DNA Repeat Atlases

Dave Ussery
Comparative Microbial Genomics Workshop
Pathumthani, Thailand
9 March, 2005
Questions

1. What is the biological significance of repeats?

2. List three levels of regulation of gene expression in bacteria.

3. Why might the amino acid usage be so similar for most of the fungal genomes sequenced so far?
Outline

DNA Repeats in Bacterial Genomes

- Introduction
- Local Repeats
- Global Repeats
The DNA sequence contains information.

But what kind of information?
**DNA Symmetry Elements Defined**

What are symmetry elements?

**REPEATS in DNA Sequences**

**Direct repeats**

5' **GGGAATNNGGGAAT** 3'

3' **CCCTTANNCCTTA** 5'

**Inverted repeats**

5' **GGGAATNNATCCC** 3'

3' **CCCTTANNTAAGGG** 5'
REPEATS in DNA Sequences

Direct repeats

5’ GGGAAATNNGGGAAT 3’
3’ CCCTTANNCCTTA 5’

Direct Repeat --> Slipped strand structure
Inverted repeats

5' GGGAAATNNATTC 3'
3' CCCTTANNTAAGGG 5'

Inverted repeat --> Cruciform
Mirror Repeat

5' GGAAAGAGAAGGGAAGAGAAAGG 3'
3' CTTTTCTCTTCCCTTCTCTTTTCC 5'

Py · Pu · Py

Mirror repeat --> Hy3-type intramolecular triplex
Everted Repeat

5' GGAAAGAGCAAAGCCTTTTCTCTTTCC
3' CCTTTCTCTTTCCGAAGAGAAG

Parallel-stranded DNA

5' ATTAAATTAAAAATTTTTTTT
3'

5' TAATTAAATTTTTTTAAAAAA
3'
Inverted repeat --> Cruciform

Direct Repeat --> Slipped strand structure

Mirror repeat --> Hy3-type intramolecular triplex

Parallel-stranded DNA
Microsatellite instability regulates transcription factor binding and gene expression

Patricia Martin*,†, Katherine Makepeace*, Stuart A. Hill‡, Derek W. Hood*, and E. Richard Moxon*

*Molecular Infectious Diseases Group, University of Oxford Department of Paediatrics, Weatherall Institute of Molecular Medicine, John Radcliffe Hospital, Oxford OX3 9DS, United Kingdom; and ‡Department of Biological Sciences, Northern Illinois University, DeKalb, IL 60115

Edited by John R. Roth, University of California, Davis, CA, and approved January 21, 2005 (received for review September 14, 2004)

Microsatellites are tandemly repeated simple sequence DNA motifs widely prevalent in eukaryotic and prokaryotic genomes. In pathogenic bacteria, instability of these hypermutable loci through slipped-strand mispairing mediates the high-frequency reversible switching of phenotype expression, i.e., phase variation. Phase-variable expression of NadA, an outer membrane protein and adhesin of the pathogen Neisseria meningitidis, is mediated by changes in the number of TAAA repeats located upstream of the core promoter of nadA. Here we report that loss or gain of TAAA repeats affects the binding of the transcriptional regulatory protein IHF to the nadA promoter. Thus, phase-variable transcription of nadA potentially incorporates interplay between stochastic (mutational) and prescriptive (classical) mechanisms of gene regulation.

Methods

N. meningitidis Strains. Strain 64/96 presents (TAAA)₄, strain BZ83 presents (TAAA)₅, strain CU385 presents (TAAA)₆, strain ISS759 presents (TAAA)₈, strains MC58, SB25, F6124, and 90/18311 present (TAAA)₉, strain 92001 presents (TAAA)₁₀, and strain BZ169 presents (TAAA)₁₂ in the repeat tract associated with the nadA gene. The strains were described in refs. 9 and 13.

DNA Constructs and Mutant Strains. Plasmid pΔ221-108 contains a 930-bp DNA fragment containing the NMB1993 gene of N. meningitidis strain MC58 amplified by PCR with primers 1993-SacI and 1993-AflIII (Table 1), and then a 1.3-kb DNA fragment containing the nadA gene of strain MC58, its core promoter and the TAAA repeat tract amplified by PCR with primers 1994-AflIII.1 and 1994-SmaI. Then the 1.25-kb kanamycin-resistance cassette is found, excised from the pUC4K vector (Pharmacia) by digestion with BamHI, followed by a 475-bp DNA fragment containing the NMB1995 gene of strain MC58 amplified by PCR with primers 1995-BamHI and 1995 (Table 1). Plasmid pΔ221-50 contains the 503-bp DNA fragment containing the IHF site.

N. meningitidis is a Gram-negative encapsulated bacterium predominantly found in the human nasopharynx, where it can be a constituent of the normal microbial flora.
that was identical in length in SIRU15, but differed by 26% from the consensus sequence (Fig. 3). N315 and Mu50 have been shown to be clonal (Kuroda et al., 2001), and investigations of their SIRUs show that they have the same number of repeat units in all loci. They also contain the same base changes in all repeat units. Similarity in the base changes observed in the repeat units has been seen in other SIRU repeat units, for example in SIRU16 of NCTC8325 and COL.

The repeat units were true tandem repeats in all SIRUs, with no intervening sequence between the repeats, apart from SIRU01. SIRU01 had either 1 or 2 bp between the first and second repeats in all genomes. When sizing a fragment using a gel-based method, the presence of 1 or 2 bp does not change the apparent number of copies present.

The published S. aureus sequences have been annotated, and we were able to establish the position of the SIRUs.

Fig. 2. Alignment of SIRU05 60 bp repeat units in the seven sequenced genomes. Deletion of a repeat is represented by a dash; a nucleotide present in all repeats is represented by a dot; repeat units are located within square brackets; flanking regions are underlined.

Fig. 1. Genome atlas of MRSA strain N315. The genomic position of each of the SIRUs is represented on the grey bands by red (positive DNA strand) and blue (negative DNA strands) markers. Between SIRU07 and 13, the genes on the forward and reverse strands are shown in red and blue respectively. In the centre circle, the number of bases around the genome is shown; 0M, the origin of replication.

Fig. 5. Dendrogram deduced from the cluster analysis of the EMRSA strains. PFGE profiles, SIRU profiles and corresponding MLST types are shown. The box highlights SIRU profiles that differ from each other by a maximum of two loci.
Inverted repeat --> Cruciform

Direct Repeat --> Slipped strand structure

Mirror repeat --> Hy3-type intramolecular triplex

Parallel-stranded DNA
phiX174
5,386 bp
Extensive DNA Inversions in the
B. fragilis Genome Control
Variable Gene Expression

Ana M. Cerdeño-Tárraga,1 Sheila Patrick,2* Lisa C. Crossman,1
Garry Blakely,3 Val Abratt,4 Nicola Lennard,1 Ian Poxton,5
Brian Duerden,6 Barbara Harris,1 Mike A. Quail,1 Andrew Barron,1
Louise Clark,1 Craig Corton,1 Jonathan Doggett,1
Matthew T. G. Holden,1 Natasha Larke,1 Alexandra Line,1
Angela Lord,1 Halina Norbertczak,1 Doug Ormond,1 Claire Price,1
Ester Rabinowitsch,1 John Woodward,1 Bart Barrell,1
Julian Parkhill1*

The obligately anaerobic bacterium Bacteroides fragilis, an opportunistic
pathogen and inhabitant of the normal human colonic microbiota, exhibits
considerable within-strain phase and antigenic variation of surface compo-
nents. The complete genome sequence has revealed an unusual breadth (in
number and in effect) of DNA inversion events that potentially control
expression of many different components, including surface and secreted
components, regulatory molecules, and restriction-modification proteins.
Invertible promoters of two different types (12 group 1 and 11 group 2)
were identified. One group has inversion crossover (fix) sites similar to the hix
sites of Salmonella typhimurium. There are also four independent intergenic
shufflons that potentially alter the expression and function of varied genes.
The composition of the 10 different polysaccharide biosynthesis gene clusters
identified (7 with associated invertible promoters) suggests a mechanism of
synthesis similar to the O-antigen capsules of Escherichia coli.
Fig. 1. Examples of invertible regions in the *B. fragilis* genome (18). (A) Restriction modification intergenic shufflon: restriction-modification (R/M) complex genes, gray boxes; other genes, open boxes; potential *hsdS* DNA binding modules, hatched boxes; different inverted repeats at the inversion ends, light gray triangles. (B) Inversion of large segments of DNA through large inverted repeats (black triangles) brings alternative outer membrane protein genes (gray boxes) downstream of an invertible promoter (gray diamond). (C) Local inversion through inverted repeats (black triangles) fuses silent alternative outer membrane protein gene cassettes (hatched boxes) to a fixed promoter and translation start.
Fig. 1. Quasipalindromes and quasipalindrome correction mutations. (a) A model quasipalindrome (i.e., an imperfect inverted repeat). The left half contains a CC dinucleotide, whereas the right half contains a TT dinucleotide (denoted by XX above the sequences). In addition, the left half contains the sequence GAAGAA (ΔΔΔΔΔΔ above the sequence). This sequence is missing in the right half, and its position is denoted by the sequence above the brackets pointing to the site where from which the sequence is missing. (b) The top strand of the quasipalindrome is shown in its hairpin conformation in which the 2-bp mismatch and the GAAGAA loop are evident. (c,d) Left–left (c) and right–right (d) quasipalindrome correction, in which perfect inverted repeats have been created. In (c) the bases that are inserted are shown in orange.
Fig. 2. Models for quasipalindrome correction. The quasipalindrome is represented as the shaded rectangles: A is a short arm, and B is a long arm. In the
Fig. 3. The frequencies of quasipalindromes and perfect inverted repeats were calculated, as described in the text. Oqp, observed quasipalindromes, as a percentage of the whole genome; Eqp, expected quasipalindromes, based on the dinucleotide base composition; Op, observed palindromes, as a percentage of the whole genome; Ep, expected palindromes, based on the dinucleotide base composition. The red line is where Oqp/Eqp is equal to Op/Ep. Open green circles indicate Bacteria, filled blue circles indicate Archaea. The orange triangle is *Escherichia coli*. 
Fig. 4. Fraction of the genome containing local quasipalindromes. (a) The observed (orange) and the predicted (green) fraction of the genome that is part of a perfect inverted repeat, plotted against the $A + T$ content of the genome. A ‘perfect palindrome’ for the purposes of this analysis is an exact match of any 7-bp piece of DNA within a 3-bp window. (Thus, these repeats can have asymmetric centers of up to 16 bp.) (b) The observed (orange) and the predicted (green) fraction of the genome that is part of a quasipalindrome, plotted against the $A + T$ content of the genome. Quasipalindromes were calculated as the best match of a 7-bp window in a 30-bp length of DNA. The cut-off value for scoring a quasipalindrome was 80%, which allows one mismatch in the 7-bp window. To calculate the expected value for quasipalindromes and perfect inverted repeats (in (a) and (b)), 20 random DNA sequences with the same nucleotide composition and length were generated and analysed in the same fashion as the entire genome. The values are mean values with a relative standard deviation of $\sim 0.5\%$. 
Fig. 5. Cruciform atlas for *Bacillus anthracis* pX01. The cruciform atlas was constructed from the pX01 GenBank file (AF065404). The quasipalindromes and perfect palindromes are calculated as described in the text, and other repeats are calculated as before [19]. The genes containing insertion sequence (IS) elements are shown in red. Abbreviations: CDS, coding sequence; avg, average.
Summary

1. DNA symmetry elements code for many DNA structures.
   
   At a local level:
   - Direct Repeats -> slipped mispaired DNA
   - Inverted Repeats -> cruciforms
   - Mirror Repeats -> triple-stranded DNA
   - Everted Repeats -> parallel stranded DNA

2. Cruciforms are often found at the end of genes. “Imperfect” cruciforms can be corrected to “perfect” cruciforms, via templating.

3. Repeats can lead to diversity.