EECS730: Introduction to Bioinformatics

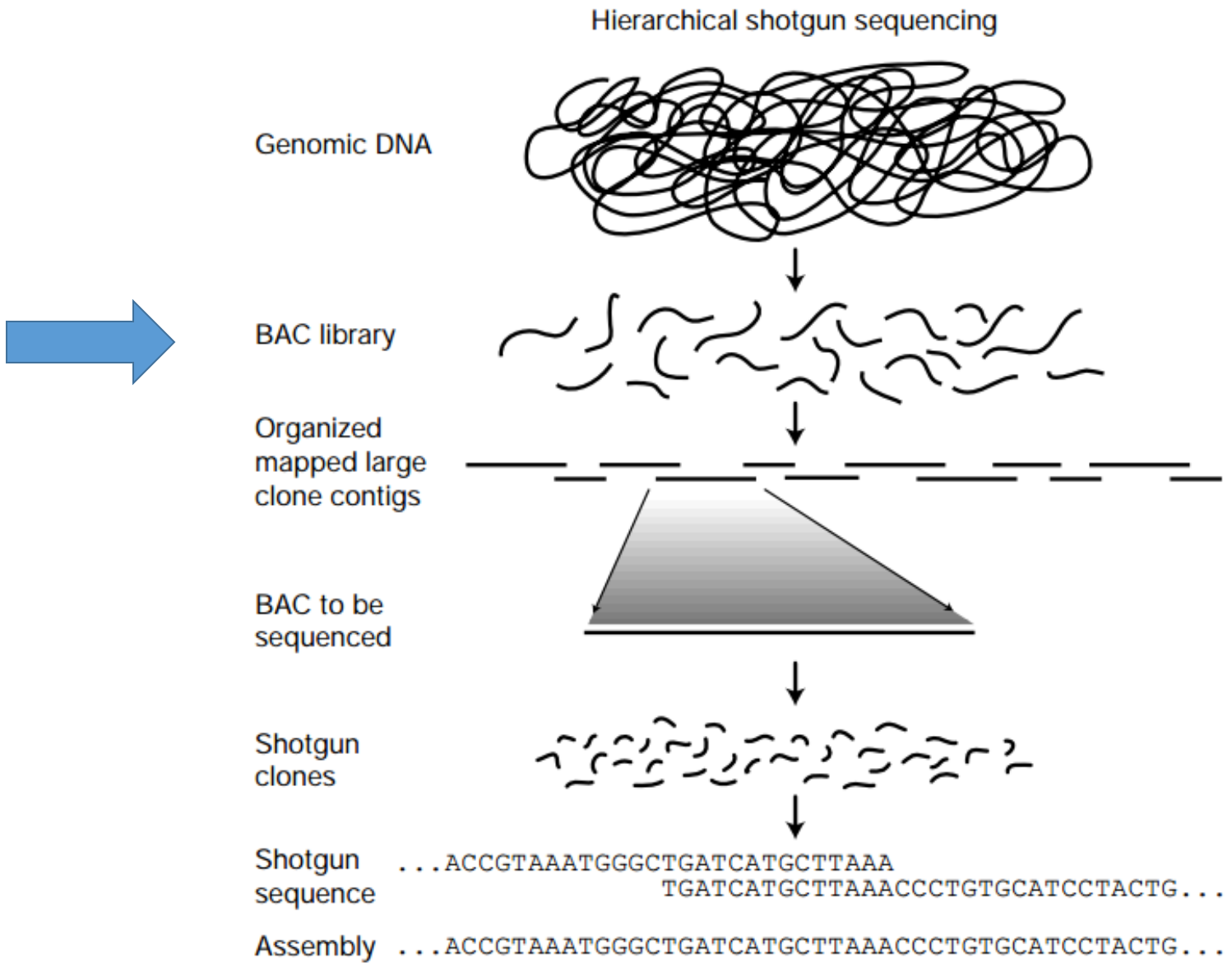
Lecture 0: Bioinformatics and the human health

Cuncong Zhong
Department of EECS
University of Kansas

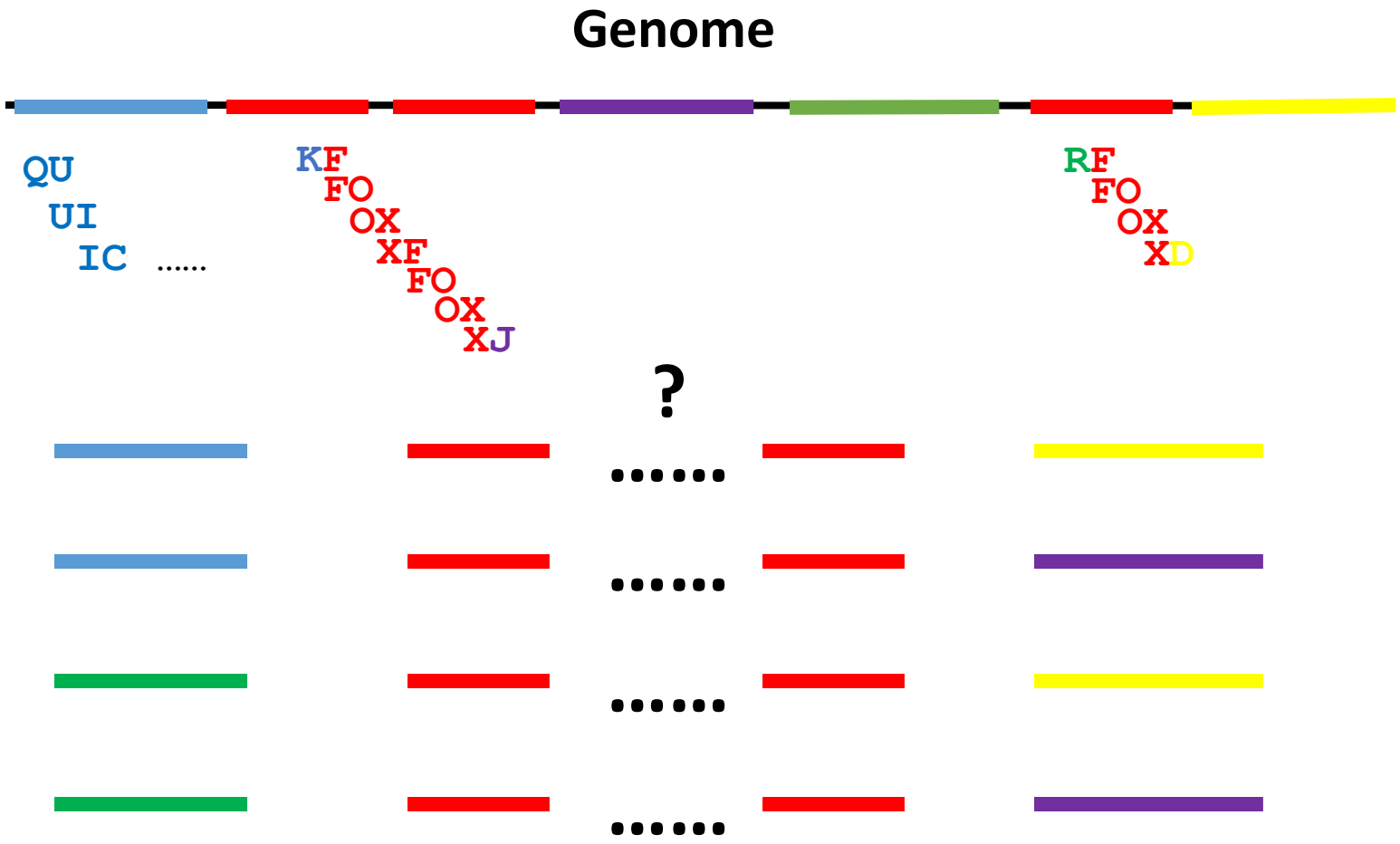
The human genome project

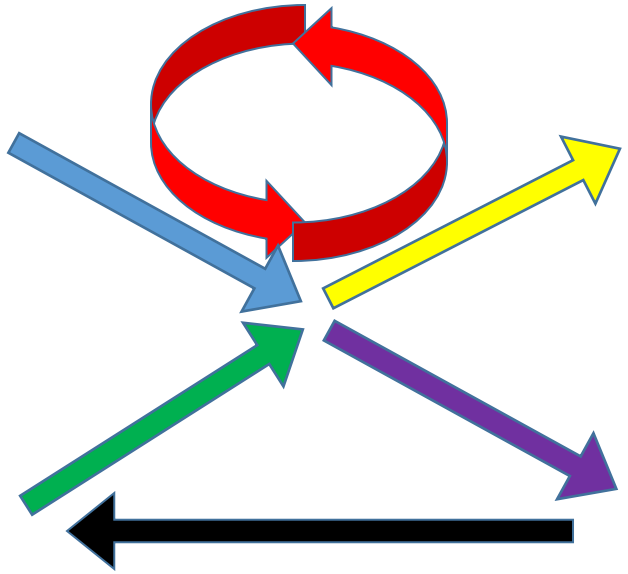
- Watch video

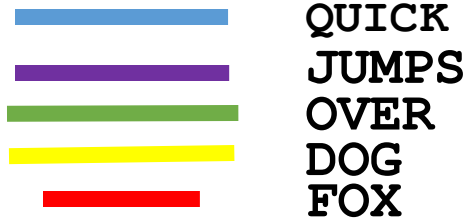
Hierarchical approach used by HGSC



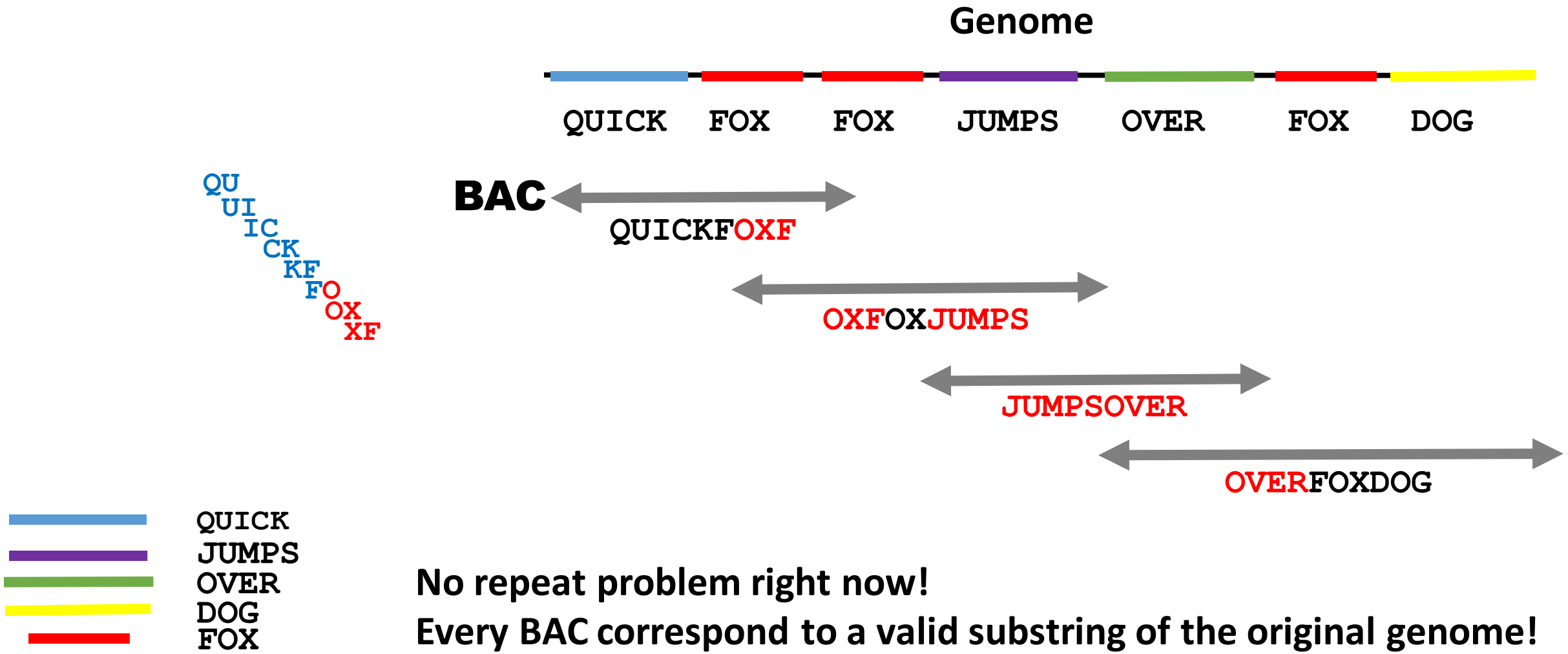
Repeats


Genome
QU
UI
IC .....
KF
FO
OX
XF
FO
OX
XJ
RF
FO
OX
XD
?
.....
.....
.....
.....




QUICK
JUMPS
OVER
DOG
FOX

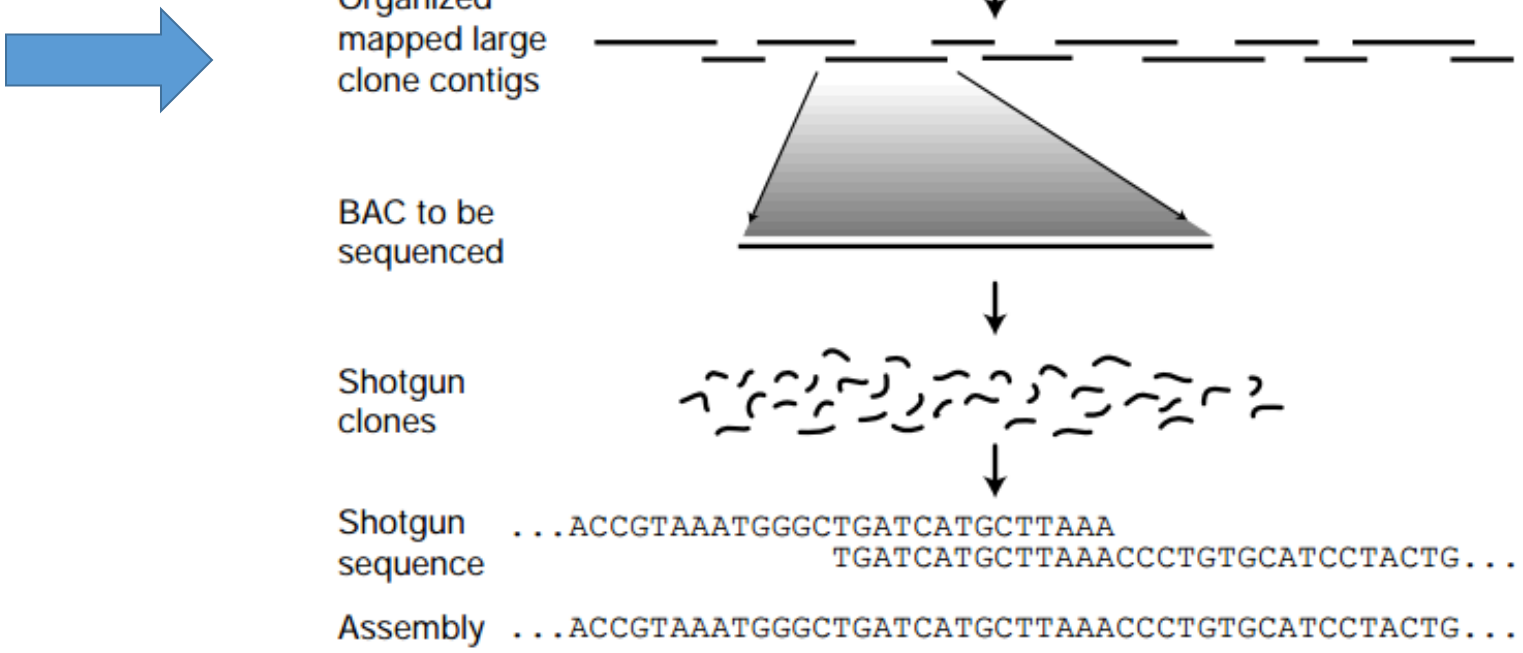
Repeats (why we use BAC)



No repeat problem right now!

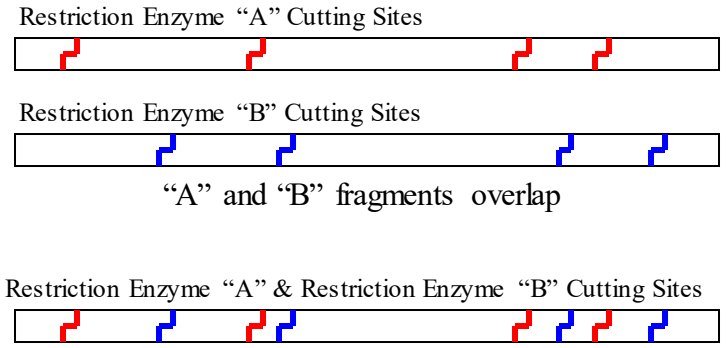
Every BAC correspond to a valid substring of the original genome!

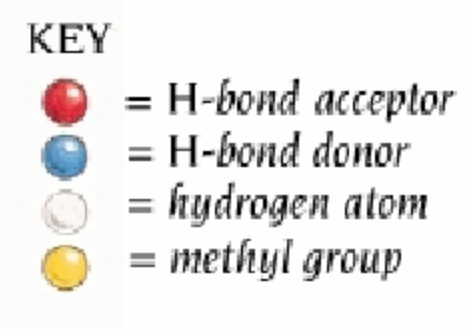
Hierarchical approach used by HGSC

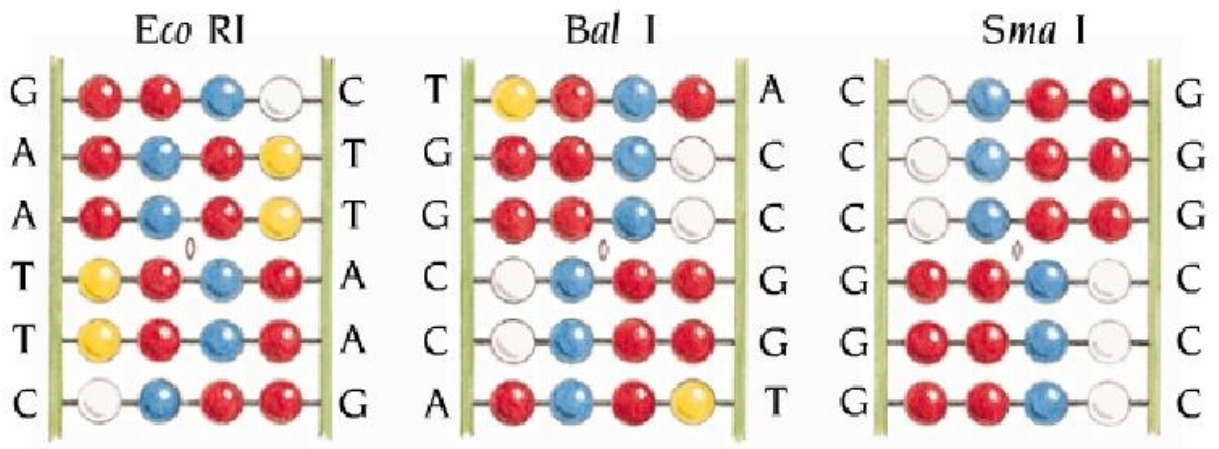


Restriction enzyme

- Restriction Enzymes cut DNA
  - Only cut at special sequences
- DNA contains thousands of these sites.
- Applying different Restriction Enzymes creates fragments of varying size.







Restriction enzymes

- Some enzymes that cut the DNA with a specific sequence pattern
- An easy way to make “barcode” for a sequence

QUICK FOX FOX JUMPS OVER FOX DOG

An enzyme that cuts “O” will give us the following fragments

QUICKF
QUICKFOXF
QUICKFOXFOXJUMPS
QUICKFOXFOXJUMPSOVERF
QUICKFOXFOXJUMPSOVERFOXD
QUICKFOXFOXJUMPSOVERFOXDOG
XF
XFOXJUMPS
XFOXJUMPSOVERF
XFOXJUMPSOVERFOXD
XFOXJUMPSOVERFOXDOG
……

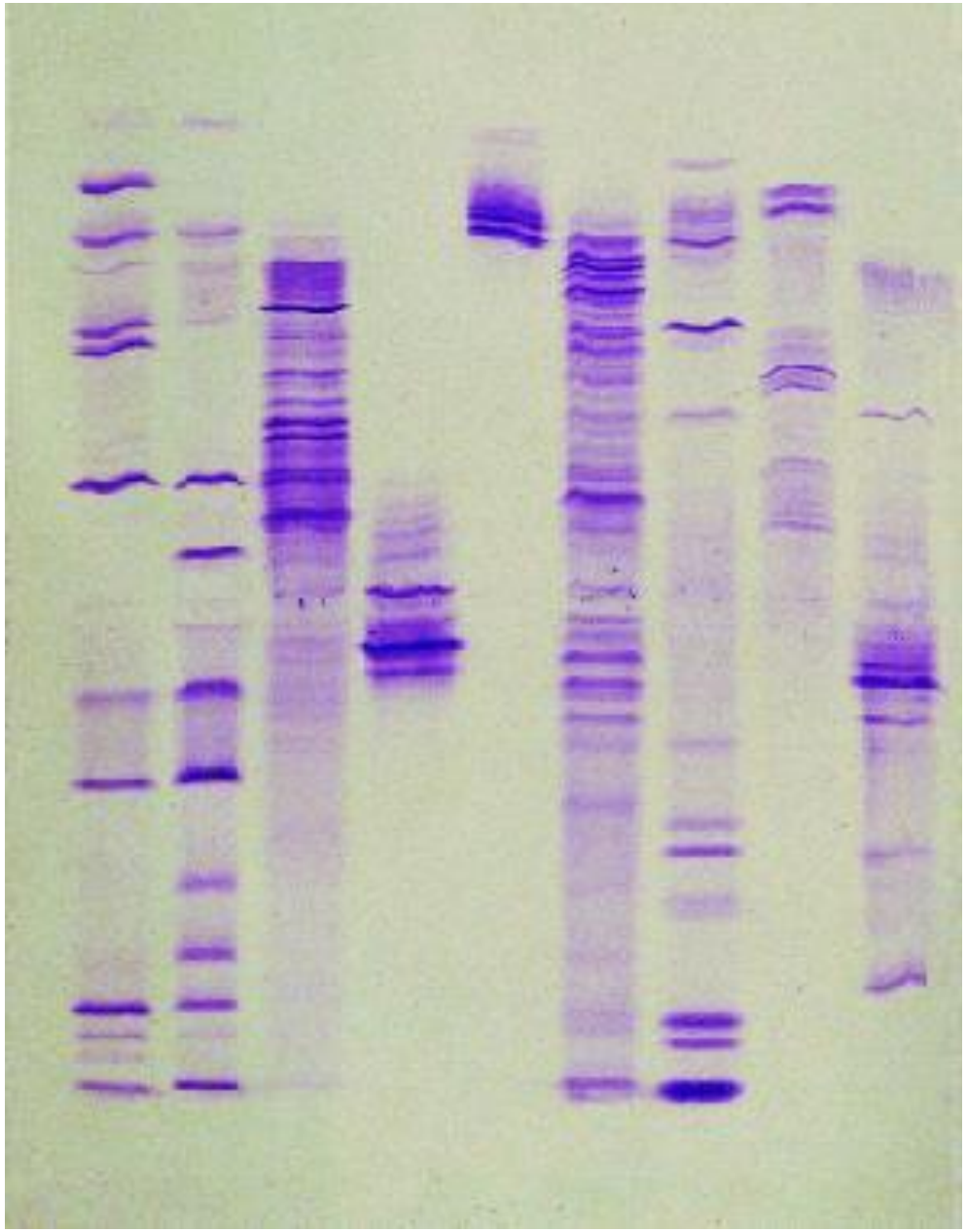
But we do not know the exact sequence!

??????
?????????
????????????????
????????????????????
????????????????????????
????????????????????????????
??
?????????
????????????????
????????????????????
????????????????????
……

Electrophoresis

- A copolymer of mannose and galactose, agaraose, when melted and recooled, forms a gel with pores sizes dependent upon the concentration of agarose

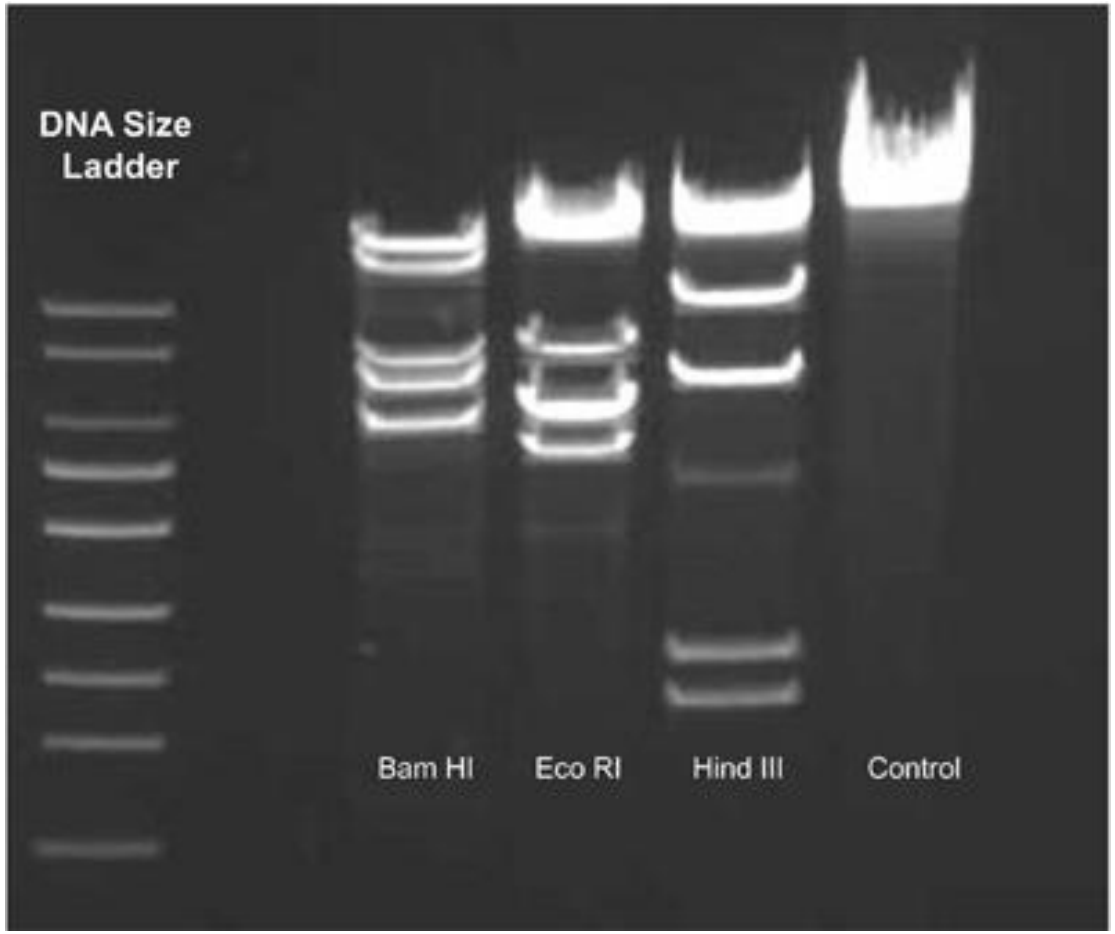
- The phosphate backbone of DNA is highly negatively charged, therefore DNA will migrate in an electric field
  - The size of DNA fragments can then be determined by comparing their migration in the gel to known size standards.



Electrophoresis cont.

But we do not know the exact sequence!

??????
?????????
????????????????
????????????????????
????????????????????????
??????????????????????????
??
?????????
????????????????
????????????????
????????????????????



http://www.apsnet.org/EDCENTER/K-12/TEACHERSGUIDE/PLANTBIOTECHNOLOGY/Pages/Activity3.aspx

Information flow

Sequence

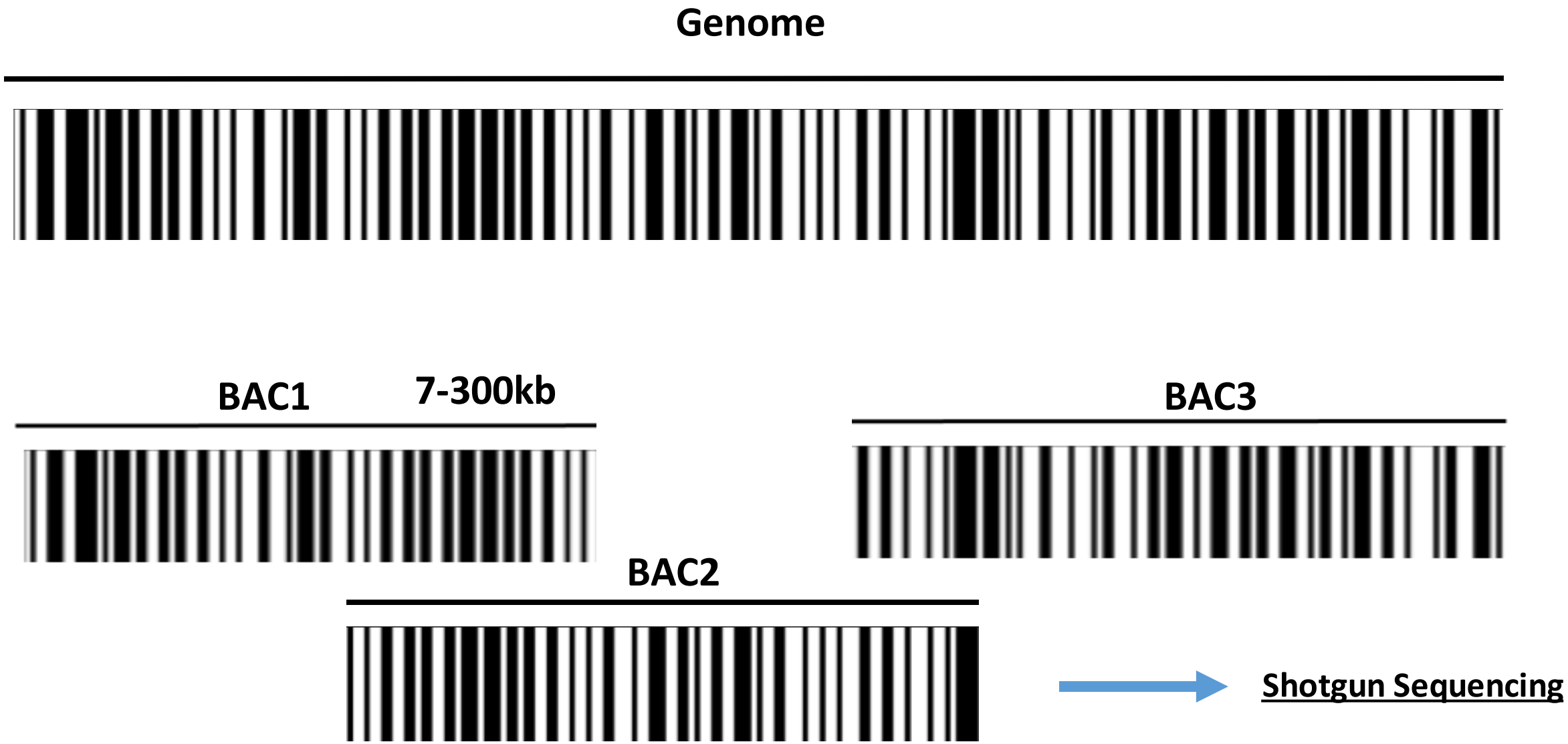
Locations of the enzyme cut sites

Lengths of the resulting fragments

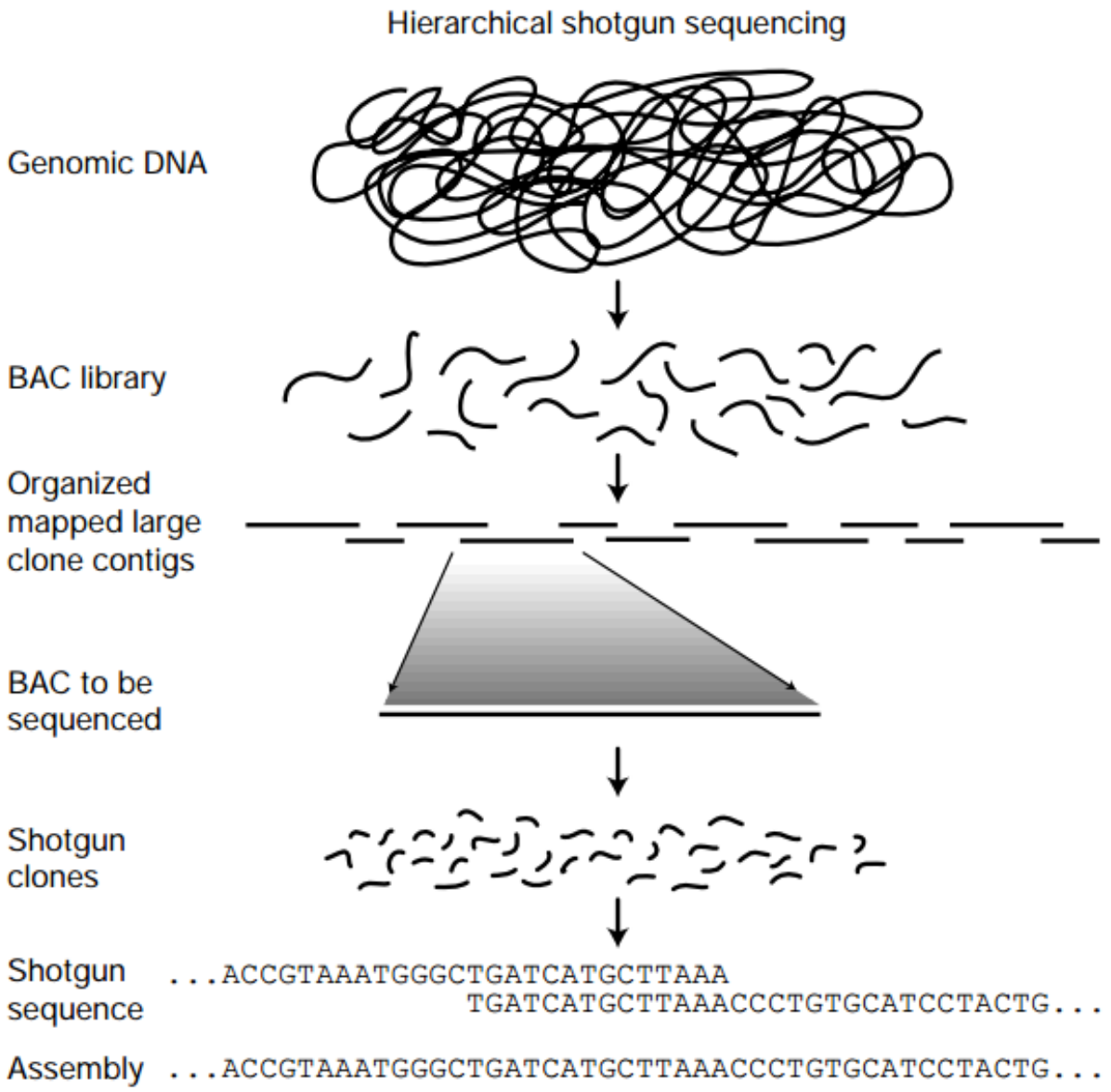
Electrophoresis pattern

For specific sequence, you get specific electrophoresis pattern

Barcode the sequence

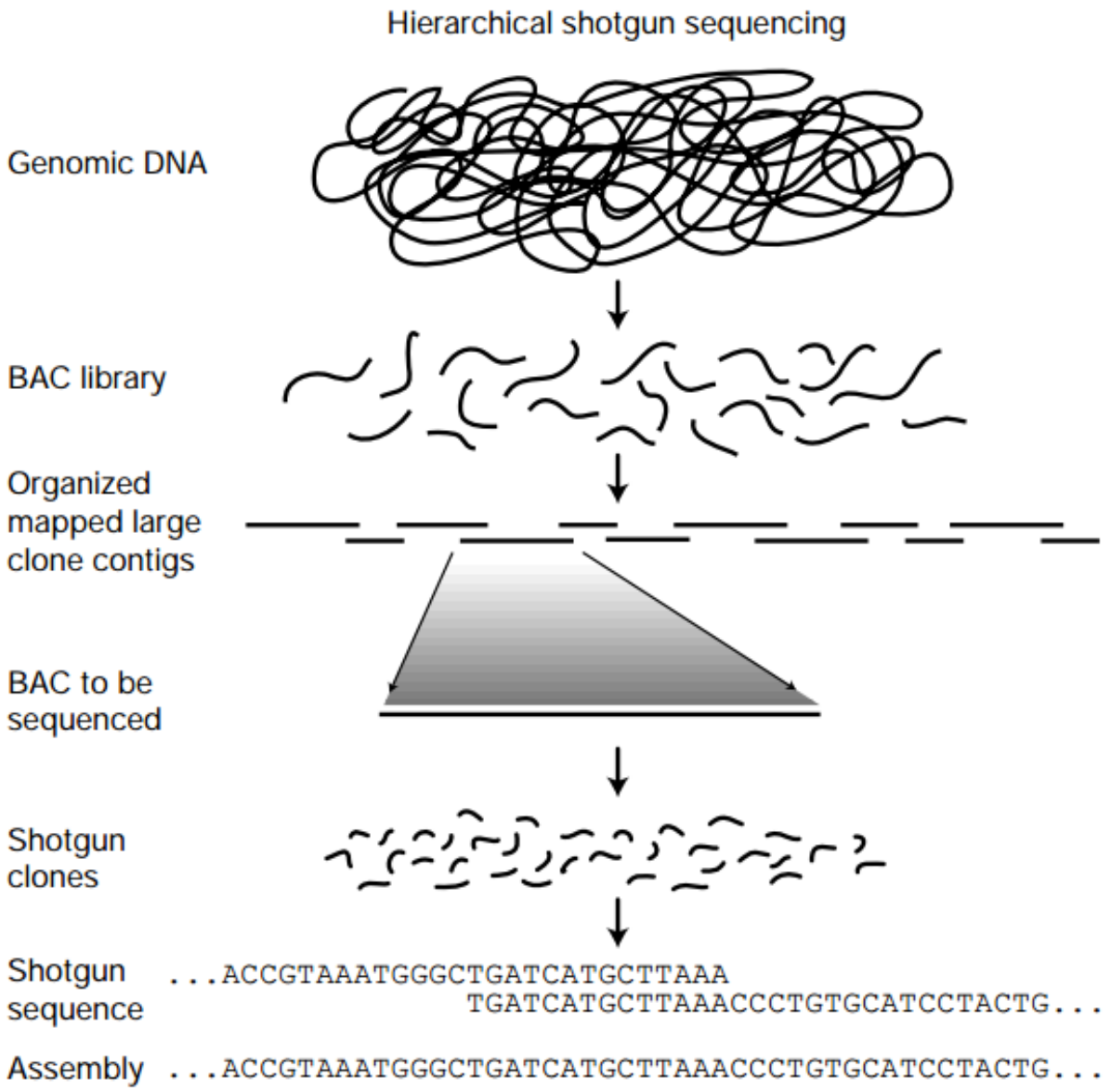


Hierarchical approach used by HGSC

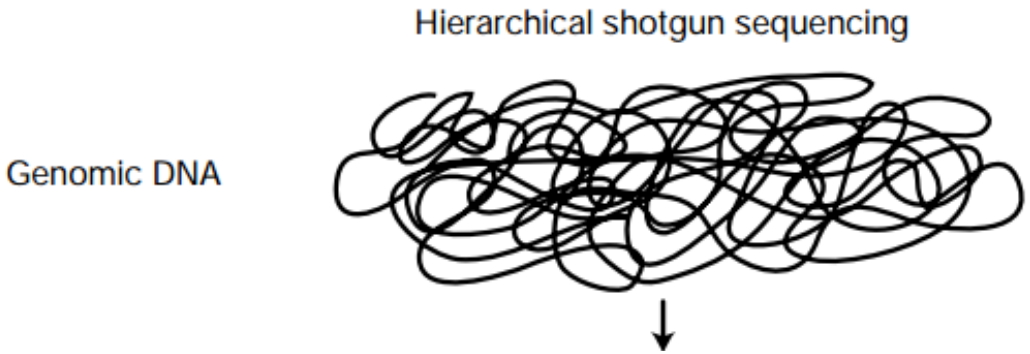


Machine in the video!!!

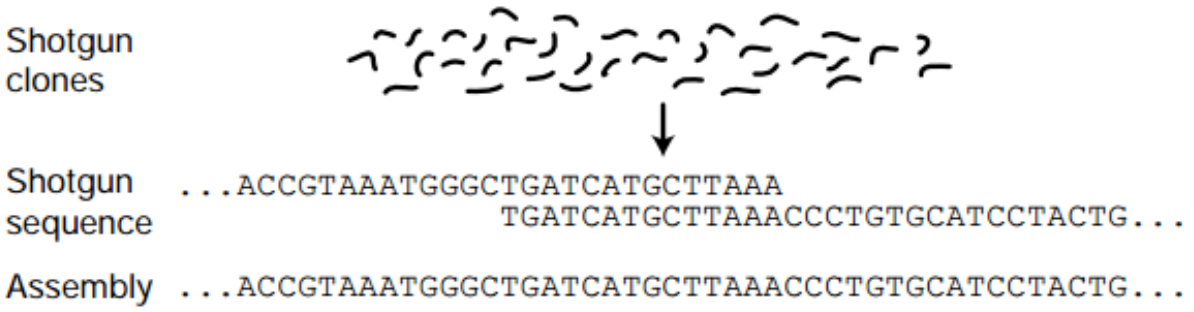
Can we simplify this pipeline?



How can we do it much faster?



We can remove these steps!!! Can we …?



Feasibility of the whole genome sequencing approach

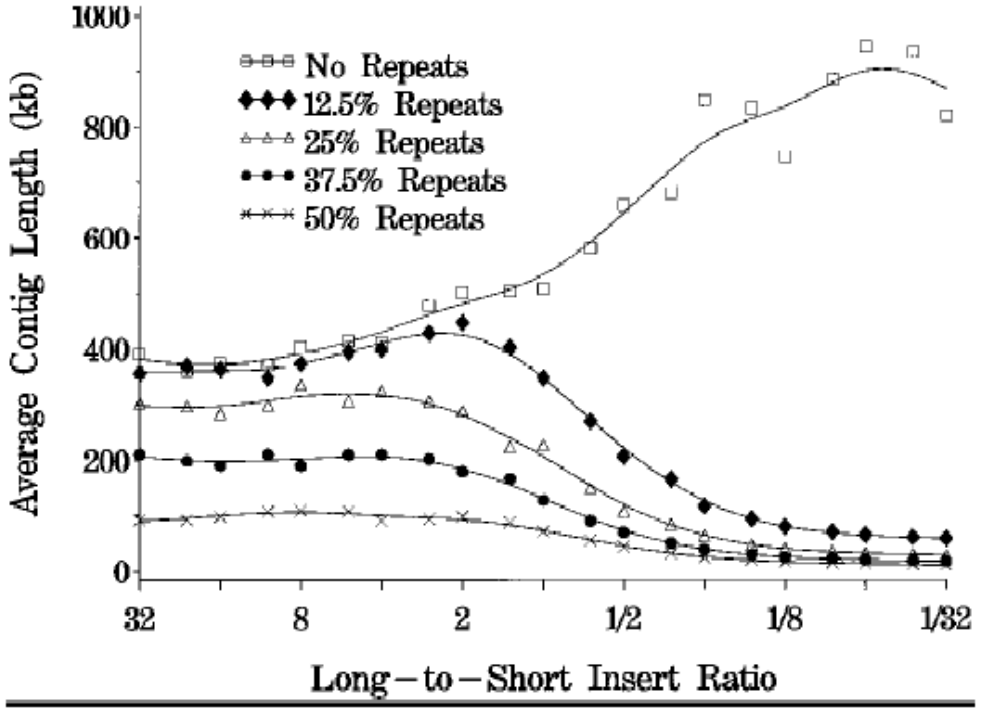
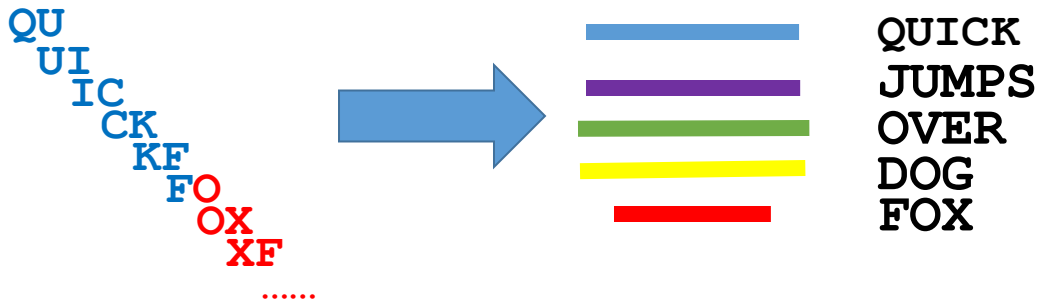


Figure 1 Average simulation contig length as a function of repeat density and long-to-short insert ratio. At each level of repetitive DNA, 80% of the repeats were assumed to be SINES and 20% LINES. All simulation parameters not specified in the plots were set to default values (see Table 2). Average contig length excluded those contigs consisting of only single reads. The single-read contigs comprised only ~0.1% of all reads.



See what we can do with these pieces…

100kb is (usually) long enough define a unique barcode!!!

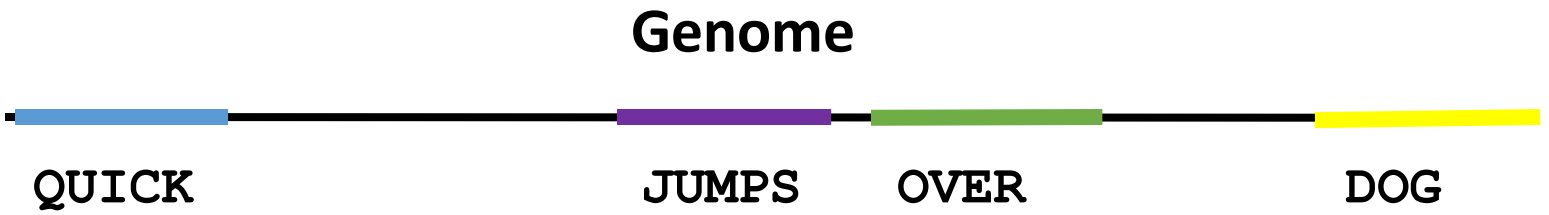
Compared to BAC library size of 7-300kb

Assembling the human genome

… Using conservative and sensitive overlap detection algorithms, it would currently be possible to span sequence-tagged sites (STSs) spaced at 100 kb at a rate of at least one STS pair per day per 100 mips (million instructions per second) workstation. With a cluster of 100 such workstations the assembly of the entire human genome would take 300 days. By using less sensitive, but faster, overlap detection software, this time could be reduced by nearly a factor of 10. Note also that the power of computer processors has doubled every 18 months for many years, and this trend is likely to continue (Patterson 1995). If contemplated machines such as the 3-teraflop supercomputer planned in 1998 for Lawrence Livermore National Laboratory (Macilwain 1996) were recruited to the task of assembly, then the human genome could be assembled, in principle, in 4 min.

-Weber and Myers, *Genome Research*, 1997

What we have right now…



- **We have the sequences of the non-repeat regions, <u>with the help of the barcode generated from the genome</u>.**
- **Fortunately, repeats <u>usually</u> do not have important biological function.**
- **We can always go back and fill the gap later!**

Summary

- The IHGC: We will spend a lot of money and effort to sequence the complete genome with high quality, because each base may have its biological meaning. Our posterity will thank us for that!

- Celera: We will sequence the majority of the meaningful pieces of the genome fast! We want to save people's life faster! We receive no support from the government and we only have a small group of people, we have to do it smartly!

- So, what do you think?

Genome and disease

- Watch video (if we have time!)

Grading

- Two take-home programming assignment
  - Local sequence alignment (aligning two sequences)
  - Alignment between a protein family profile HMM against a sequence (HMM vs sequence)
- Attendance 10%; Homework 40%; Mid-term 20%; Final 20%; Presentation 10%
  - Mid-term exam would be about alignment
  - Which topic in bioinformatics interest you the most?
  - In-class presentation in the last few weeks
  - Take-home assay as the final
- A: >=90%; B: 80-89%; C: 70-79%; D: 60-69%; F: below 60%

Announcement Sep. 13th

- Visit KU Medical Center between 11:00AM – noon
- Faculty and researchers in KUMC will introduce their work
  - Good opportunity to look for research projects!!!
- The visit is optional but encouraged
- Email me if you want to go