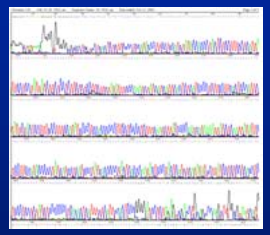
REPSA: Making DNA Binding Specificity Simple

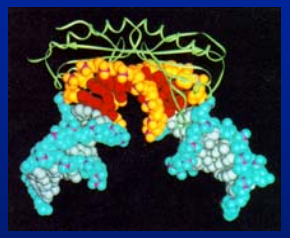
Michael Van Dyke, PhD
Western Carolina University
Cullowhee, NC

Paul Hardenbol, Ph.D.

- 1989, B.S., Microbiology, UT-Austin
- 1996, Ph.D., Biomedical Sciences, UT-Houston
- 1996–9, Postdoctoral Fellow, Stanford U. Sch. Med.
- 1999–01, Director of Mass Spectrometry facility, Stanford
- 2002–5, Chief Technology Officer, Parallele BioScience
- 2006–8, V/P, Parallele BioScience Div., Affymetrix
- 2009–, Senior Scientist, Pacific Biosciences

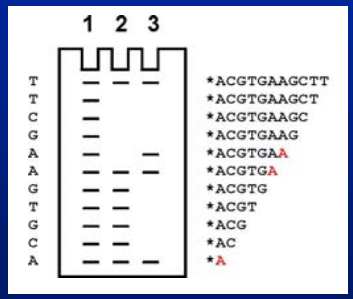
DNA is a Heterogeneous Target





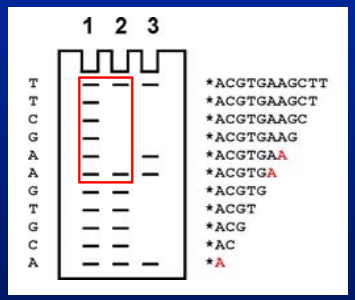
- Many ligands recognize specific DNA sequences/structures
- Important to understand ligand-DNA interaction specificity

A Footprinting Primer



- Singly end-labeled probe
- “Single-hit” chemical modification or enzymatic digestion
- Population of different length labeled DNAs
- Separate by high resolution gel electrophoresis
- Footprint = “gap” in otherwise uniform pattern

A Footprinting Primer

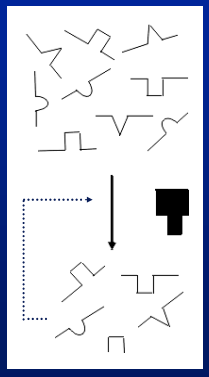


- Singly end-labeled probe
- "Single-hit" chemical modification or enzymatic digestion
- Population of different length labeled DNAs
- Separate by high resolution gel electrophoresis
- Footprint = "gap" in otherwise uniform pattern

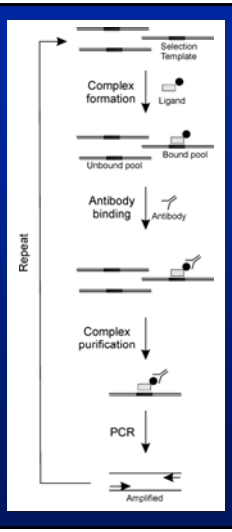
DNA Sites: A Number of Combinations

- Different ligands interact with different lengths of DNA
  - Small molecule (distamycin) 5 base pairs
  - protein (TBP) 8 base pairs
  - DNA (triplex) 19 base pairs
- The number of all possible sequence combinations (P) for a given length of DNA (n) , $P = 4^n/2$
  - $4^5/2 = 512$; $4^8/2 = 32{,}768$; $4^{19}/2 = 137$ billion
- Typically, only 100-200 base pairs can be examined per footprinting experiment

A Combinatorial Primer



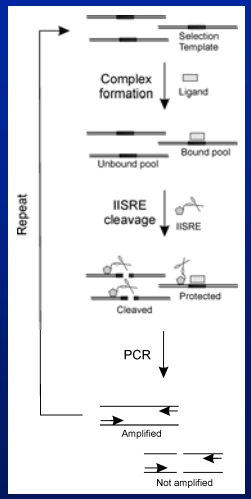
- Start with a (large) population of different possibilities
- Devise a scheme that allows selection of a subpopulation that has desired properties
- Use amplification methods to increase selected population to useable amounts
- Repeat process until a uniform subset with desired properties is obtained



CASTing: A Conventional Combinatorial Method

- Binding
  - Solution, optimal conditions
- Selection
  - Altered properties of ligand-DNA complex (*e.g.*, electrophoretic mobility, hydrophobicity) and/or affinity methods
  - Requires physical separation of complexes
- Amplification
  - PCR

REPSA: A Novel Combinatorial Method



- Binding
  - Solution, physiological conditions
- Selection
  - Resistance to type IIS restriction endonuclease cleavage
  - Enzymatic inhibition, not physical separation
- Amplification
  - PCR

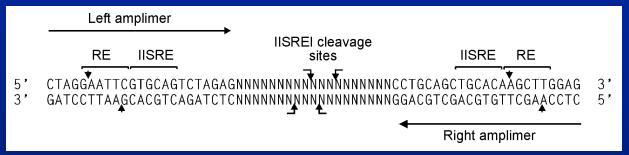
Type IIS Restriction Enzymes

ENase II (*Eco*RI)

ENase IIS (*Fok*I)

- Type IIS restriction endonucleases have cleavage sites located at a fixed distance from their recognition site (1 - 20 nt)
- Recognition sites are typically nonpalindromic and 4 - 7 bp long
- Cleavage occurs without sequence specificity

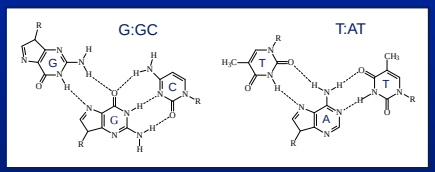
REPSA Selection Template

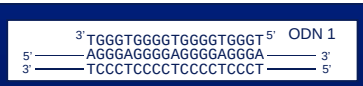


- Central region composed of randomized sequence
- Type IIS recognition sites located in defined flanks; their cleavage sites positioned within the randomized center
- Additional restriction endonuclease sites for subcloning
- Sufficient length flanks for PCR amplification

Triple-helical DNA (Triplex)

Purine Motif





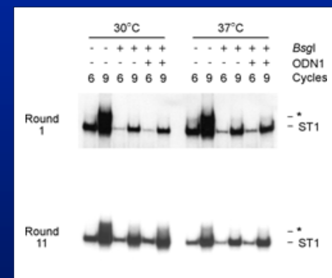
- What is the consensus duplex target for a given TFO?

Triplex/REPSA Experimental

- Selection template: ST1

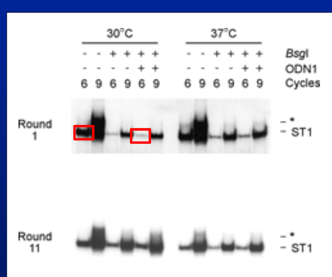
- Binding rxn: 40 mM HEPES (8.4), 12 mM Mg^{2+}, 5 µM ODN1, 23 nM ST1; 2 h @ 37 ºC
- Cleavage rxn: 4 U Bsg I, 30 min @ 37 ºC
- PCR: 94 ºC/1 min, 50 ºC/3 min; 6 or 9 cycles
- REPSA: 11 Bsg I selections

Identification of an Emergent Population



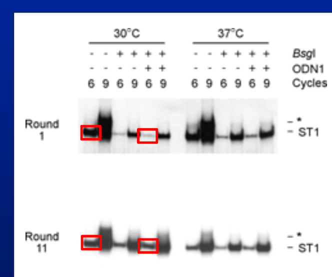
- Emergent population = one that is cleavage resistant
- Evident directly (REPA) or indirectly (PCR amplification)
- Should be ligand-dependent

Identification of an Emergent Population



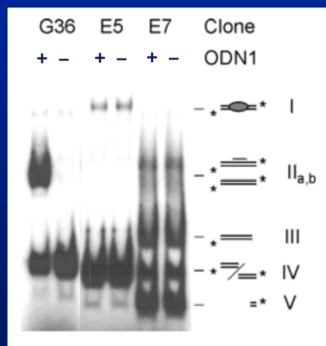
- Emergent population = one that is cleavage resistant
- Evident directly (REPA) or indirectly (PCR amplification)
- Should be ligand-dependent

Identification of an Emergent Population



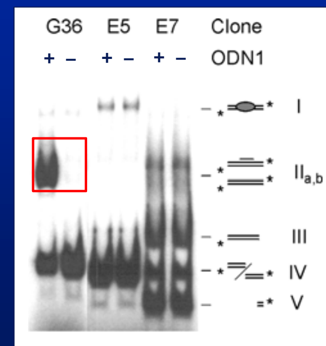
- Emergent population = one that is cleavage resistant
- Evident directly (REPA) or indirectly (PCR amplification)
- Should be ligand-dependent

REPA Analysis of Selected Clones



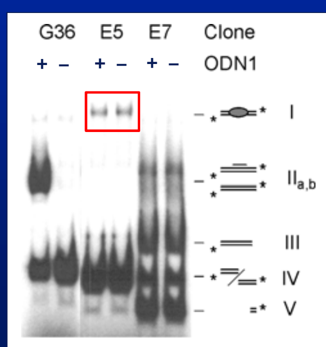
- Three types of *Bsg* I cleavage products observed:
  - (G36) ODN 1-dependent cleavage inhibition
  - (E5) ODN 1-independent pattern
  - (E7) complex, ODN 1-independent pattern

REPA Analysis of Selected Clones



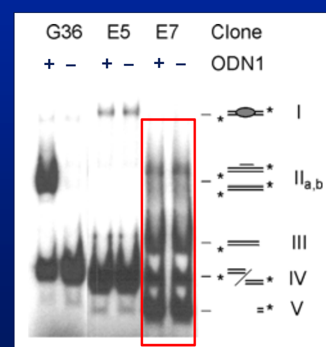
- Three types of *Bsg* I cleavage products observed:
  - (G36) ODN 1-dependent cleavage inhibition
  - (E5) ODN 1-independent pattern
  - (E7) complex, ODN 1-independent pattern

REPA Analysis of Selected Clones



- Three types of *Bsg* I cleavage products observed:
  - (G36) ODN 1-dependent cleavage inhibition
  - (E5) ODN 1-independent pattern
  - (E7) complex, ODN 1-independent pattern

REPA Analysis of Selected Clones



- Three types of *Bsg* I cleavage products observed:
  - (G36) ODN 1-dependent cleavage inhibition
  - (E5) ODN 1-independent pattern
  - (E7) complex, ODN 1-independent pattern

Triplex-forming Sequences

```
[37°C Selection]
E26       T A G G G A G G G G A G G G G G C C G C
E27       T A G G G A G G G G A G G G G T T A C C
E32  GA   A G G G A G G G G A G G G G T T G T C C
E37       G G C A T A G G G A G G G G A G G G A CA
G1        C A G G G G G G G A G G G G G T T A A AGA
G3        A G G G A G G G G A G G A C A T T T C
G8        A G G G A G G G G A G G C C T C C T C
G10       A G G G A G G G G A G G A C C C C C T
G11 GCAG  A G G G A G G G G G G G G A G C T C T
G12 ACAA  A G G G C G G G G A G G G G T C T C T
G13       A G G G A G G G G A G G G G G A G G G ATG
G20       T G G G A G G G G A G G G G T A G A C
G21       A G G G A G G G G A G G G G T C T A G TC
G22       T G G G A G G G G A G G G G C T A A T
G23       A G A G A C G G G A G G G G A G G G A GGTT
G26   TT  A G G G A G G G G A G G A T G G A C C
G28       A A G G G A G G G G A G G G T C C A T C
G29   TT  A G G G A G G G G A G G G C T C C C C
G33       G G G G A G G G G A G G G G G C A C T
G34       A G G G C G G G G A G G G G G C C A A AAA
G36       A T C G A A G G G A G G G G A G G G A

PuD       A G G G A G G G G A G G G G A G G G A
Cons      A G G G A G G G G A G G G - - - - - -
```

- ODN 1-dependent cleavage-resistant sequences
  - Fit expected target, but only on its 5' side
  - Consensus is 13 bp, not 19 bp
  - "Mismatches" are infrequent, but G*(G→A)·T predominate

Purine-Motif Triplex Consensus Structure

- Pu-motif triplexes are only stable for approximately one turn of the duplex DNA target - out of register
- Limitation for achieving high sequence specificity in complex genomes

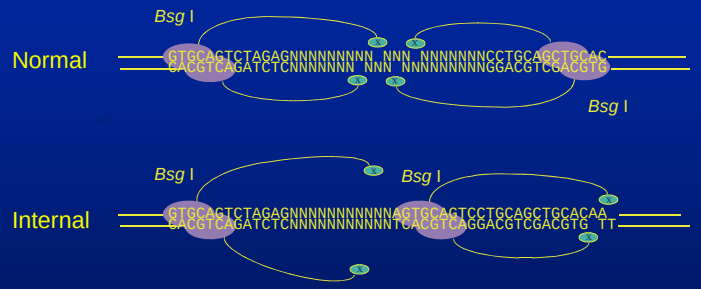
Bsg I Binding Sequences

```
          Bsg I                       G T G C A G
E3        T T T C T A A C C G T A G T G C A G T
E7        T C C T A C G A G T T A G T G C A G T
F26       T G T A A A A A A A A A G T G C A G T
17        T A T T G G C T T A C A G T G C A G A
18        G G A T A C T C G T T A G T G C A G T
110       A T A G G C A A A T T A G T G C A G T
111       A T A T T A G T G A T A G T G C A G T
113       T T T T C G C C T G T A G T G C A G T
115       T A T T T C T T A T T A G T G C A G A

      A   2 2 3 1 3 3 3 3 3 3 1 9 0 0 0 0 9 0 2
      C   0 1 1 1 1 4 2 3 1 0 1 0 0 0 0 9 0 0 0
      G   1 2 0 1 2 2 2 0 3 2 0 0 9 0 9 0 0 0 0
      T   6 4 5 6 3 0 2 3 2 4 7 0 0 9 0 0 0 0 7
Consensus - - - - - - - - - - - A G T G C A G T
```

- Complex, ODN 1-independent pattern
  - Consensus is 8 bp
  - Maps to 3´ edge of randomized cassette
  - Contains a *Bsg* I DNA-recognition site
- A better *Bsg* I site?

Type IIS Restriction Enzymes



- Internal Bsg I site interferes with internal cleavage, but cleaves flank yielding sufficient length for PCR

Unknown *B. sphaericus* Protein-DNA Binding Sequences

```
Inverted Repeat  T G G G A                 T C C C A

E5               T G G G A C T T T T A T T G T C C C A
G7               T G G G A T A A T C T C G G T C C C A
I3               T G G G A T A G G A A T T G T C C C A

           A     0 0 0 0 3 0 2 1 0 1 2 0 0 0 0 0 0 0 3
           C     0 0 0 0 0 1 0 0 0 1 0 1 0 0 0 3 3 3 0
           G     0 3 3 3 0 0 0 1 1 0 0 0 1 3 0 0 0 0 0
           T     3 0 0 0 0 2 1 1 2 1 1 2 2 0 3 0 0 0 0
   Consensus     T G G G A - - - - - - - - - G T C C C A
```

- ODN 1-independent pattern
  - Consensus is 5-6 bp on either edge of randomized cassette
  - Inverted repeat = classic dimeric protein binding site
- Binding site for unknown *B. sphaericus* protein?

Triplex REPSA: Conclusions

- Recognition between 13 bases on a G/T-rich oligonucleotide 3′ end and the duplex DNA were sufficient for purine-motif triplex formation
- The base triplets G:GC and T:AT are preferred; G:AT is the least inhibitory of the mismatches
- The preferred binding site for *Bsg* I is 5′ -AGTGCAGT-3′
- There exists a *B. sphaericus* protein in commercial *Bsg* I preparations that binds the palindromic consensus sequence 5′-TGGGANNNNNNNNTCCCA-3′

REPSA Examples

Purine-motif Triplex

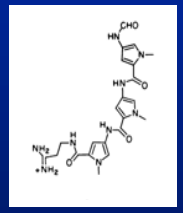
- **5′-AGGGAGGGGAGGNNNNNN-3′**
  3′-TCCCTCCCCTCCNNNNNN-5′
- Hardenbol & Van Dyke, 1996

TATA-box Binding Protein

- **5′-TATAAATA-3′**
  3′-ATATTTAT-5′
- Hardenbol *et al.*, 1997

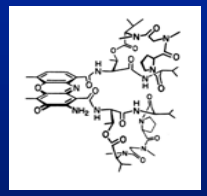
More REPSA Examples

Distamycin A



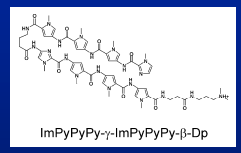
- **5′-AATTTTATT-3′**
  3′-TTAAAATAA-5′
- Hardenbol *et al.*, 1997

Actinomycin D



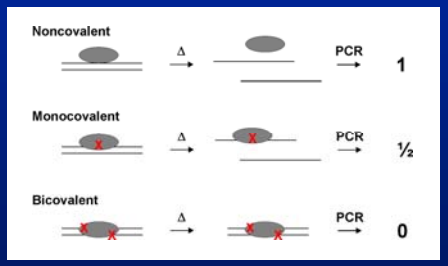
- **5′-TGCTGCA-3′**
  3′-ACGACGA-5′
- Shen *et al.*, 2001

Hairpin Polyamides



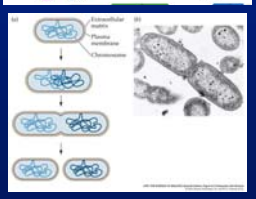
- **5′-WGWWCW-3′**
  3′-WCWWGW-5′
- Gopal & Van Dyke, 2003

REPSA of DNA-Modifying Ligands: A Special Problem



- Modifications: cleavage, cross-link, chemical alteration
- Modification interferes with PCR
- Amplification requires at least one intact strand

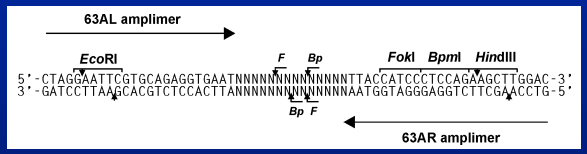
Molecular Function of the *E. coli* SlmA Protein



- 198-aa protein, contains putative N-terminal helix-turn-helix and C-terminal coiled-coil motifs. Dimeric DNA-binding protein?
- Genetically demonstrates a synthetic lethal phenotype with defective Min system strains
- Min involved in spatially controlling *E. coli* cytokinesis, specifically preventing Z-ring formation at cell poles; SlmA & nucleoid exclusion?

SlmA/REPSA Experimental

- Selection template: ST2-18

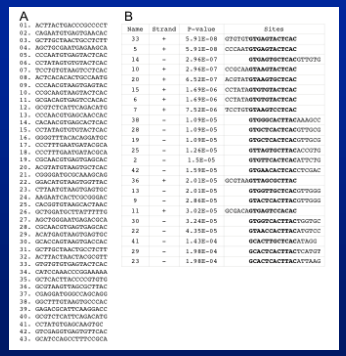


- Binding rxn: 10 mM Tris-Cl (7.9), 50 mM Na^+, 10 mM Mg^{2+}, 1 mM DTT, 40 μM SlmA, 10 nM ST2-18; 30 min @ 30 ºC
- Cleavage rxn: 0.5 U Fok I or Bpm I, 10 min @ 37 C
- PCR: 94 ºC/1 min, 50 ºC/3 min, 72 ºC/1 min; 6 cycles
- REPSA: 6 Fok I and 1 Bpm I selections

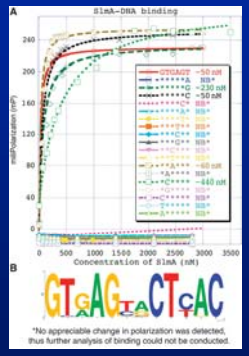
SlmA/REPSA: Emergent Population & Analysis

- Observed emergent population by convergence in DNA sequencing
- Subcloned – obtained 43 unique sequences
- Sequence analysis by Multiple Expectation Maximum for Motif Elicitation [MEME] (Bailey *et al.*, 2006)
- 23/43 sequences contain sites with high homology to a consensus 12-bp palidromic sequence (SBS) 5′-GTGAGTACTCAC-3′
- Importance of each nucleotide in SBS determined systematically using point mutants and fluorescence polarization

SlmA DNA-binding Sequence



Name	Strand	P-value	Sites
33	+	5.91E-08	GTGTGTGTGAGTACTCAC
5	+	5.91E-08	CCCAATGTGAGTACTCAC
14	–	2.94E-07	GTGAGTGCTCACGTTGTG
10	+	2.94E-07	CCGCAAGTAAGTACTCAC
20	+	4.52E-07	ACGTATGTAAGTGCTCAC
15	+	1.49E-06	CCTATAGTGTGTACTCAC
6	+	1.49E-06	CCTATAGTGTGTACTCAC
7	+	7.52E-06	TCCTGTGTAAGTCCTCAC
38	–	1.09E-05	GTGGGCACTTACAAAGCC
28	–	1.09E-05	GTGCTCACTCACGTTGCG
19	–	1.09E-05	GTGCTCACTCACGTTGCG
25	–	1.24E-05	GTTAGTGCTTACACGTTG
2	–	1.5E-05	GTGTTCACTCACATTCTG
42	–	1.59E-05	GTGAACACTCACCTCGAC
36	+	2.01E-05	GCGTAAGTTAGTGCTTAC
13	–	2.01E-05	GTGGTTGCTCACGTTGGG
9	–	2.86E-05	GTACTCACTTACGTTGGG
11	+	3.02E-05	GCGACAGTGAGTCCACAC
30	–	3.24E-05	GTGGTCACTTACTYGGTGC
22	–	4.35E-05	GTAACACACTTACATGTCC
41	–	1.43E-04	GCACTTGCTCACATAGG
29	–	1.98E-04	GCACTCACTTACTYATGT
23	–	1.98E-04	GCACTCACTTACATTAAG



- SBS = 5′-GTrAGyrCTyAC-3′

SlmA-DNA binding *in vivo*

- Putative SlmA binding sites on the *E. coli* chromosome were identified by Find Individual Motif Occurrences [FIMO] (Grant *et al.*, 2011)

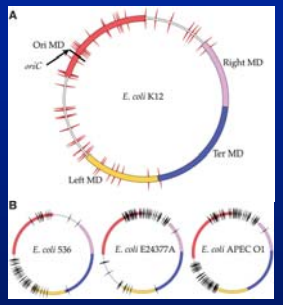
SlmA-DNA binding *in vivo*

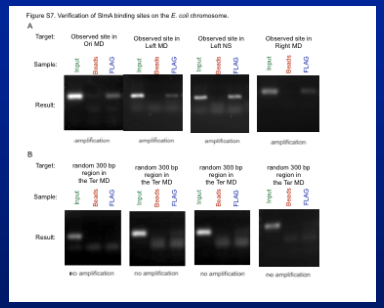
- Putative SlmA binding sites on the *E. coli* chromosome were identified by Find Individual Motif Occurrences [FIMO] (Grant *et al.*, 2011)
- Actual SlmA binding sites *in vivo* determined by ChIP-Seq and Motif Alignment & Search Tool [MAST] (Bailey & Gribskov, 1998). 50/52 contain consensus SBS!

SlmA-DNA binding *in vivo*

- Putative SlmA binding sites on the *E. coli* chromosome were identified by Find Individual Motif Occurrences [FIMO] (Grant *et al.*, 2011)
- Actual SlmA binding sites *in vivo* determined by ChIP-Seq and Motif Alignment & Search Tool [MAST] (Bailey & Gribskov, 1998). 50/52 contain consensus SBS!
- SlmA-binding sites did not map to the *E. coli* Ter macrodomain – consistent with a role in chromosome segregation & cell division

SlmA-DNA Complexes *In Vivo*





- Representative sites verified individually

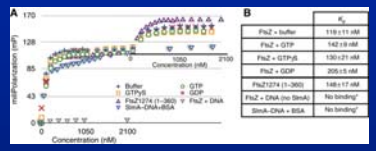
SlmA: Biological Function

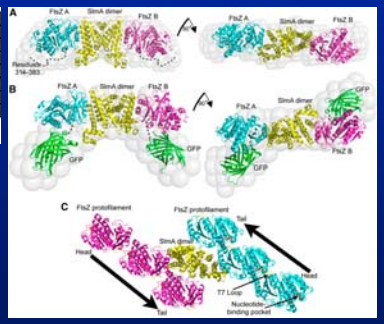
- SlmA binds the Z-ring-forming protein FtsZ, as determined by Fluorescence Polarization [FP] and Small-Angle X-ray Scattering [SAXS] analysis. Can simultaneously bind DNA!

SlmA: Biological Function

- SlmA binds the Z-ring-forming protein FtsZ, as determined by Fluorescence Polarization [FP] and Small-Angle X-ray Scattering [SAXS] analysis. Can simultaneously bind DNA!
- SlmA alters FtsZ polymerization (*parallel* → *antiparallel*), prevents formation of Z-rings

SlmA/DNA/FtsZ Complexes





- FtsZ still can polymerize antiparallel

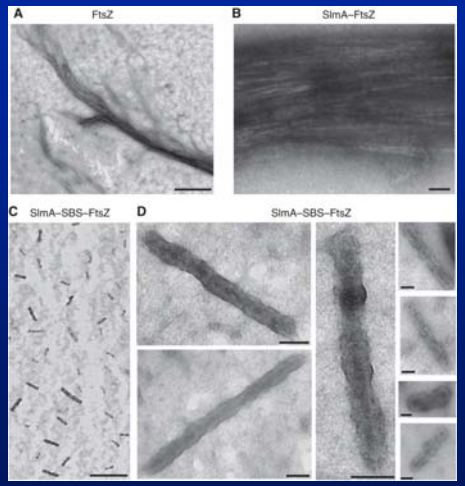
SlmA/DNA/FtsZ Complexes

Negative-stain EM shows:

A. FtsZ filaments

B. FtsZ/SlmA proto-filaments

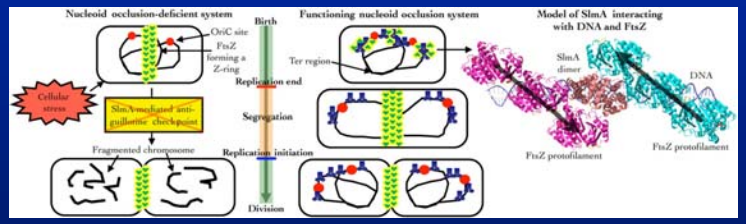
C. FtsZ/SlmA/DNA helices



SlmA: Biological Function

- SlmA binds the Z-ring-forming protein FtsZ, as determined by Fluorescence Polarization [FP] and Small-Angle X-ray Scattering [SAXS] analysis. Can simultaneously bind DNA!
- SlmA alters FtsZ polymerization (*parallel* → *antiparallel*), prevents formation of Z-rings
- Leads to a model of SlmA biological function
  - SlmA binds to non-Ter sequences
  - SlmA causes local formation of unproductive FtsZ assemblies
  - Functional FtsZ polymerization only occurs at end of replication over Ter macrodomain; leads to Z-ring formation and septation after chromosomes have segregated into daughter cells

SlmA: Biological Function



- SlmA required for nucleoid exclusion during cytokinesis

SlmA/REPSA Conclusions

- REPSA was successful in identifying the preferred binding sites of relatively little-known HLH protein SlmA
  Note: other methods were unsuccessful
- Consensus site and genome analysis led to hypothesis in nucleoid exclusion and cytokinesis
- Consensus sequence essential for subsequent demonstration of biochemical effects for SlmA-DNA on FtsZ polymerization.
- Sequence & structure yielded mechanism for SlmA biological function

REPSA Conclusions

- REPSA can identify the consensus DNA binding sites of nucleic acids, proteins and small molecule drugs
- Advantages of REPSA over other combinatorial methods:
  - Mild selection conditions
  - Mixed ligands
  - Uncharacterized ligands
  - Noncovalent and covalent ligands
  - Control over both DNA and ligand concentrations
  - Only combinatorial method that works for drugs
- Future: REPSA for identification and characterization of unknown DNA-binding proteins

Acknowledgements

Paul Hardenbol (*Sr. Scientist, Pacific Biosciences*)
Jo C. Wang (*Retired*)
Jing Shen (*Asst. Prof., UTMB*)
Gulshan Sunavala (*Asst. Prof., LSU-HSC*)
Xiaohong Yang (*Housewife*)
Vashisht Gopal Yennu Nanda (Instructor, UTMDACC)
Nam Tonthat (PhD student, Duke)

Robert A. Welch Foundation, American Cancer Society
NIH & US Army CDMRP BCRP