



Pathogen genomics: approaches and applications

Dr Nick Loman
University of Birmingham
Tuesday, 10th June 2008

Acknowledgments

- Caroline Channing / EADGENE
- Mark Pallen for this talk!
- Roy Chaudhuri, BBSRC for xBASE
- Lori Snyder for genome annotations
- Del Ala'Aldeen, Mike Barer, Chris Clew, Neil Hall, Debbie Mortiboy, Julian Parkhill and the PSU, Charles Penn, Kumar Rajakumar, George Weinstock, Adrian Simmons for figures and/or DNA and/or sequences and/or inspiration



This Talk

- Bacterial genomics in context
- What have we learnt from studying bacterial genomes so far?
- Future perspectives- next-generation sequencing and translational genomics

Bacterial Genomics is different...

- Easier
 - ~1000x smaller genomes than animal genomes
 - no introns; >80% of genome is protein-coding
 - “WYSIWYG genomes”
 - small genome size means that whole-genome analyses (e.g. transcriptomics) have been possible using current technology for several years
 - synergy with human and animal genome projects (whole-genome shotgun sequencing pioneered on bacteria)

Bacterial Genomics is different...

- Harder
 - immense variation between strains and species
 - bacteria engaged in global orgy of group sex
 - two *E. coli* strains can differ by one third of their genome (cf. humans versus slime moulds)
 - Many more protein-coding genes in human microbiome than in human genome

Bacterial genome sequencing: Milestones

- 1995: First complete genome sequence of free-living organism, *Haemophilus influenzae*, 1995
- 1997: *E. coli* K12 sequenced
- 2002: ~100 completed genomes
- 2008: 666 bacterial genomes complete, 1837 projects under way

Genomes Online Database

Animal Bacteriology

- Animal Pathogens
 - *Escherichia coli* (70)
 - Staphylococcus spp. (37)
 - Streptococcus spp. (99)
 - Pasteurella, Actinobacillus, Mannheimia (12)
 - Mycobacterium bovis (4)
- Zoonoses
 - Salmonella spp. (71)
 - Campylobacter spp. (23)
 - Yersinia spp. (27)
 - Helicobacter spp. (22)

Genomes Online Database

What have we learnt from completed Genome-Sequencing Projects?

- *Campylobacter jejuni*
 - Capsule, homopolymeric tracts, N-linked glycosylation
- *Corynebacterium diphtheriae*
 - Fimbriae, iron-scavenging systems
- *Tropheryma whippelii*
 - Variable surface protein family
- *Francisella tularensis*
 - Aromatic amino acid biosynthesis pathways



What have we learnt?

NATURE 443 XXXX INSIGHT XXXXXXXX

Bacterial pathogenomics

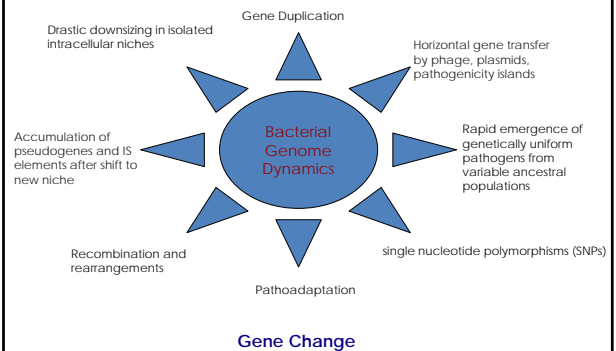
Mark J. Pallen¹ & Brendan W. Wren²

The genomes of all of the important bacterial pathogens of humans, plants and animals have now been sequenced, as have those of many important commensal, symbiotic and environmental microorganisms. Analysis of these sequences has revealed the forces that shape pathogen evolution and has brought to light unexpected aspects of pathogen biology. The finding that horizontal gene transfer and genome decay have key roles in the evolution of bacterial pathogens was particularly surprising. It has also become evident that even the definitions for 'pathogen' and 'virulence factor' need to be re-evaluated.

Nature 2007 Oct 18;449:835-42

Gene Loss

Gene Gain



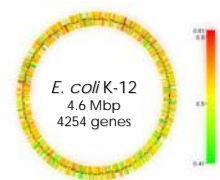
Escherichia coli a versatile bacterium

- Gut commensal
 - from kangaroos to cattle
- Model lab organism
- Biotechnology work horse
- Probiotic
- Pathogen



Escherichia coli K-12

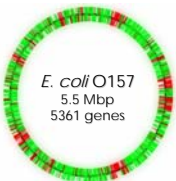
- archetypal model *E. coli* strain
- isolated from stool of a convalescent diphtheria patient in Palo Alto in 1922
 - commensal strain that has undergone extensive manipulation in the laboratory
- has one of the smallest known *E. coli* genomes
 - often taken as close to the ancestral archetype



<http://xbase.ac.uk>

Escherichia coli Genomic diversity

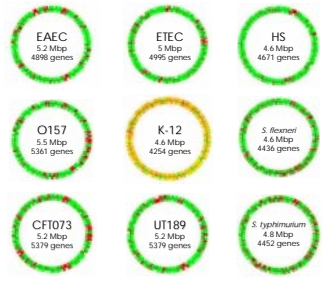
- *E. coli* O157 genome sequenced in 2001
- common "backbone" sequence with K-12
- 1.5 Mbp novel sequence cf. K-12
- extensive lateral gene transfer
- 177 "O-islands"



E. coli O157
5.5 Mbp
5361 genes

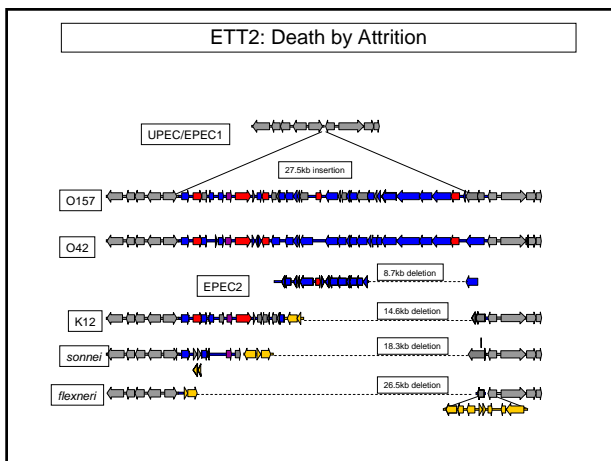
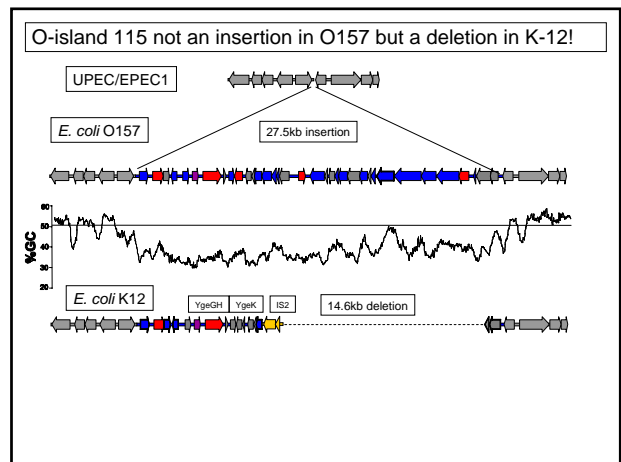
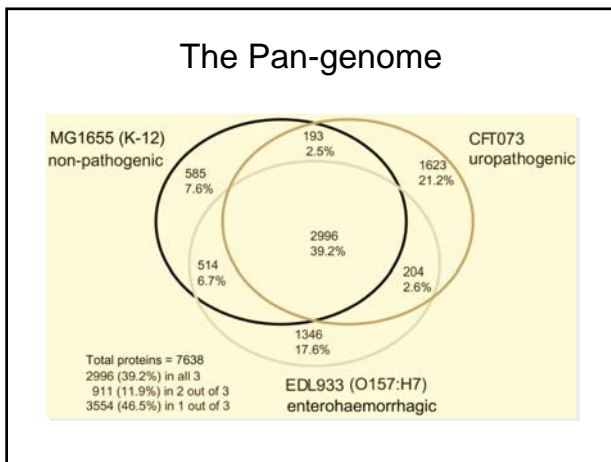

Hayashi et al. DNA Res. 2001 Feb 28;8(1):11-22
Perna et al. Nature. 2001 Jan 25;409(6851):529-33

Escherichia coli Genomic diversity



Strain	Size (Mbp)	Genes
EAEC	5.2	4996
ETEC	5	4995
HS	4.6	4611
O157	5.5	5361
K-12	4.6	4254
<i>S. flexneri</i>	4.8	4436
CFT073	5.2	5379
UT189	5.2	5379
<i>S. typhimurium</i>	4.8	4452

Welch et al. Proc Natl Acad Sci U S A. 2002 Dec 24;99(26):17020-4

xBASE

Popular Tools

Genomes

Alignments

xBASE Alignment Viewer

xbase.ac.uk



Next-Generation Sequencing



At least 3 rival technologies

- 454 (Roche)
- Solexa (Illumina)
- SOLiD (ABI)

~100x faster!
 ~100x cheaper!
 A DISRUPTIVE
 TECHNOLOGY?

[NB single molecule approaches could make even conventional next-generation sequencing look lame]

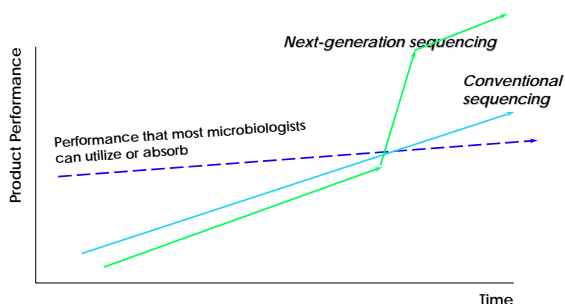
'Next generation': fundamentally new approaches

- Clonal template populations obtained by new methods:
 - PCR on solid phase to grow a 'molecular colony'
 - Massive increase in number of 'clones' compensates for shorter read length
- New chemistries for sequence reading
 - Pyrophosphate detection (PPi release upon base addition): 454
 - Single (reversibly 3'-blocked) fluorescent base (quencher) added per step: Solexa
 - Sequencing by Ligation (ABI SOLiD)

Throughput

Machine	Raw Data [base pairs]	Read Length
ABI 3700 (96 wells)	100,000	1000
ABI 3730XL	1,000,000	1000
454	200,000,000	250
Solexa	2,000,000,000	36

Could next-generation sequencing be a disruptive innovation in microbiology?



Disruptive innovations

- Digital photography versus traditional
- HD-DVD versus DVD versus video tapes
- MP3 versus CD versus audio cassette versus vinyl
- DTP versus type setting
- Desktop PC versus mainframe
- TV versus radio
- Car versus horse-drawn carriage

Opportunities Basic Science

- Elucidating the biology of the organism:
 - virulence factors
 - antibiotic resistance
 - drug and vaccine targets
 - microbiota
- 454 sequencing projects great for this

Opportunities Translational Genomics

- Genome sequencing as epidemiological tool
 - Universal digital “library” technique
 - Portable in time and space
- Ultimate in resolving power: can resolve a single SNP; indistinguishable means identical!
 - Track spread of bacterial pathogens between patients and in environment in real time
 - Improved understanding of patterns of spread will allow optimal targeting of scarce resources for infection control
- Will it replace existing techniques (PFGE, spoligo-typing, MLST, VNTR etc) or simply complement them?
 - Rather than *why* sequence the genome, *why not*?



Opportunities Translational Genomics

- Improved understanding of pathogens & their evolution
 - will be able to watch pathogen evolution in real time, between and within patients
 - detect genes under positive selection during human infection
 - comprehensive cataloguing of pathogen gene pools
- Identifying and characterising emerging pathogens
- Detection of antibiotic resistance
 - e.g. in *M. tuberculosis*
- Translational metagenomics?

Illumina (Solexa)

- Illumina Genome Analyzer 1G
- Up to 2 Gbp per 2-3 day run
- 3 day sample prep
- 8-channel flow cell allowing same sample (for large genomes) or different samples to be loaded every run.



TB



TB Strain Selection

Strain	Provenance	Collected	Antibiotics
CH	Crown Hills Index Case	16-Feb-2001	Sensitive
DP	Father of CH	04-Jan-2001	Sensitive
RY	Non-outbreak associated isolate	03-Sep-2003	Sensitive
SG	Non-outbreak associated isolate	09-Mar-2004	Sensitive

	ETR	MIRU	RFLP
CH	Indistinguishable	Indistinguishable	Indistinguishable
DP	Indistinguishable	Indistinguishable	Indistinguishable
RY	Indistinguishable	Indistinguishable	n/a
SG	Indistinguishable	Indistinguishable	n/a

Single Sequencing Run

Illumina Genome Analyzer

Lane	Strain	Raw Bases	Aligned reads	Coverage
1	<i>Mycobacterium tuberculosis</i> CH	115,422,538	88.7%	24x
2	<i>Mycobacterium tuberculosis</i> RY	75,577,454	90.5%	16x
3	<i>Staphylococcus aureus</i> MRSA 255.1	79,335,887	81.28%	24x
4	<i>Acinetobacter baumannii</i> 6014059	64,929,187	75.73%	13x
5	<i>Acinetobacter baumannii</i> 6013150	115,124,509	66.26%	22x
6	Pool 1	87,701,551	74.77%	n/a
7	Pool 2	99,355,385	12.58%	n/a
8	Control	113,700,459	100%	n/a

George Weinstock, BCM, Texas

Multiplex Lanes

Pool 1

Total Bases 87,701,551

Strain	% Aligned Reads	% All Reads	Coverage
<i>Mycobacterium tuberculosis</i> DP	13.9%	10.4%	2.2x
<i>Acinetobacter baumannii</i> 6013113	45.70%	34.20%	8.8x
<i>Staphylococcus aureus</i> MRSA 289.1	40.40%	30.20%	9.8x

£200 genome?!

George Weinstock, BCM, Texas

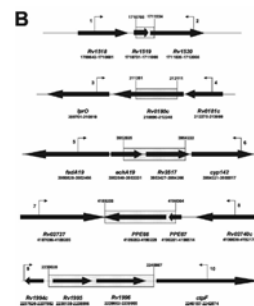
Short Read Mapping To Reference

CDC-1551 Genome 4.4 megabases

Sequence	Reference	Coverage	High Quality SNPs	Uncalled
CH	CDC-1551	24.8x	1536	9926
DP	CDC-1551	2.24x	259	550,294
RY	CDC-1551	16.4x	1495	20,839
SG	n/a	n/a	n/a	n/a

GLIP

Strain	GLIP
CH	G00000
DP	G00000
RY	G00001
SG	G01001



- 1) Whole genome microarray
- 2) PCR-based confirmation
- 3) 5 loci
- 4) Sequencing of products

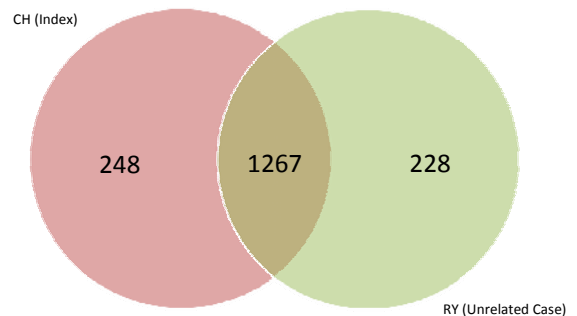
Rajkumar et al, J. Clin Micro 2004

Compare with GLIP *Mycobacterium tuberculosis* CH

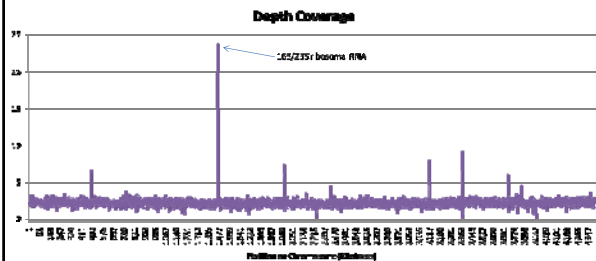
Gene	'Deletion' Length
Rv0180c	821
Rv1519	784
plcD	71
ctpG	62
Rv1995	1892
echA19	404
Rv3517	850
Rv3737	1103
Rv3785	65

Plus over 100 more 'deletions', mainly in PE/PPE/PGRS genes

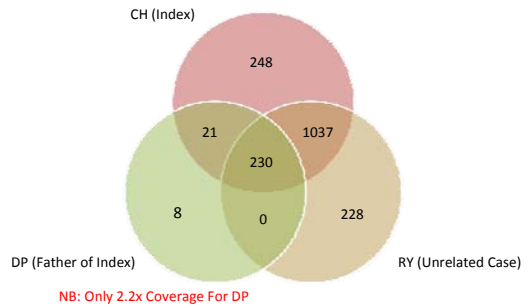
TB SNPs Summary of differences from CDC-1551



Coverage Plot DP vs CDC-1551



TB SNPs Summary of differences from CDC-1551



Unique DP SNPs

Position	Gene	Description	Ref	Cons	Qual	Depth	Change
112087	MT0110	peptide synthetase, putative	T	G	39	4	Syn
245163	MT0216	membrane protein, Mmpl family	N	C	35	3	X→G
245164	MT0216	membrane protein, Mmpl family	C	G	36	3	X→G
395969	MT0345	transcriptional regulator, TetR family	A	C	36	3	V→G
638735	MT0570	phosphate transport protein	T	G	39	4	S→P
638738	MT0570	phosphate transport protein	A	G	39	4	T→P
807740	MT0736	50S ribosomal protein L29	A	G	36	3	E→G
2697097	MT2474	hypothetical protein	T	G	35	4	V→G

Shared SNPs – CH and DP vs CDC-1551

Position	Gene	Product	Ref	Cons	Quality	Depth	Alt Change
103542	MT0102	hypothetical protein	G	T	36	3	A→S
248496	MT0218	rRNA (guanine-N(7)-methyltransferase	G	A	39	4	Syn
504983	MT0432	hydrolase	C	T	36	3	Syn
747493	MT0676	glycosyl hydrolase, family 5	C	A	39	4	Syn
1350486	MT1244	acyl-CoA synthetase	C	T	34	4	Syn
1502123	MT1373	ATP-dependent Clp protease adaptor protein ClpS	C	T	42	5	Syn
1583802	MT1451	sum protein	C	T	42	5	T→H
1616438	n/a	n/a	A	C	36	3	n/a
1720235	MT1580	acyl-CoA synthetase	T	G	39	4	F→V
2212751	MT2112	hypothetical protein	G	A	51	8	A→V
2229857	MT2141	hypothetical protein	N	C	36	3	T→D
2485618	MT2278	glutamine synthetase	G	A	39	4	Syn
2630767	n/a	n/a	G	A	45	6	n/a
2626955	MT2678	hypothetical protein	G	C	36	3	D→E
2977372	n/a	n/a	T	C	36	3	n/a
3391857	MT3120	hypothetical protein	C	G	36	3	P→A
3429700	MT3153.1	campylobacter resistance protein CrcB	T	G	36	3	L→R
4486244	n/a	n/a	G	C	36	3	n/a
3509273	MT3236	NADH dehydrogenase subunit D	G	T	39	4	G→W
3613311	MT3337	drug transporter	A	C	39	4	L→R
3803457	MT3504	bifunctional GMP synthase/glutamine amidotransferase protein	C	A	36	3	Syn

Translational genomics

- Opportunities
 - The ultimate tool for tracking and mapping our microbial adversaries....
 - Bacterial genome sequencing poised to enter medical and public health microbiology?
 - a genome sequence for £30?

Translational genomics

- Challenges
 - The problem of Malthus applied to genomics?
 - How will we cope with the data flood?



Genome sequence data

Our ability to analyse it

