Analysis of structural variation

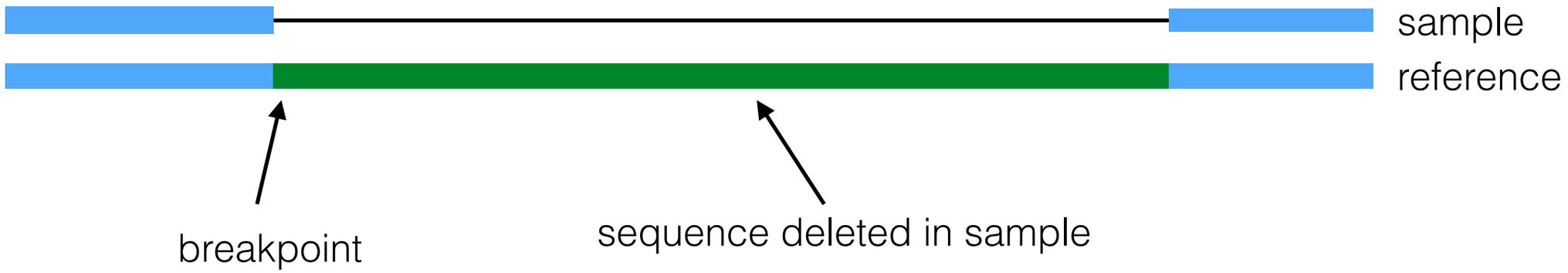
Alistair Ward - Boston College

What is structural variation?

- What differentiates SV from short variants?
- What are the major SV types?
- Summary of MEI detection

What is an SV?

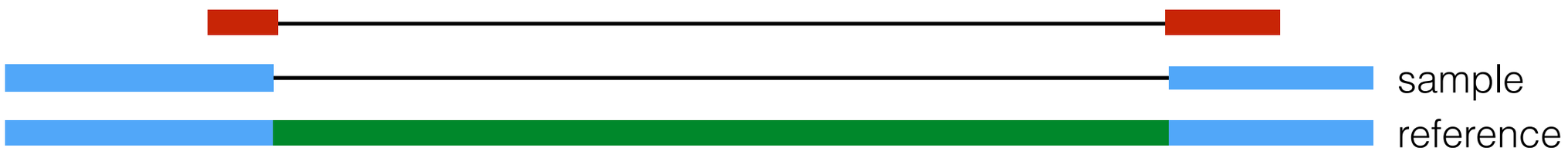
- Often considered to be >1kb or larger
- Practically, often considered >= read length



What is an SV?

- Often considered to be >1kb or larger
- Practically, often considered >= read length

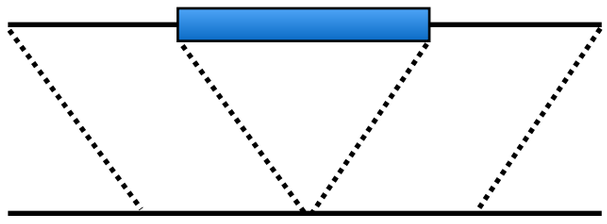
short deletion - mapping can span the gap
can be detected with short variant detectors



structural variation - mapping cannot span the gap
cannot generally be detected with short variant detectors

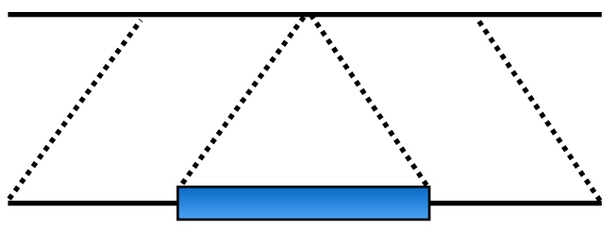
SV types

Deletion



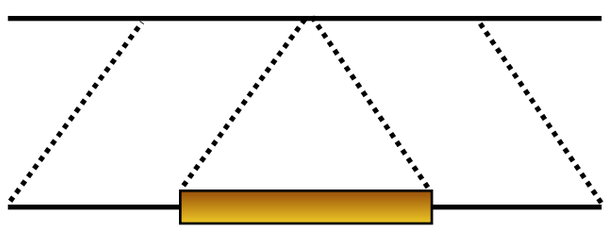
If inserted sequence/deletion ≦ 50bp
-> indel

Novel insertion



sequence ≧ 50bp
-> Copy number variation

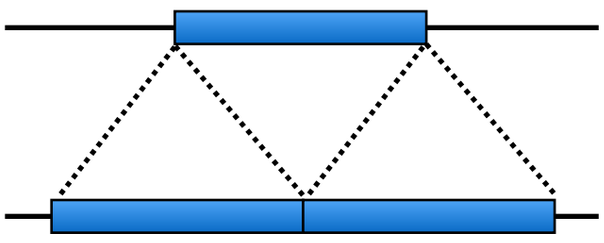
Mobile element insertion



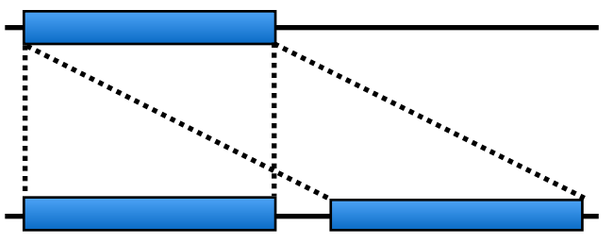
The mobile element sequence is ubiquitous in the genome

SV types

Tandem duplication



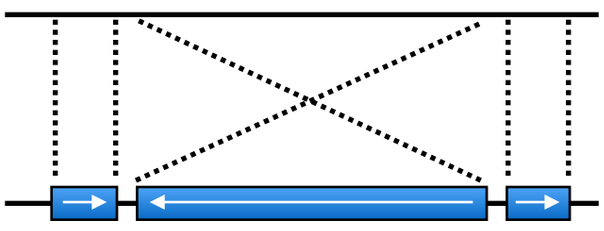
Interspersed duplication



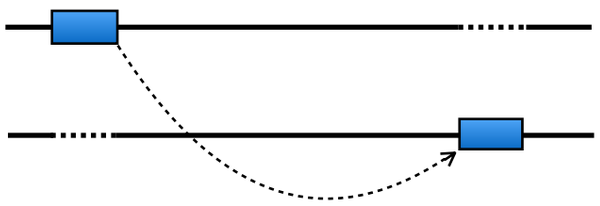
Duplications cause problems for mappers. From which copy of the duplicated sequence did the read originate?

SV types

Inversion



Mappers will not be able to place reads correctly in the inversion, or across the breakpoints



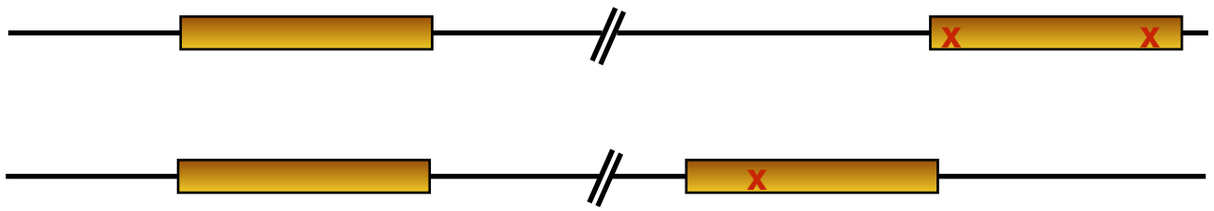
Mappers will be able to place reads in the translocated sequence, but fail at the breakpoints

Mobile element insertions (MEIs)

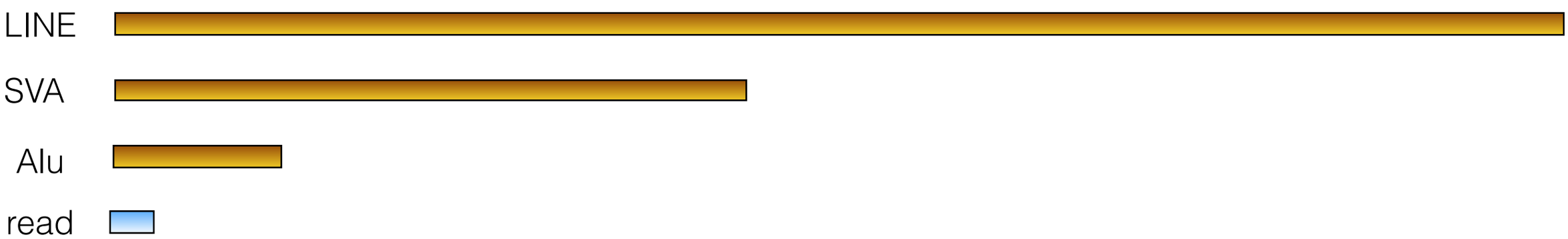
- Retrotransposons comprise nearly 50% of the human genome
- Implicated in a number of diseases, (Crohn's disease, haemophilia, …)
- non-LTR transposons are still active in the human genome
- Why can't we use short variant detectors to find MEIs?

MEI detection

MEIs are repetitive



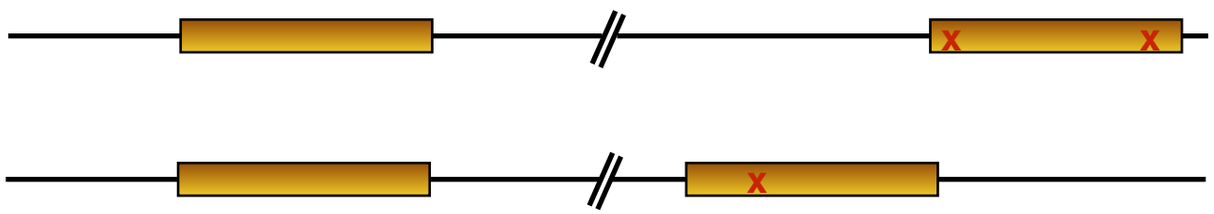
MEIs are large with respect to read length*



* sequencing technologies are closing this gap!

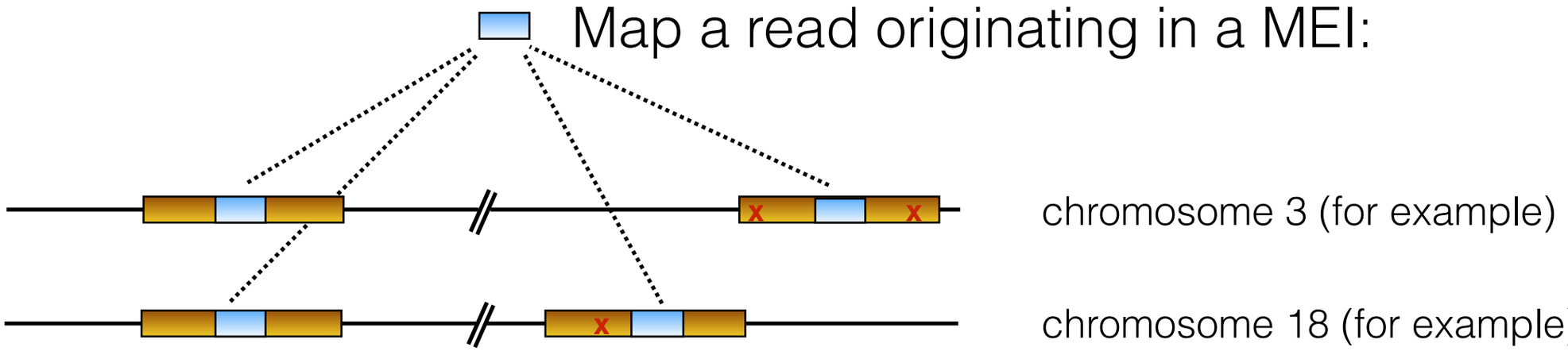
Map reads from MEIs

Map a read originating in a MEI: 



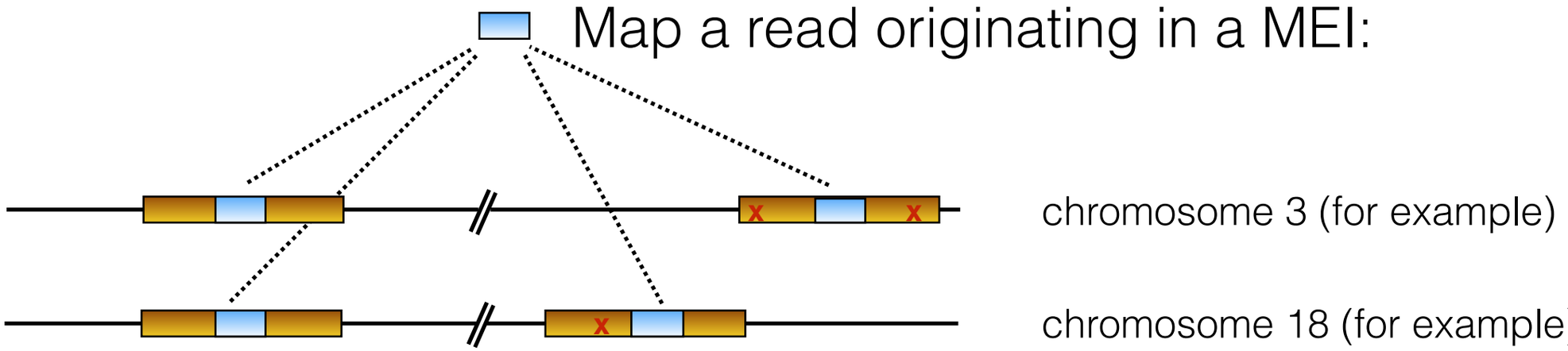
* sequencing technologies are closing this gap!

Map reads from MEIs



What did we learn? - Not much

Map reads from MEIs



What did we learn? - Not much

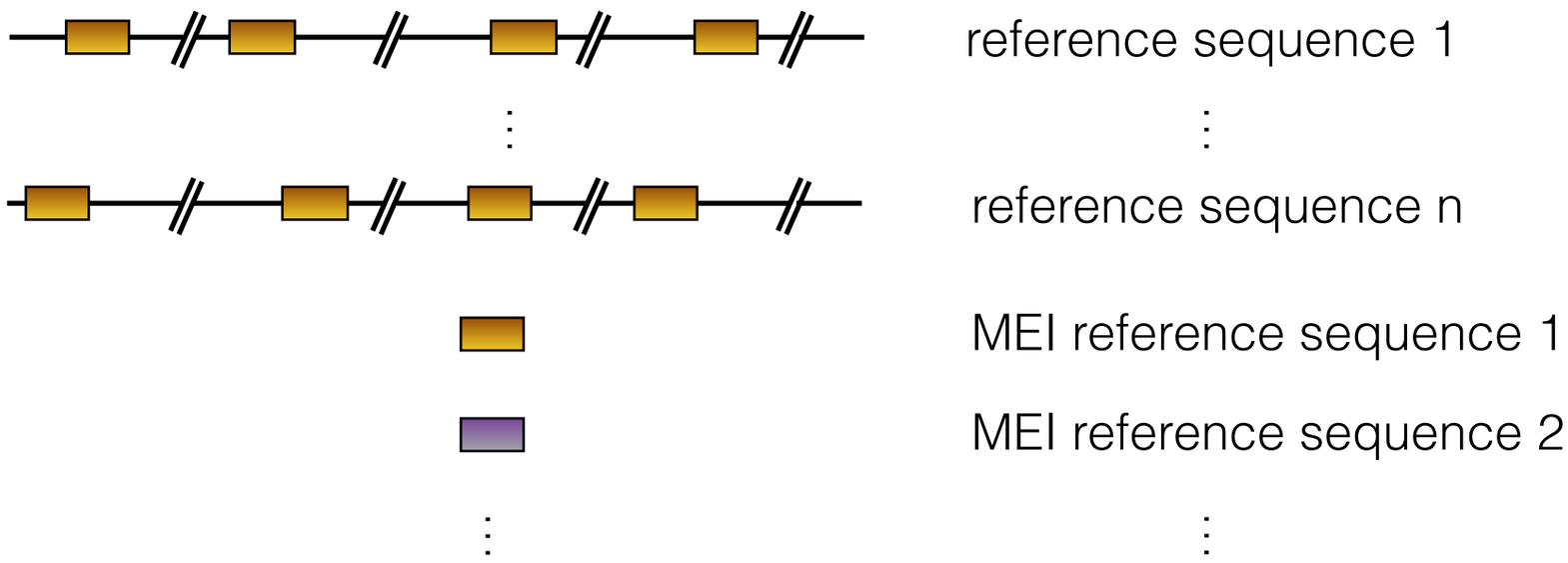
But remember, we have read pairs!

The DNA fragment isn’t so small!

Update to mapping strategy

There are well over 1,000,000 Alu elements in the human genome

Recall our mapping strategy?

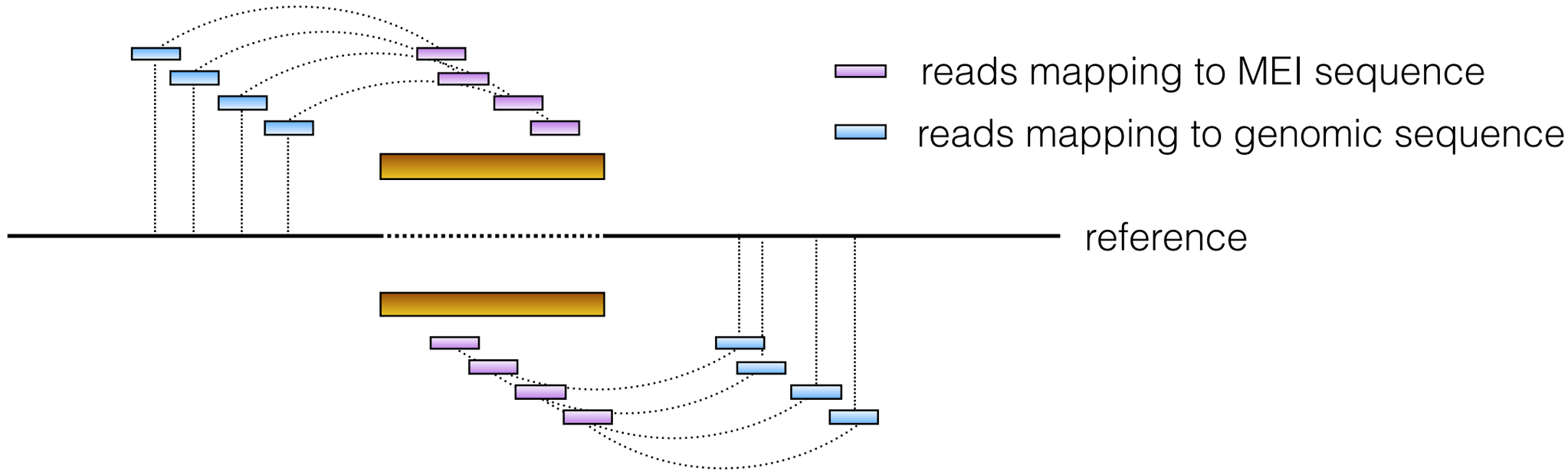


read → Hash read and find clusters in the reference

If clusters in MEI sequence, don't bother looking in the genome

Evidence for non-reference MEI

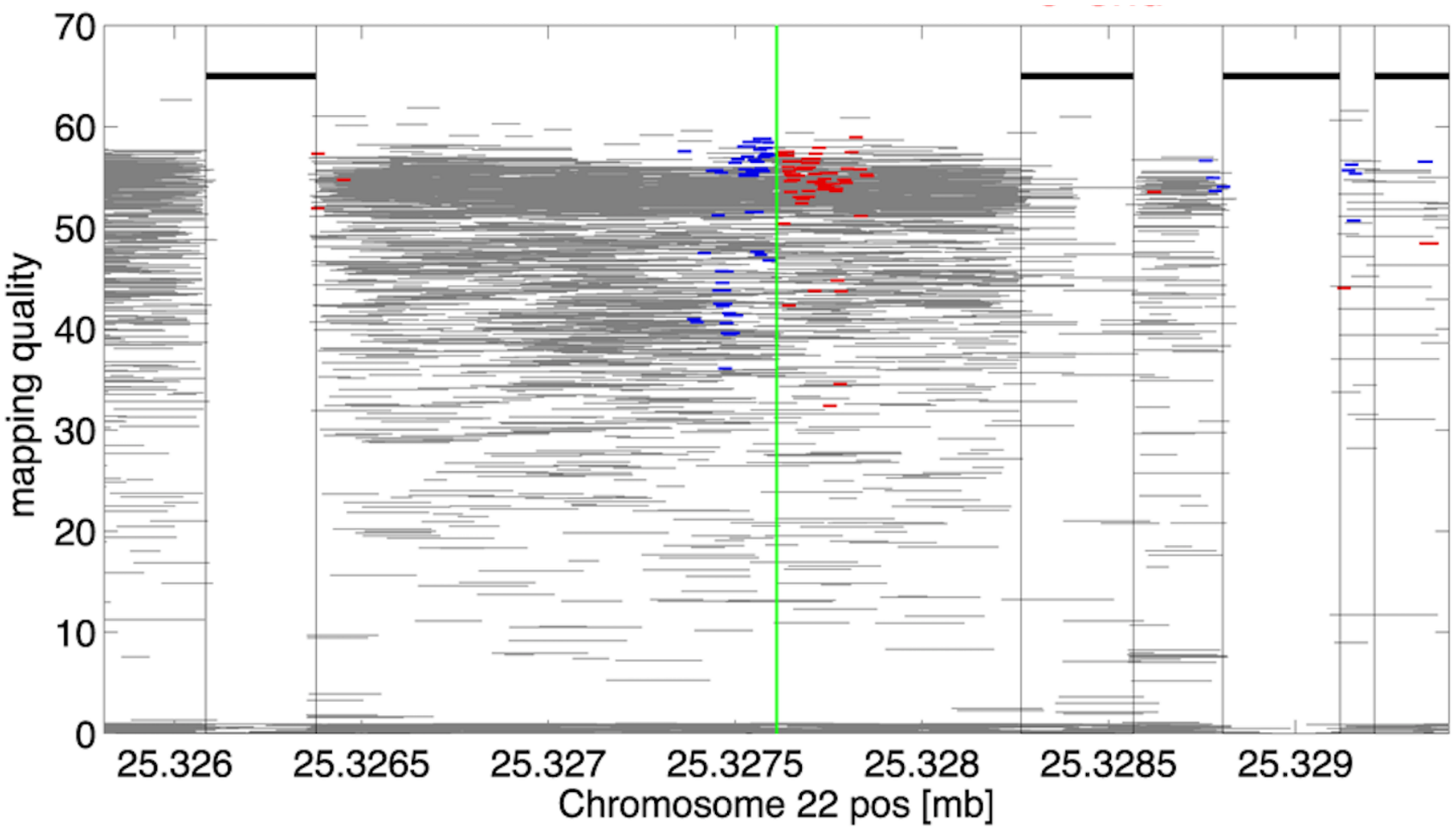
Search for fragments with one mate uniquely mapped
and the other falling within an MEI



Demand fragments spanning into MEI from both
the 3’ and the 5’ end

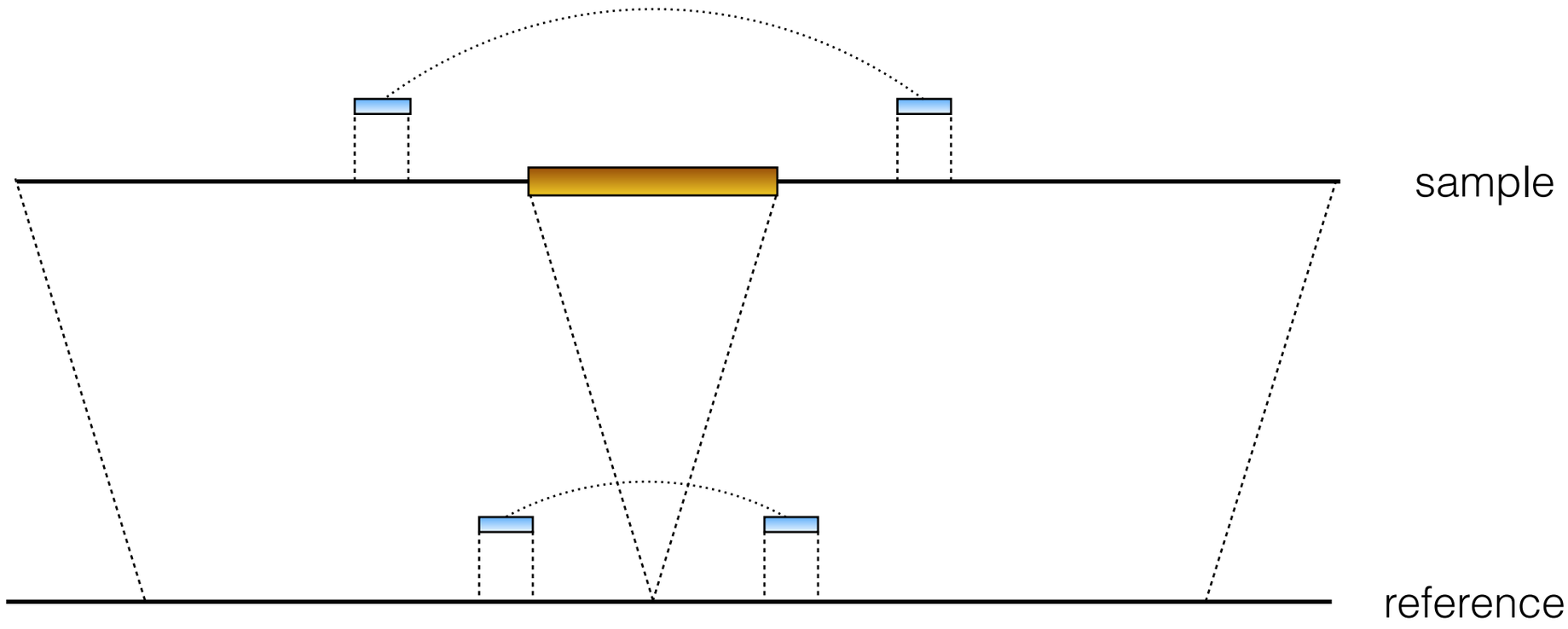
‘Spanning in’ evidence

1000 Genomes Pilot Project data



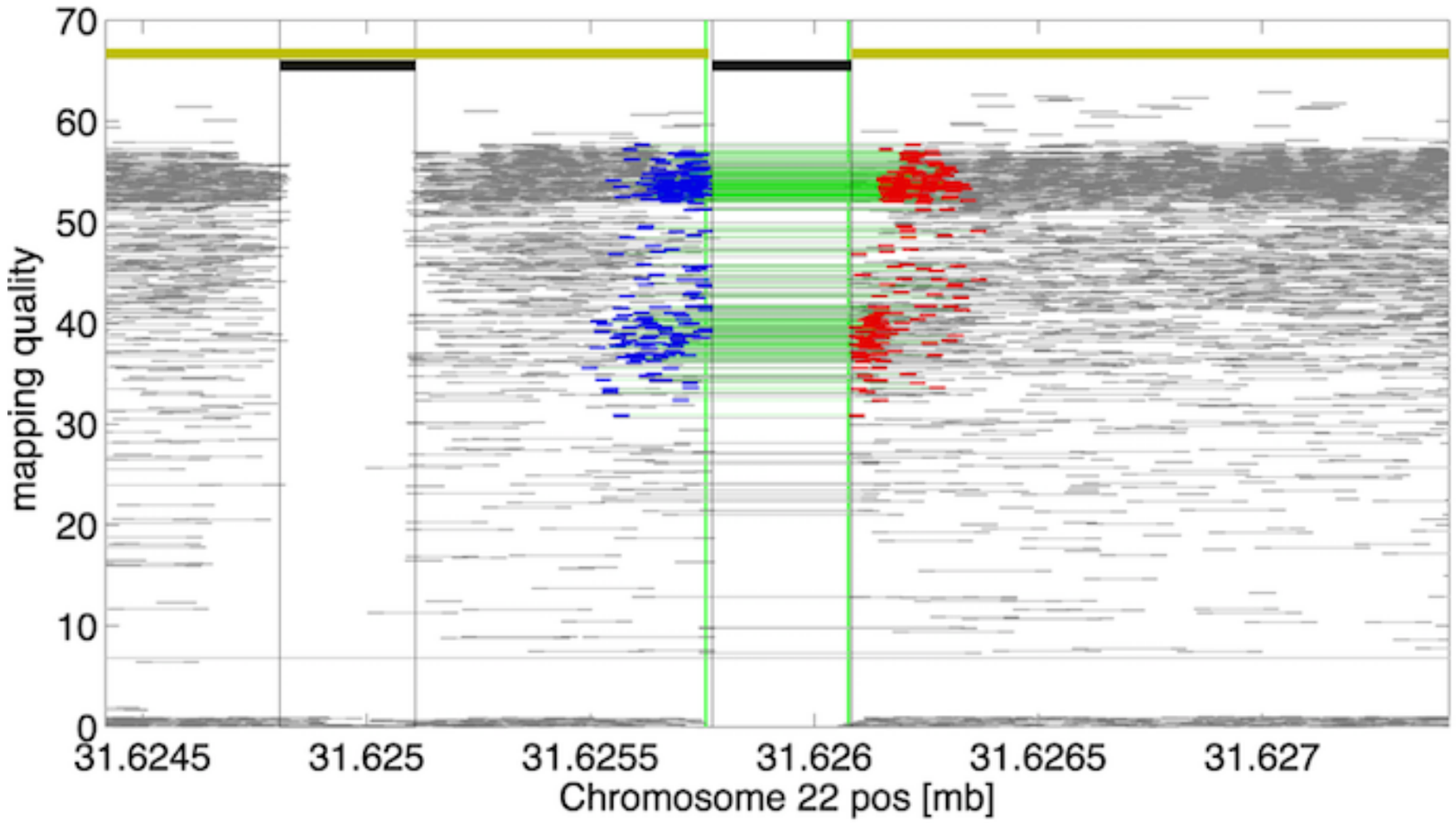
A Comprehensive Map of Mobile Element Insertion Polymorphisms in Humans, Stewart et. al., (2011)

Span across a (non)-reference MEI



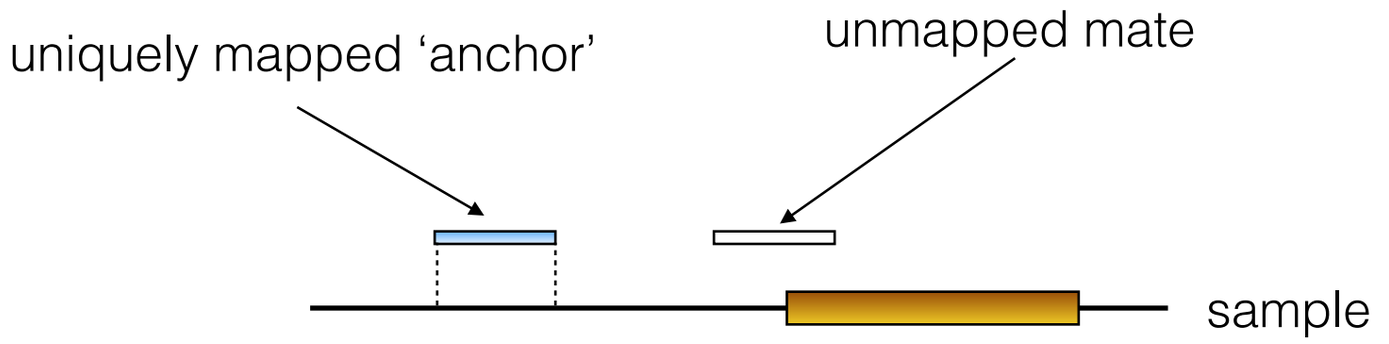
Search for mappings with abnormally short or long fragment lengths

‘Spanning across’ evidence



A Comprehensive Map of Mobile Element Insertion Polymorphisms in Humans, Stewart et. al., (2011)

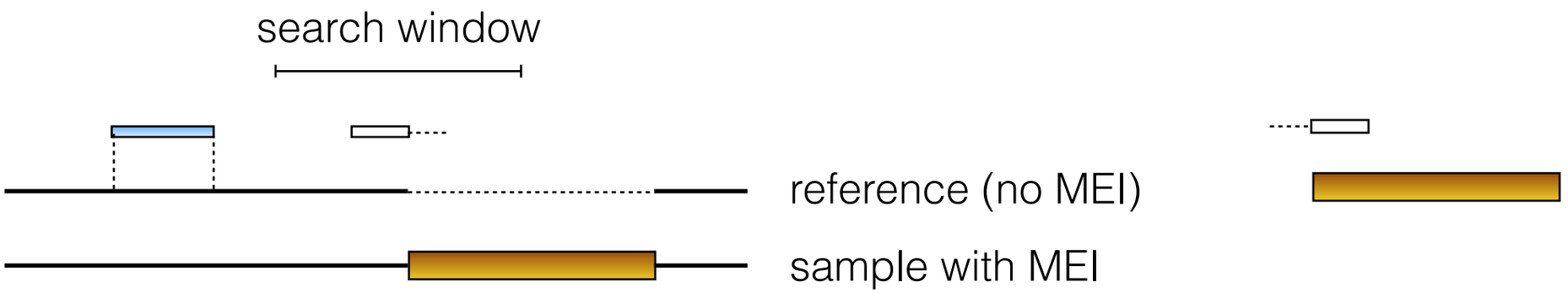
Split read evidence



Attempt to:

a) map unmapped mate to reference in a window based on anchor position and fragment length distribution

b) map unmapped mate to known MEI sequences



Summary

- Modify mapping to explicitly look for MEI mappings (Mosaik is set up to do this)

- Collate evidence from read pair and split read signals

- Leverage population to improve coverage

- Use local graph alignment to aid in genotyping