SESSION 6. HUMAN DISEASE

Cancer as a result of aberrant proteins

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 function that normally restrict 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
  - i.e., BRAC1, 2
  - Genome instability
  - Germline mutations

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

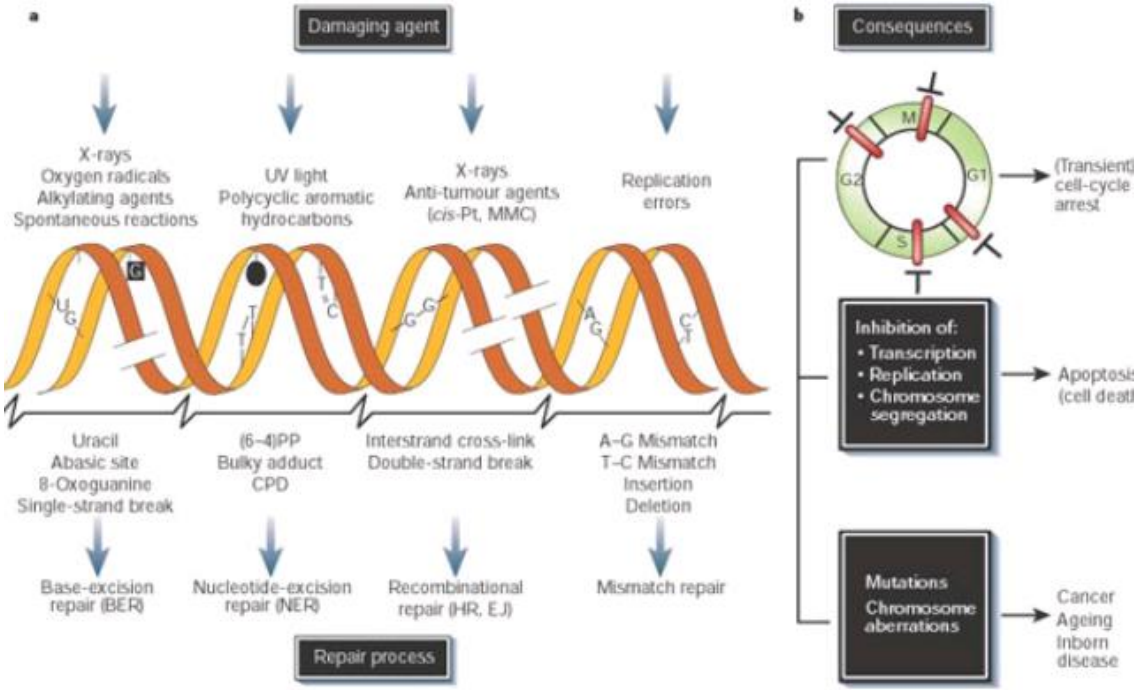
Additional gene mutations

- Mutations on Apoptotic genes
- TP53 mutations in cancers
- TGF-beta and DCC mutations in colorectal cancer

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

DNA repair processes



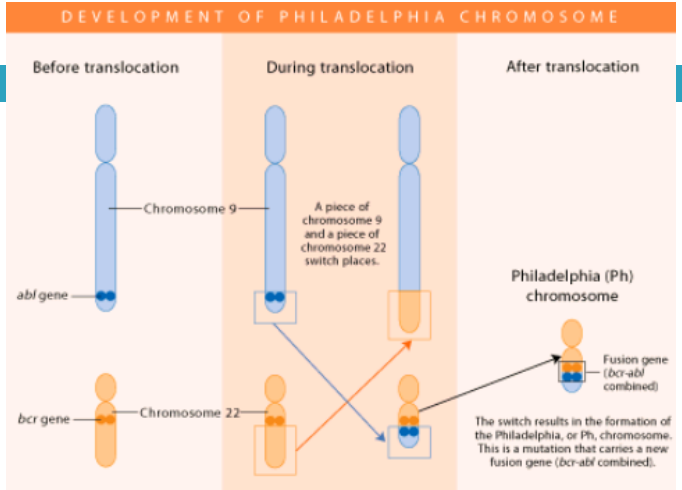
Chromosomal rearrangement in cancer

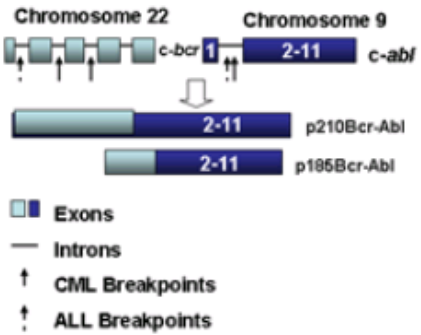
- **Aneuploidy**
  - Resulted from mis-segregation of chr. in meiosis or mitosis
- **Copy number alteration**
- **Structural variations**
  - Large insertion/deletion
  - Inversion
  - Translocation (inter/intra)
  - Duplication

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

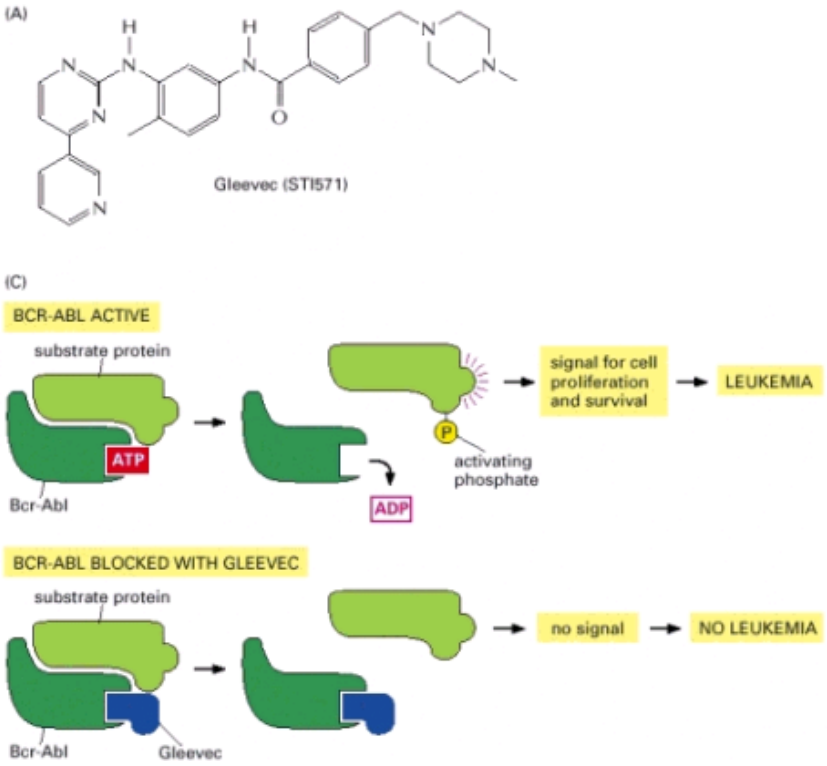
Philadelphia chromosome and fusion genes

- 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



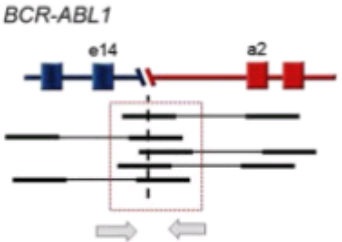


Target therapy for bcr-abl fusion gene



How to identify such fusion genes?

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



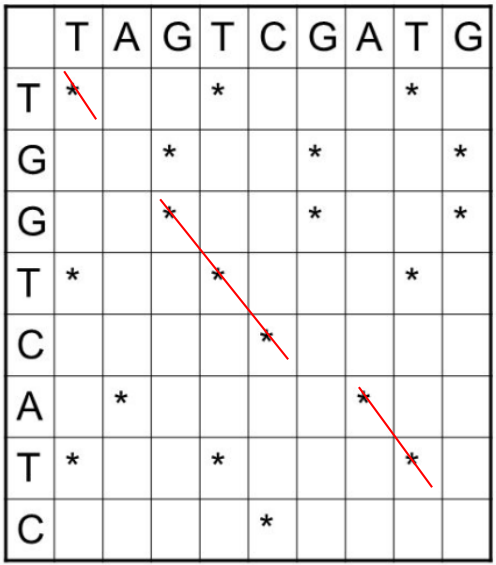
DOT plots

- Construct a simple dot plot for

  TAGTCGATG
  TGGTCATC

- The alignment is

  TAGTCGATG
  TGGTC-ATC

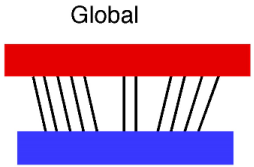


Sequence alignment

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


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

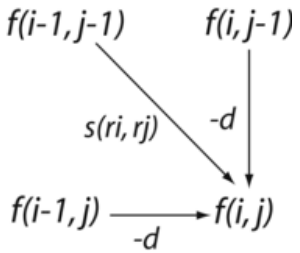
- Global alignment
  - Covers the entire lengths of the sequences involved (Needleman-Wunsch alg.)




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

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

BLAST

- Journal of Molecular Biology (Altschul et al., 1990)
- BLAST (basicl 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.

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

Scoring matrix (PAM, BLOSUM)

- **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

BLAST Execution

1. Database indexing
   -> **makeblastdb**

2. Download or generate sequence data (fasta format)
   -> **blastdbcmd**

3. Global alignment of protein sequences
   -> **blastp**