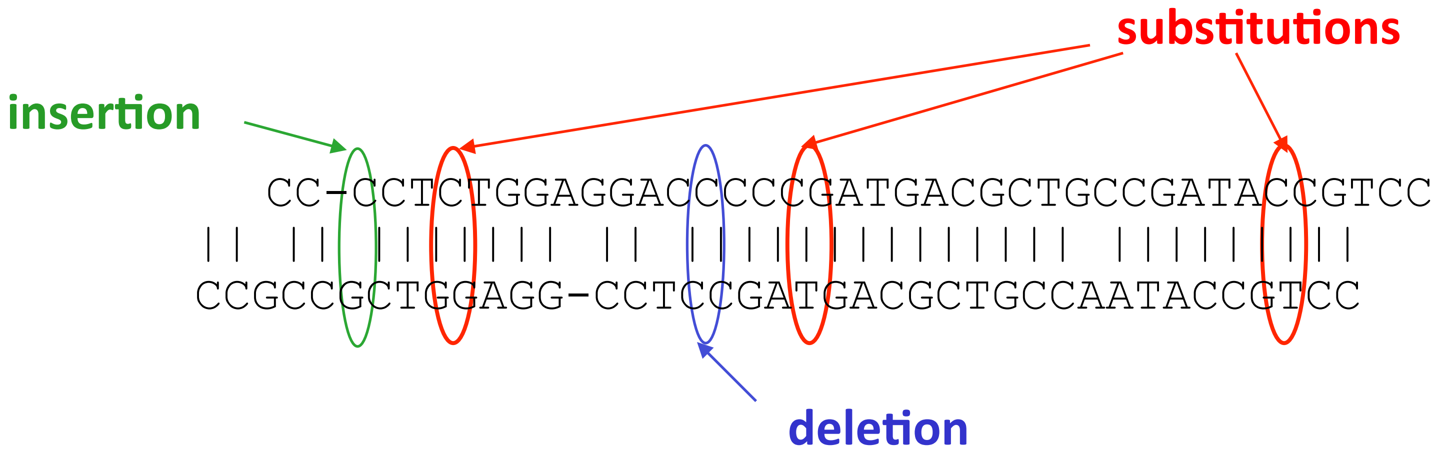
Alignments

- Sequence a fragment of the gene (RNA) or genomic region (DNA), then map (align) it to the genome
- Alignment = a mapping between the letters of the two sequences, with some spacers (indels)
- The alignment will take into account differences such as polymorphisms and sequencing errors, and introns (for genes)



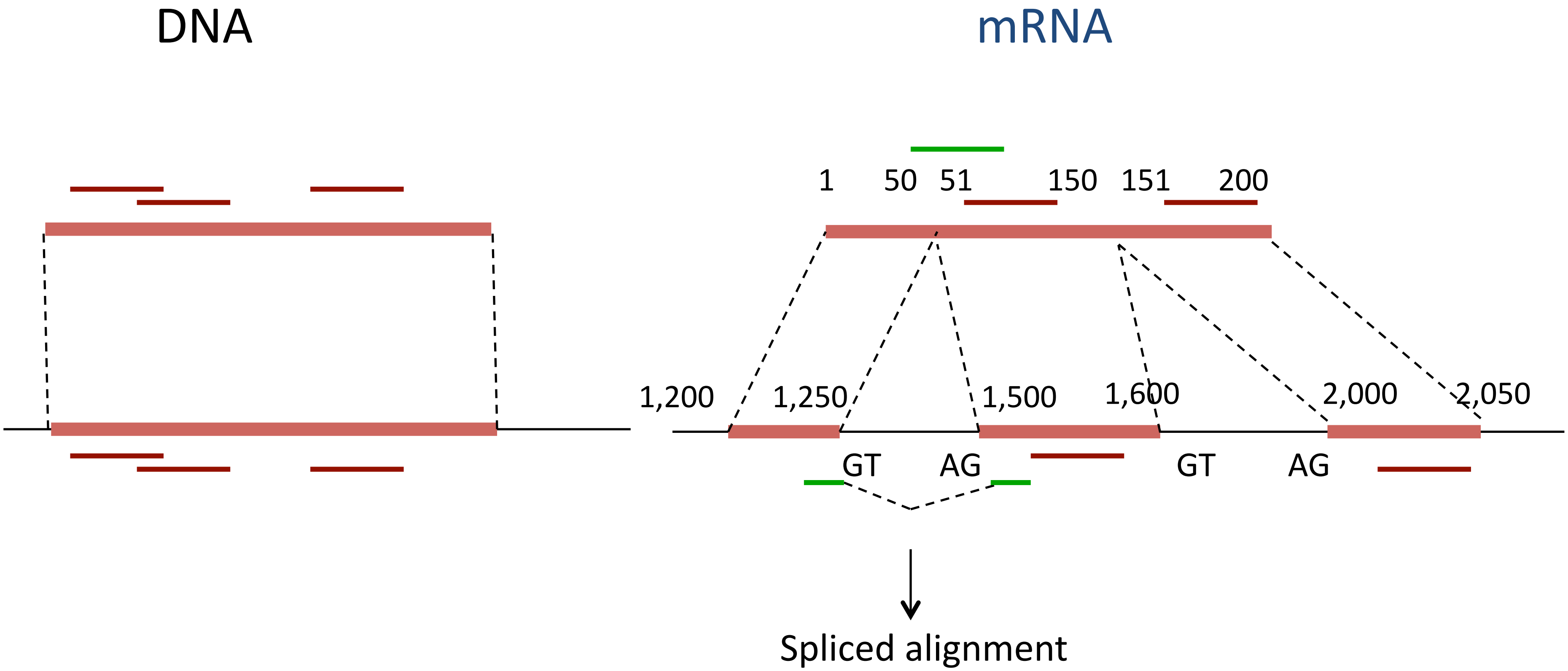
substitutions

insertion

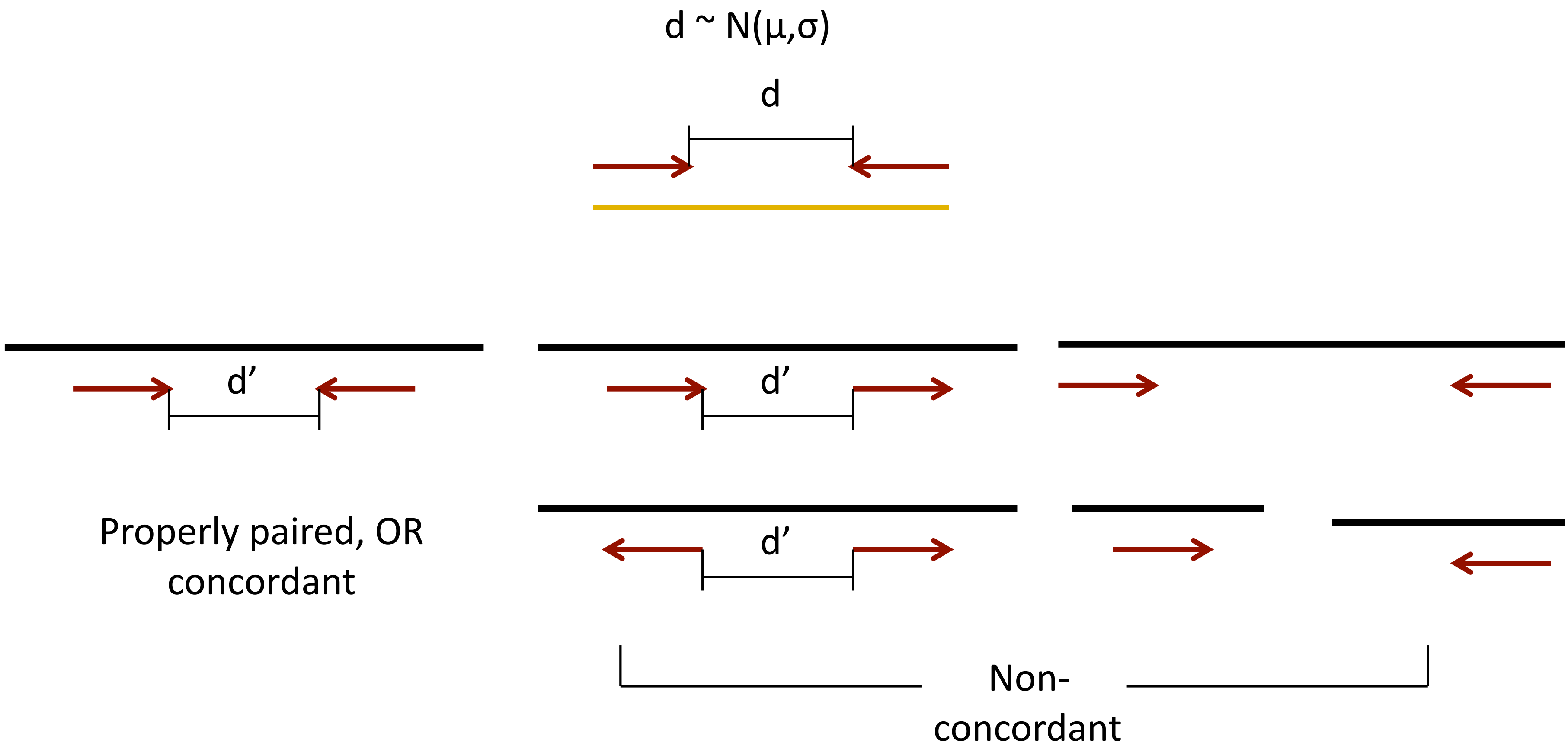
```
   CC-CCTCTGGAGGACCCCCGATGACGCTGCCGATACCGTCC
|| || ||||||| || |||||||||||||| |||||||||
CCGCCGCTGGAGG-CCTCCGATGACGCTGCCAATACCGTCC
```

deletion

Alignments


DNA
mRNA
1
50
51
150
151
200
1,200
1,250
1,500
1,600
2,000
2,050
GT
AG
GT
AG
Spliced alignment

NGS Alignments



Representation: SAM/BAM format

Header

```
@HD VN:1.0   SO:coordinate
@SQ SN:chr1  LN:248956422
@SQ SN:chr10 LN:133797422
@SQ SN:chr11 LN:135086622
…
@PG ID:TopHat VN:2.0.13 CL:/
data1/igm3/sw/packages/
tophat-2.0.13.Linux_x86_64/
tophat -p 8 -o …
```

Alignments

```
141217_CIDR4_0073_BHCFG7ADXX:2:1111:3128:29074   345
chr1  10021   0  68M  * ACCCTAA...CCCTAAC  @DC?=2...DDDD@?@
AS:i:0 XN:i:0 XM:i:0   XO:i:0 XG:i:0 NM:i:0   MD:Z:68 YT:Z:UU
NH:i:10   CC:Z:chr10   CP:i:10004   XS:A:- HI:i:0

. . .
```

Representation: SAM/BAM format

`141217_CIDR4_0073_BHCFG7ADXX:2:1111:3128:29074`	Read id
`99`	**FLAG**
`chr1`	Chr
`10021`	Start
`0`	Mapping quality
`50M`	**CIGAR** (alignment)
`=`	Mate chr
`10151`	Mate start
`180`	Mate dist
`ACCCTAACCCTAACCCTAACCCTAACCCTAACCCTAACCCTAACCCTAAC`	Query seq
`@DC?=2.FFGE@7>C62>BGABGB9HFBAFIIHEGFIIIHFAIIGDA<FC`	Query base quals
`AS:i:0`	Alignment score
`NM:i:0`	Edit distance to reference
`NH:i:10`	Number of hits
`XS:A:-`	Strand
`HI:i:0`	Hit index for this alignment

Tags: [A-Za-z][A-Za-z]:[AifZH]:.*
where A =character; i = integer; f = float; Z=string; H = hex string