Puzzles in massively parallel genome biology await GPU solutions

Oleksiy Karpenko
Bioinformatics Program
University of Illinois at Chicago

Super Computing 2011
Seattle, WA
Human genome for computer scientists

...ACGTTGGATCGAGACATGACGATG...

• 4 letter alphabet of DNA: \{A, C, G, T\}
• ~3 GB letters long
Human genome for computer scientists

...ACGTTTGGATCGAGACATGACGATG...

- 4 letter alphabet of DNA: \{A, C, G, T\}
- \(~3\) GB letters long
- 2% are genes coding for proteins
- 98% formerly known as “junk DNA”
DNA is the same but active genes are different

heart genes ≠ bone genes

Human genome for computer scientists

...ACGTTGGATCGAGACATGACGATG...

• 4 letter alphabet of DNA: \{A, C, G, T\}
• ~3 GB letters long
• 2% are genes coding for proteins
• 98% formerly known as “junk DNA”
• the non-coding regions host important regulatory sites
1) Transcription factors
1) Transcription factors
2) Histone modifications

http://www.nyas.org/image.axd?id=a261f6e-1e42-442c-8dc6-428a60574df4&t=634329511969400000
2) Histone modifications
3) DNA methylation

DNA methylation is the addition of a methyl group (M) to the DNA base cytosine (C).
Some known regulatory rules

1. DNA methylation dominant state, gene transcription silenced.

2. H3K4me2 dominant state, gene transcription inducible.

3. H3K4me2 and H3K4me3 equal state, gene moderately transcribed.

4. H3K4me3 dominant state, gene actively transcribed.
Genome site signals detected individually

Single transcription factor binding sites

Single histone modification sites

DNA methylation sites

...ACGTTGGGATCGAGACATGACGATG...
Different signals for different tissues

\[\ldots\text{ACGTTGGATCGAGACATGACGATG}\ldots\]
Complexity?

Consider a combination of

- ~2600 transcription factors
- ~210 unique cell types
- dozens of histone modifications
Genome signals
Gene neighborhood matters
Look for common patterns
Partition the data
Process in parallel
Take home message

- Biology as a science is becoming more computationally oriented

- Genome biology is massively parallel by its nature

- GPUs fit perfectly for solving problems in genome biology