Genomes: Now that we have them, what can we do with them?

Shifra Ben-Dor

Bioinformatics Unit

Weizmann Institute of Science

What do we use genome sequence for?

- Genes
- Splice Variants
- Variation analysis (SNP, mutations)
- Promoters
- Comparative Genomics
- Evolution

What do we use genome sequence for?

- Genes
- Splice Variants
- Variation analysis (SNP, mutations)
- Promoters
- Comparative Genomics
- Evolution

Looking for genes:

- In the lab:
  - Two hybrid system
  - SSH (suppression subtractive hybridization)
  - Exon trapping
  - Linkage analysis
  - Database Searches

Looking for genes

- In the browser:
  - Existing mRNAs or ESTs from large scale projects
  - Gene Prediction programs
    - Notoriously unreliable
  - Genetic Maps and Markers
  - Comparative Genomics

Looking for genes

- In the browser:
  - Existing mRNAs or ESTs from large scale projects
  - Gene Prediction programs
    - Notoriously unreliable
  - Genetic Maps and Markers
  - Comparative Genomics

Sources of mRNAs

- Experimental
    - Clone new gene
    - Clone gene from database
    - 2 hybrid system
- Database
    - "Typical" cDNA
    - Full length cDNA
    - EST

Looking for genes

In the browser:

- Existing mRNAs or ESTs from large scale projects
- Gene Prediction programs
  - Notoriously unreliable
- Genetic Maps and Markers
- Comparative Genomics

Gene Prediction Programs

- Many available, based on different algorithms, mainly checking for coding regions
- In one paper checking for accuracy, the measure used was “at least one exon correct”!

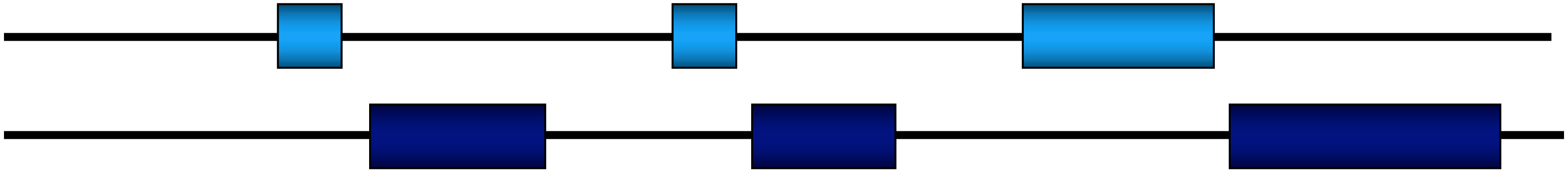
Gene Prediction Problems

- UTRs (non-coding exons) usually wrong, or not predicted at all
- Interleaved genes
- Nested Genes
- Cassette genes

Interleaved Genes

We have genes that are interleaved inside other genes.

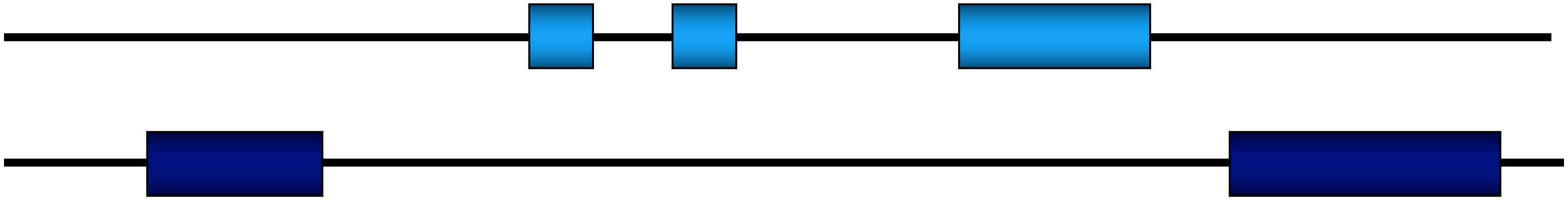
This means that there is slight overlap, but not total.



Nested genes

We have genes that are nested within other genes.

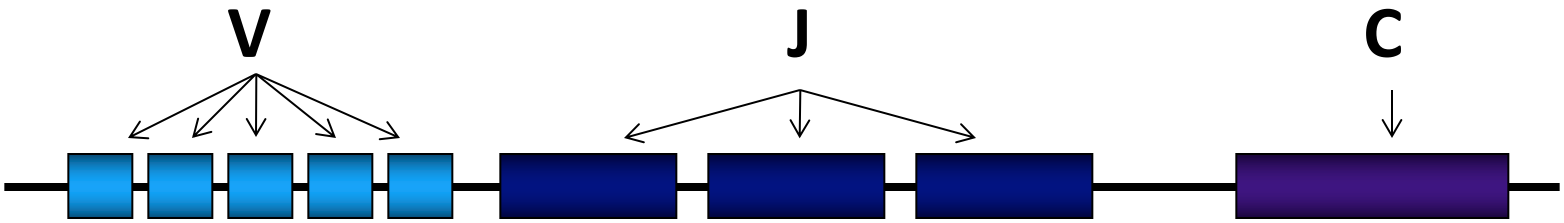
This causes problem for gene prediction programs that can ‘read’ them as one gene.



Cassette Genes

Genes that have multiple subunits, but only one combination is expressed in a given cell. For example, Immunoglobulin genes.

There are V, (D) and J regions, and they recombine with the constant region to form the expressed gene.



Looking for genes

- In the browser:
  - Existing mRNAs or ESTs from large scale projects
  - Gene Prediction programs
    - Notoriously unreliable
  - Genetic Maps and Markers
  - Comparative Genomics

Genetic Maps and Markers

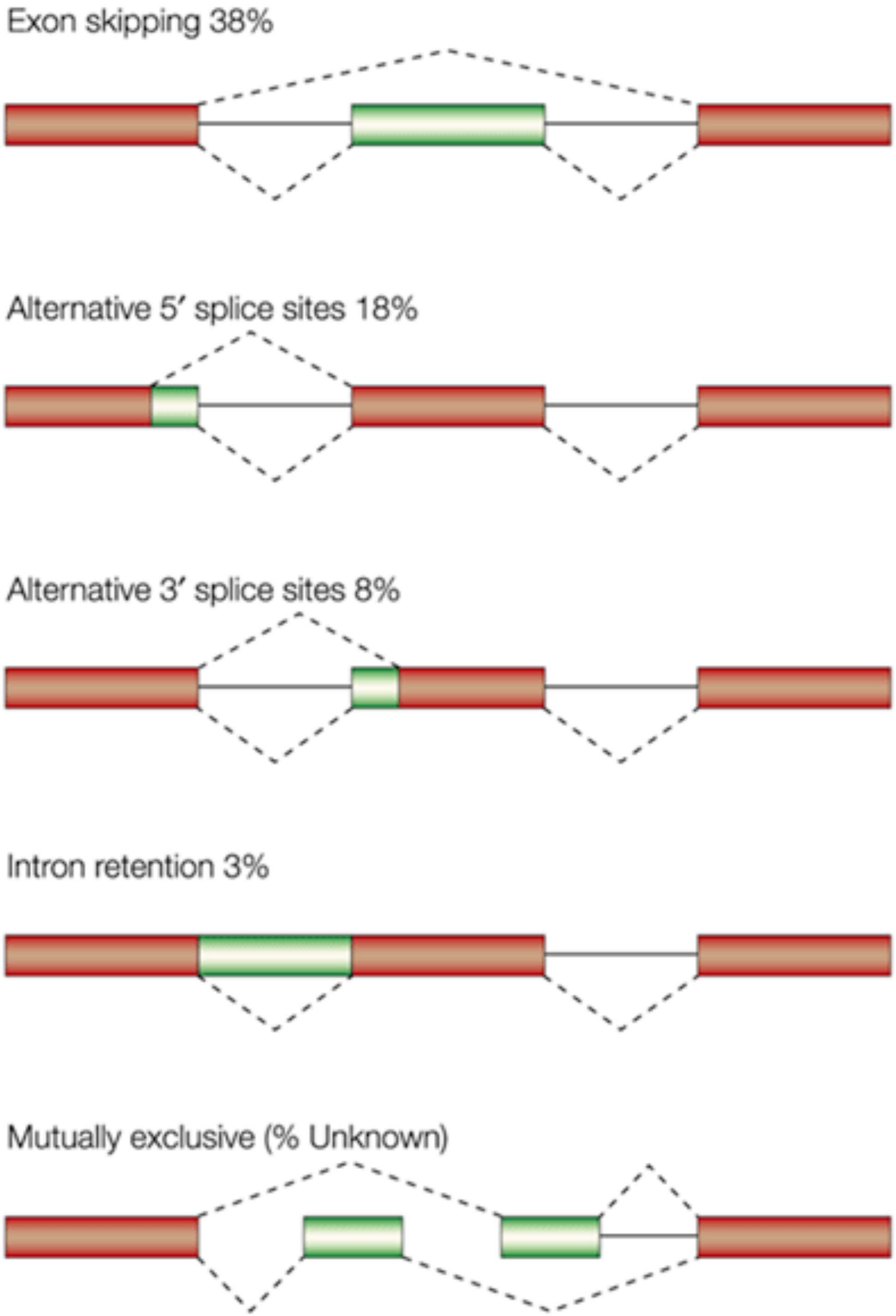
- Go to an area with linkage disequilibrium
- Pull out a list of candidate transcripts

What do we use genome sequence for?

- Genes
- Splice Variants
- Variation analysis (SNP, mutations)
- Promoters
- Comparative Genomics
- Evolution

Splice Variants

- Can generally be seen by comparison of expressed sequences (mRNA and EST) to genomic sequence



Exon skipping 38%

Alternative 5′ splice sites 18%

Alternative 3′ splice sites 8%

Intron retention 3%

Mutually exclusive (% Unknown)

What do we use genome sequence for?

- Genes
- Splice Variants
- Variation analysis (SNP, mutations, editing)
- Promoters
- Comparative Genomics
- Evolution

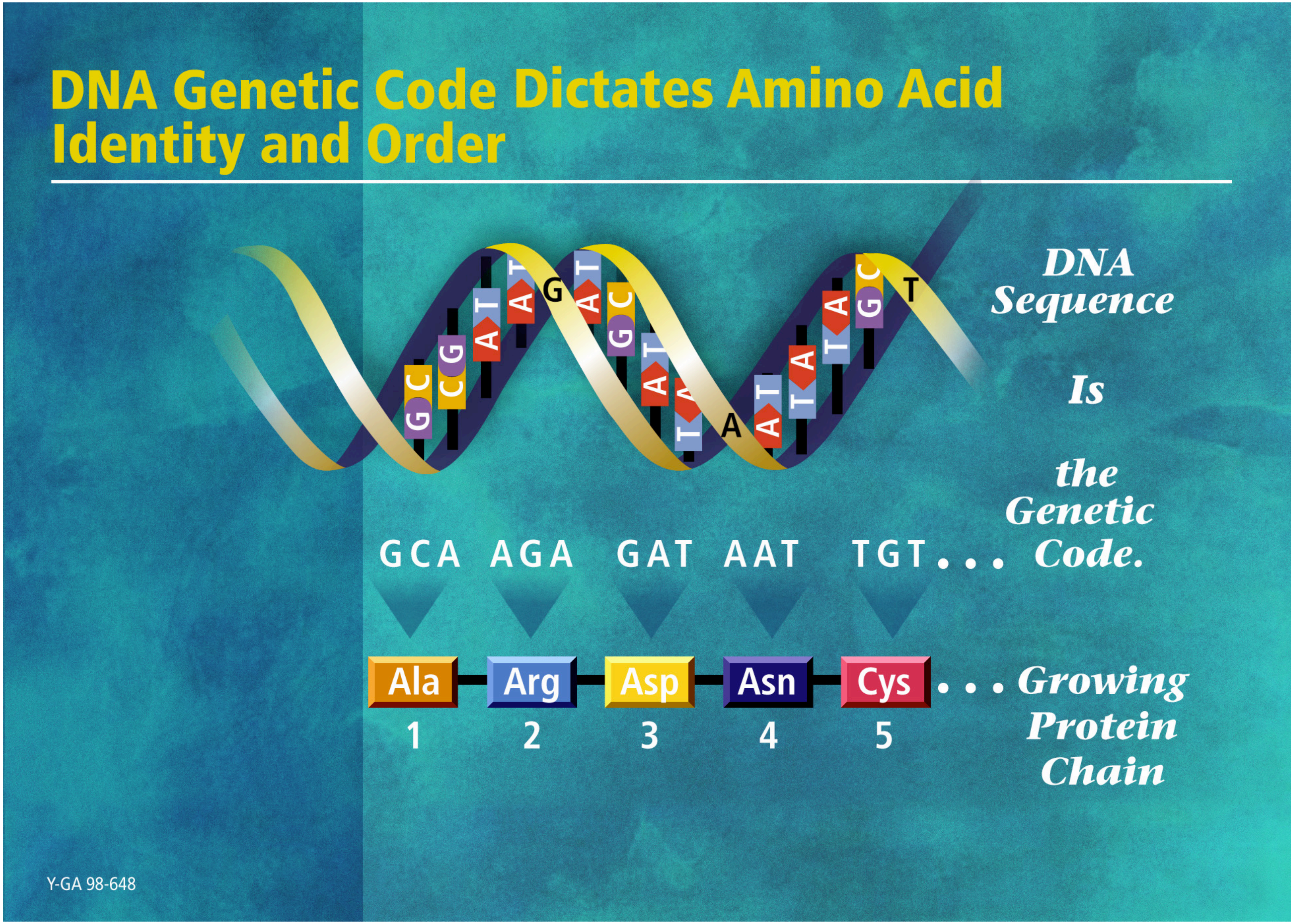
Variation (SNP, mutations, editing)

- Comparison of Genome sequence to transcribed sequences: mRNA, EST
- Look for differences

Genetic Variation

Since the DNA used in the various genome sequencing projects came from different donors, we can start looking at genetic variation.

There are many changes in our DNA. Some cause disease, some cause no change, some may be the reason people differ in reaction to various drug treatments.



DNA Genetic Code Dictates Amino Acid Identity and Order

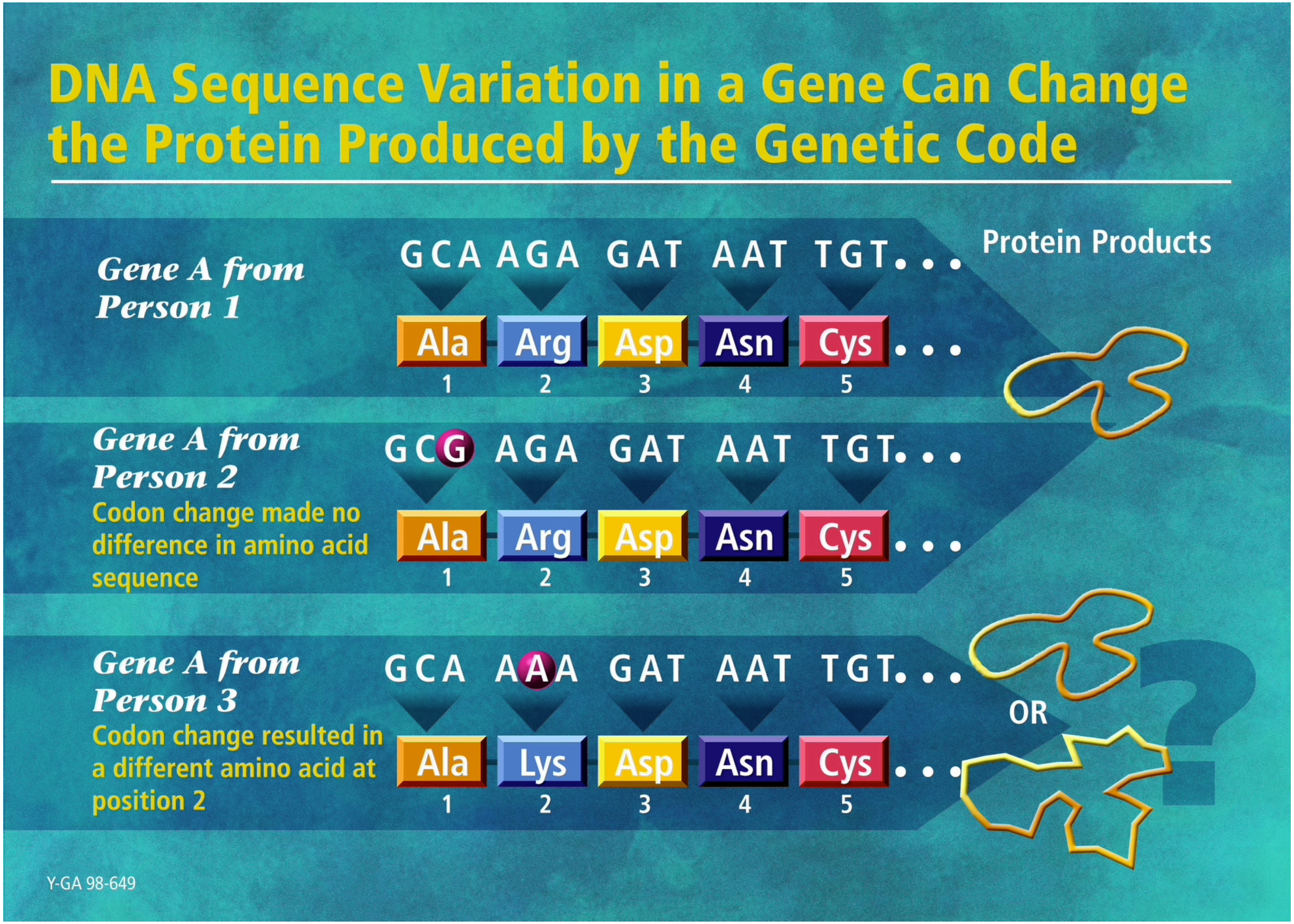
DNA Sequence Is the Genetic Code.

GCA AGA GAT AAT TGT . . .

Ala	Arg	Asp	Asn	Cys	. . . *Growing Protein Chain*
1	2	3	4	5	

Y-GA 98-648

Slide taken from DOE Human Genome Program website
http://www.ornl.gov/hgmis



DNA Sequence Variation in a Gene Can Change the Protein Produced by the Genetic Code

Protein Products

Gene A from Person 1

GCA AGA GAT AAT TGT...

Ala Arg Asp Asn Cys...

1 2 3 4 5

Gene A from Person 2

Codon change made no difference in amino acid sequence

GCG AGA GAT AAT TGT...

Ala Arg Asp Asn Cys...

1 2 3 4 5

Gene A from Person 3

Codon change resulted in a different amino acid at position 2

GCA AAA GAT AAT TGT...

Ala Lys Asp Asn Cys...

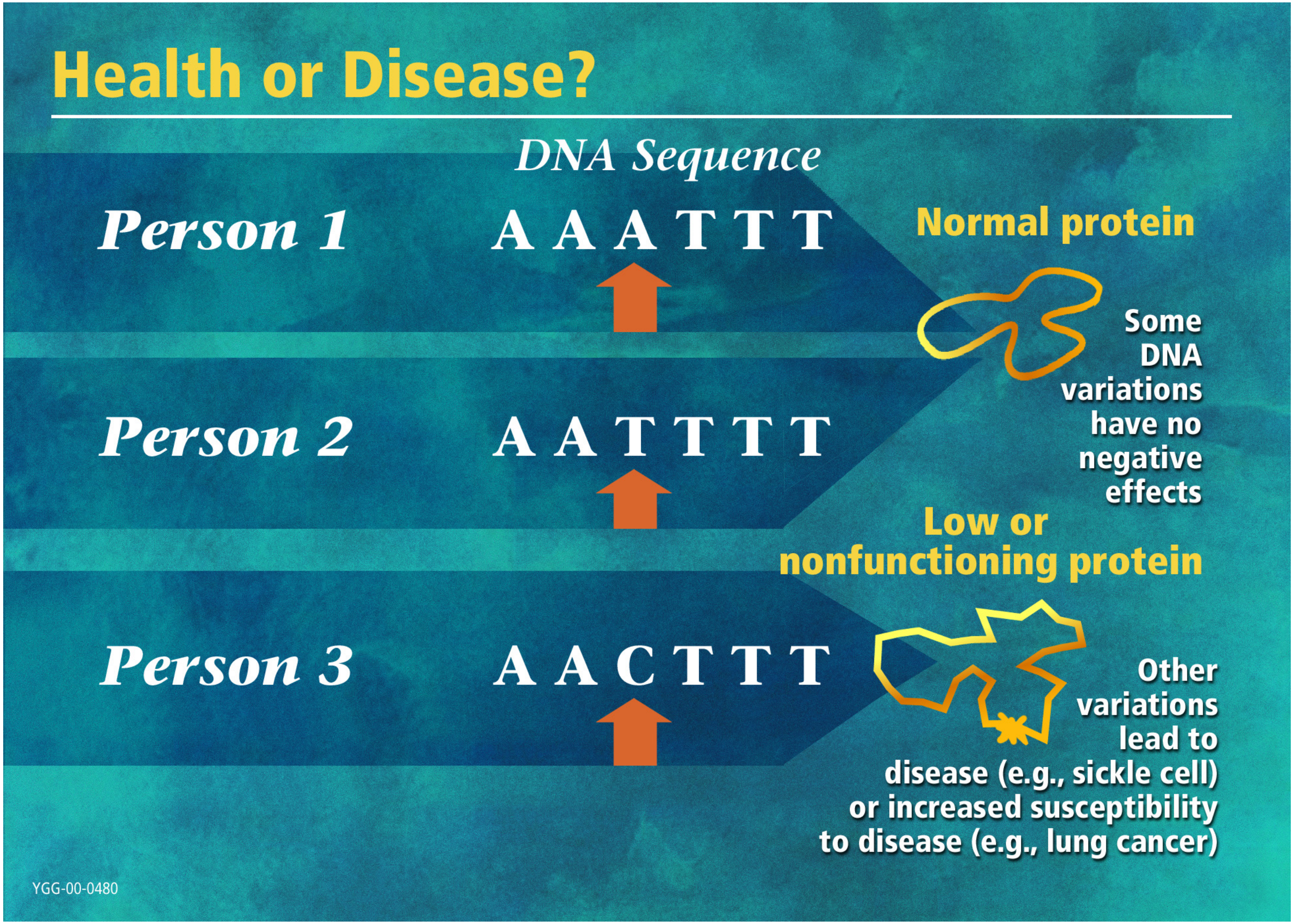
1 2 3 4 5

OR

?

Y-GA 98-649

Slide taken from DOE Human Genome Program website
http://www.ornl.gov/hgmis



Health or Disease?

DNA Sequence

Person 1 A A A T T T

Person 2 A A T T T T

Person 3 A A C T T T

Normal protein

Some DNA variations have no negative effects

Low or nonfunctioning protein

Other variations lead to disease (e.g., sickle cell) or increased susceptibility to disease (e.g., lung cancer)

YGG-00-0480

Slide taken from DOE Human Genome Program website
http://www.ornl.gov/hgmis

What do we use genome sequence for?

- Genes
- Splice Variants
- Variation analysis (SNP, mutations)
- Promoters
- Comparative Genomics
- Evolution

Promoters

- Make sure you have the 5’ end of the sequence (look at expressed sequences)
- Look for alternate starts
- Extract upstream regions

What do we use genome sequence for?

- Genes
- Splice Variants
- Variation analysis (SNP, mutations)
- Promoters
- Comparative Genomics
- Evolution

Comparative Genomics

- Look for genes in one species not found in another
- Look for genes in one species not present in another
- Gene Clusters / Synteny
- Look for conserved regulatory elements

General Considerations in choosing browsers

What do genome browsers look like?

- They have some representation of the sequence
  - Sometimes vertical
  - Usually horizontal (UCSC, Ensembl,TAIR)
  - Sometimes circular (Microbial Browsers)

- They have some form of annotation

DEFINE YOUR QUESTION

- If you don’t know what you’re looking for, it’s hard to find the answer

- Different questions require different browsers

Single organism vs. Multi organism

- Some genomes are available in multi organism browsers, others have their own web sites, and some are in both.
- For those that are in both, the question is what are you looking for, and where is it easier to find it?

Multiple genomes, same genus

- Sacchromyces
- Drosophila
- Microbes (strains)
- Dog, mouse strains

Release Date

- When was the sequence last updated?
- Always check (and WRITE DOWN) the version of the genome you worked with
- Different browsers may be using different releases of the same genome!

Genome Version (or Build)

- Each Browser has its own numbering system
- Most take the sequence from the same location
- There are two dating systems
  - Freeze date
  - Release date
- There are two types of release dates
  - Sequence Release
  - Annotation Release

Why do I get different coordinates if they use the same build?

- Different Browsers use different coordinate systems
  - Running count of the whole chromosome
  - Running count of the contig
- Double check the build!

What search options are there?

- Text search only?
- Different types of sequence searches?
  - Nucleotide, Protein, Translated
- Multi-species search
- Speed of search

Speed

- Blat is much faster than Blast
- Some sites give results back quickly, others take more time, or only by email
- Some sites just don’t load well

Data Mining

- UCSC (Table Browser)
- Ensembl (Ensmart)
- Phytozome
- IMG (Integrated Microbial Genomes and Microbiomes)

NNNNNN

- Different browsers/genome builds relate to gaps differently
- Most put at least 5 N’s to symbolize a gap
- Some put in a standard amount, no matter what size the gap
- Some try to represent the size of the gap with the number of N’s

Cookies - know who is in your lab!

- Some browsers use cookies, so they remember the last setting used on a particular computer

- If other members of your lab use the same computer, always check to make sure you are working in the correct genome and build!

Always check the assembly track

- You should always have the assembly/gap track open

- Open other tracks as needed to answer your questions

How do we look at a browser

Three main ways to enter:

- Text Search
- Sequence Search
- Browse

Text Search

Type in name of what you’re looking for: gene, marker, locus, location…

- Problems: You get where they think the best location is. If there is an alternate location, you won’t see it.
- If there are related sequences (psuedogenes, family members, duplications…) you won’t see them.
- If there are problems with the assembly, you won’t see them

Sequence Search

- If you have ANY sequence, use it!
- Run searches against NR and Genome
  - Some genome sequences aren’t in the genome assembly
  - Some mRNA sequences are deposited as DNA in Genbank
- If you are running a cross-species comparison, start with the protein
- Problem: you need sequence

Browse

- Usually, you can click on a part of the genome, and scroll in

- Usually, doesn’t give you much information unless you know where you’re going, or unless you’re preparing examples for a course

How do we look at what comes out of the search?

- Various annotations are kept in ‘tracks’ or ‘maps’
- There is generally a way to control which ones are visible
- There are generally controls to zoom in/out, or jump to particular locations

Keep in mind what the input is

- If you are looking at markers, expect a point to be highlighted

- If you are looking at an mRNA/genomic comparison there will be regions of match and regions of mismatch