Optical mapping reveals genomic structure

Steve Goldstein July 22, 2010

Laboratory for Molecular and Computational Genomics – UW Madison

(Lab of David C. Schwartz)

Outline

- How variable are DNA sequences?
- How can we “see” genomic variation?

  Using genome-wide optical maps.
- Algorithms for optical maps.
- How Condor enables our analysis
  - Makes computations feasible
  - Enables programmer efficiency (in lieu of cycles)
  - Alters how I think about algorithms.
- Some examples:
  - Cancer genomic rearrangements
  - Uncovering sequence errors

How do DNA sequences differ?

20th Century:

How similar is your DNA to mine?

Ans: About 99.9%

A few large differences (e.g. immune response loci).
Some diseased, abnormal genomes differ at chromosomal scale.

21st Century:

How much does your DNA differ from mine?

Ans: About 10% of the genome is variable.

Human Genomic Structural Variation

- Small (single base pair, 10's of base pairs)
  - Single nucleotide polymorphisms (SNPs)
    - On the order of 10^6 out of 10^9 bases.
  - High throughput methods developed in 1990s.
- Larger (100's – millions of base pairs)
  - Insertions, deletions.
  - Copy number changes.
  - Large-scale, complex rearrangements.

Genome-wide restriction maps

- SwaI (ATTT^AAAT)
  - Human genome (3 Gb) has 220,000 SwaI sites.
  - 15 kb average fragment size.
- Restriction map of a genome
  - provides information analogous to sequence (at a coarser resolution).
  - is also a sequence (sequence of the lengths between consecutive restriction sites).
- Concepts and algorithms from genomic sequence analysis have analogues.

Comparative genomics with optical maps

Discovering genomic structural variation with optical mapping

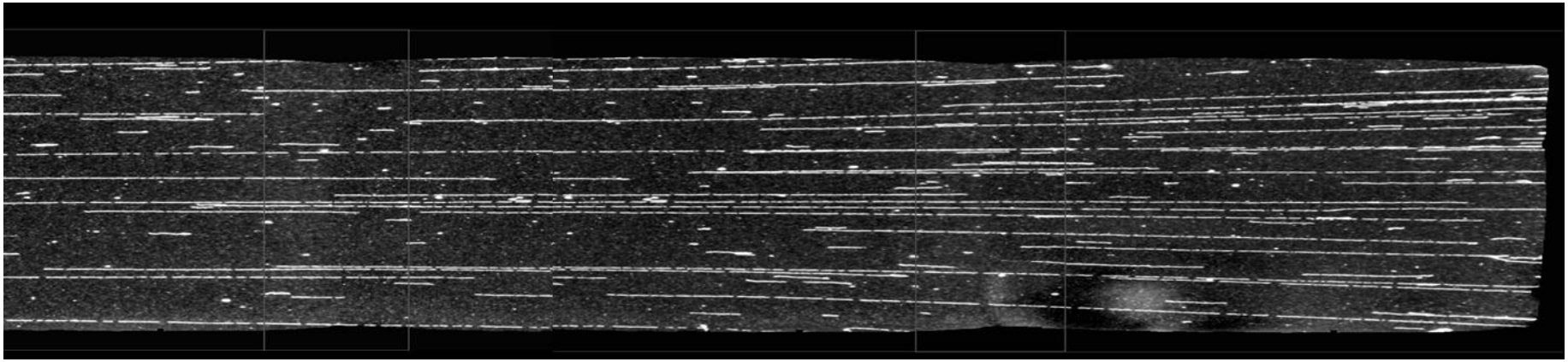
- Generate genome-wide map.
- Compare to *in silico* map to discover structural variation.
  - Identical regions (up to optical mapping error)
  - Local differences.
    - Missing and extra cuts
    - Indels (within restriction fragment (RFLPs); involving multiple fragments)
    - Other
  - Major discordances

Optical Structural Alterations (OSA)

- **MCs**
- **ECs**
- **Deletions**
- **Insertions**

Single-molecule optical maps

- Ordered restriction map obtained from a single DNA molecule.
  - Stretch DNA molecules and attach to substrate.
  - Digest with restriction enzyme.
  - Stain with fluorescent dye and process image.
  - Analogous to sequence read.



Measurement is error-prone

Missing, extra cuts; length errors
Gross errors.

Scaling assembly to mammalian genomes

- *Initialize*:

  number_of_iterations = 8;

  reference_mapset = NCBI build 36;

- *Main loop*:

  for (1 .. number_of_iterations) {

  align optical maps to reference_mapset**;** //**SOMA**: S. Kohn

  foreach (cluster of hits) {

  assemble (cluster**);** // **Gentig**: Anantharaman,Mishra,Schwartz

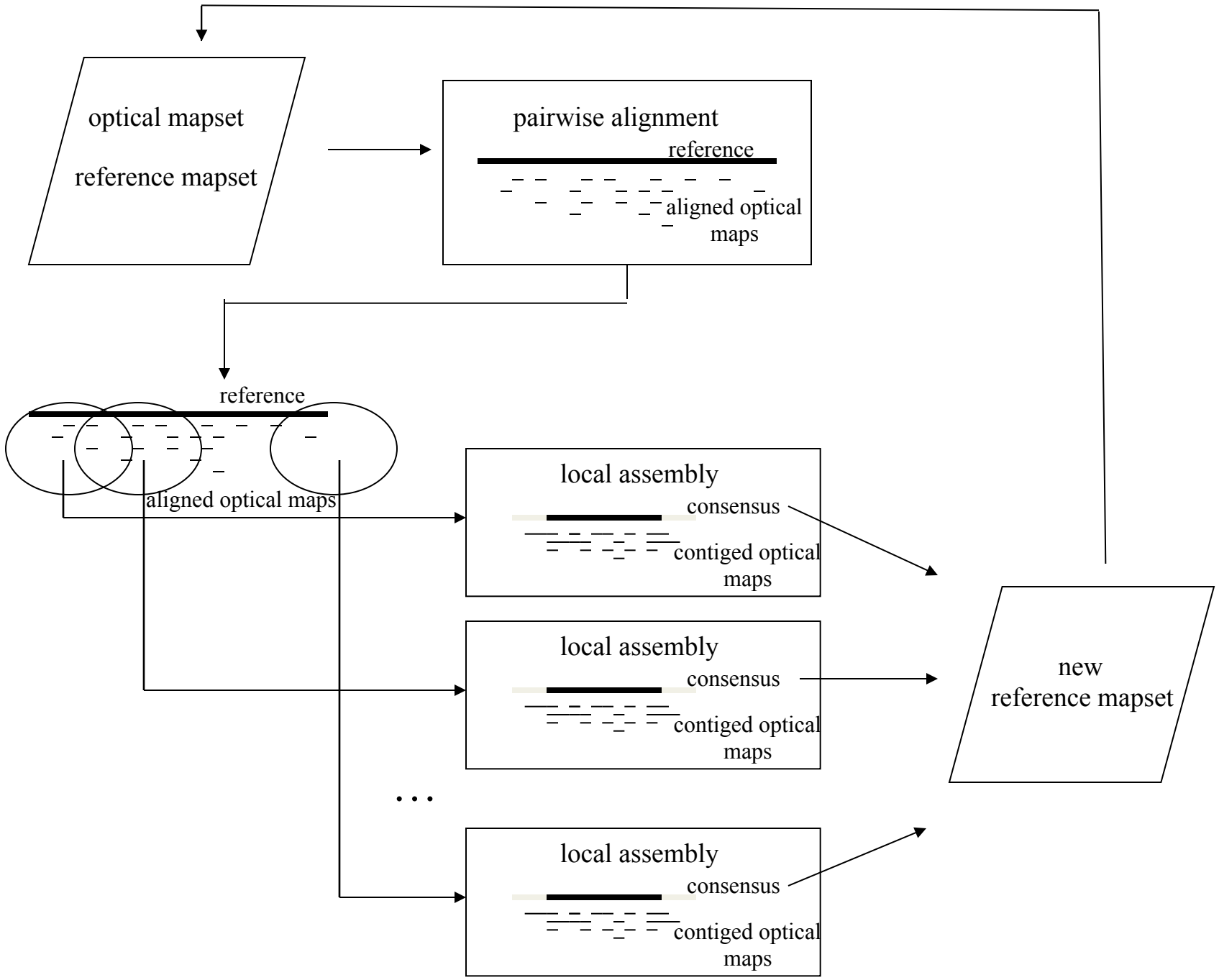
  }

  reference_mapset = set of consensus maps from assemblies;

  }

Manual editing:

- resolve redundancies in reference_mapset;


optical mapset
reference mapset
pairwise alignment
reference
aligned optical maps
reference
aligned optical maps
local assembly
consensus
contiged optical maps
local assembly
consensus
contiged optical maps
local assembly
consensus
contiged optical maps
new reference mapset

Leveraging the power of HTC

- Makes computations feasible
- Enables programmer efficiency
  - Trade programmer time and algorithmic uncertainty for inefficient compute cycles.
- Alters how I think about algorithms.

Genome-wide restriction maps

- About 12 human maps
  - 4 Lymphoblastoid cell lines ; CHM("normal")
    - GM10860, GM15510, GM18994
  - 3 stem cell lines, 2 cancer cell lines
  - Oligodendroglioma tumor slices
- Other mammalian maps:
  - Mouse (C57BL/6)
  - Rat (BN and SHR; BNLX)
- Rice
- Maize
- Many bacterial and microbial genomes

OSAs in lymphoblastoid maps

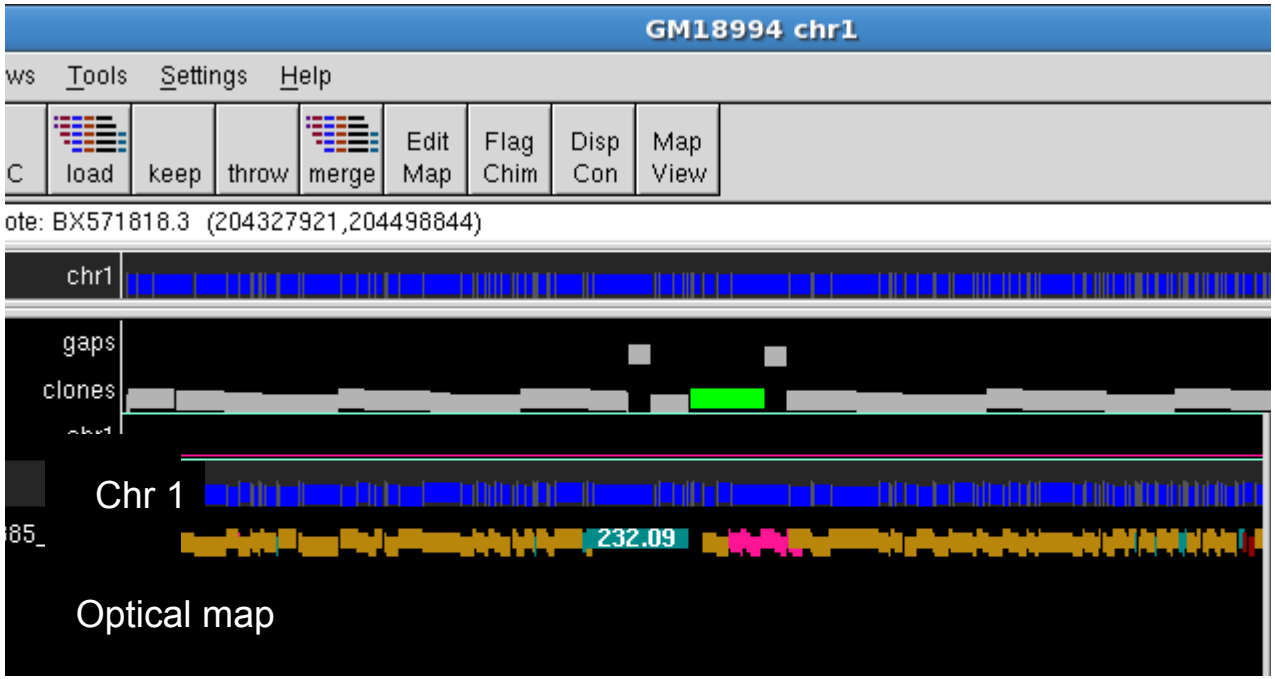
The human genome is dynamic!

Summary of Structural Variants Discerned by Optical Mapping

	EC	MC	Ins	Del	Other	Unique	Int.1	Int.2	Int.3	Total
CHM	465	446	165	183	96	471	283	273	322	1355
GM15510	556	384	447	105	105	616	387	417	322	1753
GM10860	584	352	631	350	86	777	447	411	322	2003
GM18994	535	409	523	384	90	735	443	411	322	1941
Total	2140	1555	1766	1214	377	2599	780	504	322	

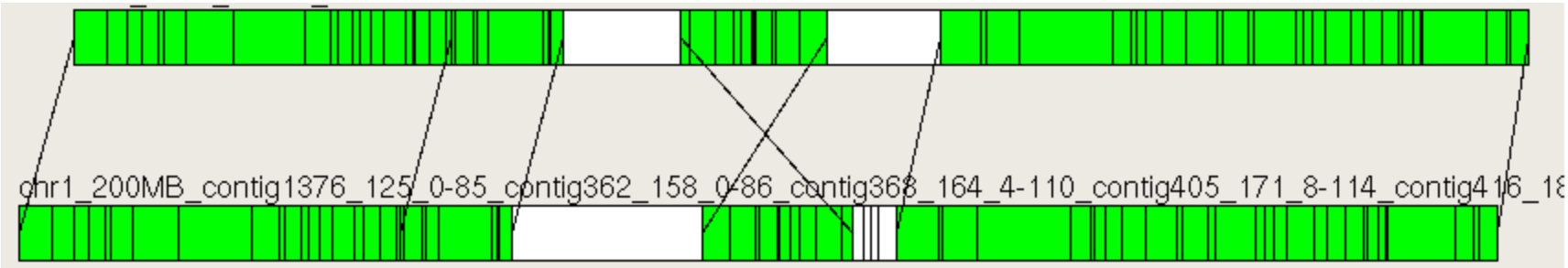
These findings are consistent with the emerging literature on human structural genomic variation. [Wigler (2003), Eichler, McCarroll, Sebat, Redon, Kidd, Redon, Conrad, Sharp, Tuzun, Iafrated, Korbel, …].

GM18994 at chr1:204.2 Mb



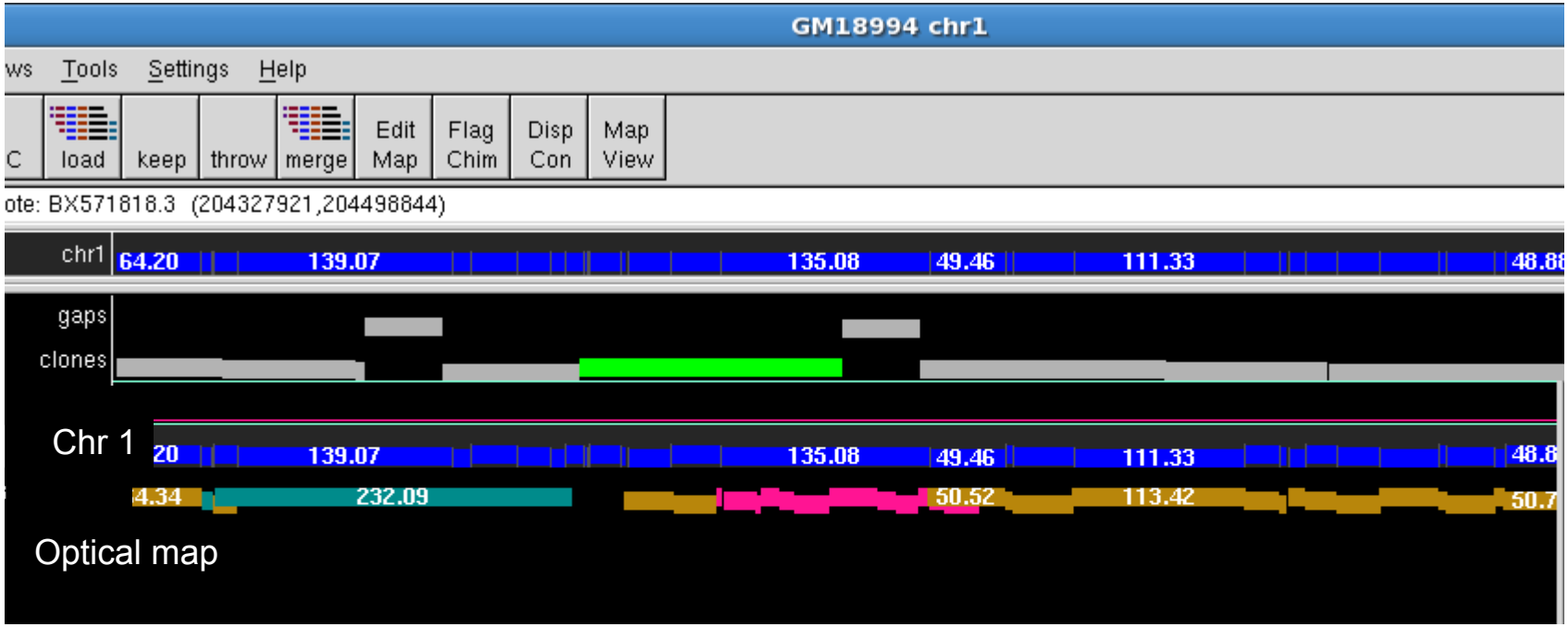
GM18994 at chr1:204.2 Mb

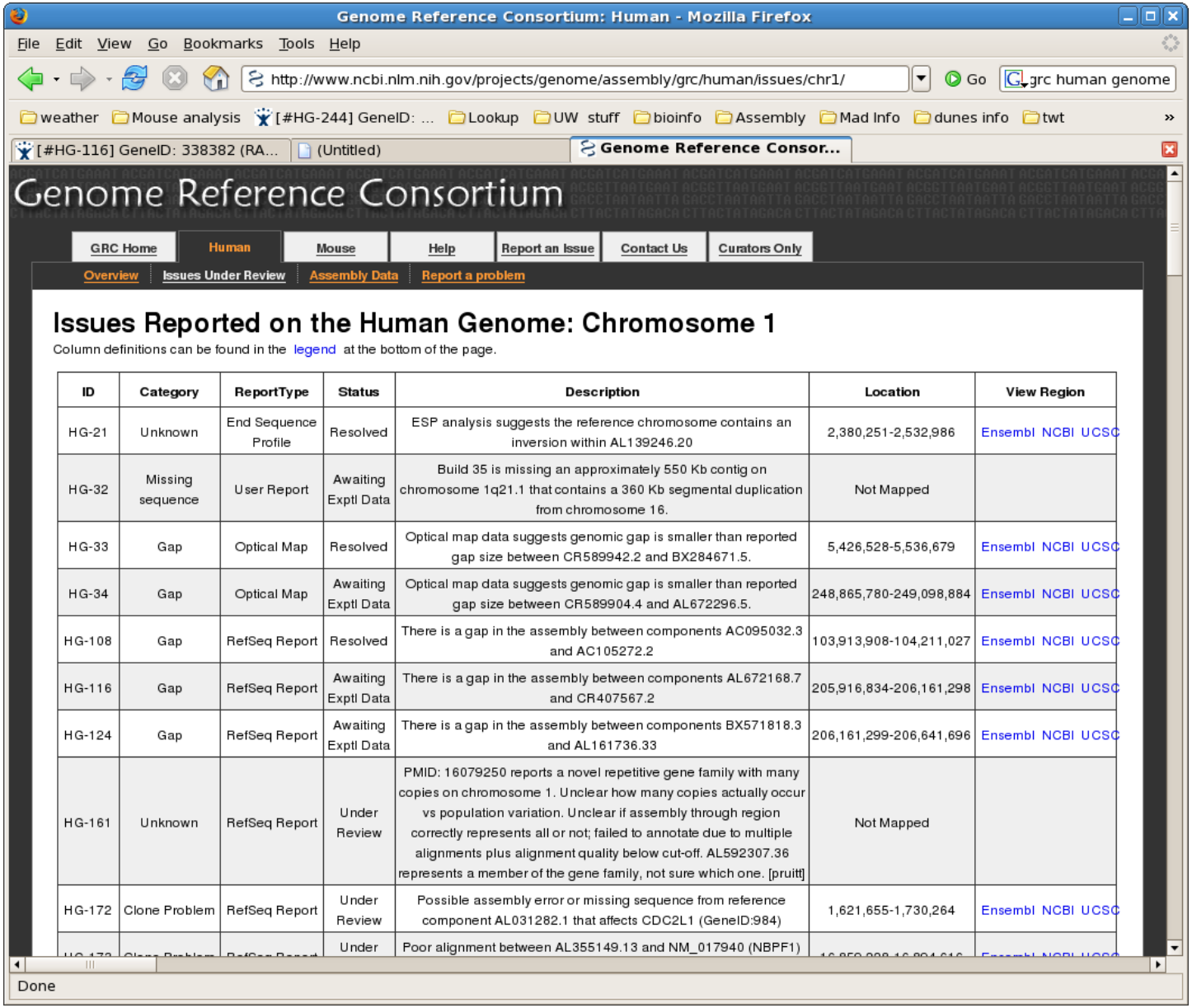
Chr 1



Optical map

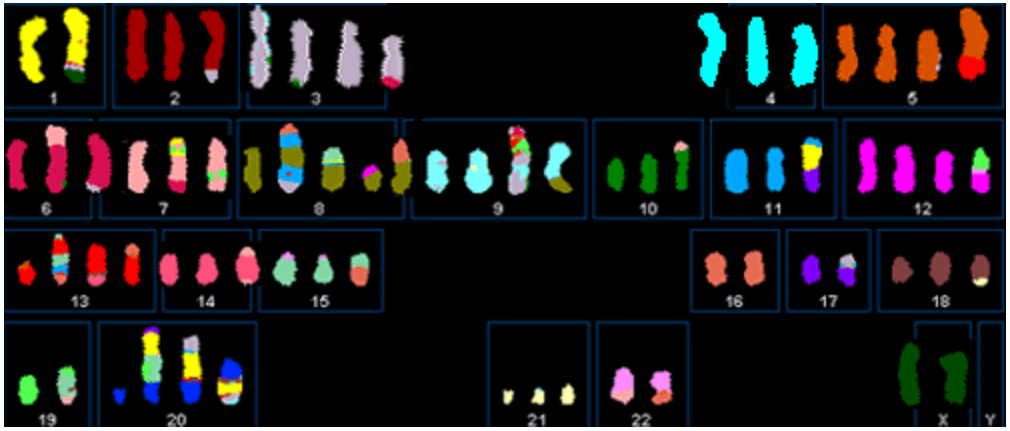
GM18994 at chr1:204.2 Mb




Genome Reference Consortium: Human - Mozilla Firefox
File Edit View Go Bookmarks Tools Help
http://www.ncbi.nlm.nih.gov/projects/genome/assembly/grc/human/issues/chr1/
Go
grc human genome
weather
Mouse analysis
[#HG-244] GeneID: ...
Lookup
UW stuff
bioinfo
Assembly
Mad Info
dunes info
twt
[#HG-116] GeneID: 338382 (RA...
(Untitled)
Genome Reference Consor...
Genome Reference Consortium
GRC Home
Human
Mouse
Help
Report an Issue
Contact Us
Curators Only
Overview
Issues Under Review
Assembly Data
Report a problem
Issues Reported on the Human Genome: Chromosome 1
Column definitions can be found in the legend at the bottom of the page.
ID
Category
ReportType
Status
Description
Location
View Region
HG-21
Unknown
End Sequence Profile
Resolved
ESP analysis suggests the reference chromosome contains an inversion within AL139246.20
2,380,251-2,532,986
Ensembl NCBI UCSC
HG-32
Missing sequence
User Report
Awaiting Exptl Data
Build 35 is missing an approximately 550 Kb contig on chromosome 1q21.1 that contains a 360 Kb segmental duplication from chromosome 16.
Not Mapped
HG-33
Gap
Optical Map
Resolved
Optical map data suggests genomic gap is smaller than reported gap size between CR589942.2 and BX284671.5.
5,426,528-5,536,679
Ensembl NCBI UCSC
HG-34
Gap
Optical Map
Awaiting Exptl Data
Optical map data suggests genomic gap is smaller than reported gap size between CR589904.4 and AL672296.5.
248,865,780-249,098,884
Ensembl NCBI UCSC
HG-108
Gap
RefSeq Report
Resolved
There is a gap in the assembly between components AC095032.3 and AC105272.2
103,913,908-104,211,027
Ensembl NCBI UCSC
HG-116
Gap
RefSeq Report
Awaiting Exptl Data
There is a gap in the assembly between components AL672168.7 and CR407567.2
205,916,834-206,161,298
Ensembl NCBI UCSC
HG-124
Gap
RefSeq Report
Awaiting Exptl Data
There is a gap in the assembly between components BX571818.3 and AL161736.33
206,161,299-206,641,696
Ensembl NCBI UCSC
HG-161
Unknown
RefSeq Report
Under Review
PMID: 16079250 reports a novel repetitive gene family with many copies on chromosome 1. Unclear how many copies actually occur vs population variation. Unclear if assembly through region correctly represents all or not; failed to annotate due to multiple alignments plus alignment quality below cut-off. AL592307.36 represents a member of the gene family, not sure which one. [pruitt]
Not Mapped
HG-172
Clone Problem
RefSeq Report
Under Review
Possible assembly error or missing sequence from reference component AL031282.1 that affects CDC2L1 (GeneID:984)
1,621,655-1,730,264
Ensembl NCBI UCSC
Under
Poor alignment between AL355149.13 and NM_017940 (NBPF1)
Done

Mapping breakpoints in MCF7

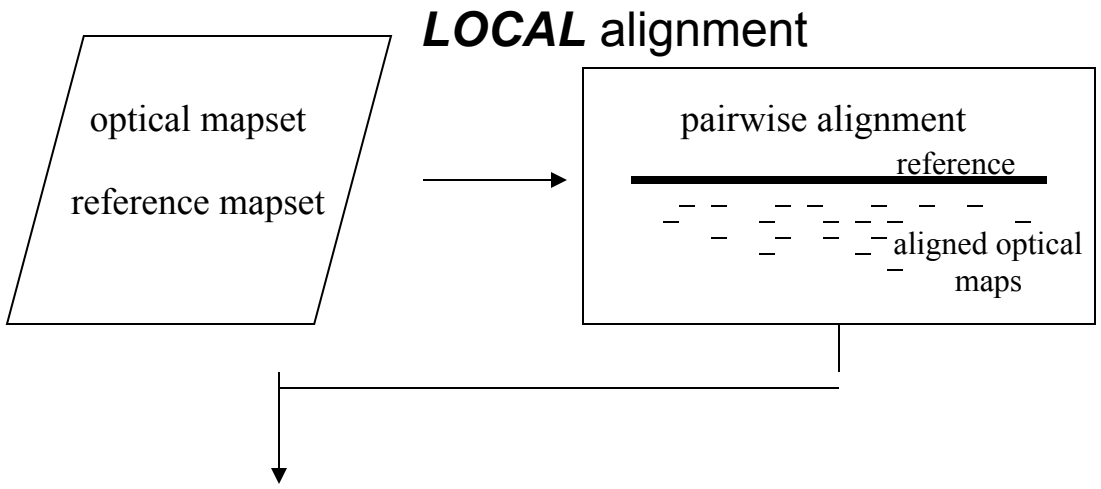
- Well-studied breast cancer cell line with abnormal karyotype.



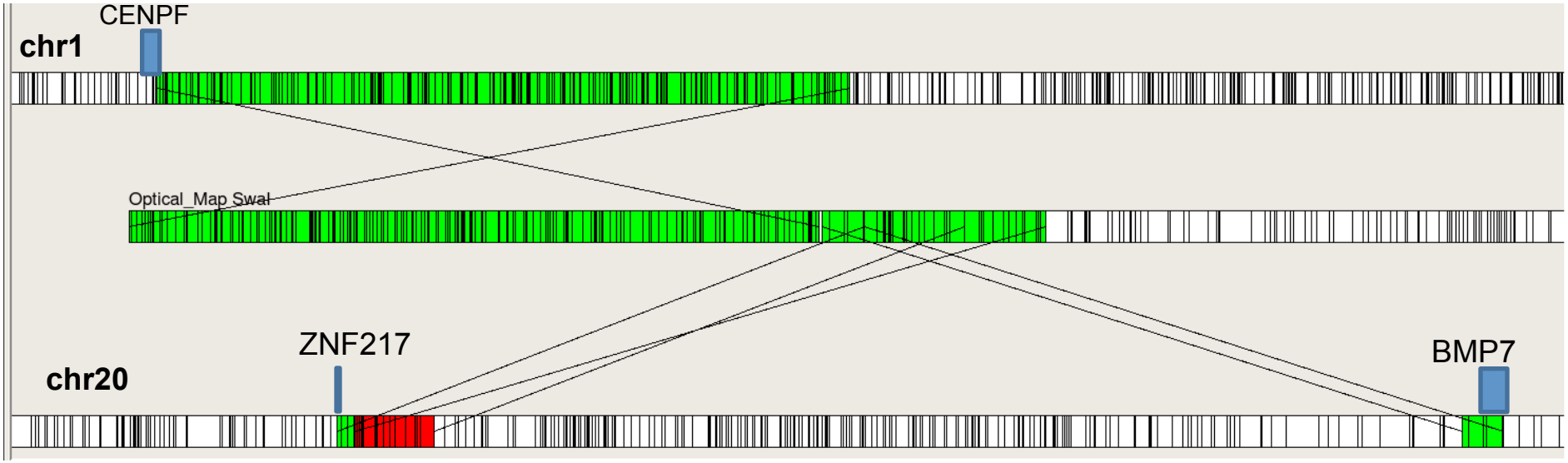
- Optical map reveals 58 breakpoints, 7 potential fusion genes.

MCF7 breakpoint assemblies

Local alignment in first iteration.



Chr1-20 rearrangement involving CENPF, ZNF217 and BMP7.



Chr1: 212884252 – 215064600 (R)
Truncates CENPF

Chr20: 55124197 – 55254799
Chr20: 51618363 - 51888308
} BMP7-ZNF217 fusion?

Chr20: 51705585 – 51888308 (R)
Inverted duplication near ZNF217

Collaborators

- Deepayan Sarkar and Michael Newton
- Rod Runnheim, Mike Bechner, Casey Lamers
- Brian Teague, Jill Herschleb, Susan Reslewic
- Adam Briska, Scott Kohn
- Shiguo Zhou, Gus Potamousis
- April Cook (Broad); Deanna Church (NCBI); Jo Wood (Sanger)
- Dave Schwartz

Partial list of references

1. **High-resolution human genome structure by single molecule analysis.** B. Teague, M. Waterman,S. Goldstein,K. Potamousis,S.Zhou,S. Reslewic, D.Sarkar,A. Valouev,C. Churas,J. Kidd,S. Kohn,R. Runnheim,C. Lamers,D. Forrest,M. Newton,E. Eichler,M. Kent-First,U. Surti,M. Livny,D. Schwartz. PNAS. 107:24. (2010). PDF
2. Deepayan Sarkar's thesis (UW Stats)
3. Human genomic variation search in Google scholar
4. USC optical mapping algorithms
5. Genome Reference Consortium
6. Anantharaman T, Mishra B, Schwartz DC. Proceedings of the 7th International Conference on Intelligent Systems for Molecular Biology, Genomics via optical mapping. III. (1999)

Many Common “OSAs” are sequence errors

Genome-specifc events (not snip SNPs)		Common to two	
GM10860	552	GM10860-GM15510	103
GM15510	418	GM10860-GM18994	128
GM18994	536	GM15510-GM18994	87

Shared by three

271