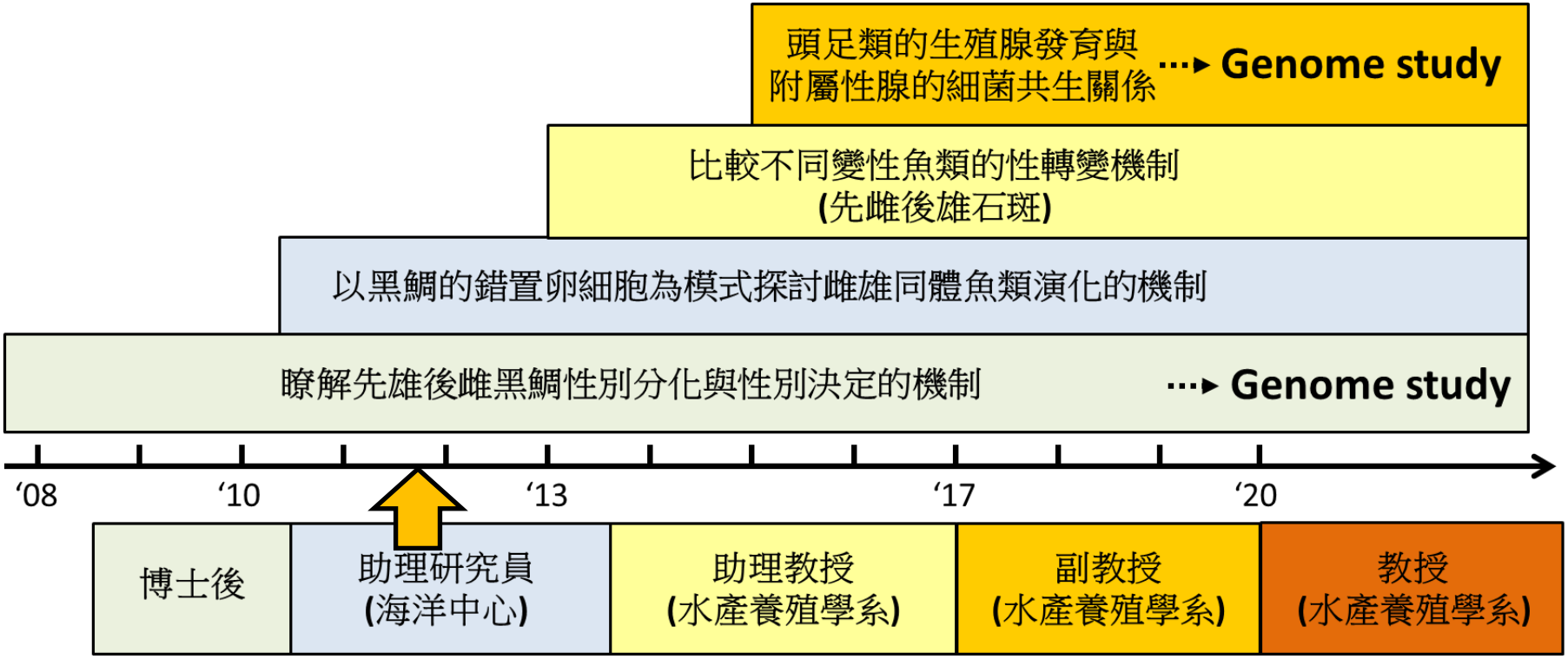
我嘗試進行的基因體組裝：從454到三世代定序的心得分享

吳貫忠

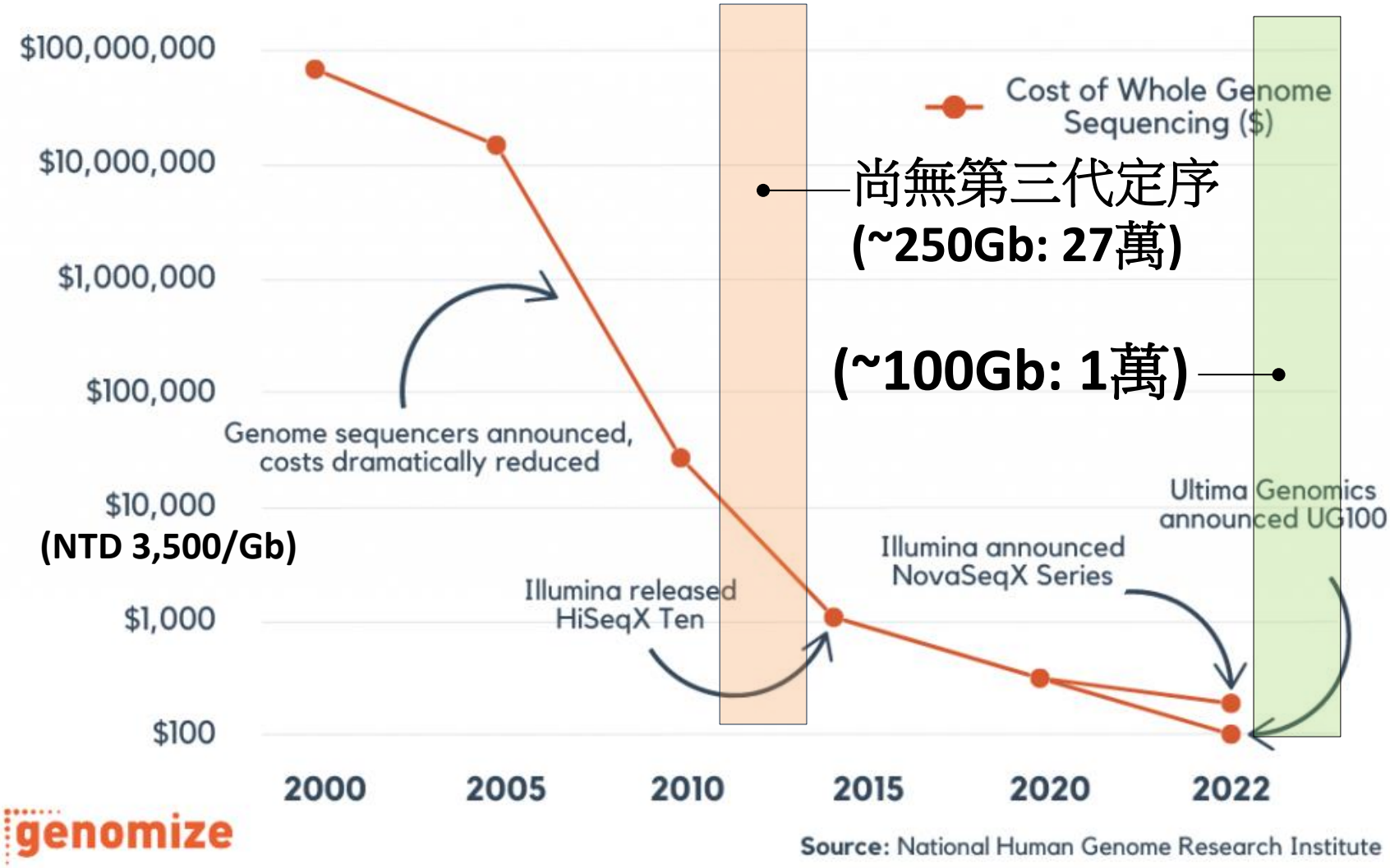
國立臺灣海洋大學 水產養殖學系

Current research directions

33 of 45 articles, H-index = 21, i10-index = 33



Genome sequencing costs



Draft genome of fish species

2014 **Draft genome**

Nature communications (rainbow trout, yellow croaker, and European seabass)

N50 = ~500 kb, ~10,000 scaffolds

Nature genetics (grass carp and tongue sole)

N50 = ~500 kb, ~500 scaffolds

2024 **Comparative genomics**

Nature communications (contains some fishes genome data)

>10 species in one article and/or with chromosomal level genome

BMC genomics…etc.

Draft genome in one fish species

基因體定序的旅途中的停看聽

1. 先射箭再畫靶? 還是先畫靶再射箭?

需要基因體才能解決問題?還是只是想要有基因體！

臺灣特有臺灣鮭魚之基因體解碼
(2014~2016年)

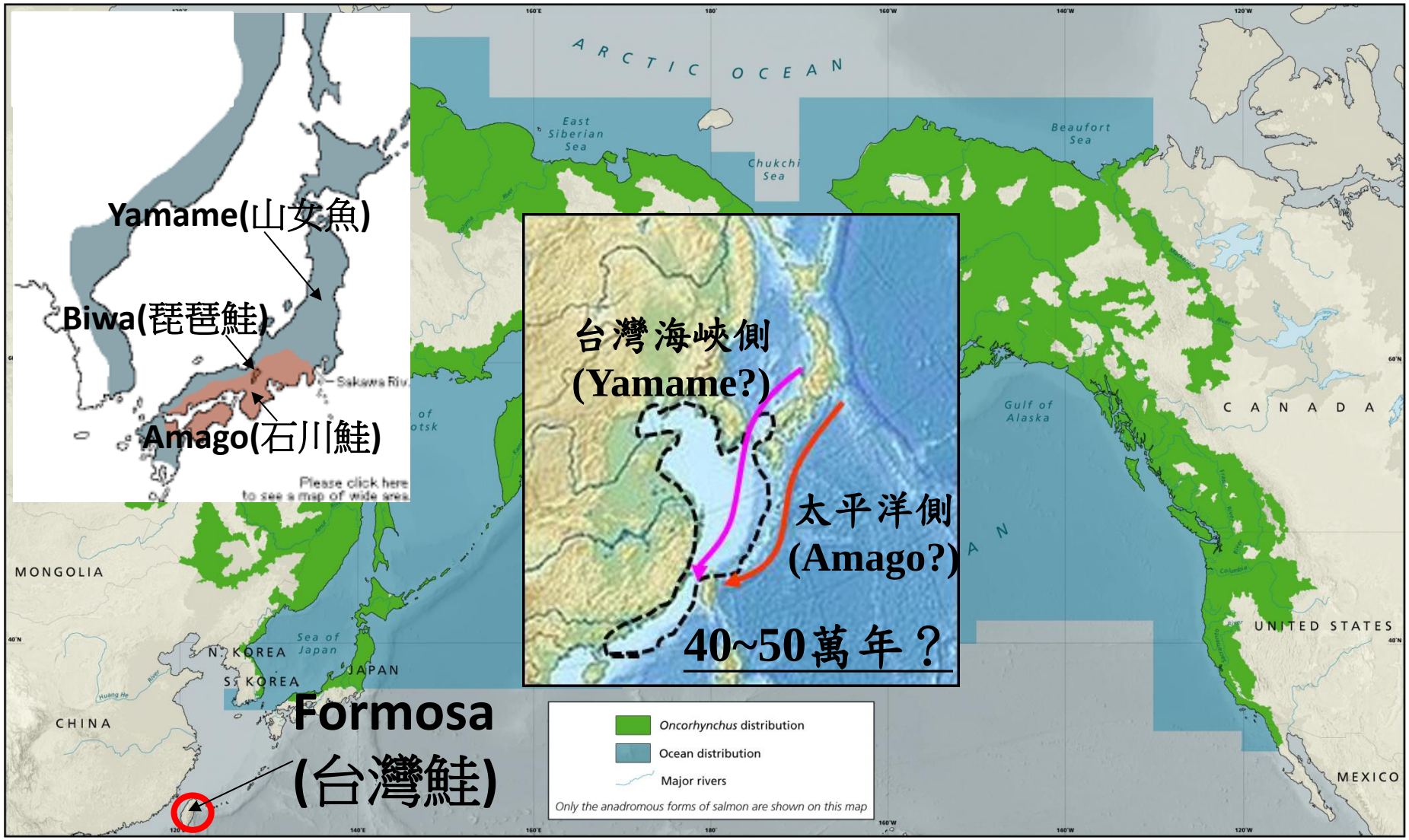
臺灣鮭在生態與學術的重要性:

1.臺灣鮭為「**陸封性鮭魚**」也是目前世界上分布最低緯度的鮭魚。

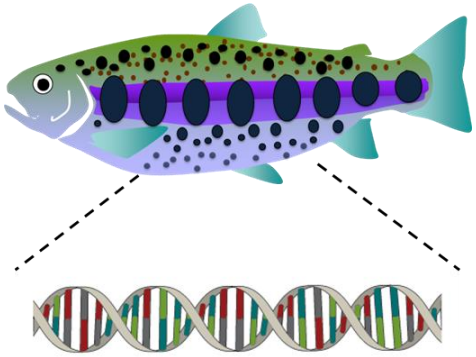
2.可能是世界上唯一無法在降海迴游(適應海水)的鮭魚。

3. 2007年，中央研究院執行「數位典藏國家型科技計畫」第二期計畫（2007年～）將臺灣鮭模式標本列為計畫重點，希望釐清命名的疑點。

Original distribution of Pacific salmon (*Oncorhynchus*)



臺灣鮭魚之基因體解碼之價值與貢獻



1. 臺灣鮭的演化歷史(比較日本產的Yamame和Amago)→**從台灣海峽或太平洋遷徙？**
2. 臺灣鮭還能**降海(適應海水)？**
3. 比較虹鱒(多次產卵)與臺灣鮭(產卵後死亡)在基因數量上的變化→**老化的原因！**
4. 鮭魚第4次基因重組後(陸生脊椎動物2次，硬骨魚類3次)**新基因的去留？**

基因體定序的旅途中的停看聽

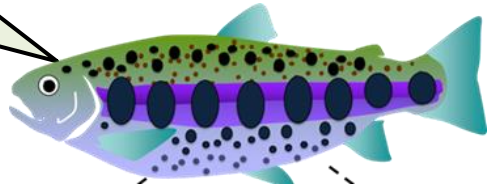
1. 先射箭再畫靶? 還是先畫靶再射箭?

需要基因體才能解決問題?還是只是想要有基因體！

2. 我需要準備多少的經費? 生物資訊部分?

0.7Gb genome size >300,000 NTD (不含分析)

計畫開始前我們對台灣鮭基因體瞭解多少？



1. 基因體大小約2.7Gb (經細胞核染色推估)
2. 基因體初探(survey)符合2.7G (150Gb data，>50X coverage)
3. 預測基因數目(>20K)

鮭魚基因體解序目標: **解序90%以上，N50>100K。**

Genome sequencing strategy

Rainbow trout (2.4Gb)

BAC-end sequences

1,000,000 reads/flowcell

Genomic Roche454 FLX libraries (8k-, 12k-, 20k mate-pairs)

Genomic Illumina libraries **Nature Communications, 2014**

200,000,000 reads/flowcell

Sea bass (0.67Gb)

BAC-end sequences

Genomic Roche454 FLX libraries (~20k mate-pairs)

Genomic Illumina libraries **Nature Communications, 2014**

Taiwan salmon (2.7Gb)

~~**BAC-end sequences**~~

Genomic Roche454 FLX libraries (4k-, 8k-, 16k- mate-pairs)

Genomic Illumina libraries

臺灣鮭基因體解序會面臨的分析問題：

1. **BAC–end sequence以2.7Gb大小的genome，需要建立約1500個libraries (coverage>100X)。經費龐大，計畫無法負擔！**
2. **除了上述原因，臺灣鮭為保育類，無法利用親代與子代比較基因交聯來建立連鎖地圖(linkage-map)。**

目前的Genome組裝工具

Short reads

Illumina NovaSeqX Series → **Genetic markers**

(20~30X coverage) **Re-mapping**

Long reads

貴

Illumina PacBio HiFi → **Scaffolds**

Nanopore

挑樣本

Illumina PacBio Hi-C → **Scaffolds**

吃經驗 **(Chromosomal levels)**

Transcriptome → **Coding gene**

Summary of genome assembly

Species (genome MW/cell)	Genome assembly	Scaffold	Size (Gb)	N50
Taiwan salmon	(> 100 bp) (v.2.1)	2,196,540	2.74	276.5 Kb
Taiwan salmon (2.7 Gb)	(> 500 bp)	**49,767**	**2.44**	**340 Kb**
Rainbow trout (2.4 Gb)	(> 500 bp)	79,941	1.88	383.6 Kb
Atlantic salmon (3 Gb)	(> 500 bp) (1n)	843,055	3.40	493.6 Kb

DNA transposon and retrotransposon (9.16%)
Others interspersed repeats (26.72%)
Total repetitive DNA: 38.64% of genome

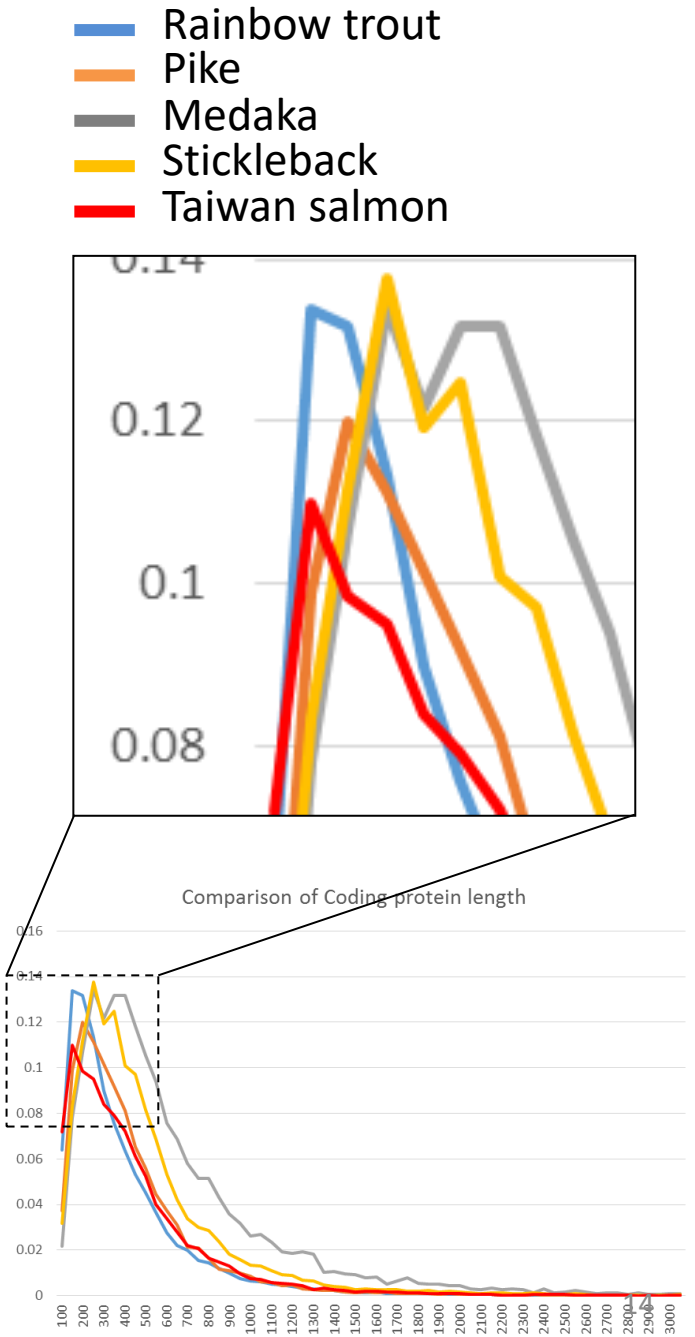
Comparison of repetitive DNA:
Medaka (9.2%) < Salmon (38%) < Zebrafish (52%)

臺灣鮭基因體組裝已符合要求 (Coverage > 90%，N50 > 100K)

基因註解 (gene annotation)

Summary of genes	Taiwan salmon
Significant gene	36,248
Average transcript length (bp)	1300
Average protein length (aa)	423
Number of exon per gene	7.53
Average exon length (bp)	176.9
Percent with Unigene identified	90.4%
Percent with GO function	71.2 %
Number of GO annotations	78,411
Percent with Protein Domains	95.0%

Gene set	Gene Count
5.8s rRNA	143
18s rRNA	15
28s rRNA	16
Sum of total	174



臺灣鮭基因模型與轉錄體之關係

Table. Summary of sequences

	Gene #	Size	N50	Mean	Max	File
Transcriptome	31,957	11,256,127	499	352	12,156	Trinity.final.cds
Genome	36,248	15,316,774	565	423	21,988	protein.nrg.fasta

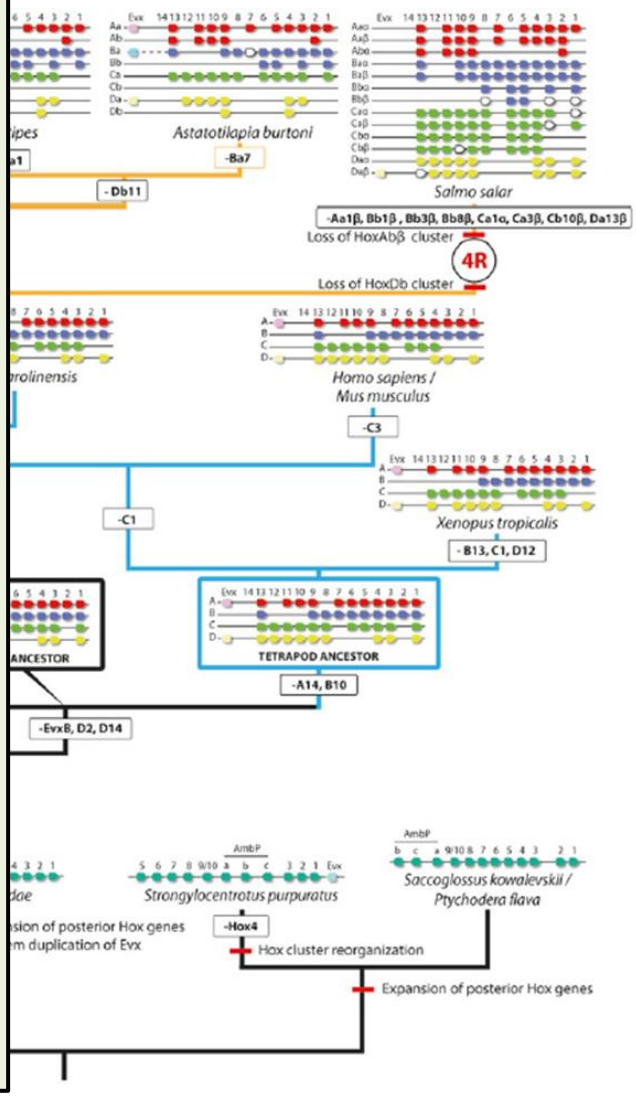
Table. Summary of Blast results

1e-10 (blx)	Significant hits	Percentage
Transcriptome (query)	26,709	83.6%
Genome (subj_t)	**31,761**	**87.6%**

Of the 36,248 predicted genes, 31,761 (88%) were supported by transcriptome data at least 1 of 3 tissues examined (muscle, brain and gill).

Comparison of the Atlantic salmon and Taiwan salmon *hox* clusters

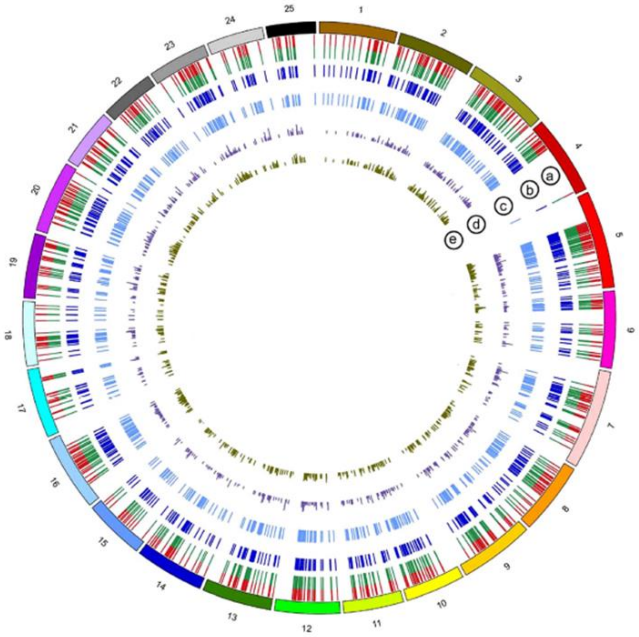
	14	13	12	11	10	9	8	7	6	5	4	3	2	1
SS Aa1														
TS Aa1		1440		1435							1438	1436	1439	1437
SS Ab1														
TS Ab1		24901		24902								24905	24903	
SS Ab2														
TS Ab2		6842		6844	13831	13833							13830	
SS Ba1														
TS Ba1		10571			10574	10573	10575			10572	10567	10569	10570	10568
SS Ba2														
TS Ba2		24287				24293	24292		24294	24291	24289	24290	24288	24295
SS Bb1														
TS Bb1							33576		33573	33575		33572		33571
SS Bb2														
TS Bb2									11409	11408				
SS Ca1														
TS Ca1		8939	8940	8941	8942	29674	29675		29672		29671			29673
SS Ca2														
TS Ca2		26390	26391	26392		26396			26395	26397	26394	26399		26398
SS Cb1														
TS Cb1		13188	23918	23919	23922	23920	23921		23925	23924	23923			
SS Cb2														
TS Cb2		30890	30892	30891		30893	30894		30888		30887			
SS Da1														
TS Da1		160	159		158						153	155		154
SS Da2														
SS Da2			24525		24524						24523	24522		



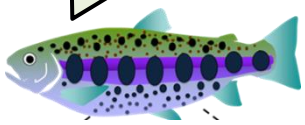
The gene annotation reveal that Taiwan salmon is following a 4th round whole genome duplication.

Comparison of the gene repertoire of the Taiwan salmon with zebrafish

zebrafish: 25,453 Taiwan salmon: 36,248	Idn (%)	Aln (%)	Ortholog (zebrafish)	ratio(%) zebrafish	ratio(%) Taiwan salmon
Ortholog filter	30	50	15689	62%	87%
	30	70	11602	46%	64%
	30	80	8271	32%	46%
	30	90	3746	15%	21%
	40	80	7632	30%	42%
	50	80	6413	25%	35%
	60	80	5129	20%	28%
	70	80	3729	15%	21%
	70	90	2157	8%	12%



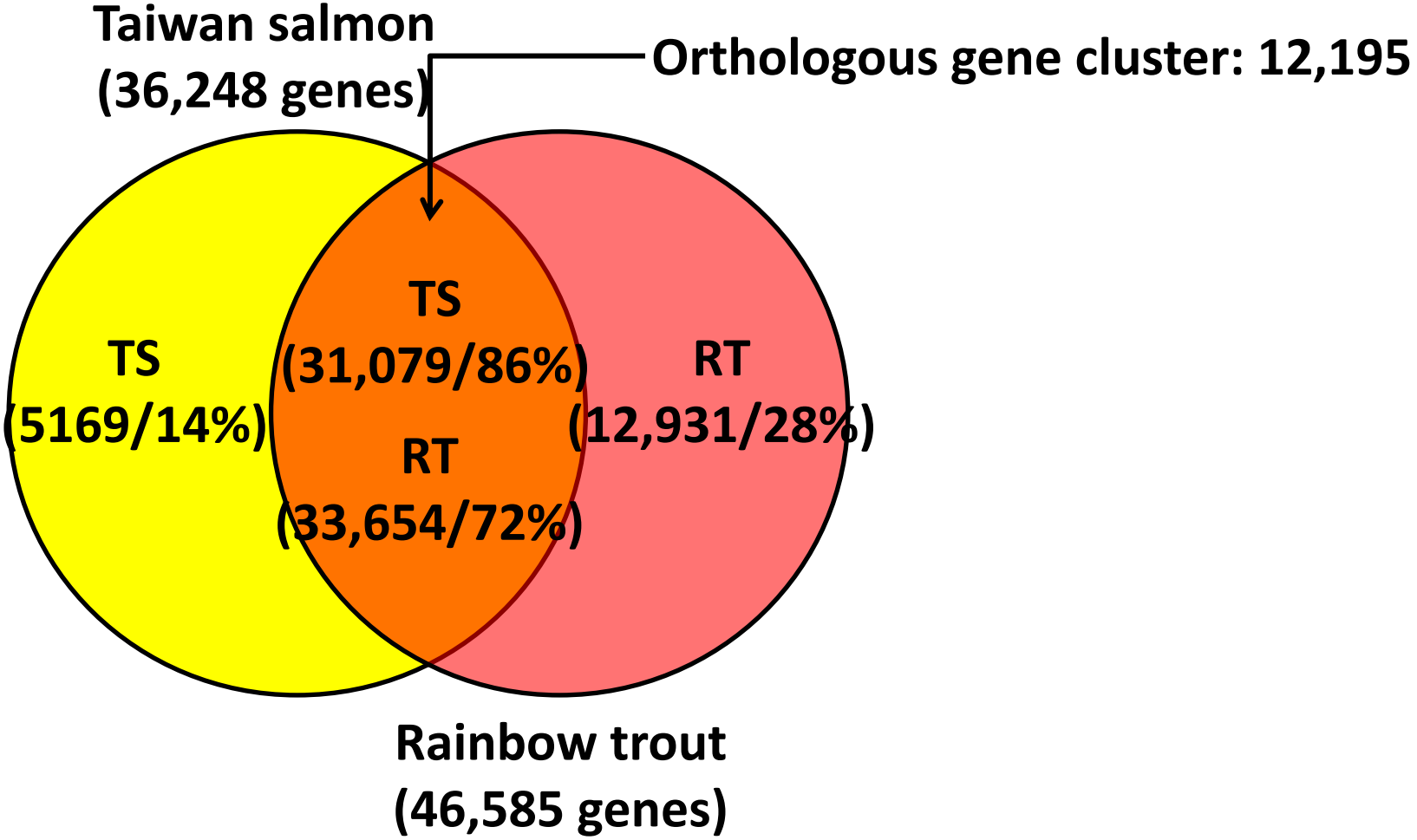
相較於虹鱒(2.4Gb)，臺灣鮭(2.7Gb)多出來的0.3Gb為何？



4R虹鱒的研究指出(相較於3R斑馬魚)，約有50%的duplicated genes會被保留下來。

8百萬年的演化對虹鱒(RT)和臺灣鮭(TS)基因體的影響為何？

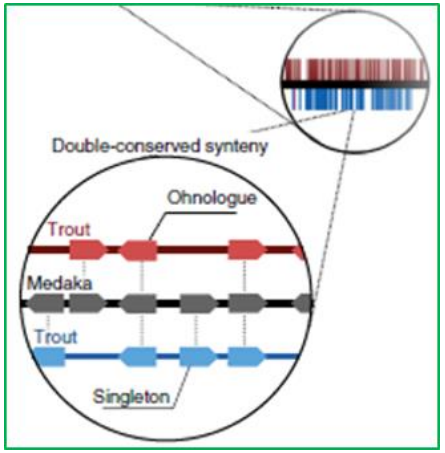
台鮭與虹鱒進行 Reciprocal BLASTP



台灣鮭與虹鱒同源基因與非同源基因比較

同源基因關係分類, 鮭形目基因體重覆於硬骨魚類

OrthoMCL 結果 (針對orthologs.txt 分析)



- 1 對1 : orthologous : 10,997
- 1 對2 ; 1對N (singleton) : 2415
- 2 對1; N對1 (singleton) : 784
- 2對2 ; N對N (ohnologous) : 8502

OrthoMCL group 結果

- 1 對1 (orthologous):: 9,052 (包含其它的17,689 同源結果)
- 2 對2 (ohnologous): 1,897

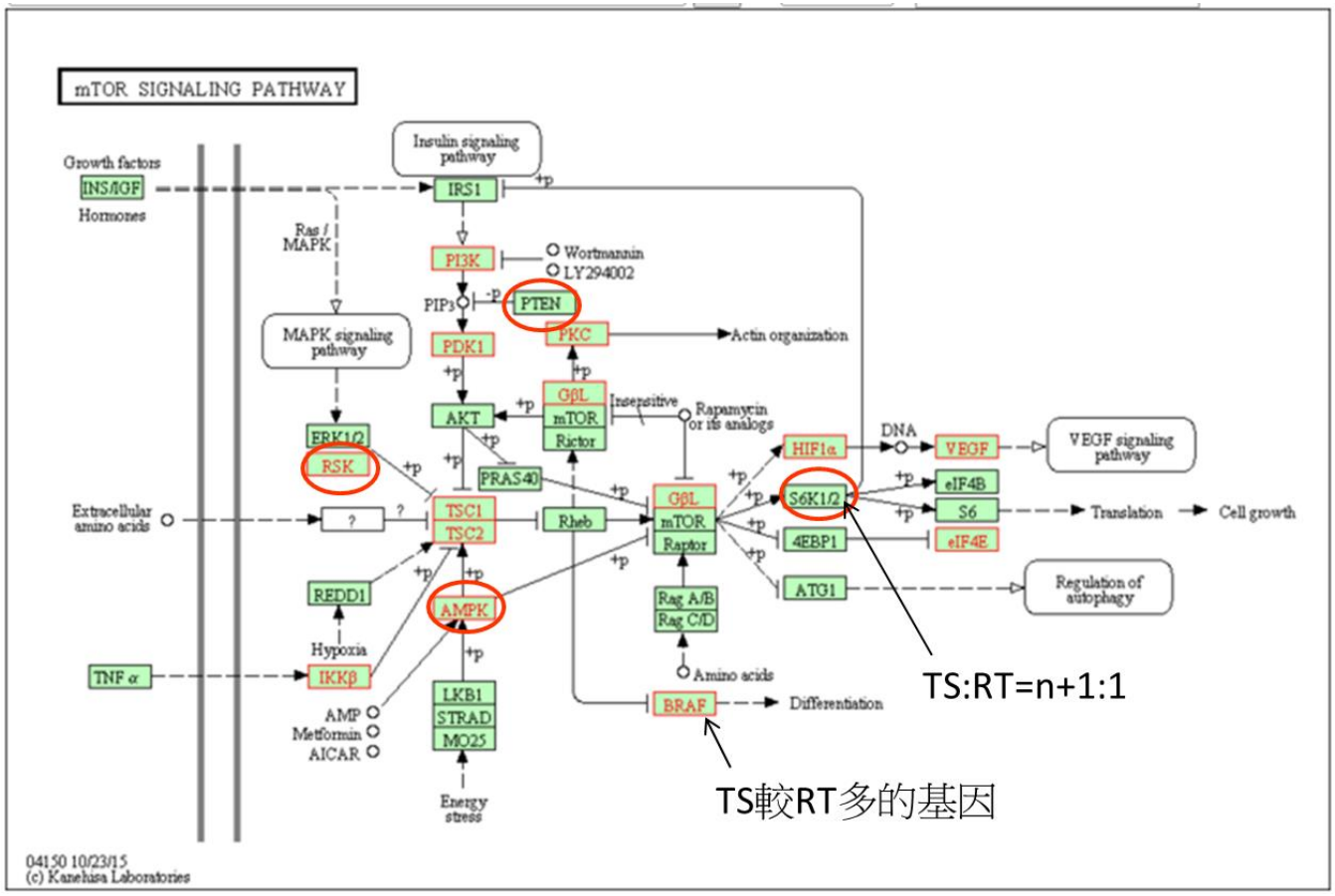
The GO (Gene ontology) of retained duplicated genes in Taiwan salmon

全部基因: 2394 種 GO 分類

GO_id	Count	GO_term
GO:0005515	6349	GO:protein binding
GO:0016021	2610	GO:integral component of membrane
GO:0016020	2534	GO:membrane
GO:0003677	1971	GO:DNA binding
GO:0005524	1912	GO:ATP binding
GO:0003676	1891	GO:nucleic acid binding
GO:0006355	1562	GO:regulation of transcription
GO:0005634	1562	GO:nucleus
GO:0008270	1510	GO:zinc ion binding
GO:0046872	1502	GO:metal ion binding
GO:0003700	1241	GO:transcription factor activity
GO:0005622	1212	GO:intracellular
GO:0007186	995	GO:G-protein coupled receptor signaling pathway

老化相關路徑的基因數量在虹鱒和臺灣鮭有不一樣嗎？

The GO of retained duplicated genes in Taiwan salmon



虹鱒(多次產卵)與臺灣鮭(產卵後死亡)的差異可能是這些基因參與所影響的嗎？也許可以藉由調控這些路徑達到臺灣鮭種魚的長期養殖(不用擔心種魚交配後死亡)

Taiwan salmon genome study

Before

Genome size: 2.7G
Sequencing data: 150G
(Coverage >50X)
Predicted genes: >20K

Aims

Sequencing data: 600G
(Coverage >200X)
Resolution: >90%
Compared to *Masu sp*.
(山女魚 and 石川鮭)

After

Taiwan salmon:
Sequencing data: 550G
(Coverage >194X)
Resolution: >90%
Predicted genes: >36K

Japan salmon:
Amago (>100G; >30X)
Yamame (>100G; >30X)

基因體定序的旅途中的停看聽

1. 先射箭再畫靶? 還是先畫靶再射箭?
   需要基因體才能解決問題?還是只是想要有基因體！
2. 需要準備多少的經費? 誰處理生物資訊部分?
   0.7Gb genome size >300,000 NTD (不含分析)
3. 新發現是否能進一步做實驗驗證推論?

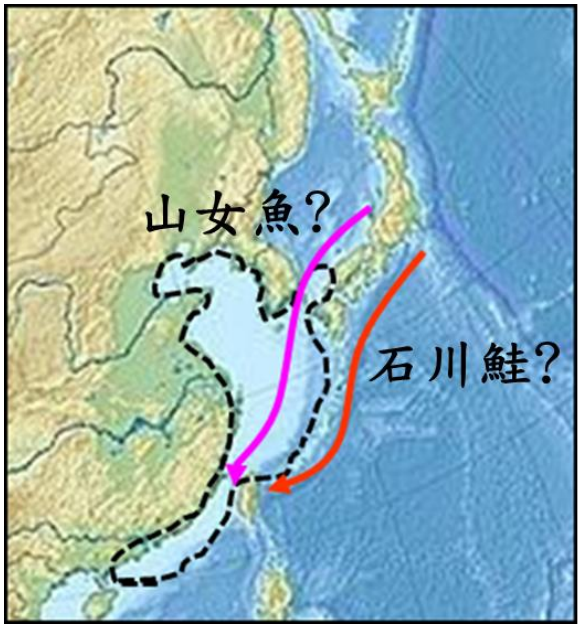
臺灣鮭魚的遺傳變異度初步分析

Species	SNP Count			
	100%	75%	50%	25%
Taiwan Salmon	439091	273297	61096	3091
Yamame	8354574	3920583	718321	135501
Amago	9065854	4072615	791114	156549

經由比較日本山女魚 (yamame, 10尾)和石川鮭 (Amago, 10尾)的SNP數量，顯示臺灣鮭 (10尾)樣本的多樣性(SNP數目)顯著地下降。**臺灣鮭是純系？**

進一步的臺灣鮭家系分析，可以經由RAD-seq的方式來建立。只是大規模之野地採樣需要林務局同意，目前尚未獲得採集許可。

台灣鮭從何而來？可能不是山女魚也不是石川鮭



Taiwan salmon genome (scaffold)

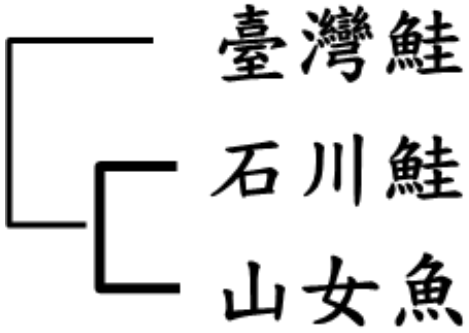
↑ **Remapping（序列重貼）**

Yamame (山女魚)
(30X coverage)

Amago (石川鮭)
(30X coverage)

Taiwan salmon (台灣鮭)
(30X coverage)

根據分析的初步結果(Ks value)顯示，臺灣鮭的基因數量與分布與兩種日本鮭的差異較大。這個結果暗示**臺灣鮭遷徙到臺灣的時間可能遠早於山女魚和石川鮭的遺傳分化**。



臺灣鮭基因體解序所解答的問題：

1. 臺灣鮭4th 基因體重組後的新基因約有半數失去功能，但仍有半數新基因獲得保留。
2. 臺灣鮭與虹鱒(演化距離8百萬年)的基因數量差異與可能扮演的角色。
3. 臺灣鮭在太平洋鮭(*O. masou sp*)中，基因數量較山女魚與石川鮭具有明顯的差異，實應獨立於兩者之外。*O. masou formosanus→O. formosanus*？
4. 臺灣鮭可能已經無法降海洄游 (適應鹽度改變)！

未來可以進一步經由實驗解答的問題：

1. 臺灣鮭失去滲透壓調控的原因。(基因功能的佚失？)
2. 臺灣鮭繁殖後死亡的調控機制。(為何虹鱒可以多次繁殖？)

Are you ready for the genome study?

Genome size

Sequencing tools

Coverage

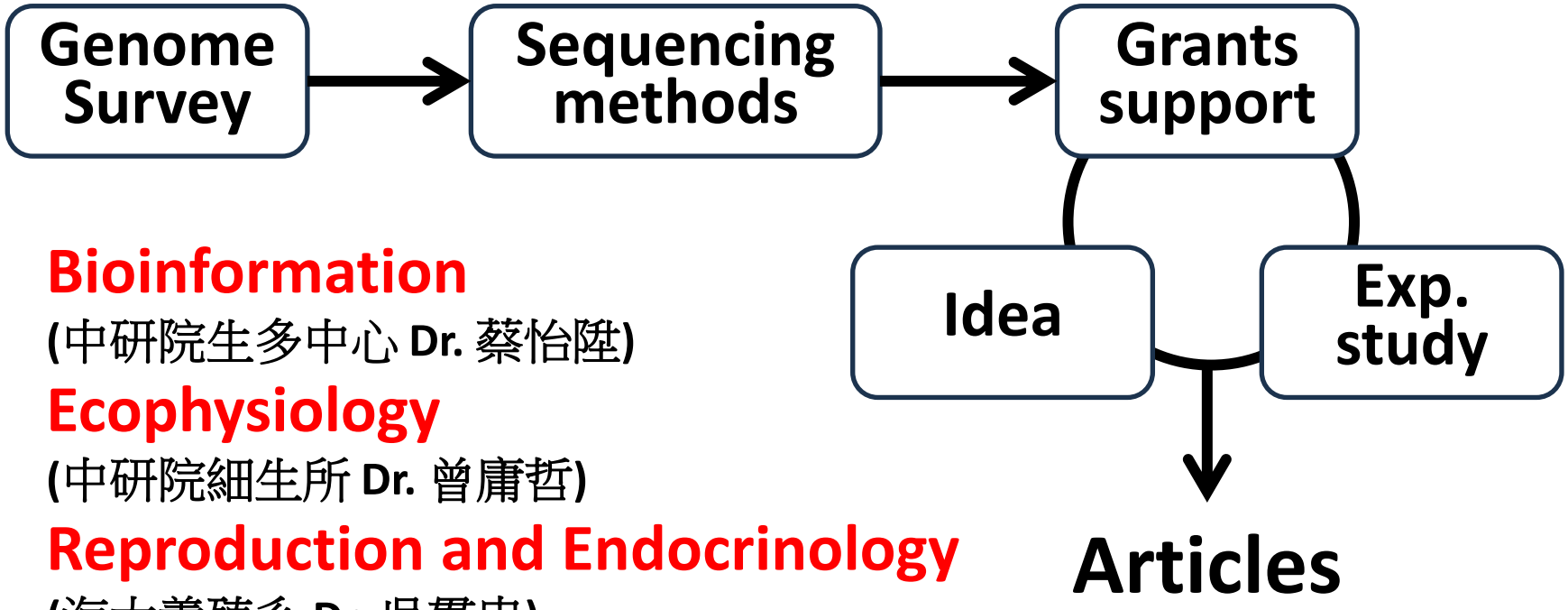
Further study

Funding

本研究團隊目前頭足類的合作模式

1. Genome size (~5.2Gb)
2. DNA size and amounts (> lambda)
3. Nanopore or PacBio (N used)

1. Team members?
2. Division of Labor



Bioinformation
(中研院生多中心 Dr. 蔡怡陞)

Ecophysiology
(中研院細生所 Dr. 曾庸哲)

Reproduction and Endocrinology
(海大養殖系 Dr. 吳貫忠)

Microbiome

Immunity