Bacteria Genome Analysis

April 22, 2011

陳怡原

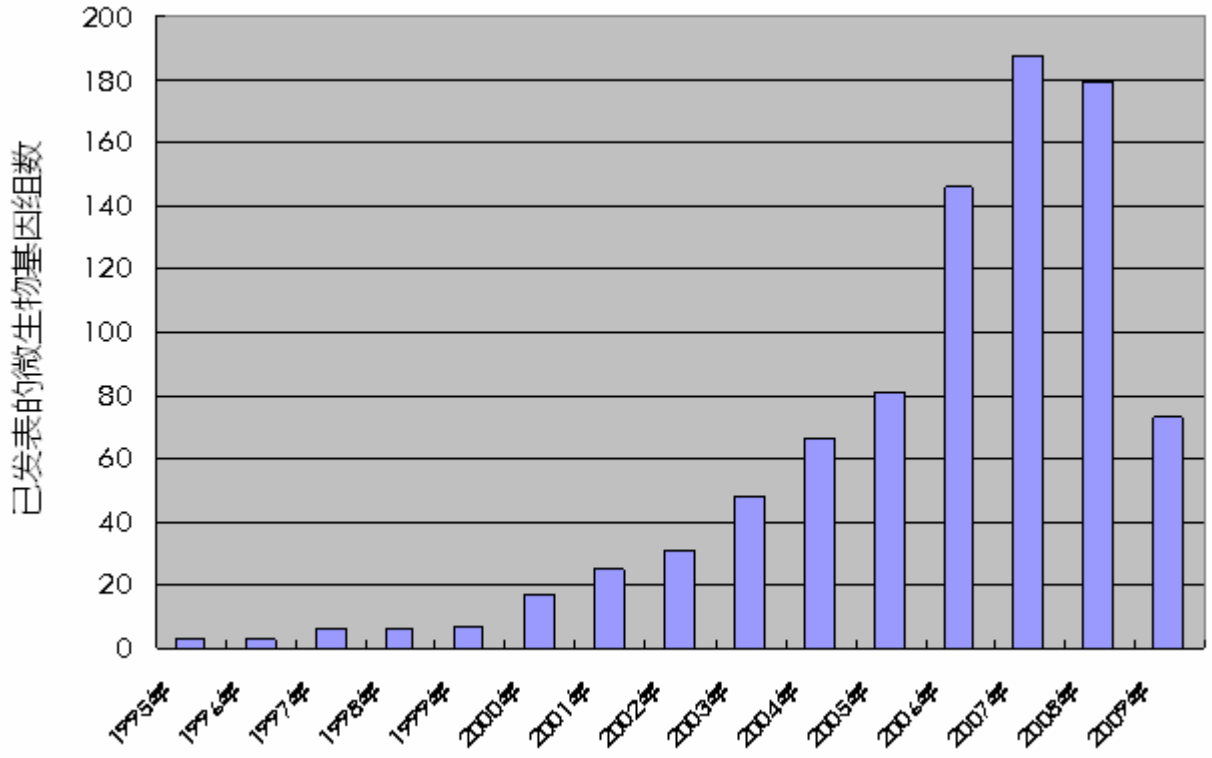
Bacterial Genomes

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

Plasmids

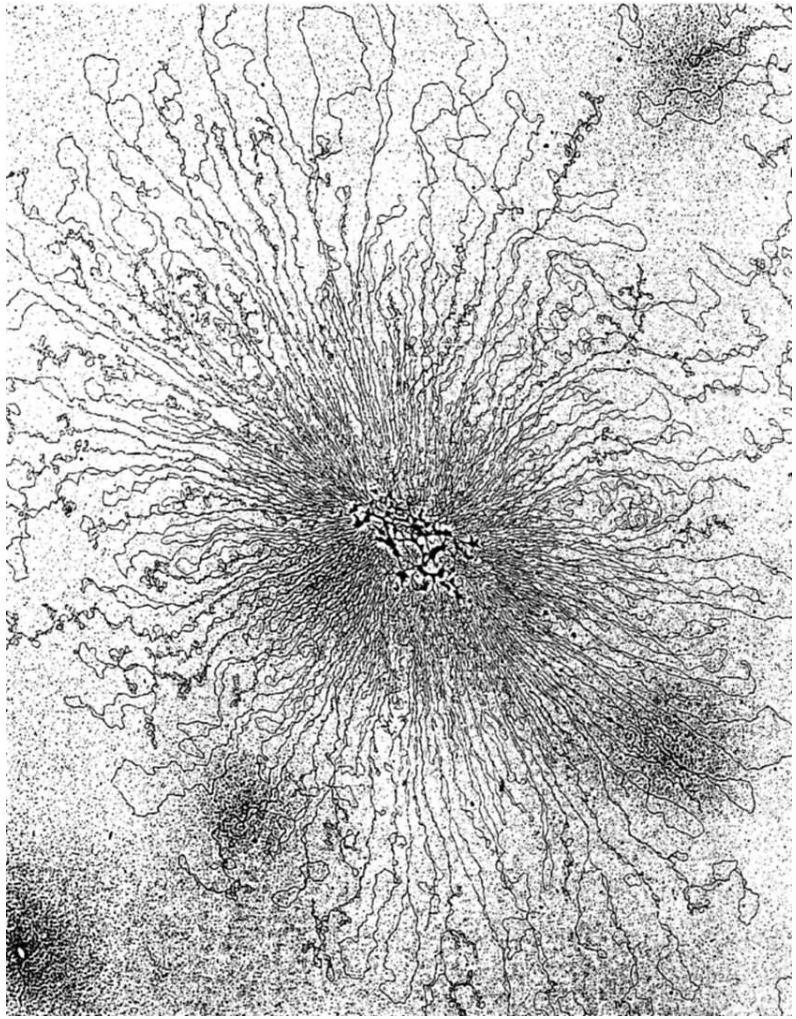
- Extrachromosomal genetic elements that can be stably inherited (autonomously replicating genetic entities)
- Variable in sizes and copy numbers
- Linear and circular
- Double- or single-stranded DNA
- Conjugative or not
- Generally do not encode functions essential to bacterial growth
- Important in bacterial evolution

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 genome

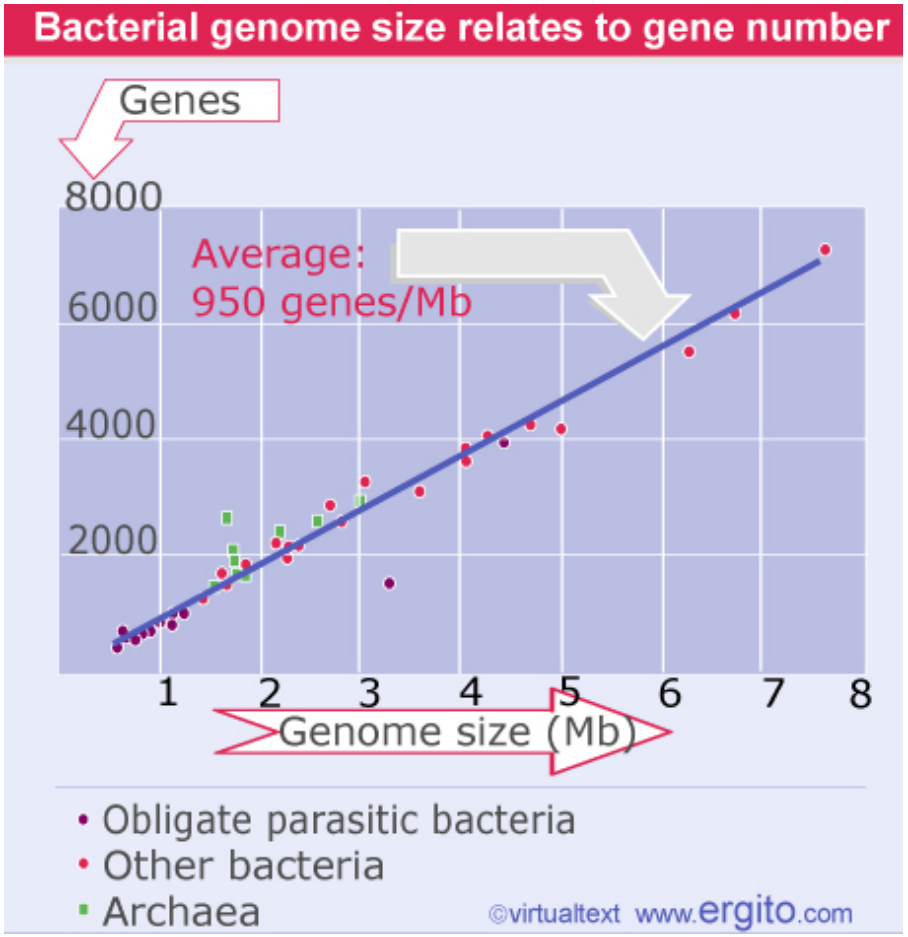
- ***Mycoplasma genitalium*** **has the smallest genome, approx. 470 genes**

- ***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

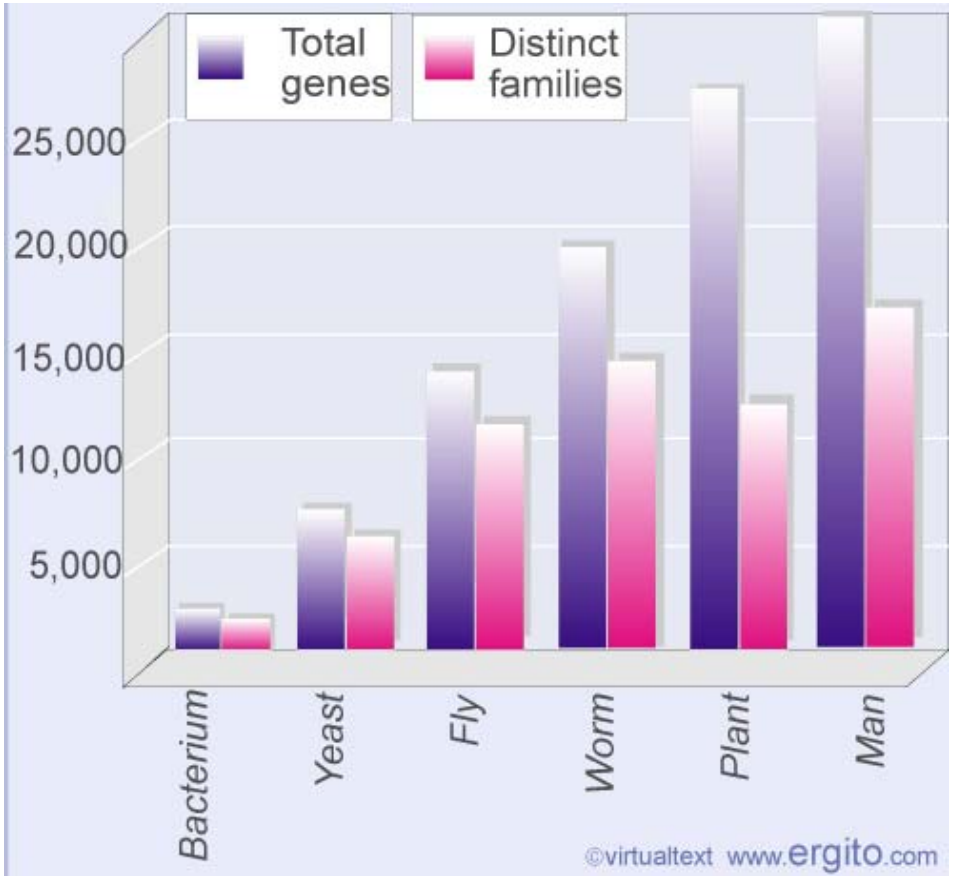
Gene numbers in microbes with different life styles

- **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

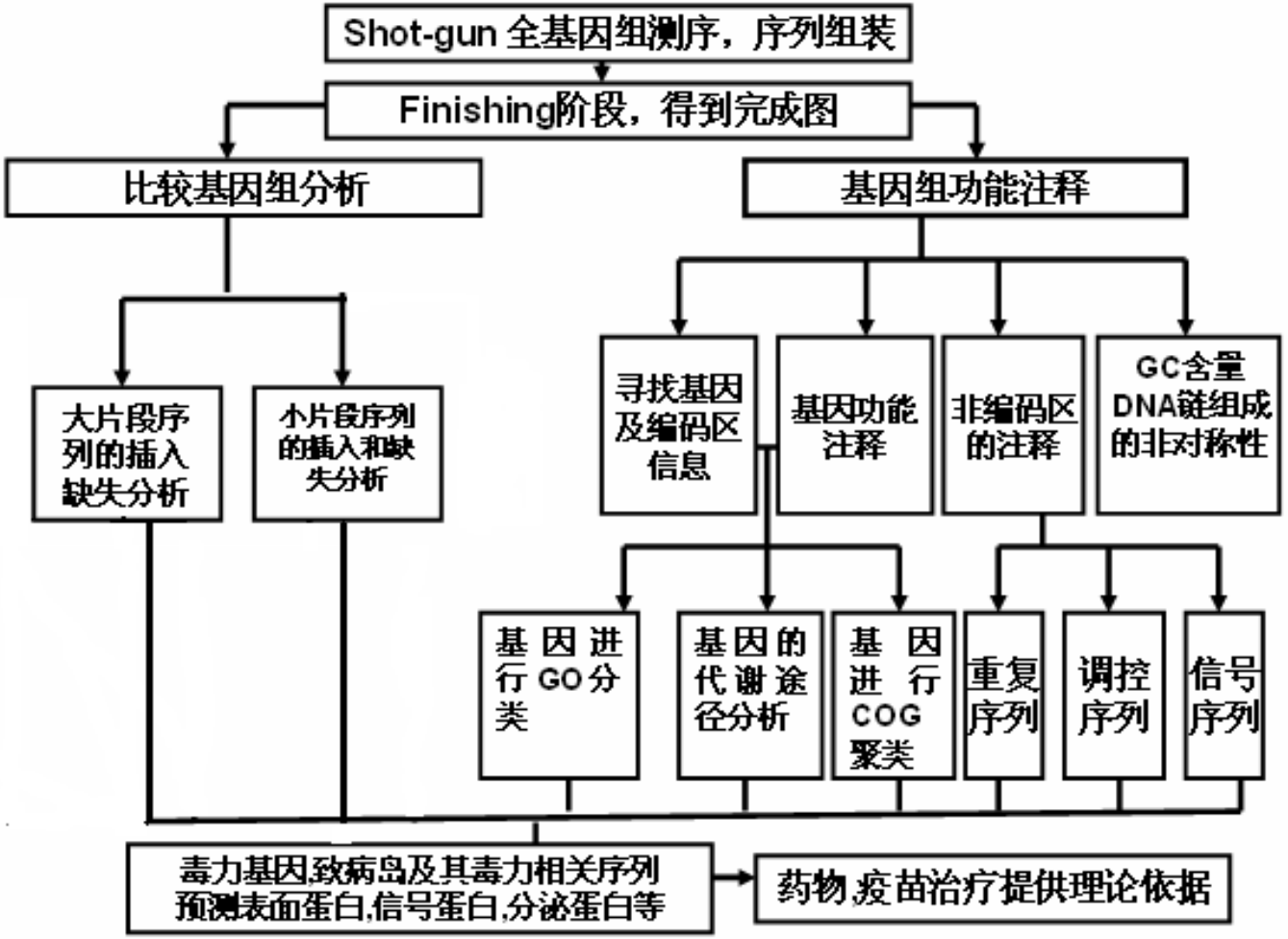
- **Most of bacterial genes are unique, so the number of distinct families is closed to the total gene number.**
- **The number of genes increase sharply in the higher eukaryotes, but the number of gene families does not change much.**



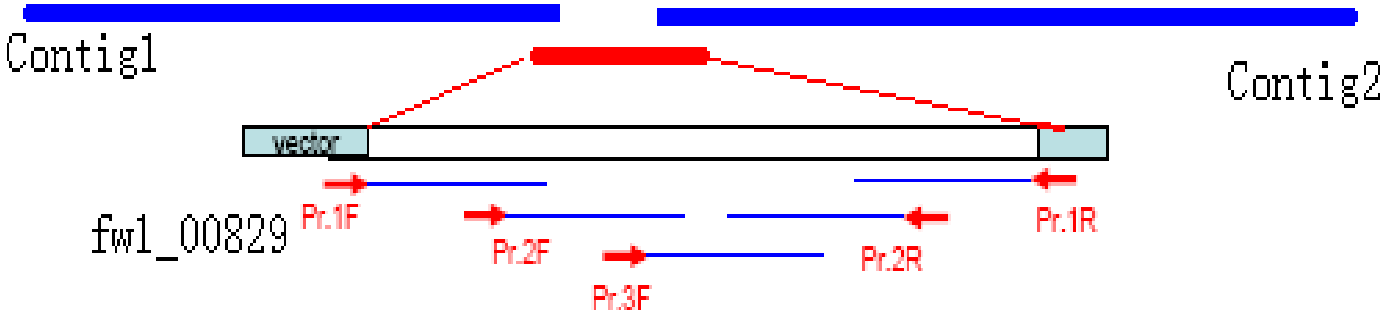
How to obtain a complete bacterial genome?

What can we learn from a genome?

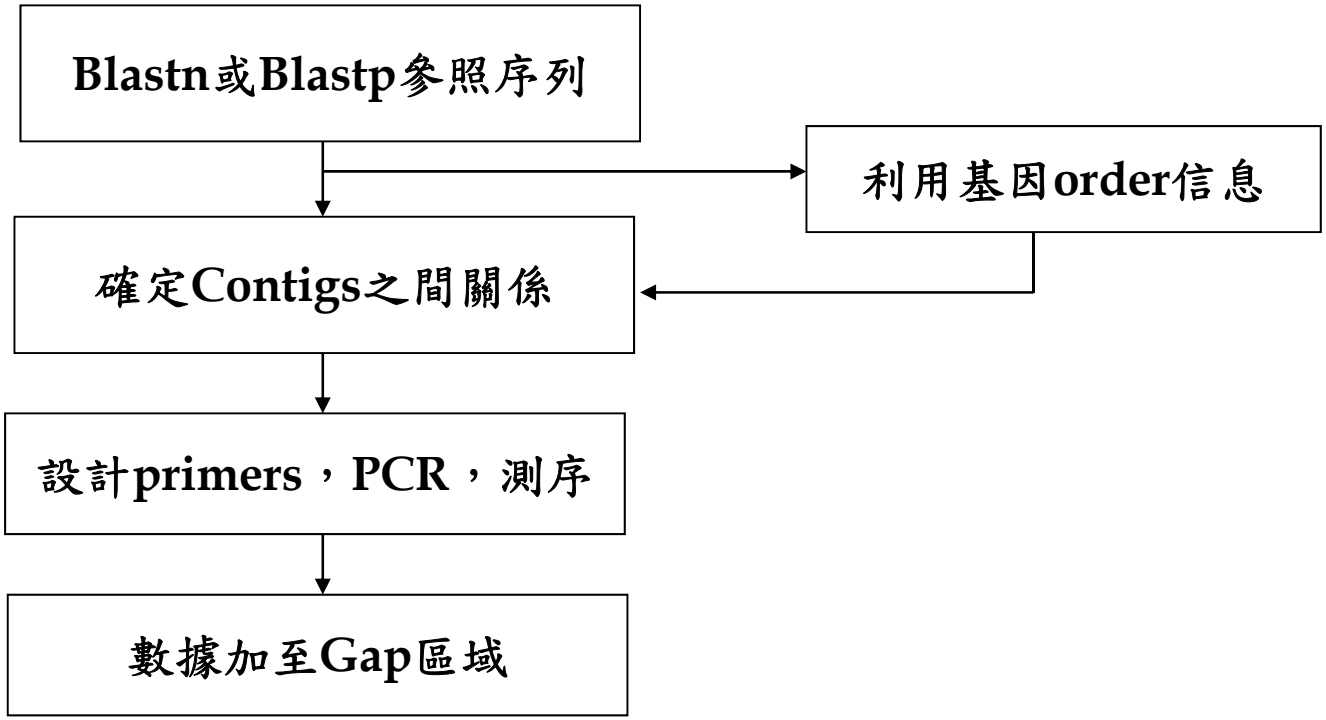
Analysis of Bacterial Genomes



To finish up a genome



To finish up a genome



The characteristics of the bacterial genomes

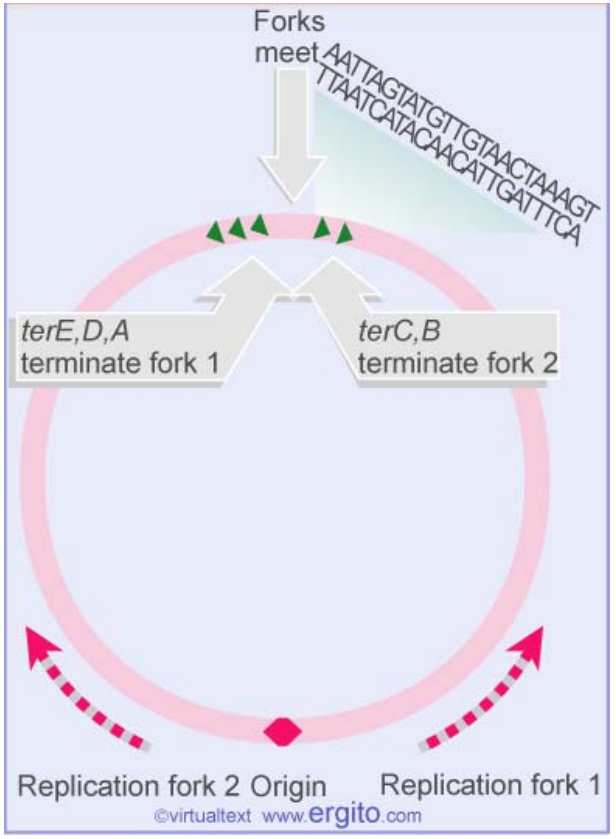
	Characteristics
Structure	**Double-stranded circular; linear**
Sizes	**0.16-13Mb**
ORF	**Occupy more than 90% of the total length; averagely 1 kb/ORF**
GC content	**16.6%-74.9%**

The characteristic of an oral streptococcus Genome

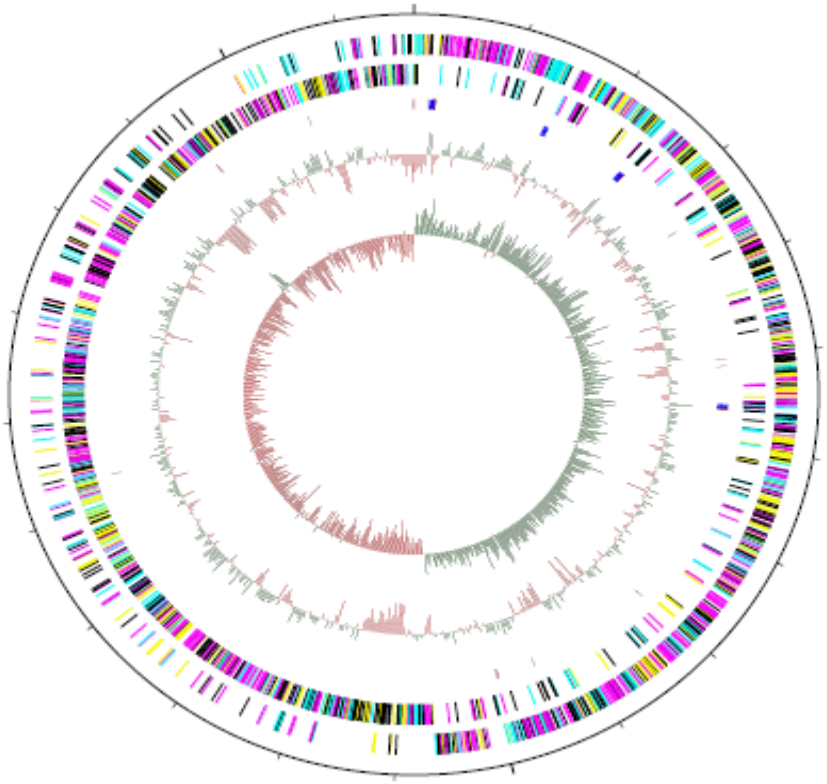
Genome size		**2,171,616 bp**
G+C content		**41.62%**
Average CDS length		**932 bp**
Protein coding		**2059**
	Protein with known or predicted fuction	**1496(72.7%)**
	Conserved hypothetical proteins	**461(22.4%)**
	Hypothetical proteins	**102(4.9%)**
tRNAs		**61**
rRNA operons		**4**
New Transcripts		
	Matched with known proteins	**10**
	Match with Rfam database	**5**
	Unknown functions	**29**

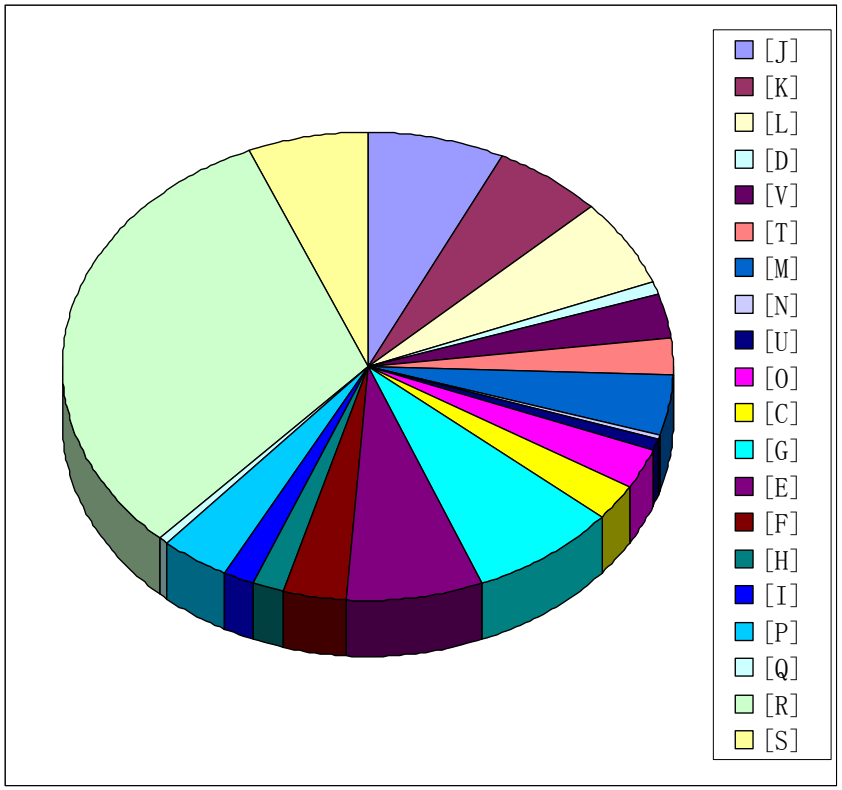
Origin (*ori*) and termination (*ter*) of replication

- **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.**



The characteristics of *S. parasanguinis* 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%)

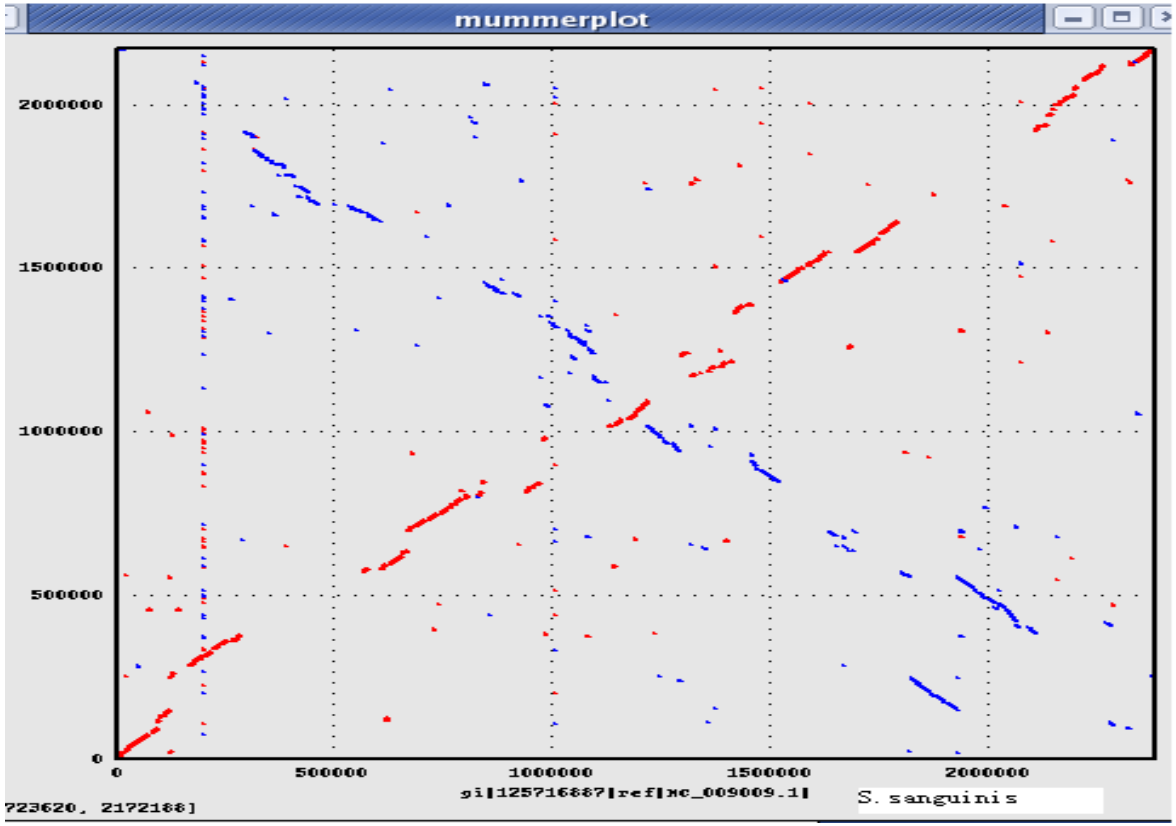
Comparison with other streptococcal species

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

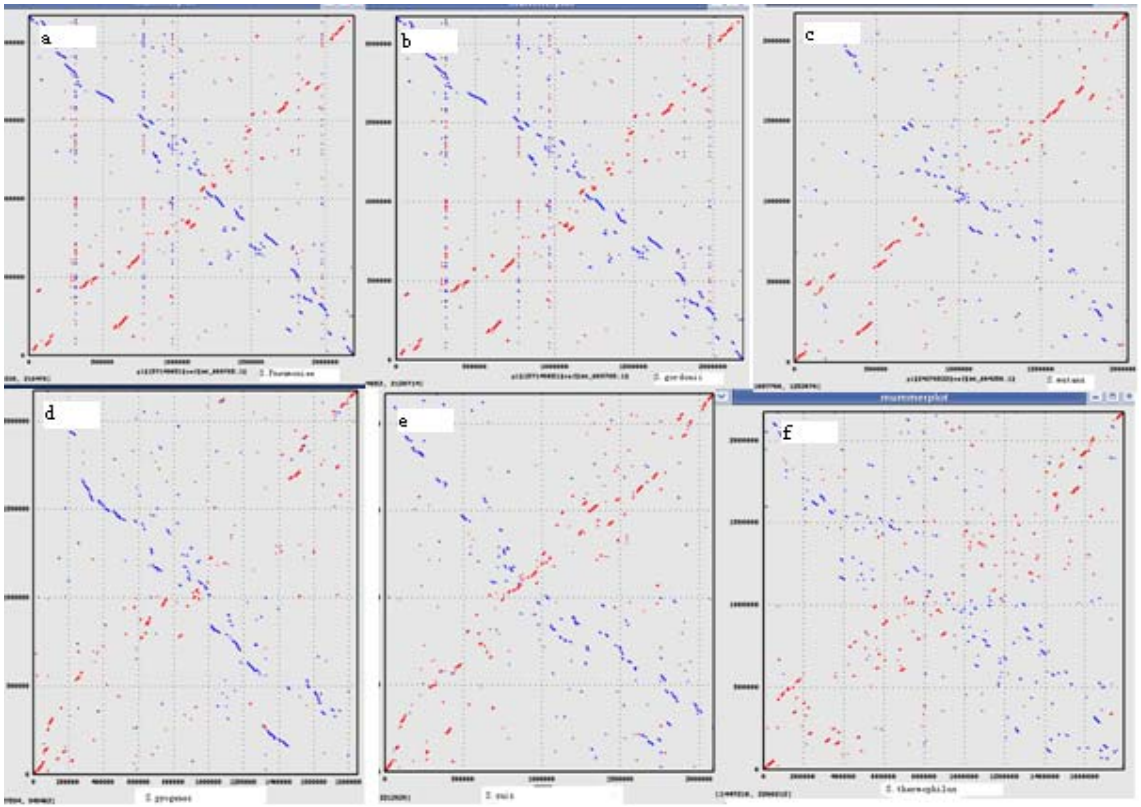
Comparison between *S. sanguinis* & *S. parasanguinis*

***Streptococcus sanguinis* SK36**



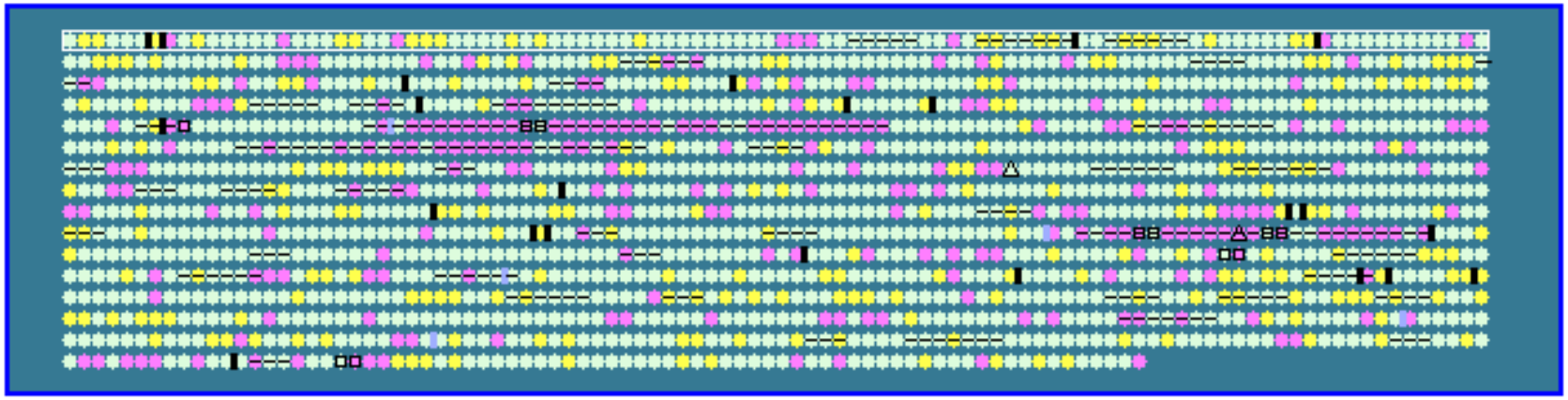
***Streptococcus parasanguinis* FW213**

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



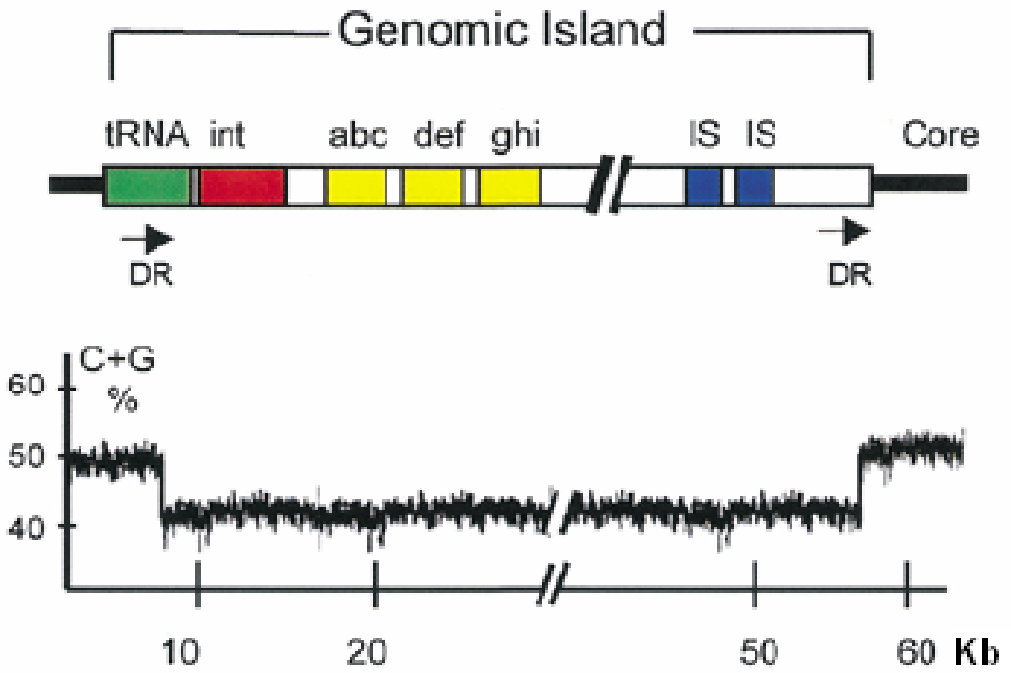
S. pyogenes *S. suis* *S. thermophilus*

Pathogenic islands



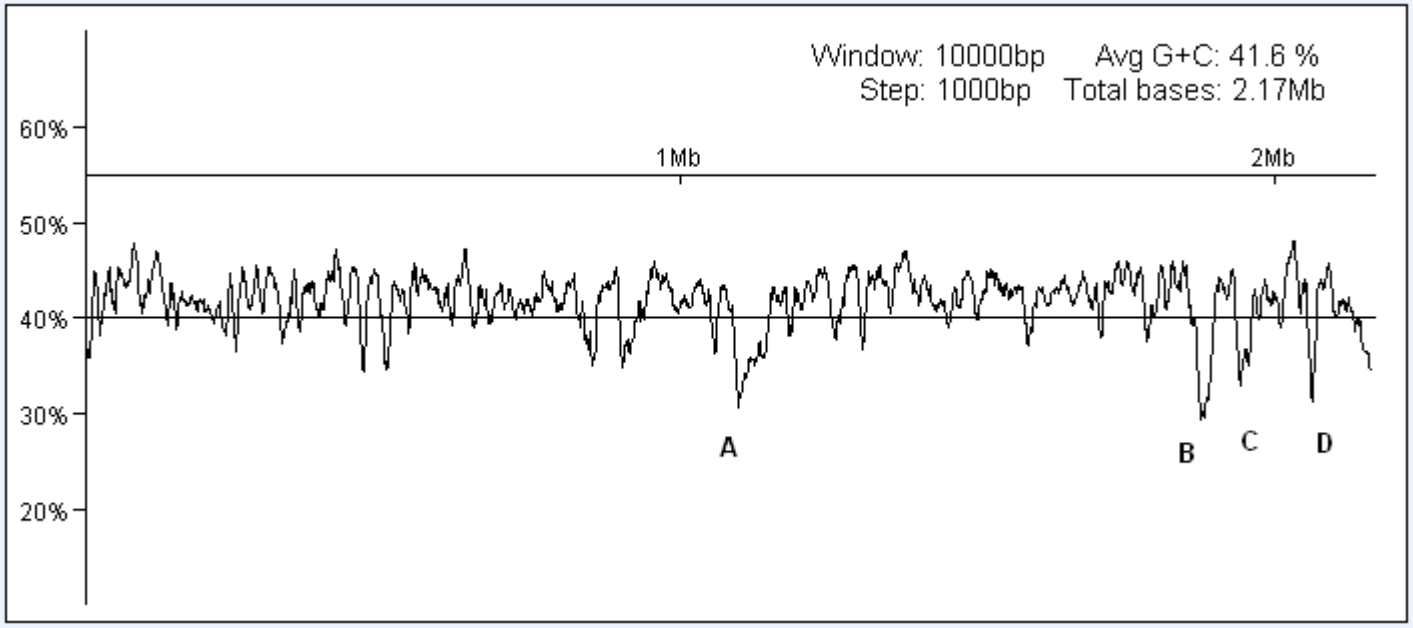
IslandPath對肺炎鏈球菌R6全基因組分析的輸出結果，每個圈代表一個預測編碼蛋白的可讀框，圈的顏色代表G+C百分比；橫線代表二核苷酸偏向性；分隔號代表轉運核糖核酸和核糖核酸；黑色方塊代表轉座酶基因；黑色三角代表整合酶基因。

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.

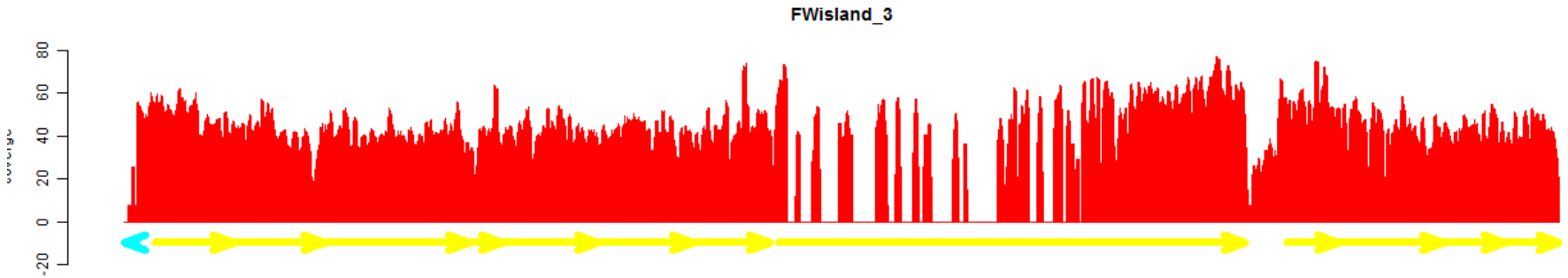
Pathogenic islands



The features of a genomics island

Locus	5'-end	3'-end	Annotation	Best BLAST match	GC content
orf1923	1940981	1941175	transposase	*Streptococcus sanguinis* SK36	38.46
orf1924	1941486	1942802	glycosyltransferase,gtf2	*Streptococcus parasanguis*	32.35
orf1925	1942798	1944309	glycosyltransferase,gtf1	*Streptococcus parasanguis*	32.74
orf1926	1944322	1946715	preprotein translocase subunit,SecA2	*Streptococcus agalactiae* A909	34.46
orf1927	1946726	1947256	glycosylation associated protein,gap3	*Streptococcus parasanguis*	31.64
orf1928	1947269	1948849	glycosylation associated protein,gap2	*Streptococcus parasanguis*	33.4
orf1929	1948866	1950440	glycosylation associated protein,gap1	*Streptococcus parasanguis*	31.68
orf1930	1950457	1951716	preprotein translocase subunit,SecY	*Streptococcus gordonii* str. C hallis substr. CH1	30.24
orf1931	1951827	1959605	fimbriae-associated protein,Fap1	*Streptococcus parasanguis*	38.34
orf1932	1960274	1961152	putative glycosyltransferase,galT2	*Streptococcus parasanguis*	34.58
orf1933	1961155	1962900	putative glycosyltransferase,galT1	*Streptococcus parasanguis*	33.28
orf1934	1962916	1963905	nucleotide sugar synthetase-like protein,nss	*Streptococcus parasanguis*	33.23
orf1935	1963926	1964816	putative glycosyltransferase (gly)	*Streptococcus parasanguis*	32.77

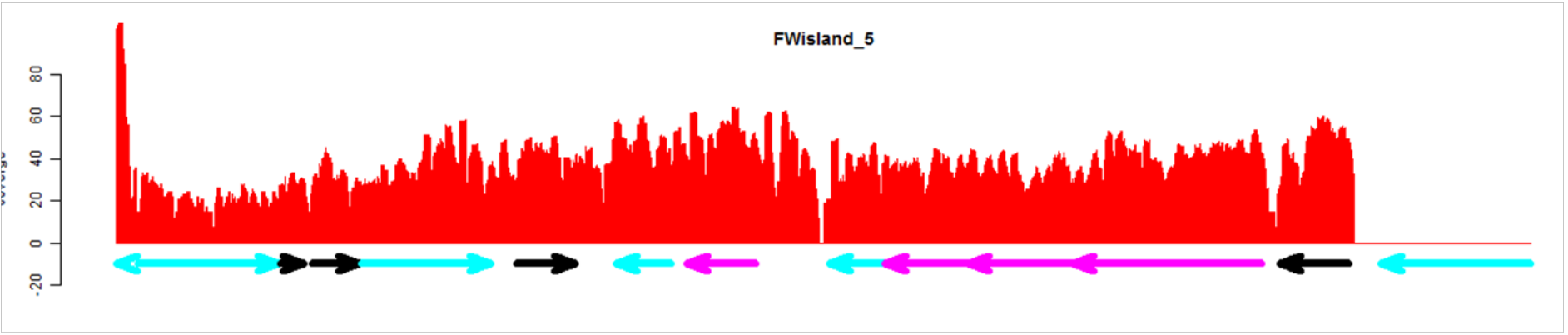
The genomic island and its transcript levels



The features of a genomics island

Locus	5'-end	3'-end	Annotation	Best BLAST match	GC content
orf2134	252272	252355	tRNA-Leu		66.67
orf242	252433	253563	Integrase		28.38
orf243	253570	253761	hypothetical protein	*Streptococcus pneumoniae* SP23-BS72	38.02
orf244	253824	254207	hypothetical protein	*Listeria monocytogenes str.* 1/ F6854	33.85
orf245	254213	255250	Tn916, transcriptional regulator	*Streptococcus agalactiae* 2603V/R	34.2
orf246	255450	255920	hypothetical protein	*Streptococcus agalactiae* 2603V/R	34.61
orf247	256243	256692	transcriptional regulator, Cro/CI family	*Streptococcus agalactiae* 515	29.56
orf248	256810	257361	phosphoglycerate mutase family protein	*Bacillus cereus*	30.98
orf249	257944	258378	putative transcriptional repressor,adcR	*Streptococcus gordonii* subsp. Challis	38.85
orf250	258383	259084	ABC transporter, Zn porter,adcC	*Streptococcus sanguinis* SK36	43.16
orf251	259050	259880	ABC transporter, Zn porter,adcB	*Streptococcus sanguinis* SK36	44.28
orf252	259886	261388	Metal-binding (Zn) permease,adcA	*Streptococcus sanguinis* SK36	40.45
orf253	261530	262084	Conserved uncharacterized protein	*Streptococcus sanguinis* SK36	44.32
orf254	262337	263536	Transposase, IS/IS1328/IS1533	*Streptococcus suis* 89/1591	39.58

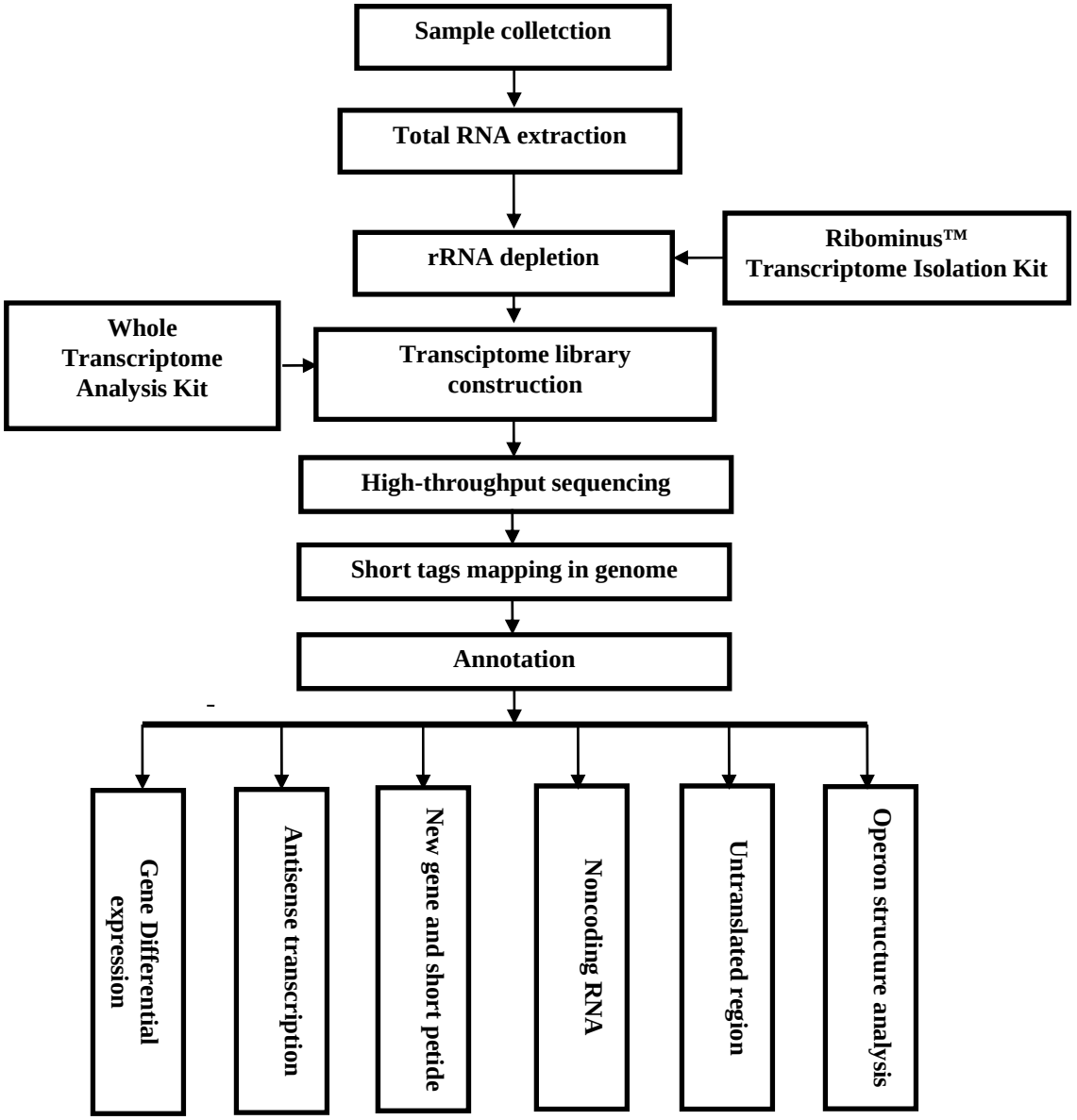
The genomic island and its transcript levels



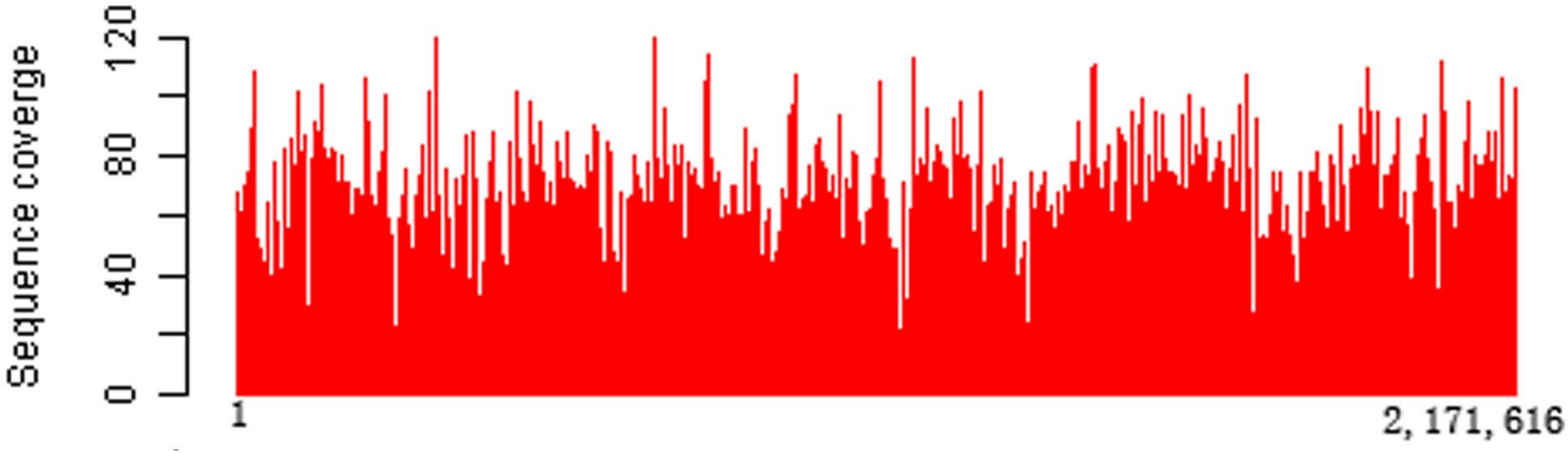
What can we learn from expression analysis by NGS?

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

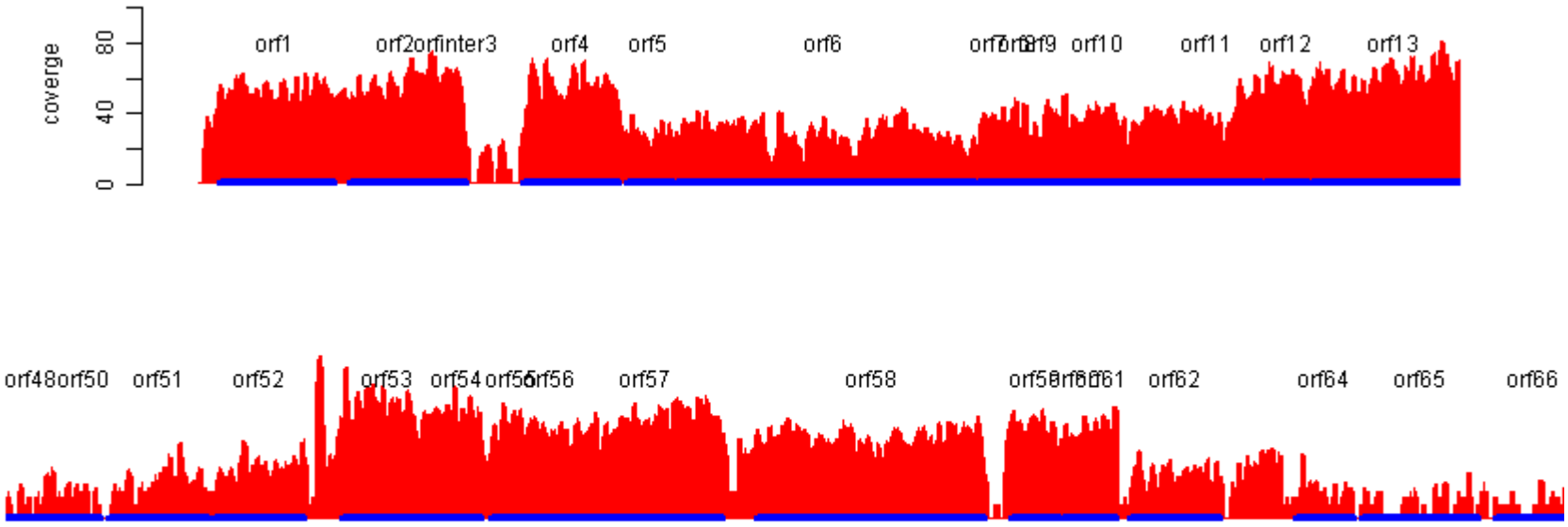
The approaches



The transcriptome of an oral streptococcus

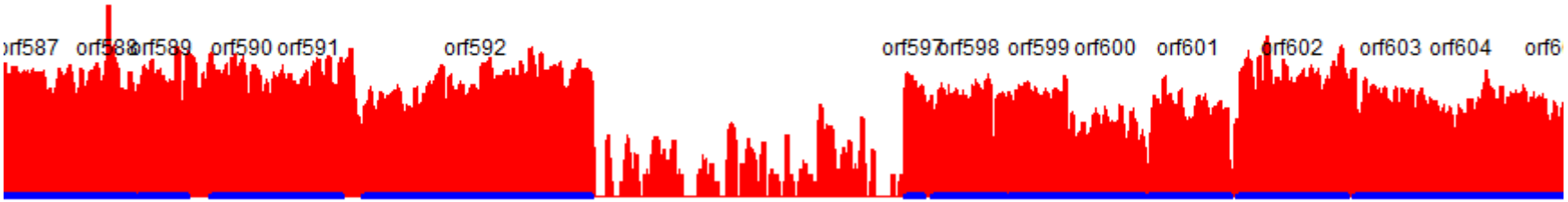


Single nucleotide mapping coverage

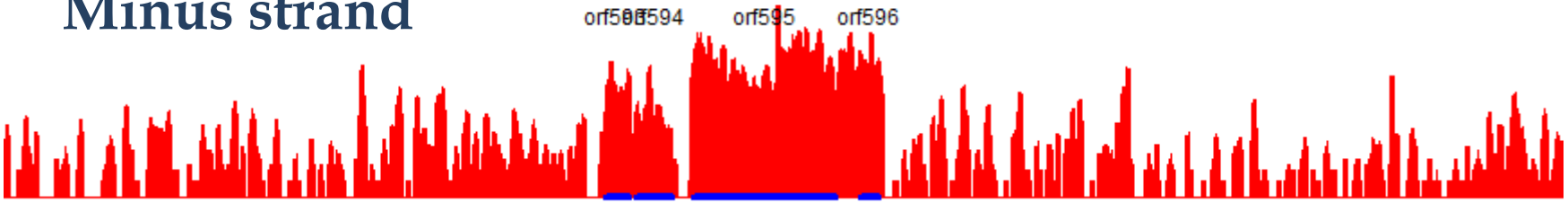


The expression of the two strands in a single region of the chromosome

Plus strand



Minus strand



The statistics of a NGS result

Map to CDs

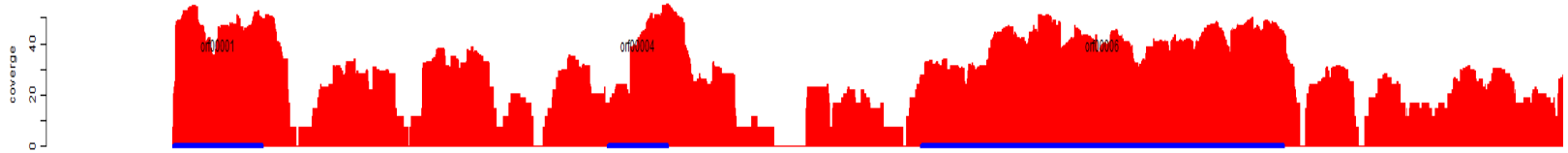
Read	Count	Avg length	Total bases
Reads	53,140,406	49.78	2,645,221,763
Matched	9,760,186	49.74	487,471,780
Not matched	43,380,220	49.74	2,157,749,983
References	2,049	934	1,913,826

Map Unmapped to Genome

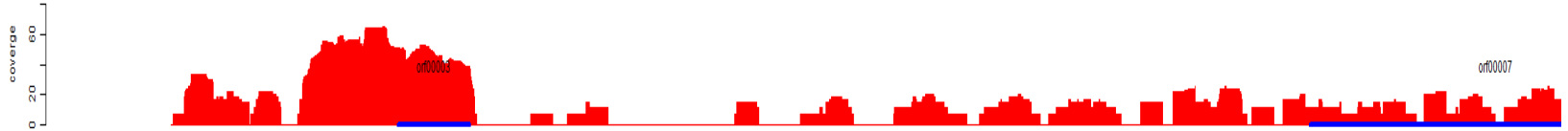
Read	Count	Avg length	Total bases
Reads	43,380,220	49.74	2,645,221,763
Matched	9,193,393	49.95	459,181,927
Not matched	34,186,827	49.68	1,698,568,056
References	1	2,171,616	2,171,616

An example: Expression analysis of pFW213

Plus strand



Minus strand



Validation of NGS expression analysis

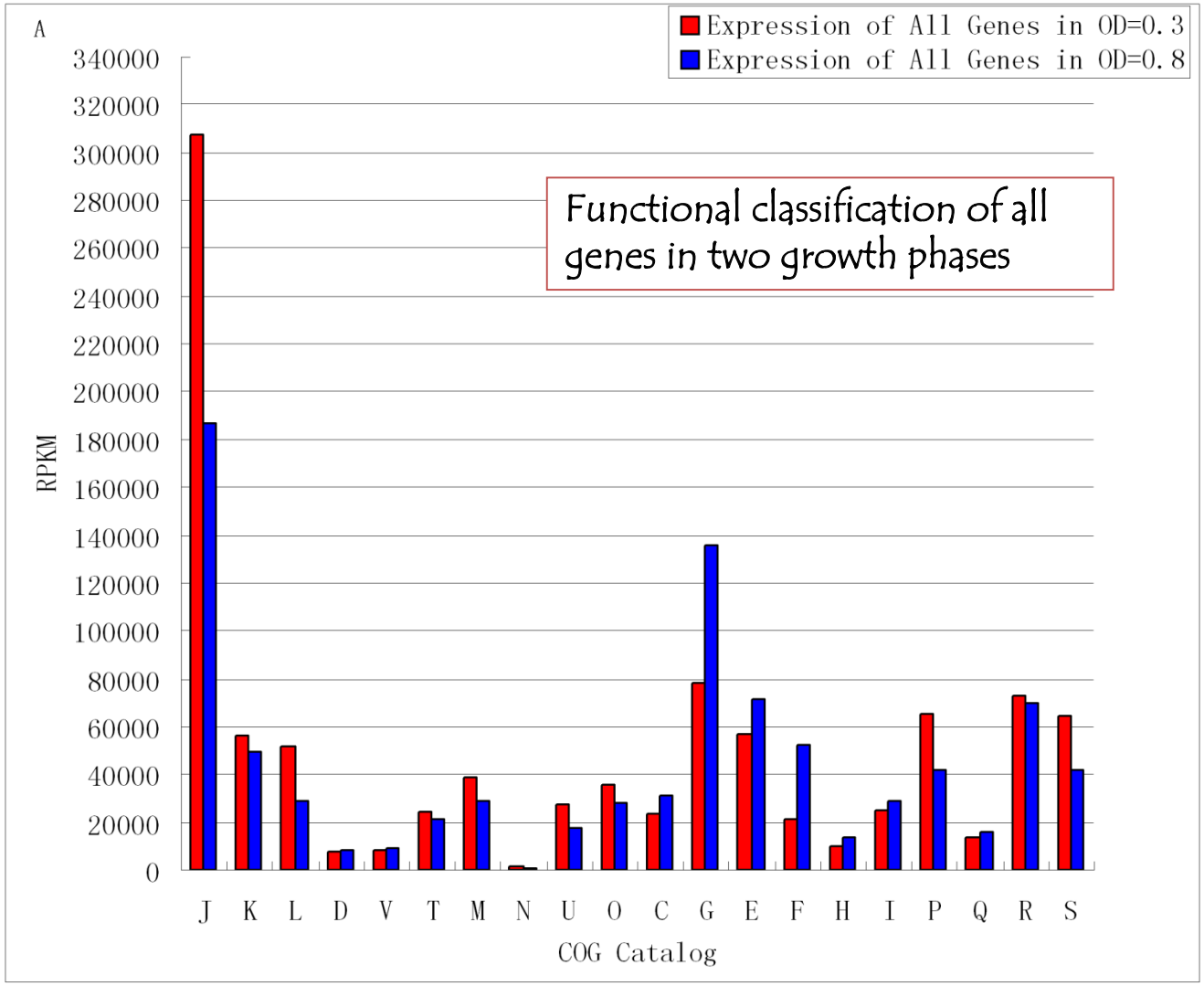
ORF	Annotation	NGS	RT-PCR
Orf1	inorganic pyrophophatase	+2.5	+3
Orf2	no nr database	+2.5	+3
Orf3	no nr database	0	Yes/No
Orf4	Mob	+0.5/0	+1
Orf5	Rep	+1.5	+2
Orf6	Hypothetic protein	+1.5	+2
Orf7	RelE	+1.25	+0.5
Orf8	RelB	+1.25	+0.5

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

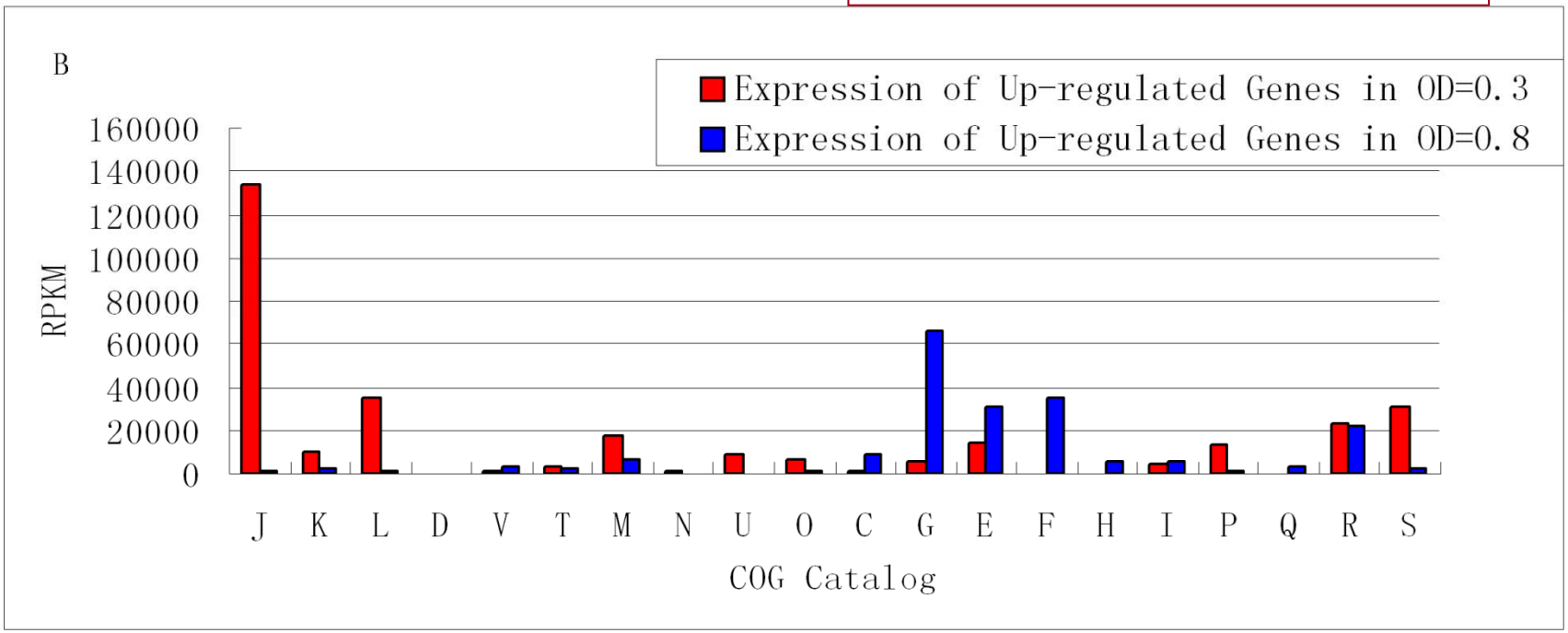
ORF	strand	gene length	fw3tag number	fw3rpkm	fw8tag number	fw8rpkm	fw3/fw8 rpkm	fw3/fw8 PT-PCR
orf48	+	825	46	4.824337	1669	191.6416	0.025174	0.2499
orf58	+	2649	4412	144.1077	34467	1232.563	0.116917	0.4841
orf309	+	435	537	106.8117	20904	4552.266	0.023463	0.8093
orf439	+	1665	89	4.624977	3039	172.9036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

Gene expression analysis generated by RNA-seq

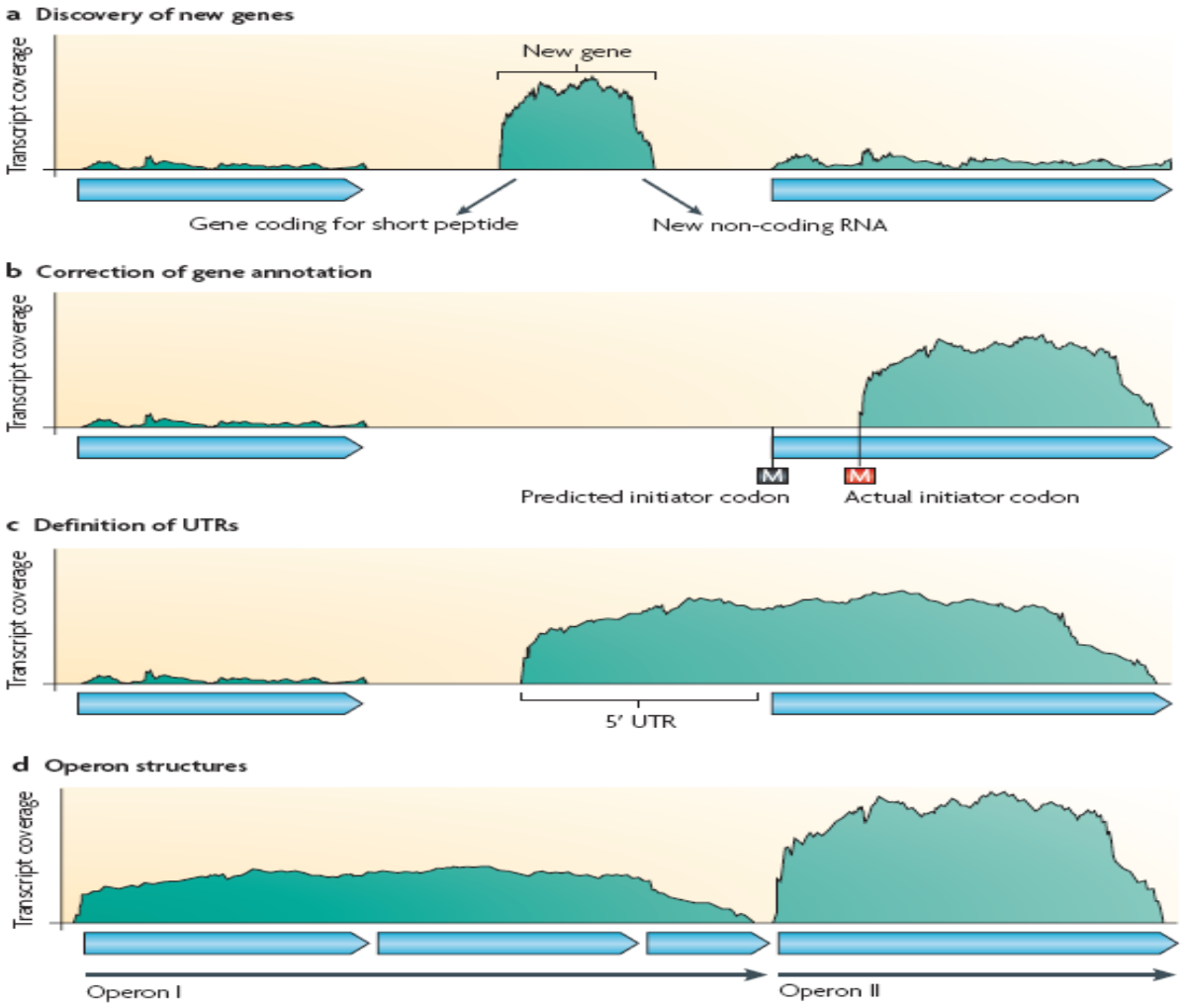


Gene expression analysis generated by RNA-seq

differentially expressed genes in two conditions

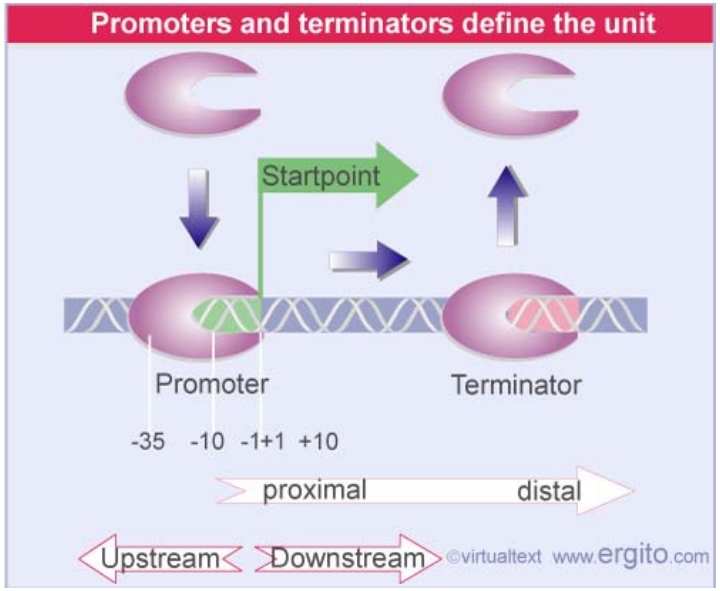


Other possibilities



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

A transcription unit is transcribed into a single RNA



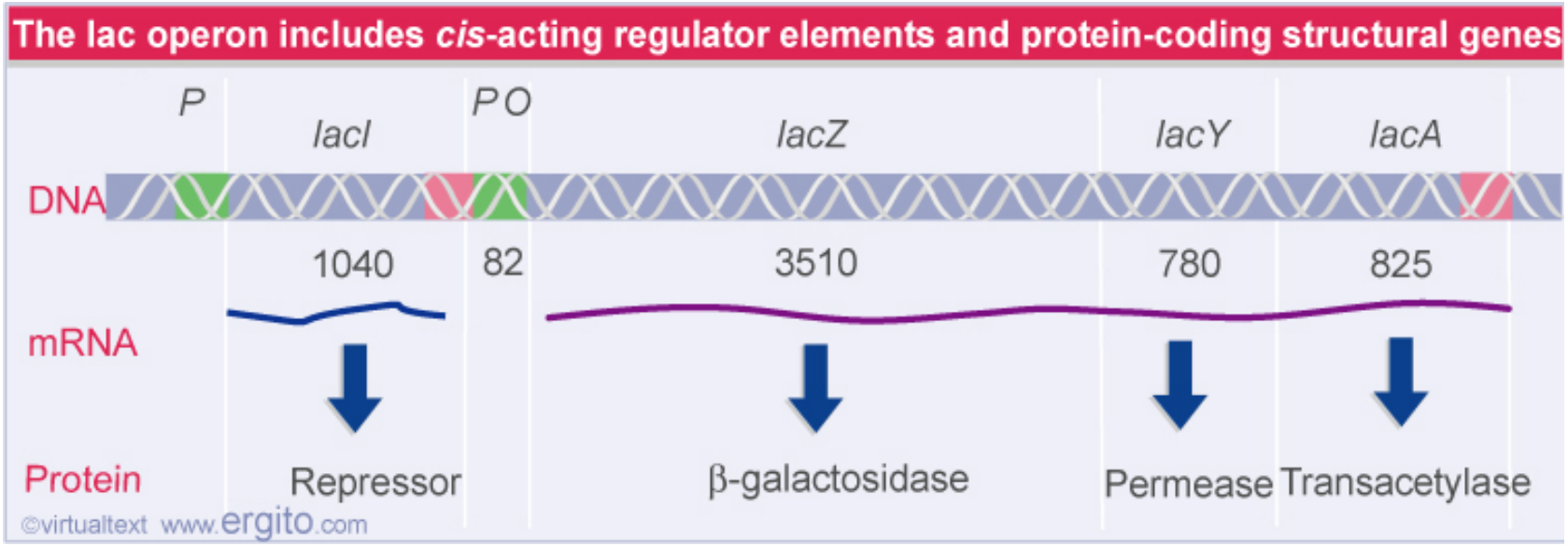
- The **start point** (start site) refers to the position on DNA corresponding to the first base incorporated into RNA.
- A **terminator** is a sequence of DNA that causes RNA polymerase to terminate transcription.
- A transcription unit is the distance between sites of initiation and termination by RNA polymerase; **may include more than one gene.**

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).**

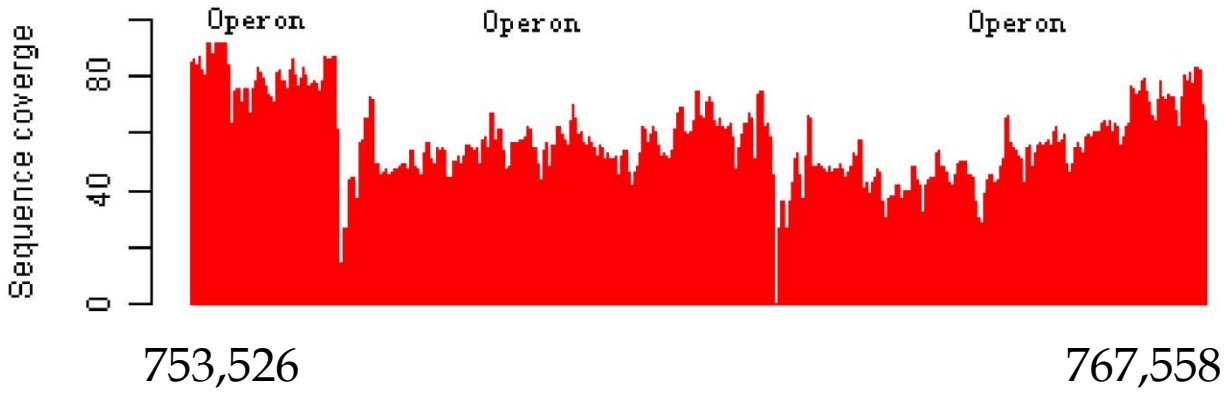
- **Structural gene clusters are coordinately controlled.**

The *lac* operon



- **Transcription of the *lacZYA* gene cluster is controlled by a repressor protein that binds to an operator that overlaps the promoter at the start of the cluster.**
- **The repressor protein is a tetramer of identical subunits coded by the gene *lacI*.**

Operon determination



What else?

- **Finding novel small RNA**
- **Potential small proteins**