Genome-wide Single Nucleotide Polymorphism Analysis of *Salmonella enterica*

P. Hu[1], D. Ballinger[2], C. Pethiyagoda[2], K. Pant[2], T. Torok[1] and G. L. Andersen[1]
[1]Lawrence Berkeley National Lab, [2]Perlegen Sciences, Inc. H-125

Abstract: The goal of this study is to identify signature single nucleotide polymorphisms (SNPs) to differentiate strains of *Salmonella* spp. Prompt identification of the source of a natural, food-borne disease outbreak or an act of bioterrorism can limit the number of affected individuals and save lives. In the event of an outbreak, DNA fingerprints of *Salmonella* samples isolated from infected individuals can be compared to those of strains associated with suspected source of contamination, helping to identify the source of the outbreak. We discovered more than 33,286 SNPs. The identified SNPs appeared to be relatively evenly distributed throughout the genomes of previously sequenced strains. The SNPs were verified on second array and were used for genotyping 217 strains of *Salmonella* serovars Enteritidis, Typhi, Dublin, Typhimurium and several others. The strains tested were clustered into five major groups each representing Typhimurium, Dublin, Typhi, Enteritidis and other serotypes. In addition, strains within Enteritidis were also clustered into distinct groups largely corresponding to major phage types. Deletion analyses also correlated phage-type specific patterns. Subsets of SNPs were identified which further delineated individual isolates within each major phage type. Molecular signatures based on single nucleotide polymorphism genotyping are a powerful tool for grouping *Salmonella enteritidis* strains for diagnostics and can be used to study the evolution of this pathogen.

Introduction: *Salmonella enterica* is a ubiquitous pathogen with more than 2,500 known serovars. *Salmonella* serotype Typhi causes typhoid fever and is specific to humans (5). By contrast, two of the most common serovars, Typhimurium and Enteritidis have a broader host range and cause a gastroenteric form of the disease (1, 6). The differences in host range and mode of infection appear to be due to about 500 kb of unique chromosomal and plasmid sequence for each of the serovars. The remainder of the genome appears to be quite stable at the individual gene level although inversions and rearrangements may be common. Molecular typing is desirable for both a more rapid typing method and a tool to study the evolution of the serotypes and strains.

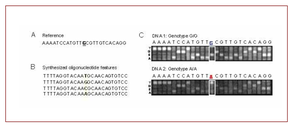
Methods:

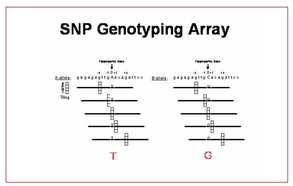
The SNP discovery and genotyping were accomplished using oligo microarrays by Perlegen sciences, Inc. (http://www.perlegen.com)

1) SNP discovery: re-sequence every base in both forward and reverse strand Genomic DNA from *S. typhy* CT18, Ty2, *S. enteritidis* PT4 and *S. typhymurium* LT2 were fractionated, end-labeled with biotin and hybridized to 60 million probes.

2) Genotyping with SNPs: discovered SNPs were used to design a 500,000-probe genotyping array. It was used for genotyping 217 strains of *Salmonella* serovars Enteritidis, Typhi, Dublin, Typhimurium and several others.

3) Data analysis: A total of 33,286 SNPs were used in analyses. Each SNP was mapped to at least one of the four reference genomes: *Salmonella typhi* CT18 (4), *Salmonella typhi* Ty2 (3), *Salmonella typhymurium* LT2 (2) and *Salmonella enteritidis* PT4 (http://www.sanger.ac.uk/Projects/Salmonella/). Genomic deletions were identified within each tested isolate by lack of oligonucleotide hybridization for at least 10 adjacent SNP loci. Agglomerative clustering was used for phylogenetic analyses.





Agglomerative Clustering algorithm: Agglomerative clustering is a top-down hierarchical algorithms (The disadvantage of nonhierarchical clustering, such as the k-means algorithm is that the clustering depends greatly on the initial choice of cluster centers. The resulting clusters are not independent of the order in which the data are processed). It begins with each object as a separate group. These groups are successively combined based on similarity until there is only one group remaining or a specified termination condition is satisfied. For n objects, n-1 mergings are done (It repeatedly links pairs of clusters until every data object is included in the hierarchy). Hierarchical algorithms are rigid in that once a merge has been done, it cannot be undone. Similarity was determined by common distance functions, such as the Euclidian distance functions.

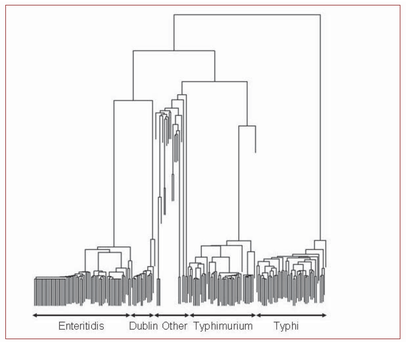
Application of the agglomerative clustering method in our *Salmonella* genotyping studies gave us satisfactory mathematical evaluation. A "good" clustering solution should have the following merits: 1) homogeneity: strains inside a cluster are highly similar; 2) separation: strains from different cluster have low similarity to each other. Our preliminary data have shown excellent results in balancing these seemingly conflicting features. We had clear separation not only with each biovars, but also at the phage type level.

Results:

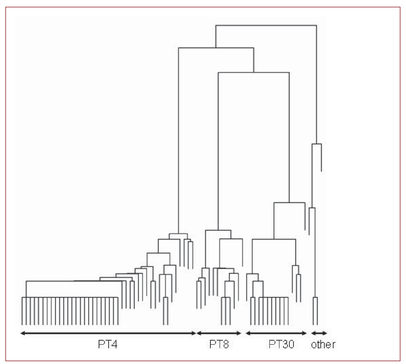
SNPs mapped onto reference genomes: A total of 33,286 SNPs were mapped to at least one reference genome

	Typhimuri um LT2	Typhi CT18	Typhi Ty2	Enteritidis PT4
intergenic SNPs	3555	4516	4600	3509
Synonymous SNPs	23398	22180	22131	23151
Non-synonymous SNPs	6110	5646	5623	6173

Serotype clusters: The 217 strains tested were grouped by agglomerative clustering method into five clusters, each representing Typhimurium, Dublin, Typhi, Enteritidis and other serotypes.



Discriminating within phage profiles: strains within Enteritidis were also clustered into distinct groups largely corresponding to major phage types.



Deletion profiles correlated phage-type specific patterns:
In addition, deletion analyses also correlated phage-type specific patterns

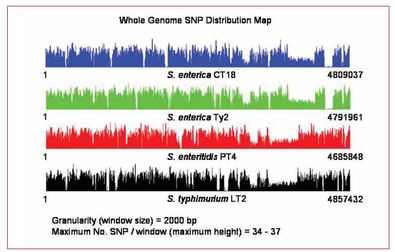
Deletion regions	No. of consecutive SNPs	Size (kb)	Starting position (CT18)	Specificity
Rhs-family protein	10	2.6	31287	All strains in PT30 cluster
monooxygenase	22	3.1	1603200	Both strains are PT9c
Membrane transporter	13	1.5	1609275	All strains of PT9c and RM4635
phage ST64B	11-43	1.2-20.8	2020789	All strains of PT9c and RM4635 (long deletion)
phage ST64B	41-42	8.8 - 20	2020848	All strains in PT8 cluster, ST2606 has short deletion
phage ST64B	21	6.3	2025535	All strains in "other" cluster (including PT3, PT33 and PT13)
phage ST64B	15-19	12.7-13	2027830	All strains of PT9c and RM4635
phage ST64B	15-17	11-12.5	2028348	All strains of PT8 cluster
phage ST64B	25-26	7.25	2044176	10 stains in PT30 cluster and A7 (PT8 cluster)
Fels-1 prophage	11	7.1	2050286	All strains of PT33 in "other" cluster

SNPs distribution on reference genomes and in functional groups:

1) Distribution of SNPs in COG (*Salmonella enterica* CT18) is similar to that of genes

COG_function	%SNP	%sSNP	%gene
RNA processing and modification	0.02	0.02	0.02
chromatin structure and dynamics	0	0	0.02
energy production	8.3	8.7	6.5
cell division	0.97	0.94	1.18
amino acid metabolism	12.2	12.3	9.6
nucleotide metabolism	2.2	2.3	1.9
carbohydrate metabolism	9.4	9.4	8.6
coenzyme metabolism	3.7	3.8	3.8
lipid metabolism	2.2	2.2	2.1
translation	3.2	3.3	4.5
transcription	6.3	6.3	7.65
DNA replication or repair	3.95	4.1	4.96
cell wall/membrane biogenesis	5.9	5.98	5.6
cell motility	2.4	2.3	2.7
posttranslational modification	3.3	3.4	3.9
inorganic ion metabolism	7.2	7.3	5.99
secondary metabolites biosynthesis, transport and catabolism	1.87	1.8	1.8
general function prediction only	11.9	11.6	12.8
function unknown	5.4	5.02	7.8
signal transduction	5.3	5.2	4.26
intracellular trafficking and secretion	2.4	2.3	3.0
defense mechanism	1.6	1.65	1.1
extracellular structures	0.06	0.05	0.046

2) Distribution of total SNPs along reference genomes shows some low density regions



3) Nonsynonymous SNPs have more uneven distribution

genome	Maximum/average Ratio	
	nsSNP	sSNP
S. typhi CT18	6.5	3.4
S. typhi Ty2	7.8	3.5
S. typhimurium LT2	6	3.3
S. enteritidis PT4	5.8	3.1

Conclusion:

1) Molecular signatures based on single nucleotide polymorphism genotyping are a powerful tool for grouping *Salmonella enteritidis* strains for diagnostics and can be used to study the evolution of this pathogen.
2) The genotyping methods can also detect deletions, some of which are group specific, and can be used as diagnostic or confirmatory tools.
3) There are no obvious concentrations of SNPs in any COG functional groups. The identified SNPs appeared to be relatively evenly distributed throughout the genomes of previously sequenced strains (except small regions on Typhi genomes), although the distribution of nsSNPs has more variation.

References:

1. Chalker, R.B. and M. J. Blaser. 1988. A review of human *salmonellosis III*. Magnitude of *Salmonella* infection in the United States. Rev. Infect. Dis. 10:111-124.
2. McClelland et al. 2001. Complete genome sequence of a multiple drug resistant *Salmonella enterica* serovar Typhimurium LT2. Nature. 413:852-856.
3. Deng et al. 2003. Comparative genomics of *Salmonella enterica* serovar Typhi strains Ty2 and CT18. J. Bacteriol. 185:2330-2337.
4. Parkhill et al. 2001. Complete genome sequence of a multiple drug resistant *Salmonella enterica* serovar Typhi CT18. Nature. 413:848-852.
5. Parry, C. M., T. T. Hien, G. Dougan, N. J. White and J. J. Farrar. 2002. Typhoid fever. N. End. J. Med. 347:1770-1782.
6. Patrick, M. E., P. M. Adcock, T. M. Gomez, S. F. Altekruse, B. H. Holland, R. V. tauxe and D. L. Swerdlow. 2004. *Salmonella* Enteritidis infections, United States, 1985-1999. Emerg. Infect. Dis. 10:1-7.

Acknowledgements:
This study was funded by NIAID. This work was performed under the auspices of the US Department of Energy by the University of California, Lawrence Berkeley National Laboratory under Contract No. DE-AC03-76SF00098. We thank S. Seybold for graphic support.