Bacterial Genome and Expression Analysis

April 20, 2012

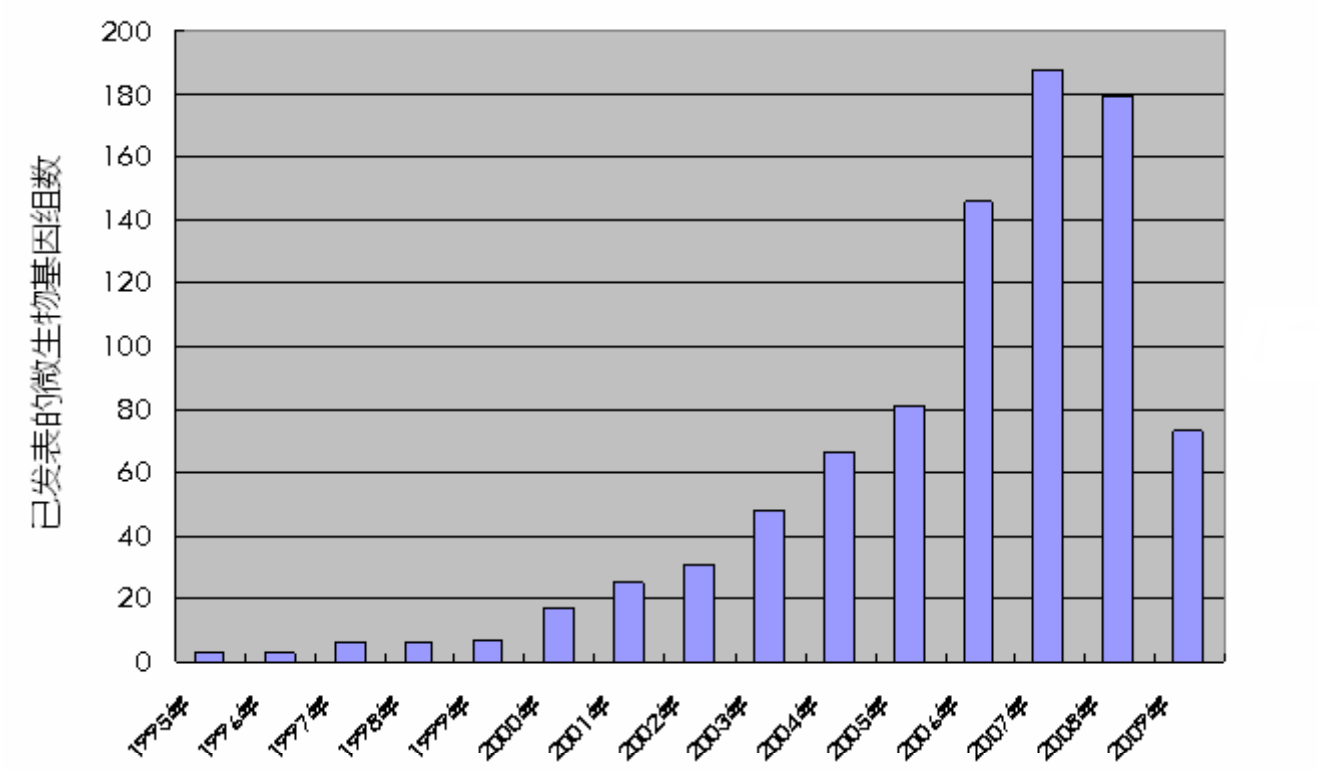
長庚大學微免科

陳怡原

Bacterial Genomes

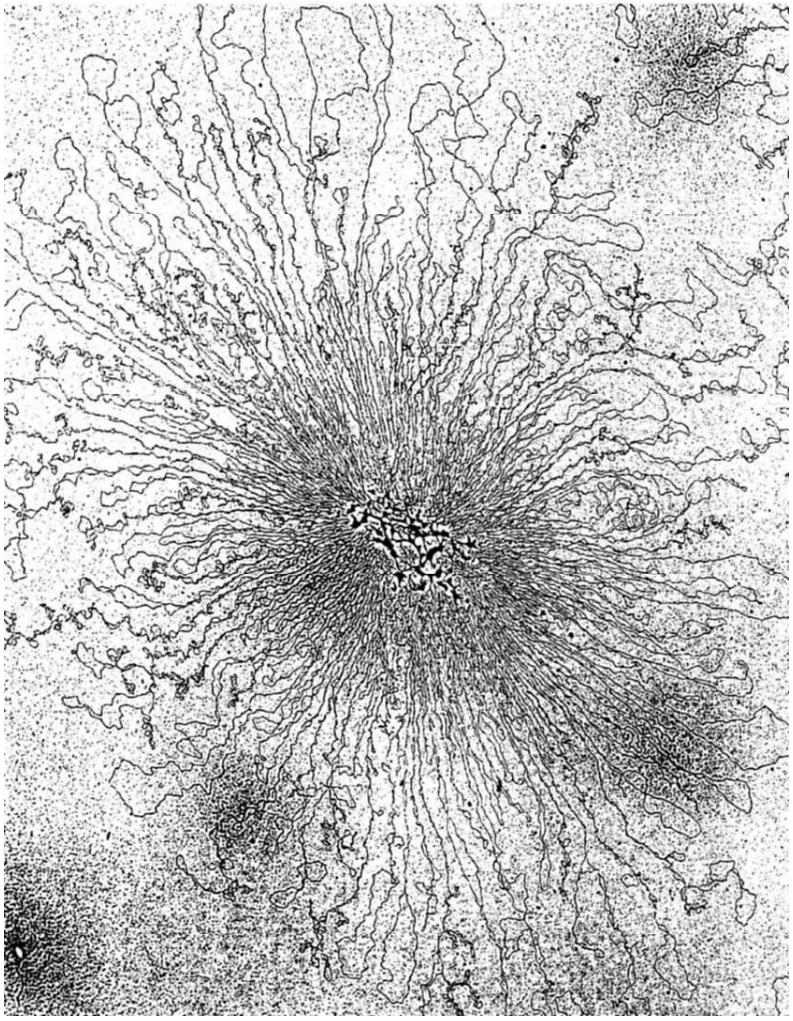
- One chromosome and possibly plasmids
- Some bacteria possess two large ‘plasmids’ as the total genome
- Circular mostly; some are in a linear form

The known bacterial genomes



Up to May 9, 2009

Structure of bacterial chromosome



- Most bacteria have one closed circular chromosome located in a distinct area called nucleoid. Chromosome usually exists as supercoiled molecule associated with basic proteins.
- Species such as Actinomycetes and spirochete *Borrelia burgdorferi* have linear chromosome. Also, *Agrobacterium tumefaciens* has one circular and one linear chromosome.

Sizes of bacterial genomes

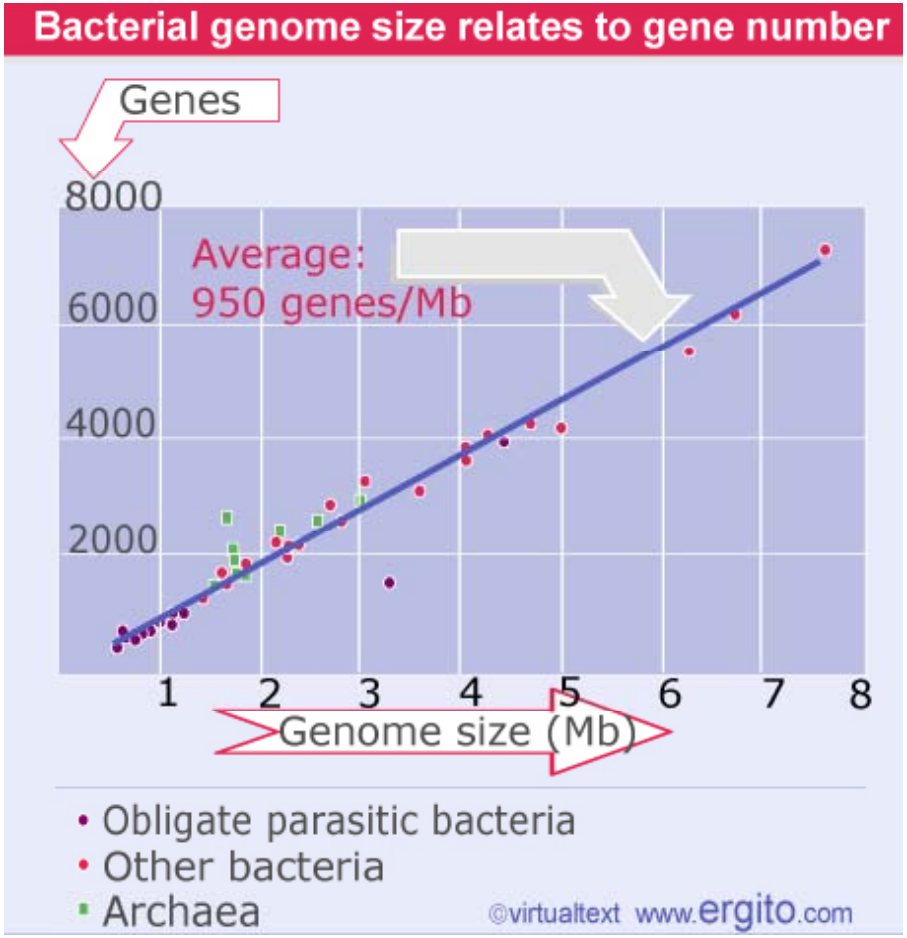
- *Mycoplasma genitalium* **has the smallest genome, approx. 470 genes**

- *S. meliloti* **and** *M. loti* **have genome sizes of approx. 7 Mb (similar to what yeasts have), both are nitrogen-fixing bacteria that live on plant roots.**

Species	Genome (Mb)	Genes	Lethal loci
Mycoplasma genitalium	0.58	470	~300
Rickettsia prowazekii	1.11	834	
Haemophilus influenzae	1.83	1,743	
Methanococcus jannaschi	1.66	1,738	
B subtilis	4.2	4,100	
E coli	4.6	4,288	1,800

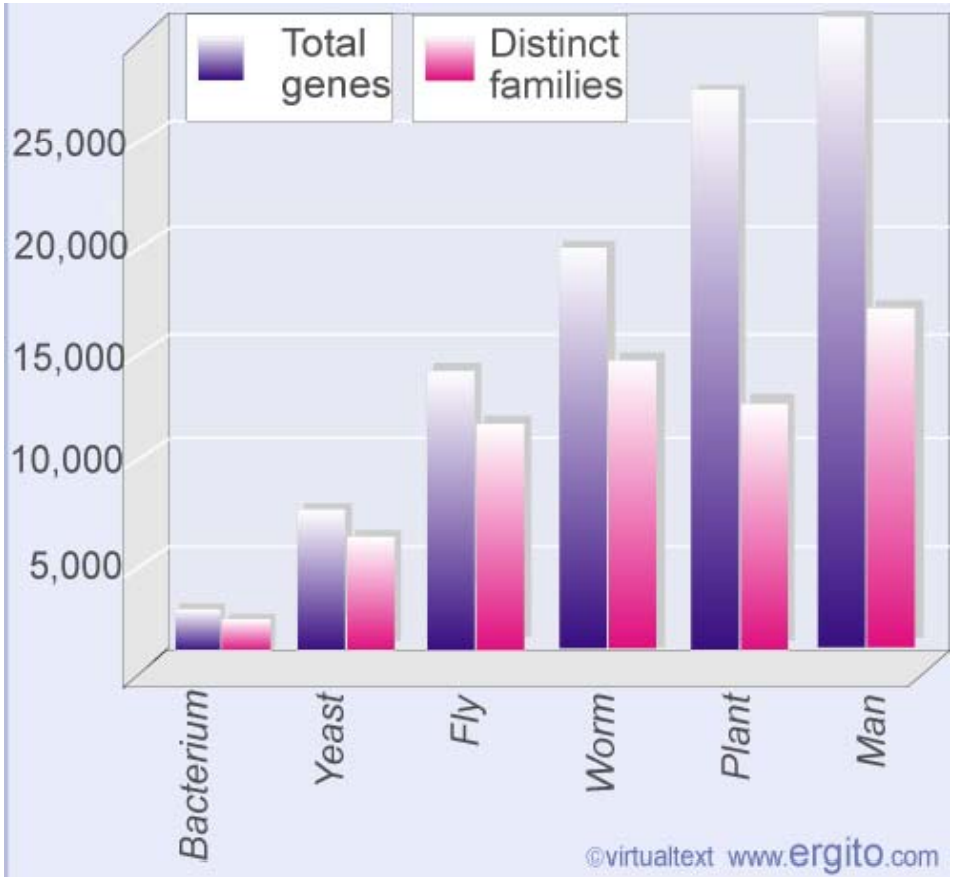
Gene numbers in microbes with different life styles

- Most of sequences of bacterial and archaeal genomes encode RNA or proteins.
- Bacteria with genome sizes below 1.5 Mb are obligate intracellular parasites.



The number of gene families plateaus with genome size

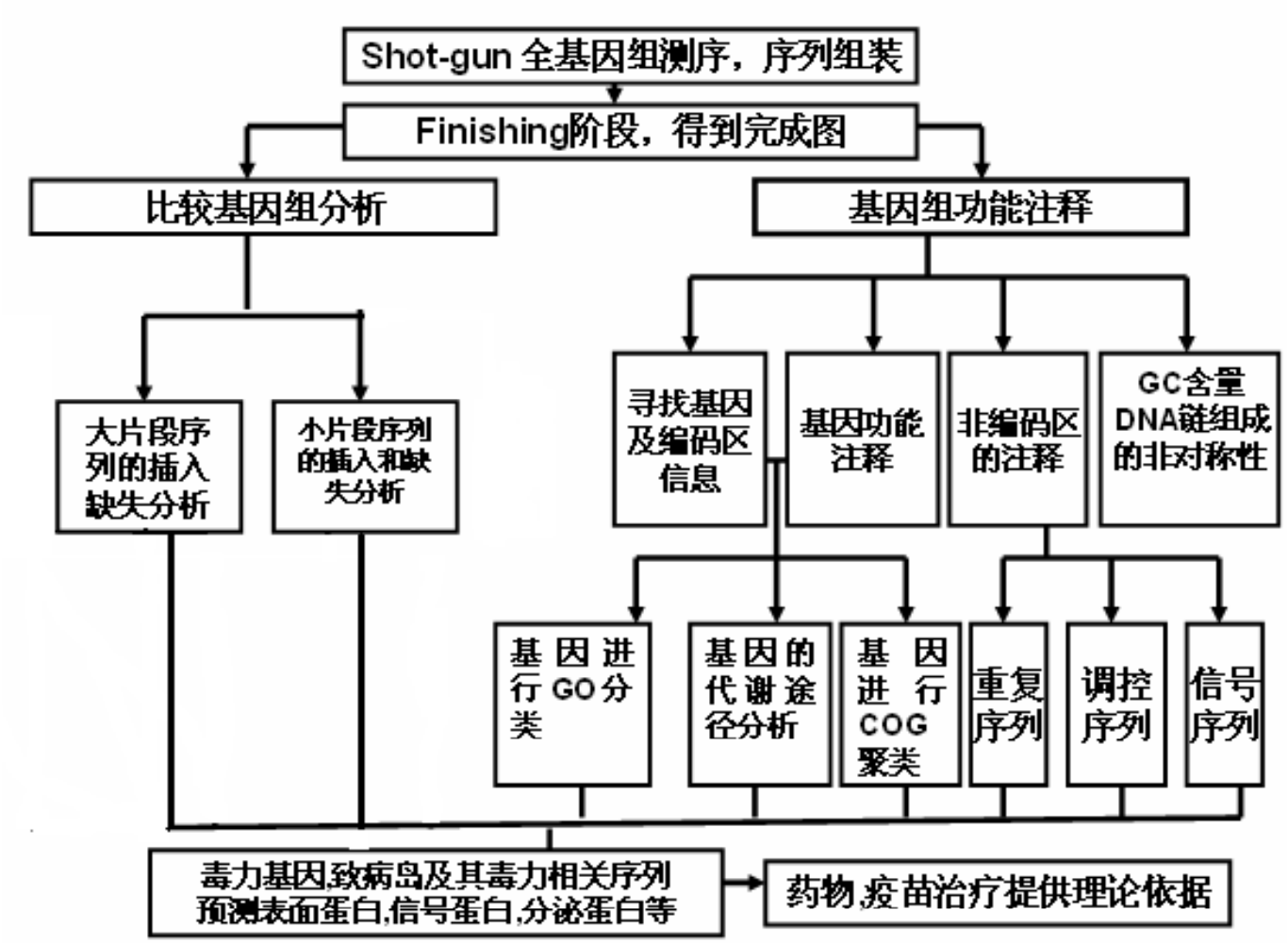
- Most of bacterial genes are unique, so the number of distinct families is closed to the total gene number.
- The number of genes increase sharply in the higher eukaryotes, but the number of gene families does not change much.



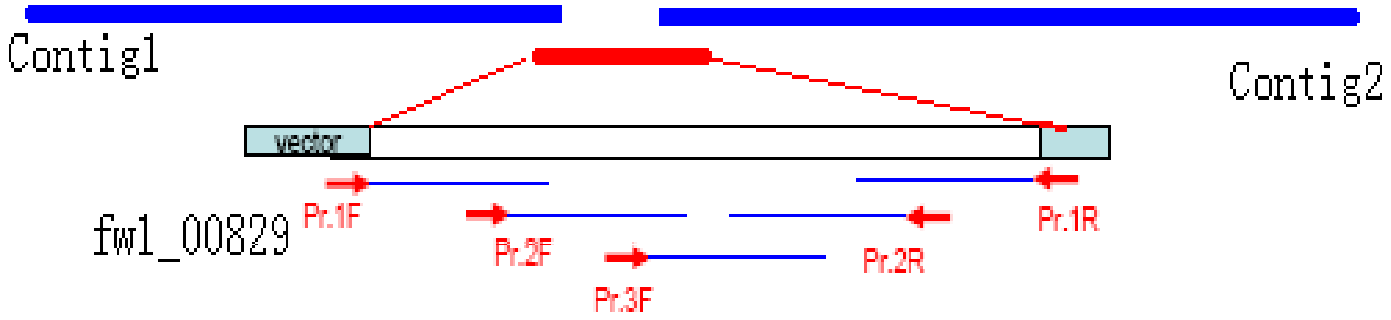
How to obtain a complete bacterial genome?

What can we learn from a genome?

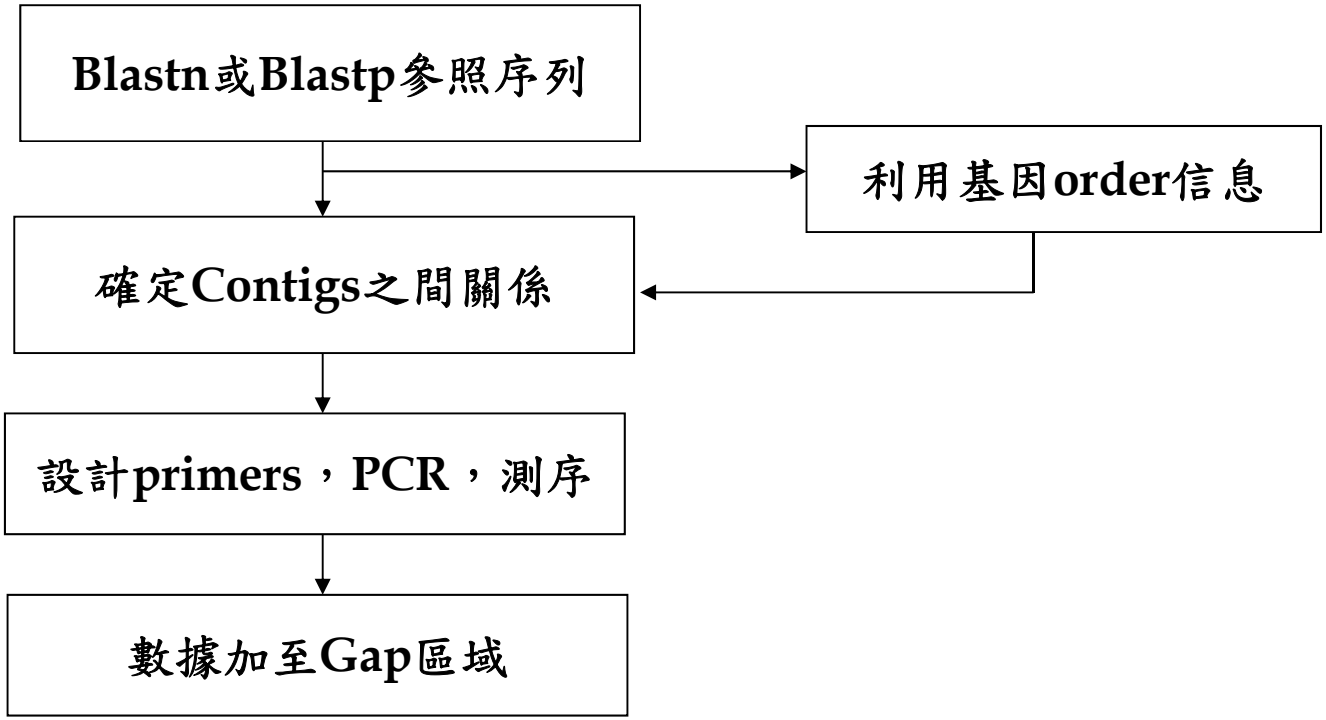
Analysis of Bacterial Genomes



To finish up a genome



To finish up a genome

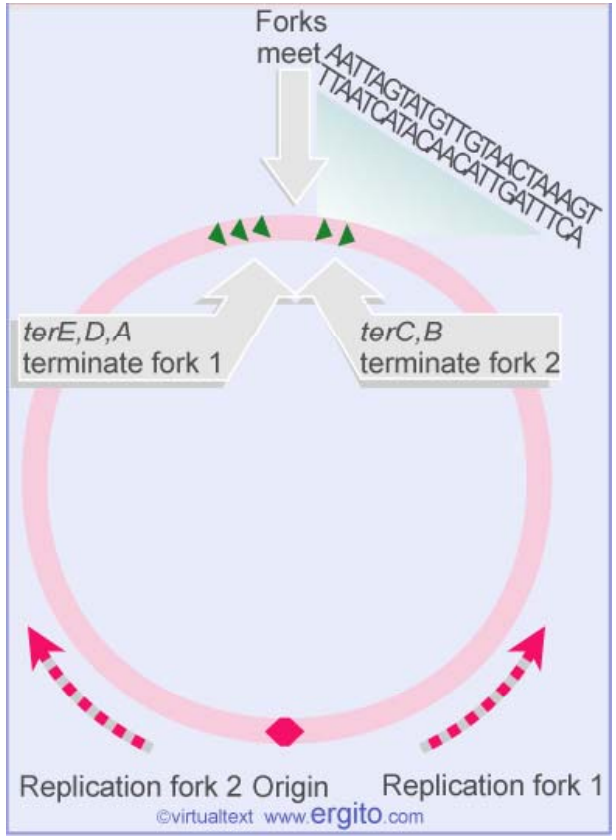


The characteristics of the bacterial genomes

	Characteristics
Structure	Double-stranded circular; linear
Sizes	0.16-13Mb
ORF	Occupy more than 90% of the total length; averagely 1 kb/ORF
GC content	16.6%-74.9%

Origin (*ori*) and termination (*ter*) of replication

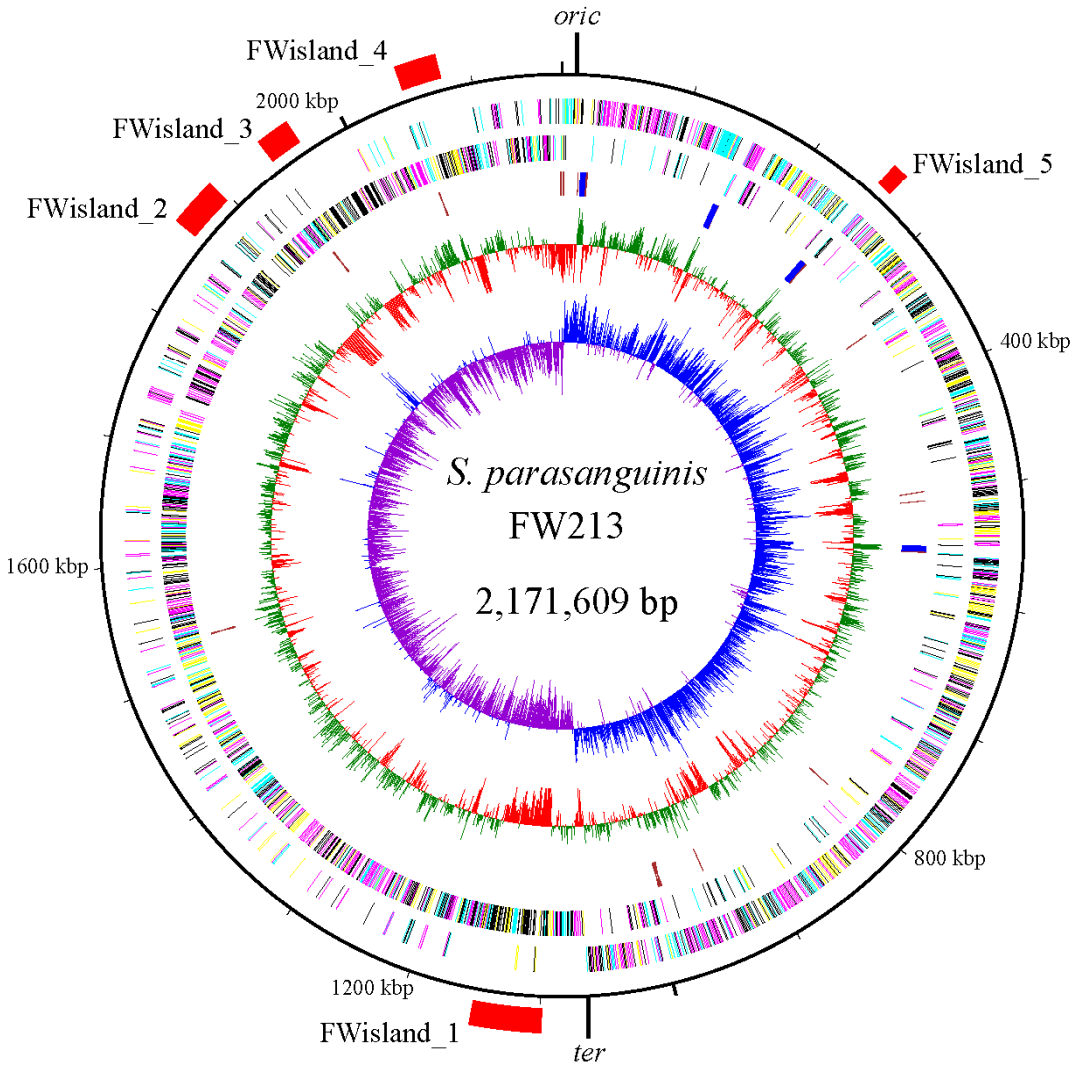
- In bacteria, chromosome is replicated bidirectionally from a single origin, called θ-type replication.
- Replicon:
  Autonomous units of replication with initiators (proteins) that act upon replication origins.
- *ter* : 23 bp asymmetrical consensus sequence for Tus protein binding; Tus-*ter* complex inhibits helicase activity and cause replication termination.



General features of *S. parasanguinis* FW213 genome

Category	Characteristics
Genome size (bp)	2,171,609
GC content (%)	41.62
Protein coding	2,019
tRNAs	61
rRNA operons	4
New transcripts	
Match with known protein	10
Pseudogene	1
Match with Rfam database	5
Putative small RNA	87
Plasmid (size in bp)	pFW213 (7078 bp)

The characteristics of *S. parasanguinis* FW213 Genome



COG Classification

COG category	
[J] Translation, ribosomal structure and biogenesis	139(7.17%)
[A] RNA processing and modification	0
[K] Transcription	113(5.83%)
[L] Replication, recombination and repair	119(6.14%)
[B] Chromatin structure and dynamics	0
[D] Cell cycle control, cell division, chromosome partitioning	19(0.98%)
[Y] Nuclear structure	0
[V] Defense mechanisms	55(2.84%)
[T] Signal transduction mechanisms	49(2.53%)
[M] Cell wall/membrane/envelope biogenesis	82(4.23%)
[N] Cell motility	5(0.26%)
[Z] Cytoskeleton	0
[W] Extracellular structures	0
[U] Intracellular trafficking, secretion, and vesicular transport	18(0.93%)
[O] Posttranslational modification, protein turnover, chaperones	53(2.73%)
[C] Energy production and conversion	47(2.42%)
[G] Carbohydrate transport and metabolism	150(7.74%)
[E] Amino acid transport and metabolism	144(7.43%)
[F] Nucleotide transport and metabolism	63(3.25%)
[H] Coenzyme transport and metabolism	32(1.65%)
[I] Lipid transport and metabolism	34(1.75%)
[P] Inorganic ion transport and metabolism	69(3.56%)
[Q]Secondary metabolites biosynthesis, transport and catabolism	12(0.62%)
[R] General function prediction only	611(31.51%)
[S] Function unknown	125(6.45%)

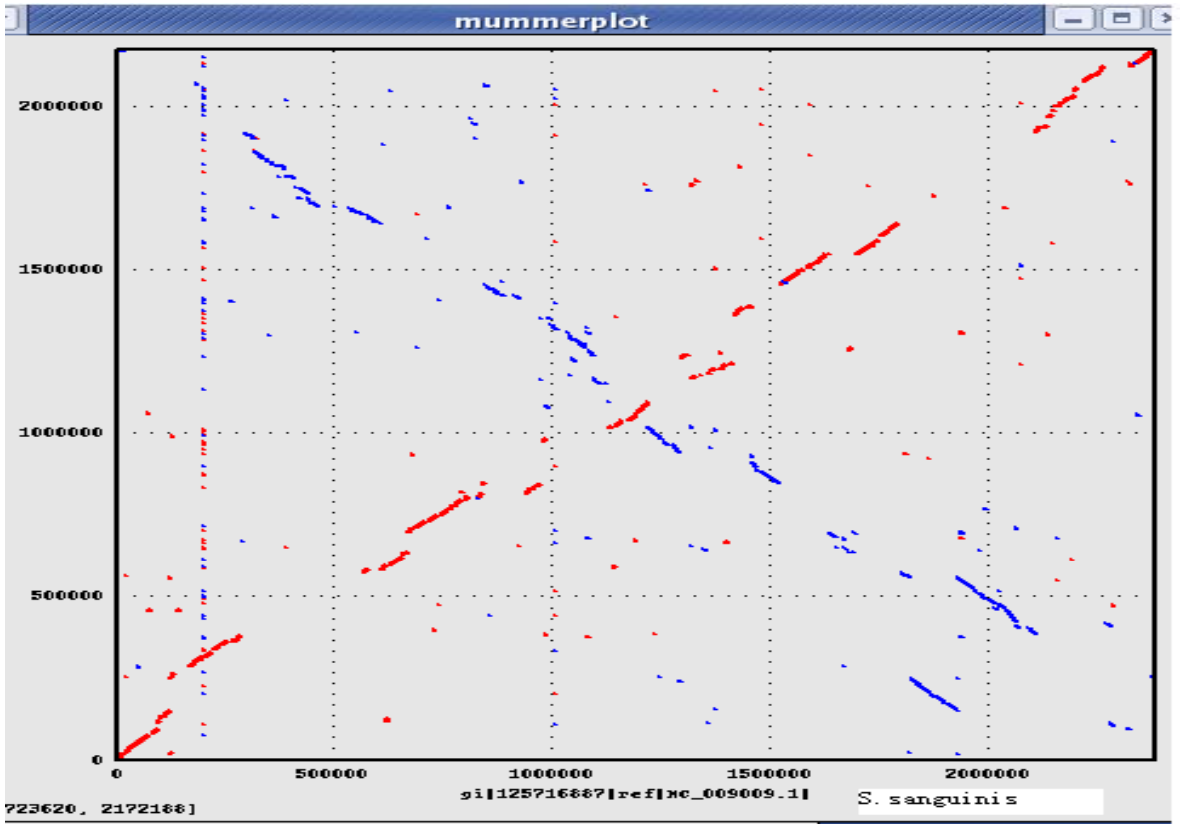
Comparison with other streptococcal species

蛋白質資料庫比對最相似基因數分佈

Species	最相似的基因數
血鏈球菌（*Streptococcus sanguinis* SK36）	557
肺炎鏈球菌（*Streptococcus pneumoniae*）	552
戈登鏈球菌（*Streptococcus gordonii* str. Challis substr. CH1）	233
嗜熱鏈球菌（*Streptococcus thermophilus*）	149
豬鏈球菌（*Streptococcus suis*）	68
變型鏈球菌（*Streptococcus mutans* UA159）	42
無乳鏈球菌（*Streptococcus agalactiae*）	31
化膿鏈球菌（*Streptococcus pyogenes*）	21

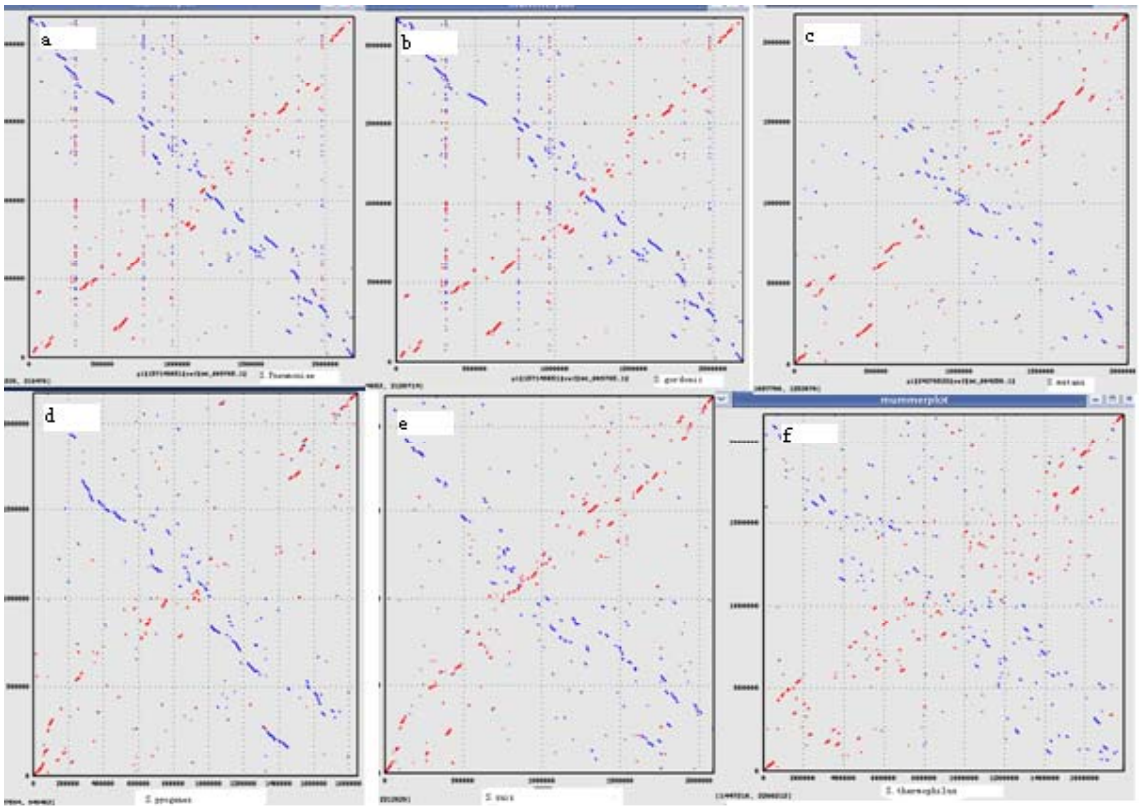
Comparison between *S. sanguinis* & *S. parasanguinis*

Streptococcus sanguinis **SK36**



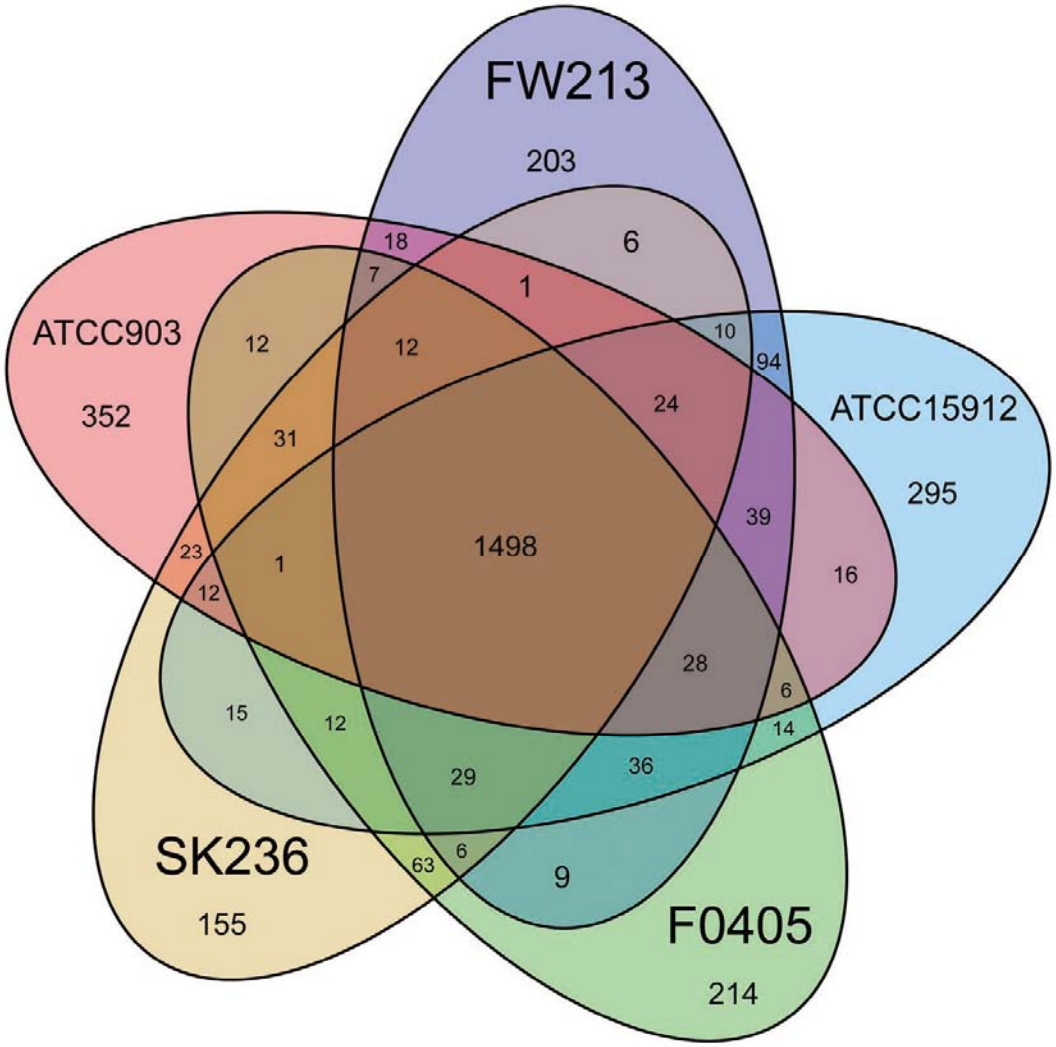
Streptococcus parasanguinis **FW213**

S. pneumoniae *S. gordonii* *S. mutans*

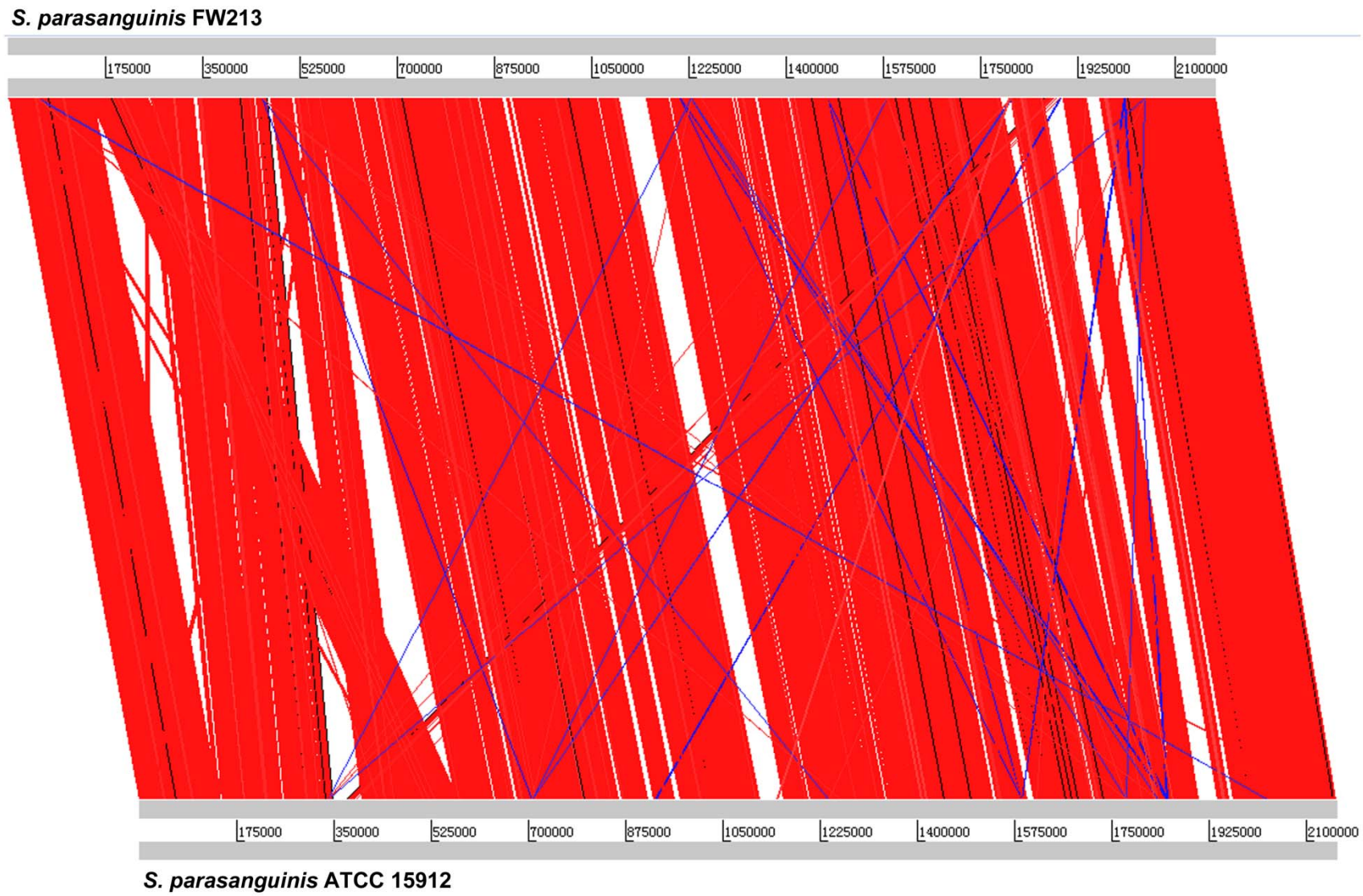


S. pyogenes *S. suis* *S. thermophilus*

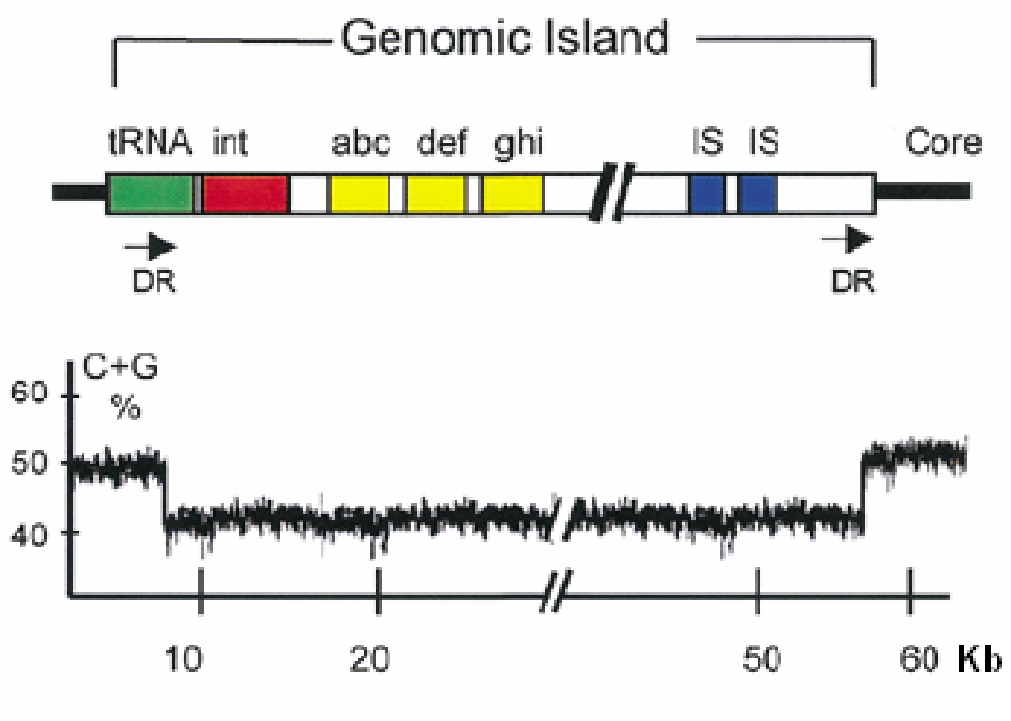
Genome comparisons of *S. parasanguinis* strains



Two *S. parasanguinis* strains



Pathogenic Islands



The average size of an island is at approximately 30Kb. The island is generally flanked with repeats and is with an unusual GC content.

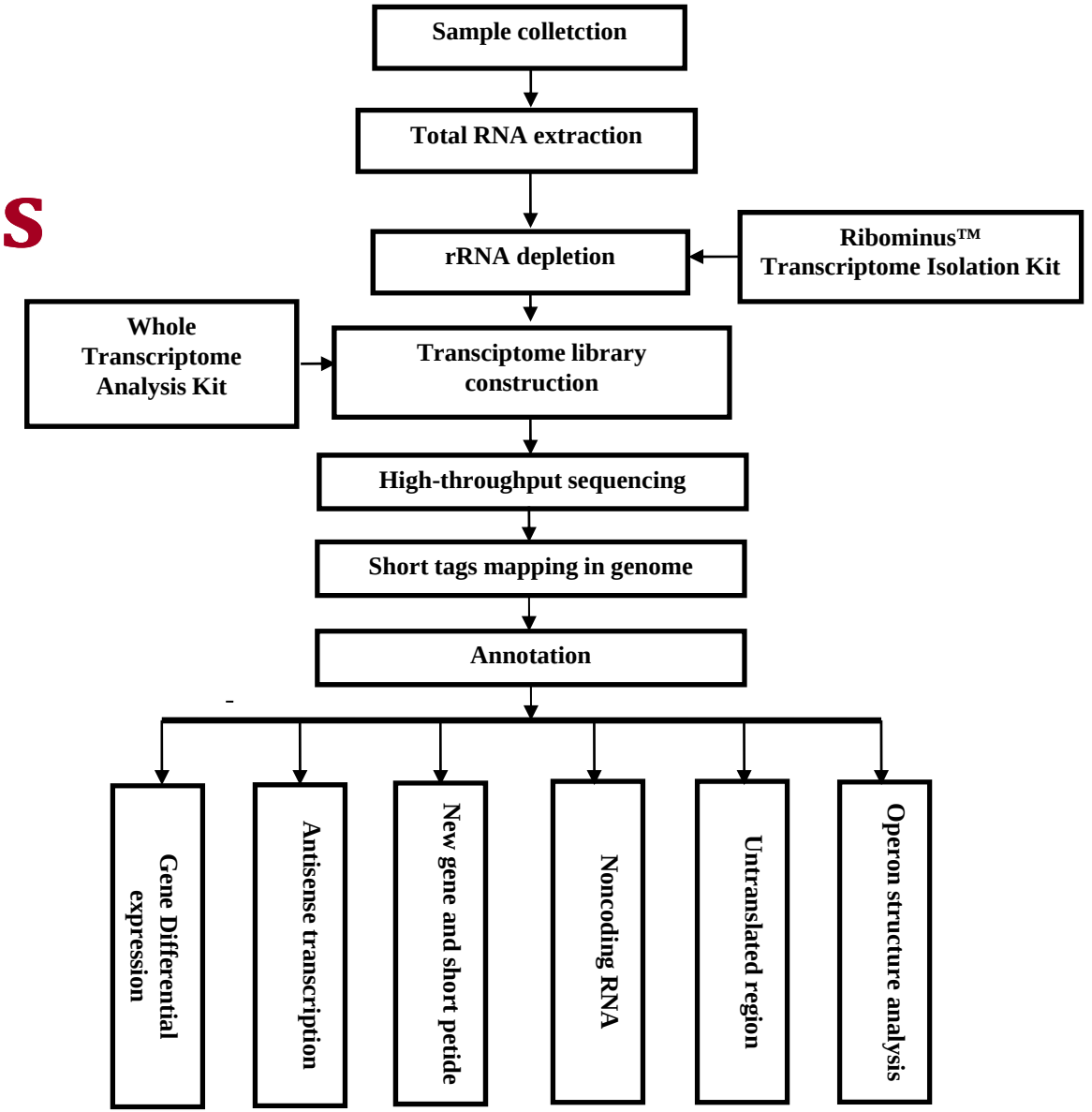
Properties of the pathogenic island in *S. parasanguinis* FW213

FWisland	Locus_tag	%GC	Putative Functions
1	Spaf_1090~ 1138	34.86	Salivaricin B and nisin production
2	Spaf_1837~ 1873	35.96	Bacterican production
3	Spaf_1923~ 1935	33.63	Fap1 production and fimbriae maturation
4	Spaf_1996~ 2016, Spaf_t58	38.20	EPS production
5	Spaf_0242~0254, Spaf_t35	36.94	adc operon

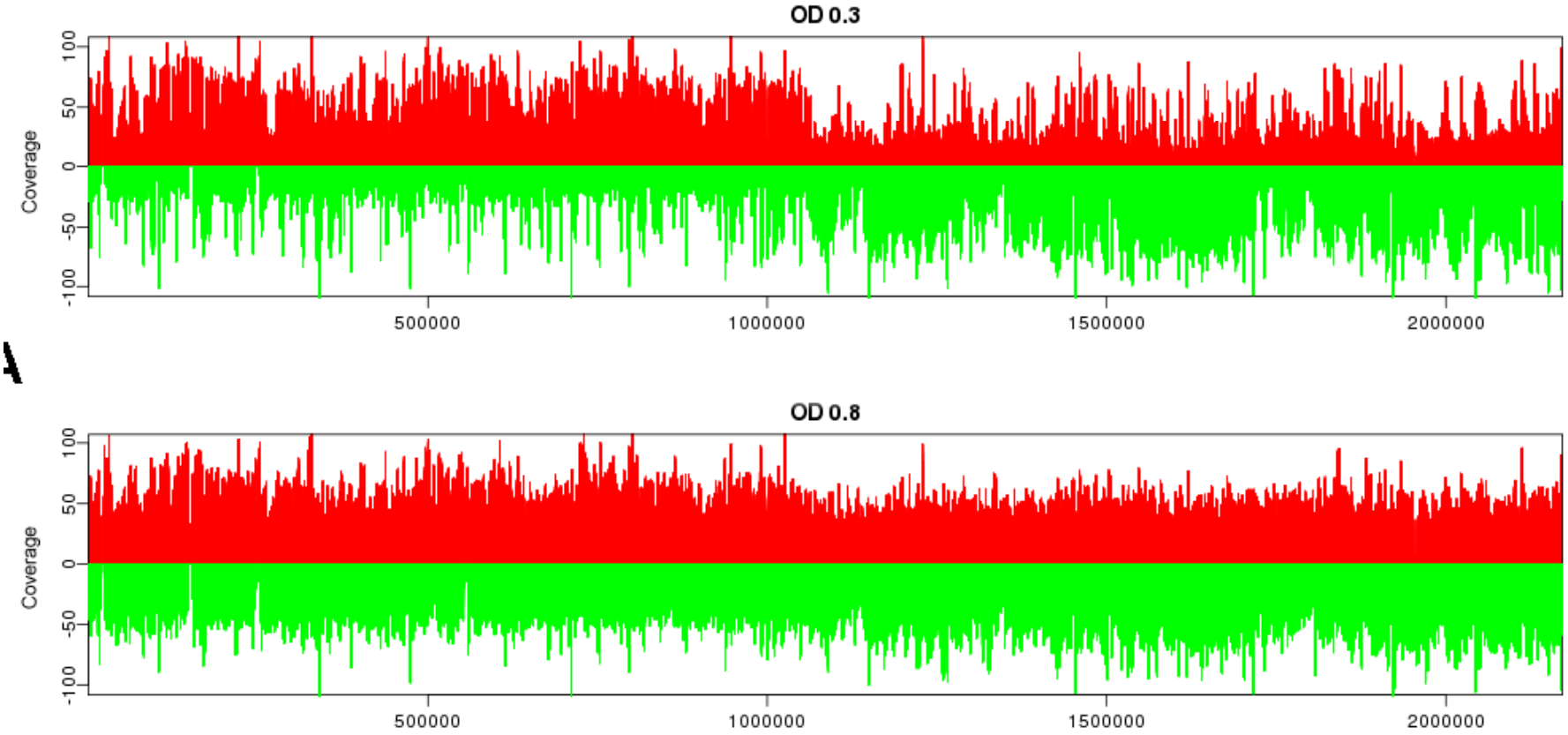
What can we learn from expression analysis by NGS?

- **Comparison of the transcriptomes of cultures from two different conditions**

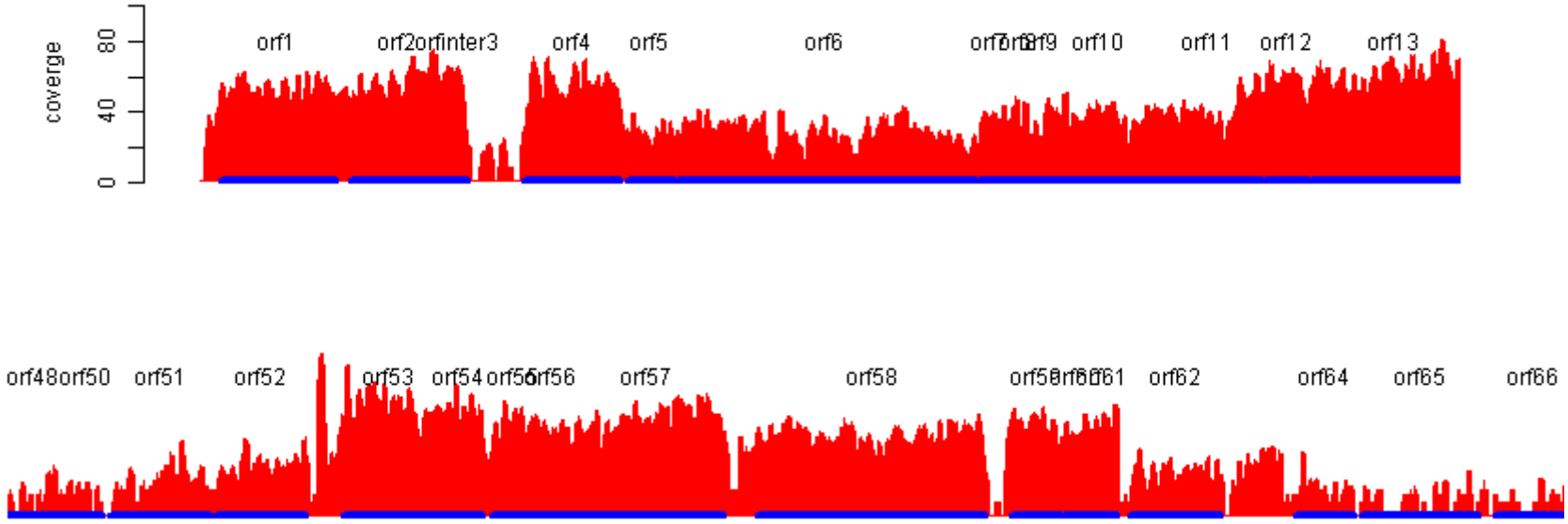
The approaches



The transcriptome of an oral streptococcus

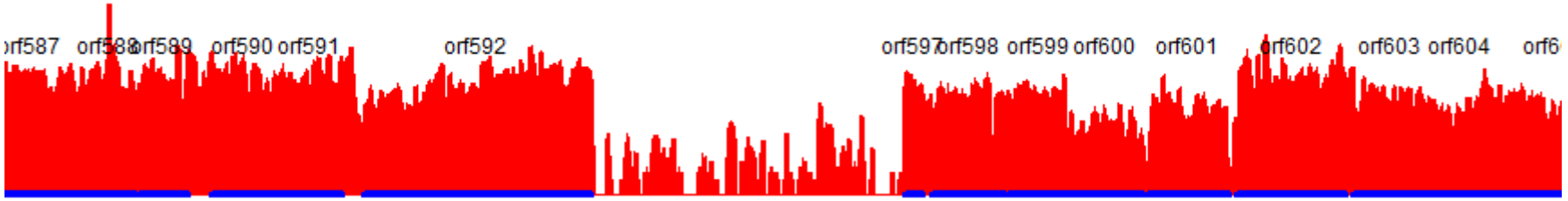


Single nucleotide mapping coverage

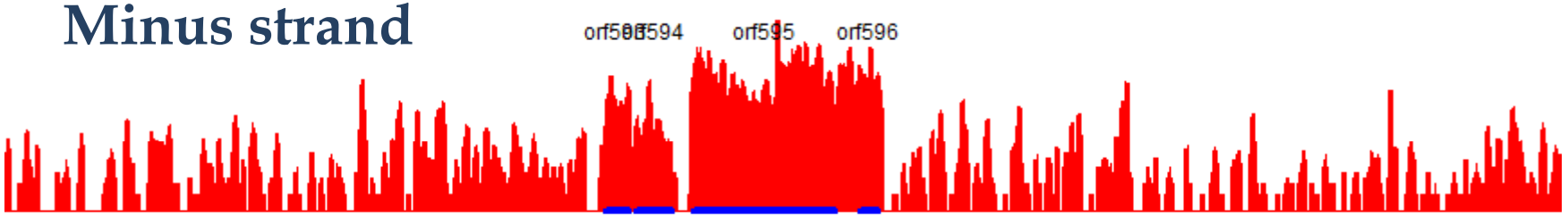


The expression of the two strands in a single region of the chromosome

Plus strand



Minus strand



The statistics of a NGS result

Map to CDs

Read	Count	Avg length	Total bases
Reads	53,140,406	49.78	2,645,221,763
Matched	9,760,186	49.74	487,471,780
Not matched	43,380,220	49.74	2,157,749,983
References	2,049	934	1,913,826

Map Unmapped to Genome

Read	Count	Avg length	Total bases
Reads	43,380,220	49.74	2,645,221,763
Matched	9,193,393	49.95	459,181,927
Not matched	34,186,827	49.68	1,698,568,056
References	1	2,171,616	2,171,616

Validation of NGS expression analysis

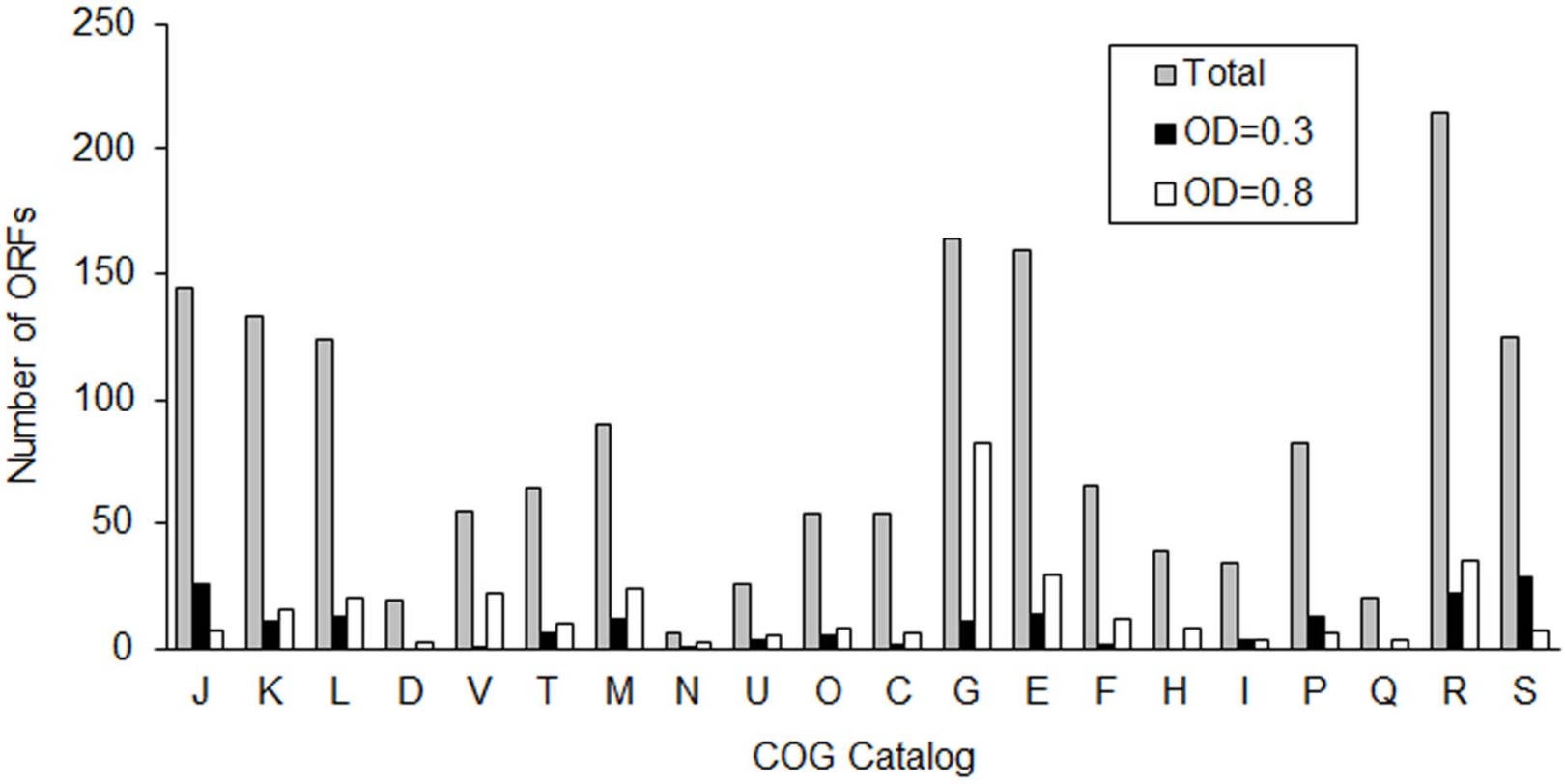
ORF	Annotation	NGS	RT-PCR
Orf1	inorganic pyrophophatase	+2.5	+3
Orf2	no nr database	+2.5	+3
Orf3	no nr database	0	Yes/No
Orf4	Mob	+0.5/0	+1
Orf5	Rep	+1.5	+2
Orf6	Hypothetic protein	+1.5	+2
Orf7	RelE	+1.25	+0.5
Orf8	RelB	+1.25	+0.5

Validation of NGS analysis

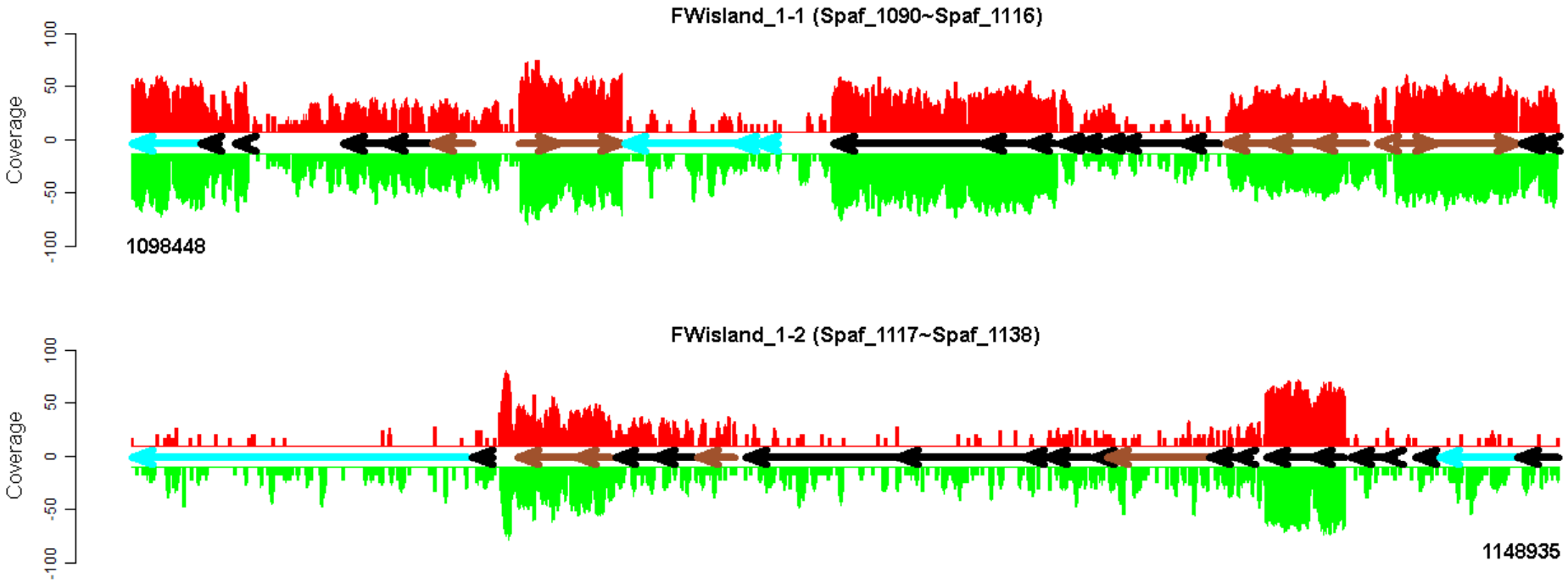
ORF	strand	gene length	fw3tag number	fw3rpkm	fw8tag number	fw8rpkm	fw3/fw8 rpkm	fw3/fw8 PT-PCR
orf1037	+	1008	13123	1126.436	1927	181.0959	6.220105	3.306
orf967	-	642	14826	1998.125	2259	333.3256	5.994515	1.6962
orf1580	-	570	26321	3995.409	4252	706.6524	5.653995	2.0825
orf1705	+	1050	3861	318.159	687	61.98047	5.133214	6.1005
orf1466	-	279	5560	1724.266	991	336.4782	5.124452	2.1842

ORF	strand	gene length	fw3tag number	fw3rpkm	fw8tag number	fw8rpkm	fw3/fw8 rpkm	fw3/fw8 PT-PCR
orf48	+	825	46	4.824337	1669	191.6416	0.025174	0.2499
orf58	+	2649	4412	144.1077	34467	1232.563	0.116917	0.4841
orf309	+	435	537	106.8117	20904	4552.266	0.023463	0.8093
orf439	+	1665	89	4.624977	3039	172.9036	0.026749	0.3561

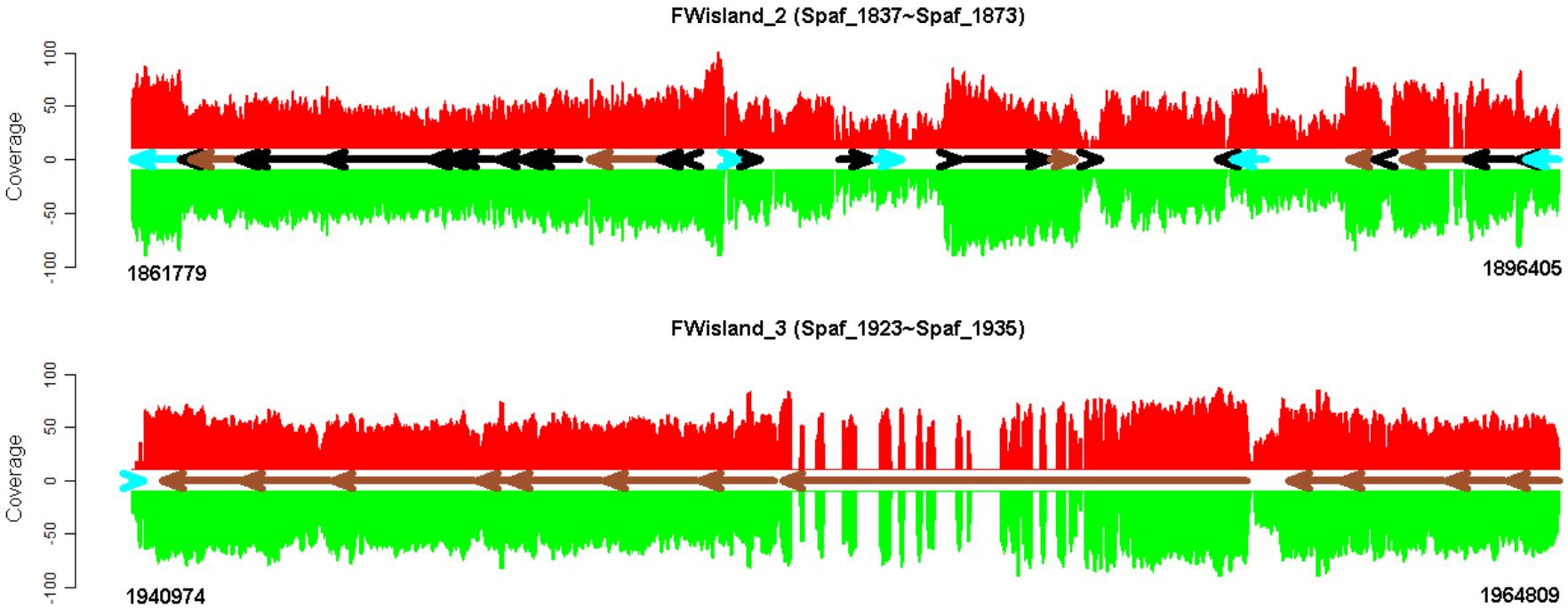
Gene expression analysis generated by RNA-seq



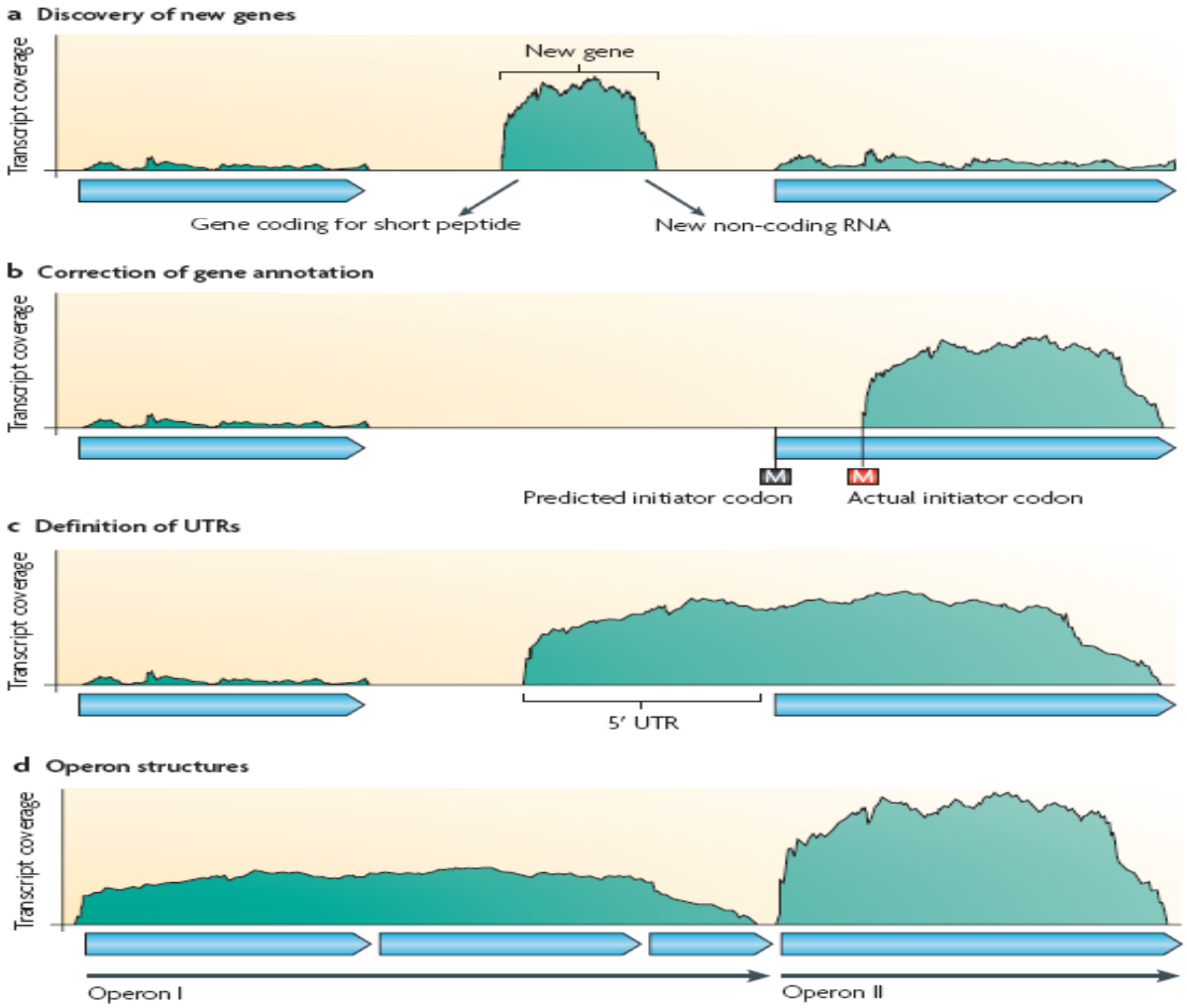
The genomic island and its transcription



The genomic island and its transcription

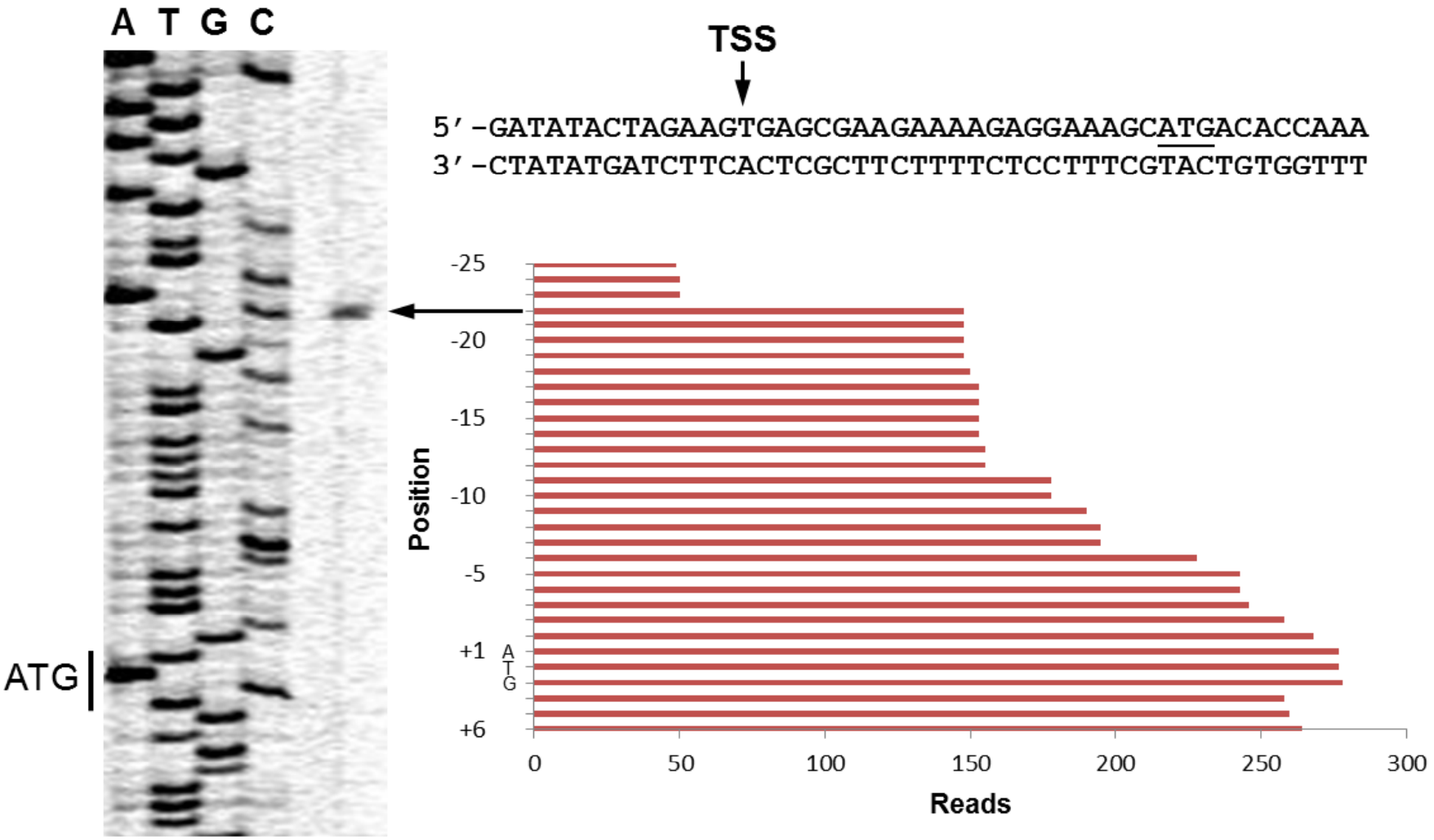


Other possibilities



Nature reviews. Genetics, Vol. 11, No. 1

The transcription initiation site of Spaf_0344

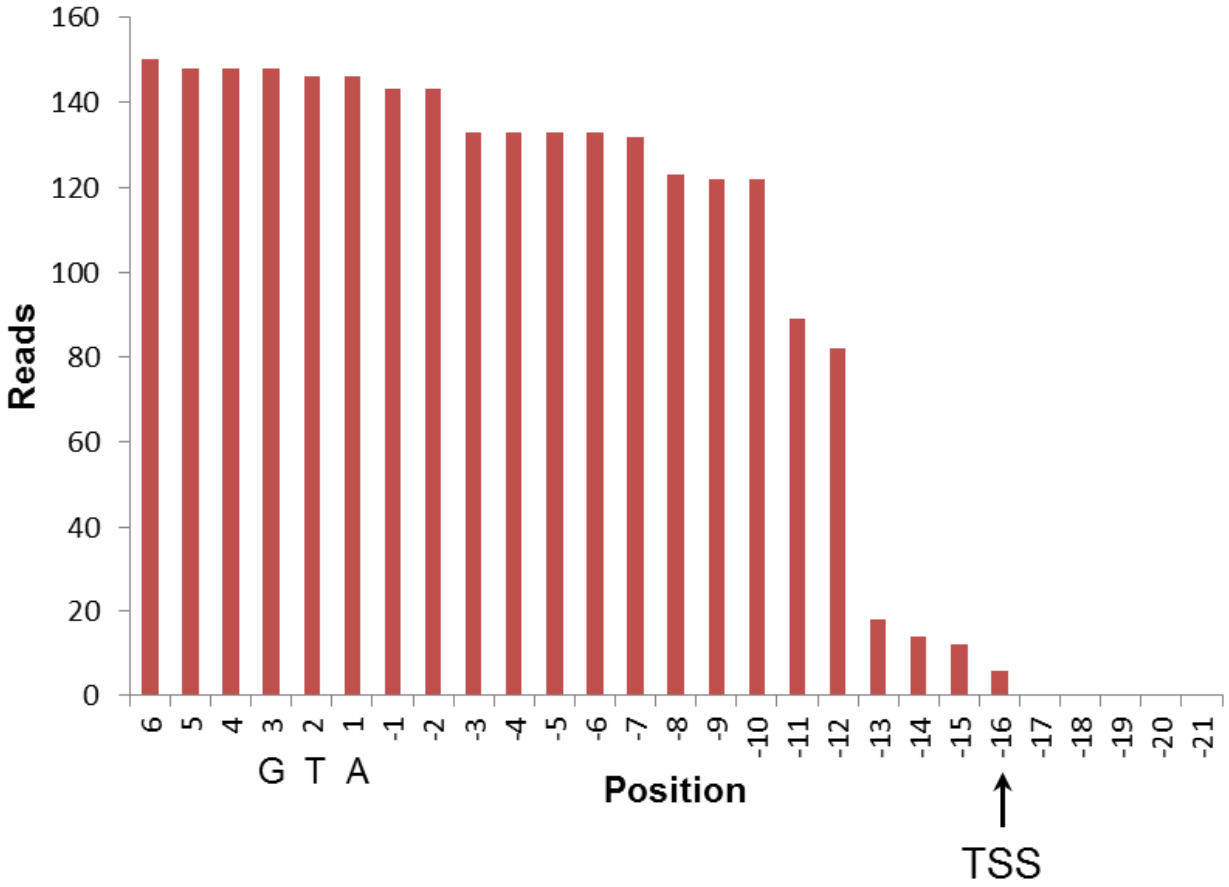


Another example

TSS

5'- GACTTAATTTTTTTATTTATGTATATTAATAGTAAAGGAGCAAACTCATG

3'- CTGAATTAAAAAAATAAATACATATAATTATCATTTCCTCGTTTGAGTAC

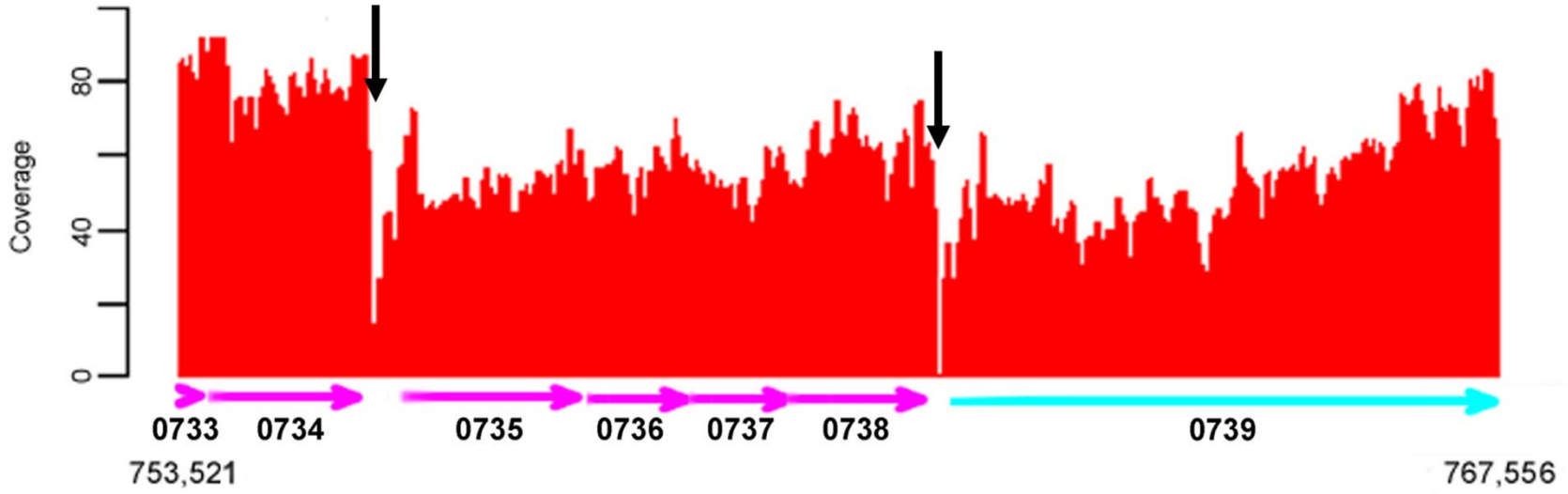


An operon

- An operon is a unit of bacterial gene expression and regulation, including structural genes and control elements in DNA recognized by regulator gene product(s).

- Structural gene clusters are coordinately controlled.

Operon determination

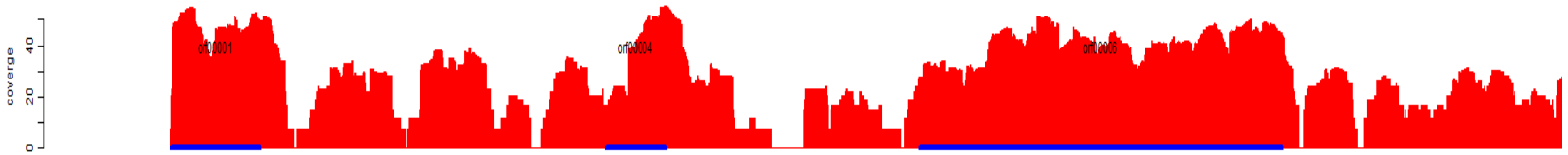


What else?

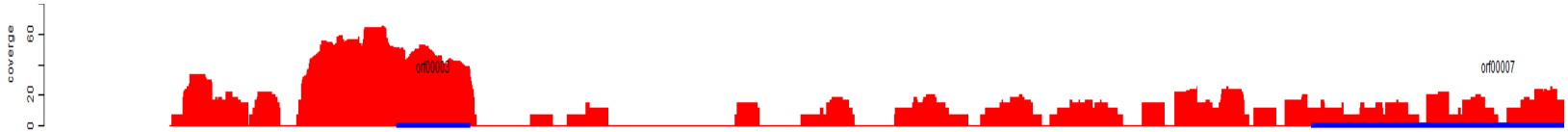
- Finding novel small RNA
- Potential small proteins

An example:
Expression analysis of pFW213

Plus strand



Minus strand



Conclusions

- **Transcriptomes analyzed by RNA-seq analysis can be used for refining annotation, defining operon structures, predicting transcription initiation sties and potentially identifying sRNA**

- **Comparative transcriptomic analysis requires adequate normalization and probably multiple samples/runs.**