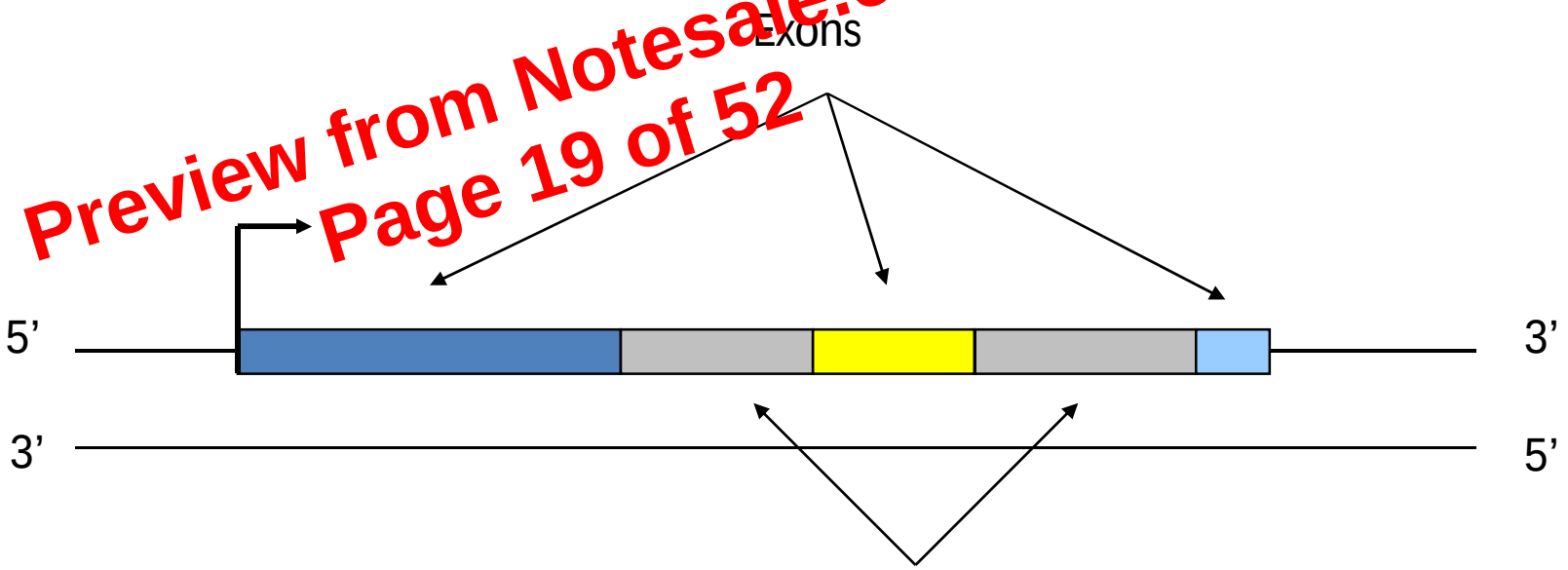
All Life depends on 3 critical molecules

- DNAs (Deoxyribonucleic acid)
  - Hold information on how cell works
- RNAs (Ribonucleic acid)
  - Act to transfer short pieces of information to different parts of cell
  - Provide templates to synthesize into protein

- Proteins
  - Form enzymes that send signals to other cells and regulate gene activity
  - Form body's major components

Exons and introns & splicing

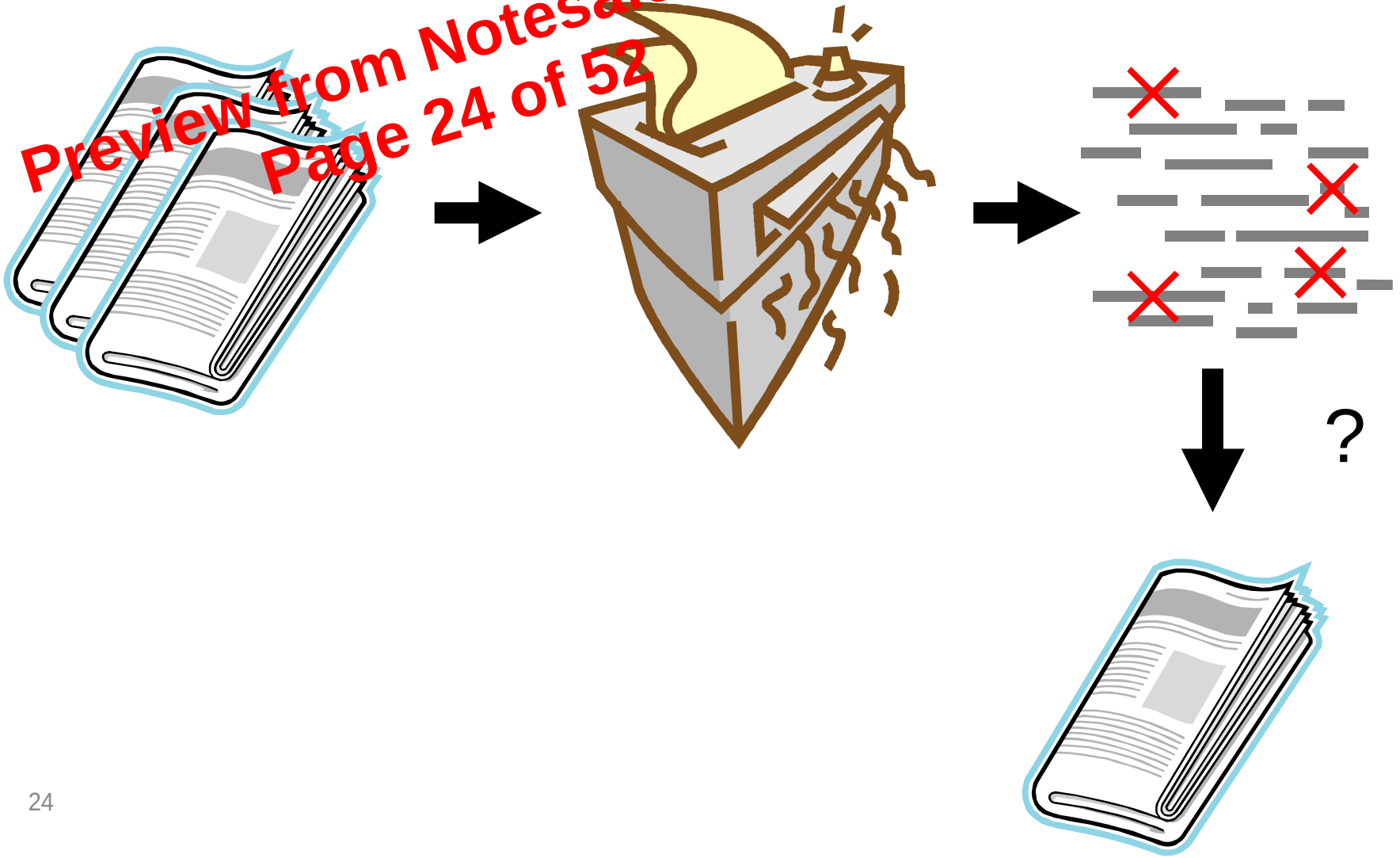


Introns are removed from RNA after transcription

Exons are joined:

This process is called *splicing*

Recovery of shredded newspaper



Problems

- Sanger sequencing error rate per base varies from 1% to 3%[1]
- Repeats in DNA
  - For example, ~300 base longs *Alu* sequence repeated is over million times in human genome
  - Repeats occur in different scales
- What happens if repeat length is longer than read length?

- Shortest superstring problem
  - Find the shortest string that ”explains” the reads
  - *Given a set of strings (reads), find a shortest string that contains all of them*

Repeats in DNA and genome assembly

Two instances of the same repeat

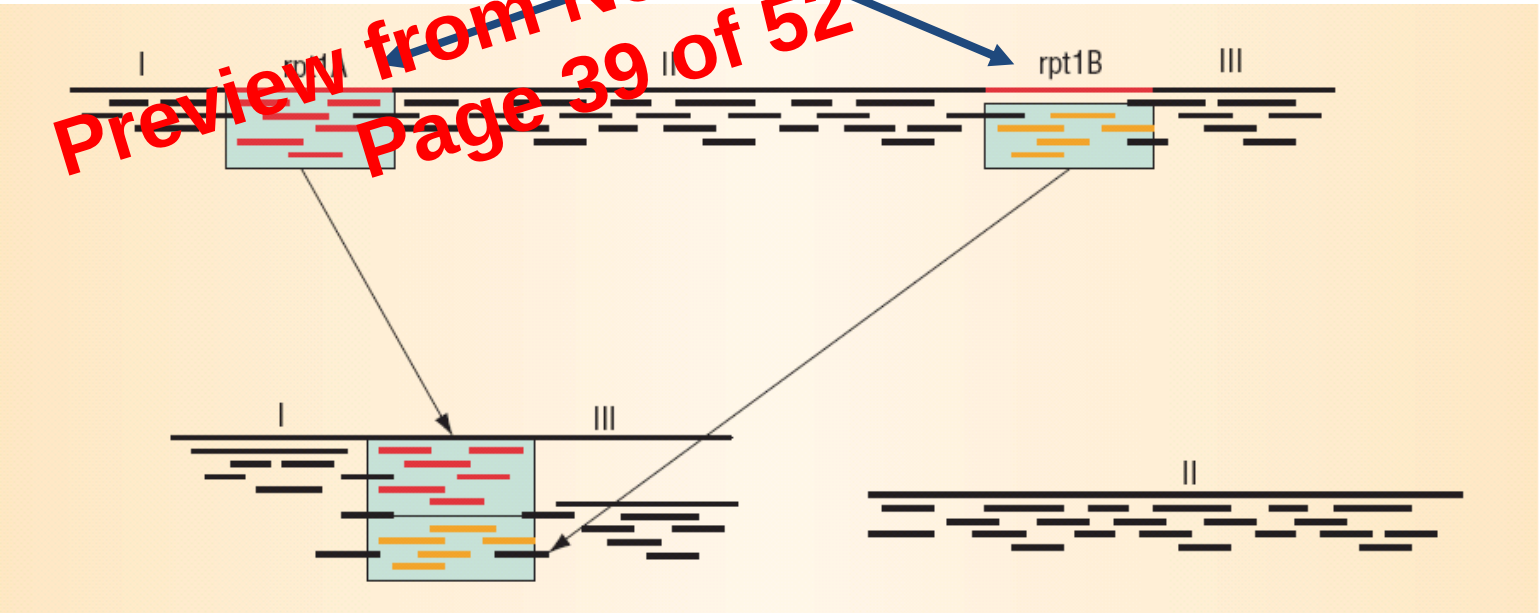


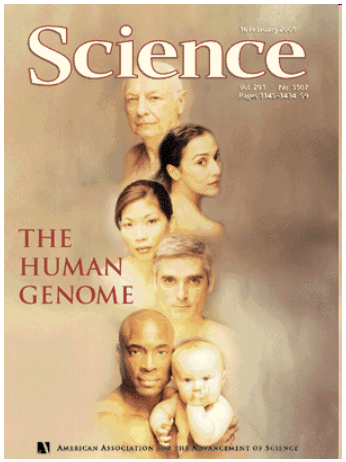
Figure 2. Repeat sequence. The top represents the correct layout of three DNA sequences. The bottom shows a repeat collapsed in a misassembly.

BAC-by-BAC sequencing

- Each BAC (Bacterial Artificial Chromosome) is about 150 kbp
- Covering the human genome requires ~30000 BACs
- BACs shotgun-sequenced separately
  - Number of repeats in each BAC is significantly smaller than in the whole genome...
  - ...needs much more manual work compared to whole-genome shotgun sequencing

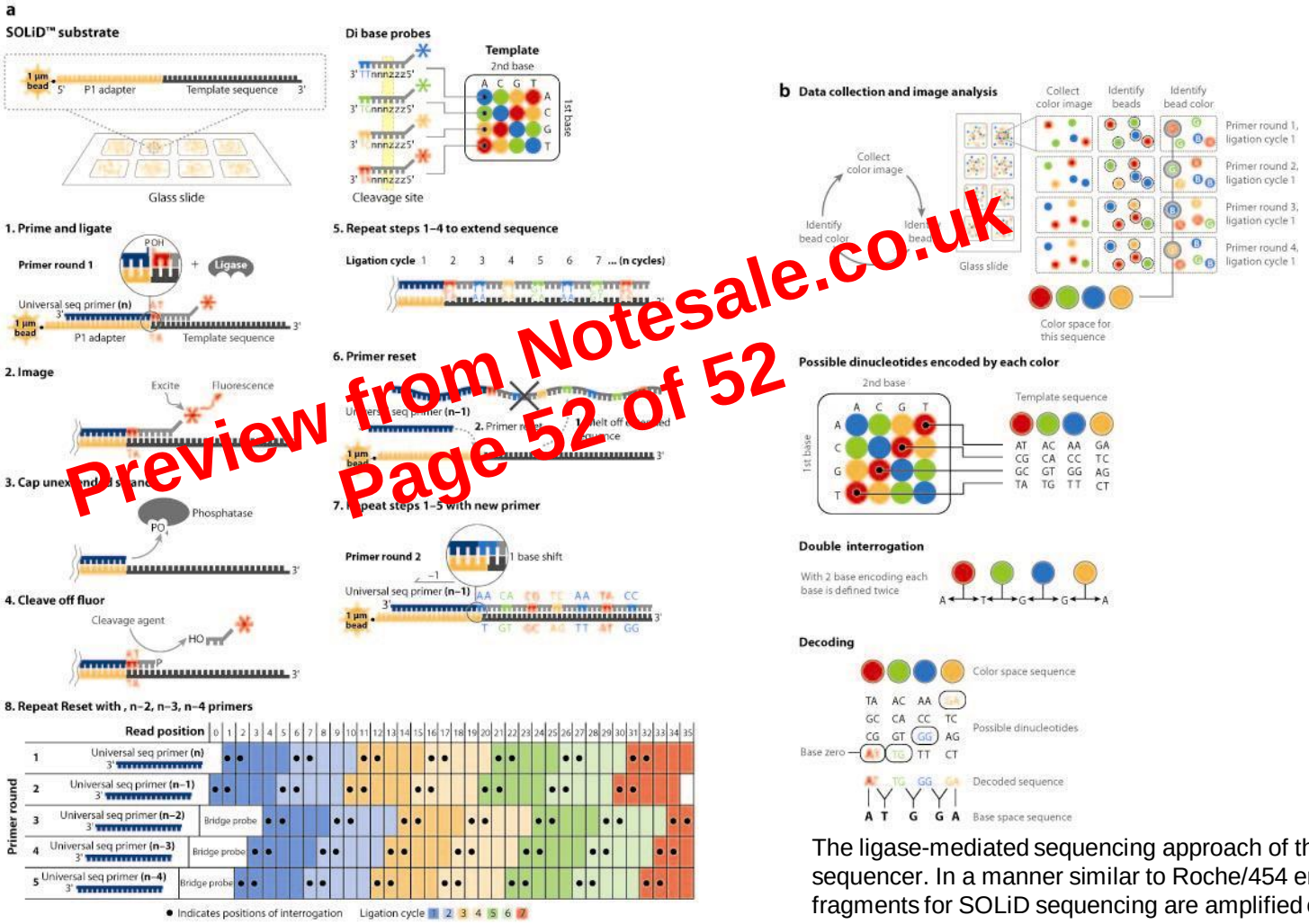
Sequencing of the Human Genome

- The (draft) human genome was published in 2001
- Two efforts:
  - Human Genome Project (public consortium)
  - Celera (private company)
- HGP: BAC-by-BAC approach
- Celera: whole-genome shotgun sequencing



HGP: Nature 15 February 2001
Vol 409 Number 6822

Celera: Science 16 February 2001
Vol 291, Issue 5507



Mardis ER. 2008.
Annu. Rev. Genomics Hum. Genet. 9:387–402

The ligase-mediated sequencing approach of the Applied Biosystems SOLiD sequencer. In a manner similar to Roche/454 emulsion PCR amplification, DNA fragments for SOLiD sequencing are amplified on the surfaces of 1-µm magnetic beads to provide sufficient signal during the sequencing reactions, and are then deposited onto a flow cell slide. Ligase-mediated sequencing begins by annealing a primer to the shared adapter sequences on each amplified fragment, and then DNA ligase is provided along with specific fluorescent-labeled 8mers, whose 4th and 5th bases are encoded by the attached fluorescent group. Each ligation step is followed by fluorescence detection, after which a regeneration step removes bases from the ligated 8mer (including the fluorescent group) and concomitantly prepares the extended primer for another round of ligation. (*b*) Principles of two-base encoding. Because each fluorescent group on a ligated 8mer identifies a two-base combination, the resulting sequence reads can be screened for base-calling errors versus true polymorphisms versus single base deletions by aligning the individual reads to a known high-quality reference sequence.