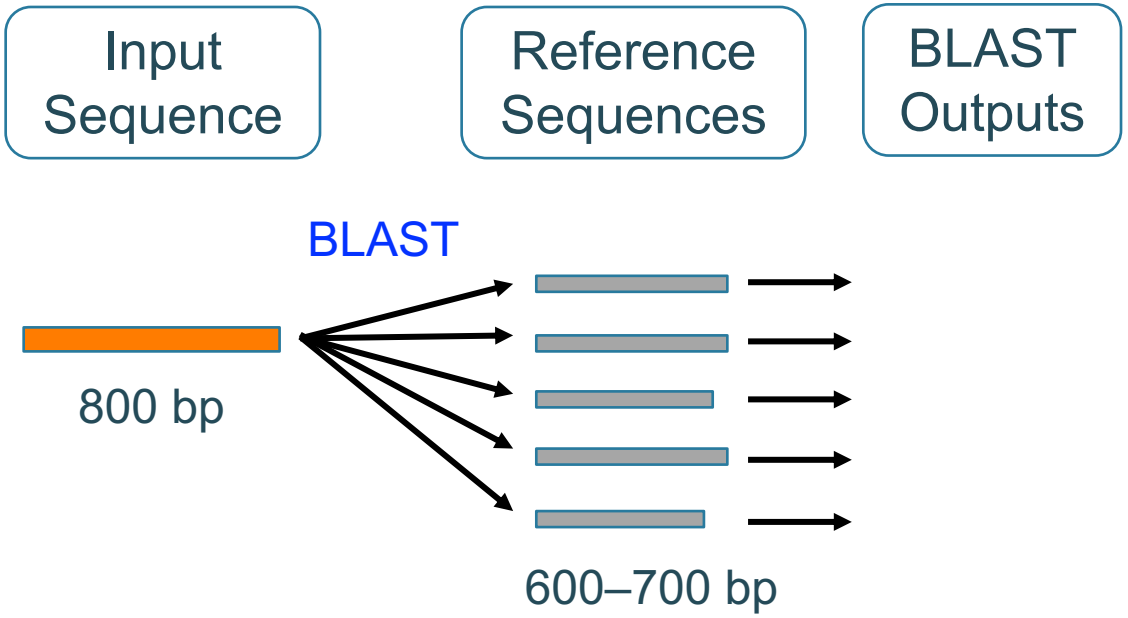
Sequence Similarity Filtering

Use BLASTn to identify and extract target sequence



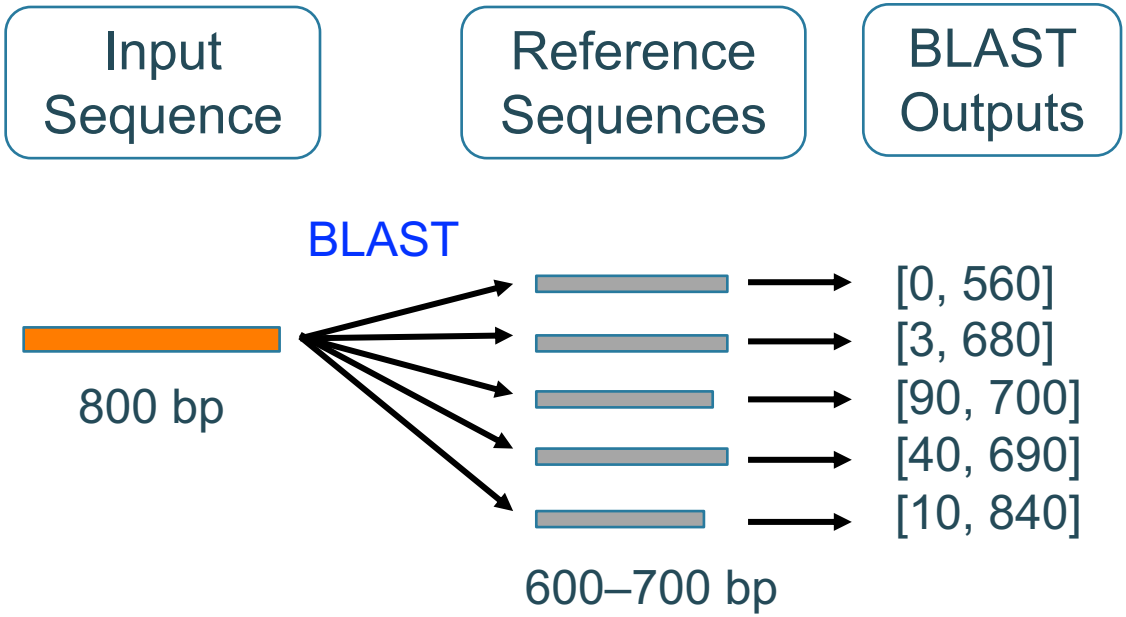
Sequence Similarity Filtering

Use BLASTn to identify and extract target sequence



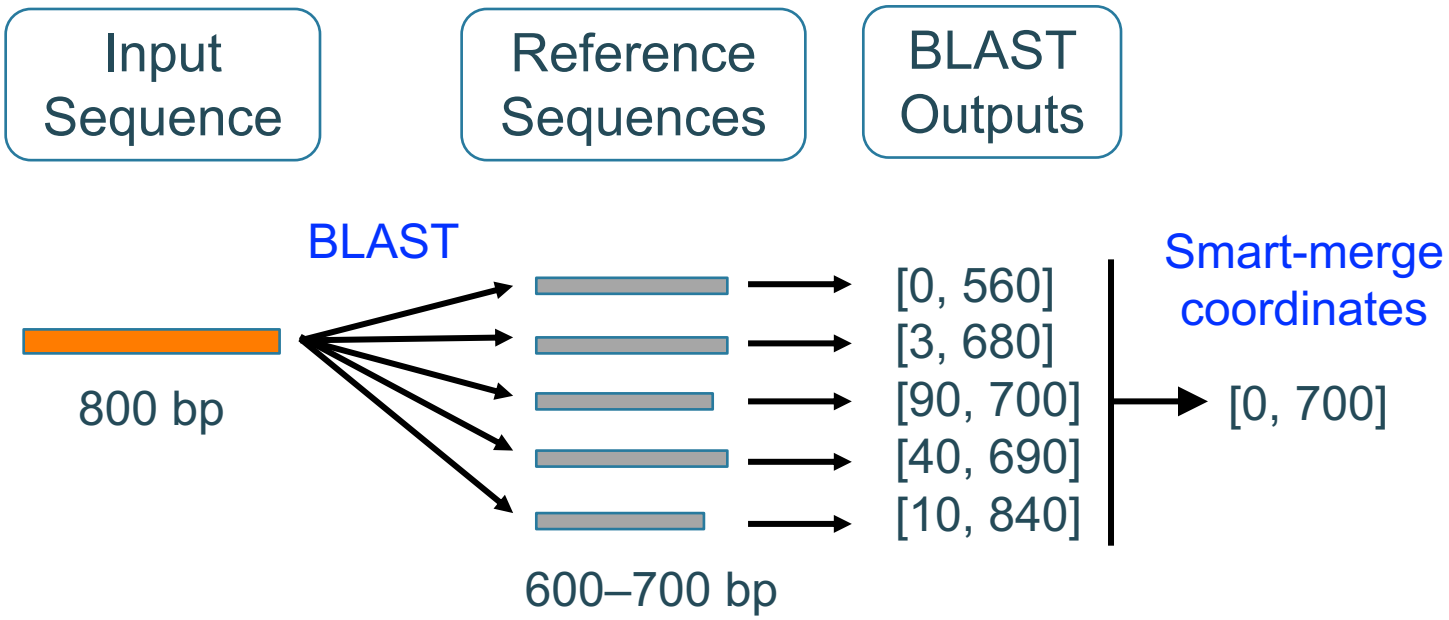
Sequence Similarity Filtering

Use BLASTn to identify and extract target sequence



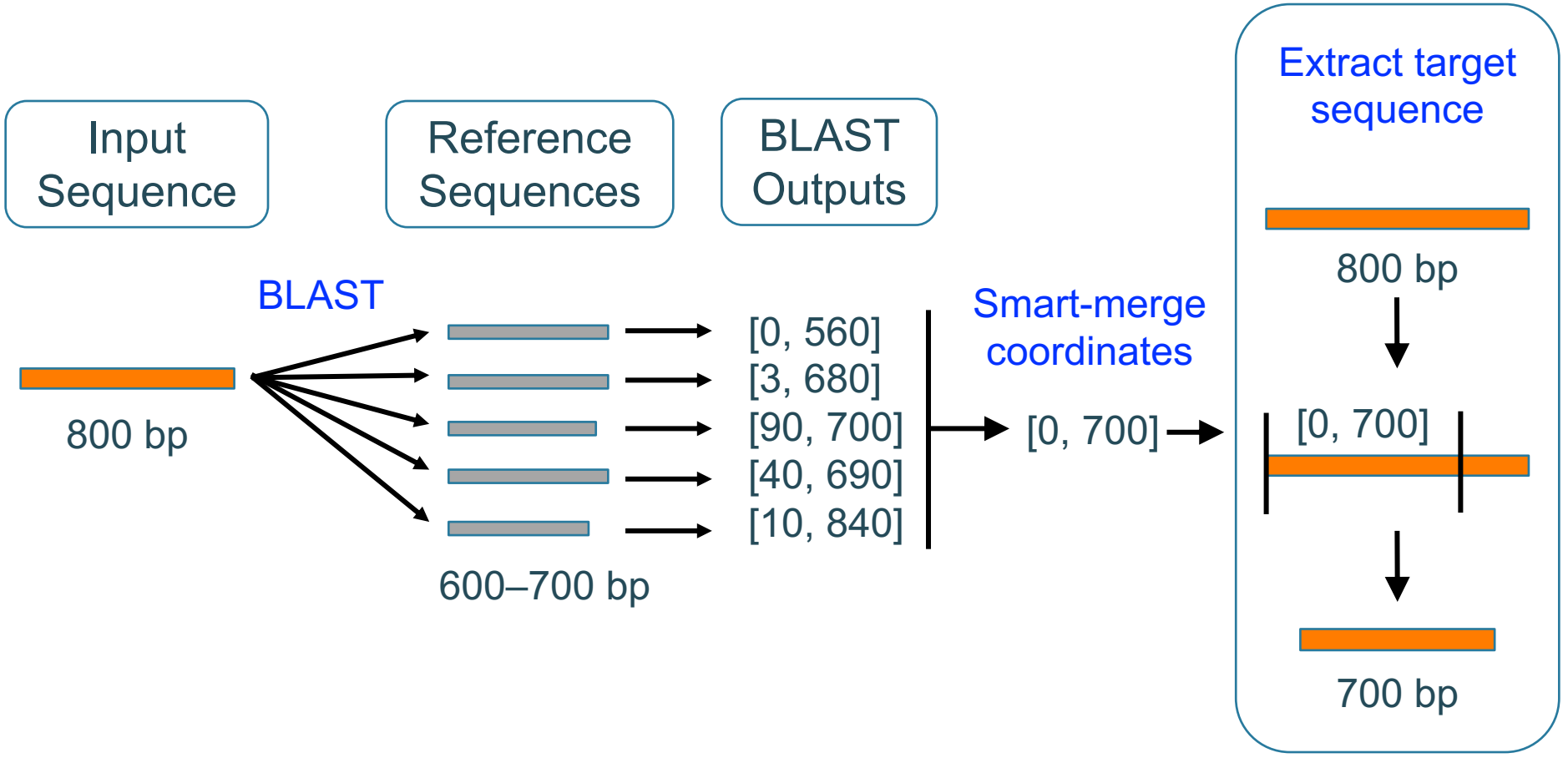
Sequence Similarity Filtering

Use BLASTn to identify and extract target sequence



Sequence Similarity Filtering

Use BLASTn to identify and extract target sequence



Sequence Similarity Filtering

Use BLASTn to identify and extract target sequence

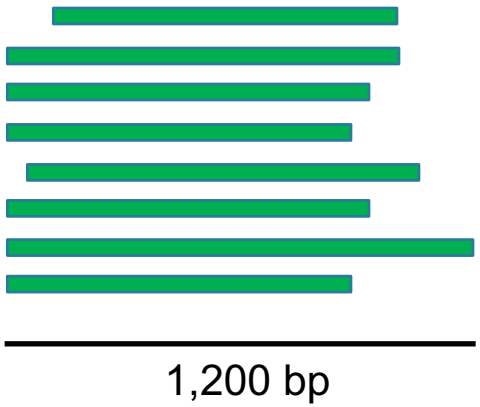
‘Simple’ Records

‘Complex’ Records

Sequence Similarity Filtering

Use BLASTn to identify and extract target sequence

‘Simple’ Records



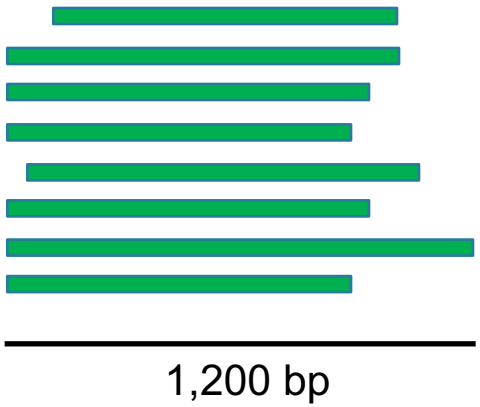
- Single gene
- Same region obtained
- Minor length variation

‘Complex’ Records

Sequence Similarity Filtering

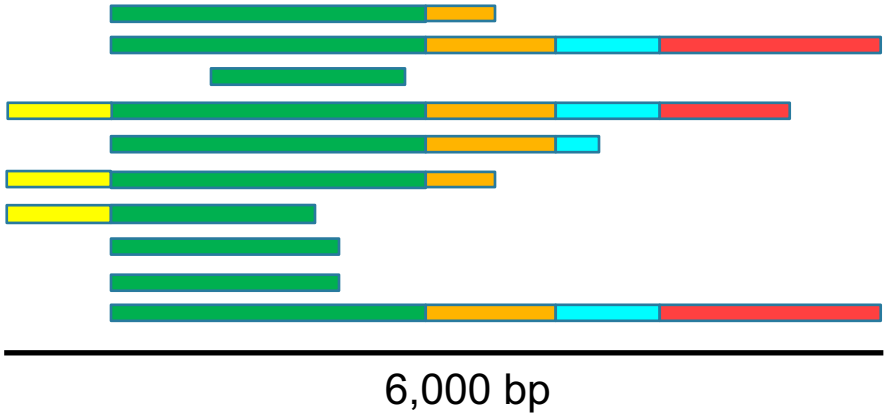
Use BLASTn to identify and extract target sequence

‘Simple’ Records



- Single gene
- Same region obtained
- Minor length variation

‘Complex’ Records

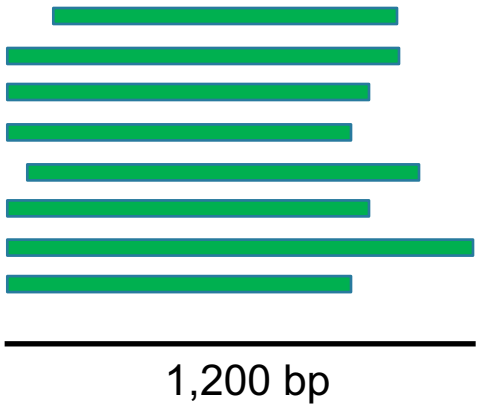


- Often multiple genes
- Variation in region obtained
- Major length variation

Sequence Similarity Filtering

Use BLASTn to identify and extract target sequence

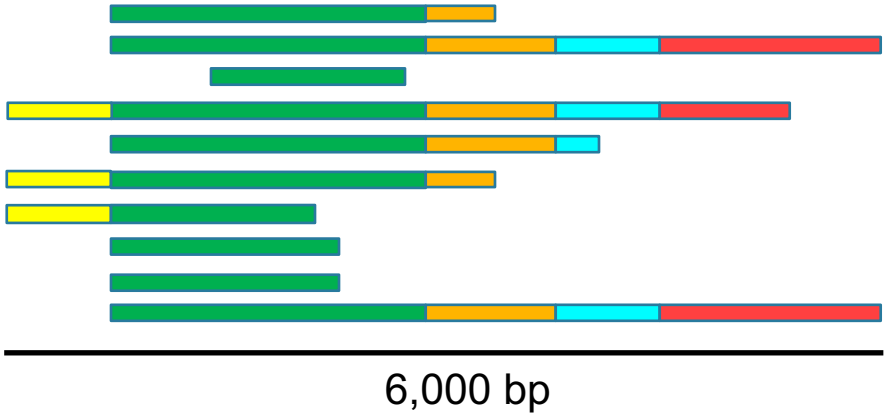
‘Simple’ Records



- Single gene
- Same region obtained
- Minor length variation

Automatic selection of reference sequences

‘Complex’ Records



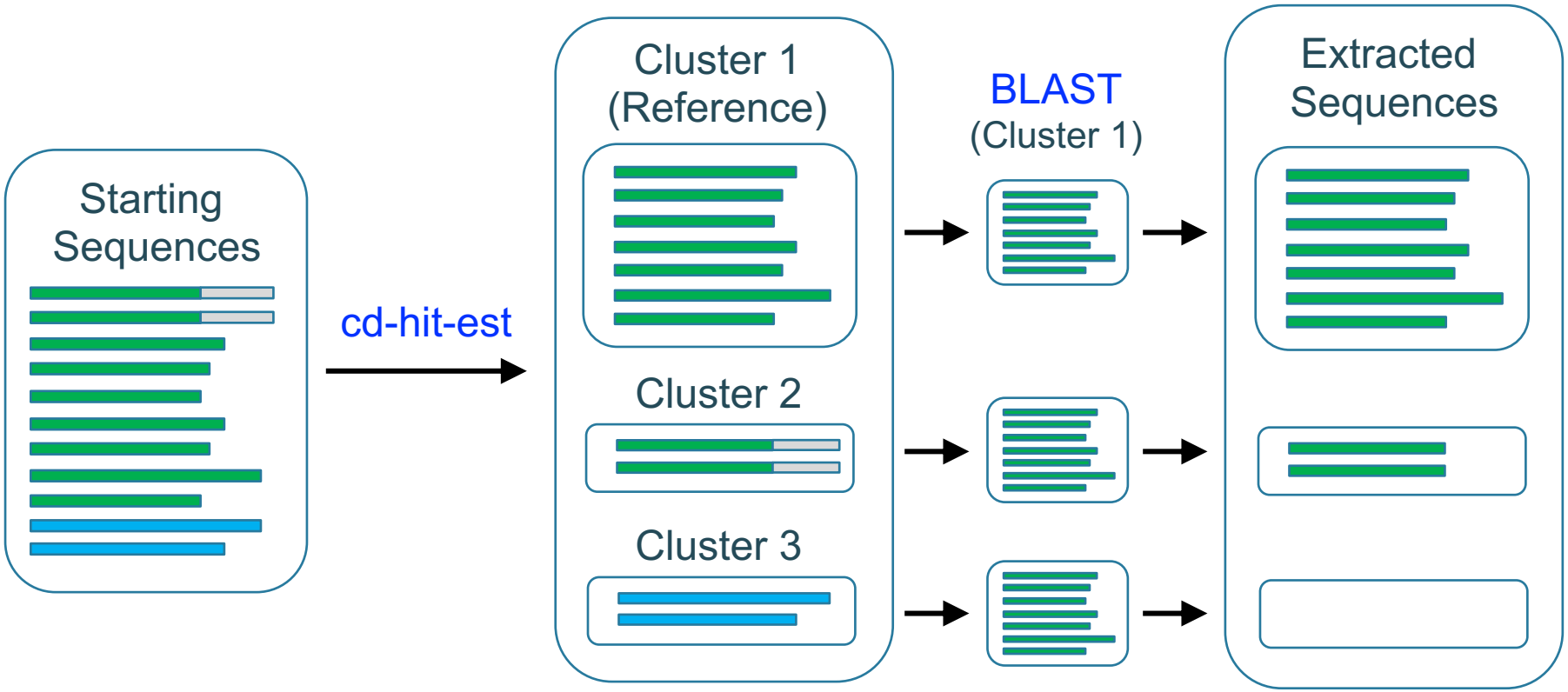
- Often multiple genes
- Variation in region obtained
- Major length variation

Requires a user-supplied reference sequences

Sequence Similarity Filtering

‘Simple’ Records

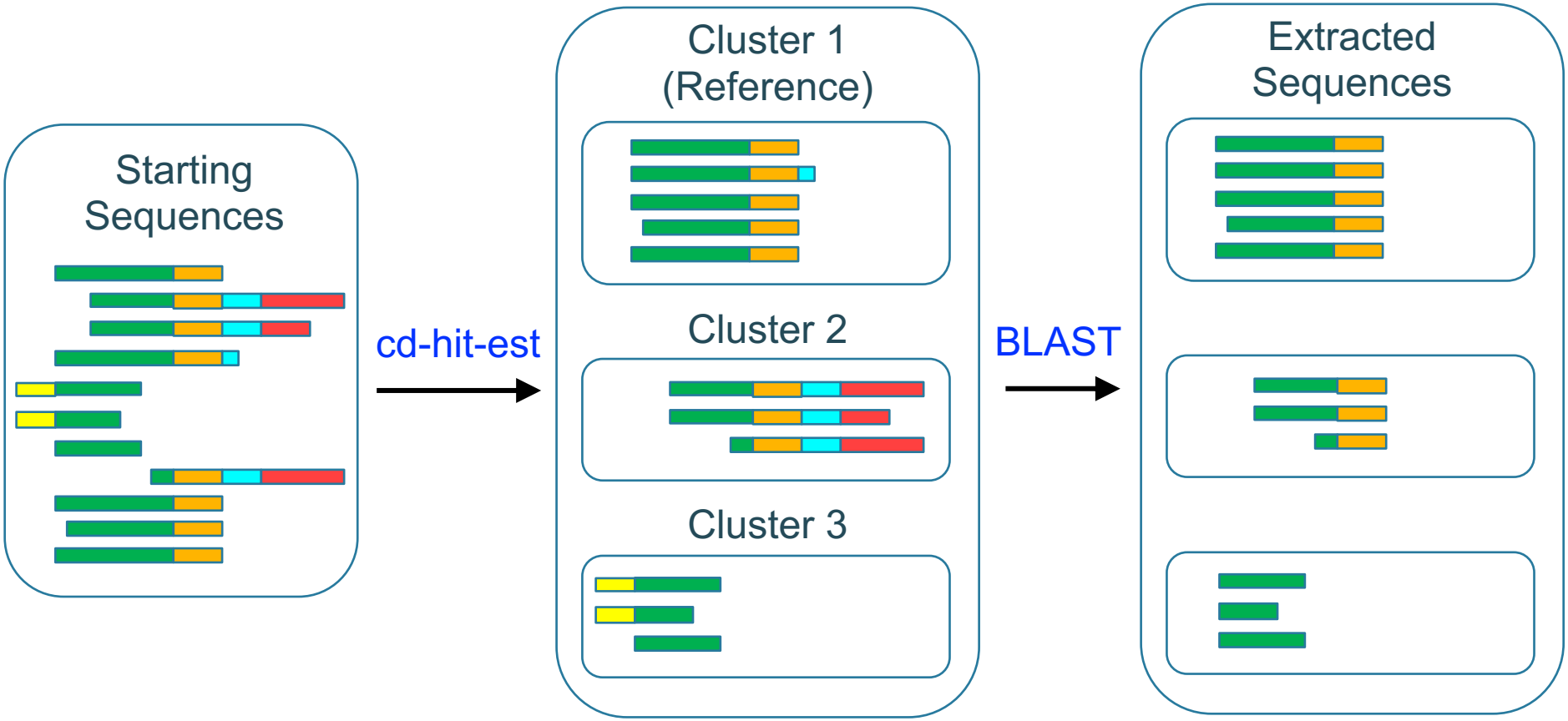
Automatic selection of reference sequences



Sequence Similarity Filtering

‘Complex’ Records

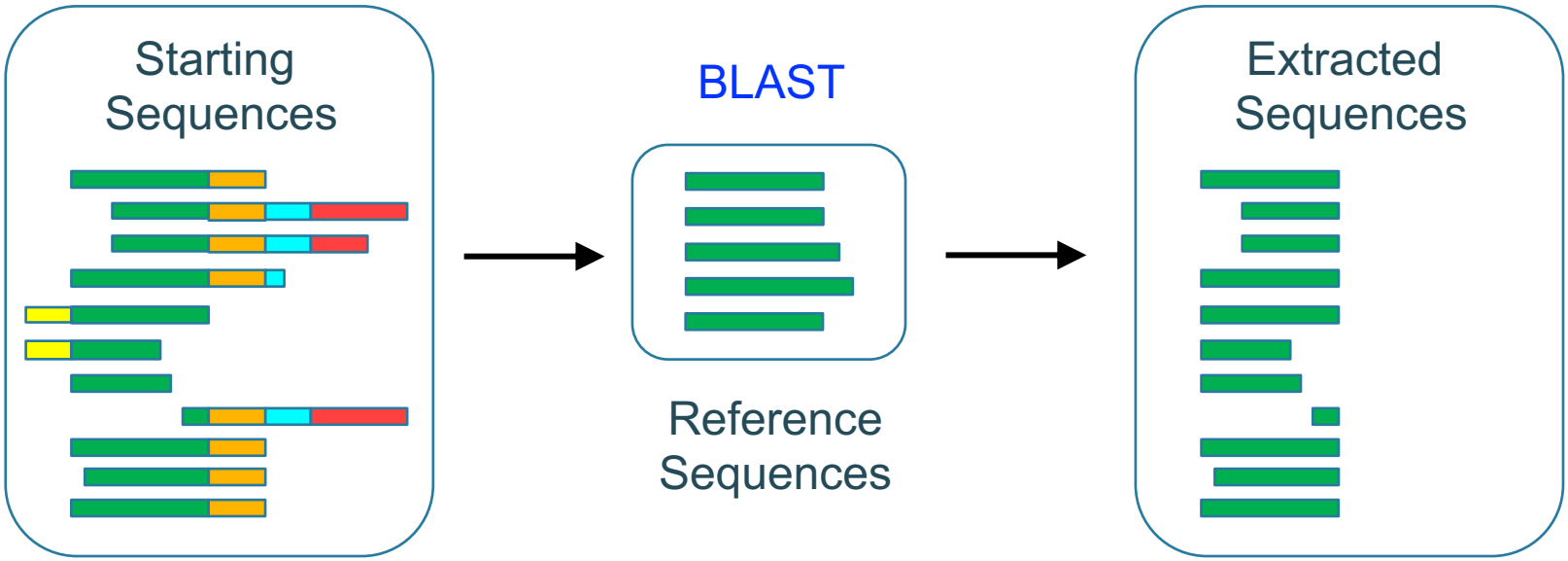
Automatic selection of reference sequences



Sequence Similarity Filtering

‘Complex’ Records

Supply validated reference sequences



Sequence Similarity Filtering

- Ability to specify BLAST algorithm for searches
  - blastn, megablast, dc-megablast

Sequence Similarity Filtering

- Ability to specify BLAST algorithm for searches
  - blastn, megablast, dc-megablast
- Detailed outputs provide complete transparency
  - sequences discarded, info for extracted sequences

Sequence Similarity Filtering

- Ability to specify BLAST algorithm for searches
  - blastn, megablast, dc-megablast
- Detailed outputs provide complete transparency
  - sequences discarded, info for extracted sequences

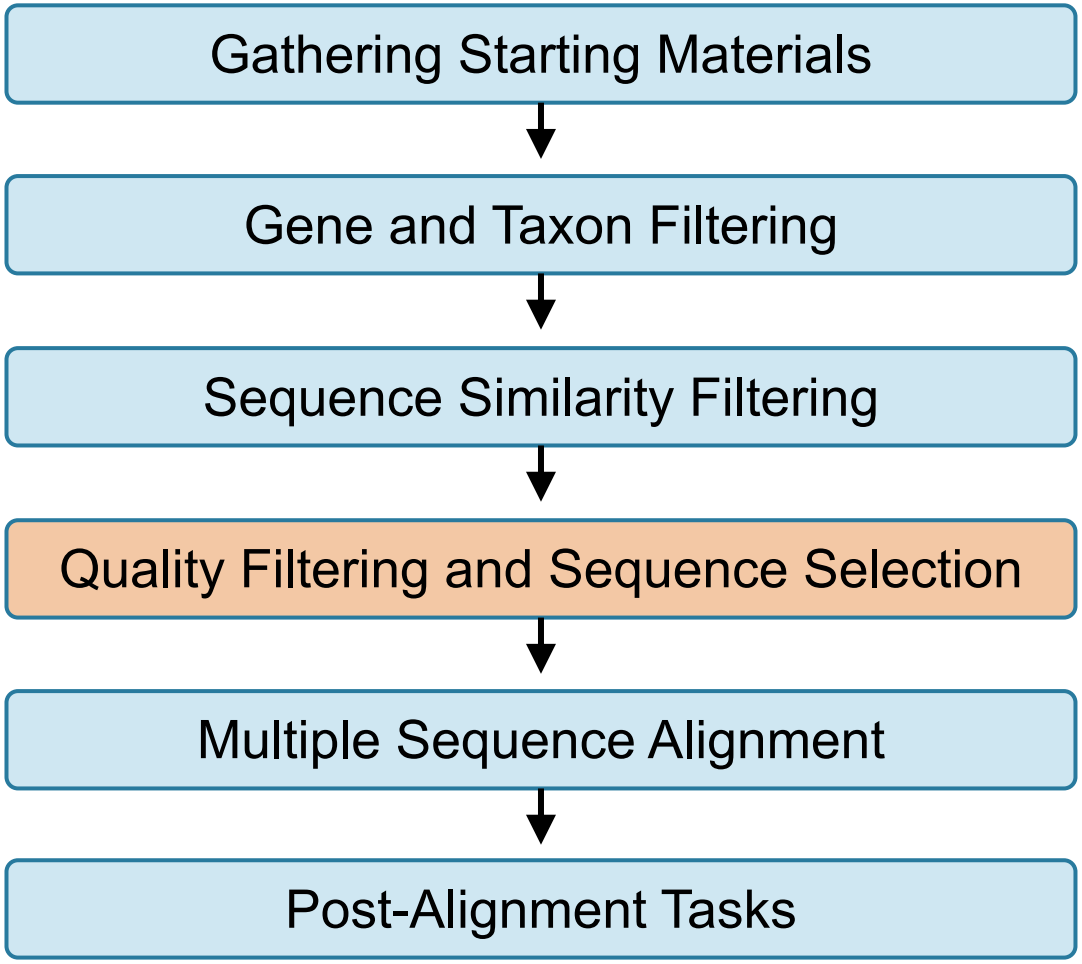
'Simple' Records

Record	Original Length	Extracted Length	Coordinates Used
GU432436.1	772	772	[0, 771]
GU432437.1	753	753	[0, 752]
GU432438.1	773	773	[0, 772]
GU432439.1	749	749	[0, 748]
JF818222.1	719	717	[2, 718]

'Complex' Records

Record	Original Length	Extracted Length	Coordinates Used
AM055651.1	510	510	[0, 509]
AM055653.1	508	508	[0, 507]
JF317635.1	17,388	1,575	[1,060, 2,634]
JF317636.1	17,344	1,574	[1,061, 2,634]
AF440085.1	3,700	1,473	[1,014, 2,486]
AF440019.1	3,693	1,470	[1,010, 2,479]
AF338325.1	1,071	41	[1,030, 1,070]
AF338329.1	1,085	84	[1,001, 1,084]

SuperCRUNCH Workflow



Quality Filtering and Sequence Selection

- What additional filters should be applied to sequences?
- If multiple sequences are available for a species, how do we choose?

Quality Filtering and Sequence Selection

- What additional filters should be applied to sequences?
- If multiple sequences are available for a species, how do we choose?

Example from Iguania dataset:

1,426 species | 66 loci | 58,642 sequences → 13,419 sequences

Quality Filtering and Sequence Selection

- What additional filters should be applied to sequences?
- If multiple sequences are available for a species, how do we choose?

Example from Iguania dataset:

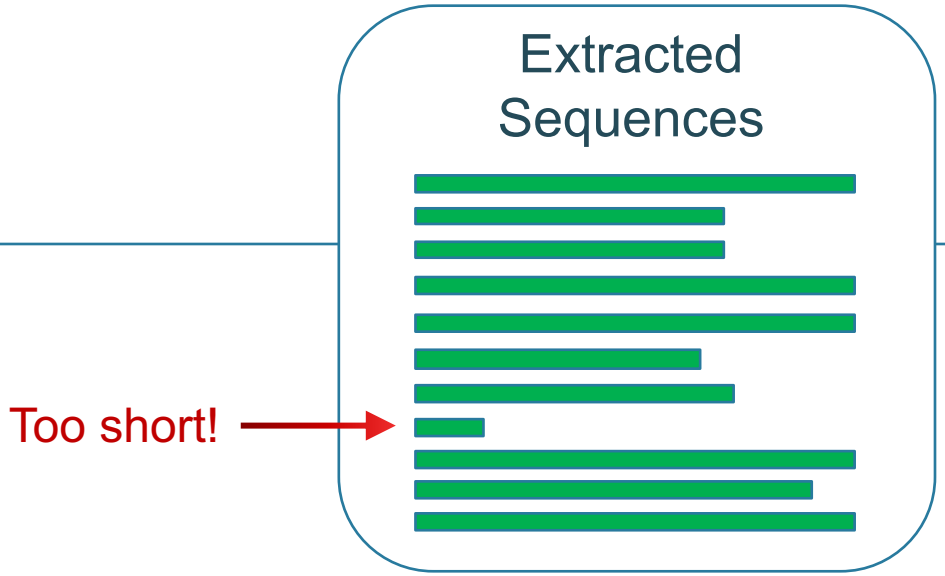
1,426 species | 66 loci | 58,642 sequences → 13,419 sequences

426,941,590,086,091,766,797,665,819,027,414,303,769,493,004,175,161,067,612,739,804,543,200,225,182,323,549,827,313,258,790,810,403,584,768,416,770,835,213,996,198,117,363,132,678,180,852,868,614,265,036,
460,700,461,471,274,375,628,253,781,270,736,717,213,100,620,869,318,857,262,839,230,802,209,013,310,881,128,210,163,684,848,705,688,760,829,587,224,314,529,722,798,841,605,227,315,801,162,497,423,940,091,
445,188,160,456,972,216,318,159,736,154,549,212,758,918,720,838,769,144,570,389,101,663,805,288,038,665,221,120,916,779,881,901,255,863,151,877,738,579,537,633,380,998,770,743,672,990,604,846,680,104,027,
167,115,116,206,452,903,240,810,448,638,340,423,170,159,510,377,667,838,928,644,400,326,345,259,425,132,817,446,024,738,134,949,167,451,283,577,159,233,975,368,043,845,814,970,551,368,290,958,170,688,277,
115,968,163,767,179,583,344,819,955,808,675,706,120,716,127,841,902,671,038,574,563,295,380,997,234,275,716,641,319,398,960,753,493,111,865,915,581,185,681,256,447,657,684,292,147,851,091,848,904,031,905,
387,435,249,180,033,266,268,639,262,607,030,240,852,855,093,777,997,576,264,609,615,265,815,897,849,810,922,612,539,919,156,687,177,799,986,699,532,506,211,815,639,620,162,600,713,718,532,134,078,981,010,
743,512,039,854,864,363,591,482,911,795,007,361,078,822,903,918,055,129,139,648,478,267,028,792,224,110,685,960,539,613,635,760,895,107,201,334,148,515,444,443,597,898,652,623,512,504,393,469,374,222,619,
268,592,490,285,076,675,903,171,062,614,102,4 … 7,018,473,530,880,562,167,592,704,337,066,145,
099,704,181,444,231,441,940,523,100,514,159,5 … 3,082,851,089,187,745,425,003,610,731,063,004,
313,660,517,535,733,182,220,861,208,603,283,8 … 5,495,812,832,004,410,010,381,734,498,921,043,
243,533,960,207,533,795,459,840,031,908,098,8 … 0,668,833,309,884,075,222,723,937,766,841,654,
763,378,019,946,453,740,819,825,184,254,388,2 … 5,185,909,718,356,128,563,576,741,832,452,217,
039,774,129,294,356,391,386,509,669,049,875,269,970,555,806,547,033,373,002,920,883,491,852,594,682,169,157,866,564,351,032,258,048,265,399,602,201,896,472,493,979,381,128,488,445,956,987,213,348,288,691,
570,345,518,215,045,551,657,443,773,109,440,951,716,740,650,626,381,855,714,027,966,269,256,541,819,875,874,148,285,355,711,249,446,998,328,263,555,431,922,877,103,912,603,707,237,294,457,126,237,335,120,
615,747,480,570,883,800,483,168,842,218,615,712,536,618,560,314,298,785,483,957,793,651,340,708,773,978,657,046,495,247,386,189,829,390,270,265,570,384,589,779,122,752,664,382,246,091,162,095,791,918,557,
333,940,793,101,639,289,334,201,951,985,795,613,687,153,555,254,560,644,705,211,775,735,050,699,748,624,983,085,112,994,257,905,508,442,819,445,795,885,787,172,418,157,918,550,456,444,746,358,297,678,622,
463,988,269,145,616,572,941,468,948,514,096,088,113,035,534,925,353,541,357,725,877,081,086,244,616,257,692,955,727,328,211,740,517,921,553,581,834,768,961,424,438,980,548,957,506,413,571,483,429,032,631,
248,265,538,814,640,371,209,808,007,080,820,517,248,513,765,513,852,972,727,410,094,462,697,676,765,921,280,000,000,000,000,000,000,000,000,000,000,000,000,000,000,000,000,000,000,000,000,000,000,000,000,
000,000,000,000,000,000,000,000,000,000,000,000,000,000,000,000,000,000,000,000,000,000,000,000,000,000,000,000,000,000,000,000,000,000,000,000,000,000,000,000,000,000,000,000,000,000,000,000,000,000,000,
000,000,000,000,000,000,000,000,000,000,000,000,000,000,000,000,000,000,000,000,000,000,000,000,000,000,000,000,000,000,000,000,000,000,000,000,000,000,000,000,000,000,000,000,000,000,000,000,000,000,000,
000,000,000,000,000,000,000,000,000,000,000,000,000,000,000,000,000,000,000,000,000,000,000,000,000,000,000,000,000,000,000,000

Combinations: 4.27×10^{3155}

Quality Filtering and Sequence Selection

- What additional filters should be applied to sequences?
  - Minimum length (defined by user)
- If multiple sequences are available for a species, how do we choose?



Quality Filtering and Sequence Selection

- What additional filters should be applied to sequences?
  - Minimum length (defined by user)
- If multiple sequences are available for a species, how do we choose?
  - 1) Sort by length, select longest sequence
  - 2) Sort by length, test translation (for protein-coding)
  - 3) Random

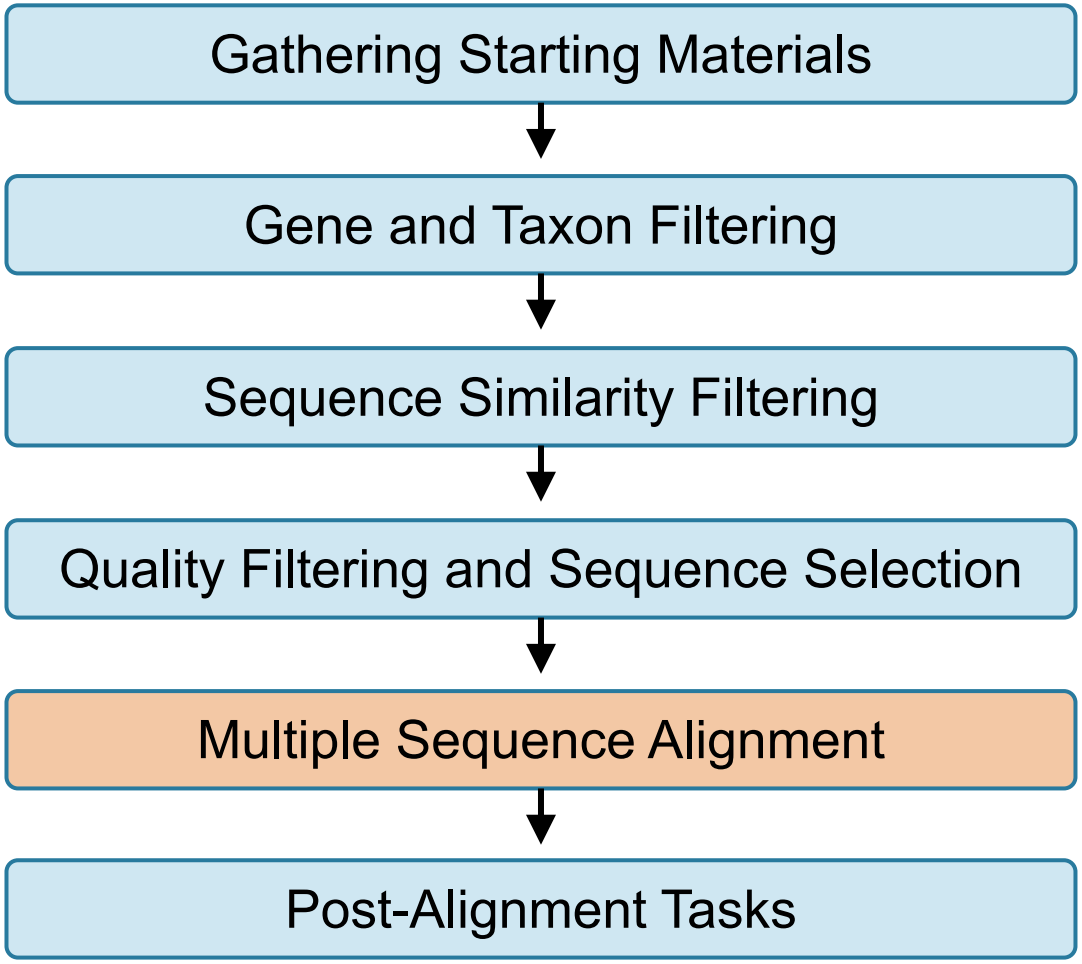
Quality Filtering and Sequence Selection

- What additional filters should be applied to sequences?
  - Minimum length (defined by user)
- If multiple sequences are available for a species, how do we choose?
  - 1) Sort by length, select longest sequence
  - 2) Sort by length, test translation (for protein-coding)
  - 3) Random

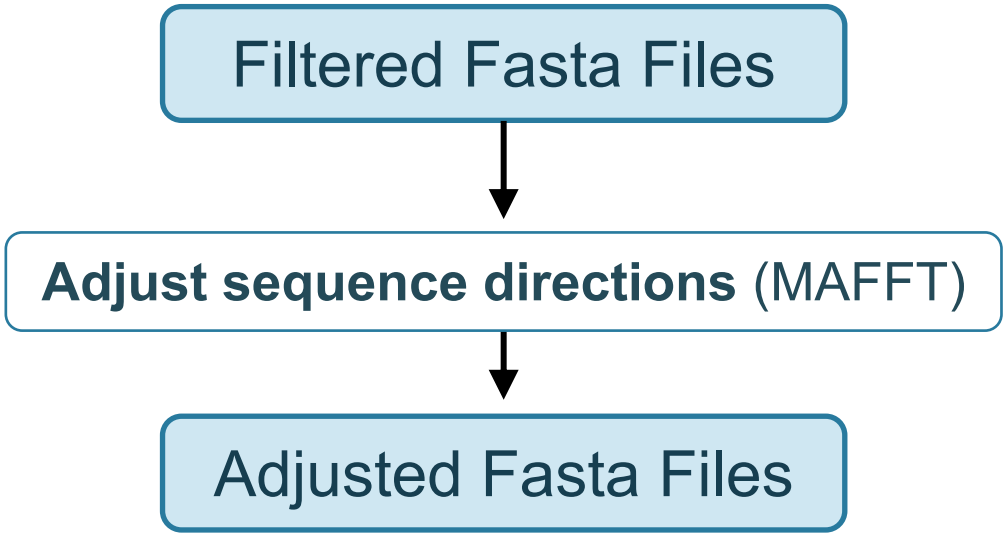
Species-level Data — **One sequence per species per gene**

Population-level Data — **All sequences per species per gene**

SuperCRUNCH Workflow

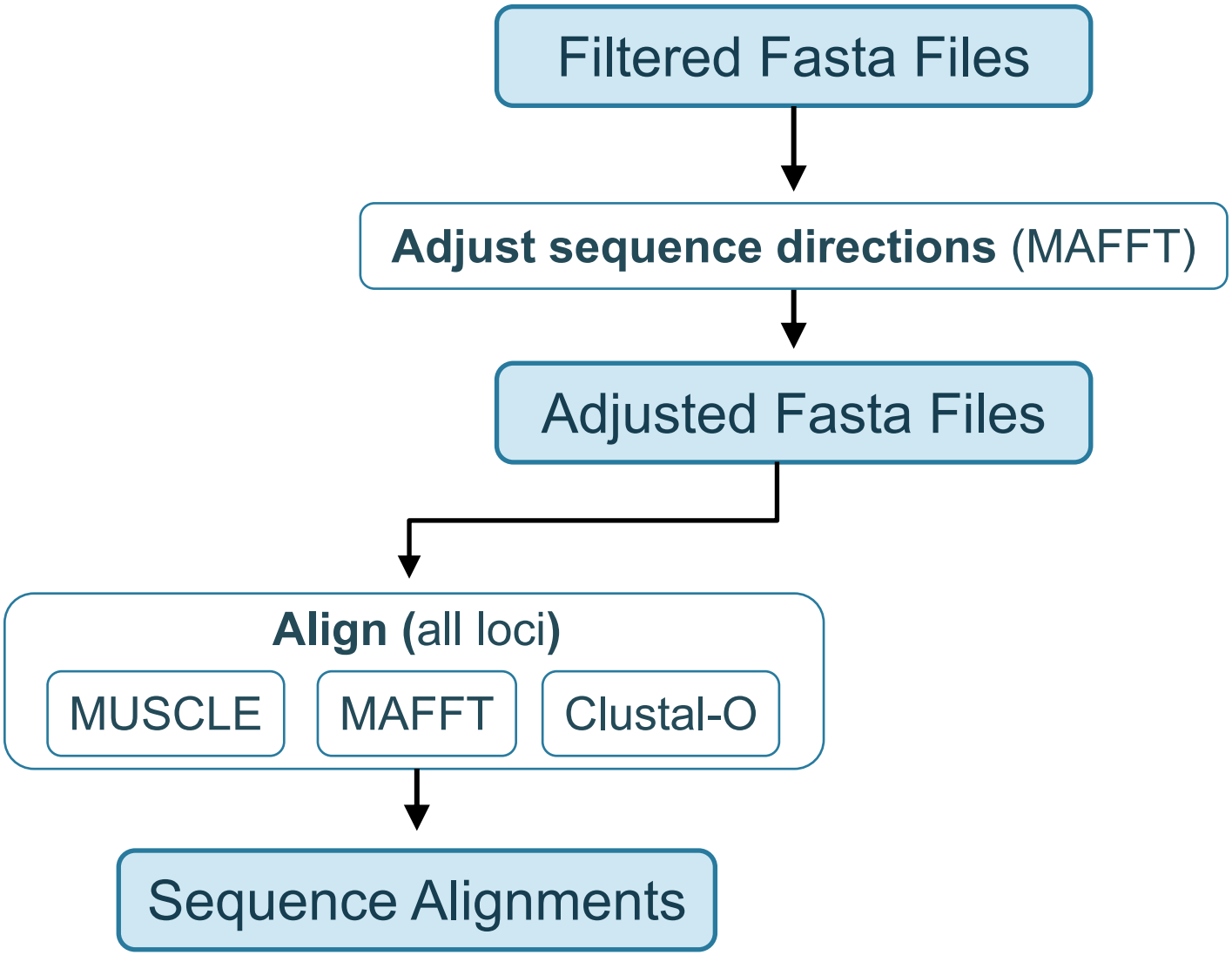


Multiple Sequence Alignment

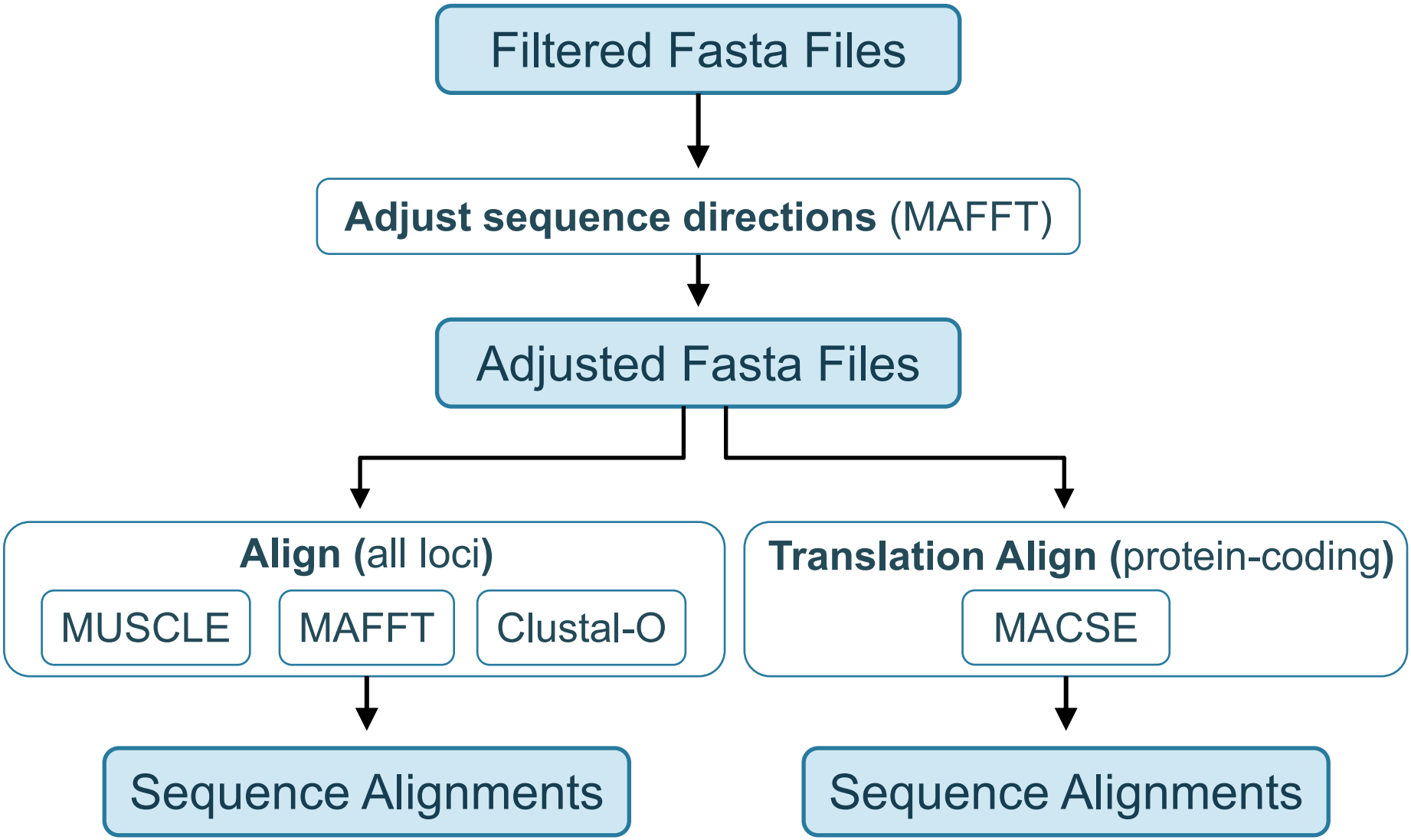


Locus	Correct Direction	Direction Adjusted
UCE-1003	66	5
UCE-1005	62	5
UCE-1012	57	11
UCE-1013	15	4

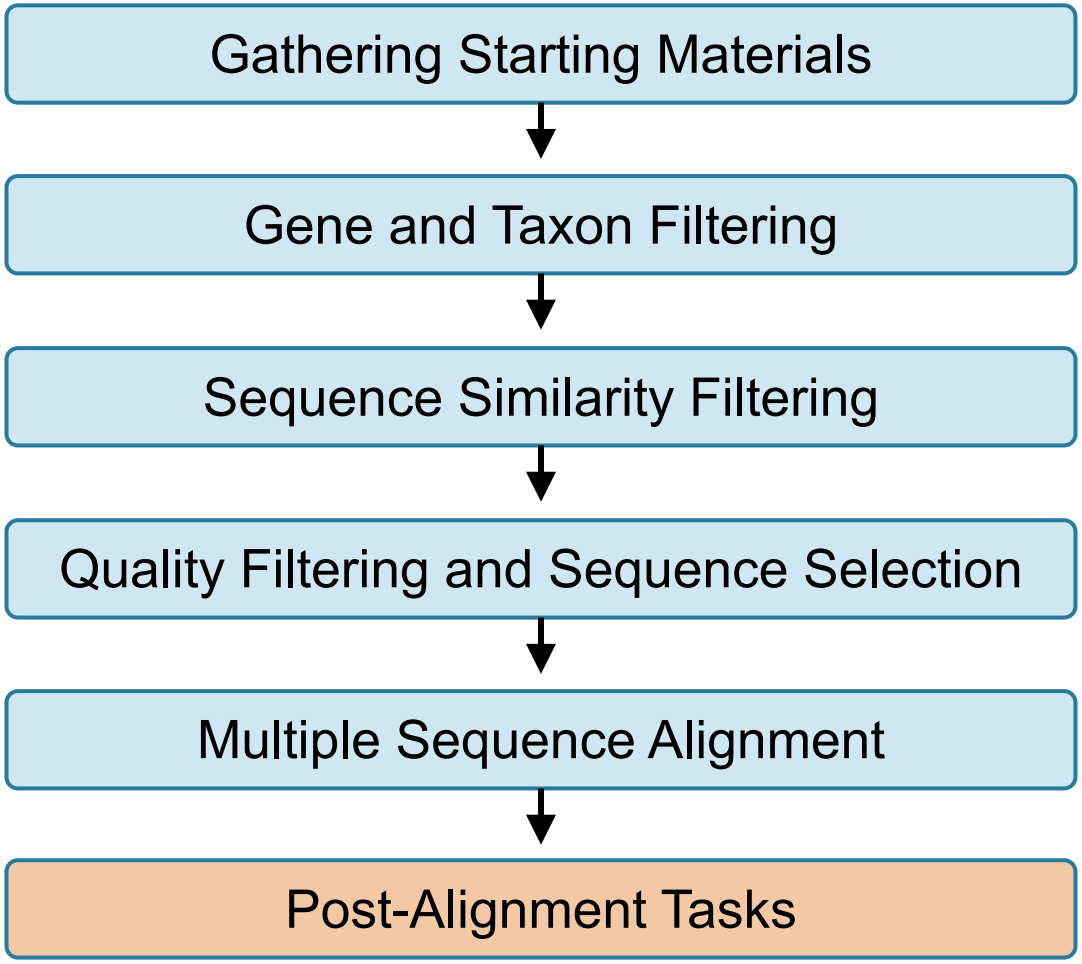
Multiple Sequence Alignment



Multiple Sequence Alignment



SuperCRUNCH Workflow



Post-Alignment Tasks

- Relabel sequence records

Post-Alignment Tasks

- Relabel sequence records

Original Labels

```
>JN881132.1 Daboia russelii activity-dependent neuroprotector...

>KU765220.1 Sceloporus undulatus voucher ADL182 activity-depe...
```

Post-Alignment Tasks

- Relabel sequence records

Original Labels

```
>JN881132.1 Daboia russelii activity-dependent neuroprotector...

>KU765220.1 Sceloporus undulatus voucher ADL182 activity-depe...
```

Species

```
>Daboia_russelii

>Sceloporus_undulatus
```

Post-Alignment Tasks

- Relabel sequence records

Original Labels

```
>JN881132.1 Daboia russelii activity-dependent neuroprotector...

>KU765220.1 Sceloporus undulatus voucher ADL182 activity-depe...
```

Species

```
>Daboia_russelii

>Sceloporus_undulatus
```

Accession

```
>JN881132.1

>KU765220.1
```

Post-Alignment Tasks

- Relabel sequence records

Original Labels

```
>JN881132.1 Daboia russelii activity-dependent neuroprotector...

>KU765220.1 Sceloporus undulatus voucher ADL182 activity-depe...
```

Species

```
>Daboia_russelii

>Sceloporus_undulatus
```

Accession

```
>JN881132.1

>KU765220.1
```

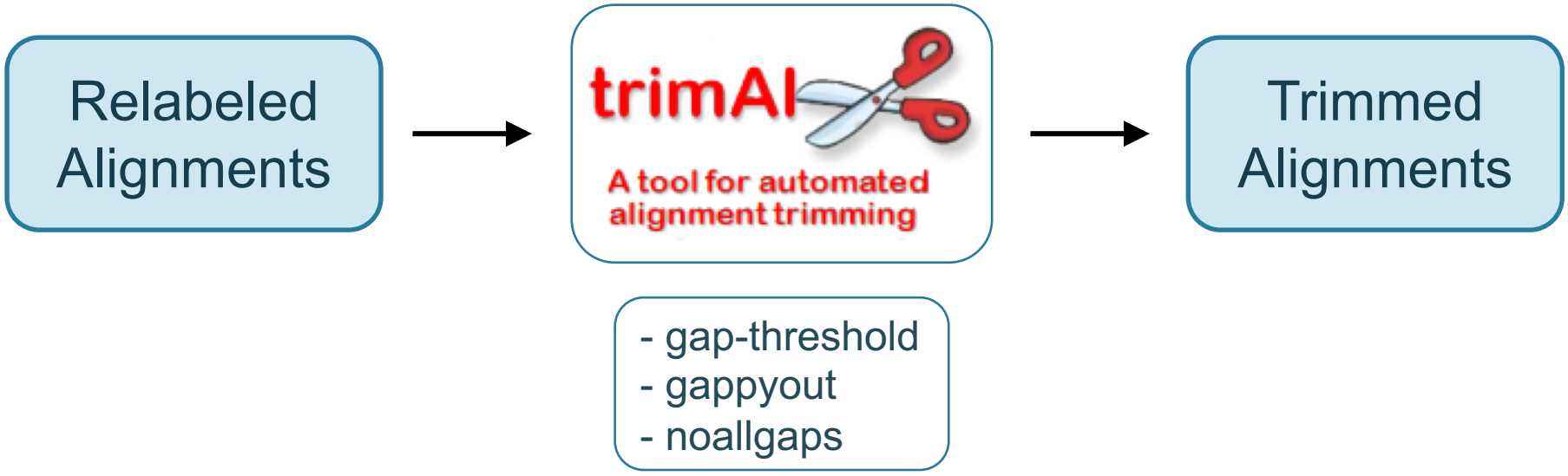
Species + Accession

```
>Daboia_russelii_JN881132.1

>Sceloporus_undulatus_KU765220.1
```

Post-Alignment Tasks

- Relabel sequence records
- Trim alignments



Post-Alignment Tasks

- Relabel sequence records
- Trim alignments
- Convert to nexus, phylip format

Post-Alignment Tasks

- Relabel sequence records
- Trim alignments
- Convert to nexus, phylip format
- Concatenate alignments

Requires Species Labeling

```
>Daboia_russelii

>Sceloporus_undulatus
```

Missing Data Options

N - ?

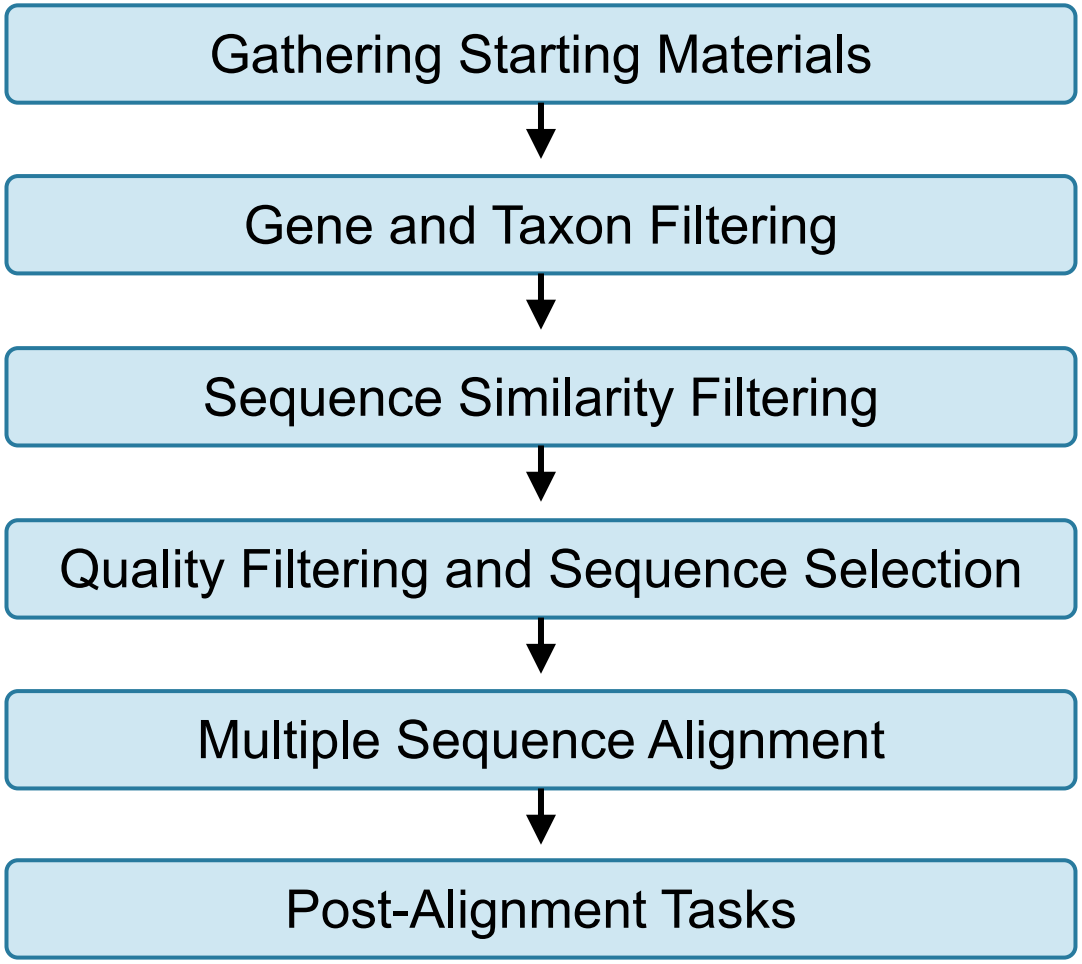
Outputs

Fasta or phylip

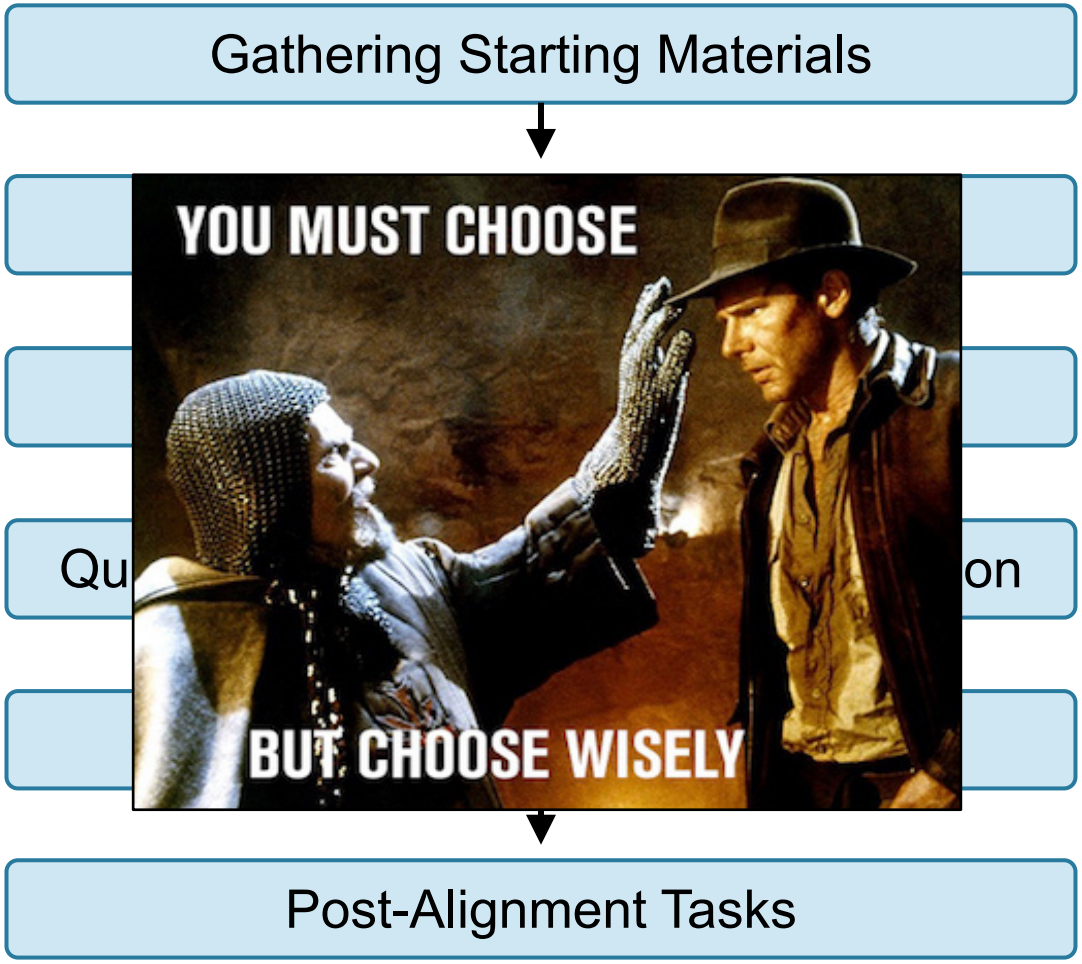
Partitions file

Loci per taxon

SuperCRUNCH Workflow



SuperCRUNCH Workflow



Benchmarking: Time

- Dataset dependent, but reasonable!
- Bottleneck is always sequence alignment

	Population level	Traditional Supermatrix	Genomic (UCE's)
File Size	52 MB	13.3 GB	205 MB
Sequences	82,557	8,785,378	338,942
Taxa	3	1,426	123
Genes	4	70	5,041
Total time	**~2.5 min**	**~4.5 hours**	**~10.5 hours**

Benchmarking: Other Programs

Comparison to PyPHLAWD

- PyPHLAWD uses GenBank release database and Taxonomy Browser to fetch starting sequences
- Default is automated clustering of all sequences (requires follow-up identification and processing)
- A ‘baited analysis’ can target specific loci, requires reference sequences

Benchmarking: Other Programs

Comparison to PyPHLAWD

- Compared supermatrix construction for Iguania, large clade of squamate reptiles (~1900 species)

- Targeted 68 loci
  - 61 nuclear – used automatic reference selection
  - 7 mtDNA loci – used validated reference sequences

Three analyses:

SuperCRUNCH NCBI direct-download	**SuperCRUNCH** PyPHLAWD database	**PyPHLAWD** PyPHLAWD database

Benchmarking: Other Programs

Comparison to PyPHLAWD

	Loci	Taxa	Sequences
SuperCRUNCH NCBI direct-download	66	1,426	13,419
SuperCRUNCH PyPHLAWD database	66	1,358	12,654
PyPHLAWD PyPHLAWD database	65	1,069	10,397

Benchmarking: Other Programs

- PyPHLAWD was comparable for nuclear loci
- PyPHLAWD failed to find 62% (2,125) of mtDNA seqs

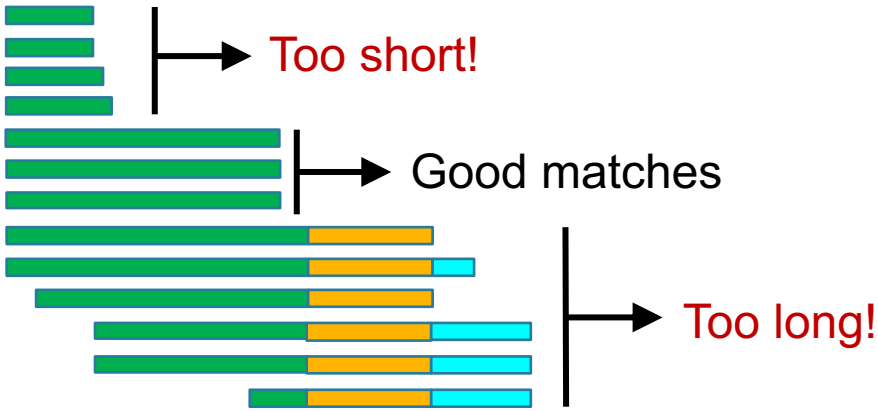
Benchmarking: Other Programs

- PyPHLAWD was comparable for nuclear loci
- PyPHLAWD failed to find 62% (2,125) of mtDNA seqs

References

Sequences must be highly similar to bait to pass filter

Entire sequence is kept or discarded (no extraction)



Looking for Testers!

SuperCRUNCH

FOR PHYLOGENETIC DATA

Pre-Print	https://doi.org/10.1101/538728
Download	https://github.com/dportik/SuperCRUNCH
Documentation	https://github.com/dportik/SuperCRUNCH/wiki
Tutorials & Data	https://osf.io/bpt94/

Benchmarking

	Population level	Genomic (UCE's)	Supermatrix
File Size	52 MB	205 MB	13.3 GB
Sequences	82,557	338,942	8,785,378
Taxa	3	123	1,426
Genes	4	5,041	70
Gene and Taxon Filtering	5s	2hr 40min	34min
Sequence Similarity Filtering	8s	---	1hr 10min
Sequence Selection	1s	7min	7min
Sequence Alignment	2min	7hr 10min	2hr 40min
Post-Alignment Tasks	2s	21min	7s
Total cpu time	**~2.5 min**	**~10.5 hours**	**~4.5 hours**