Fine-scale structural variation of the human genome

Tuzun et. al. Nature Genetics (2005)

Hilal Kosucu

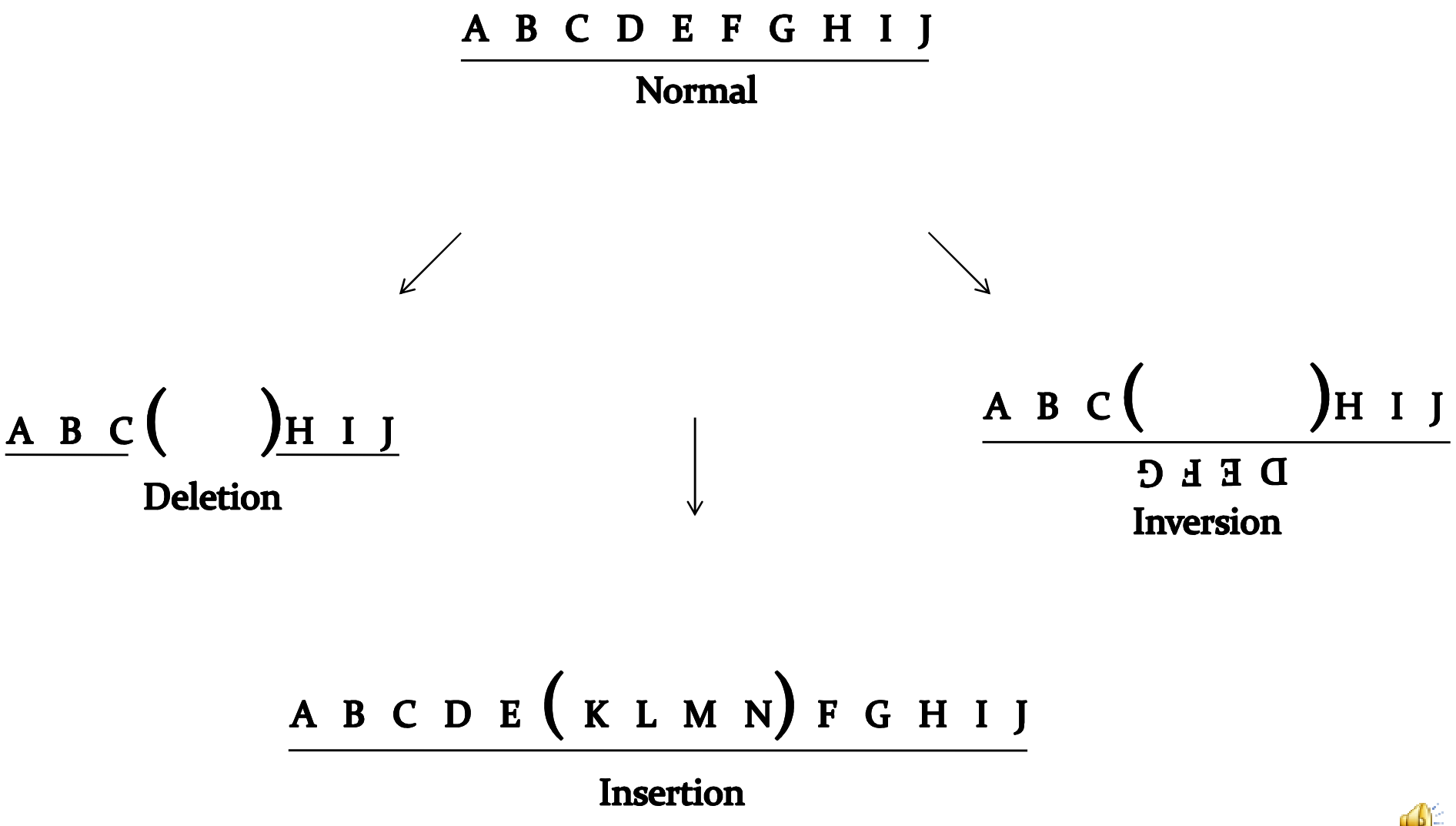
The spectrum of variation in the human genome

Variation	Rearrangement type	Size range[a]
Single base-pair changes	Single nucleotide polymorphisms, point mutations	1 bp
Small insertions/deletions	Binary insertion/deletion events of short sequences (majority <10 bp in size)	1–50 bp
Short tandem repeats	Microsatellites and other simple repeats	1–500 bp
Fine-scale structural variation	Deletions, duplications, tandem repeats, inversions	50 bp to 5 kb
Retroelement insertions	SINEs, LINEs, LTRs, ERVs[b]	300 bp to 10 kb
Intermediate-scale structural variation	Deletions, duplications, tandem repeats, inversions	5 kb to 50 kb
Large-scale structural variation	Deletions, duplications, large tandem repeats, inversions	50 kb to 5 Mb
Chromosomal variation	Euchromatic variants, large cytogenetically visible deletions, duplications, translocations, inversions, and aneuploidy	~5 Mb to entire chromosomes

[a]Size ranges quoted are indicative only of the scale of each type of rearrangement, and are not definitive.
[b]SINE, short interspersed element; LINE, long interspersed element; LTR, long terminal repeat; ERV, endogenous repeat virus.

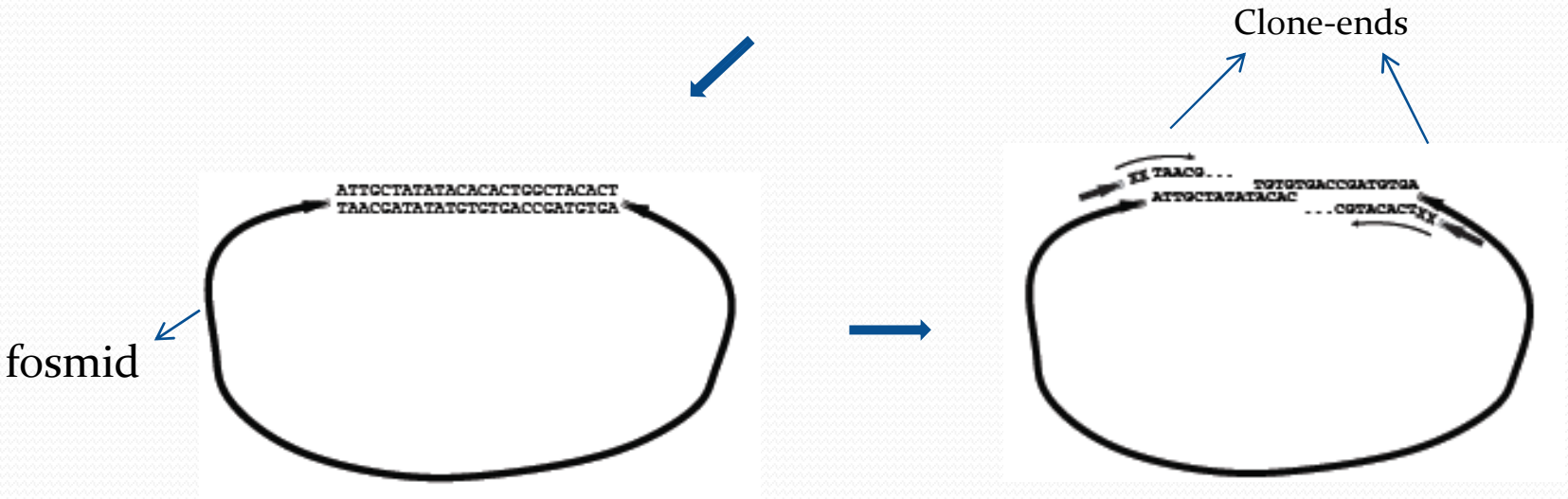
Annu.Rev.Genom.Human.Genet.2006 (7) 407-442

Types of Structural Variation



Paired-end sequence analysis

- DNA
- Fragments



Mapping the reads to the genome

1. Initial recruitment
2. Optimal realignment with quality rescoring
3. Determination of the best paired-end placements
4. Detection of rearrangements

1.Initial Recruitment

- Random genomic fosmid library from a female North American donor
- **2,298,774 end sequences**
- Alignment of all fosmid end sequences to the reference genome using **NCBI Megablast.**
  - Keep the **7 highest** scoring alignments (≥ 80 % sequence identity) for each end sequence
- **66,7%** of the clones are mapped (both ends having 1 or more alignment)

2.Optimal realignment with quality rescoring

- Global realignment with Needleman-Wunsch (no terminal gap penalties)
- Recalculation of percent identity using only base pairs with a minimum of phred Q 30
- New alignment score

  $$\text{Score} = \text{basepairs} * 2*\text{identity-}20[1\text{-identity}]]$$

- Low scoring alignments removed

- Global alignment:

```
G A A T T C A G T T A
G G A T * C * G * * A
```

- Semi-global alignment (ignore end spaces):

```
G A A T T C * A (G T T A)
G G A T * C G A (* * * *)
```

- Local alignment:

```
(*) G A A T T C A G (T T A)
(G) G A * T * C * G (A * *)
```

2.Optimal realignment with quality rescoring

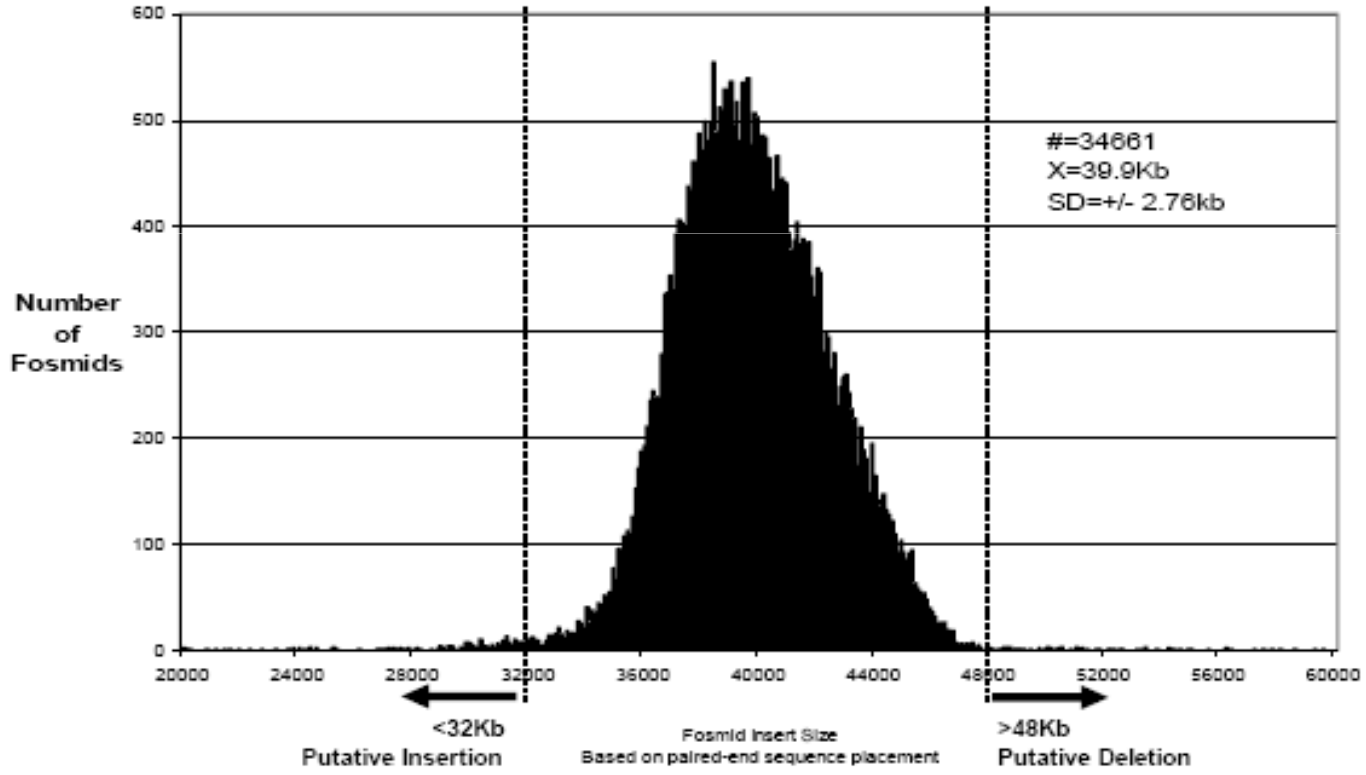
- Global realignment with Needleman-Wunsch (no terminal gap penalties)
- Recalculation of percent identity using only base pairs with a minimum of phred Q 30
- New alignment score

  $$\text{Score} = \text{basepairs} * 2*\text{identity-}20[1\text{-identity}]]$$

- Low scoring alignments are removed

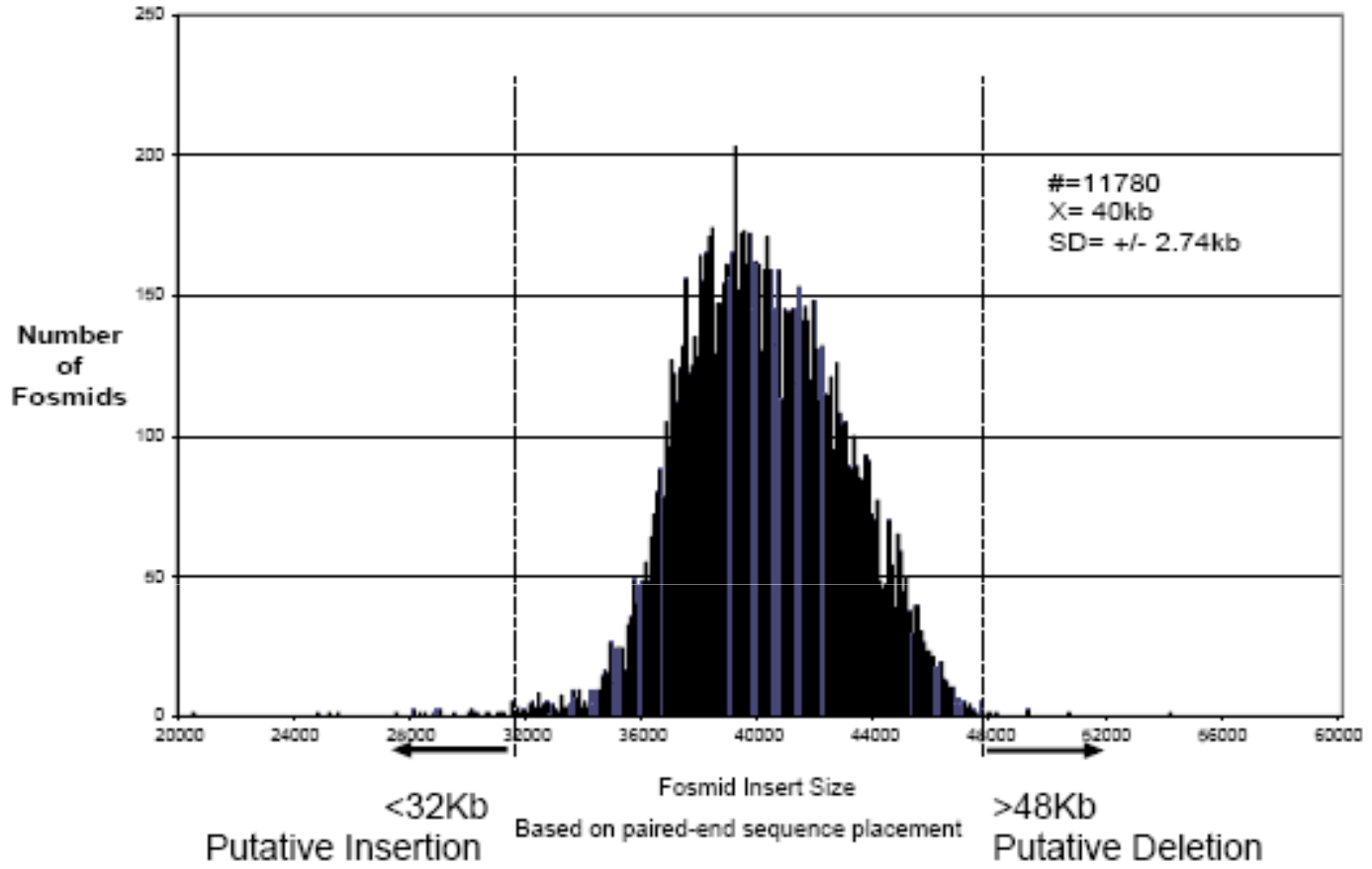
3.Determination of best paired-end replacements

Supplementary Figure 1 a) Fosmid Size Distribution in Chr7



Supplementary Figure 1 a. Fosmid size distribution. A total of 34,661 fosmids were mapped to finished human chromosome 7. The distance between the two ends was determined based on the coordinates within the human genome reference. The mean *in silico* size was 39.9 +/- 2.76 kb).

Supplementary Figure 1 b) Fosmid Size Distribution in Chr22



Supplementary Figure 1 b. Fosmid size distribution. The experiment was repeated for chromosome 22. A total of 11,780 fosmids were mapped with a mean *in silico* size of 40.2 +/- 2.74 kb).

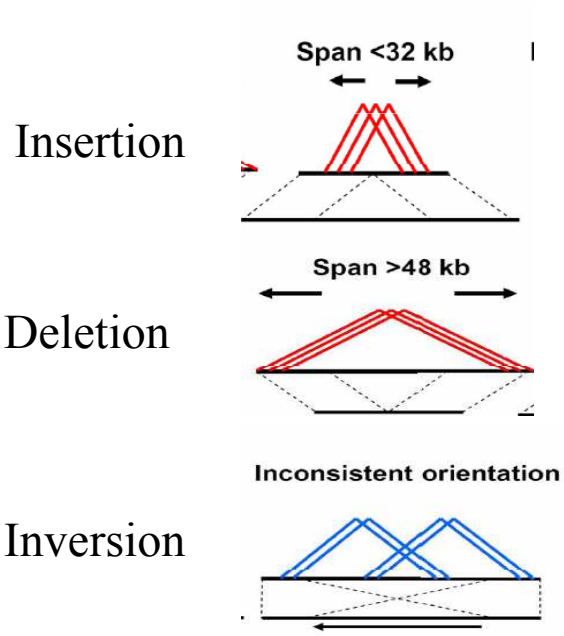
- Concordant insert size range 32-48 kb (3 sd)

Placement scores

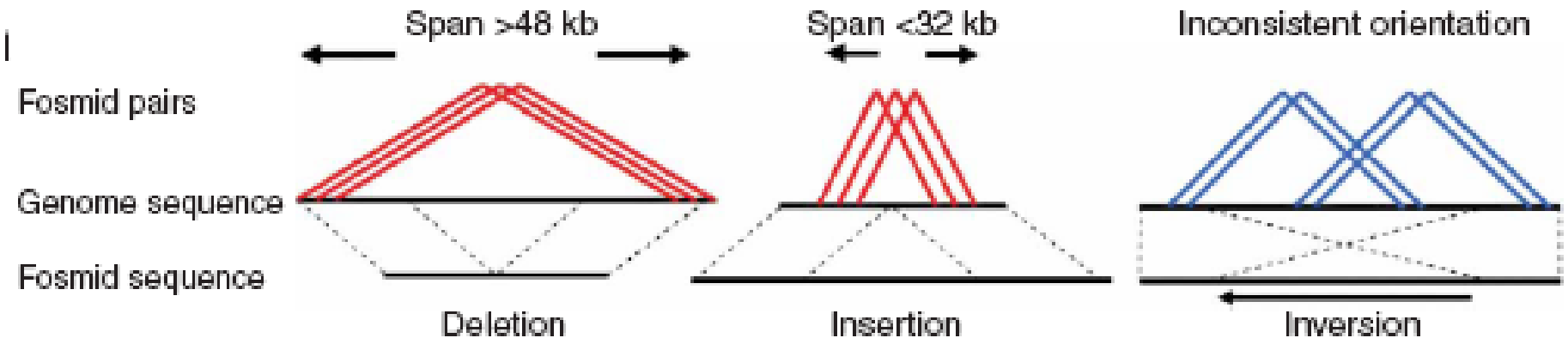
- From the set of all possible paired-end alignments
  - longest (+1 per end)
  - most identical ends(+1 per end)
  - level of identity(+2 per end if 99.5%)
  - adequate bp quality(+1 per end)
  - concordant alignments (favored)
    - proper insert size(32-48kb)(+2 per pair)
    - orientation(+1 per end)
  - discordant alignments
    - 99,5% identity
    - 400 bp in length, 150 bp unique sequence
    - remove the ones with ends located on diff. chromosomes.

Placement results

- 589,275 fosmids were mapped (8x coverage)
  - 99% concordant (583,550)
  - 1% discordant(3,189)
    - Small insert size (1638)
    - Large insert size (1531)
    - Incorrect orientation (698)



4. Detection of Rearrangements

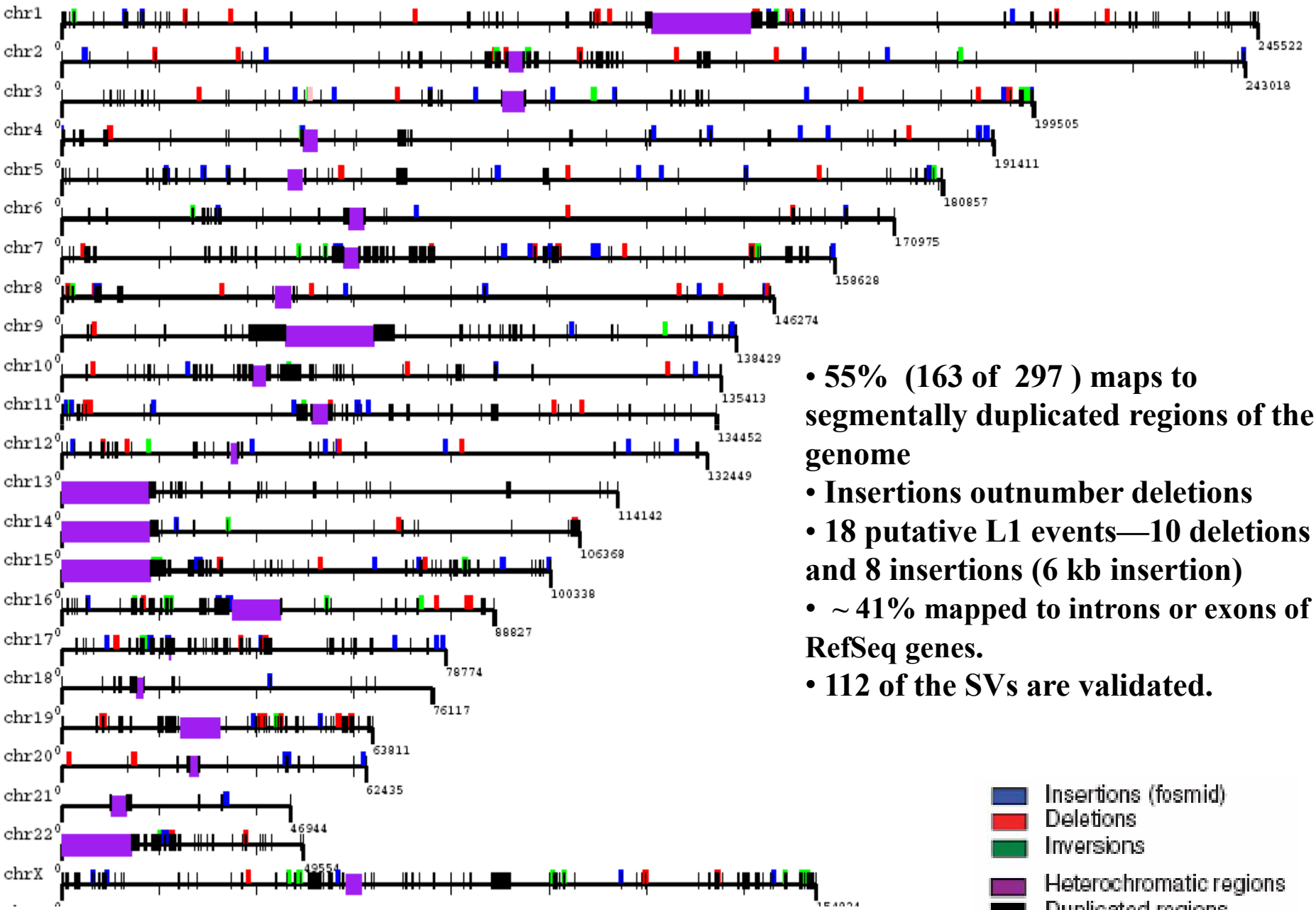


- Only when two or more independent fosmids supported the same rearrangement.
- Only paired-ends within 10 MB of each other on the same chromosome
  - Removed 27 regions where fosmid pairs span gaps.
  - Removed 88 sites where two fosmids mapped very near to each other (within 20 bp)

Results

- 297 sites of putative structural variation
  - 139 insertions
  - 102 deletions
  - 56 inversions - - 44 independent
- Size of the variations based on average discordance
  - largest deletion ~329 kb
  - largest insertion 36.5 kb, upper bound 40 kb
  - inversions from 13 kb to 1.9 Mb

Fine-Scale Structural Variation Map: (build35 vs. fosmids)



- 55% (163 of 297) maps to segmentally duplicated regions of the genome
- Insertions outnumber deletions
- 18 putative L1 events—10 deletions and 8 insertions (6 kb insertion)
- ~41% mapped to introns or exons of RefSeq genes.
- 112 of the SVs are validated.