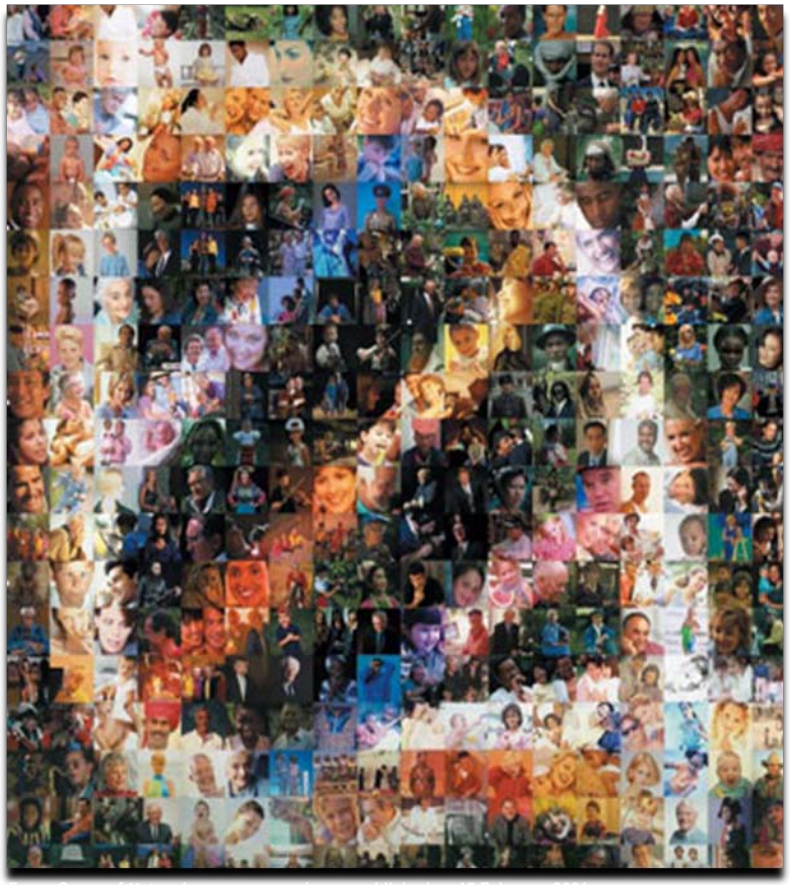
NEXT GENERATION SEQUENCING OF MULTIPLE GENES IN PARALLEL IN GENETICALLY HETEROGENEOUS DISORDERS AFTER ENRICHMENT

Hans Scheffer

SEQUENCE CAPTURE IN HETEROGENEOUS MONOGENIC DISORDERS

Disease	# (known) Major Genes
Breast cancer	2
Recessive ataxias	>20
Blindness	>200
Mental retardation	~1000



CHALLENGES

- **Sequencing of multiple genes in parallel for heterogeneous hereditary disorders**
- **Multiplexing of samples for sequencing to increase throughput and reduce costs**
- **Automated data-analysis, including**
- **Interpretation of potential pathogenicity of identified variants**

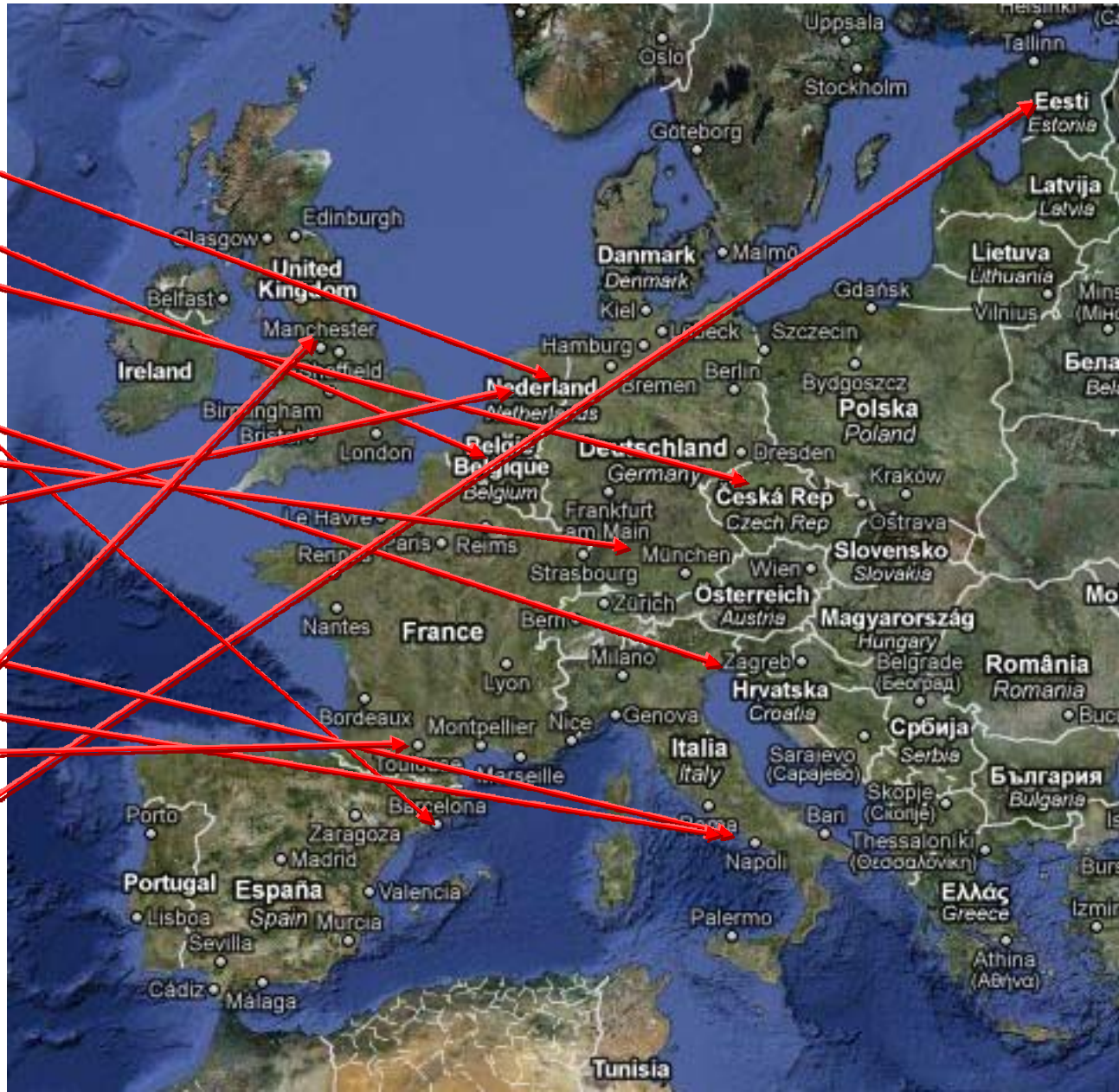
TECHGENE

- **Is a network of ISO15189 accredited clinical molecular genetic laboratories and renowned genome research laboratories**

- **Has been positioned as a technological innovation platform connected with the EUGT Network of Excellence dealing with Quality Standards in Clinical (Laboratory) Genetics**

TECHGENE - WHO ARE WE?

PARTNER
Scheffer/Veltman
Matthijs/Cuppens
Macek
Estivill
Gasparini
Riess
Bakker/Den Dunnen
Banfi
Nigro
Cambon-Thomsen
Payne
Sak

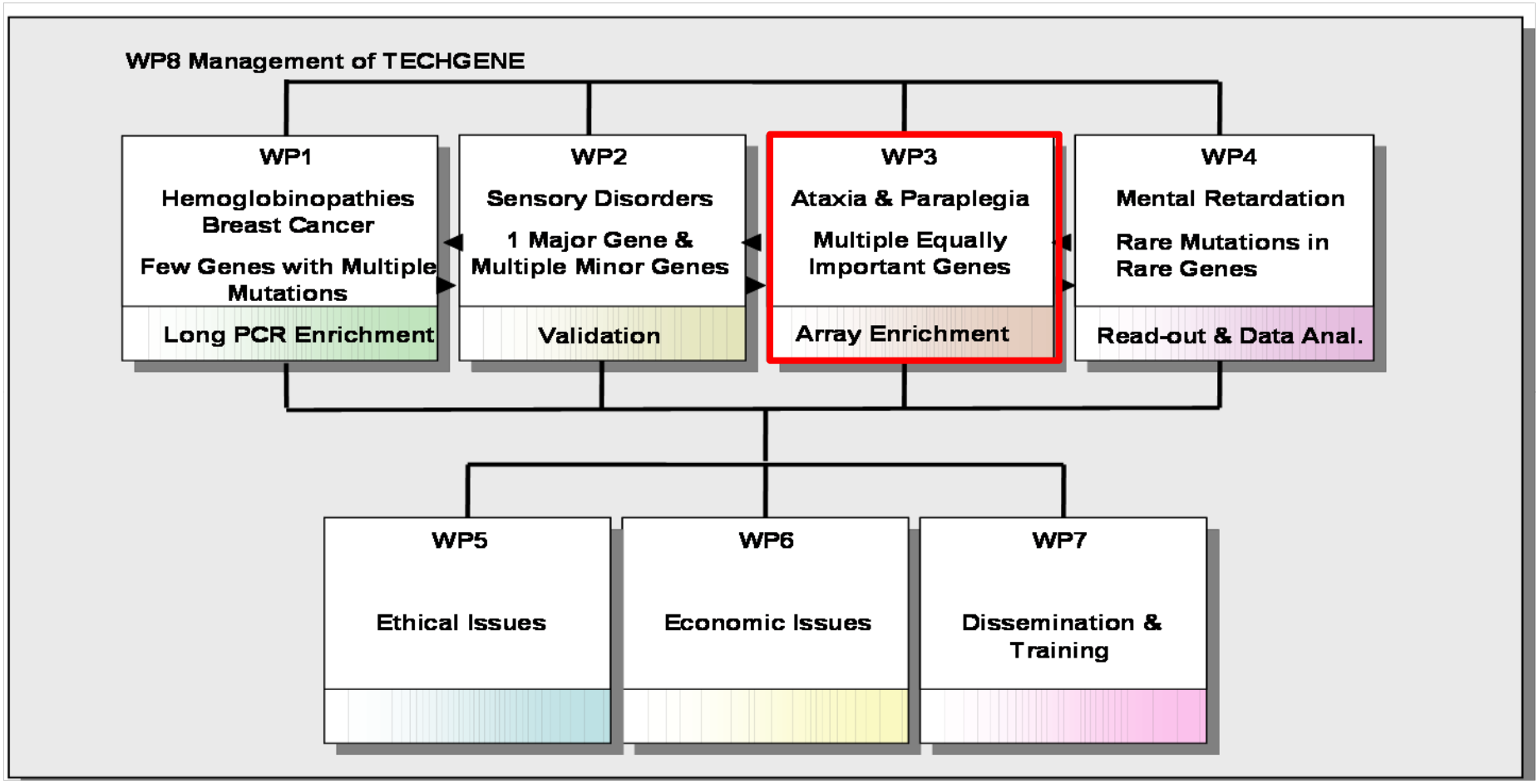


PLATFORMS

COMPANY	TECHNOLOGY	CURRENT READ LENGTH	CURRENT THROUGHPUT
Roche/454 (GS FLX Titanium)	Sequencing by synthesis	300-500 bases	500 Mb/10h
Illumina (Solexa) (Illumina Genome Analyzer System)	Sequencing by synthesis	35 bases	4.5-6.0 Gb/2 days
Life Technologies (SOLiD3)	Sequencing by ligation	50 bases	30 Gb/5 days


600 million
map-able 50bp reads
→ 30Gb
1 million
map-able 500bp reads
→ 500Mb

TECHGENE - HOW?
Overview of Plan of Investigation



→ Next Generation Sequencing in Diagnostics

THE MAIN CHALLENGE: INTERPRETATION!

- Step 1: Identify all potential genetic variation (SNPs,CNVs)
- Step 2: Exclude false positives

Data of first whole genome sequencing results look great but…..
0,1% FP = 6 million bp!
1 mutation may cause disease!

Asian Genome
Wang. Nature 2008

- Step 3: Link true genetic variants to phenotype

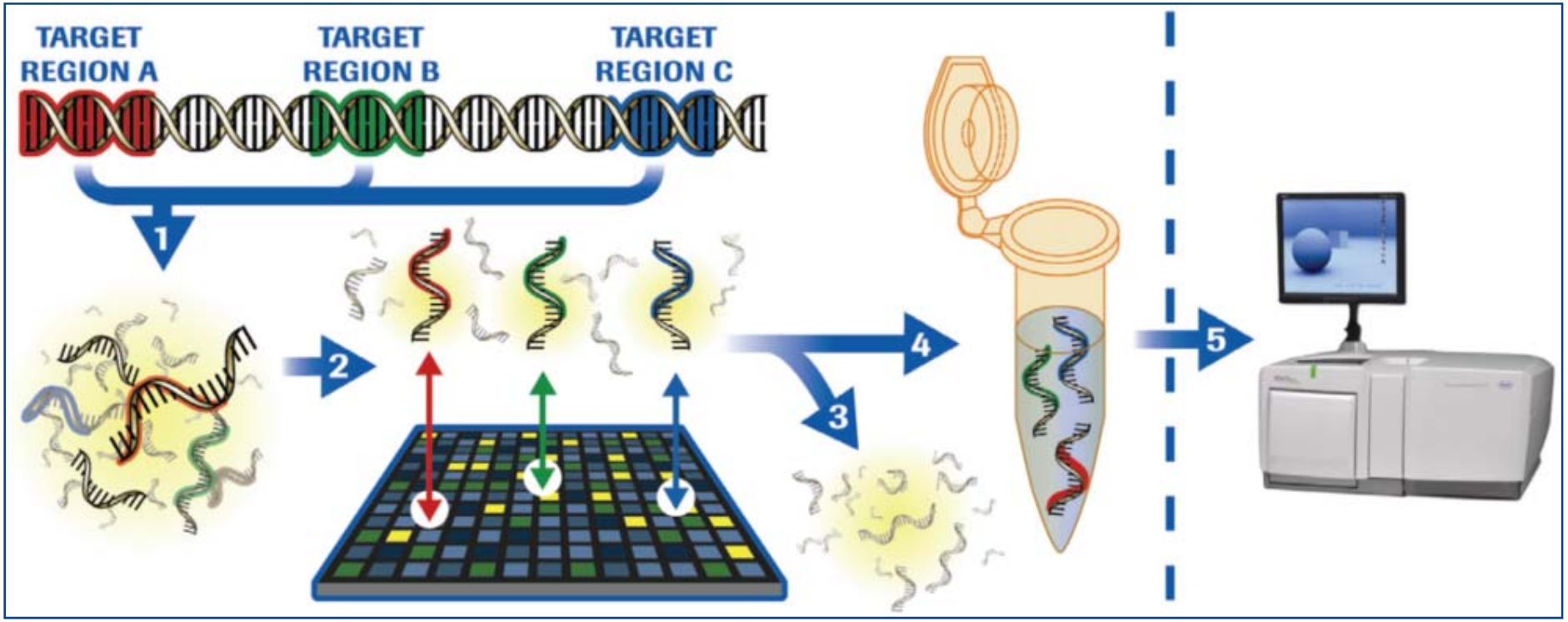
Whole genome sequencing is not suitable for diagnostics yet

Enrichment Strategies Including Sequence Capture

ENRICHMENT STRATEGIES

- Long range PCR
- Amplicon sequencing
- Array-based gene capture (e.g. Nimblegen)
- Gene capture in solution (e.g. Agilent SureSelect)
- Alternative approaches (e.g. Raindance)

ARRAY-BASED SEQUENCE CAPTURE

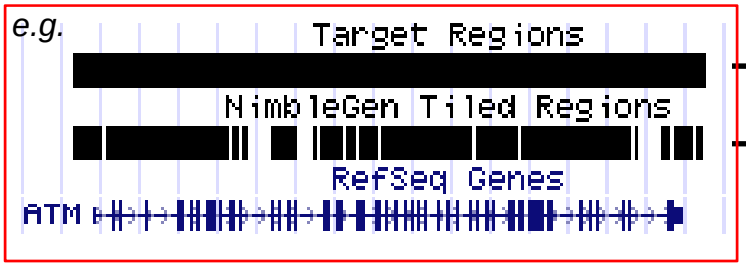


1. DNA preparation (adaptor ligation)
2. Hybridization
3. Stringent washing
4. Elution & ligation mediated (LM)-PCR
5. Sequencing

SEQUENCE CAPTURE WORKFLOW

- Select your targets

7 genes:
APTX, SYNE1, TDP1, SACS, SETX, ATM, FXN
- all exons & introns
- 886kb



Targeted region

Tiled region, i.e. represented on the array

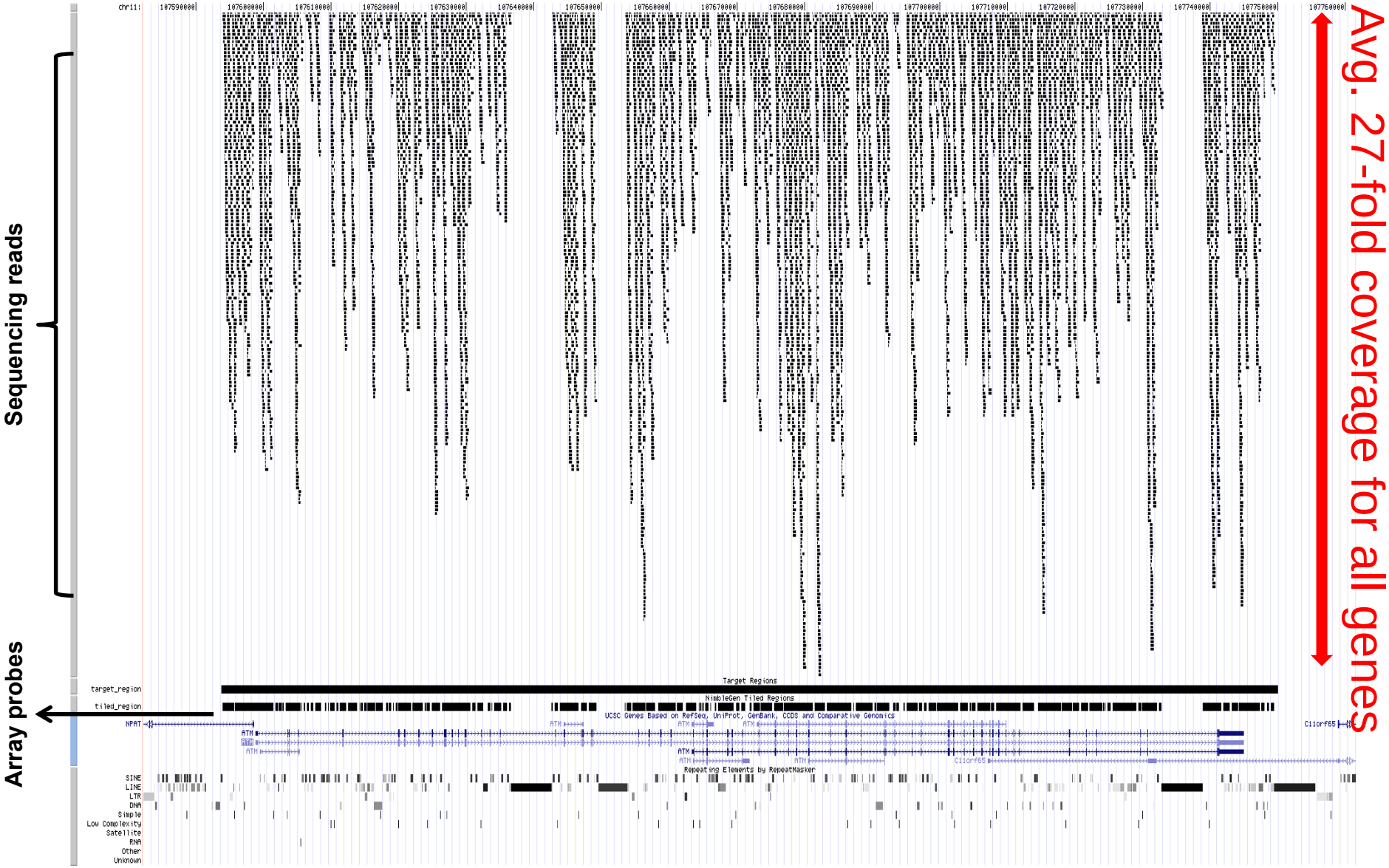
- Enrichment of 10 samples
- 8/10 samples carried known disease causing mutations (10 mutation alleles) for **AR ataxia** (severe movement disorder)
- Roche 454 GS FLX Titanium Sequencing (1/4 run each)

>65Mb map-able sequence

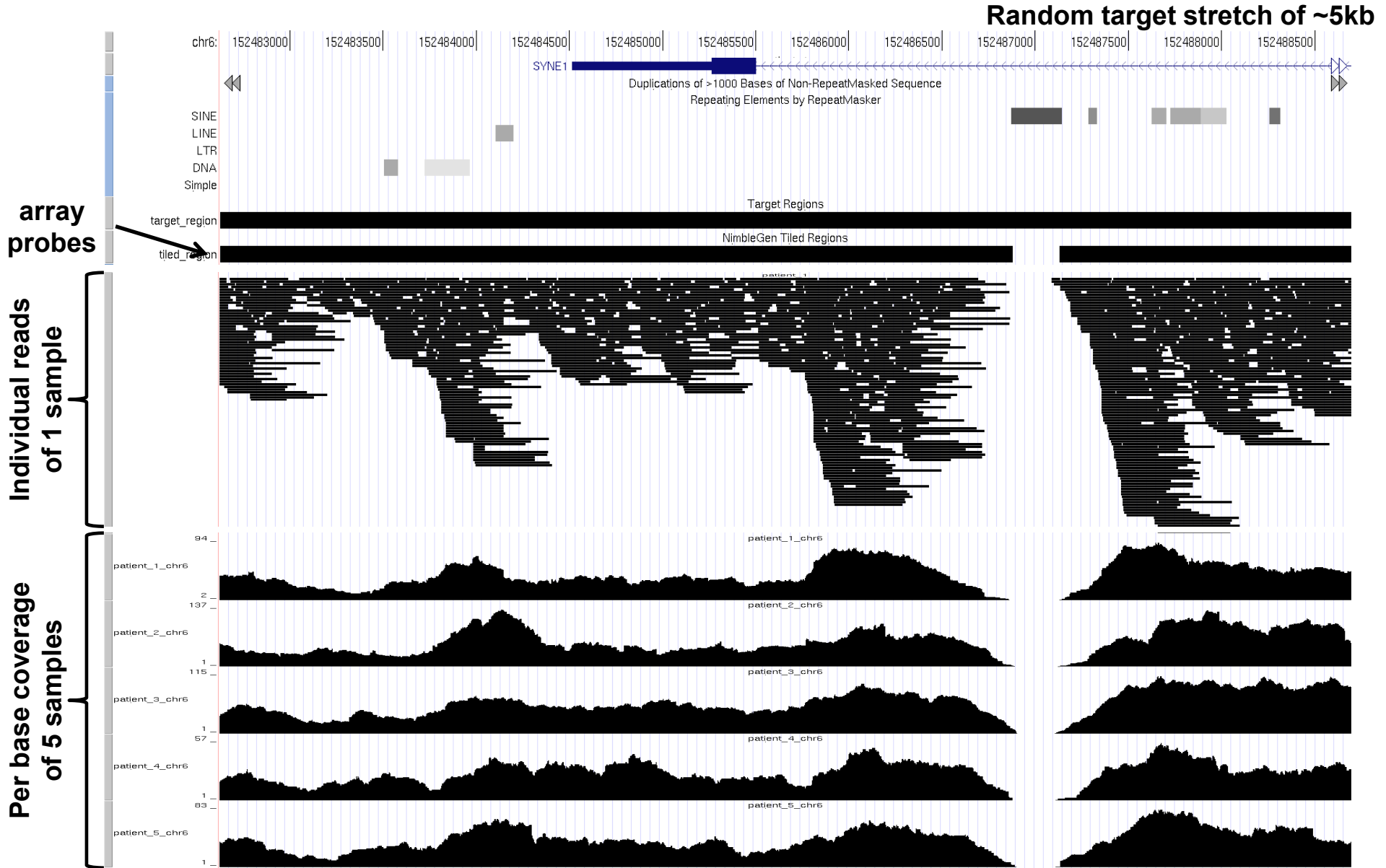
AVERAGE MAPPING DATA

Target bases covered	**98.05%**
Mean sequence depth in target regions	**27 fold**
Mean sequence depth in coding regions	**31 fold**
Mean sequence depth in non-coding regions	**24 fold**
Target region bases not covered	**1.95%**

ENRICHMENT WORKS !



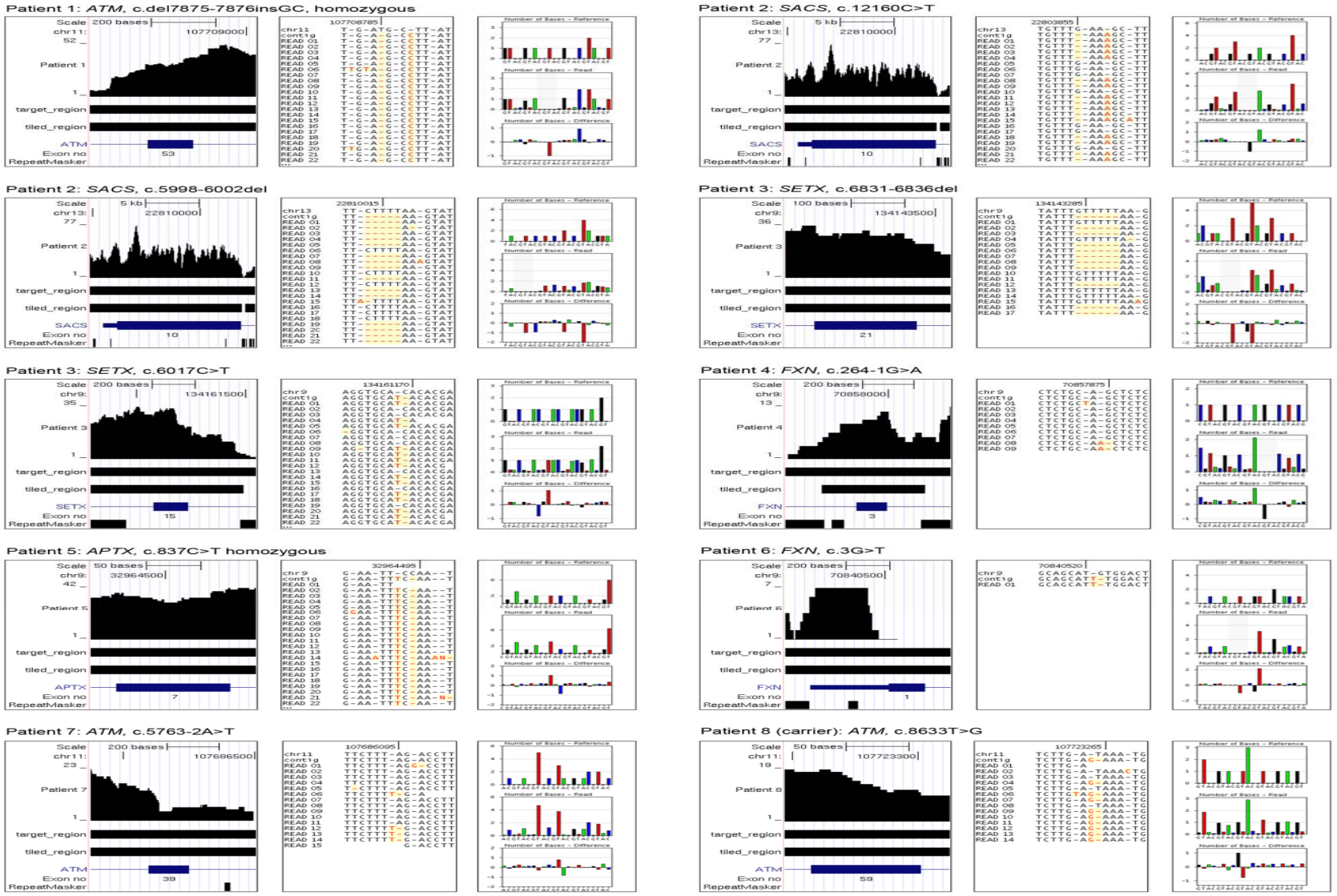
ENRICHMENT AND SEQUENCING IS REPRODUCIBLE !



ARE THERE REGIONS NOT COVERED ?

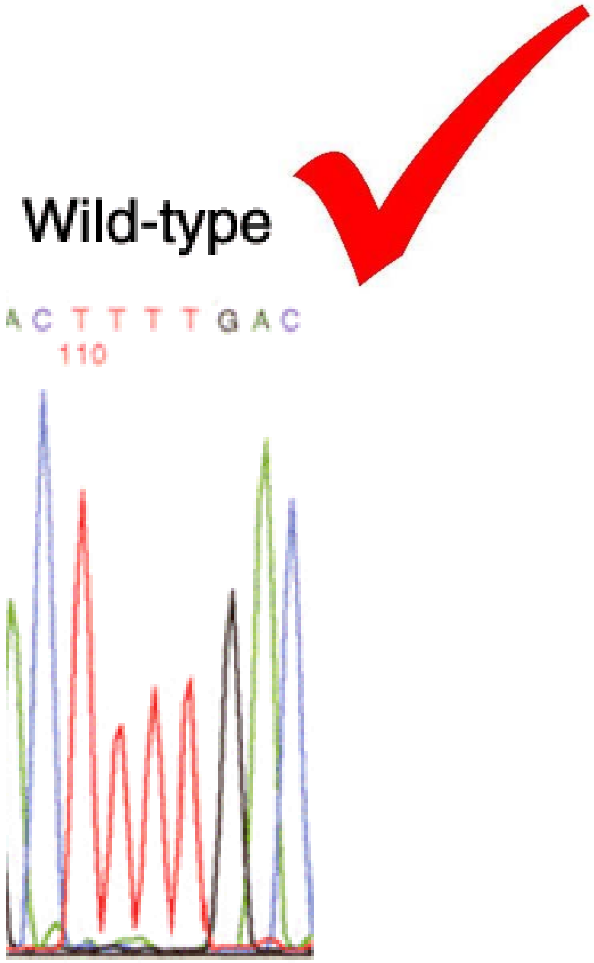
Only two exons (exon 1 of FXN, and 2nd coding exon of SACS) are not covered at all, due to very high CG content (~65%)

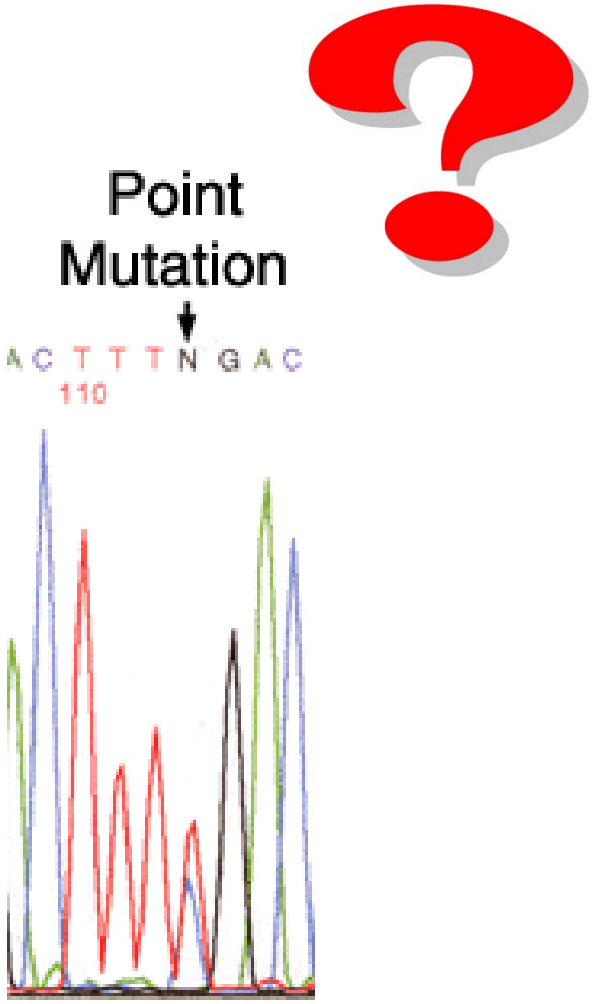
ALL MUTATIONS IDENTIFIED



Hoischen et al., Human Mutation, 2010

CAN WE DETECT PATHOGENIC MUTATIONS?



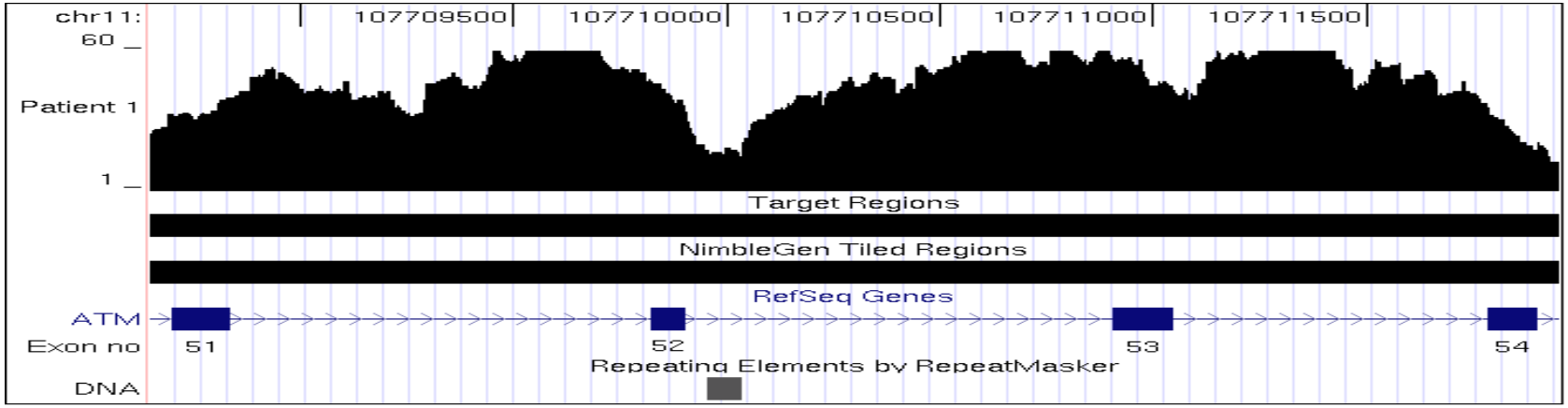


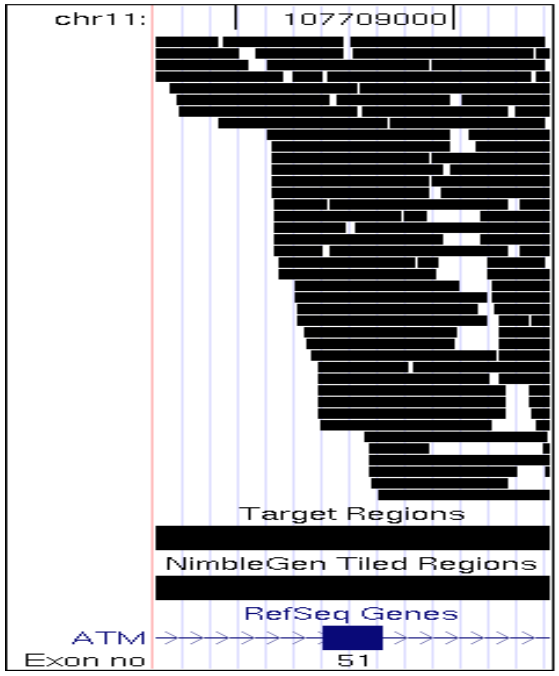
PRIORITIZATION HIERARCHIES FOR THE DISEASE CAUSING MUTATIONS

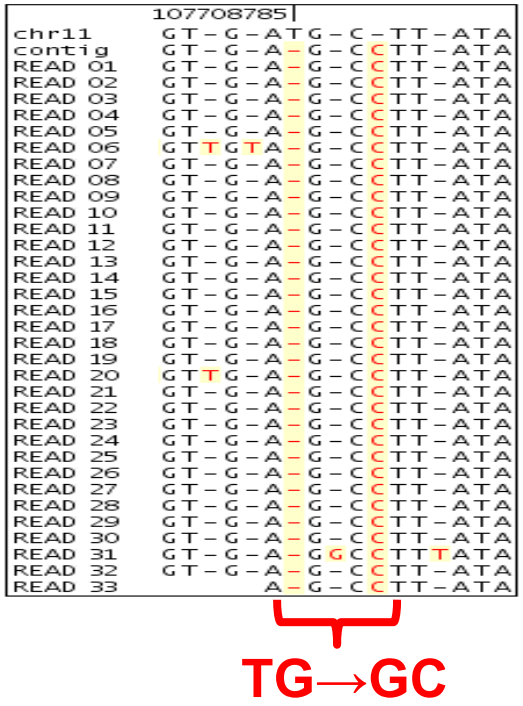
- Mutations that could give rise to premature protein truncating mutations *i.e.* stop mutations, variants at potential splice sites and exonic deletions/insertions potentially resulting in frame shifts
- All exonic missense variants not known in dbSNP with at least 20% of variant reads. (This reduced the number of variants to an average of 7 per sample)
- Ranking these variants based upon evolutionary conservation of the affected nucleotide using vertebrate PhyloP and PhastCons scores

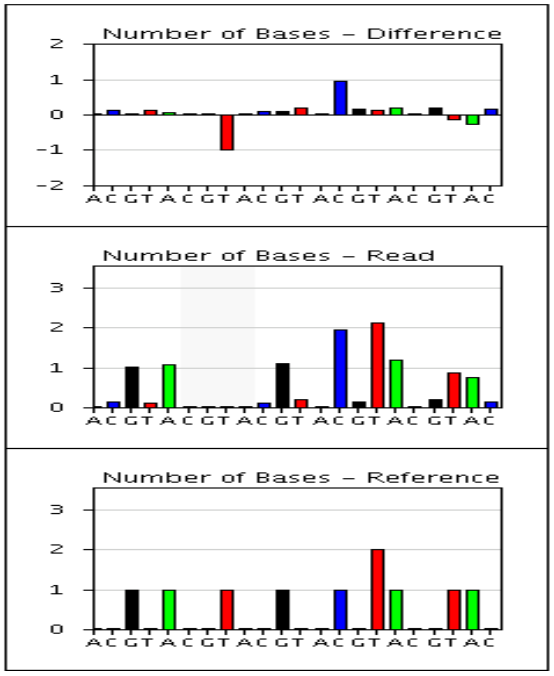
→ All known pathogenic mutations were ranked on the top position

HOMOZYGOUS MUTATION IN THE ATM GENE









MUTATION IDENTIFICATION AGAINST A BACKGROUND OF APPR. 800 HIGH CONFIDENTIALITY (HC) DIFFERENCES

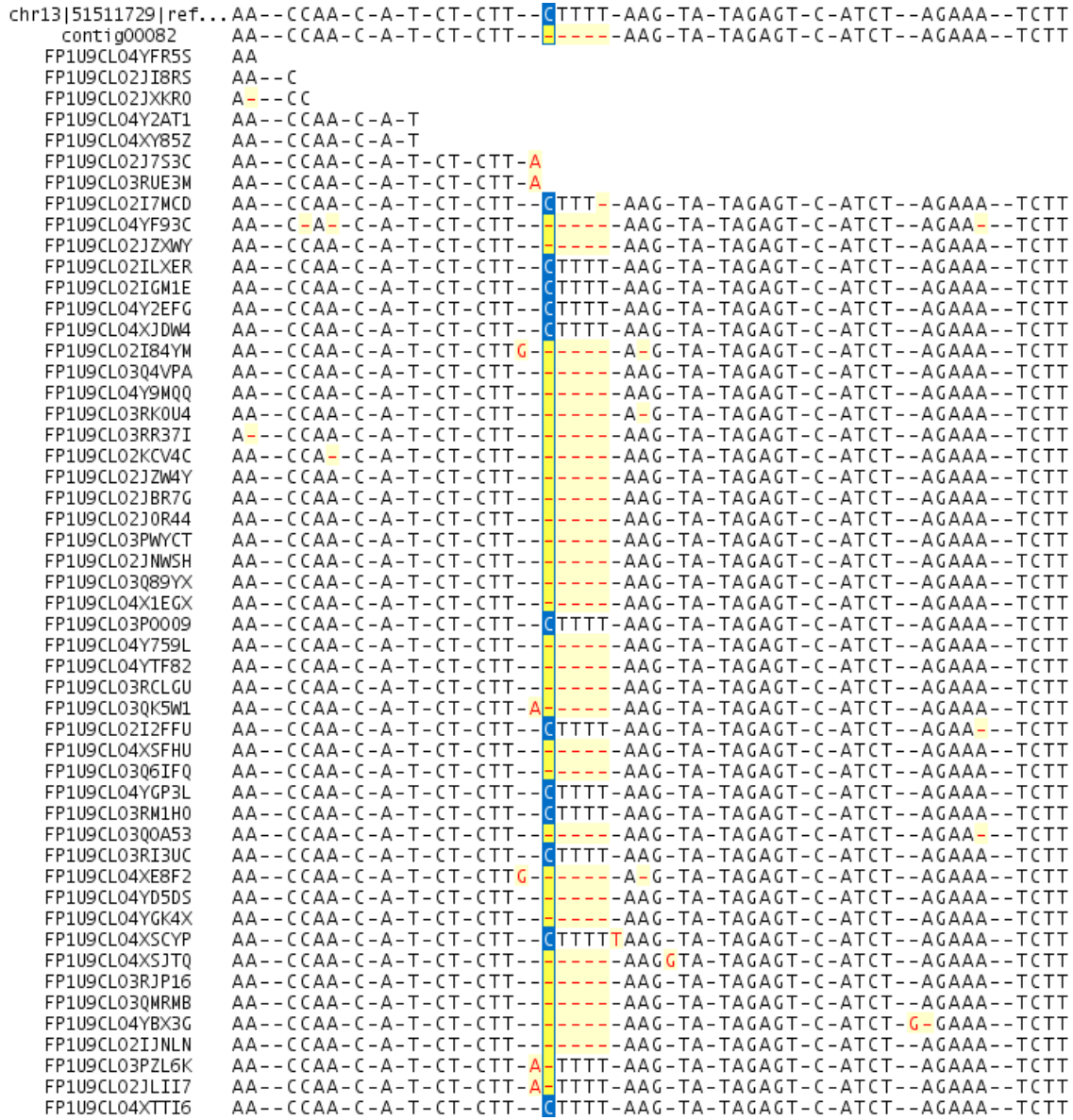
AT Patient: c.7875-7576 TG>GC (homozygous mutation *ATM* gene)

Roche Mapping software (HCdifferences) output:
* exclude variants outside coding regions
* exclude known SNPs (dbSNP)
* sort for conservation (phastCons & phyloP)

chromosome	start	stop	reference	mutation	reads	variation reads	% variation	ref amino acid	mutation amino acid	coding frame	gene name	phast Cons	phyloP
chr11	107,708,785	107,708,786	TG	GC	22	22	100	DA	EP	1	ATM	1	3.8459
chr13	22,812,629	22,812,629	T	C	21	2	14	K	R	-2	SACS	1	1.7722
chr6	152,735,877	152,735,877	T	C	12	6	50	E	E	-1	SYNE1	1	1.1568
chr14	63,491,288	63,491,288	T	A	28	3	11	M	K	3	SYNE2	0.992	4.4954
chr6	152,699,834	152,699,835	CT	AC	15	9	60	K	S	-1	SYNE1	0.992	1.9699

→ Top 1 on the list is the known mutation!

HETEROZYGOUS MUTATION IN THE SACS GENE



ARSACS patient:

A 5 bp heterozygous CTTTT deletion in exon 5 of the *SACS* gene

MUTATION IDENTIFICATION AGAINST A BACKGROUND OF APPR. 800 HIGH CONFIDENTIALITY (HC) DIFFERENCES

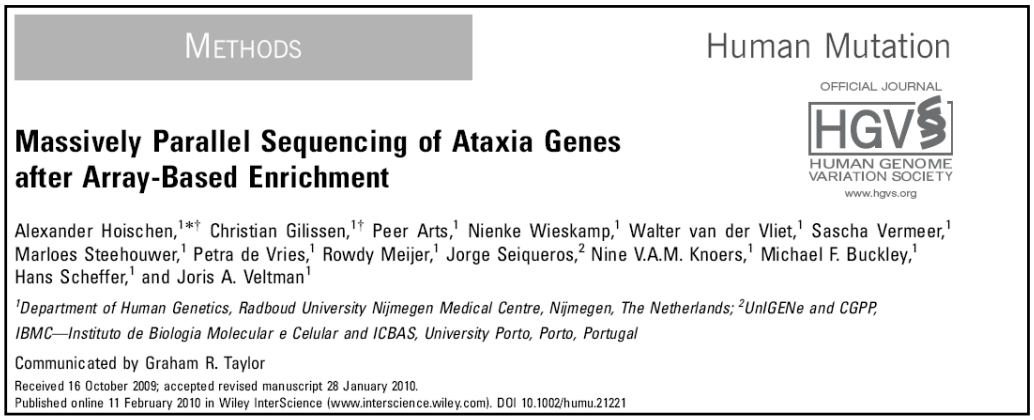
ARSACS patient: compound heterozygous for
SACS c.5559-5563 delCTTTT / c. 11719C>T (p.Q3907X)

Roche Mapping software (HC differences) output:
* exclude variants outside coding regions
* exclude known SNPs (dbSNP)
* sort for conservation (phastCons & phyloP)
→ Top 2 on the list are the known mutations!

chromosome	start	stop	reference	mutation	reads	variation reads	% variation	ref amino acid	mutation amino acid	coding frame	gene name	phast Cons	phyloP
chr13	22,803,855	22,803,855	G	A	35	21	60	Q	*	-2	SACS	1	6.5854
chr13	22,810,013	22,810,017	CTTTT	-	27	17	63			-2	SACS	1	5.3822
chr17	42,576,285	42,576,285	C	T	4	4	100	S	N	-3	CDC27	1	5.2978
chr14	63,498,047	63,498,047	A	G	30	3	10	Q	R	2	SYNE2	1	5.1516
chr14	63,695,793	63,695,793	A	G	25	3	12	D	G	1	SYNE2	1	3.4163
chr14	63,700,048	63,700,048	T	A	28	3	11	F	Y	1	SYNE2	1	3.3456

PROOF OF CONCEPT

- combination Roche 454 GS FLX Titanium and NimbleGen array works
- 7 disease genes
- Enrichment works, ~80% on or near target
- Mean coverage: 24x
- All known mutations were identified
- Detection of the disease causing mutation depends on the coverage
- Coverage: depending on the size of the targeted region, GC content, repeat elements
- At least 15x coverage
- Prioritization due to conservation



METHODS

Human Mutation

OFFICIAL JOURNAL HGVS HUMAN GENOME VARIATION SOCIETY www.hgvs.org

Massively Parallel Sequencing of Ataxia Genes after Array-Based Enrichment

Alexander Hoischen,[1*†] Christian Gilissen,[1†] Peer Arts,[1] Nienke Wieskamp,[1] Walter van der Vliet,[1] Sascha Vermeer,[1] Marloes Steehouwer,[1] Petra de Vries,[1] Rowdy Meijer,[1] Jorge Seiqueros,[2] Nine V.A.M. Knoers,[1] Michael F. Buckley,[1] Hans Scheffer,[1] and Joris A. Veltman[1]

[1]Department of Human Genetics, Radboud University Nijmegen Medical Centre, Nijmegen, The Netherlands; [2]UnIGENe and CGPP, IBMC—Instituto de Biologia Molecular e Celular and ICBAS, University Porto, Porto, Portugal

Communicated by Graham R. Taylor

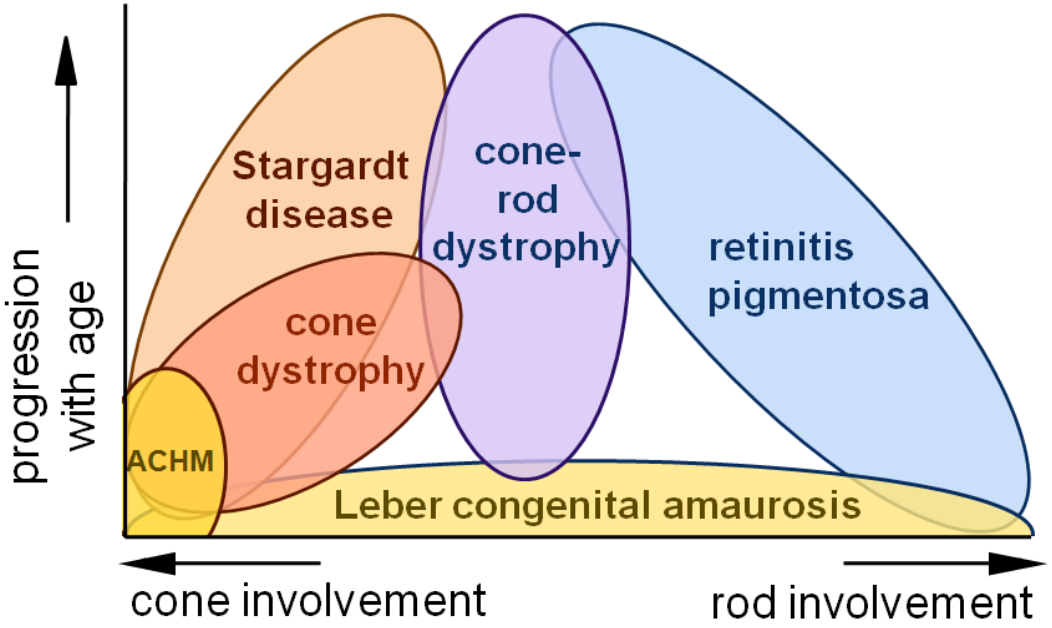
Received 16 October 2009; accepted revised manuscript 28 January 2010.
Published online 11 February 2010 in Wiley InterScience (www.interscience.wiley.com). DOI 10.1002/humu.21221

NEXT PHASE

- Design of seq cap tools for parallel sequencing of appr. 100 genes, in hereditary blindness, in hereditary movement disorders, in mitochondrial disorders

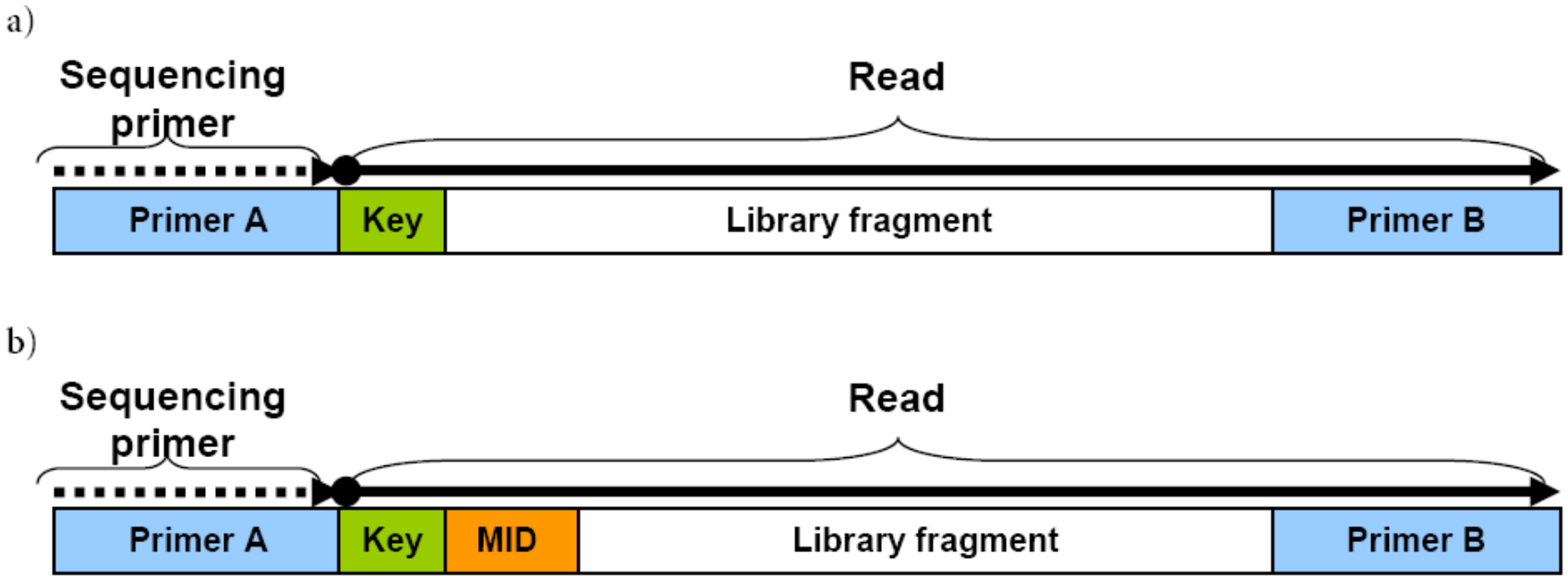
- Validation of amplicon sequencing-based NGS approach for BRCA1/2

GENE CAPTURE FOR HEREDITARY BLINDNESS DIAGNOSTICS?



- clinically and genetically heterogeneous
- focus on RP and LCA
- mutations described for >100 genes
- only known mutations detectable on microarrays from Asper
- diagnostic yield by current Sanger sequencing-based strategy is very low!

BARCODING/MULTIPLEXING
(12 samples)



MID = multiplex identifiers

DATA ANALYSIS/INTERPRETATION OF SAMPLES WITH UNKNOWN MUTATIONS?

- Design makes data analysis challenging
- Many variations found
- ***Every*** gene is a candidate gene
- ***Every*** variation potentially pathogenic
- Several known blindness mutations are in dbSNP
- Include known blindness-causing mutations (e.g. HGMD) in analysis pipeline

Data analysis/interpretation major bottleneck
(taking into account also possibility of digenic inheritance/modifier genes…)

CONCLUSIONS

- NGS will be suitable for diagnostics
- We need good coverage!
- Increase throughput (e.g. multiplexing)
- Decrease costs
- Robust handling for technicians in routine diagnostics
- Possibility for automation
- Careful data analysis/interpretation (further development of pathogenic mutation identification pipeline)

We can distinguish known mutations from 800-1000 variants; sequence capture almost suitable for diagnostics

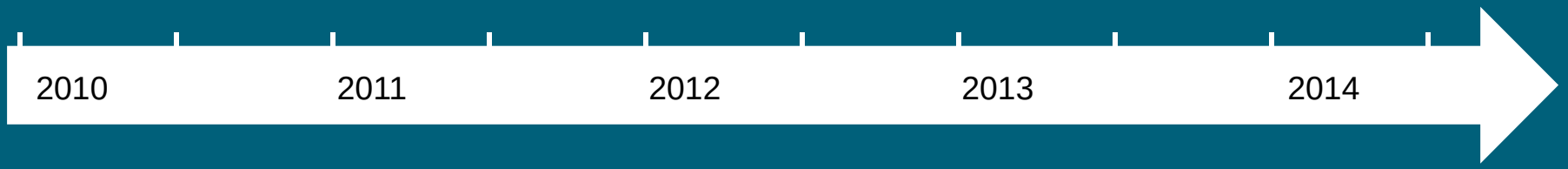
Exome sequencing successfully applied in gene identification (e.g. Schinzel-Giedion*)

-

Soon also as diagnostic test (e.g. for very heterogeneous disorders caused by *de novo* mutations)

*** Hoischen et al. (2010) De novo mutations of SETBP1 cause Schinzel-Giedion syndrome. Nat Genet 42(6):483-5**

Future perspectives in molecular genetics



Research:

- Exome sequencing
- Whole genome sequencing

Diagnostics:

- Targeted re-sequencing in complex monogenic diseases
- Exome sequencing
- Whole genome sequencing

Diagnostic testing first, then detailed clinical investigation?

SEVENTH FRAMEWORK PROGRAMME

TECHGENE

NGS DIAGNOSTICS KNOWLEDGE NETWORK

More info at: www.techgene.eu or www.techgene.org

Register at: info@techgene.org

ACKNOWLEDGEMENTS

Alexander Hoischen

Kornelia Neveling

Peer Arts

Christian Gilissen

Nienke Wieskamp

Joep de Ligt

Michael Buckley

Joris Veltman

Hans Scheffer

Sascha Vermeer

Nine van Slobbe-Knoers

Berry Kremer

Jorge Sequeiros

SEVENTH FRAMEWORK PROGRAMME

TECHGENE

ZonMw

www.techgene.eu

www.techgene.org