



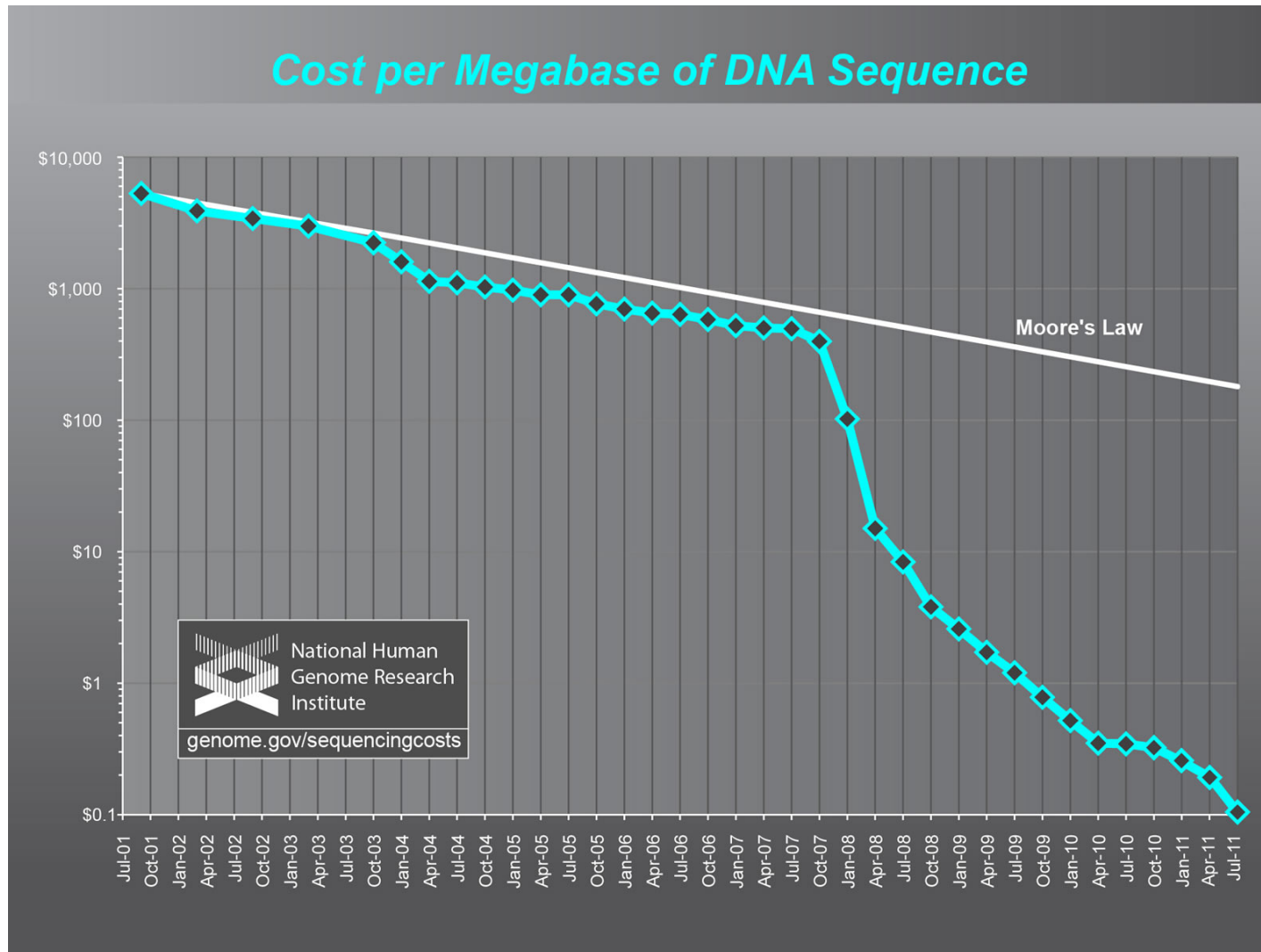
CURRENT CHALLENGES IN GENOMIC DATA VISUALIZATION

Cydney Nielsen

BC Cancer Agency
Genome Sciences Centre
Vancouver, Canada

The Data Deluge

~\$5,000
in 2001

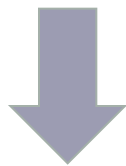


~10¢
in 2011

Sequencing Experiments

De novo assembly

AGCTTCAGATGGACAGATAA
GGCATACAGACTTAGACATA
CCAGACAAGACAGACACAGTA
TACAAGACATAAGCAATACAGA
CCAGACAAGACAGACACAGTA

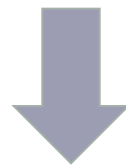


Genome Assembly

Re-sequencing

GGCATACAGACTTAGACATA
AGCTTCAGATGGACAGATAA
CCAGACAAGACAGACACAGTA
CCAGACAAGACAGACACAGTA
TACAAGACATAAGCAATACAGA

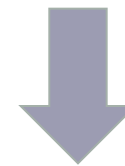
Reference Genome



Enrichment

CCAGACAAGACAGACACAGTA
AGCTTCAGATGGACAGATAA
GGCATACAGACTTAGACATA
CCAGACAAGACAGACACAGTA
TACAAGACATAAGCAATACAGA

Reference Genome



PHASE TWO: INTERPRETATION

SHENEMAN The Star Ledger



Drew Sheneman, New Jersey - The Newark Star Ledger

Challenge 1

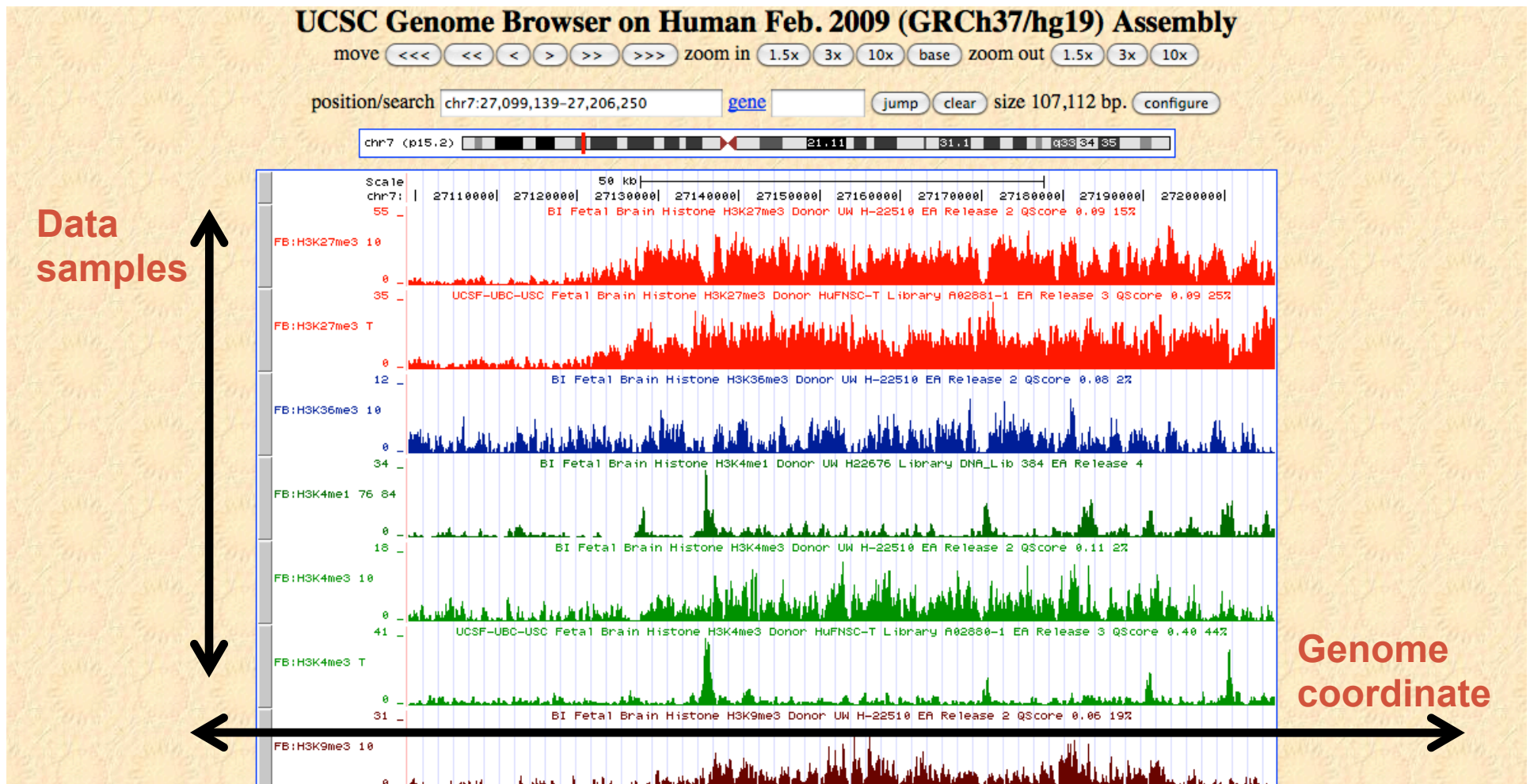
Large number of samples for comparison

“To systematically characterize the genomic changes in hundreds of tumors...and thousands of samples over the next five years”

The Cancer Genome Atlas
www.cancergenome.nih.gov

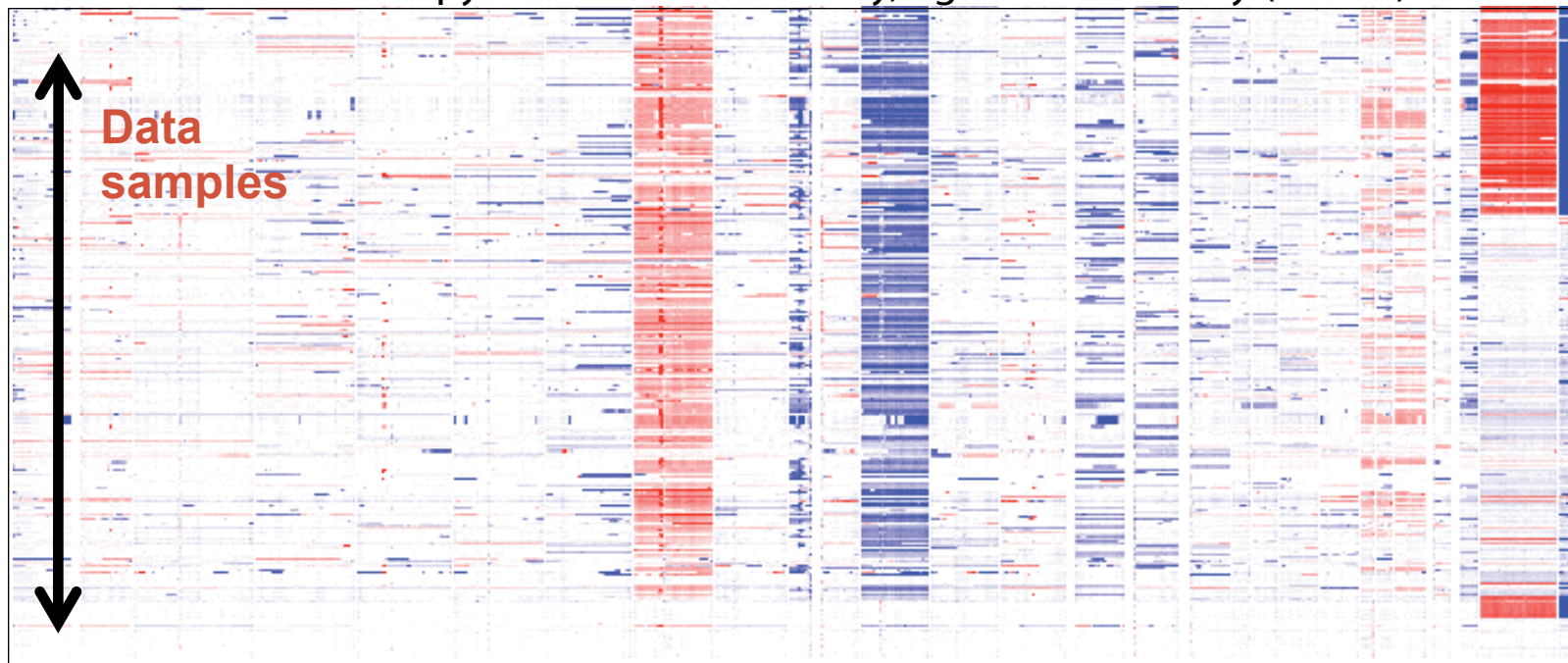
Genome Browsers

Stacked data tracks along a common genome x-axis



UCSC Cancer Genomics Heatmaps

Glioblastoma Copy Number Abnormality, Agilent 244A array (n=200)



Heatmap provides a more condensed view

Zhu *et al.*, Nature Methods, 2009

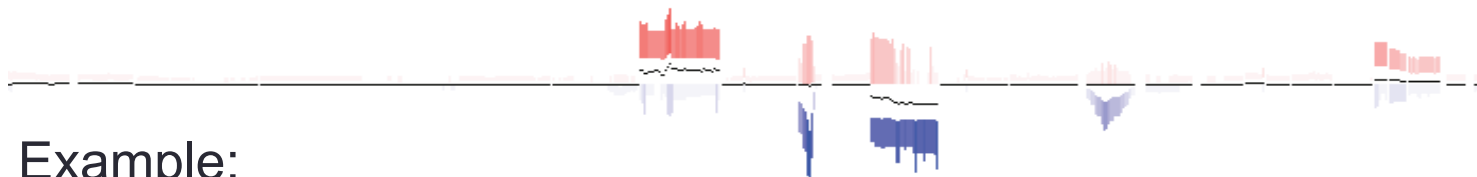
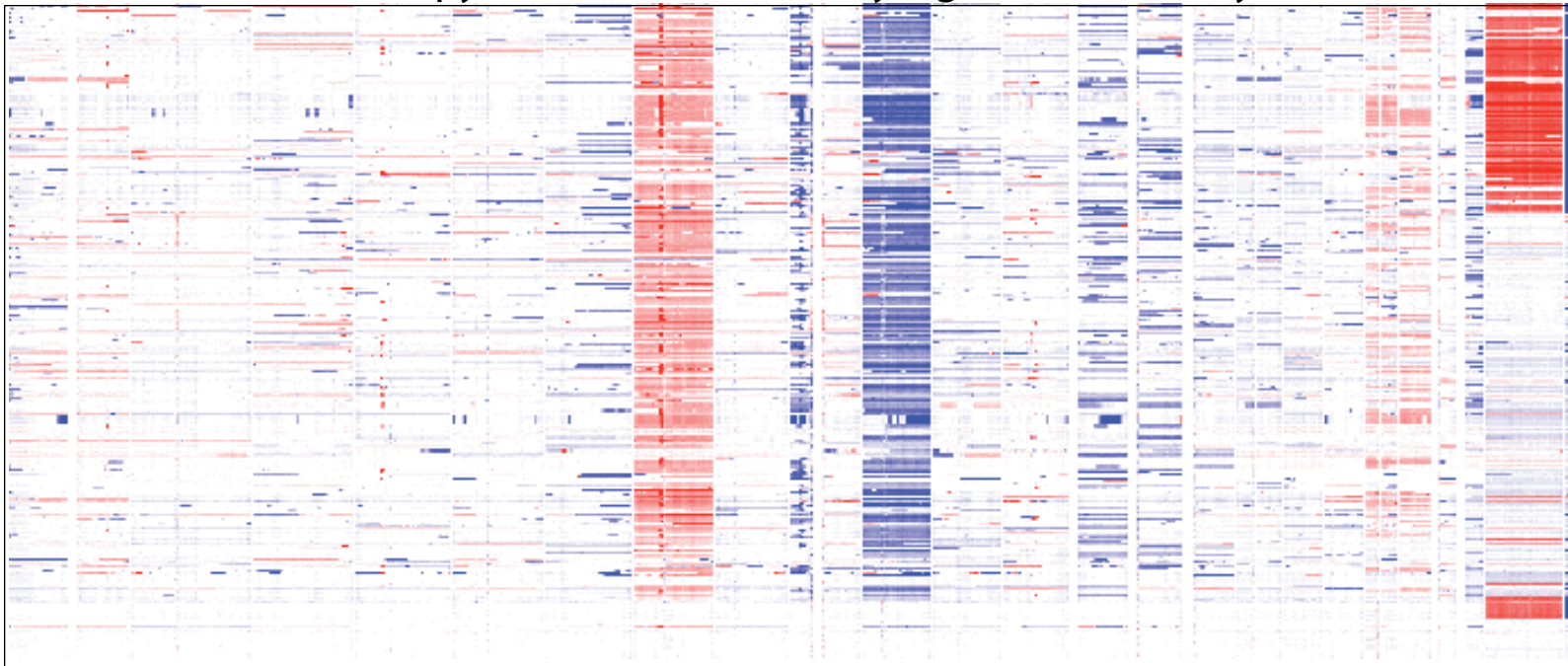
Challenge 1

Large number of samples for comparison

- *Consider what information is needed*
e.g. replace with biologically meaningful summary,
such as significant change between samples

UCSC Cancer Genomics Heatmaps

Glioblastoma Copy Number Abnormality, Agilent 244A array (n=200)



Example:
Summary view (column averages)

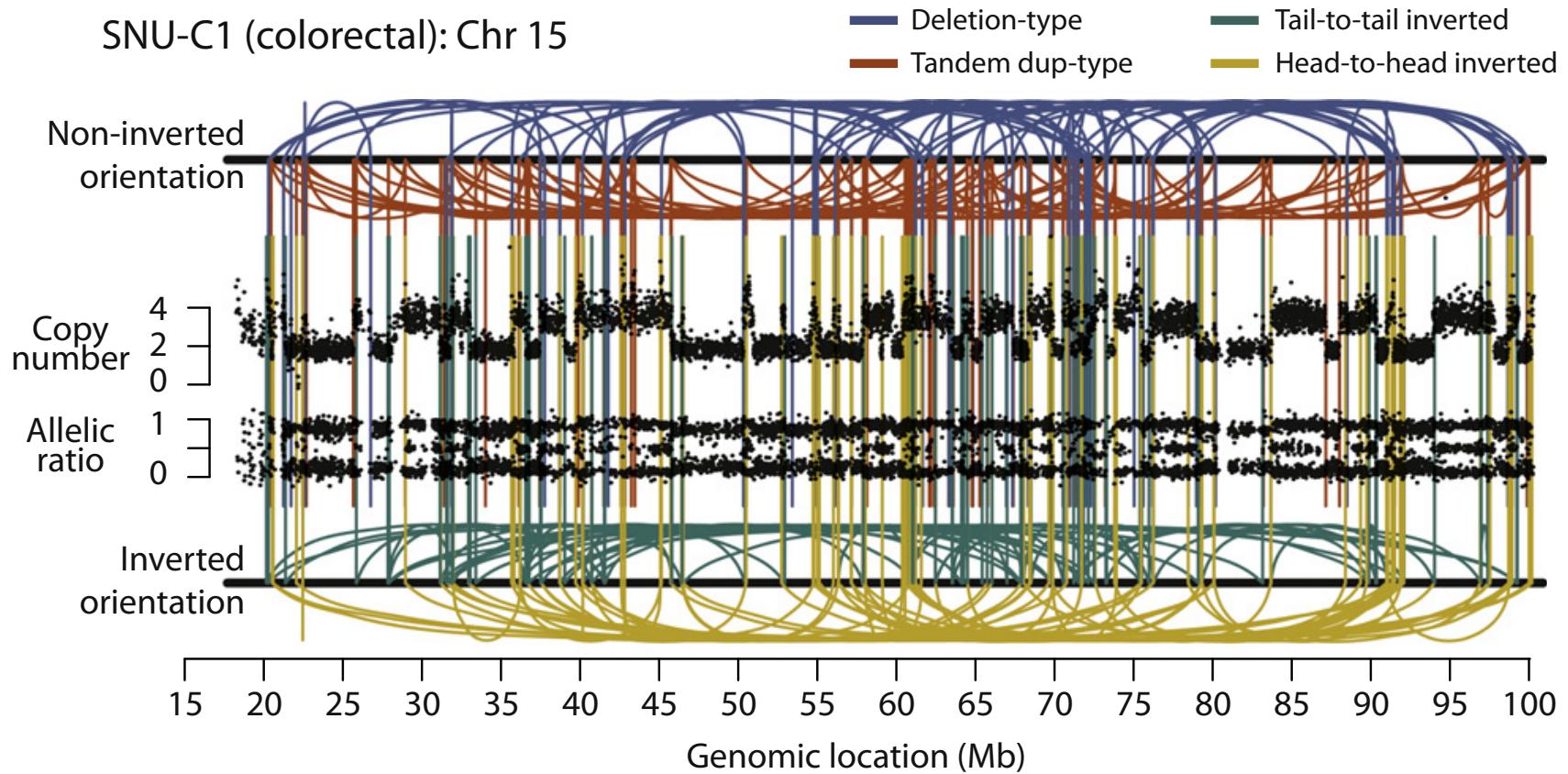


Challenge 2

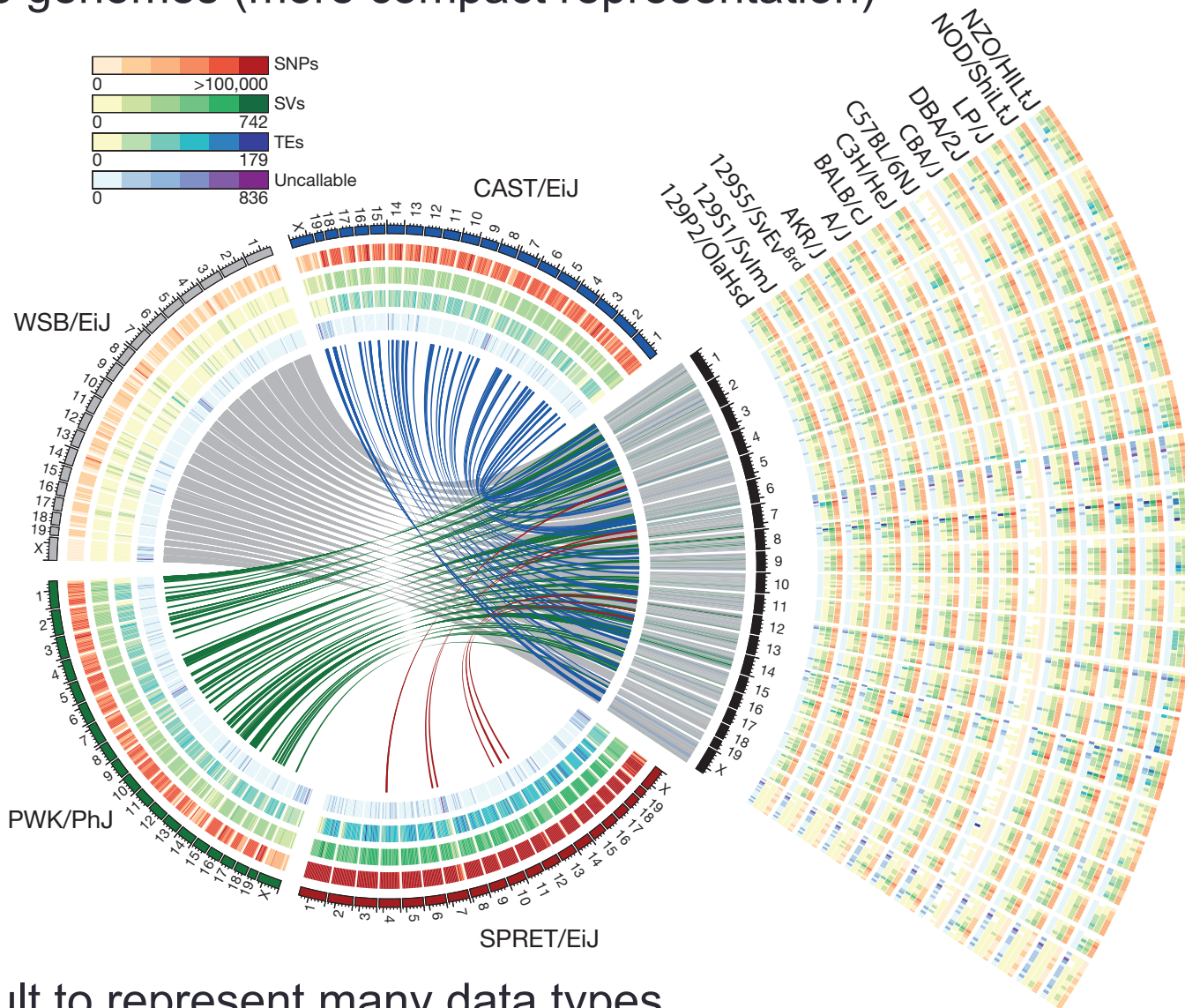
Large number of data types

Genomic rearrangements in cancer (complex representation)

SNU-C1 (colorectal): Chr 15



17 mouse genomes (more compact representation)



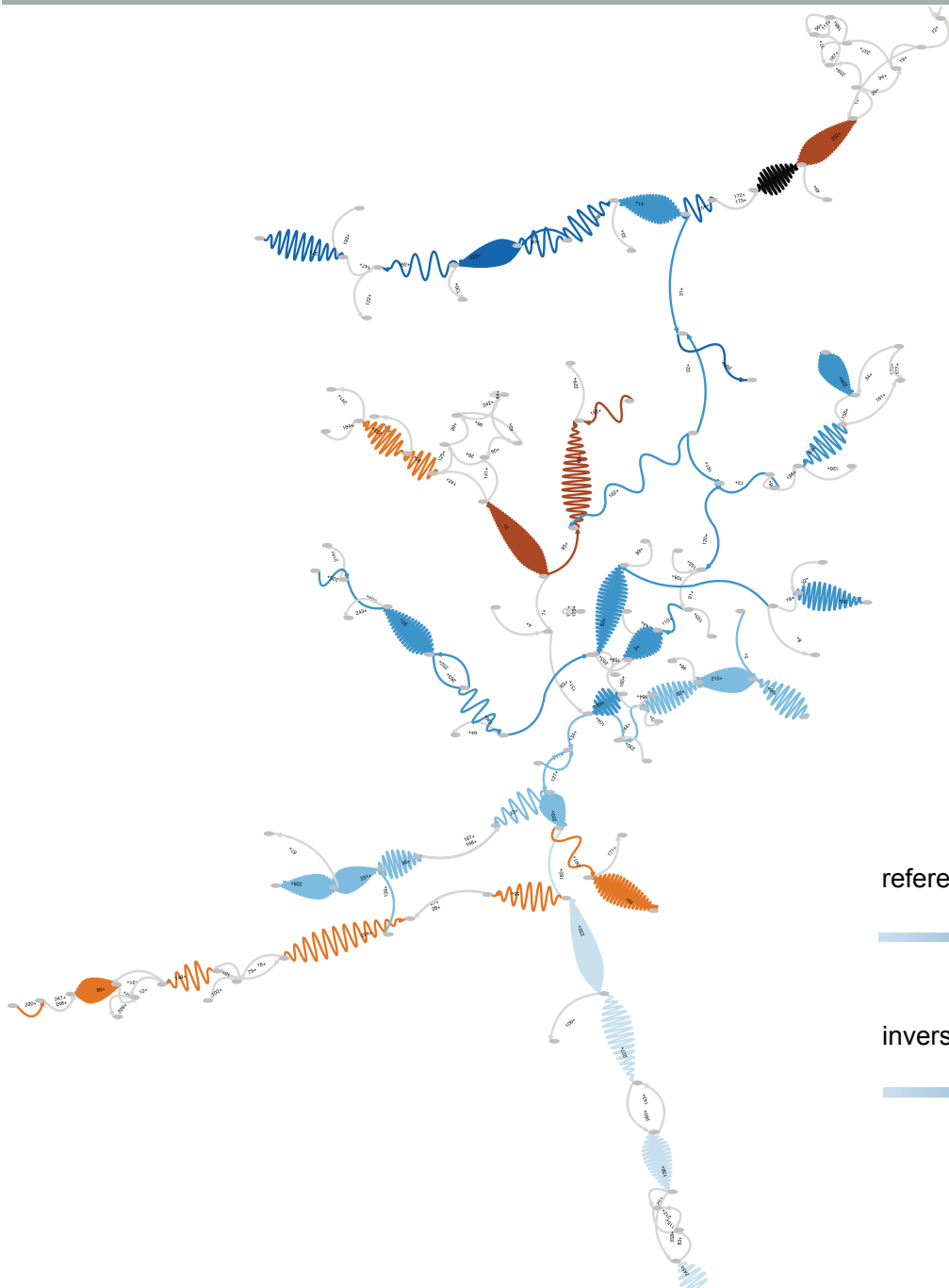
Still difficult to represent many data types in a general tool

Keane *et al.*, Nature, 2011

Challenge 2

Large number of data types

- *Compact, customized data encoding*



ABySS-Explorer

Represents sequence

- connectivity
- strand
- length
- mapping on reference

Interactively access

- sequence coverage
- scaffolding

reference human genome



inversion event in a human lymphoma genome



Nielsen *et al.*
Best Paper Award at InfoVis 2009

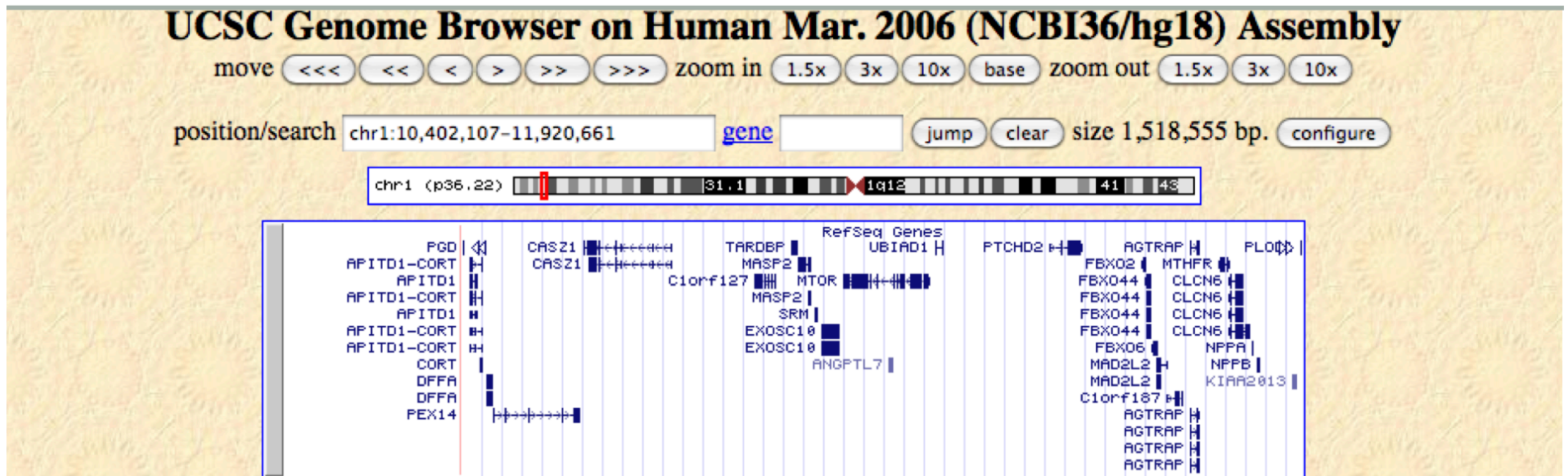


Challenge 3

Genomic features are sparse

Genome Browsers

LOCAL VIEW

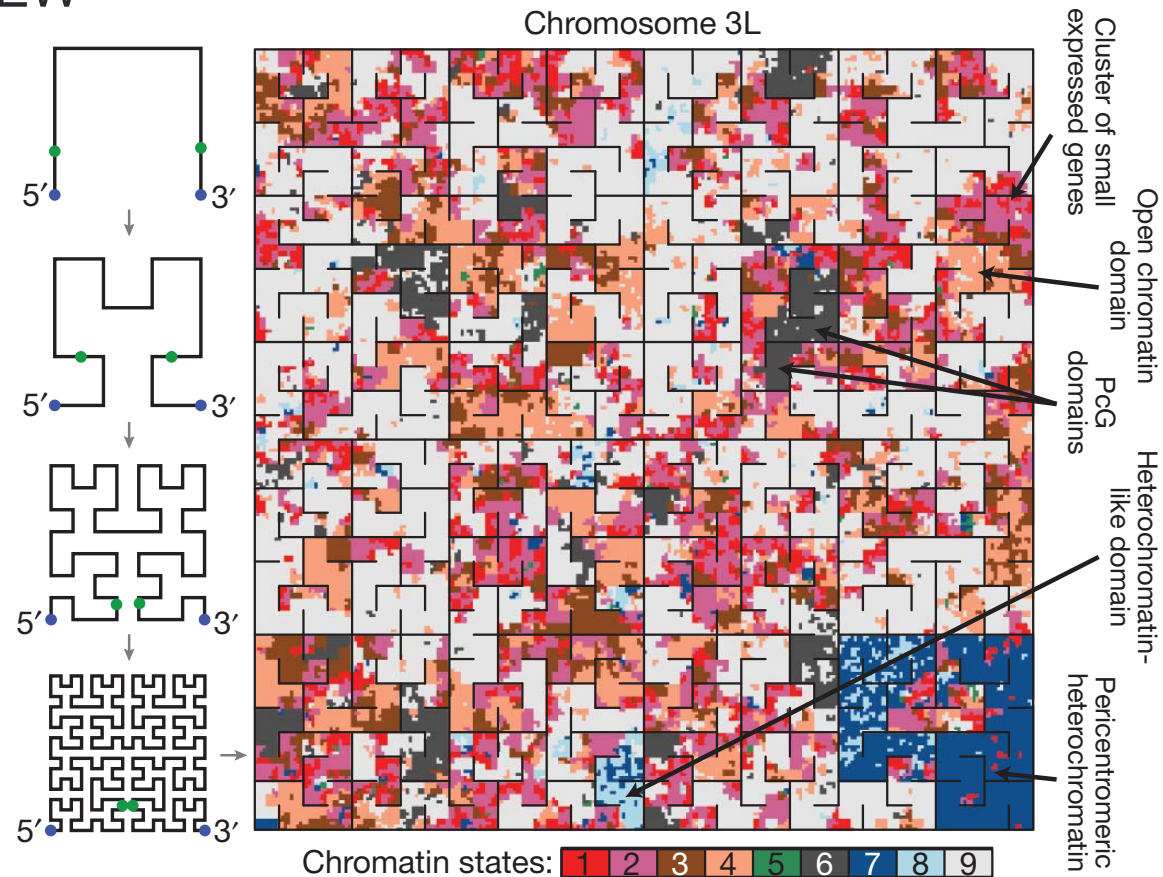


Human chr1, 1 pt corresponds to 480 kb, which is larger than 98% of all human genes!

- Martin Krzywinski

Hilbert Curve

GLOBAL VIEW



Kharchenko *et al.*, Nature, 2011
Anders, Bioinformatics, 2009

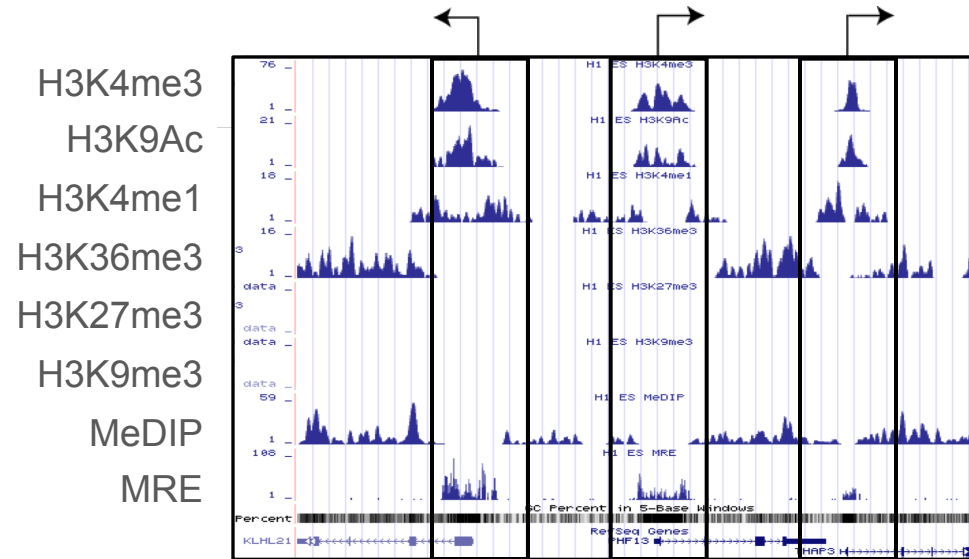
Challenge 3

Genomic features are sparse

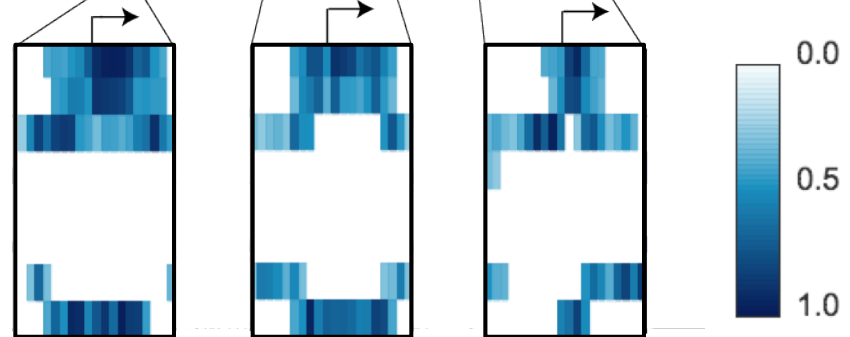
- *Need both overview and detail*
Functional axis (perhaps not full genome)

Spark – a genomic data exploration tool

1. Focus on regions of interest (e.g. transcriptional start sites)



2. Extract data matrices



3. Cluster matrices

4. Interactive cluster visualization

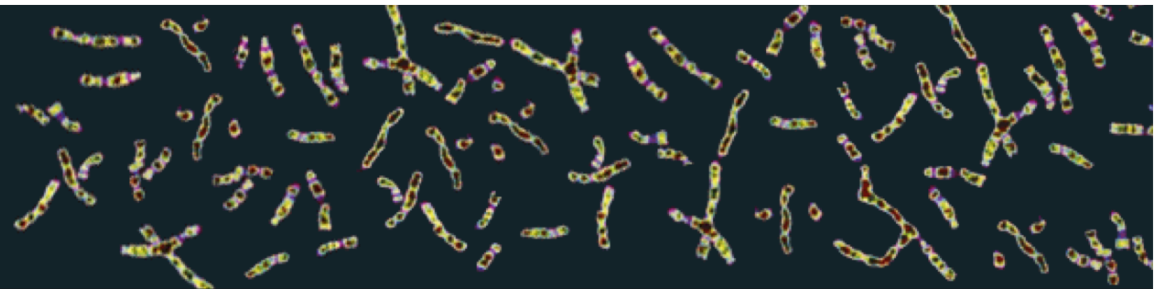
Nielsen *et al.* in preparation

Challenge 4

No longer one genome but many

1000 Genomes

A Deep Catalog of Human Genetic Variation



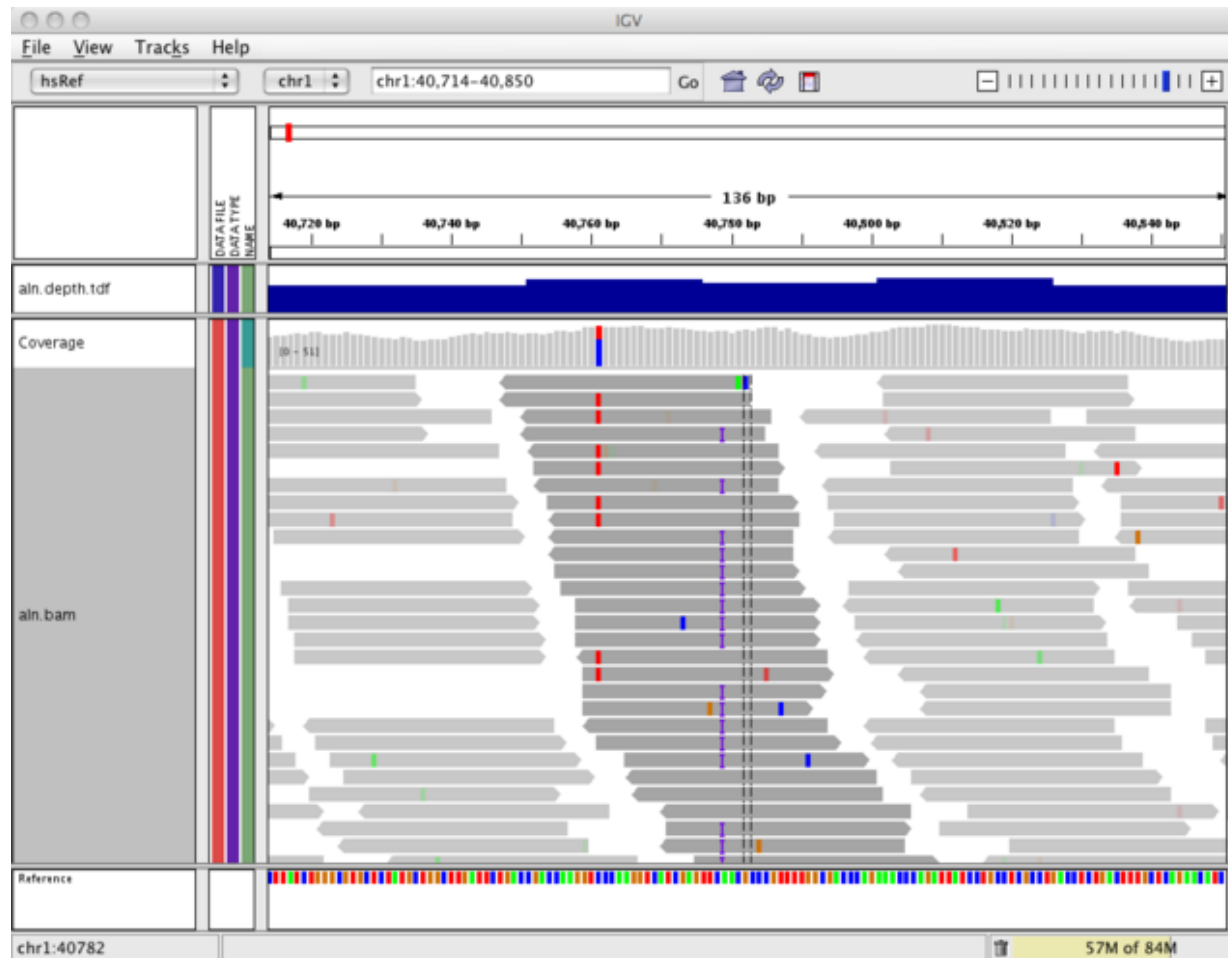
Single nucleotide variation

```
TACGTGCACCAAGACCACCAGTCTTCCCCTTTTC
TAAGCTTACGTGCACCAAGACCACCAACTTCCCAG CTCGACACAAGA
GTAAGCTTACGTGCACCAAGACCACCAGTCTTCCC CTCGACACAAGA
GTAGTAAGCTTACGTGCACCAAGACCACCAGATC CTCGACACAAGA
TTAGTAGTAAGCTTACGTGCACCAAGACCACCAGGC CTGTCTCGACACAAGA
TTAGTAGTAAGCTTACGTGCACCAAGACCACCAGTC CAGTCTTTCTCGACACAAGA
CTTAGTAGTAAGCTTACGTGCACCAAGACCACCAGT CCAGTCTTTCTCGACACAAGA
CTTAGTAGTAAGCTTACGTGCACCAAGACCACCAG CCAGTCTTTCTCGACACAAGA
CCTTAGTAGTAAGCTTACGTGCACCAAGACCACCAC CCAGTCTTTCTCGACACAAGA
CCTTAGTAGTAAGCTTACGTGCACCAAGACCACCAG TCCCAGTCTTTCTCGACACAAGA
ACCTTAGTAGTAAGCTTACGTGCACCAAGACCACCA TCCCAGTCTTTCTCGACACAAGA
ACCTTAGTAGTAAGCTTACGTGCACCAAGACCACCA TCCCAGTCTTTCTCGACACAAGA
TTACCTTAGTAGTAAGCTTACGTGCACCAAGACCAC CTTCCCAGTCTTTCTCGACACAAGA
AAAACGTTACCTTAGTAGTAAGCTTACGTGCACAAA CTTCCCAGTCTTTCTCGACACAAGA
AAAACGTTACCTTAGTAGTAAGCTTACGTGCACCAA GTCTTCCCAGTCTTTCTCGACACAAGA
CGAAAAACGTTACCTTAGTAGTAAGCTTACGTGCAC AGTCTTCCCAGTCTTTCTCGACACAAGA
ACGAAAAACGTTACCTTAGTAGTAAGCTTACGTGCC CAGTCTTCCCAGTCTTTCTCGACACAAGA
ACGAAAAACGTTACCTTAGTAGTAAGCTTACGTGTC CCACCTGTCTTCCCAGTCTTTCTCGACACAAGA
AACGAAAAACGTTACCTTAGTAGTAAGCTTACGTTC ACCACCAGTCTTCCCAGTCTTTCTCGACACAAGA
AACGAAAAACGTTACCTTAGTAGTAAGCTTACGTGC GACCACCAGTCTTCCCAGTCTTTCTCGACACAAGA
AACGAAAAACGTTACCTTAGTAGTAAGCTTACGTGC AAGACCACCAGTCTTCCCAGTCTTTCTCGACACAAG
AAAAACGAAAAACGTTACCTTAGTAGTAAGCTTACG AAGACCACCAGTCTTCCCAGTCTTTCTCGACACACG
TAAAAACGAAAAACGTTACCTTAGTAGTAAGCTTAC CCAACACCACCAGTCTTCCCAGTCTTTCTCGACACA
ATAAAAAACGAAAAACGTTACCTTAGTAGTAAGCTTA CCAACACCACCAGTCTTCCCAGTCTTTCTCGACACA
AATAAAAAACGAAAAACGTTACCTTAGTAGTAAGCT GCACCAAGACCACCAGTCTTCCCAGTCTTTCTCGA
AAATAAAAAACGAAAAACGTTACCTTAGTAGTAAGCT GTGCACCAAGACCACCAGTCTTCCCAGTCTTTCTCG
AAATAAAAAACGAAAAACGTTACCTTAGTAGTAAGCT ACGTGCACCAAGACCACCAGTCTTCCCAGTCTTTCT
AAATAAAAAACGAAAAACGTTACCTTAGTAGTAAGCT ACGTGCACCAAGACCACCAGTCTTCCCAGTCTTTTC
TTATAAAAAACGAAAAACGTTACCTTAGTAGTA TACGCGCCCAAGCCACCAGTCTTCCCAGTCTTTTC
TTCTAATAAAAAACGAAAAACGTTACCTTAGTGT TTACGTGCACCAAGACCACCAGCCCTCCCAGTCTTT
```

TTCTTTATAAAAAACGAAAAACGTTACCTTAGTAGTAAGCTTACGAGCACCAAGACCACCAGTCTTCCCAGACTTTTCGGAAAACAAGA

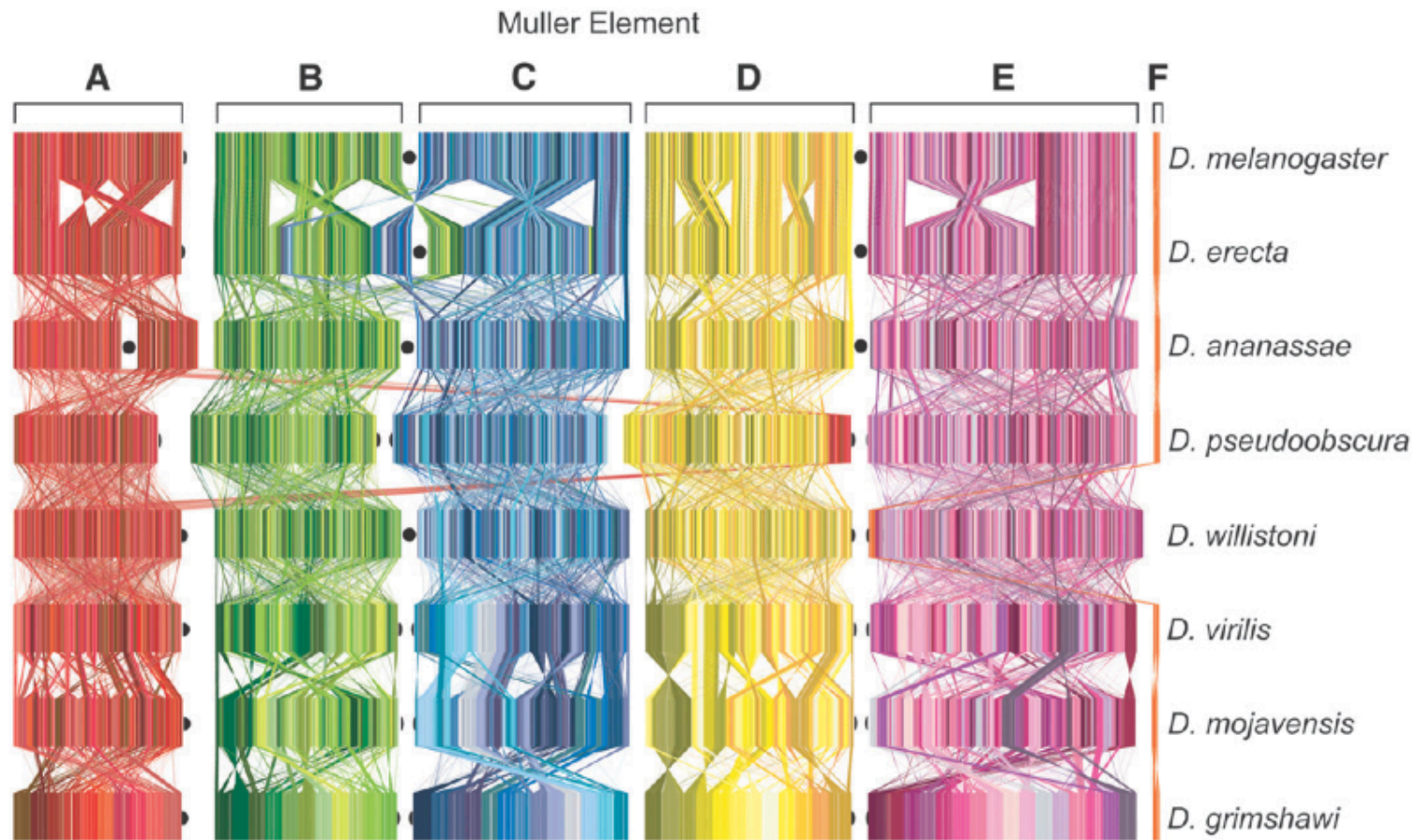
Single nucleotide variation

Integrative Genomics Viewer (IGV)



Robinson *et al.* Nature Biotechnology, 2011

Structural variation



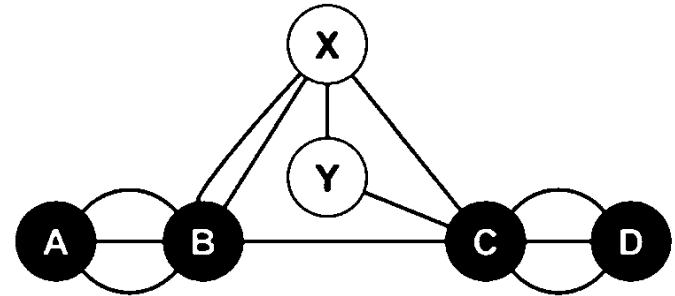
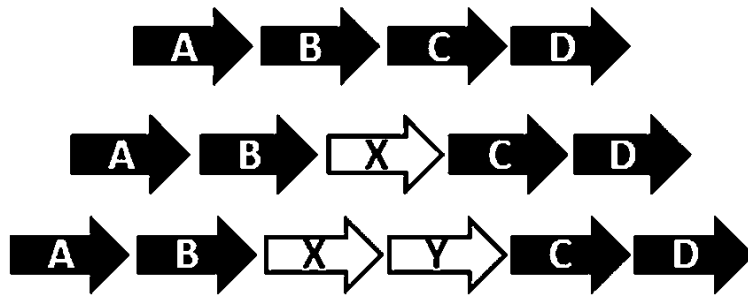
Bhutkar *et al.*, Genetics, 2008

Challenge 4

No longer one genome but many

- *Capture variation on a graph*

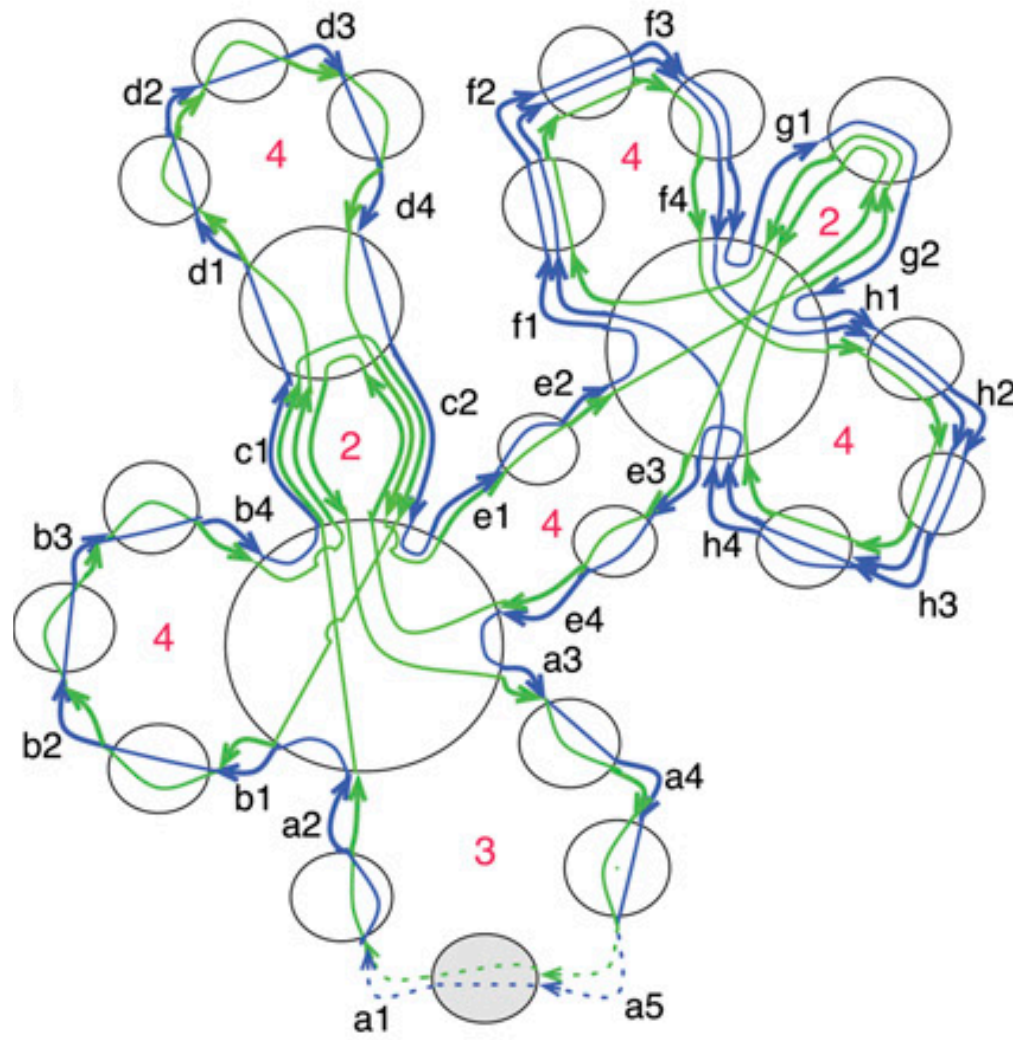
Sequence variation on a graph



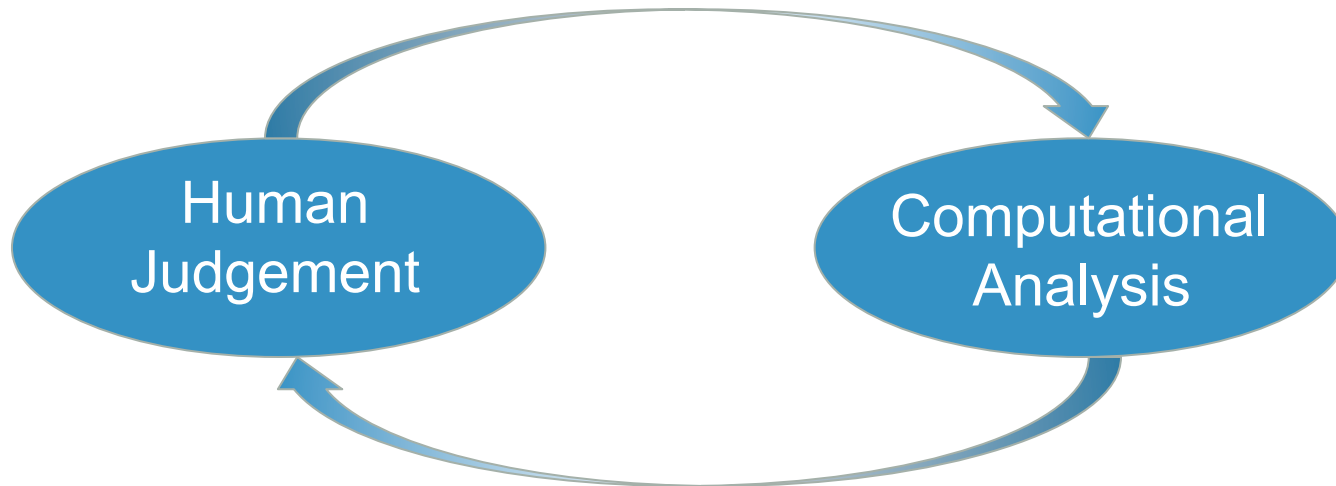
Comeau *et al.*, Mol. Biol. Evol., 2010

Users may require more time to learn how to interpret graph representations, but such graphs are likely to scale better and may prove more powerful for analysis

Sequence variation on a graph



Challenge 5



Consed Genome Assembly and Finishing Tool

File Navigate Info Color Dim Misc Sort Help

standard.fasta.screen.ace.1 Contig1 Sone Tags Pos: I clear

Search for String Compl Cont Compare Cont Find Main Min Err/10kb: 313.29

770 780 790 800 810 820 830

CONSENSUS GGGCTACAAGAAATTTT*TACTTTTAAAAAATCAGACAATAGGGATTCTAAGAGAGGCTTCATGACGGCTAAC

djs74-2361.s1 GGGCTACAAGAAATTTT*TACTTTTAAA*AAATCAGACAATAGGGATTCTAAGAGAGGCTTCATGACGGCTAAC

djs74-996.s2 GGGCTACAAGAAATTTT*TACTTTTAAAAAATCAGACAATAGGGATTCTAAGAGAGGCTTCATGACGGCTAAC

djs74-2689.s1 GGGCTACAAGAAATTTT*TACTTTTAAAAAATCAGACAATAGGGATTCTAAGAGAGGCTTCATGACGGCTAAC

djs74-2350.s1 GGGCTACAAGAAATTTT*TACTTTTAAAAAATCAGACAATAGGGATTCTAAGAGAGGCTTCATGACGGCTAAC

djs74-1180.s1 GGGCTACAAGAAATTTT*TACTTTTAAAAAATCAGACAATAGGGATTCTAAGAGAGGCTTCATGACGGCTAAC

djs74-564.s1 GGGCTACAAGAAATTTT*TACTTTTAAAAAATCAGACAATAGGGATTXXXXXXXXXXXXXXXXXXXXXXXXXXXX

djs74-423.s1 GGGCTACAAGAAATTTT*tACTTTTAAAAAATCAGACAATAGGGATTCTAAGAGAGGCTTCATGACGGCTAAC

djs74-1532.s1 gGGCTACAAGAAatTTT*tACTttttaaAAATCAGACAatagggattCTAAGAGaggcttCatgacggctaAC

djs74-1802.s1 GGGCTACAAGAAATTTT*TACTTTTAAAAAATCAGACAATAGGGATTCTAAGAGAGGCTTCATGACGGCTAAC

djs74-237.s1 gggctacaagaaatTTTatacttttaaaaaatcggacaatagggatctctaagagaggcttcatgacggctaac

djs74-1432.s1 agagtgtttttcccccTt*tttcgaaaaaancagaggtgaccctttatgggaatattXXXXXXXXXXXXXXXXXXXX

<<< << < Prev Next > >> >>> cursor reads sorted by strand and then position dismiss

David Gordon and Phil Green

Good example of integrated visualization and computational analysis functionality

Challenge 5

Need to integrate computation

- *High interactivity, low memory overhead*
Avoid storing large data sets locally
Popularity of web-based tools
Evolving sequencing technologies

Summary

- 1 Large number of samples for comparison
- 2 Large number of data types
- 3 Genomic features are sparse
- 4 No longer one genome but many
- 5 Need to integrate computational analysis

