Sequencing the unrearranged human Immunoglobulin

heavy chain locus (IgH) from hydatidiform mole

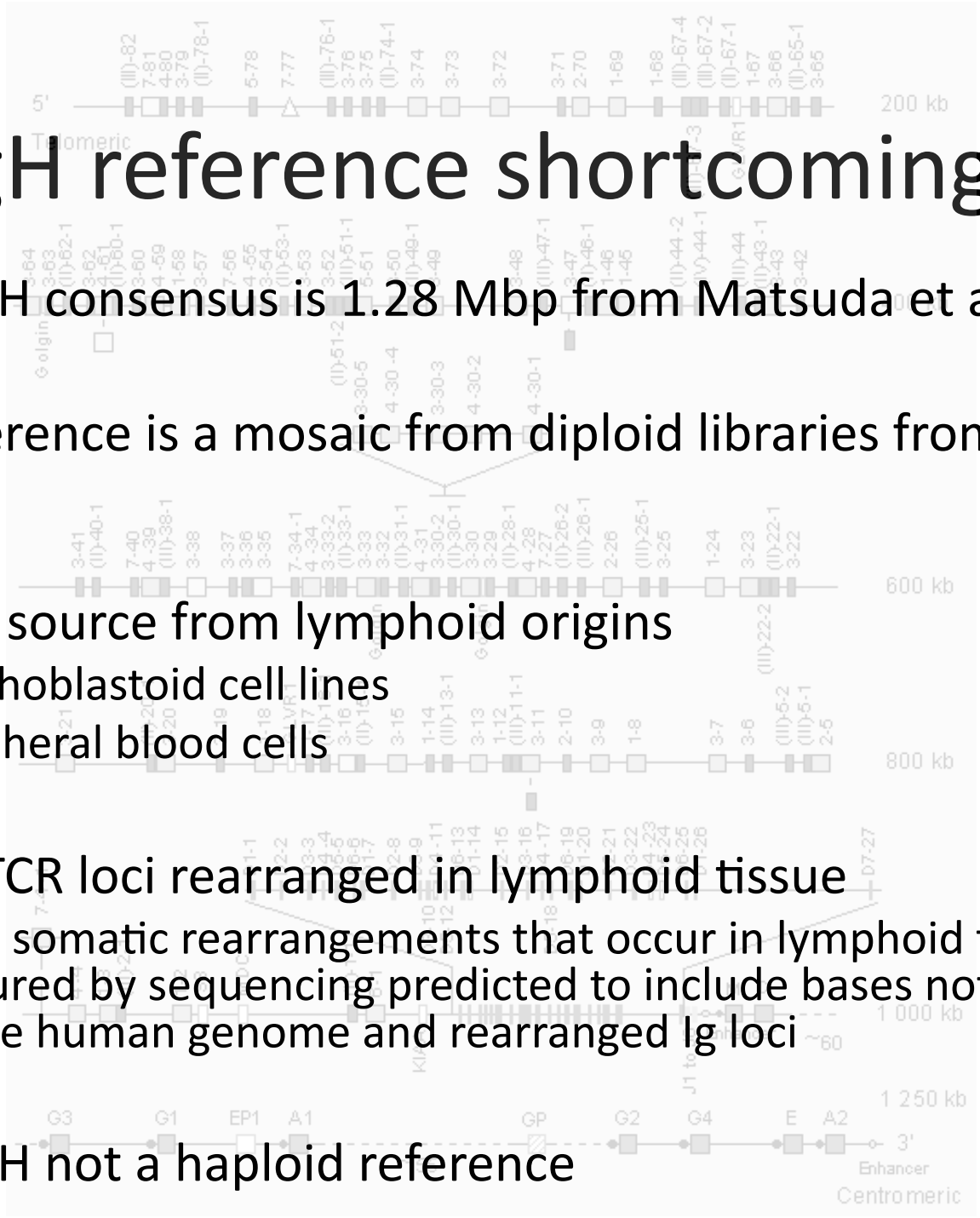
SFAF
Santa Fe, USA
René Warren – June 2010

Immunoglobulin (Ig)

- Loci comprises V,(D),J gene segments that recombine to make heavy chain (IgH) or light chain (IgK, IgL) messages
- Combinatorial rearrangements and "somatic hypermutation" (addition/removal of non-templates bases) creates astonishing theoretical antibody diversity 10^{14}
- Germline variation can confer disease susceptibility
  - Defective IgK V allele fail to protect Navajos natives against *H.influenzae* type B (Feeney, 1996)
  - Homozygous IgH-V deletions can cause nephritis (Lupus) (Cho, 2003)

IgH reference shortcomings

- hg19-IgH consensus is 1.28 Mbp from Matsuda et al. 1998
- IgH reference is a mosaic from diploid libraries from various sources
- Sample source from lymphoid origins
  - lymphoblastoid cell lines
  - peripheral blood cells
- Ig and TCR loci rearranged in lymphoid tissue
  - V(D)J somatic rearrangements that occur in lymphoid tissue and captured by sequencing predicted to include bases not encoded by the human genome and rearranged Ig loci
- hg19-IgH not a haploid reference



Hydatidiform mole

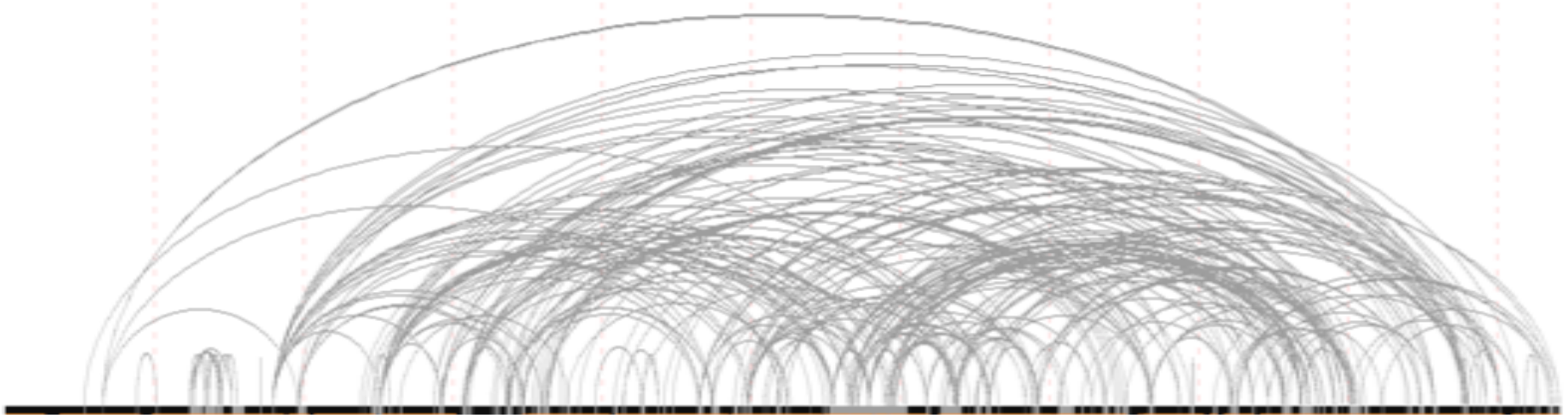
- Pregnancy abnormality (human) generated by fertilization of an empty egg

- Haploid, non-lymphoid
  - Source of unrearranged Ig gene segments
- BAC physical map exists (BCGSC) and resources available [BACPAC] (CHORI)

1.0 mm

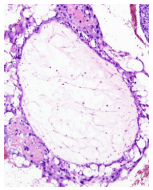
Sequencing challenges

- Human genome interspersed repeats
- Ig gene segments (VDJ) share high sequence homology
- IgH most segmentally duplicated locus in the human genome

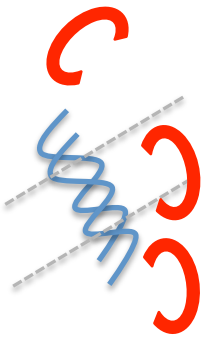


showing hg19-IgH repeats >=76bp with <=10%mismatch

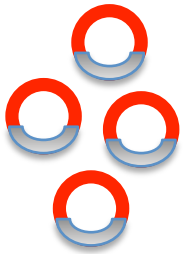
Approach



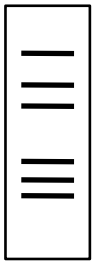
Mole cell line



Genomic DNA extraction, shearing

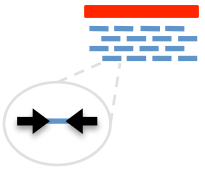


BAC library construction



Fingerprint BACs

Build map, select overlapping BACs



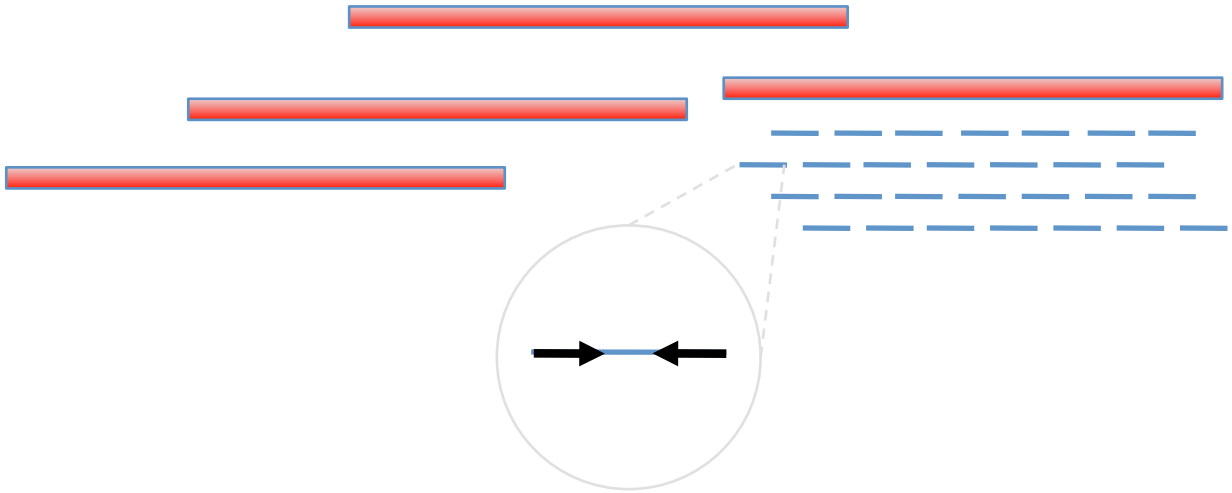
Build illumina sequencing library, index,pool, sequence, assembly

BACPAC (CHORI)

BCGSC mapping

BCGSC/SFU

High-throughput sequencing



- IgH,IgK and IgL tiling path BAC sequences sheared
- BAC Sequence library built & indexed for HT sequencing
  - hexamer barcode (Illumina)
- BAC libraries pooled
- Sequenced onto 1 lane Illumina GAIIx instrument
  - 27.6M x 76bp PET, 200nt insert size
  - ~1950-fold average sequence coverage

De novo short read assembly

- Using in-house assembler SSAKE (v3.5)
  - 1st short read assembler
  - continuously improving

BIOINFORMATICS APPLICATIONS NOTE
Vol. 23 no. 4 2007, pages 500–501
doi:10.1093/bioinformatics/btl629

Genome analysis

Assembling millions of short DNA sequences using SSAKE

René L. Warren*, Granger G. Sutton[1], Steven J. M. Jones and Robert A. Holt
British Columbia Cancer Agency, Genome Sciences Centre, 675 West 10th Avenue, Vancouver, BC V5Z 1L3, Canada and [1]J. Craig Venter Institute, 9704 Medical Center Drive, Rockville, MD 20850, USA

- Handles errors, mate pairs, reads of different lengths
  - Used in conjunction with base quality trimming

- Suited for IgH sequencing from BACs
  - Certain short read assemblers mishandle BAC overlap sequence redundancy as large repeats

http://www.bcgsc.ca/platform/bioinfo/software/ssake

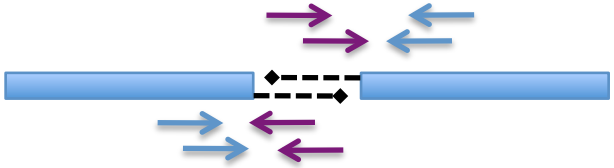
What's new in SSAKE v3.5?

❶ PET usage at run-time

– helps resolve repeats during contig extension

❷ Fill gaps between adjacent contigs in a scaffold

– fill gaps with missing/misplaced mates



E.coli K12 short read test set (SRA)

12.2M x 36bp mate pairs from Illumina GA :: 6.4M pairs and 2.9M unpaired quality trimmed

Metric	SSAKE v3.4	SSAKE v3.5
time	0:37:05	0:39:18
# contigs	896	856
N50 (bp)	12,433	11,946
largest (bp)	50,675	48,852
bases (bp)	4,569,095	4,569,002
% coverage	96.3 (880)	97.0 (842)
% accuracy	99.98	99.98
# merged contigs (fill gaps)	NA	563
N50 (bp)	NA	21,664
largest (bp)	NA	70,162
bases (bp)	NA	4,568,849
% coverage	NA	93.4 (544)
% accuracy	NA	99.88
# scaffolds	270	224
N50 (bp)	55,882	69,464
largest (bp)	134,207	221,647
Satisfied pairs (%)	99.75	99.79

❶ Fewer contigs & scaffolds providing higher coverage and higher scaffold contiguity

run-time PET usage

❷ N50 doubled, #contigs halved after automated gap closure

fill gaps

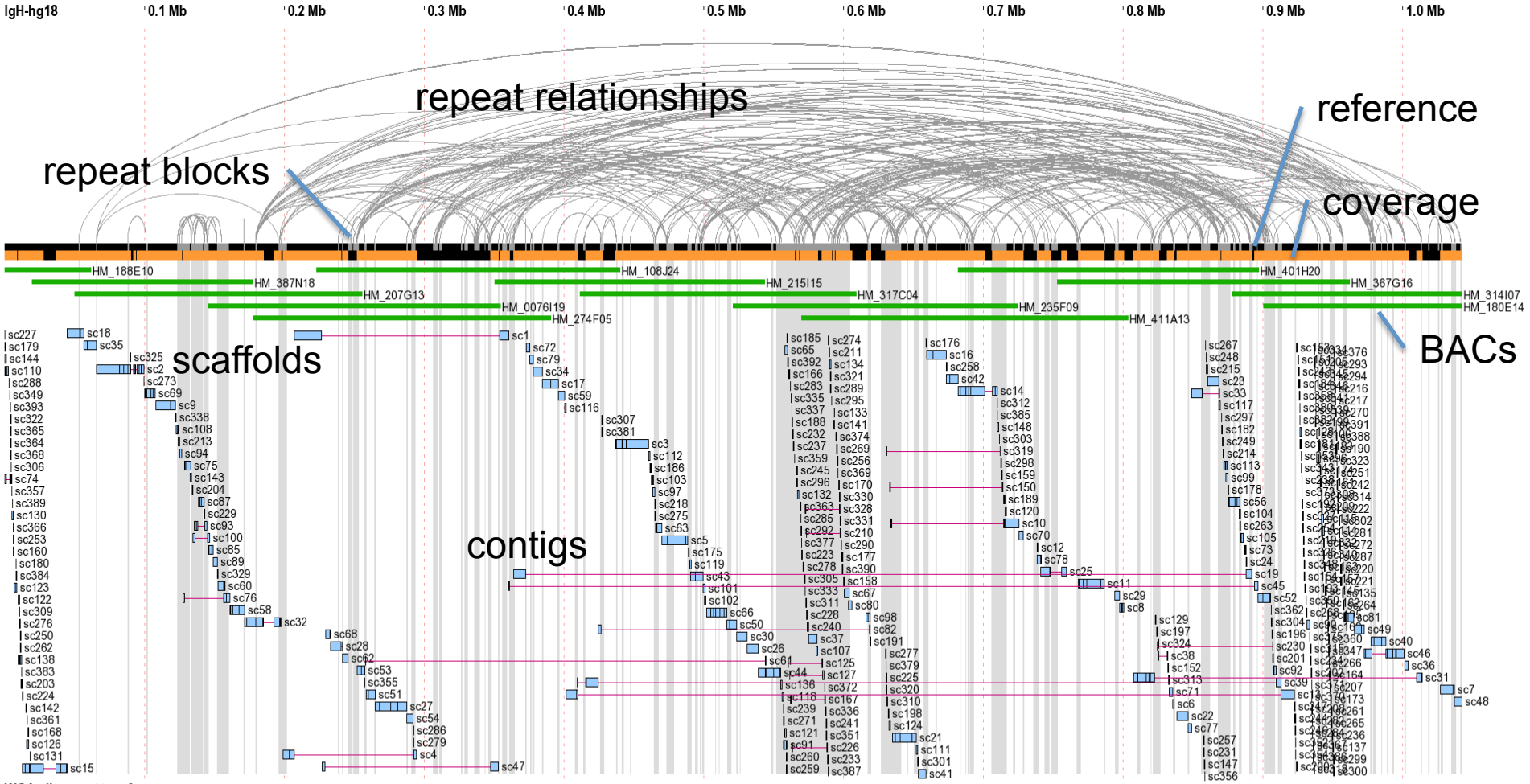
IgH assemblies

Metric	SSAKE v3.4	SSAKE v3.4 data cleaned*	SSAKE v3.4 stringency up	SSAKE v3.5 pair@RT/fill gap
# contigs^	1,817	1,293	936	725
N50 (bp)	1,444	2,553	3,181	4,704
largest (bp)	14,322	19,427	19,422	19,421
bases (bp)	1,390,273	1,222,204	1,128,894	1,119,786
# merged contigs (fill gaps)	na	na	na	508
N50 (bp)	na	na	na	8,313
largest (bp)	na	na	na	31,411
# scaffolds	720	689	431	390
N50 (bp)	9,199	9,785	11,535	12,585
largest (bp)	61,664	53,487	49,960	50,660
Satisfied pairs (%)	95.8	96.3	97.2	97.9

*vector and *E.coli* reads removed, segregate IgH reads on HQ indexes (>=Q20)
^contigs >=200bp analyzed

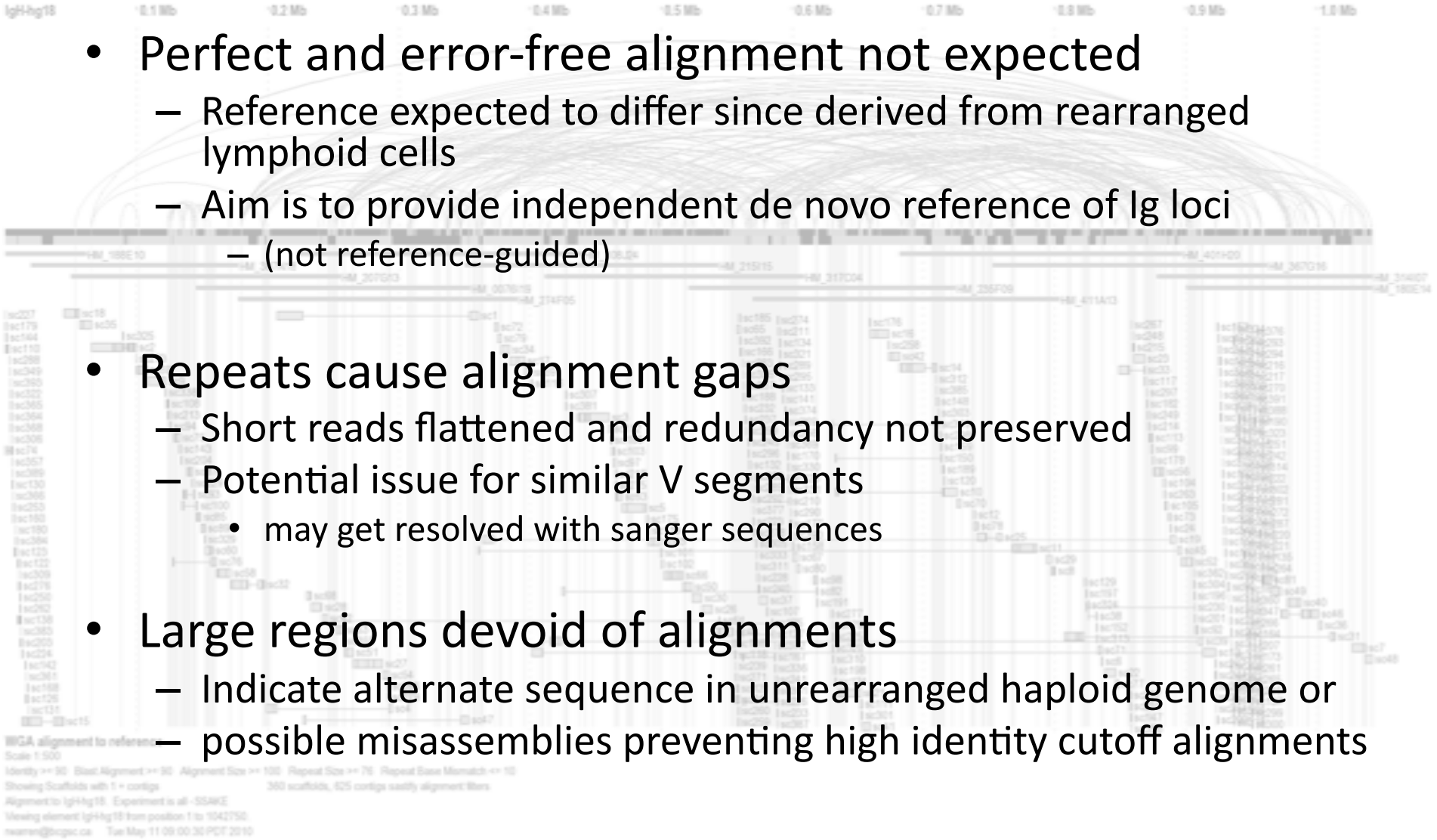
RT: run-time
SC: scaffolding

Comparison to a reference



80% coverage by contigs >=90% seq.id.

Comparison to a reference



- Perfect and error-free alignment not expected
  - Reference expected to differ since derived from rearranged lymphoid cells
  - Aim is to provide independent de novo reference of Ig loci
    - (not reference-guided)
- Repeats cause alignment gaps
  - Short reads flattened and redundancy not preserved
  - Potential issue for similar V segments
    - may get resolved with sanger sequences
- Large regions devoid of alignments
  - Indicate alternate sequence in unrearranged haploid genome or
  - possible misassemblies preventing high identity cutoff alignments

In the works

- 5kb plasmid library from minimum BAC tiling set
  - light sequence coverage from plasmid clones
- Additional SSAKE developments
  - Support for paired-end sanger sequences
    - improve [short+long range] contiguity / fix local misassemblies
  - Support for libraries of different insert sizes
- Phase I: illumina as a framework to produce draft sequence
- Phase II: Assemblies of illumina+sanger data set and directed finishing
  - Primer walking
- Goal is to get a fully finished IgH sequence

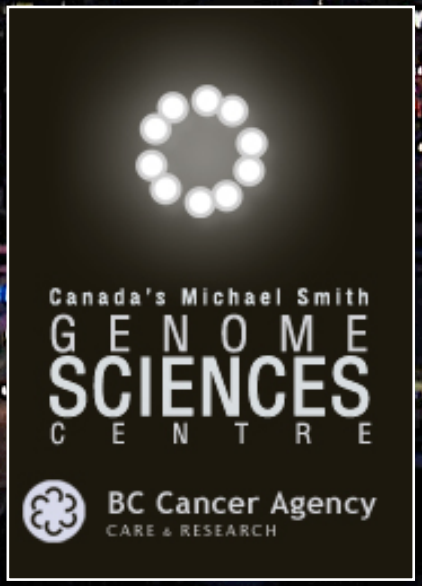
Acknowledgements

BC Genome Sciences Centre

- Rob Holt
- Jacquie Schein and BCGSC mapping
- BCGSC sequencing group

Simon Fraser University

- Felix Breden
- Corey Watson



http://www.bcgsc.ca/platform/bioinfo/software/ssake