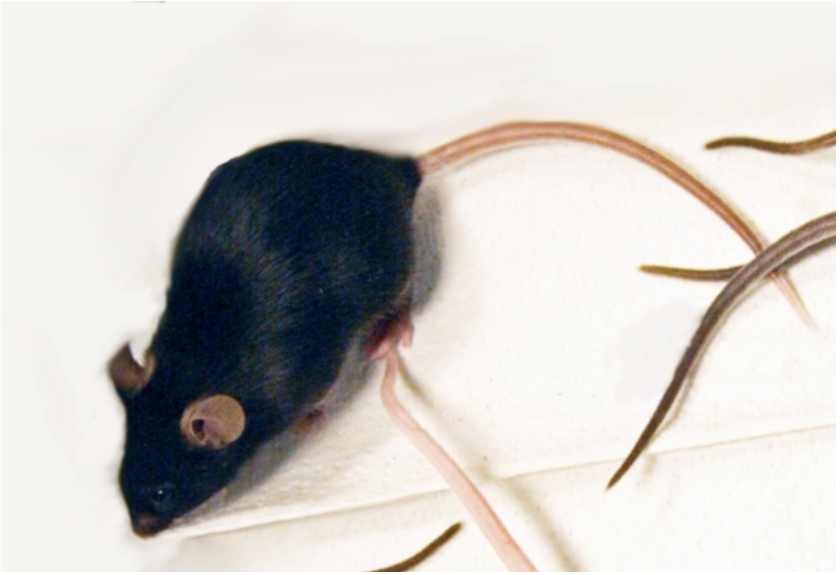
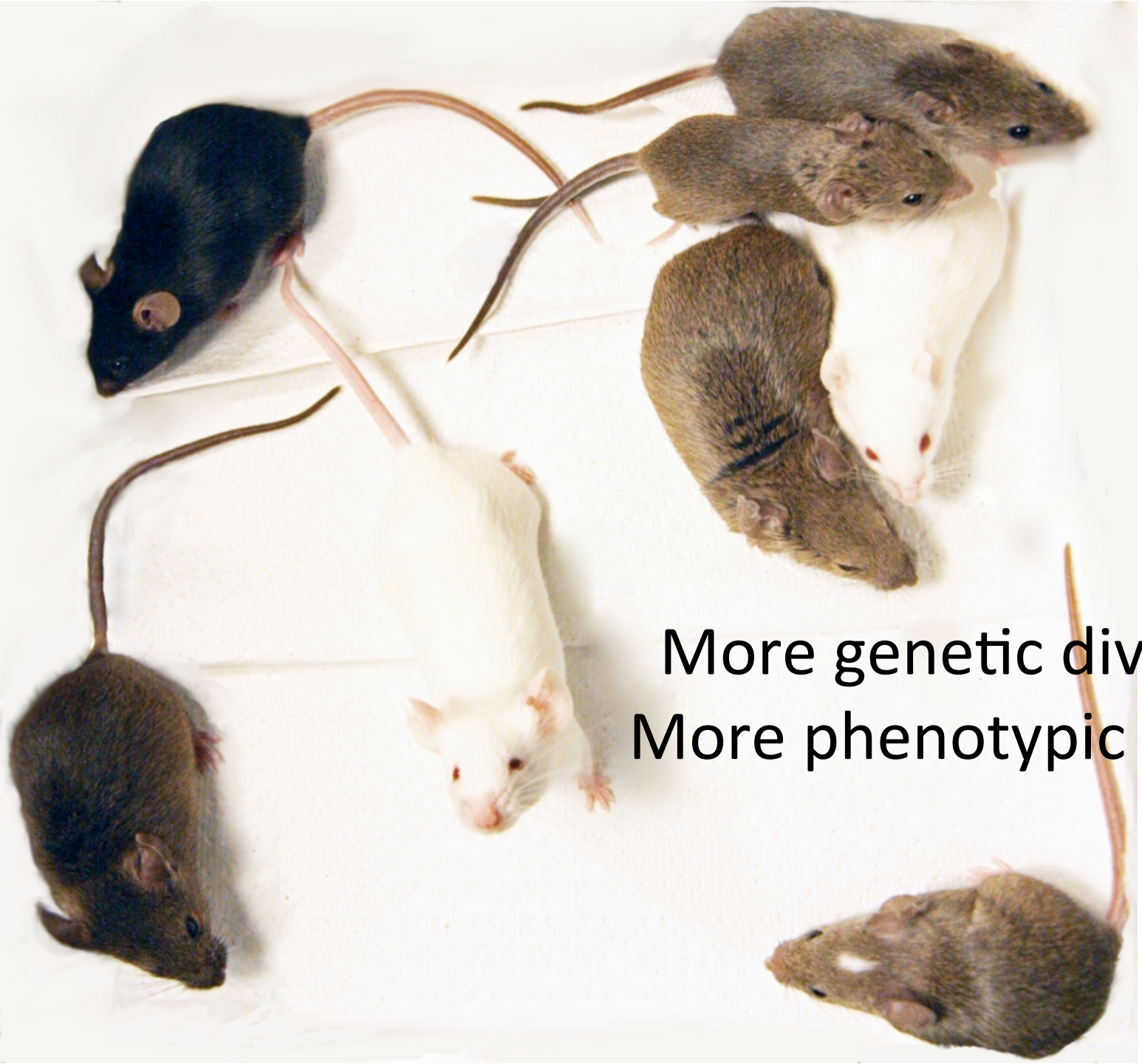
RNA-seq alignment to individualized genomes.

Steven Munger

Gary Churchill Group
The Jackson Laboratory

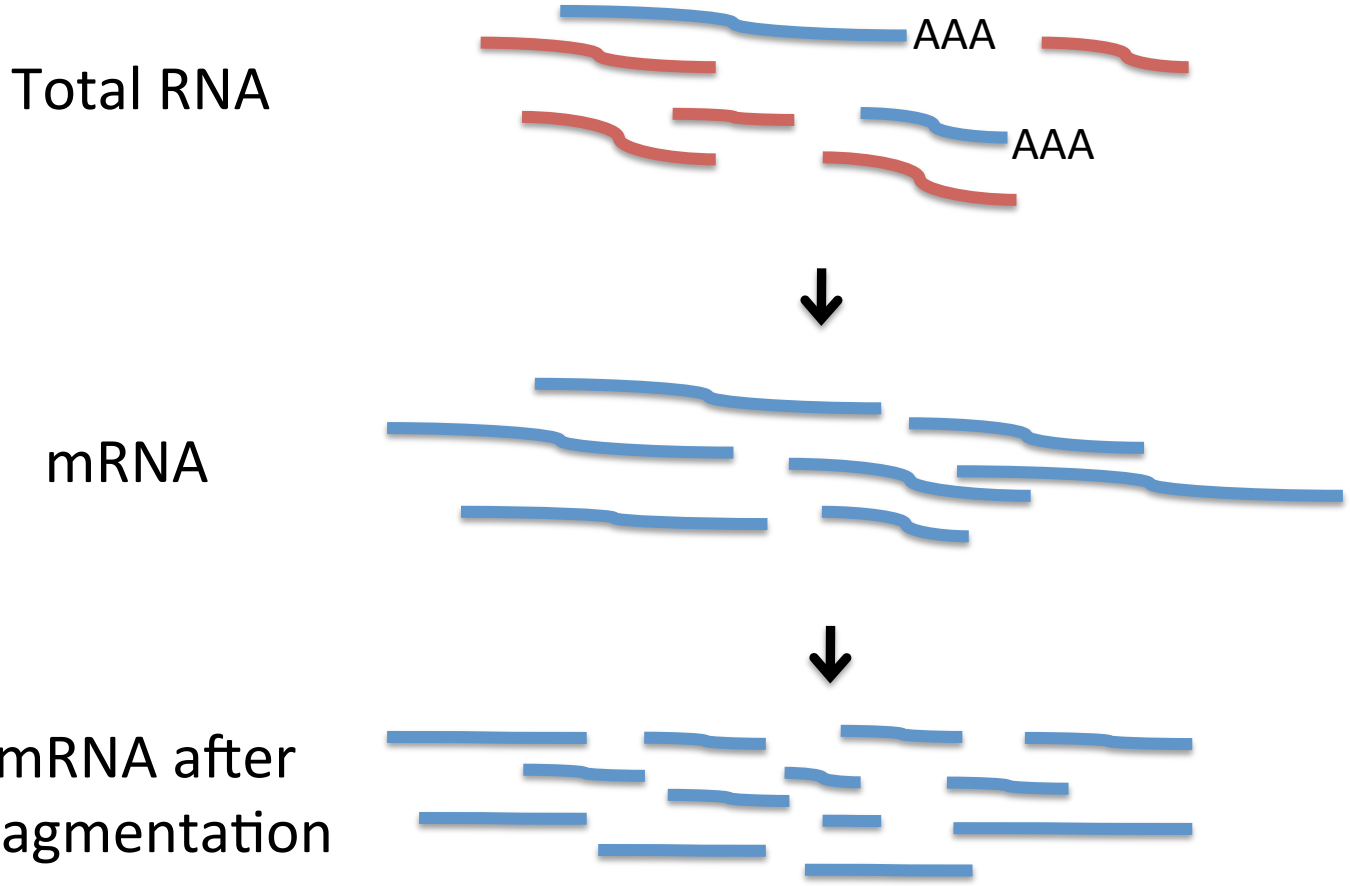
(Not to scale)





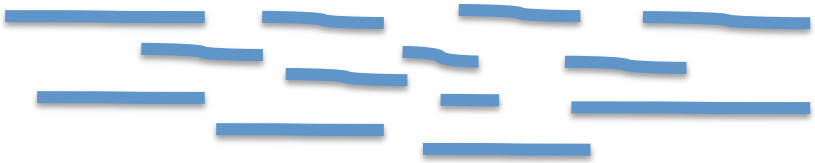
More genetic diversity =
More phenotypic diversity

Overview of RNA-seq

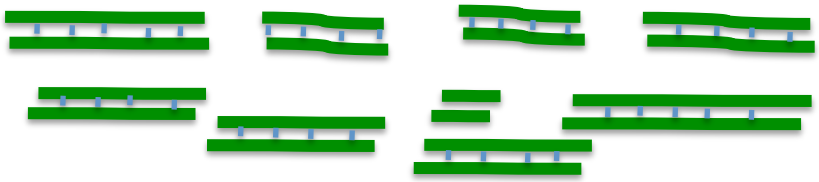


Overview of RNA-seq

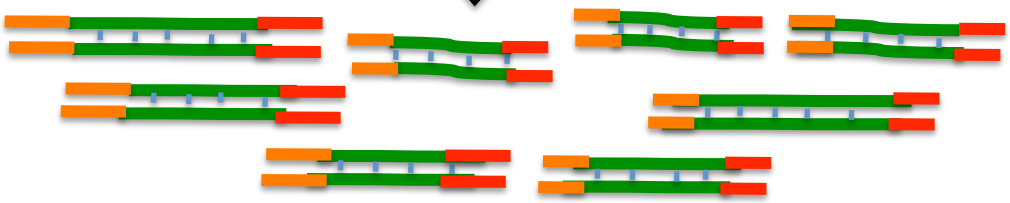
mRNA after fragmentation



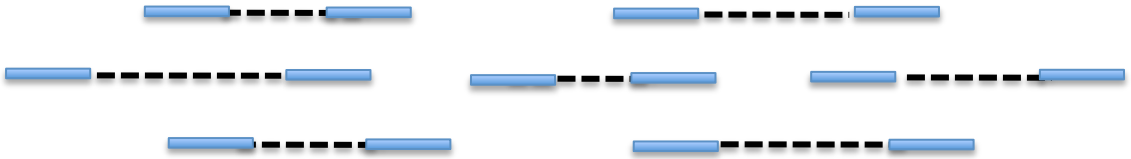
cDNA



Adaptors ligated to cDNA



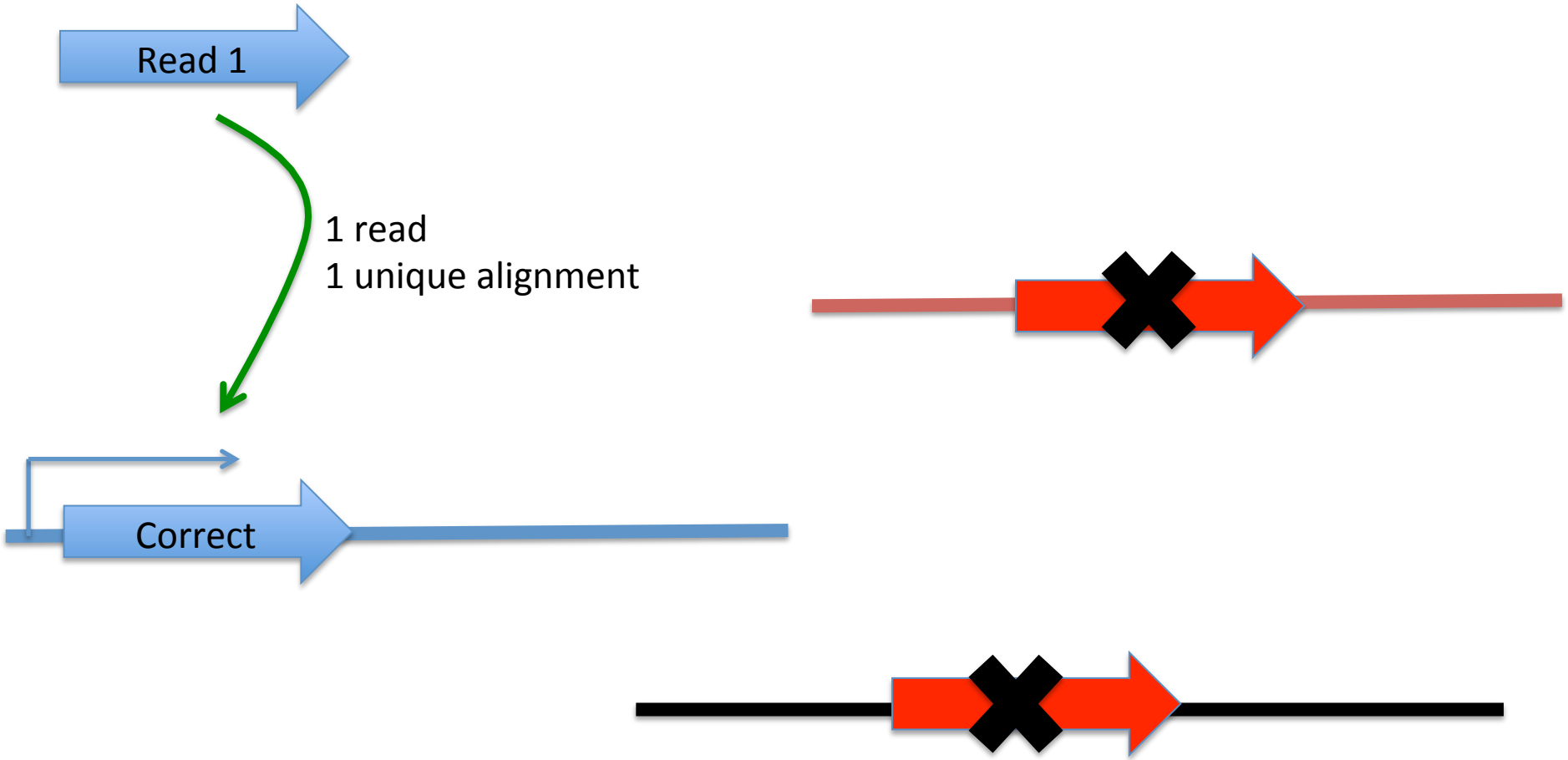
Single/ Paired End Sequencing



Millions and millions of reads…

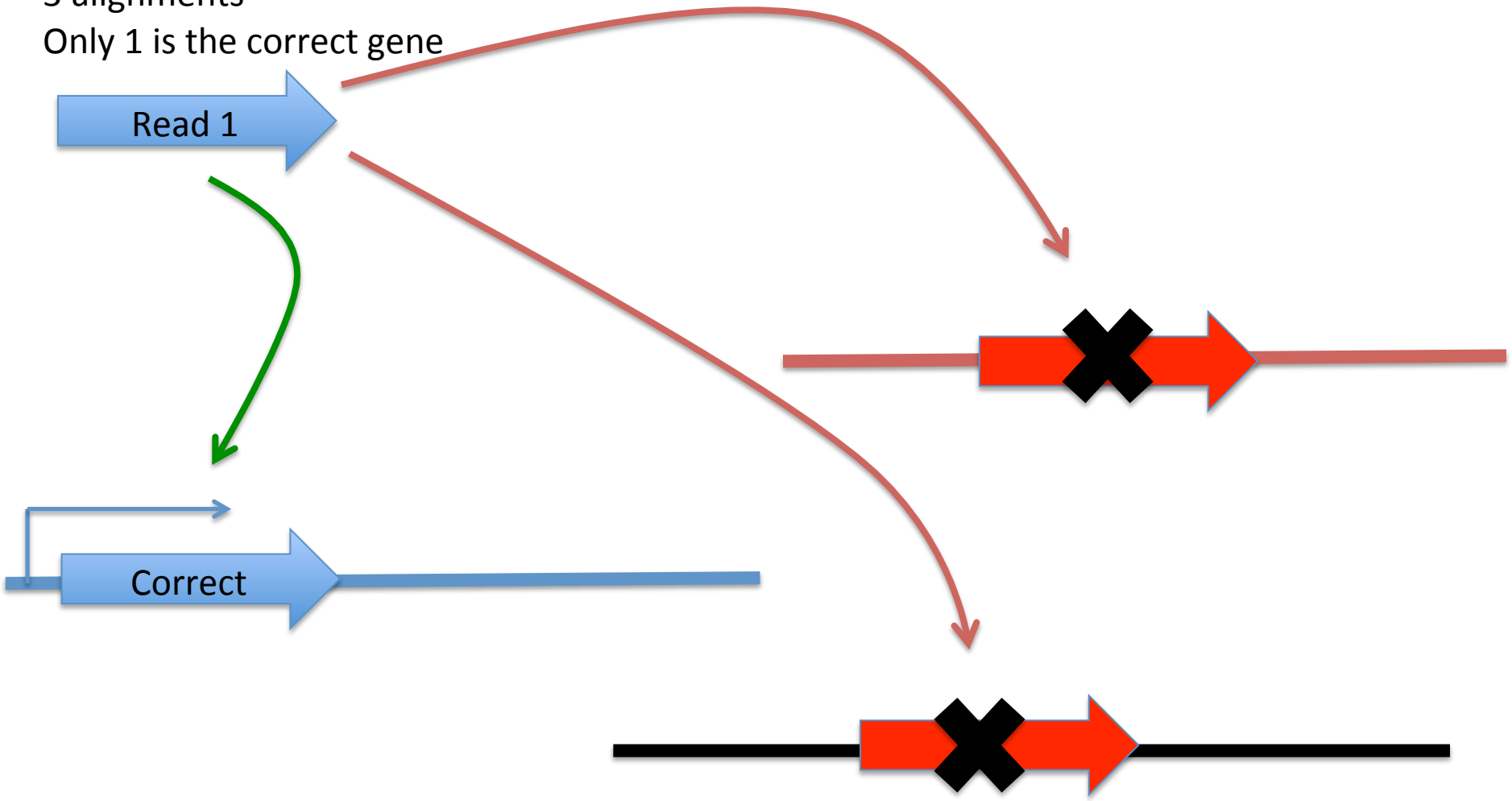
```
@HISEQ2000:160:D0WTCACXX:2:1101:1591:1957 1:N:0:GTGGCC
NTCATGCAACACAGCATCAGTGTACGGCATGTGGTTCCTGTCCTGCATGGTGGGTCTCCGATGTCTACCAATCACATGAGCAATTTCTTCCTGAATTTTA
+
#4=DDFFFHHHHGJJJJJJJIIIJJJJJJJJIIJHIJJJJJJJJJJJJJJIJJJIIIJJJFHHFFFFFFFCEEEEDDDDDDDDDDDEDDDDDDDEDDE
@HISEQ2000:160:D0WTCACXX:2:1101:1924:1989 1:N:0:GTGGCC
NGAATGGTTGACAACCACATAGCATTCTTTTTATAGAAAGGATTTTCTATGACCTCCTCCAGTGCATTGAGCACATCTGACTTTGACATTGTTCTGATAT
+
#1BDFFFFHHHHHJJJJJJJJJJJJJIJJJJJIJIJJJJJJHGIJJIJIJJJJJJJJJJIJJJJJJJJJJJGIJHHHHHHFFFFFFEDEEEEEDEDDEDD
@HISEQ2000:160:D0WTCACXX:2:1101:2175:1993 1:N:0:GTGGCC
NCTCAGTGTACAGAACCATACAGAAGTGTACATTTTAAAATCCACAATGCAATGGTATGAATTACCAGGTCTAAGAAAACATAAAAGGGGGTCAGACAGT
+
#1=DFFDFHHHGHJJJJJIJIIIJIJJJHHHIIIJJJJJJIIIJJJIIJJJJIJJFIIIJJIJJJJJJIHHIJJIIIJJHHHGHHFFFFDD=?BDDDDDA
@HISEQ2000:160:D0WTCACXX:2:1101:2320:1968 1:N:0:GTGGCC
NTACGATATGGTATAATTCTATTCATCGTCTCGGAAGTATTTTTCTTTGCAGGATTCTTCTGAGCGTTCTATCATTCTAGCCTCGTACCAACACATGATC
+
#1=BDDDDFBF<BGFFF@FI4CFFIIIGGFEGIIIIF:DBDFFF;BGE??G@DFAF<BBCFBFADFAEEFEAEFFFDDDD;BBCCC?@BBBBBBBBB@@>
@HISEQ2000:160:D0WTCACXX:2:1101:2675:1971 1:N:0:GTGGCC
NAAGTGGGTAACCTTTCTCCTCCTCCTCTTCGTCTCCGGCTCTGCTTTTTCCAGGGGTGTGTTTCGCCGAGAAGCACACAAGAGTGAGATCGCCCATCGG
+
#4=DDDFFDHHHHJJJJJIJJJJIJJIJJJJGI7F;F6D@:?F>*8?DHG;=C@F@E5=>BDEFEDCA>;B6=??ADDDDBBCA+:@@@A@BDD<BDDDD
@HISEQ2000:160:D0WTCACXX:2:1101:2970:1981 1:N:0:GTGGCC
NCGGGGGCAGTTGATGTCCAGTGACTTAAGGGGTTCCTTGGCTTTTCCAAGTGTGTACTGTCTTATTTCAGACACATTTTTCATTACTTTTCTGTTGCTT
+
#11=BBDA:?:CBD>DD?DB?*??BD>D4?B(?AACCCD;3=@CEEC=>C;=A?>@:;7>BAAA;>AD;;>A5>99AABDAA>>(,:>AAA#########
@HISEQ2000:160:D0WTCACXX:2:1101:3001:1980 1:N:0:GTGGCC
NAAACATCTTCAACTTCAAACTCAGCCGAGAAGGGCTGAGGCTGTCCAATGATTTGATGGGCTCCTATGCTGAGATGAAACTTGACCACACACACAGTCT
+
#1BDFFFFHHHHHJJJJJJJJIJJJJJJJIJJJJJJIJJJJJJJJJJJIIJIJJJICIJJJJHHHHFFFFFFFDDCEEECDCDDDDDDDDBDDDDDDACD
@HISEQ2000:160:D0WTCACXX:2:1101:3978:1965 1:N:0:GTGGCC
NTGTGTTCTAAGTGGTAAACAAATAATCTGCGCATGTGCCAAGGGTAGTTCTCCACTCCATGTGCTCTGCCTTCCCCGTGACGACAACTCGGCCGATGGG
+
#1=BDDFFHHHHHJJEGGIGJIJIJJJJJJJJIJJJGIIJJJJIJDFHFGIGIJJJJJIJJJJJJJJIJJIHHHHHFFDC@DDBDDBDCDDBDDDBDDD2
```

Alignment 101: The perfect read



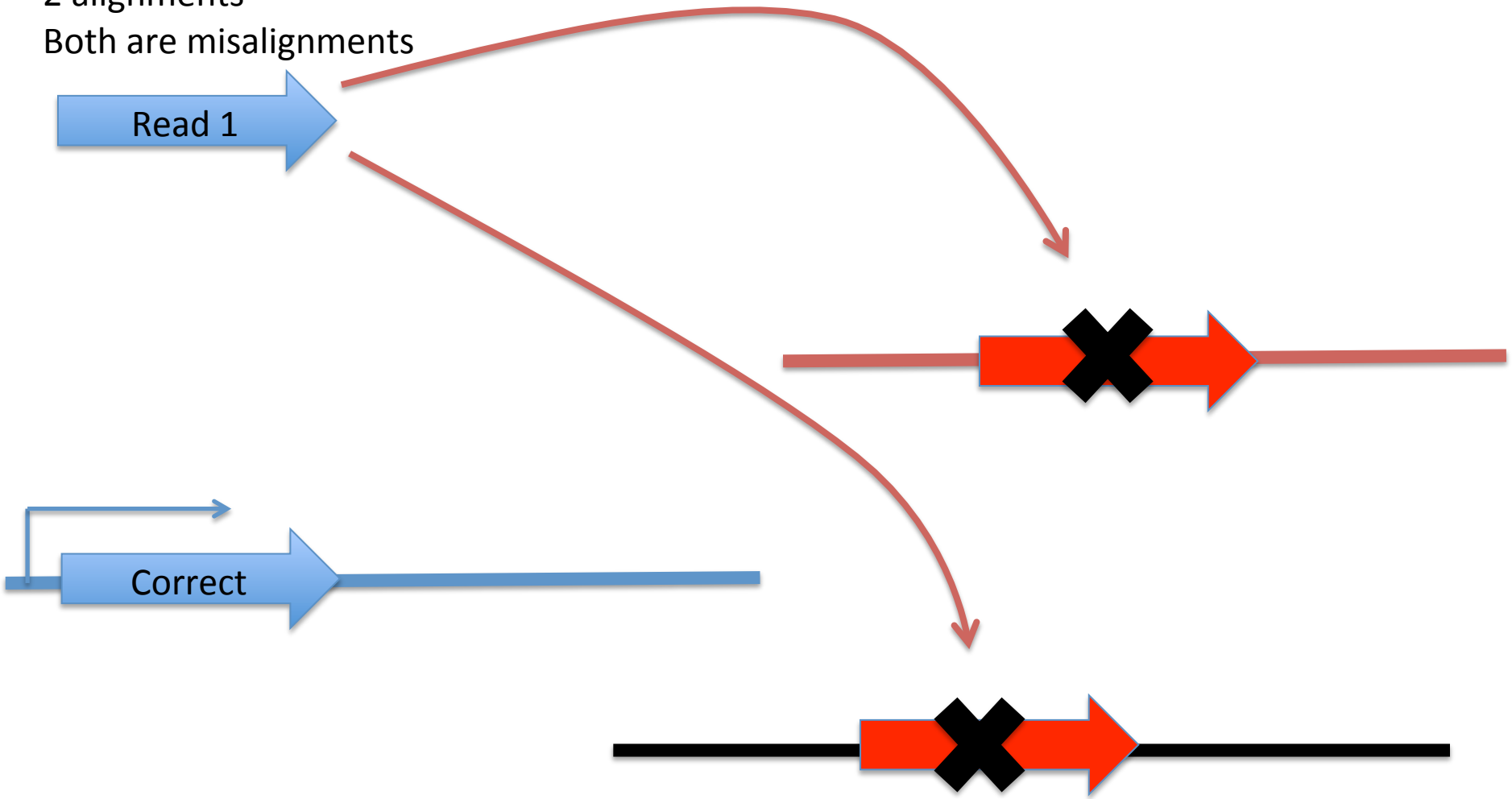
Alignment 101: Some cases

1 read
3 alignments
Only 1 is the correct gene



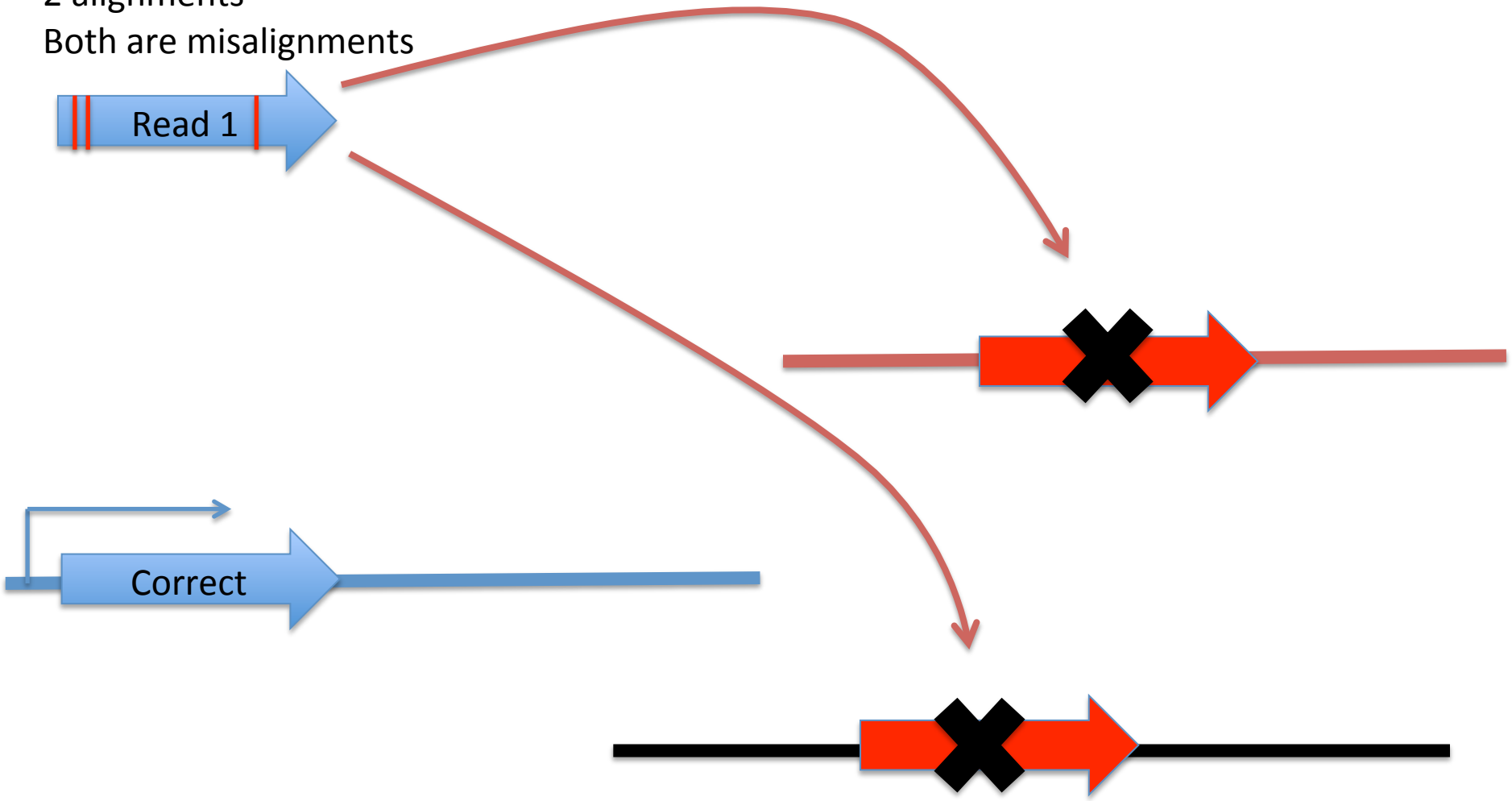
Alignment 101: Worst case

1 read
2 alignments
Both are misalignments

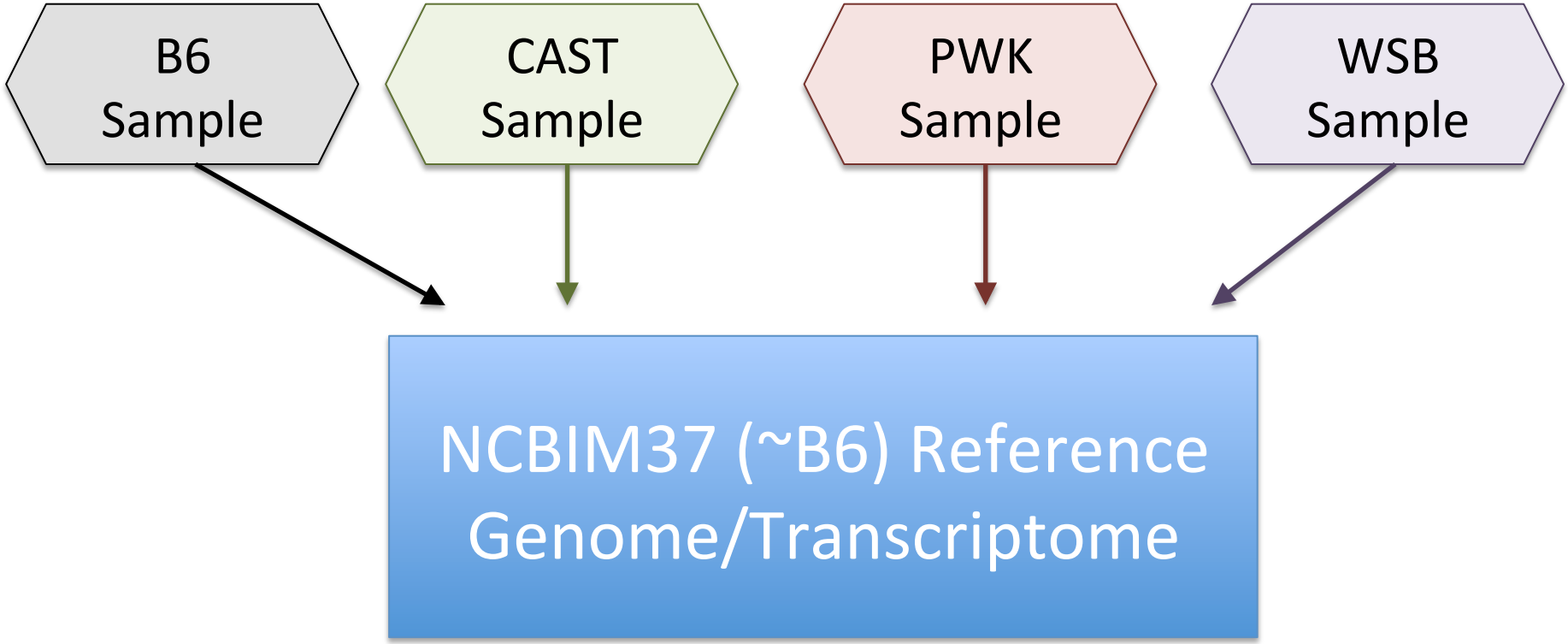


Individual genetic variation may affect read alignment.

1 read
2 alignments
Both are misalignments



A “One reference aligns all” strategy.



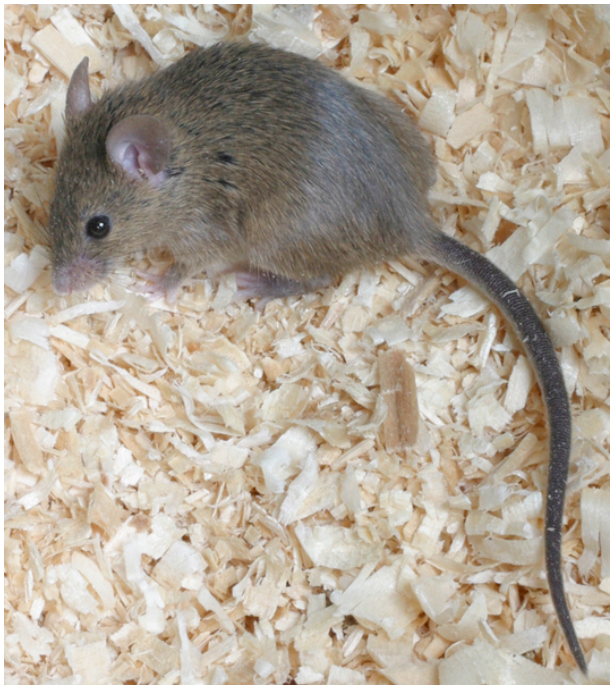
How does genetic variation affect alignment of RNA-Seq reads?

C57BL/6J

≈

≠

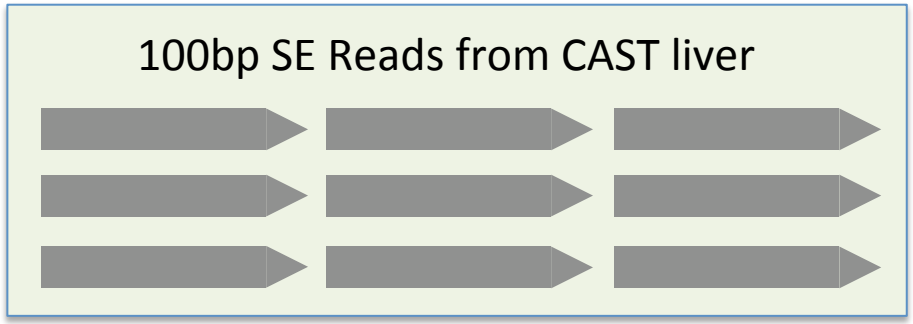
CAST/EiJ



Based on known gene annotations, we expect that >50% of 100bp CAST reads will have at least one SNP that differs from the reference.

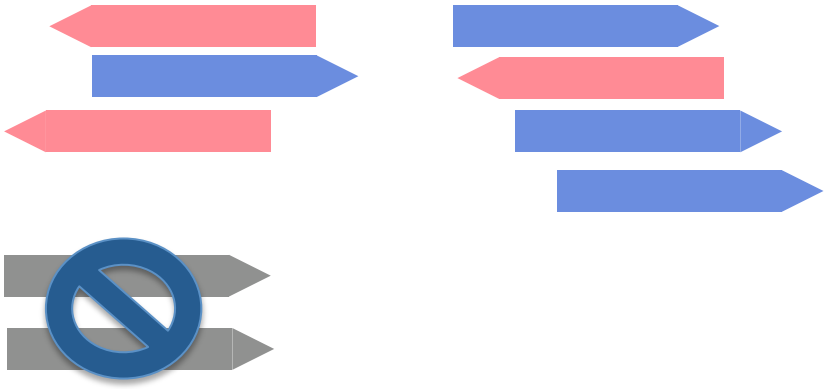
	Genome			Transcriptome		
Strain	**SNPs**	**Insertions**	**Deletions**	**SNPs**	**Insertions**	**Deletions**
A/J	4,198,324	401,264	422,424	104,358	7,846	8,394
129S1/SvImJ	4,458,004	428,081	458,055	109,598	8,154	8,875
NOD/ShiLtJ	4,323,530	389,285	407,801	108,881	7,599	8,168
NZO/HlLtJ	4,492,372	396,393	410,118	108,026	7,551	7,905
CAST/EiJ	17,673,726	1,359,607	1,367,482	410,805	26,975	27,474
PWK/PhJ	17,202,436	1,247,627	1,388,258	411,647	25,226	27,842
WSB/EiJ	6,045,573	588,061	608,945	146,495	10,966	11,559
All Strains	31,593,523	2,963,385	3,213,340	746,993	56,354	61,204

To what degree do these differences affect alignment of RNA-Seq reads and gene abundance estimates?



Align to B6 Transcriptome

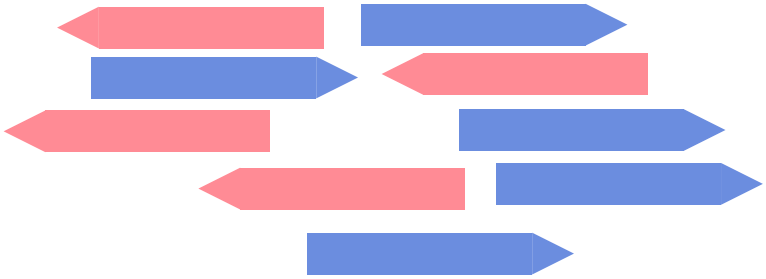
5’-ATCGGCGTCTTACATTAGCTCAAGGGTGCC-3’



Align to CAST Pseudotranscriptome

5’-ATCGGCGTCTTA~~CATTA~~GCTCAAGGGTGCC-3’

5’-ATCGGCGTCTTGCTCAAGGGTGCC-3’

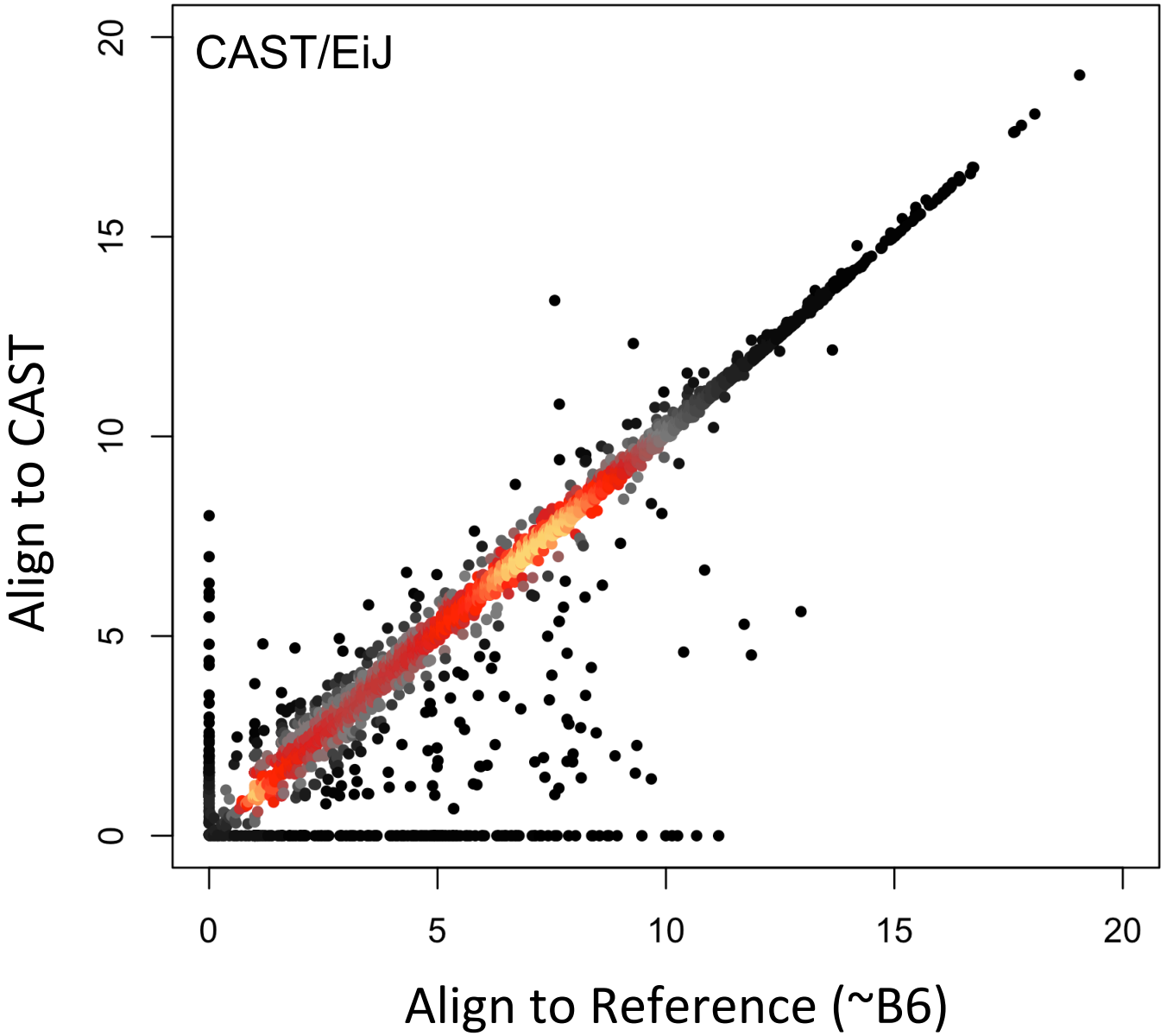


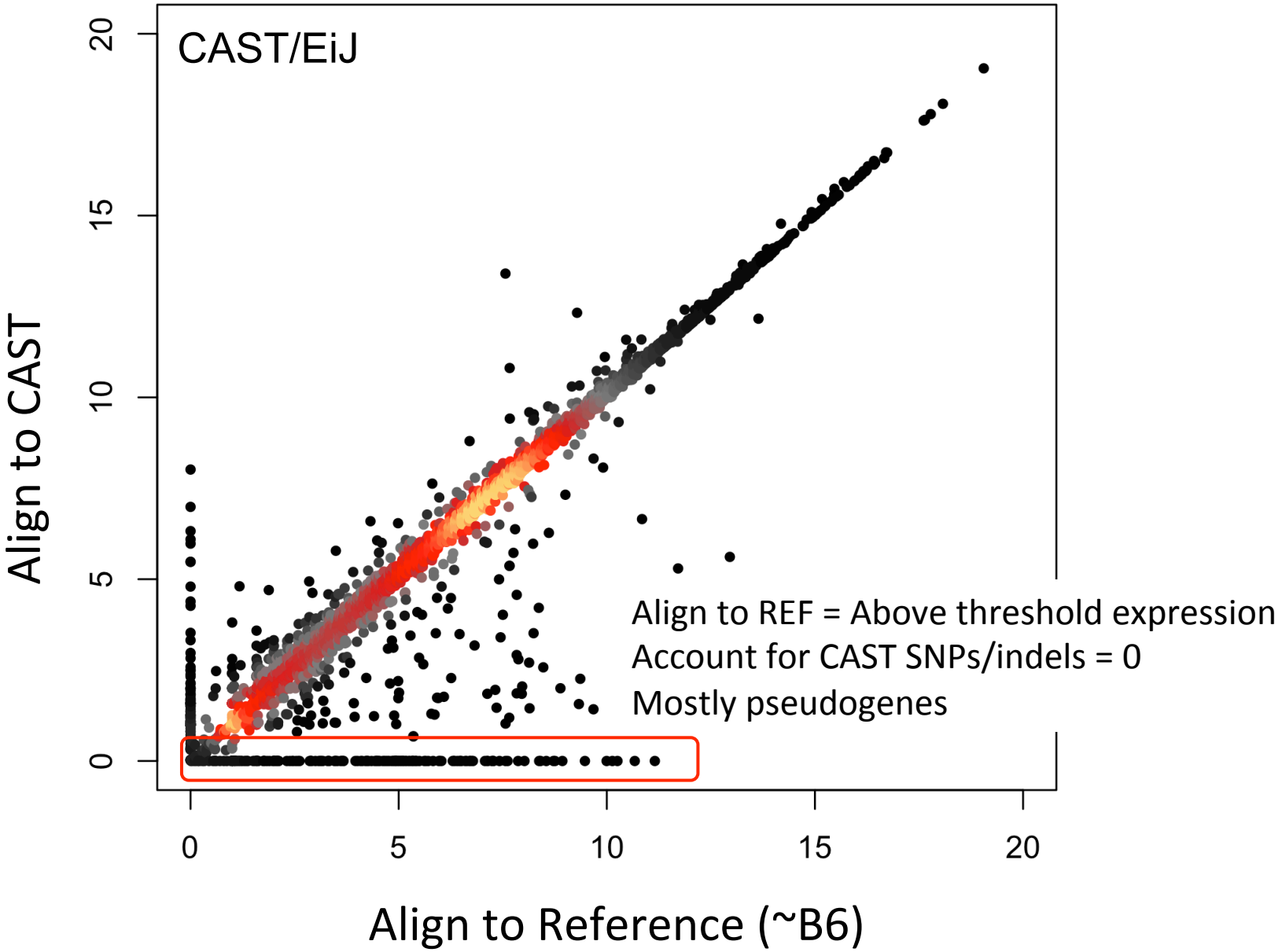
Compare Results

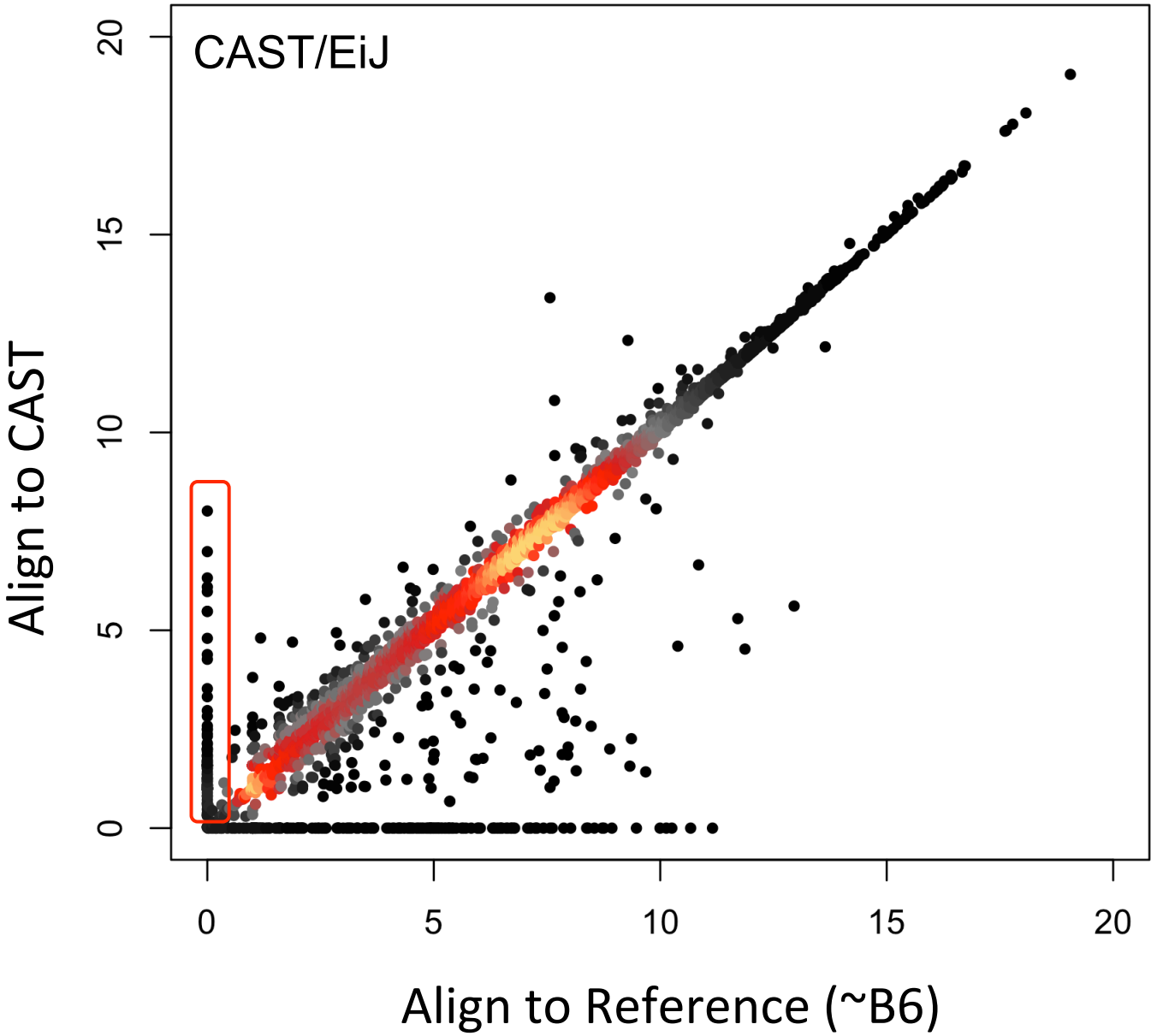
More CAST reads align with fewer mismatches to the CAST transcript sequences.

	Aligned to REF	Aligned to CAST
Total Reads	11,795,344	11,795,344
Reads with valid alignments (≤ 3 mismatches)	8,832,341 (74.9%)	9,085,246 (77.0%)
Difference		+252,905 (2.1%)
Reads with perfect matches (zero mismatches)	4,201,180 (35.6%)	5,183,409 (43.9%)
Difference		+982,229 (8.3%)
Total valid alignments	45,607,883	46,131,288

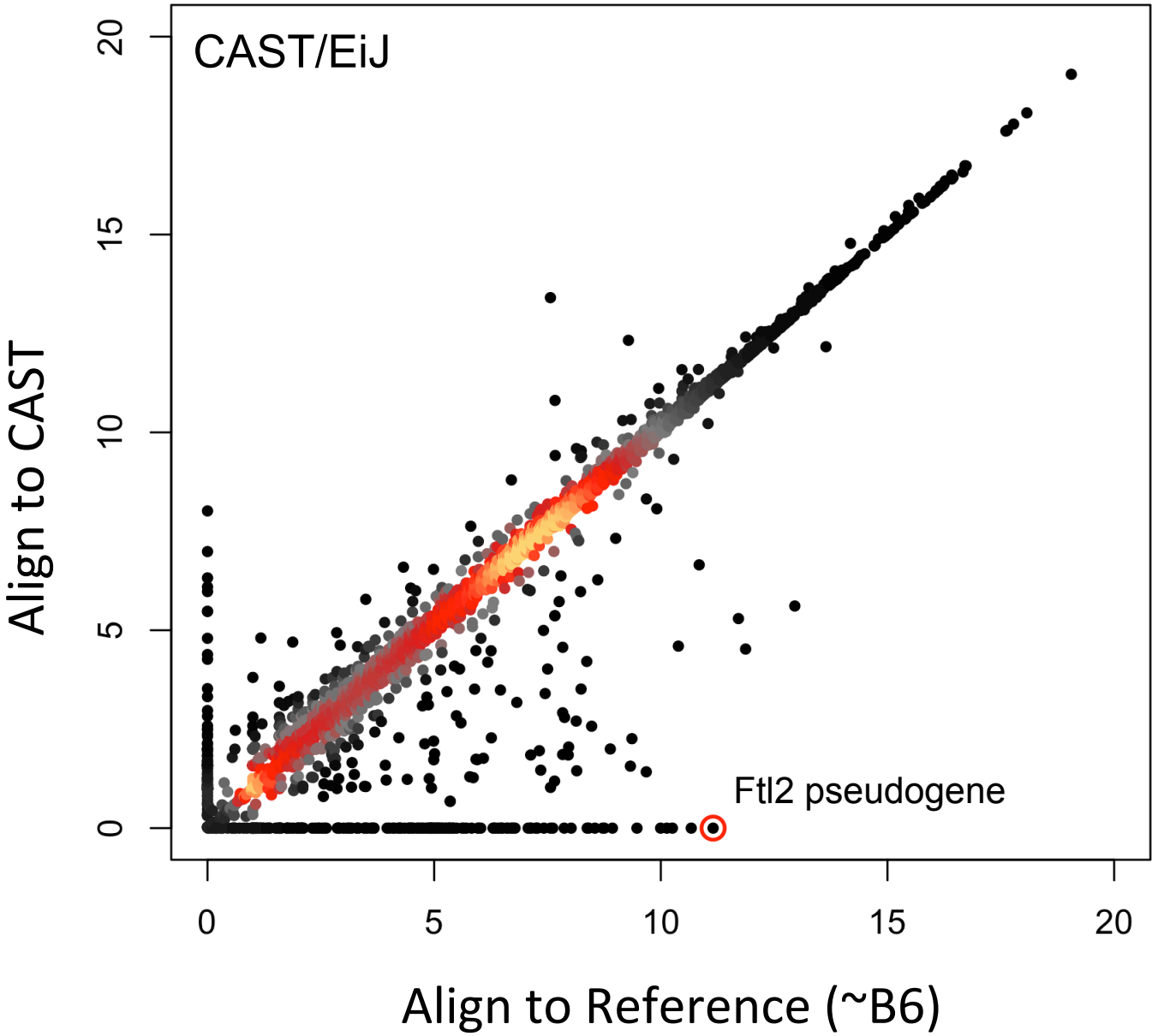
How are gene expression estimates affected by alignment to REF or CAST?

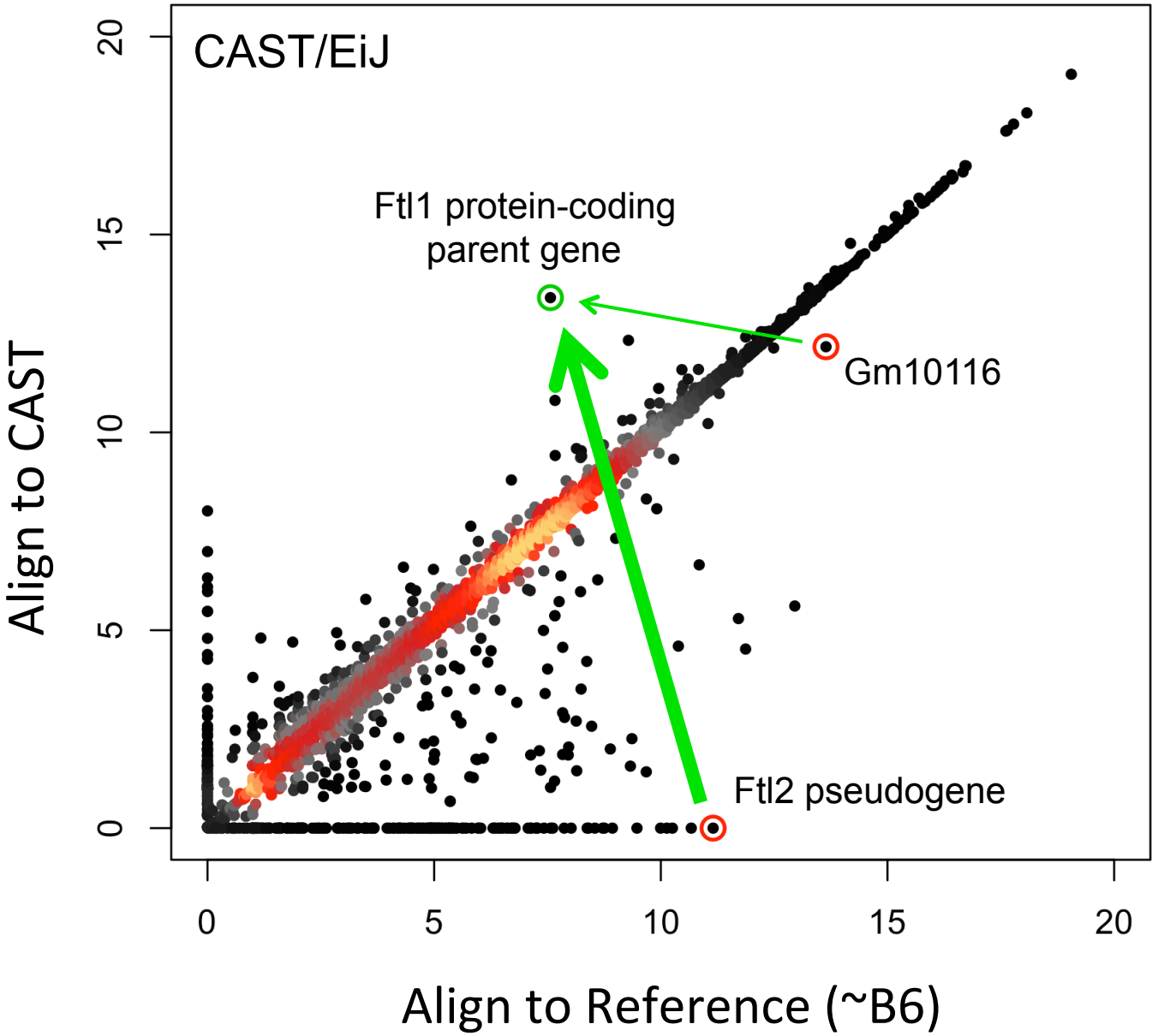



CAST/EiJ
Align to CAST
Align to Reference (~B6)
0
5
10
15
20
Align to REF = Above threshold expression
Account for CAST SNPs/indels = 0
Mostly pseudogenes

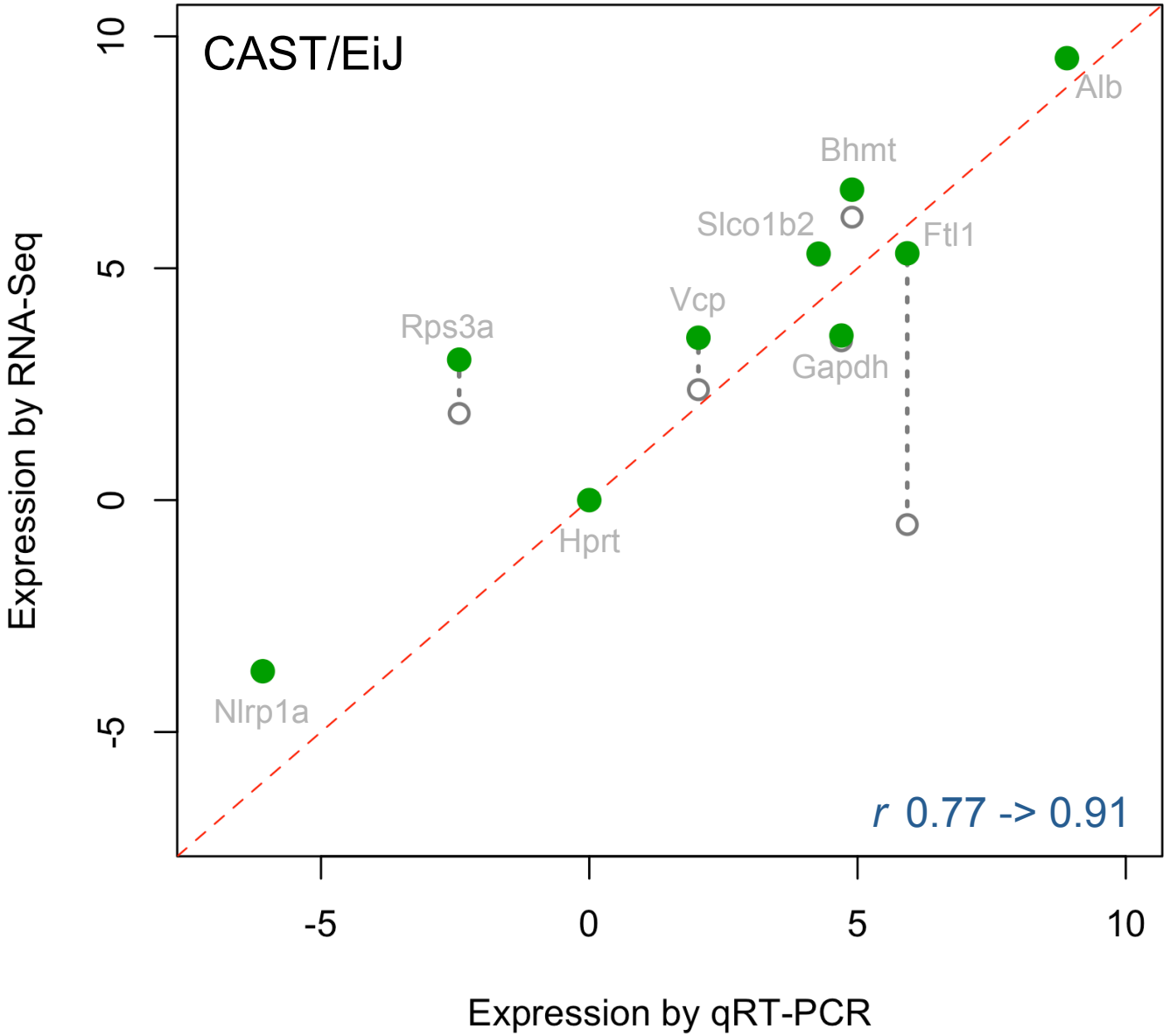

CAST/EiJ
Align to CAST
Align to Reference (~B6)
0
5
10
15
20

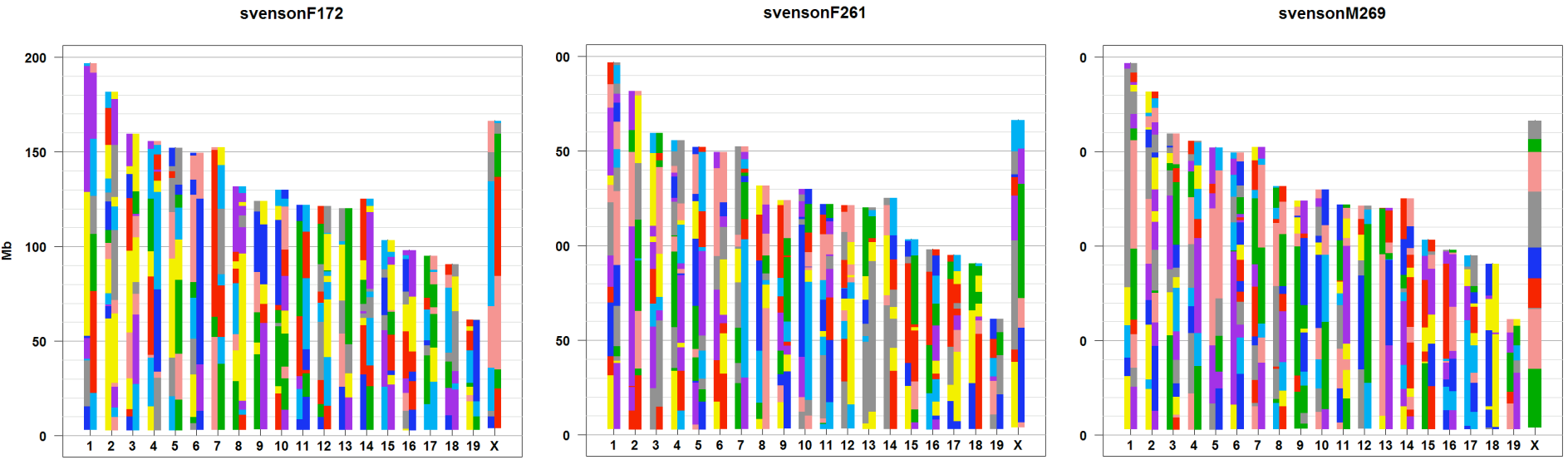
Where do the pseudogene reads go?



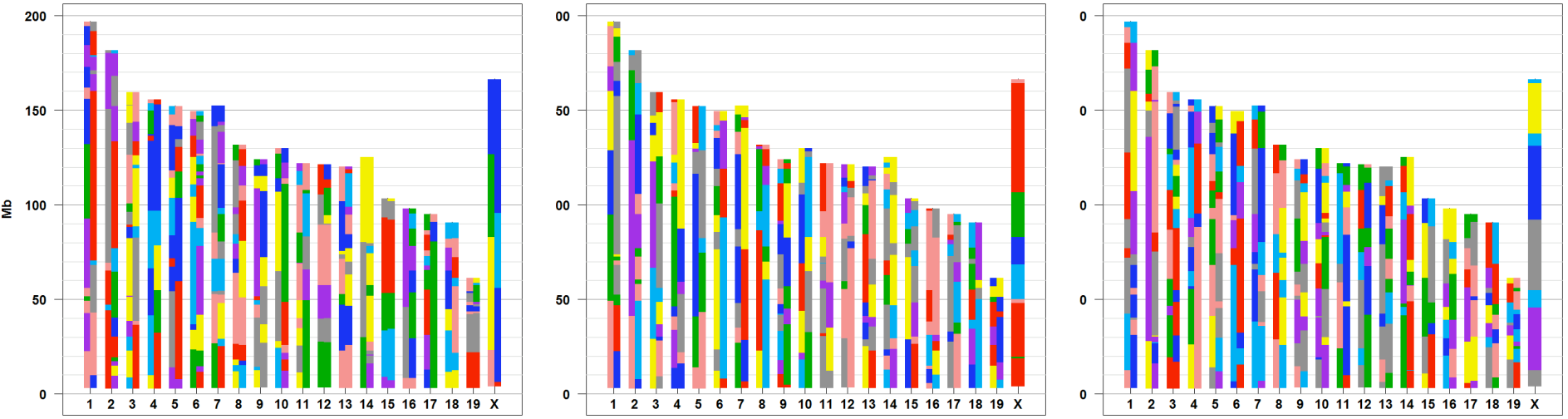

CAST/EiJ
Ftl1 protein-coding
parent gene
Gm10116
Ftl2 pseudogene
Align to CAST
Align to Reference (~B6)
0
5
10
15
20

Alignment to CAST pseudotranscriptome improves gene abundance estimates.



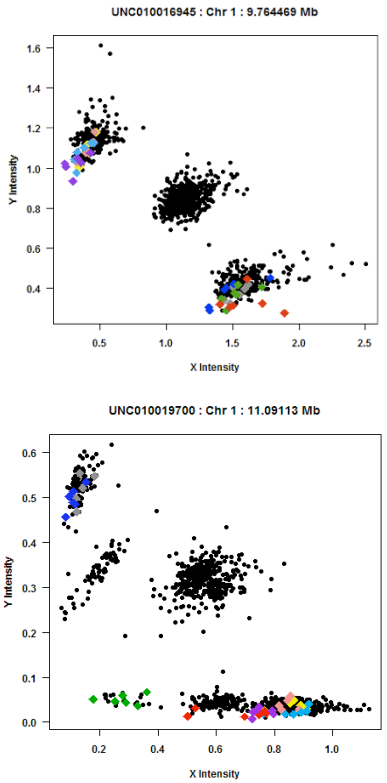


But how do you align RNA-seq reads from outbred mice? Or from humans for that matter?



Building an individualized diploid transcriptome.

Genotype DNA @ 8-80k informative SNPs



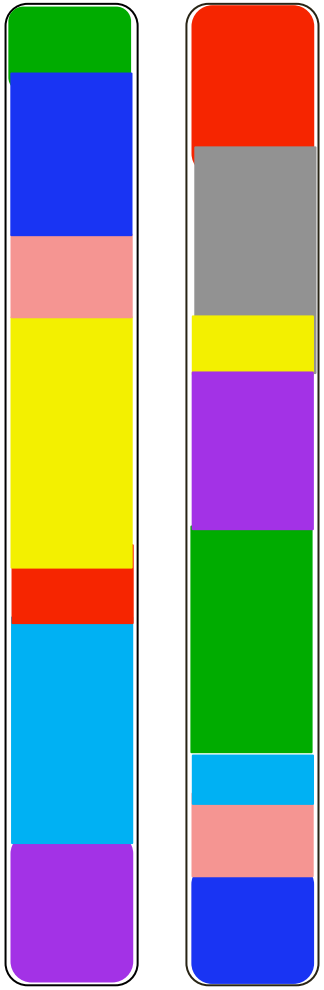
Infer 36-state genotypes

FG
FG
DG
DG
DB
DB
DB
CB
CB
AA
AA
AH
AH
AH
AH
AF
GF
GF
EF
EF
EF
EE
EE
EC
HC
HC
HD
HD

Psuedo-phased Chromosomes

F	G
F	G
D	G
D	G
D	B
D	B
D	B
C	B
C	B
A	A
A	A
A	H
A	H
A	H
A	H
A	F
G	F
G	F
E	F
E	F
E	F
E	E
E	E
E	C
H	C
H	C
H	D
H	D

Two Haploid Chromosomes

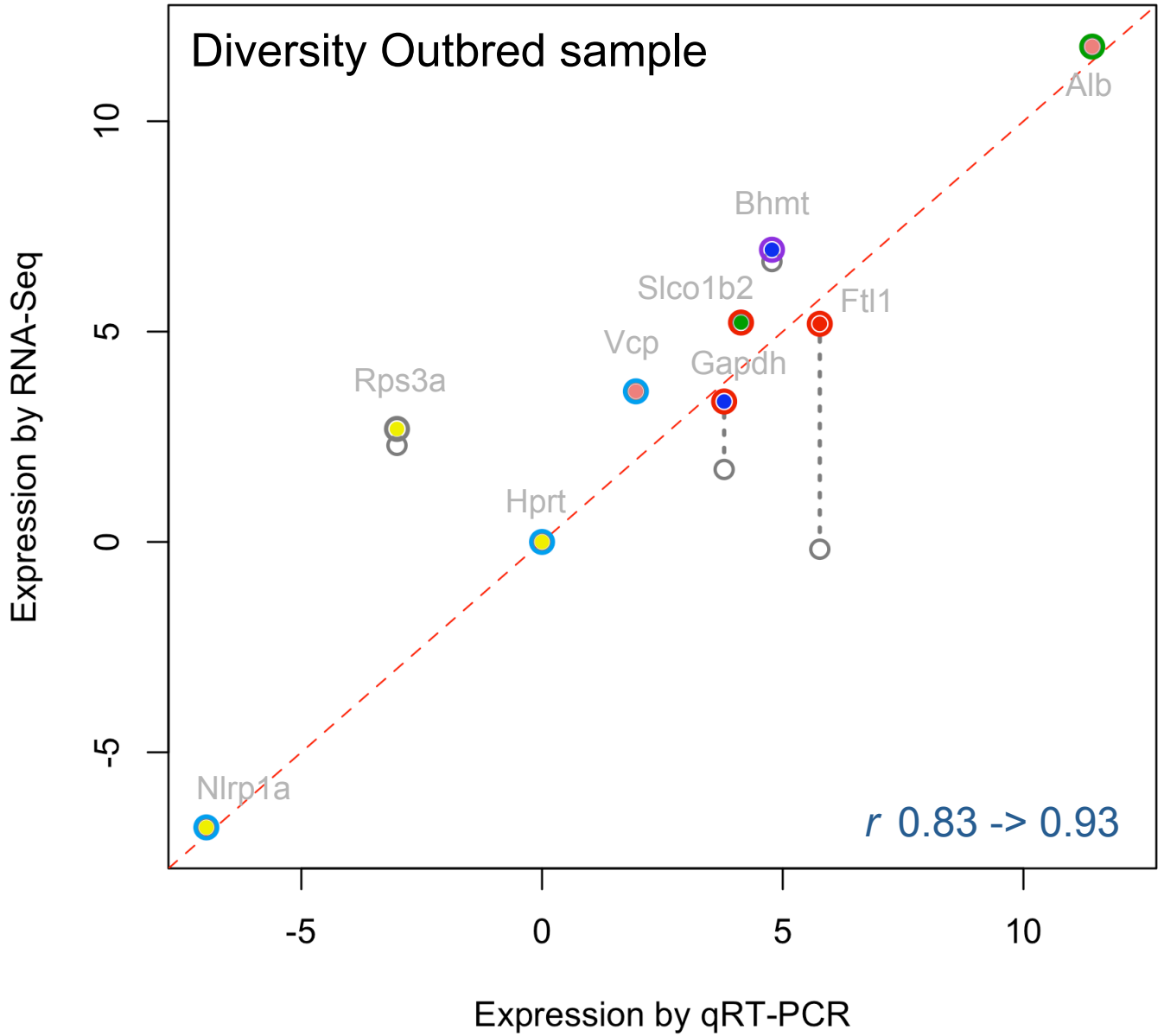


Munger et al., *submitted*

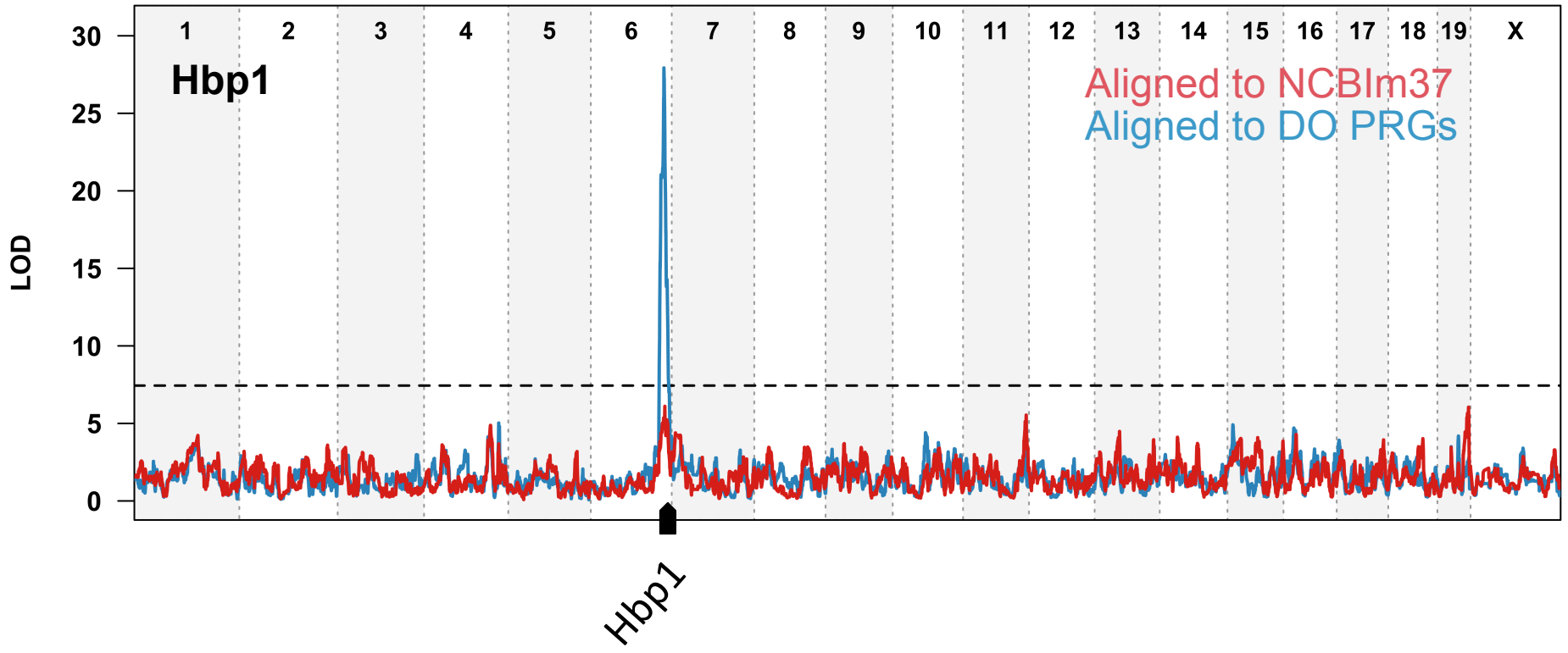
More DO reads align more uniquely to the individualized DO transcript sequences.

Sample F66	Aligned to REF	Aligned to Individ
Total Reads	13,053,816	13,053,816
Reads with valid alignments (≤ 3 mismatches)	10,835,301 (83.0%)	10,989,682 (84.2%)
Difference		+154,381 (1.2%)
Reads with perfect matches (zero mismatches)	7,143,400 (54.7%)	7,983,351 (61.2%)
Difference		+839,951 (6.5%)
Total valid alignments	57,872,246 (scaled)	54,516,518

Alignment to individualized transcriptomes improves gene abundance estimates from outbred mice.



Downstream analyses like expression QTL mapping are highly sensitive to alignment strategy.



Conclusions

- RNA-seq enables us to ask new questions in genetically diverse populations/species.
  - Increased information = Increased analytical complexity.
- Individualized reference genomes improve read alignment and gene abundance estimates.
  - Alignment to the common reference masks local genetic associations for many genes.
- Avoid the “one reference aligns all” approach.
- There is still room for improvement.

Collaborative Science @ JAX

- Gary Churchill
- Elissa Chesler
- Al Simons
- Narayanan Raghupathy
- KB Choi
- Dan Gatti
- Joel Graber
- Karen Svenson, Steve Ciciotte, Lisa Somes, et al.
- Doug Hinerfeld/ Sandy Daigle/ Sonya Kamdar/ Gene Expression Service

Thank you!