Genetics, genomics and gene regulation

Steve Twigg
Clinical Genetics Group
WIMM
Oxford

Weatherall Institute of Molecular Medicine

UNIVERSITY OF OXFORD

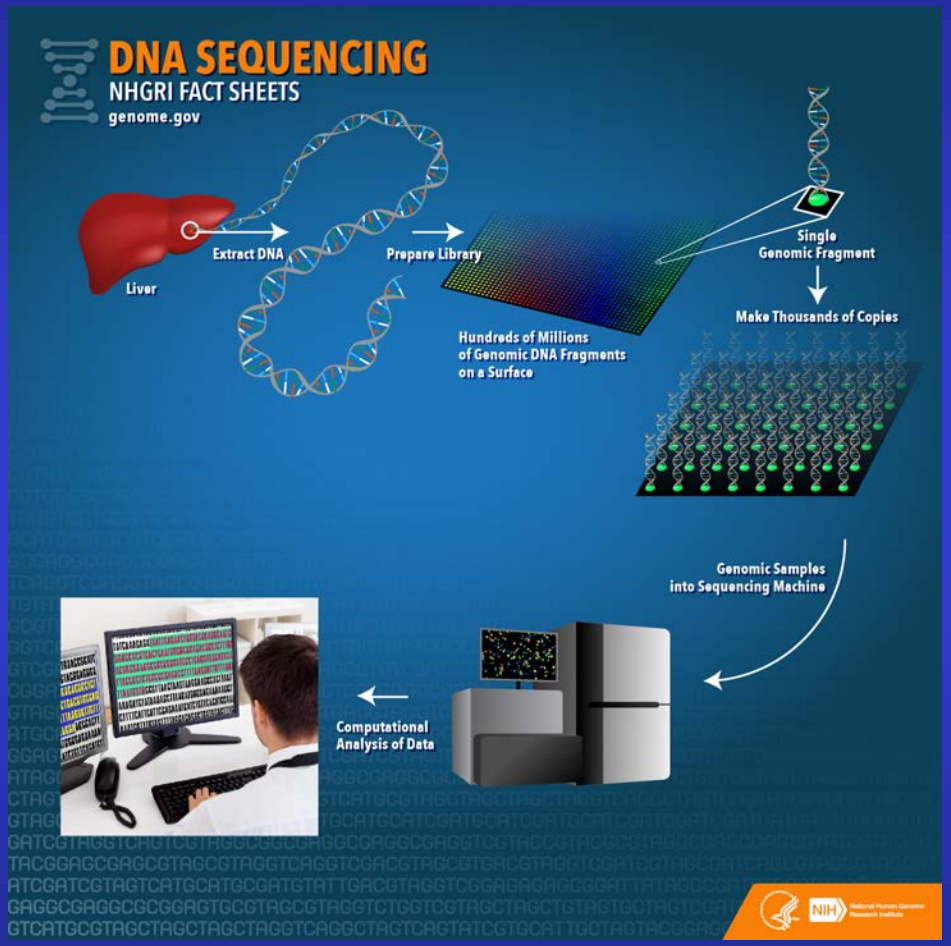
The MRC Weatherall Institute of Molecular Medicine is a strategic alliance between the Medical Research Council and the University of Oxford

MRC Medical Research Council

WIMM DPhil Course Nov 2019

Next Generation Sequencing (NGS) impact

Disease gene identification

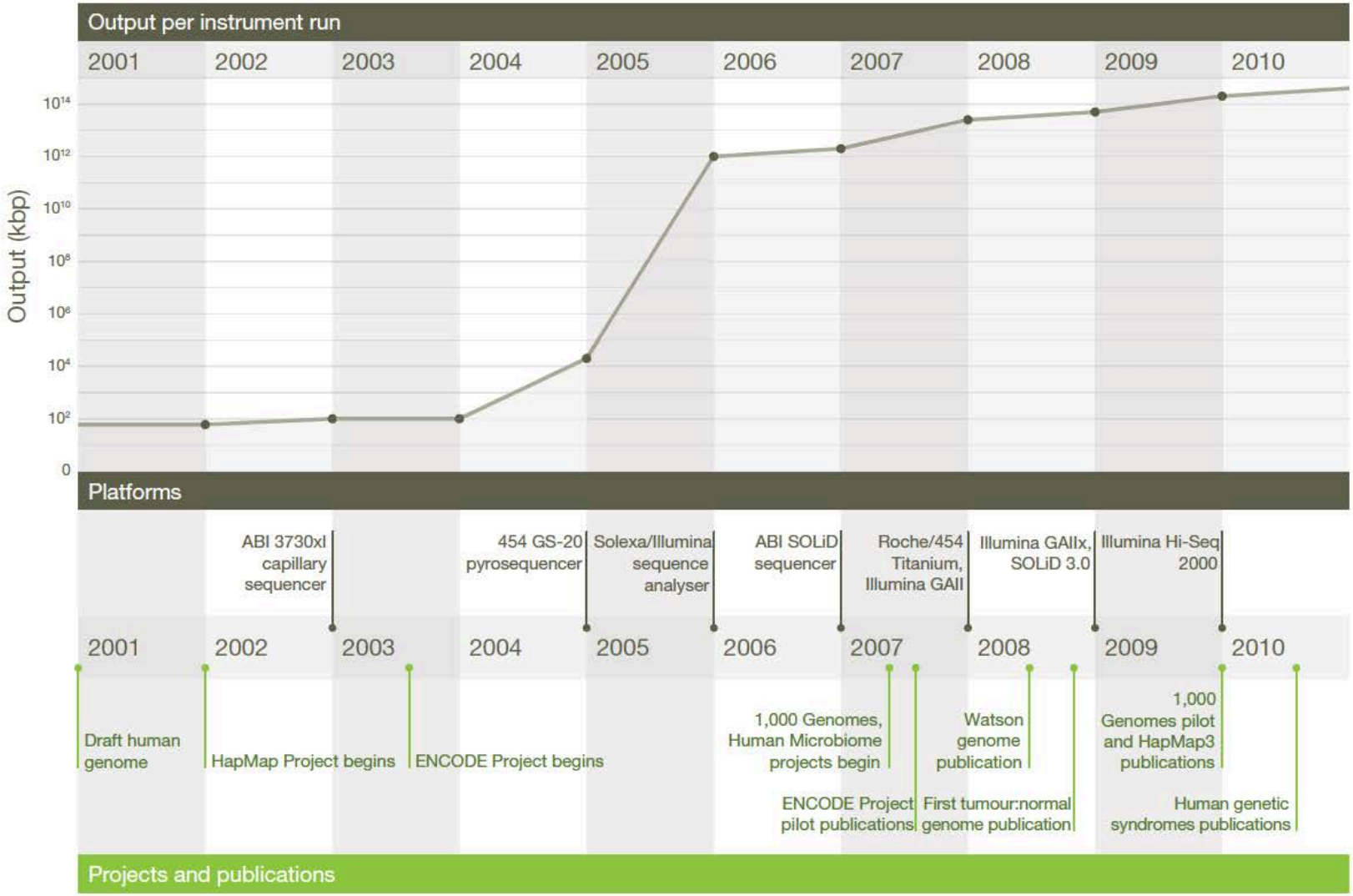


Study of gene regulation

stephen.twigg@imm.ox.ac.uk

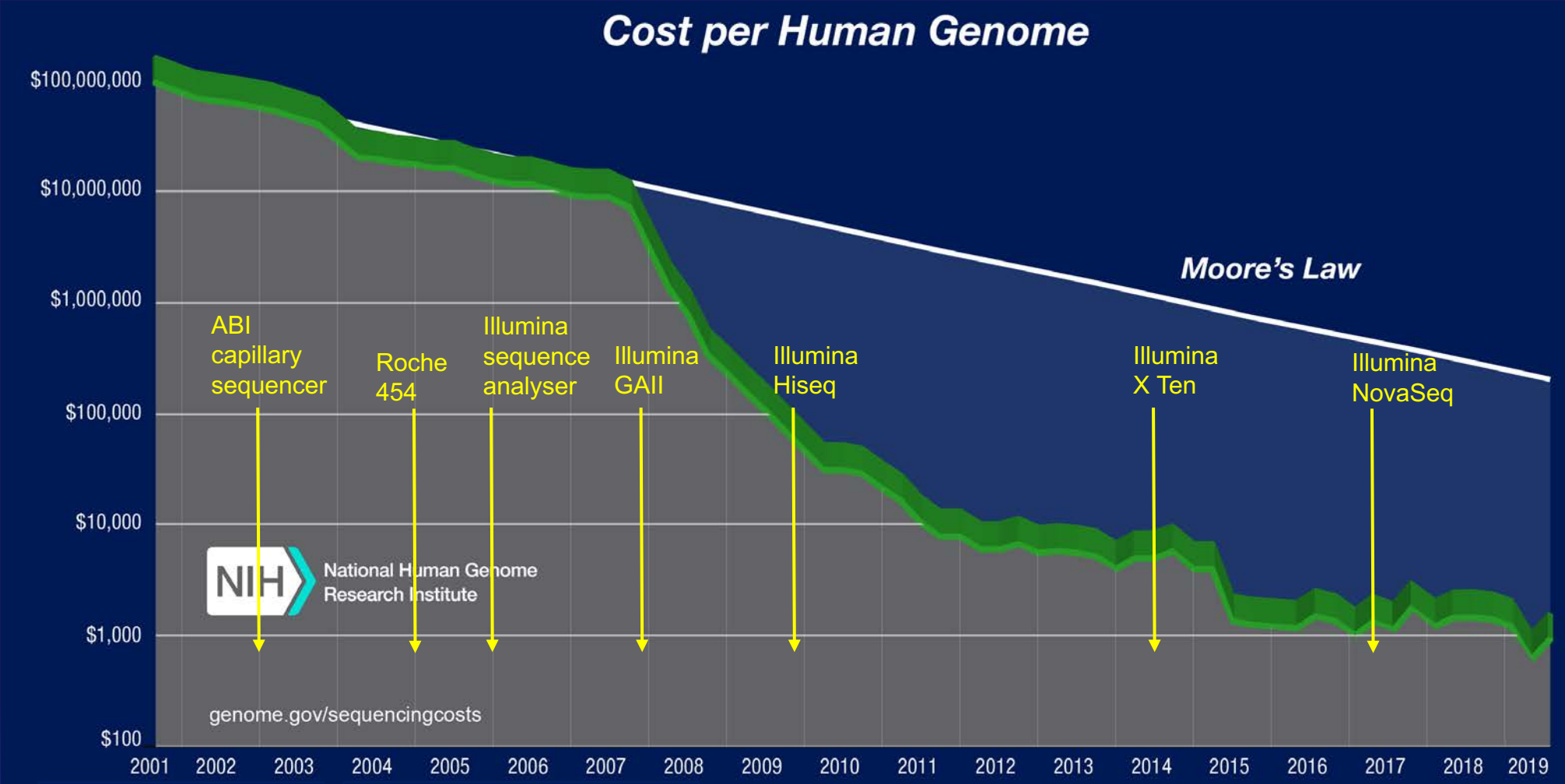
jim.hughes@imm.ox.ac.uk

DNA sequencing output



Mardis
Nature 470:198
(2011)

Figure 1 | Changes in instrument capacity over the past decade, and the timing of major sequencing projects. Top, increasing scale of data output per run plotted on a logarithmic scale. Middle, timeline representing major milestones in massively parallel sequencing platform introduction and instrument revisions. Bottom, the timing of several projects and milestones described in the text.

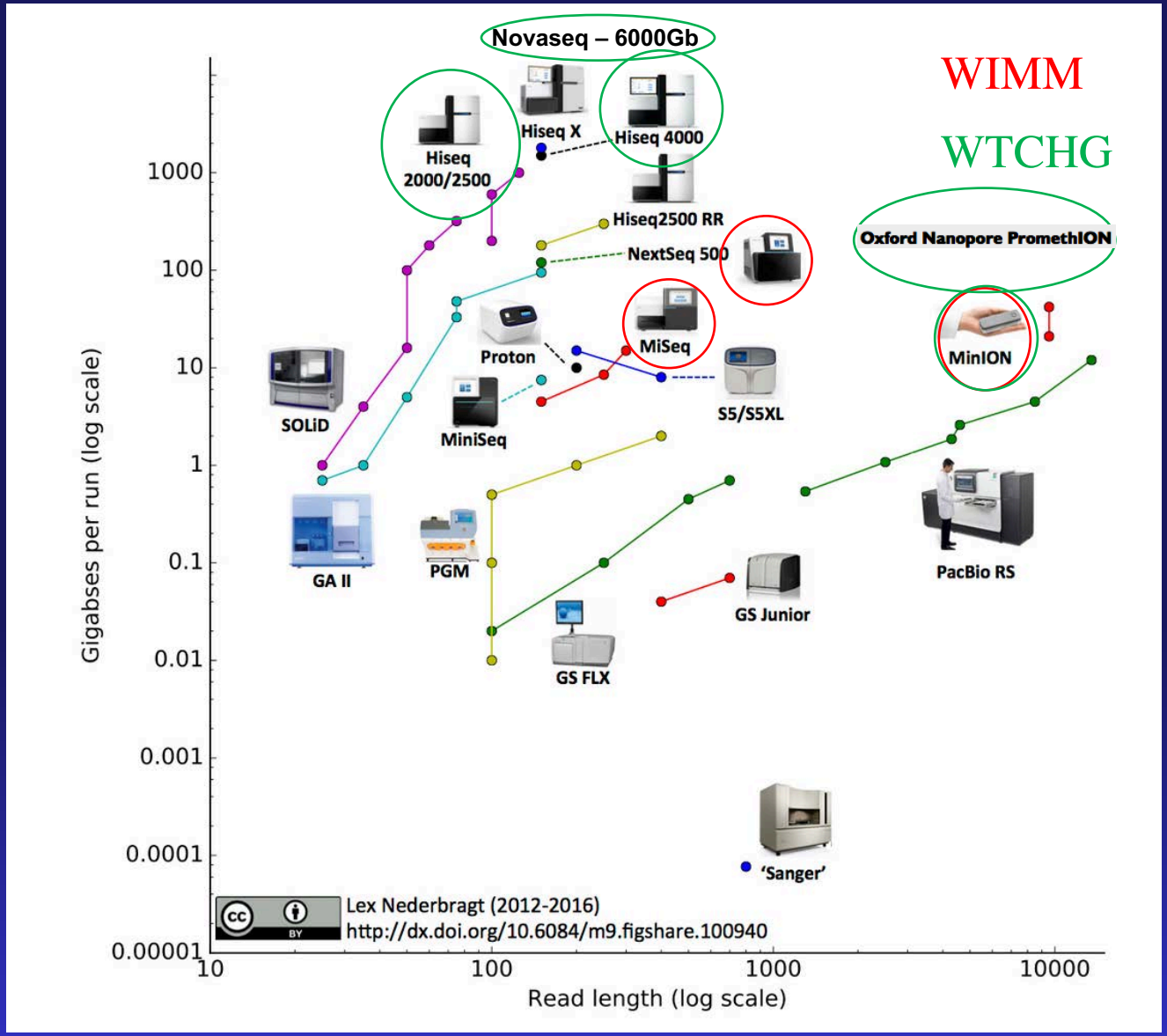


Human genome first draft **(2000)**: $300 million – took 15 years

Human Genome Project: (**1988-2003**): $2.7 billion

Illumina HiSeq X Ten system – 45 genomes in a day for $1000 each [1800 Gb/run (3 days) 18,000 genomes/year]

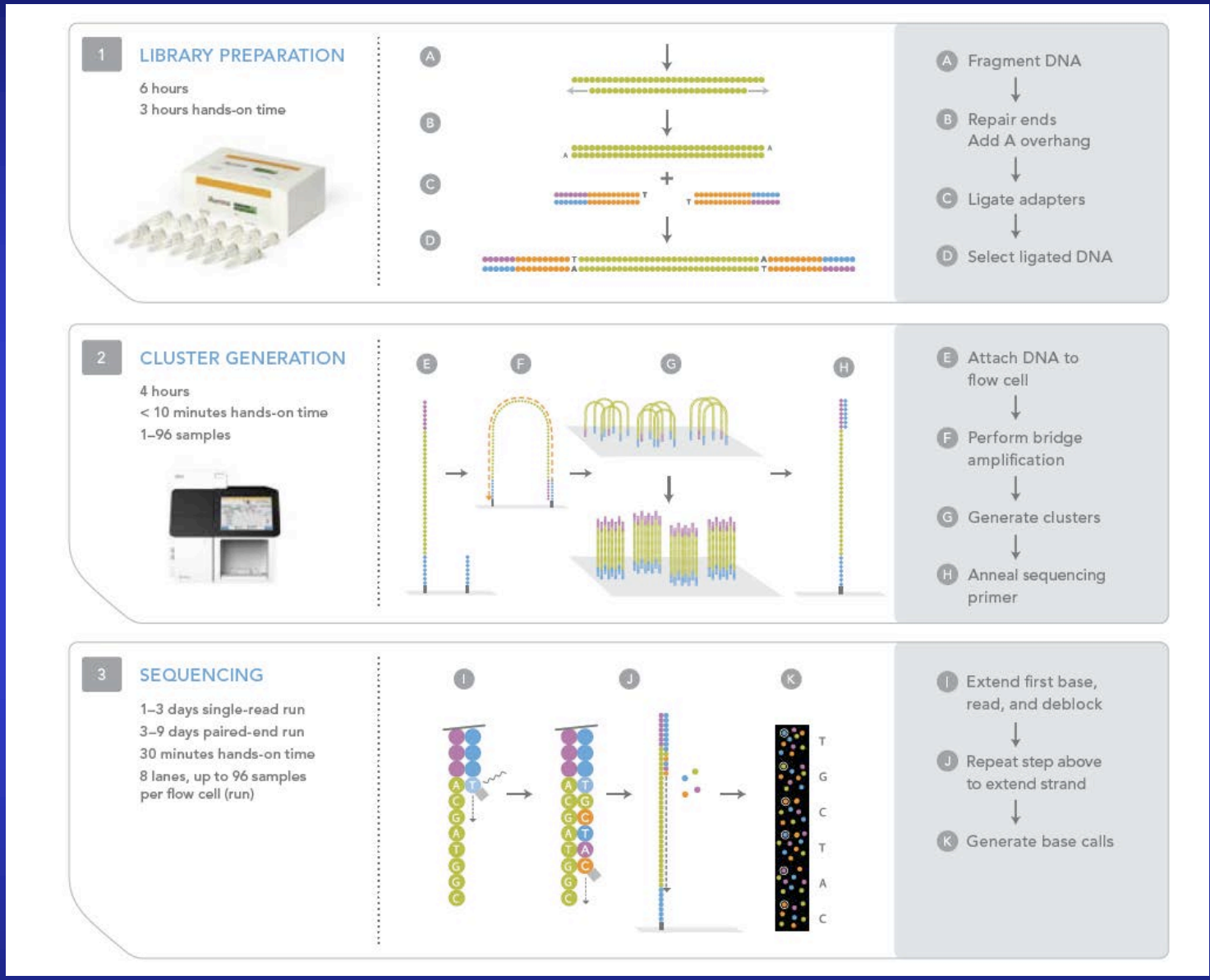
DNA sequencing output



2nd generation DNA-sequencing - Illumina

Relies on PCR

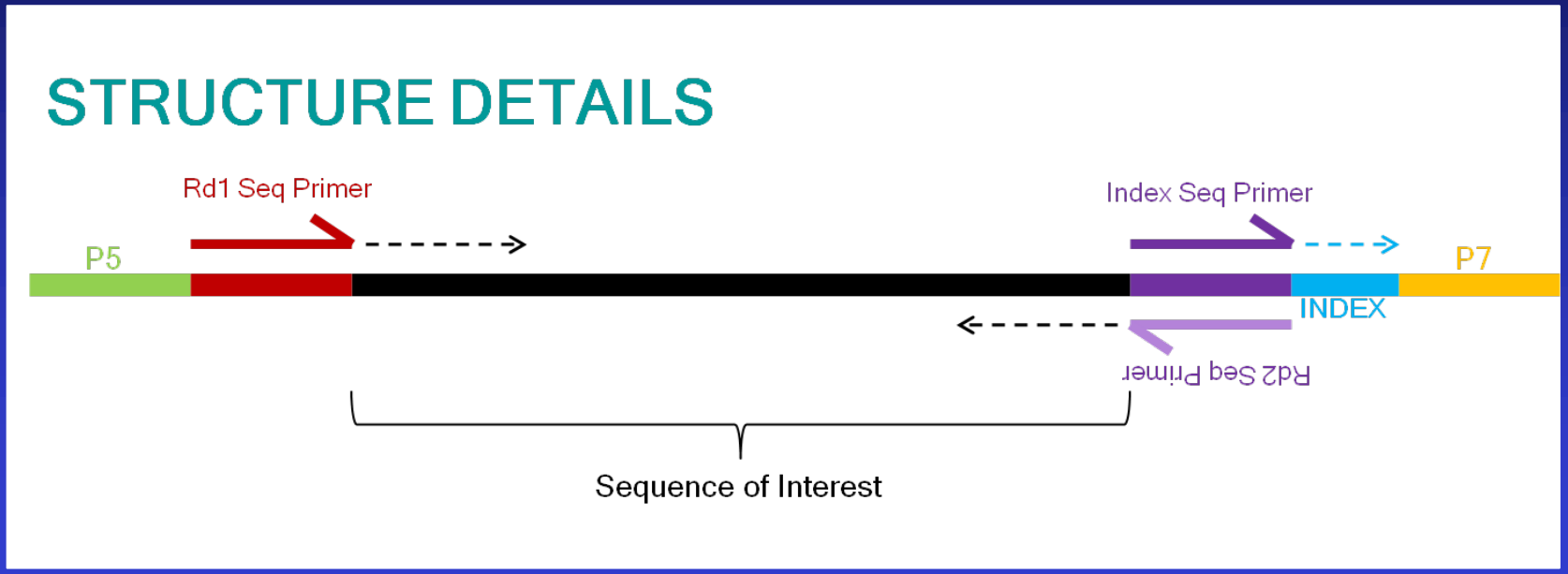
Illumina
'wash-and-scan'
- 1, sequential flooding in of labeled nucleotides
- 2, incorporation of nucleotides into the DNA strands
- 3, stopping the incorporation reaction
- 4, washing out the excess reagent
- 5, scanning to identify the incorporated bases
- 6, prepare the DNA templates for the next 'wash-and-scan' cycle

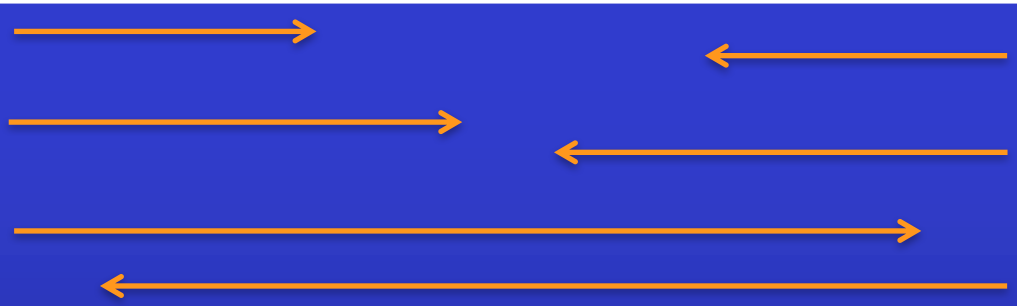


Massively parallel generation of short fragment reads

Mardis ER 2017 *Nat Protocol* **12**:213

NGS reads





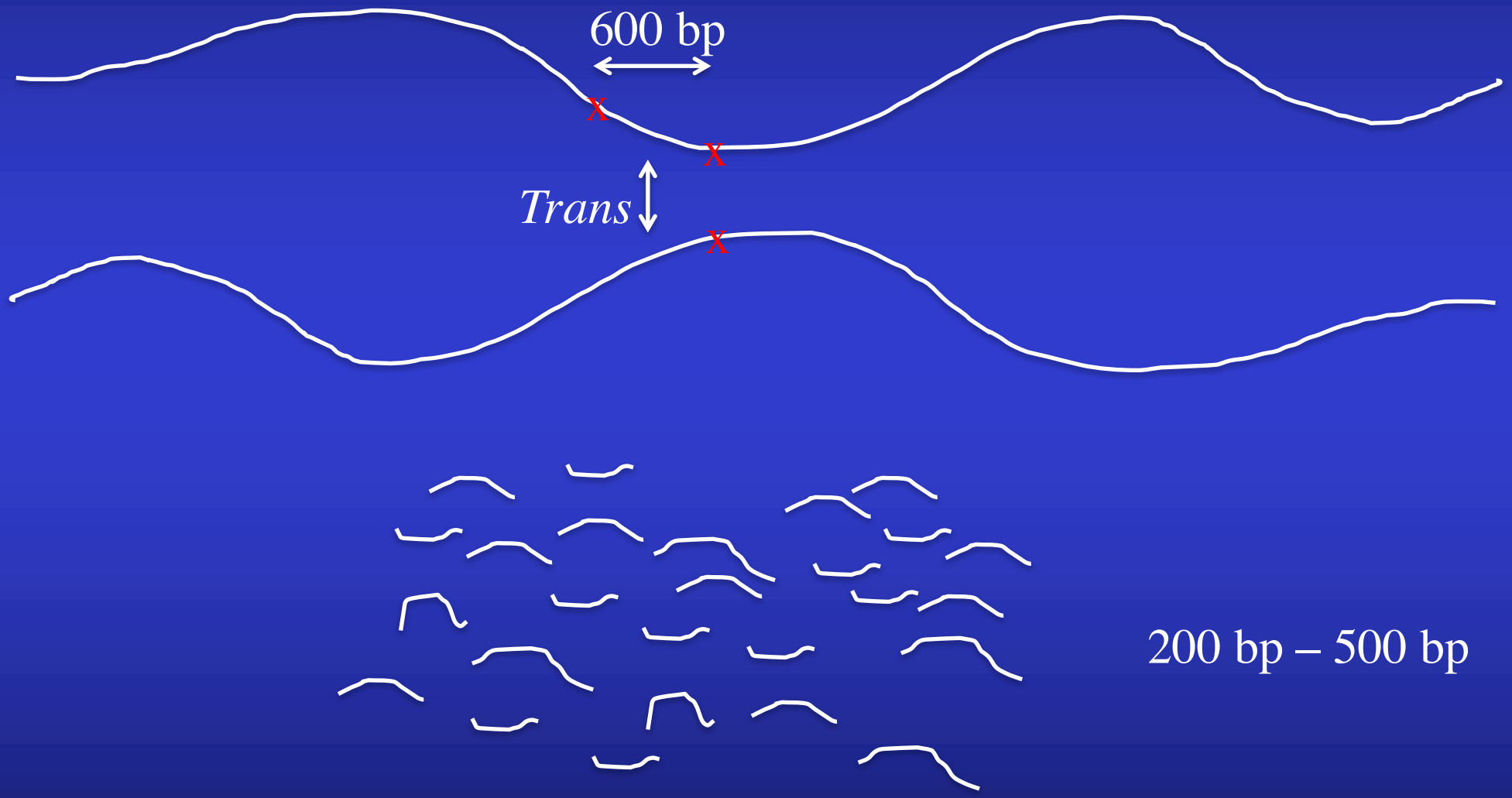
100 cycles, 50 bp PE

150 cycles, 75bp PE

300 cycles, 150bp PE

140 - 240 Million sequences from each end HiSeq2500
~ 400 Million - Nextseq
~ 20 Million - MiSeq
10 Billion - NovaSeq

NGS Sequencing Introduction



Methods to find disease genes

1. Linkage/association
2. Chromosome abnormality
3. Animal model/ Candidate gene
4. Exome/ whole genome sequencing
5. Complex disease - GWAS

A step change in disease gene identification

Table 1. OMIM Phenotypes for which the Molecular Basis Is Known, 2007 and 2013

Inheritance Pattern	January 2007	July 2013
Autosomal	1,851	3,525
X Linked	169	277
Y Linked	2	4
Mitochondrial	26	28
Total	2,048	3,834

Koboldt et al., Cell **155**, 2013

Table 2. Disease-Causing Genes Identified by Exome Sequencing Studies, 2009–2010

Gene	Disorder	Individuals	Citation
DHODH	Miller syndrome	four affected from three kindreds	(Ng et al., 2010b)
FLVCR2	Fowler syndrome	two unrelated	(Lalonde et al., 2010)
GPSM2	Nonsyndromic hearing loss	one proband	(Walsh et al., 2010)
MLL2	Kabuki syndrome	ten unrelated	(Ng et al., 2010a)
WDR62	Severe brain malformations	one proband	(Bilgüvar et al., 2010)
PIGV	Hyperphosphatasia mental retardation	three siblings	(Krawitz et al., 2010)
WDR35	Sensenbrenner syndrome	two unrelated	(Gilissen et al., 2010)
STIM1	Kaposi sarcoma	one patient	(Byun et al., 2010)
ANGPTL3	Familial combined hypolipidemia	two family members	(Musunuru et al., 2010)
ACAD9	Complex I deficiency	one patient	(Haack et al., 2010)
SETBP1	Schinzel-Giedion syndrome	four unrelated	(Hoischen et al., 2010)
TGM6	Spinocerebellar ataxia	four family members	(Wang et al., 2010)
FADD	Autoimmune lymphoproliferative syndrome	one proband	(Bolze et al., 2010)
VCP	Familial ALS	two family members	(Johnson et al., 2010)

Massively parallel sequencing – first disease gene

Exome sequencing identifies the cause of a mendelian disorder

Sarah B Ng[1,10], Kati J Buckingham[2,10], Choli Lee[1], Abigail W Bigham[2], Holly K Tabor[2,3], Karin M Dent[4], Chad D Huff[5], Paul T Shannon[6], Ethylin Wang Jabs[7,8], Deborah A Nickerson[1], Jay Shendure[1] & Michael J Bamshad[1,2,9]

We demonstrate the first successful application of exome sequencing to discover the gene for a rare mendelian disorder of unknown cause, Miller syndrome (MIM%263750). For four affected individuals in three independent kindreds, we captured and sequenced coding regions to a mean coverage of 40× and sufficient depth to call variants at ~97% of each targeted exome. Filtering against public SNP databases and eight HapMap exomes for genes with two previously unknown variants in each of the four individuals identified a single candidate gene, *DHODH*, which encodes a key enzyme in the pyrimidine *de novo* biosynthesis pathway. Sanger sequencing confirmed the presence of *DHODH* mutations in three additional families with Miller syndrome. Exome sequencing of a small number of unrelated affected individuals is a powerful, efficient strategy for identifying the genes underlying rare mendelian disorders and will likely transform the genetic analysis of monogenic traits.

VOLUME 42 | NUMBER 1 | JANUARY 2010 | NATURE GENETICS

2010

Exome sequencing

Pros:

Most disease variants are coding

Cheaper than whole genome – multiplex

Targeted capture is possible

Bioinformatics pipeline easier – fewer variants to deal with

Cons:

Some exons are omitted from capture library

Non-coding regions not covered

GC-rich regions – no baits / poor capture

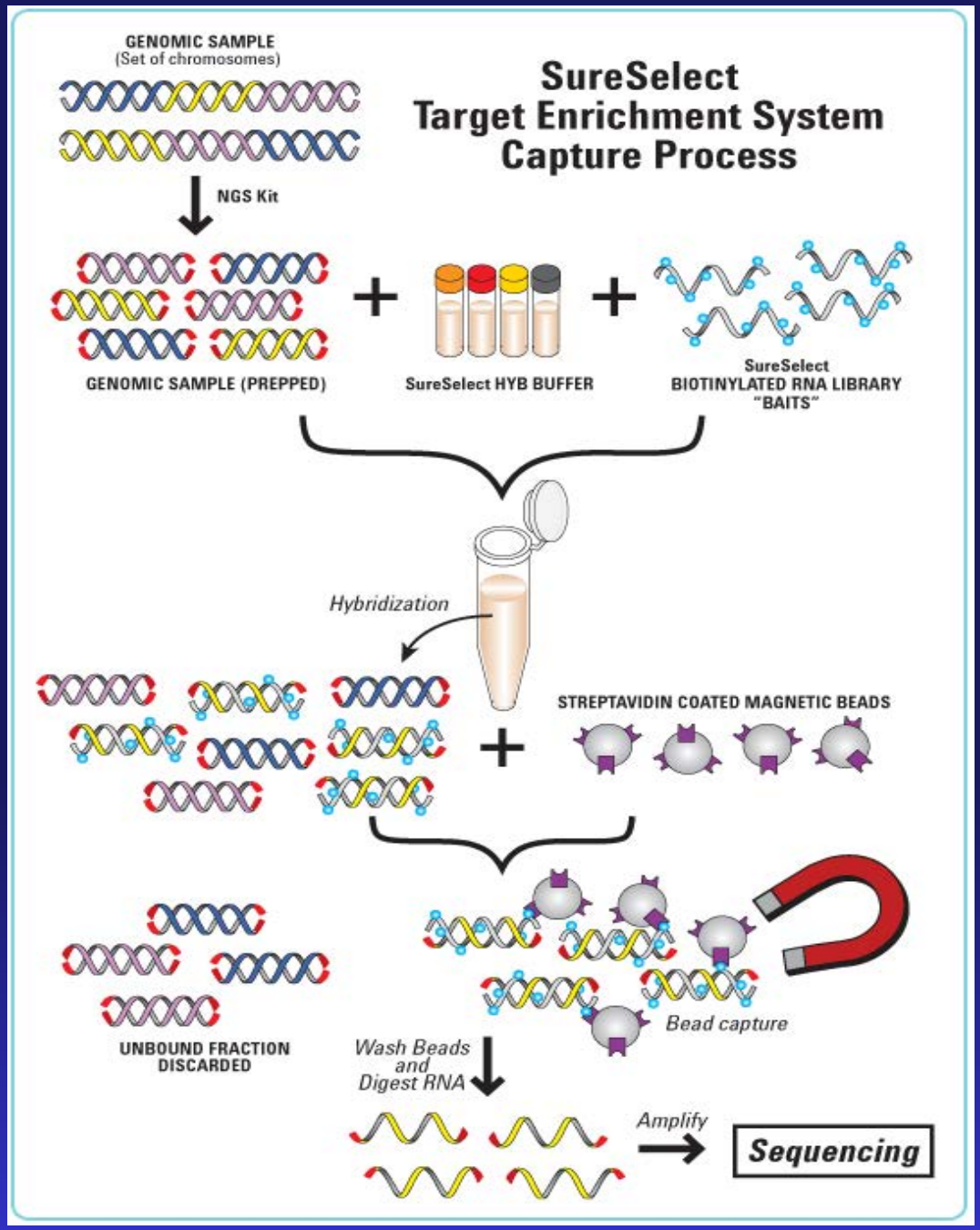
Exome sequencing as a tool for Mendelian disease gene discovery

Michael J. Bamshad[*‡], *Sarah B. Ng*[‡], *Abigail W. Bigham*[*§], *Holly K. Tabor*[*||], *Mary J. Emond*[¶], *Deborah A. Nickerson*[‡] *and Jay Shendure*[‡]

Unlocking Mendelian disease using exome sequencing

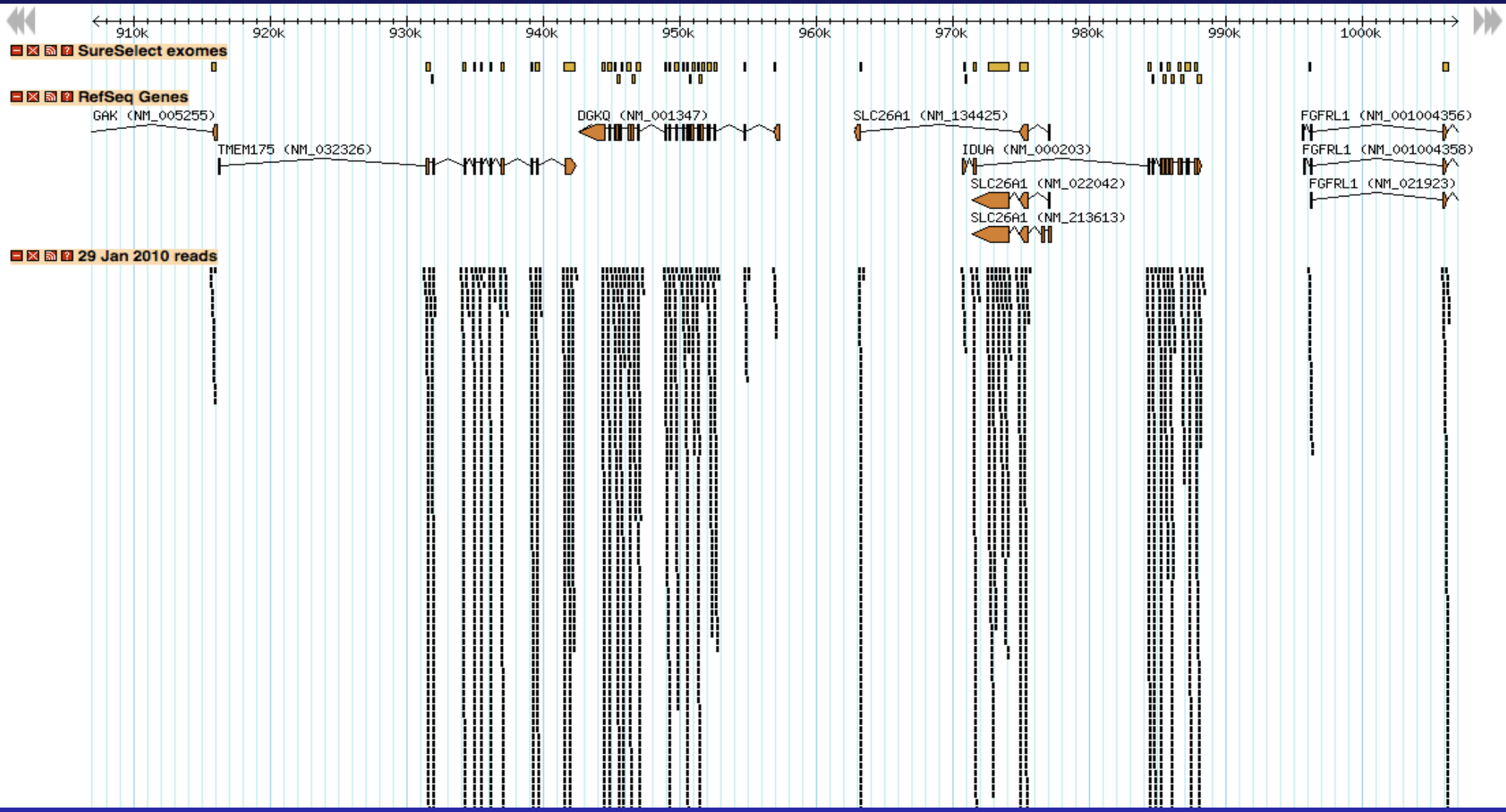
Gilissen et al. Genome Biology 2011, 12:228

Whole exome or targeted sequencing



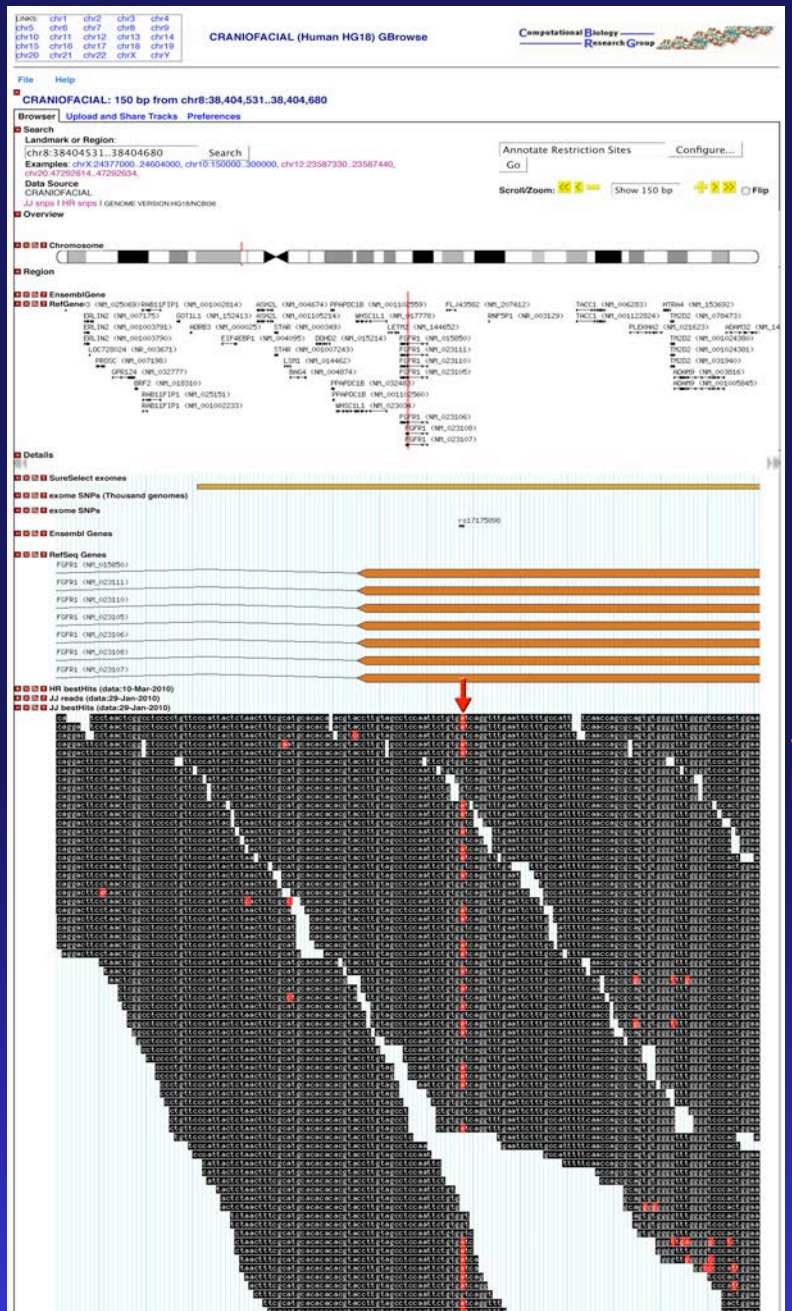
Agilent

Exome seq aligned to genome



Exome seq aligned to genome

WIMM Centre for Computational Biology
Steve Taylor
Simon McGowan



Sequencing reads aligned to the human genome
Black – reference bases
Red – variant bases

Whole genome sequencing

Pros:
Both exonic and non-coding variants are detected
Uniform coverage
CNVs / translocations detected

Cons:
Cost
Vast amount of data - ~ 4 million variants/ individual
Bioinformatic analysis is challenging – data volume/ variant calling
Sequencing errors over repeats/simple sequences

Science. 2010 Apr 30;328(5978):636-9. Epub 2010 Mar 10.

Analysis of genetic inheritance in a family quartet by whole-genome sequencing.

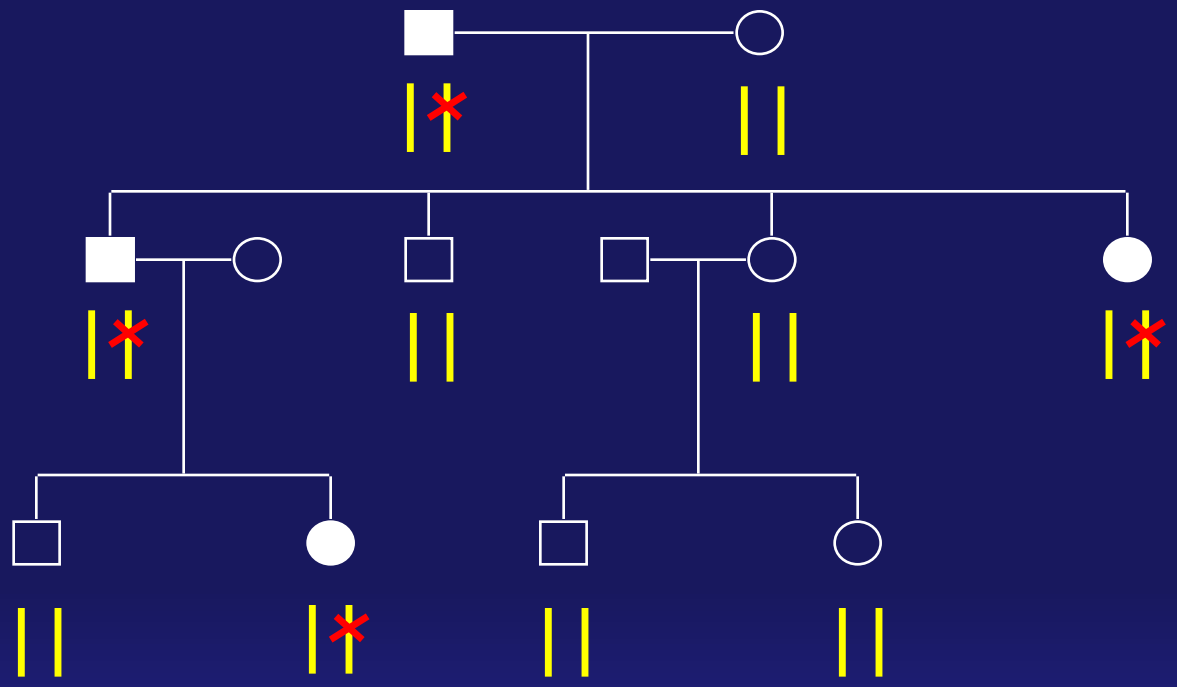
Roach JC, Glusman G, Smit AF, Huff CD, Hubley R, Shannon PT, Rowen L, Pant KP, Goodman N, Bamshad M, Shendure J, Drmanac R, Jorde LB, Hood L, Galas DJ.

Institute for Systems Biology, Seattle, WA 98103, USA.

Disease gene identification

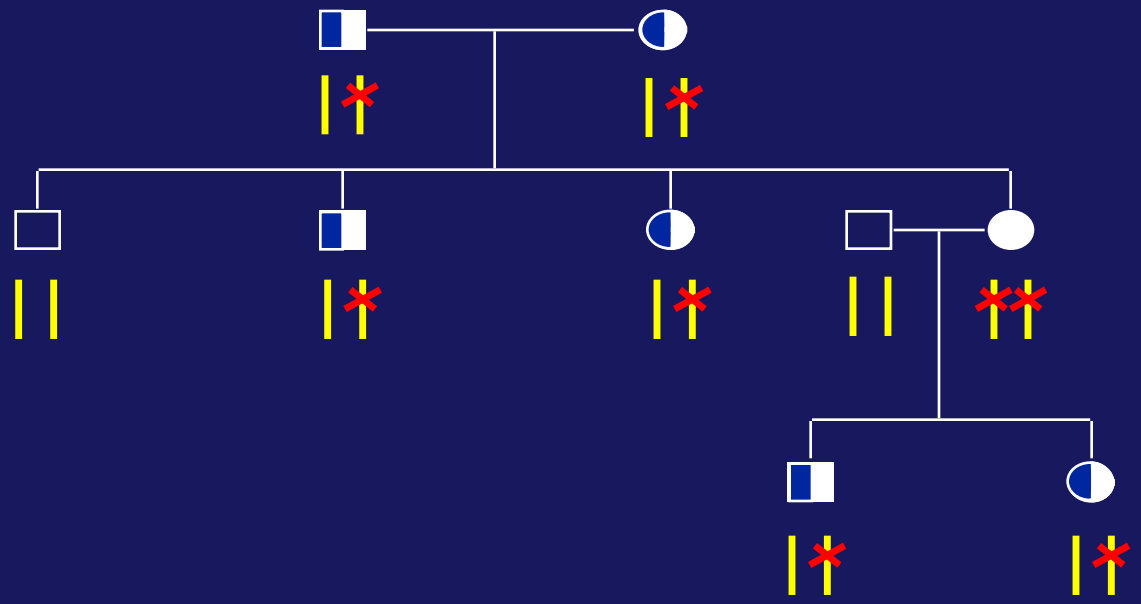
Strategy - Informed by inheritance pattern

Autosomal dominant inheritance



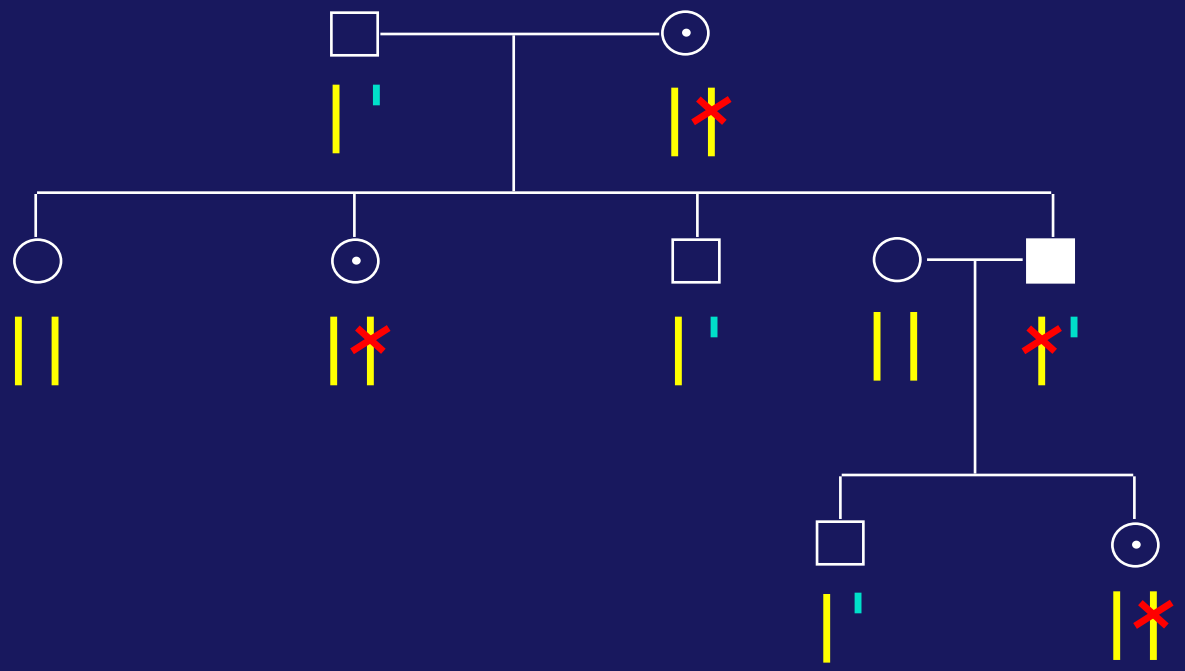
- Disease manifests when only one gene of pair carries mutation
- Sexes usually equally affected
- Risk for siblings and offspring of affected individual = 50%

Autosomal recessive inheritance



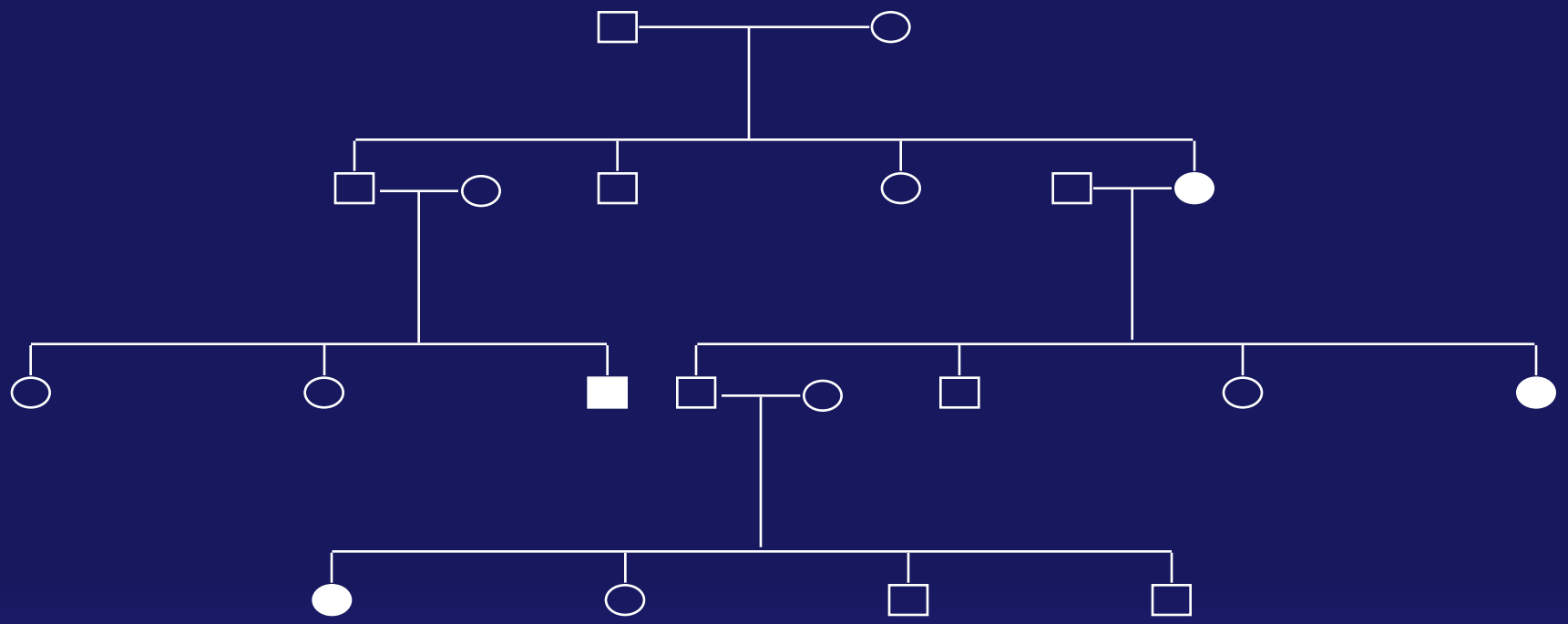
- Disease manifests when both genes of a pair carry a mutation
- Sexes usually equally affected
- More common when parents are consanguineous
- Offspring risk for parents of affected individual = 25%
- Risk to offspring of affected individuals low (usually <1%)
- Carrier risk to siblings of affected individual = 2/3 (67%)

X-linked (recessive) inheritance



- Disease manifests in males, females are carriers
- No male to male transmission; all female offspring of affected males are carriers
- Risk to offspring of carrier female for affected male 25%
- Fragile X mental retardation is a special case: about 10% of male offspring of carrier females are ‘premutation’ carriers

Polygenic inheritance

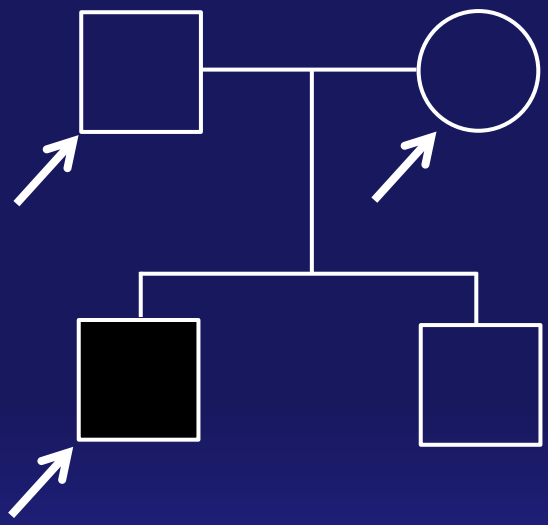


- Clustering of cases within a family, not conforming to any clear pattern of inheritance
- Risks must be based on empiric data

Disease gene identification strategies

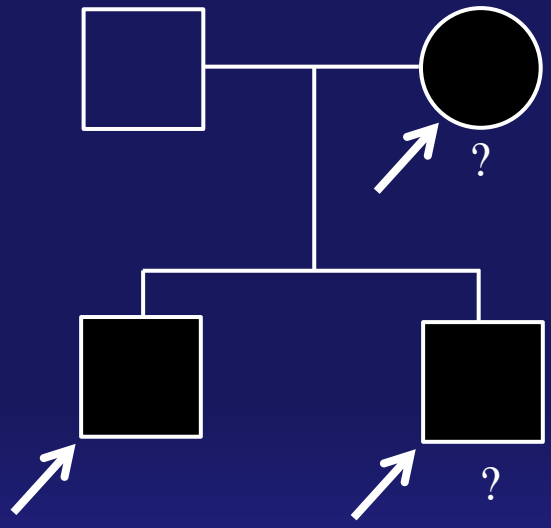
Pedigrees - mode of inheritance/ which individuals to sequence?

Trio



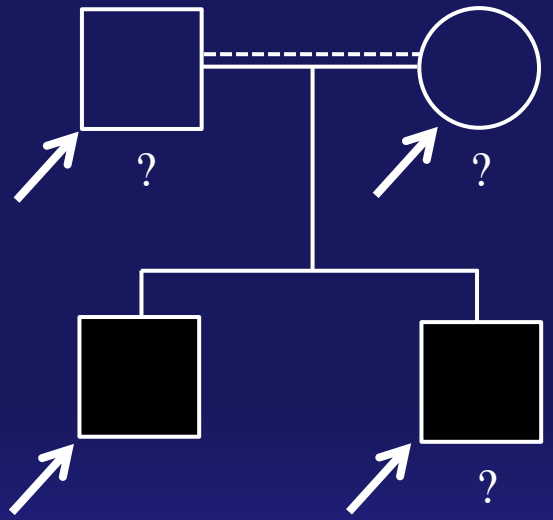
Assume *de novo*
(2-3 coding variants/ generation)
[could be recessive or dominant]

Dominant



Familial
variant inherited
from 1 parent
[underpowered, long list of variants]

Recessive



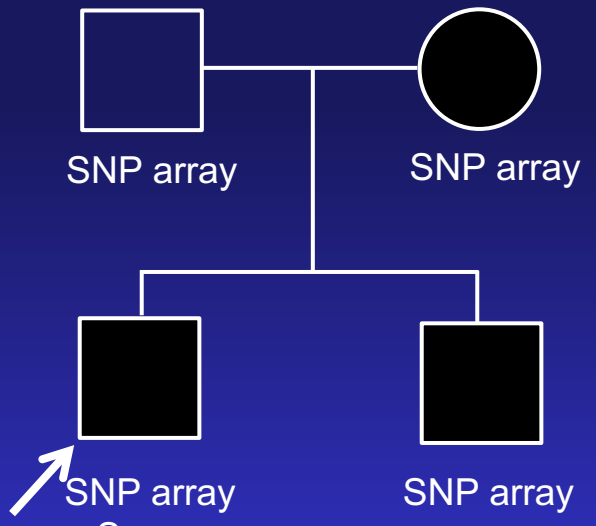
Assume recessive
(homozygous or compound heterozygous)
[could be dominant]

Disease gene identification strategies

Next generation sequencing analysis generates large numbers of variants

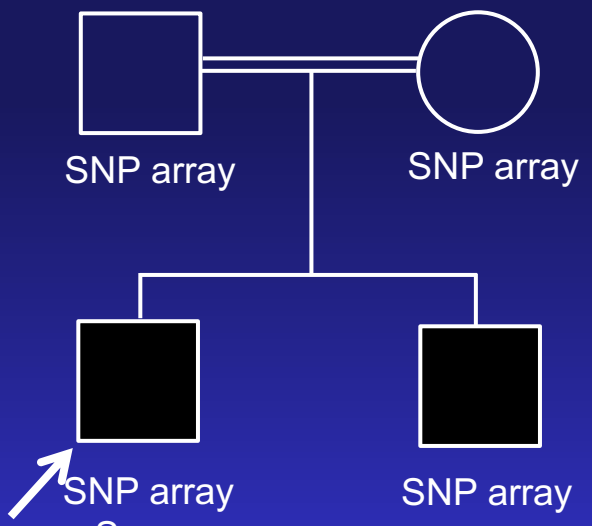
1) Segregation – SNP array - Dominant
2) Segregation – SNP array - Recessive
3) Cohort studies – shared genes
4) Trio – *de novo*/recessive

1) Dominant



Genotype (SNP array or sequencing) all family members to (1) identify genomic regions that segregate with disease or (2) to identify regions of homozygosity

2) Recessive

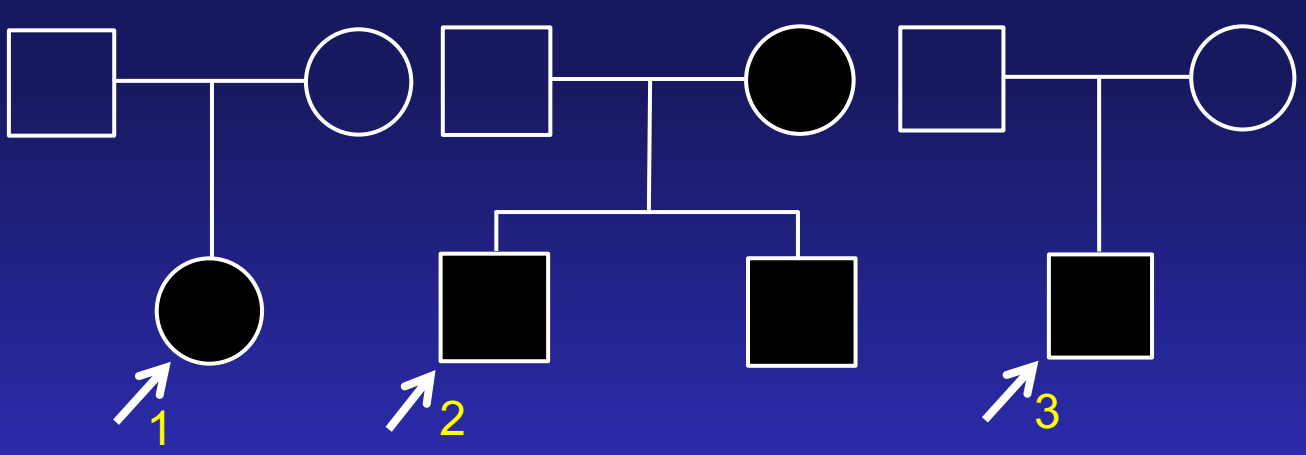


Disease gene identification strategies

Next generation sequencing analysis generates large numbers of variants

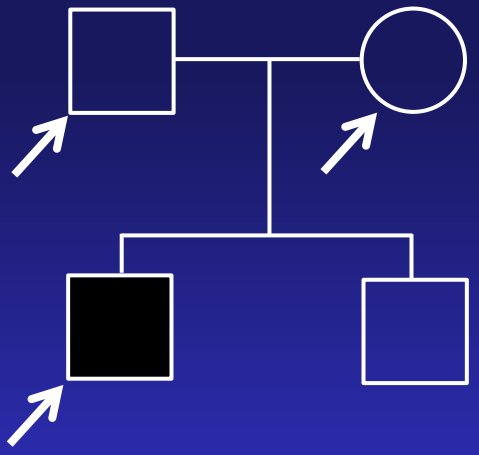
1) Segregation – SNP array - Dominant
2) Segregation – SNP array - Recessive
3) Cohort studies – shared genes
4) Trio – *de novo*/recessive

3) Cohort studies



Identify variants in shared genes

4) Trio



Identify new variants

Identifying disease causing variants - which variants to follow up?

The genome contains:

100 loss of function mutations with 20 genes having both copies inactivated (MacArthur et al., 2012)

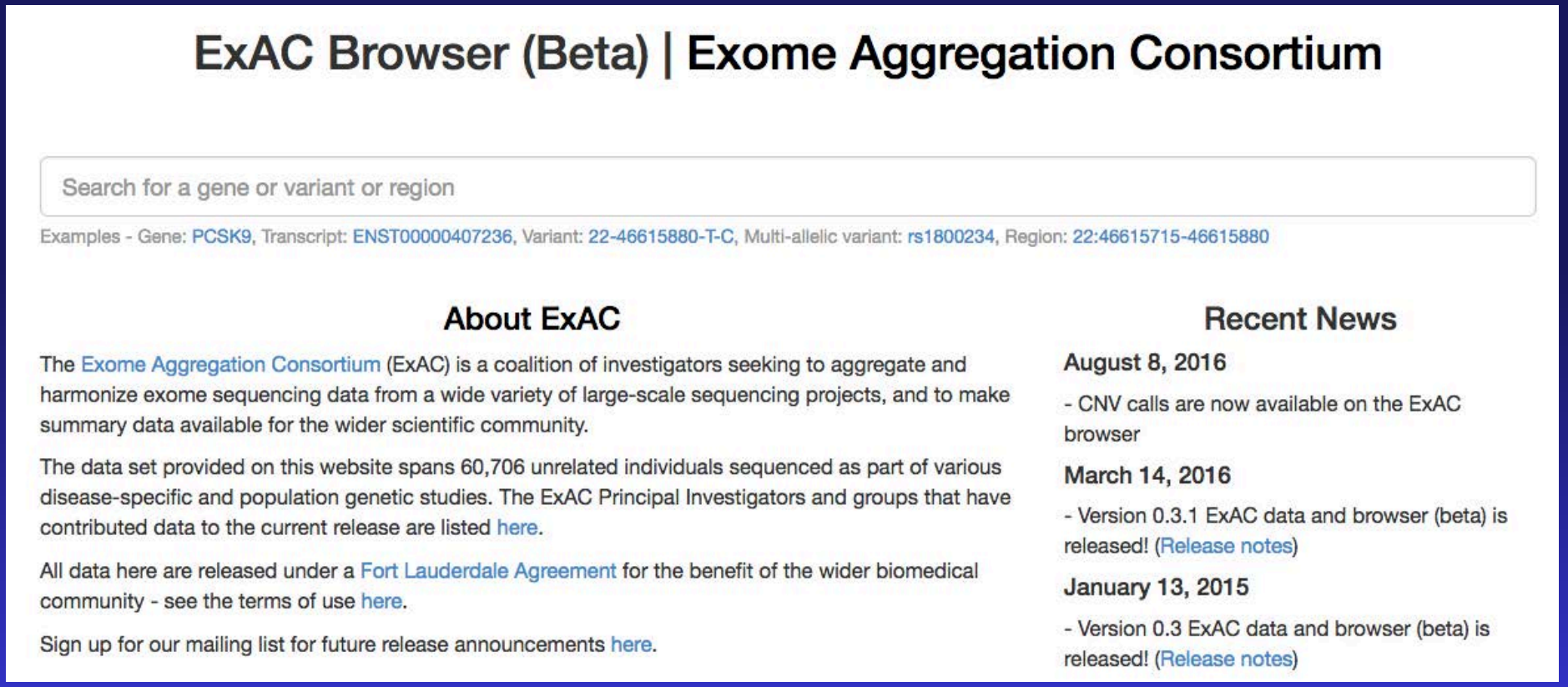
218-515 damaging missense mutations with 40-85 present in the homozygous state (Xue et al., 2012)

40-100 have previously been annotated as disease causing

Which variants to follow up?

Check that variants haven't been seen before:

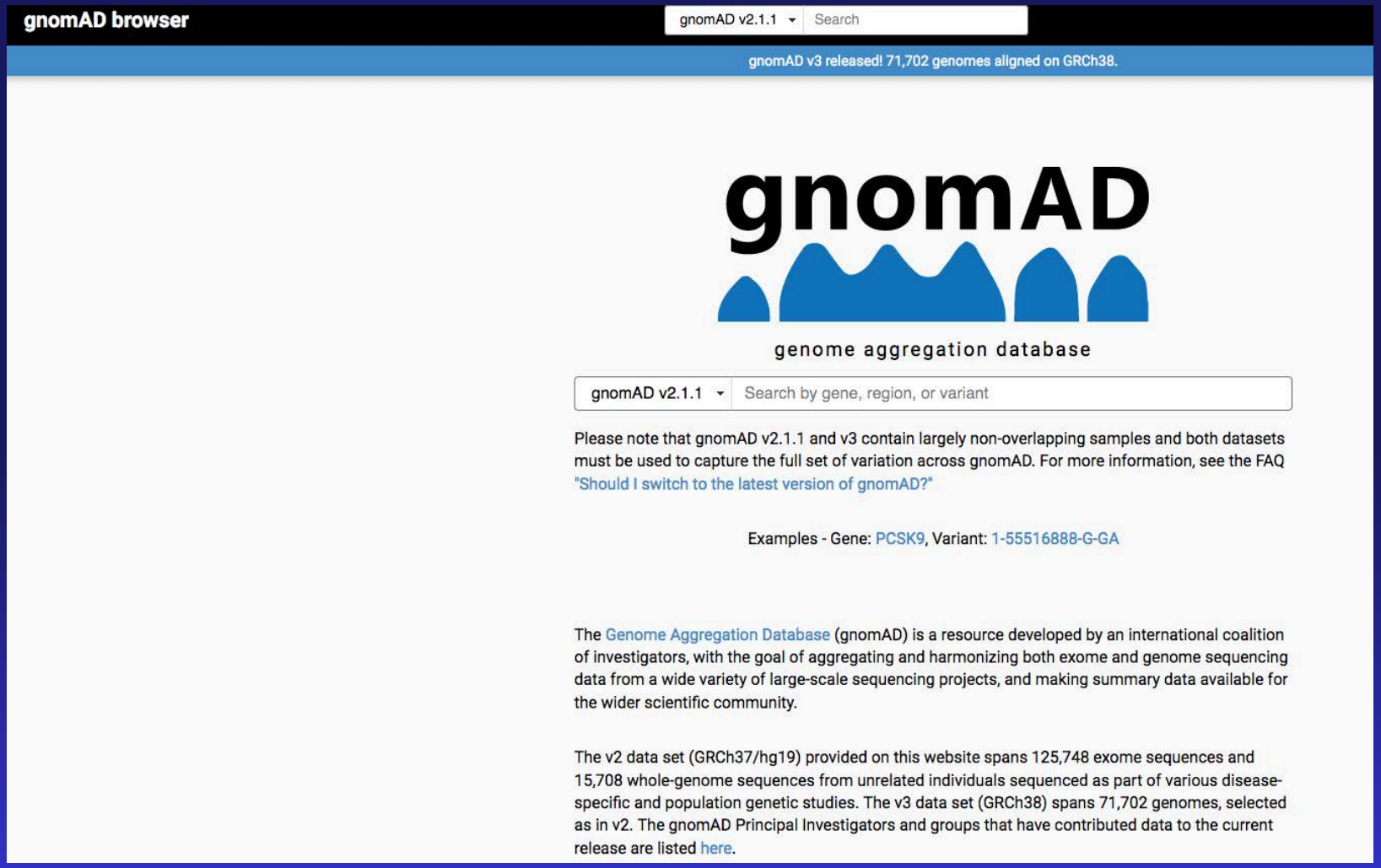
- Exome Variant Server (http://evs.gs.washington.edu/EVS/)
- Exome Aggregation Consortium (http://exac.broadinstitute.org/)
- Local solved exomes/genomes
- Filter against dbSNP – BUT contains clinically relevant variants



Which variants to follow up?

Check that variants haven't been seen before:

- gnomAD (https://gnomad.broadinstitute.org/)



Which variants to follow up?

Is the gene under constraint?
pLI metric (ExAC)
pLI >= 0.9 - extremely LoF intolerant

Constraint from ExAC	Expected no. variants	Observed no. variants	Constraint Metric
Synonymous	173.9	101	z = 3.43
Missense	270.3	91	z = 5.33
LoF	9.8	1	pLI = 0.83
CNV	3.2	0	z = 0.73

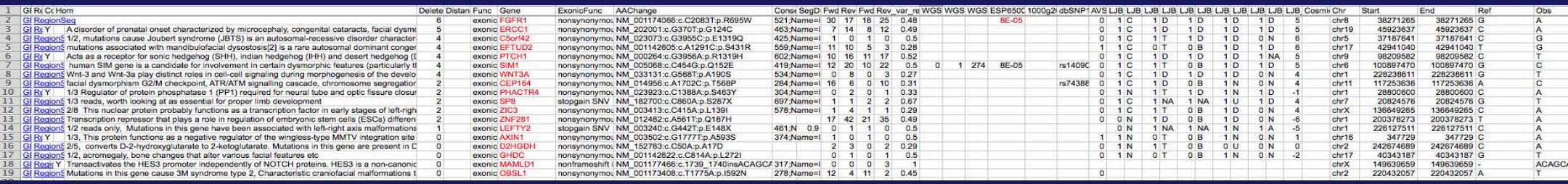
Samocha et al *Nature Genetics* **46**, 944–950 (2014);
Samocha et al bioRxiv 148353 (2017)

Prioritise coding variants and focus on damaging variants?

- nonsense/frameshift/missense - (assess missense with Polyphen2/SIFT/other algorithms - ANNOVAR output – CBRG pipeline)
- Conservation – CADD scores (https://cadd.gs.washington.edu/)

Does the affected gene fit with the disease?
Within an interesting gene/pathway?
Expression patterns
Animal models

Candidate variant list – what next?



	GI	Re	Cc	Hom	Delete	Distan	Func	Gene	ExonicFunc	AAChange	Conse	SegD	Fwd	Rev	Fwd	Rev_var_re	WGS	WGS	WGS	ESP6500	1000g2(	dbSNP1	AVS	LJB	LJB	LJB	LJB	LJB	LJB	LJB	LJB	LJB	Cosmi	Chr	Start	End	Ref	Obs
2	GI	RegionSeg			6		exonic	FGFR1	nonsynonymou	NM_001174066:c.C2083T:p.R695W	521	Name=I	30	17	18	25	0.48				8E-05			0	1	C	1	D	1	D	1	D	1	D	5		chr8	38271265	38271265	G	A
3	GI	Re	Y	A disorder of prenatal onset characterized by microcephaly, congenital cataracts, facial dysme	5		exonic	ERCC1	nonsynonymou	NM_202001:c.G370T:p.G124C	463	Name=I	7	14	8	12	0.49							0	1	C	1	D	1	D	1	D	1	D	5		chr19	45923637	45923637	C	A
4	GI	RegionS		1/2, mutations cause Joubert syndrome (JBTS) is an autosomal-recessive disorder character	4		exonic	C5orf42	nonsynonymou	NM_023073:c.G3955C:p.E1319Q	425	Name=I	1	0	1	0	0.5							0	1	C	1	T	1	D	1	D	0	N	6		chr5	37187641	37187641	C	G
5	GI	RegionS		mutations associated with mandibulofacial dysostosis[2] is a rare autosomal dominant conger	4		exonic	EFTUD2	nonsynonymou	NM_001142605:c.A1291C:p.S431R	559	Name=I	11	10	5	3	0.28							1	1	C	0	T	0	B	1	D	1	D	6		chr17	42941040	42941040	T	G
6	GI	Re	Y	Acts as a receptor for sonic hedgehog (SHH), indian hedgehog (IHH) and desert hedgehog (D	4		exonic	PTCH1	nonsynonymou	NM_000264:c.G3956A:p.R1319H	602	Name=I	10	16	11	17	0.52							0	1	C	1	D	1	D	1	D	1	NA	5		chr9	98209582	98209582	C	T
7	GI	RegionS		human SIM gene is a candidate for involvement in certain dysmorphic features (particularly th	4		exonic	SIM1	nonsynonymou	NM_005068:c.C454G:p.Q152E	419	Name=I	12	20	10	22	0.5	0	1	274	8E-05		rs14090	0	1	C	1	T	0	B	1	D	1	D	5		chr6	100897470	100897470	G	C
8	GI	RegionS		Wnt-3 and Wnt-3a play distinct roles in cell-cell signaling during morphogenesis of the develo	4		exonic	WNT3A	nonsynonymou	NM_033131:c.G568T:p.A190S	534	Name=I	0	8	0	3	0.27							0	1	C	1	D	1	D	1	D	0	N	4		chr1	228238611	228238611	G	T
9	GI	RegionS		facial dysmorphism G2/M checkpoint, ATR/ATM signalling cascade, chromosome segregation	2		exonic	CEP164	nonsynonymou	NM_014956:c.A1702C:p.T568P	284	Name=I	16	6	0	10	0.31						rs74388	0	1	C	1	D	0	B	1	N	0	N	4		chr11	117253636	117253636	A	C
10	GI	Re	Y	1/3 Regulator of protein phosphatase 1 (PP1) required for neural tube and optic fissure closur	2		exonic	PHACTR4	nonsynonymou	NM_023923:c.C1388A:p.S463Y	304	Name=I	0	2	0	1	0.33							0	1	N	1	T	1	D	1	N	1	D	-1		chr1	28800600	28800600	C	A
11	GI	RegionS		1/3 reads, worth looking at as essential for proper limb development	2		exonic	SP8	stopgain SNV	NM_182700:c.C860A:p.S287X	697	Name=I	1	1	2	2	0.67							0	1	C	1	NA	1	NA	1	U	1	D	4		chr7	20824576	20824576	G	T
12	GI	RegionS		2/8 This nuclear protein probably functions as a transcription factor in early stages of left-righ	2		exonic	ZIC3	nonsynonymou	NM_003413:c.C415A:p.L139I	578	Name=I	1	4	1	1	0.29							0	1	C	1	T	0	B	1	D	0	N	4		chrX	136649265	136649265	C	A
13	GI	RegionS		Transcription repressor that plays a role in regulation of embryonic stem cells (ESCs) differen	2		exonic	ZNF281	nonsynonymou	NM_012482:c.A561T:p.Q187H			17	42	21	35	0.49							0	0	N	1	D	0	B	1	D	0	N	-6		chr1	200378273	200378273	T	A
14	GI	RegionS		1/2 reads only, Mutations in this gene have been associated with left-right axis malformations	1		exonic	LEFTY2	stopgain SNV	NM_003240:c.G442T:p.E148X	461	N 0.9	0	1	1	0	0.5								0	N	1	NA	1	NA	1	N	1	A	-5		chr1	226127511	226127511	C	A
15	GI	Re	Y	1/3, This protein functions as a negative regulator of the wingless-type MMTV integration site	0		exonic	AXIN1	nonsynonymou	NM_003502:c.G1777T:p.A593S	374	Name=I	1	0	1	0	0.5							1	1	N	0	T	0	B	1	N	0	N	1		chr16	347729	347729	C	A
16	GI	RegionS		2/5, converts D-2-hydroxyglutarate to 2-ketoglutarate. Mutations in this gene are present in D	0		exonic	D2HGDH	nonsynonymou	NM_152783:c.C50A:p.A17D			2	3	0	2	0.29							0	1	N	1	T	0	B	0	U	0	N	0		chr2	242674689	242674689	C	A
17	GI	RegionS		1/2, acromegaly, bone changes that alter various facial features etc	0		exonic	GHDC	nonsynonymou	NM_001142622:c.C814A:p.L272I			0	1	0	1	0.5							0	1	N	0	T	0	B	1	N	0	N	-2		chr17	40343187	40343187	G	T
18	GI	Regio	Y	Transactivates the HES3 promoter independently of NOTCH proteins. HES3 is a non-canonic	0		exonic	MAMLD1	nonframeshift i	NM_001177466:c.1739_1740insACAGCA	317	Name=I	0	0	0	3	1																				chrX	149639659	149639659	-	ACAGCA
19	GI	RegionS		Mutations in this gene cause 3M syndrome type 2, Characteristic craniofacial malformations t	0		exonic	OBSL1	nonsynonymou	NM_001173408:c.T1775A:p.I592N	278	Name=I	12	4	11	2	0.45							0														chr2	220432057	220432057	A	T

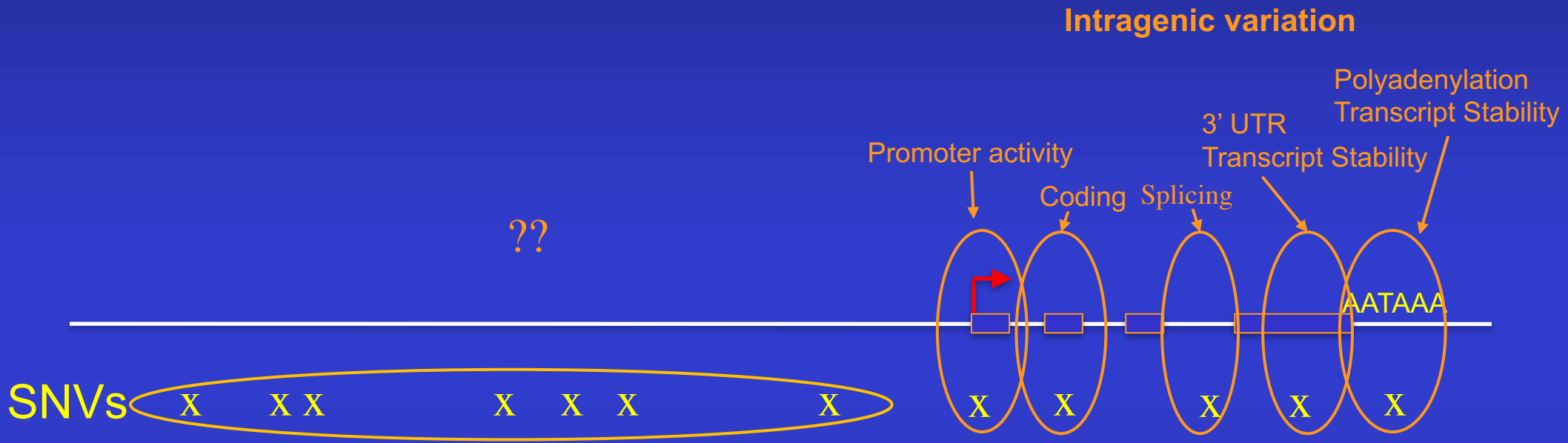
Validate in the family

Proving causation: Find more cases with mutations in the same gene

- Resequence in related cases

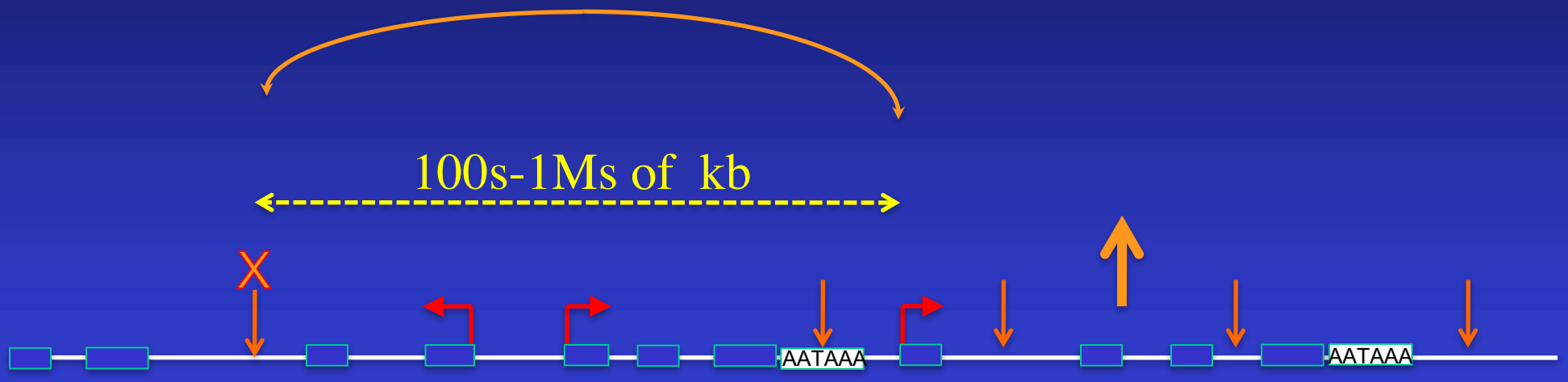
- Functional analysis
- Animal models

Interpretation of sequence variants - gene regulation



~ 90% GWAS hits are not gene associated

Distribution of Enhancers around their Target Genes



To appreciated the impact of sequence variants and mutations in health and disease

- Understand the biology of distal regulation

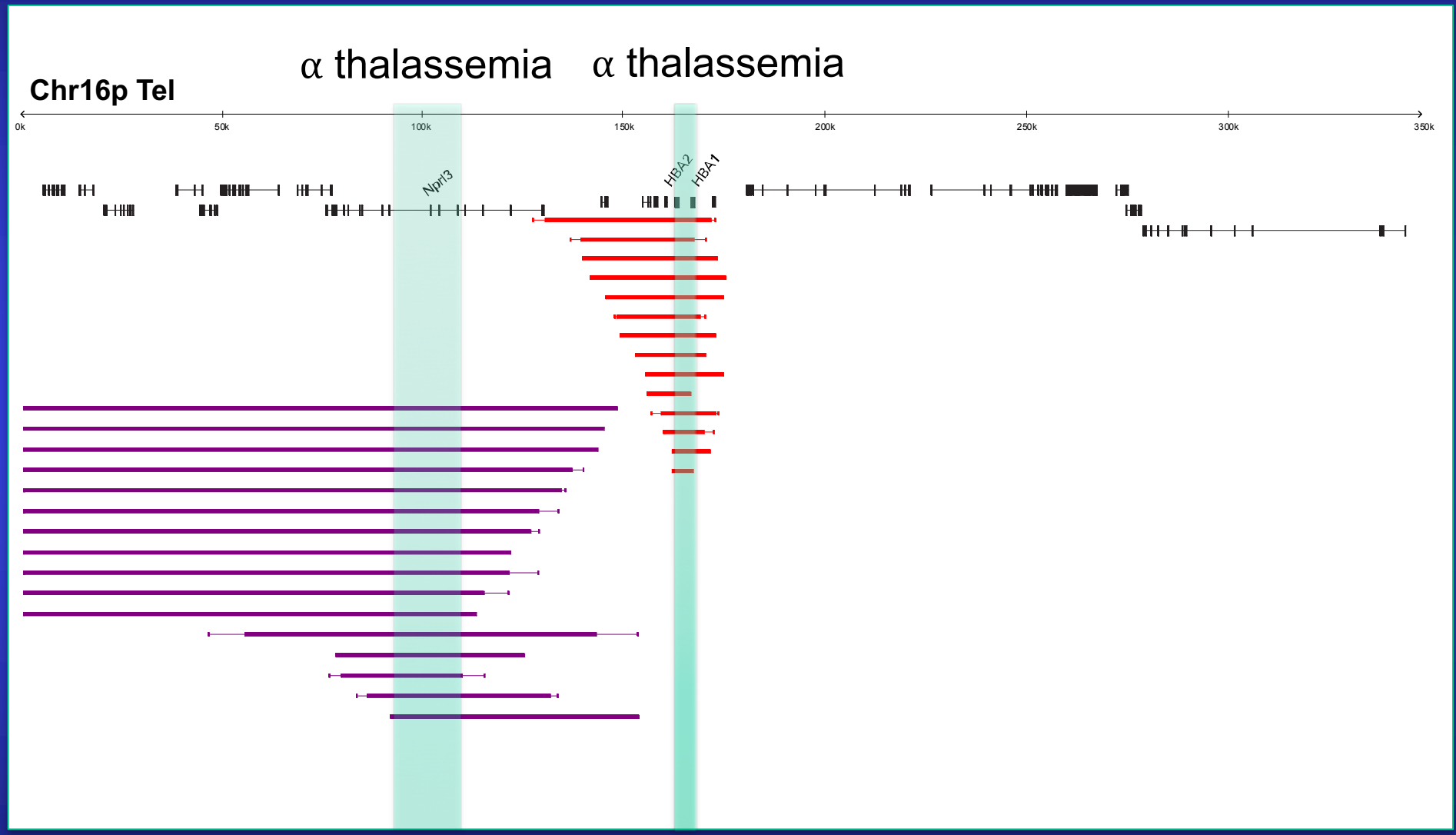
Interrogation of molecular events associated gene regulation

- Genome-wide
- High-resolution
- Dynamics

ChIPseq/RNAseq/Dnase-seq/
/ATACseq/CaptureC

Linking promoters and their regulatory elements represents a major bottleneck

Deletions associated with alpha thalassemia



Jim Hughes

ARTICLE

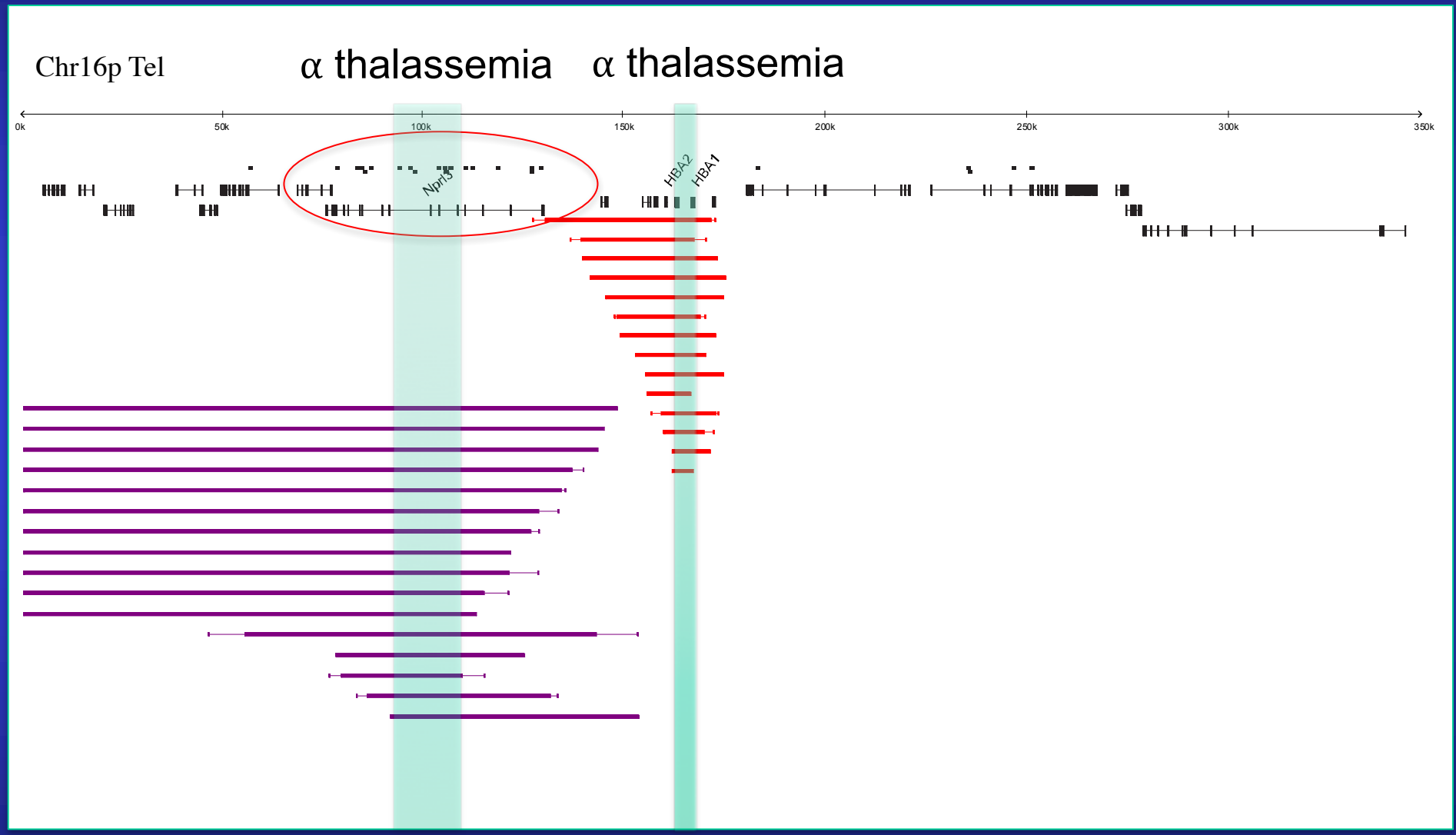
doi:10.1038/nature11677

Seventy-five genetic loci influencing the human red blood cell

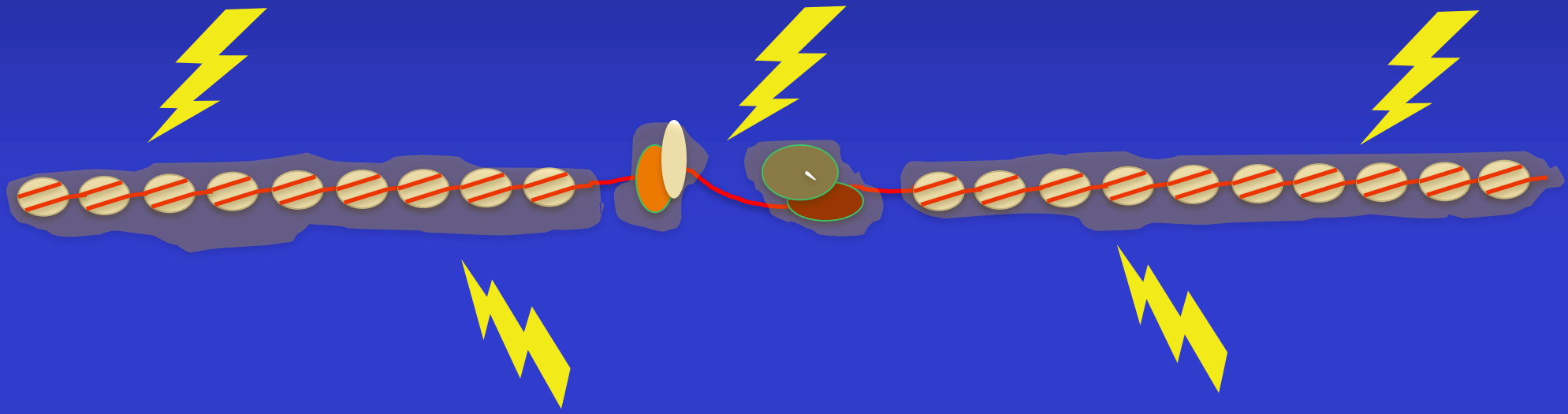
A list of authors and their affiliations appears at the end of the paper

Anaemia is a chief determinant of global ill health, contributing to cognitive impairment, growth retardation and impaired physical capacity. To understand further the genetic factors influencing red blood cells, we carried out a genome-wide association study of haemoglobin concentration and related parameters in up to 135,367 individuals. Here we identify 75 independent genetic loci associated with one or more red blood cell phenotypes at $P < 10^{-8}$, which together explain 4–9% of the phenotypic variance per trait. Using expression quantitative trait loci and bioinformatic strategies, we identify 121 candidate genes enriched in functions relevant to red blood cell biology. The candidate genes are expressed preferentially in red blood cell precursors, and 43 have haematopoietic phenotypes in *Mus musculus* or *Drosophila melanogaster*. Through open-chromatin and coding-variant analyses we identify potential causal genetic variants at 41 loci. Our findings provide extensive new insights into genetic mechanisms and biological pathways controlling red blood cell formation and function.

Non-coding variants associated with hemoglobinization

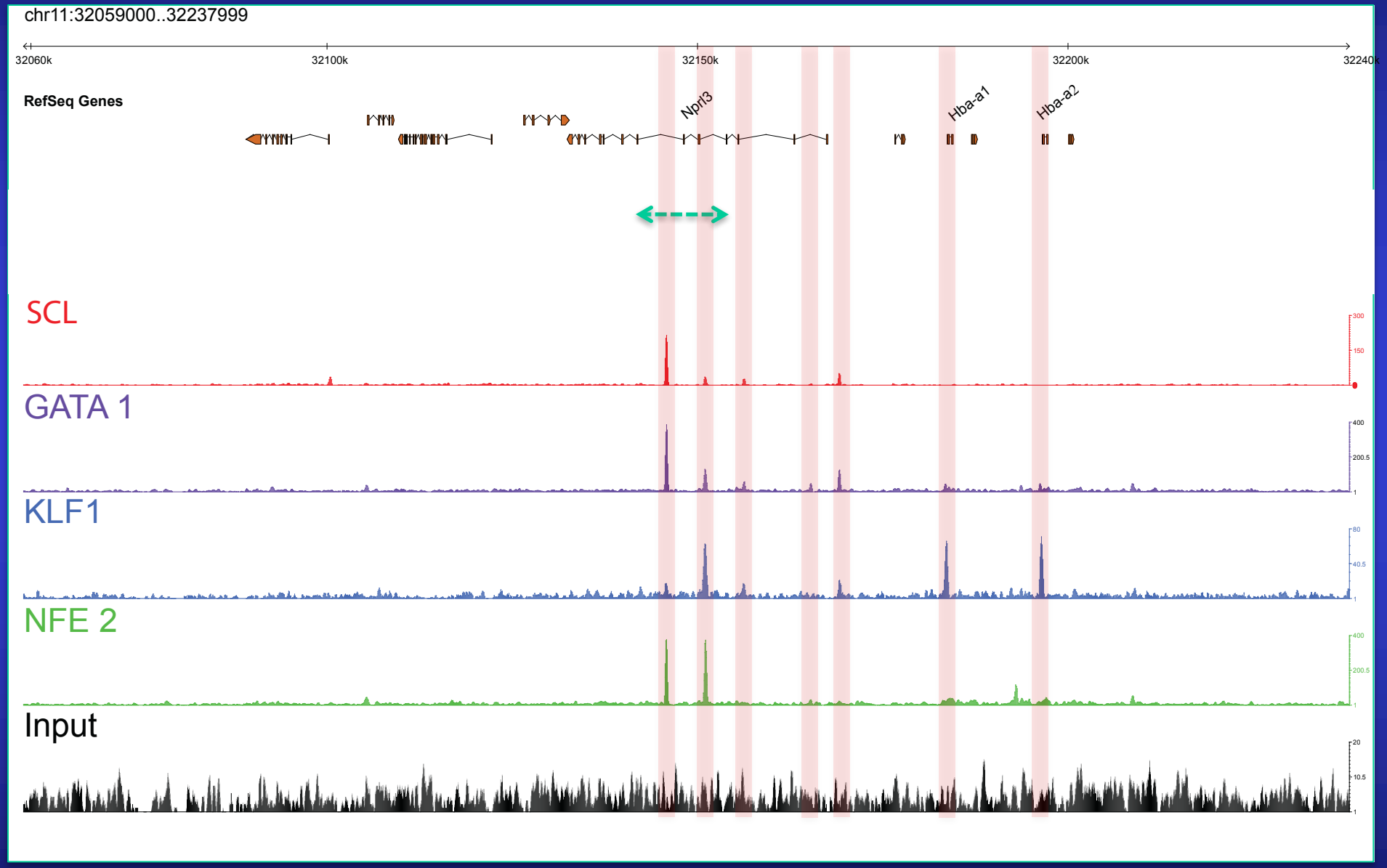


Chromatin Immunoprecipitation (ChIP)

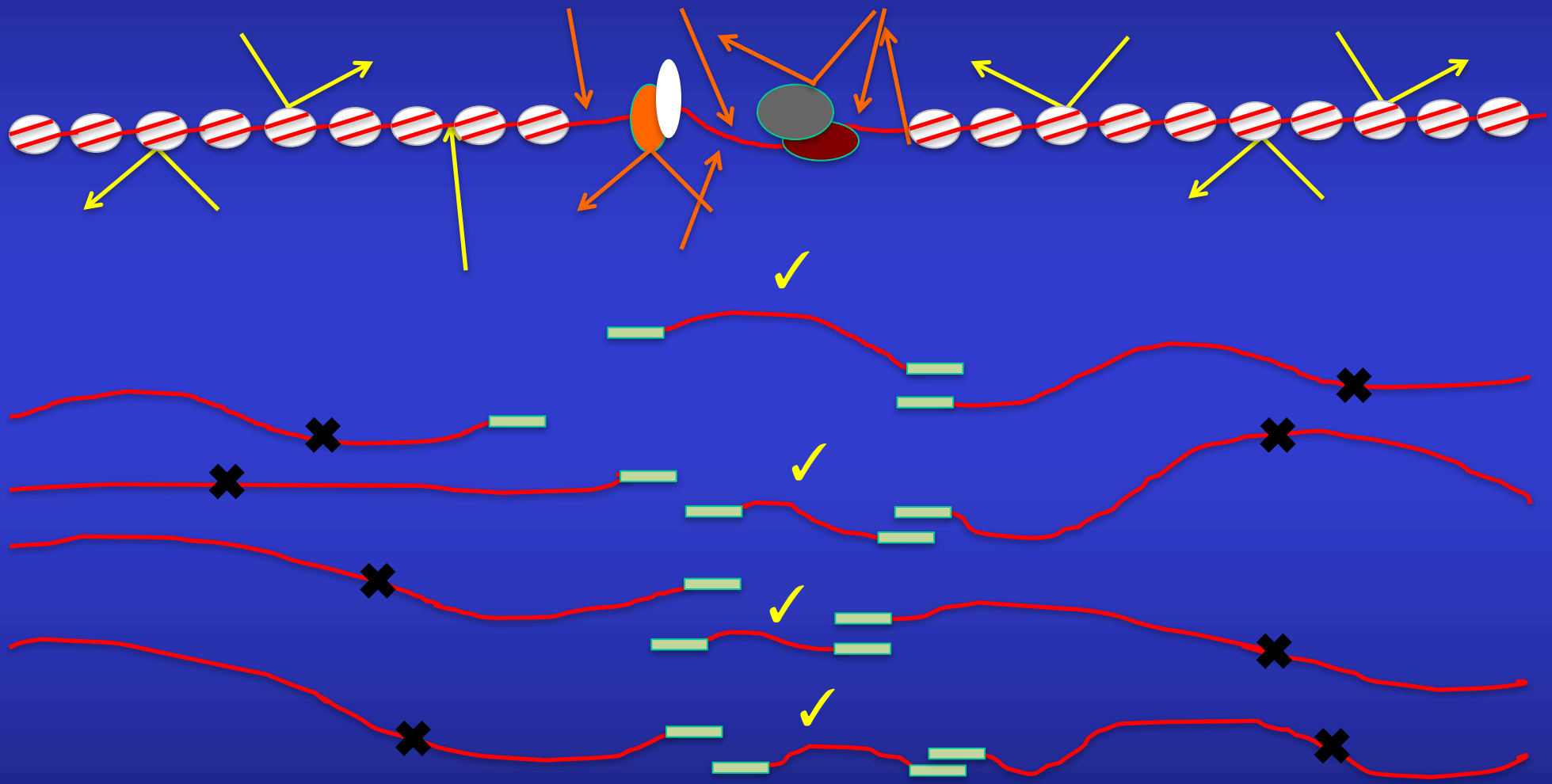


Chromatin Immunoprecipitation (ChIP)

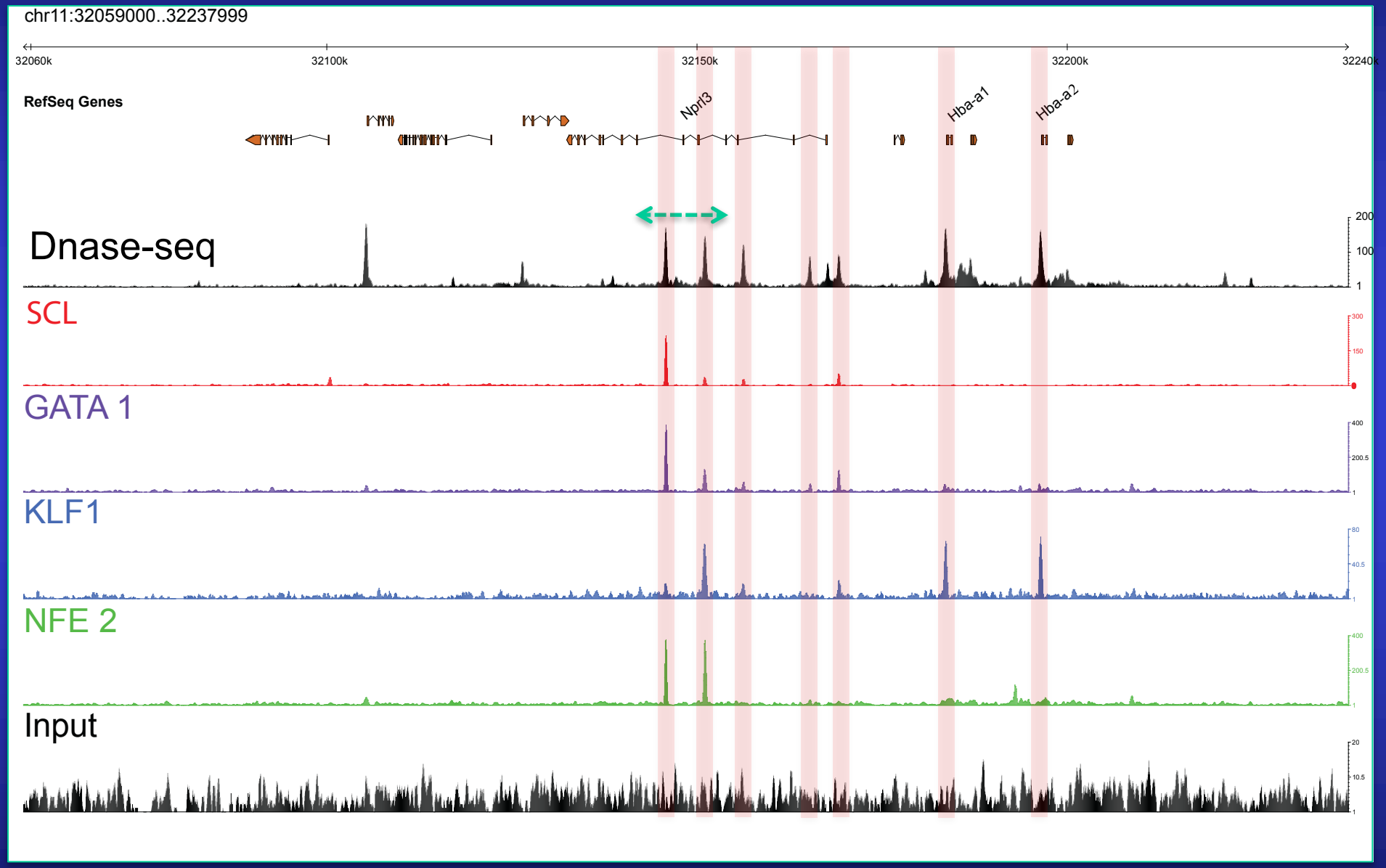
Erythroid Transcription Factors at the α Globin Locus



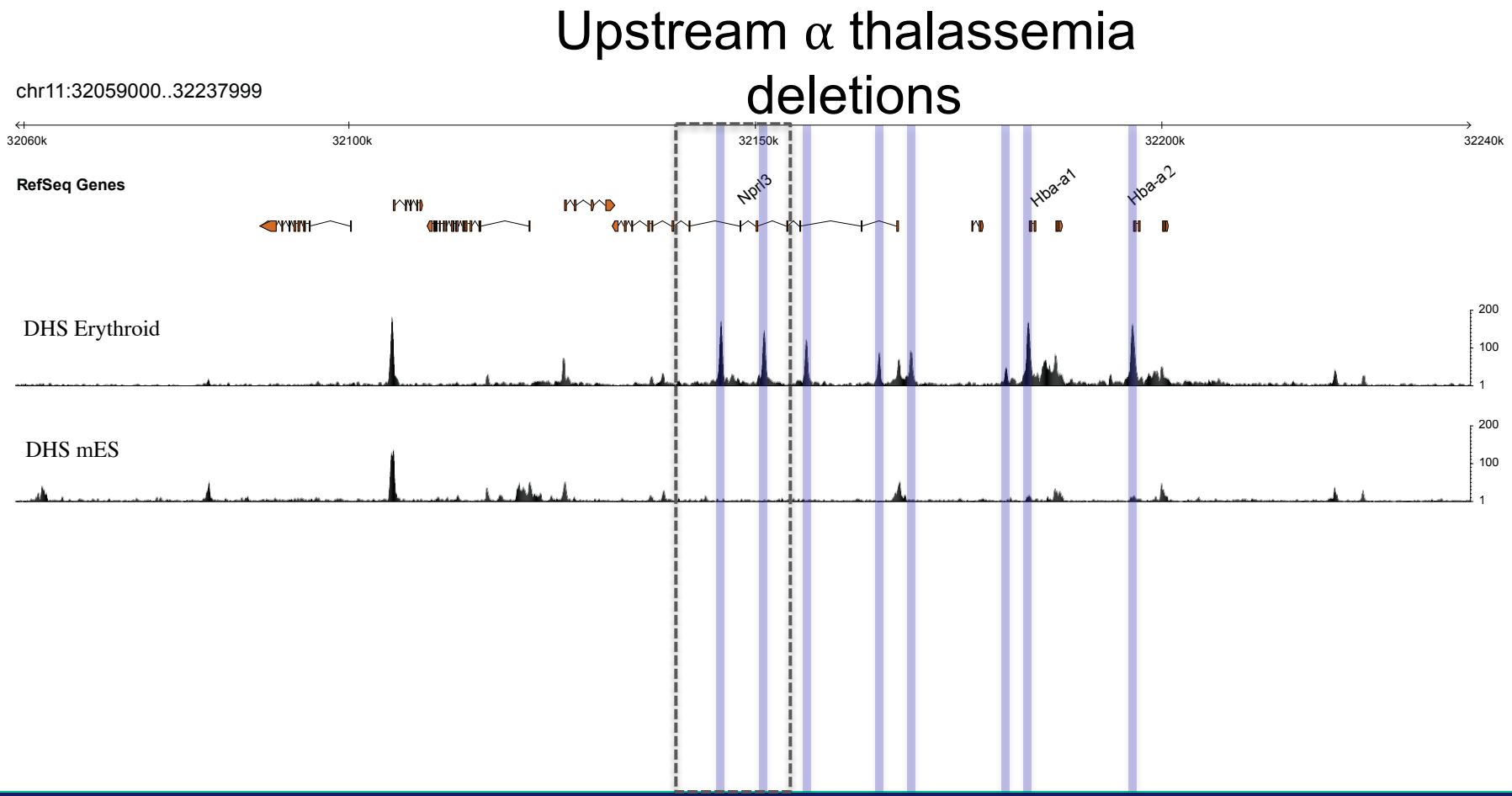
DNase 1 Hypersensitive sites



Erythroid Transcription Factors in the α Globin Locus

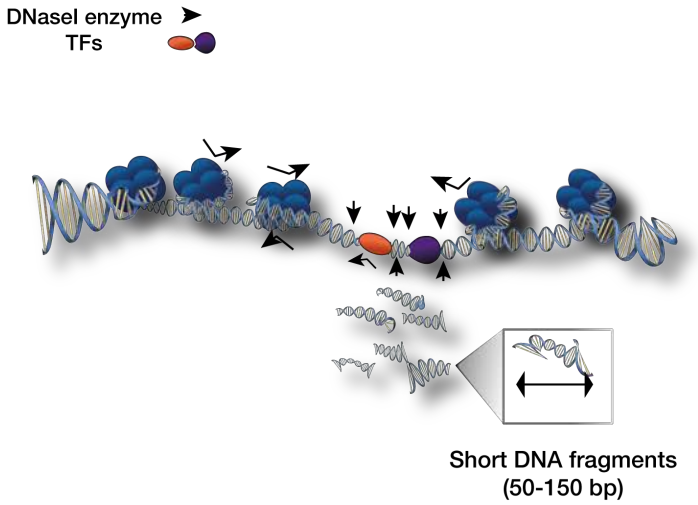


Comparative DHS Analysis - Mouse α Globin Locus



Open-chromatin assays

DNase-Seq

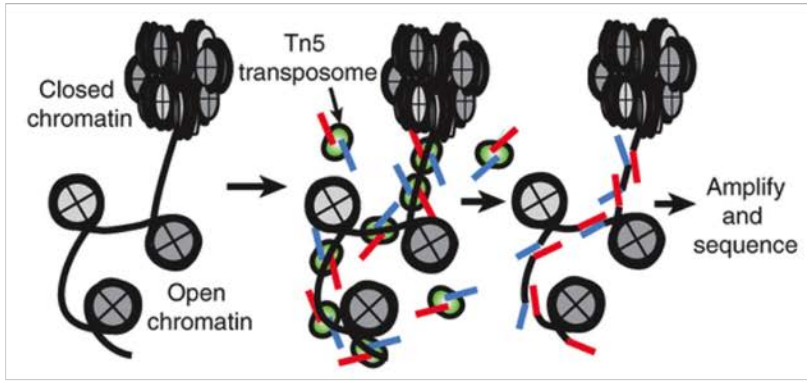


- Classic method
- Difficult to setup and optimise (cell-specific)
- Next generation sequencing protocol rather tedious

- Requires high cell number: 30-50 million cells

Boyle et al., 2008
Hesselberth et al.,2011
Neph et al., 2012
Hosseini et al.,2013

ATAC-Seq

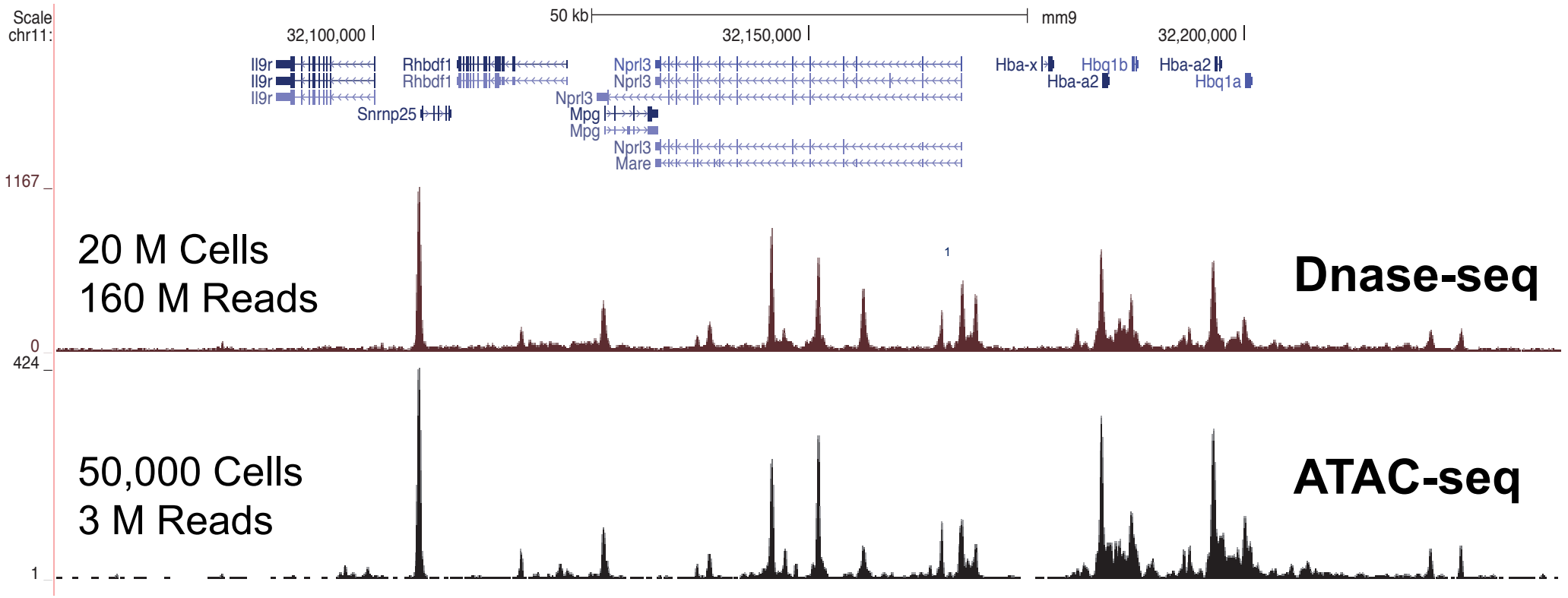


- Recent development
- Protocol relatively easy to setup
- Rapid NGS protocol

- Adapted for low cell number: 50000 cells or even 500

Buenrostro et al.,2013

Chromatin Accessibility Assay



scATAC-seq

Buenrostro et al Nat Methods 2013
Maria Suciu Unpublished

RNA-seqs; Plural

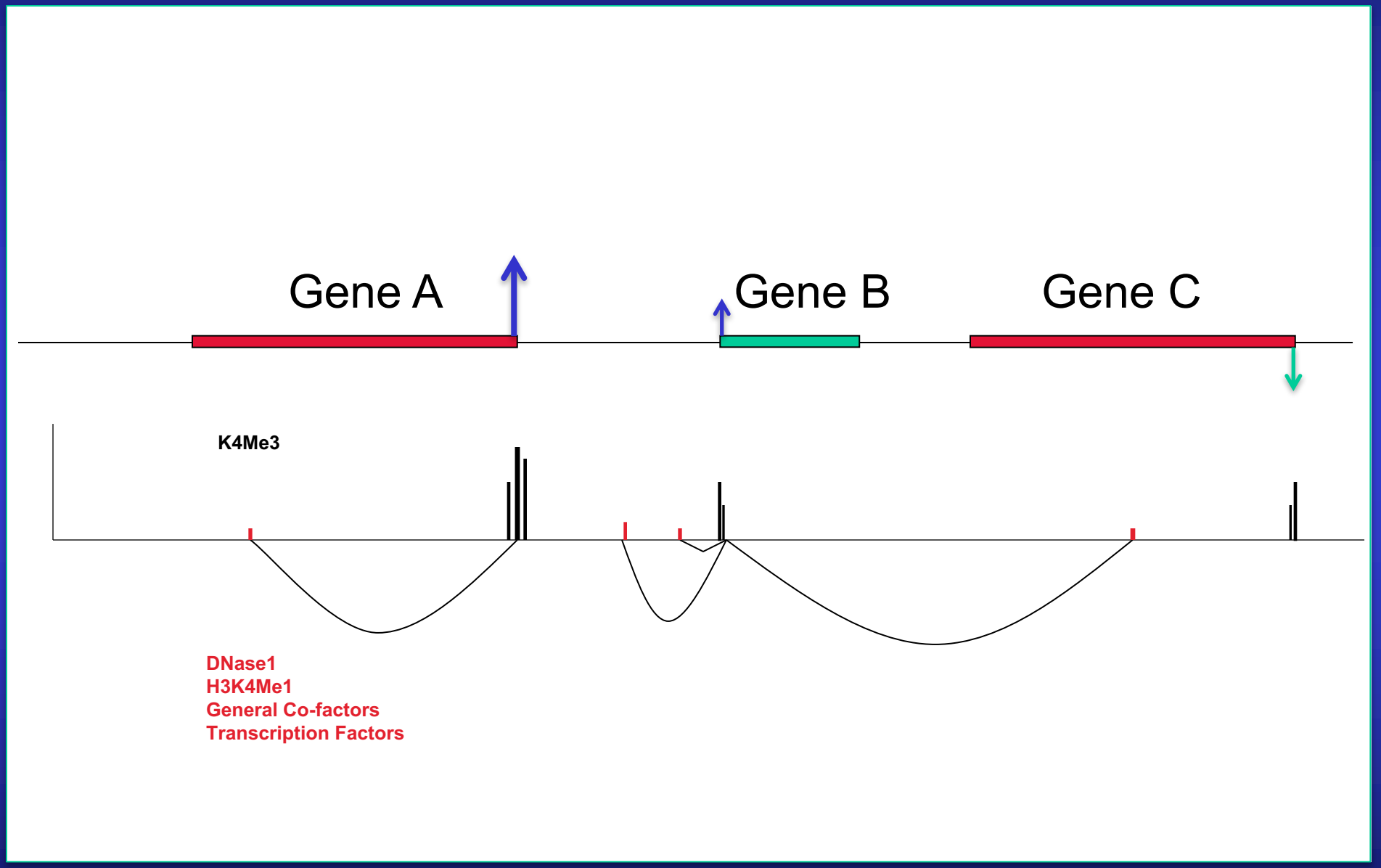
PolyA+ AAAAAAAAA

PolyA-

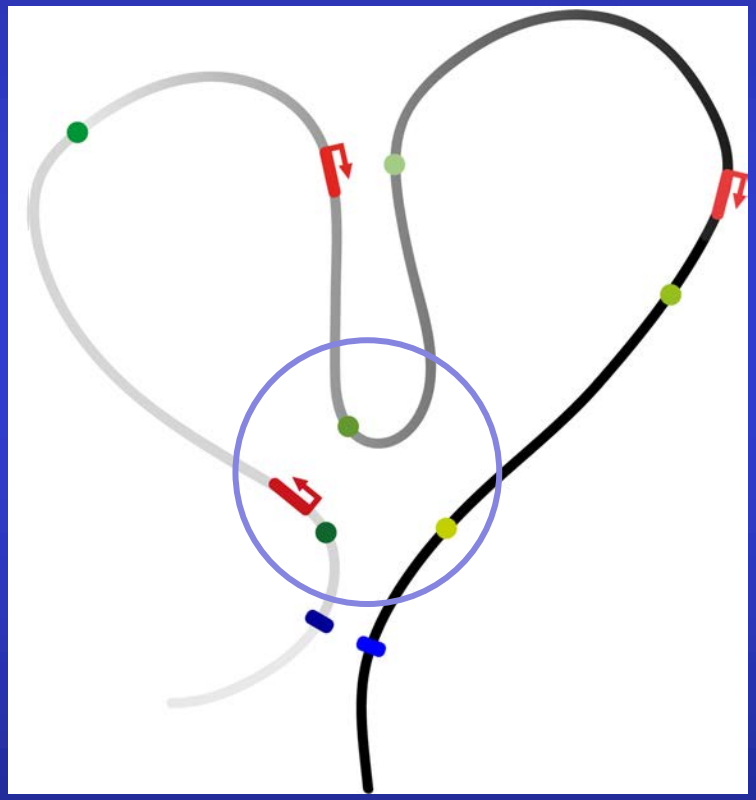
Specialist AAAAAAAAA

Nascent AAAAAAAAA

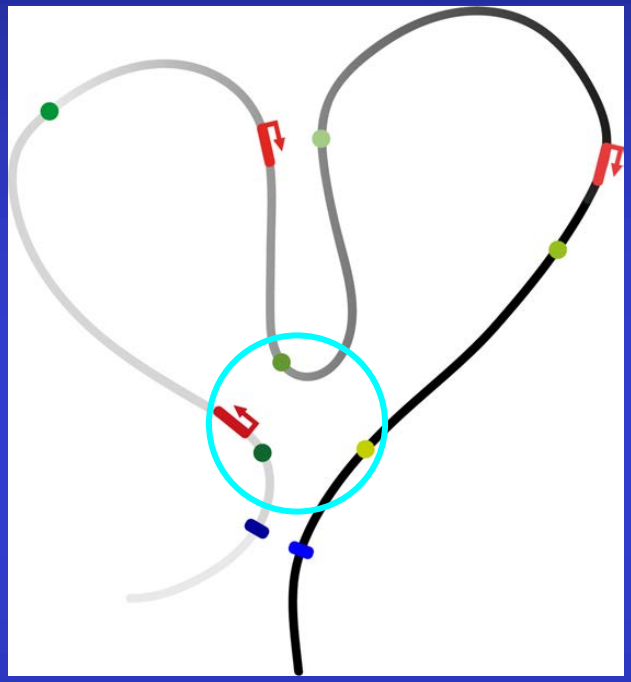
Connecting REs to genes

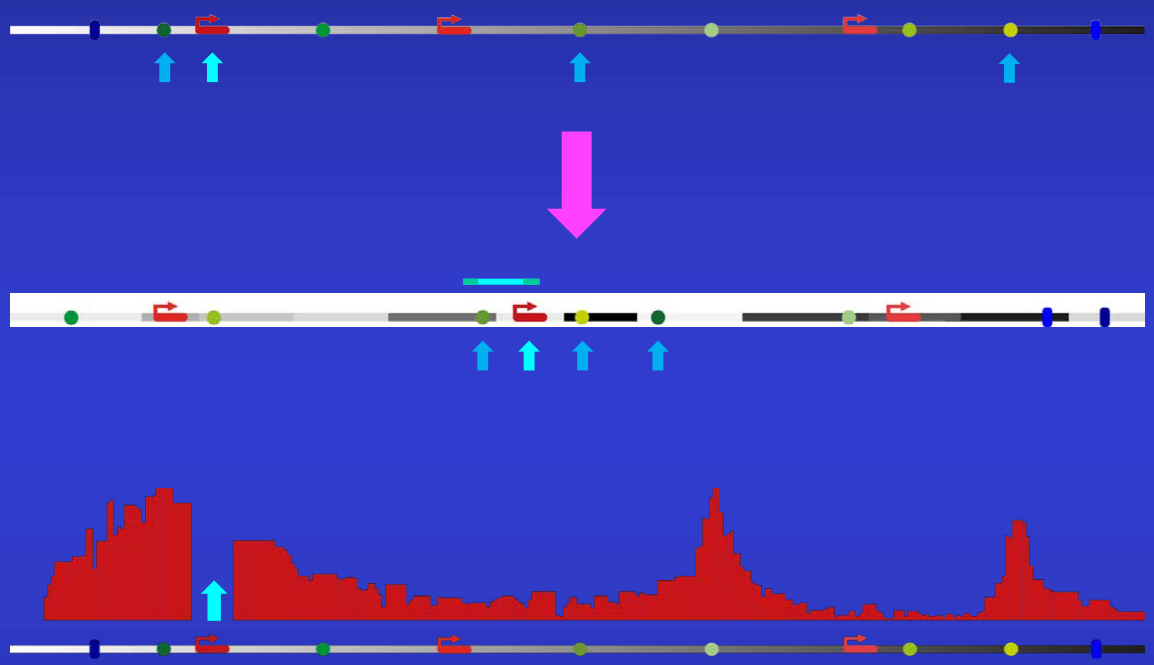


Chromosome Conformation Capture (3C)



Chromosome Conformation Capture (3C)





Capture-C interaction profile

Hughes lab

Davies JO et al. Nature Methods **13**:74 (2016)
Davies JO et al. Nature Methods **142**:125 (2017)

Increases in contact frequency at active and in active regulatory elements