出國報告（出國類別：進修）

參加「次世代基因定序資料分析-實作引介」課程心得報告

服務機關：臺北榮民總醫院 病理檢驗部
姓名職稱：何祥齡科主任，葉奕成醫師
派赴國家：德國
出國期間：106 年 12 月 6 日至 106 年 12 月 8 日
報告日期：106 年 12 月 28 日

摘要（含關鍵字）

次世代基因定序的資料分析需要高度的專業，為了研習次世代基因定序的資料分析基礎，病理檢驗部派員前往德國 ecSeq Bioinformatics 機構，學習為期三天的「次世代基因定序資料分析-實作引介」，內容涵蓋 Unix/Linux 系統介紹、次世代基因定序原理及儀器簡介、次世代定序資料分析的方法等等。課程進行的方式以上機實作為主，講演為輔，透過分析實際的資料，了解整個資料分析的流程，以及不同軟體之間差異。經由本次課程，對次世代基因定序資料的分析有進一步的認識。

關鍵字：次世代基因定序、資料分析、生物資訊

目次

一、目的

次世代基因定序(Next Generation Sequencing, NGS)是近十年來快速竄起的核酸定序技術，其原理是利用同時間大量的短序列片段定序，達到高速、高通量的特性。相較於傳統的一代定序法(Sanger 定序)，次世代基因定序的速度快了數萬倍，能迅速產生大量的資料，且單位成本更經濟，因此已廣泛用於全基因體定序 (Whole genome sequencing)、轉錄組定序 (Transcriptome sequencing)、全表現子定序(whole exome sequencing)、以及多基因檢測組合(Multi-gene panel)等用途。

由於次世代基因定序產生的資料動輒數十億至數千億個鹼基，如何正確且有效率地分析這些大量數據需要高度的專業與經驗。分析時從軟體的選擇，參數的設定、軟體串連到最後的驗證，都是很大的挑戰。為了研習次世代基因定序的資料分析基礎，病理檢驗部檢派何祥齡科主任及葉奕成主治醫師，前往德國 ecSeq Bioinformatics 機構，學習為期三天的「次世代基因定序資料分析-實作引介(Next-Generation Sequencing Data Analysis: A Practical Introduction)」。

二、過程

本次前往進修的 ecSeq Bioinformatics 是位於德國的私人教育與研發機構，專精於次世代基因定序資料分析，業務範圍包括提供資料分析服務、次世代基因定序資料分析方法研究，以及舉辦次世代定序資料分析訓練課程等等。ecSeq Bioinformatics 設址於德國萊比錫，講師及成員主要來自於德國萊比錫大學之生物資訊學博士，自 2013 年起每年於德國各大城市包括萊比錫、柏林、法蘭克福、慕尼黑等地舉辦短期的次世代定序資料分析訓練課程。本次參加的是於慕尼黑所舉辦的「Next-Generation Sequencing Data Analysis: A Practical Introduction」課程。

參與此次課程的成員共有 20 位，除來自台灣的我們之外，尚有來自美國、德國、瑞典、法國、泰國、丹麥、義大利、比利時等國家的學員。學員的背景大多是生

物醫學方面的博士級研究人員以及在醫院工作的臨床醫事人員，很特別的是也有來自藥廠的統計學家。由於這次的課程是引介課程，除了少數的學員之外，大部份的學員都沒有次世代定序資料分析的經驗。上課地點在慕尼黑的 Computer & Management Training GmbH 電腦教室，每位學員都有一部電腦可實際操作。課程進行的方式以上機實作為主，有充份的時間可以實機練習。講師共有兩位：Dr. Mario Fasold 及 Dr. Gero Doose，都是萊比錫大學的生物資訊博士。

這三天課程的內容約略可分成以下三個部份：

(一) Unix/Linux 系統介紹：

Unix 是 1969 年在 AT&T 的貝爾實驗室開發的多用戶、多工作業系統，後續又有許多由 Unix 擴充改進而衍生出的作業系統，稱為類 Unix 作業系統。其中廣為大家所使用的 Linux，即是一種自由、開放的類 Unix 作業系統。由於目前大多數的次世代定序資料分析軟體都需在 Linux/Unix 的作業系統架構下才能執行，因此要學習次世代定序資料分析的第一步，便需先認識 Linux/Unix 的作業系統的使用方式。第一天的課程主要教導學員如何操作 Linux/Unix 作業系統的終端機(terminal)文字介面，並學習以指令列進行檔案管理、資料下載、文字檔內文搜尋、檔案切割、指令串接等基本功能。

(二) 次世代基因定序的原理、應用以及儀器簡介：

次世代基因定序的原理是將欲定序的核酸打成小片段，再大量且快速地同時進行短序列片段定序，並運用生物資訊運算進行片段的接合，進而得到完整的核酸序列。主要實驗流程包含核酸片段化 (Fragmentation)、建庫 (Library Construction)、高通量定序 (High-throughput Sequencing)、數據分析 (Data Analysis) 等步驟。每一家儀器廠商所使用的技術各有不同及優缺點。例如市佔率最高的 Illumina 平台，定序準確度高，且各種不同層級的機種相當完備；PacBio 平台雖然定序錯誤

率高出許多，但擁有長讀長(long read length)的獨門測序技術，對於 de novo sequencing 及分析 splicing isoforms 具有無可比擬的優勢；而 Nanopore 平台的 MinION，雖然產出資料量較低，但是體積只有一個 USB 隨身碟的大小，易於攜帶，已被美國太空總署應用於在太空中進行核酸定序的測試。

(三) 次世代基因定序資料分析的方法：

1. 次世代基因定序儀器所產出的原始數據，到最後結果之分析，其間須進行許多步驟，各個步驟分別需使用不同的軟體。以 DNA sequencing analysis 為例，其生物資訊分析流程簡略如下圖：

Step	Process	File Format	Software
1	Sequence raw reads	FASTQ	N/A; obtained from Sequencer
2	Quality Control	FASTQ	FASTQC
3	Trimming	FASTQ	Cutadapt
4	Mapping and Alignment	SAM/BAM	Bowtie, BWA, STAR, Segemehl
5	Variant Calling	VCF	GATK, Freebayes
6	Variant Filtering and Annotation	N/A	N/A

本次上課的部分主要著重於

1) Quality Control
2) Trimming
3) Mapping and Alignment
4) Variant Calling

2. <u>Quality Control</u>

2.1. FASTQ 檔案格式:

2.1.1. 在次世代基因定序資料中，我們大部分獲取的定序結果檔案格式為

FASTQ 格式，以文字的格式儲存定序之結果。如下圖:

FASTQ format

TASK

Take a look at the FASTQ file of mate 1

```
$ zcat SRR359063_1.fastq.gz | head
Line 1 → @SRR359063.1 D042KACXX:3:1101:2690:2160 length=101
Line 2 → NCATCGTCCGGTATGTAGAACAGGGGAACCGGACGTTTTCCAAGGCGTAGCCATGTTAGACAAGGCGCAGATATAG
GTGATGCTGATGCAGAAAAACGATT
Line 3 → +SRR359063.1 D042KACXX:3:1101:2690:2160 length=101
Line 4 → #4=DBDDDHFHFFHIGHIIIJJJJJJJJJJJJJBHDAGHJGGGHIJHFFFFDDEDCCDCCCCDDDDDBDBD>CDEE>
C@CDDDDDDCACAACCDCDBDBB<1
@SRR359063.2 D042KACXX:3:1101:5202:2193 length=101
CTCTGGTACAGAACACGTGGATTATAAGAGTTGCCGCTTCGCACAGAAGTCGGAGTTCTCTCACCACTTTTGAGCT
CTTCCTCGGCTTCTTCTTCCTCTTT
```

Line 1: 始於@符號，表示此序列的基本名稱及描述

Line 2: 原始序列資訊

Line 3: 始於+符號，與 Line 1 相同，描述此序列之基本資料

Line 4: 表示每個定序酸基的定序"Phred Quality Score"

2.1.2. Phred Quality Score 為 illumina 系統之核酸定序品質指標，其需對應 ASCII 表格換算(下圖)。當 FASTQ 檔案之 Phred Quality Score 顯示為 F 時，其對應 ASCII 表為 70，70-33=37 (由 33 之後的 ASCII 表示開始算)， 因此，此鹼基之 Phred Quality Score 為 37。

Dec	Hx	Oct	Char	Dec	Hx	Oct	Chr	Dec	Hx	Oct	Chr	Dec	Hx	Oct	Chr
0	0	000	NUL (null)	32	20	040	Space	64	40	100	@	96	60	140	`
1	1	001	SOH (start of heading)	33	21	041	!	65	41	101	A	97	61	141	a
2	2	002	STX (start of text)	34	22	042	"	66	42	102	B	98	62	142	b
3	3	003	ETX (end of text)	35	23	043	#	67	43	103	C	99	63	143	c
4	4	004	EOT (end of transmission)	36	24	044	$	68	44	104	D	100	64	144	d
5	5	005	ENQ (enquiry)	37	25	045	%	69	45	105	E	101	65	145	e
6	6	006	ACK (acknowledge)	38	26	046	&	70	46	106	F	102	66	146	f
7	7	007	BEL (bell)	39	27	047	'	71	47	107	G	103	67	147	g
8	8	010	BS (backspace)	40	28	050	(	72	48	110	H	104	68	150	h
9	9	011	TAB (horizontal tab)	41	29	051	)	73	49	111	I	105	69	151	i
10	A	012	LF (NL line feed, new line)	42	2A	052	*	74	4A	112	J	106	6A	152	j
11	B	013	VT (vertical tab)	43	2B	053	+	75	4B	113	K	107	6B	153	k
12	C	014	FF (NP form feed, new page)	44	2C	054	,	76	4C	114	L	108	6C	154	l
13	D	015	CR (carriage return)	45	2D	055	-	77	4D	115	M	109	6D	155	m
14	E	016	SO (shift out)	46	2E	056	.	78	4E	116	N	110	6E	156	n
15	F	017	SI (shift in)	47	2F	057	/	79	4F	117	O	111	6F	157	o
16	10	020	DLE (data link escape)	48	30	060	0	80	50	120	P	112	70	160	p
17	11	021	DC1 (device control 1)	49	31	061	1	81	51	121	Q	113	71	161	q
18	12	022	DC2 (device control 2)	50	32	062	2	82	52	122	R	114	72	162	r
19	13	023	DC3 (device control 3)	51	33	063	3	83	53	123	S	115	73	163	s
20	14	024	DC4 (device control 4)	52	34	064	4	84	54	124	T	116	74	164	t
21	15	025	NAK (negative acknowledge)	53	35	065	5	85	55	125	U	117	75	165	u
22	16	026	SYN (synchronous idle)	54	36	066	6	86	56	126	V	118	76	166	v
23	17	027	ETB (end of trans. block)	55	37	067	7	87	57	127	W	119	77	167	w
24	18	030	CAN (cancel)	56	38	070	8	88	58	130	X	120	78	170	x
25	19	031	EM (end of medium)	57	39	071	9	89	59	131	Y	121	79	171	y
26	1A	032	SUB (substitute)	58	3A	072	:	90	5A	132	Z	122	7A	172	z
27	1B	033	ESC (escape)	59	3B	073	;	91	5B	133	[	123	7B	173	{
28	1C	034	FS (file separator)	60	3C	074	<	92	5C	134	\	124	7C	174	|
29	1D	035	GS (group separator)	61	3D	075	=	93	5D	135	]	125	7D	175	}
30	1E	036	RS (record separator)	62	3E	076	>	94	5E	136	^	126	7E	176	~
31	1F	037	US (unit separator)	63	3F	077	?	95	5F	137	_	127	7F	177	DEL

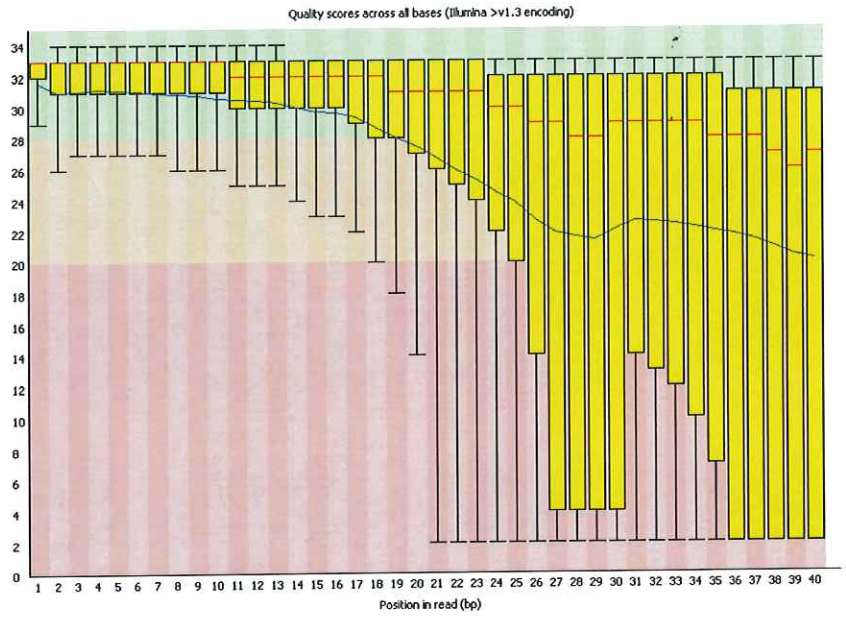
2.1.3. 當 Phred Quality Score 為 30 時，則一個鹼基被測錯的機率為 0.001，其對應表如下:

Table 1: Quality Scores and Base Calling Accuracy

Phred Quality Score	Probability of Incorrect Base Call	Base Call Accuracy
10	1 in 10	90%
20	1 in 100	99%
30	1 in 1,000	99.9%
40	1 in 10,000	99.99%
50	1 in 100,000	99.999%

2.1.4. 一般來說，鹼基定序的 Phred Quality Score 要大於 20 以上，才算是可信的定序結果。

2.2. FastQC 為一個用來檢查次世代基因定序品質的步驟，使用的軟體為 FastQC tool，其顯示的圖形為下圖所示:

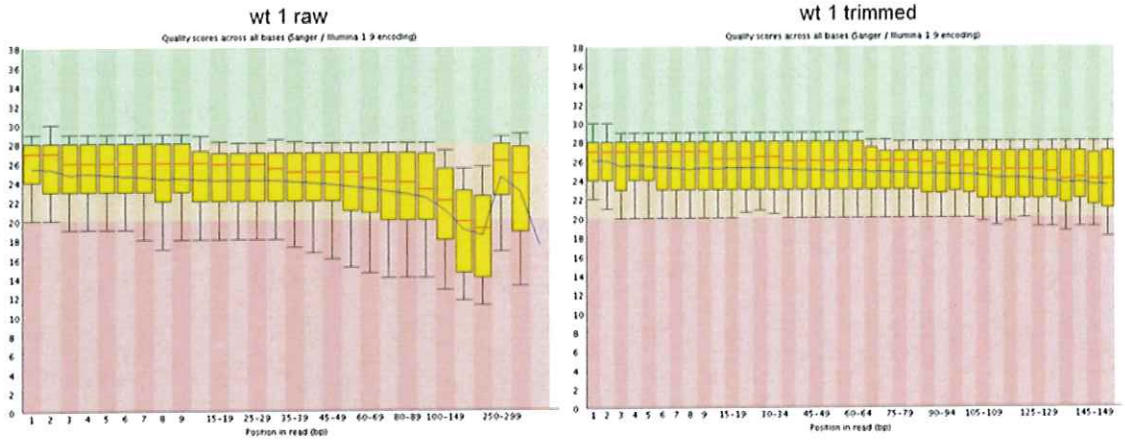


2.2.1. 橫坐標代表每一個定序鹼基，反映了讀長信息，比如定序的讀長為 150 bp,橫坐標就是 1 到 150；縱坐標代表鹼基定序的 Phred Quality Score; 中間紅線代表中位數; 黃色箱圖代表四分位數 25%-75%; 藍色的線代表平均值。

2.2.2. 一般來說，鹼基定序的 Phred Quality Score 要大於 20 以上，才算是可信的定序結果。因此，透過 FastQC，可用來評估每個片段定序的品質，鹼基定序品質不佳的部分可透過 trimming 步驟來刪除。

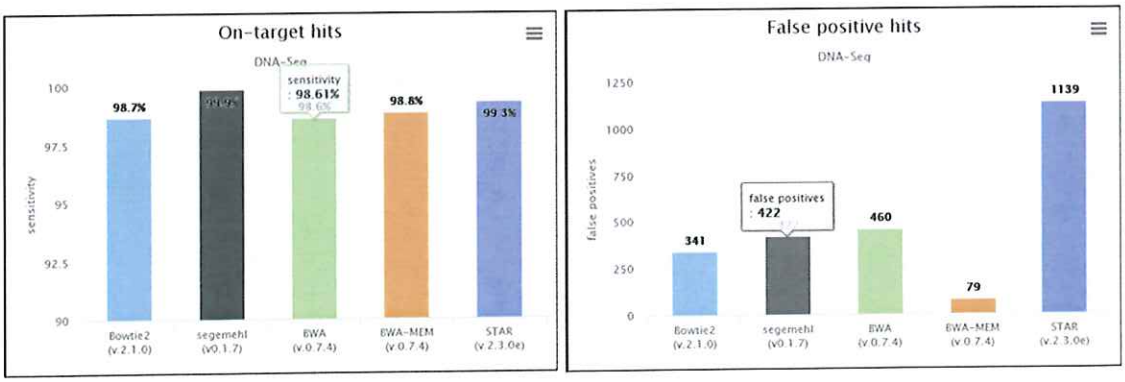
3. Trimming

Trimming 為一個用來切除定序品質不佳之鹼基或定序反應過程之 Adaptor sequence(illumine system)的步驟，使用的軟體為 Cutadapt。切除完之序列再使用 Fastqc 軟體進行檢查。如下圖，左圖為 trimming 前之結果，全長為 299 bp，但超過 150 bp 之後，定序品質不佳。右圖為 trimming 後之結果，全長為 150 bp。

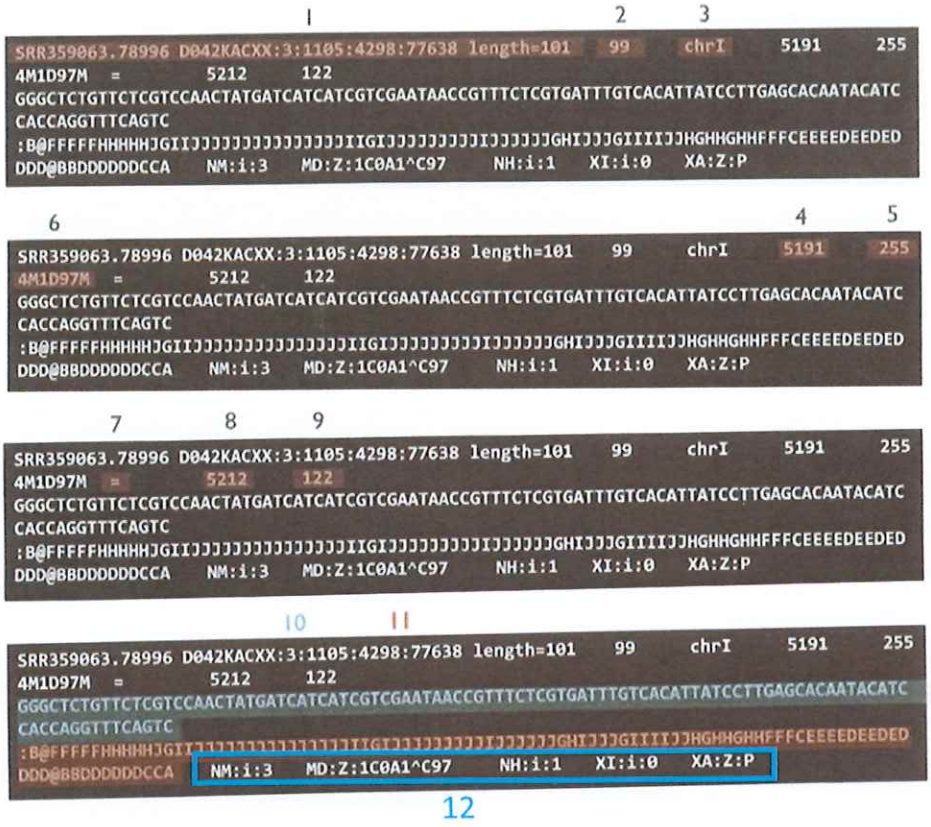


4. Mapping and Alignment

4.1. 此部分使用的軟體有很多，例如 Segmehl、Bowtie2、BWA、BWA-MEM 及 STAR 等。不同的軟體其 mapping 的靈敏度及偽陽性程度皆不相同。如下圖所示。



4.2. 此步驟使用之檔案格式為 SAM/BAM 檔案



Lane 1: 分析序列的名稱

Lane 2: Bitwise flag

Lane 3: 參考序列

Lane 4: 序列最左邊第一個核酸位置

Lane 5: mapping quality

Lane 6: CIGAR string

Lane 7: ref. name of the mate/ next segment

Lane 8: position of the mate/ next segment

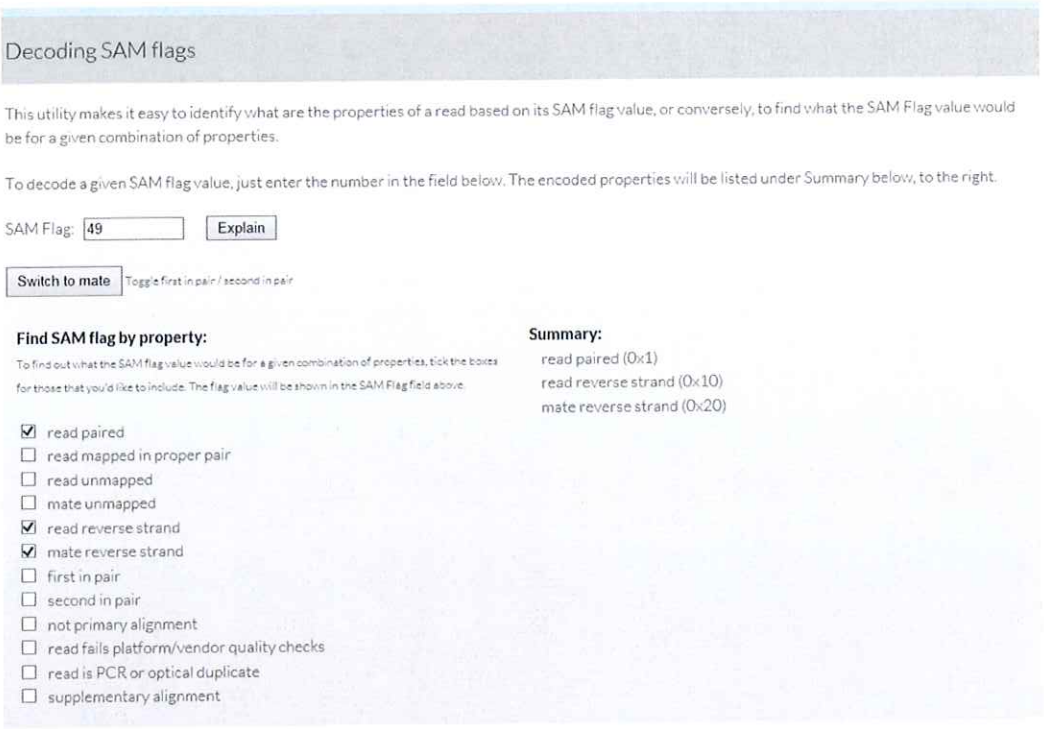
Lane 9: template length

Lane 10: segment sequence

Lane 11: Phred score quality

Lane 12: TAG

4.3. 透過 Bitwise flag 可以告訴我們此序列 mapping 結果之特性，其計算分數如下圖(https://broadinstitute.github.io/picard/explain-flags.html)



4.4. CIGAR string

GIGAR string 是用來描述此定序序列與參考序列比對後之結果。例如:

GIGAR string 為 1M1I28M1D3M 時，代表此序列由左至右為:

5.4.1. 1M: one match/mismatch

5.4.2. 1I: one insertion

5.4.3. 28M: 28 matches/mismatches

5.4.4. 1D: one deletion

5.4.5. 3M: 3 matches/mismatches

4.5. TAGs

4.5.1. Alignment 的 TAGs 描述又分成三部分:

(1) MD: 代表 mismatching positions

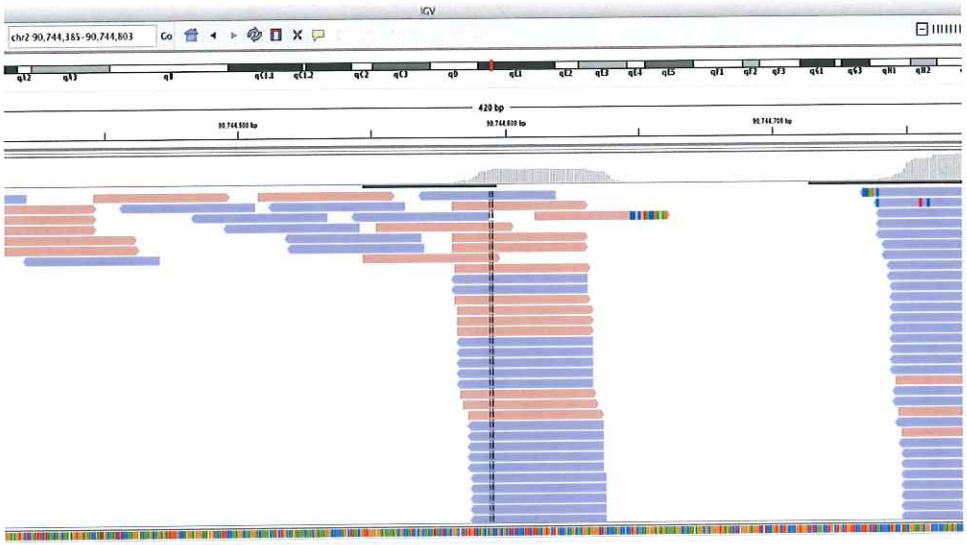
(2) NH: 此序列對應到的 hit 數目

(3) NM: edit distance to the reference，與標準序列之差異

4.5.2. 其中 MD-TAG 較常用，如 MD-TAG 為 3C3T1^GCTCAG26 代表此序列與標準序列比對後之結果為:

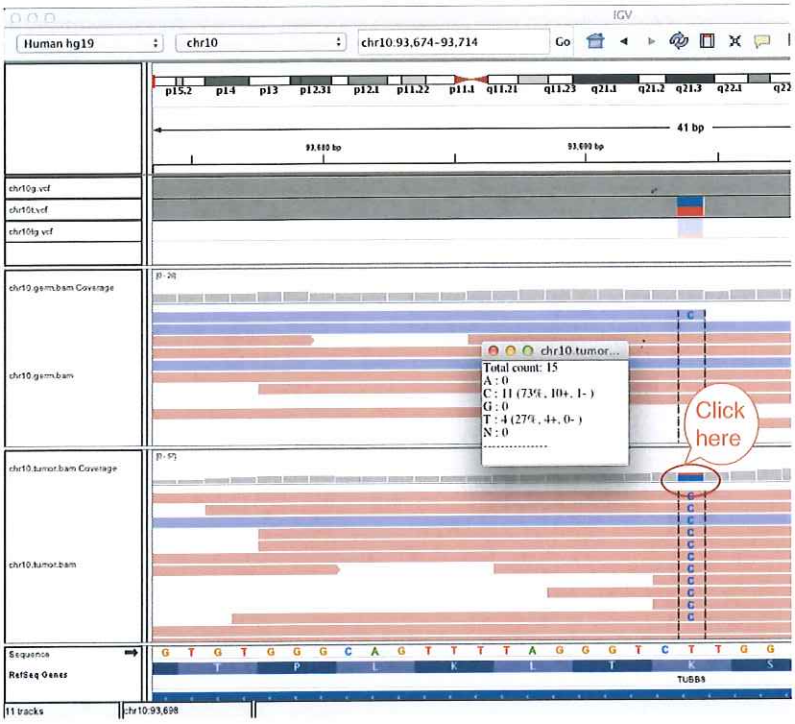
3 個 matches - C(unmatch) - 3 個 matches - T(unmatch) - 1 個 match - insertion of GCTCAG - 26 個 matches

4.6. SAM/BAM 檔案格式是用文字方式來儲存比對定序結果之資料，我們可將其 import 至 IGV genome browser 轉換成圖檔查看，如下圖:



5. Variant Calling:

Variant calling 為將比對序列與參考序列之間的差異結果，連結到參考序列之染色體與基因位置，並判斷是何種變異之步驟。其使用之軟體為 freebayes，檔案格式為 VCF 檔案。Variant calling 之後的 VCF 檔案，亦使用 IGV genome browser 轉換成圖檔查看，如下圖，參考序列為 T，定序比對之結果，此序列可判讀之 reads 共 15 條，其中為 C 的有 11 條(73%)，T 的有 4 條(27%)。



6. 上課時講師帶領著學員一步一步地實際操作分析的過程。由於次世代基因定序資料數據十分龐大，分析相當耗時，為了節省時間，本次課程讓學員們以 C. elegans 的次世代定序實驗數據資料為分析的材料，大大提升了學習的效率。此外，講師也將學員們分成幾個小組，每個小組分別以不同的軟體進行分析，並互相比較分析的結果。經由這樣的過程，發現使用不同的軟體，所得到的分析結果可能會有很大的差異。例如使用 Segemehl 做為 alignment software 的組別，所得到的 mapped reads 可在 95%以上，但使用 Bowtie 做為 alignment software 的組別，所得到的 mapped reads 卻只有 80%。講師們也針對造成這種差異的原因進行講解，提醒學員們在面對不同類型的資料時，應注意選擇合適的軟體進行分析。

三、心得及建議事項

這次的課程雖然只有短短的三天，但收穫十分豐富，心得與建議事項如下：

1. 本次課程之同學來自歐洲各個不同的國家，幾乎都是生物醫學相關之博士與

學生及藥廠研發人員，甚至有從美國來的助理教授，可以看出生物資訊發展是未來精準醫學之重要關鍵，尤其是針對基因體大數據分析領域。在次世代定序，甚至未來的第三代定序，實驗技術層面已經不是太難克服的問題，相對的是當我們得到龐大之數據後，如何去分析與運用，應是未來積極努力的方向與目標。

2. 目前國內基因體大數據分析之發展仍不是十分普級，且相較下也較缺乏相關人才。雖然國內偶有開設次世代基因定序資料分析相關的短期工作坊或課程，但是課程大多零散且片段，只針對某個軟體進行講解，缺乏像本次德國課程完整及系統性之講解與分析。且此次課程是針對完全沒有生物資訊背景的人來說可以學習非常完整之內容與概念。且課程的規畫、資料分析的材料選擇，到講師講解的內容都有值得國內課程借鏡之處。

3. 本次課程可以深切體認到，次世代基因定序資料分析甚至是大數據分析，為相當專業的領域，從軟體的選擇、參數的設定、到分析結果的判讀等，皆需要專業的知識與訓練，否則很容易導致分析錯誤，產生「Garbage In, Garbage Out」的情況。此外，這個領域的變動與進展極為迅速，三不五時就有新的分析軟體發表，十分具有挑戰性。由於正確的資料分析是一個次世代基因定序實驗室不可或缺的一環，因此實驗室中必須有專業的生物資訊人員來負責。建議本院未來應積極派遣同仁進修學習相關知識，及延攬具次世代定序資料分析經驗與專業之人才，以因應精準醫學之發展。

4. 有鑑於 Illumina 定序分析平台為全球目前市佔率最高的平台，本次課程所講授的主要是 Illumina 平台的資料分析，若是使用其他平台執行次世代基因定序，在生物資訊分析流程方面應需要再進一步設計與調整，或者使用其他合適的軟體。本部今年將採購 ThermoFisher Scientific 的 Ion S5 平台，此平台相較於 Illumina 平台，較適用於臨床檢測，且其考量臨床檢測之用途，在生物資訊分析方面有設計自行之的分析軟體套件，亦有 oncomine 資料庫聯結。未來也將進一步學習並了解 Ion S5 平台相關之資訊分析。

四、次世代定序分析並不僅止於單純定序資料的分析，若能將進一步將個定序分析之結果與其他臨床指標聯結，建置大數據資料庫，將有助於臨床重要發現，亦是發展精準醫療之一環。建議本院除發展精準醫學之檢測外，也能進一步考量如何有效統合各部門，將基因資訊與臨床資料庫進行整合與聯結，與精準醫療發展相輔相成。

五、附錄

本次課程學員合影