

---

# ASSESSING ASSEMBLY QUALITY

# MEASURES OF ASSEMBLY QUALITY

- ▶ Contig N50\*
- ▶ Mapping of "proper" pairs
- ▶ # of full length proteins
- ▶ Contig ExN50 (to be covered after transcript quantification)

### CONTIG N50

- ▶ Like a "weighted median"
- ▶ The length of the contig for which half of the total number of base pairs are in contigs of greater or equal length

\_\_\_\_\_

\_\_\_\_\_

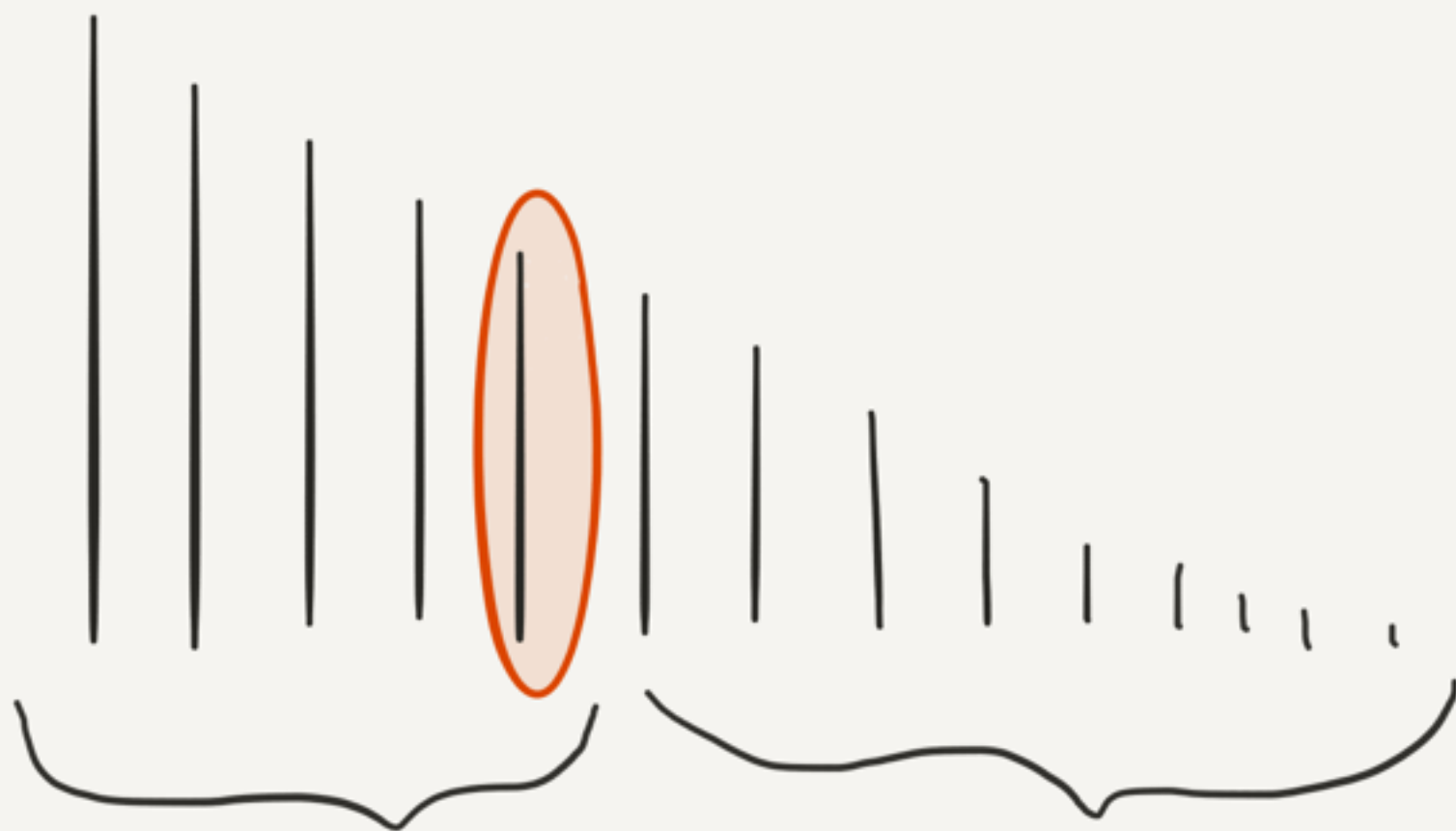
\_\_\_\_\_

\_\_\_\_\_

\_\_\_\_\_







# MAPPING OF "PROPER" PAIRS

- ▶ Map paired end reads back to transcript contigs
- ▶ A "proper" map is when both pairs map to the same contig
- ▶ An "improper" map is when the pairs map to different contigs



---

---

---

\_\_\_\_\_

\_\_\_\_\_

\_\_\_\_\_

\_\_\_\_\_

\_\_\_\_\_

\_\_\_\_\_

---

□ — □

□ — □

□ — □

---

□ — □

---

---

□ — □

□ — □

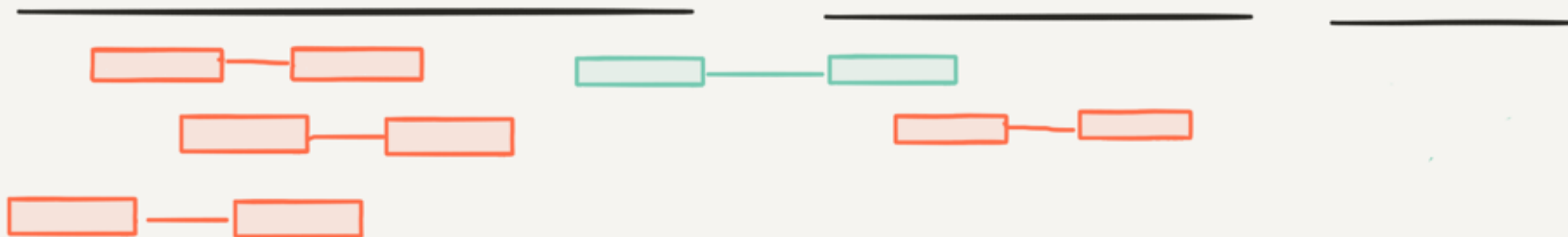
□ — □

---

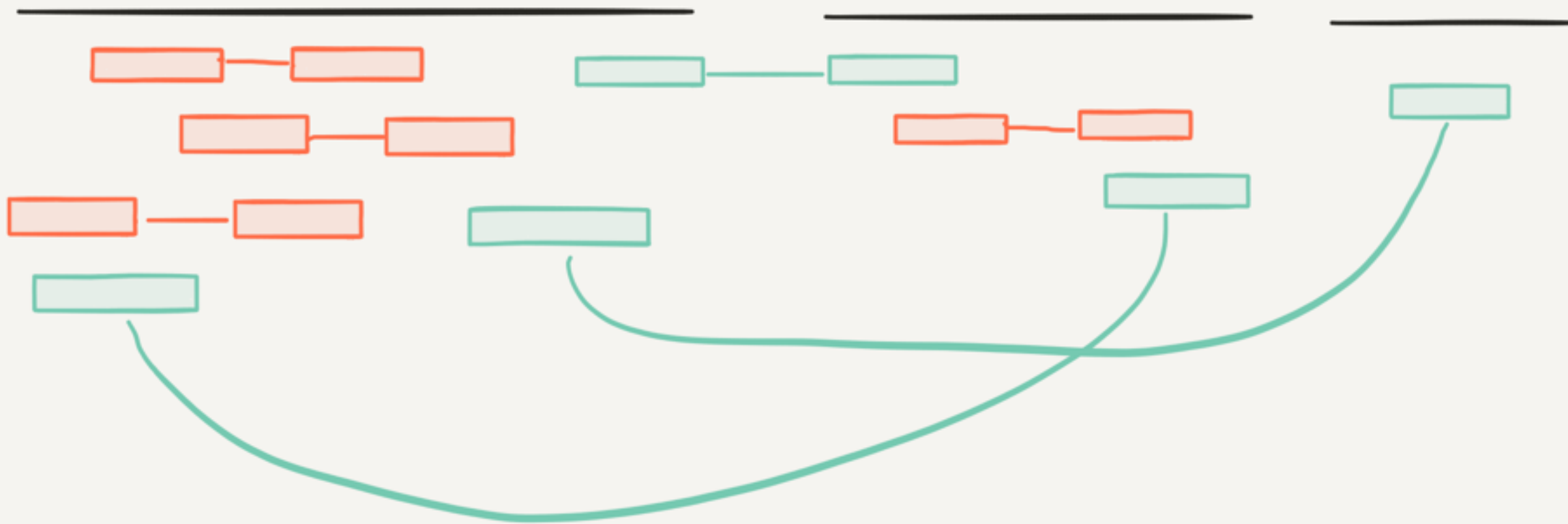
□

□ — □

---







# MAPPING OF "PROPER" PAIRS

- ▶ Map paired end reads back to transcript contigs
- ▶ A "proper" map is when both pairs map to the same contig
- ▶ An "improper" map is when the pairs map to different contigs
- ▶ The number of reads that properly map is a good measure of contiguity
- ▶ The value for a good Trinity assembly is often  $>70\%$



### NUMBER OF FULL LENGTH PROTEINS

- ▶ Use BLASTX to compare to a well-curated database of proteins (like Swiss Prot)
- ▶ Number of full length proteins can give you information on how well the transcripts are reconstructed

### HANDS-ON

- ▶ Go back to <http://github.com/SmithsonianWorkshops/RNA-seq> and follow the "4a\_Assessing Trinity assembly quality.md" tutorial