

STRUCTURE

February 9, 2007

Matthew Hufford

Kraig Kraft

Sarah Brown

Two common approaches in the study of populations:

1. Investigate the evolutionary relationships of modern populations
 2. Start with predefined populations and try to classify individuals of unknown origin
- Both approaches involve the crucial first step of defining a set of populations

What is a population?

Classic definition from Population Biology:

“A collection of individuals belonging to the same species, living in the same area”

- Could be subjective
- May not be accurate in genetic terms

Cryptic Population Structure

- Definition: population structure that cannot be detected based on physical characters
- When is it of interest?
 1. Conservation Genetics
 2. Association mapping to find disease genes
 3. DNA fingerprinting for forensics

Distance-based clustering methods:

1. Begin by calculating a pair-wise distance matrix with entries that list the distance between every pair of individuals
2. Matrix can be visualized using an illuminating graphical representation such as a tree or a multidimensional scaling plot
3. Clusters are identified by eye, not statistically or with probability

Distance-based methods in Genetics:

1. Multilocus genotype data often clustered using distance-based algorithms such as neighbor-joining
2. Problems with this approach:
 - a) Clusters identified are heavily dependent on the distance measure and graphical representation chosen
 - b) Cannot assess confidence in clusters
 - c) Difficult to incorporate additional information such as sampling location of individuals.
3. These methods are useful for exploratory data analysis rather than to find statistical inference.

Model-based clustering methods:

1. Observations from each cluster are random draws from a parametric model
2. Simultaneously infer parameters for each cluster and cluster membership of each individual using statistical methods such as maximum-likelihood or Bayesian methods

Frequentist versus Bayesian Inference

TABLE 1. Some fundamental differences between frequentist and Bayesian statistical inference in their uses and interpretations of statistical concepts and terms.

Concept or term	Frequentist interpretation	Bayesian interpretation
Probability	Result of an infinite series of trials conducted under identical conditions	The observer's degree of belief, or the organized appraisal in light of the data
Data	Random (representative) sample	Fixed (all there is)
Parameters	Fixed	Random
$k\%$ confidence interval	This interval will include the true value of a given parameter in $k\%$ of all possible samples	$k\%$ of the possible parameter values will fall within the confidence (credibility) interval
Treatment of nuisance parameters	Conditions on sufficient statistics or maximum likelihood estimate	Integrates over all possible values
Conclusion	$P(x H)$	$P(H x)$

What's so great about Bayesian Inference?

- Can incorporate background information into the specification of a model
- Can test multiple hypotheses simultaneously
- Can express a level of certainty in the alternative hypotheses

Bayesian Inference

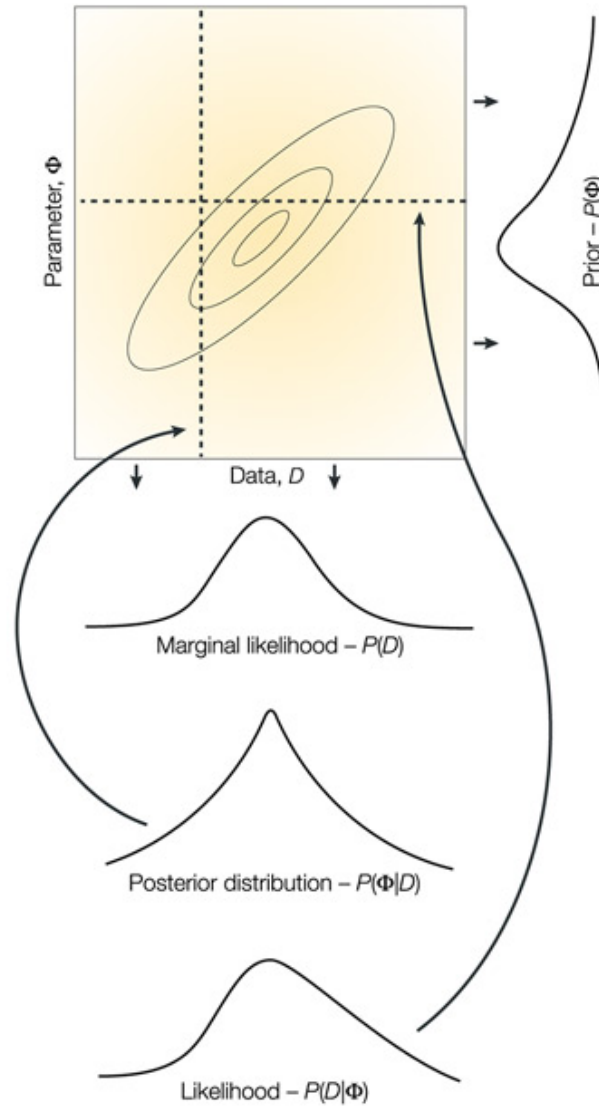
- **Bayesian Viewpoint:** There is no logical distinction between model parameters and data—both random variables with a joint probability distribution specified by a probabilistic model
- Data observed, Parameters unobserved

The Bayesian Players

1. *The Prior*: encapsulates, as a probability distribution, information about the values of a parameter before examining the data
 2. *The Likelihood*: conditional distribution that specifies the probability of the data given values of parameters—based on model of underlying process
- Together these two combine all information about parameters and produce the *Joint Distribution*

Aim of Bayesian Inference

- Calculate the *Posterior Distribution* of the parameters, which is the conditional distribution of parameters given the data
- Parameter is typically estimated as the mode or the mean of the posterior distribution



Nature Reviews | Genetics

Nature Reviews Genetics 5, 251-261 (2004)
THE BAYESIAN REVOLUTION IN GENETICS

		Data (observed variables)						
		Genotype A			Genotype B			
		Likelihood	Joint distribution	Posterior probability	Likelihood	Joint distribution	Posterior probability	Prior probability
Parameters (unobserved variables)	Immigrant	0.01		0.0012	0.99		0.69	
			0.001			0.099		0.1
	Resident	0.95		0.9988	0.05		0.31	
			0.855			0.045		0.9
Probability of data			0.856			0.144		1

Nature Reviews | **Genetics**

Nature Reviews Genetics 5, 251-261 (2004)
THE BAYESIAN REVOLUTION IN GENETICS

Uses of Bayesian Inference in Population Genetics

1. To make inferences about non-identifiable parameters (e.g., μN_e)
2. Can both partition individuals into subpopulations and assign individual migrant histories
3. Can jointly infer subpopulations within a larger population and estimate traditional fixation indices among and within the subpopulations
4. Can infer short-term migration rates using multi-locus genotypes.
5. Can infer historical changes in populations size
6. Deal with genetic data that are taken at different times, allowing for population growth
7. Identification of SNPs: can use software package, PolyBayes to identify SNPs. Bayesian method is used to eliminate false positives that result from paralogous sequences and sequencing errors.

STRUCTURE

- Uses Bayesian Inference to identify subpopulations and probabilistically assign individuals
- Assigns based on genotype while simultaneously determining allele frequencies
- Allows for presence of admixed individuals
- Various types of markers can be used: microsatellites, RFLPs, SNPs

STRUCTURE Assumptions

1. Assumes Hardy-Weinberg Equilibrium (HWE) within populations
 2. Assumes loci are unlinked and at linkage equilibrium in populations
- Model achieves HWE by introducing population structure and attempts to find population groupings that are not in disequilibrium

Components of Model

1. The Data:

X = genotypes of sampled individuals

2. The Parameters of Interest/Hypotheses:

Z = the unknown population of origin of the individuals

P = the unknown allele frequencies in all the populations

Q = the admixture proportions for each individual

K = the number of cluster

The Process of Inference

1. Obtain a sample of each parameter from the posterior distribution given X using Markov Chain Monte Carlo Method
2. A point estimate of this information is obtained via the mode of the posterior
3. The posterior distribution of K is “peculiarly dependent” on priors and assumptions and K is chosen in an *ad hoc* fashion that is “dubious” but seems to work. Yikes!

Applications of STRUCTURE: Simulated Data

- Different population models
- Different #'s of microsatellite loci
- Tested K's from 1-5
- Determined burn-in length based on summary statistic stationarity

Simulated Data: Population Models

Model 1: Single random-mating population

Model 2: Two random-mating populations
split from a common ancestor with no
subsequent migration

Model 3: Two discrete populations fused to
produce a single random-mating population
and allowed to mate for two generations

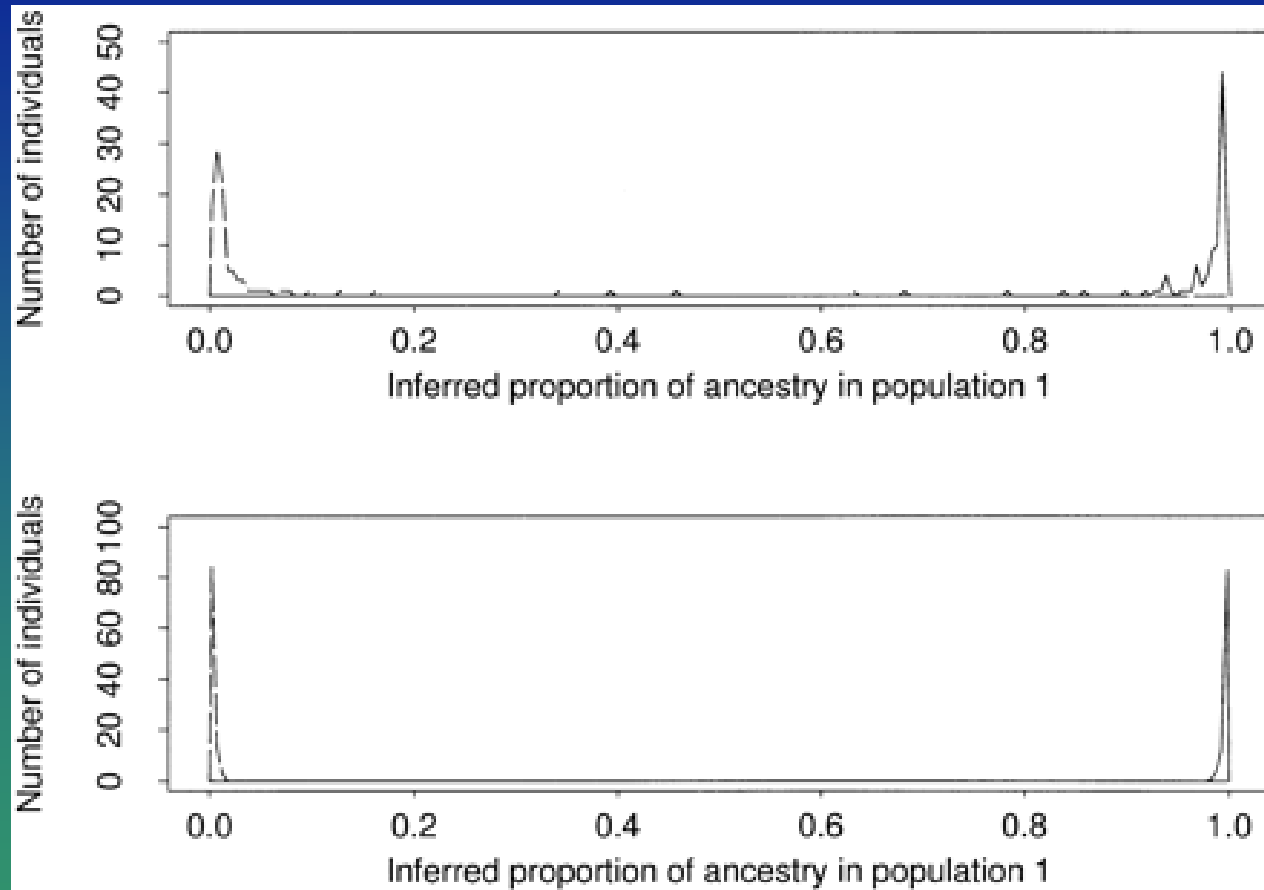
Simulated Data: Posterior Probabilities of K

Table 1. Estimated posterior probabilities of K , for simulated data sets 1, 2A, 2B, and 3 (denoted X_1 , X_{2A} , X_{2B} , and X_3 , respectively)

K	$\log P(K X_1)$	$P(K X_{2A})$	$P(K X_{2B})$	$P(K X_3)$
1	~1.0	~0.0	~0.0	~0.0
2	~0.0	0.21	0.999	~1.0
3	~0.0	0.58	0.0009	~0.0
4	~0.0	0.21	~0.0	~0.0
5	~0.0	~0.0	~0.0	~0.0

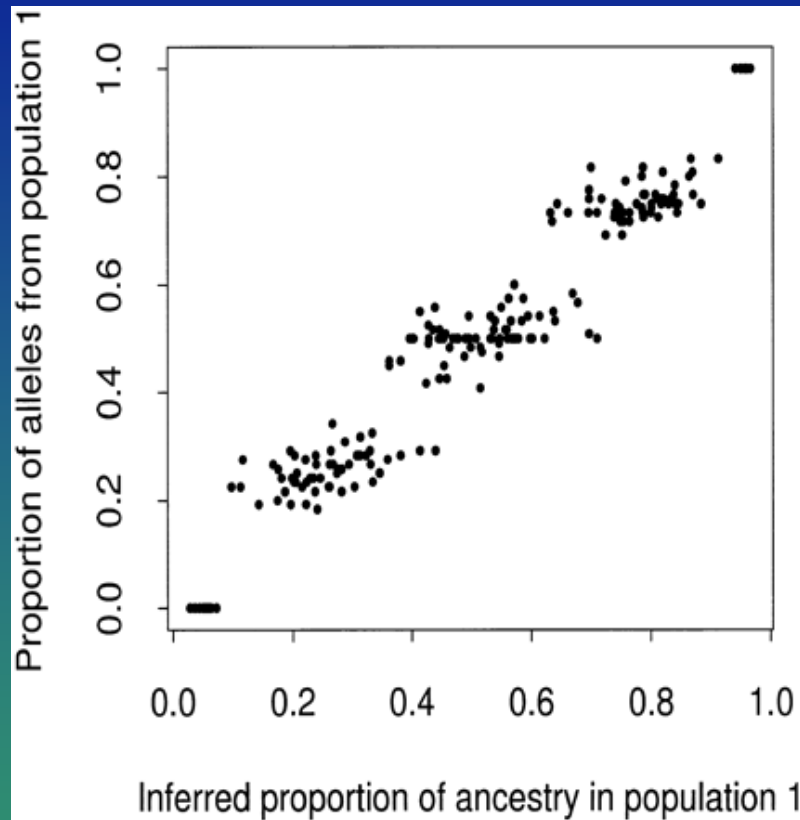
- 2A: 5 loci, 2B: 15 loci
- 3: Q accounts for admixture

Assignment of Individuals



- Performs well with discrete populations (Model 2)
- Very well with 5 loci, near perfect for 15 loci

Assignment of Individuals



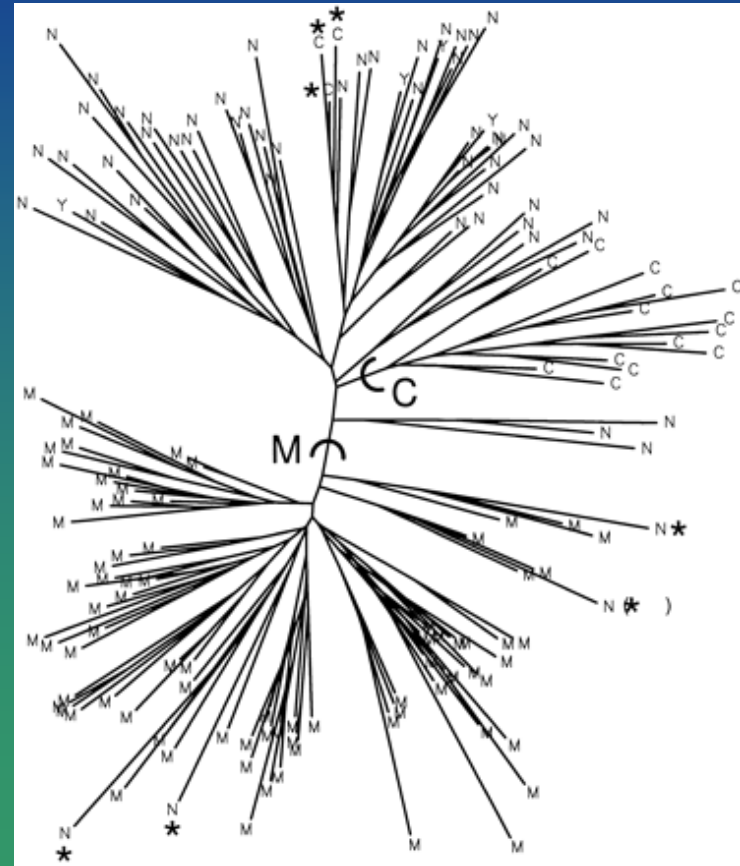
- From left to right: 0, 1, 2, 3, 4 grandparents in population 1
- Hard to get accurate measurement of q for highly admixed populations

Application to Populations of Taita Thrush

- Individuals sampled at 4 locations in Kenya
- Each individual genotyped at 7 loci
- Populations in fragments of indigenous cloud forest anthropogenically fragmented
- Low migration rate indicated in radio-tagging and ringing experiments

Neighbor-Joining Tree of Thrush Data

- Fairly distinct clusters
- Asterisked individuals appear to be classified with other groups
- One of the asterisked individuals (*) was identified as a migrant by the STRUCTURE algorithm
- Weaknesses?

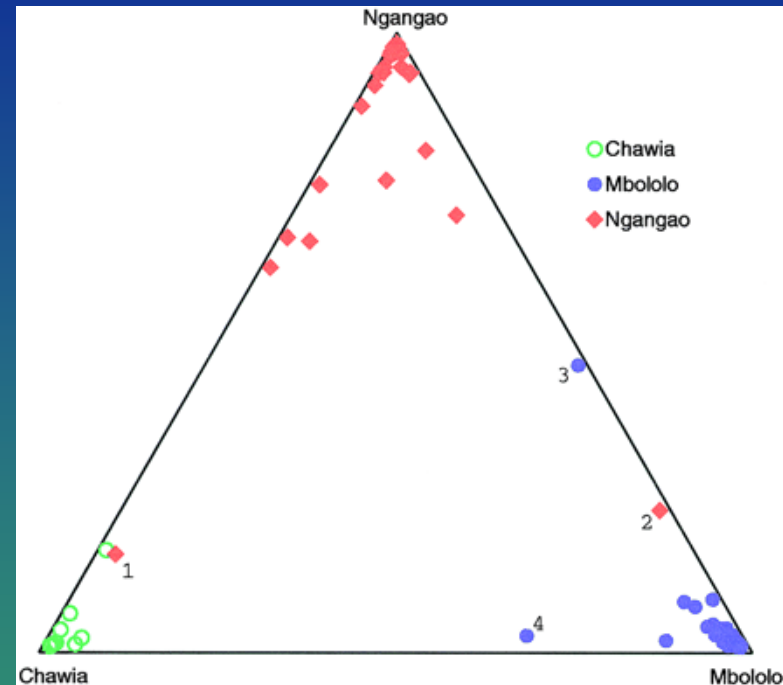


STRUCTURE Analysis of Data

Table 3. Inferring the value of K , the number of populations, for the *T. helleri* data

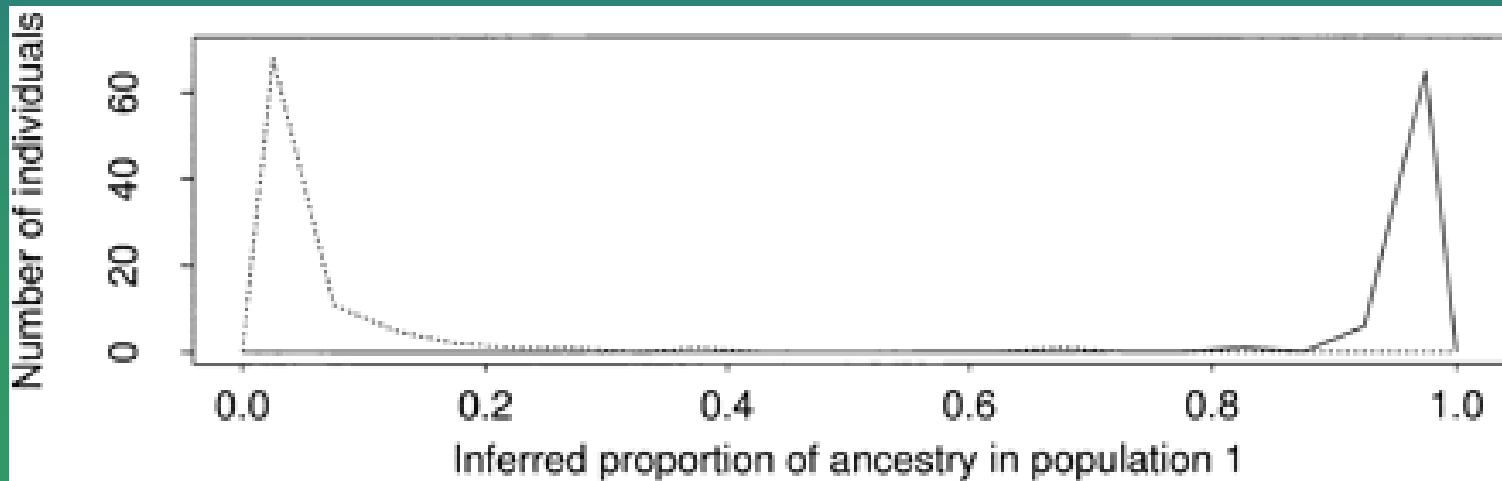
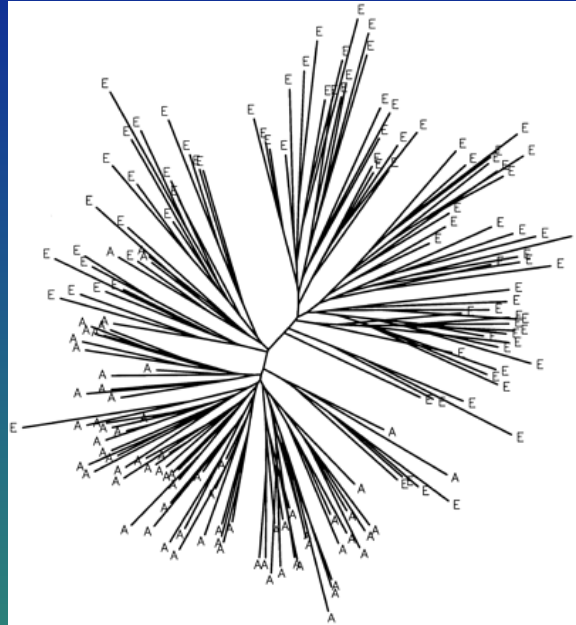
K	$\log P(X K)$	$P(K X)$
1	-3144	~ 0
2	-2769	~ 0
3	-2678	0.993
4	-2683	0.007
5	-2688	0.00005

The values in the last column assume a uniform prior for K ($K \in \{1, \dots, 5\}$).



- Yale population (4 individuals) assigned to Ngangao
- #’d individuals outside of obvious clusters

Application to European and African Populations



Incorporation of Geographic Information into the Thrush Data

Table 4. Testing whether particular individuals are immigrants or have recent immigrant ancestors

Individual	Geographic origin	Possible source	ν	No immigrant ancestry	Immigrant	Immigrant parent	Immigrant grandparent
1	Ngangao	Chawia	0.05	0.869	0.008	0.052	0.063
			0.10	0.739	0.019	0.106	0.123
2	Ngangao	Mbololo	0.05	0.673	0.029	0.126	0.168
			0.10	0.472	0.046	0.203	0.273
3	Mbololo	Ngangao	0.05	0.649	0.002	0.179	0.165
			0.10	0.464	0.003	0.271	0.253
4	Mbololo	Chawia	0.05	0.891	0.000	0.007	0.082
			0.10	0.791	0.000	0.014	0.157

- ν : probability that an individual is an immigrant from population $g^{(i)}$
- Moderate possibility 2 & 3 are immigrants
- Sensitivity to ν indicates a need for more loci

Points from Discussion:

Accuracy of Assignment Depends on:

1. # of Individuals (affects P)
2. # of Loci (affects Q)
3. The Amount of Admixture
4. Extent of Allele-Frequency Differences Among Populations

Points from Discussion:

STRUCTURE will be useful for:

1. Situations where there is little information about population structure
2. Elucidating cryptic population structure (i.e, ensure that populations defined on extrinsic basis reflect underlying genetic structure)

Points from Discussion:

Ways to modify/improve the model:

1. Incorporate more prior information
2. Change the way allele frequencies (P) are estimated (current method might group subpopulations)
3. Allow linkage among loci (especially when there is recent admixture)

Points from Discussion:

User Beware:

- K is *ad hoc* but sensible—dependent on priors and assumptions
- Clusters may not be real populations—think about a low migration species on a continuous plane