

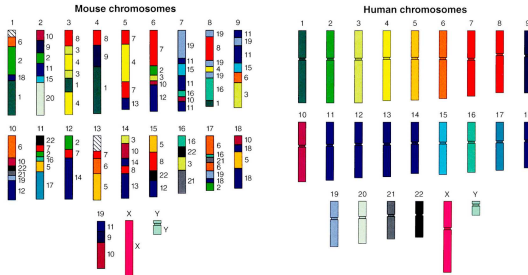
The Deletion-Insertion Model Applied to the Genome Rearrangement Problem (Work in Progress)

Abra Brisbin, Manda Riehl, Noah Williams

University of Wisconsin - Eau Claire

Permutation Patterns - July 5, 2013

Mouse and Human Genetic Similarities



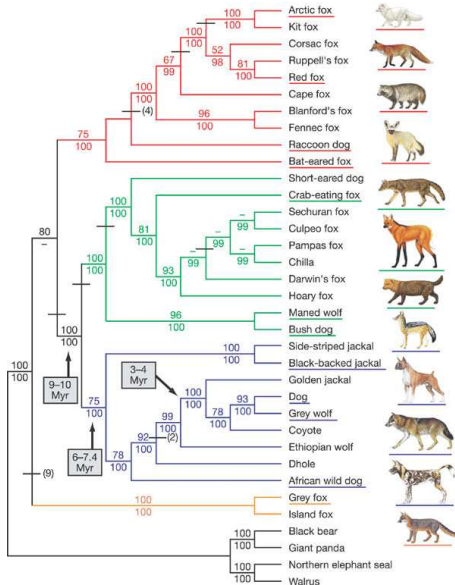
YCA 98-073R2

Courtesy Lisa Stubbs
Oak Ridge National Laboratory

[2]

- Preserved segments between mouse and human genomes.
- More than ninety percent of the mouse genome consists of shuffled pieces of the human genome [4]

Phylogenetic Trees



[3]

Neurofibromatosis (NF)

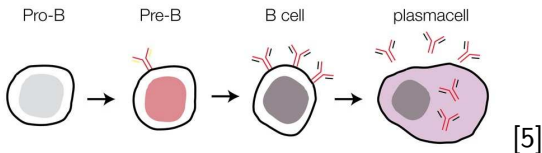
- A common autosomal dominant disorder
- Sporadic and hereditary mutations [6]
- Deletions and insertions prevalent [1]

Neurofibromatosis (NF)

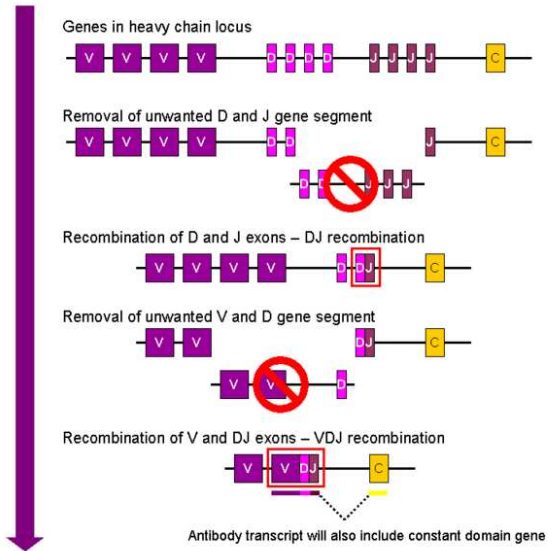
- A common autosomal dominant disorder
- Sporadic and hereditary mutations [6]
- Deletions and insertions prevalent [1]

B - Cell Formation

Key stages in B cell development and differentiation



Motivation: B-Cell Formation



The Deletion-Insertion (DI) Model

DI Genome:

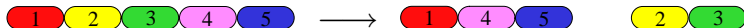


The Deletion-Insertion (DI) Model

DI Genome:



Deletion:

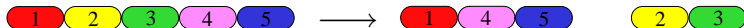


The Deletion-Insertion (DI) Model

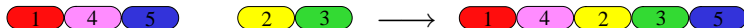
DI Genome:



Deletion:



Insertion:

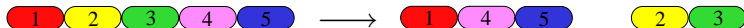


The Deletion-Insertion (DI) Model

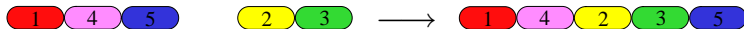
DI Genome:



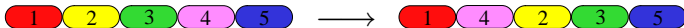
Deletion:



Insertion:



Deletion-Insertion:



The Deletion-Insertion (DI) Model

Definition (Deletion-Insertion Distance)

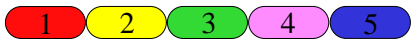
$d_{DI}(\rho, \sigma)$: The minimum number of deletion-insertions required to transform genome (permutation) ρ into genome (permutation) σ .

The Deletion-Insertion (DI) Model

Definition (Deletion-Insertion Distance)

$d_{DI}(\rho, \sigma)$: The minimum number of deletion-insertions required to transform genome (permutation) ρ into genome (permutation) σ .

Example:

 12345

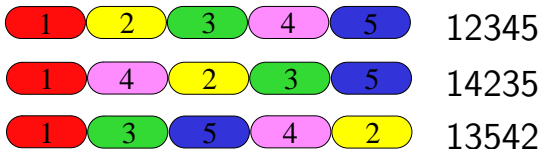
 13542

The Deletion-Insertion (DI) Model

Definition (Deletion-Insertion Distance)

$d_{DI}(\rho, \sigma)$: The minimum number of deletion-insertions required to transform genome (permutation) ρ into genome (permutation) σ .

Example:



A Non-Trivial Example

12345678 \rightarrow 87654321

A Non-Trivial Example

12345678 \rightarrow 87654321

A Non-Trivial Example

12345678 \rightarrow 87654321
43218765

A Non-Trivial Example

12345678 \rightarrow 87654321
43218765

A Non-Trivial Example

12345678 \rightarrow 87654321
43218765
18432765

A Non-Trivial Example

12345678 \rightarrow 87654321
43218765
18432765

A Non-Trivial Example

12345678 \rightarrow 87654321
43218765
18432765
12784365

A Non-Trivial Example

12345678 \rightarrow 87654321
43218765
18432765
12784365

A Non-Trivial Example

12345678 \rightarrow 87654321
43218765
18432765
12784365
12367845

A Non-Trivial Example

12345678 \rightarrow 87654321
43218765
18432765
12784365
12367845

A Non-Trivial Example

12345678 \rightarrow 87654321
43218765
18432765
12784365
12367845
12345678

Remark

Let ρ and σ be two genomes defined on N genes, let τ be another permutation, and let $\iota = 01 \dots N$ be the identity genome. Then,

Remark

Let ρ and σ be two genomes defined on N genes, let τ be another permutation, and let $\iota = 01 \dots N$ be the identity genome. Then,

- $d_{DI}(\rho, \sigma) = d_{DI}(\tau \circ \rho, \tau \circ \sigma),$

Remark

Let ρ and σ be two genomes defined on N genes, let τ be another permutation, and let $\iota = 01 \dots N$ be the identity genome. Then,

- $d_{DI}(\rho, \sigma) = d_{DI}(\tau \circ \rho, \tau \circ \sigma),$
- $d_{DI}(\rho, \sigma) = d_{DI}(\iota, \rho^{-1} \circ \sigma),$

Remark

Let ρ and σ be two genomes defined on N genes, let τ be another permutation, and let $\iota = 01 \dots N$ be the identity genome. Then,

- $d_{DI}(\rho, \sigma) = d_{DI}(\tau \circ \rho, \tau \circ \sigma),$
- $d_{DI}(\rho, \sigma) = d_{DI}(\iota, \rho^{-1} \circ \sigma),$
- $d_{DI}(\iota, \rho) = d_{DI}(\rho^{-1}, \iota).$

Remark

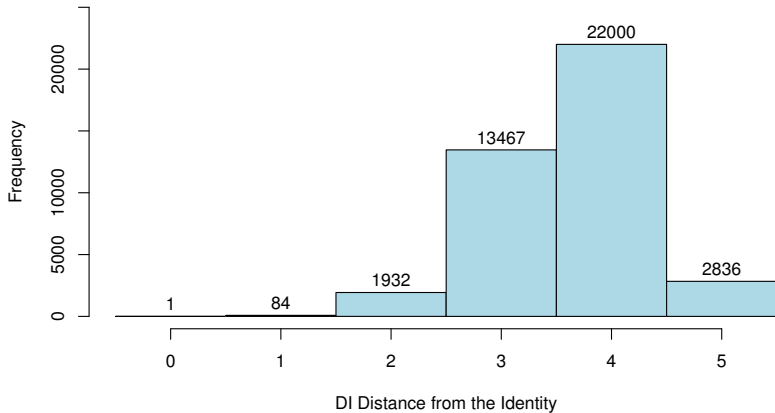
Let ρ and σ be two genomes defined on N genes, let τ be another permutation, and let $\iota = 01 \dots N$ be the identity genome. Then,

- $d_{DI}(\rho, \sigma) = d_{DI}(\tau \circ \rho, \tau \circ \sigma),$
- $d_{DI}(\rho, \sigma) = d_{DI}(\iota, \rho^{-1} \circ \sigma),$
- $d_{DI}(\iota, \rho) = d_{DI}(\rho^{-1}, \iota).$

Thus, $d_{DI}(\rho) := d_{DI}(\iota, \rho)$ is a permutation statistic.

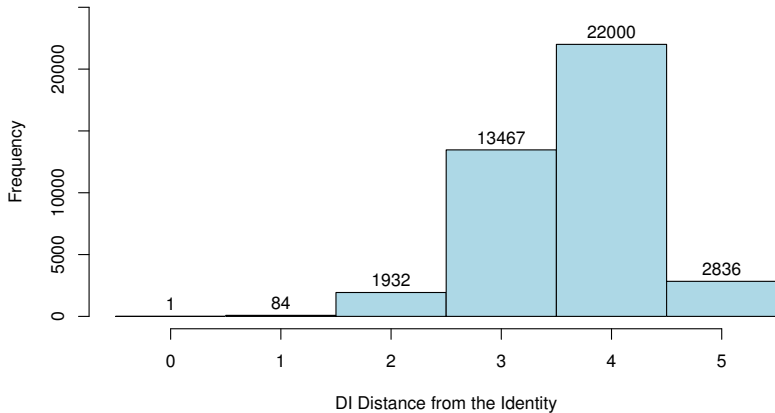
DI Distance as a Permutation Statistic

**Distribution of DI Distance
(8 genes)**



DI Distance as a Permutation Statistic

**Distribution of DI Distance
(8 genes)**



How does DI distance compare with other permutation statistics?

We denote the number of descents in a permutation σ by $des(\sigma)$.

Notation

We denote the number of descents in a permutation σ by $des(\sigma)$.

$lic(\sigma)$ and $ldc(\sigma)$ refer to the length of a longest increasing chain and a longest decreasing chain in σ , respectively.

Notation

We denote the number of descents in a permutation σ by $\text{des}(\sigma)$.

$\text{lic}(\sigma)$ and $\text{ldc}(\sigma)$ refer to the length of a longest increasing chain and a longest decreasing chain in σ , respectively.

Definition (casc)

$\text{casc}(\sigma)$ is the number of occurrences of a consecutive 12 pattern in $0 \sigma n+1$.

We denote the number of descents in a permutation σ by $\text{des}(\sigma)$.

$\text{lic}(\sigma)$ and $\text{ldc}(\sigma)$ refer to the length of a longest increasing chain and a longest decreasing chain in σ , respectively.

Definition (casc)

$\text{casc}(\sigma)$ is the number of occurrences of a consecutive 12 pattern in $0 \sigma n+1$.

Example:

$$\sigma = 134265$$

We denote the number of descents in a permutation σ by $des(\sigma)$.

$lic(\sigma)$ and $ldc(\sigma)$ refer to the length of a longest increasing chain and a longest decreasing chain in σ , respectively.

Definition ($casc$)

$casc(\sigma)$ is the number of occurrences of a consecutive 12 pattern in $0 \sigma n+1$.

Example:

$$\sigma = 134265$$

$$des(\sigma) = 2$$

$$lic(\sigma) = 4$$

$$ldc(\sigma) = 2$$

We denote the number of descents in a permutation σ by $des(\sigma)$.

$lic(\sigma)$ and $ldc(\sigma)$ refer to the length of a longest increasing chain and a longest decreasing chain in σ , respectively.

Definition (casc)

$casc(\sigma)$ is the number of occurrences of a consecutive 12 pattern in $0 \sigma n+1$.

Example:

$$\sigma = 134265$$

$$des(\sigma) = 2$$

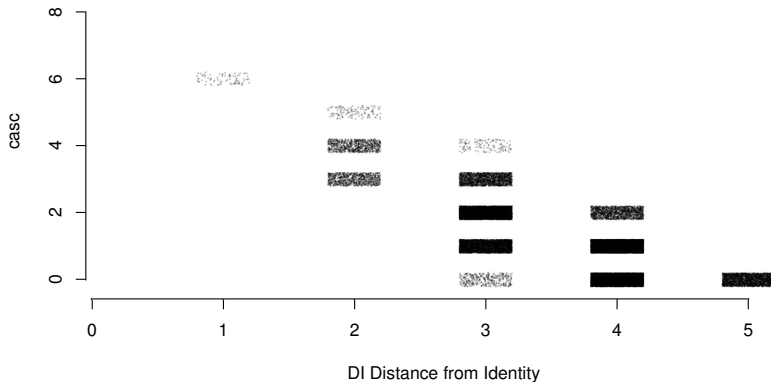
$$lic(\sigma) = 4$$

$$ldc(\sigma) = 2$$

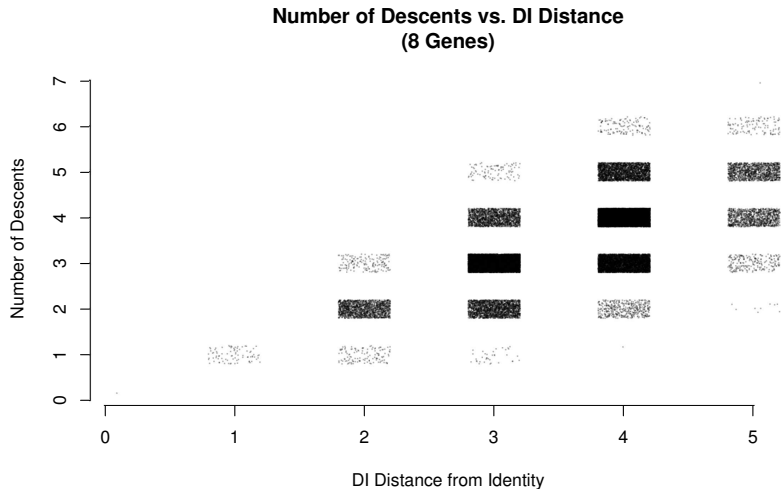
$$casc(\sigma) = 2$$

DI Distance as a Permutation Statistic

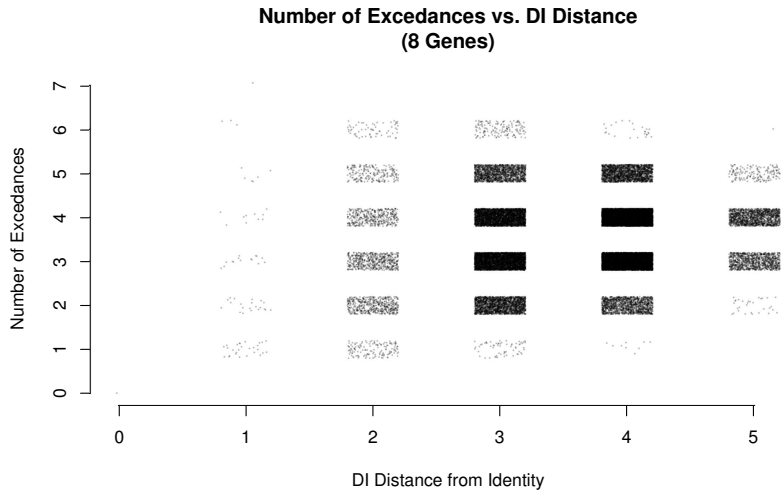
casc vs. DI Distance (8 Genes)



DI Distance as a Permutation Statistic

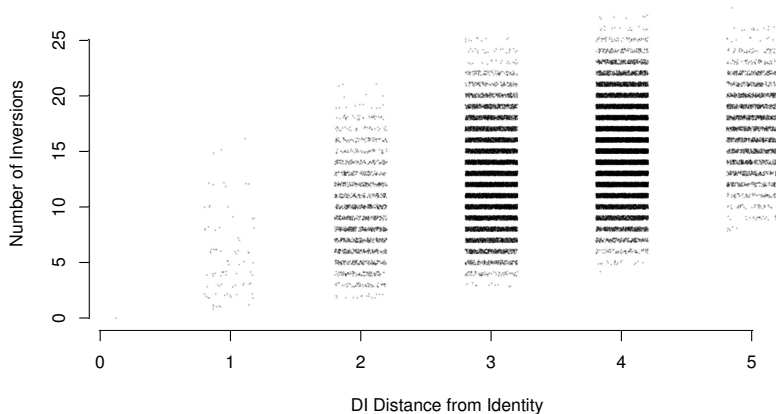


DI Distance as a Permutation Statistic



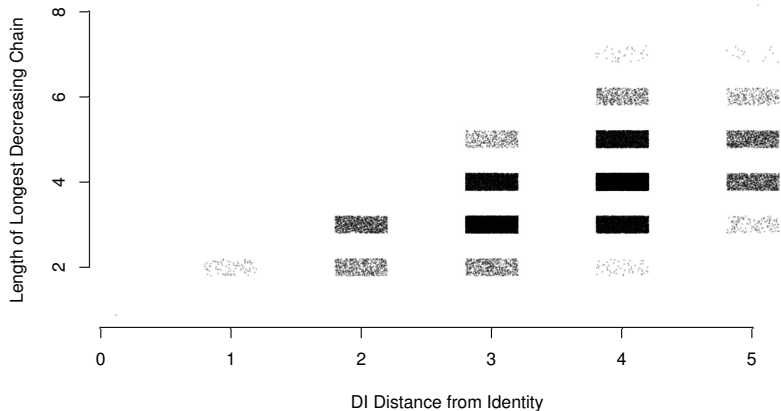
DI Distance as a Permutation Statistic

Number of Inversions vs. DI Distance
(8 Genes)



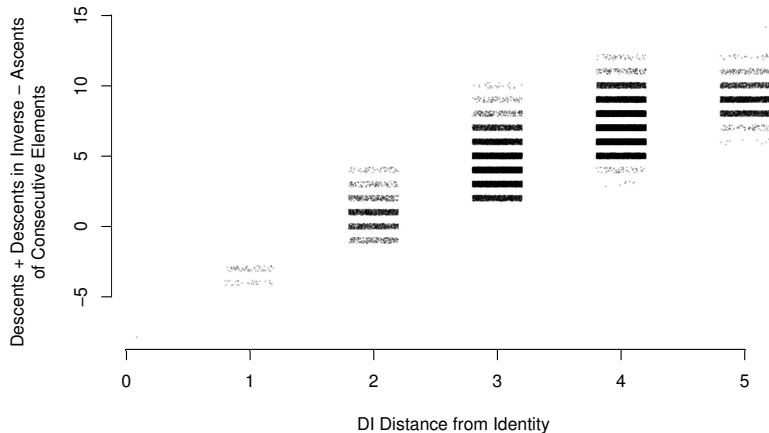
Other Permutation Statistics and DI Distance

**Length of Longest Decreasing Chain vs. DI Distance
(8 Genes)**



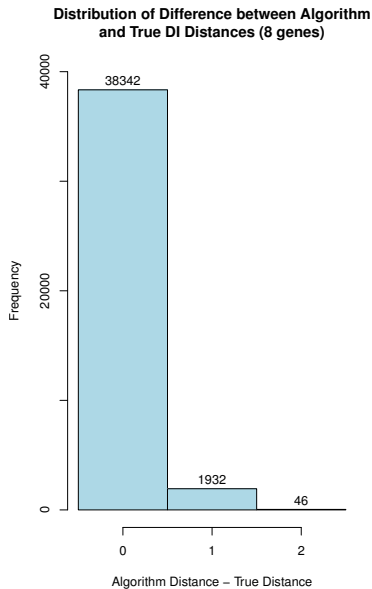
DI Distance as a Permutation Statistic

Descents + Descents in Inverse – Ascents of Consecutive Elements vs. DI Distance (8 Genes)



Idea:

- Minimize $des(\rho)$, $des(\rho^{-1})$, $ldc(\rho)$
- Maximize $casc(\rho)$



Lower Bounds:

$$d_{DI}(\rho) \geq \left\lfloor \frac{des(\rho)}{2} \right\rfloor + \chi(\rho \neq \iota)$$

$$d_{DI}(\rho) \geq \left\lfloor \frac{l dc(\rho)}{2} \right\rfloor + \chi(l dc(\rho) > 2)$$

Lower Bounds:

$$d_{DI}(\rho) \geq \left\lfloor \frac{des(\rho)}{2} \right\rfloor + \chi(\rho \neq \iota)$$

$$d_{DI}(\rho) \geq \left\lfloor \frac{ldc(\rho)}{2} \right\rfloor + \chi(ldc(\rho) > 2)$$

Upper Bounds:

$$d_{DI}(\rho) \leq \text{algorithm}(\rho)$$

$$d_{DI}(\rho) \leq n - lic(\rho)$$

$$d_{DI}(\rho) \leq n - casc(\rho) + 1$$

Bounds on DI Distance (for ρ of length n)

Lower Bounds:

$$d_{DI}(\rho) \geq \left\lfloor \frac{des(\rho)}{2} \right\rfloor + \chi(\rho \neq \iota)$$

$$d_{DI}(\rho) \geq \left\lfloor \frac{ldc(\rho)}{2} \right\rfloor + \chi(ldc(\rho) > 2)$$

Upper Bounds:

$$d_{DI}(\rho) \leq \text{algorithm}(\rho)$$

$$d_{DI}(\rho) \leq n - lic(\rho)$$

$$d_{DI}(\rho) \leq n - casc(\rho) + 1$$

* These bounds also apply to ρ^{-1} since $d_{DI}(\rho) = d_{DI}(\rho^{-1})$.

Results for 8 Genes:

Lower Bound = Dist	Upper Bound = Dist	Neither Bound = Dist
8726	38962	809
21.6 %	96.6%	2.0%

Continuing Research

- Represent genomes with other structures
(i.e. trees, lattice paths)
- Find a formula for DI Distance as a function of permutation statistics
- Consider gene repeats

Merci - Thank You!

- Lara Pudwell, Alex Burstein, and Alex Woo
- Permutation Patterns 2013 Organizers
- Dr. Manda Riehl - University of Wisconsin-Eau Claire
- Dr. Abra Brisbin - University of Wisconsin-Eau Claire
- UWEC Department of Mathematics
- UWEC Office of Research and Sponsored Programs

References

- [1] E. Ars, E. Serra, J. Garcia, H. Kruyer, A. Gaona, C. Lazaro, X. Estivill, *Mutations Affecting mRNA Splicing are the Most Common Molecular Defects in Patients with Neurofibromatosis Type 1*, Human Molecular Genetics 9, (2000) 237–247.
- [2] Human Genome Program, U.S. Department of Energy, Human Genome Program Report, 1997. http://www.ornl.gov/sci/techresources/Human_Genome/publicat/97pr/index.html
- [3] K. Lindblad-Toh et al. *Genome sequence, comparative analysis and haplotype structure of the domestic dog*. Nature, (2005): 803–819.
- [4] National Human Genome Research Institute. NIH News Advisory, December 2002. <http://www.genome.gov/page.cfm?pageID=10005831>.
- [5] http://nfs.unipv.it/nfs/minf/dispense/immunology/lectures/files/bcell_tcell_development.html
- [6] R. Rubin, D. Strayer, *Rubin's Pathology: Clinico-pathologic Foundation of Medicine (5 ed.)*, Wolters Kluwer Health: Lippincot Williams & Wilkins, Baltimore, Maryland, (2008). 201-203.