1 51 2 51 3 51 4 51 5 51	ace Name 15149259 15149260 15149261	Size 825 865	Clip L 34	Clip R	Read	
2 51 3 51 4 51 5 5	15149260 15149261	865	34	805	THE DU	
3 51 4 51 5 51	15149261	000			CCTTTGGGCTGGAAGGTTGTGTTAGCGACCCTAATTCGACAGGACCTAATGTGGCGGGTTAAT	
4 51 5 51	13149261		18	836	CGTTIGCTGGAAAGTATACGTATGGGTCCTGCTTATTTTATT	
5 5		847	18	822	CATTIGTGCTGGAAAGGTTTTAAACGCGTTATTCAAATGCAACCGGATTTAATTAICAAGCC	
5 5	15149262	871	33	519	TIGTGCTGGAAAGCGCCCATAGCTAATCAAGTGAAAATGAGTGACGCAATAAATA	End of safe year
	15149263	847	9	783	GTTTGAAGCGGCTGCGGAATCGTGTCTGAAAGGCTAAAACCTGAAATAACAAAAAAA	Maximum cut len
6 5	15149264	860	135	789	TTTTTGGGCTGGAAGCGGCGTGAACGCCGGTGATTTAATTTCCGAAGTTGCTTTAGCAATAG	Extended ou
7 5	15149265	888	18	642	CCTTTTGTGCTGGAAAGCGCATTTAACGTTCAGTGGAACAATCGATGCGGTTTGGAAAACAG	Kmer
8 5	15149266	876	16	830	ACTTGTGCTGGAAAGCTGATTGTCCGCACCTCGCAGATTTTACTACCTCCTCAATTTTAGTGAGA	Library
9 5	15149267	889	46	575	AACCTGGGCGTGGGAAAGGGTATGGTCTCCAACAGGACACGTCAYCCAAGTTTATTTTAGCT	Sz
10 5	15149268	880	35	868	GTATATAAAAAGGGTCCTTTGTGCTGGTAAAAACAAAAATTAAACAAGTGTATCCACCGCTTL	
11 5	15149269	842	30	793	TCGTGGTGGCTTGGAAAGGCGTTATGGCCTCGTAAAAAACTGAAAACTTCAACTTGAAGC	Max
12 5	15149270	825	28	799	CCTTTTGGTGCTGGTAAAGTTGCCTGTCGTCAAAATGCTTAAGTTTCTCCAGTATATTGGCGG	Percentage Min
13 5	15149271	890	3	849	GTIGTIGCCTGGAAAAAGTTCCAATTTCTCACGAATCCTCGCAGAGGAGTTGAATCTTTCAG	V S-end
14 5	15149272	877	20	868	CCGTTGTGCTGGTAAAAAAGAGGCGATGAAGGACGTAGTAGCCTGCGATAAGCTTCGGGG	
15 5	15149273	847	72	557	TTTAAGTTTTTTTTAGGATCTACTTGGTTGGTAAAGTTATTTTTTGGGAATACGCTCGATG	
16 5	15149274	749	17	323	CATTGTGCTGGAAAGGCTTGGCTGATGTTATCCGTTGAACCTAAAATGCAGTGGTATCGAGT	3 Vector
17 5	15149275	972	26	488	AGCCGTGAGTGCTGGTAAAGGTATAATCTTTAATCCGTGCAATTTGTGATCCTACCGGACTTT	Manua
10 5	15140276	882	22	852	TTTTTGTGGCTGGAAAGCTATTTTAATTAAGCTGCCATTCTTCTAAGAAAACCTGCGGGCTAC	P Specal C
10 1	15149270	870	17	766	CATTITIGTIGCTGGAAAAGAATCGGTTCACCGTATTGTAAATTTAATTGATTTATTAAATGAAAT	
19	013149277	979	15	854	CTIGIGCIGGAAAGCIGICTITIGGIAAAIAIGCAIGGGCTICCAACAIAITTICIAACACAAG	- SONG (2) D
a Sequ	ences /	<b>9</b> Qi	salities		Information / Trades / Pebbles / Hate-pairs / A HP-Pebbles /	Repeats
	-	-	-	-		
						( gabanga )
					annes A Property and the state of the state of the state of the	

100 0

• \$1 39209

# **DNA Scissor Manual**

As a part of Arapan Project

DNA Scissor is a multi-functional software which provides users with a set of easy-to-use graphical tools to perform several operations mainly: quality trimming, vector masking, vector-like contaminants removal, and repeats identification.

By

Ххххх Үууууу, Ххххх Үууууу 5/16/2011

# Manual

# Contents

Ma	nual
I.	Main interface
1	Introduction
2	Sequences
3	Qualities
4	Information
5	Reads
6	Pebbles
7	Mate-pairs
8	MP-Pebbles
9	Repeats7
1	0 Refresh buttons
II.	Statistics
1	Sequence statistics
2	Whole statistics
3	Clipping points statistics
III.	Quality values trimming
IV.	Vector sequences clipping
1	Automatic 10
2	Vector-like contaminant11
3	Manually
4	Special case
V.	Other utilities
VI.	Repeats detection
VII	. DNA Viewer
VII	I. Contact

# I. Main interface

# 1 Introduction

DNA Scissor is a multi-functional software which provides users with a set of easy-to-use graphical tools to perform several operations mainly: quality trimming, vector masking, vector-like contaminants removal, and repeats identification.



Figure 1: The main interface of DNA Scissor.

It is composed of two essential parts: the displaying interface and the panel control. We will describe each tab of the displaying interface independently.

# 2 Sequences

63	🗙 🖌 🗆 (	lip L	🗆 Clip	R Or	iginal ə 🛷 🍖		1	8			<b>N</b>	irrors	
	Trace Name	Size	Clip L	Clip R	Read	-	5	utomat	ic				
1	857348472	906	100	710	GGGGTGCCNCCGGGGCCGGGCCNTTTTGANCCCCATATNGATTTGT								
2	857348473	940	34	818	GCTTTGTGCTGGAAAGAGGGGGATTCTACTGCTAACGCAGCTAATCGA	- 11	Method				5'-e	nd + 3'-	end o
3	857348474	907	25	673	CTGTGTGTGCTGGTAAGTGGCTGTAGGTGCTGACTATCTGAAAGCTCAG	- 11							1
4	857348475	947	37	830	GGAGGGGNNATANAATGGTTCTAATGACTATCACGATTCGACGTGAC	- 11	1	Ind of si	afe zone	(E)			500
5	857348476	999	31	288	GGCTATTGTGCTGGTAAAAATTGTAACAGGCGACGCACGTTGATGCG		N	laximun	n cut len	gth (M)			100
5	857348477	893	35	652	CCTATTGTGCTGGAAAATATTGTTCAAGTCGGTATGCCAATTCGTGTG		6	Exten	ded cut	length (E)	4)		200
,	857348478	1041	81	845	GAGGCGCGCACAGGATAGTTAAAGGCCCAGGTGGGTCTGGAAAAA	- 11		mar				_	0
3	857348479	1014	82	660	CGGGCGCGCGCCGTTTGGCATGGATGGTGCTGGAACGTACCTTTC	- 11	kmer			-		•	
,	857348480	1009	80	679	GAGAGATGAAAGTATCGTAAAGGCCCCACTAGTCTGAAAGACAATTC	- 11		lorary			Non	2	0
10	857348481	938	74	840	AGGCGCGCGAGGTTAGTTTTGGTTAAGCCACTGCGTTTTTAACTAGA	- 11	May framiency found			3		0	
11	857348482	902	32	649	TTTGAGGTTGAAGGTCAGATTTCTCAGCACTTGCTATATCACCATCGA	- 11			Max freq	uency tou	nd		0 
12	857348483	926	52	774	GCAGGAGGGNGGNNNNGGACATTGAACAATGATTGCCGTTGGCAA	- 11			Min freq	Jency			
13	857348484	937	99	848	GGGCCCGAGAGTTAAGTTTCGGACAGGTATCCTGTTTTTAACCCATTA	- 11	⊻ 5'-end end-3				3no-3 (		
14	857348485	913	28	682	GGNNTTTGATTTCTTTCACTCAGCCATGCTTTCTTTCTGAATTTCATC	- 11			+	1		-	1
15	857348486	930	37	527	GCTTGGTGCTGGTAACAAGTATATAATCTTCATTAAATTTTCTTCCATCT	- 11		M	ate zoni laximum	cut lengt	h		
16	857348487	976	380	610	GGGGAGAAAAAAGTTATAGAGAAGCTGCTGATTTAAAAAAGAAAC	- 11		L	eft end c	f safe zon	e		
17	857348488	881	96	749	GCCTTTGTGCTGGCAAACCCCTGTGACTCAGTACTTTTACACGCAGC	- 11		Ri	ght end	of safe zo	ne		
18	857348489	945	31	837	CCTGGTGCTGGAAAATGTCATGACGGTATGCACTATCTGGGTAGCGA								
19	857348490	1002	33	758	CCTTTGTGCTGGAAAACTGCACAGAAACACAGCAGTGTGGTTCAGC								
20	857348491	959	70	792	GGGCAGGTGGGTGTTTANTCCCATCGCGTCCTCGNATTGAACNCCA								
21	857348492	1035	82	869	AAGGGTGAAATTATCGTGTAAGGACCGATCCGTCTGGTAAACTTCTTT		5	vector-li	ke conta	minant			
22	857348493	1007	35	813	GGACTAGGTGCTGGAAAGACCCCATTTCAGTGAGTTTGAGCCTGTGT		1	Manually	Y				
23	857348494	921	58	691	TGTATCAATAGCGCCTNTTGNTTTTGAAAACATTGGAGATAACCTGAT		0	Special o	case				
24	857348495	946	30	749	GGATCTGTGCTGAAAACGACCATAAGACAACCCGTTGCAACTGCTTA								
25	857348496	1030	65	371	TGGAGGGGGGGGGGGGTAGGTTTGTAAAACTTTGATAACANTTCGTAAATTGCT				Dafaidt	1 8 put	and I	× 400	

The loaded sequences can be shown in this tab along with the following essential information:

Trace Name	the sequence identifier.
Size	the length of the sequence
ClipL	the Left Clipping point of the sequence.
ClipR	the Right Clipping point of the sequence.
Read	DNA sequence.

Clipping points can be gotten from the original data and updated by DNA Scissor clipping/trimming operations. In case the user wants to retrieve the original clipping points or reset them, the tool provides some useful operations  $\Box \operatorname{clip} R \ \overline{\operatorname{Original}} \ \checkmark$  so that the user can work conveniently.

# 3 Qualities

The quality values are shown in this tab along with its corresponding Trace Name.



Remark: It is not suitable to display reads and quality values in case of dealing with large amount of data because of memory usage. DNA Scissor does not display them by default unless the user select this option in the preferences menu.

## 4 Information

The information concerning each sequence is shown here. It comprises from:

Trace Name, Temple ID, Library ID, Trace End, Insert Size, Insert Deviation, Insert Min, Insert Max and Insert Mean.

													1	9	-		Error		
			C. BARRIER			housed when								Automa	tie .				
	515149259	GCRAIDS	T12201	- ype	e c	750	225	0	10	a reean				Vector-I	ke conti	aminant			
	515149260	GCBEU07	111201	Ra		750	225		0	0									
	515149261	OCRAUII	111201	Pa		750	225	-	0	0		- 8	× ×	ector C	empan	son Param	eters		- 14
	515149262	GCBBX/93	T12201	Pa	c.	250	225		0	0		- 1		Mivect	or samp	005		30	1
	515149263	GCBCC59	113201	Pa		750	225		0	0		- 1		Agains	t N sam	ples		300	10
	515149264	OCBBI 33	113201	84		750	225	-	0	0		- 1		Vector	length n	node Defa	it.		0
	515149265	GCBC069	113201	Pa		750	225		6	0		- 1							5
	5151,09265	GCBEZ14	111201	Pa		750	225		0	0		- 1							1
	515149267	OCBAIDS	113201	22	0	250	225		0	0	D.			Libe	wy	LD SIZE	Score		Len
	515149268	GCBAU19	T11201	Pa	0	250	225		0	0		- 1		10473	226		43.6	10	
	515109269	GCBBL 22	T13201	Pa	0	750	225		0	0		- 1				-	2207.0	20	
12	515149270	GCBCC59	113201	22	0	250	225	0	0	0		- 1		10477	-		242	10	
0	515149271	GCBC069	T13201	Pa	R	750	225	0	0	0		- 1		TIASA		144	1526.5	33	
	515149272	GCBEZ14	T13201	Pa	n	750	225	0	0	0		- 8		T1320		24672	1310.5	23	
15	515149273	GCBAE36	113201	Pa	F	750	225	0	0	0		- 1							
16	515149274	GC88H66	T13201	Pa	F	750	225	0	0	0		- 1							
17	515149275	GCBD511	T13201	Pa	P	750	225	0	0	0		- 8							
	515149276	GCBEJ31	T13201	Pa	F	750	225	0	0	0		- 1							
19	515149277	GCBBT76	T13201	Pa	F	750	225	0	0	0		- 1							
10	515149278	GCBFA51	T13201	Pa	F	750	225	0	0	0		- 1		_					
11	515149279	GCBCK96	T13201	Pa	F	750	225	0	0	0		- 1							
12	515149280	GCBEV41	T13201	Pa	F	750	225	0	0	0		- 1	1	Manual	У				
23	515149281	GCBGD81	T13201	Pa	F	750	225	0	0	0		- 8	1	Special	case				
14	515149282	GCBFY71	T13201	Pa	F	750	225	0	0	0		- 1							
15	515149283	GCBAE 36	T13201	Pa.	R	750	225	0	0	0				1.00	Dada da	A Real		a secolo i	

# 5 Reads

We define clean reads those which are not mate-pairs or pebbles.

Each line contains a clean read ID along with its template ID.



# 6 Pebbles

Pebble is a discarded read through the different cleaning processes. As a matter of fact, the discarded reads will be moved to the pebbles tab by the will of the user. Each line contains a read pebble ID along with its template ID.

Just to note that pebbles can be very useful in the finishing part of genome assembly process in order to get much more longer scaffolds.



#### 7 Mate-pairs

Sometimes, the data also include mate-pairs file. In such a case, mate-pairs will be downloaded to this tab in which each line contains two mate-pairs' ID along with its template ID.



# 8 MP-Pebbles

The discarded mate-pairs through any cleaning process can be moved to this tab for further analyses. Each line contains two mate-pairs pebbles' ID with the template ID for each.



## 9 Repeats

DNA Scissor detects repetitive sequences by calculating the k-mer distribution for all reads. If some k-mers appears more frequently and exceeds a predetermined threshold, it may be originated from a repetitive sequence. The most frequent k-mers and the detected repeats can be displayed in this tab such that each line contains a repeat ID.

.3	K					or 😫 🥢 🐴	
Read	s Kmers					Repeats	
	Template ID	Tace Name				Mechoa	de urujn c
1	ANIN9231	50-1213706				Number of reads	35797
2	ANIN7396	504209798				Reads average length (bp)	647 🗘
3	TIH1343	504282318			1	Genome's size (Kbp)	2895.00
4	ANIN13169	504257038				Coverage (x)	
5	TIH6090	504290581				Kimers (hn)	10
6	TIH6047	504290500				Threadworld for the	100
7	ANIN1223	504202609					100 9
8	ANIN14005	504255288				Threshold	100 2
9	ANIN13839	504255103					
10	TIH1689	504282903					
11	ANIN10985	504253683					
12	ANIN9756	504256330					
13	11146578	504291467					
14	ANIN7227	504212774					
15	ANIN5562	504211230					
16	ANIN6835	504212190					
17	ANIN8452	504208637					
18	ANIN5989	504211544					
19	ANIN11046	504254830					
20	ANIN6838	504212193			_		
21	ANIN5934	504211795			_		
22	ANIN12335	504258485			_		
23	POY902	504263393					
						🕐 Default 📢 R	fresh Apply

# 10 Refresh buttons 🚺

Since we deal with huge data, the displaying process might slow the different operations if it was done automatically. For this reason, DNA Scissor does not automatically display the result in general, but it sends signals to refresh buttons so that the user can click on them in case s/he wants to see the changes.

# **II. Statistics**

Jr-

# 1 Sequence statistics

To show the statistics concerning a specific sequence, you can select the target sequence in the sequences tab and click the following button:



# 2 Whole statistics

This window gives an overall statistics concerning the data being processed by clicking on:



Frame	
Number of all reads	4307
Number of F reads	2177
Number of R reads	2121
Number of N reads	0
Number of proper reads	4307
Number of pebbles	0
Number of mate-pairs	0
Number of MP pebbles	0
Number of repeats	0
Reads average length (bp)	871
Coverage	14
Genome size (bp)	267000
Number of A nucleotides	1019055
Number of T nucleotides	1025409
Number of C nucleotides	857483
Number of G nucleotides	853514
Number of N nucleotides	0
GC-Content	45.5602
Number of reads that include N	0
Expected number of contigs	0.0115202
	•

## 3 Clipping points statistics

DNA Scissor can also give statistics on the left clipping points for the current version.

Concerning the meaning of different rows names please consult the part concerning vector trimming.

Clip L	Clip R
1773	
1	
201	
39709	
23.0506	
24.9296	
39195	
21.4014	
20.2427	
514	
148.815	
27.2398	
	Clip L 1773 1 201 201 23.0506 24.9296 24.9296 21.4014 20.2427 514 148.815 27.2398

# III. Quality values trimming

DNA Scissor can detect the longest low quality regions from the 5'-end and 3'-end of reads depending only on two parameters:

- Q the minimum quality score (20 by default),
- T the threshold (6 by default).

If DNA Scissor detects T consecutive nucleotides whose quality values are greater than Q, it stops exploring and marks the trimming points since it considers the T nucleotides as the starting section of the good quality region.

The case when there are some low quality regions within the good quality region, DNA Scissor does not detect them. However, it can deal with reads which contain some unknown 'N' nucleotides such that the user can then discard or mark them as 'pebbles' for further utilization during the genome assembly process.

Cool quality region

Low quality regions

Low quality region starting section

Parameters

Minimum quality score

20

Threshold

6

Threshold

6

Parameters

Image: score

20

Threshold

6

Image: score

20

Threshold

6

Image: score

20

Threshold

6

Image: score

Im

Remark 1 : Note that clipping points get the

maximum of the index position of new calculated points and the provided clip information (by NCBI Trace Archive). In case the user does not want to incorporate the provided clipping information, our tool can simply reset them before the process starting.

**Remark 2** : It is preferably to run quality trimming after vector sequences trimming.

# IV. Vector sequences clipping

#### 1 Automatic

The program is able to detect the endmers (the k-mer that represents the end of vector) of vector sequences at the 5'-end of reads. In case short libraries are used, it can be forced to detect vector sequences at the 3'-end as well. For getting more accurate clipping points, DNA Scissor can detect libraries automatically. It is preferable to let the software detect the clipping points for each library in order to get more accurate results.

DNA Scissor can detect vector sequences endmers without prior knowledge of cloning vectors.

The algorithm is uses two main parameters:

- M : The maximum cut length (100 by default)
- EM : The extended maximum cut length (2\*M by default).
- Kmer: the length of the short read (seed).
- ٠

However, only M the maximum cut length parameter which should be defined by the user. DNA Scissor adjust other parameters automatically.

1 8 😫 🍕	Errors								
🐜 Automatic									
Method	Method 5'-end + 3'-end								
Threshold value (T)	1 🔺								
End of safe zone (E)	500 束								
Maximum cut length (M)	100 荣								
Extended cut length (EM)	200 🛓								
Kmer	8 👤								
Library	All 🔻 🚺								
Size	39709								
Max frequency found	8210								
Percentage Min frequency	20 🚔 🔲								
💟 5'-end	3'-end 📃								
💐 Vector-like contaminant									
/ Manually									
Special case									
50% 🎲 Default 🚺 Ref	resh Apply 78%								

To give more flexibility to the algorithm implemented in DNA Scissor, the Kmer length is automatically changed according to the maximum cut length parameter. The minimum k-mer length is 4 (resp. 8) when the maximum cut length is very short (resp. very long). Table 1 shows different provided lengths.

Range	<i>k</i> -mer length
$90 \le M$	8
$50 \le M \le 90$	7
$30 \le M \le 50$	6
$20 \le M < 30$	5
M < 20	4

Table . Maximum cut length ranges with the corresponding k-mer lengths.

In fact, some shotgun raw data (from NCBI Trace Archive) have very short vector sequence or adapter (e.g. of length 5). Such situations lured us into the idea of changing the k-mer length according to the maximum cut length. Nonetheless, the user can change the k-mer length whenever it is possible.

## 2 Vector-like contaminant

Since vector sequences do not only appear at the 5'-end of reads, but also may occur within the reads. The good point of DNA Scissor is the ability to detect a very reliable vector sequence automatically for each library without prior knowledge of the vector. As a result, despite the luck of vectors, user can get a very reliable vector sequence for each library and deal with contaminated reads. Note that when a vector sequence appears

DNA Scis	sor					
e <u>T</u> ools	Help			-		
2.	E ()	3				
63	🗙 💉 🗏 di	ip L	🗐 Clip	R All	· · · · ·	y w State Construction
	Trace Name	Size	Clip L	Clip R	Read	Second Automatic
1	515149259	825	107	805	CCTTTGGGCTGGAAGGTTGTGTTAGCGACCCTAATTCGACAGGACCTAATGTGGCGGTTAAT	Vector-like contaminant
2	515149260	865	13	836	CGTTIGCTGGAAAGTATACGTATGGGTCCTGCTTATTTTATT	Vector Comparison Parameters
3	515149261	847	15	822	CATTIGIGCTGGAAAGGTTTTAAACGCGTTATTCAAATGCAACCGGATTTAATTTATCAAGCC	M vector samples 30 👘
4	515149262	871	12	519	TTGTGCTGGAAAGCGCCCATAGCTAATCAAGTGAAAATGAGTGACGCAATAAATA	Against N samples 300 文
5	515149263	847	21	783	GTTTGAAGCGGCTGCGGAATCGTGTCTGAAAGGCTAAAACCTGAAATTACAGATGAAAAGA	Vector length mode Default
6	515149264	860	16	789	TTTTTGGGCTGGAAGCGGCGTGAACGCCGGTGATTTAATTTCCGAAGTTGCTTTAGCAATAG	🗙 🦉 All 👻 🔪
7	515149265	888	16	642	CCTTTTGTGCTGGAAAGCGCATTTAACGTTCAGTGGAACAATCGATGCGGTTTGGAAAACAG	Library Lib Size Score Length Vector
8	515149266	876	14	830	ACTTGTGCTGGAAAGCTGATTGTCCGCACCTCGCAGATTTTTACTACCTCAATTTTAGTTGAGA	1 T10223 141 1902 46 AGGGCACGAATGTGTTGG
9	515149267	889	19	575	AACCTGGGCGTGGGAAAGGGTATGGTCTCCAACAGGACACGTCATCCAAGTTTATTTTAGCT	2 T1898 388 488 56 AAGGTACCGGTATCGACT
10	515149268	880	28	868	GTATATAAAAAGGGTCCTTTGTGCTGGTAAAAACAAAAATTAAACAAGTGTATCCACCGCTTT	3 T13210 14473 1610.5 21 CATGTTTGTGGCTTGGAA
11	515149269	842	20	793	TCGTGGTGGCTTGGAAAGGCGTTATGGCCTCGTAAAAAAACTGAAAACTTCAACTTTGAAGC	4 T13201 24672 643.5 26 TATAAAAAAGCAATCTGG
12	515149270	825	18	799	CCTTTTGGTGCTGGTAAAGTTGCCTGTCGTCAAAATGCTTAAGTTTCTCCAGTATATTGGCGG	
13	515149271	890	17	849	GTTGTTGCCTGGAAAAAGTTCCAATTTCTCACGAATCCTCGCAGAGGAGTTGAATCTTTTCAG	
14	515149272	877	70	868	CCGTTGTGCTGGTAAAAAAAGAGGCGATGAAGGACGTAGTAGCCTGCGATAAGCTTCGGGGG	
15	515149273	847	37	557	TTTAAGTTTTTTTTAGGATCTACTTGGTTGGTAAAGTTATTTTTTGGGAATACGCTCGATG	
16	515149274	749	14	323	CATTGTGCTGGAAAAGGCTTGGCTGATGTTATCCGTTGAACCTAAAATGCAGTGGTATCGAGT	
17	515149275	922	20	488	AGCCGTGAGTGCTGGTAAAGGTATAATCTTTAATCCGTGCAATTTGTGATCCTACCGGACTTT	
18	515149276	882	17	852	TTTTTGTGGCTGGAAAGCTATTTTAATTAGGCTGCCATTCTTCTAAGAAAACCTGCGGGGCTAC	
19	515149277	870	16	766	CATTTIGIGCIGGAAAAGAAICGGTICACCGTATIGTAAATTTAATIGATTTAATGAAAT	
20	515149278	878	13	854	CTTGTGCTGGAAAGCTGTCTTTTGGTAAATATGCATGGGCTTCCAACATATTTTCTAACACAAG	/ Manually
21	515149279	912	20	824	CCAGTTTGTGCCTGGAAAAGGGACACAATAAAGAAGAAAATGAACAAATCGCGCAGCGTAA	Special case
22	515149280	758	75	758	GACCGCATTTTTCTGATCTGAAAATTTTTCACACTGATGCCGCCGGGAAGGGAATATTCCCCTACC	
23	515149281	841	14	832	ATTTGTGCTGGAAGTGTCTCTTTAATTTTATCACAGCTCATGCAAATCCTCGTTTCTTCTGTAAA	🛟 Default 🚺 Refresh 📉 Apply
	A					
J Se	equences / 🚺	¶,Qu	alities		Information / Reads / Pebbles / _ The Mate-pairs / MP-Pebbles / Repeats /	<u></u>

beyond the 5'-end or represents the whole read, it is considered as 'contaminant' or in a specific term 'vector-like contaminant'.

DNA Scissor requires two parameters in this case:

- *p* = number of random competitive vector sequences (30 by default)
- *q* = another number of random vector sequences (300 by default)

After that, just click on  $\geq$  to start the process.

Remark: This task should be done immediately after detecting the vector sequences clipping points. Please do not trim data or run quality trimming process.

The obtained vector sequences can be treated as explained in the following.

## 3 Manually

Whether the user generated reliable vector sequences from the previous section or prefers to use his/her own vectors. In this case, we provide a set of tools to remove all vector sequences manually. Users have a full control of vector management. The affected DNA reads by vectors in the center may be discarded, cut at specific location or marked as pebbles.

User should define his/her desired parameters of the dynamic programming by clicking on the submenu called "Preferences".

💽 Frame			
System Display Output files Dynamic Programming	Gap score Match score Mismatch score Similarity score (%)	-1.00 \$ 1.00 \$ -0.50 \$ -0.50 \$	
			OK Cancel

🖌 🌮 💐 🍓 🦄 Errors
Ver Automatic
🗿 Vector-like contaminant
/ Manually
Precise vector      Vector file
Adapter
Sequence middle region
20 🛓 80 束
Vector sequence is found in the middle
<ul> <li>Discard</li> <li>Pebble</li> <li>Keep the right side part</li> </ul>
Multiple occurrences of vector sequences
<ul> <li>Discard</li> <li>Pebble</li> </ul>
Parameters
Break point distance 0 束
Method Approximate
🖉 Special case
Default S Refresh Apply

# 4 Special case

The vector sequences do not only appear in the 5'-end of reads, but also may represent the whole read or appear elsewhere in the read. DNA Scissor can also deal such situations. Before running this section the dynamic programming parameters should be defined.

🖌 🌮 😫 🍓 🏂 Errors
🛰 Automatic
💐 Vector-like contaminant
/ Manually
🖉 Special case
Contaminant  A B A B C Read Sequence parts Contaminant  Discard  Pebble
<ul> <li>Keep only A+B +C</li> <li>A+B+C is a pebble</li> </ul>
Method Approximate
Default Refresh Apply

# V. Other utilities

#### Discarder

Discard sequences of minimum length and those which include unknown 'N' nucleotides.

#### Trimmer

The user can keep a specific part of sequences.

#### Cutter

The user can remove a number of nucleotides at the head and the tail of sequences.

#### Collapser

DNA Scissor is able to delete the duplicate sequences such as the case of PCR duplicates which often appear in the NGS (Next-Generation Sequencer) data of PCR-amplified DNA sequences. The tool provides three modes to delete duplicates (Light, Medium, Heavy). The Light mode is the fastest since it only allows DNA Scissor1 to delete sequence whose have the same sequence ID. The Medium mode lets the software to check sequences IDs then the similarity of sequences. The slowest mode is the Heavy mode since it neglects checking sequences IDs and focus only on the similarity of sequences without taking into account their IDs, with which we can remove PCR duplicates.

#### Pebbles

The discarded sequences can be marked as pebbles in case the user wants to use them in the finishing part of the genome assembly project.

scarder				
Minimum sequence le	ength			15 ≑
	Discard sequenc	es with unkn	own (N) ba	ises 📃
	Move discarded	sequences t	o Pebbles S	Sets 📃
			X	Apply
Trimmer				
Keep bases betwee	n	0 🖨 and		0 🌩
Only selected sequ	iences			Apply
Cutter				
Remove	0 🚔 bas	ses at the he	ad of sequ	ences
Remove	0 🖨 ba	ses at the ta	il of sequer	nces
Only selected sequ	iences 📃		A	pply
Collapser				
Collapse duplicate	sequences 📃			
Mode	Medium	T		
Check by trace nam (Average)	e then read seque	nce.	A	pply
🗳 De	fault 🚺 Refr	esh 🚺	Apply	

# VI. Repeats detection

This part is heavily parameterizable. The user should know at least the value of one parameter: the genome size or the coverage value. If one of them is known, the other can be calculated automatically. We also facilitated calculating other parameters depends on the genome size or the coverage value. We provide two methods: 'Oligomers' and 'de Bruijn k-mers'. Oligomers method is faster but it is not as efficient as de Bruijn k-mer based method. The algorithm detects repetitive sequences by calculating the k-mer distribution for all reads. If some k-mers appears more frequently and exceeds a predetermined threshold, it may be originated from a repetitive sequence.

The critical parameters of this part is the coverage and threshold for 1x. They should be tuned carefully. However, in case the repeats were mistakenly detected. The user can delete them and tunes the parameters again.

For updating the other parameters please click on the buttons 5 before you start applying the process.



#### VII. DNA Viewer

The selected sequence can be seen in 3D mode whether in helix format or straight view as shown below. Java virtual machine must be installed in order to launch this feature.



The viewer is controlled by mouse buttons (the left and the right buttons).

# VIII. Contact

For further questions, please contact us at <u>dnascissor@hgc.jp</u>.