JABG
Journal of Animal Breeding & Genomics

JABG. 2023 March, 7(1)1-8
https://doi.org/10.12972/jabng.20230001

Print ISSN 1226-5543
Online ISSN 2586-4297

Technical Protocol

단일염기서열변이 검출을 위한 오픈소스 소프트웨어 STADEN 패키지 기반 Sanger sequencing 결과 데이터 분석

이유주[1,†], 이재봉[2,†], 임규상[1], 김봉기[1], 김지혁[1], 박희복[1,*]

[1]공주대학교 동물자원학과, [2]전북대학교 인수공통전염병연구소

Practice of STADEN package to detect SNPs

Yu-Ju Lee[1,*], Jae-Bong Lee[2,*], Kyu-Sang Lim[1], Bong-Ki Kim[1], Ji-Hyuk Kim[1], and Hee-Bok Park[1,†]

[1]Department of Animal Resources Science, Kongju National University, Yesan 32439, Korea

[2]Korea Zoonosis Research Institute, Jeonbuk National University, Iksan 54531, Korea

[*]**Corresponding author: Hee-Bok Park,** Department of Animal Resources Science, College of Industrial Sciences, Kongju National University, Yesan 32439, Republic of Korea, Tel: +82-41-330-1244, Fax:+82-41-330-1249, E-mail: heebokpark@kongju.ac.kr

[†] These authors contributed equally to this work.

ABSTRACT

Nucleotide sequences of DNA contain information about the genetic characteristics of organisms, and in particular, single nucleotide polymorphism (SNP) that shows a difference of one nucleotide sequence in DNA may also affect phenotypic variations. Therefore, it is necessary to precisely read the nucleotide sequence information for the improvement of economically important traits in domestic animals. Sanger sequencing, one of the DNA sequencing methods, is widely used for the verification of genome-wide association study results and the analysis of DNA sequences from small- to medium-scale experiments. Several programs have been developed to analyze such sequencing results, of which STADEN package (https://staden.sourceforge.net), one of the open-source programs, is a useful program with a great cost advantage compared to other commercialized programs that require a license fee. Through the Pregap4 and Gap4 programs installed in the STADEN package, sequence assembly and editing analysis with user-provided settings can be conducted to detect SNPs. The nucleotide sequence results of the STADEN package can be used for follow-up bioinformatic analyses such as database query or phylogenetic analysis.

Key words: DNA sequence, Sanger sequencing, STADEN package, Pregap4, Gap4, SNP

Introduction

유전현상을 조절하는 DNA (Deoxyribonucleic acid)의 기본단위인 개별 뉴클레오타이드(nucleotide) 서열정보를 정확히 해독하는 것은 유전현상연구와 이를 이용한 유용한 유전형질 개량에 중요한 역할을 한다. 이를 위하여 1970년대, DNA 염기서열을 해독하기 위한 기법 개발이 시작되어 단일 실험에서 300~500 bp의 DNA 서열을 해독할 수 있는 염기결합순서해독 (DNA sequencing) 방법이 개발되었다(Sanger et al, 1975; Maxam and Gilbert, 1977). 특히, Frederick Sanger (1918~2013)는 자신이 개발한 디데옥시유사체(dideoxy analog)와 DNA 중합효소를 이용한 빠르고 효율적이며 정확한 DNA 시퀀싱 방법(Sanger sequencing)을 이용해서 bacteriophage ϕX174의 DNA 염기서열 5,286개(Sanger et al., 1978), 사람의 미토콘드리아 DNA 서열 16,569개(Anderson et al., 1981), 그리고 *E. coli*를 감염시키는 λ 파지

Received December 07, 2022 **Revised** March 24, 2023 **Accepted** March 28, 2023

의 DNA 서열 48,513개를 완전하게 해독하였다(Sanger et al., 1982). 이후, 방사선동위원소 기반의 Sanger sequencing 염기서열 검출법은 형광색소기반의 자동검출법으로 발전되어 인간게놈프로젝트의 DNA 염기서열분석에서 중요한 역할을 하였고, 현재도 널리 사용되고 있다(Smith et al., 1986; Lander et al., 2001).

이와 같이 DNA sequencing 데이터의 양 또한 커짐에 따라, 이를 효율적으로 처리하기 위한 컴퓨터 프로그램 개발의 필요성이 대두되었다. 1977년에 SEQEDT, TRANSQ와 같은 간단한 DNA 시퀀스 분석 프로그램들이 보고된 후(Staden, 1977), 자동화된 sequence assembly와 같은 지속적인 발전이 있어왔다(Staden, 1982). NGS (Next Generation Sequencing)가 상용화된 현재에서도 실험실별로 Sanger sequencing 결과의 분석 및 편집은 여전히 많이 이루어지고 있다. 이와 같은 작업이 많은 경우 라이선스 비용이 많이 들어가는 상용 프로그램으로 분석되고 있는 상황에서 STADEN 패키지의 경우 동일한 기능이 무료로 제공된다는 점에서 비용적으로 매우 유용한 프로그램이다. STADEN 패키지는 Optimal alignments in linear space와 Dynamic programming에 기초한 프로그램으로(Myers et al., 1988; Huang, 1994), STADEN 내에는 고유의 기능을 수행하는 여러 프로그램으로 구성되어 있으며, 패키지 내 프로그램이 다양한 만큼 그 활용도의 폭 역시 매우 넓다. 따라서 본 고에서는 STADEN 패키지를 이용하여 벡터 DNA를 매개하지 않은 PCR 산물의 염기서열해독을 위한 Direct Sanger sequencing 결과를 편집하고 분석하여, 최종적으로는 목표로 하는 SNP의 유전자형을 결정하고자 한다.

Practice

STADEN 패키지는 염기서열 handling과 분석에 사용되고 있는 소프트웨어로서 sequence assembly부터 유전자, 제한효소 작용위치 및 motif 동정 등 다양한 생물정보학 분석수행이 가능한 라이선스 비용이 들지 않는 무료의 소프트웨어이다. STADEN 패키지는 Gap4, Pregap4, Trev, Spin 등의 여러 프로그램으로 구성되어 있으며, 아래의 URL에서 다운로드하여 설치할 수 있으며, 본고에서의 컴퓨터 운영 소프트웨어는 Microsoft Windows 11을 기준으로 한다.

Table 1. Programs in STADEN package

프로그램명	기능
Gap4	시퀀스 어셈블리, read pair 데이터를 기반으로 하는 contig 정렬, 시퀀스 비교를 기반으로 하는 contig joining, 어셈블리 검사, read pair 분석 및 contig 편집
Pregap4	시퀀스 어셈블리 또는 분석을 위한 trace 데이터를 prepare하는 데 필요한 processing 설정(trace 형식 변환, quality analysis, vector clipping, contaminant screening, repeat searching, mutation detection)
Trev	ABI, ALF, SCF 및 ZTR trace 파일 뷰어 및 편집
Prefinish	Partially completed 시퀀스 어셈블리 분석 및 가장 효율적인 set of experiments to help finish the project 제안
Tracediff and hetscan	Trace data를 reference traces와 비교하여 돌연변이를 자동으로 annotation
Spin	염기서열을 분석하여 유전자, restriction sites, 모티프 등을 찾으며, translation, find open reading frames, count codons 등의 작업 수행

https://staden.sourceforge.net/

위의 프로그램들 중 본고에서 사용될 프로그램은 Gap4와 Pregap4이며, sequence의 경우 *Carassius auratus* (NCBI accession number: MT155797~MT155801)와 *Carassius cuvieri* (NCBI accession number: MT155802~MT155806)의 *CYTB* 유전자 sequence를 사용했다(Kim et al. 2020). Pregap4 프로그램을 이용하여 원하는 설정에 맞춰 DNA sequence assembly를 수행할 것이며, Gap4 프로그램을 이용하여 PCR에 의해 생성된 DNA 단편의 염기서열을 분석하고, 목표로 하는 SNP의 chromatogram과 그 유전자형을 확인할 것이다.

Pregap4를 이용한 DNA sequence assembly 수행

Pregap4 프로그램을 실행하게 되면, 아래와 같은 화면이 보이게 된다. Add files를 클릭하여 시퀀싱 데이터를 프로그램으로 불러온다. 시퀀싱 데이터 파일은 ABI, ALF, SCF, CTF, ZTR, FASTA, txt 형식이어야 하며, 다루고자 하는 파일들의 형식이 모두 같을 필요는 없다. 하지만 시퀀싱 데이터의 경로 폴더명이 한글로 작성된 경우, 결과 파일이 만들어지는 과정에서 error가 발생하니 경로 폴더명을 영어로 작성해야 한다. 또한 시퀀싱 데이터를 입력할 때에는 Figure 1과 같이 Forward와 Reverse, 개체 순서에 주의하여 순서대로 입력하는 것을 권장한다.

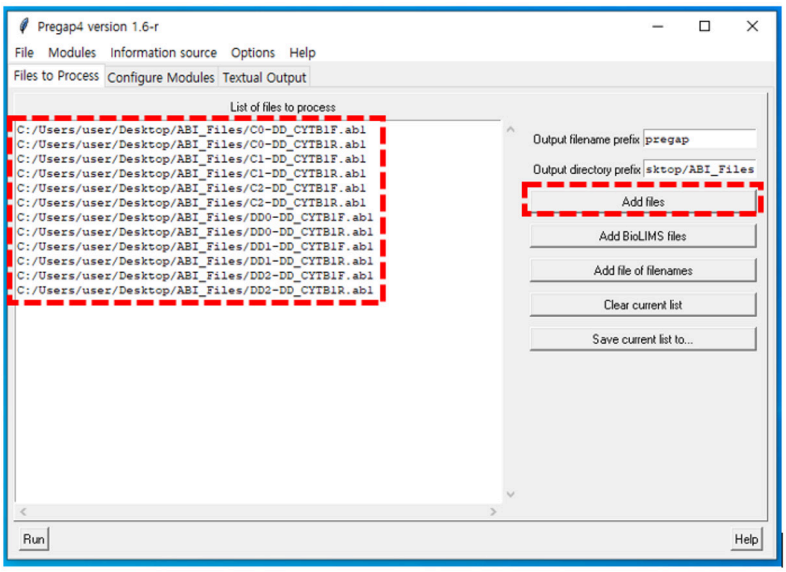

Figure 1. Display of Files to Process. The figure above is the display of pregap4's files to process panel. The sequencing data to be analyzed can be loaded by clicking “Add files (right red box)”, and you can check the imported data files in the list of files to process (left red box).

Configure Modules를 클릭하게 되면 아래와 같은 화면이 보이며, 왼쪽에는 모듈 목록이, 오른쪽에는 모듈에 대한 구성 패널이 표시된다. 모듈 앞의 [ ] 표시는 해당 모듈이 비활성화 된 상태임을 나타내며, [x] 표시는 해당 모듈이 활성화된 상태임을 나타낸다. 모듈들 중 활성화할 모듈은 Estimate Base Accuracies, Trace Format Conversion, Initialise Experiment Files, Augment Experiment Files, Quality Clip, Gap4 shotgun assembly로 Figure 2와 같이 이것들 외의 다른 모듈들은 비활성화 해준다.

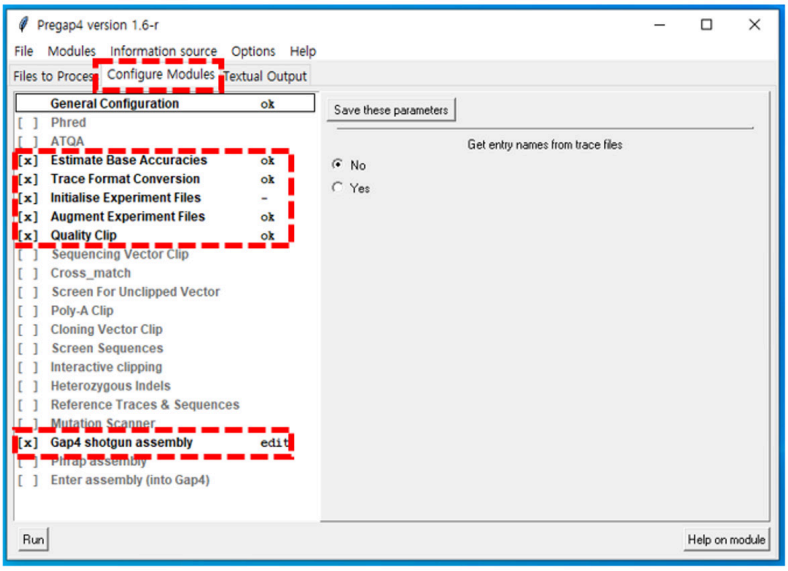

Figure 2. Display of Configure Modules. The figure above is the display of pregap4's configure modules panel. To perform the assembly of imported Sanger sequencing data files, six modules in the red box are expected to be activated, and the other modules are deactivated.

Quality Clip 모듈을 클릭하면 Figure 3와 같이 오른쪽에 구성 패널이 표시된다. Quality Clip 모듈은 sequence 품질이 낮아 assembly를 수행하기에 적합하지 않은 sequence 위치를 결정하는 기능을 하며, 이 과정에서 qclip 프로그램이 이용된다. Clip mode는 "by base calls"와 "by confidence" 두 가지로, 여기 에서는 "by confidence" 모드로 clipping을 수행하였으며, 해당 모드는 clipping quality를 결정할 때 Phred-scaled confidence 값을 사용하여 clipping을 수행한다(Ewing et al., 1998). "by confidence" 모드의 Window length는 신뢰도가 평균화 될 window의 길이를 설정해주며, 여기에서는 Window length를 25로 설정해주었다. Average confidence는 시퀀스가 좋은 품질로 받아들여지기 위한 최소 평균 신뢰도를 설정해주며, 여기에서는 Average confidence를 35로 설정해주었다.

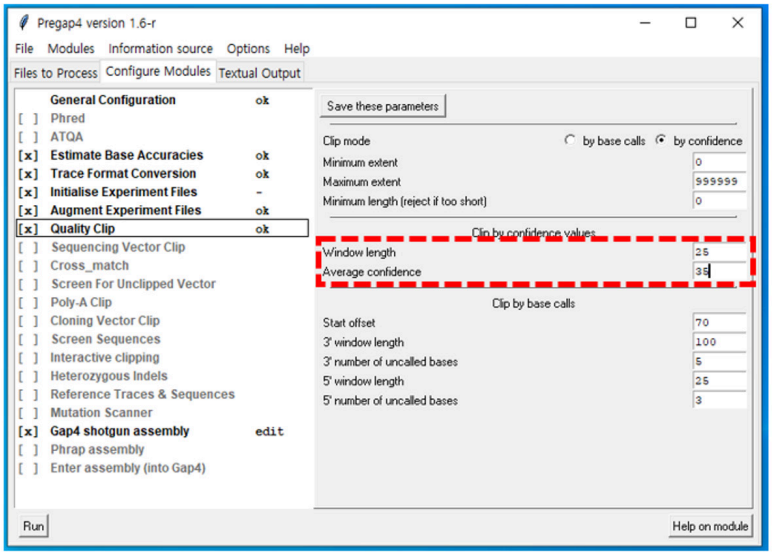

Figure 3. Configuration panel of Quality Clip. The figure above is the configuration panel of the quality clip module of Pregap4. This module determines the locations of low-quality sequences that are not suitable for assembly. The window length in the red box is the length of the window to average the reliability, and the average confidence is the minimum average confidence level for the sequence to be accepted as good quality.

Gap4 shotgun assembly 모듈을 클릭하면 Figure 4와 같이 오른쪽에 구성 패널이 표시된다. 이 모듈은 Gap4 자체 어셈블리 엔진을 사용하여 시퀀스를 assembly한다. 이 구성 패널의 Gap4 database name에 생성하고자 하는 output 파일의 이름을 입력하고 Create new database를 클릭한다. 이때 output 파일이 생성될 디렉토리에 동일한 이름의 기존 데이터베이스가 있는 경우 경고 없이 자동으로 덮어 쓰이니 주의해야한다. 그리고 Maximum percentage mismatch를 20으로 설정해준다. Maximum percentage mismatch는 gap4 내의 주 assembly 매개변수를 제어해주는데, gap4에서 각 reading을 비교할 때 maximum percent mismatch를 초과하지 않는 reading이 입력된다. 설정을 모두 끝내고 난 후 DNA sequence assembly를 진행하기 위해 하단 좌측의 Run을 클릭한다. Run을 클릭하고 난 후 상단의 Textual Output을 클릭하면 DNA sequence assembly가 잘 수행되었는지에 대한 문구를 확인할 수 있다. Sequence assembly가 잘 수행되었다면, input 파일이 있던 디렉터리에 output 파일이 생성되어 있을 것이다.

Gap4 프로그램을 이용한 assembly 결과 확인

Gap4 프로그램을 실행하여 File 탭에서 Open을 눌러 Pregap4 프로그램에서 생성된 파일을 열어준다. 그러면 database의 정보와 함께 Contig Selector탭이 열리게 되는데, Figure 5와 같이 Contig Selector 탭의 굵은 막대를 우클릭하고 Edit contig를 누르게 되면, Contig Editor탭이 열리게 된다. Contig Editor 탭의 Next Search를 클릭하면 Search 탭이 나오는데, 이 탭에서 discrepancies를 클릭한 후 Search를 눌러주면, Figure 5의 빨간색 상자 안과 같이 개체 내에서의 또는 개체 사이에서의 allele이 차이를 보이는 SNP의 위치로 이동한다. Figure 5처럼 strand를 보기 좋기 정렬하려면, Contig Editor 탭의 Settings를 누르고 Group Readings를 눌러 자신이 원하는 정렬 기준을 활성화하면 된다.

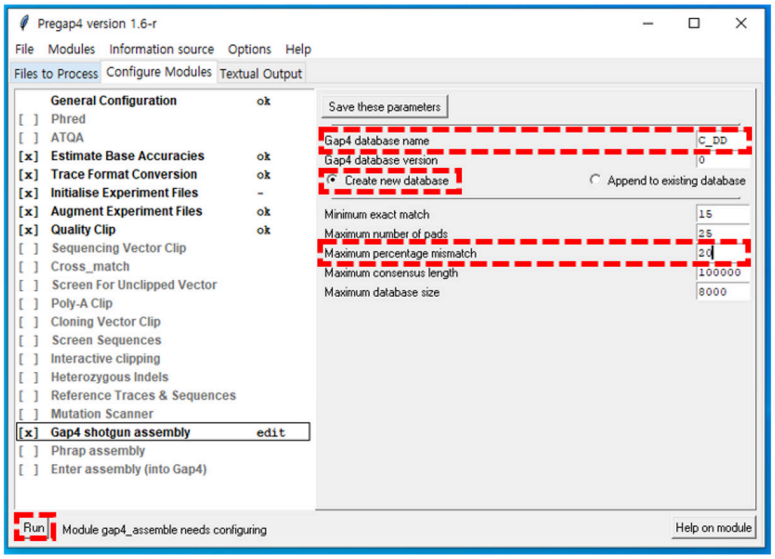

Figure 4. Configuration panel of Gap4 shotgun assembly. The figure above is the configuration panel of pregap4's gap4 shotgun assembly module. Enter the name of the output file in gap4 database name. For the maximum percentage mismatch, readings that do not exceed the maximum percentage mismatch are entered when comparing each reading in gap4. If you click "Run" at the bottom left, the assembly is executed.

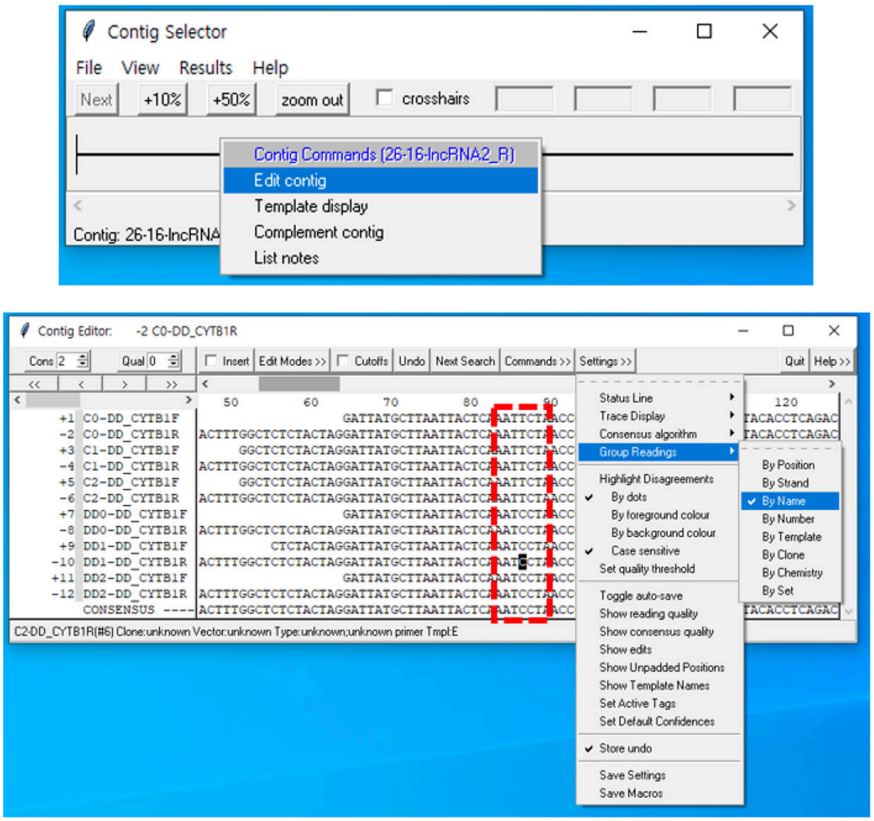

Figure 5. Displays of Contig Selector and Contig Editor. The figure above is the displays of gap4's contig selector and contig editor. If you right-click the thick bar in the contig selector and click edit contig, you can open the contig editor tab. If you click next search at the top of the contig editor tab, the search tab appears, and you can click discrepancies on the search tab. Press search to move to the SNP position that shows allele differences within or between individuals. You can change the sorting criteria by clicking group readings in settings at the top.

SNP의 allele을 텍스트가 아닌 chromatogram으로 보고자 할 때에는 Gap4의 Contig Editor 상에 있는 보고자 하는 개체의 SNP을 더블 클릭하면 된다. Figure 6과 같이 Contig Editor에서 여러 개의 strand를 더블 클릭하면 Trace display에서 한 번에 여러 개의 strand를 한 번에 볼 수 있으며, Contig Editor에서 다른 위치의 SNP을 클릭하면 Trace display의 chromatogram도 해당 위치로 이동한다. 또한 Trace display의 상단에 위치한 Show confidence를 활성화하여 allele별 confidence의 정도를 확인할 수 있다.

Figure 6. Trace display. The figure above is a trace display. Double-click the SNP of the object individual in the gap4 contig editor to check the sequence chromatogram of the corresponding strand of the object in the trace display tab.

Gap4를 이용한 assembled sequence의 FASTA file로의 저장

개체별 assembled sequence를 FASTA 파일로 저장하고자 한다면, 우선 Pregap4를 이용해 한 개체씩 따로 assembly를 수행해야 한다. 개체별 sequence assembly가 모두 완료되었다면, Gap4 프로그램을 열고 상단의 File을 누른 후 Open을 눌러 assembled sequence를 열어준다. 해당 개체의 assembled sequence를 FASTA 파일로 저장하는 방법에 대해 설명하겠다. Figure 7과 같이 File을 누른 후 Save consensus의 normal을 누르면, Save consensus: normal 탭이 열리게 된다. 이 탭 하단의 Output file 칸에 아웃풋 파일명을 입력해준 후 OK를 누르면, assembled sequence 파일이 있던 폴더에 FASTA 파일이 생성된다. 이후 다른 개체들에 대해서도 같은 작업을 반복해준 후에 생성된 FASTA 파일을 하나로 취합해준다.

Gap4의 Contig Editor에서 sequence를 눌러 원하는 알파벳을 입력하면 Figure 8과 같이 해당 sequence의 allele 수정이 가능하다. 수정한 후 Contig Editor의 Quit를 누르면 수정된 내용이 저장되며, 이후 Save consensus: normal을 이용해 assembly 및 수정된 sequence를 FASTA 파일로 저장할 수 있다.

요약

1. STADEN 패키지는 라이선스 비용을 들이지 않고 Sanger sequencing 결과의 분석 및 편집이 가능한 매우 유용한 프로그램으로 https://staden.sourceforge.net/에서 다운로드 할 수 있다.

2. STADEN 패키지의 Pregap4 프로그램을 이용하여 사용자의 설정에 맞게 DNA sequence assembly를 수행할 수 있으며, Gap4프로그램을 이용하여 DNA sequence assembly의 결과를 분석, 편집하여 SNP를 검출할 수 있다.

색인: DNA 염기서열, Sanger 시퀀싱, STADEN package, Pregap4, Gap4, SNP

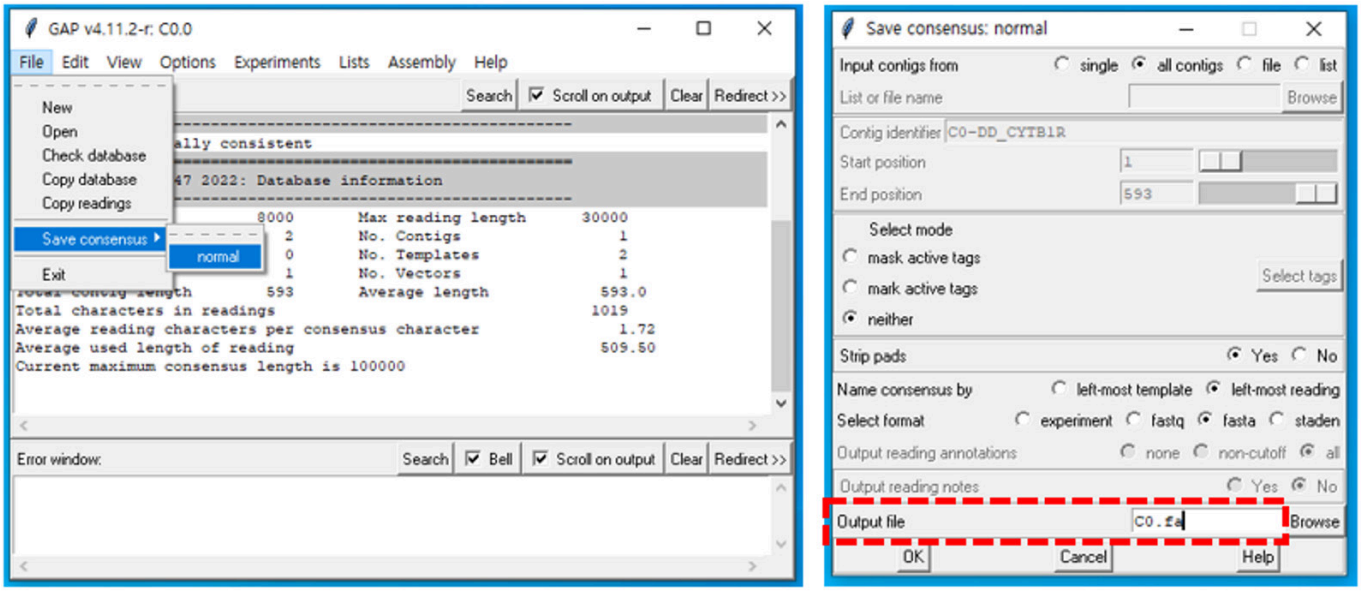

Figure 7. Displays of save consensus: normal. The figure above is the display of gap4's save consensus: normal. First find file banner located in the upper left of gap4, and click save consensus and normal. Then the save consensus: normal tab appears as shown in the right figure. Write the name of the output file in the output file (red box) of the tab and click OK at the bottom to save the data of gap4.

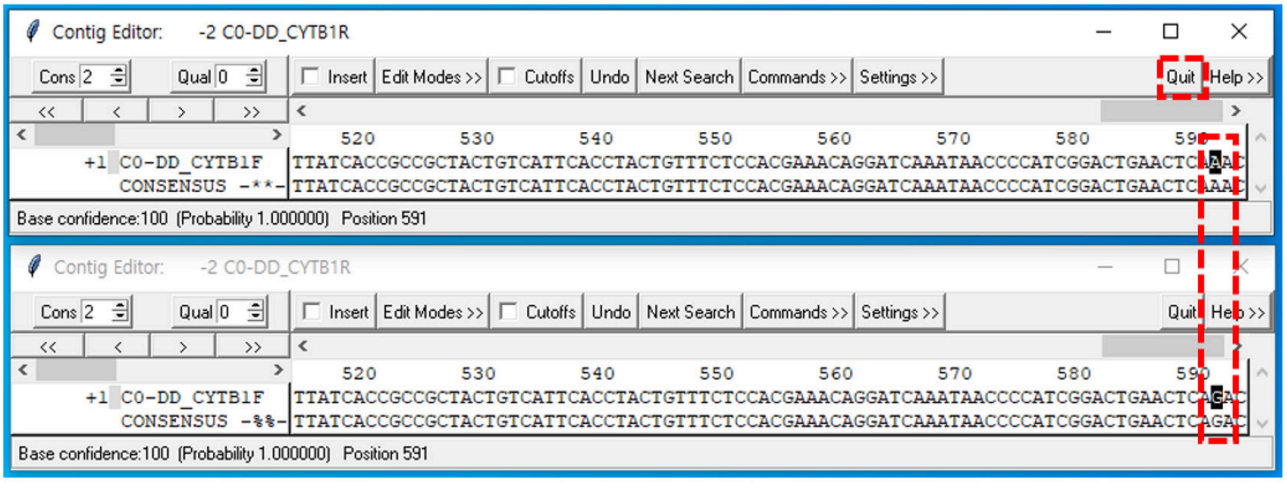

Figure 8. Edit of allele with Gap4. Allele Editing using the Gap4 program is possible in the Contig Editor tab. After clicking the allele you want to edit in Contig Editor, you can edit the allele by entering the desired alphabet. After editing Allele, click Quit on the top right to save the modified content.

References

Anderson, S.; Bankier, A.T.; Barrell, B.G.; De Bruijn, M.H.; Coulson, A.R.; Drouin, J.; Eperon, I.C.; Nierlich, D.P.; Roe, B.A.; Sanger, F.; Schreier, P.H.; Smith, A.J.; Staden, R.; Young, I.G. (1981), "Sequence and organization of the human mitochondrial genome", Nature, 290 (5806): 457–465.

Ewing B, Hillier L, Wendl MC, Green P. 1998. BaseCalling of automated sequencer traces using phred. I. Accuracy Assesment. Genome Res 8: 175-185.

Huang X. 1997. On global sequence alignment. Bioinformatics. 10:227-235.

Kim G. W., Joe S. D., Kim H. Y. and Park H. B. 2020. Phylogenetic Analysis of Carassius auratus and C. Cuvieri in Lake Yedang Based on Variations of Mitochondrial CYTB Gene Sequences. Journal of Life Science. 30:1063-1069.

Lander E.S. et al. 2001. Initial sequencing and analysis of the human genome. Nature 409.6822: 860-921.

Maxam A. M. and Gilbert W. 1977. A new method for sequencing DNA. Proceedings of the National Academy of Sciences. 560.

Myers, E. W. and Miller, W. 1988. Optimal alignments in linear space. Bioinformatics, 4(1), 11-17.

Sanger F. and Coulson A. R. 1975. A rapid method for determining sequences in DNA by primed synthesis with DNA polymerase. J. Mol. Biol. 94:441-448.

Sanger F., Coulson A. R., Friedmann T., Air G. M., Barrell B. G., Brown N. L., Fiddes J. C., Hutchison C. A., Slocombe P. M. and Smith M. 1978. The nucleotide sequence of bacteriophage ϕX174. J. Mol. Biol. 125:225-246.

Sanger, F.; Coulson, A.R.; Hong, G.F.; Hill, D.F.; Petersen, G.B. (1982), "Nucleotide sequence of bacteriophage λ DNA", Journal of Molecular Biology, 162 (4): 729–773.

Smith, L. M., Sanders, J. Z., Kaiser, R. J., Hughes, P., Dodd, C., Connell, C. R., and Hood, L. E. (1986). Fluorescence detection in automated DNA sequence analysis. Nature, 321(6071), 674-679.

Staden R. 1977. Sequence data handling by computer. Oxford University Press. 4037.

Staden R. 1982. Automation of the computer handling of gel reading data produced by the shotgun method of DNA sequencing. Nucleic Acids Res. 10:4731-4751.