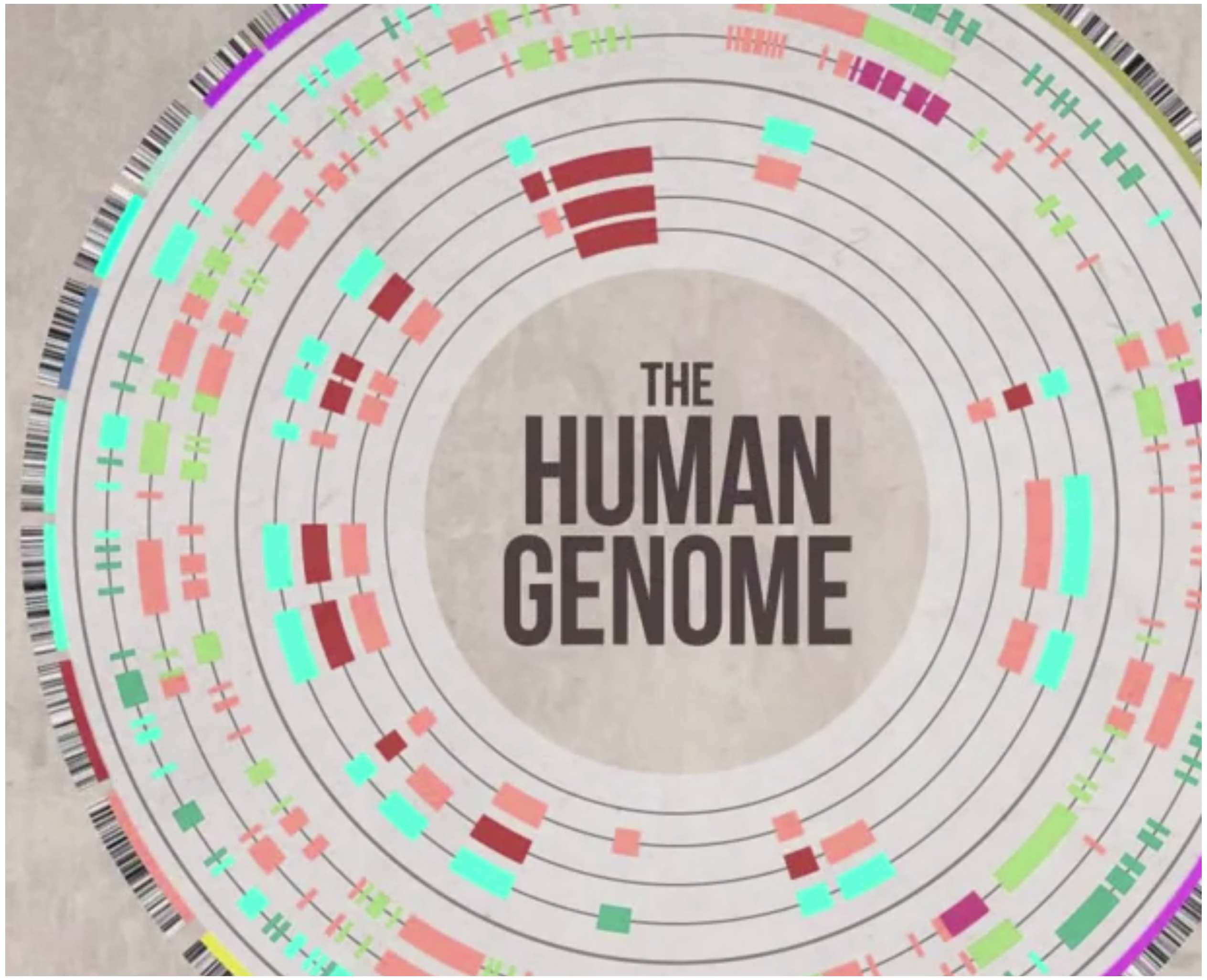
# Genome sequencing story

The sequencing of Human Reference genome provided a roadmap that is the foundation for modern biomedical research.

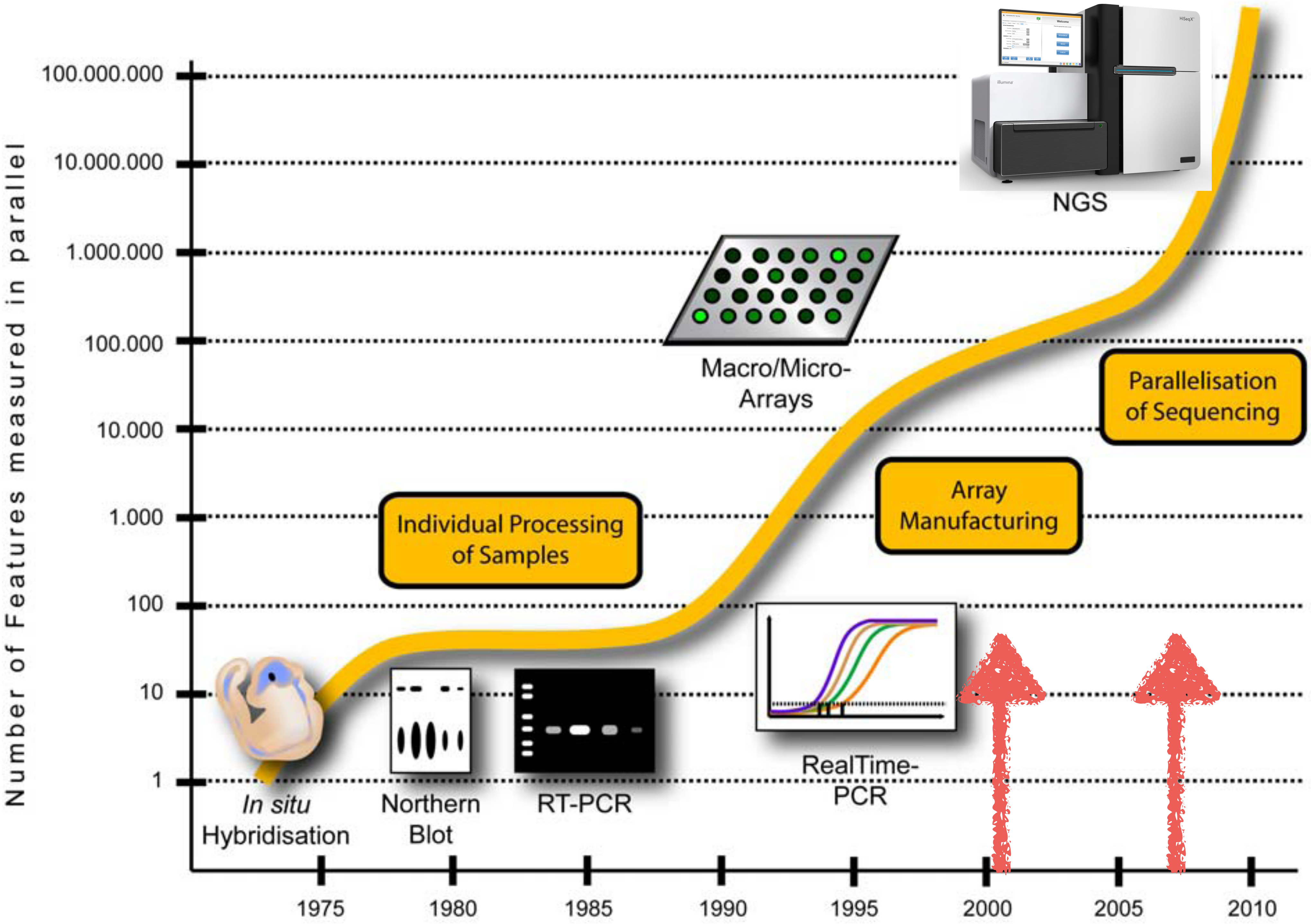


Science
THE
HUMAN
GENOME
AMERICAN ASSOCIATION FOR THE ADVANCEMENT
nature
the
human
genome
Nuclear fission
Five-dimensional energy landscapes
Seafloor spreading
The view from under the Arctic ice
Career prospects
Sequence creates new opportunities
naturejobs
genomics special

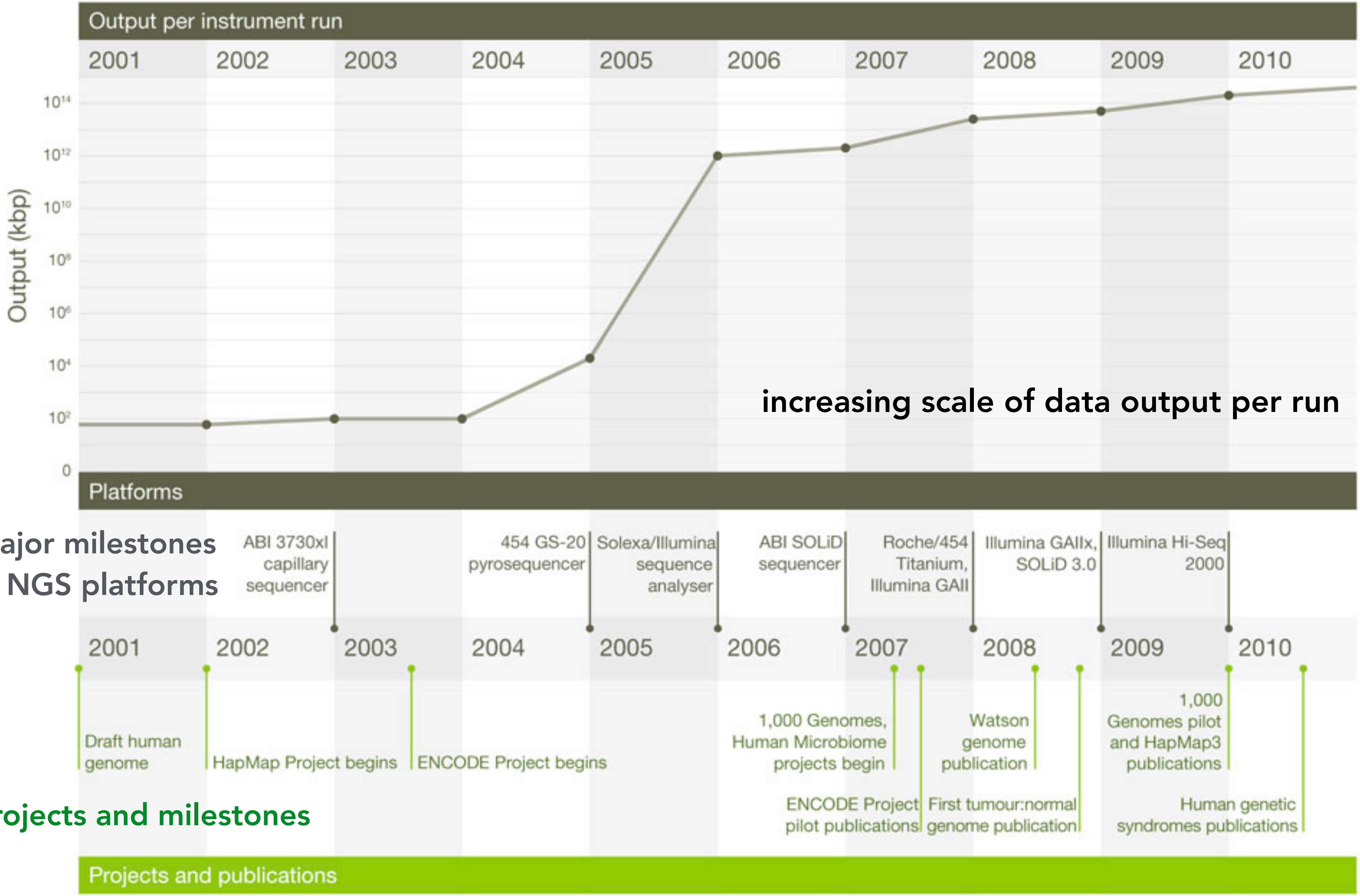
The technology that sequences the human genome was based on capillary electrophoresis of individual fluorescently labelled Sanger sequencing reaction.

500-600 bases from 96 reactions in 10 hours
24-hour reactions = 115 Kbp

Moving forward in the genomic era, the Next generation DNA sequencing technology is enable a revolutionary advances



# Genome sequencing story



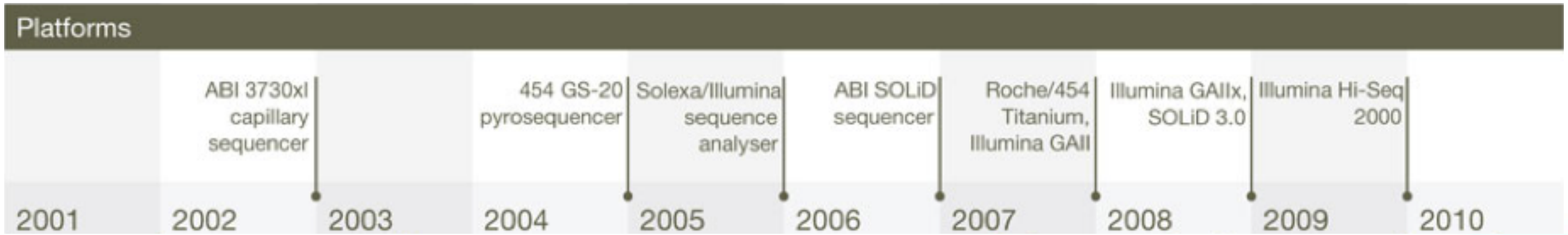
increasing scale of data output per run

major milestones in NGS platforms

projects and milestones

Mardis E.R., Nature 2012

# NGS platforms



Platforms

ABI 3730xl capillary sequencer

454 GS-20 pyrosequencer

Solexa/Illumina sequence analyser

ABI SOLiD sequencer

Roche/454 Titanium, Illumina GAII

Illumina GAIIx, SOLiD 3.0

Illumina Hi-Seq 2000

2001 2002 2003 2004 2005 2006 2007 2008 2009 2010

Each NGS instrument is distinctly different on its specifics

Shared attributes

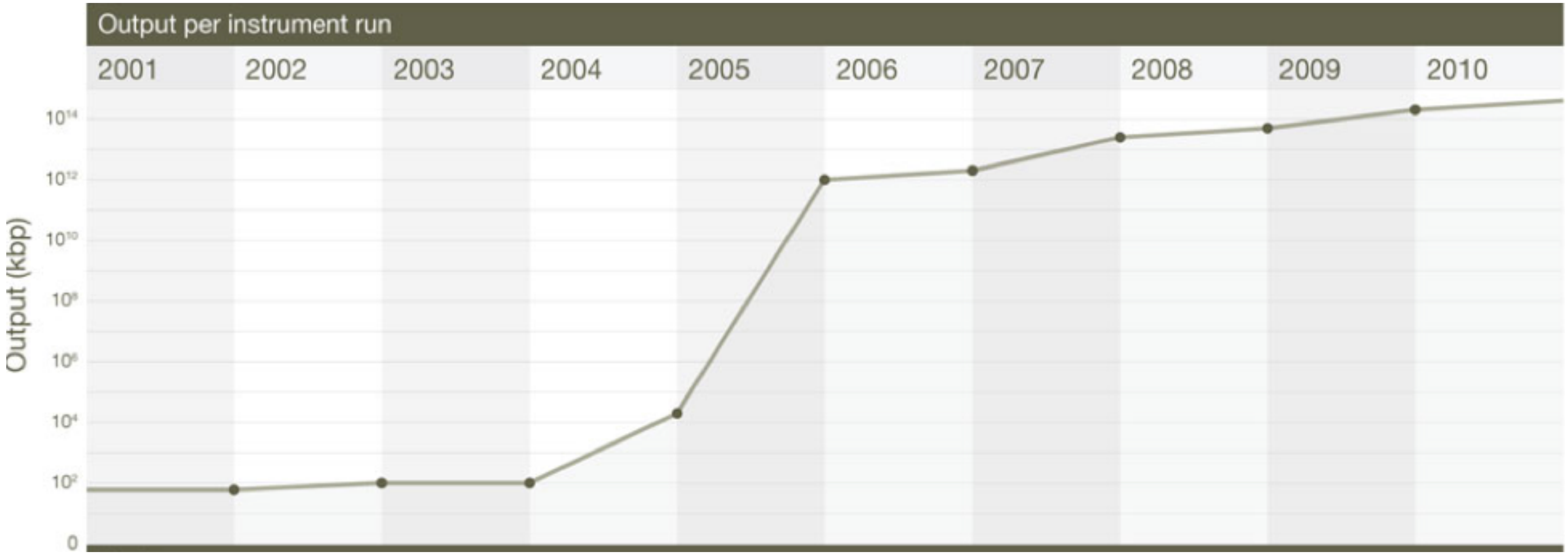
Library preparation step platform-specific adapters are ligated to the fragment to be sequenced

Amplification library fragments are amplified on a solid surface

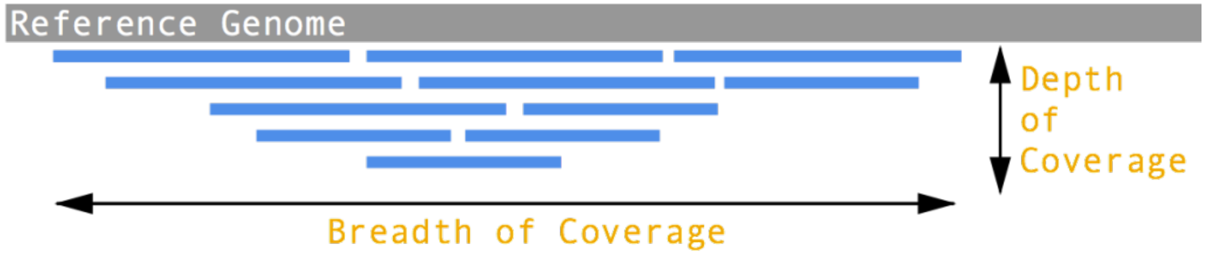
Sequencing reactions series of repeating steps that are performed to detect automatically the nucleotides

Data available it is possible obtain the sequencing information from both the ends of the fragments

# NGS main features

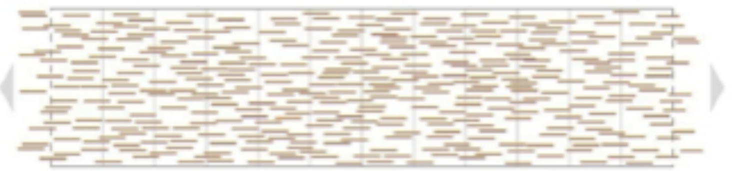


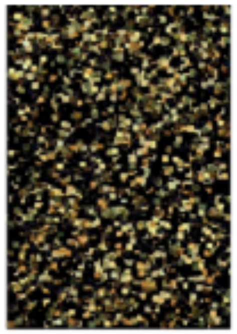
## Coverage



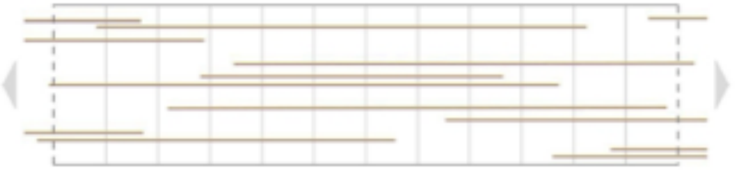
## Reads

Short reads

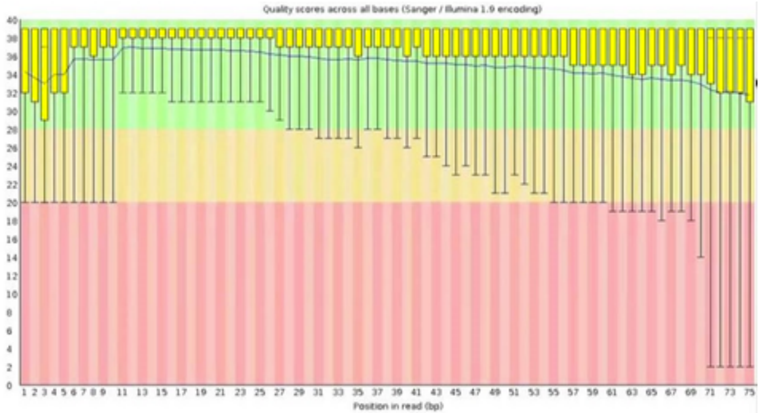




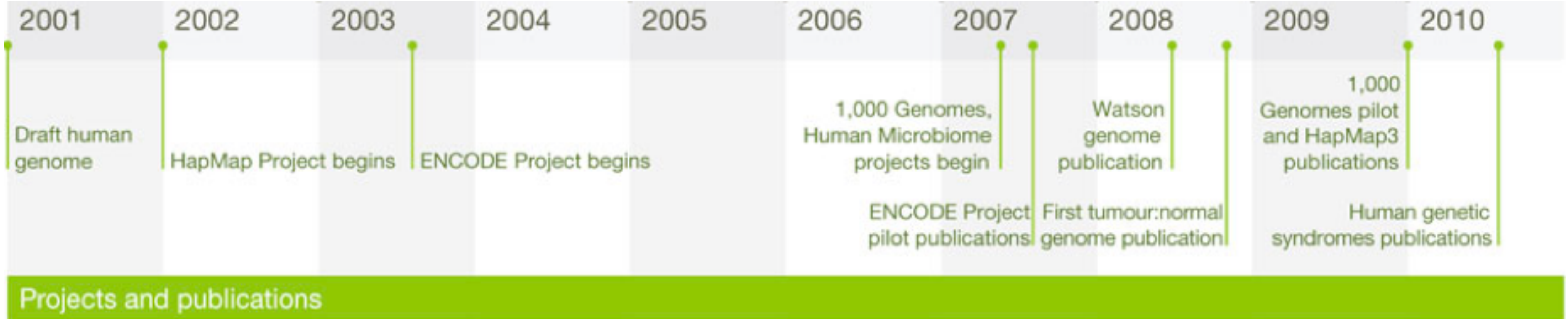
Long reads



## Base calling accuracy



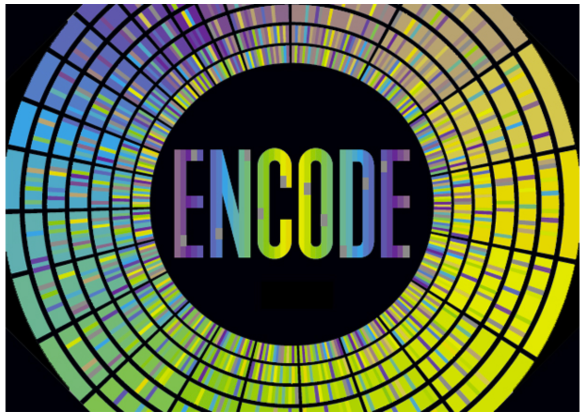
# NGS projects

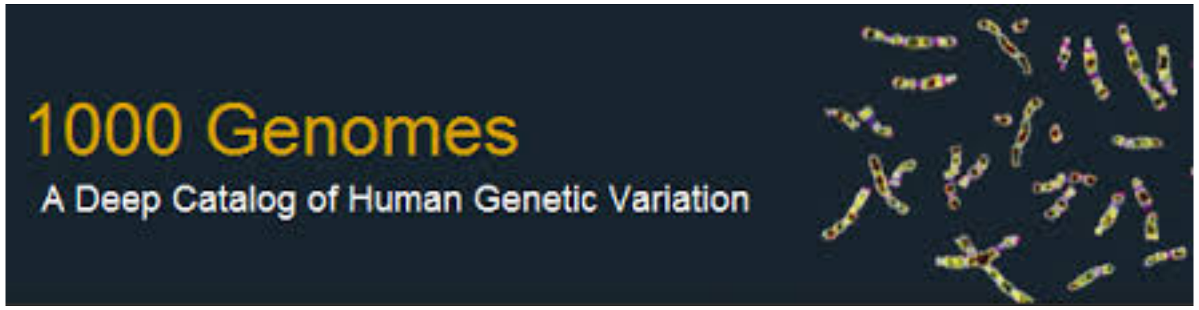

2001
2002
2003
2004
2005
2006
2007
2008
2009
2010
Draft human genome
HapMap Project begins
ENCODE Project begins
1,000 Genomes, Human Microbiome projects begin
ENCODE Project pilot publications
Watson genome publication
First tumour:normal genome publication
1,000 Genomes pilot and HapMap3 publications
Human genetic syndromes publications
Projects and publications

International project for SNP discovery to map the haplotype diversity in the human genome.

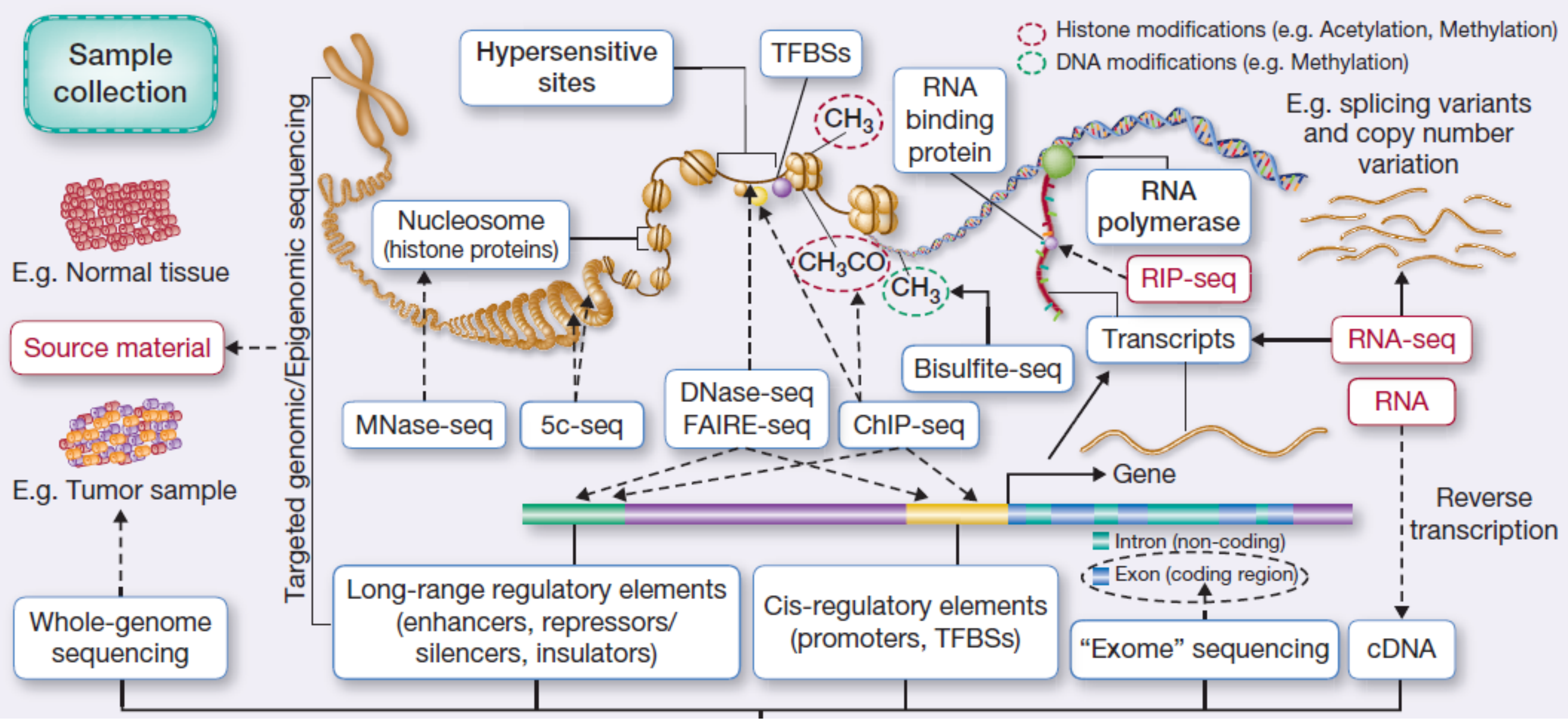
Identification of common SNP variation in multiple human populations.

Genome-wide characterisation for placement of regulatory DNA-binding proteins on genomic DNA, genome wide methylation and histone modification.



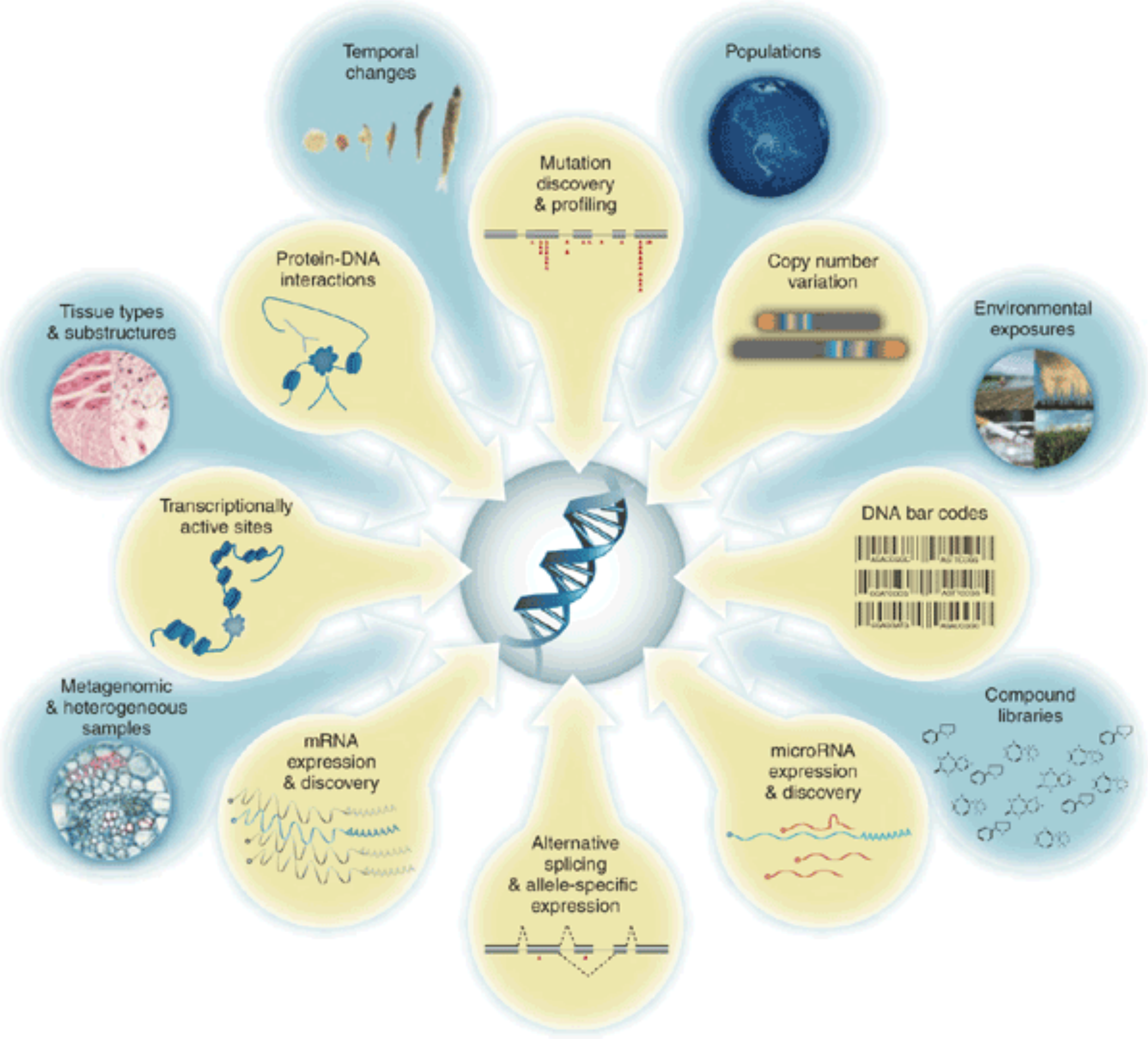


# Knowledge and sample sources

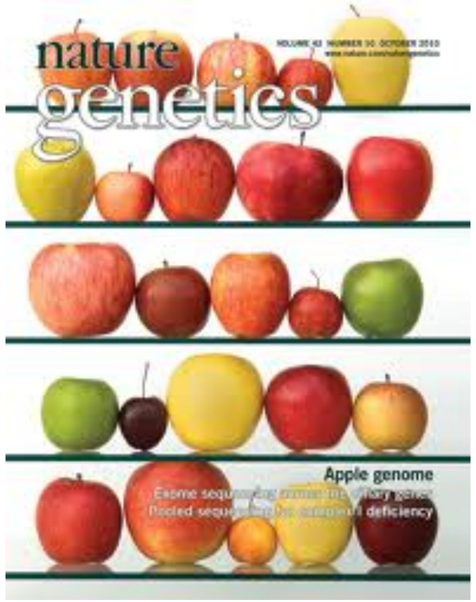


Ritchie MD. et al.,Nat Rev Genet. 2015

# Knowledge and sample sources


Temporal changes
Populations
Mutation discovery & profiling
Protein-DNA interactions
Copy number variation
Tissue types & substructures
Environmental exposures
Transcriptionally active sites
DNA bar codes
Metagenomic & heterogeneous samples
Compound libraries
mRNA expression & discovery
microRNA expression & discovery
Alternative splicing & allele-specific expression

Kahvejian A., et al. Nature Biotech 2008



# Statistical Analysis of Sequence

# Pattern discovery dimensions

- Type of learning
  - from positive examples only (unsupervised)
  - from both positive and negative examples (supervised)
  - noisy data
- Type of patterns
  - deterministic, rigid, flexible, profiles,
- Measure of statistical significance
- A priori knowledge

# Pattern discovery: TFBSs

A major challenge of understanding transcriptional control in higher eukaryotes is the incomplete catalog of regulatory elements, particularly long-range regulatory elements, such as enhancers and insulators.

Firstly, it is necessary identify the **Transcriptional Factor Biding Sites** (TFBSs).

# TFBSs Features

1. TFBSs occur several times in the same genome
2. TFBSs are evolutionarily conserved among different species
3. TFBSs are short
4. TFBSs can have some variation without loss of function

Therefore, most motifs are also found as random hits throughout the genome, and it is a challenging problem to distinguish between false positive hits and true positive binding sites.

# The gold bug problem

Given a secret message:

53++!305))6*;4826)4+.)4+);806*;48!8‘60))85;]8*:+*8!83(88)5*!;
46(;88*96*?;8)*+(;485);5*!2:*+(;4956*2(5*-4)8‘8*; 4069285);)6
!8)4++;1(+9;48081;8:8+1;48!85;4)485!528806*81(+9;48;(88;4(+?3
4;48)4+;161;:188;+?;

Decipher the message encrypted in the fragment, hints: The encrypted message is in English, each symbol correspond to one letter in the English alphabet, no punctuation marks are encoded

# The gold bug problem

Naive approach to solving the problem:

- Count the frequency of each symbol in the encrypted message
- Find the frequency of each letter in the alphabet in the English language
- Compare the frequencies of the previous steps, try to find a correlation and map the symbols to a letter in the alphabet

Most frequent

Less frequent

**e t a o i n s r h l d c u m f p g w y b v k x j q z**

# The gold bug problem

By simply mapping the most frequent symbols to the most frequent letters of the alphabet:

sfiilfcsoorntaeuroaikoaiotecrntaeleyrcooestvenpinelefheeosnlt
arhteenmrnwteonihtaesotsnlupnihtamsrnuhsnbaoeyentacrmuesotorl
eoaiitdhimtaecedtepeidtaelestaoaeslsueecrnedhimtaetheetahiwfa
taeoaitdrdtpdeetiwt

The result does not make sense

# The gold bug problem

A better approach:

- Examine frequencies of l-tuples, combinations of 2 symbols, 3 symbols, etc.
- *The* is the most frequent 3-tuple in English and *;48* is the most frequent 3-tuple in the encrypted text
- Make inferences of unknown symbols by examining other frequent l-tuples

## The gold bug problem

Mapping the to ;48 and substituting all occurrences of the symbols:

```
53++!305))6*the26)h+.)h+)te06*the!e‘60))e5t]e*:+*e!e3(ee)5*!t
h6(tee*96*?te)*+(the5)t5*!2:*+(th956*2(5*h)e‘e*th0692e5)t)6!e
)h++t1(+9the0e1te:e+1the!e5th)he5!52ee06*e1(+9thet(eeth(+?3ht
he)h+t161t:1eet+?t
```

# The gold bug problem

Make inferences:

53++!305))6*the26)h+.)h+)te06*the!e‘60))e5t]e*:+*e!e3(ee)5*!t
h6(tee*96*?te)*+(the5)t5*!2:*+(th956*2(5*h)e‘e*th0692e5)t)6!e
)h++t1(+9the0e1te:e+1the!e5th)he5!52ee06*e1(+9thet(eeth(+?3ht
he)h+t161t:1eet+?t

- *thet(ee* most likely means *the tree*  
  Infer *(* = *r*
- *th(+?3h* becomes *through*  
  Can we guess + and *?*

# The gold bug problem

The solution:

A GOOD GLASS IN THE BISHOPS HOSTEL IN THE DEVILS SEA, TWENY ONE DEGREES AND THIRTEEN MINUTES NORTHEAST AND BY NORTH, MAIN BRANCH SEVENTH LIMB, EAST SIDE, SHOOT FROM THE LEFT EYE OF THE DEATHS HEAD A BEE LINE FROM THE TREE THROUGH THE SHOT, FIFTY FEET OUT.

# The gold bug problem

Prerequisites to solve the problem:

- Need to know the relative frequencies of single letters, and combinations of two and three letters in English

- Knowledge of all the words in the English dictionary is highly desired to make accurate inferences

# The gold bug problem and Motif finding . Similarities

- Nucleotides in motifs encode for a message in the genetic language. Symbols in The Gold Bug encode for a message in English
- In order to solve the problem, we analyze the frequencies of patterns in DNA/Gold Bug message.
- Knowledge of established regulatory motifs makes the Motif Finding problem simpler. Knowledge of the words in the English dictionary helps to solve the Gold Bug problem.

# The gold bug problem and Motif finding . Similarities

- Motif Finding:
  1. In order to solve the problem, we analyze the frequencies of patterns in the nucleotide sequences
  2. Knowledge of established motifs reduces the complexity of the problem
- Gold Bug Problem:
  1. In order to solve the problem, we analyze the frequencies of patterns in the text written in English
  2. Knowledge of the words in the dictionary is highly desirable

# The gold bug problem and Motif finding . Differences

Motif Finding is harder than Gold Bug problem:

- We do not have the complete dictionary of motifs
- The *genetic* language does not have a standard *grammar*
- Only a small fraction of nucleotide sequences encode for motifs; the size of data is enormous

# 1) Deterministic Patterns

- Definition: Deterministic patterns are substrings over the alphabet $\sum$ e.g., **TATAAA** (TATA-box consensus)
- Discovery algorithms are faster on these types of patterns
- Usually not flexible enough for the needs of molecular biology

# 2) Rigid patterns

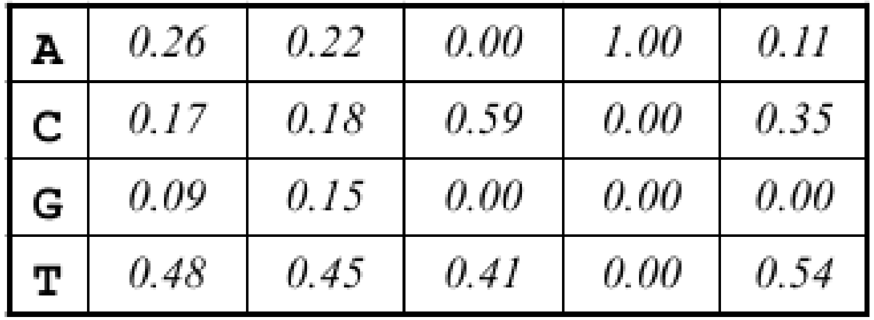
- Definition: Rigid patterns are patterns which allow substitutions/ don't care symbols
  - e.g., the patterns under IUPAC alphabet **A,C,G,T,U,M,R,W,S,Y,K,V,H,D,B,X,N** where for example R=[A—G], Y=[C—T], etc.,
  - e.g, **ARNNTTYGA** under IUPAC means A[A—G] [A—C—G—T][A—C—G—T]TT[C—T]GA
- Note that the size of the pattern is not allowed to change

# 3) Flexible patterns

- Definition: Flexible patterns are patterns which allow substitutions/ don t care symbols and variable-length gaps
  e.g., Prosite **F-x(5)-G-x(2,4)-G-*-H**
- Note that the length of these pattern is variable
- Very expressive
- Space of all patterns is huge

# 4) Profiles

- *Position weight matrices*, or profiles, are matrices containing real numbers in the interval $[0, 1]$, such that each column sums to 1



| | | | | | |
|---|---|---|---|---|---|
| **A** | *0.26* | *0.22* | *0.00* | *1.00* | *0.11* |
| **C** | *0.17* | *0.18* | *0.59* | *0.00* | *0.35* |
| **G** | *0.09* | *0.15* | *0.00* | *0.00* | *0.00* |
| **T** | *0.48* | *0.45* | *0.41* | *0.00* | *0.54* |

- Notion of *consensus*