nature genetics

ARTICLES

Personalized copy number and segmental duplication maps using next-generation sequencing

Can Alkan[1,2], Jeffrey M Kidd[1], Tomas Marques-Bonet[1,3], Gozde Aksay[1], Francesca Antonacci[1], Fereydoun Hormozdiari[4], Jacob O Kitzman[1], Carl Baker[1], Maika Malig[1], Onur Mutlu[5], S Cenk Sahinalp[4], Richard A Gibbs[6] & Evan E Eichler[1,2]

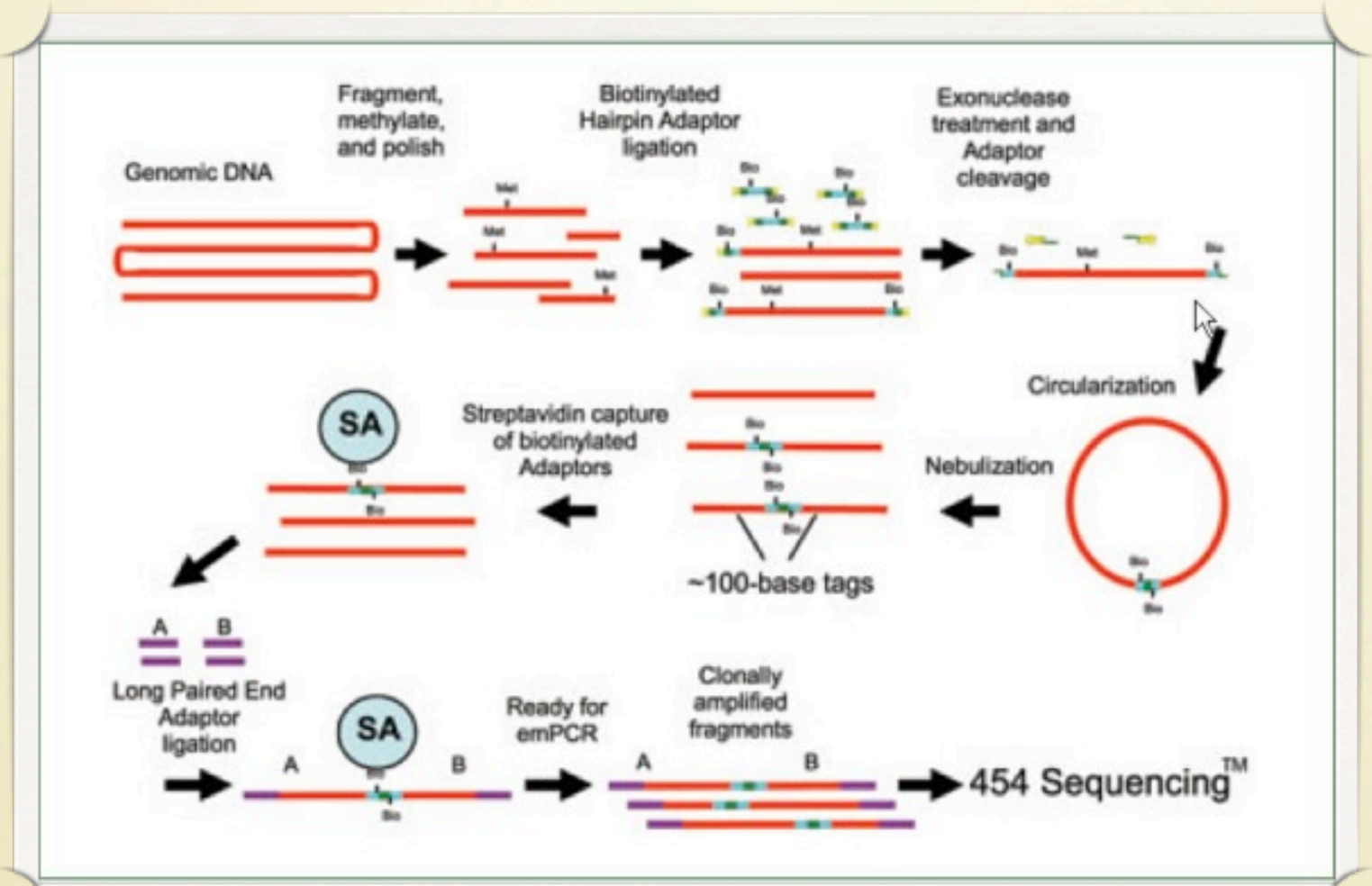
Despite their importance in gene innovation and phenotypic variation, duplicated regions have remained largely intractable owing to difficulties in accurately resolving their structure, copy number and sequence content. We present an algorithm (mrFAST) to comprehensively map next-generation sequence reads, which allows for the prediction of absolute copy-number variation of duplicated segments and genes. We examine three human genomes and experimentally validate genome-wide copy number differences. We estimate that, on average, 73–87 genes vary in copy number between any two individuals and find that these genic differences overwhelmingly correspond to segmental duplications (odds ratio = 135; $P < 2.2 \times 10^{-16}$). Our method can distinguish between different copies of highly identical genes, providing a more accurate assessment of gene content and insight into functional constraint without the limitations of array-based technology.

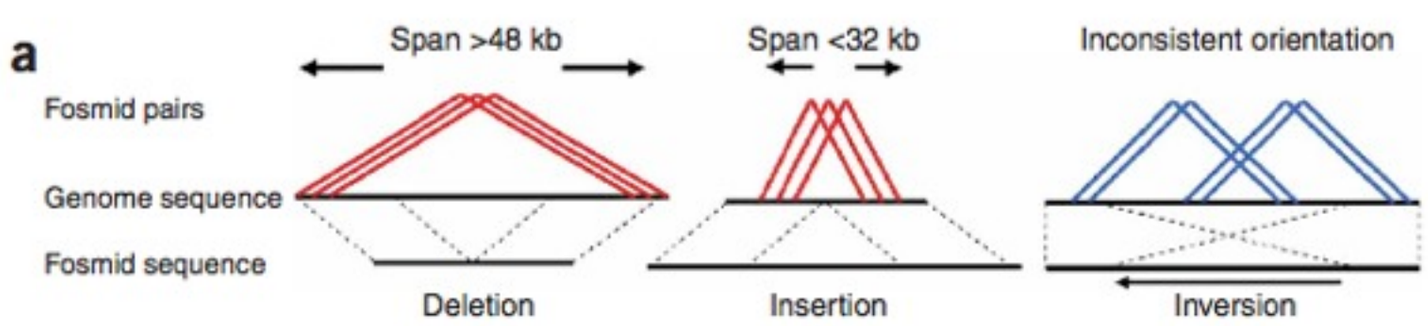
Andres Veidenberg
20.10.09

Motivation

- Detection of copy number variations in duplicated sequences is problematic
    - probe bias and limited power in copy number detection in microarray methods
- Next-generation sequencing is promising

Paired end mapping





Tuzun *et al.* (2005)

Depth of coverage

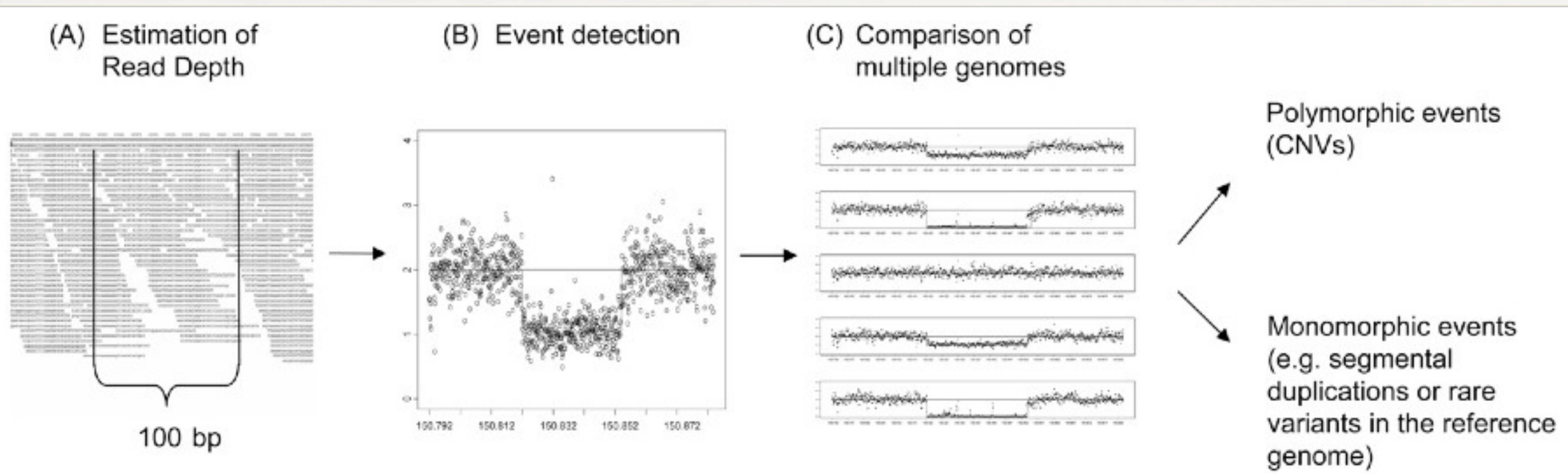


Figure 1. Pipeline for the detection of CNVs based on analysis of read depth (RD). (*A*) RD was determined by counting the start position of reads in nonoverlapping windows of 100 bp. (*B*) Events were detected using a custom CNV-calling algorithm, event-wise testing (EWT). (*C*) Each event was examined in multiple genomes in order to distinguish polymorphic events (CNVs) from the majority of events that were found to show a similar copy number change in all five genomes in this study (i.e., monomorphic events).

Yoon *et al.* (2009)

Results

- Develop a read-mapping algorithm (mrFAST) to rapidly assay copy number variation in complex and duplicated regions of human genome
- Create personal duplication maps of 3 individual genomes
- Validate results experimentally
- Find and analyze copy-number polymorphic gene regions

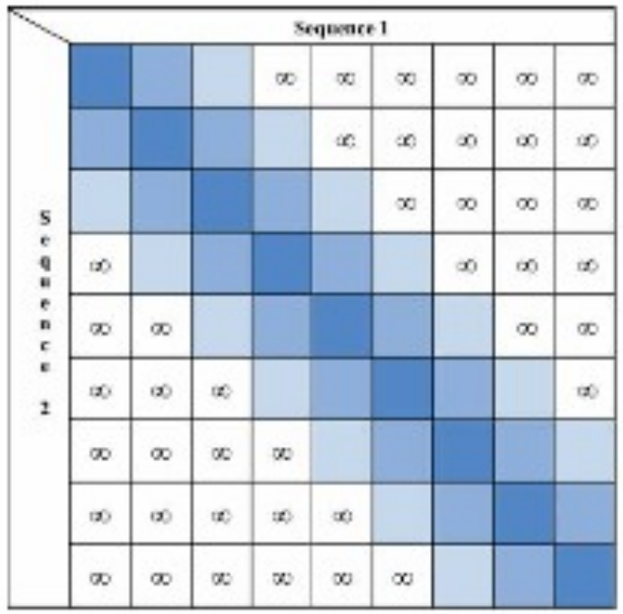
Algorithm

- Micro-read fast alignment search tool (mrFAST)
  - effectively map large amounts of short sequence data to the human genome assembly
  - calculate accurate read depth
  - return all possible single-nucleotide differences

Algorithm

- Optimized for efficient memory and CPU usage
    - Memory efficient hash to store read k-mers
    - Rapid seed extension algorithm
    - Optimized for Illumina/Solexa read properties
    - Uses SSE-2 instructions for CPU optimization

Algorithm optimization

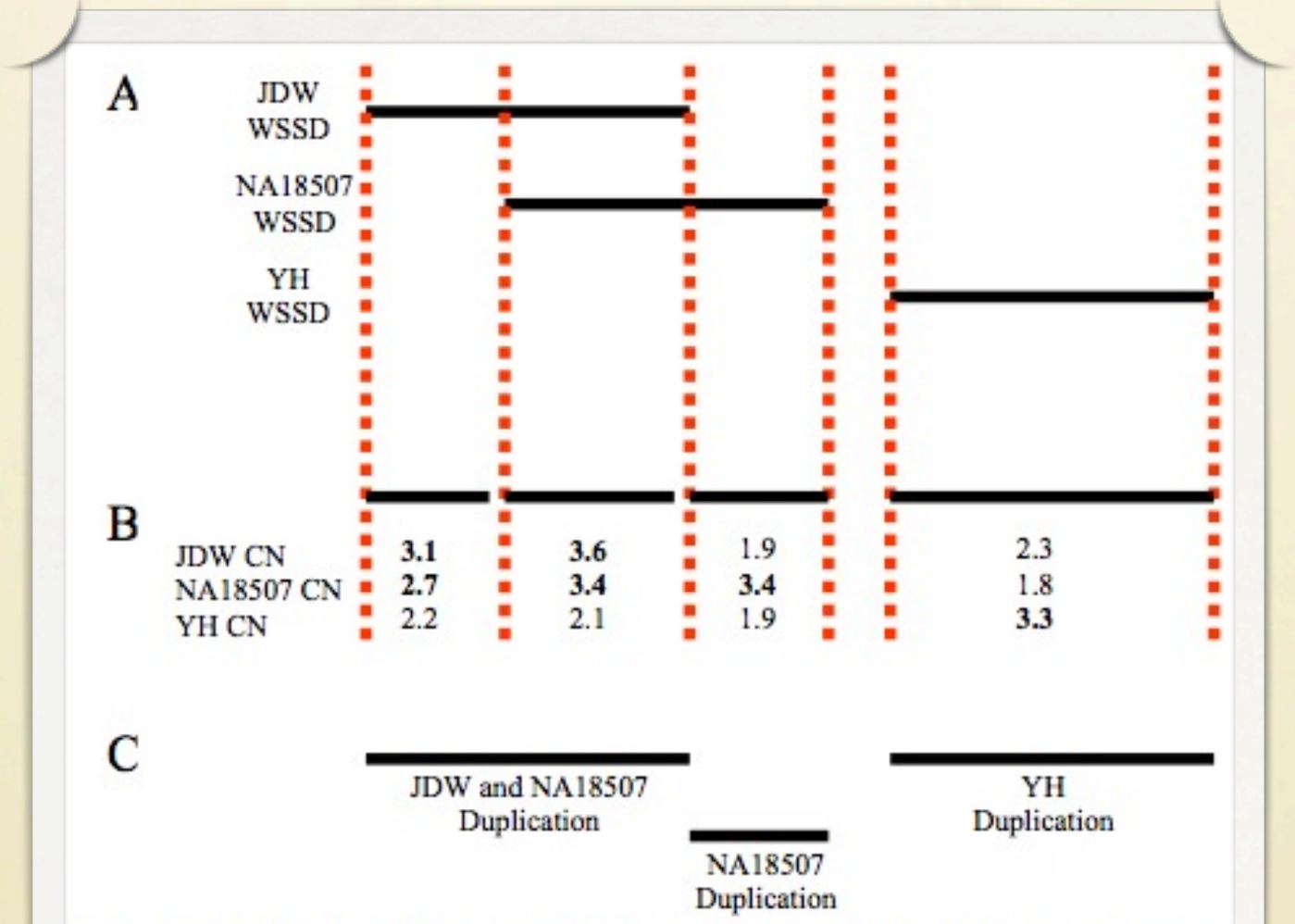


Supplementary Note Figure 1. Improvement of Levenshtein distance computation by Ukkonen's algorithm [10]. We need to calculate only the color-marked cells in the dynamic programming matrix if the maximum allowed number of gaps is bounded by a small value *t*. This figure shows the diagonals to be calculated when $t = 2$.

Personal duplication maps

- 3 genomes: JWD, NA18507, YH
- Uses read depth for calculating copy number
  - Calibration with known duplications
    - good correlation with verified copy number
  - accurately predicted duplication boundaries
    - >90% of segmented duplications (if 20x coverage)

Calculating duplications



Supplementary Note Figure 6. Refining the duplication predictions. Before this analysis we created a refined, non-redundant set of duplication predictions. We began with the intervals identified as duplicated in each sample by our standard WSSD heuristics (A). We then split the predicted intervals from the three samples into non-overlapping segments. We calculated the median diploid copy number of each segment for each of the three samples (B). We reclassified a segment as being duplicated in a sample if the median copy number was greater than 2.5. Finally, we merged together

Personal duplication maps

- Constructed duplication maps for each of 3 genomes and predicted absolute copy number (CN) for each duplication interval of >20kb length
  - defined 725 duplication intervals (84,7Mb)
  - 97% of large segmental duplications are shared

Experimental validation

- Array-based comparative genomic hybridization (arrayCGH)
    - individual-specific duplications are near shared duplications
    - shared duplication show greatest CN variation
- Fluorescence in-situ hybridization (FISH)
    - CN difference between YH and NA18507
    - FISH results highly consistent with mrFAST

CN polymorphic genes

- 68 gene families validated as being CN variable
  - complement factor H, defensin, CCL3L1 genes
  - higher CN in evolving genes
  - found 300 distinct paralogous sequences
  - calculated CN for 17610 RefSeq transcripts
  - 97% of CN variable genes mapped in segmental duplications

Discussion

- Designed mrFAST for accurate mapping of short sequences
- Provide one of first comprehensive estimates of absolute CN differences in 3 individual genomes
- Genes with most variable copy number:
    - embedded in segmental duplications (mostly tandem changes in copy)
        - correspond to rapidly evolving gene families

THANK YOU!