

Sequence alignment.

Two general human i.d. alignment situations

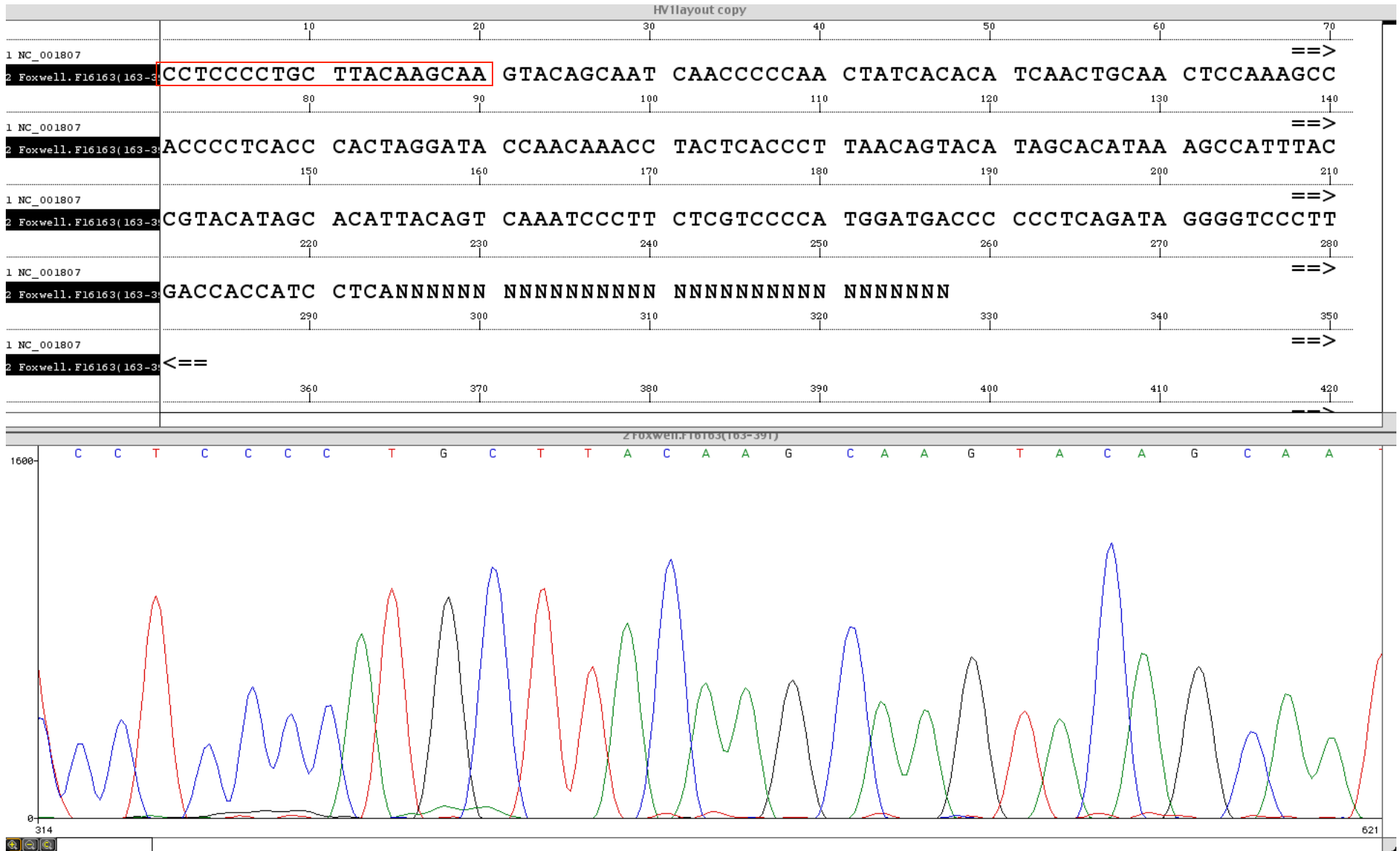
- Aligning overlapping data from the same specimen
- Aligning your sample sequence with the CRS.

Sequence editing and alignment can be done in a variety of ways.

This can be mostly manual or very automated. Forensic analysis tends to be more manual.



Initial manual editing. The letters above show the sequence as automatically called by the analytical software. The actual data, the electropherogram, for a selected region is shown below. A text file of the CRS is out of view in this layout.



After cropping 5' end. What else is wrong with the sequence as called by the computer?

		16090	16100	16110	16120
1 NC_001807		ACCGCTATGT	ATTTCGTACA	TTACTGCCAG	CCACCATGAA
2 Foxwell.F16163(163-3)					==>
		16130	16140	16150	16160
1 NC_001807		TATTGTACGG	TACCATAAAT	ACTTGACCAC	CTGTAGTACA
2 Foxwell.F16163(163-3)					==>
		16170	16180	16190	16200
1 NC_001807		TAAAAACCCA	ATCCACATCA	AAACCCCTC	CCCATGCTTA
2 Foxwell.F16163(163-3)				CCTC	CCCATGCTTA
		16210	16220	16230	16240
1 NC_001807		CAAGCAAGTA	CAGCAATCAA	CCCTCAACTA	TCACACATCA
2 Foxwell.F16163(163-3)		CAAGCAAGTA	CAGCAATCAA	CCCCCAACTA	TCACACATCA
		16250	16260	16270	16280
1 NC_001807		ACTGCAACTC	CAAAGCCACC	CCTCACCCAC	TAGGATACCA
2 Foxwell.F16163(163-3)		ACTGCAACTC	CAAAGCCACC	CCCTCACCCA	CTAGGATACC
		16290	16300	16310	16320
1 NC_001807		ACAAACCTAC	CCACCCTTAA	CAGTACATAG	TACATAAAGC
2 Foxwell.F16163(163-3)		AACAAACCTA	CTCACCTTA	ACAGTACATA	GCACATAAAG
		16330	16340	16350	16360
1 NC_001807		CATTTACCGT	ACATAGCACA	TTACAGTCAA	ATCCCTTCTC
2 Foxwell.F16163(163-3)		CCATTTACCG	TACATAGCAC	ATTACAGTCA	AATCCCTTCT
		16370	16380	16390	16400
1 NC_001807		GTCCCCATGG	ATGACCCCCC	TCAGATAGGG	GTCCCTTGAC
2 Foxwell.F16163(163-3)		CGTCCCCATG	GATGACCCCC	CTCAGATAGG	GGTCCCTTGA
		16410	16420	16430	16440
1 NC_001807		CACCATCCTC	CGTGAAATCA	ATATCCCGCA	CAAGAGTGCT
2 Foxwell.F16163(163-3)		CCACCATCCT	CA		

The edited data file is then moved to align it with the reference sequence. This establishes base position # and allows comparison between sequences. However an indel can throw off alignment.

		16090	16100	16110	16120
1 NC_001807		ACCGCTATGT	ATTTTCGTACA	T TACTGCCAG	CCACCATGAA
2 Foxwell.F16163(163-3)					==>
		16130	16140	16150	16160
1 NC_001807		TATTGTACGG	TACCATAAAT	ACTTGACCAC	CTGTAGTACA
2 Foxwell.F16163(163-3)					==>
		16170	16180	16190	16200
1 NC_001807		TAAAAACCCA	ATCCACATCA	AAACCCCCTC	CCCATGCTTA
2 Foxwell.F16163(163-3)				CCTC	CCCATGCTTA
		16210	16220	16230	16240
1 NC_001807		CAAGCAAGTA	CAGCAATCAA	CCCTCAACTA	TCACACATCA
2 Foxwell.F16163(163-3)		CAAGCAAGTA	CAGCAATCAA	CCCCCAACTA	TCACACATCA
		16250	16260	16270	16280
1 NC_001807		ACTGCAACTC	CAAAGCCACC	CCTCACCCAC	TAGGATACCA
2 Foxwell.F16163(163-3)		ACTGCAACTC	CAAAGCCACC	C ² TCACCCAC	TAGGATACCA
		16290	16300	16310	16320
1 NC_001807		ACAAACCTAC	CCACCCTTAA	CAGTACATAG	TACATAAAGC
2 Foxwell.F16163(163-3)		ACAAACCTAC	TCACCCTTAA	CAGTACATAG	CACATAAAGC
		16330	16340	16350	16360
1 NC_001807		CATTTACCGT	ACATAGCACA	TTACAGTCAA	ATCCCTTCTC
2 Foxwell.F16163(163-3)		CATTTACCGT	ACATAGCACA	TTACAGTCAA	ATCCCTTCTC
		16370	16380	16390	16400
1 NC_001807		GTCCCCATGG	ATGACCCCCC	TCAGATAGGG	GTCCCTTGAC
2 Foxwell.F16163(163-3)		GTCCCCATGG	ATGACCCCCC	TCAGATAGGG	GTCCCTTGAC
		16410	16420	16430	16440
1 NC_001807		CACCATCCTC	CGTGAAATCA	ATATCCCGCA	CAAGAGTGCT
2 Foxwell.F16163(163-3)		CACCATCCTC	A		

The Wells system for maintaining alignment faced with an insertion relative to the CRS. However you do it, the insertion is recorded with an extra decimal for the position. This sequence has a C at position 16262.1. By convention, it is located as far 3' as possible.

		key						
		16120	16130	16140	16150	16160	16170	16180
4 PRIMERS					tgacca	cctgtagtac	ataa	
1 NC_001807	GCCACCATGA ATATTGTACG GTACCATAAA TACTTGACCA CCTGTAGTAC ATAAAAACCC AATCCACATC							==>
2 Foxwell.F16163(163-3)								
3 Foxwell.R16391(163-3)					Ttgacca	cctgtagtac	ataaAAACCC	AATCCACATC
7	-----							-----
		16190	16200	16210	16220	16230	16240	16250
4 PRIMERS								
1 NC_001807	AAAACCCCCT CCCCATGCTT ACAAGCAAGT ACAGCAATCA ACCCTCAACT ATCACACATC AACTGCAACT							
2 Foxwell.F16163(163-3)	CCT CCCCATGCTT ACAAGCAAGT ACAGCAATCA ACCCCCAACT ATCACACATC AACTGCAACT							
3 Foxwell.R16391(163-3)	AAAACCCCCT CCCCATGCTT ACAAGCAAGT ACAGCAATCA ACCCCCAACT ATCACACATC AACTGCAACT							
7	-----							-----
		16260	16270	16280	16290	16300	16310	16320
4 PRIMERS								
1 NC_001807	CCAAAGCCAC CCCTCACCCA CTAGGATACC AACAAACCTA CCCACCCTTA ACAGTACATA GTACATAAAG							
2 Foxwell.F16163(163-3)	CCAAAGCCAC CCCTCACCCA CTAGGATACC AACAAACCTA CTCACCCTTA ACAGTACATA GCACATAAAG							
3 Foxwell.R16391(163-3)	CCAAAGCCAC CCCTCACCCA CTAGGATACC AACAAACCTA CTCACCCTTA ACAGTACATA GCACATAAAG							
7	-----							-----
		16330	16340	16350	16360	16370	16380	16390
4 PRIMERS								
1 NC_001807	CCATTTACCG TACATAGCAC ATTACAGTCA AATCCCTTCT CGTCCCCATG GATGACCCCC CTCAGATAGG							
2 Foxwell.F16163(163-3)	CCATTTACCG TACATAGCAC ATTACAGTCA AATCCCTTCT CGTCCCCATG GATGACCCCC CTCAGATAGG							
3 Foxwell.R16391(163-3)	CCATTTACCG TACATAGCAC ATTACAGTCA AATCCCTTCT CGTCCCCATG GATGACCCCC							
7	-----							-----
		16400	16410	16420	16430	16440	16450	16460
4 PRIMERS								
1 NC_001807	gtcccttga ccaccatcct c							
2 Foxwell.F16163(163-3)	GGTCCCTTGA CCACCATCCT CCGTGAAATC AATATCCCGC ACAAGAGTGC TACTCTCCTC GCTCCGGGCC							
3 Foxwell.R16391(163-3)	Ggtcccttga ccaccatcct cA							
7	<==							-----
	-----							-----

One way to finish the layout. The forward sequence and the reverse complement of the reverse sequence are edited and aligned. Primer sequences have been identified. The program has automatically found differences with the CRS.

There should be no conflict
between overlapping sequences
generated from the same DNA
extraction.

		key						
		16120	16130	16140	16150	16160	16170	16180
4 PRIMERS					tgacca	cctgtagtac	ataa	
1 NC_001807	GCCACCATGA ATATTGTACG GTACCATAAA TACTTGACCA CCTGTAGTAC ATAAAAACCC AATCCACATC							
2 Foxwell.F16163(163-3)								==>
3 Foxwell.R16391(163-3)					Ttgacca	cctgtagtac	ataaAAACCC	AATCCACATC
7	-----							-----
		16190	16200	16210	16220	16230	16240	16250
4 PRIMERS								
1 NC_001807	AAAACCCCCT CCCCATGCTT ACAAGCAAGT ACAGCAATCA ACCCTCAACT ATCACACATC AACTGCAACT							
2 Foxwell.F16163(163-3)		CCT	CCCCATGCTT	ACAAGCAAGT	ACAGCAATCA	ACCCCAACT	ATCACACATC	AACTGCAACT
3 Foxwell.R16391(163-3)	AAAACCCCCT CCCCATGCTT ACAAGCAAGT ACAGCAATCA ACCCCCAACT ATCACACATC AACTGCAACT							
7	-----							-----
		16260	16270	16280	16290	16300	16310	16320
4 PRIMERS								
1 NC_001807	CCAAAGCCAC CCCTCACCCA CTAGGATACC AACAAACCTA CCCACCCTTA ACAGTACATA GTACATAAAG							
2 Foxwell.F16163(163-3)	CCAAAGCCAC CCCTCACCCA CTAGGATACC AACAAACCTA CTCACCCTTA ACAGTACATA GCACATAAAG							
3 Foxwell.R16391(163-3)	CCAAAGCCAC CCCTCACCCA CTAGGATACC AACAAACCTA CTCACCCTTA ACAGTACATA GCACATAAAG							
7	-----							-----
		16330	16340	16350	16360	16370	16380	16390
4 PRIMERS								
1 NC_001807	CCATTTACCG TACATAGCAC ATTACAGTCA AATCCCTTCT CGTCCCCATG GATGACCCCC CTCAGATAGG							
2 Foxwell.F16163(163-3)	CCATTTACCG TACATAGCAC ATTACAGTCA AATCCCTTCT CGTCCCCATG GATGACCCCC CTCAGATAGG							
3 Foxwell.R16391(163-3)	CCATTTACCG TACATAGCAC ATTACAGTCA AATCCCTTCT CGTCCCCATG GATGACCCCC							
7	-----							-----
		16400	16410	16420	16430	16440	16450	16460
4 PRIMERS								
1 NC_001807	gtcccttga ccaccatcct c							
2 Foxwell.F16163(163-3)	GGTCCCTTGA CCACCATCCT CCGTGAAATC AATATCCCGC ACAAGAGTGC TACTCTCCTC GCTCCGGGCC							
3 Foxwell.R16391(163-3)	Ggtcccttga ccaccatcct cA							
7	<==							-----

- Which sequence is forward and which is reverse?
- What is the explanation for that most 3' polymorphism?

Length heteroplasmy relative to the Cambridge Reference Sequence has led to database entry discordance. The FBI has advocated rules for promoting consistency.

The common problem occurs when the sample has a deletion relative to the CRS. See Appendix 1 in Budowle et al.

A standard system of alignment with the CRS is necessary for a database search to be meaningful.

How should these be aligned?

Evidence: ATTGATGTC

CRS: ATTGAATGTC

Evidence: ATTG–ATGTC
CRS: ATTGAATGTC

 or

Evidence: ATTGA–TGTC
CRS: ATTGAATGTC

Example: Two alternative gap placements. Each implies an indel, but no other mutation hypothesis required.

Recommended standards.

1. Use the least possible number of hypothesized mutations (gap at one site or a point mutation).
2. When choosing between two scenarios with the same number of mutations, prefer a gap over a transition over a transversion.
3. An indel should be located as far 3' as possible on the light strand, and if possible indels should be combined.
4. Gaps should be combined only if this doesn't increase the total number of mutations.

Evidence: ATTG-ATGTC
CRS: ATTGAATGTC
or

Evidence: ATTGA-TGTC
CRS: ATTGAATGTC



★=the preferred alignment

Evidence: AAAACCTCCCCCATGCT
CRS: AAAACCCCCTCCCCATGCT

Evidence: AAAACCTCC—CCCCATGCT
CRS: AAAACCCCCTCCCCATGCT
gap + one transition



Evidence: AAAACC—TCCCCCATGCT
CRS: AAAACCCCCTCCCCATGCT
gap + two transitions

Evidence: AAAACCTCCCCC—ATGCT
CRS: AAAACCCCCTCCCCATGCT
gap + two transitions