Genome Analysis

- Bacterial Genome sequencing
  - does this help us in the investigation of adaptive responses/regulatory systems?
- Genome Sequencing Projects
  - strategy & methods
  - annotation
- Comparative genomics
  - regulatory systems
- Functional genomics
  - transcriptome
  - proteome
  - genome-wide mutation
- Concentrate on strategy & ideas

Bacterial genome projects

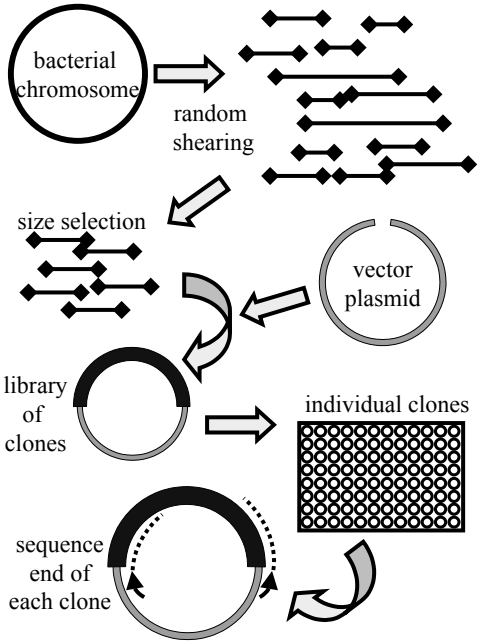
- Many completed:
  - *Haemophilus influenzae*
  - *Escherichia coli*
  - *Bacillus subtilis*
  - *Mycoplasma genitalium*
  - *Helicobacter pylori (x2)*
  - *Campylobacter jejuni*
  - *Treponema pallidum*
  - *Neisseria menigitidis*
  - *Neisseria gonnorhoea*
  - *Vibrio cholerae*
  - *E. coli O157*
  - and many more…(63 done, 137 ongoing Nov 01)
- Good link to projects:
  - http://www.tigr.org/tigr-scripts/CMR2/CMRHomePage.spl
  - http://www.ncbi.nlm.nih.gov/PMGifs/Genomes/bact.html
  - http://www.sanger.ac.uk/Projects/Microbes/

Why bother?

- piecemeal collection of sequenced genes
  - slow
  - costly
  - ever complete?
- genome project
  - rational approach
  - efficient and rapid
  - quality assurance
  - address novel questions
- problems/issues
  - ownership
  - strain choice
  - cost
  - approach
  - are these now academic?
- Post genomic era
  - Comparative genomics
  - Functional genomics

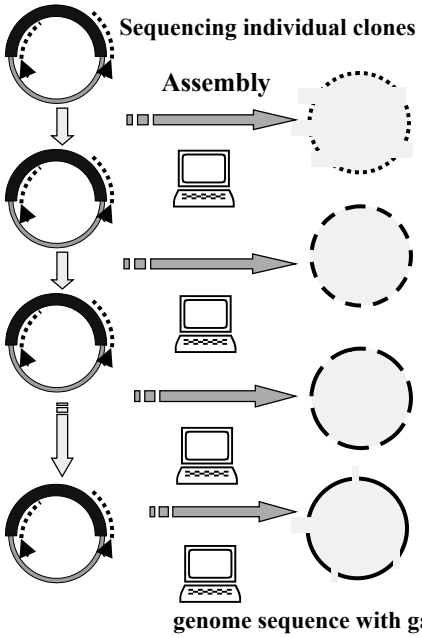
Genome sequencing strategy

- ***Strategy choice***
- large collaborative cosmid-based projects
  - now better suited for larger genomes
  - slow
- small insert shotgun approach
  - centralised
  - rapid and efficient
  - choice for bacteria
- ***Strain choice***
  - fresh isolate vs lab strain
  - clinical vs environmental
  - subsequent genetic analysis



2001-2002

E: 5



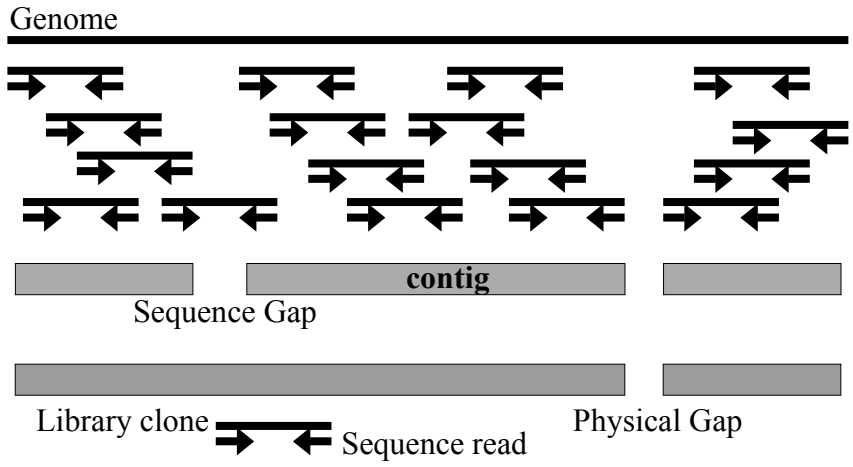
2001-2002

E: 6

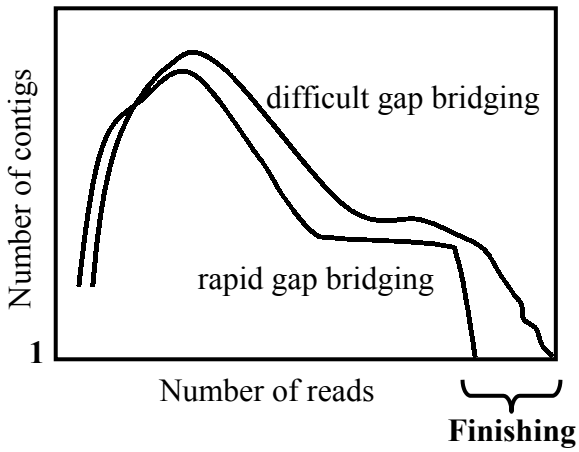
Just how much effort?

- individual sequencing reads accumulate
  - each read about 500bp
  - computing used to assemble reads
  - contiguous sequences called contigs
- Aim for 8-10 read coverage of genome for accuracy
- example:
  - *H.influenzae*
    - 19,687 templates
    - 24,304 reads assembled
    - 11,631,485 bp

Gaps



Bridging Gaps



- rise in contig number as amount of reads increases
- steady fall as accumulating sequence bridges gaps between contigs
- levels off as new reads more likely in known contig than gap
- start finishing

Finishing

- Why are gaps present?
- Gap bridging
  - sequence gaps
    - sequence gaps –choose appropriate clone and walk
  - physical gaps
    - alternative libraries (which?)
    - PCR across gap
- Mistakes/poor sequence
  - areas where sequence reads are less than 8-10
  - repeated sequences -rRNA
- closure and completion

Genome Annotation

- Find ORFs
  - look for ATG-Stop (+alternatives) over certain size
  - overlaps
  - computer based (“Glimmer” & “Orpheus”) and trained eye.
- ORF function
  - Search databases with predicted translated sequences – BLASTX
  - Consider level of similarity and context
  - Domain comparisons
    - Pfam/Prosite
- Other features

Finding regulators

- Annotation
  - sequence similarity
    - regulator families
  - domain matches
  - RR receiver, HtH etc
- ORFans
  - significant proportion of genome contains ORFs of unknown function
  - examples:
    - *H.influenzae*: 42%
    - *H.pylori*: 33%
    - *E.coli*: 38%
    - *M.tuberculosis*: 60% to 16%
  - number decreasing
  - some likely to be novel regulators

Functional genomics: regulation

- *Functional genomics* –ascribing gene function across a genome
- application to regulation
  - gene identification
  - expression pattern
  - mutant phenotype

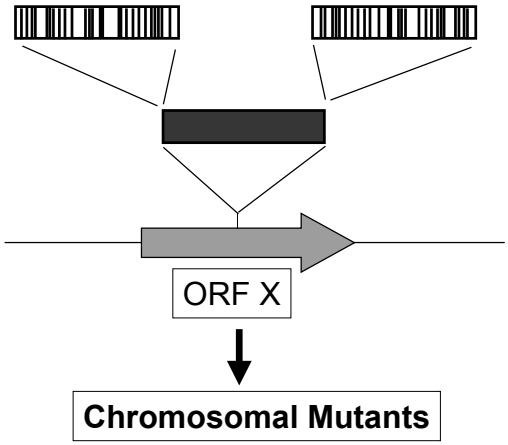
- Combination of approaches for finding regulators, stimulons and regulons:
- mass mutagenesis
- transcriptome
- proteome
- genomic SELEX

Mass Mutagenesis

- Mutate every ORF in genome
  - organism specific technology
- High throughput analysis of phenotype
  - need to analyse many 1000s of mutants under many conditions
- Signature-tagged technology
  - enables analysis of mutant pools
  - requires array technology for genome-wide projects
- Association on ORF with mutant phenotypes
- Regulators might be pleiotropic

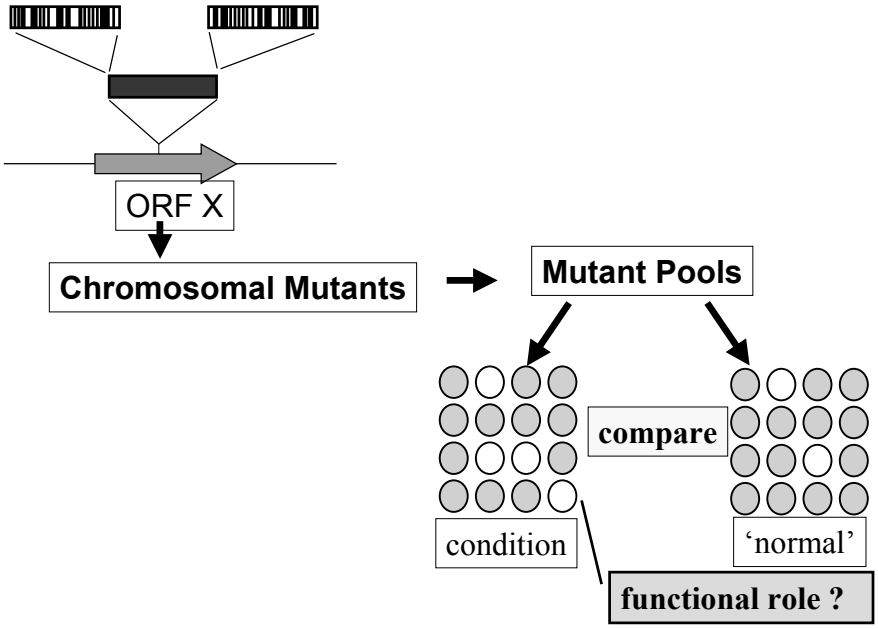
Signature Tagged

- Tags are short unique DNA sequences
- Tag linked to mutation
- Each individual mutant has unique tag
- Each mutant ORF has unique Tag



Arrays: micro and chip

- Microarrays
  - Glass slides with <10000 individual samples applied in *known* position
  - Use of robotics
  - Samples can be PCR products or oligos
  - example: oligos complementary to each unique Tag
  - example: oligo/PCR product complementary to each ORF
- Chip arrays
  - silicon based
  - >10,000 sequences
- Redundancy
- fluorescent labels



2001-2002

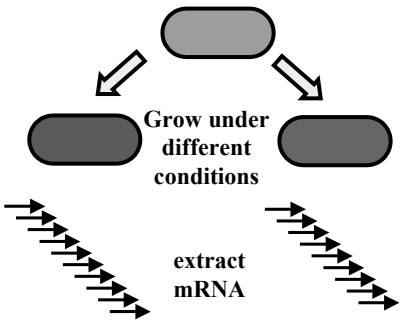
E: 17

Transcriptome

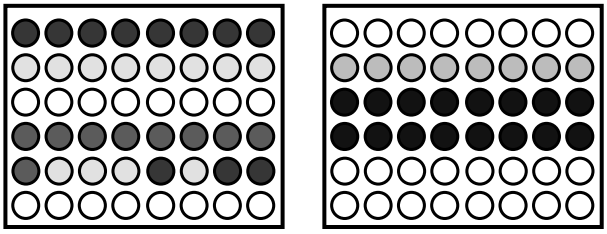
- Genome-wide determination of expression level of each ORF
- Gives information of Stimulons
- when expressed relates to role
- also assess mutants
- regulatory mutants will affect expression of several ORFs

2001-2002

E: 18



Probe array with labelled copy of mRNA



2001-2002

E: 19

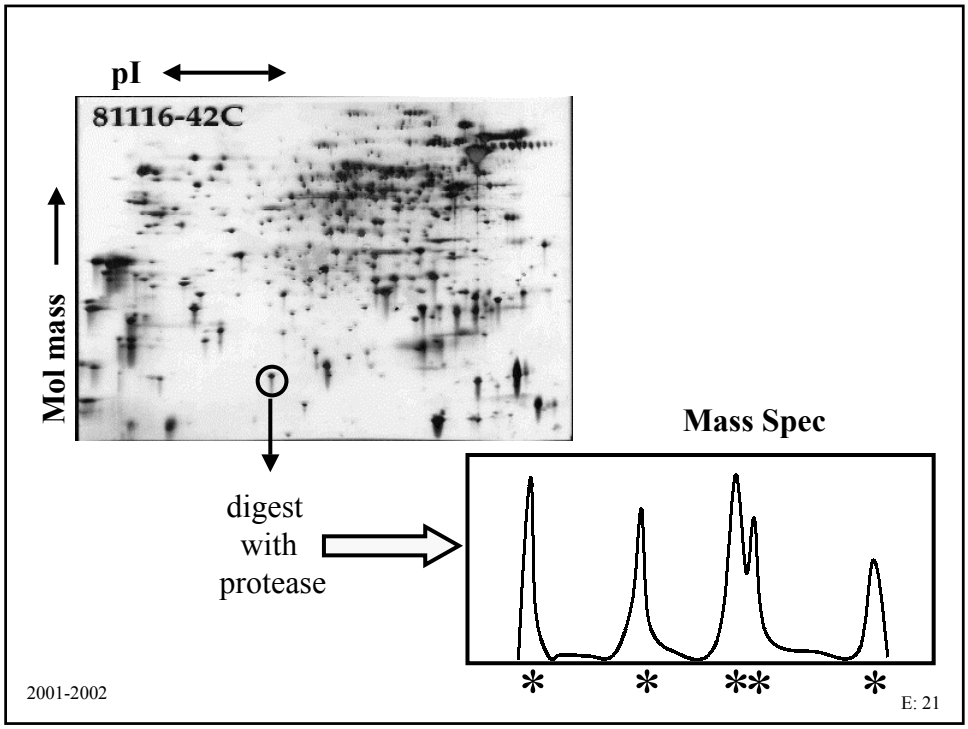
Proteome

- Genome-wide determination of protein expression
- Gives information stimulons
- protein expression linked to function
- assess mutants (regulatory mutants affect several proteins)

- Grow bacteria under defined conditions
- Extract proteins
- 2D-gel electrophoresis
- Protein spot identification
- Mass Spectrometry
- peptide size predictions from Genome data

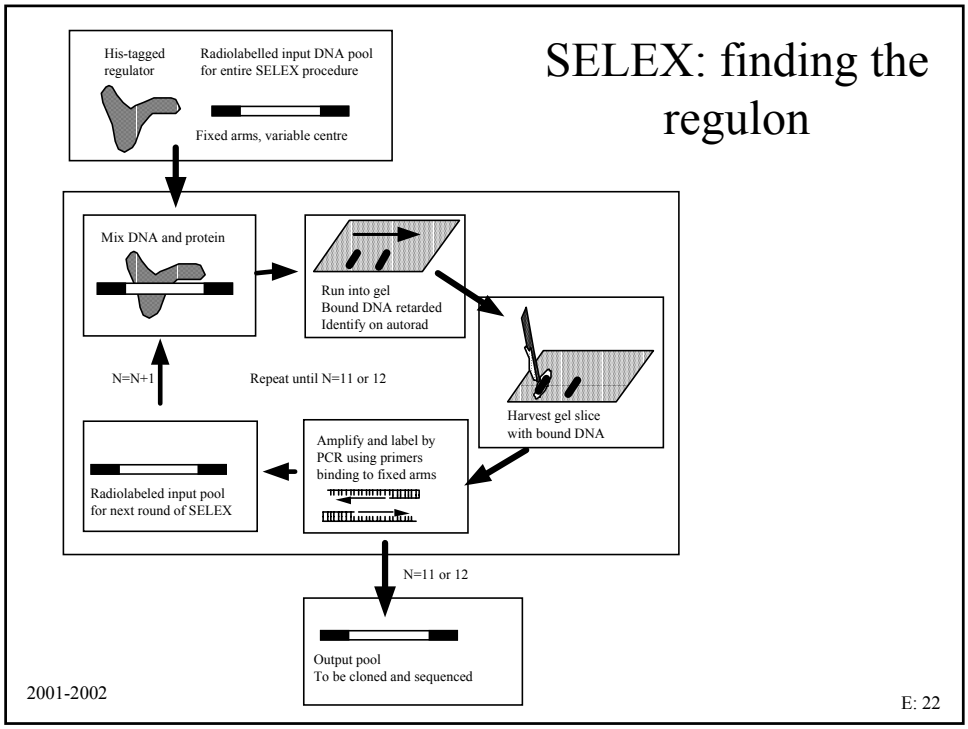
2001-2002

E: 20



2001-2002

E: 21



2001-2002

E: 22