ERIBA European Research Institute for the Biology of Ageing

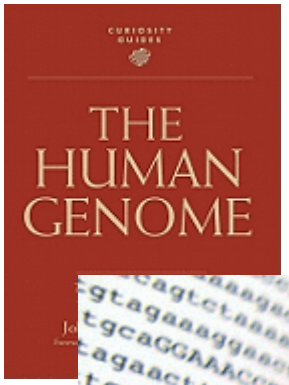
nbic netherlands bioinformatics centre

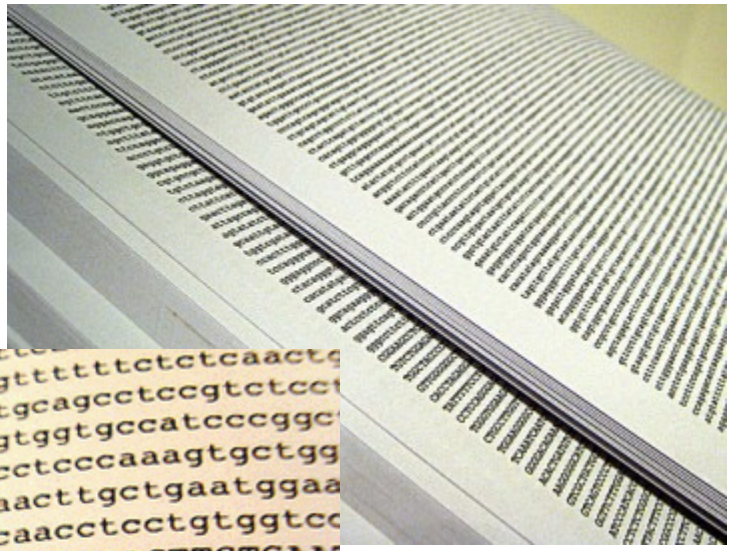
Discovery and annotation of structural variants

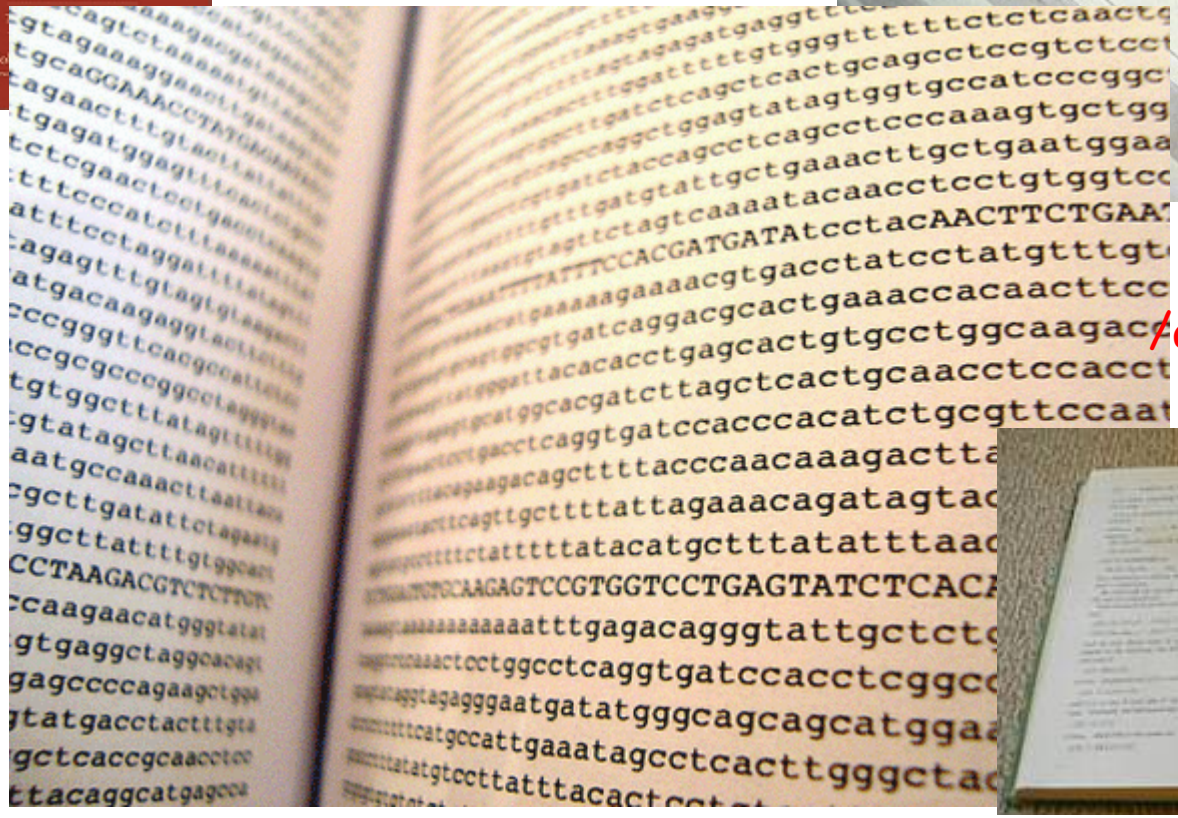
Victor Guryev
European Research Institute for the Biology of Ageing (ERIBA)
UMC Groningen, Rijksuniversiteit Groningen
NBIC

High-Throughput Next Generation Biology Course
Groningen
May 7, 2013

The human genome ‘book’

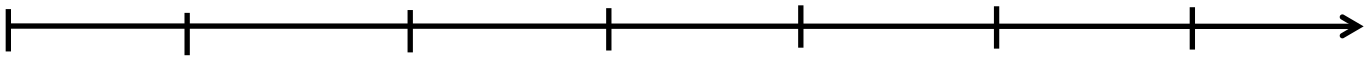






Scale of genetic differences

Scale



Chromosome 1 Mb 100 kb 10 kb 1 kb 100 bp 10 bp 1 bp

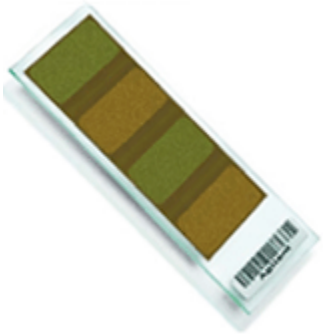
Variants

Short indels

SNPs

S t r u c t u r a l v a r i a n t s

Tools
Solutions



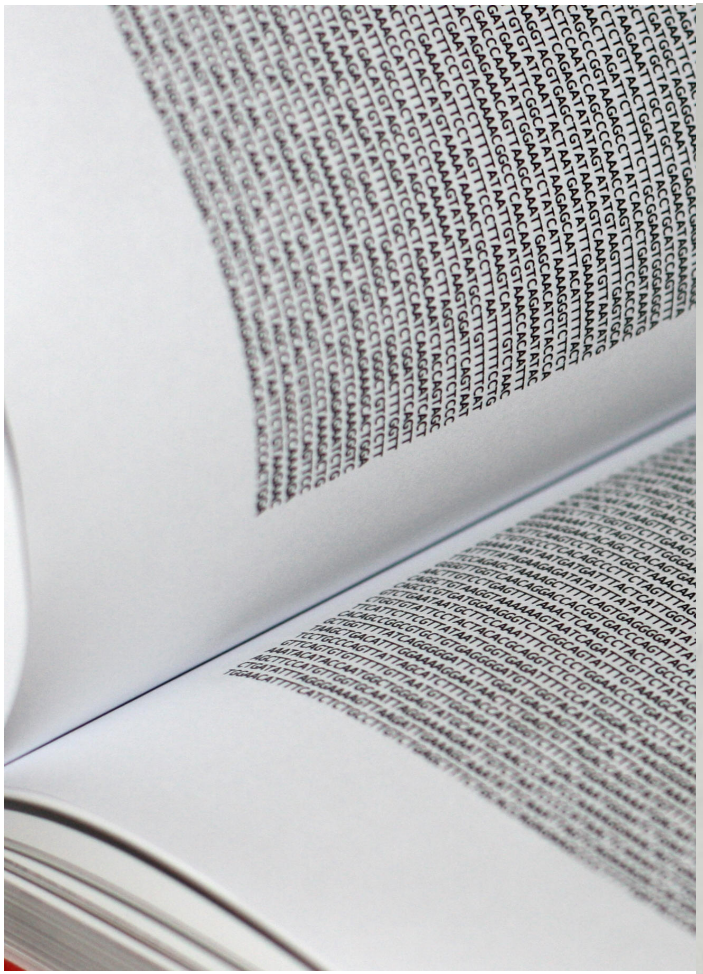
ERIBA European Research Institutes for the Biology of Ageing

$1000 genome and beyond

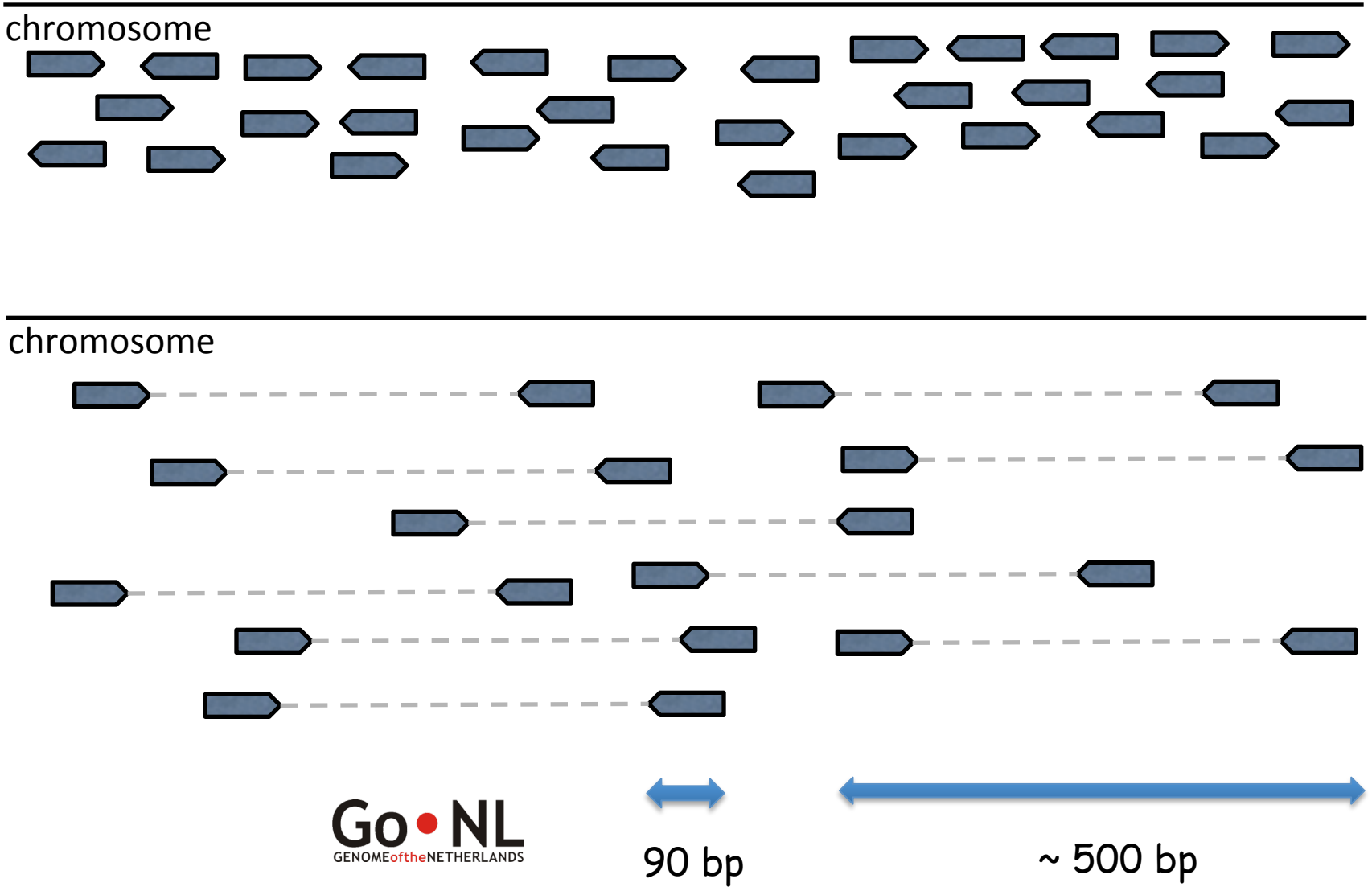
We can be confident in predicting that the $1,000 human genome will be achieved in 2013

Life sciences are ready for a revolution, but it will require collaboration on many fronts, says **Yang Huanming**, president of BGI (the Beijing Genomics Institute)

How do we get our NGS genomes?



Fragment and paired-sequencing



Advantages of paired-sequencing

1) Twice as many bases per slide !

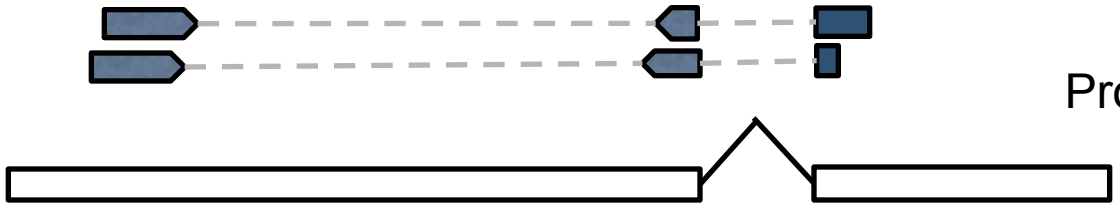
2) Structural information !!!

Molecular haplotyping (phasing)

Genome assembly

Structural variants

Better repeat coverage



Profiling of transcript isoforms

SV types and their detection

Structural Genome Variations (SVs)
ABCD

Copy-number variants

Deletion
ABD

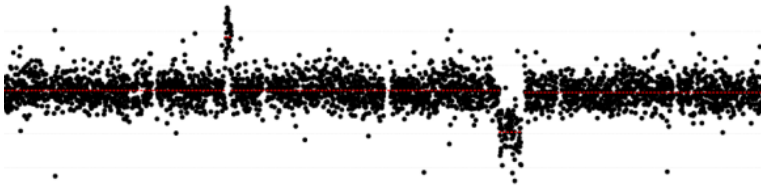
Duplication
ABCCCD

Copy-balanced variants

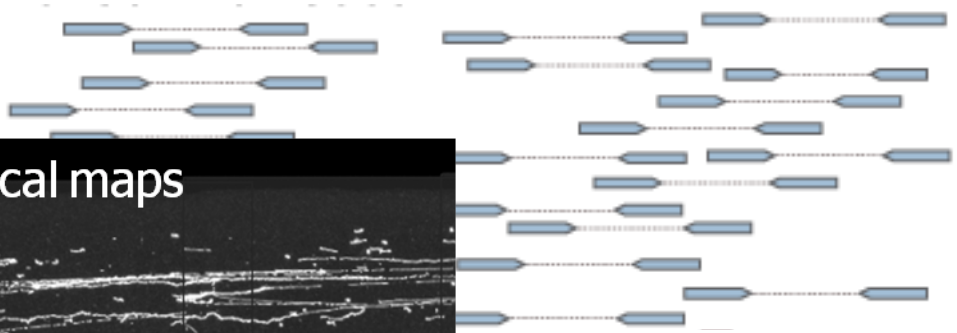
Inversion
ADCB

Translocation
AB **CD**

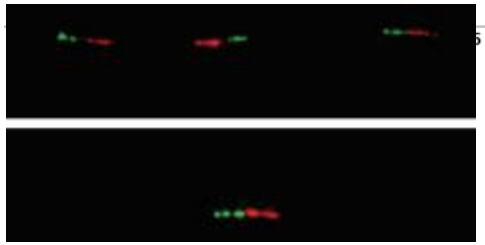
aCGH

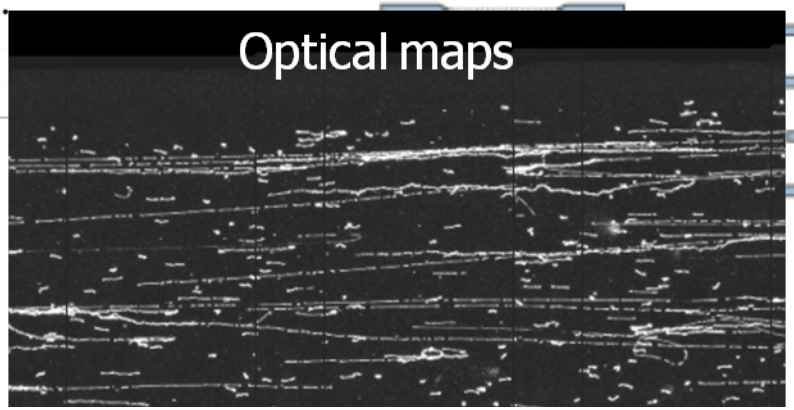


Di-tag fosmid and NGS sequencing

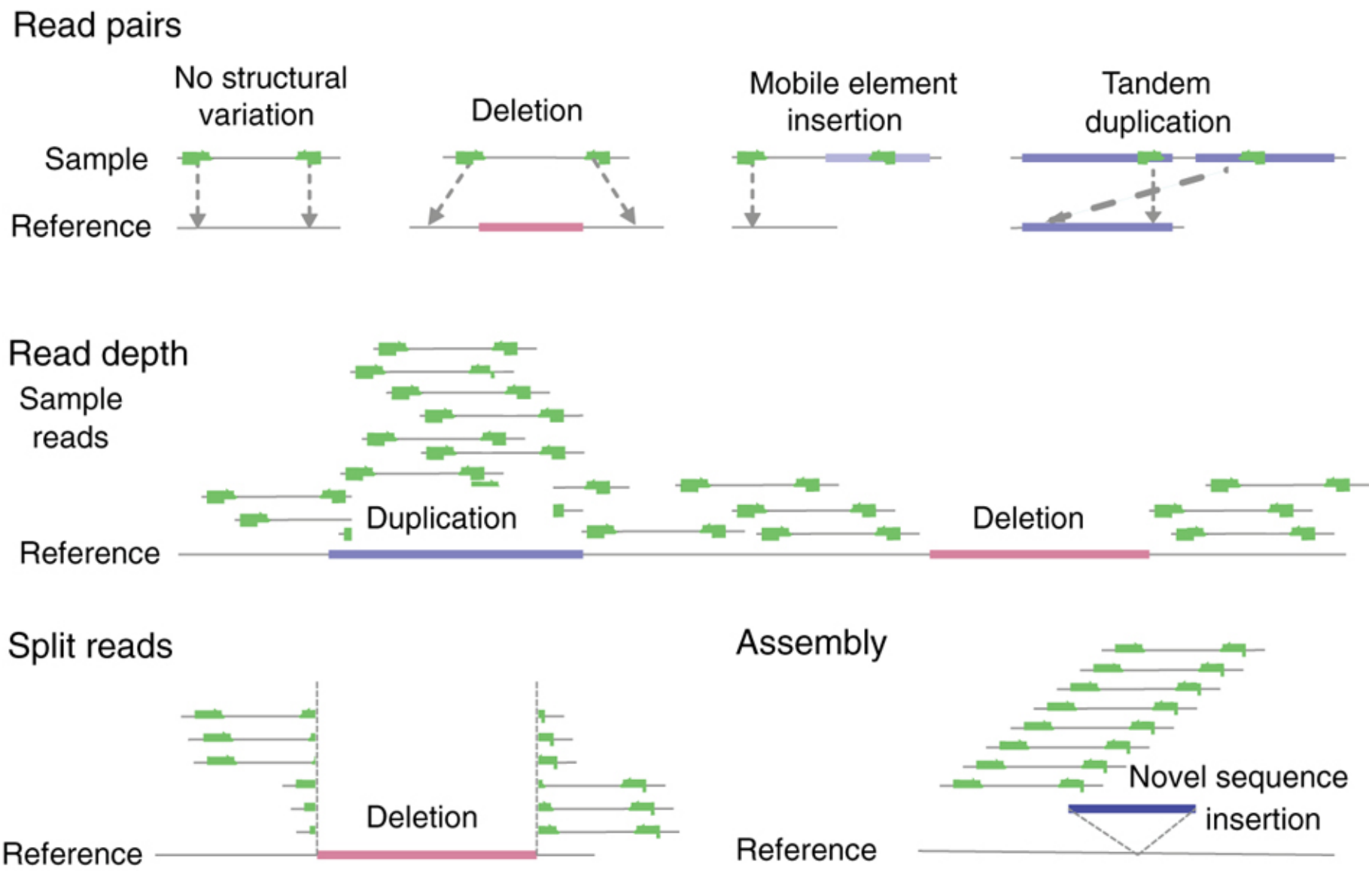


Fibre-FISH





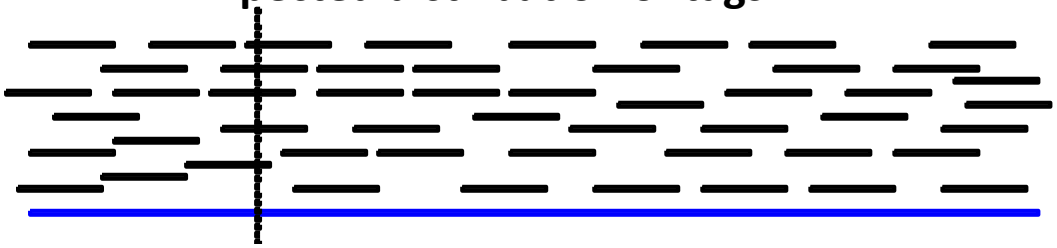
Approaches for SV detection using NGS data



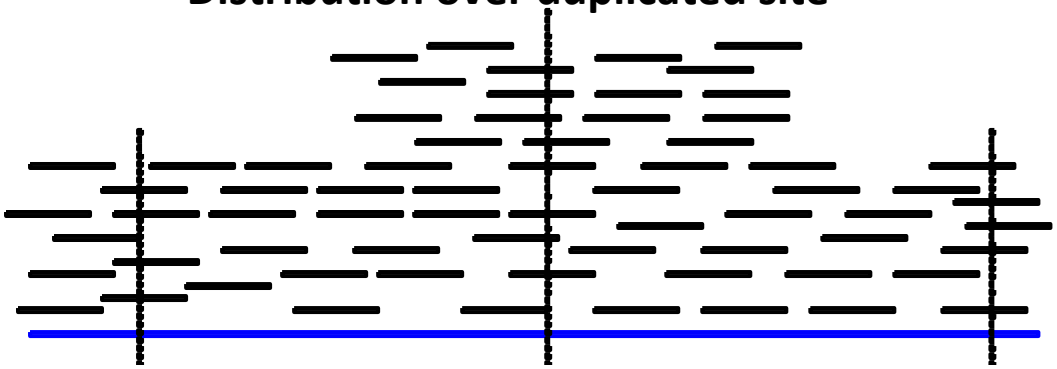
[Korbel et al, Nature Methods]

RD: Read density analysis

Expected distribution of tags



Distribution over duplicated site



Scope:
Copy-number changes

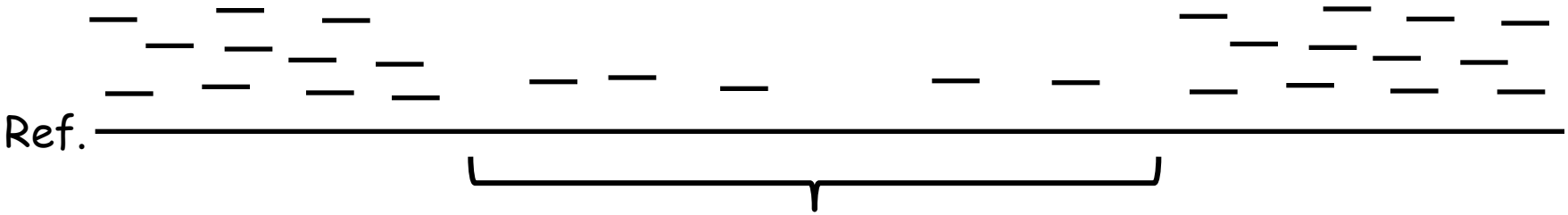
Tools:
CNV-Seq (Xie &Tammi 2009)

SegSeq (Chiang et al, 2009)

DWAC-Seq (our tool)

Real data: non-uniformity of genome coverage

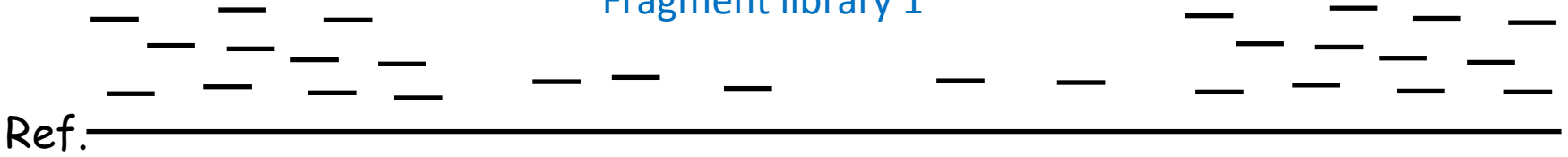
Fragment library, sample vs genome reference



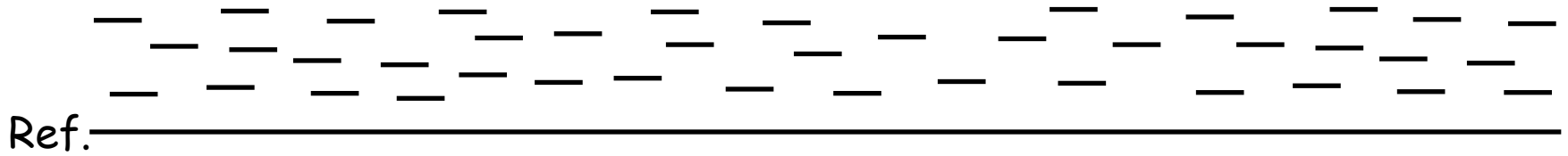
Heterozygous deletion ?
'Sequenceability' issue ?
'Mappability' issue ?

Sample vs sample comparison

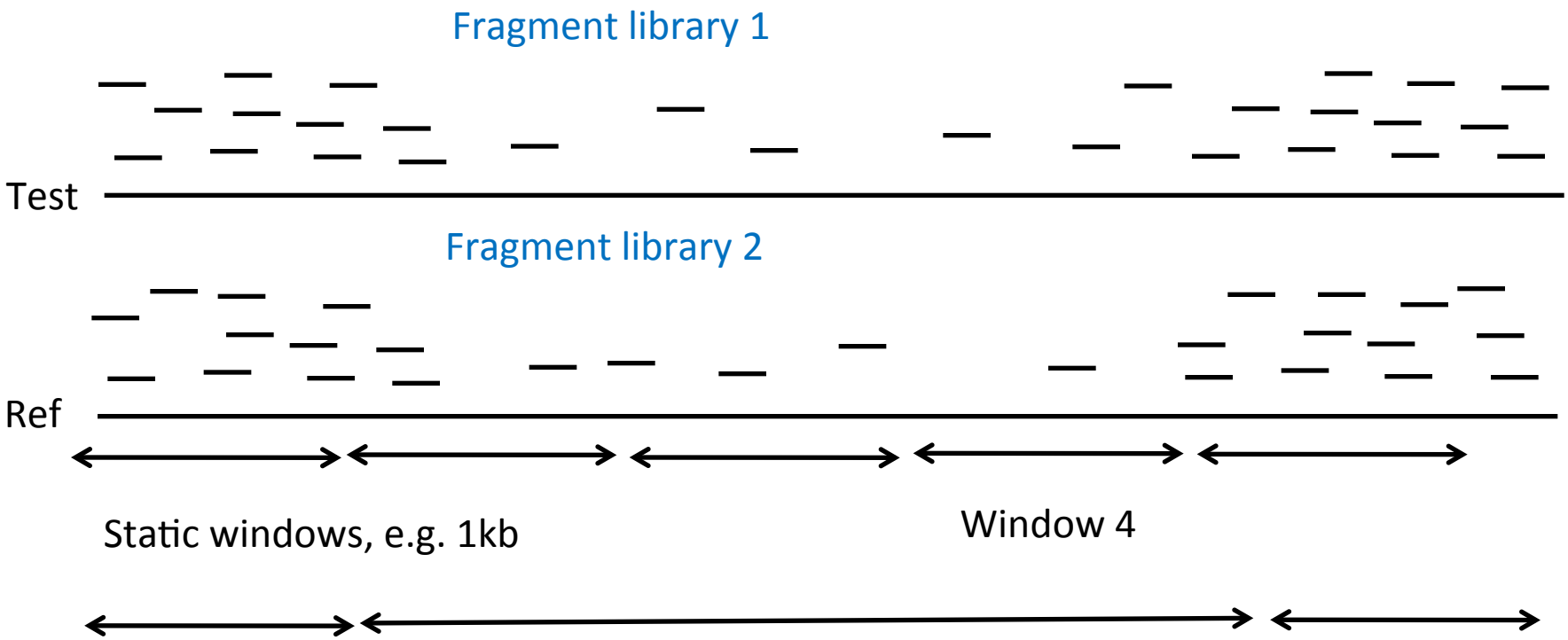
Fragment library 1



Fragment library 2



Read depth analysis with dynamic windows



Step1: Segmenting the genome, CNV calling
Step2: Fine-mapping to determine their exact breakpoints

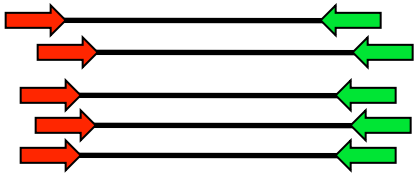
Dynamic-Window Approach for CNV calling using nextgen Sequencing
http://tools.genomes.nl

Paired-end and mate-pair libraries

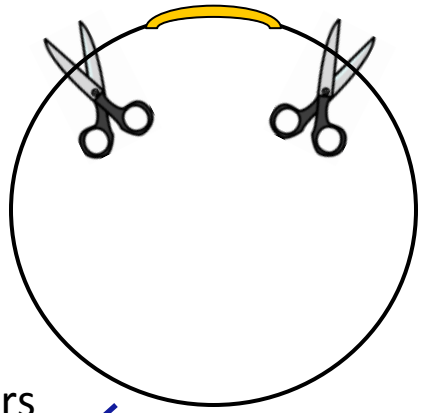
Fragmentation,
size selection

Adaptors
ligation

Internal adaptor
ligation,
circularization



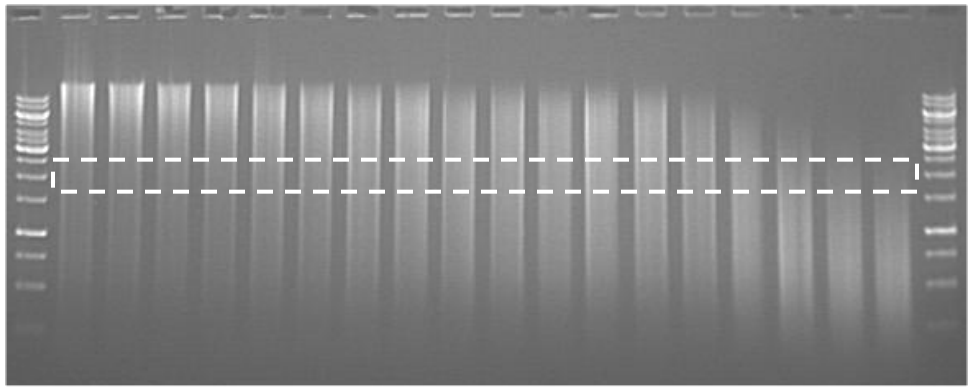
Paired-end library
Insert size
200-600 bp



Adaptors
ligation

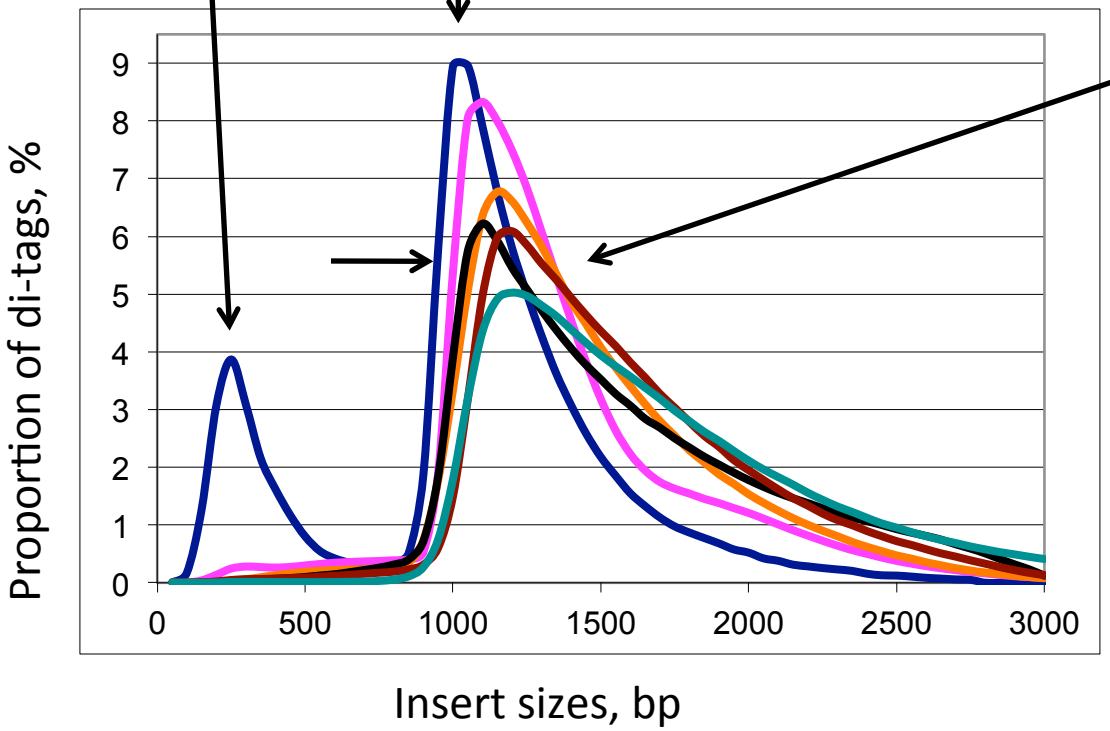
Mate-pair library
Insert size
0.6-25 kb

QC of paired libraries



Size of insert

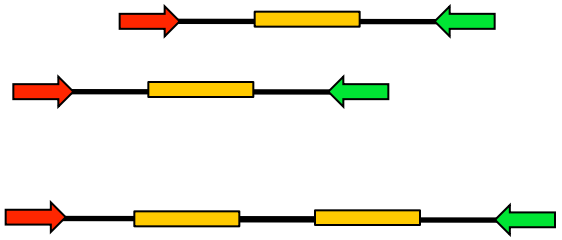
Secondary peaks



Sharpness of size distribution

Clonality
- Low DNA input/ overamplification

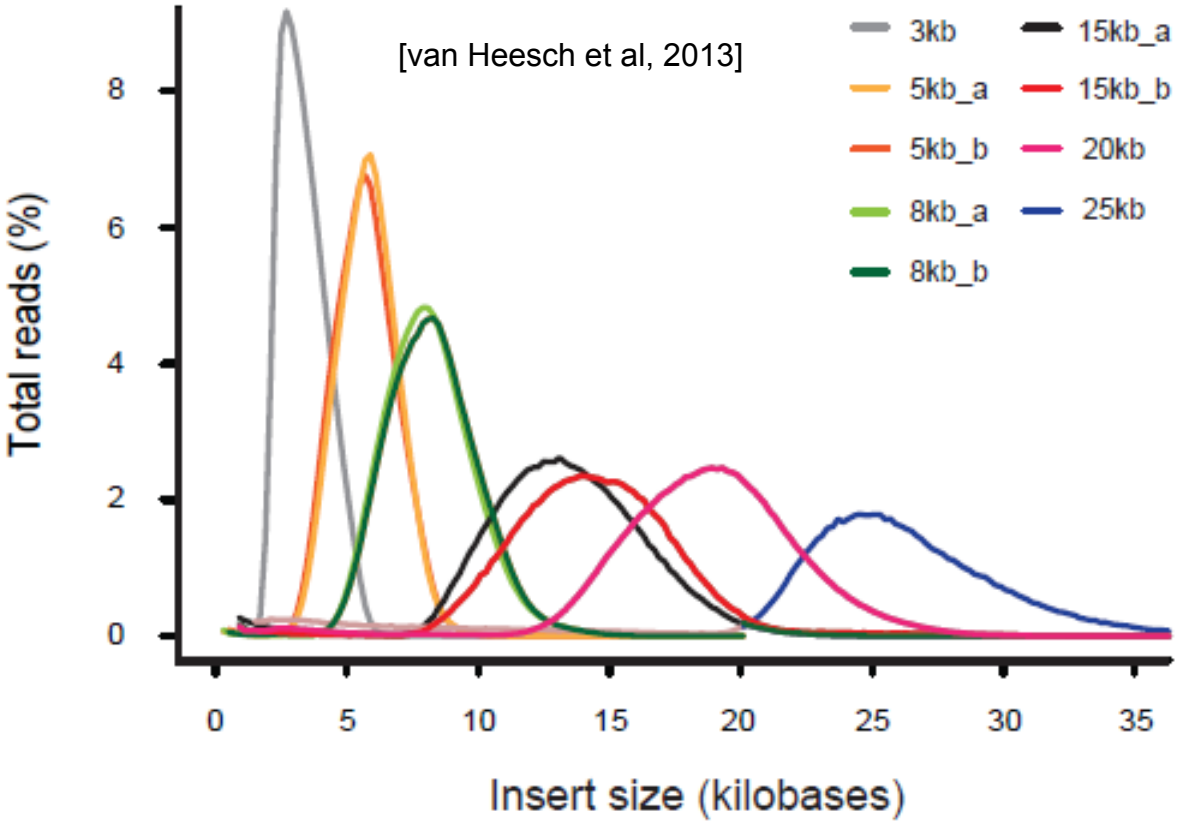
Chimerism



Di-tag libraries size: S, M, L, XL, …

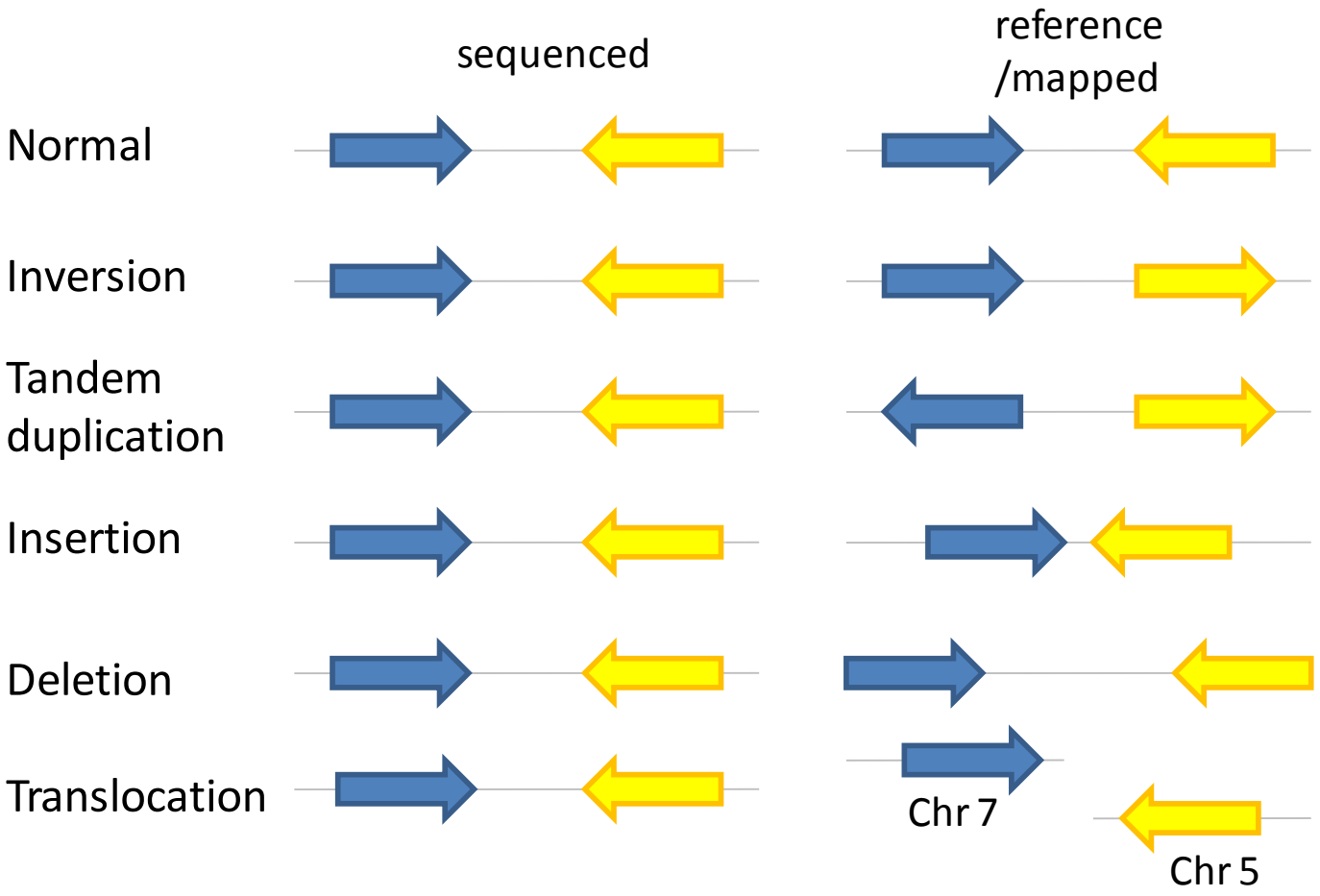
Different insert sizes to cover small and large variants with optimal precision

Di-tag libraries generated: PE (200bp), MP (2,3,5,8,15,20,25 kb)

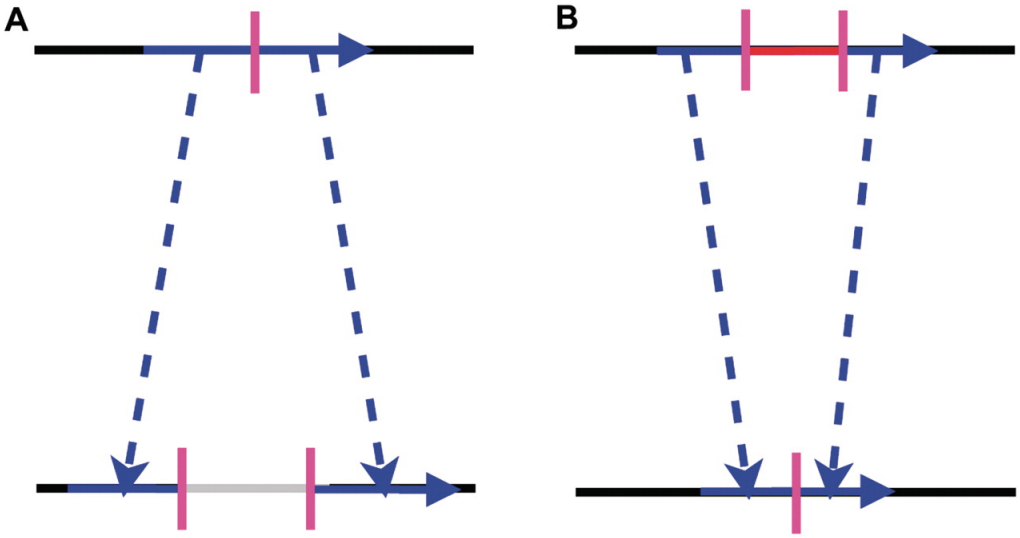


+ Fosmid libraries:
40kb tags
Protocol for Illumina
[Williams et al 2012]

Signatures of structural variation



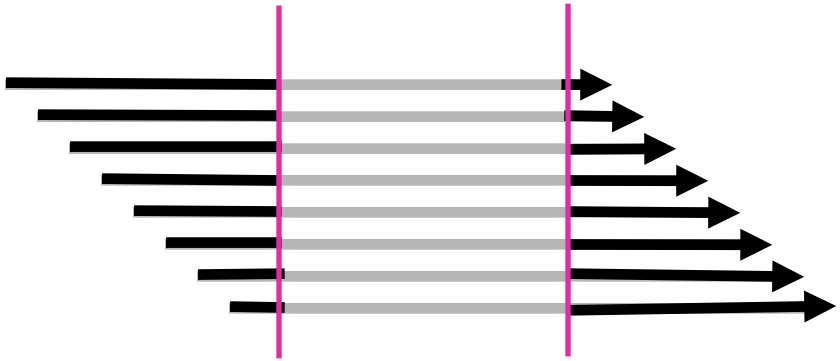
Split-reads mapping



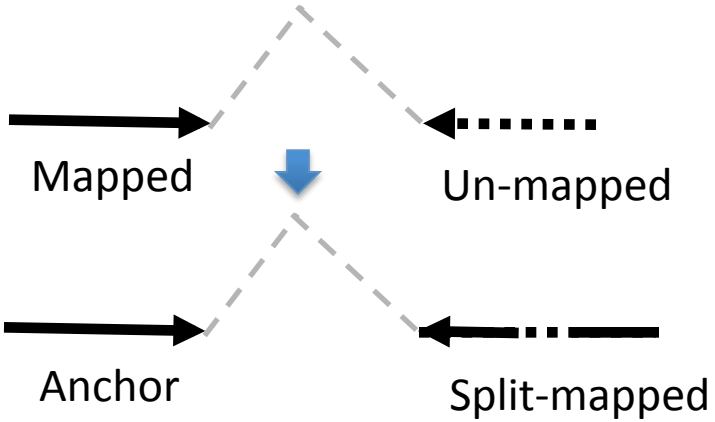
Scope: prediction of copy-number and copy-neutral SV at nucleotide resolution

Tools:
Pindel (Ye et al, 2009)
SRiC (Zhang et al, 2011)

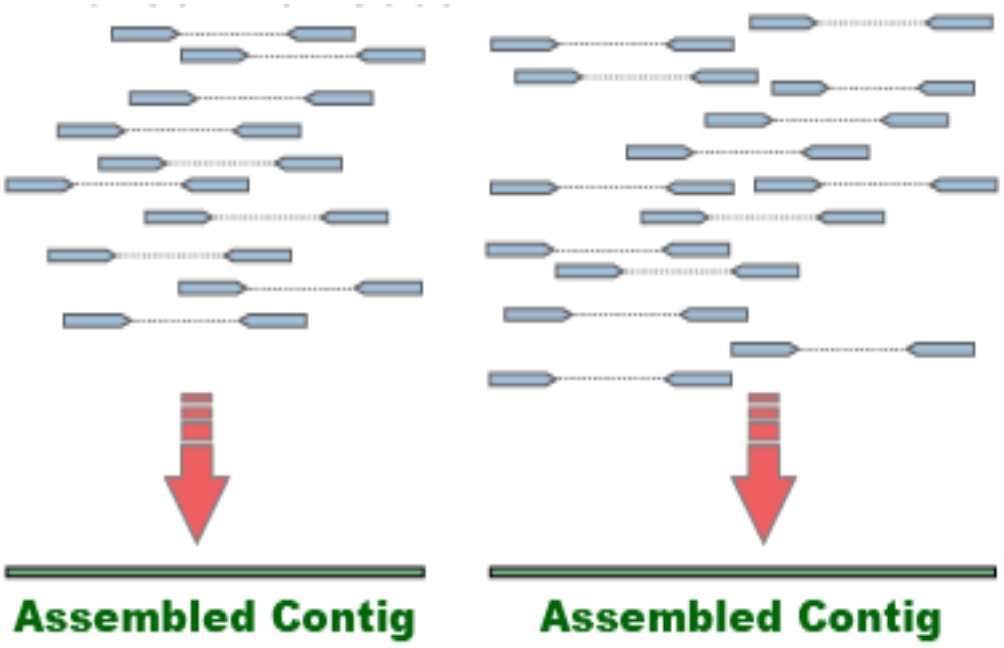
Evidence from multiple reads



Advantage of **paired reads**



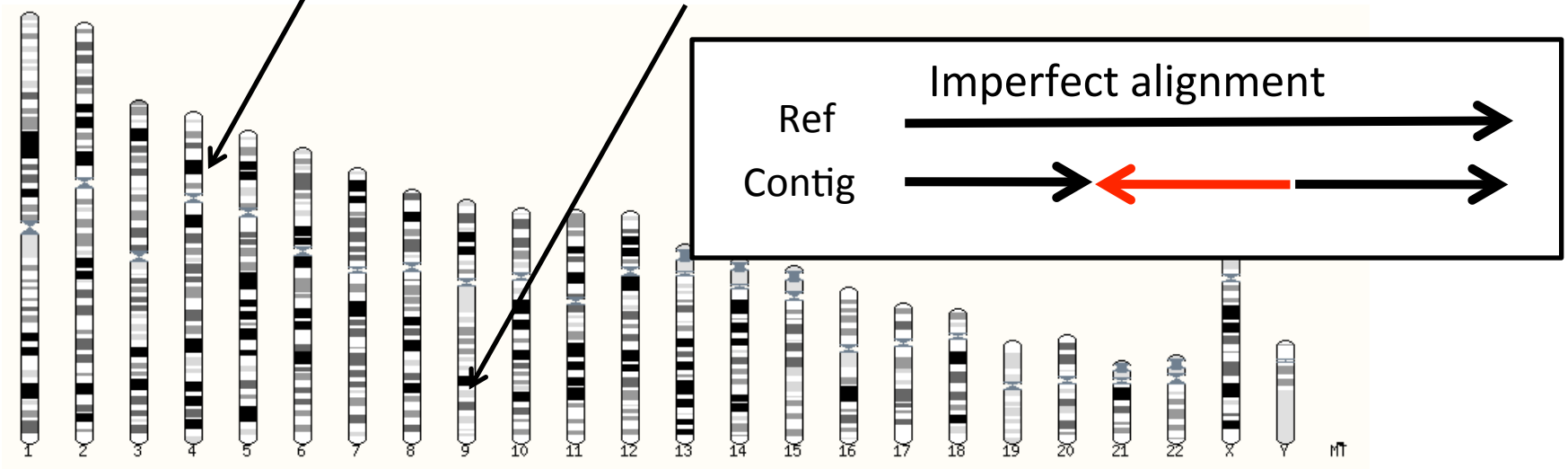
De novo assembly for SV detection



Scope: various types of SVs including large inserts

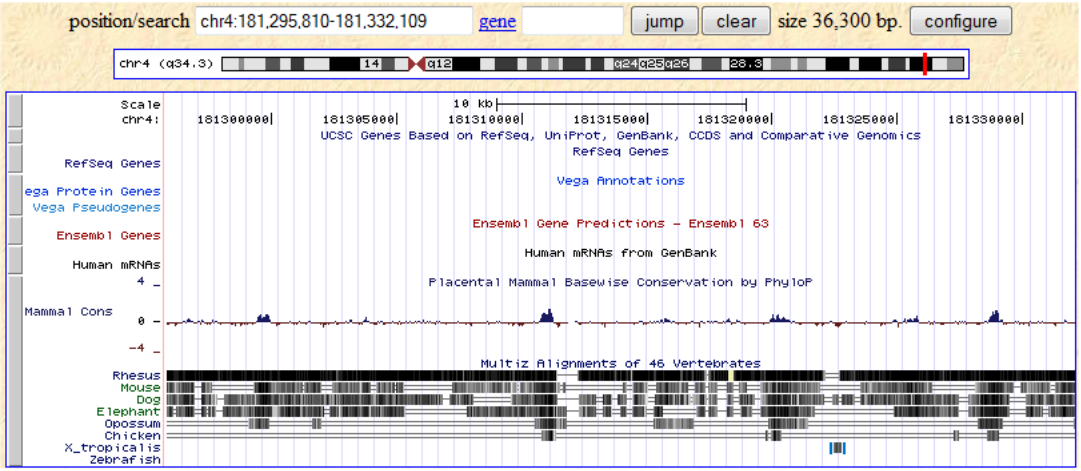
Tools: de novo assemblers SOAPdenovo, ABYSS, Allpaths-LG, Velvet

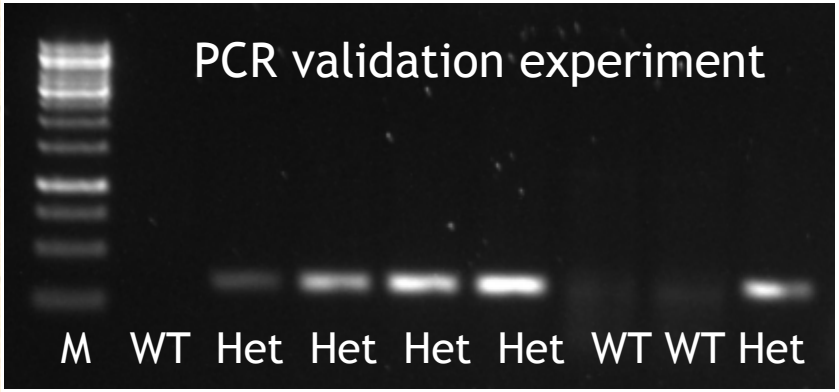
BLAST/BLAT search for comparison of contigs and genome reference

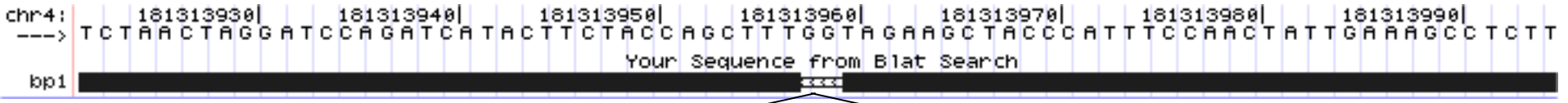


De novo assembly: sample-specific segments

Comparison of individual *de novo* assembly of GoNL data to GRCh37
~ 70 new regions (> 1kb), totaling 235 kb of NEW sequences





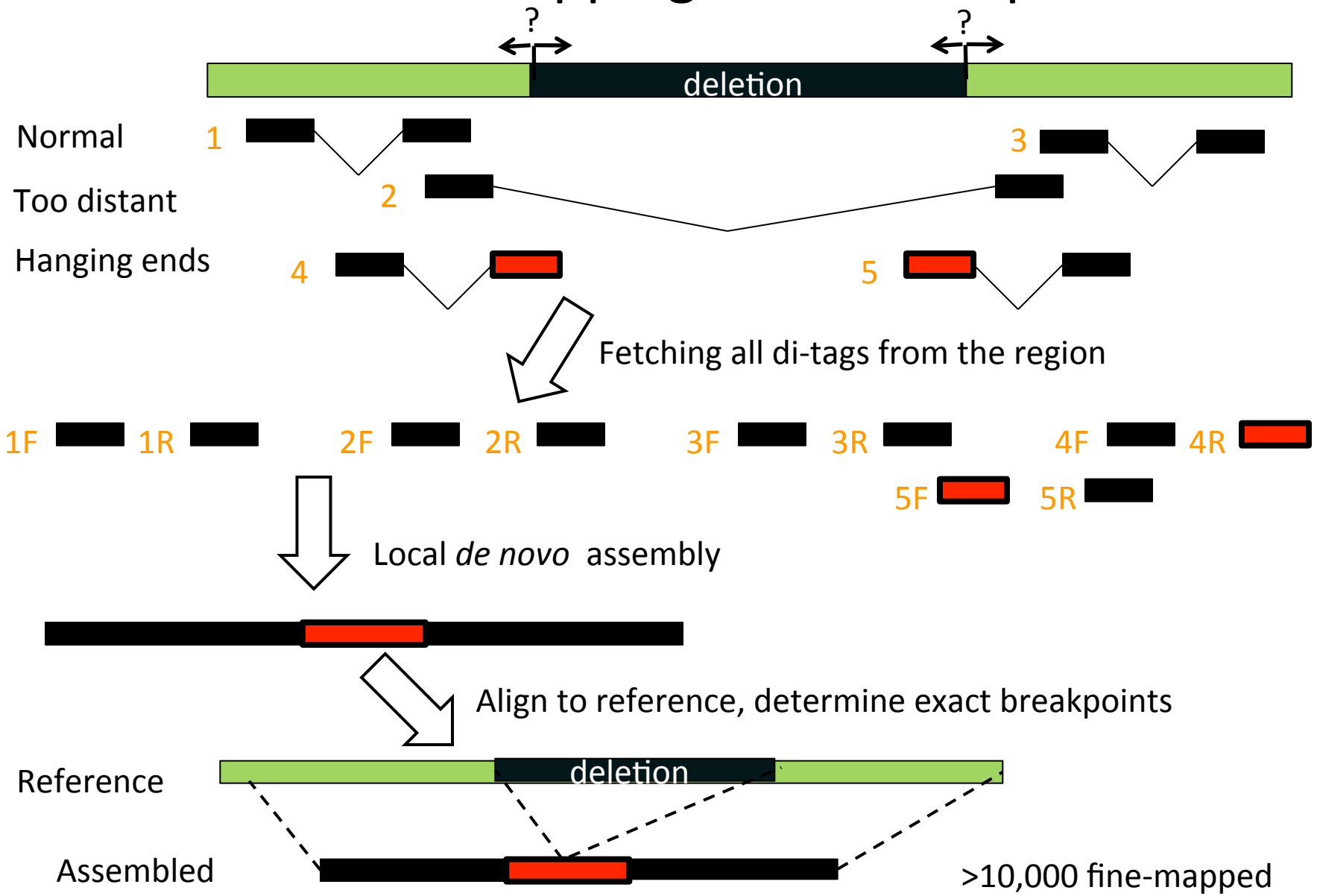


1,765 bp GCATAGGAAT…….AGAAATCAAG

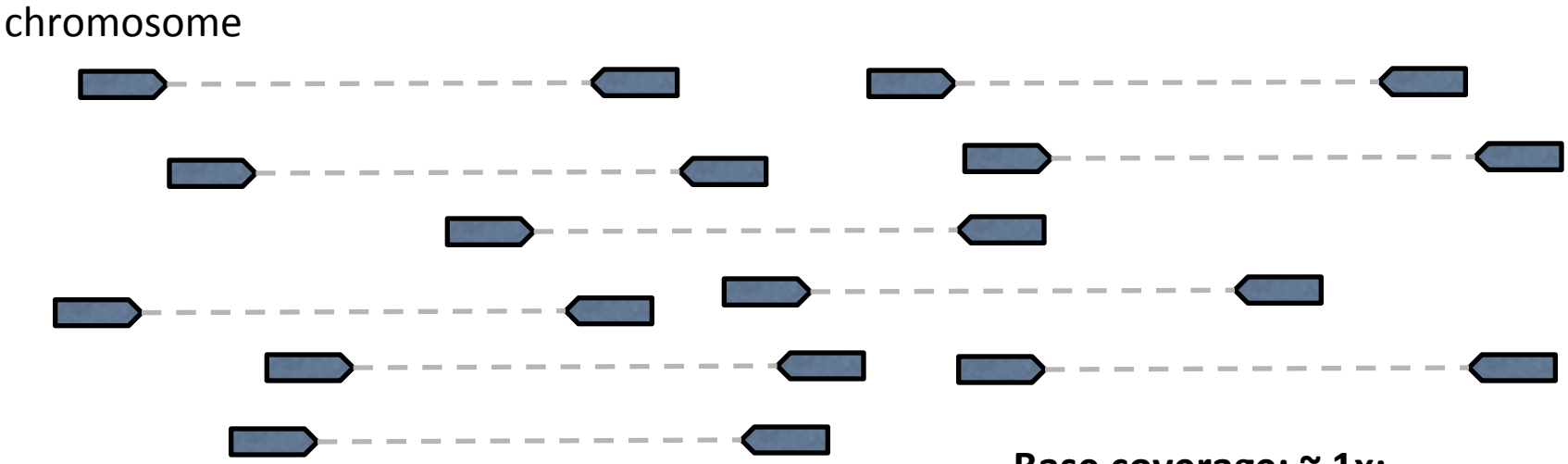
- No significant homology to known sequences on nucleotide and amino-acid levels
- Estimated frequency: ~42% in Dutch population, ~5% in 1000 Genomes

ERIBA European Research Institute for the Biology of Ageing

Local *de novo* assembly: fine-mapping of SV breakpoints



Base- and physical coverage



Base coverage: ~ 1x;
Physical coverage ~ 4x

Approach	Base coverage	Physical coverage
Depth of coverage	✔	
Discordant pairs		✔
Split-mapping	✔	
De novo assembly	✔	✔

ERIBA European Research Institute for the Biology of Ageing

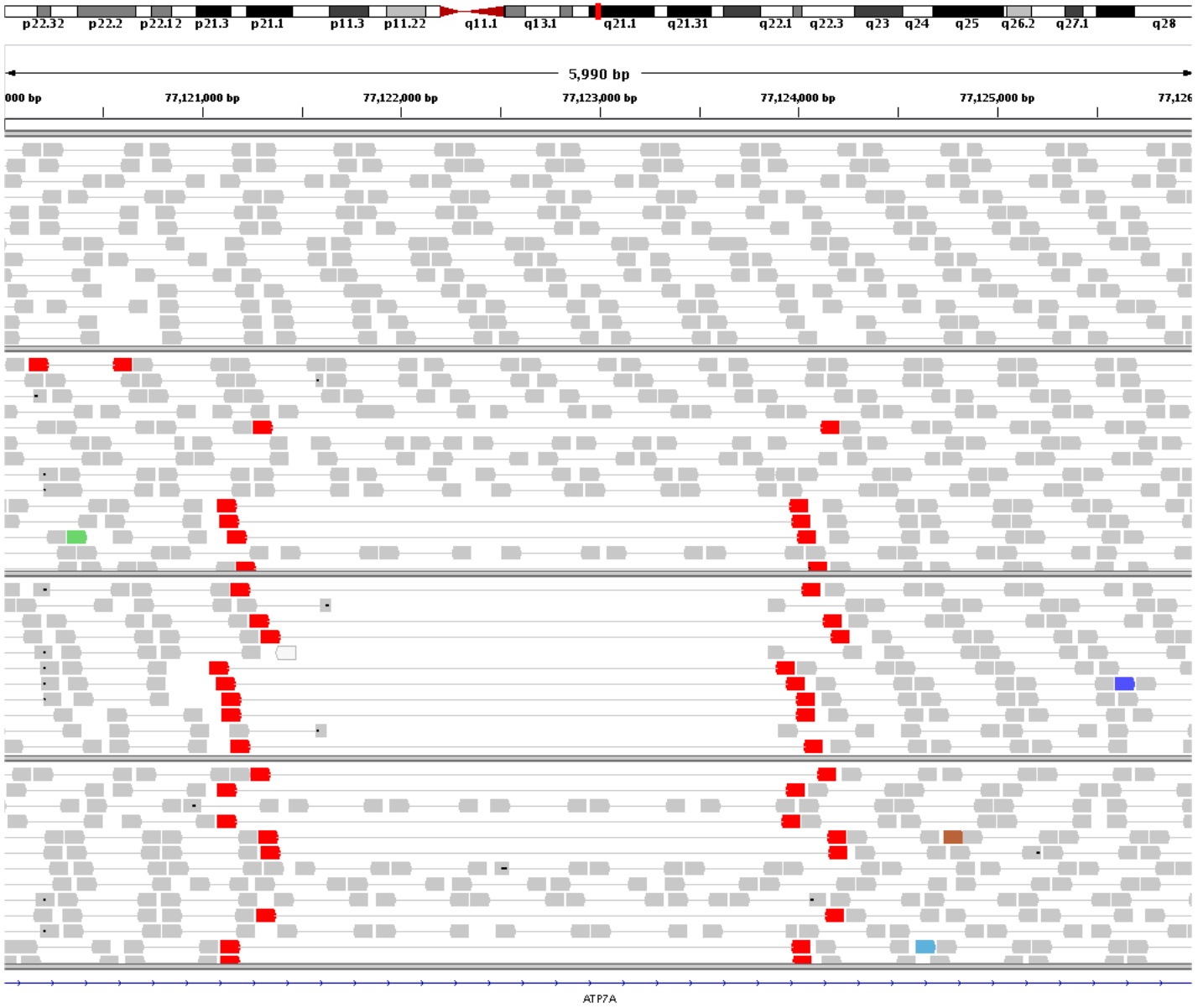
Composite patterns of SV

Father

Mother

Son

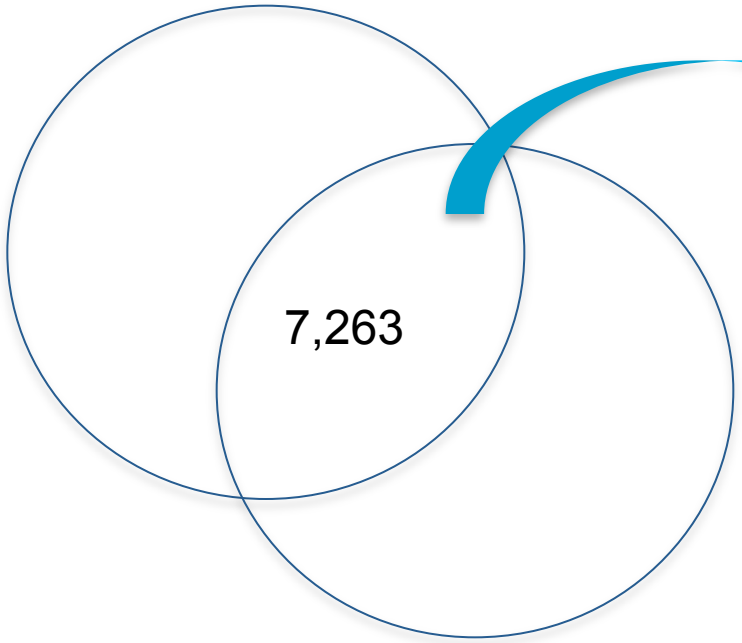
Daughter



Combining detection methods: a good idea!

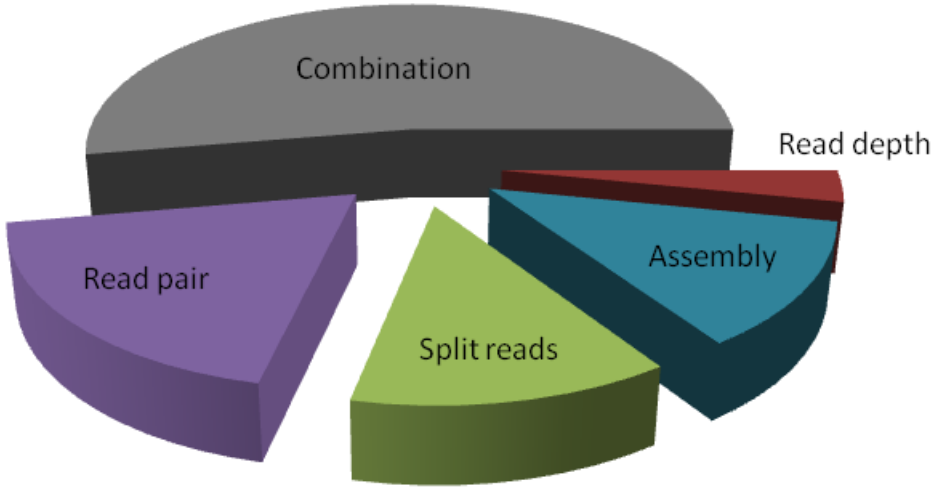
1000 Genomes
A Deep Catalog of Human Genetic Variation

1000 genomes project (phase 1)
21,541 deletions (>30bp)



Go•NL
GENOMEoftheNETHERLANDS

Detection methods



NextGen Sequencing: what do we get?

	Genome of NL	1000 genomes
Individuals	769 (250 families)	1092
SNPs	19.8 M	36.7 M
Small indels	1.4 M	1.4 M

Per individual genome (as compared to reference genome):

3.7M SNPs

360k short indels (1-20bp)

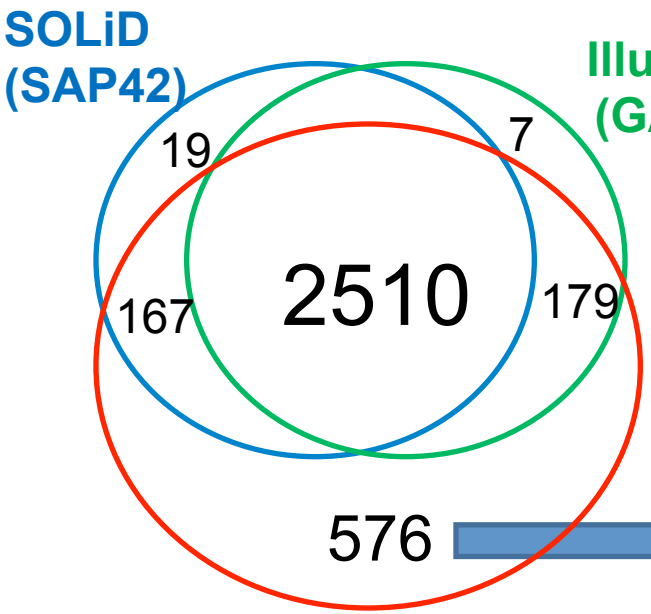
5.2k medium deletions (20 – 100 bp)

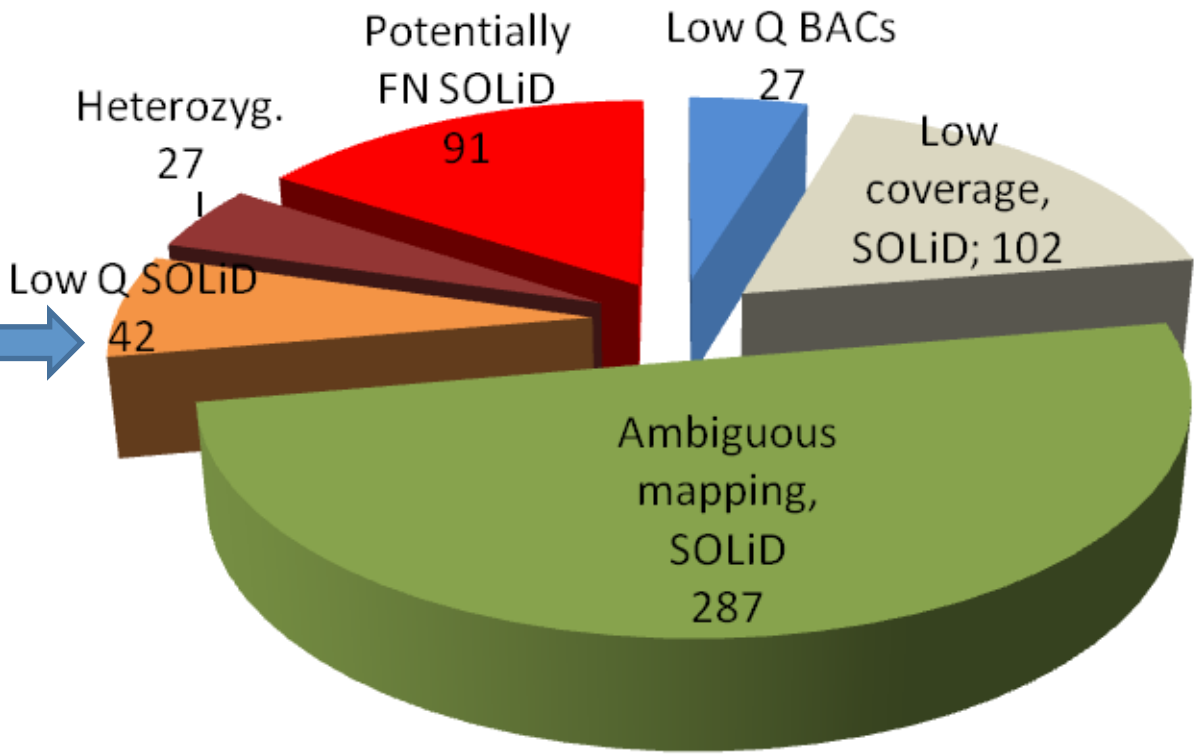
3.3k large deletions (100+ bp)

Go•NL
GENOMEoftheNETHERLANDS

NextGen Sequencing: what are we missing?

13 BACs (2.17 Mb)
Sanger Sequencing, Assembly
NGS: SOLiD and Solexa @ 20-25x

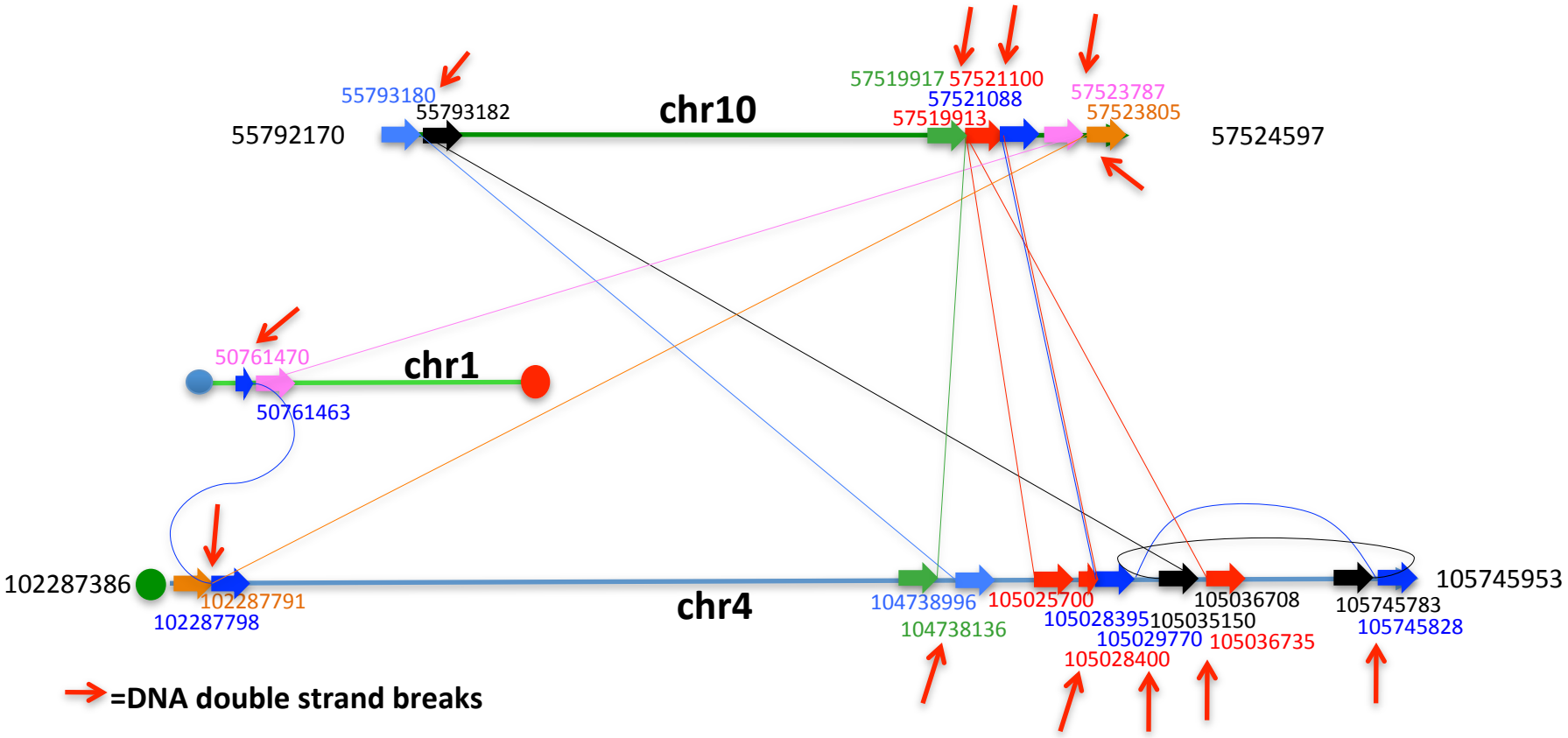




[HS rats sequencing consortium, 2013]

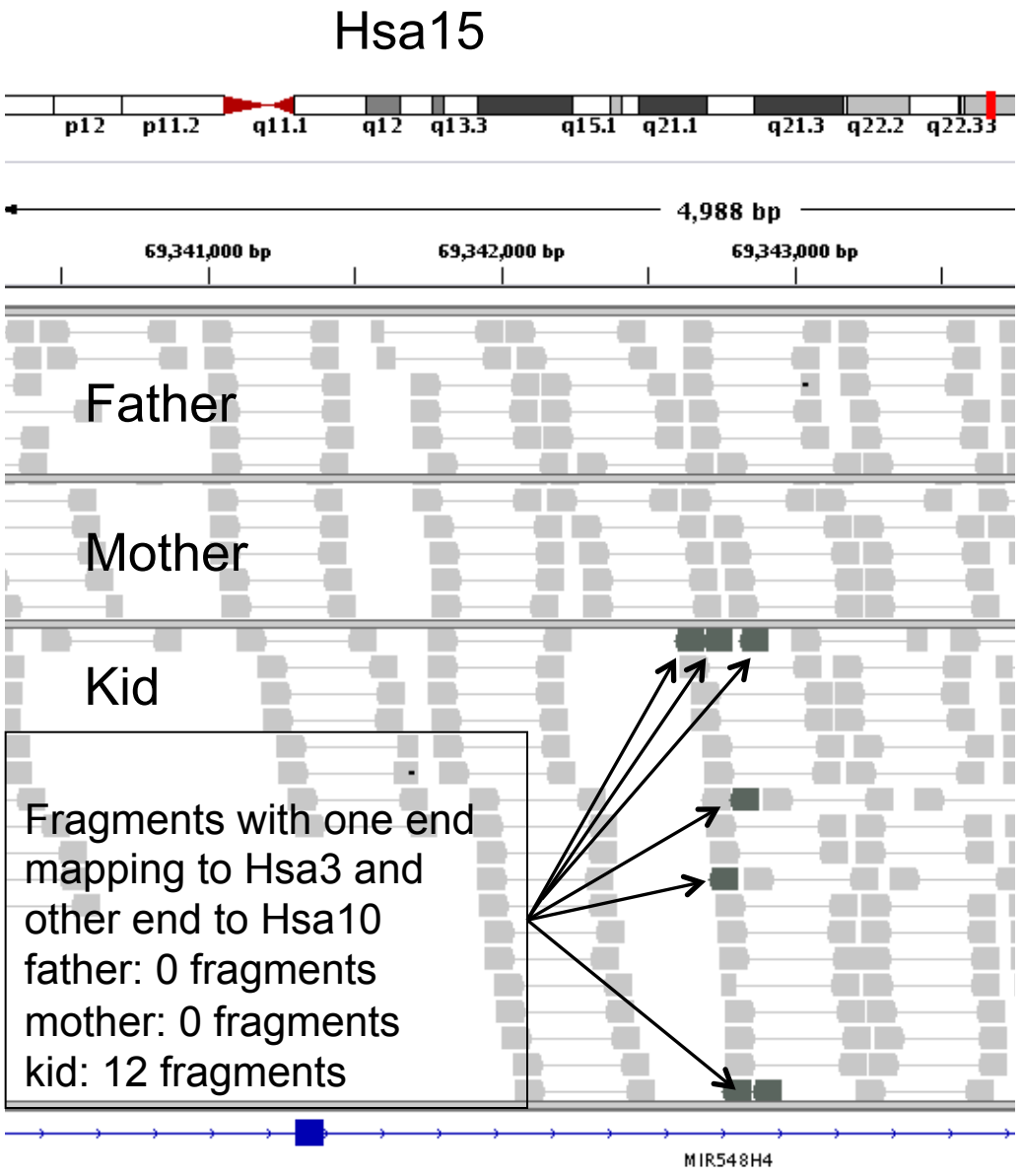
How SVs arise?

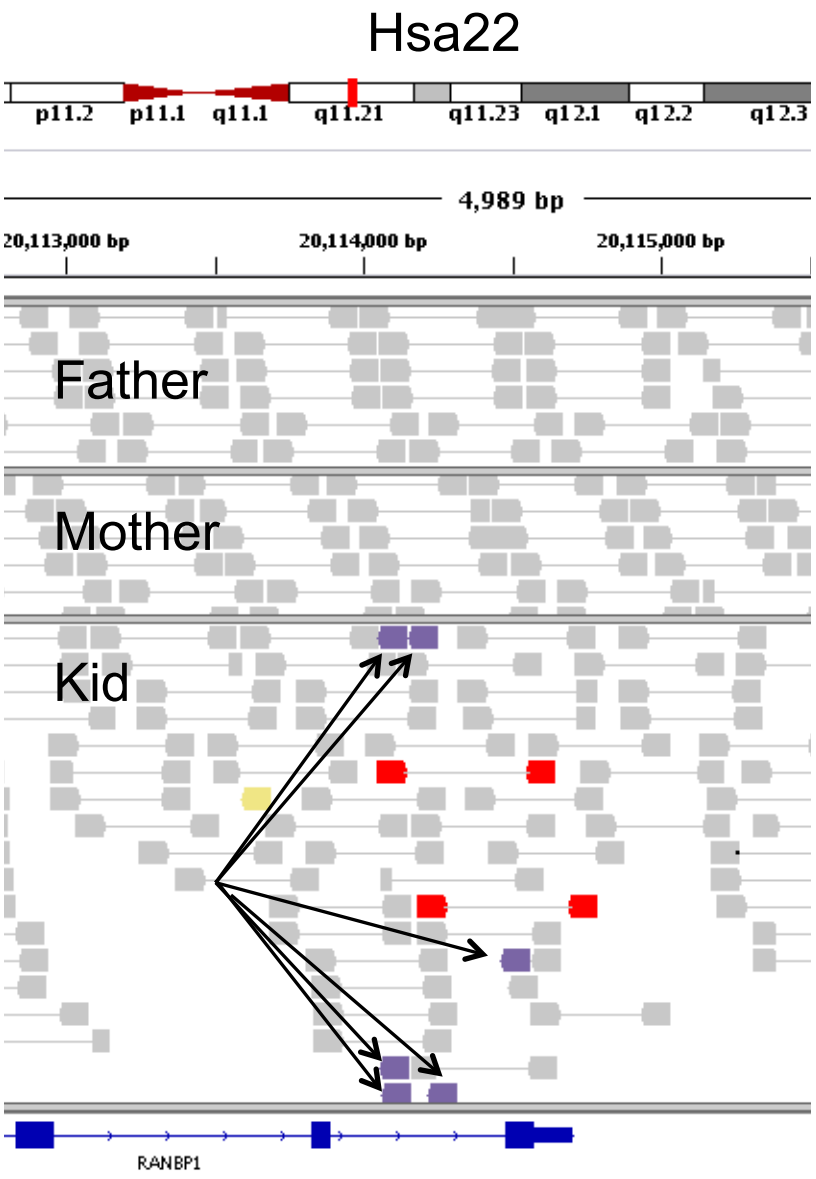
Complex structural variations, chromotrypsis



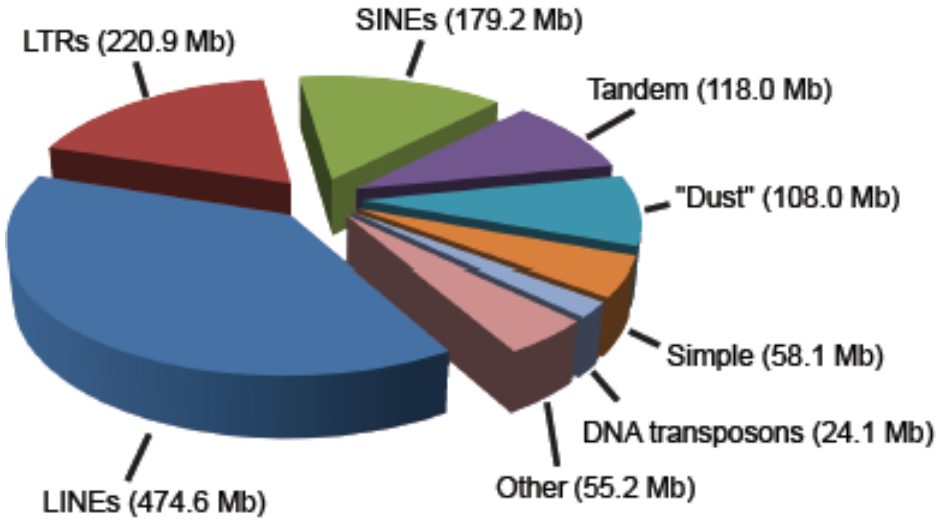
[Kloosterman et al, 2011]

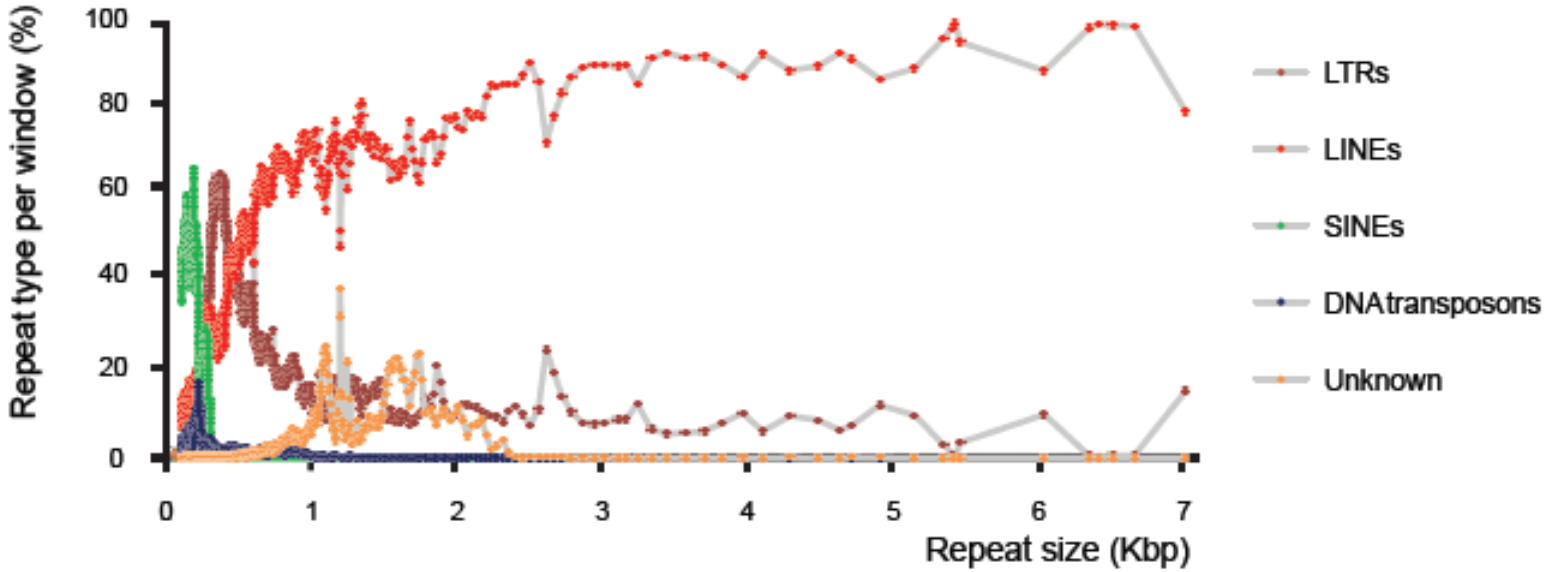
De novo SVs in healthy individuals





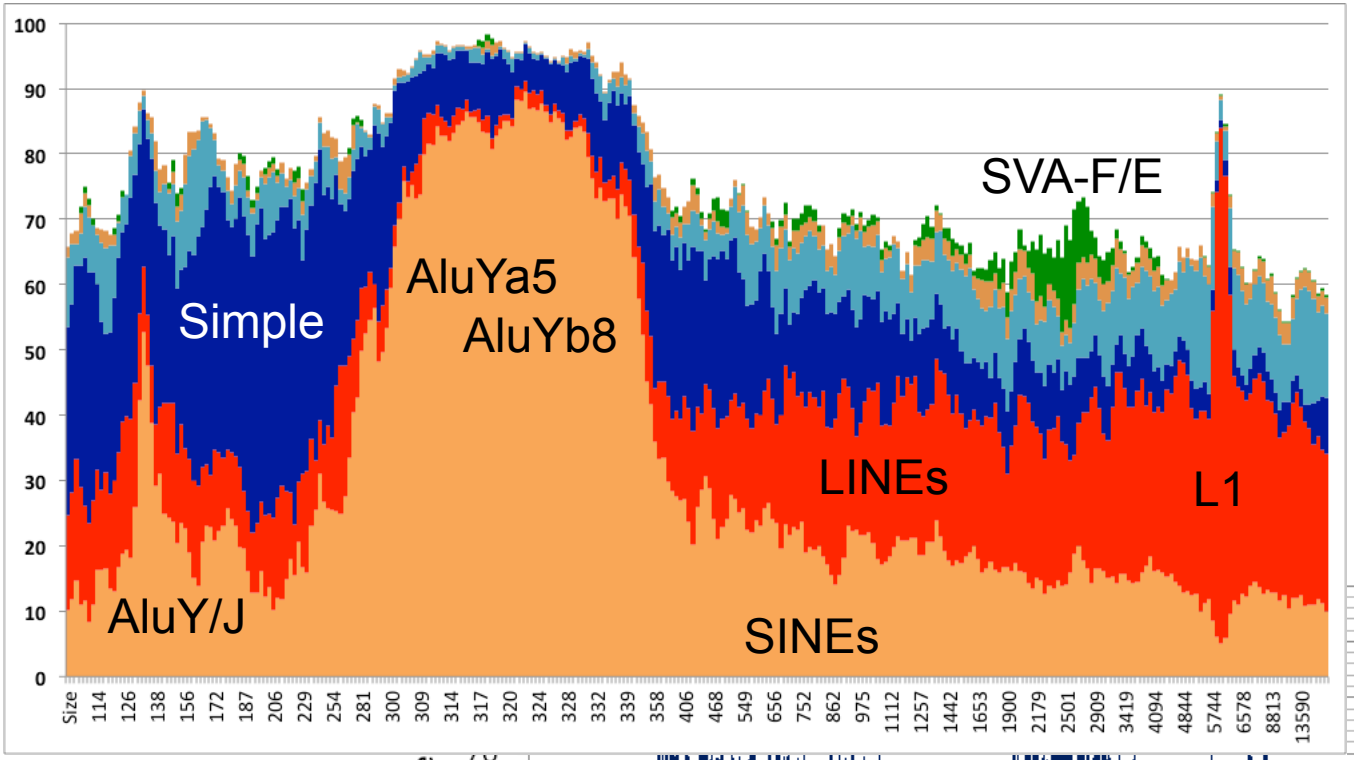
Repeats in mammalian genomes

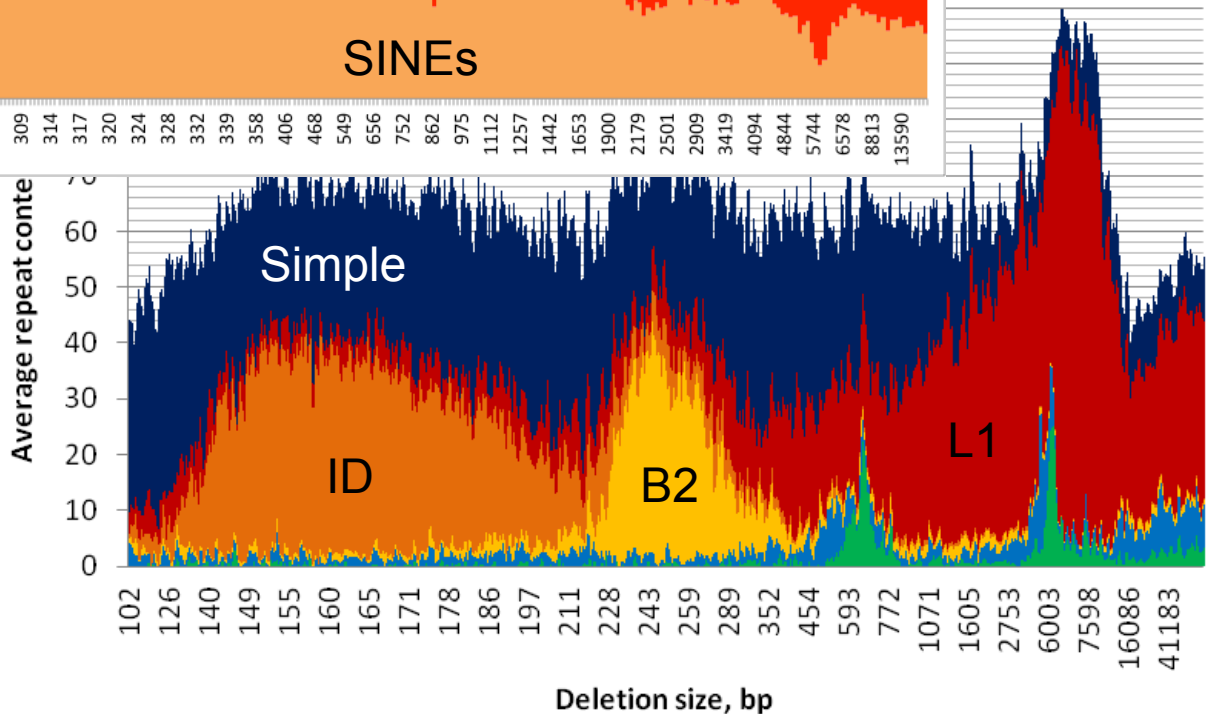




ERIBA European Research Institute for the Biology of Aging

Repeat instability: primary cause of SVs





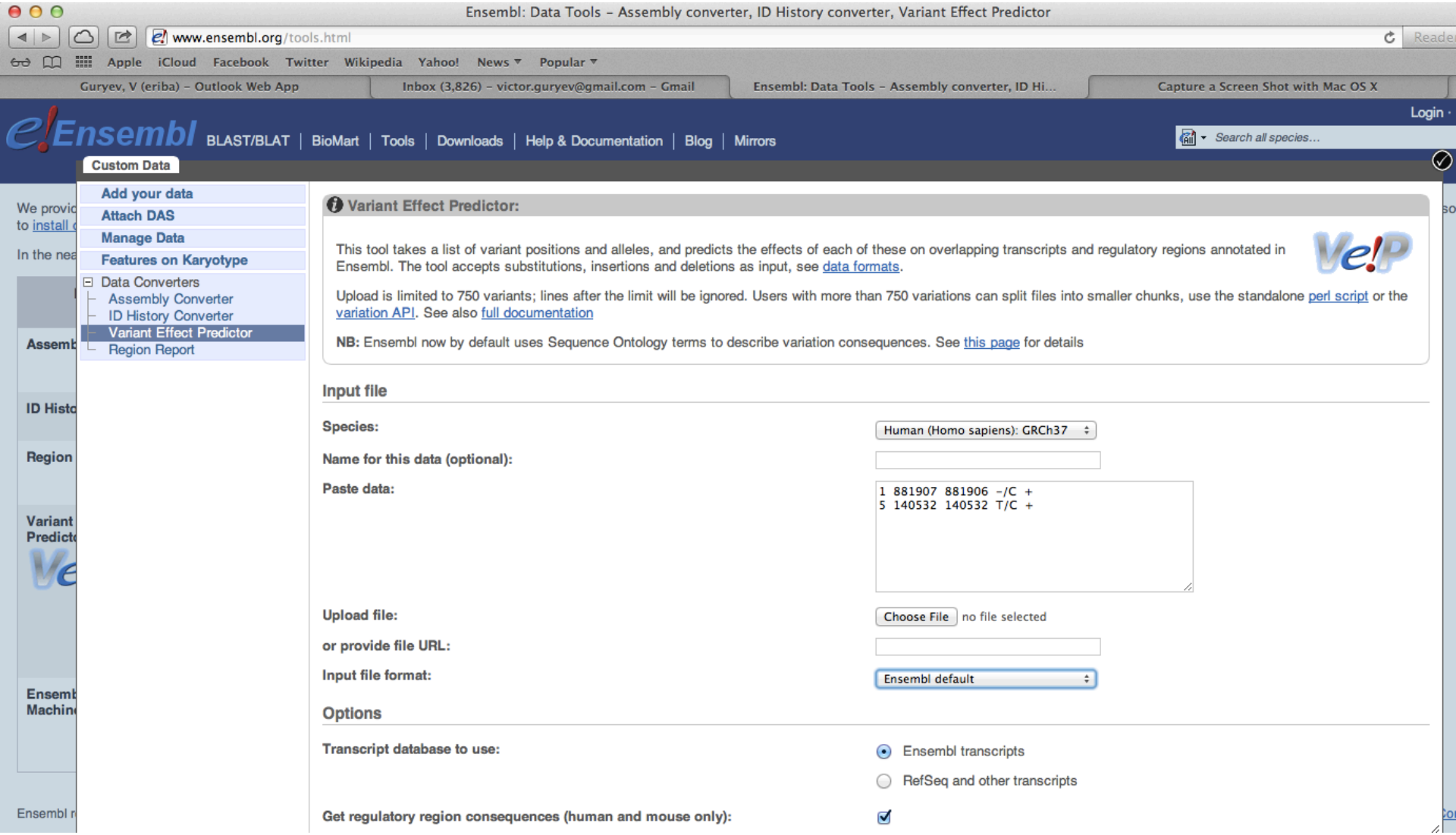
ERIBA European Research Institute for the Biology of Ageing

Representing SVs in VCF format

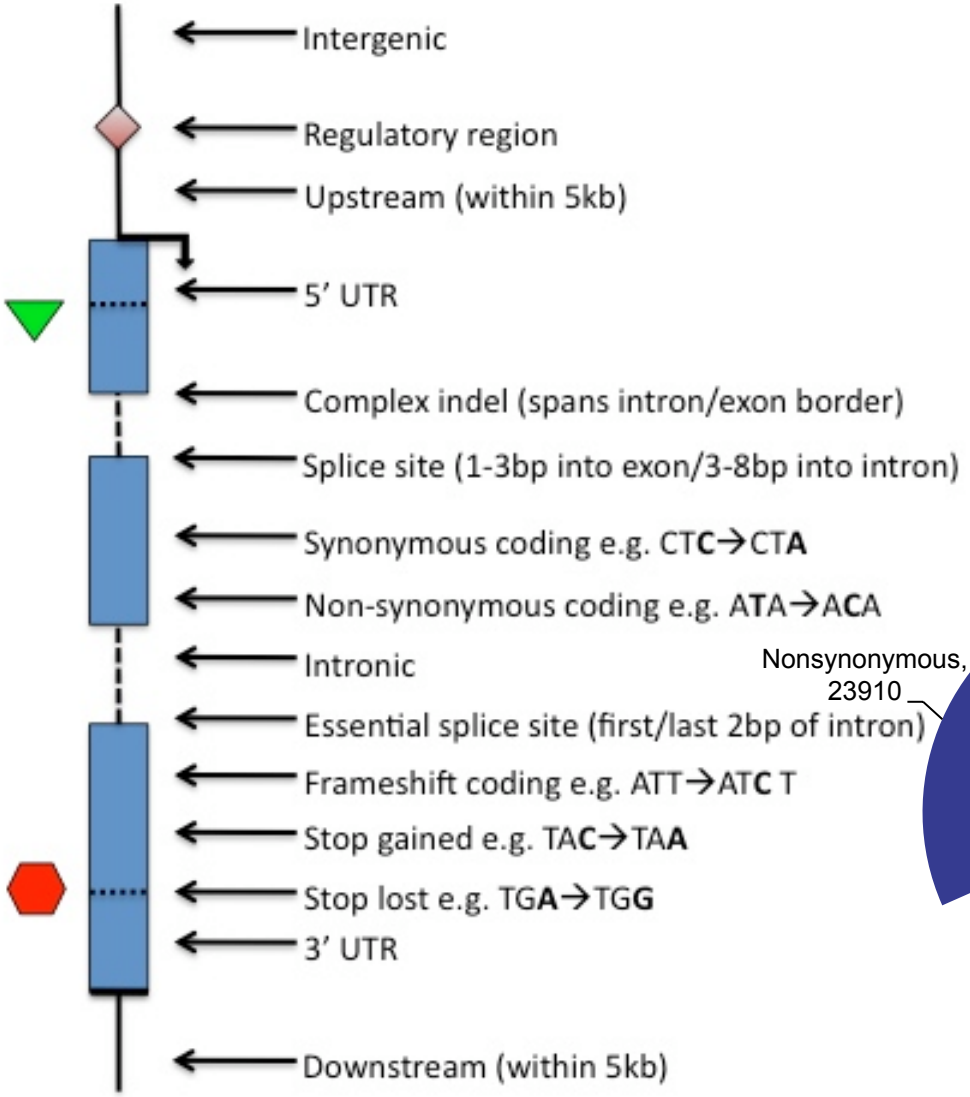
```
##ALT=<ID=DEL,Description="Deletion">
##ALT=<ID=DEL:ME:ALU,Description="Deletion of ALU element">
##ALT=<ID=DEL:ME:L1,Description="Deletion of L1 element">
##ALT=<ID=DUP,Description="Duplication">
##ALT=<ID=DUP:TANDEM,Description="Tandem Duplication">
##ALT=<ID=INS,Description="Insertion of novel sequence">
##ALT=<ID=INS:ME:ALU,Description="Insertion of ALU element">
##ALT=<ID=INS:ME:L1,Description="Insertion of L1 element">
##ALT=<ID=INV,Description="Inversion">
##ALT=<ID=CNV,Description="Copy number variable region">
##FORMAT=<ID=GT,Number=1,Type=Integer,Description="Genotype">
##FORMAT=<ID=GQ,Number=1,Type=Float,Description="Genotype quality">
##FORMAT=<ID=CN,Number=1,Type=Integer,Description="Copy number genotype for imprecise events">
##FORMAT=<ID=CNQ,Number=1,Type=Float,Description="Copy number genotype quality for imprecise">
#CHROM  POS    ID  REF ALT    QUAL  FILTER  INFO  FORMAT  NA00001
1 2827693    . CCGTGGATGCGGGGACCCGCATCCCCTCTCCCTTCACAGCTGAGTGACCCACATCCCCTCTCCCCTCGCA  C . PASS \
   SVTYPE=DEL;END=2827680;BKPTID=Pindel_LCS_D1099159;HOMLEN=1;HOMSEQ=C;SVLEN=-66 GT:GQ 1/1:13.9
2 321682     . T <DEL>   6 PASS    IMPRECISE;SVTYPE=DEL;END=321887;SVLEN=-105;CIPOS=-56,20; \
   CIEND=-10,62  GT:GQ 0/1:12
2 14477084   . C <DEL:ME:ALU>  12  PASS  IMPRECISE;SVTYPE=DEL;END=14477381;SVLEN=-297;
   MEINFO=AluYa5,5,307,+;CIPOS=-22,18;CIEND=-12,32  GT:GQ 0/1:12
3 9425916    . C <INS:ME:L1> 23  PASS  IMPRECISE;SVTYPE=INS;END=9425916;SVLEN=6027;CIPOS=-16,22;\
   MIINFO=L1HS,1,6025,- GT:GQ 1/1:15
3 12665100   . A <DUP>   14  PASS  IMPRECISE;SVTYPE=DUP;END=12686200;SVLEN=21100;CIPOS=-500,500;\
   CIEND=-500,500   GT:GQ:CN:CNQ  ./.:0:3:16.2
4 18665128   . T <DUP:TANDEM>  11  PASS  IMPRECISE;SVTYPE=DUP;END=18665204;SVLEN=76;CIPOS=-10,10;\
  CIEND=-10,10  GT:GQ:CN:CNQ  ./.:0:5:8.3
```

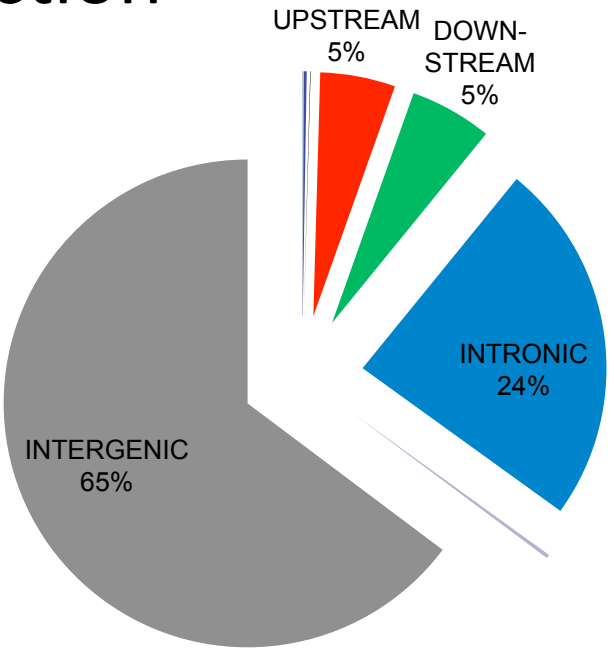
http://www.1000genomes.org

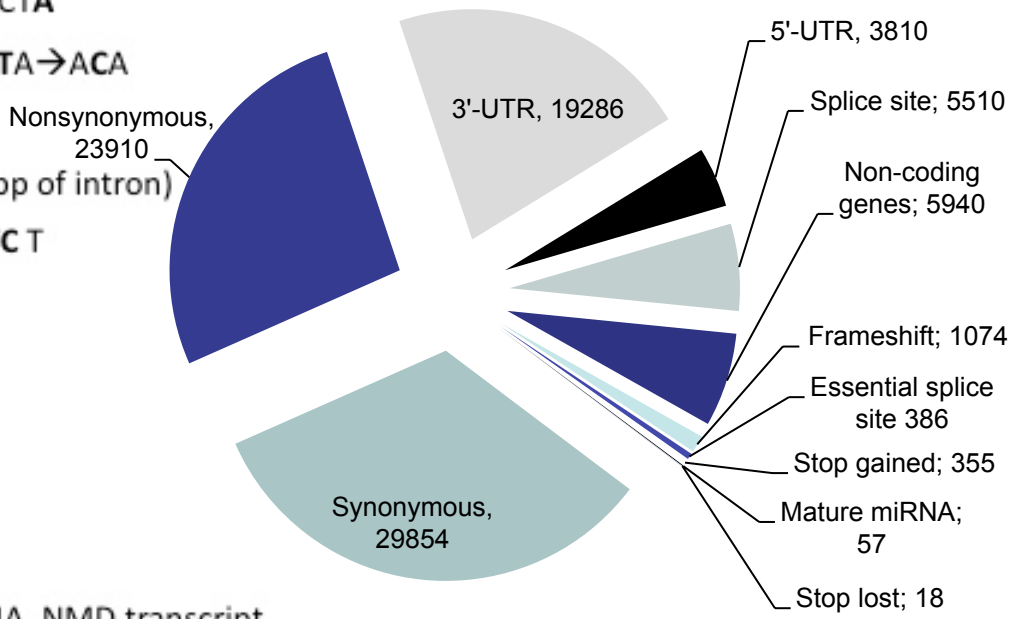
Variant annotation: Variant Effect Predictor (VEP)



Genome variation, function

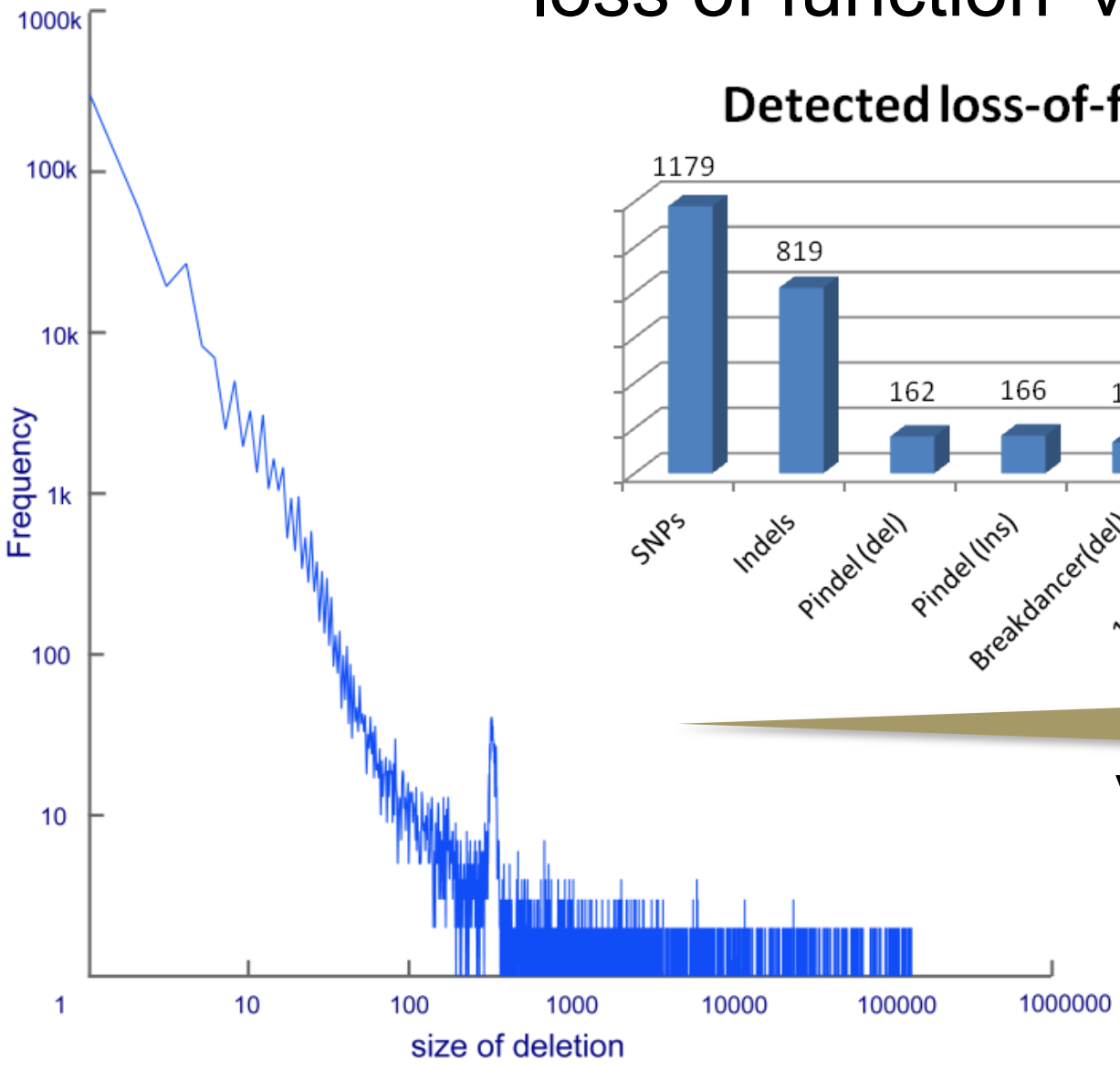






Others: Within non-coding gene, Within mature miRNA, NMD transcript

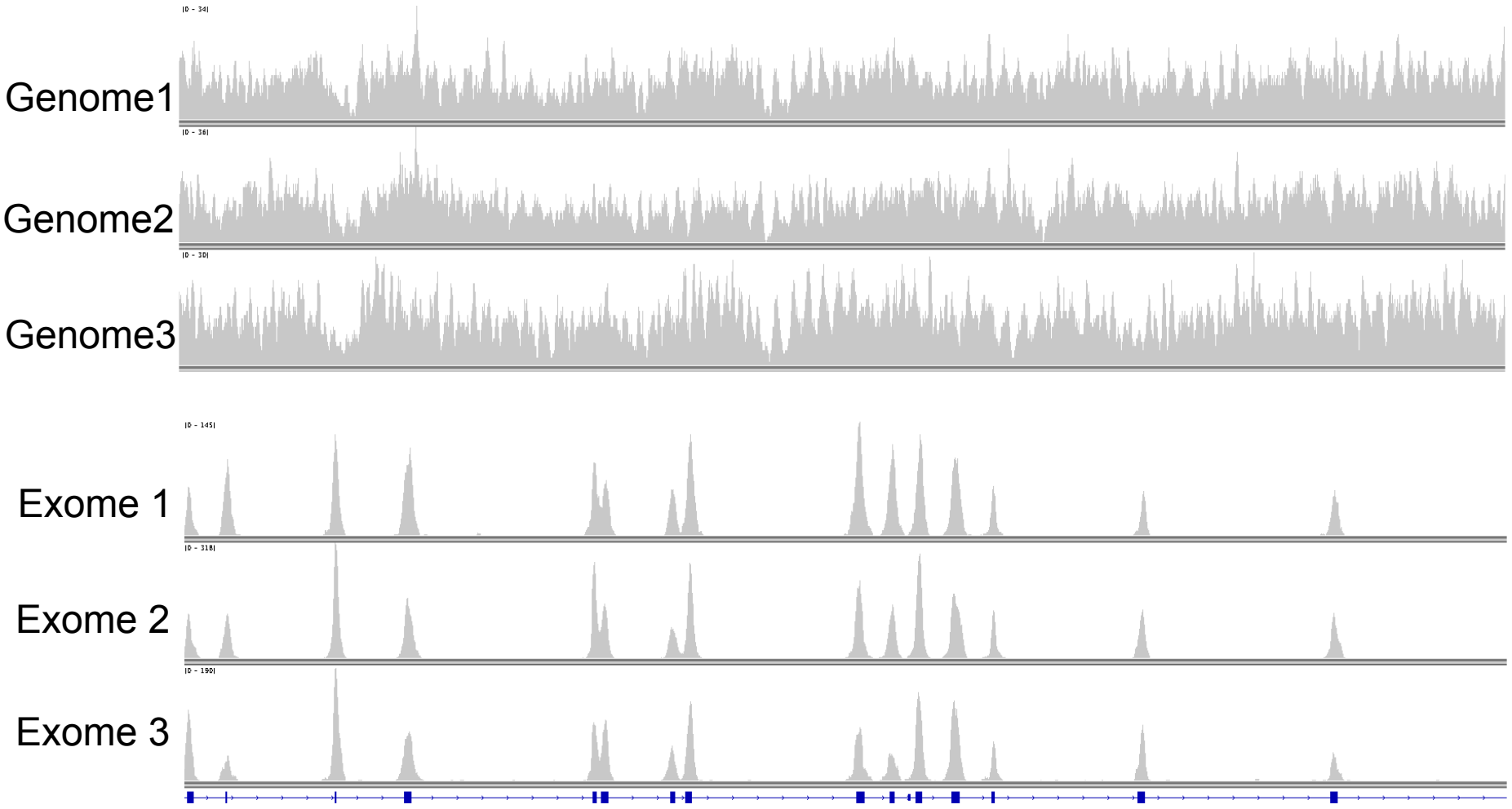
Contribution of different variation types to ‘loss of function’ variants



Variant size

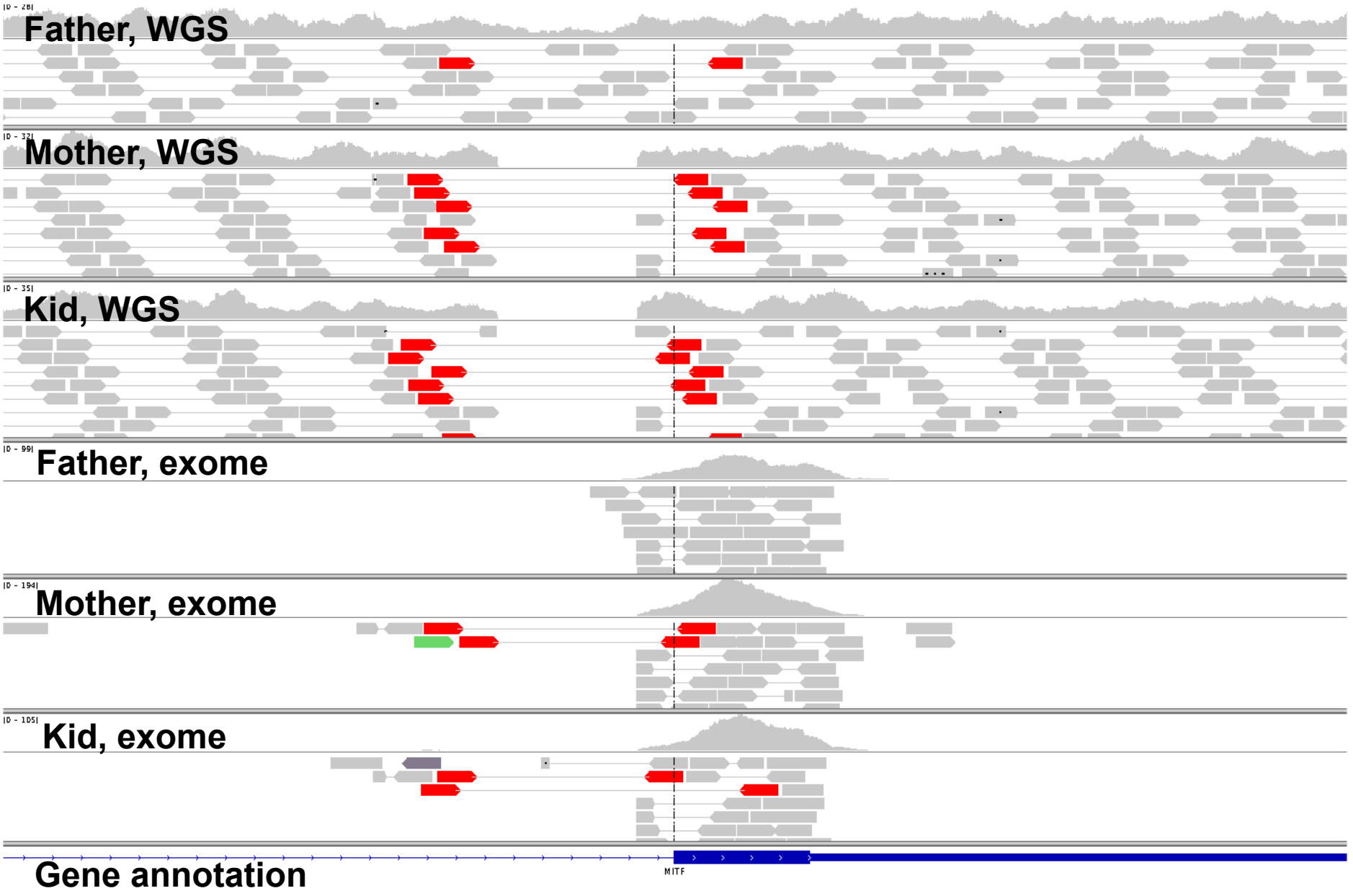
Why read whole genome if we only know function of a small proportion of our genomes?

Whole-genome sequencing vs exome-seq



ERIBA European Research Institute for the Biology of Ageing

Catching SVs with exome data



Validation strategies

Sanger Sequencing (golden standard)
de novo variants
false discovery rate
false negative rate

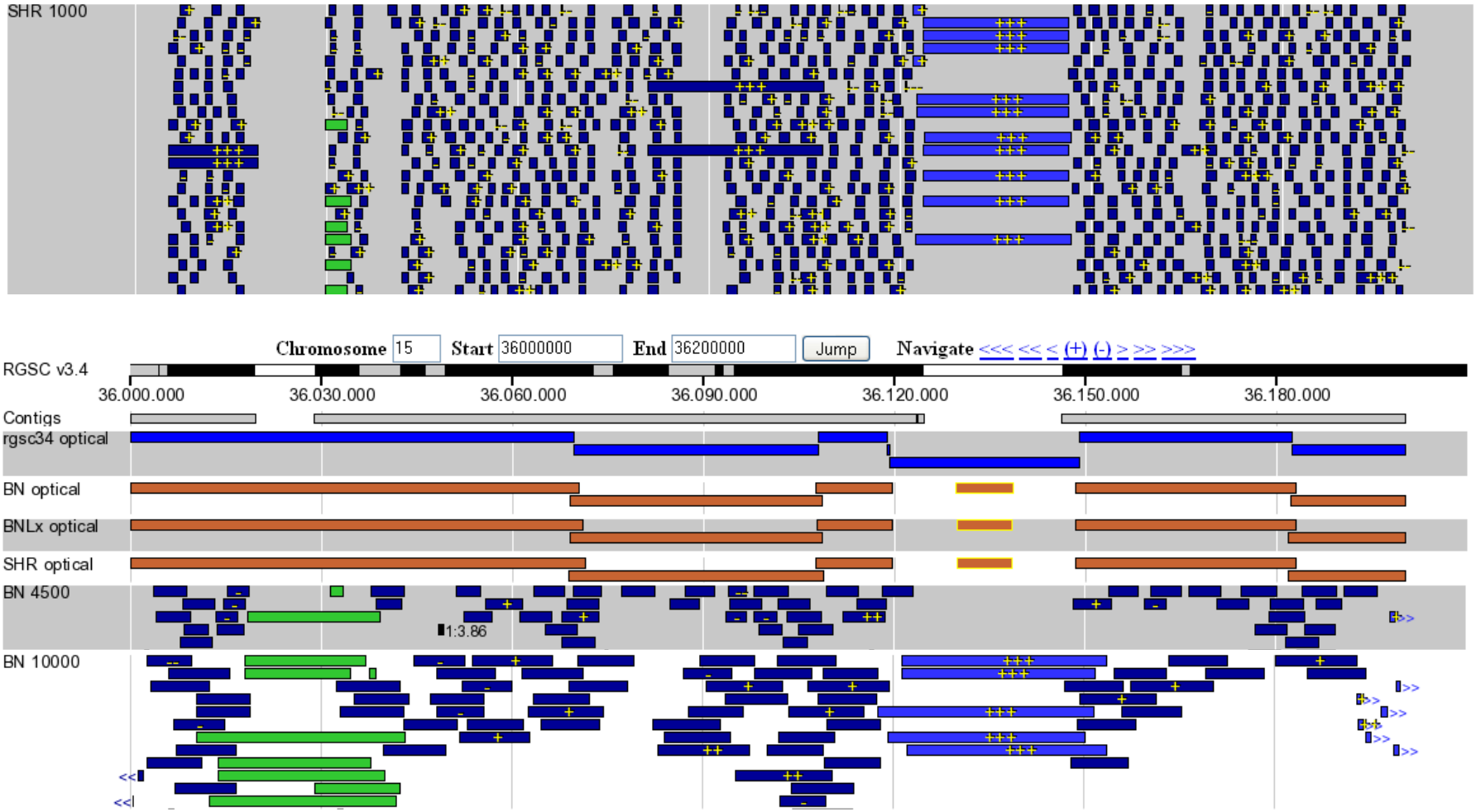
Targeted resequencing by enrichment (e.g. Ion Torrent, Fluidigm)
loss-of-function or non-synonymous alleles

SV breakpoint array
large (>30 bp) SVs with known breakpoints

FISH, aCGH
large SVs and translocations

ERIBA European Research Institute for the Biology of Ageing

Assembly quality is critical



Take home messages

1. **Prerequisites**: Genome assembly quality

2. **Study design**: # of libraries, type(PE, MP), insert sizes

3. **Quality control**: Insert size distribution, chimerism

4. **Combine methods** for SV discovery: read depth, read-pair, split-reads, *de novo* assembly

5. **Do verifications** (high FP rates)

Future directions

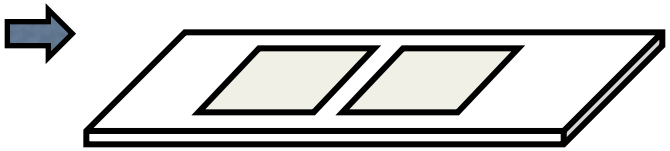
Longer reads:
PacBio (long reads, relatively low throughput)
Oxford nanopore (very long reads, easy preps)

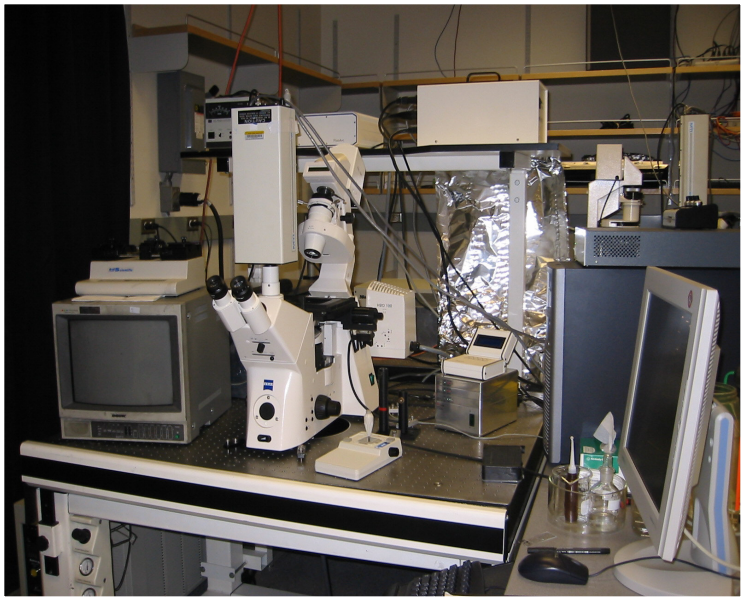
Lower prices:
Higher coverage = better calls, more samples, whole genome sequencing

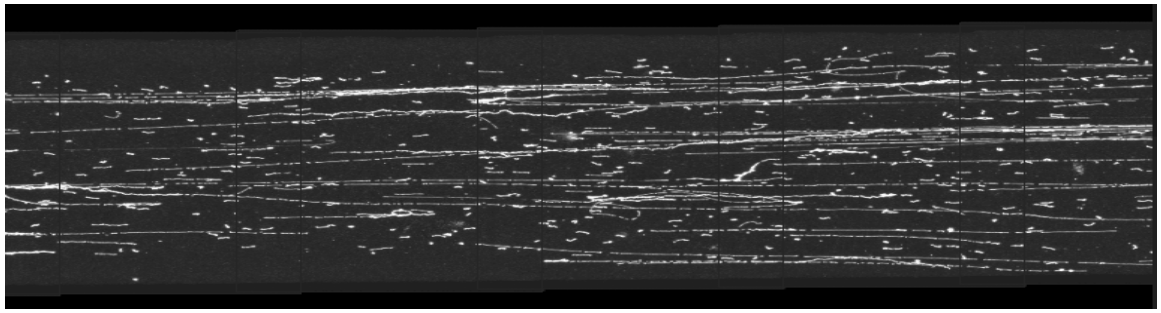
Even better and faster data analysis

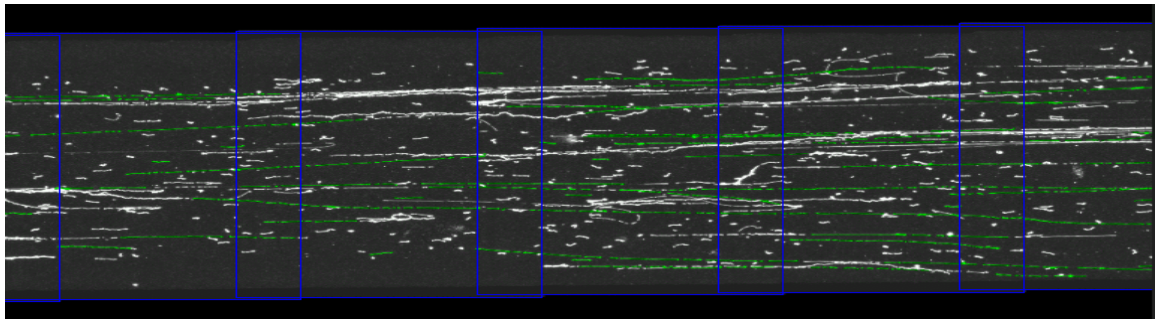
Optical mapping

High molecular weight DNA









Swa I restriction of genomic DNA

~ 20% missing cuts

1-5 extra “cuts” / Mbp

Acknowlegements

Cuppen Group

UMC Utrecht

Hubrecht Institute

Human genomes
Genome of the Netherlands

Go•NL
GENOMEofthe NETHERLANDS

nbic netherlands bioinformatics centre

ERIBA

Vacancies: PhD student, bioinformatics technician
Topic: Genome structure and ageing
Contact: v.guryev@umcg.nl