Human Genome Sequencing: The Next Step

Michael Snyder

August 11, 2010

Outline/Topics

- General introduction
- Human variation
- How to sequence a human genome
- How to interpret genome information
- Disease genome sequencing

Genotyping

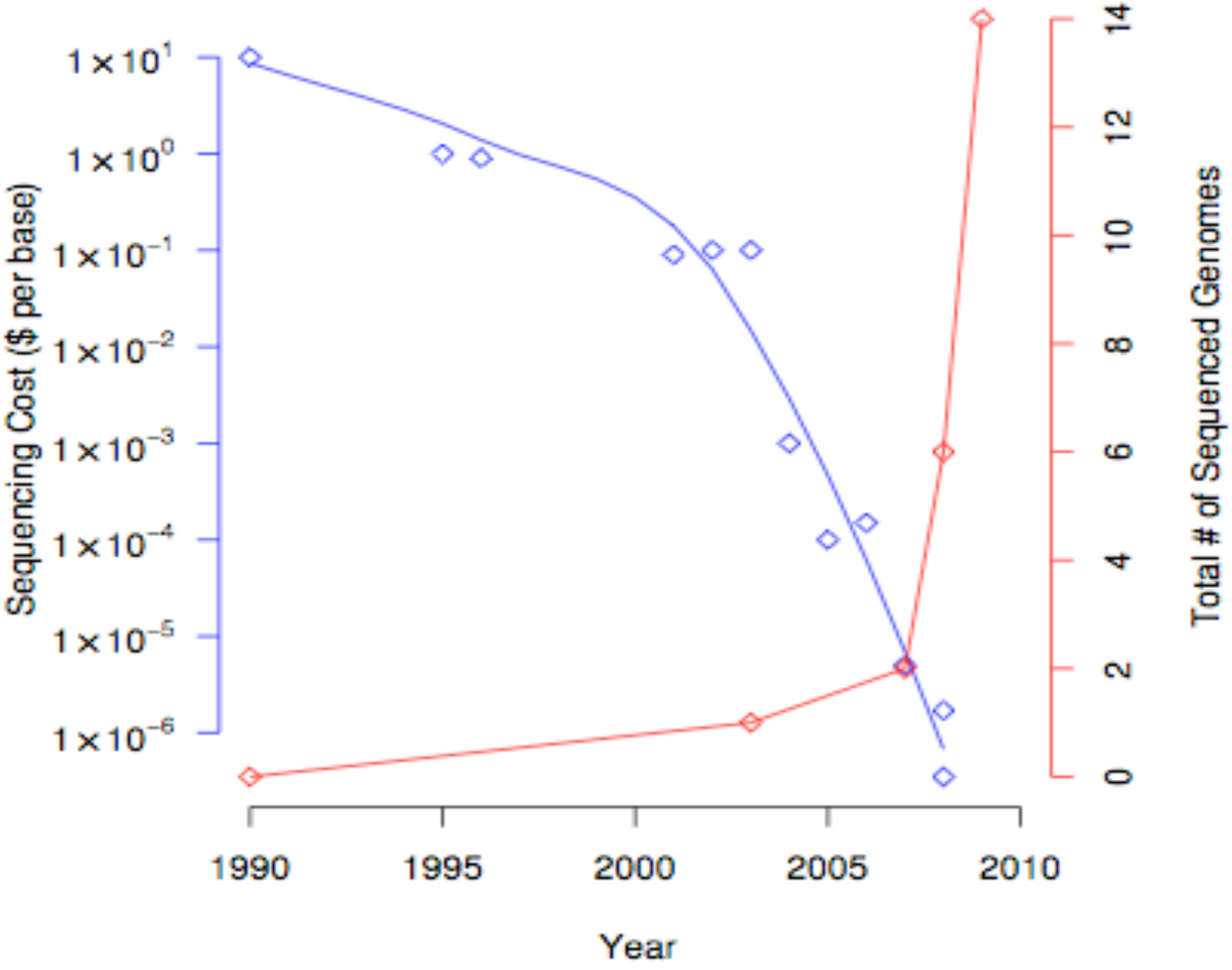
- Strengths
  - Easy and inexpensive
  - Good for ancestry
  - Some disease are linked to markers
- Limitations
  - Low resolution
  - Information limited; most phenotypes cannot be interpreted using genotyping

Phenotypes

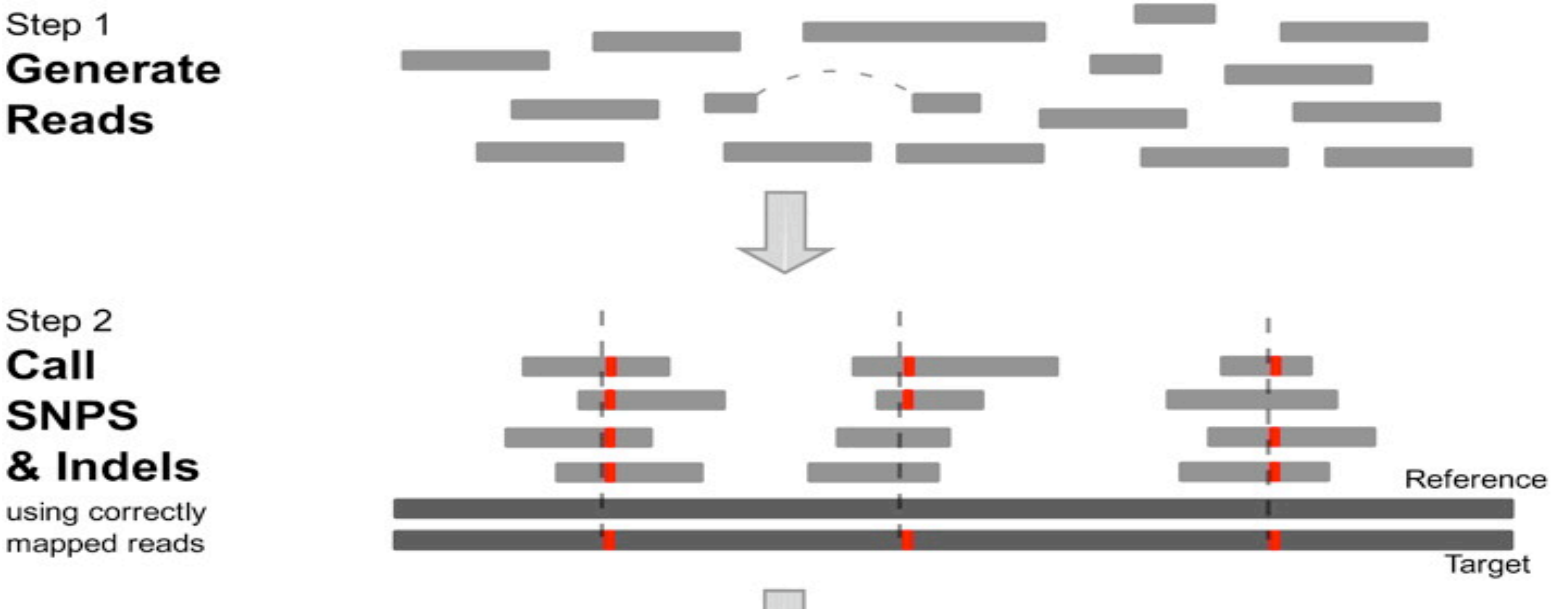
- Common alleles largely identified (Cystic fibrosis, sickle cell anemia)
  - generally apparent from family history

- Most phenotypes are complex (diabetes, neuropathies, height)
  - Likely due to rare alleles, combinations of alleles or both
  - Genome sequencing is an avenue for finding rare Mendelian alleles.

Sequencing Cost & Number of Sequenced genomes



Flow chart for determining a personal genome sequence



Snyder et al. Genes Dev 2010;24:423-431

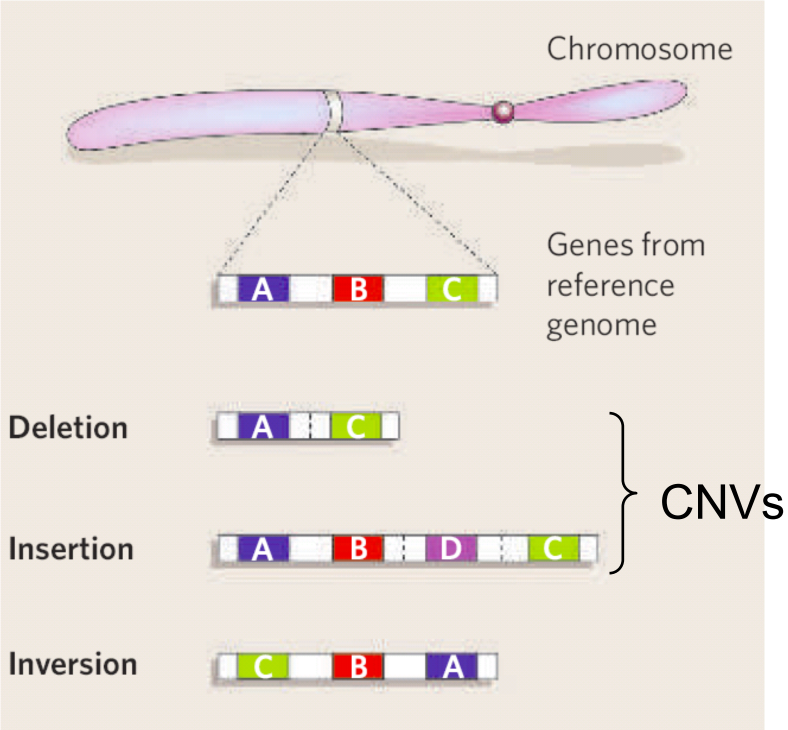
Genetic Variation Among People

Single nucleotide polymorphisms (SNPs)

GATTTAGATCGCGATAGAG
GATTTAGATCTCGATAGAG

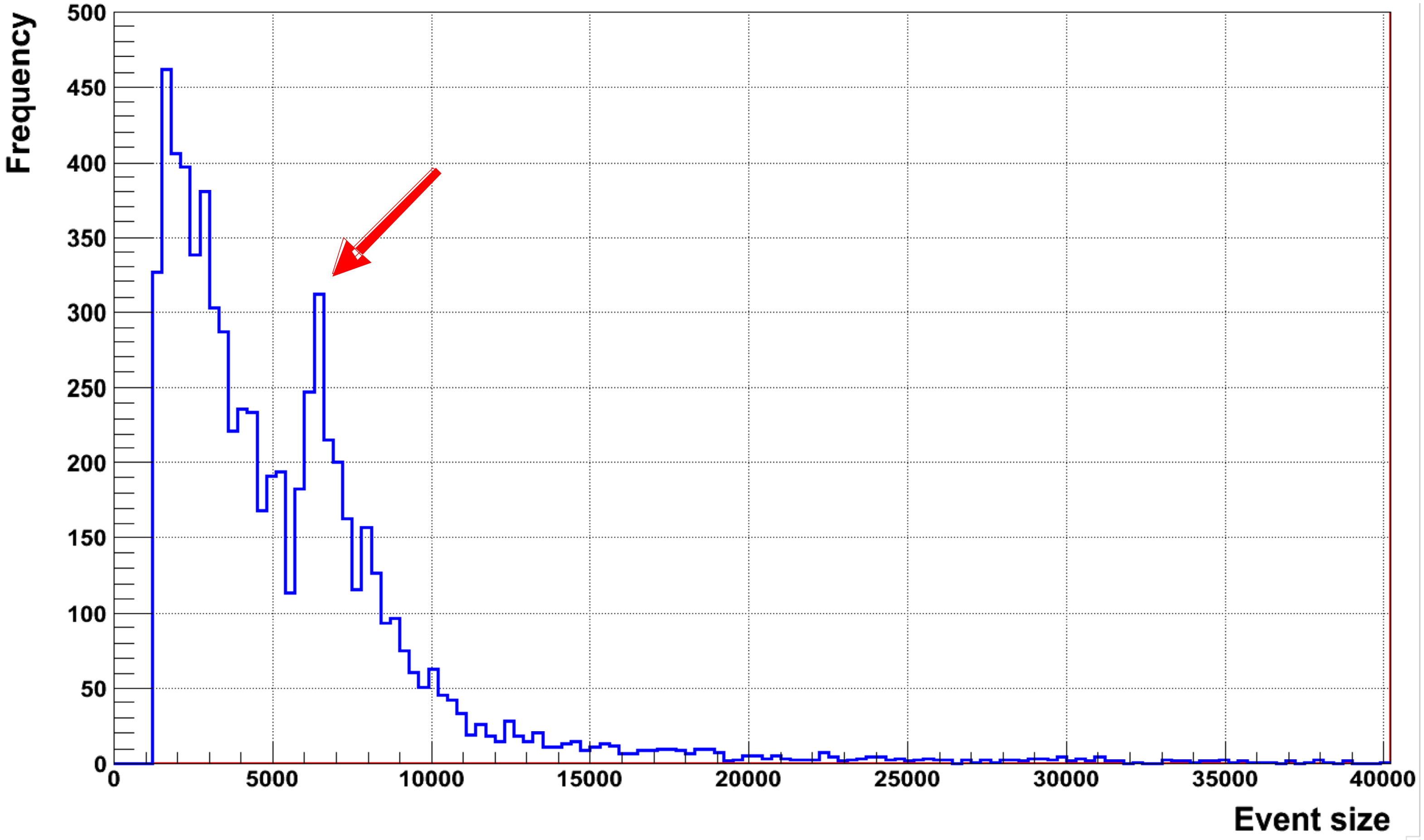
1/1200 differences among people

Mapping Structural Variation in Humans
>1 kb segments



- 3-4% of the human genome
- Likely involved in phenotype variation and disease
- Until recently most methods for detection were low resolution (>50 kb)

Size Distribution of CNV in a Human Genome



Most Genome Sequencing Projects Ignore SVs

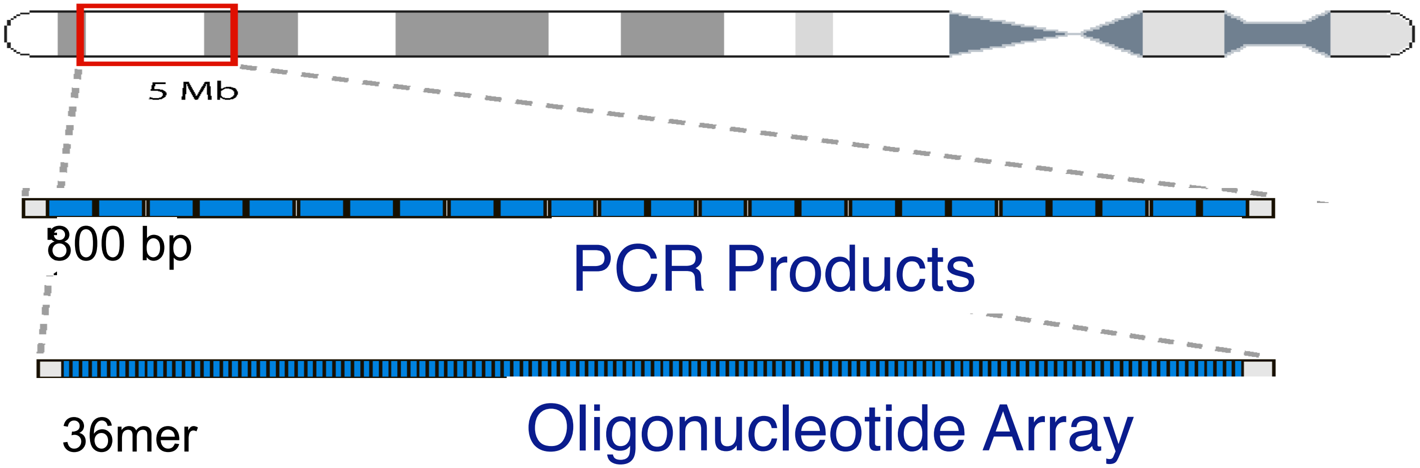
Project	Technology	Paired End	SNPs; Short Indel	SVs	New Seq.	Genotype	Reference
European-Venter	Sanger	Yes	3M; 0.3M	0.2M (> 1000bp)	1M	Limited	Levy et al., 2007
European-Watson	454	No	3M; 0.2M	Limited	No	No	Wheeler et al., 2008
European-Quake	Helicos	No	3M	Limited	No	No	Pushkarev et al., 2009
Asian	Illumina	Partially	3M; 0.1M	2.7K (>100bp)	No	No	Wang et al., 2008
HapMap Sample; Yoruban 18507	Illumina	Yes	4M; 10K	0.1K	No	No	Bentley et al., 2008
HapMap Sample; Yoruban 18507	SOLiD	Partially	4M; 0.2M	5.5K (unknown definition)	No	No	McKernan et al., 2009
Korean	Illumina	Yes	3M	Limited	No	No	Ahn et al., 2009
Korean- AK1	Illumina	Yes	3.45M; 0.17M	~300 CNVs	No	No	Kim et al., 2009
Three human genomes	Complete Genomics	Yes	3.2-4.5M; 0.3-0.5M	Limited (50-90K block substitutions)	No	Limited	Drmanac et al., 2009
AML genome & normal counterpart	Illumina	No	3.8M; 0.7K	Limited	No	No	Ley et al., 2008
AML genome	Illumina	Yes	64	Limited	No	No	Mardis et al., 2009
Melanoma genome	Illumina	Yes	32K;1K	51	No	No	Pleasance et al., 2009a
Lung cancer genome	SOLiD	Yes	23K; 65	392	No	No	Pleasance et al. 2009b

Why Are SVs Not Studied More?

- Often involves repeated regions (transposons, duplicated regions)

- Rearrangements are complex

Genome Tiling Arrays

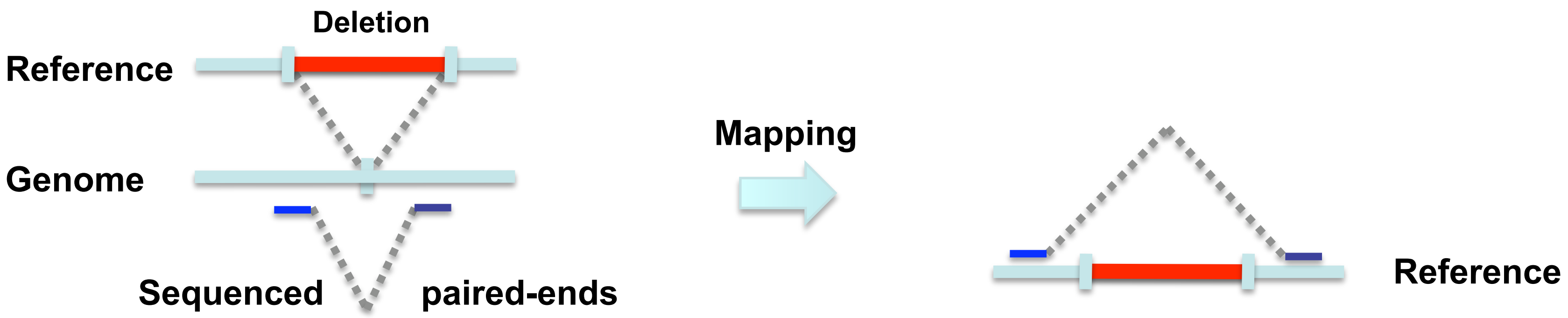


Massively Parallel Sequencing

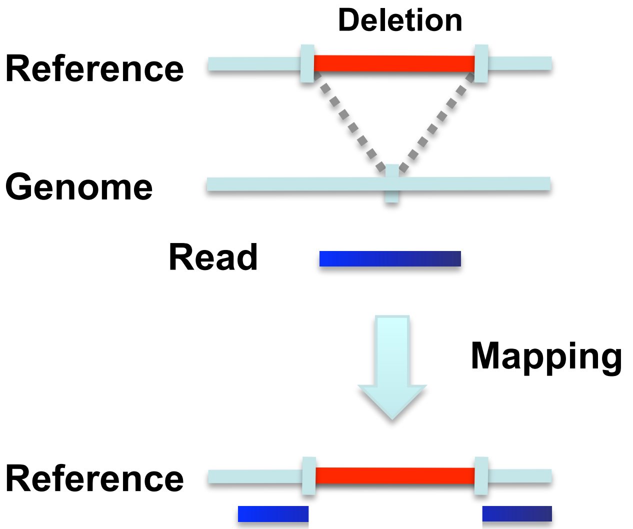
AGTTCACCTAAGA…
CTTGAATGCCGAT…
GTCATTCCGCAAT…

Methods to Find SVs

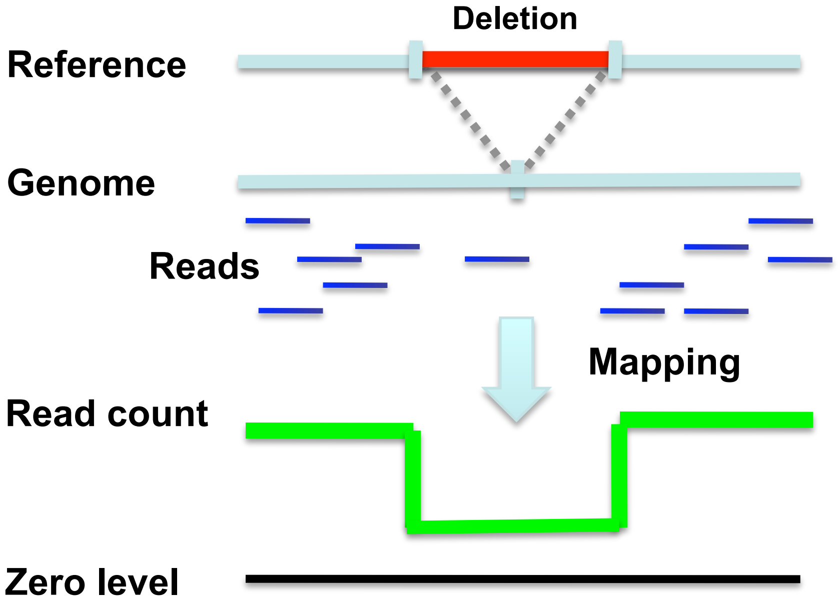
1. Paired ends



2. Split read



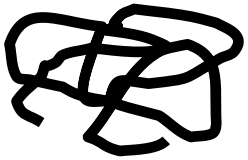
3. Read depth (or aCGH)



4. Match with database

[Snyder et al. Genes & Dev. 2010

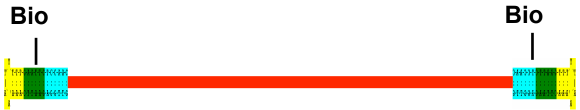
High Resolution-Paired-End Mapping (HR-PEM)



Genomic DNA

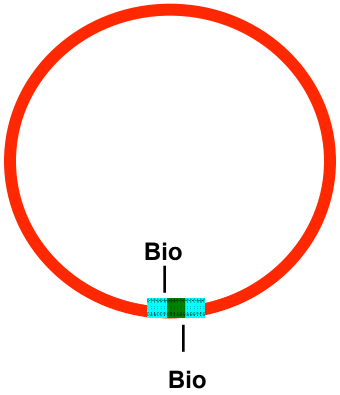
Shear to 3 kb

Adaptor ligation

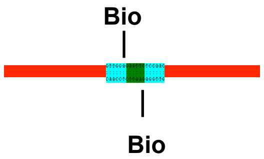


Fragments

Circularize



Random Cleavage

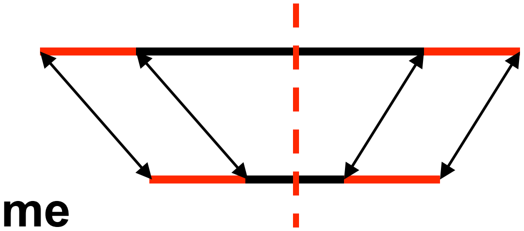


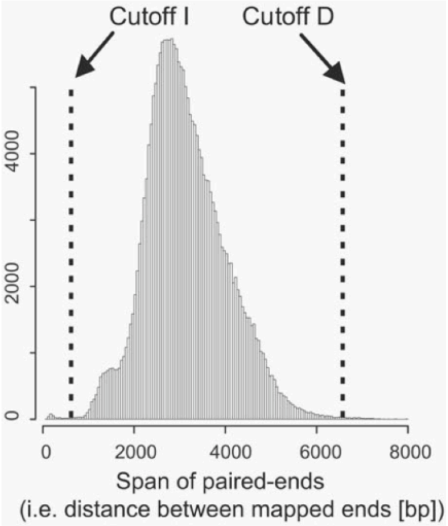
200-300bp

454 Sequencing

(250bp reads, 400K reads/run)

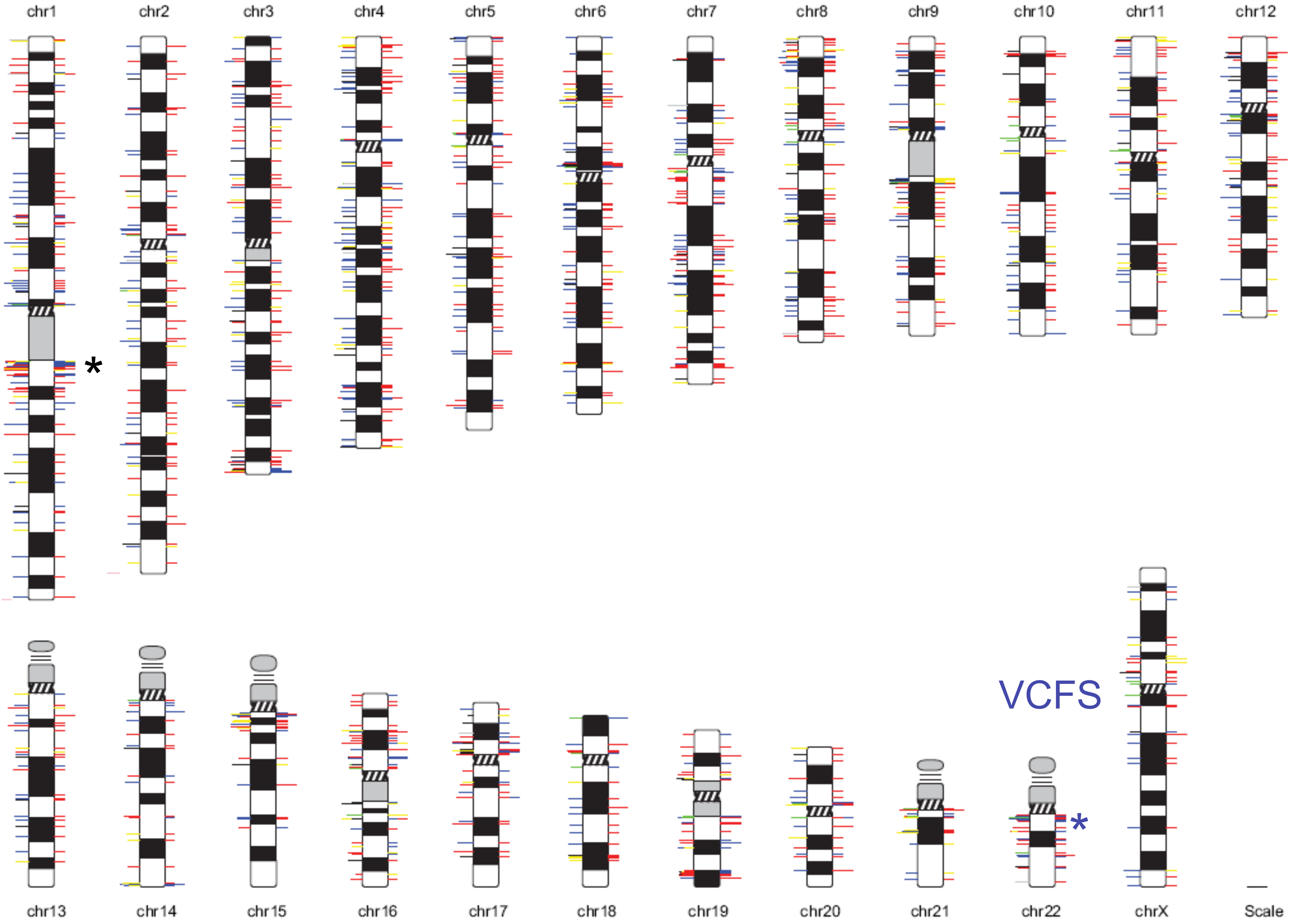
Map paired ends to reference genome



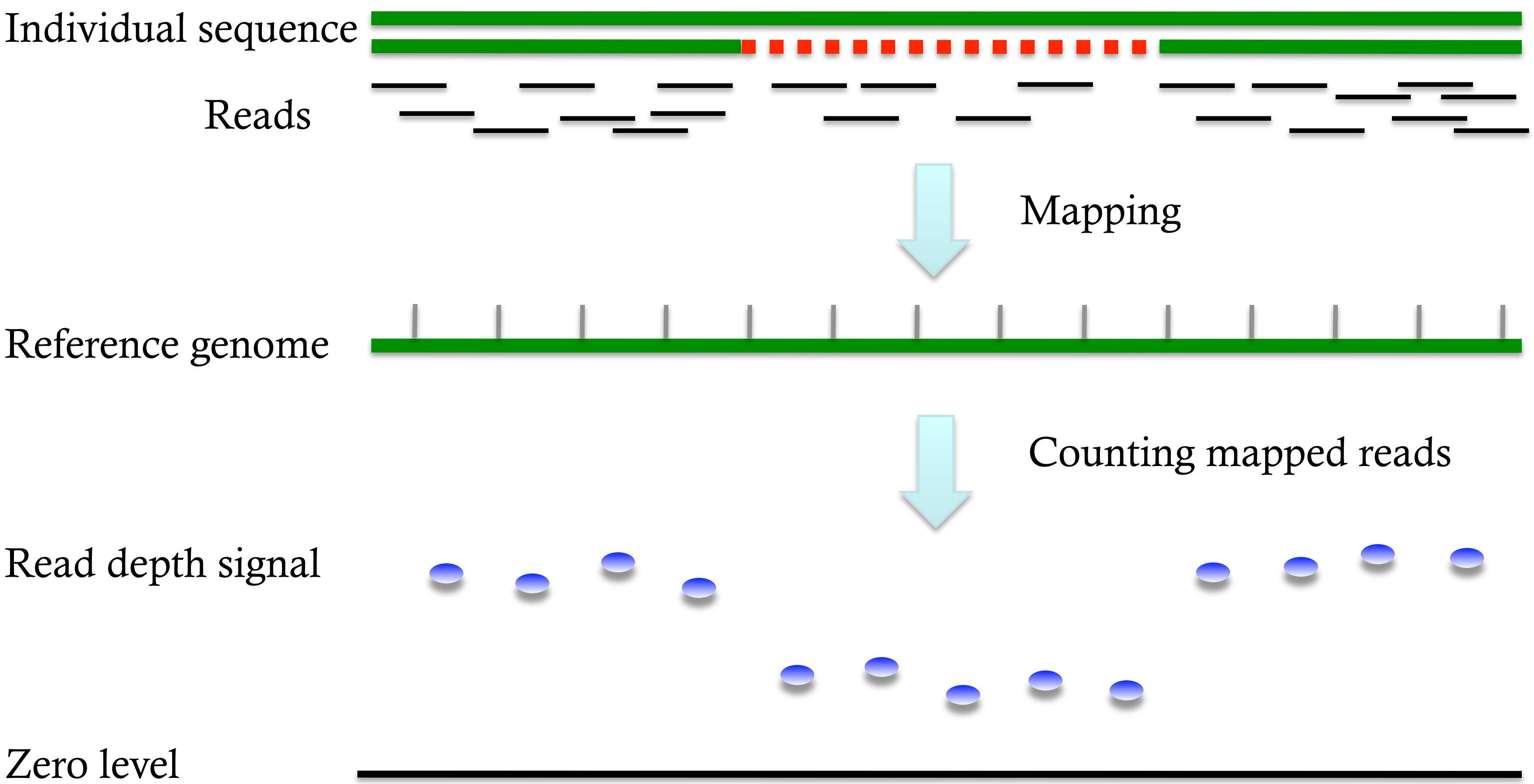


Korbel et al., 2007 Science

~1500 SVs >2.5kb per Person



Sequence Read Depth Analysis



Individual sequence

Reads

Mapping

Reference genome

Counting mapped reads

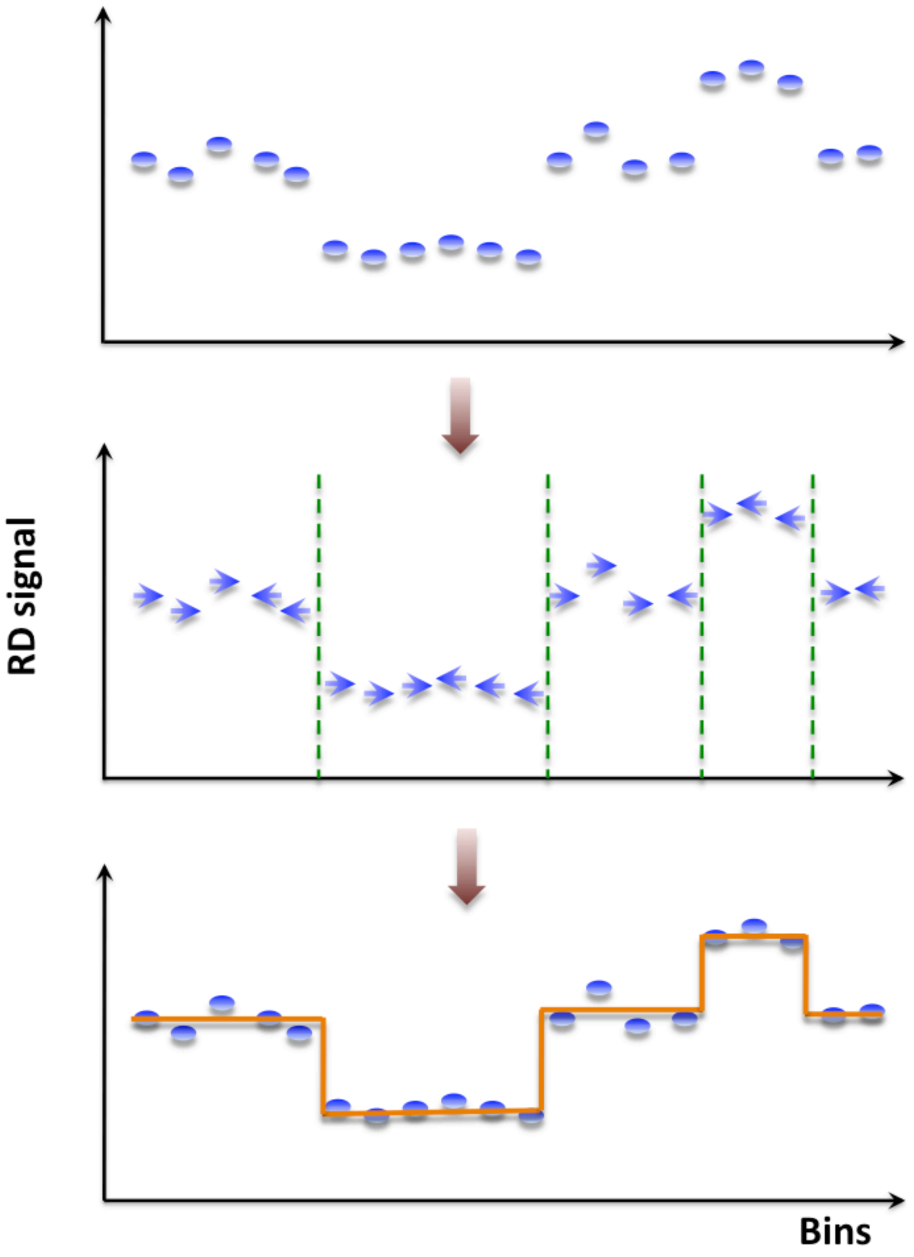
Read depth signal

Zero level

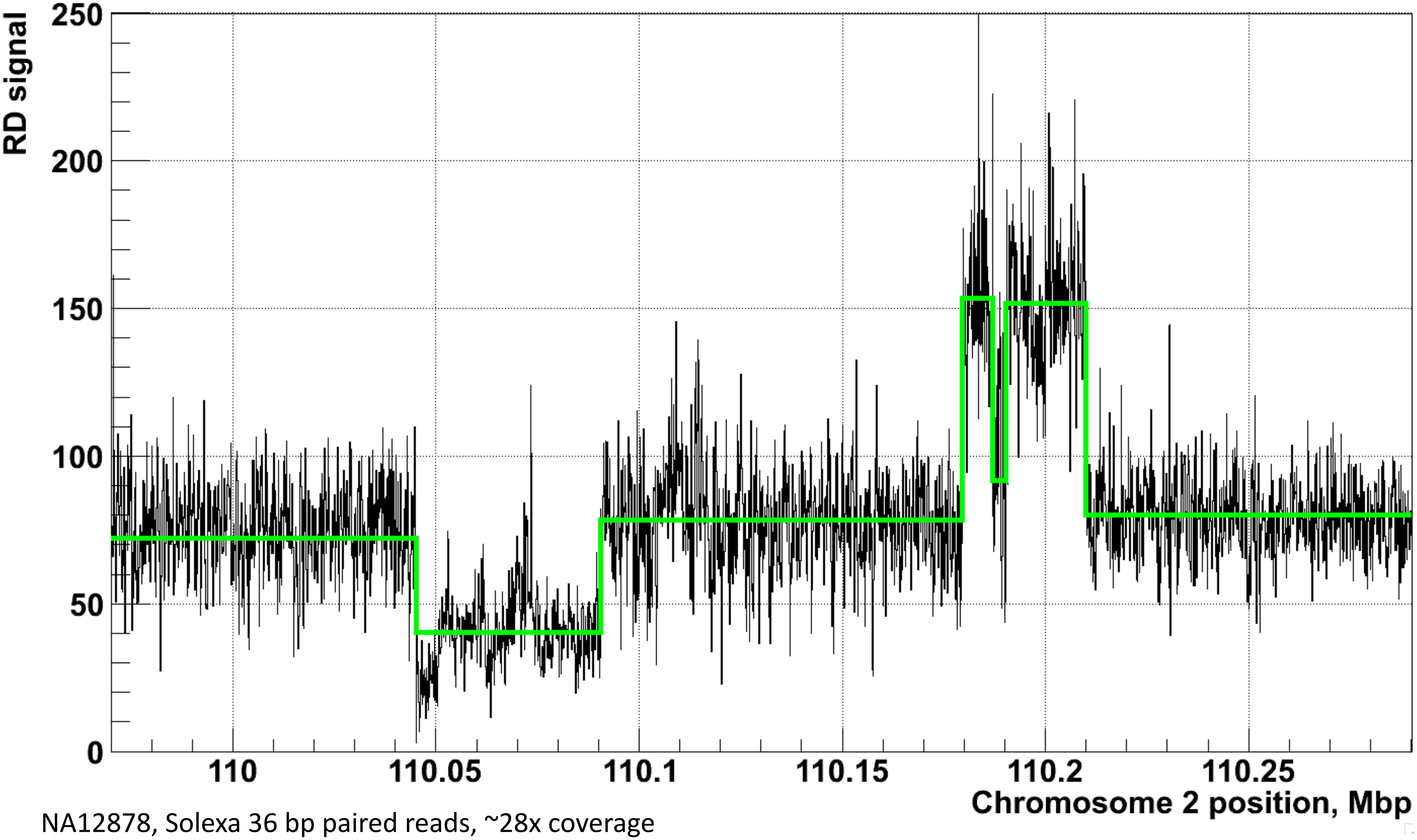
Novel method, CNVnator, mean-shift approach

- For each bin attraction (mean-shift) vector points in the direction of bins with most similar RD signal
- No prior assumptions about number, sizes, haplotype, frequency and density of CNV regions
- Achieves discontinuity-preserving smoothing
- Derived from image-processing applications

Alexej Abyzov

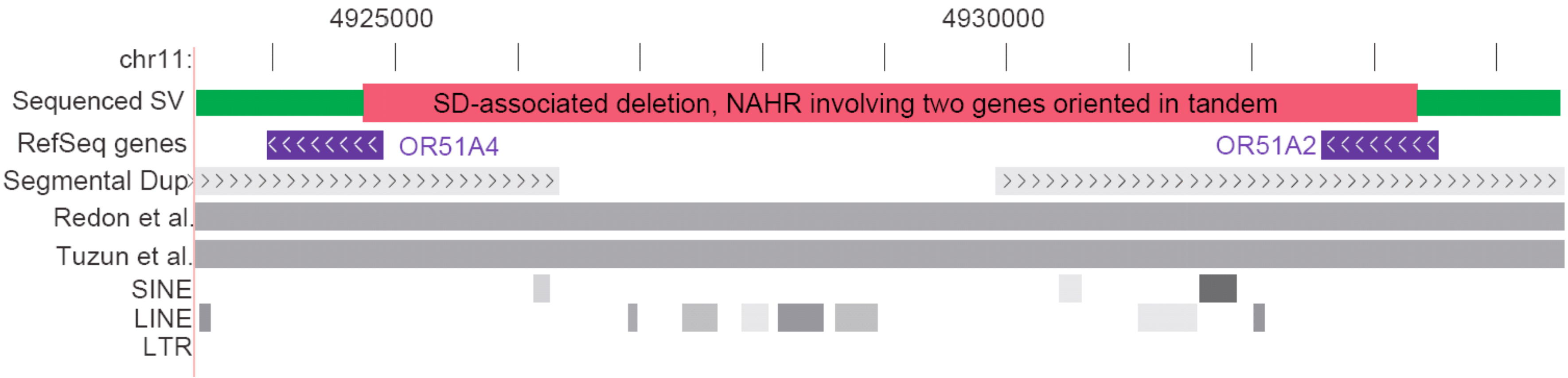


CNVnator on RD data



NA12878, Solexa 36 bp paired reads, ~28x coverage

17% of SVs Affect Genes



Non-allelic homologous recombination (NAHR; breakpoints in OR51A2 and OR51A4)

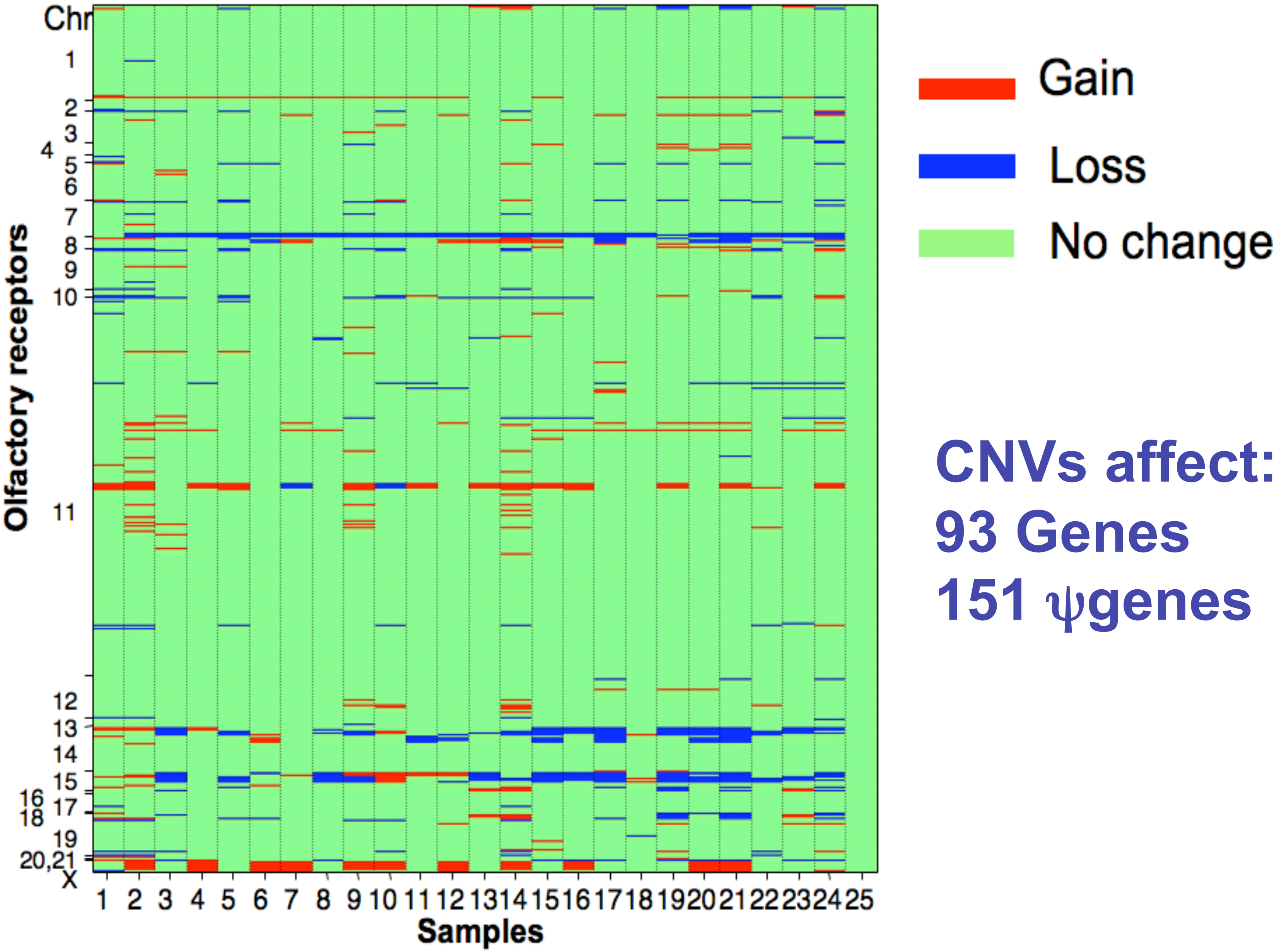
```
←M  P  E  H  L  S  P  E  T  K←     ←M  P  G  H  L  S  P  E  T  K←
CATGGGCTCATGCAAGgagggctctgtctt - - - catgggcccatgcaagGAGGGCTCTGTCTT

chr11 4924724                                               4933366
```

Olfactory Receptor Gene Fusion

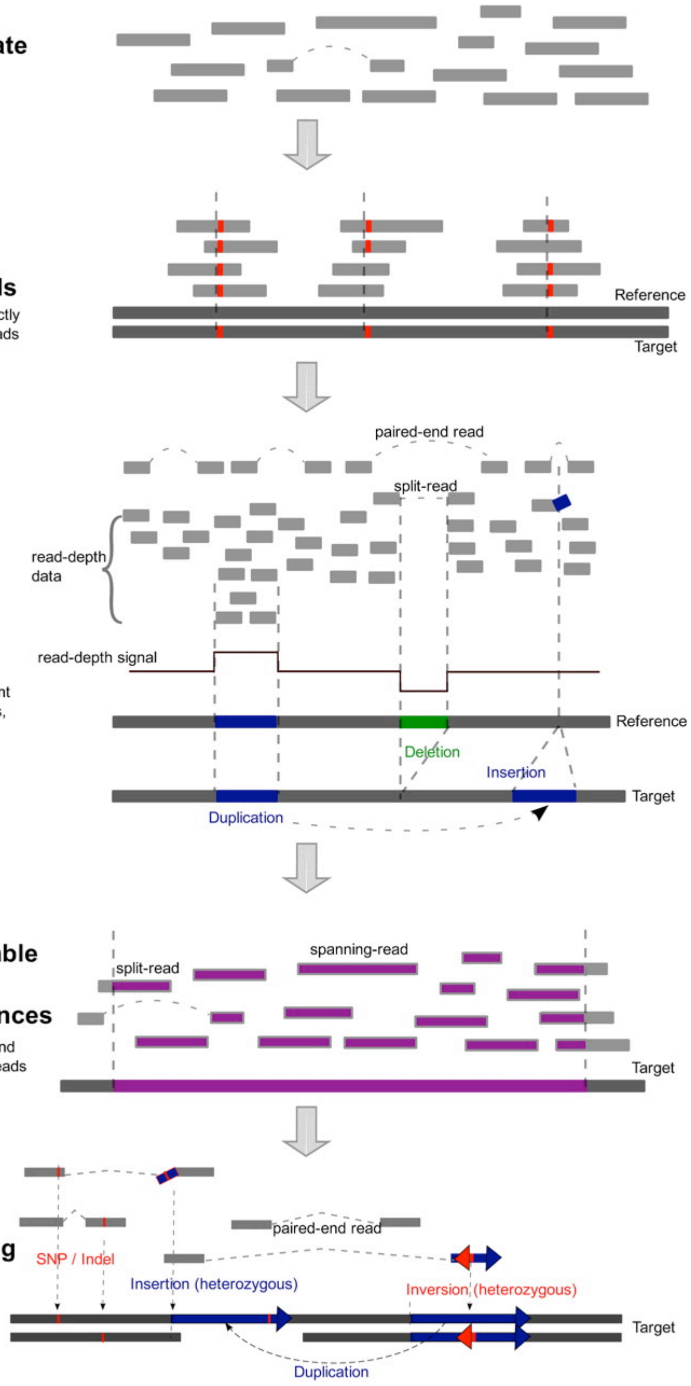
Heterogeneity in Olfactory Receptor Genes

(Examined 851 OR Loci)



CNVs affect:
93 Genes
151 ψgenes

Flow chart for determining a personal genome sequence.



Step 1
Generate Reads

Step 2
Call SNPS & Indels
using correctly mapped reads

Step 3
Find SVs
with aberrant paired-ends, split-reads, read-depth analysis & array CGH

Step 4
Assemble New Sequences
with split- and spanning-reads

Step 5
Phasing
mostly with paired-end reads

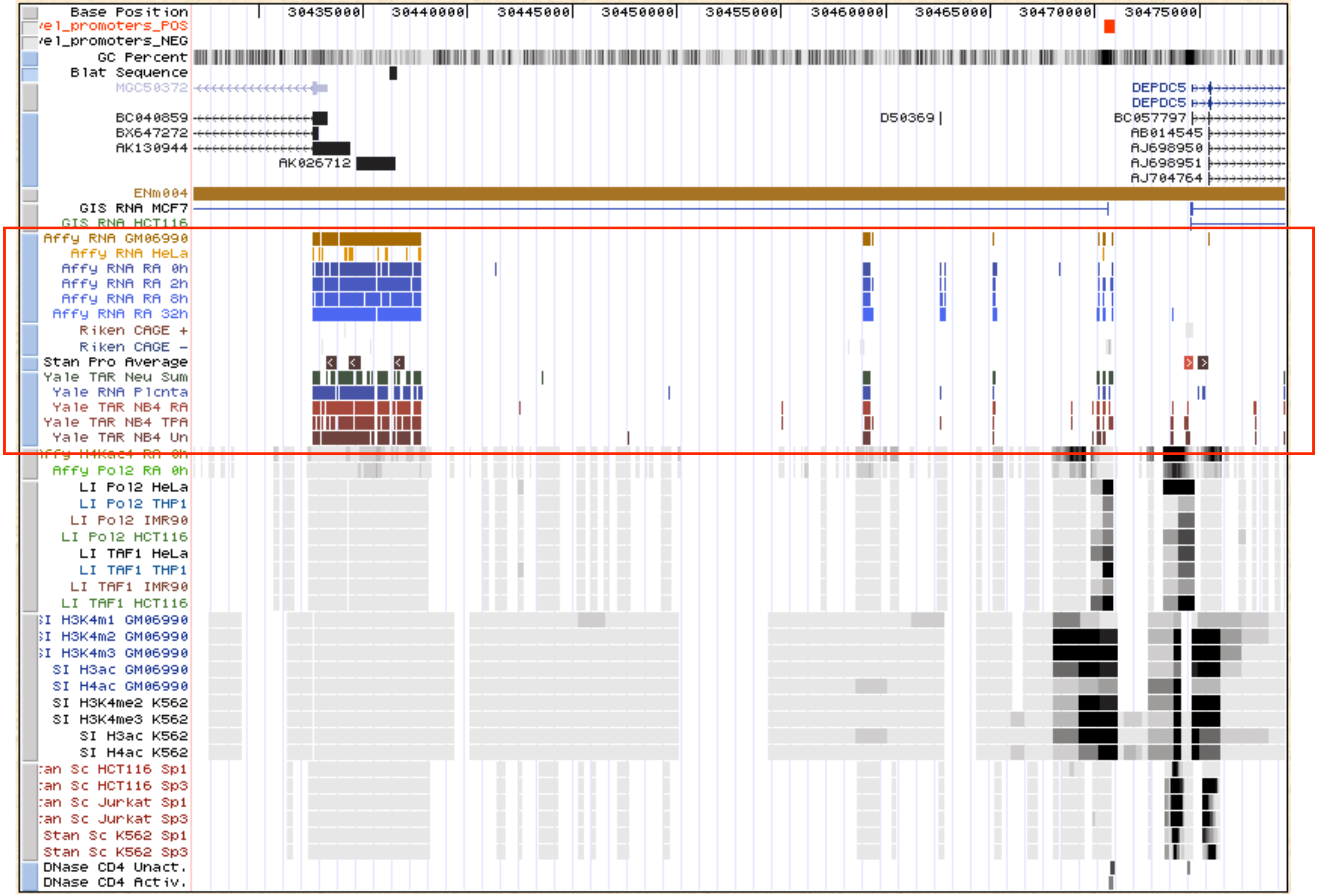
Snyder et al. Genes Dev.
2010;24:423-431

Interpreting the Genome

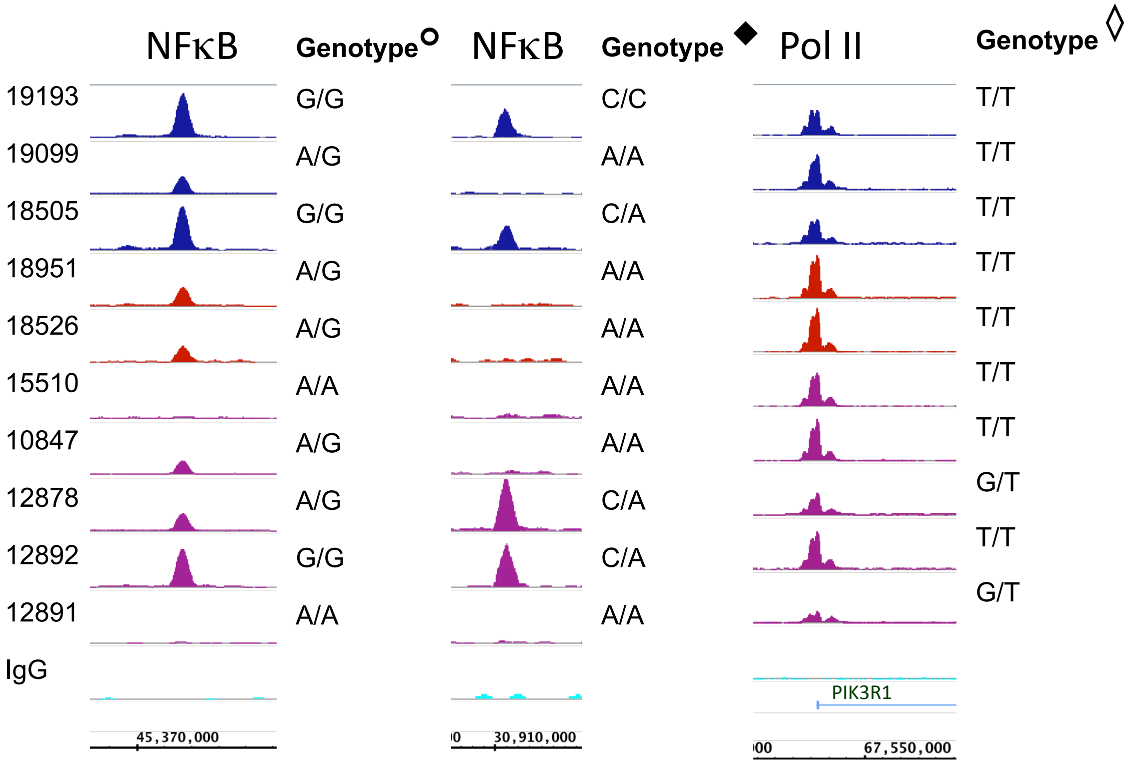
- Technical
  - Error rate = 1 X 10exp-5

- Finding phenotypic variants
  - Coding mutations (PolyPhen, SIFT)
  - Comparison with known databases
    - Existing variants associated with phenotypes (PharmaGkb; Atul Butte’s)
    - Natural variants databases (1000 Genome project)
  - Functional genomics information

ENCODE Project: Transcribed and Regulatory Regions

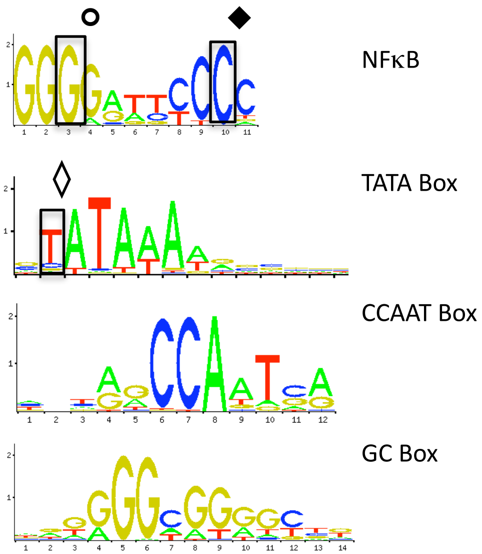


Effect of Motif Associated SNPs on Binding

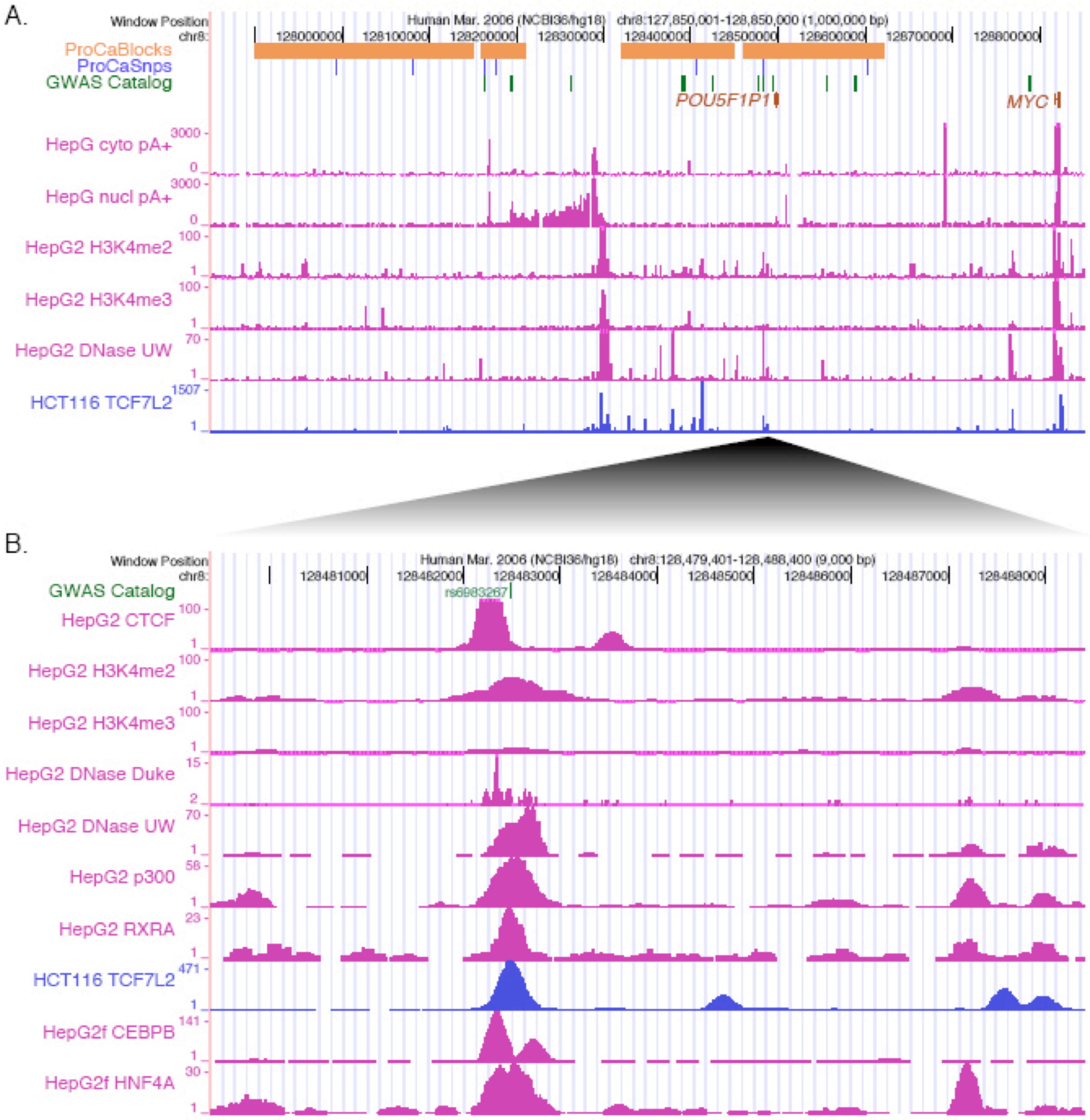


	NFκB	Genotype ○	NFκB	Genotype ◆	Pol II	Genotype ◊
19193		G/G		C/C		T/T
19099		A/G		A/A		T/T
18505		G/G		C/A		T/T
18951		A/G		A/A		T/T
18526		A/G		A/A		T/T
15510		A/A		A/A		T/T
10847		A/G		A/A		T/T
12878		A/G		C/A		G/T
12892		G/G		C/A		T/T
12891		A/A		A/A		G/T
IgG						

Position Weight Matrices



Correlate functional data with variation data



Personnel Genome Sequencing: Pinpoint Disease Mutations

- Focused on families with rare disease
- Exome sequencing
- Whole genome sequencing
  - Miller syndrome
  - Charcot-Marie Tooth Disease
  - Cancer

Charcot-Marie Disease

- Neuropathy
  - Heterogenous disease—many different genes mapped
- Sequence genome to 30X coverage
- 3.4 M SNPs: (561,719 novel)
  - 2,255,102 in intergenic
  - 1,165,204 in genes, introns etc.

174 nonsynonymous SNPs in region of interest

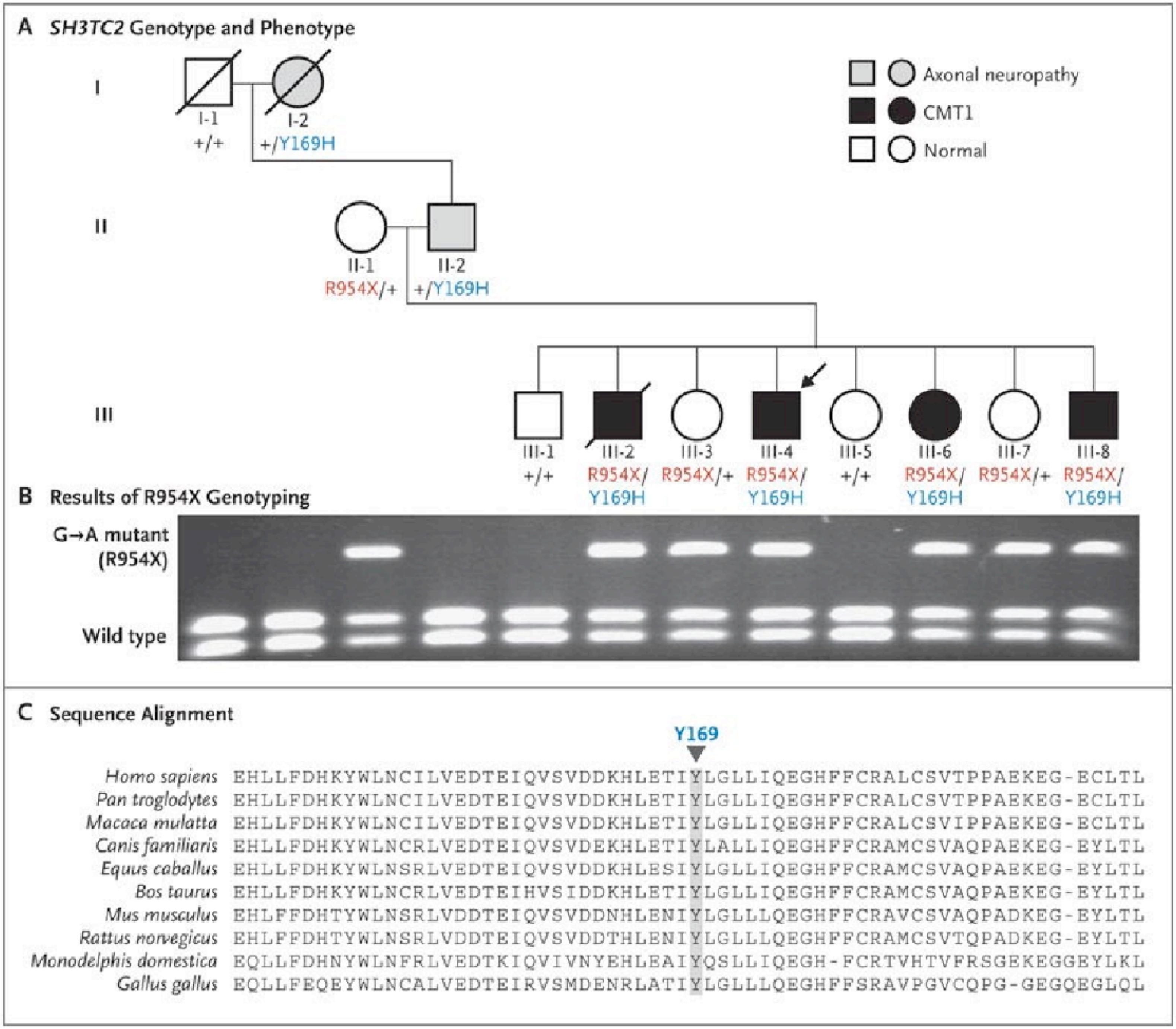
Ultimately zoomed in on SH3TC2 gene:

Full blown disease has two mutations: Y169H (missense), R954X (nonsense)*

Single heterozygotes have some phenotypes

*Implicated previously

Lupski et al 2010 NEJM


A SH3TC2 Genotype and Phenotype
I
II
III
I-1
+/+
I-2
+/Y169H
II-1
R954X/+
II-2
+/Y169H
III-1
+/+
III-2
R954X/
Y169H
III-3
R954X/+
III-4
R954X/
Y169H
III-5
+/+
III-6
R954X/
Y169H
III-7
R954X/+
III-8
R954X/
Y169H
Axonal neuropathy
CMT1
Normal
B Results of R954X Genotyping
G→A mutant
(R954X)
Wild type
C Sequence Alignment
Y169
Homo sapiens EHLLFDHKYWLNCILVEDTEIQVSVDDKHLETIYLGLLIQEGHFFCRALCSVTPPAEKEG-ECLTL
Pan troglodytes EHLLFDHKYWLNCILVEDTEIQVSVDDKHLETIYLGLLIQEGHFFCRALCSVTPPAEKEG-ECLTL
Macaca mulatta EHLLFDHKYWLNCILVEDTEIQVSVDDKHLETIYLGLLIQEGHFFCRALCSVIPPAEKEG-ECLTL
Canis familiaris EHLLFDHKYWLNCRLVEDTEIQVSVDEKHLETIYLALLIQEGHFFCRAMCSVAQPAEKEG-EYLTL
Equus caballus EHLLFDHKYWLNSRLVEDTEIQVSVDDKHLESIYLGLLIQEGHFFCRAMCSVAQPAEKEG-EYLTL
Bos taurus EHLLFDHKYWLNCRLVEDTEIHVSIDDKHLETIYLGLLIQEGHFFCRAMCSVAQPAEKEG-EYLTL
Mus musculus EHLFFDHTYWLNSRLVDDTEIQVSVDDNHLENIYLGLLLQEGHFFCRAVCSVAQPADKEG-EYLTL
Rattus norvegicus EHLFFDHTYWLNSRLVDDTEIQVSVDDTHLENIYLGLLLQEGHFFCRAMCSVTQPADKEG-EYLTL
Monodelphis domestica EQLLFDHNYWLNFRLVEDTKIQVIVNYEHLEAIYQSLLIQEGH-FCRTVHTVFRSGEKEGGEYLKL
Gallus gallus EQLLFEQEYWLNCALVEDTEIRVSMDENRLATIYLGLLLQEGHFFSRAVPGVCQPG-GEGQEGLQL

Conclusions

1) **Many phenotypes are due to rare or private mutations.**

2) **Personnel genome sequencing can help find them.**

3) **Mutations/variants that land outside of gene will be hard to predict.**