1. (a)

$$P(D \mid P) = \frac{P(P \mid D)P(D)}{P(P \mid D)P(D) + P(P \mid D^c)P(D^c)}$$
$$= \frac{p_0 p_1}{p_0 p_1 + (1-p_0)p_1}$$

(b) Here is my R code:

```
## part (b) - Write a function to calculate credible intervals
drawTheta <- function(b,X,n,a) {
    p <- rbeta(3,a+X,a+n-X)
    return(p[1]*p[2]/(p[1]*p[2]+(1-p[1])*p[3]))
}
bootThetaCI <- function(b,X,n,a) {
    vals<-sort(unlist(mclapply(1:b,drawTheta,X<-X,n<-n,a<-a,mc.cores=4)))
    return(c(vals[ceiling(b*.025)],vals[ceiling(b*.975)]))
```

(c)/(d) I have produced the results for $a = .5, 1 \,\&\, 2$ in one table, each time with confidence intervals based off of $10,000$ iterations.

	n0=n1=n2	(x0,x1,x2)	a=.5 95% CI	a=1 95% CI	a=2 95% CI
1	20.00	(2,18,2)	(0.121,0.878)	(0.142,0.855)	(0.177,0.819)
2	20.00	(10,18,0)	(0.86,1.00)	(0.805,0.999)	(0.737,0.989)
3	80.00	(20,60,20)	(0.344,0.653)	(0.345,0.653)	(0.348,0.650)
4	80.00	(40,72,8)	(0.811,0.954)	(0.805,0.950)	(0.792,0.943)

We can observe the sensitivity to a by noting that the confidence intervals get pulled towards .5 as a increases. The distinction isn't huge though; with $n = 20$ we note the bounds change by as much as .1, but with $n = 80$, the change is at most near .01.

(e) I have simulated the portion of the time the true θ (th) was below the CI lower bound (LB) and the portion the time it was above the upper bound (UB) for each a and the given p. The simulation drew the X $1,000$ times, and for each of these generated a confidence interval based on another $1,000$ random draws based on the given X.

	(p0,p1,p2)	a=0.5 th<LB	a=0.5 UB<th	a=1 th<LB	a=1 UB<th	a=2 th<LB	a=2 UB<th
1	(0.5,0.5,0.5)	0.03	0.02	0.03	0.02	0.02	0.02
2	(0.2,0.6,0.7)	0.02	0.03	0.04	0.01	0.06	0.00
3	(0.5,0.1,0.9)	0.03	0.03	0.03	0.01	0.06	0.00
4	(0.95,0.95,0.05)	0.00	0.01	0.00	0.05	0.00	0.25
5	(0.2,0.1,0.9)	0.02	0.03	0.03	0.01	0.12	0.00

In general, $a = .05$ seems to have proper, or very close to proper, frequentist coverage properties. However, in the cases where there are p near 0 or 1, this seems to be less true, with the worst case being $p_0 = .95, p_1 = .95, p_2 = 0$, where we never see θ fall below the CI. For different values of a though, we seem to consistently see either higher or lower error rates on each side of the CI.

2. i) We can derive $\mathrm{E}(\theta_j)$ as such:

$$\mathrm{E}(\theta_j) = \int \cdots \int_{\sum_{i=1}^k \theta_i = 1} \theta_i \frac{\Gamma(\sum_{i=1}^k \alpha_i)}{\prod_{i=1}^k \Gamma(\alpha_i)} \left(\prod_{i=1}^k \theta_i^{\alpha_i - 1} \right) d\theta_1 ... d\theta_k$$

$$\mathrm{E}(\theta_j) = \frac{\Gamma(1+\alpha_j)\Gamma(\sum_{i=1}^k \alpha_i)}{\Gamma(\alpha_j)\Gamma(1+\sum_{i=1}^k \alpha_i)} \int \cdots \int_{\sum_{i=1}^k \theta_i = 1} \frac{\Gamma(1+\sum_{i=1}^k \alpha_i)}{\Gamma(1+\alpha_j)\prod_{i \neq j} \Gamma(\alpha_i)} \left(\theta_i^{\alpha_i} \prod_{i \neq j} \theta_i^{\alpha_i - 1} \right) d\theta_1 ... d\theta_k$$

$$\mathrm{E}(\theta_j) = \frac{\Gamma(1+\alpha_j)\Gamma(\sum_{i=1}^k \alpha_i)}{\Gamma(\alpha_j)\Gamma(1+\sum_{i=1}^k \alpha_i)}$$

$$\mathrm{E}(\theta_j) = \frac{\alpha_j}{\sum_{i=1}^k \alpha_i}$$

Then, as a first step to get the variance,

$$\mathrm{E}(\theta_j^2) = \int \cdots \int_{\sum_{i=1}^k \theta_i = 1} \theta_i^2 \frac{\Gamma(\sum_{i=1}^k \alpha_i)}{\prod_{i=1}^k \Gamma(\alpha_i)} \left(\prod_{i=1}^k \theta_i^{\alpha_i - 1} \right) d\theta_1 ... d\theta_k$$

$$\mathrm{E}(\theta_j^2) = \frac{\Gamma(2+\alpha_j)\Gamma(\sum_{i=1}^k \alpha_i)}{\Gamma(\alpha_j)\Gamma(2+\sum_{i=1}^k \alpha_i)} \int \cdots \int_{\sum_{i=1}^k \theta_i = 1} \frac{\Gamma(2+\sum_{i=1}^k \alpha_i)}{\Gamma(2+\alpha_j)\prod_{i \neq j} \Gamma(\alpha_i)} \left(\theta_i^{1+\alpha_i} \prod_{i \neq j} \theta_i^{\alpha_i - 1} \right) d\theta_1 ... d\theta_k$$

$$\mathrm{E}(\theta_j^2) = \frac{\Gamma(2+\alpha_j)\Gamma(\sum_{i=1}^k \alpha_i)}{\Gamma(\alpha_j)\Gamma(2+\sum_{i=1}^k \alpha_i)}$$

$$\mathrm{E}(\theta_j^2) = \frac{\alpha_j(\alpha_j+1)}{\left(\sum_{i=1}^k \alpha_i + 1\right)\sum_{i=1}^k \alpha_i}$$

Which we can use to get

$$\mathrm{Var}(\theta_j) = \mathrm{E}(\theta_j^2) - \mathrm{E}(\theta_j)^2$$

$$\mathrm{Var}(\theta_j) = \frac{\alpha_j^2 + \alpha_j}{\left(\sum_{i=1}^k \alpha_i + 1\right)\sum_{i=1}^k \alpha_i} - \frac{\alpha_j^2}{\left(\sum_{i=1}^k \alpha_i\right)^2}$$

$$\mathrm{Var}(\theta_j) = \frac{\alpha_j^2\left(\sum_{i=1}^k \alpha_i\right) + \alpha_j\left(\sum_{i=1}^k \alpha_i\right) - \alpha_j^2\left(\sum_{i=1}^k \alpha_i\right) - \alpha_j^2}{\left(\sum_{i=1}^k \alpha_i + 1\right)\left(\sum_{i=1}^k \alpha_i\right)^2}$$

$$\mathrm{Var}(\theta_j) = \frac{\alpha_j\left(\sum_{i=1}^k \alpha_i\right) - \alpha_j^2}{\left(\sum_{i=1}^k \alpha_i + 1\right)\left(\sum_{i=1}^k \alpha_i\right)^2}$$

Finally,

$$\mathrm{Cov}(\theta_j, \theta_i) = \mathrm{E}(\theta_j\theta_i) - \mathrm{E}(\theta_j)\mathrm{E}(\theta_i)$$

$$\mathrm{Cov}(\theta_j, \theta_i) = \frac{\Gamma(1+\alpha_j)\Gamma(1+\alpha_i)\Gamma(\sum_{i=1}^k \alpha_i)}{\Gamma(\alpha_i)\Gamma(\alpha_j)\Gamma(2+\sum_{i=1}^k \alpha_i)} - \frac{\alpha_j\alpha_i}{\left(\sum_{i=1}^k \alpha_i\right)^2}$$

$$\mathrm{Cov}(\theta_j, \theta_i) = \frac{\alpha_j\alpha_i}{\left(\sum_{i=1}^k \alpha_i + 1\right)\sum_{i=1}^k \alpha_i} - \frac{\alpha_j\alpha_i}{\left(\sum_{i=1}^k \alpha_i\right)^2}$$

$$\mathrm{Cov}(\theta_j, \theta_i) = \frac{-\alpha_j\alpha_i}{\left(\sum_{i=1}^k \alpha_i + 1\right)\left(\sum_{i=1}^k \alpha_i\right)^2}$$

Qualitatively, as any particular α_i grows with the rest held constant, the probability of high values for θ_i grows. When all the α_j grow together, the distribution becomes more concentrated around the center where all θ_j are close together, and when all the α_j shrink together, the distribution becomes clustered around high values for one of the θ_j with the rest close to 0

ii) Let, for each $1 \le j \le k$, let $|\{j; X_j = k\}| = n_j$. Then,

a)

$$
\begin{aligned}
P(\theta \mid X_1, ..., X_n) &\propto P(X_1, ..., X_n \mid \theta) P(\theta) \\
P(\theta \mid X_1, ..., X_n) &\propto \left(\prod_{i=1}^{k} \theta_i^{n_i}\right)\left(\prod_{i=1}^{k} \theta_i^{\alpha_i - 1}\right) \\
P(\theta \mid X_1, ..., X_n) &\propto \prod_{i=1}^{k} \theta_i^{n_i + \alpha_i - 1}
\end{aligned}
$$

So $\theta \mid X_1, ..., X_n$ has a Dirichlet distribution with parameters $(n_1 + \alpha_1, ..., n_k + \alpha_k)$

b)

$$
\begin{aligned}
P(X_{n+1} = j \mid X_1, ..., X_n) &= \int \cdots \int_{\sum_{i=1}^k \theta_i = 1} P(X_{n+1} = j \mid X_1, ..., X_n, \theta) P(\theta \mid X_1, ..., X_n) d\theta_1 ... d\theta_k \\
P(X_{n+1} = j \mid X_1, ..., X_n) &= \frac{\Gamma(\sum_{i=1}^k n_i + \alpha_i)}{\prod_{i=1}^k \Gamma(n_i + \alpha_i)} \int \cdots \int_{\sum_{i=1}^k \theta_i = 1} \theta_j \prod_{i=1}^{k} \theta_i^{n_i + \alpha_i - 1} d\theta_1 ... d\theta_k \\
P(X_{n+1} = j \mid X_1, ..., X_n) &= \frac{n_j + \alpha_j}{\sum_{i=1}^k n_i + \alpha_i}
\end{aligned}
$$

So we can conclude that $X_{n+1} = j \mid X_1, ..., X_n$ has a multinomial distribution with probability vector

$$
p = \left(\frac{n_1 + \alpha_1}{\sum_{i=1}^k n_i + \alpha_i}, ..., \frac{n_k + \alpha_k}{\sum_{i=1}^k n_i + \alpha_i}\right)
$$

3. If X is Poisson with mean θ, then $P(X = x \mid \theta) \propto \theta^x e^{-\theta}$, which makes a prior that is something of the sort $P(\theta) \propto \theta^a e^{b\theta}$, for some hyperparameters a, b, a natural choice. This is satistified by the gamma distribution, which is the conjugate prior. If we parameterize it the usual way with $P(\theta) \propto \theta^{\alpha-1} e^{-\beta\theta}$, we get the posterior

$$
\begin{aligned}
P(\theta \mid X) &\propto P(X \mid \theta) P(\theta) \\
P(\theta \mid X) &\propto \theta^{\sum x_i} e^{-n\theta} \theta^{\alpha-1} e^{-\beta\theta} \\
P(\theta \mid X) &\propto \theta^{\sum x_i + \alpha - 1} e^{-(\beta+n)\theta}
\end{aligned}
$$

Giving us a gamma posterior with parameters $(\sum x_i + \alpha, \beta + n)$.

We can get the Jeffreys prior from the Fisher information:

$$
\begin{aligned}
p(\theta) &\propto \sqrt{I(\theta)} \\
p(\theta) &\propto \sqrt{-\mathrm{E}\left[\frac{d^2}{d\theta^2}\log(f(X \mid \theta))\right]} \\
p(\theta) &\propto \sqrt{-\mathrm{E}\left[\frac{d^2}{d\theta^2} x\log(\theta) - \theta + c\right]} \\
p(\theta) &\propto \sqrt{-\mathrm{E}\left[\frac{d}{d\theta}\frac{x}{\theta} - 1\right]} \\
p(\theta) &\propto \sqrt{-\mathrm{E}\left[-\frac{x}{\theta^2}\right]} \\
p(\theta) &\propto \sqrt{\frac{1}{\theta}} \\
p(\theta) &\propto \theta^{-\frac{1}{2}}
\end{aligned}
$$

With the Jeffreys prior, we get the posteriors

$$
\begin{aligned}
P(\theta \mid x) &\propto P(x \mid \theta)P(\theta) \\
P(\theta \mid x) &\propto \theta^x e^{-\theta}\theta^{-\frac{1}{2}} \\
P(\theta \mid x) &\propto \theta^{x-\frac{1}{2}} e^{-\theta} \\
P(\theta \mid cx) &\propto \theta^{cx-\frac{1}{2}} e^{-\theta}
\end{aligned}
$$

Versus, for the 'scale invariant' prior

$$
\begin{aligned}
P(\theta \mid x) &\propto P(x \mid \theta)P(\theta) \\
P(\theta \mid x) &\propto \theta^x e^{-\theta}\theta^{-1} \\
P(\theta \mid x) &\propto \theta^{x-1} e^{-\theta} \\
P(\theta \mid cx) &\propto \theta^{cx-1} e^{-\theta}
\end{aligned}
$$

The posteriors are similar, but the Jeffreys prior has a mean that is a bit closer to the MLE mean, which probably makes it preferable as far as a noninformative prior. As c grows, they both converge to the MLE.

4. We can derive the Jefferies prior from the Fisher Information:

$$
\begin{aligned}
I(p)_{i,j} &= -\mathrm{E}\left[H\left(\log(f(X \mid p))\right)_{i,j}\right] \\
I(p)_{i,j} &= -\mathrm{E}\left[H\left(c + \sum_{i=1}^{k} x_i \log(p_i)\right)_{i,j}\right] \\
I(p)_{i,j} &= \begin{cases} -\mathrm{E}\left[-\frac{x_i}{p_i^2}\right] & \text{if } i = j \\ 0 & \text{if } i \neq j \end{cases} \\
I(p)_{i,j} &= \begin{cases} \frac{n}{p_i} & \text{if } i = j \\ 0 & \text{if } i \neq j \end{cases}
\end{aligned}
$$

Which makes $|I(p)|^{\frac{1}{2}} = n^{\frac{k}{2}} \prod_{i=1}^{k} p_i^{-\frac{1}{2}}$. We can thus conclude

$$p(\theta) \propto \prod_{i=1}^{k} p_i^{-\frac{1}{2}}$$

Making it a Dirichlet prior with each α set to $\frac{1}{2}$.

5. a) If we let X be the training data, G_j be the testing data, consisting of gene $g_{l,j}$ at locus l for individual j, and P_j denote the population of training sample j, then previously our model was:

$$P(P_j \mid G_j, X) \propto P(P_j)P(G_j \mid P_j, X)$$
$$P(P_j \mid G_j, X) \propto P(P_j) \prod_{l=1}^{24} P(g_{l,j} \mid P_j, X)$$

Where $P(g_{l,j} \mid P_j = k, X)$ was the portion of the training population k which had gene $g_{l,j}$ at locus l. Now however, I implemented a new model, where

$$P(P_j \mid G_j, X, \alpha) \propto P(P_j)P(G_j \mid P_j, X, \alpha)$$
$$P(P_j \mid G_j, X, \alpha) \propto P(P_j) \prod_{l=1}^{24} P(g_{l,j} \mid P_j, X, \alpha)$$

Where $P(g_{l,j} \mid P_j, X, \alpha)$ is the posterior prediction distribution based on all the genes at locus l in population P_j with a uniform α Dirichlet distribution (as in question 2.ii.b). If W is the set of genes at locus l, and n_{w,P_j} is the count from the training sample of gene w at locus l in population P_j, then this quantity is:

$$P(g_{l,j} = v \mid P_j, X, \alpha) = \frac{\alpha + n_{v,P_j}}{\sum_{w \in W} \alpha + n_{w,P_j}}$$

b) Trying out different values of α, I got these empirical error rates

	Value 1	Value 2	Value 3	Value 4	Value 5	Value 6	Value 7	Value 8	Value 9
Alpha Values	0.001	0.01	0.2	0.5	1	2	5	10	100
Error Rate	0.261	0.257	0.231	0.235	0.22	0.216	0.216	0.243	0.299

c) I calculated the likelihood for a given alpha as the outcome which made our training data most likely. In other words:

$$P(G \mid X, \alpha) = \prod_{j=1}^{n} P(G_j \mid X, \alpha)$$
$$P(G \mid X, \alpha) = \prod_{j=1}^{n} \prod_{l=1}^{24} P(g_{l,j} \mid X, \alpha)$$
$$P(G \mid X, \alpha) = \prod_{j=1}^{n} \prod_{l=1}^{24} \sum_{k=1}^{4} P(P_j = k) P(g_{l,j} \mid P_j = k, X, \alpha)$$

I maximized α over this quantity, with $P(g_{l,j} \mid P_j = k, X, \alpha)$ defined as above.

d) When I did this, I found the optimal parameter to be $\alpha = .363$, which had a log likelihood of -13621 and yielded an error rate of .231. This rate is only a bit higher than the best rate observed in part b of .216.

e) Where A=$\{0.025, 0.05, 0.1, 0.2, 0.4, 0.8, 1.6, 3.2, 6.4\}$, I implemented the method as such:

$$
\begin{aligned}
P(P_j \mid G_j, X) &\propto P(P_j)P(G_j \mid P_j, X) \\
P(P_j \mid G_j, X) &\propto \sum_{\alpha \in A} P(\alpha)P(P_j)P(G_j \mid P_j, X, \alpha) \\
P(P_j \mid G_j, X) &\propto \sum_{\alpha \in A} \frac{1}{|A|} P(P_j) \prod_{l=1}^{24} P(g_{l,j} \mid P_j, X, \alpha)
\end{aligned}
$$

This time, I got an error rate of .243, which is higher than the best α by it self by a few percent. This isn't surprising though, we are averaging the probabilities over a few α which we know have higher error rates than the best α we've seen. Of course, it may be that this value is really closer to what α "should" be given the true variation in the data.