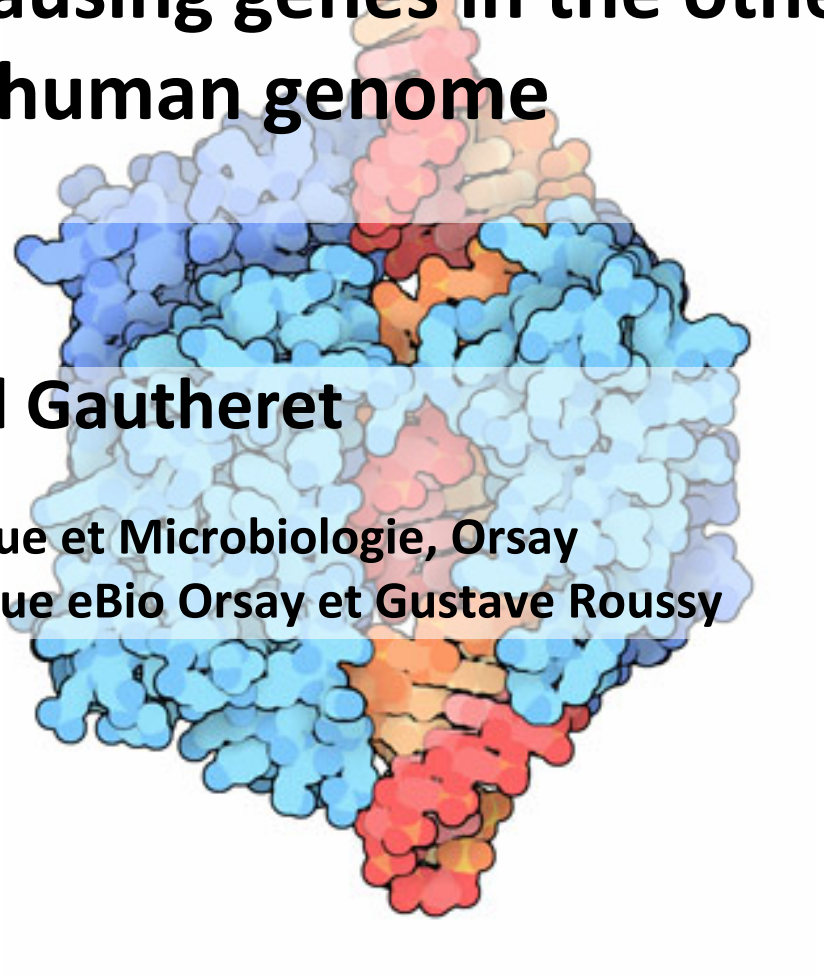
Mining for disease-causing genes in the other 98% of the human genome



Daniel Gautheret

Institut de Génétique et Microbiologie, Orsay
Plateforme Bioinformatique eBio Orsay et Gustave Roussy

UNIVERSITÉ PARIS SUD

Gene expression

RNA polymerase

Transcription factor

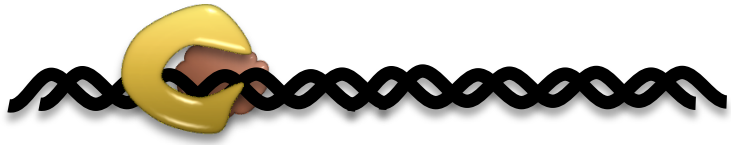
DNA

Gene expression

RNA polymerase

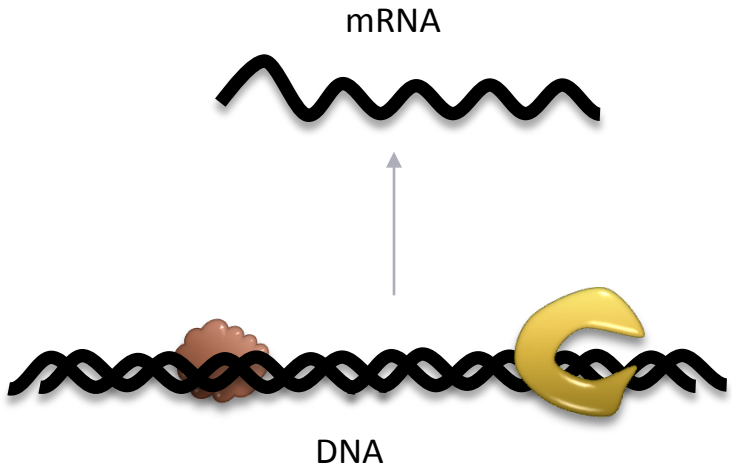
DNA

Gene expression



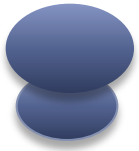
DNA

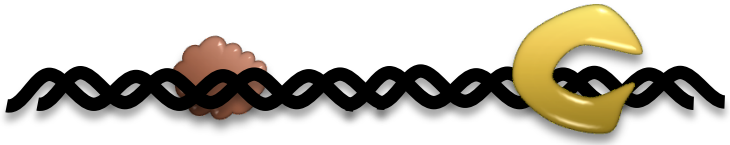
Gene expression



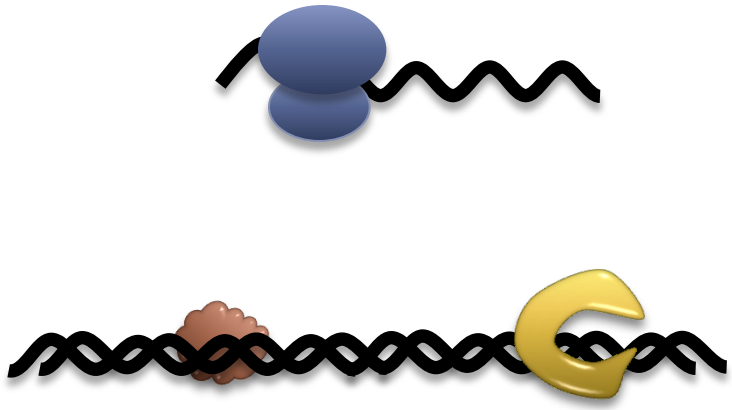
Gene expression

Ribosome

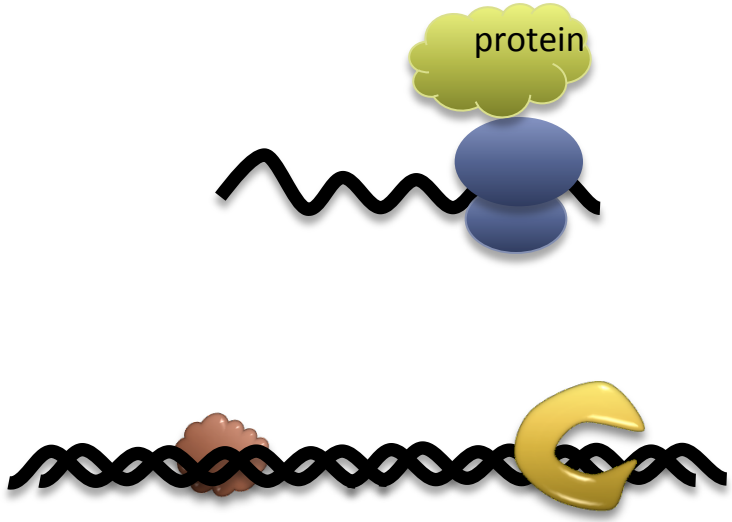




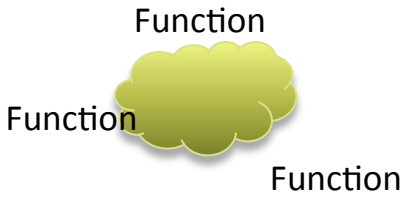
Gene expression

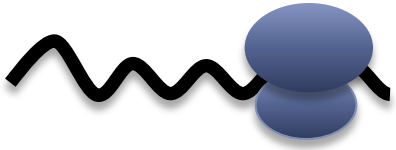


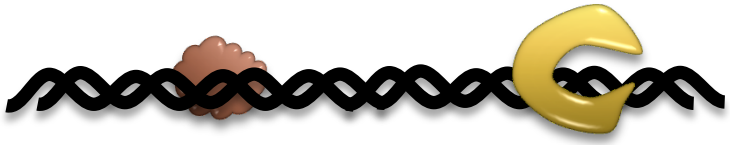
Gene expression



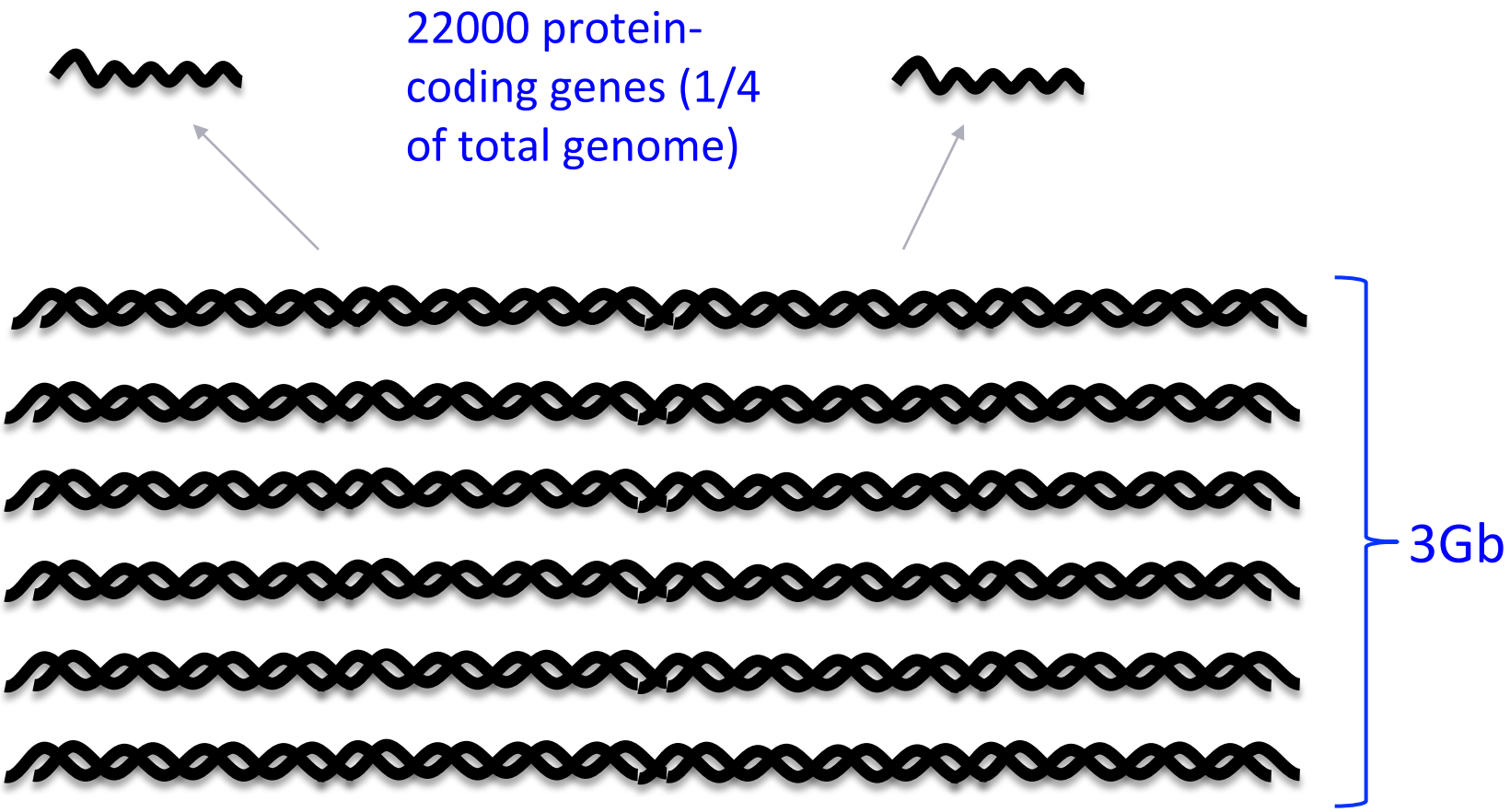
Gene expression





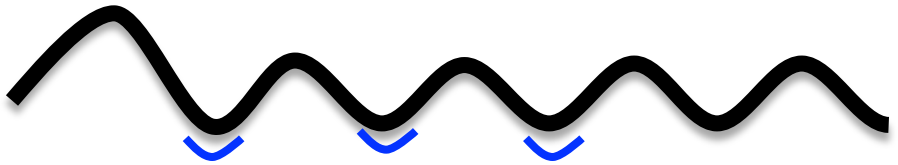


2001: Human genome has surprisingly few protein-coding genes

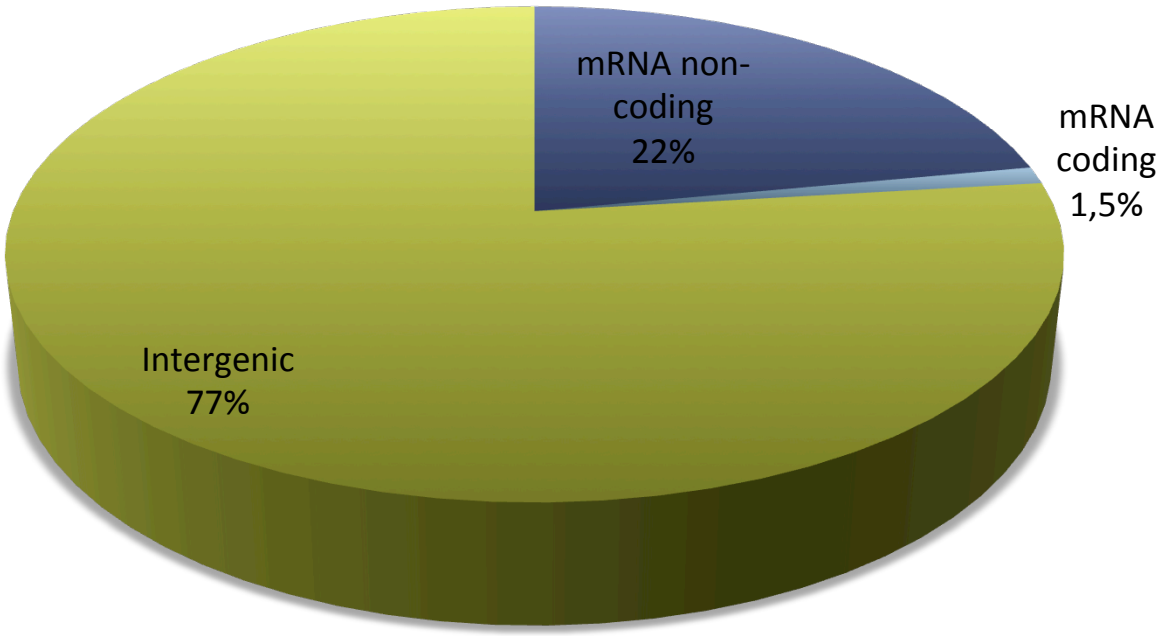


Animal/plant genes are mostly non-coding

Human gene is
5% coding



Human genome : 98% non-coding



« Pervasive » transcription…

Identification and analysis of functional elements in 1% of the human genome by the ENCODE pilot project

The ENCODE Project Consortium*

Encode 2005

« Remarkably, 93% of bases are represented in a primary transcript identified by at least two independent observations »

ARTICLE

doi:10.1038/nature11233

Landscape of transcription in human cells

Encode 2012

« 75% of the human genome is covered by primary transcripts »

OPEN ACCESS Freely available online

PLOS BIOLOGY

Most "Dark Matter" Transcripts Are Associated With Known Genes

Harm van Bakel[1], Corey Nislow[1,2], Benjamin J. Blencowe[1,2], Timothy R. Hughes[1,2]*

OPEN ACCESS Freely available online

PLOS BIOLOGY

Perspective

The Reality of Pervasive Transcription

Michael B. Clark[1], Paulo P. Amaral[1☯], Felix J. Schlesinger[2☯], Marcel E. Dinger[1], Ryan J. Taft[1], John L. Rinn[3], Chris P. Ponting[4], Peter F. Stadler[5], Kevin V. Morris[6], Antonin Morillon[7], Joel S. Rozowsky[8], Mark B. Gerstein[8], Claes Wahlestedt[9], Yoshihide Hayashizaki[10], Piero Carninci[10], Thomas R. Gingeras[2]*, John S. Mattick[1]*

OPEN ACCESS Freely available online

PLOS BIOLOGY

Perspective

Response to "The Reality of Pervasive Transcription"

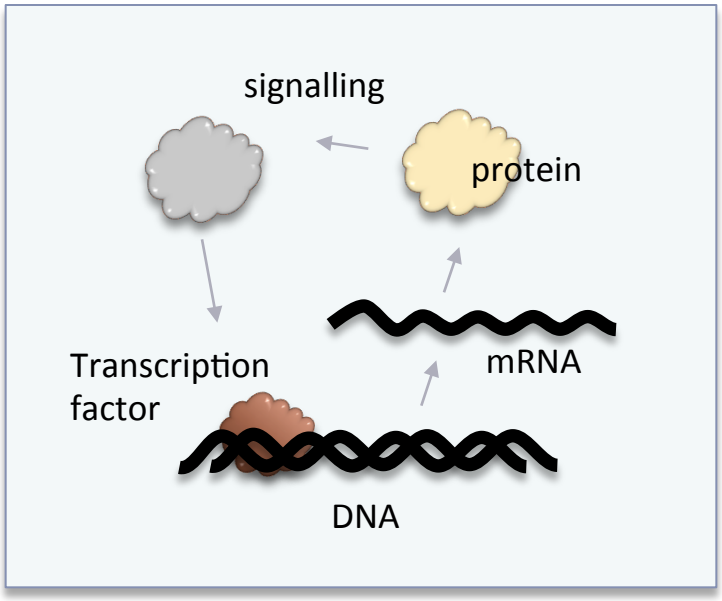
Harm van Bakel[1], Corey Nislow[1,2], Benjamin J. Blencowe[1,2], Timothy R. Hughes[1,2]*

>15000 new « real genes » in the non-coding genome

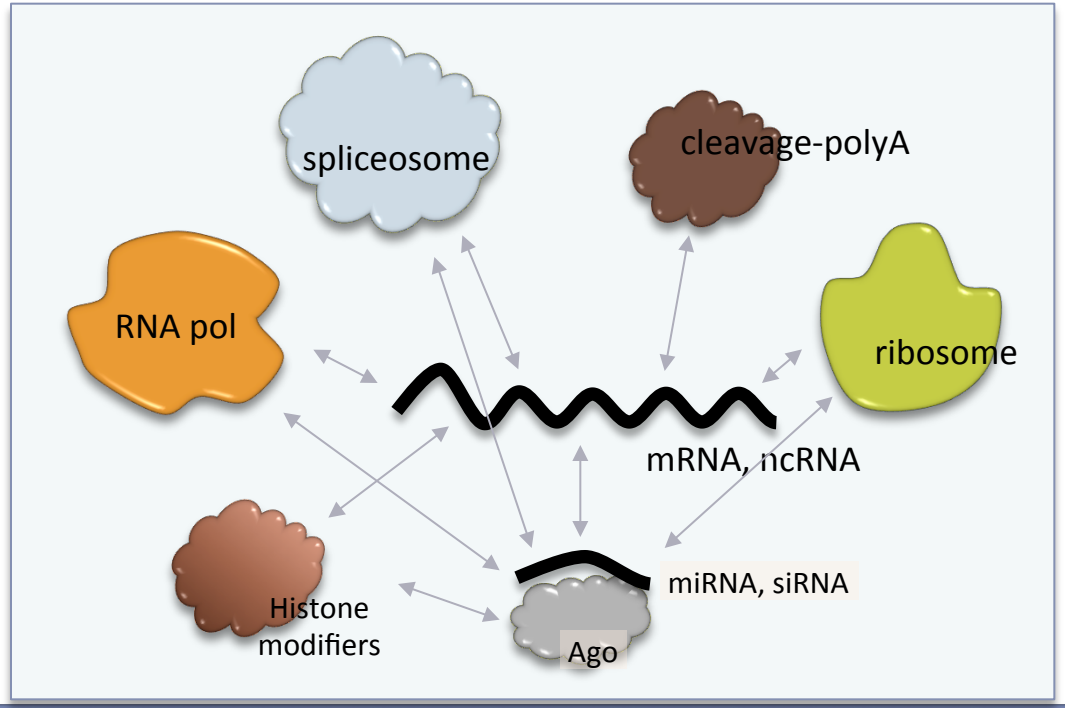
- 13000 long non-coding RNAs (lncRNAs)
  - Do not produce any protein
  - Involved in regulating protein coding genes
- 2000+ microRNAs (miRNAs)
  - Produced by specific machinery
  - Repress distant mRNAs

RNA is at the center of gene expression regulation

Conventional gene network

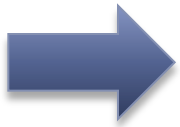


RNA-centric gene network



Cancer genes

- What is a « cancer gene »?
  - A gene whose expression, repression or mutation is causally related to tumor development, progression or metastasis
- Currently: about 500 cancer genes
- >50% are related to gene expression regulation
  - Just like ncRNA genes
  - Yet all known cancer genes are protein coding!



Our lab mission is to discover non-coding cancer driver genes

How to spot a disease gene or biomarker?

1) Population-based

General population → Extract DNA/ RNA from organ

Disease population → Extract DNA/ RNA from organ

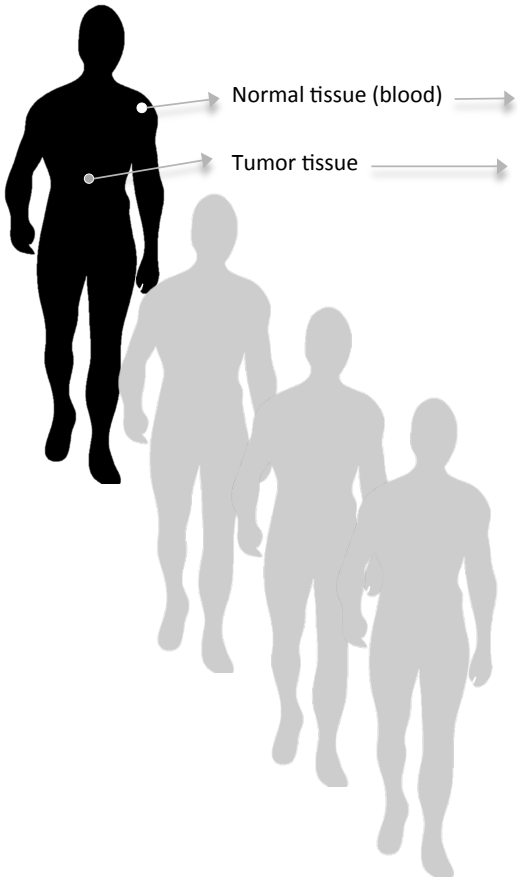
compare

RNA: expression patterns

DNA: mutation patterns

How to spot a disease gene or biomarker?

2) Somatic events



Extract DNA, compare and look for mutations

Repeat for several patients → Look for recurrent events

Next Generation Sequencing

	Illumina GA IIx	Illumina Hi-Seq 2000	Illumina MySeq	Lifetech Ion torrent PGM	Lifetech Ion proton
Read number	8x20M	16x100M	15M	1M	50M
Read size	2x100	2x100	2x250	35-400	35-400

Sequencer capacity & genome coverage

				Coverage		
	reads	**read length**	**total nt**	30Mb (exome)	90Mb (transcriptome)	3Gb (Genome)
Ion torrent 316	1 000 000	100	100 000 000	3	1,1	0,03
Ion proton	50 000 000	200	10 000 000 000	333	111,1	3,33
Hi Seq 8 lanes PE	800 000 000	160	128 000 000 000	4000	1450,0	42,64

Patient genome Data now under production

Ilumina X-ten: 15,000-20,000 genomes per year

Normal DNA +
tumor DNA + RNA

ATCGTACGTAGCTACTGAC
CGACTGATCGTAGCTAGCT
CTAGCTGATCGTAC

ATCGTACGTAGCTACTGAC
CGACTGATCGTAGCTAGCT
CTAGCTGATCGTAC

ATCGTACGTAGCTACTGAC
CGACTGATCGTAGCTAGCT
CTAGCTGATCGTAC

ATCGTACGTAGCTACTGAC
CGACTGATCGTAGCTAGCT
CTAGCTGATCGTAC

ATCGTACGTAGCTACTGAC
CGACTGATCGTAGCTAGCT
CTAGCTGATCGTAC

Clinical data

Tumor type
Grade
Age …

Tumor type
Grade
Age …

Tumor type
Grade
Age …

Tumor type
Grade
Age …

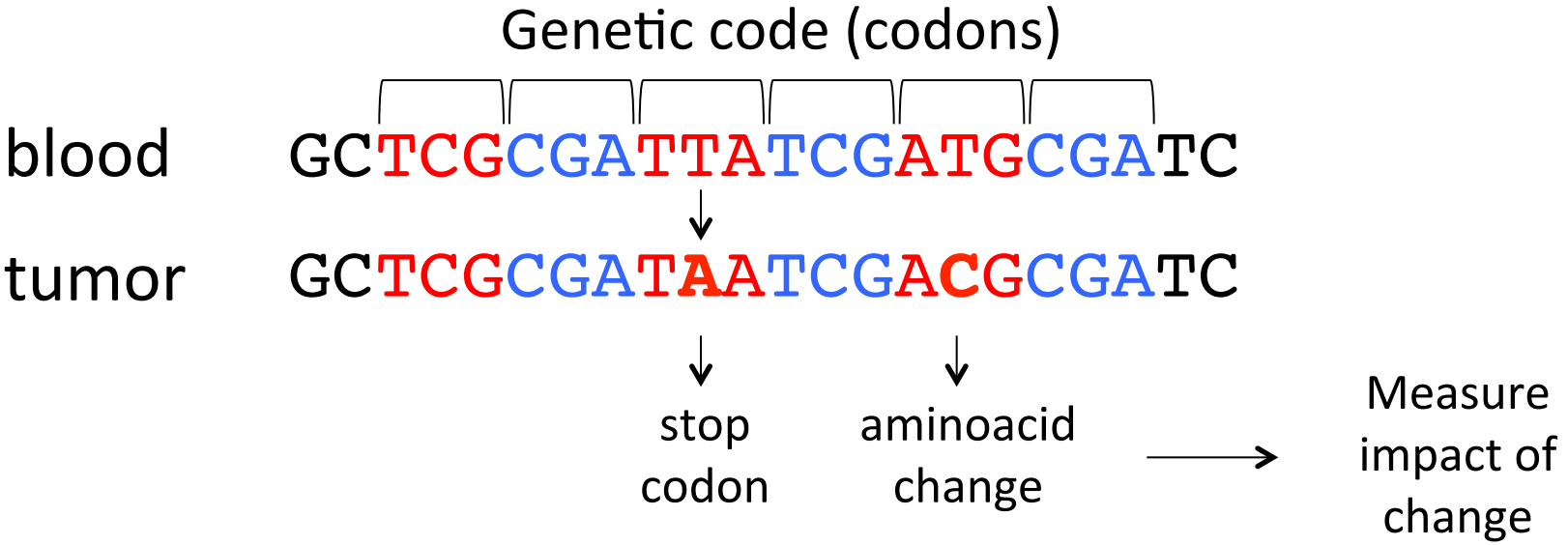
Tumor type
Grade
Age …

Let's assume we have many tumor/normal whole genome sequences. So what?

- There are typically 5M single nucleotide differences between 2 individuals
- 100-10000 differences between a tumor and normal tissue from the same person
- Most of these differences
  - are harmless
  - are located in the 98.5% non-coding part

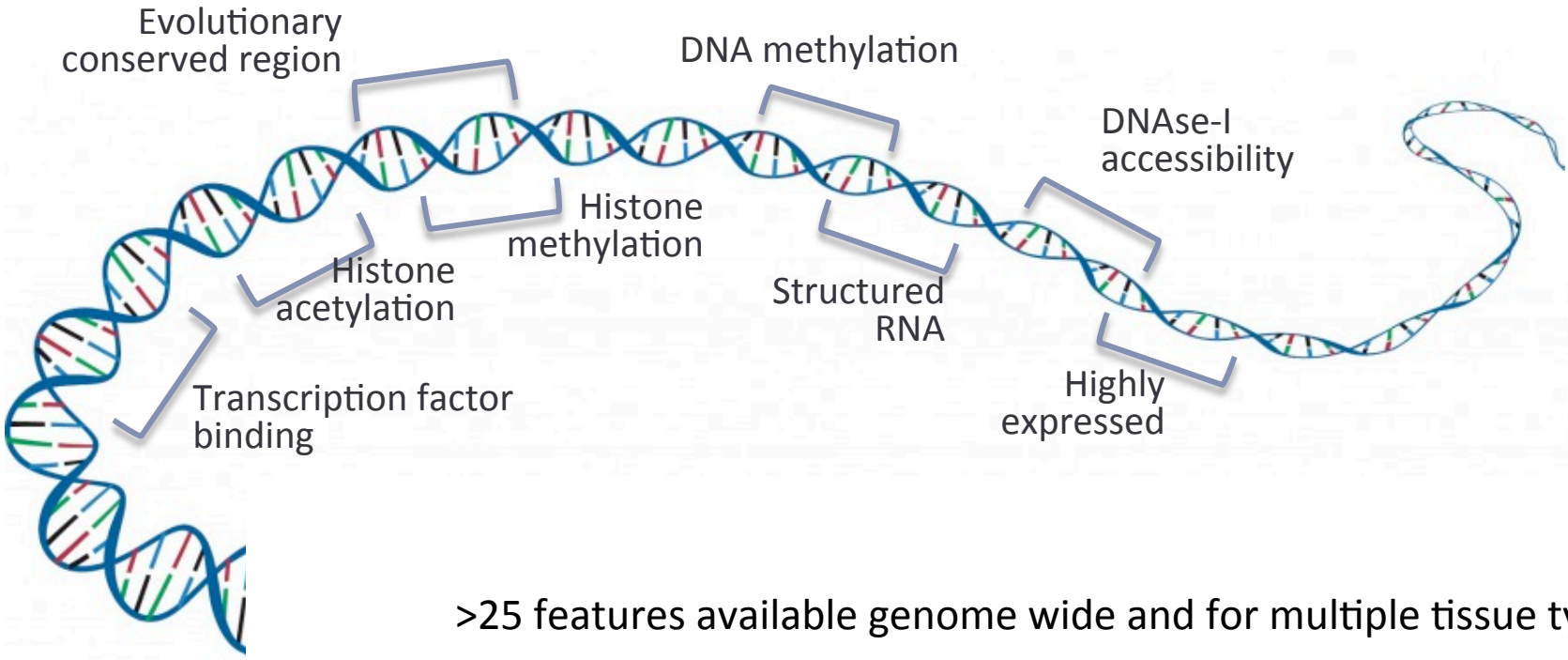
Current approach

- Focus on the 1.5% coding part
- Powerful tools are available to find deleterious mutations



Then what can we say about a mutation in the 98.5% non-coding part?

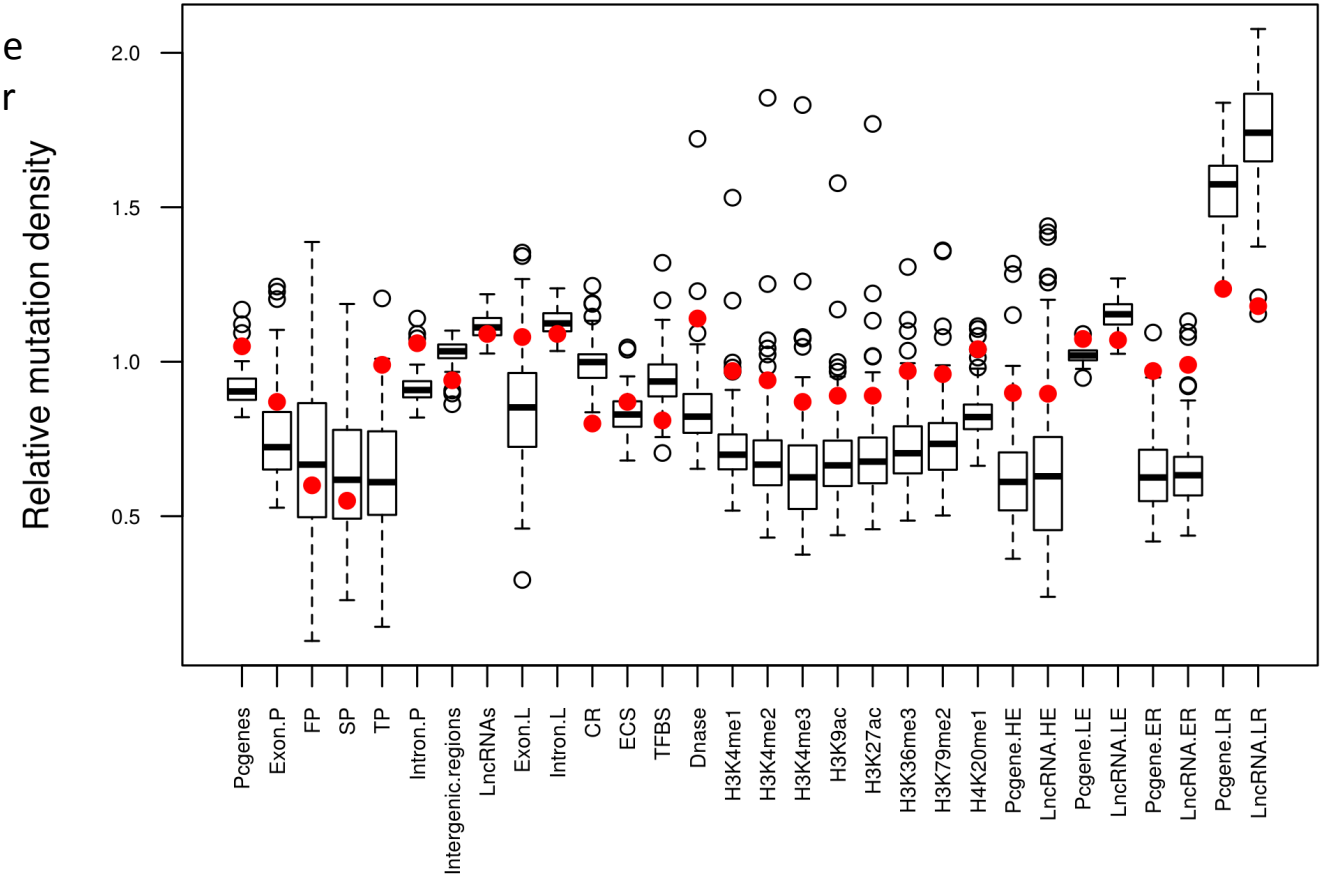
- Genome annotation data is quickly building up



>25 features available genome wide and for multiple tissue types

Mutation density in functional regions

Data from 88 whole genomes from liver cancer



A tumor behaves like a quickly evolving organism that constantly fights selection

Where mutations are rarest they are potentially most harmful

Models for scoring non-coding mutation

- Based on
  - Observed mutation densities in cancer genomes and in general population
  - Genome features
    - Evolutionary
    - Physical
    - Epigenetic
    - etc.
- Modeling techniques
  - Logistic regression
  - Random forests

Challenges

- Computational:
  - Volumes
    - 100-200Gb / sequencing library = 100 Tb/yr for a Cancer Center.
  - CPU
    - Typically 1-3 days computing per sequence library on 50-core node
  - Flexibility!
    - Our pipelines change everyday
  - Data exchange policies and standards

- Human ressources:
  - Serious bioinformatics needs scientists willing to learn both sides of the story