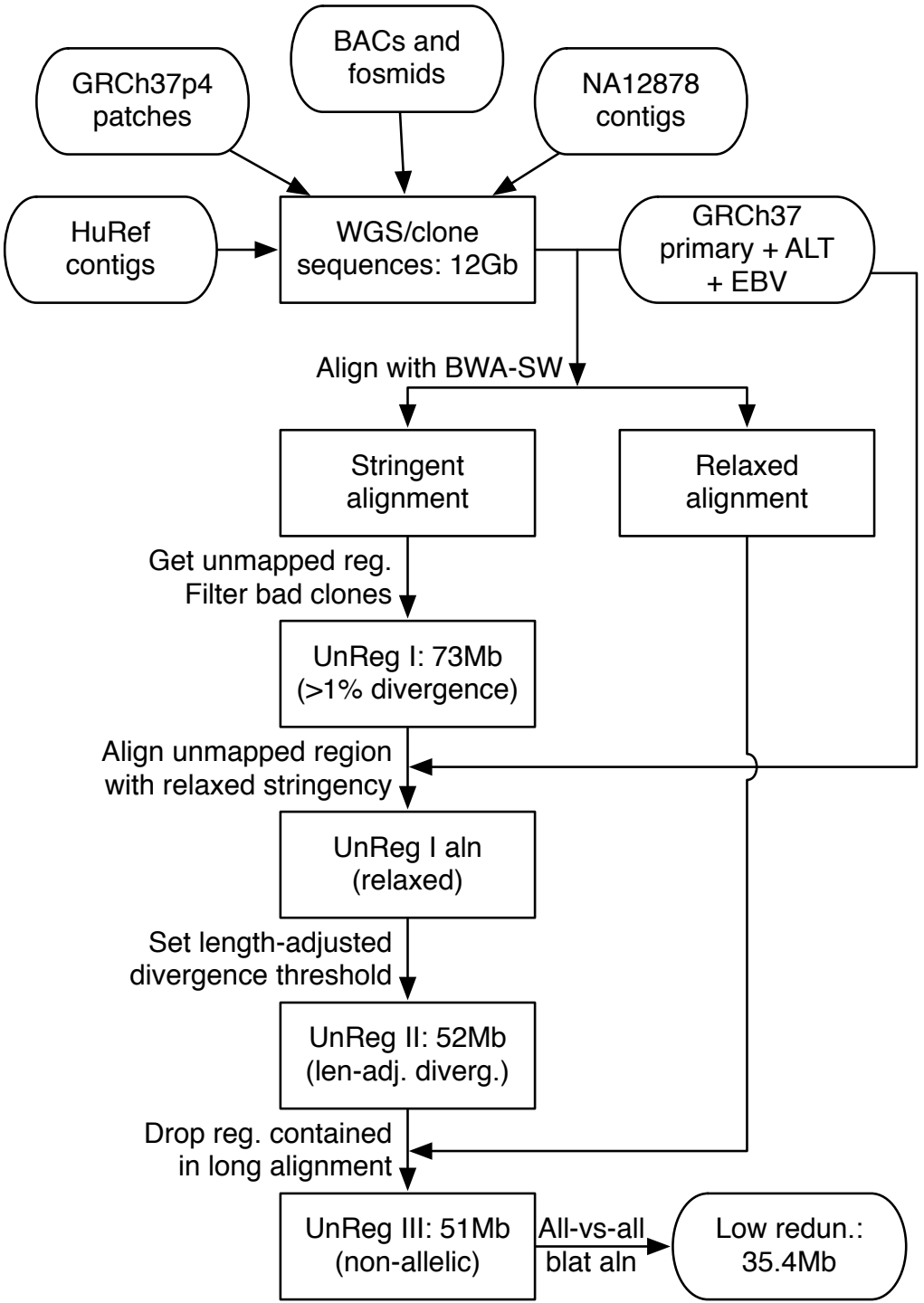
The missing human sequences (version 5)

2011-06-13

Heng Li

Summary

- Method: map WGS/clone sequences and extract regions not aligned well
- Decoy sequences: 35.4Mb segments; N50=22.9kb
  - low identity to the 1000g reference (length-dependent)
  - 50% satellite or simple repeat; 23% interspersed
  - 422kb non-repeat unaligned to human+chimp+gorilla.
  - Localized, if possible
- Including decoys helps SNP discovery
  - Improved mapping ratio: 96.8% => 99.2%
  - 1.5% unfiltered SNPs affected
  - Recommend to include decoys in the PhaseII mapping
  - Potentially useful to the SV group as well



- Stringent alignment:
  - set boundaries between high- and low-identity reg.
  - CMD: -b99 -q199 -r49
- Relaxed alignment
  - Identify allelic regions
  - Compute identity
  - CMD: -b3 -q5 -r2
- Extracting unmapped reg.:
  - Drop aln with score<500
  - Select regions >= 1000bp
- Length-adjusted threshold:
  - <95% for 500bp
  - <99% for 20kb+
  - Logistic regression for the rest of lengths
- List of problematic BACs provided by Deanna Church et al.

Characteristics of decoy sequences

- 35.4Mb; N50=22.9kb
- Repeat (based on RepeatMasker):
  - 50% are satellite or simple repeats (e.g. alphoid)
  - 23% are interspersed repeats (e.g. SINE/LINE/LTR)
- Align to human/chimp/gorilla combined:
  - 2.0/35.4Mb unaligned with BWA-SW default
  - 1.6/2.0Mb are satellites/low-complexity repeats
  - 422kb (probable) euchromatin unmappable
    - Real, assembly errors or remaining contaminations?

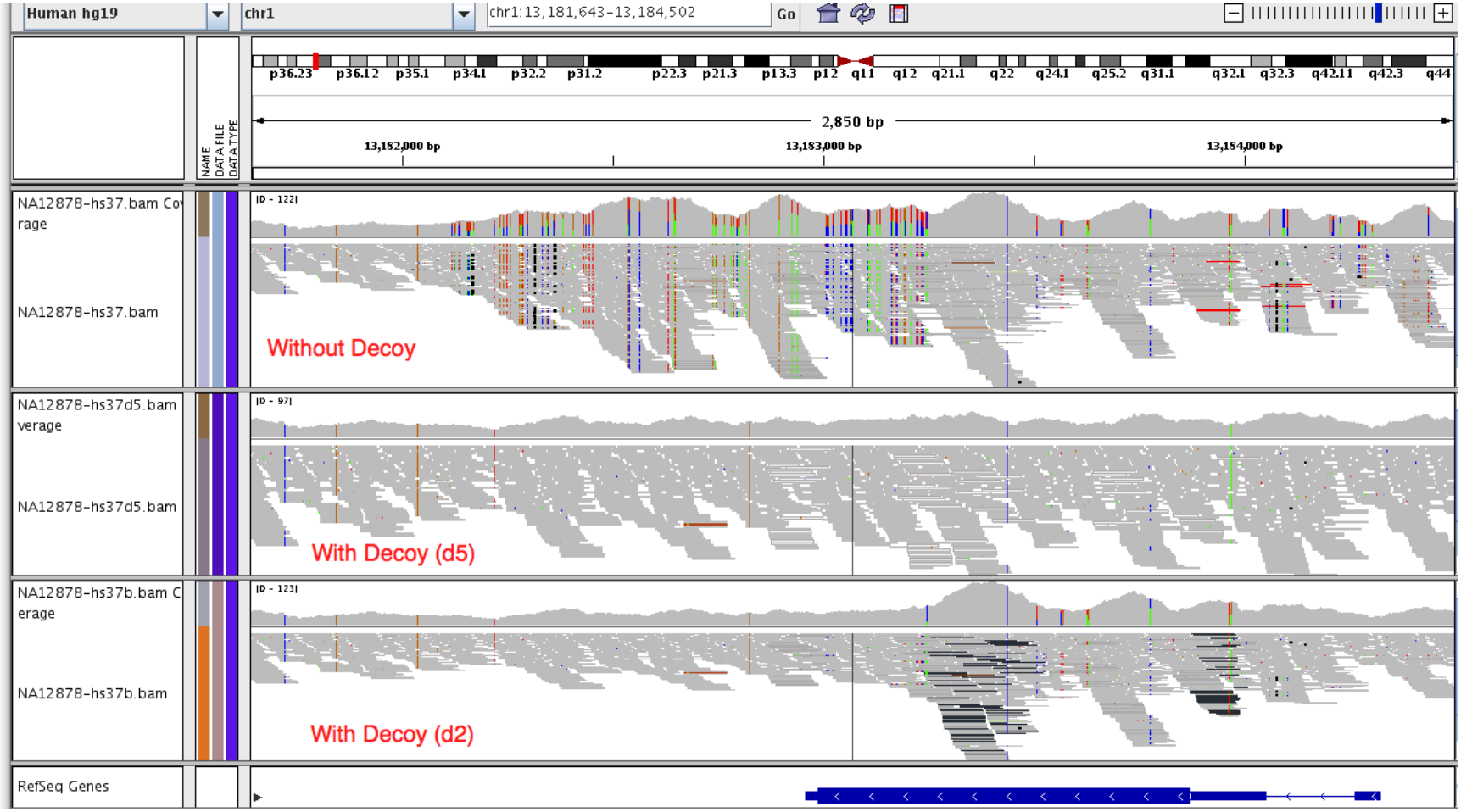
Effect on SNP calling

- Data: NA12878, 30X Hiseq
- Map to GRCh37+EBV w/ and w/o novoseq
- On autosomes: 3.5 million SNPs w/o filtering
  - Specific to alignment w/o decoy: 61,789 (~1.8%)
    - Called ***type-1 SNPs*** in the following screen shots
  - Specific to alignment w/ decoy: 5,234
    - Called ***type-2 SNPs***
- 10X read depth reduction in chr1 centromere, 40X reduction in chr10 (two ultra-deep regions)

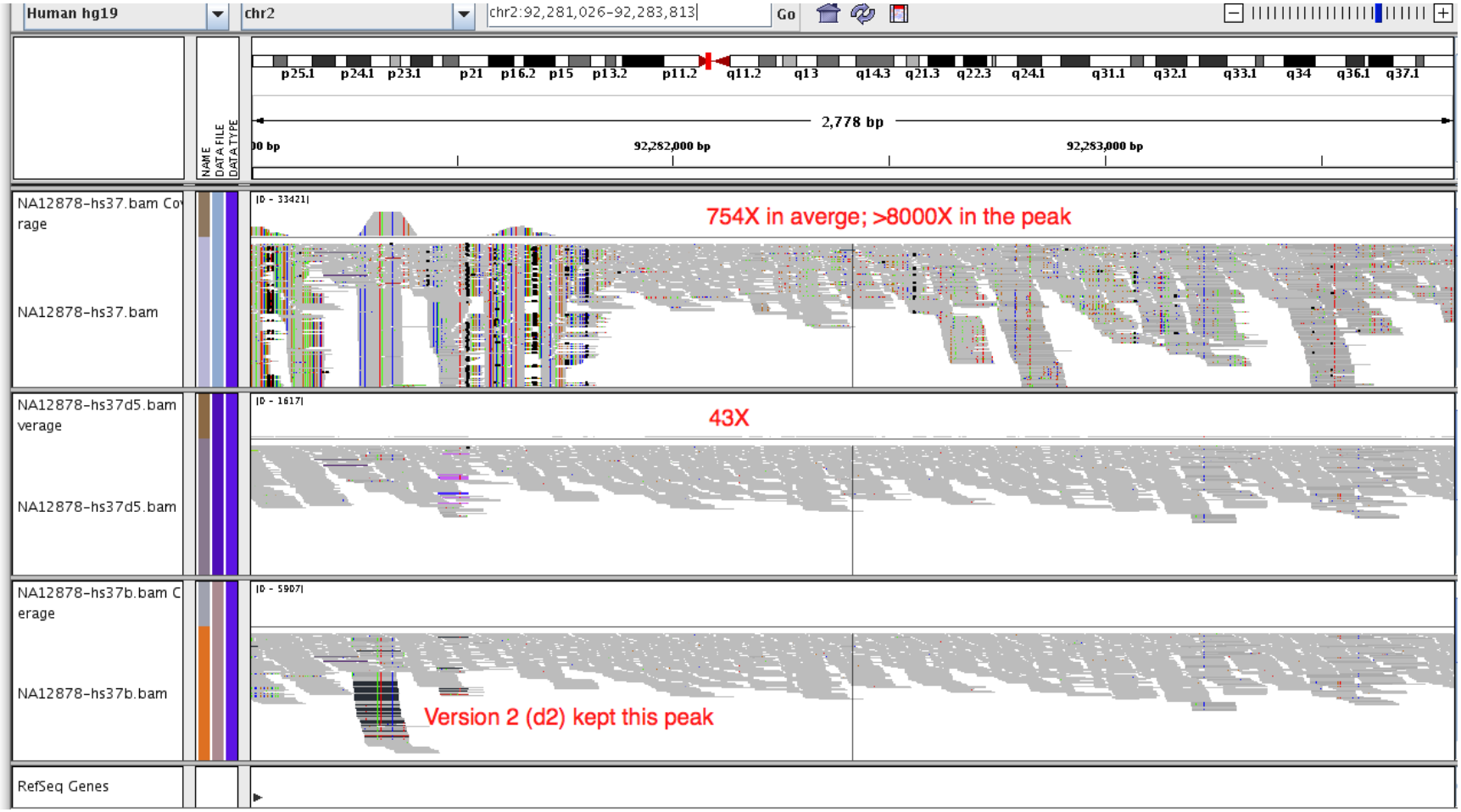
Manual confirmation

- For type-1 SNPs (called exclusively from alignment w/o decoys):
  - The majority have excessive read depth and enriched mismatches
  - A few arguable false negatives, most associated with poor flanking alignment
  - Have not found definite false negatives
- For type-2 SNPs:
  - Most have a snpQ close to 20, the threshold.
  - A minority of type-2 SNPs are FNs in type-1.

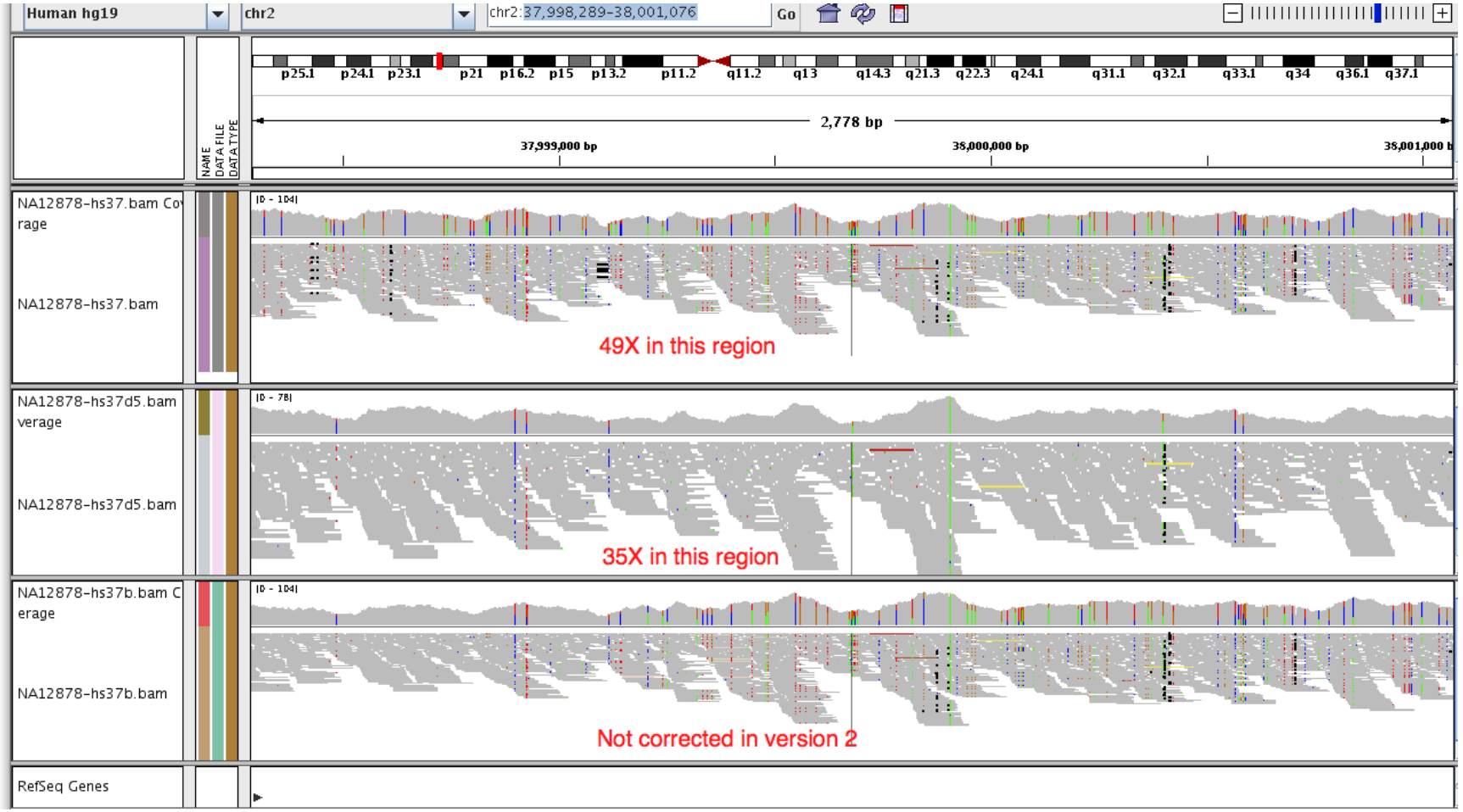
Type-1: Most look like this – excessive depth and mismatches



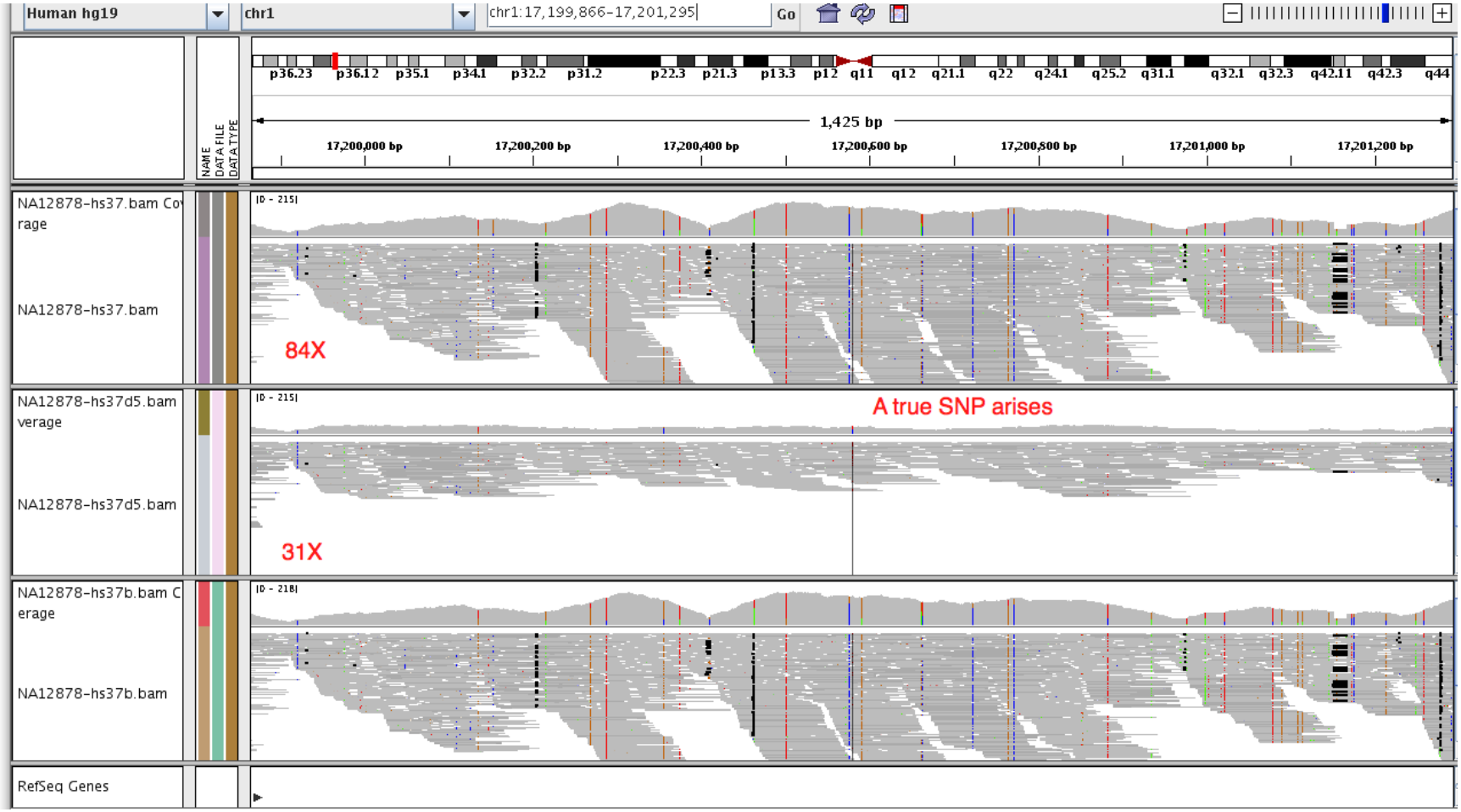
Type-1: much cleaner centromeric regions



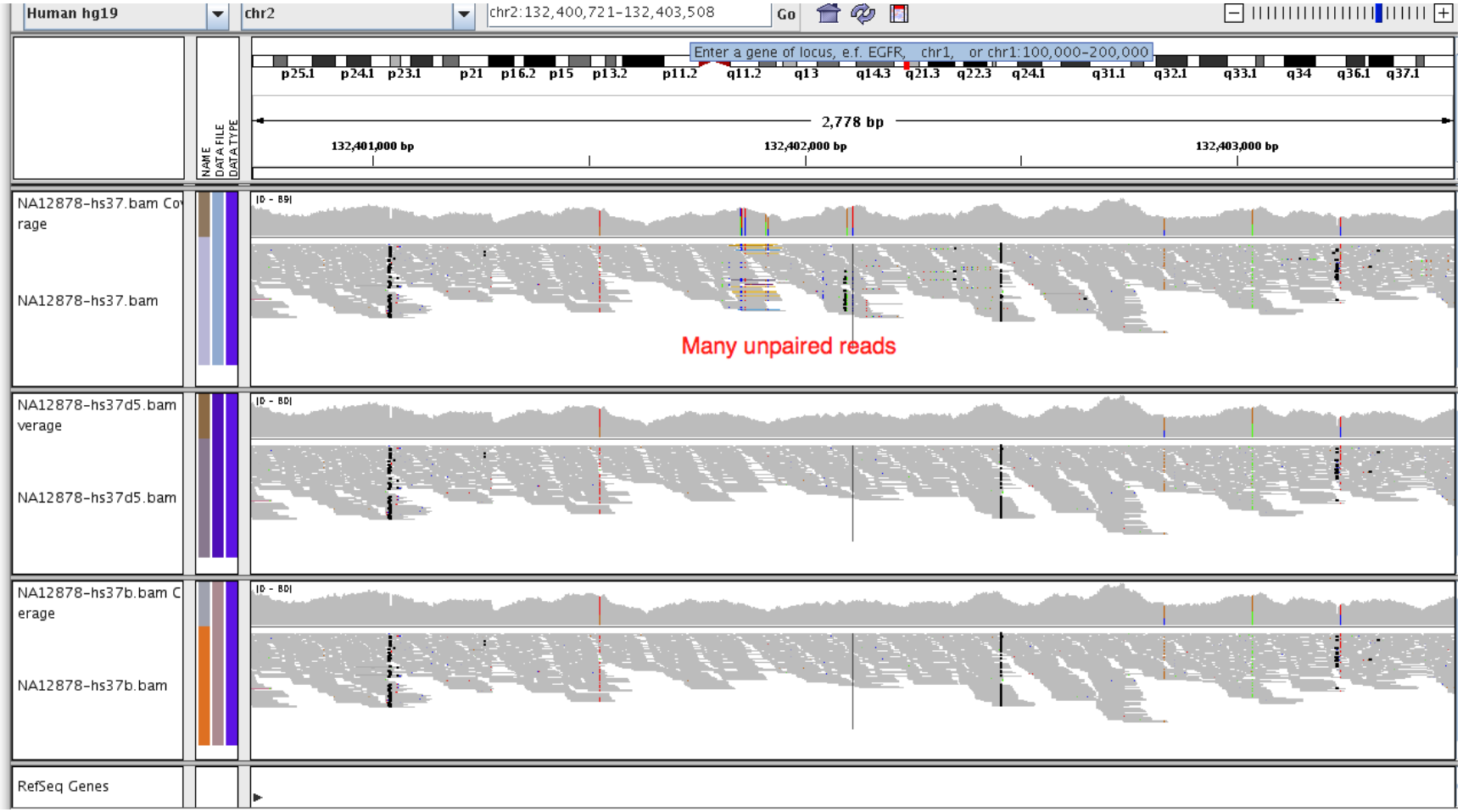
Type-1: Improvement over the previous version



Type-2: Incomplete genome may *occasionally* lead to false negatives

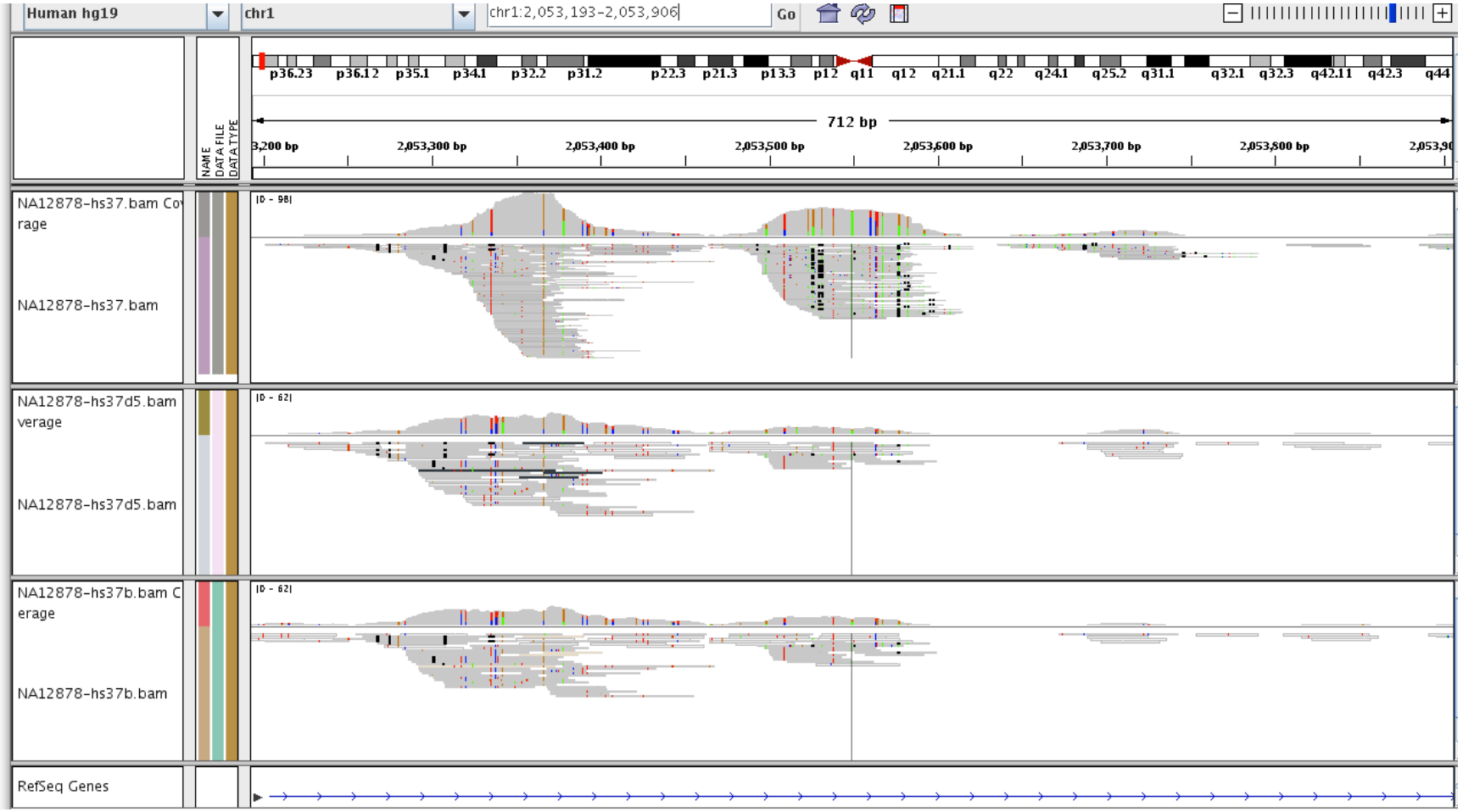


Type-1: short regions may also be affected by false alignments



Type-1: A questionable d5 false negative – very few

Type-1: Ambiguous – FN or FP?



Acknowledgements

- Deanna Church, Valerie Schneider and Nathan Bouk from NCBI

Version history

- V1: HuRef only; <90% identity; 13Mb; N50=3.8kb
- V2: BAC+HuRef+NA12878; <95-96% identity; 30.5Mb
- V3: Four sources; <~95-96% identity; control of allelic and contaminated sequences; 27.5Mb; N50=6.9kb
- V4: Length-adjusted identity (up to 99%); localized segments; improved contiguity; 37.1Mb
- V5: More careful control of allelic sequences; 35.4Mb; N50=22.9kb