



Institut-Hôpital
neurologique de Montréal

Montreal Neurological
Institute-Hospital



EDDU Protocols

Sequencing Analysis Guide

Authors: Michael Nicouleau and Thomas M. Durcan

Version 2.0

EDDU-014-01

May 2020

Sequencing Analysis Guide

Author(s): Michael Nicouleau and Thomas M. Durcan

Version	Authors/Updated by	Date	Signature
V1.0	Michael Nicouleau Thomas Durcan	2020-04-30	

The involved functions approve the document for its intended use:

Name	Function	Role	Date	Signature
Thomas Durcan	Associate Director	Associate Director, MNI Early Drug Discovery Unit (EDDU)		

Table of Contents

1	Introduction	1
1.1	Objectives	1
1.2	Protocol overview	1
1.3	Abbreviation list	1
2	Materials	2
2.1	Equipment	2
2.2	Websites	2
2.3	Software	2
3	Protocol.....	3
3.1	Data	3
3.1.1	View result on SeqStudio Genetic Analyser instrument	3
3.1.2	Export data SeqStudio Genetic Analyser instrument	3
3.1.3	Analyse data.....	4
3.2	SnapGene Software	4
3.2.1	Open .ab1 file	4
3.2.2	Alignment to a reference DNA sequence	7
3.3	Variant detection.....	9
3.3.1	Single base-pair substitution	9
3.3.2	Indel.....	10
3.4	Interpretation of sequencing data	11
3.4.1	Determine homology sequences	11
3.4.2	Translation and Open Reading Frame	12
3.4.3	Protein parameters	16
3.5	Nomenclature	16
3.5.1	Reference genome	17
3.5.2	Type of reference sequence	18
3.5.3	Code used to describe variants	18
3.5.4	Type of variation	18
3.6	Troubleshooting	20
3.6.1	Failed sequence	20
3.6.2	Weak sequence.....	21

3.6.3	Poor start gradually drops off	21
3.6.4	Multiple peaks	22
3.6.5	Truncated sequence.....	22
3.6.6	Repetitive region	23
3.6.7	GC-Rich region.....	24
3.6.8	Spikes.....	25

1 Introduction

1.1 Objectives

Sanger sequencing is a method of DNA sequencing that is based on the selective incorporation of chain-terminating dideoxynucleotides by DNA polymerase during in vitro DNA replication.

Sanger sequencing analysis software packages are used for base calling (process of assigning nucleobases to chromatogram peaks), sequence alignment, trace visualization and variant detection.

In this protocol, we describe our method used for analyzing sequence data obtained with our Applied Biosystems SeqStudio Genetic Analyzer instrument and analyzed using SnapGene software..

1.2 Protocol overview

Sequencing analysis programs are used to open. ab1 files. They provide a view of the chromatogram, the raw data and a text file data. Data can be edited and aligned to a DNA reference sequence in the aim to highlight variations in the imported sequence. DNA sequence will be translated to a protein sequence that can be analysed.

1.3 Abbreviation list

Abbreviations	Full Name
QC	Quality Check
DNA	Deoxyribonucleic Acid
cDNA	Coding DNA
RNA	Ribonucleic Acid
WT	Wild Type
bp	Base Pair
pI	Isoelectric point

2 Materials

2.1 Equipment

- Minimal computer configuration:
 - MacOS 10.8 or higher
 - Windows 7 or higher
 - Debian/Ubuntu
 - Fedora/openSUSE
- Applied Biosystems SeqStudio™ Genetic Analyzer System– ThermoFisher Scientific

2.2 Websites

- <http://www.ncbi.nlm.nih.gov/BLAST/>
- <https://www.snapgene.com/support/tutorial-videos/>
- <https://www.ncbi.nlm.nih.gov/refseq/>
- <https://www.ncbi.nlm.nih.gov/grc>
- <https://genome.ucsc.edu/index.html>
- <https://useast.ensembl.org/index.html>
- <https://web.expasy.org/translate/>
- <https://web.expasy.org/protparam/>

2.3 Software

- SnapGene software® - GSL Biotech LLC

3 Protocol

3.1 Data

Sanger Sequencing data is obtained following the protocol: [EDDU-005-01_SeqStudio DNA Sequencing](#),

3.1.1 View result on SeqStudio Genetic Analyser instrument

View results for the plate when the run is completed (all injections are finished).

Select Results to view the run results

Select List view. Each injection group displays a QC color for each capillary:

- Quality Check provides a summary of results based on the quality parameter settings and automatically flags lower-quality traces for further inspection
- All QC tests passed.
- At least 1 warning quality alert was triggered.
- At least 1 failing quality alert was triggered.

3.1.2 Export data SeqStudio Genetic Analyser instrument

1. In the instrument **Home** screen, select **Settings > Run history**.
 2. Select one or more plates from the Run History table.
 3. Select Export.
 4. Select a storage location (USB drive).
 5. Select Export.
- Sequencing experiments use base calling (the algorithms and settings required to assign nucleobases to chromatogram peaks) to determine the fragment base sequence. For each sample, an .ab1 file containing an electropherogram and the DNA base sequence is generated.
 - A sequencing file (.ab1) and plate quality check (QC) report (.CSV and .PDF) are automatically exported to the folder SeqStudio_Data > Data > Plate name on the desktop.

3.1.3 Analyse data

For each sample, open the .ab1 file in the sequencing analysis software in order to analyze the data, display sequencing electropherogram, edit, save, and print sample files.

- Non-exhaustive list of Sanger Sequencing data analysis software available:
 - SnapGene Software - GSL Biotech (Windows / MAC / Linux)
 - Sequencing Analysis Software - Applied Biosystems (Windows)
 - SeqScape Software - Applied Biosystems (Windows)
 - Variant Reporter Software – Applied Biosystems (Windows)
 - Chromaseq - Maddison, D.R., & W.P. Maddison. (Windows / MAC / Linux)
 - Genome Compiler – Genome Compiler (Windows / MAC)

3.2 SnapGene Software

3.2.1 Open .ab1 file

In order to view the raw data, the chromatogram trace and chromatogram data of .ab1 files (**Figure 2**),

- Open SnapGene software (**Figure 1**)
- Click Open to select the .ab1 file

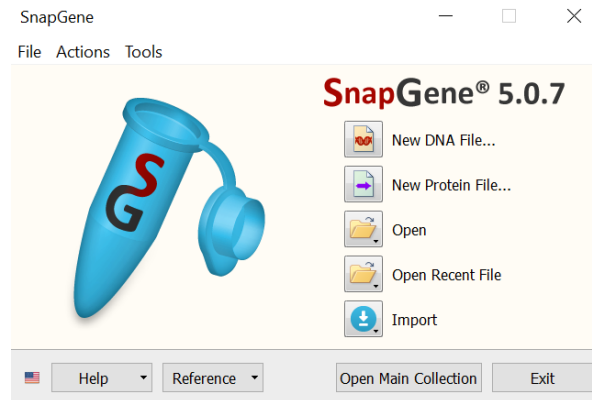


Figure 1. Start up window of SnapGene Software.

- Sequencing files can be opened directly by clicking on the file icon if SnapGene is set up as the reference software.



Figure 2. Autoscaled electropherogram sequence view on SnapGene software.

- Nucleic Acid notation:
 - Guanine (G) – black
 - Thymine (T) – red
 - Cytosine (C) – bleu
 - Adenine(A) – green
- When the base calling is not able to assign a specific nucleotide to a peak, ambiguity code is used (**Table 1**). Ambiguous nucleotide can be due to a variation or a poor quality sequence data.

Table 1. Nucleotide ambiguity code.

Code	Represents
Y	Pyrimidine (C or T)
R	Purine (A or G)
W	Weak (A or T)
S	Strong (G or C)
K	Keto (T or G)
M	Amino (C or A)
D	A, G, T (not C)
V	A, C, G (not T)
H	A, C, T (not G)
B	C, G, T (not A)
X/N	any base
-	Gap

- Complementarity of DNA strands in a double helix make it possible to use one strand as a template to construct the other following the complementary code (**Table 2**). The arrows on top right can be used to reverse the sequence (forward or reverse strand).

Table 2. Complementarity nucleotide code.

Code	Complement
A	T
G	C
C	G
T	A
Y	R
R	Y
W	W
S	S
K	M
M	K
D	H
V	B
H	D
B	V
X/N	X/N
-	-

- To ensure the quality of the chromatogram; peaks should be individual, sharp and evenly spaced (**Figure 2**). That does not concern the first 20-30 bases of DNA sequence where peaks are usually unresolved and small.
- Once the quality of the sequence data is determined to be satisfactory, the sequence may need to be edited. Base miscalls by the analysis software are common and should be expected. Occasionally, the computer could call an 'N' when a human would be confident in making a more specific basecall (**Figure 3**). Edit nucleotides as needed by clicking on bases.

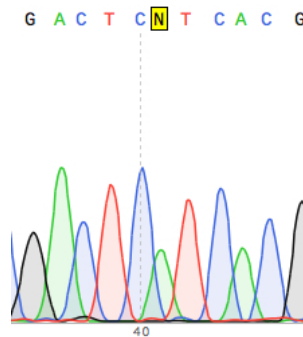


Figure 3. Example of miscall base in chromatogram

- **IMPORTANT:** Any modification of sequence may lead to an inaccurate alignment and an incorrect sequence data. Edit sequence to fix minor misanalyses only.
- Text sequences for the DNA sequence can be obtained by **selecting** chromatogram data (**Figure 3**).

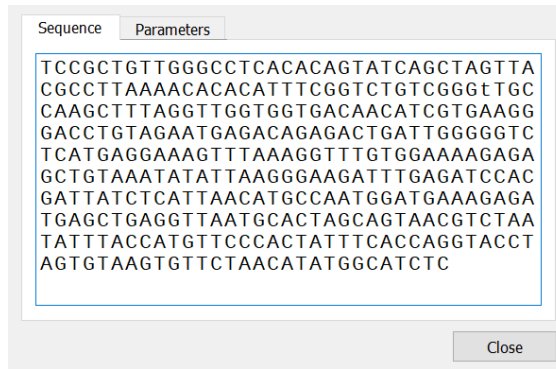


Figure 4. Chromatogram text data on SnapGene.

3.2.2 Alignment to a reference DNA sequence

Reference DNA sequence is a sequence file that is used as a reference to describe variants that are present in a sequence analysed.

Reference sequences must come from data sources that provide stable and permanent identifiers, like;

- NCBI at <https://www.ncbi.nlm.nih.gov/refseq/>
- UCSC at <https://genome.ucsc.edu/index.html>
- Ensembl at <https://useast.ensembl.org/index.html>

Reference Sequence records can be retrieved by querying with an accession number, symbol or locus, or name.

If the sequence is unknown, homology sequence can be determined using BLAST (**session 3.4.1**).

Align DNA sequences with a reference sequence;

- To verify a cloning or a mutagenesis
- To align a cDNA to genome
- To check indel or nucleotide variant

A SnapGene tutorial video is available at <https://www.snapgene.com/support/tutorial-videos/>

In order to align sequences in SnapGene;

- Open your reference sequence;
- Select New DNA file (**Figure 1**)
- Paste the text data sequence (e.g. FASTA format) and click Ok.

- In the main menu, select Tools > Align Multiple Sequences (**Figure 5**).
- In the new window, import the required files and go to the sequence view to see and edit the original sequence and alignment files.

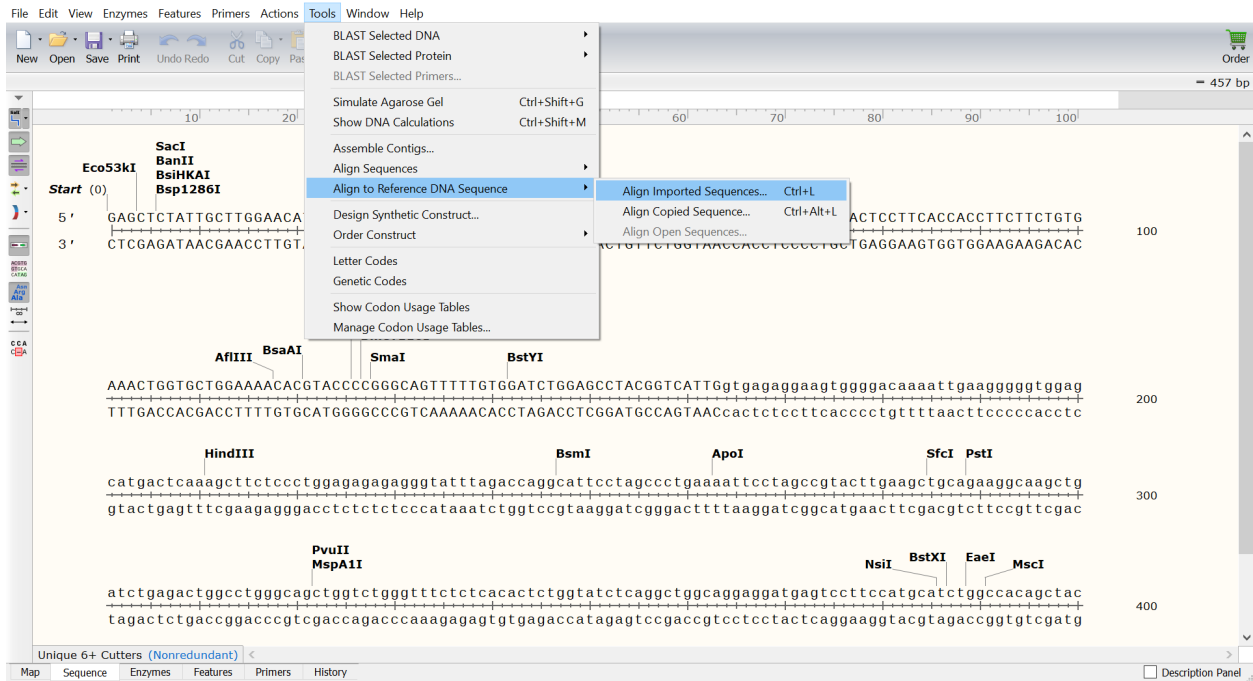


Figure 5. Launching alignment from the main menu on SnapGene.

- Click the disclosure triangle to expand the display and show sequence trace (**Figure 6**)
- Verify the quality of the chromatogram and edit nucleotides if needed by clicking on bases.
- **IMPORTANT:** Any modification of sequence may result in an inaccurate alignment and an incorrect sequence data. Edit sequence to fix minor misanalyses only.



Figure 6. Alignment sequences to a Reference DNA display on SnapGene.

3.3 Variant detection

3.3.1 Single base-pair substitution

These are also known as single nucleotide variants (SNV). This is a substitution of a single nucleotide for another. It can be any nucleic acid change (**Figure 7**).

If a SNV occurs in a protein coding region, this could result in;

- Synonymous change: A nucleotide substitution that does not result in a change in amino acid.
- Non-synonymous change (or a missense variant): A nucleotide substitution leads to an amino acid change.
- Stop gain change (or nonsense variant): A nucleotide substitution results in a stop codon and consequently premature truncation of the protein.

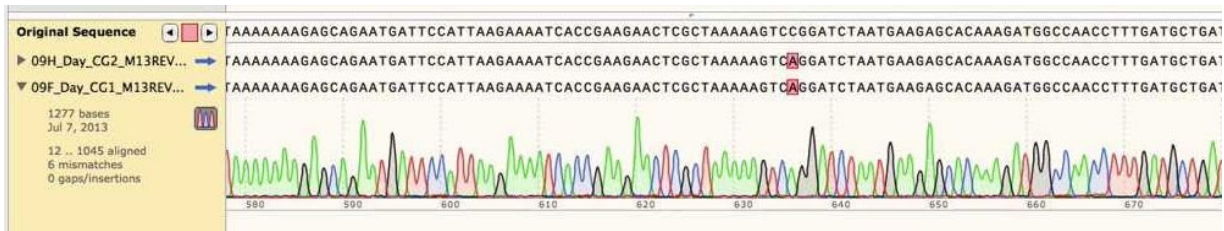


Figure 7. Example of homozygous single nucleotide substitution C>A.

- Zygosity is the degree to which both copies of a chromosome or gene have the same genetic sequence.
- Homozygous and heterozygous are used to describe the genotype of a diploid organism at a single locus on the DNA (**Figure 8**)
 - Homozygous describes a genotype consisting of two identical alleles at a given locus
 - Heterozygous describes a genotype consisting of two different alleles at a locus

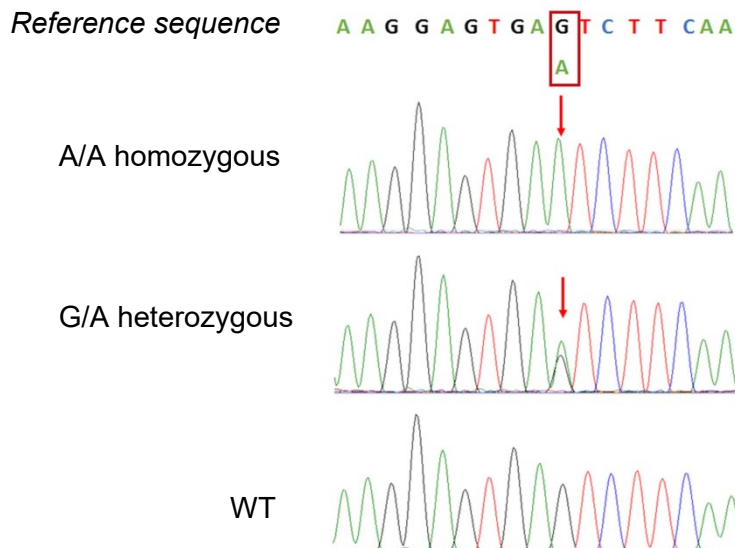


Figure 8. Example of heterozygous vs. homozygous single base pair variation in diploid organism.

3.3.2 Indel

Insertion or deletion of a single stretch of DNA sequence can range from one to 100s of base-pairs in length.

If the length of an indel is not a multiple of three, then it will likely result in a frameshift. A frameshift changes the reading frame of all the remaining bases resulting in the rest of the gene being incorrectly translated.

- Insertion

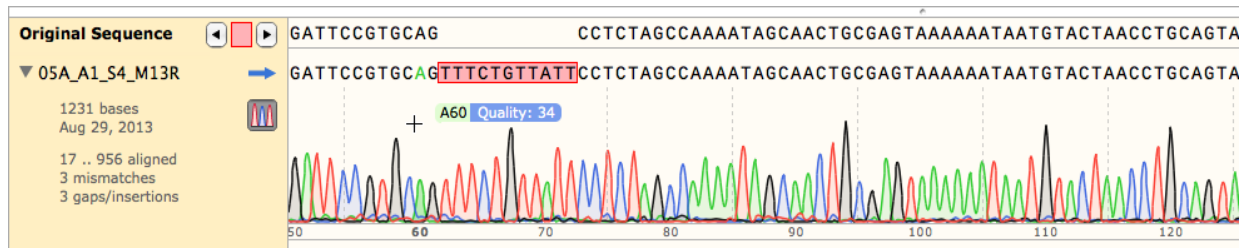


Figure 9. Example of homozygous insertion of 11 bp.

- Deletion

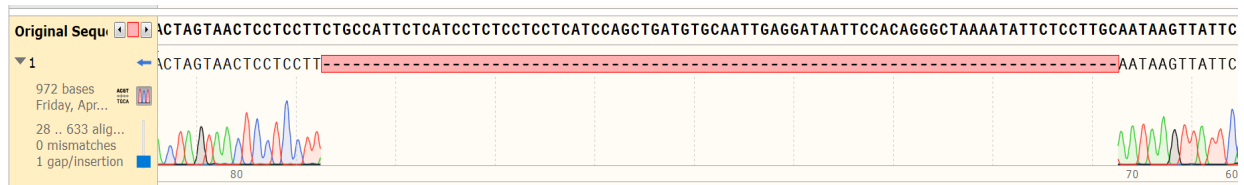


Figure 10. Example of homozygous deletion of 80 bp.

3.4 Interpretation of sequencing data

3.4.1 Determine homology sequences

Sequence similarity searching to identify homologous sequences is the first steps in any analysis of newly determined sequences.

BLAST, (Basic Local Alignment Search Tool, supported by the National Center for Biotechnology Information (NCBI) is a program made to compare nucleotide or protein sequences to sequence databases and calculates the statistical significance of matches (**Figure 11**).

This program is accessible at: <http://www.ncbi.nlm.nih.gov/BLAST/>

Standard Nucleotide BLAST

blastn blastp blastx tblastn tblastx BLASTN programs search nucleotide databases using a nucleotide query. [more...](#) [Reset page](#) [Bookmark](#)

Enter Query Sequence

Enter accession number(s), gi(s), or FASTA sequence(s) Clear Query subrange

```

atggatgattcatgaaggacttcaaaagccaaaggagggtgtgctgctgagaaaaccaaagg
gtgtgcagagcagcaggaaagcaaaagggttctctatgtaggctccaaacaaaggaggagtggt
gcotgtgtgcaacagtgctgagaaagcaaaagcaagtgcaaaagtggagagcagtgtaacggt
gtgacagcagtagcccaagacagtgaggagcaggagcattgacagccactgcttgcaaaagg
accagtggcaagaatgaagaggccccacaggagaattctgaaagatagctgtggtctgacaa

```

From

To

Or, upload file No file chosen

Job Title

Enter a descriptive title for your BLAST search

Align two or more sequences

Choose Search Set

Database Standard databases (nr etc.): rRNA/ITS databases Genomic + transcript databases Betacoronavirus

Nucleotide collection (nr/nt)

Organism Optional

Enter organism name or id—completions will be suggested

Enter organism common name, binomial, or tax id. Only 20 top taxa will be shown

Models (XM/XP) Uncultured/environmental sample sequences

Exclude Optional

Sequences from type material

Limit to Optional

Entrez Query Optional

Enter an Entrez query to limit search [Create custom database](#)

Program Selection

Optimize for

Highly similar sequences (megablast)

More dissimilar sequences (discontiguous megablast)

Somewhat similar sequences (blastn)

Choose a BLAST algorithm

BLAST

Search database Nucleotide collection (nr/nt) using Megablast (Optimize for highly similar sequences)

Show results in a new window

BLAST results will be displayed in a new format by default

You can always switch back to the Traditional Results page.

Figure 11. BLAST screen display.

3.4.2 Translation and Open Reading Frame

The DNA is first transcribed into an RNA sequence where Uracil (U) is used instead of Thymine (T). Translation of the RNA sequence can help to determine the sequence of amino acids that will appear in the final protein based on the nucleic acid sequence (**Figure 12**).

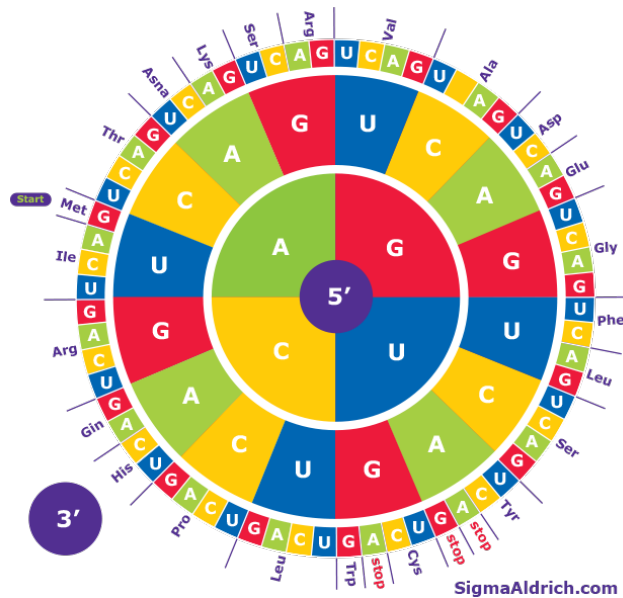


Figure 12. Amino Acid Codon Wheel Translation.

In translation, codons of three nucleotides determine which amino acid will be added next in the protein chain. Every region of DNA has six possible reading frames, three in each direction (Figure 13)

The program used to translate the sequencing data is accessible at: <https://web.expasy.org/translate/>

- A RNA sequence or directly a DNA sequence can be used
- Different output format are available; one and three letter code (Table 3), including nucleotide sequence or not.

Table 3. Amino acid code

Amino acid	Three letter code	One letter code
alanine	ala	A
arginine	arg	R
asparagine	asn	N
aspartic acid	asp	D
asparagine or aspartic acid	asx	B
cysteine	cys	C
glutamic acid	glu	E
glutamine	gln	Q
glutamine or glutamic acid	glx	Z
glycine	gly	G
histidine	his	H
isoleucine	ile	I
leucine	leu	L
lysine	lys	K
methionine	met	M

Amino acid	Three letter code	One letter code
phenylalanine	phe	F
proline	pro	P
serine	ser	S
threonine	thr	T
tryptophan	trp	W
tyrosine	tyr	Y
valine	val	V

- Determine the correct open reading frame (ORF).
 - The reading frame used determines which amino acids are to be encoded by a gene
 - One ORF is used in translating a gene and this is often the longest ORF
 - An ORF starts with an ATG (Met) in most species and ends with a stop codon (TAA, TAG or TGA).

Translate is a tool which allows the translation of a nucleotide (DNA/RNA) sequence to a protein sequence.

DNA or RNA sequence

```
atggatgtattcatgaaaggactttcaaaggccaaggaggagtttgctgctgctgagaaaacaaacaggg  
tgtggcagaagcagcaggaagacaaaagagggttctctatgtaggctccaaacaaaggaggagtggtgc  
atgggtggcaacagtggctgagaagacaaagagcaagtgacaaaatgtggaggagcagtggtgacgggtg  
acagcagtagccagaagacagtggaggagcaggagcattgcagcagccactggctttgtcaaaaaggacca  
gttgggcaagaatgaagaaggagcccccacaggaaggaattctggaagatatgcctgtggatcctgacaatgagg  
cttatgaaatgccttctgaggaagggtatcaagactacgaacctgaagcctaa
```

Output format

- Verbose: Met, Stop, spaces between residues
- Compact: M, -, no spaces
- Includes nucleotide sequence
- Includes nucleotide sequence, no spaces

DNA strands

- forward
- reverse

Genetic codes - [See NCBI's genetic codes](#)

Standard

reset

TRANSLATE!

Results of translation

- Open reading frames are highlighted in red
- Select your initiator on one of the following frames to retrieve your amino acid sequence

Download all the translated frames

5'3' Frame 1

MDVFMKGLSKAKEGVVAAAEEKTKQGVAAEAGKTEGVLYVGSKTKEGVVHVGVATVAEKTKEQVTNVGGAVVTGVTAVAQKTVEGAGSIAAATGFVKKDD
LGKNEEGAPQEGILEDMVDPDNEAYEMPSEEGYQDYEPEA-

5'3' Frame 2

WMYS-KDFQRPRRELWLLLRKPNRVWQKQERQKRVSFMA-APKPRREWCMVWQQWLRPFSK-OMLEEQW-RV-QQ-PRRQWREQALQQPLALSKRTS
WARMKKEPHRKEFWKICLWILTMRLMKCLLRKGIKRTNLKP

5'3' Frame 3

GCIHERTFKGGQSGGCC-ENQTGCGRSSRKDKRGCSLCLRLQNGGSGAWCGNSG-EDQRASDKWRSSGDGCDSSSPEDSGSREHCSHWLQKGPV
GQE-RRSPTGRNSGRYACGS-Q-GL-NAF-GRVSRRLRT-SL

3'5' Frame 1

LGFRFVVLIPFLRRHFISLIVRIHRHIFQNSFLWGSFFILAQLVLFDKASGCCNAPCSLHCLLGYCCHTRHHCSSNICHLFLGLLSHCCHTMHHSLLGF
GAYIENTLFLSCLCFHTLFGFLSSSHNSLLGL-KSFHEYIH

3'5' Frame 2

-ASGS-S-YPSSEGIS-ASLSGSTGISSRIPSCGAPSSFLPNWSFLTQPVAAAMLPAFSTVFWATAVTPVTTAPPTFVTCSLVFSATVATPCTTSLVL
EPT-RTPSFVFPAAATPCLVFSAAATPPLAFESPFMNTS

3'5' Frame 3

RLQVRS�DTPQAFHKPHCQDPQAYLPEFLPVGLLLHSCPTGPF-QSQWLLQCSLLPPLSSGLLSHPSPLLLQHLSLALWSSQPLPHHAPLPPWF
SLHREHPLLSFLLLPVWVFSQQQPQLPPWPLKVLV-IHP

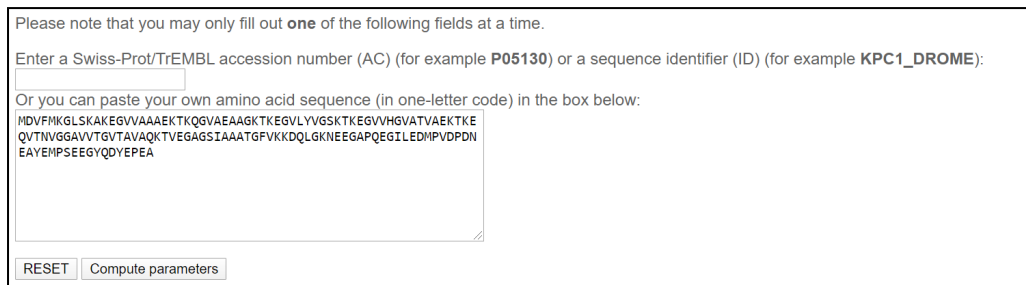
Figure 13. Example of DNA sequence translation with six possible reading frames on Expsy/Translate

3.4.3 Protein parameters

ProtParam, accessible at <https://web.expasy.org/protparam/>, is a tool which allows the computation of various physical and chemical parameters of a protein (**Figure 14**).

The computed parameters include;

- molecular weight
 - theoretical pI
 - amino acid composition
 - atomic composition
 - extinction coefficient
 - estimated half-life
 - aliphatic index
 - grand average of hydropathicity
 - instability index
- A protein whose instability index is smaller than 40 is predicted as stable, a value above 40 predicts that the protein may be unstable.
 - One letter code is used to enter the amino acid sequence



Please note that you may only fill out **one** of the following fields at a time.

Enter a Swiss-Prot/TrEMBL accession number (AC) (for example **P05130**) or a sequence identifier (ID) (for example **KPC1_DROME**):

Or you can paste your own amino acid sequence (in one-letter code) in the box below:

```
MDVFMKGLSKAKEGVAAAEEKTKQGVAAAGKTEGVLVGSKTEGVVHGATVAEKTKE
QVTNIVGGAVTGVTAQAQKTVGAGSIAAATGFVKDQLGKNEEGAPQEGILEDMPVDPDN
EAYEMPSEEGYQDYEP EA
```

Figure 14. ProtParam screen display

3.5 Nomenclature

Sequence variations are described following a certain nomenclature.

For a more detailed description of the variant nomenclature, please, refer to: den Dunnen JT1, Antonarakis SE. Nomenclature for the description of human sequence variations. Hum Genet. 2001 Jul;109(1):121-4. [DOI: 10.1007/s004390100505](https://doi.org/10.1007/s004390100505)

3.5.1 Reference genome

A reference genome (or reference genome assembly) is a digital nucleic acid sequence database, assembled by scientists as a representative example of a species set of genes. The Genome Reference Consortium (GRC) is an international collective of academic and research institutes with expertise in genome mapping, sequencing, and informatics, formed to improve the representation of reference genomes. <https://www.ncbi.nlm.nih.gov/grc>

- The human reference genome GRCh38 was released from the GRC on 17 December 2013. GRCh38 is a corrected and improved version of GRCh37. The newer assembly should be use.
- The existence of different builds, published by different groups, with different naming conventions, with some differences in sequences, has caused much confusion and plentiful errors over the years. Part of the problem is that many bioinformatic tools fail to enforce consistent use of a specific reference. This allows the unwary user to switch reference genomes halfway through a project without realizing that their comparisons suddenly become worthless.
- **IMPORTANT:** The nucleotide numbering changes between different reference genome. This is very important to use the right reference genome to avoid any confusion in the description of variant.
- Recent human genome assemblies

Table 4. Human reference genome.

Release name	Date of release
GRCh38 or hg38	Dec 2013
GRCh37 or hg19	Feb 2009
NCBI Build 36.1 or hg18	Mar 2006
NCBI Build 35 or hg17	May 2004
NCBI Build 34 or hg16	Jul 2003

- Recent mouse genome assemblies

Table 5. Mouse reference genome.

Release name	Date of release
GRCm38 or mm10	Dec 2011
NCBI Build 37 or mm9	Jul 2007
NCBI Build 36 or mm8	Feb 2006
NCBI Build 35 or mm7	Aug 2005
NCBI Build 34 or mm6	Mar 2005

3.5.2 Type of reference sequence

- To avoid confusion in the description of a variant it should be preceded by a letter indicating the type of reference sequence used (**Table 6**). The reference sequence used must contain the residue described to be changed.

Table 6. Type of reference sequence.

Full name	abbreviation	
DNA	Coding DNA	c.
	Genomic DNA	g.
	Mitochondrial DNA	m.
RNA		r.
Protein		p.

3.5.3 Code used to describe variants

Different types of variation are described using the following nomenclature (**Table 7**).

Table 7. Code used to describe variants.

Full name	abbreviation
substitution (for bases)	>
range	-
more change in one allele	;
more transcripts / mosaicism	,
allele	[]
deletion	del
duplication	dup
insertion	ins
inversion	inv
conversion	con
extension	ext
stop codon	X
frame shift	fsX

3.5.4 Type of variation

A standard variant description has the format “prefix_position(s)_change” (**Table 8**).

For a clear distinction, descriptions at DNA, RNA and protein level are unique;

- DNA level
in capitals, starting with a number referring to the first nucleotide affected (c.76A>T or g.476A>T)

- RNA level
in lower-case, starting with a number referring to the first nucleotide affected (r.76a>u)
- Protein level
in capitals, starting with a letter referring to first the amino acid affected (p.Lys76Asn)

Table 8. Different types of described variations.

Substitution	
c.123A>G	on cDNA, A in 123 is replaced by G
p.P252R	on protein, proline (P) replaced by arginine (R)
Deletion	
c.546delT	deletion of T in 546
c.586_591del	for six bases deleted
p.F508del	deletion of phenylalanine (F) in 508
Duplication	
c.546dupT	duplication of T in 546
c.586_591dup	duplication of the segment 586 to 591
p.G4_Q6dup	duplication of the segment from glycine (G) in 4 to glutamine (Q) in 6
Insertion	
c.546_547insT	insertion of T between 546 and 547
c.1086_1087insGCGTGA	insertion of GCGTGA
p.K2_L3insQS	insertion of glutamine serine between lysine (K) in 2 and leucine (L) in 3
Inversion	
c.546_2031inv	segment 546 to 2031 inverted
Frameshift	
p.R83SfsX15	arginine (R) is the first amino acid changed, it is in position 83, it makes serine (S) instead, the length of the shift frame is 15, including the stop codon (X)

3.6 Troubleshooting

The following shows some examples of failures in sequencing and their possible cause. This is a non exhaustive list and other issues can also arise.

3.6.1 Failed sequence

- Chromatogram data looks messy or is mostly blank.
- Many 'N's in the sequence, if bases are not called at all.
- Raw data has signal intensity in the low hundreds.

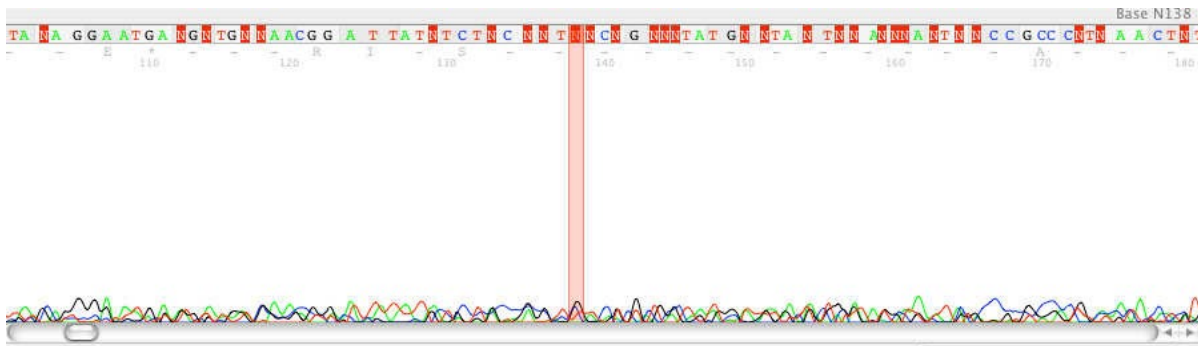


Figure 15. Failed sequence. Credit: thegenepool

Problem	Possible Cause	Recommended
Lack of sequence data	No priming site present	Make sure the primer site is present in the vector/PCR product you are using Redesign/ use a different primer
	Primers have degraded through freeze-thaw cycles Inefficient primer binding	Make up new primer stocks Redesign primer
	Insufficient amount of DNA template	Quantify DNA Increase the amount of DNA template
	DNA has degraded Inhibitory contaminant in your samples (eg. Salt, phenol, EDTA, ethanol)	Re-extract DNA Clean-up DNA template

3.6.2 Weak sequence



Figure 16. Weak sequence. Credit: thegenepool

Problem	Possible Cause	Recommended
Low peaks throughout	Insufficient amount of DNA template	Quantitate the DNA Increase the amount of DNA template
	Inhibitory contaminant in your samples (eg. Salts, phenol, EDTA, ethanol) Insufficient amount of primer Insufficient primer binding	Clean-up DNA template Check primer dilution Redesign primer

3.6.3 Poor start gradually drops off

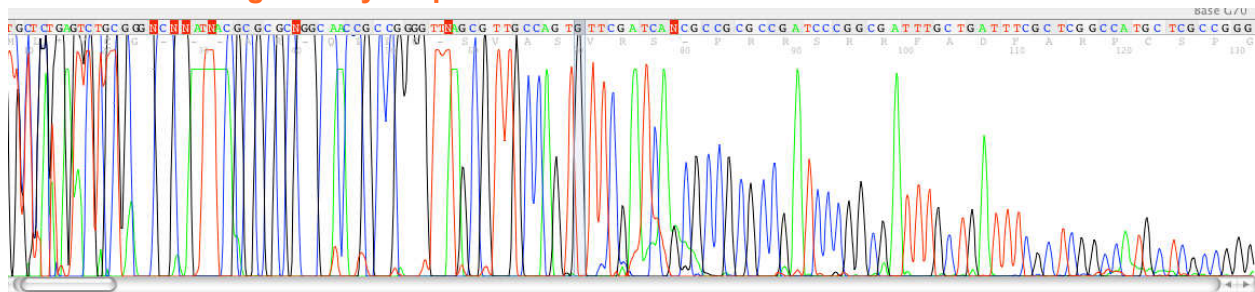


Figure 17. Poor start gradually drops off sequence. Credit: thegenepool

Problem	Possible Cause	Recommended
Poor sequence at the start followed by weak signal	Primer binding to itself	Redesign sequencing primer
	Other primers present	Check PCR clean-up has removed all other possible primers

3.6.4 Multiple peaks

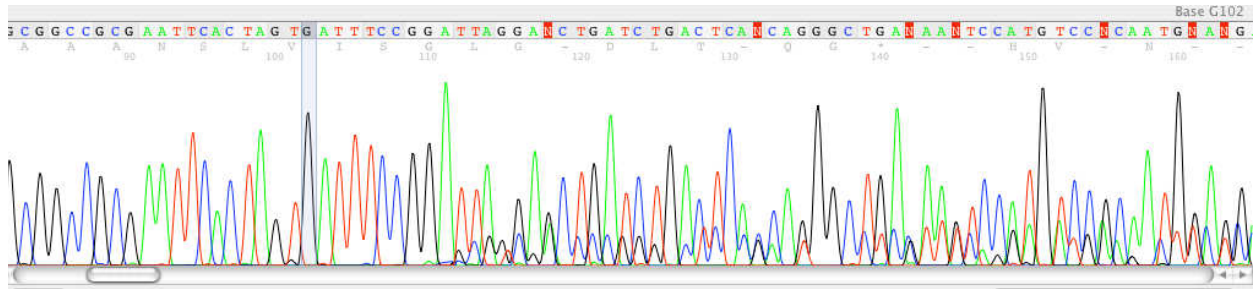


Figure 18. Multiple peaks sequence. Credit: thegenepool

Problem	Possible Cause	Recommended
Overlapping peaks in the sequence data	INDEL in PCR product	Sequence the complementary strand
	Multiple priming sites	Use a different primer
	Residual primers (PCR product has not been cleaned up)	Make sure all PCR primers and dNTPs have been removed
	Mixed plasmid/PCR product prep	Contaminated template. Clean sequence at the start with mixed peaks beginning at the cloning site Ensure single colonies are picked

3.6.5 Truncated sequence

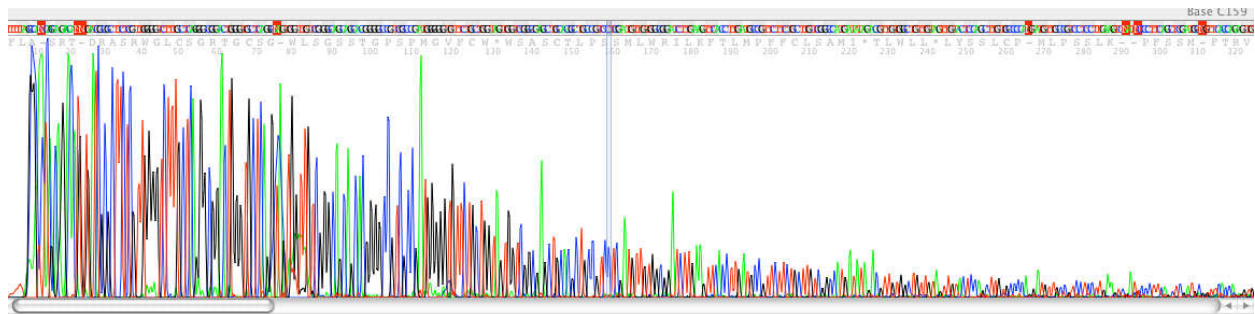


Figure 19. Truncated sequence. Credit: thegenepool

Problem	Possible Cause	Recommended
Sequence starts well but signal weakens gradually	Too much DNA template (overload of DNA lead to excessive number of short fragments)	Use less DNA template Carefully quantify your DNA template and primer

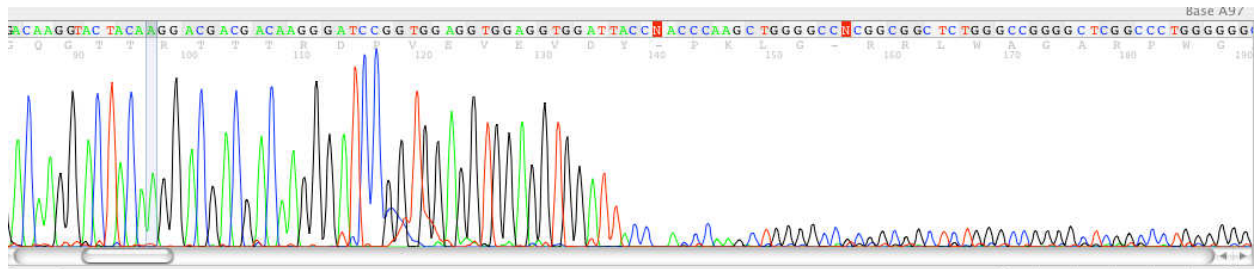


Figure 20. Truncated sequence. Credit: thegenepool

Problem	Possible Cause	Recommended
Sequence starts well but signal stops abruptly	Secondary structure (GC and AT rich templates can cause the DNA to loop and form hairpins)	Add 1 uL DMSO to the sequencing reaction to help relax the structure Design primers close to the hairpin

3.6.6 Repetitive region

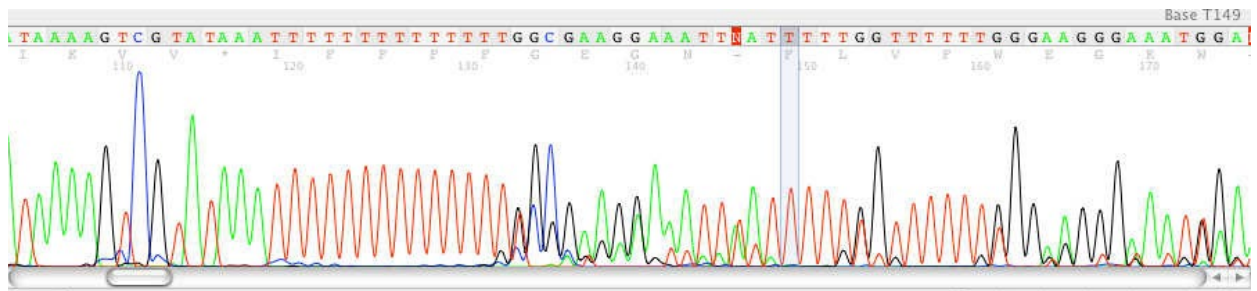


Figure 21. Mononucleotides stretched sequence. Credit: thegenepool

Problem	Possible Cause	Recommended
Overlapping peaks following stretch of single nucleotide sequence	Enzyme slippage occurs giving carrying lengths of the same sequence after this region (n-1, n-2, and n-3 populations)	Sequence the complementary strand

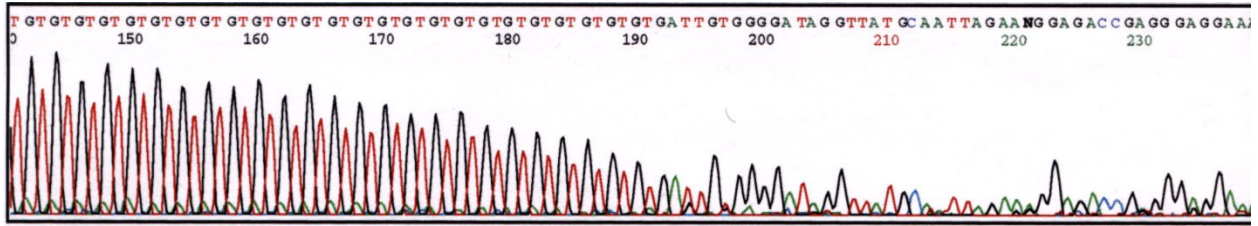


Figure 22. G/T nucleotides stretched sequence. Credit: DNAcore-University of Missouri

Problem	Possible Cause	Recommended
Decrease of signal to the point that no sequence is obtain	Enzyme dissociate from the template cause of long repetitive stretch	Sequence the complementary strand

3.6.7 GC-Rich region

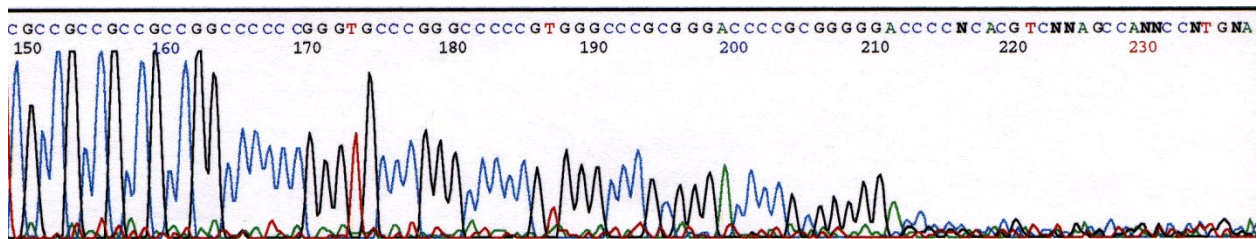


Figure 23. High GC content sequence. Credit: DNAcore-University of Missouri

Problem	Possible Cause	Recommended
Sequence starts well but rapidly loss signal strength.	High GC content	Increase denaturation temperature Sequence the complementary strand

3.6.8 Spikes

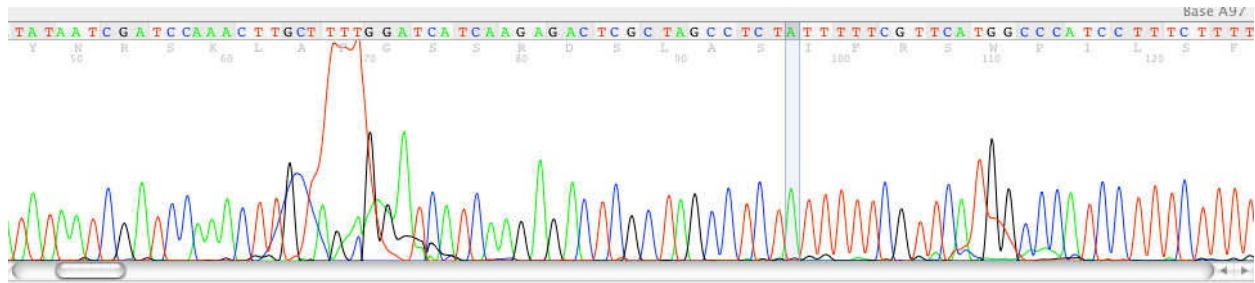


Figure 24. Sequence with spikes. Credit: thegenepool

Problem	Possible Cause	Recommended
Large peaks obscuring the real sequence	Dye blobs caused by unincorporated BigDye and typically seen around 70 and 120 bp. Real sequence can still be read underneath these blobs.	Use less BigDye for sequencing reaction
Sudden large multicoloured peak covering 1-2 bases	Small air bubble of dried polymer within the capillary If related to individual samples, this is due to a contaminant in the sample. Degradation of polymer or capillary array	