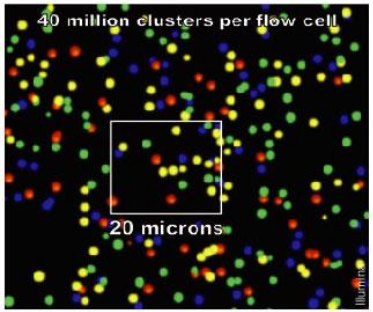
生物信息学：导论与方法
Bioinformatics: Introduction and Methods

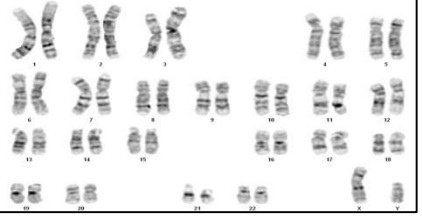
Ge Gao 高歌 & Liping Wei 魏丽萍
Center for Bioinformatics, Peking University

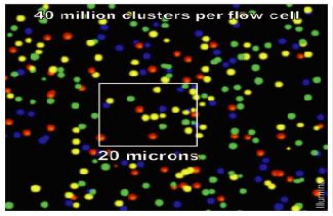
https://www.coursera.org/course/pkubioinfo

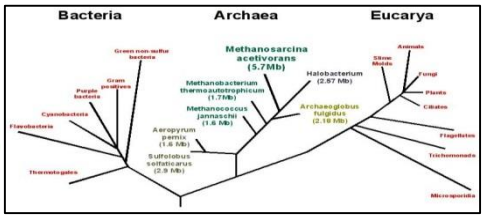
The dawning of the age of personalized medicine

- Next-generation sequencing can sequence one person's whole genome with ~$3000 in a day.
- The personal genomes hold promises for a future of personalized medicine.







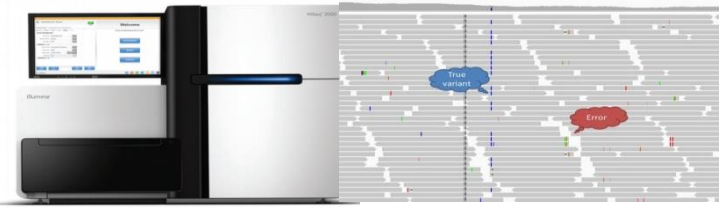


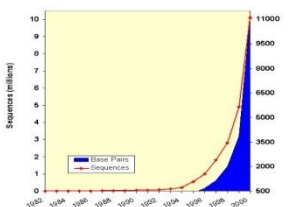
Week 6: Functional prediction of genetic variantions

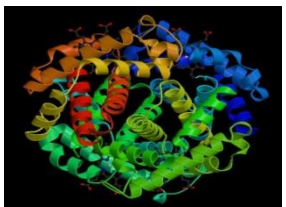
北京大学生物信息学中心 魏丽萍

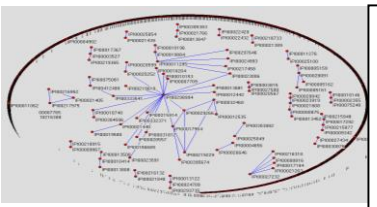
Liping Wei, Ph.D.

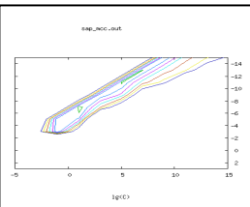
Center for Bioinformatics, Peking University

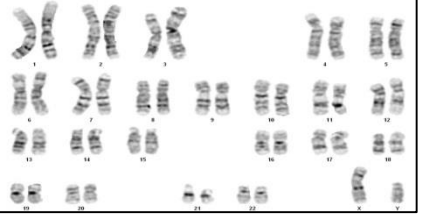


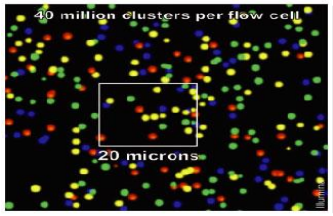


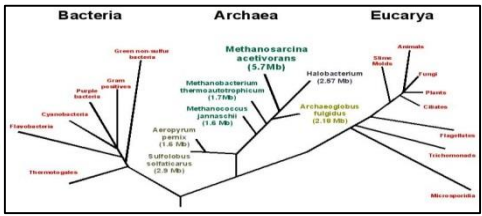










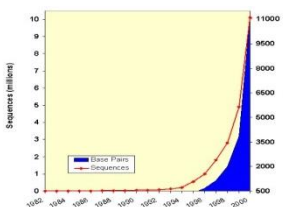


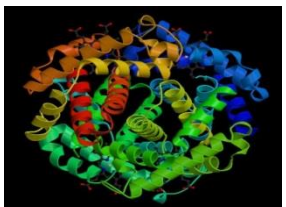
Unit 1: Overview of the problem

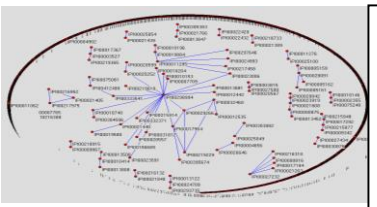
北京大学生物信息学中心 魏丽萍

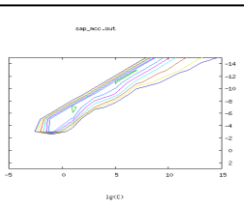
Liping Wei, Ph.D.

Center for Bioinformatics, Peking University









The New York Times

May 14, 2013

My Medical Choice

By ANGELINA JOLIE

LOS ANGELES

MY MOTHER fought cancer for almost a decade and died at 56. She held out long enough to meet the first of her grandchildren and to hold them in her arms. But my other children will never have the chance to know her and experience how loving and gracious she was.

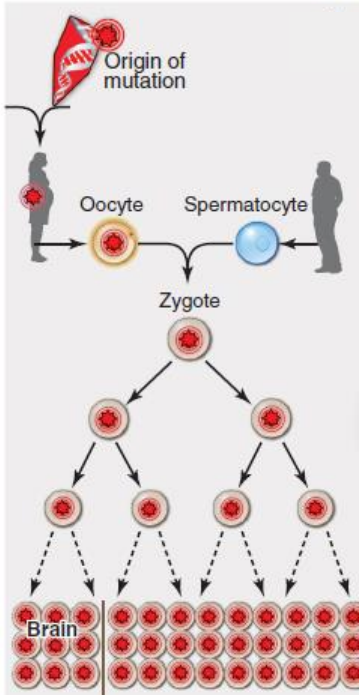
Do you think Angelina made the right decision to remove her breasts?

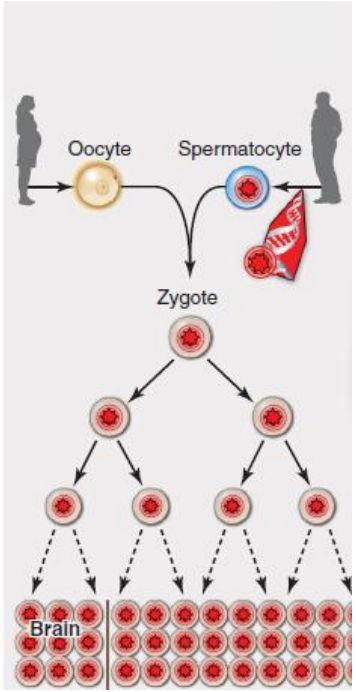
Angelina Joli has a genetic mutation in *BRCA1.*

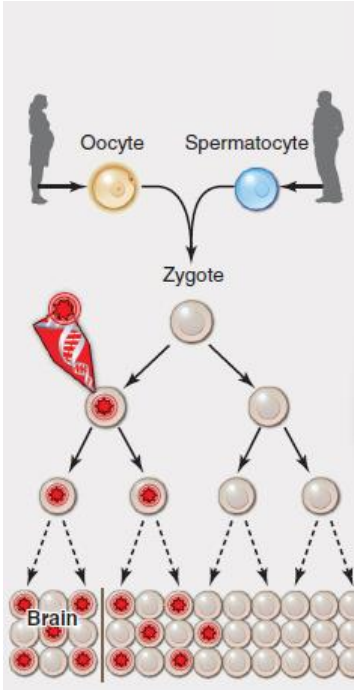
How can we predict the likelihood of her getting breast cancer given this mutation?

- **P(breast cancer|her mutation)**
- **P(breast cancer free|her mutation)**

Where did our genetic variations come from?





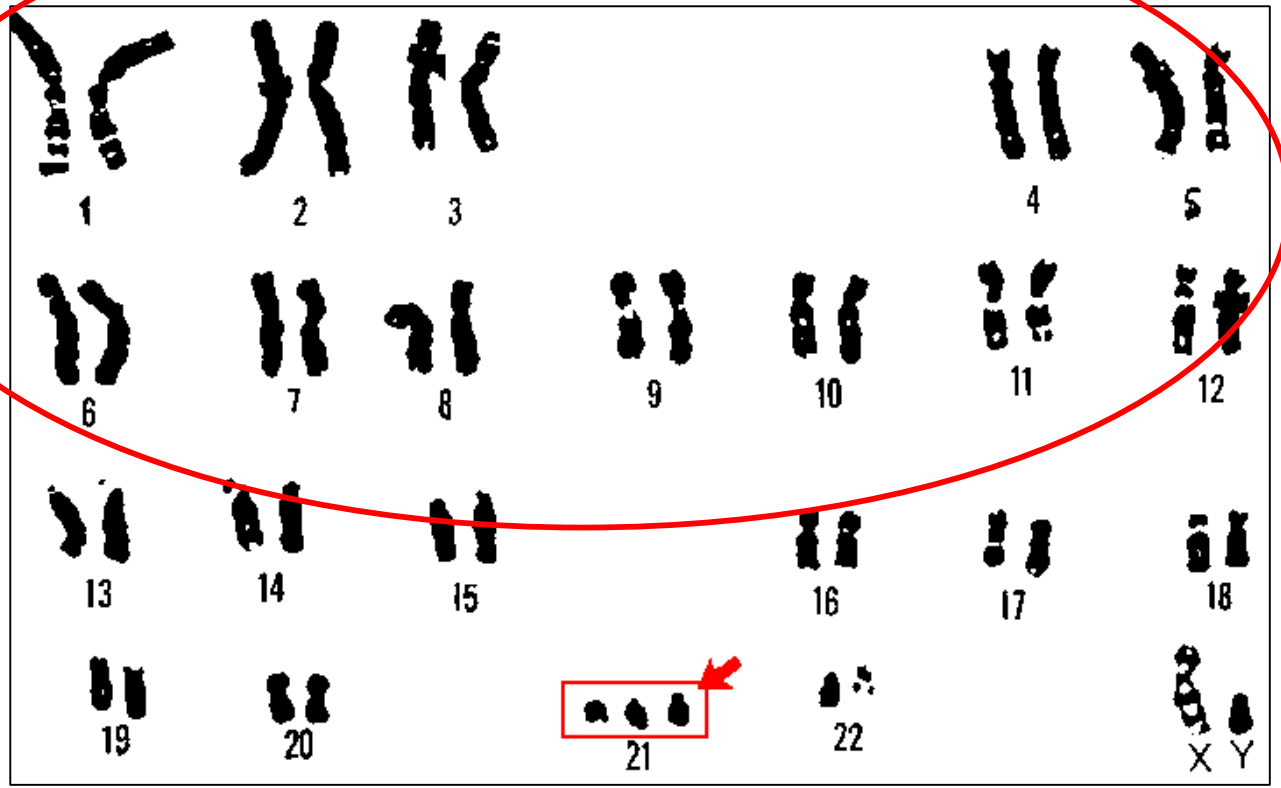


- inherited from parents
- *de novo* mutations
- somatic mutations

Annapurna Poduri et. al. *Somatic Mutation, Genomic Variation, and Neurological Disease* Science 5 July 2013: 341

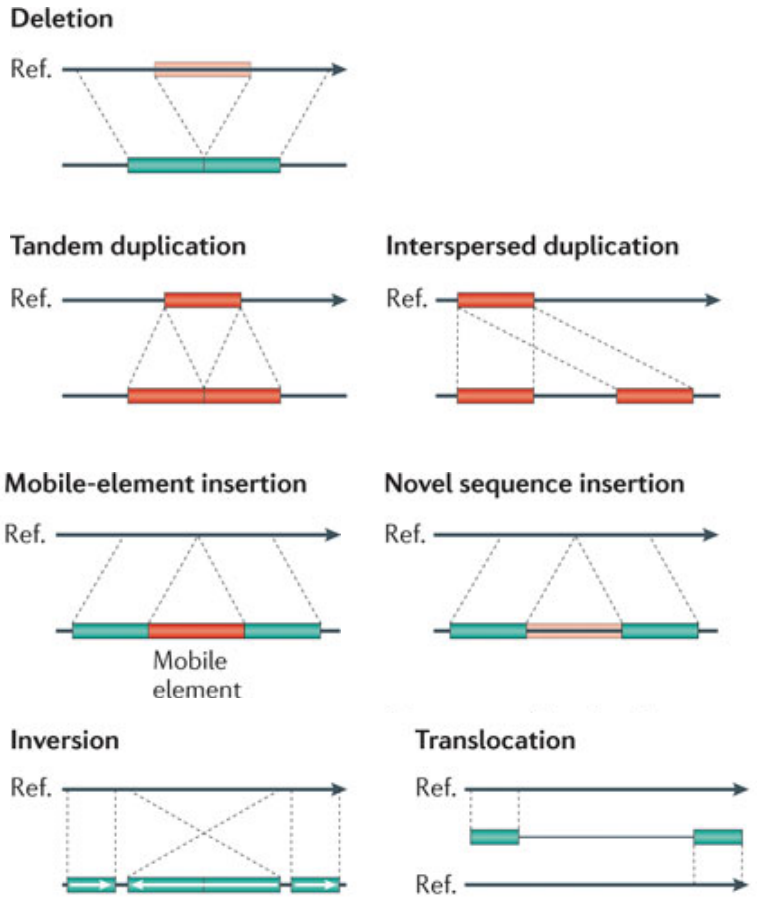
Types of genetic variations in a human genome

Chromosomal aneuploidy



http://www.esacademic.com/pictures/eswiki/50/21_trisomy_-_Down_syndrome.png

Structural Variations (SVs)



Copy Number Variations (CNVs)

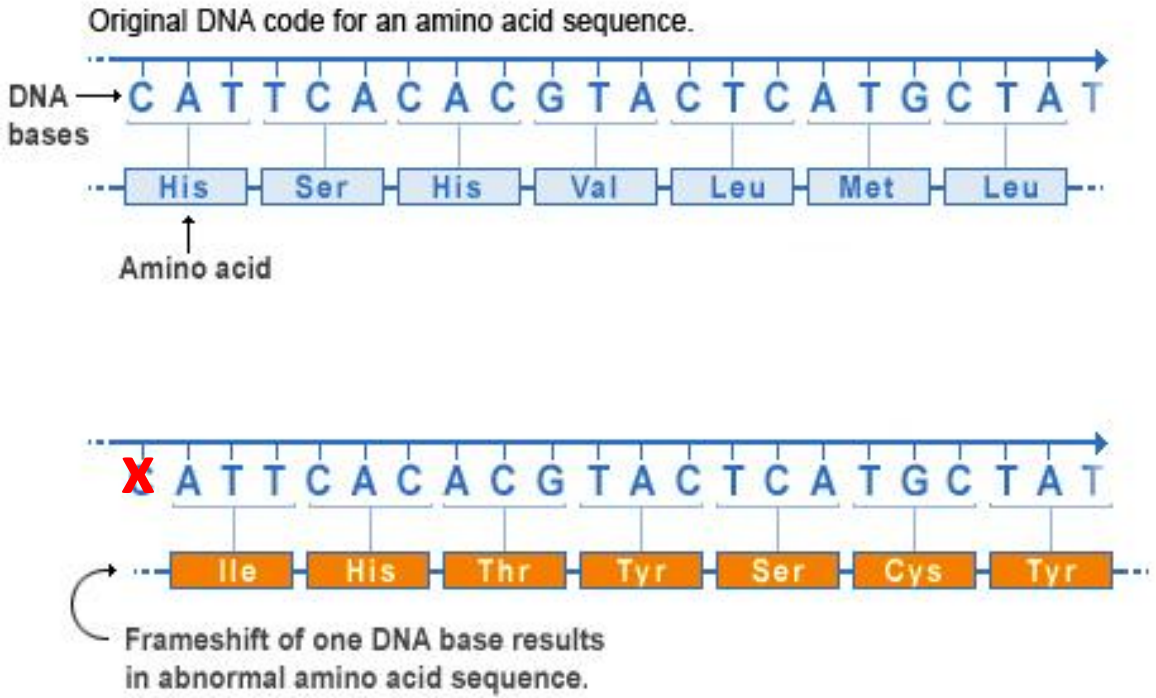
Adapted from Can Alkan et al. *Genome structural variation discovery and genotyping*. Nature Reviews Genetics 12, 363-376 (2011)

Indel – short Insertion/Deletion

Within intergenic/intronic regions

Within coding regions

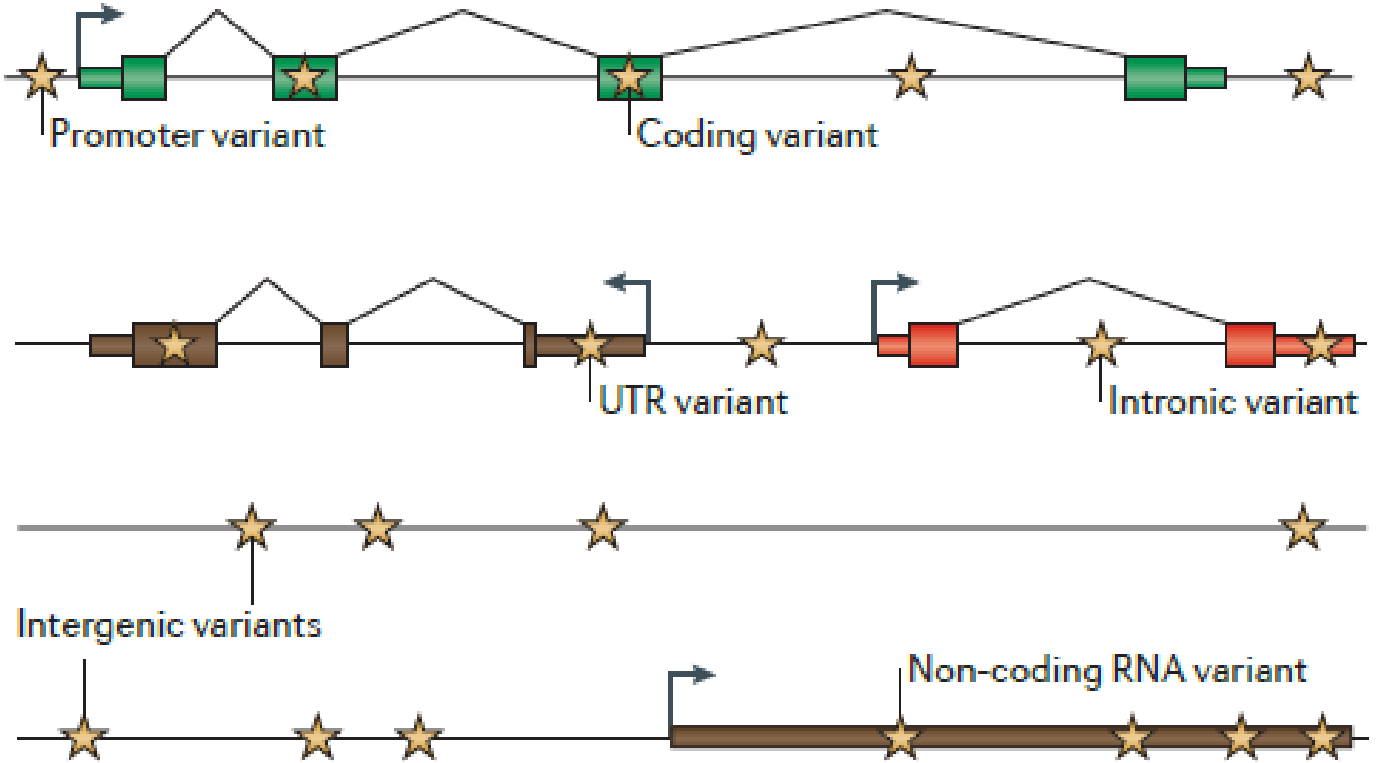
- Non-frameshifting
- Frameshifting



Adapted from http://ghr.nlm.nih.gov/handbook/illustrations

SNV – Single Nucleotide Variation

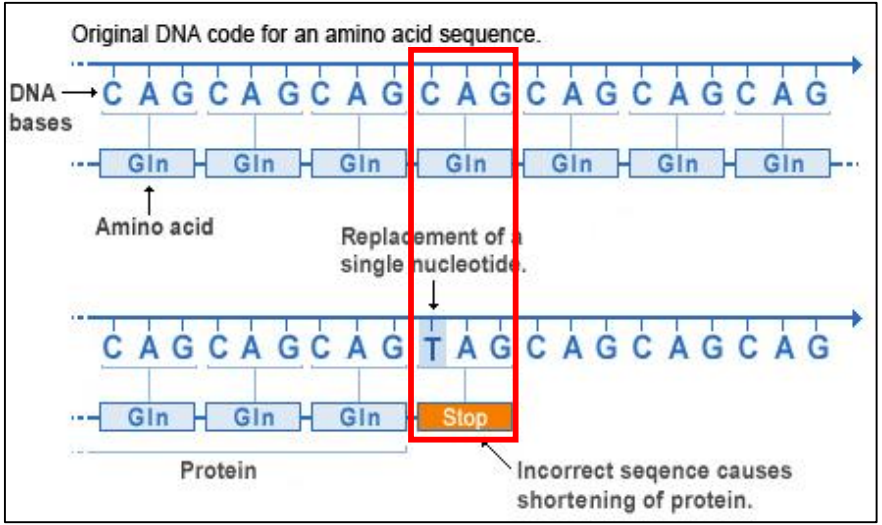
There are about 3 million SNVs in one person's genome, equivalent to ~ 1/1000 frequency.

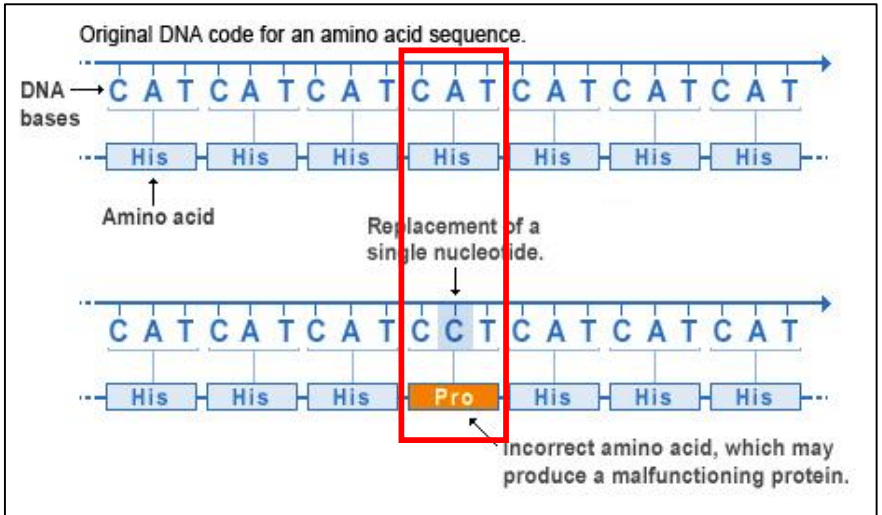


Adapted from Gregory M. Cooper and Jay Shendure. Needles in stacks of needles: finding disease-causal variants in a wealth of genomic data. Nat Rev Genet. 2011 Aug 18;12(9):628-40.

SNVs within coding regions

- Stop codon gain (nonsense)
- Non-synonymous (missense)
- Synonymous (same sense/silent)
- Affect splicing
- Stop codon loss





http://ghr.nlm.nih.gov/handbook/illustrations

- **Nonsense SNVs are usually considered deleterious.**
  - even though it is not always the case…

- **synonymous, intronic, and intergenic variations are often ignored.**
  - However, according to GWAS studies, 88% of trait-associated variants of weak effect are non-coding.
  - They remain under-studied and new methods are needed.

- **Most research so far had focused on missense SNVs.**

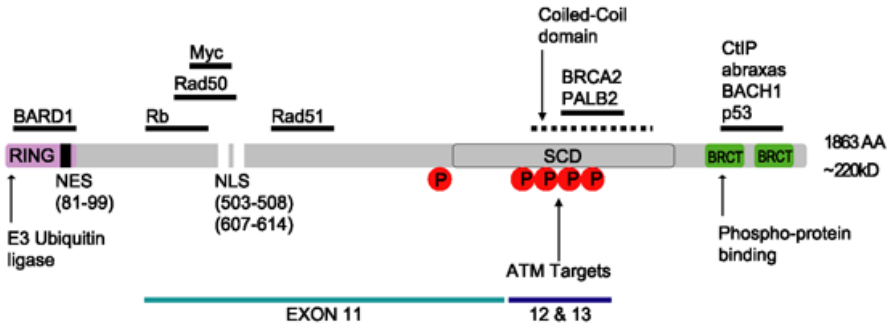
- Known deleterious mutations are enriched in missense mutations.
  - ~50 of all known mutations of Mendelian disorders are missense mutations
    - ascertainment bias?

Although missense SNVs change the protein sequence, many do not cause phenotypic changes.

In 1990, linkage analyses of large pedigrees of early-onset familial breast cancer identified *BRCA1* as the first gene associated with breast cancer.

***BRCA1* has a total of 238 known missense variations**

- 163 are present only in patients
- 62 are present only in healthy persons
- 13 in both patients and healthy persons



http://en.wikipedia.org/wiki/BRCA1

On average, a healthy individual has

Class	Number
SNP	3,019,909
Indel	361,669
Deletions	15,893
Duplications	407
mobile element insertions	4,775

Within protein-coding regions,

Class	Number
Genes disrupted by large deletions	147
Stop-introducing SNPs	1,057
Stop losses	77
Small frameshift indels	954
Small in-frame indels	714
Non-synonymous SNPs	68,300
Synonymous SNPs	60,157

1000 Genomes Project Consortium. A map of human genome variation from population-scale sequencing. Nature. 2010 Oct 28;467(7319):1061-73.

Questions

- **What features differentiate disease-causing variants from neutral ones?**
- **How can we use these features to predict whether a variation is disease-causing or not?**

Nomenclature

- Mutation
- Polymorphism
- Variation/variant

Functional/Phenotypic “effects” of human genetic variations

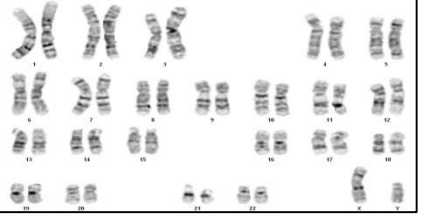
- Disease *vs.* normal
- Deleterious *vs.* neutral
- Personal trait differences (e.g., height)

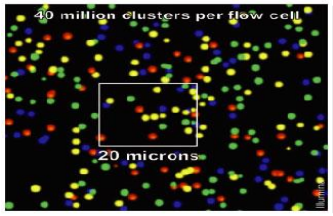
- Animal model phenotypic changes
- Cellular phenotypic changes

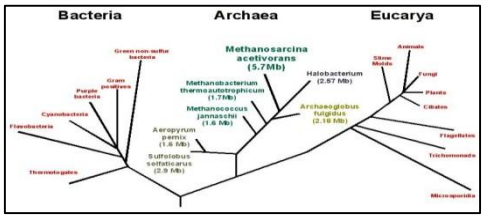
- Protein function changes
- Protein structure changes

Statistical and stochastic, not deterministic

Observations, not “truth”





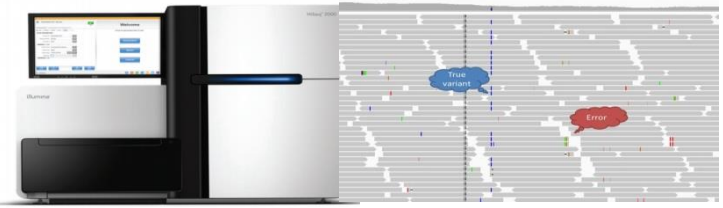


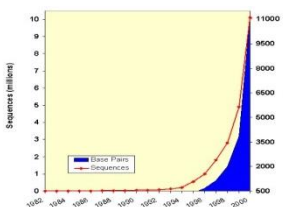
Unit 2: Databases of genetic variations

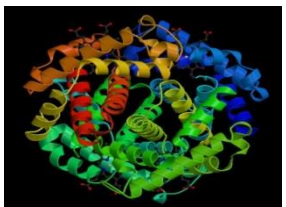
北京大学生物信息学中心 魏丽萍

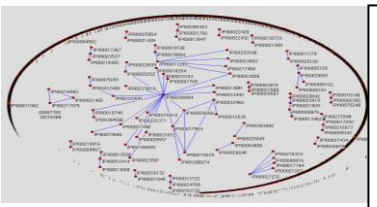
Liping Wei, Ph.D.

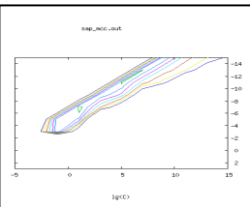
Center for Bioinformatics, Peking University











生物信息学：导论与方法
Bioinformatics: Introduction and Methods

Ge Gao 高歌 & Liping Wei 魏丽萍
Center for Bioinformatics, Peking University

https://www.coursera.org/course/pkubioinfo