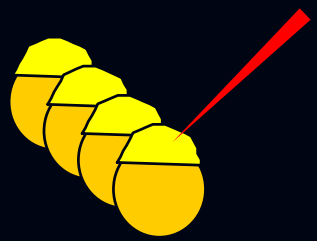
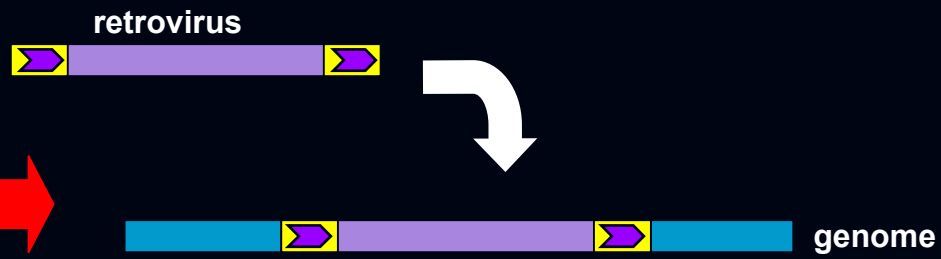
Different Global Genomic Preferences for MLV and HIV-1 Proviral Integration

Shawn Burgess
Genome Technology Branch
National Human Genome Research Institute
National Institutes of Health

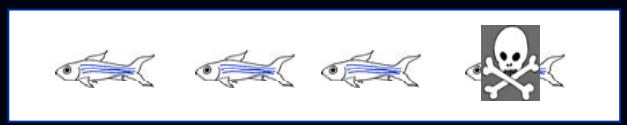
Insertional Mutagenesis Screen



Inject pseudotyped retrovirus into zebrafish embryos



Retrovirus integrates into genome causing mutation



Identify mutation

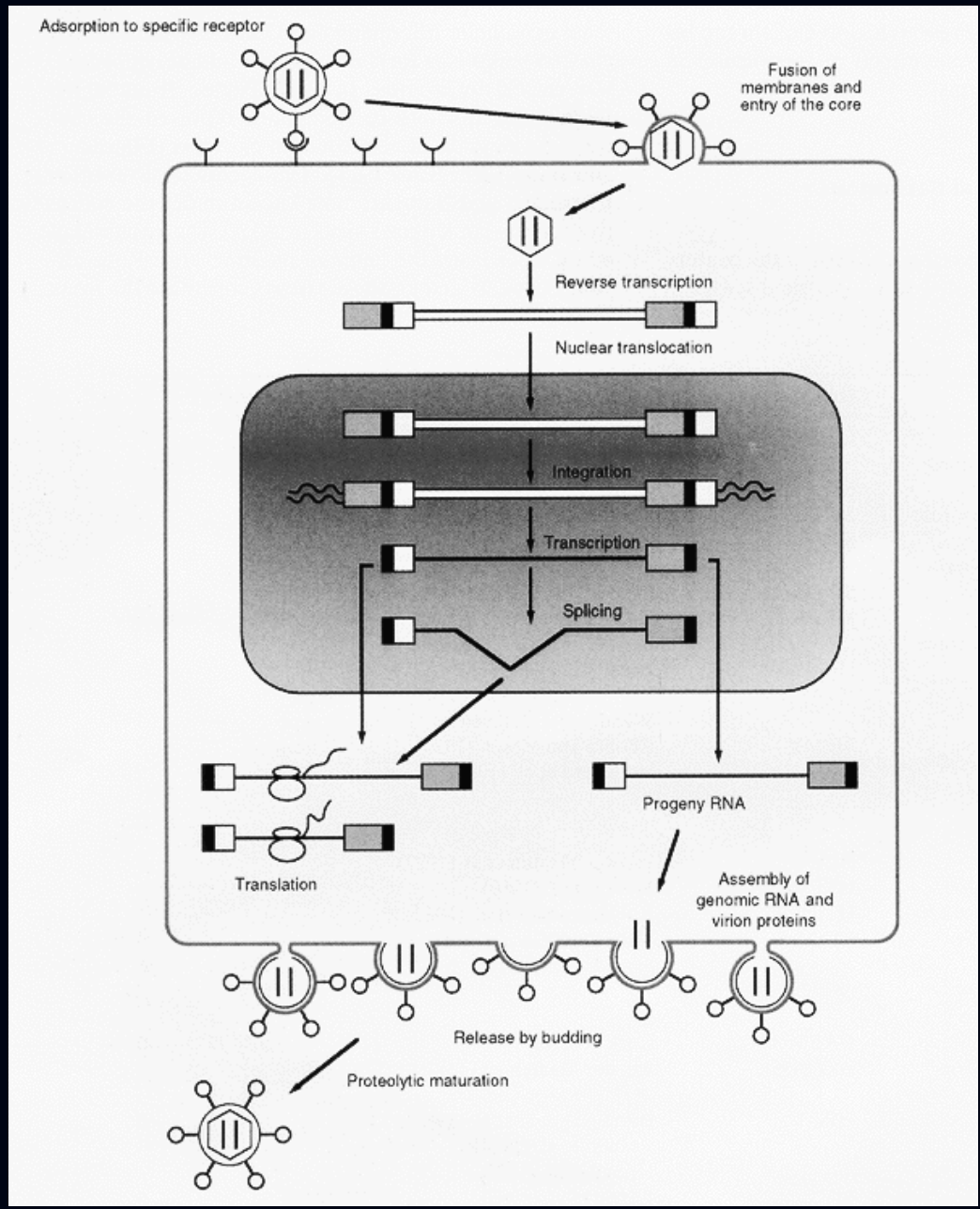
Clone gene using retroviral sequences

What is the efficiency of retroviral insertional mutagenesis?

- Does the virus integrate randomly into the genome?
- Does it like genes?
- Does it like exons?
- Any regional hot spots?
- How many integrations do we need to disrupt all genes in the zebrafish genome? Is it possible?
- How can we identify integration sites quickly and efficiently?

Retrovirus

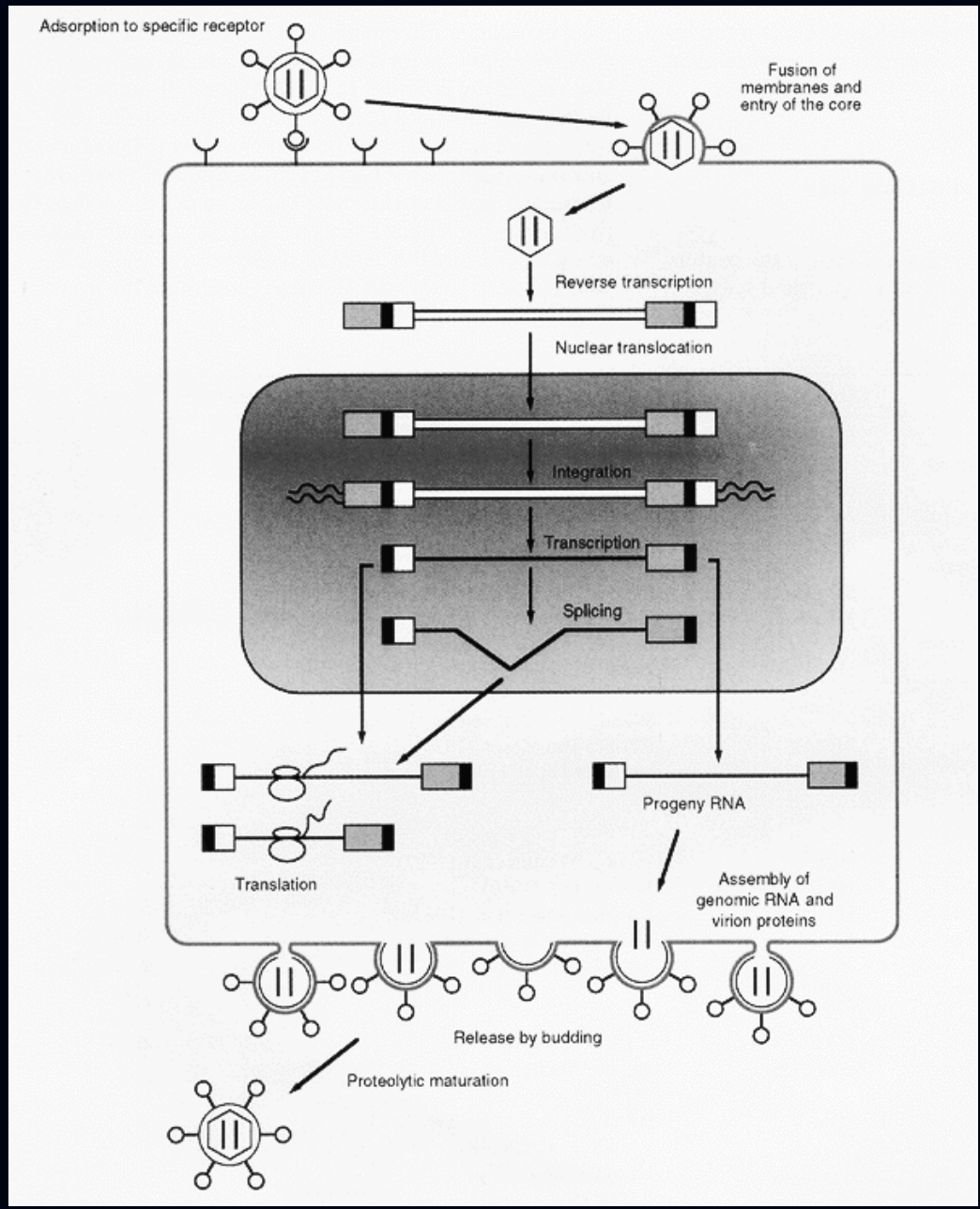
- RNA virus
- Life Cycle:
  - Entry
  - Reverse Transcription
  - Integration into the host genome
  - Transcription
  - Package



From *Retroviruses*, Coffin, Hughes, Varmus, 1997

Retrovirus

- RNA virus
- Life Cycle:
  - Entry
  - Reverse Transcription
  - Integration into the host genome
  - Transcription
  - Package



From *Retroviruses*, Coffin, Hughes, Varmus, 1997

Retroviral Integration requires specific viral sequences

- Catalyzed by integrase
- DNA sequence elements in virus LTR required: 5'NNTG----------CANN3'

No conserved sequence found in host DNA

No Clear Understanding of Target Site Selection

In vitro studies

- Typically using a specific piece of DNA such as a plasmid as target
- Findings: largely random, but nucleosomal structure or DNA binding proteins may influence target site selection

In vivo studies

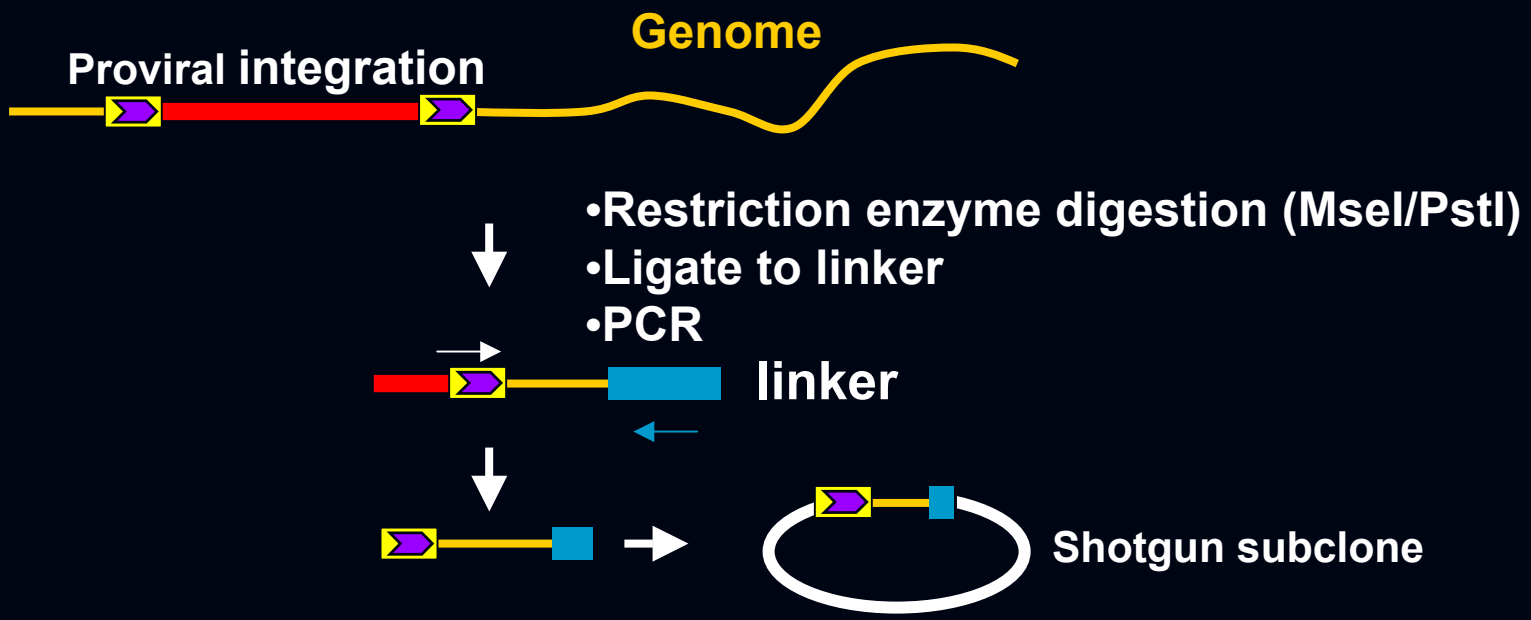
- DNase I hypersensitive regions are preferred (Vijaya et al, 1986; Rohdewohld et al, J Vir 1987)
- Transcriptionally active regions are preferred (Scherdin et al, J Vir 1990)
- Transcriptionally active DNA is disfavored (Weidhaas et al, J Vir 2000)

Limitations of early *in vivo* studies

- Clonal selection to clone junction fragment: isolation of stably integrated provirus from cell lines or tumors
- Small sample size: no studies had statistically significant numbers of integrations

Mapping Integrations

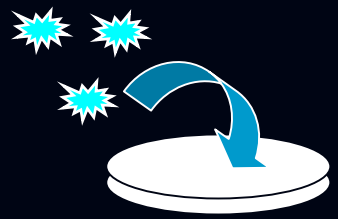
- Linker-mediated PCR to amplify junction sequences



Sequence and BLAST/BLAT against genomic sequence

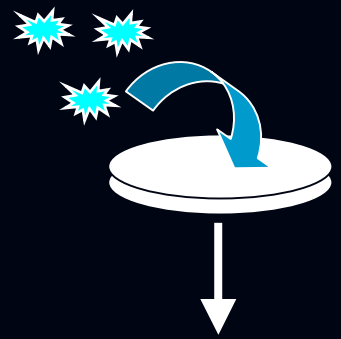
Availability of human genome make it possible to study global retroviral integration sites selection

- Infect human HeLa cell line with replication defective MLV, HIV-1



Availability of human genome make it possible to study global retroviral integration sites selection

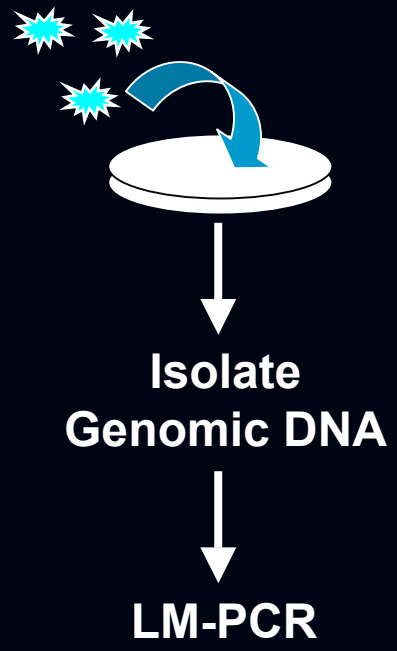
- Infect human HeLa cell line with replication defective MLV, HIV-1
- Grow for 48 hrs *without selection*
- Extract genomic DNA with thousands to millions of integrations



Isolate
Genomic DNA

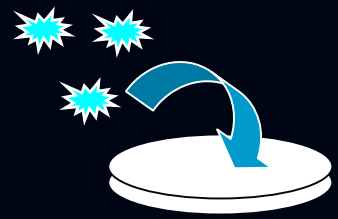
Availability of human genome make it possible to study global retroviral integration sites selection

- Infect human HeLa cell line with replication defective MLV, HIV-1
- Grow for 48 hrs *without selection*
- Extract genomic DNA with thousands to millions of integrations
- Linker mediated-PCR amplify junction fragments



Availability of human genome make it possible to study global retroviral integration sites selection

- Infect human HeLa cell line with replication defective MLV, HIV-1
- Grow for 48 hrs *without selection*
- Extract genomic DNA with thousands to millions of integrations
- Linker mediated-PCR amplify junction fragments
- Shotgun clone PCR products and sequence junction fragments
- Map to the genome and analyze distribution



Isolate Genomic DNA

LM-PCR

Sequence, BLAT genome

The information will be of interest to virologists, gene therapists, and geneticists

Two major classes of retroviruses are used as gene delivery vectors

- **Oncoretrovirus:**
  - known to cause cancer
  - Genome is simple: LTR-gag-pol-env-LTR
  - Example: MLV
- **Lentivirus:**
  - Known to cause disease slowly
  - Genome is more complex
  - Example: HIV-1
  - Advantage: can integrate into non-dividing cells

Mapping Results

Integration sites that were mapped to a unique position in the human genome (UCSC Nov 2002 freeze)

- MLV=903
- HIV-1=379(+524)=903

Schroder et al 2002

Global Integration Preferences

- **Genic region vs Non-genic region:**

  Genic region: Transcription Start-Transcription Stop of RefSeq genes UCSC Nov 2002 freeze, 18,214 RefSeq Genes.

	MLV	HIV-1	Random
Landed in RefSeq Genes	34.2%* (309/904)	57.8%* (219/379)	22.4%

*$p<0.001$ compared to random integration, χ^2 test

Wu et al. 2003

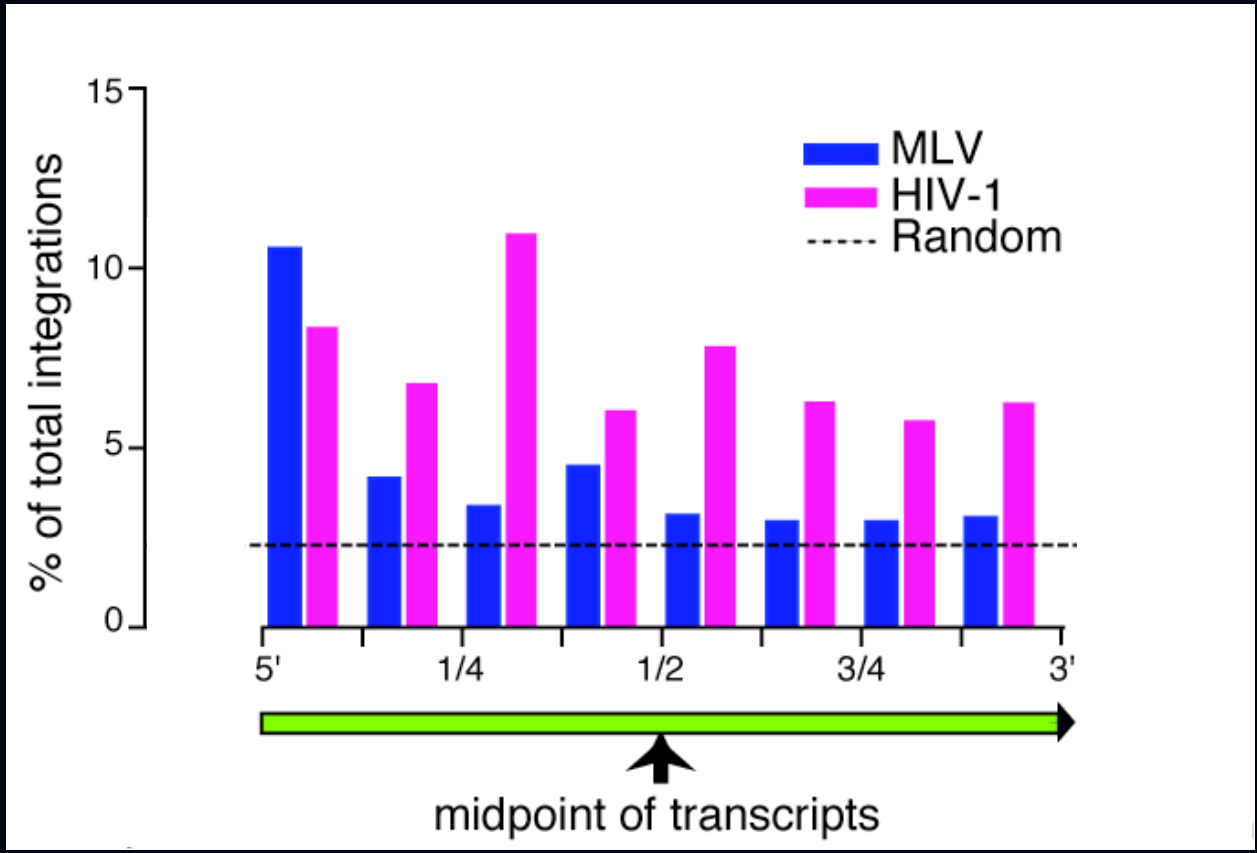
MLV prefers CpG islands
HIV-1 shows no preference

- CpG islands are commonly associated with the 5' end of genes. There are 27,704 CpG islands documented in the Nov 2002 freeze of human genome

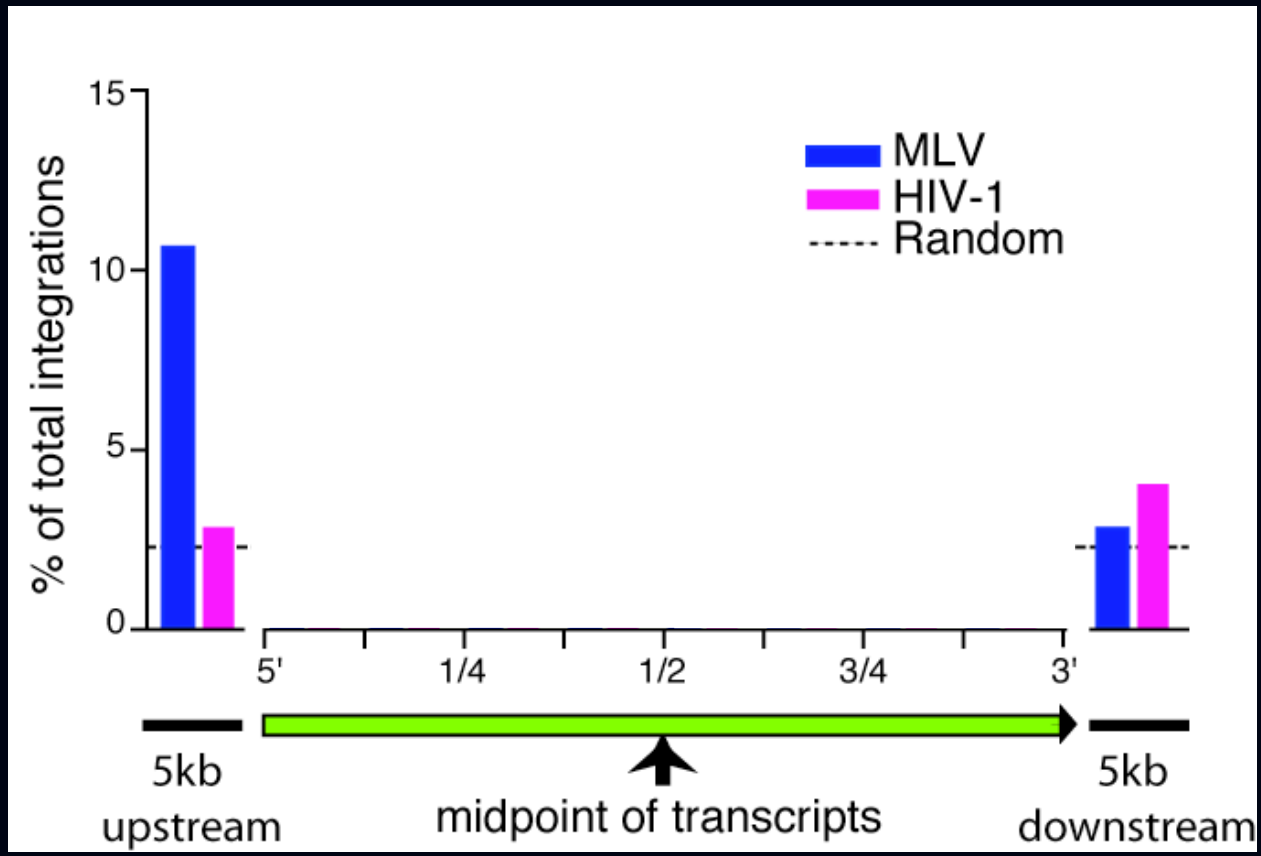
	MLV	HIV-1	Expected
% in CpG island region (+/- 1kb)	16.8%*	2.1%	2.1%

* $p < 0.001$, χ^2 test

Do retroviruses like certain regions of genes?

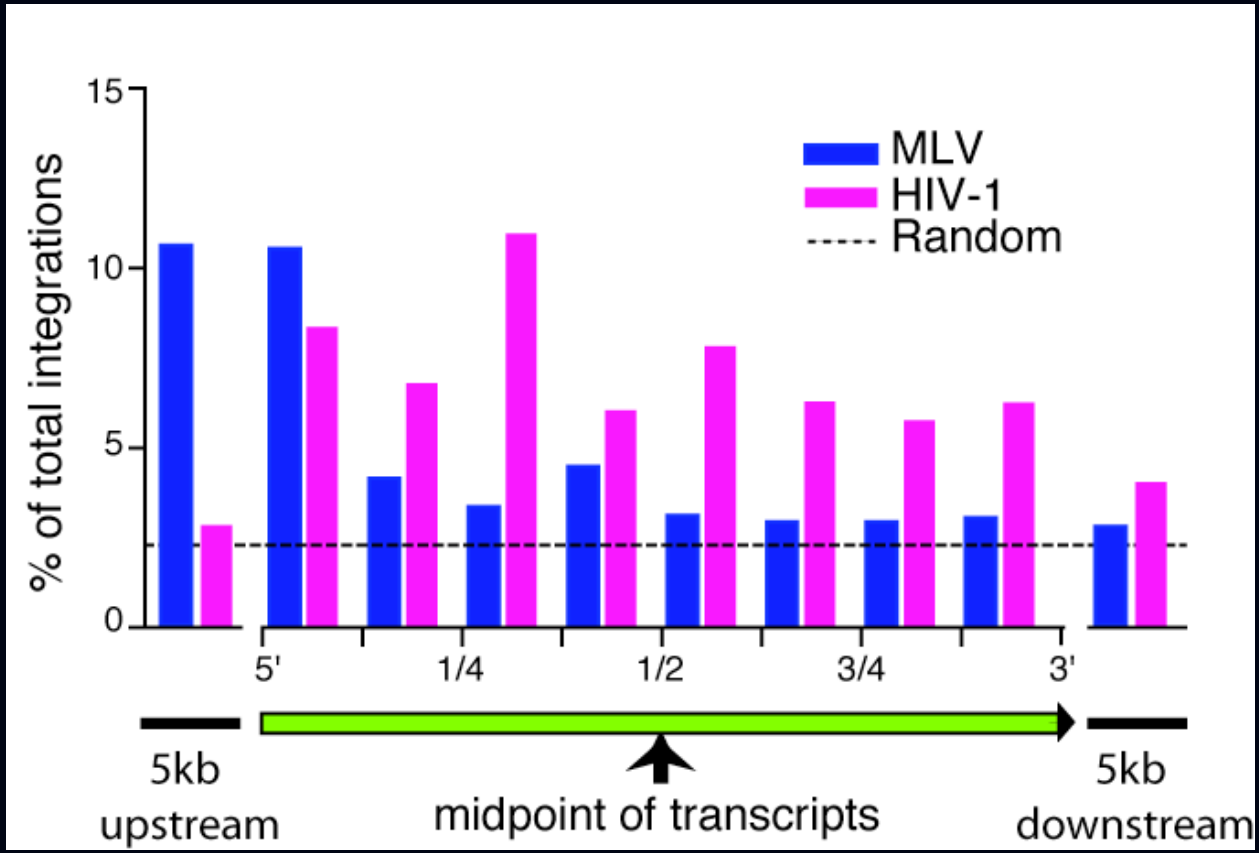


Do Retroviruses Prefer to Land in the Upstream or Downstream Regions of Genes?

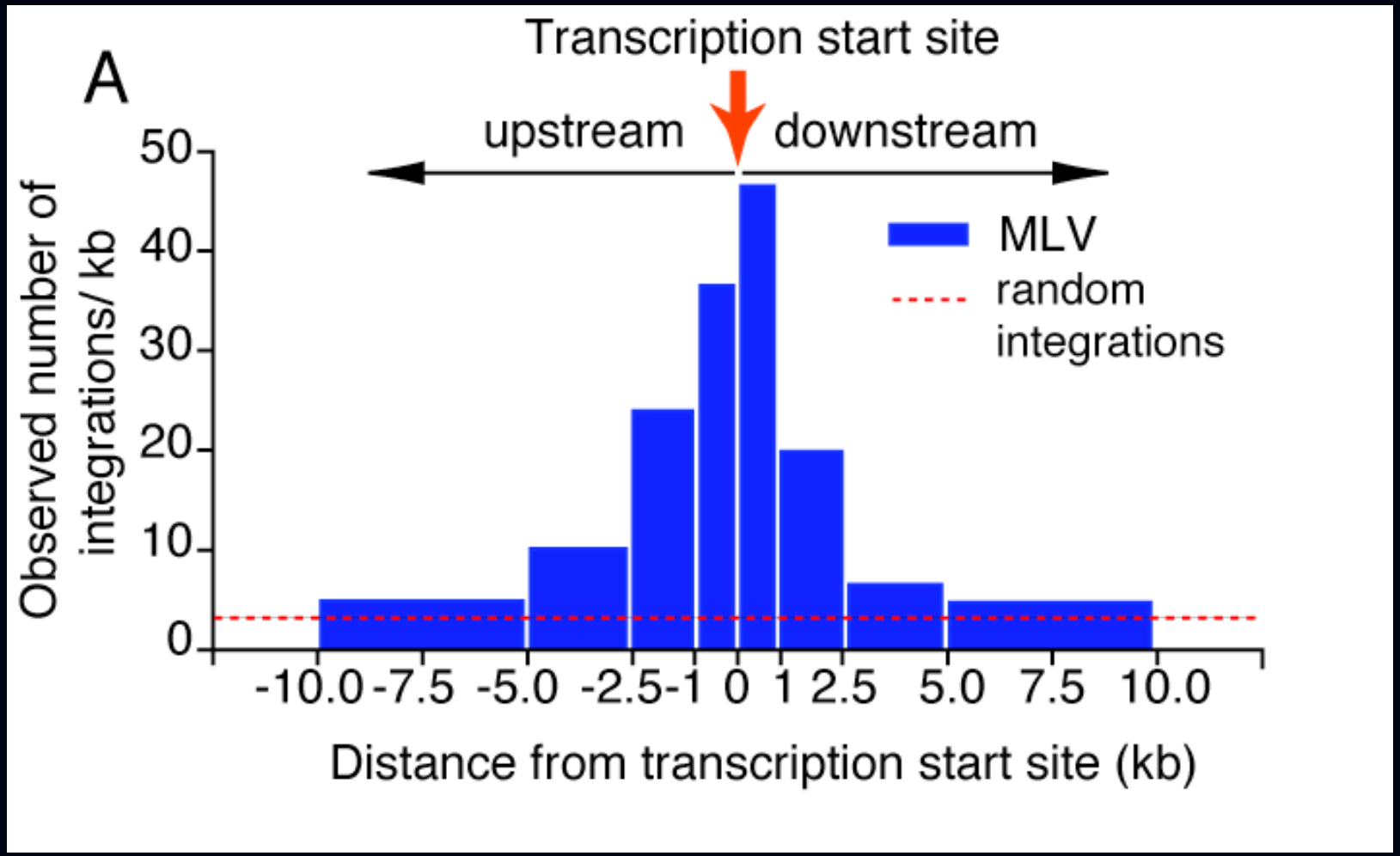


The distribution of integration sites in genic regions are different for MLV and HIV-1:

MLV shows a strong preference for the 5' end of genes
HIV-1 integrates evenly across genes
we did not see any preference for introns or exons



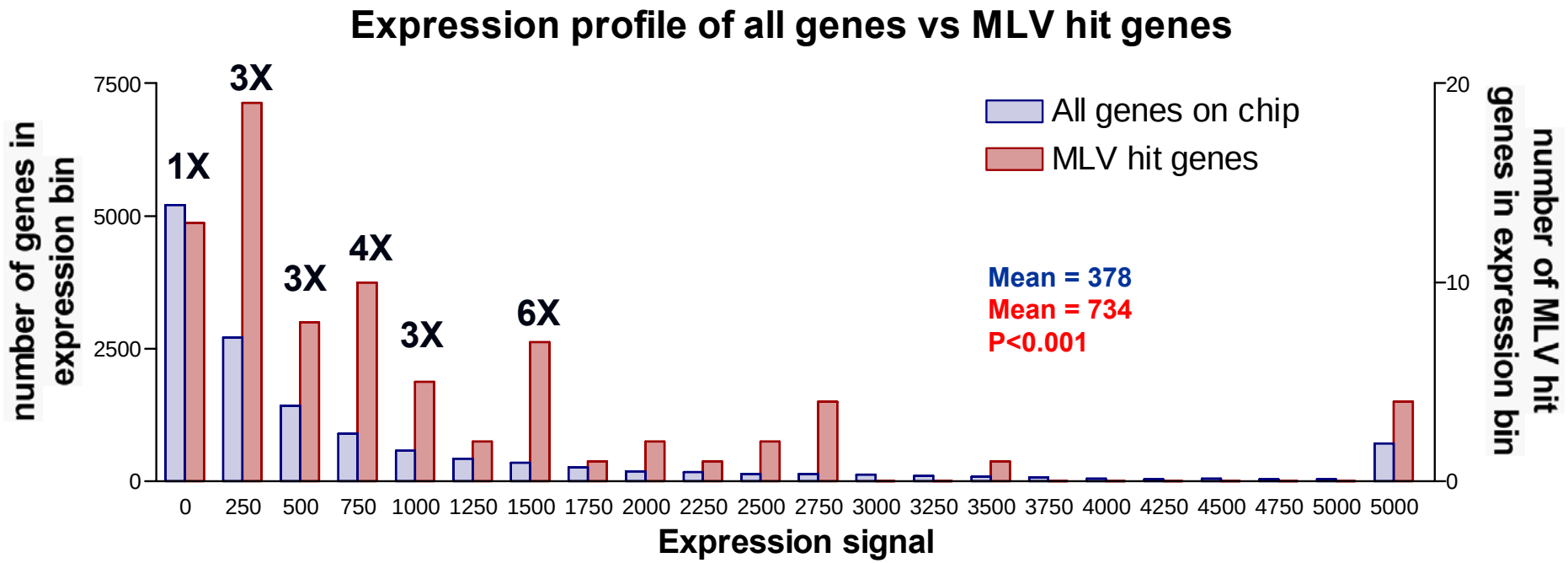
MLV prefers a small window around transcriptional start sites



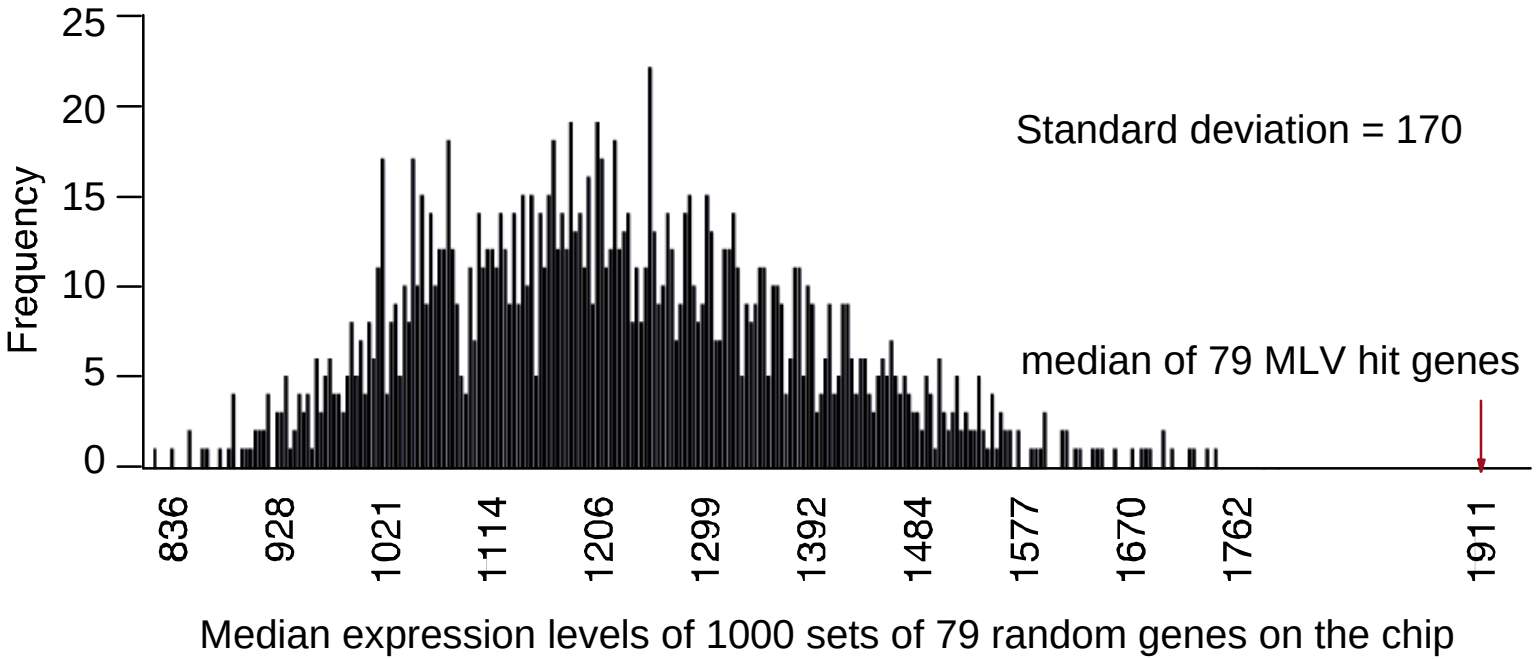
MLV target genes are more active than other genes

HeLa cell microarray expression analysis

	Dataset1	Dataset2	Dataset3
The median expression level of targeted genes	2055	1209	734
The median expression level of all genes	1228	487	378


Expression profile of all genes vs MLV hit genes
number of genes in expression bin
number of MLV hit genes in expression bin
All genes on chip
MLV hit genes
Mean = 378
Mean = 734
P<0.001
1X
3X
3X
4X
3X
6X
7500
5000
2500
0
20
10
0
0 250 500 750 1000 1250 1500 1750 2000 2250 2500 2750 3000 3250 3500 3750 4000 4250 4500 4750 5000 5000
Expression signal

Comparison of the median expression level of 79 MLV targeted genes to the median levels of 1000 sets of 79 randomly picked genes on the chip



Safety issue of retroviral vectors in gene therapy

- The chance of insertional mutagenesis in gene therapy was considered very small in the field.
- Our data show that 20% of MLV integrations landed in the +/- 5kb region of transcription start of RefSeq genes, likely higher for all genes.
- In both cases of leukemia in gene-therapy children, an integration was found near the oncogene, LMO2. Both integrations fit the preferred site profile.
- Based on our data, the number of cells and the number of integrations per cell used for gene therapy, >220 integration events will occur near the LMO2 gene in a 5,000,000 cell infection.

Relevance to CIS Analysis

Copeland Data	Our Data
1202 mapped insertions	903 Mapped Insertions
146 CIS	64 CIS (80)
17 with ≥ 5 integrations	2 with 6 integrations
11 with 4 integrations	1 with 4 integrations
18 with 3 integrations	3 with 3 integrations
95 with 2 integrations	58 (72) with 2 integrations (random≈12)
Suzuki et al. 2002	Wu et al. 2003

Conclusions

- MLV likes transcriptional start sites
- HIV likes transcriptional units
- Different retroviruses may have different integration preferences, which may reflect the involvement of different cellular factors
- Each vector will have different risk factors associated with it

Lab:
Xiaolin Wu

Lisha Xu

Tess Yannucci

Jizhou Yan

Zengfeng Wang

Deborah Davis

Martine Behra

Dr. Eric Green and the Green lab

Dr. Tyra Wolfsberg and Bioinformatics core facility

NISC