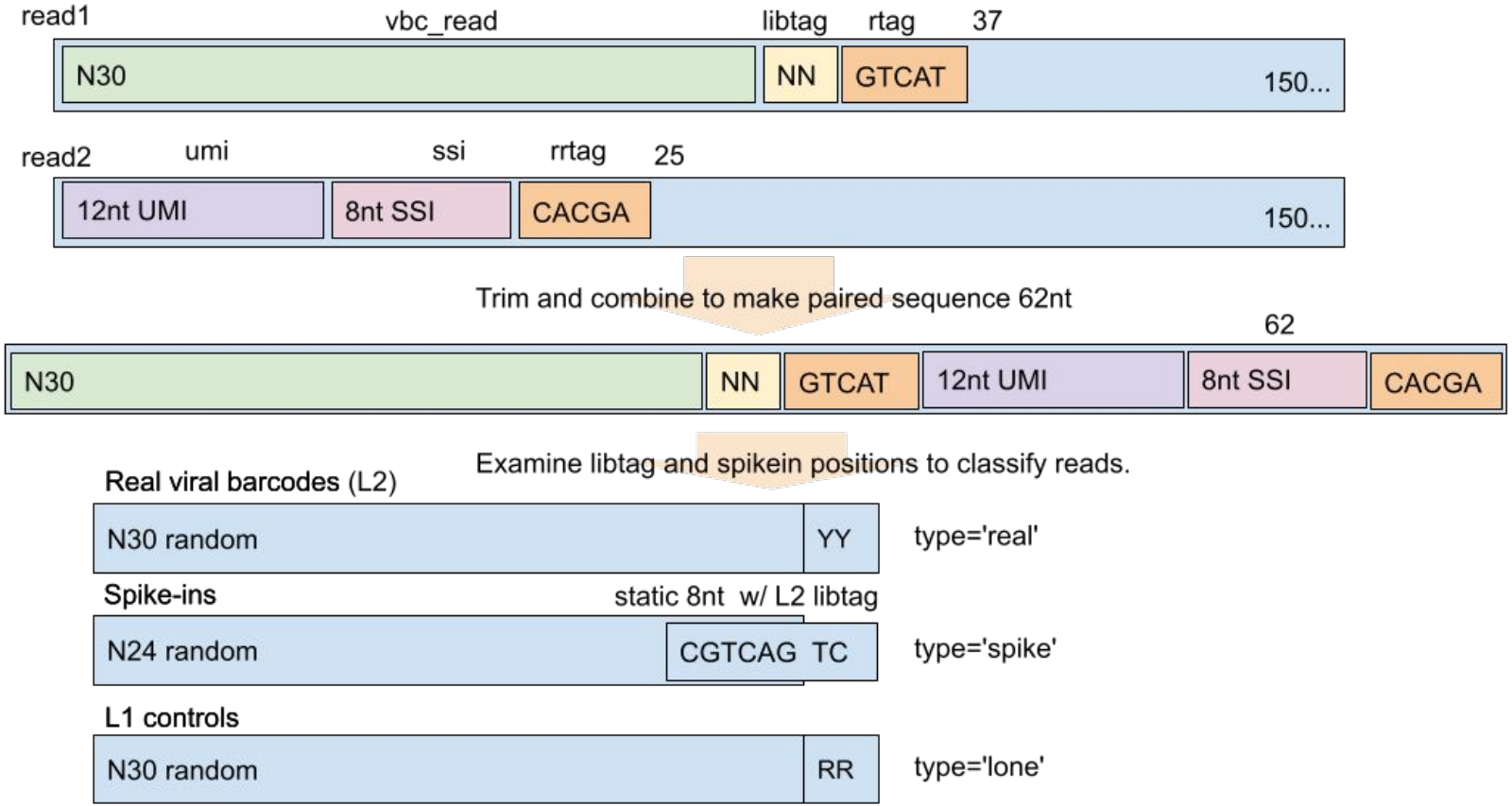
Read assembly from FASTQ



Processing Logic Recap

- Assemble reads
  - 52/62
  - 1 row per FASTQ read.
- Aggregate reads
  - collapse on identical reads
  - add read_count
- Filter, Split reads
  - remove N, homopolymers
  - split fields.

	sequence	source
0	CNGGTCATTTTGCGGTCGGGGAACACGGCATTNGCTGAGTGCCTGAGTTACC	M295-100
1	GTGCAGATATGGAGATGTTGGTGCGTTGTGTTNCGCAGCTGATAGAGTTACC	M295-100
2	GNCGAGAATTGGGTGTGGCGGCTTAGATGTAGNGGGGCAGCGTCGAGTTACC	M295-100
...		
803349987	GGCGGGCGAGGGGGGGCTGGGGTGAGTGGGATGNANGGTGACGCCCATGGTG	M295-9
803349988	ATGCCNCCAATGAGATGAAAGCATTAGGAATTCNCNGACACCCCCCATGGTG	M295-9
803349989	CGCGGGAGATAAACGGGGTACGATGTAGCGGGANGNACTCCGTTCCATGGTG	M295-9

	sequence	source	read_count
0	TTGGTGTTACCGTTCCGAGTTGTATGCCGATAGGGGGGGGGGGGGGGGGGGG	M295-40	10530
1	GGGTGCCGAAGATTCTAGTATAAGTAGCCGTAGGGGGGGGGGGGGGGGGGGG	M295-17	9842
2	GTTCTCAGATTATTTAATAATTCGTTACCTAAGGGGGGGGGGGGGGGGGGGG	M295-57	6407
...			
553739572	CTACGTAAAGAGTCCGAAGAAGGGACCAACGTGGCGGGCAGTGGAGATTCTA	M295-51	1
553739573	CTACGTAAAGAGTCCGAAGAAGGGACCAACGTGGCGGGCAACGAAGATTCTA	M295-51	1
553739574	CTACGTAAAGAGTCCGAAGAAGGGACCAACGTGGCGGGGCTTTAAGATTCTA	M295-51	1

	read_count	vbc_read	spikeseq	libtag	umi	ssi
0	1013	ATGTATGGTATCGAGGACAATGCTCAGTCA	CAGTCAAT	AT	GATCTGGAAGTA	GGGCGGGG
1	90	ACAGTTATTGACCGGTCAGGTCTTCGTGGC	CGTGGCAT	AT	GACAAGTATTGT	TGCCCCGG
2	81	GATGCTTATTATAATTACTCGATTGATGTA	GATGTAAA	AA	TACGCGCCAGGG	ACCGGGTG
...						
537350214	1	CTACGTAAAGAGTCCGAAGAAGGGACCAAC	ACCAACGT	GT	GGCGGGCAGTGG	AGATTCTA
537350215	1	CTACGTAAAGAGTCCGAAGAAGGGACCAAC	ACCAACGT	GT	GGCGGGCAACGA	AGATTCTA
537350216	1	CTACGTAAAGAGTCCGAAGAAGGGACCAAC	ACCAACGT	GT	GGCGGGGCTTTA	AGATTCTA

Processing Logic Recap (2)

- Interpret read fields
  - SSI -> label, rtprimer, site `[target, injection, controls..]` **remove mismatches**
  - libtag, spikeseq -> type [real, spike] **remove bad**
  - SSI -> other info from sample information spreadsheet

	read_count	vbc_read	umi	rtprimer	label	site	type	brain	region	ourtube
0	62	GTGGGTCAAAACTGTGACTGAGAAGGGCTC	CCCCCGCCCCCC	73	BC73	target	real	780345	ctx	1
1	55	GTGGGTCAAAACTGTGACTGAGAAGGGCTC	CCCCCCGCCCCC	73	BC73	target	real	780345	ctx	1
2	51	GTGGGTCAAAACTGTGACTGAGAAGGGCTC	CCCCCCCGCCCC	73	BC73	target	real	780345	ctx	1
...										
136227302	1	CTACGTAAAGAGTCCGAAGAAGGGACCAAC	AACATGTTTCGC	51	BC51	target	real	780345	BNST	
136227303	1	CTACGTAAAGAGTCCGAAGAAGGGACCAAC	AACATGGCAATG	51	BC51	target	real	780345	BNST	
136227304	1	CTACGTAAAGAGTCCGAAGAAGGGACCAAC	AACCCAGTGCAC	51	BC51	target	real	780345	BNST	

- Collapse `vbc_read` by Hamming
  - Performed per-brain.
  - Same format, same length, with vbc_read replaced by most common variant.

Processing Logic Recap (3)

- Aggregate VBCs on UMI
  - All previous steps read-oriented, now UMI-oriented.
  - Apply minimum read thresholds (injection, target)
  - Aggregate by label and type, sum unique UMIs, sum read_count

	vbc_read	label	type	umi_count	read_count	brain	region	site
0	AAAAAAAGTCCCTGCCCGCTATTAGAACTC	BC40	real	1	2	780345	ctx 3	target
1	AAAAAAATAACATCACATCATTCTGCGATC	BC90	real	7	14	780345	midbrain	target
2	AAAAAAATAACATCACATCATTCTGCGATC	BC92	real	24	49	780345	midbrain	target
...								
1143204	TTTTTTTGTTGTAAGCATTGAGCATATTGA	BC94	real	1	2	780345	midbrain	target
1143205	TTTTTTTGTTTTTTTCGGGGGATGTGTTCC	BC67	real	8	16	780345	ctx 2	target
1143206	TTTTTTTGTTTTTTTCGGGGGATGTGTTCC	BC70	real	1	2	780345	amyg-GPe	target

- Filter to prepare for matrix generation.
  - Apply minimum UMI thresholds, conditionally for target (any passes, keep all)
  - Require target VBCs in injection (and vice versa)
- Generate per-brain matrices
  - Pivot on label -> create real and spike-in matrices
  - Normalize real VBC by spike-ins per label