

# Statistical modeling in molecular medicine: **genomics**

**Anna Gambin**

Institute of Informatics,  
University of Warsaw

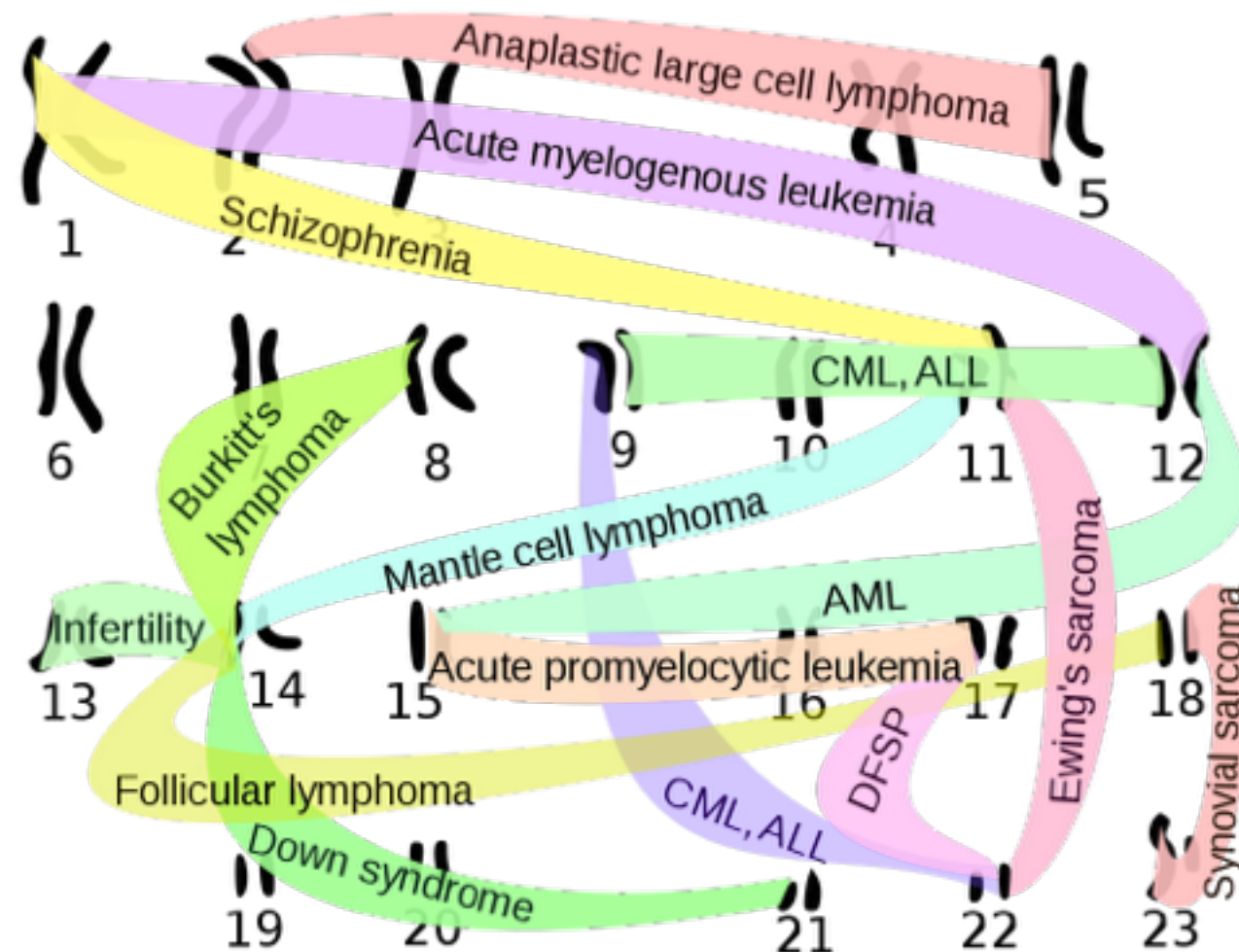


# outline

- the **NAHR mechanism**, CNVs, genomic disorders
- what drives **NAHRs** to specific genomic regions?
- genomic regions prone to **instability**
- clinical data from **array CGH — BCM database**
- breakpoint identifications by **Hidden Markov Model**
- **molecular** validation for **LINE mediation** hypothesis
- conclusions

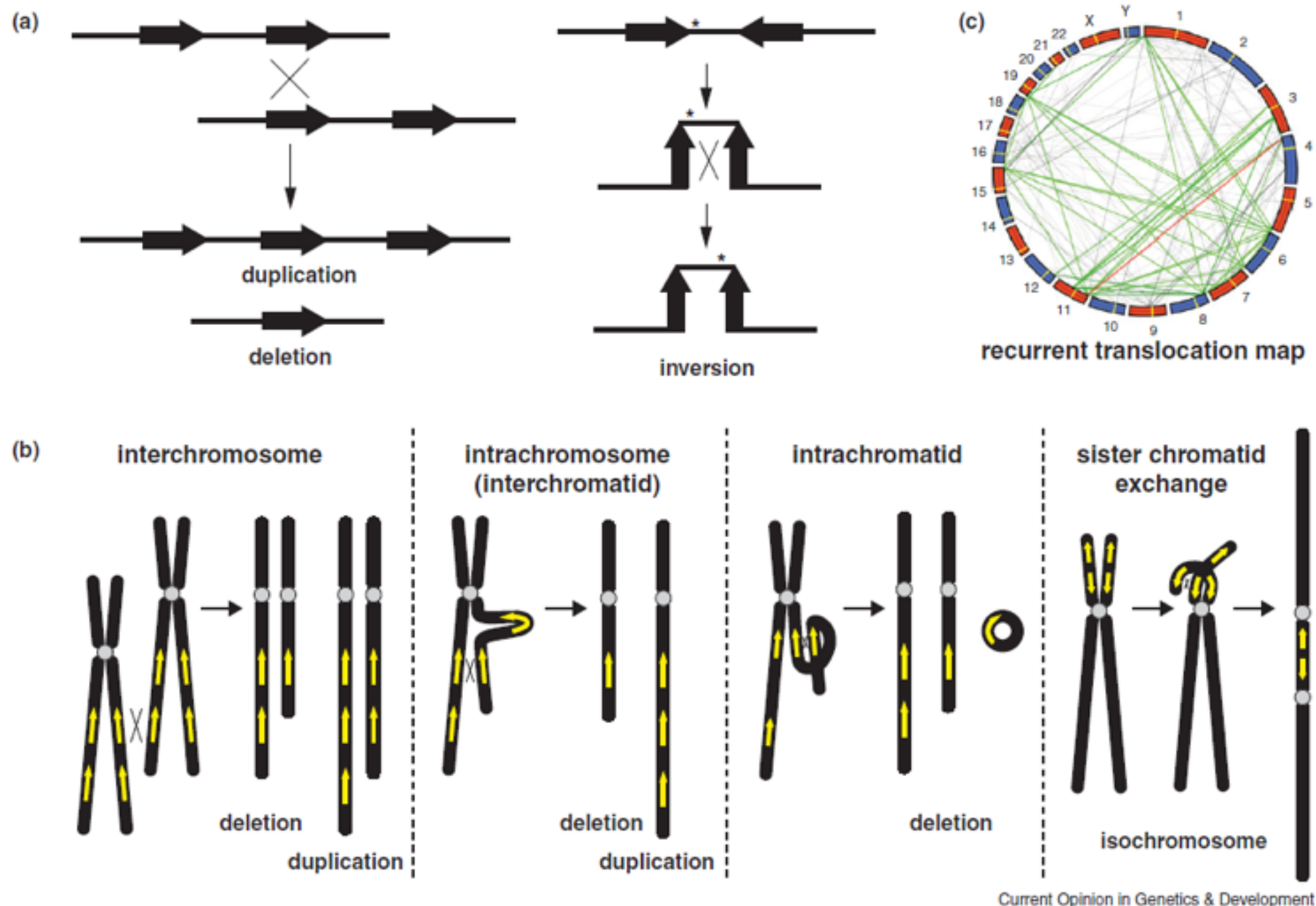
# genomic disorders

- higher-order genomic architectural features can lead to a susceptibility to DNA rearrangements (called **genomic disorders**); frequent cause of diseases in humans
- mechanism causing disorders: variation in copy number of dosage sensitive genes



# NAHR

Non-allelic homologous recombination = recombination which occurs between similar fragments of DNA which are not alleles.

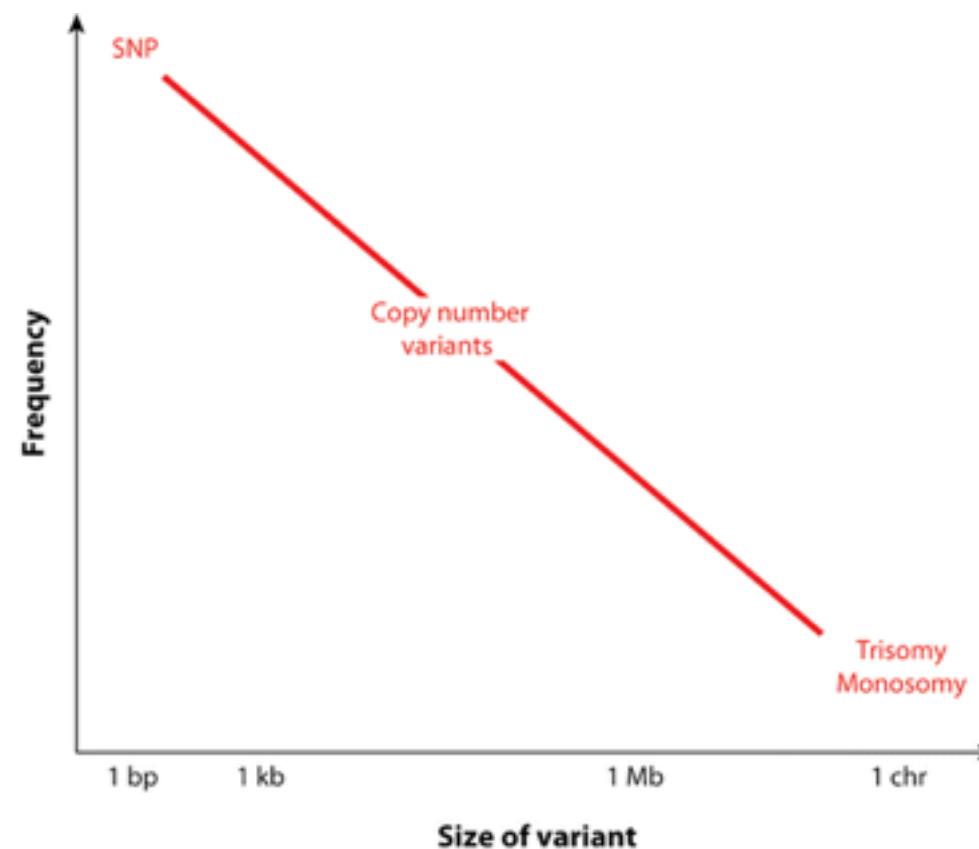




# CNVs in disorders and cancer

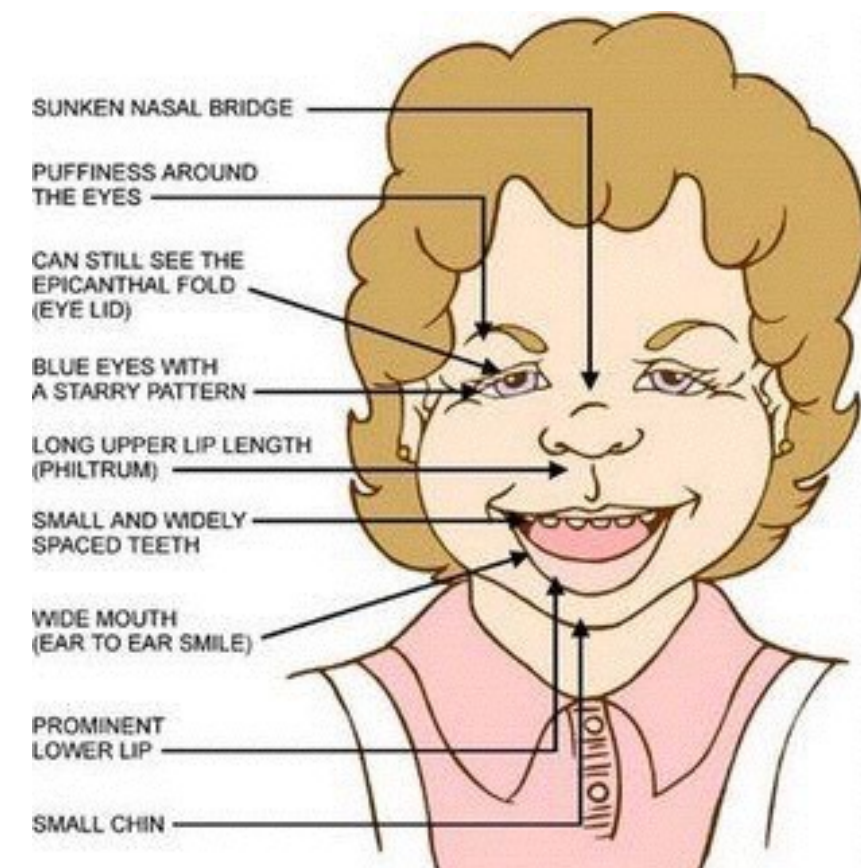
NAHR = one of the most important mechanisms causing formation of **Copy Number Variants** (CNVs).

CNVs: responsible for wide range of genetic disorders, both mild and severe.



# Known NAHR-associated syndromes

- known **recurrent rearrangements**: include the same interval occurring in unrelated individuals
  - only two known syndromes associated with **inversion** (Hunter syndrome, Hemophilia type A)
  - tens of syndromes associated with **deletion/reciprocal duplication**:  
DiGeorge, Potocki-Lupski, Smith-Magenis,...
  - usually deletions are much more serious than duplications
- “too few is worse than too many”**



[geneticsf.labanca.net](http://geneticsf.labanca.net)

# **first suspected: Low-Copy Repeats**

- LCRs also known as **Segmental Duplications**,
- DNA fragments  $> 1$  kb and  $> 90\%$  DNA sequence identity
- working hypothesis: LCRs  $>$  around 10 kb and  $>$  around 95% sequence identity can lead to local genomic instability
- may stimulate and/or mediate constitutional (both recurrent and nonrecurrent), evolutionary, and somatic genomic rearrangements
- may cause **Non Allelic Homologous Recombination (NAHR)**

# IP-LCRs - AD 2012

RESEARCH ARTICLE

Human Mutation

## Inverted Low-Copy Repeats and Genome Instability— A Genome-Wide Analysis



Piotr Dittwald,<sup>1,2†</sup> Tomasz Gambin,<sup>3\*†</sup> Claudia Gonzaga-Jauregui,<sup>4†</sup> Claudia M.B. Carvalho,<sup>4</sup> James R. Lupski,<sup>4-6</sup>  
Paweł Stankiewicz,<sup>4\*</sup> and Anna Gambin<sup>1,7</sup>

# DP-LCRs - AD 2013

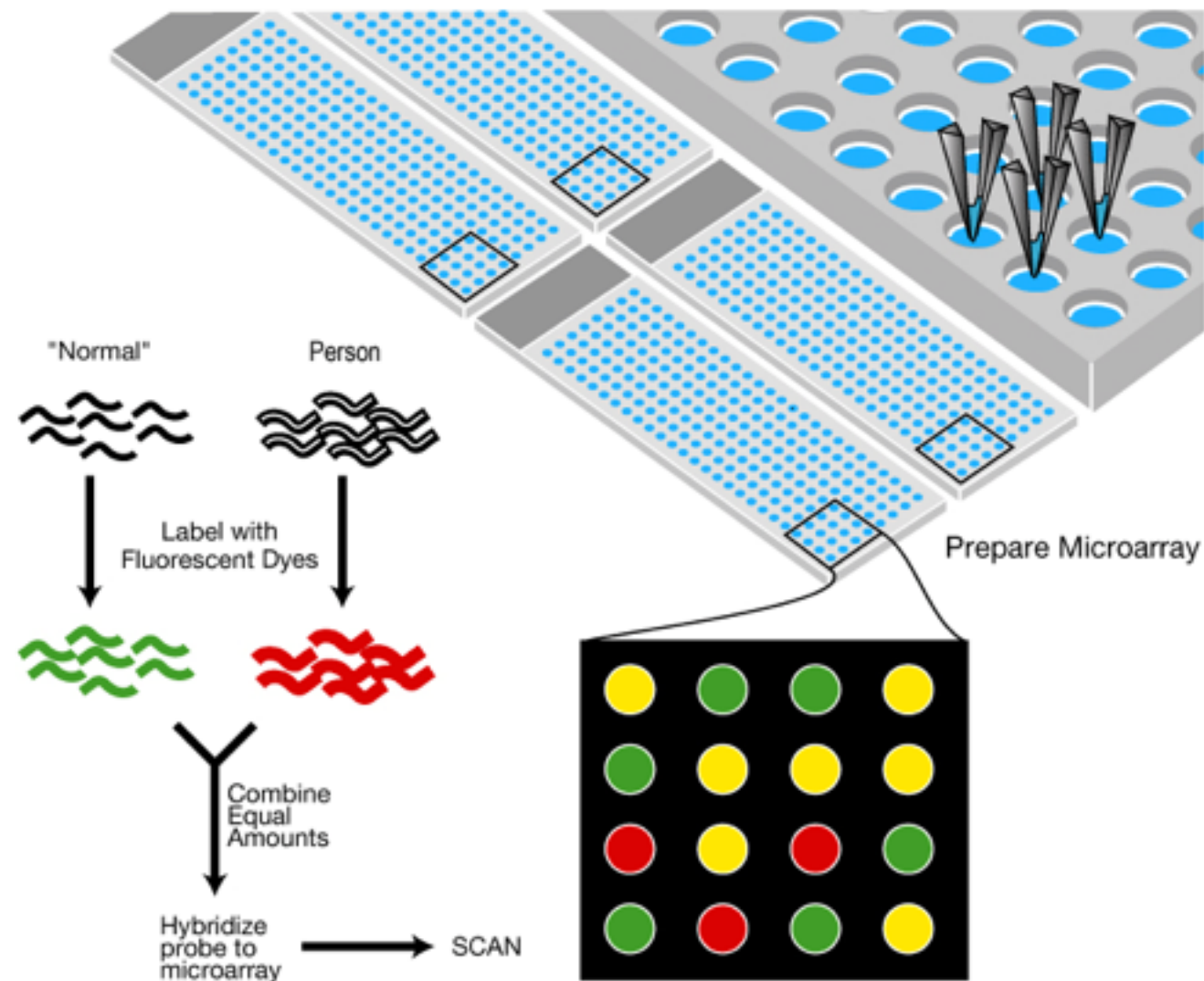


NAHR-mediated copy-number variants in a clinical population:  
mechanistic insights into both genomic disorders and Mendelizing  
traits

Piotr Dittwald, Tomasz Gambin, Przemysław Szafranski, et al.

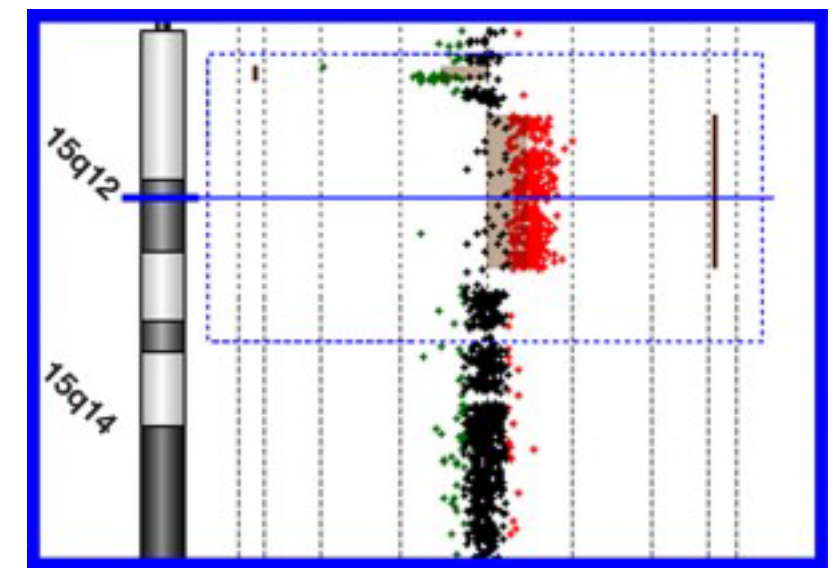
# model?

# chromosomal microarray analysis



source: [atlantichealth.dnadirect.com](http://atlantichealth.dnadirect.com)

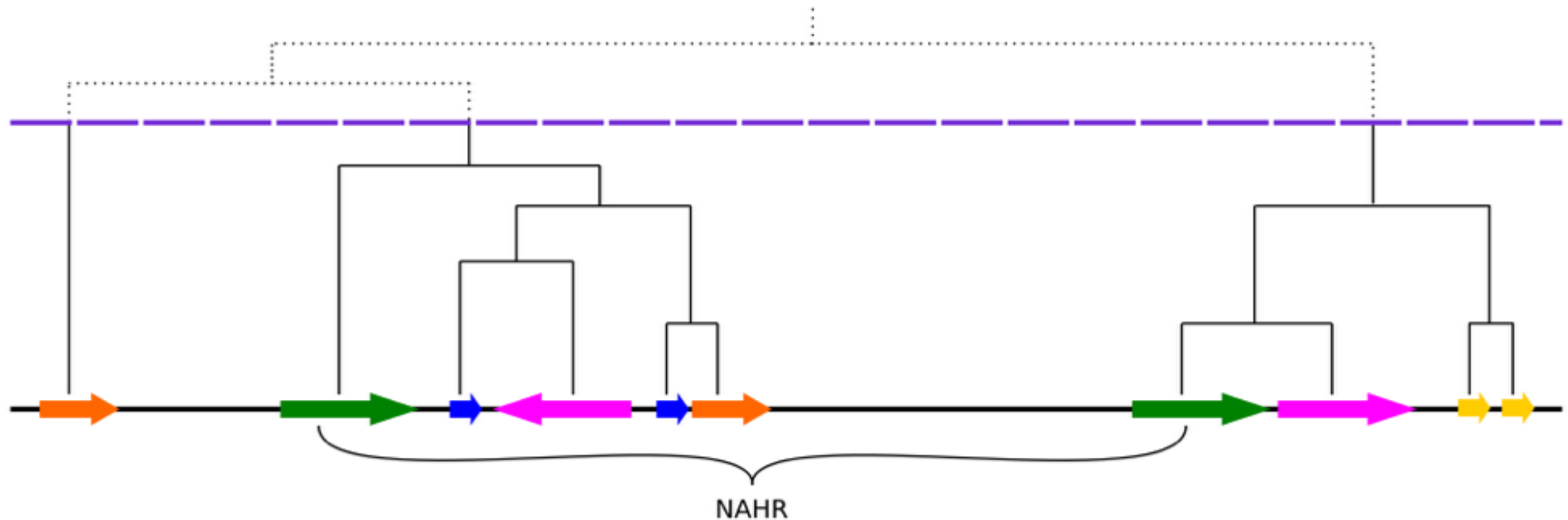
**BCM**  
Baylor College of Medicine



source: [childrenshospitalblog.org](http://childrenshospitalblog.org)



# LCRs cluster

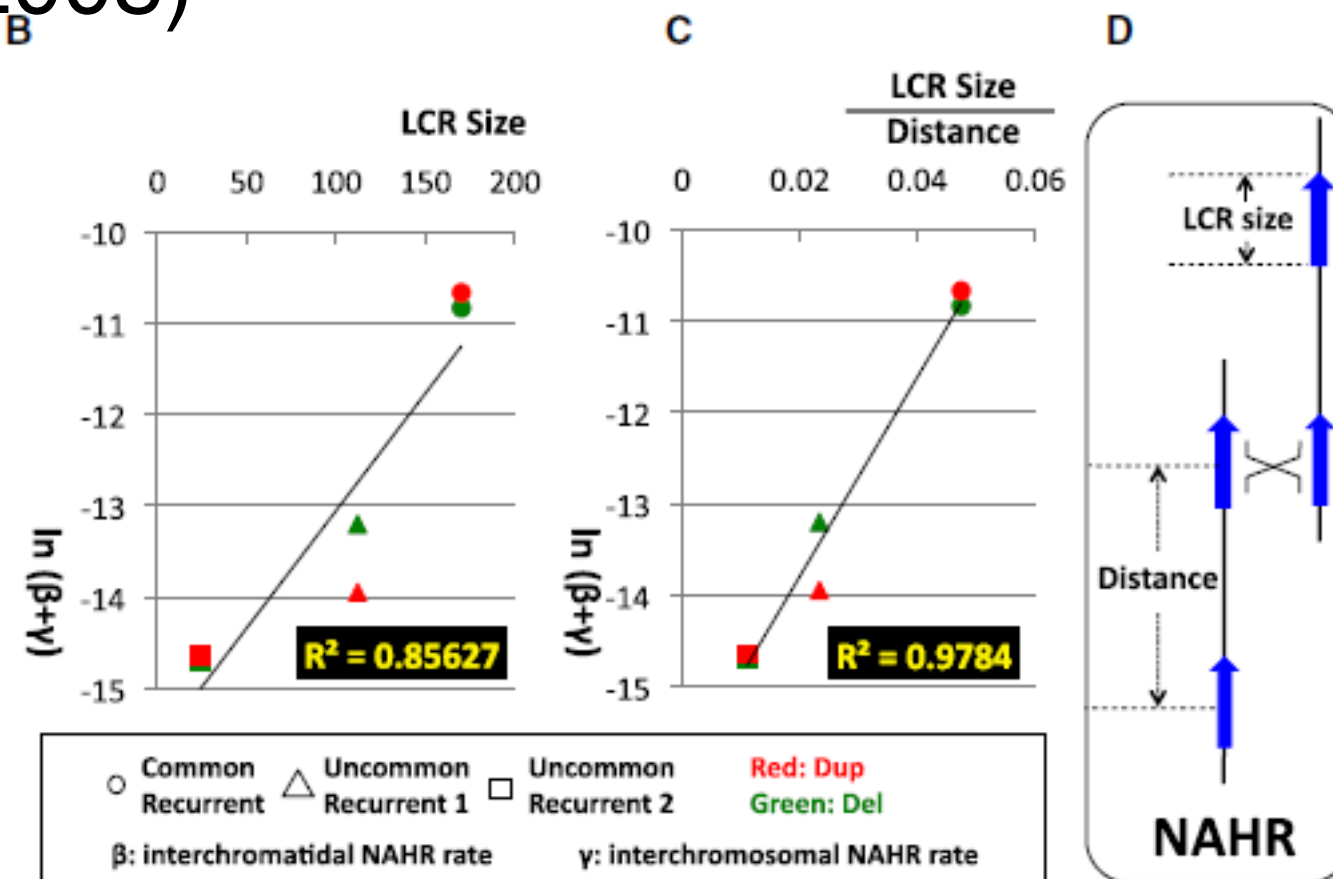


- arrows indicate LCR elements and their orientation,
- the same colour represents a pair of LCRs
- hierarchical clustering tree is depicted
- oriented paralogous LCRs within the clusters (green) potentially **mediate NAHR event**



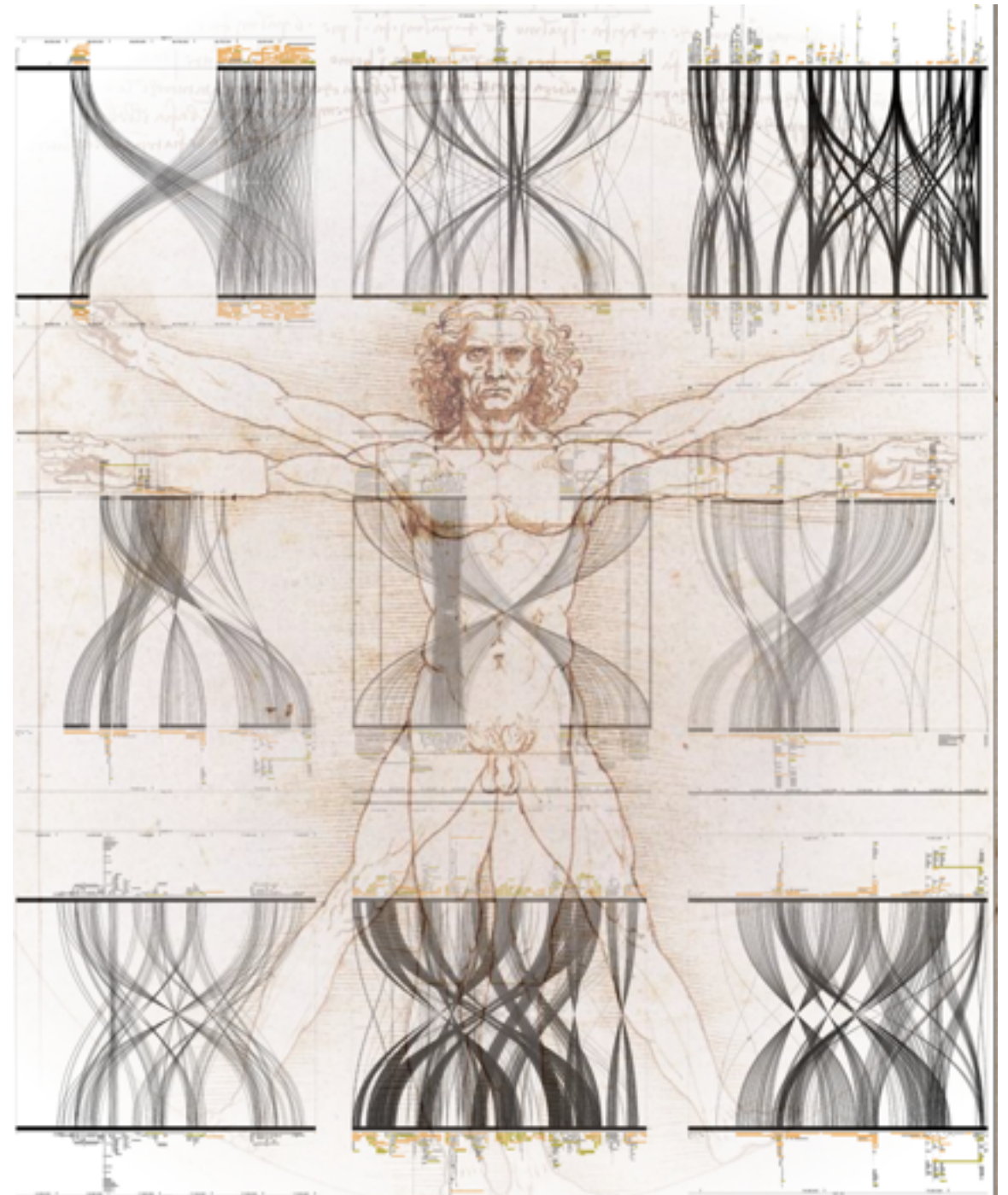
# Genomic features correlating with NAHR frequency

- LCR size, LCR size/distance (Liu et al. 2011)
- frequency of motif 5'- CCNCCNTNNCCNC- 3' the histone methyltransferase PRDM9 binding site (Myers et al. 2008)



# Poisson regression: considered parameters

- **DP-LCR:** average lengths, distances, fraction matching, presence of the 13-mer recombination hotspot motif 5'-CCNCCNTNNCCNC-3'
- **LCR clusters:** number of LCRs within the cluster, average length of LCRs, concentration of recombination hotspot motif



# Findings on genome-scale

- **DP-LCR:** length of homology (weak association,  $p=1.68e-01$ ); distance between homologous pair; inverse relationship - the further the DP-LCR are apart, the less frequent ( $p=2.19e-04$ ); percent DNA sequence identity ( $p=8.18e-05$ ).
- **LCR clusters:** the maximum length of homology among LCRs within a cluster ( $p=4.62e-02$ ); GC content within the cluster ( $p=7.04e-03$ ); occurrences of recombination hot spot motif among LCRs assigned to the cluster ( $p=6.79e-03$ ).



# Findings on genome-scale

Feature of DP-LCRs	Comparison of DP-LCRs flanking active NAHR hot spots vs. DP-LCRs flanking inactive cold spots ( <i>p</i> -values from Mann-Whitney-Wilcoxon test)			Correlation/regression of DP-LCRs feature and frequency of de novo deletions; DP-LCRs flanking reliable recurrent changes, i.e. genomic regions for which we detected at least three recurrent de novo deletions, were considered
	Feature is greater in DP-LCRs flanking active NAHR hot spots, i.e. regions for which we detected at least one de novo deletion	Feature is greater in DP-LCRs flanking inactive NAHR cold spots, i.e. regions for which we did not detect any de novo deletion	Spearman rank correlation coefficients and <i>p</i> -values	Poisson regression coefficients and <i>p</i> -values
Length of homology of paralogous DP-LCRs	<i>(p</i> =1.86e-01)		0.29 ( <i>p</i> =1.68e-01)	
Distance between paralogous DP-LCRs	** ( <i>p</i> =1.18e-03)		-0.69*** ( <i>p</i> =2.19e-04)	
Length of homology divided by distance between paralogous DP-LCRs		** ( <i>p</i> =7.64e-03)	0.60** ( <i>p</i> =2.30e-03)	4.37e+01** ( <i>p</i> =1.08e-03)
Fraction matching (percent identity) of paralogous DP-LCRs	<i>(p</i> =2.69e-01)		0.73*** ( <i>p</i> =8.18e-05)	29.72* ( <i>p</i> =1.51e-02)
Mean GC content of paralogous DP-LCRs	*** ( <i>p</i> =7.53e-06)		-0.02 ( <i>p</i> =9.05e-01)	
Number of occurrences of the 13-mer recombination motif in the pair of DP-LCRs combined	*** ( <i>p</i> =7.06e-05)		0.33 ( <i>p</i> =1.17e-01)	
Average density of the 13-mer recombination motif in the pair of DP-LCRs combined	*** ( <i>p</i> =2.57e-06)		0.04 ( <i>p</i> =8.55e-01)	

# Findings on genome-scale

Feature of LCR cluster	Comparison of LCR clusters flanking active NAHR hot spots vs. LCR clusters flanking inactive cold spots ( <i>p</i> -values from Mann-Whitney-Wilcoxon test)			Correlation/regression of LCR cluster's feature and frequency of de novo deletions; clusters flanking reliable recurrent changes, <i>i.e.</i> genomic regions for which we detected at least three recurrent de novo deletions, were considered
	Feature is greater in LCR clusters flanking active NAHR hot spots	Feature is greater in LCR clusters flanking inactive NAHR cold spots	Spearman rank correlation coefficients and <i>p</i> -values	Poisson regression coefficients and <i>p</i> -values
GC content within the cluster	***( <i>p</i> =1.11e-04)		0.54** ( <i>p</i> =7.04e-03)	2.71e+01*** ( <i>p</i> =1.34e-25)
Minimum length of homology among LCRs within the cluster		( <i>p</i> =9.96e-01)	0.12 ( <i>p</i> =5.74e-01)	
First quartile of the length of homology among LCRs within the cluster		( <i>p</i> =8.43e-01)	0.02 ( <i>p</i> =9.26e-01)	
Median length of homology among LCRs within the cluster		( <i>p</i> =5.57e-01)	0.23 ( <i>p</i> =2.71e-01)	
Third quartile of the length of homology among LCRs within the cluster		( <i>p</i> =4.81e-01)	0.15 ( <i>p</i> =4.73e-01)	
Maximum length of homology among LCRs within the cluster	( <i>p</i> =1.41e-01)		0.41* ( <i>p</i> =4.62e-02)	1.4e-05*** ( <i>p</i> =5.43e-11)
Total number of occurrences of the 13-mer recombination hot spot motif in the cluster		( <i>p</i> =2.7e-01)	0.51* ( <i>p</i> =1.17e-02)	
Minimum number of occurrences of the 13-mer recombination hot spot motif among LCRs within the cluster		( <i>p</i> =2.25e-01)	0.00 ( <i>p</i> =1.00)	
First quartile of the number of occurrences of the 13-mer recombination hot spot motif among LCRs within the cluster		( <i>p</i> =7.59e-01)	0.01 ( <i>p</i> =9.33e-01)	
Median number of occurrences of the 13-mer recombination hot spot motif among LCRs within the cluster	( <i>p</i> =7.1e-02)		0.48* ( <i>p</i> =2.01e-02)	4.45e-01** ( <i>p</i> =3.5e-03)

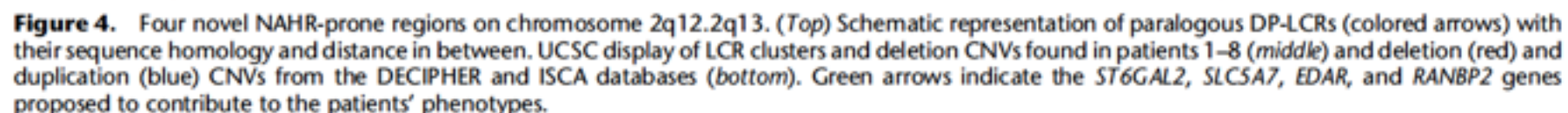


LCR length: 18.4 kb; seq ident 98.98%

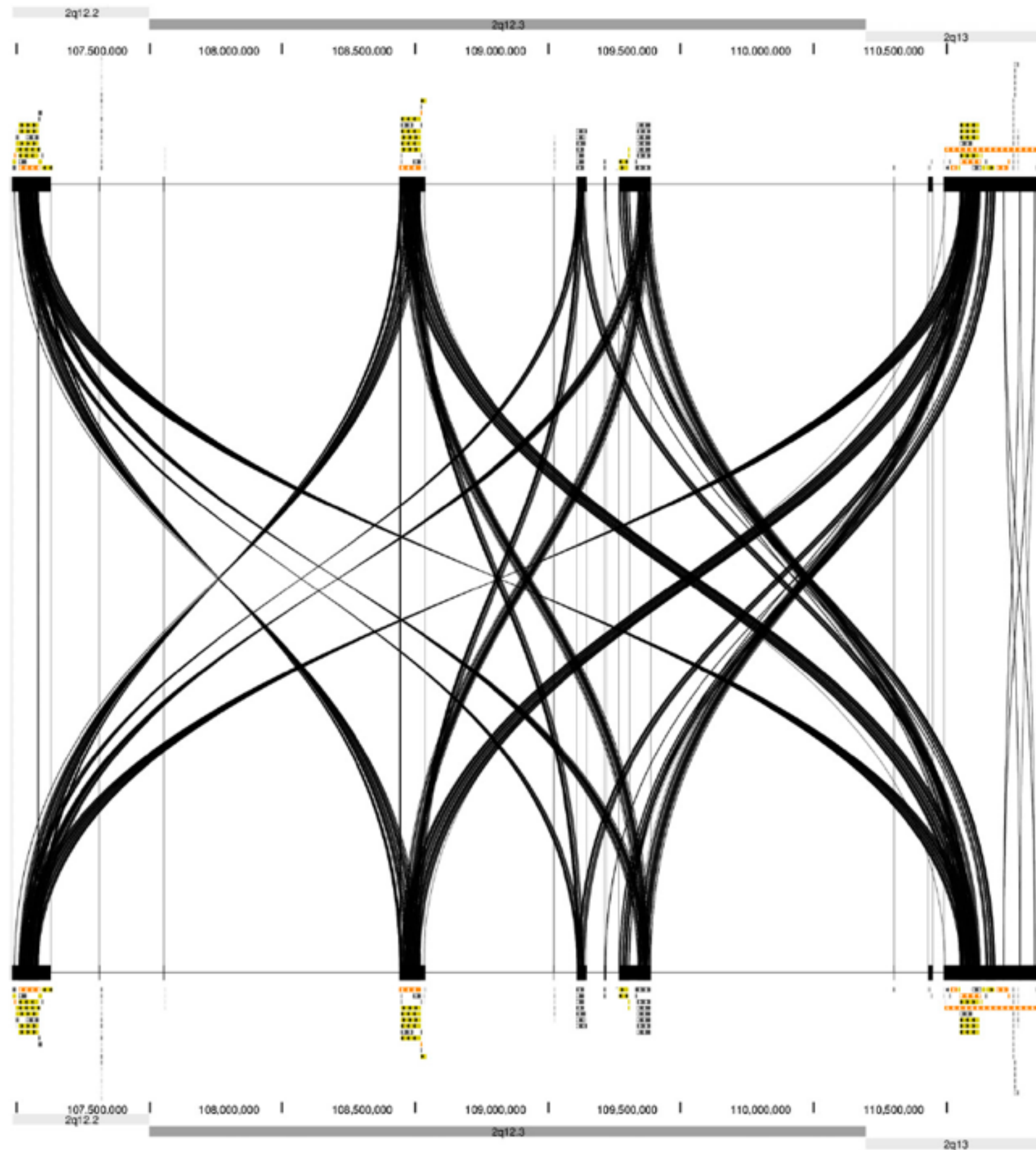
LCR length: 25.1 kb; seq ident 97.62%

LCR length: 29.2 kb; seq ident 97.53%

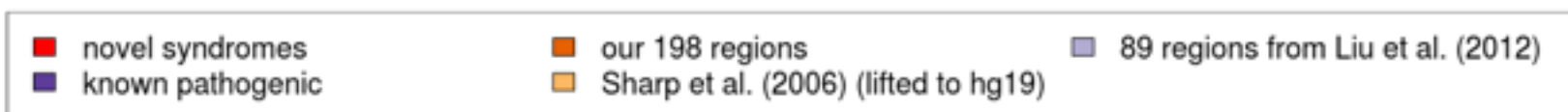
LCR length: 72.1 kb; seq ident 98.33%







**Figure 5.** DNA sequence homology between four LCR clusters in the 2q12.2q13 region (chr2:106,985,338-110,870,754) for paralogous subunits larger than 1 kb in size (hg19). (*Top and bottom*) UCSC Segmental Duplications (segdup) track representing the 2q12.2q13 region. (*Middle*) Results of *Miropeats* program analysis among all four clusters.



# more NAHR mediators !!!

Usually thought to occur between a pair of homologous (long) LCRs (up to 300 kb in size) but...

Campbell et al. *BMC Biology* 2014, 12:74  
<http://www.biomedcentral.com/1741-7007/12/74>



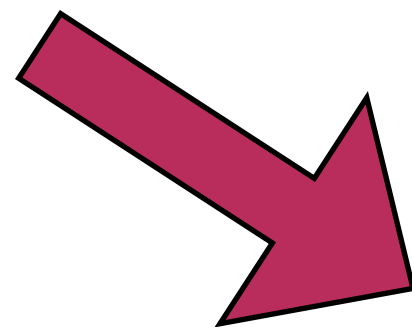
RESEARCH ARTICLE

Open Access

Human endogenous retroviral elements promote genome instability via non-allelic homologous recombination

Ian M Campbell<sup>1†</sup>, Tomasz Gambin<sup>1†</sup>, Piotr Dittwald<sup>2,3†</sup>, Christine R Beck<sup>1</sup>, Andrey Shuvarikov<sup>4</sup>, Patricia Hixson<sup>1</sup>, Ankita Patel<sup>1</sup>, Anna Gambin<sup>2,5</sup>, Chad A Shaw<sup>1</sup>, Jill A Rosenfeld<sup>1,4</sup> and Paweł Stankiewicz<sup>1\*</sup>

**AD 2014**



lower boundary on the length of the homologous region which is capable of mediating NAHRs **might be as low as few kb !!!**

# next step: LINEs





Transposable elements: short (usually  $< 10\text{kb}$ ) sequences of mobile, self-replicating DNA;

source of repeating sequences in most genomes;

main cause of genomic self-similarity (in addition to Low Copy Repeats aka Segmental Duplications).

## Types of transposable elements in the human genome

Long INterspersed Elements (LINEs): 500 000 copies, 21% of the human genome.

Element	Transposition	Structure	Length	Copy number	Fraction of genome
LINEs	Autonomous		1–5 kb	20,000–40,000	21%
SINEs	Nonautonomous		100–300 bp	1,500,000	13%
DNA transposons	Autonomous		2–3 kb	300,000	3%
	Nonautonomous		80–3000 bp		

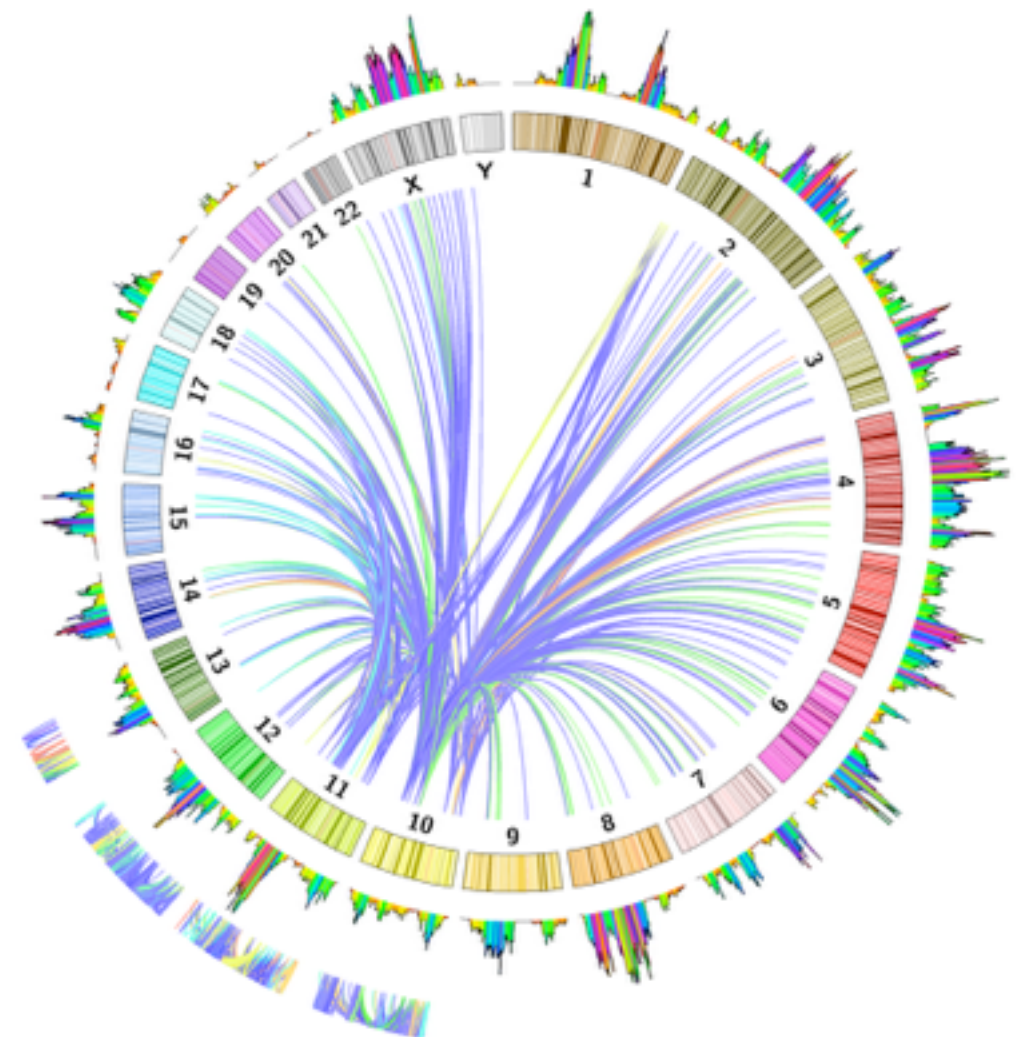
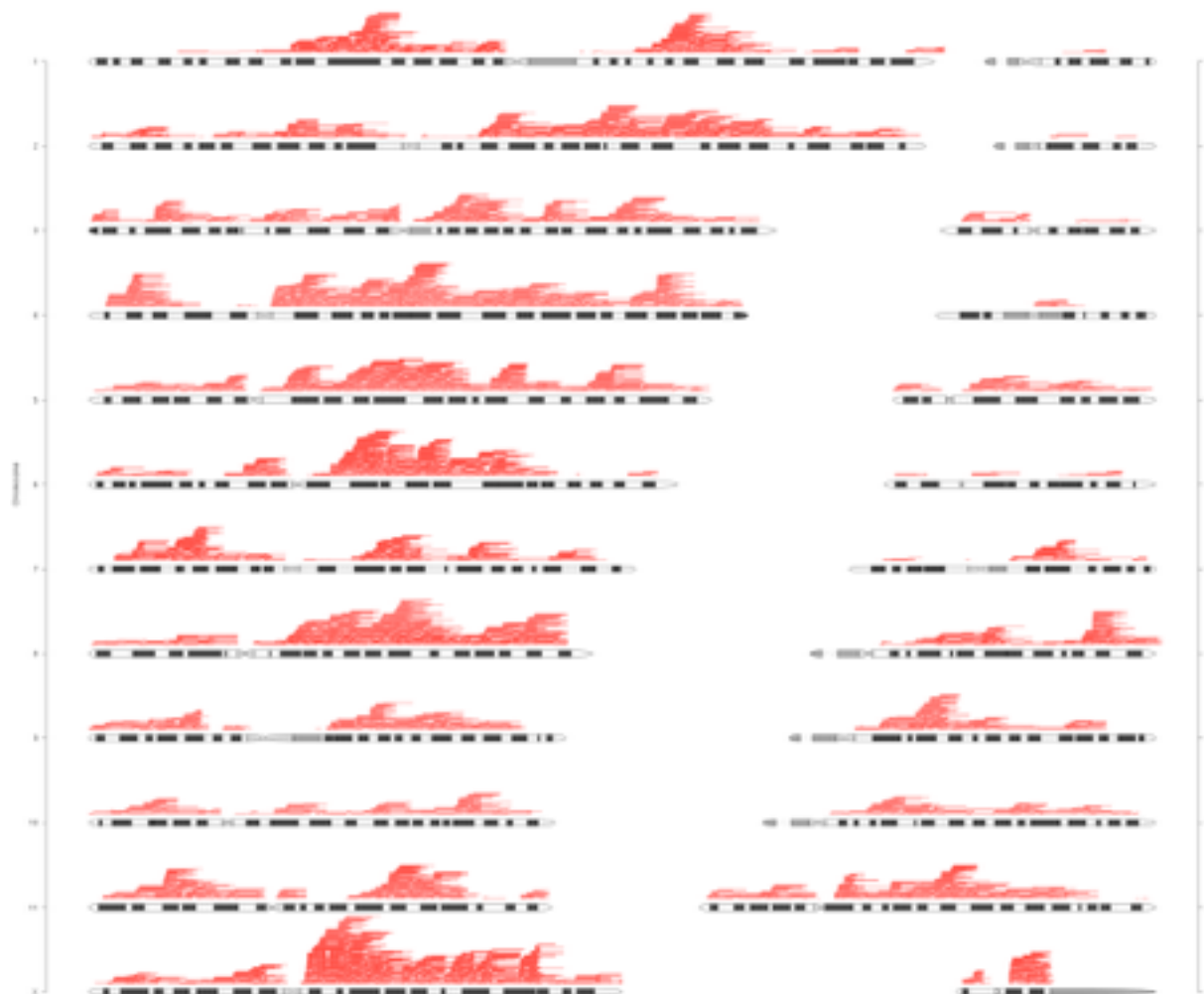


# huge instability risk

**determine LINE pairs in HG19 = mediators of NAHR**

- share a homology over more than 4kb of their length (as detected by BLAST);
- the identity over homologous region had to be over 95 %;
- on the same chromosome and spanned over a region between 10kb and 10Mb;

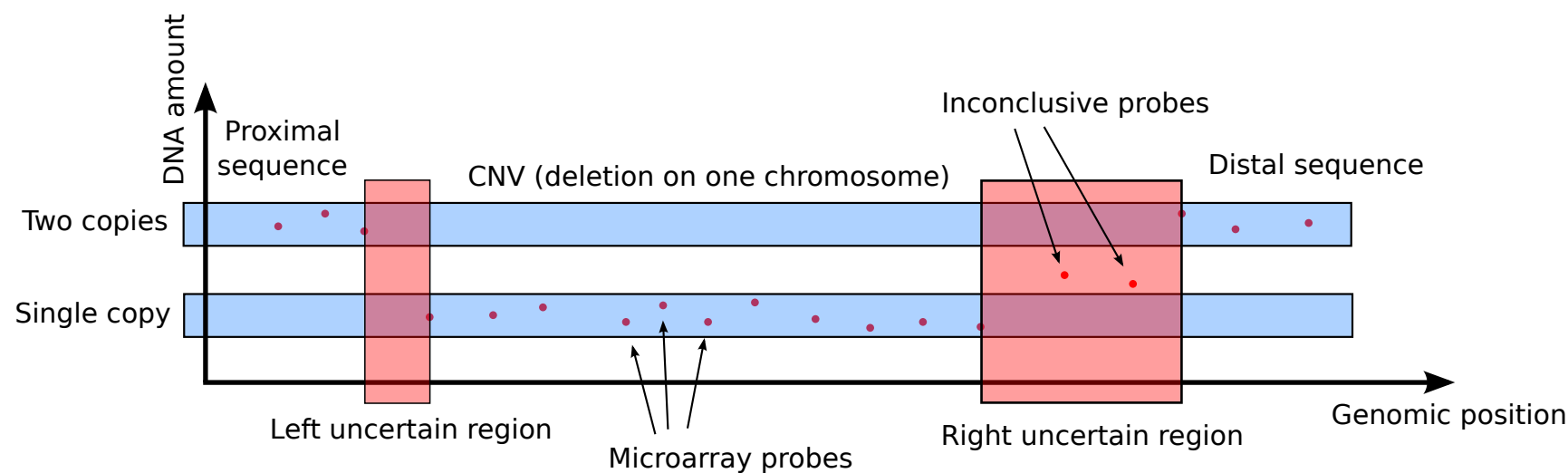
We have detected **37095** LINE pairs fulfilling the specified criteria, putting 82.8% of the human genome at risk of instability.



# chromosomal microarray analysis



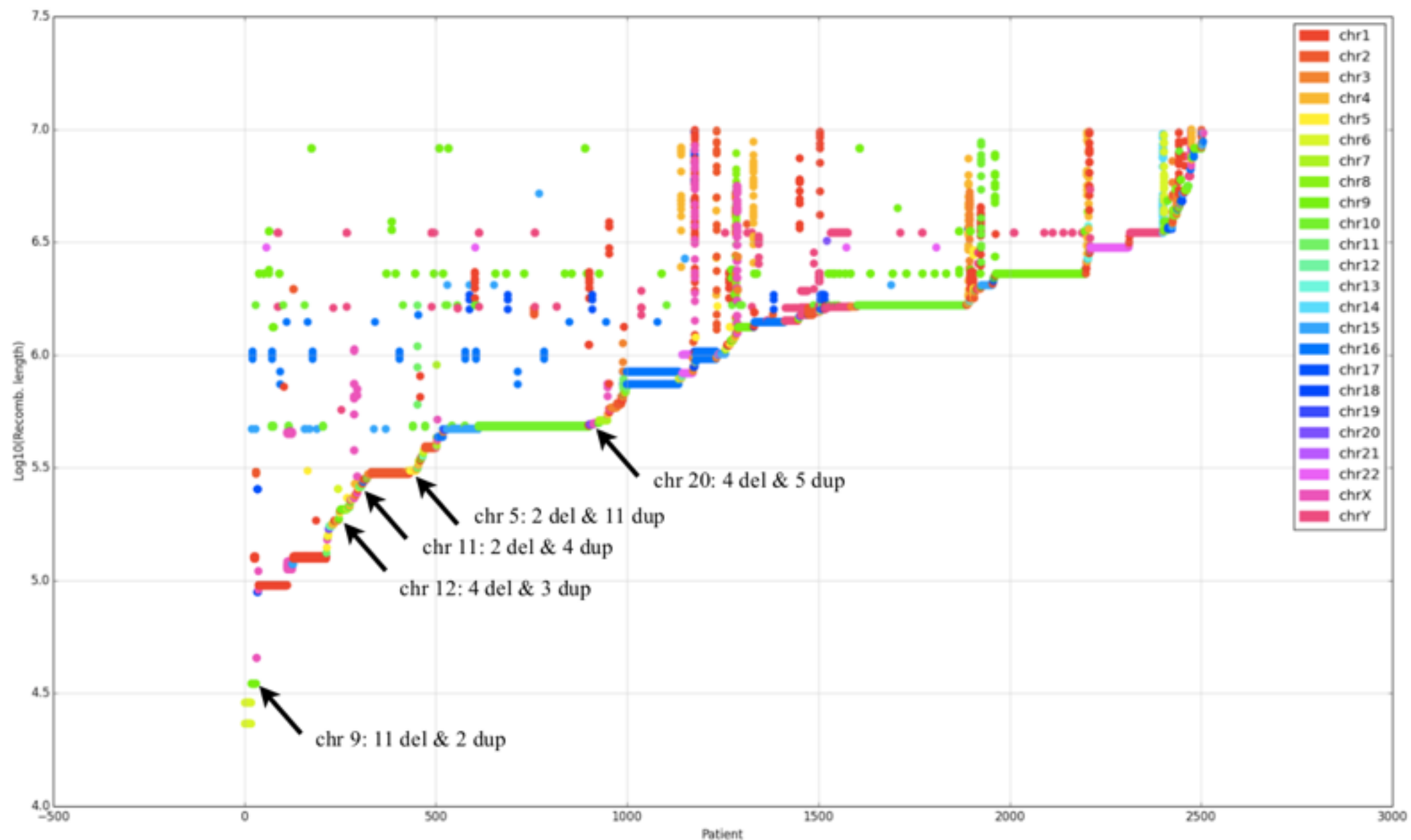
398 468 CNVs identified in 36 285 patients who underwent oligonucleotide chromosomal microarray analysis (CMA) at the Medical Genetics Laboratories at BCM.

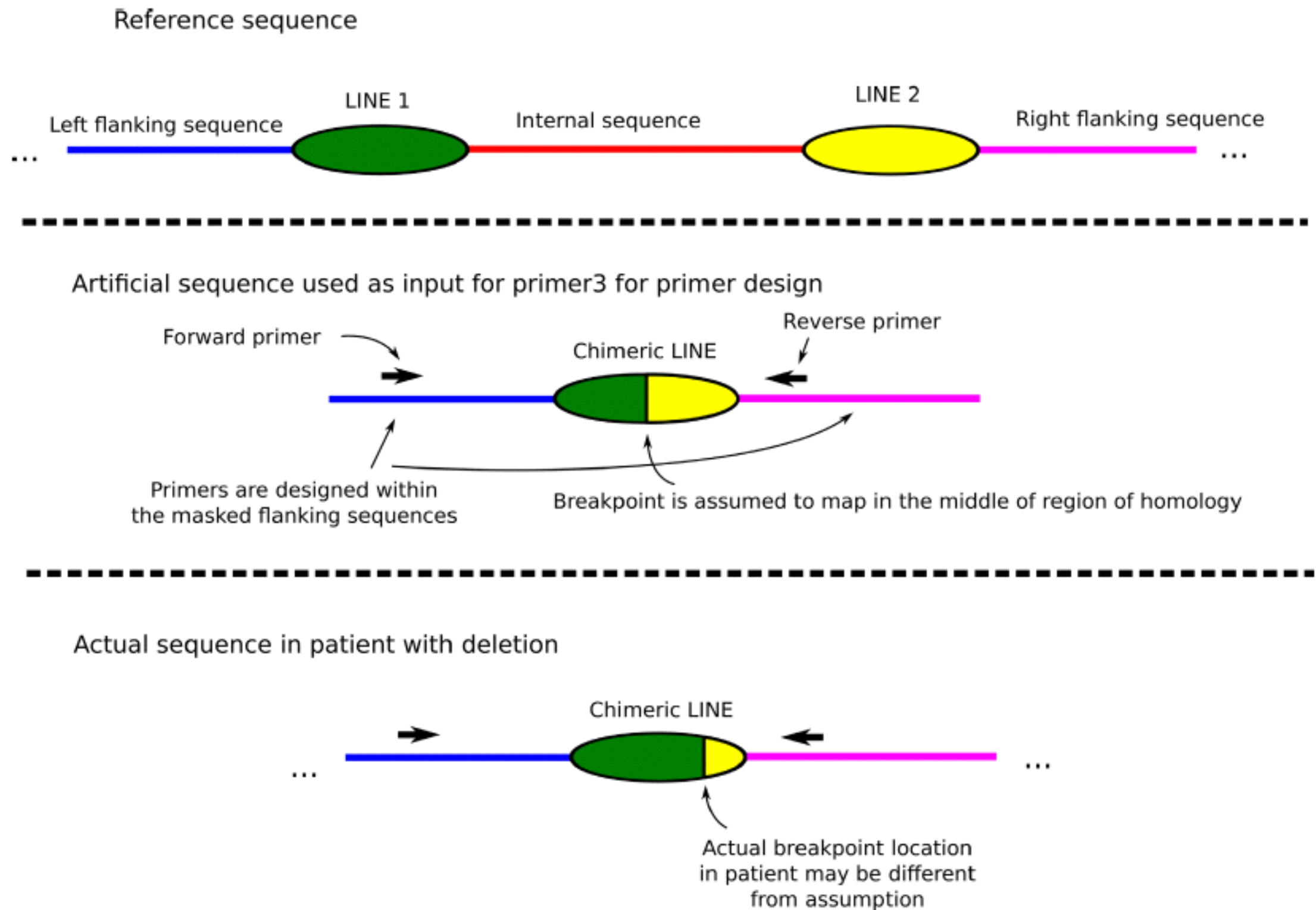




# patients

44 individuals harbouring potential LINE–LINE/ NAHR CNVs: 21 deletions and 23 duplications, from five different genomic regions.





**Figure 3.** Artificial sequences computed for primer design for detection of chimeric LINE sequences. Shown on figure is the process for deletion, with duplications and inversions being handled in similar fashion.

# **molecular validation**

Each successful amplicon was sequenced using Sanger technology.

Reads of about 1000 base pairs, starting from primers.

Each base pair is annotated with read quality

## **where are breakpoints ?**

# and healthy subjects

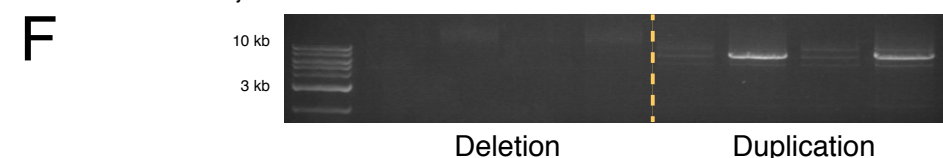
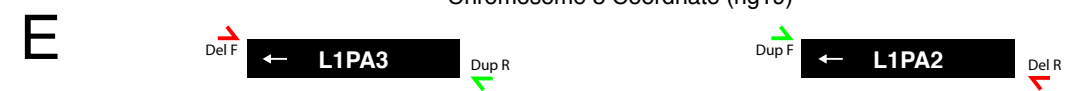
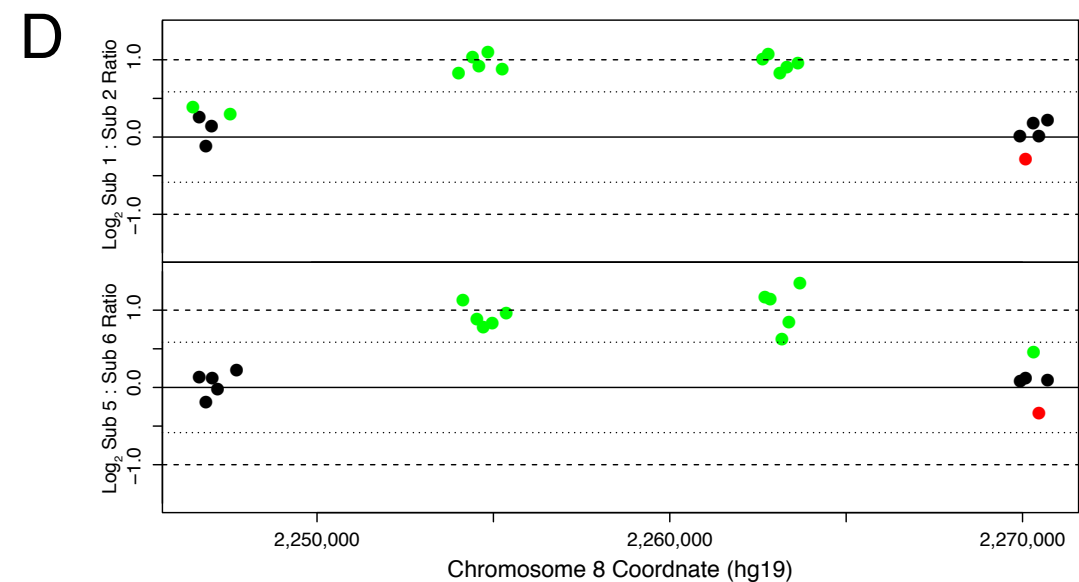
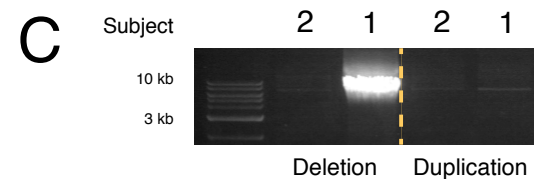
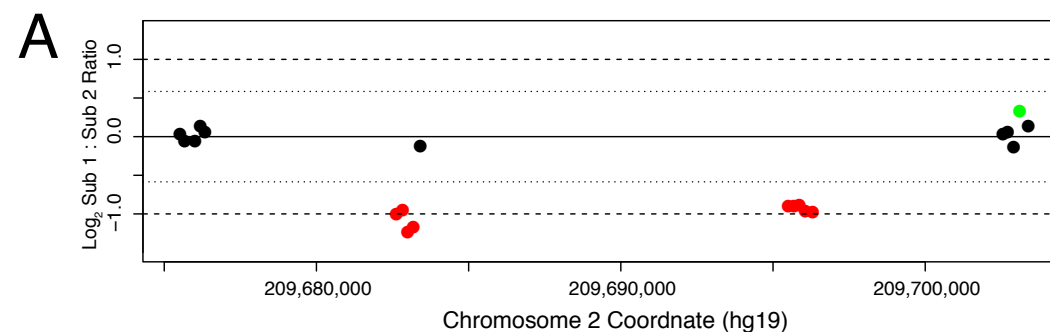


LR-PCR reactions for six healthy subjects not known to suffer from genetic disease -> 13 CNVs detected

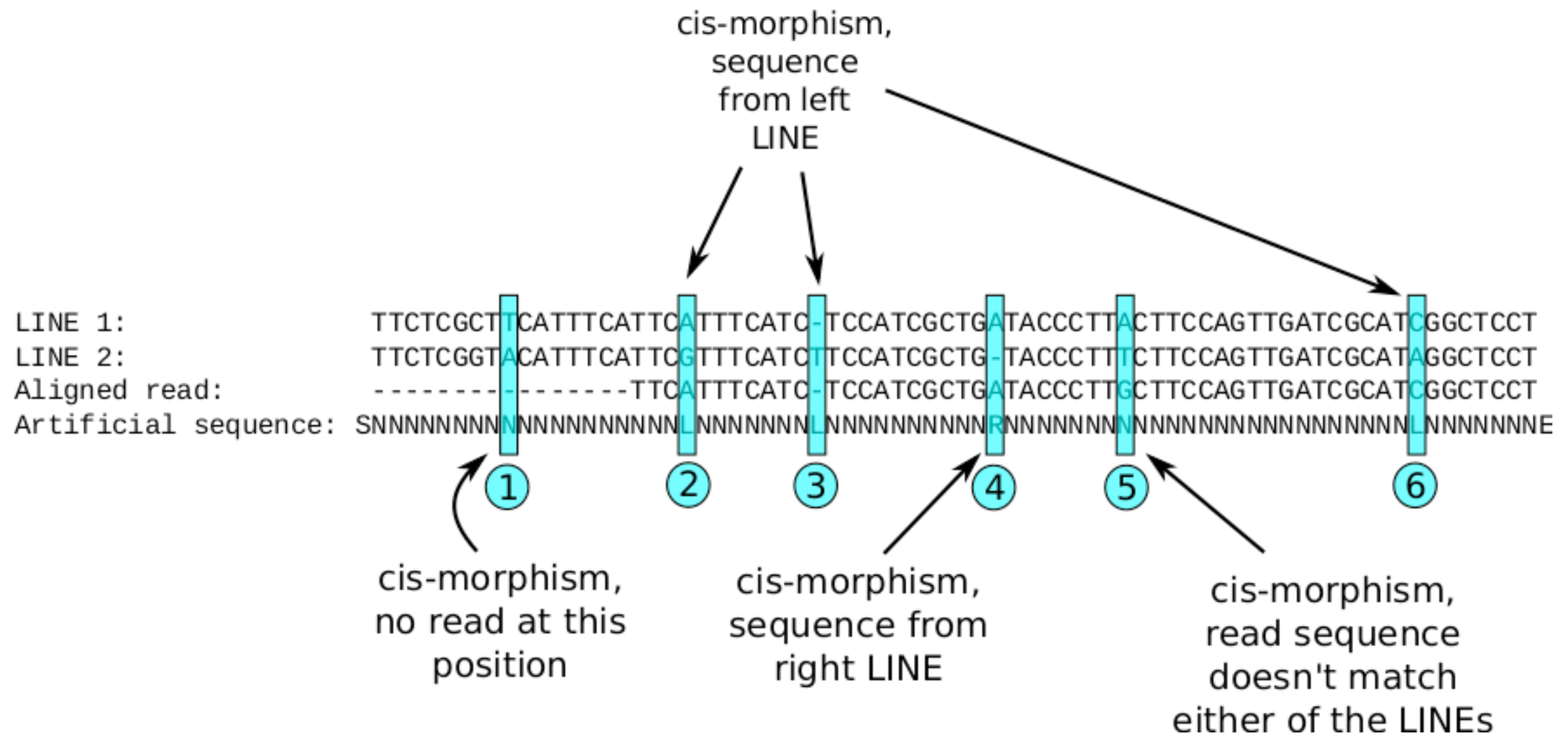
NAHRs are quite prevalent, it is expected (on average) that every person carries a several CNVs caused by NAHRs, some of them de-novo, some inherited from the parents. Most of these are benign.

# molecular validation: aCGH

- (A) Array CGH indicates a CNV.
- (B) L1PA elements that mediate the CNV and LR-PCR primers testing for the CNV.
- (C) LR-PCR identifies the presence of a deletion.
- (D) Array CGH indicates a CNV.
- (E) L1PA elements that mediate the CNVs and LR-PCR primers testing for the CNVs.
- (F) LR-PCR identifies the presence of homozygous duplications.



# breakpoint identification by HMIM

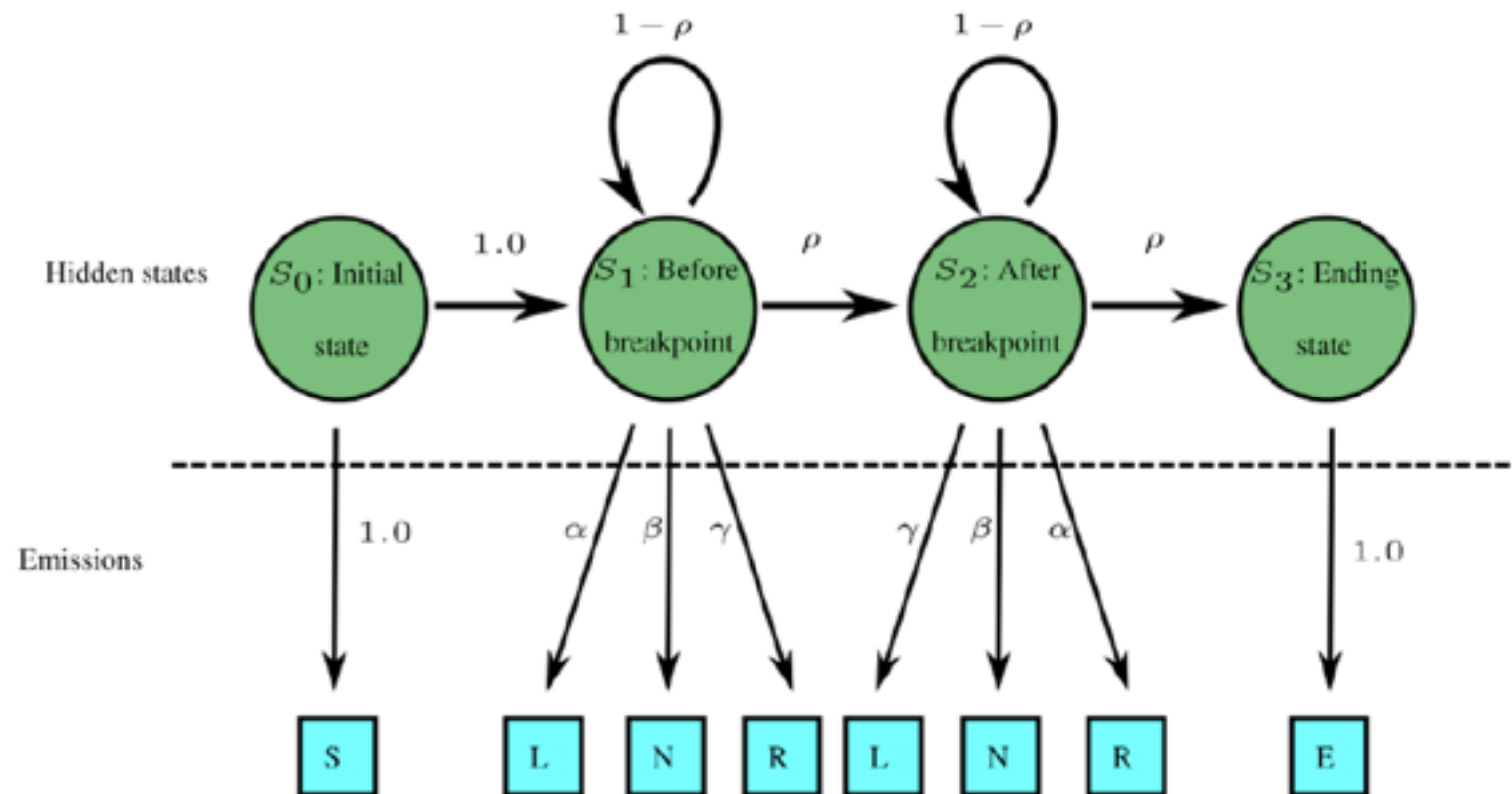


For each pair of LINEs, a consensus sequence was computed, and a custom version of the **Needleman-Wunsch** algorithm, modified to compute a semi-global alignment was used to align the Sanger reads to the consensus. An artificial sequence contains the information about sequence **cis -morphisms**



# breakpoint identification by HMM

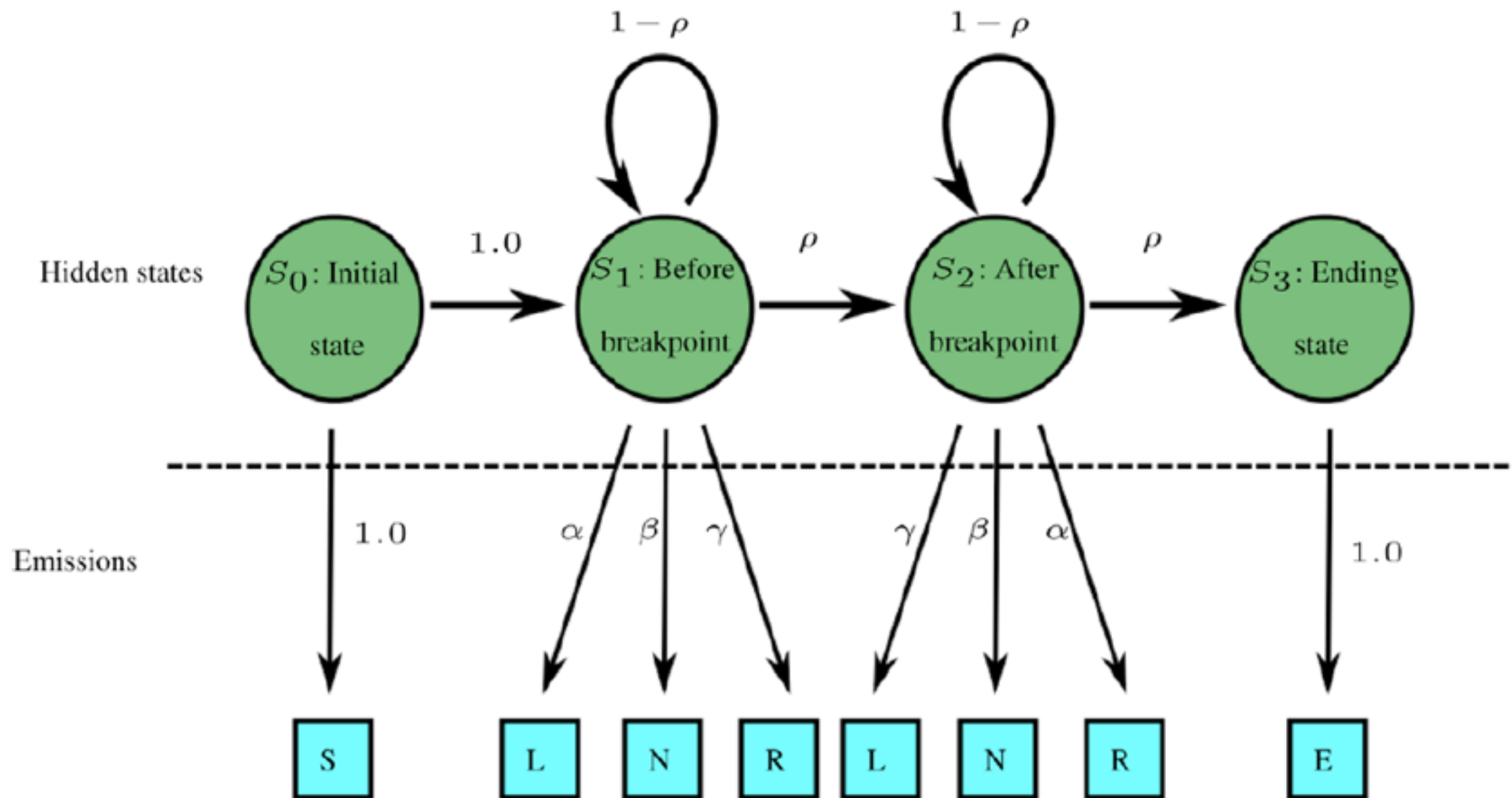
- sequences were analyzed with a Hidden Markov Model trained using a custom version of the **Baum-Welch algorithm**;
- modified algorithm differs from the standard version in that it enforced the constraints that ensures the model does not favour placement of breakpoints near the beginning or end of alignments because the training data happens to be skewed as such
- assumes that CNVs with respect to the reference sequence are equally likely to occur on either side of the breakpoint.



# breakpoint identification by HMM

- The model with parameters obtained from the Baum-Welch algorithm were then used to compute the posterior probabilities of transition from the S1 state to S2 at all locations, which correspond to the probability that the NAHR cross-over event occurred at each location.
- These were computed using a custom version of the **forward-backward algorithm**, in which the observation matrices corresponding to the L and R emissions were replaced with an affine combination of matrices for L and R with weights based on the **PHRED** quality score of the sequence from which the L or R signals originated.
- The computed locations were later confirmed by visual inspection using **Sequencher** software.

# hidden Markov model

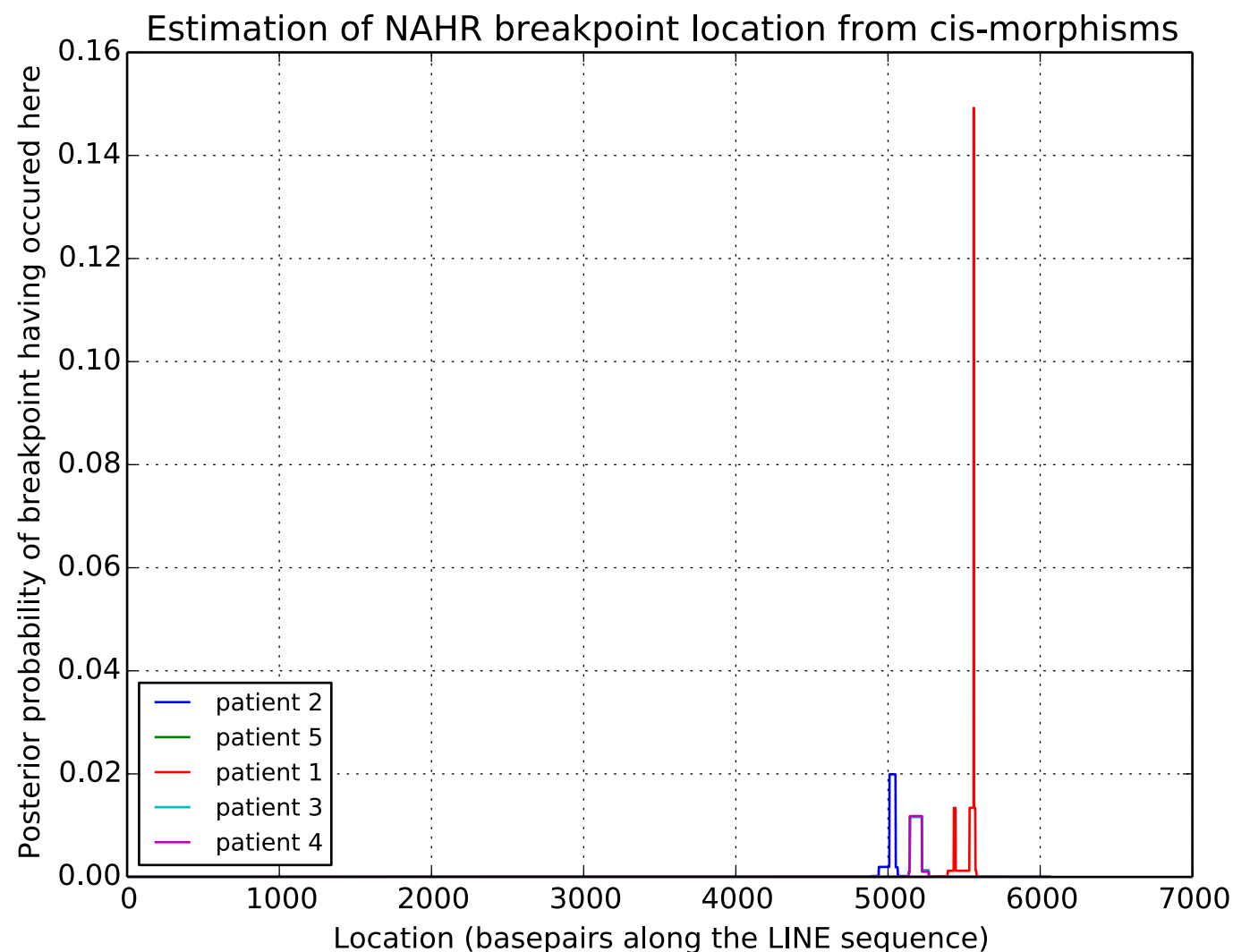


Parameter name	Prior value	Posterior value
$\alpha$	0.1	0.00924896713296794
$\beta$	0.89	0.9899202188834337
$\gamma$	0.01	0.0008308139835982798
$\rho$	0.05	0.00035456442623037844

# consensus

Estimated NAHR breakpoint location probabilities from the hidden Markov model for duplications between LINEs on chromosome 20

Three distinct NAHR loci were identified among the tested patients. For each LINE pair a consensus sequence has been computed, and each read has been aligned using Needleman-Wunsch algorithm.





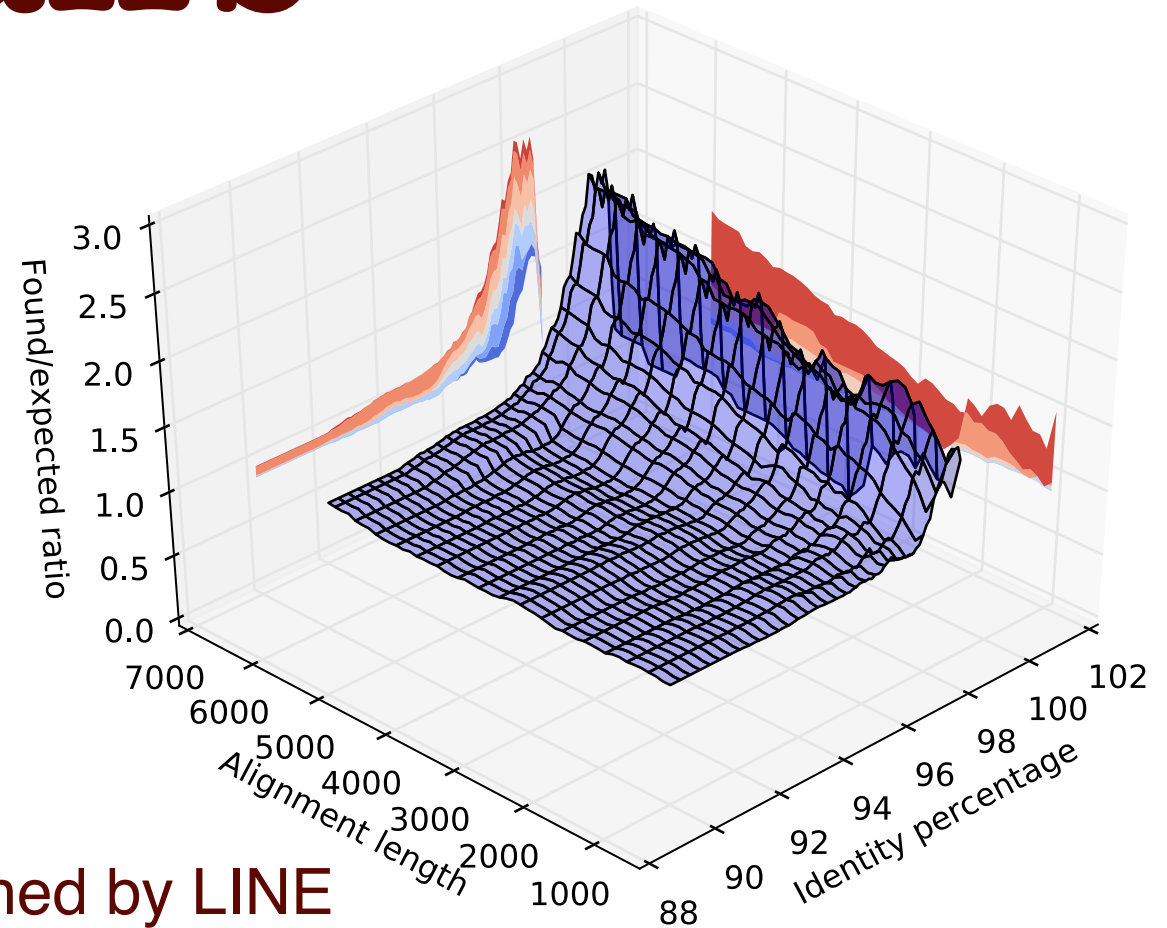
# enrichment of mediating LINE pairs

$$\mathcal{E}(l, id) = \frac{\#matched\_CNVs(l, id)}{\epsilon \cdot \#LINE\_pairs(l, id)}$$

$\#matched\ CNVs(l, id)$  - number of CNVs matched by LINE pairs with homology length of  $l$  or more and identity  $id$  or more

$\epsilon$  - expected number of matching CNVs per LINE (0.058)

$\#LINE\ pairs(l, id)$  - total number of LINE pairs with homology of  $l$  or more and identity  $id$  or more.



# conclusions

our statistical analyses showed that LINE pairs with as little as 1 kb of homology are enriched at CNV breakpoint uncertainty regions.

LINE–LINE-mediated NAHR does occur frequently and on a genome scale.

LINE elements contribute to human genetic variability by promoting NAHR in addition to well-described mechanisms of active retrotransposition.

each healthy individual carries on average three different LINE mediated NAHR CNVs.

for more details:

PRINT ISSN: 0305-1048  
ONLINE ISSN: 1362-4962

# Nucleic Acids Research

VOLUME 43 ISSUE 4 2015

www.nar.oxfordjournals.org

## NAR Breakthrough Article

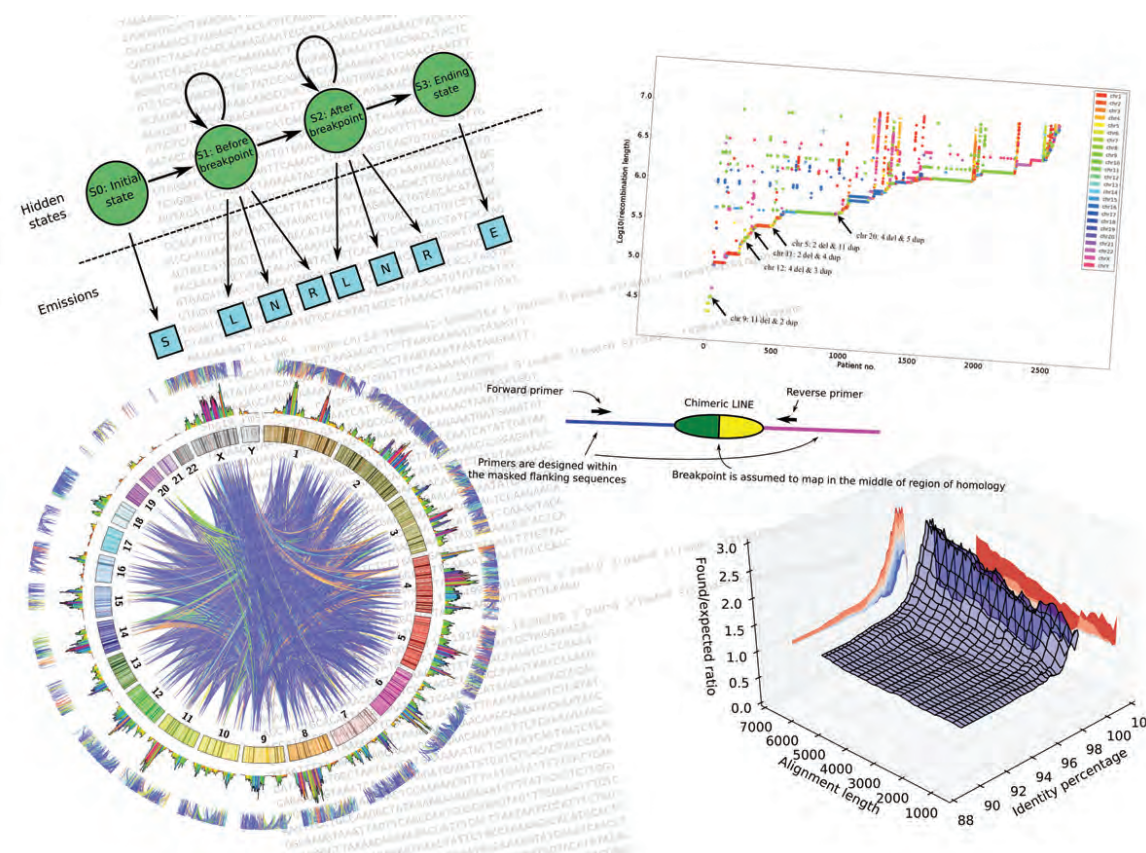
### Genome-wide analyses of LINE–LINE-mediated nonallelic homologous recombination

Michał Startek<sup>1,†</sup>, Przemysław Szafranski<sup>2,†</sup>, Tomasz Gambin<sup>2</sup>, Ian M. Campbell<sup>2</sup>, Patricia Hixson<sup>2</sup>, Chad A. Shaw<sup>2</sup>, Paweł Stankiewicz<sup>2,\*</sup> and Anna Gambin<sup>1,3,\*</sup>



### NAHR-mediated copy-number variants in a clinical population: mechanistic insights into both genomic disorders and Mendelizing traits

Piotr Dittwald, Tomasz Gambin, Przemysław Szafranski, et al.



OXFORD  
UNIVERSITY PRESS

Open Access

No barriers to access – all articles freely available online



# Many thanks to collaborators



Piotr Dittwald



Tomek Gambin



Michał Startek



Maciek Sykulski



Paweł Stankiewicz