JOHNS HOPKINS
BLOOMBERG
SCHOOL of PUBLIC HEALTH

JOHNS HOPKINS
BLOOMBERG
SCHOOL of PUBLIC HEALTH

Genomics and Infectious Disease

Gary Ketner, PhD
Johns Hopkins University

JOHNS HOPKINS
BLOOMBERG
SCHOOL *of* PUBLIC HEALTH

Section A

Overview: Genomics, the Human Genome Project, and Public Health

Genomics

- Genome: an individual’s complement of genetic information
  - Frequently taken to mean the nucleotide sequence of the genome
- Genomics: the study of the genome (or of its sequence)
- Transcriptome: the mRNAs present in an individual, organ, or cell
- Proteome: the proteins present in an individual, organ, or cell

Sequenced Genomes

- Viruses (1,674 sequences as of 10/01/04)
  - First sequenced genome: ϕX174 (1978) 5,386 nucleotides (0.005 Mb)
- Archea (19)
  - *Thermoplasma volcanium* (1.6 million bases; Mb)
- Bacteria (173)
  - First sequenced free-living organism: *Haemophilus influenzae* (1995) 1.83 Mb
  - *Vibrio cholerae* (2001) 2.96 Mb
  - *Mycobacterium tuberculosis* (2001) 4.40 Mb

http://www.genomesonline.org/CompleteGenomesList.html

Continued

Sequenced Genomes

- Eukaryotes (28)
  - *Plasmodium falciparum* (2002) 22.9 Mb
  - *Anopheles gambiae* (2003) 278 Mb
  - Human (2001 [draft]; 2003 [ref.]) 2900 Mb (2.91 billion bp)
    - Total file size, downloaded sequence: 841 Mb

http://www.genomesonline.org/CompleteGenomesList.html

Genomics and Public Health

- Genomics is a tool of extraordinary power in the investigation of biology
- Because biology is central to most public health problems, an understanding of biology is essential
- Genomics is a potential contributor of immense value to public health

Human Genome Project (HGP)—Circa 1991

- Goal
  - Complete DNA sequence of the human genome
- Schedule
  - 15 years
- Cost
  - A few billion dollars

HGP: What We Knew to Begin With

- DNA is the genetic material
- Information is encoded in DNA in the order along its length of the bases A, T, G, and C
- The human genome is big: >10^9 bp
- From genetic and molecular studies (humans and model organisms)
  - A general outline of the organization of the genome
  - The nature of genes, regulatory elements, and other components of the genome

HGP: What We Hoped to Find Out

- The raw sequence
  - A string of more than a billion As, Ts, Gs, and Cs
- From this, we hoped to deduce:
  - The DNA sequence of all of our genes
  - The amino acid sequence of all of our proteins
    - The functions of all of our proteins
  - And a lot more
    - Evolutionary relationships
    - Regulatory mechanisms
    - Bases of normal and abnormal development
    - Determinants of genetic disease and disease susceptibility

Section B

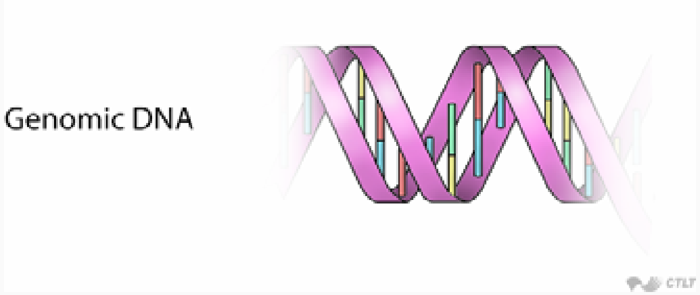
Sequencing and Finding Genes

The Approach: Shotgun Sequencing

- Generate raw sequence data in vast quantities from short, randomly produced pieces of human DNA—by highly automated procedures
- Assemble these sequences computationally to generate a complete sequence
- Annotate the sequence (identify genes, etc.), again by computation

Shotgun Cloning

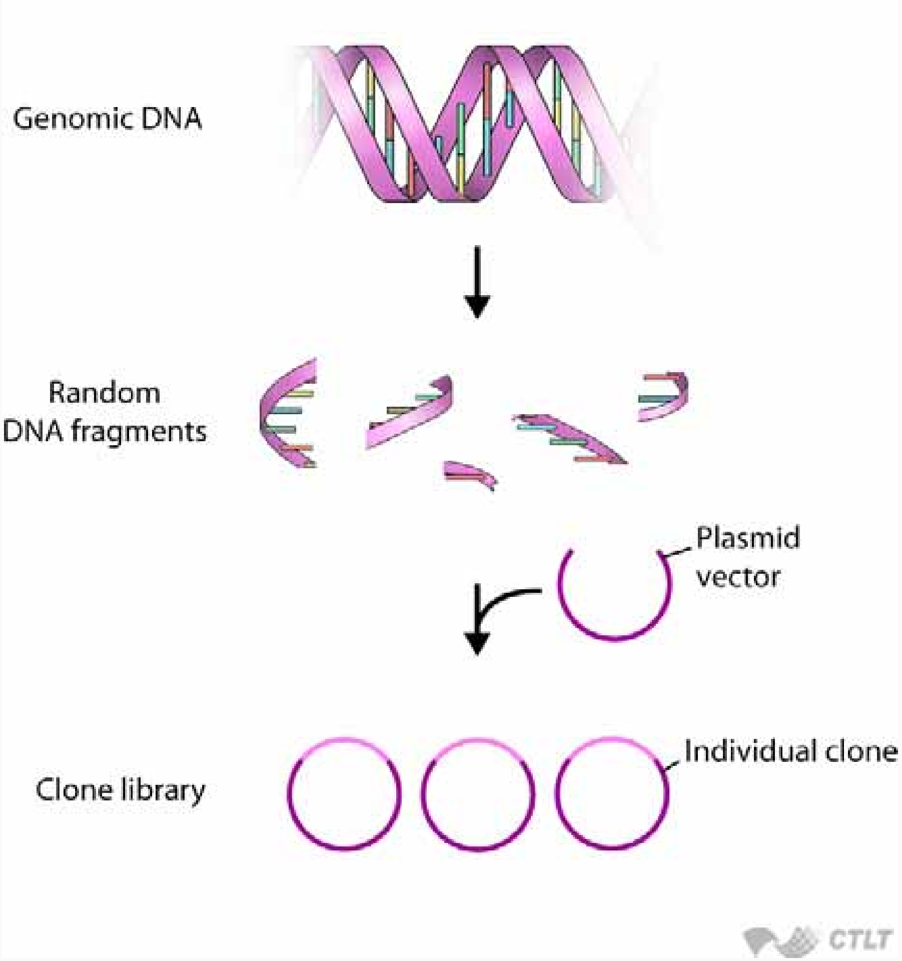
- Isolate genomic DNA
- Randomly fragment (average sizes 2kb, 10kb, 50kb)



Continued

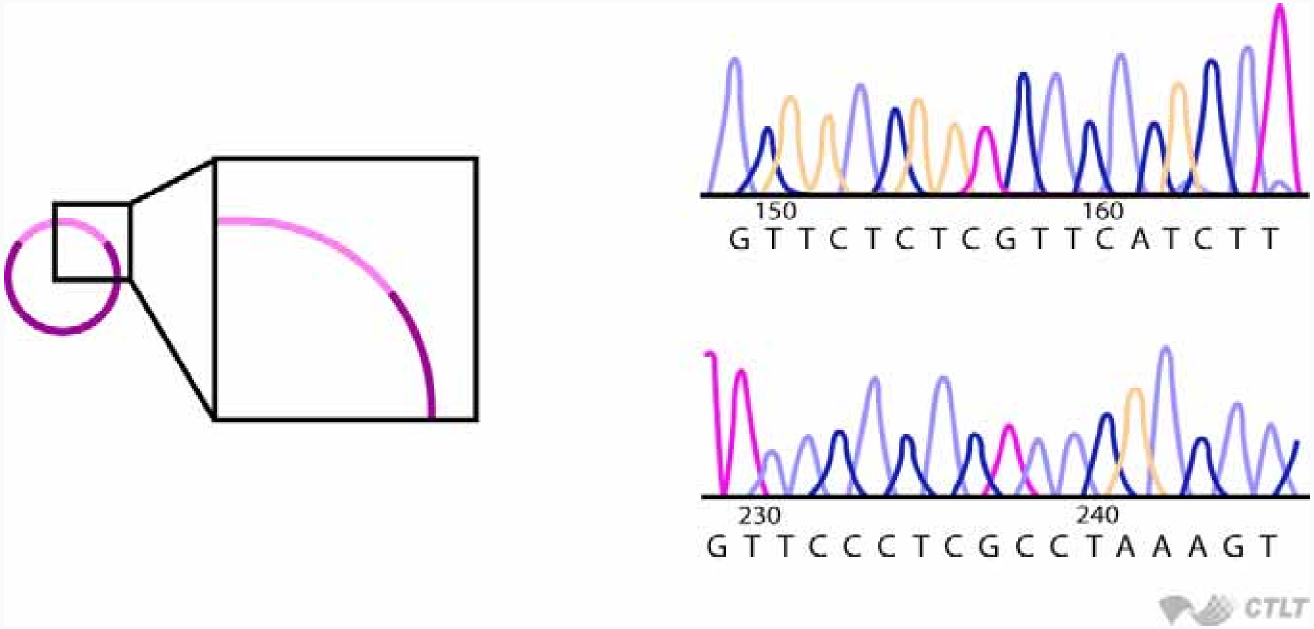
Shotgun Cloning

- Use fragments to prepare clones/clone libraries
  - 2kb: 5,000,000
  - 10kb: 2,500,000
  - 50kb: 500,000
  - 37x coverage

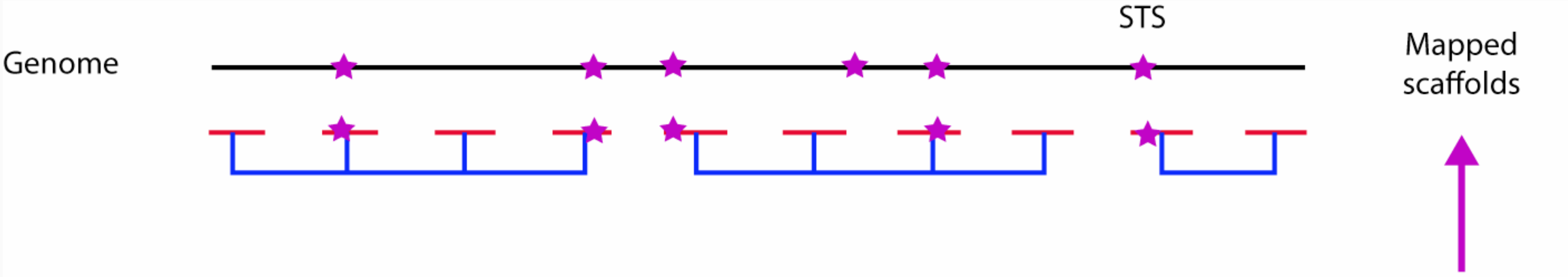


Shotgun Sequencing

- Determine 500bp sequences from each end of each clone
    - 273 x 10^6 sequence reads
    - 14.9 x 10^9 base pairs read
    - 5x sequence coverage
- Send the sequence file to the assembler

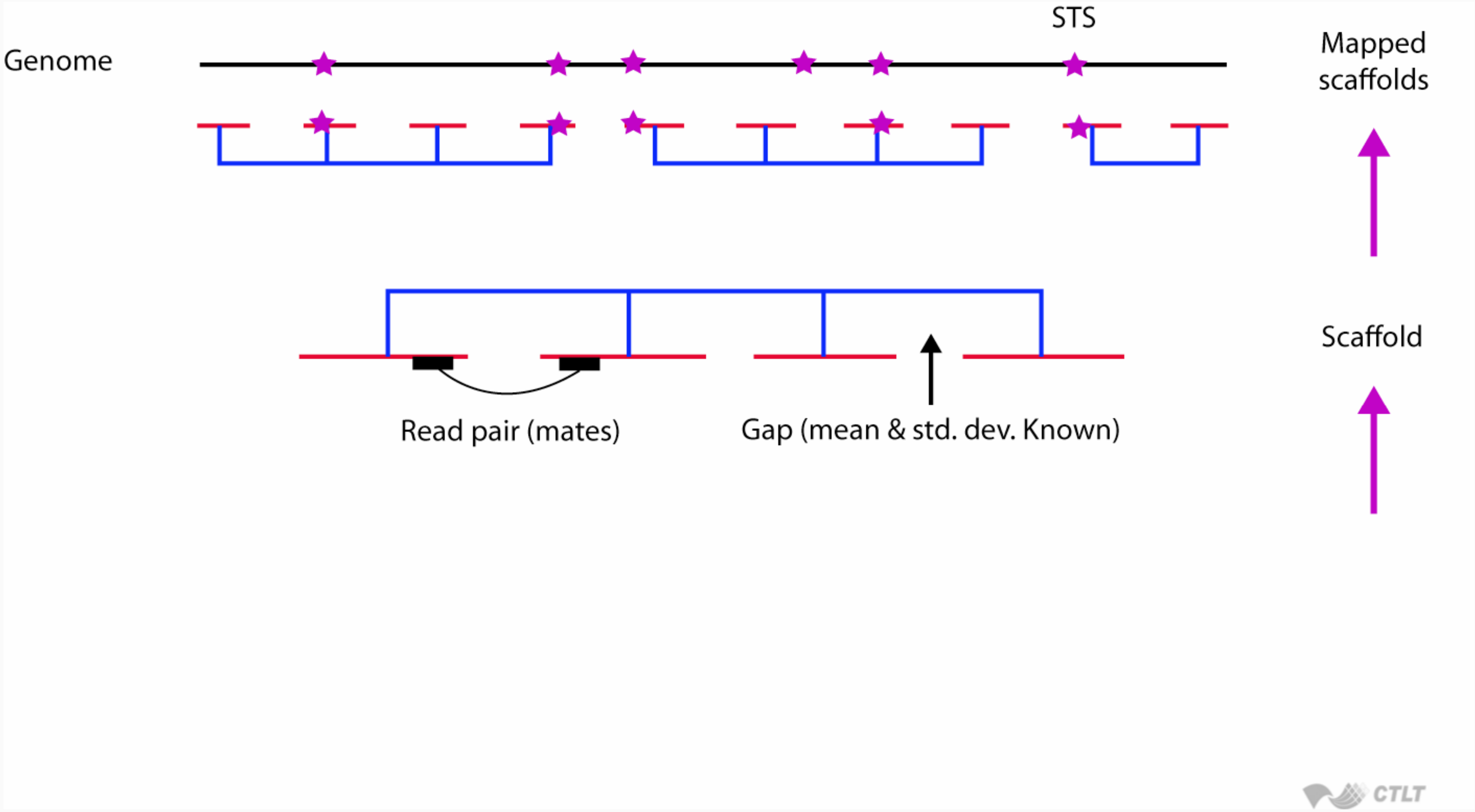

150
160
G T T C T C T C G T T C A T C T T
230
240
G T T C C C T C G C C T A A A G T
CTLT

Assembly



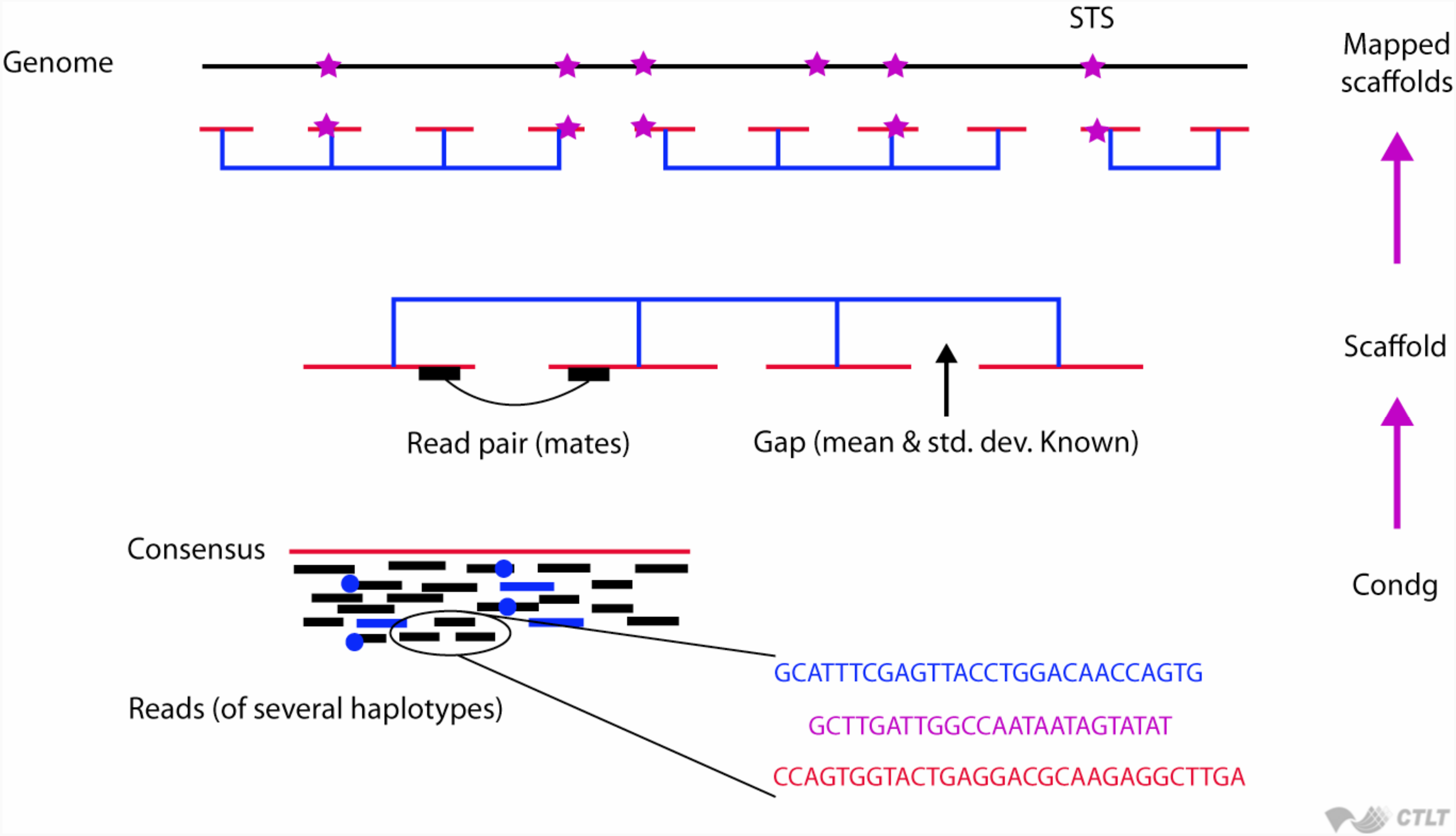
Continued

Assembly



Continued

Assembly


STS
Genome
Mapped scaffolds
Scaffold
Read pair (mates)
Gap (mean & std. dev. Known)
Consensus
Condg
Reads (of several haplotypes)
GCATTTCGAGTTACCTGGACAACCAGTG
GCTTGATTGGCCAATAATAGTATAT
CCAGTGGTACTGAGGACGCAAGAGGCTTGA
CTLT

The Reference Sequence of the Human Genome: Stats

- The total sequence consists of 3.09×10^9 bp
  - Haploid amount
  - Includes both X and Y
- There are 400 gaps in known places, containing 1% of the DNA and relatively few genes
- Accuracy: 99.99%
- Cost: $\$2.7 \times 10^9$
- Download it
  - http://www.ncbi.nlm.nih.gov/genome/seq/

Continued

The Reference Sequence of the Human Genome: Stats

- What is the reference sequence?
  - The sequence of one human's genome?
  - The “average” sequence of all human genomes?
  - The “normal” sequence of the human genome?
  - None of the above?
- The reference sequence is a composite
  - Assembled from good quality pieces of several individuals’ sequences
  - It is nobody’s genome, exactly
  - Presumably, it would give rise to a fully functional individual
- Individual genomes can be efficiently described in terms of differences from the reference sequence

Finding Genes

- Proteins mediate most of the processes that occur in living cells
  - Proteins are responsible for normal and abnormal metabolism, development, and disease susceptibility
- Genes encode proteins
- Gene identification is a primary goal of the HGP

Continued

Finding Genes

- Genes can be identified in the DNA sequence computationally
  - Genes contain open reading frames (ORFs)
    - DNA sequences that can be used to predict amino acid sequence by means of genetic code
  - Genes show characteristic usage of the genetic code words for amino acids
  - The coding sequences of a gene fall within length limits
  - Genes are flanked by punctuation for transcription and translation

Gene Function

- About 30,000 genes have been found in the human genome
  - Twice as many as in a fly
  - Five times as many as in yeast
- About half of the genes in the human genome are identical or similar enough to genes of a known function to confidently assign function
  - More exciting, half are not

Fun Genome Facts

- The total amount of DNA accounted for by genes is about 1.5% of the total DNA in the genome
  - Introns account for a substantial fraction of the rest
  - About half of the human genome consists of transposons (mobile genetic elements) with no known function
- Some of our genes entered our lineage from bacteria relatively recently (600 mega years—after the divergence of vertebrates and invertebrates)
  - These genes largely encode enzymes that deal with xenobiotic chemicals (monoamine oxidase)

JOHNS HOPKINS
BLOOMBERG
SCHOOL *of* PUBLIC HEALTH

Section C

The *Plasmodium falciparum* Genome

The Plasmodium falciparum *Genome*

- Sequenced by a joint effort including publicly and privately funded components
- Completed in 2002
- 23 Mb
- 5,268 genes/proteins
  - 40% related to genes in other organisms
  - 60% unique
  - The *Plasmodium* proteome is somewhat poor in enzymes (parasitic lifestyle)
    - But it is rich in genes involved in immune evasion and cell adhesion

The Pf *Genome and Vaccines*

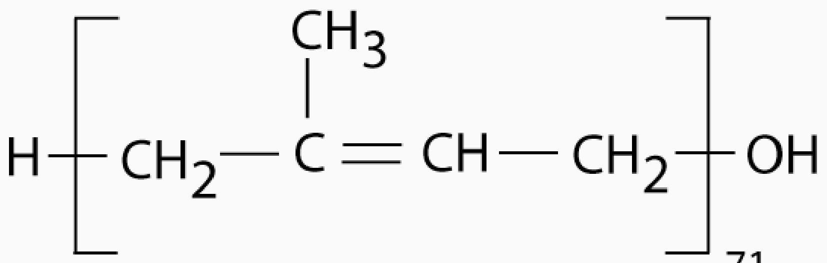
- The *P. falciparum* proteome constitutes a complete list of all of the *Pf* antigens that might induce protective immunity
- Potential targets that could be identified using genome data:
  - Merozoite surface or adhesion proteins
    - Antibody neutralization of blood stage parasites
  - Proteins expressed specifically in liver cells
    - CMI against infected hepatocytes
  - Proteins expressed on gametocytes
    - Transmission-blocking immunity

The Pf Genome and Drugs

- Drugs must be specific
- 60% of the malaria genome is unique, and all of those genes are potential drug targets once their function is known
- Several *Pf* enzyme systems have been identified from genome data as similar to bacterial enzymes
  - Most are associated with a specialized organelle called the apicoplast, whose evolutionary origin is bacterial
  - Immediately exploitable targets for drugs

Isoprenoid Metabolism in Pf

- Isoprenoids are a class of biochemicals with a wide variety of functions in all living things:
  - Cholesterol
  - Sterol hormones
  - Vitamin A
  - Dolichol
- Isoprenoid synthesis proceeds by successive additions of isopentenyl diphosphate to a growing molecule

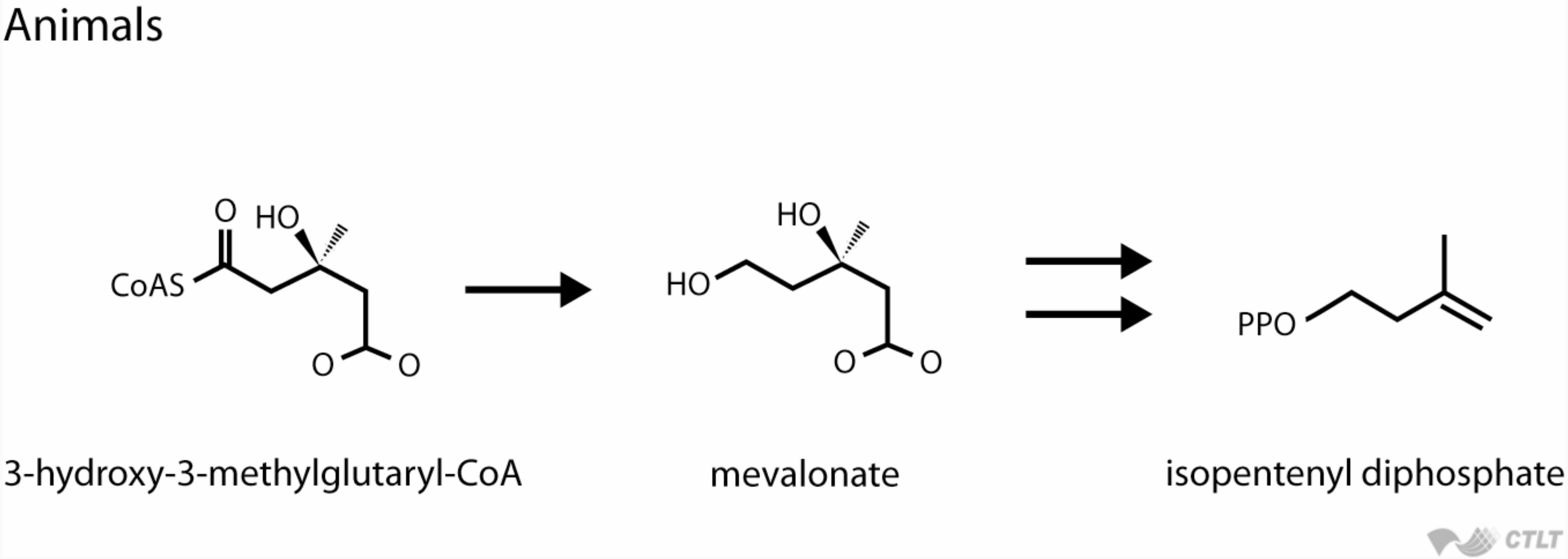


I

Continued

Isoprenoid Metabolism in Pf

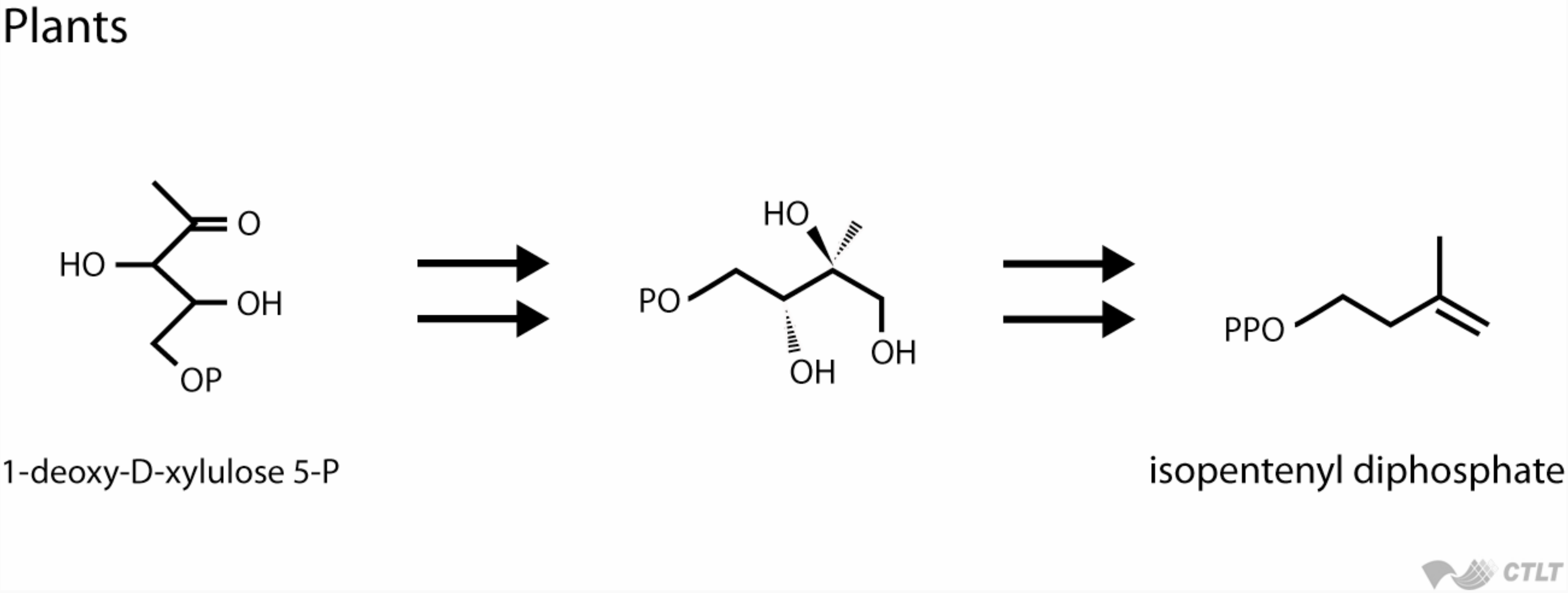
- Isopentenyl diphosphate is made in one of two ways
  - In animals (including humans)
    - Via mevalonic acid and the enzyme HMG CoA reductase



Continued

Isoprenoid Metabolism in Pf

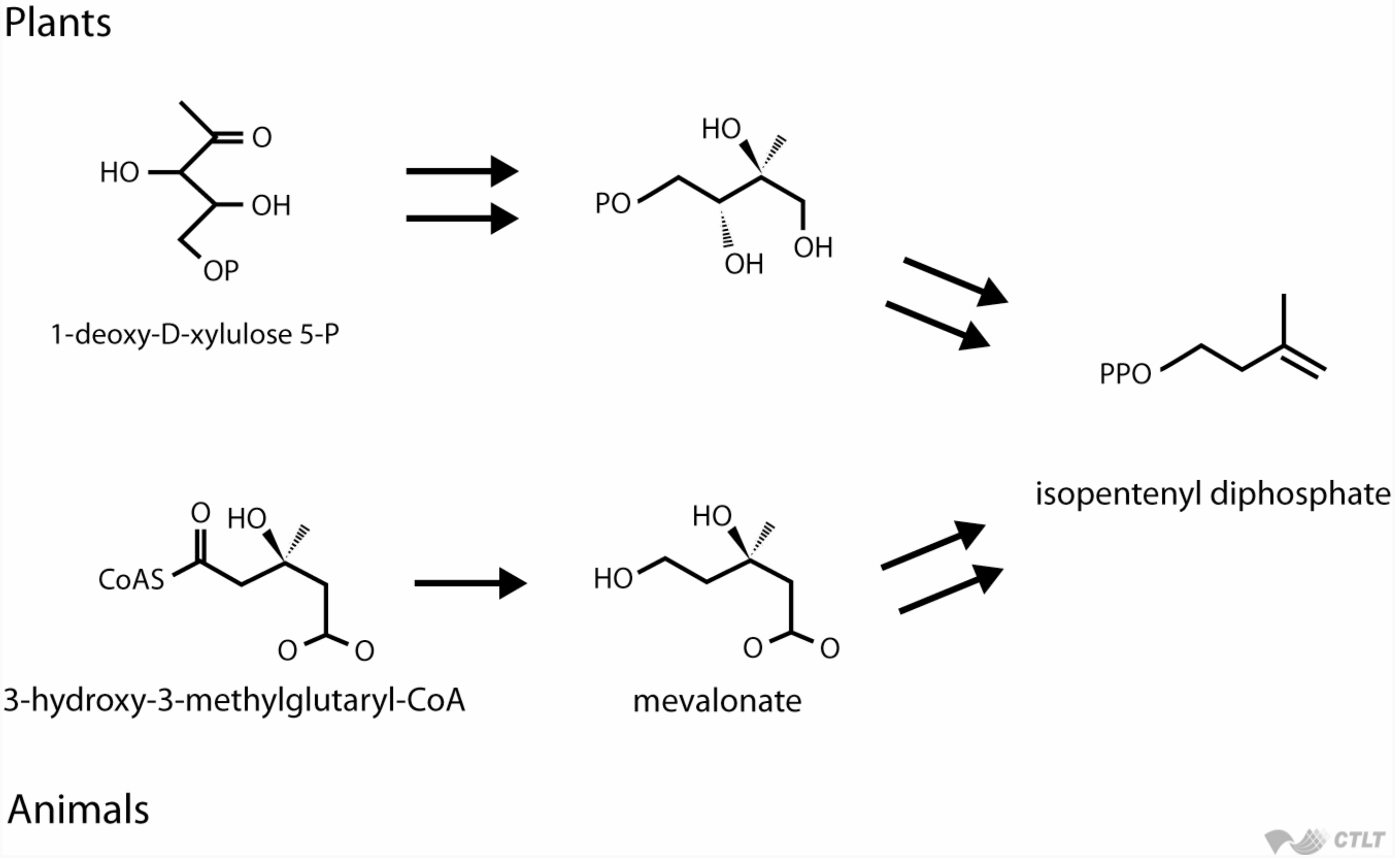
- Isopentenyl diphosphate is made in one of two ways
  - In plants and bacteria
    - Via deoxyxylulose-5-diphosphate (DOXP) and DOXP reductoisomerase



Continued

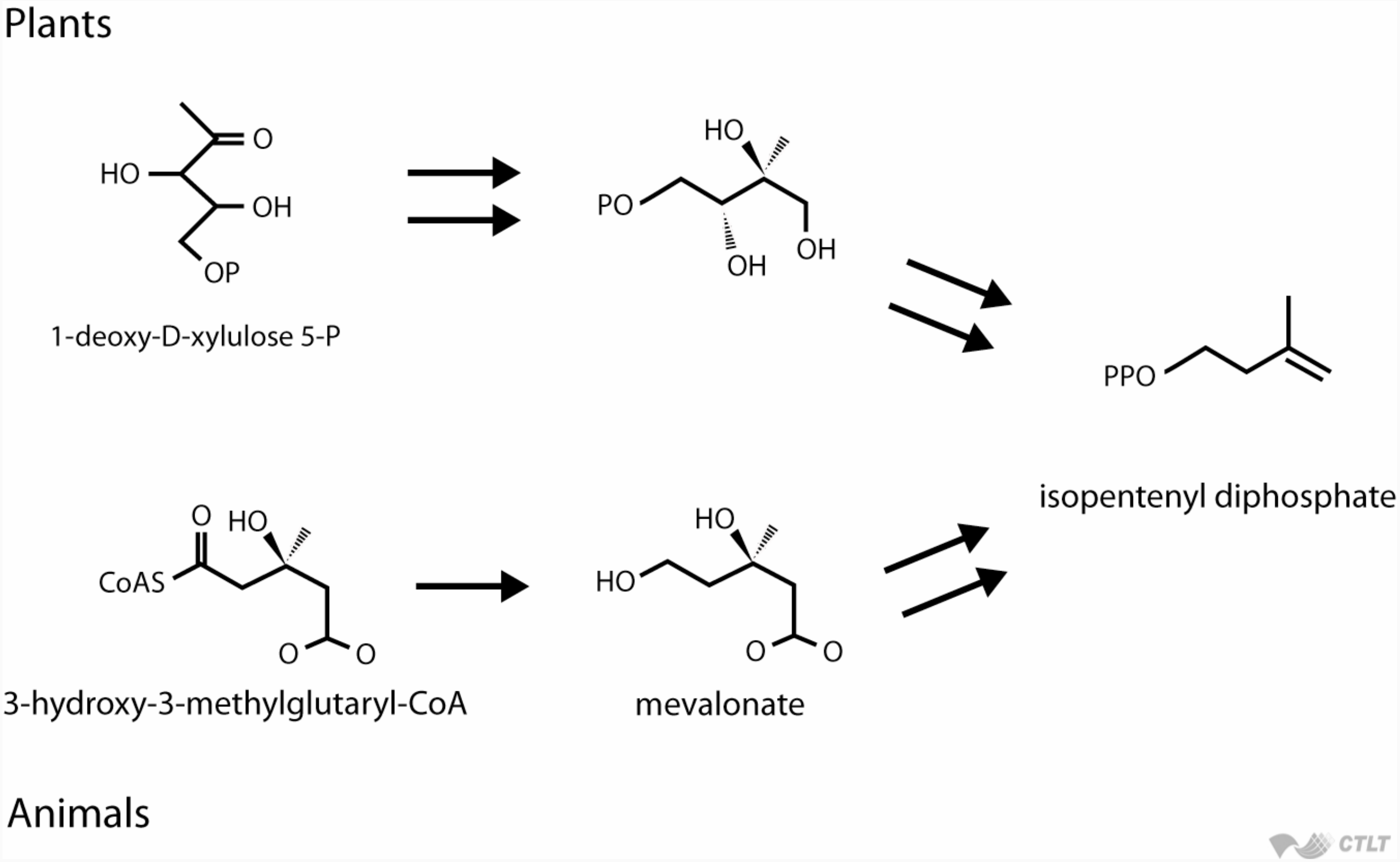
Isoprenoid Metabolism in Pf

- *P. falciparum* has no HMG CoA reductase
- However, the genome sequence revealed that it does have the enzymes required by the bacterial pathway (including DOXP reductoisomerase)



Isoprenoid Metabolism in Pf

- Inhibitors of DOXP reductoisomerase exist
- One of these, fosmidomycin, kills *Pf* in culture, and is therapeutic in *Plasmodium* infection in mice and humans



More Candidates

- Aromatic amino acid synthesis via shikimate
- Fatty acid synthesis by the type II pathway
- Both present in *Pf;* absent in humans
- The crystal structures of the type II enzymes are the targets of a JHSPH study, with the goal of designing inhibitors based on structural data

The Anopheles gambiae *Genome*

- 278 Mb
- 13,683 genes/proteins
  - 1/2 shared with *Drosophila*
- *A. gambiae* is the most important vector for malaria in Africa
  - Some strains of *A. gambiae* don’t transmit malaria because they don’t support parasite development
  - Genomic data is being used to identify the *Anopheles* genes responsible for this strain difference

Mosquito Engineering

- The plan
  - Identify a gene that will prevent malaria growth in mosquitoes
  - Genetically modify mosquitoes to carry that gene on a transposon
  - Introduce these mosquitoes into the wild, where the transposon will propagate through the population
- Efficient propagation of transposons has happened in nature in *Drosophila* populations over a few generations

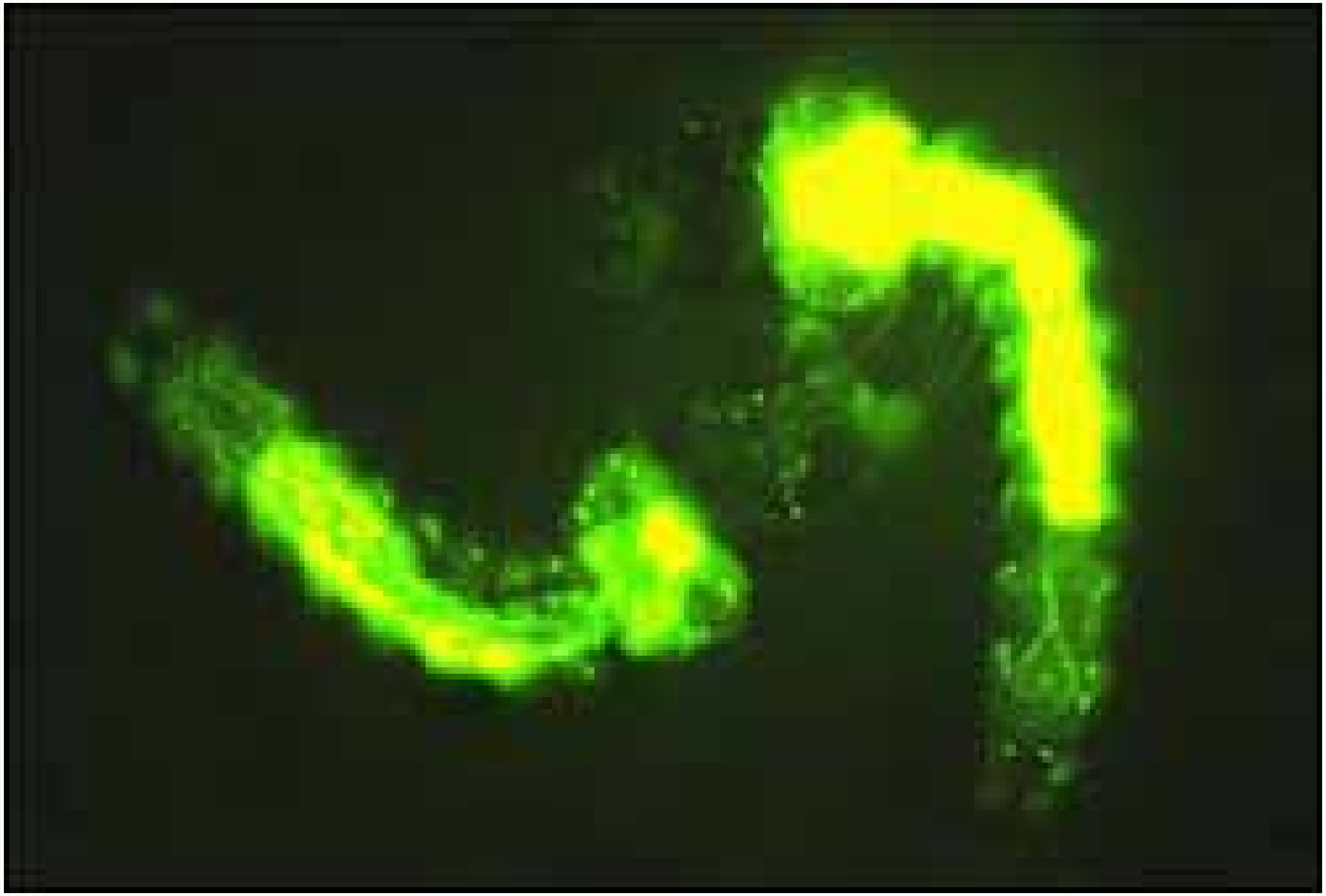
Continued

Mosquito Engineering

- The issues
  - What genes are required?
  - Population structure—will spread occur?
  - Will engineered mosquitoes be fit?
  - Can engineered mosquitoes be released?
    - Safety
    - Politics

Light at the End of the Tunnel?

- A. Crisanti, Imperial College, London



Mosquito larvae engineered to express foreign gene (green fluorescent protein).