SESSION 6. HUMAN DISEASE

Cancer as a result of aberrant proteins

[Keywords]
Cancer-driving mutations
Fusion gene
Sequence alignment
Dynamic programming

Cancer vs Tumor

Tumor (cell mass): benign or malignant tumors

Malignant tumor → cancer

Uncontrolled cell divisions (i.e., defects in cell-cycle checkpoints)

Resulted 1) by overproduction of proteins that stimulate cell growth or 2) by the inactivation of tumor suppressors that normally restrict cell growth.

Cancer is a complex disease caused by multiple gene mutations and epigenetic factors

Causal factors of cancer

- Two main non-genetic causal factors
  - Tabaco
  - Obesity
- Genetic factors
  - Germline mutations
  - i.e., BRCA1, 2
  - Genome instability
  - Somatic mutations

- Most tumors are most common results of mutation in somatic cells (somatic mutations).
- Mutations in oncogenes and/or tumor suppressor genes
- i.e., mutations in Wnt signaling pathway or in RB gene.

DNA damage and repair

- Chemicals and irradiation make DNA damages, point mutations or double-strand breaks.
- Mutations in **DNA repair systems** can cause cancer
- i.e., mutations in genes involved in nucleotide excision repair system

General mutational process during cancer development

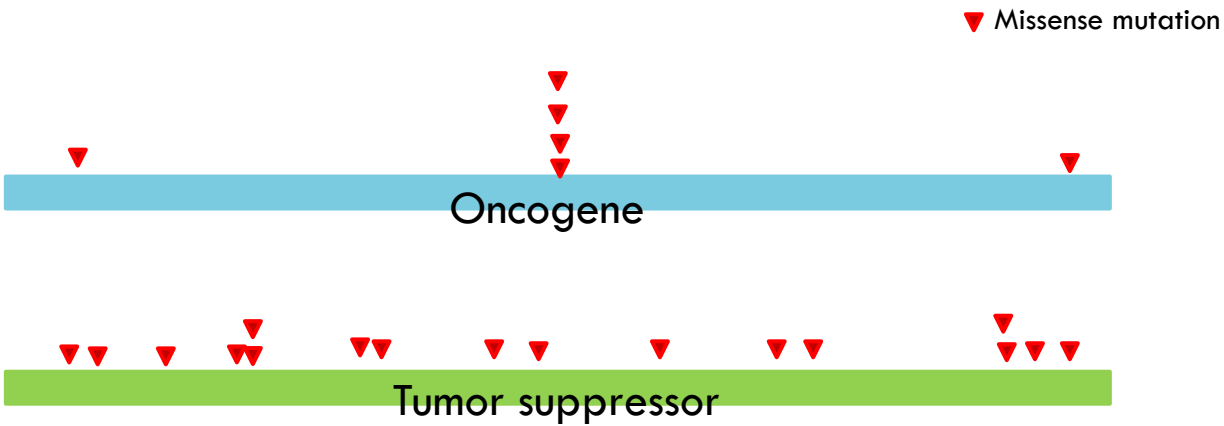
1) Sporadic mutations by replication errors or other foreign factors → 2) mutations on DNA repair systems → 3) highly frequent mutations → 4) mutations on cancer-driving genes

Additional gene mutations

- Mutations on apoptotic genes
- TGF-beta and DCC mutations in colorectal cancer

- Point mutations / small indels → highly frequent yet less impact than SV
- Structural variations → less frequent yet greater impact

Mutation patterns on OG & TS

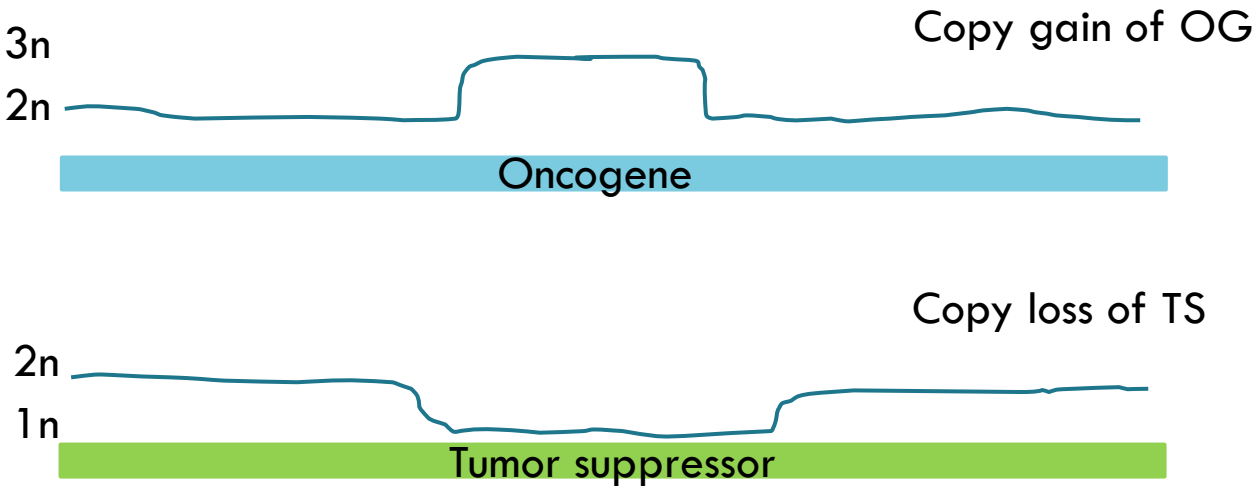


Chromosomal rearrangement in cancer

- **Aneuploidy**
  - Resulted from mis-segregation of chr. in meiosis or mitosis
- **Structural variations**
  - Large insertion/deletion
  - Inversion
  - Translocation (inter/intra)
  - Duplication
- **Copy number alteration**
  - Copy gain or loss

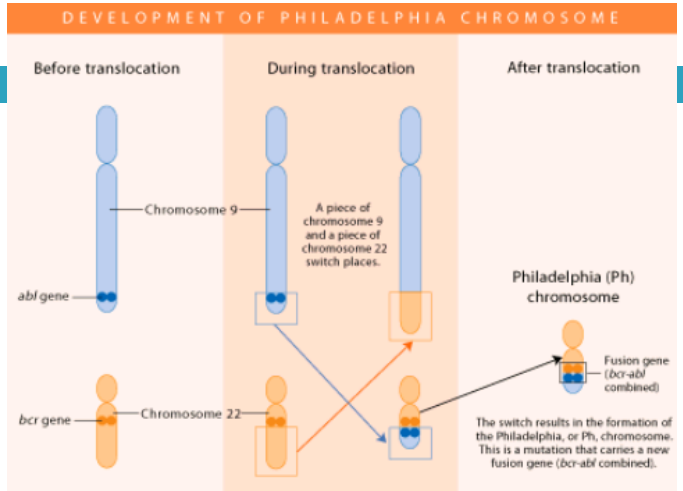
- Causes of SV
  - Nonhomologous recombination
  - Mis-replication
  - DNA ds breaks by chemicals or irradiation

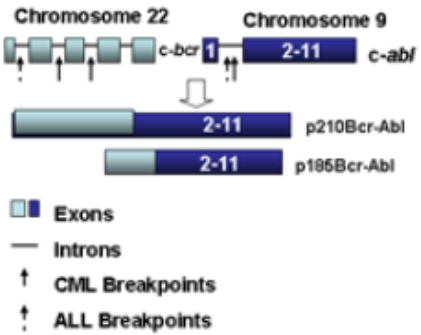
Copy number alterations on OG & TS



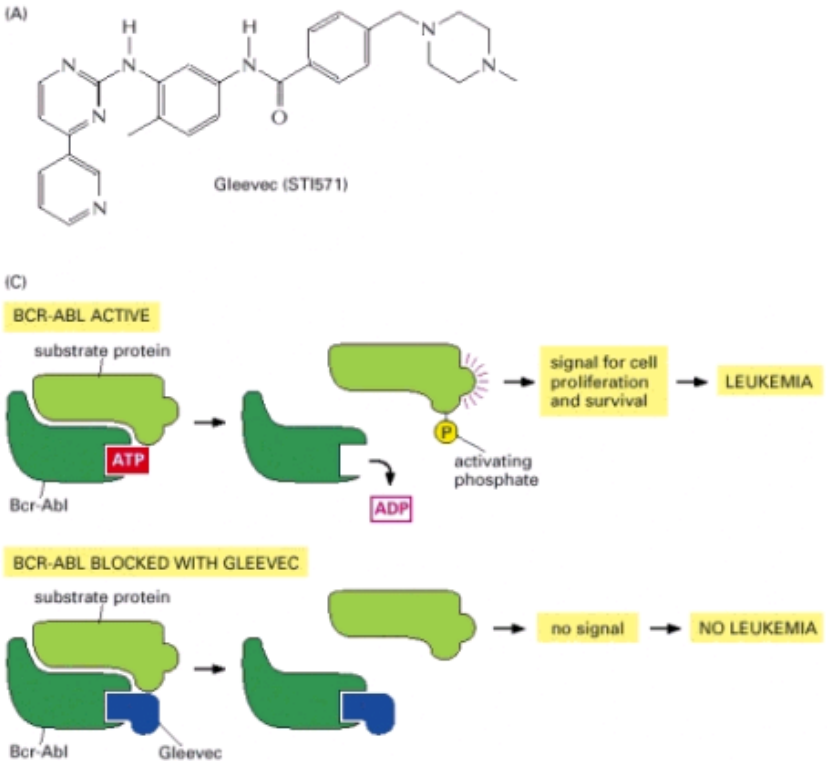
Philadelphia chromosome and fusion genes

- SVs often lead to formation of fusion genes
- Fusion genes
  - Two genes can be fused by translocation, inversion, or deletion
  - i.e., Philadelphia chromosome in Leukemia



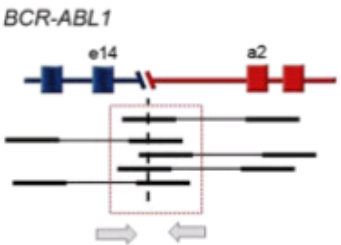


Target therapy for bcr-abl fusion gene

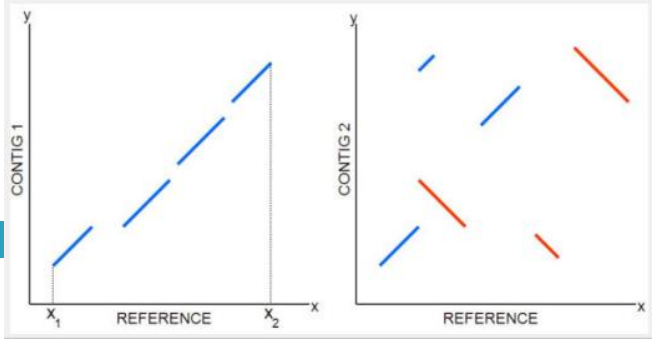


How to identify such fusion genes?

- Sequence alignment
  - Dot plot
  - Alignment
- Sequencing → Alignment → Identification of split-reads



DOT plots



- Construct a simple dot plot for

  TAGTCGATG
  TGGTCATC

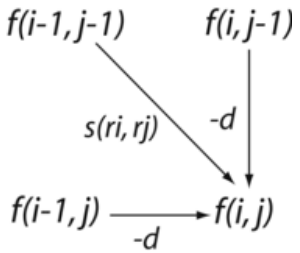
- The alignment is

  TAGTCGATG
  TGGTC-ATC

	T	A	G	T	C	G	A	T	G
T	*			*				*	
G			*			*			*
G			*			*			*
T	*			*				*	
C					*				
A		*					*		
T	*			*				*	
C					*				

Dynamic programming

- $f(i,j) = \max [0, f(i\text{-}1, j\text{-}1) + s(x_i, y_j), f(i\text{-}1, j) - d, f(i, j\text{-}1) - d]$



		C	A	A	C	A	A
	0	0	0	0	0	0	0
T	0	0	0	0	0	0	0
A	0	0	2	2	0	2	2
A	0	0	2	4	2	2	4
A	0	0	2	4	3	4	4
A	0	0	2	4	3	5	6

$s(x, y)$ = -1 (mismatch)/ 2(match)
$-d$ = -2

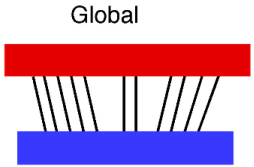
CAACAA
TAA-AA

Sequence alignment

- Local alignment
  - Covers parts of the sequences involved (Smith-Waterman alg.)


```
                tccCAGTTATGTCAGgggacacgagcatgcagagac
                |||||||||||
aattgccgccgtcgttttcagCAGTTATGTCAGatc
```

- Global alignment
  - Covers the entire lengths of the sequences involved (Needleman-Wunsch alg.)




```
--T--CC-C-AGT--TATGT-CAGGGGACACG--A-GCATGCAGA-GAC
  |  || |  ||  | | | |||    || |  | |  | ||||   |
AATTGCCGCC-GTCGT-T-TTCAG----CA-GTTATG--T-CAGAT--C
```

BLAST

- Journal of Molecular Biology (Altschul et al., 1990)
- BLAST (basic local alignment search tool)
- Compare a query sequence (DNA, RNA, protein seq.)
- K-mer-based nucleation search → Alignment extension

- BLAST algorithm
  - Remove low-complexity region or sequence repeats in the query sequence
  - Make a *k*-letter word list of the query sequence.
  - List the possible matching words.
  - Organize the remaining high-scoring words into an efficient search tree.
  - Repeat step 3 to 4 for each *k*-letter word in the query sequence.
  - Scan the database sequences for exact matches with the remaining high-scoring words.
  - Extend the exact matches to high-scoring segment pair (HSP).
  - List all of the HSPs in the database whose score is high enough to be considered.
  - Evaluate the significance of the HSP score.

$$p(S \geq x) = 1 - \exp\left(-e^{-\lambda(x-\mu)}\right)$$

BLAST Execution (make indexed_db)

Type the following in at the command prompt: “formatdb –i maize_genes.txt –p F –o F” (this command will format the target database, maize_genes.txt, so that it can be searched by BLAST)

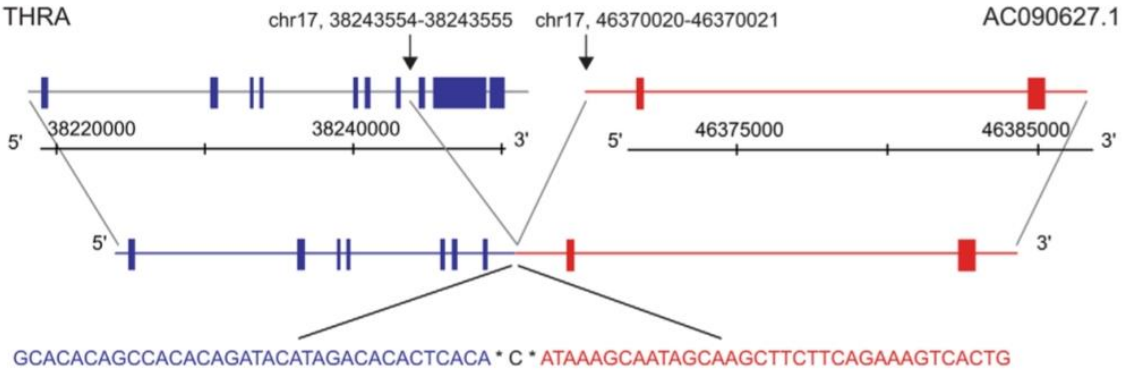
```
-i  Input file(s) for formatting [File In]  Optional
-p  Type of file
          T - protein
          F - nucleotide [T/F]  Optional
          default = T
-o  Parse options
        T - True: Parse SeqId and create indexes.
        F - False: Do not parse SeqId. Do not create indexes
```

BLAST Execution

```
blastall –p [blastn, blastp, blastx, tblastx] –d [database] –i [query.fa]
–e [Expectation_value] –m [alignment view]
```

```
-p  Program Name [String]
-d  Database [String]
   default = nr
-i  Query File [File In]
   default = stdin
-e  Expectation value (E) [Real]
   default = 10.0
-m  alignment view options:
      0 = pairwise,
      1 = query-anchored showing identities,
      2 = query-anchored no identities,
      3 = flat query-anchored, show identities,
      4 = flat query-anchored, no identities,
      5 = query-anchored no identities and blunt ends,
      6 = flat query-anchored, no identities and blunt ends,
      7 = XML Blast output,
      8 = tabular,
      9 tabular with comment lines
```

BLAST Execution



Scoring matrix (PAM, BLOSUM) for protein sequence alignment

- **BLOSUM** (**BLO**cks **SU**bstitution **M**atrix) matrix is a substitution matrix used for sequence alignment of proteins.
- scanned the BLOCKS database for very conserved regions of protein families (that do not have gaps in the sequence alignment) and then counted the relative frequencies of amino acids and their substitution probabilities

	Ala	Arg	Asn	Asp	Cys	Gln	Glu	Gly	His	Ile	Leu	Lys	Met	Phe	Pro	Ser	Thr	Trp	Tyr	Val
Ala	4																			
Arg	-1	5																		
Asn	-2	0	6																	
Asp	-2	-2	1	6																
Cys	0	-3	-3	-3	9															
Gln	-1	1	0	0	-3	5														
Glu	-1	0	0	2	-4	2	5													
Gly	0	-2	0	-1	-3	-2	-2	6												
His	-2	0	1	-1	-3	0	0	-2	8											
Ile	-1	-3	-3	-3	-1	-3	-3	-4	-3	4										
Leu	-1	-2	-3	-4	-1	-2	-3	-4	-3	2	4									
Lys	-1	2	0	-1	-3	1	1	-2	-1	-3	-2	5								
Met	-1	-1	-2	-3	-1	0	-2	-3	-2	1	2	-1	5							
Phe	-2	-3	-3	-3	-2	-3	-3	-3	-1	0	0	-3	0	6						
Pro	-1	-2	-2	-1	-3	-1	-1	-2	-2	-3	-3	-1	-2	-4	7					
Ser	1	-1	1	0	-1	0	0	0	-1	-2	-2	0	-1	-2	-1	4				
Thr	0	-1	0	-1	-1	-1	-1	-2	-2	-1	-1	-1	-1	-2	-1	1	5			
Trp	-3	-3	-4	-4	-2	-2	-3	-2	-2	-3	-2	-3	-1	1	-4	-3	-2	11		
Tyr	-2	-2	-2	-3	-2	-1	-2	-3	2	-1	-1	-2	-1	3	-3	-2	-2	2	7	
Val	0	-3	-3	-3	-1	-2	-2	-3	-3	3	1	-2	1	-1	-2	-2	0	-3	-1	4