CAP 5510: Introduction to Bioinformatics
CGS 5166: Bioinformatics Tools

Giri Narasimhan
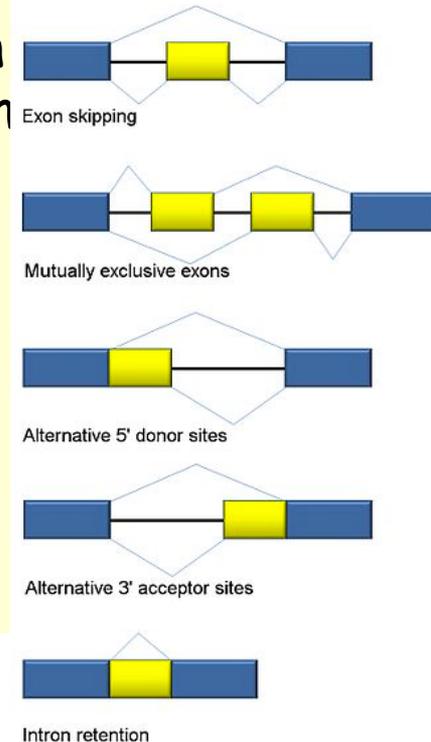
ECS 254; Phone: x3748

giri@cis.fiu.edu

www.cis.fiu.edu/~giri/teach/BioinfS15.html

Alternative Splicing

Alternative Splicing

❑ Regular transcription: exons are spliced together in order of appearance

❑ Alternative Splicing: Exons may be dropped or shuffled during splicing, causing alternatively spliced mRNA, giving rise to different proteins

❑ Short video:

- http://highered.mheducation.com/sites/9834092339/student_view0/chapter16/animation_-_exon_shuffling.html

❑ Increases the set of proteins coded for by the genes in Human genome may very well code for over 100K protein the number of coding genes is around 30K.

Exon skipping

Mutually exclusive exons

Alternative 5' donor sites

Alternative 3' acceptor sites

Intron retention

Small RNA

Small RNA

- ❑ What is **non-coding RNA**?
  - ● **Non-coding** "genes" are transcribed, but do not code for proteins
    - ➢ Types of non-coding genes include
      - ▪ tRNA
      - ▪ rRNA
      - ▪ **Small RNA**
  - ● Thus, they are not translated, but function as RNA molecules
- ❑ How can you find them?
  - ● You can find them as transcripts in the total mRNA
- ❑ How long are the small RNAs?
  - ● 50 – 250 nucleotides in length
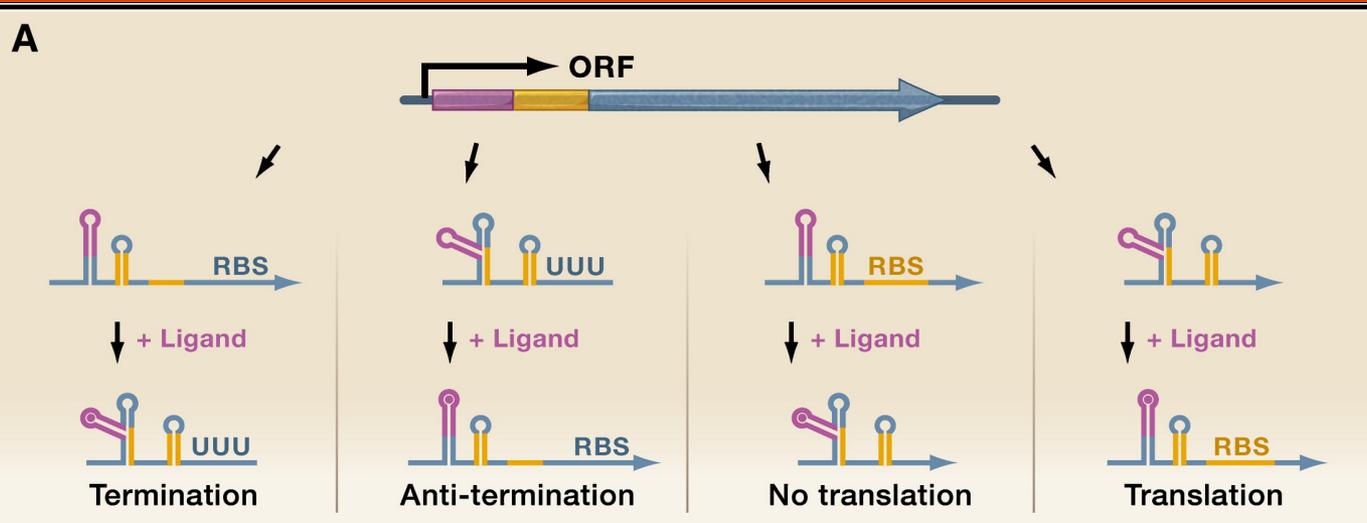- ❑ Functions of small RNA?
  - ● Mainly regulation

Functions of Non-Coding RNA

❑ tRNA and rRNA have **critical** specific functions

❑ **Interference**: bind to proteins to alter their function

- 🔴 **6S rRNA** binds to RNA polymerase and regulates overall transcription
- 🔴 **tmRNA** is involved in protein synthesis, including recycling of ribosomes,
- 🔴 **4.5S rRNA** regulates signal recognition particle (SRP), which is required for secretion of proteins
- 🔴 **RNase P** is involved in maturing tRNAs
- 🔴 Often involved in regulating translation of target RNAs

❑ **Interference**: bind to mRNA to "regulate" transcription & translation

- 🔴 DsrA and RprA both activate RpoS translation
  - ➢ This happens by base pairing to a region in leader sequence of RpoS mRNA and disrupting formation of hairpin to free up the ribosome loading site.
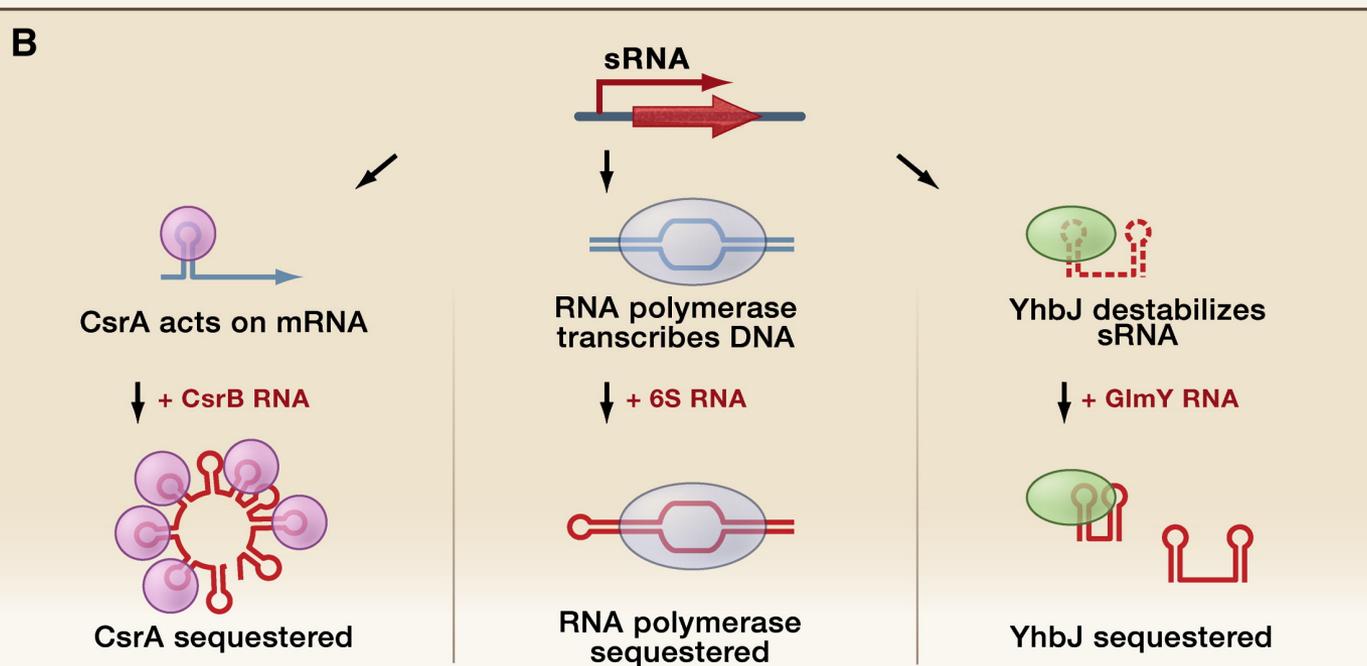- 🔴 OxyS inhibits RpoS translation

Waters, Storz, *Cell*, 136(4):615-28, 2009

❑ Bacteria possess numerous and diverse means of gene regulation using RNA molecules, including
- mRNA leaders that affect expression in cis,
- small RNAs that bind to proteins or
- small RNAs that bind to base pair with target RNAs, and
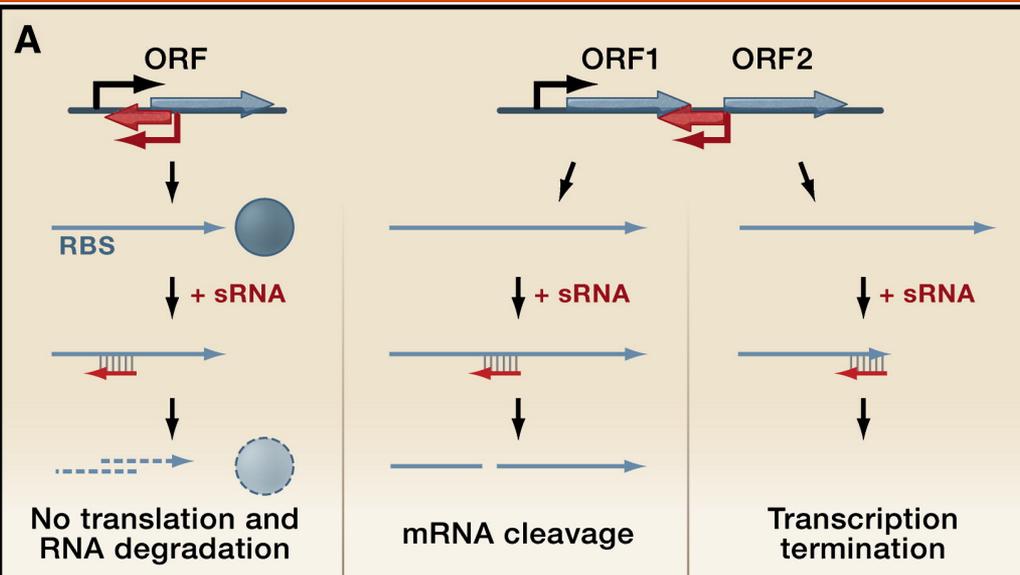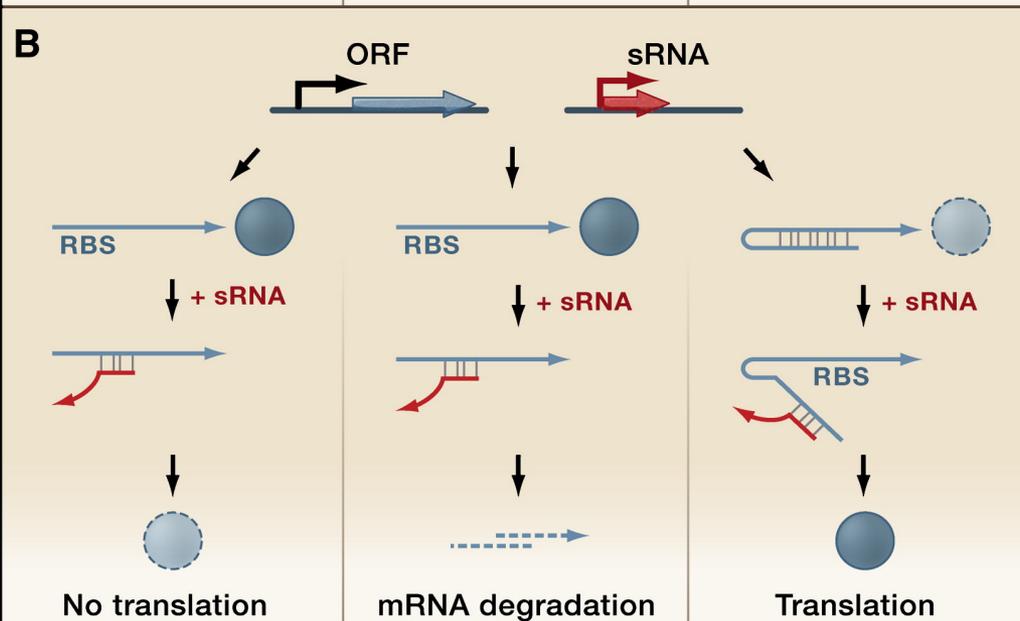- CRISPR RNAs that inhibit the uptake of foreign DNA.

CRISPR sRNA
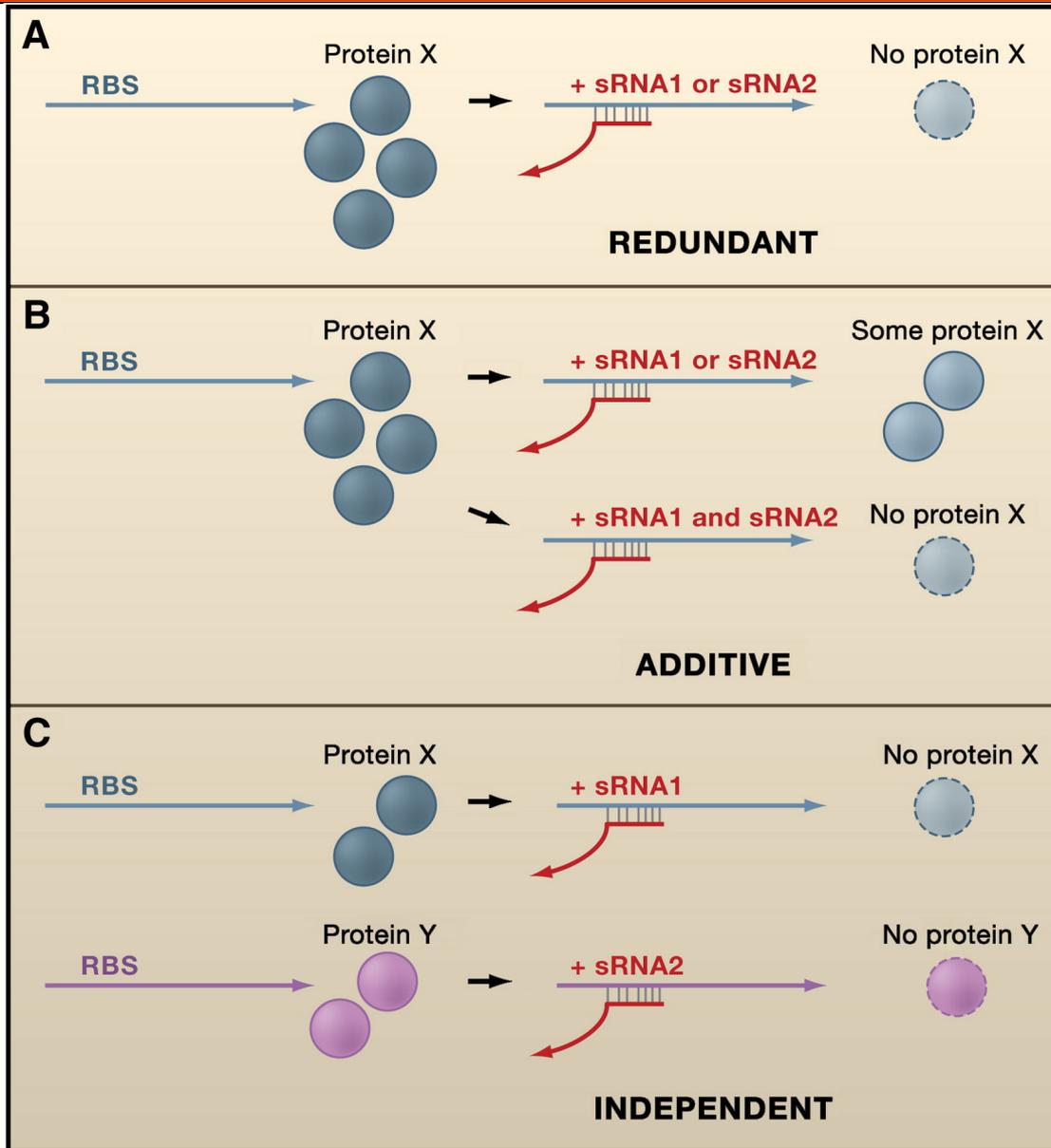
Copy Number Variation

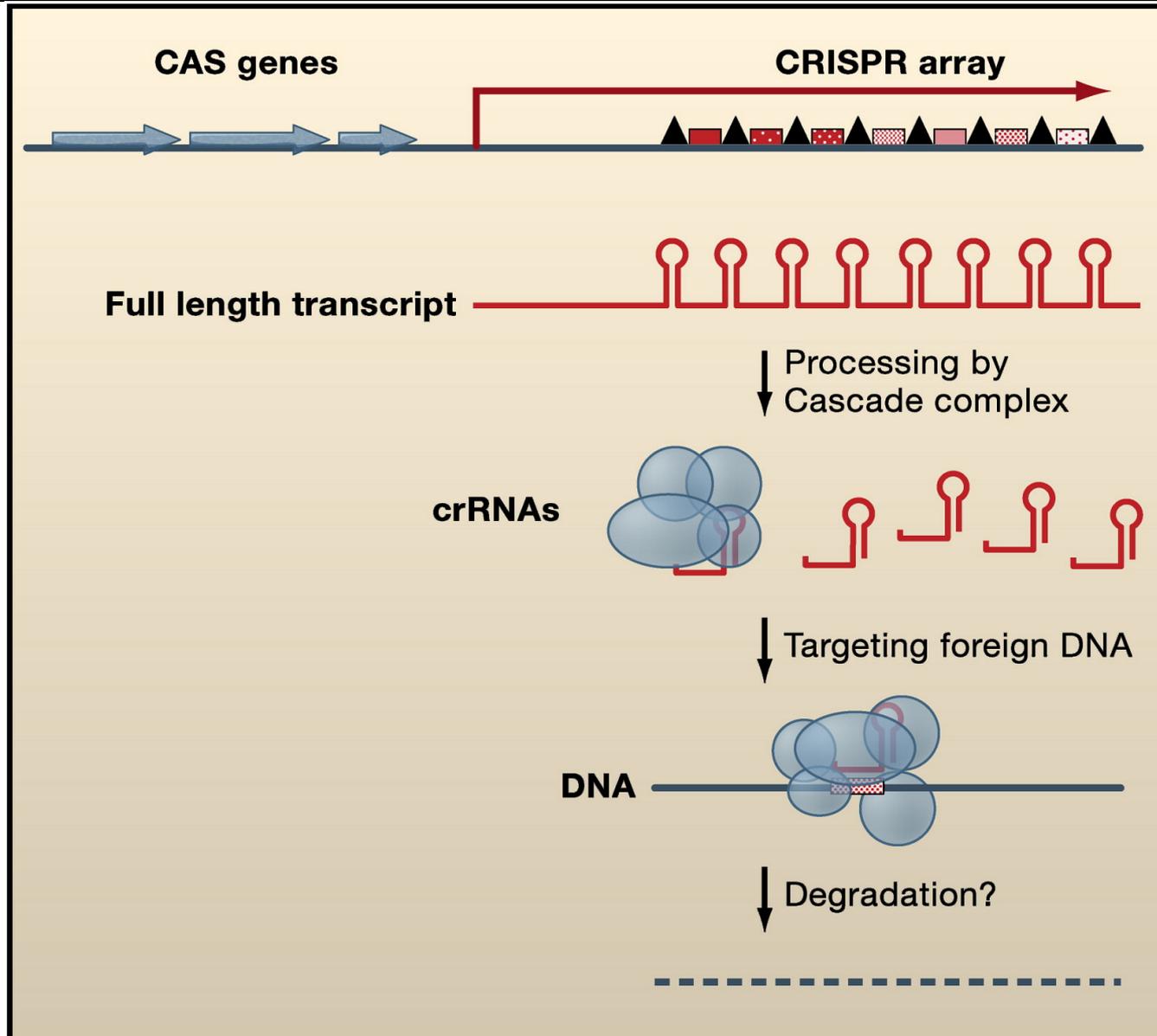# Copy-number variations (CNVs)

- "Mutation" where cells have an abnormal number of copies of one or more sections
- Most CNVs are stable and heritable
- Like other genetic variations, some CNVs are associated with susceptibility or resistance to disease.
  - EGFR (high) in non-small cell lung cancer
  - CCL3L1 (high) lower susceptibility to HIV infection
  - FCGR3B (low) high susceptibility to systemic lupus erythematosus and similar inflammatory autoimmune disorders
  - Autism, schizophrenia, and idiopathic learning disability

Duplicated area

Before duplication

After duplication
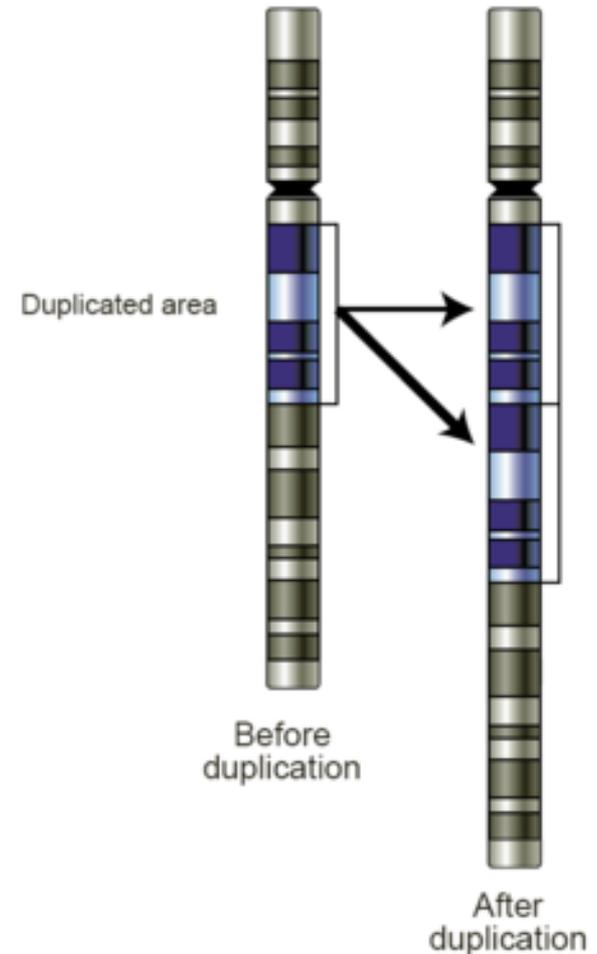
Horizontal Gene Transfer

Transfer of Genetic Material

❑ Vertical gene transfer happens during cell division
  - Normal mode of inheriting genetic material by offspring
  - Leads to tree phylogenies

❑ Horizontal gene transfer is process by which organism transfers genetic material to another organism that is not its offspring
  - Common between different species of bacteria
  - Can happen in eukaryotes between nuclear, mitochondrial and chloroplast
  - Can happen between a eukaryotic cell and a viral or bacterial cell (during endocytosis)

Horizontal Gene Transfer (HGT)

- ❑ HGT happens by non-homologous recombination
- ❑ Three mechanisms:
  - ● conjugation,
  - ● transduction, or
  - ● Transformation
- ❑ Bacteria can acquire genetic material from other co-habiting bacteria or viruses

Donor cell   Recipient cell
Conjugation

Donor cell   Recipient cell
(infected by a virus)
Transduction

Donor cell   Recipient cell
(dead)
Transformation
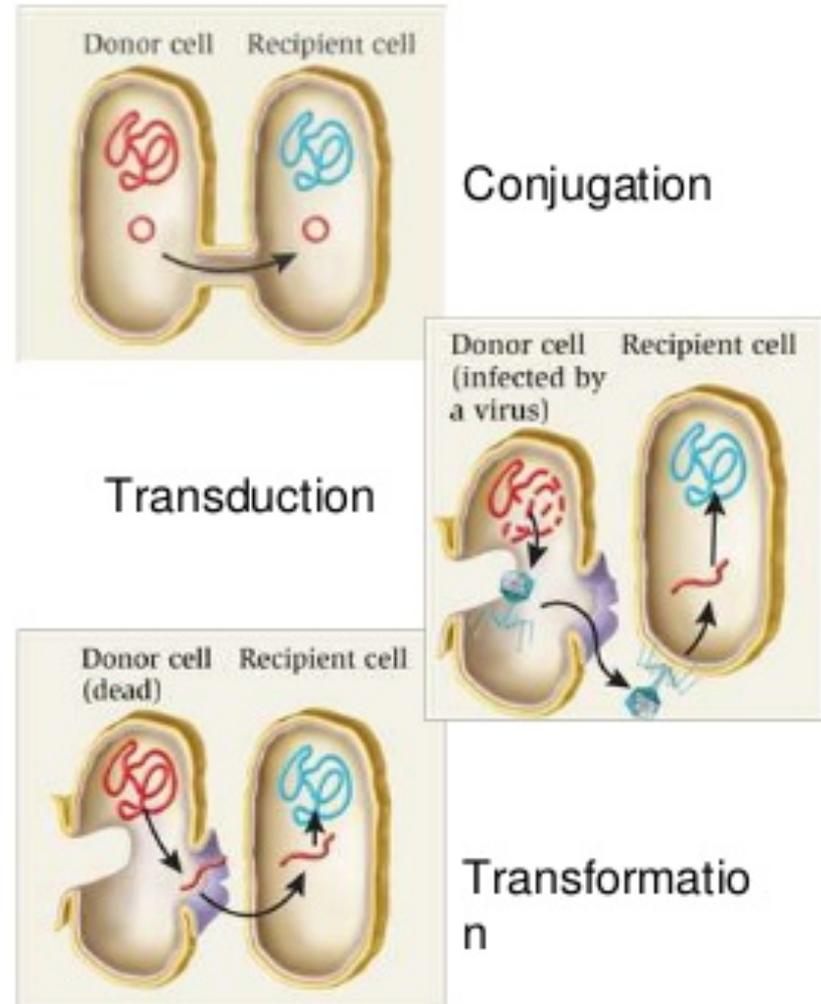
HGT in bacteria is important …

- ❑ Survival-specific genes acquired through HGT
    - 🔴 Antibiotic resistance genes
    - 🔴 Toxin resistance genes
- ❑ Environmental adaptation genes acquired through HGT
    - 🔴 "Symbiosis island" in *Mesorhizobium loti*
    - 🔴 "Metabolic island" in *Salmonella senftenberg*
    - 🔴 Pathogenicity islands carrying Virulence genes

How to detect HGT?

- ❑ Main principle: Sequences new to genome will retain (for a while) signatures of donor genome and distinguished from ancestral DNA
  - 🔴 Sequence analysis
    - ➤ Closest orthologs are in unrelated bacteria (e.g., w/ BLAST)
    - ➤ Synteny
  - 🔴 G+C Content
  - 🔴 Unusual codon usage
  - 🔴 Other sequence signatures
    - ➤ Dyad, Triad, k-mer frequencies
  - 🔴 Located close to direct and inverted repeats
  - 🔴 Located close to RNA genes
  - 🔴 Phylogenetic methods: Evolutionary history different from rest of genome
  - 🔴 Database of known insertion sequences
  - 🔴 Other factors include presence of special genes: soj, integrons, etc.
  - 🔴 Combination of above

Disordered Proteins

October 9, 2013

☐ Nobel Prize for Chemistry Awarded
- For computer models for complex chemical processes
- Martin Karplus, Michael Levitt, Arieh Warshel

Sequence ➔ Structure

> P00698 Chicken Lysozyme Protein Sequence

MRSLLILVLC FLPLAALGKV FGRCELAAAM KRHGLDNYRG YSLGNWVCAA KFESNFNTQA

TNRNTDGSTD YGILQINSRW WCNDGRTPGS RNLCNIPCSA LLSSDITASV NCAKKIVSDG

NGMNAWVAWR NRCKGTDVQA WIRGCRL
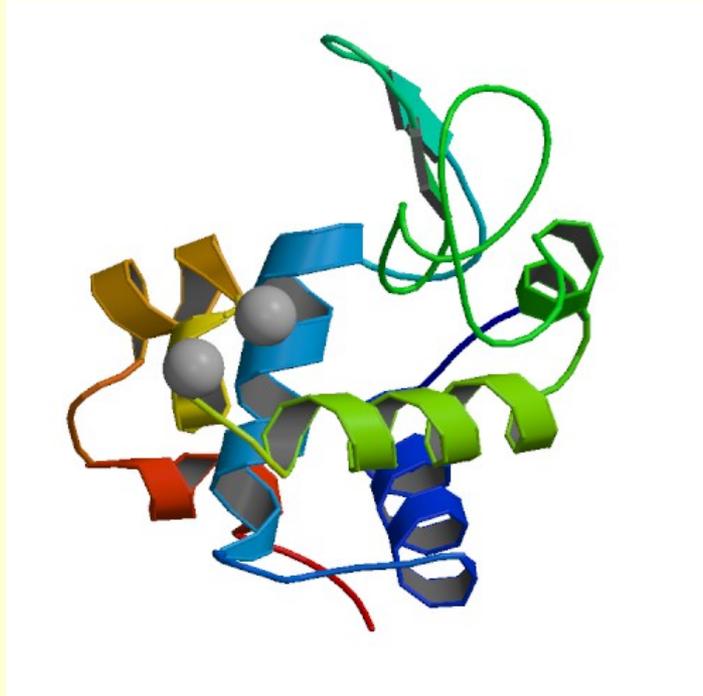
Sequence ➜ Structure

❑ Minimize total classical potential energy of the system

❑ Protein structure determined by the minimum energy configuration

❑ Potential energy is approximated by 6 terms

- $\sum K_b(b-b_0)^2 +$ Bond Length
- $\sum K_\theta(\theta-\theta_0)^2 +$ Bond Angle
- $\sum K_\Phi(\Phi-\Phi_0)^2 +$ Bond Angle
- $\sum \varepsilon_{ij}(r_{ij}'/r_{ij})^{12} - \sum 2\varepsilon_{ij}(r_{ij}'/r_{ij})^6 +$ Lennard-Jones potential
- $\sum Q_i Q_j/r_{ij}$ Coulomb Forces

❑ The r^{-12} term is Pauli repulsive term at short ranges due to overlapping electron orbitals and the r^{-6} term is attractive long-range term (van der Waals force, or dispersion force).

❑ "Computer Simulation of protein folding," M. Levitt and A. Warshel, Nature, 253(5494): 694–698, 1975.

Structure ➜ Dynamics

❑ In reality, what you get is not a static system, but a dynamic system.

❑ Understanding molecular dynamics is useful

❑ Challenges traditional paradigm that protein 3D structure can be fully determined by knowing its sequence, and that the minimum energy structure is a single stable structure

❑ Molecular dynamics (MD) is a computer simulation of physical movements of atoms and molecules in the context of N-body simulation

❑ "CHARMM: A Program for Macromolecular Energy, Minimization, and Dynamics Calculations," B. R. Brooks, R. E. Bruccoleri, B. D. Olafson, D. J. States, S. Swaminathan, and M. Karplus, J. Comp. Chem, 4: 187–217, 1983

Dynamics ➔ Function

❑ "Theoretical Studies of Enzymic Reactions : Dielectric, Electrostatic and Steric Stabilization of the Carbonium Ion in the Reaction of Lysozyme," A. Warshel and M. Levitt, J Mol Biol, 103: 227–249, 1976.

❑ http://www.dnatube.com/video/1228/ tRNARibosome-Molecular-Dynamics-Simulation

Disordered Proteins

- Protein that lacks a fixed or ordered three-dimensional structure
- Challenges traditional paradigm that protein function depends on fixed 3D structure
- Disordered proteins have regular functions
- Disorder may be associated with disease also
- Different proteins have different levels of disorder
  - Fully unstructured to partially structured
- Good prediction software exists
- DisProt, MobiDB, D^2P^2: database of such proteins

[Kozlowski LP, Bujnicki JM. MetaDisorder:

a meta-server for the prediction of intrinsic

disorder in proteins. BMC Bioinformatics.

2012 May 24;13(1):111.]

[http://www.ideal.force.cs.is.nagoya-u.ac.jp/IDEAL]
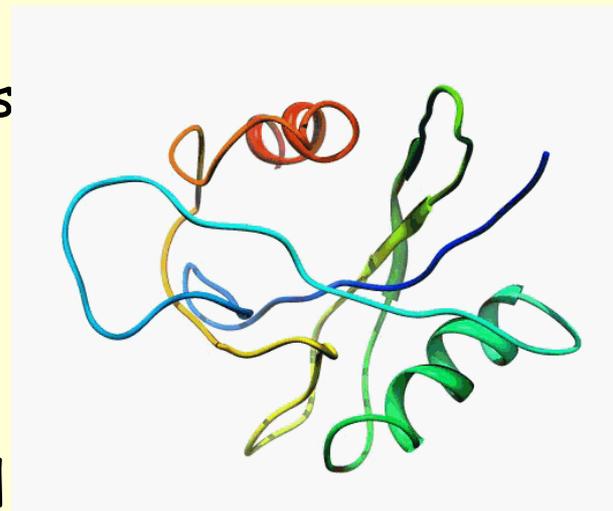
Transformer Proteins

Optical Mapping

Optical Mapping

- ❏ Technology to identify **location** and **order** of restriction sites on large DNA molecules
  - Enables efficient construction of accurate genome-scale restriction maps
  - Used to scaffold & validate genome assemblies (bacteria to human genome)
- ❏ Nanocoding: improved technology for higher accuracy and throughput
- ❏ Technology:
  - Large DNA fragments (~100Kb) immobilized on surface
  - Digested with one or more restriction enzymes
  - Stained with fluorescent dye
  - Imaging + Estimation of fragment lengths
  - Results in ordered series of fragment lengths
- ❏ Source of errors
  - Inaccurate estimates of lengths
  - Missing or extra restriction sites
  - Missing small fragments
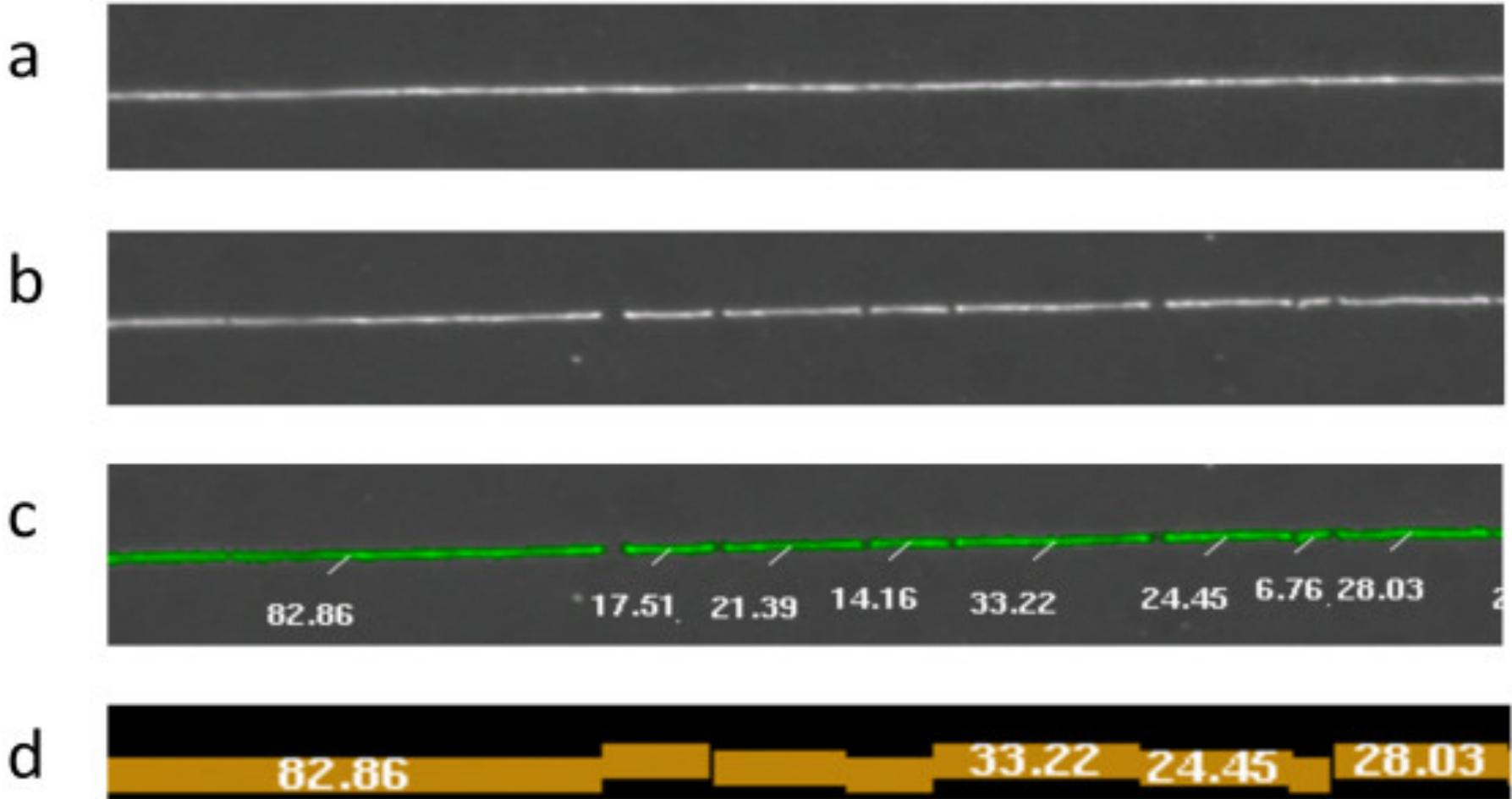
[Mendelowitz & Pop. GigaScience 2014, 3:33]
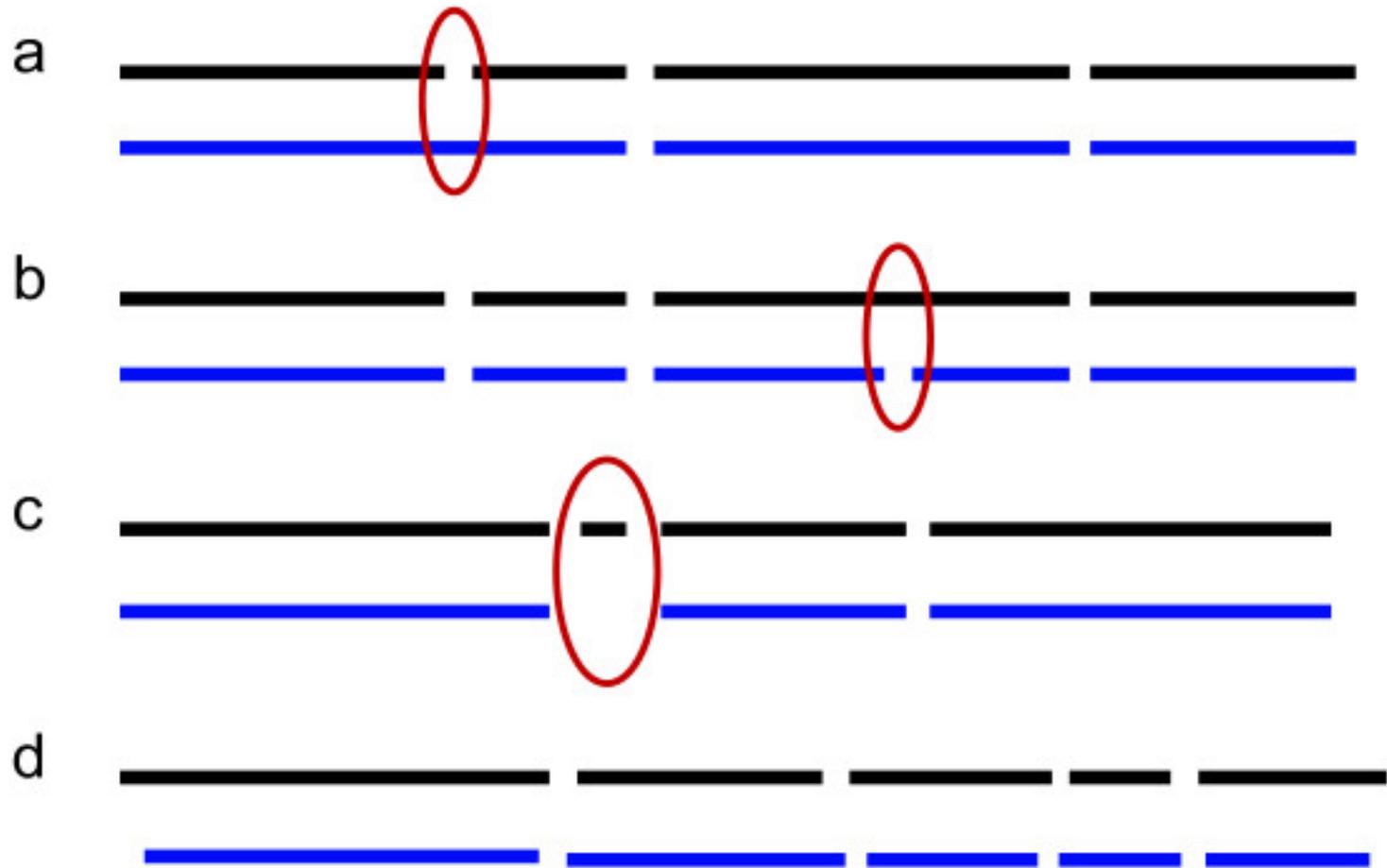
[Mendelowitz & Pop. GigaScience 2014, 3:33]

Bacterial Genome Assembly using OM

- Xavier BB, Sabirova J, Pieter M, Hernalsteens JP, de Greve H, Goossens H, Malhotra-Kumar S: Employing whole genome mapping for optimal de novo assembly of bacterial genomes. BMC Res Notes 2014, 7:484

- Lin HC, Goldstein S, Mendelowitz L, Zhou S, Wetzel J, Schwartz DC, Pop M: AGORA: Assembly Guided by Optical Restriction Alignment. BMC Bioinformatics 2012, 13:189

- Nagarajan N, Read TD, Pop M: Scaffolding and validation of bacterial genome assemblies using optical restriction maps. Bioinformatics 2008, 24:1229–1235

- Zhou S, Kile A, Bechner M, Place M, Kvikstad E, Deng W, Wei J, Severin J, Runnheim R, Churas C, Forrest D, Dimalanta ET, Lamers C, Burland V, Blattner FR, Schwartz DC: Single-molecule approach to bacterial genomic comparisons via optical mapping. J Bacteriol 2004, 186:7773–7782.