# Imputation

Matt Spangler

University of Nebraska-Lincoln

# Imputation

- Imputation creates data that were not actually collected
- Imputation allows us to retain observations that would otherwise be left out of an analysis

# Imputation—Not New

- Imputation is common when data are analyzed with regression analysis
  - Like genome-wide association studies
- Observations only included in the model if they have values for all variables (e.g. SNP)
- If there are a lot of variables (e.g. SNP)the probability of missing values increases
  - If we have 50,000 SNP there is a very good chance some of them are missing or “not called”

# General Methods for Imputation

- Random
  - Assigns values randomly based on a desired distribution
- Deterministic
  - Assigns values based on existing knowledge
    - Mean of all non-missing SNP at a locus from a population
    - Missing values replaced with a set of likely values

# Example

| Animal | Herd | Sex | Birth Weight | Calving Difficulty |
|---|---|---|---|---|
| 1 | Mine | M | 100 | 3 |
| 2 | Mine | F | 80 | 2 |
| 3 | Yours | M | 75 | 1 |
| 4 | Yours | F | 70 | 1 |
| 5 | Mine | M | 115 | ? |
| 6 | Mine | F | 90 | 2 |
| 7 | Yours | M | 70 | 1 |
| 8 | Yours | M | 65 | ? |

# Example

| Animal | Herd | Sex | Birth Weight | Calving Difficulty |
|---|---|---|---|---|
| 1 | Mine | M | 100 | 3 |
| 2 | Mine | F | 80 | 2 |
| 3 | Yours | M | 75 | 1 |
| 4 | Yours | F | 70 | 1 |
| 5 | Mine | M | 115 | 3 |
| 6 | Mine | F | 90 | 2 |
| 7 | Yours | M | 70 | 1 |
| 8 | Yours | M | 65 | 1 |

# Imputation--Genomics

- Method of assigning missing genotypes based on actual genotypes of related animals
- Requires relatives to have been genotyped with higher density assays
  - More animals genotyped with a higher density
  - Closely related animals genotyped with higher density
- Allows other animals (younger) to be genotyped with lower density (cheaper) assays

# Imputation--Genomics

- Based on dividing the genotype into individual chromosomes (maternal and paternal contributions)
- Missing SNP assigned by tracking inheritance from ancestors and/or descendants
- Allows for the merger of various SNP densities by imputing SNP not on LD (or non-target) panels.

# SNP Densities Not Static

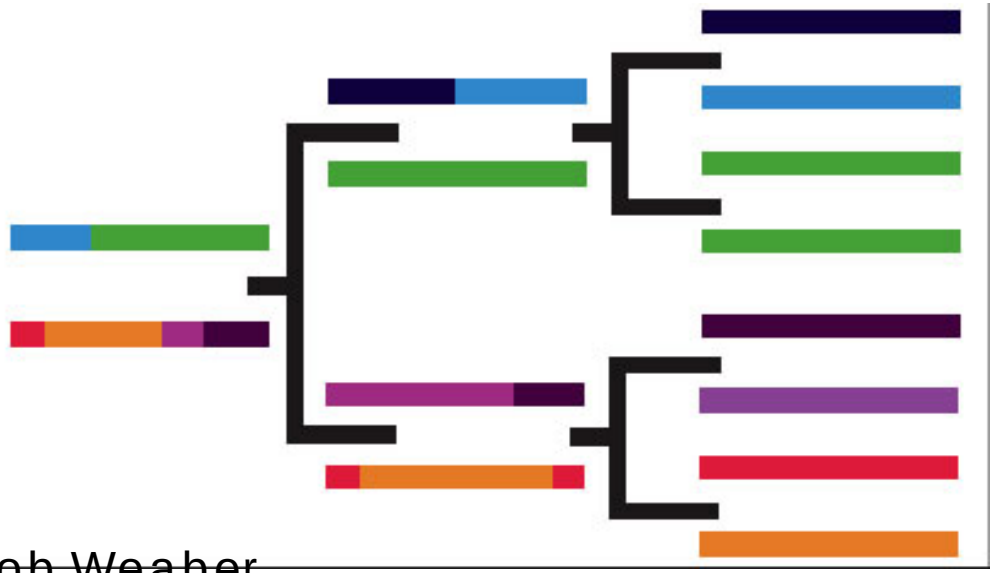
- 50K was/is the backbone
- HD; 770K
- GGP-HD; 9K (8K in common with 50K)
- GGP-LD; 77K (27K in common with 50K)
- More changes coming
- Current approaches to MBV require common SNP density

# Practical Need

- Alternative is to genotype every animal at high-density
- As SNP platforms change, animals would have to be re-genotyped
- Allows two opportunities:
  - Going from lower density to higher density
    - Allow wide-spread use of cheaper panels
  - Going from higher density to lower density
    - Allows use of large training sets built on the lower density
    - First use of imputation in NCE

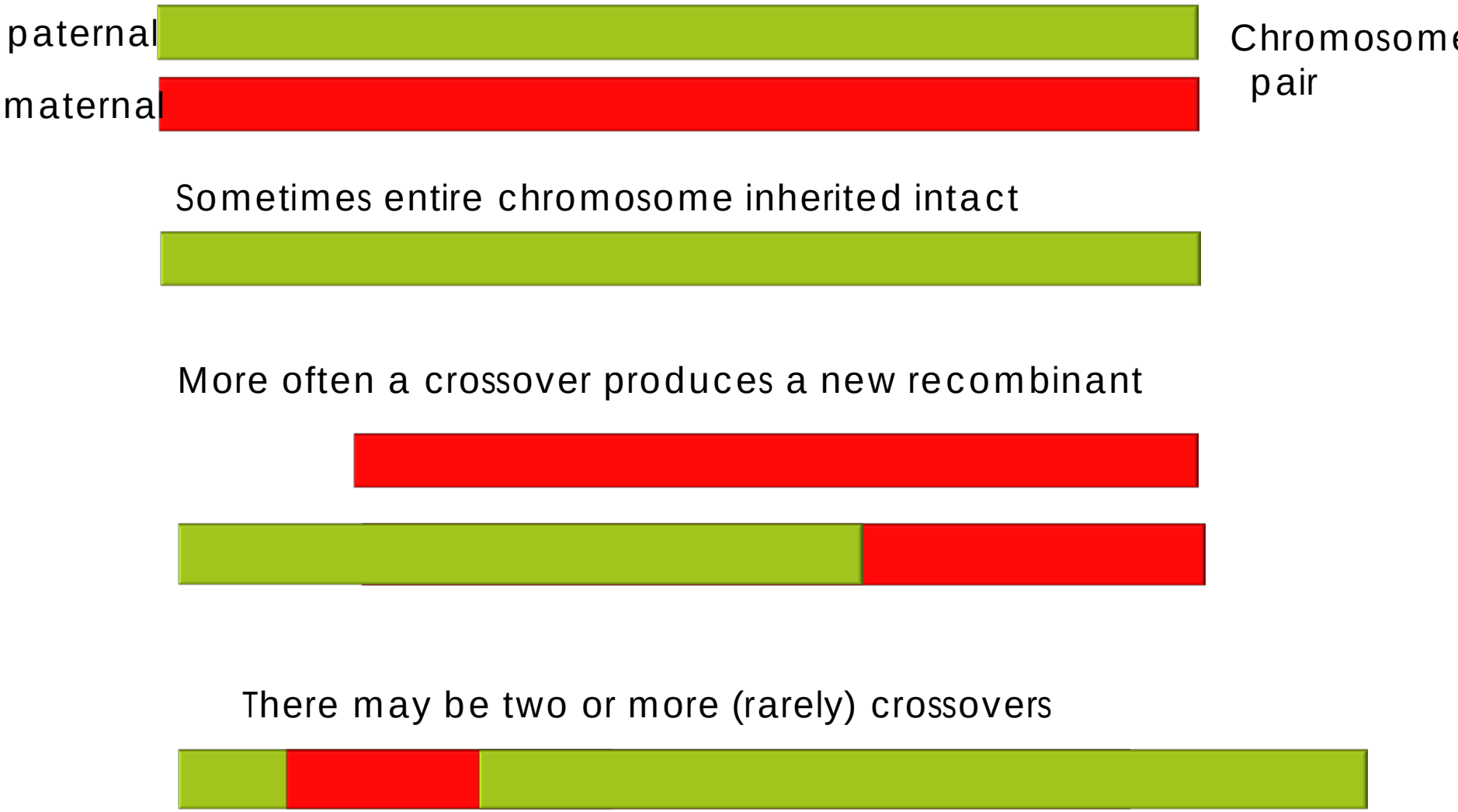
# Linkage

- The tendency of certain loci to be inherited together
- Loci that are close to each other on chromosome tend to stay together during meiosis
- Crossing over (recombination) breaks up linkage.



Slide Courtesy of Bob Weaber

# “Blocks” of Alleles Inherited Together



# Inheritance Example


Consider a small window of 1 Mb (1% of the genome)

Example Adapted from Garrick

# Single Parent Commonality in Small Sections

# Imputation

**Sire (HD)**

…..TCACCGCTGAG…..

…..CAGATAGGATT…..

**Offspring (LD)**

…..??G??????A??….

…..??T??????T??…..

**Offspring (Imputed)**

….CAGATAGGATT…..

…..??T??????T??…..

# Imputation—Parentage Check

**Sire (HD)**

…..TCACCGCTGAG…..

…..CAGATAGGATT…..

**Offspring (LD)**

…..??C??????T??….

…..??T??????T??…..

# Imputation

**MGS**

**Sire (HD)**

**Dam**

…..TCACCGCTGAG…..

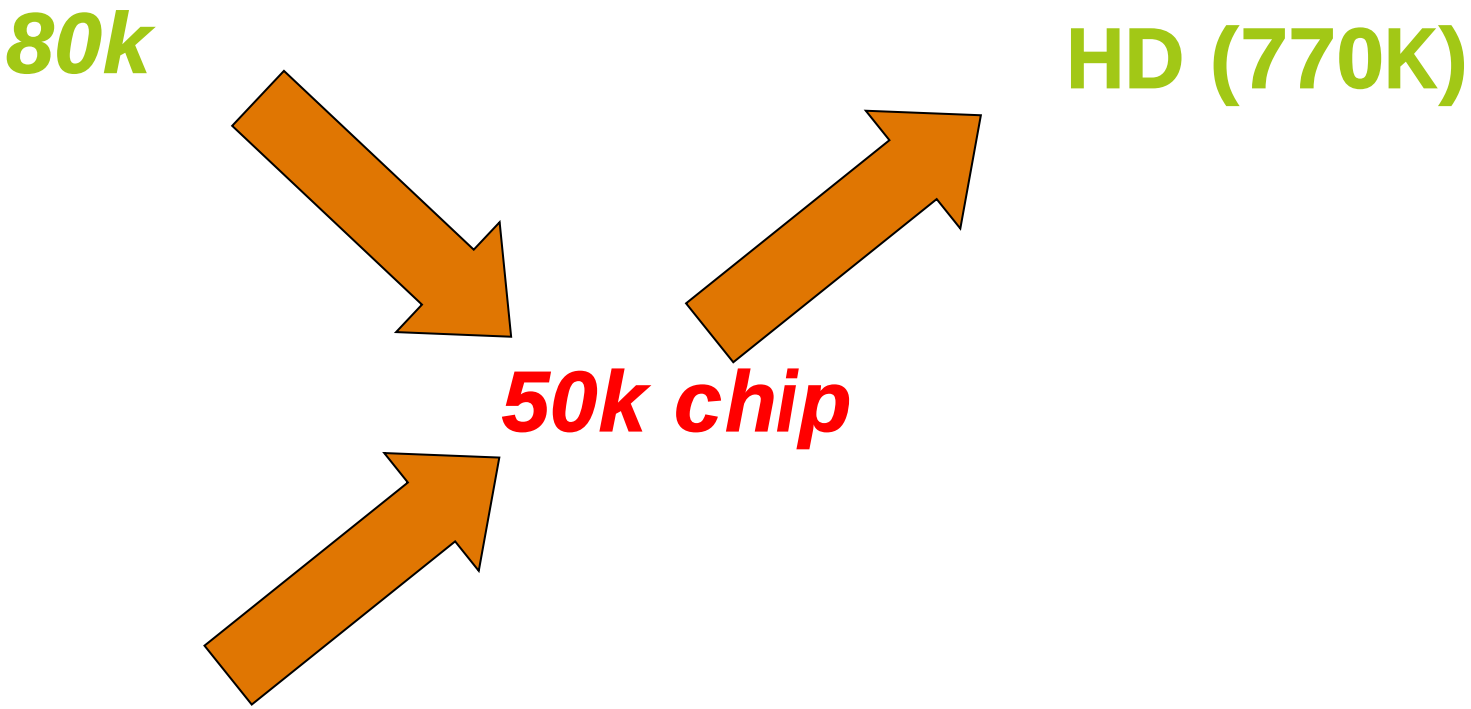
…..CAGATAGGATT…..

**Offspring (LD)**

…..??G??????A??….

…..??T??????T??…..

# Imputation Scenarios




All seek to reduce cost
--Genotyping animals with cheaper panels
--Leveraging existing training sets

# HEREFORD DATA (N=1,081)

| Trait | 50K | Imputed 50K (GGP-LD) |
|---|---|---|
| BWT | 0.44 | 0.42 |
| WWT | 0.40 | 0.38 |
| YWT | 0.20 | 0.19 |
| MILK | 0.42 | 0.40 |
| MARB | 0.27 | 0.26 |
| REA | 0.26 | 0.25 |

Saatchi et al., 2013

# HEREFORD DATA (N=1,081)

| Trait | 50K | Imputed 50K (GGP-HD) |
|---|---|---|
| BWT | 0.44 | 0.40 |
| WWT | 0.40 | 0.40 |
| YWT | 0.20 | 0.20 |
| MILK | 0.42 | 0.41 |
| MARB | 0.27 | 0.26 |
| REA | 0.26 | 0.25 |

Saatchi et al., 2013

# Practical Concerns

- When to shift to LD panels
  - Must develop adequate training population (50K or 80K) first
- After migration, how to get HD to allow for imputation in the future
  - Genotype moderate to high accuracy sires at higher density
  - Without this imputation accuracy will erode overtime
  - Breeds subsidize this in some form

# Optimize Future HD genotyping?

- Likely to be ad-hoc right now
  - Once a sire reaches x accuracy regenotype with higher density
- Need to move forward with an optimal approach
  - Ex: Does it make sense to do full sibs?
  - Conditional on limited resources ($) choice matters
  - Impute variants from sequence into existing SNP platforms

# Conclusions

- Imputation works and can decrease the cost of genotyping
- Does not negate the need for a training set with higher density
- Plan must be in place to ensure imputation accuracy persists overtime
  - Add “new” animals with higher density genotypes

# Thank You!

- http://beef.unl.edu
- www.nbcec.org
- www.beefefficiency.org

UNIVERSITY OF Nebraska Lincoln