Genome Projects

Shifra Ben-Dor

Bioinformatics and Biological Computing Unit,
Weizmann Institute of Science

WEIZMANN
INSTITUTE
OF SCIENCE

מכון
ויצמן
למדע

Outline

- What are genomes, and how were they sequenced?
- What can we do with genomes?
- How do we look at genomes?
- How do we choose a browser?
- How do the browsers work?

Why Study Genomes?

- Understand biological processes
- Understand pathological processes
- Diagnose, prevent, and cure diseases

So what is a genome?

- A genome is the full collection of genetic material (DNA) of an organism (including non-nuclear DNA, such as mitochondrial or chloroplast DNA)
- It is more than the protein coding genes (which are only a small percentage of the human genome)
- In humans there are 3,000,000,000 base pairs of DNA in the haploid genome

What else is in the DNA?
(aside from genes)

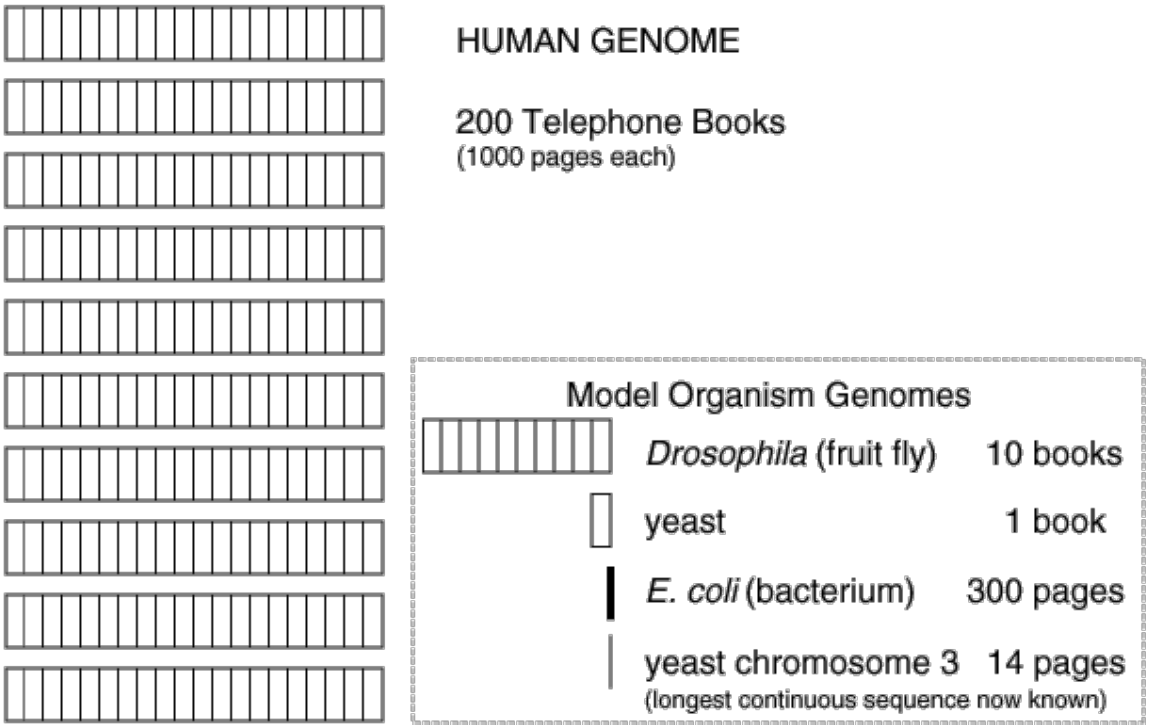
- There are areas that don’t code for genes at all
- There are areas that regulate the genes
  - When should a protein be expressed? night/day, fetus/adult
  - Where should a protein be expressed? eye, lung, muscle, brain

```
AGGTTAGATTATGCCCCGAGGGCGCCCCAGCCGAAATTTTTAATGCAGGTTTAATAGTTTAGAGC
CTGTGGCTTCCATGGCTTGGTTCTGCTGTTCTTCACTGGGGACTTGGGGGACCCTGGAGCTTG
TGATGGGGCCTGTCTCCACCTCTGTAAATCCAAGGAGTCAGATGACAAATCTGTCATTTCGGGCG
ACACACTCCCCTGAGGAAAGGGCCTTGCAGGAGGGCAGAGCAGCTTGCTGGGCATGGCAGGGAGT
GGAGAAGGGCAGGGGGCGCAGAGCAGGAGCAGCTTCCTGCCTCTGGGTGGGGACAGTGATCCCCA
CTGGGGACTGGCAAAGCCCCATGCTCTCTGTTCACCCTGGATGGGTGGCACCTGGGGGCAGGCAT
GGGGCCTGCAGGAGCCCCTGTGTGCCAGCCCTCCCCTGCCAGCATCCCATCTCCCAGGAGGCCCC
CAGGGCAGGTAAGTGCCAGGTCCCCCCTCAGCTCACCGTTGTCCTTCCCCTTGACGAACGCCTCC
CACTCCCGGAACCACTGCATGCTGATGCAGTAGATGACGCCCGGCGACTCCTCGGCCTGGAAGGC
CTTGTTCAACTGGCAGGCGGGCAGGGACGGGAGGAGACAGAGGGCAGGTCAGTCTCTATTTACCT
TCAGCAAGGATTTCCCAAATGCCCCCGCCCCAGTCCCTCACCCAAGAGTCTTACAAAAACACCAG
ATACTCAAGGTGAAAAATCAAACCCCCAGACAAGCTCAAACAAAATGAACACACCCATCACTCAG
AGAGGACGGTGGTAACATTTTGTGGGTCTCTTCAATAACTGTTTGACCCAACTGAGACCACAAGG
GAGATTCTACTTTTTGAGAAGGAATCTCACTCTGTCACCCAGGCTGGAGTGCAGTGGCGCGATCT
CGGCTTACTGCAACCTCTGCCTCCCAGGTTCAAGCGATTCTCCCACCTCAGCCTCTTGAGTAGCT
GGGATTACAGGTGTGTGCCACCACCTGGCTACTTTTTGTGTTTTTAGTAGAGACGGGGTTTCGCC
ATGTTGGCCAGGCTGGTCTTGAACTCCTGACCTCAGGTGATCCGCCCACCTCGACCTCTCAAAAG
TGCTGGGATAACAGGCATGAACCACTGCGCCCGGCCTGGGAGATGCTAATTTTCTCCGGTTGAAT
AGAATGTGCCTATCTGCTCAGAGAGGCAGCTCTCCTTCTGACAGGAGCATTTTCTTTTTCGAGAT
GGGGGGTGGTCTCACTCTGTGCCCAGGCGGGAGTGCAGCGGCGCAATCATGGCTCACTGCAGCCT
TGACCTCCTGGGCTCAAGTGATCCTCCCACCTCAGCCTCCTGAGTAGGTTGGACCACAGGTGCAT
ACCACTAGGCCCAGCCCTGACAGTCTCTTTTTCGTTTGTGTTCTGAGACAGGGTCTCACTCTATT
GCCCAGGCTGCGGTGCAGTGGCATGATCACGGCTCACTGCAGCCTCAACCTCCCAGGCTTAGGTG
ATCCTCCCAACTCACTCAGCCCTCCAGGTAGCGGGGACTACAGGTACACATCACCATGCCTGGCT
AATTTTTGTATTGTTTGTAGAGATGGGGTTTCGCCATGTTGGCCAAGTTGGTCTTGAACTCCTGG
```

Challenges of DNA

- DNA has only four letters
- They are strung together with NO obvious punctuation
- There are no signals to say “a gene starts (or ends) here”
- How do we make sense of so much information?

ORNL-DWG 91M-17472



HUMAN GENOME

200 Telephone Books
(1000 pages each)

Model Organism Genomes

Drosophila (fruit fly)	10 books
yeast	1 book
E. coli (bacterium)	300 pages
yeast chromosome 3 (longest continuous sequence now known)	14 pages

If compiled in books, the data would fill an estimated 200 volumes the size of a Manhattan telephone book (at 1000 pages each), and reading it would require 26 years working around the clock.

A T C G
MEGA
GENOME
WINNER

The Human Genome Project

- Was planned in 1988, started in 1990
- Originally planned to take 15 years, in three five year stages

Objections to the genome project

1) Fear that funding will be diverted from other areas of research

2) What is the value of sequencing a complete genome, given the high proportion of nongenic sequences ("Junk DNA").

Two alterations in the original plan helped:

1) Focus shifted from large-scale sequencing to mapping the genome, which would hasten the search for disease genes

2) Simultaneously determine the nucleotide sequence of the genomes of other organisms; this provides comparisons and points of reference for the human sequence

Stage 1 1991-1995

- Creating a Genetic Map of the genome
- Creating a Physical Map of the genome
- Creating a set of overlapping clones
- Create faster/cheaper methods of sequencing
- Create software/databases that can deal with the data

Stage 2 1994-1998

- Finish mapping (both genetic and physical)
- Start sequencing
- Start annotation - gene finding and placement on maps

Stage 3 1998-2003

Area	Goals 1993-98	Status as of Oct. 1998	Goals 1998-2003
Genetic map	Average 2- to 5-cm	1 cm map published Sept. 1994	Completed
Physical map	Map 30,000 STSs	52,000 STSs mapped	Completed
DNA sequence	Complete 80 Mb for all organisms by 1998	180 Mb human plus 111 Mb non-human	Finish 1/3 of human sequence by end of 2001 Working draft of remainder end of 2001 Complete human sequence by end of 2003
Human sequence variation	Not a goal	-	100,000 mapped SNPs
Gene Identification	Develop Technology	30,000 ESTs mapped	Full-length cDNAs
Functional analysis	Not a goal	-	Develop genomic-scale technologies

Techniques

The major breakthrough that allowed the Human Genome Project to take off were two techniques:

1) Improvements in sequencing:
   Cycle Sequencing
   Automated Sequencers
   Flourescent Dyes
   High Throughput - lower costs

2) PCR

Markers

There are many different types of markers:

RFLP: Restriction Fragment Length Polymorphism

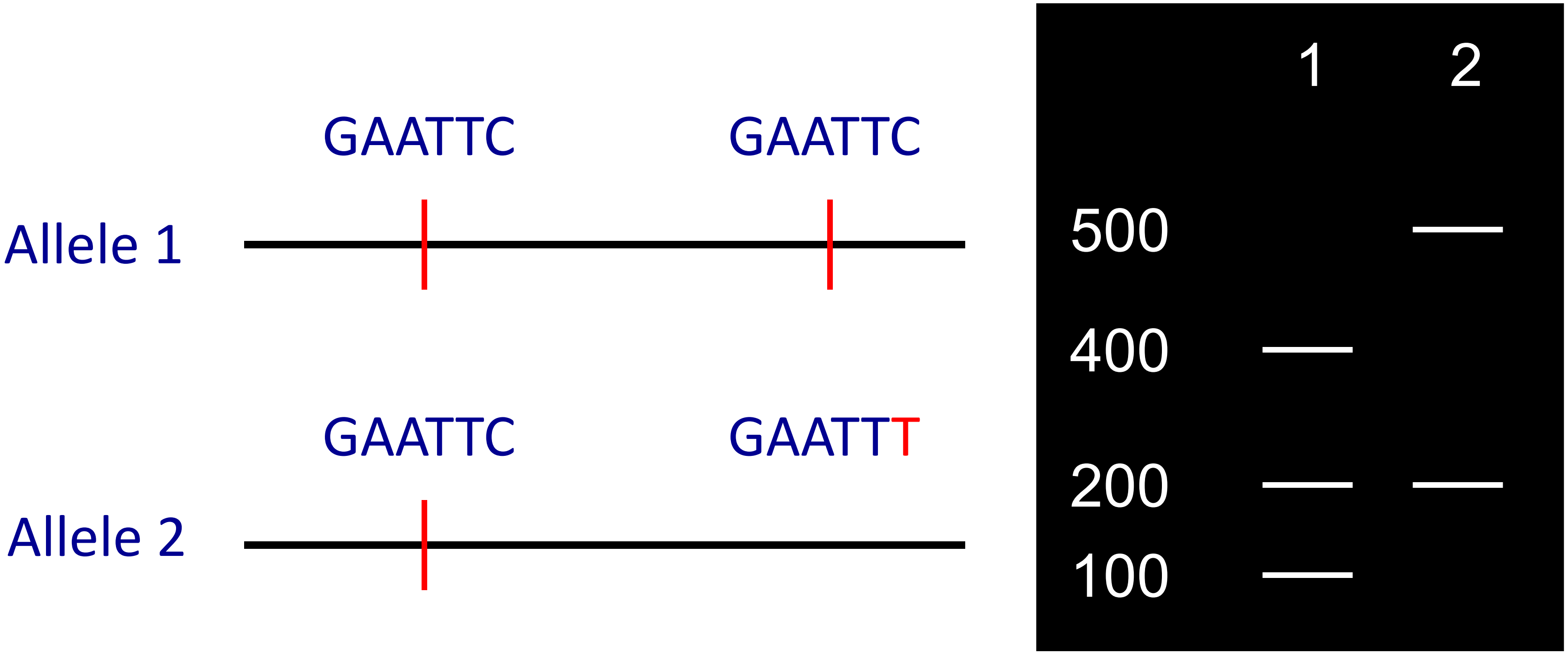
Microsatellite

VNTR: Variable Number Tandem Repeat

STS: Sequence Tagged Site

Restriction Fragment Length Polymorphism

700 base pair PCR fragment



Microsatellites

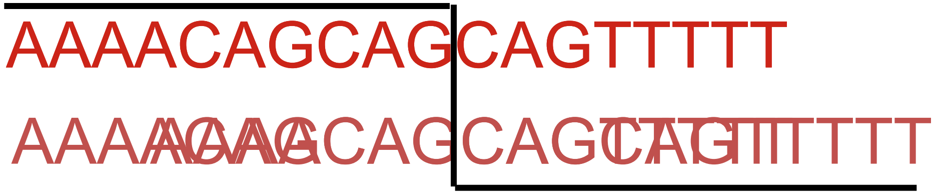
Microsatellites are small repetitive stretches of DNA, usually repeats of di, tri and tetra nucleotides.

For example CACACA, or CAGCAGCAG…..

Because these stretches repeat, when the DNA recombines, its very easy for the machinery to "slip" and add or subtract a few copies

Recombination

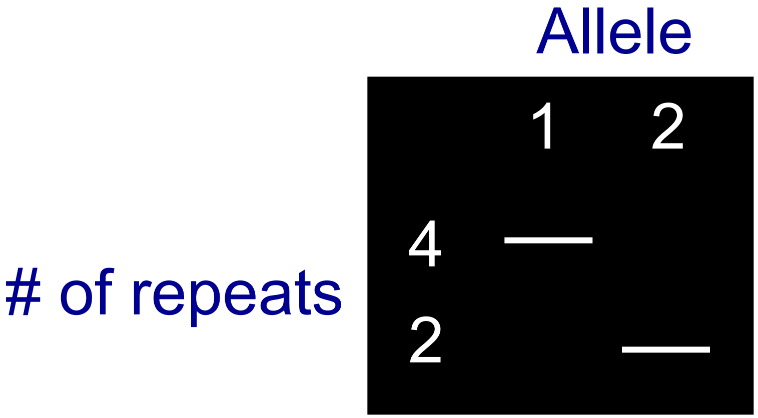
Parent chromosomes



Daughter chromosomes

Allele 1 AAAACAGCAGCAGCAGTTTTT

Allele 2 AAAACAGCAGTTTTT



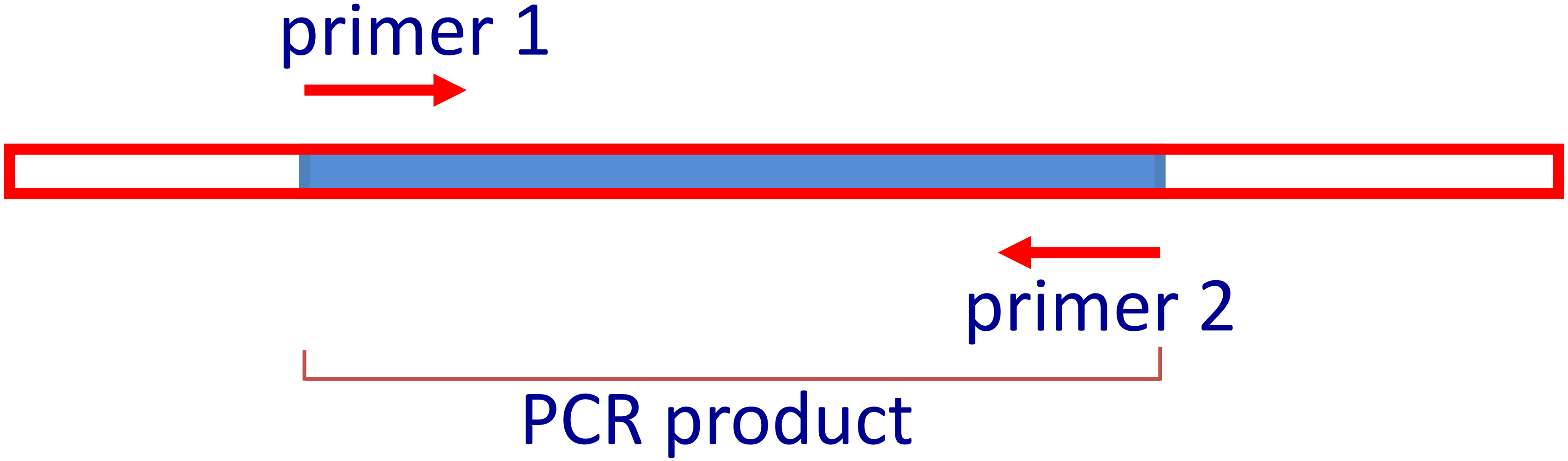
Variable Number Tandem Repeat

Regions of DNA that repeat - for example, microsatellites, although VNTR's can have more complex sequence.

This is also detected by performing PCR and looking at the number of repeats on a gel

Sequence Tagged Site

A genomic region



primer 1

primer 2

PCR product

STS is defined as:

→ primer 1

→ primer 2

PCR product size

An STS is designed to be unique in the genome

Maps

There are two types of maps:

Genetic - measures recombination distance

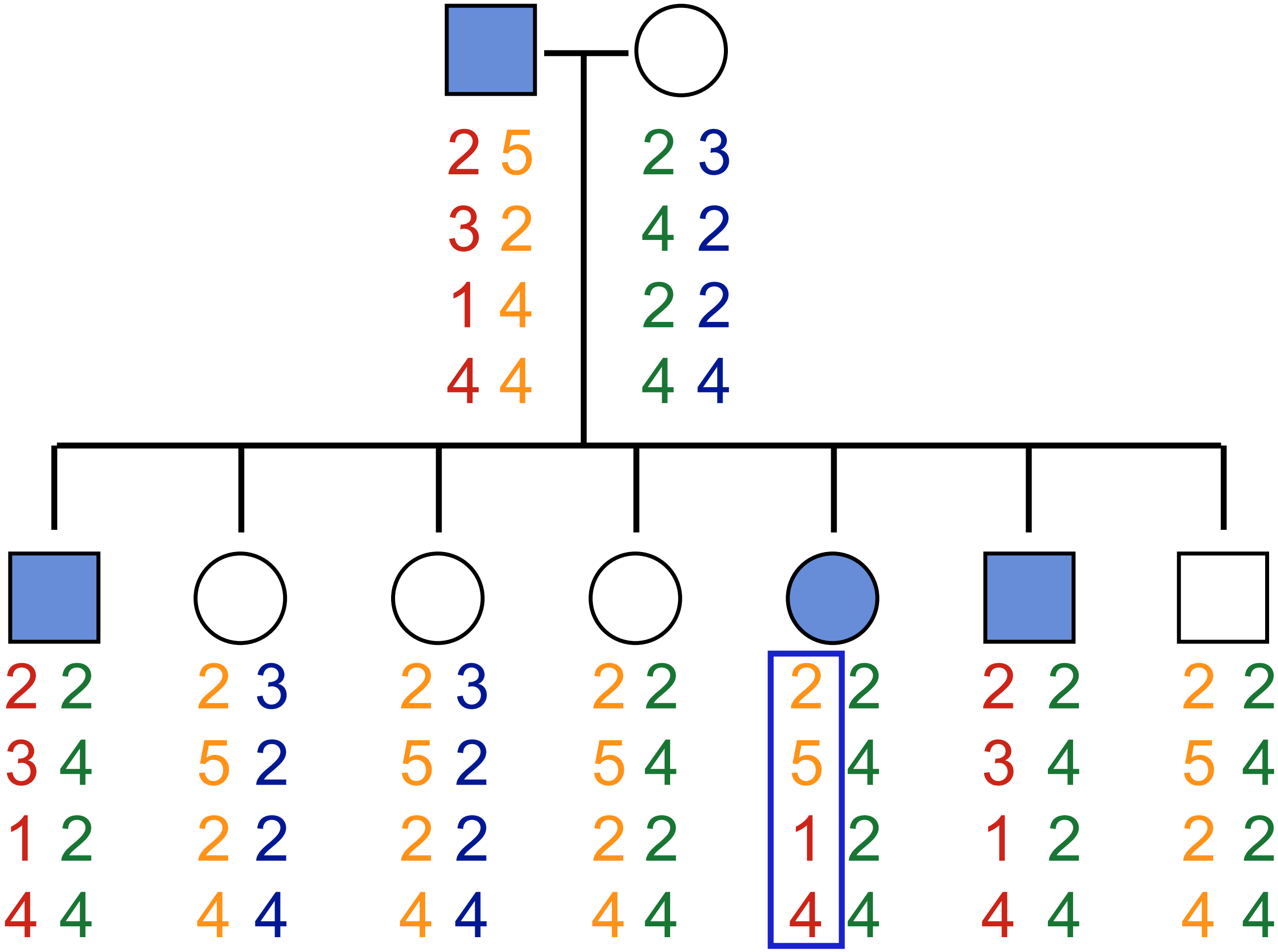
Physical- measures physical distance

Genetic Maps

A genetic map measures recombination distance and answers the question, “How often are two markers found together?”

Two markers are said to be 1 centiMorgan (cM) apart if they are separated by recombination 1% of the time. A genetic distance of 1 cM is roughly equal to a physical distance of 1 million basepairs (1 Megabase or Mb)

Genetic Maps

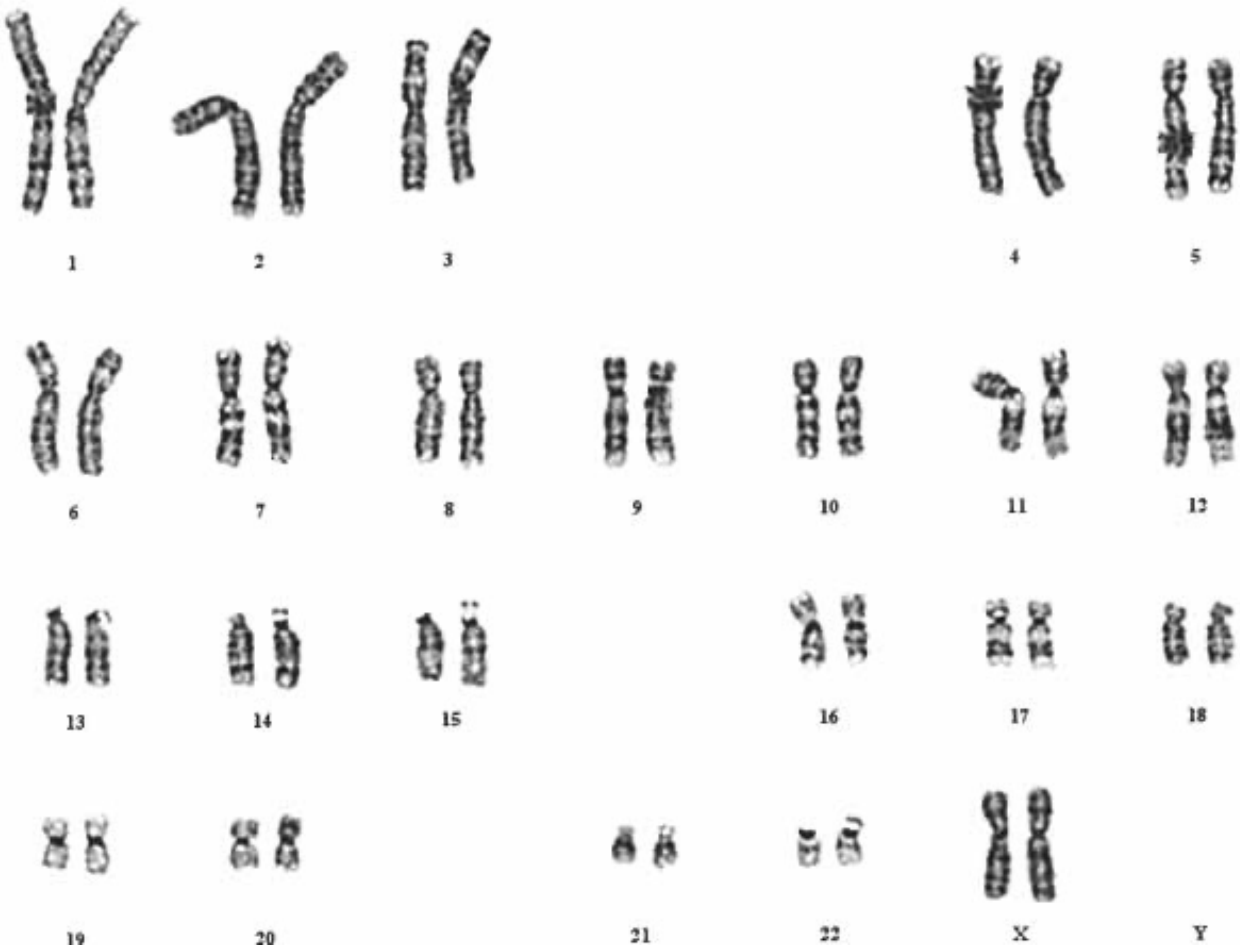


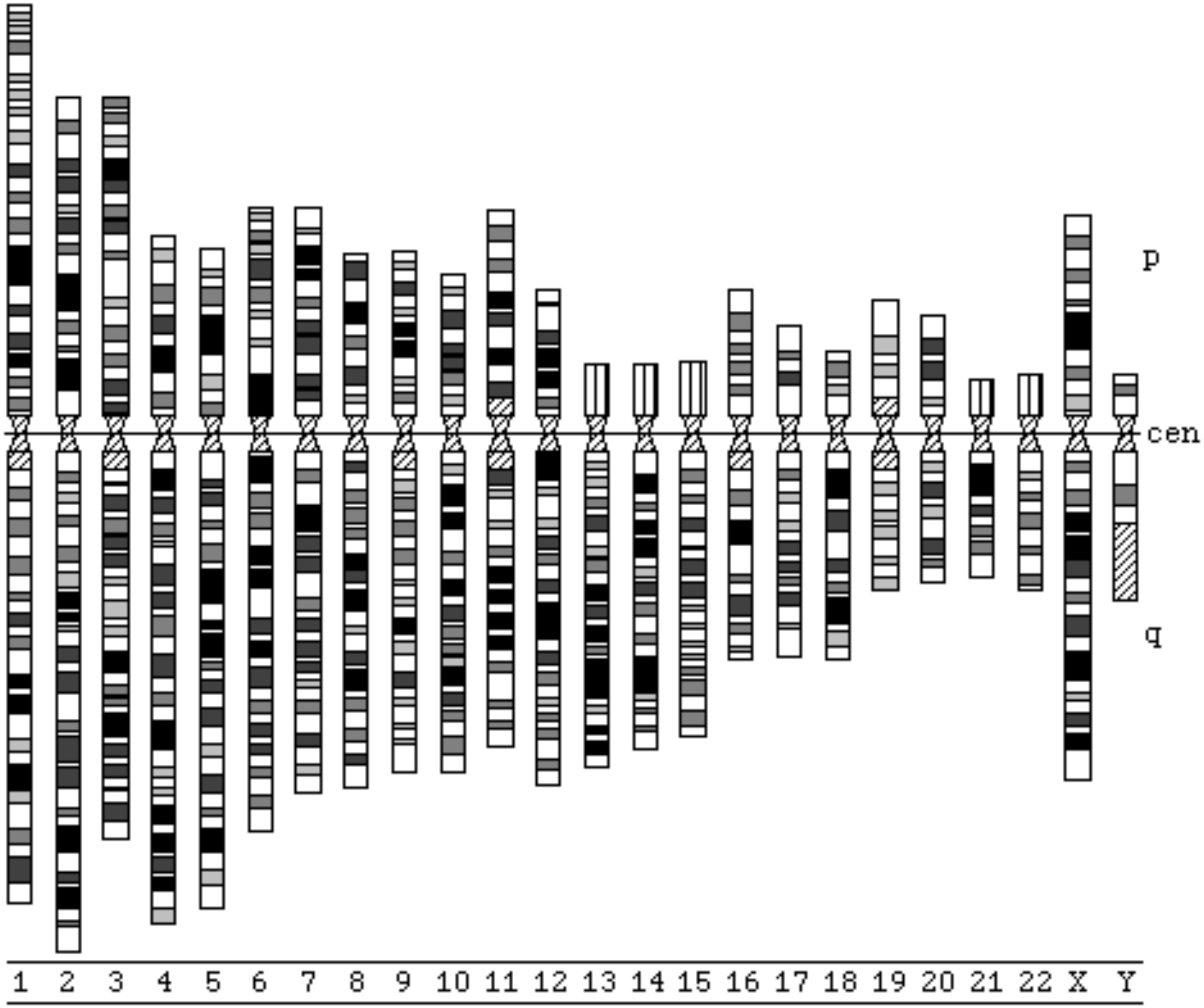
Physical Maps

Physical maps vary greatly. The lowest resolution map are the chromosome banding patterns (ideogram).

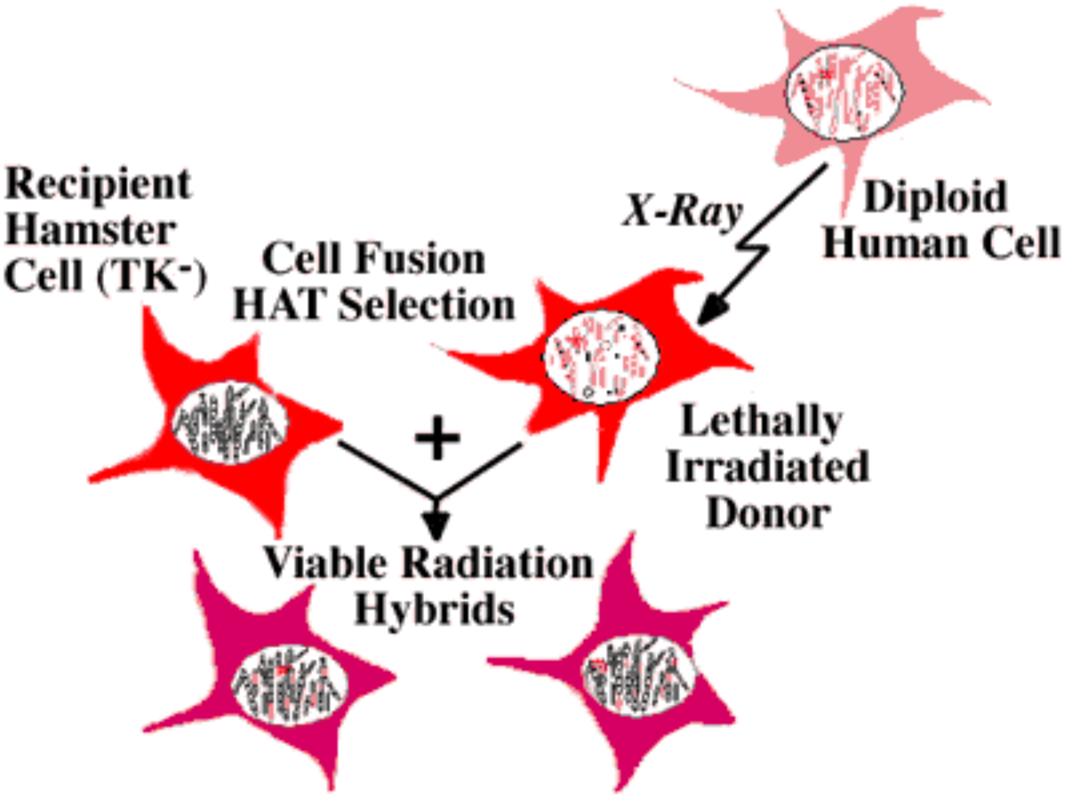
The highest resolution map is the actual sequence.

What we actually use is in between. One type of map developed as part of the genome project is the radiation hybrid map (RH map)


1
2
3
4
5
6
7
8
9
10
11
12
13
14
15
16
17
18
19
20
21
22
X
Y


p
cen
q
1
2
3
4
5
6
7
8
9
10
11
12
13
14
15
16
17
18
19
20
21
22
X
Y

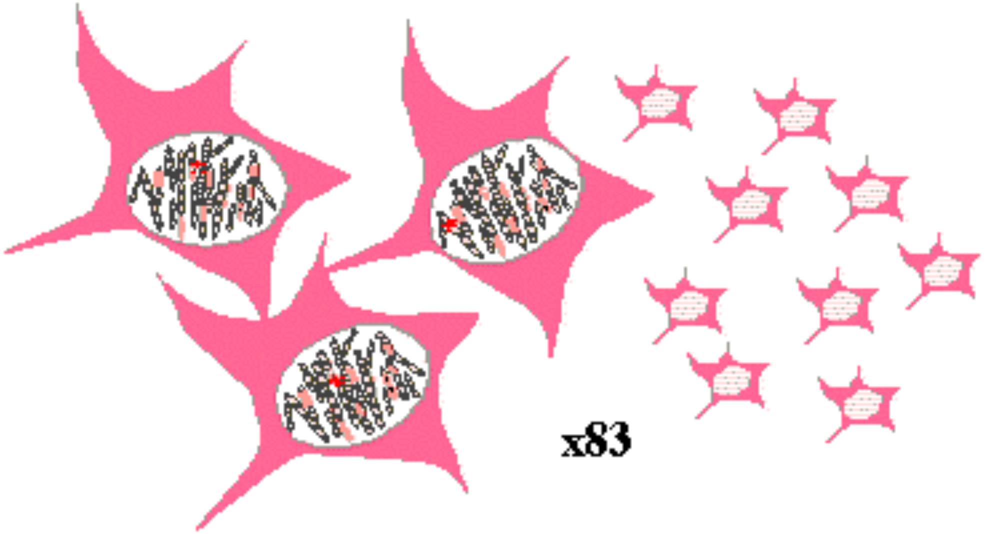
"Whole Genome" Radiation Hybrid Mapping



shgc

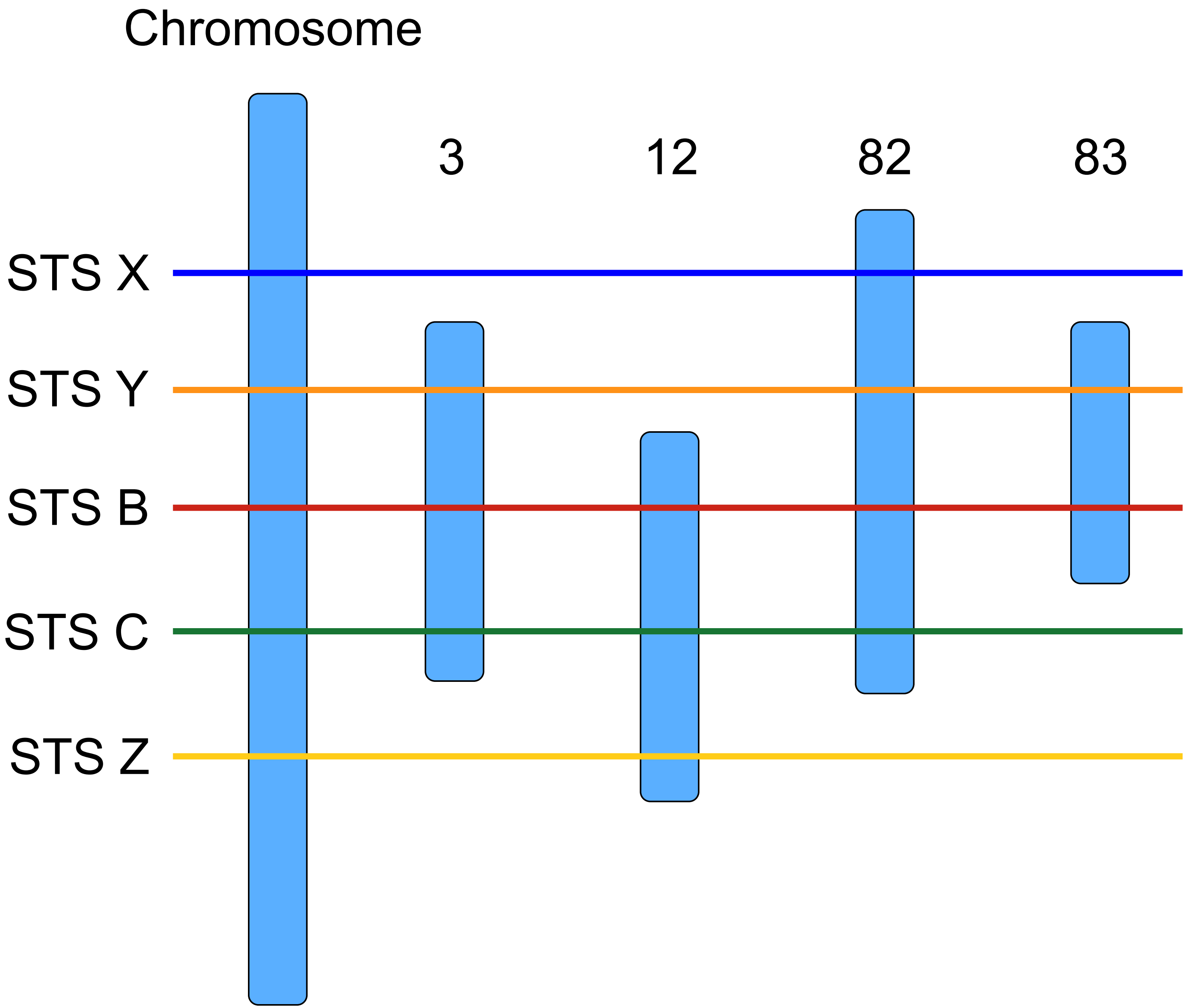
Scoring STSs on a Radiation Hybrid Panel

Radiation Hybrid Panel



Scoring Results

	1	2	3	4	5	6	7	8	9	10	11	12		81	82	83
STS-A	●	○	○	○	●	●	○	○	○	○	○	○	. . . .	○	○	○
STS-B	○	○	●	○	○	○	○	○	○	○	○	●	. . . .	○	●	●
STS-C	○	○	●	○	○	○	○	○	○	○	○	●	. . . .	○	●	○

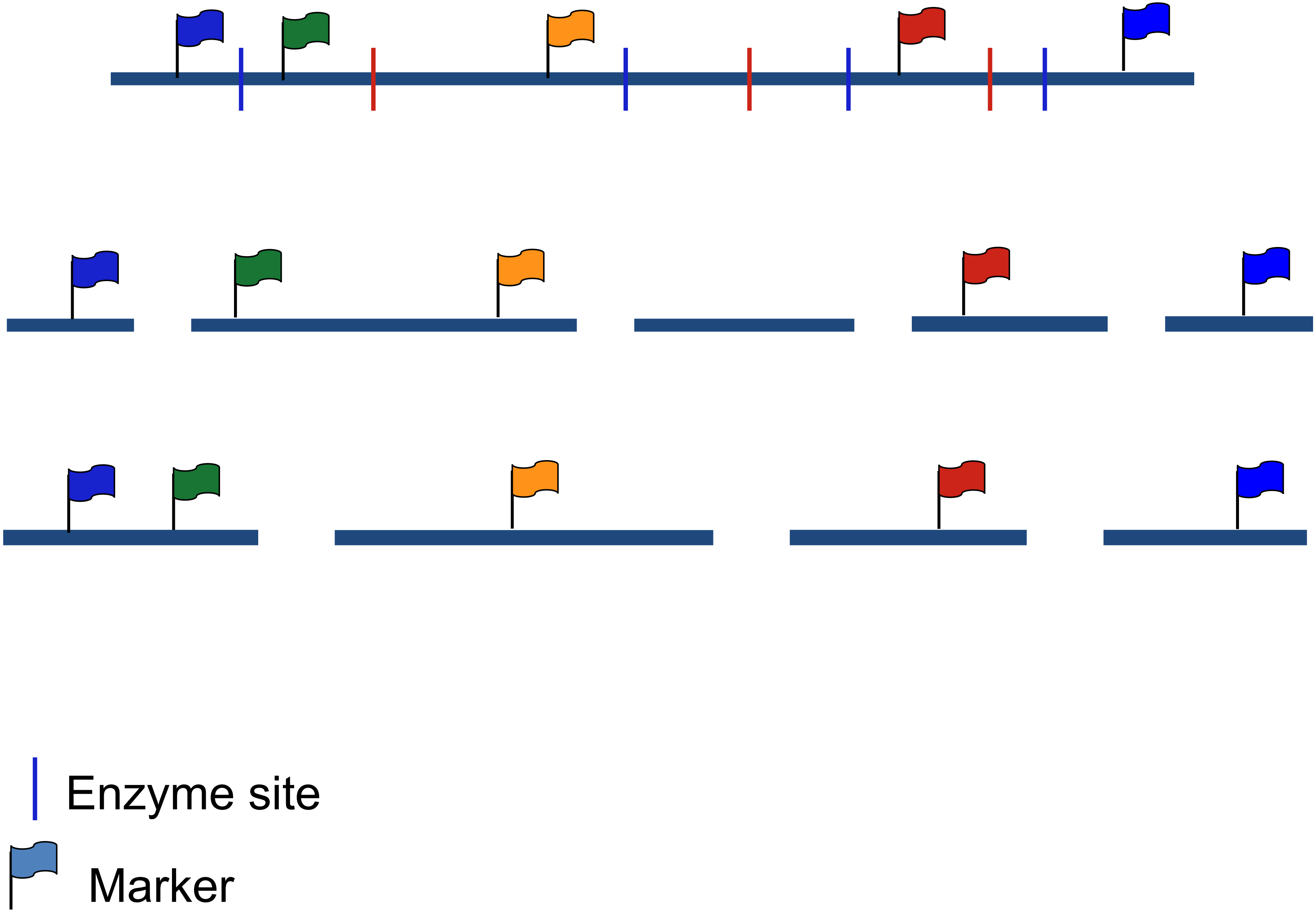

Chromosome
3
12
82
83
STS X
STS Y
STS B
STS C
STS Z

High Resolution (TNG4) Whole Genome Radiation Hybrid Mapping Panel Characteristics

Number of hybrids in mapping panel	90
Fraction of human genome retained in each hybrid	0.16
Average size of human fragments	800 kb
Relationship of X-ray breakage to distance	1% breakage = 4 kb
Average resolution of comprehensive map	60 kb
Average resolution of 1000:1 map	100 kb

How was the genome cloned?

First rare cutting enzymes were used to generate pieces ~150,000 base pairs long. Different enzymes were used to get overlapping segments.

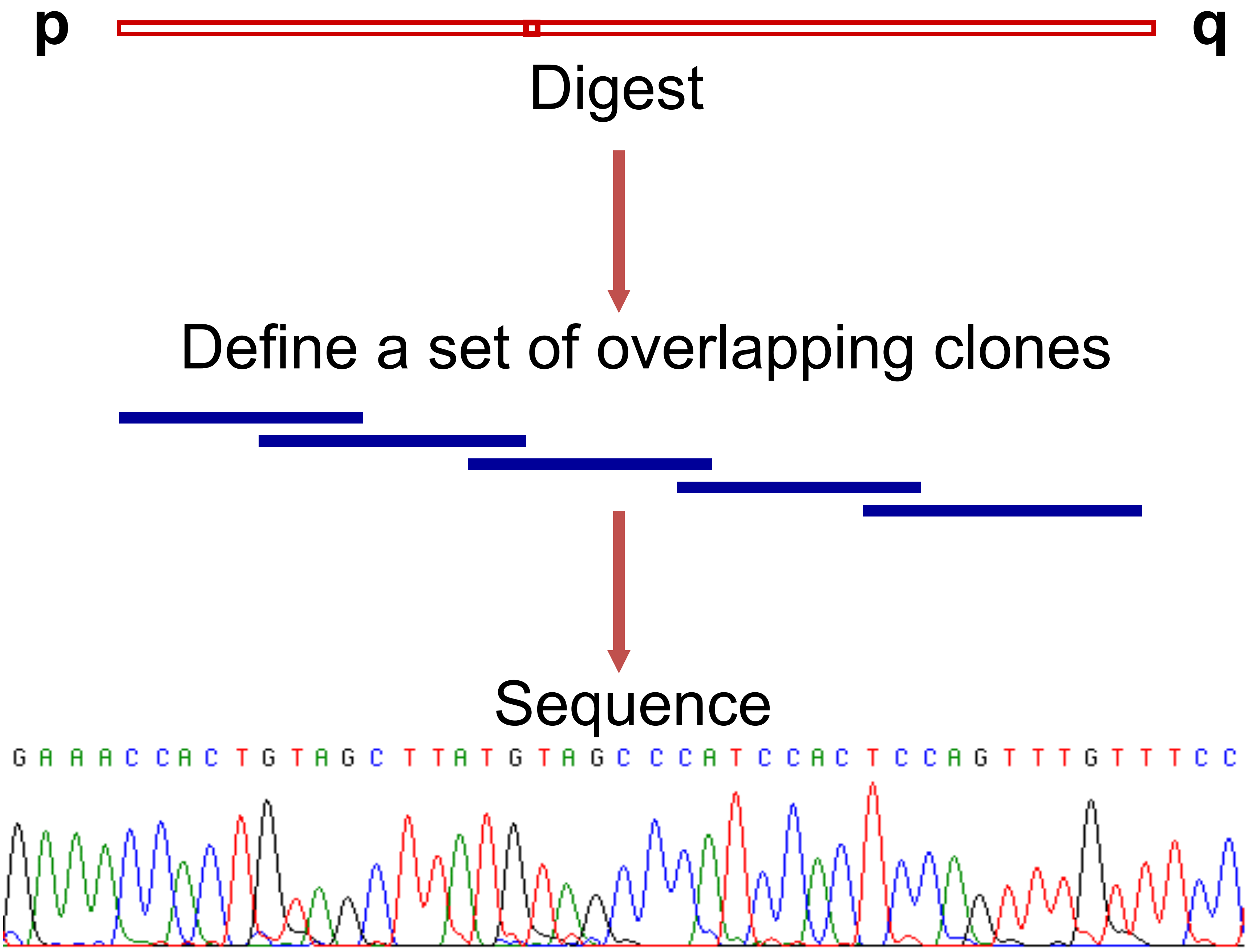


Enzyme site

Marker

BACs (Bacterial Artificial Chromosomes) were placed along the map with the help of markers, and the least redundant path was chosen.

Then each BAC was broken down the same way, the fragments were placed in order using restriction enzyme mapping, and sequenced.



p

q

Digest

Define a set of overlapping clones

Sequence

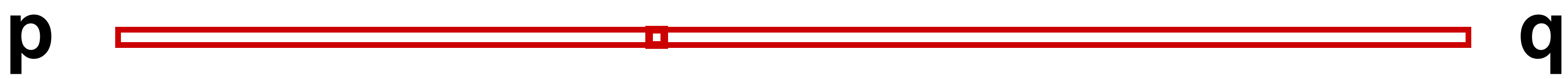
G A A A C C A C T G T A G C T T A T G T A G C C C A T C C A C T C C A G T T T G T T T C C

Then came Celera…..

Celera Corp. said that it would sequence the genome faster and cheaper than the public project could.

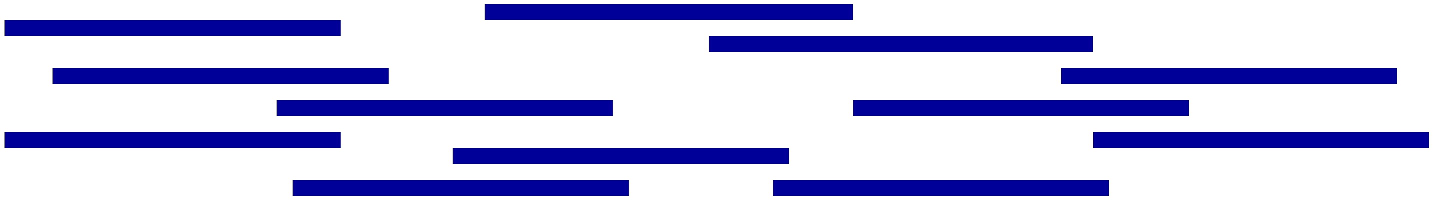
They used a method known as random shotgun sequencing.

They broke the genome up into 2000 and 10,000 bp pieces, sequenced them, and wrote a computer program to put it all back together.

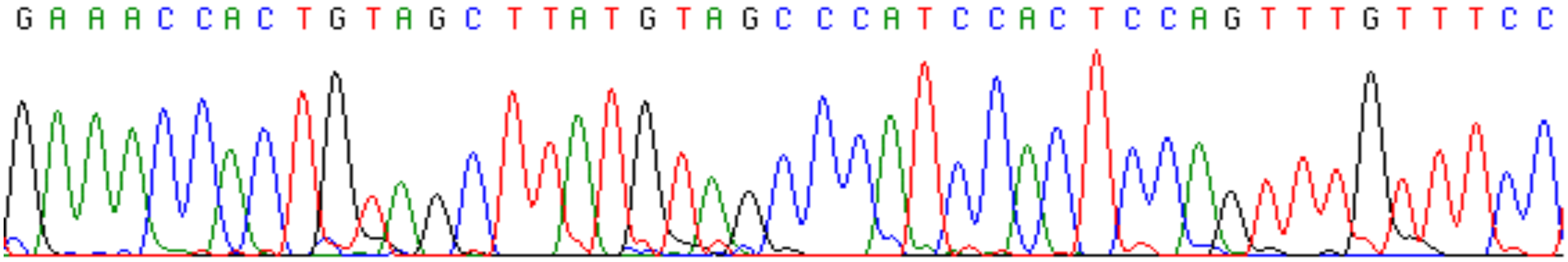


Shotgun digest

Take random clones



Sequence

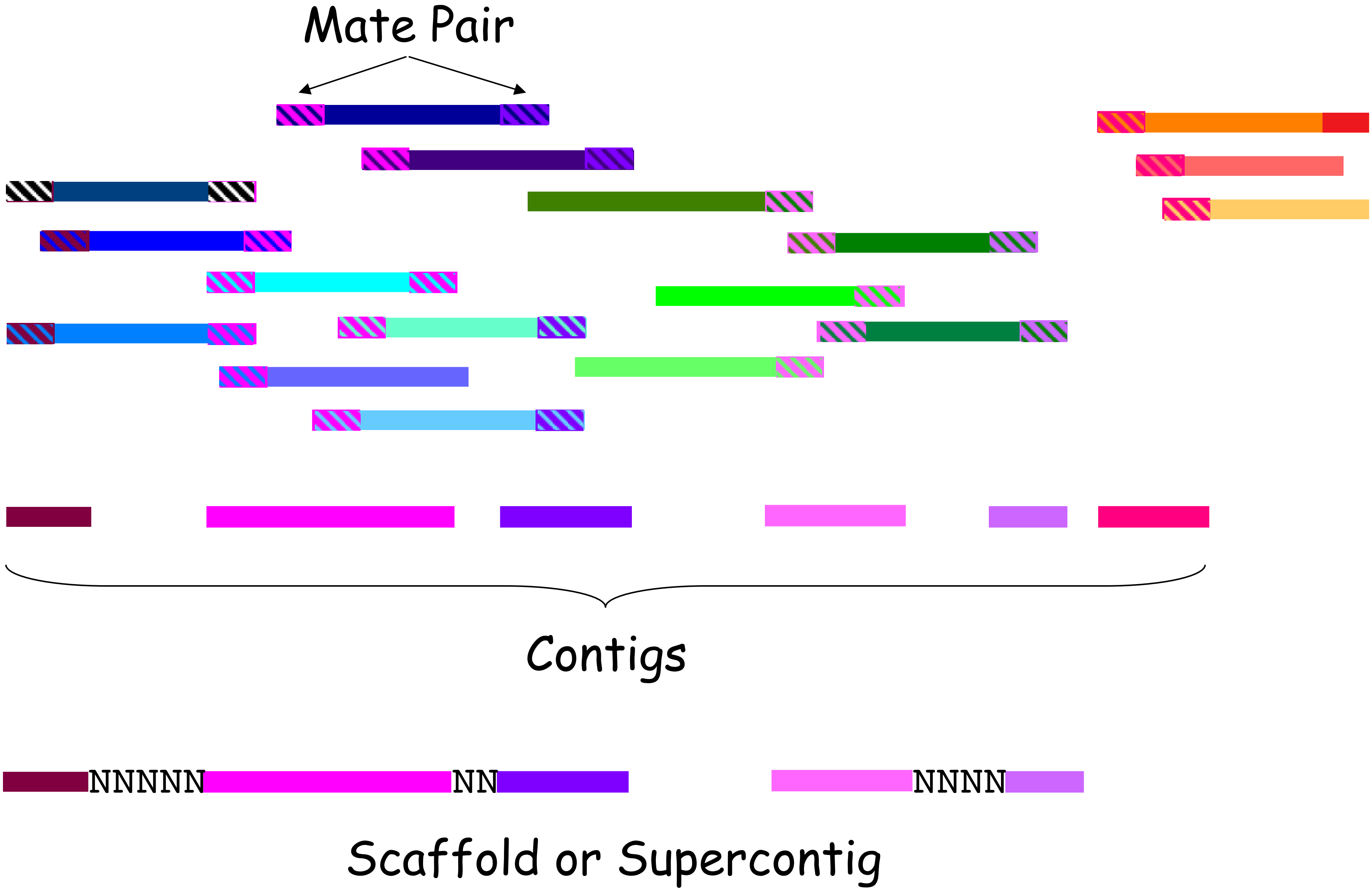


Whole Genome Shotgun

Whole genome shotgun didn’t actually sequence whole pieces of DNA, but just their ends.

The WGS reads average ~500bp.

They do however give information in terms of matepairs as to which piece falls where, and that gives a method of ordering and a measure of the size of the holes



Mate Pair

Contigs

NNNNN NN NNNN

Scaffold or Supercontig

It is important to note that not all genome sequences are organized into chromosomes. Those that are being sequenced by whole genome shotgun, where mapping has not taken place, are organized into “linkage groups”

The problems with shotgun…

There are several problems with shotgun sequencing:

- repetitive elements

- gene families

- need more sequence

“x” coverage

- What does 5x coverage of the genome mean?
- That 5 times the number of bases in that genome were sequenced (so for human 5 x $3x10^9$ bases)
- It does NOT mean that the whole genome was covered 5 times

The public reaction….

The public project was concerned that Celera would finish first, and as a commercial company, try to patent significant parts of the genome.

To make the public effort faster, they decided to shotgun the existing BACs.

This lead to HTGs

High Throughput Genome sequence

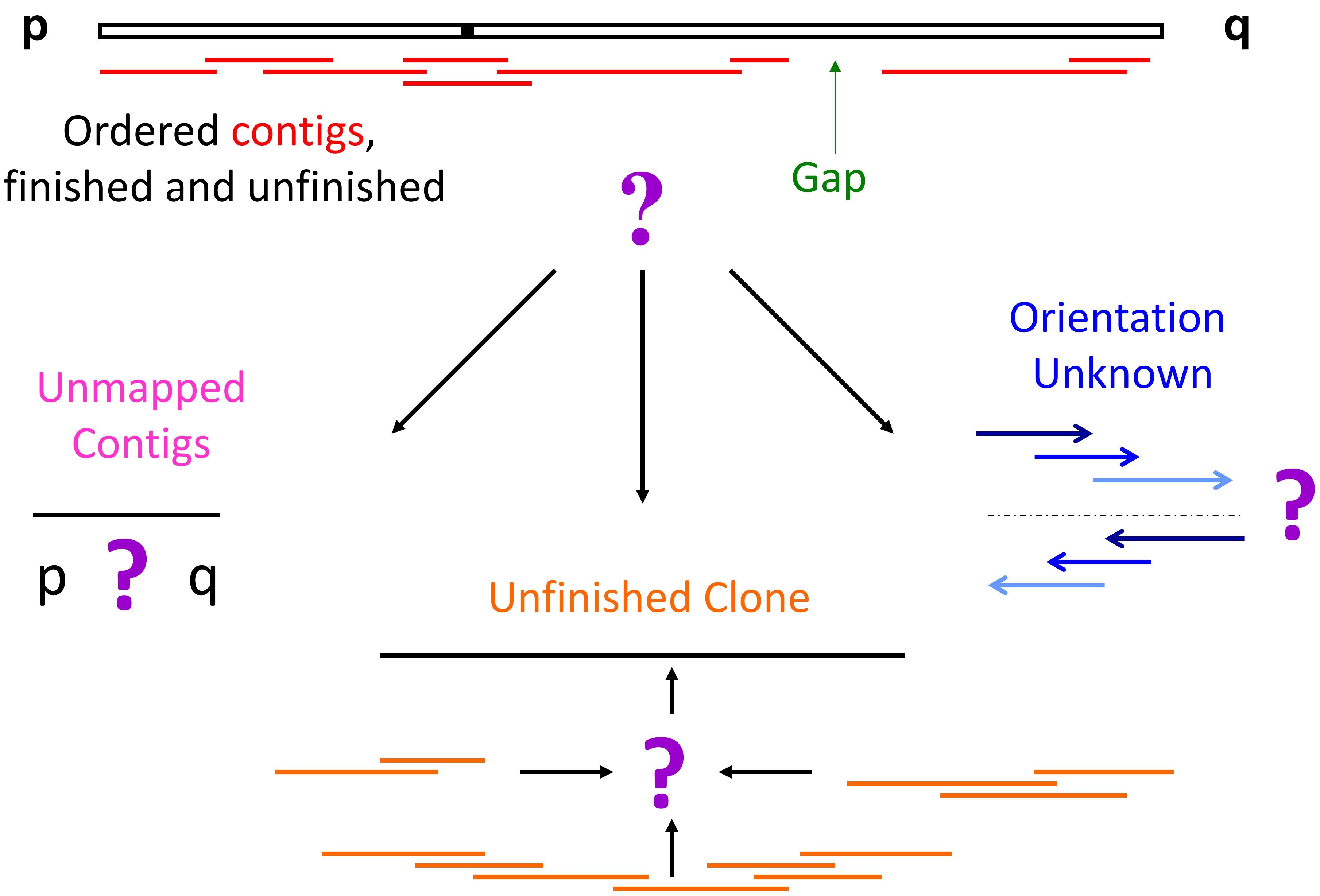
The output of the public project had been well ordered well finished sequence.

As a result of shotgun sequencing, we ended up with “draft” sequence – sequence whose general location is known, but the exact order or direction of the pieces is not.

HTGS

Status	Location	Definition
Phase 0	HTG division	single-few pass reads of a single clone (not contigs).
Phase 1	HTG division	Unfinished, may be unordered, unoriented contigs, with gaps.
Phase 2	HTG division	Unfinished, ordered, oriented contigs, with or without gaps.
Phase 3	Primary division	Finished, no gaps (with or without annotations).

Draft Sequence - Spring 2000



p

q

Ordered contigs,
finished and unfinished

Gap

?

Unmapped
Contigs

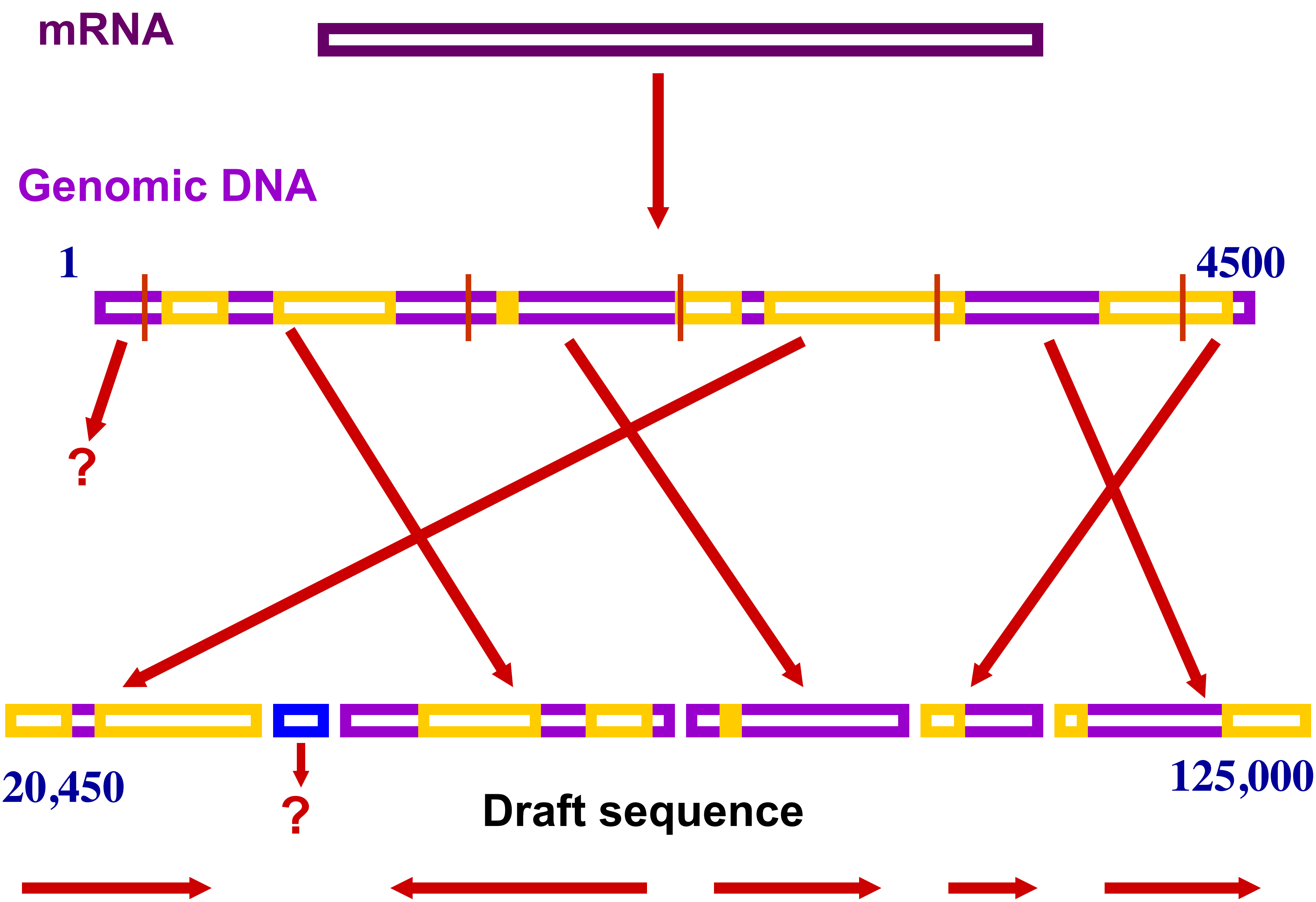
p ? q

Orientation
Unknown

?

Unfinished Clone

?

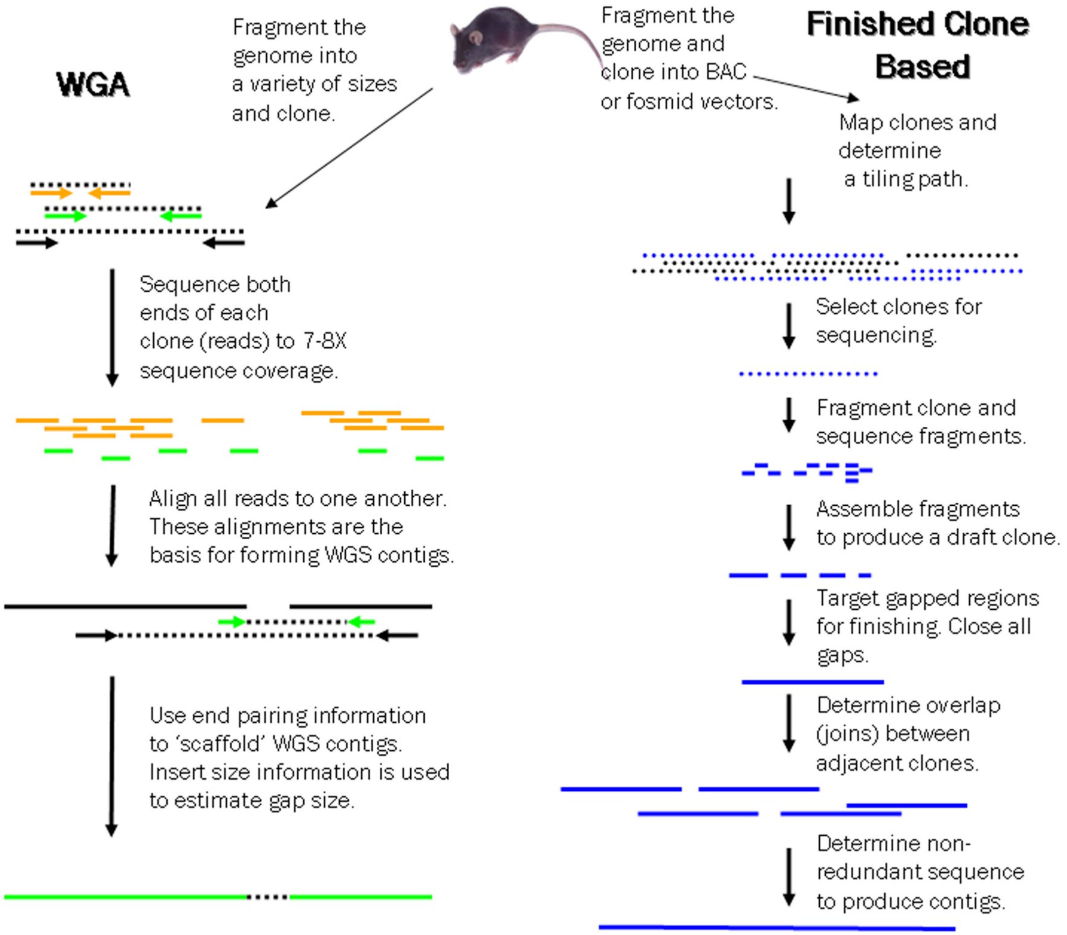

mRNA
Genomic DNA
1
4500
?
20,450
?
125,000
Draft sequence

Sequencing Problems

- Physical Holes (no clones)
- Sequence Holes (no sequence)

Assembly Problems

- Repetitive elements
- Gene families
- Pseudogenes
- Duplication



WGA

Fragment the genome into a variety of sizes and clone.

Sequence both ends of each clone (reads) to 7-8X sequence coverage.

Align all reads to one another. These alignments are the basis for forming WGS contigs.

Use end pairing information to 'scaffold' WGS contigs. Insert size information is used to estimate gap size.

Finished Clone Based

Fragment the genome and clone into BAC or fosmid vectors.

Map clones and determine a tiling path.

Select clones for sequencing.

Fragment clone and sequence fragments.

Assemble fragments to produce a draft clone.

Target gapped regions for finishing. Close all gaps.

Determine overlap (joins) between adjacent clones.

Determine non-redundant sequence to produce contigs.

Pluses:
- Relatively quick to produce a draft assembly
- No physical map required

Things to watch for:
- Lower accuracy leading to frameshifts and in-dels.
- Lineage specific regions absent or poorly represented as unplaced sequences.
- Thousands of gaps

Pluses:
- High per base accuracy rate with few frameshifts and in-dels.
- Better representation of lineage specific sequences.
- Few gaps, mostly in very difficult regions.

Things to watch for:
- Requires physical map and tiling path
- Finishing gaps requires extra effort and time.

All assemblies require:
- Analysis and comparison to other genomic resources
- Improvement and maintenance

Figure Legend

┈┈┈ Unsequenced clone

━━━ WGS contig

→ → → End sequence reads

┈┈┈ Unsequenced BAC from tiling path

━━━ Sequenced BAC

Taken from: Church DM et al (2009) Lineage-Specific Biology Revealed by a Finished Genome Assembly of the Mouse. PloS Biol 7(5): e1000112

Genome sequencing with NGS

- NGS is quicker and cheaper than older methods
- Shorter pieces makes the problem of putting everything together worse
- Long read technology can help bridge gaps, and may be the way forward, but still has a high error rate

The post-genomic era…..

Now that we have most of the genome sequence, efforts are being turned to understanding the wealth of information produced:

Defining the genes, when and where they are expressed, what controls them, who they interact with, and how they are mutated, particularly in disease.

Advantages of Genome Sequence

- Previously, it was “one gene, one postdoc”
- Now that we have a better picture of things, we can study systems, and gene interactions
- Previously, it took years to clone and sequence a gene
- Now, all we need is a little bit of sequence,
  and we can look up the rest in the genome

Advantages of Genome Sequence

- Previously, after years spent cloning, more time was needed to sequence the surrounding area of the gene, to start looking into regulatory elements
- Now, we have the surrounding sequence and can start looking for the regulatory elements directly

So we started with genomics….

Now we have “Omics”:

- Transcriptome
- Proteome
- Regulome…….

And still plenty of work…….

States of Common Genomes

No Eukaryotic Genome is Truly Finished

- Euchromatin/Heterochromatin
- Duplications
- Gaps
- Alternate alleles/regions
- Multiple sources

T2T Consortium

- Telomere to telomere coverage of a human genome
- But:
  - Didn’t sequence a normal genome (hydatidiform mole)
  - No Y chromosome
  - Only one source
- New challenge: Human Pangenome Reference Consortium

N50

- A common measure for how well put together a genome is
- The mark where 50% of the clones are above or below that length
- N50 human: 67.8 Mb (build 38)
- N50 mouse: 54.5 Mb (build 38)

Y Chromosome

“The Y chromosome in this assembly contains two pseudoautosomal regions (PARs) that were taken from the corresponding regions in the X chromosome and are exact duplicates:

```
chrY:10001-2649520 and chrY:59034050-59363566
chrX:60001-2699520 and chrX:154931044-155260560
```

UCSC genome browser, human genome page

Mouse things to remember

- Strains
  - The reference sequence is from C57Bl/6J
  - There is sequence available from: Balb/c, C3H, NOD, and differences from 15 strains
- Y chromosome - the original reference mouse was female!

http://www.genomesonline.org

Ideas and slides taken from:

Irit Orr

Vered Chalifa Caspi

Dolan DNA Learning Center

Human Genome Project at DOE

YourGenome - Sanger Center