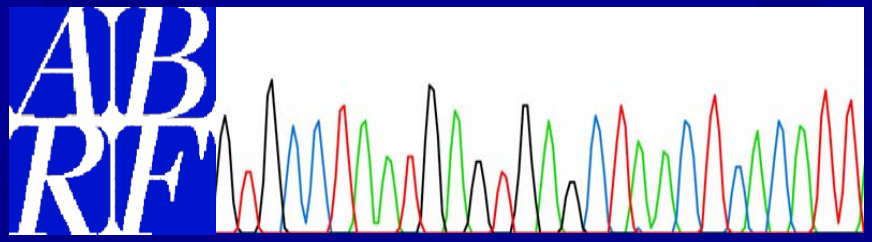
Sequencing of Difficult DNA Templates: The 2007/2008 DSRG Difficult Template Sequencing Study

February 12, 2008
DSRG/ABRF Meeting
Salt Lake City, UT

Jan Kieleczawa, Debbie Adam, Doug Bintzler, David Needleman, Sushmita Singh, Robert Steen, Michael Zianni, Peter Schweitzer, and Michelle Detwiler



DNA Sequencing Research Group

Sequencing of Difficult DNA Templates: Study Design

Invitation to Participate in the 2008 DSRG Difficult Template Study:

The DSRG has designed a study to identify a general set of guidelines that constitute the best approaches for sequencing difficult templates. This is a continuation of previous DSRG research group studies performed in 1998, 2000, and 2003.

We will be distributing a set of 8 difficult templates (with a variety of difficult regions to sequence) to participating labs with the goal of collecting the electropherograms and the conditions and formulations used by individual labs (Phase I). When these data have been collected and analyzed, an optimal protocol for each category will be selected. Then, this information will be distributed to participating labs to allow a second round of sequencing to evaluate the general applicability of the optimized protocols. Results from this second round will be collected and and analyzed (Phase II).

The results of this study will be presented at the ABRF annual meeting in Salt Lake City.

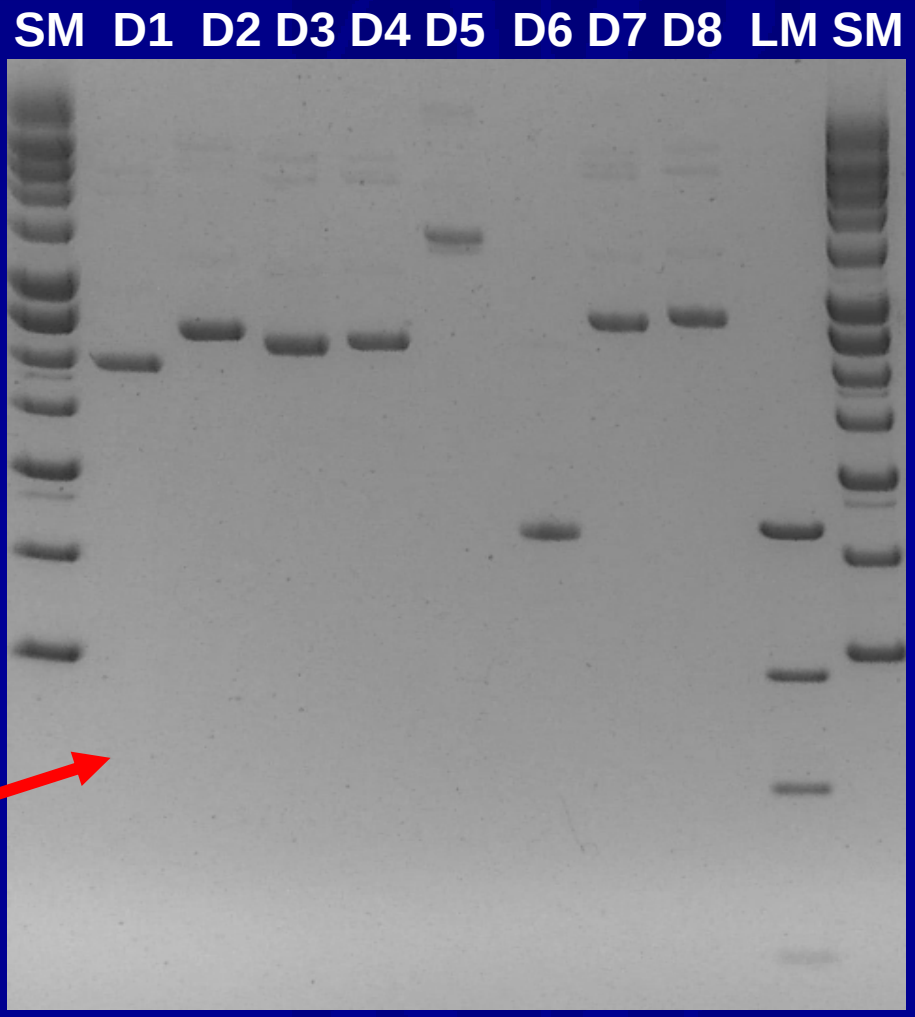
If you wish to participate, please send an email with contact information and a shipping address for your laboratory to: jkieleczawa@wyeth.com

Sample Preparation and Distribution

- **All 8 DNAs were transformed- (Electro competent TOP 10 cells from Invitrogen) - M. Mader**
- **Large scale Marligen's Maxi DNA preps (4 to 20 mg of DNA) -**

  K. Marquette
- **30 sets of DNAs, pGem3zf control and associated primers distributed and sent**
  - 100 µg of each DNA (enough to try > 30 conditions in triplicate)
  - 100 µl of 5 µM primers

  (720 tubes labeled) - E. Mazaika
- **Approximately 200 ng of each DNA run on 1% gel to check for quality/quantity**



Study Participation

Location :	USA	Canada	Europe	Australia
Sets sent/Phase I/Phase II :	27/19/10	2/2/2	0/0/0	1/0/0

Data Collection and Analysis

- Finished data from both Phase I and II were deposited at ftp://ftp.genetics.utah.edu (thanks to H. Escobar for providing this ftp) and/or emailed to JK
- Q (KB) > 20 values and signal strength were calculated using Sequence Scanner v1.0 from ABI
- All RL/SS data and associated information (ng of DNA/additives/PCR instrument/cleanups…) were assembled in Excel

Sequencing Ranges for Phase I and II

DNA #	Primer	DNA characteristics	Range for Readlength Q > 20		
			Phase I	Phase II Protocol 1	Phase II Protocol 2
1	F	94%GC over 200 bases/101 base non-repeat G/C	0-931	34-907	0-442
	R	90%GC over 200 bases/73% GC over next 400 bases	0-960	0-936	0-521
2	F	70% GC over 300 bases	0-1026	191-868	0-761
	R	78% GC over 150 bases	0-1040	0-1009	0-1068
3	F	24 base hairpin with T_m >95°C	0-891	0-889	263-903
	R	24 base hairpin with T_m >95°C	0-238	0	0
4	F	18 Cs/10 Cs, separated by 7 bases	0-767	0-531	0-252
	R	18 Gs/10 Gs, separated by 7 bases	0-830	0-674	0-602
5	F	456 base non-repeat T/C	0-569	0-601	0-611
	R	456 base non-repeat G/A	0-811	0-828	0-615
6	F	147 base non-repeat G/A	134-1093	0-1053	213-956
	R	147 base non-repeat T/C	0-1011	0-953	0-853
7	F	19 and 15 bases inverted repeats, followed by 19 Cs and 41 base non-repeat T/A	0-961	0-925	0-789
	R	19 and 15 bases inverted repeats, followed by 19 Gs and 41 base non-repeat A/T	0-910	0-746	0-607
8	F	Alu repeat + 22 base inverted repeat/84 base loop	0-1020	0-975	0-983
Control	F	pGem3zf control	683-1054	296-1001	0-1011

List of Best Sequencing Protocols From Phase I

Prot #	DNA ng	Primer µl/5µM	BD v3.1 µl	dGTP v3.0 µl	5X ABI Bffr µl	Additive/ µl	Rxn Vol µl	Preferred Cleanup	Cycling Protocol
1	200	1.0	1.5	0.5	0	Betaine/ 2.0	10	CleanSeq	[(96ºC/10sec)(50ºC/5sec)(60ºC/4min)] x35→4ºC/∞
2	100	1.0	2.0	1.0	0	Betaine/ 2.0	10	BDX	96ºC/1min→[(96ºC/10sec)(50ºC/5sec)(60ºC/4min)] x25 →4ºC/∞
3	200	0.5	1.0	0.1	0	Betaine/ 1.6	10	EtOH	95ºC/3min→98ºC/40sec→60ºC/4min→ [(98ºC/10sec)(60ºC/4min)] x24→4ºC/∞
4	300	1.0	3.0	1.0	0	DMSO/ 1.0	20	G-50 In-house #	96ºC/1min→[(96ºC/10sec)(50ºC/5sec)(60ºC/4min)] x30 →4ºC/∞
5	200	2.0	1.0	0	0	None	15	CleanSeq	95ºC/1min→[(98ºC/45sec)(50ºC/15sec)(60ºC/2.5min)] x39→4ºC/∞
6	200	1.0	1.0	0	1.5	Betaine/ 2.0	10	CleanSeq	[(96ºC/10sec)(50ºC/5sec)(60ºC/4min)] x35→4ºC/∞
7	200	1.0	1.5	0.5	0	None	10	CleanSeq	[(96ºC/10sec)(50ºC/5sec)(60ºC/4min)] x35→4ºC/∞
8	100	1.0	3.0	1.0	0	None	10	BDX	98ºC/5min→[(96ºC/10sec)(50ºC/5sec)(60ºC/4min)] x25 →4ºC/∞
9	200	1.0	0.75	0.25	1.5	None	10	CleanSeq	95ºC/1min→[(95ºC/10sec)(50ºC/5sec)(60ºC/2min)] x35 →4ºC/∞
10	100	1.0	2.0	0	0	None	20	Edge DTR v3	100ºC/2min→[(96ºC/30sec)(50ºC/15sec)(60ºC/4min)] x26 →4ºC/∞

* If you have dGTPv1.0 instead of dGTPv3.0, use it and indicate so in your notes.

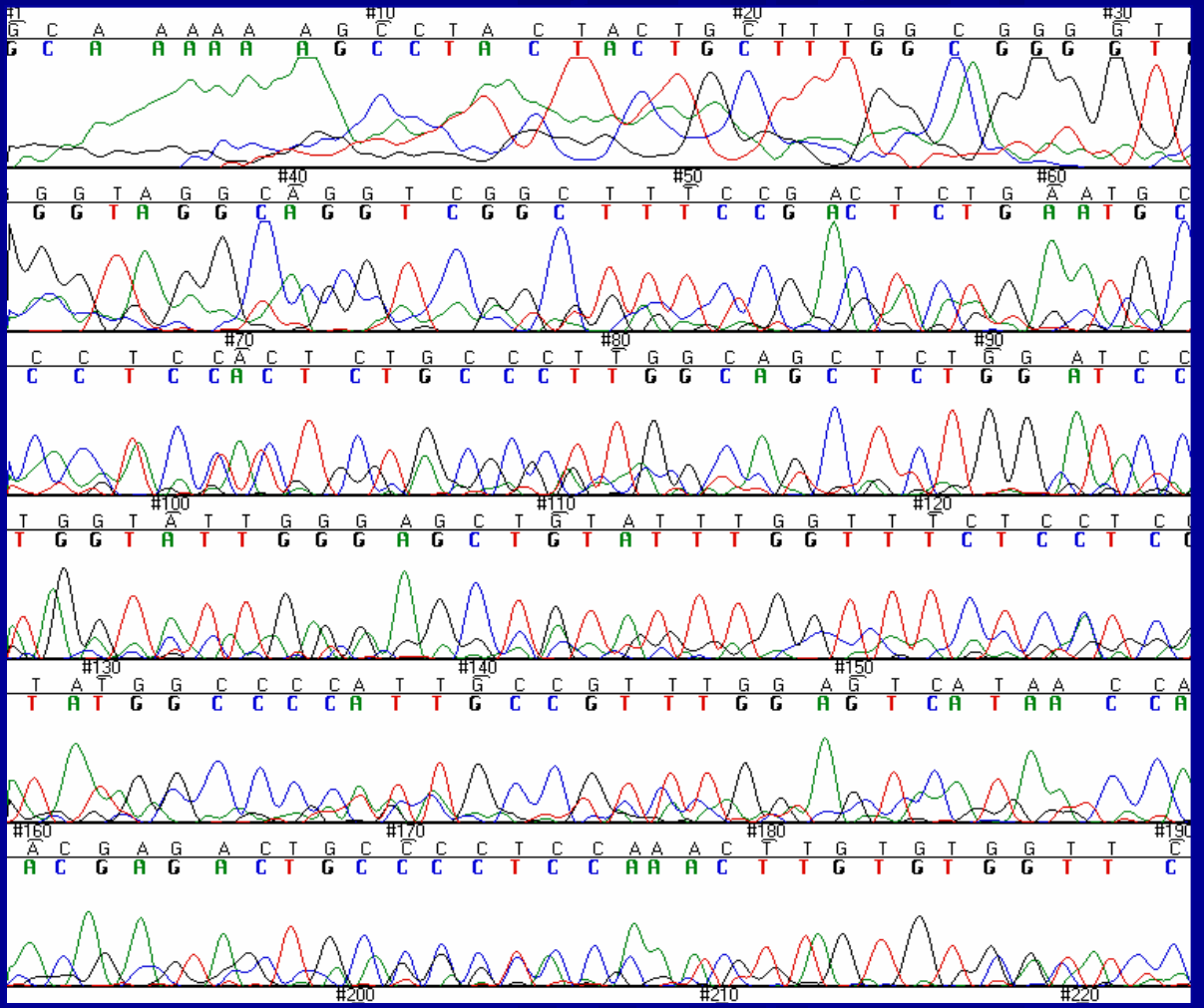
^ This is what the best protocol used originally. If you can please use it, otherwise use your own cleanup protocol and indicate so in your notes.

This was prepared in-house using G-50 Sephadex and Millipore MAHVN4550 plate.

Q >20 Not Always a Good Predictor of Sequence Quality, at Least for Difficult Templates

DNA3: 24 base hairpin with T_m > 95°C

1	GCAAAAAAGC	CTACTACTGC	TTTGGCGGGG	TGGGTAGGCA
41	GGTCGGCTTT	CCGACTCTGA	ATGCCCTCCA	CTCTGCCCTT
81	GGCAGCTCTG	GATCCTGGTA	TTGGGAGCTG	TATTTGGTTT
121	CTCCTCGTAT	GGCCCCATTG	CCGTTTGGAG	TCATAACCAA
161	CGAGACTGCC	CCTCCAAACT	TGTGTGGTTC	TACCACGCCA
201	GTGTAGCAGT	CTAGCCAAAT	GGGAAGTGCA	ACATATAGAG
241	TCTCTCCCTC	ACCCAAACCT	GACTCATGCA	A



(DNA3-P3R)
Q>20 indicated as >230

Phase II: Assignment of Protocols to Specific DNA Templates

DNA #	Primer	Protocol 1	Protocol 2
DNA 1	F	1	2
	R	1	3
DNA 2	F	6	8
	R	6	1
DNA 3	F	1	4
	R	5	10
DNA 4	F	6	1
	R	1	10
DNA 5	F	1	7
	R	1	7
DNA 6	F	9	6
	R	6	1
DNA 7	F	1	3
	R	1	3
DNA 8	F	1	7
pGem3zf	F	9	10

Phase II: Possible Reasons for Uneven Data

- **Lack of specialized reagents**
  - dGTPv3.0
  - Betaine
  - Others?
- **Various cleanup procedures**
- **Various PCR machines**
- **Experience factor**

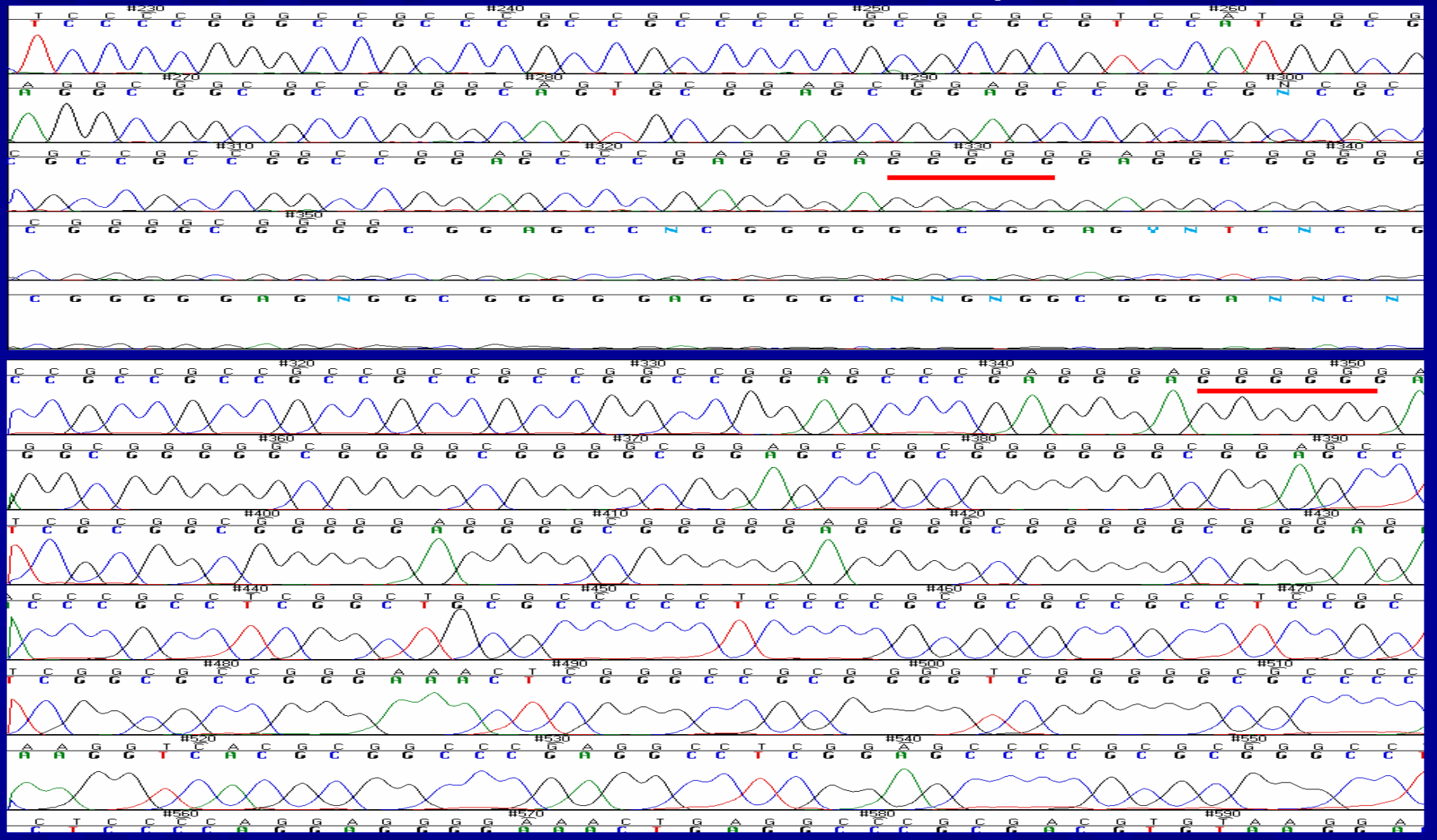
(sequencing instruments are not a factor, as all contributors used ABI platforms)

DNA1(F): Comparison of Results from Average & Best Sequencing Protocols

94%GC over 200 bases/101 base non-repeat G/C

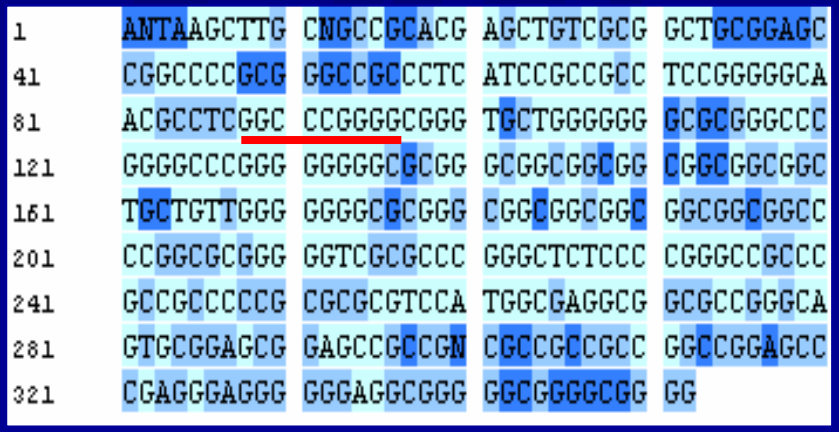
Avg

Best



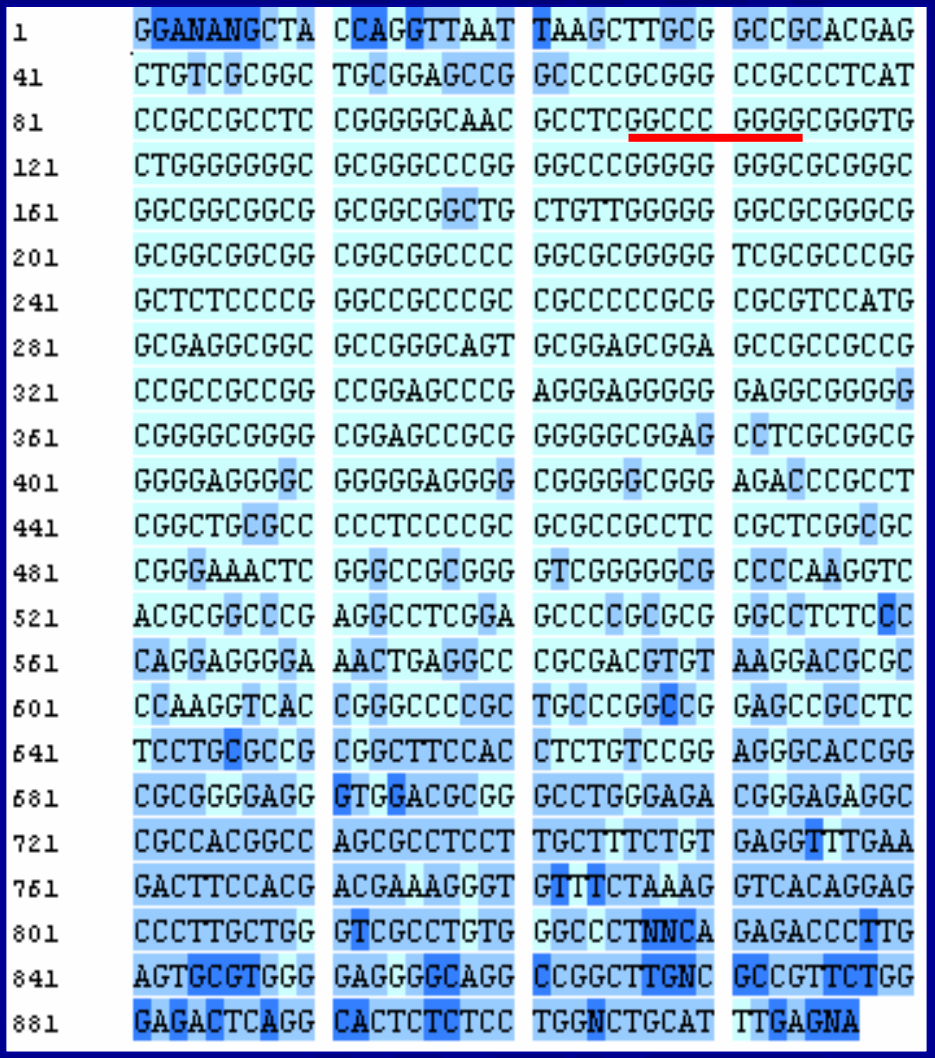
DNA1(F): Comparison of an Average and the Best Sequencing Protocols

Average



94%GC over 200 bases / 101 base non-repeat G/C

Best



DNA1: Examine Repeats Module Highlights Many Different Types of Difficult Regions

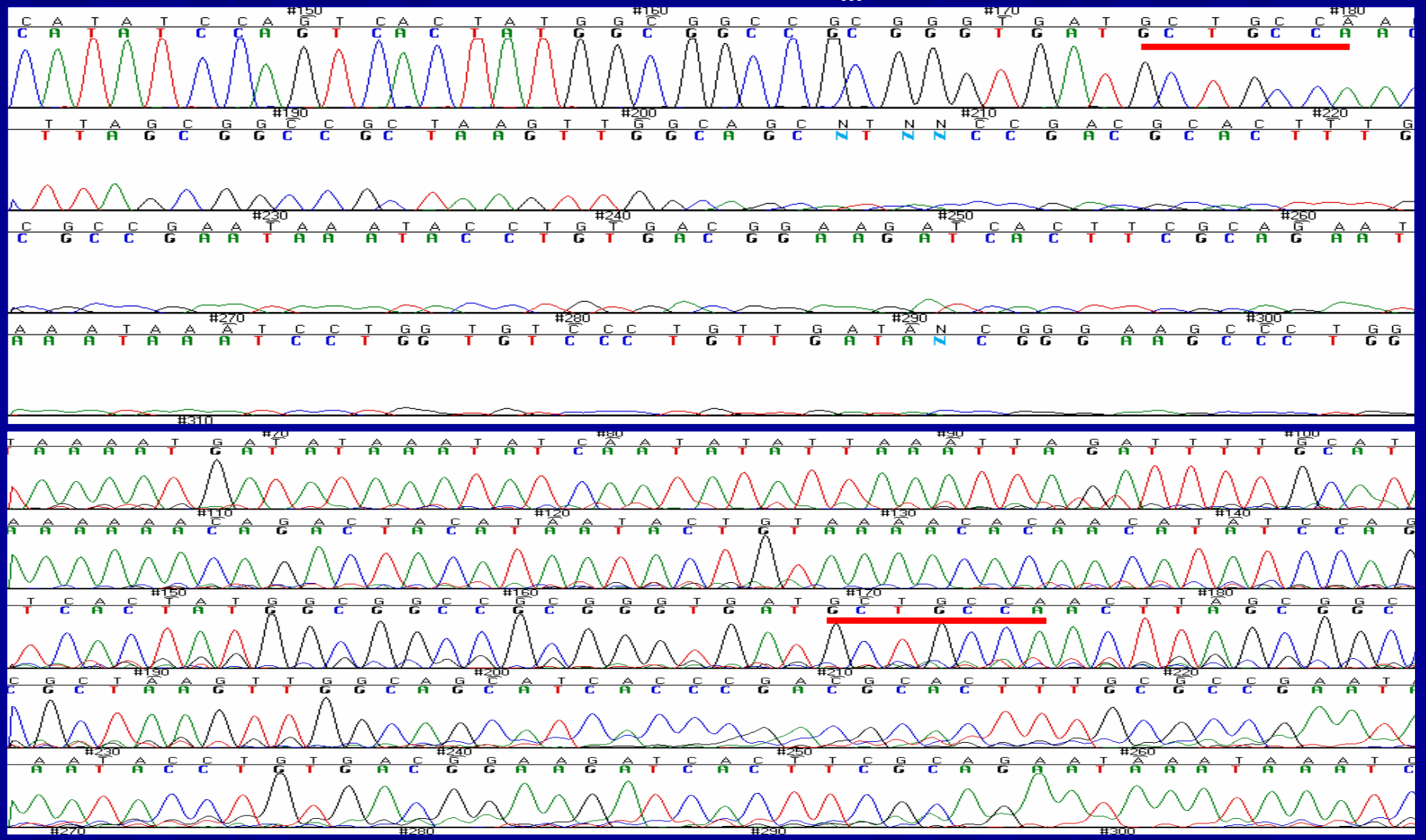
	Type	Length	Count	Positions	Sequence
2	Direct	66	2	171,174	GCAGCAGCAG CAGCAGCAGC AGCAGCAGCA GCAGCAGCAG CAGCAGCAGC AGCAGCAGCA GCAGCA
3	Direct	49	2	1369, 1372	GGCGGCGGCG GCGGCGGCGG CGGCGGCGGC GGCGGCGGCG GCGGCGGCG
4	Direct	19	18	##########	GCAGCAGCAG CAGCAGCAG
5	Direct	15	2	27,612,764	CACCACCACC ACCAC
6	Direct	13	2	253,612	AGGCAGCAGC AGC
7	Direct	12	2	19,402,757	CCAGCACCAC CA
8	Direct	11	2	3,471,314	CTTCACAGCC G
17	Invert	21	2	161,162	TGCTGCTGCT GCAGCAGCAG C
18	Invert	13	3	134,927,632,766	GTGGTGGTGG TGG
19	Invert	13	2	160,228	TTGCTGCTGC TGC
20	Invert	12	4	1,349,276,127,642,760	GTGGTGGTGG TG
21	Invert	12	21	##########	GCAGCAGCAG CA
22	Invert	11	2	24,072,408	CAGGAATTCC T
23	Invert	11	2	16,981,969	AGCTTCTGGG T
24	Invert	11	25	##########	GCTGCTGCTG C
25	Invert	11	2	7,452,750	CTGGGTGTGG A
32	Palind	22	2	161,182	TGCTGCTGCT GCAGCAGCAG CA
33	Palind	20	2	162,181	GCTGCTGCTG CAGCAGCAGC
34	Palind	18	2	163,180	CTGCTGCTGC AGCAGCAG
35	Palind	16	2	164,179	TGCTGCTGCA GCAGCA
36	Palind	14	2	165,178	GCTGCTGCAG CAGC
37	Palind	12	2	24,072,418	CAGGAATTCC TG
38	Palind	12	2	166,177	CTGCTGCAGC AG
48	Trinucleotide (TN)	69	1	171	GCAGCAGCAG CAGCAGCAGC AGCAGCAGCA GCAGCAGCAG CAGCAGCAGC AGCAGCAGCA GCAGCAGCA
49	Trinucleotide (TN)	66	1	173	AGCAGCAGCA GCAGCAGCAG CAGCAGCAGC AGCAGCAGCA GCAGCAGCAG CAGCAGCAGC AGCAGC
50	Trinucleotide (TN)	66	1	172	CAGCAGCAGC AGCAGCAGCA GCAGCAGCAG CAGCAGCAGC AGCAGCAGCA GCAGCAGCAG CAGCAG
51	Trinucleotide (TN)	51	1	1369	GGCGGCGGCG GCGGCGGCGG CGGCGGCGGC GGCGGCGGCG GCGGCGGCGG C
52	Trinucleotide (TN)	51	1	1370	GCGGCGGCGG CGGCGGCGGC GGCGGCGGCG GCGGCGGCGG CGGCGGCGGC G
53	Trinucleotide (TN)	48	1	1371	CGGCGGCGGC GGCGGCGGCG GCGGCGGCGG CGGCGGCGGC GGCGGCGG
54	Non-repeat DN	52	1	1369	GGCGGCGGCG GCGGCGGCGG CGGCGGCGGC GGCGGCGGCG GCGGCGGCGG CG

DNA3(F): Comparison of Results from Average & Best Sequencing Protocols

24 base hairpin with T_m >95°C

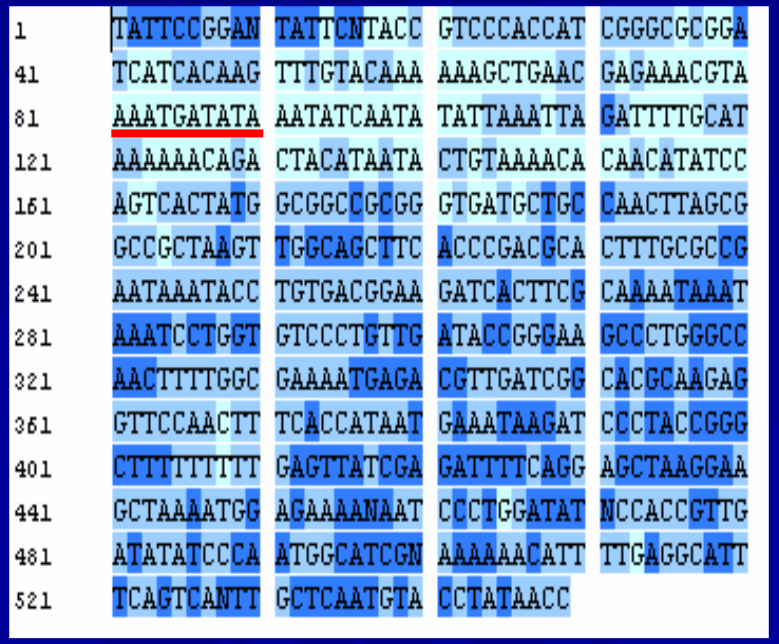
Avg

Best



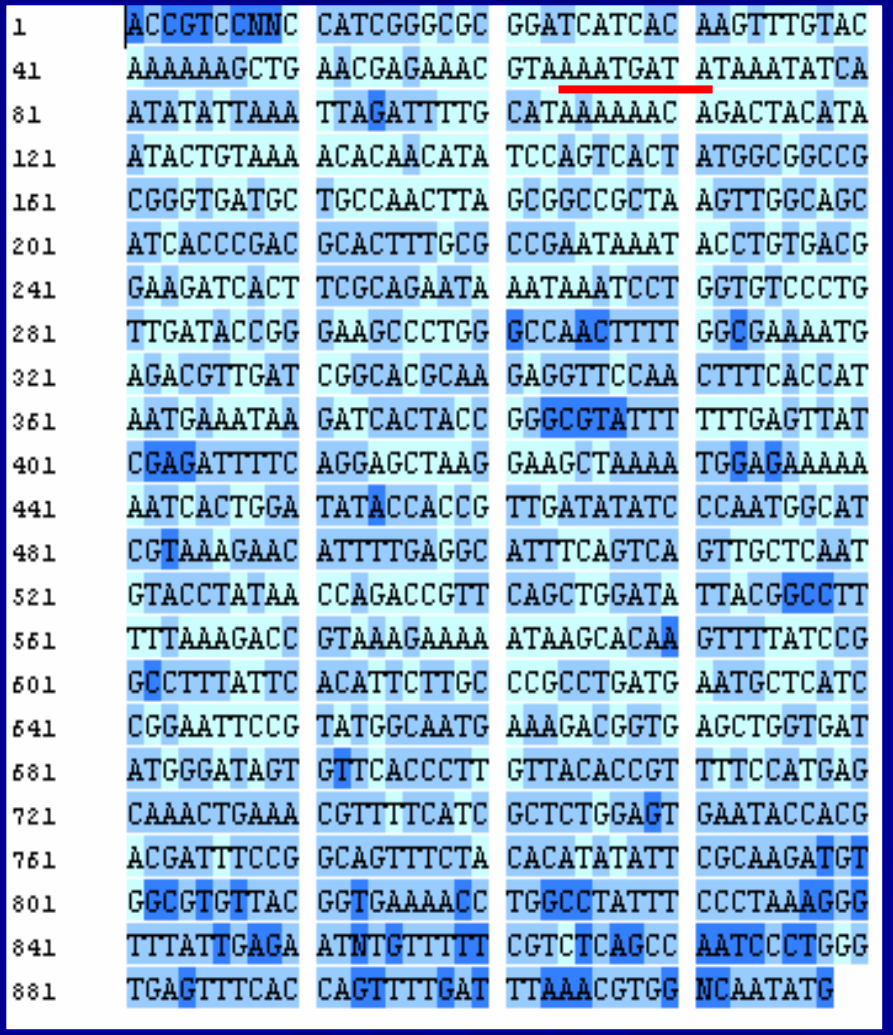
DNA3(F): Comparison of an Average and the Best Sequencing Protocols

Average



24 base hairpin with T_m >95°C

Best

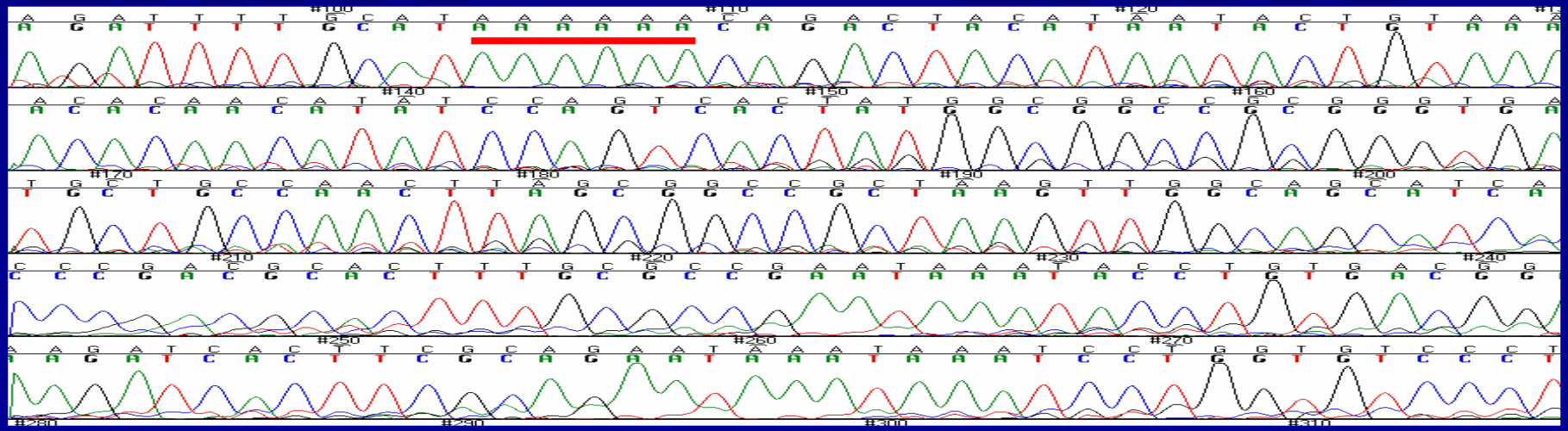


DNA3: Large Inverted Repeat (Hairpin)

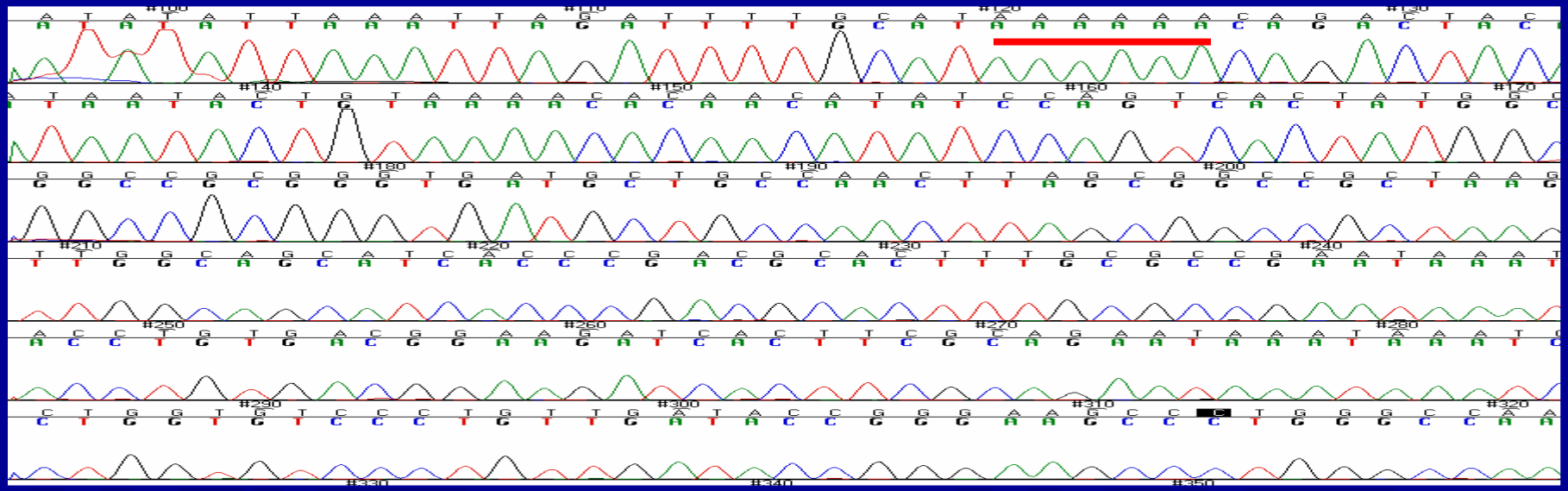
1	Type	Length	Count	Positions	Sequence pDEST 8-Examine Repeats
2					
3	**Invert**	**47**	**2**	**172,173**	**CGGGTGATGC TGCCAACTTA GCGGCCGCTA AGTTGGCAGC ATCACCC**
4					
5	Palind	48	2	172,219	CGGGTGATGC TGCCAACTTA GCGGCCGCTA AGTTGGCAGC ATCACCCG
6	Palind	46	2	173,218	GGGTGATGCT GCCAACTTAG CGGCCGCTAA GTTGGCAGCA TCACCC
7	Palind	44	2	174,217	GGTGATGCTG CCAACTTAGC GGCCGCTAAG TTGGCAGCAT CACC
8	Palind	42	2	175,216	GTGATGCTGC CAACTTAGCG GCCGCTAAGT TGGCAGCATC AC
9	Palind	40	2	176,215	TGATGCTGCC AACTTAGCGG CCGCTAAGTT GGCAGCATCA
10	Palind	38	2	177,214	GATGCTGCCA ACTTAGCGGC CGCTAAGTTG GCAGCATC
11	Palind	36	2	178,213	ATGCTGCCAA CTTAGCGGCC GCTAAGTTGG CAGCAT
12	Palind	34	2	179,212	TGCTGCCAAC TTAGCGGCCG CTAAGTTGGC AGCA
13	Palind	32	2	180,211	GCTGCCAACT TAGCGGCCGC TAAGTTGGCA GC
14	Palind	30	2	181,210	CTGCCAACTT AGCGGCCGCT AAGTTGGCAG
15	Palind	28	2	182,209	TGCCAACTTA GCGGCCGCTA AGTTGGCA
16	Palind	26	2	183,208	GCCAACTTAG CGGCCGCTAA GTTGGC
17	Palind	24	2	184,207	CCAACTTAGC GGCCGCTAAG TTGG
18	Palind	22	2	185,206	CAACTTAGCG GCCGCTAAGT TG
19	Palind	20	2	186,205	AACTTAGCGG CCGCTAAGTT
20	Palind	18	2	187,204	ACTTAGCGGC CGCTAAGT
23					

DNA3(F): Comparison of Results from Best DSRG & Best Possible Sequencing Protocols

Study Best DSRG



Current Best SFK



DNA3(F): Comparison of Best (DSRG Study) and Best Possible (SFK) Sequencing Protocols

Best (DSRG Study)

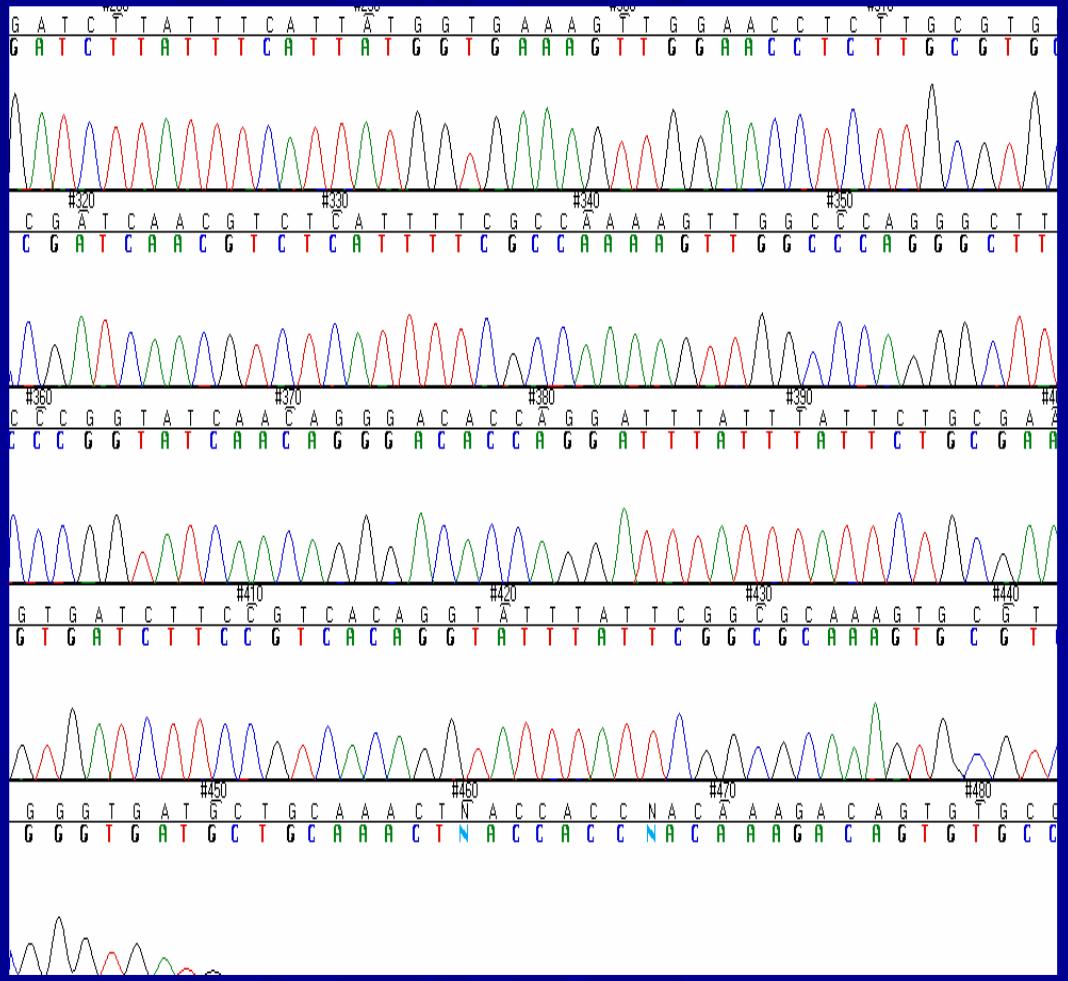
```
1    ACCGTCCNNC CATCGGGCGC GGATCATCAC AAGTTTGTAC
41   AAAAAAGCTG AACGAGAAAC GTAAAATGAT ATAAATATCA
81   ATATATTAAA TTAGATTTTG CATAAAAAAC AGACTACATA
121  ATACTGTAAA ACACAACATA TCCAGTCACT ATGGCGGCCG
161  CGGGTGATGC TGCCAACTTA GCGGCCGCTA AGTTGGCAGC
201  ATCACCCGAC GCACTTTGCG CCGAATAAAT ACCTGTGACG
241  GAAGATCACT TCGCAGAATA AATAAATCCT GGTGTCCCTG
281  TTGATACCGG GAAGCCCTGG GCCAACTTTT GGCGAAAATG
321  AGACGTTGAT CGGCACGCAA GAGGTTCCAA CTTTCACCAT
361  AATGAAATAA GATCACTACC GGGCGTATTT TTTGAGTTAT
401  CGAGATTTTC AGGAGCTAAG GAAGCTAAAA TGGAGAAAAA
441  AATCACTGGA TATACCACCG TTGATATATC CCAATGGCAT
481  CGTAAAGAAC ATTTTGAGGC ATTTCAGTCA GTTGCTCAAT
521  GTACCTATAA CCAGACCGTT CAGCTGGATA TTACGGCCTT
561  TTTAAAGACC GTAAAGAAAA ATAAGCACAA GTTTTATCCG
601  GCCTTTATTC ACATTCTTGC CCGCCTGATG AATGCTCATC
641  CGGAATTCCG TATGGCAATG AAAGACGGTG AGCTGGTGAT
681  ATGGGATAGT GTTCACCCTT GTTACACCGT TTTCCATGAG
721  CAAACTGAAA CGTTTTCATC GCTCTGGAGT GAATACCACG
761  ACGATTTCCG GCAGTTTCTA CACATATATT CGCAAGATGT
801  GGCGTGTTAC GGTGAAAACC TGGCCTATTT CCCTAAAGGG
841  TTTATTGAGA ATNTGTTTTT CGTCTCAGCC AATCCCTGGG
881  TGAGTTTCAC CAGTTTTGAT TTAAACGTGG NCAATATG
```

Best Possible (SFK)

```
1    TTCCGGNTTA TTCNTACCGT CCCACCATCG GGCGCGGATC
41   ATCACAAGTT TGNNCNNNNN AAGCTGAACG AGAAACGTAA
81   AATGATATAA ATATCAATAT ATTAAATTAG ATTTTGCATA
121  AAAAACAGAC TACATAATAC TGTAAAACAC AACATATCCA
161  GTCACTATGG CGGCCGCGGG TGATGCTGCC AACTTAGCGG
201  CCGCTAAGTT GGCAGCATCA CCCGACGCAC TTTGCGCCGA
241  ATAAATACCT GTGACGGAAG ATCACTTCGC AGAATAAATA
281  AATCCTGGTG TCCCTGTTGA TACCGGGAAG CCCTGGGCCA
321  ACTTTTGGCG AAAATGAGAC GTTGATCGGC ACGCAAGAGG
361  TTCCAACTTT CACCATAATG AAATAAGATC ACTACCGGGC
401  GTATTTTTTG AGTTATCGAG ATTTTCAGGA GCTAAGGAAG
441  CTAAAATGGA GAAAAAAATC ACTGGATATA CCACCGTTGA
481  TATATCCCAA TGGCATCGTA AAGAACATTT TGAGGCATTT
521  CAGTCAGTTG CTCAATGTAC CTATAACCAG ACCGTTCAGC
561  TGGATATTAC GGCCTTTTTA AAGACCGTAA AGAAAAATAA
601  GCACAAGTTT TATCCGGCCT TTATTCACAT TCTTGCCCGC
641  CTGATGAATG CTCATCCGGA ATTCCGTATG GCAATGAAAG
681  ACGGTGAGCT GGTGATATGG GATAGTGTTC ACCCTTGTTA
721  CACCGTTTTC CATGAGCAAA CTGAAACGTT TTCATCGCTC
761  TGGAGTGAAT ACCACGACGA TTTCCGGCAG TTTCTACACA
801  TATATTCGCA AGATGTGGCG TGTTACGGTG AAAACCTGGC
841  CTATTTCCCT AAAGGGTTTA TTGAGAATAT GTTTTTCGNC
881  TCAGCCAATC CCTGGGTGAG TTTCNNNAGT TT
```

DNA3(R): No Acceptable Data Obtained During Study, Although It Is Possible…

24 base hairpin with Tm >95°C

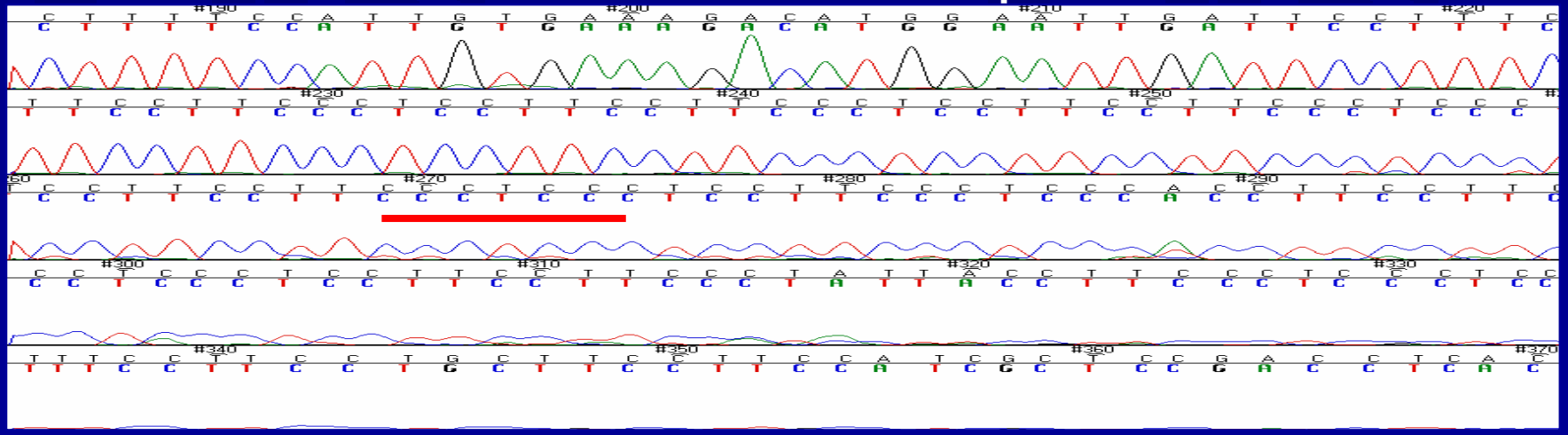


```
1    CGGANTCCGG ANGAGCATTC ATCAGGCGGG CAAGAATGTG
41   AATAAAGGCC GGATAAAACT TGTGCTTATT TTTCTTTACG
81   GTCTTTAAAA AGGCCGTAAT ATCCAGCTGA ACGGTCTGGT
121  TATAGGTACA TTGAGCAACT GACTGAAATG CCTCAAAATG
161  TTCTTTACGA TGCCATTGGG ATATATCAAC GGTGGTATAT
201  CCAGTGATTT TTTTCTCCAT TTTAGCTTCC TTAGCTCCTG
241  AAAATCTCGA TAACTCAAAA AATACGCCCG GTAGTGATCT
281  TATTTCATTA TGGTGAAAGT TGGAACCTCT TGCGTGCCGA
321  TCAACGTCTC ATTTTCGCCA AAAGTTGGCC CAGGGCTTCC
361  CGGTATCAAC AGGGACACCA GGATTTATTT ATTCTGCGAA
401  GTGATCTTCC GTCACAGGTA TTTATTCGGC GCAAAGTGCG
441  TCGGGTGATG CTGCAAACTN ACCACCNACA AAGACAGTGT
481  GCCGGTCTCC GNNATCGGGG AAGAAGTGGC TGATCGCNGC
521  CNCCGCGAAA ATGACATCNA AAACGCCANN AACCTGATGT
561  TCTGGGGAAT ATAAATGTCA GGGTCCCTTA TACAC
```

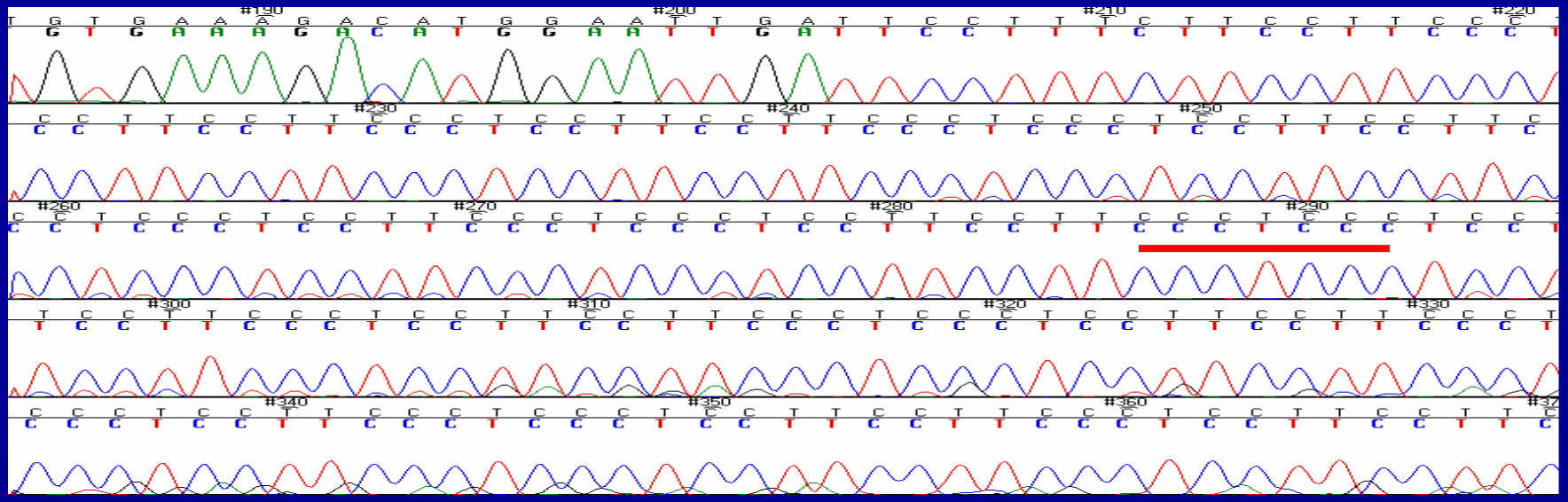
DNA5(F): Comparison of Results from Average & Best Sequencing Protocols

456 base di-nucleotide non-repeat T/C

Avg

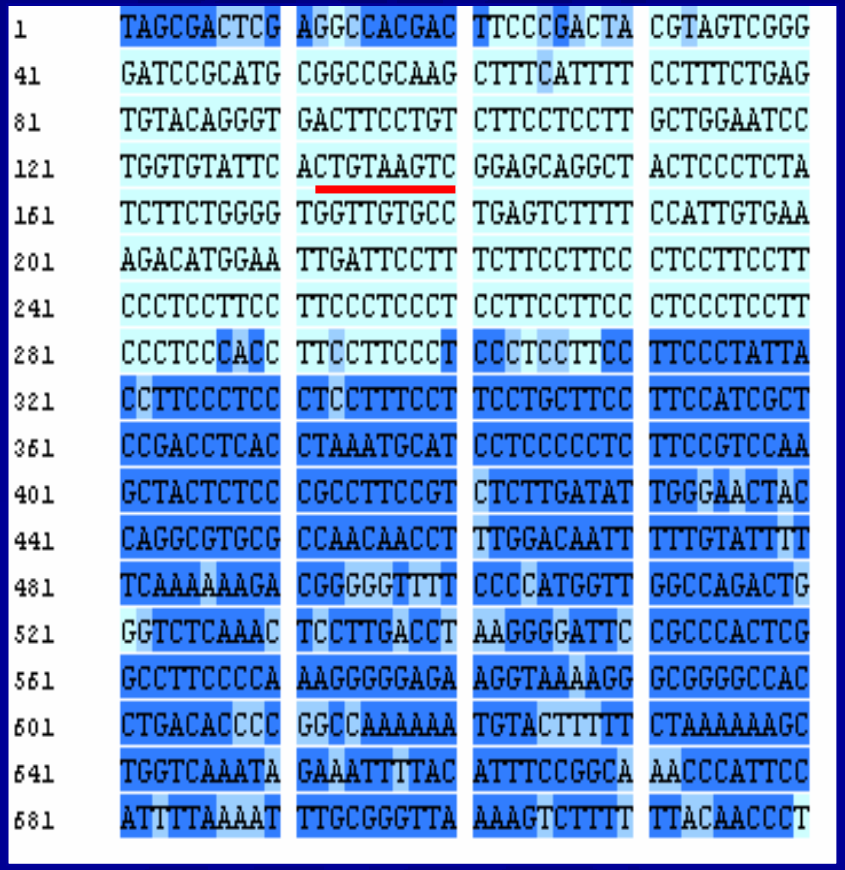


Best

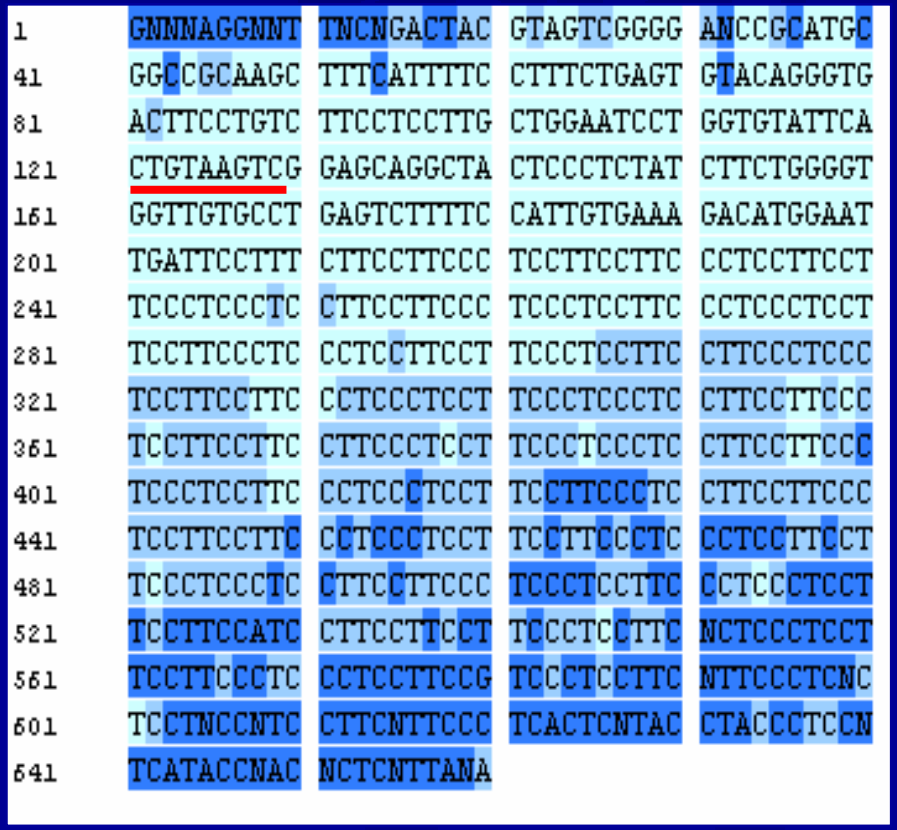


DNA5(F): Comparison of an Average and the Best Sequencing Protocols

Average



Best



456 base di-nucleotide non-repeat T/C

DNA5(F): Examine Repeats Module Highlights Many Different Kinds of Difficult Regions

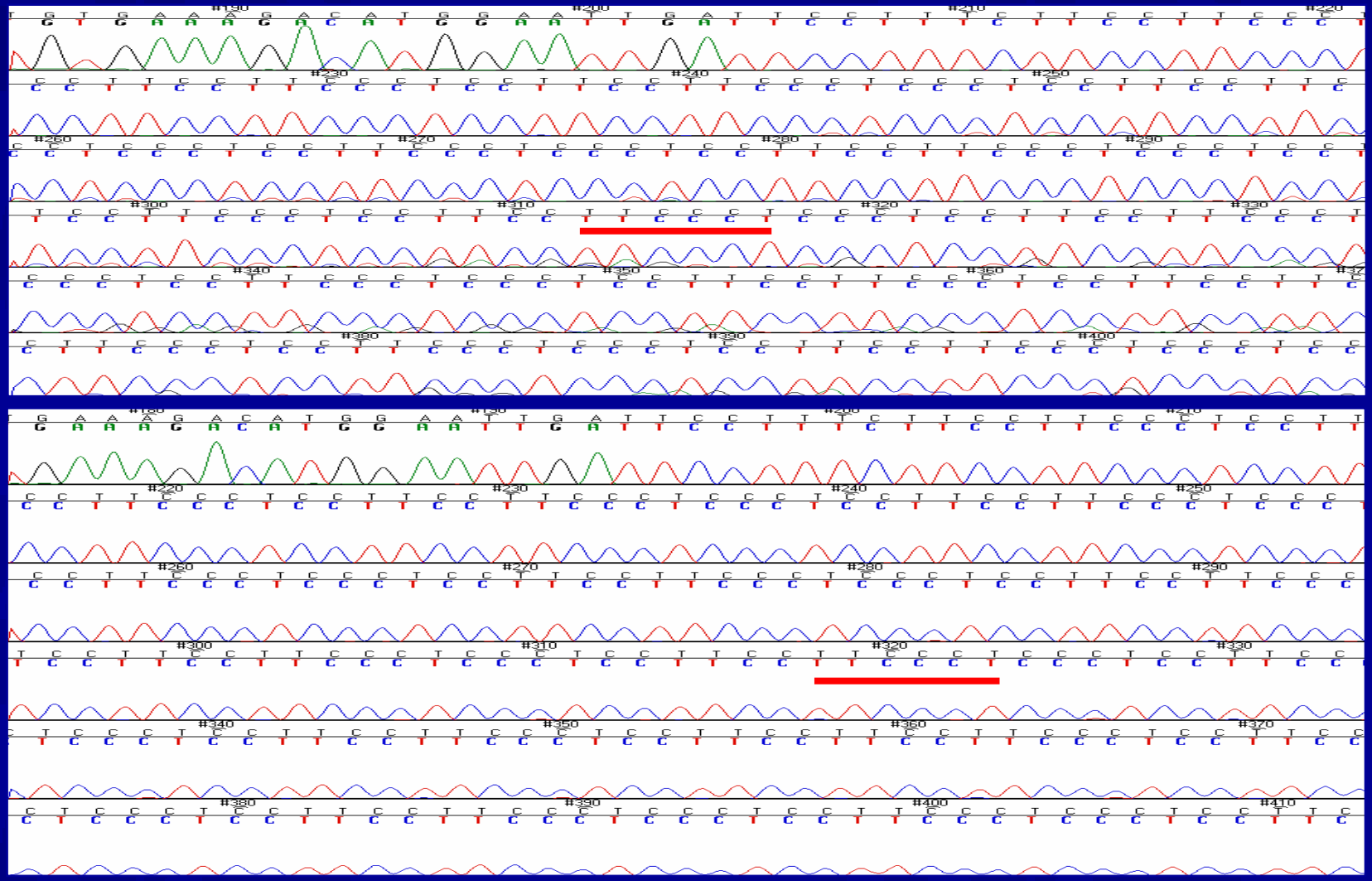
	Type	Length	#	Positions	Sequence
2	Direct	131	2	296,464	TCCCTCCTTC CTTCCCTCCC TCCTTCCTTC CCTCCCTCCT TCCCTCCCTC CTTCCTTCCC TCCTTCCTTC CTTCCCTCCT TCCCTCCCTC CTTCCTTC
3	Direct	75	2	212,284	TCCCTCCTTC CTTCCCTCCT TCCTTCCCTC CCTCCTTCCT TCCCTCCCTC CTTCCCTCCC TCCTTCCTTC CCTCC
4	Direct	63	2	232,540	TCCTTCCCTC CCTCCTTCCT TCCCTCCCTC CTTCCCTCCC TCCTTCCTTC CCTCCCTCCT TCC
5	Direct	63	3	304,372,472	TCCTTCCCTC CCTCCTTCCT TCCCTCCCTC CTTCCCTCCC TCCTTCCTTC CCTCCTTCCT TCC
6	Direct	61	2	206,414	CTTCCTTCCC TCCTTCCTTC CCTCCTTCCT TCCCTCCCTC CTTCCTTCCC TCCCTCCTTC C
7	Direct	59	2	432,448	TCCCTCCTTC CTTCCCTCCC TCCTTCCTTC CCTCCCTCCT TCCTTCCCTC CCTCCTTCC
8	Direct	55	2	268,600	TCCCTCCTTC CTTCCCTCCC TCCTTCCTTC CCTCCTTCCT TCCCTCCCTC CTTCC
9	Direct	43	3	260,440,456	TCCTTCCCTC CCTCCTTCCT TCCCTCCCTC CTTCCTTCCC TCC
10	Direct	43	3	368,536,564	TCCCTCCTTC CCTCCCTCCT TCCTTCCCTC CCTCCTTCCC TCC
11	Direct	39	4	220,292,428,596	TCCTTCCCTC CTTCCTTCCC TCCCTCCTTC CTTCCCTCC
12	Direct	39	3	276,400,608	TCCTTCCCTC CCTCCTTCCT TCCCTCCTTC CTTCCCTCC
14	Invert	17	2	7,8	CCCGACTACG TAGTCGG
17	Palind	18	2	7,24	CCCGACTACG TAGTCGGG
18	Palind	16	2	8,23	CCGACTACGT AGTCGG
19	Palind	14	2	9,22	CGACTACGTA GTCG
20	Palind	12	2	10,21	GACTACGTAG TC
21	Palind	10	2	28,37	CCGCATGCGG
22	Palind	10	2	33,42	TGCGGCCGCA
23	Palind	10	2	11,20	ACTACGTAGT
27	Non-repeat D	456	1	199	TTCCTTTCTT CCTTCCCTCC TTCCTTCCCT CCTTCCTTCC CTCCCTCCTT CCTTCCCTCC CTCCTTCCCT CCCTCCTTCC TTCCCTCCCT CCTTCCTT
28					

DNA5(F): Comparison of Results from Best DSRG & Best Possible Sequencing Protocols

456 base di-nucleotide non-repeat T/C

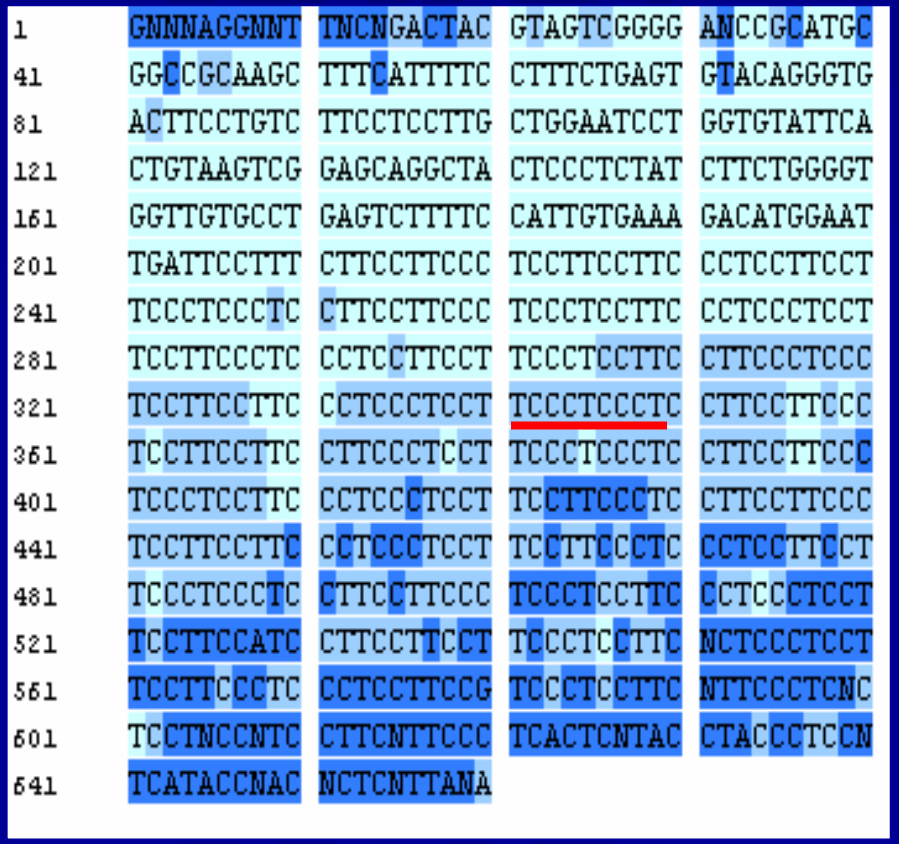
Study Best DSRG

Current Best SFK



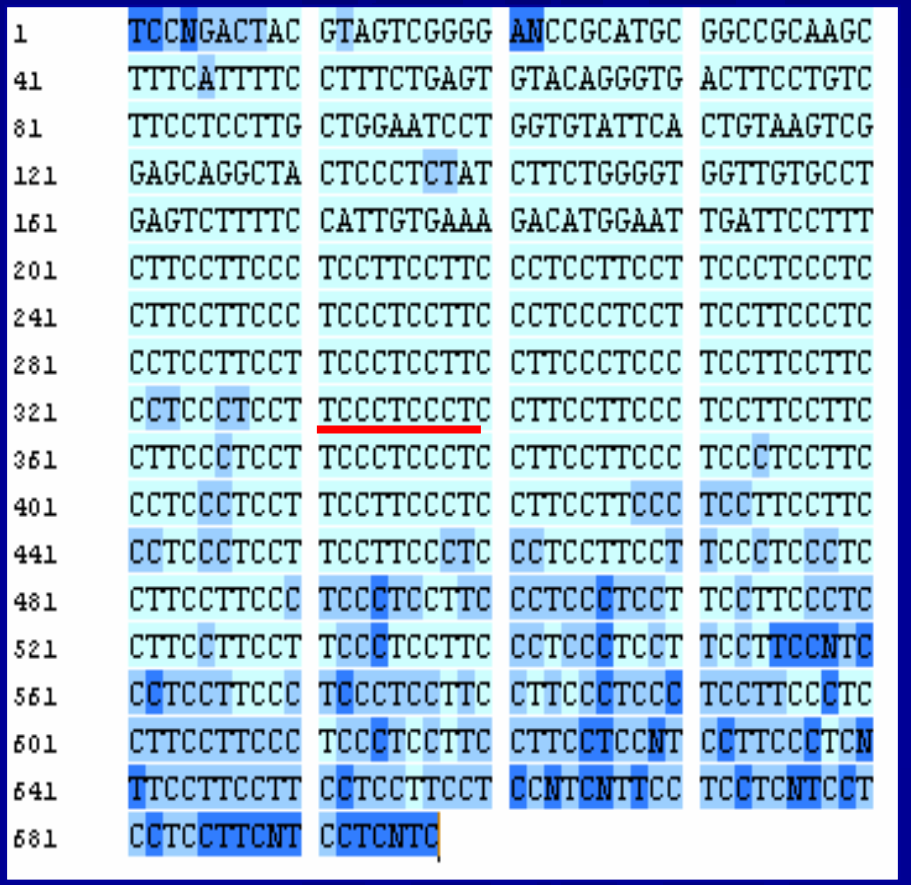
DNA5(F): Comparison of Best (DSRG Study) and Best Possible (SFK) Sequencing Protocols

Best (DSRG Study)



```
1    GNNNAGGNNT TNCNGACTAC GTAGTCGGGG ANCCGCATGC
41   GGCCGCAAGC TTTCATTTTC CTTTCTGAGT GTACAGGGTG
81   ACTTCCTGTC TTCCTCCTTG CTGGAATCCT GGTGTATTCA
121  CTGTAAGTCG GAGCAGGCTA CTCCCTCTAT CTTCTGGGGT
161  GGTTGTGCCT GAGTCTTTTC CATTGTGAAA GACATGGAAT
201  TGATTCCTTT CTTCCTTCCC TCCTTCCTTC CCTCCTTCCT
241  TCCCTCCCTC CTTCCTTCCC TCCCTCCTTC CCTCCCTCCT
281  TCCTTCCCTC CCTCCTTCCT TCCCTCCTTC CTTCCCTCCC
321  TCCTTCCTTC CCTCCCTCCT TCCCTCCCTC CTTCCTTCCC
361  TCCTTCCTTC CTTCCCTCCT TCCCTCCCTC CTTCCTTCCC
401  TCCCTCCTTC CCTCCCTCCT TCCTTCCCTC CTTCCTTCCC
441  TCCTTCCTTC CCTCCCTCCT TCCTTCCCTC CCTCCTTCCT
481  TCCCTCCCTC CTTCCTTCCC TCCCTCCTTC CCTCCCTCCT
521  TCCTTCCATC CTTCCTTCCT TCCCTCCTTC NCTCCCTCCT
561  TCCTTCCCTC CCTCCTTCCG TCCCTCCTTC NTTCCCTCNC
601  TCCTNCCNTC CTTCNTTCCC TCACTCNTAC CTACCCTCCN
641  TCATACCNAC NCTCNTTANA
```

Best Possible (SFK)

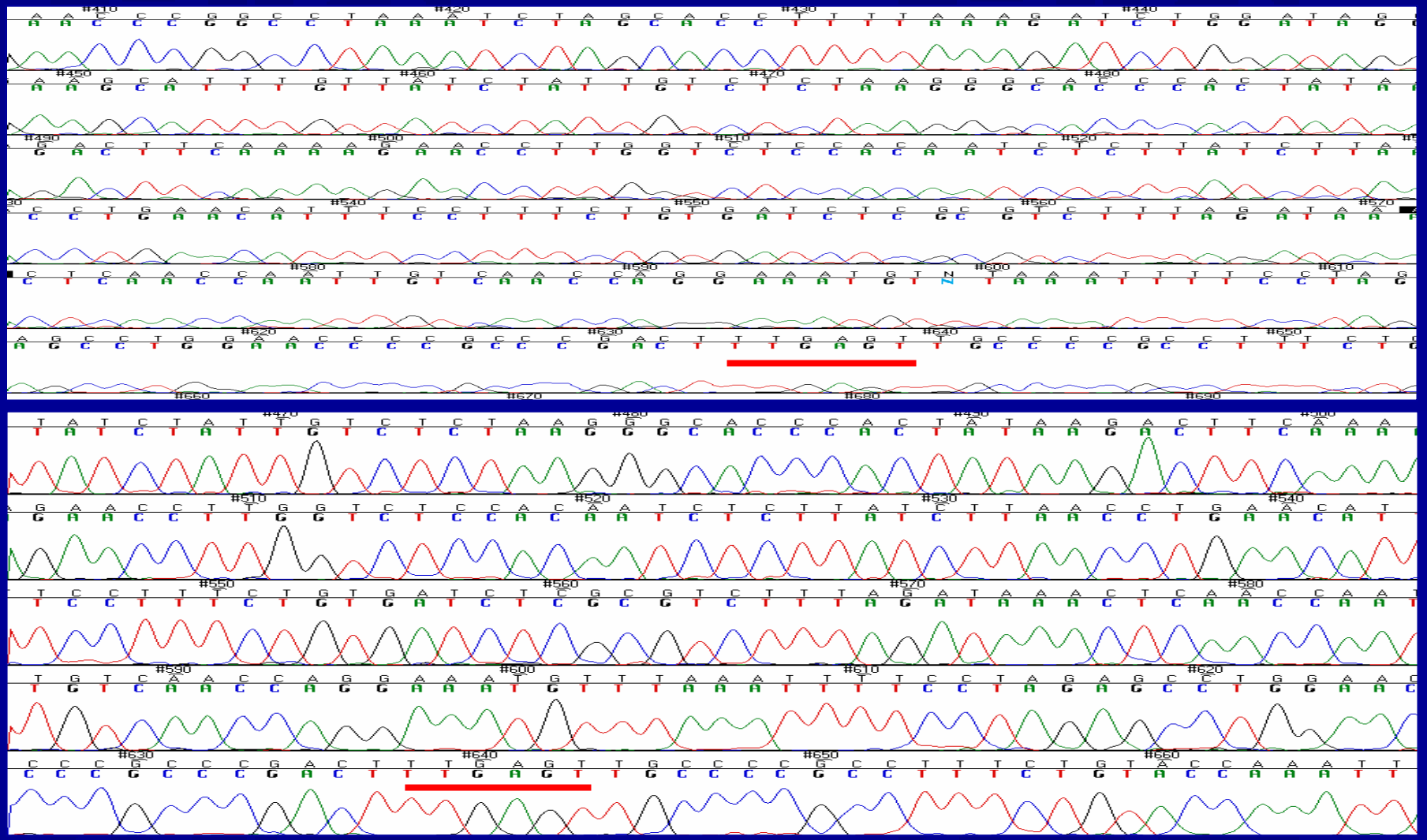


```
1    TCCNGACTAC GTAGTCGGGG ANCCGCATGC GGCCGCAAGC
41   TTTCATTTTC CTTTCTGAGT GTACAGGGTG ACTTCCTGTC
81   TTCCTCCTTG CTGGAATCCT GGTGTATTCA CTGTAAGTCG
121  GAGCAGGCTA CTCCCTCTAT CTTCTGGGGT GGTTGTGCCT
161  GAGTCTTTTC CATTGTGAAA GACATGGAAT TGATTCCTTT
201  CTTCCTTCCC TCCTTCCTTC CCTCCTTCCT TCCCTCCCTC
241  CTTCCTTCCC TCCCTCCTTC CCTCCCTCCT TCCTTCCCTC
281  CCTCCTTCCT TCCCTCCTTC CTTCCCTCCC TCCTTCCTTC
321  CCTCCCTCCT TCCCTCCCTC CTTCCTTCCC TCCTTCCTTC
361  CTTCCCTCCT TCCCTCCCTC CTTCCTTCCC TCCCTCCTTC
401  CCTCCCTCCT TCCTTCCCTC CTTCCTTCCC TCCTTCCTTC
441  CCTCCCTCCT TCCTTCCCTC CCTCCTTCCT TCCCTCCCTC
481  CTTCCTTCCC TCCCTCCTTC CCTCCCTCCT TCCTTCCCTC
521  CTTCCTTCCT TCCCTCCTTC CCTCCCTCCT TCCTTCCNTC
561  CCTCCTTCCC TCCCTCCTTC CTTCCCTCCC TCCTTCCCTC
601  CTTCCTTCCC TCCCTCCTTC CTTCCTCCNT CCTTCCCTCN
641  TTCCTTCCTT CCTCCTTCCT CCNTCNTTCC TCCTCNTCCT
681  CCTCCTTCNT CCTCNTC
```

456 base di-nucleotide non-repeat T/C

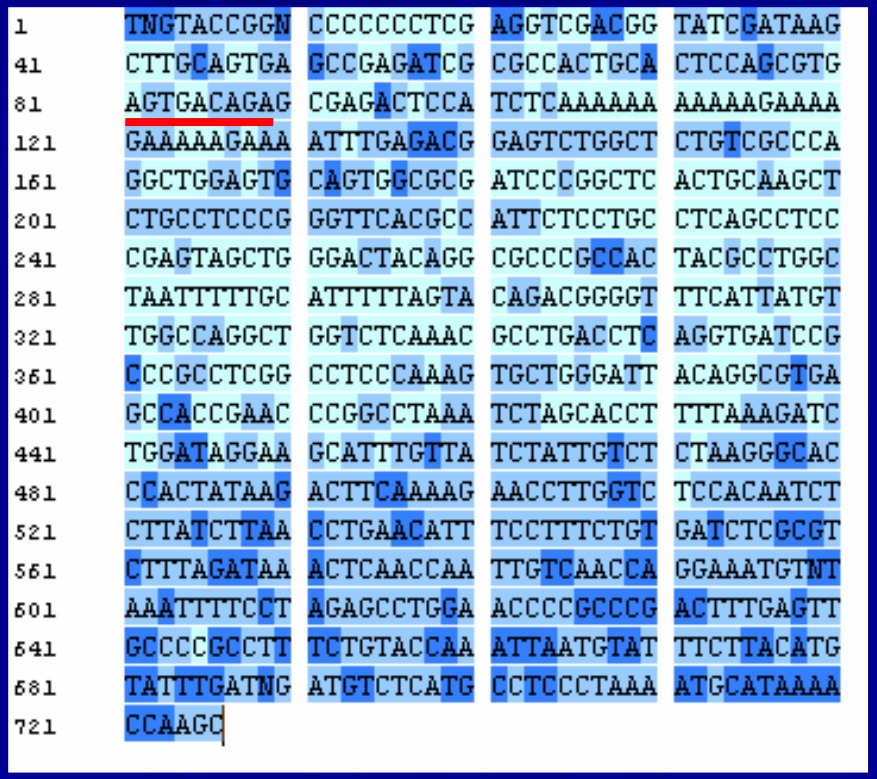
DNA8(F): Comparison of Results from Average & Best Sequencing Protocols

Alu repeat + 22 base inverted repeat/84 base loop



DNA8(Alu Repeat): Comparison of an Average and the Best Sequencing Protocols

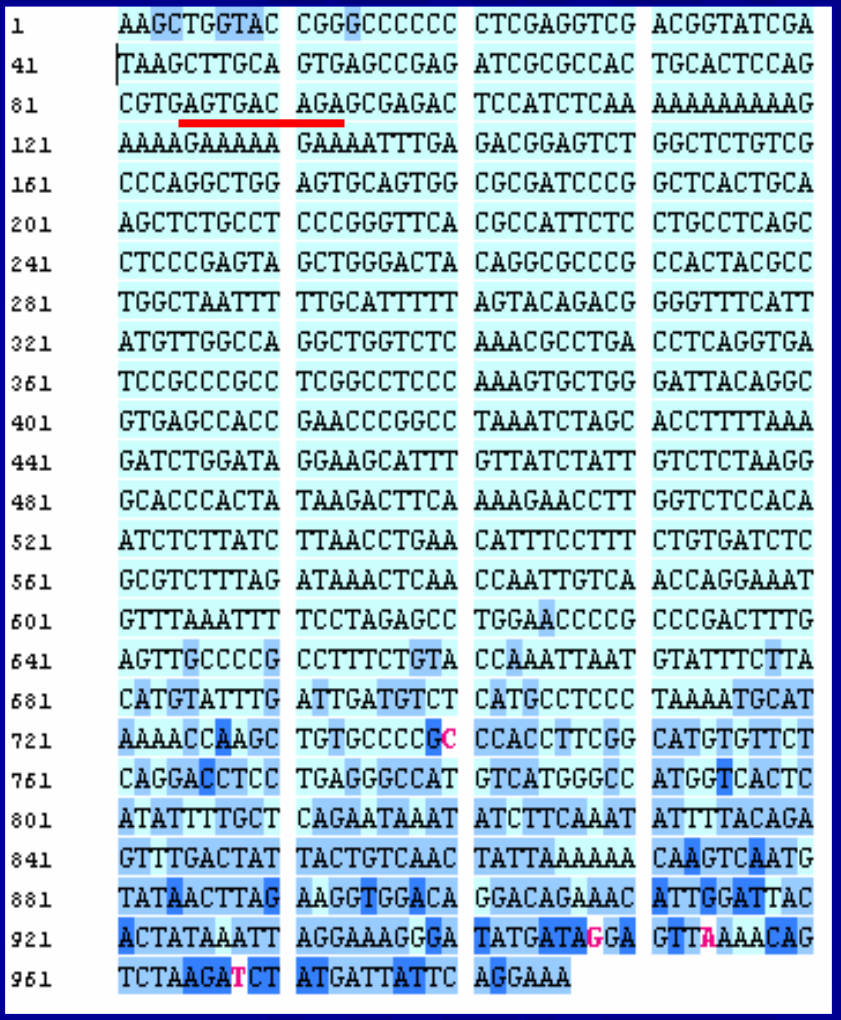
Average



```
1    TNGTACCGGN CCCCCCCTCG AGGTCGACGG TATCGATAAG
41   CTTGCAGTGA GCCGAGATCG CGCCACTGCA CTCCAGCGTG
81   AGTGACAGAG CGAGACTCCA TCTCAAAAAA AAAAAGAAAA
121  GAAAAAGAAA ATTTGAGACG GAGTCTGGCT CTGTCGCCCA
161  GGCTGGAGTG CAGTGGCGCG ATCCCGGCTC ACTGCAAGCT
201  CTGCCTCCCG GGTTCACGCC ATTCTCCTGC CTCAGCCTCC
241  CGAGTAGCTG GGACTACAGG CGCCCGCCAC TACGCCTGGC
281  TAATTTTTGC ATTTTTAGTA CAGACGGGGT TTCATTATGT
321  TGGCCAGGCT GGTCTCAAAC GCCTGACCTC AGGTGATCCG
361  CCCGCCTCGG CCTCCCAAAG TGCTGGGATT ACAGGCGTGA
401  GCCACCGAAC CCGGCCTAAA TCTAGCACCT TTTAAAGATC
441  TGGATAGGAA GCATTTGTTA TCTATTGTCT CTAAGGGCAC
481  CCACTATAAG ACTTCAAAAG AACCTTGGTC TCCACAATCT
521  CTTATCTTAA CCTGAACATT TCCTTTCTGT GATCTCGCGT
561  CTTTAGATAA ACTCAACCAA TTGTCAACCA GGAAATGTNT
601  AAATTTTCCT AGAGCCTGGA ACCCCGCCCG ACTTTGAGTT
641  GCCCCGCCTT TCTGTACCAA ATTAATGTAT TTCTTACATG
681  TATTTGATNG ATGTCTCATG CCTCCCTAAA ATGCATAAAA
721  CCAAGC
```

Alu repeat + 22 base inverted repeat / 84 base loop

Best



```
1    AAGCTGGTAC CGGGCCCCCC CTCGAGGTCG ACGGTATCGA
41   TAAGCTTGCA GTGAGCCGAG ATCGCGCCAC TGCACTCCAG
81   CGTGAGTGAC AGAGCGAGAC TCCATCTCAA AAAAAAAAAG
121  AAAAGAAAAA GAAAATTTGA GACGGAGTCT GGCTCTGTCG
161  CCCAGGCTGG AGTGCAGTGG CGCGATCCCG GCTCACTGCA
201  AGCTCTGCCT CCCGGGTTCA CGCCATTCTC CTGCCTCAGC
241  CTCCCGAGTA GCTGGGACTA CAGGCGCCCG CCACTACGCC
281  TGGCTAATTT TTGCATTTTT AGTACAGACG GGGTTTCATT
321  ATGTTGGCCA GGCTGGTCTC AAACGCCTGA CCTCAGGTGA
361  TCCGCCCGCC TCGGCCTCCC AAAGTGCTGG GATTACAGGC
401  GTGAGCCACC GAACCCGGCC TAAATCTAGC ACCTTTTAAA
441  GATCTGGATA GGAAGCATTT GTTATCTATT GTCTCTAAGG
481  GCACCCACTA TAAGACTTCA AAAGAACCTT GGTCTCCACA
521  ATCTCTTATC TTAACCTGAA CATTTCCTTT CTGTGATCTC
561  GCGTCTTTAG ATAAACTCAA CCAATTGTCA ACCAGGAAAT
601  GTTTAAATTT TCCTAGAGCC TGGAACCCCG CCCGACTTTG
641  AGTTGCCCCG CCTTTCTGTA CCAAATTAAT GTATTTCTTA
681  CATGTATTTG ATTGATGTCT CATGCCTCCC TAAAATGCAT
721  AAAACCAAGC TGTGCCCCGC CCACCTTCGG CATGTGTTCT
761  CAGGACCTCC TGAGGGCCAT GTCATGGGCC ATGGTCACTC
801  ATATTTTGCT CAGAATAAAT ATCTTCAAAT ATTTTACAGA
841  GTTTGACTAT TACTGTCAAC TATTAAAAAA CAAGTCAATG
881  TATAACTTAG AAGGTGGACA GGACAGAAAC ATTGGATTAC
921  ACTATAAATT AGGAAAGGGA TATGATAGGA GTTAAAACAG
961  TCTAAGATCT ATGATTATTC AGGAAA
```

DNA8(Alu Repeat): Various Structural Motifs

1	Type	Length	Count	Positions	Sequence pf2777-Alu repeats
2					
3	Direct	10	2	115,126	AAAAAGAAAA
4	**Invert**	**22**	**2**	**60,166**	**GATCGCGCCA CTGCACTCCA GC**
5	Invert	16	2	43,189	AGCTTGCAGT GAGCCG
6	Invert	11	2	10,141,015	TTTAAATTTA A
7	Invert	10	2	539,594	AACATTTCCT
8	Palind	12	2	10,141,025	TTTAAATTTA AA
9	Palind	10	2	10,151,024	TTAAATTTAA
10	Homopolymer (HP)	11	1	109	AAAAAAAAAA A
11					

Summary

- **This is the most comprehensive DSRG study of sequencing of difficult templates:**
  - 8 different DNA templates, both F/R direction (# 8 only F)
    - GC-rich
    - Hairpins
    - Di-nucleotide non-repeats
    - 19 G/C Homopolymer
    - Alu-repeat/22-bp inverted repeat
  - > 50 various protocols in Phase I
  - 10 most common protocols selected for Phase II
- **30 sets of DNAs/primers distributed**
  - 20 data sets received for Phase I
  - 08 data sets received for Phase II
- **Each type of difficult template requires a separate chemistry/treatment**
  - In fact, the direction matters
- **We have collected a set of sequencing protocols that have the best chance to sequence through many kinds of difficult DNA templates**

List of Participating Institutions (Data Submitted for Phase I and Phase II)

- Beckman Research Institute at the City of Hope - DNA Sequencing Core Lab
- Centre de recherche du CHUL/CHUQ - Plateforme de séquençage et de génotypage des génomes
- Cornell University - DNA Sequencing and Genotyping Lab
- Dartmouth Medical School - Pharm/Tox, Remsen
- Duke University - DNA Analysis Facility
- Eastern Regional Research Center - IBR-Genetic Analysis
- Edge BioSystems, Inc.
- Massachusetts Institute of Technology - MIT Biopolymers Laboratory
- Ohio State University - Plant-Microbe Genomics Facility
- Oklahoma State University - Recombinant DNA/Protein Core Facility
- Penn State College of Medicine - Milton S. Hershey Medical Center
- St. Jude Children's Research Hospital - High Throughput DNA Sequencing and Genotyping
- Stowers Institute for Medical Research - Molecular Biology Facility
- Trudeau Institute - Molecular Biology Core Facility
- University of British Columbia - Michael Smith Laboratories
- University of Illinois - Sequencing Core
- University of Minnesota - BMGC Sequencing and Analysis Facility
- University of Missouri - DNA Core Facility
- University of Texas Medical Branch - UTMB Protein Chemistry Lab
- Wadsworth Center – NYSDOH
- Wyeth Research - DNA Sequencing Group

Acknowledgments

Wyeth-BT

Mader, Michelle

Marquette, Kim

Mazaika, Erica

Wyeth-DSG members

20 participants in this study:

Phase I & II