Introduction to NGS Technologies

Ignacio Medina

im411@cam.ac.uk

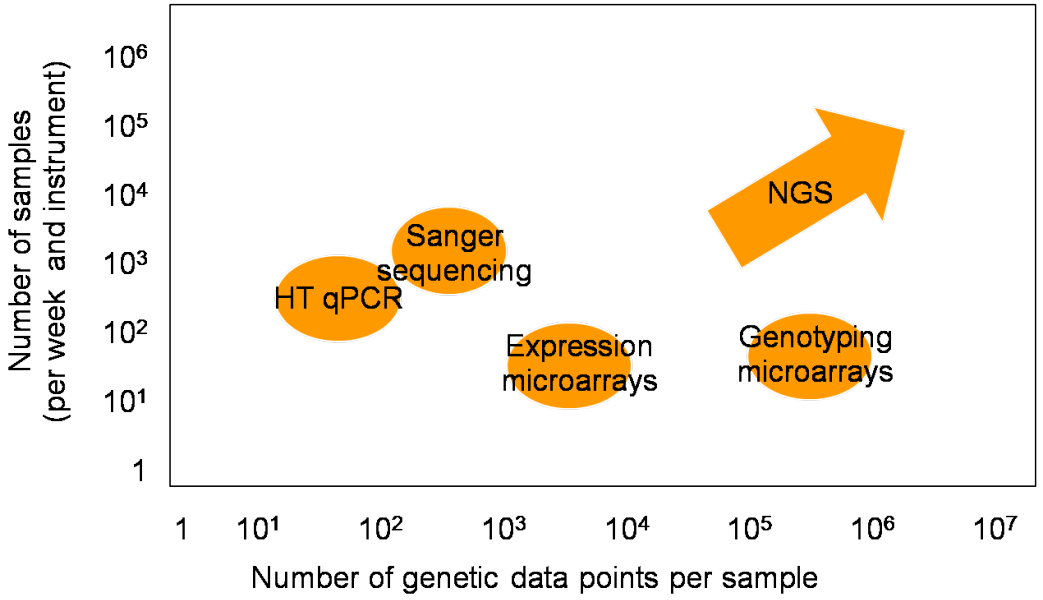
Head of Computational Biology Lab
HPC Service, University of Cambridge, UK

EMBL-EBI Scientific collaborator
Genome Campus, Hinxton, Cambridge, UK

UNIVERSITY OF CAMBRIDGE

Relative throughput of the different HT technologies

NGS emerges with a potential of data production that will, eventually wipe out conventional HT technologies in the years coming



Too many sequences to be handled in a standard computer

	Sanger (1st-gen) Sequencing	Next-Gen Sequencing, and 3rd generation
Whole Genome	Human (early drafts), model organisms, bacteria, viruses and mitochondria (chloroplast), low coverage	New human (!), individual genome, exomes, 2,500 normal (1K genome project), 25,000 cancer (TCGA and ICGC initiatives), CNV, matched control pairs, time course, rare-samples
RNA	cDNA clones, ESTs, Full Length Insert cDNAs, other RNAs	RNA-Seq: Digitization of transcriptome, alternative splicing events, miRNA, allele specific transcripts
Communities	Environmental sampling, 16S RNA populations, ocean sampling,	Human microbiome, deep environmental sequencing, Bar-Seq
Other		Epigenome, rearrangements, ChIP-Seq

NGS technologies

Cost-effective
Fast
Ultra throughput
Cloning-free
Short reads

Differences between the various platforms:

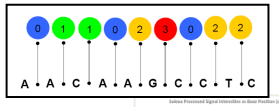
- Nanotechnology used.
- Read length
- Chemistry and enzymology.
- Signal to noise detection in the software
- Software/images/file size/pipeline
- Cost
- Applications

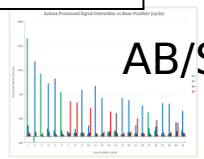
Similarities- LOTS of DATA General ways of dealing at the sequences

- Assemble them and look at what you have
- You map them (align against a known genome) and then look at what you have.
- Or a mixture of both!
- Sometimes you select the DNA you are sequencing
- or you try to sequence everything
- Depends on biological question, sequencing machine you have, and how much time and money you have.
- **NGS is relatively cheap but think what you want to answer, because the analysis won't do magic**

Next-gen sequencers

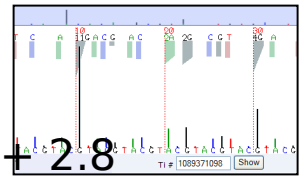
From John McPherson, OICR





Illumina/HiSeq short-read sequencers 10-600Gb, 150/150 bp, 11 days

AB/SOLiD5500, 200+Gb,75/35 bp reads, 10 days

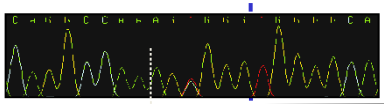


Roche 454 GS FLX+ 2.8

(0.2-1Gb in 800-1000 bp reads,
1M reads, 20 hours)

ABI capillary sequencer

(0.04-0.08 Mb in 450-800 bp reads,
96 reads, 1-3 hours)



bases per machine run

100 Gb
10 Gb
1 Gb
100 Mb
10 Mb
1 Mb

10 bp
100 bp
1,000 bp

read length

Next Generation Sequencers

In the past 3 main platforms:

- **Solexa/illumina**
- **Roche 454**
- **ABI SOLiD**

- Follow an approach similar to Sanger sequencing, but do away with separation of fragments by size and "read" the sequence as the reaction occurs
- Several different "next generation" sequencing platforms developed and commercialized, more on the way.
- Simultaneously sequence entire libraries of DNA sequence fragments

454 (Roche)

- First next generation method to be commercially available

- Uses a “sequencing by synthesis” (SBS) approach:
  - DNA is broken into pieces of 500-1,400 bp, ligated to adaptors, and amplified on tiny beads by PCR (emulsion PCR)
  - Beads (with DNA attached) are placed into tiny wells (one bead per well) on a PicoTiter Plate that has millions of wells. Each well is connected to an optical fibre.
  - DNA is sequenced by adding polymerase and DNA bases containing pyrophosphate. The different bases (A,C,G,T) are added sequentially in a flow chamber
  - When a base complementary to the template is added, the pyrophosphate is released and a burst of light is produced
  - The light is detected and used to call the base

- Initially 100-150 bp, but they have been improved to 600-1000 bp

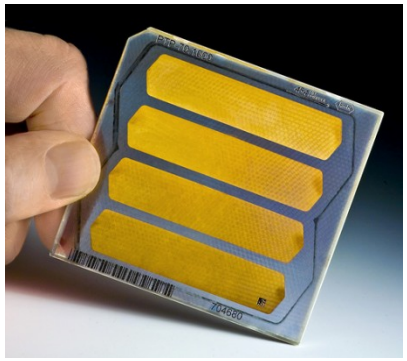
- >1 million, filter-passed reads per run (20 hours)

- 1 billion bases per day

Roche 454: GS FLX System

- Good for
  - “*de novo*“ sequencing (longer reads)
  - Resequencing (expensive)
  - New bacterial genomes.
  - Amplicons
- Pyrosequencing. Bias with long polinucleotide streches

Roche 454 GS FLX



Throughput	400-600 million high-quality, filter-passed bases per run* 1 billion bases per day
Run Time	10 hours
Read Length	Average length = 400 bases
Accuracy	Q20 read length of 400 bases (99% at 400 bases and higher for prior bases)
Reads per run	>1 million high-quality reads
Data	Trace data accepted by NCBI since 2005
Computing Requirements	Cluster recommended (Roche GS FLX Titanium Cluster available)
Robustness	No complex optics or lasers; reagents have long shelf life

GS Junior, benchtop

System Performance	
Throughput	35 million high-quality, filtered bases per run*
Run Time	10 hours sequencing 2 hours data processing
Avg. Read Length	400 bases*
Accuracy	Q20 read length of 400 bases (99% accuracy at 400 bases)
Reads per Run	100,000 shotgun, 70,000 amplicon
Sample Input	gDNA, amplicons, cDNA, or BACs depending on the application
Physical Dimensions	40 cm wide x 60 cm deep x 40 cm high (the size of a laser printer) Weight = 55 lbs.
Computing	Linux-based OS on HP desktop computer included. All software is point-and-click.

***Typical results**. Average read length and number of reads depend on specific sample and genomic characteristics*

Illumina

- Over 90% of all sequencing data is produced on Illumina systems.
- Uses a “sequencing by synthesis” approach:
  - DNA is broken into small fragments and ligated to an adaptor.
  - The fragments are attached to the surface of a flow cell and amplified.
  - DNA is sequenced by adding polymerase and labeled reversible terminator nucleotides (each base with a different color).
  - The incorporated base is determined by fluorescence.
  - The fluorescent label is removed from the terminator and the 3’ OH is unblocked, allowing a new base to be incorporated
- Started with 35 bp, increased now to up to 150 bp
- One run can give up to 10-600 Gb, 300-6000 million paired-end reads
- 75-85% of bases at or above Q30

Illumina Systems

Illumina HiSeq 2500

600 Gb/run in 11 days
2x100 bp fragments
3-6 billion reads per run

Illumina MiSeq

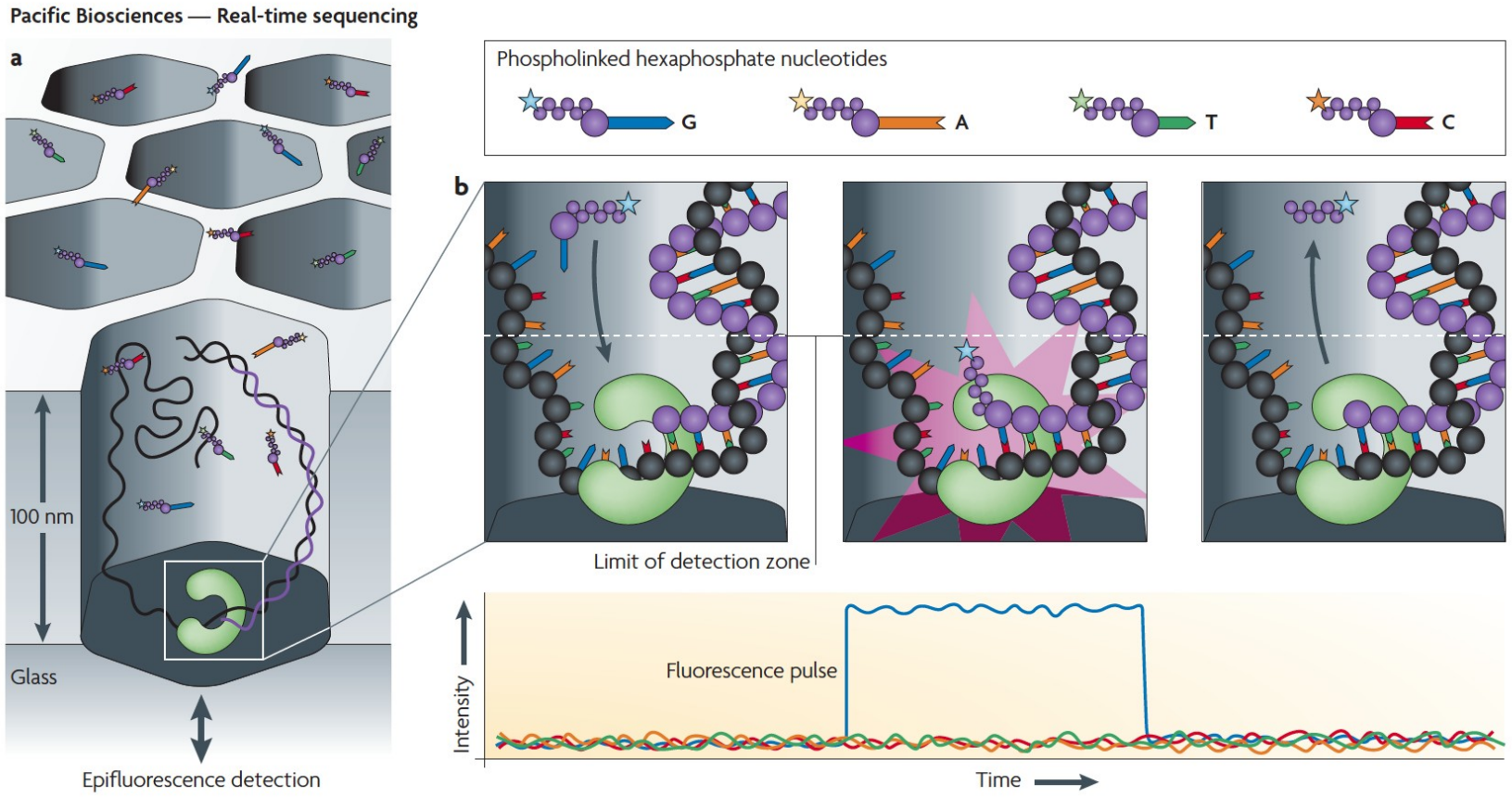
175-245 Mb 4h 1x 36bp
1.5-2.0 Gb 27h 2x150 bp

Illumina HiSeq X Ten

1800 Gb/run in <3 days
2x150 bp fragments
6 billion reads per run

Consists of a set of **10 HiSeq X** ultra-high-throughput instruments that deliver over **18,000 human genomes** per year at the price of **$1000** per genome.

PacBio

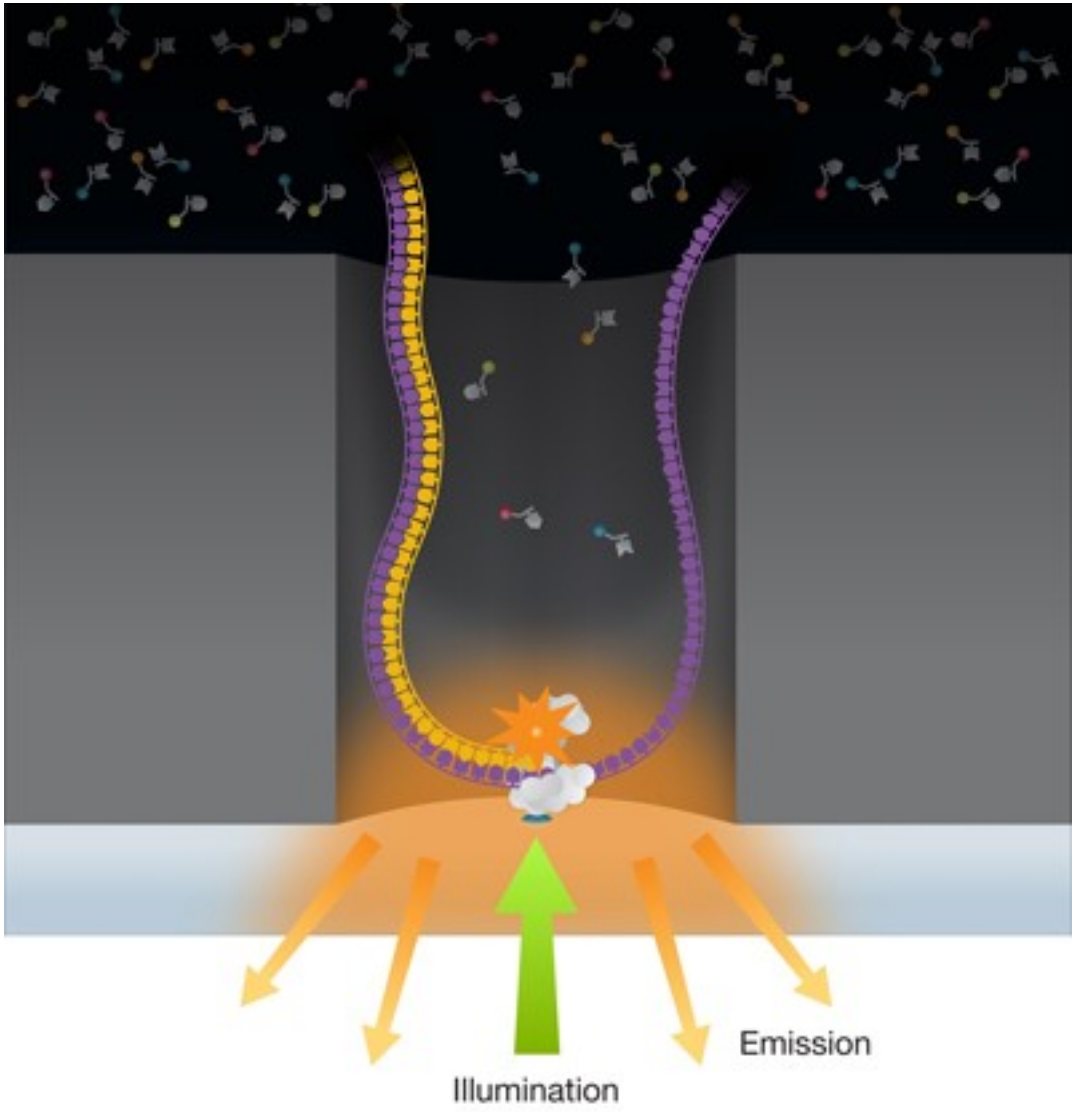


From Michael Metzker, http://view.ncbi.nlm.nih.gov/pubmed/19997069

Pacific Bioscience

SMRT: Singel Molecule Real time DNA synthesis
Up to 12000 nt
50 bases/second

ZMW: Zero Mode Waveguide



Ion Torrent

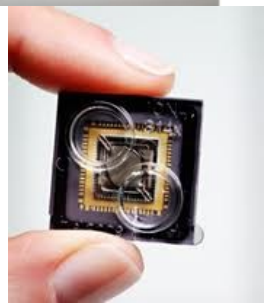
$ 50.000

$ 500 /sample

1 hour/run

> 200 nt lengths

Reads H+ released by DNA polymerase



Comparison

Roche 454

- Long fragments
- Errors: poly nts
- Low throughput
- Expensive

- De novo sequencing
- Amplicon sequencing
- RNASeq

Illumina

- Short fragments
- Errors: Hexamer bias
- High throughput
- Cheap

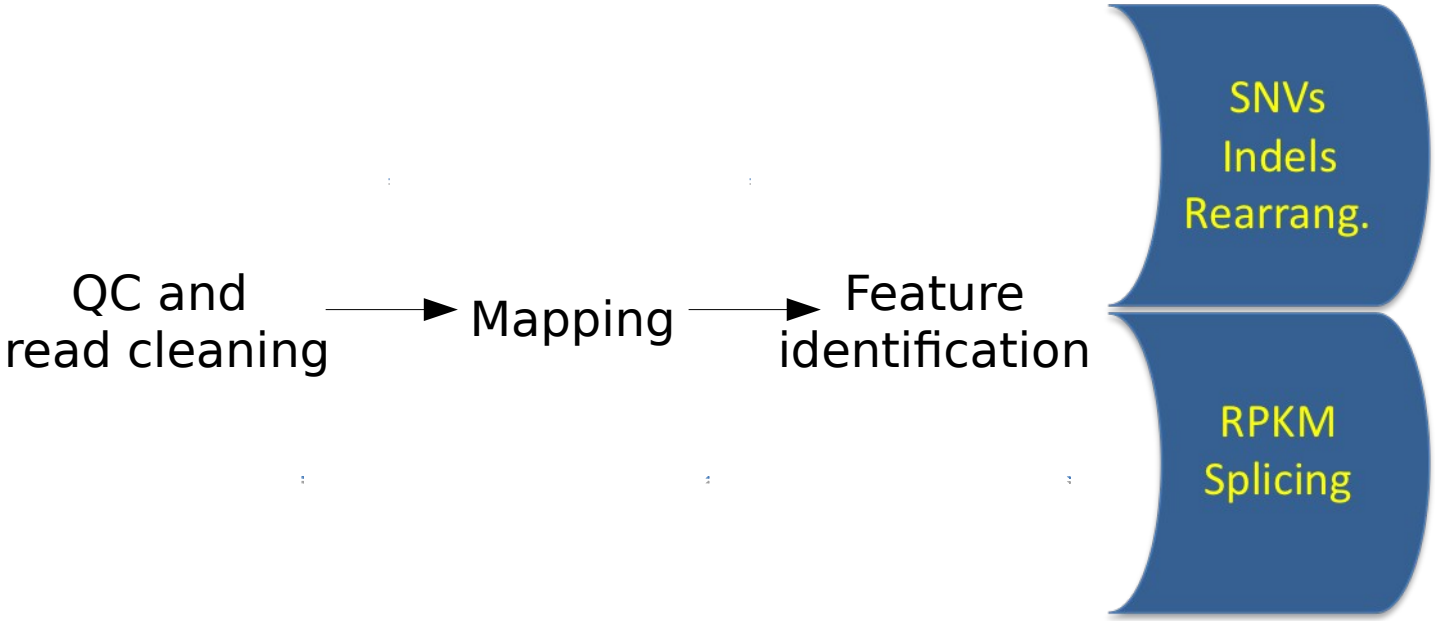
- Resequencing
- De novo sequencing
- ChipSeq
- RNASeq
- MethylSeq

SOLiD

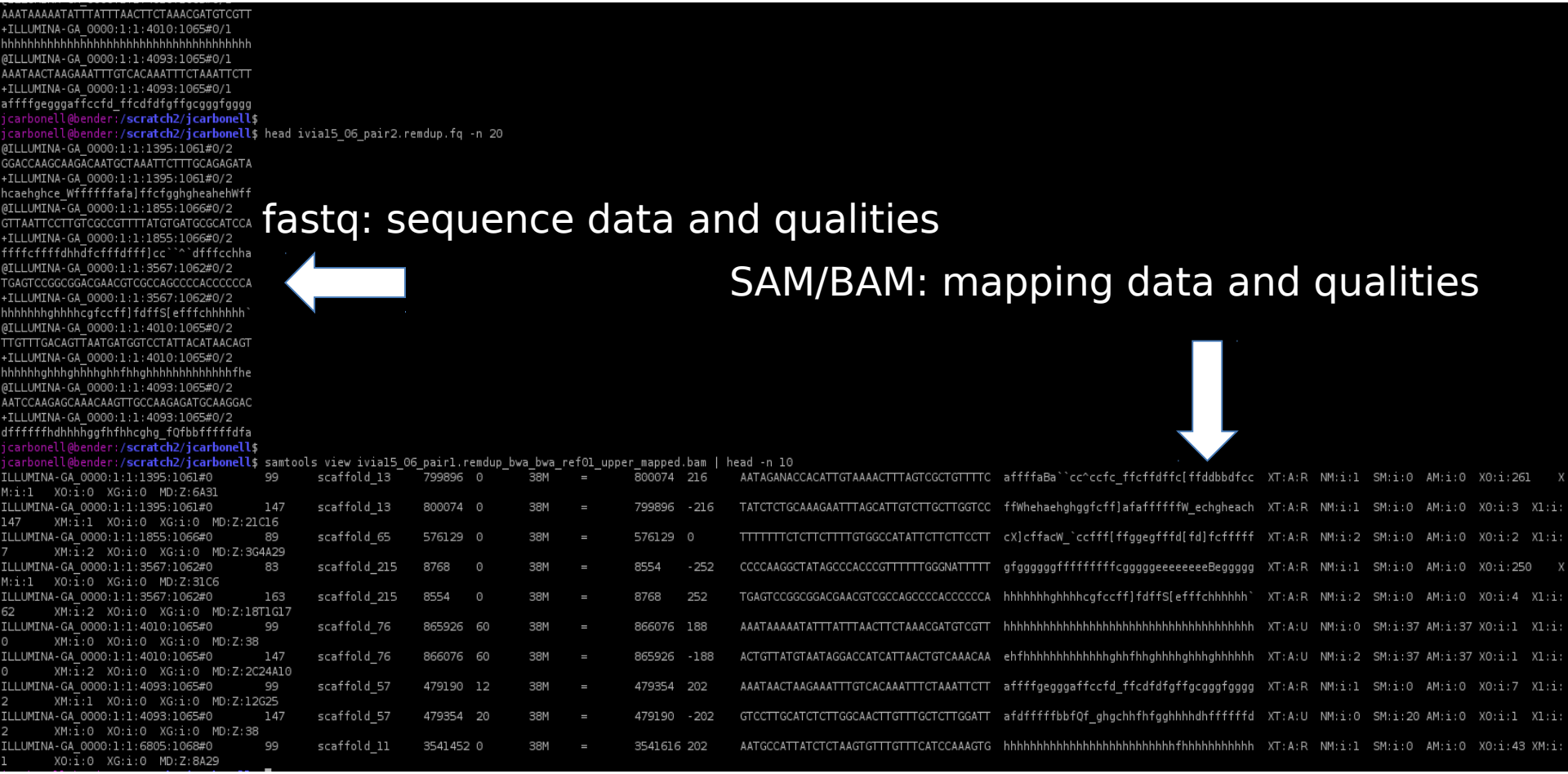
- Short fragments
- Color-space
- High throughput
- Cheap

- Resequencing
- ChipSeq
- RNASeq
- MethylSeq

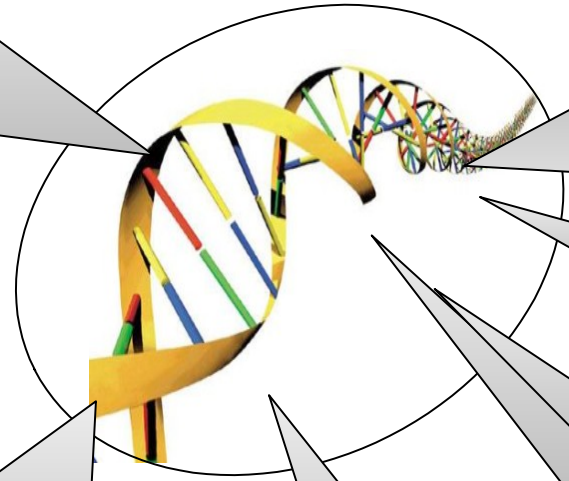
Basic steps NGS data processing



File formats



Most common applications of NGS



RNA-seq /Transcriptomics
- Quantitative
- Descriptive Alternative splicing
- miRNA profiling

Resequencing
- Mutation calling
- Profiling
- Genome annotation

***De novo* sequencing**

Exome sequencing
Targeted sequencing

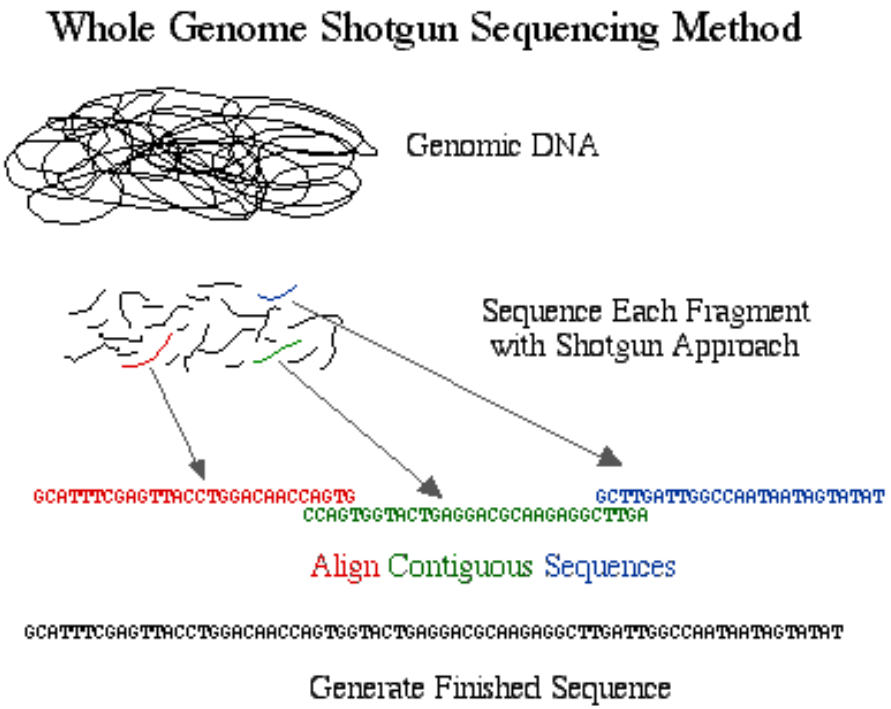
Copy number variation

Metagenomics
Metatranscriptomics

ChIP-seq /Epigenomics
- Protein-DNA interactions
- Active transcription factor binding sites
- Histone methilation

DNA sequencing - 1

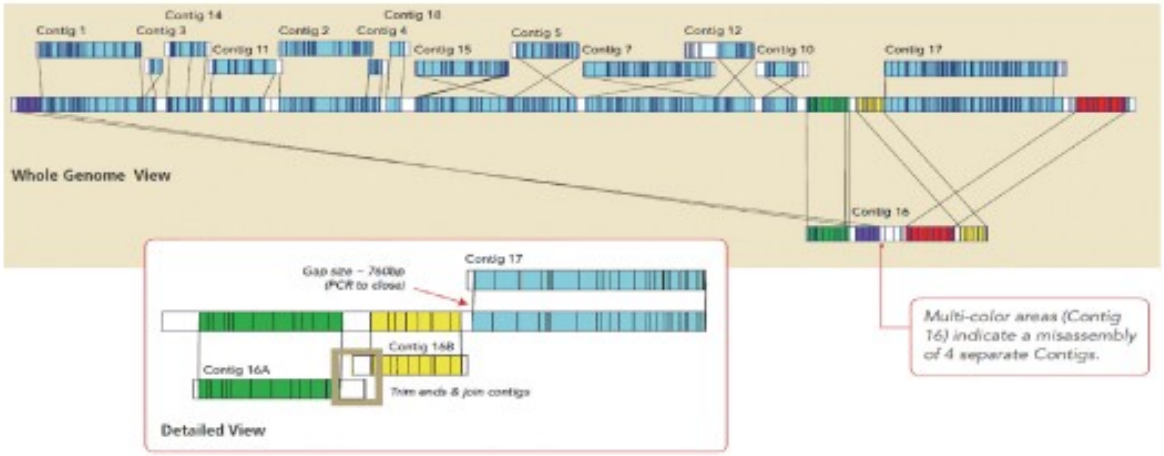
- **Whole GENOME Resequencing**
  - Need reference genome
  - Variation discovery



DNA sequencing - 2

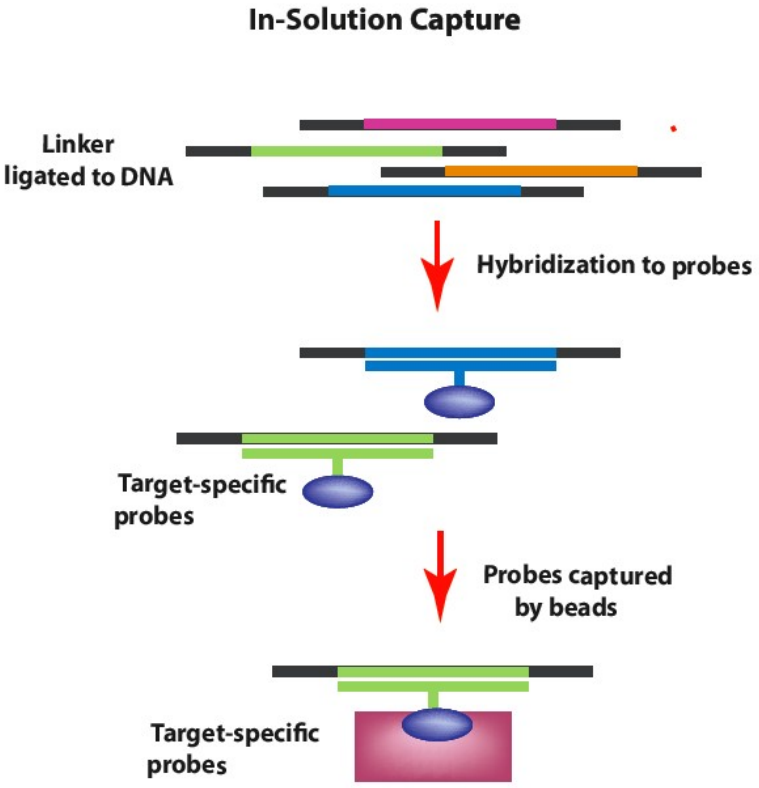
- **Whole GENOME “de novo” sequencing**
  - Uncharacterized genomes with no reference genome available
  - known genomes where significant structural variation is expected.
  - Long reads or mate-pair libraries. Sequencing mostly done by Roche 454 and also Illumina.
  - Assembly of reads is needed: Computational intensive

  - E.g. Genome bacteria sequencing



DNA sequencing - 3

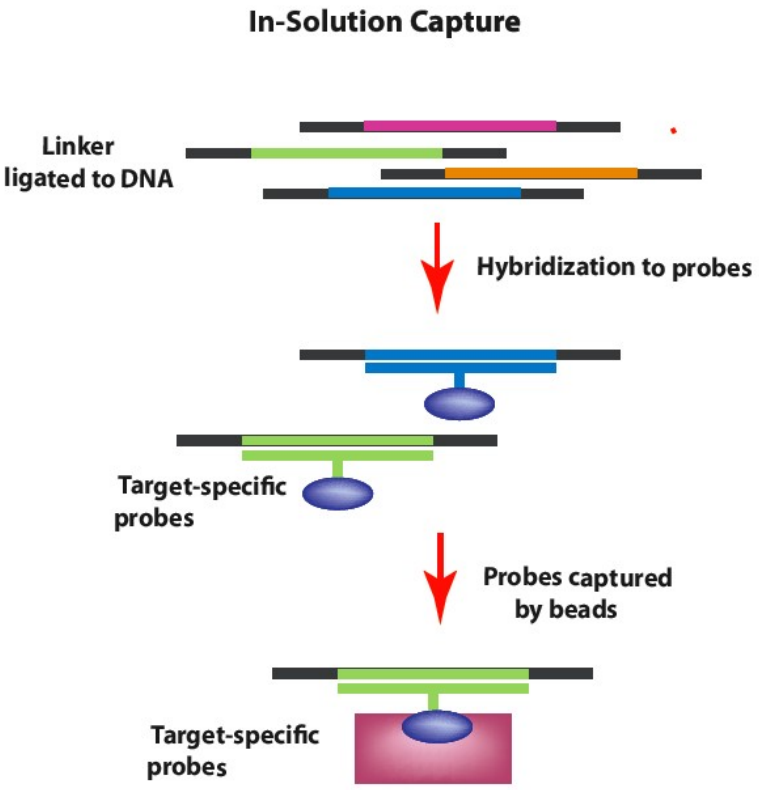
- **Whole EXOME Resequencing**
  - Need reference genome
    - Available for Human and Mouse
  - Variation discovery on ORFs
    - 2% of human genome (lower cost)
    - 85% disesease mutation are in the exome
  - Need probes complementary to exons
    - Nimblegen
    - Agilent



- E.g. Human exome

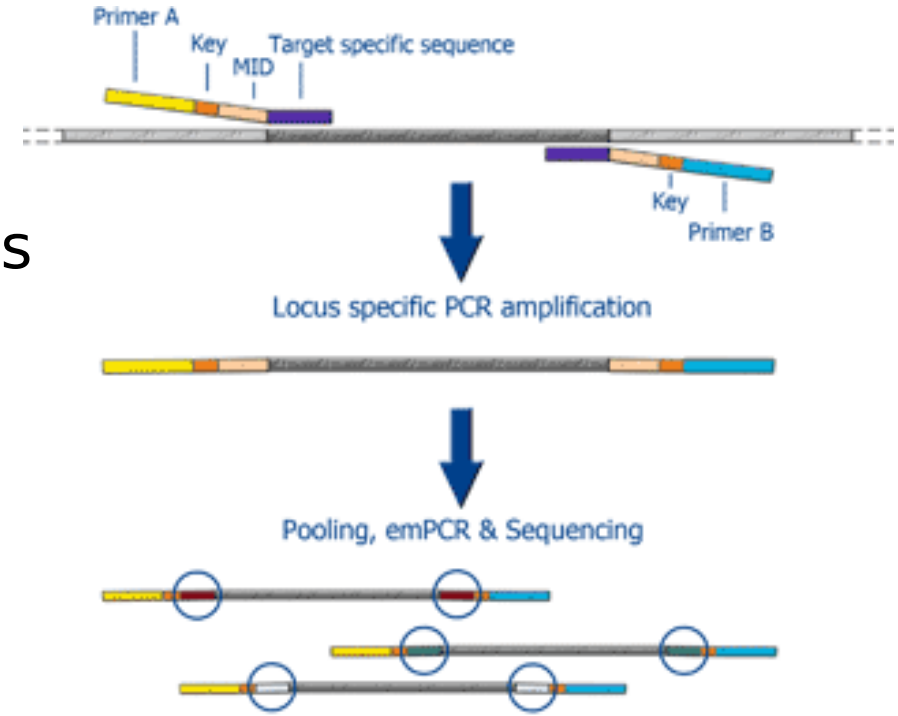
DNA sequencing - 4

- **Targeted Resequencing**
  - Capture of specific regions in the genome
- **Custom genes panel sequencing**
  - Allows to cover high number of genes related to a disease
  - *E.g. Disease gene panel*
- Low cost and quicker than capillary sequencing
- Multiplexing is possible
- Need custom probes complementary to the genomic regions
  - Nimblegen
  - Agilent



DNA sequencing - 5

- **Amplicon sequencing**
  - Sequencing of regions amplified by PCR.
  - Shorter regions to cover than targeted capture
  - No need of custom probes
  - Primer design is needed
  - High fidelity polymerase
  - Multiplexing is needed



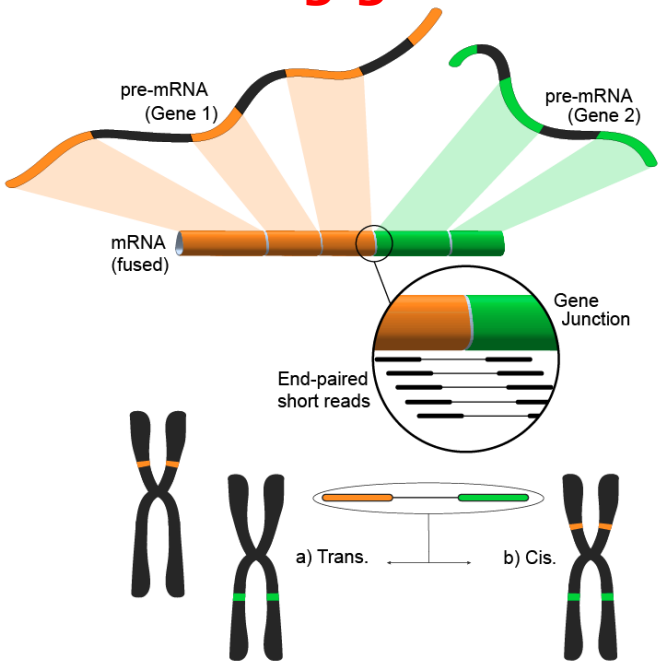
- *E.g. P53 exon amplicon sequencing*

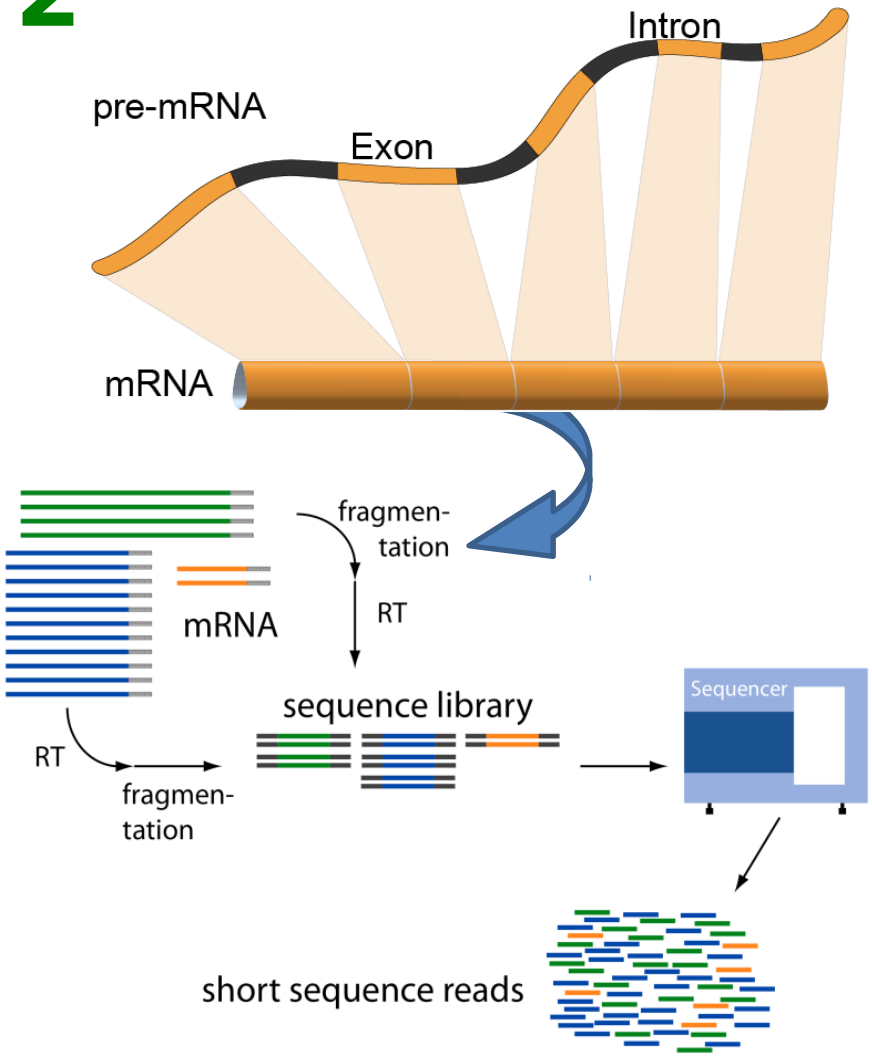
Transcriptomics - 1

- **RNA-Seq**
  - Sequencing of mRNA
  - rRNA depleted samples
  - Very high dynamic range
  - No prior knwoledge of expressed genes
  - Gives information about (richer than microarrays)
    - Differential expression of **known or unknown** transcripts during a treatment or condition
    - **Isoforms** and
    - New **alternative splicing** events
    - **Non-coding** RNAs
    - Post-transcriptional mutations or **editing**,
    - **Gene fusions**.

Transcriptomics - 2

- **RNA-Seq**
  - Sequencing of **mRNA**
  - **Detecting gene fusions**





Applications of RNAseq

Qualitative:
* Alternative splicing
* Antisense expression
* Extragenic expression
* Alternative 5’ and 3’ usage
* Detection of fusion transcripts
....

Tophat/Cufflinks
Scripture
Alexa

Quantitative:
* Differential expression
* Dynamic range of gene expression
....

edgeR
DESeq
baySeq
NOISeq

Advantages of RNAseq?

RNAseq

* Non targeted transcript detection
* No need of reference genome
* Strand specificity
* Find novels splicing sites
* Larger dynamic range
* Detects expression and SNVs
* Detects rare transcripts

....

microarrays

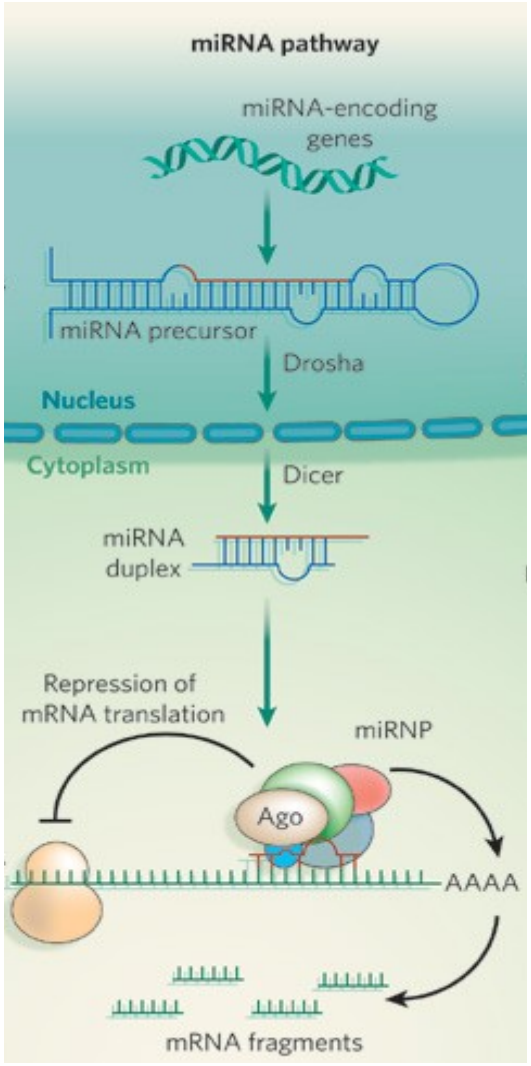
* Restricted to probes on array
* Needs genome knowledge
* Normally, not strand specific
* Exon arrays difficult to use
* Smaller dynamic range
* Does not provide sequence info
* Rare transcripts difficult

....

and…. are there any disadvantages?????

Transcriptomics - 3

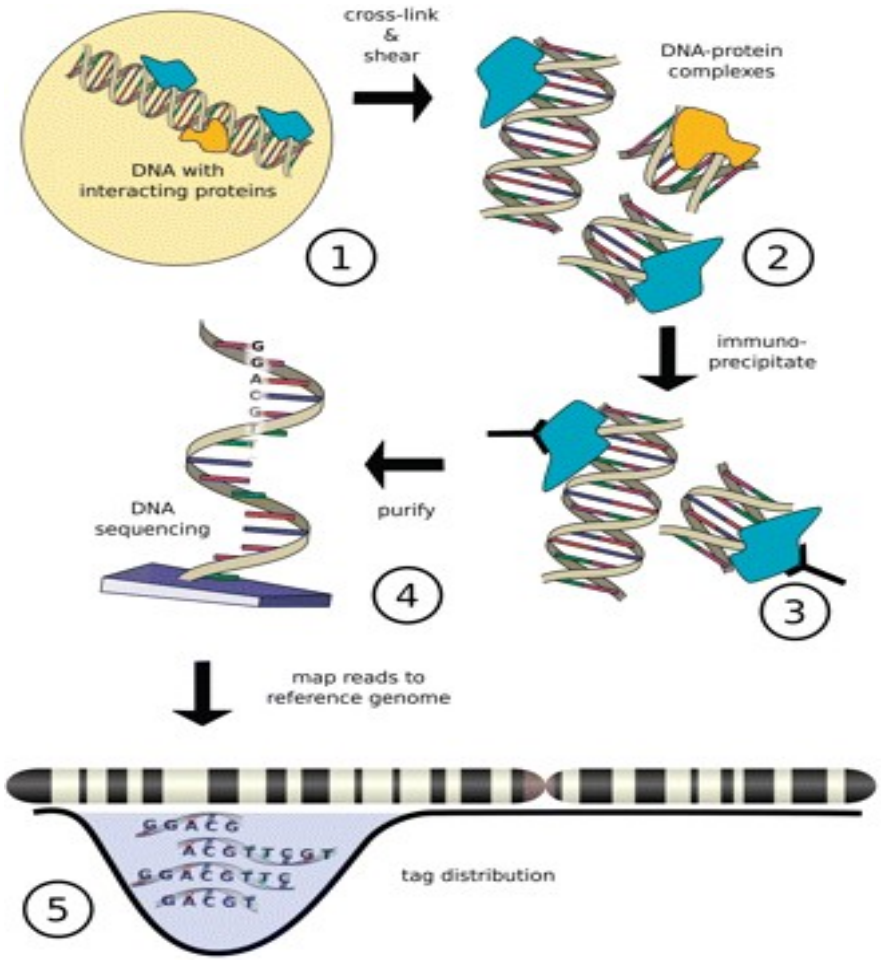
- **miRNA/small nonCoding RNA sequencing**
  - RNA Size selection step
    - 18-40 bp
  - Profiling of known miRNAs
  - miRNA discovery



TFBS detection

ChIP-Seq

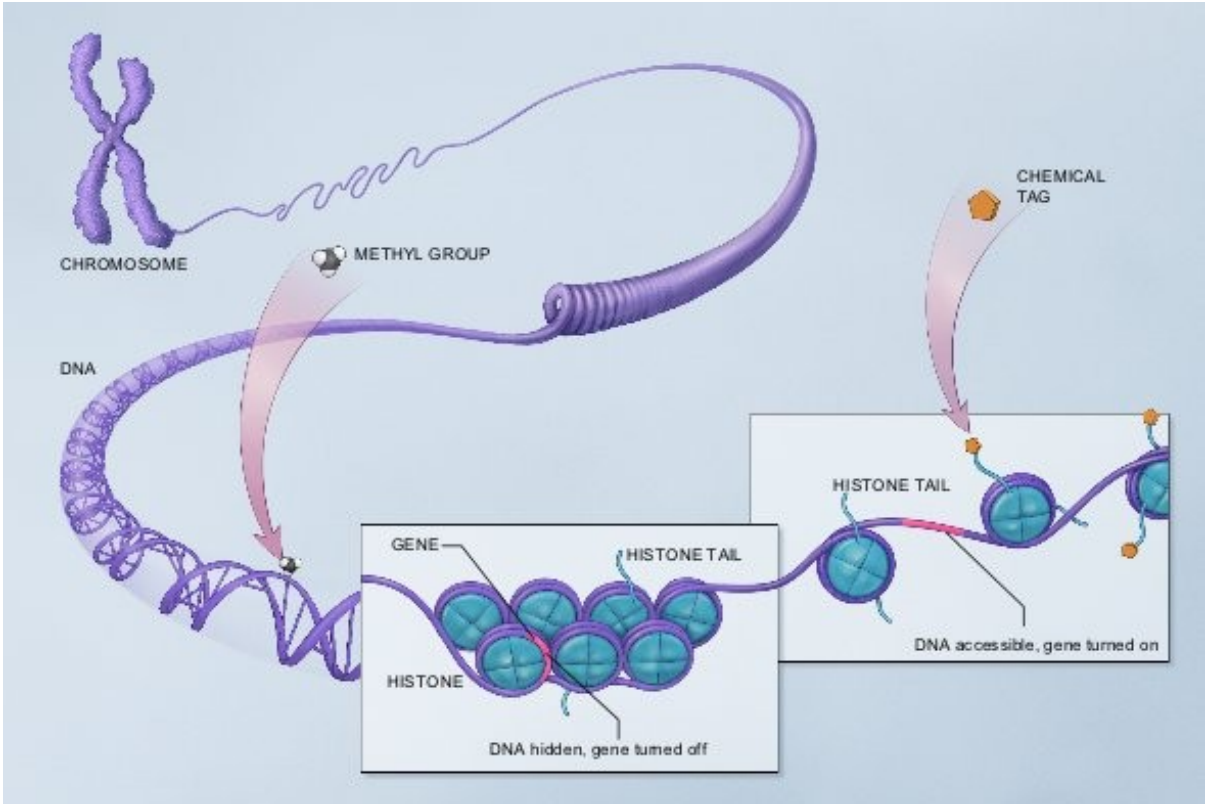
- Identification of genomic region for gDNA binding proteins:
  - Transcription Factor binding site detection



Epigenomics - I

Epigenomics refers to functionally relevant modifications to the genome that do not involve a change in the nucleotide sequence

- *Play a role in turning genes off or on*



Epigenomic Marks.

a) Methyl groups attach to the backbone of a DNA molecule.

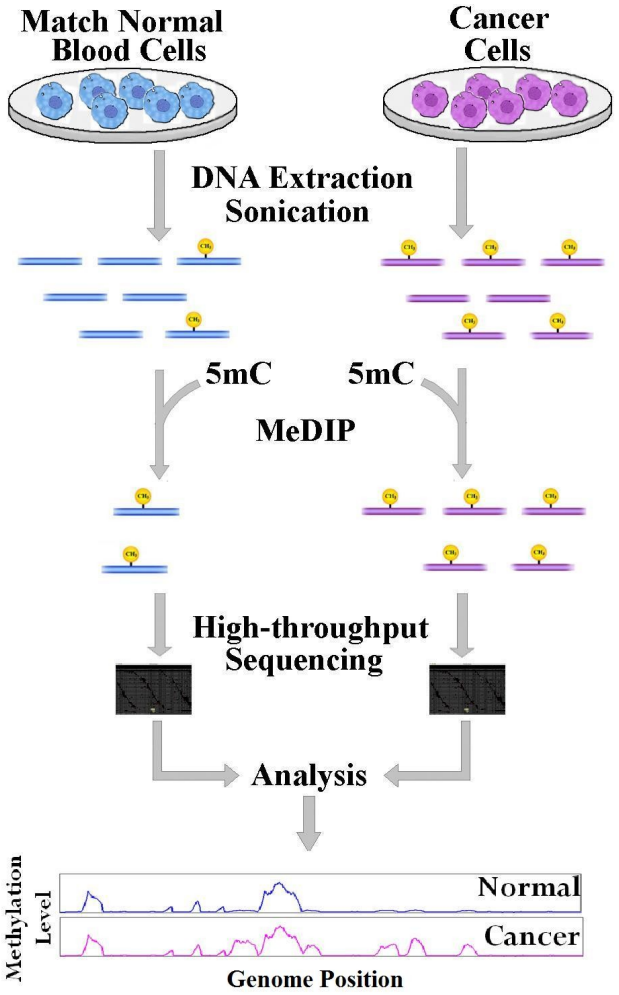
b) A variety of chemical tags attach to the tails of histones. This action affects how tightly DNA is wound around the histones.

ChIP-Seq: Histone methylation detection

Epigenomics - 2

- **Methyl-Seq**
  - CpG island methylation
  - Bisulfite sequencing-based method

  > E.g. Cancer studies.
    - Different degree of chromatin methylation affects expression of genes



New huge projects coming

- **Many big projects during the last years:**
  - ENCODE http://genome.ucsc.edu/ENCODE/
  - 1000 Genomes projects http://www.1000genomes.org/
  - ICGC http://icgc.org/
  - ...
- **New projects coming soon:**
  - BRIDGE https://bridgestudy.medschl.cam.ac.uk/index.shtml
  - Genomics England (http://www.genomicsengland.co.uk/) will produce tens of PB of data (1PB == 1000TB)

Successful NGStories

NATURE | ARTICLE OPEN

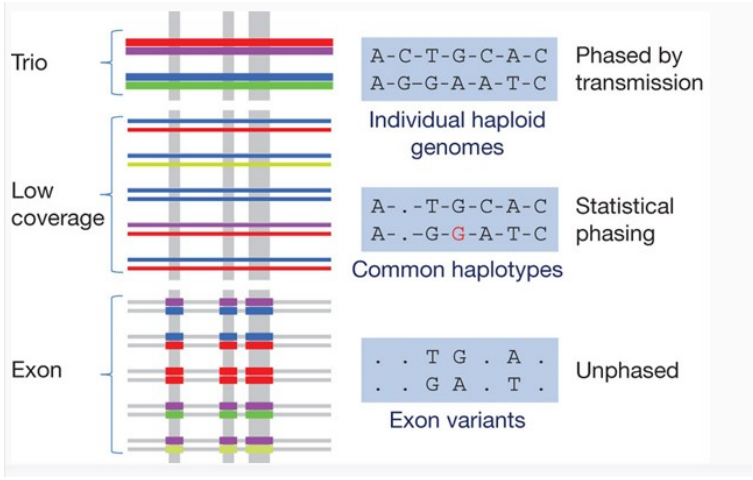
previous article next article

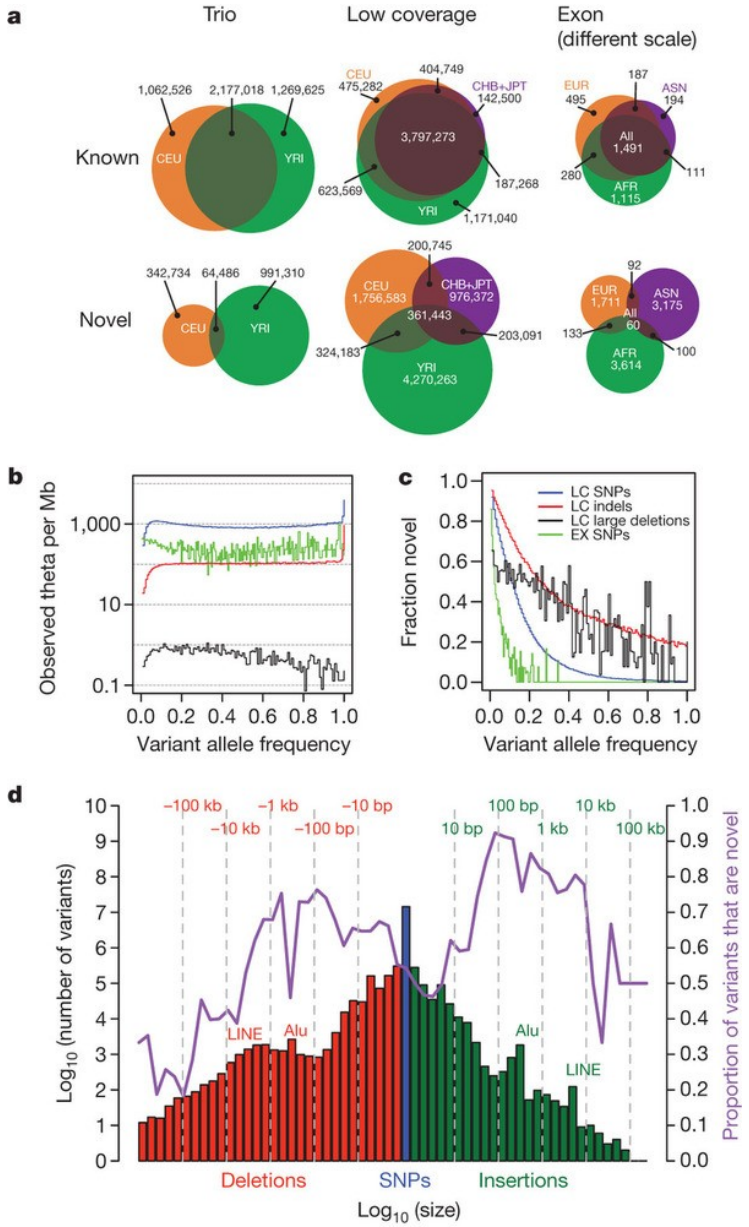
A map of human genome variation from population-scale sequencing

The 1000 Genomes Project Consortium

Affiliations | Contributions | Corresponding author

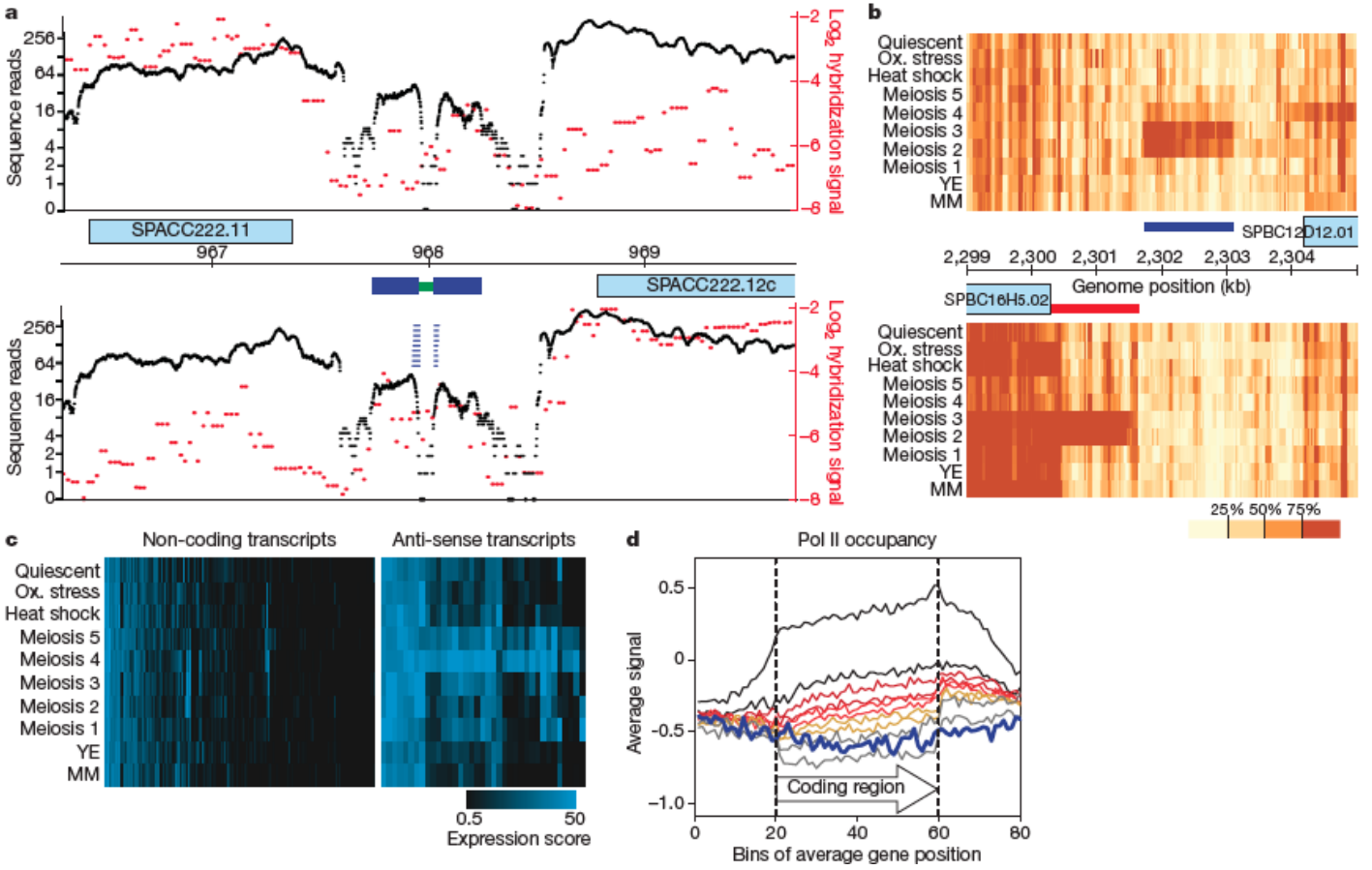
Nature **467**, 1061–1073 (28 October 2010) | doi:10.1038/nature09534
Received 20 July 2010 | Accepted 30 September 2010 | Published online 27 October 2010





Dynamic repertoire of a eukaryotic transcriptome surveyed at single-nucleotide resolution

Brian T. Wilhelm[1]*†, Samuel Marguerat[1]*†, Stephen Watt[1]†, Falk Schubert[1]†, Valerie Wood[1], Ian Goodhead[1]†, Christopher J. Penkett[1]†, Jane Rogers[1] & Jürg Bähler[1]†



Published in final edited form as:
Nat Genet. 2010 January ; 42(1): 30–35. doi:10.1038/ng.499.

Exome sequencing identifies the cause of a Mendelian disorder

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

[1]Department of Genome Sciences, University of Washington, Seattle, Washington, USA [2]Department of Pediatrics, University of Washington, Seattle, Washington, USA [3]Department of Pediatrics, University of Utah, Salt Lake City, Utah, USA [4]Department of Human Genetics, University of Utah, Salt Lake City, Utah, USA [5]Institute of Systems Biology, Seattle WA, USA [6]Department of Genetics and Genomic Sciences, Mount Sinai School of Medicine, New York, New York, USA [7]Department of Pediatrics, Johns Hopkins University, Baltimore, Maryland [8]Seattle Children's Hospital, Seattle, Washington, USA

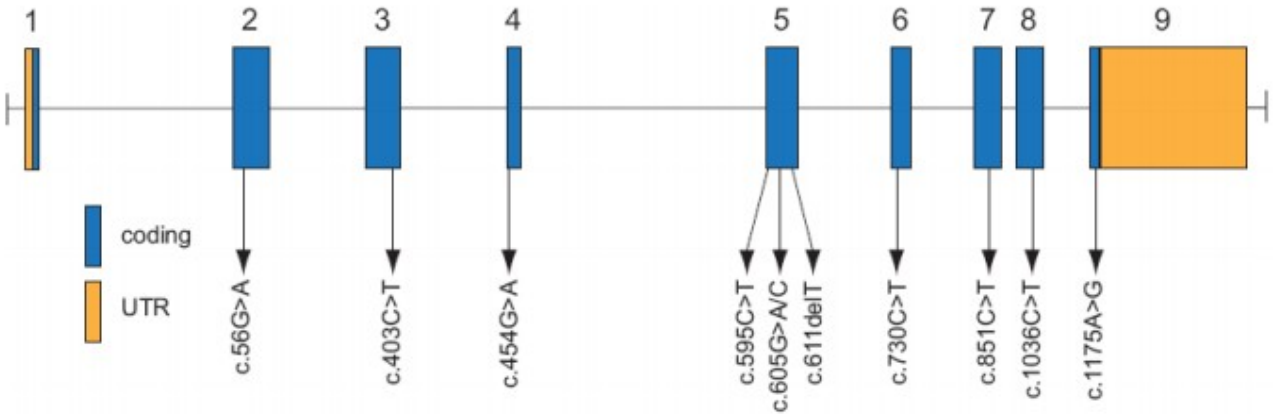


Figure 2. Genomic structure of the exons encoding the open reading frame of *DHODH*

DHODH is composed of 9 exons that encode untranslated regions (orange) and protein coding sequence (blue). Arrows indicate the locations of 11 different mutations found in 6 families with Miller syndrome.

liller syndrome

Method

Open Access

Targeted next-generation sequencing of a cancer transcriptome enhances detection of sequence variants and novel fusion transcripts

Joshua Z Levin*, Michael F Berger†, Xian Adiconis*, Peter Rogov*, Alexandre Melnikov*, Timothy Fennell‡, Chad Nusbaum*, Levi A Garraway†§ and Andreas Gnirke*

Addresses: *Genome Sequencing and Analysis Program, Broad Institute of MIT and Harvard, 320 Charles Street, Cambridge, MA 02141, USA. †Cancer Program, Broad Institute of MIT and Harvard, 5 Cambridge Center, Cambridge, MA 02142, USA. ‡Sequencing Platform, Broad Institute of MIT and Harvard, 320 Charles Street, Cambridge, MA 02141, USA. §Department of Medical Oncology and Center for Cancer Genome Discovery, Dana-Farber Cancer Institute, Harvard Medical School, 44 Binney Street, Boston, MA 02115, USA.

Figure 3
Sequences from NUP214-XKR3 fusion transcripts detected after hybrid selection. After hybrid selection, 152 reads were aligned to the transcriptome and detected as *NUP214-XKR3* fusions. From top to bottom, we observed 137, four, eight, and three reads for these transcripts. The *NUP214* (exon 27) to *XKR3* (exon 4) has a stop codon downstream (not shown). Only *NUP214* (exon 29) to *XKR3* (exon 4) retains an open reading frame downstream of the fusion. Before hybrid selection, eight reads were aligned to the transcriptome and detected as *NUP214-XKR3* fusions; only the *NUP214* (exon 29) to *XKR3* (exon 2) transcript was detected. Sequence from *NUP214* DNA is shown as lower case, and from *XKR3*, as bold and upper case.

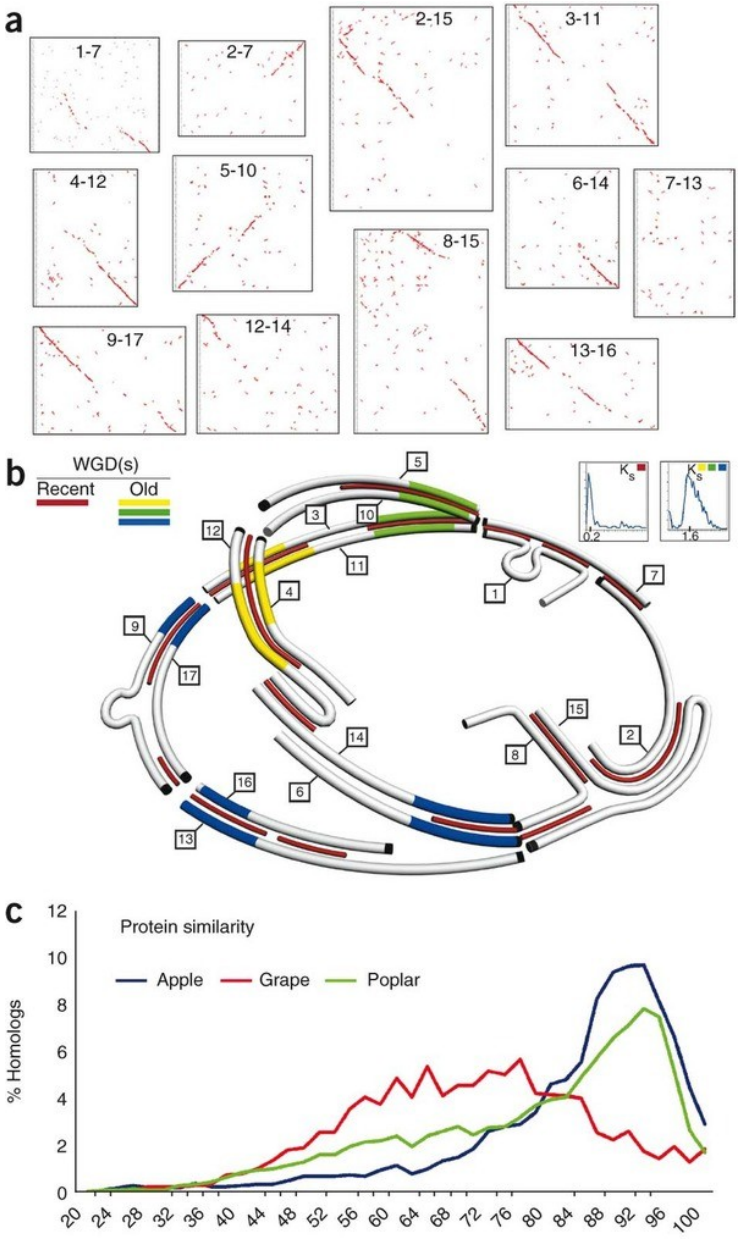
NATURE GENETICS | ARTICLE

The genome of the domesticated apple (*Malus* × *domestica* Borkh.)

Riccardo Velasco, Andrey Zharkikh, Jason Affourtit, Amit Dhingra, Alessandro Cestaro, Ananth Kalyanaraman, Paolo Fontana, Satish K Bhatnagar, Michela Troggio, Dmitry Pruss, Silvio Salvi, Massimo Pindo, Paolo Baldi, Sara Castelletti, Marina Cavaiuolo, Giuseppina Coppola, Fabrizio Costa, Valentina Cova, Antonio Dal Ri, Vadim Goremykin, Matteo Komjanc, Sara Longhi, Pierluigi Magnago, Giulia Malacarne, Mickael Malnoy + *et al.*

Affiliations | Contributions | Corresponding author

Nature Genetics **42**, 833–839 (2010) | doi:10.1038/ng.654
Received 19 November 2009 | Accepted 03 August 2010 | Published online 29 August 2010

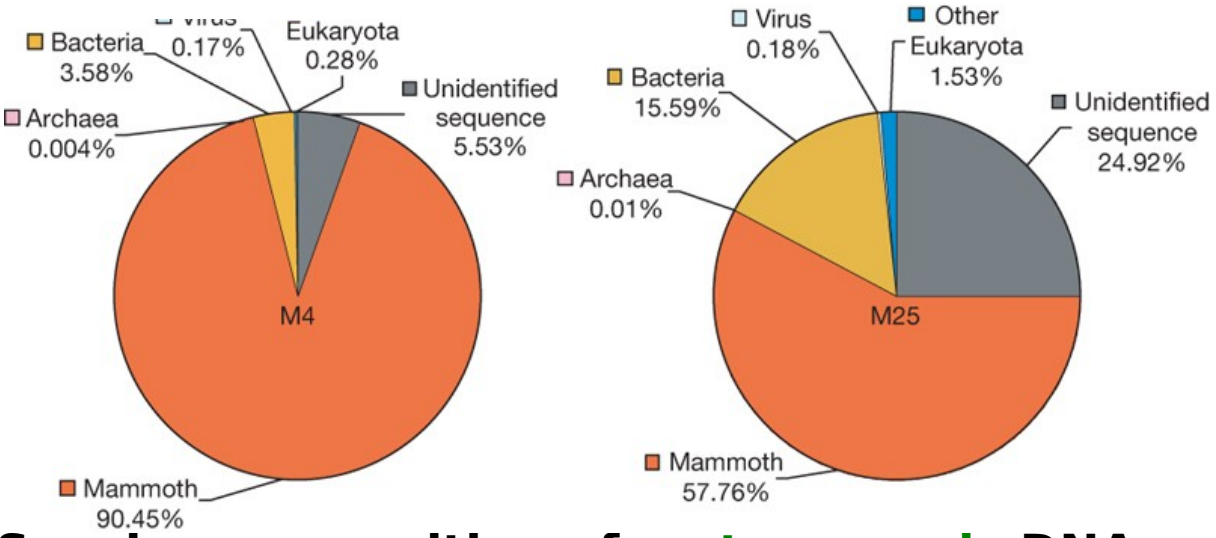


Letter

Nature **456**, 387-390 (20 November 2008) | doi:10.1038/nature07446; Received 12 May 2008; Accepted 22 September 2008

Sequencing the nuclear genome of the extinct woolly mammoth

Webb Miller[1], Daniela I. Drautz[1], Aakrosh Ratan[1], Barbara Pusey[1], Ji Qi[1], Arthur M. Lesk[1], Lynn P. Tomsho[1], Michael D. Packard[1], Fangqing Zhao[1], Andrei Sher[2,9], Alexei Tikhonov[3], Brian Raney[4], Nick Patterson[5], Kerstin Lindblad-Toh[5], Eric S. Lander[5], James R. Knight[6], Gerard P. Irzyk[6], Karin M. Fredrikson[7], Timothy T. Harkins[7], Sharon Sheridan[7], Tom Pringle[8] & Stephan C. Schuster[1]



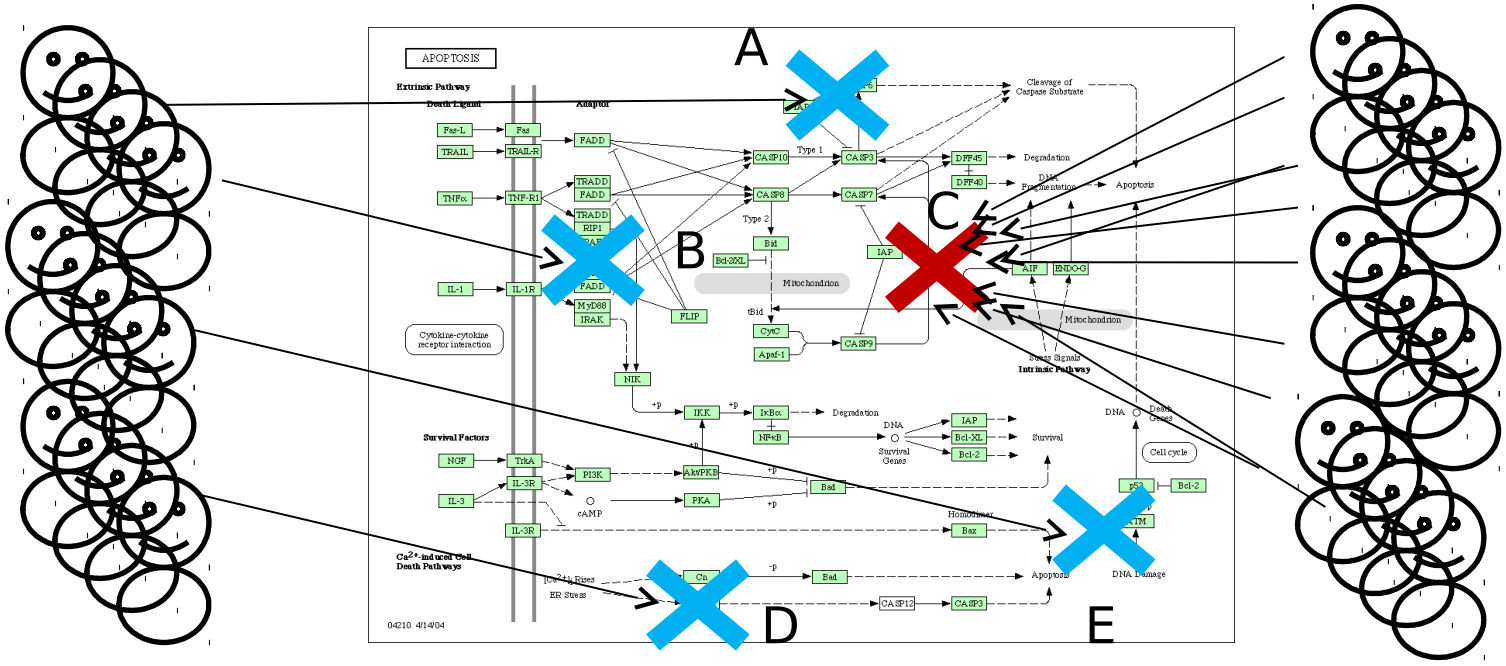
Species composition of metagenomic DNA extracted from mammoth hair

nature

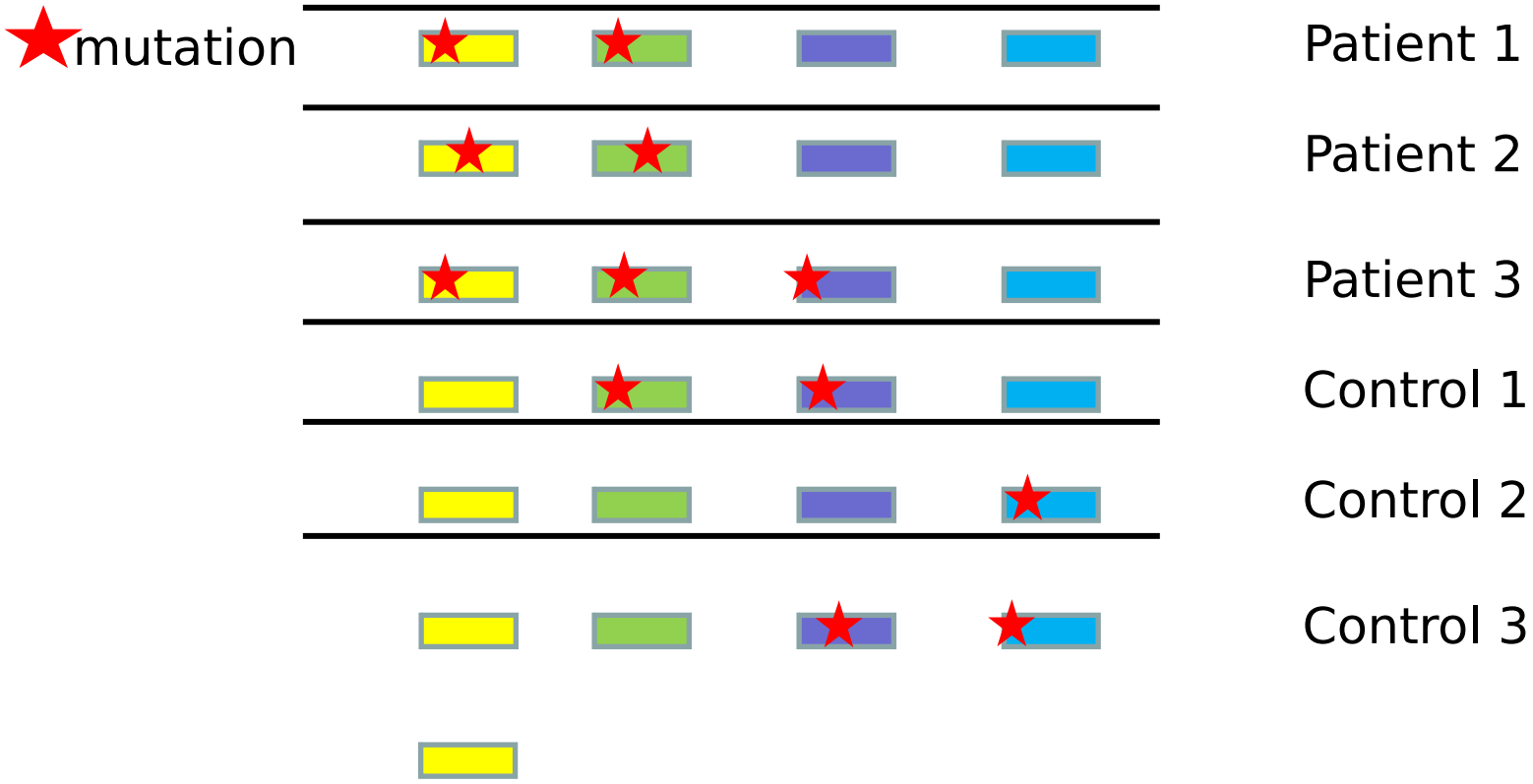
Not that easy, some challenges

Secondary analysis:
Finding the mutations causative of diseases

The simplest case: monogenic disease due to a single gene



The principle: comparison of patients (or families) and reference controls



candidate gene (shares mutation for all patients but no controls)

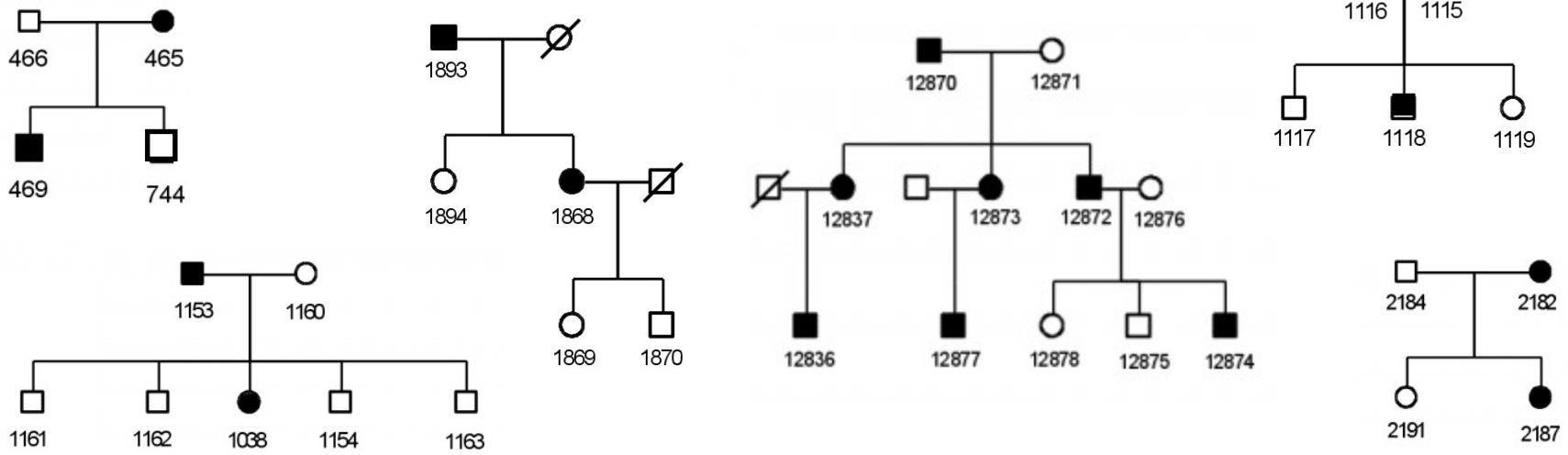
Is this approach realistic? Can we detect such rare variants so easily?

a) Interrogating 50Mb produces too many variants

b) In many cases we are not hunting new but known variants

c) Same phenotype can be due to different mutations and different genes

Filtering with multiple family information



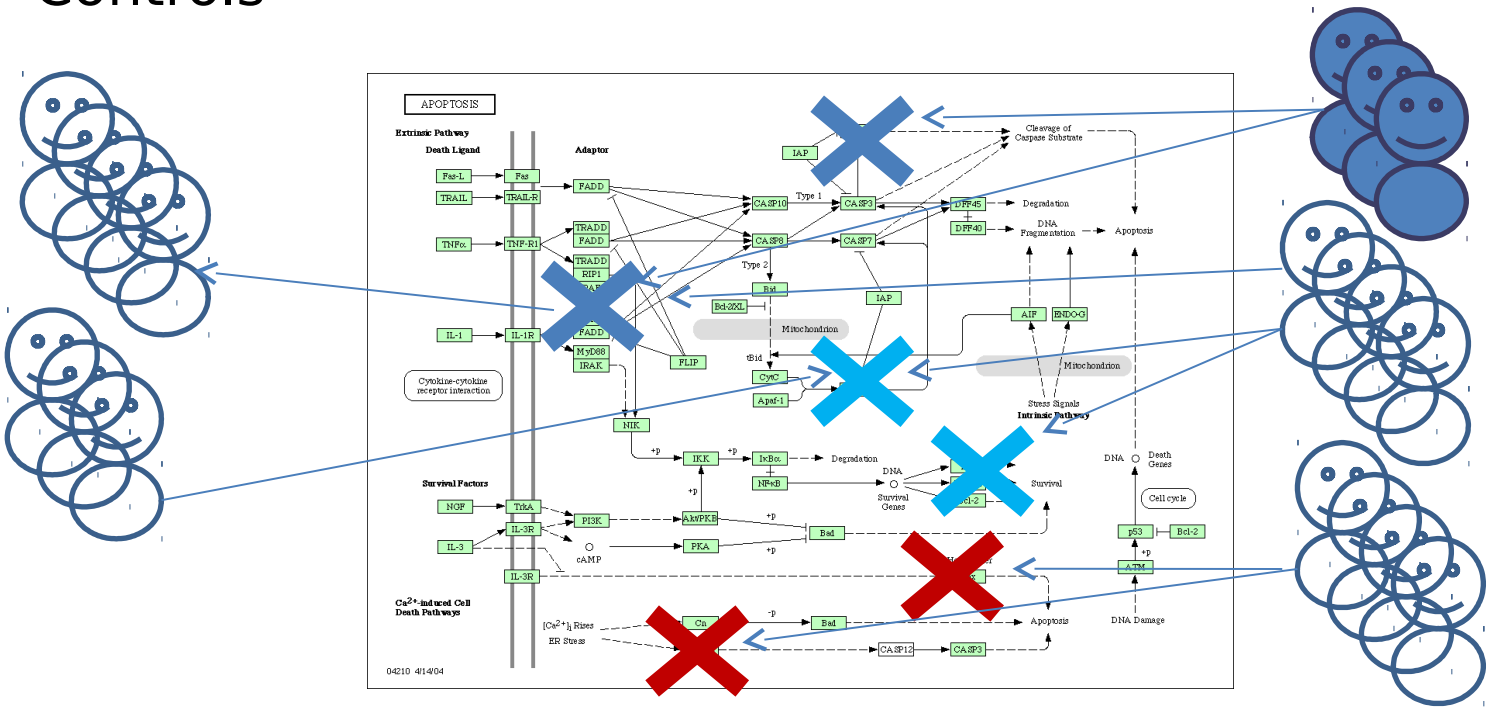
	Families					
	1	2	3	4	5	6
Variants	3403	82	4	0	0	0
Genes	2560	331	35	8	1	0

Problem: how to prioritize putative candidate genes

Clear individual gene associations are difficult to find in some diseases

Controls

Cases



They can have different mutations (or combinations).

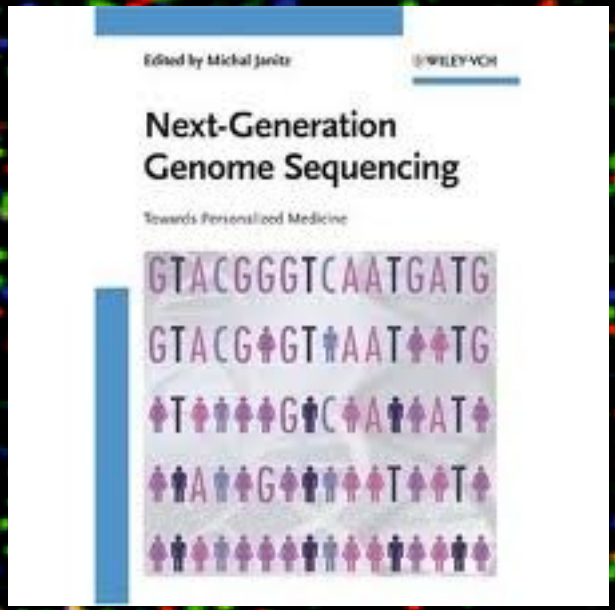
Many cases have to be used to obtain significant associations to many markers.

The only common element is the pathway (yet unknow) affected.

Conclusions

NGS is revolutionizing how we do genome research

But it will also revolutionize our lives….



If we manage to process and analyze ALL the DATA