Identification of disease genes
Mutational analyses
Monogenic diseases

Objectives : identify the disease causing mutation among millions of polymorphisms.

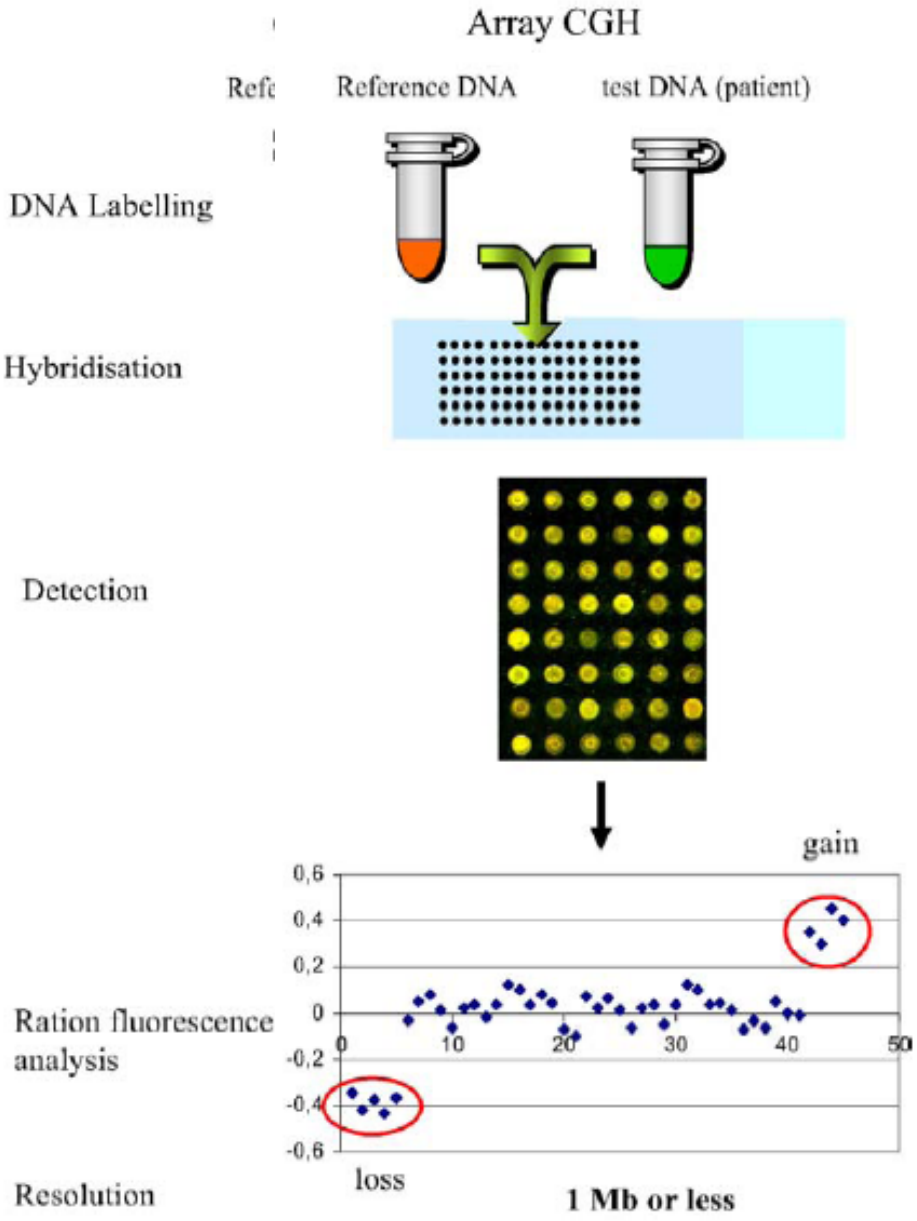
Contents :

- Copy number variations (CNV)
- Next Generation Sequencing/exome/dbSNP/EVS
- Loss of function, pathogenicity
- Strategies inherited diseases
  - Dominant mutations
  - Recessive mutations
- (Neomutations)

Copy Number Variations (CNV)

- Deletions or duplications
- Array CGH (Comparative Genomic Hybridisation) :

Slides covered with millions of ordered oligonucleotides covering the entire human genome



Sanlaville D, 2005

CNV

The resolution of the DNA arrays depends on the density of oligonucleotides

www.nimblegen.com

Mechanisms causing CNVs

- Random breaks (DNA double strand break repair, DSBR)
  - NHEJ : non homologous end joining

  Due to micro-homologies (2 to15 nt)

  —> insertion of a few nucleotides during repair

- Homologous recombination
  - NAHR : non allelic homologous recombination

  Segmental duplications, repeated sequences

Repeated sequences : account for 60 % of the human genome

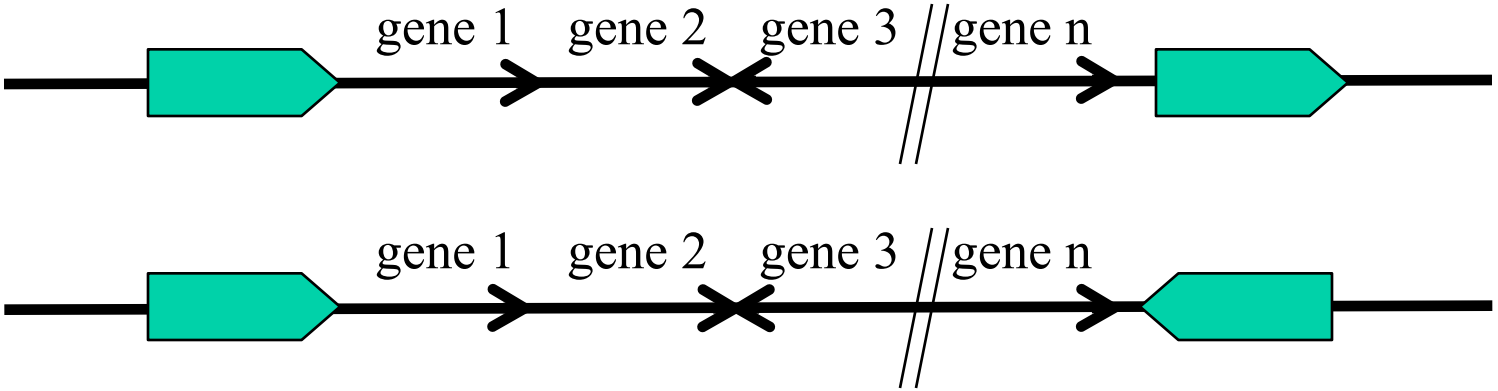
45 % interspersed sequences : example Alu sequences

2 to (5?) % segmental duplications

10 % others

Segmental duplications :

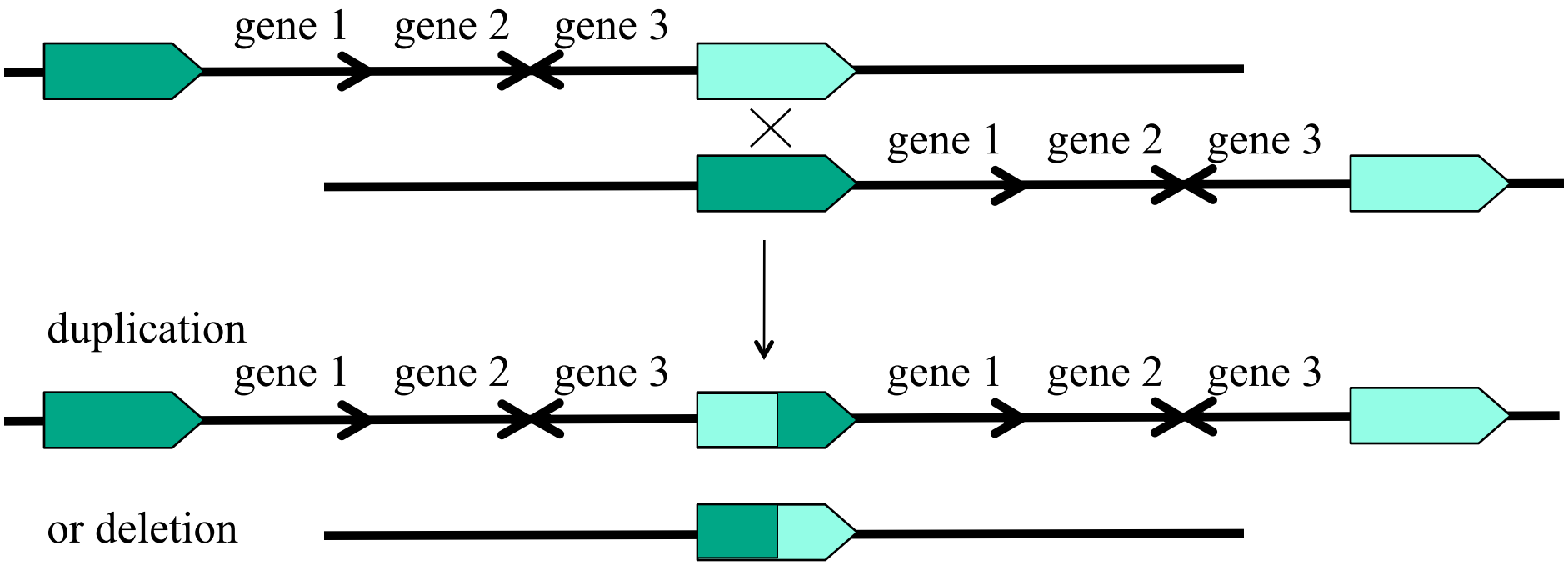
- 10 to 300 kb duplicated elements
- On a same chromosome or on different chromosomes
- In the same orientation (head to tail) or in opposite orientations



NAHR non allelic homologous recombination :

- interspersed repeated sequences (ex Alu sequences)
- segmental duplications : —> recurrent deletions

  (micro-deletional syndromes: Williams, Di-Georges …)



Interpretation of CNVs

- Polymorphic CNVs : Database of Genomic Variants
- Pathologic CNVs : DECIPHER database

Segmental duplications are causing recurrent rearrangements (often *de novo* rearrangements for CNVs dominants diseases)

Demonstration a new CNV is disease causing :

- Need to have several patients with identical or overlapping CNVs and having a share clinical picture
- or Confirmation with a patient having the same clinical picture and a point mutation in one of the genes included in the CNV

Analysis of small mutations
Next generation sequencing

Nucleotide substitutions

Small insertions or deletions of 1, 2 or several nucleotides

Sanger sequencing

Next generation sequencing (NGS or MPS)

Limitation of the size of the detectable insertions/deletions
due to the length of sequencing reads

Reads of 100 nt —> detection of max 50 nt insertion/deletion

Next generation sequencing

Illumina (Solexa) technology

HiSeq 2500 / 2000 / 1500 / 1000

HiScanSQ

Genome Analyzer IIx

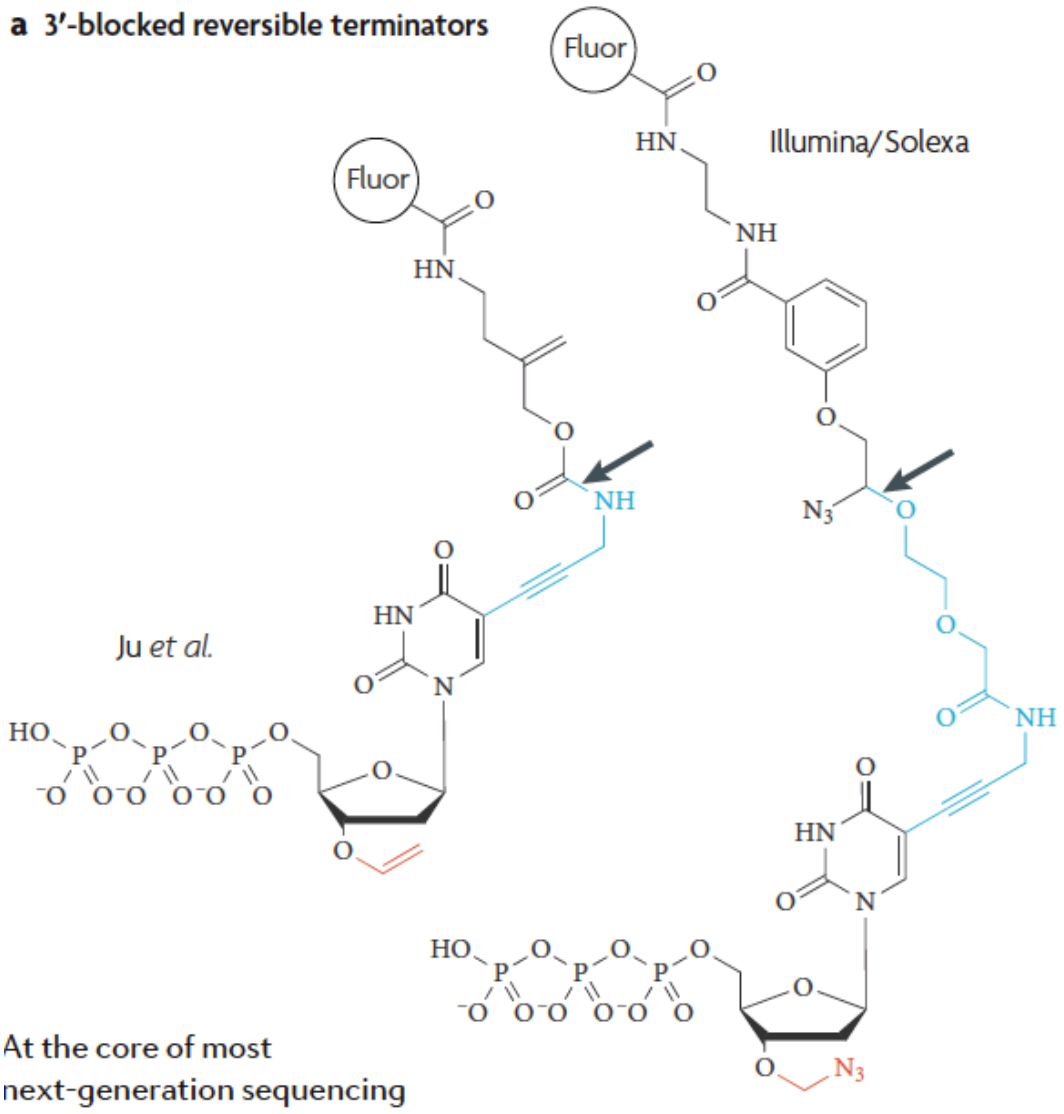
MiSeq

HiSeq 2000: Up to 600 Gb per run, in 11 days
2 x 100 nt read length, two billion paired-end reads/run.
In a single run, sequence two human genomes at ~30x coverage (read depth) for $3-5,000 (USD) per genome.

Next generation sequencing

Illumina (Solexa) technology

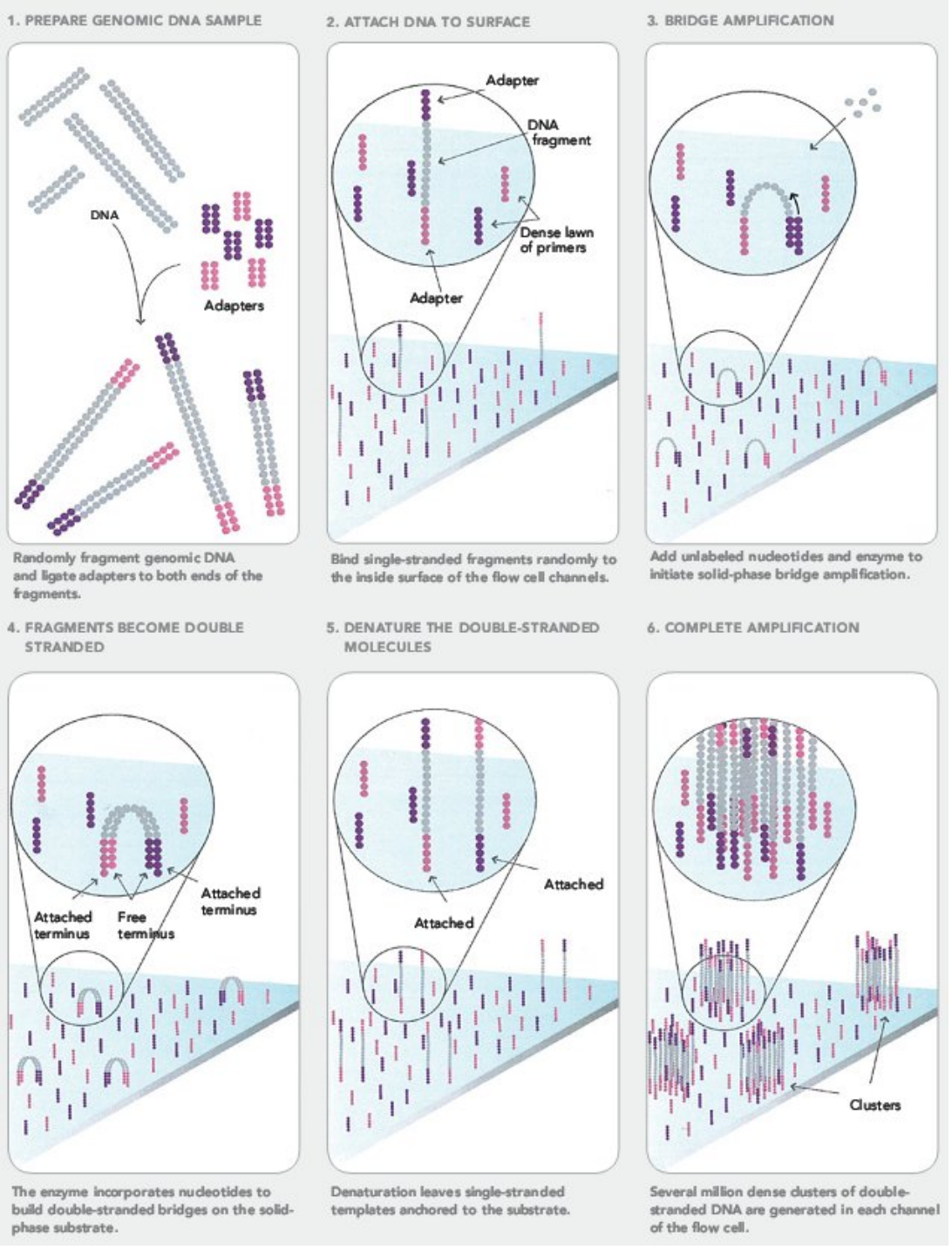
Sequencing by elongation with fluorescent terminators



- Illumina (Solexa) technology

Clonal amplification on a slide

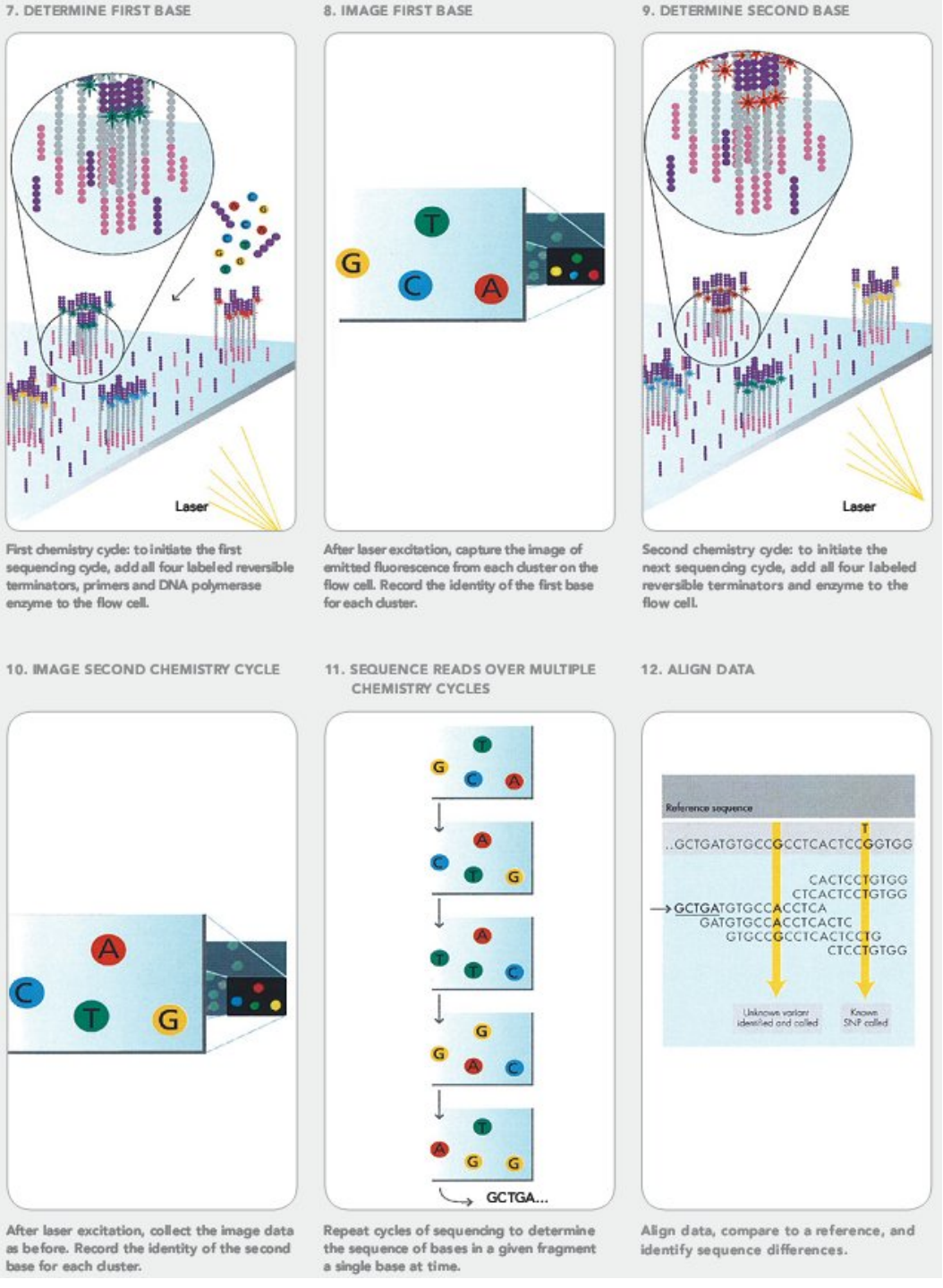
—> molecular clones



Cycle sequencing, nucleotide by nucleotide

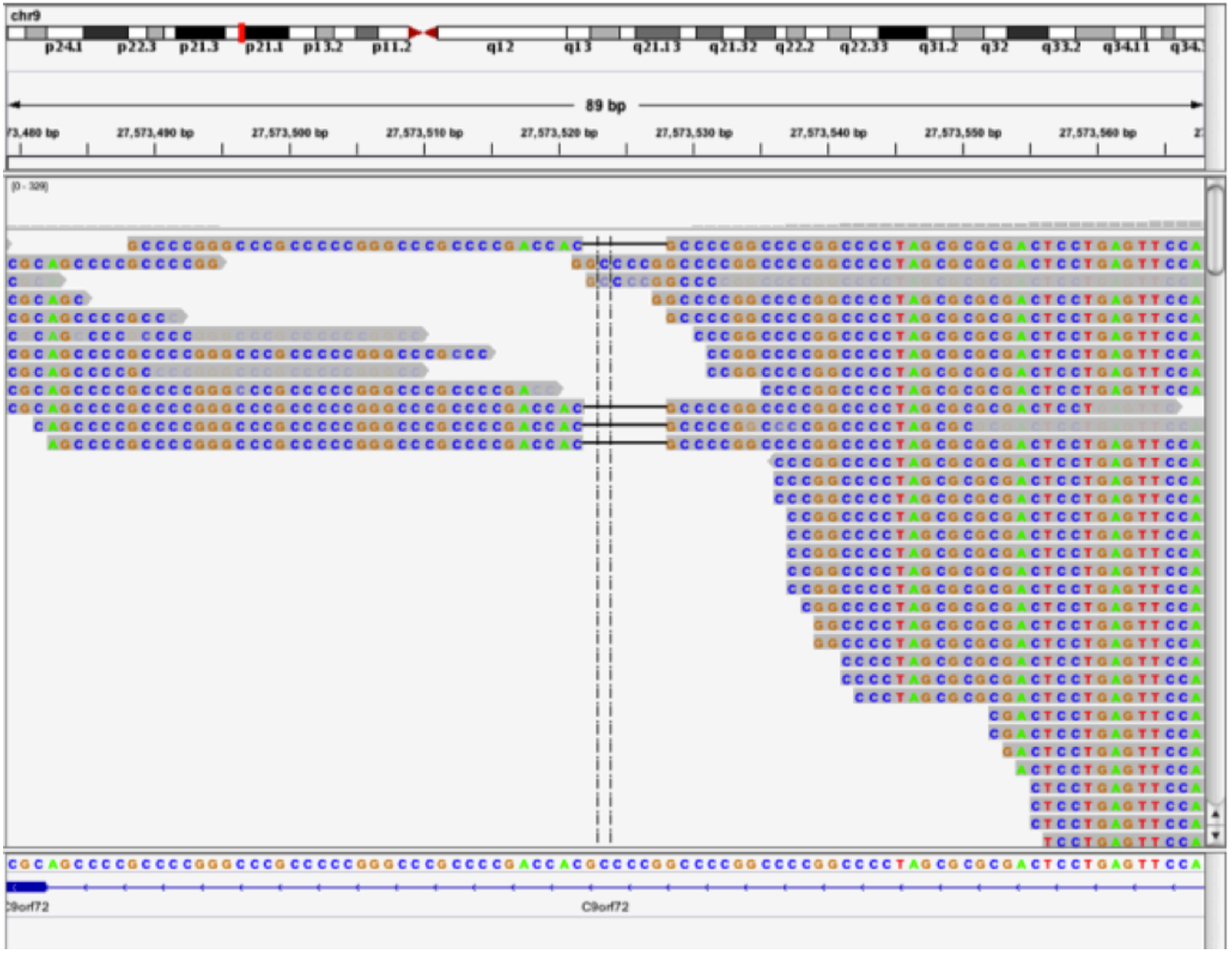
Picture scanning of the slide after each cycle (CCD camera)

–> .fastq files



Alignment of the reads with respect to the human genome reference :

–> .bam files



Targeting coding sequences

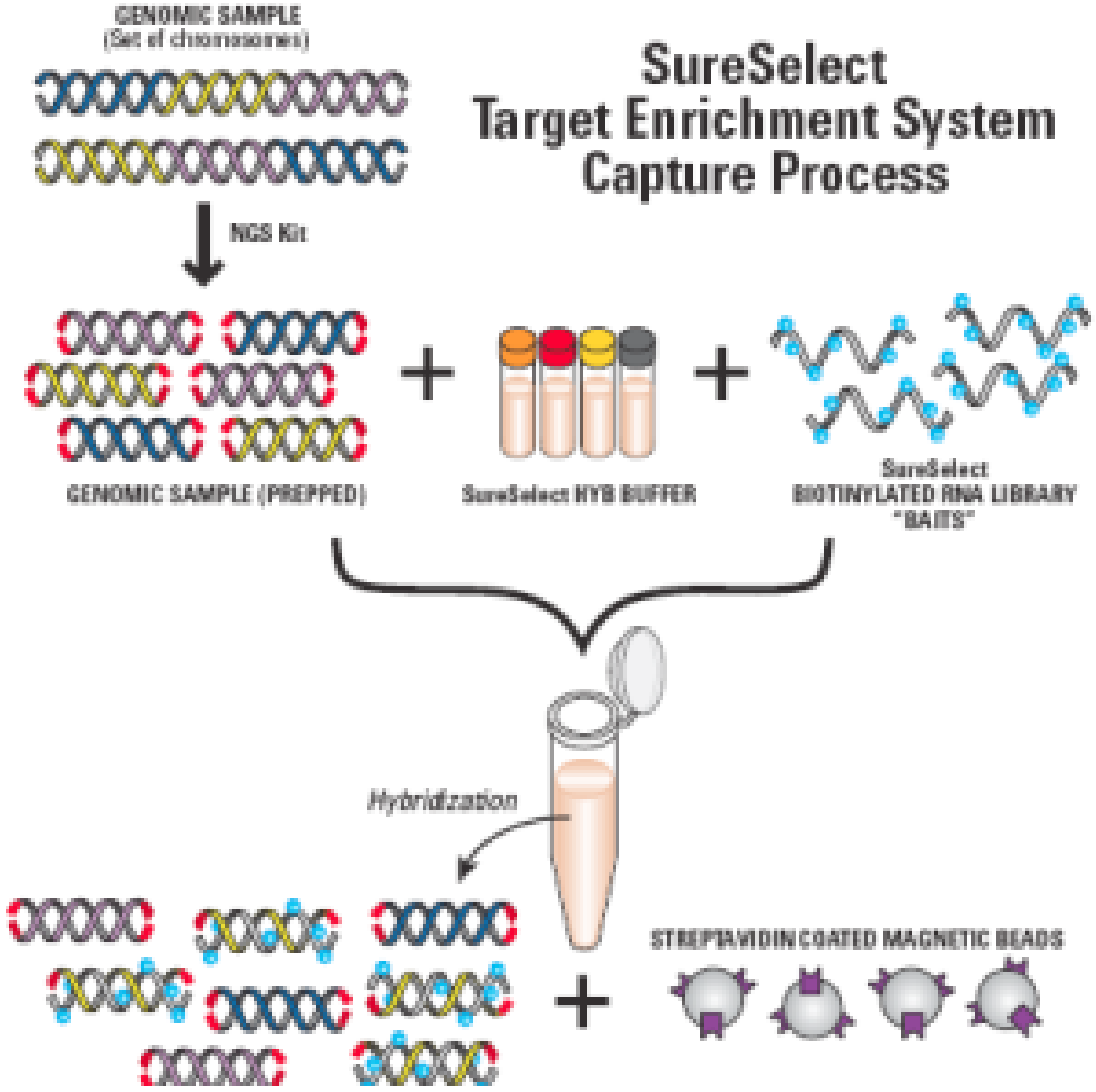
- Limited interest for introns and intergenic sequences, most interest for coding sequences and flanking splicing sequences
  - 2 % of the human genome

- The solution : enrichment of the sequences of interest by targeted capture of all exons of the genome (exome)

—> reduces the cost per sample

—> better coverage (read depth)

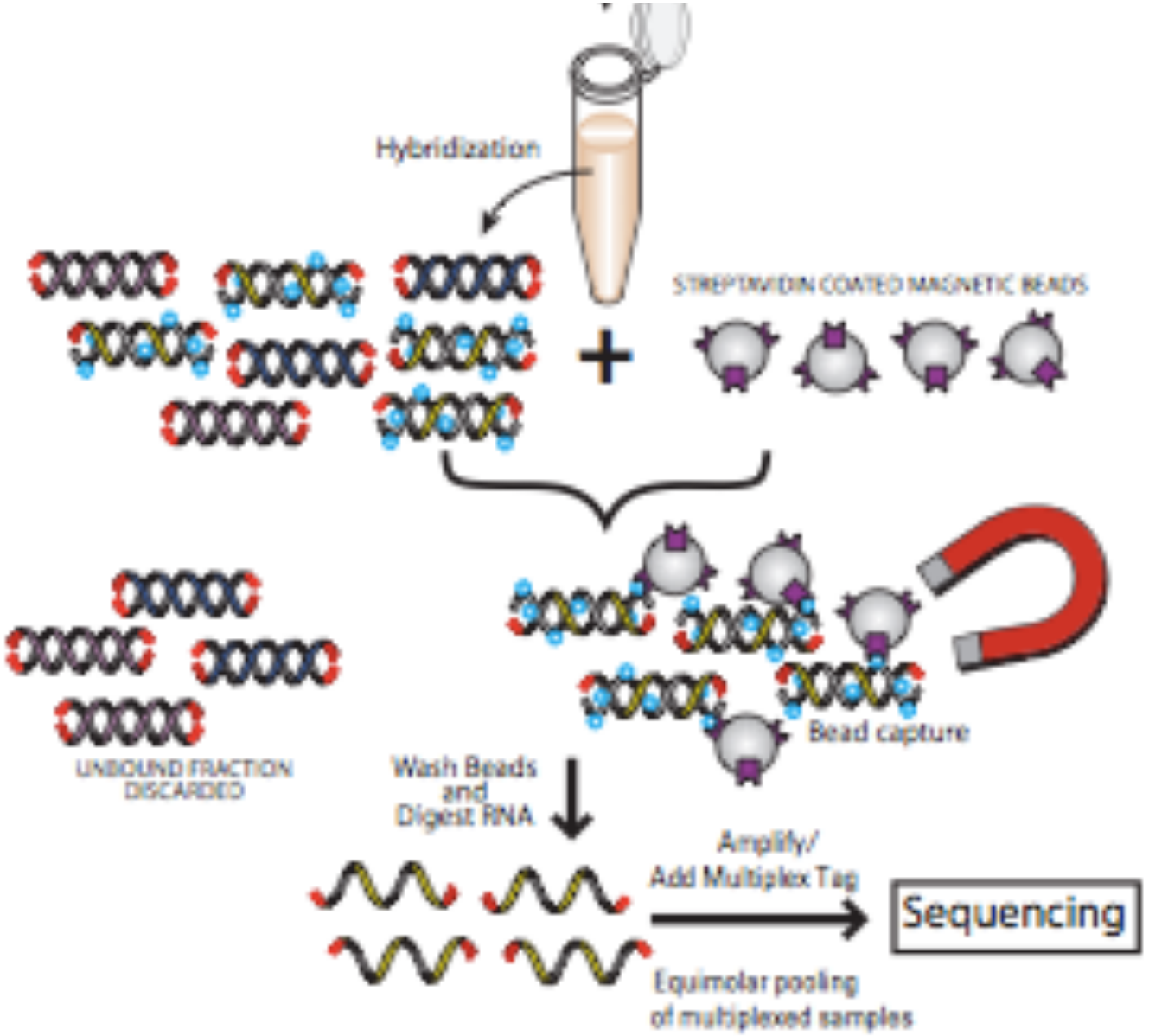
Enrichment by exome capture



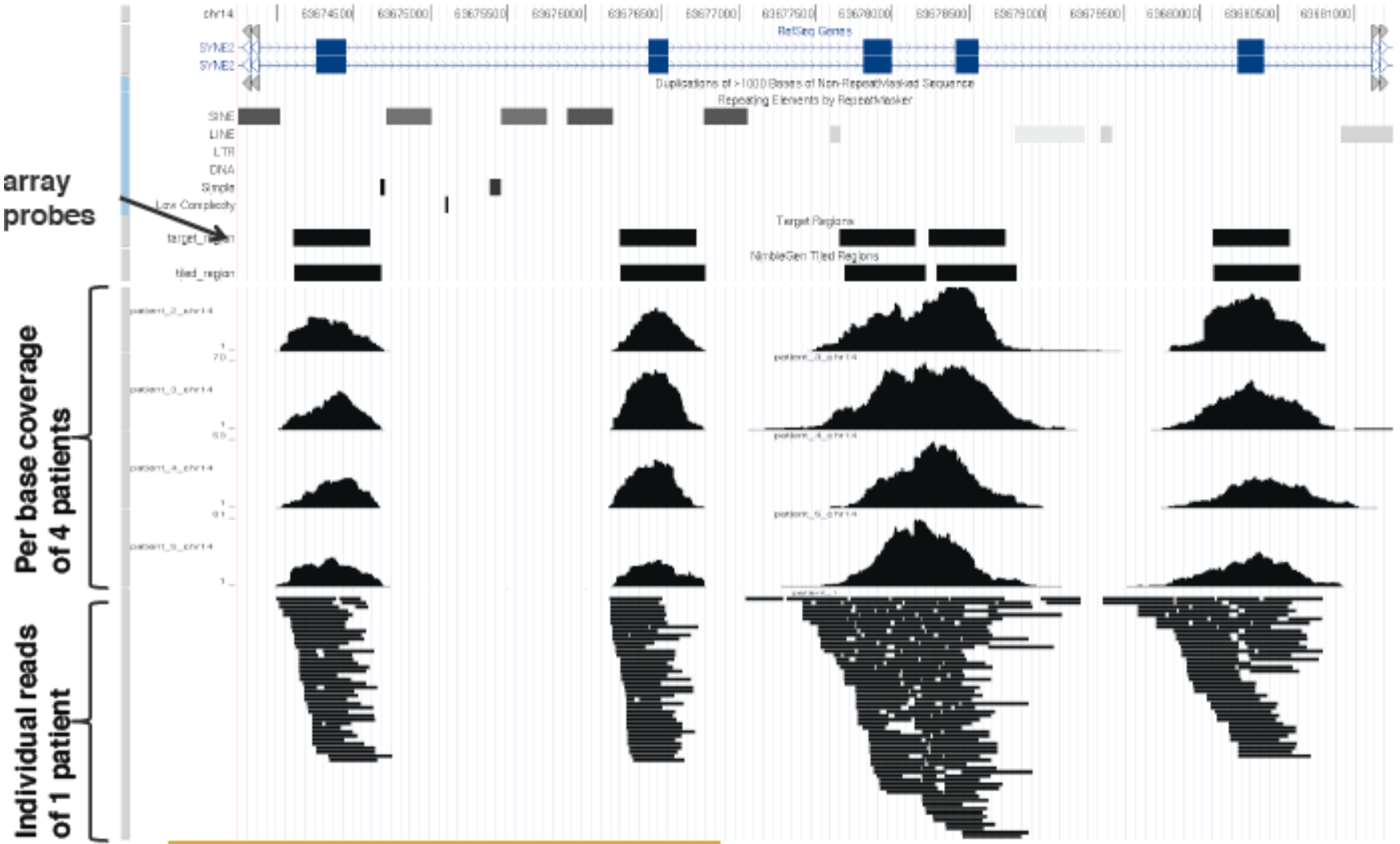
1 million of oligos of 120 nt covering the exons of the 21 000 human genes

Up to 50 mega-nt covered

Enrichment by capture (exome)(2)



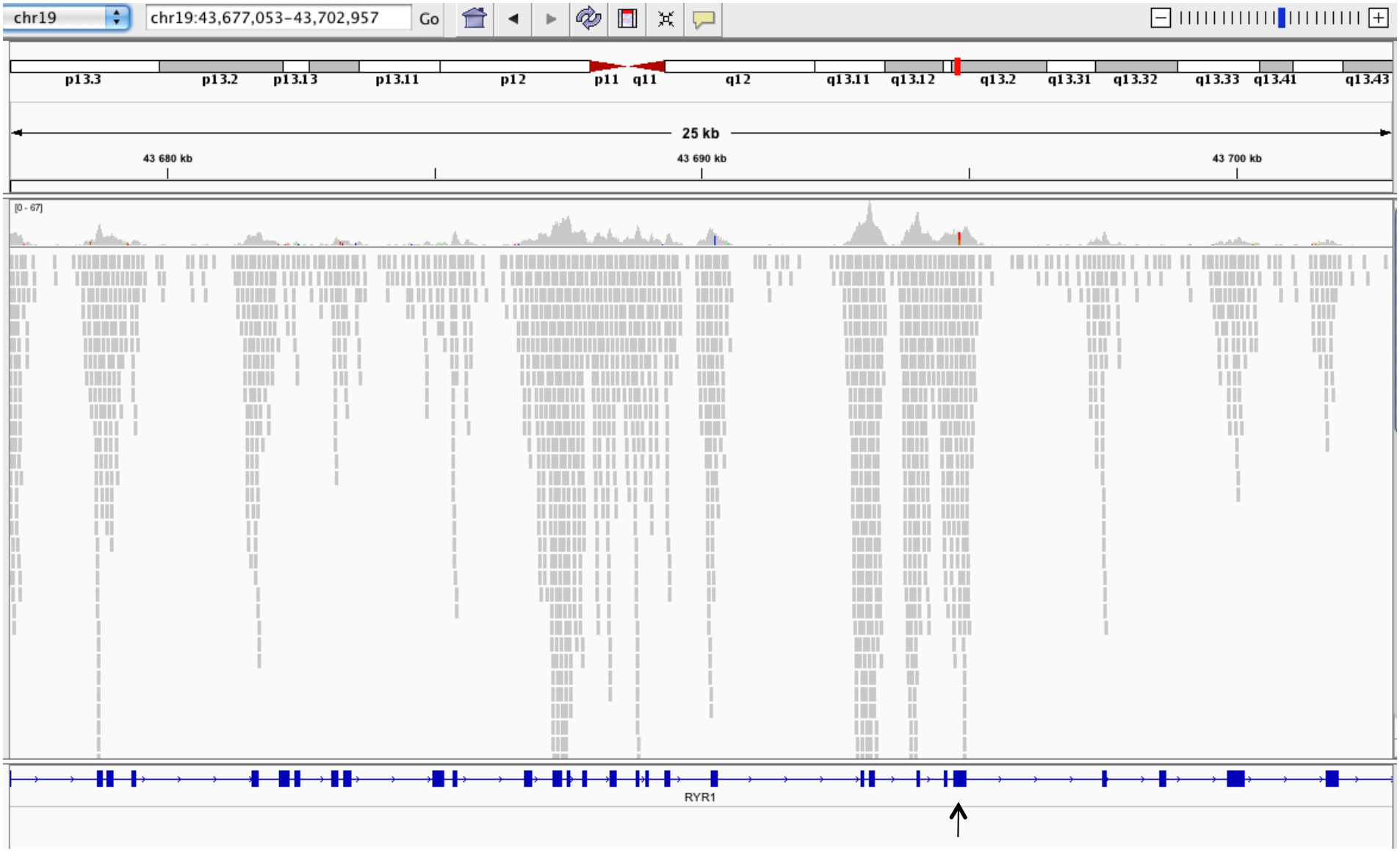
Exonic enrichment and sequencing is reproducible!



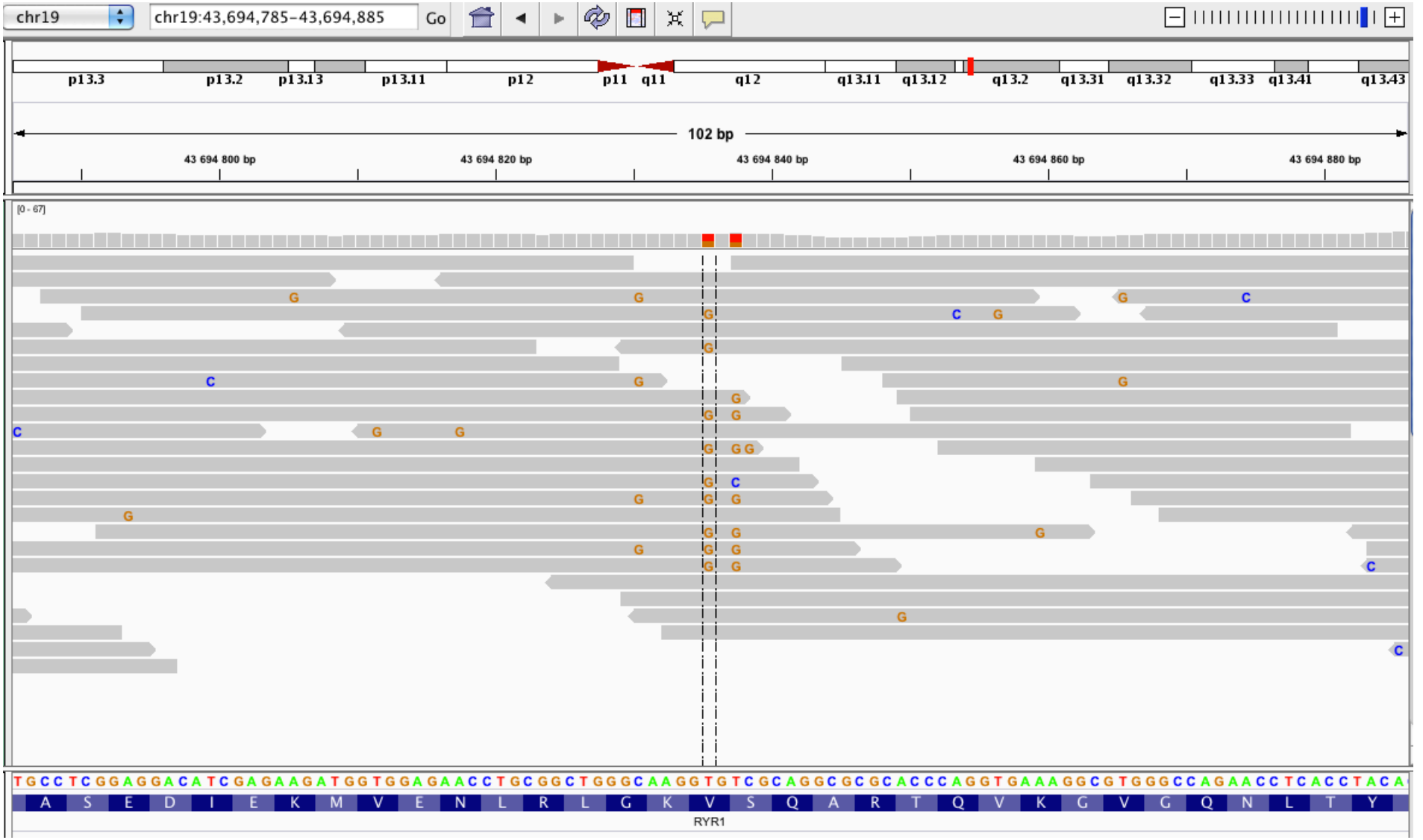
Human Genetics Nijmegen

IGV Analysis (Integrative Genomics Viewer)

with .bam files



IGV Analysis (Integrative Genomics Viewer)



Bio-informatics analysis –> .vcf files

About 70 000 SNPs et 2 000 indels per exome

- Elimination of artifacts related to technology
- Elimination of common polymorphisms :

usually all SNPs > 1%

- dbSNP138 (includes data from 1000 genome, 1KG)
- Exome Variant Server (EVS) :
  - 4,300 European American individuals
  - 2,200 African American individuals
- ExAC (Exome Aggregation Consortium)

  60,706 unrelated individuals

Bio-informatics analysis
Loss of function, pathogenicity (1)

Gain of function are not predictable.

Partial loss of function are sometimes difficult to predict.

- The non-sense, indel and -1, -2, +1, +2 splice mutations are generally deleterious.

  Exceptions : some indels which maintain the reading-frame

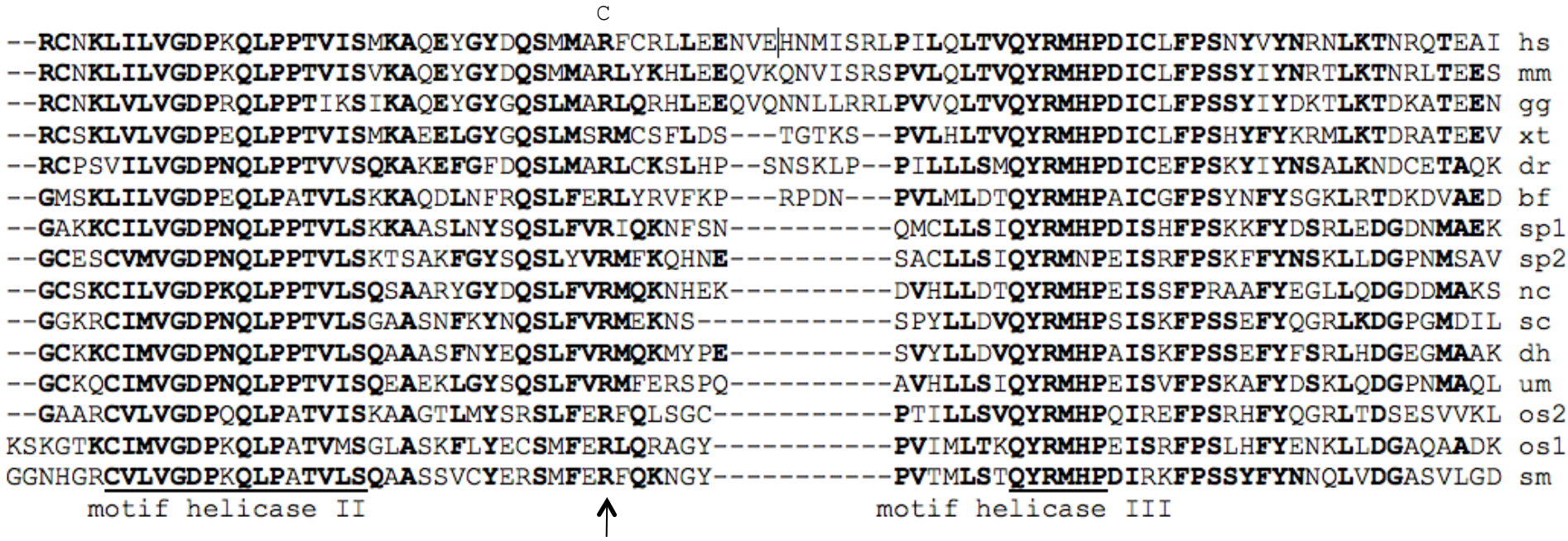
- For the other splice mutations, need to use prediction programs based on splice site consensus matrices. (Spliceport, Human Splicing Finder, NNSPLICE, MaxEntScan, HSF …)

  Need to confirm splice site alteration by RT-PCR studies on patient cells.

Loss of function, pathogenicity (2)

- Silent variations (synonymous) are in general non deleterious —> excluded
- Missense mutations : 9000 missense variations per exome
  - Grantham score (degree of physico-chemical change)
  - Conservation scores : PolyPhen-2 et SIFT programs .

  These programs use databases of conserved protein sequences



hs homo sapiens, mm mus musculus, gg gallus gallus, xt xenope tropicalis, dr danio rerio, bf branchiostoma floridae, sp schizosaccharomyces pombe, nc neurospora crassa, sc saccharomyces cerevisiae, dh debaryomyces hansenii, um ustilago maydis, os oriza sativa, sm selaginella moellendorffii

Strategies for inherited diseases

- Dominant mutations

Several independent individuals with different mutations (in this case, mutations are often clustered in a same domain —> gain of function), or with the same mutation (founder effect), are needed.

If large families with a high LOD, two families (mutations) may be sufficient.

- Recessive mutations

Two independent non-consanguineous individuals —> 4 mutations in the same gene

If large consanguineous families with high LOD score, two families (mutations) may be sufficient.

If only ONE large consanguineous family with high LOD score, there is a need to demonstrate that the mutation causes a loss of function (easier for non-sense, truncating (frame shift) or splice mutations; functional studies for missense mutations)

- Neo-mutations (next lecture/Dr Jean-Baptiste Rivière)

1st success of WES (Whole Exome Sequencing)

Miller syndrome (Ng et al. Nature Genetics 2010)

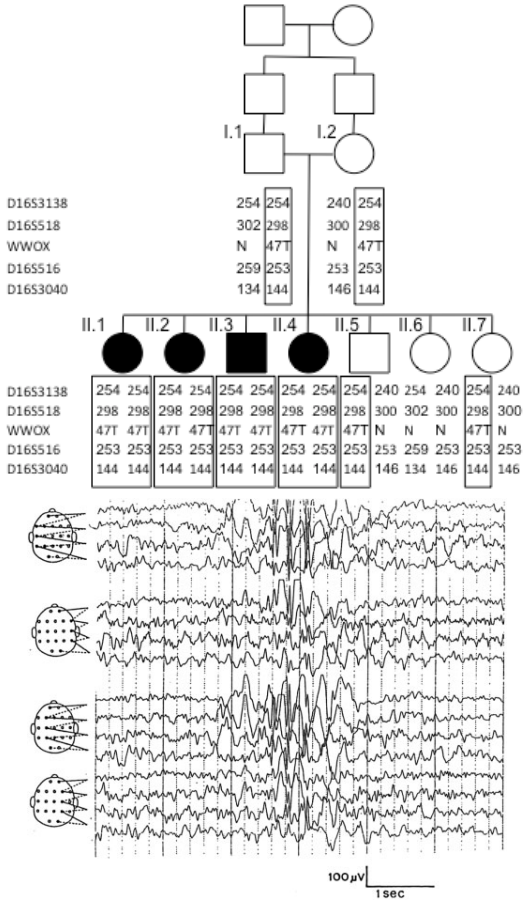
Syndromic dysmorphic disease

Exome sequencing of 4 patients from 3 families —> *DHODH* gene

---

Other examples : 1. Mallaret et al. Brain 2014

Ataxia – epilepsy – mental retardation, linkage to 16q21-23 (2007)

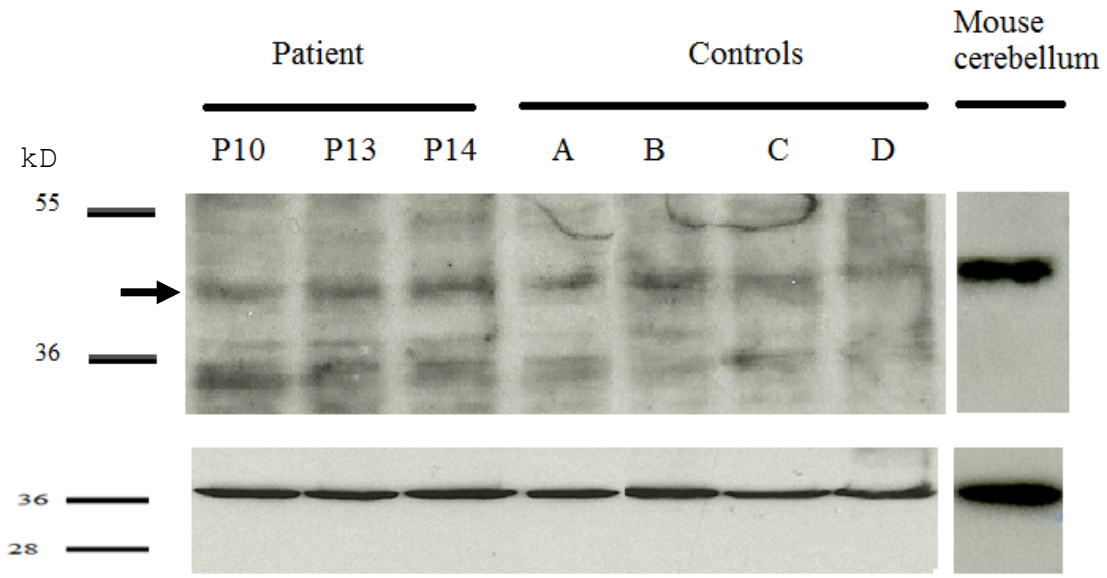


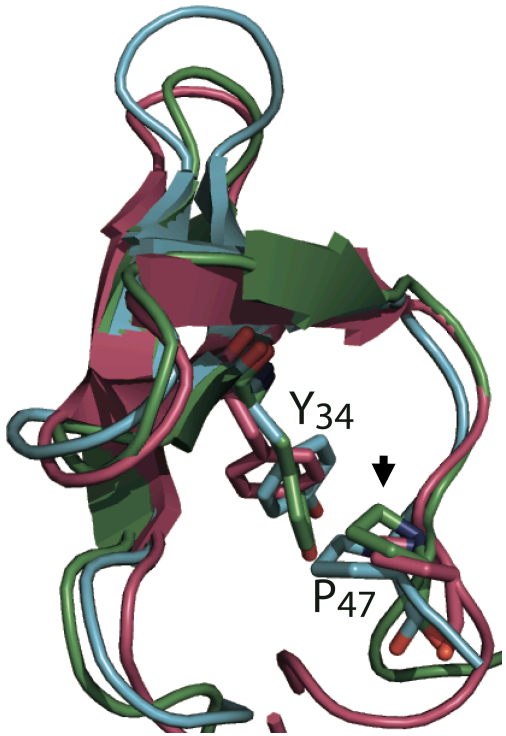
Nb of sequenced bases	101.323.813
Total SNP	68497
Novel SNP	13616
Novel homozygous SNP	807
Novel homozygous SNP in linkage region	10
Novel homozygous SNP in linkage region, affecting coding or splice sequence	1 c.139C>A, p.Pro47Thr in *WWOX* gene
Confirmation : exome on 2nd family Consanguinity, 2 affected	c.1114G>C, p.Gly372Arg in *WWOX* gene

```
                                                T
Homo s.          DSEDELPPGWEERTTKDGWVYYANHTEEKTQWEHPKTGKRK
Mus m.           DSEDELPPGWEERTTKDGWVYYANHTEEKTQWEHPKTGKRK
Gallus g.        DSEEELPPGWEERTTKDGWVYYANHLEEKTQWEHPKSGKRK
Xenopus l.       DSEDELPPGWEERSTKDGWVYYANHFDEKTQWEHPKTGKRK
Danio r.         DSEDELPPGWEERSTKDGWVYYANHEEMKTQWEHPKTGKKK
Drosophila m.    DSEDELPPGWEERATDDGTVCYVNQQGKTSQWTHPRTGRSK
Acromyrmex e.    DSEDELPPGWEERTTLDGNVYYVNHYTKGTQWTHPRTGRKK
Nematostella v.  DSDEELPVEWEVRTTDTGRVYYANHLTKTTQWQHPKTGKIR
Trichoplax a.    DSDPELLPGWEKSKTSTGRTFYVDHNTQTTQWEHPQRSHKK
```

p.Pro47Thr homozygous missense mutation in the *WWOX* gene(WW domain/oxido-reductase)

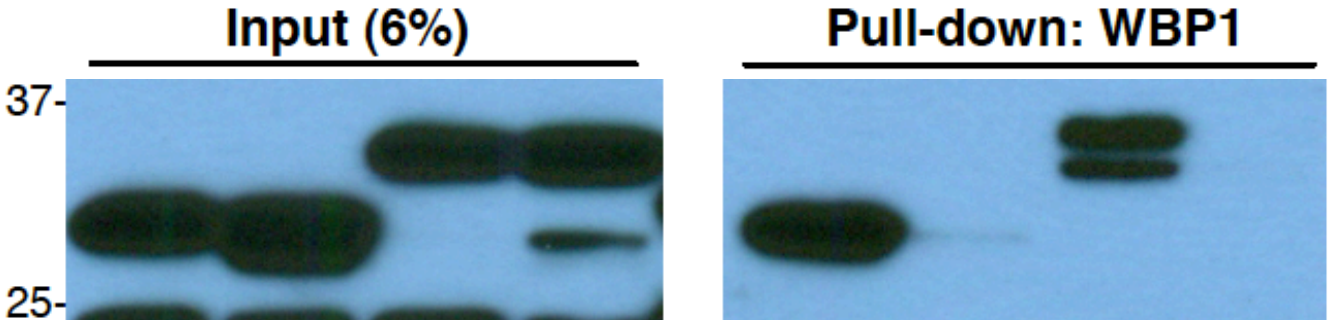
(Mallaret et al. Brain 2014)



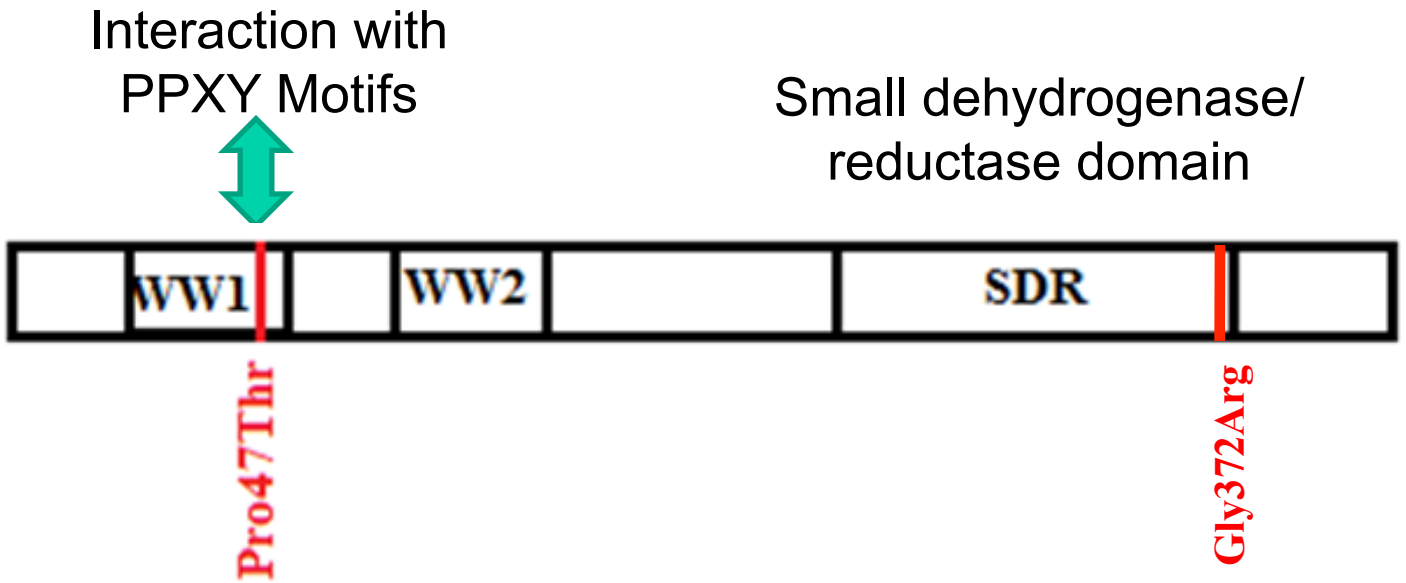


Ataxia/epilepsy syndrome (SCAR12) (Mallaret et al. Brain 2014, 137: 411-9)

Mutation in the tumor-suppressor gene WWOX: WW domain oxido-reductase
p.Pro47Thr in family 1 (Saudi Arabia)
p.Gly372Arg in family 2 (Israeli Palestinians, ataxia/epilepsy + spastic paraplegia)
(Exome NGS)

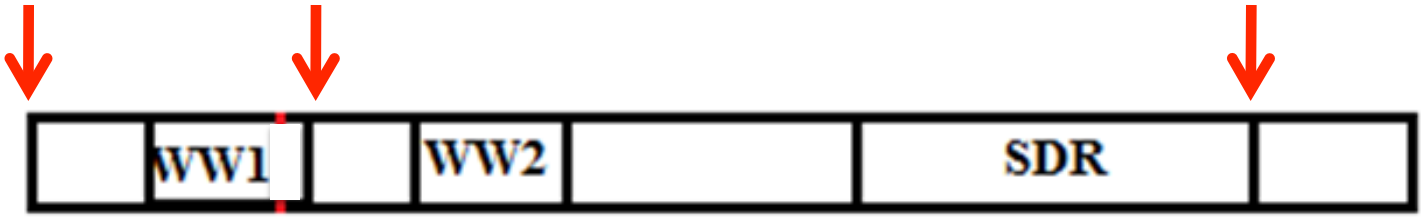


J. Lee &
C. M. Aldaz



Complete loss of function of WWOX causes a lethal-dwarfism syndrome

Truncations :



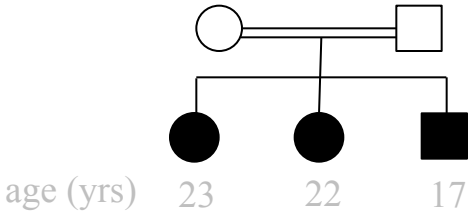
lde (*lethal dwarfism-epilepsy*) rat spontaneous mutation, *death at 3-12 weeks* :
deletion of 13 nt in exon 9 —> truncation
Suzuki et al. Comp Med. 2007; 57: 360-9.
Suzuki et al. Genes Brain Behav. 2009; 8: 650-60.

Mice knock out : *lethal dwarfism-epilepsy, death at 3-4 weeks* :
Mallaret et al. Brain 2014; 137: 411-9.

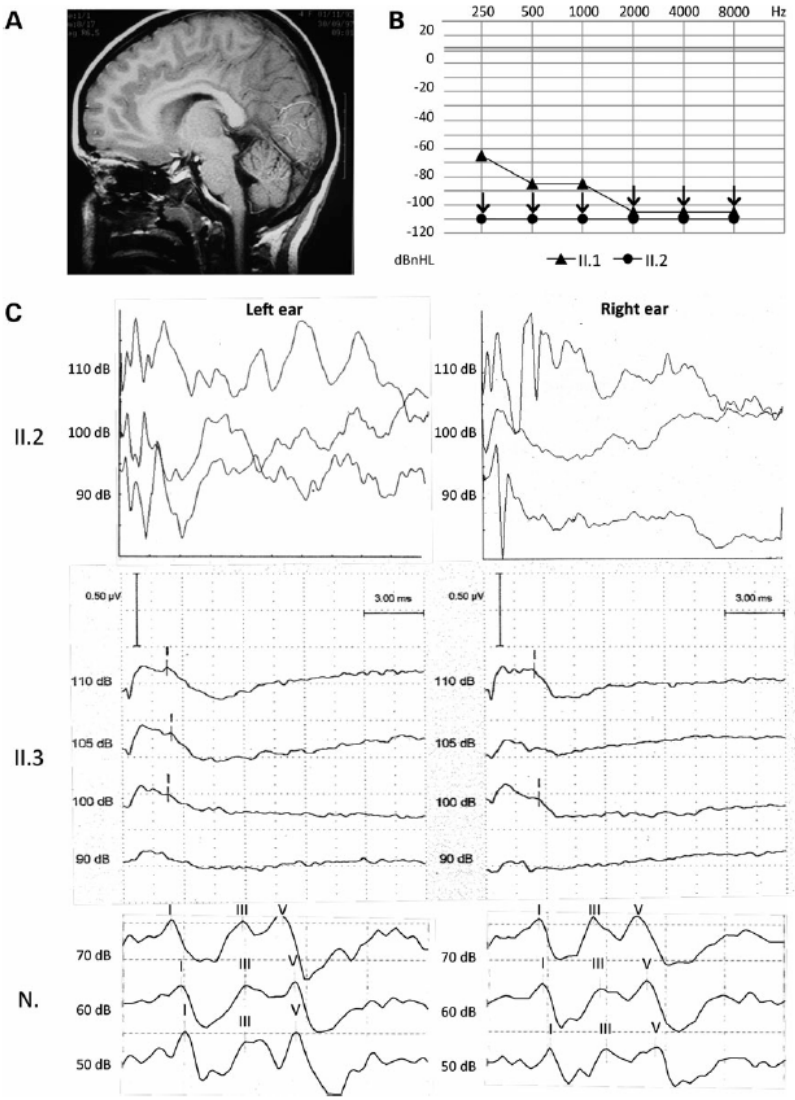
Early lethal microcephaly syndrome with epilepsy, growth retardation and retinal degeneration, *death at 4-16 month* :
homozygous non-sense mutation p.Arg54*
Abdel-Salam et al. Orphanet J. of Rare Dis. 2014; 9:12.

2. Lichtenstein-Knorr syndrome

Family CA



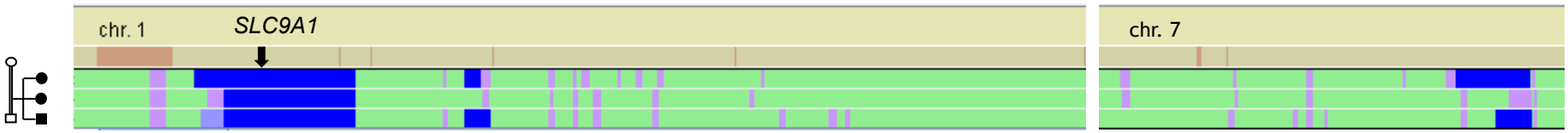
- Turkish origin, 3 affected, Pr B. Leheup, Nancy
- 1st degree consanguinity
- Delayed walking at ages ranging from 18 months to 5 years
- Cerebellar and posterior column ataxia
- Cerebral MRI revealed very mild anterior vermis atrophy (youngest sister)
- Deafness (profound in the sisters, moderate in the younger brother)
- Language: none
- Growth retardation and microcephaly (brother only)



Homozygosity mapping with Genechip SNP 50K XbaI, Affymetrix

Two regions of shared homozygosity:

Chr. 1	20 Mb	248 SNP
Chr. 7	3.5 Mb	37 SNP

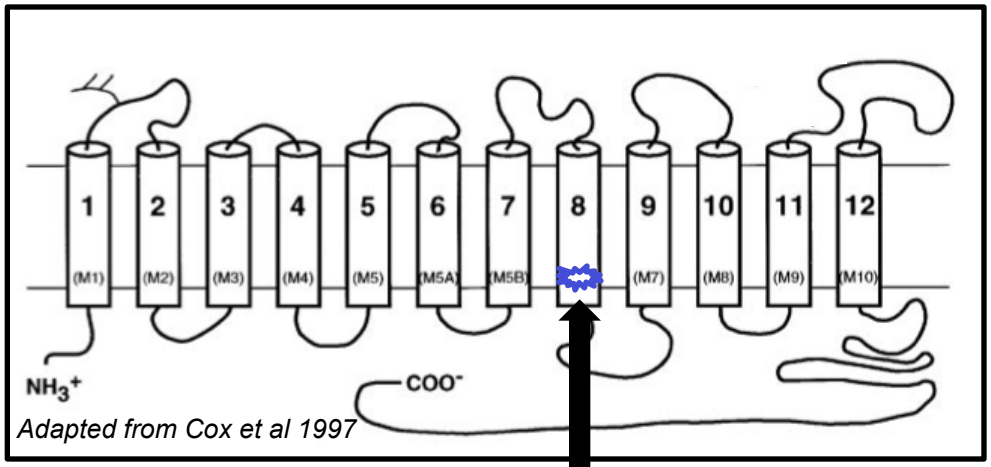


Exome NGS

SLC9A1

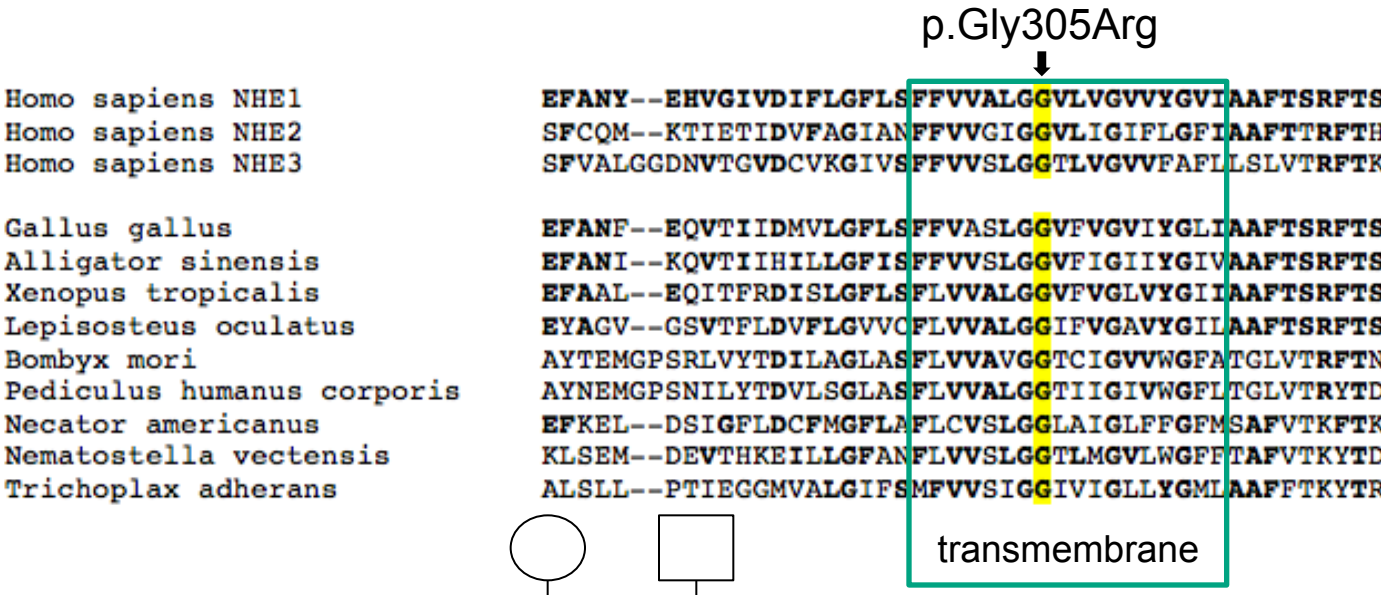
SLC9A1 encodes for NHE1 (Na+/H+ exchanger family member 1), a 815 amino acid protein with 12 transmembrane helices

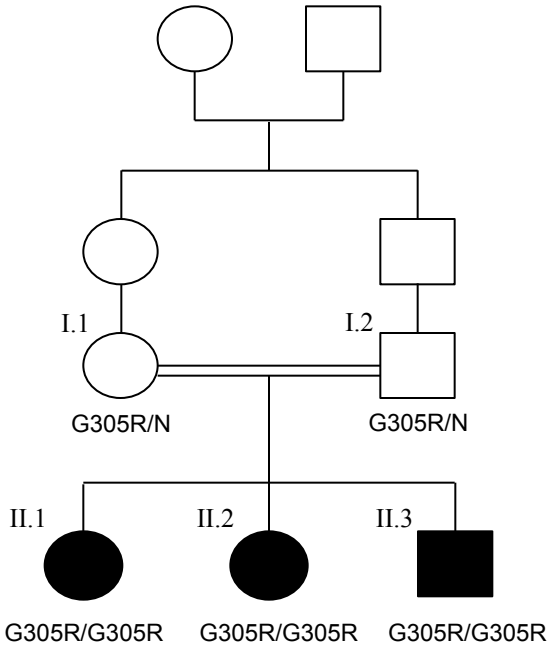
- ubiquitous at the surface of mammalian cells
- regulates the pH in inner ear



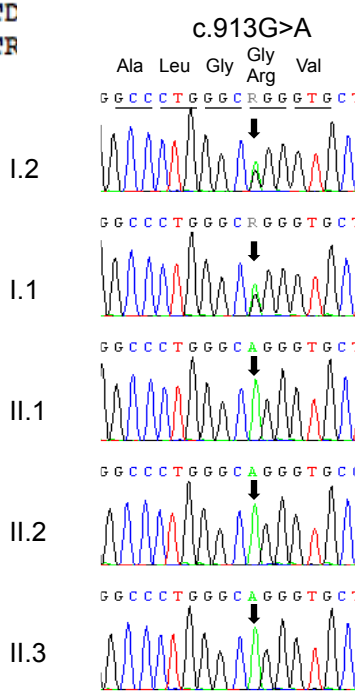
p.Gly305Arg

Segregation and *in silico* predictions





Predicted deleterious by SIFT (score=0.02)



Mouse models

Progressive Neurodegeneration in Deep Cerebellar Nuclei

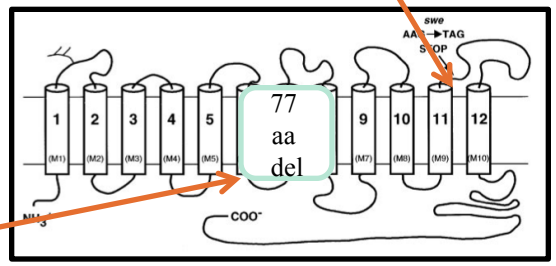
Cell, Vol. 91, 139–148, October 3, 1997, Copyright ©1997 by Cell Press

Sodium/Hydrogen Exchanger Gene Defect in Slow-Wave Epilepsy Mutant Mice

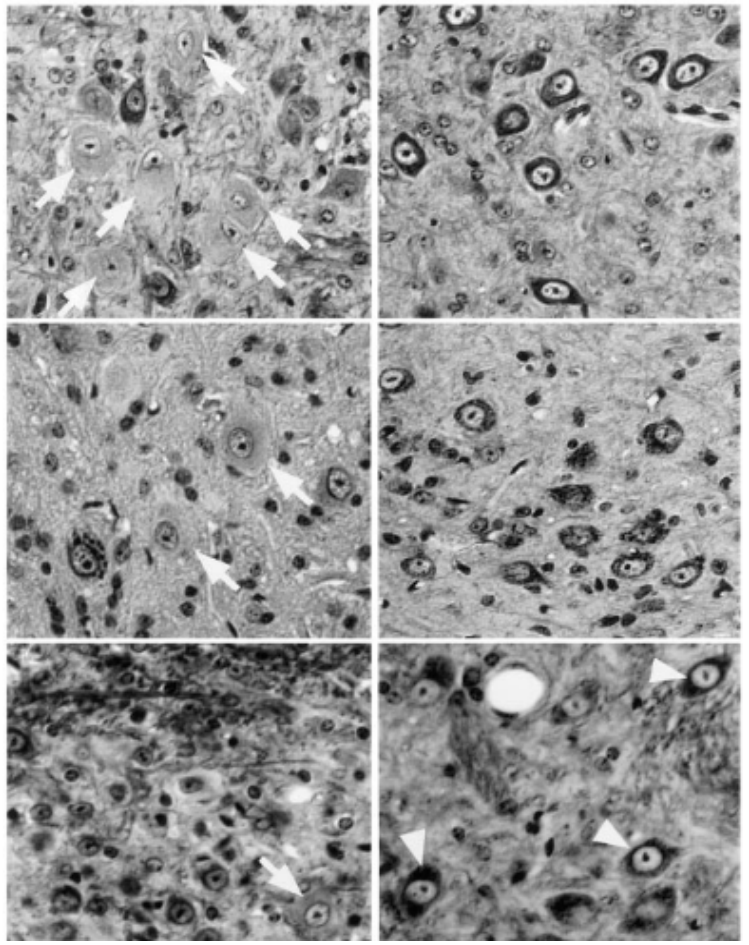
Gregory A. Cox,* Cathleen M. Lutz,* Chao-Ling Yang,† Daniel Biemesderfer,† Roderick T. Bronson,‡ Audrey Fu,* Peter S. Aronson,† Jeffrey L. Noebels,§ and Wayne N. Frankel*‖
*The Jackson Laboratory

Discussion

Here we describe the phenotype, genetic mapping, and identification of the defective gene in the *swe* mutant mouse. The disease appears to be restricted to the CNS, where *swe* mice are severely ataxic and display a unique pattern of neurodegeneration in cerebellar, vestibular, and cochlear nuclei. In addition, *swe* mice are the first to model essential elements of human generalized absence epilepsy, including a combination of 3-Hz spike-wave and tonic-clonic seizures. The ubiquitously expressed *Nhe1* gene was identified as a positional candidate. A single base A-to-T transversion was found that generates a premature stop codon severely reducing mRNA and protein expression and abolishing NHE1 activity in cultured fibroblasts from mutants. This functional null allele of *Nhe1* is the first disease-causing mutation identified in an *Nhe* gene.



Spontaneous mouse mutant p.Lys442*



Occasionally survive beyond 8 weeks

Am J Physiol. 1999 Apr;276(4 Pt 1):C788-95.

Targeted disruption of the murine Nhe1 locus induces ataxia, growth retardation, and seizures.

Bell SM[1], Schreiner CM, Schultheis PJ, Miller ML, Evans RL, Vorhees CV, Shull GE, Scott WJ.

FUNCTIONAL STUDY RESULTS

Collaboration with NHE1 specialists:
Larry Fliegel and **Xiuju Li**

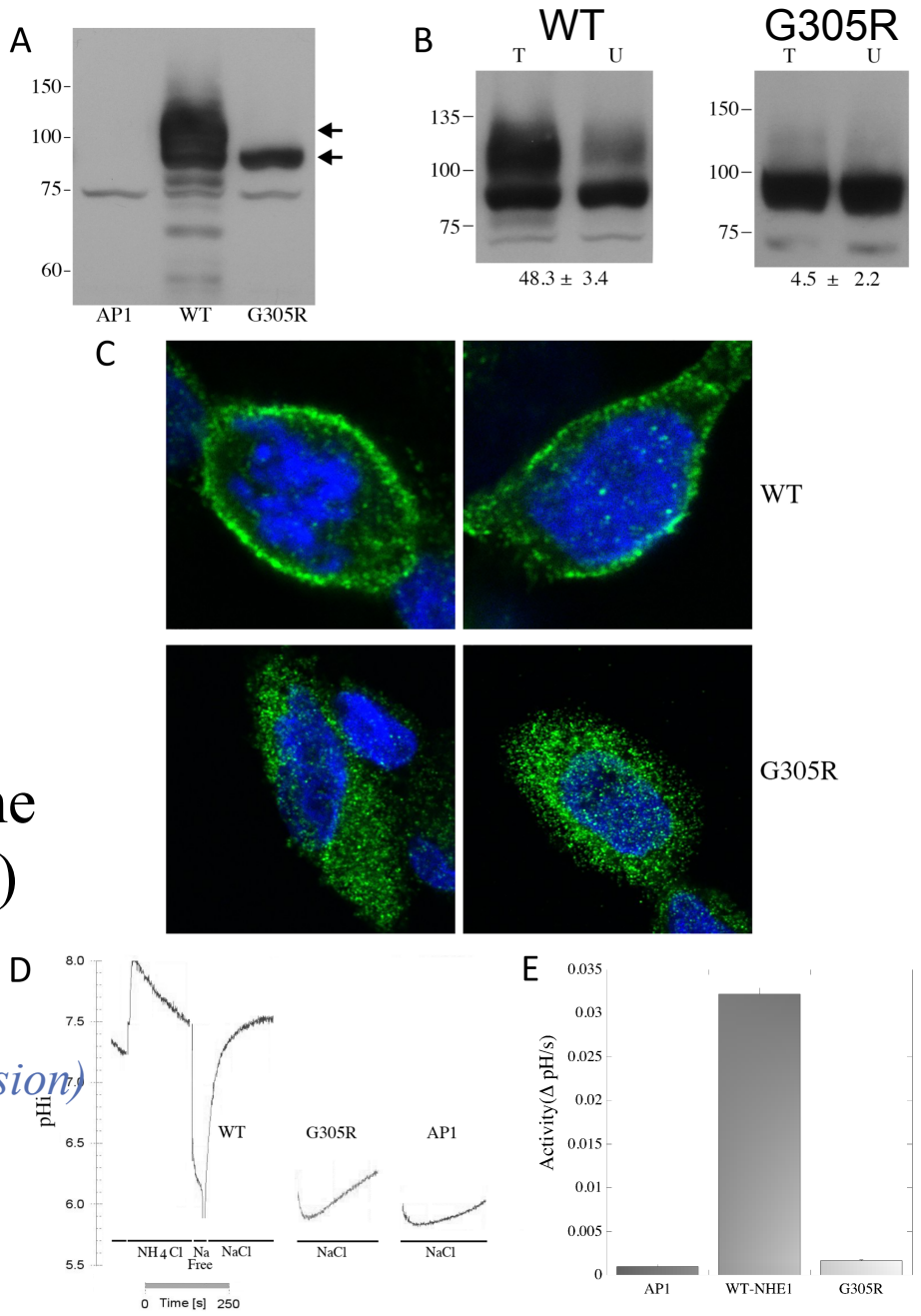
Stably transfected AP-1 cells
WT or mutant cDNA

Mutant protein:

- de-glycosylated
- almost completely absent from the cell surface (>90% miss-targeted)
- Very low residual activity (2%)

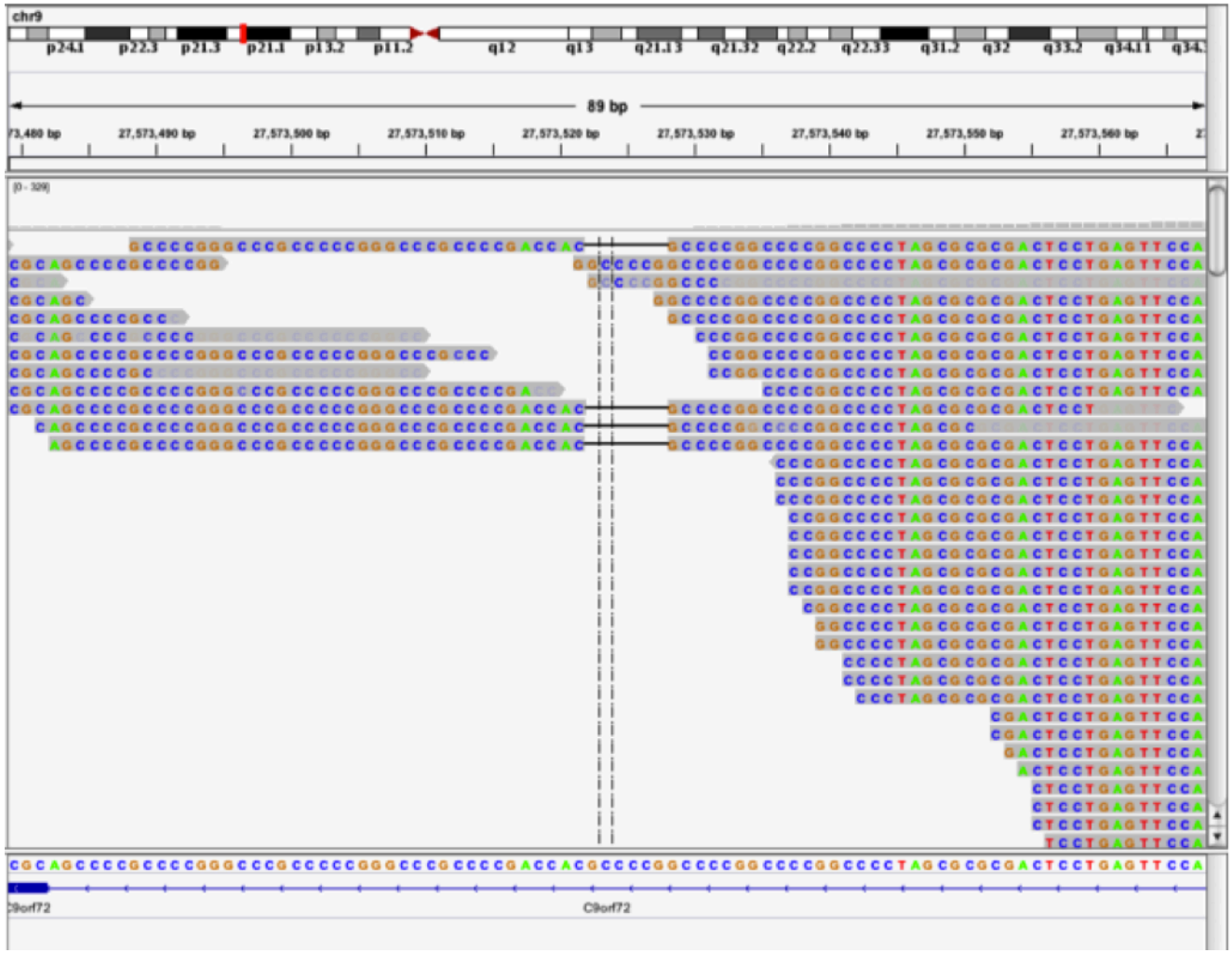
intracellular pH measurement
(cell-permeant dye with pH dependent emission)

Conclusion → almost complete loss of function



CGH arrays and next generation sequencing are not suited for identification of trinucleotide expansions

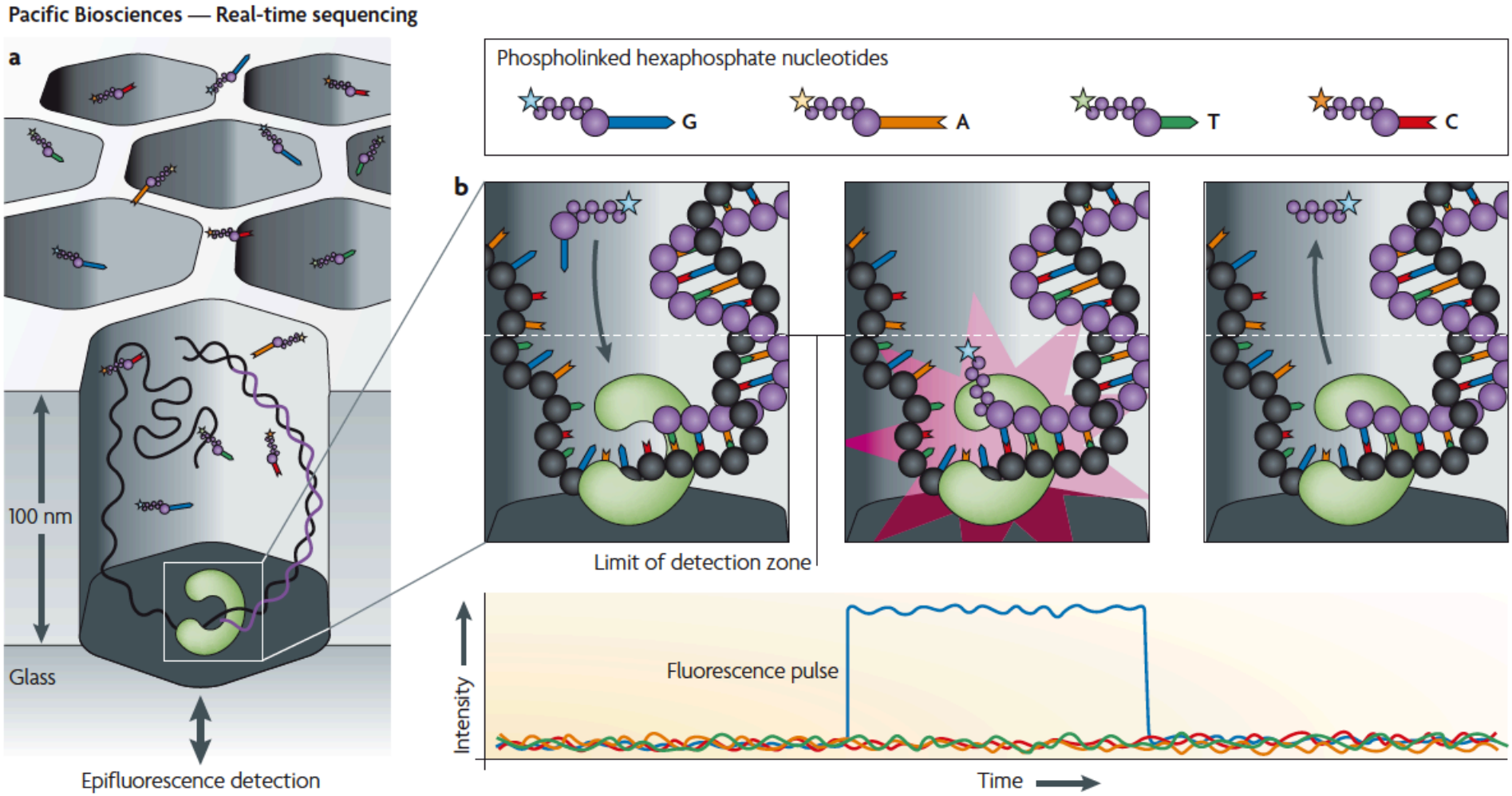
—> need for genetic linkage studies



Hexanucleotide Repeat Expansion in C9ORF72 As the Cause of familial ALS-FTD,
Renton et al Neuron 2011

Future

- exome or whole genome (WES versus WGS)
- third generation sequencing : single DNA molecule sequencing



LOD score